Analysis to Determine

Impact of Venues on Collision Occurrence in the City of Ottawa

Sonia Singh
March 26, 2020

Image: Canadian Flag projected on the city of Ottawa's Parliament Building

Abstract

This report summarizes a data science-based analysis of Ottawa collisions to determine whether or not there is a co-relation between collision occurrence and nearby venues.

A co-relation would allow city planners to take these into consideration when approving and designing roads and intersections near specific types of venues.

Table of Contents

1 Introduction

1.1 Background

Collisions represent a significant impact in major cities and a large number of studies have been undertaken to analyze causes of collisions with a view towards actions that can be taken to reduce the number of collisions.

Many of these studies observe that collisions are impacted by a variety of factors including but not limited to:

- Road conditions
- Weather conditions
- Traffic volume
- Driver behavior
- Driver distractions

Some of the factors are relatively well studied, and it is understood that driver behavior and distracted driving are strong contributors towards collision occurrences.

1.2 Gap

Amongst published papers, there is currently minimal analysis of the potential impact of nearby venue types such as restaurants, coffee shops, bars, or stores on collision occurrences.

Given that distracted driving is considered to be a strong factor contributing to collisions, the types of venues in specific locations could be impacting driver distraction.

My hypothesis is that the number of collisions at each location have a reasonable probability of being co-related to the number of types of venues at each location. For example, the number of venues related to food, liquor, or shopping could impact driving behaviors such as individuals being distracted or being in a rush due to their visits to or from these venues.

1.3 Focus of this Analysis

Collisions in the Ottawa area will be analyzed to determine whether or not such a co-relation exists. If this analysis indicates that collision occurrences are impacted by nearby venues, this will be useful for city planners when approving and designing roads and intersections near specific types of venues.

2 Data

2.1 Data Sources

The analysis requires both collision data and venue data which was obtained from the following sources:

1. The City of Ottawa's open data site includes 2018 tabular transportation collision data. The dataset includes a tabular list of all the collisions that occurred within the Ottawa area in 2018.
    a. Accuracy of the dataset:
        - Each reportable collision occurring on public roadways is sent to the City of Ottawa and is validated at least once.
        - Approximately 50% of the records are validated once again by a senior staff. Additionally, many queries are run on the database looking for errors.

    b. Features included in the dataset:
        - X and Y coordinate format is projected in MTM Zone 9, NAD83 (CSRS)
        - KML and CSV/XLS formats are projected in latitude, longitude (WGS84)
        - Date
        - Time
        - Location description (RD1 @ RD2 or RD from RD 1 to RD 2)
        - Classification of collision (non-fatal, fatal, property damage only)
        - Collision location (Intersection, non-intersection, at/near private driveway)
        - Pedestrians involved
        - Road surface condition (Ice, wet, dry snow...)
        - Environment (Clear, rain, snow…)
        - Light (daylight, dawn, dusk…)
        - Initial impact type (Angle, turning movement, rear-end…)
        - Traffic control (stop, traffic signal, no control…)
        - Latitude and longitude

    c. Location of dataset:

- https://open.ottawa.ca/datasets/ec6f9c7d3e214fcebd8d97111f1804df_0?geometry=-77.066%2C45.217%2C-74.542%2C45.890

2. Foursquare API offers real-time access to their global database of rich venue data and can be used to extract venue data for specific geographical locations. The "explore" endpoint was used to extract a list of venues near each collision location.
   a. Features included in the dataset that were used for this analysis:
      - Venue name
      - Venue category
      - Venue location in latitude and longitude

3. GoogleMaps was used to obtain latitudes and longitudes for specific venues that were investigated.

2.2 Data Cleansing

Some data cleansing was required to ensure effective analysis.

- Collision data had a feature called "TRAFFIC_CONTROL_CONDITION" with missing data. Since this feature included a value of "00 – Unknown", all missing data was set to "00 – Unknown".
- Venue data had a feature called "Venue category" which had a high level of granularity, and included values such as "Restaurant", "Japanese Restaurant", "Noodle House", etc. This level of granularity was too high, so I created a new feature called "Venue type" to allow grouping at a lower level of granularity, such as grouping all of the above examples into "Eatery". A total of 238 venue categories were grouped into 16 venue types for ease of analysis.

2.3 Feature Selection

Collision data features included in the analysis:

1. The City of Ottawa's open data site includes 2018 tabular transportation collision data. The dataset includes a tabular list of all the collisions that occurred within the Ottawa area in 2018.
   - Location description (RD1 @ RD2 or RD from RD 1 to RD 2)
   - Classification of collision (non-fatal, fatal, property damage only)

- Collision location (Intersection, non-intersection, at/near private driveway)
- Pedestrians involved
- Road surface condition (Ice, wet, dry snow...)
- Environment (Clear, rain, snow…)
- Light (daylight, dawn, dusk…)
- Initial impact type (Angle, turning movement, rear-end…)
- Traffic control (stop, traffic signal, no control…)
- Latitude and longitude

2. Venue data features included in the analysis:
   - Venue type, which is derived from rolled up values of venue category
   - Latitude and longitude

2.4 How the Data is Used

The collision data includes 14485 incidents at 5585 locations. Since FourSquare places a limitation on the number of calls that can be made in a day and in an hour, a subset of the collision data is used for analysis instead of using the full dataset.

A subset of the collision locations close to the Ottawa city center are selected for analysis. For each of these locations, corresponding venue data are obtained to create a dataset that can then be analyzed.

3 Methodology

3.1 FourSquare Limitations

Since FourSquare has a daily and hourly limit on the API calls that can be made to obtain venue data, as well as the number of calls that can be made in burst mode, the first challenge that needed to be addressed was to find a workaround for this limitation.

I started by limiting the number of collision locations that would be included in the analysis to FourSquare's daily limit of 950, and extracted a subset of collision locations that were close to the Ottawa city center. However, I still kept running into failures, and after testing over multiple days (since the daily limit would halt progress till the next day), realized that FourSquare support would need to be contacted. They explained that "In addition to a daily call limit, we have a burst limit as well. Keep the calls to a maximum of 50 per second.”, which was not apparent from their online documentation.

After discovering that there is an hourly limit and a burst mode limit, I decided to further divide the selected data subset into smaller datasets so that I could split up the FourSquare calls. This was done through trial and error to come up with subsets as follows that were well within FourSquare’s daily limit of 950 and the hourly/burst limit:

```
<=.5 km (236, 37) 78
>.5 and <=.75 km (251, 37) 95
>.75 and <=1 km (324, 37) 142
>1 and <=1.25 km (375, 37) 124
>1.25 and <=1.5 km (279, 37) 122
>1.5 and <=2 km (366, 37) 167
>2 and <=2.5 km (494, 37) 202
```

Once I was finally able to make all the FourSquare calls to collect the necessary venue data, I saved my dataframes into external files so that I could avoid this step every time I needed to restart my notebook.

3.2 Analysis Approach

3.2.1 First Stage

My intention was to use K-Means to analyze the co-relation between different types of venues and the number of collisions at each collision location. With the results of this analysis, I would then explore further co-relations with additional features such as involvement of pedestrians, impact type, light conditions, and type of traffic control present.

After analyzing the K-Means results and exploring co-relation with various venue types, there did not appear to be a clear co-relation between the venue types and number of collisions at particular locations. In fact, there appeared to be a negative co-relation, i.e. the locations with higher number of collisions seemed to have less venues related to restaurants, bars, and shopping.

3.2.2 Second Stage

I then decided to use a different approach to determine if there was indeed any kind of co-relation between venue data and number of collisions nearby. Based on anecdotal observations of increased distracted driver behavior near popular drive-thru venues such as Tim Horton’s and McDonald’s, I conducted an analysis of collision locations in proximity to a Tim Horton's located at 2145 Robertson Rd, Nepean, ON K2H 5Z2. As a result of this second approach, I decided to forego the initial intent of analysis with features such as involvement of pedestrians, impact type, light conditions, and type of traffic control present.

