Proyecto Integrador

Módulo 3: Python for Data Analytics

Análisis de Project RideFare

Sector: Transporte / Economía Digital

Integrantes:

OLALLA SACANCELA MIGUEL SEBASTIÁN

PONCE PINCAY JHON LENIN

TERÁN GARCÍA LUIS MATTEO

TULCAN ALVAREZ JOSE DAVID

VELIZ RIVERA FELIX JAIR

Curso:

Paralelo - 04

Coding Bootcamps - MINTEL

Programa Data-Driven-Decision Specialist

Profesor:

Eduardo Cruz, PhD

13 de noviembre de 2025

Índice

1. Project RideFare

El proyecto RideFare tiene como objetivo analizar el comportamiento de los precios y la demanda de servicios de transporte por aplicación —como Uber y Lyft— considerando factores geográficos, temporales y climáticos.

Para ello, se emplean dos fuentes de datos principales:

Dataset de viajes (`rides.csv`)

Contiene información detallada de los desplazamientos realizados a través de plataformas digitales, incluyendo variables como distancia, tipo de servicio, origen, destino, tarifa aplicada y momento del viaje. Estos datos permiten estudiar cómo varía el precio o la disponibilidad del servicio según la zona o la hora del día.

Dataset meteorológico (`weather.csv`)

Registra las condiciones del clima en diferentes zonas de la ciudad, con variables como temperatura, presión, lluvia, humedad, nubosidad y velocidad del viento. Este conjunto de datos resulta esencial para analizar el impacto de las condiciones meteorológicas sobre la demanda y el precio de los viajes.

El estudio combina técnicas de análisis de datos, visualización y modelado estadístico utilizando herramientas de Python (pandas, matplotlib, seaborn), con el fin de identificar patrones y tendencias que ayuden a comprender mejor el funcionamiento de la economía digital del transporte urbano.

2. Objetivo General

Analizar la relación entre los precios y la demanda de servicios de transporte digital (Uber y Lyft) y las condiciones climáticas, geográficas y temporales, mediante el uso de técnicas de

análisis de datos y visualización, con el fin de identificar patrones y factores que influyen en la variación de las tarifas y la actividad de los viajes.

3. Objetivo Específico

- Integrar y depurar los datasets de viajes (`rides.csv`) y condiciones meteorológicas (`weather.csv`), asegurando la correcta correspondencia entre las variables temporales y geográficas.

- Explorar y describir las características principales de los datos, identificando la distribución de precios, demanda y condiciones del clima por zonas y momentos del día.

- Analizar la relación entre las variables climáticas (temperatura, humedad y nubosidad) mediante el cálculo del coeficiente de correlación, para determinar el grado de asociación entre ellas.

- Identificar y tratar los valores atípicos de temperatura superiores a 45 °C, con el fin de mejorar la calidad de los datos y reducir el impacto de valores extremos en los resultados del análisis.

- Visualizar las correlaciones entre las variables meteorológicas y la cantidad de viajes a través de un mapa de calor (heatmap), con el propósito de identificar si existen relaciones significativas entre el clima y el comportamiento de los viajes.

- Examinar la evolución temporal de los viajes y las condiciones climáticas promedio, para observar tendencias, variaciones y posibles patrones a lo largo del tiempo.

- Interpretar los resultados obtenidos de las correlaciones y las gráficas, con el fin de determinar si las condiciones climáticas influyen o no de manera significativa en la cantidad de viajes realizados.

4. Fases del proyecto

4.1 Importación de librerías

En el desarrollo del proyecto RideFare, se utilizan diversas librerías de Python que permiten realizar el procesamiento, análisis y visualización de los datos obtenidos de los archivos *rides.csv* (viajes) y *weather.csv* (condiciones meteorológicas).

Las principales librerías empleadas son las siguientes:

- **Pandas:** se utiliza para la manipulación y análisis de datos. Permite cargar, limpiar, combinar y transformar los datasets de viajes y clima, facilitando la organización de la información en estructuras tipo DataFrame.

- **NumPy:** se aplica para realizar operaciones matemáticas y estadísticas sobre los datos numéricos, como el cálculo de promedios, correlaciones o desviaciones estándar.

- **Matplotlib:** sirve para la creación de gráficos y visualizaciones básicas, permitiendo representar la variación de los precios, la demanda o las condiciones meteorológicas de forma clara.

- **Seaborn:** complementa a Matplotlib con gráficos más elaborados y estéticamente atractivos, útiles para explorar relaciones entre variables y mostrar patrones o tendencias de manera visual.

Estas librerías en conjunto proporcionan las herramientas necesarias para realizar un análisis integral de los datos, desde su limpieza hasta la visualización de los resultados, contribuyendo al cumplimiento de los objetivos del proyecto.

4.2 Inicialización de datos

En esta etapa del proyecto RideFare, se procede a la carga e inspección inicial de los conjuntos de datos que servirán como base para el análisis. Estos datasets contienen

información sobre los viajes realizados mediante plataformas de transporte digital (Uber y Lyft) y las condiciones meteorológicas registradas en diferentes zonas y momentos del día.

Los principales archivos utilizados son:

- **rides.csv:** incluye los datos de los desplazamientos, con variables como tipo de servicio (Uber o Lyft), punto de origen, destino, distancia recorrida, tarifa, fecha y hora del viaje. Este conjunto de información permite estudiar la dinámica del precio y la demanda de los servicios de transporte en función del tiempo y la ubicación.

- **weather.csv:** contiene los registros climáticos asociados a distintas zonas geográficas, con variables como temperatura, humedad, presión, velocidad del viento, lluvia y nubosidad. Este archivo es esencial para evaluar la influencia del clima sobre el comportamiento del precio y la frecuencia de los viajes.

Durante la inicialización de los datos, se realiza la carga de ambos archivos en estructuras que permitan su manejo y análisis, asegurando que los tipos de datos (numéricos, categóricos, temporales) sean correctamente interpretados. Asimismo, se lleva a cabo una revisión preliminar para verificar el número de registros, las columnas disponibles y la correspondencia entre los datasets según el tiempo y la ubicación.

Esta fase garantiza que la información esté correctamente preparada para los procesos posteriores de integración, depuración, análisis exploratorio y visualización.

4.3 Exploración inicial de los datos

Una vez cargados los conjuntos de datos **rides_original** y **weather_original**, se realiza una exploración inicial con el fin de conocer su estructura, tipo de información y las variables disponibles para el análisis. Esta etapa permite familiarizarse con el contenido de los archivos y detectar posibles problemas de formato, valores faltantes o inconsistencias.

4.3.1 Vista rápida

Para obtener una primera impresión de los datos, se utiliza el método .head(), el cual muestra las primeras filas de cada dataset. Esto facilita una revisión rápida de las columnas, sus nombres y el tipo de información que contienen.

En el caso del dataset **rides_original**, se observa que incluye información detallada sobre los viajes realizados a través de plataformas digitales. Las variables principales son:

- **distance:** distancia recorrida en cada viaje.
- **cab_type:** tipo de servicio utilizado (por ejemplo, Lyft, Lyft Lux, Lyft XL).
- **time_stamp:** marca de tiempo en la que se registró el viaje.
- **destination** y **source:** zonas de destino y origen del desplazamiento.
- **price:** tarifa total aplicada al viaje.
- **surge_multiplier:** factor de incremento del precio en momentos de alta demanda.
- **id** y **product_id:** identificadores únicos de cada registro.
- **name:** nombre comercial del servicio ofrecido.

Por otro lado, el dataset **weather_original** presenta las condiciones meteorológicas registradas en distintas ubicaciones de la ciudad. Las variables principales son:

- **temp:** temperatura medida en grados Fahrenheit.
- **location:** nombre del sector o zona donde se registró el clima.
- **clouds:** nivel de nubosidad.
- **pressure:** presión atmosférica en hectopascales.
- **rain:** nivel de precipitación.

- **time_stamp:** marca temporal que permite asociar el registro climático con el momento del viaje.

- **humidity:** porcentaje de humedad en el ambiente.

- **wind:** velocidad del viento.

Esta vista preliminar confirma que ambos datasets están estructurados de manera tabular, comparten la variable **time_stamp** y pueden ser vinculados posteriormente para analizar la relación entre el clima y las tarifas de transporte. Asimismo, la exploración inicial permite identificar que los datos son adecuados para continuar con las fases de limpieza, integración y análisis estadístico.

4.3.2 Información general

A partir del análisis del resultado de **`rides_original.info()`**, se obtiene la siguiente información general sobre el dataset de viajes:

- El dataframe `rides_original` contiene un total de 693.071 registros, cada uno correspondiente a un viaje realizado mediante plataformas digitales como Uber o Lyft.

- El conjunto de datos está compuesto por 10 columnas, que incluyen variables tanto numéricas como categóricas relacionadas con el tipo de servicio, origen, destino, precio y momento del viaje.

- La columna `time_stamp` almacena la información temporal de cada registro, sin embargo, se encuentra en formato numérico (epoch). Por este motivo, será necesario convertirla a un formato de fecha y hora legible para poder realizar análisis basados en el tiempo, como la variación de precios o la demanda por hora y día.

El dataset **`weather_original`** contiene **6.276 registros y 8 columnas** que describen las condiciones meteorológicas en diferentes zonas y momentos del día.

Las principales variables son:

- **temp:** temperatura registrada.
- **location:** zona geográfica.
- **clouds, pressure, humidity, wind:** indicadores del estado del clima.
- **rain:** nivel de precipitación, con muchos valores faltantes (solo 894 no nulos).
- **time_stamp:** marca temporal en formato numérico que debe convertirse a fecha y hora legible.

El conjunto de datos presenta una estructura limpia, con la mayoría de las variables completas y un tamaño manejable (392.4 KB). Solo requiere ajustes en el formato temporal y tratamiento de los valores faltantes en la columna *rain*.

4.3.3 Transformación inicial del tiempo para el Dataset Rides

Para comenzar el proceso de análisis y limpieza, se crea una copia del dataset original denominada `rides_exploratorio`. Esta copia permite realizar transformaciones y pruebas sin alterar los datos originales, manteniendo así la integridad del archivo fuente.

Uno de los primeros pasos consiste en formatear la columna `time_stamp`, que inicialmente se encontraba en formato numérico (milisegundos desde la época Unix). Mediante su conversión a formato de fecha y hora legible, es posible analizar la información temporal con mayor precisión, identificando patrones por día, hora o periodo.

Esta transformación es fundamental para los análisis posteriores, ya que permitirá relacionar los precios y la demanda de los viajes con las variables temporales y meteorológicas de forma coherente.

4.3.4 Transformación inicial del tiempo para el Dataset Weather

Se creó una copia del conjunto de datos original weather_original bajo el nombre weather_exploratorio, con el propósito de preservar la integridad de los datos originales y trabajar sobre una versión destinada al análisis. En esta nueva copia, la columna time_stamp fue convertida al formato de fecha y hora (datetime), ya que inicialmente se encontraba expresada como una marca temporal numérica en segundos. Esta transformación permite sincronizar y relacionar los registros meteorológicos con los datos de viajes según el momento exacto en que ocurrieron.

Adicionalmente, se generó una nueva columna denominada temp_c, que representa la temperatura en grados Celsius, calculada a partir de la conversión de las temperaturas originalmente registradas en grados Fahrenheit. Esta conversión facilita la interpretación y comparación de los datos climáticos, alineándolos con las unidades comúnmente utilizadas en análisis científicos y contextos locales.

4.3.5 Estadísticas descriptivas

- **Rides**

  El conjunto de datos **rides** contiene un total de 693,071 registros y 10 variables. A partir del análisis estadístico descriptivo se obtienen los siguientes resultados principales:

  - **Distancia:** Los viajes tienen una distancia promedio de 2.19 millas, con un rango entre 0.02 y 7.86 millas, lo que sugiere que la mayoría de los trayectos son de corta duración.

