Fatal Police Shooting Analysis

Charanpreet Kaur ID: 23191482

Professor: Dr. Mariam Adedoyin-Olowe

1. Contents

1. Introduction:

Introduction to the domain:

We have chosen law enforcement and public safety domain, as it plays a crucial role in preserving a social order as well as it guarantees community safety. In past recent year, there's major stress on data driven way to address different difficulties defying law enforcement agencies. One of the main issue being is fatal police shootings, this created series of questions regarding police and community relations and public's trust. After close examining data on fatal police shooting, these law enforcement agencies can acquire important analysis, and analysis can contribute huge to discover what are the paters or trends and also this analysis will be helpfile to reduce risks and improving community safety.

Data analysis is a critical step to improve law enforcement tactics and also ensuring public's safety. These law enforcement agencies can then use the analysis provided by data driven approaches in order to make educated judgements and ca execute involvements which will be helpful I am addressing issues around fatal police shootings.

Data analysis also enables them to detect trends, patters and any risk factors and because of already being aware of these issues, taking practice measure to reduce them would be easier. Furthermore, data analysis also enables openness, accountability, and increases trust within communities. If properly focussed on this domain, we can help addressing, social challenges, like police shootings. By using data analysis on fatal police shootings, we hope to shed light on core factors, and trends which might be interrelated. This domain also provides the opportunity in make contributions in creating safer and more decent communities.

2. Problem Statement:

Problem statement: Understanding Demographic factors in Fatal police shootings.

Fatal police shootings ins a serious and complex issue, which often leads to question like what the underlying issues were and if biases was involved. By Understanding demographic characteristics, it would be easier to address concerns and policies with the law enforcement. Furthermore, after exploring factors like age, gender, race, signs of mental illness, trends and patterns can be identified. This analysis also provides, uneven impact of fatal police shooting on couple groups.

3. Objective:

Ther objective of this analysis is to investigate demographic factors which are deeply associated with fatal police shootings and how their implications are relate to law enforcement policies and public safety. Our aim is to uncover relations between these patterns in fatal police shooting incidents. Our main objective is to produce evidence-based insights that can address system and can promote accountability as well as, improve the outcomes of communities' interactions with law enforcement agencies. After thorough analysis and understanding, we pursue to contribute development of more reasonable, clear and ana effective policing practices.

4. Literature Review:

Theres complex dynamics that surrounds fatal police shooting data and researching that beforehand was a crucial step and understanding what the underlying issues might be. In order to get through these issues, research uses different methods, qualitative analysis of large scale of dataset, qualitative assessment of individual cases also mixed-method approaches. Some of the existing datasets shows missing aspect of mental illness and young men of colour in dataset.

To understand my domain in detail and to do analysis properly, I first researched about previous findings related to my domain. These resources provided valuation insights regarding the topic and guided me in analysis. Some of the like resources are following:

- Lowery, W., Kelly, K., Rich, S., Tate, J., & Jenkins, J. (2019). Fatal Force: 2019 Police Shootings Database. The Washington Post[1] Fatal Force 2019: 2019 police shootings database: This report actually tracks down US police shooting. In 2019, it tracked down each and every occasion where a police officer might or have to fatally shot someone. This database provide deep information regarding frequency, what might be geographical distribution or any circumstances of fatal police shootings. It also provided demographic of victims, like their age, gender and the reasons that led to their deaths.
- Burghart, D. B. (2020). Fatal Encounters [2]Fatal encounters: Database of people killed during interactions with law enforcement: this report provided me with the information about peoples who have been killed during their interaction s with law enforcement. It also provided data tgo understand the circumstances that lead to the incidents.
- Gupta, P. (2018). Data Mining: Employee Turnover in a Company[3]: this report discusses application of data mining techniques, machine learning models using pandas, NumPy, seaborn etc
- Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013)[4] .Predictive policing : role of crime forecasting in law enforcement operations: This particularly explores, data mining techniques specifically k means clustering under predictive policing. Also show these methods can be applied to crime dataset.
- Nix, J., Pickett, J. T., Wolfe, S. E., & Campbell, B. A. (2017)[5]. Demeanour, race, and police perceptions of procedural justice: Evidence from two randomized experiments. The Journal of Politics, 79(4), 1154-1169.: this examined how race and public demeanour impact on procedural justice.
- Lewie, J., & Fagan, J. (2019). Aggressive policing and the educational performance of minority youth. American Sociological Review, 84(2), 220-247[6] : This report showed aggressive policing on minority youth, and races.
- Goff, P. A., Obermark, D., La Vigne, N., Yahner, J., & Geller, A. (2016). The science of justice: Race, arrests, and police use of force. Center for Policing Equity[7]: this report showed racial inequality in police forces.
- Goff, P. A., & Kahn, K. B. (2012). Racial bias in policing: Why we know less than we should. Social Issues and Policy Review, 6(1), 177-210.[8] : this article displayed the present literature on bias based on races, minorities..
- Legewie, J. (2016). Racial profiling and use of force in police stops: How local events trigger periods of increased discrimination. American Journal of Sociology, 122(2), 379-424[9]: this report shows how high profile police shootings are impacted on local events..

