Traffic Data Analysis

By:
Prerna Desai
Jash Shah
BIA-786-C
Prof: Denghui Zhang

Content

01 Introduction

02 Exploring Our Data Set

03 Our Goal

04 Exploratory Analysis

05 Statistical Analysis

06 Data Visualization

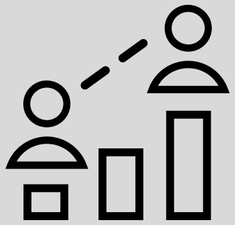
07 Modeling and Predictions

08 Conclusions

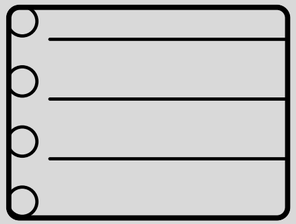
Introduction

Traffic congestion is a pressing issue affecting cities globally, driven by factors such as population growth and aging infrastructure. In response, cities are increasingly reliant on data-driven solutions to manage traffic effectively. This analysis focuses on a dataset containing hourly vehicle count observations at four junctions.

Exploring out Data-set



The dataset comprises records of the number of vehicles at a junction at a particular date and time. This Dataset contains 4 columns and 48k rows.



1. Datetime (Categorical/Numeric)
2. Junction (Numeric)
3. Vehicles (Numeric)
4. Id(Numeric)

The dataset is structured and stored in CSV format. This dataset does not have many data-cleaning requirements.

Our Goal

The goal of this analysis is to develop a robust predictive traffic analysis model that accurately predicts vehicle congestion patterns at four different junctions based on historical data.

The model aims to provide actionable insights for city planners and traffic management authorities to optimize signal timing, implement congestion mitigation strategies, and improve overall traffic flow efficiency in urban areas.

Through the utilization of machine learning techniques and statistical modeling, the goal is to contribute to the development of effective strategies and technologies for managing urban traffic congestion, ultimately enhancing the sustainability and livability of cities worldwide.

Exploratory Data Analysis

	DateTime	Junction	Vehicles	ID
0	2015-11-01 00:00:00	1	15	20151101001
1	2015-11-01 01:00:00	1	13	20151101011
2	2015-11-01 02:00:00	1	10	20151101021
3	2015-11-01 03:00:00	1	7	20151101031
4	2015-11-01 04:00:00	1	9	20151101041
...	...	...	...	...
48115	2017-06-30 19:00:00	4	11	20170630194
48116	2017-06-30 20:00:00	4	30	20170630204
48117	2017-06-30 21:00:00	4	16	20170630214
48118	2017-06-30 22:00:00	4	22	20170630224
48119	2017-06-30 23:00:00	4	12	20170630234

48120 rows × 4 columns

```
df.isnull().sum()
```

```
DateTime    0
Junction    0
Vehicles    0
ID          0
dtype: int64
```

