**Colorado Traffic Stops**

Springboard DSC Capstone 2

Lisa Moller

**1 Introduction**

One of the most common interactions between police and the public are traffic stops. With over 50,000 stops per day[[1]](#footnote-1) in the United States, there is a lot of data to track, and analyze, to have a better idea of who is being stopped, why they were stopped and what the outcome was. This information is helpful to know to understand if resources are being used wisely, and to draw attention to conscious or unconscious biases in police behavior. We will look specifically at Colorado traffic stop data.

There are two groups who may be clients for this type of analysis, government watchdog groups, and police departments. Government watchdog groups would use the analysis to raise awareness for issues found. Police Departments may invest in the analysis to ensure resources are being used wisely and to proactively combat any biases through trainings or other necessary actions.

**2 Dataset**

The dataset used in this project came from the Stanford Open Policing Project.[[2]](#footnote-2) The dataset was retrieved from the following website: <https://openpolicing.stanford.edu/data/>.

The data contains one record for each stop between January 2010 and March 2016. The data features are…

* stop date
* stop time
* county name
* police department
* driver gender
* driver age
* driver race
* violation
* an indicator for whether a search was conducted
* search type
* whether or not contraband was found
* stop outcome
* whether an arrest was made
* the vehicle type
* whether the driver was from out of state

The dataset included ‘raw’ columns that recorded the data received from police departments, from which Stanford Open Policing Project inferred standardized values. For example, ‘location\_raw’ contained “[t]he original data value from which we compute the county (or comparably granular location) in which the stop occurred”[[3]](#footnote-3) We chose use the standardized values provided by the Stanford Open Policing Project as opposed to the raw values.

**3 Data Wrangling**

Generally Data Wrangling is finding, structuring, cleaning and adding to your data, validating and preparing for analysis. Because we used the dataset from the source mentioned above, much of the finding/structuring/cleaning/validating was already completed. A fairly clean dataset was provided.

To get a sense of the data, we counted features and rows, detailed the type of each feature, highlighted the categorical columns, looked for missing values and counted unique values within each feature.

After dropping the ‘raw’ columns, the dataset contained 18 features and 2,584,744 rows. Of the 18 features, two were datetime data types, one was age, which we treated as categorical, and the rest were categorical. Ten columns had varying numbers of null values. We waited until further into the analysis to decide what to do with the nulls. The number of unique values in each feature was interesting to look at. Both 'violation' and 'vehicle\_type' had quite a lot of unique values, 1,953 and 154,137 respectively. While a large number of unique values, we kept the columns as they were through the exploratory analysis. Officer ID was mostly unique, and the majority of Officer gender was null, so neither feature was used in the analysis or model.

**4 Exploratory Analysis**

Exploratory analysis helps form an intuition about the dataset. With our goal to better understand who is getting stopped, to uncover the conscious or unconscious biases, the exploratory analysis was a very important task. We looked at the data in two parts: the information present when the stop was made, and the actions after the stop, for example, whether a search was conducted.

Below are several interesting observations. More exploratory analysis completed for the dataset can be found at: [https://github.com/lmoller/Springboard-Capstone-2/3 Exploratory Data Analysis Cap 2.ipynb](https://github.com/lmoller/Springboard-Capstone-2/3%20Exploratory%20Data%20Analysis%20Cap%202.ipynb)**.**

**4.1 Data Present When the Traffic Stop Was Made**

4.1.1 Traffic Stops by Age

We counted the number of stops for each age and ordered them in descending order. The top-12 ages stopped were ages 19-30 years old. This may be because they were newer drivers who most likely drove faster. 16 and 17 year olds may not be included in the top-12 because they generally still live at home where a parent may do more driving. 16 and 17 year olds are also generally in a school close to their home, so they are not driving as far for their daily activities. Once they are 18, they have a bit more responsibility and freedom to drive farther/more.

