Project Report – Edward Wong (Group A)

GitHub URL
https://github.com/edwawong/UCDPA_Edwardwong

Abstract (Short overview of the entire project & features)
In this project, I chose 2 dataset – 1. Uber vs. Lyft rides data for Data Analysis, and 2. Countries' Quality of Life from World Population Data for Machine Learning. Both datasets were found in Kaggle.com.

I used most of the concepts covered in the course: File importing, using Regex & iterators to replace the values, dropped column & replaced missing values, merged dataframes from 2 csv files, defined a custom function, used NumPy array to do the calculation, created a dictionary to distinguish the trip types, produced charts with Seaborn, and used Machine Learning to do the predictions.

Introduction (Explain why you chose this project use case)
The reason of choosing Uber vs. Lyft rides dataset is because cab ordering is a real-life example that most of the people should have used. It would be interesting to know the pricing difference / customer habits among these 2 companies, for example, which one got a higher price per mile, which trip type got ordered frequently, etc. The dataset provides a detailed view on cab type, product, distance, origin & destination, price of each order, which is good for analyzing and comparing the difference between Uber & Lyft.

For machine learning dataset, I chose Countries' Quality of life is because this dataset provides a structural, clear numeric data, which is suitable to do the machine learning and would be interesting to know if there is any correlation between different aspects in a society.

Dataset (Provide a description of your dataset and source, also justify why you chose this source)

Dataset – Uber Lyft cab prices
https://www.kaggle.com/datasets/ravi72munde/uber-lyft-cab-prices

This dataset contains 2 files:

1) Cab_rides.csv – contains the cab type (Uber / Lyft), distance, timestamp, trip origin & destination, surge multiplier, price, id, product id, product name.
2) Weather.csv – contains the locations' weather conditions such as pressure, rain level of the day.

This dataset provides all necessary information of a trip which allows me to do the comparison of price vs. products vs. distance among Uber & Lyft. The weather.csv could be used for merging the dataframe to have a boarder view on how weather conditions affect the trips.

Dataset – World Population data
https://www.kaggle.com/datasets/madhurpant/world-population-data

This dataset has 5 files containing different aspect of the world population, while I only picked "Quality_of_life.csv") for the machine learning part, as I would like to create a regression model of Safety vs. Stability. The csv file contains the data of different aspect in a society of each country, all data in numeric format which is good for supervised machine learning.

Implementation Process (Describe your entire process in detail)

- Firstly, I imported all necessary packages such as numpy, pandas, matplotlib, sklearn, etc. – in **[1]**

  Importing data, Analysing data, Python

- Then I import the "cab_rides.csv" as "df" – in **[2]**

- Understand the columns & values in the csv file – in **[5]**

- Found that the "Product ID" of Uber is a string of random number, so I used Regex to find the pattern, and replaced the product id as "uber_product" which is similar as Lyft, and use iterator to loop it for every row – in **[6]**

- Changed the timestamp into a readable data format in df, so the data would be shown as date format – in **[7]**

- Check how many product types in the dataset by using .unique() – in **[8]**

- Create a dictionary to categorize the trip type by product, this allows analyzing the behavior by different trip type such as carpool / luxury trips – in **[9]**

- Create a column to store the trip type of each order – in **[10]**

- Check missing data in df, found there are 55,095 missing price data – in **[11]**

- Groupby cab_type, product, trip_type to see where the missing price data come from, it shows all come from "Uber Taxi", this might be because Taxi would be billed by meter price, therefore no information in this dataset – in **[12]**

- As Lyft doesn't get Taxi product, I decided to drop the "Uber Taxi" so in my analysis it would only contains Private cars' trip – in **[13]**

- Store the values (product, price, cab type, trip type, distance) into Numpy Array for the calculations – in **[14]**

- I would like to know the Price per Distance of Uber & Lyft, so I used Numpy Array to calculate, and use if, elif, else function to tell me the result – in **[15]**

