Trabajo Práctico 1 - Grupo 04

Integrantes:

- Agustin Braida
- Ignacio Carrera
- Matias Etchegoyen
- Agustin Trombetta
- Tomas Ghiglione

EJ1: Análisis Exploratorio

Descripción general del dataset

El dataset utilizado corresponde a los viajes realizados en taxis amarillos de la ciudad de Nueva York durante octubre de 2024 e incluye información detallada de cada viaje.

- Cantidad de registros: aproximadamente 11 millones de viajes.

- Cantidad de columnas: 19, cada una aportando distinta información sobre cada viaje, como la duración, método de pago, cantidad de pasajeros, etc.

Features destacables

A continuación se describen algunas de las variables más relevantes y por qué se consideran clave en el análisis.

Variable	Tipo	Descripción
tpep_pickup_datetime	datetime	Fecha y hora de inicio del viaje. Permite análisis temporales detallados.
tpep_dropoff_datetime	datetime	Fecha y hora de fin del viaje. Junto con la anterior, permite calcular duración.
trip_distance	float	Distancia recorrida en millas. Es central para análisis de eficiencia y costos.
fare_amount	float	Tarifa base cobrada por el viaje, sin incluir recargos.

total_amount	float	Monto total abonado, incluyendo extras y propina.
tip_amount	float	Propina registrada (en general solo con pagos con tarjeta).
payment_type	categórica	Tipo de pago (tarjeta, efectivo, otros). Permite analizar su relación con propinas.
PULocationID / DOLocationID	entero	ID de zona de inicio y fin. Se usó con el diccionario taxi_zone_lookup para obtener nombres de zona y borough.

Hipótesis y supuestos considerados

- Se asumió que los viajes con las variables correspondientes igual a cero (como el valor neto o la duración) eran inválidos o mal registrados, y fueron eliminados o filtrados del análisis.

- Se interpretó que los valores como RatecodeID, payment_type, store_and_fwd_flag, entre otros, son categóricos codificados numéricamente, y se mapearon según documentación oficial o criterios consistentes.

- Para evitar distorsiones y poder realizar visualizaciones de manera lógica,, se eliminaron outliers extremos en duración, distancia y monto total, tanto univariadamente como multivariadamente (LOF, Isolation Forest, Mahalanobis).

- Los registros con recargos inusuales o proporciones atípicas fueron tratados como posibles errores o casos excepcionales.

Preprocesamiento de Datos

Durante el análisis se realizaron diversas tareas de limpieza, transformación y enriquecimiento del dataset original. A continuación se detallan las más relevantes:

¿Se eliminaron columnas?

No se eliminaron columnas completas, ya que todas las variables originales resultaron relevantes en al menos una parte del análisis o para representar alguna visualización. Pero hay dos columnas que no aportaron mucha información a los análisis y si podrían haber sido eliminadas por completo:

- VendorID: Se utilizó para realizar una visualización para mostrar la proporción de proveedores de viaje.

- improvement_surcharge: tenía un valor fijo en la mayoría de los registros por lo cual no aportó información muy valiosa.

¿Detectaron correlaciones interesantes?

Se calculó una matriz de correlación entre variables numéricas. Algunas relaciones destacables fueron:

- trip_distance y trip_duration: es interesante ver la poca correlación, lo cual tiene sentido debido a todos los factores que influyen en lo que tarda un viaje de taxi (tráfico, horario de viaje, velocidad, etc).

- fare_amount y total_amount: correlación muy alta, lo cual es esperable dado que la tarifa base es el valor más influyente sobre el valor total.

- tip_amount y total_amount: también mostraron buena correlación (pero obviamente bastante menor), indicando que las propinas influyen de forma significativa en el valor total del viaje.

¿Generaron nuevos features?

Sí, se generaron varias variables derivadas para enriquecer el análisis y facilitar la visualización de patrones. Algunas de las más importantes fueron:

- trip_duration: duración del viaje en minutos.

- avg_speed_mph: velocidad promedio del viaje.

- tip_pct: porcentaje de propina sobre el valor base (fare_amount).

- fare_per_min: tarifa base dividida por duración.

- total_extra_fees: suma de recargos (extra, mta_tax, congestion_surcharge, airport_fee).

- hour_of_day, day_of_week, day_name: variables temporales.

- pickup_borough, dropoff_borough: zonas geográficas agregadas a partir del diccionario de zonas.

¿Encontraron valores atípicos? ¿Qué técnicas utilizaron y qué decisiones tomaron?

Sí, se detectaron valores atípicos tanto univariadamente como multivariadamente.

Univariado:

- Se utilizaron reglas basadas en percentiles y método del IQR (rango intercuartílico) para variables como trip_duration, trip_distance y total_amount.

- Se eliminaron registros con valores extremos muy alejados (ej: viajes de más de 200 millas o duraciones mayores a 1000 minutos).

Multivariado:

- Se aplicaron técnicas como:
    - Local Outlier Factor (LOF)
    - Isolation Forest
    - Distancia de Mahalanobis
- Se utilizaron para pares como:
    - trip_distance vs trip_duration
    - fare_amount vs tip_amount
    - trip_duration vs total_amount

Decisión: se eliminaron únicamente los outliers muy severos, los que podríamos concluir que son valores mal ingresados o algún tipo de error, pero que no tendría ningún sentido que represente a un viaje real.

¿Qué columnas tenían datos faltantes? ¿En qué proporción? ¿Qué se hizo con estos registros?

Variable	% de datos faltantes
RatecodeID	10.96
passenger_count	9.80
store_and_fwd_flag	9.80
payment_type	9.80
congestion_surcharge	9.80
Airport_fee	9.80
VendorID	0.01

En todos los registros se completó con la moda debido al alto porcentaje representativo de la misma, excepto en payment_type, que se utilizó la moda pero para completar de forma proporcional, esto se hizo para evitar eliminar una alta cantidad de registros.

¿Se detectaron registros con pagos inconsistentes?

Sí. Se identificaron registros con valores muy alto de propina de manera porcentual respecto del valor total. Además que se pudo observar que las propinas son registradas únicamente en los pagos con tarjeta, posiblemente ya que al recibir una propina en efectivo y no registrarla, el taxista puede quedarse con el total del valor sin impuestos.

¿Se unificaron o mapearon variables categóricas codificadas numéricamente?

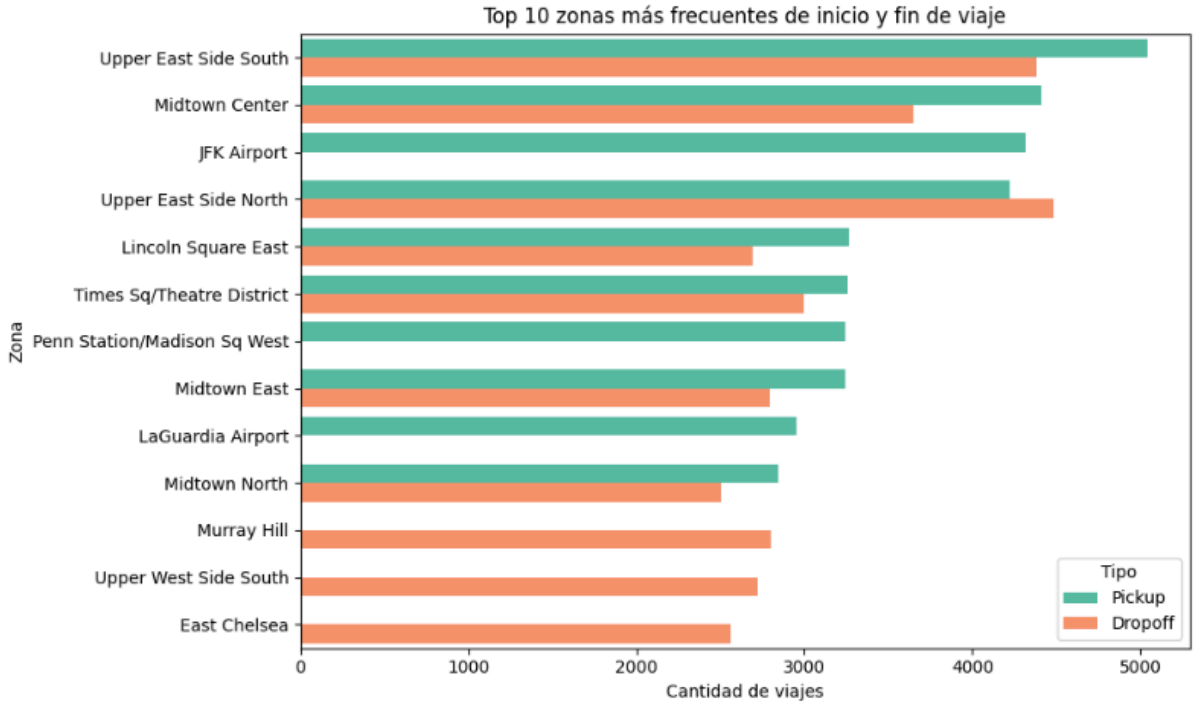
Sí. Se reemplazaron los valores numéricos en columnas como payment_type, RatecodeID y store_and_fwd_flag por sus descripciones correspondientes, utilizando los mapeos oficiales provistos por TLC.
Esto facilitó el análisis e interpretación de resultados.

