

UNIVERSIDAD NACIONAL AUTÓNOMA DE  
MÉXICO

---

---

FACULTAD DE ESTUDIOS SUPERIORES  
ACATLÁN

**Diplomado de Ciencia de Datos**

**Proyecto Ecobicis**

Bonilla Cruz José Armando

# Índice general

<b>1. Introducción y Objetivo</b>	<b>1</b>
1.1. Introducción . . . . .	1
1.2. Objetivo . . . . .	1
<b>2. Estructura de los datos</b>	<b>4</b>
2.1. Diccionario de datos . . . . .	5
<b>3. Lectura de datos</b>	<b>7</b>
<b>4. Tratamiento de los Datos</b>	<b>9</b>
4.1. Cruce de viajes con estaciones . . . . .	9
4.2. Missings . . . . .	9
<b>5. Visualizaciones</b>	<b>11</b>
<b>6. Generación de la tabla Viajes por día</b>	<b>17</b>
<b>7. Perdición de la demanda diaria</b>	<b>19</b>
7.1. K vecinos más cercanos . . . . .	19
7.2. Red elástica . . . . .	20
7.3. Red Neuronal Recurrente con GRU . . . . .	21
<b>8. Distribución de las estaciones con modelos no supervisados</b>	<b>25</b>
8.1. K means, Agrupación natural de las estaciones . . . . .	26
8.2. DBSCAN para estaciones de difícil acceso . . . . .	28
8.3. K means, agrupación por número de viajes en el día . . . . .	30

*ÍNDICE GENERAL* III

**9. Conclusiones** 35

# **Capítulo 1**

## **Introducción y Objetivo**

### **1.1 Introducción**

Ecobici es el sistema de bicicletas compartidas de la ciudad de México, auspiciado por el gobierno estatal de la CDMX.

El sistema Ecobici tiene poco más de una década que comenzó, inició en el año 2010 con cerca 84 cicloestaciones y 1,200 bicicletas, al día de hoy, según la página oficial, se informa que ahora se cuenta con 687 estaciones y cerca de nueve mil bicicletas para su disposición. Trece años y una pandemia después, gracias a la gran demanda del servicio así como de la infraestructura; en agosto del 2022 el gobierno de la Ciudad de México decidió hacer un cambio en el sistema ecobici, con la renovación de infraestructura, gestión y servicio.

### **1.2 Objetivo**

El objetivo de este trabajo es vislumbrar el estatus actual del nuevo sistema y su constante evolución en su corto periodo de vida. Este nuevo sistema ha traído demasiados cambios, ya que no solo se renovaron las bicis, sino las estaciones de carga, algunas de estas estaciones se dejaron en desuso y otras muchas están en transición de ser renovadas para ser compatibles con el nuevo proyecto.

## CAPÍTULO 1. INTRODUCCIÓN Y OBJETIVO

---

Dada la problemática de la transición y el gradual aumento de la demanda, en dicho trabajo se presenta un análisis del sistema actual con el fin de tomar mejores decisiones sobre el rumbo y la gestión del mismo, para esto se va a abordar los siguientes puntos que creemos son importantes.

- Predicción de la demanda por día
- Prevalencia del género masculino y tendencia del femenino
- Identificación de estaciones de difícil acceso
- Segmentación de las estaciones por sectores
- Segmentación e identificación de estaciones críticas

Estas preguntas son de vital importancia, pues con estas podemos ver que tan exitoso es el sistema y con la predicción la gestión se facilitaría, ya que podríamos censar si la infraestructura y los recursos actuales son suficientes.

Por otro lado, la Ciudad de México en sus campañas más recientes ha implementado muchos programas enfocados en la mujer y queremos ver cómo es que la proporción de usuarias ha evolucionado, y si es que las mujeres se sienten cómodas usando este nuevo sistema con respecto al anterior. En este punto sólo se busca un bosquejo general respecto a la problemática con tal de hacerle frente con alguna campaña que promueva el sistema.

En lo que respecta a los puntos de las estaciones, es más que claro que conocer bien como se van desarrollando el sistema en estos puntos fijos es de mucha ayuda para la administración del mismo, ya que así se pueden enfocar esfuerzos en puntos críticos, atender problemas de conectividad, incluso este conocimiento nos podría ayudar a planificar la extensión a las demás zonas de la ciudad. Este último punto es de suma importancia, ya que según la última encuesta del 2020 Ecobicis la principal problemática del sistema es que no se ha extendido a otras zonas .

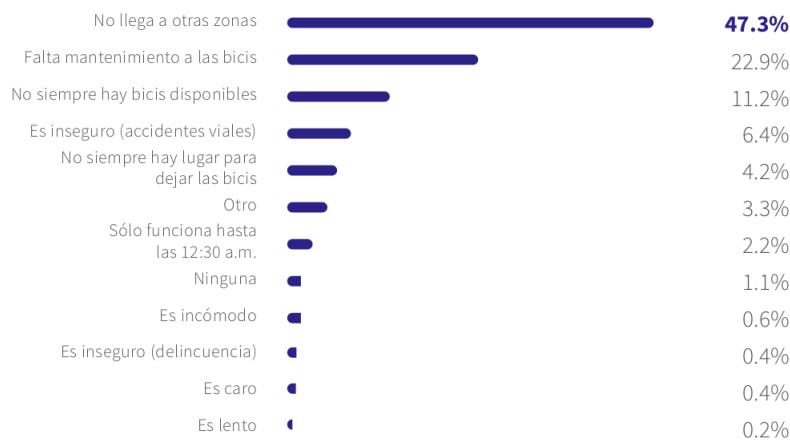
Con lo ya mencionado es claro pensar que lo que busca este documento es llegar a manos de los encargados del sistema y por lo tanto estos serían nuestro

## CAPÍTULO 1. INTRODUCCIÓN Y OBJETIVO

---

usuario final.

- »» 18 En tu opinión, ¿cuál es la principal desventaja que tiene ECOBICI?



(a) Desventajas del sistema.

# **Capítulo 2**

## **Estructura de los datos**

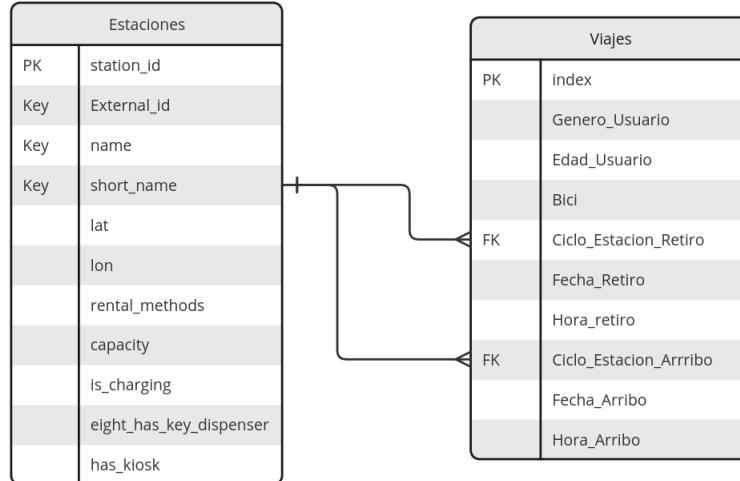
En el presente proyecto se busca trabajar con los datos públicos del sistema ecobici. Los datos pueden ser obtenidos de la siguiente liga. De esta liga podemos obtener la información de ambas tablas, la información de las estaciones puede ser consultada sin necesidad de descargar la información, por medio de una API se puede acceder y consultar la información perfectamente.

[https://ecobici.cdmx.gob.mx/datos-abiertos/.](https://ecobici.cdmx.gob.mx/datos-abiertos/)

Como se mencionó, la información con la que se quiere trabajar es la referente al nuevo sistema que comenzó en agosto del año 2022, a pesar de que en el portal tenemos información desde el 2010, no se tomará información previa, pues no está dentro de los alcances del proyecto.

A continuación un pequeño diagrama que ilustra la relación entre las diferentes tablas.

# Sistema Ecobici



(a) Diagrama entidad relación, sistema ecobici

Nota 1: A pesar de que las tablas no son una base de datos en estricto sentido, porque no cuentan con estructura, llaves primarias, secundarias, tablas entidad-relación ni otros elementos importantes para considerar a los datos como una base de datos, la información se presenta así para poder tener una mayor compresión de como es que se relacionan las tablas y algunas atributos de los que se dispone.

## 2.1 Diccionario de datos

### Viajes

Esta es nuestra tabla principal en la que se registran los datos de cada viaje de la manera más atómica posible, y se cuenta con los siguientes atributos como lo son, índice, datos del usuario y del viaje. Dentro de los que se encuentra el género y la edad del usuario, la bici que se utilizó, de qué estación se retiró, la fecha y hora del retiro, así como la estación de arribo.