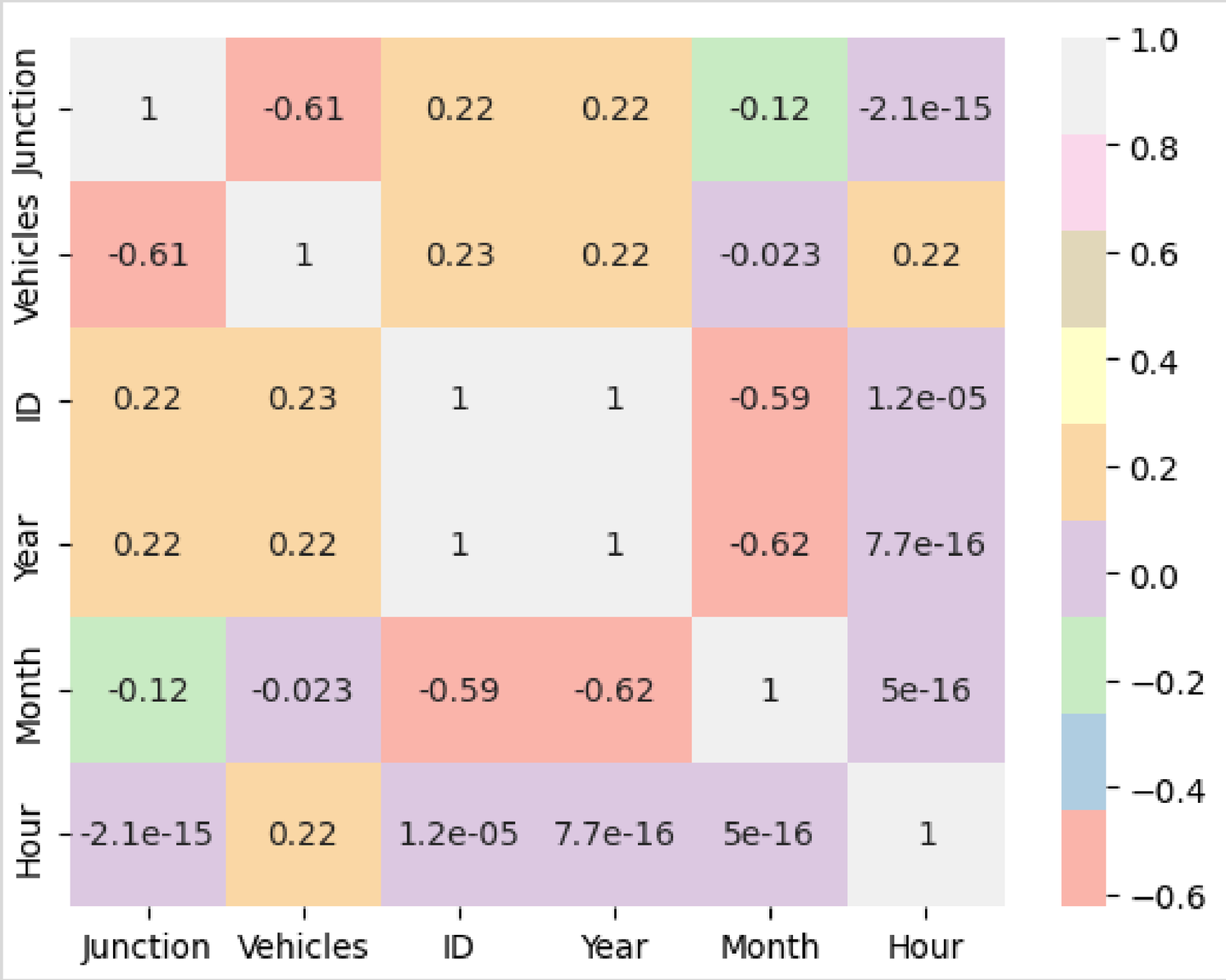
There are no null values in our data set, hence there is no need for data cleaning procedures.