3.3 Analysis of Collisions Within 2.5 Km to Ottawa City Center

3.3.1 Mapping of Selected Data

The following set of 890 collision locations are included in this analysis, with 2325 collisions in total, ranging from 1 to 31 at each location.

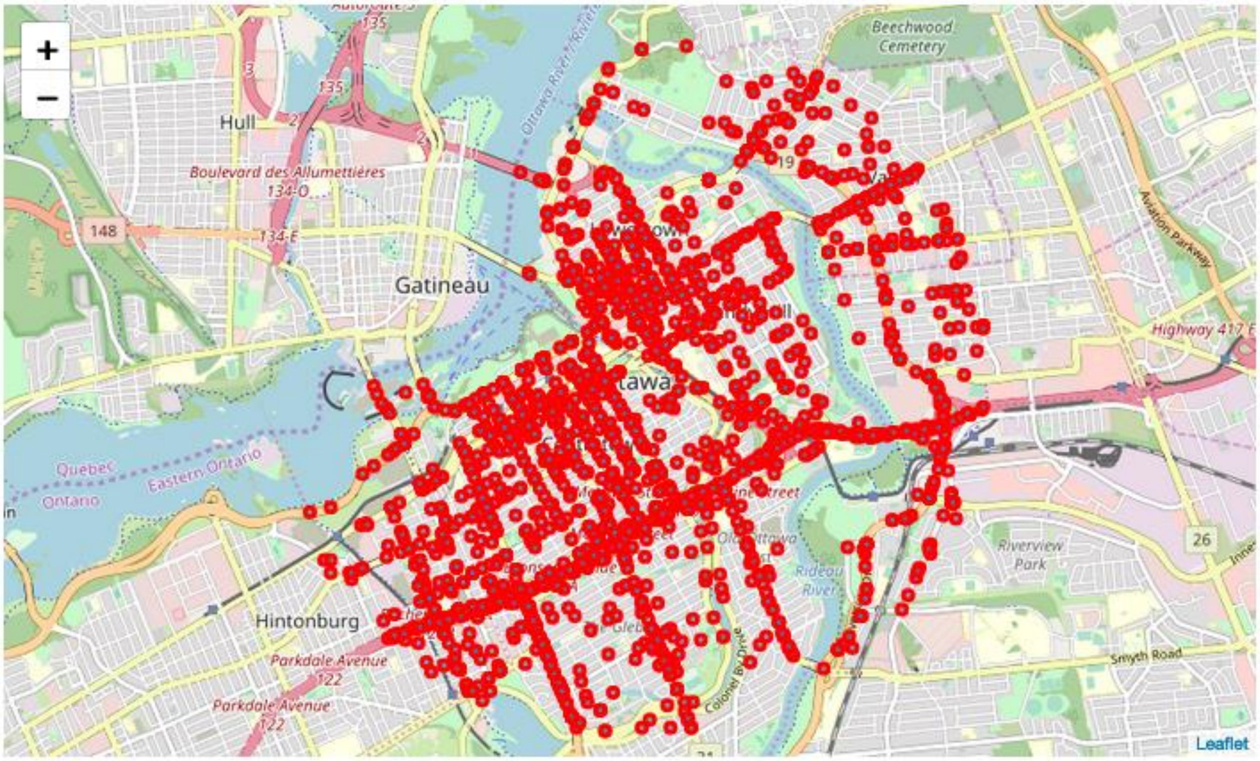

Figure 1 Map of collision locations within 2.5 km from Ottawa city center

3.3.2 K-Means Clustering

K-Means clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data and is particularly useful for unsupervised learning problems.

I applied K-Means clustering using the following features:

- Number of collisions

- Number of pedestrians
- Number of venues in each venue type. Note that as indicated in Section 2.1 Data Cleansing, 238 venue categories were grouped into the following 16 venue types:
    - Auto
    - Coffee
    - Eatery
    - Entertainment
    - Grocery
    - Hotel
    - Liquor
    - Music
    - Nature
    - Nightlife
    - Not Known
    - Other
    - Park
    - Shopping
    - Sport or Fitness
    - Transportation

To determine the ideal value of K, I compared the Squared Error or Cost for different values of K and used the “elbow” method to determine that 3 was the ideal value of K.

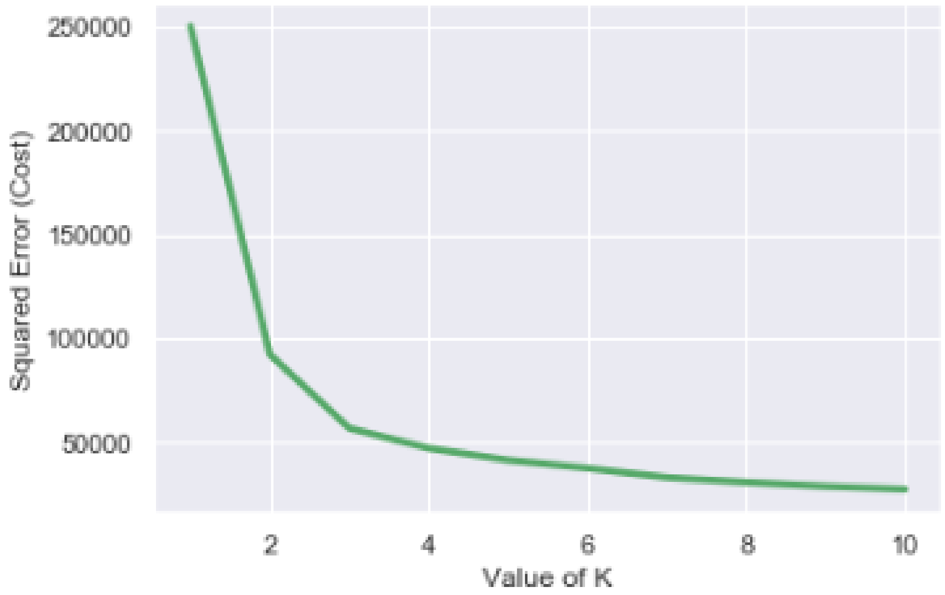

Figure 2 Compare the Squared Error(Cost) for different values of K

3.3.3 K-Means Clustering Grouping

The 3 K-Means clusters seem to be grouped as follows:

- Cluster 0: Medium number of Coffee and Eatery type of venues
- Cluster 1: Low number of Coffee and Eatery type of venues
- Cluster 2: High number of Coffee and Eatery type of venues

3.3.5 Visualization using Parallel Coordinates

The first visualization used all the features as follows:

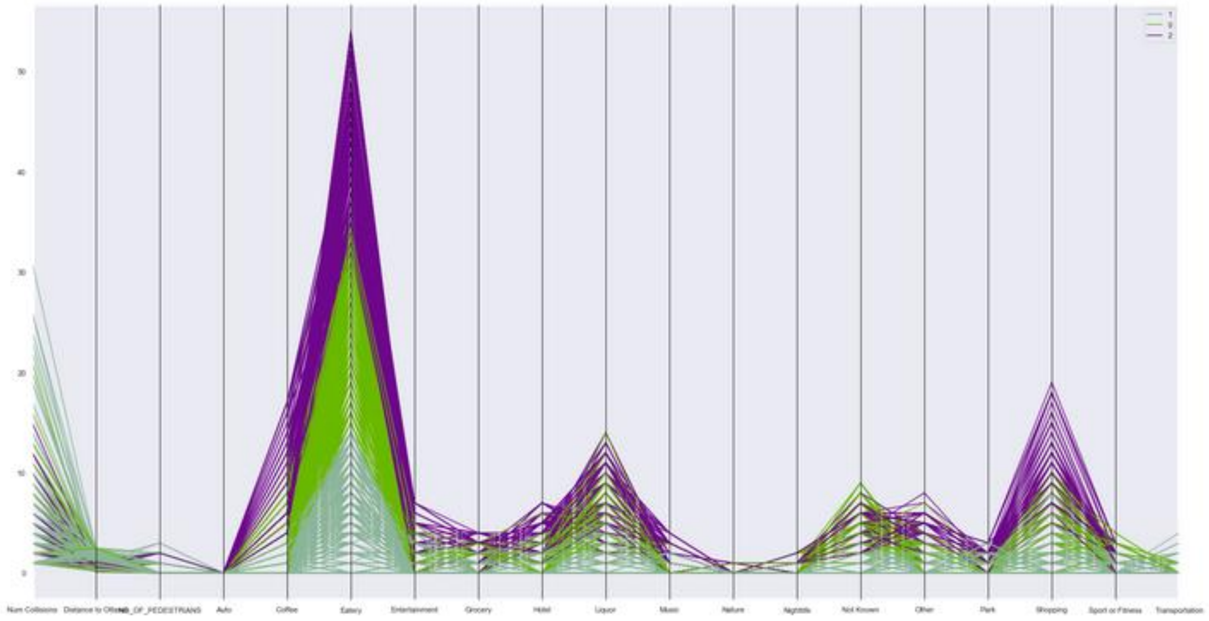

Figure 3 Analysis using Parallel Coordinates of K-Means results for all venue types

