Análisis de viajes en taxi – Zuber

Introducción

Zuber, una nueva empresa de viajes compartidos, quiere analizar los viajes en taxi realizados en Chicago durante noviembre de 2017 para entender:

- Las empresas de taxi más activas.
- Los barrios con mayor número de llegadas.
- El impacto del clima en la duración de los viajes.

Este análisis se basa en datos extraídos de una base de datos y archivos CSV proporcionados. Además, se probará una hipótesis sobre cómo las condiciones climáticas afectan los viajes hacia el aeropuerto en un día específico.

Tabla de contenidos

1. Carga y revisión de datos

Importamos los archivos CSV con información sobre las empresas de taxis y los barrios donde finalizaron los viajes. Estos datos fueron obtenidos previamente mediante consultas SQL.

Antes de analizarlos, verificamos que:

• Los archivos se hayan cargado correctamente.

• Las columnas tengan los tipos de datos adecuados.

• No haya valores nulos que interfieran con el análisis.

Esta revisión nos permite asegurarnos de que los DataFrames están listos para

aplicar filtros, agrupaciones y visualizaciones.

In [1]:
```
import pandas as pd

# Importamos los datasets
df_companies = pd.read_csv('../Data/moved_project_sql_result_01.csv')
df_neighborhoods = pd.read_csv('../Data/moved_project_sql_result_04.cs

# Mostrar información general sobre los DataFrames
df_companies.info()
print()
df_neighborhoods.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64 entries, 0 to 63
Data columns (total 2 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   company_name  64 non-null     object
 1   trips_amount  64 non-null     int64 
dtypes: int64(1), object(1)
memory usage: 1.1+ KB

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 94 entries, 0 to 93
Data columns (total 2 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   dropoff_location_name  94 non-null     object 
 1   average_trips          94 non-null     float64
dtypes: float64(1), object(1)
memory usage: 1.6+ KB
```

2. Análisis de barrios por viajes finalizados

Los datos están completos y tienen el tipo de dato correcto. En este paso vamos a identificar los **10 barrios de Chicago con mayor número promedio de viajes finalizados** durante noviembre de 2017. Para eso, usamos el archivo 'project_sql_result_04.csv', que contiene:

- **dropoff_location_name:** nombre del barrio donde terminó el viaje
- **average_trips:** promedio de viajes finalizados por barrio

Este análisis es importante para conocer las **zonas de mayor demanda**, lo cual puede dar pistas útiles para **optimización de flotas, campañas de marketing o diseño de rutas estratégicas**.

Vamos a ordenar los datos de forma descendente por número de viajes promedio

y luego seleccionar los primeros diez.

In [2]:
```
# Identificamos los 10 barrios principales por finalización del recorr
top_neighborhoods = df_neighborhoods.sort_values(by='average_trips', a

# Mostramos el resultado
top_neighborhoods
```

Out[2]:

	dropoff_location_name	average_trips
0	Loop	10727.466667
1	River North	9523.666667
2	Streeterville	6664.666667
3	West Loop	5163.666667
4	O'Hare	2546.900000
5	Lake View	2420.966667
6	Grant Park	2068.533333
7	Museum Campus	1510.000000
8	Gold Coast	1364.233333
9	Sheffield & DePaul	1259.766667

3. Análisis de empresas de taxi

Ahora vamos a identificar las **10 empresas de taxis con mayor número de viajes registrados** durante los días 15 y 16 de noviembre de 2017. Usaremos el archivo `'project_sql_result_01.csv'`, que contiene:

- `company_name`: nombre de la empresa de taxis
- `trips_amount`: número total de viajes realizados por la empresa el 15 y 16 de noviembre de 2017

Este análisis nos permite conocer qué empresas tuvieron mayor participación en el mercado durante esos días. Con estos datos, podemos **observar tendencias** que podrían estar relacionadas con cobertura, reputación o promociones específicas.

Vamos a ordenar los datos en orden descendente por número de viajes y luego seleccionaremos las 10 primeras empresas.