  - **Precio:** El precio promedio de los viajes es de $16.55, con un mínimo de $2.50 y un máximo de $97.50, mostrando una alta variabilidad entre las tarifas.

  - **Multiplicador de tarifa dinámica (`surge_multiplier`):** Presenta un valor promedio de 1.01, con un máximo de 3.0, lo que indica que, aunque la mayoría de los viajes mantienen tarifas estándar, existen casos en los que se aplican aumentos por alta demanda.

- **Marca temporal (`time_stamp`):** Inicialmente se encontraba en formato numérico (milisegundos desde época UNIX), por lo que fue convertida al formato de fecha y hora para facilitar el análisis temporal de los viajes.

En conjunto, estos resultados muestran que los viajes registrados son en su mayoría cortos, con tarifas moderadas y un bajo impacto de precios dinámicos, aunque con presencia ocasional de picos de tarifa.

➢ **Weather**

Al observar el resumen estadístico, no se identifican valores atípicos significativos, ya que la mayoría de las variables presentan medias y medianas cercanas, lo que indica una distribución relativamente normal de los datos.

- **Temperatura (temp):** promedio de 39 °F con una dispersión moderada (desviación estándar ≈ 6). Los valores mínimos (19.6 °F) y máximos (55.4 °F) reflejan variaciones normales dentro del rango climático esperado.

- **Nubosidad (clouds):** promedio bajo (≈ 0.68), lo que sugiere predominio de cielos parcialmente despejados.

- **Presión (pressure):** valor medio de 1008 hPa con poca variabilidad (std ≈ 12.9), representando estabilidad atmosférica.

- **Lluvia (rain):** pocos datos registrados (≈ 894 observaciones), con una media muy baja (0.057 mm), indicando lluvias ligeras o poco frecuentes.

- **Humedad (humidity):** promedio del 76 %, coherente con un ambiente húmedo moderado y sin presencia de valores extremos.

- **Viento (wind):** velocidad promedio de 6.8 m/s, con cierta variabilidad (std ≈ 3.6), dentro de los valores típicos.

- **Temperatura en °C (temp_c):** se mantiene en la misma tendencia que `temp`, ya que corresponde a su conversión directa de Fahrenheit a Celsius.

En conjunto, los datos meteorológicos presentan coherencia, sin anomalías marcadas, y describen condiciones ambientales estables con variaciones normales propias del clima.

5. Evaluación de calidad de datos

5.1 Valores Duplicados

➢ **Rides**

Durante la fase de exploración de datos se realizó una verificación para identificar la existencia de registros duplicados en el dataset rides_exploratorio. El resultado obtenido fue cero duplicados, lo que indica que todos los registros son únicos. Esto garantiza la integridad y consistencia de la información, por lo que no es necesario aplicar ningún proceso de limpieza adicional en este aspecto.

➢ **Weather**

Al realizar la verificación de duplicados en el dataset weather, se comprobó que no existen registros repetidos. Esto confirma que cada observación climática es única y que los datos meteorológicos presentan una estructura limpia y confiable, adecuada para su posterior análisis y correlación con la información de los viajes.

5.2 Nulos

➢ **Rides**

En el análisis de valores nulos del dataset rides_exploratorio, se identificó que la columna price presenta aproximadamente un 8% de datos faltantes. Dado que esta variable es fundamental para el estudio del comportamiento de las tarifas y su relación con otros factores, se considera conveniente imputar los valores faltantes en lugar de eliminarlos. Esta decisión permite preservar la integridad del conjunto de datos y asegurar un análisis más completo y representativo de las tendencias de precios en los servicios de transporte digital.

➢ **Weather**

Al analizar el porcentaje de valores nulos en el dataset weather_exploratorio, se identificó que la columna rain presenta aproximadamente un 86% de datos faltantes. Inicialmente, se consideró la posibilidad de eliminar esta variable debido a su alto nivel de incompletitud y al riesgo de que aportara poca información estadística al análisis.

No obstante, tras una revisión más detallada, se concluyó que los valores nulos probablemente representan la ausencia de lluvia en los registros meteorológicos. Por este motivo, se decidió conservar la columna e imputar los valores faltantes con cero (0), interpretándolos como días o momentos sin precipitación. Esta decisión permite mantener la coherencia del conjunto de datos y aprovechar la variable en los análisis posteriores sobre la influencia de las condiciones climáticas en la demanda y los precios de los viajes.

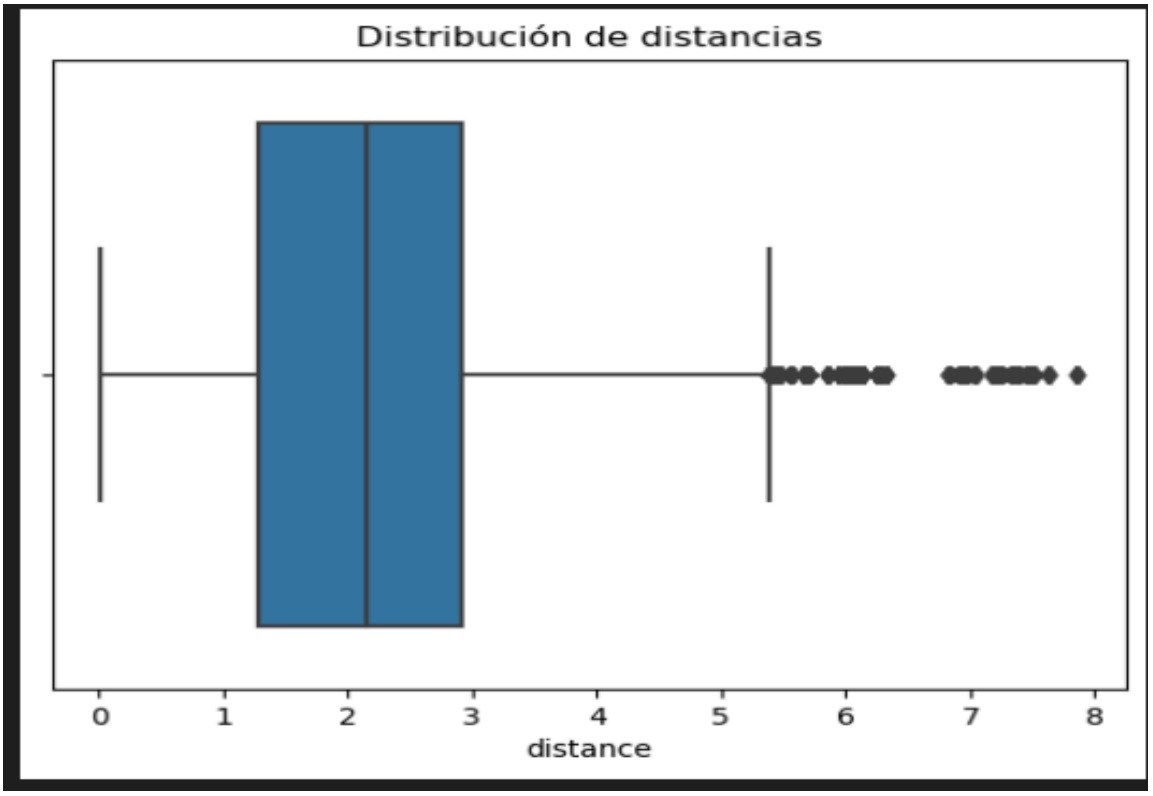
Dataset Rides

6.Exploración de Distance

6.1 Distribución de distancias

En el análisis exploratorio de la variable distance, se observó que la mayoría de los viajes registrados se concentran en un rango comprendido entre 1 y 3 millas, lo que sugiere que los usuarios de servicios de transporte por aplicación, como Uber y Lyft, suelen utilizarlos principalmente para traslados cortos dentro de la ciudad. La mediana, ubicada alrededor de los 2 millas, confirma esta tendencia hacia desplazamientos de corta distancia, típicos de trayectos urbanos o interzonales.

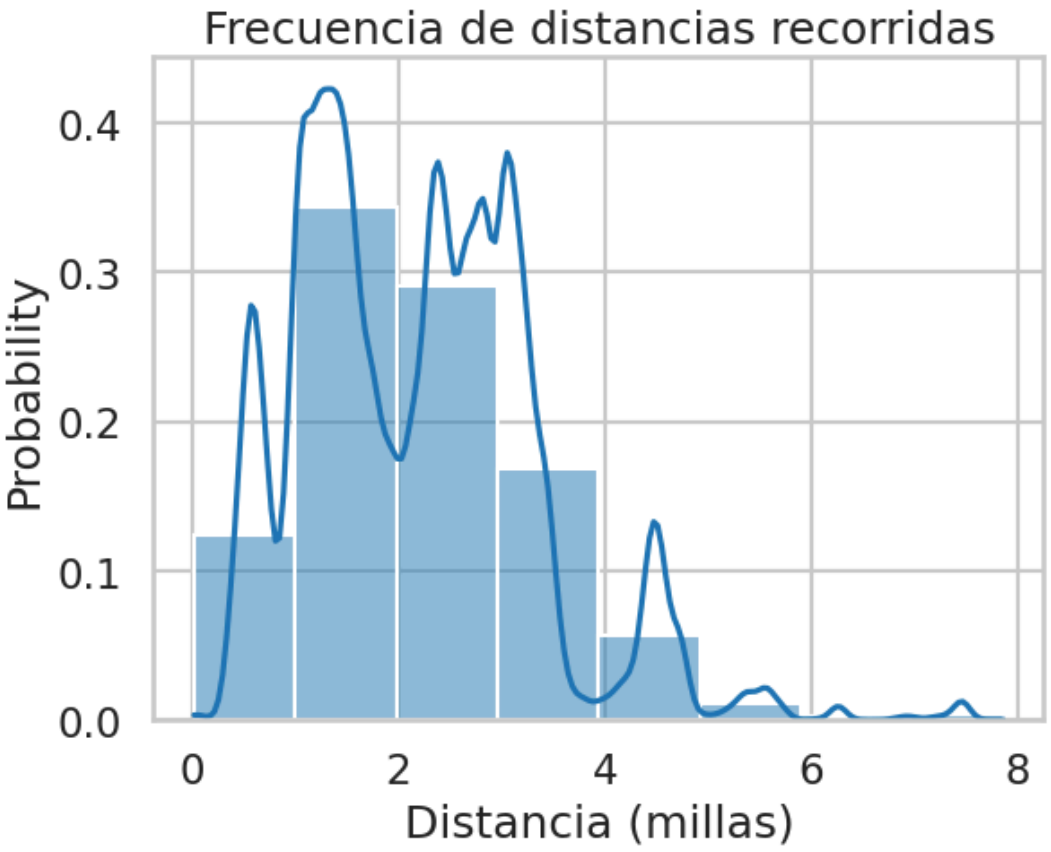
Asimismo, se identificaron algunos viajes con distancias superiores a los 5 millas, considerados valores atípicos o outliers, que podrían corresponder a desplazamientos hacia zonas periféricas o viajes excepcionales de mayor duración. En términos generales, la distribución de la distancia es equilibrada, sin un sesgo marcado hacia ningún extremo, lo que evidencia una dispersión moderada de los datos y un comportamiento coherente con el uso habitual de estas plataformas en entornos urbanos.



6.2 Frecuencia de distancias recorridas

El análisis de la distribución de la variable ‘distance’ mediante un histograma permitió visualizar con mayor detalle la frecuencia de los trayectos realizados. Se observó que la mayor parte de los viajes se concentran en el rango de 1 a 3 millas, con un pico pronunciado alrededor de los 2 millas, lo que indica que los usuarios recurren principalmente a estos servicios para desplazamientos cortos y rápidos dentro de la ciudad.

