# CODING BOOTCAMPS ESPOL:
# Data-Driven Decisions Specialist

## Python for Data Analytics

## Grupo 3

## Proyecto: RideFare

## Entregable: Avance


**Integrantes:**

Forero Villota Katherine Sheila

Heredia Villamar Kimberly Elizabeth

# Índice

# 1. Introducción

Analizamos dos conjuntos de datos relacionados con servicios de transporte y condiciones climáticas.

Se utilizaron herramientas de análisis de datos en Python, principalmente las siguientes librerías:

- Pandas
- Matplotlib
- Seaborn
- Numpy,

Esto último, con el fin de explorar la información, detectar patrones, identificar valores atípicos y comprender cómo se comportan las variables principales.

**Los datasets trabajados fueron:**

- PFDA_rides.csv → Información de viajes
- PFDA_weather.csv → Información del clima

# 2. Descripción de los Datos

## 2.1 Dataset de Viajes (rides)

Este dataset contiene 693,071 registros y 10 columnas, relacionados con viajes realizados en dos plataformas de movilidad: Uber y Lyft.

Columnas y tipos de datos:

- ✓ distance (float64): Distancia recorrida en cada viaje.
- ✓ cab_type (object): Empresa proveedora del servicio (Uber o Lyft).
- ✓ time_stamp (float64): Registro temporal del viaje.
- ✓ destination (object): Lugar al que se dirige el pasajero.
- ✓ source (object): Punto de inicio del viaje.
- ✓ price (float64): Costo del viaje.
- ✓ surge_multiplier (int64): Multiplicador aplicado por alta demanda u otros factores.
- ✓ id (object): Identificador único del viaje.
- ✓ product_id (object): Identificador del tipo de servicio.
- ✓ name (object): Nombre/categoría del viaje (UberX, Shared, Taxi, etc.).

Se pueden analizar precios, distancias, comportamiento por tipo de vehículo, zonas origen–destino o variaciones por demanda (surge_multiplier).

## 2.2 Dataset de Clima (weather)

Este dataset contiene 6,276 registros y 8 columnas, relacionadas con condiciones climáticas.

Columnas y tipos de datos:

- ✓ temp (float64): Temperatura registrada.
- ✓ location (object): Zona donde se tomó la medición.
- ✓ clouds (float64): Nivel de nubosidad.
- ✓ pressure (float64): Presión atmosférica.
- ✓ rain (float64): Nivel de lluvia (presenta varios valores nulos que se imputaron).
- ✓ time_stamp (int64): Registro temporal de la medición climática.
- ✓ humidity (float64): Nivel de humedad.
- ✓ wind (float64): Velocidad del viento.

Permite estudiar temperatura, lluvia, presión y otras variables ambientales para identificar patrones o posibles efectos sobre la demanda de viajes.

# 3. Exploración Inicial de los Datos

Se realizó una revisión preliminar de los datos para entender su estructura antes del análisis de su distribución.

## 3.1 Exploraciones del Dataset de Viajes

Se exploraron las principales variables categóricas. Esta revisión permitió identificar cuántas categorías existían en cada columna y cuáles eran sus valores. Se utilizaron funciones como nunique() para contar categorías y unique() para listarlas.

### *Tipos de Viajes*

Se analizó la columna name, que representa los tipos de viaje disponibles.

El resultado mostró que existen 13 tipos de viajes, entre ellos: Shared, Lux, UberX, UberPool, Taxi, Black SUV, entre otros.

Esto permite identificar la oferta de servicios disponibles dentro de la plataforma.

### *Lugares de Partida*

Se exploró la columna source para conocer desde dónde inician los viajes.

Se encontraron 12 puntos de partida, como Haymarket Square, Back Bay, North End, Fenway, South Station, etc.

Esta información ayuda a entender qué zonas generan mayor demanda.

### *Lugares de Destino*

De igual manera, se examinó la columna destination, obteniendo también 12 puntos de destino.

Las ubicaciones coinciden con áreas importantes de la ciudad, lo que sugiere rutas frecuentes entre zonas comerciales, residenciales y universitarias.

### *Empresas de Transporte*

Finalmente, se revisó la columna cab_type, identificando 2 empresas principales que ofrecen los servicios: Uber y Lyft.

Esto permite comparar comportamientos entre ambas compañías.

## 3.2 Exploraciones del Dataset de Clima

### *Lugares*

Se revisó la columna location del dataset weather utilizando nunique() y unique() para identificar los lugares donde se registran datos climáticos.