¿Se validó la coherencia temporal entre pickup y dropoff?

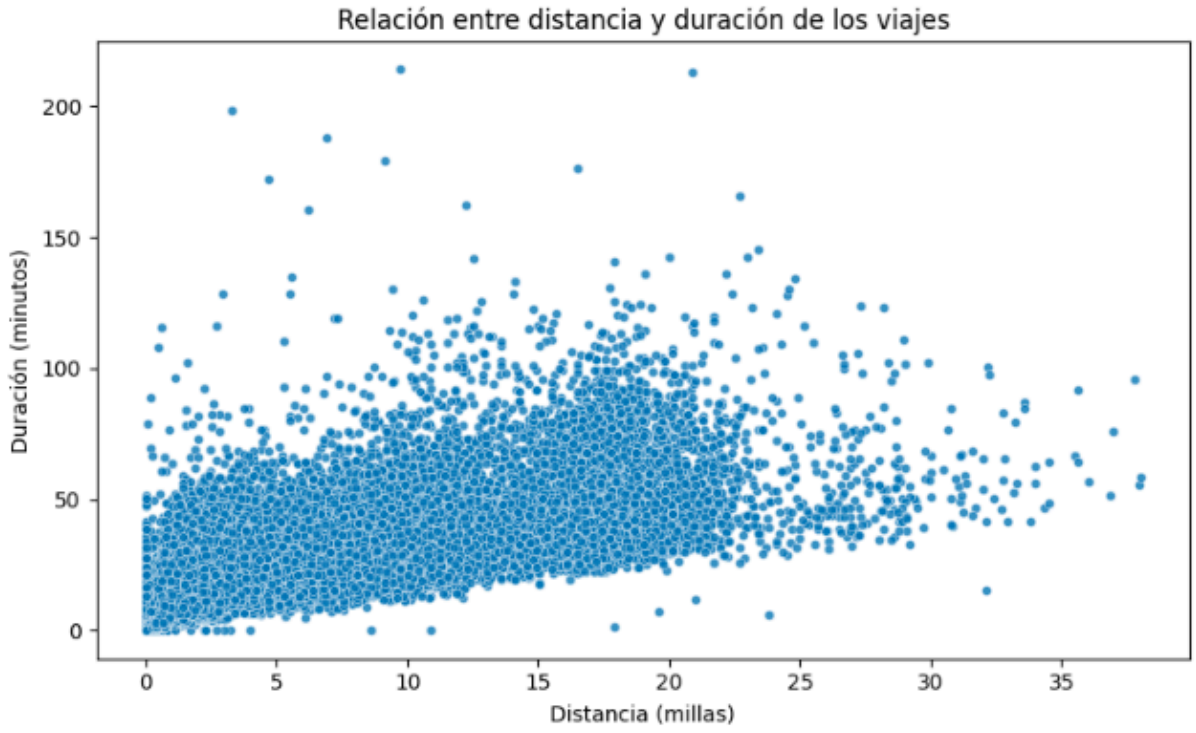
Sí. Se verificó que la fecha/hora de dropoff sea siempre posterior a la de pickup.
Todos los registros que no cumplían esta condición fueron descartados por inconsistencia temporal.

Visualizaciones

¿Cuáles son los lugares más frecuentes de inicio y fin de viaje?

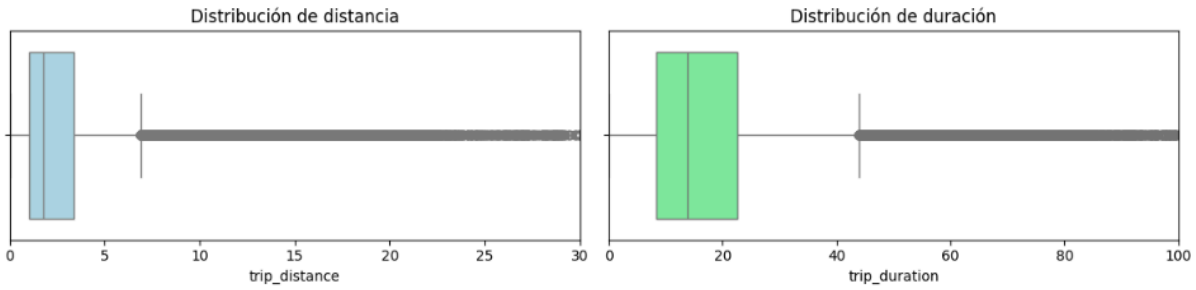


¿Como son los viajes típicamente en distancia y duración?

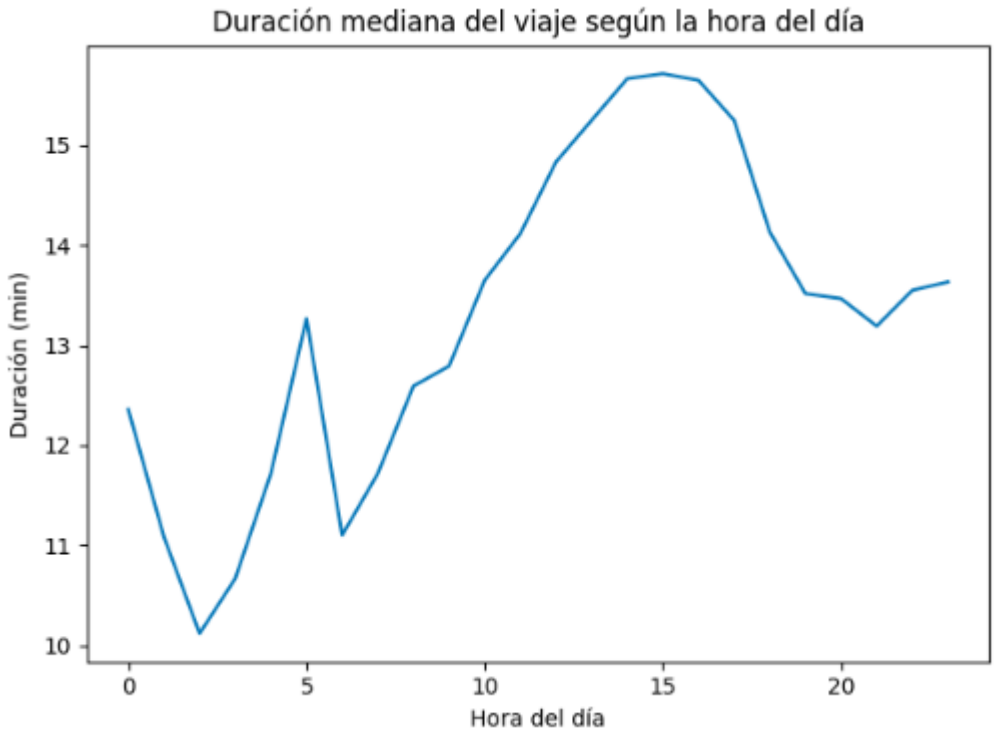


Siguen una estructura general lineal lo cual hace sentido, se observa además como sucede que para una misma distancia la duración tiene una gran variación, lo cual podría suceder por el tráfico o algunas causas externas.

