d31b4867-8870-44f1-a923-d728048022d7-Copy1

November 6, 2024

1 SPRINT 8 SQL

2 Índice

2.1 Introducción

Este proyecto tiene como objetivo analizar los datos de viajes en taxi en Chicago con el fin de identificar patrones de comportamiento de los pasajeros y comprender el impacto de factores externos, como las condiciones climáticas, en la demanda y duración de los viajes. A través del análisis de una base de datos que incluye información sobre viajes, taxis, barrios y condiciones meteorológicas, se buscará responder a preguntas clave como: ¿Cuáles son las empresas de taxi más populares? ¿Cómo afecta la lluvia a la duración de los viajes entre el Loop y el Aeropuerto O'Hare? Los resultados de este estudio servirán como base para la toma de decisiones estratégicas en el sector del transporte compartido.

2.2 Analisis de datos

[47]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import scipy.stats as stats
```

[48]:
```
df_cant_viajes = pd.read_csv ('/datasets/project_sql_result_01.csv')
df_viajes_chicago = pd.read_csv ('/datasets/project_sql_result_04.csv')
df_hipotesis = pd.read_csv("/datasets/project_sql_result_07.csv")
```

[49]:
```
df_cant_viajes.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
```

[50]:
```
duplicados_cant = df_cant_viajes.duplicated().sum()
print(f"Hay {duplicados} filas duplicadas en el DataFrame.")
```

```
Hay 0 filas duplicadas en el DataFrame.
```

[51]:
```
df_viajes_chicago.info()
```

```
<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
```

[52]:
```
duplicados = df_viajes_chicago.duplicated().sum()
print(f"Hay {duplicados} filas duplicadas en el DataFrame.")

print(df_viajes_chicago.duplicated(subset=['dropoff_location_name']).sum())
print(df_viajes_chicago.duplicated(subset=['average_trips']).sum())
```

```
Hay 0 filas duplicadas en el DataFrame.
0
0
```

[53]:
```
df_viajes_chicago = df_viajes_chicago.round(2)
print(df_viajes_chicago.head())
```

```
  dropoff_location_name  average_trips
0                  Loop       10727.47
```

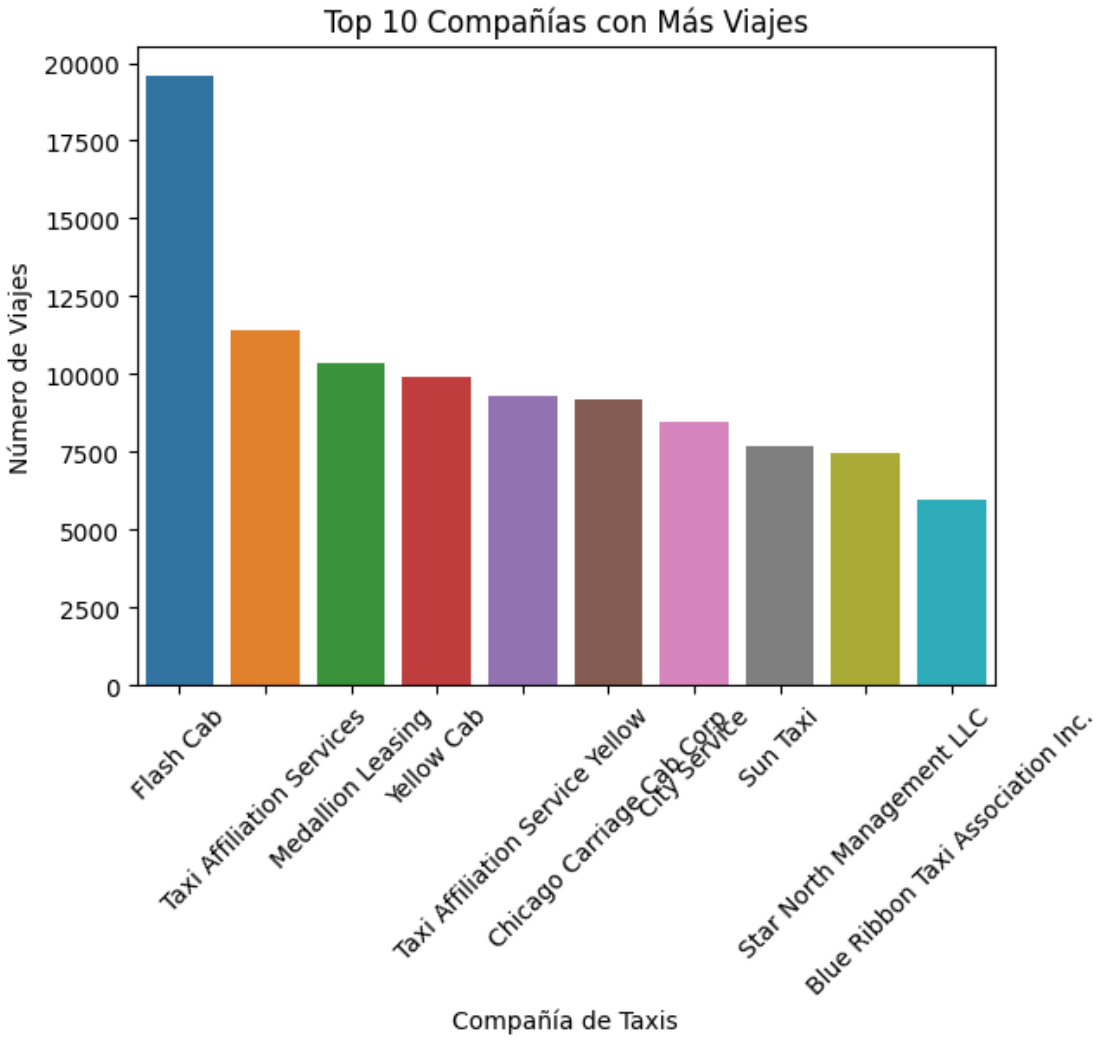
```
1           River North        9523.67
2         Streeterville        6664.67
3             West Loop        5163.67
4                O'Hare        2546.90
```

[54]:
```
df_ordenado = df_viajes_chicago.sort_values(by='average_trips', ascending=False)

