**Toronto Police Services Data Visualization**

May 1, 2021

**CDSA 1050 – Advanced Analytics**

**Capstone**

**Group 4:**

1. Introduction

# The Toronto Police Services are dedicated to delivering police services in partnership with their communities to keep the city of Toronto safe. They have approximately 7,900 full-time and part-time uniform and civilian members.

# Over the last years the city of Toronto has become a large and complex city, with this comes new demands and pressures on policing. One of the key goals of the Toronto Police Service is to enhance confidence and strengthen ties with society by providing access to open data for public safety in Toronto.

# This will promote transparency, accountability and efficiency by the police departments. Data visualization helps police and the general public to look deeper into data, discover hidden opportunities, identify key data relationships, and make quicker, more informed decisions.

2. Business Problem Statement

# Visualization is a critical method for analyzing and comprehending large-scale data and helping the stakeholders target where the issues exist and prioritize the deployment of safety measures. The aim of this project is to recommend potential improvements of data visualization for Cyclists data available on the Toronto Police Service’s Public Safety Data Portal. Visual analytics can be used by police and other stakeholders to discover collision and theft causes, find the hotspots, and make potentially life-saving adjustments.

# We aim to derive insights for the development of new visualizations for PSDP such as analytical dashboards, interactive visual exploration, heat maps that would enhance the “analytics journey” of users.

# This will lead to better data storytelling and pattern discovery that will allow users to quickly find answers to key questions within the public data, identify trends, patterns, and outliers within large data sets.

## 3. Analytical Problem Statement

## Our two main analytical questions are:

## How can data visualization be used to inform the public about the dangers while cycling in Toronto

## What variables lead to cyclist-motorist accidents in Toronto?

4. Data Understanding

# We will be using two datasets from the Toronto Police Service Public Safety Data portal:

# 

# <https://data.torontopolice.on.ca/search?tags=Cyclist>

<https://data.torontopolice.on.ca/search?tags=bicycle%20thefts>

First dataset is a subset of the Killed and Seriously Injured (KSI) dataset from 2006-2019. These events include any serious or fatal collision where a cyclist is involved. Second dataset contains occurrences related to bicycle thefts from 2009 to 2020. (Find the Data dictionary in the Appendix b).

5. Data Exploration

Upon importing the datasets for cyclist collision and bicycle thefts in Toronto, the shape was viewed to identify how many records and attributes exist in the dataframe. The collision dataset contained 1681 records and 56 attributes while the theft dataset had 25569 records and 35 attributes.

A list of the column names and the total number of null values in each column was then produced for both tables. This allows for easier determination of what attributes will be dropped either due to extremely high null values or having little relevance for the desired analysis that will be done. Value counts were used for multiple columns to identify the frequency of certain values and the number of unique values found in each column. Attributes of interest received a deeper look with visualization.

We used different visualizations like graphs and charts to show patterns and reveal connections between factors, times, and places. Preliminary analysis revealed that collisions were highest in 2012 and 2013, and have been trending in a decreasing fashion until 2019. This may be the result of increased awareness or safety measures put into place by the city.

One such policy is Toronto's cycling network plan of 2016, which looks to grow and renew infrastructure for Toronto's cycling routes. It was also discovered that 68.41% of collisions happen in daylight and that the majority of collisions occur in afternoon/evening rush hour between 5PM and 7PM.

This is logical as most cyclists do not ride at night and the increase in vehicles on the road at rush hour makes it more likely that a collision will occur. It was also discovered that 64.03% of collisions occur in the 'Toronto and East York' district. This indicates that perhaps better signage or a better layout of bicycle lanes should be implemented in this district, as it accounts for almost two-thirds of the total number of collisions.

With regards to bicycle thefts, only 1.2% of bicycles stolen in Toronto were recovered. In an attempt to understand this low percentage, an article in Bloomberg in which NYPD officers were interviewed was found and it explains why bicycle thieves are extremely hard to identify.

According to the officers, this is mainly due to a lack of relationship between victim and perpetrator, as well as a lack of proof of ownership of the bike. Just like with collisions, bicycle thefts were highest during rush hour.

The high availability of bicycles at this time, combined with the low likelihood of getting caught, may tempt thieves into stealing the bikes. Finally, very few thefts were reported between 2009-2013 but the figures have been consistently high since 2016. This kind of jump is likely due to a lack of data collection or storage by the police department.

6. Data Preparation

Preparation of the data began with identifying the columns to drop due to high null values or lack of relevance for the desired analysis. Once the attributes were dropped from both datasets, column names for the theft dataset were renamed to match the collision data in some cases and to make the table look cleaner in others. This would allow for smoother joining of the tables to create a new dataframe.

The ‘Division’ column appeared in both datasets but were two different types, so some formatting was done to make this attribute an integer. There were many vehicle-based columns in the collision dataset so a new column was created which identified the vehicle type, allowing for the previous columns to be dropped. A new column was created in both datasets to identify if the incident was a theft or a collision.

The 'REDLIGHT', 'ALCOHOL', 'SPEEDING', AG\_DRIV' and 'PEDESTRIAN' columns all contained either 'Yes' or null values. The null values were converted to 'No' as the absence of a 'Yes' means the police did not believe these factors were involved in the accident. Upon merging the tables, the 'DATE' column was converted to date-time format to simplify any potential time based analysis which may be conducted.

