1.0 Introduction

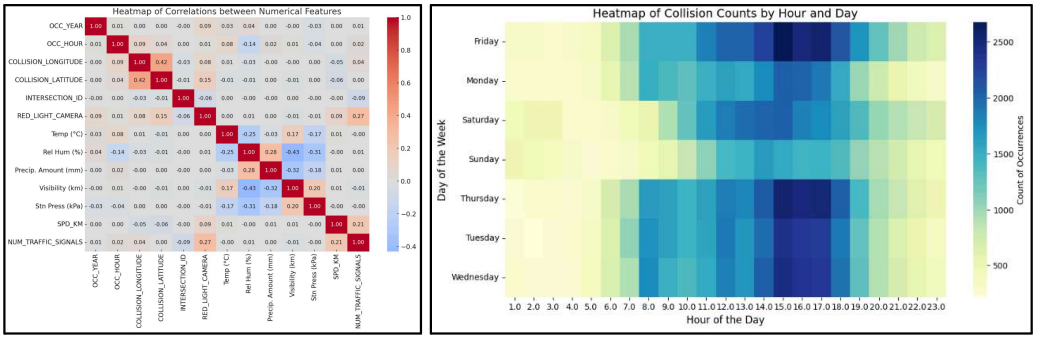
Traffic collisions remain a significant public safety concern in Toronto. According to a report produced by the Toronto Police Service, from the year 2022 to 2023, 600 people were killed or seriously injured as a result of motor vehicle collisions [1]. Studies highlight that timely emergency medical services (EMS) response plays a critical role in mitigating the severity of collision outcomes and reducing fatalities [2]. Given this, identifying the contributing factors to motor vehicle collisions and exploring effective mitigation strategies is crucial for improving road safety. This project will analyze traffic and EMS data to identify factors increasing traffic collision risk and determine how optimizing EMS station locations can reduce response time and minimize accident severity. By gaining these insights, our team seeks to develop evidence-based solutions to enhance road safety in Toronto.

This project addresses two key objectives: understanding the risk of collisions through predictive modeling of contributing factors and minimizing collision impact by optimizing EMS response times. Considering this scope, the aim of this project is to explore the following questions: (1) Based on the features available and the data present, how can we predict the risk of collisions? (2) Which features have the highest importance and relevance when determining the risk of a collision? (3) How can we optimize EMS placements to minimize the distance to high-risk collision areas, thereby decreasing response times?

2.0 Data

For this project, data was merged from multiple sources including the City of Toronto, Environment Canada, and reputable traffic sources to create a single comprehensive dataset consisting of 183,617 rows and 23 columns for the analysis. Historical collision data from Toronto Open Data Catalog for traffic collisions provided details on each incident such as accident location (latitude, longitude), date, time of day, and road factors (speed limits, traffic signals, and intersection type) [3]. Weather data from Environment Canada (i.e. temperature, precipitation, and visibility) was integrated to assess the impact of weather conditions on accident occurrences [4]. This combined dataset was merged on common factors (intersection ID, location, datetime) and resulted in a final set of 13 numerical and 10 categorical features. Current ambulance stations [5] and city-owned properties (i.e. fire stations) [6] were used to inform the current and candidate location dataframes in the optimization model (see Section 3.3 for details).

Initial Exploratory Data Analysis (EDA) revealed missing values in the final dataset. These were addressed by imputing values using patterns from historical data (i.e. seasonal or temporal patterns) or completely removing entire rows where appropriate to preserve the integrity of the dataset. Figure 1 shows a correlation heatmap, where no significant correlations between features were found. In contrast, weekday and time had a significant impact on collision frequency, as shown in Figure 2. Additional EDA results are presented in Appendix A.



Figures 1 & 2: Correlation Matrix (left) and Heatmap of Collision by Week and Hour of Day (right)

3.0 Methods

Three prediction models were used to estimate risk at collision locations: Kernel Density Estimation (KDE), Logistic Regression (LR), and Random Forest Classifier (RFC). KDE is an appropriate method for estimating traffic accident risk as it can effectively visualize the density of accidents across a geographic area [7]. The LR and RFC models were more suitable for assessing the impact of non-spatial features on collision risk, and to estimate the collision risk probabilities.

