**1. Introduction**

**1.1 Real-world problems**

Uber and Lyft are the two most popular transportation apps that can allow passengers to get to their destination from their preferred pickup point and drivers to get flexible pay from each ride. Unlike traditional taxi services, fares of these transportation methods can greatly fluctuate depending on the demand and supply of the ride, and the levels of the service you choose. So, for this particular project, we will solve these real-world business problems:

Cab usage patterns: The analysis of cab usage patterns in the Financial District could be further investigated to gain insights on preferred cab types and ride preferences. This could help companies better understand the market and cater to their customers' needs.

Price comparison: The comparison of price increments for Uber and Lyft based on ride distance and cab type could be refined. This would help customers to make more informed decisions about which cab service to use and when.

Time and day analysis: The analysis of the most expensive and cheapest times and days to take cabs could be refined to identify the reasons for these trends. This could help customers to plan their trips more cost-effectively and help cab companies to plan their pricing strategies more effectively.

Weather and demand analysis: The analysis of how precipitation affects cab demand and pricing could be refined. This could help companies to plan their operations during different weather conditions and help customers to plan their trips more efficiently.

**1.2 Dataset Selection**

We select two datasets from Kaggle. One is Cab rides collected for a week in Nov to Dec in 2018 from both Uber and Lyft. The data is queried for every 5 mins. This dataset contains the relevant information from each ride: Distance, Cab type, Time stamp, Destination, Source, price, and Surge multiplier. This dataset can provide some insight about peak hours, popular destinations, price fluctuations, and customer preferences and habits. From those insights, we can optimize the supply and demand of cabs and predict prices during each hour.

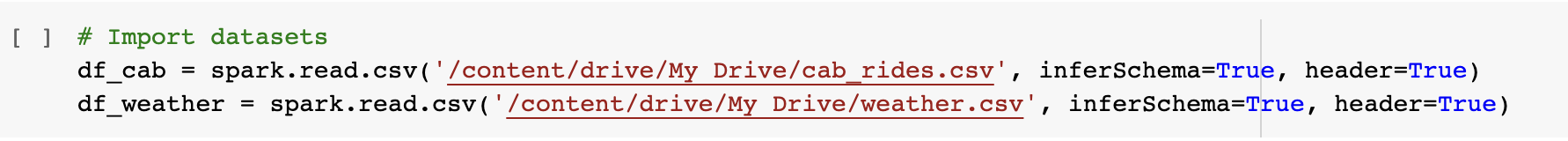
The second dataset is the weather dataset. It is collected for every 1 hour for the same week as the Cab rides. It contains variables such as temperature, rain, cloud, pressure, humidity, and wind. We can obtain insights such as, weather impact on cab rides prices, and customer preference. From those insights, we could predict the surge multiplier during specific weather conditions to optimize the supply of cabs and propose potential weather-based promotions to customers.

**1.3 Methodology for data analysis**

The methodology for analyzing the Cab\_rides and Weather datasets will involve the following steps. Due to the size of the two datasets (70000 rows in Cab\_rides and 6276 rows in Weather), we will use big data analysis tools like PySpark, an open-source big data processing framework built on Apache Spark, to perform distributed data processing tasks.

The first step in the analysis will involve exploratory data analysis (EDA) to understand the data and identify patterns and relationships between dependent variables and independent variables. This will include examining the price distribution and patterns by each Cab type, rides preference, time throughout the day and week, and across weather attributes to answer our real-world questions.

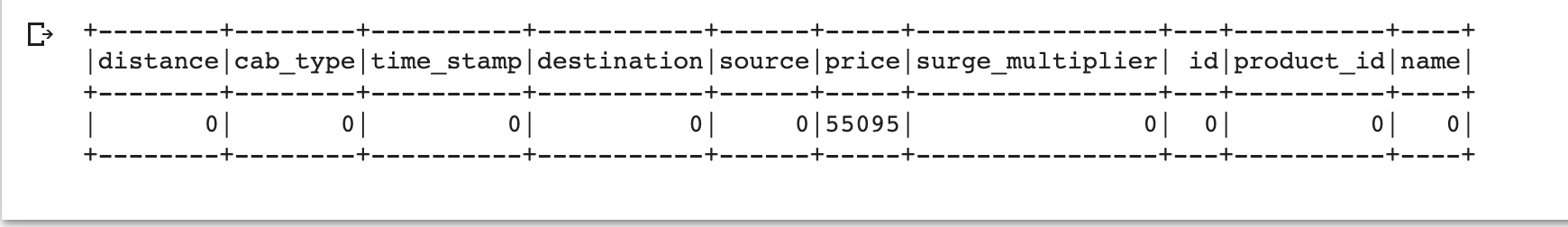
**2. Data Preparation**

We download two datasets: “weather.csv” and “cab\_rides.csv” from Kaggle and put them into Google Drive so that they can be inputted directly into Google Colab using PySpark:  


*Figure 1. Import two datasets.*

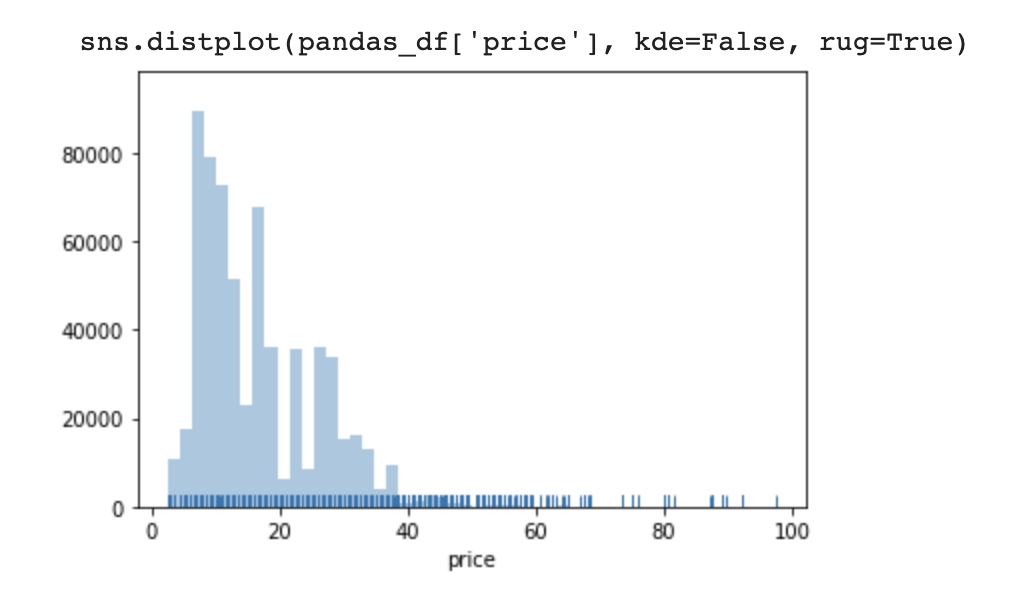
Firstly, we do data preprocessing on “cab\_rides.csv” which covers different types of cabs of Uber and Lyft and the price from different locations, the dataset contains information for every 5 minutes for a week from Nov 18th. (RaviMunde, 2019)

The first thing we have done is check the miss values on each column, which is shown in the figure below:



*Figure 2. Missing values in the “cab\_rides” dataset.*

We found that all the missing values are in the “price” column, to fill in those **missing values**, we need to look at the distribution of the “price” variable. As shown in the figure below, we could see that the distribution is right skewed. Therefore, we decide to fill in the missing values with the median of the price.