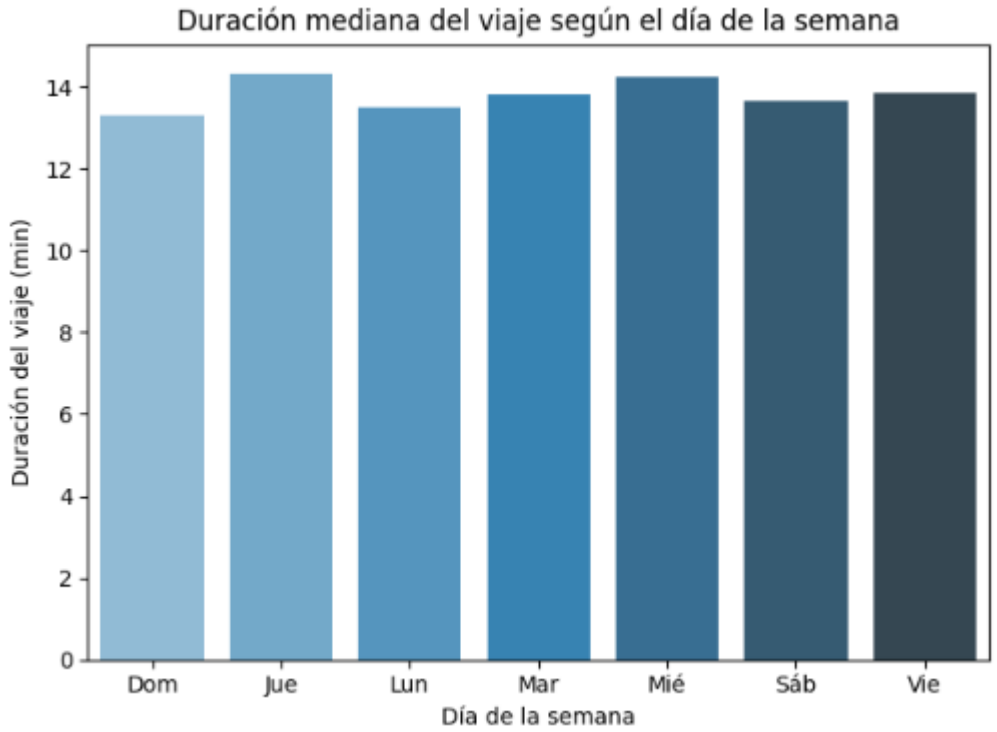
Podemos analizar ambas variables de forma independiente.



¿Cómo varía la duración según el día o la hora?

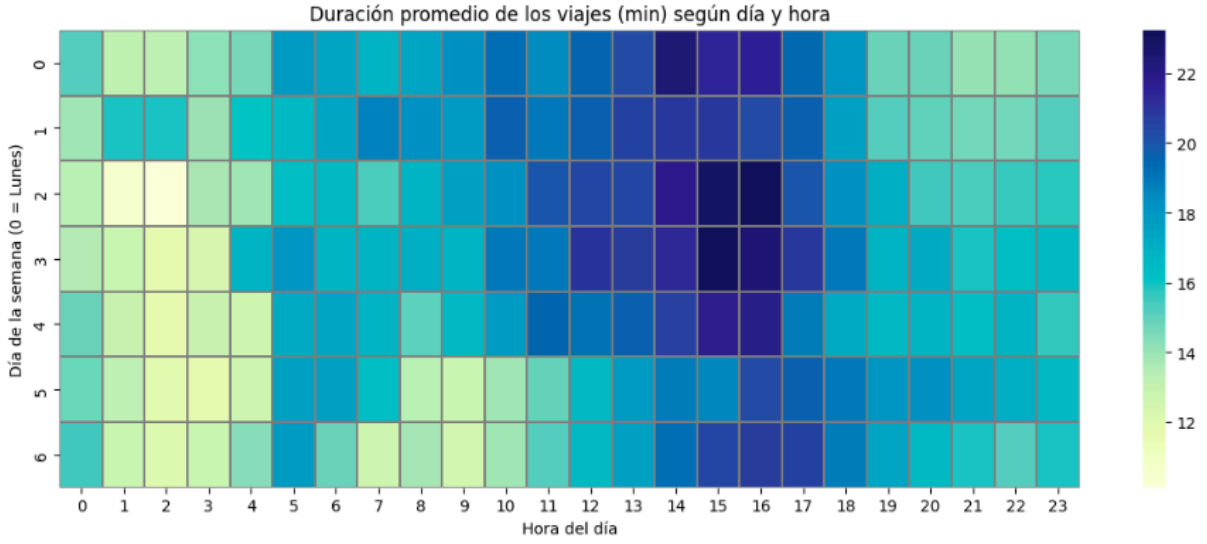


Se observa una mayor duración de los viaje por la tarde, y una duración menor durante la madrugada, aunque la diferencia no es tanta, lo cual tiene sentido ya que los viajes son en Nueva York, una ciudad con mucho movimiento a todas horas.

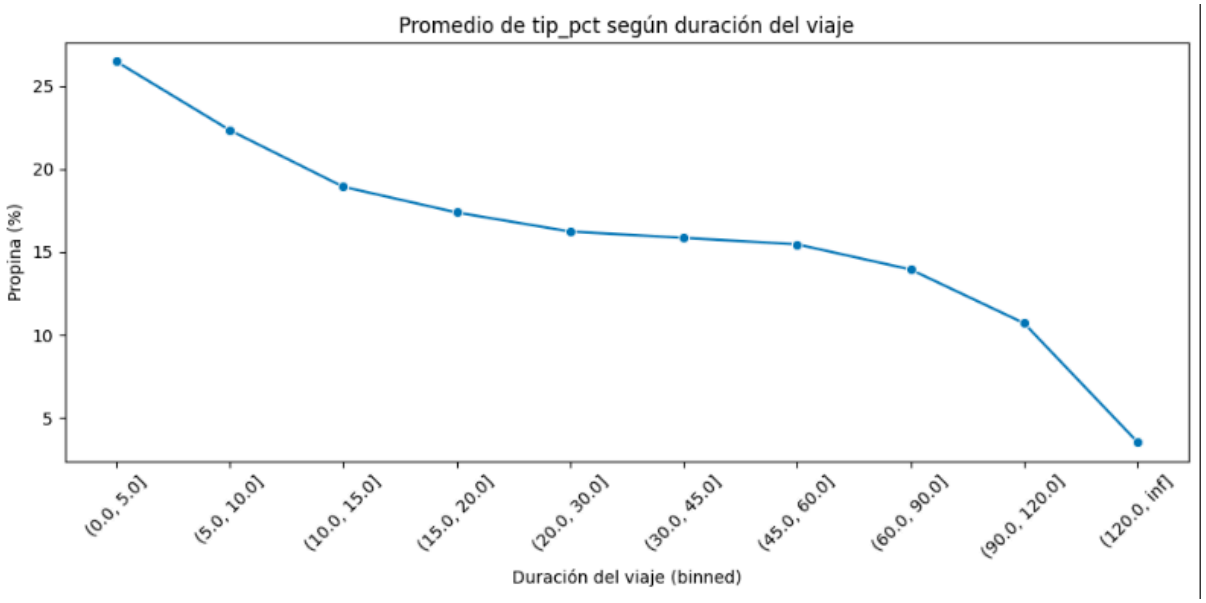


Una duración promedio casi igual para todos los días.

