Phase 4: Data exploration

Team name: **THE TRAILBLAZERS**
Team members:

1) Vyshnavi Basude– vbasu1@unh.newhaven.edu
2) Manasa Sukavasi(msuka1@newhaven.edu)
3) Pooja Donuru(pdonu1@unh.newhaven.edu)

GIT link: https://github.com/VyshnaviBasude/Team-TrailBlazers

1. Introduction

Machine learning algorithms power Uber's dynamic pricing strategy by processing historical ride data, weather forecasts, event calendars, and traffic conditions. These algorithms predict when and where demand for rides will increase, allowing Uber to allocate more drivers to high-demand areas and optimize pricing accordingly. Surge pricing incentivizes drivers to meet surging demand while ensuring efficient service for passengers. This data-driven approach maximizes driver earnings during peak times and enhances overall user satisfaction by reducing wait times and providing reliable transportation options.

2. Dataset:

The dataset is called the " Uber and Lyft Dataset Boston, MA" and it contains information about the ride information between Uber and Lyft . Here are some details about the dataset:

a. The selected dataset contains extensive details regarding Uber rides taken over a two month period in Boston, Massachusetts.
b. It provides not only ride details but also contextual features such as the weather conditions, time specifics, and environmental settings during each trip. Accessibility: The dataset is publicly available on Kaggle, a platform for predictive modeling and analytics competitions. Users can freely download this dataset after creating an account on Kaggle.
c. The data in this dataset has been collected using APIs provided by both Uber and Lyft, amalgamated with weather data APIs to provide context regarding environmental conditions during each ride.
d. The dataset is likely to have categorical data representing the type of ride (e.g., UberX, UberXL), although this is inferred and not explicitly mentioned in the provided data.

Exploring Uber and Lyft , ride sharing details, pricing , temperature , hours to Drive Sales. Identification of Deviations in sales and supply chain based on Customer reviews and sentiment analysis.

The columns which we have in our dataset are:
Temporal Data: hour, day, month, sunriseTime, sunsetTime, windGustTime, temperatureHighTime, temperatureLowTime, apparentTemperatureHighTime, apparentTemperatureLowTime, uvIndexTime, temperatureMinTime, temperatureMaxTime, apparentTemperatureMinTime, apparentTemperatureMaxTime,
Numerical Data: price, distance, temperature, apparentTemperature, precipIntensity, precipProbability, humidity, windSpeed, windGust, visibility, temperatureHigh, temperatureLow, apparentTemperatureHigh, apparentTemperatureLow, dewPoint, pressure, windBearing, cloudCover, uvIndex, ozone, moonPhase, precipIntensityMax.
Categorical Data: The dataset is likely to have categorical data representing the type of ride (e.g., UberX, UberXL), although this is inferred and not explicitly mentioned in the provided details.

3. Exploration Techniques:

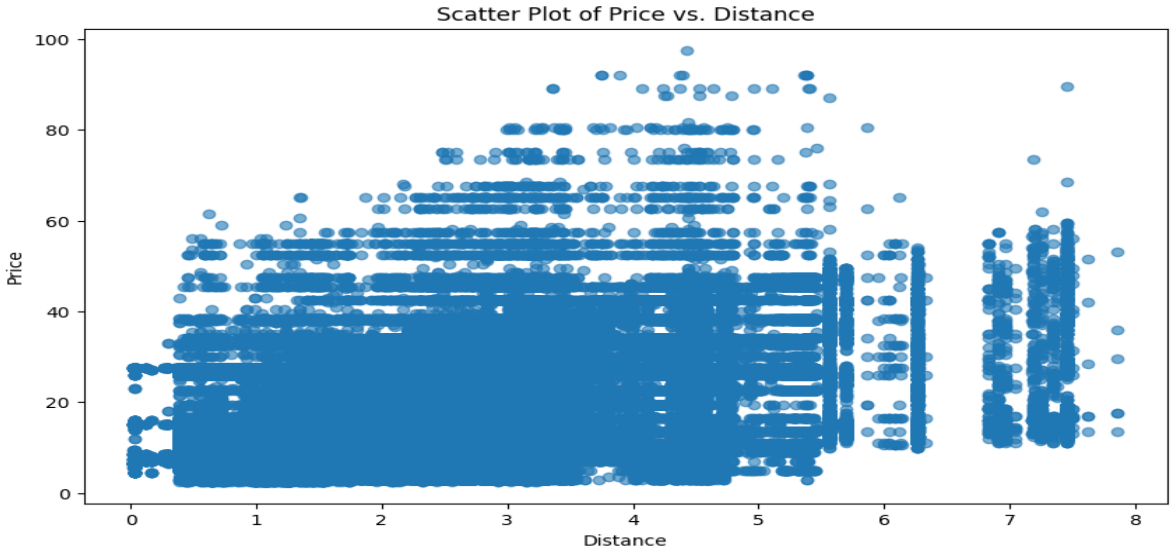
A summary statistics and exploratory data analysis was performed to gain insights of the data that included varied visualization techniques such as scatter plot, bar graph and histogram, 2-Y Axis Plot , Pie Chart, Co-relation