In [3]:
```
# Identificamos las 10 empresas con más viajes
top_companies = df_companies.sort_values(by='trips_amount', ascending=
```

```
# Mostramos el resultado
top_companies
```

Out[3]:

	company_name	trips_amount
0	Flash Cab	19558
1	Taxi Affiliation Services	11422
2	Medallion Leasin	10367
3	Yellow Cab	9888
4	Taxi Affiliation Service Yellow	9299
5	Chicago Carriage Cab Corp	9181
6	City Service	8448
7	Sun Taxi	7701
8	Star North Management LLC	7455
9	Blue Ribbon Taxi Association Inc.	5953

4. Gráficos de resultados

Ya que identificamos los 10 barrios con mayor número promedio de viajes finalizados y las 10 empresas de taxis con mayor número de viajes durante los días 15 y 16 de noviembre, ahora vamos a visualizar estos datos con gráficos de barras para comparar de forma clara y rápida los valores, detectar patrones y facilitar la interpretación de los datos. En este caso:

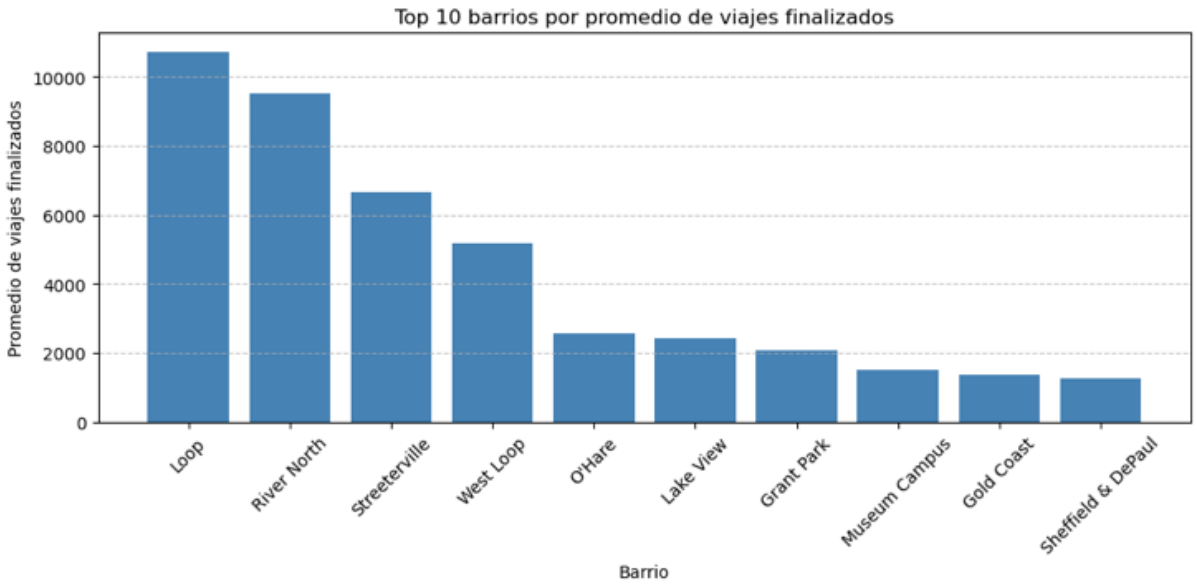
- La primera gráfica mostrará los **barrios más populares como destino final**.
- La segunda gráfica mostrará las **empresas de taxis más activas** en esos días.

