Modelo de predicción de usuarios y asignación de flota del sistema de transporte masivo de Bogotá (Transmilenio S.A) para usuarios que se movilizan hacia el Portal El Dorado

Trabajo de grado para optar por el título de Magister en Analítica para la Inteligencia de Negocios

BLANDÓN LOMBANA LINA MARCELA

MORENO SÁNCHEZ MARCO ANTONIO

OCAMPO HERNÁNDEZ MANUELA

DIRECTOR:

LUIS MANUEL PULIDO MORENO

PONTIFICIA UNIVERSIDAD JAVERIANA

FACULTAD DE INGENIERÍA

MAESTRÍA EN ANALÍTICA PARA LA INTELIGENCIA DE NEGOCIOS

BOGOTÁ D.C, 2022

Tabla de contenido

1. ENTENDIMIENTO DE NEGOCIO

1.1 Determinación de objetivos de negocio

1.1.1 Antecedentes

Transmilenio

TransMilenio es un sistema de transporte de tipo BRT (Bus Rapid Transit) que hace parte del sistema de transporte masivo de Bogotá y Soacha. La entidad gestora del sistema de Transmilenio nació en 1999 siendo la Empresa de Transporte del Tercer Milenio S.A: Transmilenio S.A la encargada de coordinar los diferentes actores, planear, gestionar y controlar la prestación del servicio público de transporte masivo urbano de pasajeros, y tiene la responsabilidad de la prestación eficiente y permanente del servicio (TRANSMILENIO S.A, 2022).

La construcción del sistema inició en 1998 y fue inaugurado el 4 de diciembre de 2000. Entró en operación el 18 del mismo mes con las troncales (líneas) de la avenida Caracas y la calle 80. Desde entonces se han abierto nuevas troncales. Forma parte del SITP, junto con los servicios urbano, complementario y especial, que circulan por los barrios y vías principales de la ciudad. De 2001 al 2003 se establecieron proyectos prioritarios para la creación de tres nuevas troncales de transporte masivo: Américas, NQS y Avenida Suba (TRANSMILENIO S.A, 2022).

Transmilenio en cifras (TRANSMILENIO S.A, 2022).:

- Actualmente el sistema cuenta con 1.114,4 Km de vía en troncal en operación, 12 troncales en operación (ver la tabla 1 y la imagen 1), 138 estaciones, 9 portales y 53 patio garajes. Adicionalmente el Sistema troncal tiene 2.364 buses.
- Así mismo el componente zonal cuenta con 7.048 buses zonales, 2400 Kilómetros de cobertura, 38 patio talleres.
- El volumen de pasajeros ha aumentado a través de los años, para el 2016 se registraban 702 millones de abordajes

Troncal	Inicio	Fin	Transferencias	Número de estaciones	Longitud troncal	Longitud pre-troncal	Longitud total
A Caracas	Calle 76	Tercer Milenio	**F** Avenida Jiménez	14	8,0 km		8,0 km
B Autopista Norte	Terminal	Héroes		17	12,0 km		12,0 km
C Avenida Suba	Portal de Suba	San Martín		14	9,5 km	1,5 km	11,0 km
D Calle 80	Portal de la 80	Polo		13	7,5 km	2,5 km	10,0 km
E NQS Central	La Castellana	Tygua - San José	**F** Ricaurte	13	11,5 km		11,5 km
F Avenida Las Américas	Portal de Las Américas	Avenida Jiménez	**A** Avenida Jiménez **E** Ricaurte	18	12,5 km		12,5 km
G NQS Sur	Comuneros	San Mateo		17	13,0 km		13,0 km
H Caracas Sur *(Usme)*	Hospital	Portal de Usme		13	10,0 km	0,25 km	10,25 km
H Caracas Sur *(Tunal)*	Biblioteca	Portal del Tunal	TransMiCable	3	6,75 km		6,75 km
J Eje Ambiental	Museo del Oro	Universidades		3	1,5 km		1,5 km
K Avenida Eldorado	Portal Eldorado	Centro Memoria		13	11,0 km	2,0 km	13,0 km
L Carrera Décima	San Diego	Portal 20 de Julio		10	6,5 km		6,5 km

M Carrera Séptima	Museo Nacional		1	0,5 km	12,0 km	12,5 km

Tabla 1 Características de las troncales del sistema de TransMilenio

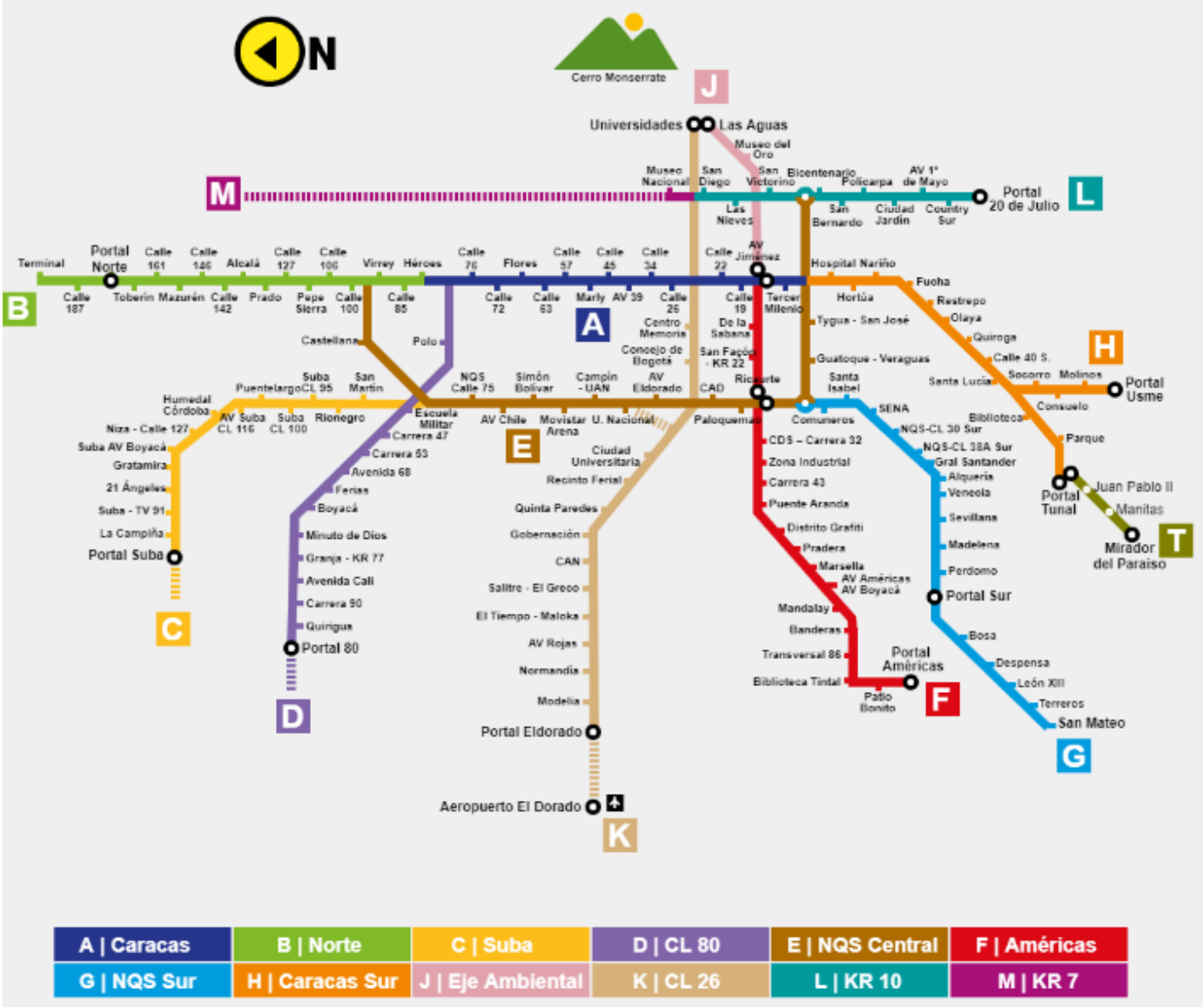

Figura 1 Mapa del sistema de TransMilenio (TransMilenio, 2022)

Troncal calle 26 y portal El Dorado

En el 2012, se inauguró la Troncal Calle 26 (Avenida El dorado), esta troncal nace en la carrera 3a. y termina en el Portal El Dorado el cual está ubicado en el occidente de la ciudad, sobre la Avenida El Dorado entre la Avenida Cali y la transversal 93.

La importancia de este portal radica en que es el punto de llegada y de alimentación de una gran población de la ciudad, el portal El Dorado atiende la tercera localidad más grande de Bogotá en cuanto a habitantes: Engativá, y además la localidad de Fontibón, específicamente atiende directamente a los barrios Santa Cecilia, Los Álamos y sus alrededores por medio de las rutas alimentadoras o complementarias.

En las cercanías del portal se encuentran, el hotel Aloft Bogotá Airport, hotel Movich Buró 26, hotel Habiltel, la sede principal de Carvajal S.A., los laboratorios GlaxoSmithKline, el almacén Hipercentro Corona Dorado y el centro empresarial Connecta, además de varias sedes de varias multinacionales y de oficinas de visados para las embajadas de Australia y Estados Unidos.

Adicional a lo anterior la troncal Calle 26 es un eje vital para la llegada al Aeropuerto Internacional El Dorado Luis Carlos Galán Sarmiento, el cual es el principal aeropuerto de Colombia con un área aproximada de 7 km². Para entender la importancia del transporte vial que alimenta el Aeropuerto cabe resaltar las siguientes cifras y reconocimientos de este:

Es el primer aeropuerto de Latinoamérica en volumen de carga, Es el tercer aeropuerto más importante de América Latina en volumen de pasajeros, después del Aeropuerto Internacional de la Ciudad de México y el Aeropuerto Internacional de São Paulo-Guarulhos, y uno de los más importantes hubs de Sudamérica debido a que posee una posición privilegiada y estratégica, dado que se encuentra en la parte media del continente americano, facilitando su comunicación con todos los continentes (Aeropuertos del mundo, s.f.).

El aeropuerto internacional El Dorado, de Bogotá, fue reconocido por la firma británica Skytrax como la terminal con el mejor personal de Suramérica y obtuvo el premio a la excelencia aeroportuaria por la implementación de protocolos de bioseguridad, además, El Dorado ocupa el puesto 43 en el listado de los mejores aeropuertos del mundo, entre las 500 terminales aéreas evaluadas (Portafolio, 17 de agosto de 2021).

A estos reconocimientos se suma la calificación de cinco estrellas otorgada por Skytrax a El Dorado por la implementación de las medidas de seguridad durante las épocas más críticas de la emergencia sanitaria del coronavirus. Para el 2021 el aeropuerto El Dorado tuvo un volumen de 22.091.102 pasajeros (UAEAC, 07 de febrero de 2021), la recuperación frente al covid ha sido tan buena que para el 2022 de acuerdo con cifras de la Aerocivil, entre enero y febrero de se han transportado poco más de cinco millones de pasajeros en Colombia, es decir, 10,2% más que en el mismo periodo de 2020, antes de que iniciara la pandemia (Valora Analitik, 2022).

Rutas del portal El Dorado

A continuación, se muestran las rutas que salen y llegan al portal El Dorado y sus horarios:

- **1** Universidades - Portal Eldorado: lunes a viernes 04:30 AM - 11:00 PM – sábados 05:00 AM - 11:00 PM – domingos y festivos 05:30 AM - 10:00 PM
- **1** Portal El Dorado – Universidades: lunes a viernes 04:30 AM - 11:00 PM – sábados 05:00 AM - 11:00 PM – domingos y festivos 05:30 AM - 10:00 PM

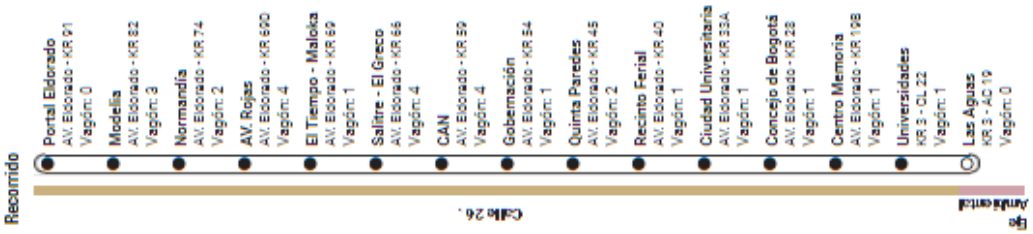

Figura 2 Ruta 1

- **G43** Portal El Dorado - San Mateo: lunes a sábado 05:00 AM - 11:00 PM
- **K43** San Mateo - Portal El Dorado: lunes a viernes 04:00 AM - 11:00 PM - sábados 04:30 AM - 11:00 PM

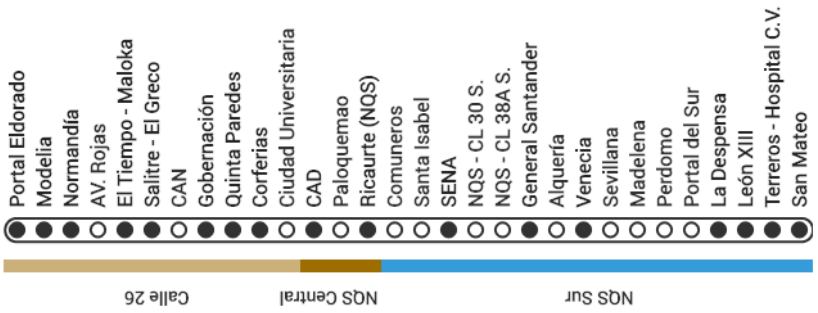

Figura 3 Ruta G43 – K43

- **B16** Portal El Dorado – Terminal: lunes a viernes 05:30 AM - 10:30 PM - sábados 05:30 AM - 10:00 PM
- **K16** Terminal - Portal El Dorado: lunes a viernes 05:30 AM - 10:30 PM - sábados 05:30 AM - 10:00 PM

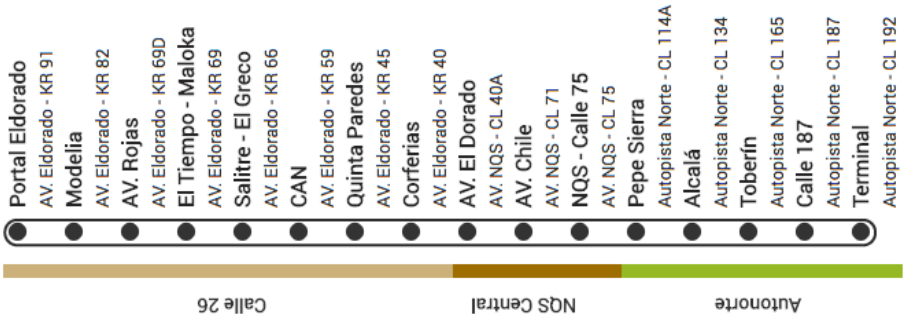

Figura 4 Ruta B16 – K16

- **B23** Portal El Dorado – Alcalá: lunes a viernes 05:00 AM - 10:00 PM - sábados 05:30 AM - 10:00 PM
- **K23** Alcalá - Portal El Dorado: lunes a viernes 05:30 AM - 10:00 PM

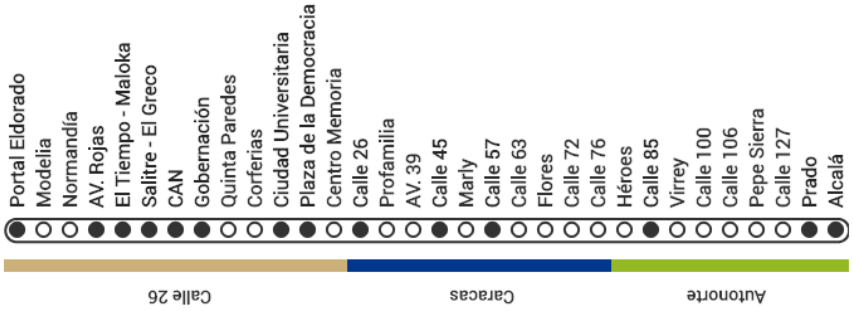

Figura 5 Ruta B23 – K23

- **M86** Portal Dorado Aeropuerto - Hacienda Santa Bárbara: lunes a viernes 04:30 AM - 10:00 PM - sábados 05:00 AM - 10:00 PM domingo y festivos 06:00 AM - 09:00 PM
- **K86** Fundación Santa Fe - Aeropuerto El Dorado: lunes a viernes 05:30 AM - 11:00 PM - sábados 06:00 AM - 11:00 PM – domingos y festivos 07:00 AM - 10:00 PM

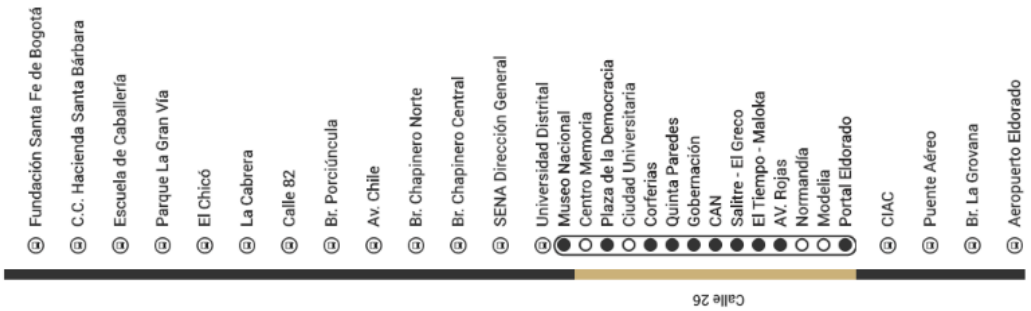

Figura 6 Ruta M86 – K86

- **K86** Fundación Santa Fe - Aeropuerto El Dorado (Ciclovía): Domingo y festivos 07:00 AM - 02:00 PM

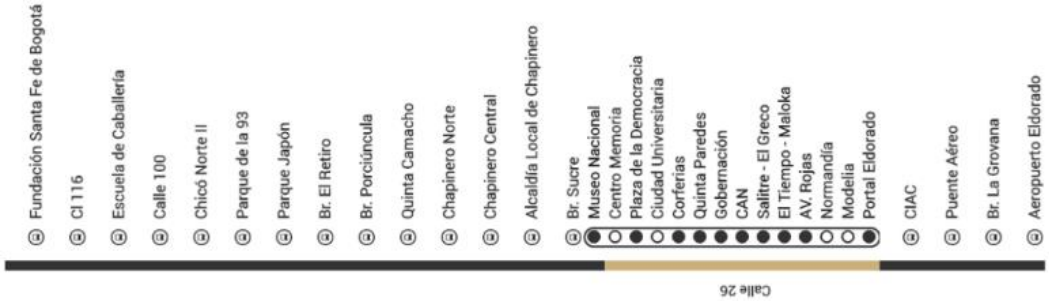

Figura 7 Ruta K86 ciclovía

- **L10** Portal Eldorado - Portal 20 de Julio: lunes a sábados 05:00 AM - 11:00 PM – domingos y festivos 05:30 AM - 10:00 PM
- **K10** Portal 20 de Julio - Portal El Dorado: lunes a viernes 04:30 AM - 11:00 PM – sábados 05:00 AM - 11:00 PM – domingos y festivos 05:30 AM - 10:00 PM

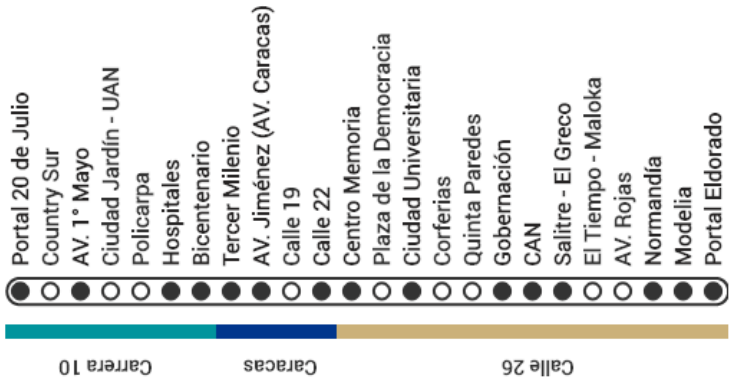

Figura 8 Ruta L10 – K10

- **H54** Portal Eldorado - Portal Usme: lunes a viernes 05:30 AM - 10:30 PM
- **K54** Portal Usme - Portal Eldorado: lunes a viernes 05:30 AM - 10:30 PM

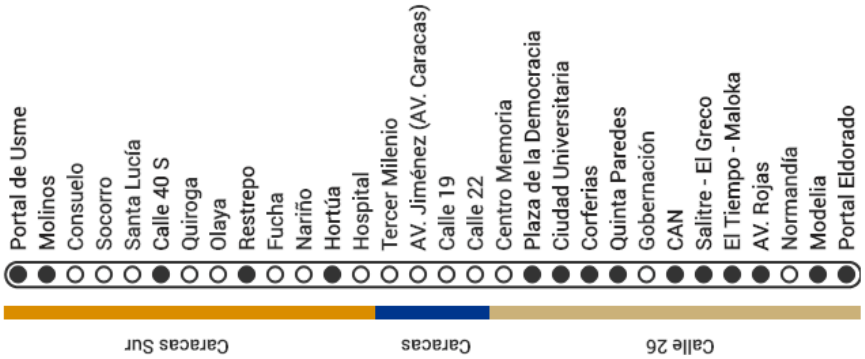

Figura 9 Ruta H54 – K54

La gestión de TransMilenio S.A como compañía prestadora de servicios de transporte masivo es vital para la movilidad de la ciudad, por ese motivo es importante administrar de forma correcta las rutas que entran y salen en los puntos más importantes del sistema, especialmente por su ubicación estratégica las rutas del portal El Dorado; para entender si se debe mejorar o si se está cumpliendo con ese objetivo se realizan encuestas que permitan conocer la percepción de los ciudadanos frente a la movilidad, algo que mejora esta percepción es la disminución del tiempo de permanencia de los usuarios en el sistema, ya que esto indica que los usuarios están llegando a sus destinos en un menor tiempo, cuando este objetivo no se cumple se disparan indicadores de insatisfacción del cliente, esto se puede ver en la tabla 2 donde se ve una relación proporcional frente a la percepción de que los tiempos son mayores con la percepción de que el servicio empeoró: a mayor tiempo de viaje mayor sensación de que el servicio está siendo malo.

Con base en lo anterior, TransMilenio S.A reconoce la importancia de buscar estrategias en pro de lograr dichas disminuciones de tiempos de permanencia de modo que la calidad del servicio pueda percibirse como mejorada frente a los usuarios.

Año	% de encuestados que piensan que sus trayectos habituales duran más tiempo en Bogotá	% de encuestados que piensan TransMilenio ha empeorado en Bogotá
2017	65 %	60 %
2018	61 %	55 %
2019	55 %	47 %

Tabla 2 Percepción de los ciudadanos frente al tiempo de viaje y del servicio de TransMilenio para el 2017, 2018 y 2019.

