Module 4

Capstone Project Draft Report

College of Professional Studies, Northeastern University

ALY6140: Python and Analytics Systems Technology

Prof. Vivian Clements Edwin

December 06, 2023

Group 05

Aarushi Sharma

Denish Borad

Shawn Njoroge

Thu Tran

Introduction

This dataset covers all reported dates for theft from motor vehicles as well as associated offenses from 2014 to 2023.

The project intends to analyze the temporal trends, distribution, and hotspots of motor vehicle theft to empower communities and promote crime prevention. The objectives encompass pinpointing high points, comprehending the distribution of theft, predicting patterns, and projecting the probability of an incident. We'll use techniques like regression analysis, clustering, time series analysis, spatial analysis, classification, anomaly detection, and feature importance. The goal of the project is to give policymakers, law enforcement, and community organizations useful information so they can address the underlying causes of theft and make the community safer.

Goals for the dataset:

To better allocate resources during high vehicle theft, it is necessary to analyze temporal trends and pinpoint peak motor vehicle theft periods. This objective entails figuring out the trends in theft incidents over time, such as weekly, monthly, or daily patterns, to advise legislators and law enforcement on the best times and locations to commit theft.

To develop targeted crime prevention strategies, it may be necessary to identify high-frequency crime hotspots and determine whether any particular demographic or location exhibits greater vulnerability. The objective entails identifying particular regions or groups of people who experience auto theft at a disproportionate rate. This will facilitate the creation of focused interventions and strategies for preventing crime that target the root causes and points of vulnerability in these areas.

Questions to Investigate:

1. What are the temporal trends of motor vehicle theft, and can peak periods be identified?
2. How is theft distributed across divisions, neighborhoods, or specific areas?

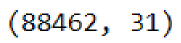
3. Can crime hotspots with high frequency be identified, and do certain demographics or locations show higher vulnerability?
4. Can we forecast future theft trends using time series analysis?
5. How accurately can we predict the likelihood of theft based on various features?
6. What are the key features impacting theft occurrences?
7. How can anomaly detection be leveraged to identify unusual patterns or outliers?

Exploratory Data Analysis

An important step in the data analysis process is exploratory data analysis (EDA), which offers important insights into the variables in the dataset, their distributions, and their interactions. This stage includes activities like data preparation, in-depth research, data cleaning, and visualization, all essential for developing a thorough grasp of the dataset and setting the stage for additional analysis and modeling.

Proposed Machine Learning Models To anticipate future car theft, we plan to train models with the "report date" field and the total number of thefts as inputs. The variables will be forecasted using models such as Logistic Regression, Clustering, and Time Series.

Data Exploration Segment



```
(88462, 31)
```

Figure 1: Number of rows and features in the dataset

As illustrated in Fig. 1, the pertinent dataset consists of 88462 rows and 31 columns. The object data type, numerical variables, and float data type make up most of the dataset's columns, with the remaining columns being the integer data type.

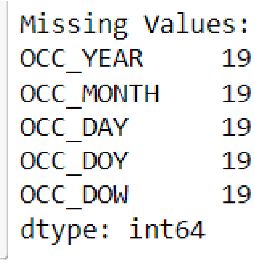
Data Cleaning

To guarantee the accuracy and consistency of the dataset, the data cleaning procedure entails several crucial steps. Missing value handling, data type correction, categorical variable conversion, handling outliers, change verification, duplicate removal, scaling/normalization, and summary statistics computation are some of these steps. Each step is vital to ensure that the data is accurate, consistent, and in the right format for analysis and modeling.

Handling missing values:

Four columns have missing values and the missing information shown in Fig. 2 is already there in another column but in a date form. To overcome this missing value, we simply extract the particular field from the date and assign that value to its relatable place.

It aids in guaranteeing that the dataset is free of inaccurate or faulty data. Here are the missing values because it's related to the date as shown in Fig. 2 and they cannot be replaced with mod or mean.



```
Missing Values:
OCC_YEAR      19
OCC_MONTH     19
OCC_DAY       19
OCC_DOY       19
OCC_DOW       19
dtype: int64
```

Figure 2: Missing Values

Correcting data types:

In this step, specific columns must be converted to the proper data types. For example, date columns must be converted to datetime format using the pd.to_datetime method.

Converting categorical variables:

The "astype" method converts categorical variables to the appropriate data type. For example, the 'LOCATION_TYPE' column is converted to a categorical data type.

Dealing with outliers:

To ensure that extreme values do not unnecessarily influence the analysis, outliers in the dataset are addressed, for instance, by removing outliers in the 'REPORT_HOUR' column using z-scores. Firstly, we calculated z values and stored them in the column then with the use of a box plot and scatter plot analyzed the extremes.

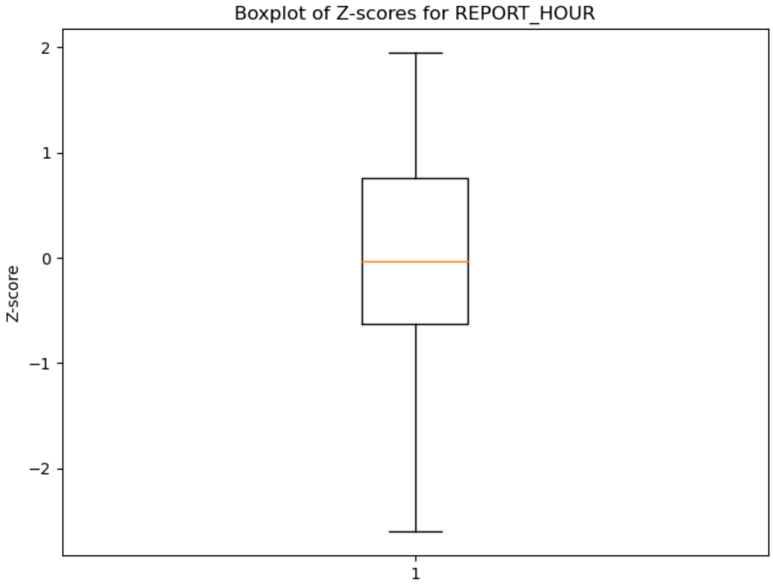


Figure 3: Dealing with outliers with boxplot

The Z-scores minimum, first quartile, median, third quartile, and maximum can all be seen visually with the box plot shown in Figure 3. It makes it possible to spot anomalies and gives information about the distribution and central tendency of the z-scores for the 'REPORT_HOUR' column.

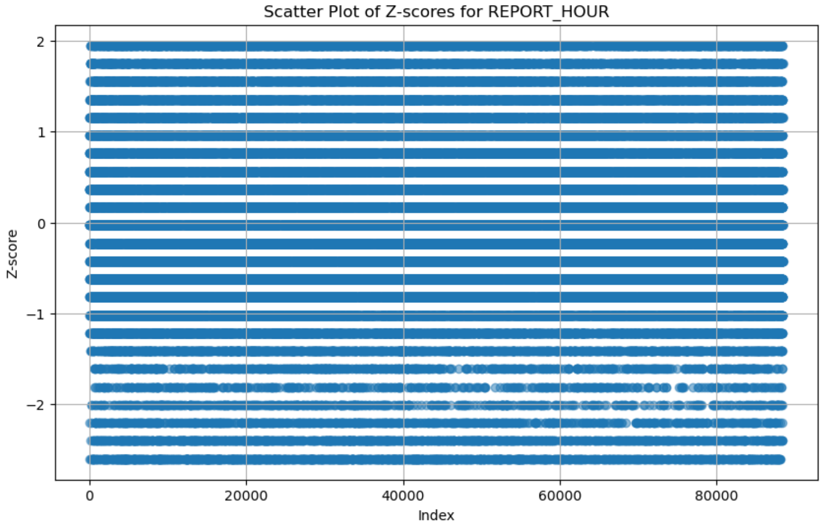


Figure 4: Dealing with outliers with Scatterplot

In Fig. 4, the relationship between the z-scores and the associated report hours is shown graphically. A distinct report hour and the corresponding z-score are represented by each data point on the scatter plot. To identify any trends or outliers in the data, this kind of plot helps examine the distribution and patterns of z-scores about the report hours.

Verifying changes:

Following the cleaning process, the modifications are confirmed by looking through the data's information using "info()" and looking for any missing values.

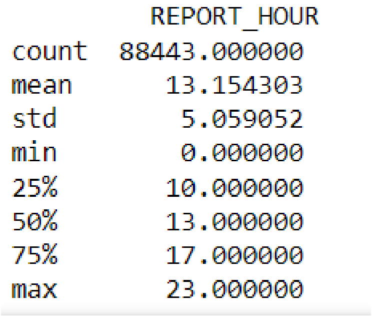
Removing duplicates:

To prevent duplication and guarantee data consistency, duplicate rows are found and eliminated from the dataset. Unfortunately, there is no repeated information in this dataset, so we just check it for data consistency.

Scaling/Normalization:

To bring the values within a given range, some numerical columns are scaled or normalized using methods like Min-Max scaling.

Summary statistics:



```
        REPORT_HOUR
count  88443.000000
mean      13.154303
std        5.059052
min        0.000000
25%       10.000000
50%       13.000000
75%       17.000000
max       23.000000
```

Figure 5: Summary statistics of Report hour

Summary statistics, such as descriptive statistics for the 'REPORT_HOUR' column, are calculated to provide an overview of the cleaned dataset. According to the 'REPORT_HOUR' column summary statistics as shown in Fig. 5, there are 88,443 data points. With a standard deviation of roughly 5.06 and an average report hour of roughly 13.15, these numbers indicate moderate variability. There are two report hours: zero and twenty-three. With 25% of the data falling below 10 and 75% falling below 17, the median report hour is 13.

Data Visualizations

It is evident from Fig. 6 below that there is a little decline in crime incidents as the year comes to an end. This graphic depiction of crime data over time offers important insights into patterns and trends in a particular area's crime rate. After the coronavirus pandemic, the lowest number of incidents occurred in 2017 and the highest in 2020.

