Resurgence of NYC Taxis

Apurva Sharma
Sanman Yadav
Shama Kamat
Shivanshi Bajpai

Table of Content

Problem Description

Issues that need to be addressed in a reform process for Yellow Cab drivers:

- Information Unavailability:
  - No get real-time updates on availability of passengers.
  - Majority of their lost time in the day is spent searching or waiting for customers
- Fare negotiation:
  - Availability of smartphone apps for uber/lyft etc makes passenger informed
  - Time spent in making a proposal and waiting to make a deal

Data Source

City of New York Dataset

https://data.cityofnewyork.us/Transportation/2015-Yellow-Taxi-Trip-Data/ba8s-jw6u

Data Dictionary:

http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

Federal Holiday Dataset

https://catalog.data.gov/dataset/federal-holidays

Weather Dataset

https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2015

Data Description

Full Merged Dataset : NYC Yellow Taxi Trip Records + USA Federal Holidays Dataset + NYC Weather Data

(Number of data points) : 12 Million rows

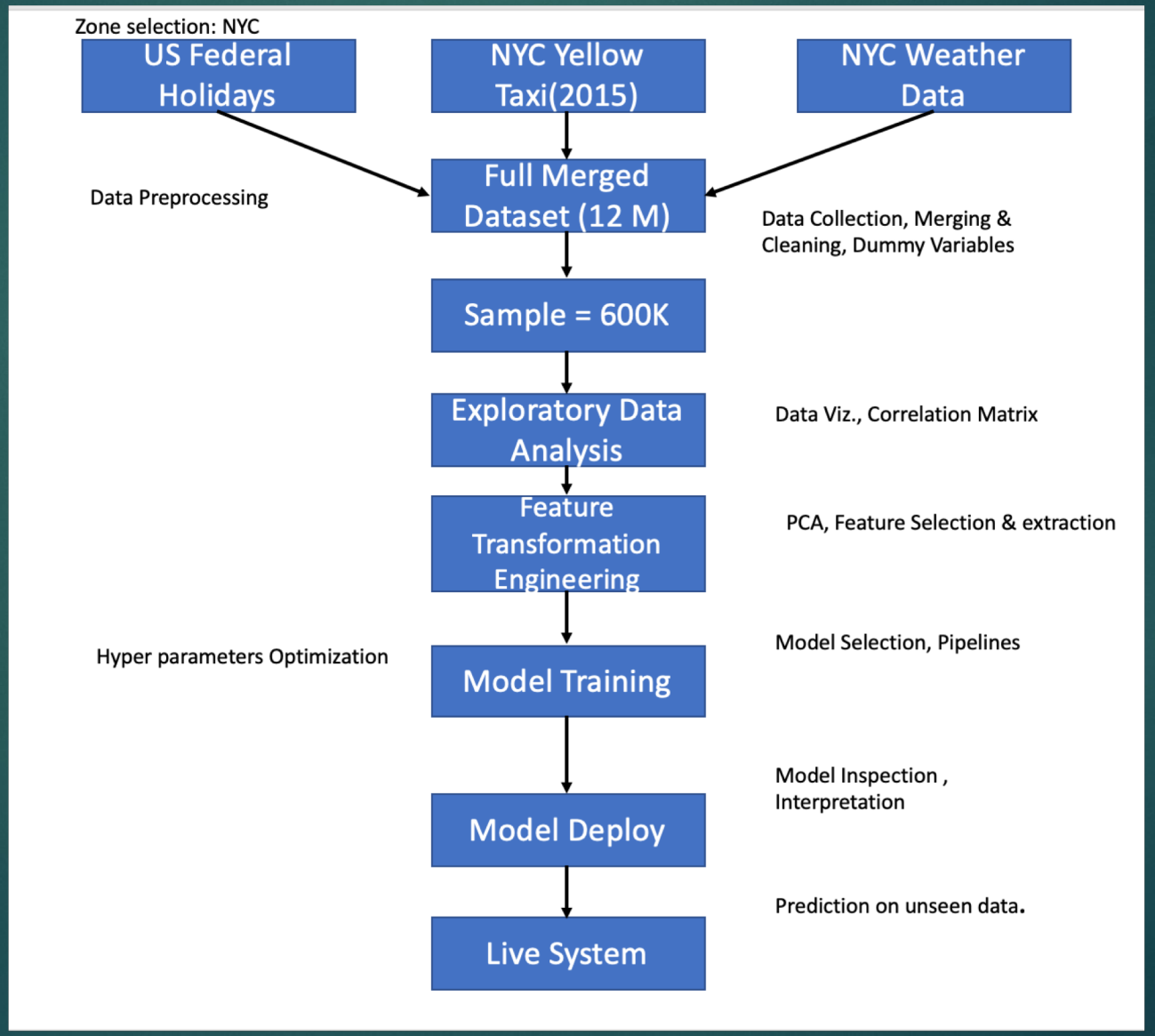
Training data points out of Sample Dataset: 600 K

Total Features: 61 (after cleaning and dummy variables creation)

Cleanliness: Merging, Date time conversion, Imputation, Dummy variables, Standardization

Label/output to predict: Duration and Total Amount

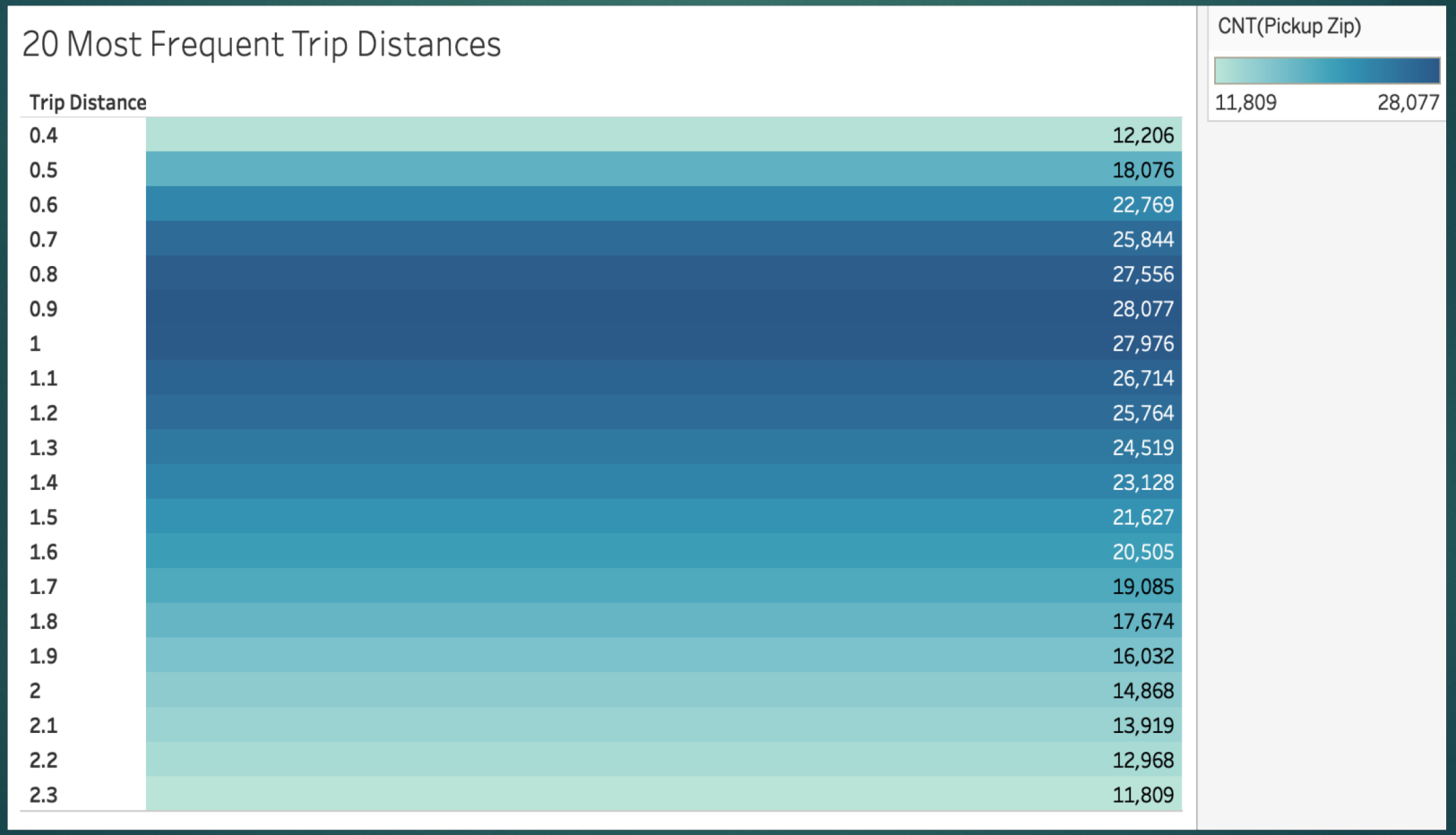
Project Approach



Feature Description

Descriptive statistics and Exploratory Data analysis for factors Affecting Duration and Total Amount

