Portafolio - Proyecto 3

April 20, 2025

1 Impact of Weather on Ride Duration – Zuber Taxi Data (Chicago)

Victor Uriel Leyva
Analista de Datos Jr.

1.0.1 Objetivo del Proyecto:

Analizar los patrones de viaje de una empresa ficticia de transporte compartido (Zuber) en Chicago para entender el comportamiento de los pasajeros, estudiar a los competidores y evaluar el impacto de las condiciones climáticas —específicamente la lluvia— en la duración de los viajes entre Loop y el Aeropuerto Internacional O'Hare durante los sábados de noviembre de 2017.

1.0.2 Herramientas utilizadas:

- SQL (consultas para análisis preliminar)
- Python
- pandas
- matplotlib
- scipy.stats
- numpy

1.0.3 Principales habilidades aplicadas:

- Limpieza e integración de datos de múltiples fuentes
- Análisis exploratorio utilizando SQL y Python
- Agrupación y comparación de datos por empresa y ubicación
- Prueba de hipótesis con métodos estadísticos
- Visualización de resultados y validación con p-valor

1.1 Descripción del proyecto

En este proyecto, trabajo como analista para Zuber, una nueva empresa de viajes compartidos que se está lanzando en Chicago. Mi tarea es identificar patrones en los datos disponibles para comprender las preferencias de los pasajeros y el impacto de factores externos en los viajes. Analizaré una base

de datos para estudiar a los competidores y evaluar una hipótesis sobre cómo el clima influye en la frecuencia de los viajes.

1.2 Descripción de los datos

Cuento con una base de datos con información sobre viajes en taxi en Chicago, compuesta por las siguientes tablas: - **neighborhoods**: detalles sobre los barrios de la ciudad. - **cabs**: información sobre los taxis. - **trips**: registros de los viajes. - **weather_records**: datos sobre el clima.

1.3 Plan de trabajo para completar el proyecto

Paso 1. Se extrajeron los datos sobre el clima en Chicago en noviembre de 2017 desde el siguiente sitio web haciendo uso del lenguaje Python: https://practicum-content.s3.us-west-1.amazonaws.com/data-analyst-eng/moved_chicago_weather_2017.html

Paso 2. Análisis exploratorio de datos (SQL) - Se obtuvo el número de viajes en taxi por empresa entre el 15 y 16 de noviembre de 2017, ordenados en orden descendente por número de viajes. - Se calculó la cantidad de viajes entre el 1 y 7 de noviembre de 2017 para empresas cuyo nombre contiene "Yellow" o "Blue". Los resultados fueron agrupados por company_name. - Se analizaron los viajes de las empresas Flash Cab y Taxi Affiliation Services, las más populares en noviembre de 2017. Se agruparon los viajes del resto de las compañías bajo la categoría "Other", mostrando los resultados en orden descendente por trips_amount.

Paso 3. Prueba de hipótesis. Se evaluó si la duración de los viajes entre Loop y el Aeropuerto Internacional O'Hare varía los sábados lluviosos. - Se recuperaron los identificadores de los barrios Loop y O'Hare. - Se clasificaron las condiciones climáticas por hora en dos categorías: - "Bad": si la descripción incluía "rain" o "storm". - "Good": en los demás casos. - Se seleccionaron los viajes que comenzaron en Loop y finalizaron en O'Hare un sábado, junto con su duración y las condiciones climáticas correspondientes. - Se descartaron los viajes sin datos climáticos disponibles.

Paso 4. Análisis exploratorio de datos (Python)

Después de obtener nuestras consultas, terminamos con dos archivos de tipo ".csv". Los cuales contienen los siguientes datos: - project_sql_result_01.csv: - company_name: nombre de la empresa de taxis - trips_amount: el número de viajes de cada compañía de taxis el 15 y 16 de noviembre de 2017. - project_sql_result_04.csv: - dropoff_location_name: barrios de Chicago donde finalizaron los viajes - average_trips: el promedio de viajes que terminaron en cada barrio en noviembre de 2017.