Since some of the venue types had minimal data, I removed these venue types and recreated the visualization to explore the results:

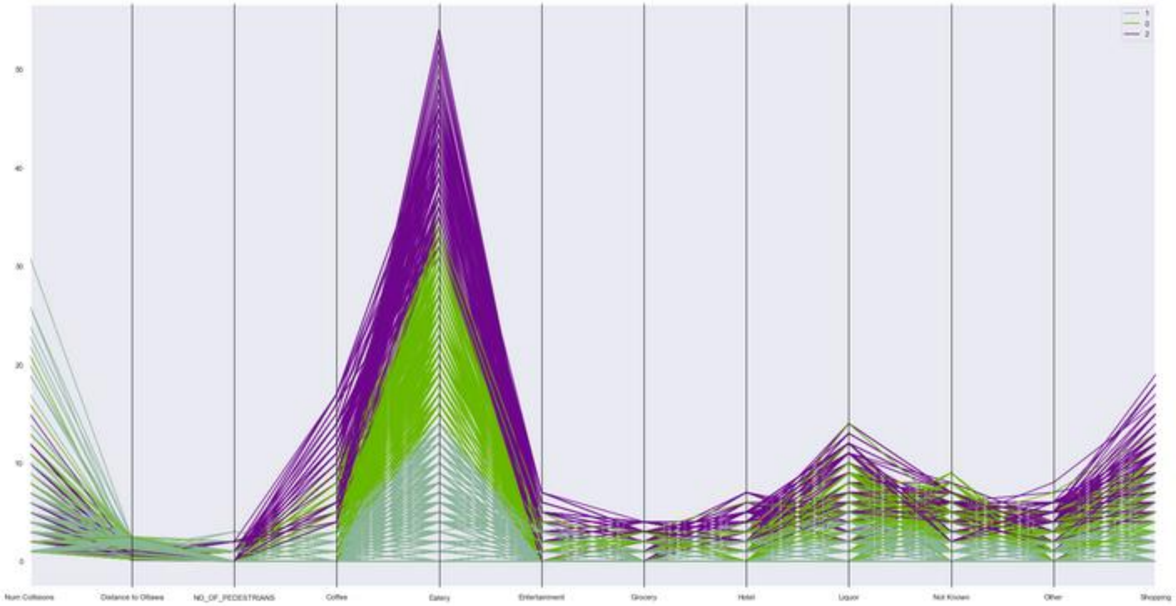

Figure 4 Analysis using Parallel Coordinates of K-Means results for more prominent venue types

There did not appear to be a clear co-relation between the venue types and number of collisions at particular locations. In fact, there appears to be a negative co-relation, i.e. the locations with higher number of collisions have less venues related to restaurants, bars, and shopping.

I then tried a deeper dive by analyzing the locations by breaking them down by number of collisions. These are the results for collisions > 15:

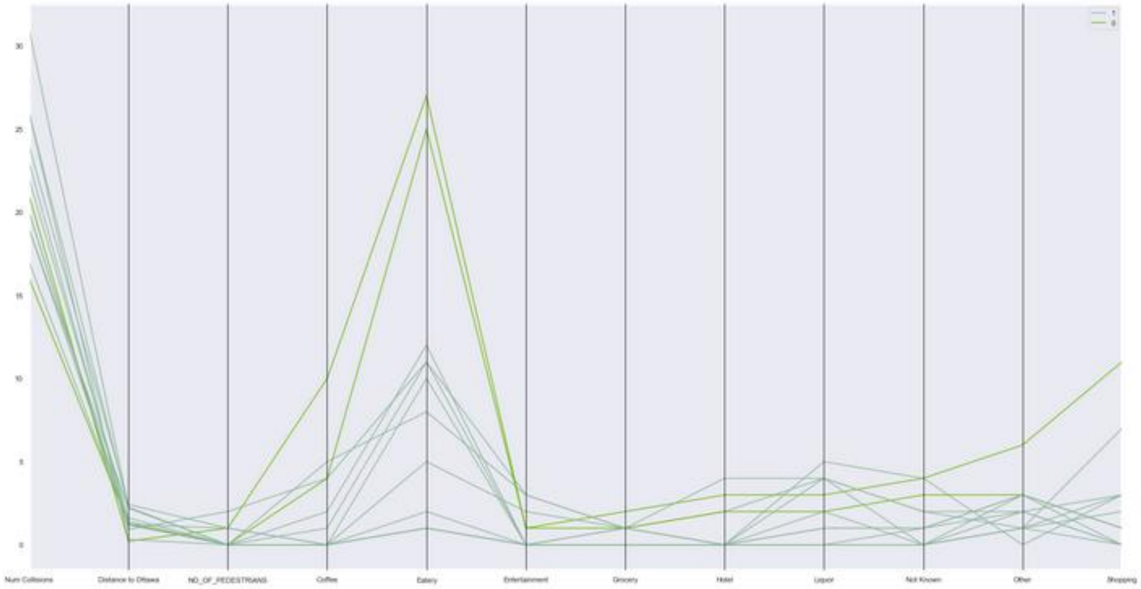

Figure 5 Analysis using Parallel Coordinates of K-Means results for more prominent venue types, where # of collisions > 15

These are the results for collisions <=10:

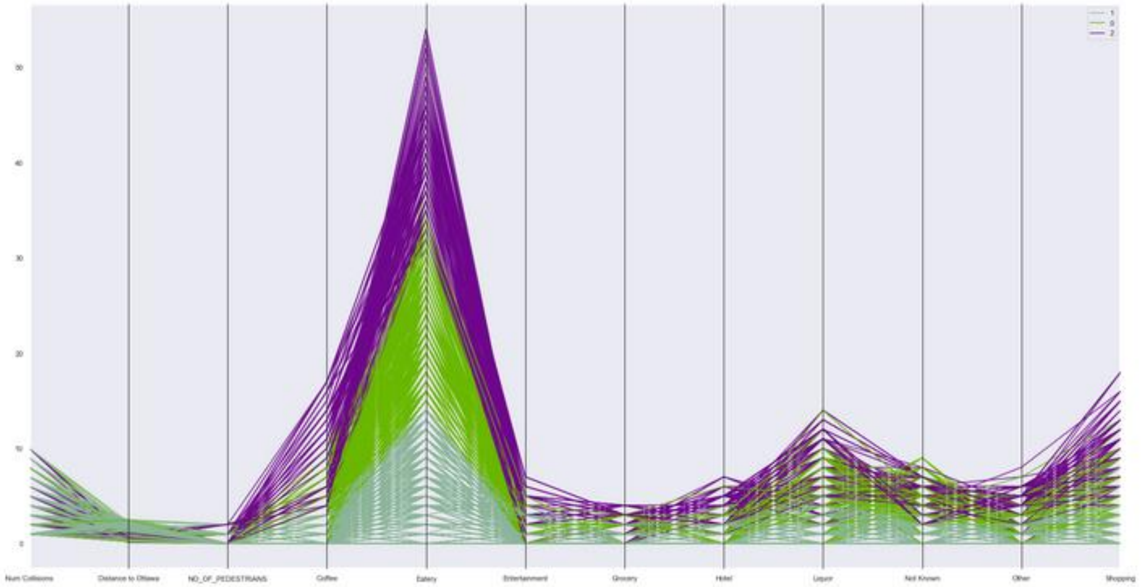

Figure 6 Analysis using Parallel Coordinates of K-Means results for more prominent venue types, where # of collisions <= 10