4. Data Exploration Techniques

Below are data exploration techniques we used in this project:

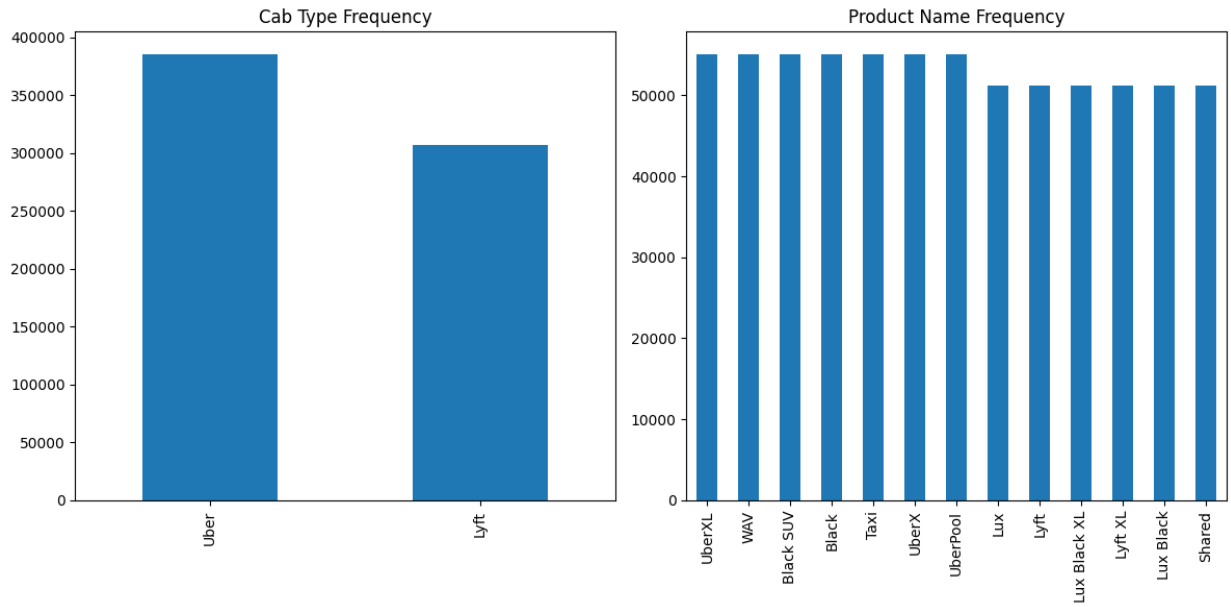
Scatter Plot :

Scatter plots help in visualizing the relationship between two numeric variables. To observed the data distribution and a relationship (if any), Scatter plot of Price vs Distance
Let's plot price against distance:
x



Bar Graph:

Bar graphs are great for visualizing the distribution of categorical data. Assuming cab_type is a categorical variable for one plot and Product Name for another plot , let's visualize its frequency:

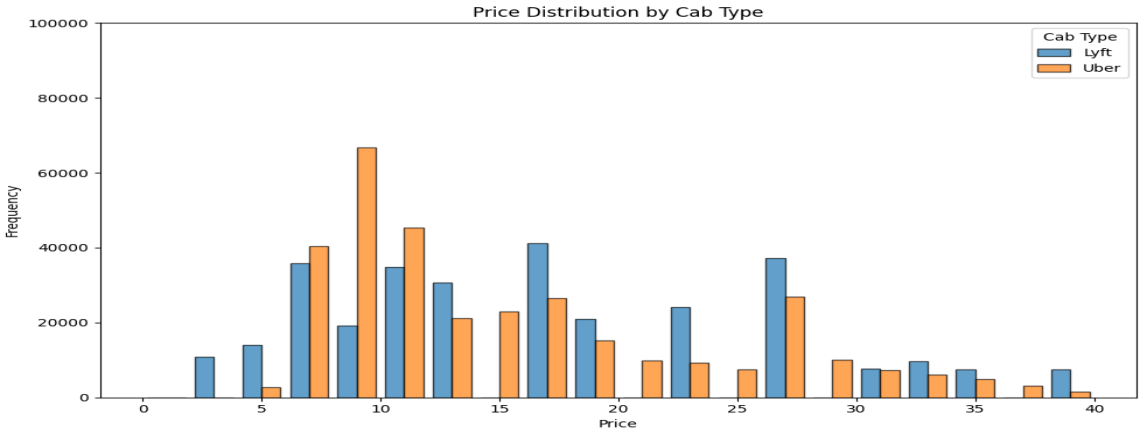


So , we can see very clearly that we have two different Cab types from plot one , and different product names from plot two.

