**Bike Search and Rentals Prediction**

**INFO6105- DATA SCIENCE ENGINEERING METHODS AND**

**TOOLS**

**Date: 04.15.2020**

**Avani Iddalgi Devina Jaiswal Gaurav Lohani Rachit Agrawal**

**Contents**

Topic…………………………………………………………………………………………...3

Professor……………………………………………………………………………………….3

Group Members……………………………………………………………………………......3

Introduction……………………………………………………………………………………4

Objectives……………………………………………………………………………………...4

Methodology…………………………………………………………………………………..5

Specification of the dataset…………………………………………………………………….5

Working………………………………………………………………………………………..8

Data Cleaning and

Preprocessing………………………………………………………….8

Missing Value

Analysis……………………………………………………………………9

Outlier

Analysis…………………………………………………………………………..10

Exploratory

Analysis……………………………………………………………………..10

Algorithms Used……………………………………………………………………………...14

Comparing the Best Two Models…………………………………………………………….17

Feature Importance and Correlation………………………………………………………….18

Result and Analysis…………………………………………………………………………..22

Conclusion……………………………………………………………………………………23

References……………………………………………………………………………………24

**Topic:**

 Bike Search and Rentals Prediction

**Professor:**

 Dr. Liu Handan

**Group Members:**

 Avani Iddalgi (001087272)

 Devina Jaiswal (001306506)

 Gaurav Lohani (001058907)

 Rachit Agrawal (001306902)

**Introduction:**

Bike-sharing is a short-distance bike rental service that allows customers to pick up a bike at one station, use it for transportation as needed, and return it to another station. This is automated via a network of kiosk locations throughout the city. The bike-share system has a relatively young history. This transportation system came into existence in 1965 whereas bicycles, which were invented in the early

1800s and cars, have been around since the late 1800s. Around the world, there are over 500 bike- sharing systems. Heavy street traffic in busy cities and a desire for an environmentally friendly form of transportation makes biking an attractive alternative to traveling by car. Customers get the convenience of a bike without the headaches of bike ownership including parking, theft, routine maintenance, and storage through this transportation model. Demand for bike-sharing services has exploded in recent times; as of June 2014, public bike-sharing systems can be found in 712 cities, with approximately

806,200 bicycles in circulation at a total of 37,500 stations [1].

Researchers are attracted to the data generated by these bike-sharing systems because they function as a sensor network. Using the data generated by these systems, we can study the mobility of a city. Also, demand prediction is crucial to ensure that there are neither too many bikes (and hence not enough bikes at some other station that needs them) nor too few bikes thus resulting in a loss. With the help of bike- sharing system data, we can understand the business more accurately and forecast sales. We can answer the following questions after the analysis, that can be used to increase sales.

 How weathers impact bike trips?

 How bike trip patterns vary by time of day and the day of the week?

**Objectives**:

In this project, we put ourselves in the mindset of an administrator of the bike-sharing system and better understand the system’s user to provide better service. In particular, we are:

1. Finding the duration of bike trips according to the area so that we can understand the pattern of customers to increase sales.

2. Predicting the total number of customers and subscribers. Also finding the trip duration of

both types of customers.

3. Predicting the number of stations and docks available in the bay area.

4. Discovering how weather conditions impact the bike rental service

5. Finding particular properties of different stations. For example, trips taken to and fro from various stations.

**Methodology**

Machine learning is the tool of choice when working with problems that are either too complex for a human to comprehend or are infeasible to program by hand, and it seems an appropriate tool to apply to the bike-sharing system data so that it can help in the growth of the business.

The first job as a data scientist is to explore the dataset we have and determine the baseline of our predictions. Below are the stages of our analysis.

1. Data pre-processing

2. Missing Value Analysis

3. Outlier Analysis

4. Exploratory Analysis

5. Model Analysis

6. Feature importance and Correlation Analysis

We are intent to use the following supervised ML algorithms to train the model for predictive analysis.

1. Linear Regression Model

2. Decision Tree Model

3. Random Forest Regression

4. AdaBoost Classifier

5. Gradient Boosting Classifier

**Specification of the Dataset**

The data used for this project is taken from Kaggle and the span of this dataset is from June 14, 2016, to November 11, 2019. This dataset is a transformed version of the Bay-area bike-sharing system data. The dataset of the bike-sharing system is in correspondence with environmental and seasonal settings. For instance, weather conditions, precipitation, day of the week, season, an hour of the day, etc. can affect the rental behaviours.

**Dataset Key Specifications:**

1. The SF Bay Area Bike Share data consists of four excel files.

• **station.csv**- This file consists of station data where users can pick up and return the bikes. It has 7 columns that mention unique id, station name, latitude, longitude, dock count, etc. of the station.

• **status.csv-** There are 4 columns in this file that contains the data about the number of bikes

and docks available for the given station and minute.

• **trip.csv-** Data about individual bike trips are present in this file. It consists of 11 columns that show bike id, duration, starting station and date, ending station and date, zip code, etc.