The optimization model determined the optimal placement of ambulances using the risk level as a weight factor, by minimizing the distance from ten thousand randomly sampled unique intersections from the dataset. To inspire the construction of the optimization model, academic publications were used to find a variety of models, including the maximal covering location problem, which maximizes demand coverage [8][9]. The number of new stations to place was determined as a pilot value and can be adjusted in alignment with the City of Toronto's budget constraints. Gurobipy was used to solve this optimization problem. The results will: (1) guide the optimal placement of ambulances, and (2) allow prediction of future accident risk based on input factors.

3.1 Prediction Models - Kernel Density Estimation

To identify high-risk collision areas, a multi-model approach was employed incorporating spatial and non-spatial data analysis techniques. KDE was used to analyze the spatial distribution of accidents and identify areas with high concentrations of collisions as displayed in Figure 3. Collision density values were estimated based on the latitude and longitude of collision points. This prediction model is nonparametric, meaning it does not make assumptions about the underlying distribution of the data, making it flexible and robust [7]. The model utilized a Gaussian kernel and an initial bandwidth of 0.02. Smaller KDE bandwidths capture finer details but can lead to noise, while larger bandwidths smooth the data and may mask important patterns. The bandwidth of 0.02 was selected to account for this bias-variance tradeoff. The Gaussian kernel was selected, as it is a common kernel choice [10]. This spatial analysis provided a valuable understanding of the geographical patterns of accidents.

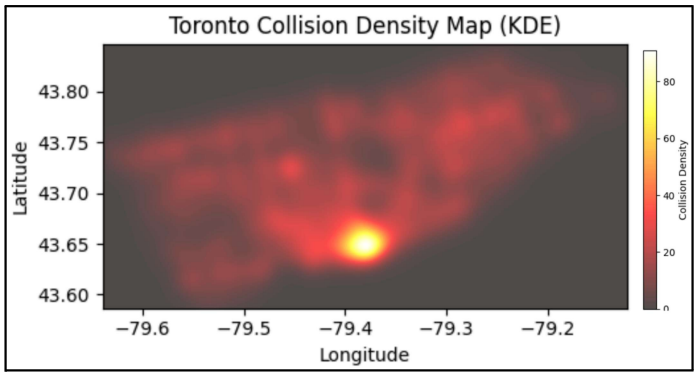


Figure 3: Toronto Collision Density Map produced by Kernel Density Estimation

3.2 Prediction Models - Logistic Regression and Random Forest Classifiers

To account for non-geographical features and improve predictions, LR and RFC were applied. LR offered a clear interpretation of each feature's influence on collision risk. RFC captured non-linear patterns in the data and evaluated feature importance. Both models produced collision risk probabilities, which were inputted into the optimization model as the risk constant. Evaluation of the models was based on the optimization model's performance on unseen data (i.e. the outputted optimal solution) using risk determined by the prediction models. This allows for direct comparison with KDE.

To prepare the categorical data for input into the LR and RFC models, one-hot encoding was used. This technique converts categorical variables into numerical representations, where each category was represented by a binary column indicating its presence or absence. This transformation allowed the models to effectively process and interpret the categorical information. Additionally, continuous variables within the training dataset were standardized to ensure all features had a similar scale. This prevents features with larger magnitudes from dominating the model's predictions, thus allowing models to learn feature importance more accurately.

Positive data samples were randomly selected from the training set and for each of these samples, either the date, time, or location were altered randomly. These sample points were then classified as negative samples if no collisions were associated with them. The ratio of negative to positive samples was set as 2:1 based on the optimal ratio from similar studies [11].

3.3 Optimization Model

The optimization model aims to minimize the response time to collisions by strategically placing additional ambulances within the city. To inform ambulance placement locations, a dataset consisting of current ambulance placements and a separate dataset consisting of fire station locations were used. Fire station locations served as candidate locations for new ambulance placements, because they are categorized as city-owned property. Each prediction model returned an array of risk values for ten thousand randomly sampled unique intersections, serving as the demand points for the optimization model. The optimization model considers the Haversine distance between demand locations and both existing and candidate ambulance locations. The Haversine distance (straight-line distance on a spherical surface) was used in place of the Euclidean distance to account for the Earth's curvature.

