**Business Memo**

Toronto is one of the major cities in Canada, and its larger population increases the risk of traffic collisions. Through the utilization of historical data, our team has performed various analysis to develop a multi-purpose dashboard. Our dashboard visualizes historical data to portray various trends of traffic collisions and also provides the audience with our forecasting analysis and predictions. Our analytics project consists of five processes as listed below:

*1.* *Data Processing and Aggregation*

We utilized public historical data obtained from two primary sources to analyze traffic collisions in relation to weather conditions. The traffic collision data was sourced from the Toronto Police Service, which provides a comprehensive dataset through the City of Toronto Open Data portal. Additionally, weather-related data was retrieved from the Government of Canada's climate database, offering detailed daily climate measurements across various stations. These datasets were meticulously merged based on the date of each event, allowing for a robust dataset that combines traffic and weather variables.

*2.* *Feature Engineering*

To enhance our understanding of traffic collisions and to facilitate deeper analysis of potential causative factors, we transformed and enriched the raw data through several feature engineering steps:

1. Seasonal Analysis: Recognizing the potential impact of seasonal variations on collision rates, we classified dates into meteorological seasons. This categorization helps identify seasonal trends in traffic collisions and examines whether certain weather conditions linked to specific seasons influence the frequency or severity of accidents.
2. Day Type Classification: To determine if the risk of collisions varies between weekdays and weekends, we categorized each date accordingly. This distinction is crucial for understanding traffic patterns and planning city traffic management and emergency response strategies more effectively.
3. Rush Hour Identification: We introduced a binary feature indicating whether a collision occurred during typical rush hours, which is known for higher traffic volumes and potentially higher accident rates. This feature is essential for assessing the impact of traffic density on collision occurrences.

These engineered features are expected to provide valuable insights into the dynamics of traffic collisions and support the development of targeted strategies to enhance road safety.

*3.* *Classification (Logistic Regression, KNN, and Decision Tree)*

To further understand the classification of the predicted value of traffic collision data, we built the model with 3 different methods to test the results before making a business decision on the appropriate one. The one with the best fit is selected based on the criteria of accuracy, precision, and recall. Upon thorough analysis on Logistic Regression, KNN, and Decision Tree, the data shows that KNN is the best fit to use to classify our business case: traffic collision data. The results obtained can be proposed to the City of Toronto to aid in ambulance usage, and only dispatch ambulances to collisions that are injury involved.

*4.* *Forecasting (Time series/Arima)*

To better predict traffic accident trends in Toronto, we used timeseries analysis and ARIMA models. In the timeseries analysis, we visualized data from 2014 to 2024 using Single Exponential Smoothing, Holt’s Linear Trend Method, and Holt-Winters Triple Exponential Smoothing, showing long-term trends and fluctuations. Holt’s Linear Trend Method performed the best, successfully capturing a linear downward trend. Additionally, we used the ARIMA model for monthly forecasting. By converting date formats, analyzing trends and seasonality, and using the ADF test to ensure data stationarity, our final forecast shows that 2024 data nearly returns to pre-pandemic levels, with similar patterns expected in the coming years.

*5.* *Dashboard*

We have presented our story through two different dashboards: one displaying our analysis of the historical data and one with our predictive analysis. The dashboards are easy to navigate and interpret. The historical data dashboard consists of a map of Toronto, neighbourhood data table, two fatalities tree maps, bar graph displaying number of collisions per season, and a side-by-side circle graph displaying the number of collisions over months. The map allows the user to hover over each location to identify the number of automobile, bicycle, and motorcycle collisions per neighbourhood. The size and colour of the points on the maps correlate with the number number of fatalities in the neighborhood. The data table allows the users to look at the number of various types of collisions per neighbourhood and they may filter out the data by their column of choice. For example, filtering by automobile collisions allows us to see that West Humber-Clairville has the highest number of collisions. The first tree-map displays the number of fatalities based on rush hour. We can see that there were more fatalities during off peak hours. The second tree-map displays the number of fatalities per month. January, November, and March had the highest number of fatalities. The bar graph displays the number of automobile, bicycle, and motorcycle collisions per season. Automobile collisions were highest for all seasons, second being bicycle, and last motorcycle. The side-by-side circle graph depicts the number of hit and run, injury, and pedestrian collisions over the months. January and February have the highest number of collisions. The second dashboard depicts the graphical results obtained from our forecasting (time series and Arima) models as stated in process number four.

Overall, our project goal is to help the police force identify which neighbourhoods have the highest number of traffic collisions and the correlations between when the different types of collisions occur. Through this information, Toronto Police can gain a deeper understanding of what and where proactive measures need to be implemented. As mentioned, the results from our KNN classification may be utilized to increase efficiency in ambulance usage by only dispatching ambulances to injury collisions. Furthermore, our forecasting results show a downward trend in upcoming traffic collisions, and that the upcoming collision cases resemble 2024 and pre-pandemic times. This analysis could potentially help the City of Toronto decide how much funds should be allocated for collision aid. Lastly, our historical dashboard is able to provide Toronto Police with information on which areas in the city are likely to have collisions, when these collisions are likely to occur, and what type of collisions they may be. This data can help them stay proactive in such areas to control traffic.

This concludes the business memo. For further information, please refer to our dashboard images and Technical Report below.

