# Group Project Report

COMP309-006 Group 9

Data Warehouse & Mining – HCIS

11/29/2020

Liping Wu 300958061

Xinglong Lu

Dayoon Lee 301034178

Yoonseop Lee 301037418

Nayul Kim

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

Table of Contents

[0. Executive summary 3](#_Toc57472792)

[1. Data exploration: a complete review and analysis of the dataset including: 4](#_Toc57472793)

[1.1 Load and describe data elements (columns), provide descriptions & types, ranges and values of elements as appropriate. 4](#_Toc57472794)

[1.2 Statistical assessments including means, averages, correlations 8](#_Toc57472795)

[1.3 Missing data evaluations – use pandas, NumPy and any other python packages 8](#_Toc57472796)

[1.4 Graphs and visualizations – use pandas, matplotlib, seaborn, NumPy and any other python packages, you also can use power BI desktop. 8](#_Toc57472797)

[2. Data modelling: 9](#_Toc57472798)

[2.1 Data transformations – includes handling missing data, categorical data management, data normalization and standardizations as needed. 9](#_Toc57472799)

[2.2 Feature selection – use pandas and sci-kit learn. 9](#_Toc57472800)

[2.3 Train, Test data splitting – use NumPy, sci-kit learn. 9](#_Toc57472801)

[2.4 Managing imbalanced classes if needed. Check here for info: https://elitedatascience.com/imbalanced-classes 9](#_Toc57472802)

[3. Predictive model building 10](#_Toc57472803)

[4. Model scoring and evaluation 11](#_Toc57472804)

[5. Deploying the model 12](#_Toc57472805)

[6. Prepare a report explaining your project and detailing all the assumptions, constraints you applied should have the following sections: 13](#_Toc57472806)

# 0. Executive summary

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 Load and describe data elements (columns), provide descriptions & types, ranges and values of elements as appropriate.

import pandas as pd

import os

path = "D:/CentennialWu/2020Fall/COMP309Data/GroupProject2/"

filename = 'Bicycle\_Thefts.csv'

fullpath = os.path.join(path,filename)

print(fullpath)

data\_bicycle\_group09 = pd.read\_csv(fullpath)

data\_bicycle\_group09.columns.values

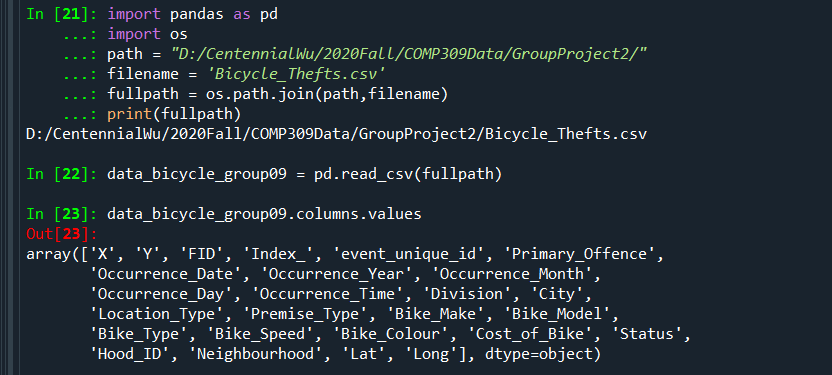
data\_bicycle\_group09.shape

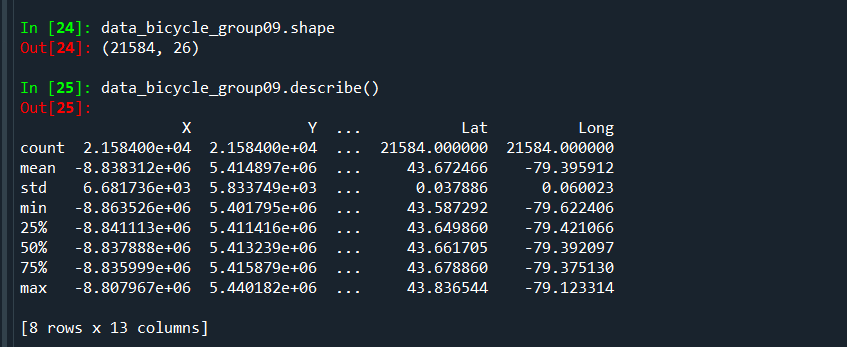
data\_bicycle\_group09.describe()