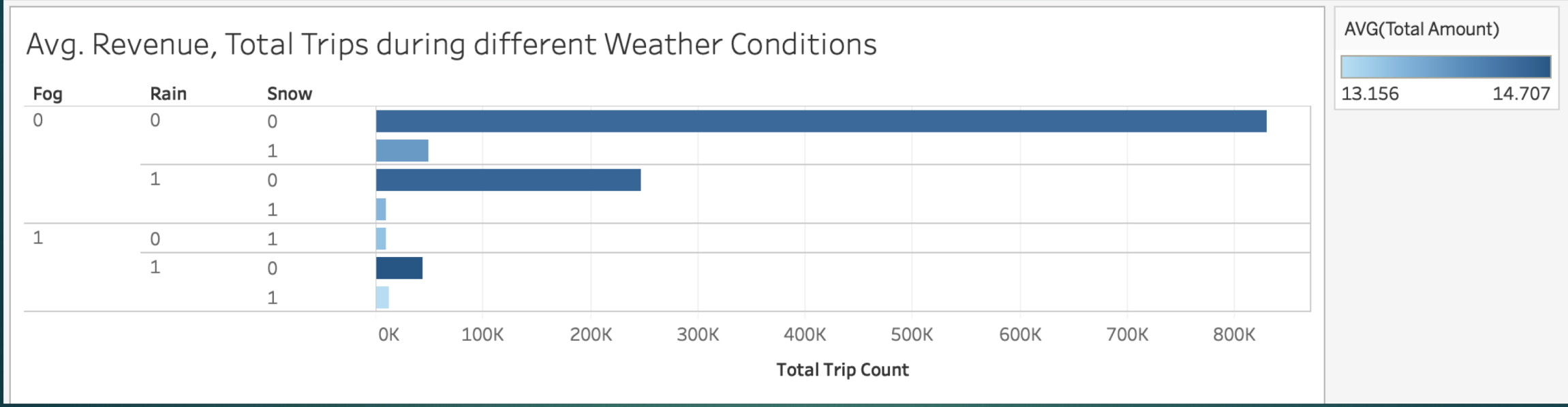
A) Trip Distance (miles



Factors Affecting Duration and Total Amount

B) Weather

For yellow cab drivers, a bright sunny day is the best day to earn higher revenues on a trip