## CAPÍTULO 2. ESTRUCTURA DE LOS DATOS

---

### Estaciones

En esta tabla podremos encontrar toda la información relevante necesaria para caracterizar las estaciones de ecobicis, desde la ubicación, hasta la capacidad máxima de bicis que pueden alojar. En un principio esta tabla nos va a ayudar a identificar las estaciones que están adaptadas al nuevo sistema, porque no todas las 687 estaciones están adaptadas para acoger a las bicis 2.0. Durante la vida del proyecto hemos visto como las estaciones poco a poco se han ido adaptando pues ya que al 22 de noviembre contábamos con 571 estaciones, con la actualización del 25 de diciembre teníamos 619, con la ultima actualización de enero 19 contamos con un total de 651 estaciones.



(b) Ecobici 1.0 y 2.0.

# Capítulo 3

## Lectura de datos

En las siguientes tablas se muestra una pequeña muestra de los datos que contienen las dos tablas previamente mencionadas.

### Viajes

La tabla viajes es la que necesitó más trabajo, limpieza y paciencia. A lo largo del trabajo se tuvo que ir y venir para poder modificar datos, de tal manera que toda la información fuera consistente, este tema se va a tocar con mayor profundidad en el siguiente capítulo. Aquí una extracción de los datos.

Genero_Usuario	Edad_Usuario	Bici	Ciclo_Estacion_Retiro	Fecha_Retiro	Hora_Retiro	Ciclo_Estacion_Arribo	Fecha_Arribo	Hora_Arribo	
0	O	44	7344476	284	2022-08-27	15:35:47.5490	1000	2022-10-01	14:41:05.3850
1	M	35	8908451	081	2022-09-03	18:45:42.3370	1000	2023-02-02	19:09:36.7840
2	M	27	8335416	067	2022-09-16	20:05:41.6240	1000	2022-10-05	14:37:49.5730
3	M	30	2668598	360	2022-09-19	16:26:55.0630	1000	2022-10-01	14:40:56.8900
4	M	34	5679634	083	2022-09-20	14:25:23.2570	1000	2022-10-01	14:39:25.6220
...	...	...	...	...	...	...	...	...	
5924205	M	33	6523906	095	2023-06-30	11:51:20.5730	172	2023-06-30	13:26:38.1710
5924206	F	42	8360976	178	2023-06-30	12:18:51.7830	064	2023-06-30	13:26:40.6750
5924207	M	46	2508539	113	2023-06-30	12:19:54.9460	096	2023-06-30	13:26:43.1730
5924208	M	26	5842372	369	2023-06-30	12:20:04.0770	386	2023-06-30	13:26:17.4740
5924209	M	52	2702280	114	2023-06-30	10:31:05.7880	271-272	2023-06-30	11:41:50.1060

(a) DataFrame Viajes.

Al momento de la última actualización del proyecto que considera los viajes hasta diciembre 2023, se cuenta con información de trece millones de viajes.

## CAPÍTULO 3. LECTURA DE DATOS

---

### Estaciones

El data frame Estaciones es una tabla llena de valiosa información, sobre la capacidad, ubicación y características de las estaciones en el nuevo sistema. Esta información se actualiza automáticamente del repositorio y desde el inicio del proyecto el número de estaciones ha aumentado, contando ahora con 651 estaciones.

	station_id	external_id	name	short_name	lat	lon	rental_methods	capacity	electric_bike_surcharge_waiver	is_charging	eightd	has_key_dispenser	has_kiosk
0	5	3ea89109-d2f3-46eb-9c41-c43742050340	CE-407 Prolongación Xochicalco-General Emilia...		407	19.3672	-99.1587	[KEY, CREDITCARD]	19	False	False	False	True
1	6	ba78b703-4e5a-44bd-ab2c-1eedc71e11c3	CE-428 Prolongación Uxmal-Av. Popocatépetl (E...		428	19.3634	-99.1604	[KEY, CREDITCARD]	27	False	False	False	True
2	7	6563d263-2342-46e3-98b3-461e68d2d615	CE-427 Avenida México-Coyoacán-Av. Popocatépet...		427	19.3649	-99.1630	[KEY, CREDITCARD]	19	False	False	False	True
3	8	ec55e597-c8fc-4e86-bcfe-b0e81a494790	CE-443 Bruno Traven-Golondrinas		443	19.3596	-99.1621	[KEY, CREDITCARD]	31	False	False	False	True
4	9	a98c7fac-12ce-4895-923b-6039f0421ca	CE-423 Moras-José María Rico (Eje 8)		423	19.3678	-99.1753	[KEY, CREDITCARD]	23	False	False	False	True

(b) DataFrame Estaciones.

# **Capítulo 4**

## **Tratamiento de los Datos**

### **4.1 Cruce de viajes con estaciones**

A pesar de que el sistema ecobicis 2.0 comenzó oficialmente en agosto de 2022, hasta principios del año 2023 todavía coexistían ambos servicios al mismo tiempo, así que cruzamos la información de los viajes con las estaciones para poder tener información del sistema 2.0 y se llegó a que nuestros datos se reducían a 13.23 millones de datos.

### **4.2 Missings**

A partir de lo mencionado en la sección anterior obtuvimos la siguiente matriz de nulos, dónde se puede observar los datos que se conservaron a partir de los cruces.

## CAPÍTULO 4. TRATAMIENTO DE LOS DATOS

---

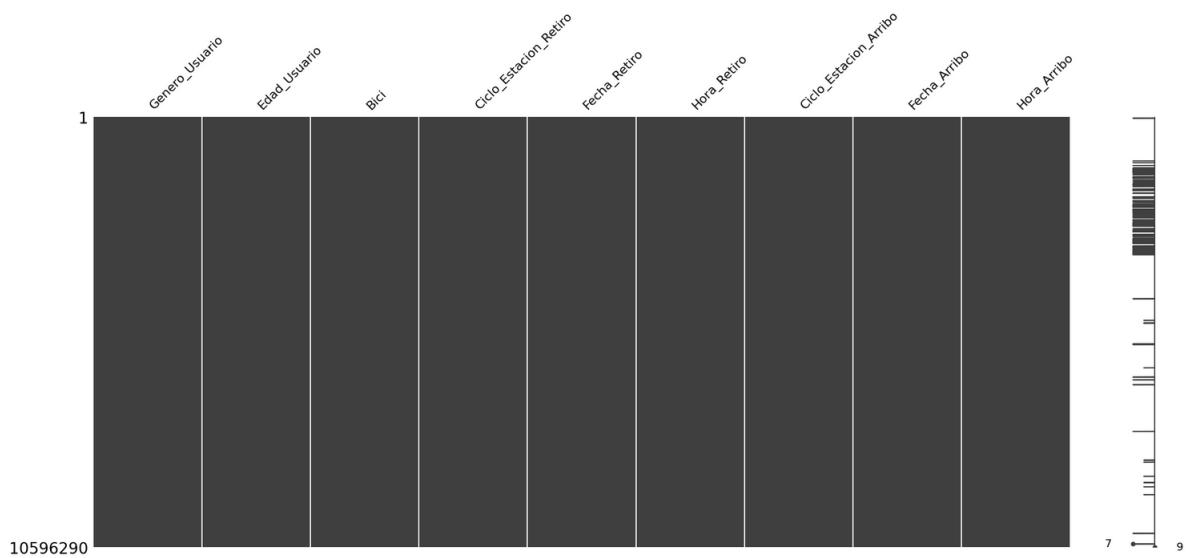


Figura 4.1: Missing matrix

Además de verificar que la tabla viene muy limpia, sólo se tiene a lo más dos datos ausente por registro y casi siempre vienen en pares esas dos variables las cuales son la edad del usuario y el género. Los registros los cuales tenemos nulos son 147, por lo que se procedió a eliminarlos.

