69989e16-903b-41d4-ac0c-ac8b47fe5a7f

April 5, 2025

1 Analysis of Taxi Rides in Chicago and the Impact of Weather on Trip Duration

2 Análise de Corridas de Táxi em Chicago e o Impacto do Clima na Duração das Viagens

[3]:
```
import pandas as pd
import seaborn as sns
import math
from matplotlib import pyplot as plt
import numpy as np
from scipy import stats as st
```

[4]:
```
df_trips = pd.read_csv('/datasets/project_sql_result_01.csv')
df_location = pd.read_csv('/datasets/project_sql_result_04.csv')
```

[5]:
```
df_trips.info()
print()
df_location.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
```

Na tabela df_trips, a coluna 'company_name' armazena os nomes das empresas de táxi como texto, enquanto 'trips_amount' registra a quantidade de corridas em números inteiros.

Já na tabela df_location, a coluna 'dropoff_location_name' guarda os nomes dos bairros onde as corridas terminaram, também em formato de texto. A coluna 'average_trips' contém valores decimais que representam as médias de viagens para cada local.

In the df_trips table, the 'company_name' column stores taxi company names as text, while 'trips_amount' records the number of trips as whole numbers.

In the df_location table, the 'dropoff_location_name' column contains the names of neighborhoods where trips ended, also in text format. The 'average_trips' column holds decimal values representing trip averages for each location.

[6]:
```
df_trips.head(10)
```

[6]:

	company_name	trips_amount
0	Flash Cab	19558
1	Taxi Affiliation Services	11422
2	Medallion Leasing	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

[7]:
```
df_location.head(10)
```

[7]:

	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

[8]:
```
df_top10_location = df_location.sort_values(by = 'average_trips', ascending =␣
↪False).head(10)
df_top10_location
```

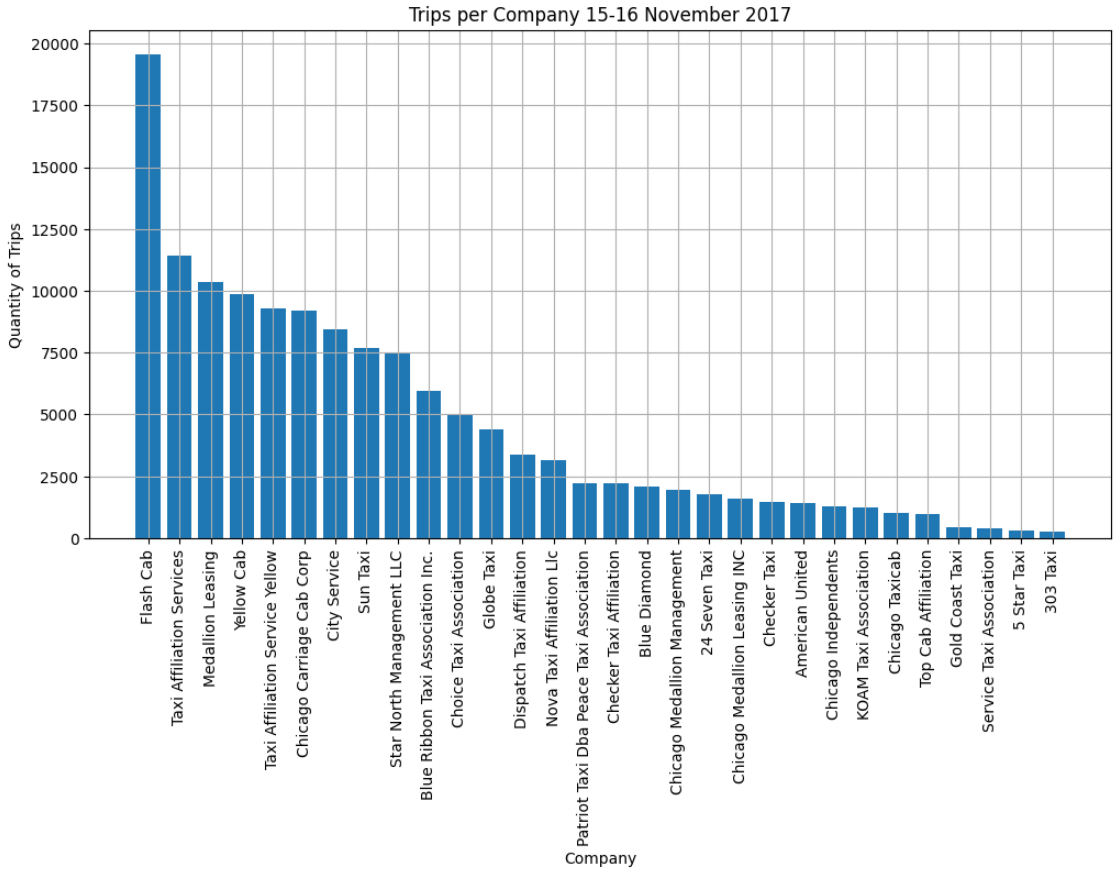
[8]:

	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

Executei o código para ordenar a coluna 'average_trips' em ordem decrescente e verificar se os valores estavam organizados do maior para o menor.

[9]:
```
trips_per_company = df_trips.sort_values(by="trips_amount", ascending=False).
↪head(30)


plt.figure(figsize = (12, 6))
plt.bar(trips_per_company["company_name"], trips_per_company["trips_amount"])
plt.xlabel('Company')
plt.ylabel('Quantity of Trips')
plt.title('Trips per Company 15-16 November 2017')
plt.grid(True)
plt.xticks(rotation = 90)
plt.show()
```

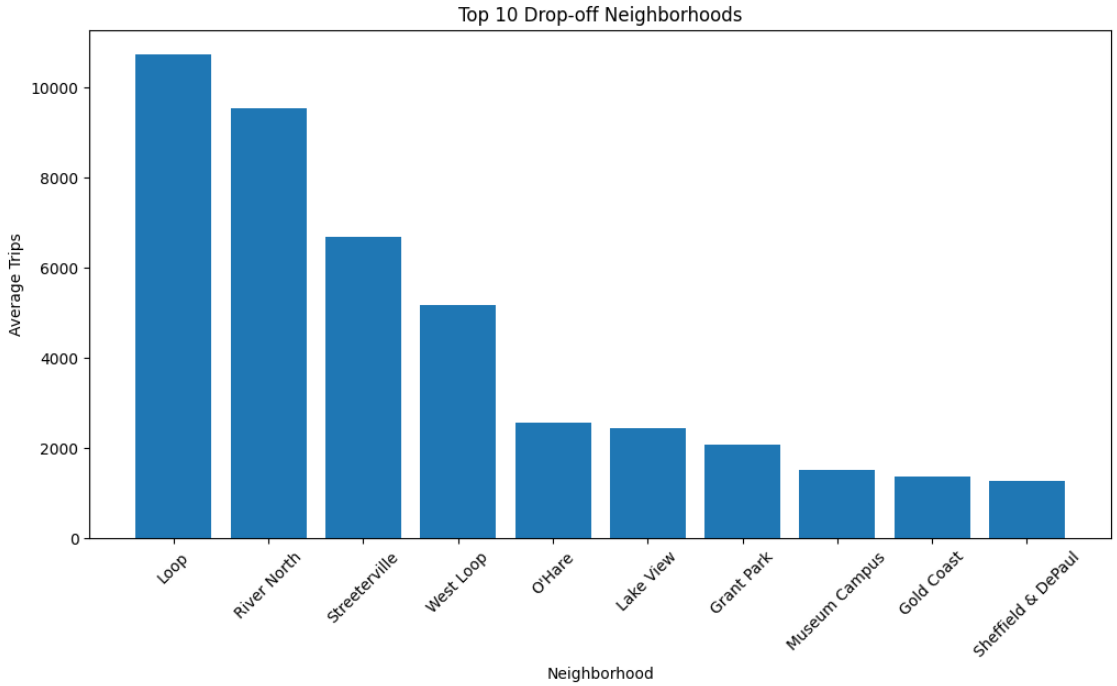


Analisamos as 30 principais empresas de táxi com mais viagens registradas. Observamos que, após a 15ª colocada, o número de viagens começa a cair significativamente. A empresa "Flash Cab" lidera o mercado com quase o dobro de corridas em comparação com a segunda colocada, "Taxi Affiliation Services", provavelmente devido a uma combinação de fatores: maior frota de veículos, estratégias de marketing mais agressivas, preços competitivos ou maior tempo de atuação no mercado, que geram maior reconhecimento e preferência pelos passageiros. Além disso, as demais empresas apresentam números bastante próximos entre si, indicando uma competição equilibrada entre elas.