5. Data-Set Description:

The fatal police shootings dataset I chose is from Kaggle. This database documents fatal police shootings of on duty police officers in US, also including fatal police shootings from 2015 to 2020 also including almost 5000 instances which provides significant number of records of criminals . Dataset also provide different features which can be useful for the analysis.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5338 entries, 0 to 5337
Data columns (total 14 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   id                       5338 non-null   int64  
 1   name                     5338 non-null   object 
 2   date                     5338 non-null   object 
 3   manner_of_death          5338 non-null   object 
 4   armed                    5098 non-null   object 
 5   age                      5089 non-null   float64
 6   gender                   5336 non-null   object 
 7   race                     4731 non-null   object 
 8   city                     5338 non-null   object 
 9   state                    5338 non-null   object 
 10  signs_of_mental_illness  5338 non-null   bool   
 11  threat_level             5338 non-null   object 
 12  flee                     5088 non-null   object 
 13  body_camera              5338 non-null   bool   
dtypes: bool(2), float64(1), int64(1), object(10)
memory usage: 511.0+ KB
```

Figure 1 Columns of Fatal Police shootings dataset.

	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_mental_illness	threat_level	flee	body_camera	year	month
0	3	Tim Elliot	2015-01-02	shot	gun	53.0	M	A	Shelton	WA	True	attack	Not fleeing	False	2015	1
1	4	Lewis Lee Lembke	2015-01-02	shot	gun	47.0	M	W	Aloha	OR	False	attack	Not fleeing	False	2015	1
2	5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23.0	M	H	Wichita	KS	False	other	Not fleeing	False	2015	1
3	8	Matthew Hoffman	2015-01-04	shot	toy weapon	32.0	M	W	San Francisco	CA	True	attack	Not fleeing	False	2015	1
4	9	Michael Rodriguez	2015-01-04	shot	nail gun	39.0	M	H	Evans	CO	False	attack	Not fleeing	False	2015	1

Figure 2 Representation of dataset Values

Database contains reco0rds of fatal police shootings in the United States, with 5338 entries. Each column is detailed with 14 further entries regarding that column, covering both incident-specific and victim-specific information. These data values include unique variables “id”, “name”, “date” and “manner of death”, which distinguish between shootings and other fatal interactions. This dataset serves as a foundation for the application of Principal Component Analysis and K-Means clustering, to uncover the underlying trends and patterns of this dataset.

6. Dataset Pre-Processing:

Before leading to analysis of dataset, several pre-processing steps are crucial to perform in order to make dataset suitable for the analysis.

1. Handling Missing Values:

Missing values is common issue and if this issue is not resolved, then analysis result would show bias nature. Also by not resolving null values it could cause skewness in dataset and changing whole analysis. Also errors can occur while training the dataset. By handling missing values properly, it

makes sure that dataset introduce no bias to any specific variable. In our dataset, we found missing values in various columns such as “Age”, “Armed” and more. In order to handle these values, we can use different methods to impute values.

- For missing numerical values like “age”, we chose to impute missing values with mean age value. We first assessed if mean is suitable or median based on the skewness of dataset but it turned out dataset has evenly distributed dataset which also means in simpler words, data is spread out in same amount on both sides, creating a bell like shape. So, we went with mean. In order to impute missing values
- For another variable like” race” and “gender”, missing values were imputed using mode method, under this method, The most occurring elements are placed in place of missing values which is suitable as well for race.
- For the categorical variables, such as “armed” and “flee”, these variable’s missing values were imputed by, “unknown”, first these were few values that were missing, by adding unknown it kept integrity of dataset also avoided biasing dataset towards any specific category.

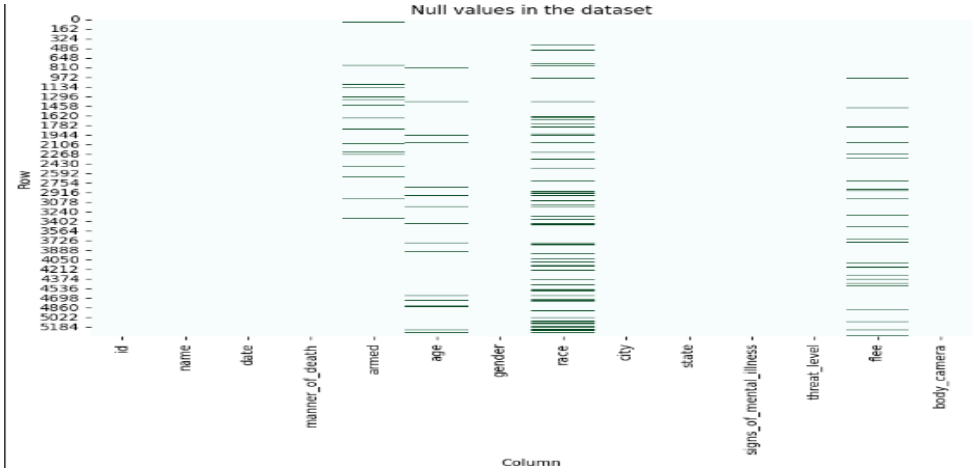

Figure 3 Null values in dataset.

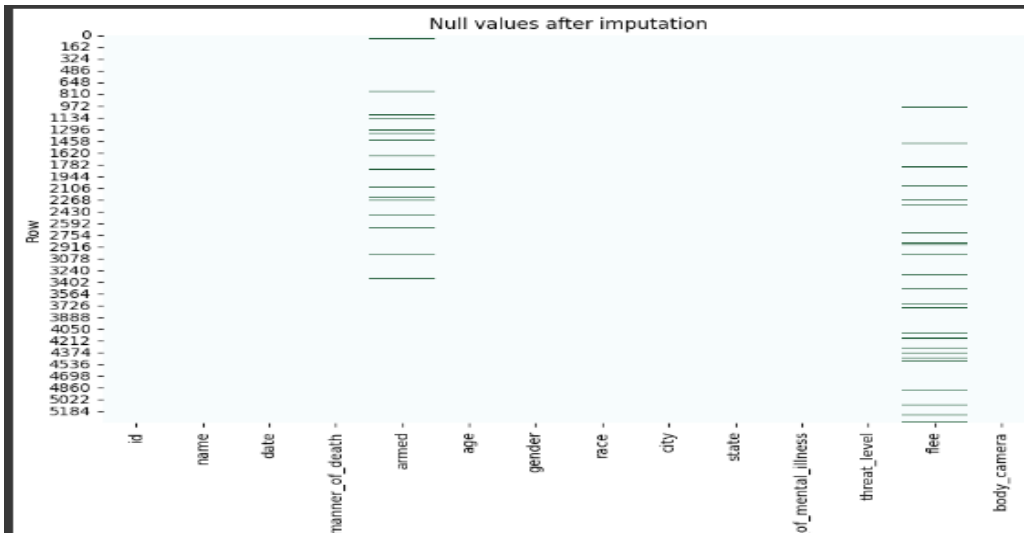

Figure 4 Null values in "race " and "age" has been imputed

