# Bias Incidents

## Datasets Overview

### Bias incidents dataset:

Montgomery County bias incident reports from 2016 to 2021 (updated weekly).

Source: https://data.montgomerycountymd.gov/Public-Safety/MCPD-Bias-Incidents/7bhj-887p

Variables of interest: incident date, bias code, offense, case status, victim count, and suspect count.

### Hate crimes dataset:

Hate crimes reported in the United States from 1991 to 2018.

Source: https://www.kaggle.com/louissebye/united-states-hate-crimes-19912017?select=hate_crime.csv

Variables of interest: year, state, bias code, offense, victim count, offender count, and offender race.

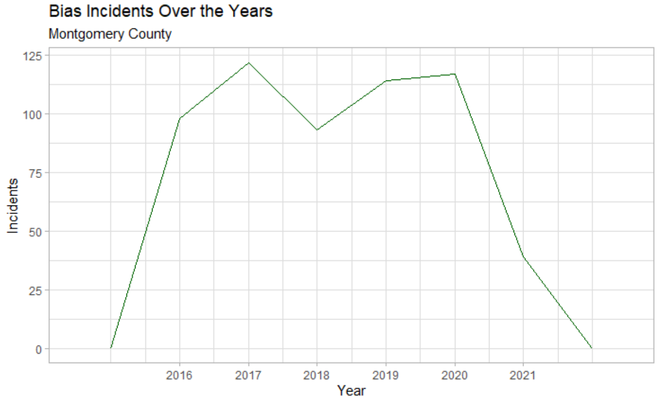
## Data Cleaning

The process of data cleaning was done through R's dyplr package: reordering columns, removing unwanted columns, renaming variables, assigning appropriate data types, adjusting missing value notations, and changing labels for categorical variables.

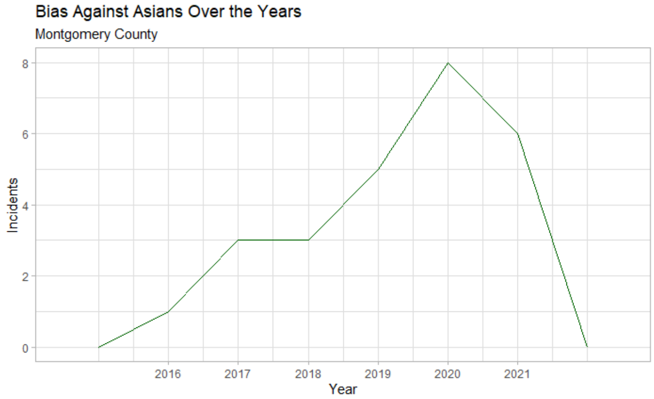
## Research Questions

1- How did hate crimes change throughout the years?
2- Who are the most common victims of hate crimes?
3- Is there an association between the bias code and the offense?
4- Is there an association between the bias code and the offender's race?
5- Are there more open or closed cases?
6- Is the number of incidents for a state proportional to that state's population size?
7- Can the offense be predicted using the information given by the data?
8- What does the distribution of the suspect count look like?
9- What does the distribution of the victim count look like?
10- What recommendations can be given based on the study of these datasets?