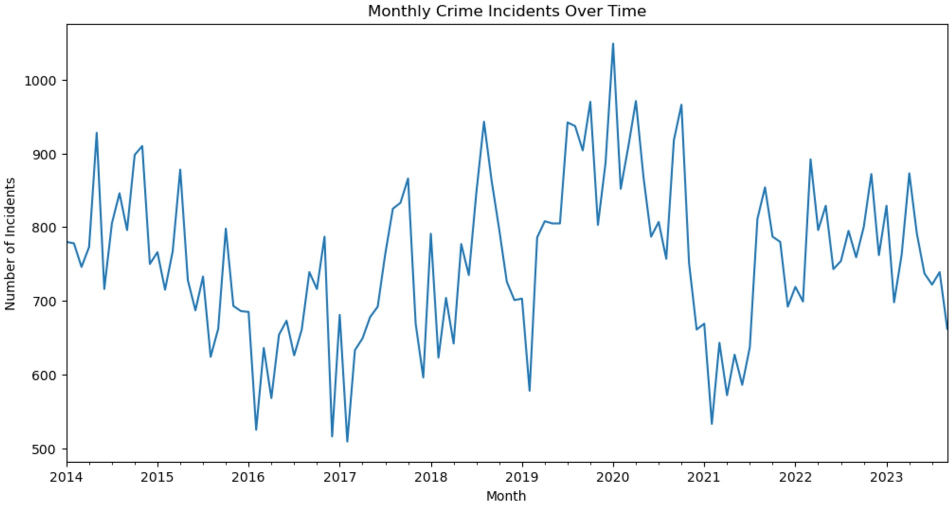


Figure 6: Crime Incidents Over Time

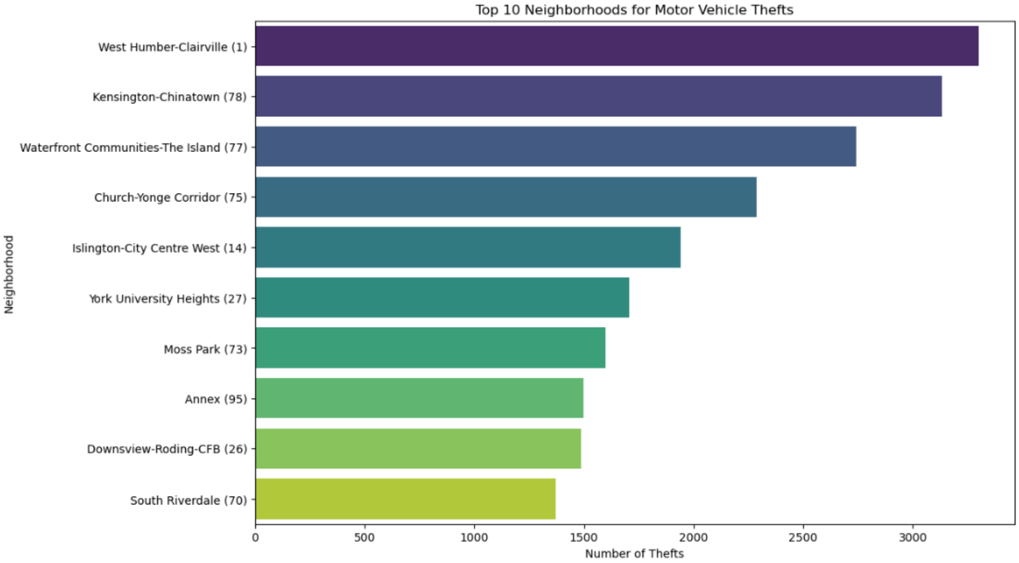


Figure 7: Top 10 Neighbourhoods for Theft

The top ten neighborhoods for auto theft are shown in Fig. 7. It provides a comparison of the frequency of car theft in various locations by displaying the number of thefts in each neighborhood. Additionally, the least and the majority of the vehicles were being stolen from the same location. It can also provide insight into the area that a car thief would like to target to steal vehicles.

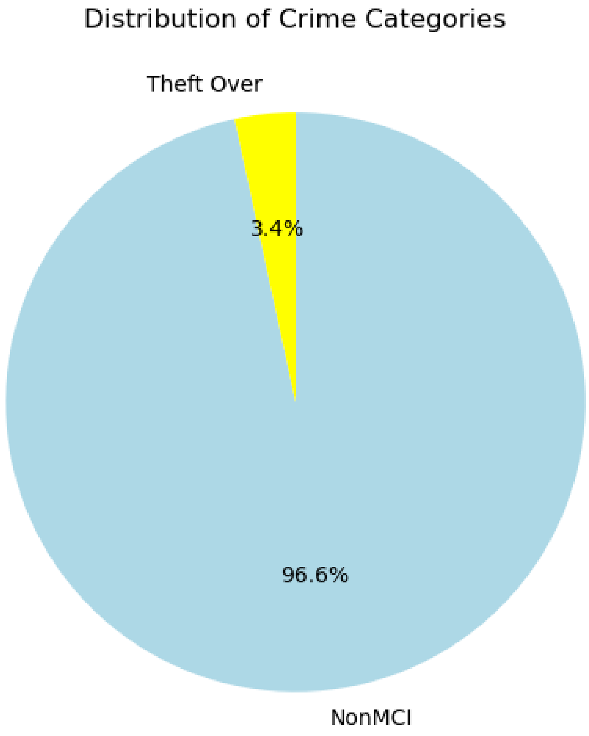


Figure 8: Crime Categories Distribution

Figure 8's pie chart illustrates how theft is distributed among the various categories. It provides insights into the frequency of theft across different kinds of goods or assets by graphically displaying the percentage of theft within each category. As can be observed, nonMCI categories, or "non-major crime incidence," account for nearly 96.6% of all theft incidences. In contrast, "Theft Over" crimes, or those exceeding $5000, represent only 3.4% of all theft incidences, according to one report.

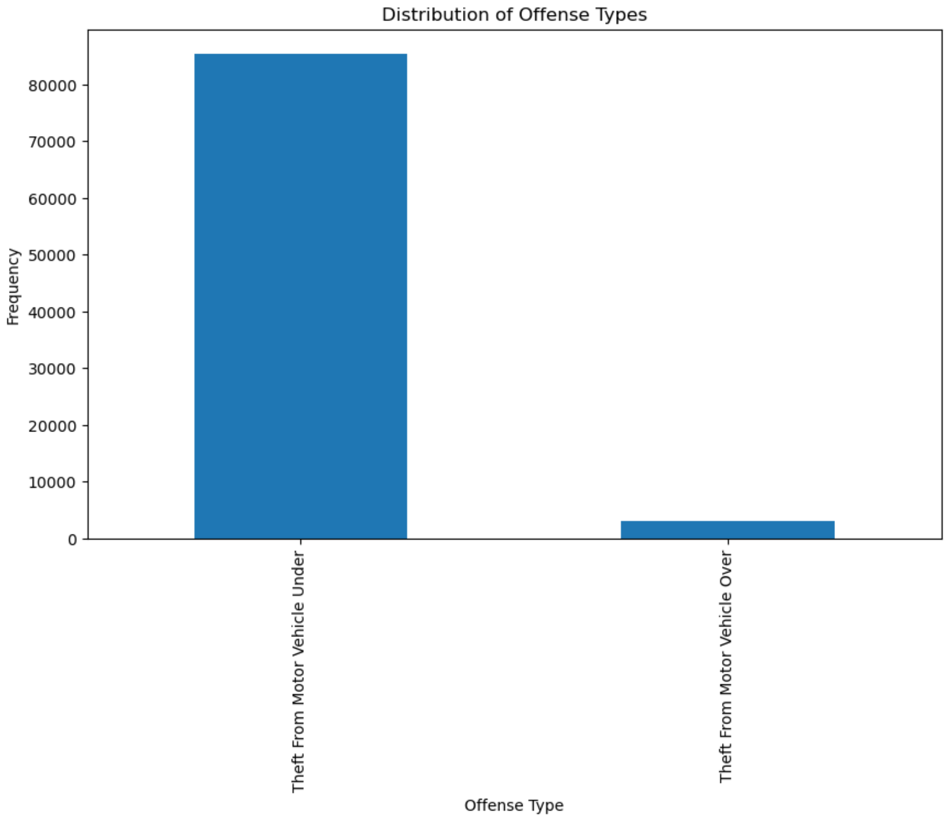


Figure 9: Distribution of Offence Types

The distribution of various offense types is shown in the chart above (Fig. 9), which gives an idea of how common different criminal activities are. It is evident that the category with the highest percentage of theft incidents was theft; however, there is a small amount of data with the other category.

As can be observed from Fig. 10 below, the majority of the cars were stolen from apartment complex parking lots, both commercial and residential, then from houses, houses, and finally from roads and streets. These three locations are the main ones that are ranked from high volume to low, and theft incidences from these locations have almost doubled the national average when compared to other locations that have relatively low levels of theft.

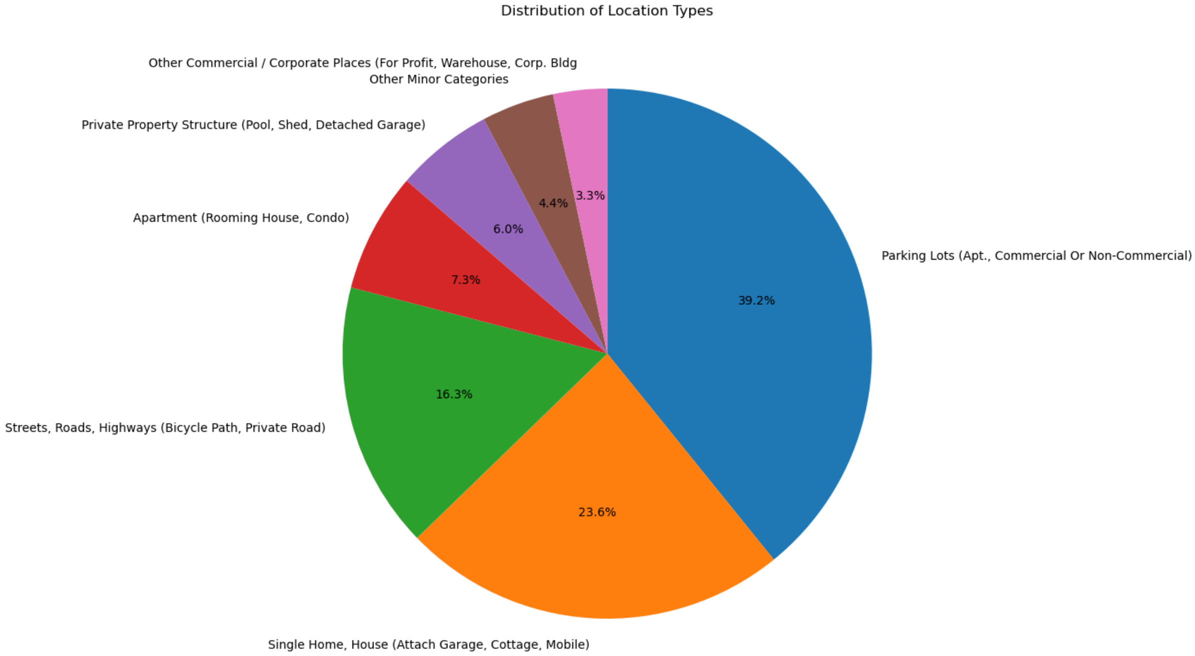


Figure 10: Distribution of Location Types

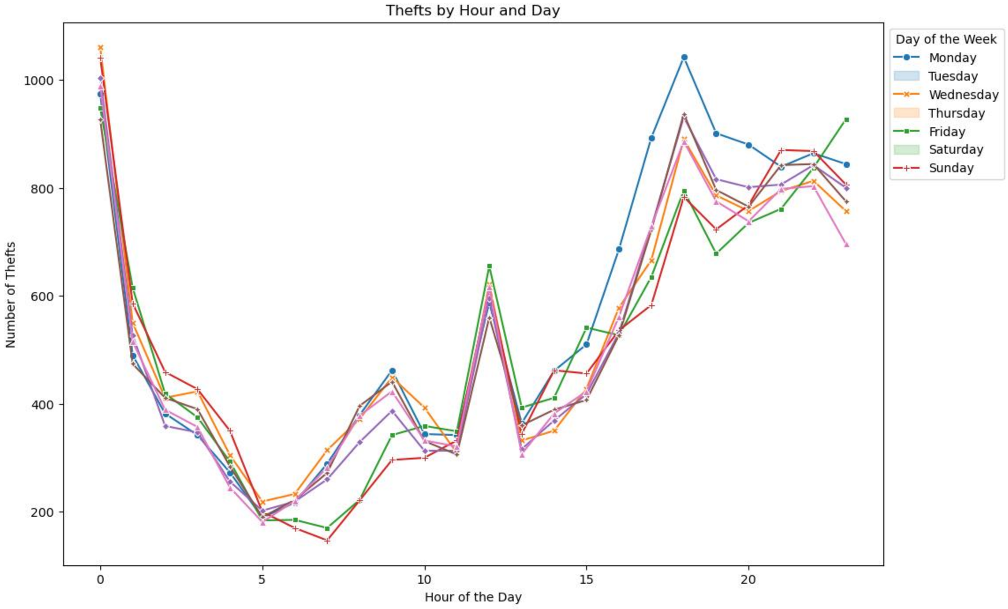


Figure 11: Thefts by Hour and Day

The number of theft incidents that occur in a given day is represented by the chart above in Figure 11. The figure has seven distinct lines, each representing data for a different week. Data indicates that while thefts appear to be distributed uniformly across the day at specific times, there is a higher likelihood of vehicle theft on Mondays around 5:00 PM when compared to other times of the day. Aside from that, the graph indicates a high likelihood of the car being taken overnight.

