**Deaths in Police Custody and Officer Involved Shootings in Texas**

Capstone 2 Project for Springboard Data Science Career Track

1. Define the problem

From 2005-2016 more that 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 death 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?

1. 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 you the informatiuon to improve officer training in order to reduce deaths and improve outcomes.

1. Describe your data set, and how you cleaned/wrangled it

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> )

1. List other potential data sets you could use

The FBI has the Uniform Crime Report, which I could use 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.

1. 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.

1. Findings

All th thinggsf’kn

1. 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 repaced with simply “2”. For the Subject Age, a couple entries were “Juvenile”, which were replaced with “15” as a median age guess for a sunject 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 men 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--