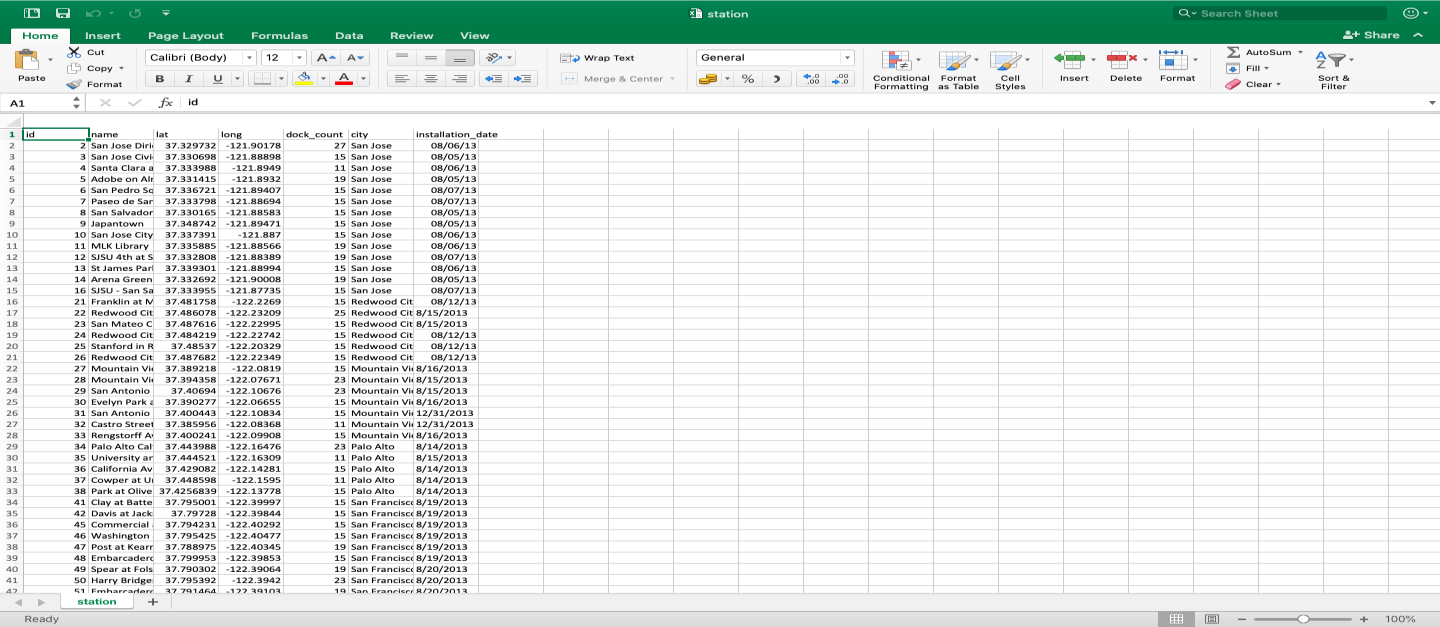
• **weather.csv-** In this file, each row represents the weather for a specific day and zip code in the bay area. It has 24 columns in total showing different aspects of a weather condition such as minimum temperature, maximum temperature, humidity, wind speed, precipitation, etc. Temperatures are in Fahrenheit.