The objective function minimizes the weighted sum of distances between potential collision locations and their nearest ambulance, with weights determined by the predicted risk of each location.

$$\min_{y,z} \quad \sum_i R_i \sum_j z_{ij} D_{ij}$$

In order to describe the model, the following notation is used; R_i for the risk weight derived from the KDE, LR, or RFC output at location i, from the set of demand locations L; D_{ij} for the distance between location point i and candidate ambulance location j, where j is from the set of locations A; x for the fixed number of ambulances to dispatch; y_j for the binary decision variable, equal to 1 if and only if an ambulance is located at potential deployment location j; z_{ij} for the second continuous decision variable, equal to 1 if and only if ambulance j is assigned to demand location i, where each demand location is assigned to at most one ambulance.

$$\begin{aligned}
\text{s.t.} \quad & \sum_j z_{ij} = 1 \quad \forall i, \\
& \sum_j y_j \leq x, \\
& z_{ij} \leq y_j \quad \forall i, j, \\
& y_j \in \{0, 1\}, \\
& z_{ij} \geq 0 \quad \forall i, j.
\end{aligned}$$

The model constraints ensure that the number of deployed ambulances does not exceed the available fleet size. The binary decision variable (y_j) is used to indicate whether an ambulance is placed at a particular location. The third constraint ensures that ambulance *j* is assigned to demand location *i* if and only if there is an ambulance located at *j*.

3.4 Sensitivity Analysis

The goal of sensitivity analysis was to determine the optimal number of ambulances based on the diminishing returns observed in the evaluation metrics. The optimization model was run with the constraint, where *x* is the proposed number of candidate locations:

$$\sum y_j \leq x, \; for \; x \in [2, 72]$$

The *x* value was incremented until the improvement in the evaluation metric (see Section 4.0) was less than one percent of the metric value for the previous *x*, indicating further increases in ambulances yield minimal benefits. For each increment in *x*, the evaluation output was recorded in Appendix B. This was repeated for all three prediction models.

4.0 Results

The evaluation method for the three predictive models assessed their performance by leveraging the optimal solution determined from each model's risk predictions. The output metric (in kilometers) was the total sum of distances from approximately 37,000 unseen data points to the nearest ambulance bay (including both existing and proposed locations), and a smaller value indicated better performance.

$$\sum_{i \in L} \left(\min_{j \in A} D_{ij} \right)$$

- L is the set of collision points.
- A is the set of ambulance locations.
- D_{ij} represents the distance between collision point i and ambulance j.
- $\min_{j \in A} D_i$ selects the smallest distance between point i and any ambulance j.

The ambulance placement recommendations for LR and RFC, both of which incorporated non-spatial features, have overlaps in their optimal locations, while KDE suggested different results. LR was the most effective at 5 additional ambulance placements with an estimated 3498 km improvement when compared to the original scenario. Random Forest and Kernel Density Estimation performed best at 2 additional placements, reducing 1304 and 930 kilometers respectively (see Table 1).

Table 1: Benchmark Comparison of Evaluation Metric for Prediction Models

Model	Num. of Ambulances	EMS Coordinates (lon, lat)	Evaluation Metric
Original	N/A	N/A	53621 km
Kernel Density Estimation	2	(-79.430752, 43.680105) (-79.404782, 43.656825)	52691 km
Random Forest	2	(-79.502938, 43.628019) (-79.441141, 43.694520)	52317 km
Logistic Regression	5	(-79.502938, 43.628019) (-79.441141, 43.694520) (-79.346726, 43.745814) (-79.402157, 43.726217) (-79.227788, 43.764644)	50123 km

Minimal improvements occurred when adding the third ambulance for KDE and RFC, and the sixth ambulance for LR (see Figure 4).

