Industrial Internship Report on

”Forecasting of Smart city traffic patterns”

Prepared by

Kanike Sai Datta

Executive Summary

This report provides details of the Industrial Internship provided by upskill Campus and The IoT Academy in collaboration with Industrial Partner UniConverge Technologies Pvt Ltd (UCT).

This internship was focused on a project/problem statement provided by UCT. We had to finish the project including the report in 6 weeks’ time.

My project is to develop a robust traffic management system for smart cities by analyzing traffic patterns at four junctions, incorporating variations on holidays and other occasions, to enhance efficiency and provide input for future infrastructure planning.

This internship gave me a very good opportunity to get exposure to Industrial problems and design/implement solution for that. It was an overall great experience to have this internship.

TABLE OF CONTENTS

1 Preface

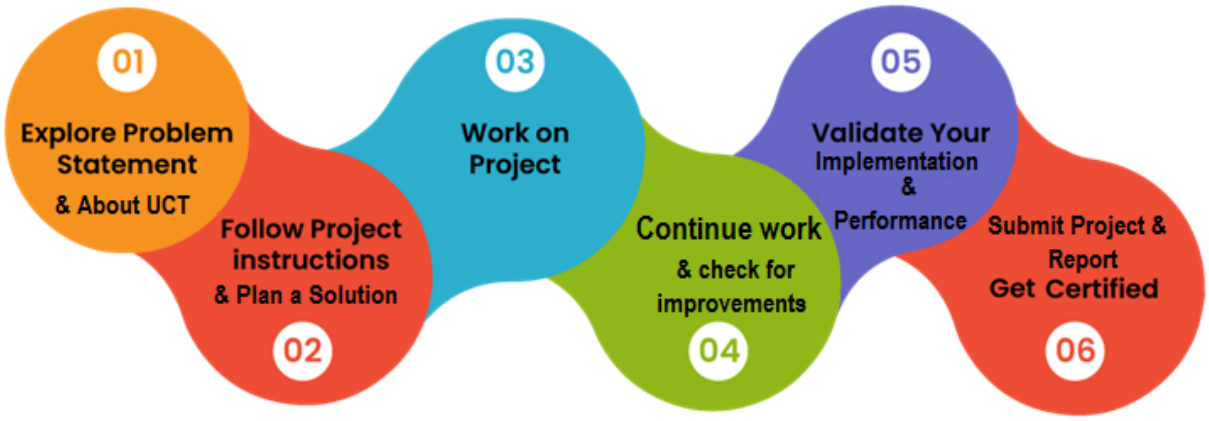
Summary of the whole 6 weeks' work.

About need of relevant Internship in career development.

Brief about Your project/problem statement.

Opportunity given by USC/ UCT.

How Program was planned



Your Learnings and overall experience.

Thanks to all, who have helped you directly or indirectly.

Your message to your juniors and peers.

2 Introduction

2.1 About UniConverge Technologies Pvt Ltd

A company established in 2013 and working in Digital Transformation domain and providing Industrial solutions with prime focus on sustainability and RoI.

For developing its products and solutions it is leveraging various **Cutting Edge Technologies e.g. Internet of Things (IoT), Cyber Security, Cloud computing (AWS, Azure), Machine Learning, Communication Technologies (4G/5G/LoRaWAN), Java Full Stack, Python, Front end** etc.

i. UCT IoT Platform (UCT Insight)

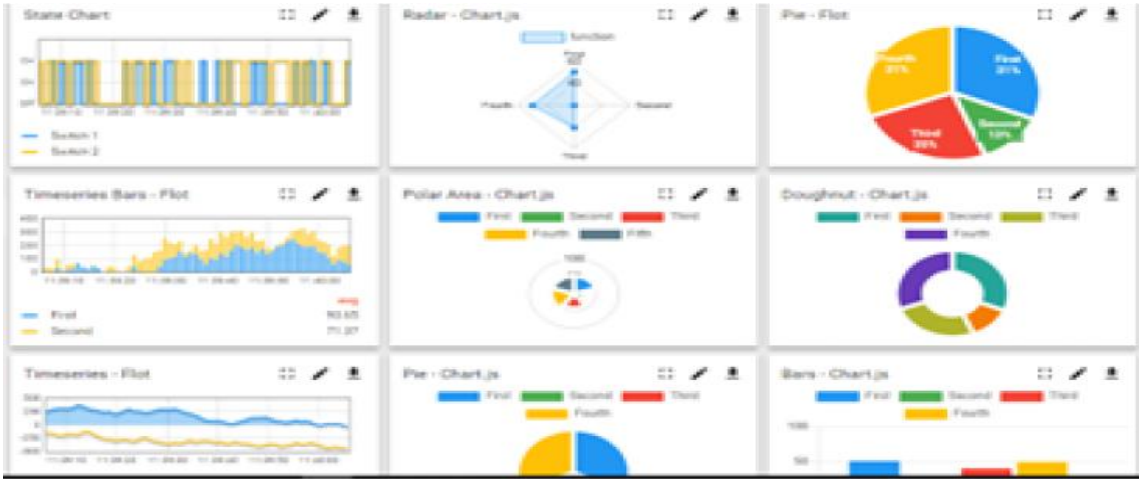
UCT Insight is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable "insight" for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

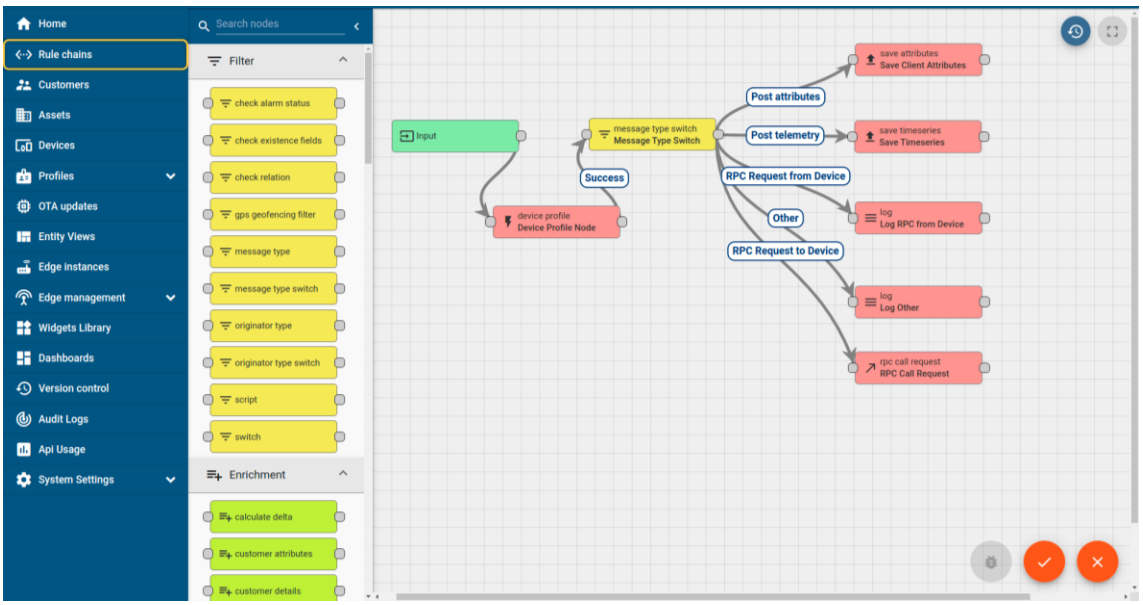
- It enables device connectivity via industry standard IoT protocols - MQTT, CoAP, HTTP, Modbus TCP, OPC UA

- It supports both cloud and on-premises deployments.

It has features to

- Build Your own dashboard
- Analytics and Reporting
- Alert and Notification
- Integration with third party application (Power BI, SAP, ERP)
- Rule Engine





ii. Smart Factory Platform (FACTORY WATCH)

Factory watch is a platform for smart factory needs.

It provides Users/ Factory

- with a scalable solution for their Production and asset monitoring
- OEE and predictive maintenance solution scaling up to digital twin for your assets.
- to unleased the true potential of the data that their machines are generating and helps to identify the KPIs and also improve them.
- A modular architecture that allows users to choose the service that they what to start and then can scale to more complex solutions as per their demands.

Its unique SaaS model helps users to save time, cost and money.