Histogram:

A histogram is a graphical representation of the distribution of a dataset. It's a way to visualize the frequency (or count) of different values within a dataset and understand the underlying pattern or structure.

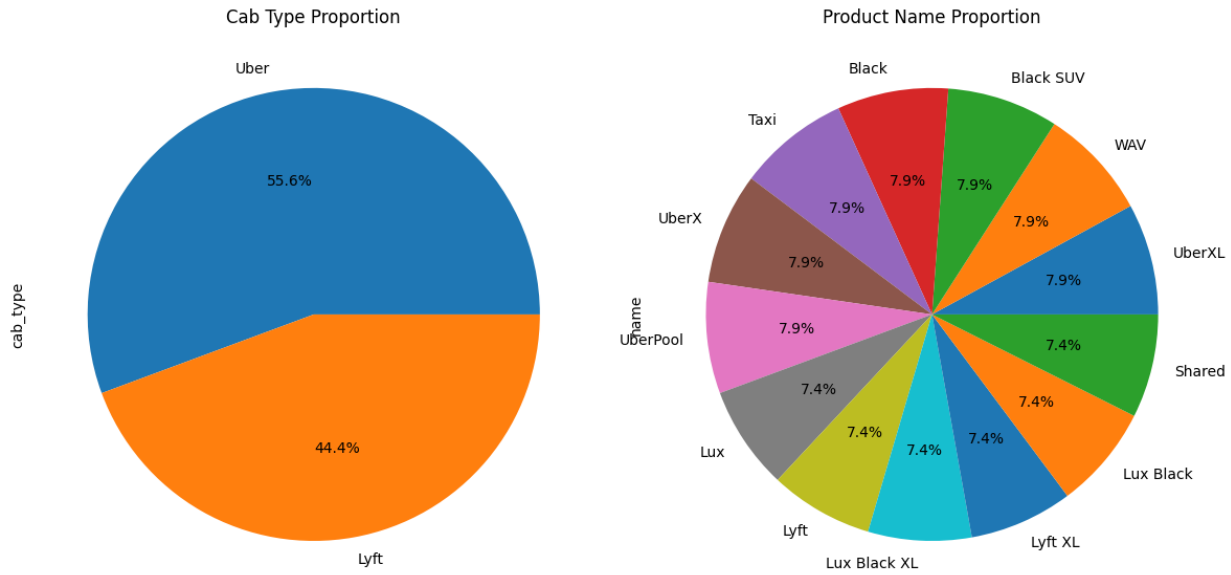
In our exploration technique , we have plotted an histogram for Price Distribution of cab type between Lyft and Uber, so that we coud easily predict the pricing and get an easy estimation which cab type is cheaper.



Pie Chart:

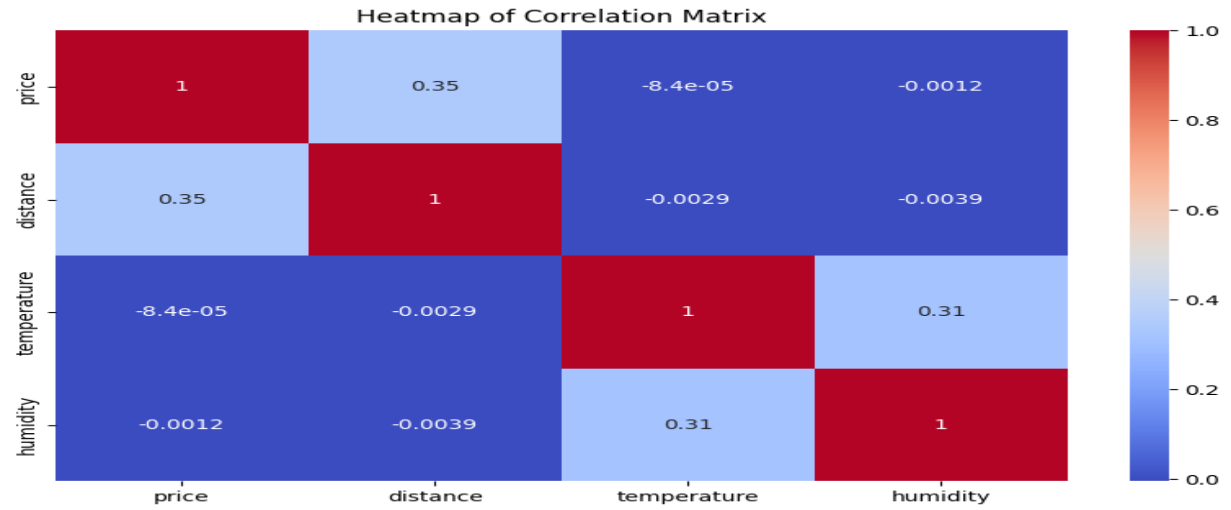
A pie chart is a circular graph that is divided into slices to represent numerical proportions. Each slice (or sector) corresponds to a category within a dataset, and the size of each slice is proportional to the value it represents relative to the whole.