Combinando las variables:

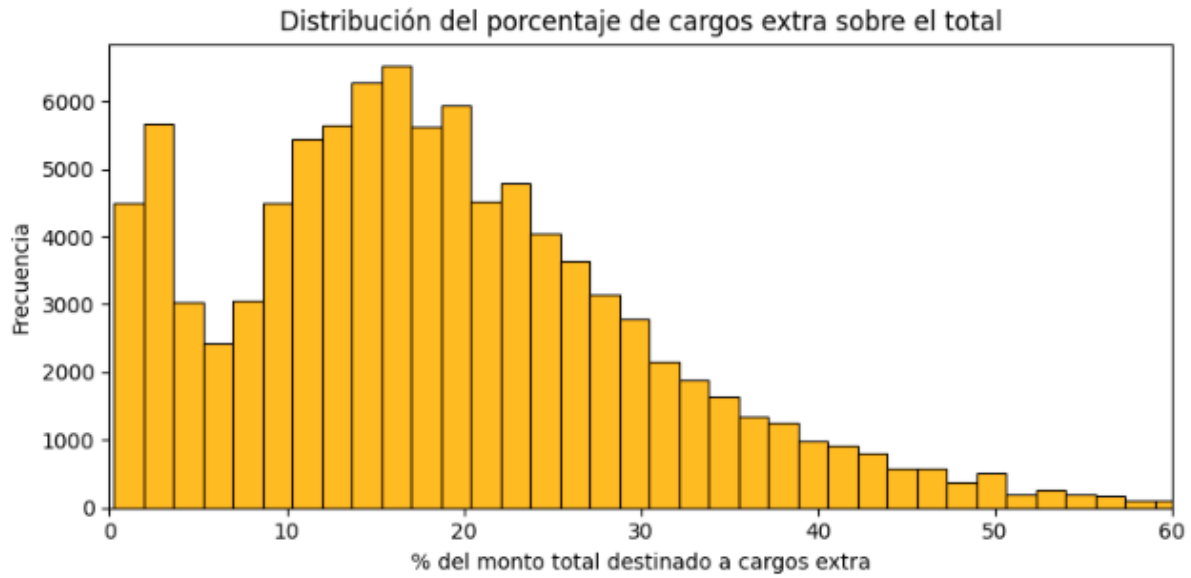


¿Los viajes más largos tienen propinas más altas?

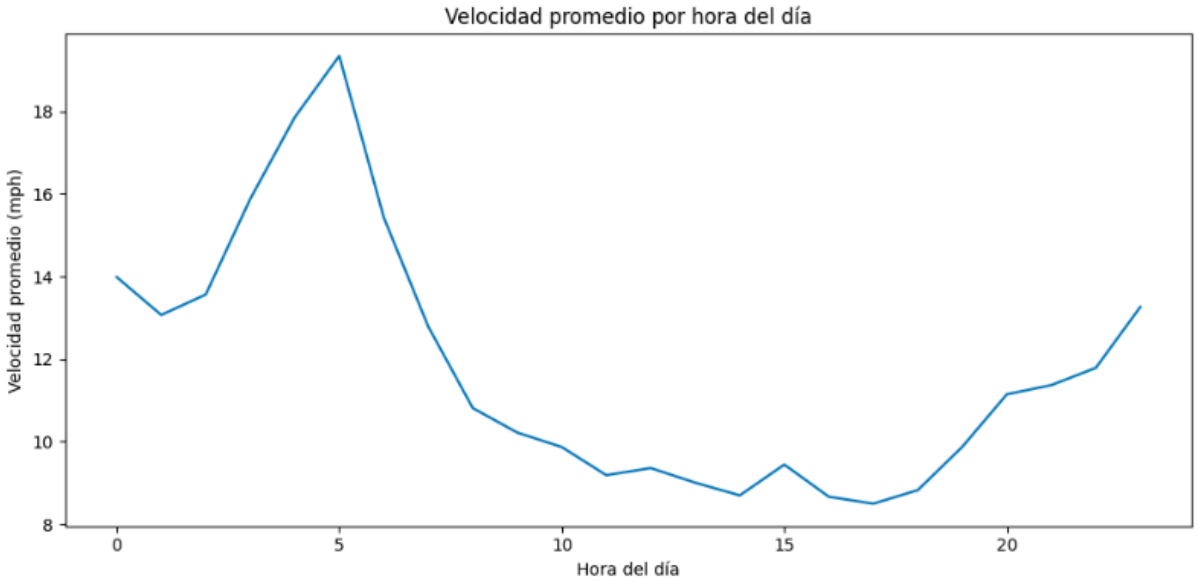


De manera porcentual podemos observar que no, pero lo que vemos en realidad es que las personas que realizan viajes cortos tienden a dejar un valor similar a las de viajes largos, por eso se ve reflejado que los viajes cortos tienen un mayor porcentaje respecto del total.

¿Qué proporción del costo total se destina a cargos extra?

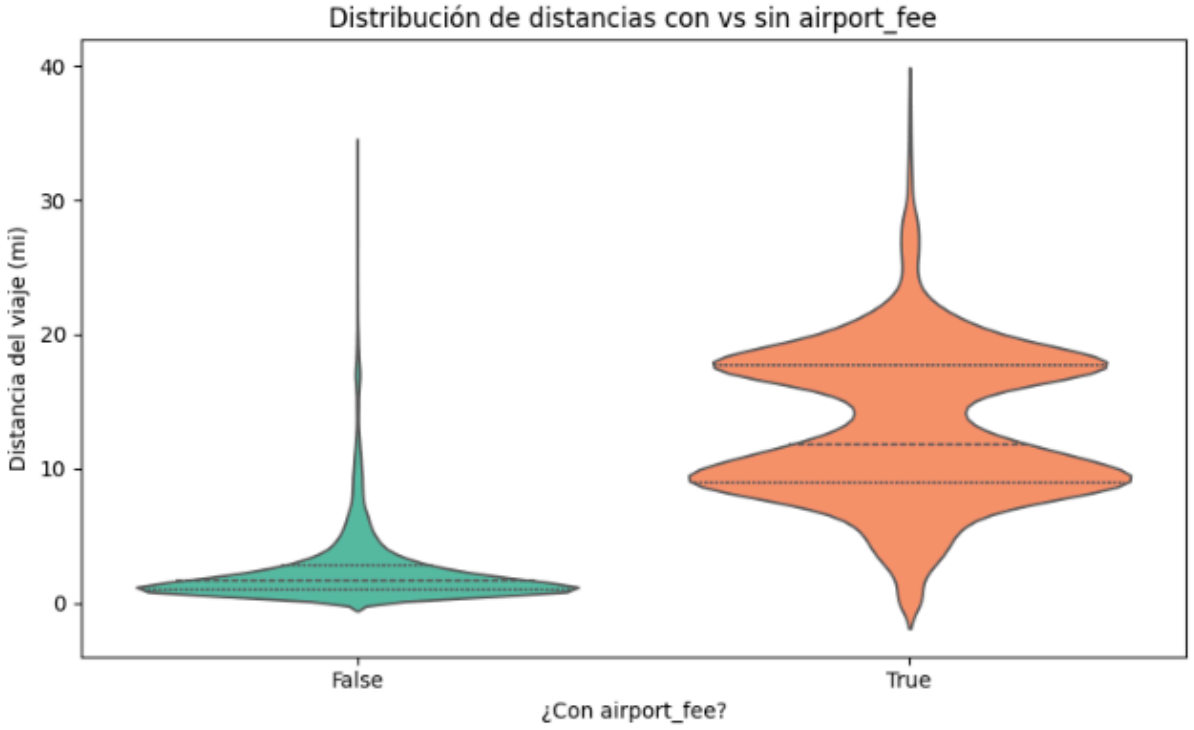


¿Cómo varía la velocidad promedio según la hora?



De forma esperable, se ve una mayor velocidad en horas de madrugada.

¿Qué tan frecuentes son los viajes de aeropuerto y en qué contextos se aplica?



EJ2: Clasificación

Descripción general del dataset

El dataset utilizado es un subset del dataset Rain in Australia. Este contiene datos meteorológicos diarios de estaciones meteorológicas de varias localidades de Australia. En particular, nos quedamos con los datos correspondientes a las locaciones de Queensland, Victoria, Australia Meridional y Australia Occidental

- Cantidad de registros: aproximadamente 76000 filas.
- Cantidad de columnas: 24

Features destacables

A continuación se describen algunas de las variables más relevantes y por qué se consideran clave en el análisis.

Variable	Tipo	Descripción
RainTomorrow (target)	int	0 o 1
Rainfall	float	Valor en mm indicando la precipitación del día.
Sunshine	float	Valor en horas indicando las horas despejadas del día.
Humidity (9am y 3pm)	float	Dos features indicando la humedad en distintos momentos del día.
Pressure (9am y 3pm)	float	Dos features indicando la presión atmosférica en hPa en distintos momentos del día.
Temperature (9am y 3pm)	float	Dos features indicando la temperatura en *C en distintos momentos del día.

Transformaciones

Features Eliminadas

Se eliminaron Locación y Región (se determinó que tenían poco poder discriminatorio) y Pressure9am, MaxTemp y Temp9am (por colinealidad con otras variables)

Además, para cada modelo, previo al tuneo de hiperparametros se eliminan las features con un valor de feature importance bajo.