We analyzed the top 30 taxi companies with the highest number of recorded trips. We noticed that after the 15th company, the number of trips drops significantly. "Flash Cab" dominates the market, with nearly twice as many trips as the second-place company, "Taxi Affiliation Services", likely due to a combination of factors: larger fleet size, more aggressive marketing strategies, competitive pricing, or longer market presence, which generate greater brand recognition and customer preference. Additionally, the remaining companies have very similar trip numbers, indicating a competitive balance among them.

[10]:
```
plt.figure(figsize = (12, 6))
plt.bar(df_top10_location["dropoff_location_name"],␣
  ↪df_top10_location["average_trips"])
```

```
plt.xlabel('Neighborhood')
plt.ylabel('Average Trips')
plt.title('Top 10 Drop-off Neighborhoods')
plt.xticks(rotation = 45)
plt.show()
```



Ao analisarmos os 10 bairros com maior número de desembarques, percebemos que os destinos Loop e River North apresentam uma média de corridas significativamente maior em comparação aos demais. Isso pode ser explicado pelo fato de esses bairros possuírem uma concentração maior de escritórios, aeroporto, hotéis ou atrações turísticas, tornando-os pontos de alta demanda para corridas de táxi.

When analyzing the top 10 drop-off neighborhoods, we noticed that the destinations Loop and River North have a significantly higher average number of trips compared to the others. This is likely due to these areas hosting a concentration of offices, airport, hotels, or tourist attractions, making them key locations for taxi demand.

[11]:
```
df_trips_airport = pd.read_csv('/datasets/project_sql_result_07.csv')
df_trips_airport.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
```

Hipótese Nula: O tempo médio das corridas em sábados chuvosos e não chuvosos é o mesmo.

Hipótese Alternativa: O tempo médio das corridas em sábados chuvosos e não chuvosos são diferentes.

[18]:
```
rain_weather = df_trips_airport[df_trips_airport['weather_conditions'] ==␣
 ↪'Bad']['duration_seconds']
sun_weather = df_trips_airport[df_trips_airport['weather_conditions'] ==␣
 ↪'Good']['duration_seconds']

var_equal = rain_weather.var() == sun_weather.var()

t_stat = st.ttest_ind(rain_weather,sun_weather,equal_var = var_equal)

p_value = t_stat.pvalue

print(var_equal)

alpha = 0.05

print(t_stat)

if p_value < alpha:
    print("Rejeitamos a Hipótese Nula.")
else:
    print("Não Rejeitamos a Hipótese Nula.")
```

```
False
Ttest_indResult(statistic=7.186034288068629, pvalue=6.738994326108734e-12)
Rejeitamos a Hipótese Nula.
```

Escolhemos um nível de significância = 0,05, pois é um valor comum em testes de hipóteses. Isso significa que aceitamos uma margem de erro de 5% ao rejeitar a hipótese nula quando ela pode ser verdadeira.

A Hipótese Nula diz que não há diferença significativa no tempo médio das viagens entre sábados chuvosos e não chuvosos. Ou seja, independentemente do clima, o tempo médio das corridas entre o Loop e o Aeroporto Internacional O'Hare seria o mesmo.

Já a Hipótese Alternativa sugere que o tempo médio das viagens é diferente nos sábados chuvosos e não chuvosos, indicando que a chuva pode impactar a duração da corrida.

Para testar essa hipótese, utilizamos o teste t de Student para amostras independentes, que compara as médias de dois grupos e verifica se a diferença entre elas é estatisticamente significativa. Se o p-valor (p-value) for menor que 0,05, rejeitamos a hipótese nula, o que significa que a chuva realmente

afeta a duração das viagens. Caso contrário, não há evidências suficientes para afirmar que a chuva tem impacto.

We chose a significance level of $= 0.05$, as it is a common threshold in hypothesis testing. This means we accept a 5% margin of error when rejecting the null hypothesis, even if it might be true.

The Null Hypothesis states that there is no significant difference in the average trip duration on rainy and non-rainy Saturdays. In other words, regardless of the weather, the average travel time between the Loop and O'Hare International Airport would remain the same.

The Alternative Hypothesis, on the other hand, suggests that the average trip duration is different on rainy and non-rainy Saturdays, meaning that rain might impact travel time.

To test this, we used the Student's t-test for independent samples, which compares the means of two groups to check if the difference is statistically significant. If the p-value is less than 0.05, we reject the null hypothesis, meaning rain does have an effect on trip duration. Otherwise, there isn't enough evidence to conclude that rain makes a difference.