Thank you,

Team New York

MMA 2025B

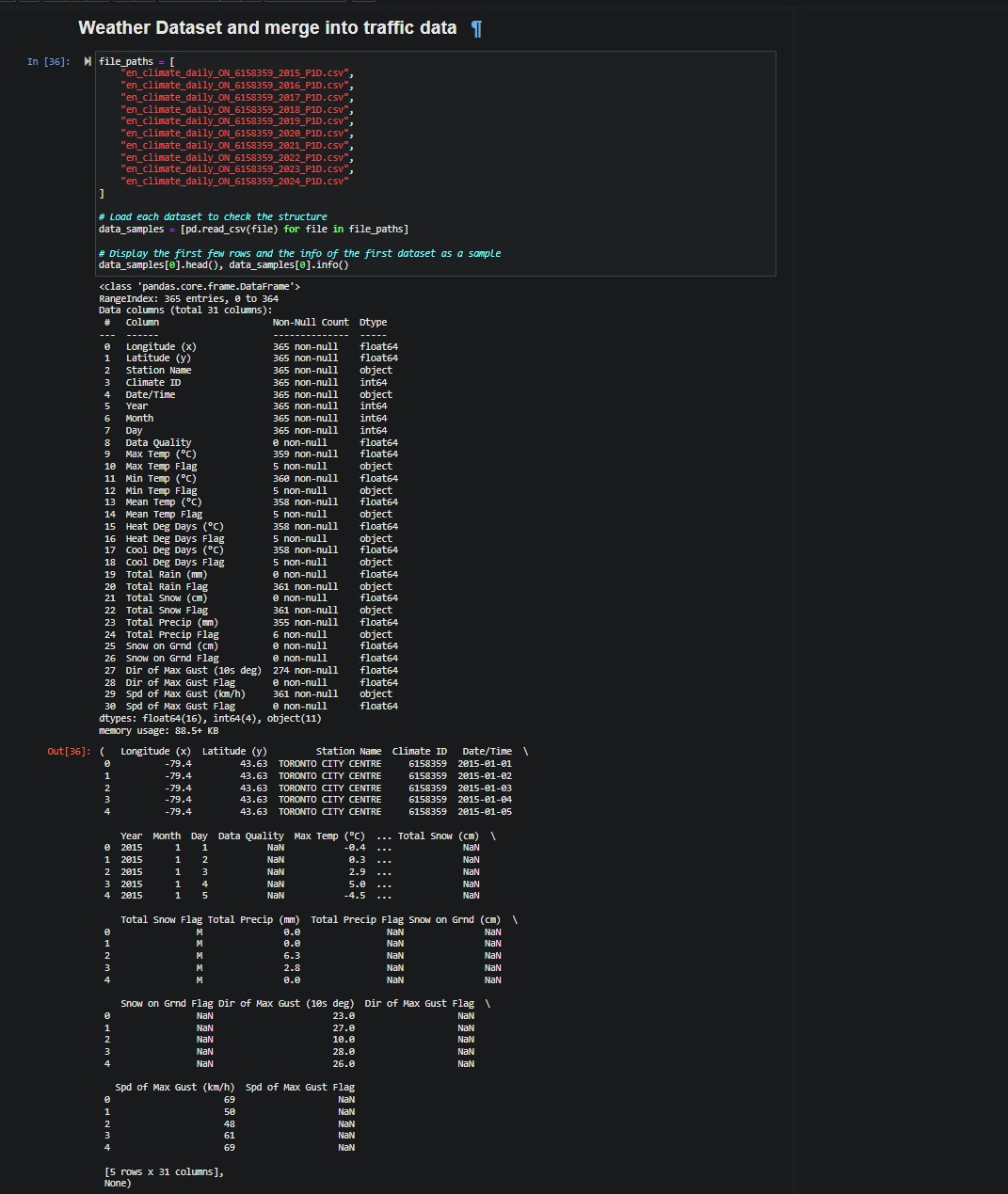
**DASHBOARD**

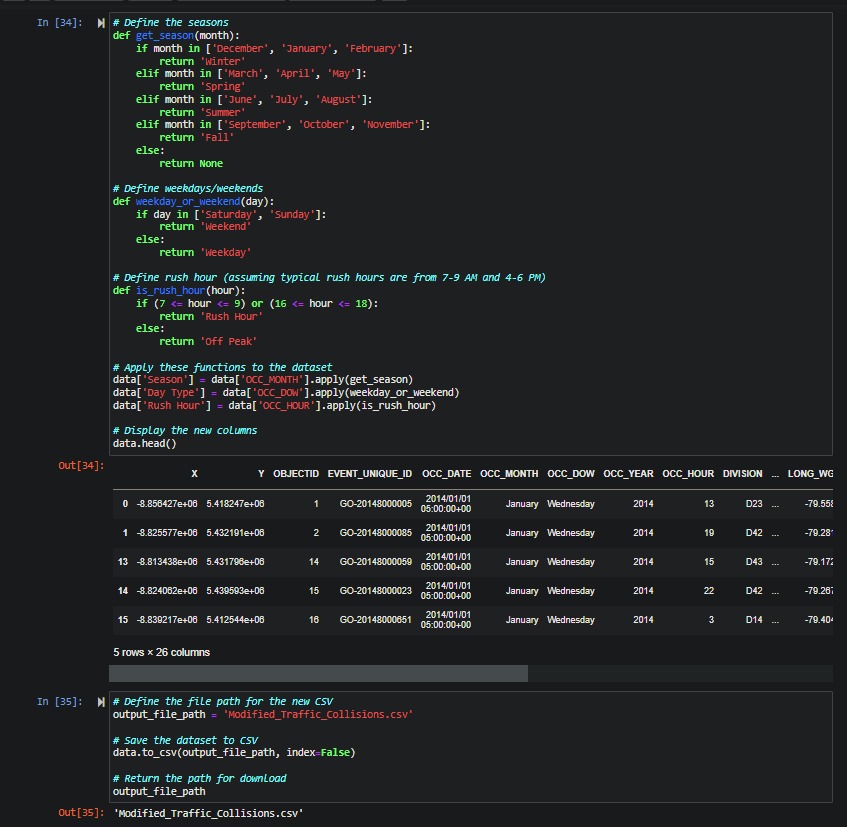
|  |
| --- |
|  |
|  |