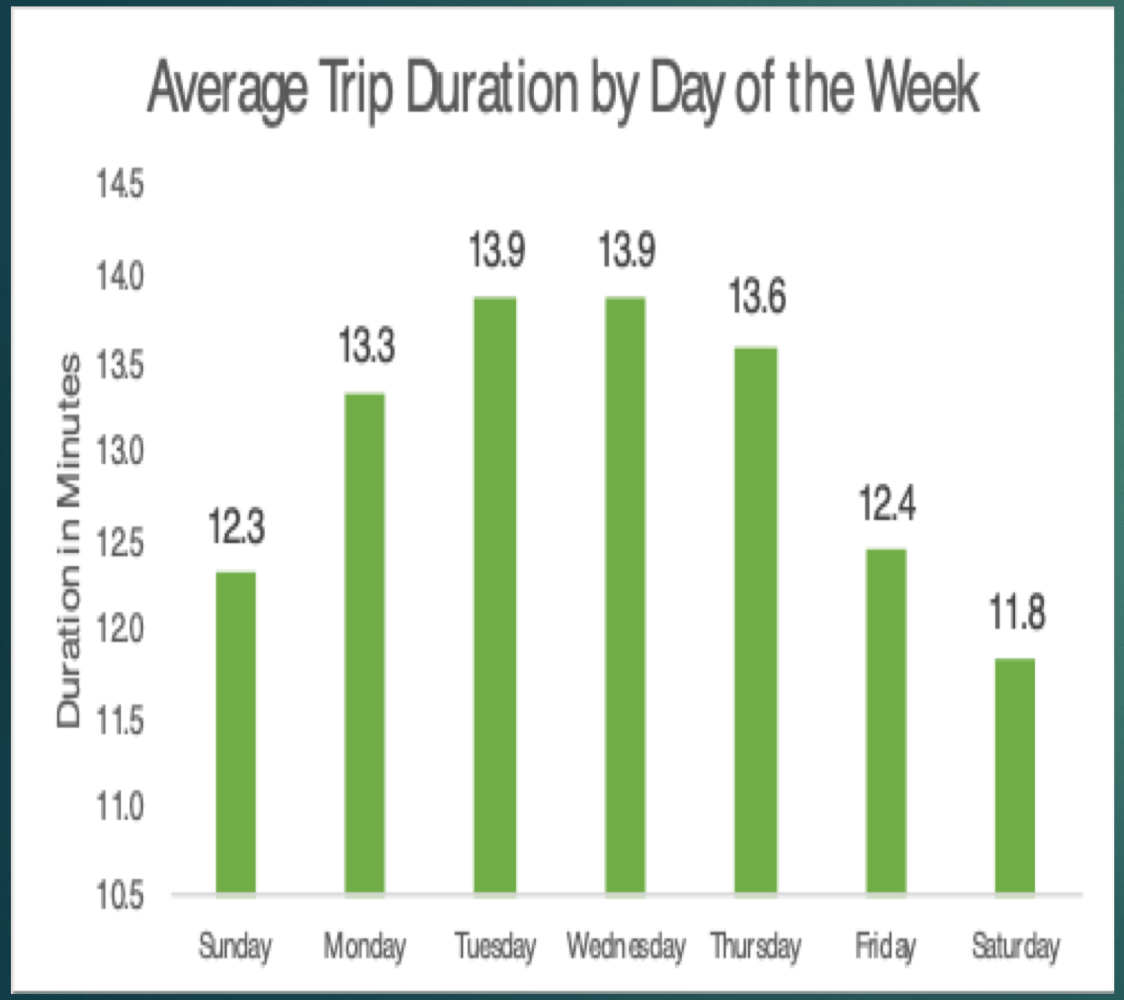
Caption

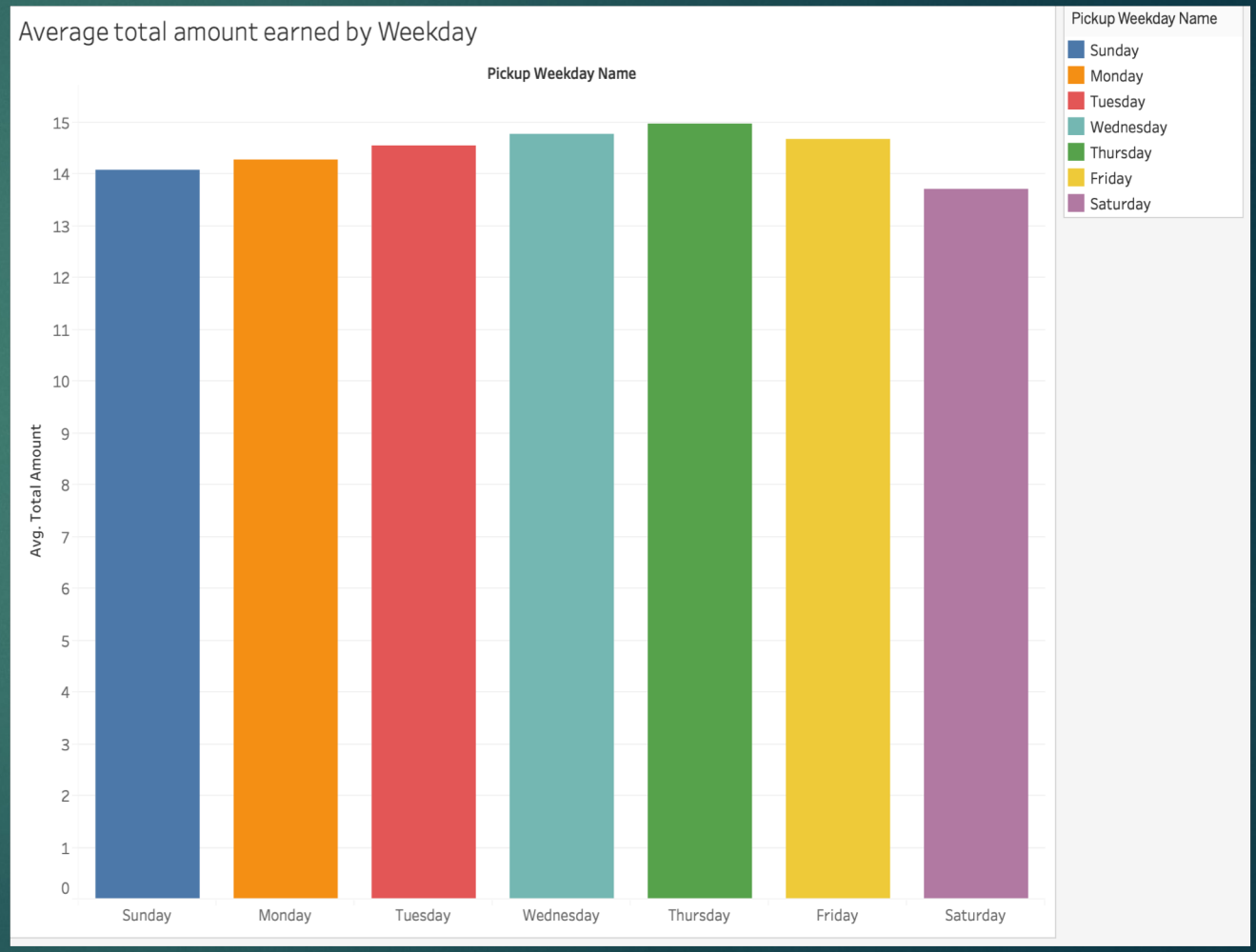
Count of Dropoff Zip for each Snow broken down by Fog and Rain. Color shows average of Total Amount.

Factors Affecting Duration and Total Amount

C) Day of the Week

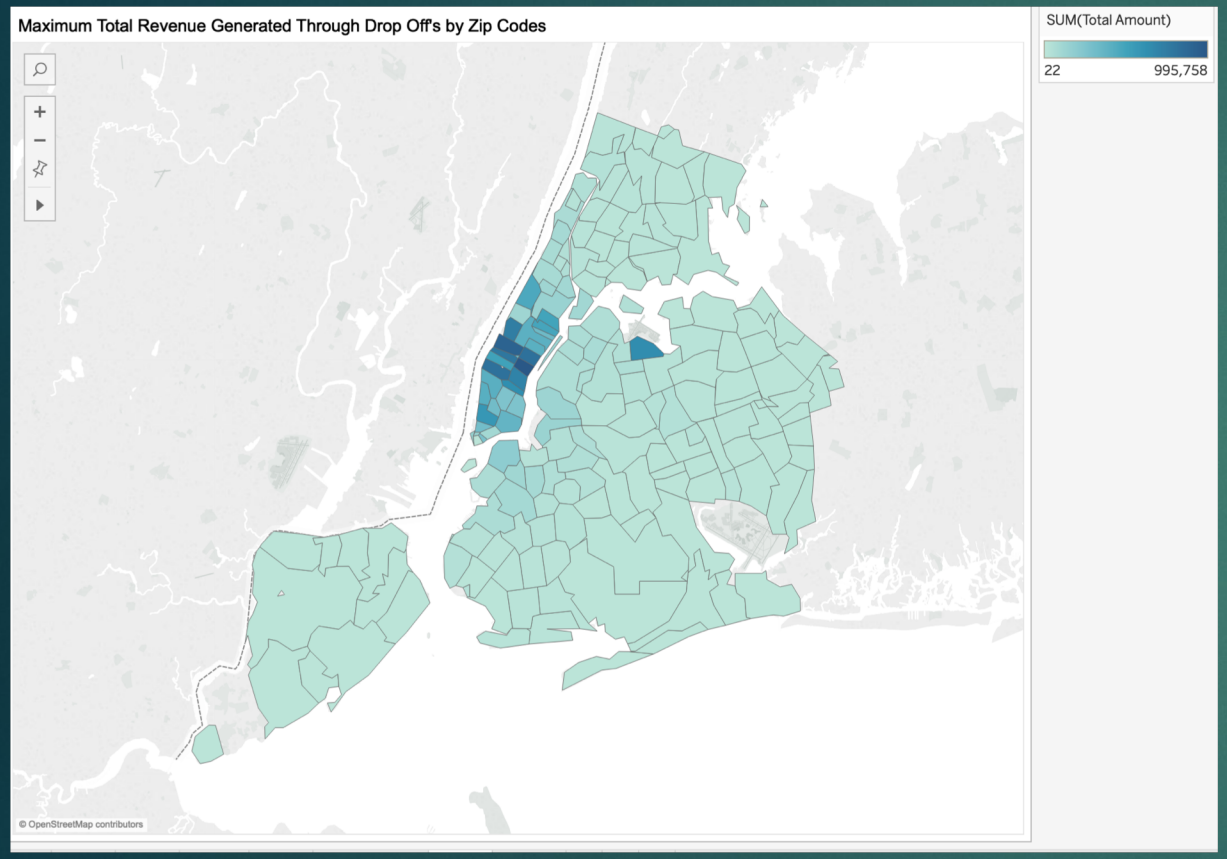
Day of the week wasn't found to have a significant impact on trip duration per trip or total amount per trip

