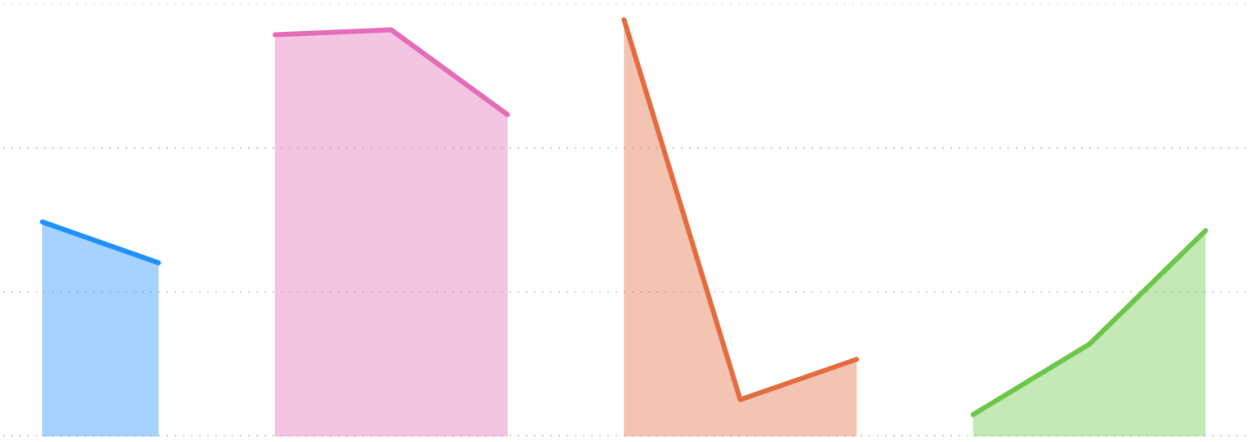
Rachel UWIMANA

23684

Uber Data Analysis

Big Data Assignment I Report



27th JULY 2025

Introduction

This assignment project is about a car hailing business and this report is about the uber ride dataset extracted from **Kaggle**. I wanted to understand how people use the service, if there were any interesting patterns in their trips, and how things like time and distance affect the fare. I started with Python for data preparation, and then I used Power BI to make some more advanced visuals that helped me see everything clearly as per the assignment requirements.

