# Group Project Report

COMP309-006 Group 9

Data Warehouse & Mining – HCIS

11/29/2020

Liping Wu 300958061

Xinglong Lu 300967280

Dayoon Lee 301034178

Yoonseop Lee 301037418

Nayul Kim 300888467

Github source link: <https://github.com/foxpeer/comp309group09.git>

Table of Contents

[0. Executive summary & Overview of solution 5](#_Toc57563708)

[1. Data exploration: a complete review and analysis of the dataset including: 6](#_Toc57563709)

[1.1 Original Data download 6](#_Toc57563710)

[1.2 Data Loading 6](#_Toc57563711)

[1.3 Original Data Description 6](#_Toc57563712)

[1.3.1 Columns and Values 6](#_Toc57563713)

[1.3.2 Data Shape 6](#_Toc57563714)

[1.3.3 Data Describe 7](#_Toc57563715)

[1.3.4 Data Type 8](#_Toc57563716)

[1.3.5 First 5 rows 9](#_Toc57563717)

[1.3.6 Last 5 rows 10](#_Toc57563718)

[1.3.7 Understanding of the Columns 10](#_Toc57563719)

[1.4 Data transformation 11](#_Toc57563720)

[1.4.1 From Occurrence\_Date to get dayofweek 11](#_Toc57563721)

[1.4.2 From Occurrence\_Time to get hour of day 11](#_Toc57563722)

[1.4.3 Drop unnecessary columns 11](#_Toc57563723)

[1.4.4 New DataFrame(df\_g9) columns 12](#_Toc57563724)

[1.4.5 New DataFrame(df\_g9) data shapes 12](#_Toc57563725)

[1.4.6 New DataFrame(df\_g9) data describe 12](#_Toc57563726)

[1.4.7 New DataFrame(df\_g9) data types 13](#_Toc57563727)

[1.4.8 New DataFrame(df\_g9) data first 5 rows 14](#_Toc57563728)

[1.4.9 New DataFrame(df\_g9) data last 5 rows 15](#_Toc57563729)

[1.4.10 New DataFrame(df\_g9) data columns unique values 15](#_Toc57563730)

[1.5 Data missing and replacement 19](#_Toc57563731)

[1.5.1 Check the missing data 19](#_Toc57563732)

[1.5.2 Fill the Bike\_Model null by “UNKNOWN”, 19](#_Toc57563733)

[1.5.3 Fill the Bike\_Colour null by “UNKNOWN”, 19](#_Toc57563734)

[1.5.4 Fill missing or 0 of Cost\_of\_Bike with median 20](#_Toc57563735)

[1.5.5 Drop row with empty space ‘ ’ in Status 20](#_Toc57563736)

[1.6 Group the data with different group base on features 20](#_Toc57563737)

[1.6.1 Group the data with location information 20](#_Toc57563738)

[1.6.2 Group the data with time/date information 21](#_Toc57563739)

[2.6.3 Group the data with bike information 22](#_Toc57563740)

[1.6.4 Group the data with geo information 22](#_Toc57563741)

[1.6.5 Group the data with Primary\_Offence 23](#_Toc57563742)

[1.7 Graphs and Visualization 23](#_Toc57563743)

[1.7.1 Stolen Status Pie Chart 23](#_Toc57563744)

[1.7.2 Bike Theft Map 24](#_Toc57563745)

[1.7.3 Division and Stolen 26](#_Toc57563746)

[1.7.4 Hood\_ID/Neighbourhood, Lat, Long and Stolen 27](#_Toc57563747)

[1.7.5 Premise\_Type and Stolen 31](#_Toc57563748)

[1.7.6 Location Type and Stolen 32](#_Toc57563749)

[1.7.7 Bike\_Make and Stolen 36](#_Toc57563750)

[1.7.8 Bike\_Colour and Stolen 37](#_Toc57563751)

[1.7.9 Bike\_Type and Stolen 38](#_Toc57563752)

[1.7.10 Cost\_of\_Bike and Stolen 40](#_Toc57563753)

[1.7.11 Bike\_Speed and Stolen 41](#_Toc57563754)

[1.7.12 Occurrence\_Year and Stolen 41](#_Toc57563755)

[1.7.13 Occurrence\_Month and Stolen 42](#_Toc57563756)

[1.7.14 Occurrence\_Day and Stolen 43](#_Toc57563757)

[1.7.15 Day Of Week and Stolen 43](#_Toc57563758)

[1.7.16 Hour of a Day and Stolen 44](#_Toc57563759)

[1.7.17 Primary\_Offence and Stolen 45](#_Toc57563760)

[1.8 Data Further Selection or Combination Suggestion 45](#_Toc57563761)

[2. Data modelling: 47](#_Toc57563762)

[2.1 Data Cleaning Strategies 47](#_Toc57563763)

[2.2 Check missing values from columns 47](#_Toc57563764)

[2.3 Data transform and category 47](#_Toc57563765)

[3.4 Categorical data management : Created dummy values 48](#_Toc57563766)

[3.5 Feature selection 49](#_Toc57563767)

[3.6 Train, Test data splitting 49](#_Toc57563768)

[3. Predictive model building 51](#_Toc57563769)

[3.1 Build logistic regression model 51](#_Toc57563770)

[3.2 Build the decision tree model and validate the parameters 51](#_Toc57563771)

[4. Model scoring and evaluation 53](#_Toc57563772)

[4.1 Cross validation score 53](#_Toc57563773)

[4.2 Confusion\_matrix, accuracy, recall, F1 score, precision score and ROC-AUC 53](#_Toc57563774)

[5. Deploying the model 54](#_Toc57563775)

[5.1 Logistic Regression Prediction Example 54](#_Toc57563776)

[5.2 Decision Tree Prediction Example 54](#_Toc57563777)

[6. References 55](#_Toc57563778)

# 0. Executive summary & Overview of solution

Police and residents are having a hard time because of the cases of bicycle theft in different regions. This project is a necessary service for public safety and awareness from local bicycle theft crimes. It helps people to analyze whether a stolen bicycle will be returned or not. Toronto police will be able to further strengthen their solutions to prevent theft in certain areas, and residents will be extra careful and seek preventive measures such as anti-theft locks. Therefore, this will gradually reduce the number of bicycle theft cases.

Based on the open data provided by the Toronto government and police, our team developed an analytic service using data exploration, data modeling, model building etc. It can also be learned and used easily by police because we deployed API service.

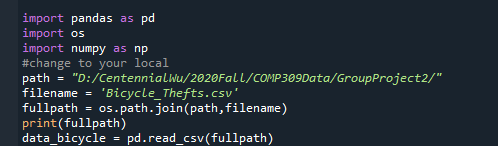
Through this project, our team learned more about how to extract useful data and use them well as well as the use of API.

# 1. Data exploration: a complete review and analysis of the dataset including:

## 1.1 Original Data download

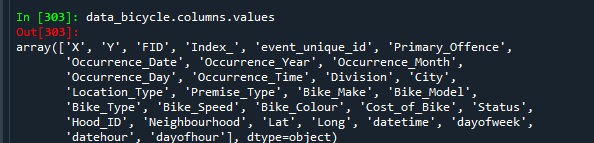
<https://data.torontopolice.on.ca/datasets/bicycle-thefts>