I then tried a deeper dive by analyzing the locations by breaking them down by involvement of pedestrians. These are the results for pedestrians > 1:

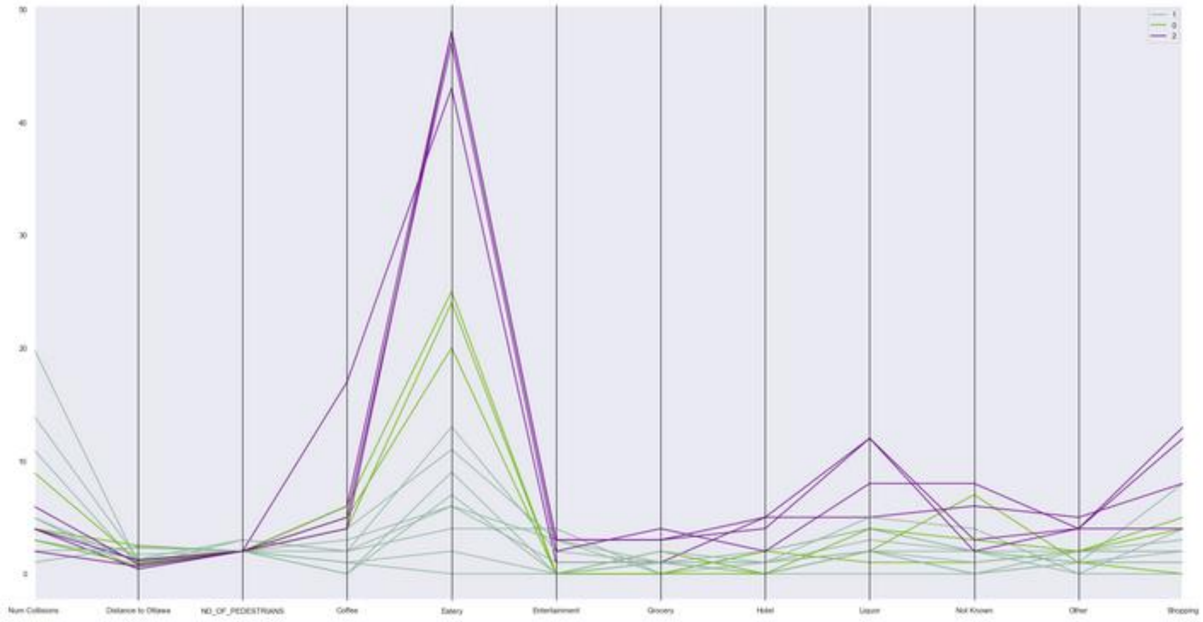

Figure 7 Analysis using Parallel Coordinates of K-Means results for more prominent venue types, where # of pedestrians > 1

These are the results for pedestrians <=0 :

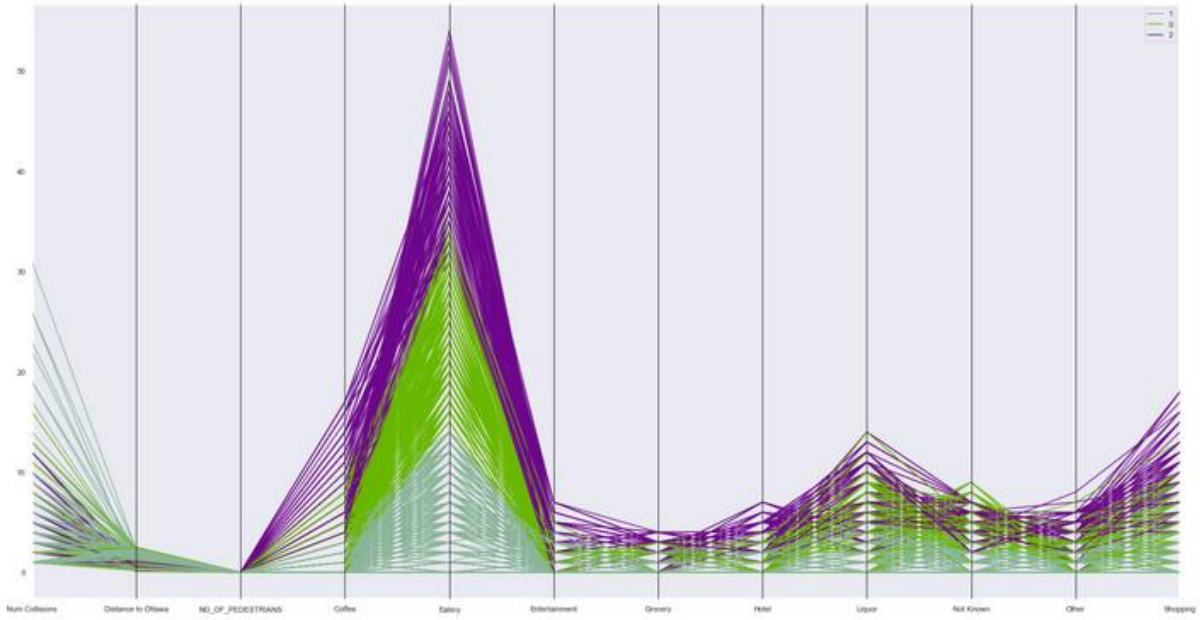

Figure 8 Analysis using Parallel Coordinates of K-Means results for more prominent venue types, where # of pedestrians <= 0

I was still unable to discern an clear co-relation that would support the initial hypothesis and tried analyzing using other visualization methods.

3.3.6 Visualization using Pair Plots, Scatter Plots, and Box Plots

In an attempt to determine if there were any other insights that could be gleaned from different ways of looking at the results, I tried a few other visualizations, with similar results.