A medida que la distancia aumenta, la frecuencia de los viajes disminuye de manera notable, siendo los trayectos de más de 4 millas considerablemente menos comunes. Esta tendencia confirma los resultados obtenidos en el boxplot, donde también se identificó la presencia de outliers o valores atípicos que representan viajes más largos y menos frecuentes. En conjunto, la distribución muestra un comportamiento típico del transporte urbano, donde predomina la demanda de recorridos cortos asociados a la movilidad cotidiana.



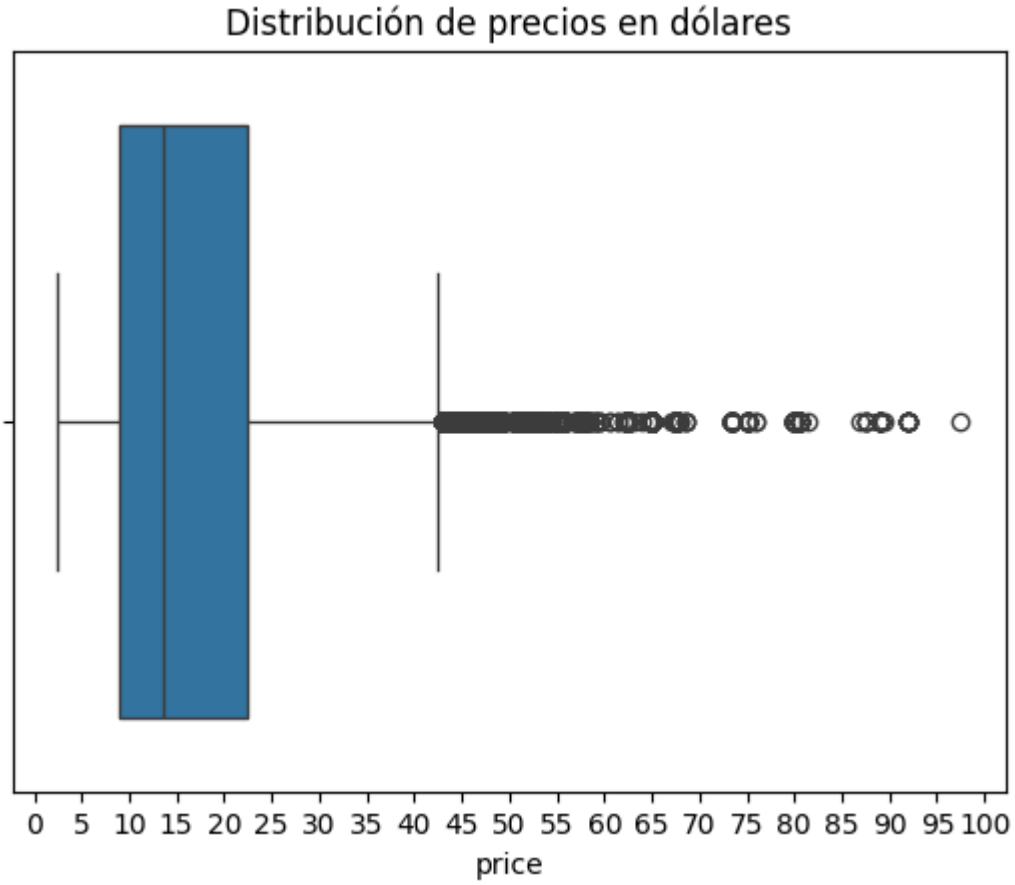
7.Exploración de precios

7.1 Distribución de precios en dólares

En el análisis de la variable price, se observó que la mayoría de los viajes tienen un costo que oscila entre 9 y 22.5 d, rango que corresponde al intervalo intercuartílico y concentra la mayor parte de los valores. El precio típico o mediana se sitúa alrededor de los 13.5 dólares, lo que representa el costo más común de los viajes urbanos.

Sin embargo, también se detectaron valores extremos que alcanzan tarifas de hasta casi 100 dólares, los cuales, aunque poco frecuentes, incrementan el promedio general de los precios. La diferencia entre la mediana (13.5) y la media (16.5) revela que la distribución está sesgada hacia la derecha, es decir, existen algunos viajes significativamente más caros que influyen en el promedio total. Este comportamiento

es típico en servicios con tarifas dinámicas, donde factores como la distancia, la demanda o las condiciones climáticas pueden elevar el precio de manera puntual.

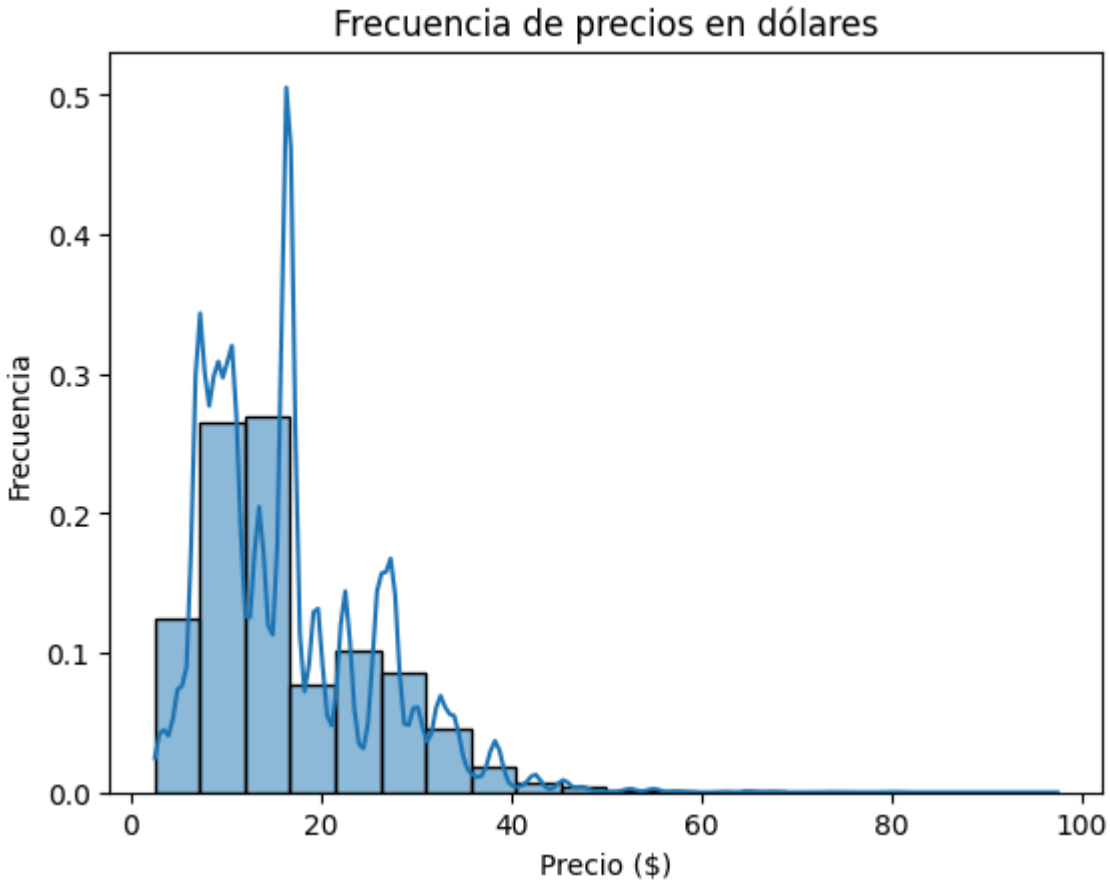


7.2 Frecuencia de precios en dólares

El análisis de la distribución de la variable price mediante el histograma permitió observar que la mayor parte de los viajes presentan tarifas bajas, concentradas principalmente entre 0 y 20 dólares. El pico de frecuencia se encuentra en el rango de 10 a 15 dólares, lo que coincide con la mediana identificada en el boxplot, representando el costo más habitual de los servicios de transporte urbano.

A medida que el precio aumenta, la frecuencia de los viajes disminuye rápidamente, lo que indica que los trayectos más costosos son poco frecuentes y corresponden a situaciones específicas, como recorridos largos o momentos de alta demanda. La curva suavizada (KDE) muestra claramente que la distribución está sesgada hacia la derecha, con una cola larga de precios elevados que aparecen como outliers. Este

patrón es característico de mercados donde predominan los servicios económicos, pero existen casos excepcionales con tarifas significativamente más altas.



8.Correlación

8.1 Correlación entre precio y distancia

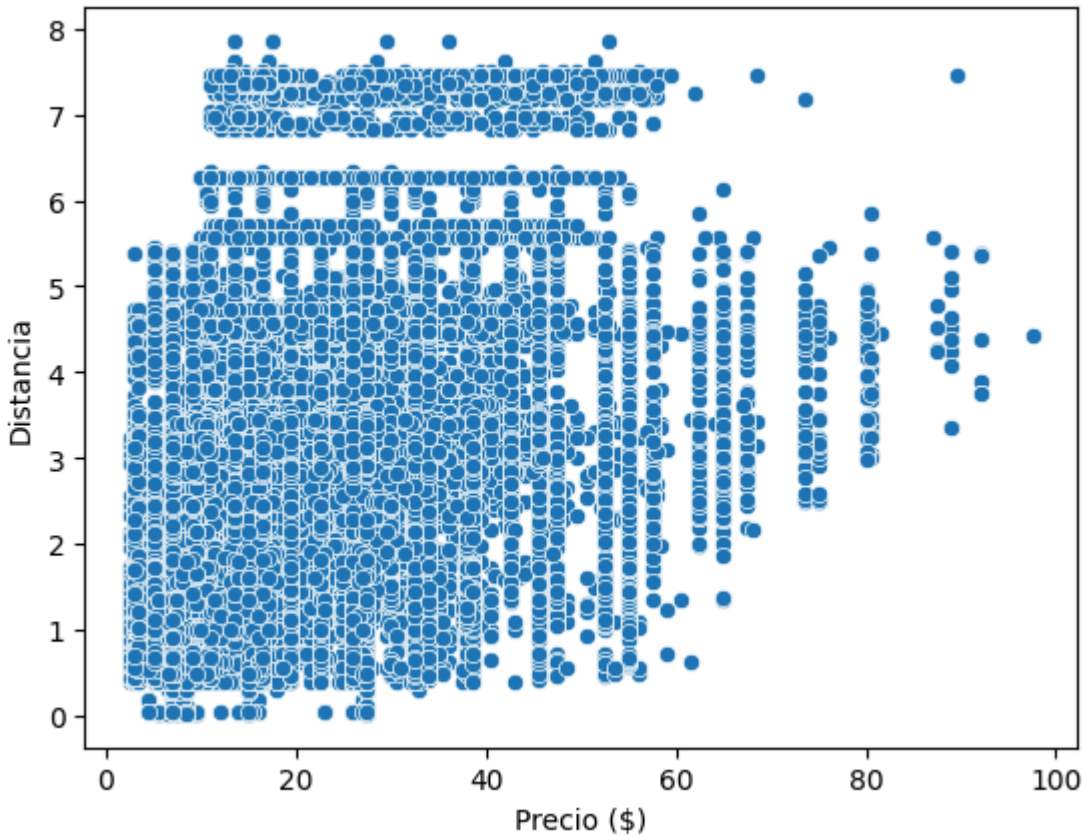
El análisis de la correlación entre el precio y la distancia mostró un coeficiente de 0.3451, lo que indica una relación positiva pero débil entre ambas variables. En otras palabras, aunque los viajes más largos tienden a tener precios más altos, la relación no es lo suficientemente fuerte como para afirmar que la distancia sea el único factor determinante del costo del viaje.

Esta correlación moderadamente baja sugiere que otros factores, como el tipo de servicio (por ejemplo, económico o de lujo), el momento del día, las condiciones del tráfico o la aplicación de tarifas dinámicas, también influyen de manera significativa en el precio final.

Por esta razón, se considera apropiado imputar los valores faltantes de la variable “price” utilizando el promedio general, ya que esta estrategia permite mantener la coherencia del conjunto de datos sin introducir sesgos relevantes derivados de la distancia.