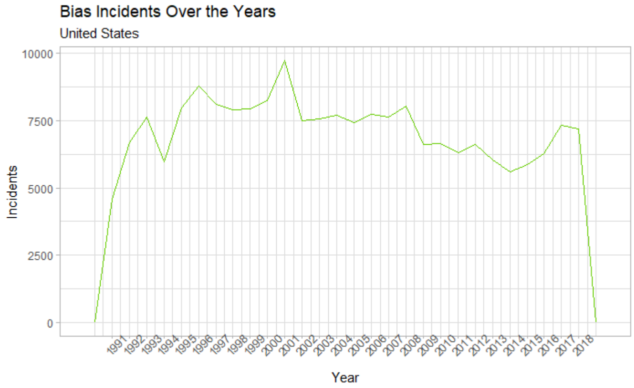
## Year of the Incident



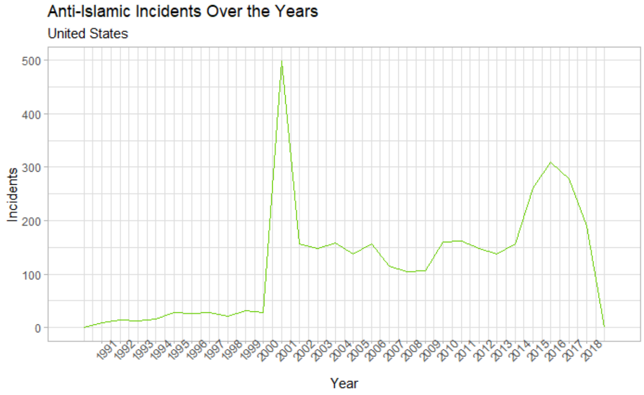
2017 and 2018 are the years with the highest and lowest number of incidents, respectively. Between 2018 and 2020 the incident count increases, reaching a high point. Crimes, in general, decreased in 2020 in Montgomery County because of the coronavirus pandemic, so why is it different for hate crimes? This will be investigated in the next chart.



After filtering the data for hate crimes against Asians, it is visible that the 2020 peak is due to the increased bias against Asians as a result of the pandemic.

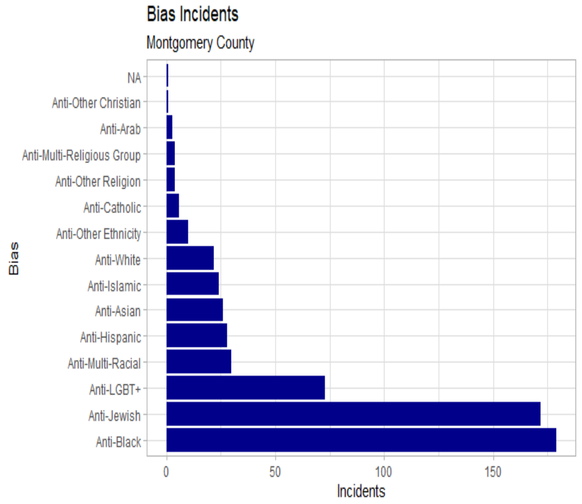


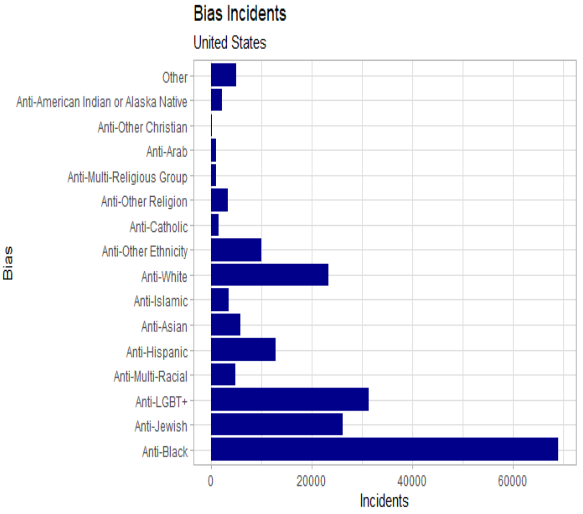
The plot above shows the changes in bias incidents over the years for the United States. It is possible to see that there were many increases and decreases during the available time frame, just as there were for Montgomery County. What catches the eye here is the spike in 2001. Could it be related to 9/11? This will be investigated in the next plot.



As predicted, the 2001 spike is present and stands out even more in the anti-Islamic plot. This means that discrimination against Muslims increased significantly because of 9/11.

## Bias Code of the Incident





For both Montgomery County and the United States anti-black incidents are by far the most common ones. Nationally there are more bias incidents against people from the LGBT+ community than against Jews. Locally the discrimination against Jews is more common, which is probably related to the fact that the Jewish population represents around 10% in Montgomery County (for the US this proportion is less than 3%).

## Bias Code vs Offense

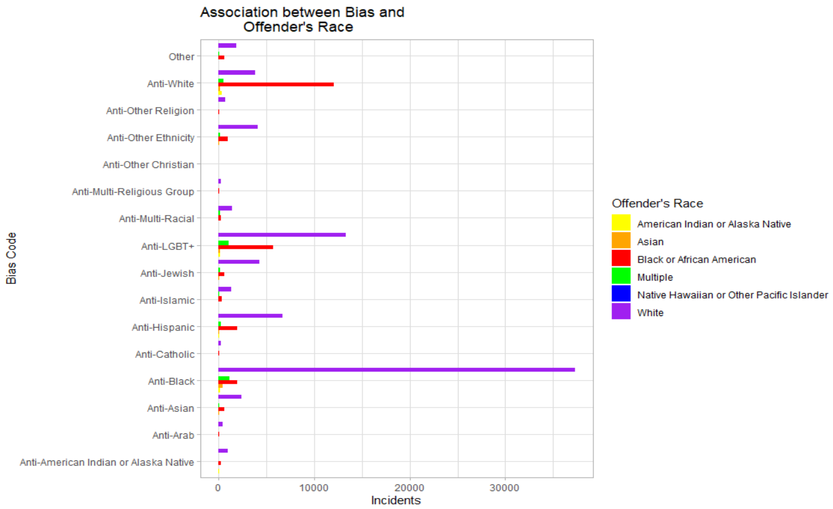
The Montgomery County dataset was used for the following analysis.