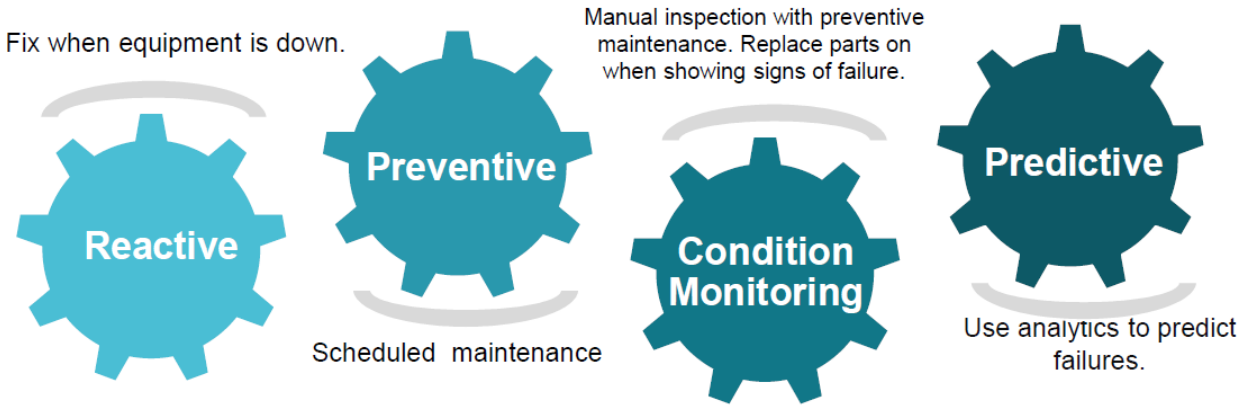
Machine	Operator	Work Order ID	Job ID	Job Performance	Job Progress		Output		Rejection	Time (mins)				Job Status	End Customer
					Start Time	End Time	Planned	Actual		Setup	Pred	Downtime	Idle		
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i

iii. LoRaWAN based Solution

UCT is one of the early adopters of LoRAWAN technology and providing solution in Agri-tech, Smart cities, Industrial Monitoring, Smart Street Light, Smart Water/ Gas/ Electricity metering solutions etc.

iv. Predictive Maintenance

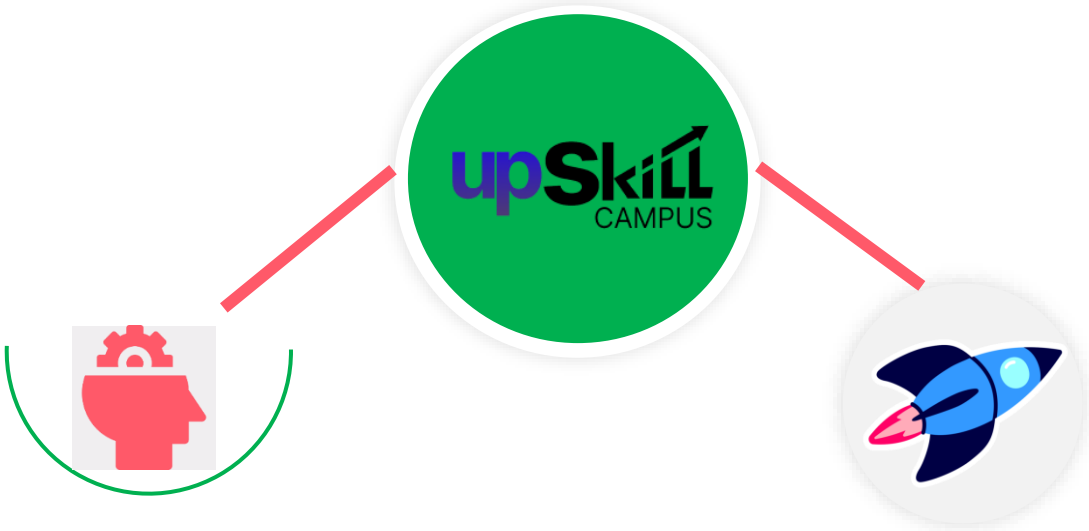
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



2.2 About upskill Campus (USC)

upskill Campus along with The IoT Academy and in association with Uniconverge technologies has facilitated the smooth execution of the complete internship process.

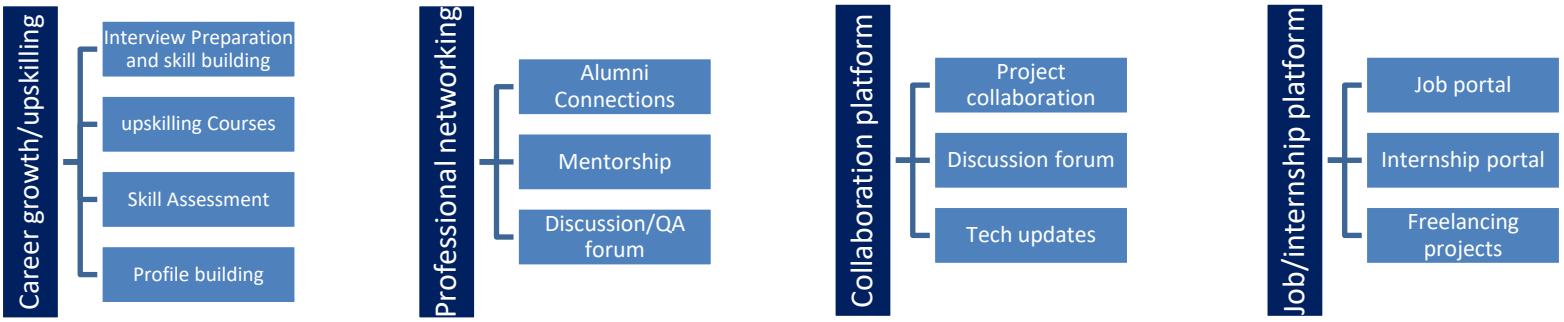
USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.



Seeing need of upskilling in self paced manner along-with additional support services e.g. Internship, projects, interaction with Industry experts, Career growth Services

upSkill Campus aiming to upskill 1 million learners in next 5 year

https://www.upskillcampus.com/



2.3 The IoT Academy

The IoT academy is EdTech Division of UCT that is running long executive certification programs in collaboration with EICT Academy, IITK, IITR and IITG in multiple domains.

2.4 Objectives of this Internship program

The objective for this internship program was to

- ☛ get practical experience of working in the industry.
- ☛ to solve real world problems.
- ☛ to have improved job prospects.
- ☛ to have Improved understanding of our field and its applications.
- ☛ to have Personal growth like better communication and problem solving.

2.5 Reference

[1] Introducing Data Science Book (By Davy Cielen , Arno D.B. Meysman ,Mohamed Ali)

[2] Introduction to Machine Learning Book (Alex Smola and S.V.N. Vishwanathan)

[3] An Introduction to Probability and Statistics (WILEY Series)

3 Problem Statement

The government has provided two datasets: a training dataset and a test dataset. The training dataset contains historical data from November 1, 2015, to June 30, 2017, and includes information such as the date and time of traffic recordings, the junction number (1, 2, 3, or 4), the number of vehicles, and an ID for each entry. The test dataset covers the period from July 1, 2017, to October 31, 2017, and has a similar structure to the training dataset.

As a data scientist, my task is to utilize time series forecasting techniques to predict the traffic patterns at each of the four junctions for the next four months. This prediction will help the government better understand the traffic dynamics and plan infrastructure accordingly. It is important to consider variations in traffic patterns on different occasions throughout the year, including holidays and other special events.

By analyzing the historical data and identifying patterns and trends, I will develop a predictive model that can forecast the traffic volume at each junction for the given time frame. This information will assist the government in making informed decisions regarding traffic management, optimizing road networks, and improving overall transportation efficiency in the city.

4 Existing and Proposed solution

Existing solutions in traffic management have made significant contributions to understanding and addressing traffic-related issues. These solutions often involve the analysis of traffic flow data, the development of predictive models, and the integration of infrastructure planning techniques. However, when it comes to predicting traffic patterns at four specific junctions and accounting for variations during holidays and special occasions, there are certain limitations to be considered.

1. Lack of granularity: Many existing solutions provide aggregated traffic predictions for the entire city or specific regions. However, they may not offer a detailed understanding of traffic patterns at individual junctions, which is essential for effective management.

2. Limited consideration of temporal dynamics: While some solutions consider temporal patterns, such as daily and weekly variations, they may not adequately capture the complex dynamics associated with holidays and special occasions. These variations can significantly impact traffic flow and require a more nuanced approach.

3. Inadequate incorporation of external factors: External factors such as weather conditions, major events, and road construction play a crucial role in influencing traffic patterns. Existing solutions may not fully account for these factors, leading to suboptimal predictions.

To address these limitations and provide more accurate and actionable insights, my proposed solution involves the use of a GRU (Gated Recurrent Unit) model for traffic pattern prediction. The GRU model is a type of recurrent neural network (RNN) that excels at capturing temporal dependencies and predicting sequences.

4.1 Code submission (Github link):

https://github.com/KanikeSaidatta/UpSkill-Campus

4.2 Report submission (Github link) :

https://github.com/KanikeSaidatta/UpSkill-Campus/blob/main/Smart%20City%20Traffic/ForecastingofSmartcitytrafficpatterns_KanikeSaidatta_USC_UCT.pdf

5 Proposed Design/ Model

we will be exploring the dataset of four junctions and building a model to predict traffic on the same. This could potentially help in solving the traffic congestion problem by providing a better understanding of traffic patterns that will further help in building an infrastructure to eliminate the problem.

Data collection:

Data Collection: Gather historical data on traffic patterns in each of the four junctions. The data should cover a significant time period, including various types of days such as working days, holidays, weekends, and special occasions. The dataset should include relevant information such as date, time, traffic volume, and any other factors that may influence traffic (e.g., weather conditions, events in the city, roadwork, etc.).

About the Data

This dataset is a compilation of hourly counts of automobiles at four intersections. There are four features in the CSV file:

- **DateTime**
- **Junctions**
- **Vehicles**
- **ID**

The traffic data comes from several time periods since the sensors on each of these intersections were gathering data at different times. Data from several of the intersections were scarce or restricted.

Data Cleaning:

Data cleaning, also known as data cleansing or data preprocessing, is a crucial step in the data science pipeline that involves identifying and correcting or removing errors, inconsistencies, and inaccuracies in the data to improve its quality and usability. Data cleaning is essential because

raw data is often noisy, incomplete, and inconsistent, which can negatively impact the accuracy and reliability of the insights derived from it.

- Clean and preprocess the collected data to handle missing values, outliers, and ensure data quality.
- Perform data normalization or scaling to bring the features to a consistent scale.
- Explore the data to identify any patterns, trends, or anomalies.

Feature Engineering:

Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictive model using machine learning or statistical modeling. The goal of feature engineering and selection is to improve the performance of machine learning (ML) algorithms.