## 1.2 Data Loading



## 1.3 Original Data Description

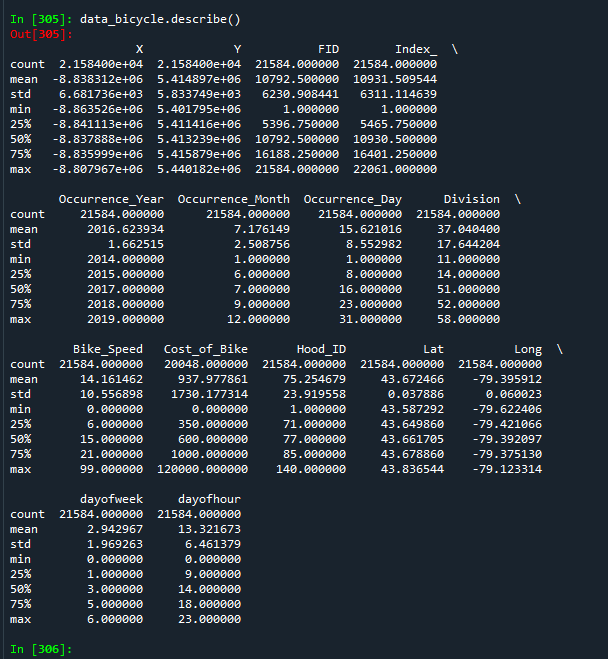
### 1.3.1 Columns and Values



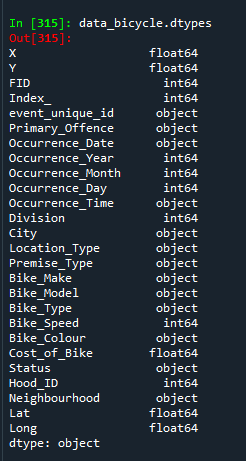
### 1.3.2 Data Shape



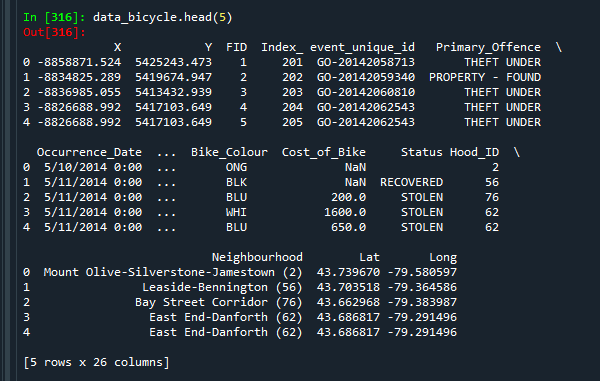
### 1.3.3 Data Describe



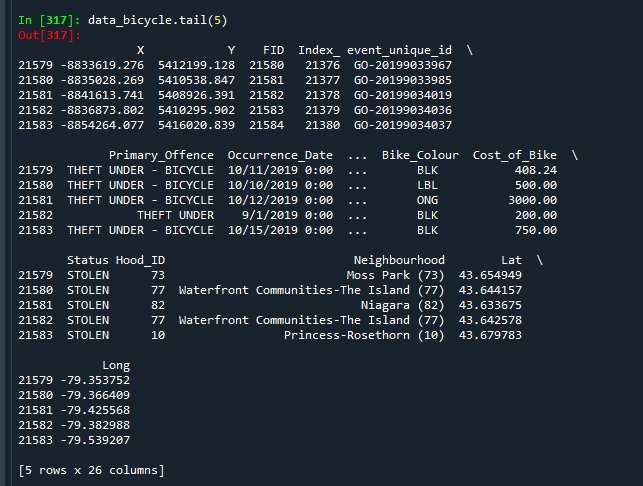
### 1.3.4 Data Type



### 1.3.5 First 5 rows



### 1.3.6 Last 5 rows



### 1.3.7 Understanding of the Columns

#### 1.3.7.1 Geography related Information

X, Y - For Grid drawing

Lat, Long - For Map Information

One pair of data, either (X, Y) or (Lat, Long) is enough for one dataframe

#### 1.3.7.2 Bike Feature Related Information

Bike\_Made, Bike\_Speed, Bike\_Style, Bike\_Speed, Cost\_Of\_Bike are bike feature information

#### 1.3.7.3 Location Related Information

City, Division, Hood\_ID, Neighbourhood, Primis\_Type, Location\_Type

All city information is Toronto

In which, Hood\_ID and Neighbourhood is the same information with different data types. We can select one of them as analysis is needed.

Primise\_Type and Location\_Type are quite similar information.

#### 1.3.7.4 Time of Date Related Information

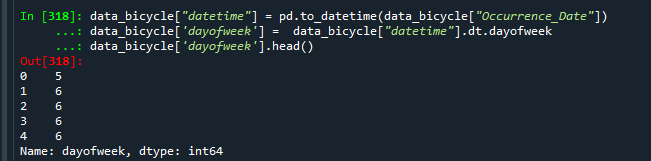
Occurence\_Year, Occurrence\_Month, Occurrence \_Day, Occurrence\_Time

#### 1.3.7.5 Target or Predict Feature

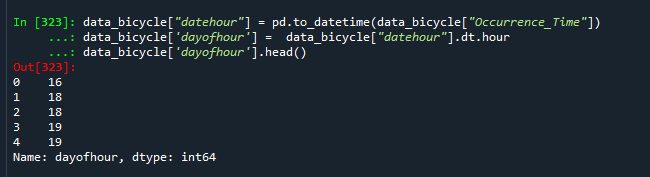
In our case, we would like to know if the factors will cause the bike theft or not, so the Column “Status” is very important to set as Target or Predict Feature

## 1.4 Data transformation

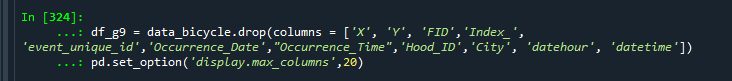
### 1.4.1 From Occurrence\_Date to get dayofweek



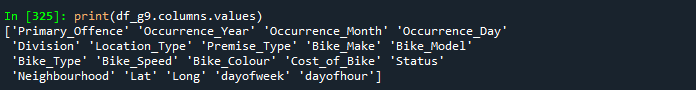
### 1.4.2 From Occurrence\_Time to get hour of day



### 1.4.3 Drop unnecessary columns



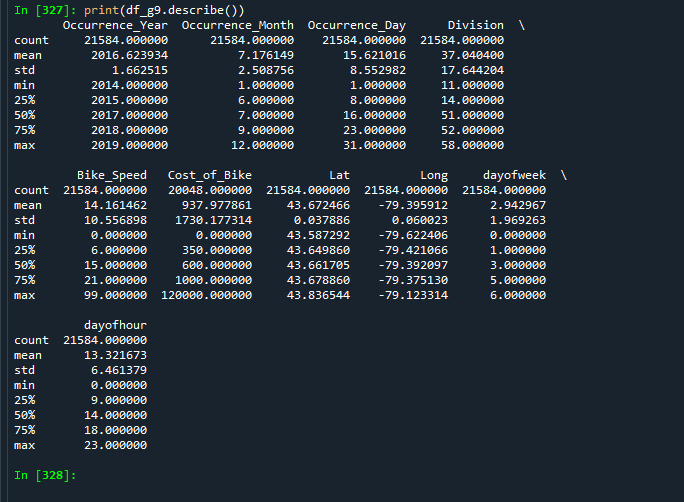
### 1.4.4 New DataFrame(df\_g9) columns



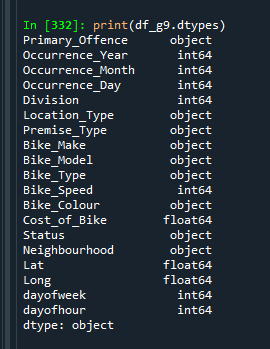
### 1.4.5 New DataFrame(df\_g9) data shapes



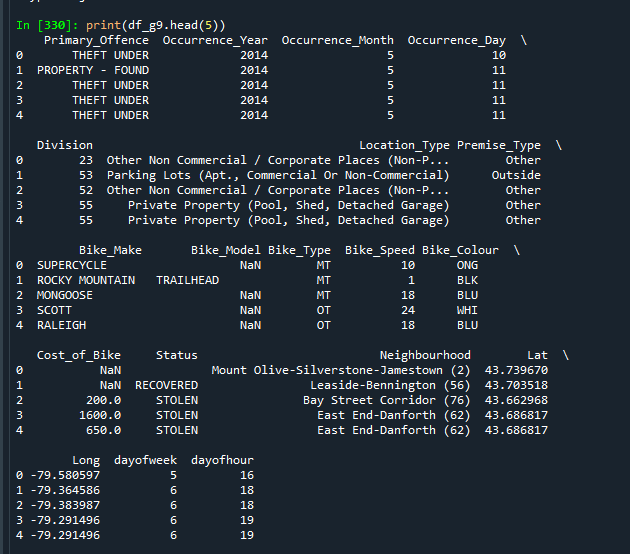
### 1.4.6 New DataFrame(df\_g9) data describe



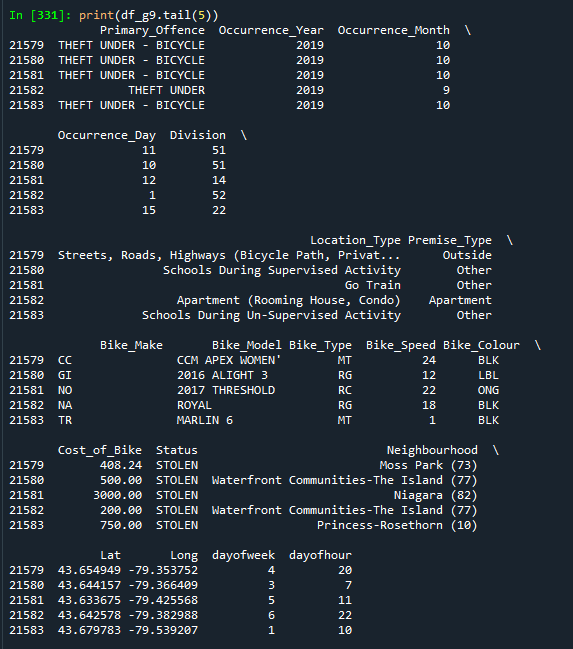
### 1.4.7 New DataFrame(df\_g9) data types



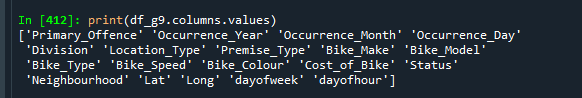
### 1.4.8 New DataFrame(df\_g9) data first 5 rows



