Deaths in Police Custody and Officer Involved Shootings in Texas

Capstone 2 Project for Springboard Data Science Career Track

The Problem

From 2005-2016 more than 7700 people died while in police custody in Texas. From 2010-2016 there were 640 reported incidents of officer involved shootings by major police departments in Texas, over 200 of which were fatal.

How could these deaths have been prevented? What factors contributed to the fatalities? Could discovering these indicators help improve officer training and reduce deaths in custody or during an officer encounter?

Potential Clients

Texas Justice Initiative, ACLU of Texas, other non-profit or civil rights organizations, the police departments in question.

All of these organizations may find the analysis and models useful in predicting future deaths and preventing them. The non-profit activist groups could use the report as a jumping off point for further investigation to expose biases within the police departments or to report on overall trends in policing.

The individual police departments or even the overarching Texas government could use the information to improve officer training in order to reduce deaths and improve outcomes.

Data sets

Two data sets were used in this project. The first is the Custodial Deaths reported to the Attorney General of Texas. Police departments are required by law to report deaths in custody. The data was obtained by the Texas Justice Initiative and published on their website (http://texasjusticeinitiative.org/), downloadable as a CSV file. The second dataset was downloaded as a CSV file from VICE news, who collected the information by contacting the 50 largest police departments in the country. (https://news.vice.com/en_us/article/a3jjpa/nonfatal-police-shootings-data)

Other potential data sets

The FBI has the Uniform Crime Report could be used to obtain statistics on the greater population in custody in Texas, since general population demographics are not a particularly comparable set for many reasons. Data on all people in custody could possibly be obtained from the individual departments in question, but this would likely be an arduous, time-consuming process.

Initial Data Wrangling

A “state” column was manually created in the Officer Involved Shootings CVS, aka “shoot”, by entering the state abbreviation for each unique city, then using pandas’ forward fill method to create a “state” label for each instance. The “shoot” dataframe was then restricted to only rows with “TX” in the “state” column.

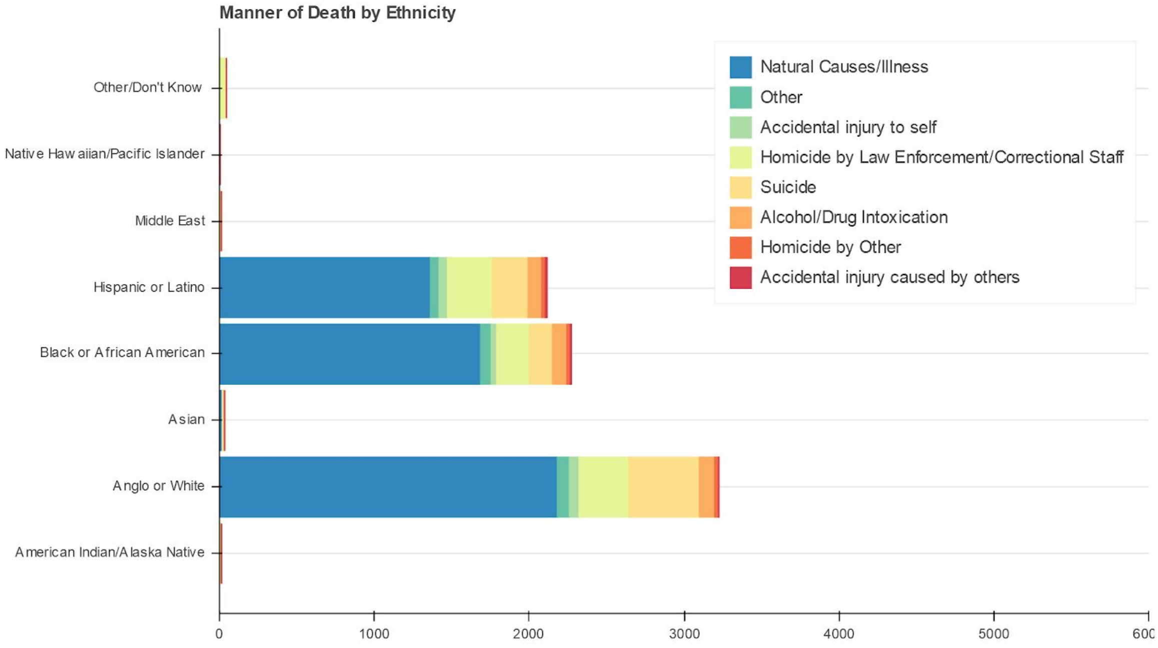
In the Custodial Death dataframe, aka “cust”, one report (John Daniel Hanson-Dallas) was deleted because it was not in the correct format. The report was spread out across 180 rows and 3 columns, as opposed to being in one row, broken up into uniform columns and the summary written under one column.

A “year” and “month” column were created in both the “cust” and “shoot” dataframes with info extracted from their “date” columns.

Exploratory Findings

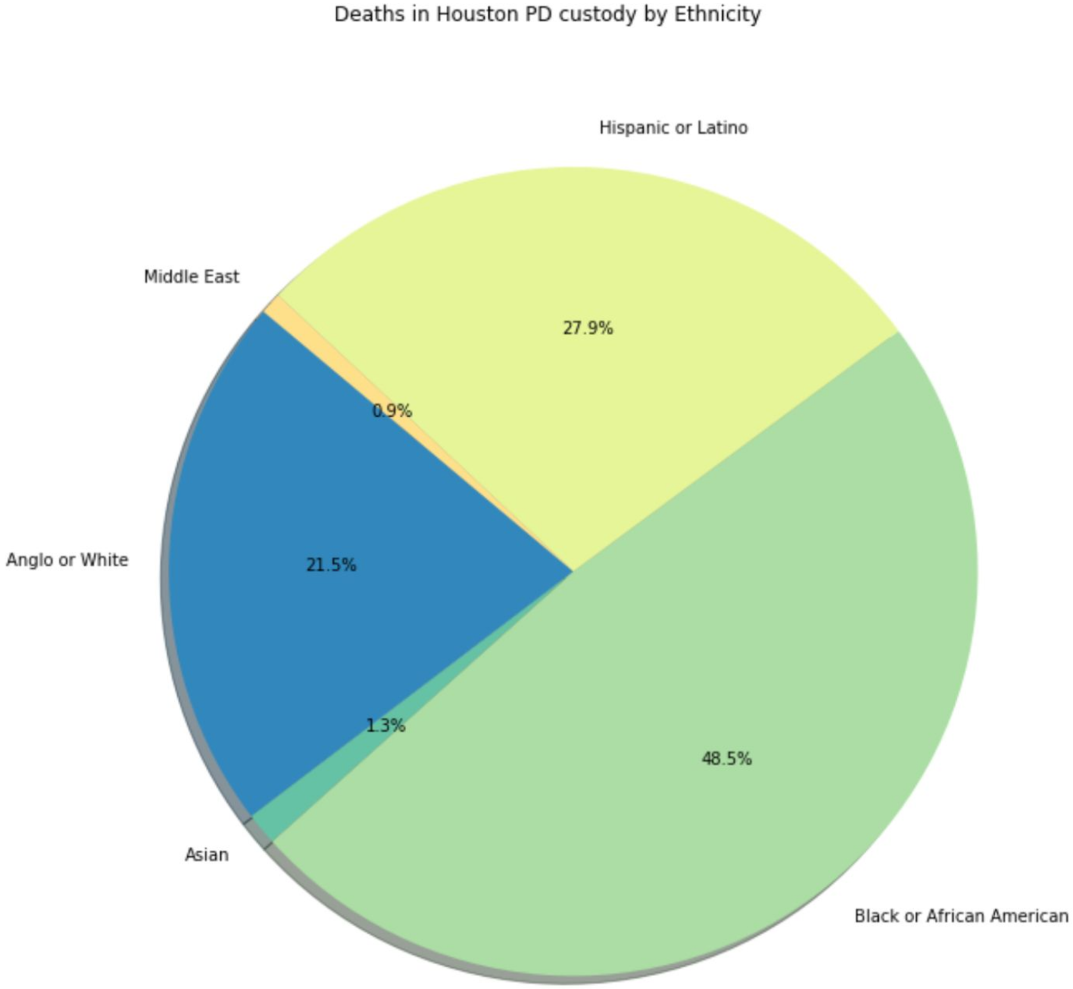
First, the Deaths in Custody were divided by the Manner of Death and plotted in a bar graph, showing that “Natural Causes/Illness” is by far the highest cause of death. Surprisingly, the next highest Manner of Death is “Homicide by Law Enforcement/ Correctional Staff”, though still less than ⅕ of the total of deaths by “Natural Causes/Illness”. Some of the deaths in this category were armed suspects killed while officers attempted to apprehend them. Further analysis of these deaths could show if any particular department has a higher rate of homicide.