**Table 1: Association between bias code and offense.**

| Categories | Arson | Assault | Intimidation | Other | Vandalism | Total |
|---|---|---|---|---|---|---|
| Anti-Arab | 0 | 1 | 1 | 0 | 1 | 3 |
| Anti-Asian | 0 | 6 | 14 | 1 | 5 | 26 |
| Anti-Black | 0 | 24 | 82 | 14 | 59 | 179 |
| Anti-Catholic | 1 | 0 | 0 | 0 | 5 | 6 |
| Anti-Hispanic | 0 | 13 | 11 | 1 | 3 | 28 |
| Anti-Islamic | 0 | 5 | 14 | 1 | 4 | 24 |
| Anti-Jewish | 0 | 1 | 53 | 14 | 104 | 172 |
| Anti-LGBT+ | 0 | 26 | 24 | 5 | 18 | 73 |
| Anti-Multi-Racial | 0 | 0 | 6 | 8 | 16 | 30 |
| Anti-Multi-Religious Group | 0 | 0 | 0 | 1 | 3 | 4 |
| Anti-Other Christian | 0 | 0 | 1 | 0 | 0 | 1 |
| Anti-Other Ethnicity | 0 | 1 | 6 | 1 | 2 | 10 |
| Anti-Other Religion | 0 | 0 | 1 | 0 | 3 | 4 |
| Anti-White | 0 | 10 | 7 | 0 | 5 | 22 |
| Total | 1 | 87 | 220 | 46 | 228 | 582 |

Since African Americans, Jews, and LGBT+ people are the most common victims of bias incidents, it is unsurprising that they present the highest counts on almost all categories of offense. Black people usually suffer more intimidation than other offenses and Jews suffer more vandalism (and almost no assault). For the remaining bias codes: Hispanics usually suffer more assaults, Asians, and Muslims suffer more intimidation, and multi-racial people usually suffer more vandalism. To investigate the relationship further, an Independency Test was performed. Since the resulting p-value was exceedingly small, the hypothesis of independency was rejected, and it was concluded that offense and bias code are associated. To quantify this association, the Contingency Coefficient was calculated and the value of 0.562 was found, which represents a moderate association between the two variables.

## Bias Code vs Offender's Race



There are 89,302 missing values for the offender's race variable, but since this national dataset is large (201,376 rows), this does not affect the analysis much—there is still plenty of information to work with. The most prominent color in this chart is purple, which represents white offenders, and then red, which represents black offenders. White offenders are spread out across all bias codes (but have an especially high count for Anti-Black), while black offenders are concentrated on Anti-White, Anti-LGBT+, and Anti-Hispanic bias.

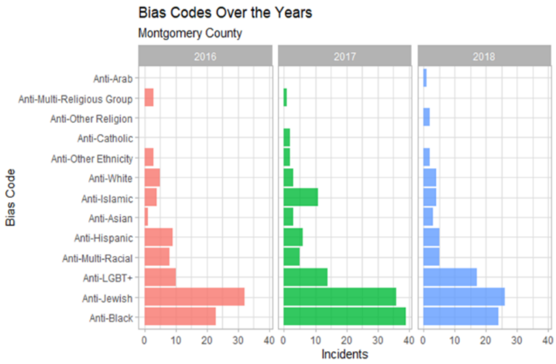
## Case Status

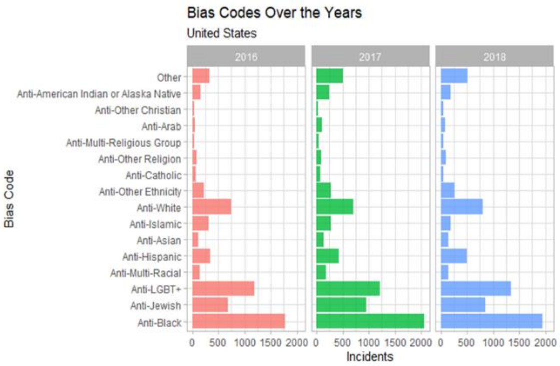
### Table 2: Distribution of frequency for the status of the case.