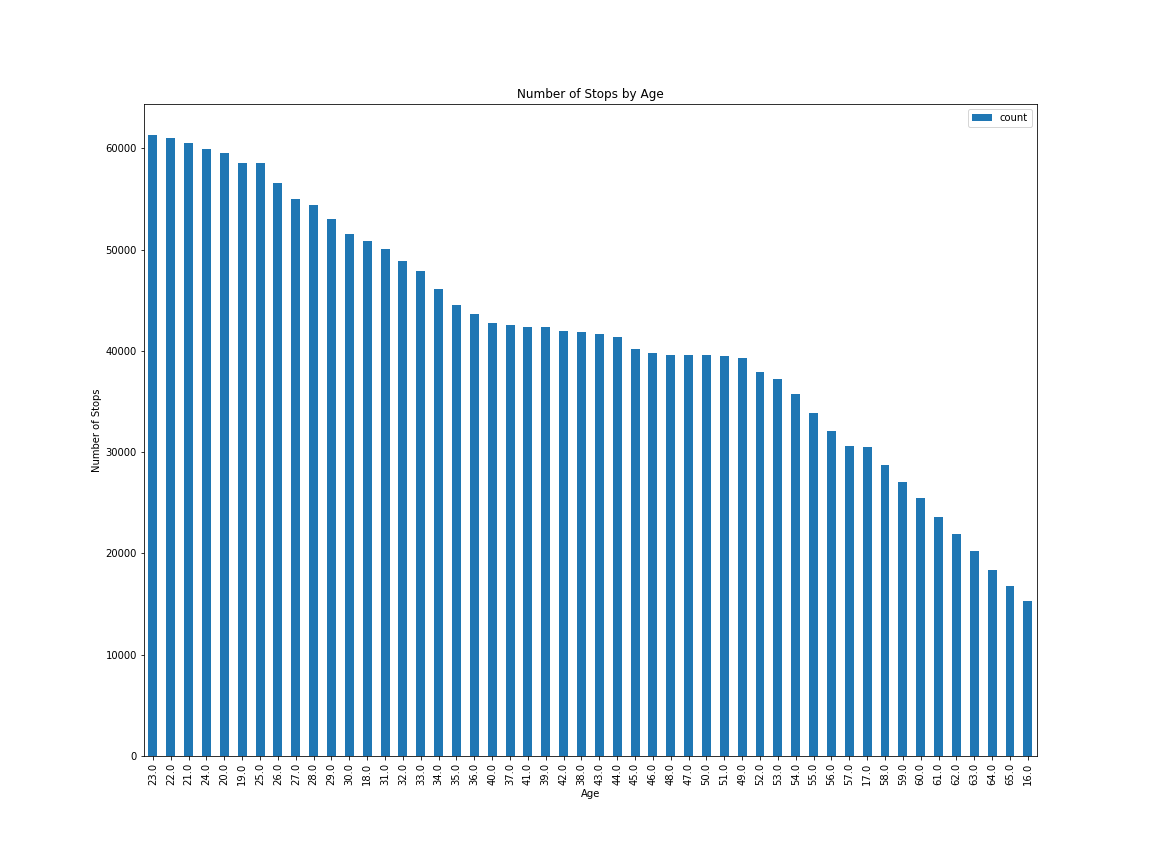


Figure 2: Traffic Stops Ranked by Age

4.1.2 Traffic Stops by Gender

Looking at traffic stops by gender was very interesting. More than twice as many males (68.9%) were pulled over than females (31.1%). With the large difference between genders, we broke it down by age to see if that trend was consistent for all age groups or not.

Figure 2 shows the number of stops of each age, with the colors distinguishing gender. There is a smaller disparity between men and women who are pulled over in the younger (18-33) age range. Men are 60-65% of the stops, while women are 35-40%. The older group split is closer to 70% men and 30% women. It would be interesting to know the division of drivers on the road to understand if that partially explains the disparity. We expected the opposite - a larger difference between the genders at a younger age and as age increased, the disparity to lessen, but that does not seem to be the case.