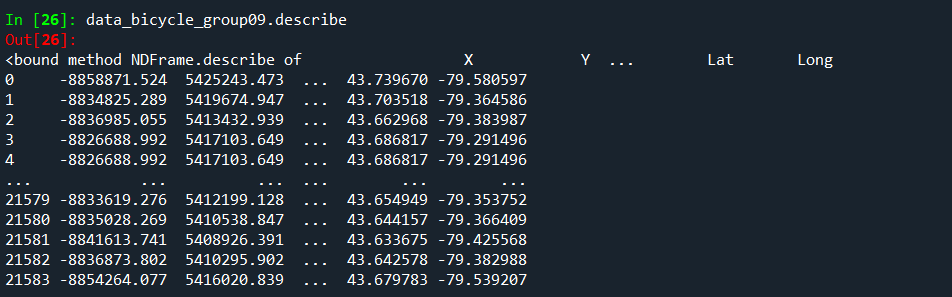
data\_bicycle\_group09.describe

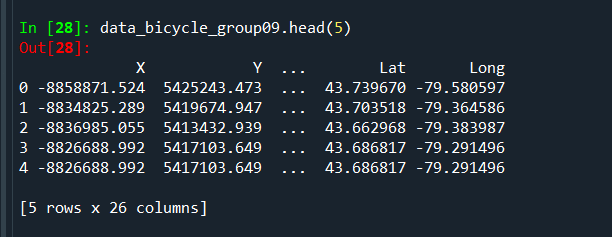
data\_bicycle\_group09.dtypes

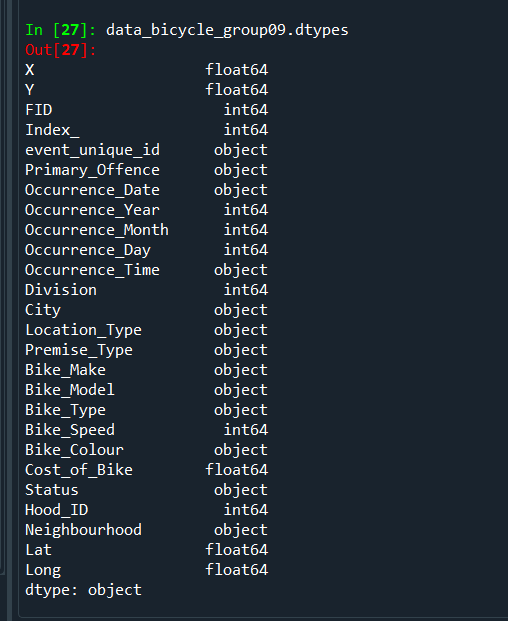
data\_bicycle\_group09.head(5)



