Trabajo Práctico 1

Grupo 01

Florencia Dellisola - Beltrán Malbrán - Gian Luca Spagnolo - Marcela Cruz

Análisis de Datos

Análisis exploratorio de datos: Taxi Yellow cab

Descripción del dataset

- Meses enero, febrero y marzo de 2023.
- 9.384.487 filas y 20 columnas
- Columnas con valores predeterminados.
- Variables cualitativas y cuantitativas.

Limpieza de datos

- 1.180.895 valores nulos totales
- Imputación con la mediana para las númericas
- Store_and_fwd_flag
- VendorID
- RatecodeID
- Columnas eliminadas

Generación de nuevas Features

- Fecha.
- Tipos de viajes.
- Cantidad de pasajeros abordo.
- Franja horaria.
- Quincenas.
- Distritos.

Analisis y correlación de valores atipicos

Correlación de valores

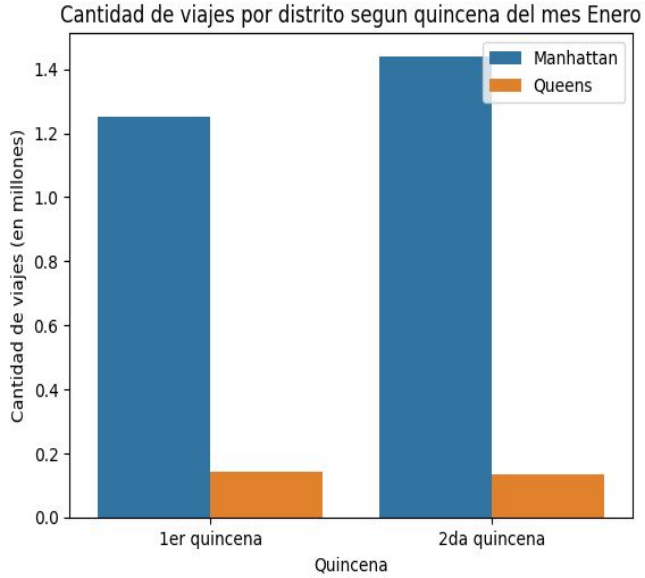
- Relación entre variables cuantitativas
  - Fare_amount
  - Tip_amount
  - Total_amount
  - Tolls_amount

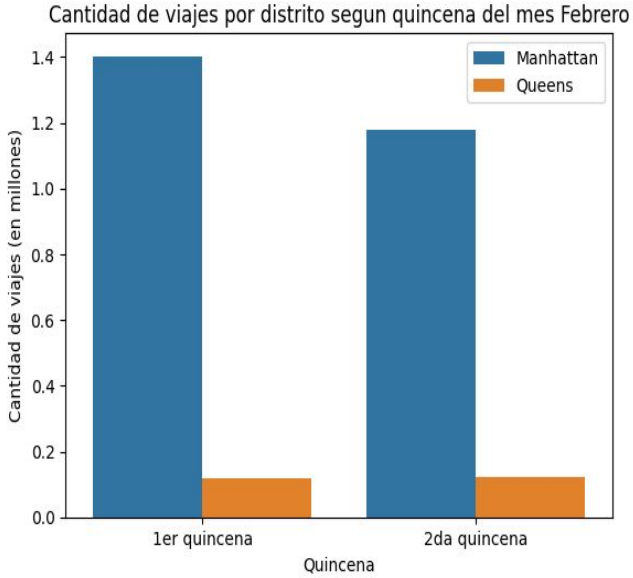
Analisis valores atípicos

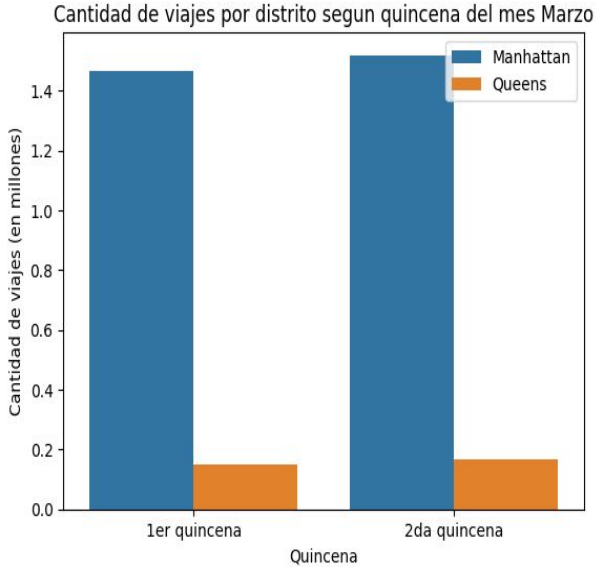
- Multivariado
  - Mahalanobis
- Univariado
  - Rango intercuartículo