- Defined a reusable custom function to extract the total / average of Uber & Lyft. I utilize this function to calculate the Price, Distance, Price per Distance of Uber & Lyft – in **[16]**

- Merge the cab_rides.csv & weather.csv together – import the weather.csv as weather_df – in **[25]**

- Add a "date" column to convert the timestamp into readable format – in **[26]**

- Fill missing values - Create a new dataframe in weather_df to fill the missing rain value as 0 – in **[27]**

- Add a new column for having the date & origin in both cab ride df & weather df to create a common column for merging – in **[28]**

- Merge 2 dataframes together to have a combined view of trip order & weather conditions – in **[29]**

Visualise

- Used Seaborn Countplot to show the number of trips of Uber & Lyft in the file, set x & y labels to present the chart clearly – in **[17]**

- Used Seaborn Countplot to show the number of trips of different products, set x & y labels to present the chart clearly – in **[18]**

- Used Seaborn Scatterplot to show the price & distance of the trips, to see the distribution of Uber & Lyft – in **[19]**

- Used Seaborn Countplot to show the number of trips of each trip type – in **[20]**

- Used Seaborn Scatterplot to show the price & distance, to see the distribution of each trip type – in **[21]**

- Used Seaborn Violinplot to show the distance, and the density distribution of the distance of the origins & destinations, I setup the order sequence in both Origin & Destination chart so it would be easier to compare – in **[22] & [23]**

- Used Seaborn Relplot to see the distance & price of each trip type, separated by Uber & Lyft by putting them into 2 columns – in **[24]**

Machine Learning

- Firstly, import the "quality_of_life.csv" – in **[32]**

- Create feature & target arrays (Safety) – in **[33]**

- Making predictions for Stability – in **[34]** & **[35]**

- Plotting safety vs. Stability – in **[36]**

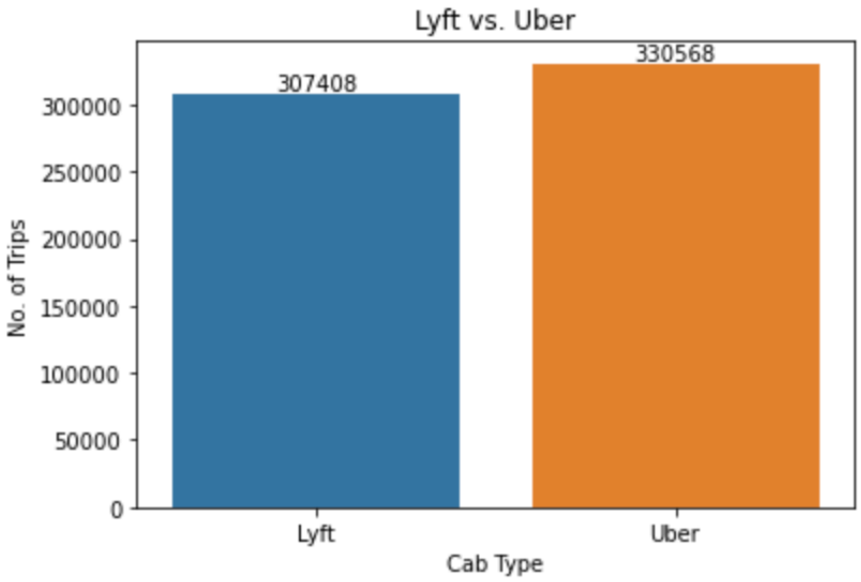
- Fitting a regression model – in **[37]**