By this Pie chart , we have a circular graph representation of cab-types and product- name with the numerical proportions.



Correlation:

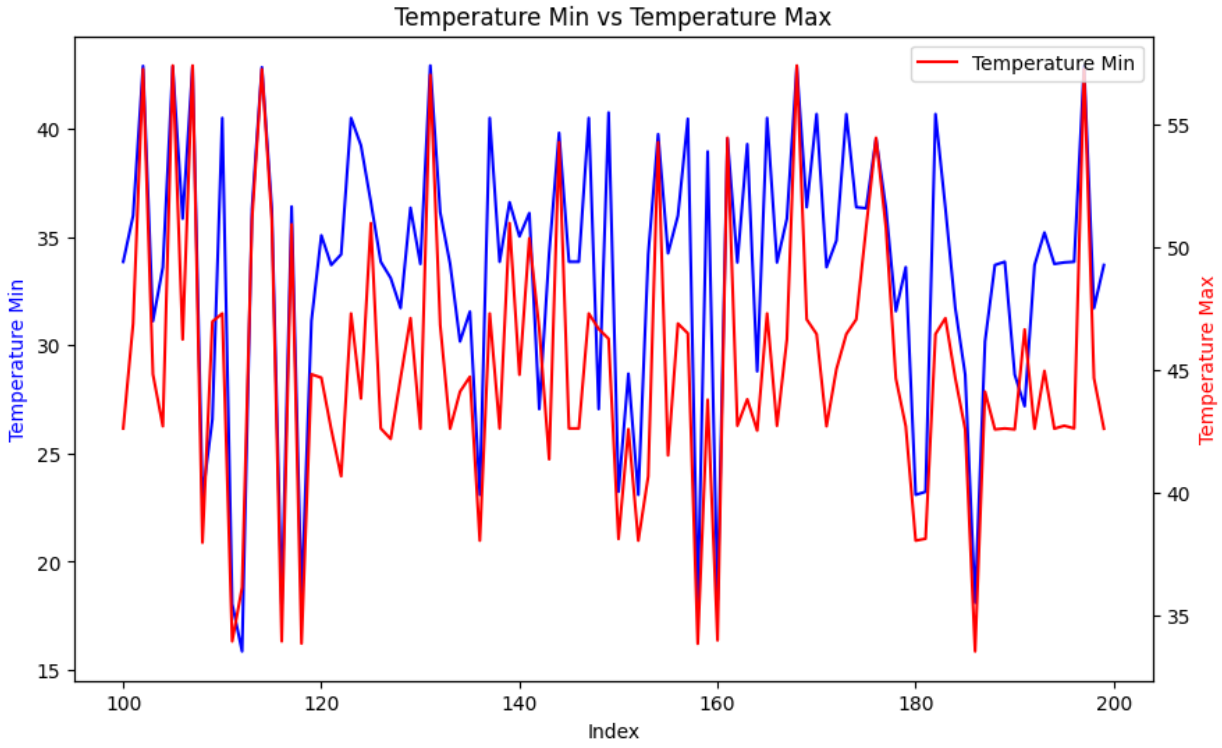
Heat maps are commonly used to visualize correlation matrices. Let's create a heatmap for the correlation between several numerical variables we have used humidity,temperature,distance and price.



2-Y Axis Plot:

A plot with two y-axes, also known as a dual-axis plot, is a graphical representation that displays two different sets of data on the same chart, each with its own y-axis. This allows for the visualization of two related but potentially different scales or units of measurement.

So, here the visualization was between the Minimum Temperature and Maximum Temperature.



4 Conclusion

The data analysis provided a detailed understanding of the data distribution. A significant finding is that the Uber has more sales compared to Lyft , in terms of price, days, temperature or humidity.