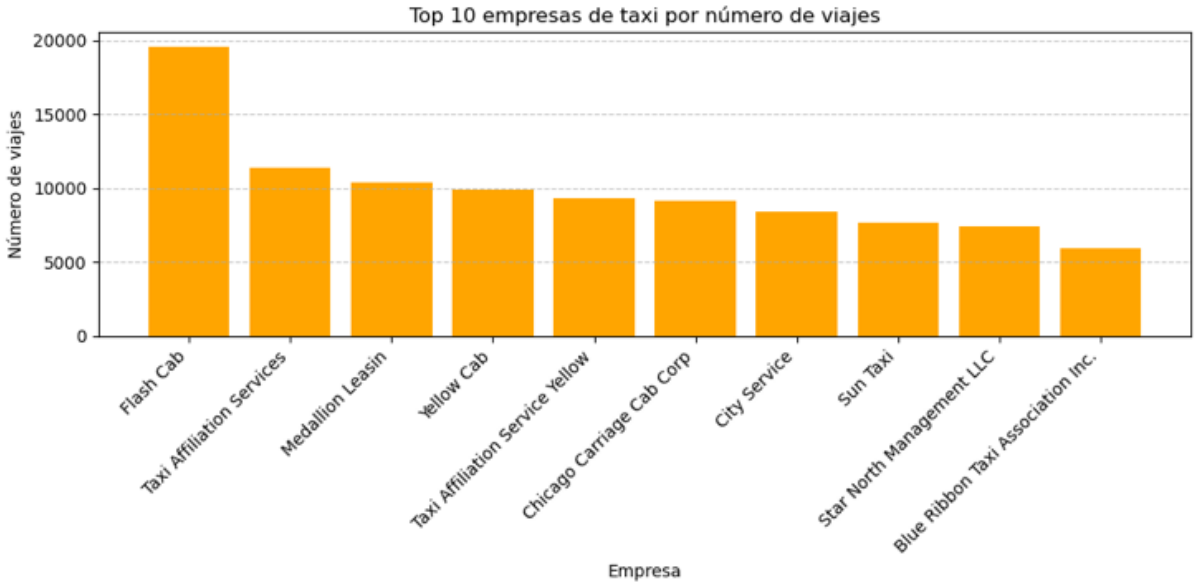
In [4]:

```
import matplotlib.pyplot as plt

# Gráfica 1: Barrios más populares
plt.figure(figsize=(10, 5))
plt.bar(top_neighborhoods['dropoff_location_name'], top_neighborhoods[
plt.title('Top 10 barrios por promedio de viajes finalizados')
plt.xlabel('Barrio')
plt.ylabel('Promedio de viajes finalizados')
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

```
# Gráfica 2: Empresas más activas
plt.figure(figsize=(10, 5))
plt.bar(top_companies['company_name'], top_companies['trips_amount'],
plt.title('Top 10 empresas de taxi por número de viajes')
plt.xlabel('Empresa')
plt.ylabel('Número de viajes')
plt.xticks(rotation=45, ha='right')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```





Conclusión parcial

Los gráficos permiten visualizar más facilmente los datos obtenidos:

- **Loop**, **River North** y **Streeterville** destacan como los barrios más populares como destino final de los viajes.
- **Flash Cab** y **Taxi Affiliation Services** fueron las empresas con mayor

número de viajes registrados durante los días 15 y 16 de noviembre de 2017.

Con esta información identificamos zonas de alta concentración de viajes y empresas líderes en volumen de servicio. Esto será útil más adelante para interpretar si el impacto del clima afecta de alguna forma diferente a ciertos factores.

5. Prueba de hipótesis

Vamos a analizar si las condiciones climáticas afectan la duración de los viajes en taxi desde **Loop** hacia el **Aeropuerto O'Hare**, específicamente los **sábados** de noviembre de 2017.

> Ya filtramos estos viajes y los combinamos con los datos del clima en pasos anteriores. Ahora vamos a trabajar directamente con esa tabla (`project_sql_result_07.csv`), que ya incluye el campo `weather_conditions` .

Objetivo

Determinar si existe una diferencia significativa en la duración promedio de los viajes entre días con **buen clima (Good)** y **mal clima (Bad)**.

Hipótesis

- **Hipótesis nula (H_0):** No hay diferencia en la duración promedio entre climas buenos y malos.
- **Hipótesis alternativa (H_1):** Sí hay una diferencia significativa en la duración promedio entre climas buenos y malos.

Metodología

1. Cargamos los datos de viajes ya filtrados y clasificados por clima.
2. Dividimos la duración de los viajes en dos grupos según `weather_conditions` .
3. Revisamos estadísticas descriptivas de cada grupo (mínimo, máximo, media, etc.).
4. Aplicamos una **prueba de hipótesis (t-test de dos muestras)** para determinar si la diferencia en duraciones es estadísticamente significativa.