top_10_barrios = df_ordenado.head(10)


print(top_10_barrios[['dropoff_location_name', 'average_trips']])
```

```
  dropoff_location_name  average_trips
0                  Loop       10727.47
1           River North        9523.67
2         Streeterville        6664.67
3             West Loop        5163.67
4                O'Hare        2546.90
5             Lake View        2420.97
6            Grant Park        2068.53
7         Museum Campus        1510.00
8            Gold Coast        1364.23
9    Sheffield & DePaul        1259.77
```

2.3 Graficos

[55]:
```
top_cinco_compañias = df_cant_viajes.sort_values(by='trips_amount',␣
↪ascending=False).head(10)

sns.barplot(x='company_name', y='trips_amount', data=top_cinco_compañias)
plt.xticks(rotation=45)
plt.xlabel('Compañía de Taxis')
plt.ylabel('Número de Viajes')
plt.title('Top 10 Compañías con Más Viajes')
plt.show()
```



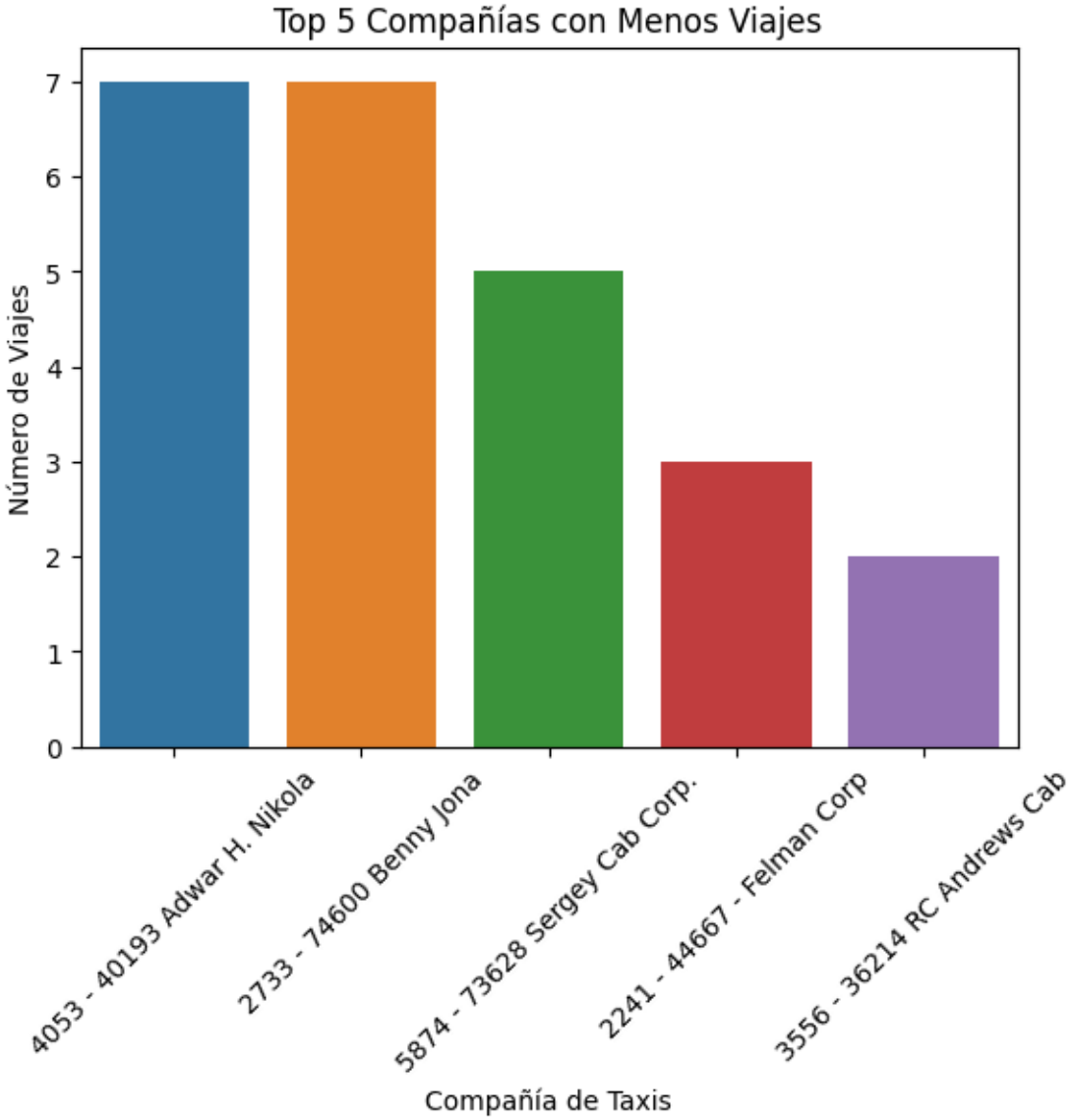