7. Modeling

For the cyclist KSI dataset, we are using Clustering Analysis for modelling. Clustering would help us to explore further and understand the relationships between the data.

The Elbow method was used to determine the number of Clusters which was found to be 4. K-means and Hierarchical clustering using Dendrogram were considered for the analysis. Clustering helped to further investigate the parameters Speeding, Aggressive Driving, Red-light and Alcohol against the Neighbourhoods, accident incidents and Fatal injuries.

Random Forest Regressor was executed to fit the model on the training dataset. The model score was 0.8784 on the training set.

When grouped by neighbourhoods, it was found that aggressive driving was the most common reason for accidents. Regions in cluster 3 had the highest numbers in injury due to aggressive driving. Also, the majority of the accidents and injuries due to it happened in daylight in all the clusters.

Further analysis was carried out by selecting only Fatal injuries and it was found that aggressive driving and speeding contribute to the majority of accidents leading to fatal injury in cluster 2 and cluster 3. Again, daylight was the time when most of the fatalities occurred but in cluster 3 dark light too contributed to fatalities in higher percentage.

Cluster 3 had no fatal injuries due to Speeding and Alcohol. Cluster 4 had no fatal injury reported due to Redlight and Alcohol while those due to Aggressive Driving and Speeding were also lower as compared to other regions or Clusters. It can be said that Cluster 4 was relatively safe as compared to others.

8. Challenges

There were a lot of Nan values in the datasets. We handled that in the data cleaning through dropping unwanted columns or replacing the Nan values. There are some columns that are not shared by both datasets so Nan values remain in those fields.

The other issue with the development of the dashboard in Tableau is we are not connected to the data source directly so will not be able to automate. The dashboard ideally would refresh daily with the latest data for the user.

9. Recommendations

Every stakeholder needs to take responsibility in improving cyclist safety and experience in the city of Toronto.

* For the police

The police can not only carry out repressive measures but also aim to eradicate the triggers. We identified hotspots of cyclist incidents, including thefts and collisions from our research and attempted to determine the causes. Although the majority of accidents are non-fatal, it is still vital to use our analysis to identify dangerous intersections to improve road safety. We recommend that the fatal hot spot areas be prioritized.

We propose that the police and the parties concerned install more cycling lanes, bike paths, underpasses, and overpasses so cyclists have more space and options especially during rush hour.

Since most of the accidents happened when the cyclists were going straight or making turns, we recommend more signage or advance green for cyclists to indicate who has the right of way. Reducing conflict areas between cyclists and motor vehicles will most likely reduce a lot of bike - motorist collisions.

Also educating and urging the public to be cautious when cycling or when storing their bikes, especially in clear weather, during the day, in the summer times, and mostly rush hour can also bring the number of thefts and accidents down.

Police can also work with the public to capture and crowd source real time road information so as to improve public safety as well as keep their data up to date. This can also help the police to communicate real time traffic control through mobile alerts, offer fastest routes, or alternative routes to increase efficiency and reduce incidents especially during rush hours.

* The developers

We recommend apartment developers and landlords to include safer bike storage in their plans so as to assure potential clients that their bikes and property is secure because the data shows a lot of bikes are stolen in and around apartment buildings. Also better surveillance systems in the areas where bikes are stored to help with identifying thieves.

* Civilians/public

We recommend the public to use this data to do their part in reducing cyclist incidents. For accidents and thefts, it is important to be aware of the factors that are most likely to lead to either incident happening. So, we recommend vigilance especially during the summer time. The public can store their bikes near transit stations if possible since the data show less possibility of their bike being stolen.

Civilians should be urged to report incidents and help the police in situations where a theft has occurred, similar to “Neighbourhood Watch”. This may be also useful for cyclist safety as well, as these people may be able to identify unsafe behaviour of drivers or cyclists that could lead to collisions.