### 1.4.9 New DataFrame(df\_g9) data last 5 rows



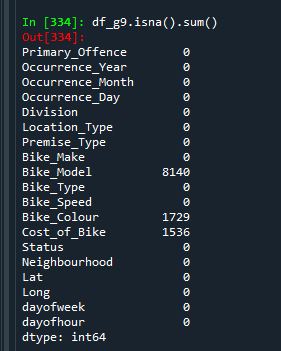
### 1.4.10 New DataFrame(df\_g9) data columns unique values



|  |  |  |  |
| --- | --- | --- | --- |
|  | Column Name | unique | Details |
| 1 | Primary\_Offence | 65 | ['THEFT UNDER', 'PROPERTY - FOUND', 'MISCHIEF UNDER', 'B&E',  "B&E W'INTENT", 'THEFT OVER', 'THEFT FROM MOTOR VEHICLE UNDER',  'THEFT UNDER - BICYCLE', 'ROBBERY - MUGGING',  'CARELESS DRIVING- HTA', 'POSSESSION PROPERTY OBC UNDER',  'THEFT UNDER - SHOPLIFTING', 'THEFT OF MOTOR VEHICLE',  'FRAUD OVER', 'DOMESTIC INCIDENT', 'ROBBERY WITH WEAPON',  'ROBBERY - OTHER', 'THEFT FROM MOTOR VEHICLE OVER',  'FTC PROBATION ORDER', 'POSSESSION PROPERTY OBC OVER',  'PROPERTY - RECOVERED', 'PROPERTY - LOST',  'UNLAWFULLY IN DWELLING-HOUSE', 'INCIDENT',  'MISCHIEF - INTERFERE W-PROP', 'FRAUD UNDER',  'MISCHIEF - ENDANGER LIFE', 'SUSPICIOUS INCIDENT',  'ROBBERY - SWARMING', 'ASSAULT WITH WEAPON', 'ASSAULT',  'ROBBERY - HOME INVASION', 'ARR/WARR EXECUTED NO ADDED CHG',  'AGGRAVATED ASLT PEACE OFFICER', 'THEFT OVER - BICYCLE',  'THEFT OF EBIKE UNDER $5000', 'OTHER FEDERAL STATUTE OFFENCES',  'LIQUOR - INTOXICATED', 'THEFT OF EBIKE OVER $5000',  'THREAT - PERSON', 'TRAFFICKING PROPERTY OBC UNDER',  'MHA SEC 16 (FORM 2)', 'OBSTRUCT PEACE OFFICER',  'TRESPASS AT NIGHT', 'INFORMATION ONLY',  'FRAUD - IDENTITY/PERS W-INT', 'PUBLIC MISCHIEF',  'FTC WITH CONDITIONS', 'POSSESSION HOUSE BREAK INSTRUM',  'FIRE - DETERMINED', 'DRUG - POSS COCAINE (SCHD I)',  'CARRYING CONCEALED WEAPON', 'MISCHIEF TO VEHICLE',  'INVALID GO - RMS ONLY', 'DRUG - TRAF CANNABIS (SCHD II)',  'DRUG - TRAF OTHER (SCHD I)', 'ASSAULT - RESIST/ PREVENT SEIZ',  'THEFT FROM MAIL / BAG / KEY', 'ROBBERY - BUSINESS',  'DRUG - POSS METH (SCHD I)', 'WEAPON - POSS DANGEROUS PURP',  'ASSAULT - FORCE/THRT/IMPEDE', 'B&E OUT', 'ASSAULT PEACE OFFICER',  'ASSAULT BODILY HARM'] |
| 2 | Occurrence\_Year | 6 | [2014, 2015, 2016, 2017, 2018, 2019] |
| 3 | Occurrence\_Month | 12 | [ 5, 4, 6, 8, 7, 9, 3, 2, 1, 10, 11, 12] |
| 4 | Occurrence\_Day | 31 | 1- 31 |
| 5 | dayofweek | 7 | [5, 6, 4, 3, 0, 1, 2] |
| 6 | dayofhour | 24 | 0-23 |
| 7 | 'Bike\_Colour' | 234 |  |
| 8 | 'Bike\_Model' | 7010 |  |
| 9 | 'Bike\_Type' | 13 |  |
| 10 | 'Bike\_Make' | 725 |  |
| 11 | 'Bike\_Speed' | 62 |  |
| 12 | 'Cost\_of\_Bike' | 1457 |  |
| 13 | 'Division' | 18 | [23, 53, 52, 55, 11, 51, 14, 42, 54, 31, 22, 32, 13, 43, 12, 41, 33, 58] |
| 14 | 'Location\_Type' | 44 | ["Other Non Commercial / Corporate Places (Non-Profit, Gov'T, Firehall)",  'Parking Lots (Apt., Commercial Or Non-Commercial)',  'Private Property (Pool, Shed, Detached Garage)',  'Apartment (Rooming House, Condo)', 'Universities / Colleges',  'Other Commercial / Corporate Places (For Profit, Warehouse, Corp. Bldg',  'Single Home, House (Attach Garage, Cottage, Mobile)',  'Streets, Roads, Highways (Bicycle Path, Private Road)',  'Schools During Un-Supervised Activity', 'Ttc Subway Station',  'Commercial Dwelling Unit (Hotel, Motel, B & B, Short Term Rental)',  'Open Areas (Lakes, Parks, Rivers)',  'Schools During Supervised Activity', 'Bar / Restaurant',  'Construction Site (Warehouse, Trailer, Shed)',  'Bank And Other Financial Institutions (Money Mart, Tsx)',  'Other Train Tracks', 'Convenience Stores',  'Hospital / Institutions / Medical Facilities (Clinic, Dentist, Morgue)',  'Jails / Detention Centres', 'Go Station',  'Other Train Admin Or Support Facility',  'Religious Facilities (Synagogue, Church, Convent, Mosque)',  'Police / Courts (Parole Board, Probation Office)',  'Homeless Shelter / Mission', 'Retirement / Nursing Homes',  'Go Train', 'Gas Station (Self, Full, Attached Convenience)',  'Unknown',  'Group Homes (Non-Profit, Halfway House, Social Agency)',  'Other Passenger Train',  'Private Property Structure (Pool, Shed, Detached Garage)',  'Ttc Bus Stop / Shelter / Loop',  'Dealership (Car, Motorcycle, Marine, Trailer, Etc.)',  'Ttc Admin Or Support Facility', 'Ttc Bus',  'Other Regional Transit System Vehicle', 'Ttc Subway Train',  'Ttc Light Rail Transit Station', 'Pharmacy', 'Go Bus',  'Other Passenger Train Station', 'Retirement Home',  'Ttc Street Car'] |
| 15 | 'Premise\_Type' | 5 | ['Other', 'Outside', 'Apartment', 'Commercial', 'House'] |
| 16 | 'Neighbourhood' | 140 |  |
| 17 | 'Long' | 4885 |  |
| 18 | 'Lat' | 4874 |  |
| 19 | 'Status' | 4 | [' ', 'RECOVERED', 'STOLEN', 'UNKNOWN' |

## 1.5 Data missing and replacement

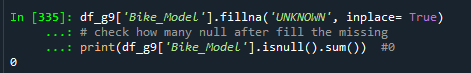
### 1.5.1 Check the missing data



From the results above, only Bike\_Model and Bike\_Color and Bike\_Cost has some null values

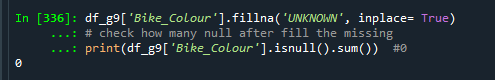
### 1.5.2 Fill the Bike\_Model null by “UNKNOWN”,

Check the null values of Bike\_Model

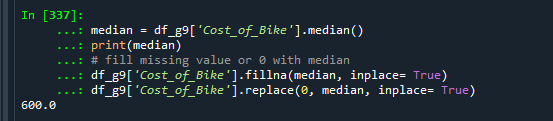


### 1.5.3 Fill the Bike\_Colour null by “UNKNOWN”,

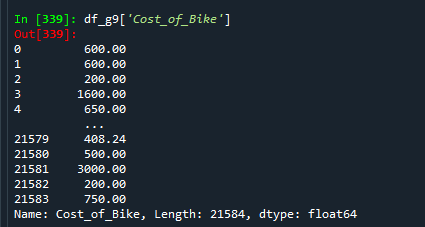
Check the null values of Bike\_Colour



### 1.5.4 Fill missing or 0 of Cost\_of\_Bike with median



Check the Cost\_Of\_Bike after filling median



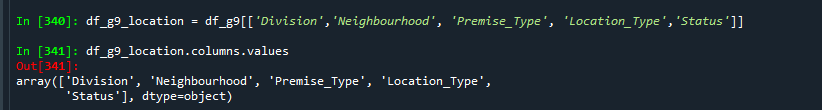
### 1.5.5 Drop row with empty space ‘ ’ in Status

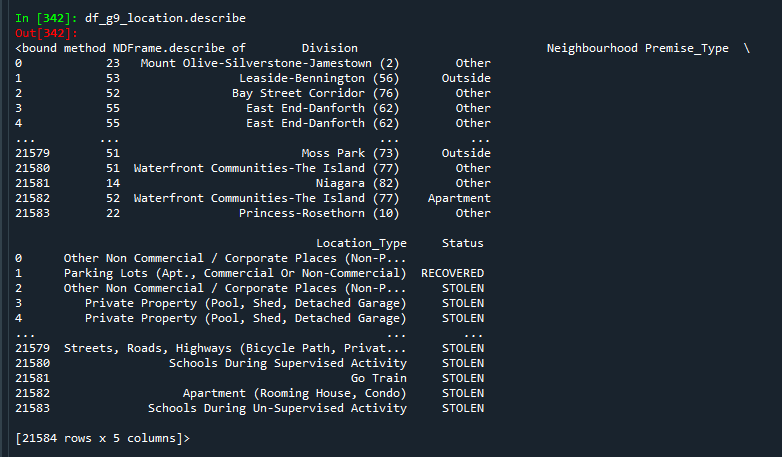
df\_g9 =df\_g9[~df\_g9["Status"].str.contains(' ')]