Para manejar los valores faltantes en la columna price, se aplicó una imputación utilizando el promedio general de los precios calculado a partir del conjunto de datos disponible. Este valor promedio se utilizó para rellenar las observaciones con datos nulos, garantizando así la completitud del dataset sin eliminar registros relevantes.

Esta decisión metodológica se justifica porque la correlación entre precio y distancia (0.3451) es relativamente baja, lo que indica que el precio no depende exclusivamente de la distancia. Por lo tanto, el uso del valor medio permite mantener la consistencia estadística del análisis y evita distorsionar las distribuciones generales de tarifas, facilitando un estudio más robusto y equilibrado de las relaciones entre las variables.



Dataset Weather

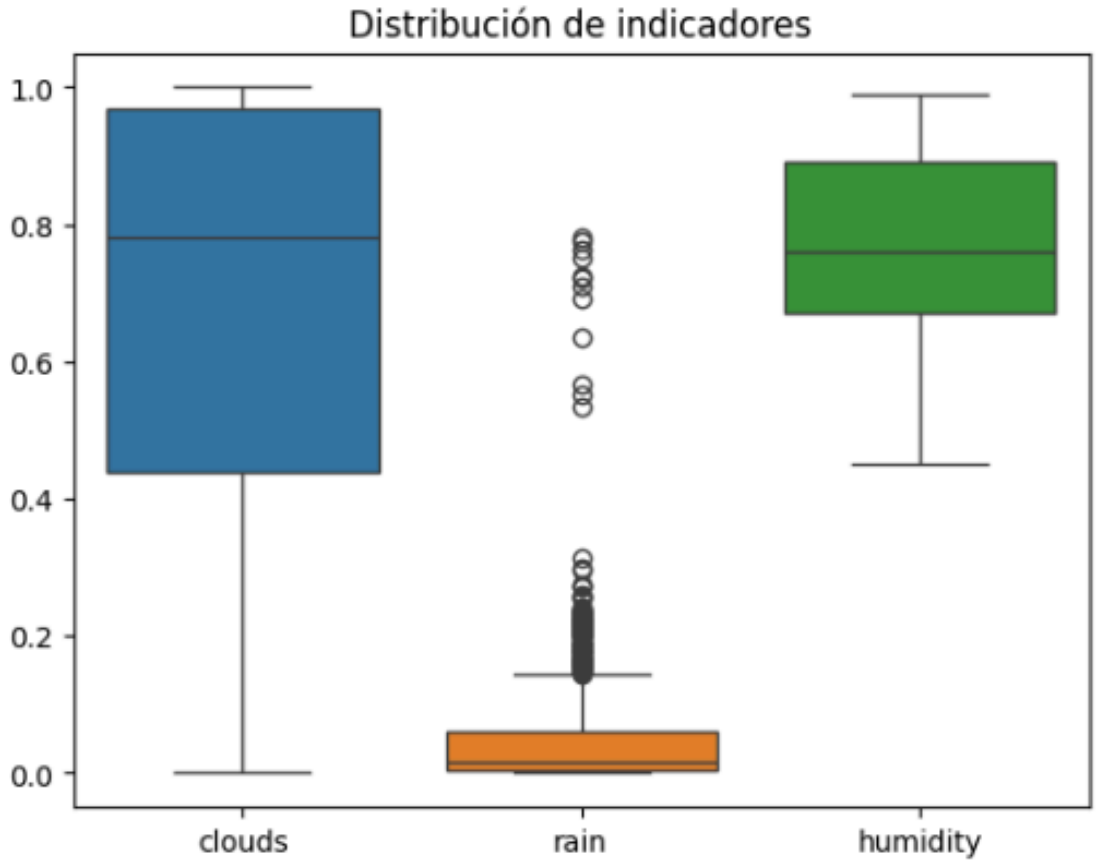
9.Exploración

9.1 Distribución de indicadores

Se realizó un boxplot para las variables *clouds*, *rain* y *humidity* con el fin de observar su dispersión y posibles valores atípicos.

- Rain (lluvia): se confirma que los valores mínimos son mayores a 0, lo que permite imputar los valores nulos con 0 (representando ausencia de lluvia). Sin embargo, el 75 % de los registros no supera los 0.1 mm, lo que demuestra que la mayoría de los días presentaron poca o nula precipitación.
  Por esta razón, la variable *rain* no aporta información significativa al análisis global y se decide eliminar la columna.

Por último, se aplicó `weather_exploratorio.dropna(axis=1, inplace=True)` para limpiar el conjunto de datos, eliminando las columnas con valores nulos restantes y optimizando el dataset para análisis posteriores.



10. Análisis de distribución de variables

10.1 Exploración de temperatura

Se compararon las estadísticas descriptivas de las columnas `temp` (°F) y `temp_c` (°C), confirmando que la segunda corresponde correctamente a la conversión de la primera, aplicando la fórmula.

- **Temperatura en Fahrenheit (`temp`):**
  - **Promedio:** 39.09 °F
  - **Rango:** 19.6 °F – 55.4 °F
  - La dispersión (std = 6.02) indica una variabilidad moderada.
  - La mayoría de los valores (entre el 25 % y 75 %) se concentran entre 36.1 °F y 42.8 °F, es decir, temperaturas frías pero estables.
- **Temperatura en Celsius (`temp_c`):**
  - **Promedio:** 3.94 °C, equivalente al promedio anterior.
  - **Rango:** –6.88 °C a 13.00 °C, coherente con las condiciones de clima frío.
  - La desviación estándar (3.34) mantiene la misma proporción de variabilidad.
- Los valores son consistentes y no presentan anomalías. La conversión a grados Celsius permite interpretar más fácilmente el comportamiento térmico

del dataset, mostrando un clima predominantemente frío con variaciones leves.

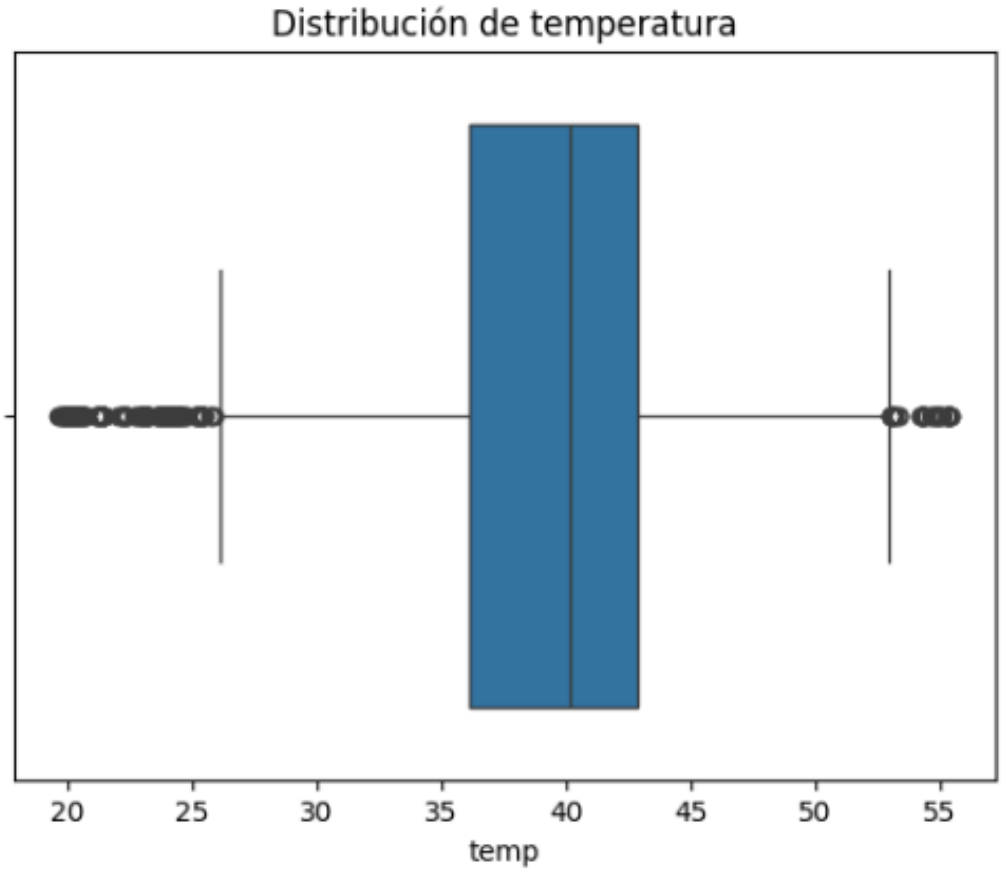
11.Valores únicos y su frecuencia

11.1 Distribución de temperatura

El análisis de la variable temperatura mediante el boxplot muestra que la mayoría de los registros se concentran entre 36 °F y 43 °F, lo cual representa el rango típico de las condiciones climáticas observadas. La mediana se sitúa alrededor de 40 °F, reflejando que la mitad de los días presentan temperaturas iguales o inferiores a ese valor.

Se identifican algunos valores atípicos en los extremos de la distribución: por un lado, días con temperaturas superiores a 50 °F, poco frecuentes dentro del conjunto de datos; y por otro, días con temperaturas inferiores a 26 °F, que también se consideran inusuales.

En general, la distribución es relativamente equilibrada, aunque la presencia de estos valores extremos sugiere la existencia de variaciones climáticas ocasionales fuera del rango habitual. Estos resultados permiten concluir que las condiciones térmicas son estables en su mayoría, con pocas excepciones que pueden atribuirse a fenómenos meteorológicos aislados.

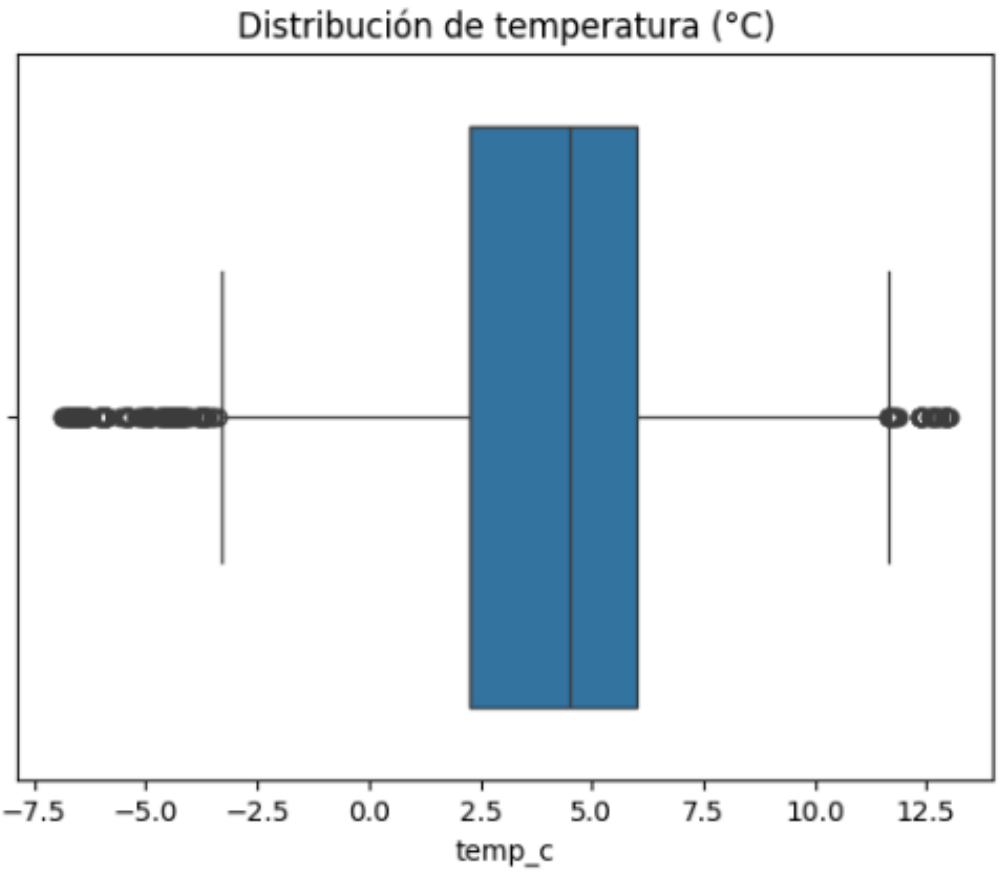


11.2 Distribución de temperatura en grados Celsius

El boxplot presentado muestra la distribución de la temperatura en grados Celsius, convertida desde la escala Fahrenheit. Se observa que la mayoría de los valores se concentran entre aproximadamente 2 °C y 6 °C, rango donde se ubica el 50% central de las observaciones.