Preguntas de Investigación

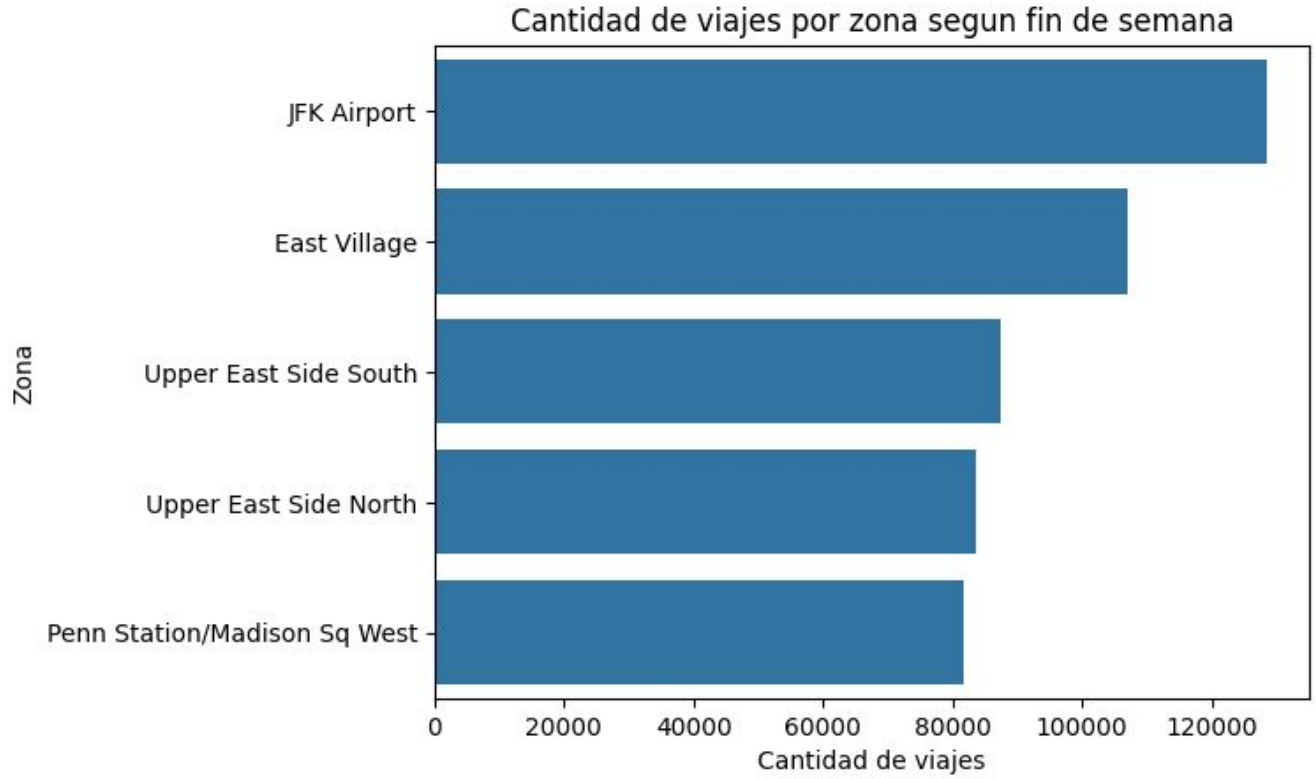
¿Qué distritos solicitaron la mayor cantidad de viajes por quincena?