A more complex stacked bar graph was created to plot Deaths in Custody, broken down by Ethnicity and Manner of Death:



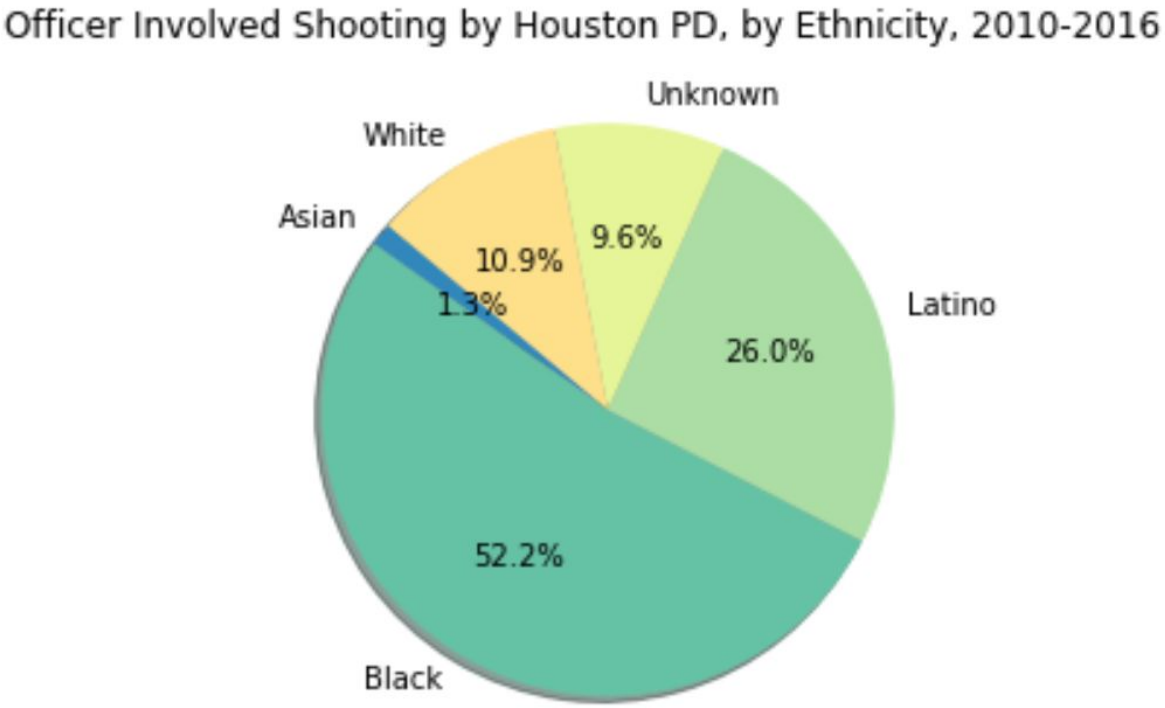
Despite the fact that more Anglo/White people died in custody, about the same number of Hispanic/Latino people died by “Homicide by Law Enforcement/Correctional Staff” as Anglo/White people did.

When looking at the racial breakdown of Houston Police Department’s deaths in custody, it was surprising to find that nearly 50% are Black/African American, when only about 25% of the city population is. However, since the federal prison population is about 50% Black, it’s unclear if HPD’s percentages show bias, especially considering that the Black population nationally is closer to 13%. More analysis would be needed, with statistics on HPD’s total custody population to compare to:



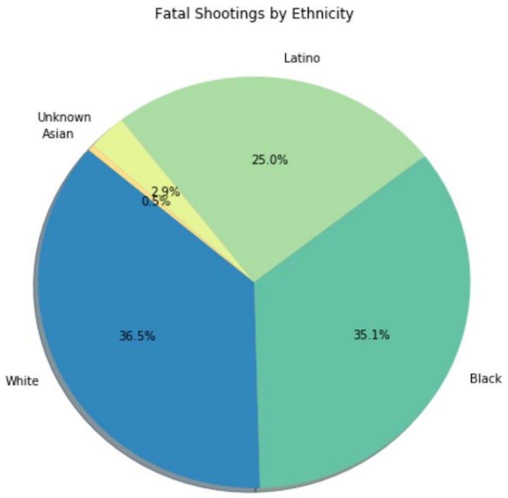
The same graphs were generated for a few different departments, but they all face the same issue of not having non-fatal custody population statistics to compare to.

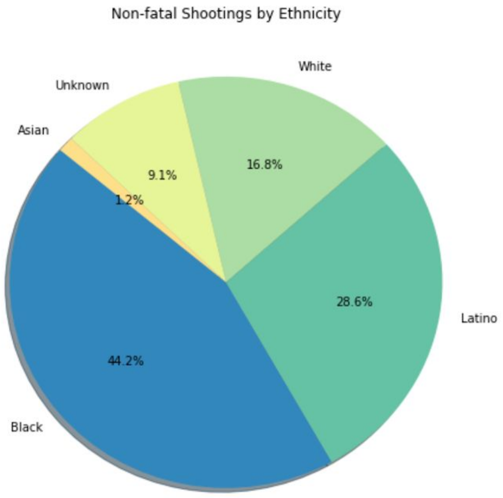
Analysis of Officer Involved Shootings offers slightly more insight, since officers would be expected to be interacting with the general population. However, looking again at the racial breakdown of subjects shot at by officers in Houston PD, shows a similar racial breakdown:



Again, around 50% of subjects in shootings were Black, despite that the city's Black population is only around 25% Black. Officers are either engaging with and being called on Black subjects significantly more, officers are more likely to shoot at Black subjects, or Black subjects are more likely to be armed. It is unclear without data on all Officer interactions over this same time period.

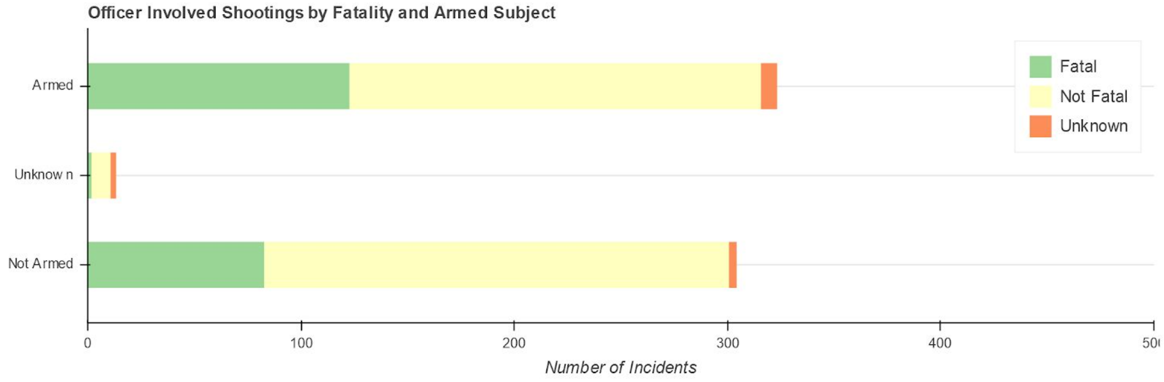
Pie charts were also created of the racial breakdown of Fatal vs. Non-fatal shootings:





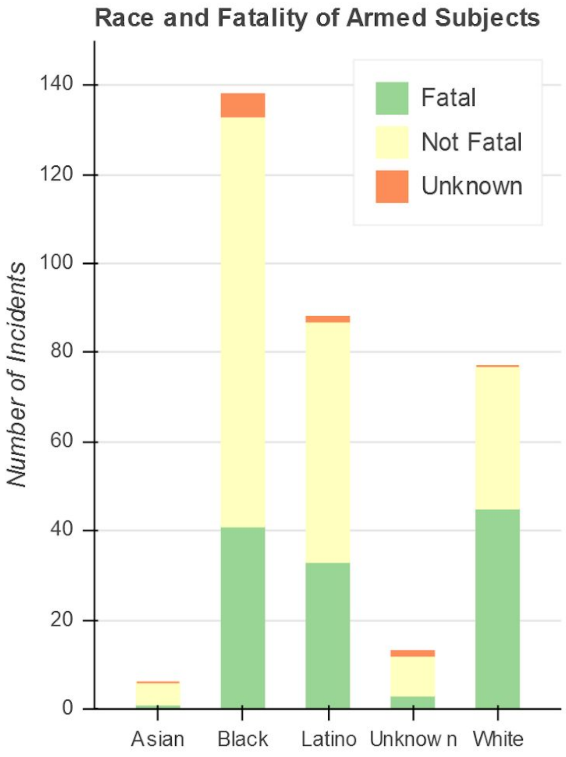
It will be more insightful however, to look at the racial breakdown of fatalities vs the total of each race, to see if any particular race is more likely to experience a fatality once involved in a shooting with officers. Breaking this down by department as well could help identify any potential biases, vs. racial reflections of the city's population.

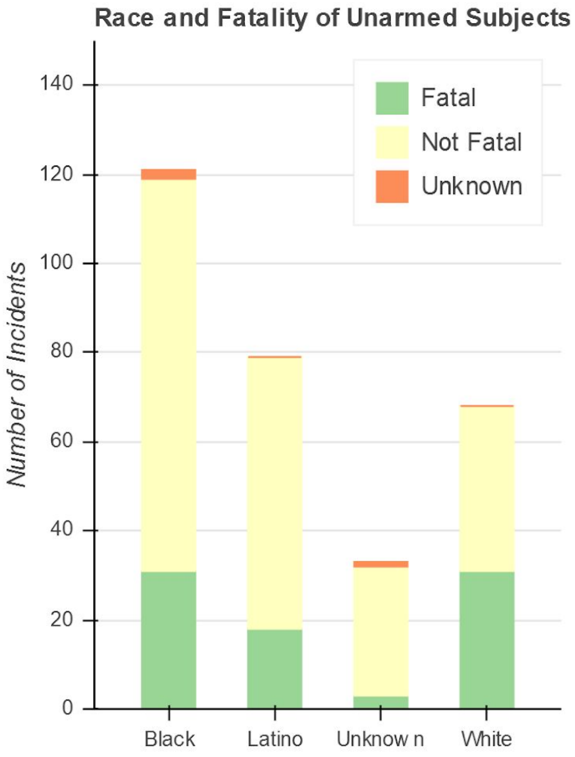
Officer Involved shootings were also broken down by whether the subject was armed and where the incident was fatal:


Officer Involved Shootings by Fatality and Armed Subject

It is surprising to see that of the total shootings, nearly half were unarmed. Within this, fewer of the unarmed shootings were fatal, though about ¼ still were fatal.

The same data was divided up again by race of the subject:


Race and Fatality of Armed Subjects


Race and Fatality of Unarmed Subjects

In both “Armed” and “Unarmed” groups, despite that the largest group involved in shootings are Black, the largest fatal group is White subjects. It is unclear what is causing this phenomenon.

Wrangling for Modeling

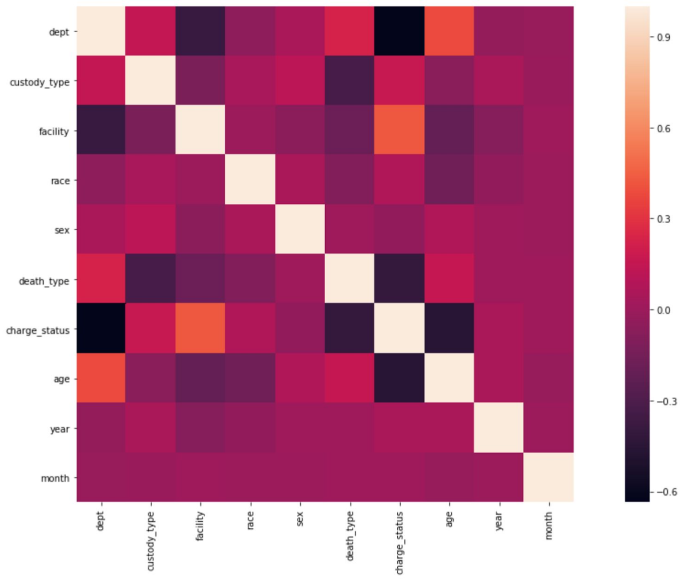
New data frames were created from both the “cust” and “shoot” data frames to give numeric values to all categorical features to prepare for Machine Learning.

Scikit-Learn’s “LabelEncoder” was used to encode the columns in the new dataframes with numbers representing the various str values in the original categorical columns. Columns that were already numeric were simply copied to the new dataframes.

During this process, a few inappropriate values were found in the number columns “NumberOfOfficers” and “SubjectAge” in the “shoot df. One of the values for Number of Officers was listed as “2 or More”, which was replaced with simply “2”. For the Subject Age, a couple entries were “Juvenile”, which were replaced with “15” as a median age guess for a subject under 18 years old. Many entries listed age as “U” which were convert to NaN values instead. After converting all the dtypes in the “newshoot” dataframe to be numeric (all columns in the “newcust” dataframe were already numeric), all the NaN values in the age column were filled with the mean age of the data in the column. NaN values in the Number of Officers column were filled with “1”, which was the integer value of the mean of the data in that column.