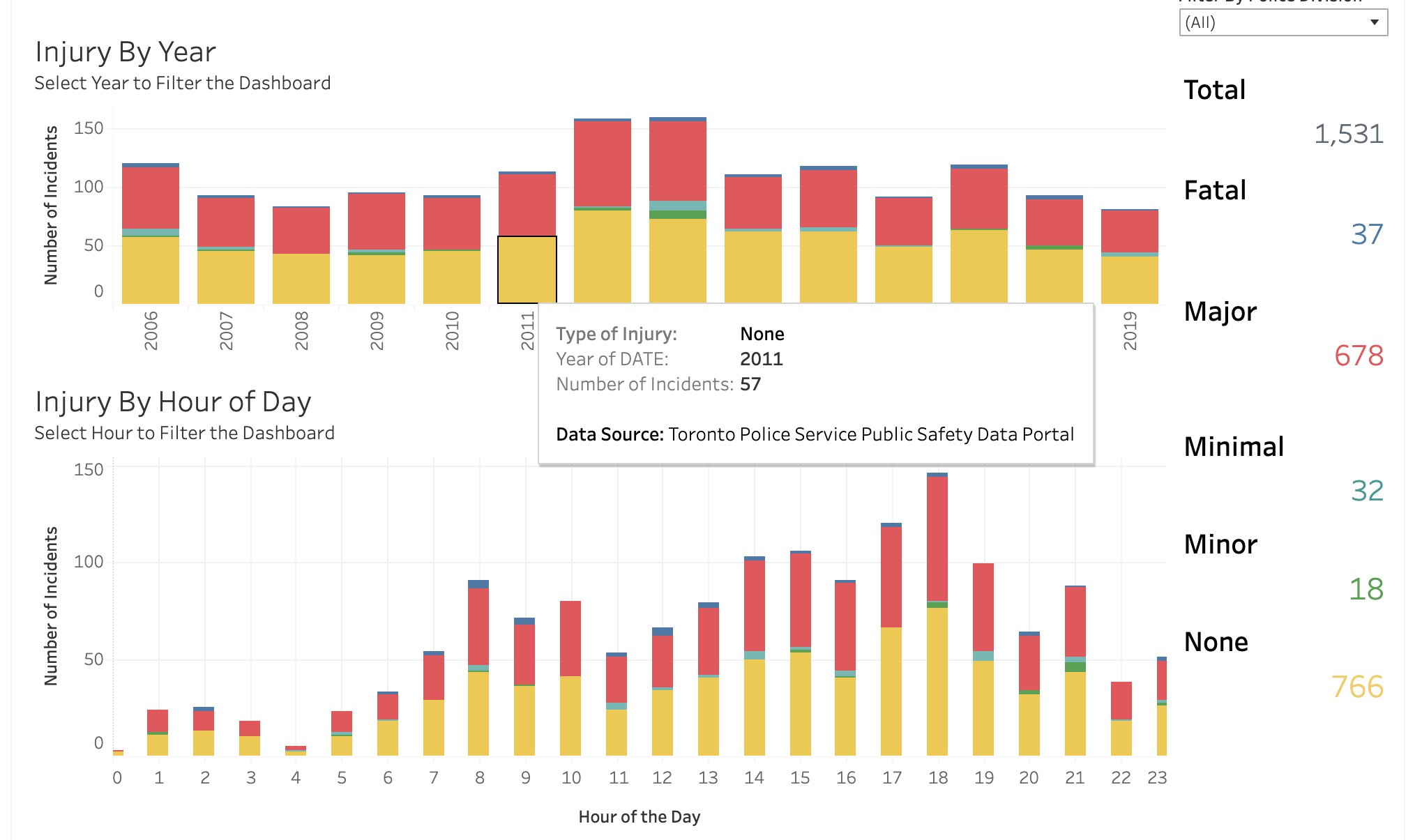
* The city/ the city council

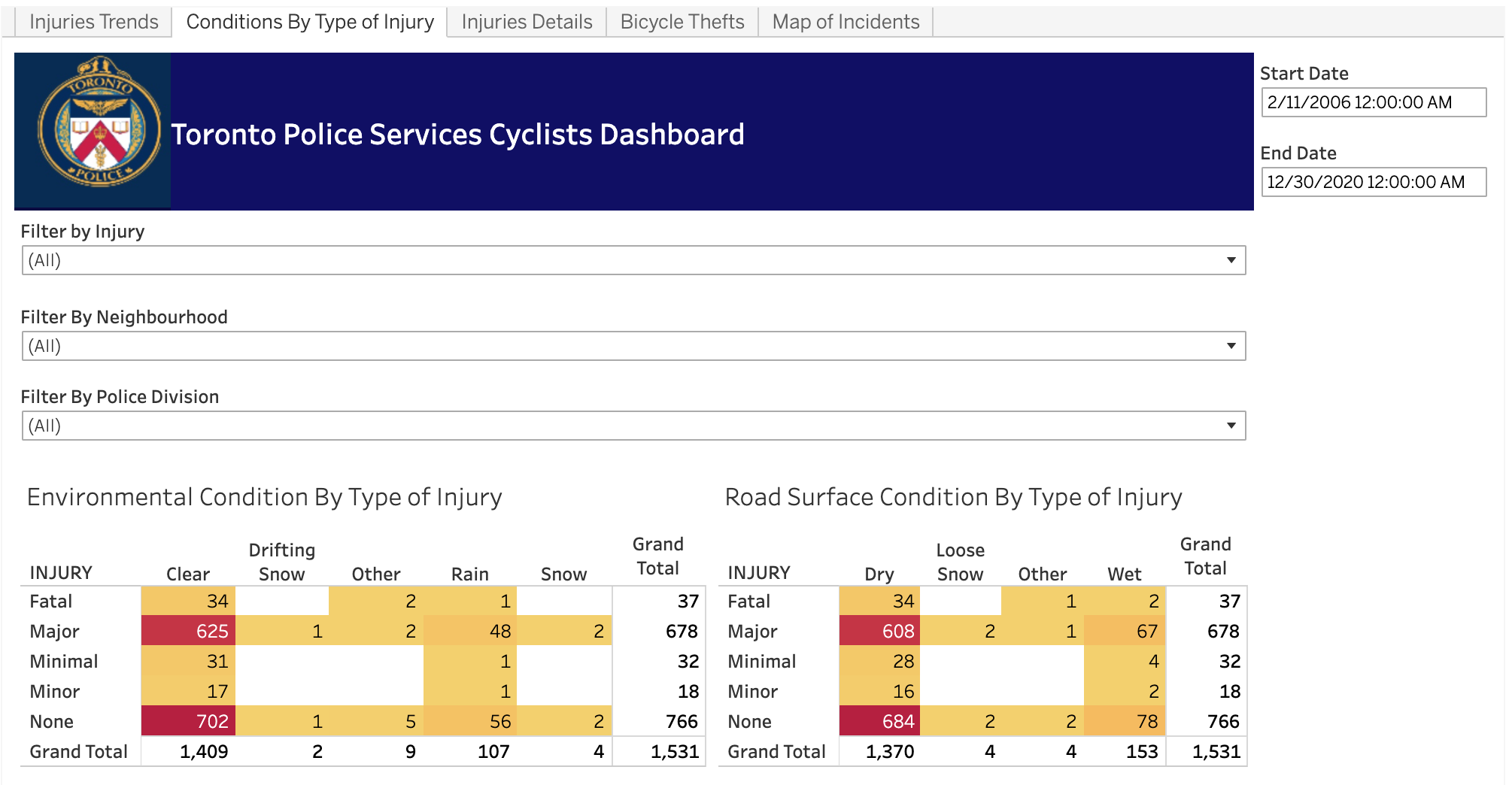
When planning the city’s infrastructure, the city needs to use the data to design safer and more conducive roads and spaces for cyclists. The ‘Cycling Network Plan’ was created by the city of Toronto in 2016 and was originally intended to be a 10-year plan. It has been updated in 2019 and will be operating as a detailed three-year rolling implementation program that will be updated as required for the foreseeable future.