## 1.6 Group the data with different group base on features

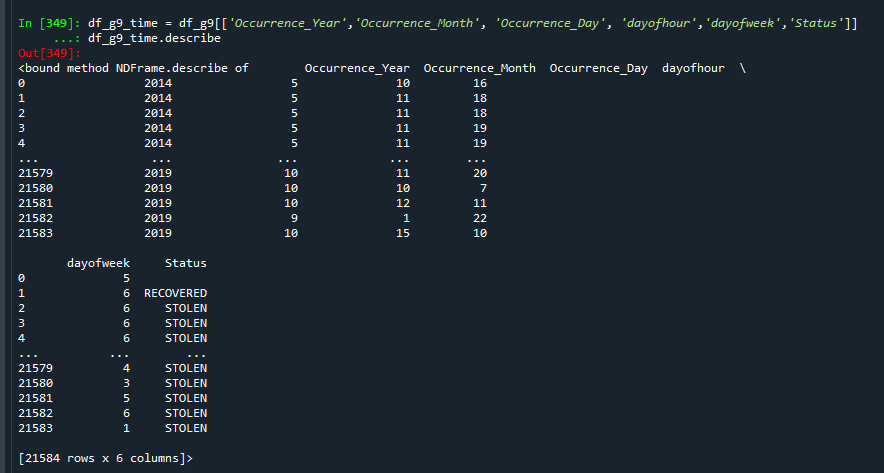
### 1.6.1 Group the data with location information

df\_g9\_location = df\_g9[['Division','Neighbourhood', 'Premise\_Type', 'Location\_Type','Status']]

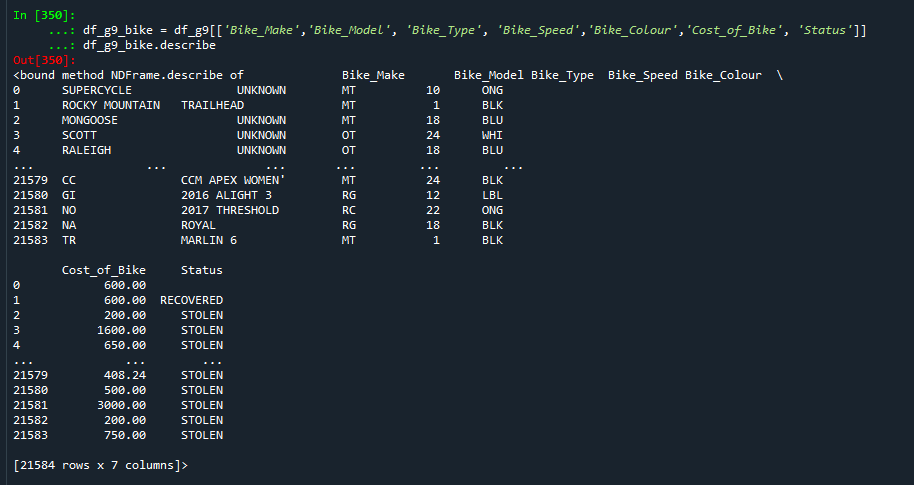




### 1.6.2 Group the data with time/date information



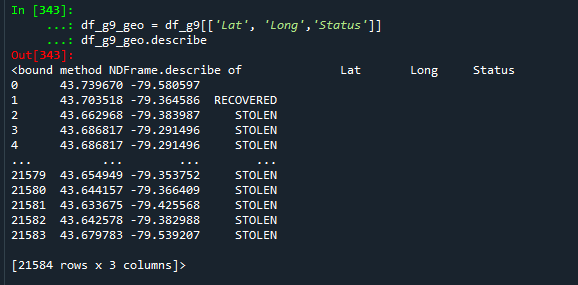
### 2.6.3 Group the data with bike information

****

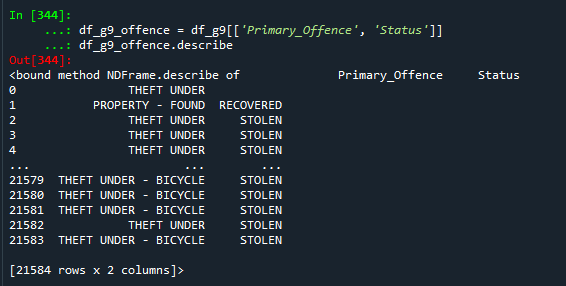
### 1.6.4 Group the data with geo information

df\_g9\_geo = df\_g9[['Lat', 'Long','Status']]

df\_g9\_geo.describe

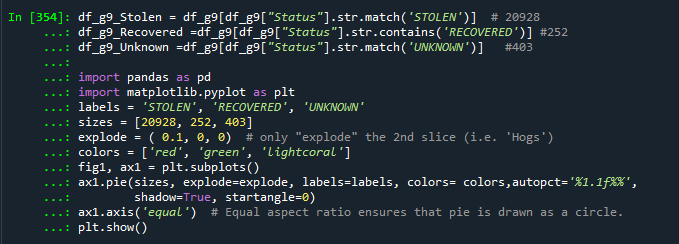


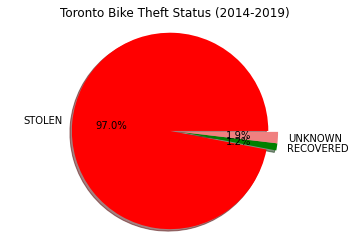
### 1.6.5 Group the data with Primary\_Offence



