# Capstone Project 1 – Milestone Report
# Predicting Automobile Accidents in Montgomery County

**Problem Statement**

Automobile accidents are a part of society today. It is reasonable to believe if there is a way to predict what causes their increased likelihood, the overall society would benefit. The goal of this project is to look at connections in accident frequency for automobiles based on particular factors such the color of the automobile or the time of the year.

**Proposed Solution**

Perform exploratory analysis and predictive modeling from a large dataset of a Montgomery county, Maryland. Insurance agencies, automobile dealerships and manufacturers can benefit in knowing what factors correlate with accidents occurring in automobiles.

- Accurate predictive models can give dealerships a way to detect what colors will sell better based on accident data.
- Accurate predictive models can give insurance companies data to work with if determining cost of particular color vehicles and time of year being driven.
- There is the potential for this to lead to safer automobiles on the roads if there is a high correlation of accidents with particular colors. Automobile manufacturers and dealerships may sway the colors to be safer, or sway their marketing campaigns for having vehicles with lower accident rates.

**Datasets**

Data from Maryland's Montgomery county traffic stop database is used to look at variables that could potentially help predict increased accident likelihood. This is continually updated and easily publicly accessible on the internet.

https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q

dataMontgomery

**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. Additional dataframes were created; Automobile stops marked as being in an accident, Automobiles stops marked as not accident, dates of stops broken down by monthly totals and years. This made counts for each 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.

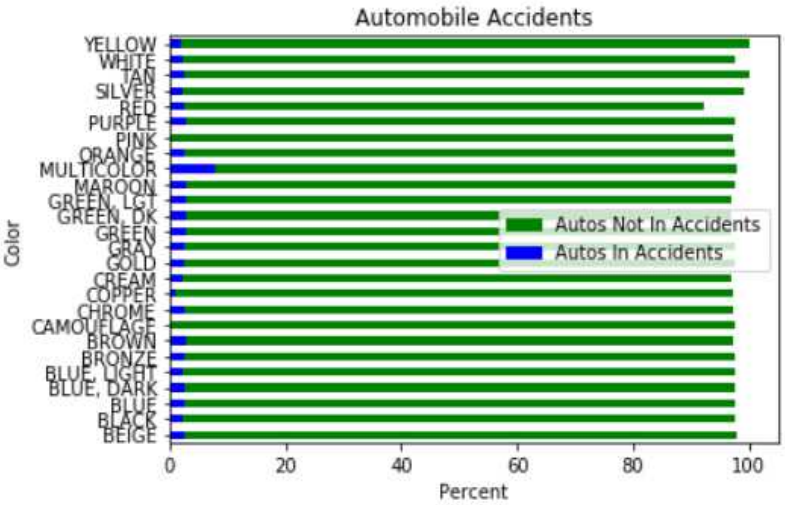
*Features to be used: Vehicle Type, Color, Accident, Date of Stop, Alcohol*

**Dealing with Outliers**

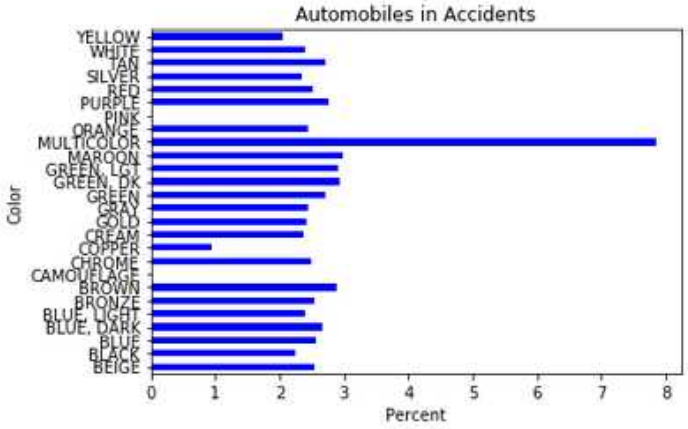
The outliers for more popular colors of Automobiles stood out. Those are not eliminated from the dataset because that may be of interest once viewing the likelihood of each color along with the other variables.

**Exploratory Data Analysis (EDA)**

Once the data was cleaned and wrangled, EDA was performed using statistics, visualizations and hypothesis testing. As an example, this plot below shows the percentages of Automobile Accidents categorized by color, as reported in traffic stops over a period of five years from 2012-2019. It is unknown what automobiles are placed into the multicolor category that appears to be quite an outlier. It is interesting to see that pink and camouflage had no reported accidents.