# Capítulo 5

## Visualizaciones



Figura 5.1: Distribución de las estaciones

La visualización de nuestros datos es una de los puntos clave del proyecto, en esta primer imagen presentada puede corroborarse que las estaciones ecobi-

## CAPÍTULO 5. VISUALIZACIONES

---

cis están concentradas en la parte poniente de la CDMX. Además, en esta misma gráfica ploteamos un círculo de cien metros a la redonda para localizar los sitios de mayor acceso. Este mapa es accesible y se puede navegar sobre él para visualizar la distribución con más detalle. El archivo para su manipulación se puede encontrar en el repositorio como *estaciones.html*

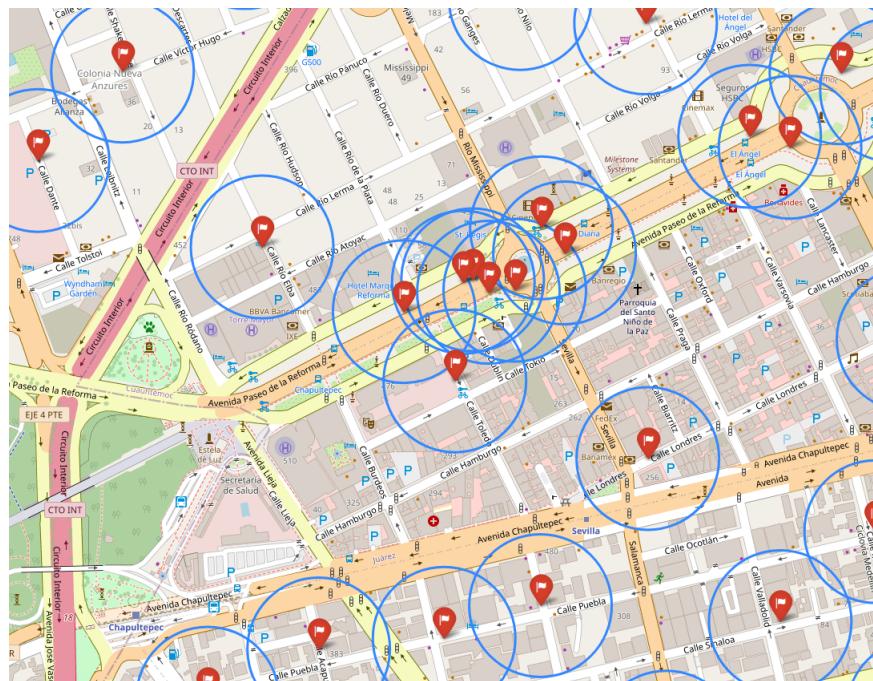


Figura 5.2: Distribución de las estaciones detalle

A continuación, la distribución de usuarios respecto al género.

## CAPÍTULO 5. VISUALIZACIONES

---

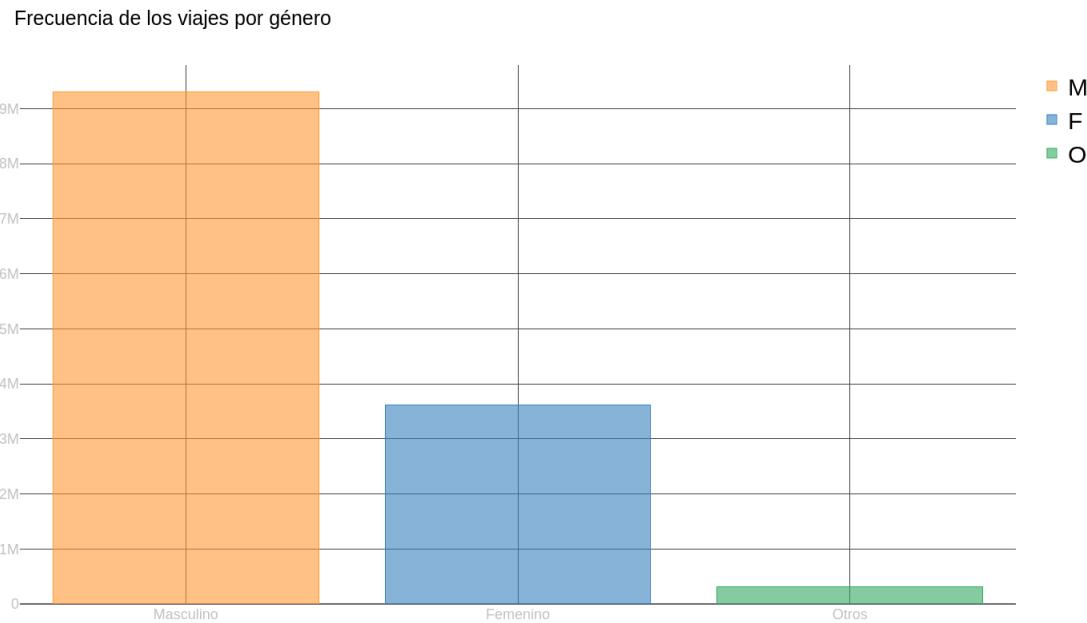


Figura 5.3: Frecuencia de los viajes por género

El género que más usa el sistema, es el Masculino, de hecho estos representan el 70% de los datos, mientras que el Femenino se lleva sólo el 27%, además, el sistema tienen la facilidad que el usuario proporcione o no su género, con lo que se forma una pequeña categoría de Otros que se lleva el resto. Cabe mencionar que la proporción de usuario y usuarias en la encuesta del 2020, la proporción de hombres era más baja, mientras que la proporción de mujeres era más alta con 33.6%.

## CAPÍTULO 5. VISUALIZACIONES

---

### »» 01 ¿Con qué género te identificas?



Figura 5.4: Frecuencia de los viajes por género, encuesta 2020

A continuación, unas gráficas que nos ayudan a observar la demanda del sistema a través del tiempo.

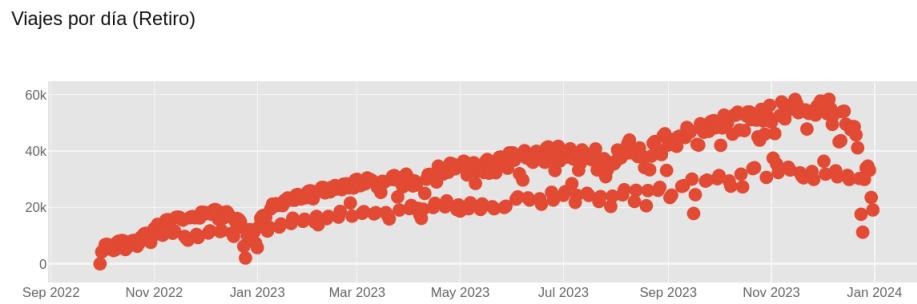


Figura 5.5: Viajes por día de Fecha Retiro

## CAPÍTULO 5. VISUALIZACIONES

---

Viajes por día (Arribo)

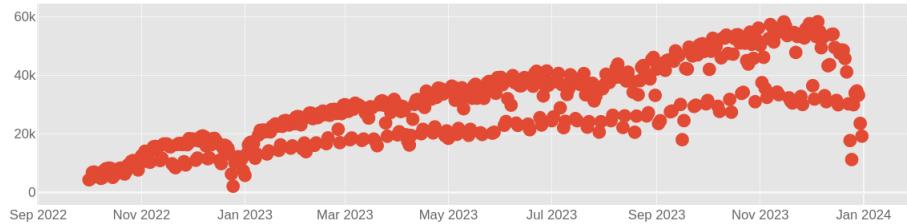


Figura 5.6: Viajes por día de Fecha Arribo

Estas dos gráficas son dos de las más importantes pues podemos inferir la siguiente información, los viajes suelen comenzar y finalizar el mismo día, pues casi tenemos la misma gráfica. De hecho, menos del 1 % de los viajes comienzan un día y terminan en días diferentes.

Vemos que hay una clara diferencia en la demanda entre semana y los fines de semana, contrario a lo que suponía, el sistema es menos solicitado los fines de semana. Una última observación que tenemos es que cuando comenzó el sistema este se utilizaba con la misma frecuencia todos los días de la semana, y como era de esperarse se ha vuelto más popular día con día, y la distribución mucho más dispersa.

Como última observación, con una segunda muestra de los viajes en el mes de diciembre corroboramos que estas fechas son las de menor demanda, este comportamiento se observó en 2022, pero en 2023 se volvió a observar el mismo patrón e incluso fue aún más abrupto.

## CAPÍTULO 5. VISUALIZACIONES

---



Figura 5.7: Proporción de viajes por tipo de género

Y por último, pero no menos importante, la gráfica de la proporción de los viajes por género a través del tiempo. Donde podemos apreciar un hecho interesante, y es que a pesar de que entre semana es cuando se tiene mayor afluencia en el sistema, el género Femenino prefiere usar más el sistema los fines de semana, lo que nos da un claro indicio de que el fin de semana se utiliza como medio recreativo y aquí es cuando las mujeres están más dispuestas a salir en bicicleta y no entre semana como un método de transporte.

Aunado a lo ya mencionado, hemos aplicado una regresión lineal a la proporción de viajes femeninos a lo largo del tiempo. Los resultados obtenidos indican que la proporción del género femenino ha experimentado un pequeño aumento a lo largo del tiempo. La tendencia de esta métrica fue positiva, ya que  $\beta_1 > 0$ ; sin embargo, cabe destacar que este valor es del orden de las centésimas. En consecuencia, se concluye que el crecimiento es prácticamente nulo y muy lento.

