ASSIGNMENT: DATA CLEANING & DOCUMENTATION

Our assigned dataset is *Traffic Violations in Montgomery County, Maryland*. It contains information on electronic traffic violations issued within the county as of January 1, 2013 and is updated daily. The dataset contains thirty-five (35) attributes and 823,129 rows as of 10/6/2016 when the dataset was downloaded. The file contains a robust amount of information that is easy to decipher. Information such as time/date/description of violation, vehicle make/model/year, agency(district) of issuance, race/gender of driver, whether or not the violation ended in arrest, fatality, or alcohol related are some of elements contained. The dataset does not contain any information that could uniquely identify the traffic vehicle owner or officer.

The information was obtained from the website, dataMontgomery (<https://data.montgomerycountymd.gov>) and provides the community and any interested parties with various public, open data sets. The county offers a data catalog of approximately 600 datasets from 15 categories such as Business, Elections, Health and Human Services, and Public Safety. The website states, “*The dataMontgomery program seeks to provide residents and constituents with direct access to County datasets in consumable formats, so they may be viewed, sorted and used in various ways, including being potentially leveraged in the development of new applications and services by interested parties. Providing this information offers the public an opportunity to review and analyze raw data, and the opportunity to use it for a variety of purposes.”.1* That being said there are no licensing or intellectual policy constraints on use of the data. The dataset is provided by the Montgomery County Department of Police.

Each record in the file provides a detailed account (metadata) of a traffic violation. Upon review, one can determine when and where the incident occurred, what state the driver’s license was issued, whether or not he/she was wearing a seatbelt and if personal injuries were sustained. The file also identifies if hazardous materials were involved and if the violation contributed to an accident. Specific vehicle information is recorded such as commercial, school bus, motorcycle, and heavy duty. There is also a plain text field describing the specific charge, i.e. driving wrong way on one-way street, failing to stop at red traffic signal, and driving while impaired. Finally, the arrest type of the violation shows if it was done by a marked/unmarked patrol car, radar, station, VACAR (Visual Average Speed Computer and Recorder). Each attribute is documented with a description, data type, and the API field name. The link to the data can be found at: <https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q>.

We decided to use data from 2013 – 2015 since this provides a complete and equally comparable set of data for our research questions. Records from 2016 were removed since the year is still in progress and data still being recorded. The following five additional attributes were added for efficiency or to aid in our research:

1. Index: Unique/Primary Identifier
2. Season: Assigned based on the month of Date\_of\_Stop
   * **Spring**: March –May
   * **Summer**: June- August
   * **Fall**: September – November
   * **Winter**: December – February
3. Rush\_Hour: Assigned based on time of Time\_of\_Stop
   * **Rush Hour – AM:** 06:00:00 AM -09:00:00 AM
   * **Rush Hour – PM:** 04:00:00 PM-6:30:00 PM
   * Non Rush Hour: Times occurring outside of above segments
4. Make\_Color: To provide consistent reporting of car color and to group similar colorings (Mapping table provided under Data Cleaning)
5. Make\_Clean: Car Make appeared to be text field which made entries hugely inconsistent. In all there were 2,386 variations of car make. This field was added to provide some type of consistency in the car make based on matching patterns which whittled down the list to 43 (including a value of *Unknown* where the Make could not be determined).

A group of 1524 records were removed due to missing values. One group of records (843) did not contain entries for Date\_of\_Stop which also corresponded with the Time\_of\_Stop missing. Another group of records were identified (681) where Gender and a majority of other attributes were null except for Date of Stop/Time of Stop/Agency/Description. The Make field, as stated in #5 above, contained unstandardized content due to abbreviations, misspellings, and invalid data. Queries were run based on matching patterns and a new standardized value written to “Make\_Clean”. Examples of pattern matching:

Make

|  |  |
| --- | --- |
| Make: Original Entries | Make Clean: New Mapping |
| Acu,Acrura,Acru,Acira,Accur,Accu | Accura |
| Chrysley,Crystler,Chtysler,Chryl, | Chrysler |
| Toyotaq,Toyova,Toyoval2000,Toy | Toyota |

DATA CLEANING

*NOTE: Prior to starting an original copy of the dataset was saved. All deleted records were copied to a separate table.*

1. Data was imported to MS Access table
2. An object titled, INDEX, was added
3. Query 1: Identify and remove records where DATE OF STOP is null (843 Records; New Record Total: 822286).
4. Query 2: Identify and delete records where TIME OF STOP is null (0 Records; New Record Total: 822286). Records that were missing a TIME OF STOP were same records in Query 1.
5. Query 3: Query and delete records with DATE OF STOP in 2016 (172697 records removed; 649589)
6. Add additional fields for: SEASON and RUSH HOUR
7. Query 4: Run update query to assign seasons by month of DATE OF STOP (175800 records)
   * Spring= 175800
   * Summer=161082
   * Autumn=169830
   * Winter=142877
8. Query 5: Run Update Query to populate RUSH HOUR for
   * 06:00:00 AM -09:00:00 AM = Rush Hour – AM (69317)
   * 04:00:00 PM-6:30:00 PM= Rush Hour – PM (75080)
9. Query 6: Run Update Query to populate RUSH HOUR where IS NULL with NON-RUSH HOUR
   * 505192 Records
10. Extract “MAKE” into a separate table, tbl\_Make with Index Key
11. Add New Object/Field titled, Make\_Clean
12. Query 7: Run update queries to clean Make columns based on matching patterns
13. Add new column, “Color\_Clean”, and remap colors for consistency based on the following mapping structure:

|  |  |
| --- | --- |
| **Original Field: Color** | **New Field**  **Color\_Clean** |
| BLACK | Black |
| "BLUE, DARK" | Blue |
| "BLUE, LIGHT" | Blue |
| BLUE | Blue |
| BROWN | Brown |
| BRONZE | Bronze\_Copper\_Gold |
| COPPER | Bronze\_Copper\_Gold |
| GOLD | Bronze\_Copper\_Gold |
| GRAY | Gray |
| "GREEN, DK" | Green |
| "GREEN, LGT" | Green |
| GREEN | Green |
| MAROON | Maroon |
| CAMOUFLAGE | Multicolor |
| MULTICOLOR | Multicolor |
| N/A | N/A |
| ORANGE | Orange |
| PINK | Pink |
| PURPLE | Purple |
| RED | Red |
| CHROME | Silver |
| SILVER | SIlver |
| BEIGE | Tan\_Beige |
| TAN | Tan\_Beige |
| CREAM | White\_Cream |
| WHITE | White\_Cream |
| YELLOW | Yellow |

1. Query 8: Identified 681 incomplete records where Gender and majority of other attributes were null except for Date of Stop/Time of Stop/Agency/Description.

1Montgomery County Government Open Data Operations Manual, Fiscal Year 2017 Annual Update, dataMontgomery (2016). http://montgomerycountymd.gov/open/Resources/Files/OpenDataOperationsManual.pdf (Links to an external site.) Retrieved October 30, 2016.