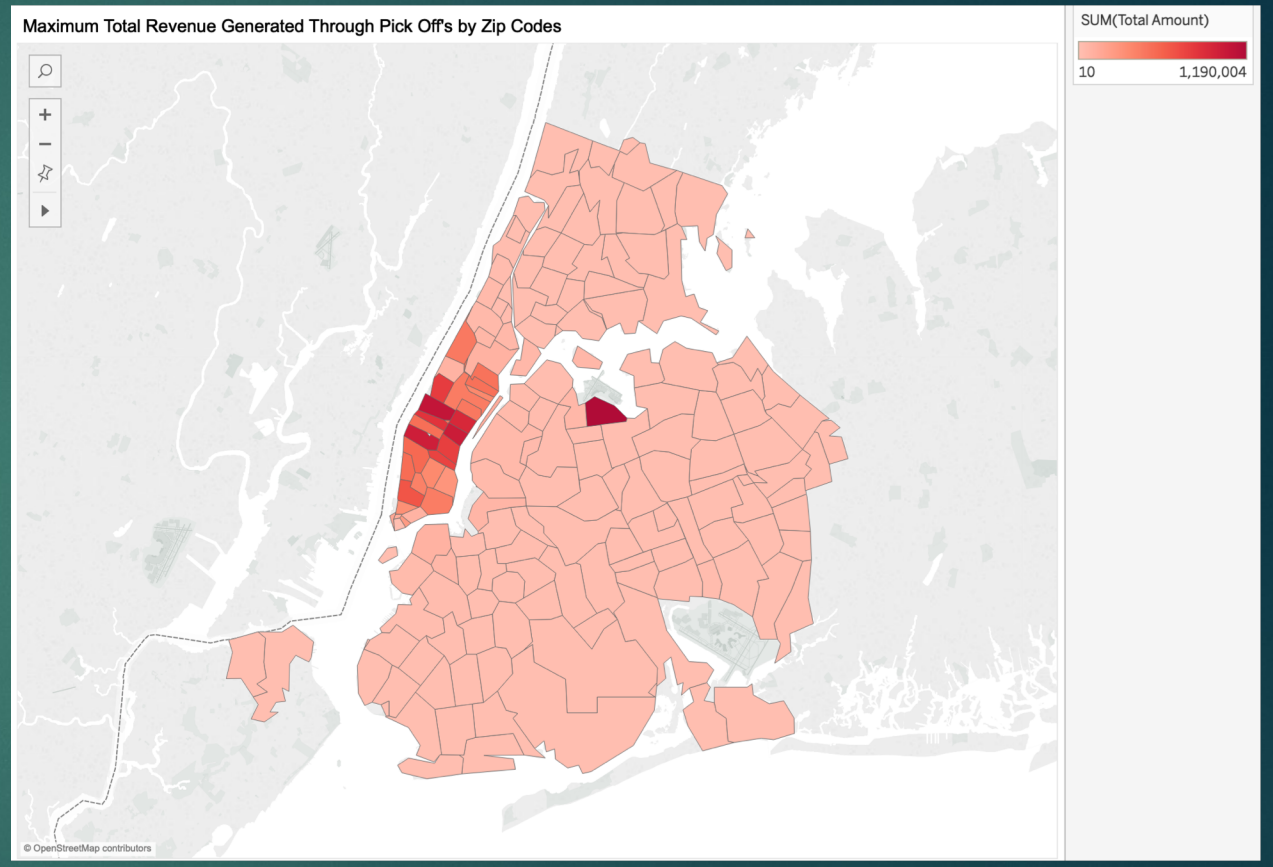




Factors Affecting Total Amount

D) Pick up and Drop off Zip Codes



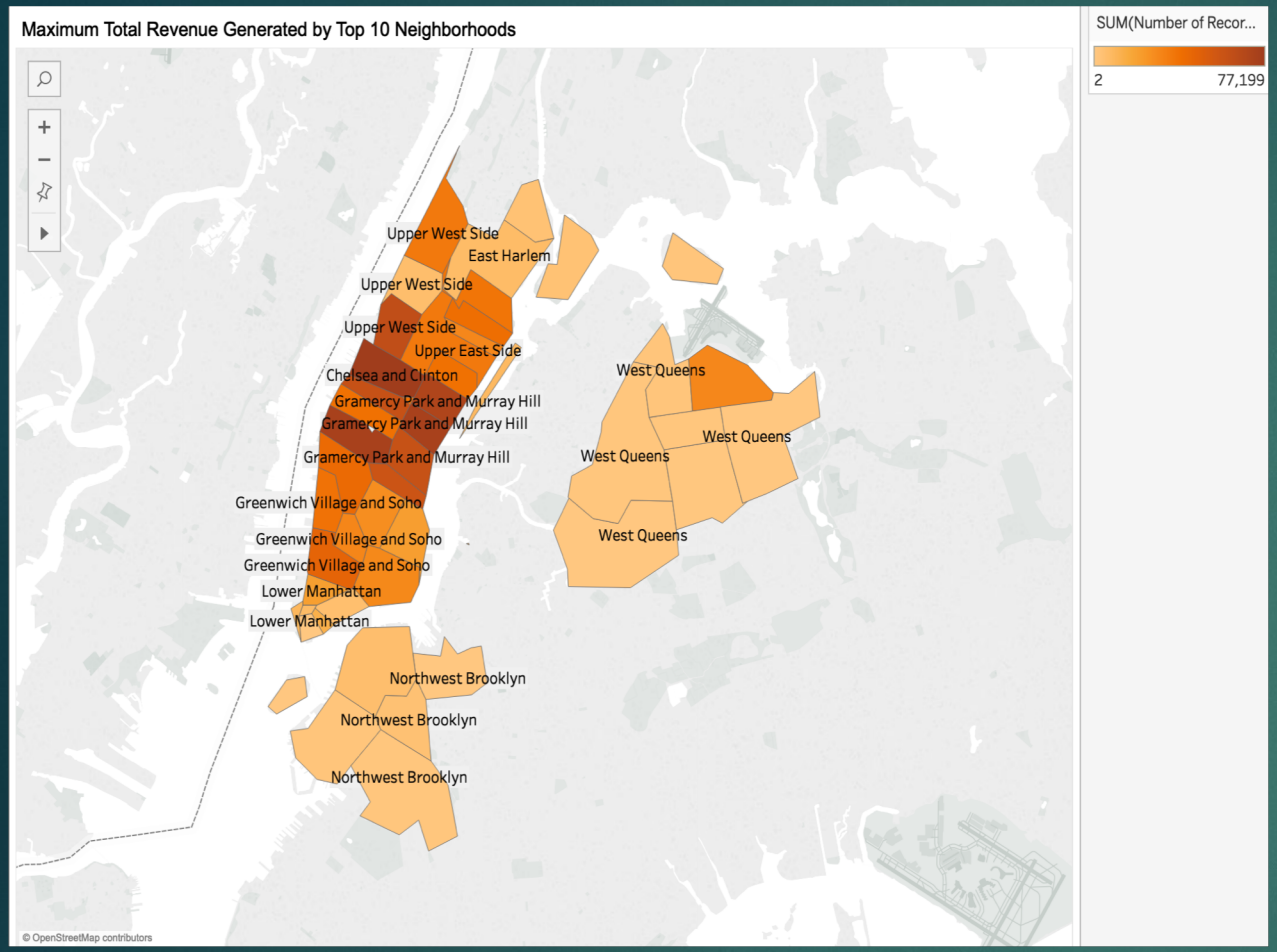


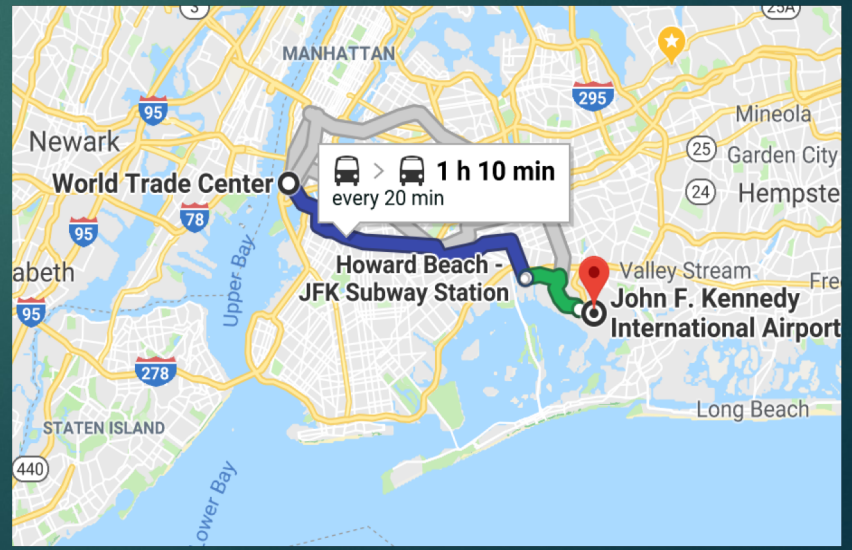
Caption

Map based on Longitude (generated) and Latitude (generated). Color shows sum of Number of Records. The marks are labeled by Pickup Neighborhood.