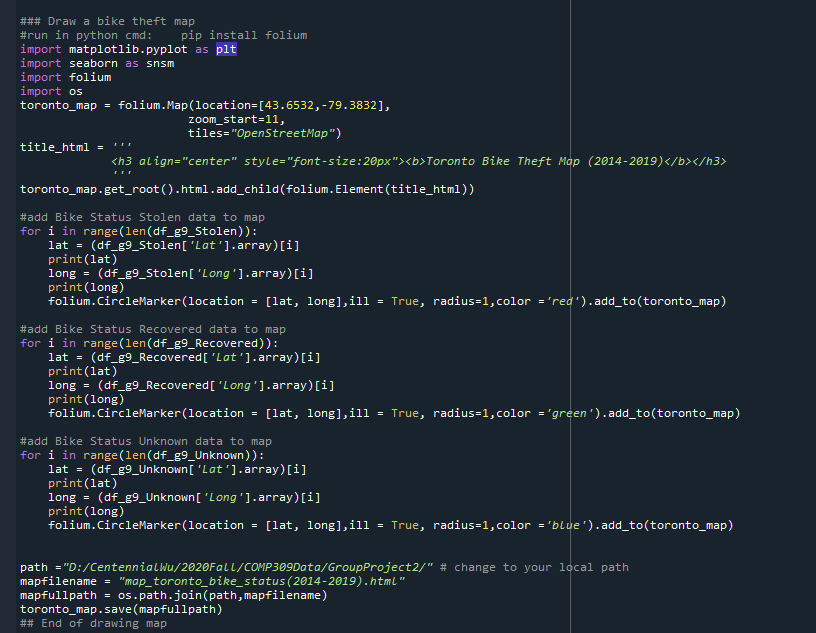
## 1.7 Graphs and Visualization

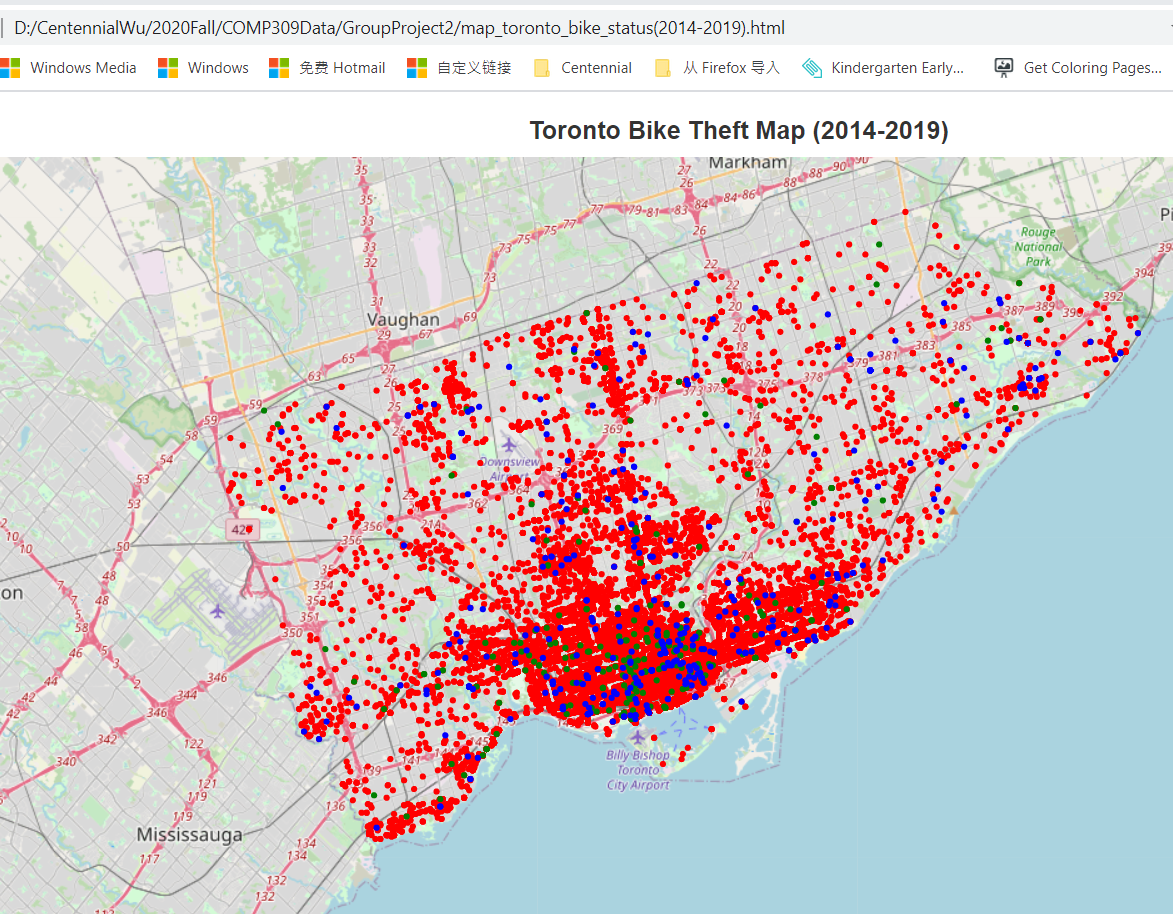
### 1.7.1 Stolen Status Pie Chart



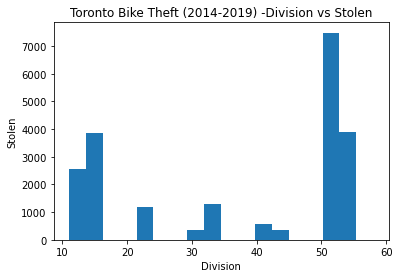


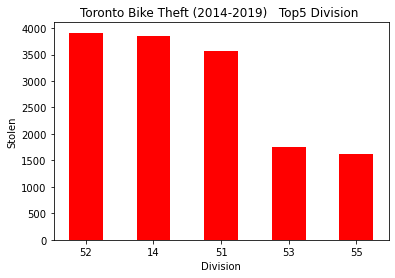
### 1.7.2 Bike Theft Map

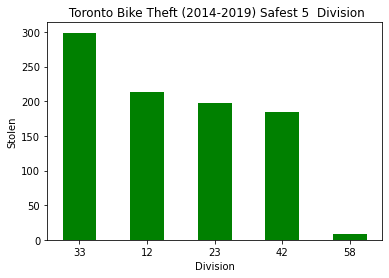




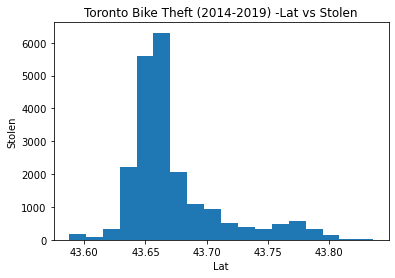
### 1.7.3 Division and Stolen

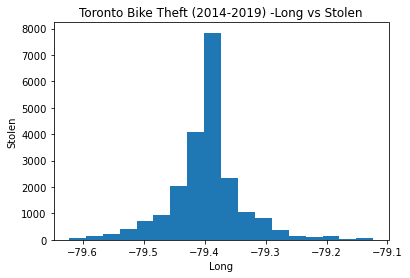


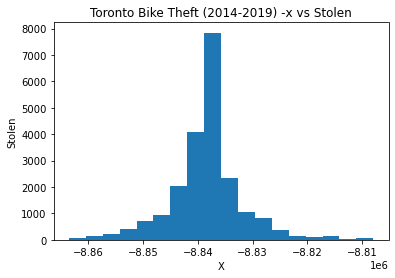


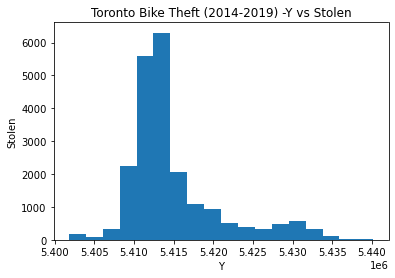


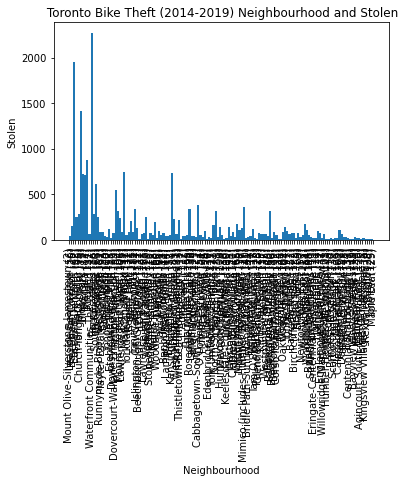
### 1.7.4 Hood\_ID/Neighbourhood, Lat, Long and Stolen