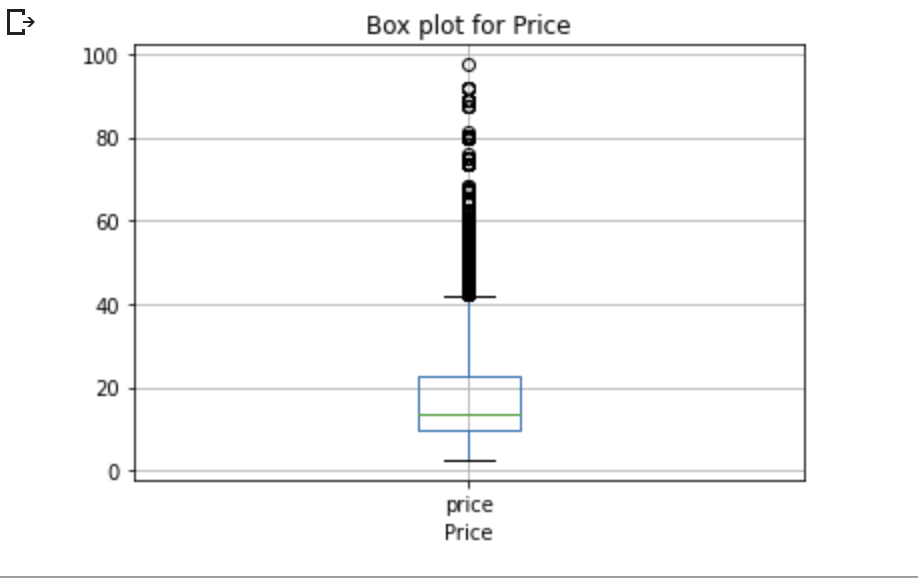
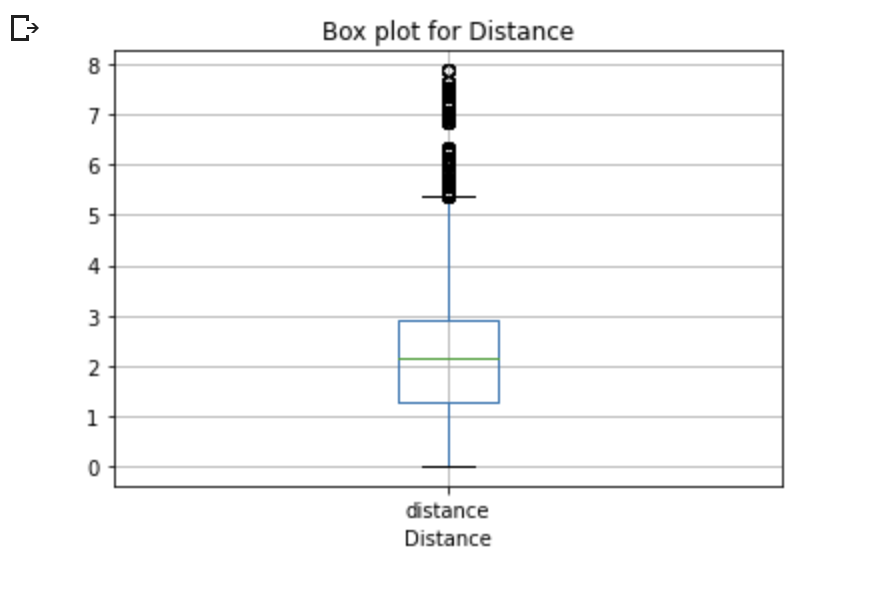


*Figure 3. The distribution of the “price” variable in the “cab\_rides” dataset.*

Then we check if there is any **duplicate value** in the “cab\_rides” dataset, before removing duplicate values, there are 693071 records in the dataset. After cleaning the duplicate values, there are still 693071 records, which means that there are no duplicate values in the dataset.

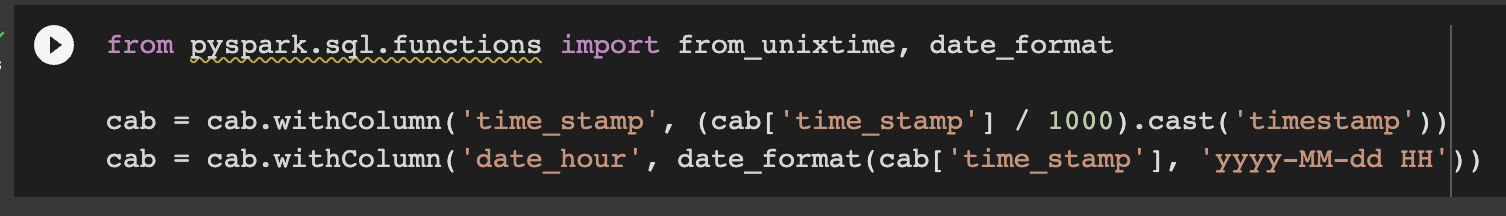
The next step is to check and **clean potential outliers** in every column. First, we look at the boxplot of the “distance” variable in the dataset. We found that there are some outliers that exist, more specifically, there are 10872 outliers out of a total of 693071 records. We decided to remove those outliers because the number of outliers is only 1.57% compared to the total number.

For the next variable “price”, we found that there are 6401 outliers, counting 0.77% of the total number of records, as shown in the box plot, also we remove those outliers as well.

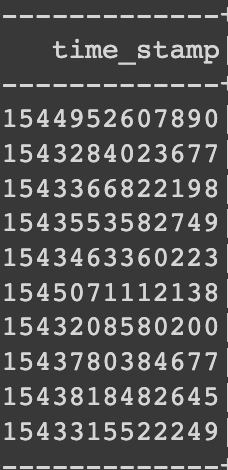
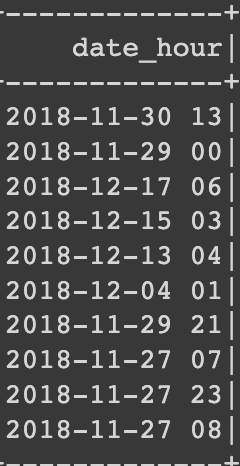


*Figure 4. The boxplot of the “distance” variable and “price” variable in the “cab” dataset*

After removing the potential outliers, the next variable we would focus on in the data cleaning step is the **“timestamp”** variable in the “cab\_ride” dataset. Before any process, the value in the timestamp variable is a long string of numbers containing how many seconds from 1970-01-01 to the time the data collected, which makes little sense in the first glance and for the analysis afterward. Therefore, we change the long string number into date time values. Since the data collected every 5 seconds in the “cab\_rides” dataset and collected every hour in the “weather” dataset, for the purpose of further merging these two datasets, here we transfer the timestamp in the “cab\_ride” dataset into hours in date time, the process and result are shown in the figure5 and figure 6 below.