Encoding:

WindDir (WindGusDir, WindDir9am y WindDir3pm): Se utilizó OHE.

Imputación de nulos

Se eliminaron los registros con valores nulos en categorías con una proporción de nulos menor al 1%.

WindDir (WindGusDir, WindDir9am y WindDir3pm): Se utilizó la media en la misma columna.

WindGustSpeed: Se inputo mediante una regresión lineal en función de WindSpeed9am y WindSpeed3am.

Pressure3pm: Se imputó a través de la media.

Cloud3pm y **Cloud9am**: Se imputaron a través de la media.

Sunshine: Se imputo como regresión lineal de Cloud3pm, Cloud9am y Humidity3pm

Evaporation: Se input usando la mediana de la columna (Debido a poseer una distribución sesgada)

Nuevas Features

Se crearon las siguientes features:

PressureDiff: Diferencia entre la presión a las 3pm y la presión a las 9am.

HumidityDiff: Diferencia entre la humedad a las 3pm y la humedad a las 9am.

WindSpeedDiff: Diferencia entre la velocidad del viento a las 3pm y la velocidad del viento a las 9am.

TempDiff: Diferencia entre la temperatura a las 3pm y la temperatura mínima del día.

PressureTemp: Producto de la presión a las 3pm y la temperatura a las 3pm.

HumidityTemp: Producto de la humedad a las 3pm y la temperatura a las 3pm.

PressureSunshineRatio: Relación entre la presión a las 3pm y la cantidad de sol (Sunshine).

Cloud3pmRainfallRatio: Producto entre la cantidad de nubes a las 3pm y la cantidad de lluvia. (Hubo un error al nombrarla)

Normalización

No se consideró necesario normalizar debido a que ninguno de los tres modelos analizados son sensibles a la escala de los features.

Outliers

Se detectaron outliers para varias de las variables, en particular Rainfall. Aun así, se decidió no quitar/capear esos valores ya que no se consideró que disminuían la performance del modelo.

Modelos

Arbol de decision

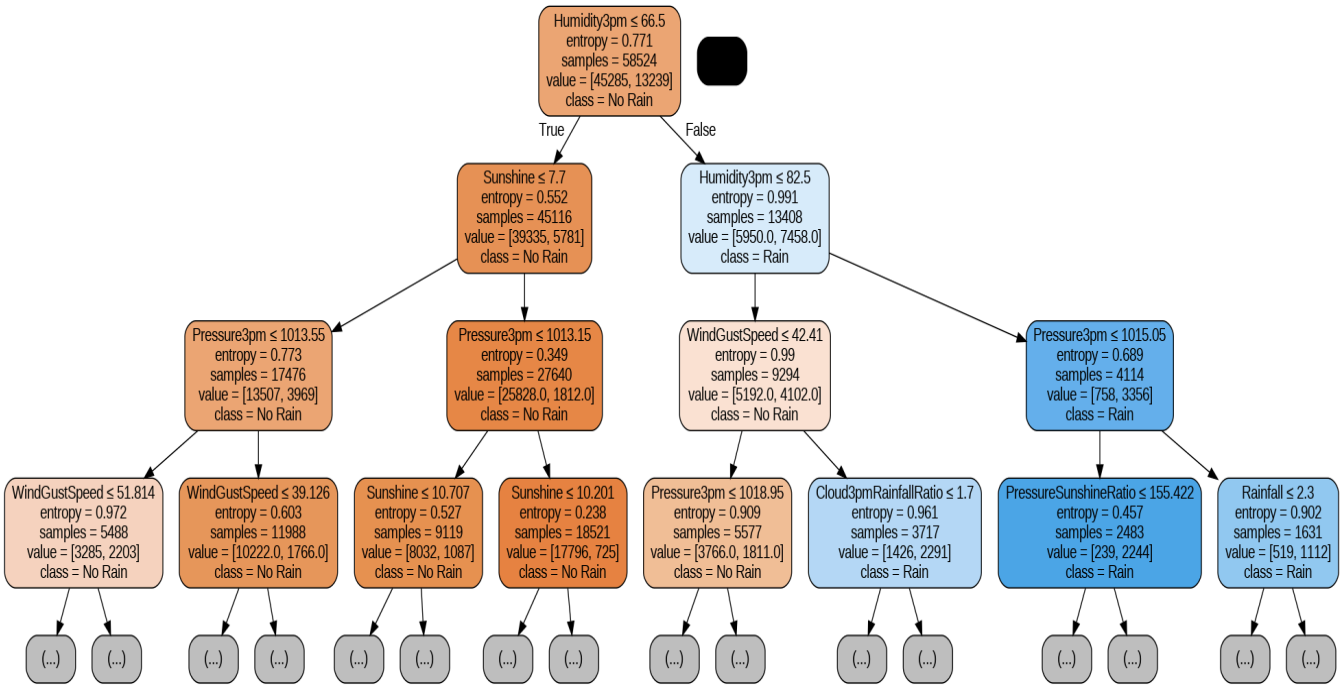
Se optimizaron los siguientes hiperparametros:

'max_depth', 'min_samples_split', 'min_samples_leaf', 'criterion', 'class_weight'

Se utilizó K-fold Cross Validation usando time split y 5 folds.

Para optimizar los parámetros, se consideró la métrica f1-score. Se eligió esa métrica teniendo en cuenta el desbalance en la variable RainTomorrow y la necesidad de lograr un balance entre falsos positivos y falsos negativos.

Se obtuvo el siguiente árbol de decisión:



La primera regla evaluada por el modelo es Humidity3pm ≤ 66.5%. Las ramas izquierdas (cumplen la condición) tienen mayor proporción de No Rain que las derechas.

En la rama izquierda del primer nodo, el siguiente criterio es Sunshine ≤ 7.7 hs. Independientemente del resultado de esto, los nodos hijos deciden según Pressure3pm ≤ 1013.15 hPa. A partir de ahí, los hijos del nodo izquierdo deciden en base a WindGustSpeed < X, y los del derecho a partir de Sunshine. Estas decisiones se corresponden con las esperadas: Baja humedad, cielo despejado y presión baja son indicadores comunes de ausencia de lluvia. En cuanto al viento, vientos fuertes podrían traer frentes secos en las regiones específicas analizadas.

Por otro lado, en la rama derecha del nodo raíz (Humidity3pm > 66.5%), el modelo evalúa vuelve a decidir en base a la misma variable (Humidity3pm ≤ 82.5%). Si la humedad es menor a 82.5%, el modelo continúa con WindGustSpeed ≤ 42.41 km/h. En la rama izquierda a esa se decide en base a Pressure3pm (≤ 1018.95 hPa). En la derecha se utiliza el producto de dos variables: Cloud3pmRainfallRatio ≤ 1.7 octas*mm. En la rama donde Humidity3pm > 82.5%, se vuelve a utilizar Pressure3pm ≤ 1015.05 hPa. A partir de ahí se utiliza el índice PressureSunshineRatio ≤ 155.422 hPa/hs a la izquierda y Rainfall a la derecha.

En síntesis se observa un alto poder discriminatorio para las variables Humidity3pm, Sunshine, Pressure3pm, WindGustSpeed y Rainfall y sus derivados.