Horarios de operación de TransMilenio

- Lunes - viernes: 04:00-23:00
- Sábado: 04:00-23:00
- Domingo - festivos: 04:00-22:00

Eventos que afectaron la operación de TransMilenio en el 2017:

- Respecto al 2017 se encuentran los siguientes fenómenos que podrían afectar el comportamiento de un día normal de operación de Transmilenio:
  - Domingos y Festivos: Estos días la operación de TransMilenio tiene una operación disminuida en horario y con rutas diferentes a las que operan de lunes a viernes, adicional a esto la Calle 26 comparte en sus vías unos carriles para la ciclovía de
  - Festivos del 2017:
    - 9 de enero
    - 20 de marzo
    - 13 y 14 de abril
    - 01 y 29 de mayo
    - 19 y 26 de junio
    - 03 y 20 de julio
    - 07 y 21 de agosto

- 16 de octubre
- 06 y 13 de noviembre
- 08 y 25 de diciembre

- Temporadas vacacionales:
  - **01 de enero al 23 de enero:** Primer ciclo de vacaciones calendario escolar:
  - **10 de abril al 16 de abril:** Semana santa:
  - **16 de junio al 3 de julio:** Vacaciones de mitad de año calendario escolar
  - **9 de octubre al 13 de octubre:** Vacaciones semana de receso escolar
  - **23 de diciembre al 31 de diciembre:** Fin de año y navidad
- Principales días de Paro o manifestaciones:
  - **22 de enero:** Afectación principalmente del centro de la ciudad
  - **28 de marzo al 15 de junio:** Paro del docente, este periodo de tiempo tuvo marchas por toda la ciudad o manifestaciones como cacerolazos en las vías durante distintos días, cabe resaltar que este paro tuvo una especial afectación para la calle 26 ya que sobre esta vía se encuentra la secretaria de educación; las marchas que más resaltan por su tamaño fueron en las siguientes fechas:
    - 28 de febrero
    - 16 de mayo
    - 09 de junio
    - 15 de junio
  - **01 de abril:** Afectación principalmente del centro de la ciudad
  - **10 de mayo:** Paro de taxistas en toda Bogotá
  - **24 de agosto:** Afectación principalmente del centro de la ciudad, la marcha se dio por recorte de presupuesto a ciencia por lo que afecto también la calle 26 donde se encuentra la secretaria de Educación
  - **20 de septiembre de 2017 al 9 de noviembre:** Aunque no fue una manifestación que genero todos los días bloqueos a vías de la ciudad, afecto la operación del Aeropuerto El Dorado ya que fue una huelga de los pilotos de Avianca, ocasionando la cancelación de muchos vuelos ya que esta empresa es una de principales prestadores de servicios aéreos, como consecuencia para el sistema de Transmilenio se esperó la disminución de la recurrencia hasta al portal el Dorado, esta huelga paso por 3 etapas (U. Externado, s. f):
    - 20 de septiembre: más de 100 pilotos salen a huelga, Avianca sin un plan b debe cancelar sus vuelos.
    - 30 de septiembre: Avianca contrata pilotos extranjeros tratando de activar los vuelos
    - 09 de octubre 60 pilotos de la huelga vuelven a trabajar ayudando a la reactivación de los vuelos.
  - **27 de septiembre:** Llamado “Paro del sur”, tuvo afectación sobre los portales de Usme, el Tunal, y Del Sur de Transmilenio, además de afectar las vías principales de la zona sur y algunas estaciones de la calle 26.
  - **04 de octubre:** Afectación principalmente del centro de la ciudad, la marcha se dio por recorte de presupuesto de la educación nacional por lo que afecto también la calle 26 donde se encuentra la secretaria de Educación
  - **23 de octubre:** Paro de taxistas en toda Bogotá
  - **10 de noviembre:** Cierre de la estación y de la Calle 72
  - **21 de noviembre:** Marchas generales por la ciudad
  - **25 de noviembre 2017:** Marcha día Internacional de la No Violencia contra la Mujer
- Época de lluvias en Bogotá: Este factor es importante ya que ralentiza el flujo vehicular de la ciudad, la época de lluvia en Bogotá es de casi 9 meses del 16 de marzo a 9 de diciembre, en esta época se tiene una probabilidad de más del 50% cada día de que haya una precipitación, sin embargo, se destacan 2 periodos donde las lluvias son

intensas: de marzo a mayo y de septiembre a noviembre ver inundaciones en distintos sectores de la ciudad en estos periodos (Weather Spark, 2022).
    - Días sin carro: jueves 2 de febrero

1.1.2 Objetivos de negocio

Disminuir los tiempos de permanencia de los pasajeros en el sistema que se dirigen hacia el Portal El Dorado, entendiendo el comportamiento del flujo de personas por jornadas y días en las troncales de Transmilenio y la afectación de variables externas para realizar la asignación de la flota del sistema dirigida hacia esta troncal y portal, con el fin de mejorar la calidad del servicio.

1.1.3 Criterios de éxito del negocio

Para la compañía es de gran importancia comprender el comportamiento de las entradas y salidas de las personas a las estaciones que componen las rutas que se dirigen al Portal del Dorado, los tiempos de permanencia dentro del sistema son fundamentales al momento de evaluar la calidad del servicio, por ello, se considera un proyecto exitoso si se logra disminuir este tiempo entendiendo la variación del flujo de personas según los días de la semana y sus diferentes jornadas, además, como factores externos tienen influencia en este comportamiento con el fin de asignar y despachar la flota óptima de servicios en esta troncal.

1.2 Evaluación de la situación

1.2.1 Inventario de recursos, requerimientos, suposiciones, y restricciones

- **Recursos:**

TIPO DE RECURSO	RECURSO	CANTIDAD / DESCRIPCIÓN
Personal	Consultores de analítica	3 personas
	Experto en el negocio	1 persona
	Asesores de trabajo de grado	1 persona
Datos	Datos de entradas a las estaciones 2017	12 bases de datos
	Datos de salidas de las estaciones 2017	12 bases de datos
	Histórico de lluvias 2017	1 base de datos
	Histórico de paros 2017	1 bases de datos
	Rutas del Portal el Dorado 2017	1 bases de datos
Recursos computacionales	Computadores de 16 GB de RAM, 512 GB SSD, 1 TB de almacenamiento, Procesador Intel 7- 8ª generación	3
Software	Excel	Haciendo uso de la licencia Office 365 institucional
	Word	
	Power Point	
	Anaconda con Python 3.9	NA
	NetLogo 6.2	NA

Tabla 3 Recursos de proyecto

- **Requerimientos:**
    - Fecha máxima de ejecución del proyecto: 08 de julio de 2022.
    - El modelo deberá estar alineado al objetivo de negocio y al de analítica.
    - La calidad del modelo se deberá probar con el histórico de entradas y salidas de las estaciones del 2017 entregado por la empresa.
    - Se deberá dar cumplimiento a la siguiente legislación:
        - Ley de transparencia 1712 de 2014
        - Ley de Habeas Data 1266 de 2008
        - Ley de protección de datos 1581 de 2012
    - El proyecto cumplirá además con el reglamento estudiantil de la Pontificia Universidad Javeriana y demás normas y estatutos que haya a lugar.
    - La simulación deberá replicar el comportamiento del flujo de personas según información entregada por la empresa, toda la información se trabajará solo con fines académicos.

- El alcance del proyecto se limitará al análisis de las rutas que se dirigen al Portal del Dorado y que estaban vigentes en el año 2017, dado que es la información histórica que se tiene, sin embargo, se podría ajustar para que funcione bajo las rutas vigentes.
- Los entregables del proyecto deberán ser los siguientes:
  - Documento Word del proyecto con la metodología CRISP-DM
  - Bases de datos originales y modificadas
  - Códigos fuente de los desarrollos realizados del modelo
  - Simulación final en Netlogo
  - Instructivo de simulación
  - Resultados de la simulación y resultados de tiempos de espera promedio, archivos xlsx.

- Suposiciones:
  - La información compartida por la compañía corresponde a las entradas y salidas reales de las estaciones.
  - No se tienen en cuenta los trasbordos realizados en estaciones intermedias.
  - Las lluvias tienen influencia en el flujo de personas que ingresan al sistema.
  - No se tienen en cuenta los ingresos no pagos.
  - No se tienen en cuenta las salidas no registradas en un torniquete.
- Restricciones:
  - El tiempo del proyecto será de 3 meses aproximadamente, en el que se desarrollarán todas las fases comprometidas, durante este tiempo se programan reuniones cada 15 días para revisar el estado del proyecto.
  - Solo se cuenta con 1 año de información que es el 2017, si se quisiera analizar el comportamiento hoy en día, se deberán actualizar rutas y estaciones nuevas y descontinuadas.

1.2.2 Riesgos y contingencias

A continuación, se presenta el análisis de riesgos realizado:

CATEGORÍA	RIESGO	MAGNITUD DE DAÑO (1-4)	PROBABILIDAD DE OCURRENCIA (1-4)	CALIFICACIÓN DE RIESGO	PLAN DE CONTIGENCIA
Datos	Recibir información de mala calidad, con datos faltantes, registros duplicados o con errores	4	2	Medio	-Realizar un acompañamiento en la extracción de datos. -Realizar el tratamiento de datos requerido para evitar imprecisión en el modelo.
	Recibir datos que no estén alineados con el objetivo de negocio y de analítica.	4	1	Bajo	-Definir con claridad el objetivo de negocio y de analítica, solicitando aprobación por parte del negocio -Solicitar los datos necesarios para alcanzar estos objetivos.
Negocio	Perder comunicación con alguno de los involucrados en el proyecto. Validez de la SBA	4	1	Bajo	-Establecer un plan de seguimiento desde el inicio para identificar quienes y en qué momento se deberán involucrar en el proyecto. -Realizar validación del comportamiento vs. La realidad bajo métricas estadísticas.
Técnico	No contar con los recursos computacionales necesarios para alcanzar el objetivo de analítica	4	1	Bajo	-En cualquier caso, se podrá solicitar un servidor de la universidad que cuente con las características necesarias para realizar todo el proceso de analítica.

Tabla 4 Análisis de riesgos

1.2.3 Terminología

- **Agente:** Un agente constituye la unidad elemental e indivisible de un sistema, por ejemplo, una persona, un grupo de personas, un hogar, o una organización. Los agentes, por definición, tienen un propósito y responde decidiendo o actuando conforme a las reglas que rigen su comportamiento. Éstas diferencian a cada agente, definen las interrelaciones entre los agentes y el ambiente, y establecen la secuencia de acciones con el tiempo.

- **Comportamiento:** Este término hace referencia a la forma en la que se comportan las personas ante determinadas situaciones, personas o eventos, pero siempre teniendo en cuenta la influencia que existe en relación con los factores sociales o ambientales.
- **Flujo de entradas**: Cantidad de personas que ingresan a las estaciones de Transmilenio por las diferentes entradas incluso las que sean para personas con movilidad reducida.
- **Flujo de salidas:** Cantidad de personas que salen a las estaciones de Transmilenio por las diferentes entradas incluso las que sean para personas con movilidad reducida.
- **Factores externos:** Sucesos ambientales o sociales que puedan afectar el comportamiento de las entradas y salidas de personas a las estaciones de Transmilenio.
- **Parámetros:** Valores que definen y caracterizan cada entorno de simulación, basados en esos parámetros se analizan los resultados y las interacciones que surjan.
- **Parcela o parche:** Es la representación del entorno de la simulación, también funcionan como medidas para el análisis de los resultados.
- **Simulación Basada en Agentes (SBA):** Es un tipo de modelo computacional que permite la simulación de acciones e interacciones de individuos autónomos dentro de un entorno, y permite determinar qué efectos producen en el conjunto del sistema, es un intento de recrear y predecir las acciones de fenómenos complejos. ElKady, S. y Abdelsalam, H. (2015)
- **Tiempo en sistema o de proceso:** Tiempo total que pasan los pasajeros desde el momento en que el usuario ingresa al sistema hasta que llega a su destino y sale de la estación.

1.2.4 Costos y beneficios

De acuerdo con los informes de gestión que se encuentran disponibles en la página oficial de Transmilenio, una de las metas es la inversión para la expansión y mejora en la calidad del servicio, esto implica que las estrategias en expansión deberán ser reflejadas en dos frentes, uno en el que se pueda ampliar la flota del sistema de acuerdo con la demanda y por otro lado en la calidad, que se pueda disminuir el tiempo de permanencia de los usuarios en el sistema.

Las cifras de quejas para el sistema Transmilenio representan el 9% aproximadamente, y al año se reciben alrededor de 500.000 PQRs, para un total de 45.000 quejas anuales por algunas irregularidades y problemas en la prestación del servicio, muchas de las cuales referentes a los tiempos y daños presentados en los buses. Significativamente esto se representa en una carga administrativa para la entidad, por lo que ha tenido que utilizar recursos financieros en contratación de personal legal y administrativo que den respuesta a cada uno de estos requerimientos.

Estos costos y gastos administrativos han representado la contratación de 60 personas adicionales certificados en gestión administrativa y calidad del servicio a través del SENA. En el año 2017 significó una inversión cercana a 800 millones COP únicamente en nómina, para dar gestión en los tiempos correspondientes y no tener sanciones y multas por las distintas entidades reguladoras del sistema. Trayendo estos mismos valores al 2022 con inflación y aumentos de salarios esta cifra sería alrededor de 1.100 millones COP.

Por lo tanto, si al prestar un mejor servicio para el usuario disminuyendo los tiempos de permanencia en el sistema en al menos los trayectos que finalizan en el Portal El Dorado, se redujera al menos un 1% las quejas frente al sistema, se podría disminuir el costo administrativo en al menos 1 persona que pudiese trabajar en mejoras e implementación de estrategias para las demás PQRs.

1.3 Determinación de objetivos de analítica

1.3.1 Objetivos de analítica

A partir del entendimiento del negocio se contemplan los siguientes objetivos de analítica:

- Diseñar un modelo con capacidad de predecir el número de usuarios que ingresan a cada una de las estaciones que tienen conexión para llegar al Portal El Dorado en intervalos de tiempo.
- Simular el comportamiento del flujo de personas en las diferentes estaciones del sistema de transporte masivo de Bogotá Transmilenio de las rutas que se dirigen hacia el Portal el Dorado según el día de la semana y las jornadas.

1.3.2 Criterios de éxito de analítica

Se establecen los criterios de analítica los siguientes:

- ✓ Entregar el pronóstico de cantidad de personas que habrá en cada estación de las rutas que se dirigen hacia el Portal el Dorado según el día de la semana y la jornada, teniendo en cuenta las métricas de evaluación del modelo, las cuales son:
  - **RMSE:** Es la desviación estándar de los valores residuales (errores de predicción). Los valores residuales son una medida de la distancia de los puntos de datos de la línea de regresión; RMSE es una medida de cuál es el nivel de dispersión de estos valores residuales. En otras palabras, le indica el nivel de concentración de los datos en la línea de mejor ajuste (Oracle).

$$RMSE = \sqrt{\sum \frac{(y_i - \hat{y}_i)^2}{T}}$$

  - **MAE:** Mide la magnitud promedio de los errores en un conjunto de predicciones, sin considerar su dirección. Es el promedio sobre la muestra de prueba de las diferencias absolutas entre la predicción y la observación real donde todas las diferencias individuales tienen el mismo peso.

$$MAE = \sum \frac{|y_i - \hat{y}_i|}{T}$$

  - **MASE:** Es una medida para determinar la efectividad de los pronósticos generados a través de un algoritmo al comparar los pronósticos con el resultado de un enfoque de pronóstico ingenuo.

$$MASE = \frac{MAE}{MAE_{naive}}$$

- ✓ Entregar los resultados de los tiempos en el sistema de diferentes escenarios (30 ejecuciones).
- ✓ Entregar los resultados de la cantidad de viajes generadas por día por ruta según los tiempos de espera promedio de diferentes escenarios (30 ejecuciones).
- ✓ Entregar la cantidad de buses necesarios por día por ruta para cumplir con los viajes requeridos.

1.4 Planeación de proyecto

1.4.1 Plan de proyecto

El desarrollo del plan del proyecto se realizará bajo la metodología CRISP-DM, se establecen diferentes etapas donde la primera será el entendimiento del negocio, establecimiento de objetivos y conocimiento de las diferentes problemáticas y expectativas, en etapas posteriores se realizará la recolección de información, análisis, limpieza y estructuración de esta para la ejecución del modelo y evaluación de resultados.

Para ello se establecen reuniones periódicas quincenales con el encargado de la compañía, en estas reuniones se resuelven dudas relacionadas al negocio, se aterrizan objetivos y expectativas del proyecto, se identifican requerimientos, supuestos y premisas. A continuación, se presenta el plan de trabajo y las diferentes etapas, desglosando una a una las actividades y sus tiempos establecidos para el desarrollo, todo graficado en un diagrama de Gantt.

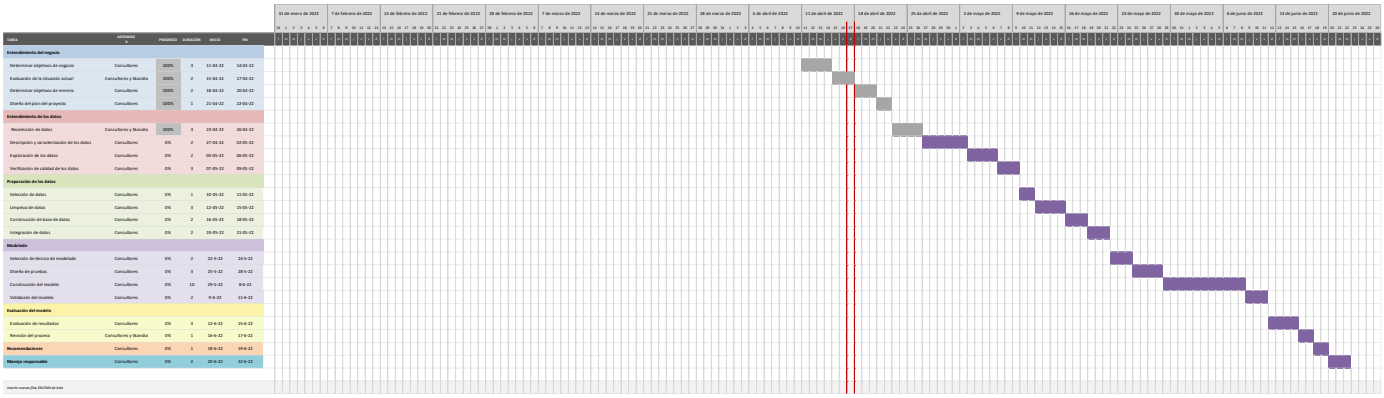

Tabla 5 Plan de proyecto

1.4.2 Evaluación inicial de herramientas y técnicas

Herramientas:

- **Nombre:** Python
- **Descripción:** Python es un lenguaje de programación interpretado que busca desarrollar una sintaxis que priorice la legibilidad del código. Este lenguaje de programación es conocido como multiparadigma ya que soporta diferentes orientaciones. En Python podrás orientar el código a objetos, a programación imperativa y funcional.
- **Versión:** 3.9
- **Editor de texto:** Júpiter Notebook
- **Suite de código abierto:** Anaconda

Según el índice de popularidad del lenguaje de programación, por sus siglas en inglés PYPL, Python es la herramienta más popular para el desarrollo de modelos de machine learning (PYPL,2022).

Herramientas:

- **Nombre:** R
- **Descripción:** Este entorno es comúnmente utilizado para la computación estadística y gráfica, ya que dispone de una amplia variedad de técnicas estadísticas (modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupamiento, etc.) y gráficas. Funciona en plataformas UNIX y sistemas similares (incluidos FreeBSD y Linux), Windows y MacOS (UNIR,2019).
- **Versión:** 2022.02.3
- **Editor de texto:** RStudio

Herramientas:

- **Nombre:** NetLogo
- **Descripción:** NetLogo es especialmente adecuado para modelar sistemas complejos que se desarrollan con el tiempo. Los modeladores pueden dar instrucciones a cientos o miles de "agentes" que operan independientemente. Esto permite explorar la conexión entre el comportamiento a nivel micro de los individuos y los patrones de nivel macro que surgen (emergen) de su interacción.
- **Versión:** 6.2.2

NetLogo permite a los estudiantes abrir simulaciones y "jugar" con ellas, explorando su comportamiento bajo diversas condiciones. También es un entorno de autoría que permite a los estudiantes, profesores y desarrolladores de planes de estudio crear sus propios modelos. NetLogo es lo suficientemente simple para los estudiantes y profesores, pero lo suficientemente avanzado como para servir como una poderosa herramienta para los investigadores en muchos campos. (Miranda, 2018)

- **Técnica:** Algoritmos de predicción

  Los algoritmos de predicción permitirán establecer el valor de entradas y salidas de pasajeros para cada una de las estaciones. Estos algoritmos dependerán de la naturaleza y comportamiento de los datos, en estos se evaluarán métricas de error que permitan conocer la capacidad del modelo en la predicción dadas unas condiciones como lo son el intervalo de tiempo, el día, y la lluvia.