| Case Status | Frequency | Relative Frequency (%) | Relative Frequency for non-missing data (%) |
|---|---|---|---|
| **Open** | 229 | 39.3 | 56.7 |
| **Closed** | 112 | 19.2 | 27.7 |
| **Inactive** | 63 | 10.8 | 15.6 |
| **NA** | 179 | 30.7 | - |

This is a variable that is available in the Montgomery County dataset. Around 30% of its data are missing. Regarding only the non-missing data: more than half of the cases are still open, 27.7% of them are closed and over 15% are inactive.

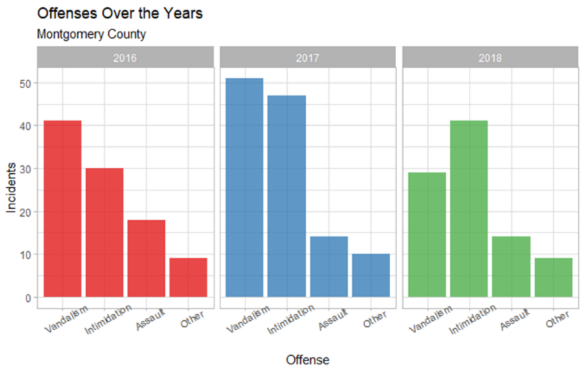
## Year vs Bias Code

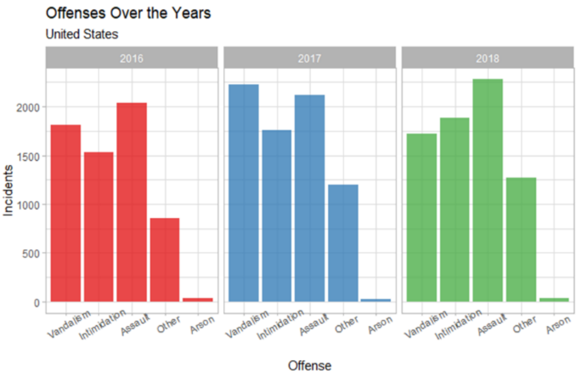




The overall distribution of bias codes in the US remains the same during the different years of the 2016-2018 time frame. For Montgomery County, it is possible to see that Anti-Black is the most common bias code only in 2017, since in 2016 and 2018 it is Anti-Jewish.