La mediana, cercana a 4.5 °C, indica que la mitad de los registros corresponden a días con temperaturas frías, pero no extremas. En los extremos del gráfico se identifican valores atípicos por debajo de -5 °C y por encima de 10 °C, que representan condiciones menos frecuentes, posiblemente asociadas a eventos climáticos puntuales.

En general, la distribución sugiere un comportamiento térmico estable y uniforme, con una ligera dispersión hacia ambos lados, lo cual refleja la consistencia del conjunto de datos meteorológicos y confirma la ausencia de anomalías significativas más allá de los valores extremos naturales.

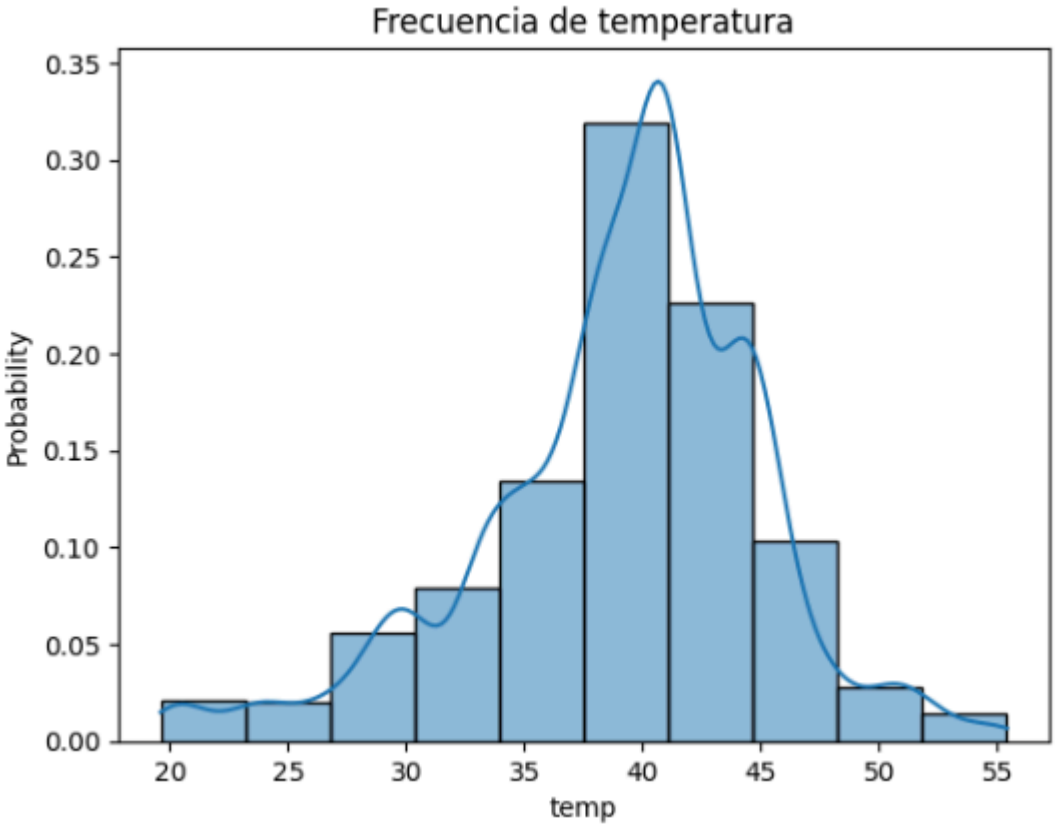


11.3 Frecuencia de temperatura

El histograma de temperaturas revela una distribución asimétrica, con una ligera inclinación hacia valores más altos. La mayor concentración de observaciones se ubica entre 35 °F y 45 °F, donde se encuentra la mayoría de los días registrados.

La curva KDE confirma esta tendencia, mostrando un pico pronunciado en torno a los 40 °F, lo que sugiere que la mayoría de los registros corresponden a condiciones frías o templadas.

Esta información es relevante para el análisis general, ya que permite comprender el contexto climático en el que se realizaron los viajes y evaluar posibles relaciones entre la temperatura y la demanda de transporte, por ejemplo, si los días más fríos presentan una mayor frecuencia de viajes o una variación en los precios.



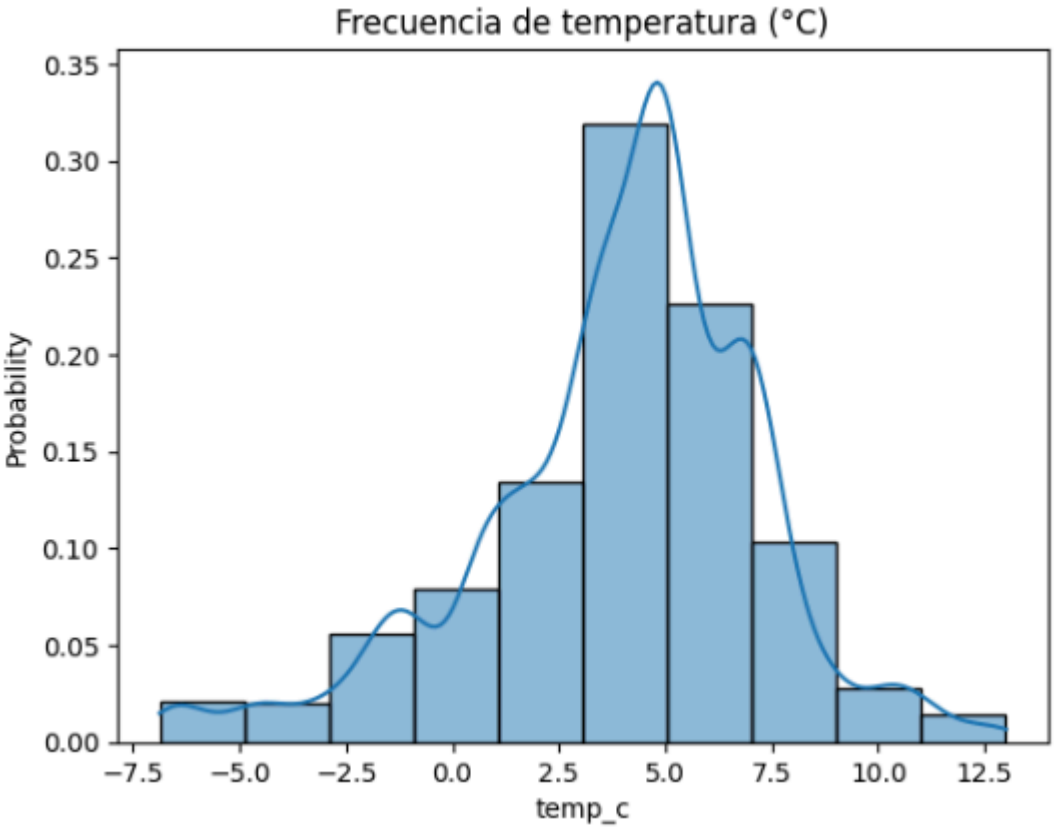
11.4 Frecuencia de temperatura en grados Celsius

El histograma de temperatura en grados Celsius muestra una distribución unimodal con un claro pico alrededor de los 4 °C, que corresponde al valor medio y mediano

del conjunto de datos. La mayoría de las observaciones se concentran entre 2 °C y 6 °C, reflejando un clima predominantemente frío.

La curva KDE indica una forma ligeramente sesgada hacia la derecha, lo que sugiere la presencia de algunos días más cálidos, aunque poco frecuentes (por encima de los 10 °C).

En general, la dispersión es moderada y no se aprecian valores atípicos significativos. Este comportamiento permite contextualizar las condiciones climáticas del entorno, siendo útil para analizar cómo la temperatura podría influir en la frecuencia o el costo de los viajes registrados.



12.Exploración de Humedad

El análisis descriptivo de la variable humedad muestra una media de 0.76 y una mediana idéntica (0.76), lo que refleja una distribución equilibrada y simétrica. Los valores se encuentran dentro del rango esperado (de 0.45 a 0.99), lo que confirma que no existen datos anómalos o fuera de escala.

El rango intercuartílico (0.67 – 0.89) indica que la mayoría de las observaciones se concentran en niveles moderados a altos de humedad, lo que sugiere un entorno climático

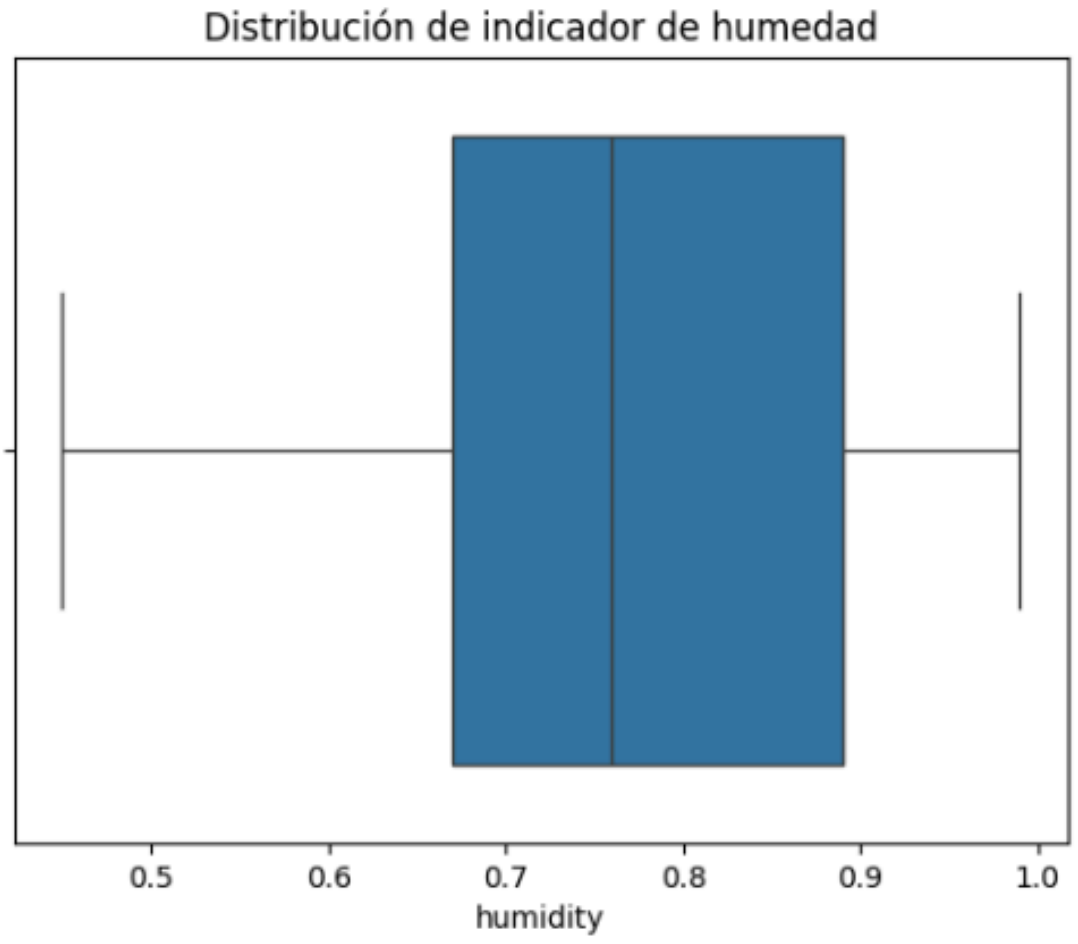
húmedo y estable. Esta variable será relevante al momento de explorar si condiciones de humedad elevada tienen alguna relación con la demanda o frecuencia de los viajes.