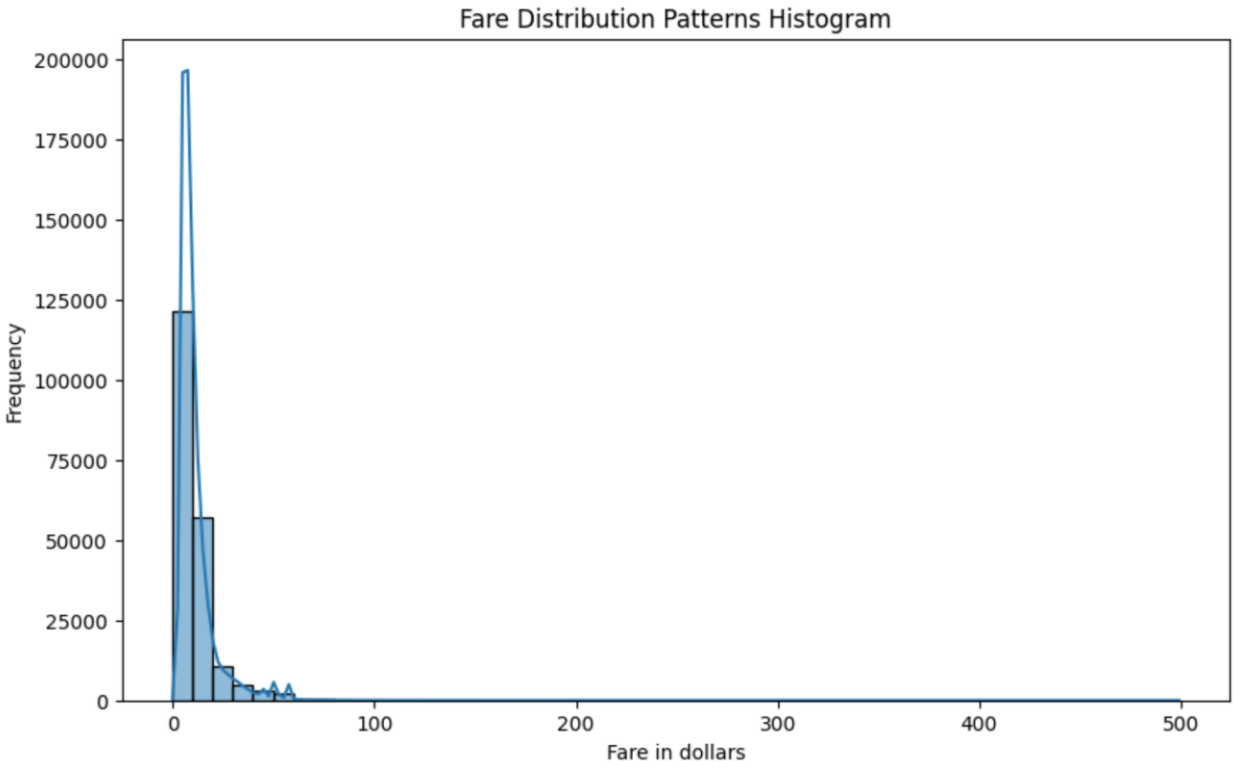
Methodology

My process was a step-by-step journey. First, I had to do some **cleanup on the data** in python, **explore the data (EDA**), and then **extract a clean dataset** that I would **insert in Power BI.** However, I met a challenge while getting Power BI to sort the months and days of the week columns in the right order, I just needed to create a specific column for sorting, but it took me a lot of time to figure it out.

```
import pandas as pd

#1. A,B,and C

#Loading the dataset into a data frame variable called dataFrame
dataFrame = pd.read_csv('uber.csv')
#Testing if it is well loaded
print(dataFrame.head())
#Checking the column names and data types
print(dataFrame.dtypes)
#Checking the rows and columns
print(dataFrame.shape)
#Checking how many null cells there are
print(dataFrame.isnull().sum())
#Checking how many duplicates there are
print(dataFrame.duplicated().sum())
#Checking the statistical standpoint of the dataset
print(dataFrame.describe())
```

```python
#2.B

import matplotlib.pyplot as plot
import seaborn as sns

#Creating a histogram for the fare amount
plot.figure(figsize=(10,6))
sns.histplot(dataFrame['fare_amount'], bins=50, kde=True)
plot.title('Fare Distribution Patterns Histogram')
plot.xlabel('Fare in dollars')
plot.ylabel('Frequency')
plot.show()
```