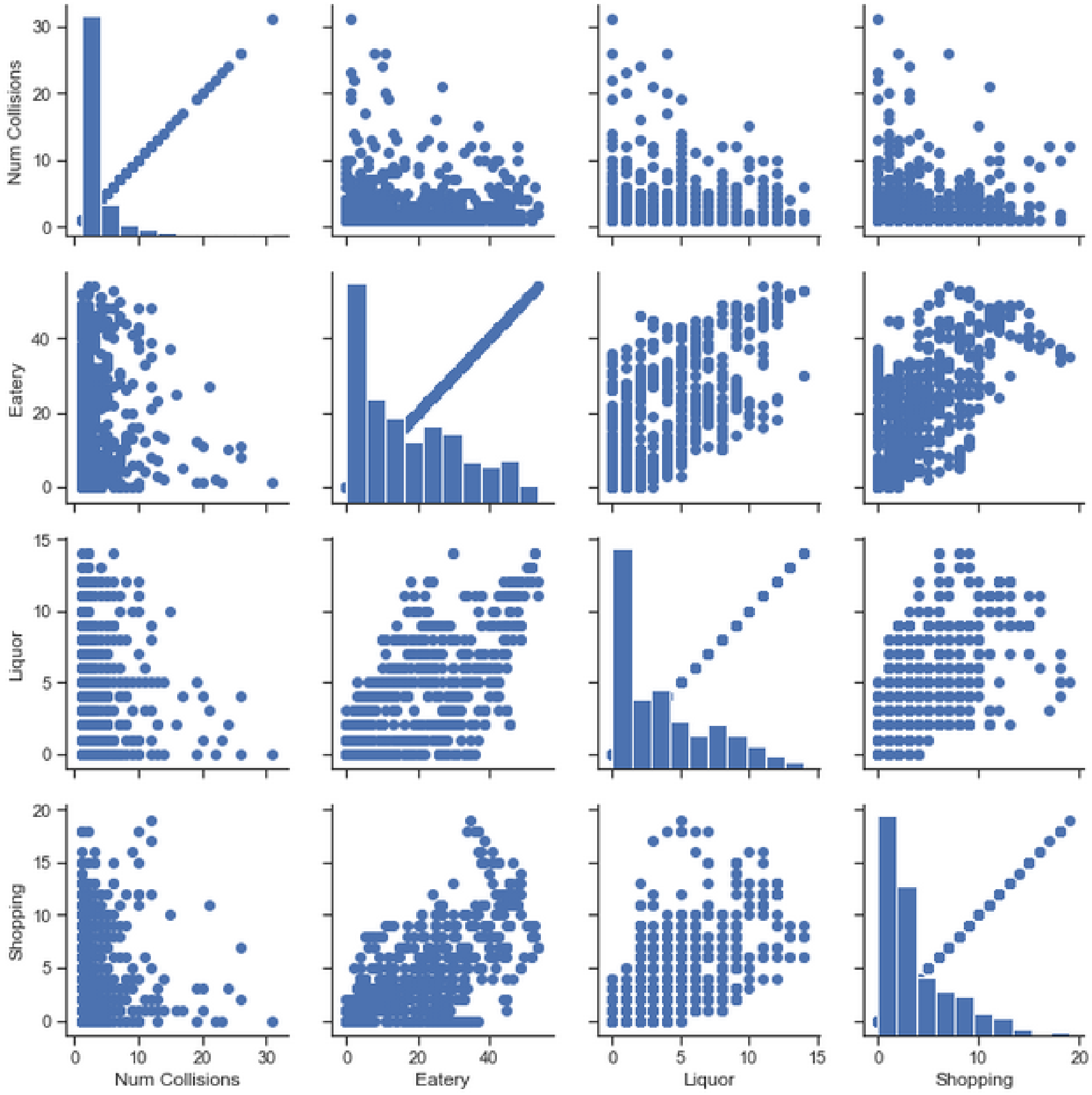

Figure 9 Analysis using Pair Plots, using Number of Collisions, against Venue Types: Eatery, Liquor, Shopping

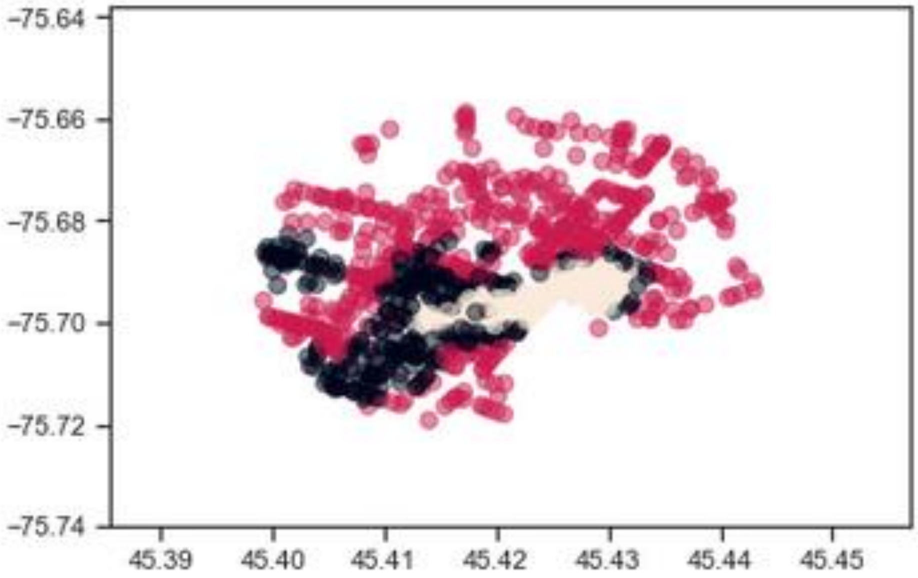

Figure 10 Analysis using a Scatter Plot, using K-Means Clusters against collision location latitudes and longitudes

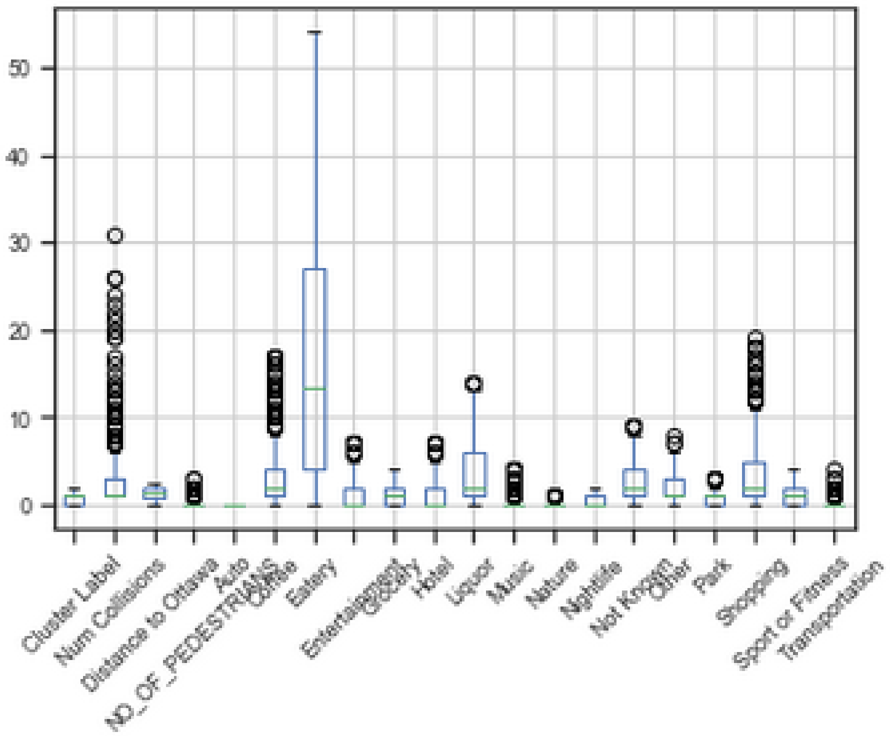

Figure 11 Analysis using a Box Plot, using Number of Collisions, against all Venue Types

3.3.7 Visualization using Heat Maps

As a last-ditch effort, I tried to see if Heat Maps would reveal any insights. These clearly show that the higher the number of collisions at a specific location, the less the number of eateries, liquor, shopping, or coffee related venues were close to those locations.

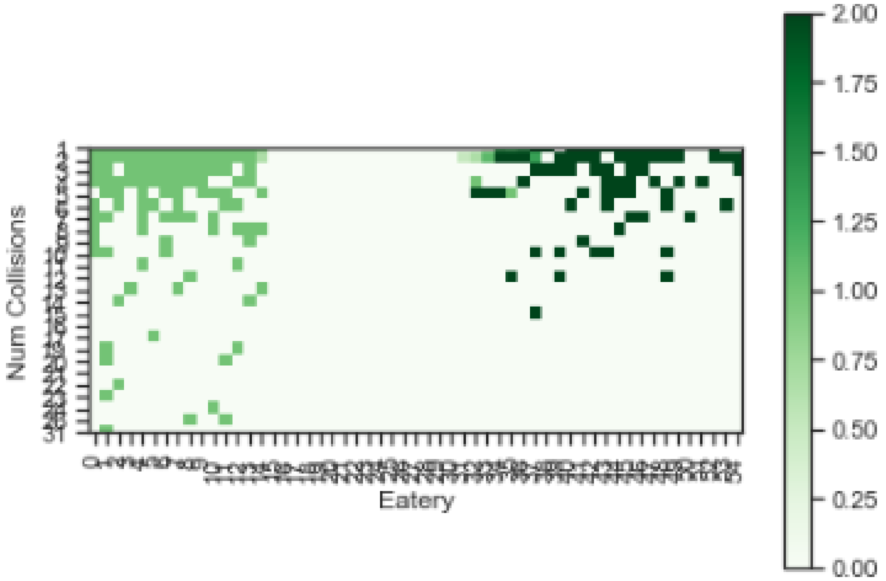

Figure 12 Analysis using a Heat Map to determine co-relation between number of collisions and eateries