- **Técnica:** Simulación Basada en Agentes

  Se utiliza para simular cómo las conductas individuales determinan la evolución de un sistema. Consiste en una colección de agentes, un conjunto de reglas de comportamiento y un entorno o ambiente. Un agente constituye la unidad elemental e indivisible de un sistema, por ejemplo, una persona, un grupo de personas, un hogar, o una organización. Los agentes, por definición, tienen un propósito y responde decidiendo o actuando conforme a las reglas que rigen su comportamiento. Éstas diferencian a cada agente, definen las interrelaciones entre los agentes y el ambiente, y establecen la secuencia de acciones con el tiempo. Las reglas se activan bajo condiciones distintas. El comportamiento puede ser reactivo (llamado cambio pasivo) o proactivo (llamado anticipatorio). Al actuar, cada agente va modificando el ambiente hasta que se alcanza el estado de equilibrio y se genera un patrón general de comportamiento que ya no cambia. (Cardoso, 2014)

2. ENTENDIMIENTO DE LOS DATOS

2.1 Recolección de los datos

Los datos recolectados y la fuente se presentan a continuación:

Datos	Fuente	Variables	Contexto	Número de tablas
Entrada de pasajeros enero a diciembre de 2017	Transmilenio	Fase Estación Intervalo Acceso de Estación Fecha	Registra el ingreso de pasajeros en intervalos de 15 minutos para cada una de las estaciones que conectan la troncal del Portal El Dorado en todo el sistema.	12
Salida de pasajeros enero a diciembre de 2017	Transmilenio	Línea Estación Intervalo Fecha	Muestra la salida de pasajeros en intervalos de 15 minutos para cada una de las estaciones que conectan la troncal del Portal El Dorado en todo el sistema.	12
Precipitaciones	Datos Abiertos	Código Estación Código Sensor Fecha observación Valor observado Nombre Estación Departamento Municipio Zona Hidrográfica Latitud Longitud Descripción Sensor Medida	Registro de precipitaciones en Colombia en intervalos de 10 minutos medidos en milímetros.	1
Coordenadas Estaciones	Google Maps	Nombre Estación Latitud Longitud	Muestra las coordenadas de cada una de las estaciones del sistema Transmilenio	1

Tabla 6 Datos recolectados

2.2 Descripción

De acuerdo con los datos recolectados, a continuación, se muestra su correspondiente descripción y características:

Datos	Variables	Tipo	Cantidad de registros
Validaciones enero 2017	Fase Estación Acceso estación Registro Fecha del día (01/01/2017-31/01/2017)	Caracter Caracter Caracter Entera	16.230
Validaciones febrero 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/02/2017-28/02/2017)	Caracter Caracter Carácter Hora Entera	958.328
Validaciones marzo 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/03/2017-31/01/2017)	Caracter Caracter Carácter Hora Entera	1.038.376
Validaciones abril 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/04/2017-30/04/2017)	Caracter Caracter Carácter Hora Entera	891.690
Validaciones mayo 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/05/2017-31/05/2017)	Caracter Caracter Carácter Hora Entera	923.304
Validaciones junio 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/06/2017-30/06/2017)	Caracter Caracter Carácter Hora Entera	978.300
Validaciones julio 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/07/2017-31/07/2017)	Caracter Caracter Carácter Hora Entera	1.543.273
Validaciones agosto 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/08/2017-31/08/2017)	Caracter Caracter Carácter Hora Entera	1.607.009
Validaciones septiembre 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/09/2017-30/09/2017)	Caracter Caracter Carácter Hora Entera	1.510.950
Validaciones octubre 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/09/2017-30/09/2017)	Caracter Caracter Carácter Hora Entera	1.748.183
Validaciones noviembre 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/09/2017-30/09/2017)	Caracter Caracter Carácter Hora Entera	1.681.470
Validaciones diciembre 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/09/2017-30/09/2017)	Caracter Caracter Carácter Hora Entera	1.636.924
Salidas enero 2017	Fase Estación	Caracter Caracter	1.492.495

	Acceso estación Hora Registro Fecha del día (01/01/2017-31/01/2017)	Carácter Hora Entera	
Salidas febrero 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/02/2017-28/02/2017)	Caracter Caracter Carácter Hora Entera	1.350.412
Salidas marzo 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/03/2017-27/03/2017)	Caracter Caracter Carácter Hora Entera	1.492.061
Salidas abril 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/04/2017-23/04/2017)	Caracter Caracter Carácter Hora Entera	1.446.320
Salidas mayo 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/05/2017-31/05/2017)	Caracter Caracter Carácter Hora Entera	1.491.782
Salidas junio 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/06/2017-30/06/2017)	Caracter Caracter Carácter Hora Entera	1.443.870
Salidas julio 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/07/2017-31/07/2017)	Caracter Caracter Carácter Hora Entera	1.493.766
Salidas agosto 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/08/2017-31/08/2017)	Caracter Caracter Carácter Hora Entera	1.494.293
Salidas septiembre 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/09/2017-30/09/2017)	Caracter Caracter Carácter Hora Entera	1.445.460
Salidas octubre 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/10/2017-29/10/2017)	Caracter Caracter Carácter Hora Entera	1.397.017
Salidas noviembre 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/11/2017-30/11/2017)	Caracter Caracter Carácter Hora Entera	1.755.510
Salidas diciembre 2017	Fase Estación Acceso estación Hora Registro Fecha del día (01/12/2017-31/12/2017)	Caracter Caracter Carácter Hora Entera	1.501.950
Precipitaciones 2017	Código Estación Código Sensor Fecha Observación Valor Observado	Numérico Numérico Fecha Decimal	40.888.821

	Nombre Estación Departamento Municipio Zona Hidrográfica Latitud Longitud Descripción Sensor Unidad Medida	Carácter Carácter Carácter Carácter Decimal Decimal Carácter Caracter	
Coordenadas Estaciones	Estación Latitud Longitud	Caracter Decimal Decimal	49

Tabla 7 Descripción de datos

2.3 Exploración

Dentro de la exploración de datos para la tabla de precipitaciones encontramos lo siguiente:

Esta tabla tiene registros desde el 01 de enero de 2017 hasta el 02 de diciembre 2019. Dado que los demás datos están dentro del año 2017, no contemplaremos los demás registros.

Se tienen 488 estaciones en el territorio nacional que han permitido recabar esta información.

La información preliminar estadística de estos datos nos muestra que la variable Código Sensor tiene el mismo valor en todos los registros.

	CodigoEstacion	CodigoSensor	ValorObservado	Latitud	Longitud
count	4.088882e+07	40888821.0	4.088882e+07	4.088882e+07	4.088882e+07
mean	1.238973e+08	240.0	6.937692e-02	5.363691e+00	-7.497744e+01
std	4.383356e+08	0.0	9.822763e-01	2.495178e+00	1.708364e+00
min	1.101702e+07	240.0	0.000000e+00	-4.194000e+00	-8.173100e+01
25%	2.121550e+07	240.0	0.000000e+00	4.204417e+00	-7.587806e+01
50%	2.612530e+07	240.0	0.000000e+00	4.991111e+00	-7.517328e+01
75%	3.502550e+07	240.0	0.000000e+00	6.470000e+00	-7.396033e+01
max	2.805500e+09	240.0	3.000000e+01	1.579700e+01	0.000000e+00

Figura 10 Descripción estadística de las variables de Precipitaciones 2017

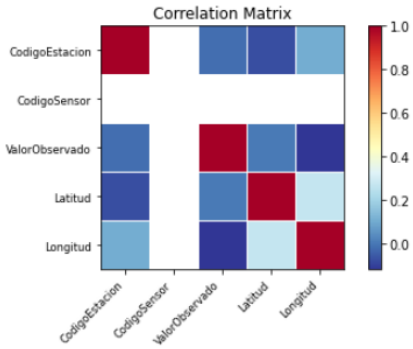

Figura 12 Matriz de correlación de datos de Precipitaciones 2017

La variable que es de mayor interés, Valor Observado, muestra que la mayoría de los registros están en 0, así que de los 40 millones de registros tomarán relevancia solo unos miles. Además, podemos visualizar que el valor máximo que se registró fue de 30 mm, lo que implicó que hubo inundaciones en algún lugar.

A partir de la matriz de correlación entre las variables no se encuentra nada significativo, lo que dará paso a trabajar con los datos directamente sin realizar algún tipo de transformación que elimine esa correlación.

Figura 11 Histograma de Valor Observado

Figura 13 Mapa geográfico de estación de sensores de medición de mm de lluvia

De acuerdo con el histograma, la mayor parte de los datos están concentrados en 0, es decir que hubo muchos registros de sequía que de lluvia. Sin embargo, los siguientes datos más frecuentes son de 2 a 10 mm registrados en los sensores, que implican lluvias fuertes.

Los sensores se encuentran distribuidos en las regiones andina, caribe y pacífica mayoritariamente. Para este análisis será importante filtrar y revisar la información solamente del área de Bogotá, ya que las demás zonas no tienen incidencia en el contexto de negocio.

Dentro de la información encontramos que existen aparentemente 39 departamentos, posiblemente se deba a errores ortográficos. El

departamento en el que más se presenta lluvia es en el Valle del Cauca, se deberá revisar si se tienen datos atípicos o erróneos en esto.

Figura 14 Lluvia acumulada en milímetros por departamento

Las bases de las entradas son 12, estas tienen las entradas al sistema de Transmilenio para los meses de enero a diciembre. La mayoría de estas bases son archivos de Excel que contienen 3 hojas: 2 segregadas por los proveedores de las tarjetas que operaron en ese mes, y una tercera hoja con el consolidado de las entradas, enero no tiene la hoja consolidada y la información de las Fase I, II y III la tienen por día, pero no por intervalos de horas. Para el propósito de este trabajo de grado nos enfocaremos en la hoja que consolida toda la información capturada del mes.

Al revisar la estructura encontramos lo siguiente:

Las bases de Ene a Jun tienen filas que agrupan totales por troncales, estaciones y portales, y además las entradas por cada acceso, mientras que esto no pasa en las bases de Jul a Dic, ya que esta información se segrega por columnas distintas, es decir, una columna tiene la troncal, otra la estación y otra los accesos por lo que no se requiere agrupar los totales en una fila, la estructura de las bases es la siguiente:

Bases de entradas Ene - Jun		**Bases de entradas Jul - Dic**	
Columna	**Descripción**	**Columna**	**Descripción**
Etiqueta de datos	Posee información de la troncal, portal, estación y accesos además de los intervalos de tiempo (cada 15 minutos)	Fase	Nombre de la fase de construcción del sistema
		Línea	Ruta del Transmilenio
		Estación	Nombre de la estación o portal del sistema
Fechas	Es un conjunto de columnas que tienen las entradas por cada día del mes	Acceso de Estación	Nombre del acceso que marca la entrada al sistema
Total general	Sumatoria de las entradas por fila, es decir, por troncal, portal, estación, acceso e intervalo de tiempo	Intervalo	Marca la hora con intervalos de 15 minutos
		Fechas	Es un conjunto de columnas que tienen las entradas por cada día del mes
		Total general	Sumatoria de cada Fila

Tabla 8 Estructura de las bases de entrada

Al realizar un análisis estadístico encontramos que las bases difieren seguramente por la forma como están construidas:

- Las bases de Ene a Jun tienen desviaciones estándar de 4 dígitos, esto se debe a que algunas de sus filas muestran las entradas por un intervalo especifico de 15 minutos y otras tienen todo el valor de la operación de una troncal completa a lo largo del día por lo que la diferencia es alta y hay datos atípicos, de manera general para estas bases el valor mínimo en casi todas las variables es de una (1) entrada y los máximos están cercanos a las 500.000, para la variable de total general estos valores son mayores ya es la suma de las demás columnas.
- Las bases de Jul a Dic tienen desviaciones estándar de dos dígitos en la mayoría de las variables aunque algunas llegan a valores cercanos a las 120 desviaciones, para este caso el valor mínimo es de 0 entradas lo que es coherente ya que dependiendo del horario o por condiciones externas puede que la estación no reciba personas en ese intervalo de tiempo, los valores máximos son variables pero son del orden de 500 hasta 2000 entradas por variable a excepción del total general que al sumar todas las columnas será mayor.

A nivel de formato de los datos encontramos que las bases de enero a junio tienen la primera columna con información alfanumérica y las demás columnas son numéricas en formato float, mientras que las bases de julio a diciembre tienen 5 columnas alfanuméricas y las demás columnas son numéricas en formato int. Adicional a

lo anterior se ve que las bases de Ene a Jun tienen valores NaN, mientras que las bases de Jul a Dic no tienen valores Nan ya que los espacios vacíos se diligenciaron con el valor "0",

La matriz de correlaciones nos indica para todos los meses fuertes correlaciones en las variables con valores cercanos a 1, sin embargo, al aterrizar estos resultados al caso de negocio se puede explicar este suceso en que al ser bases de transporte los comportamientos de los días de lunes a viernes serán parecidos en su operación y los de sábados, domingos y festivos serán días que parezcan alejados frente a los días hábiles, a continuación se ilustra esto con 2 ejemplos: el comportamiento de una mes “normal” como lo es Feb donde vemos un patrón de 5 días correlacionados y 2 días con tonos distintos y abril con cambios de tono en semana santa y los fines de semana:

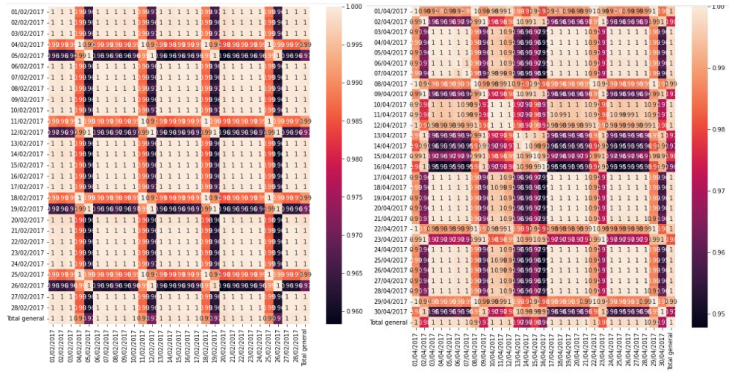

Figura 15 Matriz de correlación de datos de entrada de Feb y de Abr

Esto también se puede ver cuando se analiza el comportamiento de las entradas por día, un ejemplo es el comportamiento del mes de diciembre donde los días festivos, sábados y domingos las entradas son menos que los días hábiles, este comportamiento se repite cada semana en todos los meses y rara vez se afecta:

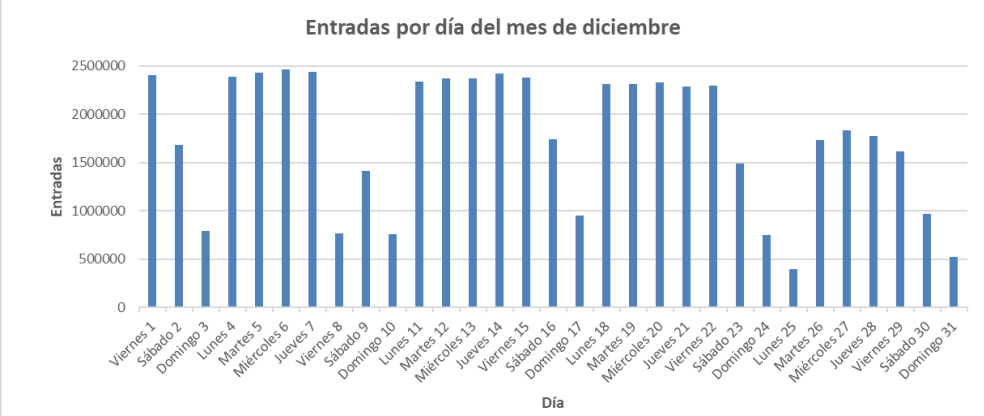

Figura 16 Número de entradas en el mes de diciembre

Se encuentran filas en las bases con los valores: (0) y (0) (Unknown), estos representan accesos a la estación que no se logran identificar.

Similar a las bases de las entradas también las bases de las salidas son 12, estas tienen los datos de las salidas del sistema de Transmilenio para los meses de enero a diciembre. Estas bases son archivos de Excel que en su mayoría contienen 3 hojas (las bases de mayo, junio y agosto solo tienen la hoja consolidada): 1 con los datos de la Fase I – II del sistema, otra con la fase 3 y una tercera hoja que consolida todas las salidas por mes, para el propósito de este trabajo de grado nos enfocaremos en la hoja que consolida toda la información capturada del mes.

Al revisar la estructura encontramos las siguientes columnas:

TABLAS DE SALIDAS DE ENERO A DICIEMBRE	
Columna	**Descripción**
Línea	Troncal de Transmilenio
Estación	Nombre de la estación o portal del sistema
Acceso de Estación	Nombre del acceso que marca la entrada al sistema o parada de servicios duales
Intervalo	Marca la hora con diferencias de 15 minutos
Fechas	Es un conjunto de columnas por cada día del mes
Total	Sumatoria de cada Fila

Tabla 9 Estructura de las bases de salidas

Al realizar un análisis estadístico encontramos que estas bases no varían tanto como las de las entradas, las bases de salidas tienen las siguientes características:

- Tienen desviaciones estándar de dos dígitos en la mayoría de las variables aunque algunas llegan a valores cercanos a las 120 desviaciones, para este caso el valor mínimo es de 0 entradas lo que es coherente ya que dependiendo del horario o por condiciones externas puede que la estación no salgan personas en ese intervalo de tiempo, los valores máximos son variables pero son del orden de 500 hasta 2000 salidas por variable a excepción del total general que al sumar todas las columnas será mayor, este comportamiento coincide con las cifras de las salidas lo que es lógico ya que el balance del sistema a nivel de usuarios será que lo que entra debe ser igual a lo que sale.
- A nivel de formato de los datos encontramos que todas las bases tienen las primeras 4 columnas en formato alfanumérico, las restantes son numéricas, la diferencia entre los formatos se da porque las bases de Ene a Jun tienen las columnas numéricas en formato float mientras que las de Jul a Dic las tiene en Int.
- Adicional a lo anterior se ve que la mayoría de las bases tienen valores NaN, solo las bases: Jul, Ago, Nov y Dic no tienen valores Nan ya que los espacios vacios se diligenciaron con el valor "0".
- La matriz de correlaciones al igual que el caso de las entradas nos indica para todos los meses fuertes correlaciones en las variables con valores cercanos a 1, sin embargo, como ya vimos el comportamiento se explica por la operación entre semana vs los días no hábiles:

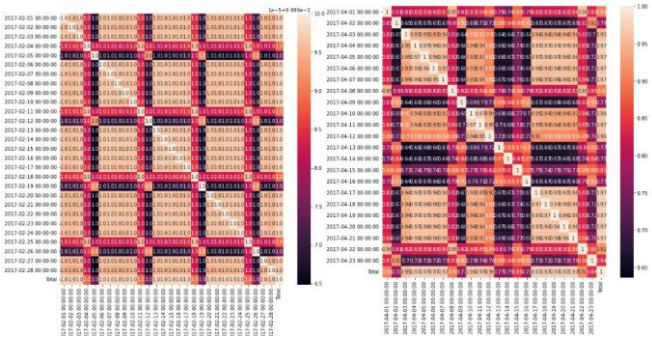

Figura 17 Matriz de correlación de datos de salidas de Feb y de Abr

Esto se constata cuando se analiza el comportamiento de las salidas por día:

Figura 18 Número de salidas en el mes de diciembre

Se puede confirmar con los datos que los puntos de alimentación y salida más altos son los portales, siendo el de más movimiento el portal del norte también llamado cabecera autopista norte, aunque resalta también las estaciones conectoras de troncales por ejemplo la Avenida Jiménez.

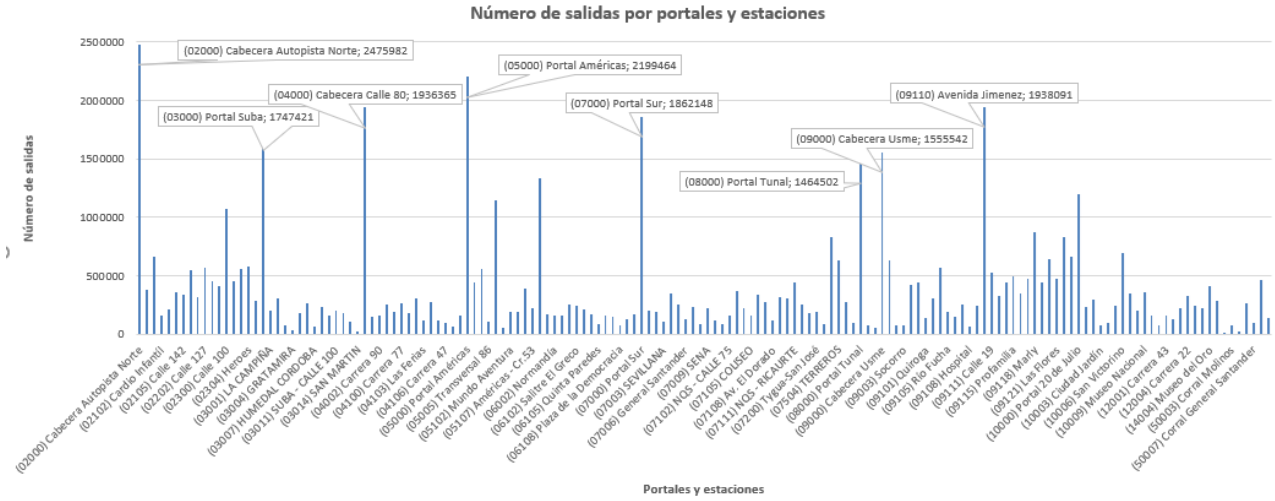

Figura 19 Número de salidas en el sistema por estaciones en el mes de diciembre

Sobre las franjas horarias encontramos que Transmilenio tiene 2 etapas con alto volumen, estos horarios coinciden con las horas en las que los usuarios salen a trabajar o estudiar y las horas de regreso a sus hogares:

Entradas: De las 5:45 am a las 7:45 am y de las 5:00 pm a las 6:30 pm, los picos más altos se registraron alrededor de las 6:15 am y las 5:15 pm.

Salidas: De las 6:45 am a las 9:00 am y de las 5:30 pm a 8:00 pm, los picos más altos se registraron alrededor de las 8:00 am y a las 6:30 pm.

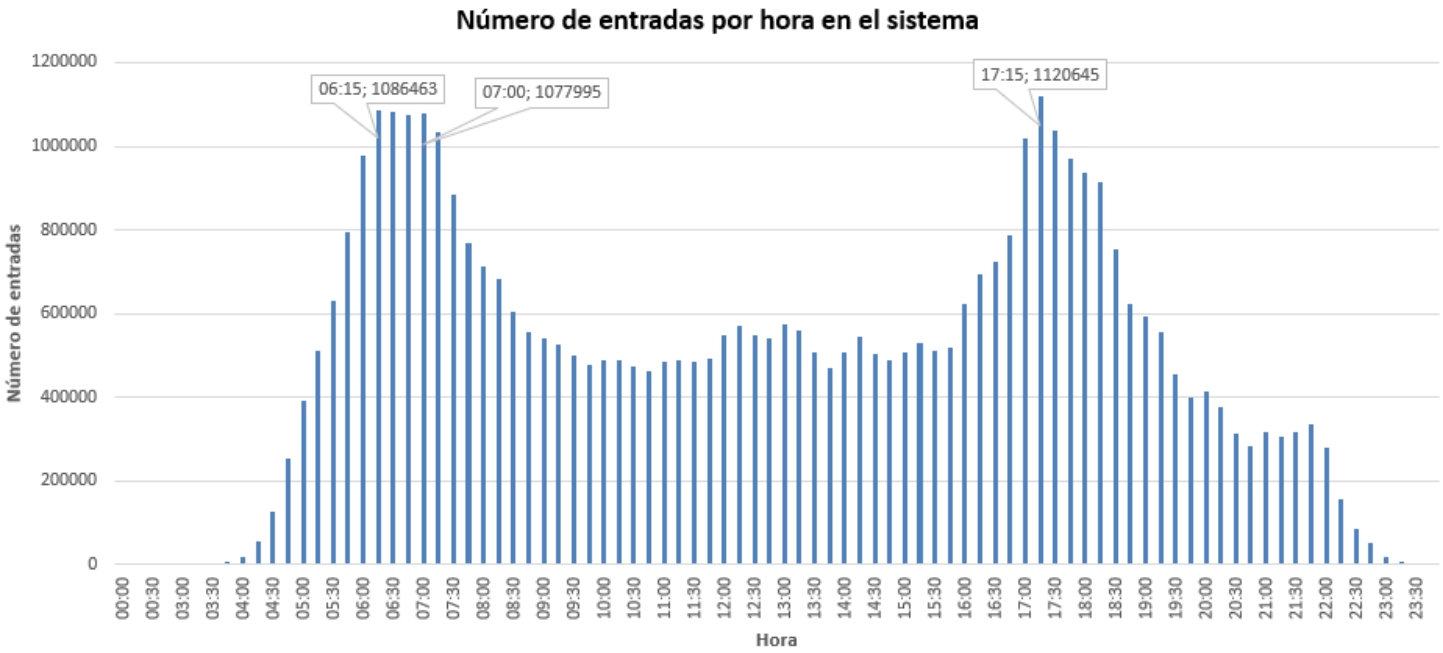

Figura 20 Número de entradas en el sistema por hora en el mes de febrero

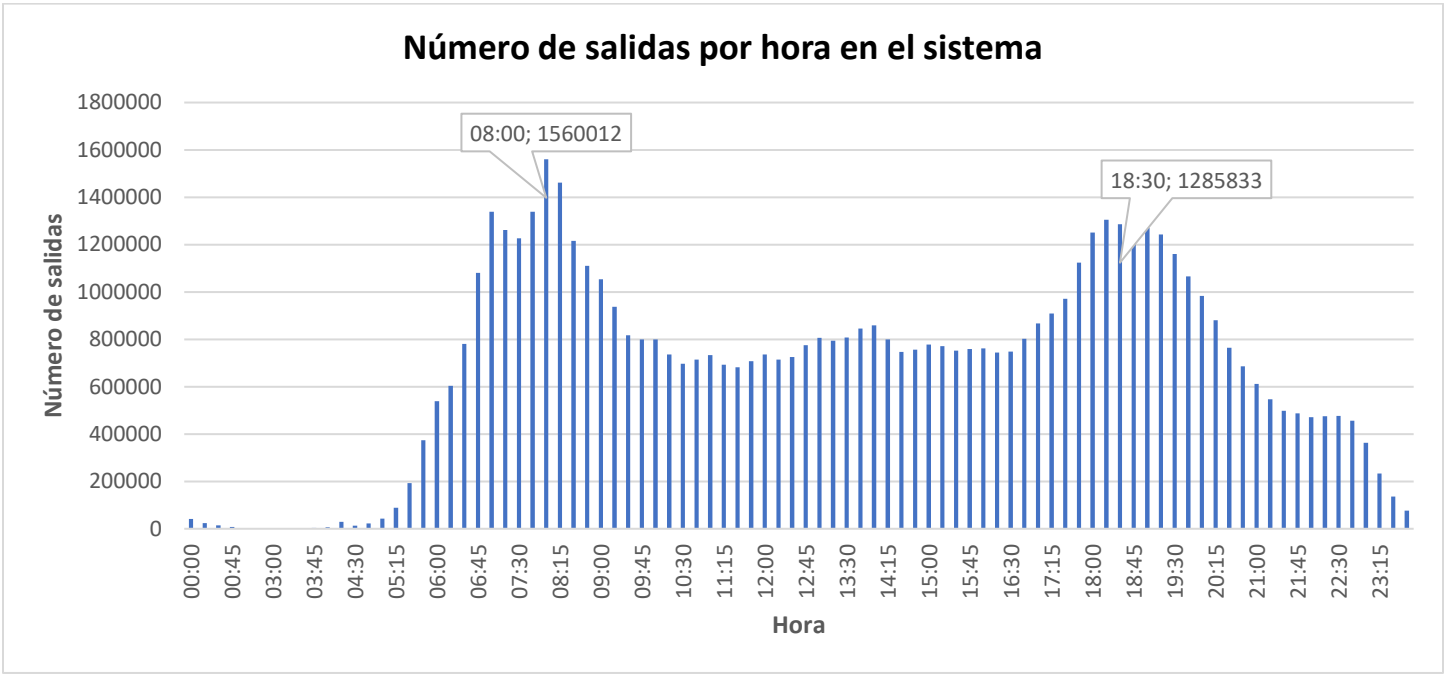

Figura 21 Número de salidas en el sistema por hora en el mes de febrero

Vemos registros de antes de las 4:00 am y después de las 12:00 pm, estos son datos erróneos ya que Transmilenio opera desde las 4:00 am y hasta las 12:00 pm.