✓ 19.9s Python



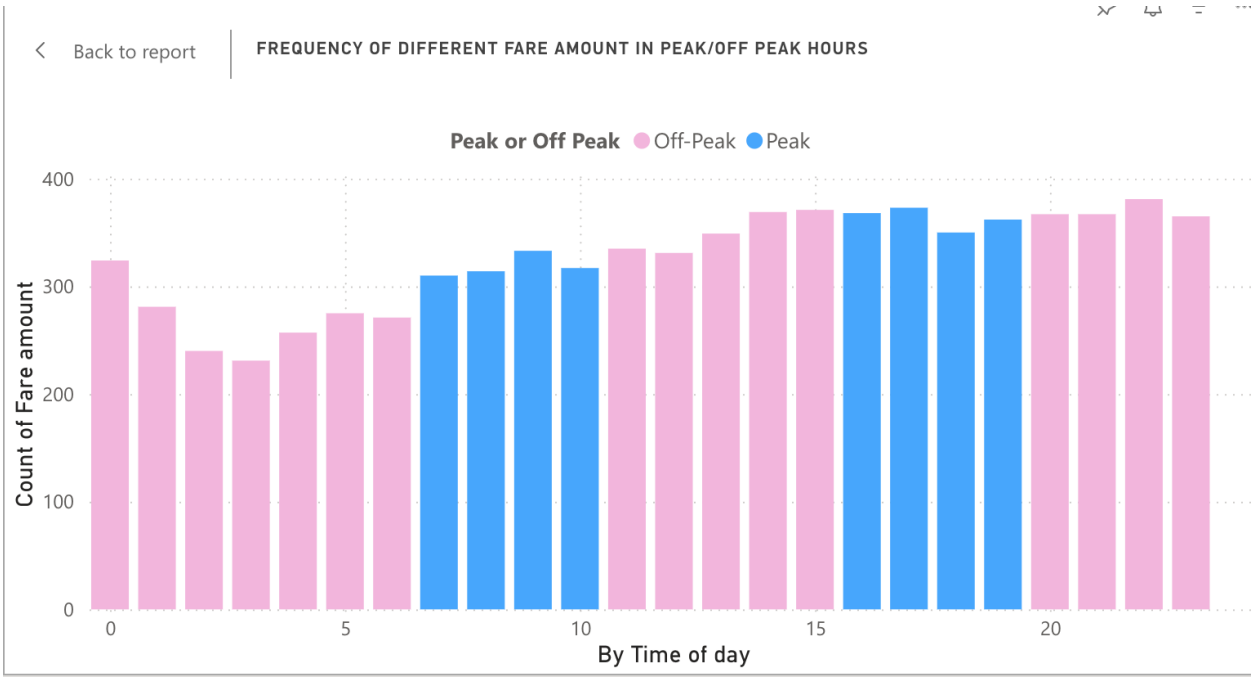
Next, I **generated visuals** to **gain insights** from the data but to display I also had to learn that for certain charts, like the scatter plot and map, I needed to tell set the axis field to "Do **not** summarize data", or it would just show me one big dot which was confusing.

After creating visuals, the findings helped me **formulate conclusions** about different user behavior, demand drivers, and pricing trends.

Analysis

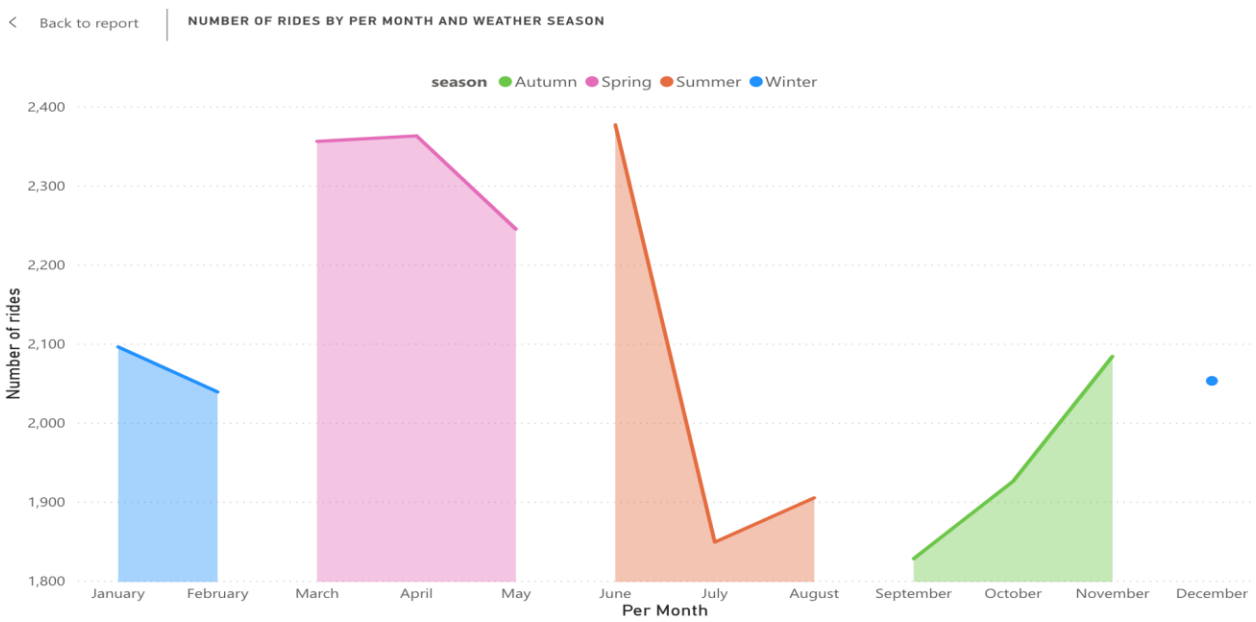
- **Weekly Trends:** The first thing I noticed was that my charts showed a big jump in rides on **Fridays and Saturdays**. This made me think that it is because people are