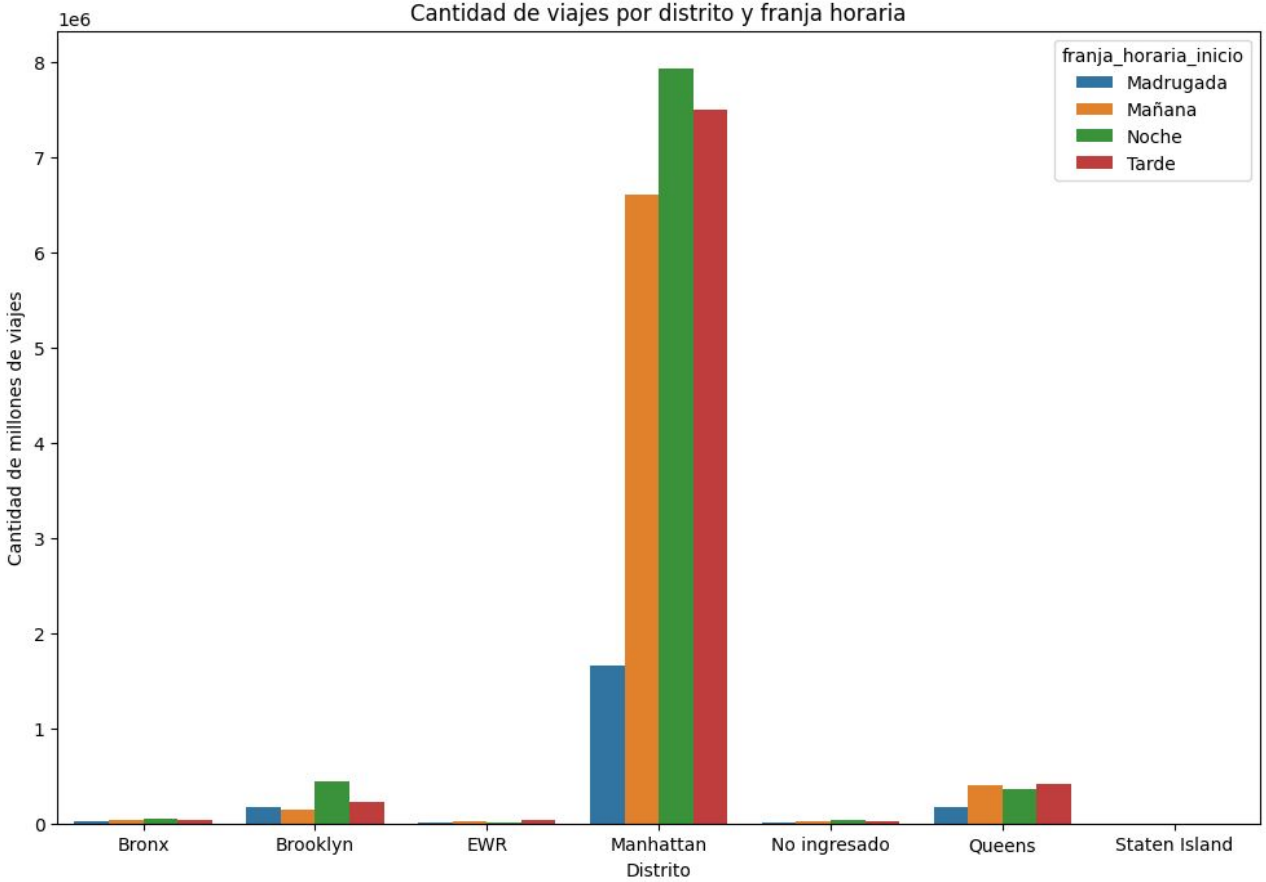




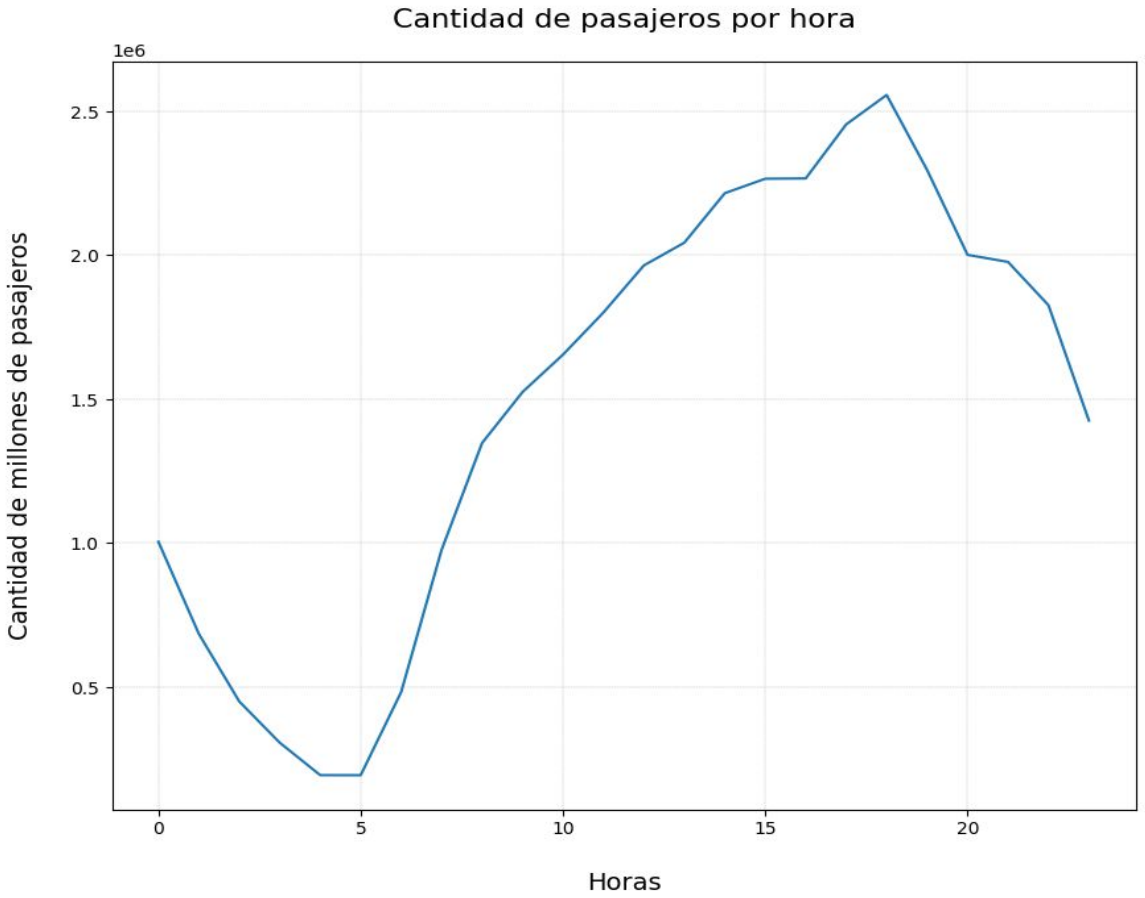
¿Cuál es la zona con mayor cantidad de viajes en el fin de semana?



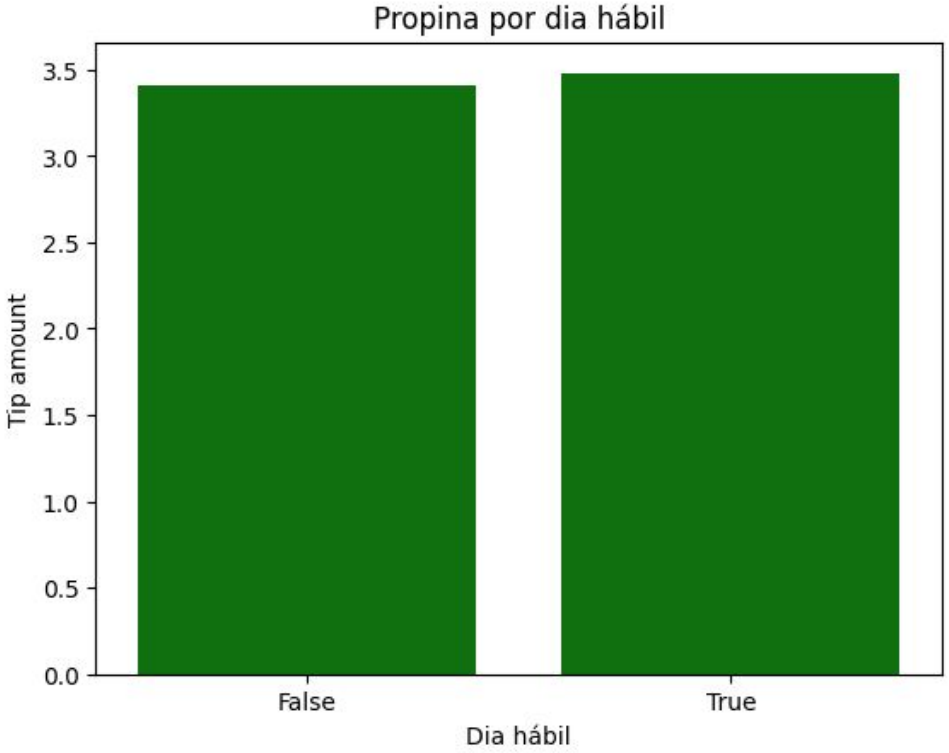
¿Cuál es la zona con mayor cantidad de viajes en el fin de semana?