# Capítulo 6

## Generación de la tabla Viajes por día

Dado que queremos hacer una predicción de la demanda lo que vamos a hacer es un conteo de los viajes por día (Retiro) y ver si es que podemos predecir la demanda del sistema una semana después de lo observado en el día. Con base en esta variable y con funciones de *rolling* y *shift* generamos la siguiente tabla con medias, máximos, mínimos, pct's, std's y diferencias, móviles de uno a catorce días. Tomamos métricas móviles en esta venta de tiempo porque en las gráficas de la demanda que previamente mostramos se aprecia una periodicidad semanal en las series de tiempo. Por la misma razón, agregamos una variable categórica que te dice si el día es fin de semana o no. finalmente, agregamos nuestra variable target, que para nuestro caso de estudio es la demanda de siete días posteriores, con lo que conseguimos la siguiente tabla de datos.

## CAPÍTULO 6. GENERACIÓN DE LA TABLA VIAJES POR DÍA

---

Fecha_Retiro	Es_Fin_de_semana	Viajes	Viajes_t-1	Viajes_diff_1	Viajes_pct_1	Viajes_ma_2	Viajes_min_2	Viajes_max_2	Viajes_std_2	Viajes_median_2	Viajes_t-2	Viajes_diff_2	Viajes_pct_2	Viajes_ma_3
2022-09-30		0	13	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2022-10-01	1	4228	13.0000	4215.0000	NaN	2120.5000	13.0000	4226.0000	2980.4551	2120.5000	NaN	NaN	NaN	NaN
2022-10-02	1	4667	4228.0000	439.0000	358.0000	4447.5000	4228.0000	4667.0000	310.4199	4447.5000	13.0000	4654.0000	NaN	2969.3333
2022-10-03	0	6763	4667.0000	2096.0000	0.5996	5715.0000	4667.0000	6763.0000	1482.0958	5715.0000	4228.0000	2535.0000	519.2308	5219.3333
2022-10-04	0	6922	6763.0000	159.0000	0.4832	6842.5000	6763.0000	6922.0000	112.4300	6842.5000	4667.0000	2255.0000	0.6372	6117.3333
2022-10-05	0	6482	6922.0000	-440.0000	-0.0415	6702.0000	6482.0000	6922.0000	311.1270	6702.0000	6763.0000	-281.0000	0.3889	6722.3333
2022-10-06	0	5341	6482.0000	-1141.0000	-0.2284	5911.5000	5341.0000	6482.0000	806.8088	5911.5000	6922.0000	-1581.0000	-0.2103	6248.3333
2022-10-07	0	6553	5341.0000	1212.0000	0.0110	5947.0000	5341.0000	6553.0000	857.0134	5947.0000	6482.0000	71.0000	-0.0533	6125.3333
2022-10-08	1	4720	6553.0000	-1833.0000	-0.1163	5636.5000	4720.0000	6553.0000	1296.1267	5636.5000	5341.0000	-621.0000	-0.2718	5538.0000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2023-10-23	0	50972	34037.0000	16935.0000	0.5098	42504.5000	34037.0000	50972.0000	11974.8533	42504.5000	33760.0000	17212.0000	-0.0025	39589.6667
2023-10-24	0	45017	50972.0000	-5955.0000	0.3226	47994.5000	45017.0000	50972.0000	4210.8209	47994.5000	34037.0000	10980.0000	0.3334	43142.0000
2023-10-25	0	43852	45017.0000	-1165.0000	-0.1397	44434.5000	43852.0000	45017.0000	823.7794	44434.5000	50972.0000	-7120.0000	0.2884	46613.6667
2023-10-26	0	54736	43852.0000	10884.0000	0.2159	49294.0000	43852.0000	54736.0000	7696.1502	49294.0000	45017.0000	9719.0000	0.0738	47668.3333
2023-10-27	0	50696	54736.0000	-4040.0000	0.1561	52716.0000	50696.0000	54736.0000	2856.7114	52716.0000	43852.0000	6844.0000	0.1262	49761.3333
2023-10-28	1	45919	50696.0000	-4777.0000	-0.1611	48307.5000	45919.0000	50696.0000	3377.8491	48307.5000	54736.0000	-8817.0000	0.0471	50450.3333
2023-10-29	1	30616	45919.0000	-15303.0000	-0.3961	38267.5000	30616.0000	45919.0000	10802.8551	38267.5000	50696.0000	-20089.0000	-0.4407	42410.3333
2023-10-30	0	52100	30616.0000	21484.0000	0.1346	41358.0000	30616.0000	52100.0000	15191.4821	41358.0000	45919.0000	6181.0000	0.0277	42878.3333
2023-10-31	0	56064	52100.0000	3964.0000	0.8312	54082.0000	52100.0000	56064.0000	2802.9713	54082.0000	30616.0000	25448.0000	0.2209	46260.0000

Figura 6.1: Tabla de Viajes por día

# **Capítulo 7**

## **Perdición de la demanda diaria**

Como se mencionó, uno de los objetivos es modelar la demanda a una semana, y para el modelado tenemos estos tres algoritmos diferentes a testear, los primeros dos más sencillos que el tercero. Los modelos que probamos fueron K vecinos más cercanos, Red elástica y una Red Neuronal Recurrente con GRU (Gated Recurrent Units), uno no paramétrico, el otro proveniente de la estadística clásica y el último desarrollado con una de las técnicas más novedosas del Machine Learning.

### **7.1 K vecinos más cercanos**

Para K Vecinos más cercanos obtuvimos métricas nada satisfactorias, en el conjunto de entrenamiento obtuvimos métricas sobre el r2 score de 0.90, mientras con la validación, la métrica resultó menor al 0.50, lo que significa que sería mejor irnos por el promedio de todas las observaciones y esto aún así sería una mejor estimación. Aquí nuestros resultados.

#### **Conjunto de entrenamiento**

r2\_score: 0.90  
error cuadrático medio: 0.005  
error cuadrático absoluto: 0.053

### Conjunto de validación

r2\_score: 0.04  
error cuadrático medio: 0.070  
error cuadrático absoluto: 0.218

## 7.2 Red elástica

Por otro lado, para la red elástica obtuvimos que el mejor modelo que pudimos ajustar obtuvo las siguientes métricas.

### Conjunto de entrenamiento

r2\_score: 0.90  
error cuadrático medio: 0.005  
error cuadrático absoluto: 0.05

### Conjunto de validación

r2\_score: 0.68  
error cuadrático medio: 0.02  
error cuadrático absoluto: 0.11

La red elástica que obtuvimos la generamos con la penalización estándar, dándole el mismo peso tanto a Ridge como a Lasso. El modelo que ajustamos a claramente sufre de sobreajuste, pues en validación nos da un r2 score menor

## CAPÍTULO 7. PERDICIÓN DE LA DEMANDA DIARIA

---

que en el entrenamiento, sin embargo al evaluar el modelo con el conjunto de validación, es decir las fechas más recientes, tenemos que la demanda es bien estimada, incluso con su alta volatilidad, esto lo debemos principalmente a la buena ingeniería de variables que generamos, donde llegamos a tener más de 100 variables predictoras, además de haber considerado variables dummies para distinguir el patrón observado en los fines de semana. Queremos destacar que las métricas de validación están recogiendo el periodo observado en diciembre 2023, periodo con la mayor volatilidad y atipicidad observada.

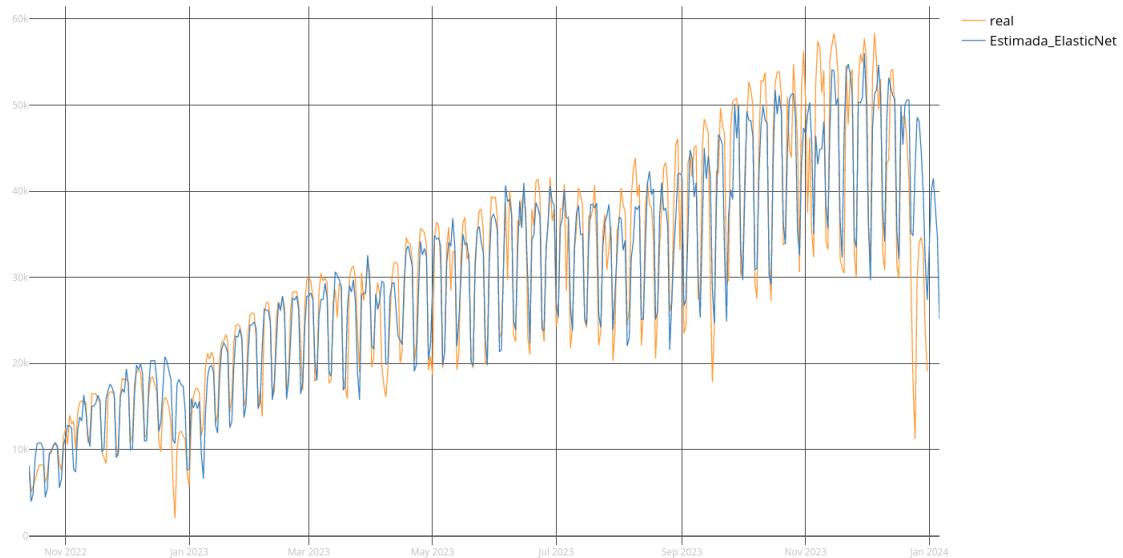


Figura 7.1: Comparación de la demanda estimada con la red elástica contra la real