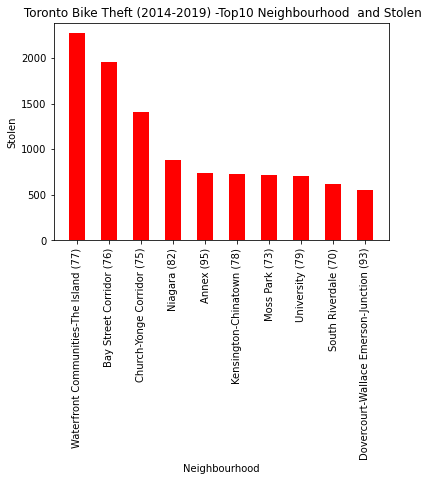


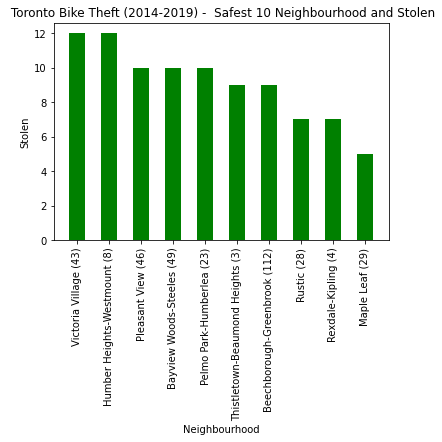




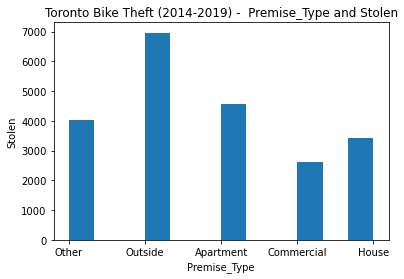


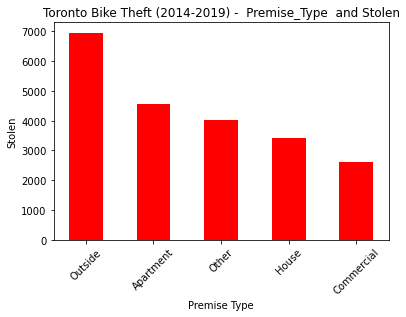




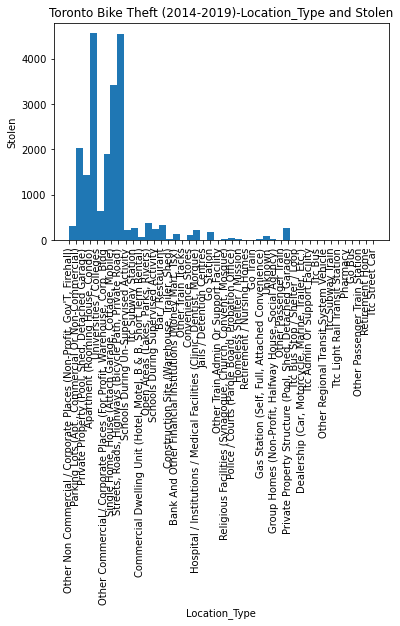


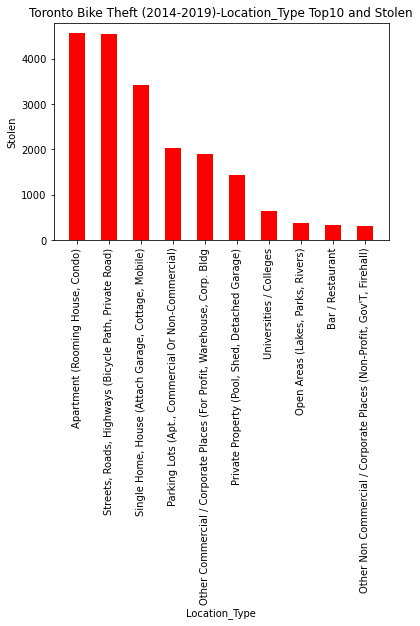
### 1.7.5 Premise\_Type and Stolen

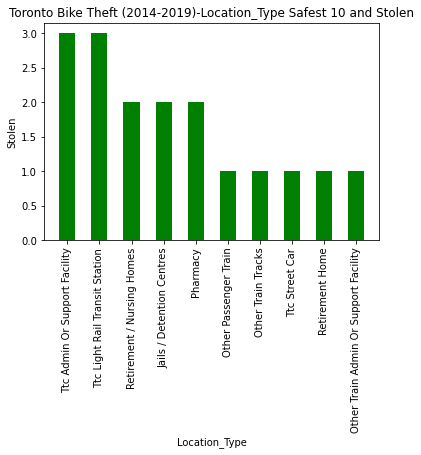




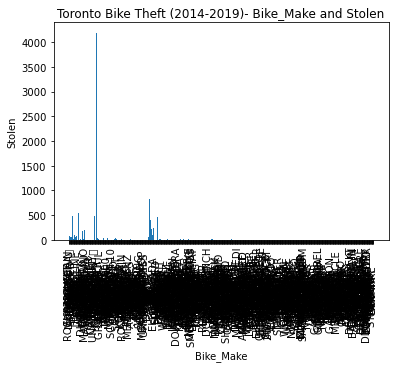
### 1.7.6 Location Type and Stolen

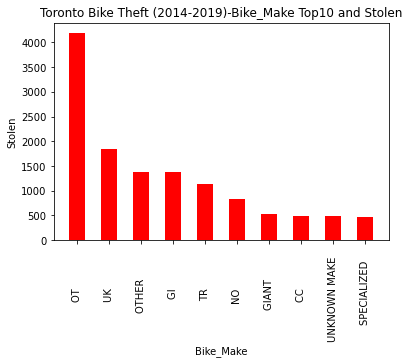


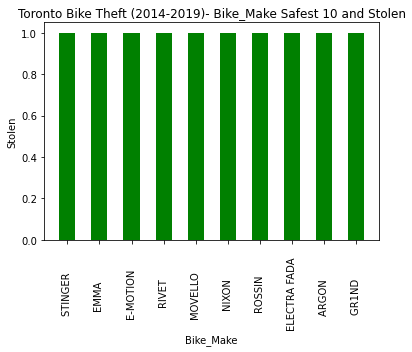




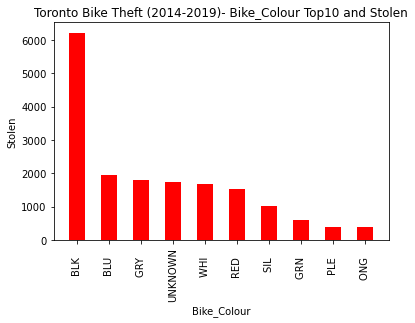
### 1.7.7 Bike\_Make and Stolen

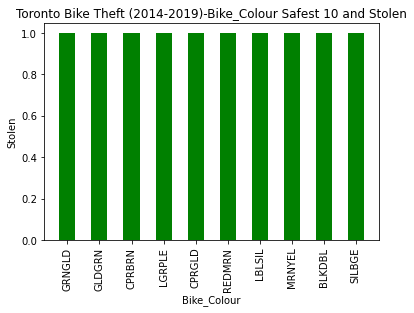




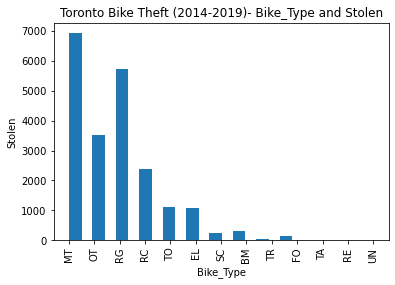


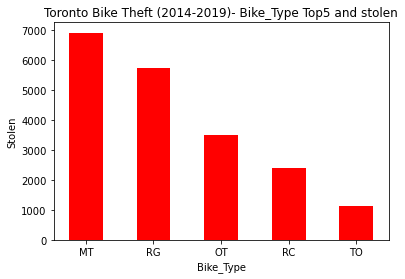
### 1.7.8 Bike\_Colour and Stolen

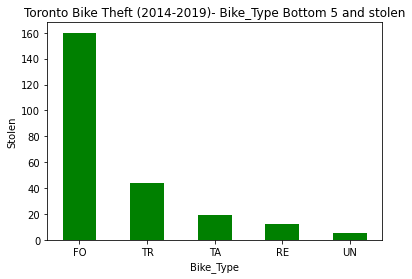




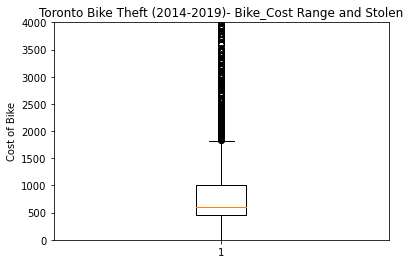
### 1.7.9 Bike\_Type and Stolen

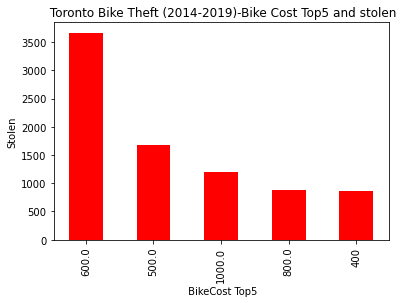




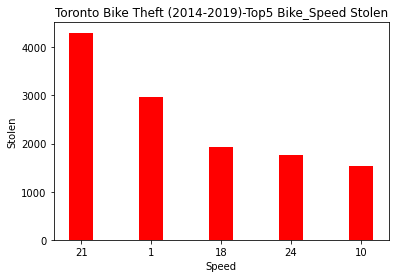


### 1.7.10 Cost\_of\_Bike and Stolen

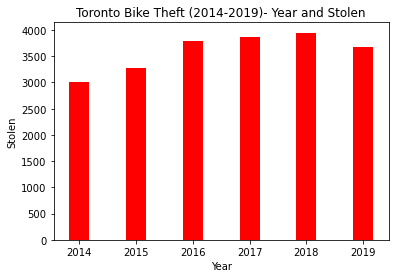




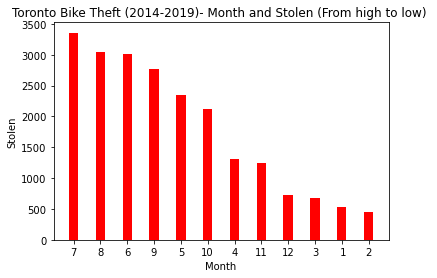
### 1.7.11 Bike\_Speed and Stolen

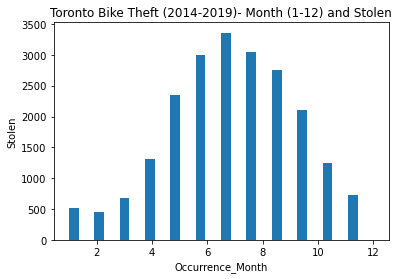


### 1.7.12 Occurrence\_Year and Stolen

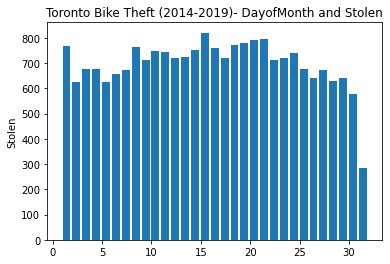


### 1.7.13 Occurrence\_Month and Stolen

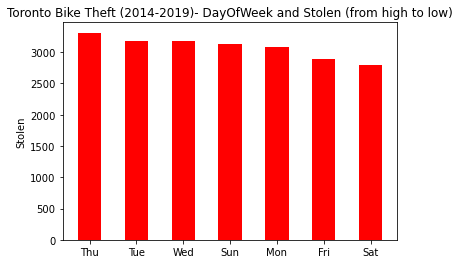




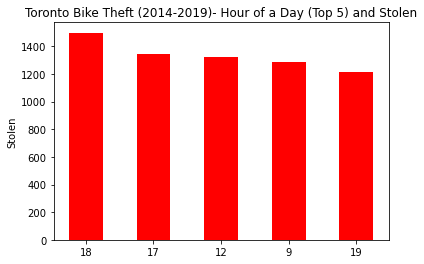
### 1.7.14 Occurrence\_Day and Stolen

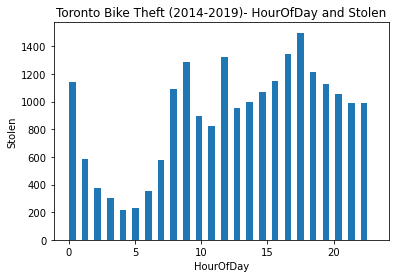


### 1.7.15 Day Of Week and Stolen

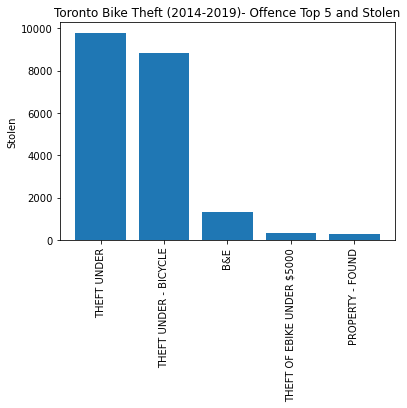


### 1.7.16 Hour of a Day and Stolen





### 1.7.17 Primary\_Offence and Stolen



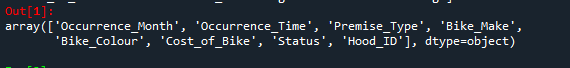
## 1.8 Data Further Selection or Combination Suggestion

|  |  |  |  |
| --- | --- | --- | --- |
|  | Column Name | unique | Suggestion |
| 1 | Primary\_Offence | 2 | Keep ’[‘THEFT UNDER’, ‘THEFT’], and Combined as ‘THEFT’  Other minors Combined as ‘OTHERS’  Or:  Just drop the rows of others |
| 2 | Occurrence\_Year | 6 | No much difference,  [2014, 2015, 2016, 2017, 2018, 2019] |
| 3 | Occurrence\_Month | 12 | Month（6,7,8, 9）- High Risk  Other- Medium or LowRisk |
| 4 | Occurrence\_Day | 31 | No much difference, Drop this column |
| 5 | dayofweek | 7 | No much difference, Drop this column |
| 6 | dayofhour | 24 | 0-23  Hour Sensitive  HighRisk Hour - 9, 12, 17, 18, 19 |
| 7 | 'Bike\_Colour' | 234 | BLK and OTHER |
| 8 | 'Bike\_Model' | 7010 | Topest one is UNKNOWN  So consider drop this column |
| 9 | 'Bike\_Type' | 13 | High Risk [MT, RG, OT, RC]  OTHER |
| 10 | 'Bike\_Make' | 725 | High Risk- [‘OT’]  LowRisk- Other |
| 11 | 'Bike\_Speed' | 62 | High Risk -[1, 10,18,21,24] |
| 12 | 'Cost\_of\_Bike' | 1457 | Most Price range （400-1000） |
| 13 | 'Division' | 18 | HighRisk[52,14, 51, 53]  Low Risk -Other |
| 14 | 'Location\_Type' | 44 | Duplicated with Premise\_Type- Drop this column |
| 15 | 'Premise\_Type' | 5 | ['Other', 'Outside', 'Apartment', 'Commercial', 'House']  Consider group for less catergeries |
| 16 | 'Neighbourhood' | 140 | HighRisk [Top 10 Neighborhood]  Other- Low Risk |
| 17 | 'Long' | 4885 | Keep for Map Analysis |
| 18 | 'Lat' | 4874 | Keep for Map Analysis |
| 19 | 'Status' | 4 | Drop or Combine to [ 'RECOVERED', 'STOLEN'] |