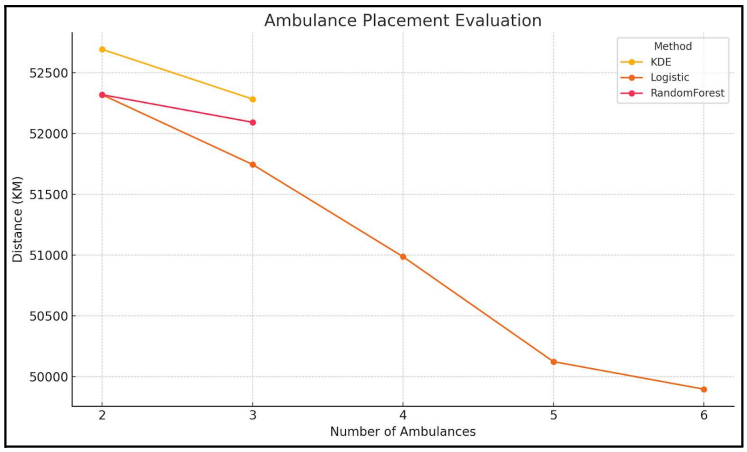
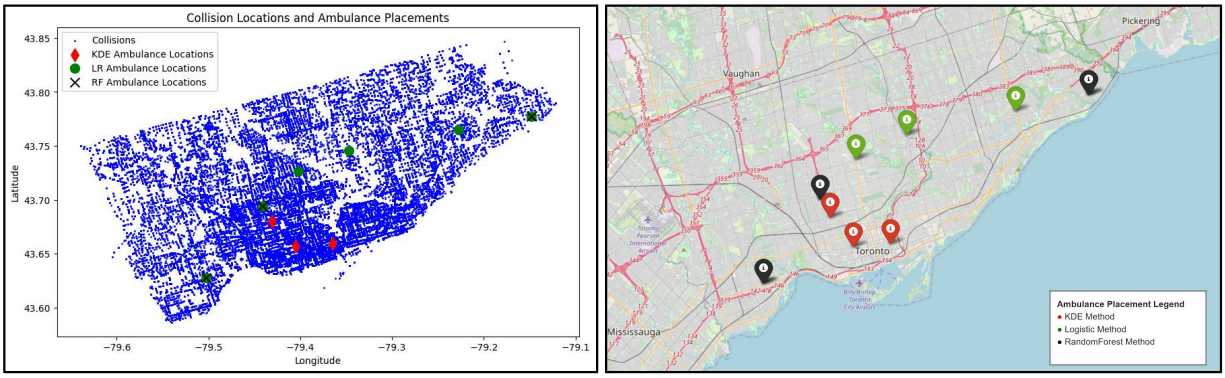


Figure 4: Evaluation Metric vs Number of Ambulances

LR identified time as the most important factor based on the absolute coefficient values, specifically during the periods from 12 to 6 PM, along with the presence of zero traffic signals within 100 m of the collision site. RFC, which captured non-linear relationships, determined the top 10 important features as follows: coordinates, temperature, pressure, relative humidity, visibility, precipitation amount, presence of one traffic signal within 100 m of the collision site, and the presence of a 40 km/h or 60 km/h speed limit.

5.0 Discussion

The selected KDE ambulance locations are concentrated in Downtown Toronto, focusing on the areas with the highest density of accidents (see Figure 8). LR and RFC had overlapping locations laterally spaced across the city, following the flow of major highways such as Ontario Highway 401 and Gardiner Expressways (see Figure 9). The similarity in results from LR and RFC may be due to both models accounting for non-spatial features and using the same training dataset, thus allowing them to identify overlapping high-risk areas despite differences in how they evaluate feature importance.



Figures 8 & 9: City of Toronto Collision & Ambulance location (left), Major Road Networks (right).

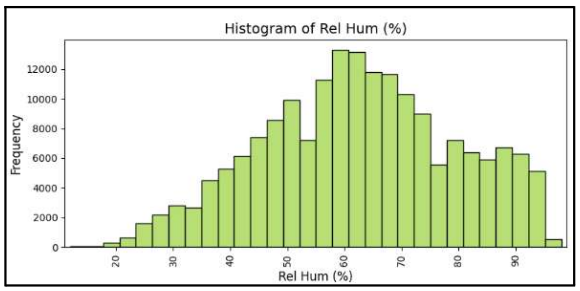


Figure 5: Frequency of Relative Humidity

LR emphasized the impact of the occurrence hour on collision risk, with 3 PM – 5 PM having the highest weighting and the features having a positive contribution to the risk. This aligns with patterns seen from EDA in Figure 5.