### 7.3 Red Neuronal Recurrente con GRU

Los resultados de la red neuronal recurrente se obtuvieron a partir de muchas pruebas generadas tanto con redes neuronales convencionales, como con redes con neuronas de memoria de corto y largo plazo (LSTM), sin embargo, se obtuvo el mejor performance con las (Gated Recurrent Units), cabe mencionar que el modelo no ocupó todas las variables de shift y rolling que se mencionaron pre-

## CAPÍTULO 7. PERDICIÓN DE LA DEMANDA DIARIA

---

viamente, pues gracias a la naturaleza de estas redes, estas se encargan de generar toda la ingeniería de variables en el entrenamiento, por lo que sólo fue necesario alimentar a la red con las variables más importantes; *Viajes* y *Es fin de semana* para obtener un buen performance. A continuación se muestra el resumen del modelo, así como los resultados obtenidos.

```
Model: "sequential"
-----
Layer (type)          Output Shape       Param #
-----
gru (GRU)              (None, 50)        8100
dropout (Dropout)     (None, 50)        0
dense (Dense)          (None, 200)       10200
dense_1 (Dense)        (None, 1)         201
-----
Total params: 18501 (72.27 KB)
Trainable params: 18501 (72.27 KB)
Non-trainable params: 0 (0.00 Byte)
```

Figura 7.2: Summary RNN GRU

### Conjunto de entrenamiento

r2_score:	0.86
error cuadrático medio:	0.08
error cuadrático absoluto:	0.05

### Conjunto de validación

r2_score:	0.63
error cuadrático medio:	0.16
error cuadrático absoluto:	0.12

Como comentarios adicionales queremos concluir en esta sección que sin duda alguna nos decantamos por las predicciones realizadas con la red elástica que al tener mejor r2 score que la recurrente es un mejor partido, aún con ese resultado, sí queremos hacer notar el buen desempeño del modelo deriva naturalmente

## CAPÍTULO 7. PERDICIÓN DE LA DEMANDA DIARIA

---

de la ingeniería de variables, cosa de la que carece la red neuronal recurrente y aún así obtiene muy buenos resultados, incluso diríamos que la red neuronal recurrente obtiene mejores resultados en puntos de quiebre como lo son Navidad y Año Nuevo que la tendencia cambia abruptamente.

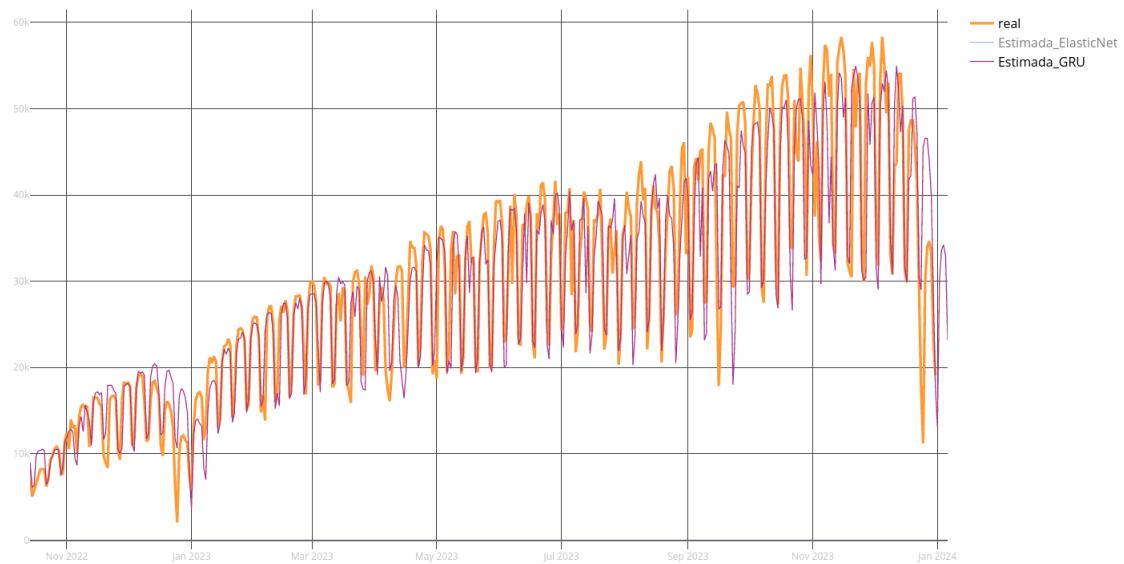


Figura 7.3: Comparación de la demanda estimada con la RNN GRU contra la real

Para concluir con esta sección del modelado de la predicción de la demanda queremos hacer notar que las predicciones sobre la demanda tanto con la red elástica como la RNN con GRU está en la ruta de Outputs del proyectos como [\*Predicciones Demanda.html\*](#), acompañado de una gráfica interactiva donde se pueden consultar las predicciones de la semana posterior, así como los datos observados. He aquí una muestra.

## CAPÍTULO 7. PERDICIÓN DE LA DEMANDA DIARIA

---

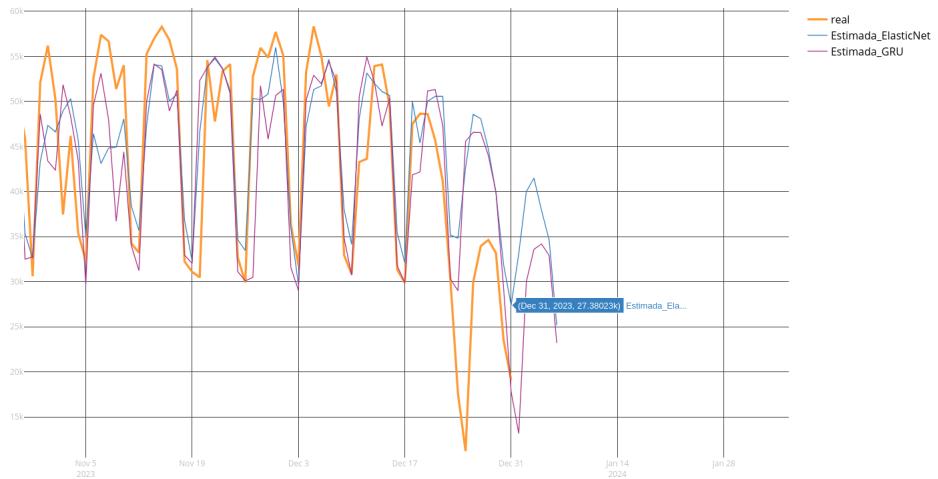


Figura 7.4: Comparación interactiva de las demandas

# **Capítulo 8**

## **Distribución de las estaciones con modelos no supervisados**

Al contar con la información detallada de la localización de las estaciones decidimos hacer varios modelos no supervisados que nos ayudaran a identificar la estructura propia que han generado las estaciones y si es que podemos separarlas en diferentes grupos. Con la modelación no supervisada queremos responder las siguientes preguntas.

1. ¿Hay una agrupación natural en términos de dispersión de las estaciones?
2. ¿Todas las estaciones cumplen con un criterio básico de cercanía?, y si es así, ¿Cuáles no y por qué?
3. ¿Cómo se distribuye la demanda diaria dentro de las diferentes estaciones del sistema, existen a caso puntos críticos o estaciones poco utilizadas y por qué están caracterizadas?

Para las primeras dos preguntas tomaremos como base la distribución espacial de las estaciones, por lo que sólo utilizaremos las variables latitud y longitud. Mientras que para la tercer pregunta tenemos que agrupar los viajes Retiro/Arribo y obtener un promedio de los viajes diarios que parten y llegan por cada estación del sistema

## CAPÍTULO 8. DISTRIBUCIÓN DE LAS ESTACIONES CON MODELOS NO SUPERVISADOS

---

Cabe mencionar que al trabajar con sólo la latitud y longitud de las estaciones no se necesita la ayuda de algoritmos que nos ayudarán a visualizar las distancias entre las estaciones en dimensiones reducidas, porque de por sí ya estamos trabajando en dos dimensiones. Con el mismo argumento es que justificamos el proceder con la resolución de la tercer pregunta.