probably going out and socializing on the weekends. Then, on **Sunday**, the rides drop off, which makes sense because everyone needs to **rest for the next week** to prepare for work.



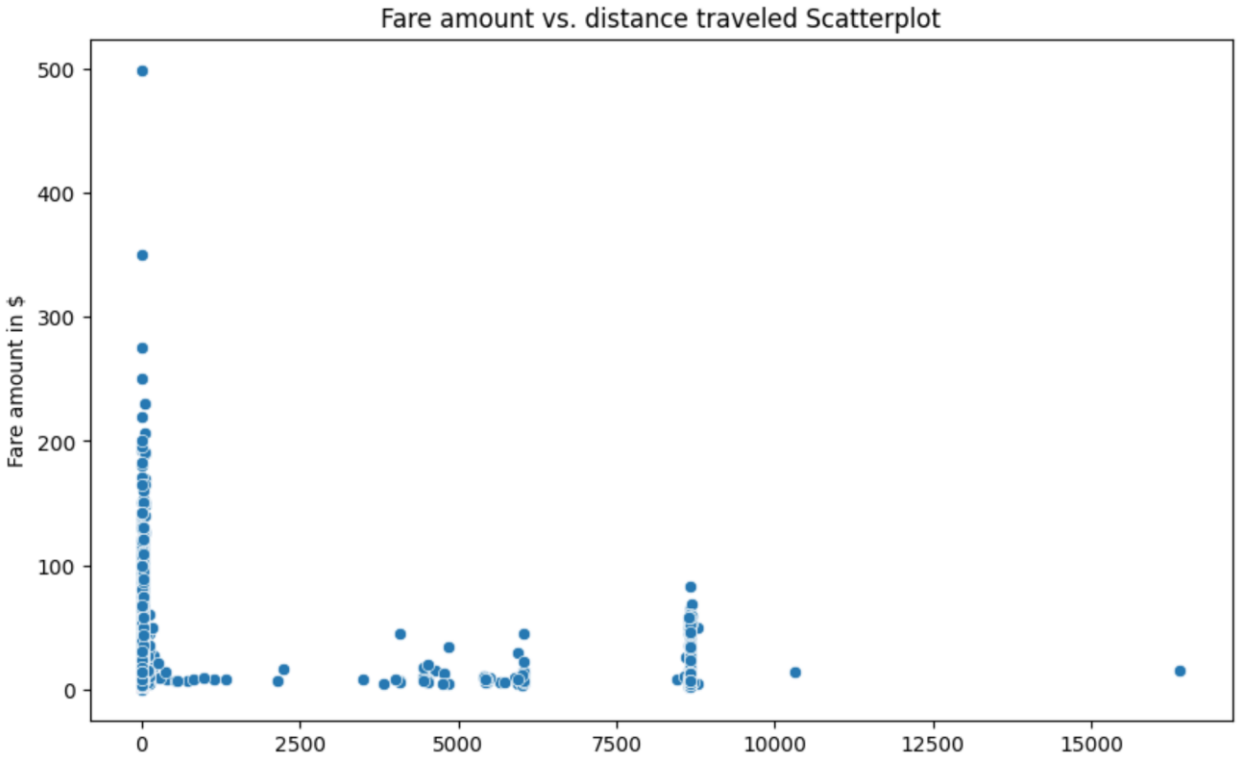
- **Time of Day:** I also found an interesting pattern when I looked at the hours of the day. The most expensive trips, and the ones that go the farthest, all happen between **9 AM and 10 PM**. This tells me that morning commutes and late night trips are the busiest and most profitable times for the company.