RFC resulted in a wider range of important features compared to LR as it captures non-linear patterns in the dataset. These features primarily focused on the weather conditions, indicating that they may have a non-linear relationship with collision risk, shown in Figure 6.

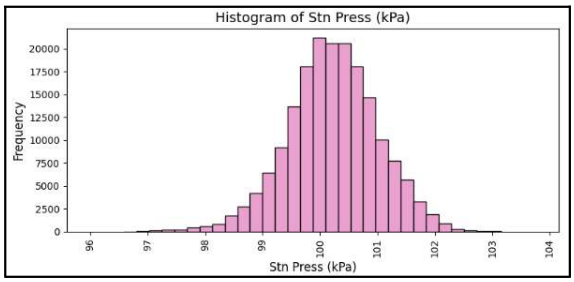


Figure 6: Frequency of Air Pressure

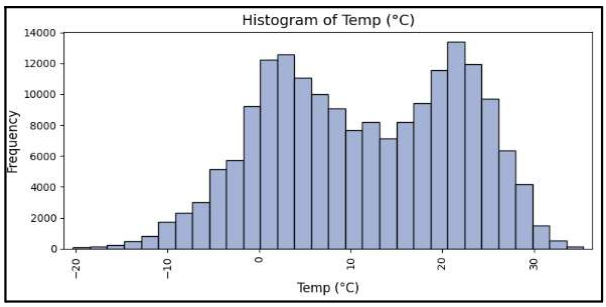


Figure 7: Frequency of Temperature

While KDE effectively identified collision hotspots, the lack of consideration for non-spatial features, such as time of day, weather conditions, and road factors. These features showed important relationships in the data as shown in Figure 7, thus limiting KDE's precision of placement recommendations.

Additionally, the influence of time, weather, and traffic signals, which were determined as important features by the LR and RF models, shows opportunity for policy interventions and infrastructure changes. Traffic signal timing can be adjusted during periods of high risk (i.e. 3 PM – 5 PM) to improve safety and congestion. Infrastructure improvements such as fog warning systems can tackle heightened risk of collisions from higher levels of humidity, which contributes to fog and reduced visibility [12].

There are several weaknesses present in the optimization and prediction models that can impact the derived insights. The use of Haversine distance to optimize ambulance placement underestimated response times, as it does not consider road networks. This affects the optimization model and the evaluation model outputs, both of which directly use the distance calculations. Random negative sampling introduced potential biases in the training data for both the LR and RF models, which likely impacts the reliability of risk predictions and the derived top 10 features. For example, EDA showed that the proportion of Cul de Sac-Single Level (CDSSL), or "dead end", intersections was less than 0.02% in the collisions dataset prior to sampling. However, this feature appeared in the top 15 most important features selected by the LR model out of 109 features, which suggests that the sampling process may have overrepresented certain rare features.

Additionally, data biases are present in features such as the speed limit, which is primarily represented by 40 km/h and 60 km/h. A similar issue occurs with visibility, precipitation amounts, and number of traffic signals within a 100-meter radius. These biases can skew the model's understanding of how these factors relate to collision risk, thus impacting the accuracy of risk predictions.

6.0 Conclusion

Three primary goals guided this report on road accidents within the City of Toronto: (1) predicting collision risk (2) gaining an understanding of the most influential factors and (3) mitigating collision impact through optimizing ambulance placement. Datasets from the City of Toronto, Environment Canada, and additional traffic sources were collected and analyzed for feature impact on collision frequency. Once compiled, a portion of the dataset was used to train three predictive models: Logistic Regression (LR), Random Forest Classifiers (RFC), and Kernel Density Estimator (KDE). LR is a linear model that is useful for identifying key predictors, while RFC captures nonlinear relationships between variables, and KDE estimates continuous collision density values based on location. The optimization model returned optimal ambulance locations that minimized the weighted distance to collision demand points, where weights were determined by risk factors outputted by each prediction model.