- Get the R-squared score – in **[38]**

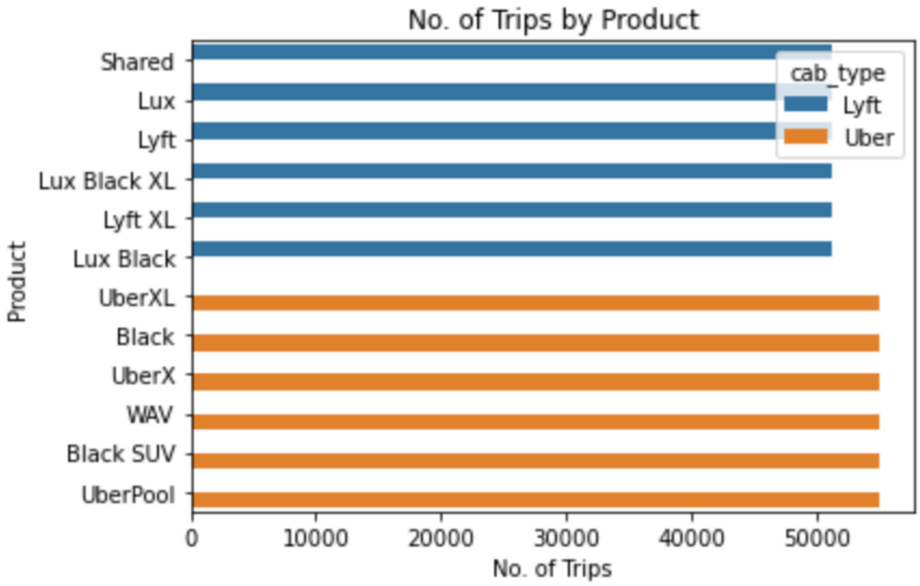
- Perform the Cross-validation – in **[39]** & **[40]** & **[41]**

Results (Include the charts and describe them)

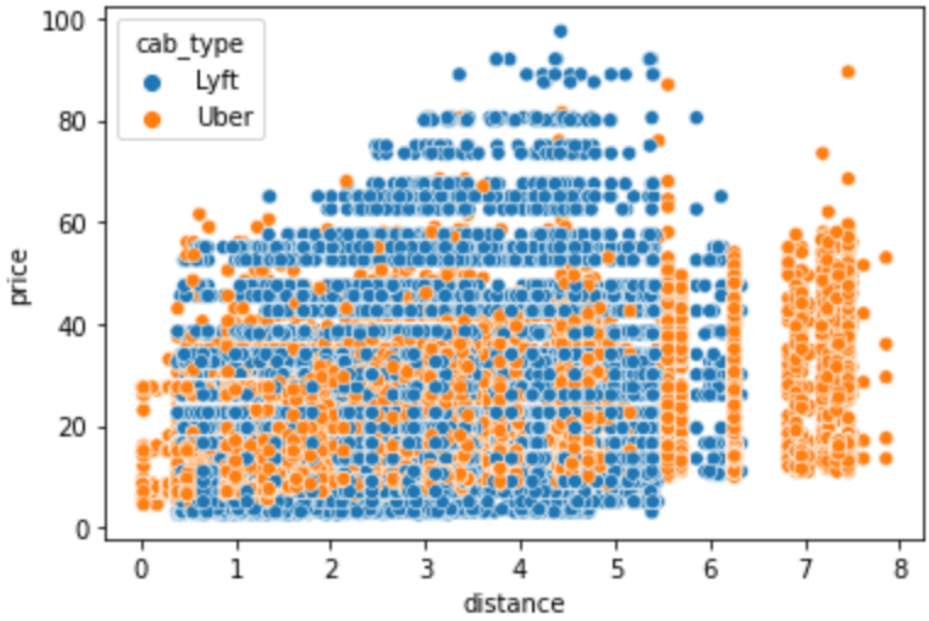
1. Seaborn Countplot - show the number of trips of Uber & Lyft in the file, Uber got 23,160 more trips than Lyft.



