Analysis of Seattle 911 Police Incident Response Time Using Machine Learning

Cheryl Kung   
School of Science, Technology, Engineering & Mathematics  
University of Washington BothellBothell, USA  
cckung@uw.edu

***Abstract*** – In this paper, I analyze the [Seattle 911 Police Incident Response dataset](https://www.kaggle.com/sohier/seattle-police-department-911-incident-response) to see if there are any patterns between the time at the scene of an event, the specific type of event, and the characteristics of the location of the event. The dataset was augmented by features from the US Census as well. The dataset was transformed for analysis by using unsupervised machine learning. Principle Component Analysis (PCA) was used to reduce the dimensions of the dataset, then the data as sorted using k-means clustering. The clusters were then manually analyzed for distinct variations.

## Introduction

Due to the history of poor police relationships in America with the BIPOC community, more needs to be done to understand this behavior. Finding patterns in police behavior can lead to effective policies that can prevent unnecessary police violence that endanger the public. Though the dataset only shows a small glimpse into police behavior, this information can direct new areas of studies.

## Dataset

The Seattle 911 Incident Response dataset was provided by city of Seattle. It contains information on the 911 calls that police responded to from 2010 to 2017. Each row contains information for one specific event.

### Features Used

The dataset contains 18 features, however the two features “Initial Type” and “At Scene Time” are only available for events in 2011 and later. Many of features also represent the similar information. For example, the location of incident is described by the “Census Tract”, “Longitude”, “Latitude, “Incident Location”, “Hundred Block Location”, and “District/Sector”. The core information from the dataset was the type, location, and time. The features chosen to be used for machine learning were represented by the “Event Clearance Code”, “Census Tract”, “Event Clearance Time”, and “At Scene Time”. Two additional features, the percentage of white residents and the median income of the area, were supplemented from the US Census American Community Survey using the “Census Tract” feature.

### Missing Data

To limit the scope of paper, the missing data was removed from this set. The most common areas of missing data were in the “Event Clearance Time”. There were also census tracts that did not correspond to any data retrievable from the US Census, so those were removed as well.

### Data Preprocessing

The “Event Clearance Code” and “Census Tract” feature was chosen because it was already numerically encoded. The time the police spent at the scene was calculated from the delta of the “At Scene Time” and “Event Clearance Time”. The year the incident occurred was also included because the US Census data is updated yearly.

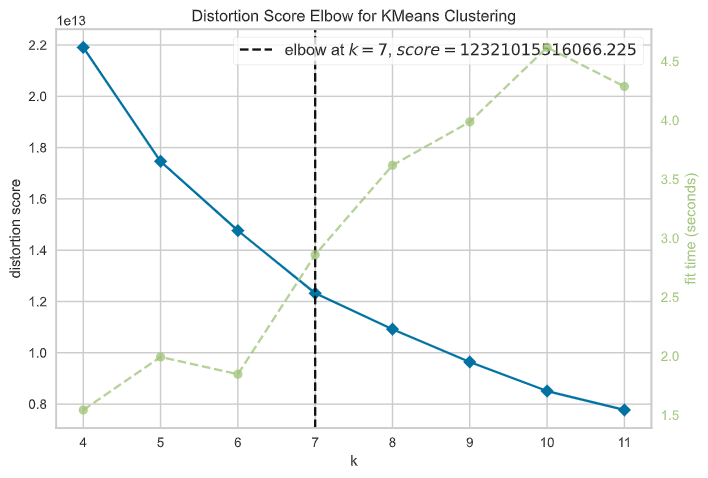
Using the [US Census API Python wrapper](https://github.com/datamade/census), the data for the percentage of white residents and the median income for each census tract was logged into lookup tables. These were then used to create the two additional features. Using PCA, the six features were reduced to two for simple graphical representation.

## Clustering

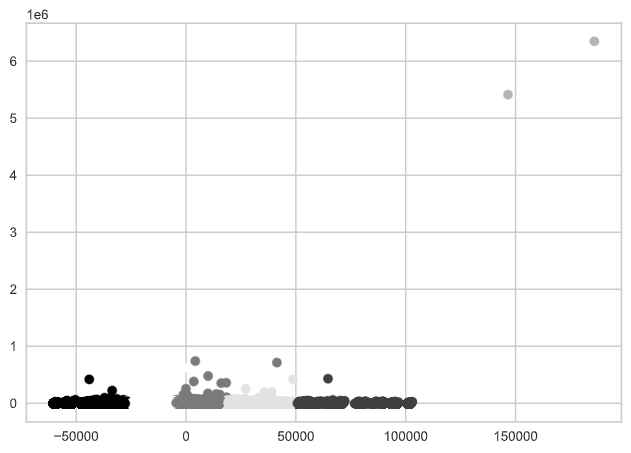
The k-means algorithm was chosen for its simplicity. Using the [Yellowbrick Python library](https://www.scikit-yb.org/), the optimal number of clusters was chosen to be seven.