La compañía “Flash Cab” es la que cuenta con un número significativamente mayor de viajes en comparación con las demás. Esto sugiere que tiene una mayor cuota de mercado o es preferida por los usuarios por diversas razones podria ser la cobertura, precios, etc.

Las siguientes cuatro compañías (“Taxi affiliation services”, “Medallion leasing”, “Yellow Cab” y “Taxi afiiliantion service yellow”) presentan números de viajes bastante similares, lo que indica una competencia. Existe una clara diferencia entre el líder del mercado (“Flash Cab”) y el resto de las compañías. Esto podría deberse a factores como una estrategia de marketing más efectiva, una mayor flota de vehículos o una mejor cobertura geográfica

Es importante tener en cuenta el período de tiempo al que corresponden estos datos ya que las tendencias podrían variar si analizamos datos de otros meses o años.

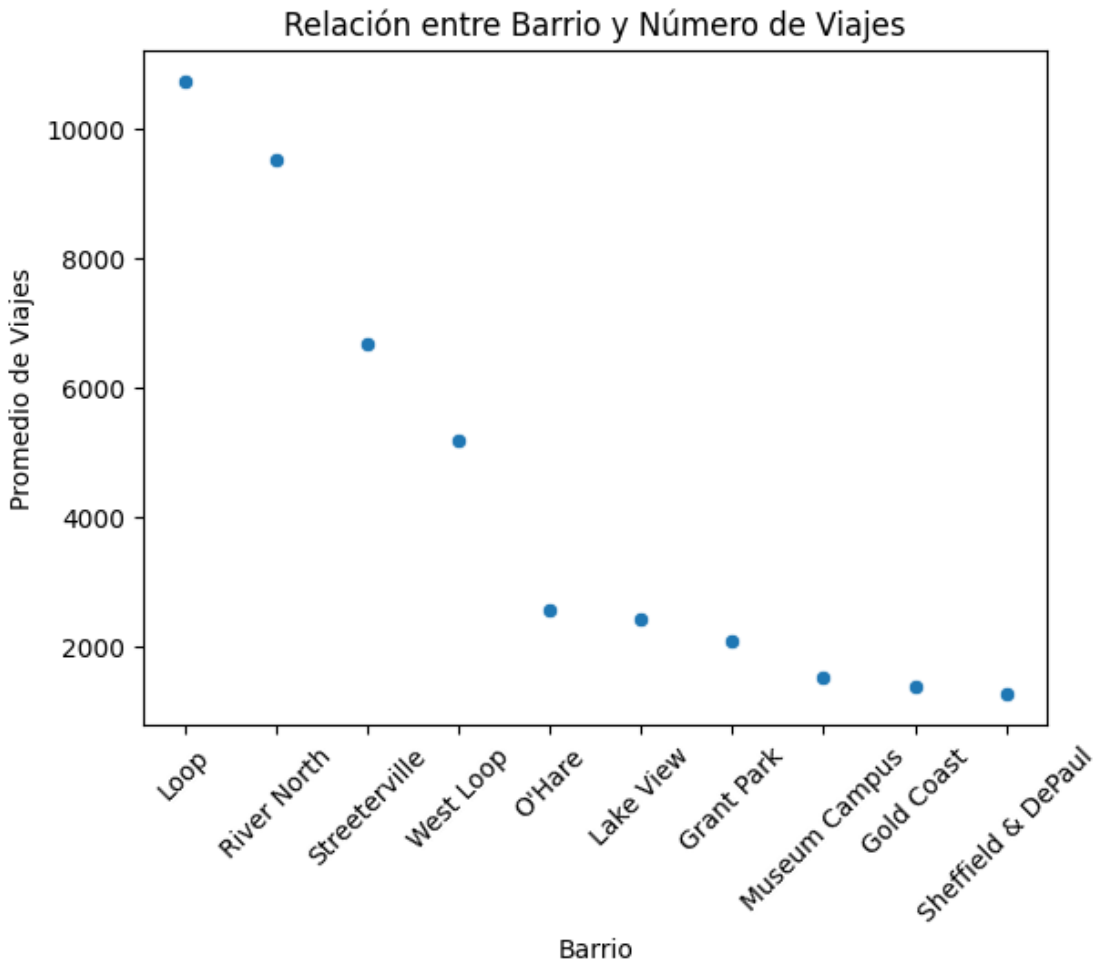
[56]:
```
compañias_con_menos_viajes = df_cant_viajes.sort_values(by='trips_amount',␣
↪ascending= False).tail(5)
```