Statistical Analysis

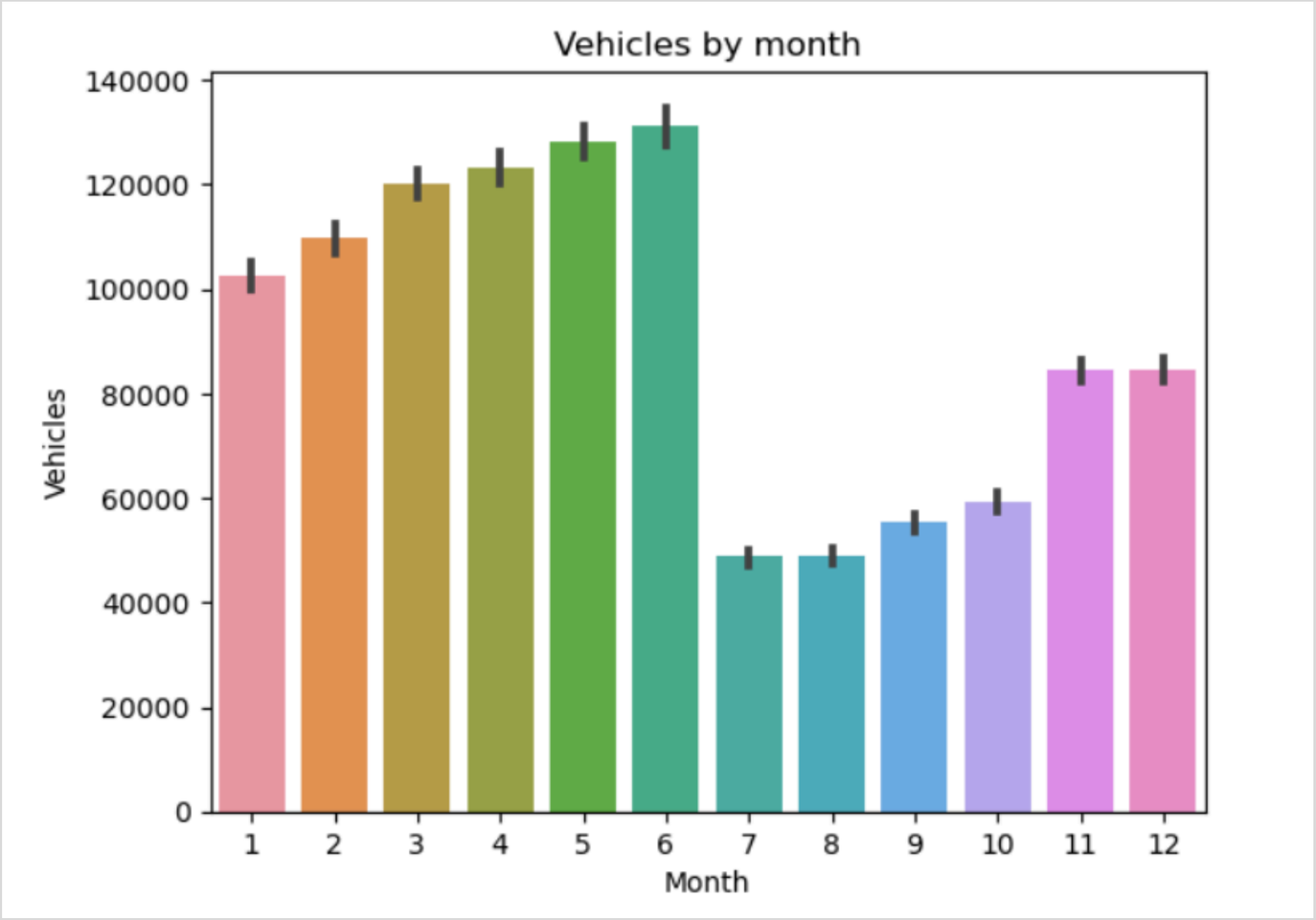
	Junction	Vehicles	ID
count	48120.000000	48120.000000	4.812000e+04
mean	2.180549	22.791334	2.016330e+10
std	0.966955	20.750063	5.944854e+06
min	1.000000	1.000000	2.015110e+10
25%	1.000000	9.000000	2.016042e+10
50%	2.000000	15.000000	2.016093e+10
75%	3.000000	29.000000	2.017023e+10
max	4.000000	180.000000	2.017063e+10

Data Visualization

The correlation matrix quantifies the strength and direction of linear relationships between the variables Junction, Vehicles, ID, Year, Month, and Hour, providing insights into potential associations or dependencies within the dataset