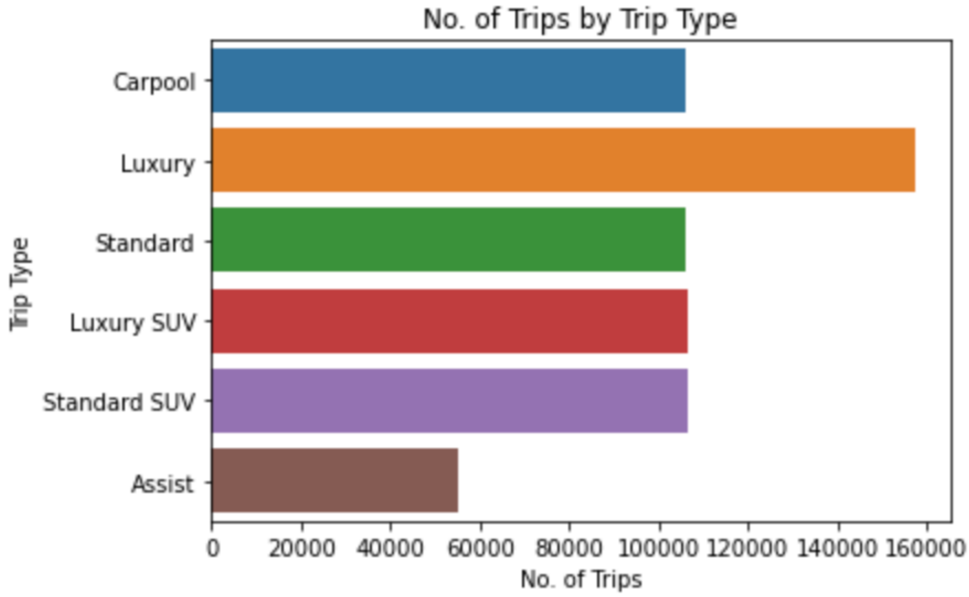
2. Seaborn Countplot to show the number of trips of different products, in this dataset, all products got a similar no. of trips – Lyft got around 50k trips for each product type, while Uber got around 55k trips for each product type.



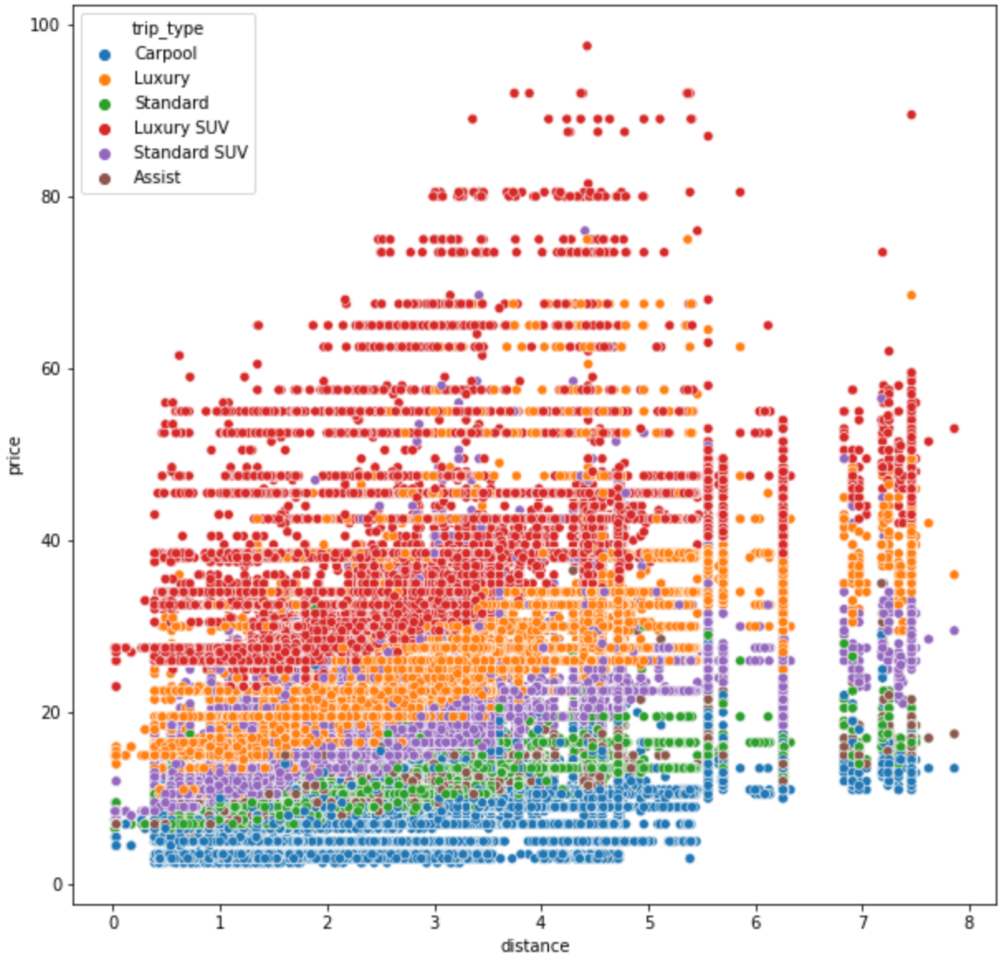
3. Seaborn Scatterplot to show the price & distance of the trips, to see the distribution of Uber & Lyft. In this chart it shows that Lyft trip distance is shorter than Uber in general, while the price is higher.