2.4 Verificación de la calidad de los datos

En la tabla de precipitaciones se verificó la calidad de los datos de la siguiente manera:

```
CodigoEstacion       0
CodigoSensor         0
FechaObservacion     0
ValorObservado       1
NombreEstacion       1
Departamento         1
Municipio            1
ZonaHidrografica     1
Latitud              1
Longitud             1
DescripcionSensor    1
UnidadMedida         1
dtype: int64
```

Figura 22 Número de valores nulos en la tabla de Precipitaciones

- Solamente se tiene un valor faltante para las variables Valor Observado, Nombre Estación, Departamento, Municipio, Zona Hidrográfica, Latitud, Longitud, Descripción Sensor, Unidad Medida. Por lo que, al no representar un porcentaje significativo, se podrá despreciar este registro.
- En la variable Código Estación encontramos 488 registros diferentes, sin ninguna alteración. Sin embargo, en el nombre de estación tan solo existen 486 nombres distintos, lo que puede considerarse como un error.
- La variable Código Sensor tiene el mismo dato, por lo que se podrá despreciar esta columna ya que no aporta ninguna información relevante.
- La columna Fecha de observación tiene datos de 2017 hasta 2019 en formato fecha y hora, por lo que se deberá filtrar la información solo para el año de interés.
- La columna valor observado no tiene errores ni datos atípicos.
- La columna Departamento tiene 39 categorías, al revisarlas, se encuentra que tenemos BOGOTÁ D.C y BOGOTA, por lo que se deberá reemplazar una de estas y así unificarlas. También se tiene categorías ATLÁNTICO y ATLANTICO, BOLIVAR y BOLÍVAR, CHOCÓ y CHOCO, CORDOBA, CÓRDOBA, NARINO y NARIÑO. Por otro lado, tenemos 175.961 registros con un valor <nil>.
- La columna Municipio tiene 333 categorías, se tienen 175.961 registros con un valor <nil>.
- En la columna Zona Hidrográfica se tienen 32 categorías con 400.101 registros con un valor <nil>.
- Las columnas de latitud y longitud muestran un dato fuera del territorio nacional, cercano al continente africano, por lo que se considerará esto como un error.
- La columna Descripción Sensor tiene una única categoría la cual es Precipitación, por lo que no representa un valor significante para el conjunto de datos.
- La columna Unidad Medida solamente tiene un valor el cual es mm (milímetros), por lo que no es relevante para el conjunto de datos.

Ahora bien, debemos verificar la calidad de los datos de entradas y salidas de usuarios de Transmilenio en sus diferentes estaciones:

- Entradas:
  - Enero 2017: La información de entradas y salidas de este mes se encuentra en dos hojas de Excel discriminadas por las fases del sistema (Fase I y II y Fase III). Se evidencia que no se tiene la información discriminada por horas o intervalos de tiempos, sino que está totalizado por cada uno de los días. Los datos nulos dentro de la tabla son ceros, ya que se toman como no ingresos a la estación por ese acceso. Para las Fases I y II tenemos 601 datos nulos y para la Fase III 145.
  - Febrero 2017: La información se encuentra consolidada por estación y hora. Se tienen 993.186 datos nulos.
  - Marzo 2017: La información se encuentra consolidada por estación y hora. Se tienen 734424 datos nulos.
  - Abril 2017: La información se encuentra consolidada por estación y hora. Se tienen 702.727 datos nulos.
  - Mayo 2017: La información se encuentra consolidada por estación y hora. Se tienen 725.920 datos nulos.
  - Junio 2017: La información se encuentra consolidada por estación y hora. Se tienen 806.070 datos nulos.
  - Julio 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.

    - Agosto 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.
    - Septiembre 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.
    - Octubre 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.
    - Noviembre 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.
    - Diciembre 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.
- Salidas:
    - Enero 2017: La información se encuentra consolidada por estación y hora. Se tienen 500.078 datos nulos.
    - Febrero 2017: La información se encuentra consolidada por estación y hora. Se tienen 456.102 datos nulos.
    - Marzo 2017: La información se encuentra consolidada por estación y hora. Solamente se tienen datos hasta el 27 de marzo, faltan 4 días del 28 al 31 de ese mes. Se tienen 461.362 datos nulos.
    - Abril 2017: La información se encuentra consolidada por estación y hora. Solamente se tienen datos hasta el 23 de abril. Se tienen 441.791 datos nulos.
    - Mayo 2017: La información se encuentra consolidada por estación y hora. Se tienen 458.712 datos nulos.
    - Junio 2017: La información se encuentra consolidada por estación y hora. Se tienen 443.606 datos nulos.
    - Julio 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.
    - Agosto 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.
    - Septiembre 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.
    - Octubre 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.
    - Noviembre 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.
    - Diciembre 2017: La información se encuentra consolidada por estación y hora. No se tienen datos nulos.

3. PREPARACIÓN DE LOS DATOS

3.1 Selección

Para la base de precipitaciones se utilizarán las siguientes variables teniendo en cuenta las siguientes consideraciones:

- Código Estación: No es relevante para el modelo, ya que solo identifica la estación en la que fue medida el nivel de lluvia.
- Código Sensor: Esta variable tiene un único valor, por lo cual no es relevante para el modelo.
- Fecha Observación: Se utilizará esta variable, ya que con esta se puede relacionar la información de entradas y salidas de usuarios.
- Valor Observado: Se utilizará esta variable ya que permitirá conocer si hubo lluvia en la ciudad o no.

- Nombre Estación: No es relevante para el modelo, ya que solo menciona el nombre de la estación en la que se tomó la medición.
- Departamento: Se utilizará únicamente para filtrar la información por Bogotá D.C, pero no será relevante para el modelo.
- Municipio: No es relevante para el modelo ya que, para el caso de Bogotá, al ser un distrito especial, el único municipio será el mismo distrito.
- Zona Hidrográfica: No es relevante para el modelo ya que menciona la zona de los ríos en las que se tomó la medición.
- Latitud: Se utilizará para poder comparar la información geográfica de las estaciones de Transmilenio y reconocer si en ese punto de la ciudad hubo lluvia y como esta afectó el sistema.
- Longitud: Se utilizará para poder comparar la información geográfica de las estaciones de Transmilenio y reconocer si en ese punto de la ciudad hubo lluvia y como esta afectó el sistema.
- Descripción Sensor: Esta variable no es relevante para el modelo, ya que indica el tipo de sensor y tiene un único valor el cual es Precipitación.
- Unidad Medida: No es relevante para el modelo ya que solo indica un único valor el cual es milímetros.

Para la base de entradas de usuarios al sistema utilizaremos:

- Las bases de entradas de febrero a diciembre, específicamente la hoja de información consolidada llamada “VALIDACIONES CONSOLIDADO”.
- La base de enero se descarta ya que no tiene la información consolidada ni discriminada por intervalos de horas o de 15 minutos lo que no permite su análisis correcto, adicional a esto esta base tiene una muestra muy qué pequeña de datos que hace que no sea significativa para el caso, tiene 16.000 registros cuando todas las bases tienen valores con cerca de un millón de registros cada una.
- Las variables que se van a usar para las entradas son las siguientes:
  - De las bases de Feb a Jun:
    - Etiqueta de datos: Posee información de la troncal, portal, estación, paradas y accesos además de los intervalos de tiempo (cada 15 minutos)
    - Fechas: Número de entradas por cada día del mes
    - Total general: Número total de entradas por estación o portal e intervalo de tiempo
  - De las bases de Jul a Dic:
    - Línea: Troncal del sistema
    - Estación: Nombre de la estación o portal del sistema
    - Acceso de estación: Accesos al sistema y paradas de servicios duales (CAC, aeropuerto, etc.)
    - Intervalo: Hora en intervalos de 15 minutos
    - Fechas: Número de entradas por cada día del mes

Para la base de salidas de usuarios del sistema se identifica que no se requerirá para el modelo ni para la simulación, dado que el modelo se enfocará en la predicción de las entradas, en la simulación las salidas será un comportamiento emergente de acuerdo con las decisiones tomadas por los agentes en la misma.

Teniendo en cuenta la necesidad del negocio la cual, está relacionada con las rutas que se dirigen netamente hacia el Portal el Dorado, se realiza la selección de las estaciones con base a la siguiente tabla:

1	K43	K16	K23	K86	K10	K54
(06111) Universidades	(07503) SAN MATEO	(02502) Terminal	(02200) Alcalá	(10009) Museo Nacional	(10000) Portal 20 de Julio	(09000) Cabecera Usme
(06109) Centro Memoria	(07504) TERREROS	(02001) Centro Comercial Santa Fe	(02201) Prado	(06108) Plaza de la Democracia	(10002) Av. Primero de Mayo	(09001) Molinos
(06108) Plaza de la Democracia	(07505) LEON XIII	(02101) Toberín	(02303) Calle 85	(06106) Corferias	(10010) Hospitales	(09100) Calle 40 Sur
(06107) Ciudad Universitaria	(07506) DESPENSA	(02200) Alcalá	(09119) Calle 57	(06105) Quinta Paredes	(10005) Bicentenario	(09104) Restrepo
(06106) Corferias	(07004) VENECIA	(02204) Pepe Sierra	(09117) Calle 45	(06104) Gobernación	(09109) Tercer Milenio	(09107) Hortúa
(06105) Quinta Paredes	(07006) General Santander	(07102) NQS - CALLE 75	(09114) Calle 26	(06103) CAN	(09110) Avenida Jimenez	(06108) Plaza de la Democracia
(06104) Gobernación	(07009) SENA	(07103) AV. CHILE	(06108) Plaza de la Democracia	(06102) Salitre El Greco	(09113) Calle 22	(06107) Ciudad Universitaria
(06103) CAN	(07111) NQS - RICAURTE	(07108) Av. El Dorado	(06107) Ciudad Universitaria	(06101) El Tiempo	(06109) Centro Memoria	(06106) Corferias
(06102) Salitre El Greco	(07109) CAD	(06106) Corferias	(06104) Gobernación	(06100) Av. Rojas	(06107) Ciudad Universitaria	(06105) Quinta Paredes
(06101) El Tiempo	(06106) Corferias	(06105) Quinta Paredes	(06103) CAN	(06000) Portal Eldorado	(06104) Gobernación	(06103) CAN
(06100) Av. Rojas	(06105) Quinta Paredes	(06103) CAN	(06102) Salitre El Greco		(06103) CAN	(06102) Salitre El Greco
(06002) Normandía	(06104) Gobernación	(06102) Salitre El Greco	(06101) El Tiempo		(06102) Salitre El Greco	(06101) El Tiempo
(06001) Modelia	(06102) Salitre El Greco	(06101) El Tiempo	(06100) Av. Rojas		(06002) Normandía	(06100) Av. Rojas
(06000) Portal Eldorado	(06101) El Tiempo	(06100) Av. Rojas	(06000) Portal Eldorado		(06001) Modelia	(06001) Modelia
	(06002) Normandía	(06001) Modelia			(06000) Portal Eldorado	(06000) Portal Eldorado
	(06001) Modelia	(06000) Portal Eldorado				
	(06000) Portal Eldorado					

Tabla 10 Estaciones que conforman cada una de las rutas

En total, se tienen 49 estaciones que pertenecen a las rutas que tienen como destino el Portal el Dorado, la base de datos inicial cuenta con la discriminación de los ingresos por las diferentes entradas de cada estación (discapacitados, sur, norte, occidente, etc.), para efectos del proyecto se consolidaron todas las estaciones en un solo valor por cada intervalo de 15 minutos, las estaciones a analizar se presentan a continuación:

ESTACIONES FINALES
(02001) Centro Comercial Santa Fe
(02101) Toberín
(02200) Alcalá
(02201) Prado
(02204) Pepe Sierra
(02303) Calle 85
(02502) Terminal
(06000) Portal Eldorado
(06001) Modelia
(06002) Normandía
(06100) Av. Rojas
(06101) El Tiempo
(06102) Salitre El Greco
(06103) CAN
(06104) Gobernación
(06105) Quinta Paredes
(06106) Corferias
(06107) Ciudad Universitaria
(06108) Plaza de la Democracia
(06109) Centro Memoria
(06111) Universidades
(07004) VENECIA
(07006) General Santander
(07009) SENA
(07102) NQS - CALLE 75
(07103) AV. CHILE
(07108) Av. El Dorado
(07109) CAD

(07111) NQS - RICAURTE
(07503) SAN MATEO
(07504) TERREROS
(07505) LEON XIII
(07506) DESPENSA
(09000) Cabecera Usme
(09001) Molinos
(09100) Calle 40 Sur
(09104) Restrepo
(09107) Hortúa
(09109) Tercer Milenio
(09110) Avenida Jimenez
(09113) Calle 22
(09114) Calle 26
(09117) Calle 45
(09119) Calle 57
(10000) Portal 20 de Julio
(10002) Av. Primero de Mayo
(10005) Bicentenario
(10009) Museo Nacional
(10010) Hospitales

Tabla 11 Estaciones que conforman las rutas que se dirigen hacia el Portal El Dorado

Se tiene además una tabla la cual contiene los horarios de funcionamiento de cada una de las rutas, dada la información contenida allí se filtran los horarios de la base de datos, dejando solamente desde las 4:00 hasta las 23:45 horas.

RUTA	ORIGEN	DESTINO	INICIO-SEM	FIN-SEM	INICIO-SAB	FIN-SAB	INICIO-D_FES	FIN-D_FES
1	(06111) Universidades	(06000) Portal Eldorado	4:30	23:00	5:00	23:00	5:30	22:00
K43	(07503) SAN MATEO	(06000) Portal Eldorado	4:00	23:00	4:30	23:00	NA	NA
K16	(02502) Terminal	(06000) Portal Eldorado	5:30	22:30	5:30	22:00	NA	NA
K23	(02200) Alcalá	(06000) Portal Eldorado	5:00	22:00	NA	NA	NA	NA
K86	(10009) Museo Nacional	(06000) Portal Eldorado	5:30	23:00	6:00	23:00	7:00	22:00
K10	(10000) Portal 20 de Julio	(06000) Portal Eldorado	4:30	23:00	5:00	23:00	5:30	22:00
K54	(09000) Cabecera Usme	(06000) Portal Eldorado	5:30	22:30	NA	NA	NA	NA

Tabla 12 Horarios y días de funcionamiento de cada ruta

3.2 Limpieza

En la limpieza de datos para la base de datos de precipitaciones se realizó lo siguiente:

- Se filtra la información de Departamento por Bogotá D.C y Bogotá, por lo cual se trabajan con 503.622 datos
- Se descartan las siguientes variables: Código Estación, Código Sensor, Nombre Estación, Departamento, Municipio, Zona Hidrográfica, Descripción Sensor, Unidad Medida
- Se divide la variable Fecha Observación en dos columnas para discriminar la fecha y la hora
- Se ajusta el formato de la hora para tenerla en 24 horas

En la limpieza de datos de las bases de entradas se realizó lo siguiente:

- Se tomó para cada base la hoja que consolida toda la información capturada del mes.
- Se eliminaron las variables: fase y total general.
- Se separa en columnas la información de la troncal, portal, estación e intervalos de tiempo de las bases de febrero a junio para que coincidan con las columnas de las bases de julio a diciembre.
- De las bases de febrero a junio se eliminan las filas que son sumatorias totales, por ejemplo, la sumatoria de todas las entradas por troncal.
- A nivel de formato de los datos se unifican los formatos de las columnas numéricas a formato int.
- Los valores NaN se diligencian con el valor "0".
- Se eliminan las filas en las bases con los valores: (0) y (0) (Unknown) ya que representan accesos a la estación que no se logran identificar.
- Se eliminan los registros de antes de las 4:00 am y después de las 12:00 pm, estos son datos erróneos ya que Transmilenio opera desde las 4:00 am y hasta las 11:00 pm.
- Como las bases de entradas y salidas deben coincidir en las fechas se eliminan las columnas de las bases de entradas de las fechas que no se encuentren en el siguiente rango:
  - Febrero: 01/02/2017-28/02/2017
  - Marzo: 01/03/2017-27/03/2017
  - Abril: 01/04/2017-23/04/2017
  - Mayo: 01/05/2017-31/05/2017
  - Junio: 01/06/2017-30/06/2017
  - Julio: 01/07/2017-31/07/2017
  - Agosto: 01/08/2017-31/08/2017
  - Septiembre: 01/09/2017-30/09/2017
  - Octubre: 01/10/2017-29/10/2017
  - Noviembre: 01/11/2017-30/11/2017
  - Diciembre: 01/12/2017-31/12/2017

3.3 Construcción

La base final de entradas tiene consolidadas las 49 estaciones seleccionadas, discriminando los días de cada mes y en los intervalos de una hora en el rango horario ya mencionado anteriormente, se tienen las siguientes variables:

	DESCRIPCIÓN	TIPO DE DATO
FASE	Fase de construcción a la que pertenece cada estación	Varchar(1)
COD_LINEA	Código de identificación de la linea a la que pertenece la estación	Varchar(10)
LINEA	Nombre de la linea a la que pertenece la estación	Varchar(50)
COD_ESTACION	Código de identifficación de la estación	Varchar(10)
ESTACION	Nombre de la estación	Varchar(50)
FECHA	Fecha del registro	Date
INTERVALO	Hora en la que se almacena la información	Time
ENTRADAS	Cantidad de personas que ingresan a la estación por alguna de las entradas	Integer
SALIDAS	Cantidad de personas que salen de la estación por alguna de las salidas	Integer
NETO	Cantidad de personas que se encuentran en la estación (Entradas - Salidas)	Integer

Tabla 13 Estructura de la base final de entradas

Dado que el proyecto contempla no solo analizar el flujo de personas en cada estación sino también la relación que tienen los factores externos como la lluvia en el comportamiento de las entradas y salidas, se generó una tabla con las coordenadas de cada estación, de esta manera se cruzara la información con la tabla de precipitaciones ya explicada anteriormente.

ESTACIONES FINALES	LONGITUD	LATITUD
(02001) Centro Comercial Santa Fe	4.763457713332859	-74.0444585326047
(02101) Toberín	4.746527757568758	-74.04728142787408
(02200) Alcalá	4.721962082046572	-74.0508282025168
(02201) Prado	4.714381101117608	-74.0526001352324
(02204) Pepe Sierra	4.698815094140516	-74.05524003485964
(02303) Calle 85	4.6719745033113576	-74.05975714016311
(02502) Terminal	4.768948312938854	-74.04343124089364
(06000) Portal Eldorado	4.681409810129376	-74.12126823063599
(06001) Modelia	4.676079880551295	-74.11758267013691
(06002) Normandía	4.669581503865497	-74.11348864137857
(06100) Av. Rojas	4.661651067137003	-74.10849808605975
(06101) El Tiempo	4.656209448931246	-74.10505377046721
(06102) Salitre El Greco	4.650826881616169	-74.10149548679819
(06103) CAN	4.646566252319917	-74.09871671483283
(06104) Gobernación	4.642363341043774	-74.0961356147643
(06105) Quinta Paredes	4.637127449190034	-74.09221208592845
(06106) Corferias	4.634257344124987	-74.08970145767846
(06107) Ciudad Universitaria	4.631591269480822	-74.08387651279595
(06108) Plaza de la Democracia	4.627153877430468	-74.0806992711593
(06109) Centro Memoria	4.622147660361698	-74.07743699935085
(06111) Universidades	4.605346689799241	-74.06698175580632
(07004) VENECIA	4.595595883786341	-74.14267682236724
(07006) General Santander	4.593690415613592	-74.12869927116238
(07009) SENA	4.597476651477029	-74.11102401262711
(07102) NQS - CALLE 75	4.669871857690101	-74.07180235580743
(07103) AV. CHILE	4.666197645699065	-74.07485068596037
(07108) Av. El Dorado	4.629518407524084	-74.07981667115898
(07109) CAD	4.623281564491598	-74.08438961414339
(07111) NQS - RICAURTE	4.612736581779376	-74.0930427134776
(07503) SAN MATEO	4.585480616255142	-74.20689669072298
(07504) TERREROS	4.590050176966589	-74.19858048298565
(07505) LEON XIII	4.592568328558069	-74.1933493522023
(07506) DESPENSA	4.594640709763832	-74.1883890576752
(09000) Cabecera Usme	4.501134104176606	-74.11732276841936
(09001) Molinos	4.557036597258704	-74.12179204382284
(09100) Calle 40 Sur	4.576267359508239	-74.12017927038427
(09104) Restrepo	4.581837101123641	-74.10182040161955
(09107) Hortúa	4.591239198410128	-74.0900491129129
(09109) Tercer Milenio	4.597608668430209	-74.08439361274019
(09110) Avenida Jimenez	4.6034236780753455	-74.08018157055132
(09113) Calle 22	4.612624957580788	-74.07433234157033
(09114) Calle 26	4.61716780404637	-74.07221121482632
(09117) Calle 45	4.632980907204341	-74.06768135681017
(09119) Calle 57	4.643157053343403	-74.06587919934647
(10000) Portal 20 de Julio	4.56739625221805	-74.09952809707126
(10002) Av. Primero de Mayo	4.5763262698665566	-74.09471640675785
(10005) Bicentenario	4.595209458832477	-74.08127122883745
(10009) Museo Nacional	4.61683678003939	-74.06856012832793
(10010) Hospitales	4.596123557692326	-74.08648097040819

Tabla 14 Ubicación geográfica de cada una de las estaciones, longitud y latitud

3.4 Integración

A continuación, se describe la integración de los datos:

- La integración de la tabla de pasajeros con la de precipitaciones se realiza a través del dato de fecha y el intervalo de tiempo, ya que la primera tiene un intervalo de hora y la segunda de 10 minutos, además de la latitud y longitud que permitirán establecer la cercanía de la precipitación en las estaciones.
- La información de festivos se relacionará directamente con la fecha en la tabla de demanda de usuarios por estación.