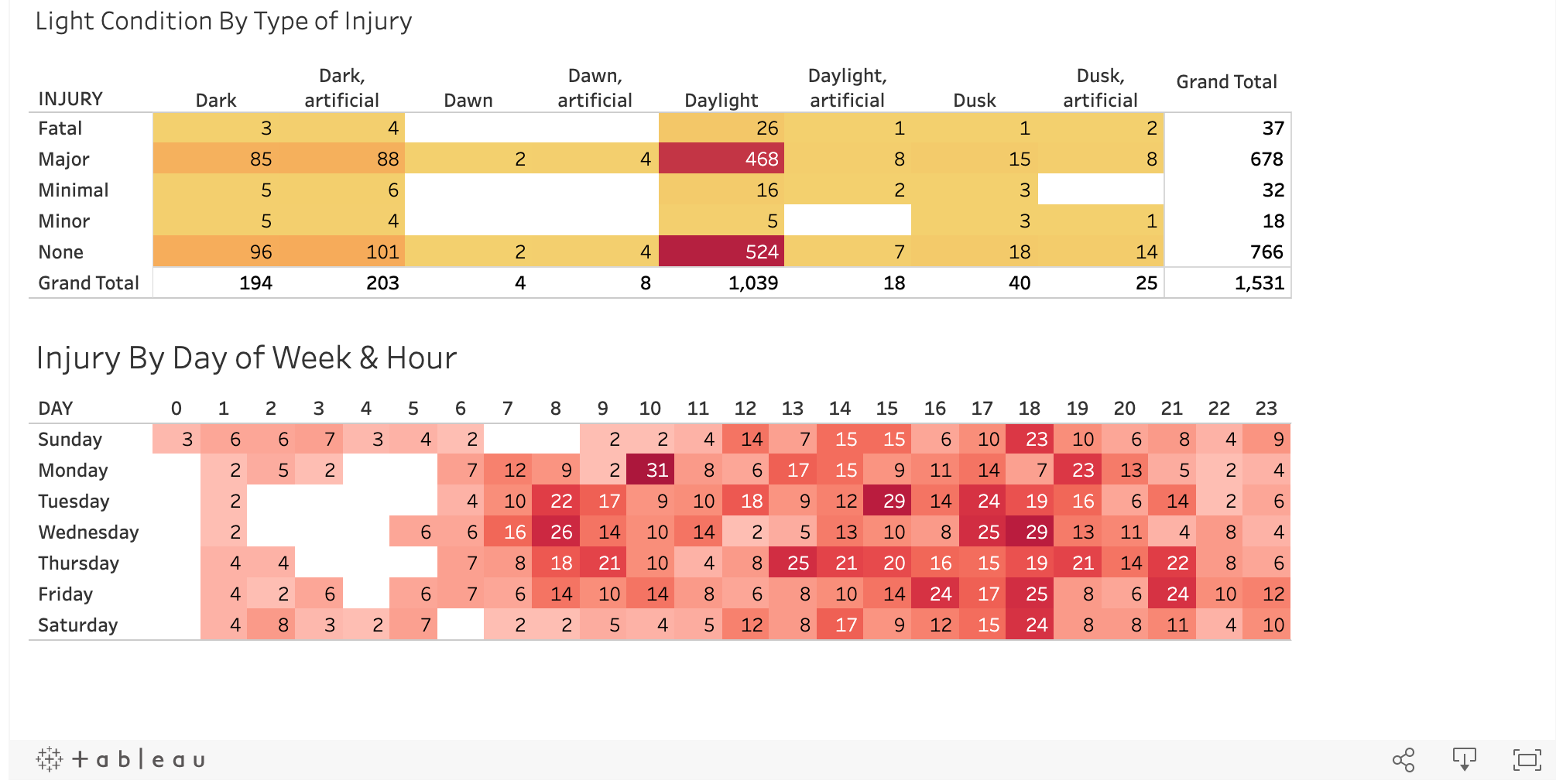
This plan will create new bike lanes and work to better coordinate road work in a manner that will improve cyclist safety. In addition to this plan, we recommend the city to instal more safe spaces where cyclists can leave/store their bikes around the city to decrease the number of thefts.