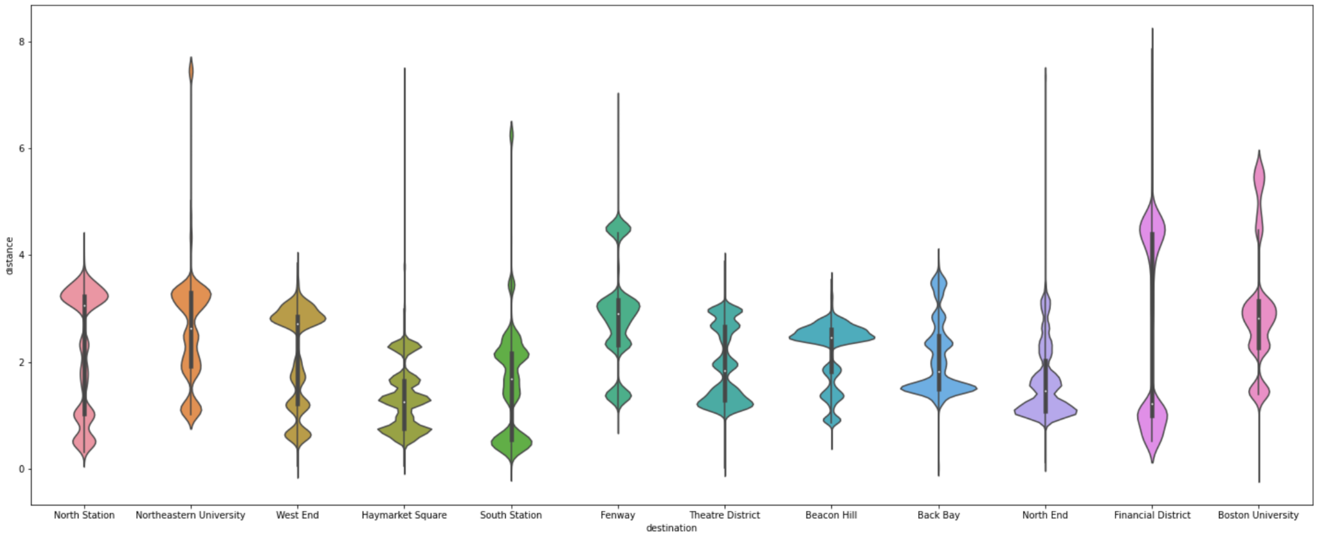
4. Seaborn Countplot to show the number of trips of each trip type, in the dataset the no. of trips per trip type is similar, Luxury got a higher no. of trip because it contains 3 products (Uber Black, Lyft Black, Lyft Lux)