El resultado mostró 12 ubicaciones distintas, como Back Bay, Beacon Hill, Boston University, Fenway y otras.

Esta revisión permitió conocer las zonas donde se recopila la información del clima.

### *Análisis de Media en Todas las Variables Climáticas*

Además, se aplicó groupby("location").mean(numeric_only=True) para calcular la media de las variables numéricas por cada lugar. Este análisis mostró que, en general, las ubicaciones presentan condiciones climáticas muy similares en promedio.

# 4. Evaluación de Calidad de Datos

## 4.1 Evaluacion de calidad de datos del Dataset de Viajes

### *Duplicados*

Para verificar si el dataset contenía registros repetidos, se utilizó la función duplicated().sum() sobre el DataFrame rides.

El resultado fue 0 registros duplicados, lo que indica que todos los datos de viajes son únicos y no requieren limpieza en este aspecto.

### *Valores Nulos*

Se aplicó isnull().sum() para identificar la cantidad de valores faltantes en cada columna.

El análisis mostró que la columna price tenía 55,095 valores nulos, lo que representa aproximadamente 7.95% del total de registros.

El resto de las columnas no presentaron datos faltantes.

### *Imputación de Datos*

Para completar los valores faltantes en la columna price, se imputaron usando la mediana (más robusta) del precio por cada empresa (Uber o Lyft).

Este método es adecuado porque reduce el impacto de valores extremos.

Código utilizado:

*rides["price"] = rides.groupby("cab_type")["price"].transform(lambda x: x.fillna(x.median()))*

Tras aplicar la imputación, se verificó que ya no quedan valores nulos en el dataset.

### *Outliers*

Para identificar valores atípicos en las columnas numéricas del dataset (distance, time_stamp, price y surge_multiplier), se calcularon los cuartiles y el rango intercuartílico (IQR). Con estos valores se establecieron límites inferiores y superiores que permiten determinar qué registros se consideran outliers.

Los resultados obtenidos fueron:

- Distance: 8,662 outliers (1.25%).
- Time_stamp: 0 outliers (0%).
- Price: 5,589 outliers (0.81%).
- Surge_multiplier: 9,890 outliers (1.43%).

Estos outliers representan valores que se alejan del comportamiento normal del dataset y pueden corresponder a viajes inusualmente largos, precios muy altos o momentos de alta demanda. Detectarlos permite mejorar la calidad del análisis en etapas posteriores.

## 4.2 Evaluacion de calidad de datos del Dataset de Clima

### *Duplicados*

Se revisó si el dataset weather contenía registros repetidos.

El resultado mostró que no existen valores duplicados, por lo que no fue necesario realizar limpieza en este aspecto.

### *Valores Nulos*

Luego se verificó la presencia de datos faltantes.

Se encontró que la única columna con valores nulos es rain, con 5382 registros faltantes, lo que representa aproximadamente 85.76% del total.

Este porcentaje elevado sugiere que, en la mayoría de los casos, simplemente no hubo lluvia registrada en ese momento.

### *Imputación de Datos*

Antes de imputar, se comprobó si existían registros donde el valor de lluvia fuera 0, lo cual confirmaría que un valor nulo podría equivaler a ausencia de lluvia.

Al no encontrarse registros con lluvia igual a 0, se asumió que los valores faltantes representan momentos sin precipitación y, por ello, se imputaron reemplazándolos por 0.

Con esta imputación, el dataset ya no presenta valores nulos en la columna de lluvia.

### *Outliers*

Para detectar valores atípicos en el dataset weather, se analizaron las columnas numéricas: temperatura, presión, lluvia, humedad y viento. Se calcularon los cuartiles y el rango intercuartílico (IQR), lo que permitió establecer límites inferiores y superiores para definir qué valores se consideran outliers.

