**Objective**

The objective of this assignment is to determine what are the best conditions to get an Uber or Cab in NYC? Another question to answer through the analysis of this data is what conditions affect fare calculation for Cabs and Ubers in the NYC area?

**Obtain**

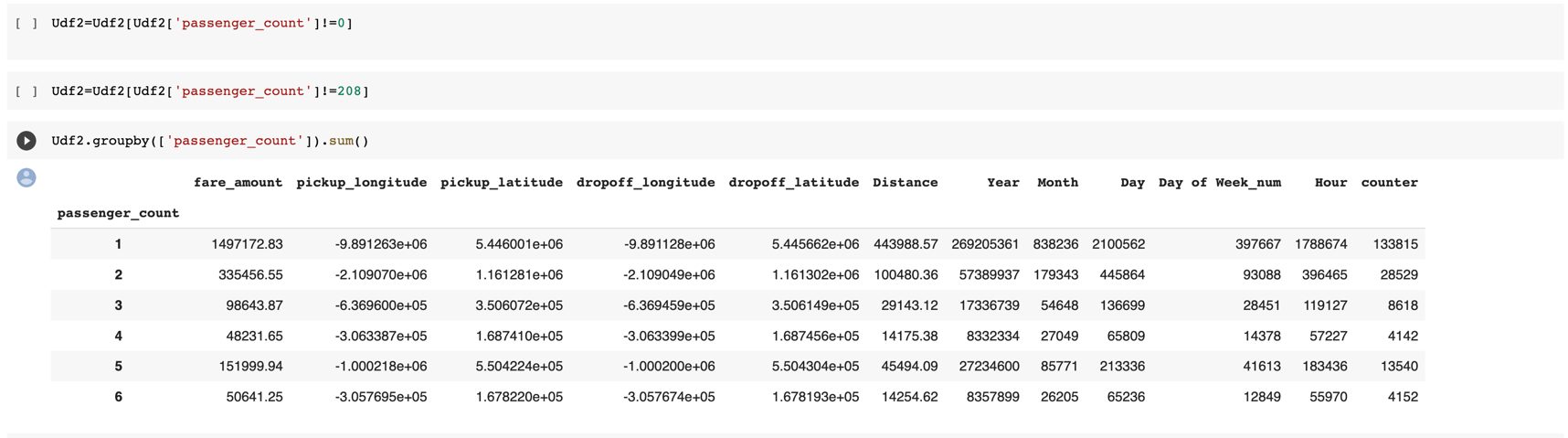
We gathered the data from Kaggle. We used an UBER and Taxi dataset to complete this project.

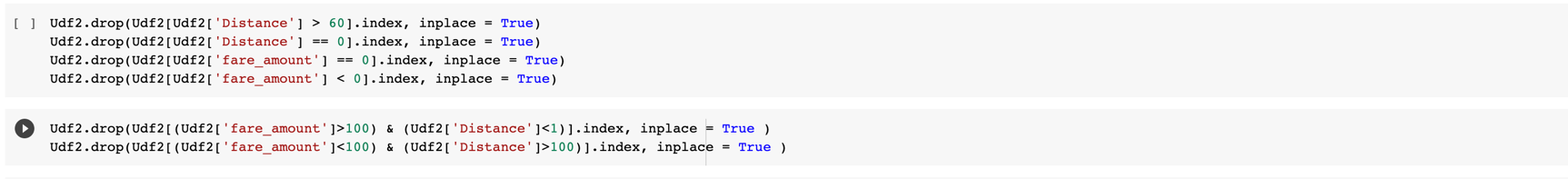
<https://www.kaggle.com/datasets/yasserh/uber-fares-dataset>

**Scrub**

A main tactic used in cleaning data was getting the date and time in a readable format. This was necessary to track what years, months, days, times were the most popular. Also, it was important to clean the latitude and longitudes so we could calculate the distance of the trip.

Then we got rid of the outlier data and data that made no sense to get ready for our modeling. We took out all trips that had zero passengers or the one trip with 208 passengers. We also trimmed any trip that had a distance over 60 kilometers or 0 kilometers. We also trimmed any trip that had a fare amount of 0 or less than 0. Then finally we trimmed any trip with a fare amount of more than 100 and less than 1 kilometer and any trip with a fare amount of less than 100 and a distance of more than 100 kilometers.





**Explore**

Getting those initial visuals of the breakdown between years, months, days, times were extremely helpful to compare between the two datasets. Also the use of maps and graphs of trip distance helped us additionally explore the data.