```
sns.barplot(x='company_name', y='trips_amount', data=compañias_con_menos_viajes)
plt.xticks(rotation=45)
plt.xlabel('Compañía de Taxis')
plt.ylabel('Número de Viajes')
plt.title('Top 5 Compañías con Menos Viajes')
plt.show()
```



De igual forma que el grafico anterior, podemos concluir que estas empresas tuvieron menor cantidad de viajes durante el 15 y 16 de noviembre de 2017. Esto pudo deberse al costo de las trarifas, disponibilidad de conductores, etc.

[57]:
```
sns.scatterplot(x='dropoff_location_name', y='average_trips',␣
  ↪data=top_10_barrios)
plt.xticks(rotation=45)
plt.xlabel('Barrio')
plt.ylabel('Promedio de Viajes')
plt.title('Relación entre Barrio y Número de Viajes')
plt.show()
```

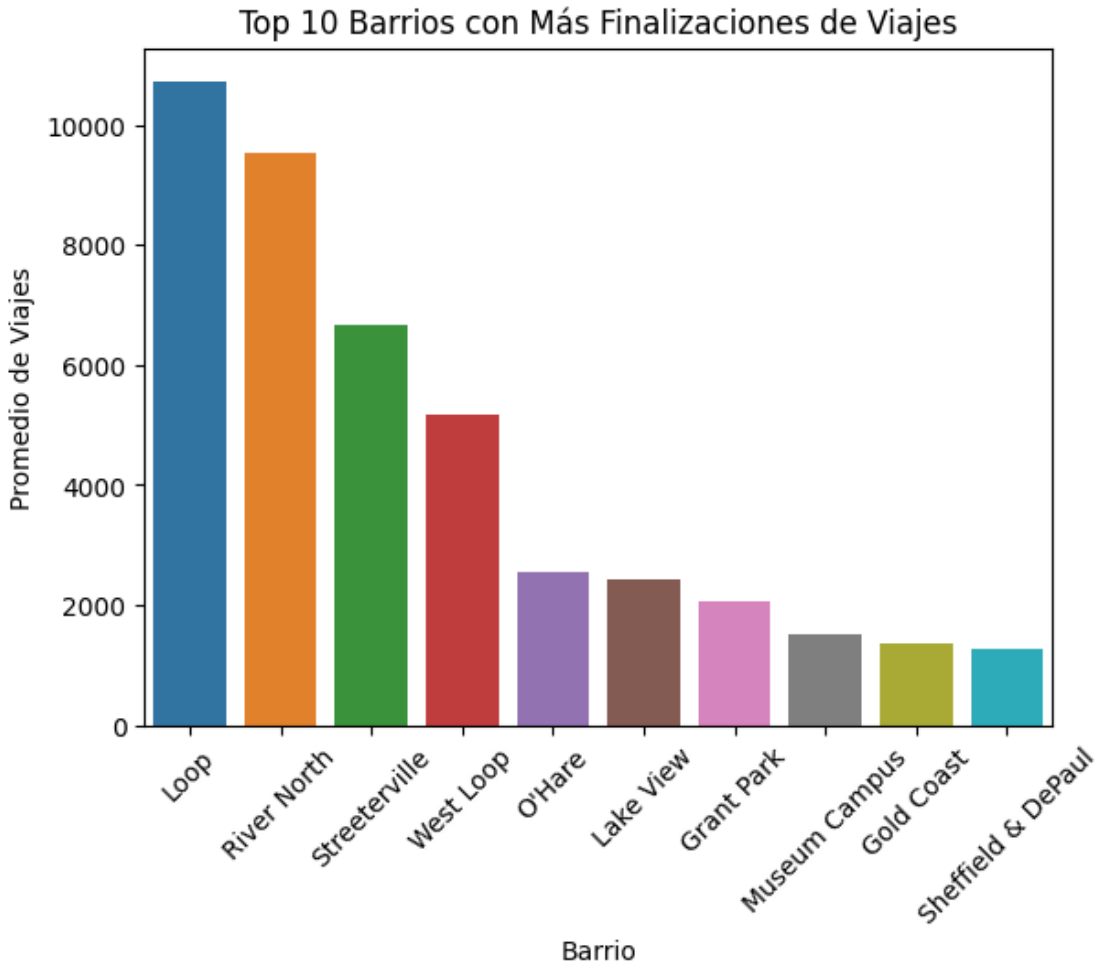


Existe una variabilidad significativa en el número promedio de viajes entre los diferentes barrios. Algunos barrios concentran un número mucho mayor de viajes en comparación con otros. Esto sugiere que factores como la densidad poblacional, la disponibilidad de transporte público, la presencia de puntos de interés (centros comerciales, zonas turísticas, etc.) y la conectividad vial pueden influir significativamente en la demanda de viajes en cada barrio.

Este grafico nos permite observar que los barrios como “Loop” son aquellos donde se registra un mayor número de viajes en promedio. Estos barrios podrían ser considerados como “hotspots” de movilidad urbana. Es importante tener presente, que estos son datos solo para dos dias del mes,

asi que, se debe tomar en consideracion que tipo de eventos ocurrieron en los diferentes barrios.

[58]:
```
sns.barplot(x='dropoff_location_name', y='average_trips', data=top_10_barrios)
plt.xticks(rotation=45)
plt.xlabel('Barrio')
plt.ylabel('Promedio de Viajes')
plt.title('Top 10 Barrios con Más Finalizaciones de Viajes')
plt.show()
```



En Chicago en noviembre del 2017, podemos observar claramente que existe una gran disparidad en el número promedio de viajes que finalizan en cada barrio, donde en el barrio "Loop" se finalizaron 10700 viajes en promedio y para el barrio "Sheffield & DePaul" se realizaron 1200 viajes. La diferencia, sin duda es significativa.