## Year vs Offense

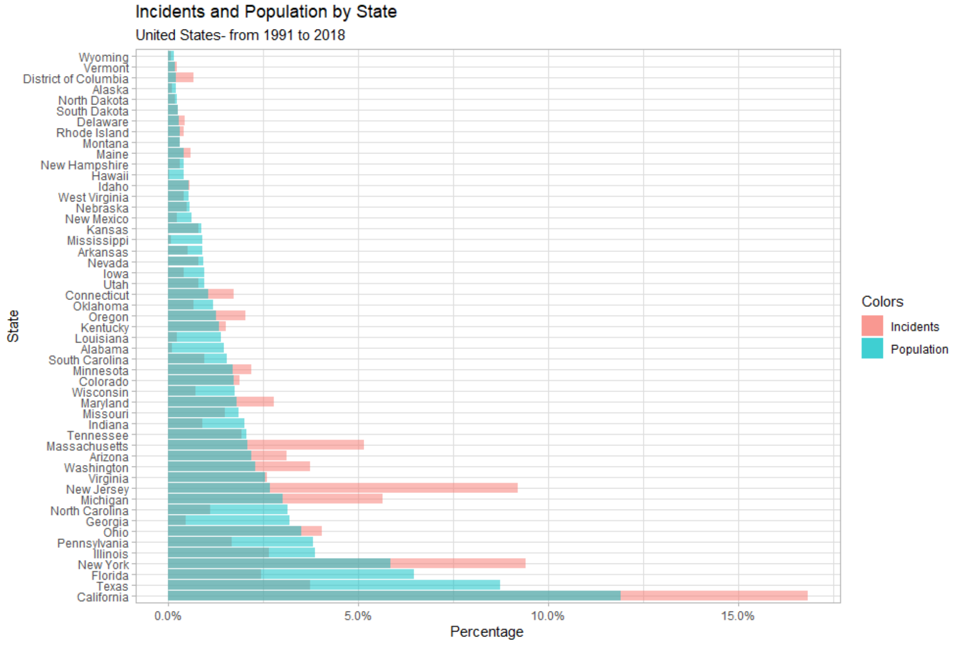




For MC, vandalism is the most common offense in 2016 and 2017, but not in 2018 (it is intimidation). For the US: in 2016 the most frequent offense is assault, followed by vandalism, and then intimidation; in 2017 the most frequent is vandalism, followed by assault, and then intimidation; in 2018 the most frequent is assault, followed by intimidation, and then vandalism.

## States

A Goodness of Fit Test was performed on the state variable. The objective of this test is to check if a variable follows a certain probability distribution. In this case, the probability distribution in question is the distribution of each state's population in proportion to the United States population. With an exceedingly small p-value, the hypothesis stating that the variable fits the distribution was rejected—meaning that the frequency of bias incidents for each state is not proportional to that state's population. The next figure shows the existing difference between the frequency of incidents and the population proportion.



The states that have the blue bar exceeding the pink bar are the ones that supposedly have a smaller proportion of hate crimes. The results were not what most would have expected, so after some research, it was concluded that these differences are due to the use of different systems for reporting incidents. The hate crime dataset from Kaggle contains information collected through the National Incident-Based Reporting System (NIBRS), but many states make use of other systems. Michigan, for example, is one of the states that reports 100% of its crimes via NIBRS. Georgia, on the other hand, only started using NIBRS in 2018 and therefore is underrepresented in the data.

## Predicting Offenses

A logistic regression model was built to make predictions about the offense from Montgomery County's dataset. The resulting accuracy (correct predictions over total predictions) is 66.9%, which is not great. The following table shows the confusion matrix of the model:

**Table 3: Confusion Matrix for Logistic Regression Model with Offense as the Target Variable.**

| Offense | Assault | Intimidation | Other | Vandalism | Total (True Count) |
|---|---|---|---|---|---|
| **Assault** | 15 | 14 | 0 | 0 | 29 |
| **Intimidation** | 5 | 46 | 1 | 9 | 61 |
| **Other** | 0 | 10 | 2 | 5 | 17 |
| **Vandalism** | 2 | 12 | 0 | 54 | 68 |
| **Total (Predicted Count)** | 22 | 82 | 3 | 68 | 175 |

These values led to the following measures of quality for the model:

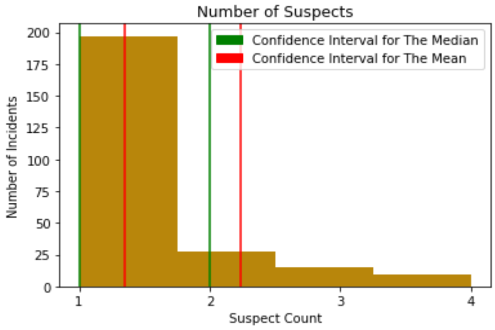
**Table 4: Classification Report for the Logistic Regression Model with Offense as the Target Variable.**

| Offense/ Measure | Precision | Recall | F1 Score |
|---|---|---|---|
| **Assault** | 0.68 | 0.52 | 0.59 |
| **Intimidation** | 0.56 | 0.75 | 0.64 |
| **Other** | 0.67 | 0.12 | 0.20 |
| **Vandalism** | 0.79 | 0.79 | 0.79 |

The precision indicates how many of the values that were predicted as being part of that category were actually part of it. The recall indicates how many of the values that were actually a part of that category were correctly predicted as such. The F1 Score derives from the first two measures and is always a number in between them. The best precision was for vandalism: 79% of the offenses that were predicted as vandalism were actually vandalisms. The category that had the second-highest recall was intimidation: 75% of the incidents that involved intimidation had their offenses correctly predicted.

## Number of Suspects

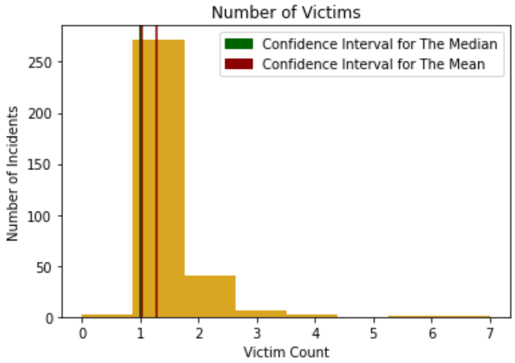
The following analysis was done using the Montgomery County dataset.