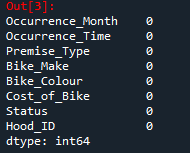
# 2. Data modelling:

## 2.1 Data Cleaning Strategies

Choose columns for using modeling. We choose 'Status','Premise\_Type','Hood\_ID','Occurrence\_Time','Occurrence\_Month','Bike\_Make','Bike\_Colour','Cost\_of\_Bike' because based on data exploring these columns have better feature that can predict result.

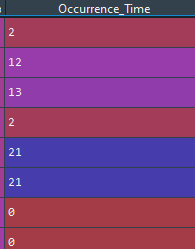


## 2.2 Check missing values from columns

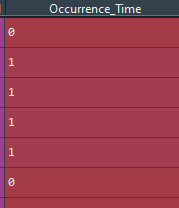


## 2.3 Data transform and category

From Occurrence\_Time to get hour of day



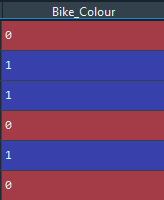
Category Occurrence\_Time. Time in 9, 12, 17, 18, 19 is peak time 1, otherwise is unpeak time 0



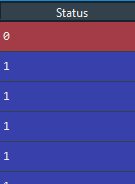
Category bike make into OT, UK, GI, TR, NO, GIANT, CC and other



Convert BikeColor into Black and NonBlack then make it into 1 and 0

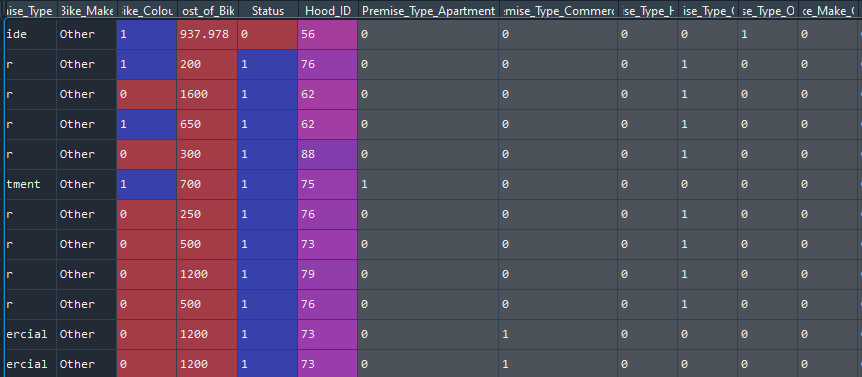


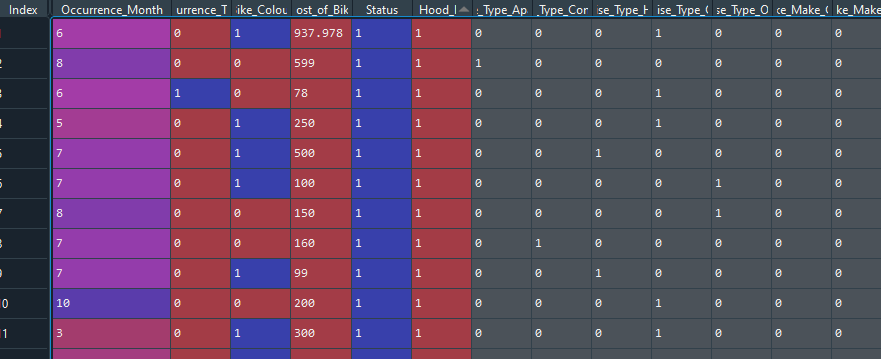
Change Status Column from Object to int stolen and unknow is 1, recover is 0 and drop first row



## 3.4 Categorical data management : Created dummy values

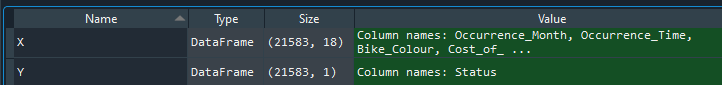
Create dummy column for categorical column

Drop original categorical column



## 3.5 Feature selection

Prepare the data for the model build as X (inputs, predictor) and Y(output, predicted)



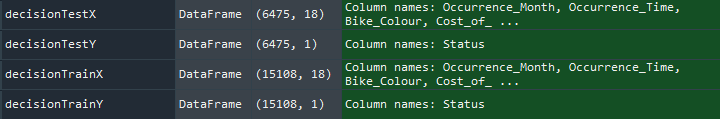
## 3.6 Train, Test data splitting

Split the data into 70% training and 30% for testing for logistic regression





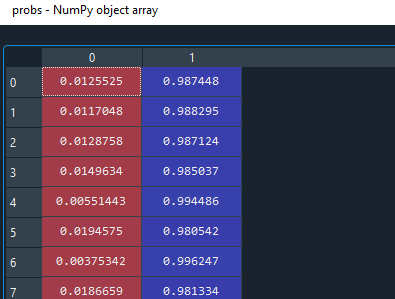
Split the data into 70% training and 30% for testing for decision tree

