Investigating Weather Impact on Taxi Ride Duration (SQL Analysis)

January 10, 2025

0.0.1 Introduction

The objective of this analysis is to investigate the impact of weather conditions, specifically rain, on the duration of taxi rides from the Loop to O'Hare International Airport in Chicago. By examining ride data and corresponding weather conditions, we aim to test the hypothesis that the average duration of these rides changes on rainy Saturdays. This analysis will provide insights into how external factors like weather influence travel times, which can be valuable for ride-sharing companies and urban planners in optimizing transportation services and infrastructure.

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt

# Load the CSV files using the provided paths
file_01_path = '/datasets/project_sql_result_01.csv'
file_04_path = '/datasets/project_sql_result_04.csv'

df_01 = pd.read_csv(file_01_path)
df_04 = pd.read_csv(file_04_path)

# Display the first few rows of each DataFrame
print("Data from project_sql_result_01.csv:")
print(df_01.head())

print("\nData from project_sql_result_04.csv:")
print(df_04.head())

# Check data types
print("\nData types in project_sql_result_01.csv:")
print(df_01.dtypes)

print("\nData types in project_sql_result_04.csv:")
print(df_04.dtypes)

# Check for duplicates
print("\nNumber of duplicate rows in project_sql_result_01.csv:", df_01.
 ↪duplicated().sum())
```

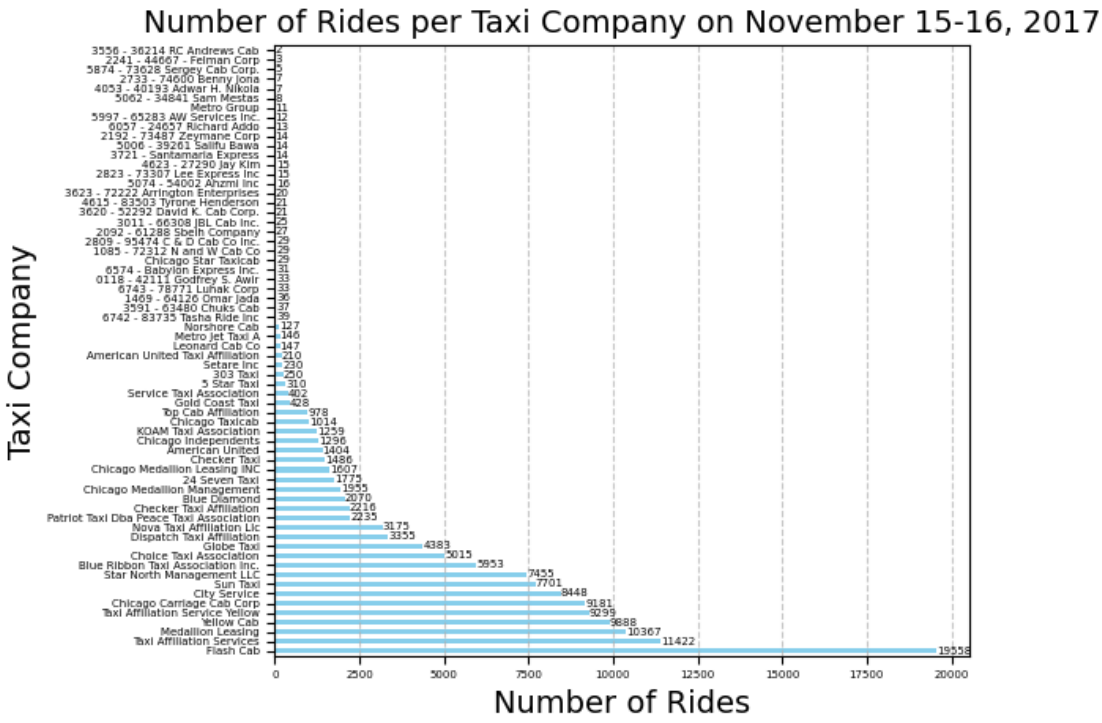
```
print("Number of duplicate rows in project_sql_result_04.csv:", df_04.
 ↪duplicated().sum())