### 8.1 K means, Agrupación natural de las estaciones

Para hallar una distribución natural de las estaciones jugamos con diferentes algoritmos no supervisados, para resolver la primer pregunta trabajamos con los algoritmos de K Mean y GMM. Como se obtenían resultados muy similares, pues como sabemos ambos algoritmos tienen la misma base, simplemente para el modelo de Gaussianas necesitamos estimar las varianzas de los clusters, así que por simplicidad nos decantamos por el modelo de K means y con ayuda de la inercia, encontramos que;

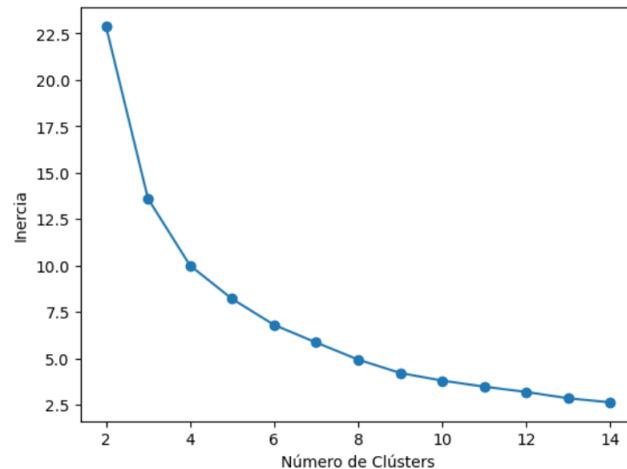


Figura 8.1: Inercia con K Means

El número óptimo de clusters en los que deberíamos agrupar a las estaciones es de tres o cuatro. Al calcular el score de silhouette obtenemos:

## CAPÍTULO 8. DISTRIBUCIÓN DE LAS ESTACIONES CON MODELOS NO SUPERVISADOS

### Score de silhouette

Con tres grupos: 0.39  
Con cuatro grupos: 0.36

Con base en los resultados anteriores tenemos que cualquiera de las dos agrupaciones nos arroja buenos resultados, sin embargo visualmente con cuatro grupos tenemos una mejor interpretación y segmentación, a continuación la misma.

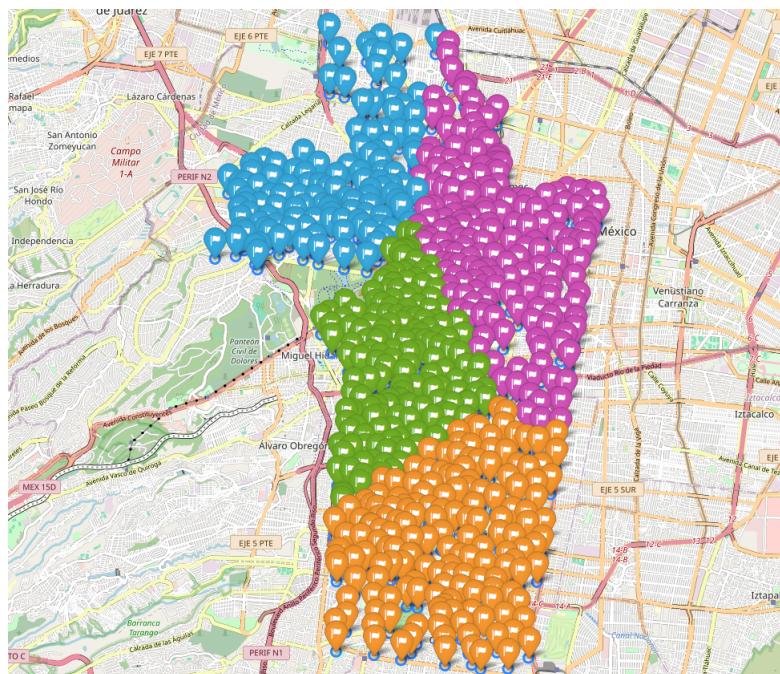


Figura 8.2: Clusters con K Means

El primer grupo se extiende de sur a Norte desde Taxqueña, pasando por Coyoacán hasta llegar a Xola (*Banderas naranjas*). Mientras el segundo Grupo se caracteriza por comenzar desde Av. Xola en el sur hasta el cruce de Av. Reforma con Circuito Interior, podríamos decir que toda esta zona circunda las colonias Nápoles, Roma y termina en Reforma (*Banderas verdes*). De Reforma hacia el Oriente es que se despliega el tercer grupo que podríamos caracterizarlo por ser

## CAPÍTULO 8. DISTRIBUCIÓN DE LAS ESTACIONES CON MODELOS NO SUPERVISADOS

---

el generado por las estaciones que pasan por la Doctores y terminan en el Zócalo de la Ciudad de México y se extiende al norte hasta llegar al Forum Buenavista (*Banderas rosas*). Por último tenemos al grupo que se extiende al norte del desde el Bosque de Chapultepec que engloba toda la zona de Polanco y termina después de la plaza Carso por metro Popotla (*Banderas azules*). Con las descripciones dadas podemos concluir nombrando a nuestros clusters;

1. Cluster Coyoacán
2. Cluster Roma
3. Cluster Centro
4. Cluster Polanco

Más adelante en este mismo análisis y con ayuda de la pregunta número tres es que podremos darle un perfilamiento un poco más profundo a estas segmentaciones encontradas.

## 8.2 DBSCAN para estaciones de difícil acceso

Adicionalmente, con los mismos datos de la distribución de las estaciones quisimos aplicar un modelo basado en DBSCAN con el fin de hallar todas estas estaciones de difícil acceso, para esto, tuvimos que definir dos parámetros, que las estaciones fueran asequibles en un cluster de a lo menos 3 estaciones con un radio de  $200\text{ m}$  a la redonda para considerar a una estación asequible la una a la otra. Tomamos en cuenta estos parámetros porque son los requisitos mínimos necesarios para poder encontrar agrupaciones con el algoritmo de DBSCAN, pero más que agrupaciones, lo que queremos hallar son estos atípicos que no pueden cumplir la condición de conexidad, ahora, pertenecer a un cluster que cumpla estos requerimientos mínimos de conectividad y fácil acceso es esencial, ya que como lo presentamos previamente, uno de los grandes problemas, según los usuarios del sistema es que falta mantenimiento a las bicis y que no siempre hay bicis disponibles, por lo que al estar en alguna de estas dos problemáticas, tener un mínimo de estaciones a la redonda que tengan espacios o bicicletas disponibles podría sacarte del apuro. Con lo que encontramos la siguiente información.

## CAPÍTULO 8. DISTRIBUCIÓN DE LAS ESTACIONES CON MODELOS NO SUPERVISADOS

---

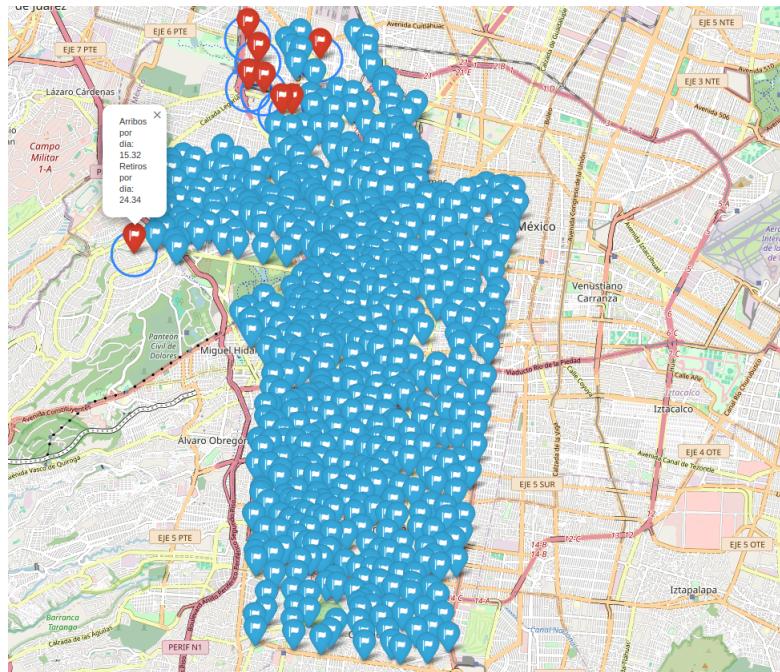


Figura 8.3: Estaciones de difícil acceso en octubre 2023

Según las restricciones colocadas tenemos que 8 estaciones de las 651 que actualmente se encuentran adaptadas para el sistema se consideran de difícil acceso y de conectividad mínima (*Banderas rojas*), mientras que el resto están muy bien conectadas y fácilmente se puede encontrar una dos nuevas estación a lo más a cuatrocientos metros a la redonda.

Como observación adicional y a manera de seguimiento, queremos presentar el algoritmo de detección de estaciones de difícil acceso realizado con las estaciones que estaban vigentes a finales de agosto donde encontrábamos más de 30 estaciones de difícil acceso, además, podemos corroborar visualmente que conforme se han agregado más estaciones al sistema se ha intentado seguir con este principio de conectividad reduciendo cada vez más el número de estaciones no conexas.