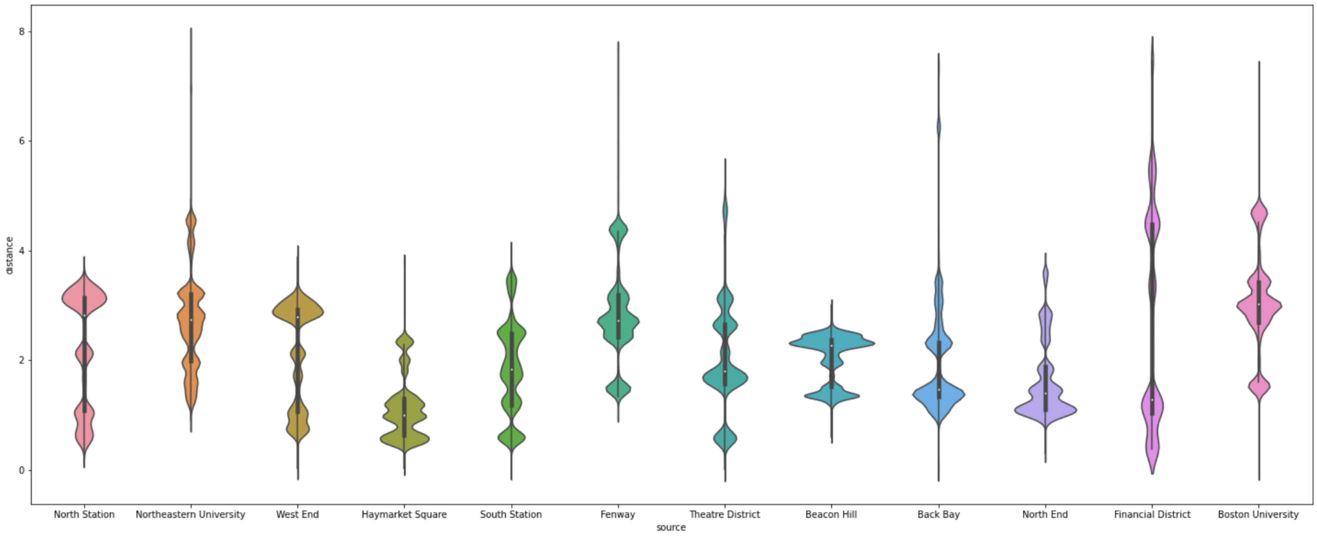


5. Seaborn Scatterplot to show the price & distance, to see the distribution of each trip type. From the chart it shows Luxury SUV is the most expensive trip type.