3.5 Formato

- Se obtiene una tabla con las siguientes características:

DATO	TIPO DE DATO
COD_LINEA	Integer
COD_ESTACION	Integer
ESTACIÓN	Varchar(50)
FECHA_HORA	Date time
FECHA	Date
INTERVALO	Time
FESTIVO	Binary
ENTRADA_HORA	Integer
PRECIPITACION_HORA	Float

Tabla 15 Tipos de datos de las variables

- Luego de obtener esta matriz se realizó un filtro a los datos por el campo COD_ESTACION para obtener los datos para cada una de las estaciones.
- Posteriormente, se realizó un filtro por los días de lunes a viernes y los sábados y Domingos.

4. MODELACIÓN

4.1 Selección del modelo

4.1.1 Técnicas de modelado

Para la predicción de los pasajeros en cada una de las estaciones se utilizarán modelos de predicción de series de tiempo, ya que estos modelos permiten absorber el comportamiento de la variable objetivo de acuerdo con el paso del tiempo. Como se pudo observar en la exploración de los datos, las entradas y salidas de usuarios dependen totalmente de dos factores principales: el día de la semana y el horario, esto indica que a lo largo de los datos encontraremos una serie de tiempo agrupada para cada una de las estaciones que conectan a los usuarios al Portal El Dorado.

La técnica de modelado para las series de tiempo de cada estación se utilizarán modelos ARIMA y SARIMA de acuerdo con el comportamiento de cada de estas.

Los modelos ARIMA o modelo autorregresivo integrado de promedio móvil son modelos estadísticos que usan variaciones y regresiones de datos con el objetivo de encontrar patrones para lograr una predicción futura, al ser un modelo de series de tiempo dinámico tiene en cuenta la dependencia existente entre los datos (Hernández, s.f)., dicho de otra forma, logra que las estimaciones futuras se expliquen por los datos del pasado y no por variables independientes, lo que lo hace ideal para el modelamiento y la estimación de datos en los modelos de transporte de pasajeros.

El modelo ARIMA deriva su nombre de sus componentes: AR (Autoregresivo), I(Integrado) y MA (Medias Móviles), este modelo permite entender los valores como una función lineal de los datos del pasado y los errores

asociados al azar, y además puede ajustarse para considerar el efecto de la estacionalidad, cuando esto sucede, se habla de un modelo SARIMA (seasonal autoregressive integrated moving average) (Hernández, s.f).

Los modelos ARIMA y SARIMA son ideales para realizar estimaciones cuando los datos forman series de tiempo con estacionariedad, como lo son las bases de datos de transporte, lo que los convierte en modelos adecuados para la estimación. Si comparamos los métodos ARIMA y SARIMA vs otros métodos de generación de series de tiempo como la suavización exponencial, encontramos que este último método estima basado en la media de los consumos históricos para un periodo dado, dando una mayor ponderación a los valores más cercanos en el tiempo, lo que no es lo mejor para modelar flujos de transporte dando menor rendimiento en las estimación pues al manejar rango de tiempo por hora hay puntos donde aunque se evalué una hora valle se dé un alto peso a horas cercanas que sean horas pico alterando la estimación y dando peores rendimientos vs los ARIMA.

Otra ventaja de los modelos ARIMA es que permiten asociar variables que puedan afectar las estimaciones, este modelo se conoce como modelo ARIMAX, para el caso de estudio es muy útil porque además de evaluar los datos pasados para predecir el futuro, podemos revisar, por ejemplo, la incidencia de una variable como la lluvia en las predicciones.

De acuerdo con la empresa Transmilenio es de suma importancia poder describir esta serie de tal manera que las características de demanda le permitan tomar decisiones en la flota que hoy tienen y poder tener un control sobre la oferta del servicio a los usuarios. De este modo estaremos contribuyendo a mejorar la percepción que tienen los usuarios frente a la calidad del servicio.

Para la asignación de la flota en cada una de las estaciones, se ha decidido tomar un enfoque en Simulación basada en agentes, que durante los últimos años se han explorado estas soluciones para el ruteo de vehículos, así como para determinar una programación para empresas de manufactura y servicios. La simulación permitirá obtener un acercamiento a la realidad del sistema con ciertas restricciones que posteriormente serán definidas de acuerdo con la empresa y a variables que también la afectan. Una vez obtenida una simulación muy cercana a la realidad, podrá alterarse al incluirse nuevas demandas al sistema y obtendremos resultados en tiempo real, y estos resultados serán obtenidos con tiempos de computación razonables y la calidad de estos serán muy buenas en comparación a otros enfoques de solución.

4.1.2 Suposiciones del modelo

- **MODELOS ARIMA Y SARIMA:**
  - Para los modelos ARIMA y SARIMA se debe considerar que las series de tiempo son estacionarias, es decir que un comportamiento se repite en intervalos específicos de tiempo. Lo anterior basado en la premisa de que la media y la varianza no cambian con el tiempo.
  - Los residuales siguen una distribución Gaussiana estable.
  - Los parámetros son constantes (Gupta,2018).

4.2 Generación del diseño de prueba

4.2.1 Modelos ARIMA y SARIMA:

- Se realizó una división del 80% de los datos de la serie de tiempo para entrenar el modelo y el 20% para probarlo. Esto no se realizó de manera aleatoria, ya que los datos dependen del tiempo, por lo cual se hizo que los datos más antiguos fueran para entrenar y los siguientes para probar. Así se realizó para cada una de las estaciones, teniendo datos de 11 meses del año 2017.

4.3 Construcción del modelo

Para la construcción del modelo se realizó la verificación de cada una de las series de tiempo de las estaciones, validando si existían atípicos y algunos comportamientos propios por algunas fechas particulares que representaban festivos y eventos que afectaron la movilidad de la ciudadanía por manifestaciones y de otro tipo

como el día sin carro. A continuación, se muestran las series de la estación Bicentenario para la base de entrenamiento:

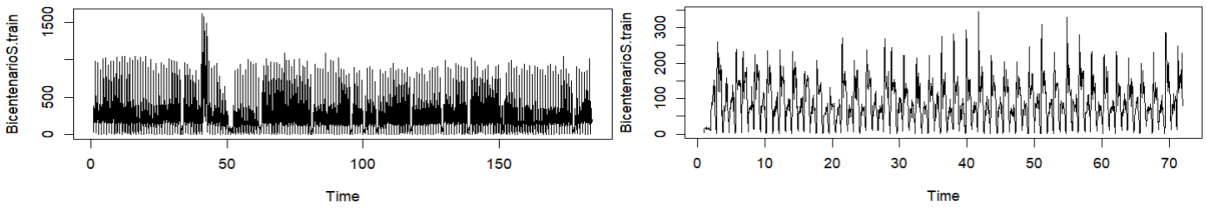

Figura 23 Series de lunes a viernes y de sábados, domingos y festivos.

Cada uno de los dígitos del eje tiempo representa un día, que para el caso particular de Transmilenio son 20 horas, que es cuando el ciclo inicia nuevamente, desde las 4 am hasta las 11 pm.

En muchas de las series se encontró un pico cercano llegando al ciclo 40 en las series de lunes a viernes correspondientes a los días 30 y 31 de marzo, sin embargo, no se encontró alguna explicación dentro de la revisión de noticias y fechas importantes del 2017 que pudiese dar indicios de esta situación anormal frente a los demás datos.

Una vez realizada la revisión de cada una de las series de entrenamiento de la serie, se procedió a validar si la serie de tiempo era estacionaria o no, para lo cual se encontró mediante las pruebas de Box-Ljung y Dickey Fuller que ninguna cumplía con este criterio.

Recordemos que la estacionariedad se refiere a que las propiedades de la serie no varían con respecto al tiempo, esto tiene una importante implicación a la hora de predecir ya que nos indica que el para el modelo las características estadísticas de nuestra serie de tiempo serán las mismas en el futuro como en el pasado, este concepto esta ligado altamente a la obtención de información significativa como media, varianza y autocorrelaciones y permite tener estos datos bien definidos; si la media y varianza no están bien definidas, tampoco lo estarán las autocorrelaciones.

Por lo anterior se usó la prueba de Ljung-Box la cual busca si un grupo cualquiera de autocorrelaciones de una serie de tiempo son diferentes de cero. En lugar de probar la aleatoriedad en cada retardo distinto, esta prueba la aleatoriedad "en general" basado en un número de retardos.

La prueba de Ljung-Box se puede definir de la siguiente manera:

H0: Los datos se distribuyen de forma independiente

Ha: Los datos no se distribuyen de forma independiente.

Para este caso se desea que el valor p de la prueba sea mayor que 0.05 porque esto significa que los residuos de nuestro modelo de series de tiempo son independientes

La prueba Dickey – Fuller busca determinar la existencia o no de raíces unitarias en una serie temporal:

H0: No hay estacionariedad

Ha: Existe estacionariedad aumentada

Esta prueba es una prueba de raíz unitaria para una muestra de una serie de tiempo, donde se busca un número negativo. Cuanto más negativo es, más fuerte es el rechazo de la hipótesis nula de que existe una raíz unitaria para un cierto nivel de confianza.

```
        Augmented Dickey-Fuller Test

data:  BicentenarioS.train
Dickey-Fuller = -17.427, Lag order = 15, p-value = 0.01
alternative hypothesis: stationary
```

```
                Box-Ljung test

data:  BicentenarioS.train
X-squared = 8442.7, df = 20, p-value < 2.2e-16
```

Figura 24 Resultados tests Dickey Fuller y Box-Ljung para la serie estación Bicentenario de lunes a viernes

Como se ha mencionado previamente la estacionariedad se requiere para que las estimaciones de los parámetros sean útiles. De otra forma, no se podrían calcular medias y variancias conforme la serie va creciendo, por tanto, se tomó la determinación de realizar una transformación para poder cumplir con el supuesto de los modelos auto regresivos. La transformación realizada para la serie fue una logaritmación junto con una diferenciación de 5 ciclos para las series de tiempo que constituían los días de lunes a viernes, y una logaritmación junto con una diferenciación de 2 ciclos para las series de tiempo que constituían los sábados, domingos y festivos. Las autocorrelaciones luego de realizar esta transformación muestran que aunque disminuye el efecto a largo plazo, sigue existiendo una fuerte relación con los datos del pasado:

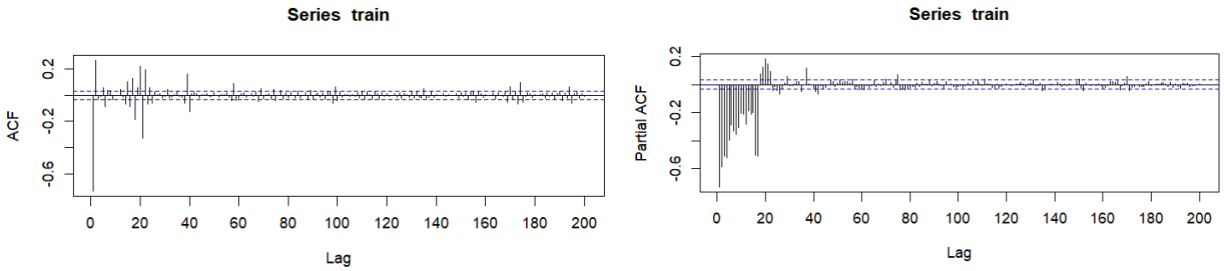

Figura 25 ACF y PACF para las series de entrenamiento de Bicentenario de lunes a viernes

De acuerdo con estos gráficos se evidencia que el término autorregresivo será de orden superior, ya que los patrones de autocorrelación se van repitiendo a lo largo que aumenta el lag.

4.3.1 Configuración de parámetros

Se toman tres modelos base para realizar el pronóstico de cada una de las estaciones en los datos de entrenamiento, los cuales fueron: ingenuo, ARIMA y ARIMAX. Se escogieron los mejores parámetros para cada modelo haciendo una búsqueda a través de funciones de auto.arima y auto.arima con xreg para encontrar los parámetros que disminuyeran el error.

4.3.2 Modelos

Los modelos obtenidos de acuerdo con la búsqueda de ARIMA, SARIMA y ARIMAX, fueron los siguientes:

PARÁMETROS			
ESTACIÓN	DATOS	ARIMA	ARIMAX
Alcalá	L-V	(1,0,0)(1,1,0)[20] with drift	(4,0,3)(0,1,2)[20] errors
	S-D	(5,0,0)(2,0,2)[20] with zero mean	(5,0,0)(2,0,2)[20] errors
AV. Chile	L-V	(1,0,0)(1,1,0)[20] with drift	(1,0,0)(1,1,0)[20] errors
	S-D	(1,0,0)(1,0,2)[20] with zero mean	(1,0,0)(1,0,2)[20] errors
AV. El Dorado	L-V	(2,0,0)(1,0,1)[20] with zero mean	(2,0,0)(1,0,1)[20] errors
	S-D	(5,0,0)(1,0,1)[20] with non-zero mean	(5,0,0)(1,0,1)[20] errors
AV. Primera de Mayo	L-V	(5,0,0)(0,1,2)[20] with drift	(5,0,0)(2,1,0)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(0,0,4)(0,0,1)[20] errors
AV. Rojas	L-V	(1,0,0)(1,0,0)[20] with non-zero mean	(1,0,0)(1,0,0)[20] errors
	S-D	(2,0,0)(0,0,2)[20] with zero mean	(2,0,0)(0,0,2)[20] errors
Avenida Jimenez	L-V	(1,0,0)(1,0,0)[20] with non-zero mean	(0,0,3)(0,0,2)[20] errors
	S-D	(0,0,1) with non-zero mean	(0,0,1) errors
Bicentenario	L-V	(5,0,0)(2,0,2)[20] with non-zero mean	(5,0,0)(2,0,2)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(2,0,2)(2,0,0)[20] errors
Cabecera Usme	L-V	(5,0,0)(0,1,2)[20]	(5,0,0)(0,1,2)[20] errors
	S-D	(4,0,0)(1,0,1)[20] with zero mean	(4,0,0)(1,0,1)[20] errors
CAD	L-V	(5,0,0)(1,0,0)[20] with zero mean	(5,0,0)(1,0,0)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(1,0,0)(1,0,0)[20] errors
Calle 22	L-V	(1,0,0)(2,0,0)[20] with zero mean	(1,0,0)(2,0,0)[20] errors
	S-D	(4,0,3)(0,0,1)[20] with zero mean	(2,0,4)(0,0,2)[20] errors
Calle 26	L-V	(1,0,0)(2,0,0)[20] with zero mean	(1,0,0)(2,0,0)[20] errors
	S-D	(5,0,0)(1,0,1)[20] with non-zero mean	(5,0,0)(1,0,1)[20] errors
Calle 40 Sur	L-V	(5,0,0)(0,1,2)[20]	(5,0,0)(0,1,2)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(1,0,0)(1,0,0)[20] errors
Calle 45	L-V	(1,0,0)(1,0,0)[20]	(1,0,0)(1,0,0)[20] errors
	S-D	(5,0,0)(1,0,1)[20] with non-zero mean	(5,0,0)(1,0,1)[20] errors
Calle 57	L-V	(2,0,0)(1,0,0)[20] with zero mean	(2,0,0)(1,0,0)[20] errors
	S-D	(5,0,0)(2,0,2)[20] with non-zero mean	(5,0,0)(2,0,2)[20] errors
Calle 85	L-V	(5,0,0)(1,0,0)[20] with zero mean	(5,0,0)(1,0,0)[20] errors
	S-D	(1,0,0)(0,0,2)[20] with zero mean	(2,0,2)(1,0,1)[20] errors
CAN	L-V	(5,0,0)(1,0,1)[20] with zero mean	(5,0,0)(1,0,1)[20] errors
	S-D	(1,0,0)(0,0,2)[20] with zero mean	(1,0,0)(0,0,2)[20] errors
Centro Comercial Santafé	L-V	(5,0,0)(2,0,2)[20] with zero mean	(5,0,0)(2,0,2)[20] errors
	S-D	(2,0,3)(0,0,1)[20] with non-zero mean	(1,0,0)(1,0,0)[20] errors
Centro Memoria	L-V	(5,0,0)(1,0,1)[20] with zero mean	(5,0,0)(1,0,1)[20] errors
	S-D	(1,0,0)(0,0,2)[20] with non-zero mean	(2,0,0)(0,0,1)[20] errors
Ciudad Universitaria	L-V	(1,0,0)(1,0,1)[20] with zero mean	(1,0,0)(1,0,1)[20] errors
	S-D	(0,0,4)(0,0,1)[20] with non-zero mean	(0,0,4)(0,0,1)[20] errors
Corferias	L-V	(2,0,3)(0,0,1)[20] with zero mean	(1,0,0)(1,0,0)[20] errors
	S-D	(5,0,0)(0,0,1)[20] with zero mean	(5,0,0)(0,0,1)[20] errors
DESPENSA	L-V	(5,0,0)(2,0,2)[20] with non-zero mean	(5,0,0)(2,0,2)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(5,0,0)(0,0,2)[20] errors
El Tiempo	L-V	(2,0,0)(2,0,0)[20] with zero mean	(2,0,0)(2,0,0)[20] errors
	S-D	(0,0,4) with non-zero mean	(1,0,0)(1,0,0)[20] errors
General Santander	L-V	(2,0,3)(1,1,1)[20] with drift	(3,0,3)(1,1,0)[20] errors
	S-D	(2,0,0)(2,0,0)[20] with zero mean	(2,0,0)(2,0,0)[20] errors
Gobernación	L-V	(5,0,0)(2,0,0)[20] with zero mean	(5,0,0)(2,0,0)[20] errors
	S-D	(5,0,0)(2,0,1)[20] with non-zero mean	(5,0,0)(2,0,1)[20] errors
Hortúa	L-V	(5,0,0)(0,1,2)[20]	(5,0,0)(0,1,2)[20] errors
	S-D	(1,0,0)(0,0,2)[20] with zero mean	(1,0,0)(0,0,2)[20] errors
Hospitales	L-V	(5,0,0)(0,1,2)[20]	(5,0,0)(0,1,2)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(1,0,0)(1,0,0)[20] errors
LEON XIII	L-V	(5,0,0)(2,0,2)[20] with zero mean	(5,0,0)(2,0,2)[20] errors
	S-D	(5,0,0)(2,0,1)[20] with zero mean	(5,0,0)(2,0,1)[20] errors
Modelia	L-V	(5,0,0)(2,0,1)[20] with zero mean	(5,0,0)(2,0,1)[20] errors
	S-D	(5,0,0)(2,0,1)[20] with non-zero mean	(5,0,0)(2,0,1)[20] errors
Molinos	L-V	(5,0,0)(0,1,2)[20]	(5,0,0)(0,1,2)[20] errors
	S-D	(0,0,3)(1,0,2)[20] with non-zero mean	(1,0,0)(0,0,2)[20] errors
Museo Nacional	L-V	(5,0,0)(2,0,2)[20] with zero mean	(5,0,0)(2,0,2)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(0,0,5)(2,0,1)[20] errors
Normandía	L-V	(5,0,0)(1,0,0)[20] with zero mean	(5,0,0)(1,0,0)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(3,0,0)(1,0,1)[20] errors
NQS - CALLE 75	L-V	(2,0,0)(2,0,0)[20] with zero mean	(2,0,0)(2,0,0)[20] errors
	S-D	(1,0,0)(0,0,2)[20] with zero mean	(1,0,0)(0,0,2)[20] errors
NQS - RICAURTE	L-V	(1,0,0)(2,0,0)[20] with zero mean	(1,0,0)(2,0,0)[20] errors
	S-D	(0,0,0)(0,0,2)[20] with non-zero mean	(0,0,0)(0,0,2)[20] errors
Pepe Sierra	L-V	(5,0,0)(1,1,2)[20]	(5,0,0)(1,1,2)[20] errors
	S-D	(5,0,0)(2,0,1)[20] with zero mean	(5,0,0)(2,0,1)[20] errors
Plaza de la Democracia	L-V	(5,0,0)(2,0,0)[20] with zero mean	(5,0,0)(2,0,0)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(1,0,0)(1,0,0)[20] errors
Portal 20 de Julio	L-V	(5,0,0)(0,1,2)[20]	(5,0,0)(0,1,2)[20] errors
	S-D	(3,0,2)(1,0,0)[20] with zero mean	(1,0,0)(1,0,0)[20] errors
Portal Eldorado	L-V	(1,0,4)(0,1,1)[20] with drift	(1,0,0)(1,1,0)[20] errors
	S-D	(0,0,4)(1,0,0)[20] with zero mean	(5,0,0)(2,0,1)[20] errors
Prado	L-V	(2,0,3)(1,1,0)[20]	(1,0,0)(1,1,0)[20] errors
	S-D	(0,0,5) with non-zero mean	(0,0,4) errors
Quinta Paredes	L-V	(5,0,0)(2,0,0)[20] with zero mean	(5,0,0)(2,0,0)[20] errors
	S-D	(5,0,0)(0,0,2)[20] with non-zero mean	(5,0,0)(0,0,2)[20] errors
Restrepo	L-V	(1,0,0)(2,0,0)[20] with zero mean	(1,0,0)(2,0,0)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(2,0,2)(1,0,1)[20] errors
Salitre El Greco	L-V	(5,0,0)(1,0,1)[20] with zero mean	(5,0,0)(1,0,1)[20] errors
	S-D	(5,0,0)(0,0,1)[20] with zero mean	(5,0,0)(0,0,1)[20] errors
SAN MATEO	L-V	(5,0,0)(2,0,2)[20] with zero mean	(5,0,0)(2,0,2)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(1,0,0)(1,0,0)[20] errors
SENA	L-V	(5,0,0)(0,1,2)[20]	(5,0,0)(0,1,2)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(1,0,0)(1,0,0)[20] errors
Tercer Milenio	L-V	(1,0,0)(1,0,0)[20] with non-zero mean	(1,0,0)(1,0,0)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(1,0,0)(1,0,0)[20] errors
Terminal	L-V	(5,0,0)(2,0,2)[20] with zero mean	(5,0,0)(2,0,2)[20] errors
	S-D	(0,0,3)(0,0,1)[20] with zero mean	(1,0,0)(1,0,0)[20] errors
TERREROS	L-V	(5,0,0)(2,0,1)[20] with zero mean	(5,0,0)(2,0,1)[20] errors
	S-D	(5,0,0)(2,0,0)[20] with non-zero mean	(5,0,0)(2,0,0)[20] errors
Toberín	L-V	(5,0,0)(0,1,2)[20]	(5,0,0)(0,1,2)[20] errors
	S-D	(1,0,0)(1,0,1)[20] with non-zero mean	(1,0,0)(1,0,1)[20] errors
Universidades	L-V	(5,0,0)(2,0,2)[20] with zero mean	(5,0,0)(2,0,2)[20] errors
	S-D	(1,0,0)(1,0,0)[20] with non-zero mean	(1,0,0)(1,0,0)[20] errors
VENECIA	L-V	(5,0,0)(0,1,2)[20]	(5,0,0)(0,1,2)[20] errors
	S-D	(1,0,0)(2,0,2)[20] with zero mean	(1,0,0)(2,0,2)[20] errors