[1]:
```
# Como primer paso importaré las librerias necesarias para trabajar en nuestro
↪código
import pandas as pd
from matplotlib import pyplot as plt
from scipy import stats as st
import numpy as np
import math as mt
```

[2]:
```
# Cargaré, limpiaré y analizaré los datos para poder formular nuestra prueba de
↪hipótesis
```

```
data_01 = pd.read_csv('project_sql_result_01.csv')
data_02 = pd.read_csv('project_sql_result_04.csv')
```

Se identificarán valores ausentes y duplicados para tratarlos adecuadamente, asegurando la coherencia y homogeneidad de los datos.

[3]:
```
# Para la limpieza de datos, se analizarán muestras de los DataFrames y los tipos de datos de sus columnas

print(data_01.dtypes)
print()
print(data_02.dtypes)
display(data_01.sample(10))
display(data_02.sample(10))
```

```
company_name    object
trips_amount     int64
dtype: object

dropoff_location_name     object
average_trips            float64
dtype: object
```

	company_name	trips_amount
33	Metro Jet Taxi A	146
44	2092 - 61288 Sbeih company	27
57	Metro Group	11
34	Norshore Cab	127
11	Globe Taxi	4383
13	Nova Taxi Affiliation Llc	3175
37	1469 - 64126 Omar Jada	36
40	6574 - Babylon Express Inc.	31
54	2192 - 73487 Zeymane Corp	14
47	4615 - 83503 Tyrone Henderson	21

	dropoff_location_name	average_trips
87	Pullman	3.896552
79	Fuller Park	8.166667
84	West Pullman	6.466667
12	Little Italy, UIC	863.700000
8	Gold Coast	1364.233333
60	New City	22.933333
15	Garfield Ridge	745.400000
22	Lincoln Square	356.733333
20	Rush & Division	395.533333
28	Wicker Park	182.600000

Nuestros datos parecen tener un tipo de dato adecuado y no presentan valores ausentes. Verifiquémoslo y busquemos posibles valores duplicados

[4]:
```
print(data_01.isna().sum())
print()
print(data_02.isna().sum())
```

```
company_name    0
trips_amount    0
dtype: int64

dropoff_location_name    0
average_trips            0
dtype: int64
```

[5]:
```
print(data_01.duplicated().sum())
print()
print(data_02.duplicated().sum())
```

```
0

0
```

Nuestros datos parecen estar limpios y completos, lo que nos proporciona una base de datos consistente y lista para el análisis.

Continuando con nuestro análisis exploratorio de datos, voy a identificar los 10 principales barrios por finalización de recorrido.

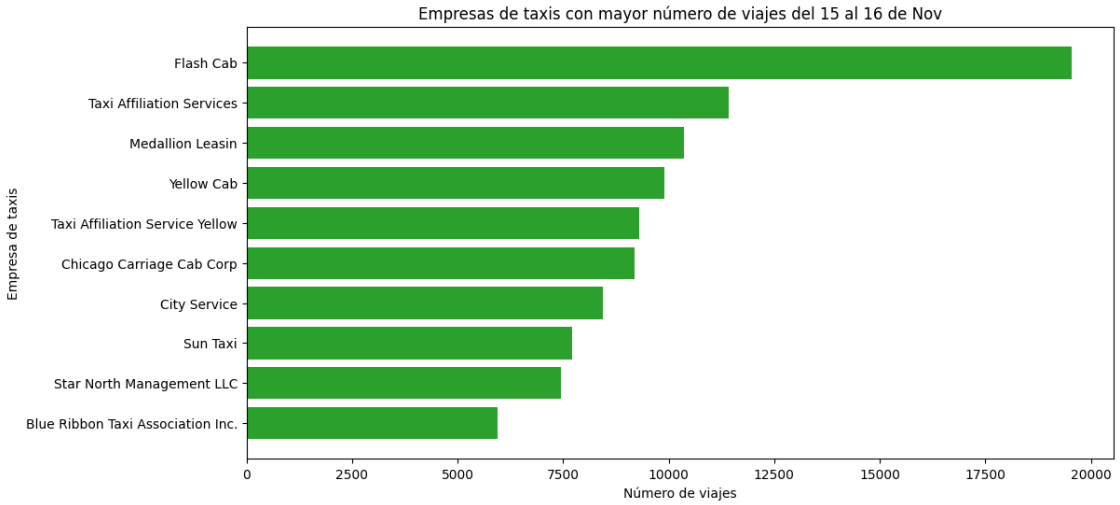
[6]:
```
df_top_10 = data_02.sort_values(by='average_trips', ascending=False).head(10)
display(df_top_10)
```