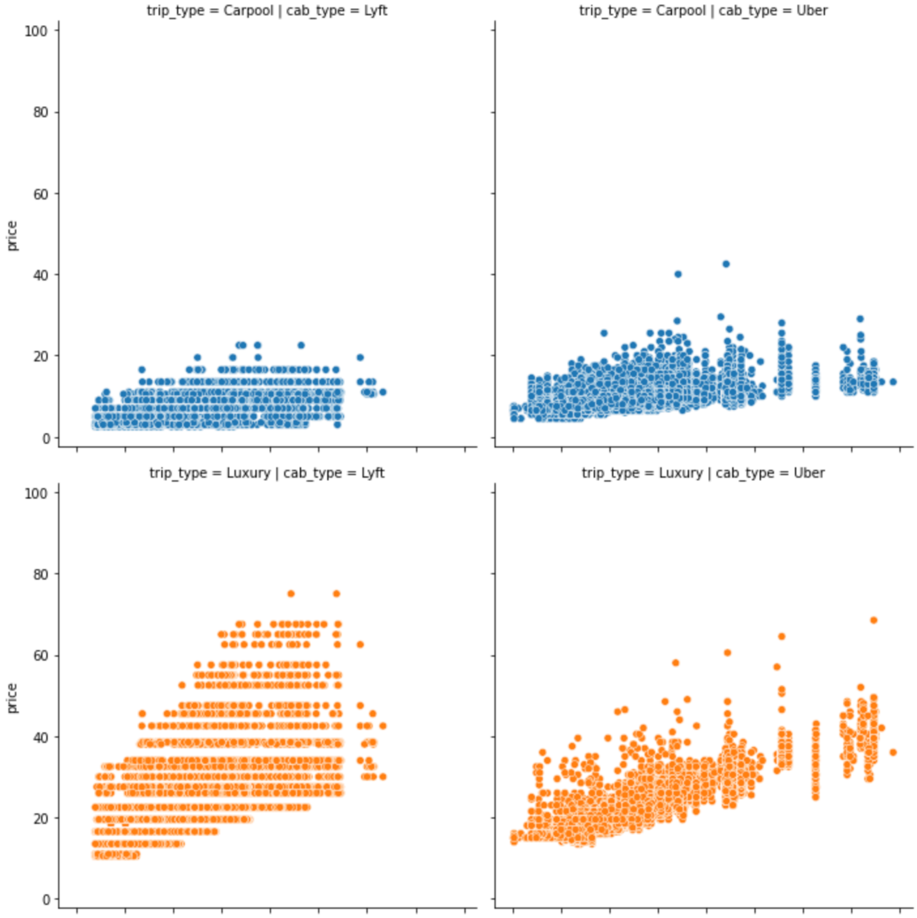
6. Seaborn Violinplot to show the distance, and the density distribution of the distance of the origins & destinations. In these charts we can see the difference of from / to each location.