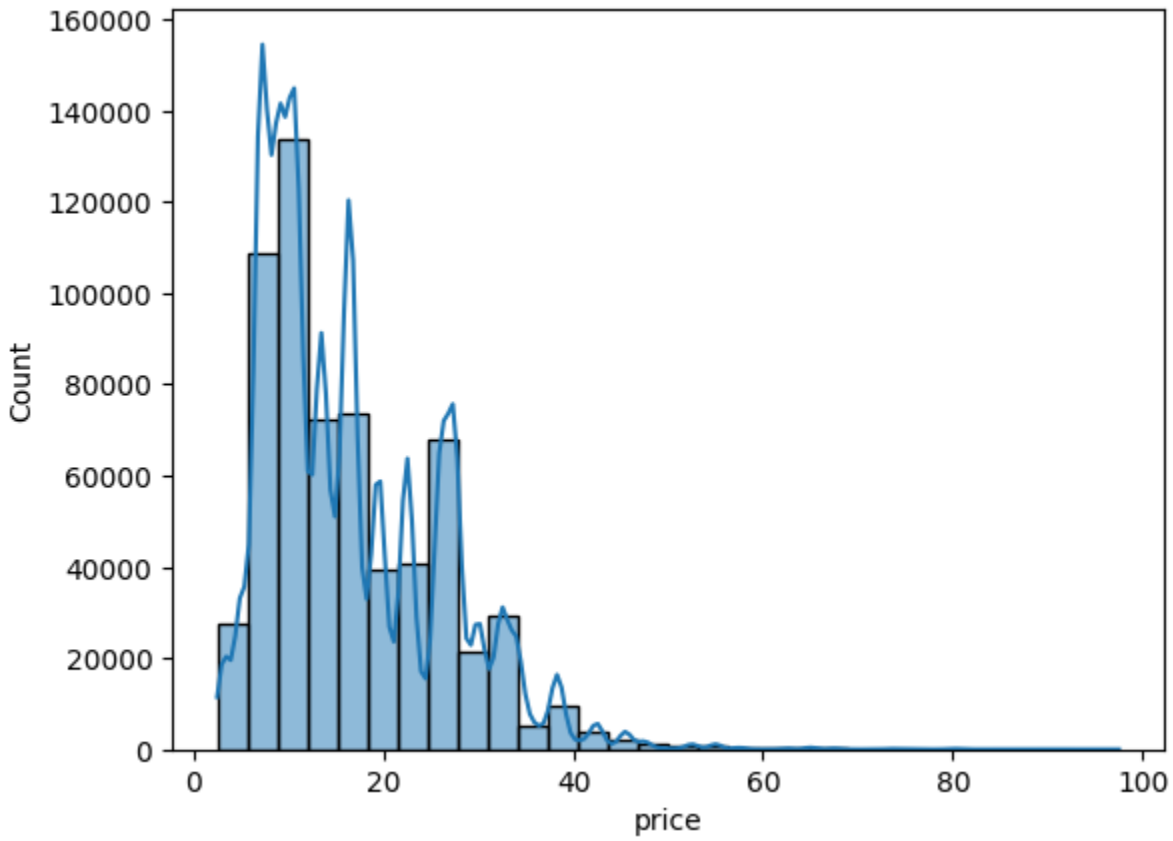
Los resultados mostraron lo siguiente:

- Temperatura: 266 outliers (4.24%).
- Presión: 0 outliers (0%).
- Lluvia: 800 outliers (12.75%).
- Humedad: 0 outliers (0%).
- Viento: 0 outliers (0%).

Estos valores atípicos reflejan mediciones inusuales del clima, especialmente en las variables de temperatura y lluvia. Su identificación permite evaluar si se deben tratar o conservar de acuerdo con el análisis posterior.