Tabla 16 Modelos definidos por estación y días

4.4 Evaluación del modelo

4.4.1 Resultados del modelo

De acuerdo con cada una de las configuraciones de parámetros obtenidas para las 98 series de tiempo, se obtuvieron los siguientes resultados en los datos de prueba determinadas en el objetivo de analítica:

		MODELOS								
		NAIVE			ARIMA			ARIMAX		
ESTACIÓN	MEDIDA DATOS	RMSE	MAE	MASE	RMSE	MAE	MASE	RMSE	MAE	MASE
Alcalá	L-V	8,922549	5,074013	2,553201	2,847098	1,291873	0,6500598	0,6833328	0,3183944	0,1602134
	S-D	2,217654	1,03106	1,10437	1,057908	0,6007109	0,6434227	1,057929	0,6012596	0,6440104
AV. Chile	L-V	10,690810	5,979295	2,337223	2,702341	1,572001	0,614473	2,702459	1,571937	0,6144483
	S-D	1,815861	0,9821854	1,148454	0,9834158	0,5145886	0,6017006	0,9831929	0,5142275	0,6012783
AV. El Dorado	L-V	9,489179	5,907321	1,865091	2,242488	1,362671	0,4302299	2,242504	1,362665	0,4302279
	S-D	1,734278	1,064774	1,214059	0,7762313	0,4896159	0,5582619	0,7759379	0,4895053	0,5581358
AV. Primera de Mayo	L-V	15,554350	7,721033	3,56015	1,457436	0,6845477	0,3156433	1,458403	0,6857076	0,3161781
	S-D	2,194021	1,019456	1,418052	1,171728	0,5306817	0,7381722	0,814441	0,4724602	0,6571867
AV. Rojas	L-V	11,199700	6,434768	2,017584	3,523488	1,979343	0,6206113	3,523099	1,978845	0,6204553
	S-D	1,950779	1,144698	1,148607	0,9666635	0,5806213	0,5826038	0,9666686	0,580509	0,5824911
Avenida Jimenez	L-V	10,902930	6,237041	2,2458	3,298007	1,86358	0,6710276	0,9693152	0,5352699	0,1927371
	S-D	1,722903	0,8342178	1,121089	0,9409218	0,5200676	0,6989084	0,9407731	0,5209807	0,7001355
Bicentenario	L-V	11,435820	6,628901	1,362614	1,851576	1,277931	0,2626872	1,851574	1,277904	0,2626816
	S-D	1,966935	1,130687	1,165407	1,112268	0,6393574	0,6589903	0,7642579	0,4801322	0,4948758
Cabecera Usme	L-V	27,925060	13,84102	4,770534	2,056881	0,895273	0,3085705	2,056889	0,8951801	0,3085385
	S-D	3,669585	1,632443	1,478971	1,556749	0,7299679	0,6613409	1,556741	0,7300312	0,6613983
CAD	L-V	10,016930	5,791118	1,759702	1,520752	0,9581343	0,2911408	1,52075	0,9581327	0,2911403
	S-D	1,948865	1,28861	1,18462	1,121162	0,7121729	0,6547011	1,121193	0,7122208	0,6547452
Calle 22	L-V	6,009244	4,127531	1,907191	1,919586	1,232822	0,5696449	1,919575	1,232884	0,5696736
	S-D	1,549077	0,8996817	1,246623	0,6421843	0,3537692	0,4901921	0,6461109	0,3586225	0,496917
Calle 26	L-V	9,425576	5,421735	2,095203	2,930495	1,654554	0,6393942	2,930875	1,654584	0,6394058
	S-D	1,902663	1,137874	1,182977	0,8993426	0,5484338	0,5701724	0,8993719	0,5487956	0,5705486
Calle 40 Sur	L-V	16,344150	8,156055	3,361407	1,458948	0,6835111	0,2816998	1,458946	0,6835115	0,2816999
	S-D	2,143180	1,081295	1,443614	1,228291	0,5979181	0,798268	1,228295	0,5979158	0,7982649
Calle 45	L-V	7,842471	4,384595	2,242236	2,436035	1,265297	0,6470597	2,435831	1,264932	0,6468729
	S-D	1,807167	0,9039149	1,121777	0,833223	0,4547075	0,5643015	0,8331533	0,454834	0,5644585
Calle 57	L-V	8,342291	4,698427	2,08186	1,894018	1,050856	0,4656315	1,894038	1,050874	0,4656392
	S-D	1,845816	0,9684463	1,108519	0,8406213	0,4971092	0,5690096	0,8407747	0,4973689	0,5693068
Calle 85	L-V	11,830290	5,917033	1,909198	1,720559	0,9422336	0,3040223	1,720441	0,9423108	0,3040473
	S-D	2,006620	1,064523	1,233788	1,216578	0,6531185	0,7569681	0,8385813	0,4865118	0,5638699
CAN	L-V	11,636270	6,848384	1,644435	1,98237	1,3281	0,3189035	1,982401	1,328081	0,3189037
	S-D	2,038617	1,306736	1,111029	1,15652	0,7163845	0,6090934	1,156446	0,7162045	0,6089404
Centro Comercial Santafé	L-V	16,349800	8,042741	0,002541	2,306784	1,414443	0,3521786	2,306797	1,414431	0,3521754
	S-D	2,562219	1,172861	1,134123	0,9563259	0,550467	0,5322855	1,453469	0,7133843	0,6898217
Centro Memoria	L-V	10,312740	6,343035	2,518825	1,352949	0,919888	0,3652884	1,352959	0,9198426	0,3652704
	S-D	1,906056	1,098511	1,156518	0,9837319	0,5751098	0,6054783	0,88432	0,5304796	0,5584913
Ciudad Universitaria	L-V	10,066730	6,192197	2,180171	3,035745	1,658565	0,5839534	3,035741	1,658522	0,5839384
	S-D	1,763418	1,053548	1,107912	0,7229523	0,4437475	0,4666455	0,727171	0,4504472	0,473691
Corferias	L-V	11,950040	7,189706	1,978862	1,007037	0,584505	0,1608765	3,591508	2,079138	0,5722524
	S-D	1,956699	1,138848	1,142534	0,8877331	0,5295065	0,53122	0,8876453	0,5294782	0,5311917
DESPENSA	L-V	12,079630	7,111969	1,173319	1,941481	1,35242	0,2231197	1,941448	1,352401	0,2231165
	S-D	1,909997	1,153584	1,337973	0,9624184	0,5758792	0,6679273	0,8177787	0,5218891	0,6053075
El Tiempo	L-V	9,918714	5,468338	2,124537	2,096381	1,230616	0,4781141	2,096282	1,23049	0,478065
	S-D	2,043768	1,111138	1,231403	0,7937605	0,4699566	0,5208227	1,110136	0,5948683	0,6592543
General Santander	L-V	14,630680	7,507025	3,011897	0,6988656	0,3189299	0,127958	0,680707	0,3302041	0,1324814
	S-D	1,660785	0,8967664	1,278098	0,7868928	0,4486362	0,6394095	0,7868884	0,448642	0,6394178
Gobernación	L-V	9,894566	6,069817	1,701685	1,668079	1,201226	0,3367662	1,668075	1,201241	0,3367702
	S-D	1,849886	1,360965	1,225776	0,7763125	0,5701928	0,5135536	0,7762737	0,5704094	0,5137487
Hortúa	L-V	13,103020	6,941739	2,633536	1,29042	0,7567592	0,287097	1,290414	0,756764	0,2870988
	S-D	2,028519	1,080664	1,115196	1,124332	0,593971	0,681086	1,124289	0,617860	0,681103
Hospitales	L-V	1,363655	7,649226	2,096328	1,832495	1,128447	0,3092594	1,832532	1,128544	0,3092861
	S-D	2,293212	1,278392	1,166812	1,143035	0,651747	0,594862	1,142934	0,651963	0,595059
LEON XIII	L-V	17,485150	8,415913	1,283159	2,601868	1,663417	0,253618	2,601885	1,663334	0,253606
	S-D	2,570951	1,191524	1,285385	1,096576	0,612256	0,660486	1,096507	0,611893	0,660094
Modelia	L-V	12,978210	7,8324661	2,160689	1,998061	1,412616	0,389580	2,008190	1,420479	0,391748
	S-D	2,086421	1,304235	1,059877	0,949899	0,620181	0,503985	0,949772	0,619846	0,503714
Molinos	L-V	18,357650	9,247761	3,926354	1,538788	0,694499	0,294866	1,538900	0,694396	0,294822
	S-D	2,383136	1,136429	1,492741	0,866180	0,484528	0,636445	1,246071	0,594164	0,780456
Museo Nacional	L-V	15,884620	8,691214	1,953720	2,241205	1,419219	0,320932	2,241271	1,419311	0,320953
	S-D	2,560184	1,384240	1,134622	1,399446	0,771461	0,632345	0,948423	0,542026	0,444283
Normandía	L-V	9,131531	5,772793	2,223524	1,310986	0,859290	0,330975	1,310996	0,859303	0,330998
	S-D	1,915006	1,152126	1,162858	1,070161	0,629641	0,635508	0,929985	0,568189	0,573482
NQS - CALLE 75	L-V	8,895859	5,735591	1,931875	2,150349	1,317477	0,4437556	2,150338	1,31747	0,4437534
	S-D	1,851032	1,092775	1,213979	0,9804624	0,5852719	0,650184	0,9834637	0,5862513	0,6512751
NQS - RICAURTE	L-V	7,954168	5,257724	2,0632	2,502513	1,434804	0,5630361	2,502733	1,434997	0,5631119
	S-D	1,632544	0,9842305	1,212452	0,9870744	0,5975332	0,736088	0,9870149	0,5975647	0,7361268
Pepe Sierra	L-V	8,707045	5,118978	2,13268	1,353218	0,7357029	0,3065102	1,353237	0,7357058	0,3065114
	S-D	1,789355	0,984267	1,097795	0,8155088	0,4639795	0,5175012	0,8154395	0,4638303	0,5173348
Plaza de la Democracia	L-V	9,323459	6,112389	1,619667	1,767739	1,347357	0,3570251	1,76776	1,347369	0,3570271
	S-D	1,875376	1,286624	1,182324	0,9994514	0,657625	0,6043151	0,9994501	0,6576474	0,6043357
Portal 20 de Julio	L-V	26,814950	13,11319	4,518892	2,118706	0,965679	0,3327793	2,118793	0,965766	0,3328093
	S-D	3,292700	1,461192	1,509208	1,067872	0,6039366	0,6237827	1,669783	0,758161	0,7830751
Portal Eldorado	L-V	21,851290	10,58977	3,891231	0,8588035	0,3830684	0,1407592	3,816677	1,772614	0,65135
	S-D	3,433006	1,508555	1,344675	1,079773	0,5691867	0,5073539	1,32557	0,7574	0,6751209
Prado	L-V	10,524850	5,563594	2,164302	0,8144619	0,3980171	0,1548332	3,200439	1,630663	0,6343469
	S-D	1,895042	1,047506	1,088657	0,7545577	0,4557948	0,4737008	0,7631843	0,4647888	0,4830481
Quinta Paredes	L-V	12,901590	7,512903	1,623532	2,076165	1,439282	0,3110275	20,76185	1,43931	0,3110336
	S-D	2,180874	1,394783	1,197858	0,8909675	0,5607091	0,481544	0,890591	0,5602579	0,4811569
Restrepo	L-V	8,043325	3,967277	2,115139	2,518427	1,079698	0,5756371	2,518428	1,079719	0,5756482
	S-D	2,339482	1,019245	1,118287	1,232599	0,5384437	0,5907653	0,8446001	0,4267586	0,4682275
Salitre El Greco	L-V	10,957580	6,145908	1,880434	1,783155	1,046378	0,3201552	1,783153	1,046368	0,3201521
	S-D	1,783445	0,9446567	1,158783	0,8527979	0,5009497	0,6145003	0,8527274	0,501576	0,6152685
SAN MATEO	L-V	28,169940	12,74947	1,771898	3,4111	2,120968	0,2947682	3,411175	2,120386	0,2946873
	S-D	3,610455	1,448246	1,333273	1,726018	0,7445833	0,6854728	1,726008	0,7445861	0,6854752
SENA	L-V	10,771590	5,69071	2,466118	1,310694	0,7256699	0,3144753	1,310684	0,7256996	0,3144882
	S-D	1,607778	0,8117423	1,291038	0,9102523	0,4712437	0,7494912	0,9271034	0,4814612	0,7657416
Tercer Milenio	L-V	11,780600	6,470994	1,6682	3,520551	2,07863	0,5358636	3,520552	2,078634	0,5358647
	S-D	1,987820	1,03846	1,178893	1,025637	0,5434519	0,6169439	1,02587	0,5435277	0,6170299
Terminal	L-V	9,952507	5,930219	1,215001	2,021352	1,351962	0,276994	2,021313	1,351803	0,2769615
	S-D	2,406830	1,148407	0,9784249	0,9501464	0,6043443	0,514892	1,275493	0,6368159	0,5425573
TERREROS	L-V	24,026880	11,13193	1,915119	2,94251	1,904214	0,3275977	2,951266	1,90398	0,3275575
	S-D	3,389835	1,421817	1,392809	1,351726	0,7078225	0,6933812	1,3517	0,7077923	0,6933516
Toberín	L-V	17,186790	8,5129	2,881941	1,859616	1,003608	0,3397594	1,85608	1,003648	0,339773
	S-D	2,415648	1,167152	1,184754	1,311308	0,6786207	0,6888609	1,311198	0,6787633	0,6890056
Universidades	L-V	14,600780	7,717058	1,566398	2,409253	1,461781	0,2967104	2,422584	1,472709	0,2989286
	S-D	2,186133	1,129513	1,233885	1,219381	0,6404172	0,6995946	1,219279	0,6404071	0,6995835
VENECIA	L-V	13,588540	6,892418	3,366881	1,108274	0,6205839	0,3031494	1,108317	0,6204552	0,3030865
	S-D	2,066628	0,9474039	1,373964	0,9587508	0,4892911	0,7095901	0,9588461	0,4893487	0,7096735

Tabla 17 Resultados de los modelos Naive, ARIMA, SARIMA, ARIMAX por estación y días

Se observa que el modelo base o ingenuo es un mal predictor, esto se ve en los indicadores para todas las estaciones y jornadas.

El MASE muestra que para el modelo ingenuo todos los valores son superiores a 1, contrario a este comportamiento los modelos ARIMA y ARIMAX tiene valores de MASE menores a 1 para todos los casos, lo que indican que es un mejor predictor vs el modelo ingenuo.

Similar a lo que pasa con el MASE pasa con los valores del RMSE y del MAE, donde se disminuye drásticamente el valor de cada indicador, por ejemplo, para el portal Usme (cabecera usme) el RMSE y el MAE es aproximadamente 13 veces menor con los modelos ARIMA y ARIMAX en comparación con el modelo ingenuo.

4.4.2 Revisión de los parámetros configurados

A partir de estos resultados se escogió por estación y serie de tiempo el modelo que disminuyera la mayor cantidad de medidas de desempeño, asegurando así que se cumpliera con el objetivo de analítica junto con los criterios de éxito. Obteniendo los siguientes modelos:

ESTACIÓN	MEDIDA DATOS	MODELO SELECCIONADO	PARÁMETROS
Alcalá	L-V	ARIMAX	(4,0,3)(0,1,2)[20] errors
	S-D	ARIMA	(5,0,0)(2,0,2)[20] with zero mean
AV. Chile	L-V	ARIMAX	(1,0,0)(1,1,0)[20] errors
	S-D	ARIMAX	(1,0,0)(1,0,2)[20] errors
AV. El Dorado	L-V	ARIMAX	(2,0,0)(1,0,1)[20] errors
	S-D	ARIMAX	(5,0,0)(1,0,1)[20] errors
AV. Primera de Mayo	L-V	ARIMA	(5,0,0)(0,1,2)[20] with drift
	S-D	ARIMAX	(0,0,4)(0,0,1)[20] errors
AV. Rojas	L-V	ARIMAX	(1,0,0)(1,0,0)[20] errors
	S-D	ARIMAX	(2,0,0)(0,0,2)[20] errors
Avenida Jimenez	L-V	ARIMAX	(0,0,3)(0,0,2)[20] errors
	S-D	ARIMA	(0,0,1) with non-zero mean
Bicentenario	L-V	ARIMAX	(5,0,0)(2,0,2)[20] errors
	S-D	ARIMAX	(2,0,2)(2,0,0)[20] errors
Cabecera Usme	L-V	ARIMAX	(5,0,0)(0,1,2)[20] errors
	S-D	ARIMA	(4,0,0)(1,0,1)[20] with zero mean
CAD	L-V	ARIMAX	(5,0,0)(1,0,0)[20] errors
	S-D	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean
Calle 22	L-V	ARIMA	(1,0,0)(2,0,0)[20] with zero mean
	S-D	ARIMA	(4,0,3)(0,0,1)[20] with zero mean
Calle 26	L-V	ARIMA	(1,0,0)(2,0,0)[20] with zero mean
	S-D	ARIMA	(5,0,0)(1,0,1)[20] with non-zero mean
Calle 40 Sur	L-V	ARIMA	(5,0,0)(0,1,2)[20]
	S-D	ARIMAX	(1,0,0)(1,0,0)[20] errors
Calle 45	L-V	ARIMAX	(1,0,0)(1,0,0)[20] errors
	S-D	ARIMA	(5,0,0)(1,0,1)[20] with non-zero mean
Calle 57	L-V	ARIMA	(2,0,0)(1,0,0)[20] with zero mean
	S-D	ARIMA	(5,0,0)(2,0,2)[20] with non-zero mean
Calle 85	L-V	ARIMA	(5,0,0)(1,0,0)[20] with zero mean
	S-D	ARIMAX	(2,0,2)(1,0,1)[20] errors
CAN	L-V	ARIMA	(5,0,0)(1,0,1)[20] with zero mean
	S-D	ARIMAX	(1,0,0)(0,0,2)[20] errors
Centro Comercial Santafé	L-V	NAIVE	N/A
	S-D	ARIMA	(2,0,3)(0,0,1)[20] with non-zero mean
Centro Memoria	L-V	ARIMAX	(5,0,0)(1,0,1)[20] errors
	S-D	ARIMAX	(2,0,0)(0,0,1)[20] errors
Ciudad Universitaria	L-V	ARIMAX	(1,0,0)(1,0,1)[20] errors
	S-D	ARIMA	(0,0,4)(0,0,1)[20] with non-zero mean
Corferias	L-V	ARIMA	(2,0,3)(0,0,1)[20] with zero mean
	S-D	ARIMAX	(5,0,0)(0,0,1)[20] errors
DESPENSA	L-V	ARIMAX	(5,0,0)(2,0,2)[20] errors
	S-D	ARIMAX	(5,0,0)(0,0,2)[20] errors
El Tiempo	L-V	ARIMAX	(2,0,0)(2,0,0)[20] errors
	S-D	ARIMA	(0,0,4) with non-zero mean
General Santander	L-V	ARIMA	(2,0,3)(1,1,1)[20] with drift
	S-D	ARIMA	(2,0,0)(2,0,0)[20] with zero mean
Gobernación	L-V	ARIMA	(5,0,0)(2,0,0)[20] with zero mean
	S-D	ARIMA	(5,0,0)(2,0,1)[20] with non-zero mean
Hortúa	L-V	ARIMA	(5,0,0)(0,1,2)[20]
	S-D	ARIMA	(1,0,0)(0,0,2)[20] with zero mean
Hospitales	L-V	ARIMA	(5,0,0)(0,1,2)[20]
	S-D	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean
LEON XIII	L-V	ARIMAX	(5,0,0)(2,0,2)[20] errors
	S-D	ARIMAX	(5,0,0)(2,0,1)[20] errors
Modelia	L-V	ARIMA	(5,0,0)(2,0,1)[20] with zero mean
	S-D	ARIMAX	(5,0,0)(2,0,1)[20] errors
Molinos	L-V	ARIMAX	(5,0,0)(0,1,2)[20] errors
	S-D	ARIMA	(0,0,3)(1,0,2)[20] with non-zero mean
Museo Nacional	L-V	ARIMA	(5,0,0)(2,0,2)[20] with zero mean
	S-D	ARIMAX	(0,0,5)(2,0,1)[20] errors
Normandía	L-V	ARIMA	(5,0,0)(1,0,0)[20] with zero mean
	S-D	ARIMAX	(3,0,0)(1,0,1)[20] errors
NQS - CALLE 75	L-V	ARIMAX	(2,0,0)(2,0,0)[20] errors
	S-D	ARIMA	(1,0,0)(0,0,2)[20] with zero mean
NQS - RICAURTE	L-V	ARIMA	(1,0,0)(2,0,0)[20] with zero mean
	S-D	ARIMA	(0,0,0)(0,0,2)[20] with non-zero mean
Pepe Sierra	L-V	ARIMA	(5,0,0)(1,1,2)[20]
	S-D	ARIMAX	(5,0,0)(2,0,1)[20] errors
Plaza de la Democracia	L-V	ARIMA	(5,0,0)(2,0,0)[20] with zero mean
	S-D	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean
Portal 20 de Julio	L-V	ARIMA	(5,0,0)(0,1,2)[20]
	S-D	ARIMA	(3,0,2)(1,0,0)[20] with zero mean
Portal Eldorado	L-V	ARIMA	(1,0,4)(0,1,1)[20] with drift
	S-D	ARIMA	(0,0,4)(1,0,0)[20] with zero mean
Prado	L-V	ARIMA	(2,0,3)(1,1,0)[20]
	S-D	ARIMA	(0,0,5) with non-zero mean
Quinta Paredes	L-V	ARIMA	(5,0,0)(2,0,0)[20] with zero mean
	S-D	ARIMAX	(5,0,0)(0,0,2)[20] errors
Restrepo	L-V	ARIMA	(1,0,0)(2,0,0)[20] with zero mean
	S-D	ARIMAX	(2,0,2)(1,0,1)[20] errors
Salitre El Greco	L-V	ARIMAX	(5,0,0)(1,0,1)[20] errors
	S-D	ARIMA	(5,0,0)(0,0,1)[20] with zero mean
SAN MATEO	L-V	ARIMAX	(5,0,0)(2,0,2)[20] errors
	S-D	ARIMA	(1,0,1,0)(0,0)[20] with non-zero mean
SENA	L-V	ARIMA	(5,0,0)(0,1,2)[20]
	S-D	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean
Tercer Milenio	L-V	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean
	S-D	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean
Terminal	L-V	ARIMAX	(5,0,0)(2,0,2)[20] errors
	S-D	ARIMA	(0,0,3)(0,0,1)[20] with zero mean
TERREROS	L-V	ARIMAX	(5,0,0)(2,0,1)[20] errors
	S-D	ARIMAX	(5,0,0)(2,0,0)[20] errors
Toberín	L-V	ARIMA	(5,0,0)(0,1,2)[20]
	S-D	ARIMA	(1,0,0)(1,0,1)[20] with non-zero mean
Universidades	L-V	ARIMA	(5,0,0)(2,0,2)[20] with zero mean
	S-D	ARIMAX	(1,0,0)(1,0,0)[20] errors
VENECIA	L-V	ARIMAX	(5,0,0)(0,1,2)[20] errors
	S-D	ARIMA	(1,0,0)(2,0,2)[20] with zero mean

Tabla 18 Modelos y parámetros por estación y días

El 42% de los modelos que mejores resultados da son ARIMAX, lo que indica que la lluvia si es un factor que incide sobre las entradas de los usuarios al sistema y vale la pena tenerla en cuenta para las predicciones, pero solo sobre estaciones específicas. El 58% restante son modelos ARIMA.