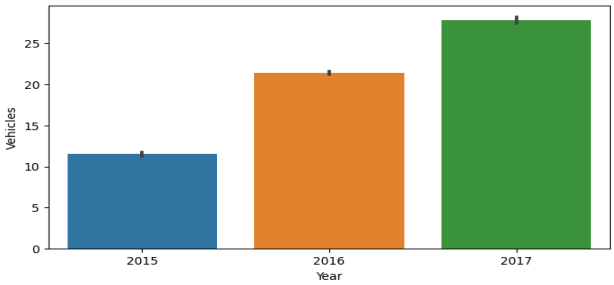
At this stage, We are using DateTime to build a few additional functionalities. Namely:

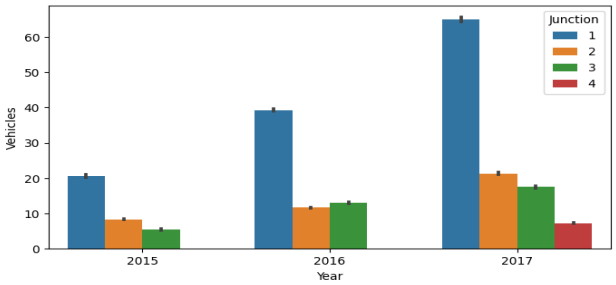
- Year
- Month
- Date in the given month
- Days of week
- Hour

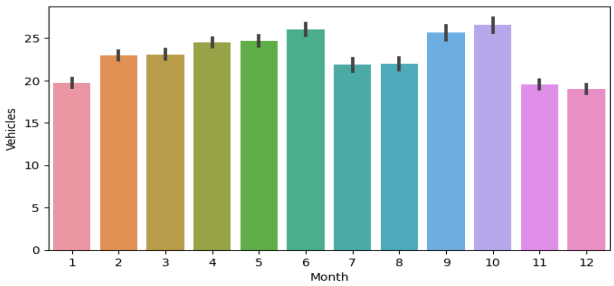
Data Visualization:

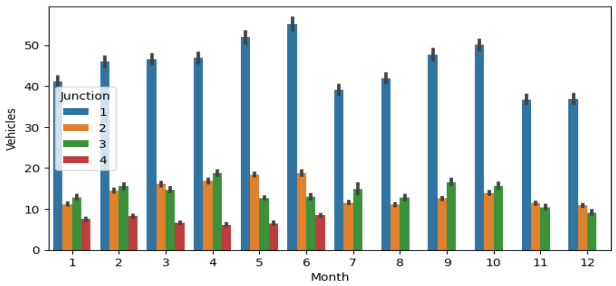
Data visualization in the given project involves using visual elements such as charts, graphs, and maps to represent traffic data and patterns. It aims to visually communicate insights about traffic flow, congestion levels, and variations on holidays and special occasions. Through interactive visualizations, stakeholders can compare traffic patterns, forecast future conditions, and make informed decisions about infrastructure planning and traffic management. Data visualization plays a vital role in presenting complex information in a clear and intuitive manner, enabling better understanding and facilitating effective decision-making.

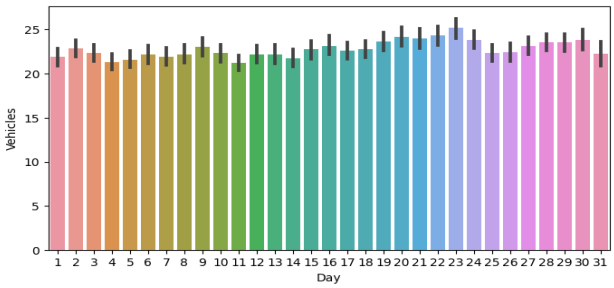
Dataplots of date vs vehicles(overall data) at different junctions