# 5. Análisis de Distribución de Variables

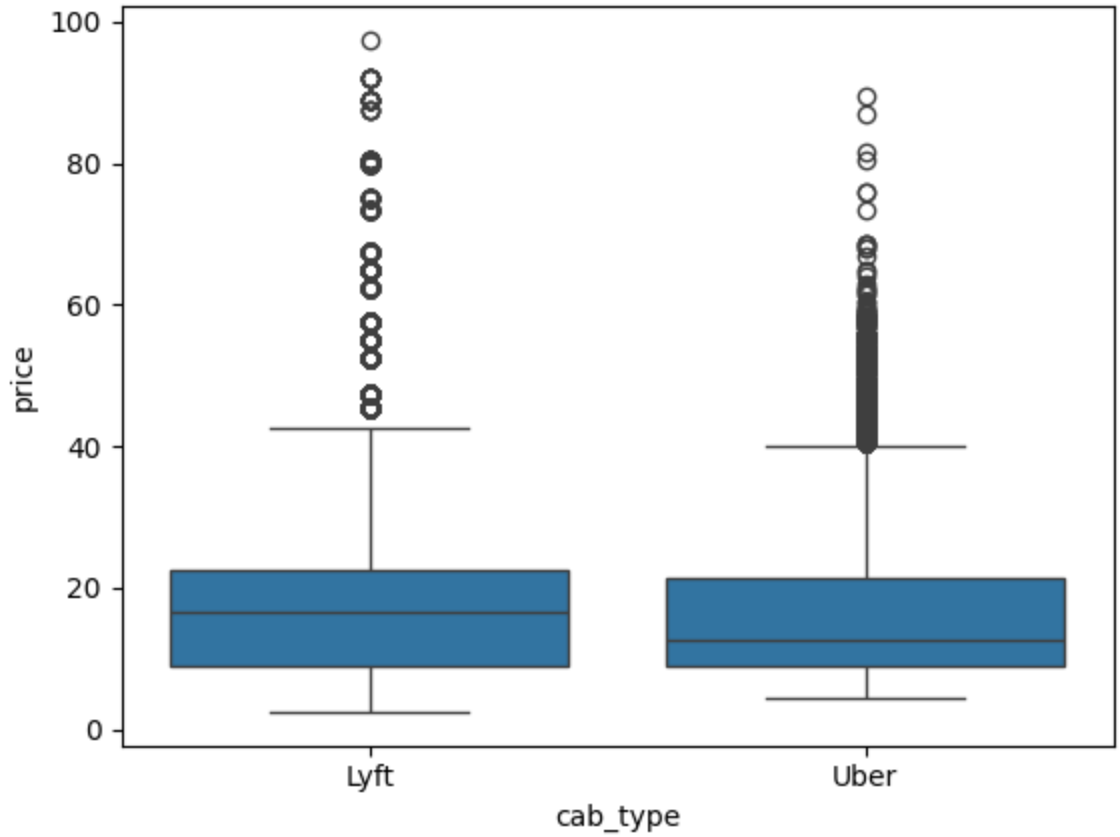
## 5.1 Histograma de la Variable Precio




El histograma muestra que la mayor parte de los precios de los viajes se encuentra concentrada en el rango de 0 a aproximadamente 25 dólares, mostrando una "cola" hacia la derecha, lo que indica una distribución sesgada positivamente. Esto significa que:

- La mayoría de los viajes tienen un precio relativamente bajo.
- Existe una proporción considerablemente menor de viajes con precios altos.
- El hecho de que haya precios dispersos y bajos pero también valores más amplios (con menor frecuencia) sugiere la presencia de grupos diferenciados en cuanto a tarifas, posiblemente ya sea por factores externos, servicios de transporte “premium” (Lux Black, Lux Black XL, Black SUV, etc), distancia del viaje, entre otros.

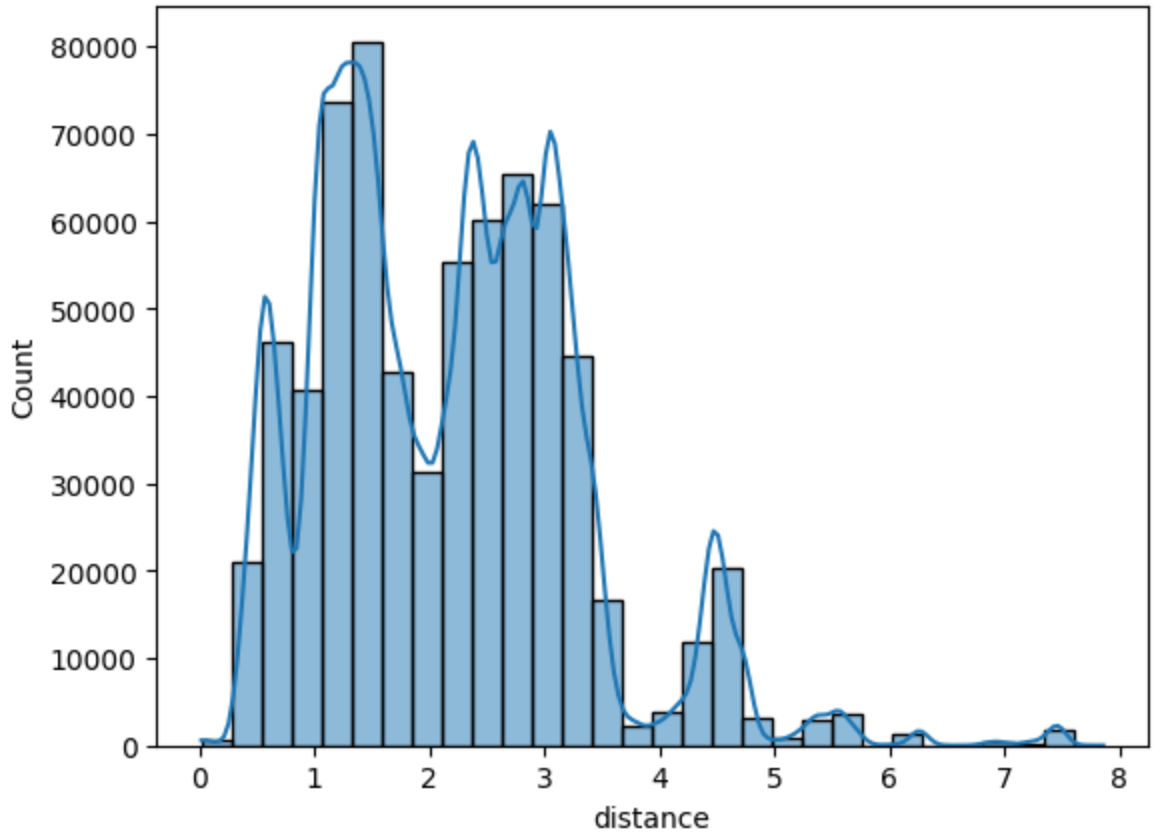
## 5.2 Diagrama de Cajas de la Variable Precio por Empresa de Viajes