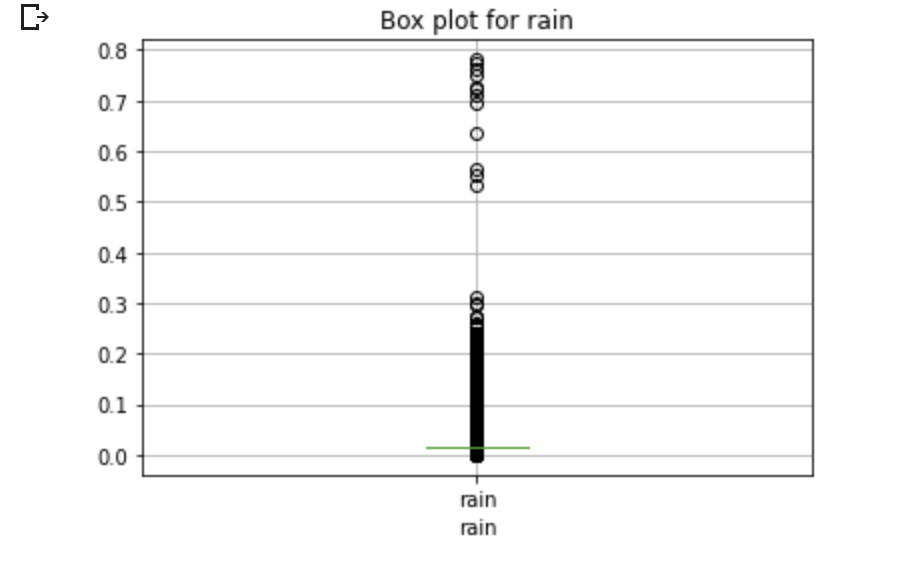
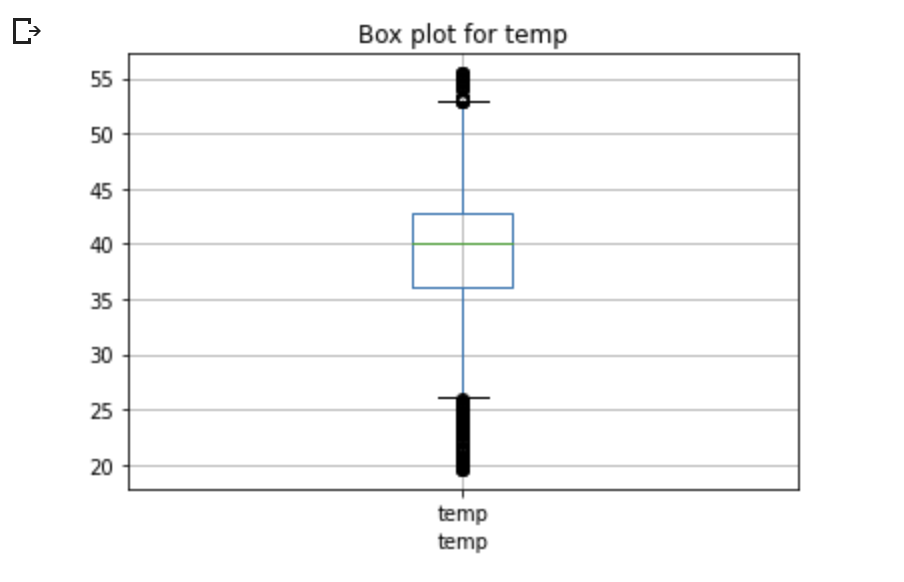
*Figure 5. The process of transferring timestamp into date time.*

**

*Figure 6. Before and after date time transformation.*

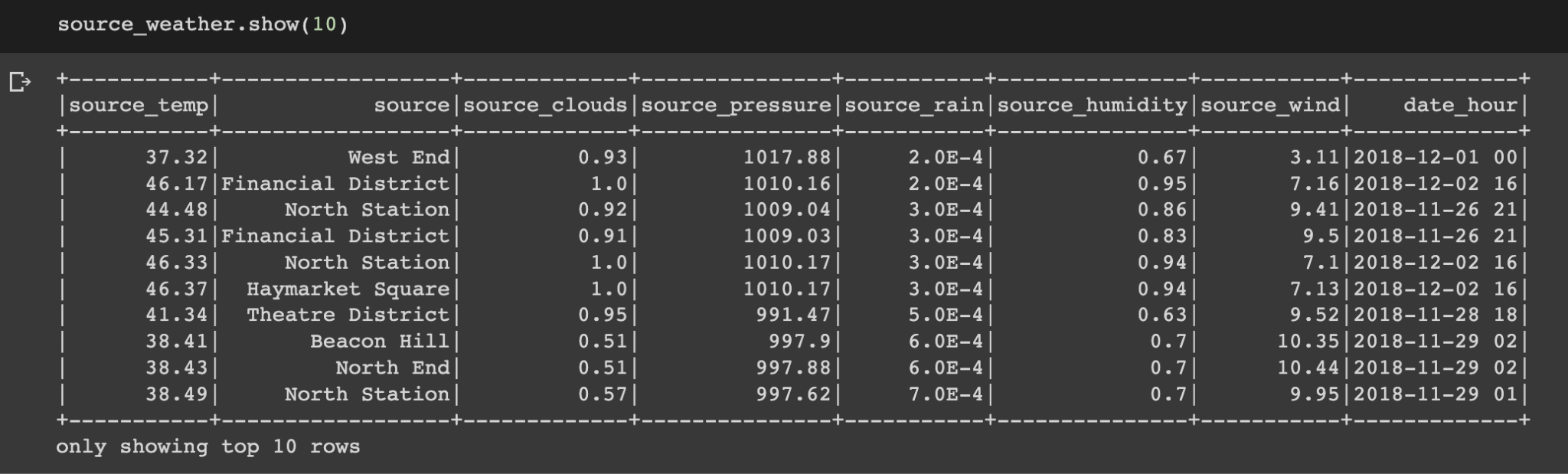
The last step we have done in the “cab\_ride” dataset is to remove the unused column “id”.

For the “weather” dataset, we go through the same procedure as we did in the “cab\_ride” dataset, we did from checking and filling in the missing values, and checking duplicate values, to removing outliers, change the timestamp to date time format and remove unused columns. The difference is that when we check outliers in both the “temperature” and “rain” variables, we found that outliers count 6.9% and 14.23% of the total number of records respectively. Therefore, we decide to replace those outliers with the upper and lower bounds instead of removing them.