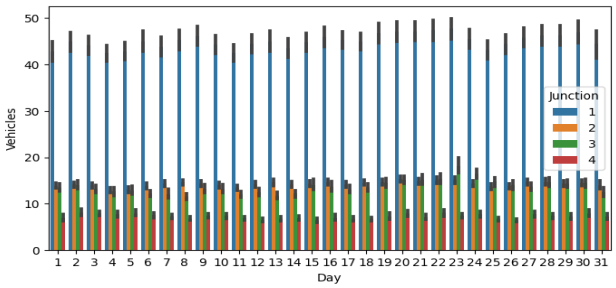


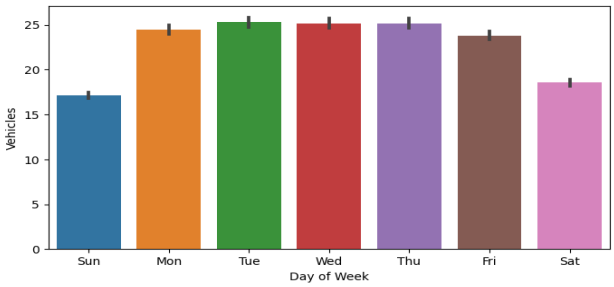


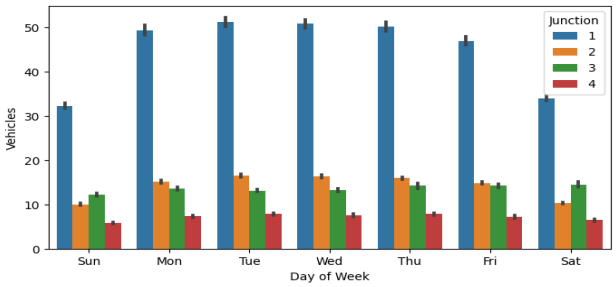


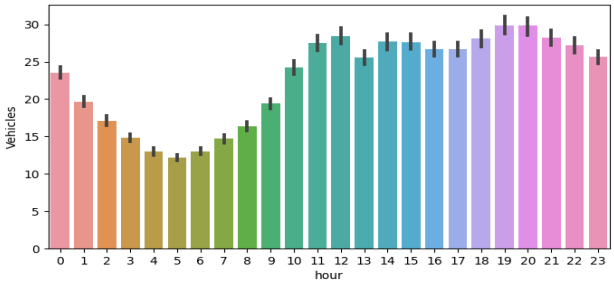


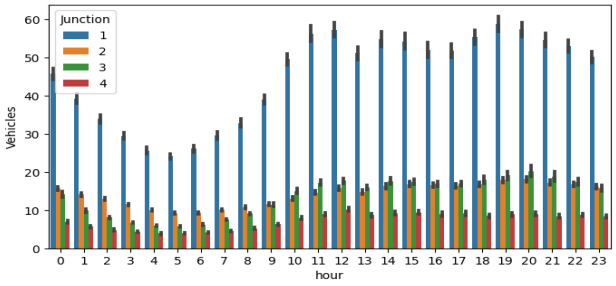


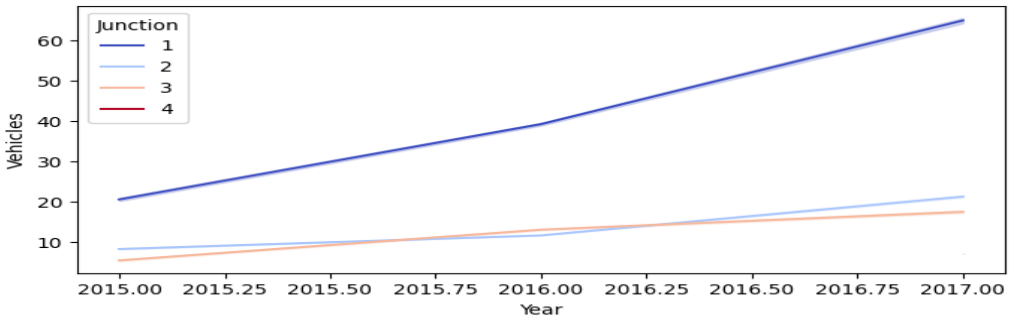


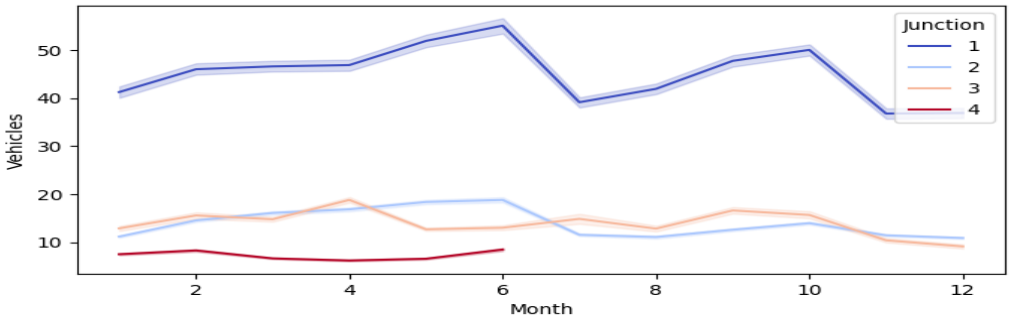


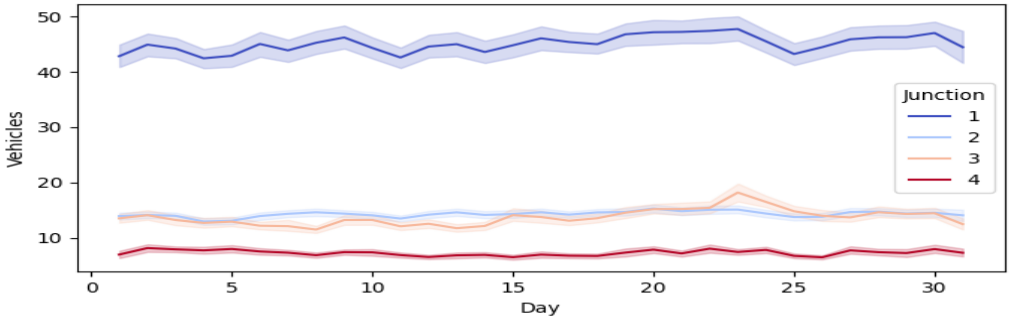


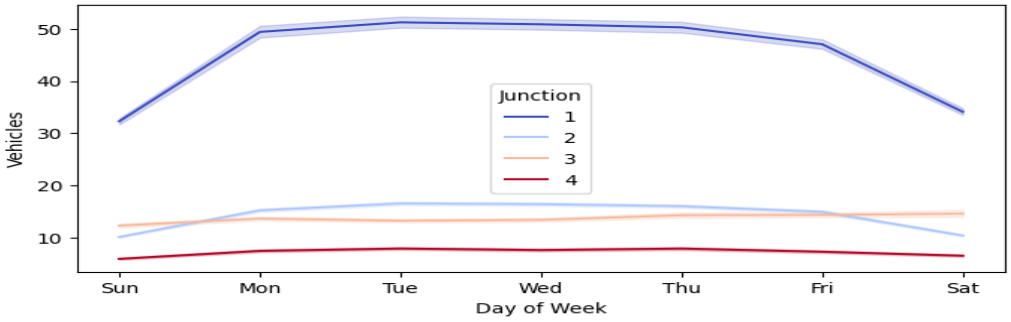


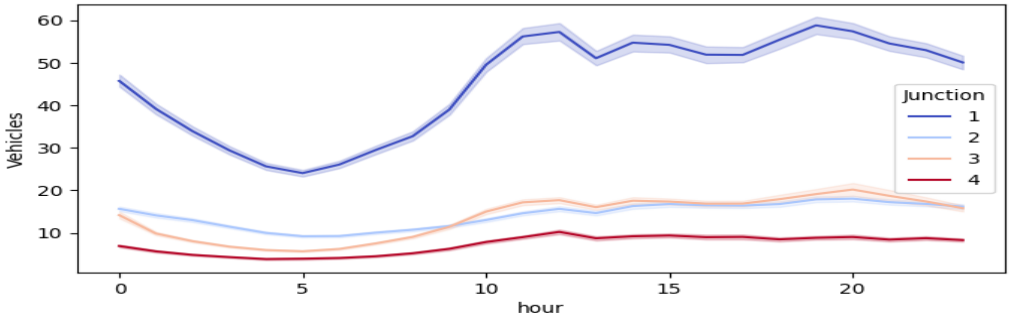



Junction
1
2
3
4
Vehicles
Year
2015.00
2015.25
2015.50
2015.75
2016.00
2016.25
2016.50
2016.75
2017.00


Junction
1
2
3
4
Vehicles
Month


Junction
1
2
3
4
Vehicles
Day


Junction
1
2
3
4
Vehicles
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Day of Week


Junction
1
2
3
4
Vehicles
hour

The plot described above leads to the following conclusions:

With the exception of the fourth junction, all junctions have shown a rising yearly tendency. As was already stated above, the fourth junction contains scant data that doesn't go back more than a year.

We can observe that around June, there is an increase in traffic at the first and second crossroads. This, we assume, may be related to summer vacation and such other related activities.

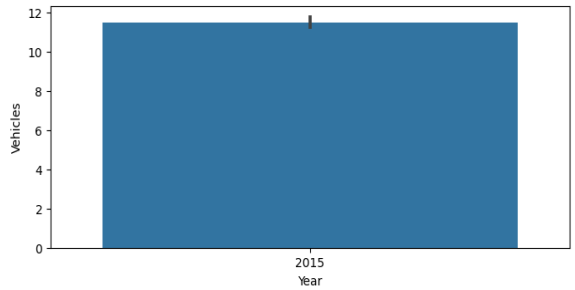
There is considerable consistency in the data on a monthly basis across all dates

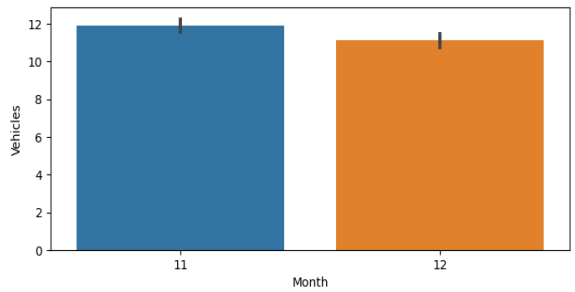
We may observe that there are peaks in the morning and evening and a fall in activity throughout the night for a given day. This is what was predicted.

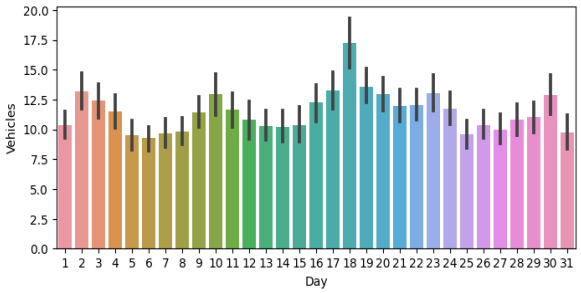
Due to fewer automobiles on the roadways on Sundays than on other days of the week, traffic flows more smoothly. The traffic is consistent from Monday through Friday.

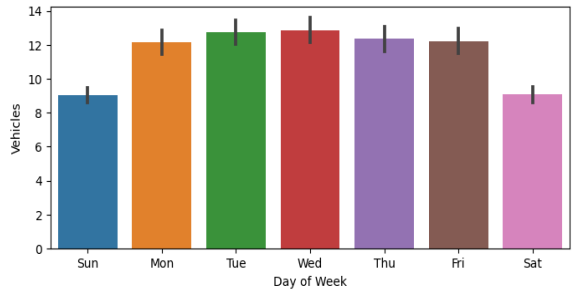
Dataplots of date(yearwise) vs vehicles at different junctions

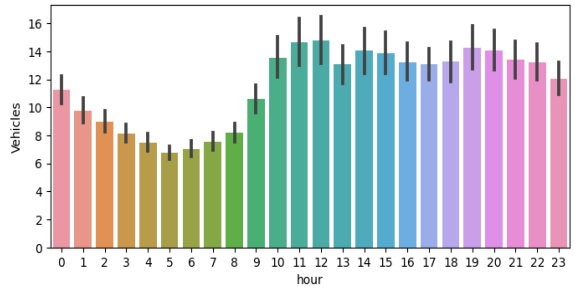
Year(2015)



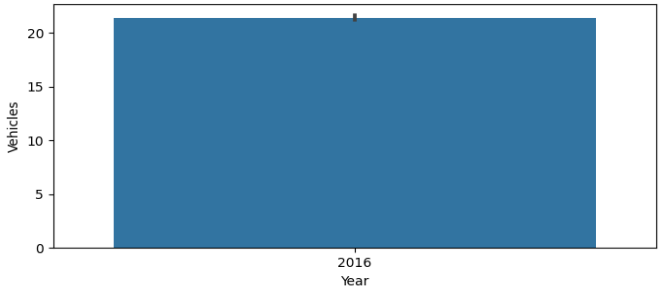


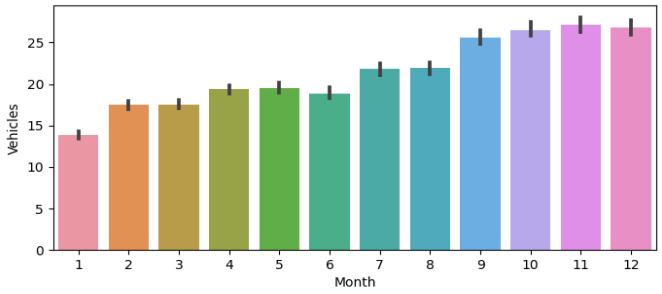


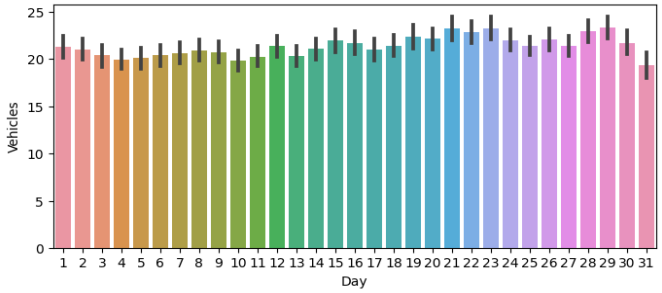


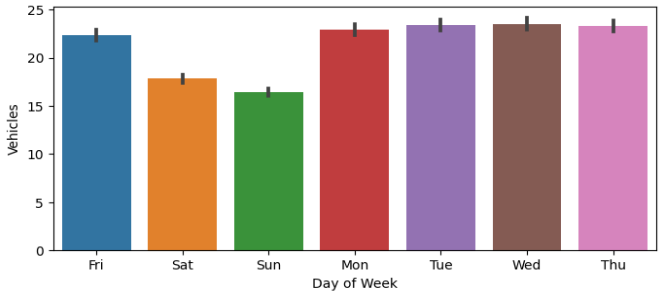


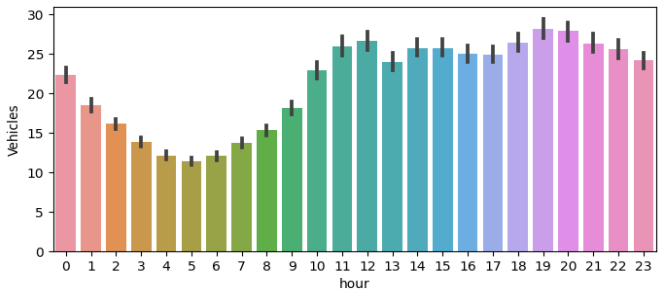
Year(2016)