Los boxplot nos permiten comparar los precios de viajes para Lyft y Uber por separado, permitiendo analizar las características y tendencias propias de cada empresa. Ambos servicios muestran precios concentrados en rangos bajos a medios, con algunos viajes considerablemente más caros identificados como valores atípicos.

- La mediana de precios en Uber es más baja que la de Lyft, lo que sugiere que en promedio los viajes de Uber tienden a ser menos costosos.
- Lyft muestra una mayor dispersión en sus precios; esto se observa en una caja más amplia (rango intercuartílico mayor), indicando que el costo de los viajes varía más en Lyft que en Uber.
- Ambos servicios tienen valores atípicos, pero Lyft presenta más casos extremos y algunos viajes llegan a precios considerablemente más altos que los experimentados en Uber. Por otro lado, no parecen afectar significativamente la distribución de los precios en general.

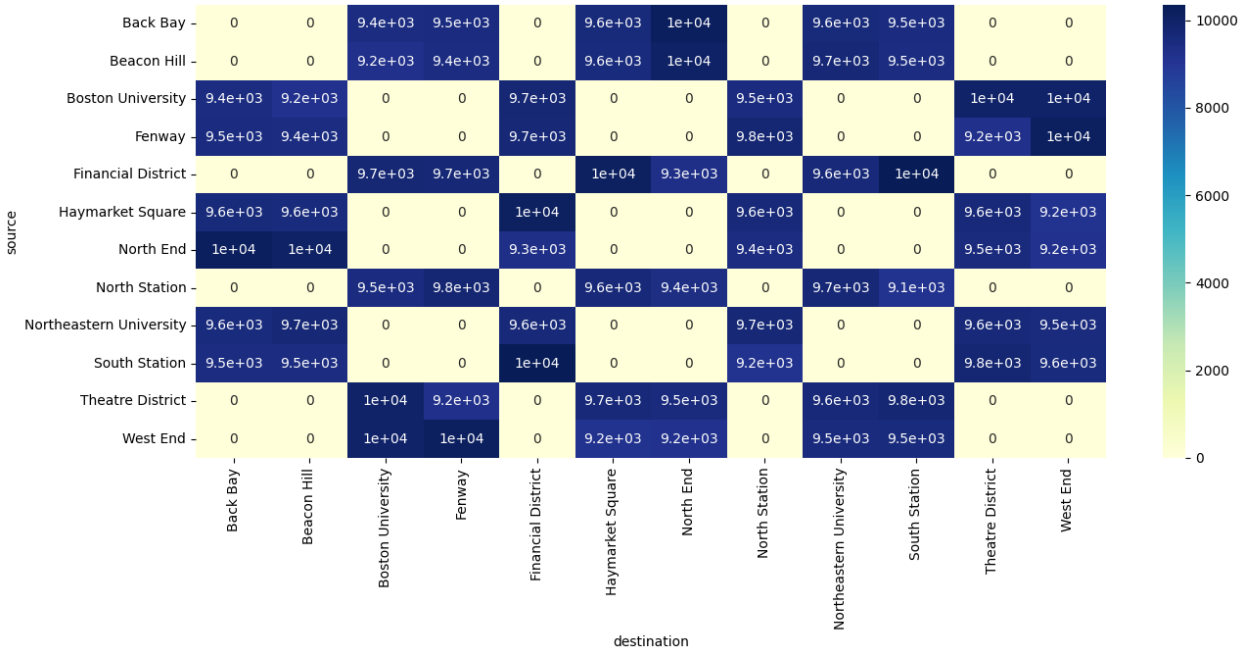
## 5.3 Histograma de la Variable Distancia