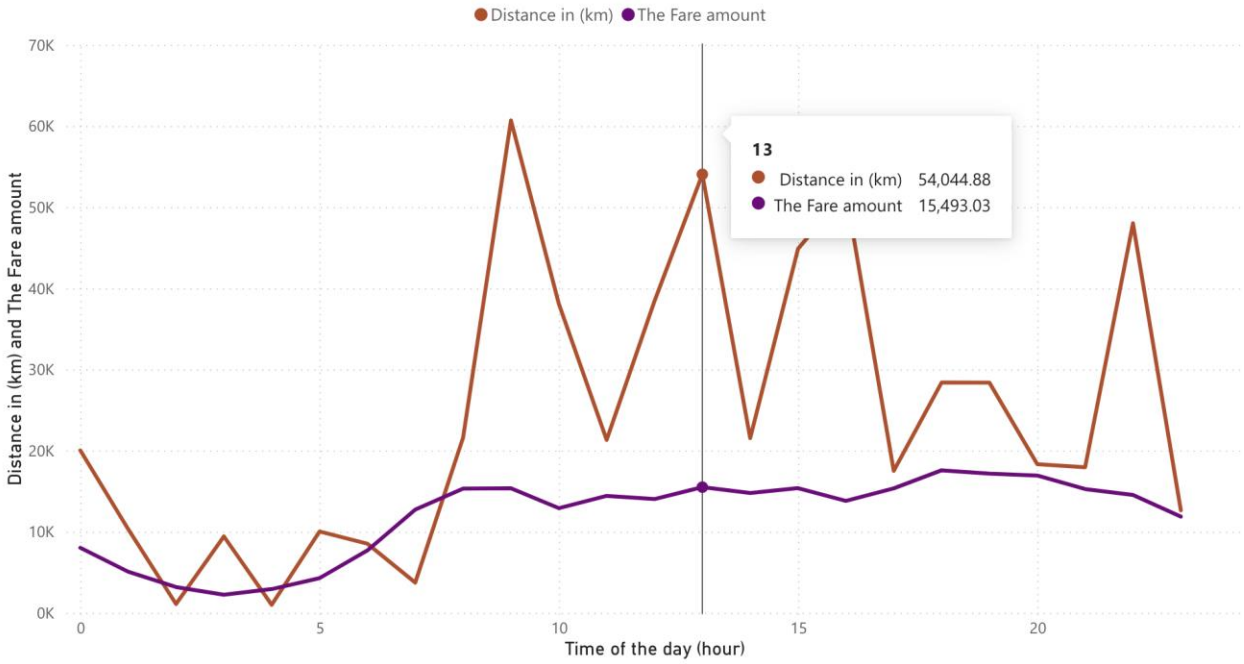
- **Seasonal Insights:** Third, my line chart showed the number of rides increased in the **spring season (March, April, and May)**. After searching for the temperature on weather seasons in the west my conclusion was that there was an increase because the weather gets nicer, and people want to go out and about more often. It was a great finding which led me to create a Season column so that I can highlight it in my diagrams.



