Análisis de Movilidad Ecobici 2023-2024

Karla Andrea Palma Villanueva

4 de junio 2025

Índice

1 Análisis Exploratorio de Datos (EDA)

Se integraron más de 11 millones de registros de viajes. Algunos hallazgos clave:

- Estaciones únicas: 680
- Distribución de género: aproximadamente 69 % hombres, 29 % mujeres y 2 % otro.

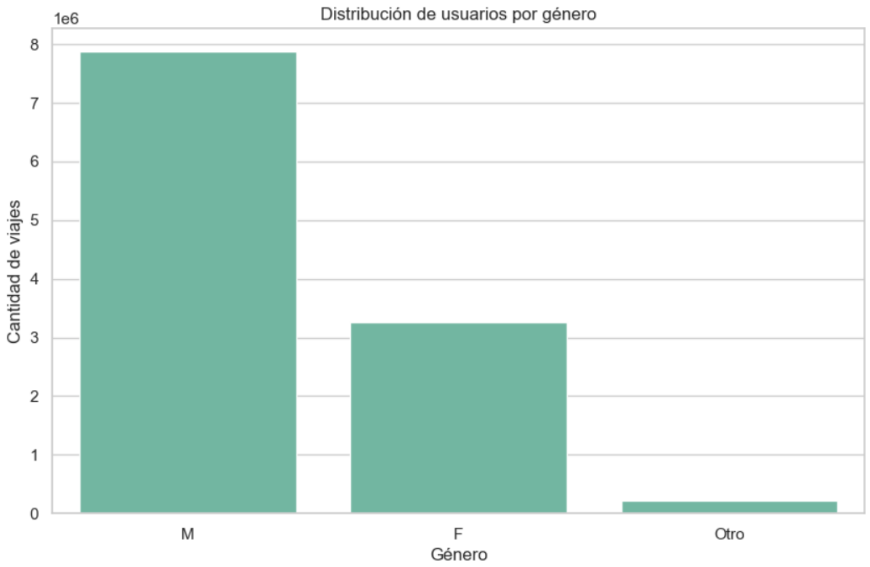


Figura 1: Distribución de usuarios por género

- Edad promedio: concentración entre 25 y 40 años.

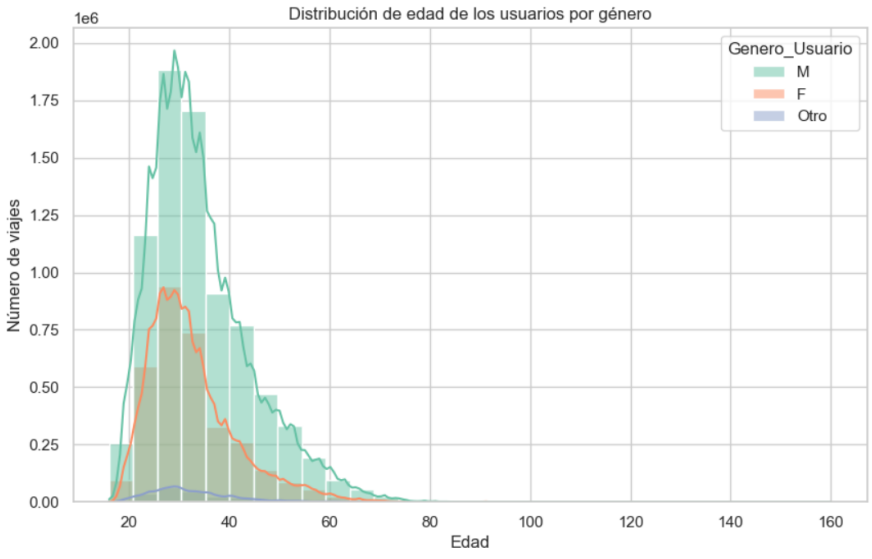


Figura 2: Distribución de edad de los usuarios por género

- Mayor uso entre semana que fines de semana.