Chart

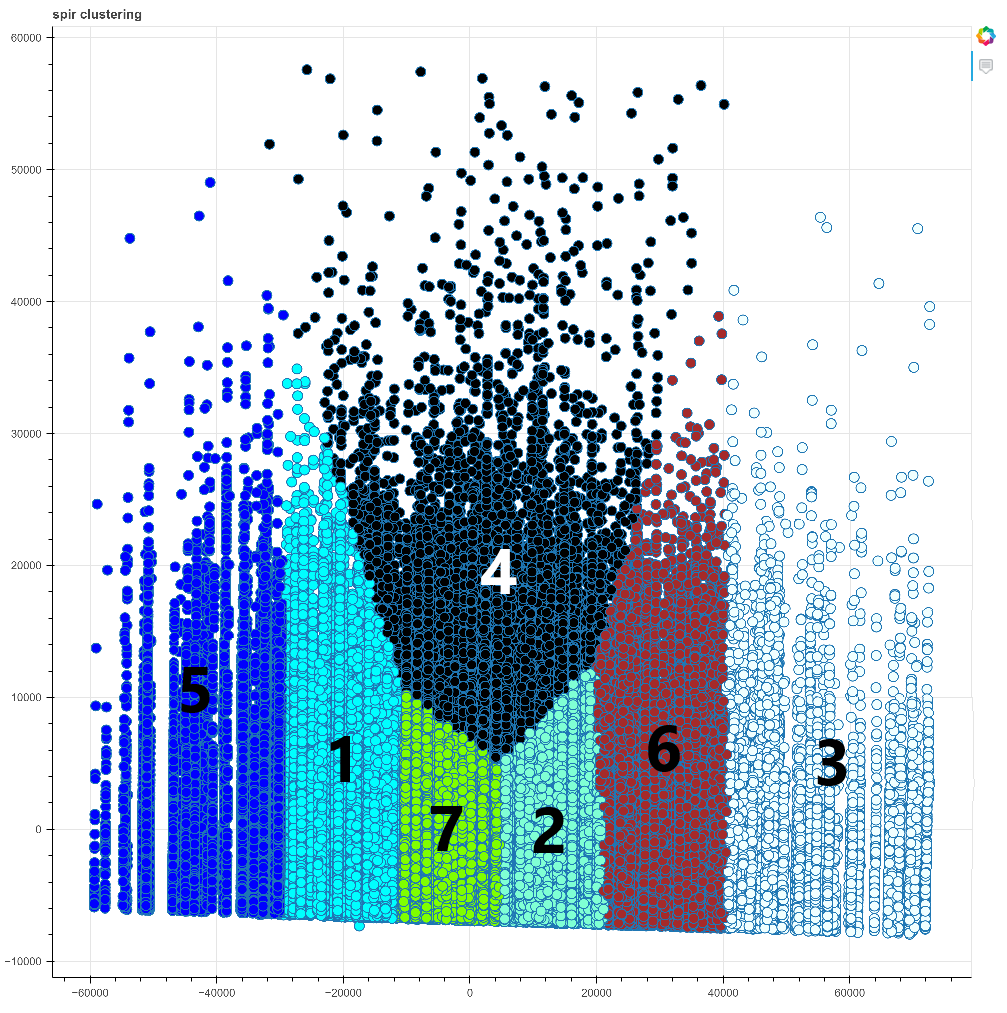
Description automatically generated



Below is the resulting graph of the PCA-transformed data.



From this visual, there are some obvious outliers. These outliers were removed by only retaining values within three standard deviations, and the k-means algorithm was run again.



The graph above is the data without outliers, processed through PCA, and labeled with the k-means clustering algorithm.

## Data Analysis

The clusters were separated into their own datasets and analyzed. First, the averages of each feature were compared in each other and the averages of the total dataset.