Factors Affecting Total Amount

E) Top Neighborhoods for Business



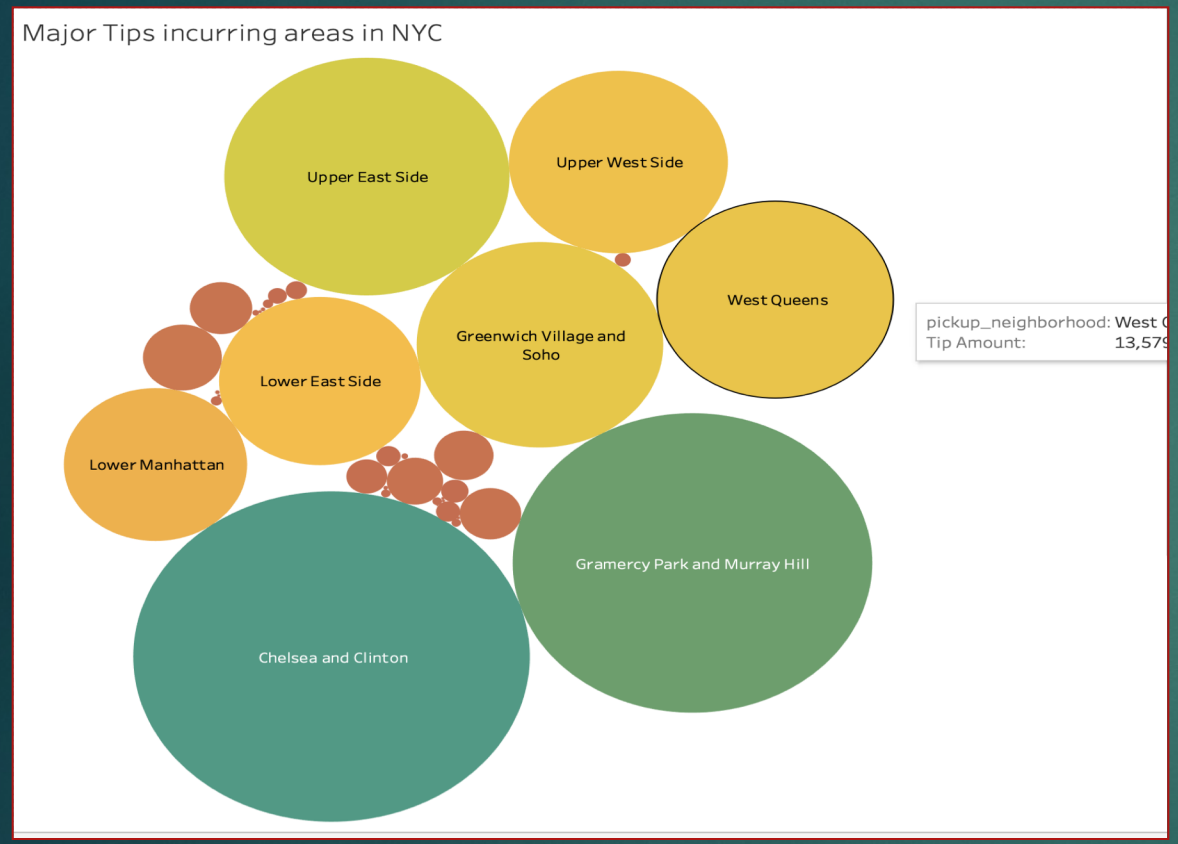


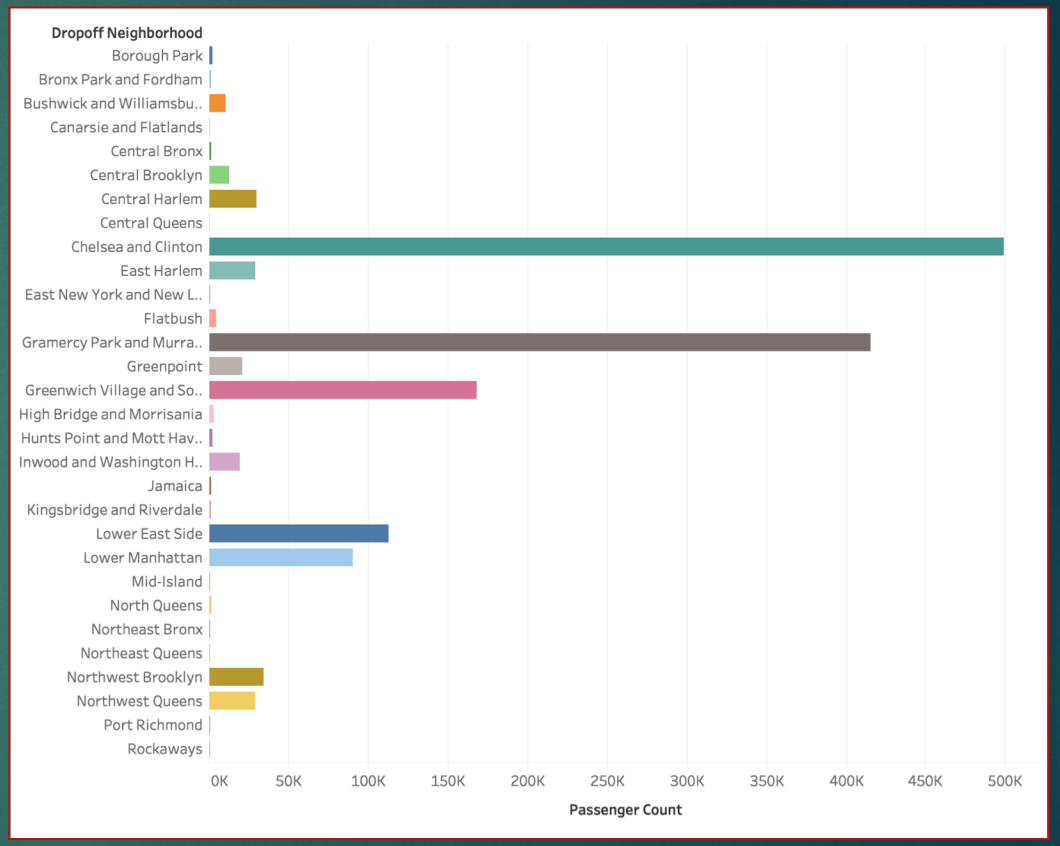
Caption

Map based on Longitude (generated) and Latitude (generated). Color shows sum of Number of Records. The marks are labeled by Pickup Neighborhood. Details are shown for Pickup Zip. The view is filtered on Pickup Neighborhood, which keeps 10 of 40 members.