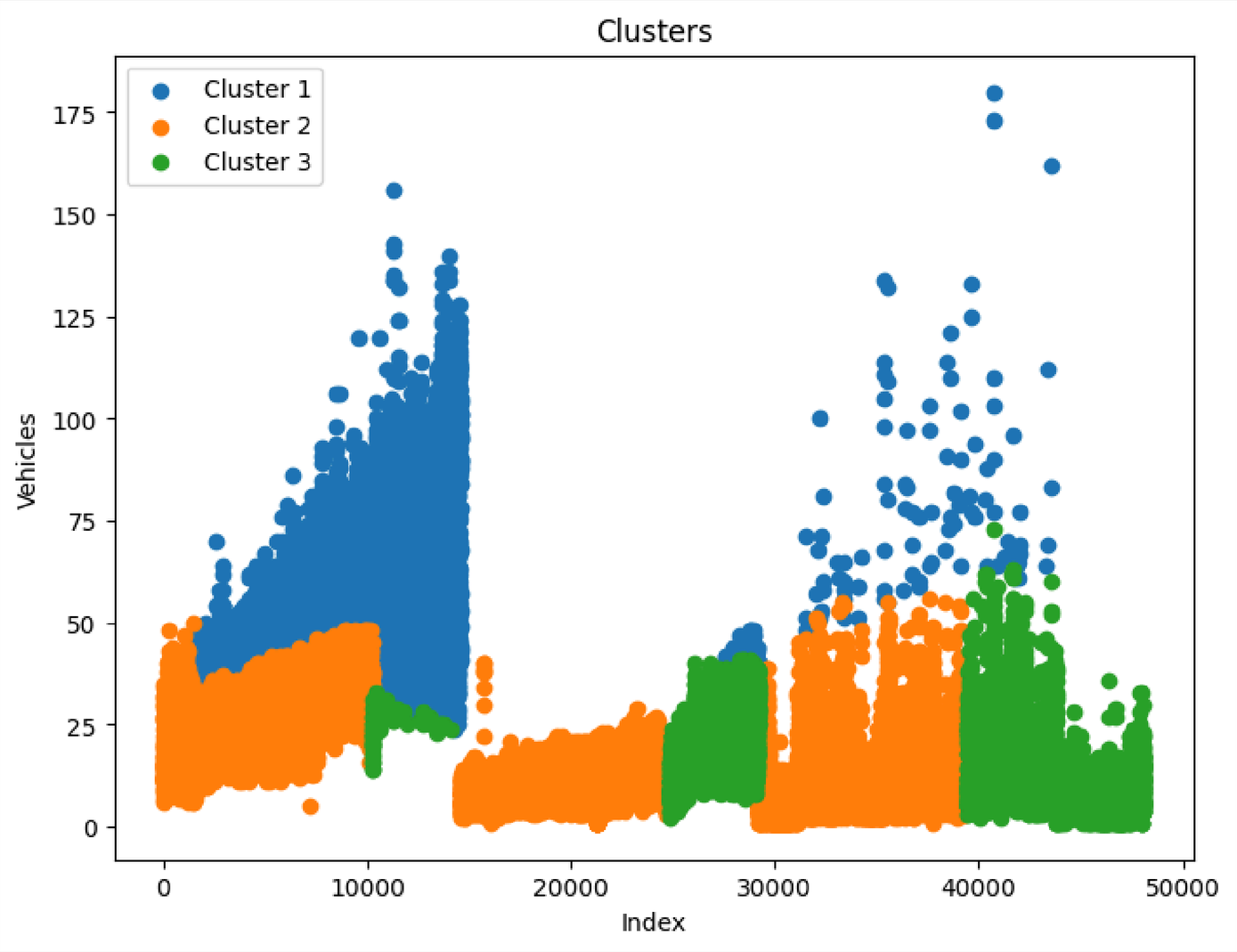


Data Visualization



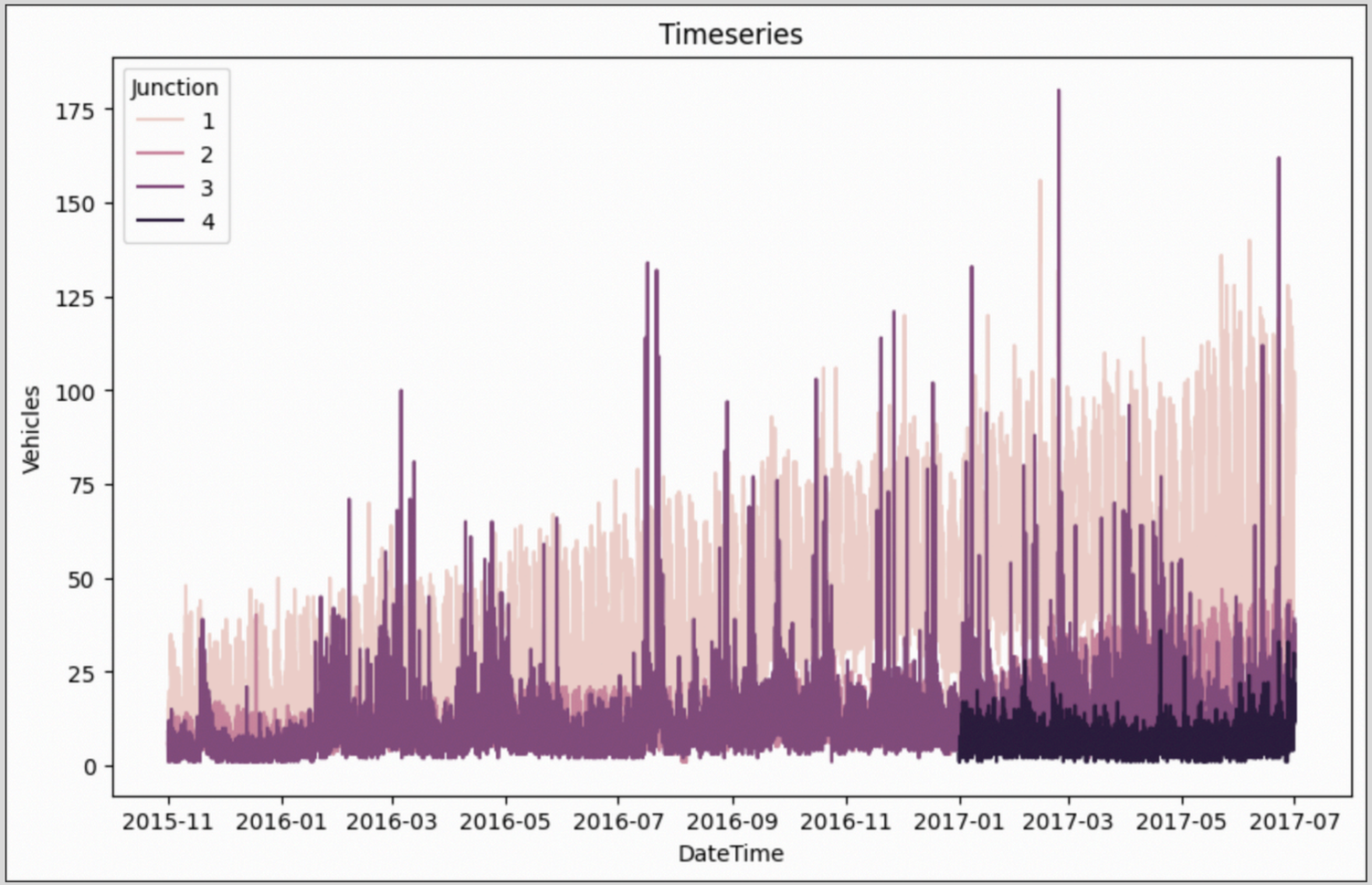
The code shows the quantity of vehicles in relation to the month. The y-axis represents the number of vehicles, and the x-axis displays the months.