2. Type of Data- Text

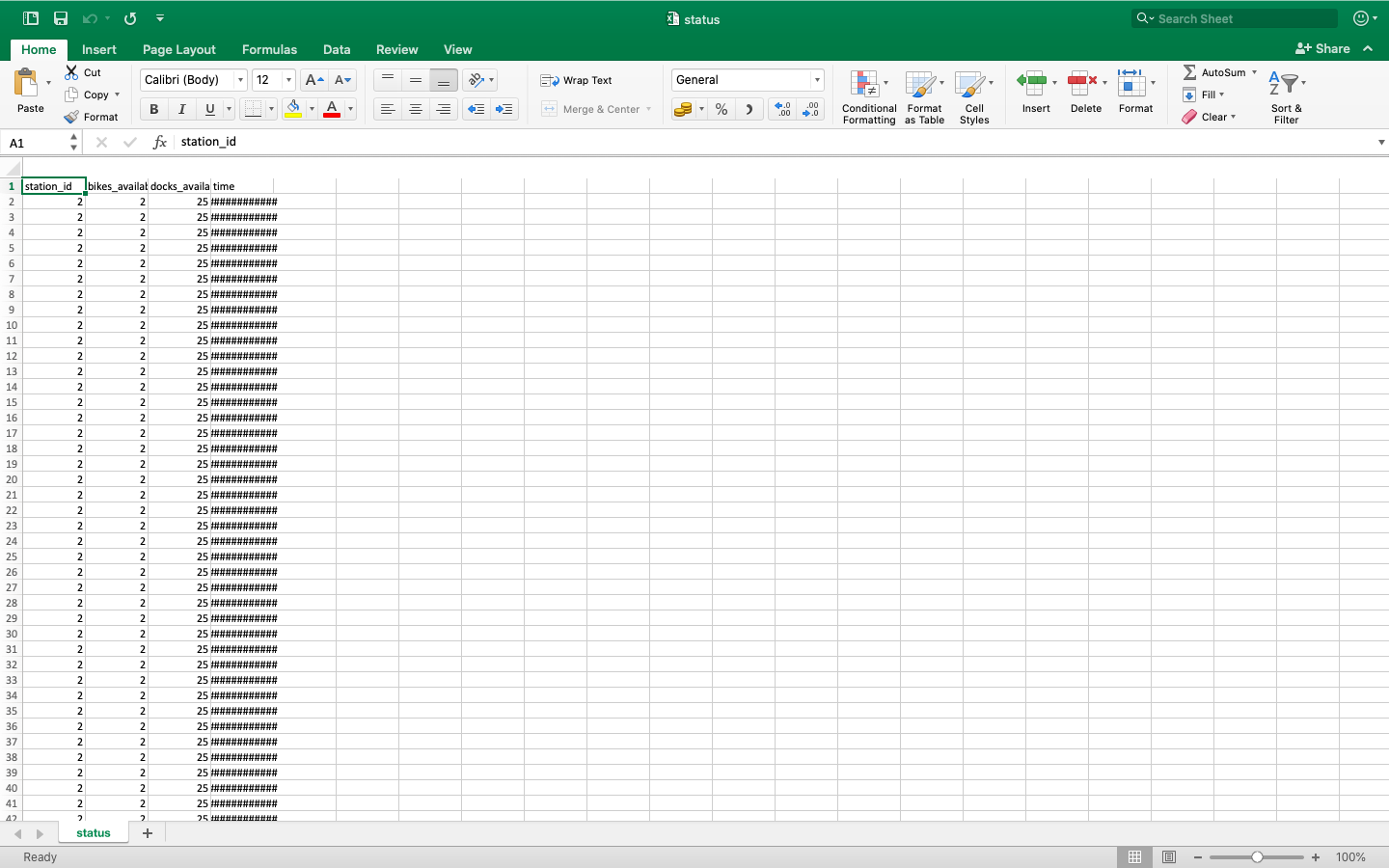
3. Dataset Size- 4 GB

4. Data Sources

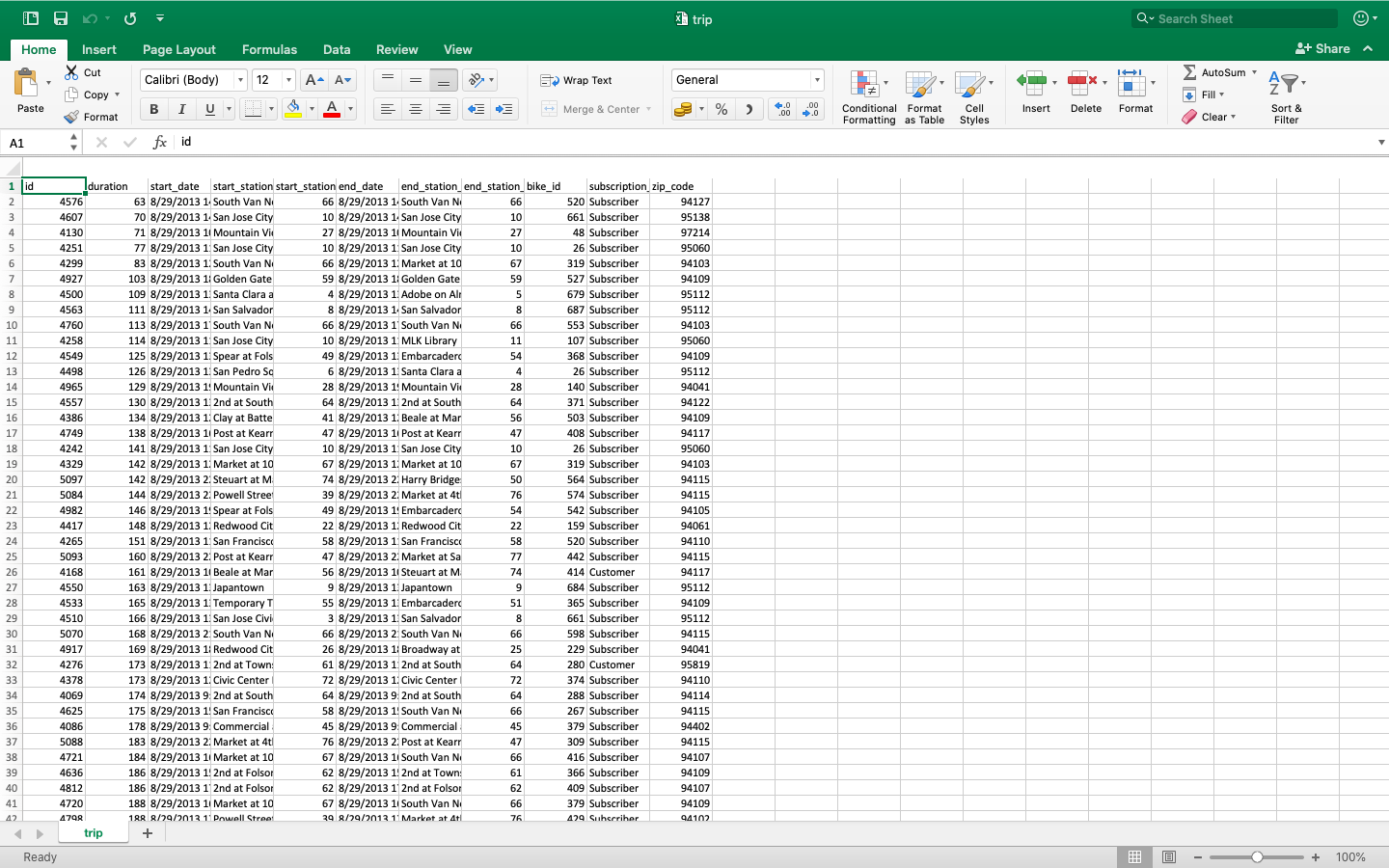
Dataset: <https://www.kaggle.com/benhamner/sf-bay-area-bike-share#weather.csv>