Random Forest:

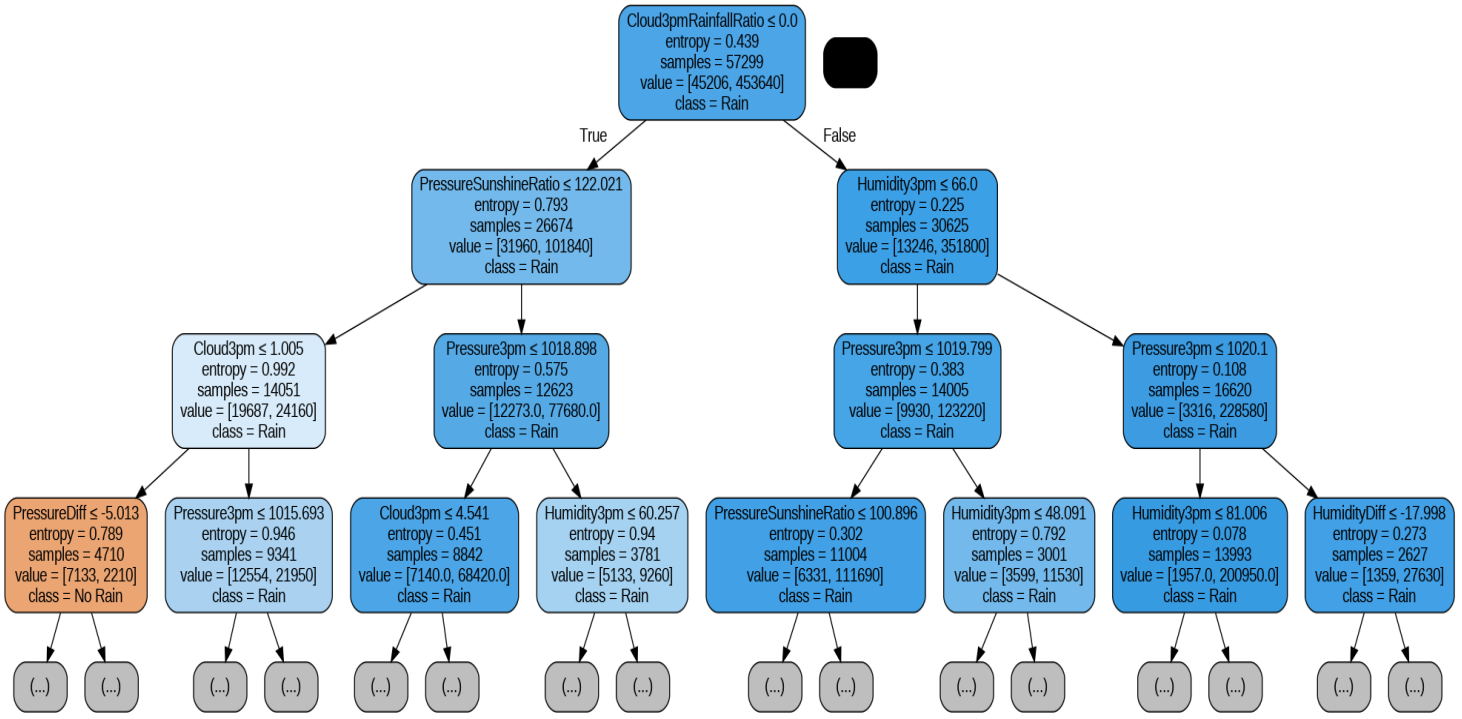
Se optimizaron los siguientes hiperparametros:

'max_depth', 'min_samples_split', 'min_samples_leaf', 'criterion', 'class_weight' "max_features", "n_estimators",

Se utilizó K-fold Cross Validation usando time split y 5 folds.

Para optimizar los parámetros, se consideró la métrica f1-score. Se eligió esa métrica teniendo en cuenta el desbalance en la variable RainTomorrow y la necesidad de lograr un balance entre falsos positivos y falsos negativos.

Se muestra el esquema de uno de los árboles generados por el modelo. Debido al parámetro weight_class = [No rain: 1, Rain :10], en la mayoría de los nodos se observa que la clase con mayor cantidad de votos es Rain.



La primera regla en ser evaluada es Cloud3pmRainfallRatio <= 0 octas * mm. Como el producto no puede ser negativo, y Cloud3pm es rara vez cero, por lo que el modelo básicamente se está preguntando si llovió o no ese dia (Rainfall = 0)

En la rama de la izquierda (Cloud3pmRainfallRatio <= 0, mayor proporción de no lluvia que en la rama derecha) la siguiente decisión es PressureSunshineRatio < 12 hPa/hs.

Si se cumple la condición, se decide en base a Cloud3pm <= 1 octas y luego en base a PressionDiff < -5 hPa a la izquierda y Pressure3pm < 1015 a la derecha. Si no se cumple, se decide en base a Pressure3pm < 1018 y luego en base a Cloud3pm 4<5.4 a la izquierda y Humidity3pm a la derecha. En resumen, baja nubosidad, presión baja o disminuyendo, y poca humedad parecen ser determinantes de que no llueva mañana.

En cuanto a la rama derecha del nodo raíz, el modelo decide en base a Humidity 3pm < 66%. Sea cual sea ese valor, se vuelve a separar en base a Pressure3pm < 1020hPa. En los

cuatro nodos hijos de esos dos se decide en base a PressureSunshineRatio, Humidity3pm y HumidityDiff.

Se observa que las mismas variables tienden a aparecer varias veces y con el mismo criterio.

Hist Gradient Boosting Classifier:

Se optimizaron los siguientes hiperparametros:

'max_depth', 'min_samples_split', 'min_samples_leaf', 'criterion', 'class_weight', ‘learning_rate’, ‘max_iter’, ‘l2_regularization’, ‘model__max_bins’

Se utilizó K-fold Cross Validation usando time split y 5 folds.

Para optimizar los parámetros, se consideró la métrica f1-score. Se eligió esa métrica teniendo en cuenta el desbalance en la variable RainTomorrow y la necesidad de lograr un balance entre falsos positivos y falsos negativos.

Cuadro de resultados

	F1-Test	Precision Test	Recall Test	Accuracy Test	ROC- AUC
Arbol de decision	0.5923	0.6614	0.5363	0.8373	0.8326
Random Forest	0.6209	0.6371	0.6055	0.8371	0.8536
HistGradient Boosting	0.6450	0.6142	0.6790	0.8353	0.8746

Arbol de decision:
- min_samples_split: 20
- min_samples_leaf: 2
- max_depth: 10
- criterion: 'gini'
- class_weight: None

Random Forest:
- n_estimators: 10
- min_samples_split: 2
- min_samples_leaf: 2
- max_features: None
- max_depth: None
- criterion: 'entropy'

class_weight: {1: 10}

HistGradient Boosting:
min_samples_leaf: 5
max_leaf_nodes: 15
max_iter: 100
max_depth: 7
max_bins: 100
learning_rate: 0.1
l2_regularization: 10.0

En el caso del árbol de decisión, se obtuvo un f1-score: 0.60 en el fold con el mejor score y f1-score 0.68 al entrenar con el set de train entero. En el set de test se obtuvo 0.59.

En el caso del Random Forest, se obtuvo un f1-score: 0.63 en el fold con el mejor score y f1-score 0.97 al entrenar con el set de train entero. En el set de test se obtuvo 0.62.

En el caso del HistGradient Boosting, se obtuvo un f1-score: 0.66 en el fold con el mejor score y f1-score 0.68 al entrenar con el set de train entero. En el set de test se obtuvo 0.64.

Estos resultados parecen indicar que en el caso del Random Forest hay overfitting del modelo a los datos de entrenamiento, a diferencia de los otros dos modelos. El árbol de decisión podría no tener la suficiente capacidad de expresión como para sobre ajustarse a estos datos (pero tampoco logra generalizar bien), mientras que el HistGradientBoosting logra que no haya overfitting pero a su vez consigue predecir mejor en el set de test.

Elección de modelo

Se elige el modelo HistGradient Boosting ya que es el que obtuvo las mejores métricas en f1-score (balance entre precisión y recall) y ROC-AUC

EJ3: Regresión

Descripción del Dataset

El dataset contiene información de propiedades de AirBnB correspondientes a Hawaii. Inicialmente poseía 36125 registros y 79 columnas. El mismo fue limpiado hasta quedar con 36125 registros y 23 columnas. Los features más relevantes son los siguientes:

- bedrooms: número de habitaciones (int).
- beds: cantidad de camas que tiene el alojamiento(int)
- bathrooms: número de baños(int).
- accommodates: cantidad de personas máximas que puede alojar la propiedad(int).
- room_type: tipo de habitación (string). Se le aplicó one-hot encoding para transformarlo en una variable categórica (4 variables dummies).
- neighbourhood_cleansed: barrio o zona del alojamiento(string). Se le realizó una clusterización para agrupar los barrios en relación de la cantidad de oferta de alojamientos y las características en los precios.
- neighbourhood_cleansed_group: Ciudad del alojamiento(string). Se le aplicó one-hot encoding para transformarlo en una variable categórica (4 variables dummies.
- máximum_nights: cantidad de noches máximas que se puede reservar el alojamiento. Se escaló la variable ya que poseía una escala muy variable.
- mínimum_nights: cantidad de noches mínimas que se puede reservar el alojamiento. Se escaló la variable ya que poseía una gran cantidad de outliers.
- Precio: valor del alojamiento (int) (variable a predecir).