	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

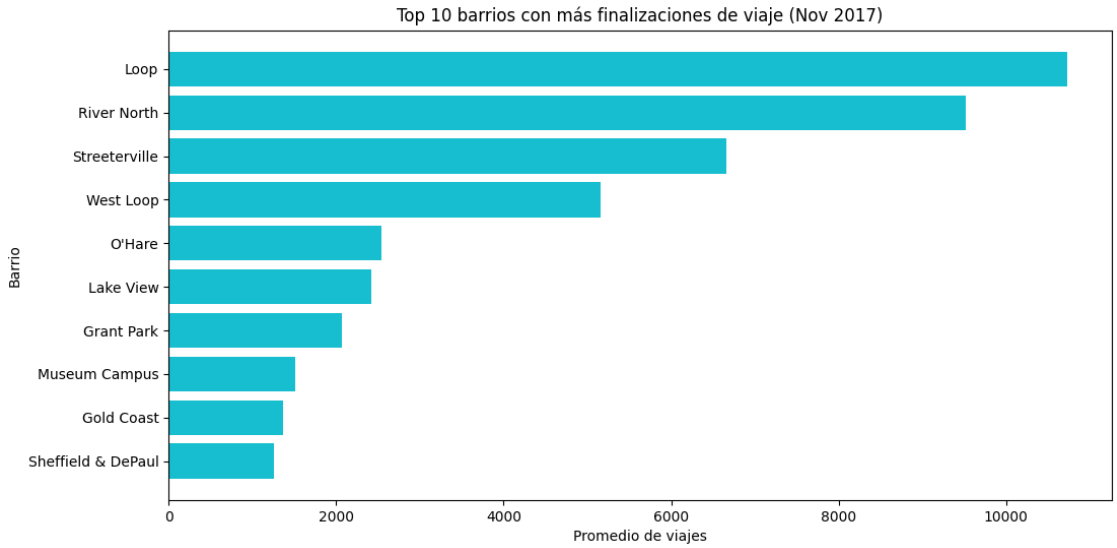
Parece ser que el barrio de Loop suele ser el que más viajes registró en 2017. Para complementar nuestros hallazgos vamos a graficar las empresas de taxis con un mayor número de viajes, así como los resultados que acabamos de obtener.

[7]:
```
# Para el gráfico de empresas de taxis con mayor número de viajes del 15 al 16␣
↪de Noviembre de 2017

df_taxis_top_10 = data_01.sort_values(by='trips_amount', ascending=False).
↪head(10)
```

```
plt.figure(figsize=(12,6))
plt.barh(df_taxis_top_10['company_name'], df_taxis_top_10['trips_amount'],␣
 ↪color='tab:green')
plt.xlabel('Número de viajes')
plt.ylabel('Empresa de taxis')
plt.title('Empresas de taxis con mayor número de viajes del 15 al 16 de Nov')
plt.gca().invert_yaxis() # Se invertirá el eje para mejor visualización
plt.show()
```