La mayor parte de los viajes registrados tienen distancias cortas, distribuyéndose principalmente entre 1 y 3 kilómetros. La frecuencia disminuye a medida que la distancia aumenta, reflejando que los trayectos largos son poco frecuentes. La distribución es asimétrica hacia la derecha (positiva), y aunque existen viajes de larga distancia, estos representan solo una pequeña fracción del total.

- La moda de la distribución está en el rango de 1 a 2 kilómetros, donde se agrupa la mayoría de los viajes, indicando alta demanda en trayectos urbanos o de corta duración.
- El énfasis en recorridos cortos puede orientar a las plataformas a optimizar recursos y tarifas, así como promover servicios específicos para viajes urbanos.
- La probabilidad de que un usuario solicite un viaje aumenta a menor distancia, por lo que el negocio se sostiene principalmente en trayectos cortos y eficientes.

## 5.4 Mapa de Calor de Frecuencias de Rutas




El mapa de calor muestra la frecuencia de viajes entre distintas combinaciones de origen y destino en una ciudad. No hay una ruta que sobresalga notablemente sobre las demás; en cambio, varias rutas populares tienen frecuencias muy similares, indicando que los usuarios distribuyen sus viajes de manera equilibrada entre diferentes zonas.

- El hecho de que ninguna ruta tenga una frecuencia que se destaque ampliamente sobre el resto sugiere un uso diversificado del servicio; es decir, no existe una ruta "estrella" que absorba la mayor parte de la demanda, sino que los flujos de viajes se reparten equitativamente entre los puntos principales de la ciudad.
- Los orígenes y destinos populares corresponden generalmente a sectores clave urbanos: universidades, estaciones de transporte y zonas residenciales, lo cual indica que el servicio responde bien a las necesidades cotidianas de movilidad de los usuarios.
- Dado que las frecuencias son similares y hay una demanda amplia en varias rutas, las empresas de transporte pueden diseñar estrategias que aseguren cobertura eficiente en todos los trayectos principales, sin necesidad de priorizar en exceso unas rutas sobre otras.

# 6. Propuestas de Análisis

## 6.1 Análisis Temporal y Patrones de Demanda a lo Largo del Día y la Semana

Este análisis busca identificar cómo varían la demanda de viajes y los precios en las plataformas Uber y Lyft en distintos momentos del día y días de la semana. Permite descubrir fenómenos como horas pico, días de mayor actividad y posibles diferencias estacionales o vinculadas a eventos especiales.

### *Variables Principales Para el Análisis*

- Hora del día (viene de time_stamp)
- Fecha (viene de time_stamp)
- Empresa del viaje (cab_type)
- Precio del viaje (price)
- Multiplicador aplicado al precio (surge_multiplier)

### *Preguntas que se Busca Responder Luego del Análisis*

- ¿En qué horas y días se concentran la mayor cantidad de viajes?
- ¿Cómo fluctúan los precios durante el día y la semana?
- ¿Existen diferencias en los patrones diarios o semanales entre Uber y Lyft?
- ¿Se observan picos de demanda asociados a horarios laborales, nocturnos o de eventos masivos?

## 6.2 Análisis del Impacto Climático en la Demanda y el Precio de los Viajes

Este análisis examina cómo las condiciones climáticas (lluvia, temperatura, nieve, etc.) afectan tanto la cantidad de viajes solicitados como los precios cobrados por los servicios de ridesharing. Busca revelar hasta qué punto el clima puede modificar patrones de movilidad y política de precios adaptable.

### *Variables Principales Para el Análisis*

- Temperatura (temp)
- Nubosidad (clouds)
- Precipitación (rain)
- Precio del viaje (price)
- Hora y día (time_stamp)
- Multiplicador aplicado al precio (surge_multiplier)

### *Preguntas que se Busca Responder Luego del Análisis*

- ¿Aumenta la demanda de viajes en días lluviosos o con mal tiempo?
- ¿Se incrementan los precios promedio bajo condiciones climáticas adversas?
- ¿Qué tipo de clima produce los mayores cambios en la demanda y el precio?
- ¿La sensibilidad al clima es diferente entre Uber y Lyft?