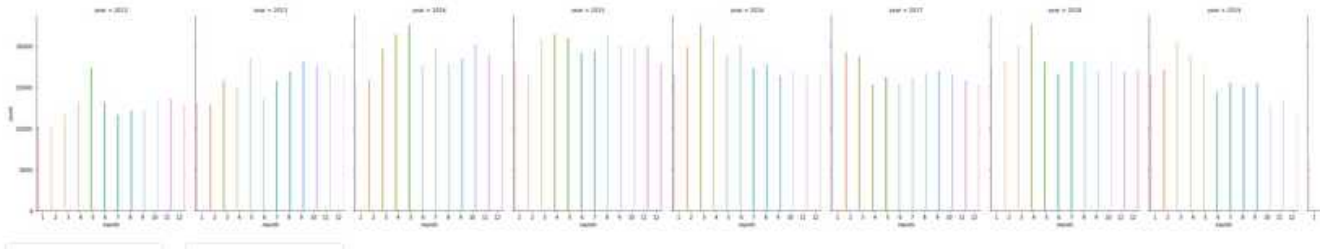



| | |
|---|---|
| BEIGE | 2.527184 |
| BLACK | 2.240856 |
| BLUE | 2.552388 |
| BLUE, DARK | 2.649421 |
| BLUE, LIGHT | 2.385771 |
| BRONZE | 2.547971 |
| BROWN | 2.881844 |
| CAMOUFLAGE | NaN |
| CHROME | 2.500000 |
| COPPER | 0.930233 |
| CREAM | 2.366864 |
| GOLD | 2.408658 |
| GRAY | 2.443378 |
| GREEN | 2.720941 |
| GREEN, DK | 2.930122 |
| GREEN, LGT | 2.898736 |
| MAROON | 2.984191 |
| MULTICOLOR | 7.843137 |
| ORANGE | 2.442415 |
| PINK | NaN |
| PURPLE | 2.758107 |
| RED | 2.517158 |
| SILVER | 2.334393 |
| TAN | 2.708199 |
| WHITE | 2.392629 |
| YELLOW | 2.054795 |



There is interest in looking over the total number of accidents reported in traffic stops each year broken down by month. A spike in 2017 and a low in 2012 is noted. Several months in 2012 are low compared to all months. September, October, and December of 2016 are significantly higher than most months recorded. The causes for this would require further investigation to see if this may have a correlation to being related to less vehicles on the road due to weather, the economy, or other factors. Later in the report, hypotheses are explored to include seeing if this can be predicted, leading to potentially reducing annual and monthly accidents.

| | count | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Year | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 |
| Month | | | | | | | | |
| 1 | 246.0 | 365.0 | 325.0 | 415.0 | 376.0 | 523.0 | 458.0 | 353.0 |
| 2 | 205.0 | 284.0 | 324.0 | 409.0 | 348.0 | 437.0 | 354.0 | 398.0 |
| 3 | 224.0 | 317.0 | 348.0 | 297.0 | 494.0 | 448.0 | 314.0 | 439.0 |
| 4 | 223.0 | 404.0 | 327.0 | 335.0 | 464.0 | 510.0 | 382.0 | 325.0 |
| 5 | 430.0 | 497.0 | 492.0 | 528.0 | 505.0 | 529.0 | 407.0 | 353.0 |
| 6 | 317.0 | 415.0 | 426.0 | 472.0 | 441.0 | 428.0 | 483.0 | 462.0 |
| 7 | 259.0 | 387.0 | 386.0 | 363.0 | 435.0 | 454.0 | 431.0 | 454.0 |
| 8 | 325.0 | 385.0 | 406.0 | 349.0 | 455.0 | 449.0 | 386.0 | 367.0 |
| 9 | 367.0 | 345.0 | 358.0 | 368.0 | 573.0 | 457.0 | 393.0 | 431.0 |
| 10 | 309.0 | 443.0 | 460.0 | 542.0 | 580.0 | 533.0 | 502.0 | 436.0 |
| 11 | 332.0 | 441.0 | 393.0 | 432.0 | 400.0 | 477.0 | 445.0 | 386.0 |
| 12 | 406.0 | 341.0 | 486.0 | 529.0 | 621.0 | 536.0 | 472.0 | 448.0 |



Below shows distribution for the number of automobile stops that were considered accidents each month over the entire span of the years from 2012-2019. The maximum number of accidents in one year through the years 22,601. The minimum being 7,372. The average, mean, per year over all years was about 16,932 accidents.

| | count |
|---|---|
| **count** | 97.000000 |
| **mean** | 408.164948 |
| **std** | 83.727320 |
| **min** | 203.000000 |
| **25%** | 353.000000 |
| **50%** | 409.000000 |
| **75%** | 458.000000 |
| **max** | 621.000000 |

**Hypothesis Testing**

After noticing trends and visualizing the datasets in meaningful formats, it is time to perform hypothesis testing on a couple of the major questions for this study. All hypothesis testing carried out is at the 5% significance level.
Detailed testing can be found in Capstone_1_ Data_Story notebook.

*-Is there a connection between certain colors of automobiles being in more accidents due to their color?*

**Null Hypothesis:**
There is no statistical significance in the likelihood of an Automobile getting into an accident related to color.
**Alternative Hypothesis:**
Certain colors of Automobiles show a higher likelihood for getting in an accident.

A Chi-Square Test was run to discover if there is a significant association between the color of an automobile and its likelihood for being in an accident.

The outcomes were:
Significance level: 0.05

```
Degree of Freedom:  1
chi-square statistic: 488.4484795288744
critical_value: 3.841458820694124
p-value: 0.0
```

Comment:
Therefore the Null Hypothesis is rejected. Accepting the Alternative Hypothesis that certain colors show a greater or reduced likelihood for an Automobile to get into an accident.

*-Can accidents be predicted based on the month?*

**Null Hypothesis:**
There is no statistical significance in the likelihood of an Automobile getting into an accident related to the month.
**Alternative Hypothesis:**
Certain months of the year show a greater or reduced likelihood for an Automobile to get into an accident.

A Paired T-Test was run to discover if there is an association in the likelihood of an Automobile getting in an accident related to the month.

```
6.672627141580349e-69
reject null hypothesis
```

Comment:
The test shows to reject the null hypothesis. Therefore certain months of the year do show a greater or reduced likelihood for an Automobile to get in an accident.

**Conclusion:**
Certain colors of vehicles have a slightly higher risk for being in an accident when looking at the percentages related to each individual color. The risks overall vary by <1%, and therefore it doesn't seem a strong statement to make for clients to base decisions from.
Certain months of the years showed a variance in accidents that could vary by more than 200 compared to other months in the same year. This merits further investigation by a client or company who wishes to use this information. The weather may have played a factor or perhaps the economy.