Un pequeño grupo de barrios acaparan la mayor parte de los viajes. Esto sugiere que estos barrios tienen ciertas características que los hacen más atractivos para los usuarios de servicios de transporte.

Es probable que estos barrios tengan una mayor concentración de habitantes, lo que aumenta la demanda de transporte.La presencia de centros comerciales, restaurantes, lugares de entretenimiento, etc., atrae a un mayor número de personas. podria deberse a que es facil el desplazamientoa otras zonas de la ciudad.

Por otro lado los barrios con menor cantidad de viajes podría deberse a factores como zonas residenciales más dispersas o con menor número de habitantes, escasa oferta de comercios, restaurantes o lugares de interés o dificultades para llegar a estos barrios debido a una mala conexión vial o al transporte público.

2.4 Prueba de hipotesis

Hipotesis nula (H0) = Los sábados lluviosos no afectan la duración promedio de los viajes entre el Loop y el Aeropuerto O'Hare

Hipotesis alternativa (H1) = Los sábados lluviosos sí afectan la duración promedio de los viajes entre el Loop y el Aeropuerto O'Hare

Explicacion del planteamiento de hipotesis : la hipótesis nula asume que no hay diferencia significativa en la duración de los viajes, independientemente de las condiciones climáticas y la hipotesis alternativa plantea lo contrario, es decir, que si afecta. Siempre partimos de la idea de que algo no ocurre (hipótesis nula) y luego buscamos evidencia para rechazarla y aceptar la hipótesis alternativa.

[59]:
```
df_hipotesis.info()
```

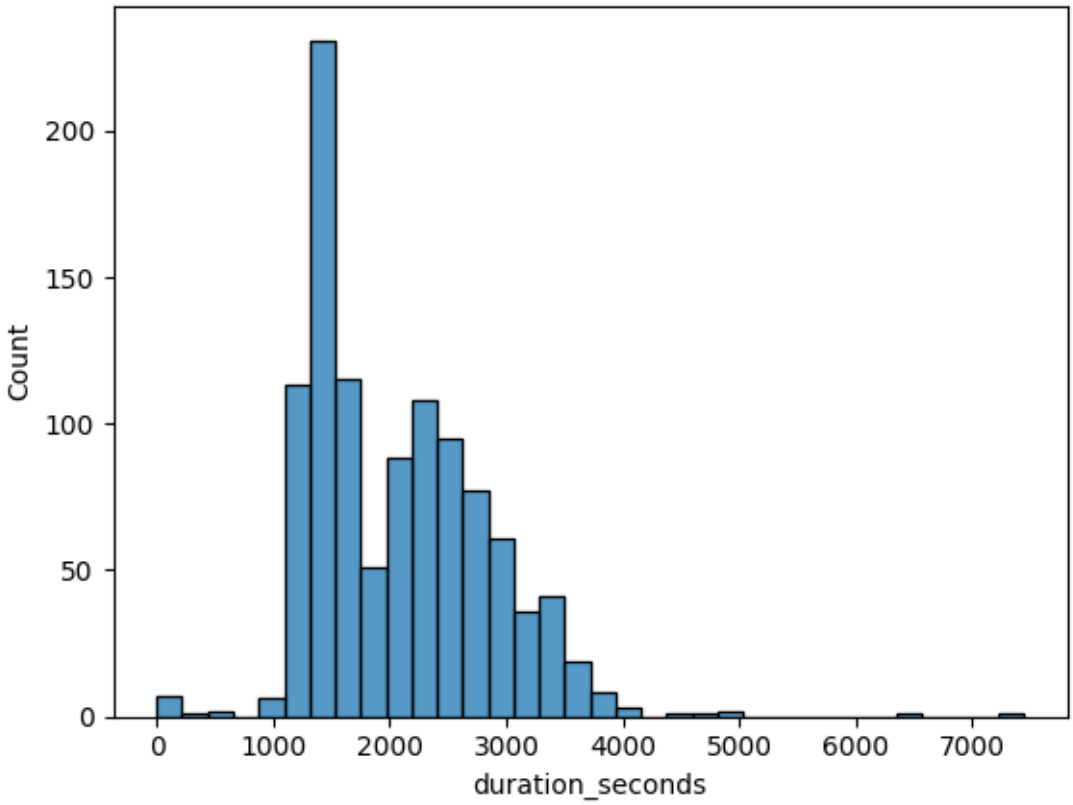
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1068 entries, 0 to 1067
Data columns (total 3 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   start_ts            1068 non-null   object 
 1   weather_conditions  1068 non-null   object 
 2   duration_seconds    1068 non-null   float64
dtypes: float64(1), object(2)
memory usage: 25.2+ KB
```