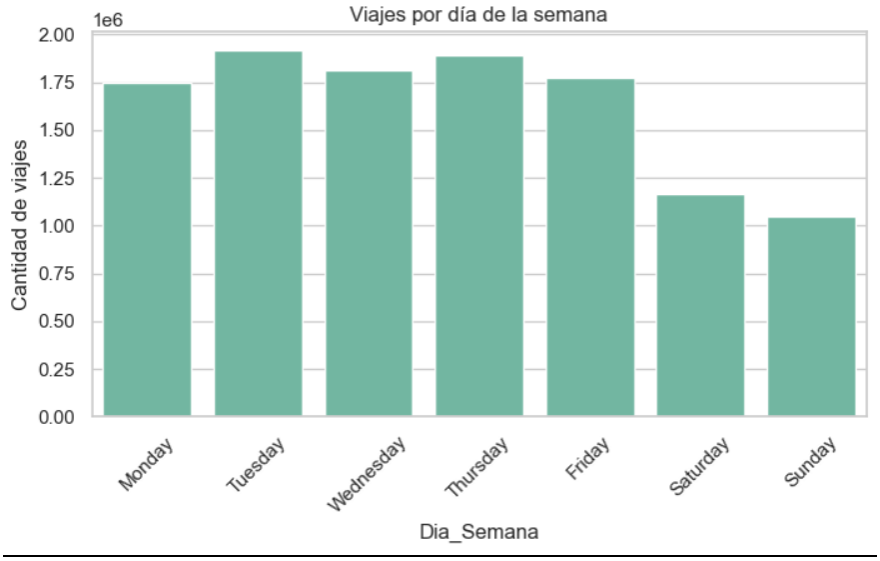


Figura 3: Viajes por día de la semana

- Horas pico: 7-9 a.m. y 5-7 p.m., relacionadas con traslados laborales.