Este análisis nos ayuda a identificar si el clima tiene un impacto tangible en los tiempos de traslado, lo cual puede ser útil para planificación operativa, estimaciones de viaje y decisiones estratégicas para empresas de transporte.

In [5]:
```
from scipy import stats

# 1. Cargar el archivo y revisar su estructura.
df = pd.read_csv('../Data/moved_project_sql_result_07.csv')

# Revisar si hay valores ausentes que puedan afectar
print(df.isna().sum())

# Separar los viajes por clima
good_weather = df[df['weather_conditions'] == 'Good']['duration_second
bad_weather = df[df['weather_conditions'] == 'Bad']['duration_seconds'
```

```
start_ts              0
weather_conditions    0
duration_seconds      0
dtype: int64
```

In [6]:
```
# 2. Análisis descriptivo
# Revisar estadísticas básicas de cada grupo
print("Viajes con buen clima:")
print(good_weather.describe())

print("\nViajes con mal clima:")
print(bad_weather.describe())
```

```
Viajes con buen clima:
count     888.000000
mean     1999.675676
std       759.198268
min         0.000000
25%      1389.750000
50%      1800.000000
75%      2460.000000
max      7440.000000
Name: duration_seconds, dtype: float64

Viajes con mal clima:
count     180.000000
mean     2427.205556
std       721.314138
min       480.000000
25%      1962.000000
50%      2540.000000
75%      2928.000000
max      4980.000000
Name: duration_seconds, dtype: float64
```

In [7]:
```
# 3. Prueba estadística e interpretación

# Prueba de hipótesis: t-test
alpha = 0.05
t_stat, p_val = stats.ttest_ind(good_weather, bad_weather, equal_var=F
```

```
print("\nResultados del t-test:")
print()
print("Estadístico t:", t_stat)
print("p-valor:", p_val)


# Interpretación
if p_val < alpha:
    print("Rechazamos la hipótesis nula. El clima afecta significativa
else:
    print("No se puede rechazar la hipótesis nula. No hay evidencia su
```

```
Resultados del t-test:

Estadístico t: -7.186034288068629
p-valor: 6.738994326108734e-12
Rechazamos la hipótesis nula. El clima afecta significativamente la dur
ación de los viajes.
```

6. Conclusiones y siguientes pasos

Conclusión del análisis

Durante noviembre de 2017, encontramos que:

- Los **barrios con más viajes finalizados** fueron Loop, River North y Streeterville, lo que sugiere zonas de alta atracción (por trabajo, turismo o conexión).
- Las **empresas más activas** fueron Flash Cab y Taxi Affiliation Services.
- Al comparar viajes desde Loop al Aeropuerto O'Hare en sábados, descubrimos que **las condiciones climáticas sí afectan significativamente la duración del trayecto**.

> En promedio, un viaje en clima malo dura **21.3% más** que uno en clima bueno (de 1999 a 2427 segundos).

Esto sugiere que el clima debe considerarse al **planear tiempos estimados de viaje** o **optimizar rutas**.

Siguientes pasos sugeridos

1. **Modelos predictivos**: Incorporar variables como clima, hora y día para anticipar la duración de los viajes con mayor precisión.

2. **Optimización de flota**: Aunque inicialmente se pensó en reforzar destinos con más finalizaciones, sería más útil ubicar taxis donde **empiezan los**

viajes, ya que eso activa la demanda.

3. **Análisis de la competencia**: Observar el comportamiento de las empresas más activas (Flash Cab, etc.) para identificar si operan con mejor cobertura, tarifas, reputación u horarios específicos.

4. **Identificación de corredores estratégicos**:

    - Analizar los tramos más frecuentes entre barrios (por ejemplo, Loop → O'Hare).
    - Implementar **estrategias de marketing específicas** o **rutas compartidas con tarifa reducida** en estos corredores.
        - Ejemplo: promociones grupales o servicios tipo shuttle en rutas de alta demanda para mejorar rentabilidad por kilómetro.