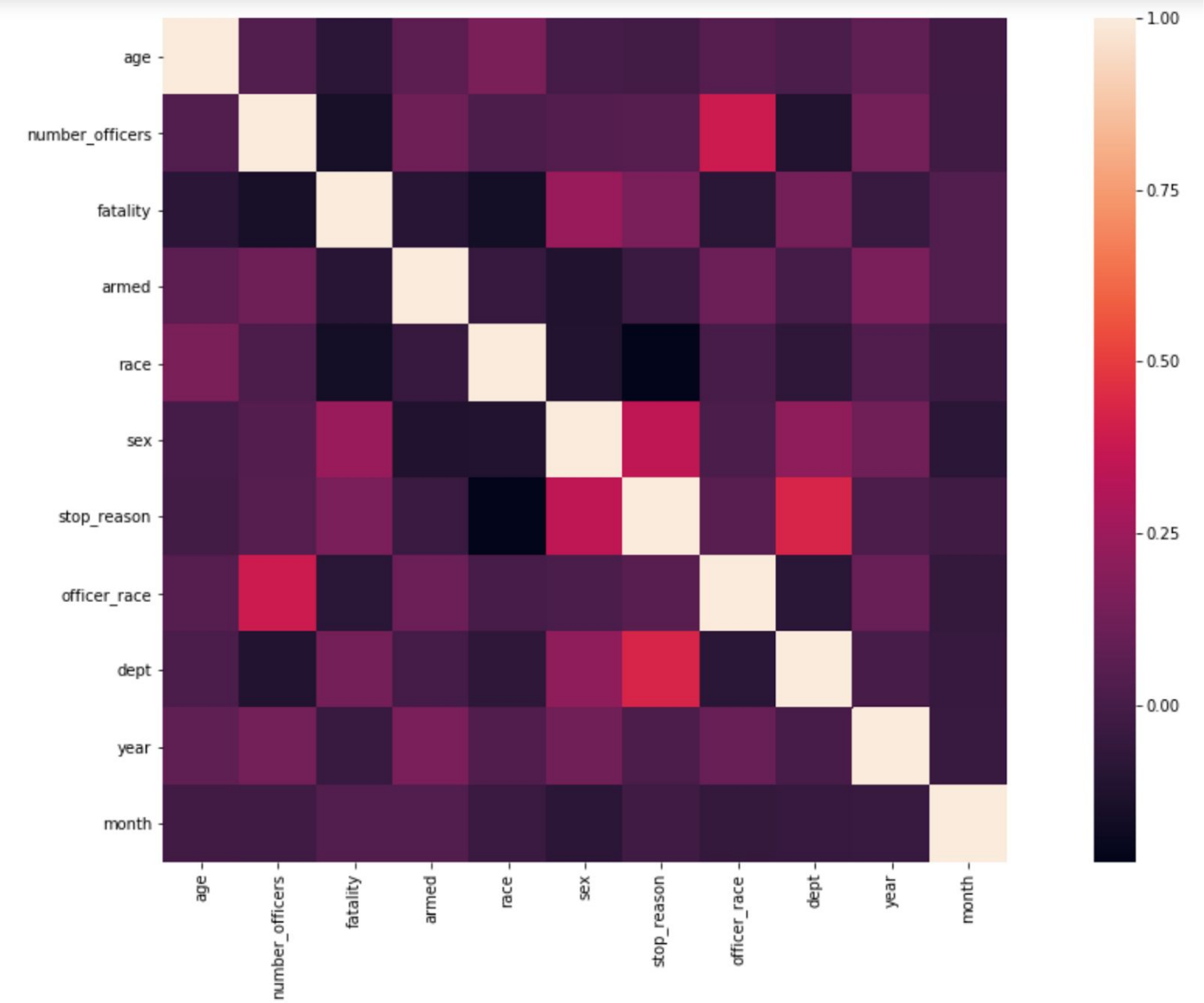
Heatmaps were created for both datasets to look at which of the features tend to correlate with each other. Because the features are categorical (except for age) whether correlations are positive or negative is not really significant since the numbers assigned to each value are arbitrary.

<u>Deaths in Custody:</u>



This visualization shows “Facility” and “Charge Status” as the highest positively correlated features, which makes sense because someone’s charge status will likely determine which facility they are placed in. Another moderate correlation is “Age” and “Department”. This correlation was unexpected as it is unclear why there would be any correlation between those two factors. “Facility” and “Age” as well as “Facility” and “Manner of Death” also appear to be slightly negatively correlated. “Facility” and “Department” are even more strongly correlated, since they are similar features. “Charge_status_ and “Department” appear to be the most correlated, but it’s unclear why. “Age” and “Charge_status” are also heavily correlated, which may have to do with the time it take to move through different stages of charges, meaning certainly levels of status would be held by older people.

Officer Involved Shootings:



The only features that seem to correlate much here are “stop_reason” and “department” but it’s still fairly weak, slightly under .5 and seems like it might have happened because all entries for Houston PD are NaN.

Feature Engineering

Clustering, an unsupervised machine learning method, was chosen to analyze the wrangled data. The benefit to this method is that it can often tease out relationships that are not immediately obvious. However, if irrelevant variables are included in the modeling, it can lead to overfitting. In order to keep extraneous information from influencing the clustering, multiple features were dropped or modified.

Deaths in Custody

"Facility" was dropped because the location of death doesn't seem particularly relevant to other features of interest. The values for "Facility" were also quite diverse and seemed as though the would not converge to show any particular relationship.

"Charges Status" was dropped because it appeared to be less significant than other features, and of the 4 categories in this feature, nearly 70% was in a single category, meaning that the variance was not particularly large.

"Year" and "Month" likewise did not seem as though they would have a significant effect on the sought after relationships.

"Department" was removed because there was concern that the clustering would weight it too much, and just end up clustering by department, or size of department.

The value counts for "Sex" were 7313 men and 416 women. Since this variable has so little variance, it was also dropped.

Officer Involved Shootings

The "Number of Officers" variable was replaced with a variable that instead indicates whether each incident involved only one officer, or more than one. This allowed the feature to be binary (instead of categorical) and eliminated unnecessary variance between incidents with multiple officers.

"Nature of Stop/Reason" variable was removed because there are so many different values that are largely inconsistent and would need significant wrangling to group values that are the same/similar in a manner that would be useful.

Value counts for "Sex" are 465 men, 11 women and 164 NaN/unknown. Because there is a significant amount of missing values, and the data that is not missing has so little variance, "Sex" was dropped.

"Officer Race" was dropped because there are so many different values that are inconsistent, the variance in the values might have caused overfitting from all the different inputs, when really the most relevant part of the variable is whether all the officers were white or not.

Again, "Year" and "Month" likewise did not seem as though they would have a significant effect on the sought after relationships.

"Department" was dropped from this dataset also because of concerns the clustering would just group to the four departments included.

Clustering - Model Building

Multiple algorithms exist for Clustering data. This projected looked at several algorithms to see if any of them might perform better than others. The algorithms utilized in this project were KMeans, Spectral Clustering, Agglomerative Clustering, and DBSCAN. Both datasets were scaled before modeling. All models were evaluated with a silhouette score, which measures how similar a data point is to its own cluster compared to other clusters.

Deaths in Custody

Kmeans was run on the Custody dataset, with the number of clusters set to a range from 2-10. The best performing model of these was n_clusters = 7, with a silhouette coefficient of 0.44

Spectral Clustering was tuned to this dataset using a range of 2-10 for n_clusters, and a set for gamma of [0.001, 0.01, 0.1, 0.25, 0.5, 1.0]. Gamma is the threshold of sensitivity to differences in features, when using the RBF kernel, which is the default setting for Spectral Clustering. The best performing model had 8 clusters and a gamma value of 0.01, resulting in a silhouette coefficient of 0.45

Agglomerative Clustering was fit to the data with an n_cluster range of 2-10, and the best performing model had 4 clusters with silhouette coefficient of 0.40

DBScan was tuned with a set of values for epsilon of [0.25, 0.5, 0.75, 1, 1.5] and "min_samples " set of [3, 4, 5, 6]. Epsilon is the maximum distance measure used to consider whether a data point is noise or part of a dense region. "Min_samples" is the minimum number of samples required to create a dense region that is the center of a cluster. In this particular case, the model found to have the highest score was had an epsilon of 1.5 with min_samples of 6, resulting in a silhouette coefficient of 0.33. Above these hyperparameters, the DBScan algorithm only grouped the data points into one cluster.

Officer Involved Shootings

Again, Kmeans was tuned with an n_cluster range of 2-10, resulting in a best fitting model with 9 clusters and silhouette coefficient of 0.38

Spectral Clustering was tuned with n_clusters of 2-20 and gamma set to [0.1, 0.5, 1.0, 1.5]. Although the best performing model scored .44 with 19 clusters, there was concern that the large number of clusters scored high because of overfitting. The best model was instead determined to be 11 clusters with a gamma of 0.1, because its silhouette coefficient was 0.40 but the subsequent few models had slightly lower silhouette scores, before starting to rise again.

Agglomerative Clustering was tuned to this model with an n_cluster range of 2-10. Though the best score of 0.32 was slightly higher with a 10 clusters model, the best model was determined to be 7 clusters with a silhouette coefficient of 0.31, because it was less likely to be overfitting.