13. Valores únicos y su frecuencia

13.1 Distribución de indicador de humedad

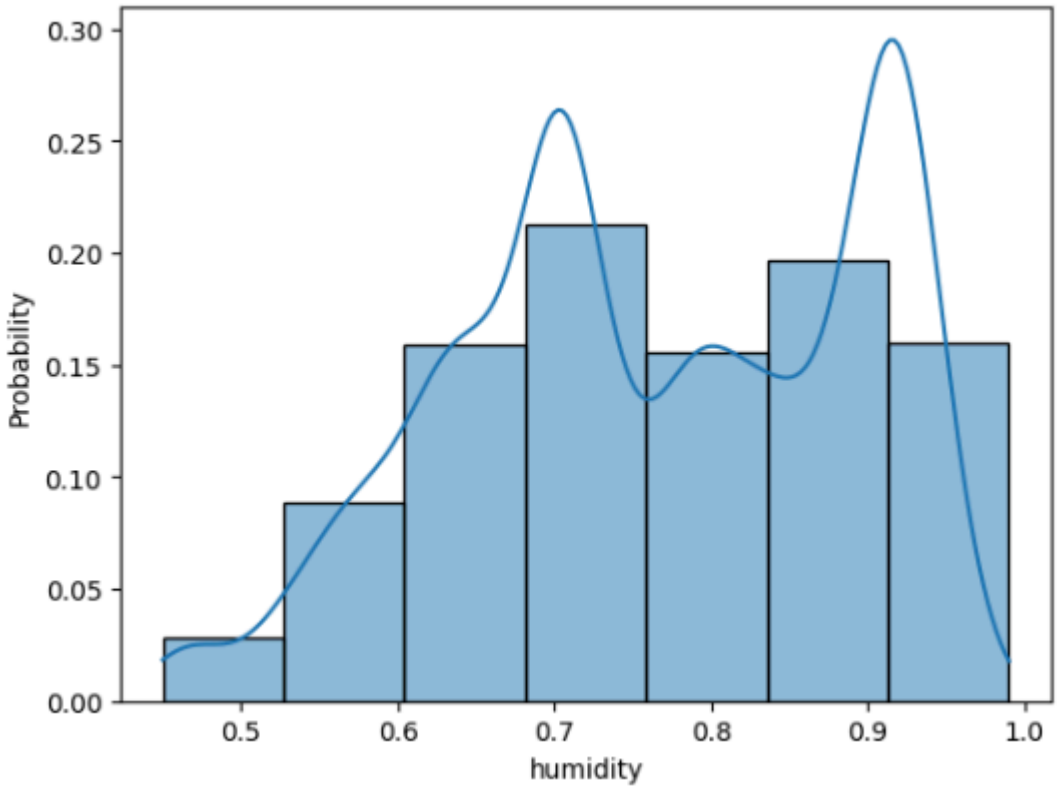
El diagrama de caja de la variable humedad muestra que la mayoría de los valores se concentran entre 0.67 y 0.9, reflejando un ambiente predominantemente húmedo. La mediana de 0.76 confirma una leve asimetría hacia la derecha, lo que indica que existen algunos registros con valores algo menores, aunque no significativos.

El bigote inferior se extiende hacia valores más bajos (alrededor de 0.45), lo que sugiere ligeras variaciones de humedad, pero sin presencia de valores atípicos. En general, la distribución es estable y homogénea, lo que refuerza la validez de esta variable para futuros análisis climáticos relacionados con la demanda de transporte.



13.2 Distribución de la humedad

El histograma de la variable humedad muestra una concentración clara de valores entre 0.65 y 0.9, lo que refleja un clima con alta humedad relativa durante la mayor parte del tiempo. El pico de densidad se ubica entre 0.75 y 0.8, coincidiendo con la media general (0.76), lo que sugiere una distribución bastante equilibrada y sin presencia de valores atípicos. En conjunto, estos resultados indican condiciones atmosféricas predominantemente húmedas, con pocas variaciones extremas.



14. Correlación dataset Weather

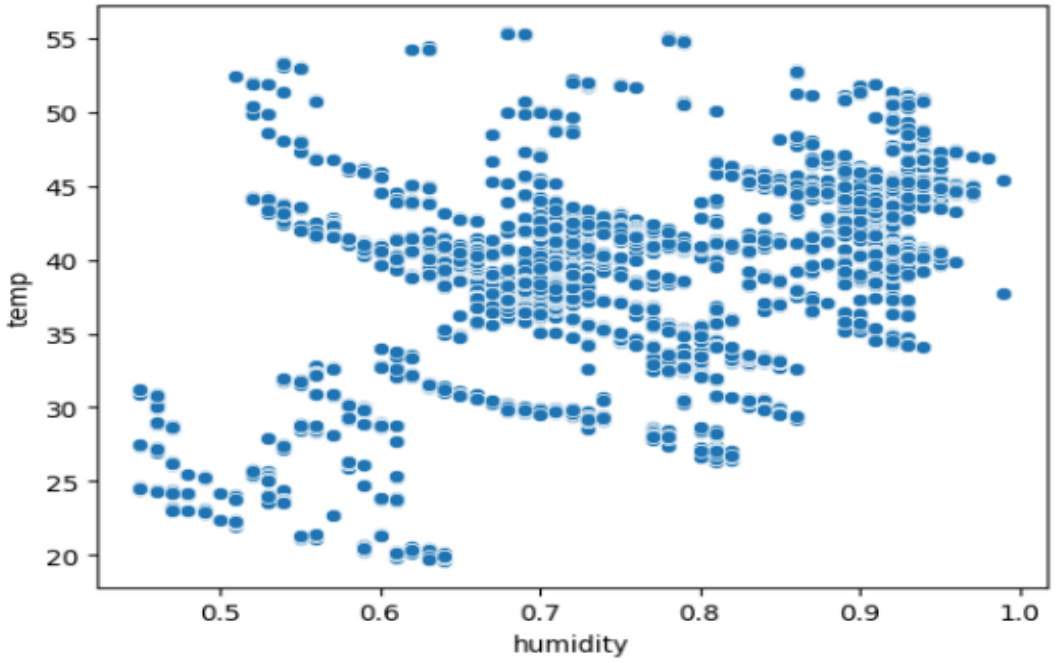
14.1 Correlación entre temperatura y humedad

La correlación entre la temperatura y la humedad es positiva y débil, con un coeficiente de 0.3873. Esto significa que, en general, cuando la humedad aumenta, la temperatura también tiende a subir ligeramente, aunque la relación no es fuerte ni constante. En el gráfico de dispersión se observa que la mayoría de los puntos se concentran en el rango de humedad entre 0.65 y 0.9 y temperaturas entre 35 °F y 45 °F. No existe una tendencia lineal clara que indique una dependencia directa entre ambas variables, por lo que se concluye que la temperatura y la humedad están

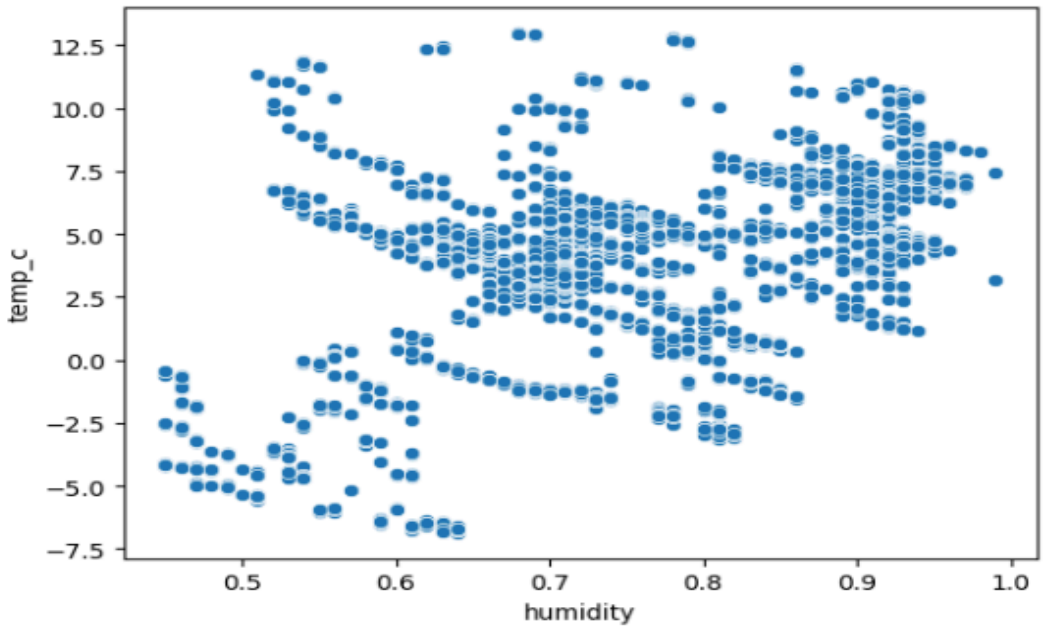
débilmente relacionadas y otros factores climáticos probablemente influyen más en la variación de la temperatura.

Cabe destacar que se realizaron dos gráficos distintos: uno utilizando `y=temp` y otro con `y=temp_c`. La diferencia radica en que `temp` representa la temperatura en grados Fahrenheit, mientras que `temp_c` la muestra en grados Celsius. Esta variación se aplicó para analizar si el cambio de unidad afectaba la relación entre temperatura y humedad; sin embargo, el resultado fue el mismo, mostrando una correlación débilmente positiva en ambos casos, lo que confirma que la relación entre las variables es independiente del sistema de medida.

y= temp



y= temp_c



15. Preguntas de Análisis

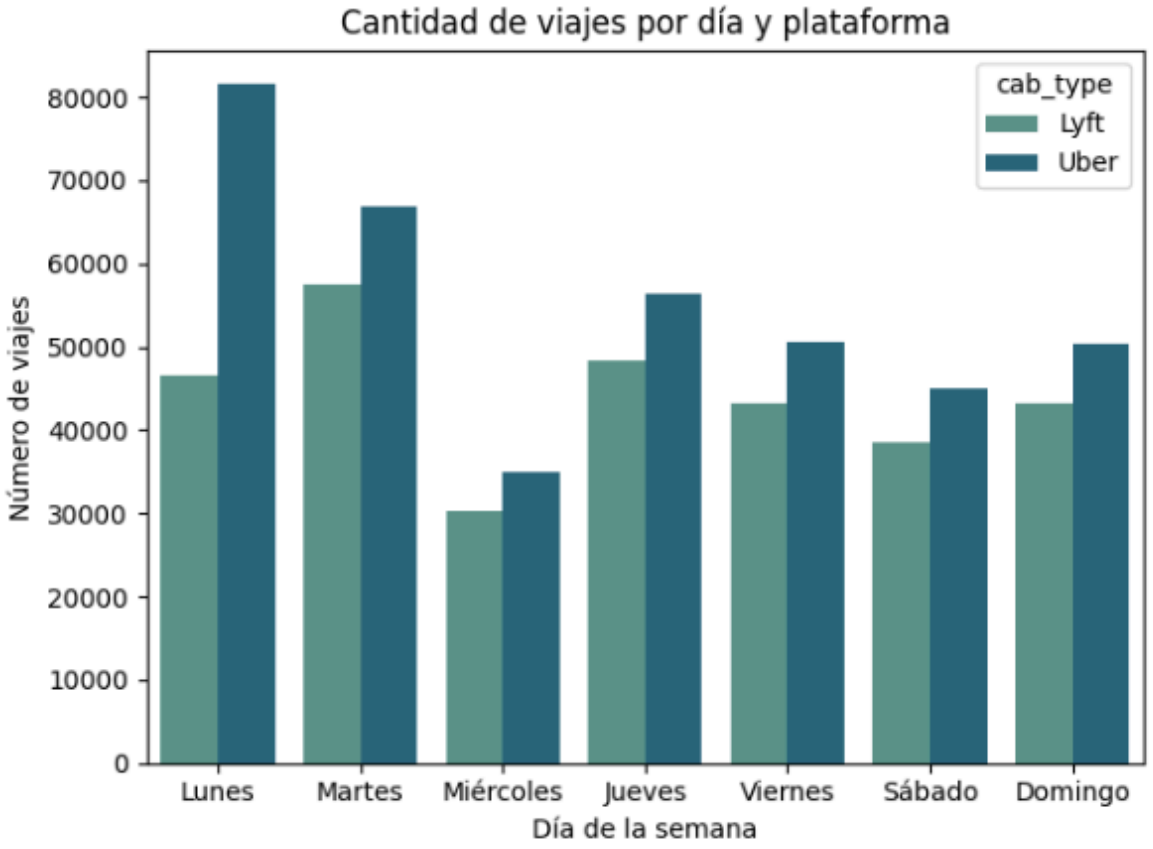
15.1 ¿Cuál es el día con mayor viajes ? ¿ Cual es la app más usada?

