**FAST National University of Computer and Emerging Sciences**



**Project Topic:** Uber Fare Price Prediction (Cleaning, Visualization, Algorithms (Linear Regression, Improving accuracy (Adaboost, KFold Approach, Holdout Approach)))

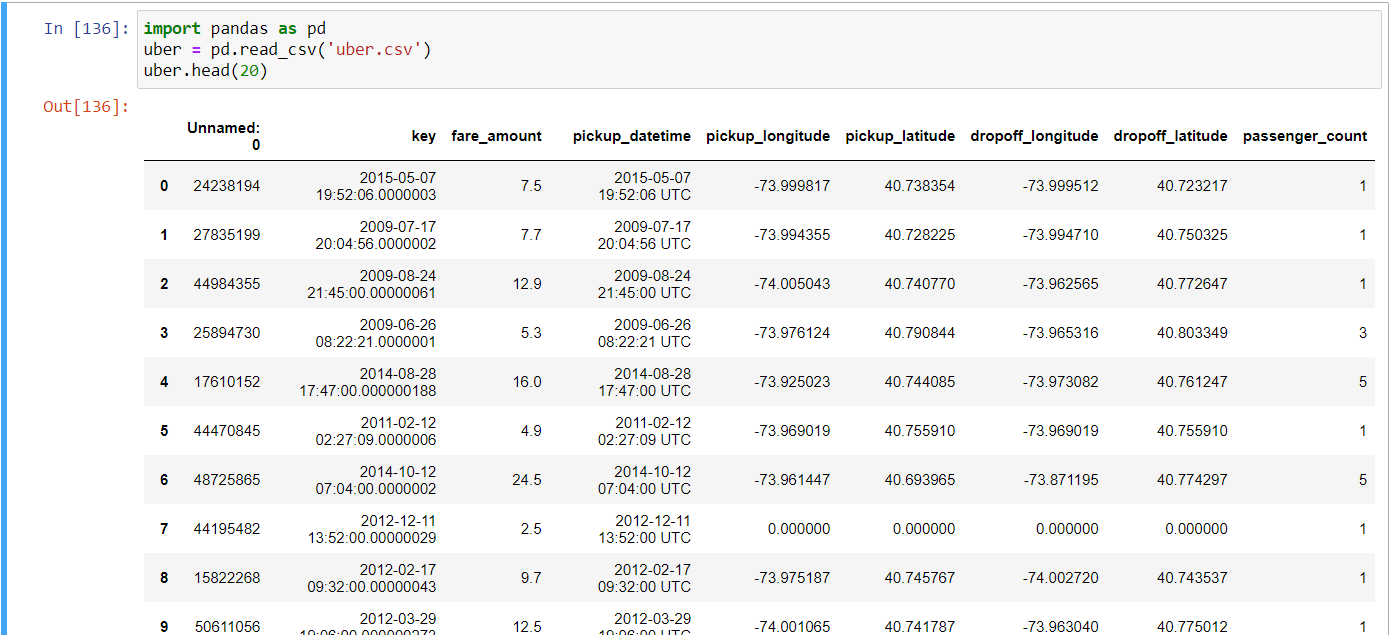
**Instructor:** Dr. Noman Durani

**Group Members:**

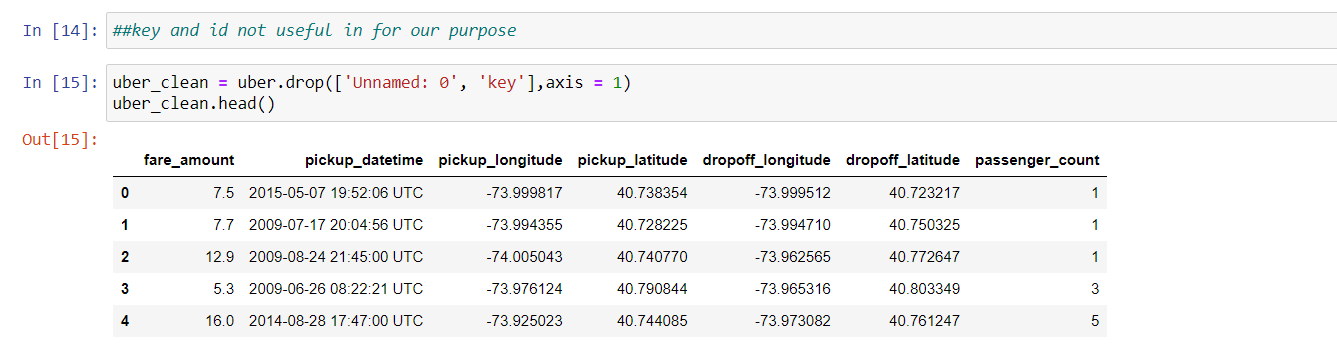
1. Aahil (20K-0308)
2. Faizan Hakim (20K-0290)