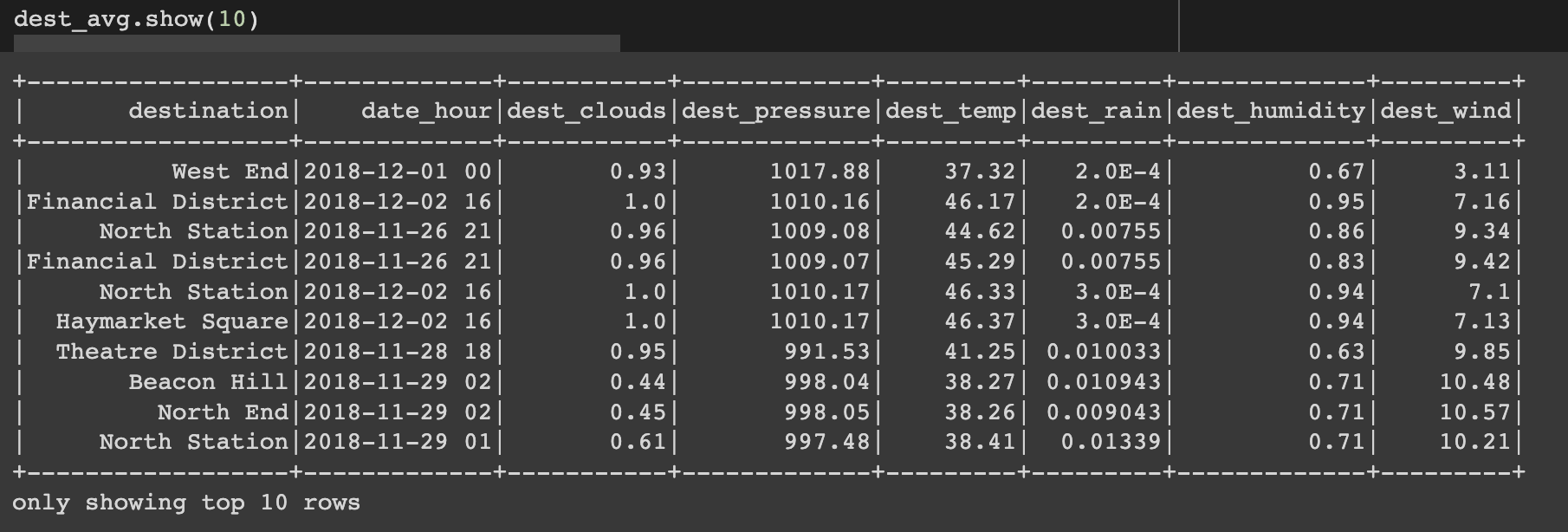


*Figure 7. The boxplot of “temperature” and “rain” variables in the “weather” dataset.*

After cleaning two dataset, it comes to the last step of the data cleaning process: we decide to **merge the “cab\_rides” and “weather” datasets** according to both the date time and location variables. Before merging two datasets, since there are “source” locations and “destination” locations in the “cab\_ride” dataset, to correctly match the location when join these two dataset, we make a copy of the “weather” dataset and named one as “source\_weather” dataset and another as “destination\_weather” dataset. Also, since there are multiple trips records on the same location at the same time, to avoid duplications, we calculate the average weather conditions for example the average temperature, average humidity, average air pressure and so on by grouping the data according to the date time and location. The resulting “source\_weather” and “destination\_weather” datasets are shown in figure 8 and figure 9 below.

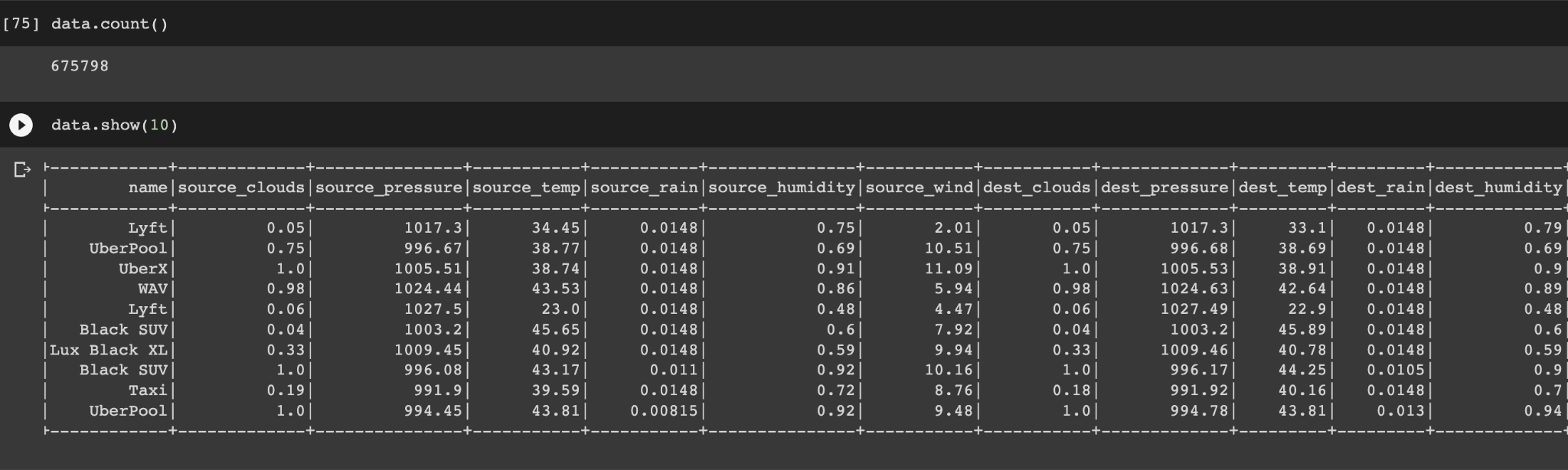


*Figure 8. Top 10 rows of the “source\_weather” dataset.*



*Figure 9. Top 10 rows of the “destination\_weather” dataset.*

After creating two copies of the “weather” dataset, we join them with the “cab\_rides' ' dataset by left join firstly on the “source” column and then left join on the “destination” column. The number of records after joining is still 675798 which means there is no data loss during merging the dataset. Since the complete dataset after merging has too many columns to show, we only show a part of columns here to show that the dataset has both the information of weather conditions on the “source” and “destination” locations.



*Figure 10. The top 10 rows and part of columns in the merged dataset.*

**3. Analysis**

**Dashboard**

图表

描述已自动生成

PySpark is a popular big data processing framework that is built on top of the Apache Hadoop ecosystem. It is designed to handle and analyze large-scale datasets in a distributed computing environment. In the case of the Uber & Lyft cab price dataset, the dataset consists of millions of rows, which can be computationally expensive to handle with traditional data analysis tools such as Pandas. Therefore, PySpark is an excellent choice for processing and analyzing this dataset because it can handle large volumes of data and perform parallel processing to optimize the analysis process. Additionally, PySpark provides various built-in functions and libraries for data manipulation and analysis, making it a powerful tool for processing and analyzing complex datasets.