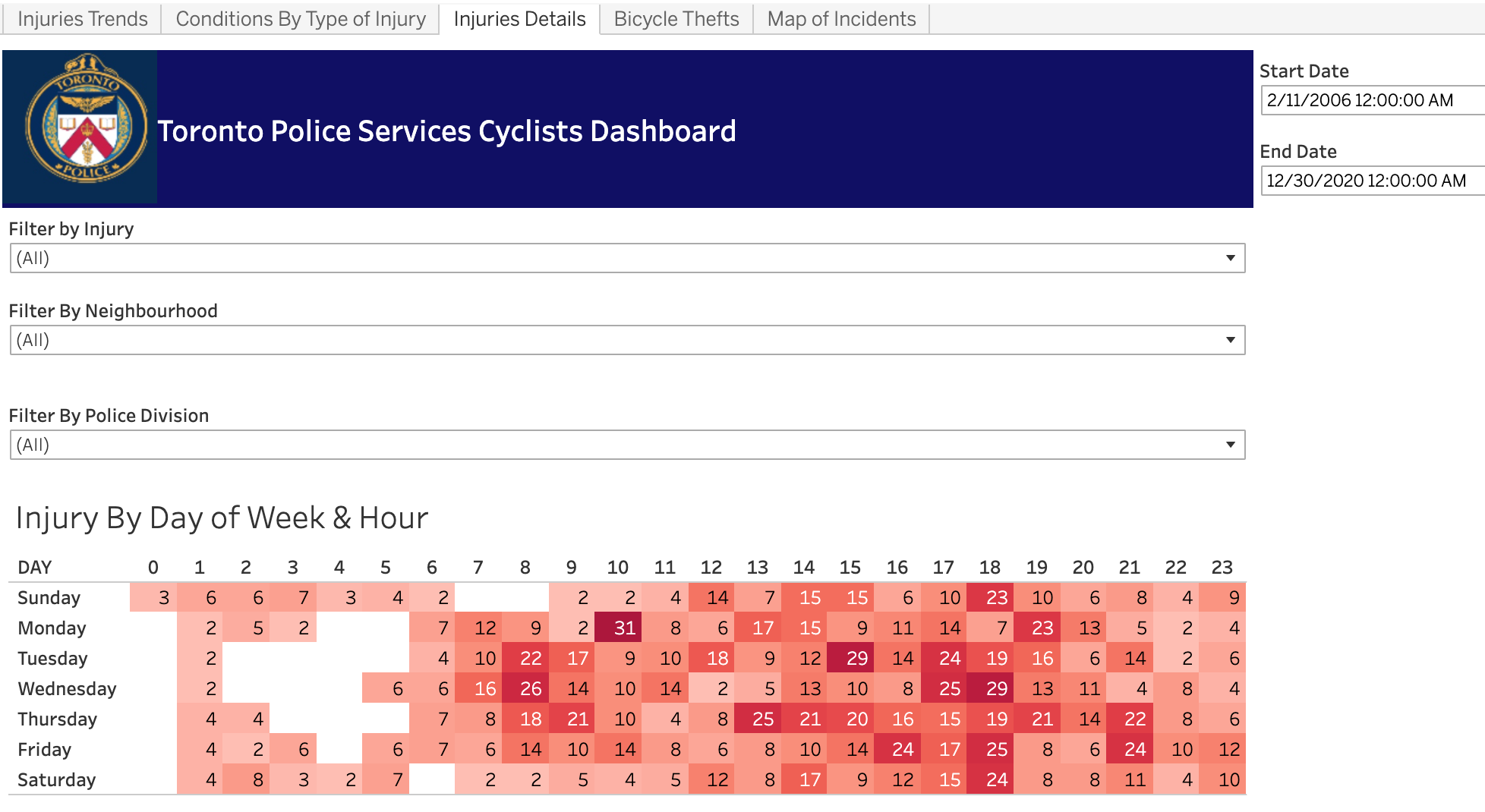
Appendix A: Tableau Visualization

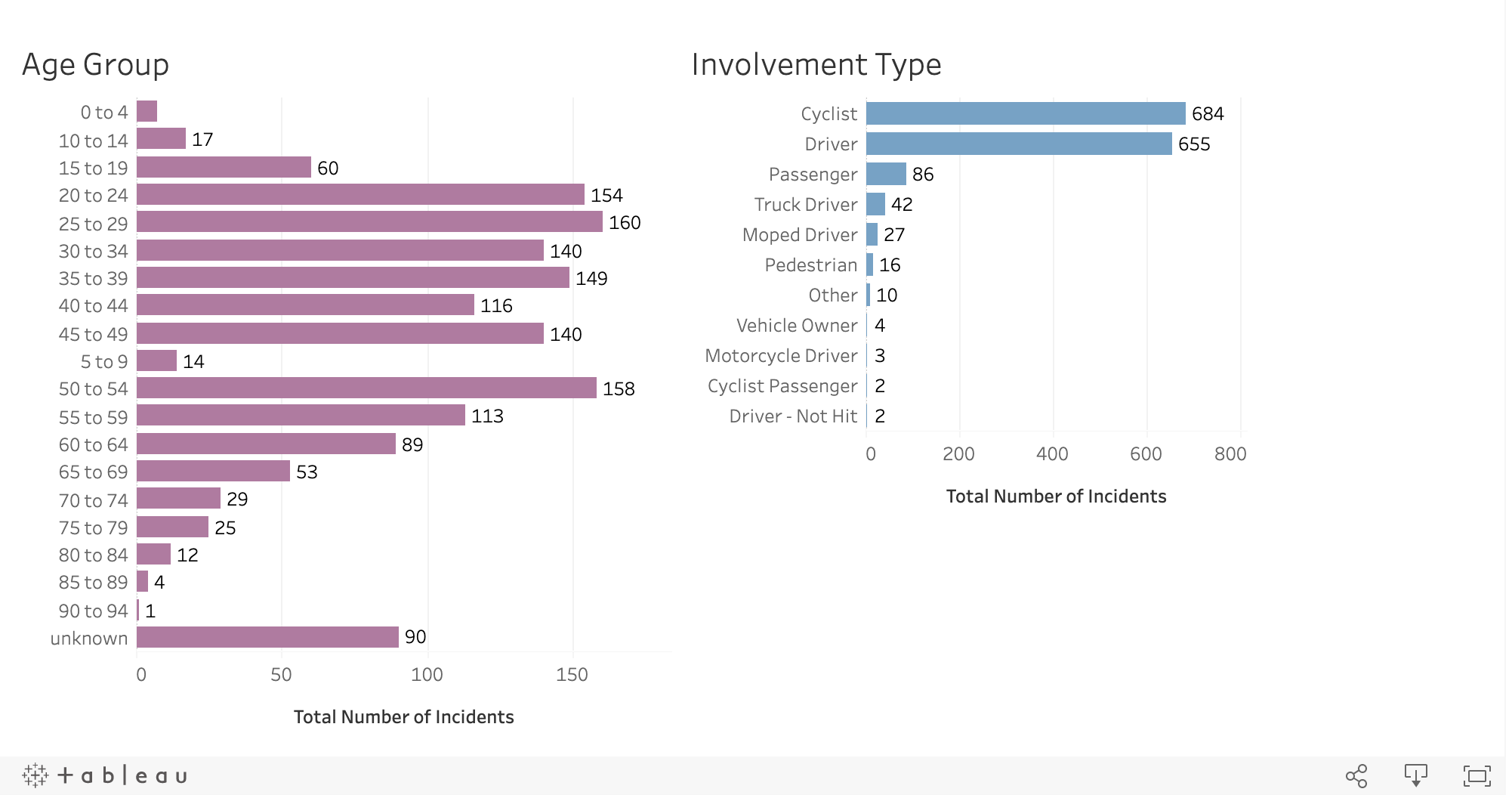


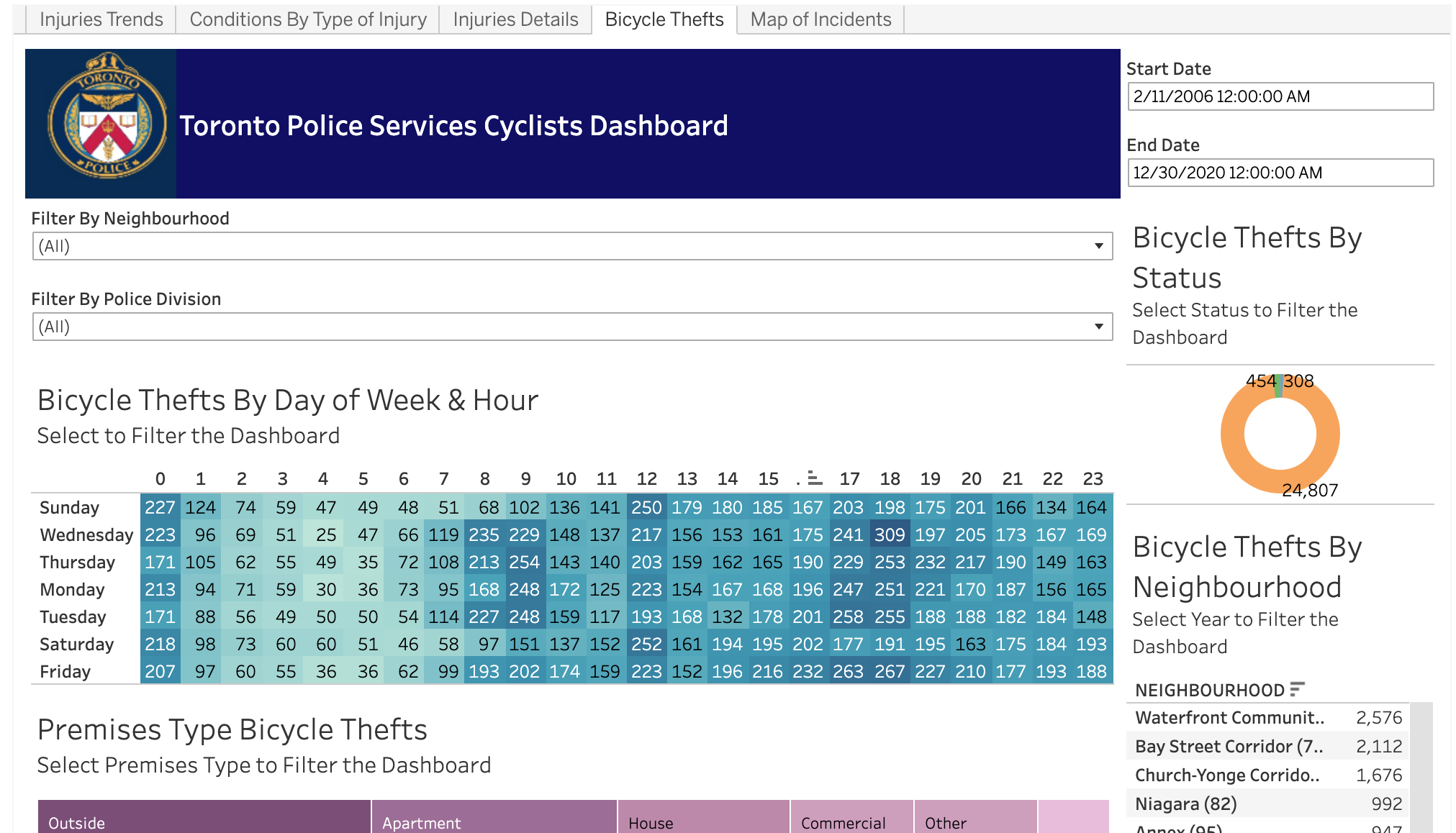


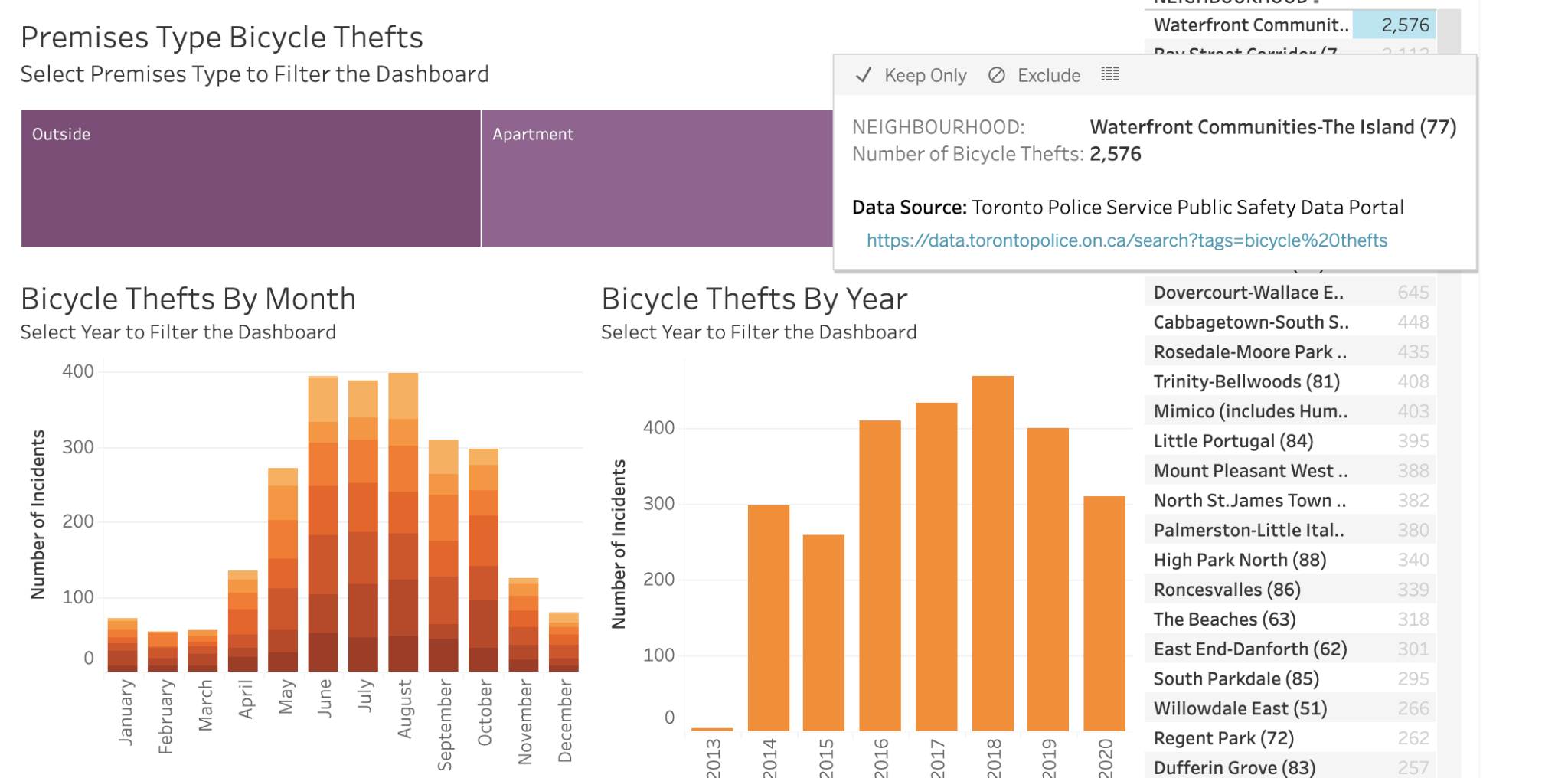