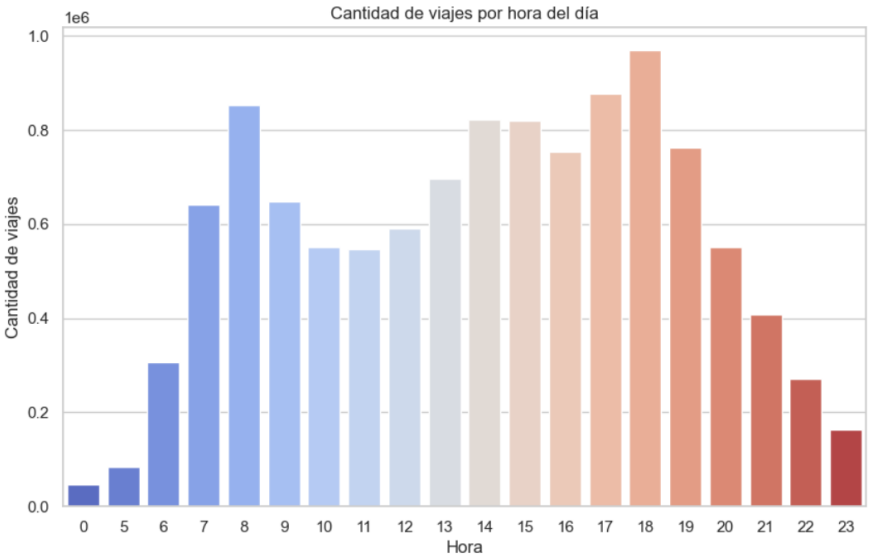


Figura 4: Cantidad de viajes por horas del día

2 Análisis de Horarios y Estaciones

- Las estaciones con mayor afluencia incluyen zonas céntricas y corporativas

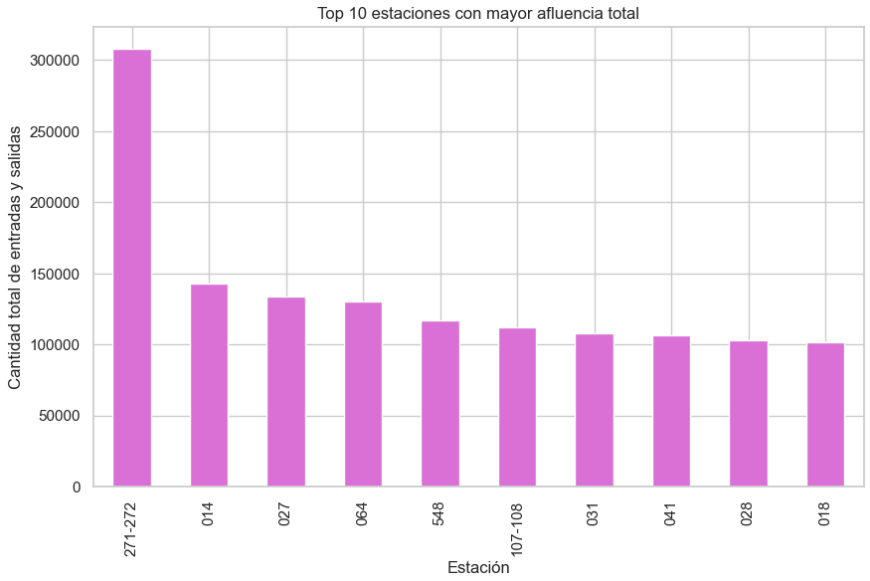


Figura 5: Top 10 estaciones con mayor afluencia total

- Se identificaron estaciones outliers con comportamientos extremos de uso.
- La hora del día es un fuerte indicador de propósito del viaje.

3 Clustering (Aprendizaje No Supervisado)

Se utilizó KMeans y DBSCAN para segmentar estaciones según:

- Cantidad de viajes
- Proporción de usuarias mujeres
- Edad promedio
- Clasificación por tipo de zona (alta o baja demanda)

Para encontrar los patrones en los datos de uso, se implementaron dos modelos de de aprendizaje no supervisado: K-Means y DBSCAN.

3.1 Algoritmo K-Means

Este modelo es muy conocido por la clusterización que se encarga de agrupar observaciones casados en la similitud de sus *features*. En este caso, se usó para identificar patrones de afluencia de bicicletas según variables como hora del día, día de la semana, y estación de origen

3.1.1. Resultados

Clúster	Viajes promedio	Proporción femenina	Edad promedio
0	20,594	0.29	33.97
1	9,045	0.28	34.30
2	43,246	0.29	33.26
3	9,962	0.29	32.24

Cuadro 1: Resumen estadístico de los clústeres generados con K-Means.

Los resultados muestran que la agrupación con las estaciones de mayor volumen de viaje es el *clúster 2* seguido por el *clúster 0*. La agrupación con baja de manda es el *clúster 1*. Recordemos que se evalúo la proproción de usuarias mujeres, entre los clústeres es relativamente constante. La edad promedio es entre 32–34 años.

3.2 DBSCAN

Density-Based Spatial Clustering of Applications with Noise es un también es un algoritmo de clustering que agrupa puntos que están densamente conectados entre sí y puede detectar outliers o ruido.

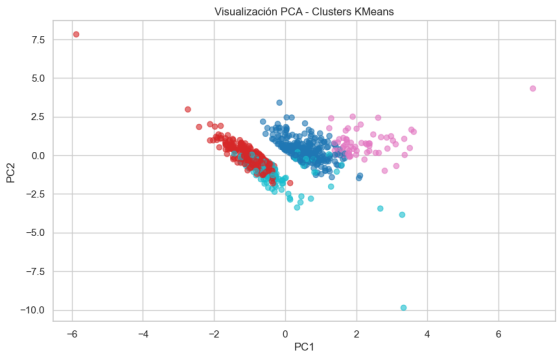
En este caso se utilizó para detectar los patrones espaciales de las estaciones y la detección de outliers de forma natural, de forma que se podrían identificar comportamientos atípicos.