**Date:** 10th June, 2023

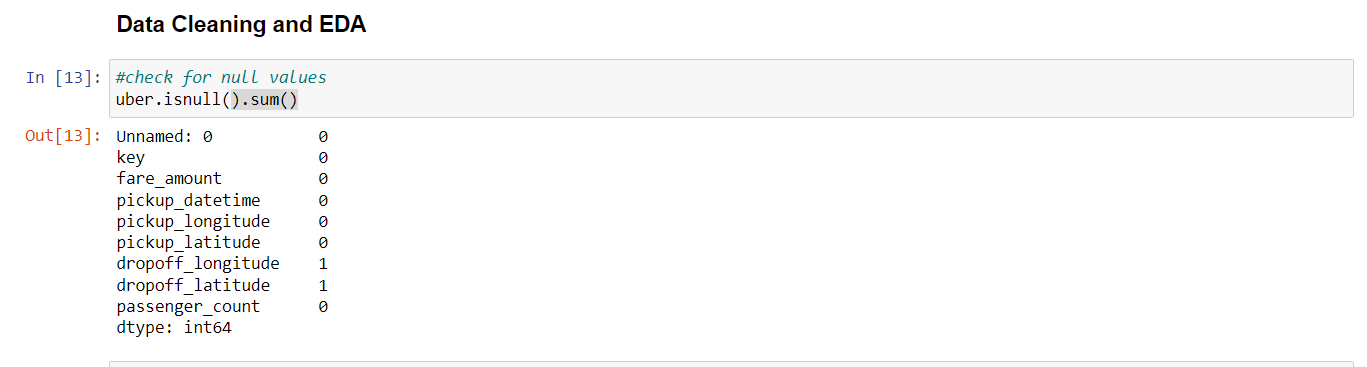
### **Data Cleaning and EDA:**

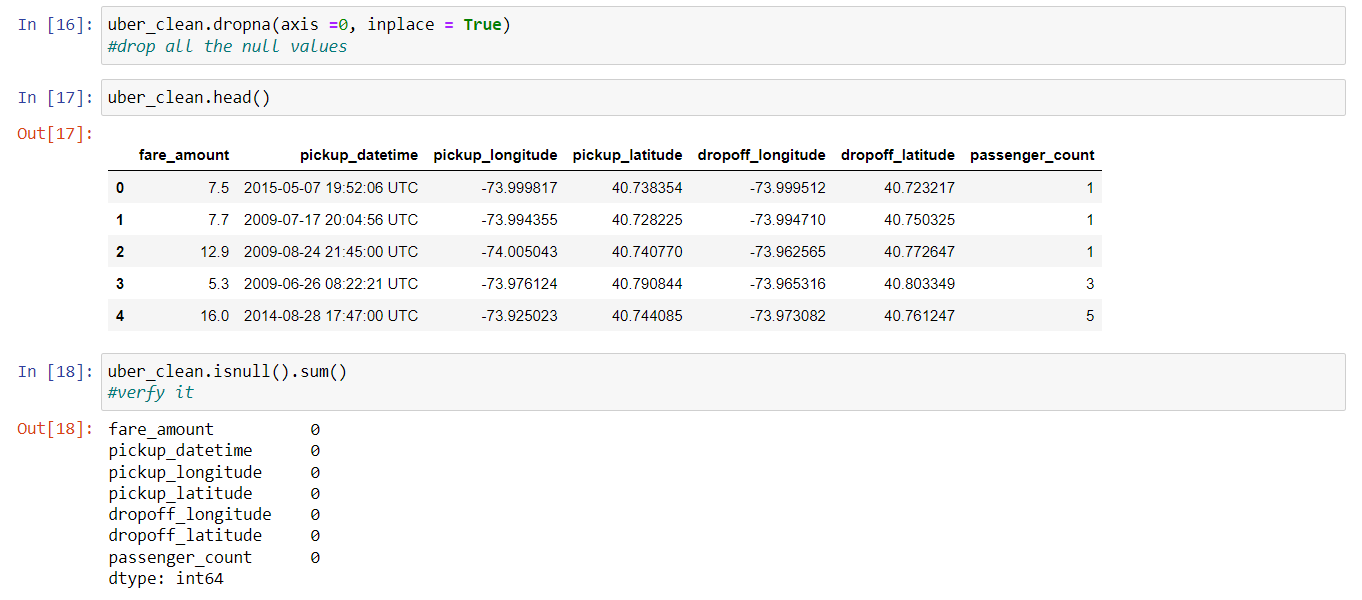


File contained columns that were not needed for our purpose, such as key and id. Thus, they were removed

