# UBER FARE PREDICTION

Author: Aswathi Sasikumar

# Summary

- The project is on world's largest taxi company Uber inc. In this project,I have tried to predict the fare for their future transactional cases.
- Uber provides service to a large number of customers daily. So it is important to manage their data properly to come up with new business ideas to get best results. And also it is important to estimate the fare prices accurately.

# Outline

- Business Problem
- Data
- Methods
- Results
- Conclusions

# Business Problem

- This project will help the stakeholder to predict the fare amount for the future transactions.

# Data

- Uber data set contained 200,000 records

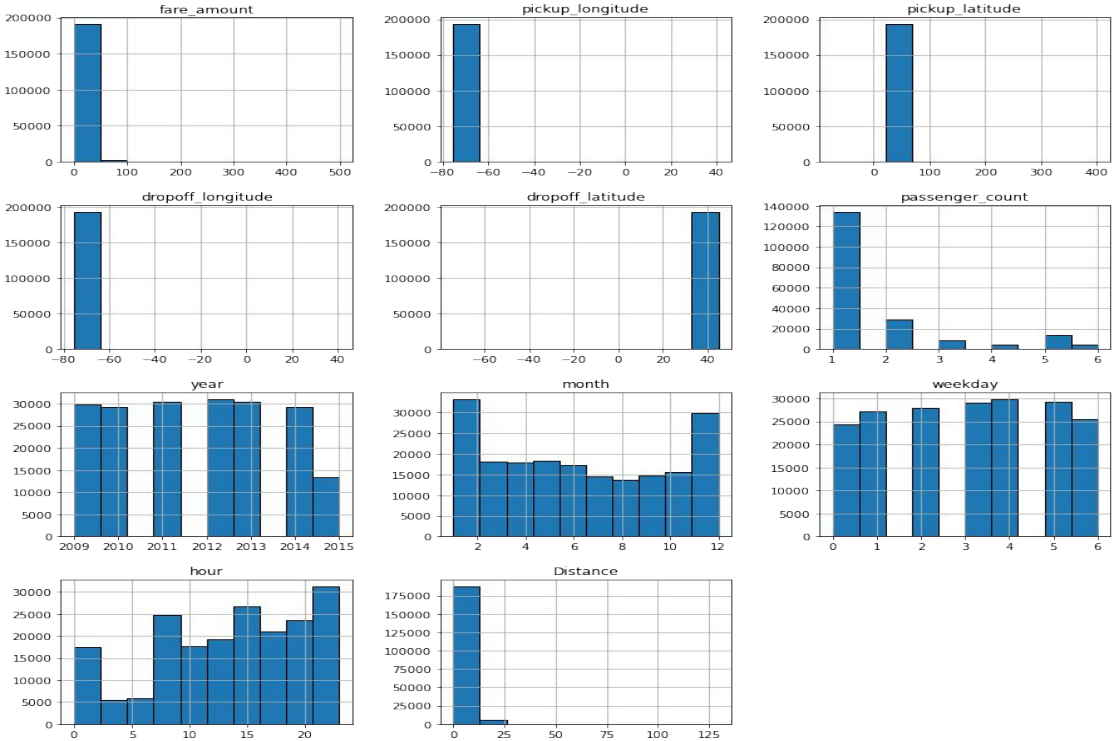
The datset contains the following fields:

- key - a unique identifier for each trip
- fare_amount - the cost of each trip in usd
- pickup_datetime - date and time when the meter was engaged
- passenger_count - the number of passengers in the vehicle (driver entered value)
- pickup_longitude - the longitude where the meter was engaged
- pickup_latitude - the latitude where the meter was engaged
- dropoff_longitude - the longitude where the meter was disengaged
- dropoff_latitude - the latitude where the meter was disengaged