Factors Affecting Total Amount

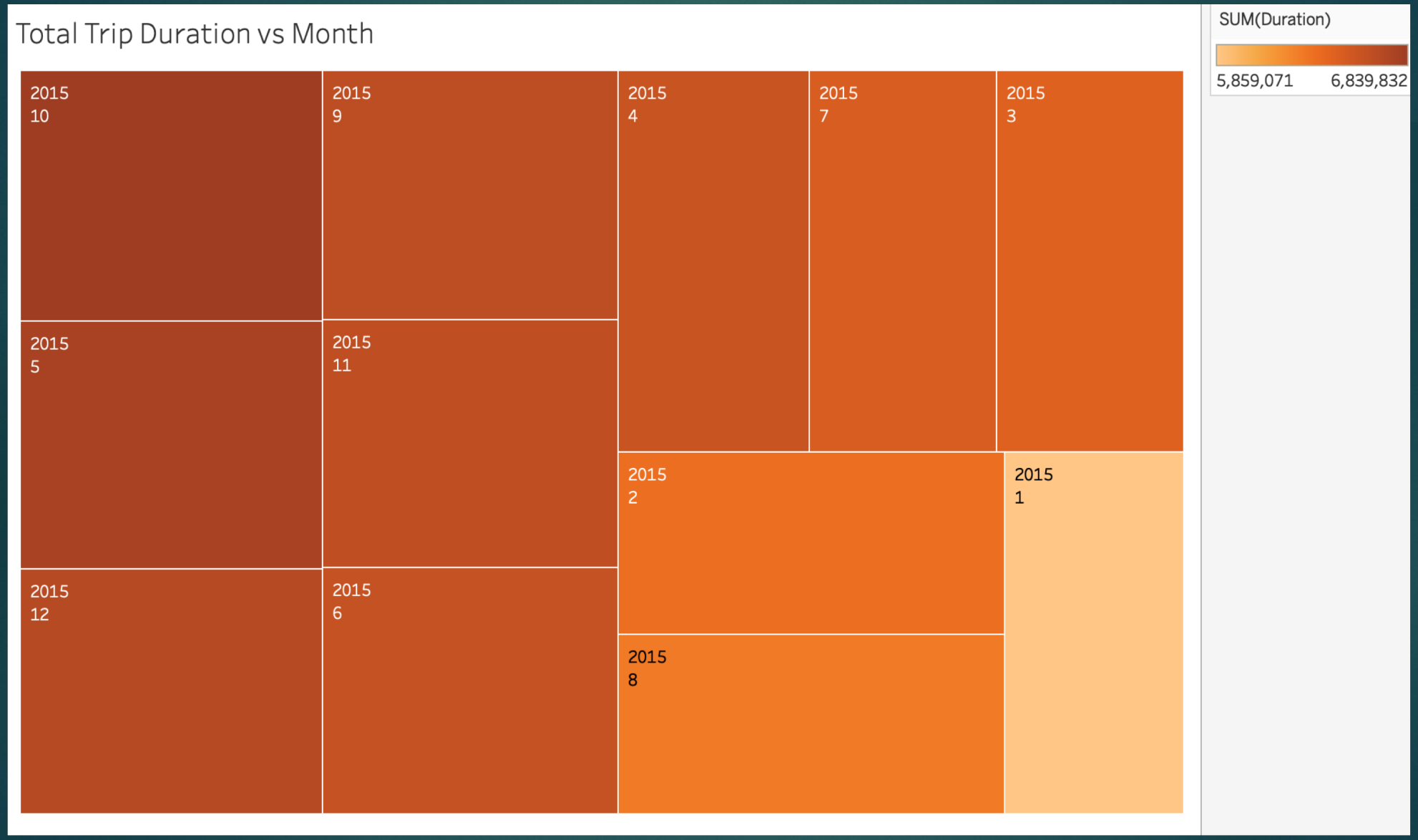
E) Top Neighborhoods for Business (Continued..)





Factors Affecting Duration and Total Amount

F) Seasonality



Feature Engineering

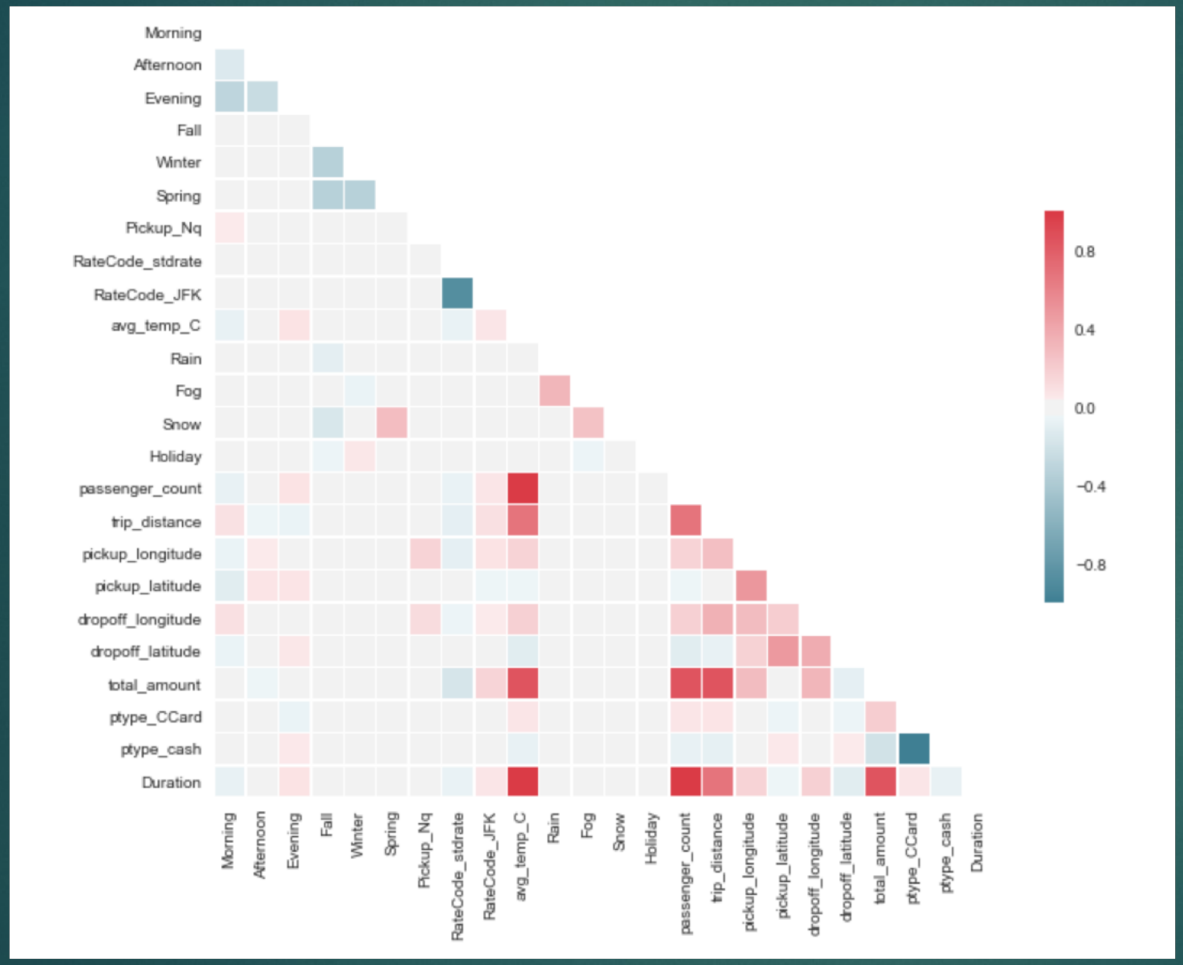
[122]:
```
features_sample_df = va.transform(dummy_df1)

from pyspark.ml.stat import Correlation
r1 = Correlation.corr(features_sample_df, 'features').head()
```

[123]:
```
r1[0].toArray()
x[18]
```

[123]:
```
array([ 0.02460368,  0.00459694, -0.00242413, -0.02030871, -0.01128363,
       -0.00815524,  0.01416242,  0.89698299,  0.00163576,  0.29461282,
       -0.02203972,  0.31885632, -0.09000388,  0.03929161, -0.06652559,
        0.64608975,  0.58869309,  0.01163316,  1.        ,  0.36270077,
        0.41295884, -0.01658767,  0.00465715,  0.02822905,  0.39567379,
        0.02765049, -0.34314617,  0.0011365 ,  0.32284087,  0.21704112,
       -0.41059884,  0.12499476,  0.19046307, -0.19082707, -0.00516868,
        0.00803957,         nan, -0.18506046,  0.15910981,  0.00852571,
        0.02194359,  0.10076259,         nan,  0.82628991])
```

[121]:
```
pipeline_estimator = Pipeline(stages=[
    feature.VectorAssembler(inputCols=[ 'trip_distance','pickup_longitude','dropoff_longitude',
                                       'tip_amount','tolls_amount','pickup_zip','dropoff_zip',
                                        'pickup_dayofweek', 'pickup_hour', 'pickup_month','pickup_Queens','pickup_Brooklyn',
                                                                                          ,'dropoff_Brooklyn',
                                        'dropoff_Manhattan','dropoff_Bronx','ptype_CCard','ptype_cash','ptype_nocharge',
                                        'ptype_dispute','ptype_voidtrip','RateCode_stdrate','RateCode_JFK',
                                        'RateCode_Newark','RateCode_NW','RateCode_Negotiated','RateCode_Group',
                                        'Duration'],
                            outputCol='features2'),
    regression.LinearRegression(featuresCol='features2', labelCol='total_amount')
])
pipe_model = pipeline_estimator.fit(features_sample_df)
pipe_model.transform(features_sample_df).show(3)
```

Correlation Matrix and PCA



`feature_engineering_duration`

	Feature	pc_best
5	total_amount	0.690945
3	dropoff_longitude	0.067075
1	pickup_longitude	0.001308
2	pickup_latitude	0.004390
4	dropoff_latitude	0.021739
0	trip_distance	0.719446

Novel Question

Focus: Resurgence of Yellow Cabs

Whether the ride will be profitable for the driver?