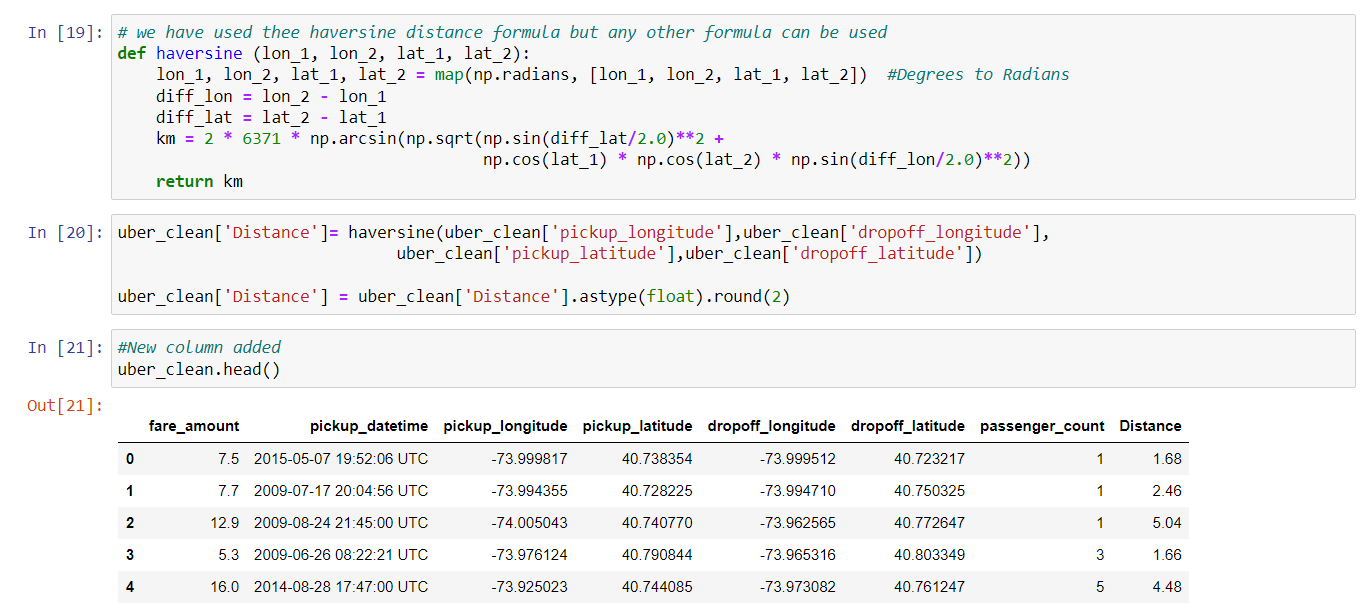


Null values were checked in every column and dropped

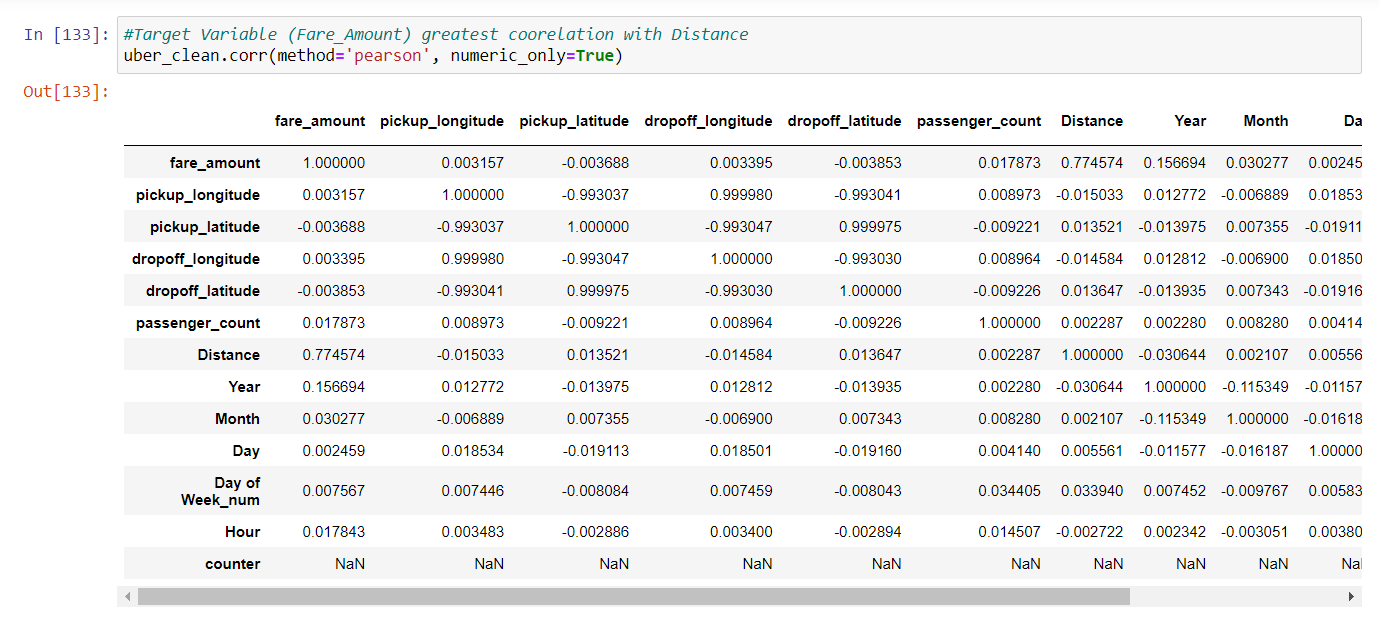




“isnull.sum()” was used to verify the presence of null values



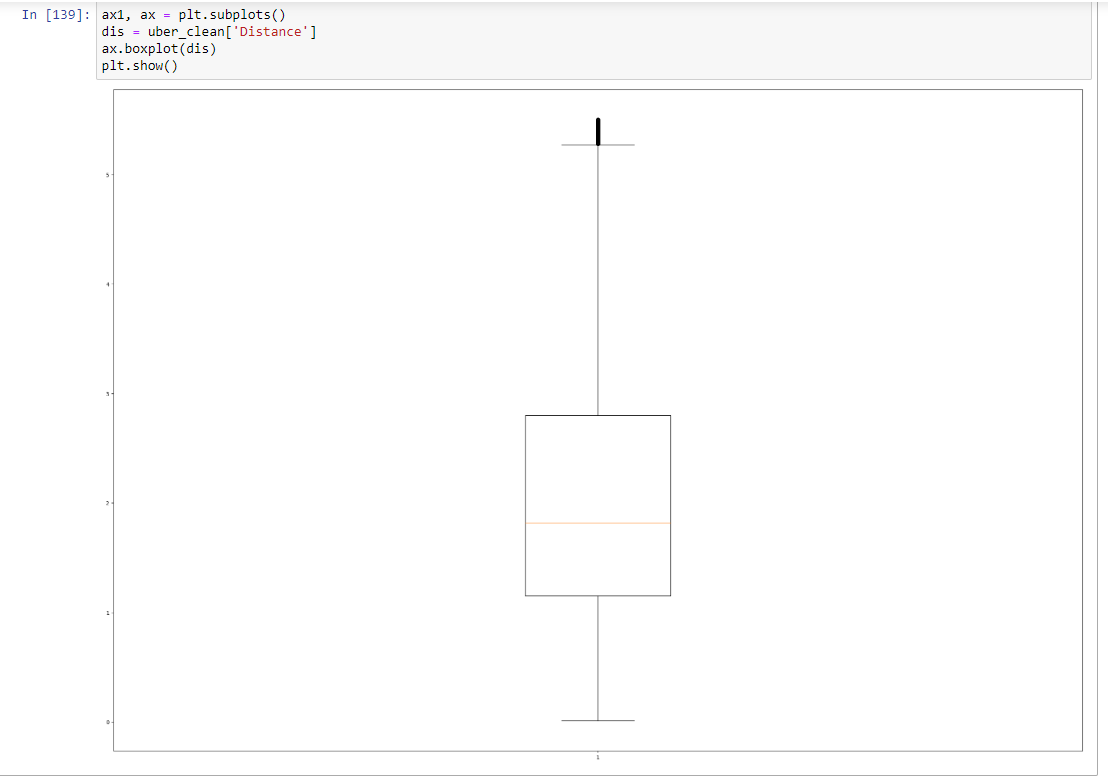
Haversine formula was defined in a function to calculate the distance from the pickup and dropoff columns to make use of the data and make it meaningful to identify the trends. The new Distance column was added to the table