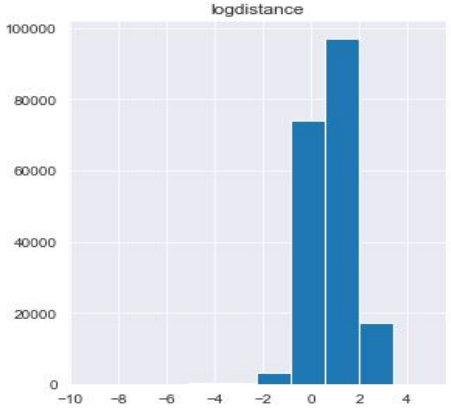
# Methods

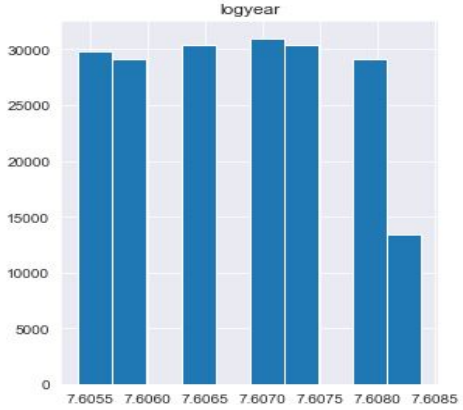
1. **Exploratory data analysis was used to analyse the data and to draw conclusions about the fare amount**
    - Invalid data was found in the dataset like fare_Amount<=0,distance >130kms which was practically impossible.
    - The dataset had lot of outliers and most of the variables were not normally distributed which was challenge in the project.
    - The data set had only latitudes and longitudes from which distance was calculated.
    - Dummy variables were created for categorical variables like passenger count,day of the week,month
    - Log transformations were done to non-normal distributed variables.
2. **After cleaning the data ,Multiple regression using OLS statsmodels was used to develop a model with fare_amount as dependent variable.**
3. **Assumptions of Linear Regression:**
    - Linearity
    - Normality (Residuals)
    - Homoscedasticity