**3.1 Insight 1: Analysis of Cab Usage Patterns in the Financial District-Insights on Preferred Cab Types and Ride Preferences**

Table 1 shows the top 5 most popular destinations, with Back Bay, Haymarket Square, and Theatre District being the top three. It is likely that these areas are popular tourist destinations or have many offices and businesses.

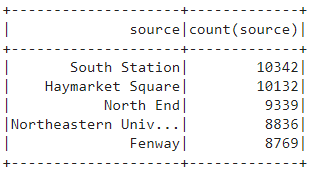
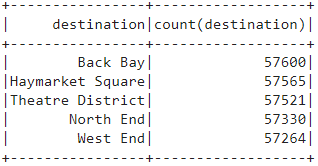
Table 1 shows the top 5 sources of rides, with South Station being the most popular. This could be due to the station's location and the fact that many commuters use it as a transportation hub. Haymarket Square and the North End are also popular starting points, possibly due to their central locations.

Table 2 shows that Uber is the most popular cab type, with Lyft coming in second. It is possible that Uber has a larger market share or a stronger presence in the areas where these rides originate or terminate.

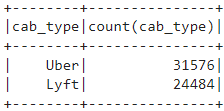
Table 3 shows the average price per mile and distance traveled for both Uber and Lyft. The average price per mile is higher for Lyft, but the average distance traveled is also longer for Lyft, resulting in a similar average price per trip for both Uber and Lyft. This may indicate that the pricing strategies of the two companies are different, with Lyft charging more per mile but offering longer rides.

Cabs are mostly used to take people to "Financial District", those people are mostly taking cabs from "South Station", and more people choose to take Uber to the "Financial District" than Lyft because on average, not only the price but also the distance by taking Uber from South Station to Financial District are lower than taking Lyft.

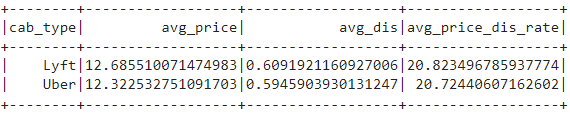
*Table 1. Top 5 popular destinations and sources*



*Table 2. Count of each cab type*



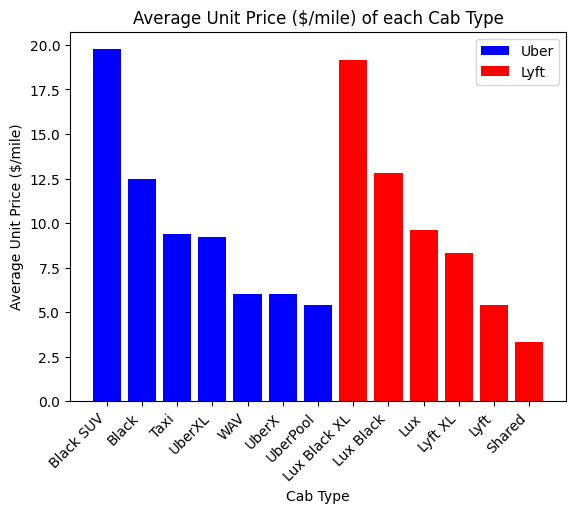
*Table 3. Average price by cab types*



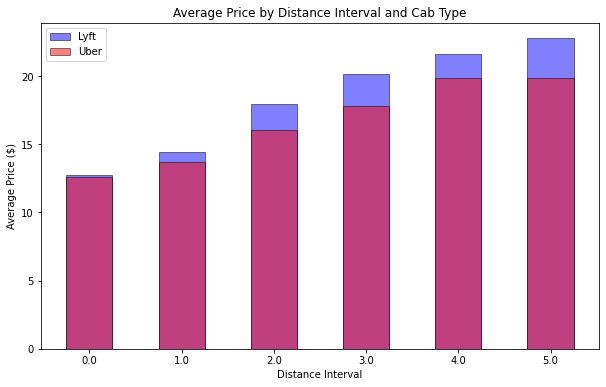
**3.2 Insight 2: Comparison of Price Increment for Uber and Lyft based on Ride Distance & Average unit price of each cab type.**

Figure 11 shows that ‘Black SUV’, ‘Lux Black XL’, and ‘Lux Black’ are the top 3 cab types with the highest average unit price($/mile), which means it is more expensive to take these cab types than others. These taxis may therefore be marketed to clients who are willing to pay more for luxury and comfort. While, ‘Shared’, 'UberPool’ and ‘Lyft’ are the cheapest cab types. We can conclude that for passengers who are more concerned with cost than luxury, they should choose 'Shared', 'UberPool', or 'Lyft' as their cab type. These cab types have the lowest average unit price per mile.

Figure 12 shows that for both Uber and Lyft, as the distance of ride increases, the fare charged to the passengers increases as well. However, the degree of increment is different between Uber and Lyft and the distance interval. Lyft has a slightly higher cab price increment than Uber. There is also no significant increment for Uber between “3-4” km and “4-5” km.



*Figure 11. Average unit price of each cab type*

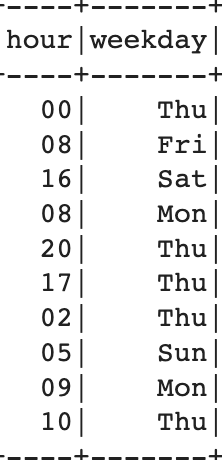
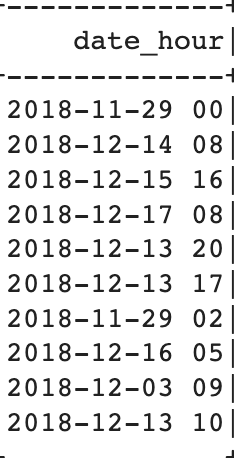


*Figure 12. Average price by distance interval and cab type*

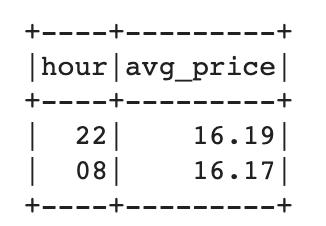
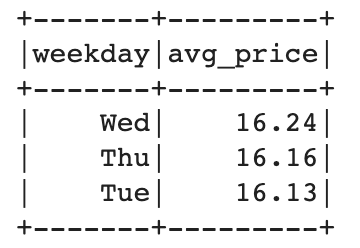
**3.3 Insight 3: Wednesday is the most expensive day to take cabs, followed by Tuesday and Thursday and Tuesday; 10 pm and 8 am are two of the most expensive hours to take cabs; and Monday morning is the cheapest moment to take cabs.**

To investigate the relationship between cab prices and time, the first thing we have done is to separate hours and weekdays from the original date time in the combined dataset, as shown in the tables below.