5. EVALUACIÓN

5.1 Evaluación de resultados

5.1.1 Evaluación de resultados de analítica frente a los criterios de éxito de negocio

De acuerdo con cada uno de los modelos y parámetros obtenidos, los resultados en las medidas de desempeño de RMSE, MAE y MASE fueron los siguientes:

				MEDIDAS EN DATOS DE PRUEBA		
ESTACIÓN	MEDIDA DATOS	MODELO SELECCIONADO	PARÁMETROS	RMSE	MAE	MASE
Alcalá	L-V	ARIMAX	(4,0,3)(0,1,2)[20] errors	0,6833328	0,3183944	0,1602134
	S-D	ARIMA	(5,0,0)(2,0,2)[20] with zero mean	1,057929	0,6012596	0,6440104
AV. Chile	L-V	ARIMAX	(1,0,0)(1,1,0)[20] errors	2,702459	1,571937	0,6144483
	S-D	ARIMAX	(1,0,0)(1,0,2)[20] errors	0,9831929	0,5142275	0,6012783
AV. El Dorado	L-V	ARIMAX	(2,0,0)(1,0,1)[20] errors	2,242504	1,362665	0,4302279
	S-D	ARIMAX	(5,0,0)(1,0,1)[20] errors	0,7759379	0,4895053	0,5581358
AV. Primera de Mayo	L-V	ARIMA	(5,0,0)(0,1,2)[20] with drift	1,458403	0,6857076	0,3161781
	S-D	ARIMAX	(0,0,4)(0,0,1)[20] errors	0,814441	0,4724602	0,6571867
AV. Rojas	L-V	ARIMAX	(1,0,0)(1,0,0)[20] errors	3,523099	1,978845	0,6204553
	S-D	ARIMAX	(2,0,0)(0,0,2)[20] errors	0,9666686	0,580509	0,5824911
Avenida Jimenez	L-V	ARIMAX	(0,0,3)(0,0,2)[20] errors	0,9693152	0,5352699	0,1927371
	S-D	ARIMA	(0,0,1) with non-zero mean	0,9407731	0,5209807	0,7001355
Bicentenario	L-V	ARIMAX	(5,0,0)(2,0,2)[20] errors	1,851574	1,277904	0,2626816
	S-D	ARIMAX	(2,0,2)(2,0,0)[20] errors	0,7642579	0,4801322	0,4948758
Cabecera Usme	L-V	ARIMAX	(5,0,0)(0,1,2)[20] errors	2,056889	0,8951801	0,3085385
	S-D	ARIMA	(4,0,0)(1,0,1)[20] with zero mean	1,556741	0,7300312	0,6613983
CAD	L-V	ARIMAX	(5,0,0)(1,0,0)[20] errors	1,52075	0,9581327	0,2911403
	S-D	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean	1,121193	0,7122208	0,6547452
Calle 22	L-V	ARIMA	(1,0,0)(2,0,0)[20] with zero mean	1,919575	1,232884	0,5696736
	S-D	ARIMA	(4,0,3)(0,0,1)[20] with zero mean	0,6461109	0,3586225	0,496917
Calle 26	L-V	ARIMA	(1,0,0)(2,0,0)[20] with zero mean	2,938875	1,654584	0,6394058
	S-D	ARIMA	(5,0,0)(1,0,1)[20] with non-zero mean	0,8993719	0,5487956	0,5705486
Calle 40 Sur	L-V	ARIMA	(5,0,0)(0,1,2)[20]	1,458946	0,6835115	0,2816999
	S-D	ARIMAX	(1,0,0)(1,0,0)[20] errors	1,228295	0,5979158	0,7982649
Calle 45	L-V	ARIMAX	(1,0,0)(1,0,0)[20] errors	2,435831	1,264932	0,6468729
	S-D	ARIMA	(5,0,0)(1,0,1)[20] with non-zero mean	0,8331533	0,454834	0,5644585
Calle 57	L-V	ARIMA	(2,0,0)(1,0,0)[20] with zero mean	1,894038	1,050874	0,4656392
	S-D	ARIMA	(5,0,0)(2,0,2)[20] with non-zero mean	0,8407747	0,4973689	0,5693068
Calle 85	L-V	ARIMA	(5,0,0)(1,0,0)[20] with zero mean	1,720441	0,9423108	0,3040473
	S-D	ARIMAX	(2,0,2)(1,0,1)[20] errors	0,8385813	0,4865118	0,5638699
CAN	L-V	ARIMA	(5,0,0)(1,0,1)[20] with zero mean	1,982401	1,328081	0,3189037
	S-D	ARIMAX	(1,0,0)(0,0,2)[20] errors	1,156446	0,7162045	0,6089404
Centro Comercial Santafé	L-V	NAIVE	N/A	2,306797	1,414431	0,3521754
	S-D	ARIMA	(2,0,3)(0,0,1)[20] with non-zero mean	1,453469	0,7133843	0,6898217
Centro Memoria	L-V	ARIMAX	(5,0,0)(1,0,1)[20] errors	1,352959	0,9198426	0,3652704
	S-D	ARIMAX	(2,0,0)(0,0,1)[20] errors	0,88432	0,5304796	0,5584913
Ciudad Universitaria	L-V	ARIMAX	(1,0,0)(1,0,1)[20] errors	3,035741	1,658522	0,5839384
	S-D	ARIMA	(0,0,4)(0,0,1)[20] with non-zero mean	0,727171	0,4504472	0,473691
Corferias	L-V	ARIMA	(2,0,3)(0,0,1)[20] with zero mean	3,591508	2,079138	0,5722524
	S-D	ARIMAX	(5,0,0)(0,0,1)[20] errors	0,8876453	0,5294782	0,5311917
DESPENSA	L-V	ARIMAX	(5,0,0)(2,0,2)[20] errors	1,941448	1,352401	0,2231165
	S-D	ARIMAX	(5,0,0)(0,0,2)[20] errors	0,8177787	0,5218891	0,6053075
El Tiempo	L-V	ARIMAX	(2,0,0)(2,0,0)[20] errors	2,096282	1,23049	0,478065
	S-D	ARIMA	(0,0,4) with non-zero mean	1,110136	0,5948683	0,6592543
General Santander	L-V	ARIMA	(2,0,3)(1,1,1)[20] with drift	0,680707	0,3302041	0,1324814
	S-D	ARIMA	(2,0,0)(2,0,0)[20] with zero mean	0,7868884	0,448642	0,6394178
Gobernación	L-V	ARIMA	(5,0,0)(2,0,0)[20] with zero mean	1,668075	1,201241	0,3367702
	S-D	ARIMA	(5,0,0)(2,0,1)[20] with non-zero mean	0,7762737	0,5704094	0,5137487
Hortúa	L-V	ARIMA	(5,0,0)(0,1,2)[20]	1,290414	0,756764	0,2870988
	S-D	ARIMA	(1,0,0)(0,0,2)[20] with zero mean	1,124289	0,6178598	0,681103
Hospitales	L-V	ARIMA	(5,0,0)(0,1,2)[20]	1,832532	1,128544	0,3092861
	S-D	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean	1,142934	0,6519634	0,5950592
LEON XIII	L-V	ARIMAX	(5,0,0)(2,0,2)[20] errors	2,601885	1,663334	0,2536056
	S-D	ARIMAX	(5,0,0)(2,0,1)[20] errors	1,096507	0,6118926	0,6600939
Modelia	L-V	ARIMA	(5,0,0)(2,0,1)[20] with zero mean	2,00819	1,420479	0,3917481
	S-D	ARIMAX	(5,0,0)(2,0,1)[20] errors	0,9497716	0,6198464	0,5037136
Molinos	L-V	ARIMAX	(5,0,0)(0,1,2)[20] errors	1,5389	0,694396	0,2948221
	S-D	ARIMA	(0,0,3)(1,0,2)[20] with non-zero mean	1,246071	0,5941644	0,7804564
Museo Nacional	L-V	ARIMA	(5,0,0)(2,0,2)[20] with zero mean	2,241271	1,419311	0,3209532
	S-D	ARIMAX	(0,0,5)(2,0,1)[20] errors	0,9484233	0,5420256	0,4442831
Normandía	L-V	ARIMA	(5,0,0)(1,0,0)[20] with zero mean	1,310996	0,8593027	0,33099802
	S-D	ARIMAX	(3,0,0)(1,0,1)[20] errors	0,9299852	0,5681893	0,5734818
NQS - CALLE 75	L-V	ARIMAX	(2,0,0)(2,0,0)[20] errors	2,150338	1,31747	0,4437534
	S-D	ARIMA	(1,0,0)(0,0,2)[20] with zero mean	0,9834637	0,5862513	0,6512751
NQS - RICAURTE	L-V	ARIMA	(1,0,0)(2,0,0)[20] with zero mean	2,502733	1,434997	0,5631119
	S-D	ARIMA	(0,0,0)(0,0,2)[20] with non-zero mean	0,9870149	0,5975647	0,7361268
Pepe Sierra	L-V	ARIMA	(5,0,0)(1,1,2)[20]	1,353237	0,7357058	0,3065114
	S-D	ARIMAX	(5,0,0)(2,0,1)[20] errors	0,8154395	0,4638303	0,5173348
Plaza de la Democracia	L-V	ARIMA	(5,0,0)(2,0,0)[20] with zero mean	1,76776	1,347369	0,3570271
	S-D	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean	0,9994501	0,6576474	0,6043357
Portal 20 de Julio	L-V	ARIMA	(5,0,0)(0,1,2)[20]	2,118793	0,965766	0,3328093
	S-D	ARIMA	(3,0,2)(1,0,0)[20] with zero mean	1,669783	0,758161	0,7830751
Portal Eldorado	L-V	ARIMA	(1,0,4)(0,1,1)[20] with drift	3,816677	1,772614	0,65135
	S-D	ARIMA	(0,0,4)(1,0,0)[20] with zero mean	1,32557	0,7574	0,5751209
Prado	L-V	ARIMA	(2,0,3)(1,1,0)[20]	3,200439	1,630663	0,6343469
	S-D	ARIMA	(0,0,5) with non-zero mean	0,7631843	0,4647888	0,4830481
Quinta Paredes	L-V	ARIMA	(5,0,0)(2,0,0)[20] with zero mean	20,76185	1,43931	0,3110336
	S-D	ARIMAX	(5,0,0)(0,0,2)[20] errors	0,890591	0,5602579	0,4811569
Restrepo	L-V	ARIMA	(1,0,0)(2,0,0)[20] with zero mean	2,518428	1,079719	0,5756482
	S-D	ARIMAX	(2,0,2)(1,0,1)[20] errors	0,8446001	0,4267586	0,4682275
Salitre El Greco	L-V	ARIMAX	(5,0,0)(1,0,1)[20] errors	1,783153	1,046368	0,3201521
	S-D	ARIMA	(5,0,0)(0,0,1)[20] with zero mean	0,8527274	0,501576	0,6152685
SAN MATEO	L-V	ARIMAX	(5,0,0)(2,0,2)[20] errors	3,411175	2,120386	0,2946873
	S-D	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean	1,726008	0,7445861	0,6854752
SENA	L-V	ARIMA	(5,0,0)(0,1,2)[20]	1,310684	0,7256996	0,3144882
	S-D	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean	0,9271034	0,4814612	0,7657416
Tercer Milenio	L-V	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean	3,520552	2,078634	0,5358647
	S-D	ARIMA	(1,0,0)(1,0,0)[20] with non-zero mean	1,02587	0,5435277	0,6170299
Terminal	L-V	ARIMAX	(5,0,0)(2,0,2)[20] errors	2,021313	1,351803	0,2769615
	S-D	ARIMA	(0,0,3)(0,0,1)[20] with zero mean	1,275493	0,6368159	0,5425573
TERREROS	L-V	ARIMAX	(5,0,0)(2,0,1)[20] errors	2,951266	1,90398	0,3275575
	S-D	ARIMAX	(5,0,0)(2,0,0)[20] errors	1,3517	0,7077923	0,6933516
Toberín	L-V	ARIMA	(5,0,0)(0,1,2)[20]	1,85608	1,003648	0,339773
	S-D	ARIMA	(1,0,0)(1,0,1)[20] with non-zero mean	1,311198	0,6787633	0,6890056
Universidades	L-V	ARIMA	(5,0,0)(2,0,2)[20] with zero mean	2,422584	1,472709	0,2989286
	S-D	ARIMAX	(1,0,0)(1,0,0)[20] errors	1,219279	0,6404071	0,6995835
VENECIA	L-V	ARIMAX	(5,0,0)(0,1,2)[20] errors	1,108317	0,6204552	0,3030865
	S-D	ARIMA	(1,0,0)(2,0,2)[20] with zero mean	0,9588461	0,4893487	0,7096735

Tabla 19 Resultados en datos de prueba por estación y series de tiempo

El valor del indicador MASE para todos los casos es menor a 1, siendo el valor más bajo 0,13 para la estación General Santander, esto nos confirma que los modelos funcionan para hacer una mejor predicción que la de un modelo ingenuo como lo puede ser una predicción usando Naive. El valor más alto es de 0,7 que sigue siendo un buen valor de predicción al compararlo contra el Naive.

Los modelos de lunes a viernes tienen en promedio un RMSE de 2,09, un MAE de 1,22 y un MASE de 0,39, mientras que los modelos de fines de semana tienen un RMSE de 1,02, un MAE de 0,57 y un MASE de 0,61, esto nos indica que, aunque los modelos de sábados y domingos tienen un peor MASE tienen unos indicadores de error más bajos al compararlos contra los modelos de lunes a viernes.

Tanto los modelos ARIMA como los ARIMAX predicen mejor que el modelo ingenuo y además han disminuido los índices de error indicando que las estimaciones son aceptables.

5.2 Pronóstico

A partir de los datos obtenidos en la parametrización de los modelos se realizó el pronóstico del mes siguiente para cada una de las estaciones. Estos datos fueron exportados en formato .xlsx para que se hiciera el cargue de información dentro de la simulación. En total fueron pronosticados 620 datos correspondientes a 20 horas por 31 días.

6. SIMULACIÓN BASADA EN AGENTES

Los sistemas de transporte pueden considerar incertidumbre dentro de la operación, ya que hay factores ajenos, cambios e interrupciones que pueden generar disturbios dentro de estos. Los problemas que incorporan la incertidumbre se clasifican dentro del tipo estocástico. Algunas de las variables como el número de pasajeros que ingresan, los tiempos de desplazamiento de la flota, el estado de las vías y el tráfico en general varían en función del tiempo, siendo alterados por otros factores adicionales.

Al ser este un problema de tipo estocástico, las soluciones y/o el entendimiento al que se quiere llegar del sistema deben ser aproximaciones a la realidad del sistema utilizando metodologías que permitan cambiar parámetros, establecer escenarios y encontrar posibles soluciones de acuerdo con un análisis de costo beneficio. Algunos métodos de solución pueden ser de tipo exacto, para lo cual se esperaría encontrar el óptimo de una función objetivo establecida, también los métodos aproximados podrían ser utilizados para la búsqueda de soluciones a través de metaheurísticas; sin embargo, existen otros tipos de soluciones como los modelos de sistemas basados en agentes que se han utilizado a lo largo de los últimos años para solucionar problemas en sistemas de transporte como la minimización de tiempos, asignación de flota entre otros.

La complejidad del problema constituye que las soluciones deben ser dinámicas de acuerdo con el comportamiento específico de cada una de las 49 estaciones analizadas, por lo cual, los sistemas basados en agentes permitirán la adecuación de escenarios y cambios en parámetros de manera frecuente de tal manera que las soluciones puedan cambiar y acoplarse a las necesidades reales del sistema en demanda, demoras y la minimización de los tiempos de espera.

De acuerdo con lo anterior, en este proyecto se utilizó la simulación basada en agentes como método de solución y llegar a cumplir el objetivo de negocio relacionado con el entendimiento del sistema. A continuación, se describe cada uno de los componentes y el desarrollo de la simulación para este problema:

AGENTES: Los agentes, más conocido como tortugas en SBA, son elementos que se pueden mover por el mapa o mundo de acuerdo con una seria de instrucciones, el mundo es un espacio bidimensional y se divide en parcelas, en el proyecto se manejan 3 tipos de agentes, cada uno recibe una “raza” para ser identificados dentro de la simulación, cada uno tiene ciertas características y son:

ESTACIONES	
VARIABLE	DESCRIPCIÓN
codigo	Código de la estación asignado por la entidad Transmilenio
nombre	Nombre de la estación asignado por la entidad Transmilenio
c_entradas	Vector con cantidad de entradas según la predicción
prom_entradas	Promedio de entradas por hora
ingresos	Cantidad de personas que ingresan a la estación y podrian dirigirse hacia el Portal el Dorado
c_rutas_totales	Cantidad de rutas que paran en la estación
c_rutas	Cantidad de rutas que paran en la estación y tienen como destino el Portal El Dorado
per_int	Cantidad de personas que se encuentran en la estación
t_entradas	Total de ingresos cada 15 min

Tabla 20 Características que tienen los agendes de raza estaciones

TRANSMILENIOS	
t_inicio	Tick en el que arranca la ruta
t_fin	Tick en el que llega a una estación
ruta	Vector con las estaciones a visitar según la ruta
dis_r	Distancia entre cada una de las estaciones de la ruta asignada
nom_r	Nombre de la ruta (1, K23, K16, K43, K86, K10, K54)
estaciones_visitadas	Cantidad de estaciones visitadas
capacidad	Capacidad de personas según el tipo de transmilenio
cant_paradas	Cantidad de estaciones que componen la ruta
pasajeros	Número de pasajeros abordo
distancia	Vector con las distancias entre cada estación

Tabla 21 Características que tienen los agendes de raza transmilenios

PERSONAS	
origen	Código de la estación de origen
destino	Código de la estación de destino
t_espera	Tiempo de espera desde que entro a la estación hasta que tomo un transmilenio
t_desplazamiento	Tiempo de desplazamiento entre la estación de origen y destino
abordo?	Bandera para indicar si ya abordo algún transmilenio
llego?	Bandera para indicar si llego a la estación destino
transmi	Ruta a la que pertenecia el transmilenio que abordo

Tabla 22 Características que tienen los agendes de raza personas

TICK: Un tick es la unidad de medida más pequeña de la simulación, en este caso, cada tick representa un minuto, en este minuto se generan ingresos de personas, abordajes, salidas de estaciones, inicio de nuevas rutas, etc, en el siguiente diagrama se muestra cómo funciona la simulación en cada tick.

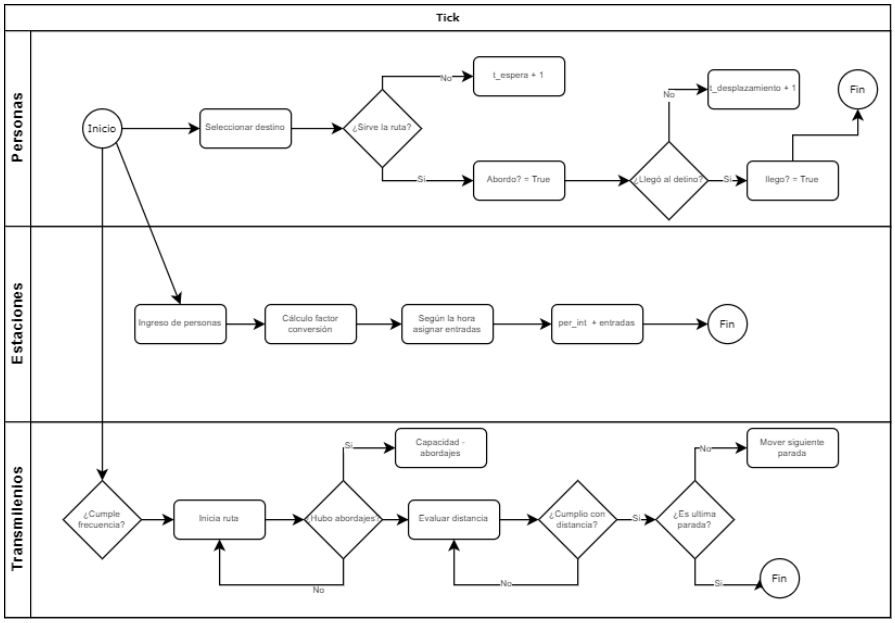

Figura 26 Diagrama de interacción entre los agentes en la simulación

Como se mencionó anteriormente, la medida más pequeña de la simulación es de 1 minuto, dado que se va a simular el mes de enero del 2018 en el intervalo horario de 04:00 a 23:45, el cual comprende 20 horas, se generan 2 ciclos adicionales al que cuenta los ticks, uno de 20 repeticiones representando las horas y otro de 31 repeticiones representando los días del mes.

FUNCIÓN CARGAR ENTORNO: Esta función es la encargada de parametrizar el mundo de la simulación, allí se carga el mapa de las diferentes rutas del sistema de transporte masivo de Bogotá Transmilenio, además, crea los primeros agentes de la raza ESTACIONES, solo se crean las estaciones de interés, es decir, por las que pasan las 7 rutas con destino al Portal el Dorado, por último, se crean las listas de distancias entre cada una de las estaciones.

Figura 27 Entorno de la simulación con las características de cada estación

PARAMETROS:

Los parámetros que caracterizan la simulación son 7:

1. Cantidad_ejecu: Corresponde a la cantidad de veces que se va a ejecutar la simulación completa, esto con el fin de estabilizar los resultados obtenidos por la misma, el valor de este parámetro es de 30.
2. Cantidad_dias: Corresponde a la cantidad de días que se van a simular, en este caso, como se quiere simular el mes de enero, serán 31 días.

3. P_estación: Corresponde a la probabilidad de que un pasajero tenga como destino el Portal El dorado, donde se genera un valor aleatorio llamado X, si X > P_estación, el destino será el Portal, de lo contrario puede elegir cualquier otra estación.
4. T_espera_max: Corresponde al tiempo máximo promedio de espera desde que el usuario ingresa al sistema hasta que aborda uno de los buses, este tiempo es de 15 minutos para todos los días excepto el sábado, el cual, tiene un tiempo máximo promedio de espera de 25 minutos.
5. T_espera_min: Corresponde al tiempo mínimo promedio de espera desde que el usuario ingresa al sistema hasta que aborda uno de los buses, estos 5 minutos aplican para cualquier día de la semana.

