Data Mining
Programming Assignment

Fall 2022

Due: Wednesday, September 21, 2022

1 Maryland Traffic Violations

Perform data exploration on the Kaggle Maryland Traffic Violations dataset. Answer the following questions:

1. Which colors of vehicles are more likely to get involved in a traffic violation?

2. Which models of vehicles are more likely to get involved in a traffic violation?

This is an open-ended question. I encourage you to try as many data preprocessing and exploratory analyses as you can possibly do within the time frame. I am ready to be impressed.

Create a GitHub account if you do not have one. Create a public repository for this class, then push your Notebook to that repository.

The purpose of this exercise is to get a warm-up using real-world data.

2 Comments:

1. You can download the data here: https://www.kaggle.com/rounak041993/traffic-violations-in-maryland-county. It's about 500 MB uncompressed. Kaggle Notebook has a limit of 100 GB per dataset, and Google Colab has a limit of 70 GB storage.

2. You may use pluto as it is a powerful server with few restrictions. To work on a data science project on pluto, the easiest way is to install an anaconda under your own directory. Then use SSH Tunneling to access your pluto Notebook from a browser at any place, such as your home. You may Google 'SSH Tunnel Jupyter Notebook' for instructions. If you do not have a pluto account, you may write an email to Mr. Jones. Mr. Jones is the system manager for the CS Department. He can help you create a pluto account.

3. You can also use your own computer.

4. R is also allowed for this homework.

5. The most relevant skill-set you may need for this assignment is Pandas. You may find a quick tutorial here: https://www.kaggle.com/learn/pandas.

6. Try some visualization