*Table 4. Before and after extracting hours and weekdays from date time.*

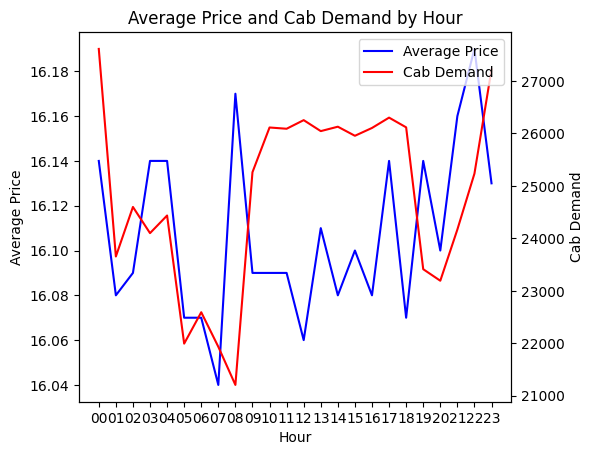


If we look at the average price of cabs calculated by each day in a week, we could find that the most expensive day to take cabs are Wednesday, followed by Thursday and Tuesday, as shown in the table below. If we see the average price hourly, 10 pm and 8 am are the two most expensive hours to take cabs.

*Table 5 & 6. The top 3 average prices of cabs on weekdays and the two most expensive hours taking cabs.*   


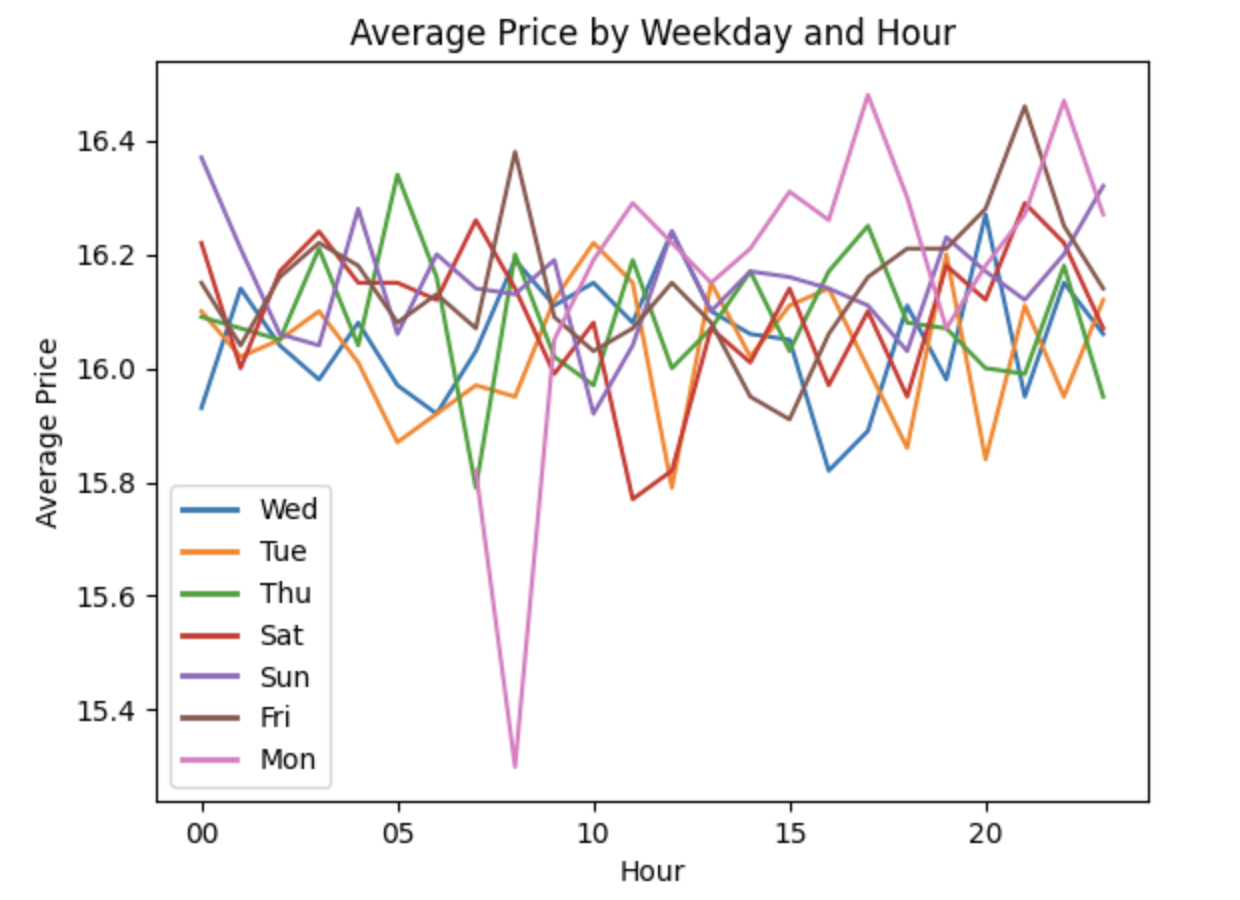
By investigating the trend line of both the average price and cab demands hourly, as shown in the figure below, we got some other interesting insights:

* After 8 am, the average price of taking a cab decreases sharply while the demand increases dramatically, it might because there are enough cabs during the morning period.
* After 6pm, the demand for cab rides decreases significantly while the average price keeps increasing. It might be because people prefer to take public transport to avoid traffic jams after 6pm. And the number of cabs decreasing after 6pm leads to the price increases.



*Figure 13. The hourly average price and cab demands.*

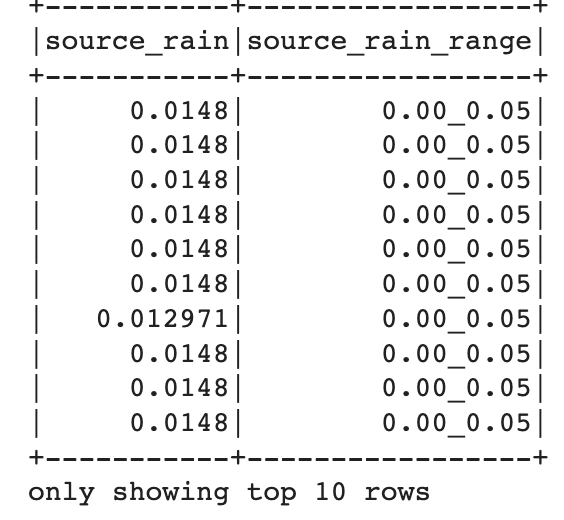
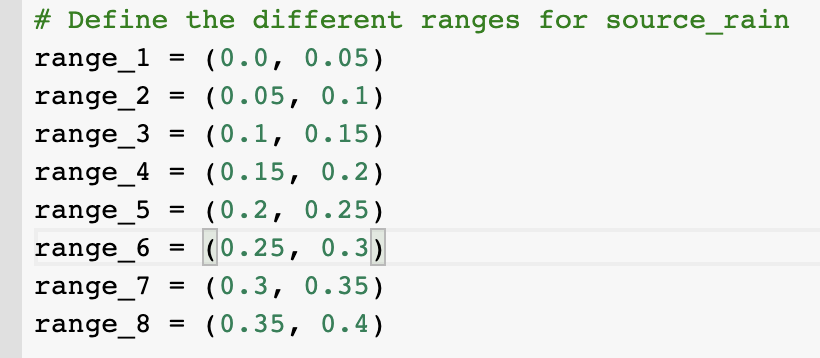
We also combine both the hour and weekdays to see a more general trend between time and cab average price. We found that most of the time, the average cab price fluctuates from 15.8 to 16.4, and it reaches its minimum on Monday morning.