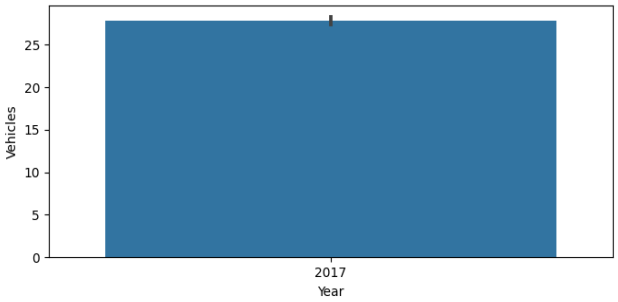


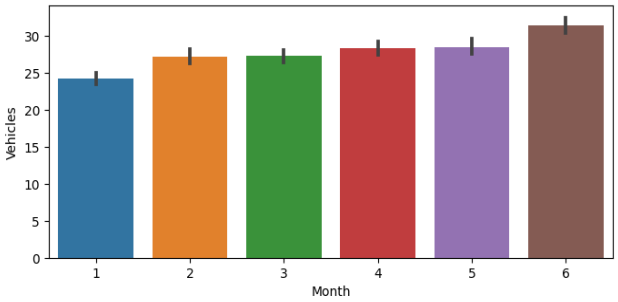


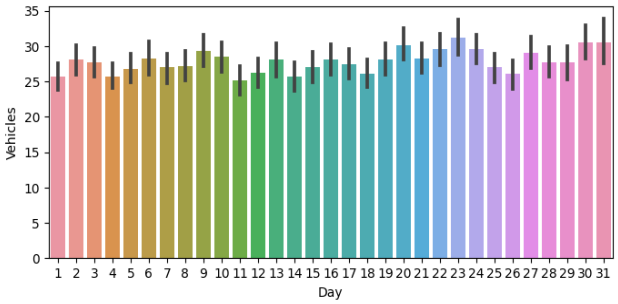


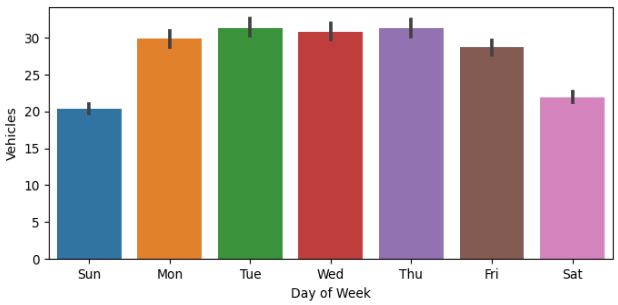


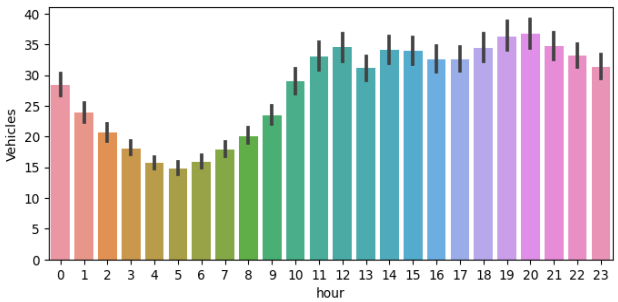
Year (2017)


Vehicles
0
5
10
15
20
25
2017
Year

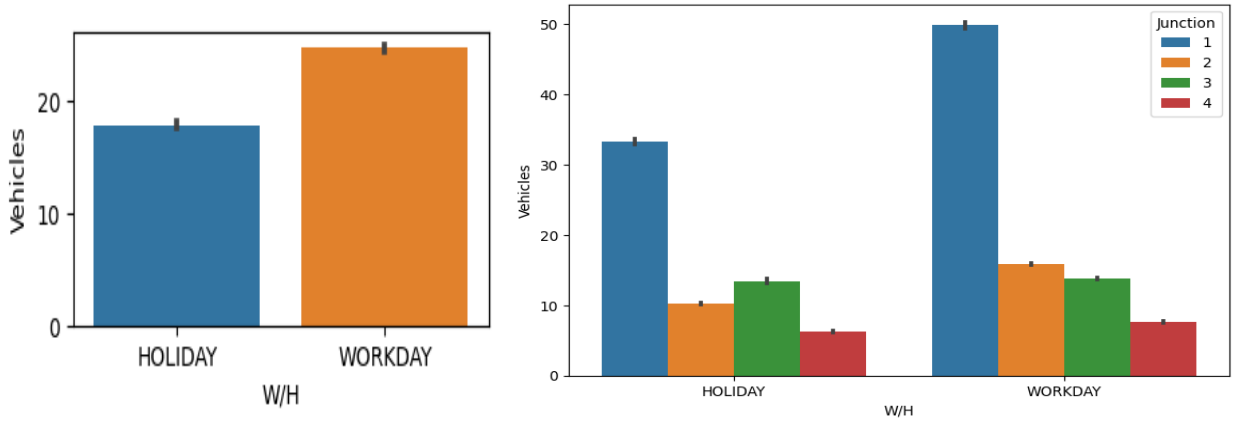

Vehicles
0
5
10
15
20
25
30
1
2
3
4
5
6
Month


Vehicles
0
5
10
15
20
25
30
35
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Day

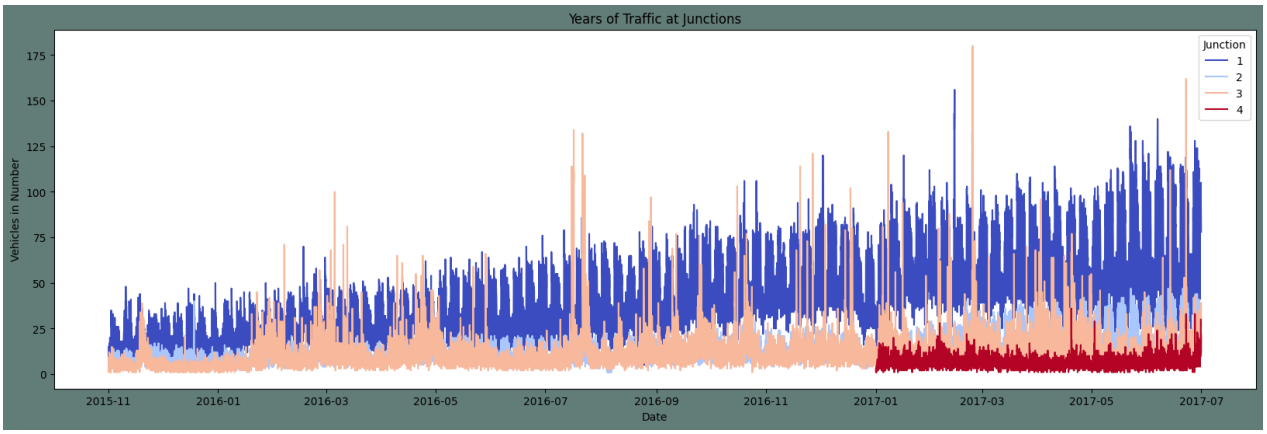

Vehicles
0
5
10
15
20
25
30
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Day of Week


Vehicles
0
5
10
15
20
25
30
35
40
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
hour

Vehicles on working days vs holidays(at each junction):



Timeseries plot:



Observations from data Visualization:

Annual Increase in Traffic Rate: The traffic rate shows a yearly increase, indicating a growth in the number of vehicles on the road over time.

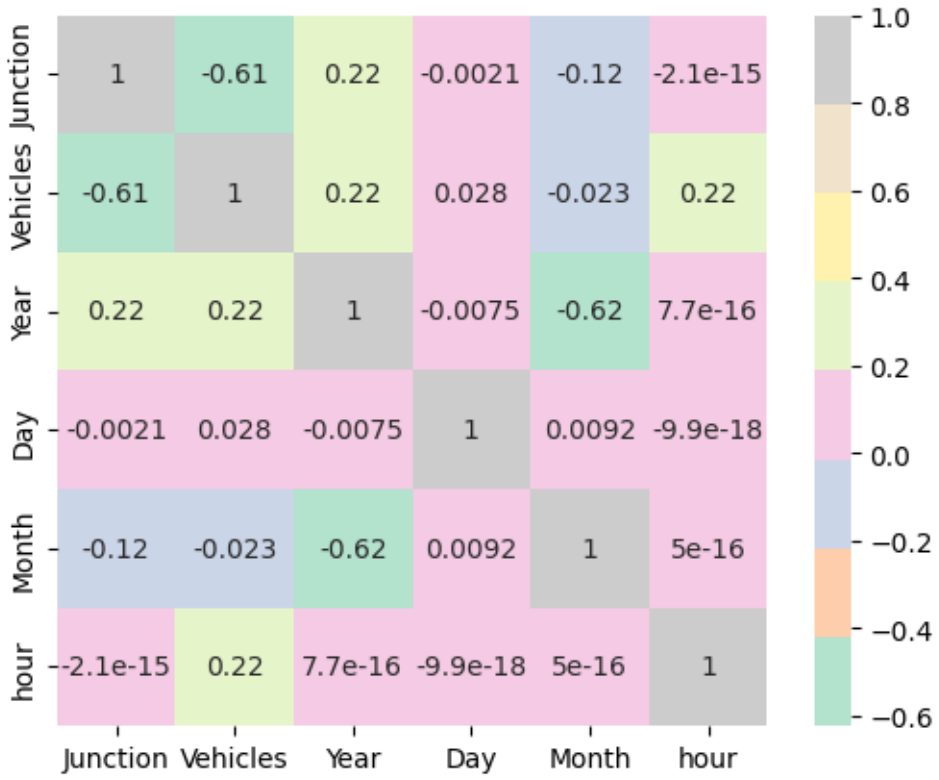
Traffic Rate During Holidays: On holidays, the traffic rate is slightly lower compared to working days. This suggests that factors such as reduced commuting to workplaces or schools contribute to a decrease in traffic volume.