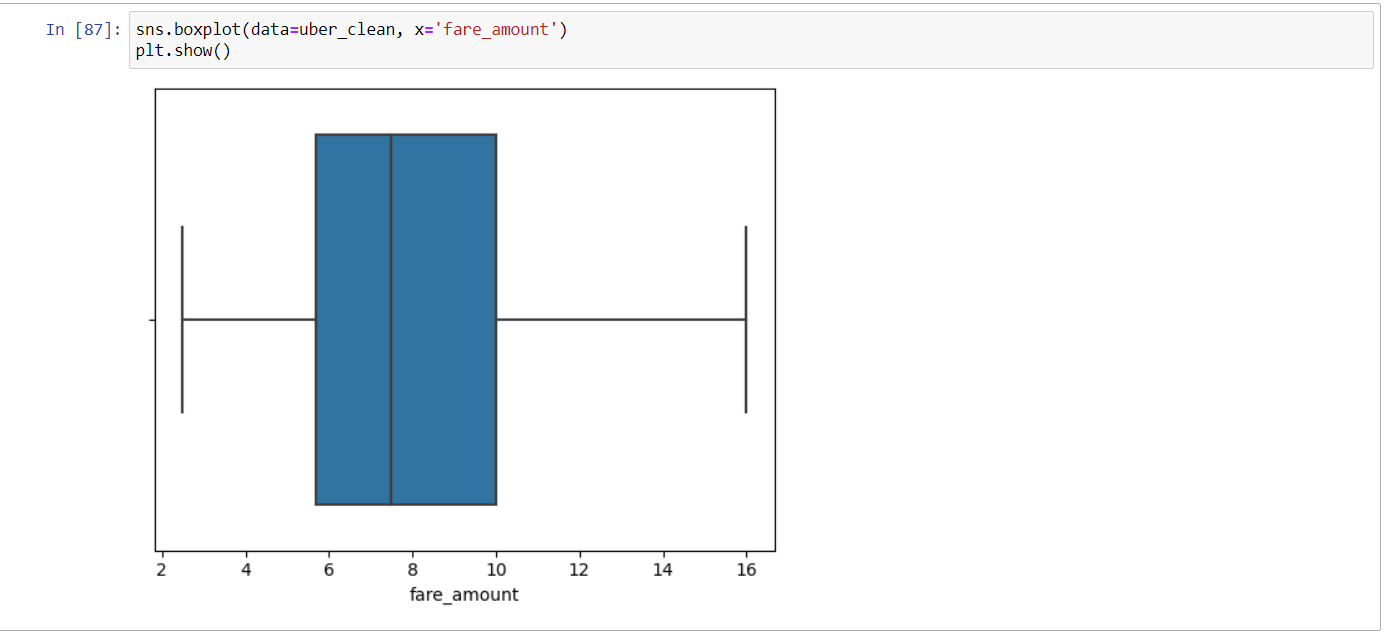


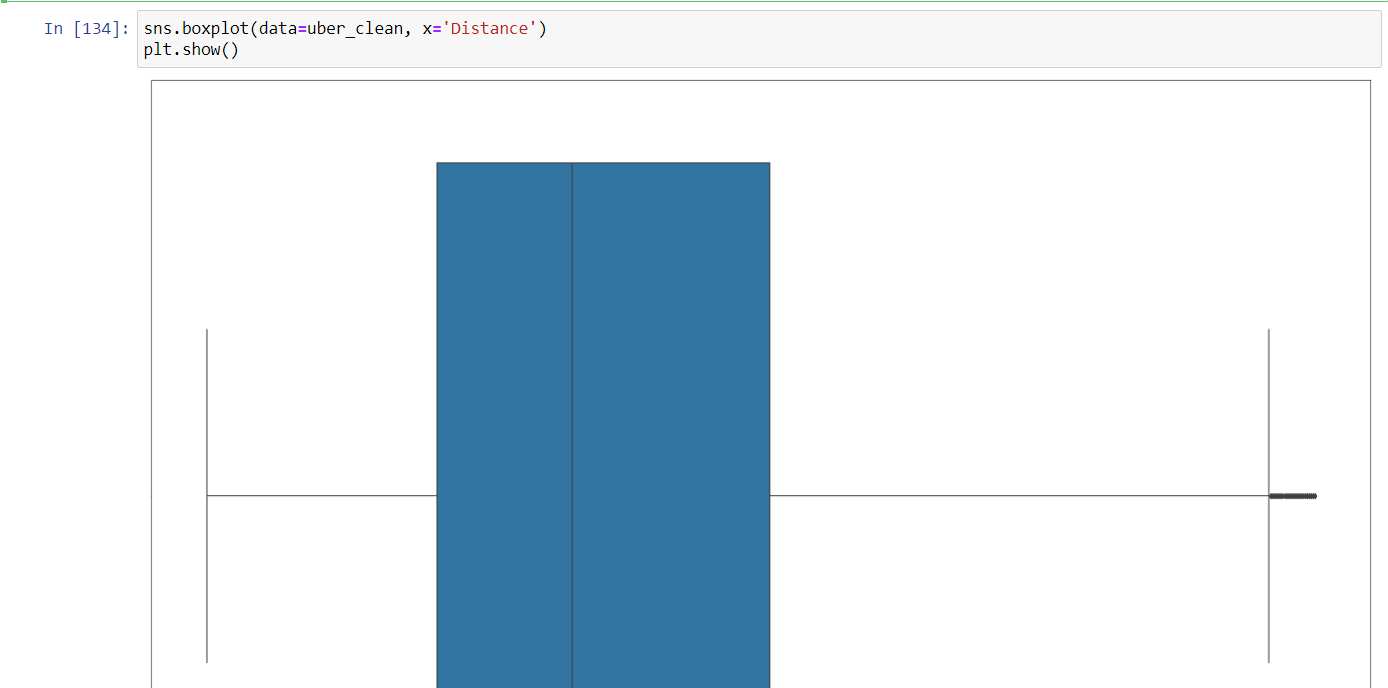
Pearson correlation was calculated for each variable against every other variable. **It was observed that Fare amount had the strongest relation with the Distance variable (0.774574)**