Leveraging these models to reduce the response time to potential crash sites, and boost survival rates, the results indicate: (1) The time of day played the most significant role in predicting collisions. The highest risk traffic periods often occurred between 3-5pm. (2) Non-linear relationships were evident in features such as temperature, humidity, and precipitation, aligning with our initial findings from EDA. (3) The locations suggested by LR were optimal, with 5 ambulances balancing the tradeoff between ambulance quantity and a reduction of the evaluation metric by 3498 km. Predictive modelling in combination with optimization can enhance post-collision response by identifying areas of high risk, contributing factors to collisions, and informing strategic placement of additional EMS services, thereby reducing the severity of impact and servicing communities by maximizing coverage with limited resources.

7.0 Future Directions

Due to resource constraints, such as computing power, insufficient data, and timing limitations, further iteration is needed to strengthen the accuracy and robustness of current results. Thus, this project's future direction involves enhancing initial findings through several key areas:

1. **Improving Negative Sample Accuracy:**
    - Incorporating real-world traffic volume data to better represent actual traffic patterns and their impact on collision likelihood.
    - Considering seasonal variations, weekday/weekend differences, and the effect of holidays to refine the model's accuracy.
2. **Additional Model Tuning:**
    - Optimizing the KDE model's bandwidth, grid resolution, and kernel type through hyperparameter tuning to enhance collision hotspot identification accuracy.
    - Perform hyperparameter tuning for the LR and RFC models, exploring techniques such as Grid Search or Random Search to identify the optimal parameters (i.e. regularization strength, n_estimators, max_depth).
    - Utilizing real-world road network data, rather than the current Haversine distances, to more accurately model EMS response times in the optimization model.
3. **Considering Alternative Optimization Approaches:**
    - Exploring the impact of redistributing existing ambulance locations as an alternative project scope.

By addressing these areas, the goal is to further enhance the accuracy of collision prediction models and the effectiveness of ambulance placement strategies, ultimately leading to improved emergency response times and reduced collision severity.

References

[1] Transport Canada, "Canadian Motor Vehicle Traffic Collision Statistics: 2022," Transport Canada, May 02, 2024. https://tc.canada.ca/en/road-transportation/statistics-data/canadian-motor-vehicle-traffic-collision-statistics-2022

[2] J. P. Byrne et al., "Association between emergency medical service response time and motor vehicle crash mortality in the United States," JAMA Surgery, vol. 154, no. 4, p. 286, Apr. 2019. doi:10.1001/jamasurg.2018.5097

[3] "Total KSI," Toronto Police Service Public Safety Data Portal, https://data.torontopolice.on.ca/pages/total-ksi

[4] E. and C. C. Canada, "Daily Data Report for October 2022 - Climate - Environment and Climate Change Canada," climate.weather.gc.ca, Oct. 31, 2011. https://climate.weather.gc.ca/climate_data/daily_data_e.html?StationID=51459

[5] "Ambulance Stations Locations" open.toronto.ca, 2024. https://www.google.com/url?q=https://open.toronto.ca/dataset/ambulance-station-locations/&sa=D&source=docs&ust=1733175777767057&usg=AOvVaw3oOYxmBVRHr4oOkEqT_KaH

[6] "Fire Station Locations," open.toronto.ca, 2024. https://open.toronto.ca/dataset/fire-station-locations/

[7] N. Z. Zhang, N. Y. Liu, N. B. Chen, and N. K. Chen, "Using GIS and KDE analysis spatial distribution on public housing households: A case study," Apr. 2013, doi: https://doi.org/10.1109/iccse.2013.6554044.

[8] S. E. Hashemi, M. Jabbari, and P. Yaghoubi, "A mathematical optimization model for location Emergency Medical Service (EMS) centers using contour lines," Healthcare Analytics, p. 100026, Feb. 2022, doi: https://doi.org/10.1016/j.health.2022.100026.

[9] Edoardo Fadda, D. Manerba, and R. Tadei, "How to locate services optimizing redundancy: A comparative analysis of K-Covering Facility Location models," Socio-Economic Planning Sciences, vol. 94, pp. 101938–101938, May 2024, doi: https://doi.org/10.1016/j.seps.2024.101938.

[10] J. Heer, "Fast & Accurate Gaussian Kernel Density Estimation." Available: https://idl.cs.washington.edu/files/2021-FastKDE-VIS.pdf

[11] P. Way, J. Roland, M. Sartipi, and O. Osman, "Spatio-Temporal Crash Prediction: Effects of Negative Sampling on Understanding Network-Level Crash Occurrence," Transportation Research Record: Journal of the Transportation Research Board, vol. 2675, no. 6, pp. 225–234, Feb. 2021, doi: https://doi.org/10.1177/0361198121991836.