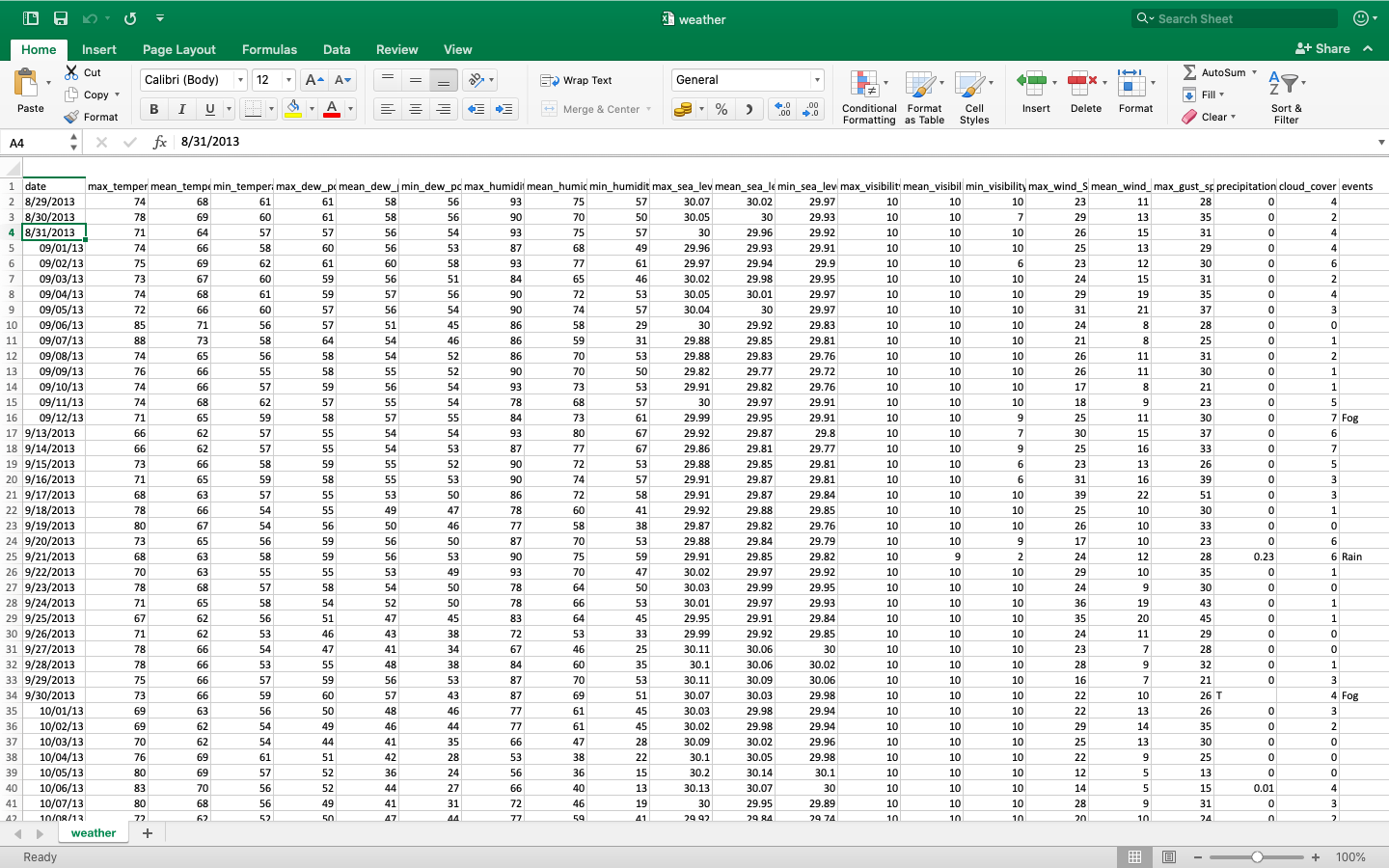
**station.csv file**



**status.csv file**



**trip.csv file**



**weather.csv file**

**Working**

**Data Cleaning and Pre-processing**

To start with the project, we transformed the raw dataset into an understandable format to improve data efficiency. It consisted of the following stages.

• Removing columns containing null values.

• Cleaning data that does not have uniform values.

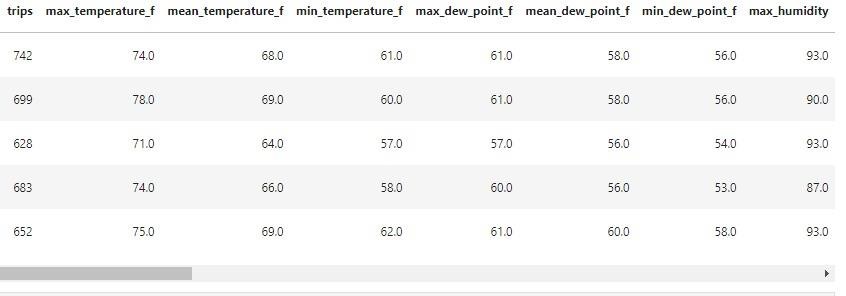
• Changing the duration column value from seconds to minutes to get a better understanding.

• Converted dates into DateTime format so that it can be manipulated more easily.

• Removing features that are not required. For example, events and zip code columns were removed from the weather file.

• Changing NA values from the dataset to the median value.

• Finding correlation for the entire dataset and elimination of features with the same meaning to avoid duplicity

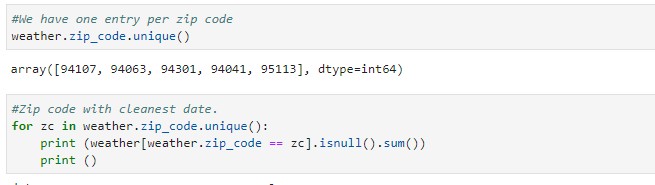


**Final Trained Data**

**Missing Value Analysis**

After analysing all the datasets, we found out that multiple values for zip code and events were null.

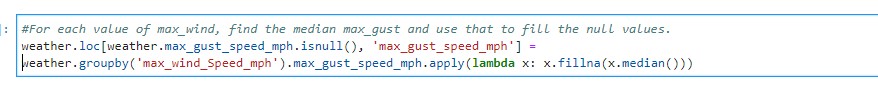
1. Zip Code - We have modelled the data according to the unique zip codes of the bay area. Thus eliminating the null values.



2. Events - The filtered data still had 574 null records in the Events column. To overcome this issue, we have found out the unique values of events and modeled the data according to it.



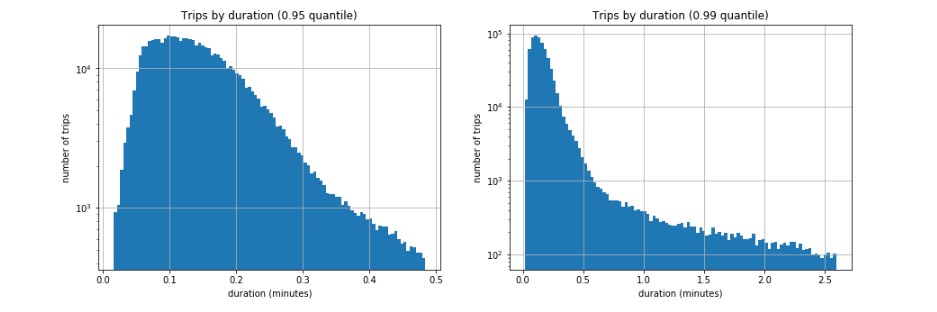
3. Max\_gust\_speed\_mph - Upon further filtering of the data, there were still 12 null records for Max\_gust\_speed\_mph column. We have to find the median of max\_gust\_speed for each value of max\_wind and filled the null values.



**Outlier Analysis**

An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as an outlier analysis.

In this project we are trying to get rid of 4% outlier data.