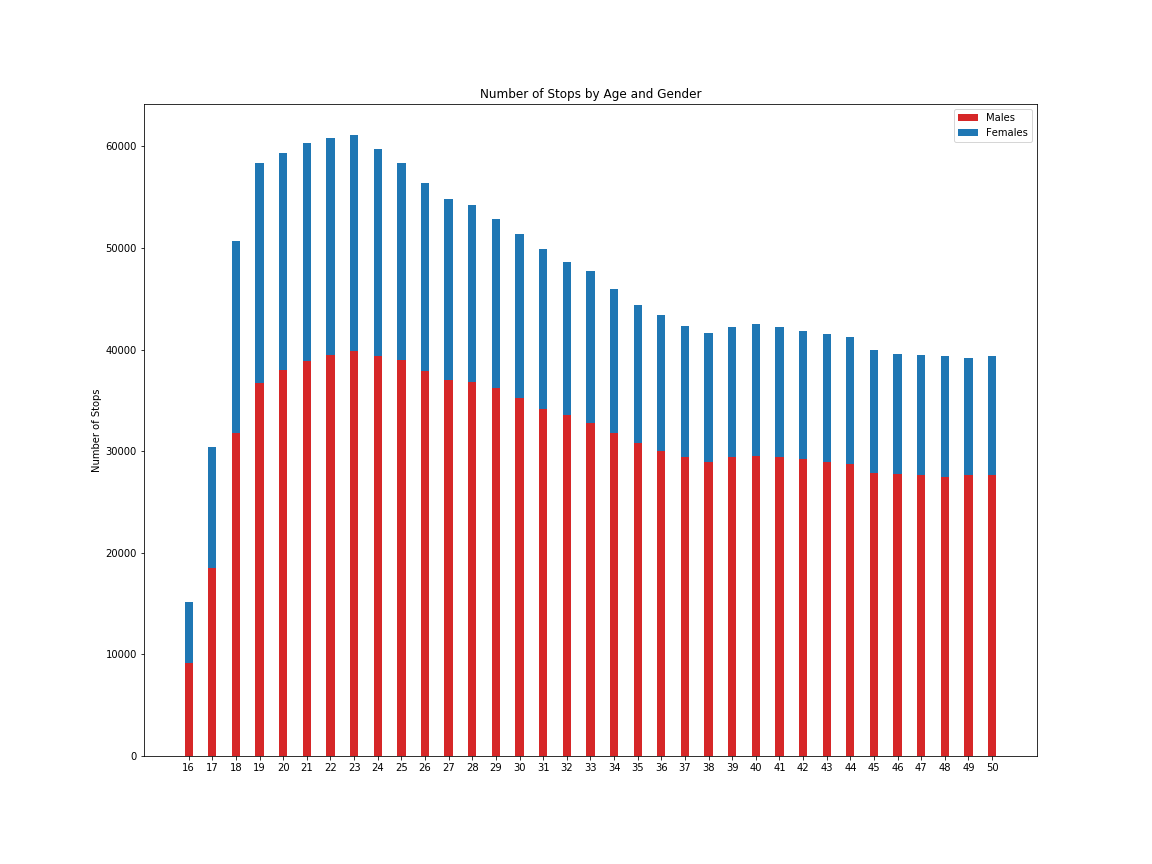


Figure 2: Traffic Stops by Ranked by Age and Distinguished by Gender

4.1.3 Traffic Stops by Race

The percentages by race for drivers pulled over looks to be fairly close to the demographics of Colorado. Figure 3 shows the breakdown from the traffic stop data.

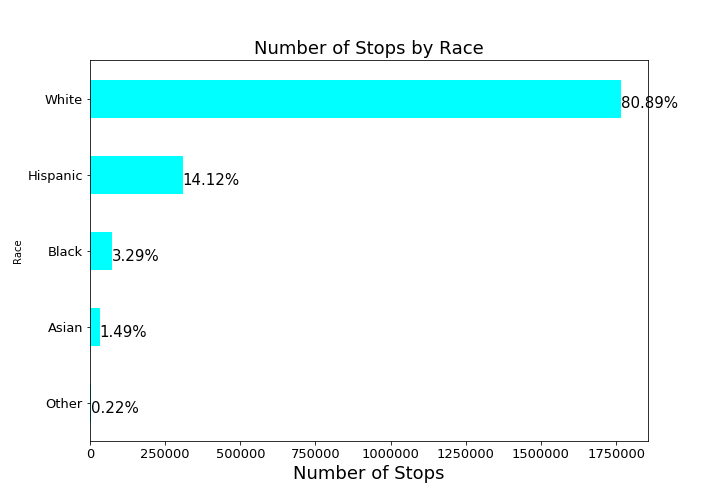


Figure 3: Traffic Stops by Race

According to the 2010 United States Census, Colorado had a population of 5,029,196. Racial composition of the state's population was:

81.3% White American (70.0% Non-Hispanic White, 11.3% Hispanic white) 20.7% Hispanic and Latino American (of any race) heritage 7.2% Some Other Race4.0% Black or African American 3.4% Multiracial American 2.8% Asian American 1.1% American Indian and Alaska Native 0.1% Native Hawaiian and Other Pacific Islander.*[[4]](#footnote-4)*

When we looked closer at the breakdown of stops by race and age together, the story changed a bit. We saw above that younger drivers were pulled over more often, and this remained true for all races. Figure 4 below shows younger Hispanic and Black drivers were pulled over at higher rates than their White, Asian and Other counterparts.