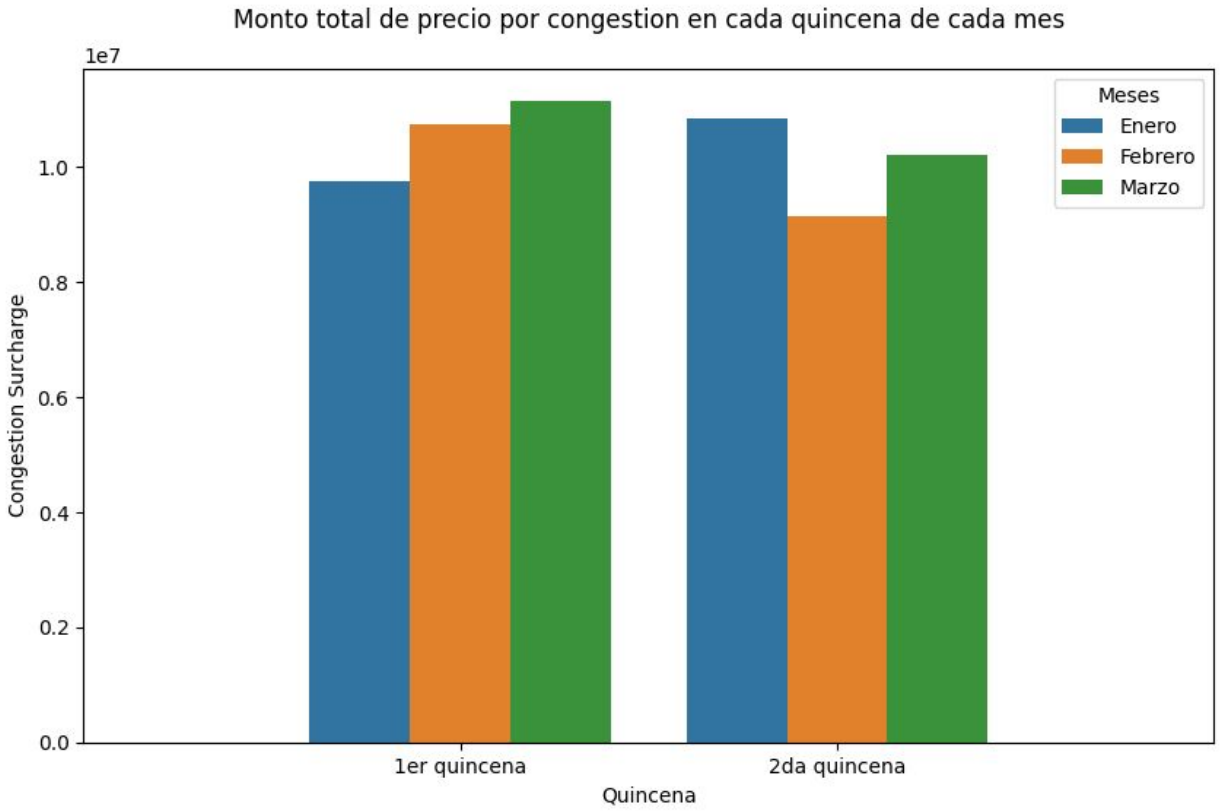
¿Cuál es la distribución de la cantidad de pasajeros en un día?



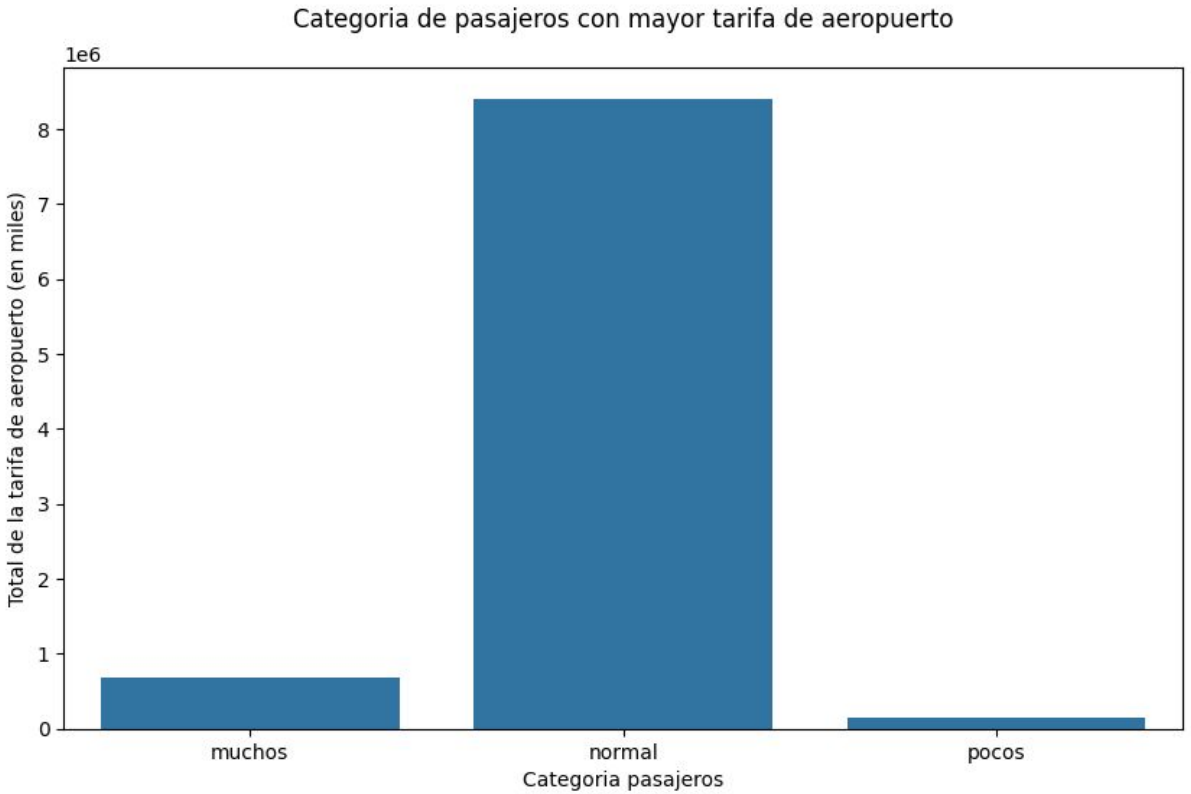
¿Se deja más propina, en promedio, en los días hábiles o los fines de semana?