Interpreting Clusters Formed with K-means Clustering



Each cluster represents a subset of observations sharing common characteristics in terms of vehicle counts, enabling the identification of different traffic patterns or congestion levels.

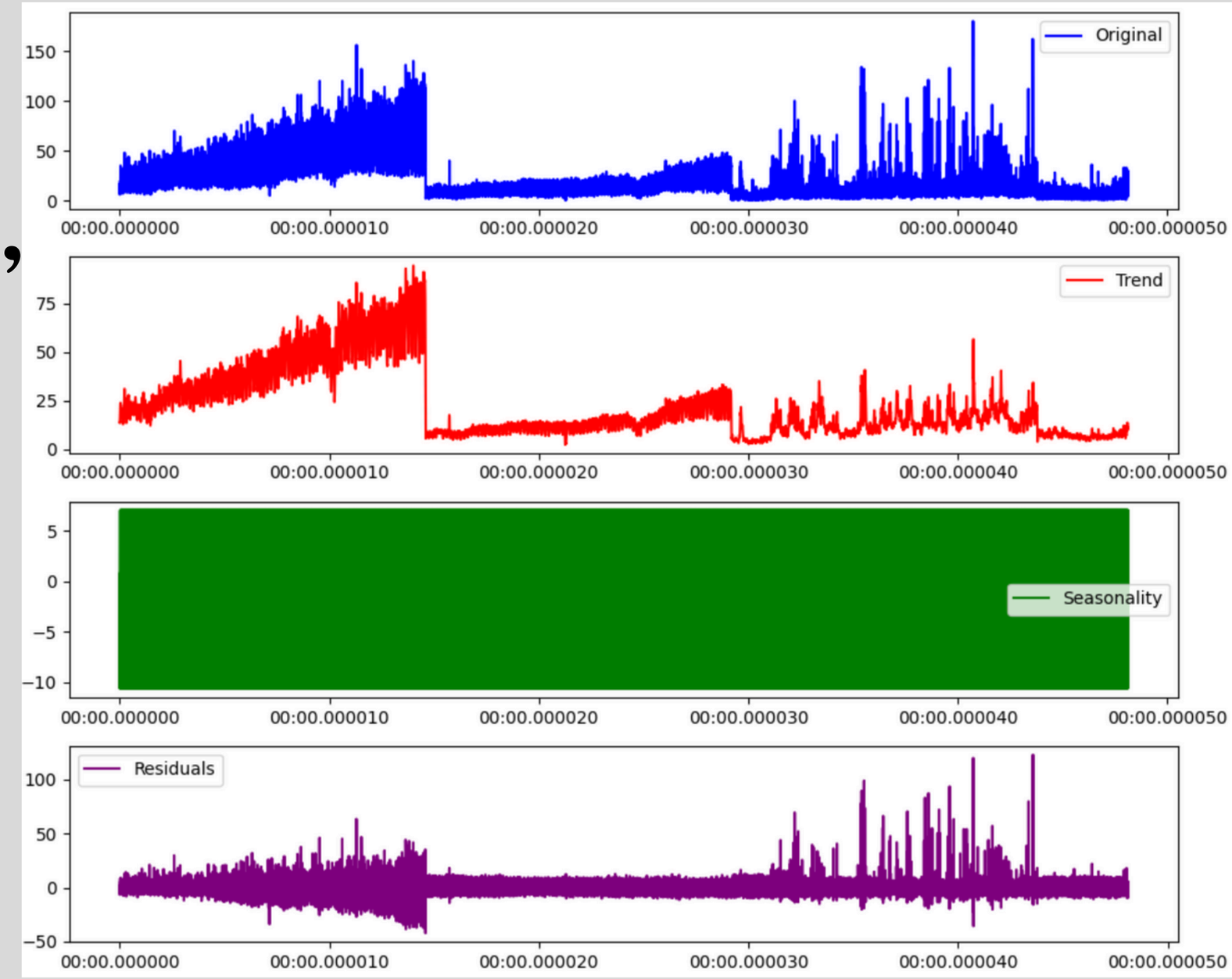
Time Series Analysis



a time series plot that shows the number of vehicles over the date, broken down by Junctions. For certain junctions, each line shows the number of vehicles over date.

"Time Series Decomposition: Trend, Seasonality, and Residual Analysis"

The Graph provides insights into the underlying patterns of a time series dataset by separating it into trend, seasonality, and residual components, aiding in understanding and forecasting future trends.



Modeling and Predictons

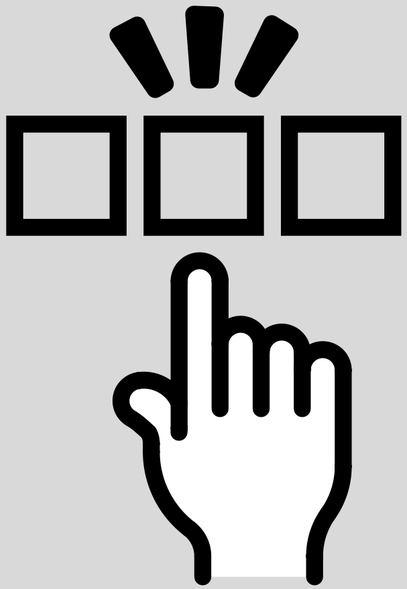
```
Training Random Forest...
Random Forest - Mean Squared Error: 0.0015575644222776326
Random Forest - Mean Absolute Error: 0.000696176226101409
Random Forest - Root Mean Squared Error: 0.039465990704372705
Random Forest - Accuracy: 1.0
Random Forest - Precision: 1.0
Random Forest - Recall: 1.0
Random Forest - AUC: 1.0

Training XGBoost...
XGBoost - Mean Squared Error: 0.07493467388205773
XGBoost - Mean Absolute Error: 0.1606024728030437
XGBoost - Root Mean Squared Error: 0.2737419841421073
XGBoost - Accuracy: 0.9773482959268496
XGBoost - Precision: 0.9673359304764759
XGBoost - Recall: 1.0
XGBoost - AUC: 0.9655934343434344

Training KNN...
KNN - Mean Squared Error: 0.07676410824789878
KNN - Mean Absolute Error: 0.1443959545580493
KNN - Root Mean Squared Error: 0.27706336504110174
KNN - Accuracy: 0.9980257689110557
KNN - Precision: 0.9975266656361107
KNN - Recall: 0.9995353159851301
KNN - AUC: 0.9972424054673126

Training Decision Tree...
Decision Tree - Mean Squared Error: 0.005429135494596842
Decision Tree - Mean Absolute Error: 0.0010910224438902742
Decision Tree - Root Mean Squared Error: 0.07368266753176653
Decision Tree - Accuracy: 1.0
Decision Tree - Precision: 1.0
Decision Tree - Recall: 1.0
Decision Tree - AUC: 1.0
```