Most Common Type of Event:

The most common type of event the police responded to among the whole dataset was “Accident Investigation”. This was also the most common event in each cluster. There were no apparent distinctive differences in the most common types of events between the clusters. This does not seem to be a feature that the algorithm used heavily to create the clusters.

|  |  |
| --- | --- |
| Cluster | Eight most common event types |
| All Clusters |  |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |

|  |  |
| --- | --- |
| Event Clearance Code | Event |
| 430 | Accident Investigation |
| 245 | Disturbance – Other |
| 280 | Suspicious Person |
| 63 | Theft – car prowl |
| 460 | Traffic (Moving) Violation |
| 65 | Theft – Miscellaneous |
| 50 | Burglary – residual, unoccupied |
| 450 | DUIs |

### Location of Events

Though the census tract numbers are not independent from each other due to the sequential numbering of neighboring areas, that did not seem affect the clustering of data points. Based on the lack of pattern between the clusters’ top five most common census tract numbers among their incidents, the clusters do not appear to be location dependent.

|  |  |
| --- | --- |
| Cluster | Five most common census tracts (locations) where incidents occurred |
| All Clusters |  |
| 1 |  |
| 2 |  |
| 3 | Map  Description automatically generated |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |

### Year the Event Occurred

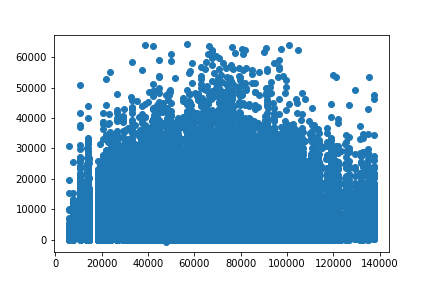
There was also no obvious skew in the clustering based on the year the incident occurred either. However, the year 2014 is notable absent, with only one incident recorded that year, which should be further investigated.

This is table contains the distribution of events by year, which a similar among the clusters

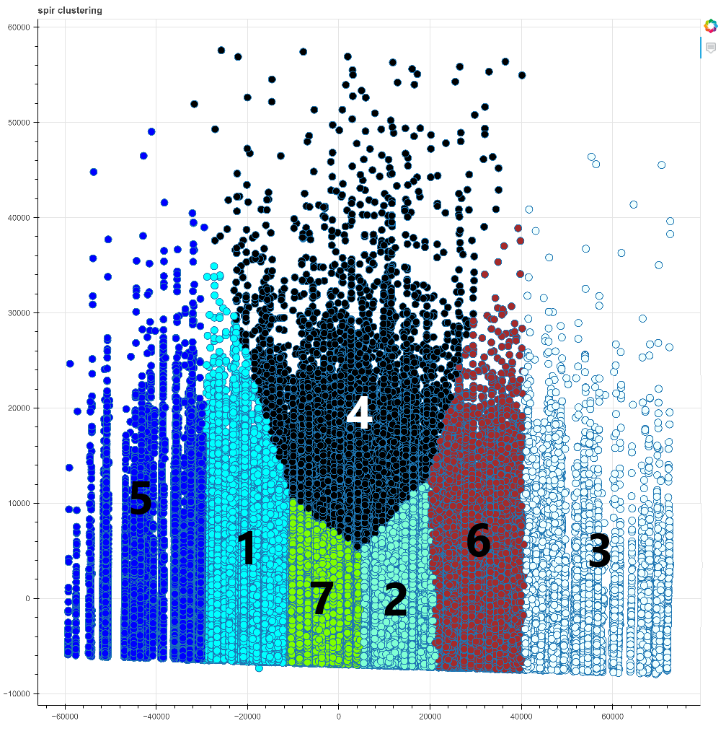
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | All | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 |
| 2017 | 0.122257 | 0.07462 | 0.143777 | 0.235558 | 0.142963 | 0.104074 | 0.187295 | 0.093643 |
| 2016 | 0.426079 | 0.324214 | 0.438117 | 0.481911 | 0.497924 | 0.416979 | 0.484102 | 0.458534 |
| 2015 | 0.208456 | 0.174387 | 0.255256 | 0.193867 | 0.267463 | 0.157329 | 0.191472 | 0.212706 |
| 2014 | 4.45E-06 | 1.96E-05 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2013 | 0.005379 | 0.0091 | 0.003764 | 0.001493 | 0.001928 | 0.004381 | 0.004247 | 0.0059 |
| 2012 | 0.160518 | 0.297481 | 0.109334 | 0.056966 | 0.074522 | 0.165847 | 0.099983 | 0.147275 |
| 2011 | 0.077307 | 0.120178 | 0.049752 | 0.030206 | 0.015201 | 0.15139 | 0.032902 | 0.081943 |
| Sum of Events | 224585 | 51099 | 51274 | 8707 | 13486 | 20543 | 28965 | 50511 |