We divide the 24 hours into four distinct sections—morning, afternoon, evening, and night - to more thoroughly examine this pattern.

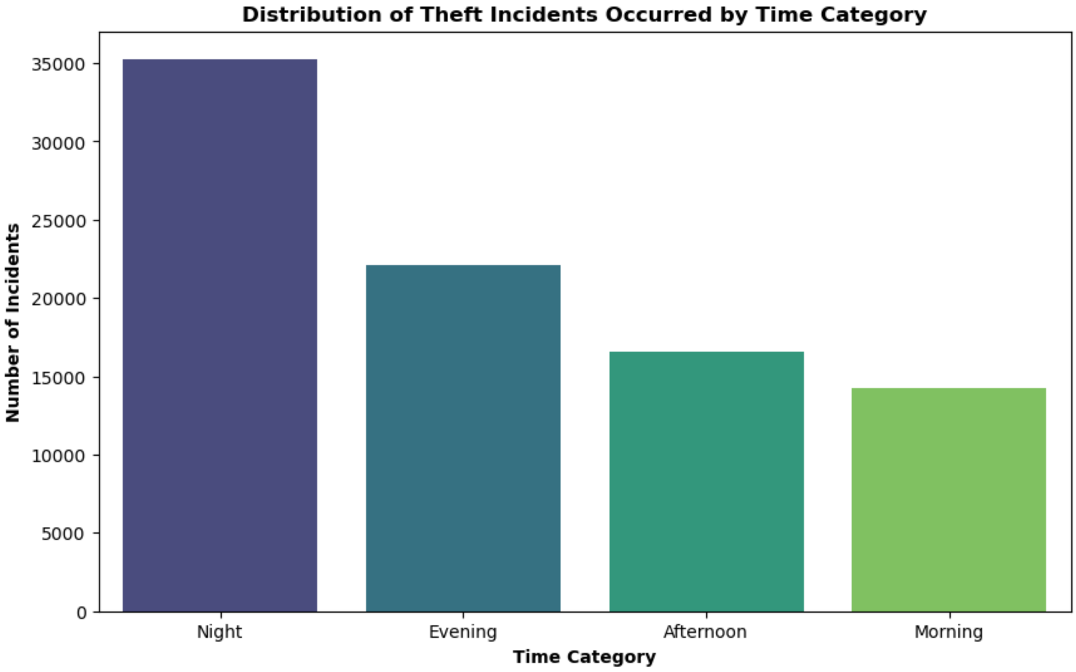


Figure 12: Distribution of Theft Incidents Occurred by Time Category

Based on Figure 12, it can be observed that nighttime is the busiest period and is considered the most convenient time for motor vehicle theft. The afternoon and evening hours follow suit. There's not much chance of the car being stolen in the morning. This knowledge can be useful in identifying trends in criminal activity and inputting specific security measures in place when theft incidents are most likely to happen.

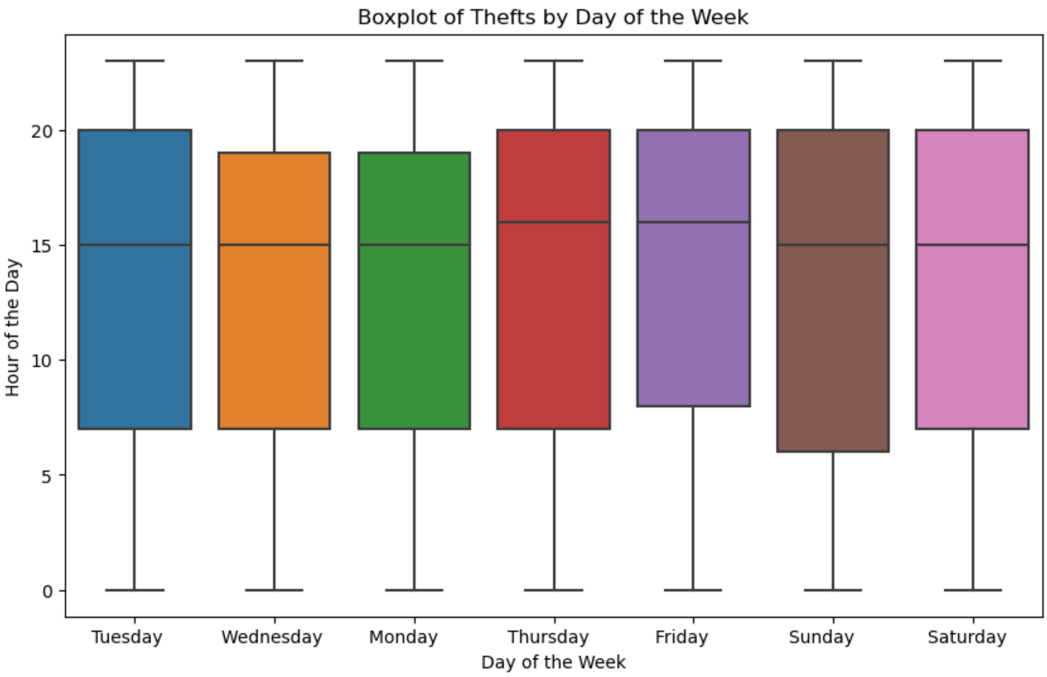


Figure 13: Thefts by Day of the Week

The data distribution across all weekdays to determine which day is riskier for people is shown in the boxplot of Figure 13. It is evident that Sunday carries a higher risk than any other day; Thursday and Saturday also carry a slightly lower risk than Sunday.

To check theft based on hour bases group data by hour and then prepare a line chart to check at which particular hour has the highest risk for theft of motor vehicle.

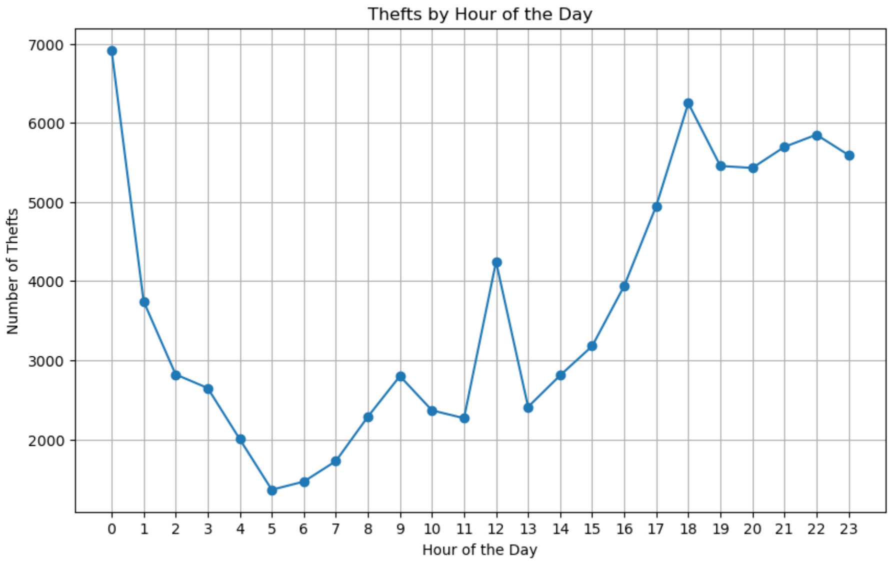


Figure 14: Thefts by Hour of the Day

A distinct pattern in the distribution of theft incidents throughout the day can be seen in the data from Figure 14. The fact that thefts gradually rise throughout the day and peak at 6:00 PM suggests that theft incidents are more likely to happen during specific hours, primarily at night. Knowing this pattern can help you put targeted security measures in place at times when theft incidents are most likely to happen.

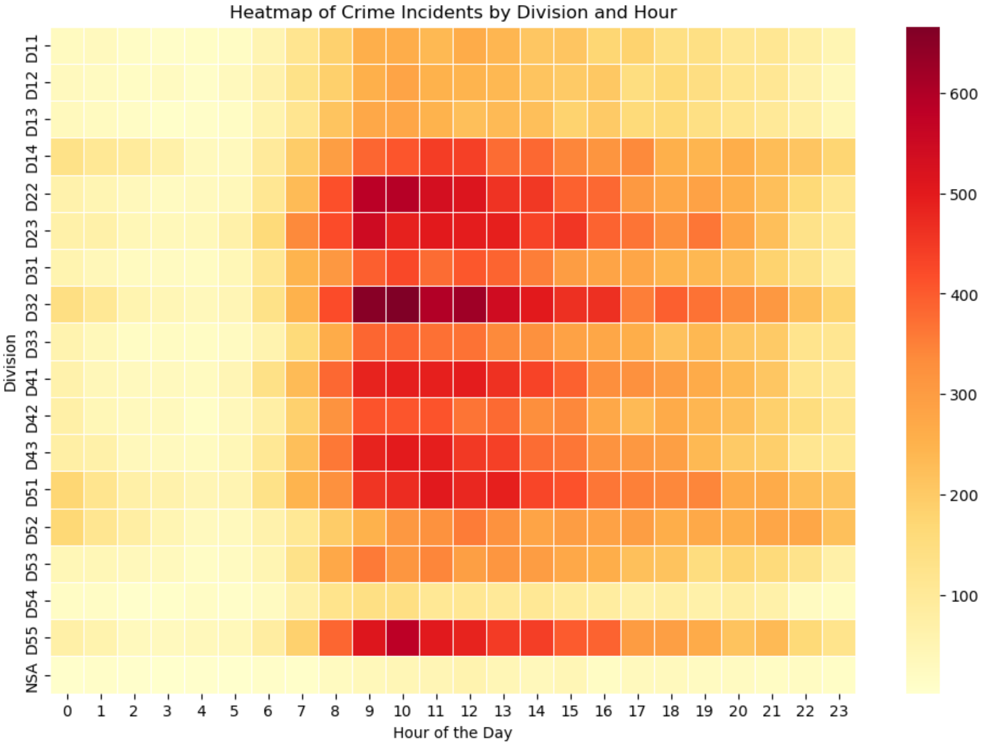


Figure 15: Crime Incidents by Division and Hour

The map above in Figure 15 shows the intensity of the crime incidents as colors, with red denoting a higher concentration of crimes and yellow denoting a lower concentration. In terms of theft, it is evident that divisions D32, D22, and D55 are the most highlighted areas, respectively. When it comes to hourly data, thieves prefer the hours of 8 to 14 when compared to other divisions.