*Figure 14. The average hourly cab price grouped by weekdays.*

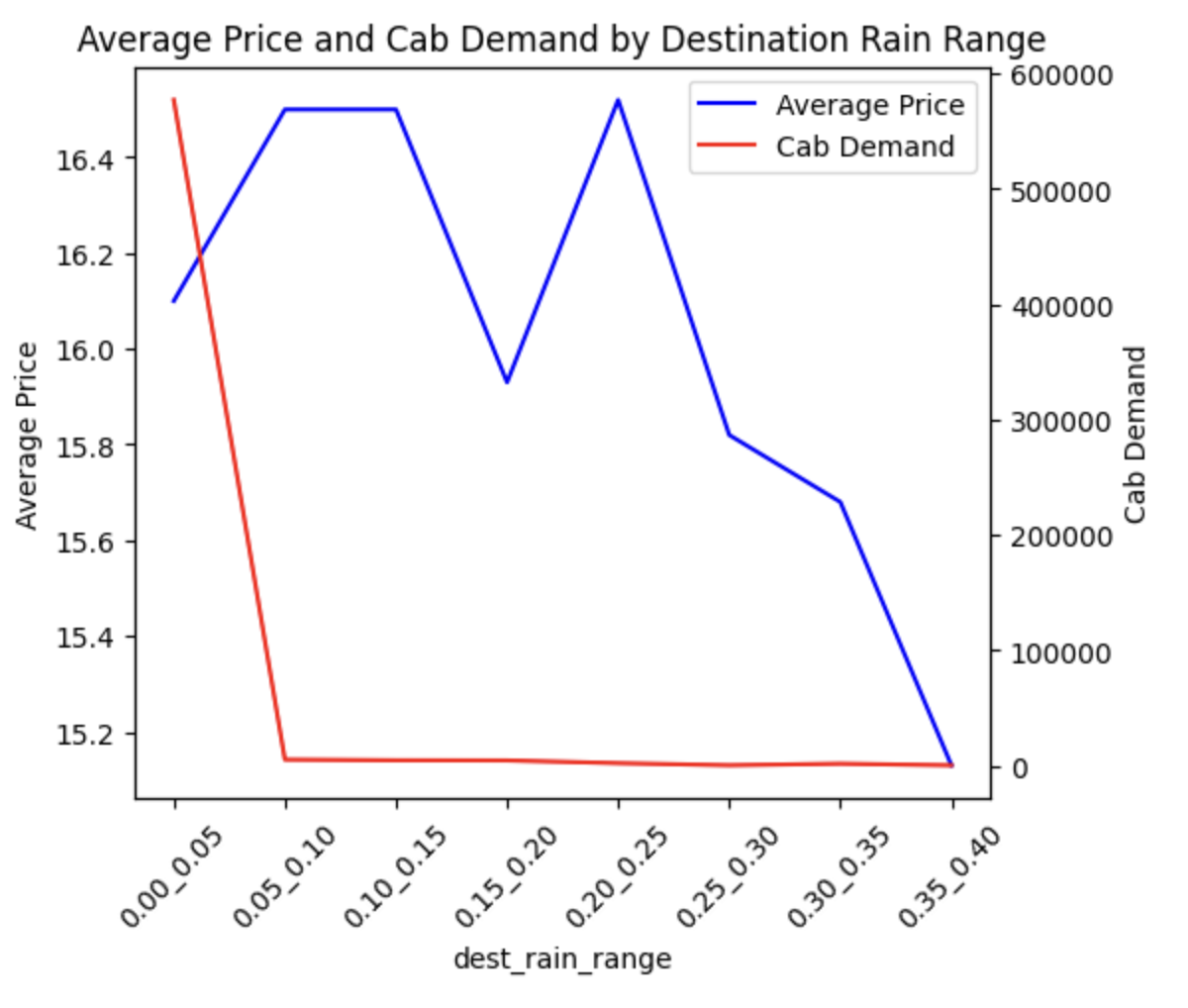
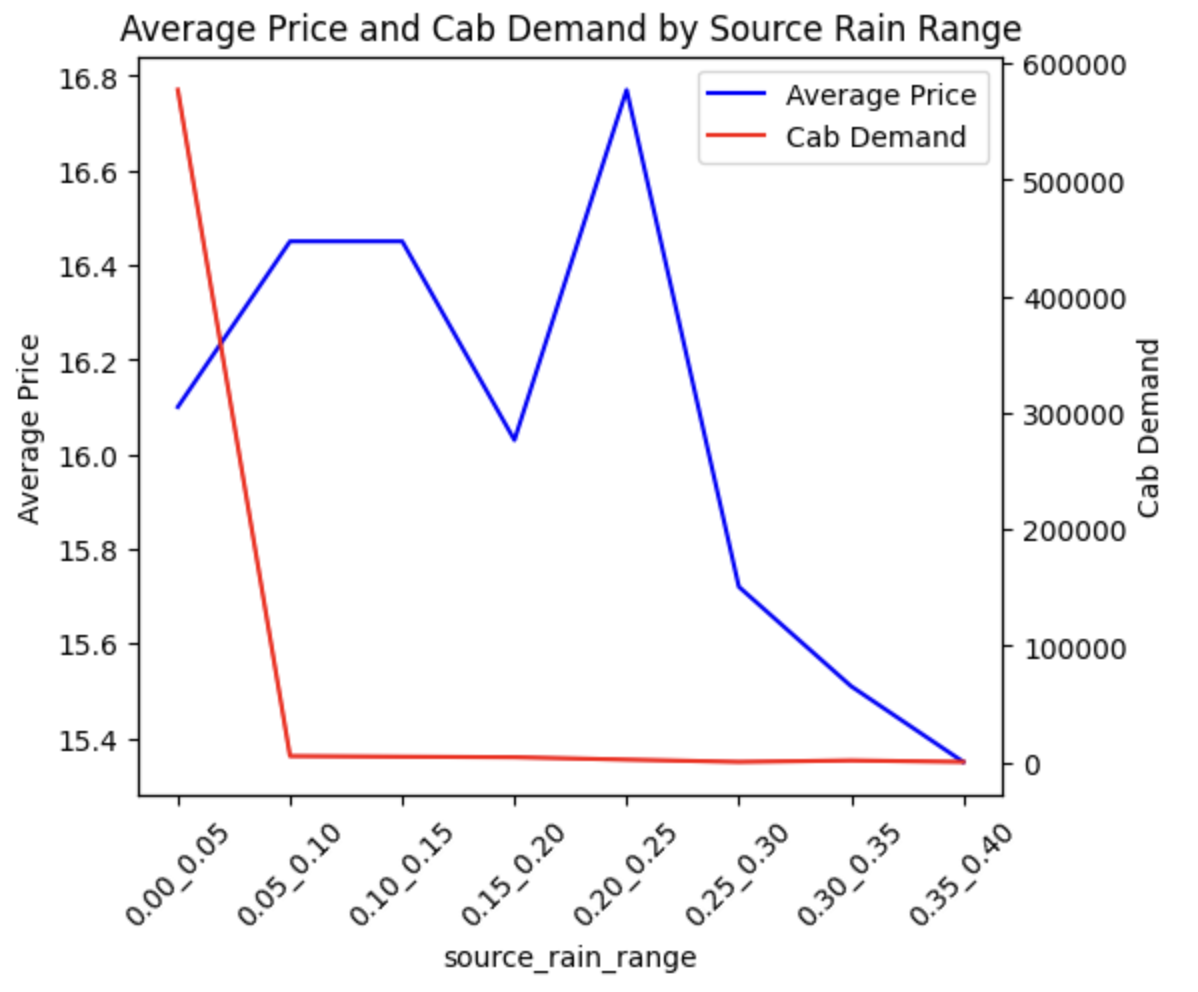
**3.4 Insight 4: The cab demand drops significantly when precipitation from 0 to 0.1, then remains roughly stable afterwards; The average price of cabs reaches a peak when precipitation is from 0.20 to 0.25 on both the source and destination locations, and both average prices drop after.**