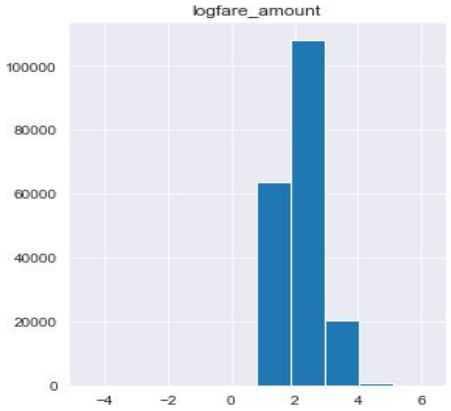
# Distribution visualization

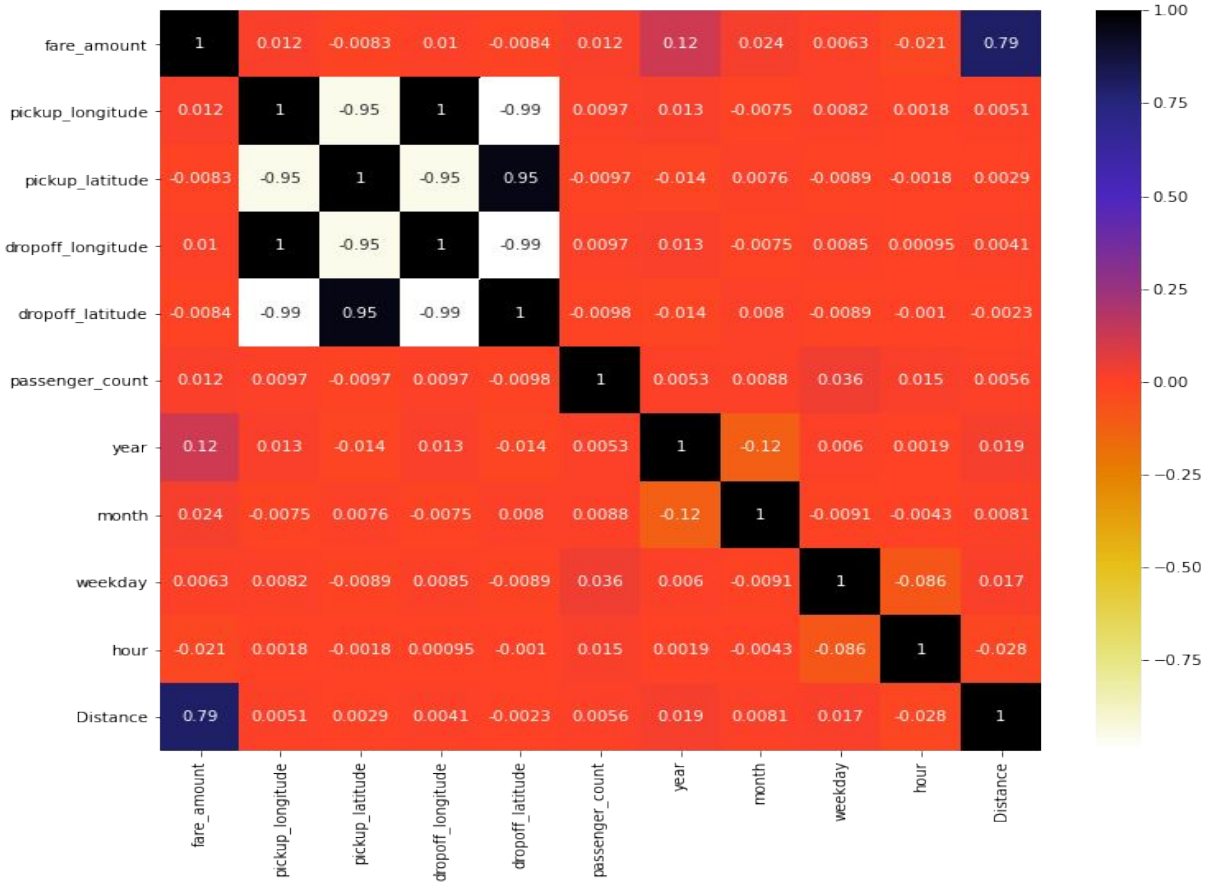


# Log transformation




fare_amount 1 0.012 -0.0083 0.01 -0.0084 0.012 0.12 0.024 0.0063 -0.021 0.79
pickup_longitude 0.012 1 -0.95 1 -0.99 0.0097 0.013 -0.0075 0.0082 0.0018 0.0051
pickup_latitude -0.0083 -0.95 1 -0.95 0.95 -0.0097 -0.014 0.0076 -0.0089 -0.0018 0.0029
dropoff_longitude 0.01 1 -0.95 1 -0.99 0.0097 0.013 -0.0075 0.0085 0.00095 0.0041
dropoff_latitude -0.0084 -0.99 0.95 -0.99 1 -0.0098 -0.014 0.008 -0.0089 -0.001 -0.0023
passenger_count 0.012 0.0097 -0.0097 0.0097 -0.0098 1 0.0053 0.0088 0.036 0.015 0.0056
year 0.12 0.013 -0.014 0.013 -0.014 0.0053 1 -0.12 0.006 0.0019 0.019
month 0.024 -0.0075 0.0076 -0.0075 0.008 0.0088 -0.12 1 -0.0091 -0.0043 0.0081
weekday 0.0063 0.0082 -0.0089 0.0085 -0.0089 0.036 0.006 -0.0091 1 -0.086 0.017
hour -0.021 0.0018 -0.0018 0.00095 -0.001 0.015 0.0019 -0.0043 -0.086 1 -0.028
Distance 0.79 0.0051 0.0029 0.0041 -0.0023 0.0056 0.019 0.0081 0.017 -0.028 1
fare_amount
pickup_longitude
pickup_latitude
dropoff_longitude
dropoff_latitude
passenger_count
year
month
weekday
hour
Distance
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

# Results

It was found that fare_amount increases with distance which is obvious.It was found that fare_Amount tends to increase during non-peak hours.

Regression models:

1. Model 1:

   Independent variables : all fields

   R-squared: 0.614

   Normality assumption was not satisfied.

2. Model 2:

   Independent variables : all fields except correlated variables dropoff_latitude', 'dropoff_longitude', 'pickup_latitude'

   R-squared: 0.806

   Normality assumption was not satisfied.

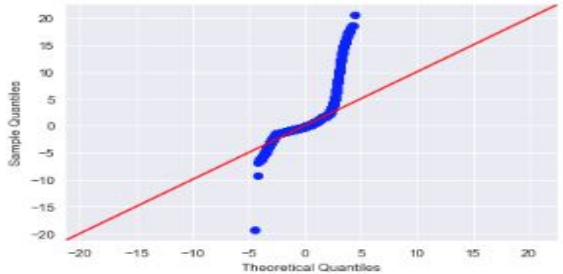
# Model 3:Independent variables: distance and hour

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          logfare_amount   R-squared:                       0.593
Model:                             OLS   Adj. R-squared:                  0.593
Method:                  Least Squares   F-statistic:                1.409e+05
Date:                 Tue, 21 Jun 2022   Prob (F-statistic):               0.00
Time:                         20:47:01   Log-Likelihood:                -88107.
No. Observations:               193218   AIC:                         1.762e+05
Df Residuals:                   193215   BIC:                         1.763e+05
Df Model:                            2
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.8296      0.002    854.810      0.000       1.825       1.834
logdistance    0.4843      0.001    530.703      0.000       0.483       0.486
hour          -0.0002      0.000     -1.469      0.142      -0.000    6.55e-05
==============================================================================
Omnibus:                   178710.333   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         18549336.934
Skew:                           4.126   Prob(JB):                         0.00
Kurtosis:                      50.286   Cond. No.                         37.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [72]:

```
import scipy.stats as stats
residuals = model3.resid
sm.graphics.qqplot(residuals, dist=stats.norm, line='45', fit=True, )
plt.show;
```



# Model validation

In [82]:

```
from sklearn.metrics import mean_squared_error

train_mse = mean_squared_error(y_train, y_hat_train)
test_mse = mean_squared_error(y_test, y_hat_test)
print('Train Mean Squarred Error:', train_mse)
print('Test Mean Squarred Error:', test_mse)
```

```
Train Mean Squarred Error: 0.1469148176444233
Test Mean Squarred Error: 0.14109311700096908
```

The Mean Squarred Error for the train and test subsets are similar. This suggests that the model will perform similarly on different data.

Distance,hour are the best fit for a multiple regression model. These features are highly correlated with fare amount, have relatively low multicollinearity, and can together account for more than half of the variability of price. All multiple regression assumptions are satisfied with these features included.

# Conclusions

- It was found that fare_Amount increases with distance which is obvious.It was found that fare_Amount tends to increase during non-peak hours.

# Limitations

Modeling the data with other types of regression analysis techniques like logistic regression might give more accurate model.

# Thank You!

Email: kukkuaswathi@gmail.com
GitHub: https://github.com/AswathiSasikumar
LinkedIn:
https://www.linkedin.com/in/aswathi-sasikumar-21989333/