¿Cuál es la quincena de cada mes en la que hay más congestión?

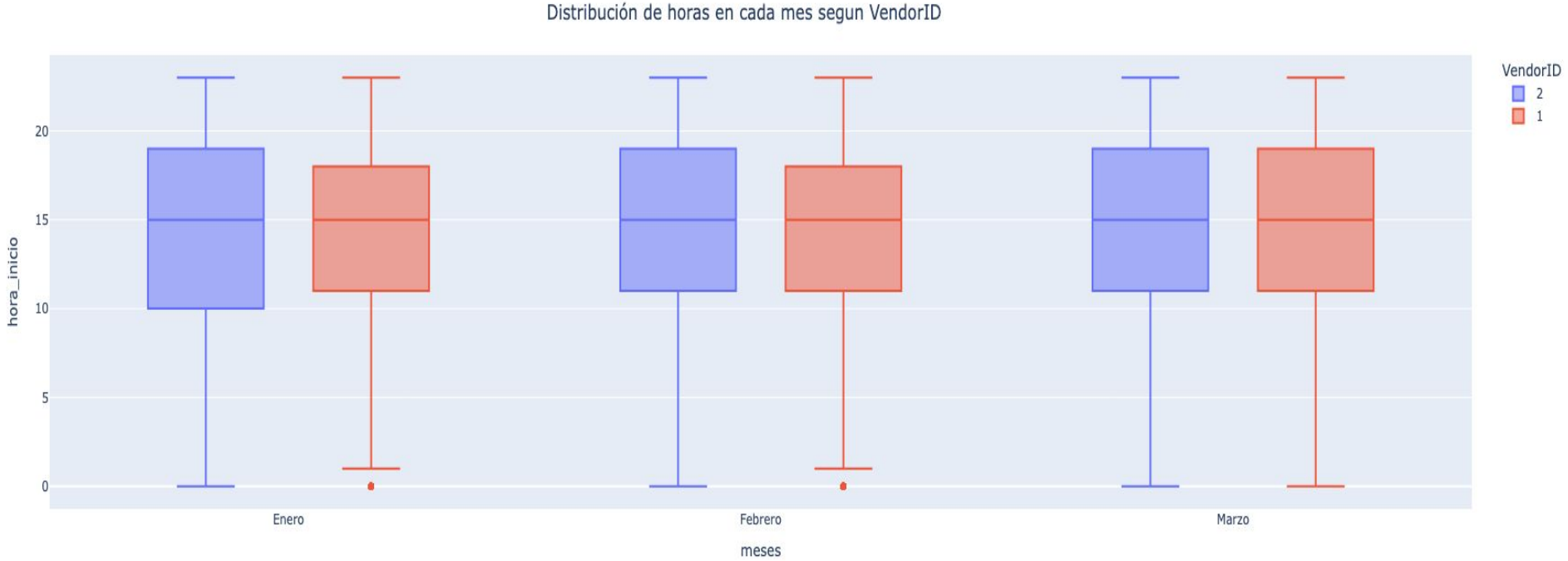


¿Cuántos pasajeros suelen abordar un taxi cuando parte del aeropuerto?



¿Cuál es el promedio de extras según el tipo de pago?

¿En qué horas presenta más viajes cada VendorID y en qué mes?



Modelos de Clasificación

Prepocesamiento de los datos: Modelos de clasificación.

Descripción del dataset

- Predicción de si lloverá mañana o no.
- Distritos Queensland, Nueva Gales del Sur y Terr. de la capital.
- 58.357 filas y 23 columnas.

Limpieza de datos

- No hay datos duplicados.
- Amplio rango de valores nulos.
  - Evaporation
  - Clouds
  - Sunshine
- Imputación de los valores con la moda

Generación de nuevas Features

- Fecha.
- Estación.
- Distrito.