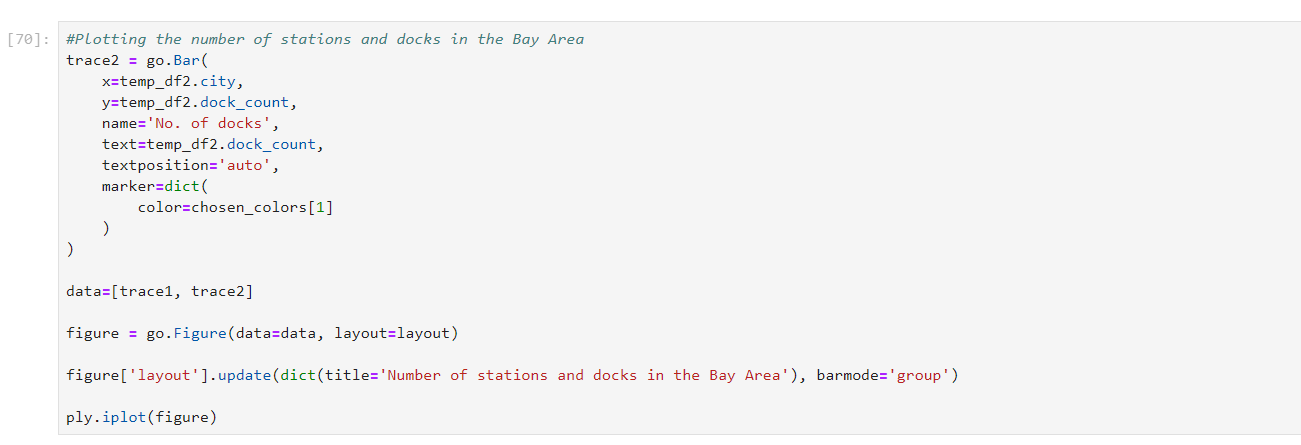
Above figure shows the number of trips by duration when 95% and 99% data is used.

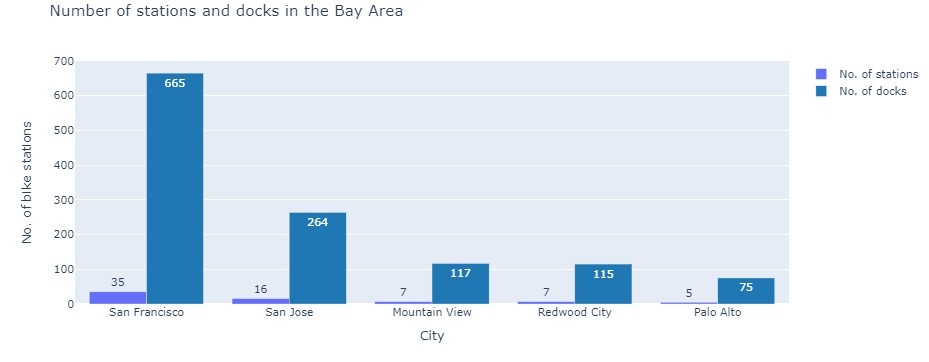
**Exploratory Analysis**

Exploratory Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations [2].

In this project, we have performed exploratory analysis on the station and trip dataset to analyse various aspects of bike-sharing system.

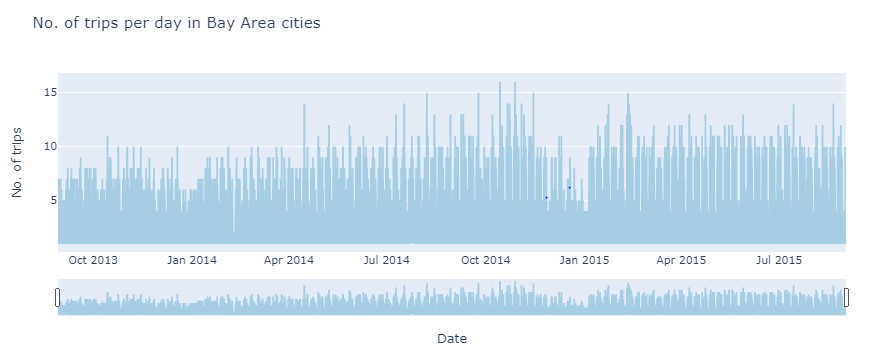
1. ***Plot of the number of stations and docks in the Bay Area***





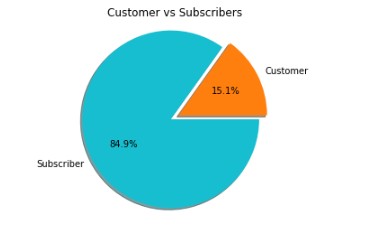
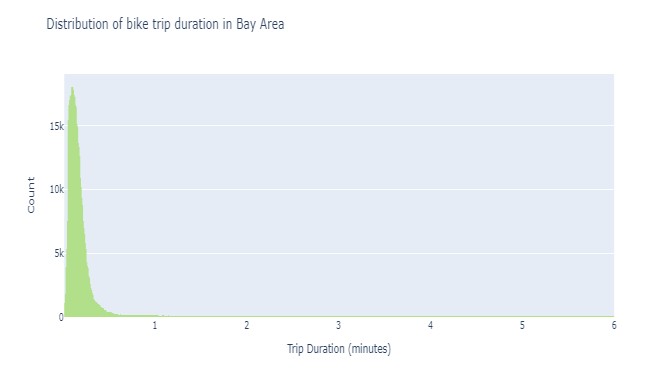
1. ***Plot showing the number of trips taken per day by customers in the Bay Area cities***

****



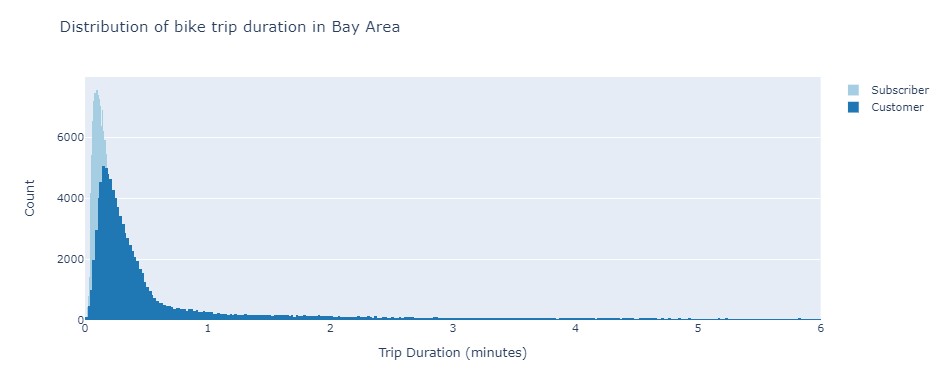
1. ***Plot of bike trip duration 4. Pie Chart of Customer Vs***