[60]:
```
df_hipotesis['start_ts'] = pd.to_datetime(df_hipotesis['start_ts'])
df_hipotesis["duration_seconds"] = df_hipotesis["duration_seconds"].astype(int)

print(df_hipotesis.head())
print(df_hipotesis["weather_conditions"].unique())
```

```
             start_ts weather_conditions  duration_seconds
0 2017-11-25 16:00:00               Good              2410
1 2017-11-25 14:00:00               Good              1920
2 2017-11-25 12:00:00               Good              1543
3 2017-11-04 10:00:00               Good              2512
4 2017-11-11 07:00:00               Good              1440
['Good' 'Bad']
```

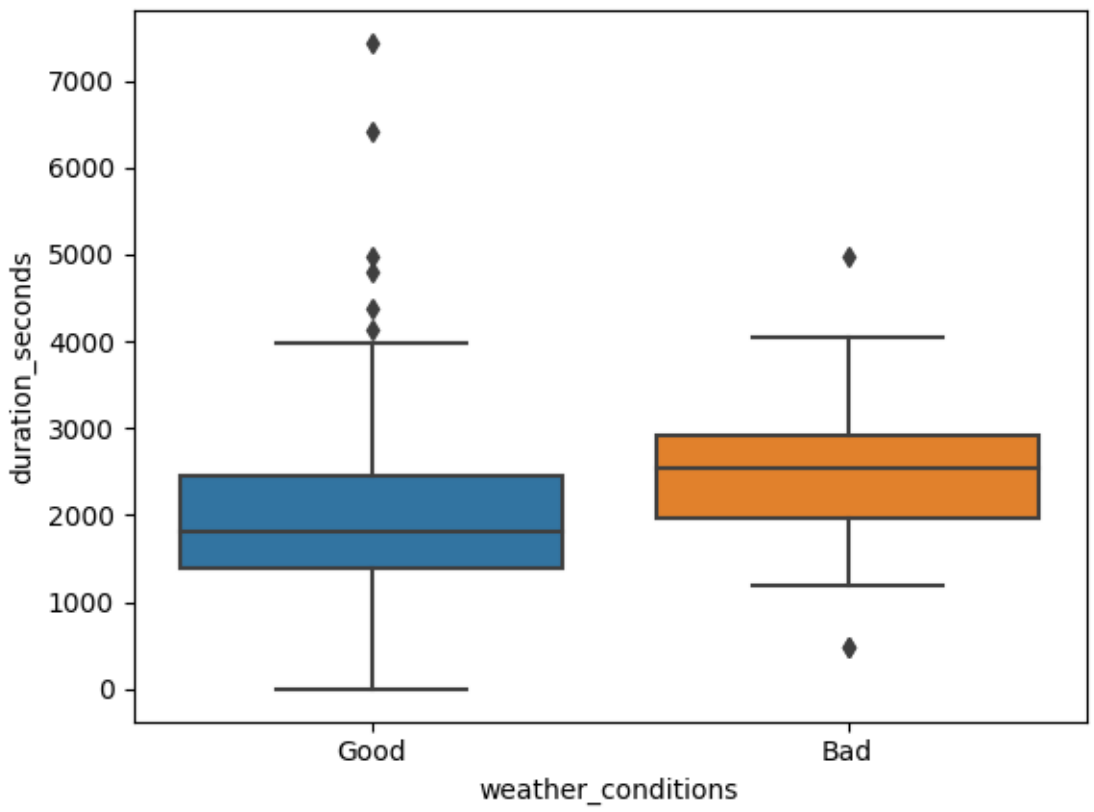
[61]:
```
# Distribución de la duración de los viajes
sns.histplot(data=df_hipotesis, x='duration_seconds')
```