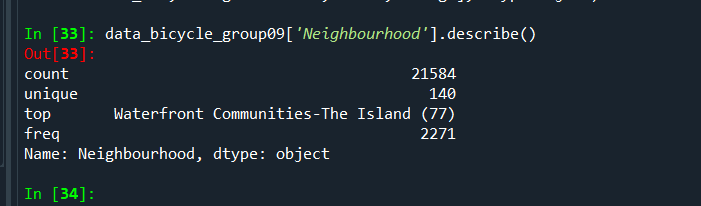


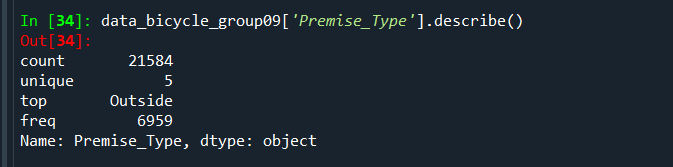


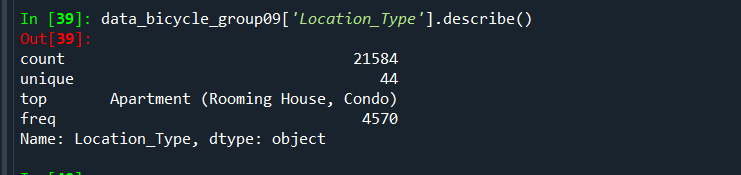


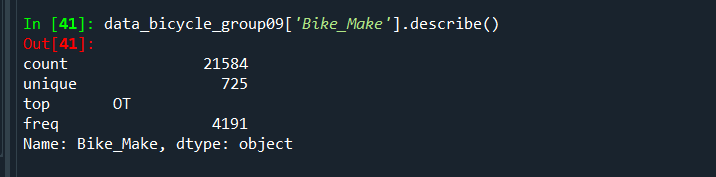


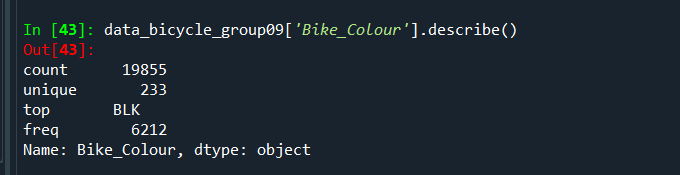
data\_bicycle\_group09['Neighbourhood'].describe()

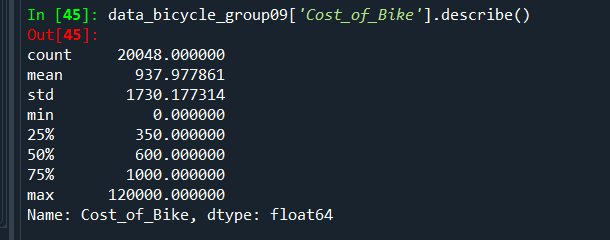


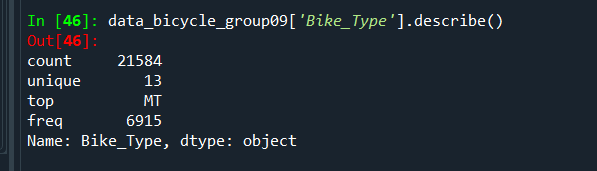


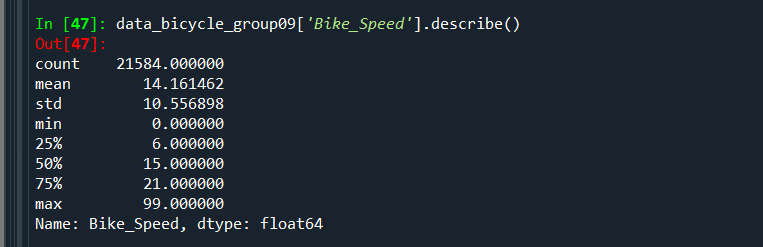












## 1.2 Statistical assessments including means, averages, correlations

## 1.3 Missing data evaluations – use pandas, NumPy and any other python packages

## 1.4 Graphs and visualizations – use pandas, matplotlib, seaborn, NumPy and any other python packages, you also can use power BI desktop.

# 2. Data modelling:

## 2.1 Data transformations – includes handling missing data, categorical data management, data normalization and standardizations as needed.

## 2.2 Feature selection – use pandas and sci-kit learn.

## 2.3 Train, Test data splitting – use NumPy, sci-kit learn.

## 2.4 Managing imbalanced classes if needed. Check here for info: <https://elitedatascience.com/imbalanced-classes>

# 3. Predictive model building

* 1. Use logistic regression and decision trees as a minimum– use scikit learn

# 4. Model scoring and evaluation

* 1. Present results as scores, confusion matrices and ROC - use sci-kit learn
  2. Select and recommend the best performing model

# 5. Deploying the model

1. Using flask framework arrange to turn your selected machine-learning model into an API.
2. Using pickle module arrange for Serialization & Deserialization of your model.
3. Build a client to test your model API service. Use the test data, which was not previously used to train the module. You can use simple Jinja HTML templates with or without Java script, REACT or any other technology but at minimum use POSTMAN Client API.

# 6. Prepare a report explaining your project and detailing all the assumptions, constraints you applied should have the following sections:

* + Executive summary (to be written once nearing the end of project work, should describe the problem/solution and key findings)
  + Overview of your solution(to be written once nearing the end of project work)
  + Data exploration and findings (dataset field descriptions, graphs, visualizations, tools and libraries used….etc.)
  + Feature selection (tools and techniques used, results of different combinations…etc.)
  + Data modeling (data cleaning strategy, results of data cleaning, data wrangling techniques, assumptions and constraints)
  + Model building (train/ test data, sampling, algorithms tested, results: confusion matrixes ...etc.)

7. 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>