From the above heatmap, it can be seen that the highest thief record we got is in division D32 at 10 AM in the day and the place is most probably the parking plot of apartments predictable as per the above visualizations.

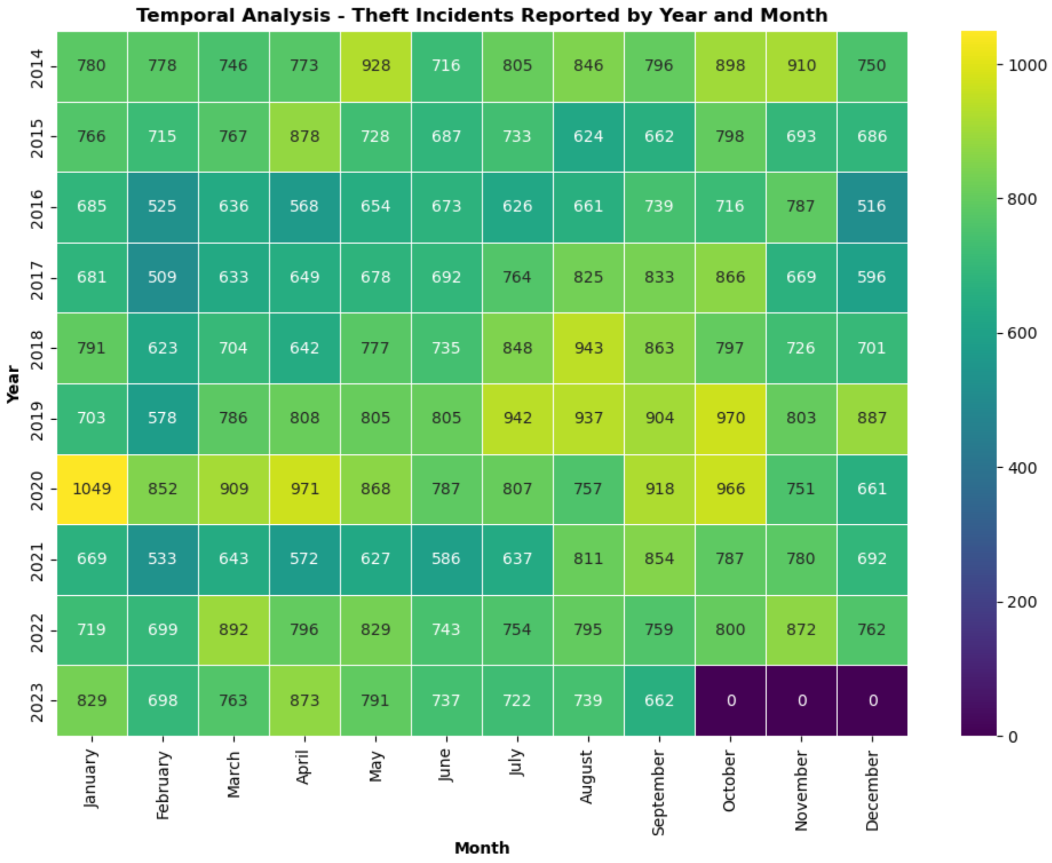


Figure 16: Temporal Analysis

Above figure 16 illustrates the temporal analysis of theft incidents reported by year and month. Each colored square represents a different category or data point, with the grid divided into rows and columns, corresponding to specific months and years.

Trends and patterns in the frequency and distribution of theft incidents reported over time can be found by examining this heatmap. For instance, clusters of a particular color during particular months and years might point to times when theft incidents were more common. By revealing important information about the temporal patterns of theft incidents, this analysis can help identify times when thefts are most likely to happen.

When implementing targeted security measures and allocating resources during periods of increased theft incidents, law enforcement, and community members may find it helpful to understand temporal patterns. Plans to address and lessen theft incidents during particular months and years can also be made with the help of this analysis.

Justification for Models:

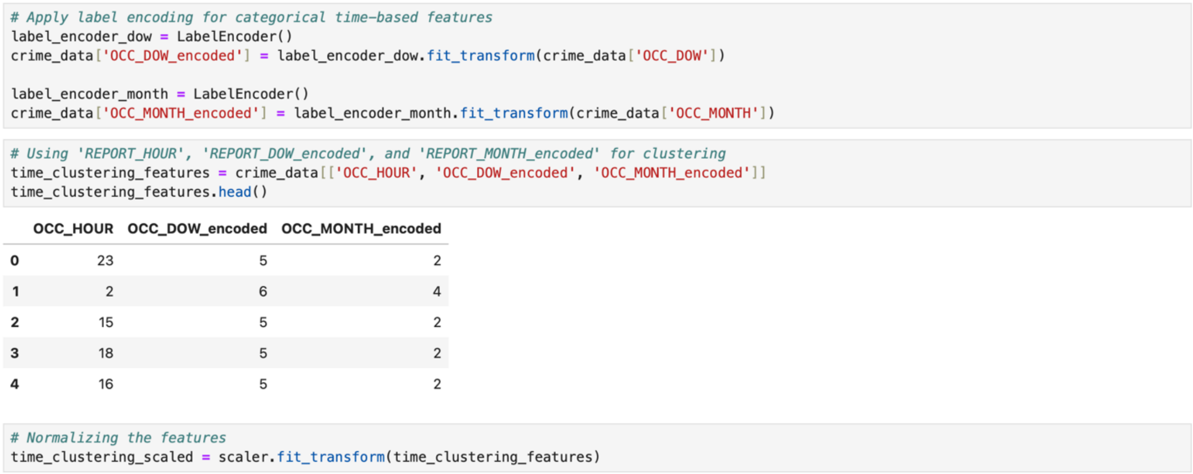
A Random Forest model can be a useful tool for understanding and forecasting theft incidents because of its capacity to manage big datasets and intricate relationships while producing feature importance metrics. Moreover, the use of a time series model makes it possible to recognize seasonality, patterns, and trends in the temporal patterns of theft incidents. This information is useful for forecasting future incidents and allocating resources during periods of high demand. Furthermore, by identifying innate patterns in theft incident data, clustering techniques can help identify hotspots and recurring themes in criminal activity. This information can then be used to inform both law enforcement and community members of targeted interventions and crime prevention strategies.

Clustering Model:

The first model applied in this dataset is clustering. Clustering is a suitable approach, especially in uncovering inherent groupings or patterns in the data without predefined labels. For the thefts from motor vehicles dataset, clustering can help identify naturally occurring clusters based on various attributes like location, time, and other characteristics of the incidents. Particularly, K-Means clustering will be used. It is not only relatively efficient and can handle large datasets well but also can help identify different 'hotspots' or categories of thefts, which could be valuable for resource allocation and preventive measures.

In this case, ‘OCC_DAY’, ‘OCC_DOW’, and ‘OCC_MONTH’ will be used to identify patterns or groups based on the timing of incidents, such as certain hours of the day, days of the week, or months of the year when thefts are most frequent.

First of all, time-based categorical data (like a day of the week or month) need to be converted into a numerical format that is suitable for clustering, then normalize these features as clustering algorithms like K-Means are sensitive to the scale of the data.



```
# Apply label encoding for categorical time-based features
label_encoder_dow = LabelEncoder()
crime_data['OCC_DOW_encoded'] = label_encoder_dow.fit_transform(crime_data['OCC_DOW'])

label_encoder_month = LabelEncoder()
crime_data['OCC_MONTH_encoded'] = label_encoder_month.fit_transform(crime_data['OCC_MONTH'])
```

```
# Using 'REPORT_HOUR', 'REPORT_DOW_encoded', and 'REPORT_MONTH_encoded' for clustering
time_clustering_features = crime_data[['OCC_HOUR', 'OCC_DOW_encoded', 'OCC_MONTH_encoded']]
time_clustering_features.head()
```

	OCC_HOUR	OCC_DOW_encoded	OCC_MONTH_encoded
0	23	5	2
1	2	6	4
2	15	5	2
3	18	5	2
4	16	5	2

```
# Normalizing the features
time_clustering_scaled = scaler.fit_transform(time_clustering_features)
```

Figure 17: Encoded data for the model

Let's proceed with the number of clusters. The Elbow Method helps us determine the optimal number of clusters for time-based clustering. We're looking for a point where the inertia starts to decrease at a slower rate, indicating a suitable number of clusters that balance between too many and too few.

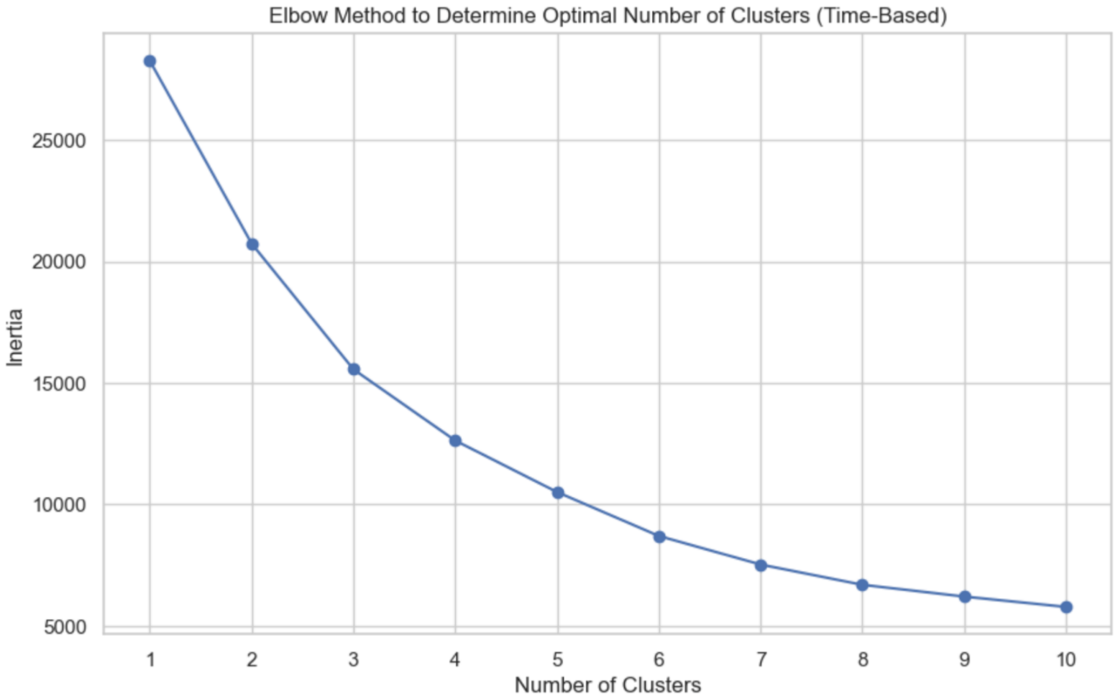
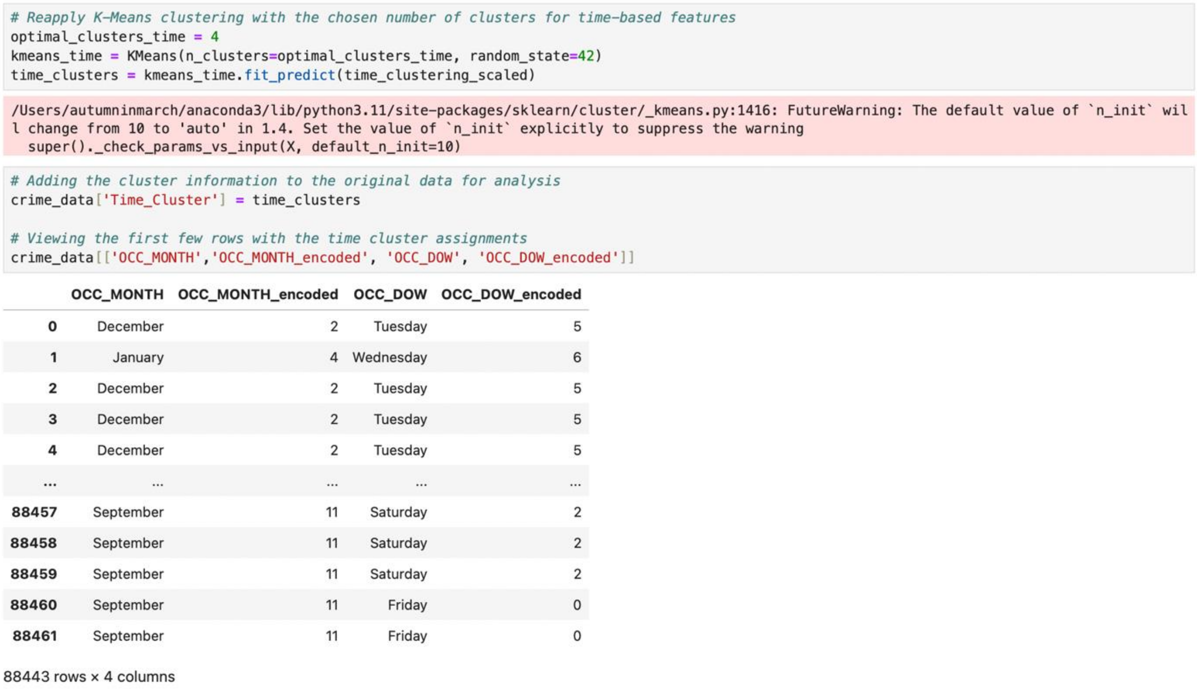