[61]: <AxesSubplot:xlabel='duration_seconds', ylabel='Count'>



[62]:
```
# Duración promedio de los viajes en diferentes condiciones climáticas
sns.boxplot(x='weather_conditions', y='duration_seconds', data=df_hipotesis)
```

[62]: <AxesSubplot:xlabel='weather_conditions', ylabel='duration_seconds'>



[63]:
```
df_hipotesis['sabado_lluvioso'] = ((df_hipotesis['start_ts'].dt.dayofweek == 5)␣
  ↪& (df_hipotesis['weather_conditions'] == 'Bad'))
print(df_hipotesis.sample(10))
```

```
                start_ts weather_conditions  duration_seconds  sabado_lluvioso
704  2017-11-04 19:00:00               Good              2400            False
913  2017-11-11 08:00:00               Good              1200            False
479  2017-11-25 19:00:00               Good              2280            False
855  2017-11-11 10:00:00               Good              1500            False
931  2017-11-25 17:00:00               Good              2220            False
226  2017-11-18 06:00:00               Good              1440            False
730  2017-11-11 13:00:00               Good              2280            False
714  2017-11-11 12:00:00               Good              1747            False
132  2017-11-18 15:00:00               Good              3480            False
119  2017-11-04 14:00:00               Good              3300            False
```