***Subscribers***

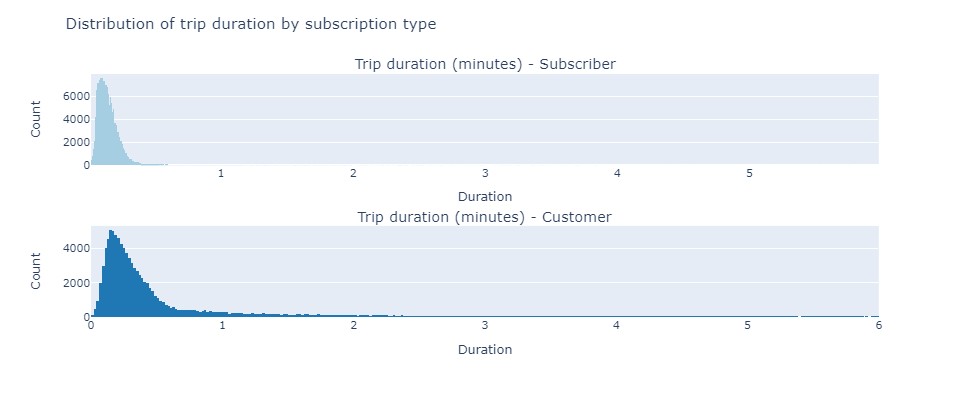


1. ***Distribution of bike trip duration in Bay Area***

******

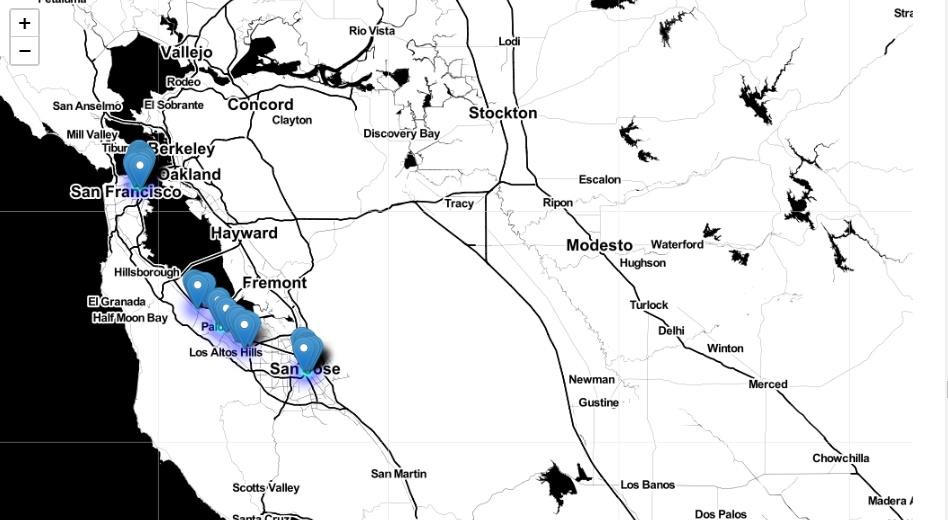


1. ***Plot of trip duration according to the subscription type (i.e. Subscribers Vs Customers)***



1. ***Heatmap of bike stations located in the Bay Area***

****

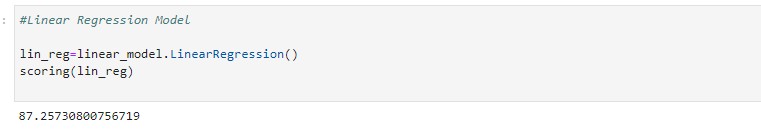
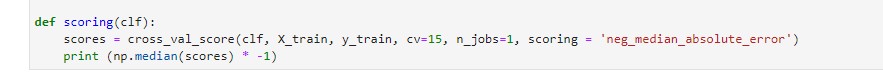


**Algorithms Used**

**Linear Regression**

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable [3].

In this project, the motive to use Linear Regression model is to predict the number of rented bikes in a given time frame and try to predict the demand of bikes in the market.

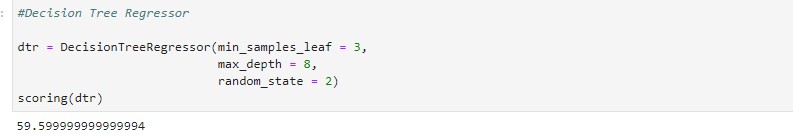
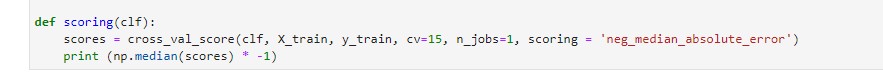


The accuracy of Linear regression model is 87.25%.

**Decision Tree Model**

The decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into a smaller subset while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes [4].

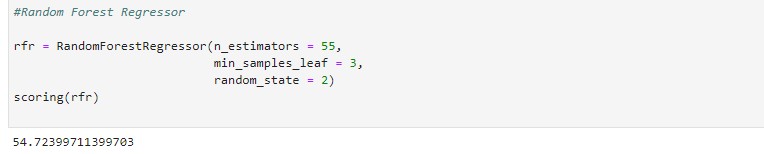
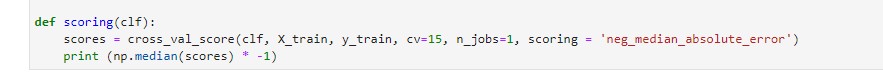
We will be using the decision Tree Model in our project to improve our predictions. The linear regression is good for datasets with lots of continuous data, but we also have categorical data set for which we will be using the decision tree model on our data. This will help us to reduce the errors significantly and we will also use the forest of decision trees to reduce overfitting.