The confidence intervals shown by the plot were obtained through bootstrapping simulations for both the median and mean number of suspects. As the histogram shows, most bias incidents have only one suspect and higher values are rare. Therefore, the distribution is right-skewed, and the median is a better representative of the center since it is not influenced by outliers like the mean. In this case, the point estimator for the median has a value of 1 and the confidence interval is between 1 and 2.

## Number of Victims

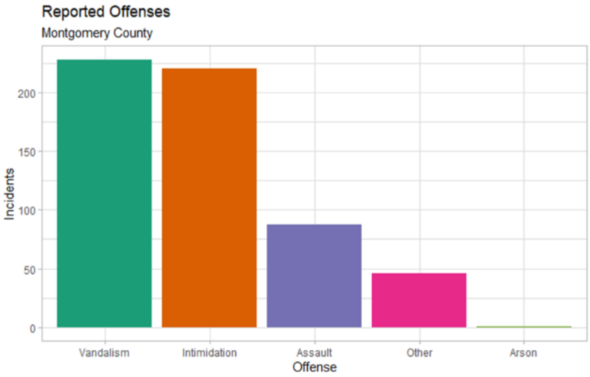
The following analysis was done using the Montgomery County dataset.

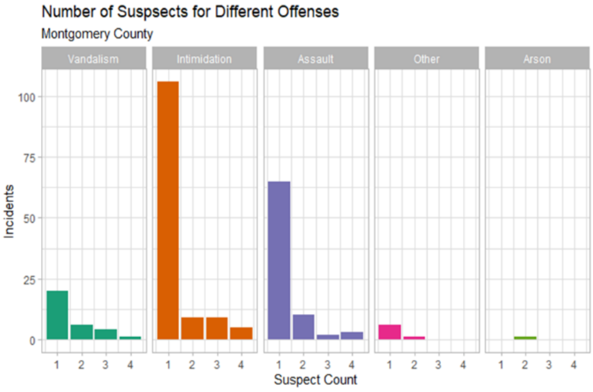


The same process conducted for the suspect count was also used for the victim count. Again, the distribution is right-skewed and has 1 as the most frequent number of victims. Since the median is the better measure of center, looking at its point estimator (1) and confidence interval (between 1 and 1) is informative, showing that the dataset is basically composed of one-victim incidents.

## Recommendations

1. For dataMontgomery: collect more information on the incidents' locations and add it to the dataset.





From the first image, it is possible to see that vandalism is the most frequent offense in bias incidents. However, suspect counts are surprisingly low for this type of offense, as seen on the second plot. This means that there are not many suspects being identified for vandalism cases. A way of both lowering the chances of vandalism and facilitating the identification of suspects is to install cameras and better illumination in locations likely to be vandalized. Adding a column about where the incidents took place (parks, malls, schools, etc.) makes it possible to pinpoint these common vandalism locations and take the recommended precautions.

2. Spread information about support services for hate crime victims (Maryland's helpline, Victim Assistance and Sexual Assault Program – VASAP, local Office of Victim Services, etc.) through fliers in Black neighborhoods, pride centers, and synagogues or through social media groups/ pages that are usually accessed by African

Americans, Jews, and people from the LGBT+ community (since they are the most frequent victims of hate crimes).

3. Focus awareness campaigns on middle/ high school students, offering mini-courses about diversity.

**Table 5: Proportion of Suspects from Each Age Group.**

| **Age Groups** | **Younger than 18** | **Between 18 and 35** | **Between 36 and 45** | **Between 46 and 55** | **Older than 55** |
|---|---|---|---|---|---|
| **Incident Count** | 32.45% | 27.66% | 12.23% | 11.17% | 16.49% |

As the table shows, the most frequent age group for suspects is "younger than 18 years old". That is why it is so important to create awareness for kids that are still in middle school or high school.

## Acknowledgments

I thank Prof. Iapalucci for guiding me through all phases of the project, Prof. Saidi for helping me improve my presentation skills, Prof. Mohamed for introducing me to the machine learning universe, Victoria Lewis for managing the amazing dataMontgomery portal from where I chose my main dataset, and the Montgomery College staff for being so welcoming and kind to me from the start.

**Juliana Rosa.**

**05/04/2021.**