Figure 18: Elbow Method to Determine Optimal Number of Clusters

For demonstration purposes, we will select 4 clusters, which seems like a reasonable choice based on the plot. Let's proceed with training the K-Means model using this number of clusters and then analyze the resulting clusters.



```
# Reapply K-Means clustering with the chosen number of clusters for time-based features
optimal_clusters_time = 4
kmeans_time = KMeans(n_clusters=optimal_clusters_time, random_state=42)
time_clusters = kmeans_time.fit_predict(time_clustering_scaled)
```

```
/Users/autumninmarch/anaconda3/lib/python3.11/site-packages/sklearn/cluster/_kmeans.py:1416: FutureWarning: The default value of `n_init` wil
l change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to suppress the warning
  super()._check_params_vs_input(X, default_n_init=10)
```

```
# Adding the cluster information to the original data for analysis
crime_data['Time_Cluster'] = time_clusters

# Viewing the first few rows with the time cluster assignments
crime_data[['OCC_MONTH','OCC_MONTH_encoded', 'OCC_DOW', 'OCC_DOW_encoded']]
```

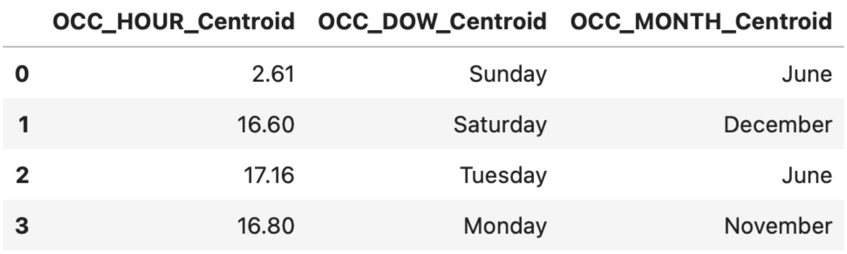
	OCC_MONTH	OCC_MONTH_encoded	OCC_DOW	OCC_DOW_encoded
0	December	2	Tuesday	5
1	January	4	Wednesday	6
2	December	2	Tuesday	5
3	December	2	Tuesday	5
4	December	2	Tuesday	5
...	...	...	...	...
88457	September	11	Saturday	2
88458	September	11	Saturday	2
88459	September	11	Saturday	2
88460	September	11	Friday	0
88461	September	11	Friday	0

88443 rows × 4 columns

Figure 19: K-means Clustering Result

The time-based clustering model using K-Means has successfully assigned each theft incident to one of four clusters (Cluster 0, 1, 2, or 3), based on the hour of the day, day of the week, and month of the year.

To interpret these clusters, we can examine cluster centroids and cluster distribution. Cluster centroids analyze the central values of each cluster and can help understand the typical time patterns (like the most common hours, days, or months) for each cluster. Cluster distribution explores the distribution of incidents within each cluster and how they differ in terms of time.



	OCC_HOUR_Centroid	OCC_DOW_Centroid	OCC_MONTH_Centroid
0	2.61	Sunday	June
1	16.60	Saturday	December
2	17.16	Tuesday	June
3	16.80	Monday	November

Figure 20: Cluster Centroids and Distribution

From the table above, it's clear that Clusters 1, 2, and 3 are centered around late afternoon hours but differ in their most common days and months. Cluster 0 indicates a night timing pattern, predominantly in the weekend and summer months. All four clusters are characterized by weekends and the first two days of the week.