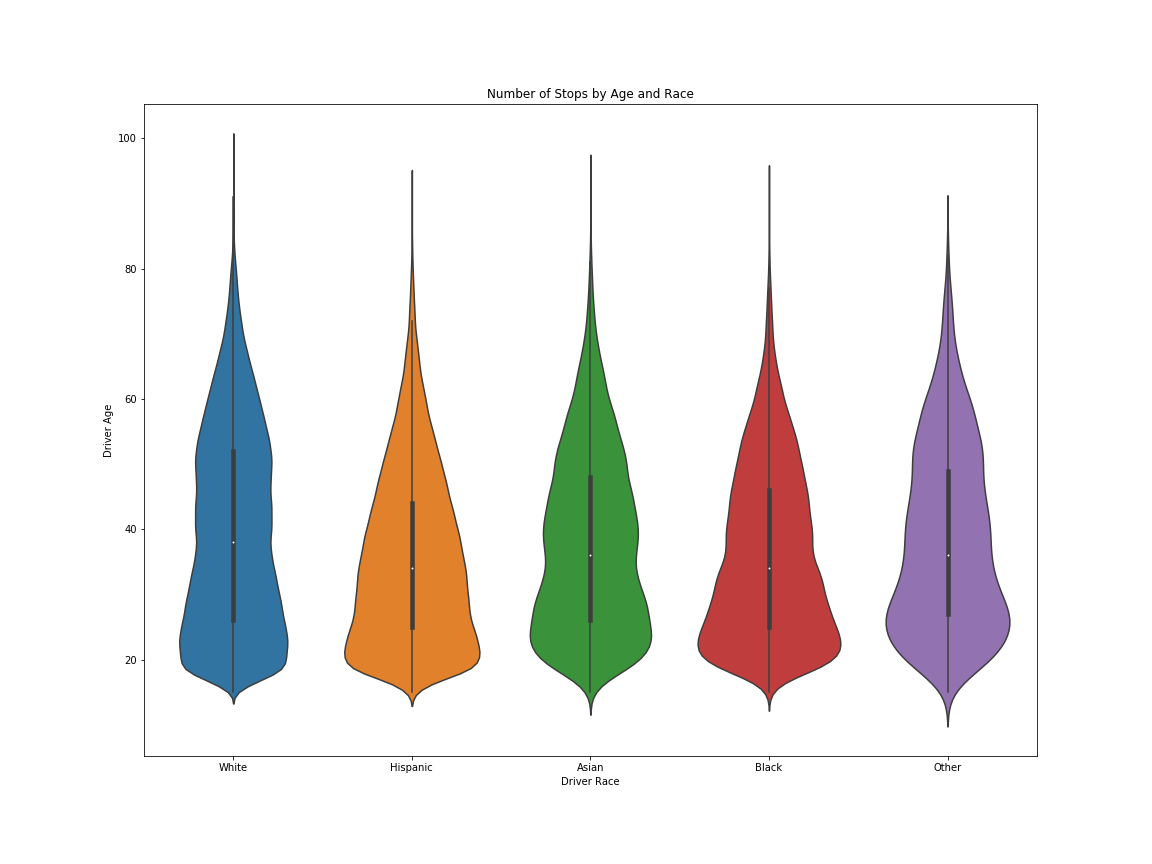


Figure 4: Number of Traffic Stops by Race and Age

4.1.4 Other Features Present for the Traffic Stop

The top 3 counties where people were pulled over made up 22% of the traffic stops. El Paso County had the highest number of stops. This county is in the mountains, on a major North/South interstate and contains Colorado Springs, the second largest city in Colorado. Jefferson County was second. The county encompasses a major entrance from the Front Range to the mountains for skiing and outdoor activity with major highways.