3.2.1. Resultados

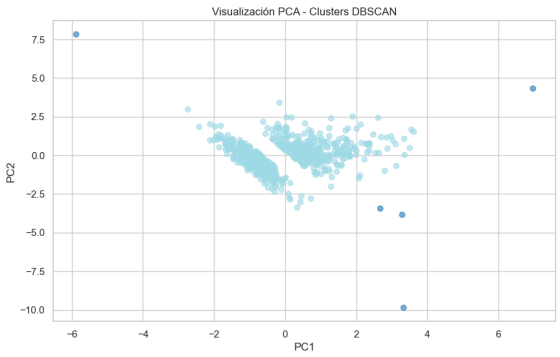
Clúster	Estaciones	Viajes promedio	Proporción femenina	Edad promedio
0	674	16,596	0.29	33.74
-1	5	34,326	0.34	32.28

Cuadro 2: Resumen de los clústeres generados con DBSCAN. El clúster -1 representa outliers.

El clúster -1 representa un grupo reducido de estaciones con un comportamiento atípico: alta demanda, mayor proporción de mujeres usuarias y un perfil más joven. Posiblemente el patrón encontrado tenga que ver con alguna zona estudiantil, comercial o recreativa.



(a) PCA con K-Means



(b) PCA con DBSCAN

Figura 6: Visualización de clústeres generados por K-Means y DBSCAN utilizando PCA.

3.3 Detección de outliers

Se identificaron distintos tipos de outliers, tanto en términos temporales como espaciales. Para detectarlos se aplicaron criterios estadísticos y modelos no supervisados.

Outliers temporales

Se analizaron los tiempos de duración de los viajes, combinando `Fecha_Retiro + Hora_Retiro` y `Fecha Arribo + Hora_Arribo`. Posteriormente, se calculó la duración en minutos para cada trayecto.

Mediante la aplicación del criterio de **IQR (rango intercuartílico)**, se consideraron como outliers aquellos viajes cuya duración superaba los 90 minutos, puesto que :

- El percentil 75 (Q3) estaba alrededor de los 30–40 minutos.
- Las duraciones superiores a 90 minutos exceden el uso esperado de una bicicleta compartida urbana.

4 Análisis Temporal (Regresión Lineal)

Se ajustó un modelo lineal para detectar tendencia mensual de viajes por estación, transformando el campo temporal en una variable numérica ordenada para aplicar la regresión.

- Estaciones con tendencia positiva: en crecimiento sostenido.
- Estaciones con tendencia negativa: podrían requerir diagnóstico técnico o reevaluación de ubicación.

Posteriormente, se entrenó una regresión lineal individual por estación. La pendiente de cada modelo indica la dirección de la tendencia:

- **Pendiente positiva**: indica que el número de viajes ha ido en aumento con el paso del tiempo.

- **Pendiente negativa**: sugiere una disminución en la actividad de la estación, lo cual puede implicar problemas de operación, cambios de movilidad local o desuso.

Finalmente, se ordenaron las estaciones por pendiente para identificar:

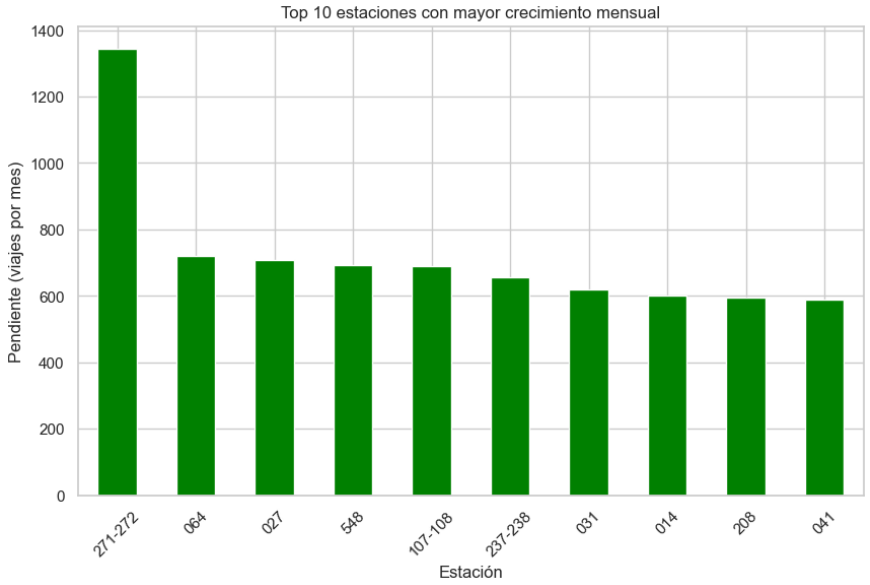


Figura 7: 10 estaciones con mayor crecimiento

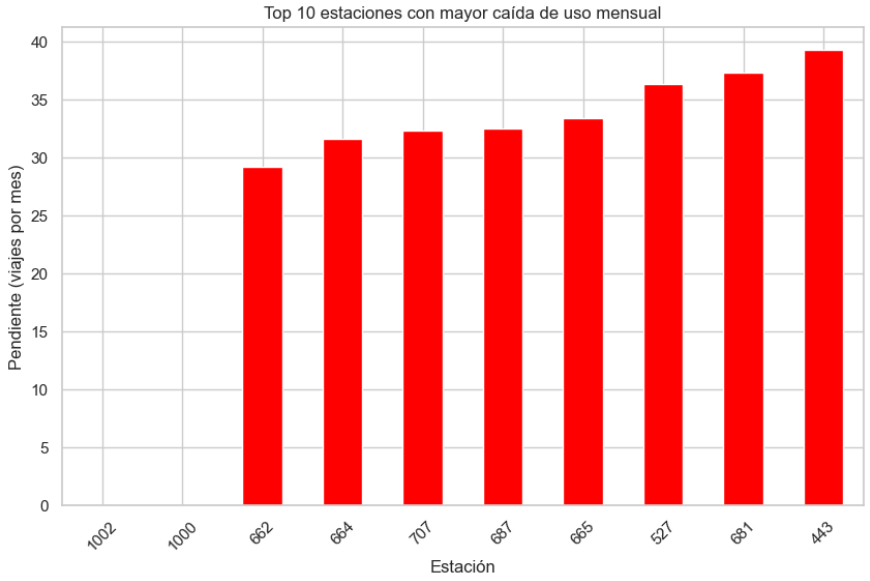


Figura 8: 10 estaciones con mayor caída de uso

5 Conclusiones

El análisis de los datos de Ecobici correspondientes a los años 2023 y 2024 permitió identificar patrones consistentes de uso caracterizados principalmente con la movilidad laboral, reflejadndo una alta demanda durante los horarios pico matutinos y vespertinos. .

6 Anexo: Justificación del Diseño Gráfico

- Se utilizaron paletas suaves (Set2, coolwarm) para facilitar la lectura y accesibilidad.
- Los gráficos de barras permiten identificar rápidamente valores extremos.
- El uso de PCA en visualizaciones de clustering ayuda a interpretar dimensiones complejas.
- Las gráficas están ordenadas e identificadas con títulos y etiquetas claras.

7 Fuentes de Datos

Los datos utilizados en este análisis fueron descargados desde el portal de datos abiertos de la Ciudad de México:

- `https://ecobici.cdmx.gob.mx/datos-abiertos/`

Para la implementación de los modelos de clustering se utilizó la biblioteca `scikit-learn`, cuya documentación oficial puede consultarse en:

- K-Means: `https://scikit-learn.org/stable/modules/clustering.html#k-means`
- DBSCAN: `https://scikit-learn.org/stable/modules/clustering.html#dbscan`

Archivos procesados:

- ecobici_2023_01.csv a ecobici_2023_12.csv
- 2024-01.csv a 2024-12.csv