## CAPÍTULO 8. DISTRIBUCIÓN DE LAS ESTACIONES CON MODELOS NO SUPERVISADOS



Figura 8.4: Estaciones de difícil acceso en agosto 2023

### 8.3 K means, agrupación por número de viajes en el día

Con ayuda de los viajes realizados pudimos cruzar esta información con las estaciones para poder obtener un promedio de los viajes que parten y llegan por día en cada una de las estaciones del sistema. Al tener esta información, lo que aplicamos fue una segmentación no supervisada de las estaciones respecto a estas dos variables, para encontrar agrupaciones naturales. Los resultados obtenidos nos marcan que el número de agrupaciones ideales deberían de ser cuatro.

## CAPÍTULO 8. DISTRIBUCIÓN DE LAS ESTACIONES CON MODELOS NO SUPERVISADOS

---

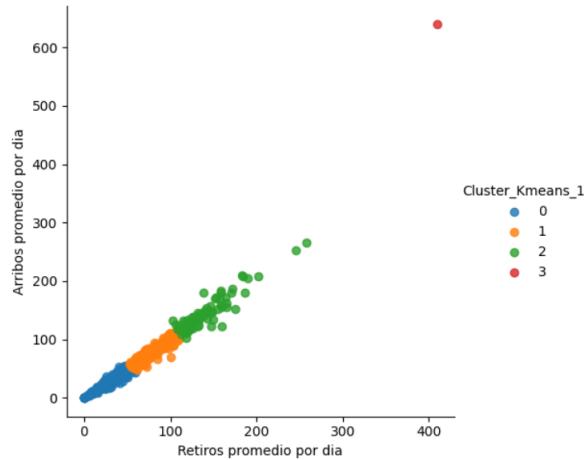


Figura 8.5: clusters, Retiros y Arribos

En esta gráfica podemos observar cómo los datos nos arrojan tres agrupaciones naturales, el cluster 0 es el de las estaciones de poco uso, el cluster 1 son las estaciones de uso medio, mientras que el cluster 2 son las estaciones de mayor demanda y finalmente, el cluster 3 que sólo es formado por una estación es un dato atípico de la estación con mayor demanda diaria, más adelante comentaremos más sobre esta estación.

Al cruzar esta información con la localización espacial de las estaciones tenemos un mayor overview de las cosas, a continuación los resultados, seguido de su análisis.

## CAPÍTULO 8. DISTRIBUCIÓN DE LAS ESTACIONES CON MODELOS NO SUPERVISADOS

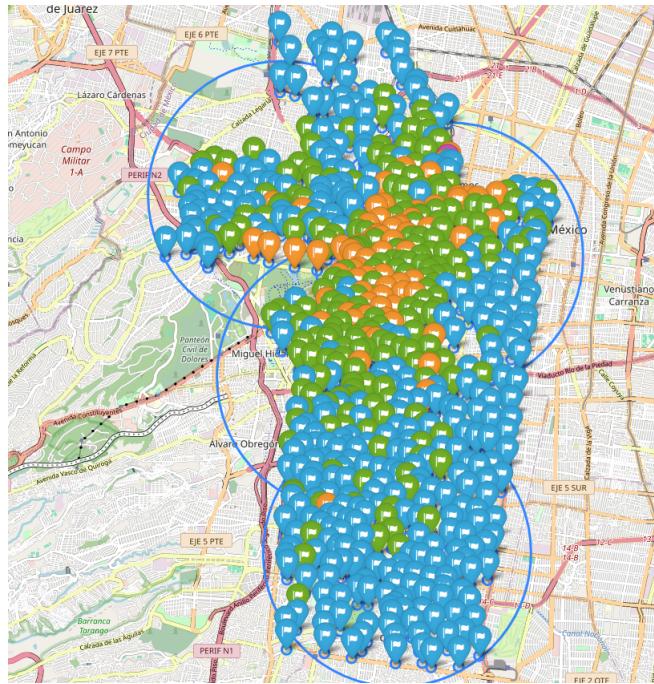


Figura 8.6: clusters, Retiros y Arribos

Primero comenzaremos comentando que los clusters que visualizamos, estos siguen la misma distribución que ya se ha comentado, azul para los de menor demanda, verdes para los demanda moderada y naranja para los de mayor demanda. Como podemos apreciar las estaciones de menor demanda son las que más imperan en el sistema y así mismo sucede lo mismo con las estaciones de demanda media y alta demanda.

Como queremos relacionar esta información con los clusters hallados en la primer sección de la agrupación natural de las estaciones y obtener un perfilamiento de estos grupos, graficamos cuatro círculos de 4 km de radio para poder identificar rápidamente a los centroides de los clusters (Coyoacán, Roma, Centro y Polanco) en donde podemos ver fácilmente de qué tipos de estaciones se compone cada cluster, por ejemplo, que el cluster Coyoacán vemos que hay estacione de menor demandado en el día a día, pues casi todas sus estaciones pertenecen al grupo de estaciones de menor demanda (estaciones), esto es predecible tomando en cuenta que ese cluster esta compuesto en su mayoría de áreas residenciales y

## CAPÍTULO 8. DISTRIBUCIÓN DE LAS ESTACIONES CON MODELOS NO SUPERVISADOS

según la encuesta de ocupación del sistema ecobicis 2020, sólo 8 % de los viajes se utilizan con la finalidad de llegar a casa, mientras que el 48 % de los viajes tienen como finalidad conducir al trabajo. Justo sobre esta última premisa tenemos como soporte que la mayor parte de las estaciones de alta demanda se concentran en la intersección de los clusters Roma, Centro y Polanco, en la zona de Insurgentes y Chapultepec, la cuál es una zona de la ciudad con un gran número de oficinas y por lo tanto a ser ocupadas por oficinistas que intentan llegar al trabajo.

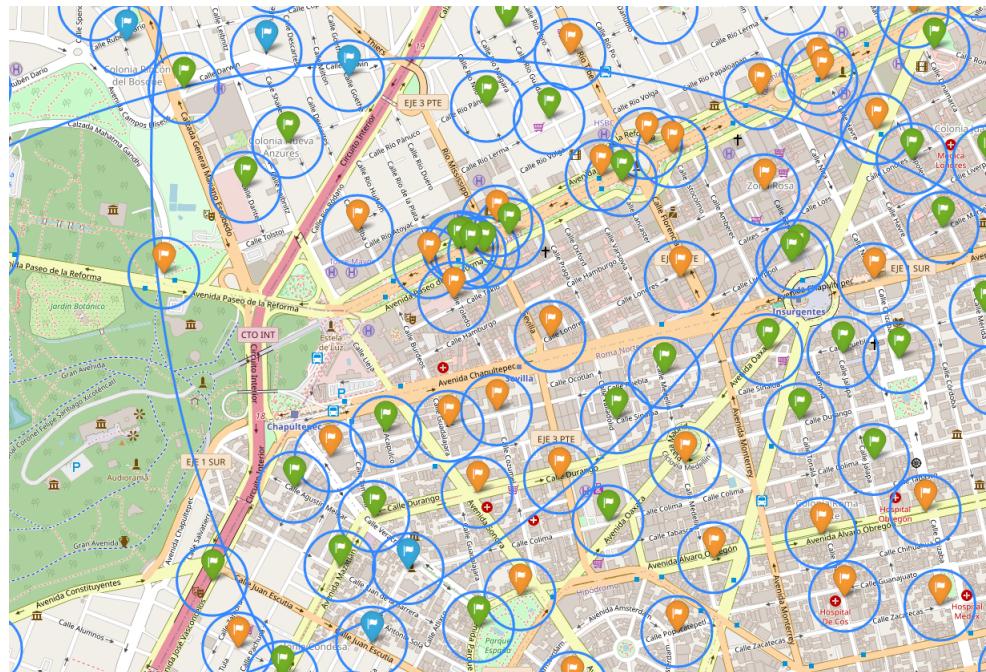


Figura 8.7: Intersección Roma, Centro y Polanco

Otro hecho interesante y que no está de más comentar es que, el sistema apremia a las estaciones que están en puntos estratégicos de la ciudad, es decir, puntos que estén cerca de terminales importantes/centros de movilidad, como lo es Buenavista, de hecho, en ese mismo lugar se encuentra la estación con el mayor número de viajes promedio por día.