Speeding accounted for over 36% of the traffic stops. This does not include when speeding was combined with other violations, for example, paperwork and speeding together made up their own class of violations. The combinations of violations created almost 2000 different categories of violations, although the top 11 violations made up just over 80% of the traffic stops.

The vehicle type was mostly not tracked. Over 70 of the data had the vehicle type recorded as 'NA NA 0' or 'NA NA NA', so was not obvious when we were searching for null values above. We did not continue using vehicle type as a feature.

Out-of-State drivers accounted for 18% of the stops, which seemed high. It would be interesting to know how many of the cars on the roads were out-of-state. Officers may monitor drivers more on major highways where there are more travelers.

Looking at which police departments made the most stops, the data was a bit surprising. Golden’s police department accounted for almost 11% of the stops. Golden is in Jefferson County, which is situated in an area much of the Front Range has to pass through to get to the mountains. There are several major highways in the area. What was surprising was that Denver is the capital city of Colorado, but a very small portion of the stops (less than 1%). This brings up a very interesting question of which departments participated in sharing their data. The information from Stanford Open Policing Project states: “Denver County has many fewer stops than expected given the residential population; this is because it only contains a small section of highway which is policed by the state patrol.” We would have to do more research to understand the difference of state patrol referenced here and the distinction of police departments within the data.

**4.2 Data Gathered During/After the Traffic Stop**

4.2.1 When a Person Was Searched, Was Contraband Found

We subset the data and looked specifically at drivers who were searched. We plotted drivers who were searched by race and whether or not contraband was found. Looking at whether contraband was found may point to biases from officers. For example, Figure 5 shows Blacks and Other had a higher rate of searches conducted without finding contraband, indicating officers are targeting those groups unfairly.