**Data Visualization**

**Box Plots:**

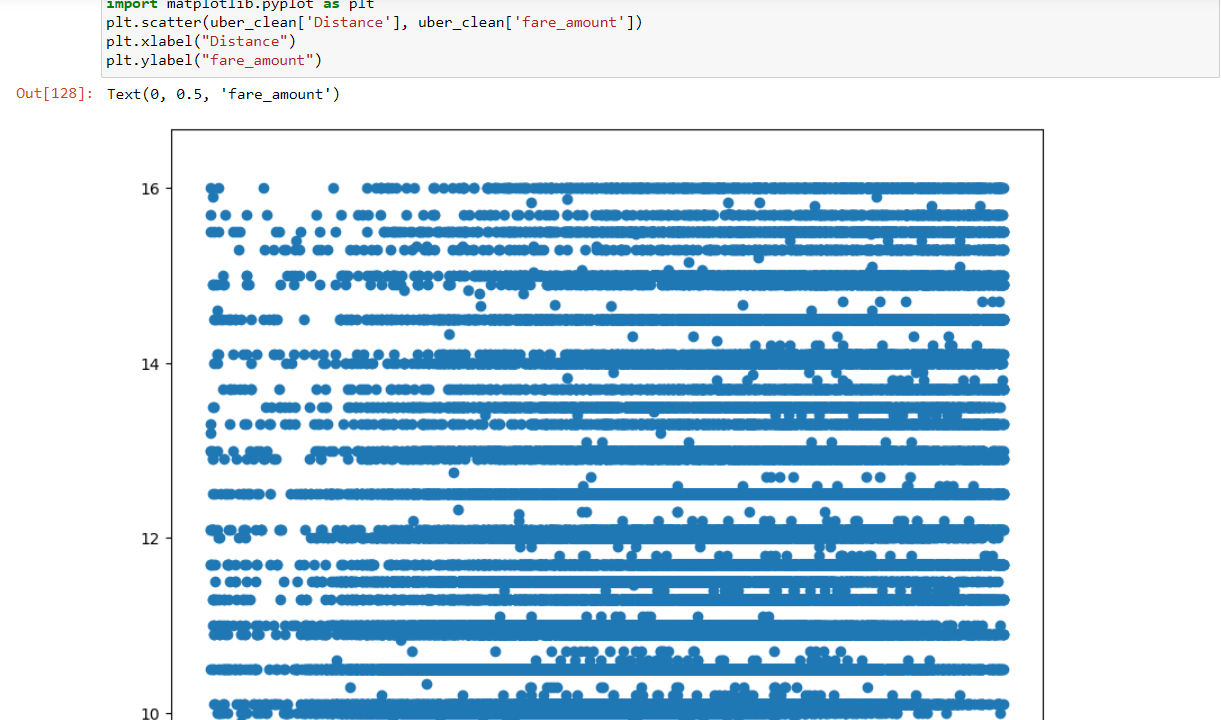




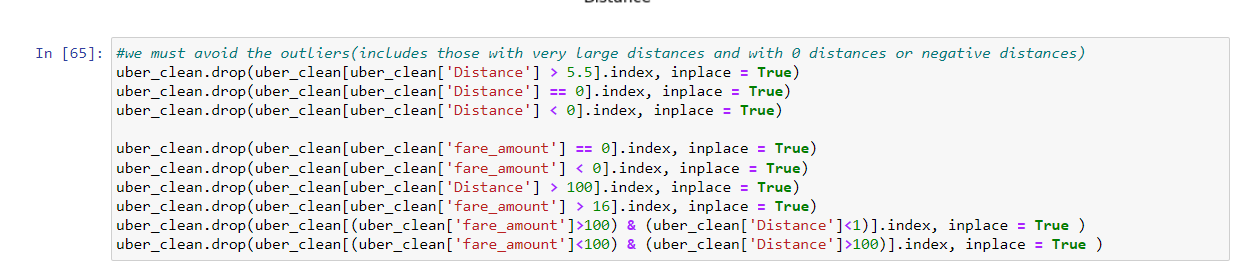


Box plots were displayed to help visualize the distribution which helped us to identify the outliers

