Class and Racial Inequality in Police Killings

Abirami Varatharajan

APID: A57842201

CMSE 381

Final Report

Introduction

Police officers play an important role in this society to protect citizens by assuring their safety. Historically and even in the present, it is notable how there are deep-rooted issues within racism with the service of the officers. This project aims to develop a model to find a relationship of police killings and the intersectionality between class and racism using methods such as K-means clustering and Principal component analysis. Features of race ethnicity, house income and college will be used to see which class group gets shot.

Related Work

There are studies shown that there is a relationship between countries and police killings. "In the United States, police officers fatally shoot about three people per day on average, a number that's close to a yearly totals for other wealthy nations" [1]. In a SQL analytics project, Eng and Wenig [2] tried to analyze trends around police killings in shootings by race and state, use of body cameras, and mental illness.

When they first tried to find the fatalities by race, they found out that more white people were shot by the police than any other race. There were 51% Whites, 26% Blacks, 19% Hispanic, 2% Asian, and 2% Native Americans. However, they also found out how there were different results in some states. Blacks made up 10% in Utah, were 10 times higher than white people in Illinois, and the fatality rate for them were higher than other races in Alaska, California, and New York. California also had the highest racial disparity among the Alaska, California and New York. Not only did they compare to states, they also found out that 4% of victims are females.

When replicating the results by using the five thirty eight dataset, there were some different results which was likely due to the fact that a different dataset was used. However, when first replicating, there were some similarities in the results such as how there were overall more white people shot than any other race by the police, about similar percentages by race in the histograms, California had the highest racial disparity among the three states, and a small amount victims were female. There were also differences in the results such as how there were only white people shot in Utah, it was equal in white and blacks shot in Illinois, and the fatality rate for white people were higher than other races in Alaska, California, and New York.

Dataset

The main dataset that was used in this project is the five thirty eight police killings dataset [3]. The dataset includes 467 entries, each with 34 features consisting of both numerical and categorical values. The programming language used for this project was R. For the initial preprocessing, there were changes done in order to fit a model such as removing all the na values by using ‘na.rm = TRUE’ and changing categorical values into numerical values by using ‘as.numeric’.

Method

There are many drawbacks from the SQL analysis blog such as how they only tried to find a trend between states, race, and body cameras. This analysis also didn't include features such as income, city, armed or college. However, these features along with many others, were included in the police killings dataset which lead to many questions being asked that the blog didn’t answer such as “Is there a relationship between police killings and income of the people?”, “Is there a relationship between police killings and if they attained a college degree?” This however lead to the question of this project which is “Is there a relationship of police killings and the intersectionality between class and racism?” In order to find the relationship and patterns from the data, there are two unsupervised learning methods being used which are principal component analysis and k-means clustering. Both principal component analysis and k-means clustering are helpful in simplifying the data.

Principal component analysis is primarily used for dimensionality reduction. Having high dimensionality means that there are large number of features which can lead to overfitting. But overfitting is an issue since this could create problems when predicting unknown future data. So, principal component analysis reduces the dimensionality of large data sets by transforming it into a smaller data set. The algorithm used for principle component analysis [4]:

1. The first principal component of a set $X_1, X_2, \ldots, X_p$ features is the normalized linear combination of the features:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

2. The principal component loading vectors then solve the optimization problem:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

K-means clustering is primarily used for identifying groups within your data that have not been explicitly labeled which then would be easy to find patterns within your data. The number of groups is determined by the variable K where then based on the similarities of the features that are provided, the data points would be assigned to one of the K groups. The algorithm used for the K-means clustering [4]:

```
1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial
   cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   2.1. For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the
        vector of the p feature means for the observations in the kth cluster.
   2.2. Assign each observation to the cluster whose centroid is closest (where closest is defined
        using Euclidean distance).
```

Experiments and Discussion

First, principal component analysis was used in this project to reduce the dimensionality. NA values were omitted in order to do a PCA and a bi-plot was created in order to see the results.

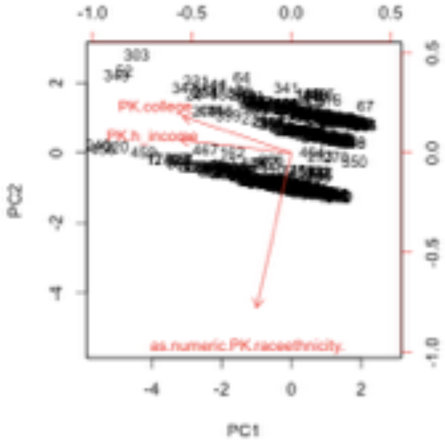

Figure 1

Figure 1 shows the bi-plot of the PCA between race ethnicity, college graduates, and house income. Looking at the above plot, the vectors of college graduates and house income are similar together meaning that there is a relationship between the two in terms of police killings. However, there is not a significant relationship in terms of race and the two factors.