El análisis de la cantidad de viajes realizados por día de la semana, diferenciándose entre las plataformas Uber y Lyft, permite identificar patrones claros de comportamiento en la demanda de transporte.

De acuerdo con los resultados obtenidos:

- El lunes es el día con mayor número de viajes registrados, lo que indica un incremento significativo en la movilidad al inicio de la semana laboral.
- A medida que avanza la semana, la cantidad de viajes presenta una tendencia decreciente, con valores más bajos entre miércoles y sábado, y una ligera recuperación el domingo.
- En todas las jornadas analizadas, Uber se posiciona como la aplicación más utilizada, superando a Lyft de forma consistente.

En conclusión, el patrón de uso sugiere que la demanda de transporte compartido es más alta al inicio de la semana, y que los usuarios muestran una preferencia marcada por Uber, posiblemente por su mayor cobertura, disponibilidad o reconocimiento en el mercado.



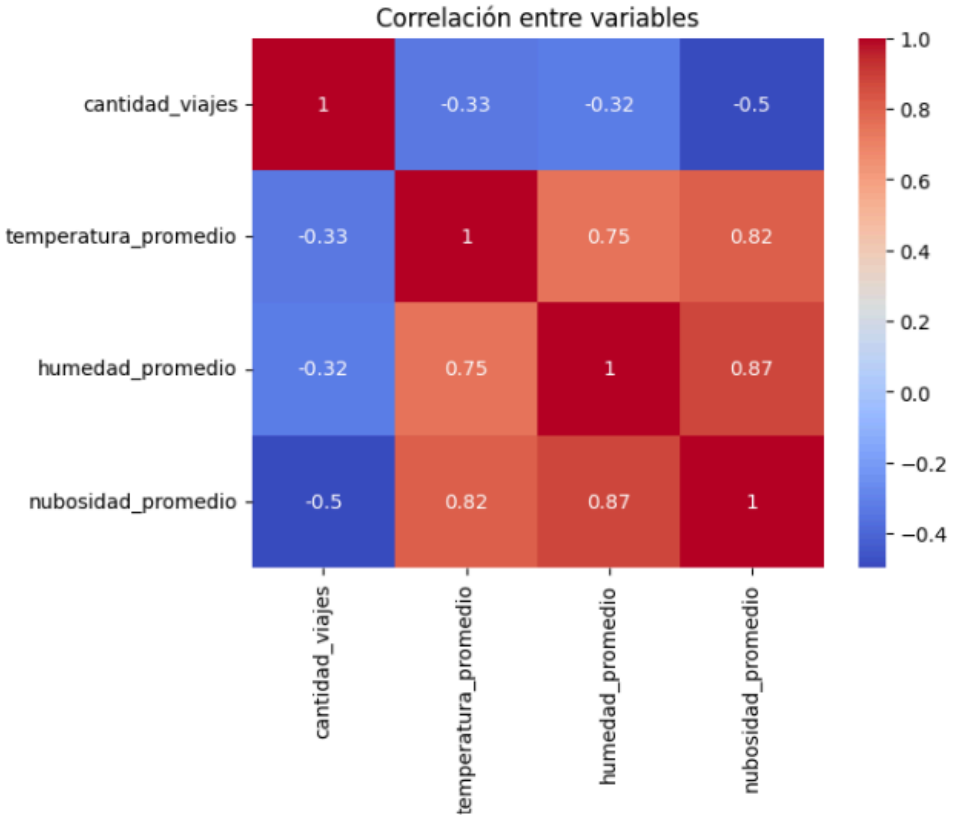
15.2 ¿ Influye el clima en la cantidad de viajes?

En este análisis se unieron los datos de viajes y condiciones climáticas a través de la columna fecha, permitiendo combinar información de ambas fuentes y estudiar cómo el clima influye en la demanda de transporte.

Para identificar si las condiciones climáticas influyen en la cantidad de viajes realizados, se elaboró un mapa de calor (heatmap) que muestra las correlaciones entre variables como la temperatura, la humedad, la velocidad del viento y las precipitaciones frente al número total de viajes.

Al analizar el *heatmap*, la nubosidad muestra una correlación positiva con la temperatura y la humedad, lo que indica que a mayor presencia de nubes, estas condiciones climáticas también tienden a incrementarse. Sin embargo, la nubosidad presenta una correlación negativa con la cantidad de viajes, sugiriendo que cuando el cielo está más cubierto, la demanda de transporte disminuye.

En contraste, los datos permiten concluir que los días con cielo más despejado registran un mayor número de viajes. Esto evidencia una tendencia donde las mejores condiciones climáticas favorecen la movilidad y el uso de aplicaciones de transporte.

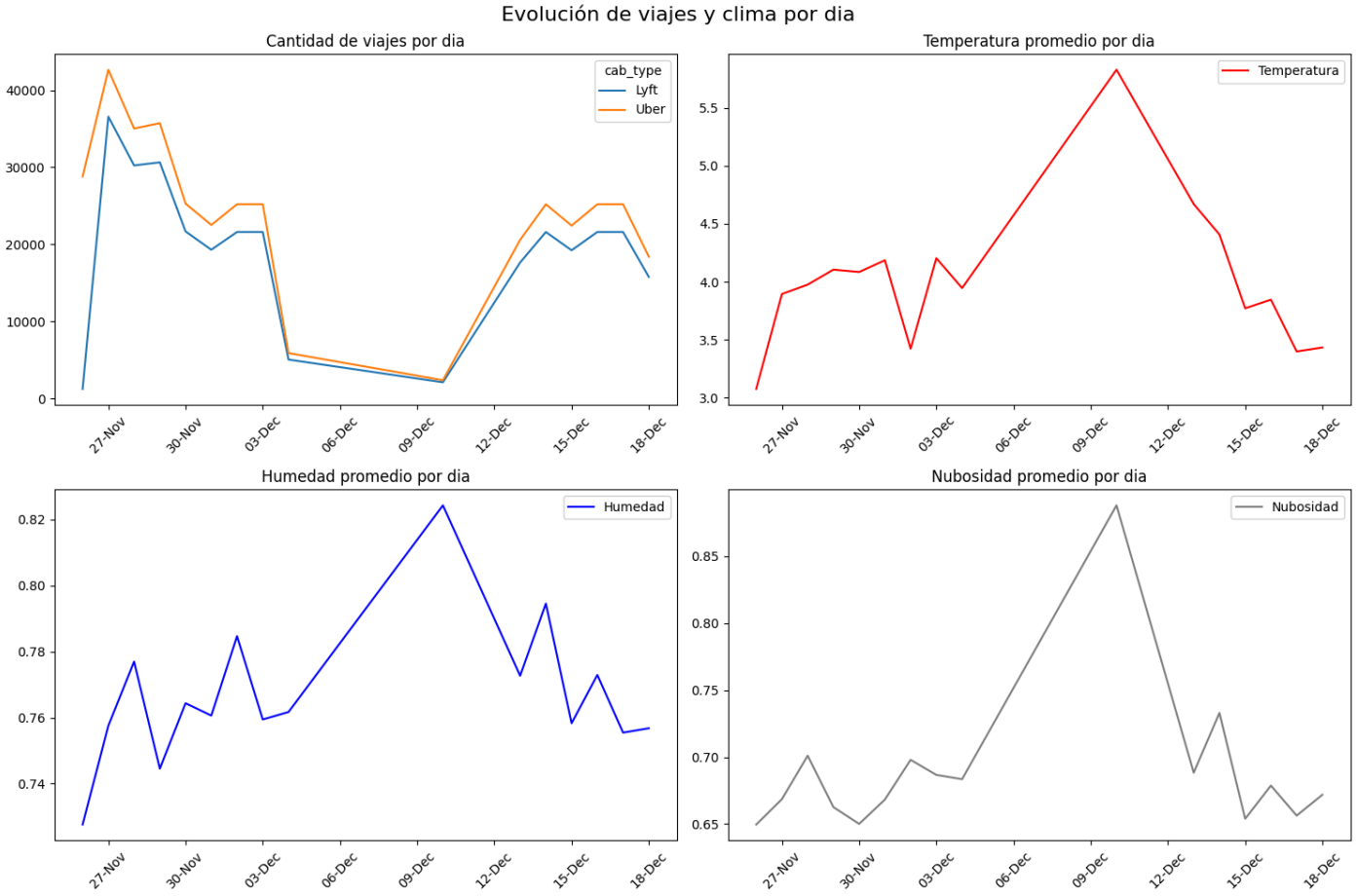


Evolución de Viajes y Variables Climáticas por Fecha

El gráfico muestra la evolución de la cantidad de viajes tanto de Uber como de Lyft, junto con las variables climáticas temperatura promedio, humedad promedio y nubosidad promedio a lo largo del tiempo.

Se puede observar que la cantidad de viajes presenta una variación considerable, con picos altos a finales de noviembre y descensos marcados a inicios de diciembre. Sin embargo, las líneas correspondientes a las variables climáticas se mantienen prácticamente estables y con valores mucho menores en comparación con el volumen de viajes, lo cual indica que no existe una relación evidente entre las condiciones del clima y el comportamiento de los viajes.

Esto sugiere que los cambios en la temperatura, humedad o nubosidad no afectan significativamente la cantidad de viajes realizados. Es posible que la variación en los desplazamientos responda a otros factores externos, como la actividad económica, los días laborales o los fines de semana, más que a las condiciones meteorológicas.