Peak Traffic Hours: The traffic rate is higher during the hours of 11 to 23 (11:00 AM to 11:00 PM) compared to the remaining hours of the day. This indicates that there is a concentration of traffic during these hours, potentially due to factors such as rush hour, increased social activities, or events happening during that time frame.

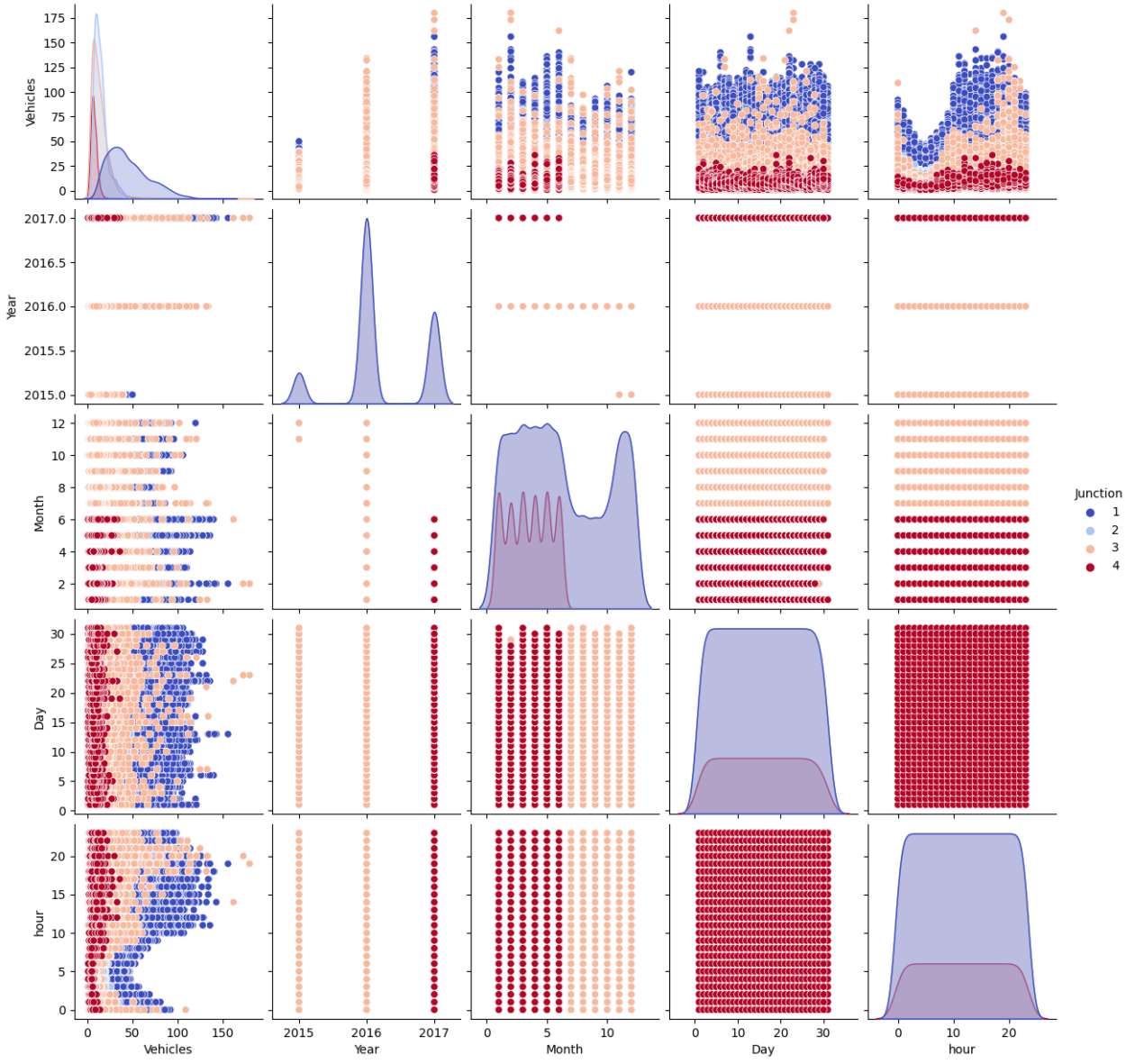
Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is an important step in the data analysis process. In the context of the given project on traffic management, EDA involves examining and visualizing the collected traffic data to gain insights, identify patterns, and understand the characteristics of the data. Here's an overview of the EDA:

Heatmap of data:



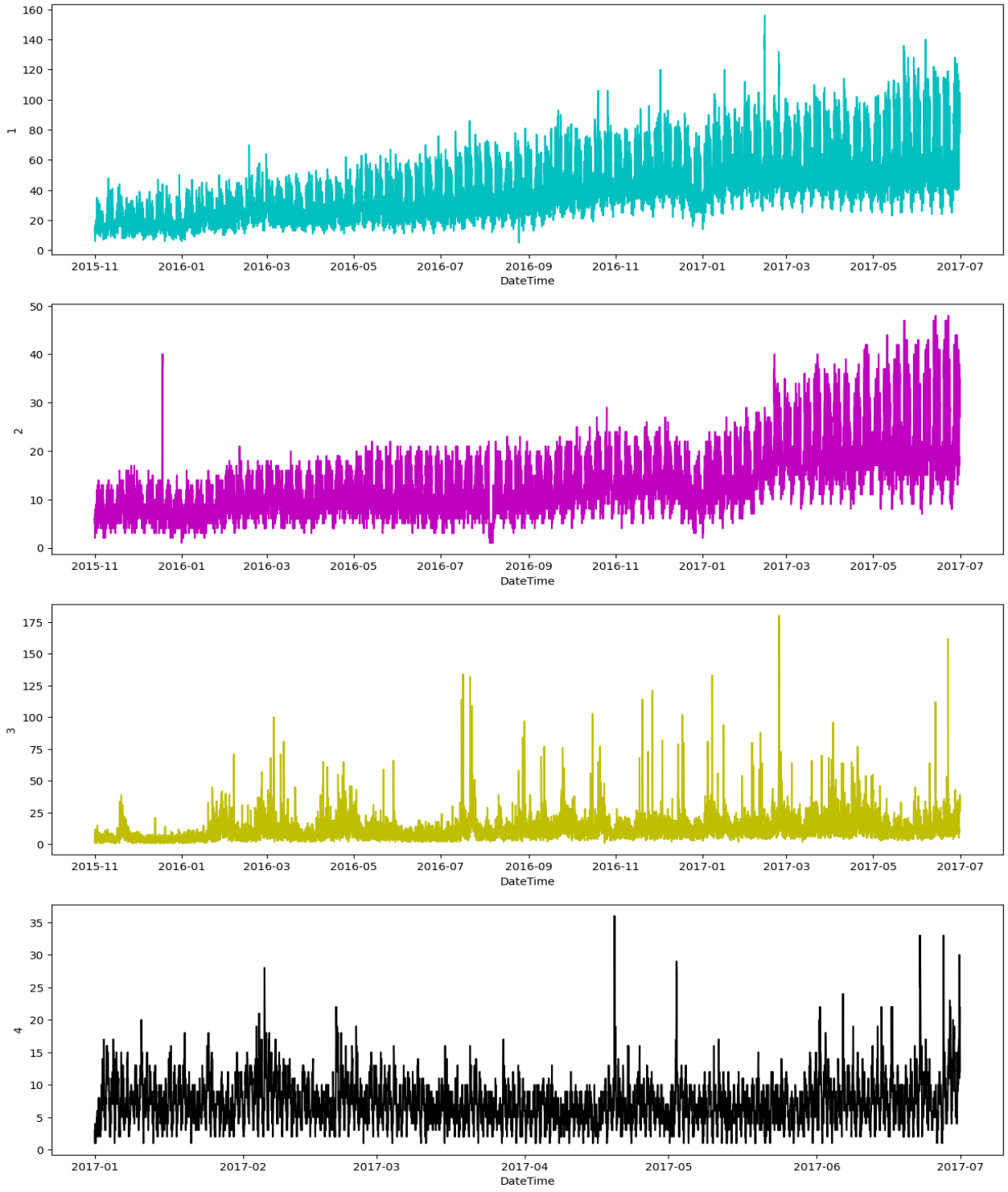
Pairplot of data:



Observations from above:

- Each of the four intersections has a different range of data. Just 2017's data are available for the fourth junction.
- The annual trend for Junctions 1, 2, and 3 has varying slopes.

At each junction, make unique frames



Model selection and evaluation:

Regression models:

Linear Regression**:**

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In the context of traffic management, it predicts traffic flow based on factors like time, Junctions, and day of the week. By analyzing historical data, the model determines the coefficients representing the impact of each factor, aiding in infrastructure planning and efficient traffic management.

Logistic Regression:

Logistic Regression is a statistical modeling technique used to predict binary or categorical outcomes based on a set of independent variables. In the context of traffic management, Logistic Regression can be utilized to predict traffic conditions, such as traffic congestion or the likelihood of accidents, by considering factors like time, weather conditions, Junctions and road characteristics. By analyzing historical data, the model estimates the probabilities of different outcomes, enabling effective decision-making in traffic management strategies.