K-means clustering was also used in this project to find similarities within the data and cluster them into groups. NA values were also omitted in order to do a clustering and a plot was created in order to see the results.

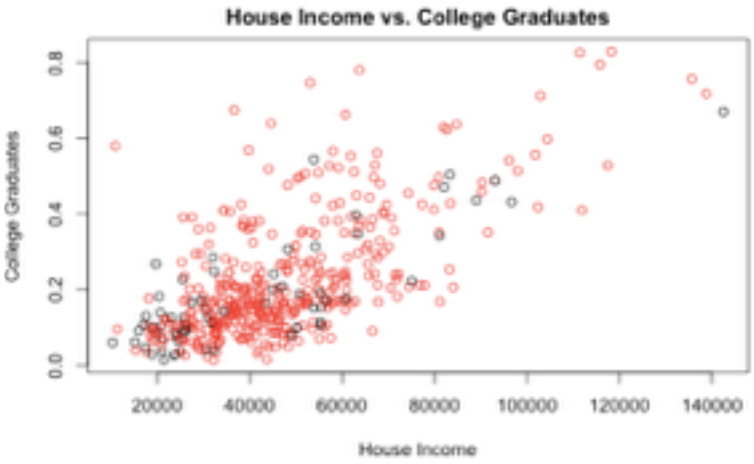

Figure 2

Figure 2 shows the plot of the K-means clustering between house income and college graduates. Looking at the above plot, there isn't a clear group but it is clearly shown that most of the points are towards the bottom left of the graph meaning that there are police killings towards people who have low income and who are in the lower part of the percentage of people who graduated with a BA or higher.

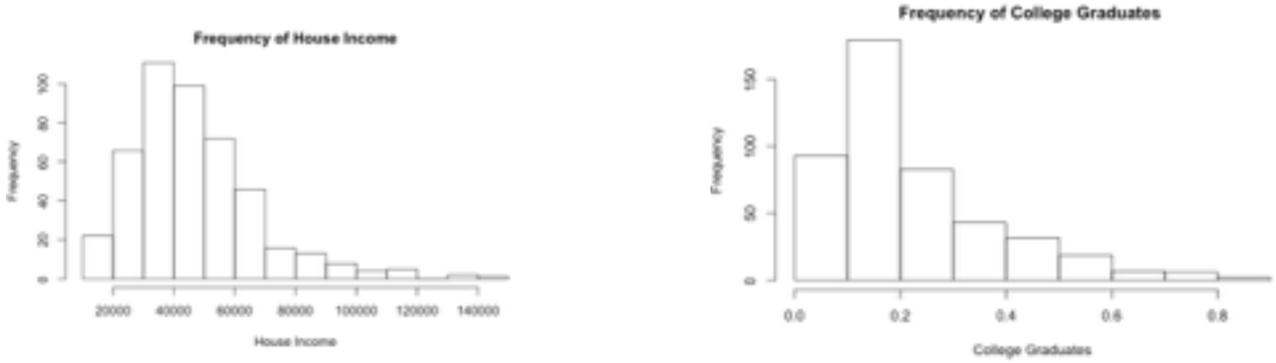

Figure 3

Figure 3 shows histograms of the frequency of both house income and college graduates. This shows how there are high frequency towards the people who have a low house income as well as people who are in the lower part of the percentage of people who graduated with a BA or higher.

Looking at all three figures, it is clear that house income and college graduates are very similar and have alike values. Not only can we see that their vectors were close together in the PCA bi-plot, their clusters as well as the histograms show that they do get shot towards the low part of their plots which suggests that there is an intersection between house income and college graduates.

Conclusions and Future Work

This project attempts to come up with finding a relationship of race and the intersection between house income and college graduates. The machine learning techniques such as K-means clustering and principal component analysis were used to achieve the best results in terms of finding a relationship. Although the models did show that there was an intersection between house income and college graduates, it did not show if there was a relationship between race and that. Among the models tested, PCA showed more of the intersection between the two features than K-means clustering. With the plots shown, it does lead to the claim that there is a relationship between police killings of people who are in the lower part of the percentage of graduating with a BA or higher and people who have a lower house income.

The future work of this project may include (i) using other models such as SVM or KNN, (ii) using other features from the dataset such as using age and being armed, (iii) getting more data from different years.

References

[1] Peeples, Lynne. “What the Data Say about Police Shootings.” Nature News, Nature Publishing Group, 4 Sept. 2019, www.nature.com/articles/d41586-019-02601-9.

[2] Eng, Chengyin, and Brooke Wenig. “Analysis of Police Fatal Shootings in the U.S.” Databricks, 2 Dec. 2020, databricks.com/blog/2020/11/16/fatal-force-exploring-police-shootings-with-sql-analytics.html.

[3] Fivethirtyeight. “Fivethirtyeight/Data.” GitHub, github.com/fivethirtyeight/data/tree/master/police-killings.