Goal – understand the conditions that determine fares for UBER and Cabs in NYC

Uber Data Set – Variables Used

* Pick-up time
* Pick up/Drop Off Longitude & Latitude
* Passenger count

**A picture containing text, monitor, screenshot, screen

Description automatically generated**

**TAXI Data Set – Variables Used**

* **Pick-up/Drop off**
* **date & time**
* **Trip Distance**
* **Passenger count**

**A black screen with white text

Description automatically generated with low confidence**

**Analysis to understand relationship between Passenger Count and Fare Amount for Taxi and UBER**

**Chart, scatter chart

Description automatically generated**

**Table

Description automatically generated**

**\* Uber and Taxi Data set confirm that passenger count does not impact Fare amount**

**Visualization to understand if distance impacts fare amount.**

Chart, scatter chart

Description automatically generated

**Visualization to understand UBER data per in a yearly, monthly and daily basis.**

A screenshot of a computer

Description automatically generated with medium confidence

**Analysis of Taxi and UBER data of trips per hour**

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

* In the Uber data set, trips increase as the time progresses throughout the day with its peak being in the night time. This is easily explained by people going out to eat in the evening and most likely having something to drink so the need for alternate transportation in turn increases
* In the case of the Taxi data set seems like trip per hour exhibits the same behavior as the UBER dataset an increase in trips as time progresses but what is interesting is that the numb er of trips decreases very late at night. This could be explained by the availability to NYC cabs at that time during the day or more competitive fares by UBER.

**Analysis of Taxi and UBER data of trips per weekday**

Chart

Description automatically generated Chart, bar chart

Description automatically generated

* Overall volume of trips in by weekday is higher for UBER than Taxis.
* The reason for this could be the availability of UBER vs Taxis. In nature, UBERS will have more reach and availability of vehicles given the number of signed up drivers which can use their own vehicle this gives UBER more opportunities abd greater reach in therms of vehicle availability. In the case of TAXIS their vehicle availability is limited given that they use certain type of vehicle and a pool of taxi drivers in which this model is more restrictive yielding less drivers and vehicles, in turn less trups.
* Also, UBER is more convenient than a TAXI givne their application.

**Analysis of Taxi and UBER data of trips by day**

Graphical user interface

Description automatically generated Chart, bar chart

Description automatically generated

* Uber trips by day of the month is consistent compared to trips by day for Taxis.

UBER Descriptive Stats

Graphical user interface, text

Description automatically generated

\* Average fare amount for BER is $11.31 with a standard deviation of $9.50

\* Average passenger count is 1-2 people per trip

\* Average distance per trip is 3.35 miles

Correlation Matrix

Chart

Description automatically generated

Here is a correlation plot. Here we are trying to see if any other variable has a positive or negative correlation to fare amount. Here we saw a fairly large correlation of distance at 0.89 to fare amount.

**HEATMAP**

Chart

Description automatically generated with low confidence

Here is a heat map. Unfortunately, the data shown here completely contradicts the graphs we made earlier. This map shows that 4am and 5am to be the busiest during most days.

Chart

Description automatically generatedHere we made a graph showing the target variable distribution of fare amount. We wanted to see if the data was skewed in any way. As you can see, the data is skewed to the left showing that most of the trips were on the cheaper side rather than the expensive side.

**FOLIUM PICTURE**

Map

Description automatically generated

Here is a picture of a Folium plugin I made to show where the bulk of the trip data was taking place in New York. While a few trips do end up outside of New York mostt of them stay in the New York City area.

**Linear Regression Model**

Chart, scatter chart

Description automatically generated

Here is a picture of our linear regression. We split our data into a training and test set. It seems like more of the data lands farther above the regression line in the training set than the test set, which is quite interesting.

**MODELING**

Graphical user interface

Description automatically generated with medium confidence

Here are the outputs for the models that we ran. We decided to record root mean square error and the adjusted R squared. Root mean squared error measures the average difference between values that our models predict and the actual values. We got a root mean squared error of 0.44 for our K-Nearest Neighbors, linear regression and decision tree models. We got an adjusted R squared of 0.62 and 0.77 for our linear regression and Random Forest models. These scores tells us that there are other variables that we don’t have the do have a good say in the fare amount for an Uber trip.

**Graphical user interface, application

Description automatically generated**

The taxi dataset was able to have more accurate models. Both predictors were extremely high, with the linear regression getting the slight edge.

**OVERALL OBSERVATIONS**

* 2012 was the busiest year of the data​
* Most trips take place in the spring months​
* Fare amount and trip distance are highly correlated​
* Taxi models were better predictors overall

**UBER SPECIFIC OBSERVATION**

* Thursday – Saturday are the most popular days​
* 6pm – 12am are the busiest times​
* Most trips take place in the middle of the month​
* Random Forest was a better predictor of fare amount

**TAXI SPECIFIC OBSERVATION**

* Wednesday – Friday were the busiest days for taxi​
* ​4pm – 7pm were the most travelled hours​
* No clear observations from days of the month​
* Models were better for predicting fare amount than Uber
* Linear Regression was a better predictor than Random Forest

**RECOMMENDATION/CONCLUSION**

* The best option in terms of vehicle availability at different times and days during the day is UBER which in turn makes UBER fares more competitive.
* Additionally, you can expect the spring months, in the middle of the month, and in the evening hours to be the most expensive scenario to get an Uber
* Having more passengers doesn’t necessarily mean you will pay more, unless it is a sizable group
* Shorter rides = less expensive

**APPENDIX**

[**https://www.kaggle.com/datasets/yasserh/uber-fares-dataset**](https://www.kaggle.com/datasets/yasserh/uber-fares-dataset)

[**https://www.kaggle.com/datasets/anandaramg/taxi-trip-data-nyc**](https://www.kaggle.com/datasets/anandaramg/taxi-trip-data-nyc)