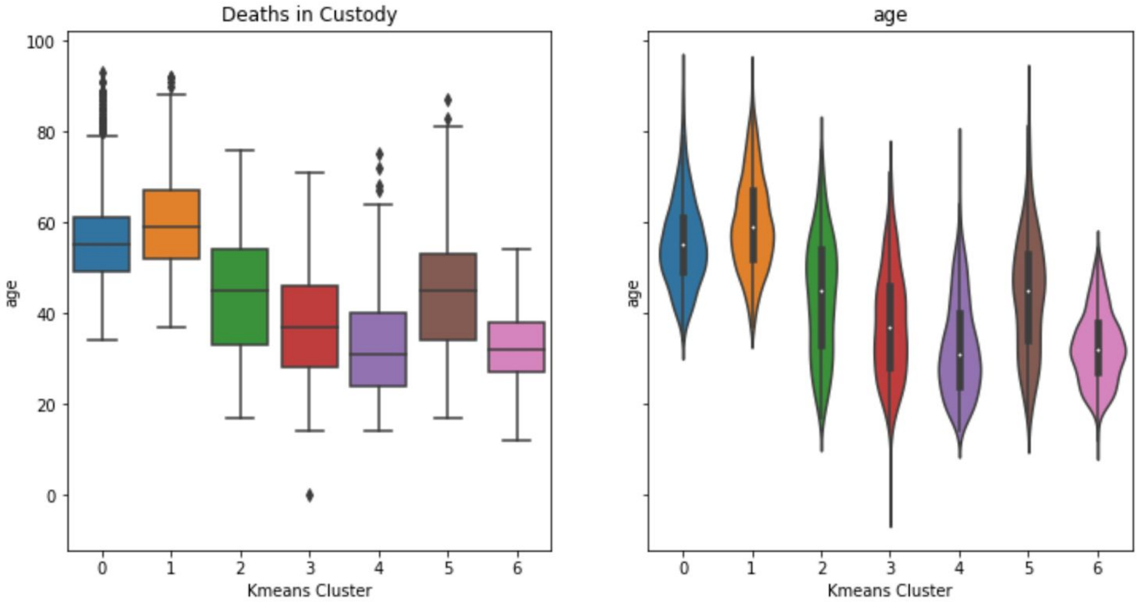
DBScan was fit to the Shooting data with epsilon values of [0.1, 0.25, 0.5, 0.75, 1, 2] (at 3 the algorithm grouped all data into one cluster) and "min_samples" = [1, 2, 3, 4, 5]. The best scoring model had an epsilon of 0.25 and min_samples of 1, which resulted in a Silhouette Coefficient of 0.598. This was clearly overfit however, as there were many clusters with only one data point and a total of 156 clusters.

In both the "Custody" and "Shooting" data sets Spectral clustering performed only slightly better than Kmeans, but had 1-2 more groups, which suggests it performs only slightly better because of overfitting. Due to this, the Kmeans models were chosen to be the clustering labels for final analysis.

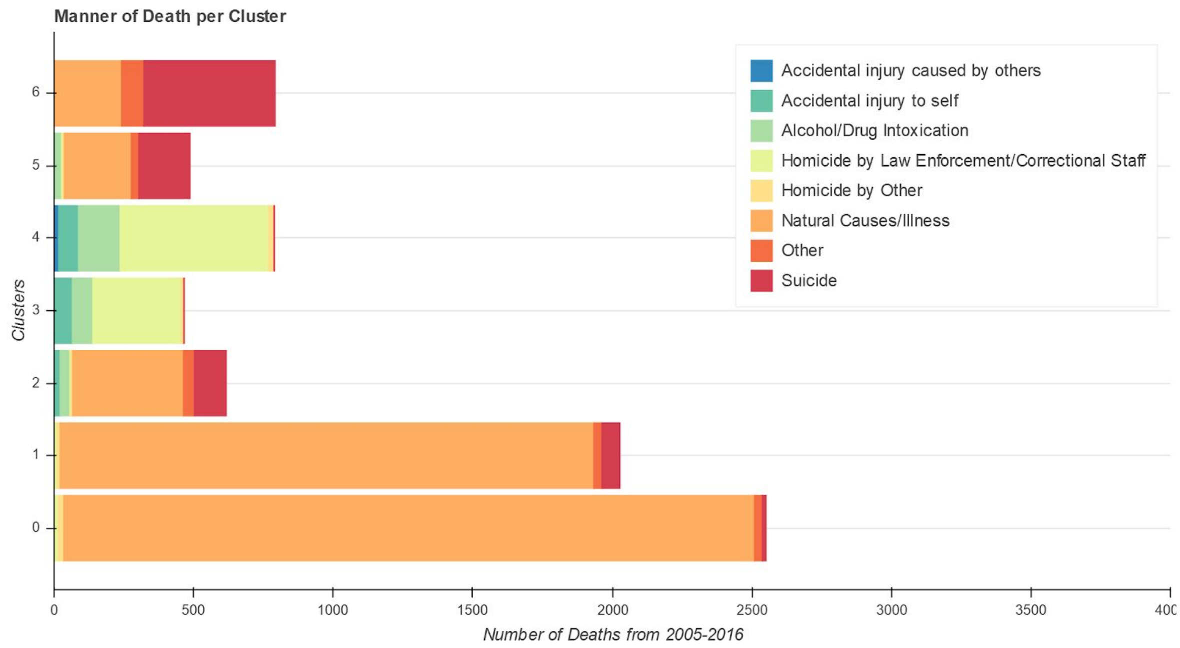
Cluster Visualization

In order to understand the differences found between the clusters into which the data were grouped, the Kmeans Cluster labels were added to the dataframes. Then box and violin plots were created to look at the distribution in each cluster of each feature. It quickly became evident that this visualization method only worked well for continuous variables, which in each data set was only "Age". For the other categorical variables stacked bar graphs were created.