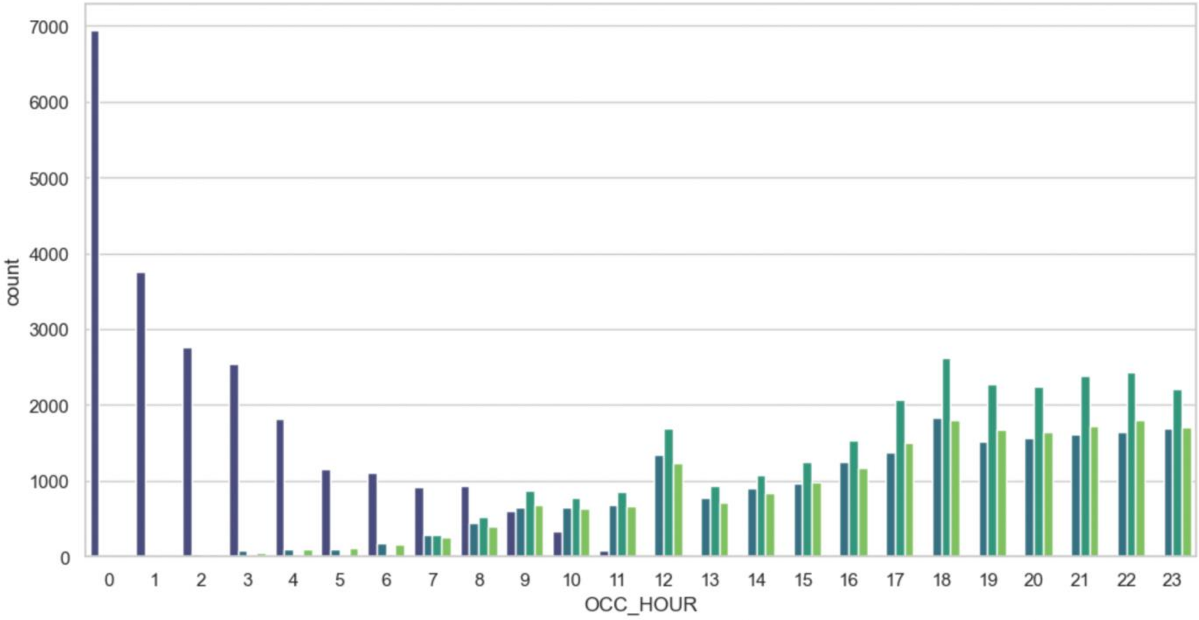


Figure 21: Cluster Model - OCC_HOUR vs count

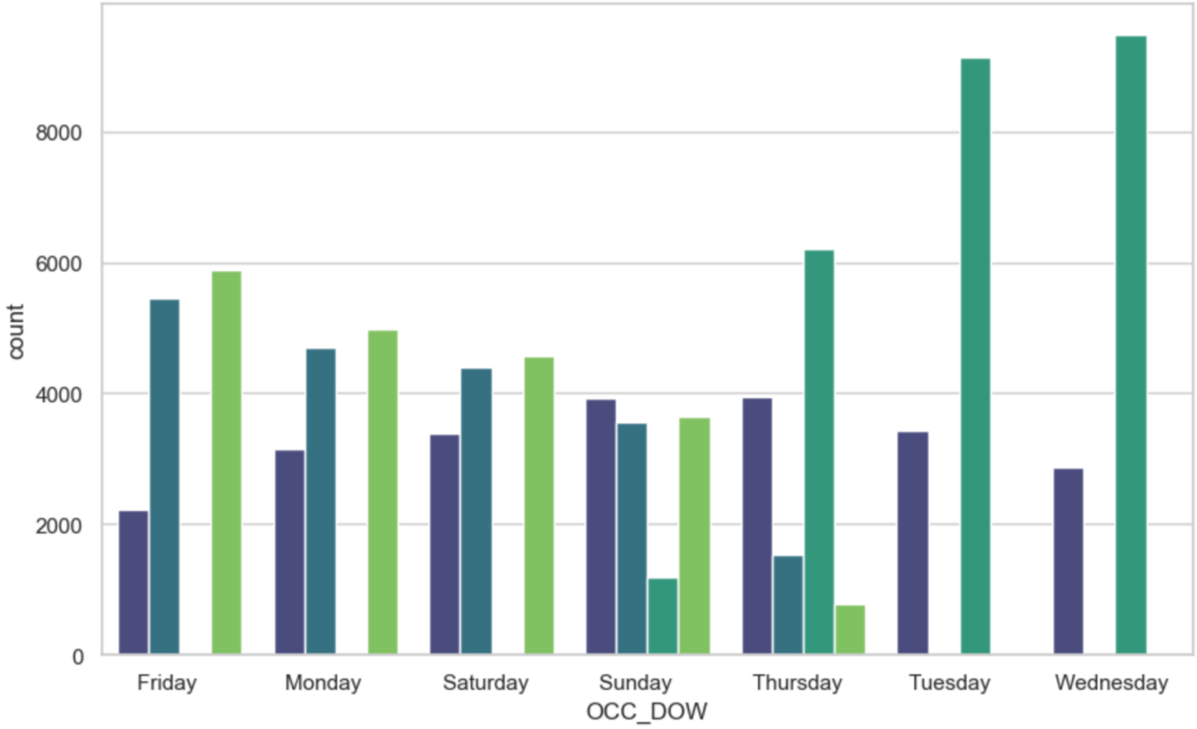


Figure 22: Cluster Model - OCC_DOW vs count

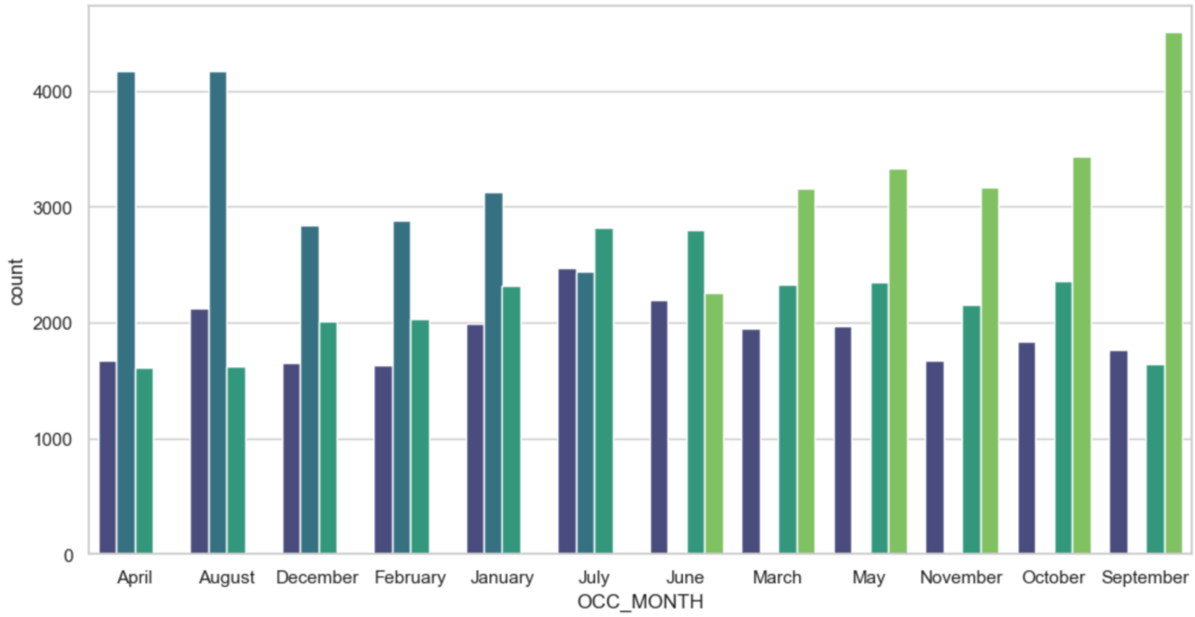
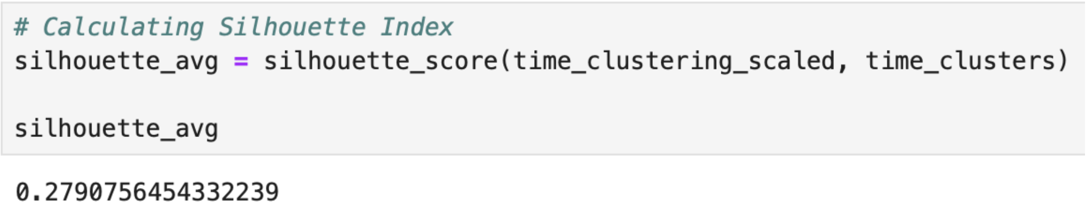


Figure 23: Cluster Model - OCC_MONTH vs count

These visualizations provide a clear view of how the time-based clusters are distributed across different time dimensions. Based on these plots, we can identify specific time-related patterns associated with each cluster, such as whether certain clusters represent periods with higher or lower incident rates. The variations in cluster distribution across hours, days, and months can provide valuable insights for understanding and predicting the timing of theft incidents.

These insights can be instrumental for law enforcement and urban planning, as they provide specific time windows and periods when increased vigilance or preventive measures might be necessary. Besides, raising public awareness about the most common times for theft could help in reducing the incidence of such crimes.

To evaluate this model, the Silhouette Score will be applied. It measures how similar an object is to its cluster compared to other clusters. The silhouette score ranges from -1 to 1, where a high value indicates that the object is well-matched to its cluster and poorly matched to neighboring clusters.



```
# Calculating Silhouette Index
silhouette_avg = silhouette_score(time_clustering_scaled, time_clusters)

silhouette_avg
```

```
0.2790756454332239
```

Figure 24: Silhouette Index

A score of 0.279 suggests that the clustering is moderate in quality. The data points are, on average, closer to their cluster center than to the centers of other clusters, but there's still considerable room for improvement.

Time Series Analysis:

The time series analysis focuses on monthly counts of vehicle theft from January 2013 to September 2022. To capture patterns and make predictions, the SARIMA (Seasonal Autoregressive Integrated Moving Average) model was used. Data preprocessing, model selection, and validation are all part of the analysis, which concludes with forecasting for the next 12 months. The analysis seeks to comprehend the temporal patterns in incidents of theft from motor vehicles and to make future predictions to aid decision-making and resource allocation.

The dataset is indexed by occurrence date, and monthly counts have been aggregated to identify trends. A plot was created to plot the time series data.

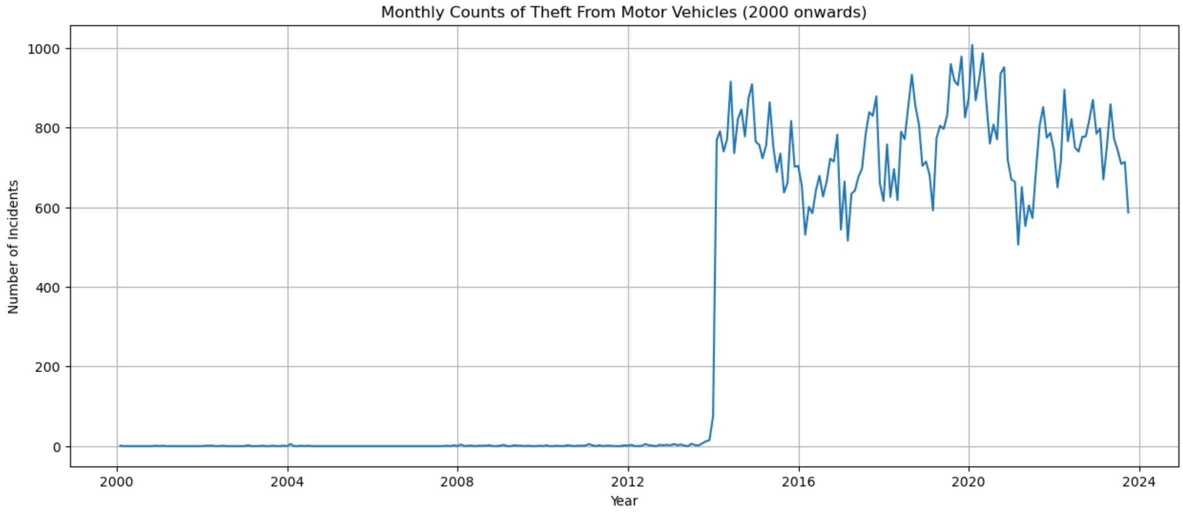


Figure 25: Monthly Counts of Theft from Motor Vehicles

Despite the data starting from the year 2000 with zero incidences, a more consistent start year is required, therefore, the annual number of incidences is calculated for each year. Having evaluated that the threshold for the number of crimes should be 50 and above, a more consistent start year is identified as 2013.

Figure 26 below shows the number of incidents from 2013 onward, we notice the number of incidents rise in 2020 compared to any other year and reduce gradually thereafter.

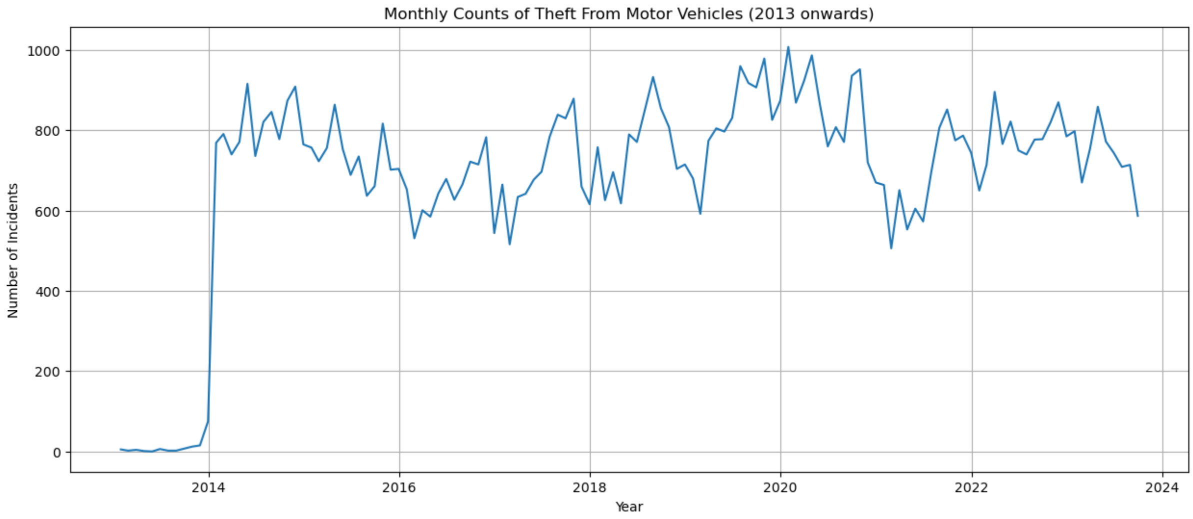
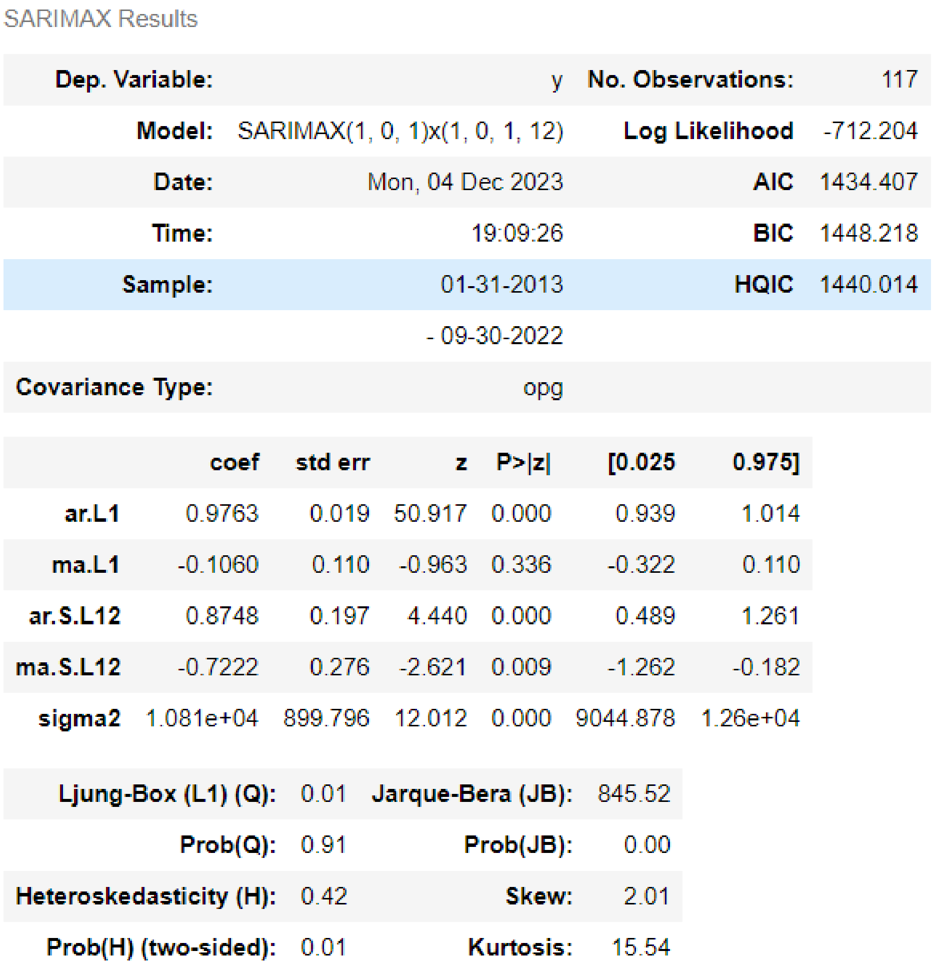


Figure 26: Monthly Counts of Theft from Motor Vehicles(2013 onwards)

A stationary time series has statistical properties that do not change over time, such as mean, variance, and autocorrelation. Because many statistical models and methods assume that the underlying data is stationary, stationarity is an important concept in time series analysis. To check if the data is stationary, the Augmented Dickey-Fuller test (Adler) is conducted. The resulting adf-statistic of **-3.3096776493626447** indicates that the test statistic is less than the critical values. This is an encouraging sign of stationarity. The p-value is 0.014451386674494908. The p-value is less than 0.05 (assuming a significance level of 5%), indicating that the null hypothesis of a unit root can be rejected. As a result, the data is most likely stationary.