Modelos

Regresión Lineal

Features seleccionados:

'accommodates', 'bathrooms', 'bedrooms', 'host_is_superhost', 'has_license', 'has_description', 'minimum_nights_scaled', ‘is_instant_bookable', 'is_bathroom_shared', 'room_type_Hotel room', ‘room_type_Private room', 'room_type_Shared room', 'neighbourhood_group_cleansed_Honolulu', 'neighbourhood_group_cleansed_Kauai', 'neighbourhood_group_cleansed_Maui', 'maximum_nights_scaled', 'neighbourhood_group_1', 'neighbourhood_group_2', 'neighbourhood_group_3

	MSE	RMSE	R^2
Train	83825.5	289.53	0.4345
Test	83584.9	289.11	0.4534

R 2 es el coeficiente de determinación y nos indica que tan bien entiende la variabilidad del problema nuestro modelo, que tan bueno es nuestro modelo, y tiene un rango de valores de 0 a 1. En esté caso la explica en un 43/45% aproximadamente lo que nos indica que no es ni un modelo bueno ni malo.

Podemos ver que los valores entre el conjunto de training y test son casi iguales, los que no lleva a concluir que el modelo generaliza bien el problema y no produce ni overfitting ni underfitting.

MSE (Error Cuadrático Medio) representa la distancia entre los valores reales a los predichos al cuadrado. Es una métrica que penaliza altamente los errores del modelo. En este caso es un valor elevado lo que nos marca que nuestro modelo esta cometiendo muchos errores, probablemente por presencia de outliers que suelen perjudicar más a los modelos de regresión lineal.

RMSE(raíz del Error Cuadrático Medio) nos indica la distancia entre los valores reales y predichos en unidades de la variable a predecir. En este caso nos marca que la media de error es de 289 un valor que puede llegar a ser razonable teniendo en cuenta que hay varios valores altos en el dataset.

Viendo los gráficos podemos decir que el modelo tiende a fallar con los valores altos y estos son los que están perjudicando nuestro modelo.

XG Boost

Se utilizó K-fold Cross Validation usando 5 folds.

Para buscar los parámetros se utilizó MSE(mean squared error) para minimizar el error y se consideró necesario ya que los precios del dataset son muy variados.

	MSE	RMSE	R^2
Train	19825.7	140.80	0.86
Test	54742.2	230.06	0.65

La diferencia entre los valores entre el conjunto de train y test nos indica que nuestro modelo tiende un poco al overfitting pero no parece ser tan grave. Por lo tanto, se puede decir que generaliza peor que el modelo de regresión lineal.

El coeficiente de determinación nos indica que esté modelo entiende mejor la variabilidad de los datos ya que esté caso la explica en un 86% para el conjunto de train y un 65% para el conjunto de test.

El MSE se redujo en esté modelo, lo cual indica que comete menos errores y se ve menos afectado por los outliers.

RMSE nos está indicando que el error promedio está en $. Se puede ver que el modelo también reduce el error.

Viendo el gráfico podemos notar que en nuestro conjunto de entrenamiento logra una predicción bastante uniforme. Al contrario de el conjunto de prueba que a medida que los valores aumentan mayor es el error. Y también podemos ver igual que en el gráfico de regresión lineal que comienza a fallar más cuanto más altos son los precios.

Random Forest

Se utilizó K-fold Cross Validation usando 3 folds.

Para buscar los parámetros se utilizó MSE(mean squared error) para minimizar el error y se consideró necesario ya que los precios del dataset son muy variados.

	MSE	RMSE	R^2
Train	18467.6	135.8	0.87
Test	54742.2	233.9	0.64

Random Forest también presenta una diferencia entre el conjunto de train y test. Indicando cierto grado de overfitting.

Al igual que XP Boost presenta un buen rendimiento en el conjunto de de training, explicando un 87% la variabilidad. Y un rendimiento aceptable en el conjunto de test con un 64%.

Al igual que XP Boost, Random Forest se ve menos perjudicado por la presencia de outliers y que su error promedio es menor como nos indican sus valores de MSE y RMSE respectivamente.

Viendo el gráfico podemos concluir lo mismo que en los modelos anteriores. Todos tiene problemas para predecir los precios altos.

Cuadro de Resultado

Modelo	MSE	RMSE	R^2
Regresión Lineal	83584.9	289.11	0.4534
XG Boost	124.86	230.06	0.65
Random Forest	54742.2	233.9	0.64

Regresión Lineal es un método para predecir valores en rango continuos. Estimando una relación lineal entre una variable dependiente (variable a predecir) y una o más variables independientes (predictoras).

XG Boost es un algoritmo de aprendizaje supervisado basado en árboles de decisión que utiliza boosting, una técnica de ensamble que entrena modelos secuencialmente para corregir los errores del anterior

Random Forest es un modelo de ensamble que construye múltiples árboles de decisión y promedia sus predicciones. Utiliza bagging (bootstrap aggregating) para reducir la varianza y mejorar la generalización.

Modelo elegido: XG Boost

Aunque Random Forest y XG Boost tienen rendimientos muy similares, el modelo elegido por nosotros va a ser XG Boost. Ya que más allá de esa mínima mejora en el rendimiento la gran diferencia está marcada por la diferencia en la complejidad computacional y la diferencia en los tiempos de entrenamiento de ambos modelos. Siendo XG Boost ampliamente superior en estos campos.

EJ4: Clustering

Tendencia al clustering:

Primera observación: tendencia al clustering por encima del umbral 0.75. Se considera que hay tendencia.

Resultado:
Hopkins Statistic: 0.7862017846082693

Cantidad apropiada de grupos

Cantidad de grupos que podrían formarse: 7.

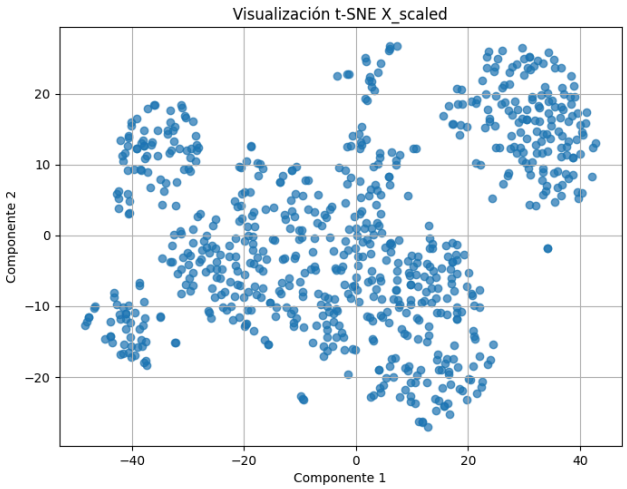
1. Se consideró, además, un grupo de targets relevantes a la hora del análisis. Este último grupo con una tendencia al clustering por encima de 0.75 también.

   target columns = 'acousticness','danceability','energy','instrumentalness', 'liveness','loudness','speechiness','tempo','valence'

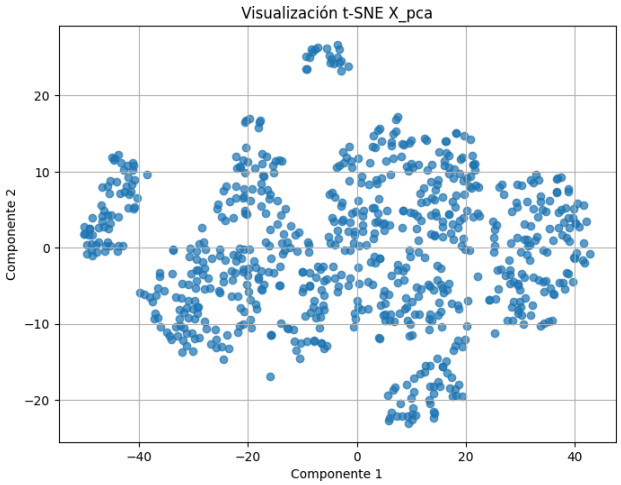
   Hopkins Statistic: 0.7865562436066339

2. Se realizó una reducción con PCA buscando reducir varianza y covarianza. Se consideran ambos grupos (dataset original y target group).