Random Forest Regression

Random Forest Regression is a machine learning algorithm that combines multiple decision trees to predict continuous numeric values. In the context of traffic management, Random Forest Regression can be employed to predict traffic flow by considering factors like time, weather conditions, and day of the week. By training on historical data, the model learns complex relationships between these factors and traffic flow. It leverages the ensemble of trees to provide accurate predictions and handle non-linear relationships, enabling efficient infrastructure planning and effective traffic management strategies.

To achieve improved results, the evaluation of feature-target relationships in the dataset was conducted. Furthermore, a time series forecasting approach was implemented to enhance the accuracy of predictions."

Time Series Forecasting Approach models:

Random forest classifier:

While Random Forest classifiers are primarily used for classification tasks, they can be adapted for time series forecasting by treating it as a regression problem. By incorporating lagged variables as features and transforming the time series into a supervised learning format, Random Forest classifiers can predict future values. The ensemble model consists of multiple decision trees, each trained on a random subset of features and samples. The final forecast is obtained by aggregating the predictions from individual trees. This approach allows for capturing complex patterns, considering different features' importance, and providing accurate traffic pattern predictions for the city's junctions in a smart city infrastructure.

Decision tree time series:

Decision tree time series forecasting involves utilizing decision tree-based regression models for predicting future values in a time series. By incorporating lagged variables as features and treating the problem as a regression task, decision trees can capture patterns and relationships in the historical data. The decision tree algorithm recursively splits the feature space based on conditions, creating a tree structure. When making predictions, the algorithm traverses the tree from the root node to a leaf node, following the conditions, and provides the forecasted values. This approach allows for interpretability and handling of non-linear relationships, making it useful for understanding and forecasting traffic patterns in smart city infrastructure planning

SVM:

SVM (Support Vector Machine) is a machine learning algorithm that can also be used for time series forecasting. In the context of traffic prediction, SVM can be applied by treating the problem as a regression task. By incorporating lagged variables as features, SVM can capture temporal dependencies and patterns in the historical traffic data. It aims to find a hyperplane that best separates the data points and predicts future traffic patterns based on their characteristics. SVMs can handle non-linear relationships and are effective in capturing complex patterns in the time series data, making them suitable for forecasting traffic patterns in the city's junctions for the next four months.

6 Performance Test

Linear regression:

Junction	MAE	RMSE
1	9.475002580485857	12.09939099947197
2	3.367256282161323	4.430184915879733
3	5.629009482253869	8.749302530348356
4	2.3576551360401847	3.2634036132930206

Logistic Regression:

Junction	MAE	RMSE
1	19.69304556354916	26.183150739842148
2	5.53100376841384	7.879031456381392
3	6.592668722165125	10.515099503785583
4	2.4016110471806673	3.46027929455378

Random forest Regression:

Junction	MAE	RMSE
1	2.852182254196643	3.977159489122666
2	1.7779239465570402	2.248420038320724
3	2.928283658787256	5.565888558936187
4	1.8990678941311852	2.6096021018604585

Accuracy of these time forcasting models:

Model Name	Accuracy Score
Random Forest	20.69201995012469
Decision Tree Classifier	100.0
SVM	8.2356608478803

6.1 Test Plan/ Test Cases

Test Plan:

Objective: Validate the accuracy and reliability of the time series forecasting model for predicting traffic patterns in the city's four junctions over the next four months.

Test Data: Utilize the train dataset for model training and the test dataset for evaluating the model's performance.

Test Environment: Specify the programming language, libraries, and hardware requirements for running the model.

Test Cases:

Test Case 1: Normal Working Days

Test Case 2: Holidays and Occasions

Test Case 3: Traffic Peaks

Test Case 4: Long-Term Forecasting

Test Execution: Define steps for executing each test case, including setup, model execution, and result comparison.

Performance Criteria: Set thresholds for accuracy metrics, such as MAE or RMSE, to determine the model's success.

Test Reporting: Document test results, overall model performance, issues or limitations, and recommendations for improvement.

6.2 Test Procedure

Prepare the test environment by setting up the required programming language, libraries, and hardware specifications for running the forecasting model.

Load the train dataset and perform any necessary preprocessing steps, such as data cleaning, normalization, or feature engineering.

Train the time series forecasting model using the prepared train dataset and appropriate algorithms or techniques.

Split the test dataset into subsets corresponding to different test cases, such as normal working days, holidays, traffic peaks, and long-term forecasting.

For each test case, feed the test dataset subset into the trained model and generate traffic predictions for the specified time period and junctions.

Compare the predicted traffic patterns with the actual traffic patterns available in the test dataset for the corresponding dates and junctions.

Calculate accuracy metrics, such as mean absolute error (MAE), root mean squared error (RMSE), or mean absolute percentage error (MAPE), to evaluate the model's performance for each test case.

Record and document the accuracy metrics and any discrepancies or observations during the comparison process.

Repeat steps 5-8 for each test case in the test plan.

Summarize the overall performance of the forecasting model based on the collected accuracy metrics.

Identify any issues, limitations, or areas for improvement observed during the testing process.

Generate a test report that includes the test results, performance analysis, issues, and recommendations for enhancing the forecasting model.

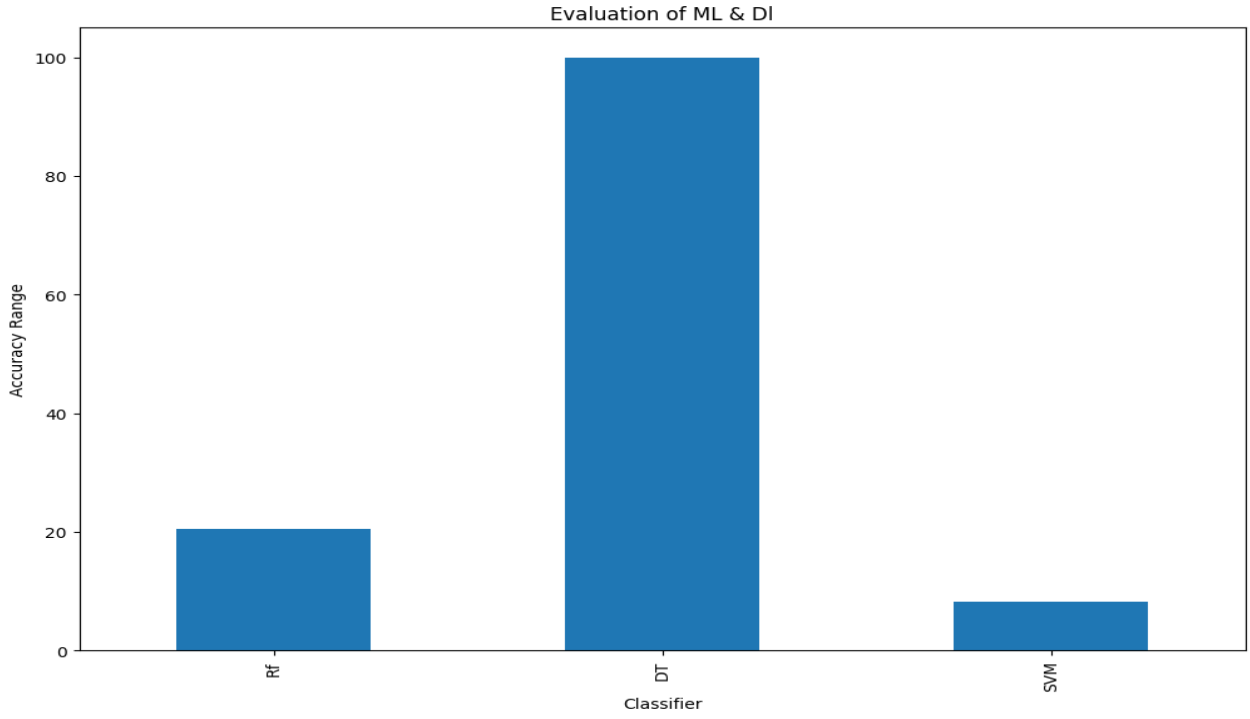
6.3 Performance Outcome

So we have performed the evaulations of both the regression and the time forecasting models.

Among the regression models we oberved that the random forest regressor has the pretty more accuracy than the linear and the logistic regression.

"To achieve improved results, the evaluation of feature-target relationships in the dataset was conducted. Furthermore, a time series forecasting approach was implemented to enhance the accuracy of predictions.

So we had made the timeseries forecasting model evaluation of the different models and we obtained the following accuarcy scores.