## CAPÍTULO 8. DISTRIBUCIÓN DE LAS ESTACIONES CON MODELOS NO SUPERVISADOS

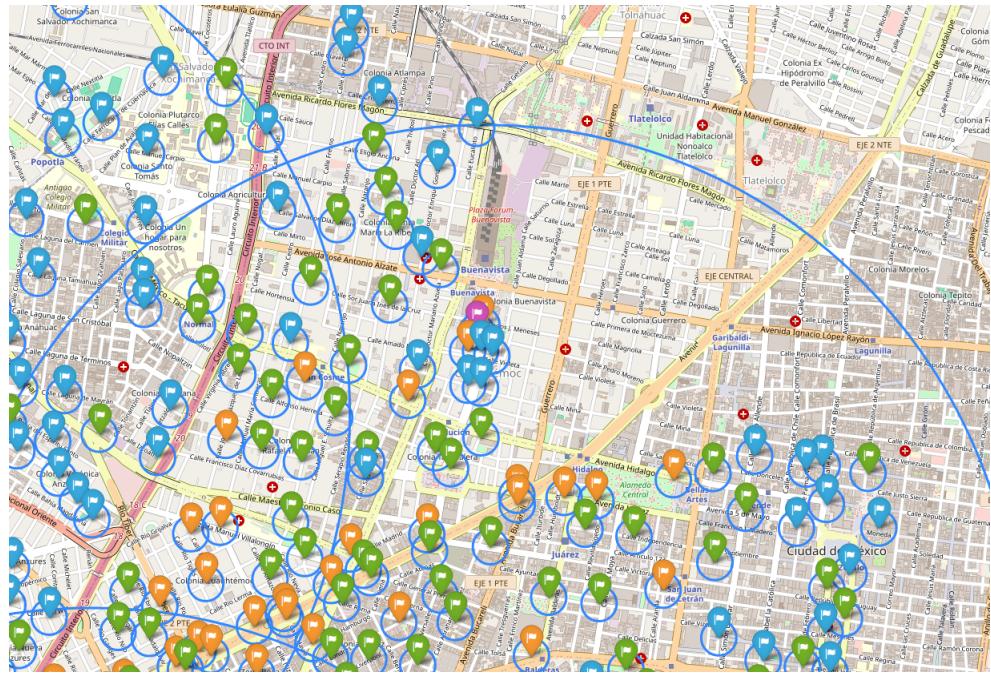


Figura 8.8: Estaciones aledañas a Buenavista

Es claro que los modelos presentados presentan información muy relevante, pero sí que necesitan el criterio de un juicio experto para sacarle mayor provecho a lo presentado, es por eso mismo que toda esta información puede ser consultada en los archivos html's que tenemos en la ruta de outputs, todos son mapas interactivos en los que se puede consultar el detalle, así como desplegar un cuadro informativo sobre la estación y tener un mayor información al respecto.

# **Capítulo 9**

## **Conclusiones**

Sin duda alguna el reto más grande de este proyecto ha sido el preprocesamiento de los datos, desde que la información vienen contaminada, en diferentes formatos y de diferentes sistemas a la vez ni siquiera de un sólo repositorio que concentre toda la información, hizo que la tarea fuera nada fácil. Aunque tengo que confesar que esa fue las partes más enriquecedoras y retadoras. Esto me hace pensar que si uno quiere hacer ciencia de datos de verdad tiene que enfrentar dificultades desde el momento cero y eso fue lo que pasó

Si bien, en un principio contábamos con más información relevante para hacer un buen análisis no sólo de los viajes, usuarios, y las estaciones, también de las bicicletas, un factor de suma importancia, si es que se quieren analizar todas las aristas de lo que está pasando en el sistema, pues se podría obtener un análisis sobre el desgaste y uso de cada bicicleta... eventualmente decidimos desechar esa idea, esto gracias a la limitante de la actualización de los datos. De todos los repositorios que tiene la ciudad de México no encontramos información útil y actualizada sobre los datos que han recolectado las bicicletas en todo su tiempo de vida y trabajar con información que no es almacenada creemos que sería mucho más pernicioso.

He de confesar que en un principio no tenía muy en claro lo que quería hacer con todos estos datos, sé que esta información bien utilizada valía oro, y que algo así podría ayudar a mejorar el servicio que se está ofreciendo a la comunidad, co-

## CAPÍTULO 9. CONCLUSIONES

---

munidad a la que claramente pertenezco y quisiera que se viera beneficiada, pero no sabía con qué empezar; ¿Predicción de la demanda, análisis del desgaste de la infraestructura, análisis de centralidad e importancia de las estaciones?, vaya un par de cosas muy interesantes que se podrían estudiar, así que hice un poco de todo, de tal manera que pudiéramos contar con vistazo general del sistema y cómo este se está desarrollando. Si bien la predicción de la demanda es un problema el cual muchas veces se quiere estudiar, no sólo quería darle ese enfoque, sino agregarle esta otra dimensión de equidad de género, porque desde como yo lo veo, que las mujeres tengan más restringido consciente o inconscientemente otro medio más de transporte es algo que a mí me preocupa y como sociedad debería importarnos. Los datos sobre el tema fueron claros, y si el género femenino utiliza más el sistema de forma creativa que como medio de transporte es un claro indicio de que hay algo que no está bien el sistema, mi tesis es que no se sienten seguras al utilizarlo, aunque lo que me desanima aún más es que el aumento relativo de mujeres usuarias es algo que dista de verse mejor en el futuro, de hecho, podríamos decir que casi es lo contrario pues en la encuesta del 2020 se contaba con más presencia de este segmento de la población.

Si bien quedó claro que los algoritmos entrenados para la predicción de la demanda quedaron cortos en esta última predicción lo que yo propondría sería un re-entrenamiento con las mismas arquitecturas, esto porque como mencionamos, los datos últimamente han aumentado su varianza, se han vuelto más y el sistema sigue manteniendo esta tendencia a la alta. Aunado a que la temporalidad de los datos esta claramente afectada por la estacionalidad anual, como lo comentamos, navidad fue un claro ejemplo de un cambio esperado. Yo creo que con un reentrenamiento del modelo nos podríamos acercar a mejores métricas sin duda alguna. Otra cosas a destacar es que el modelo simplemente se ha ejecutado con información de cierre de mes, justo porque a cierre de mes es que libera la información consolidada de los viajes realizados durante el mismo, así que para sacarle el mejor provecho al código lo lógico sería ligarlo vía API al repositorio y día a día se entregara una estimación de la predicción.

En lo que respecta a la distribución de las estaciones y los modelos no supervisados que generamos, me dejaron más que satisfecho, de hecho fueron mi parte

## CAPÍTULO 9. CONCLUSIONES

---

favorita del trabajo, pues si bien los resultados por sí solos no nos brindan información pura y dura, con mi parte de expertís como usuario, concluimos cosas muy interesantes y yo creo que sería mucho más enriquecedora esta información si llegaran a mano de los administradores. Como yo lo veo, estos resultados dan pauta a resolver n cantidad de problemas relacionados con la gestión y futuras extensiones del sistema, pero por lo mientras se me ocurre que podrían ayudar a resolver grandes problemas de conectividad y demanda, algunas ya se mencionaron en el trabajo, pero aquí vienen dos de las principales ideas por las que comencé con la sección de algoritmos no supervisados.

De la agrupación surgida al implementar K means tenemos el siguiente supuesto; imaginemos que un usuario necesita asistencia, cursos o apoyo vial por algún factor externo, desde una falla técnica hasta un accidente vial, es claro que la mejor forma de apoyar a estos sería con centros de atención estratégicamente distribuidos que pudieran auxiliar a las estaciones más cercanas y que mejor que colocar estos lugares de apoyo que en los centroides que naturalmente surgen de la agrupación de las estaciones.

Continuando con el tema de la conectividad y outliers, dos conceptos que el algoritmo de DBSCAN maneja perfectamente, hallamos las estaciones con menor conectividad. Como observamos, el sistema se en su mayoría se encuentra bien conectado, pero sí existen estaciones de difícil acceso, o con pocos vecinos a la redonda. Intuitivamente lo que propondríamos sería hacer estas estaciones mucho más alcanzables de tal manera que no llegaran a afectar al usuario. Con las actualizaciones realizadas he podido ver en carne propia como es que este sistema se está desarrollando y como lo comenté en la sección de DBSCAN, las nuevas estaciones que han nacido se han ido acoplando de tal manera que intentan mantener un patrón que en un principio presentó el sistema y espero que así siga siendo y en caso de que no, se puedan hallar fácilmente por medio del algoritmo.

Por otro lado y dejando la parte analítica del proyecto un poco de lado y enfocándonos en la parte técnica, sí que queremos mencionar que el proyecto con cada actualización está demandando más y más recursos. Esto porque las tabla viajes ya cuenta con más de 13 millones de datos sobre los viajes, que según la

## CAPÍTULO 9. CONCLUSIONES

---

última vez que corrí el código ya estaba desbordando mi memoria RAM de 16 GiB, por lo que es indispensable montar todo este código bajo un paradigma de Big Data o en su defecto, si es que no se cuenta con los recursos, intentar procesamiento en disco, que no pareciera ser lo mejor ya que de por sí el código tarda cerca de 5 minutos en ejecutarse y con esto podríamos extender este parámetro de aceptable a inaceptable.