3. Se analizaron con t-SNe los datos (previamente escalados). Se observa una mayor dispersión en el gráfico correspondiente a los targets. Se decide no seguir con ese grupo.

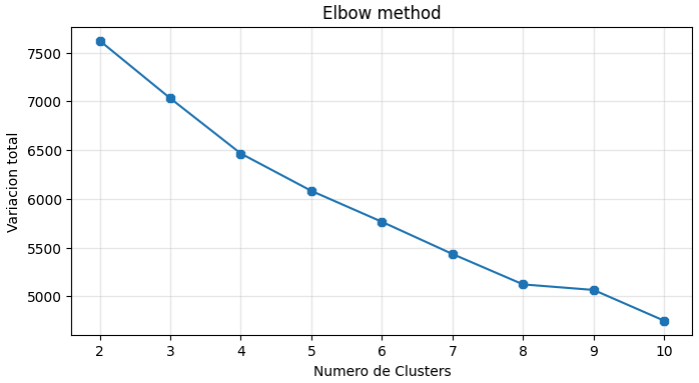


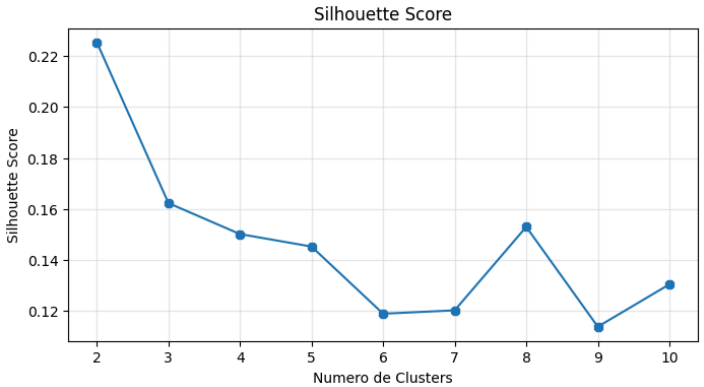
Data set original (PCA aplicado)



Dataset targets (PCA aplicado)

4. Se aplica elbow method y silhouette para analizar cantidad de cluster. Se decide testear el dataset con y sin el PCA aplicado. Se encontró claridad en el grupo con el dataset con PCA, el otro grupo no mostraba una tendencia clara.





Dataset original (PCA aplicado)

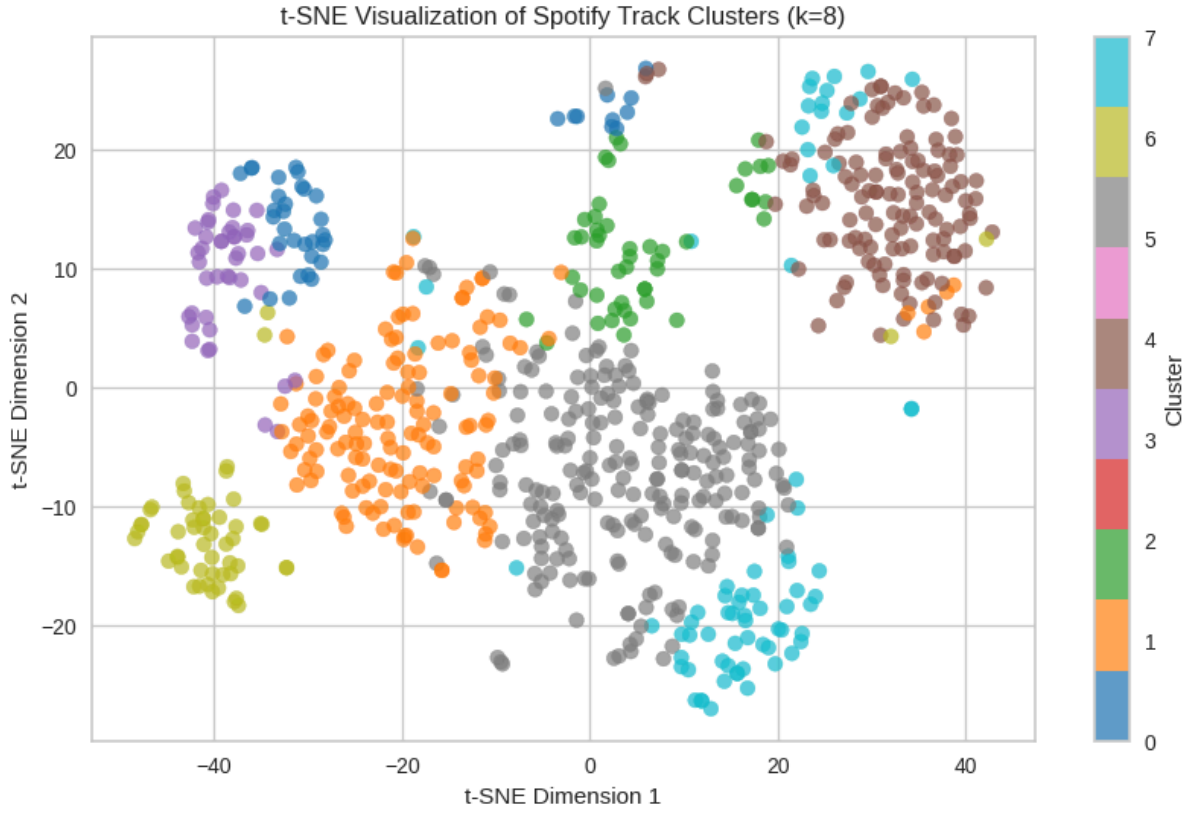
Gran coincidencia en 8.

SilhouetteVisualizer me indica una cierta estabilidad con k entre 7 y 9

Por útlimo, de determina utilizar K-Means con k = 8.

Visualización de lo grupos

Se observa ruido y ciertas dispersiones en el gráfico final.



Descripción de los grupos.

Cluster 0:
- Características destacadas: Alta instrumentalidad (0.79), moderada acústica (0.54)
- Energía moderada (0.55) y bailabilidad media (0.57)
- Tempo relativamente alto (126.58 BPM)

Cluster 1:

- Características destacadas: Alta acústica (0.61), baja energía (0.38), baja instrumentalidad (0.01)
- Volumen bajo (-10.78 dB) y valencia emocional moderada-baja (0.40)
- Tempo medio (117.98 BPM)

Cluster 2:

- Características destacadas: Alta energía (0.73), nivel extremadamente alto de liveness (0.74)
- Baja acústica (0.23), sugiriendo un sonido más procesado/eléctrico
- Mayor volumen (-7.28 dB) y valencia emocional positiva (0.52)

Cluster 3:

- Características destacadas: Alta acústica (0.95), mínima energía (0.11)
- Alta instrumentalidad (0.70) y duración extendida (aprox 4:55 min)
- Volumen muy bajo (-22.14 dB) y valencia emocional baja (0.22)
- Tempo lento (96.23 BPM)

Cluster 4:

- Características destacadas: Alta energía (0.72), buena bailabilidad (0.65)
- Muy baja acústica (0.16) e instrumentalidad (0.03)
- Único cluster en modo completamente menor (0.00)
- Volumen alto (-6.10 dB)

Cluster 5:

- Características destacadas: Alta energía (0.77), máxima bailabilidad (0.68)
- Mínima acústica (0.15), completamente en modo mayor (1.00)
- Mayor valencia emocional (0.62), indicando un carácter positivo
- Volumen más alto (-5.64 dB)

Cluster 6:

- Características destacadas: Time signature marcadamente bajo (2.77)
- Moderada-alta acústica (0.68), baja energía (0.34)
- Baja bailabilidad (0.45) y valencia emocional baja (0.37)

Cluster 7:

- Características destacadas: Speechiness alto (0.35)
- Alta bailabilidad (0.69), energía considerable (0.65)
- Tempo más rápido (137.16 BPM)
- Baja acústica (0.25) y duración más corta (3:23 min)

Tiempo dedicado

Integrante	Tarea	Prom. Hs Semana
Agustin Braida	Punto 2	4
Ignacio Carrera	Punto 1	4
Matias Etchegoyen	Punto 3	4
Agustin Trombetta	Análisis valores faltantes Punto 4	4