### Median Household Income of Census Tract of Event

The PCA transformation heavily favored the time and income features of this dataset. The distribution of the time versus income looks like the PCA graph, as shown below.



|  |  |  |
| --- | --- | --- |
| Cluster | Mean | Median |
| All | 65304.8 | 65960 |
| 1 | 45067.64 | 45511 |
| 2 | 77424.04 | 76320 |
| 3 | 116435.7 | 113432 |
| 4 | 68683.11 | 69130 |
| 5 | 26044.41 | 26886 |
| 6 | 94868.49 | 93155 |
| 7 | 62773.74 | 62980 |



Looking at the averages of the median income of each cluster, the lower income clusters are on the lower side of the x-axis, while the higher income clusters are on the higher side of the x-axis.

### Percentage of White Residents

The percentage of white residents did seem to be correlated with median income of the location of the incident, but it was not correlated with the amount of time the police spent at the incident in the same way median income was. There is a slight trend toward more police time in areas with a higher percentage of white residents.

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Mean Income | Median  Income | Percentage of White Residents |
| All | 65304.8 | 65960 | 66% |
| 1 | 45067.64 | 45511 | 60% |
| 2 | 77424.04 | 76320 | 71% |
| 3 | 116435.7 | 113432 | 83% |
| 4 | 68683.11 | 69130 | 67% |
| 5 | 26044.41 | 26886 | 47% |
| 6 | 94868.49 | 93155 | 77% |
| 7 | 62773.74 | 62980 | 66% |

Chart, scatter chart

Description automatically generated

Percentage of White Residents vs Police Time at Scene

Percentage of White Residents

Time (s)

Chart, scatter chart

Description automatically generated

Percentage of White Residents vs Median Income

Median Income ($)

Percentage of White Residents

### Time Spent at the Scene of the Incident

The time spent at the scene of the incident was higher in certain clusters than other. However, it was not correlated with any notable distinction of a certain feature in that cluster, but rather only with the areas where the median household income was closest to the entire Seattle median income.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Mean Income | Median  Income | Mean Time (s) | Median Time (s) |
| All | 65304.8 | 65960 | 6864.852033751141 | 5141.0 |
| 1 | 45067.64 | 45511 | 6129.277754946281 | 4736.0 |
| 2 | 77424.04 | 76320 | 5459.731891406951 | 4715.0 |
| 3 | 116435.7 | 113432 | 7168.200987711037 | 5382.0 |
| 4 | 68683.11 | 69130 | 21555.964333382766 | 19917.5 |
| 5 | 26044.41 | 26886 | 6315.294358175534 | 4676.0 |
| 6 | 94868.49 | 93155 | 7020.226238563784 | 5430.0 |
| 7 | 62773.74 | 62980 | 5195.053176535804 | 4560.0 |

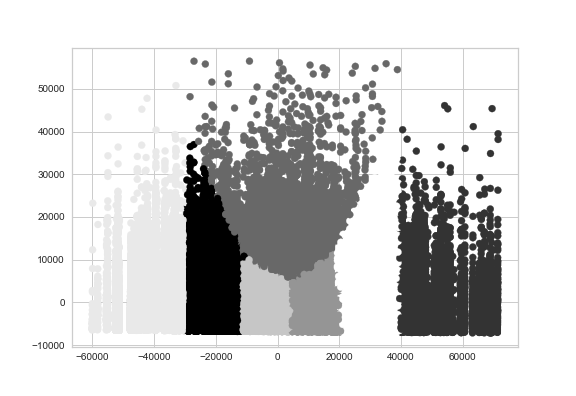
## Conclusion

The Seattle Police spend the most time in areas with median income levels closest to the overall Seattle median income. This was the main correlation found. More research into this domain will be needed to determine why this is.

### Future improvements

This analysis could benefit from add more features to find more correlations between the incident time and other factors, especially since there are many features available from the US Census API. The dataset could also be updated from using more recent data from data.seattle.gov. Additionally, the outliers initially removed during preprocessing need to be analyzed to ensure the analysis does not overlook important patterns. The analysis could also benefit from exploring using different clustering algorithms.

## Addendum

Because of feedback from the presentation, I have included this clustering graph where if the incident had a response time less than fifteen minutes, it was removed. 

The graph still has similar distribution and clustering. Though it does not seem logical that the police could spend less than fifteen minutes to clear up an incident, there is not enough information provided on how the data is recorded to regard these data points as incorrect. They may be the result of human error if the times are recorded manually.