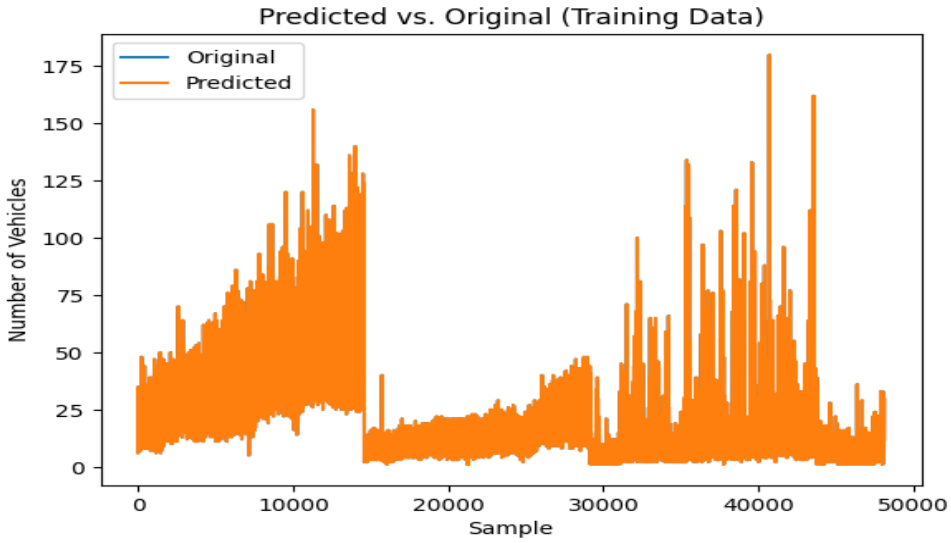


So from the above model evaluation we get the highest accuracy of 100 for the Decision Tree;

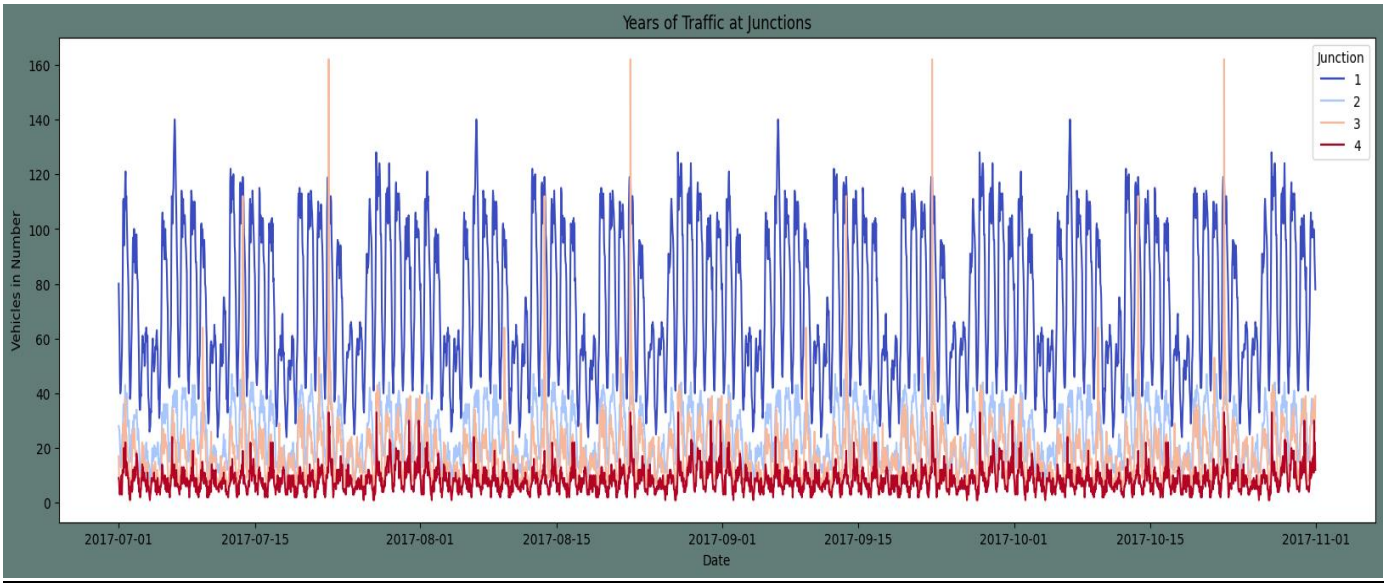
So we are going to predict the vehicles by using the decision tree classifier model.

Predictions using the Decision Tree Classifier:

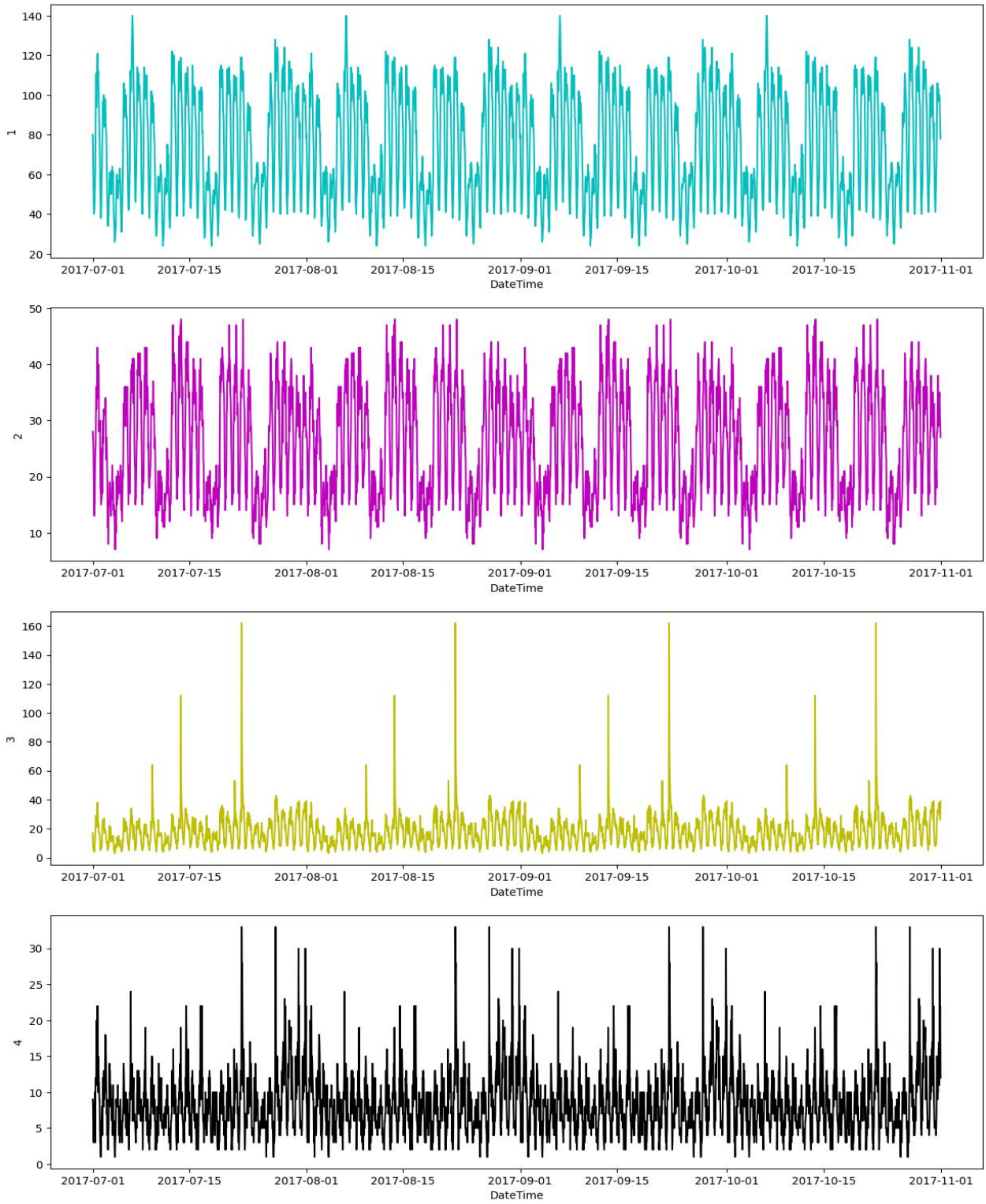
Original vs Predicted Values of training dataset:



Traffic Predictions for next 4 months of given testdataset:



Traffic Predictions at different junction for next 4 months of given testdataset:



7 My learnings

Throughout this project internship, I have learned valuable skills in machine learning and data science that will greatly benefit my career. Here is a summary of what I have learned and how it will help me:

1. Technical Skills: I have gained expertise in machine learning techniques, data preprocessing, feature engineering, and time series forecasting. These skills are essential for solving complex problems using data and will be valuable in future projects.

2. Domain Knowledge: Working on traffic management in a smart city has given me knowledge about urban infrastructure and transportation systems. This knowledge will be useful for projects related to smart cities and transportation planning.

3. Problem-Solving Abilities: I have improved my ability to solve problems by finding creative solutions. This skill will help me tackle challenging data science problems effectively.

4. Collaboration and Communication: I have learned how to work well with others, including government officials and stakeholders. I can now communicate my findings and recommendations in a way that non-technical people can understand. This is important for successful collaboration and presenting technical concepts.

5. Project Management: I have gained experience in managing data-driven projects from start to finish. This includes tasks like data collection, preprocessing, model development, testing, and reporting. These project management skills will be valuable in future projects.

6. Real-World Application: Working on a project with real-world implications has given me practical experience in applying machine learning to solve societal challenges. This experience will be beneficial in my career as it demonstrates my ability to use data science in real-world situations.

Overall, the skills and knowledge I have acquired during this project internship will greatly contribute to my career growth in machine learning and data science. I am now equipped to handle complex data challenges, collaborate effectively, and manage projects successfully. These skills will enable me to make meaningful contributions to data-driven decision-making in various industries and domains.

8 Future work scope

There were some ideas we couldn't explore due to limited time, but they can be considered for future improvements. Here are some areas to focus on:

1. Ensemble Models: Combining multiple forecasting models to improve prediction accuracy.
2. Online Learning: Creating a system that continuously updates the model using real-time data for more up-to-date predictions.
3. IoT Sensors: Using data from sensors placed in the city to gather additional information on traffic flow and road conditions.
4. Multi-City Analysis: Expanding the project to analyze traffic patterns in multiple cities for regional planning.
5. Social Media Data: Incorporating social media data to capture real-time information on events and accidents for better predictions.
6. Predictive Maintenance: Including maintenance schedules and historical data to plan road repairs and infrastructure maintenance proactively.
7. Optimal Route Recommendation: Developing a feature that suggests the best routes for drivers based on predicted traffic patterns.
8. Integration with Navigation Systems: Integrating the forecasting model with existing navigation systems or creating a standalone traffic prediction app.

Exploring these areas will improve the accuracy and reliability of the traffic forecasting system, leading to better traffic management and infrastructure planning in the city.