# Table of Content

[**Table of Content**](#_8vezt4l2j7qv) **1**

[**Introduction**](#_7c52ymyugo0u) **1**

[**Methodology**](#_nkuckj25xrxr) **1**

[Retrieving and Preparing the data](#_dapo8m34632w) 1

[Feature Engineering](#_42i5jp89eoeb) 2

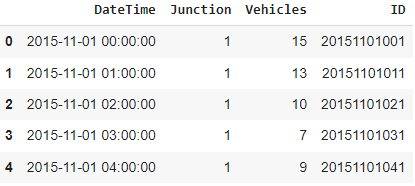
[Data Exploration](#_60wntzwderck) 2

[Data Modeling](#_1myaxdkq73ag) 5

[**Results**](#_ik3qe47z5ci) **6**

# Introduction

This task involves applying regression techniques to a traffic congestion dataset. The dataset has 4 features that are DateTime, Junction, Vehicles and ID. The datetime represents the time at which the data is collected, the junction feature gives us the junction from which the data was collected. There are four junctions in the dataset. The vehicle feature tells us about the number of vehicles at the specified junction while the ID is a unique identifier for each row. The total number of rows in the dataset is 48120.

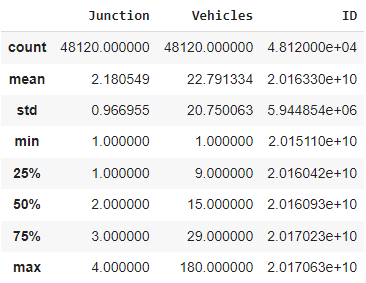


*First 5 rows of the dataset**.*

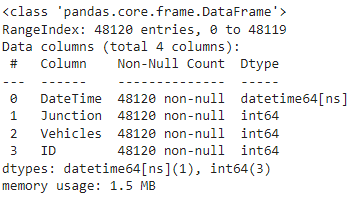
# Methodology

## Retrieving and Preparing the data

The traffic prediction dataset is provided in a CSV format so I have used Pandas for data loading and preparing the data. After loading the data, I have calculated some statistics about the features which include std, min max value, mean, etc.



I have checked the columns for missing and null values. For this dataset no missing or null values were found.