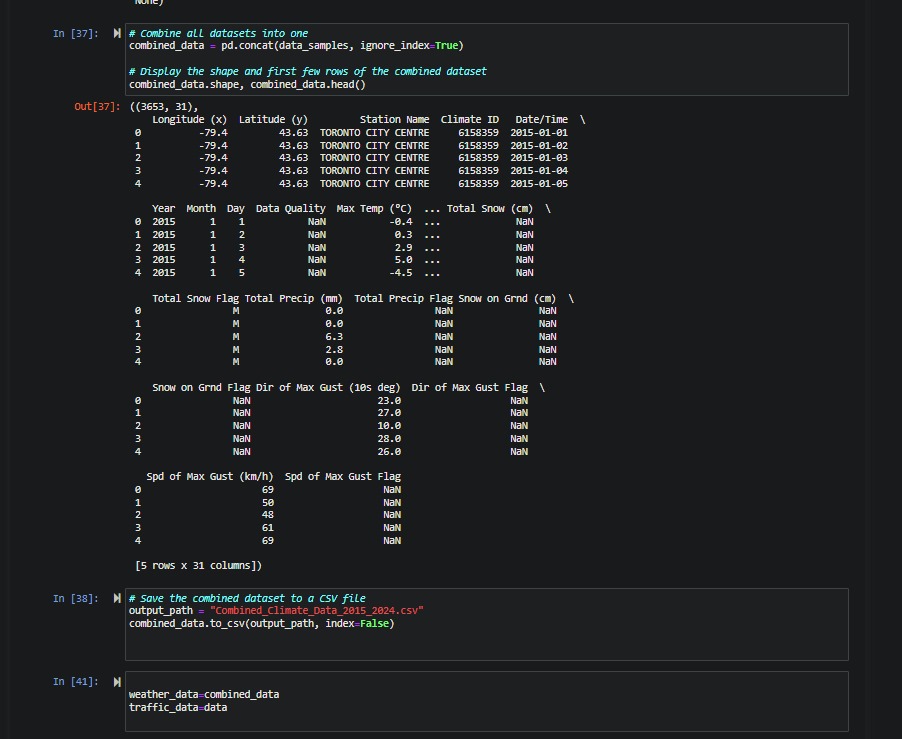
**TECHNICAL REPORT**

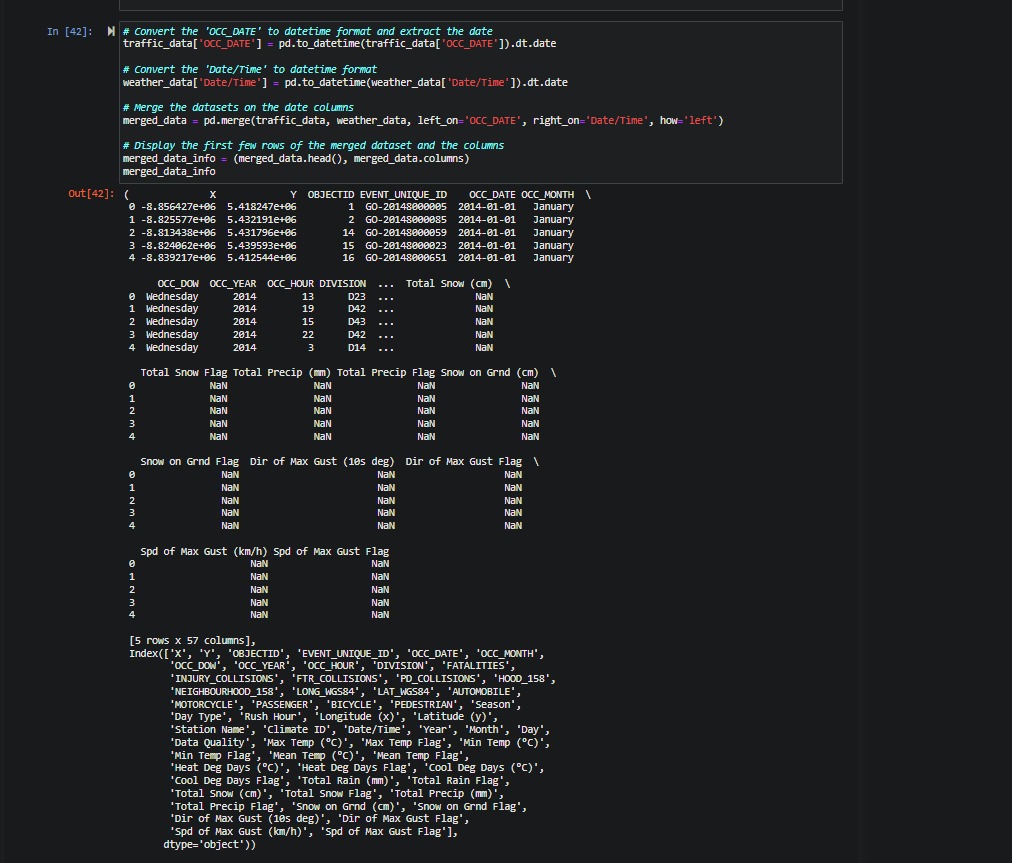
1. **Data processing and aggregation**

Merging data sets:

