Capstone Project 1 Data Wrangling

Datasets: The dataset contains all electronic traffic violations issued in Montgomery County, Maryland. This is a freely accessed public dataset at https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q.

Wrangling Steps Performed:

Traffic Accidents

- This will include looking to see if particular colors of vehicles and timeframes of the year can be predictors for the probability of an automobile getting in an accident.

1. A csv file was downloaded and read into a normalized pandas dataframe, along with parsing dates so that data can be utilized for time series portions of the study.

2. This was also an opportunity to explore which columns to use for this particular project. At this stage in the game it is Vehicle Type, Color, and Accident.

3. Further investigation into the Vehicle Type column showed the counts for the types of vehicles. This led to a focus on only Automobiles because this had the most value and well over one million items. This set would appeal to the clientele of consumers, insurance companies, and car dealerships.

4. The data was checked and scanned for any null information present within the colors listed for automobiles. Those Automobile rows with nan for color were dropped. The data was rechecked to be sure the new dataset was clean, and all Automobiles were identified in a color category.

5. Two additional dataframes were created to have one with Automobile stops marked as being in an accident and the other were those not in an accident. This was the counts for each color were more apparent and easily viewed.

Automobile Data

There now is a list of Automobiles, with rows labeled by Color. There are total counts for the number of each color in total, the number in accidents, and the number not in accidents. Rather than continuously updating the study by query of the API, this set used contains information for all of the years 2012-2019. It was downloaded and read as the csv file and that allows for quicker retrieval in the future.

Features to be used: Vehicle Type, Color, Accident, Date of Stop, Alcohol

Dealing with Outliers

The outliers for more popular colors of Automobiles stood out. Those are not eliminated from the dataset because that may be of interest once viewing the likelihood of each color along with the other variables.