FUNCIÓN GO: Es la función encargada de generar las relaciones entre los diferentes agentes de acuerdo a los parámetros, variables y comportamientos emergentes que vayan surgiendo al pasar los ticks, en esta función se generan las siguientes acciones:

- Ajuste cada hora de la frecuencia con la que se despachan los buses para cada una de las rutas dependiendo del tiempo de espera de los usuarios, por defecto la frecuencia es de 20 minutos.
- Despacho de buses por rutas según la frecuencia, horarios y días de la semana incluyendo si es festivo o no.

RUTA	ORIGEN	DESTINO	INICIO-SEM	FIN-SEM	INICIO-SAB	FIN-SAB	INICIO-D_FES	FIN-D_FES
1	(06111) Universidades	(06000) Portal Eldorado	4:30	23:00	5:00	23:00	5:30	22:00
K43	(07503) SAN MATEO	(06000) Portal Eldorado	4:00	23:00	4:30	23:00	NA	NA
K16	(02502) Terminal	(06000) Portal Eldorado	5:30	22:30	5:30	22:00	NA	NA
K23	(02200) Alcalá	(06000) Portal Eldorado	5:00	22:00	NA	NA	NA	NA
K86	(10009) Museo Nacional	(06000) Portal Eldorado	5:30	23:00	6:00	23:00	7:00	22:00
K10	(10000) Portal 20 de Julio	(06000) Portal Eldorado	4:30	23:00	5:00	23:00	5:30	22:00
K54	(09000) Cabecera Usme	(06000) Portal Eldorado	5:30	22:30	NA	NA	NA	NA

Tabla 23 Horarios de servicio de las rutas de Transmilenio

- Ingresos a las estaciones de acuerdo con la información generada en los modelos de predicción.
- Recorrido de los buses por las estaciones según la ruta a la que pertenece, este recorrido está ligado a unos tiempos de desplazamiento, los cuales fueron definidos según la distancia en km entre estaciones y suponiendo una velocidad promedio de 30 Km/h teniendo en cuenta los tiempos de parada, semáforos, trancones, etc.
- Abordaje de los usuarios según el origen y destino, además, de la capacidad con la que cuenten los buses para recoger pasajeros, dicha capacidad varía dependiendo de los buses de cada ruta.
- Generación de archivos resultado de la simulación.

INTEGRACIÓN MODELO DE PREDICCIÓN Y SIMULACIÓN: Los resultados obtenidos en el modelo de predicción, tienen como estructura en sus filas los nombres de las estaciones y en sus columnas intervalos de una hora, desde las 04:00 hasta las 23:00 que como se mencionó anteriormente, son los horarios de funcionamiento de las rutas evaluadas en este proyecto.

La creación de los agendes “personas” se realiza con base en los valores obtenidos en el modelo, dado que el tick corresponde a un minuto y los resultados son cada hora, se crean las personas con la siguiente función:

```
ask estaciones with [nombre != "Portal Eldorado"][
  let factor_conversion prom_entradas * (c_rutas / c_rutas_totales)
  let ingresos_intervalo random-poisson factor_conversion / 60
  set ingresos precision ingresos_intervalo 0
  set codigo_estacion codigo
  set p_ingresos
  set posx xcor
  set posy ycor
  set per_int count(personas with [xcor = posx and ycor = posy])
  set t_entradas t_entradas + p
```

```
;Crear los agentes de la raza personas y asignarle su destino de forma aleatoria
create-personas p [
  set destinos []
  setxy posx posy
  set origen codigo_estacion
  let probabilidad random-float 1
  set destino 6000
  set t_espera 0
  set t_desplazamiento 0
  set color blue
  set shape "person"
  set size 15
  set abordo? False
  set llego? False
```

Figura 26 Código creación agentes “personas” con base a resultados de predicción por intervalo de hora

Se evalúa la proporción de rutas analizadas vs el total de rutas que tienen parada en dicha estación, además se genera un valor aleatorio con una función de probabilidad poisson dado que corresponde a arribos y

posteriormente se divide en 60, esto con el fin de ajustar el resultado de la predicción que es por hora, al valor mas pequeño en la simulación, es decir, un tick que equivale a 1 minuto.

Se ha realizado el estudio comparativo de 4 escenarios:

1 escenario: Funcionamiento Transmilenio al 2017:

Estadísticas de transporte público en Bogotá publicado por moovit insights dicen que la media del tiempo de permanencia en el sistema de transporte público es de 83 minutos = 64 de trayecto + 19 de espera lo que representa más de una hora y 20 minutos, esa cifra es la media, sin embargo, estas estadísticas también nos dicen que el 41 % de las personas en Bogotá que viajan en transporte publico pueden durar 2 horas viajando cuando se incluye alimentadores, Transmilenio o SITP.

Se ha confirmado que con estos tiempos de espera y de permanencia en el sistema más de la mitad de los usuarios consideran que el sistema no presta un correcto servicio, incluso las estaciones colapsan ya que se acumulan los usuarios sobre todo en las horas pico.

2. escenario: Con el modelo de agentes se probaron tiempos bajos de espera (aproximadamente 5 minutos), con este modelo se tuvieron rutas que en algunos casos requerían 202 viajes, el tiempo promedio de un viaje completo (portal a portal) es de 45 minutos, para cubrir esos 202 viajes se requeriría una flota de 27 buses que operarían cerca de 6 horas sin parar a lo largo del día. Sin embargo, lo anterior en la realidad es sustentable ya que los buses no pueden operar de forma continua las 6 horas, ya que requieren revisión, tanqueo y cambios de turno de sus conductores además muchos buses deben pasar a mantenimientos largos. Transmilenio realiza un mantenimiento correctivo que implica que los buses estén fuera de servicio entre 2 y 10 días, dicho mantenimiento abarca la corrección de todas las fallas en vía que tienen los buses a diario, así como las fallas técnicas y afectaciones por las cuales se requiere un cambio de bus, se ha podido evidenciar que cerca de la mitad de los buses al mes requieren mantenimiento (Caracol radio, 2018), esto nos dice que se puede llegar a necesitar hasta 40 buses para cubrir la demanda completa.

Sobre los buses de Transmilenio se tienen 3 opciones con los siguientes precios (valores aproximados que cambian año a año de acuerdo con el comportamiento del mercado) (Autos de primera, s. f):

- Diesel: $ 1.034.000.000
- Gas: 1.103.000.000
- Eléctrico: $ 1.842.000.0000

Calculamos cuanto valdría la flota completa y obtenemos los siguientes montos:

- Diesel: $ 41.360.000.000
- Gas: $ 4.4120.000.000
- Eléctrico: $ 73.680.000.000

El valor total de un bus se calcula al sumar el costo del vehículo más su mantenimiento durante 12 años que es la vida útil que se les asigna a estos vehículos, es valor calculado termina siendo muy similar para los buses: un articulado a diésel valdrá $4.608 millones y uno eléctrico cerca de $4.500 millones (Garcia, 2018).

Calculando el valor total para la flota $180.000.000.000.

Este costo es bastante alto, aunque no aplica para todas las rutas el hecho que aplique para 3 ya hace que sea un modelo de alto costos, pues tendremos que triplicar el valor calculo y además sumar las flotas de las otras rutas.

3. escenario: Con el modelo de agentes se probaron tiempos de espera menores a los actuales, pero no tan bajos como los del escenario 2, para este caso separamos los tiempos del día de lunes a viernes de los fines

de semana, los tiempos de espera fueron de máximo 15 minutos de lunes a viernes y domingos, y de 25 minutos los sábados bajo la presunción de que el sábado es un día donde los usuarios pueden tener menos afán.

Para este modelo encontramos que las necesidades de buses varían según el día; el promedio diario del tiempo de espera arrojado por el modelo es de 8 minutos de lunes a viernes, y de 20 minutos en promedio para los sábados, domingos y festivos; esto nos dice que se requiere una flota de 20 buses para cubrir cada ruta, si incluimos la carga por mantenimiento tenemos una flota aproximada de 25 a 30 buses dependiendo de que tan nueva es la flota, al realizar el análisis de costos final tenemos que el costo de la flota será de:

$4.500.000.000 * 30 = $135.000.000.000

7. DETERMINACIÓN DE PRÓXIMOS PASOS Y CONCLUSIONES

- Es importante que la información inicial esté recolectada de forma más limpia, gran parte del proyecto tuvo una complejidad alta en poder organizar la información, es importante estandarizar los formatos de los datos al recolectarlos. El tratamiento de datos se hizo mucho más complejo al tener tablas que no tenían la misma lógica en estructura, y al ser copias de tablas dinámicas no permitía de manera sencilla entender a primera vista lo que se estaba buscando.

- El periodo de tiempo para poder generar pronósticos en series de tiempo debería ser de al menos 5 años para poder absorber el comportamiento mensual y de horas que tiene el sistema de manera general, así poder pronosticar periodos de tiempo con mayor exactitud.

- Para futuros estudios será interesante explorar datos transaccionales de los usuarios buscando entender el porqué de los trayectos que toman, por ejemplo, un usuario que viaje bastante por trabajo con regularidad nos mostrará un patrón constante para dirigirse hacia el portal del dorado, lo que nos permitirá proponer incluso ruta para los usuarios que transiten las mismas estaciones.
    - Cuando comparamos los 3 escenarios encontramos lo siguiente:
        - El Primer escenario refleja la operación real del Transmilenio, y tal como lo vemos hoy es un modelo que tiene bastantes problemas en la operativa y percepción frente al público, con tiempos de permanecía promedio de 83 minutos.
        - El segundo escenario, representa una buena disminución de permanecía en el sistema, en tiempos de espera por ejemplo se reducen un 70% pasando de 19 minutos en promedio a solo 5, sin embargo, es muy alto en costo de adquisición de la flota, solo para cubrir 3 rutas se requieren $540.000.000.000 COP, es decir, $180.000.000.000 COP por ruta.
        - El mejor escenario es el número 3, en este escenario tenemos una reducción de tiempos de permanencia donde el tiempo de espera se encuentra entre los 5 y 15 minutos y el tiempo del trayecto es en promedio de 45 minutos, teniendo un tiempo total de permanencia máximo de 60 minutos, y en promedio de 53 minutos, disminuyendo el tiempo total de permanencia mínimo en un 30%, además este modelo representa un equilibrio vs los costos, si se comparan las flotas del escenario 2 encontramos que el escenario 3 ahora un 25% de costos para cubrir cada ruta.

    - Con base en las simulaciones realizadas en los escenarios anteriores, encontramos que el mejor modelo es el escenario número 3. Por ende, recomendamos a Transmilenio que las rutas que se dirigen hacia el portal el dorado tengan un mínimo de 30 buses de modo que se garantice la operación de mínimo 20 buses de lunes a viernes con los que se pueda enviar viajes que mantengan un tiempo de espera promedio de 8 minutos. Los 10 restantes buses dan holgura al sistema para

que se puedan realizar los mantenimientos sin afectar el servicio prestado, y además en las horas pico se puede incrementar el número de viajes para disminuir el tiempo de espera a 4 o 5 minutos garantizando la disminución del tiempo de permanencia en el sistema. Cada flota por ruta requerirá una inversión de cerca de $135.000.000.000 COP, que si se compara con el escenario 2 representa un ahorra del 25% por ruta, es decir, $315.000.000.000 COP

- Recomendamos también basados en el escenario 3 programar las jornadas de mantenimiento para aprovechar los fines de semana y festivos, para estos días como el tiempo de espera aceptable es mayor (20 minutos) es posible tener un número de 12 buses operando en todo el día cubriendo los viajes requeridos, de igual manera para estos días la flota tendrá buses disponibles para aumentar la operación por si se presentan eventos especiales que aumenten el número de usuarios en el sistema.

- Las precipitaciones no corresponden a un valor relevante dentro de los modelos de predicción, aunque en algunos casos se daban valores más pequeños de error, no eran significantes para poder decir que el comportamiento pudiese influir, además la estabilidad de los datos al ser explorados mostró que la lluvia no hace cambiar las horas pico ni el número de usuarios que ingresa y sale. Para posteriores análisis será importante revisar si existe algún otro factor que pueda influir dentro del sistema.

- Los modelos de pronóstico para predecir las entradas tuvieron como resultados valores de indicadores estadísticos buenos, lo que indica que son modelos aceptables para la predicción, ejemplo de ellos es que para todos los indicadores el valor del indicador MASE es menor a 1 confirmando que los modelos predicen mejor que cualquier modelo ingenuo. Adicional a esto, los indicadores de RMSE y MAE fueron bajos.

8. MANEJO RESPONSABLE DE LA INFORMACIÓN

Para este proyecto no se suministraron datos de carácter personal, sin embargo, dentro de las recomendaciones dadas para futuros proyectos un elemento que puede aportar estimaciones más acertadas son datos de los usuarios que permita entender por qué usan los trayectos como los usan, por ejemplo, dirección de la casa, dirección de trabajo, DNI para identificarlo, etc. Con eso en mente damos las siguientes recomendaciones para el tratamiento de datos personales basados en le ley 1581 de 2012, de modo que se garantice en futuros estudios el respeto constitucional de los usuarios respecto a la protección y tratamiento de datos personales.

Para establecer las políticas es importante tener en cuenta los siguientes conceptos:

- Consentimiento previo, expreso e informado del Titular para llevar a cabo el Tratamiento de datos personales.
- Dato personal: Cualquier información vinculada o que pueda asociarse a una o varias personas naturales determinadas.
- Titular: Persona natural cuyos datos personales sean objeto de Tratamiento
- Tratamiento: Cualquier operación o conjunto de operaciones sobre datos personales, tales como la recolección, almacenamiento, uso, circulación o supresión.

En el marco de la Ley 1581 de 2012 (Función pública, 2017) es necesario contar con una serie de políticas de tratamientos de datos para garantizar el derecho a la privacidad de cada usuario. Es importante cumplir con los siguientes puntos:

- La información usada solo será aplicada en la finalidad del proyecto, y no deberá ser utilizada para otras actividades.
- La información no será divulgada a terceros sin la autorización del titular.

- Se debe garantizar que no se pierda la información, tomando medidas especiales si el proyecto se desarrolla en equipos de cómputo personales que pueden estar expuestos fácilmente a ataques externos.
- Para el tratamiento de los datos personales todas las actividades deben ser enmarcadas en los principios de legalidad, finalidad, libertad, veracidad o calidad, transparencia, acceso y circulación restringida, seguridad y confidencialidad.
- Se debe solicitar una autorización, de forma previa e informada, al titular de los datos para poder recopilarlos y tratarlos, esta autorización debe ser explícita y expresada de cualquier manera que se evidencie una aceptación del tratamiento de los datos por parte del titular.
- Se debe informar debidamente al Titular sobre la finalidad de la recolección y los derechos que le asisten por virtud de la autorización otorgada.
- La información debe ser conservada bajo las condiciones de seguridad necesarias para impedir su adulteración, pérdida, consulta, uso o acceso no autorizado o fraudulento.

En caso de realizar proyectos con data sensible se recomienda ofuscar los datos buscando asegurar la privacidad y confidencialidad de estos sin perder la calidad de los datos.

Para el presente trabajo de grado no se requirió firmar acuerdos de confidencialidad con la compañía de Transmilenio, sin embargo, como estudiantes de la de la Maestría en Analítica para la Inteligencia de Negocios y en cumplimiento con políticas de tratamiento de datos se garantizó que la data no se compartiera con terceros y que solo se usará para el objetivo del proyecto, la recepción de la data se realiza específicamente con la universidad y a ellos se les entregaran los resultados del proyecto garantizando que no existan flujos de datos fuera de los canales establecidos ya que podría generarse riegos de perdida y robo de la misma.

9. ANEXOS

- RESULTADOS SIMULACION VIAJES.xlsx: Allí se encuentran los resultados de la cantidad de viajes por ruta y día de las 30 ejecuciones de la simulación.
- RESULTADOS SIMULACION TIEMPOS DE ESPERA.xlsx: Allí se encuentran los resultados de los tiempos de espera promedio por día de las 30 ejecuciones de la simulación.
- INSTRUCTIVO SIMULACION.pdf: Se encuentra el paso a paso y los requerimientos para ejecutar la simulación.

10. REFERENCIAS

- Aeropuertos del mundo. (s.f.) Aeropuerto El Dorado de Bogotá (BOG). Recuperado de: https://aeropuertosdelmundo.com.ar/aeropuerto-BOG/
- Autos de primera. (s. F). Diésel, eléctricos o a gas, las tres opciones para los nuevos buses de Transmilenio en Bogotá. Recuperado de: https://autosdeprimera.com/transmilenio-opciones-tecnologias-nuevos-buses/
- Bogotá Cómo Vamos. (2017). encuesta de percepción ciudadana 2017. Recuperado de: https://s3.documentcloud.org/documents/4222235/Encuesta-de-Percepci%C3%B3n-Ciudadana-2017.pdf
- Bogotá Cómo Vamos. (2018). encuesta de percepción ciudadana 2018. Recuperado de: https://s3.documentcloud.org/documents/6402320/EPC-2018-FINAL-ISSN.pdf
- Bogotá Cómo Vamos. (2019). encuesta de percepción ciudadana 2019. Recuperado de:
- https://s3.documentcloud.org/documents/6551608/Encuesta-de-Percepci%C3%B3n-Ciudadana-2019.pdf
- Caracol radio. (2018). Al mes se varan, en promedio, 872 buses de Transmilenio. Recuperado de: https://caracol.com.co/emisora/2018/09/20/bogota/1537461547_480267.html
- Cardoso, C., F., Bert, & G., Podestá. (2014). Modelos basados en agentes (MBA): definición, alcances y limitaciones http://www.iai.int/wp-content/uploads/2014/03/Cardoso_et_al_Manual_ABM.pdf.

- Chávez Quisbert, Nicolás. (1997). MODELOS ARIMA. Revista Ciencia y Cultura, (1), 23-30. Recuperado en 27 de mayo de 2022, de http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2077-33231997000100005&lng=es&tlng=es.
- ElKady, S. y Abdelsalam,H. (2015). A simulation-based optimization approach for healthcare facility location allocation decision. Science and Information Conference (SAI), London, p.500-505. Recuperado de https://ieeexplore-ieee-org.ezproxy.javeriana.edu.co/document/7237189
- Función pública. (2017). ley 1581 de 2012. Recuperado de: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981
- García-Valdecasas, J. I. (2011). La simulación basada en agentes: una nueva forma de explorar fenómenos sociales. Reis 136:91-100 doi:10.5477/cis/reis.136.91.
- Garcia, Felipe. (2018). Nueva flota de Transmilenio: ¿ajustar el presupuesto o apostarle a un aire de calidad? El espectador. Recuperado de: https://www.elespectador.com/bogota/nueva-flota-de-transmilenio-ajustar-el-presupuesto-o-apostarle-a-un-aire-de-calidad-article-750163/
- Gupta, D. (2018). Applied Analytics through Case Studies Using SAS and R.
- Heppenstall, A. J. J., Crooks, A. T. T., See, L. M. M., Batty, M. (Eds.). (2012). Agent based models of geographic systems. Springer Netherlands.
- Hernandez, Santiago. (s. f.). SERIES TEMPORALES, MODELO ARIMA METODOLOGÍA DE BOX – JENKINS. Recuperado de: https://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf
- Moovit insights. (s. f). Estadísticas de transporte público en Bogota Recuperado de: https://moovitapp.com/insights/es/Moovit_Insights_%C3%8Dndice_de_Transporte_P%C3%BAblico_Colombia_Bogota-762#:~:text=Los%20usuarios%20del%20transporte%20p%C3%BAblico,espera%20m%C3%A1s%20de%2020%20minutos.
- NetLogo (2018). ¿Qué es Netlogo? Recuperado de: https://ccl.northwestern.edu/netlogo/resources/Que%20es%20NetLogo.pdf
- Oracle. (n.d.). Previsión y descripciones estadísticas de Planificación predictiva. In Trabajo con Planning. Oracle. https://docs.oracle.com/cloud/help/es/pbcs_common/PFUSU/insights_metrics_RMSE.htm#PFUSU-GUID-FD9381A1-81E1-4F6D-8EC4-82A6CE2A6E74
- Portafolio. (17 de agosto de 2021). Aeropuerto El Dorado, el de mejor personal en Suramérica. Recuperado de: https://www.portafolio.co/negocios/empresas/aeropuerto-el-dorado-reconocido-por-tener-el-mejor-personal-en-suramerica-555198
- Python. (2022). About Python. Recuperado de: https://www.python.org/about/
- Railsback, S. F., & V., Grimm. (2012). Agent-based and individual-based modeling, a practical introduction. Princeton University Press, New Jersey
- SITP. (2022). Ruta 1 Transmilenio. Recuperado de: https://sitp-bogota.com/ruta-1-transmilenio/
- SITP. (2022). Ruta G43Transmilenio. Recuperado de: https://sitp-bogota.com/ruta-g43-transmilenio/
- SITP. (2022). Ruta B16 Transmilenio. Recuperado de: https://sitp-bogota.com/ruta-b16-transmilenio/
- SITP. (2022). Ruta B23 Transmilenio. Recuperado de: https://sitp-bogota.com/ruta-b23-transmilenio/
- SITP. (2022). Ruta K16 Transmilenio. Recuperado de: https://sitp-bogota.com/ruta-k86-transmilenio/
- SITP. (2022). Ruta L10 Transmilenio. Recuperado de: https://sitp-bogota.com/ruta-l10-transmilenio/
- SITP. (2022). Ruta H54 Transmilenio. Recuperado de: https://sitp-bogota.com/ruta-h54-transmilenio/
- Transmilenio, S. A. (2018). Informe de Gestión 2017 Transmilenio. https://www.transmilenio.gov.co/publicaciones/149939/publicacionesinforme_de_gestion_2017_transmilenio_sa/
- TRANSMILENIO S.A. (2022). Historia de TransMilenio. Recuperado de: https://www.transmilenio.gov.co/publicaciones/146028/historia-de-transmilenio/
- TRANSMILENIO S.A. (2022). Mapa Interactivo de TransMilenio. Recuperado de: https://www.transmilenio.gov.co/publicaciones/150402/publicacionesmapa-interactivo-de-transmilenio/

- Unidad Administrativa Especial de Aeronáutica Civil - UAEAC. (07 de febrero de 2021). Colombia superó la meta de 30 millones de pasajeros movilizados y 835 mil toneladas de carga transportada en 2021. Recuperado de: https://www.aerocivil.gov.co/prensa/noticias/Pages/Colombia-supero-la-meta-de-30-millones-de-pasajeros-movilizados-y-835-mil-toneladas-de-carga-transportada-en-2021.aspx
- Universidad Externado de Colombia. (s. f). Cronología del caso Avianca – ACDAC. Recuperado de: https://www.uexternado.edu.co/derecho/cronologia-del-caso-avianca-acdac/
- Universidad Internacional de La Rioja. (2019). Lenguaje R, ¿qué es y por qué es tan usado en Big Data? UNIR. https://www.unir.net/ingenieria/revista/lenguaje-r-big-data/#:~:text=R%20es%20un%20entorno%20de,a%20instrucciones%20en%20lenguaje%20m%C3%A1quina.
- Valora Analitik. (2022). Aeropuerto El Dorado (Bogotá) podría presentar “demoras significativas” este viernes. Recuperado de: https://www.valoraanalitik.com/2022/04/08/como-esta-el-aeropuerto-dorado-bogota-viernes-8-abril/
- Weather Spark. (2022). El clima y el tiempo promedio en todo el año en Bogotá: Precipitación – lluvia. Recuperado de: https://es.weatherspark.com/y/23324/Clima-promedio-en-Bogot%C3%A1-Colombia-durante-todo-el-a%C3%B1o