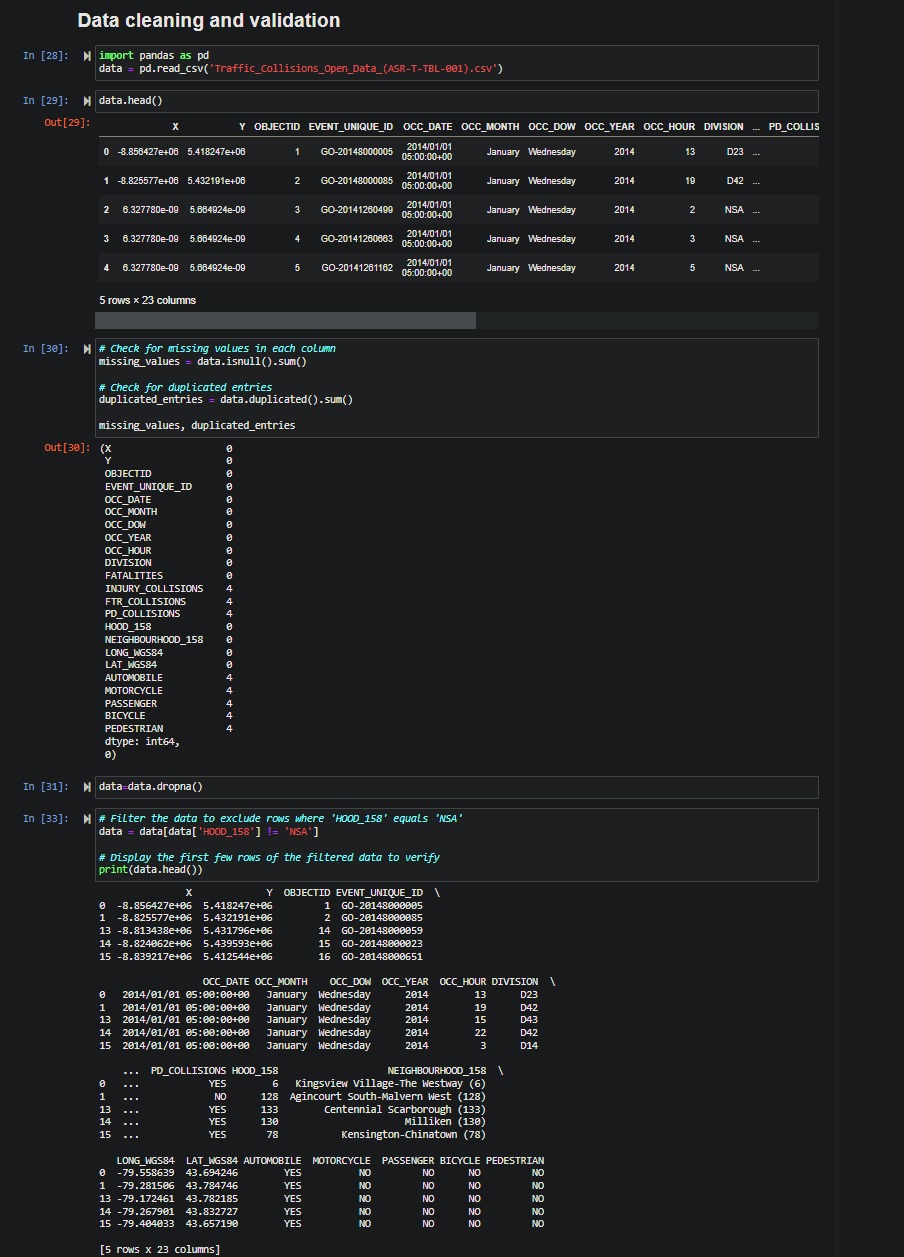


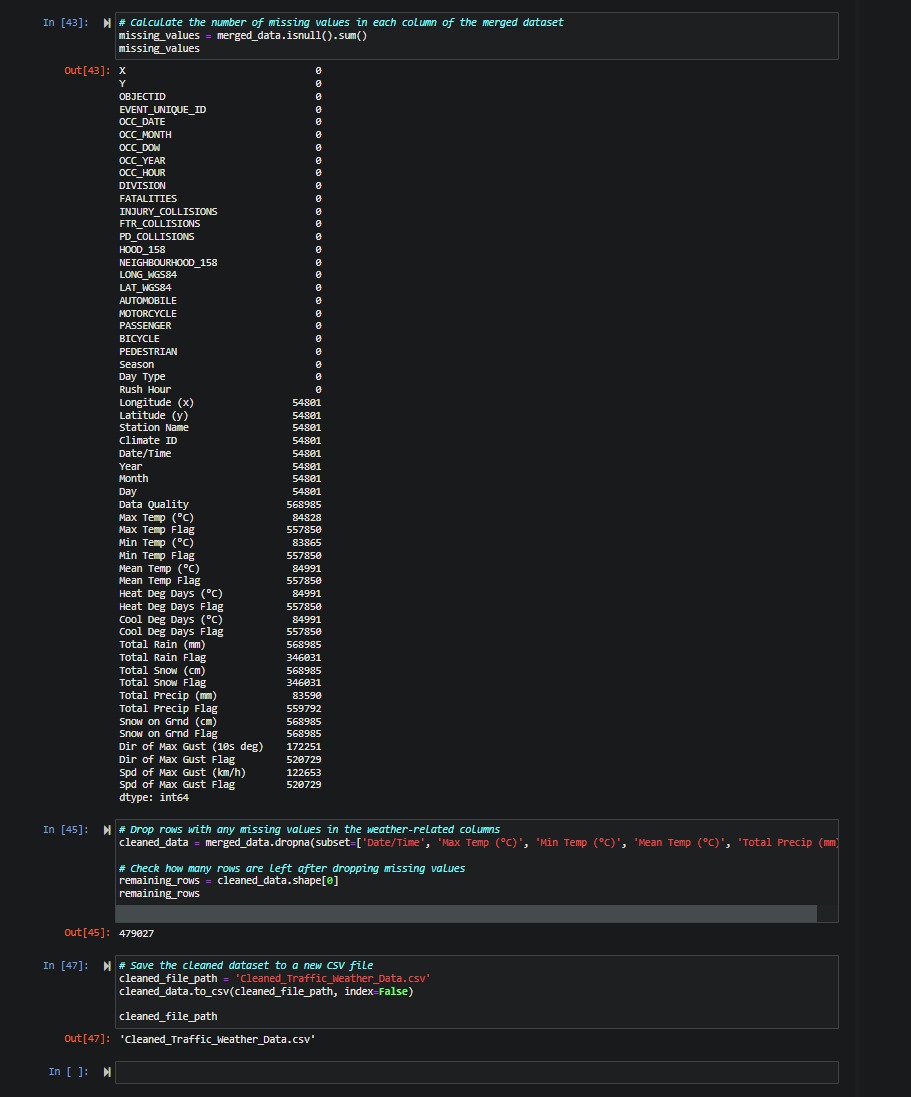


****

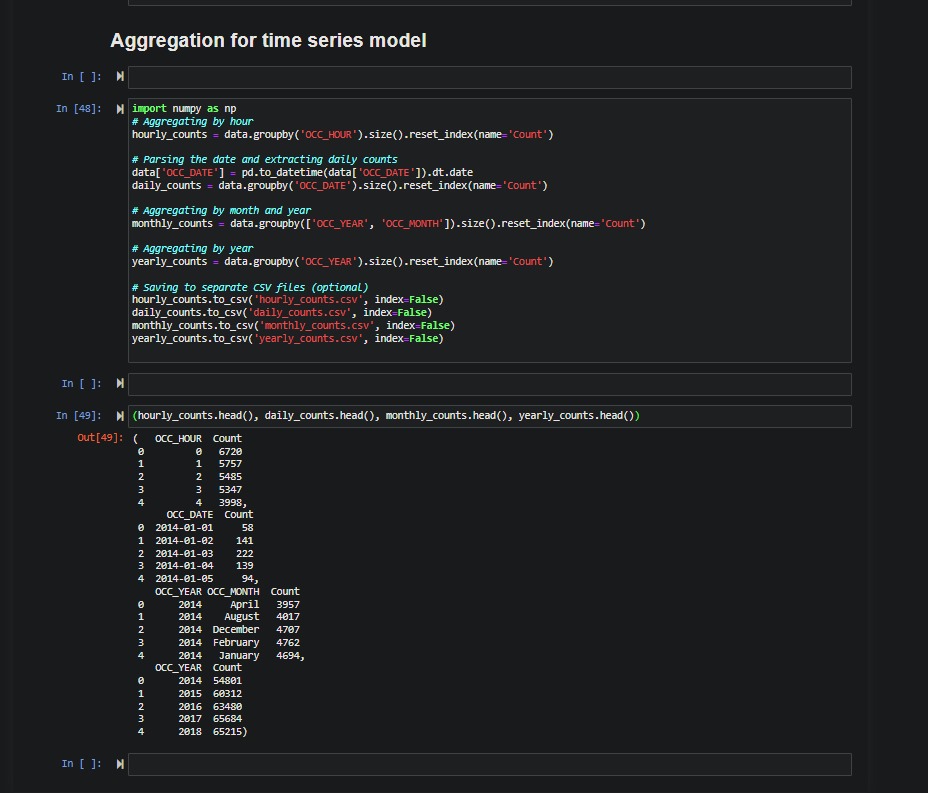
****

Cleaning and validating data:

****

****

Data Aggregation for Time Series Model:

****

1. **Feature Engineering**
2. Seasonal Analysis
   * We created a new column for ‘Season’ by grouping dates according to meteorological seasons.
3. Day Type Classification
   * We created a new column for ‘Day Type’ by categorzing each date as either ‘weekend’ or ‘weekday’.
4. Rush Hour Identification
   * We created a new column for ‘Rush Hour’ indication if a collision occurred ‘off peak hours’ or during ‘rush hour’ based on the hours data that was provided.
5. **Classification**

We have classified the ‘Injury Collisions’ from the Traffic Collisions Open Data obtained from the Toronto Police Service. We have used Logistic Regression, KNN, and Decision Tree classification methods. Here is the summary of our analysis:

**Accuracy:** All three models have very high train and test accuracy, with KNN having the highest train accuracy (0.9942) and Decision Tree having the highest test accuracy at 0.9929.

**Precision:** Logistic Regression shows the highest test precision (0.9926), while Logistic Regression has the highest train precision (0.9915).

**Recall:** KNN provides highest in both train recall (0.9755) and test recall (0.9678).

**Confusion Matrix:** The confusion matrices show the number of true positives, false positives, true negatives, and false negatives. While all models perform well, Logistic Regression has the lowest number of false positives and false negatives, suggesting better performance in identifying positive cases.

**Conclusion:**

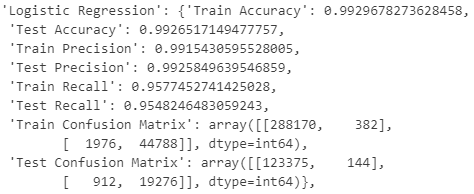
While all models perform well, the KNN model strikes a good balance between precision and recall, making it a strong candidate based on the given metrics.