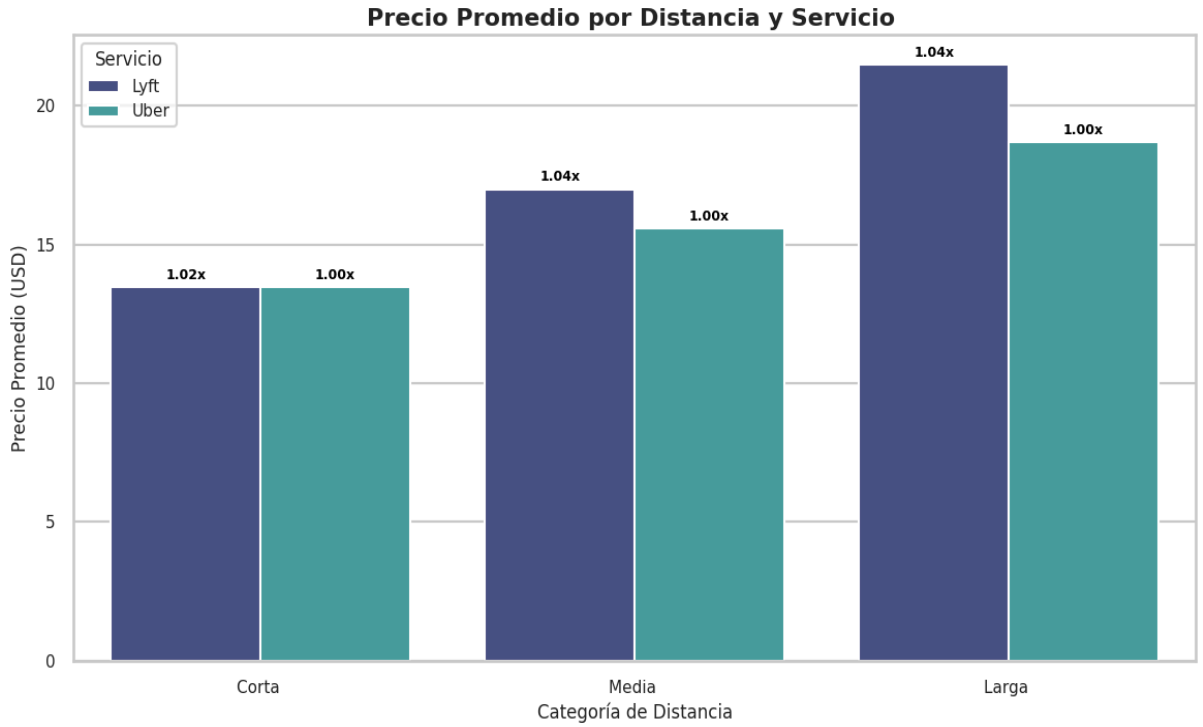
15.3 ¿Cuál es el precio promedio del servicio y cuál es el nivel de riesgo de volatilidad asociado a ese precio?

El análisis del precio promedio por categoría de distancia (corta 1.28mi, media 2.59mi y larga 7.55mi) muestra un patrón claro y consistente: Lyft mantiene tarifas más elevadas que Uber en todos los rangos de distancia.

En los viajes cortos, ambos servicios tienen precios muy similares; sin embargo, Lyft presenta un costo ligeramente superior. Es crucial notar que, en este segmento, Lyft también tiene un Multiplicador de Sobrecarga (`Surge_Promedio`) consistentemente más alto que Uber. Esto implica que el cliente de Lyft enfrenta una mayor volatilidad de precios y una aplicación de tarifas dinámicas para trayectos cortos.

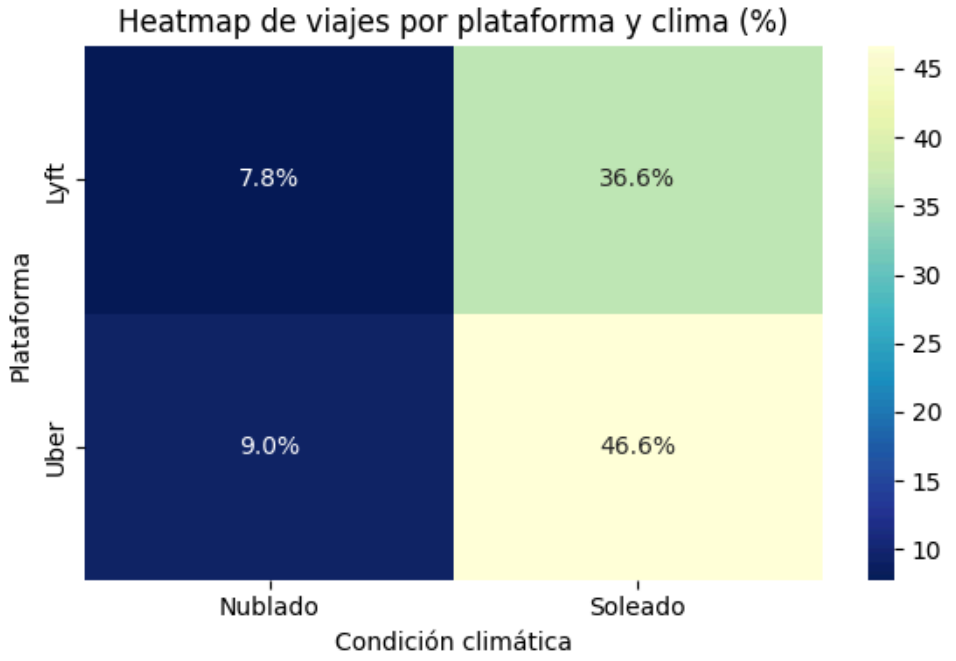
Esta diferencia de precios se vuelve más evidente en los viajes de distancia media y los viajes largos, donde Lyft cobra un promedio notablemente más alto que Uber, reflejando una política de precios más elevada para trayectos intermedios y de largas distancias. Lyft mantiene su ventaja en el `Surge_Promedio`, indicando que su mayor precio es resultado tanto de tarifas base elevadas como de un factor de riesgo de precio superior en comparación con Uber.

Esto sugiere que, conforme aumenta la distancia del recorrido, la diferencia de precios se incrementa. En estas categorías, Lyft exhibe el `Surge_Promedio` más elevado del análisis, por ende su posicionamiento como el servicio más costoso se basa en la combinación de la prima de precio y en su algoritmo de sobrecarga dinámica.



15.4 Viaje por plataforma en días soleados y nublados

El análisis revela una diferencia notable en la cantidad de viajes según las condiciones climáticas. En los días soleados, tanto Uber como Lyft registran un volumen considerablemente mayor de viajes, destacándose Uber como la plataforma más utilizada por los usuarios. Sin embargo, cuando el clima se vuelve nublado, la demanda disminuye drásticamente para ambas aplicaciones, lo que sugiere que las condiciones menos favorables pueden influir en la movilidad de las personas. A pesar de esta caída, Uber mantiene su liderazgo en número de viajes incluso en días nublados, aunque con un volumen considerablemente menor en comparación con los días soleados. Este comportamiento indica que, aunque el clima afecta la cantidad total de desplazamientos, la preferencia por Uber se mantiene constante independientemente de las condiciones atmosféricas.



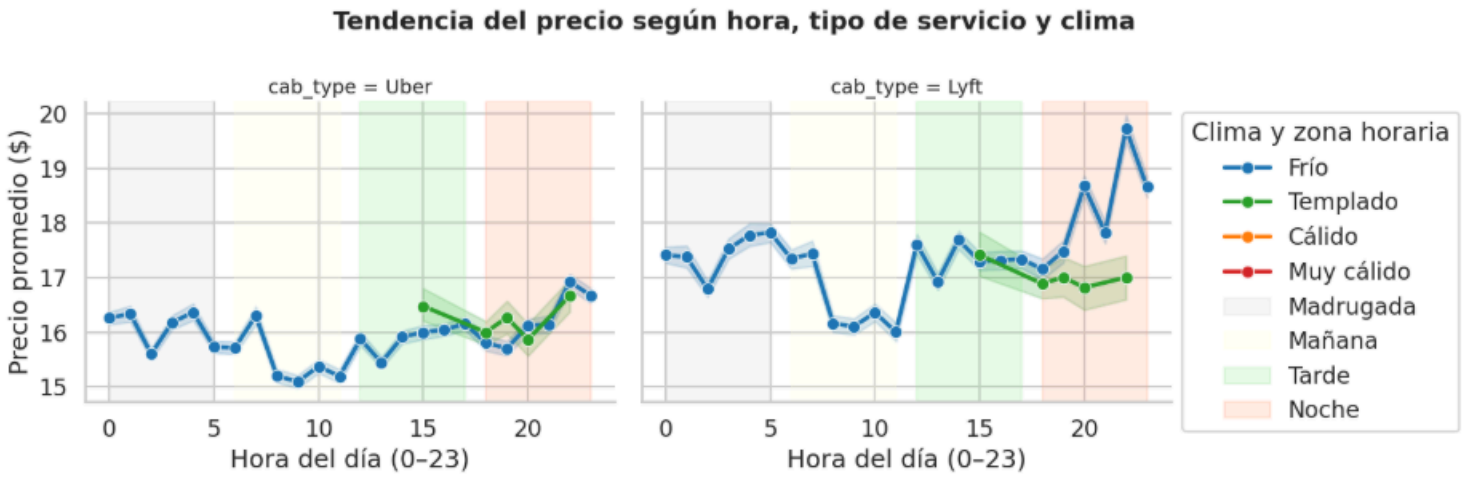
15.5 ¿Cómo interactúan clima, hora del día y tipo de servicio en el precio?

El gráfico permite analizar simultáneamente tres factores clave que influyen en el precio de los viajes por aplicación: la hora del día, el tipo de servicio (Uber o Lyft) y el clima, representado mediante rangos de temperatura (Frío, Templado, Cálido y Muy cálido). En primer lugar, se observa que la hora del día es el factor que más impacto tiene en las tarifas. Durante la madrugada (0–5 h), los precios se mantienen relativamente bajos y estables, lo que coincide con una menor demanda general en ese periodo. Conforme avanza la mañana (6–11 h), los precios tienen un incremento leve asociado probablemente a desplazamientos laborales. En la tarde (12–17 h), las tarifas continúan subiendo de manera constante, y este incremento se vuelve más evidente durante la noche (18–23 h), donde ambos servicios alcanzan sus valores más altos del día. Esto indica que la demanda nocturna, combinada con posibles incrementos por tráfico o eventos, juega un papel importante en el aumento de precios.

Al comparar los servicios, se evidencia que Lyft suele ser más costoso que Uber, especialmente en las horas de mayor demanda. Uber mantiene una línea más estable entre $15 y $17 durante prácticamente todo el día, lo que sugiere un sistema

de precios más controlado o menos afectado por factores externos. Por otro lado, Lyft presenta una mayor volatilidad: sus precios aumentan con más fuerza en la tarde y alcanzan picos significativos en la noche, llegando a valores cercanos a los $20. Esta diferencia podría deberse a distintas estrategias de pricing, variaciones en la oferta disponible de conductores o un mayor nivel de sensibilidad a la demanda.

En cuanto al clima, aunque su efecto es menos fuerte que el horario, igual influye notablemente. Los viajes realizados bajo clima frío muestran precios más altos, especialmente en Lyft. Esto puede estar relacionado con menor disponibilidad de conductores, mayor tráfico o condiciones adversas que dificultan la operación. En climas templados, las tarifas se mantienen más equilibradas y con menor variación, lo que sugiere un ambiente más estable tanto para la demanda como para la oferta. Los climas cálidos y muy cálidos presentan patrones más irregulares, pero en general tienden a mostrar precios más bajos durante el día y un pequeño incremento en la noche, probablemente porque el calor no afecta tanto la movilidad como el frío.



Conclusiones

El análisis permitió entender que la mayoría de los viajes son cortos y económicos, con precios concentrados entre 10 y 15 dólares, y que Uber domina en volumen de viajes frente a Lyft en todos los escenarios. También se evidenció que Lyft mantiene precios más altos, especialmente en trayectos largos, lo que sugiere una estrategia distinta de posicionamiento. Además, se confirmó que factores externos como el

clima y la hora del día influyen en la demanda y en los precios, aunque de manera moderada.

Con estos hallazgos, las plataformas podrían diseñar estrategias específicas para días nublados, como promociones o ajustes de precios que compensen la caída en la demanda. Lyft podría reconsiderar su política de precios en trayectos largos si busca competir más directamente con Uber. A nivel de negocio y planificación urbana, estos patrones sugieren que entender el contexto climático y horario es clave para anticipar la movilidad y mejorar la experiencia del usuario, además de optimizar la cobertura y disponibilidad de servicios.