[8]:
```
# Para el gráfico de los 10 barrios principales por número de finalizaciones

plt.figure(figsize=(12,6))
plt.barh(df_top_10['dropoff_location_name'], df_top_10['average_trips'],␣
 ↪color='tab:cyan')
plt.xlabel('Promedio de viajes')
plt.ylabel('Barrio')
plt.title('Top 10 barrios con más finalizaciones de viaje (Nov 2017)')
plt.gca().invert_yaxis()
plt.show()
```



¡Qué interesante! A través de nuestros gráficos, hemos identificado la empresa con el mayor número de viajes registrados y el barrio con más finalizaciones de trayecto desde el Aeropuerto Internacional O'Hare.

Los resultados muestran que **FlashCab** fue la empresa de taxis con el mayor número de viajes registrados entre el 15 y 16 de noviembre de 2017. Por otro lado, el barrio de **Loop** fue el destino más frecuente para los viajes que partieron desde O'Hare durante noviembre de 2017.

A partir de estas conclusiones, podemos avanzar con nuestra prueba de hipótesis para analizar cómo los factores externos, como la lluvia, pueden influir en las preferencias de los pasajeros.

Paso 5. Puebas de hipótesis (Python).

Ahora que conocemos las preferencias de los clientes, realizaremos una prueba de hipótesis para analizar si los días lluviosos afectan la duración de los viajes.

Durante nuestro análisis exploratorio de datos, obtuvimos un tercer archivo .csv, que contiene información sobre viajes desde el barrio Loop hasta el Aeropuerto Internacional O'Hare, la ruta más frecuente en noviembre de 2017. Este archivo incluye las siguientes columnas: - start_ts: fecha y hora de recogida - weather_conditions: condiciones climáticas en el momento en el que comenzó el viaje - duration_seconds: duración del viaje en segundos

Con estos datos, formularemos y probaremos la siguiente hipótesis:

- Hipótesis nula (H0): No hay diferencia en la duración promedio de los viajes entre sábados lluviosos y sábados sin lluvia.
- Hipótesis alternativa (H1): La duración promedio de los viajes es diferente en sábados lluviosos.

[15]:
```
# Primero cargaré los datos
data_03 = pd.read_csv('project_sql_result_07.csv')
```

```
# Filtraré los datos en dos conjuntos uno para los días lluviosos y otro para␣
 ↪los demás días
rainy = data_03[data_03['weather_conditions'] == "Bad"]
clear = data_03[data_03['weather_conditions'] != "Bad"]

array01 = rainy['duration_seconds']
array02 = clear['duration_seconds']

# Realizaré una prueba de Levene para comprobar el supuesto de igualdad de␣
 ↪varianzas
alpha = 0.05

levene_test = st.levene(array01, array02, center='median')

if levene_test.pvalue < alpha:
    print('Los grupos **no** tienen varinzas iguales')
else:
    print('Los grupos tienen varianzas iguales')
```

```
Los grupos tienen varianzas iguales
```

[17]:
```
# Probaré si nuestra hipótesis nula se cumple o se rechaza

results = st.ttest_ind(array01, array02, equal_var=True)

print('valor p:', results.pvalue)

if results.pvalue < alpha:
    print('Rechazamos la hipótesis nula')
else:
    print('No rechazamos la hipótesis nula')
```

```
valor p: 6.517970327099473e-12
Rechazamos la hipótesis nula
```

1.4 Conclusión

Rechazamos la hipótesis nula, lo que sugiere que existe evidencia estadísticamente significativa de que los días lluviosos afectan la duración de los viajes en taxi durante los sábados de noviembre de 2017. El valor p obtenido fue muy bajo, lo que refuerza esta conclusión. Este resultado podría explicarse por el tamaño de la muestra o por la diferencia entre medias, aunque ya hemos verificado que las medias de los datos son muy similares. No obstante, esto podría seguir respaldando la idea de que las condiciones meteorológicas, específicamente la lluvia, tienen un impacto en la duración de los viajes.