Modelos de clasificación

Encoding

Se utilizo one hot encoder y label encoder

1

Entrenamiento

80% de los datos fueron al train y un 20% al test

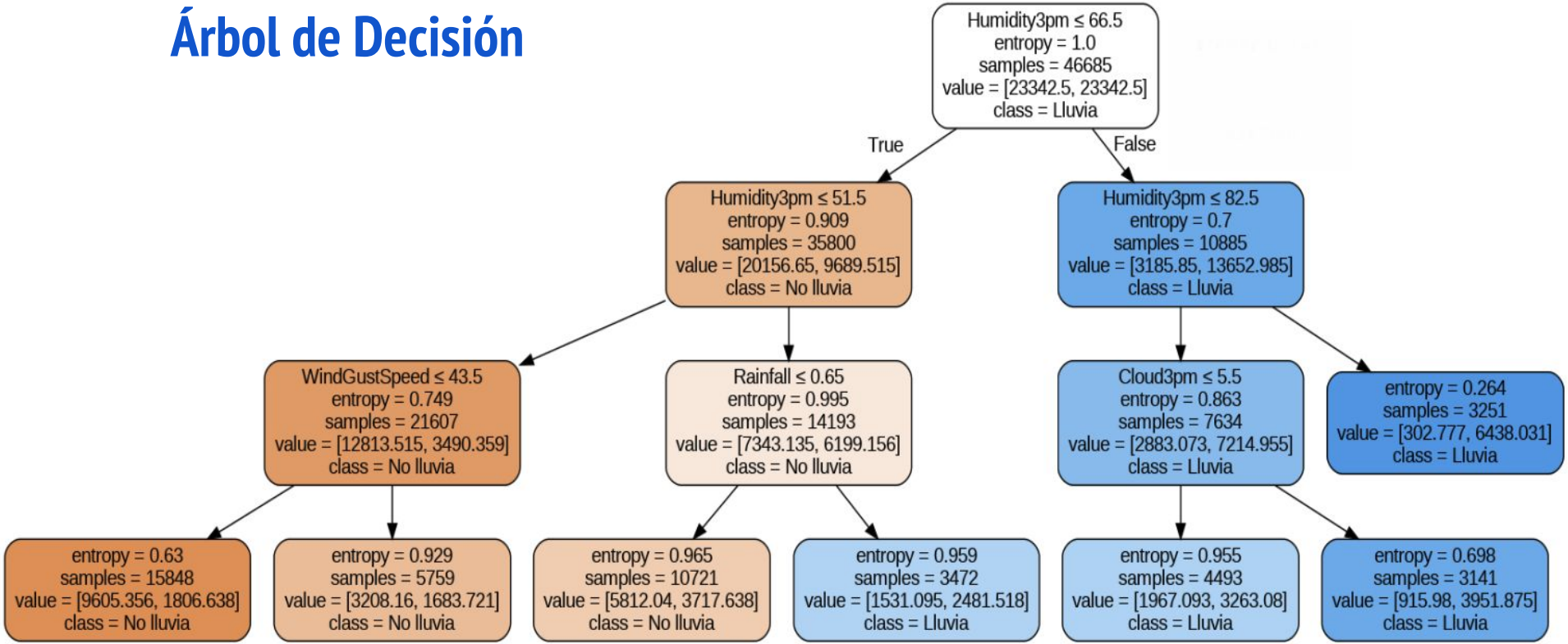
2

Predicción

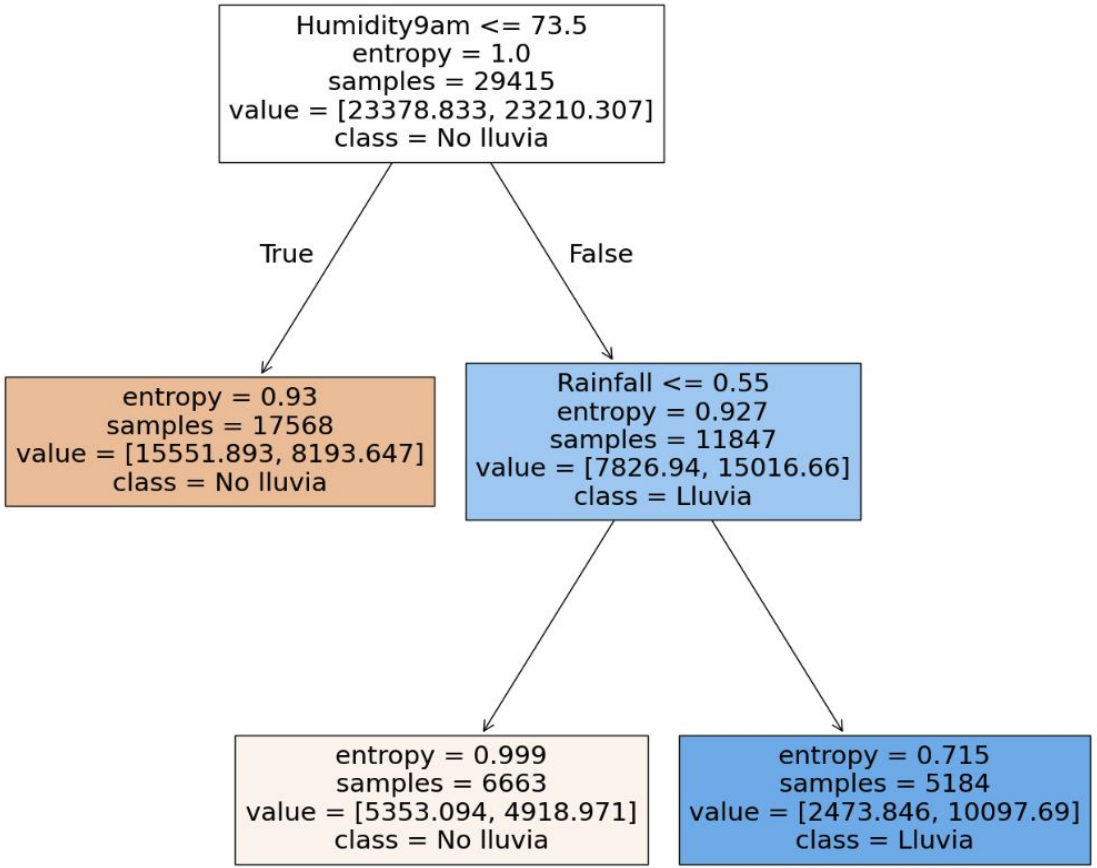
Obtenemos el modelo optimizado y sus predicciones