After splitting the data into training and testing, the Seasonal AutoRegressive Integrated Moving Average (SARIMA) model fits the training set.



SARIMAX Results

Dep. Variable:	y	**No. Observations:**	117
Model:	SARIMAX(1, 0, 1)x(1, 0, 1, 12)	**Log Likelihood**	-712.204
Date:	Mon, 04 Dec 2023	**AIC**	1434.407
Time:	19:09:26	**BIC**	1448.218
Sample:	01-31-2013	**HQIC**	1440.014
	- 09-30-2022		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar.L1	0.9763	0.019	50.917	0.000	0.939	1.014
ma.L1	-0.1060	0.110	-0.963	0.336	-0.322	0.110
ar.S.L12	0.8748	0.197	4.440	0.000	0.489	1.261
ma.S.L12	-0.7222	0.276	-2.621	0.009	-1.262	-0.182
sigma2	1.081e+04	899.796	12.012	0.000	9044.878	1.26e+04

Ljung-Box (L1) (Q):	0.01	**Jarque-Bera (JB):**	845.52
Prob(Q):	0.91	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.42	**Skew:**	2.01
Prob(H) (two-sided):	0.01	**Kurtosis:**	15.54

Figure 27: SARIMAX Results

SARIMAX (1, 0, 1) x (1, 0, 1, 12) was applied to the dataset containing 117 observations. The log-likelihood of -712.204 and low AIC (1434.407) and BIC (1448.218) values of the model indicate a robust fit to the data, indicating an effective balance between the goodness of fit and model complexity. The coefficients reveal the effect of non-seasonal and seasonal autoregressive and moving average terms on the predicted values. Notably, the autoregressive term (AR.L1) and seasonal autoregressive term (AR.S.L12) have significant z-statistics (p 0.05) highlighting their importance in capturing temporal dependencies. The residual diagnostic tests show a non-significant Ljung-Box statistic (Q), indicating no residual autocorrelation, while the Jarque-Bera test indicates potential residual non-normality.

Forecast for the next 12 months.

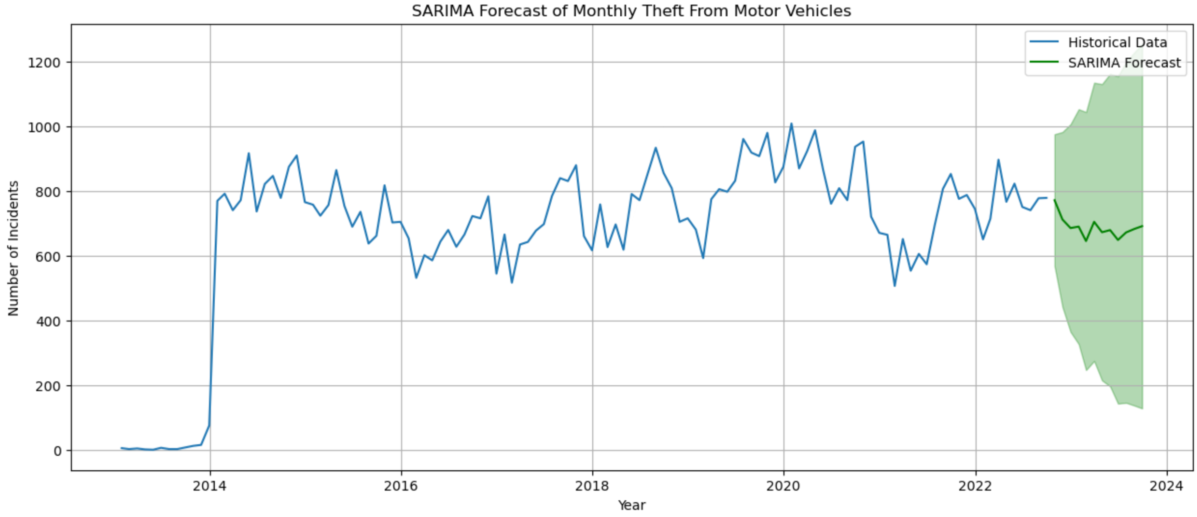


Figure 28: SARIMA Forecast of Monthly Theft from Motor Vehicles

In the figure above, the rise and fall of crime has been consistent from 2013 to 2022 where the historical data ends. However, after using the SARIMA model, a forecast of the number of incidences for the period 2022 to 2024 was created.

The Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are three metrics commonly used to assess the performance of time series forecasts. Here's how to interpret these findings and understand what they mean for the forecasted data:

a) MAE (Mean Absolute Error):

The average absolute difference between observed and predicted values is defined as MAE. In this case, an MAE of 86.78 indicates that the model's predictions deviate from the actual values by approximately 86.78 units on average.

b) MSE (Mean Squared Error):

The mean squared error (MSE) is the average of the observed and predicted values' squared differences. An MSE of 9885.52 indicates that the average squared difference between predictions and actual values is 9885.52.

c) RMSE (Root Mean Squared Error):

The RMSE is the square root of the MSE and represents the average magnitude of the errors. An RMSE of 99.43 indicates that the model's predictions deviate from the actual values by approximately 99.43 units on average.

The MAE, MSE, and RMSE values are measures of forecast accuracy, and lower values indicate better performance. In this context, an MAE of 86.78 suggests that, on average, the model's predictions have relatively small errors.

In the figure above, it is observed that the rate of crime remains consistent with past occurrences. There was a small decrease in the number of incidents in 2013, however, they increased again in the year 2024. This implies that without effective measures from the Toronto Police Service, the number of motor vehicle thefts will continue to rise.

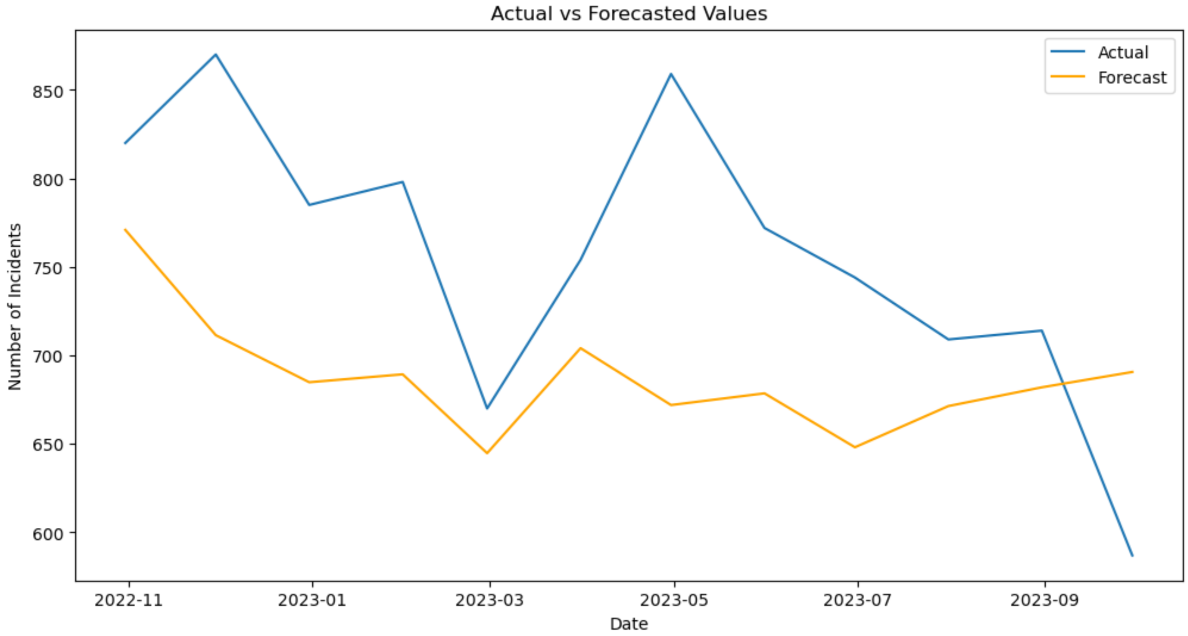


Figure 29: Comparison between Actual and Forecasted Values

The actual vs. predicted graph depicts the performance of a time series forecasting model by displaying the alignment or divergence between actual observed values and corresponding predictions. The Mean Absolute Error (MAE) of 86.78 and the Root Mean Squared Error (RMSE) of 99.43 indicate relatively small average deviations between predicted and actual values in this context. One would expect to see a close correspondence between the two lines when examining the actual vs. predicted graph, with the predicted values generally tracking the observed values. Any differences or discrepancies between the lines could indicate the model's ability to capture the underlying patterns and trends in the time series data. The huge difference in the actual vs. the predicted in the period 2023-03 to 2023-07 indicates the areas in the model struggled to accurately forecast the outcomes. The same is noted for the period immediately after 2023-09. However, the model accurately forecasts the actual occurrences in the period after 2023-09.

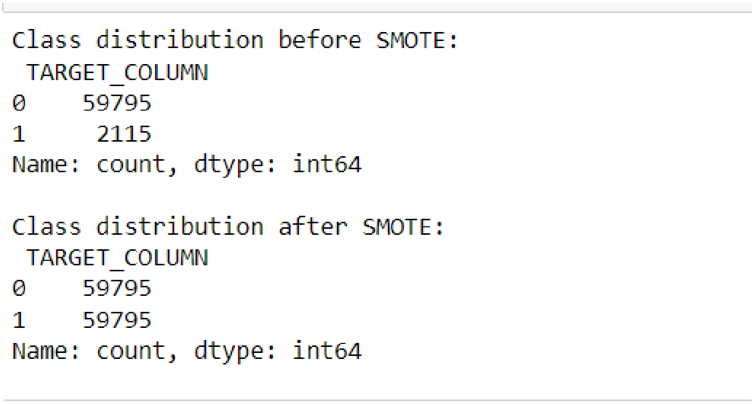
Random Forest Model:

Data Preprocessing:

- The target column is defined as 'TARGET_COLUMN' with binary values (0 or 1) based on whether the offense category is 'Theft Over'.
- Feature columns are selected, including 'REPORT_HOUR', 'REPORT_DOW', 'LOCATION_TYPE', 'NEIGHBOURHOOD_158', 'OCC_HOUR', and 'OCC_DOW'.
- One-hot encoding is applied to categorical columns.
- The data is split into training and testing sets (70-30 split).