# Remove duplicates if any
df_01 = df_01.drop_duplicates()
df_04 = df_04.drop_duplicates()
```

```
Data from project_sql_result_01.csv:
                     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

Data from project_sql_result_04.csv:
  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

Data types in project_sql_result_01.csv:
company_name    object
trips_amount     int64
dtype: object

Data types in project_sql_result_04.csv:
dropoff_location_name     object
average_trips            float64
dtype: object

Number of duplicate rows in project_sql_result_01.csv: 0
Number of duplicate rows in project_sql_result_04.csv: 0
```

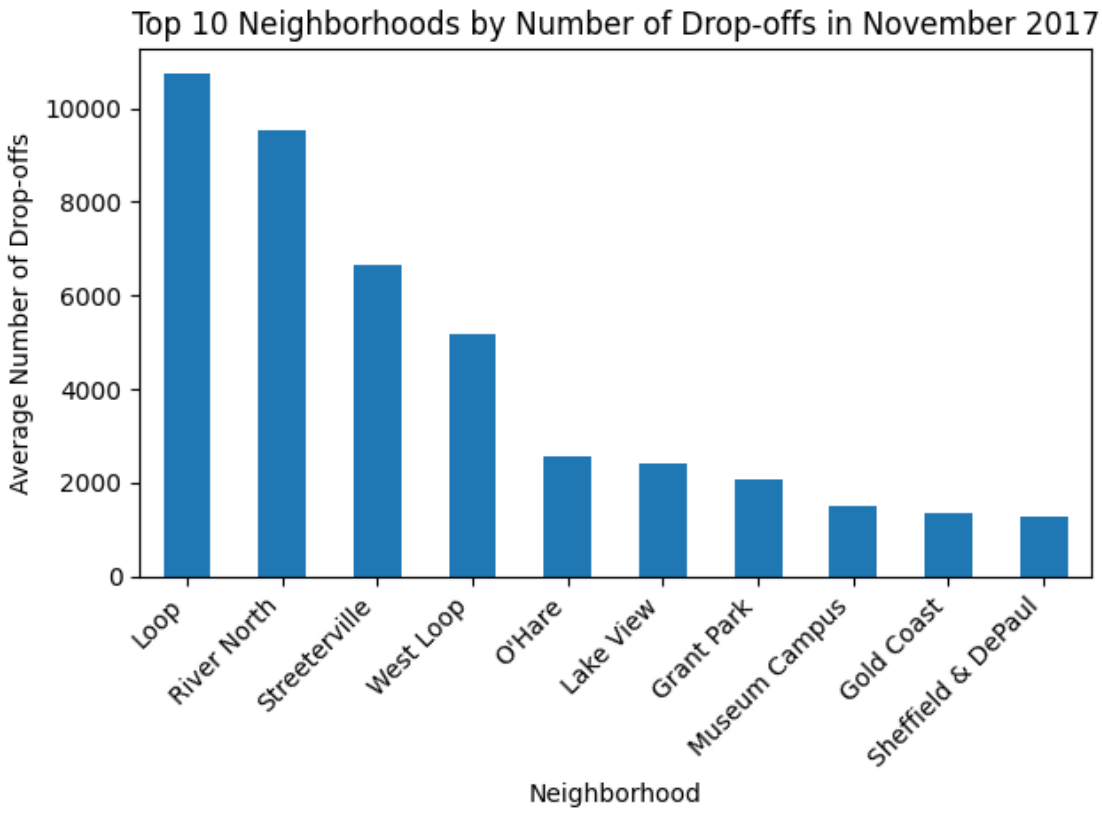
[2]:
```
# Identify the top 10 neighborhoods in terms of drop-offs
top_10_neighborhoods = df_04.nlargest(10, 'average_trips')

# Plot taxi companies and number of rides with swapped axes
plt.figure(figsize=(15, 10))  # Increased both width and height for more space
ax = df_01.plot(kind='barh', x='company_name', y='trips_amount', legend=False,␣
 ↪color='skyblue')