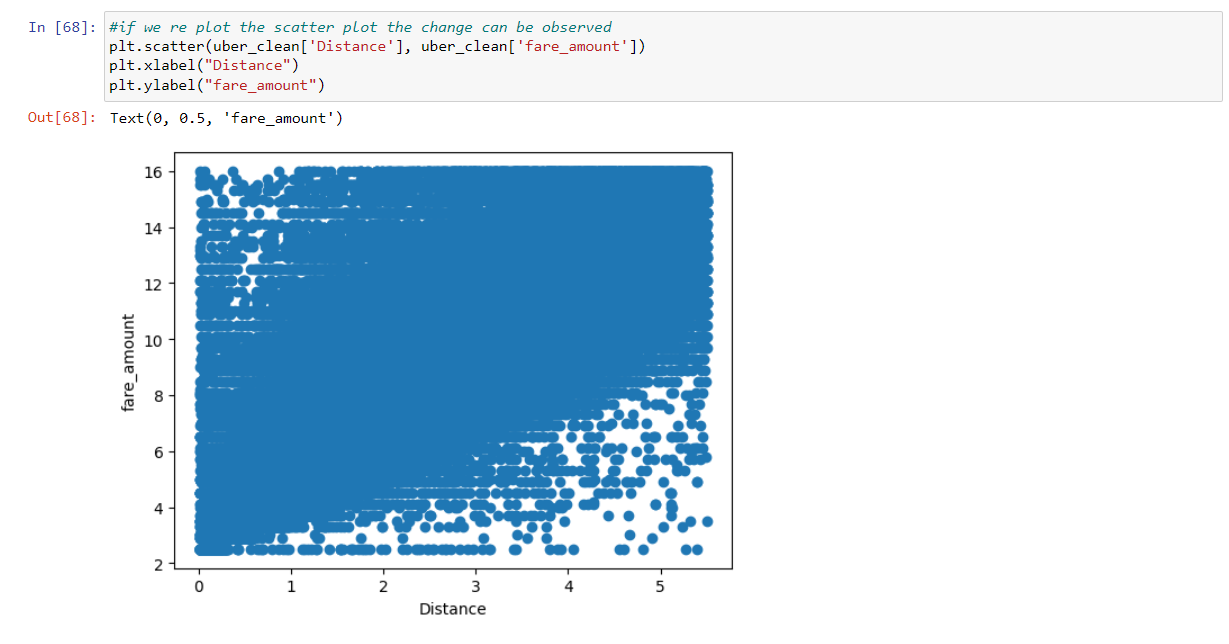
**Scatter Plots:**



Scatter Plots of fare vs distance were displayed to show the distribution and identify the relation between these two variables to check whether they fit for our purpose



Particular set of data was dropped carefully which were acting as outliers, affecting the distributions

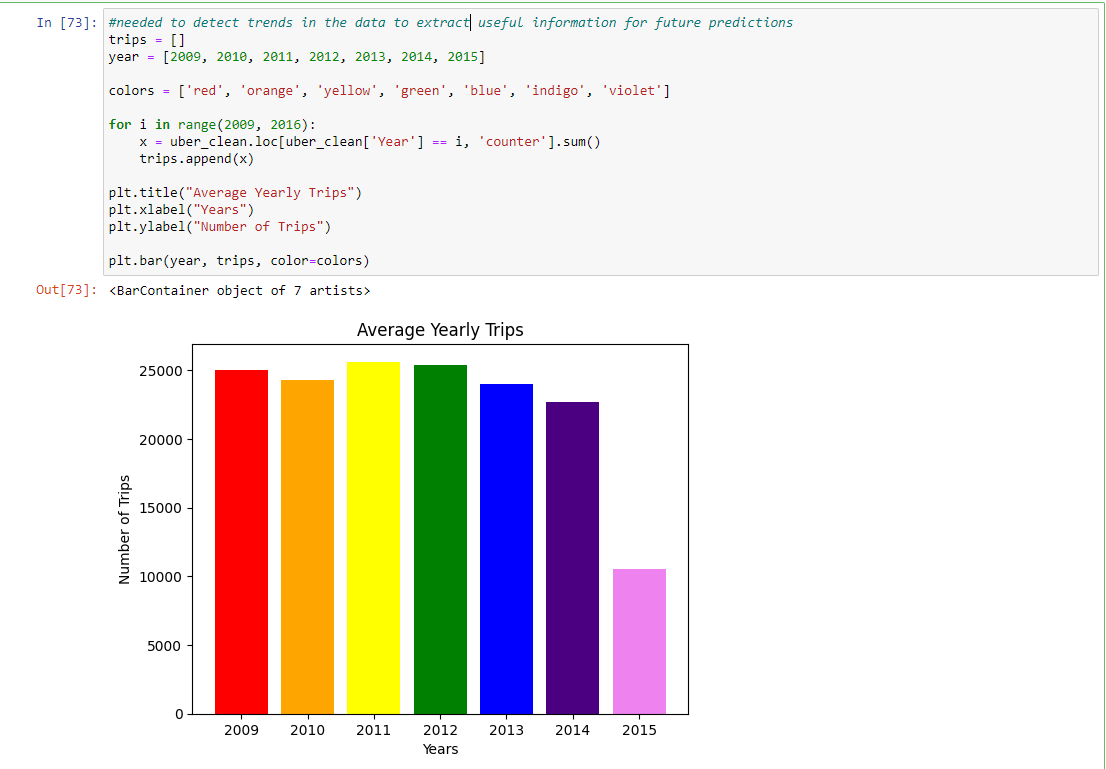


Scatterplot was redrawn. This shows better stronger relation and better distribution as outliers were removed