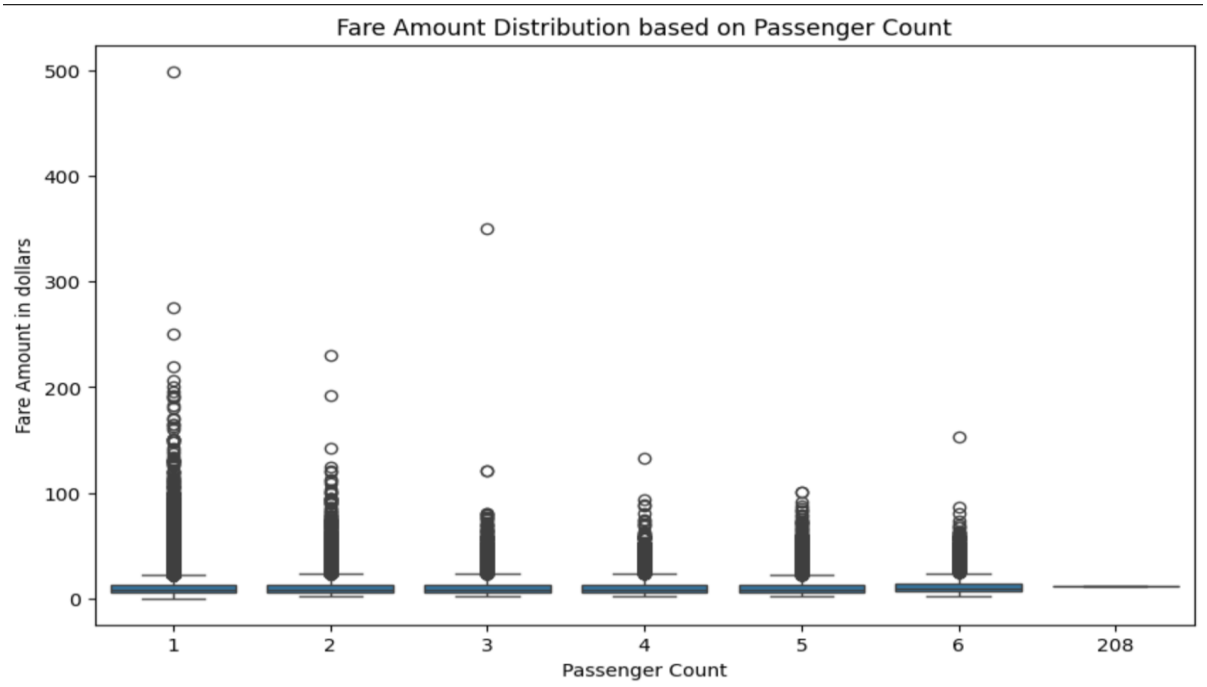
- **Fares and Distance:** The line plot showed a very clear relationship, the longer the trip, the higher the pay. This was exactly what I expected, but seeing it laid out in the visual made it really clear. To calculate the distance between longitudes and latitudes, I had to create a function with a dedicate formula.

```python
#Function with formula for calculating distance between latitudes & longitudes
def haversine(lon1, lat1, lon2, lat2):
    lon1, lat1, lon2, lat2 = map(np.radians, [lon1, lat1, lon2, lat2])
    distance_longitude = lon2 - lon1
    distance_latitude = lat2 - lat1
    a = np.sin(distance_latitude/2.0)**2 + np.cos(lat1) * np.cos(lat2) * np.sin(distance_longitude/2.0)**2
    c = 2 * np.arcsin(np.sqrt(a))
    r = 6371
    return c * r

#Adding distance as column to my data frame
dataFrame['distance'] = haversine(dataFrame['pickup_longitude'], dataFrame['pickup_latitude'], dataFrame['dropoff_longitude'],
dataFrame['dropoff_latitude'])
#Checking if the new column was created successfully
print(dataFrame[['distance']].head())
```