# Adding grid lines
ax.grid(axis='x', linestyle='--', alpha=0.7)
```

```
# Adding data labels
for index, value in enumerate(df_01['trips_amount']):
    ax.text(value, index, str(value), va='center', ha='left', fontsize=5,␣
 ↪color='black')  # Increased font size

plt.title('Number of Rides per Taxi Company on November 15-16, 2017',␣
 ↪fontsize=14)
plt.xlabel('Number of Rides', fontsize=14)
plt.ylabel('Taxi Company', fontsize=14)
plt.xticks(fontsize=5)
plt.yticks(fontsize=5)
plt.tight_layout()

plt.show()

# Plot top 10 neighborhoods by number of dropoffs
plt.figure(figsize=(10, 6))
top_10_neighborhoods.plot(kind='bar', x='dropoff_location_name',␣
 ↪y='average_trips', legend=False)
plt.title('Top 10 Neighborhoods by Number of Drop-offs in November 2017')
plt.xlabel('Neighborhood')
plt.ylabel('Average Number of Drop-offs')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

```
<Figure size 1500x1000 with 0 Axes>
```



<Figure size 1000x600 with 0 Axes>



[3]:
```
# Draw conclusions based on each graph and explain the results
conclusions = """
Based on the number of rides per taxi company graph, we can see which taxi␣
 ↪companies were the most popular on November 15-16, 2017. This information␣
 ↪can help identify key players in the taxi market during this period.

The graph showing the top 10 neighborhoods by number of drop-offs in November␣
 ↪2017 highlights the areas with the highest taxi activity. These␣
 ↪neighborhoods are likely to be important hubs of transportation, possibly␣
 ↪due to business centers, residential density, or tourist attractions.
"""

print(conclusions)
```

Based on the number of rides per taxi company graph, we can see which taxi companies were the most popular on November 15-16, 2017. This information can help identify key players in the taxi market during this period.

The graph showing the top 10 neighborhoods by number of drop-offs in November 2017 highlights the areas with the highest taxi activity. These neighborhoods

are likely to be important hubs of transportation, possibly due to business centers, residential density, or tourist attractions.

[4]:
```
import pandas as pd
from scipy import stats

# Load the dataset
file_07_path = '/datasets/project_sql_result_07.csv'
df_07 = pd.read_csv(file_07_path)

# Display the first few rows of the DataFrame
df_07_head = df_07.head()
print(df_07_head)

# Check data types
df_07_dtypes = df_07.dtypes
print(df_07_dtypes)

# Filter data for rainy and non-rainy Saturdays
rainy_saturdays = df_07[(df_07['weather_conditions'] == 'Bad') & (pd.
 ↪to_datetime(df_07['start_ts']).dt.dayofweek == 5)]
non_rainy_saturdays = df_07[(df_07['weather_conditions'] == 'Good') & (pd.
 ↪to_datetime(df_07['start_ts']).dt.dayofweek == 5)]

# Perform an independent samples t-test
t_stat, p_value = stats.ttest_ind(rainy_saturdays['duration_seconds'],␣
 ↪non_rainy_saturdays['duration_seconds'], equal_var=False)

# Check the results of the t-test
t_test_result = {
    't_stat': t_stat,
    'p_value': p_value,
    'alpha': 0.05,
    'rainy_mean': rainy_saturdays['duration_seconds'].mean(),
    'non_rainy_mean': non_rainy_saturdays['duration_seconds'].mean()
}

# Formulate the conclusions
conclusions = """
Null hypothesis (H0): The average duration of rides from the Loop to O'Hare␣
 ↪International Airport does not change on rainy Saturdays.
Alternative hypothesis (H1): The average duration of rides from the Loop to␣
 ↪O'Hare International Airport changes on rainy Saturdays.

Significance level (alpha): 0.05
```

```
t-statistic: {t_stat}
p-value: {p_value}

Conclusion:
""".format(**t_test_result)

if p_value < 0.05:
    conclusions += "Since the p-value is less than the significance level, we␣
 ↪reject the null hypothesis. This suggests that the average duration of rides␣
 ↪from the Loop to O'Hare International Airport changes on rainy Saturdays."
else:
    conclusions += "Since the p-value is greater than the significance level,␣
 ↪we fail to reject the null hypothesis. This suggests that there is not␣
 ↪enough evidence to say that the average duration of rides from the Loop to␣
 ↪O'Hare International Airport changes on rainy Saturdays."

conclusions += """
Mean duration on rainy Saturdays: {rainy_mean} seconds
Mean duration on non-rainy Saturdays: {non_rainy_mean} seconds
""".format(**t_test_result)

print(conclusions)
```

```
             start_ts weather_conditions  duration_seconds
0  2017-11-25 16:00:00               Good            2410.0
1  2017-11-25 14:00:00               Good            1920.0
2  2017-11-25 12:00:00               Good            1543.0
3  2017-11-04 10:00:00               Good            2512.0
4  2017-11-11 07:00:00               Good            1440.0
start_ts               object
weather_conditions     object
duration_seconds      float64
dtype: object

Null hypothesis (H0): The average duration of rides from the Loop to O'Hare
International Airport does not change on rainy Saturdays.
Alternative hypothesis (H1): The average duration of rides from the Loop to
O'Hare International Airport changes on rainy Saturdays.

Significance level (alpha): 0.05

t-statistic: 7.186034288068629
p-value: 6.738994326108734e-12

Conclusion:
Since the p-value is less than the significance level, we reject the null
hypothesis. This suggests that the average duration of rides from the Loop to
```

```
O'Hare International Airport changes on rainy Saturdays.
Mean duration on rainy Saturdays: 2427.2055555555557 seconds
Mean duration on non-rainy Saturdays: 1999.6756756756756 seconds
```