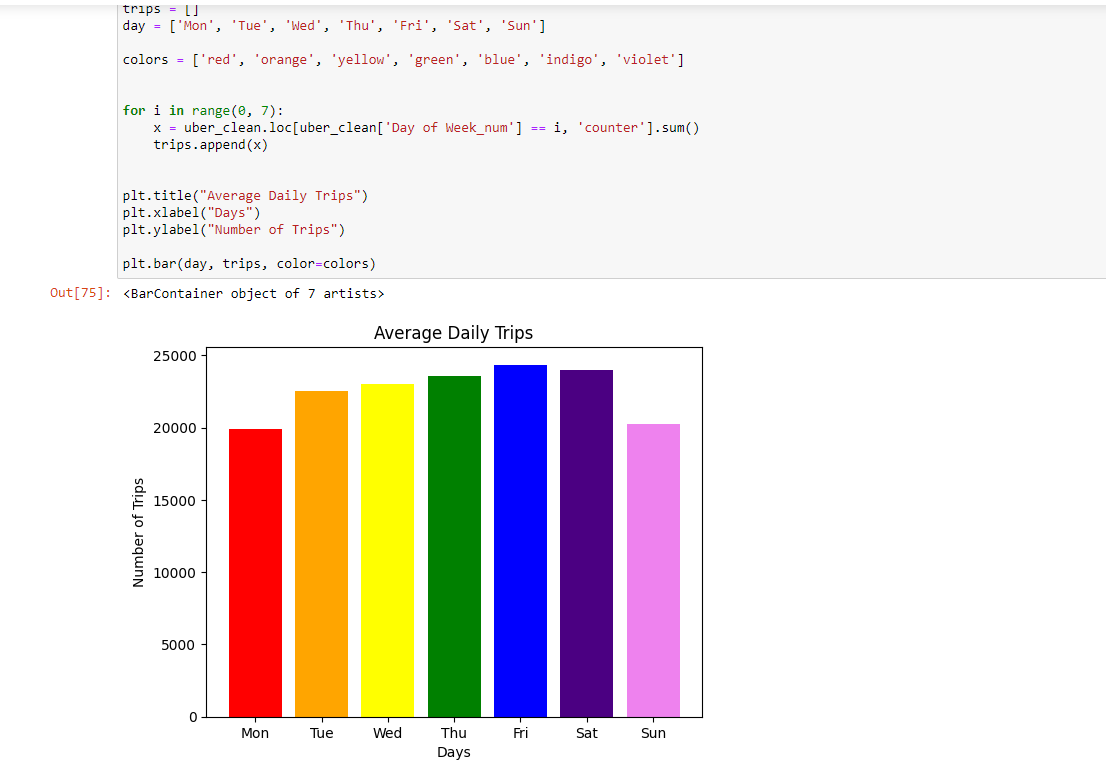


Date and time were separated into days, months and years to make data useful and identify trends clearly.

**Bar charts:**

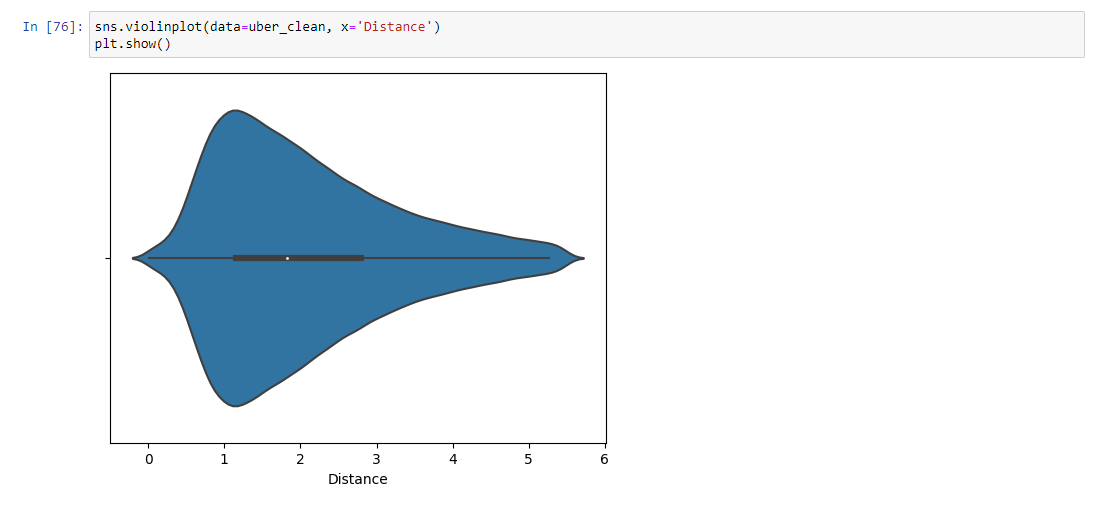






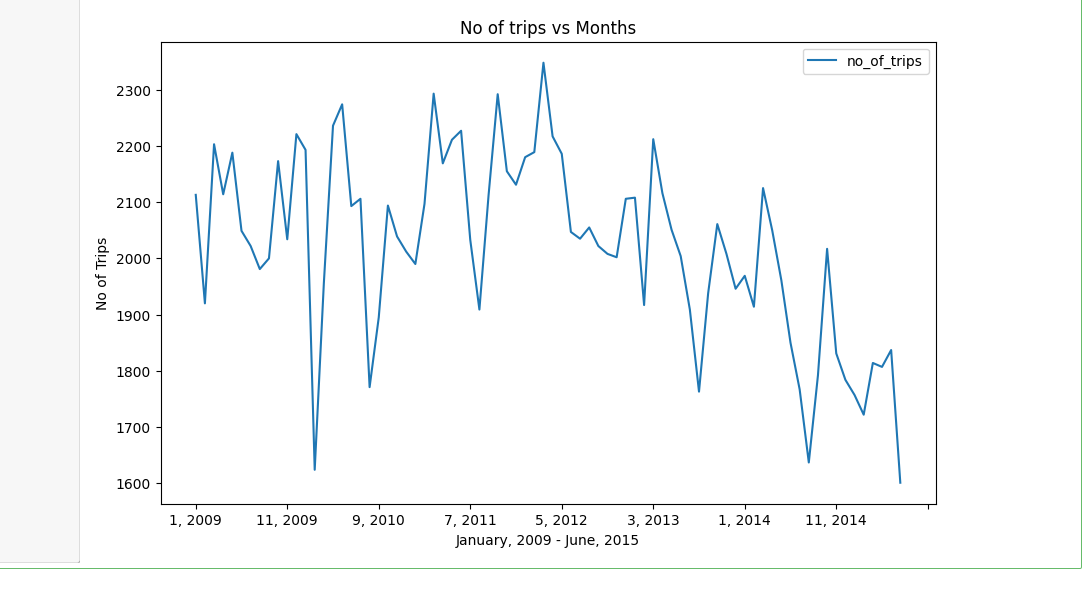
Bar charts were plotted to visualize the trips completed against days, months and years to help identify the trends