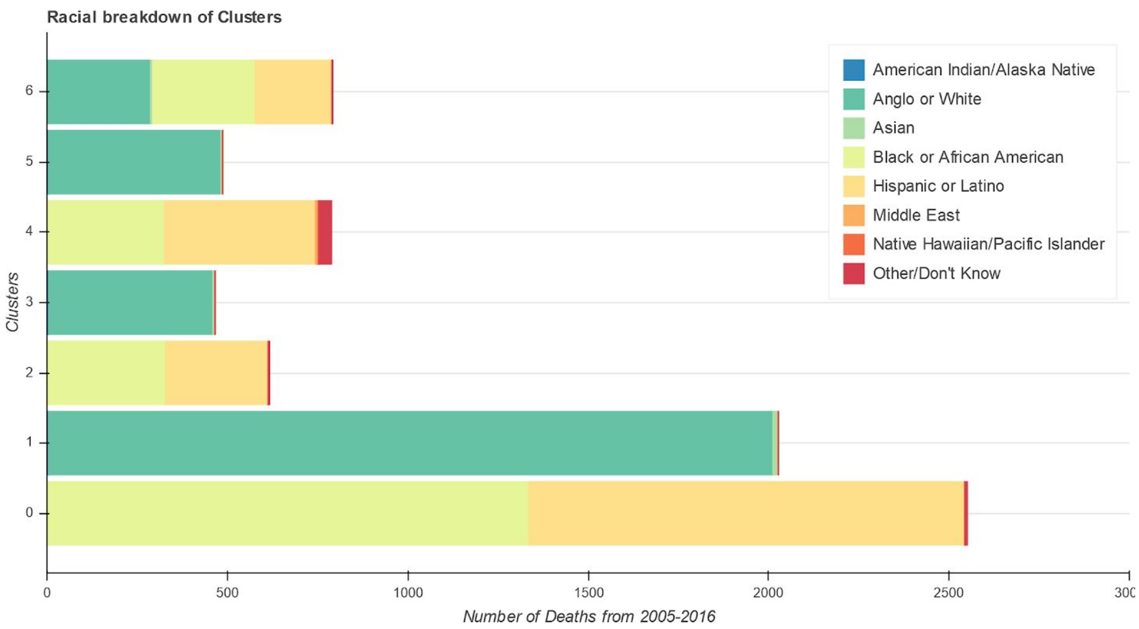
<u>Deaths in Custody</u>



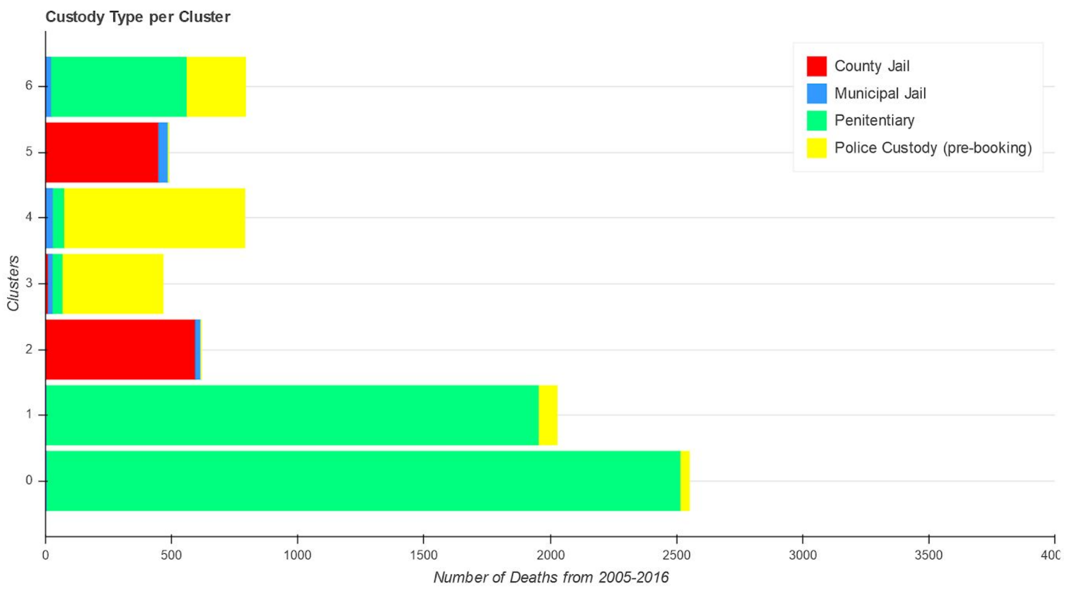
Distribution of "Age" appears to be relatively even across clusters. The cluster with the highest mean age is "1", though "0" has a very similar distribution and outliers. Clusters "3" and “5” have the widest ranges, thought “3” has a much lower mean than “5” and includes the youngest death in the data set as an outlier. Cluster “4” has the lowest mean, though cluster “6” has the youngest concentration/range.



Kmeans Cluster groups “0” and “1” appear to be made up mostly of incidents were “Natural Causes” were the Manner of Death. Cluster “2” is mostly “Natural Causes” as well, which another large portion made up of “Suicide” deaths. The majority of incidents in Cluster “3” have as their Manner of Death “Homicide by Law Enforcement/Correctional Staff”, with two smaller portions of “Accidental Injury to Self” and “Alcohol/Drug Intoxication”. Cluster “4” has a very similar makeup as “3”, but is a larger group, so each Manner of Death is also more numerous. Cluster “5” has “Natural Causes” as about half its incidents with “Suicide” making up most of the other half. Cluster 6 is close to two-third “Suicide” incidents, and “Natural Causes” and “Other” making up the last third.

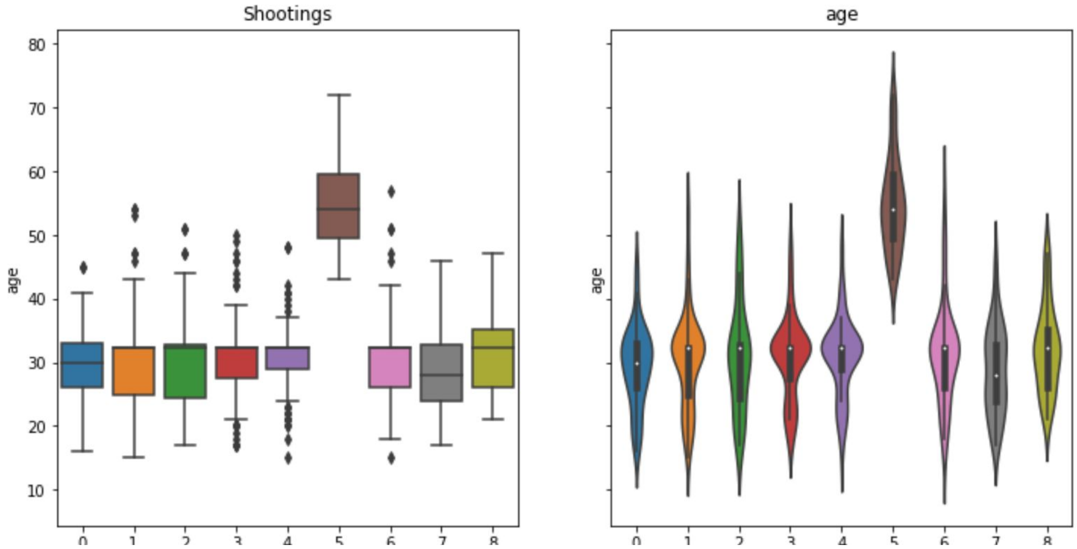


Cluster “0” appears to be slightly over half “Black or African American”, with “Hispanic or Latino” making up most of the rest of the groups. Cluster “1” is almost entirely “Anglo or White”. Cluster “2” has almost the exactly same makeup, parentage wise, as “0”, but is a much smaller group. Similarly, Cluster “3” seems to have the exact same percentages as “1”, but is much smaller. Cluster “4” is roughly half “Hispanic or Latino”, with “Black or African American” a little under half, and “Other/Don’t Know” as a small but significant portion. Cluster “5” appears to have the same percentages as Cluster “1” and “3”, with only a little higher total number of incidents than 3. Cluster “6” is broken up almost evenly into thirds with “Anglo or White”, “Black or African American” and “Hispanic or Latino”, with “Hispanic and Latino” being the smaller of the three groups.