[64]:
```
sabados_lluviosos =␣
  ↪df_hipotesis[df_hipotesis['sabado_lluvioso']]['duration_seconds']
otros_dias = df_hipotesis[~df_hipotesis['sabado_lluvioso']]['duration_seconds']
```

[65]:
```
stat, p = stats.levene(sabados_lluviosos, otros_dias)

print('Estadístico de Levene:', stat)
print('p-valor de Levene:', p)

if p > 0.05:
    results = stats.ttest_ind(sabados_lluviosos, otros_dias, equal_var=True)
else:
    results = stats.ttest_ind(sabados_lluviosos, otros_dias, equal_var=False)

print('valor p (prueba t):', results.pvalue)

alpha = 0.05
if results.pvalue < alpha:
    print("Rechazamos la hipótesis nula. Los sábados lluviosos sí afectan la␣
 ↪duración promedio de los viajes.")
else:
    print("No podemos rechazar la hipótesis nula. Los sábados lluviosos no␣
 ↪parecen afectar significativamente la duración promedio de los viajes.")
```

```
Estadístico de Levene: 0.38853489683656073
p-valor de Levene: 0.5332038671974493
valor p (prueba t): 6.517970327099473e-12
Rechazamos la hipótesis nula. Los sábados lluviosos sí afectan la duración
promedio de los viajes.
```

Este valor p es muchísimo menor que nuestro nivel de significancia (alpha). Por lo tanto, rechazamos la hipótesis nula y concluimos que sí existe una diferencia estadísticamente significativa en la duración promedio de los viajes entre los sábados lluviosos y otros días.En otras palabras, los viajes realizados en sábados lluviosos tienden a ser significativamente más largos en comparación con los viajes realizados en otros días.

Las condiciones climáticas, específicamente los días sabados lluviosos, son un factor determinante en la duración de los viajes en el área estudiada.

Al planificar viajes o realizar análisis de movilidad, es fundamental considerar variables climáticas como la lluvia, ya que pueden afectar significativamente los tiempos de desplazamiento. mucho mas cuando se trata de viajes hacia el aeropuerto, que requieren de una hora puntual de llegada.

Algunas recomendaciones es que las empresas de transporte pueden aumentar sus tarifas, por la alta demanda, otra recomendacion es mantener a los conductores informados de las zonas con mayor fluencia de clientes, informar los tiempos de espera a los clientes, etc.

[66]:
```
df_hipotesis['duracion_minutos'] = df_hipotesis['duration_seconds'] / 60

promedio_sabados = df_hipotesis[df_hipotesis['start_ts'].dt.dayofweek ==␣
 ↪5]['duracion_minutos'].mean().round(2)
```

```
print("El promedio de duración de los viajes los sábados en minutos es:",␣
 ↪promedio_sabados)
```

```
El promedio de duración de los viajes los sábados en minutos es: 34.53
```

Se uso una prueba-t para probar la hipotesis porque estamos comparando la duración promedio de los viajes en dos grupos distintos: sábados lluviosos y otros días

2.5 Conclusion general

"Flash Cab" es la compañía que mas destaca como líder indiscutible en el mercado, lo que sugiere una estrategia comercial sólida y una alta preferencia por parte de los usuarios.

La demanda de servicios de taxi varía significativamente entre los diferentes barrios de Chicago, lo que refleja una distribución desigual de la población, actividades económicas y factores sociodemográficos.Sin embargo, los eventos especiales pueden generar picos de demanda en ciertos barrios y horarios. Las condiciones climáticas adversas, como la lluvia, aumentan la demanda de taxis y pueden prolongar los tiempos de viaje, por esto las empresas de taxis deben buscar formas de diferenciarse de la competencia, ya sea a través de precios más competitivos, servicios adicionales o una mejor experiencia al cliente.

Los datos analizados revelan una compleja interacción entre factores geográficos, sociodemográficos y climáticos que influyen en la demanda de servicios de taxi en Chicago. Las empresas de taxis que logren comprender y adaptarse a estas dinámicas tendrán una mayor probabilidad de éxito en este mercado competitivo