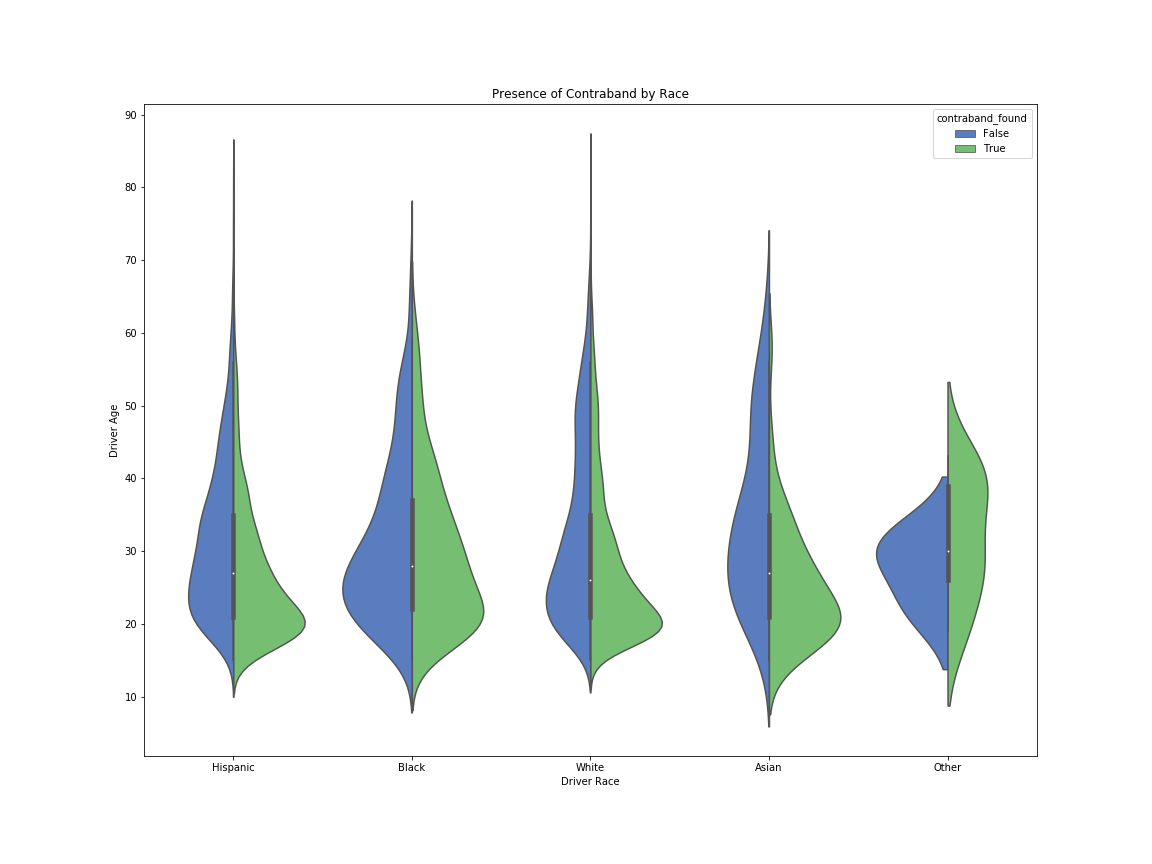


Figure 5: Whether or Not Contraband Was Found by Race

4.2.2 Number of Searches Over Time

The search rate dropped dramatically in January 2013. Recreational marijuana became legal in Colorado at that time and is most likely a significant factor in the drop. Figure 6 below shows the trend in searches between January 2010 and March 2016.

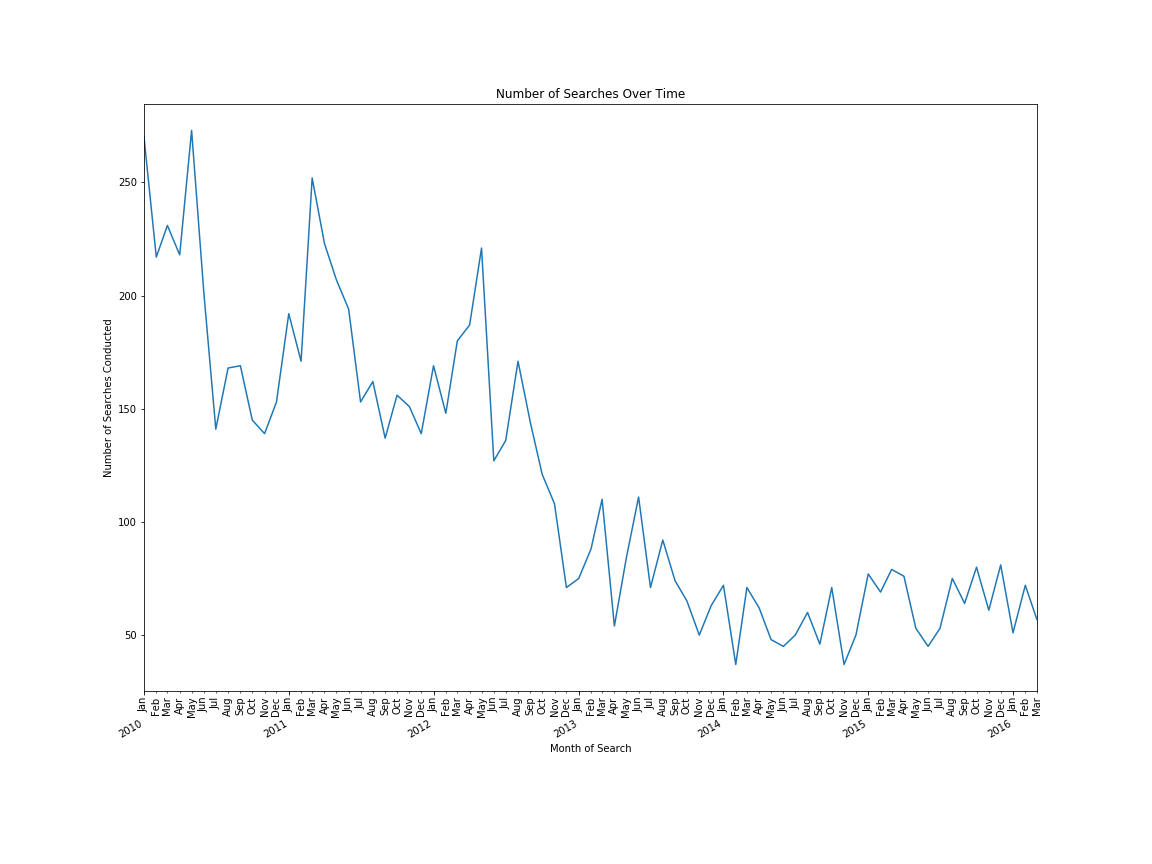


Figure 6: Number of Searches Over Time

**5 Data Analysis**

After completing the exploratory work we verified our observations through statistical inference. After that, we used a clustering algorithm to create groups of drivers who were pulled over. By aggregating all of the information, from the exploratory work, statistical inference, and machine learning, we were able to provide a more complete view of who was stopped, the stop outcome, and potential conscious and unconscious biases.