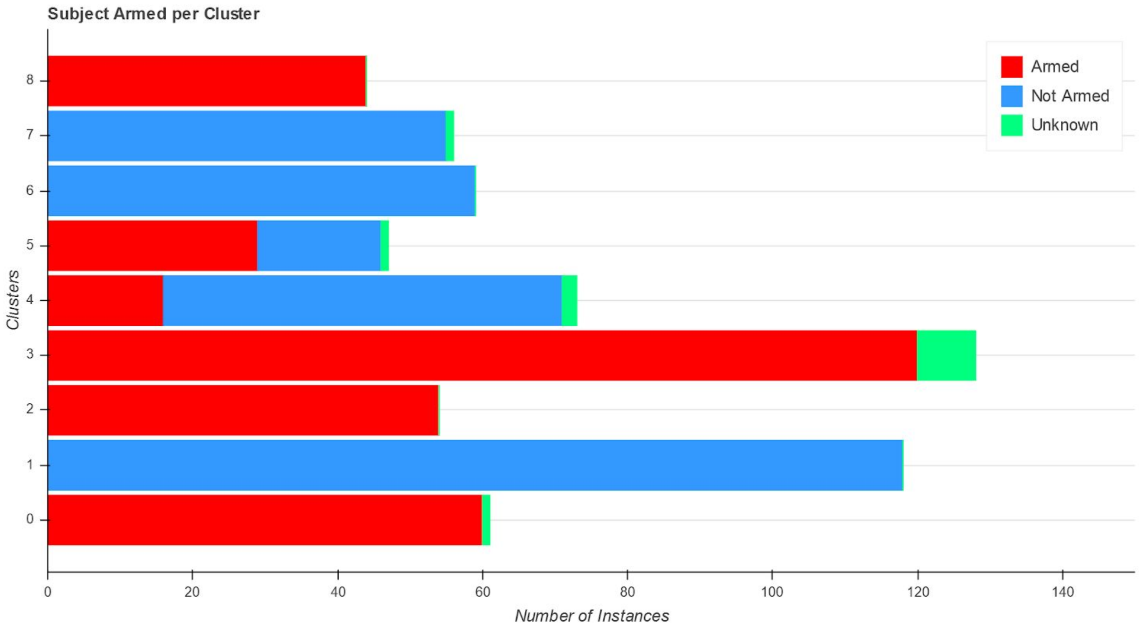


Cluster “0” is almost entirely Custody Type “Penitentiary”, with a very small portion “Police Custody (pre-booking)”. Cluster “1” has a very similar makeup to “0”, with less incidents overall, and a slightly larger portion of “Police Custody (pre-booking)”. Cluster “2” is almost entirely “County Jail” with a very small portion of “Municipal Jail”. Cluster “3” is mostly “Police Custody (pre-booking), with small numbers from each other the option options. Cluster “4” is also mostly “Police Custody” with small portions of ‘ Municipal Jail” and “Penitentiary”. Cluster “5” is mostly made up of “County Jaul” incidents, with a small portion from “Municipal Jail” as well. Cluster “6” is made up of a majority “Penitentiary” as well as a smaller, but significant portion “Police Custody” and much smaller portion of “Municipal Jail”.

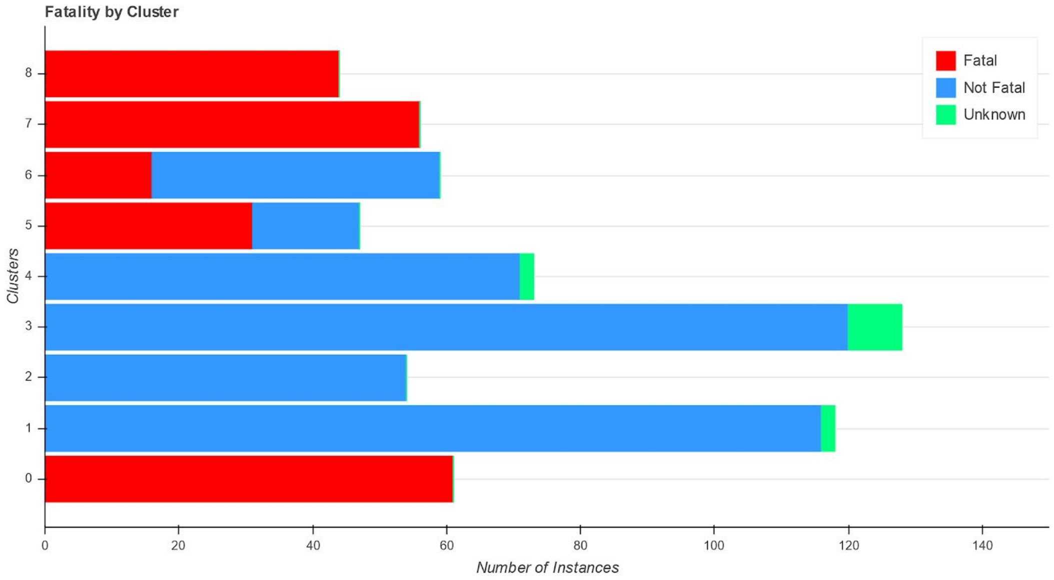
Officer Involved Shootings



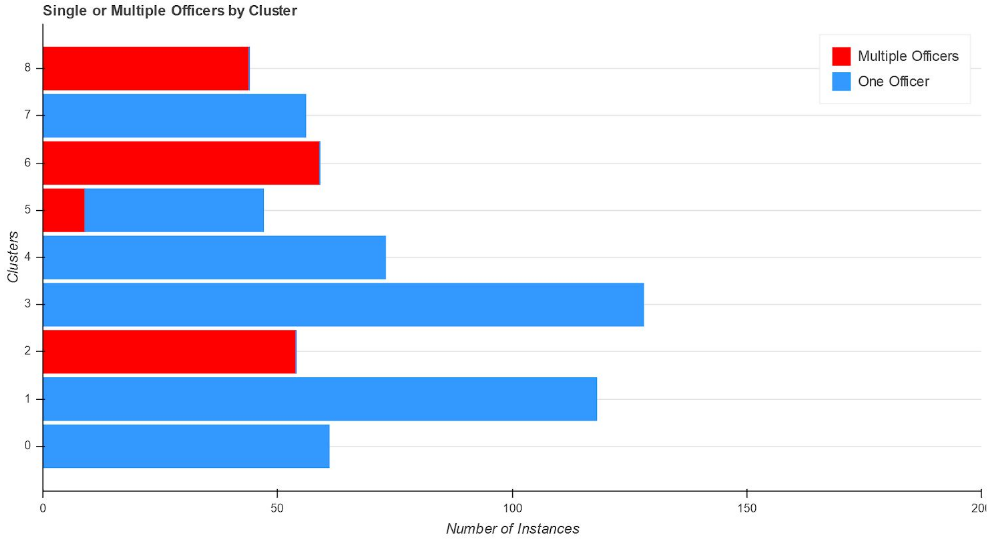
All clusters have roughly the same mean and distribution of "Age" with one exception. Cluster "5" appears to have all of the older subjects, as the lower end of its range is still older than the mean of all the other clusters.



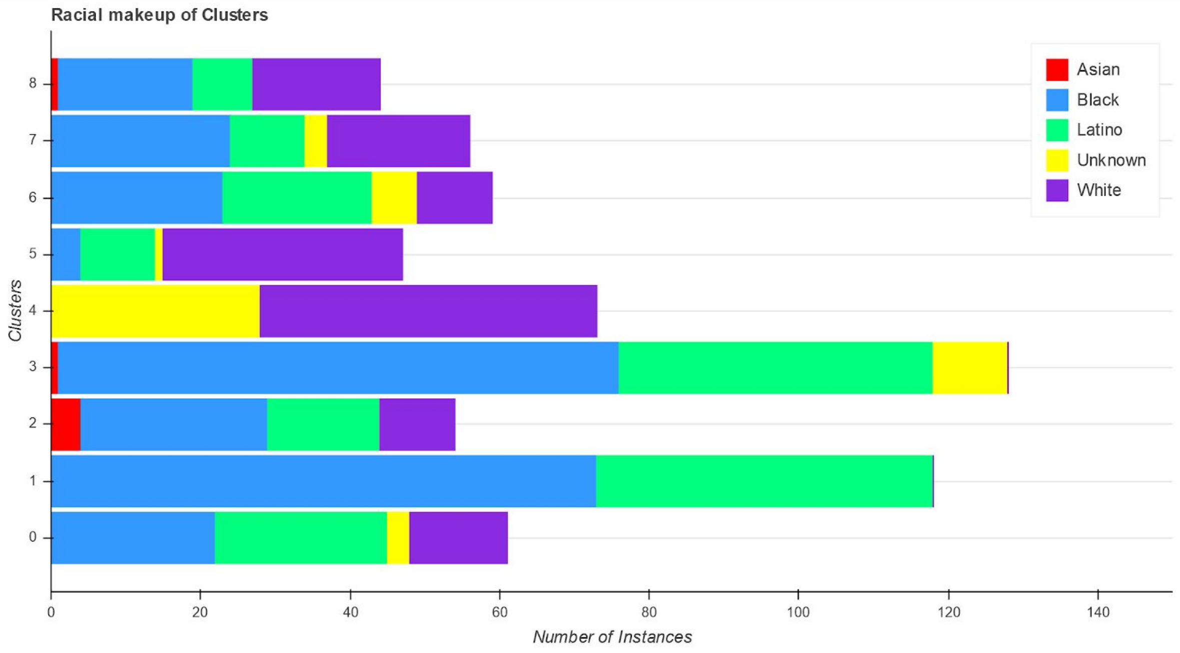
For the feature of whether subjects were armed or not, cluster "0" is made up almost entirely of incidents where the subject was armed, with only one incident where it is unknown. Cluster "1" is entirely incidents were the subject was not armed. Cluster "2" is entirely incidents where the subject was armed. Cluster "3" is mostly incidents where the subject was armed, with just a few where the subject was not armed. Cluster "4" has a majority of incidents where subjects were not armed, with a portion that were armed and a couple incidents that were unknown. Cluster "5" has a majority of incidents that were armed, with a fairly large portion that were not armed, and once incident that was unknown. Cluster "6" is entirely made up of instances where the subject was not armed. Cluster "7" is also made up of instances where the subject was not armed, with one case that was unknown. Cluster "8" is only instances where subjects were armed.