The accuracy of Decision Tree Regression model is 59.59%

**Random Forest Regression**

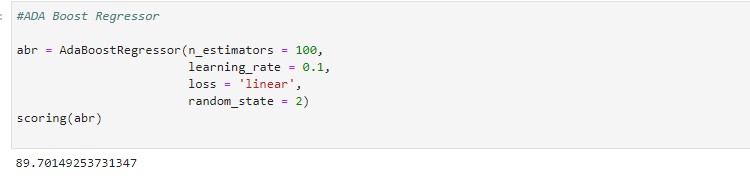
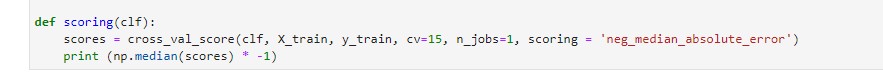
Random Forest [5] is a meta-algorithm that combines a large number of decision-tree models, each individually built on bootstrapped samples of the data. This process of sampling the data and combining the individual decision-trees is called bagging, and is able to reduce the variance of the predictions without increasing the bias. The final predictions are formed by taking the mean of the individual decision-tree prediction.



The accuracy of Random Forest Regression model is 54.72%.

**AdaBoost Classifier**

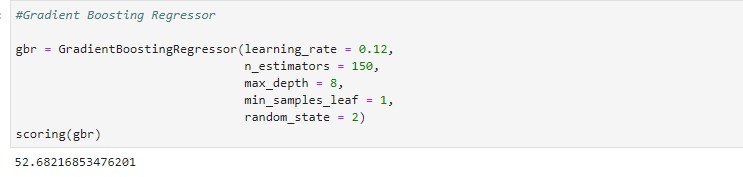
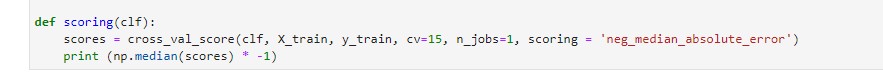
AdaBoost Model is used to boost the performance of decision trees on binary classification problems. This will help to boost the performance of any algorithm. This model is commonly used in decision trees with one level. In bike rental prediction Ada boost will be useful with stationary data.



The accuracy of AdaBoost Regression is model is 89.70%.

**Gradient Boosting Classifier**

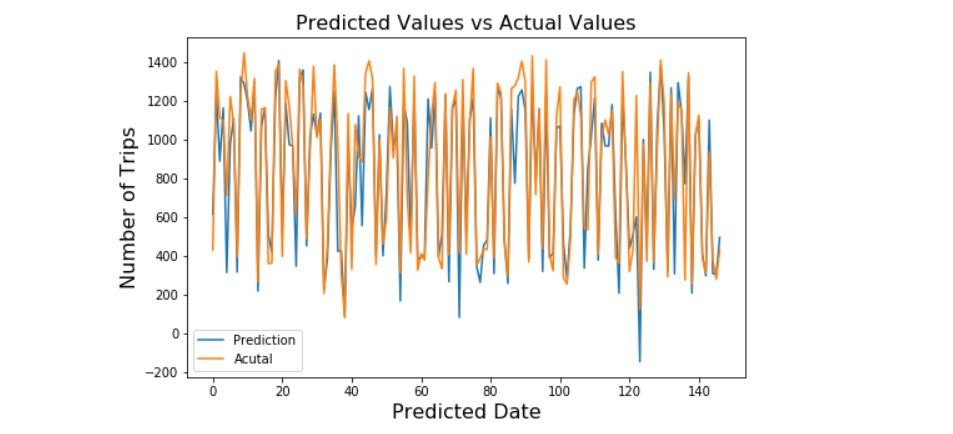
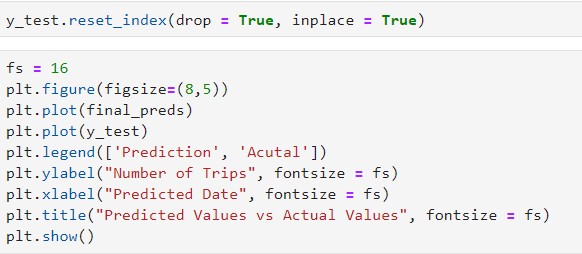
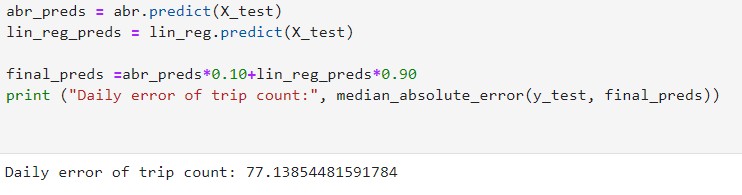
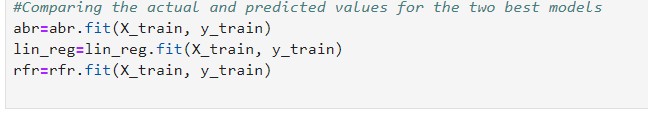
Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set.



The accuracy of Gradient Boosting Classifier is 52.68%.

**Comparing the best two models**

In this project, the two best models are Linear Regression and AdaBoost Classifier. Below is the comparison of these two models.

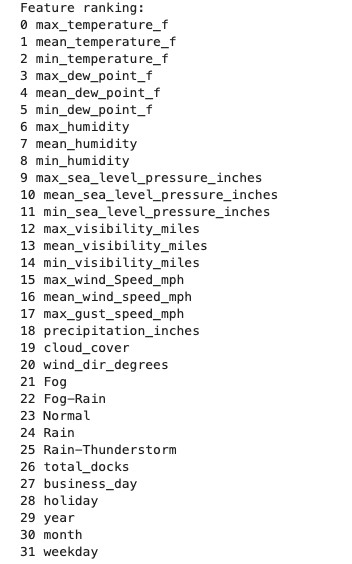


The above plot shows us that the predicted and actual values are almost similar to each other which tends to justify the high efficiency of the two models- AdaBoost Classifier and Linear Regression.