Does weather and holidays really affect the profitability of driver?

Why these questions:

- Uber/Lyft Drivers and smart apps
- Stringent Interviews and Background Checks
- Adverse effect on employment rate and crime rate

Statistical Models and Comparison

Outputs are continuous values and not categorical

Models Run: (Fitted and applied after Standardization)

- Linear Regression
- Random Forest
- Gradient Boosting
- XG Boosting

Evaluation metric and Results

For predicting both Duration and total amount we split the dataset into training, validation and testing in the 60%, 30% and 10% of the data respectively

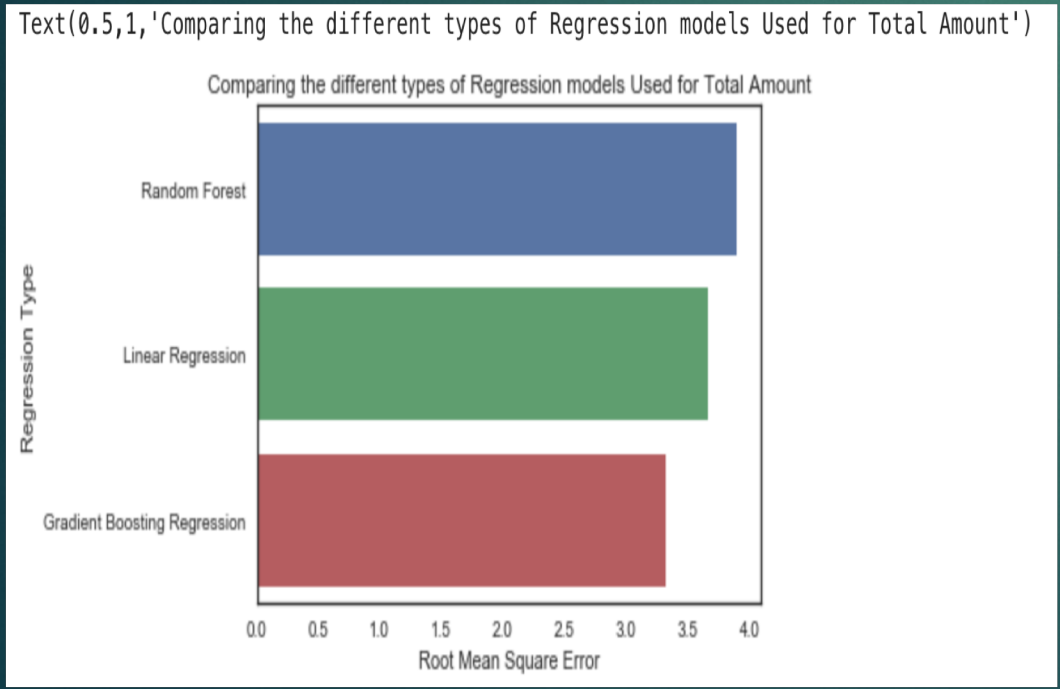
For predicting Duration and Total Amount, following models were implemented and the generalization performances were evaluated:

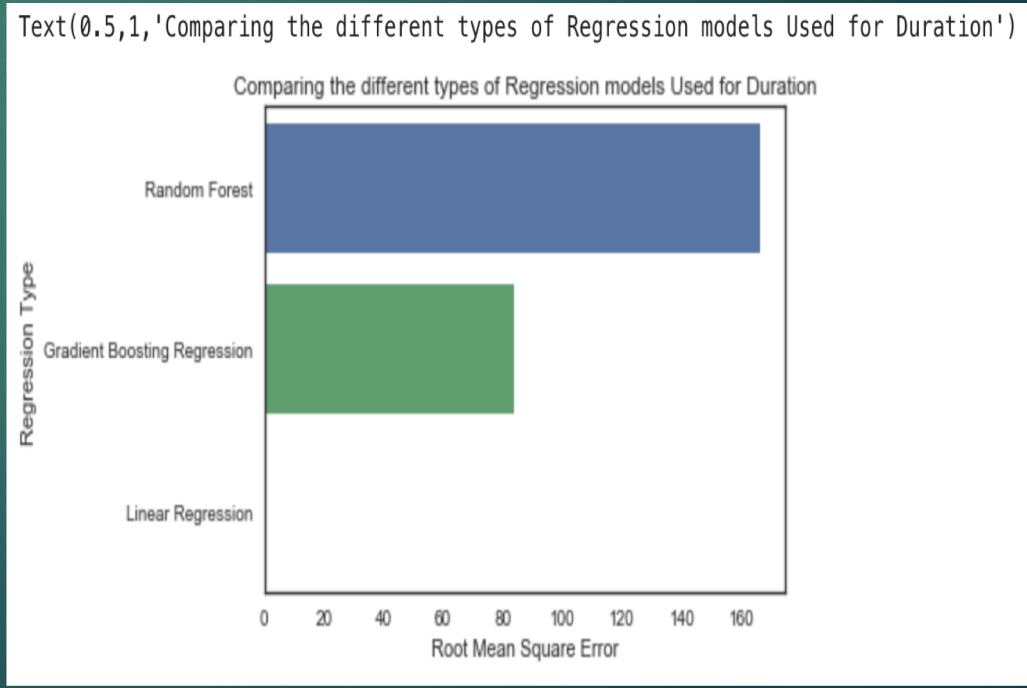
Linear regression : root mean squared error