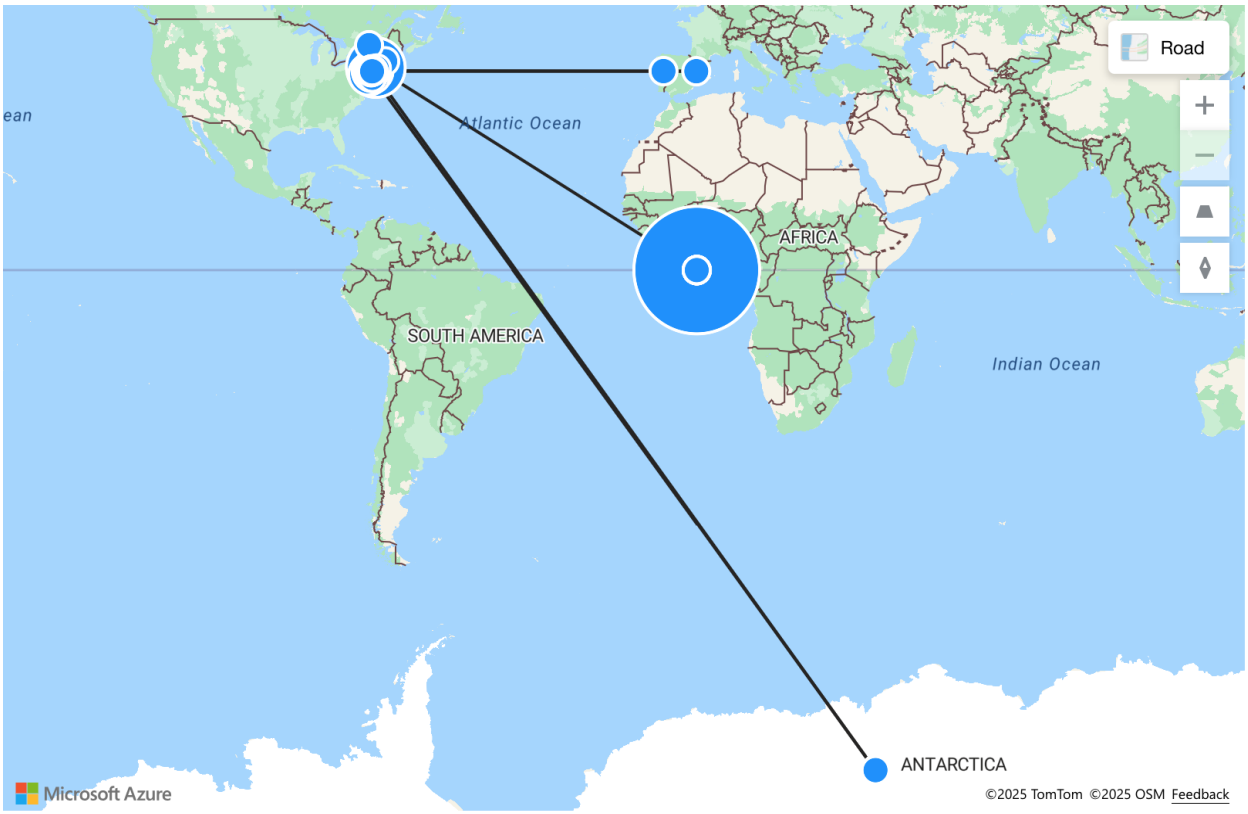




- **Fare and Passengers:** This was the most difficult one for me to understand. When I used the column bar plot the average fare for trips for more than one passenger was almost non-existent, which confused me. But when I used a stacked bar chart and lengend for peak hours, I saw that there were very few trips with two people and above, so that average wasn't reliable. Below is a scatterplot created in Jupyter and a stacked bar chart done in Power BI



- **Location of Rides:** The map showed that most of the rides are clustered together in one specific area in **North America in the US**, which makes me think the data is from a company that operates in the US. There are a few outliers of long distances like the ones going in Antartica and an African island.



Results

The main things I learned are:

- The uber business has a **weekend rush** and a **peak spring season**.
- **Trip distance** is the most important factor in determining the fare.
- The most important trips happen during the day and late at night.

Conclusion

This data analysis project was more interesting than I thought it would be. I learned that Uber's business is very predictable based on when and where people want to travel. From seeing the patterns in weekly social life to guessing how the weather affects ride demand, I feel like I've gained some really valuable skills in turning raw data into meaningful insights.

Recommendations

Based on my observation, here are a few ideas for the uber company:

- **Dynamic Pricing:** They should definitely focus on having more drivers available on **Fridays and Saturdays** and during the evening hours (9 PM onwards) because that's when demand is highest.

- **Marketing:** They could run promotions during the slower seasons, like winter, to try and boost ride numbers because that is when they are low. For example, by providing extra services like heating, winter essentials, etc…

- **Driver Strategy:** They could offer special bonuses to drivers who work in the areas where the most trips happen, or during the spring season to make sure they can handle all the extra demand.