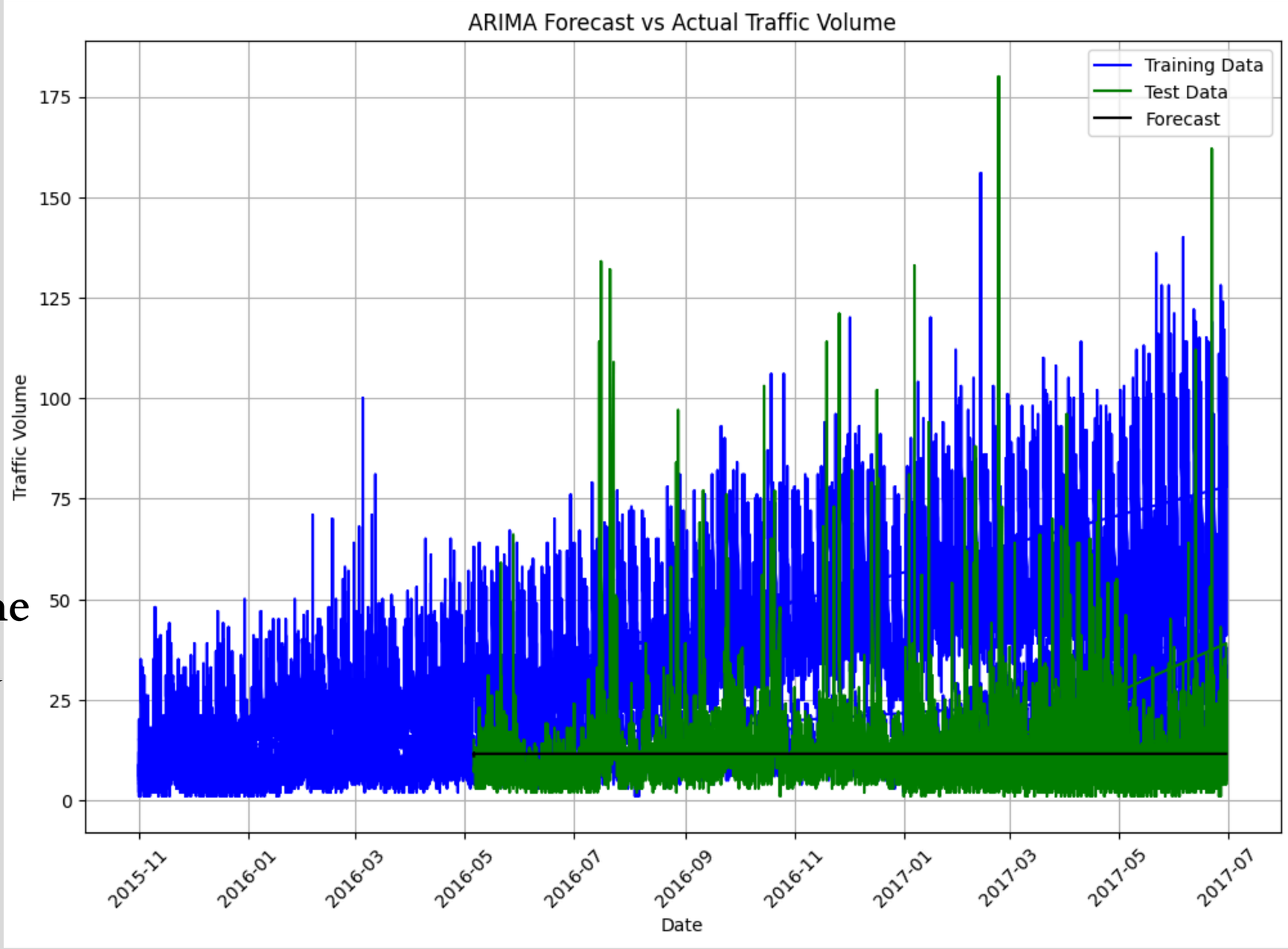
Optimal Model



```
=== Best Model ===
Name: Random Forest
Parameters: {'max_depth': None, 'n_estimators': 100}
MSE: 0.0015575644222776326
MAE: 0.0006961762261014 09
RMSE: 0.03946599070437 2705
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
AUC: 1.0
```

Analyzing Traffic Volume Forecasting with ARIMA

The blue line represents the training data, the green line represents the test data, and the red line shows the forecasted values, allowing for an assessment of the model's predictive performance over time.



Our Team

Prerna Desai

Masters in Data Science

+1-(610)-235-9718

pdesai21@stevens.edu

Jash Shah

Masters in Data Science

+1-724-680-3819

jshah67@stevens.edu

THANK YOU!