**Hyperparameters used:**



**Classification results:**

Logistic Regression:

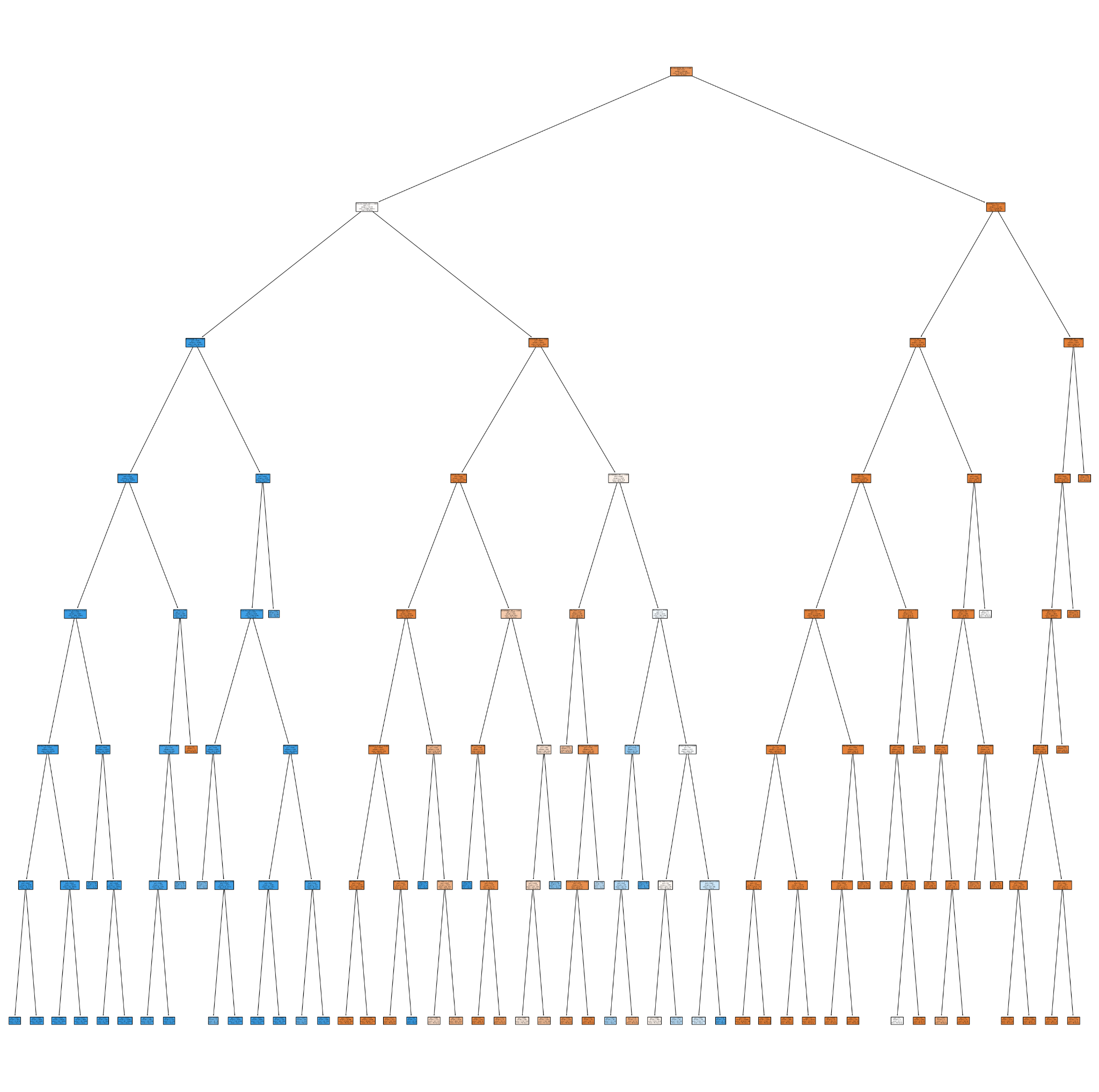


KNN:



Decision Tree:



****

Different Classification Model Performance

Accuracy: The ratio of correct predictions vs all total predictions.

Logistic Regression:

* Train: 99.29%
* Test: 99.26%

Decision Tree:

* Train: 99.33%
* Test: 99.29%

KNN:

* Train: 99.42%
* Test: 99.26%

All 3 models have high accuracy for both train and test set. The accuracy score for train and test are very close indicating there is no signs of overfitting.

Precision: The ratio of correct positive prediction vs the total positive predictions.

Logistic Regression:

* Train: 99.15%
* Test: 99.25%

Decision Tree:

* Train: 98.79%
* Test: 98.84%

KNN:

* Train: 98.30%
* Test: 97.94%

Precision score is also high for all 3 models indicating that the models made correct positive predictions. Logistic Regression has the highest score shows that it made more true positive predictions than Decision Tree and KNN.

Recall: The ratio of correct positive prediction vs all true positive predictions.

Logistic Regression:

* Train: 95.77%
* Test: 95.48%

Decision Tree:

* Train: 96.39%
* Test: 96.12%

KNN:

* Train: 97.55%
* Test: 96.78%

Recall is high for all 3 models indicating that all models are good at making predictions. KNN has the highest recall score shows that KNN made more true positive predictions than Logistic Regression and Decision Tree.

Analysis:

For our business problem, KNN is recommended to be the best fit model classification method to use to classify traffic collisions to be injury or non-injury related.