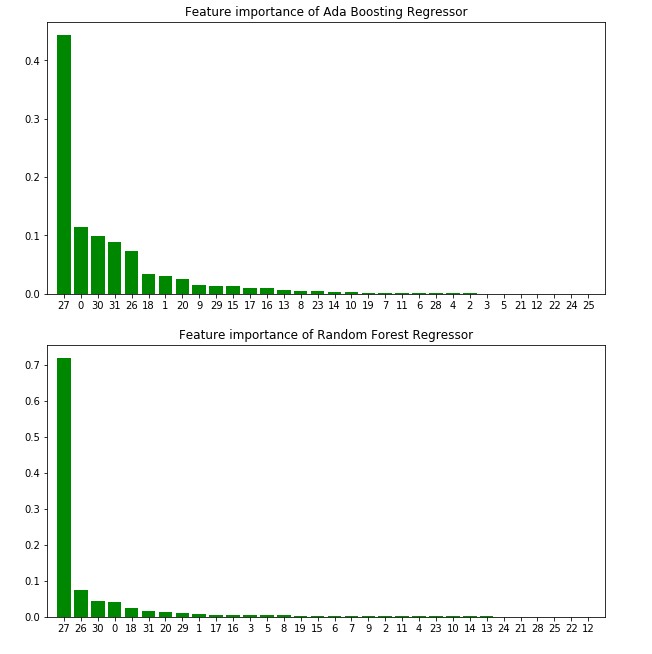
**Feature Importance and Correlation**

Feature importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model.

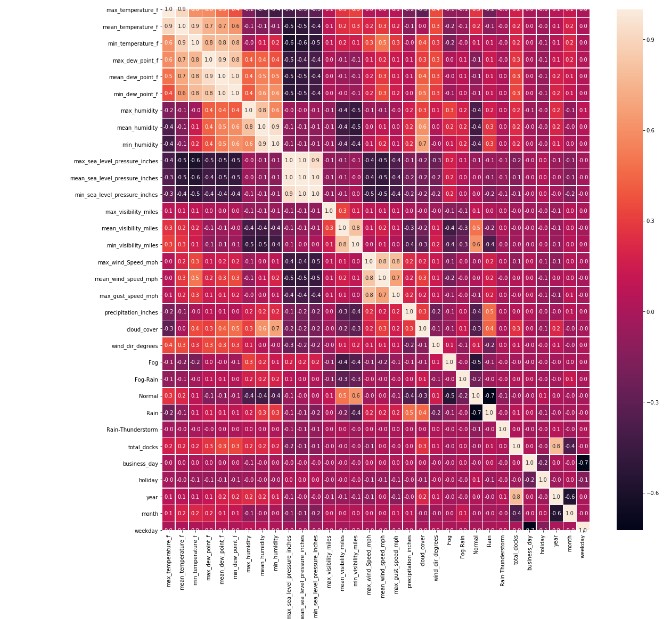
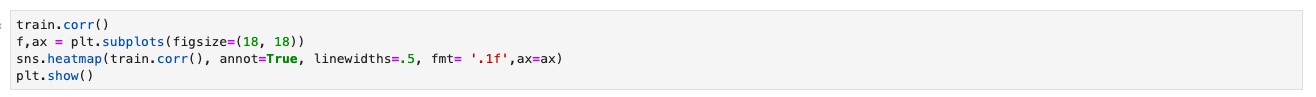
In this project, we found out feature importance in Ada Boosting Regressor and Random Forest Regressor.



20



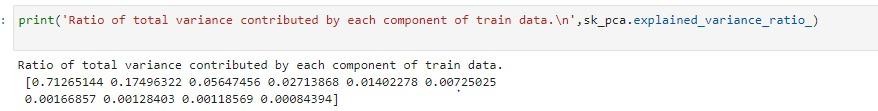
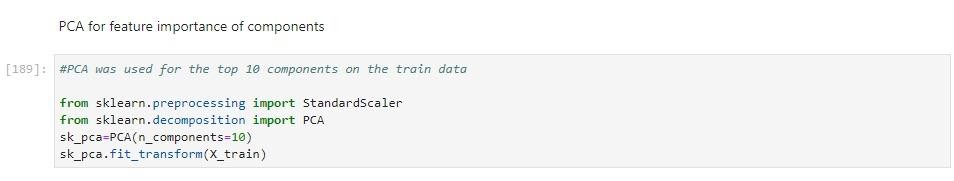
A correlation matrix is a table showing correlation coefficients between sets of variables. Each random variable (Xi) in the table is correlated with each of the other values in the table (Xj). This allows you to see which pairs have the highest correlation.



**This plot shows a trend of correlation between all the features of the dataset**

**Principal Component Analysis**

PCA was performed to identify top 10 components explaining the variance in the data.71% of the variance can be explained by only one component that is cloud\_cover.



**Result and Analysis**

**Bike-Sharing System Accuracy:**

Linear Regression- 0.8725

Decision Tree Regression- 0.5959

Random Forest Regression- 0.5472

**AdaBoost Classifier- 0.89701**

**Gradient Boosting Classifier- 0.5268**

Accuracy for bike rental prediction is 89.79% using AdaBoost Classifier. It is the best model when given the bike-sharing rental dataset for predictions.

**Conclusion**

Using different combinations of models and features of data science, we found out how the bike-sharing system dataset will impact the business model. We discovered various factors that play an important role when predicting sales from the given dataset to increase our revenues such as weather conditions, number of bikes available, location of a station, trends of customers, etc.

In this project, we attempt to find out the number of trips taken per day by the customers in the Bay Area cities which will help in taking further sound business decisions to enhance the profits of the bike-sharing system.

In addition, AdaBoost Classifier in particular and Linear Regression model were able to best capture the relationships within data, which led to the best performance on the leaderboard.Also, we have found out feature importance of the for these two models.

**References**

[1] S. A. S. et al., “Public bike sharing in north America during a period of rapid expansion: Understanding business models, industry trends and user impacts”, in Mineta Transportation Institute (MTI), 2015, p. pp. 5.

[2] “What is Exploratory Data Analysis?”, May 23, 2018. Accessed on: April 03, 2020. [Online]. Available: <https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>

[3] “Linear Regression”. Accessed on: April 09, 2020. [Online]. Available:

<http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>

[4] “Decision Tree Regression”. Accessed on: April 12, 2020. [Online]. Available:

<http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>

[5] Andy Liaw and Matthew Wiener. Classification and regression by random forest. R News,

2(3):18-22, 2002.

[6] “A Model to Predict Number of Daily Trips”, Kaggle, Feb. 16, 2017. Accessed on: March 10,

2020. [Online]. Available: [https://www.kaggle.com/currie32/a-model-to-predict-number-of-daily-](https://www.kaggle.com/currie32/a-model-to-predict-number-of-daily-trips) [trips](https://www.kaggle.com/currie32/a-model-to-predict-number-of-daily-trips)