Montgomery County Traffic Violations Data:

An Exploratory and Predictive Study

Patrick Aquino, Adam Imran, Thomas P. Malejko, and Douglas Post

Georgetown University

Montgomery County Traffic Violations Data: An Exploratory and Predictive Study

Abstract—To Be Written

**Background**

To be Written

**The Data**

**The Raw Data**

The principal data used in this research comes from the ‘Traffic Violations’ dataset, which is part of Montgomery County Maryland’s Digital Government Strategy. This dataset contains information about every electronic traffic violation issued in Montgomery County, Maryland from January 1, 2012 to March 4, 2020 (the date of collection for this study). The dataset, in its original form, contains 1.66 million records spanning 43 features—or more than 71 million individual data points. The features include information about the stop itself (i.e. date of the traffic stop, a description of the violation, and location), data about the vehicle involved (i.e. model year, make, and color), and demographics about the driver (i.e. race, gender, state of residence). Features that could be used to identify the specific vehicle, its operator and/or owner, or the ticketing officer are removed by the county prior to the dataset’s publication (*Traffic Violations,* March 4, 2020).

Supplemental data used in this research comes from Visual Crossing Corporation’s Weather Forecast and Historical Weather Data API. This data contains information about hourly weather conditions for Gaithersburg, Maryland from January 1, 2017 through December 31, 2019 and contains information about the temperature, humidity, precipitation, wind speed, etc. for the specified weather station (*Weather Forecast*, March 12, 2020). In total the retrieved dataset has 26,258 records spanning 16 features.

A detailed description of the data (and its features) as applied in this study can be found in Appendix A (Raw Data).

**Data Wrangling and Munging**

As stated above, the Montgomery County Traffic Violations dataset contained more than 1.66 million records spanning 43 features; however, that expansive dataset was too large to process using the techniques learned in Analytics 512. As a result, the dataset was abridged to a three-year period—from January 1, 2017 to December 31, 2019. This decision resulted in a dataset that was still large, but appropriate enough to apply the techniques learned in this course using local computing. Once reduced in size, weather data from the Weather Forecast and Historical Weather Data API was integrated into the traffic violations dataset to provide a more comprehensive set of features, which would allow assessments about the effects that environmental factors have on community policing and traffic accidents. The integration of this data resulted in one minor change to the original—traffic violations—dataset, which was the loss of minute-specificity in the *Time.of.Stop* feature.

The data wrangling process was fairly intensive for this dataset, requiring the removal or modification of several columns due to unclear definitions and large quantities of missing or erroneous values. The table below summarizes the columns removed from the original dataset and the reasons.

|  |  |
| --- | --- |
| **Feature Name** | **Reason for Removal** |
| **Traffic Violations Dataset** | |
| *Date.Of.Stop* | Merged with the *Time.Of.Stop* to form a consolidated *Date.time* feature |
| *Time.Of.Stop* | Merged with the *Date.Of.Stop* to form a consolidated *Date.time* feature |
| *Agency* | Limited values—all records contained ‘MCP’ |
| *Geolocation* | Redundant—contained data about the *Latitude* and *Longitude*, which are individually listed in their own column |
| *Search.Outcome* | Reformatted to indicate whether an arrest occurred due large quantity of missing values. Now called *Arrest.* |
| *Search.Reason.For.Stop* | Large quantity of missing values |
| *Search.Arrest.Reason* | Renamed to *Arrest.Reason* |
| **Weather API Dataset** | |
| *Location* | Unnecessary. All records contained Gaithersburg, Maryland |
| *Resolved.Address* | Unnecessary for this assessment |

*Table 1. Features removed from the dataset and the reasons*

The original merged dataset contained 9.27% missing or erroneously recorded values, which was too large for any future analysis. After the dropping the columns (as detailed in *Table 1)* and conducting additional munging, the dataset cleanliness increased to 92.27% with no values occurring outside their specified range. While this number may still seem large, consider that preponderance of this remaining value are ‘true’ missing values. For example, if a pedestrian is issued a citation for jay-walking there would be no model year recorded since a vehicle was not involved. Similarly, if a search was not conducted pursuant to the traffic stop, there could not be a reason for the search. As a result, the final dataset as applied in this research has an effective cleanliness in excess of 99%.