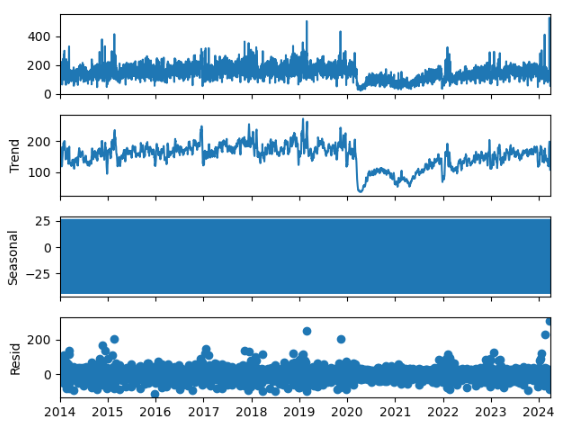
Best Model Selection

The KNN is the best model. The high accuracy on the testing data set, the balancing on precision and recall effectively. It demonstrated credibility and reliability for new, unseen data prediction.

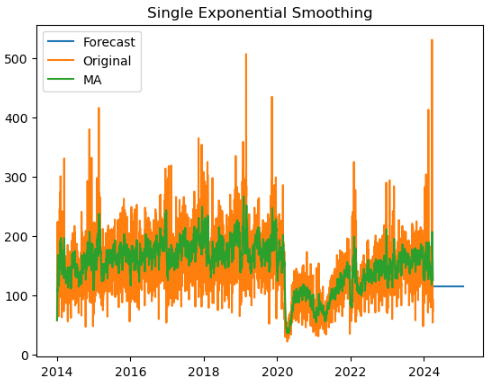
1. **Forecasting**

Time-series

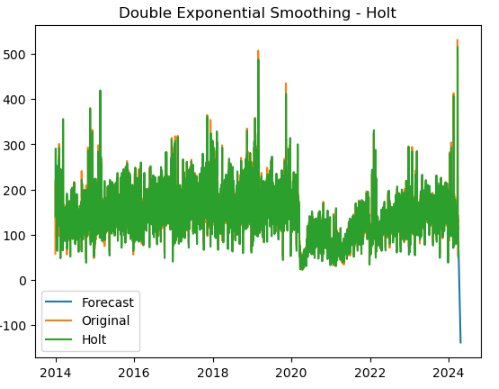
We visualized the traffic accident data from 2014 to 2024, illustrating the long-term trends and fluctuations.



We initially employed time series methods to analyze the traffic accident data, utilizing three smoothing methods: Single Exponential Smoothing, Holt’s Linear Trend Method, and Triple Exponential Smoothing (Holt-Winters). These methods were used to capture potential trends and seasonal variations in the data and to visualize the results.



This method predicts that the future number of traffic accidents will remain stable, failing to capture any trend or seasonal variations. The forecasted values appear as a horizontal line, indicating the method's limitation in reflecting changes in the data trends.



The Double Exponential method successfully captures the linear downward trend in the data, predicting a gradual decrease in future traffic accidents. The fitted values show the highest alignment with the actual data values, demonstrating the method's superior performance in capturing long-term trends.



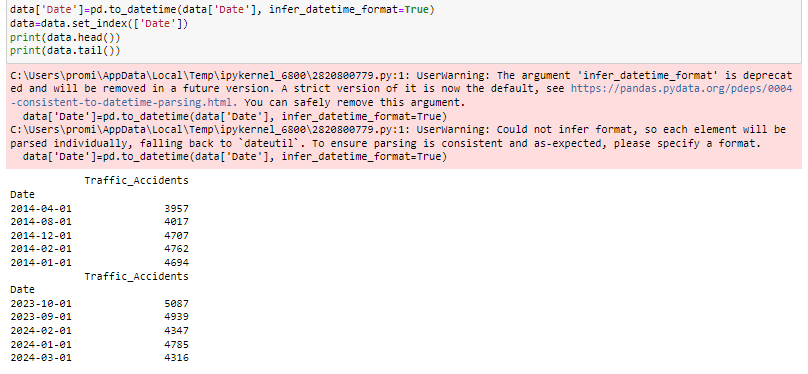
Holt Winter method captures both trend and seasonal variations in the data. However, due to the relatively weak seasonal pattern in the dataset, its performance is slightly inferior to Holt’s Linear Trend Method. The fitted values align well with the actual data but are slightly less accurate than those from the Holt method.

By comparing the forecasting results of the three smoothing methods, we found that Holt’s Linear Trend Method performs the best in capturing long-term trends with the smallest prediction error. Although the Triple Exponential Smoothing method also captures trend and seasonal variations, its performance is slightly lower due to the weak seasonality in this dataset. Therefore, we recommend using Holt’s Linear Trend Method for practical applications.

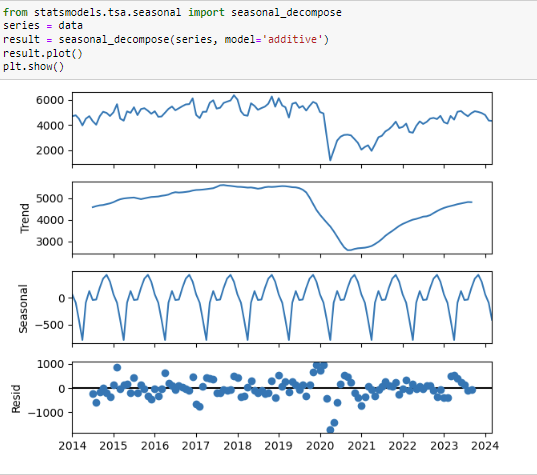
**ARIMA Model**

We set up a monthly forecasting flow for our Arima Model as opposed to daily for the time series forecasting above.