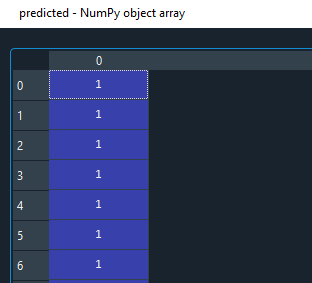


# 3. Predictive model building

## 3.1 Build logistic regression model

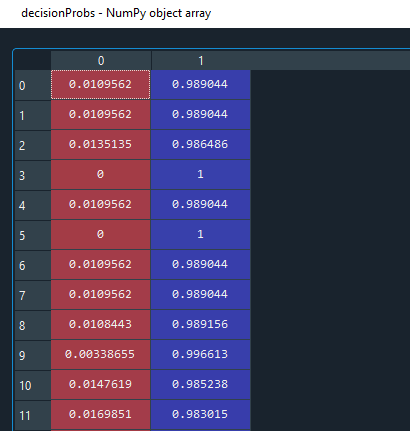


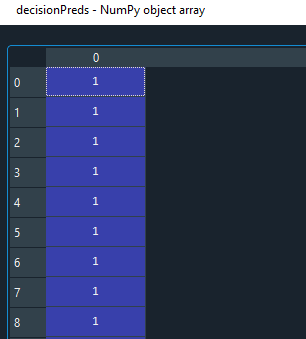


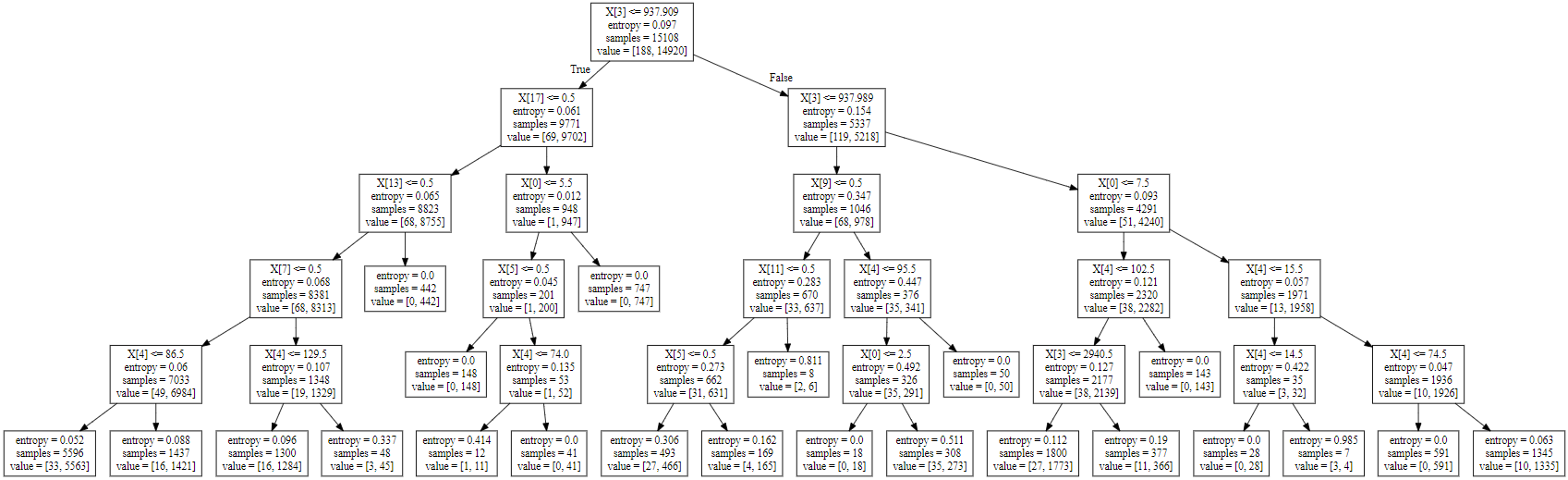


## 3.2 Build the decision tree model and validate the parameters







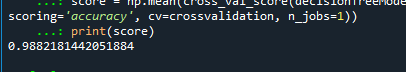


# 4. Model scoring and evaluation

## 4.1 Cross validation score

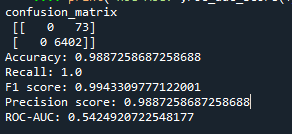


Logistic regression model

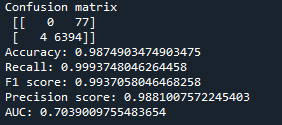


Decision tree model

Both has similar cross validation score

4.2 Confusion\_matrix, accuracy, recall, F1 score, precision score and ROC-AUC

Logistic regression model

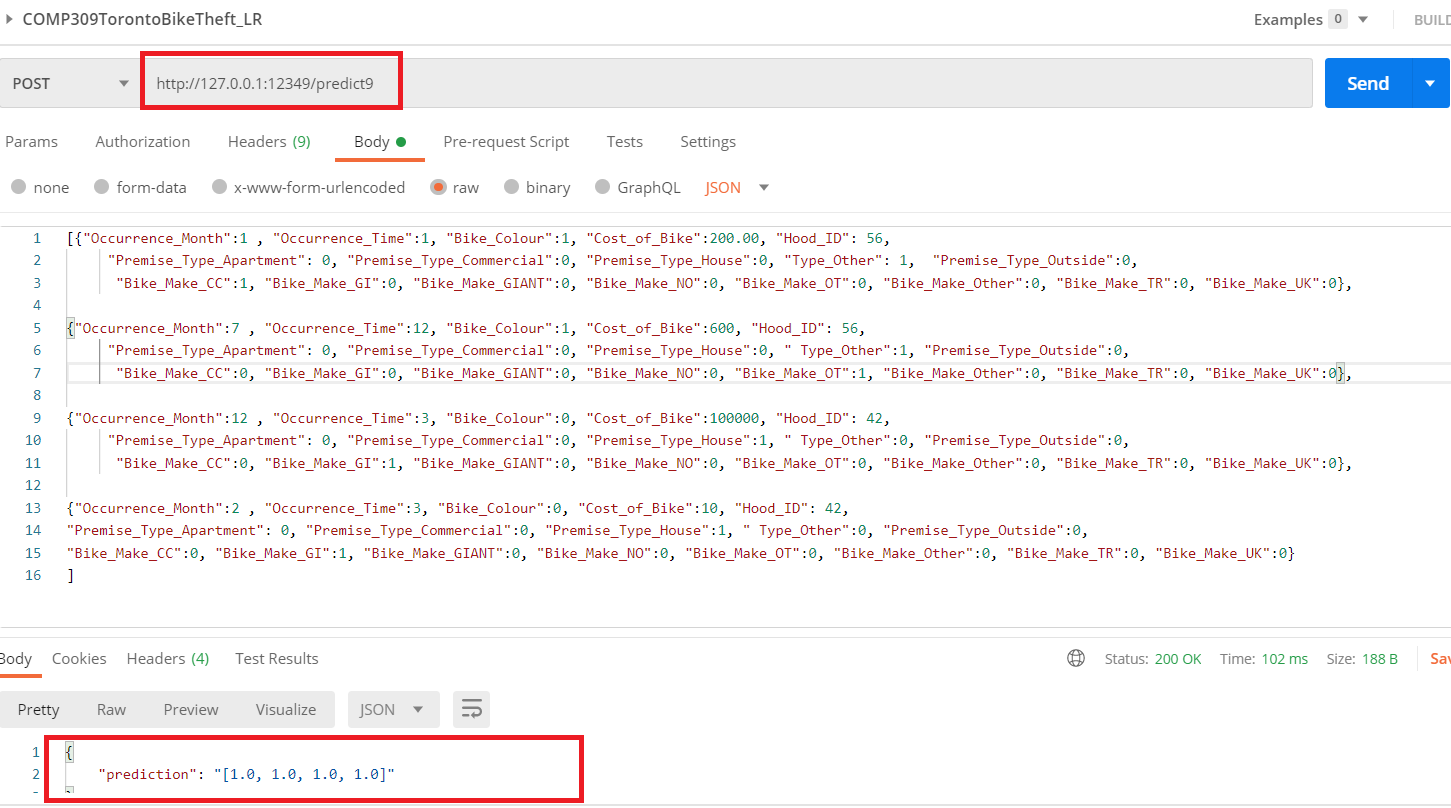


Decision tree model

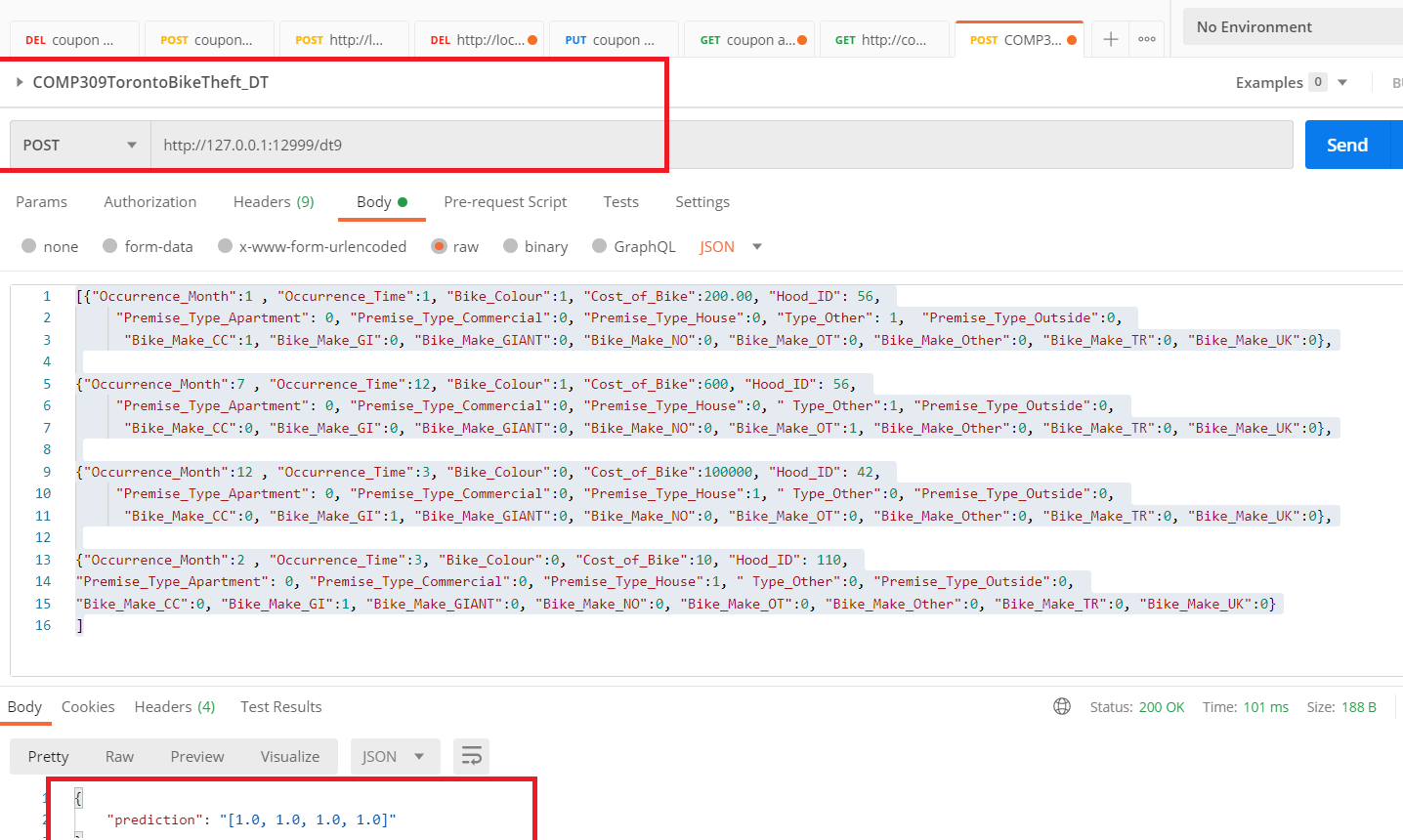
Based on the result of scores, the decision tree model has a better performance. Therefore, we selected the decision tree model as our final model.

# 5. Deploying the model

## 5.1 Logistic Regression Prediction Example



## 5.2 Decision Tree Prediction Example



# 6. References

<https://towardsdatascience.com/a-demonstration-of-carrying-data-analysis-crimes-in-denver-eda-5e852bc75bee>

<https://www.jiristodulka.com/post/toronto-crime/>

<https://medium.com/@kvnamipara/a-better-visualisation-of-pie-charts-by-matplotlib-935b7667d77f>

<https://dreampuf.github.io/GraphvizOnline/>