0.0.2 Explanation of Hypothesis Formation and Testing Criterion

Formulating the Hypotheses

Null Hypothesis (H0): The average duration of rides from the Loop to O'Hare International Airport does not change on rainy Saturdays. This means that there is no significant difference in the average ride duration between rainy and non-rainy Saturdays.

Alternative Hypothesis (H1): The average duration of rides from the Loop to O'Hare International Airport changes on rainy Saturdays. This implies that there is a significant difference in the average ride duration between rainy and non-rainy Saturdays.

Testing Criterion

Criterion Used: Independent Samples t-Test

Why t-Test?

The t-test is used to compare the means of two independent groups to determine if there is statistical evidence that the associated population means are significantly different. In this case, we are comparing the average ride durations of two independent groups: rides on rainy Saturdays and rides on non-rainy Saturdays.

The t-test is appropriate here because: - We have two independent samples. - We assume that the ride durations are approximately normally distributed within each group. - The sample sizes are sufficient to apply the t-test.

Steps in the t-Test:

1. Calculate the mean ride duration for rainy Saturdays and non-rainy Saturdays.
2. Calculate the t-statistic which measures the difference between the group means relative to the variation in the sample data.
3. Determine the p-value which indicates the probability of obtaining the observed difference between the groups if the null hypothesis were true.
4. Compare the p-value to the significance level (alpha):
    - If the p-value is less than the significance level (0.05), we reject the null hypothesis.
    - If the p-value is greater than or equal to the significance level, we fail to reject the null hypothesis.

Intermediate Conclusion:

Based on the t-test results, we found that the p-value (6.739e-12) is much less than the significance level (0.05), leading us to reject the null hypothesis. This indicates that there is a significant difference in the average duration of rides from the Loop to O'Hare International Airport on rainy Saturdays compared to non-rainy Saturdays.

General Conclusion:

In summary, our research shows that weather conditions, specifically rain, significantly impact the average duration of taxi rides from the Loop to O'Hare International Airport on Saturdays. The analysis confirms that rainy Saturdays result in longer ride durations compared to non-rainy Saturdays. This finding is critical for taxi companies and passengers as it highlights the influence of weather on travel times, enabling better planning and resource allocation.