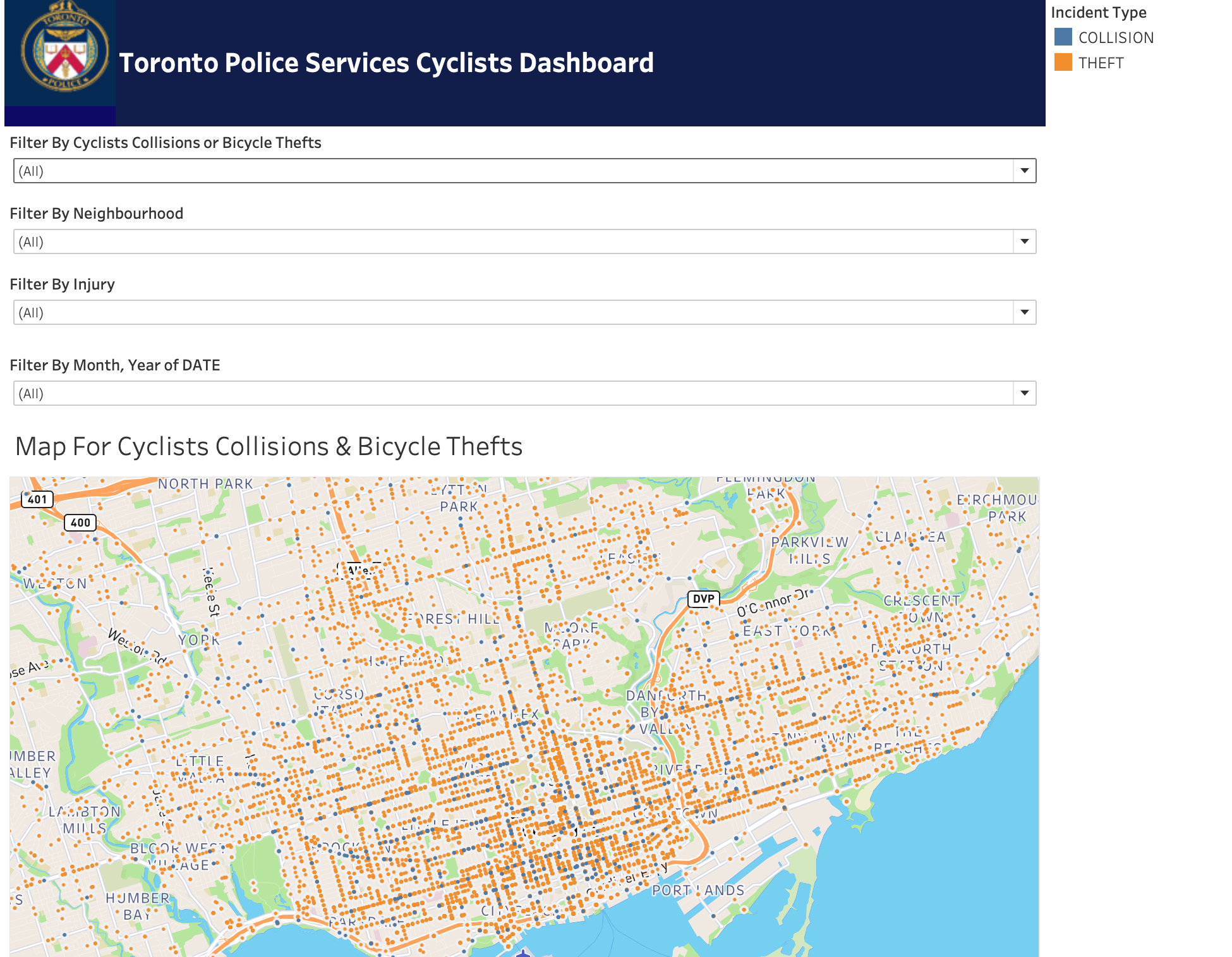












The user can hover over the numbers and be able to see the graph of all the incidents.

Appendix B: Data dictionary

|  |  |
| --- | --- |
| Bicycle data dictionary | Bicycle codes |
|  |  |

|  |  |
| --- | --- |
| KSI Cyclist data dictionary | KSI Metadata |
|  |  |

Resources

<https://www.toronto.ca/services-payments/streets-parking-transportation/cycling-in-toronto/cycle-track-projects/cycling-network-10-year-plan/>

<https://www.bloomberg.com/news/articles/2012-09-14/why-bike-theft-is-so-hard-to-stop>