**Violin Plots:**



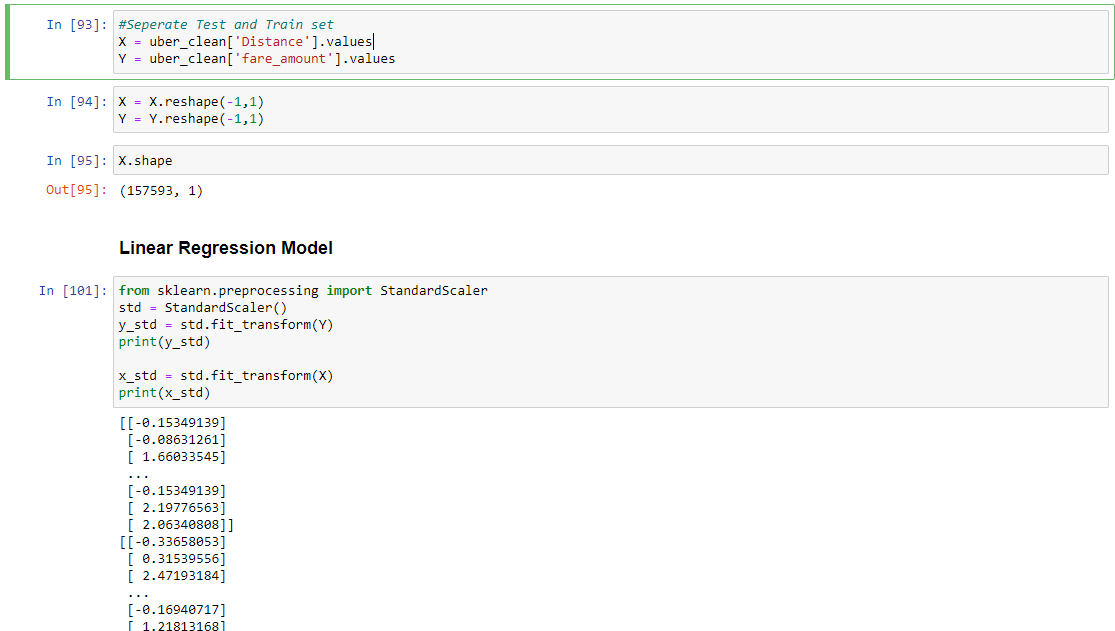
A violin plot is more informative than a plain box plot. While a box plot only shows summary statistics such as mean/median and interquartile ranges, the violin plot shows the full distribution of the data.

**Line Plots:**



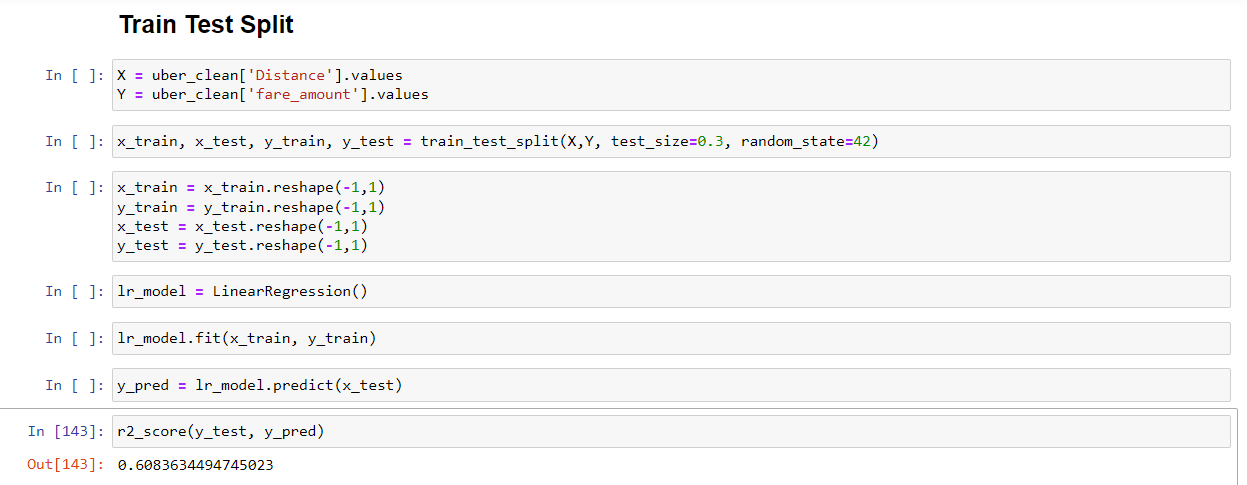
Line Plot of No. of trips vs months were plotted to check the trends.

## **Model Training**



Data was reshaped and scaled using StandardScaler to make it fit to perform linear regression on it.

**Train Test Split**



We used the train test split holdout approach as well. The accuracy output was 60.8%

**Ada Boosting with K-Fold Cross Validation Approach:**

We also performed Ada boosting with K-Fold approach to see if it increases accuracy and but there wasn’t much of a difference.