**5.1 Statistical Inference**

During the exploratory analysis, we noticed the difference in the number of males and females pulled over. We also saw younger drivers are pulled over more often than older drivers. We chose to verify whether there was a statistically significant difference in the mean ages between the genders. The mean age of men pulled over was 39.68 and the mean age for women was 37.56. We verified whether the observation was statistically significant using the z-statistic.

Generally the z-statistic is used if the sample size is over 30 and you know the population standard deviation. The t-statistic is used if you do not know the population standard deviation and the sample size is less than 30. Our sample size was over 30, so we used the z-statistic. Our threshold to reject the null hypothesis was .05.

The p-value was .075 so we did not reject the null hypothesis (that the mean age of males and females who are stopped are the same). We could not show that there is a statistically significant difference in mean ages between males and females.

**5.2 Machine Learning**

5.2.1 Prepare Data

We had a large dataset, so had to be conscious of the specific data we wanted to use for the K-Means Clustering model, but went through several iterations to get there. We first changed the null values in the remaining features to their own category of ‘unknown’ within the feature. All features were treated as categorical so we used get\_dummies() to flatten the categories so each category had its own column with a 1/0 value. For the first attempt at running the model we used the following features…

* county\_name
* police\_department
* driver\_gender
* driver\_age
* driver\_race
* violation
* search\_conducted
* contraband\_found
* is\_arrested
* out\_of\_state

The model did not finish after running over night, so we subset the data. The ‘violations’ had the most categories at 1,953. When ranking the violations, we noticed 80% of the data contained only 11 violations, so we chose to use only the top 11 violations. After this subset we had 10 features (that were then flattened out) and 1,709,208 observations.

The model with the subset data still did not run over night. We chose to subset the data again to only include features present before and up to the stop. We removed ‘search\_conducted’, ‘contraband\_found’ and ‘is\_arrested’. The focus of the model could then be clustering who was being targeted in traffic stops. We also chose to remove ‘police\_department’ because it is somewhat duplicative of County. The resulting model used 6 features and 1,709,208 observations.

5.2.2 Number of Clusters and Model

With K-Means Clustering we want to maximize the distance between the centroids of each cluster and minimize the distance between data points and their respective centroids. To achieve that goal, we wanted to make an informed decision on how many clusters to use. We chose to use the Elbow method to decide the number of clusters to use. The Elbow method plots the sum-of-squares error in each cluster against the number of clusters. The ‘elbow’ in the plot shows the optimal number of clusters to use. Figure 7 shows our elbow plot.

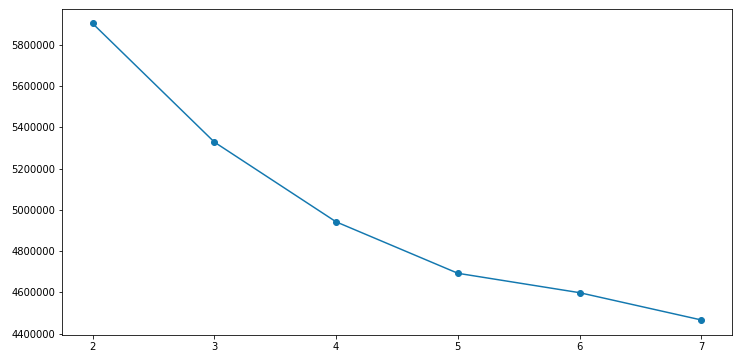


Figure 7: Elbow Plot to Choose the Optimal Number of Clusters

We chose to use five clusters. Elbow plots can be somewhat difficult unless there is a clear bend, but the difference between four and five was much larger than the difference between five and six, so we chose five clusters.