Random Forest : root mean squared error

Gradient boosting regression : root mean squared error

XGBoost : accuracy





Description of validation and testing performance

Validate models on validation datasets

Identify the best model based on evaluation metric

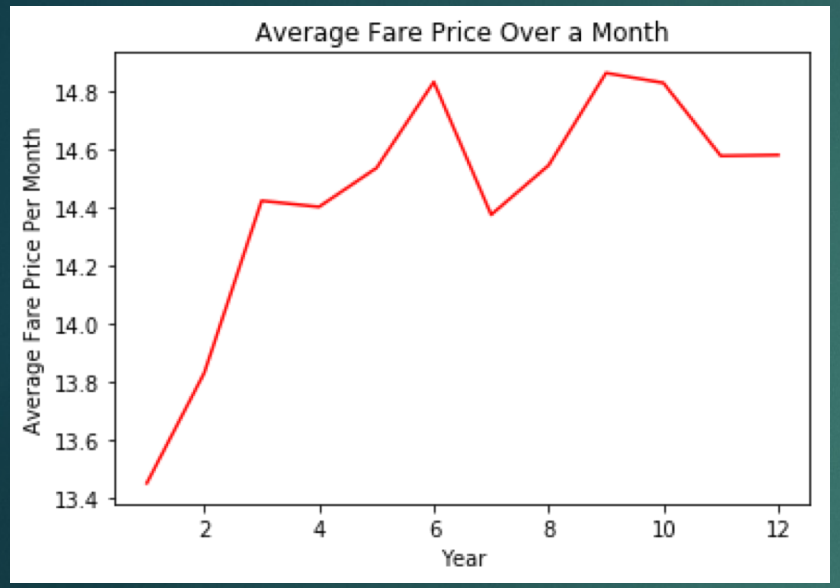
Test the best model on testing dataset

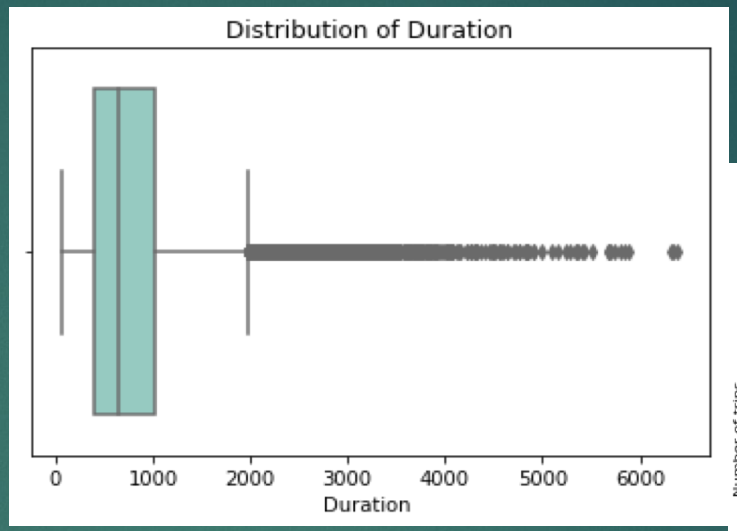
Estimate the generalization error of our models.

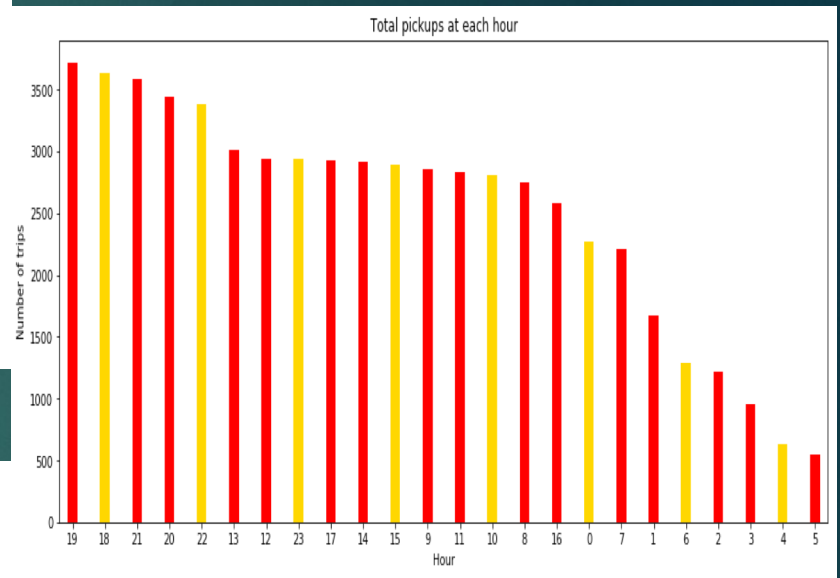
Perform inference on testing as well as full data

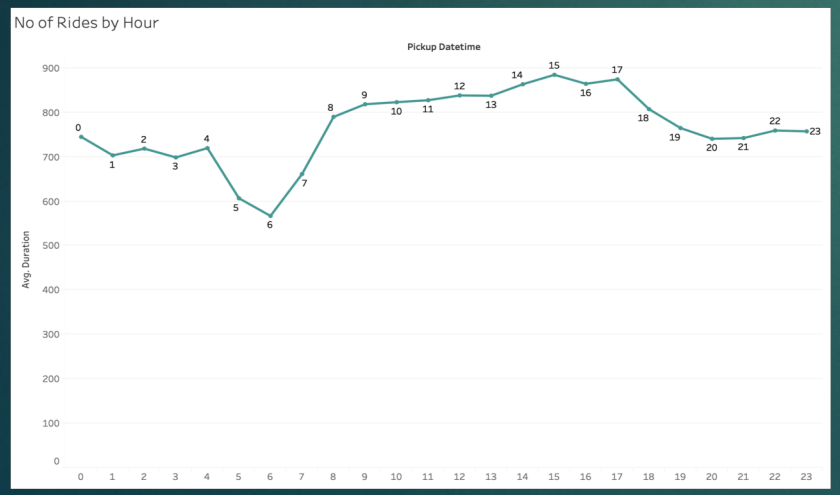
Aim is to predict the duration and prices to help drivers better plan their day, finances and earnings on their trips

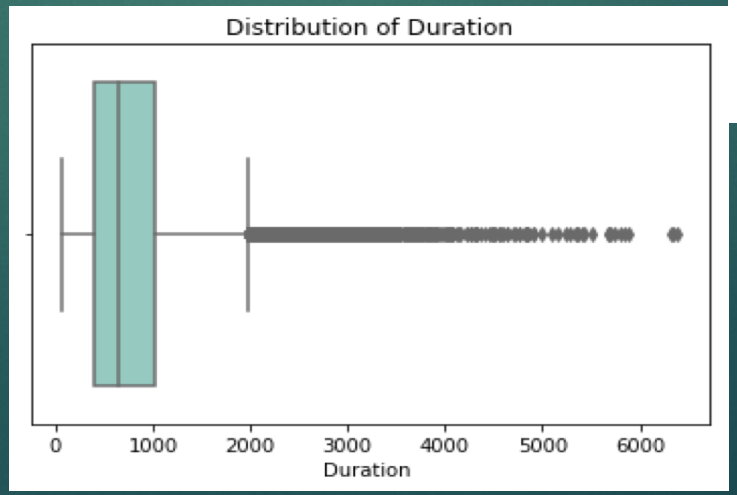
Visual Inferences











Results-Inference

Exploration Findings:

A bright sunny day is the best day for yellow cab drivers. In an event when there is no Snow, & no Fog but rain then, the total amount incurred is more which makes rain a significant factor

On Thursdays, the average total amount incurred is the most

Maximum revenue earned is in Queens possibly because of drop-offs at LaGuardia Airport

Similarly, Queens borough having LaGuardia Airport had highest revenue because of the number of pickups and drop offs rides at the airport

The average trip duration was the highest during Tuesdays and Wednesday

Most number of trips occurred during the month of October

Significance of Correlation:

The correlation matrix between each variable to every other variable identifies the highly correlated features

This helps in identifying the significant features while efficiently reducing the number of features required to run the model

Significance of PCA:

Principal Component Analysis is a significant method to identify features through dimensionality reduction

The features chosen by PCA can be cross-validated through logical understanding of correlation matrix

Best Model:

The best model is chosen by comparing the generalization performance values of all the implemented models

The gradient boosting model worked the best due to low root mean squared error value

Conclusion and future work

Overall, our models for predicting taxi pickups' total amount and duration in New York city performed well
The gradient boosting regression model closely followed by the random forest model performed the best
This was likely due to their unique ability to capture complex feature dependencies
The Rmse for gradient boosting regression model was achieved to be ~$1.71178 for amount and ~87.3163 sec for total duration. The average duration of a trip is about 15-18 minutes. The average total amount is between $10-$19. The average fare price is highest for the month of October. Longest rides occur at 3:00 PM in the evenings and at 6:00 AM in the morning, the distance of the rides is the shortest

Our results and error analysis for the most part supported our intuitions for the usefulness of features with the exception of holiday feature which was not found important for model performance
Our model can be used by city planners and yellow taxi drivers in determining where to position yellow taxi cabs and understanding patterns in ridership

Future Work:

- k-means clustering, we aim to achieve better results in the future
- Provide drivers with a smart interface with predicted duration and prices to help drivers better plan their day, finances and earnings on their trips

Problems found so far and plans to solve them

Problems:

- Data cleaning in short period of time
- Data processing because of storage restrictions
- Data integration
- Software installations for handling big data

Solutions:

- All the NA were removed
- Three datasets were merged : Taxi data, Weather data, and Holiday data
- Installed Github for shared source code maintenance
- Leveraged Dropbox resources for big data file storage and sharing

Thank You

QUESTIONS?