3

Árbol de Decisión



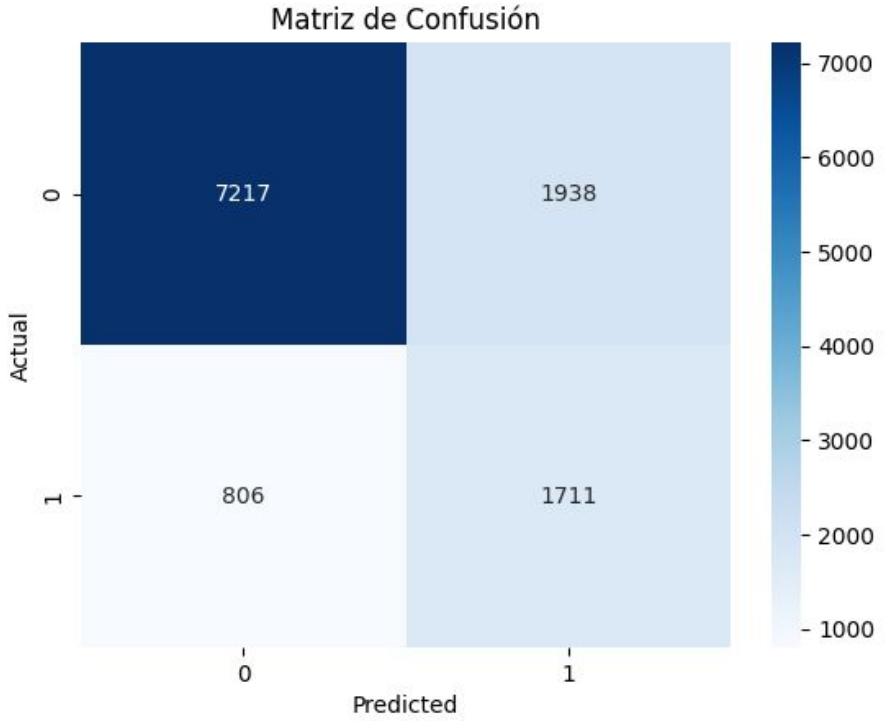
Random Forest



Regresión Logística

Escalado de datos

Validación cruzada del score



Cuadro de resultados de los modelos

Modelo	F1 - Test	Precision Test	Recall Test	Accuracy Test
Arbol de Decisión	0.78	0.81	0.77	0.77
Random Forest	0.78	0.81	0.76	0.76
Regresión logistica	0.80	0.83	0.79	0.79

Modelos de Regresión

Prepocesamiento de los datos: Modelos de clasificación.

Descripción del dataset

- Predicción del precio del alojamiento
- datos de alojamiento de airBnB de Estambul
- 31.758 filas y 18 columnas.

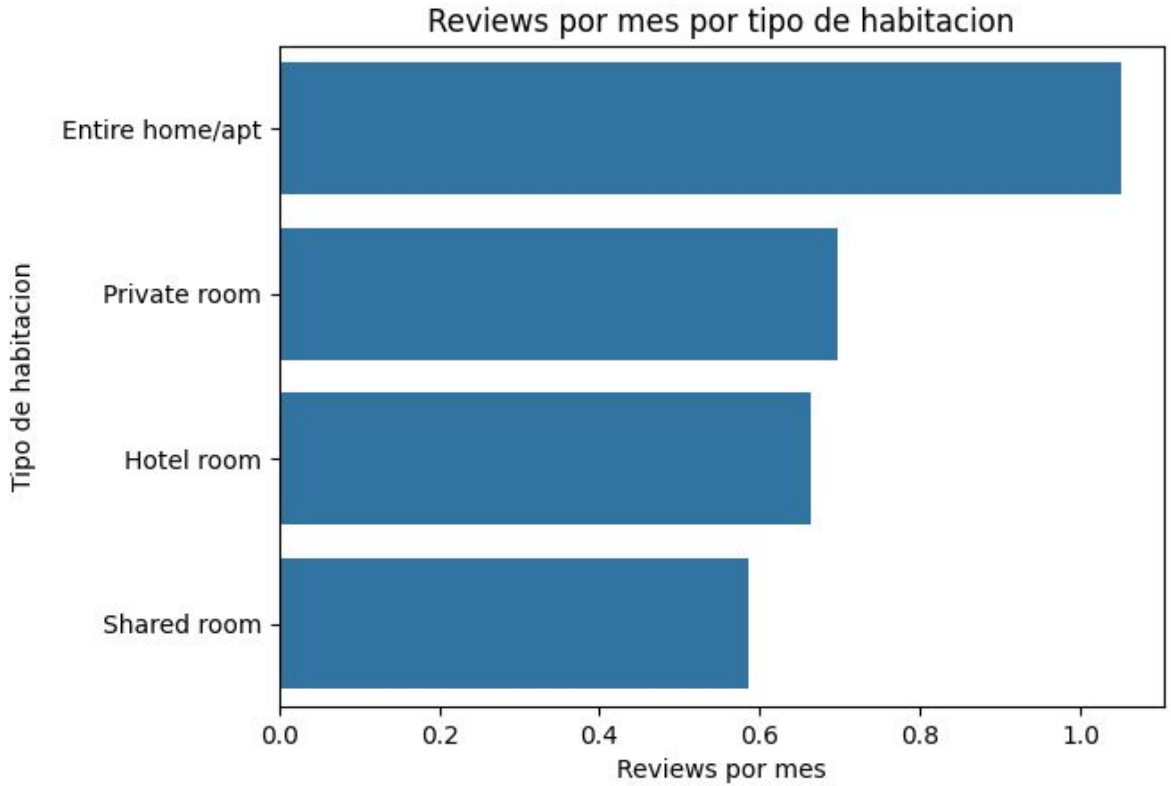
Limpieza de datos

- No hay datos duplicados.
- neighbourhood_group.
- Imputación de los datos
  - Para los numéricos con la mediana.
  - Para las fechas con interpolación.
  - Para los strings con Desconocido.
- Eliminación de last_review.