```
id                          0
name                        0
date                        0
manner_of_death             0
armed                     240
age                       249
gender                      2
race                      607
city                        0
state                       0
signs_of_mental_illness     0
threat_level                0
flee                      250
body_camera                 0
dtype: int64
```

```
id                          0
name                        0
date                        0
manner_of_death             0
armed                       0
age                         0
gender                      0
race                        0
city                        0
state                       0
signs_of_mental_illness     0
threat_level                0
flee                        0
body_camera                 0
dtype: int64
```

Figure 5 Null values(left), no null values left(right)

2. Encoding Categorical values:

Categorical values are variables that represents some specific groups or categories in a dataset. These are qualitative data and are represented as text labels. Now, if they are left like that, because these variables are text labels and not numerical so, most machine learning algorithms would be unable to interpret that data value. So it's a best practice as part of preprocessing dataset to encode all the text label values in to numerical values. After encoding, model can easily learn patterns and trends of dataset which results in accurate analysis. So based on our dataset we have used label encoding on variables i.e., threat level, body camera, gender, signs of mental health wellbeing.

```
Before Label Encoding:
   id              name        date   manner_of_death       armed   age  \
0   3        Tim Elliot  2015-01-02              shot         gun  53.0
1   4   Lewis Lee Lembke  2015-01-02              shot         gun  47.0
2   5  John Paul Quintero  2015-01-03  shot and Tasered     unarmed  23.0
3   8     Matthew Hoffman  2015-01-04              shot  toy weapon  32.0
4   9   Michael Rodriguez  2015-01-04              shot    nail gun  39.0

  gender      race           city state  signs_of_mental_illness threat_level  \
0      M     Asian        Shelton    WA                     True       attack
1      M     White          Aloha    OR                    False       attack
2      M  Hispanic        Wichita    KS                    False        other
3      M     White  San Francisco    CA                     True       attack
4      M  Hispanic          Evans    CO                    False       attack

         flee  body_camera
0  Not fleeing        False
1  Not fleeing        False
2  Not fleeing        False
3  Not fleeing        False
4  Not fleeing        False
```

```
After Label Encoding:
   id              name        date   manner_of_death       armed   age  \
0   3        Tim Elliot  2015-01-02              shot         gun  53.0
1   4   Lewis Lee Lembke  2015-01-02              shot         gun  47.0
2   5  John Paul Quintero  2015-01-03  shot and Tasered     unarmed  23.0
3   8     Matthew Hoffman  2015-01-04              shot  toy weapon  32.0
4   9   Michael Rodriguez  2015-01-04              shot    nail gun  39.0

  gender      race           city state  signs_of_mental_illness threat_level  \
0      M     Asian        Shelton    WA                     True       attack
1      M     White          Aloha    OR                    False       attack
2      M  Hispanic        Wichita    KS                    False        other
3      M     White  San Francisco    CA                     True       attack
4      M  Hispanic          Evans    CO                    False       attack