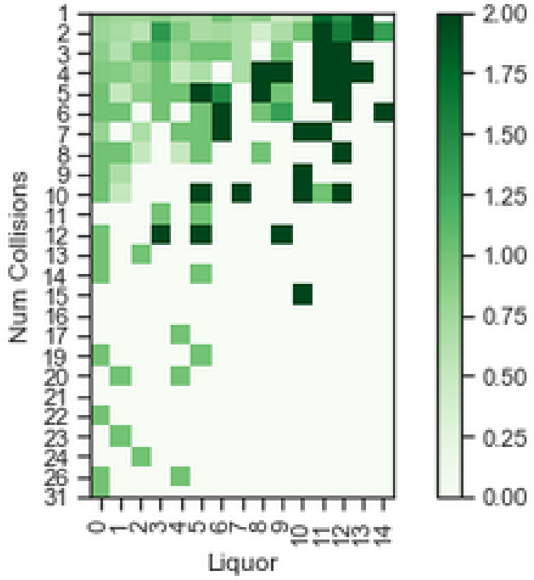

Figure 13 Analysis using a Heat Map to determine co-relation between number of collisions and liquor places

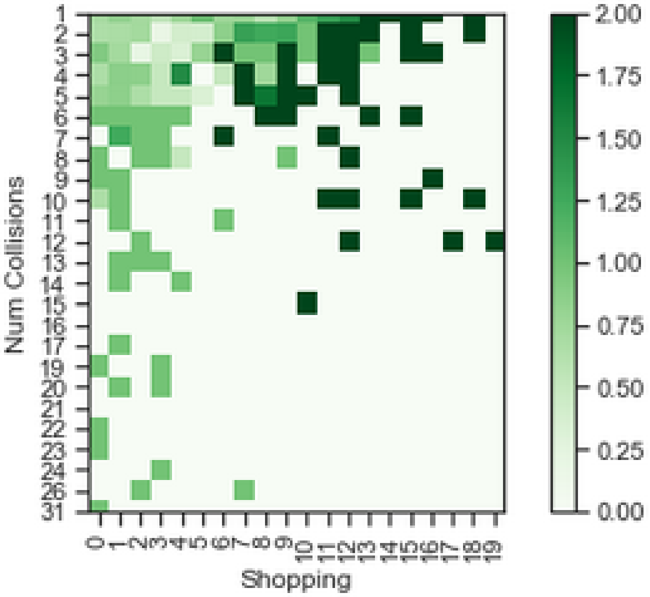

Figure 14 Analysis using a Heat Map to determine co-relation between number of collisions and shopping places

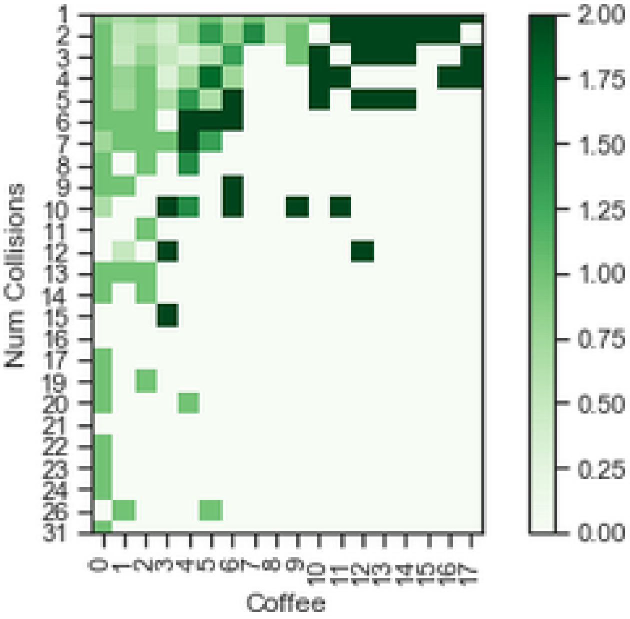

Figure 15 Analysis using a Heat Map to determine co-relation between number of collisions and coffee places

Since none of these visualizations show any co-relation to the original hypothesis, I decided to switch gears and try a different approach as described in the next section.

3.4 Analysis of Collisions relative to selected Tim Horton's

In an attempt to explore if there were any other hidden insights embedded in the data, I decided to apply a different approach to determine if there was indeed any kind of co-relation between venue data and number of collisions nearby. Based on anecdotal observations of increased distracted driver behavior near popular drive-thru venues such as Tim Horton's and McDonald's, I conducted an analysis of collision locations in proximity to a Tim Horton's located at 2145 Robertson Rd, Nepean, ON K2H 5Z2.

Starting with this as a base location point, I explored all collision locations within a 4 km distance, and determined which ones had a Tim Horton's nearby.

3.4.1 Mapping of Selected Data

The following set of 215 collision locations are included in this analysis, with 620 collisions in total, ranging from 1 to 27 at each location.

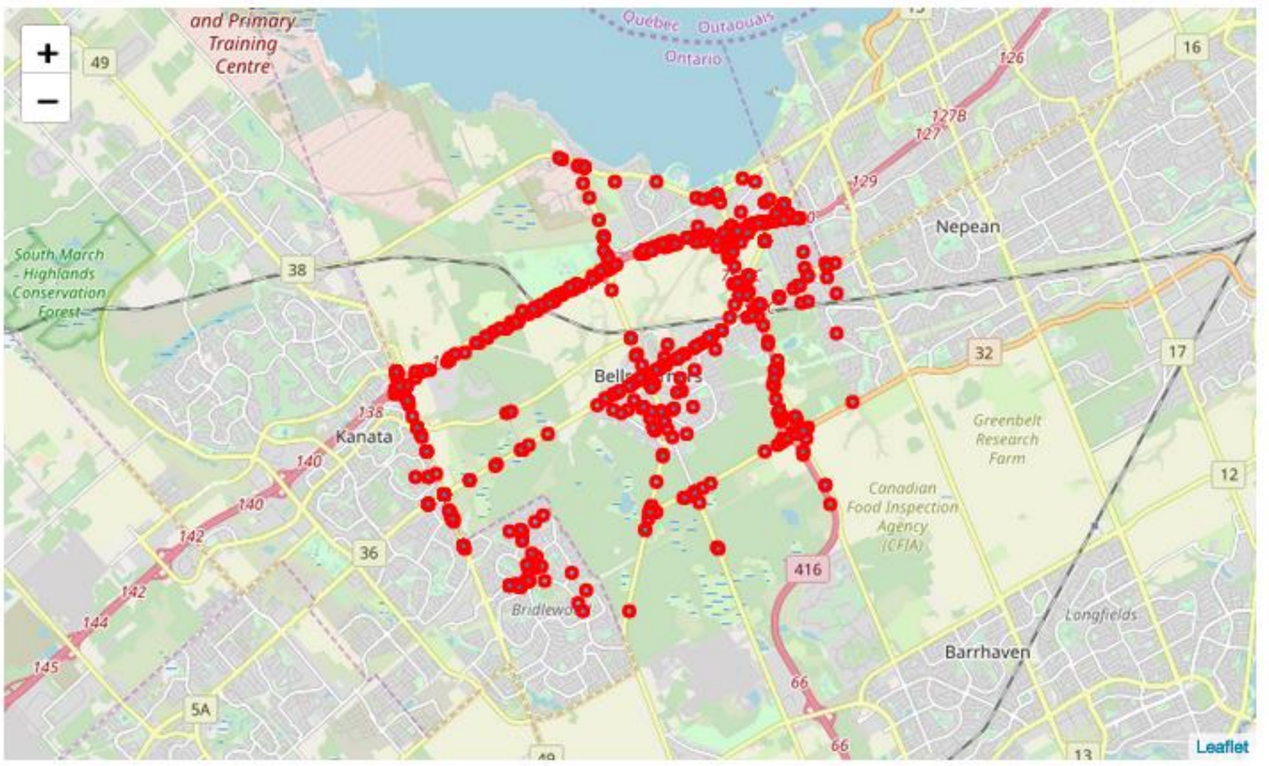

Figure 16 All collision locations within a 4 km distance from 2145 Robertson Rd, Nepean, ON K2H 5Z2

3.4.2 Data Preparation

Venue data was obtained from FourSquare for each of these points, and in addition to calculating the total of each venue type for each collision location, a new feature was added to identify if a Tim Hortons was included in the venue data for each location. This new feature was called "TH Nearby".