Generación de nuevas Features

- Fecha.
- descontinuado.

¿Qué tipo de alojamiento posee más reseñas por mes?



Modelos

Regresión lineal

Validación cruzada de los puntajes.

XGBoost

Optimización mediante Grid Search.

LightGBM

Optimización mediante Randomized Search.

Cuadro de resultados de los modelos

Modelo	MSE	RMSE	R2
Regresión Lineal	0.54	0.45	0.30
XGBoost	0.32	0.57	0.50
LightGBM	0.32	0.57	0.50

Clustering

Análisis de tendencia al clustering.

Descripción dataset

- Estadísticas de las playlist de spotify.
- 736 filas y 13 columnas

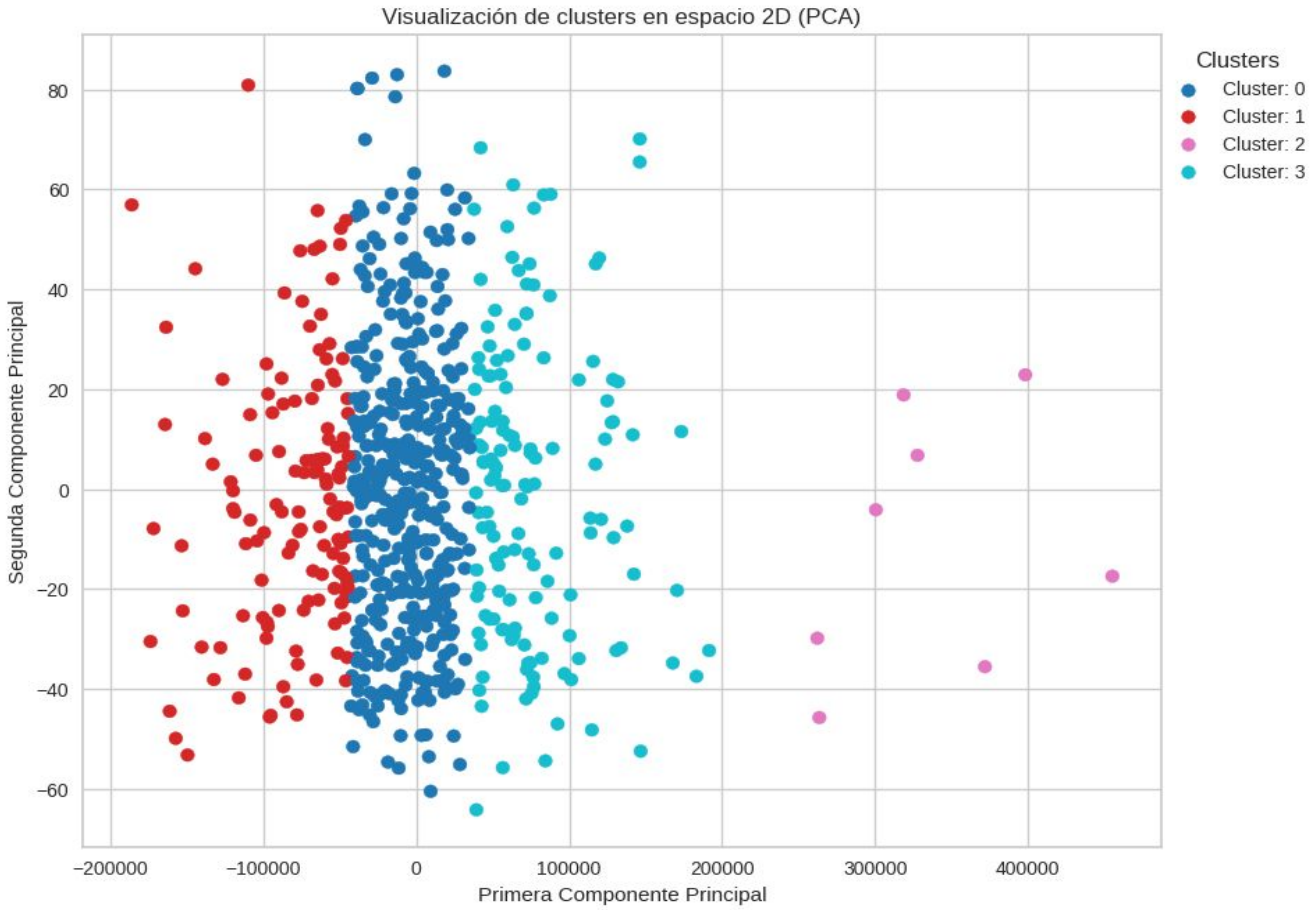
Formación clusters

- Tendencia a agruparse.
- Cantidad optima de clusters.
- Coeficiente de silhouette

Limpieza datos

- Valores duplicados

Distribución de los clusters



Descripción clusters

Cluster 0

Canciones energéticas y bailables.

Cluster 1

Canciones cortas y bailables

Cluster 2

Canciones largas y acústicas.

Cluster 3

Canciones largas y melancólicas.