[12] N. US Department of Commerce, "How Fog Forms," National Weather Service, https://www.weather.gov/lmk/fog_tutorial

Appendix

Appendix A: Additional Exploratory Data Analysis

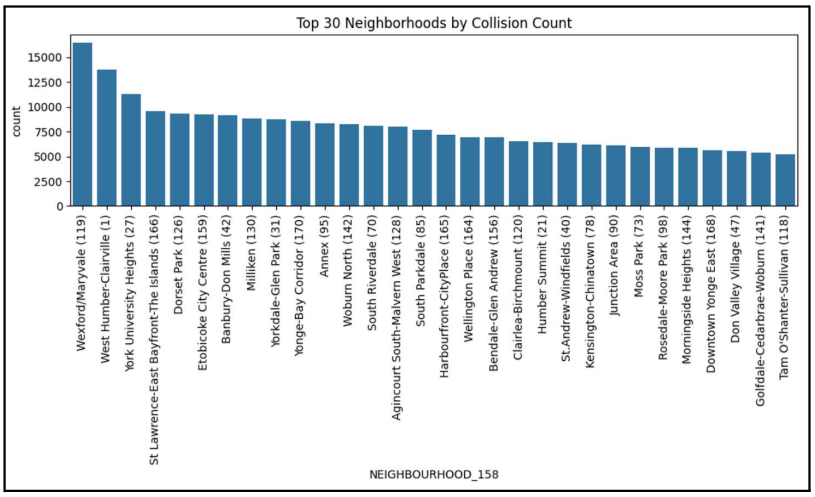


Figure 1: Histogram of Top 30 Neighbourhoods by Collisions Count

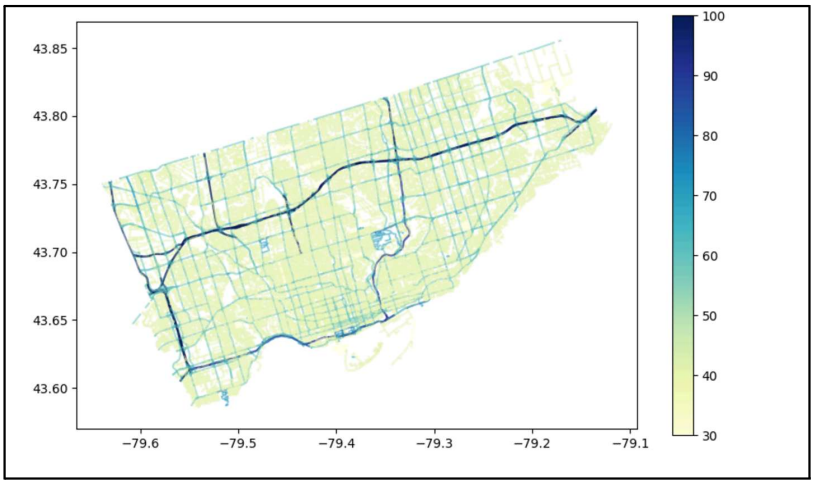


Figure 2: Speed Limit Choropleth