Handling Class Imbalance with SMOTE:

- SMOTE is a technique used to address class imbalance in machine learning datasets. It works by generating synthetic samples for the minority class, creating a more balanced distribution. This helps prevent the model from being biased toward the majority class and improves its ability to generalize to the minority class.
- SMOTE (Synthetic Minority Over-sampling Technique) is applied to handle the class imbalance issue in the training set.
- The class distribution before and after SMOTE is displayed, showing an equal number of instances for both classes after SMOTE.



```
Class distribution before SMOTE:
 TARGET_COLUMN
0    59795
1     2115
Name: count, dtype: int64

Class distribution after SMOTE:
 TARGET_COLUMN
0    59795
1    59795
Name: count, dtype: int64
```

Figure 30: SMOTE Analysis

Figure 31 below visualizes the class distribution after SMOTE analysis and after applying that both the categories are now divided equally in the taken sample of the dataset.

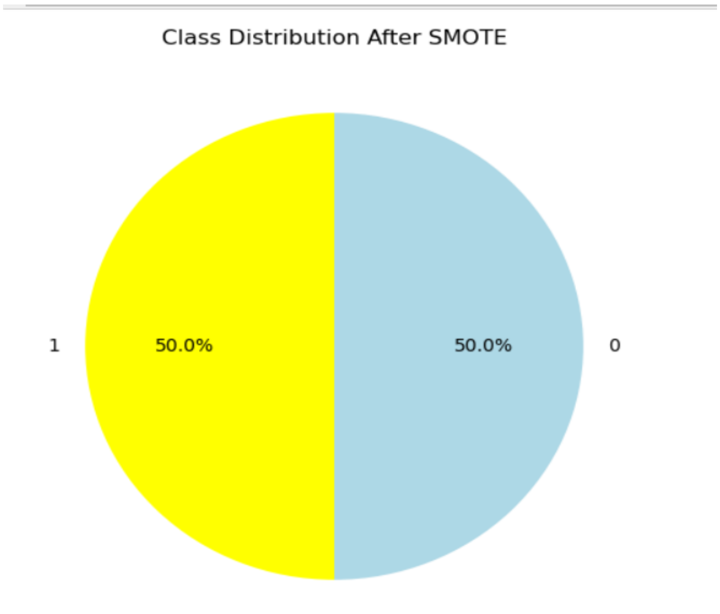


Figure 31: Graphical representation after SMOTE analysis

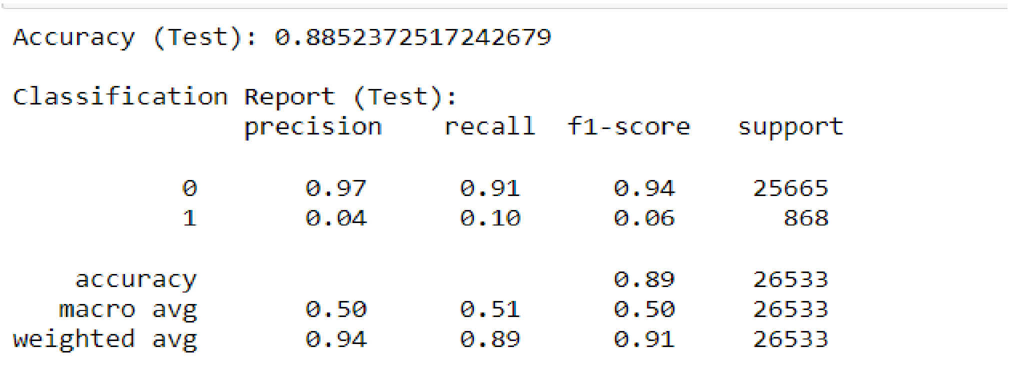
Test and Train Data:

- Training data is used to train the model, and SMOTE is applied to address class imbalance in this set.
- Testing data is used to evaluate the model's performance on unseen data, providing insights into its generalization capabilities.

Random Forest Model:

- A Random Forest classifier is built and trained on the resampled data.
- The model is trained with 100 decision trees, a random state of 42, and a minimum of 2 samples per leaf.
- Confusion matrices are generated for both the training and testing sets.

Evaluation Metrics:



```
Accuracy (Test): 0.8852372517242679

Classification Report (Test):
              precision    recall  f1-score   support

           0       0.97      0.91      0.94     25665
           1       0.04      0.10      0.06       868

    accuracy                           0.89     26533
   macro avg       0.50      0.51      0.50     26533
weighted avg       0.94      0.89      0.91     26533
```

Figure 32: Accuracy and Classification Report

- The confusion matrix for the training set shows the number of true positives, true negatives, false positives, and false negatives.
- The confusion matrix for the testing set is displayed along with the predicted class distribution.
- The accuracy, precision, recall, and F1-score are calculated and presented in the classification report for the testing set.

Summary and Justification:

- The Random Forest model is chosen due to its ability to handle complex relationships in data, handle non-linearity, and provide feature importance.
- SMOTE is applied to address the class imbalance problem, ensuring the model is not biased toward the majority class.

- On the test set, the RandomForest classifier achieved an overall accuracy of 88.52%, demonstrating a strong performance in predicting incidents classified as "Non-Theft Over" (Class 0). However, the model has difficulty properly predicting "Theft Over" occurrences (Class 1), as evidenced by the poor precision (4%), recall (10%), and F1-Score (6%). With a considerably larger number of Class 0 data, the class imbalance contributes to the model's bias towards predicting the majority class. While the precision and recall for Class 0 are high (97% and 91%, respectively), the model's limitations in identifying "Theft Over" incidents necessitate further refinement, potentially involving class imbalance strategies or fine-tuning model parameters for improved performance on the minority class.

Summary

Clustering Analysis:

Objective: Identify patterns or groups based on the timing of theft incidents.

Features Used: Timing features such as hours of the day, days of the week, or months of the year.

Outcome: Uncover insights into when theft incidents are most frequent and identify distinct patterns in the temporal distribution of thefts.

Applications: Useful for allocating resources and law enforcement during specific times when theft incidents are more likely to occur.

Time Series Analysis with SARIMA:

Objective: Capture patterns and make predictions based on monthly counts of vehicle theft from January 2013 to September 2022.

Model Used: SARIMA (Seasonal Autoregressive Integrated Moving Average).

Outcome: Provides a predictive model for understanding and forecasting the temporal trends and seasonality of vehicle theft over the given time period.

Applications: Enables law enforcement and relevant authorities to anticipate and prepare for potential increases or decreases in theft incidents in the future.

Random Forest Model for Classification:

Objective: Build a predictive model for classifying criminal incidents into two categories: "Theft Over" and "Non-Theft Over" (or "NonMCI").

Features Used: Various features, including timing, location, and other relevant variables.

Outcome: A classification model that can distinguish between different types of criminal incidents, particularly focusing on the distinction between "Theft Over" and other incidents.

Applications: Supports law enforcement in identifying high-risk areas and predicting the likelihood of specific types of criminal incidents.

Overall Strategy:

Data-Driven Decision Making: All three approaches leverage data to drive decision-making processes, providing insights into patterns and predictive capabilities.

Comprehensive Approach: The combination of clustering, time series analysis, and classification modeling offers a comprehensive strategy for understanding and addressing theft-related challenges.

Resource Optimization: These analyses can inform the optimal allocation of resources, both in terms of law enforcement personnel and preventive measures.

Key Findings

- Temporal Trends: Different temporal patterns in motor vehicle theft were found by the analysis, with higher incidents occurring during specific hours, days, and months. There was a higher chance of theft on Sunday evenings, and nighttime was found to be the busiest time for car theft.

- Geographical Distribution: Spatial analysis revealed hotspots that indicated which neighborhoods and divisions were more vulnerable to auto theft. The top ten neighborhoods for theft were identified, offering useful data for focused interventions.
- Demographic and Location Vulnerability: Hotspots were discovered through spatial analysis, and certain divisions and neighborhoods showed greater susceptibility to auto theft. A valuable source of information for focused interventions was the highlighting of the top 10 neighborhoods for theft.
- Time Series Analysis: Future trends in theft were predicted by the Time Series Analysis employing SARIMA. Despite the model's consistency with historical data, care should be taken, particularly in cases where the model had difficulty making accurate predictions.
- Random Forest Model: By utilizing SMOTE to address class imbalance, the Random Forest model proved to be highly effective in forecasting incidents that were categorized as "Non-Theft Over." Nevertheless, more improvement is required to predict the minority class of "Theft Over" incidents more accurately.

Recommendations

- Enhanced Policing during Peak Periods: Using temporal analysis, law enforcement organizations can strategically deploy resources during times of high theft. A higher level of awareness during particular days, hours, and months can aid in discouraging and preventing auto theft.
- Targeted Community Interventions: Community organizations can create focused interventions by using the data on vulnerable demographics and high-frequency crime hotspots. This could involve joint efforts with law enforcement, neighborhood watch initiatives, and public awareness campaigns.

- Predictive Policing: Predictive policing initiatives can be built upon the Time Series Analysis. Forecasted trends can be used by law enforcement to proactively allocate resources and identify potential hotspots for theft in the future.
- Refinement of Predictive Models: Time Series Analysis can be used as a foundation for predictive policing projects. Law enforcement can proactively allocate resources and identify potential hotspots for theft in the future by using forecasted trends.
- Collaborative Efforts: Together, community organizations, law enforcement, and policymakers should address the various factors that contribute to motor vehicle theft. Sustainable crime reduction can be achieved by combining social initiatives, community engagement, and effective policing holistically.

Conclusion

Conclusively, this thorough examination of car theft occurrences spanning from 2014 to 2023 offers significant perspectives on the historical trends, regional distribution, and patterns linked to these offenses. To address the underlying causes of theft and improve community safety, the project seeks to arm communities, law enforcement, and policymakers with knowledge. With the help of this project, stakeholders will be able to take proactive steps to lower motor vehicle theft and improve community safety by using evidence-based decision-making as a foundation. The long-term viability of crime prevention initiatives will depend on the ongoing assessment, analysis, and modification of tactics in light of new developments.

BIBLIOGRAPHY

- Toronto Police Service

  https://data.torontopolice.on.ca/datasets/TorontoPS::theft-from-motor-vehicle-open-data/about

- A Complete Guide to Data Visualization in Python With Libraries & More

  https://www.simplilearn.com/tutorials/python-tutorial/data-visualization-in-python#:~:text=Matplotlib%20and%20Seaborn%20are%20python,primarily%20used%20for%20statistical%20graphs.

- Random Forest Classification with Scikit-Learn

  https://www.datacamp.com/tutorial/random-forests-classifier-python

- Time Series Analysis and Forecasting | Data-Driven Insights

  https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/

- Understanding K-means Clustering in Machine Learning

  https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1