**Feature Engineering**

To be Written

**Exploratory Data Analysis**

To be Written

**Visualization of the Problem**

To be Written

**Methodology**

**Statistical Analysis**

To be Written

**Linear Models and Comparison**

To be Written

**Results and Interpretations**

To be Written

**Conclusions**

To be Written

References

*Traffic Violations*. (March 4, 2020). dataMontgomery. Retrieved March 4, 2020, from https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q#

*Weather Forecast & Historical Weather Data*. (March 12, 2020). Visual Crossing Corporation. Retrieved March 12, 2020, https://www.visualcrossing.com/weather-data

Appendix A

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Format** | **Description** |
| SeqID | String | Unique identifier for each traffic stop (multiple rows can have the same SeqID if multiple citations were issued, etc.) |
| Date.time | POSITIX | Date and time of traffic stop (time rounded to the nearest hour) |
| SubAgency | Factor | Court code representing the district of assignment of the officer |
| Description | String | Text description of the specific charge |
| Location | String | Text description of the violation (usually an address, intersection, or highway exit) |
| Latitude | Float | Latitude location of traffic violation |
| Longitude | Float | Longitude location of traffic violation |
| Maximum.Temp… | Float | Maximum temperature on the day of the traffic stop |
| Minimum.Temp… | Float | Minimum temperature on the day of the traffic stop |
| Temperature | Float | Temperature at time of the traffic stop |
| Wind.Chill | Float | Windchill at time of the traffic stop (if applicable) |
| Heat.Index | Float | Heat Index at time of the traffic stop (if applicable) |
| Precipitation | Float | Total precipitation during the hour of the traffic stop |
| Snow.Depth | Float | Snow depth at time of the traffic stop |
| Wind.Speed | Float | Average wind speed at time of the traffic stop |
| Wind.Gust | Float | Maximal wind gust at time of the traffic stop |
| Cloud.Cover | Float | Average cloud cover at time of the traffic stop |
| Relative.Humidity | Float | Average relative humidity at time of the traffic stop |
| Conditions | Factor | Description of weather conditions at the time of the stop |
| Accident | Factor | YES if traffic stop involved an accident |
| Belts | Factor | YES if seat belts were used in accident cases |
| Personal.Injury | Factor | YES if traffic violation involved personal injury |
| Property.Damage | Factor | YES if traffic violation involved property damage |
| Fatal | Factor | YES if traffic violation involved a fatality |
| Commercial.License | Factor | YES if driver holds a commercial driver’s license |
| HAZMAT | Factor | YES if traffic violation involved hazardous material |
| Commercial.Vehicle | Factor | YES if vehicle committing the violation is a commercial vehicle |
| Alcohol | Factor | YES if traffic violation included an alcohol related suspension |
| Work.Zone | Factor | YES if traffic violation was in a work zone |
| Search.Conducted | Factor | YES if a person or property search was conducted |
| Search.Disposition | Factor | Resulting outcome of the search |
| Search.Type | Factor | Type of search conducted (person, property, both, etc.) |
| State | Factor | State issuing the vehicle registration (including Canadian Provinces and US Territories) |
| Search.Reason | Factor | The reason for the search (Probable Cause, Warrant, etc.) |
| VehicleType | Factor | Type of vehicle involved in the traffic stop (automobile, light truck, motorcycle, etc.) |
| Year | Int | Year the vehicle was made |
| Make | String | Manufacturer of the vehicle (Ford, Lexus, Mack Truck, Indian, etc.) |
| Model | String | Model of the vehicle |
| Color | Factor | Color of the vehicle |
| Violation.Type | Factor | Violation type: Warning, Citation, or ESERO (Emergency Safety Equipment Repair Order) |
| Charge | String | Numeric code for the specific charge (legal citation) |
| Article | Factor | Article of state law (TA = Transportation Article, MR = Maryland Rules) |
| Race | Factor | Race of the driver |
| Contributed.To.Acc... | Factor | YES if traffic violation was contributing factor to the accident |
| Gender | Factor | Gender of the drive |
| Driver.City | String | City of the driver's home address |
| Driver.State | Factor | City of the driver's home state |
| DL.State | Factor | State issuing the driver's license |
| Arrest | Factor | Did the traffic stop result in an arrest (TRUE if yes) |
| Arrest.Reason | Factor | Reason for the arrest |
| Asset.Type | Factor | Type of asset used to generate the citation (A=Marked Car, Q=Marked Laser, etc.) |

Appendix B

R Code Script

To be Written