First, we converted our **Date** column to datetime.

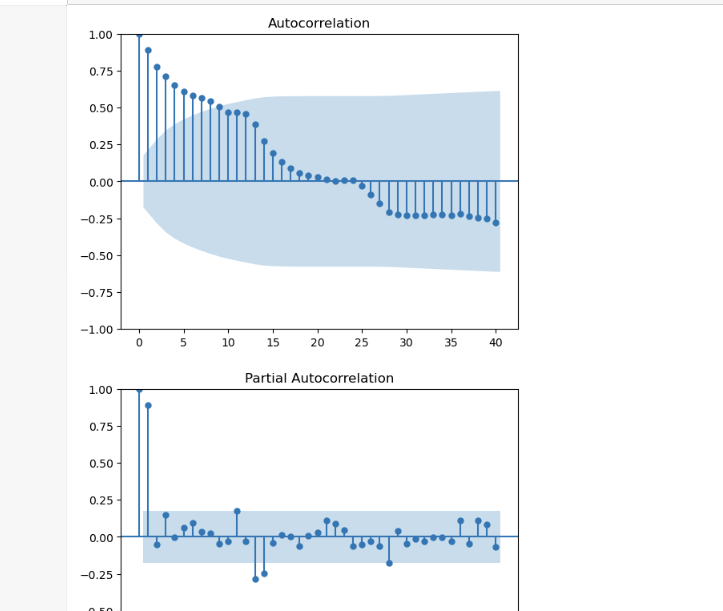
****

Then we took a look at our Trend, Seasonality and Randomness.



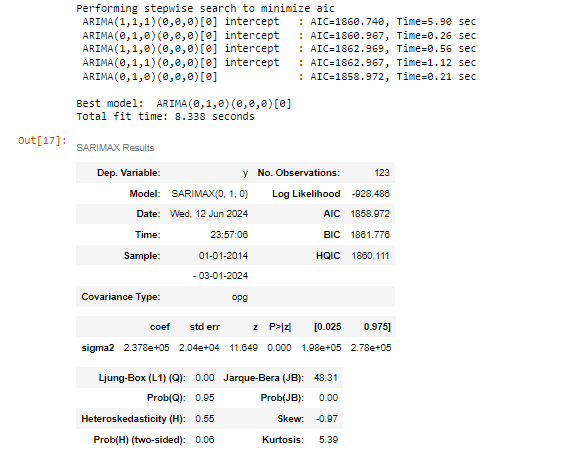
ADF Statistic: -2.355987 p-value: 0.154546- With a P value greater than the significance level we will fail to reject the null hypothesis. This implies no stationerity exists.

Next we analyze the Autocorrelation/PACF to identify lags in the series.



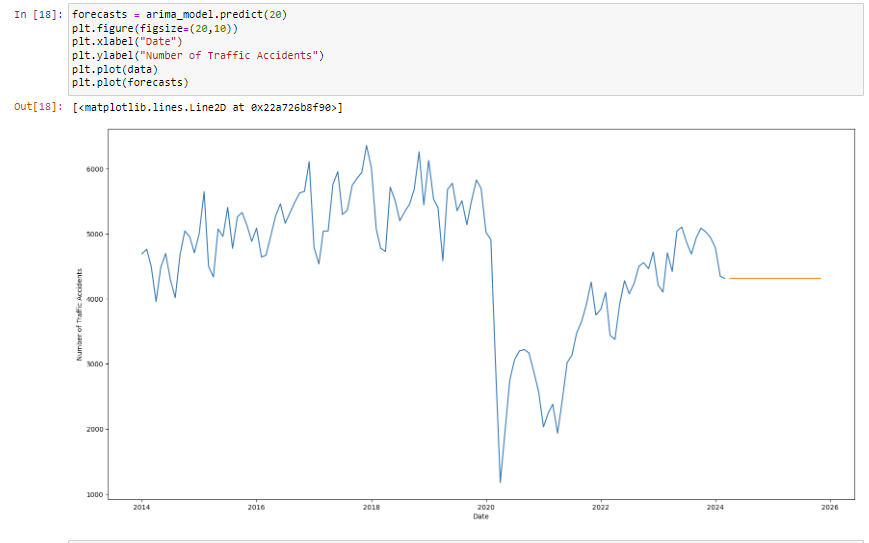
The above helps dig into a better understanding of Auto Regression and Moving Average. 1.00 is where we have the most significance.

Next Arima;



The P,D,Q helps us understand the differentiation, the lag order and the moving average. Leveraging the stepwise method, we are able to view statistics that help us understand our model. Overall, the model indicates that our series might be non stationary.

Finally the forecast;



That drop in 2020 might also be skewing our results, it might be better to take it out and look at the other years. The results might change. Due to covid, traffic was non existent so was accidents. However in 2024, we are almost back to pre-covid numbers and we can expect the numbers to follow the same pattern for the next few years.

1. **Dashboarding**

Dashboard 1: Historical Data

Map of Toronto Tableau inputs:

|  |
| --- |
|  |

Number of Fatalities Tableau inputs:

|  |  |
| --- | --- |
|  |  |

Types of Collisions over the Seasons Tableau inputs:

|  |
| --- |
|  |

Types of Collisions over the Months Tableau inputs:

|  |
| --- |
|  |

Neighbourhood Table Tableau inputs:

|  |
| --- |
|  |

Final Dashboard:

|  |
| --- |
|  |

Dashboard 2: Forecasting Analysis

Significant graphical results taken from the forecasting section were combined to create the second dashboard using the image and text features on Tableau.

Final Dashboard:

|  |
| --- |
|  |