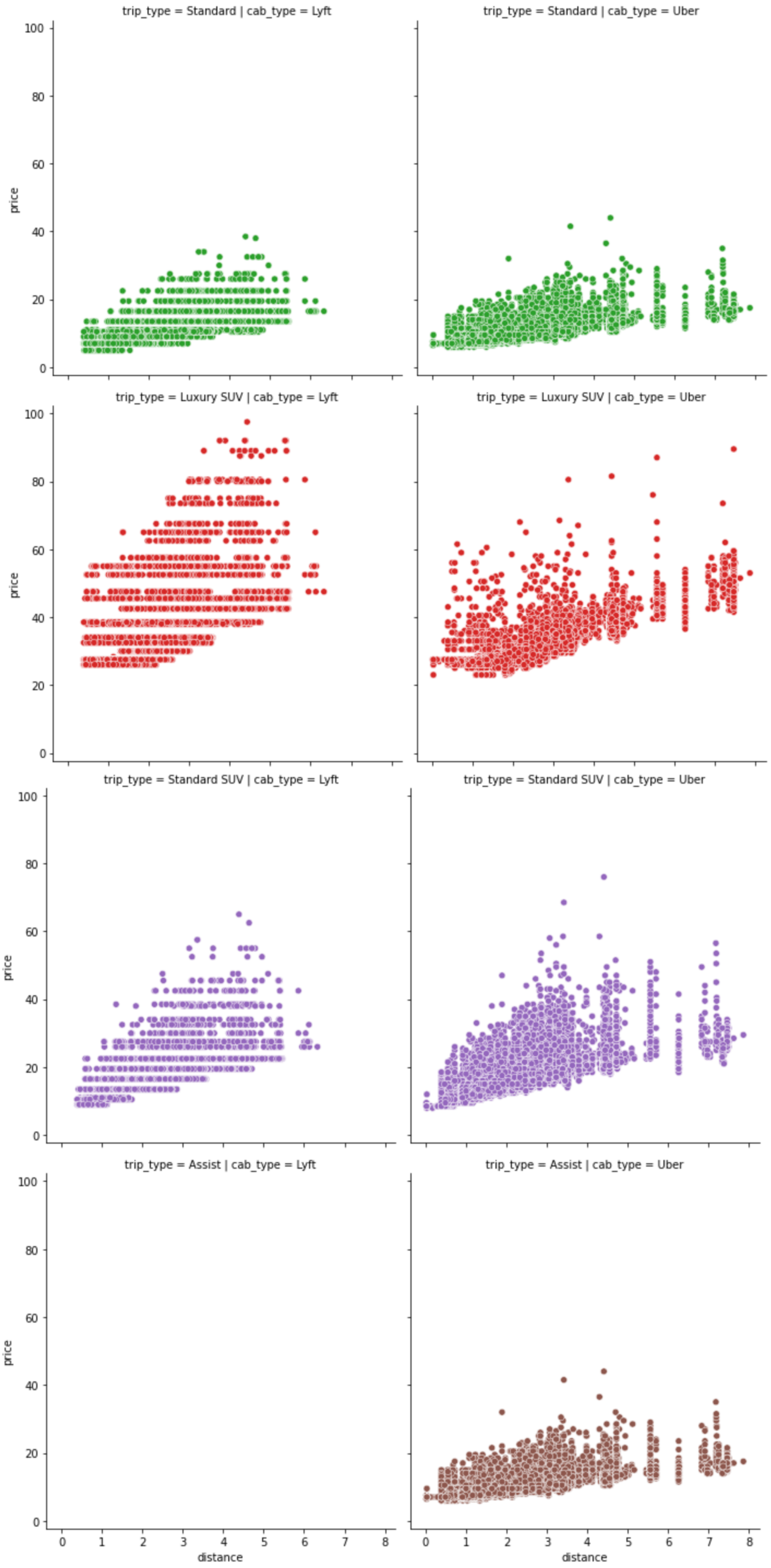




7. Used Seaborn Relplot to see the distance & price of each trip type, separated by Uber & Lyft by putting them into 2 columns. This shows the trip types by cab type side-by-side, which is easier to do an apple-to-apple comparison.
From the charts:

- Carpool – Uber usually got a higher starting price, while also got longer distance trips compared to Lyft
- Luxury – Prices of Uber's trips usually at around $20 - $50, while Lyft's trips were more expensive in the mid-length trips with the surge multiplier
- Luxury SUV – Lyft's trips got a higher price in medium distance range, while Uber price & distance increased in proportion.




trip_type = Standard | cab_type = Lyft
trip_type = Standard | cab_type = Uber
trip_type = Luxury SUV | cab_type = Lyft
trip_type = Luxury SUV | cab_type = Uber
trip_type = Standard SUV | cab_type = Lyft
trip_type = Standard SUV | cab_type = Uber
trip_type = Assist | cab_type = Lyft
trip_type = Assist | cab_type = Uber
price
0
20
40
60
80
100
distance
0
1
2
3
4
5
6
7
8

Insights (Point out at least 5 insights in bullet points)

- Total 638k Uber & Lyft trips in this dataset, 52% was Uber trips & 48% was Lyft trips. The number of trips per trip type was quite even, while Luxury was the most popular trip type in this dataset.

- Lyft got a higher price per distances comparatively; Uber's trips got $7.21 price per distance, whereas Lyft's trips got $7.93 price per distance.

- Long distance trips were dominated by Uber, >6 miles trips usually by Uber; while Lyft got more trips priced above $60 than Uber.

- Trip type key difference between Uber & Lyft:

  ○ Carpool – Uber usually got a higher starting price, while also got longer distance trips compared to Lyft

  ○ Luxury – Prices of Uber's trips usually at around $20 - $50, while Lyft's trips were more expensive in the mid-length trips with the surge multiplier

  ○ Luxury SUV – Lyft's trips got a higher price in medium distance range, while Uber price & distance increased in proportion.

- In the Machine Learning part (Quality of Life dataset), R-squared score recorded at 0.56 that said 56% of the variants follow the higher safety = higher stability regression model.