3.4.3 K-Means Clustering

For this more specific analysis, I applied K-Means clustering to a restricted set of features as follows:

- Number of Collisions
- Tim Horton's Nearby
- Number of venues in the following venue types:
  - Coffee
  - Eatery

To determine the ideal value of K, I compared the Squared Error or Cost for different values of K and used the "elbow" method to determine that 4 was the ideal value of K.

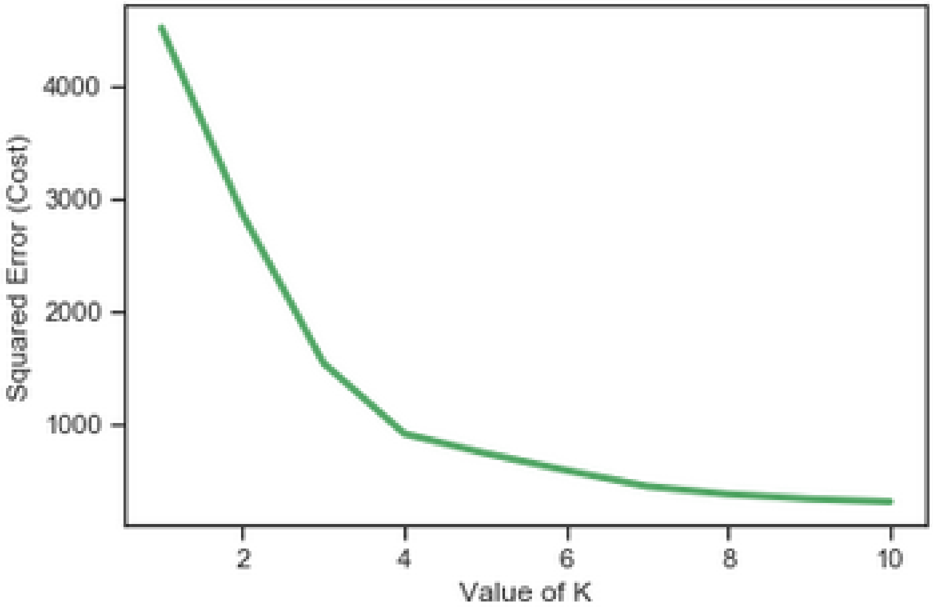

Figure 17 Compare the Squared Error(Cost) for different values of K for locations near 2145 Robertson Rd, Nepean

3.4.4 K-Means Clustering Grouping

The 4 K-Means clusters seem to be grouped as follows:

- Cluster 0: Have a low range of Coffee and Eatery type of venues
- Cluster 1: Have a medium range of Coffee and Eatery type of venues
- Cluster 2: Have 0 Coffee or 0 Eatery type of venues
- Cluster 3: High number of collisions in the range of 20-27

3.4.5 Visualization using Parallel Coordinates

The visualization using number of collisions, whether or not there is a Tim Horton's nearby, and the number of coffee and eatery type of venues is as follows:

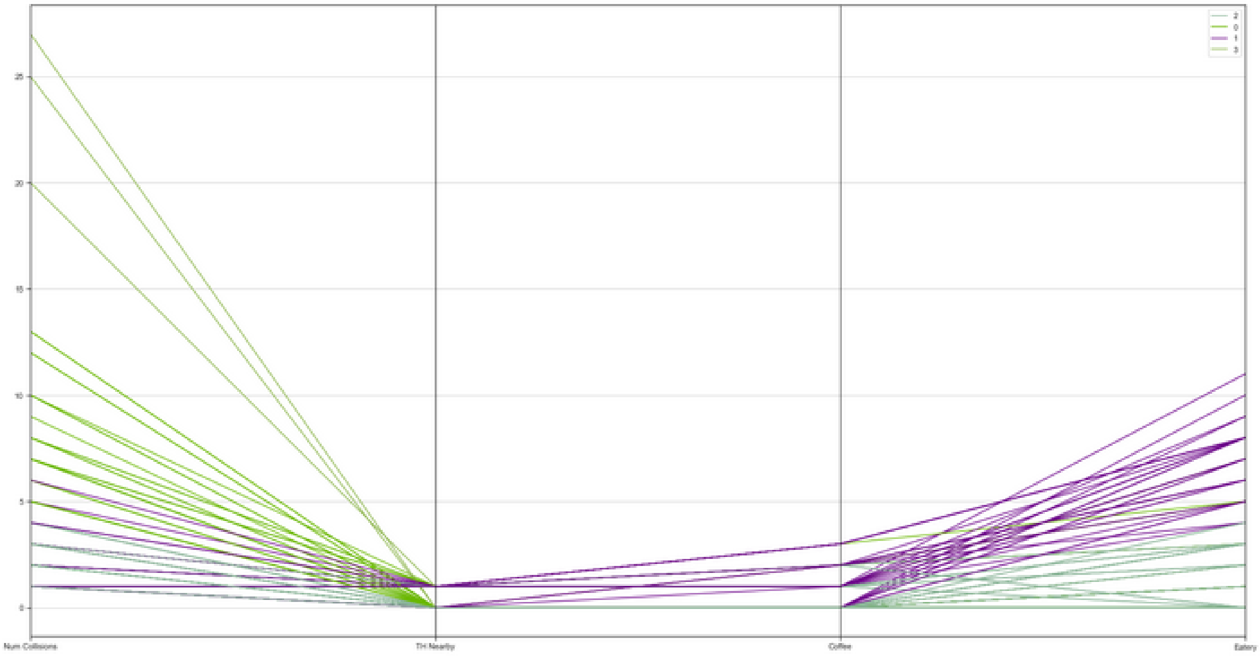

Figure 18 Analysis using Parallel Coordinates of K-Means results for locations near 2145 Robertson Rd, Nepean

The results do not clearly indicate whether or not the collision locations that have a larger number of collisions are near a Tim Horton's.

3.4.6 Mapping of K-Means Clustering Results

This visualization shows the separate clusters as circles in 4 different colors, superimposed with larger dark blue circles indicating all the Tim Horton's locations, and red markers to indicate all locations with collisions >= 8. This map focuses on locations with high number of collisions.