Cluster “0” is made up of only incidents that were fatal. Cluster “1” is made up mostly of non-fatal incidents with just a couple that are unknown. Cluster “2” is entirely non-fatal incidents. Cluster “3” is mostly non-fatal incidents, with just a couple that are unknown. Cluster “4” is almost entirely non-fatal incidents, with a couple that are unknown. Cluster “5” is about two-thirds fatal and one-third non-fatal. Cluster “6” is about a quarter fatal and the rest non-fatal incidents. Clusters “7” and “8” are both made up entirely of fatal incidents.



Clusters “0”, “1”, “3”, “4” and “7” are all incidents that involved only one officer. Clusters “2”, “6” and “8” are made up of incidents that involved more than one officer. Cluster “5” includes mostly incidents that only involved one officer, but has a portion of incidents with multiple officers.



The Kmeans Clustering did not appear to use "Race" as a focal point for grouping the data. All clusters are made up of incidents with multiple different subject races. Cluster "0" is about an equal amount of Black and Latino with a smaller portion of White and a few instances are unknown. Cluster "1" is a little over half Black and the rest are Latino. Cluster "2" has a little under half of its instances as Black with smaller portions of Latino and White with a few instances of an Asian subject. Cluster "3" is more than half Black with a large portion Latino, a small portion of unknown instances and one occurrence of an Asian subject. Cluster "4" is over half White and the rest unknown. Cluster "5" is about two-thirds White with a small portion of Latino subjects, a few Black subjects and one unknown. Cluster "6" is a little over one-third Black, about one-third Latino with small portions of White and unknown subjects. Cluster "7"is a little under half Black, with a large portion of White subjects, a smaller portion of Latino subjects and a couple instances of unknown. Cluster "8" has about equal portions Black and White with a smaller portion of Latino and one Asian subject.

Conclusion

Exploratory analysis found some interesting relationships. For example, it was shocking to find that the second leading cause of death in custody was "Homicide by Law Enforcement/Correctional Staff" (though it was a much lower percentage than the leading manner of death"Natural Causes/Illness"). Though after clustering, it became more obvious that most of these incidents occurred while the subject was in "Police Custody (pre-booking)", so the deaths likely happened during a confrontation with officers before or while being arrested. While police violence and shooting of civilians is an issue that has become increasingly vocalized, the fewer "Homicide by Law Enforcement/Correctional Staff" deaths that occurred while the subject was in Penitentiary, County or Municipal Jail are somewhat more concerning and are certainly much more overlooked than killings by officers in the field.

Another unexpected finding was that, in Officer Involved Shootings, whether the subject was armed or not only had a slight effect on the fatality of the incident: nearly 28% of unarmed subjects were killed vs 38% of armed subject.

It was also surprising to find that more white people were killed in both armed and unarmed Officer Involved Shootings, despite the fact that Black and Latino people were involved in more total shootings. It remains unclear why this trend was found. This author believes it might be possible that White subjects were more likely to be engaging in more violent/serious crimes, but without further digging, exploration and perhaps additional data it remains a mystery. To showcase this point, when Houston Police Department is excluded from the "Officer Involved Shooting" data, White subjects become the majority, so it makes sense that White subjects would have the highest total fatalities. However upon further examination it still appears that a higher percentage of White subjects were killed than other races. Since there is no data on "Nature of Stop" for Houston PD, it cannot be examined in relation to Race/Fatality, at least not with this data set (the same effect was found whether or not subjects were armed).

Once the data were clustered into groups, the nature of the incidents became more clear, even without analyzing any narratives of the situation. Below are summarized tables of the trends found in each cluster:

Deaths in Custody		
Cluster 0 ● Older (but not much) ● Mostly natural causes ● Half black, Half Latino ● Penitentiary mostly	Cluster 1 ● Older (but not much) ● Mostly natural causes ● All White ● Penitentiary mostly	Cluster 2 ● Mostly natural causes, some suicide, bits of other manners of death ● Half black, half Latino ● County jail mostly
Cluster 3 ● Large % death by Law Enforcement ● All white ● Police custody mostly	Cluster 4 ● Largest % with death by Law Enforcement ● Mostly Latino, large Black % also ● Police custody mostly	Cluster 5 ● Half Natural cause, half Suicide, a few other manners of death ● White ● County jail mostly

Cluster 6 - More than half Suicide - Pretty evenly Black, White, and Latino - Mostly penitentiary, but good portion in Police Custody		

Officer Involved Shootings		
Cluster 0 - Mostly armed - One Officer - Evenly Black and Latino, some white	Cluster 1 - All not armed - One officer - Mostly Black, good portion Latino	Cluster 2 - All armed - Multiple officers - About half Black, large portion Latino, some White, largest # of Asian of all clusters (but still low)
Cluster 3 - Almost all armed, rest unknown - One officer - More than half Black, good portion Latino, some unknown, one Asian	Cluster 4 - Mostly not armed, some armed - One officer - More than half white, rest unknown	Cluster 5 - Significantly older than other clusters - Majority armed, about 1/3 not armed - Most just one officer - Two-third white, some Latino, a few Black
Cluster 6 - Not armed - Multiple officers - Over 1/3 Black, about 1/3 Latino. Some White, a few unknown	Cluster 7 - Not armed, except 1 unknown - One officer - Over 1/3 Black, about 1/3 White, some Latino, a few unknown	Cluster 8 - Armed - Multiple officers - Equal of Black and White, some Latino

These groupings could potentially be useful in identifying biases or recognizing the strength of previously known or unknown trends. Also, in clusters where the cluster is homogeneous on a particular feature except for one or a few “unknown” instances, these instances could probably be interpolated to be of the same category as the rest of the points in that cluster, at least for that feature.

Overall, these data sets have many yet undiscovered insights within them, and this project has only begun to scratch the surface. There are unanswered questions that can only be answered by additional data, which may not have been collected or is not currently in an accessible form. For instance, with the Deaths in Custody data, some of the rates of occurrence in regards to race are hard to determine to be biased or show discrimination, since the rates of those incarcerated or that interact with police are not equal to population statistics. While there may be an overall bias in the criminal justice system, it is unclear whether there is a racial bias in these particular instances of death without comparing the rates to the overall rates of all people in police custody (who did not die). Comparing this data simply to population ratios may just show a reflection of the already well-known and well-documented racial imbalance, and not necessarily that being in custody is more deadly for people of particular races.

There are many specific ways to expand this analysis and ways in which the analysis could be improved or corrected. Better feature engineering would have helped the accuracy an usefulness of the clustering. For example in the Officer Involved Shootings data, “Officer Race” could have been re-coded as “White only” vs “at least one non-white officer” as this is a likely place to look for biases and a way to reduce the feature to a less-sparse set of values.

Another way in which the analysis could be improved is if the unsupervised machine learning clustered the data on a “Department” level, meaning that the data would be grouped by department so that each individual Police Department’s behavior could be examined separately from the others.

One last potential source of valuable information that was not utilized in the project was the “Summary” data available in the Deaths in Police Custody data. These fields are the written description of each incident by a police officer. These field could have been analyzed with Natural Language Processing to gain more insight into the incidents, and potentially to detect nefarious behavior by officers or other involved people. Although there was not such a feature in the Officer Involved Shooting data, it seems likely that such data could be collected or requested.