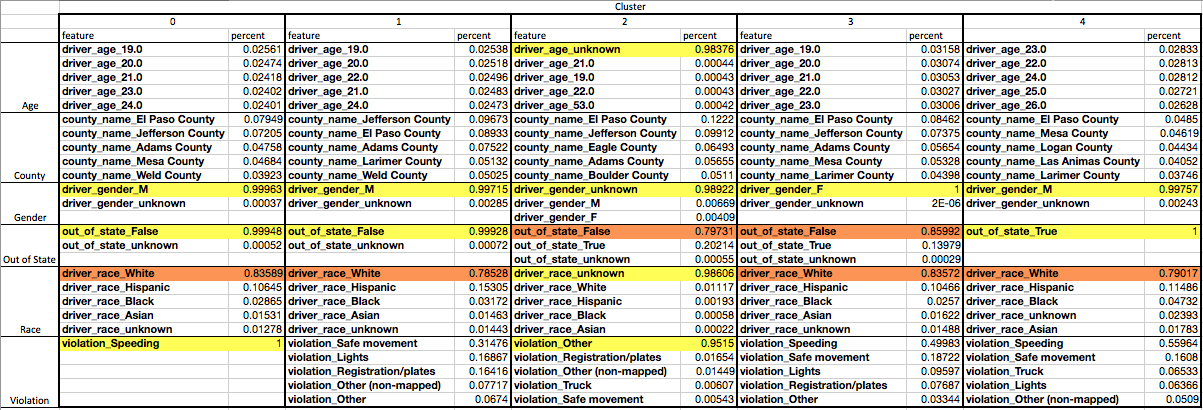
After running the K-Means clustering model with five clusters, we counted the number of observations in each cluster.

|  |  |
| --- | --- |
| Cluster | Number of Observations |
| 3 | 513,971 |
| 1 | 461,478 |
| 0 | 388,161 |
| 4 | 232,652 |
| 2 | 112,946 |

Figure 8: Clusters by Number of Observations

**6 Evaluation/Recommendations**

To evaluate the clusters, we had to collapse the “dummy” variables back down to their original features. Once the data frame was back to its original shape, we subset the data by cluster and analyzed each cluster. Figure 9 shows the top five results for each feature by cluster. The yellow highlighted descriptions show when the result was 95% of the total for that feature for the cluster, meaning a very strong descriptor for that cluster. The orange highlighted descriptions are the descriptions that made up 75-94% of the feature, a strong descriptor for the cluster.



To further summarize the clusters, gender, in/out-of-state status and violation seemed to be the strongest features for deciding clusters. This made sense given that we saw in the exploratory analysis younger drivers were stopped more often, and Whites made up a majority of the stops, so those features were present across all clusters.

The clusters seemed to be less helpful than the exploratory analysis in telling the story of who was stopped and why. The more detailed analysis was able to look at the interactions between features, for example, when race was broken down by age, we saw younger Black and Hispanic drivers were pulled over at a higher rate.

Our recommendation would be that Colorado police departments invest in a structured data collection and analysis. There are questions about who participated in supplying the current data, and there are issues with completeness of the data (null or NA values). We do not have confidence that this data set provides a full picture of who is being stopped in Colorado. Further results could also be compared to census data to understand whether the results are proportional to the state population.

**7 Next Steps/Improvements**

For specific improvements to this analysis, I would…

1. Try a different method to choose the number of clusters in the model, such as the silhouette score. The Elbow method can be difficult to interpret so another method of picking the number of clusters may be helpful.
2. We would also try another method for clustering, such as DBSCAN. DBSCAN focuses on dense regions of the data and may find more insightful patterns in the data.

1. https://openpolicing.stanford.edu/findings/ [↑](#footnote-ref-1)
2. “The Stanford Open Policing Project — a unique partnership between the Stanford Computational Journalism Lab and the Stanford School of Engineering… Starting in 2015, the Open Policing Project began requesting… data from state after state. To date, the project has collected and standardized more than 100 million records of traffic stop and search data from 31 states.” https://openpolicing.stanford.edu/findings/ [↑](#footnote-ref-2)
3. [https://github.com/5harad/openpolicing/blob/master/DATA-README.md](https://github.com/5harad/openpolicing/blob/master/DATA-README.md" \t "_blank) [↑](#footnote-ref-3)
4. *[https://en.wikipedia.org/wiki/Colorado#Demographics](https://en.wikipedia.org/wiki/Colorado" \l "Demographics" \t "_blank)* [↑](#footnote-ref-4)