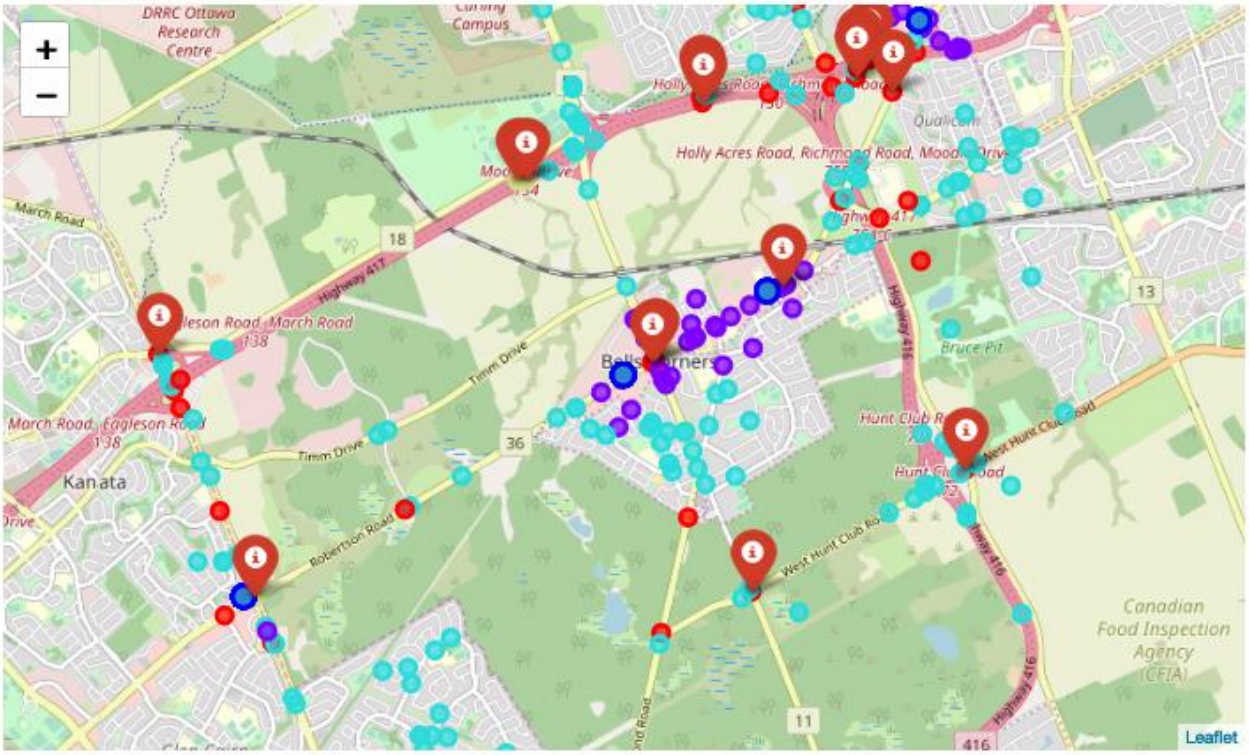

Figure 19 All collision locations within a 4 km distance from 2145 Robertson Rd, Nepean, ON K2H 5Z2, with blue circles indicating Tim Horton's locations and red markers indicating locations with >= 8 collisions

In order to change perspective, I decided to create a view with a focus on the Tim Horton's locations. This visualization shows the separate clusters as circles in 4 different colors, with black lines around the circles where the number of collisions is >= 8. It is superimposed with red markers indicating all the Tim Horton's locations.

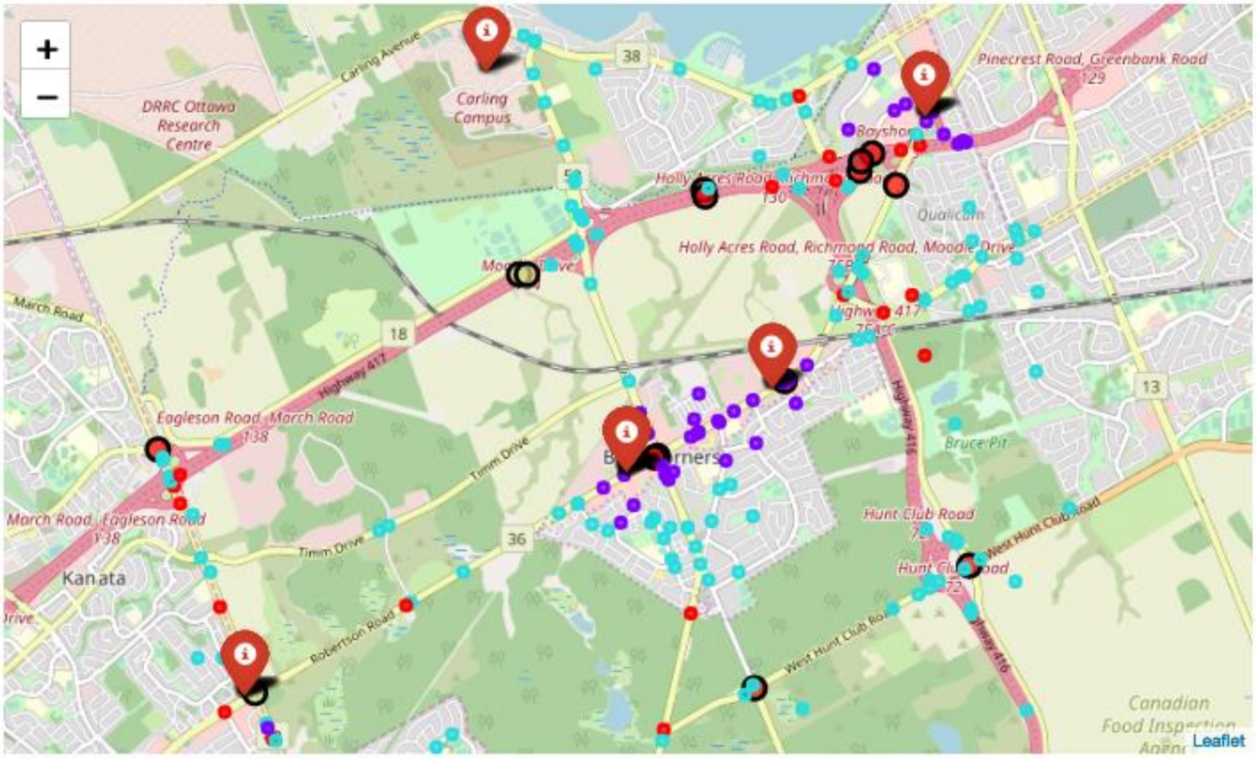

Figure 20 All collision locations within a 4 km distance from 2145 Robertson Rd, Nepean, ON K2H 5Z2, with red markers indicating Tim Horton's locations and black lines around the circles indicating locations with >= 8 collisions

A pattern has emerged indicating that the locations with higher number of locations are primarily close to either highways or Tim Horton's locations.

5 Results

5.1 First Stage

The initial hypothesis was that the number of collisions at each location would have a reasonable probability of being co-related to the number of specific types of venues at each location. For example, the number of venues related to food, liquor, or shopping could impact the driving behaviors such as individuals being distracted or being in a rush due to their visits to or from these venues.

The collisions in the Ottawa area were analyzed and the results could not find such a co-relation. In fact, results indicated that there appeared to be a negative co-relation, i.e. the locations with higher number of collisions have less venues related to restaurants, bars, and shopping.

5.2 Second Stage

Before concluding the analysis, I decided to apply another approach to perform an additional and deeper validation on these results. Based on anecdotal observations of increased distracted driver behavior near popular drive-thru venues such as Tim Horton’s and McDonalds, I conducted an analysis of collision locations in proximity to a few Tim Horton's locations. The results indicated that the higher number of collisions are primarily close to highways and near Tim Horton’s locations.

5.3 Insights

There are 2 key insights here:

1. The second stage results support some of the first stage results, as most of the highway locations in the Ottawa area do not typically have a lot of coffee, eatery, liquor, or shopping types of venues nearby. Since the higher number of collisions are at highway locations without these venues, both sets of results are in sync on this analysis.

2. These second stage results also disclosed a new pattern indicating that there is a co-relation between locations with higher number of collisions and their proximity to specific venues such as Tim Horton’s drive-thrus.

6 Discussion

Based on the two-stage approach, the analysis shows that there is a basis to support the hypothesis about venue types impacting collisions. However, since the data in the second stage was limited, broader analysis with a larger dataset would be needed to firm up this hypothesis. Additionally, the venue type as categorized may need to be reconsidered. For example, it is possible that all drive-thrus may have a higher co-relation as opposed to all coffee shops or eateries in general.

Additional analysis is recommended to explore this co-relation using a larger dataset with more specific features that are not currently available via FourSquare such as whether or not the venue has a drive-thru.

7 Conclusion

This analysis indicates that there is a reasonable probability that collision occurrences are impacted by nearby venues, although additional analysis is required to better understand what specific features of these venues can impact collisions.

The results of this analysis, as well as any additional deeper analysis that may be conducted, will be useful for city planners when approving and designing roads and intersections near specific types of venues.