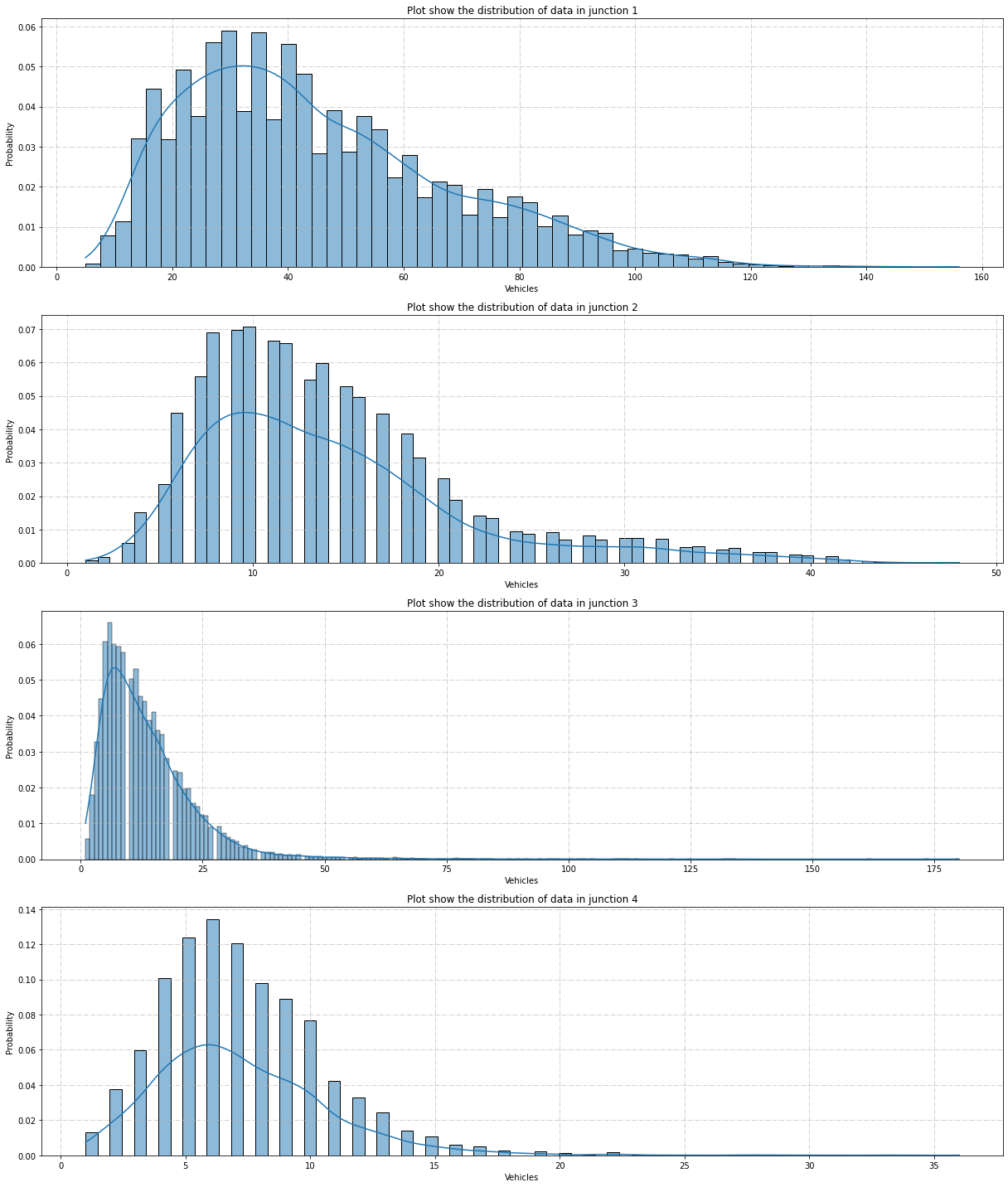


## Feature Engineering

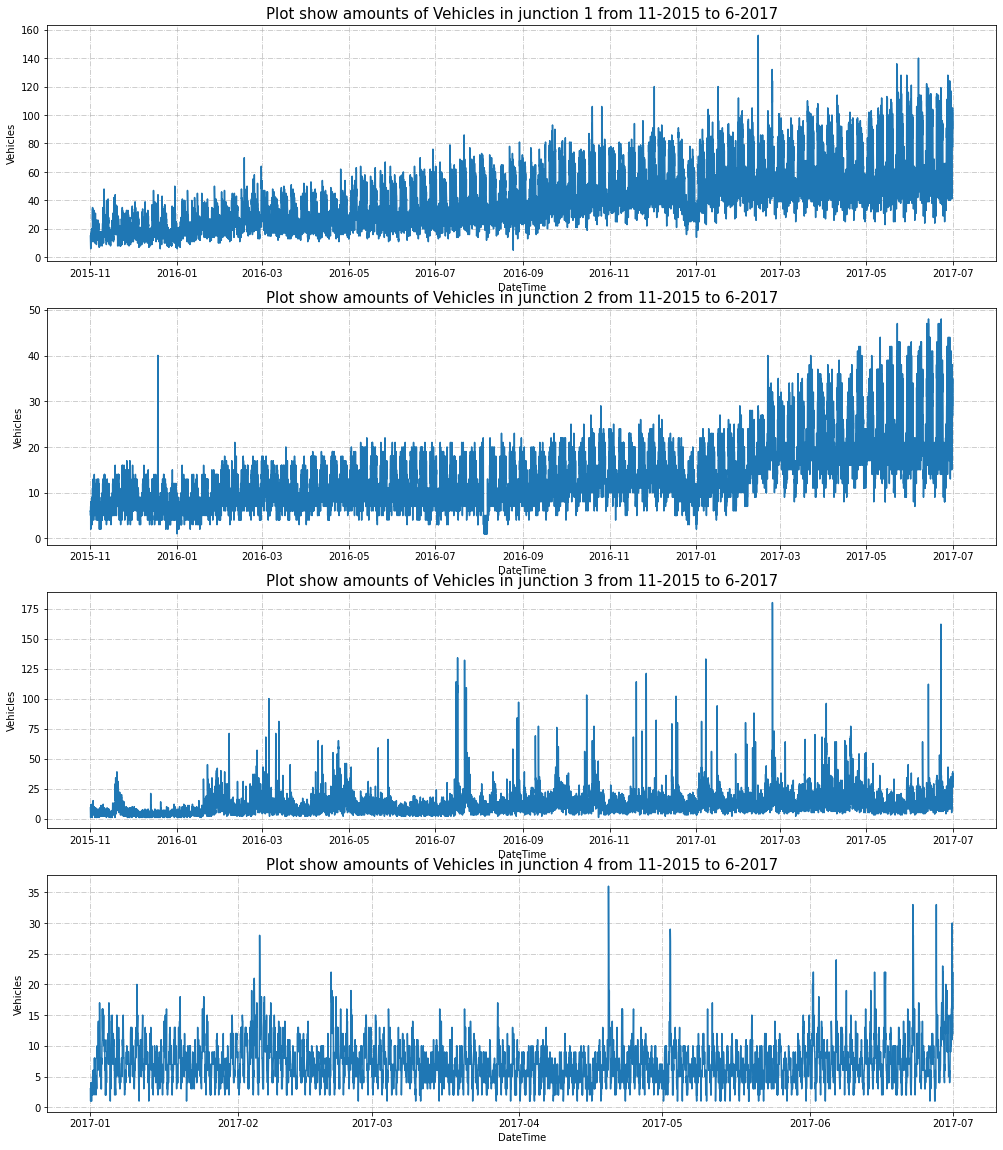
The dataset has a DateTime feature which gives both the date and time at which the data was collected. This feature in its current form is not of any use to us since both the date and time are concatenated together. I have separated this feature into four different features that are year, month, day, hour. Furthermore, the ID feature is not required for this task so I have dropped it.

## Data Exploration

For the task of data exploration, I have drawn the histograms of vehicles at each junction and have also displayed the probability for the number of vehicles. This gives us an idea of how many vehicles are to be expected at a junction



Moreover, I have also plotted the number of vehicles present at the junction from 11-2015 to 06-2017. These plots give us an annual trend and also display the increase in vehicles as the years progress along the x-axis.

I constructed a correlation heatmap to find the collinearity in-between features and between the features and the target variable. According to the correlation heatmap the most correlated events to vehicles are junction, year and hour.



## **Data Modeling**

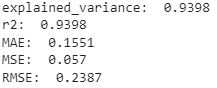
As part of preprocessing I have standardized the values of the vehicle feature. For the purpose of training and testing the model, I have split the dataset into two parts i.e., train set and test set. The testing data is 20% of the original dataset. Now to do the regression task for this dataset I have decided to use a Random forest regressor. I used Random Grid Search with 3-fold cross validation and 100 iterations to tune some of the hyperparameters of the random forest regressor

| **Hyperparamters** | **Values** |
| --- | --- |
| Number of estimators | 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 |
| Max features | auto, sqrt |
| Max depth | 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None |
| Min samples split | 2, 5, 10 |
| Min samples leaf | 1, 2, 4 |
| Bootstrap | True, False |

After running the random grid search, I found the best values to be bootstrap: True, max depth :20, max features: sqrt, minimum samples leaf: 1, minimum samples split: 2, number of estimators: 500. I used these hyperparameters to train the model.

# Results

For finding the evaluation metrics I have used explained\_variance R2-score, mean absolute error, mean square error, root mean square error. The results of these are as follows:



The model has given satisfactory evaluation metrics.