Here we would like to investigate if there is any relationship between the weather conditions and the average price, demands of cab rides. The most commonly thought regarding the weather and the cab rides might be that: if the rainy day would affect both the price and demand of the cab?

To investigate the relationship, firstly we check the description of precipitation: the range of both the source and destination precipitation is from 0.0002 (almost 0) to 0.40. Therefore, we decide to divide the precipitation range into 8 smaller gaps of 0.05, the details of subgroup ranges and the result are shown in the figure below:  
 

*Figure 15. The subgroup range of precipitations and sample result.*

Then we draw a trend line of both the average price and cab demands for each range of precipitation. We got the demand for cab rides by calculating the number of prices used to calculate the average price in each range of precipitation. By dividing the precipitation of both the source and destination locations, we got two trend lines graphs as shown in the figure below:



*Figure 16 & 17. The average price and cab demand by precipitation on source and destination locations.*

By looking at the trend line of precipitation of both the source and destination locations, we could get two insights here:

1. The cab demand drops significantly when precipitation from 0 to 0.1, then remains roughly stable afterward.
2. The average price of cabs reaches its peak when precipitation is from 0.20 to 0.25 on both the source and destination locations, and both average prices drop after.

**4. Conclusions and Recommendations**

**4.1 Conclusions**

Based on the analysis provided, several insights can be drawn for the business problems that were posed at the beginning of the project.

The study of taxi usage trends for the first issue showed that South Station is the most popular starting point in the Financial District and Uber is the most popular cab type there. Back Bay, Haymarket Square, and Theatre District were the most well-liked travel locations. These places might be well-liked vacation spots or host a large number of offices and businesses. Companies can use this knowledge to plan their operations and cater to the needs of their customers.

The top 3 taxi types with the highest average unit price were found to be "Black SUV," "Lux Black XL," and "Lux Black," while the most affordable cab types were "Shared," "UberPool," and "Lyft." The analysis also showed that while the amount of the increase varies between Uber and Lyft, both businesses charge more as the distance of the trip increases. Customers can use this information to make more informed choices about when and which taxi service to use.

In relation to the third issue of time and day analysis, it was discovered that Wednesday, along with Tuesday and Thursday, is the most expensive day to use a cab. The most expensive times to hire a cab are at 10 p.m. and 8 a.m., while Monday morning is the least expensive. Intriguing patterns in the connection between taxi prices and time were also discovered through the research, including a sharp drop in average cab prices and a rise in demand after 8 a.m. Customers can use this information to plan their trips more cost-effectively, and cab firms can use it to better plan their pricing strategies.

In the 4th insights, it implies a strong correlation between precipitation and the average fare and desire for cab rides. The analysis demonstrates that when precipitation is between 0 and 0.1, the demand for taxis considerably decreases before stabilizing. This suggests that people are less likely to use cabs during light rain, which makes sense given that people may choose to walk or take public transportation in such weather circumstances instead of taking a cab.

The analysis also shows that when precipitation is between 0.20 and 0.25 on both the source and destination locations, the average price of taxis hits its peak and then declines. This suggests that there is a rise in cab demand during moderate rain, which in turn causes a rise in cab rates. But when the amount of precipitation exceeds this range, there is a decline in demand, which lowers the average price.

**4.2 Recommendations**

Cab usage patterns: Analysis of the Financial District's cab usage trends can help businesses better understand the market and serve customers' needs. To meet the needs of the Financial District market, taxi companies could provide more Uber choices.

Price comparison: Customers can use the information on the comparison of price increments for Uber and Lyft based on ride distance and cab type to make more informed decisions about which cab service to use and when. Cab companies can adjust their pricing strategies accordingly to attract more customers.

Time and day analysis: Understanding the most and least costly times and days to take cabs can both help customers and cab companies plan their pricing strategies more skillfully. To maximize their income, taxi companies can also change the number of cabs on the road based on demand at various times and days.

Weather and demand analysis: Based on the analysis of how precipitation affects cab demand and pricing, cab firms can plan their operations for various weather conditions. When it's raining, taxi firms can change their pricing strategies to increase profits and draw more customers.

**Reference:**

* Google. (n.d.). Google collaboratory. Google Colab. Retrieved March 23, 2023, from <https://colab.research.google.com/github/justinbois/bootcamp/blob/gh-pages/2021/lessons/l29_dashboards.ipynb>
* RaviMunde. (2019, June 23). Uber &amp; Lyft Cab Prices. Kaggle. Retrieved March 31, 2023, from <https://www.kaggle.com/datasets/ravi72munde/uber-lyft-cab-prices?datasetId=195655&amp;sortBy=voteCount&amp;select=cab_rides.csv>