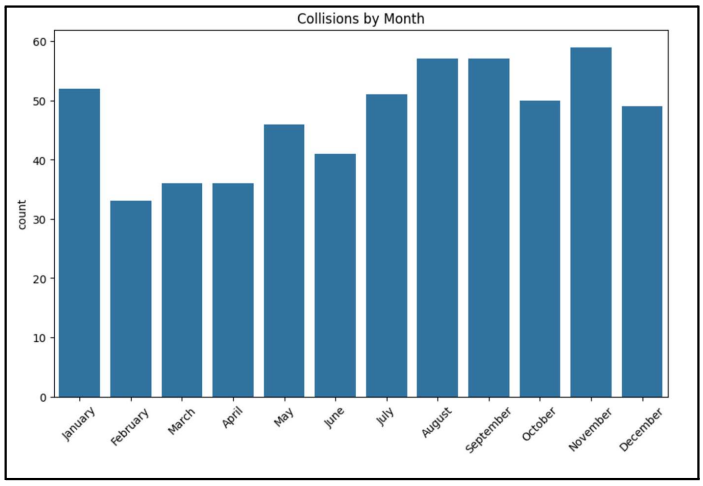


Figure 3: Histogram of Collisions by Month

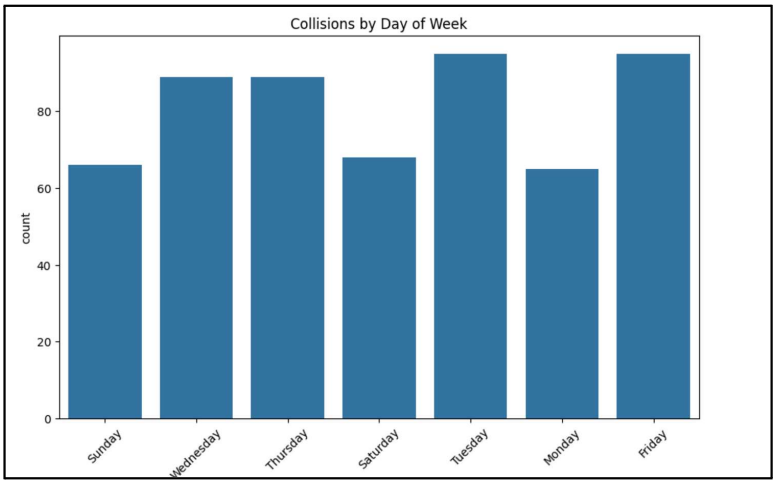


Figure 4: Histogram of Collisions by Day of Week

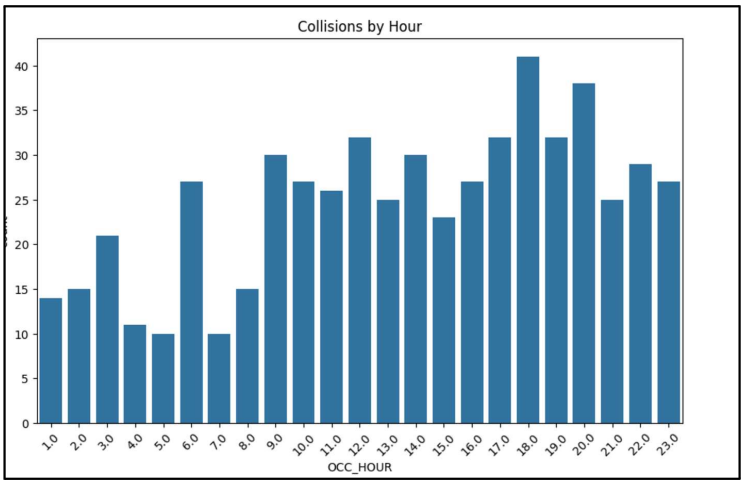


Figure 5: Histogram of Collisions by Hour

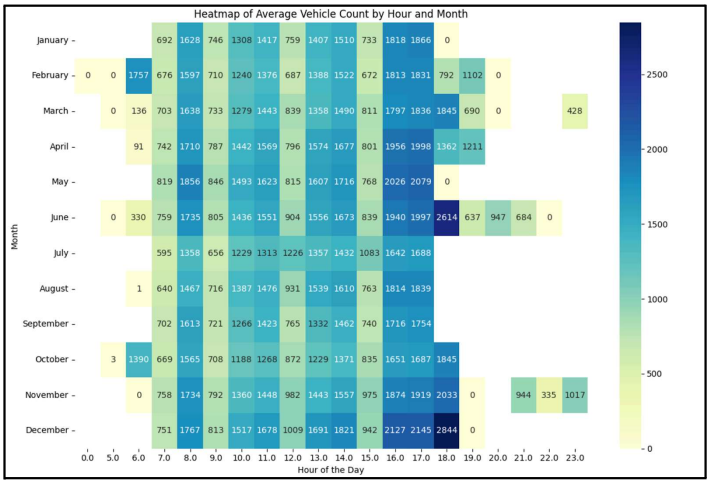


Figure 6: Heatmap of Traffic Volume Data