         flee  body_camera
0  Not fleeing        False
1  Not fleeing        False
2  Not fleeing        False
3  Not fleeing        False
4  Not fleeing        False
```

Figure 6 Before label encoding(left), After label encoding(right)

Label encoding is form of data pre-processing method which converts label values into numerical values. It can map each category with distinct number typically starting from 0. For example, if column name is colour, and has data values as orange, green, red, it would assign unique numbers to those values as 0,1, 2.

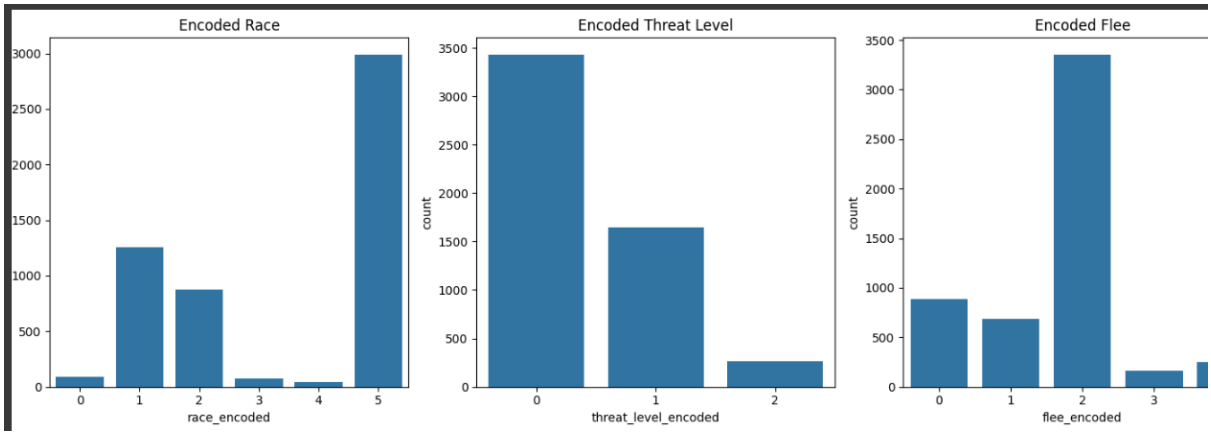

Figure 7 Visual representation of encoded values.

By encoding label values, it ensures, that algorithms can interpret and can easily process these variables correctly. By having label encoding, it also ensures in preventing bias and error.

3. Scaling Numerical values:

Scaling numerical values is essential specially for the algorithm which are sensitive to scale of features. In fatal police shootings dataset, we have numerical variables as “age” and “year”, because of having different scales, this can negatively impact analysis. In order to address this issue, we applied scaling technique. We used, standard scaler in order to scale numerical values which would provide mean of 0 and standard deviation of 1. Standardization of data values ensures al numerical variables equally contribute to the analysis and prevents larger scale of data value to dominate in the analysis.

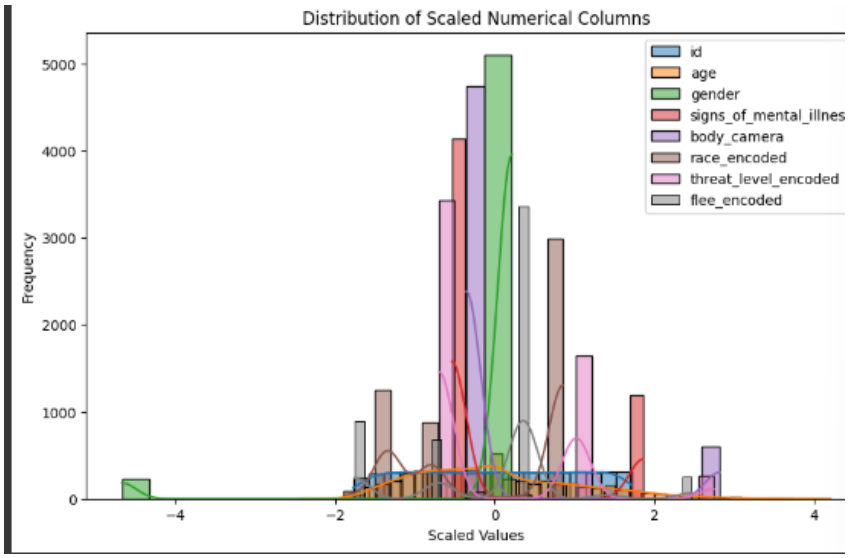

Figure 8 Scaled values representation.

In case of not performing scaling, the variables are sensitive to scale and if using k means clustering which is decided to use, would produce biased results.

By performing these preprocessing steps, it’s always ensured that dataset is prepared for the analysis. Performing data preprocessing also ensures precision and reliability of analysis, which also allows us to draw insightful analysis from output.

7. Statistical techniques using PCA & K means clustering:

We have applied Principal Component Analysis (PCA) and K means clustering, to the pre processed data in order to explore structure and identifying trends within data.

1. Applications of PCA:

Principal Component Analysis (PCA) is basically dimensionality reduction technique which is used to transform higher dimensional data to lower dimensional data form while also preserving dataset's variability in data. In our analysis, PCA was applied to reduce dataset's dimensionality, whilst also identifying most important features that might be contributing to the variance of data. In simpler words, PCA simplifies dataset , making it suitable to explore and visualize.

After applying PCA, we received principal components which represent original features but in lower dimensional. These components are linear combinations of original features of dataset. This is the best method to reduce dimensionality while also keeping important information.

2. Applications of k means Clustering:

K means clustering is unsupervised machine learning algorithm which is used to divide data into predefined number of clusters. We have applied k means clustering to identify natural groupings or clusters within dataset. After assigning data value to the cluster with nearest centroid, k means will identify hidden trends and patterns between datasets.

After allowing k means clustering, each data point in dataset was assigned to one of the k clusters based on its proximity to cluster centroids. This would allow us to identify varieties of groups of analysis within dataset.

Implementations Statistical techniques:

1. Principal Component Analysis:

We used PCA on the pre-processed dataset. The number of principal components we chose based on cumulative explained variance ratio, ensures that selected components analysis capture most of the data.

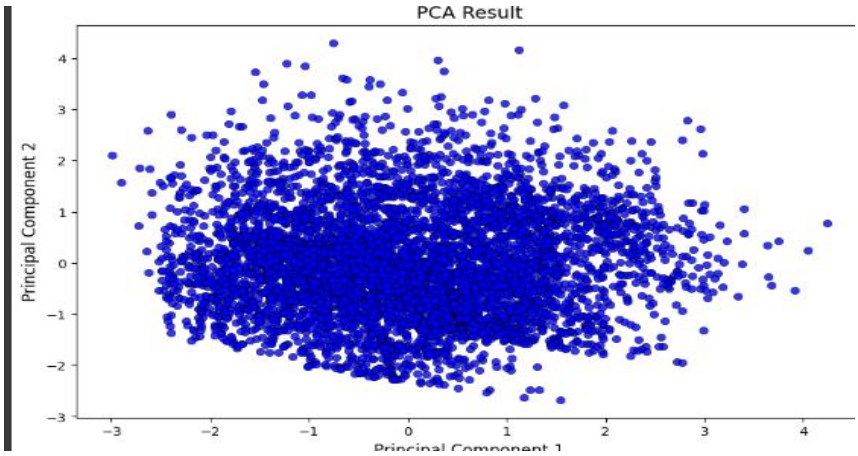

Figure 9 PCA representation between two principal components

Above 2d scatter plot visualise data in reduced dimensionality which is displayed by first two principal components that majority of data represents. This scatter pot allows the clusters to be closely packed which also represents the plots are closely similar to the original data values. This data also reveals outliners of data which are basically isolated points far from the main clusters. These might display unique observations, but as the scatter plot shows, there are less tendencies to have anomalies in our dataset as all the plots, closely packed. Because principal components have captured most of the data, there might be the furthermore reason due to which this scatter plot has been produced. Police shootings dataset includes distinct subgroups or categories. Principal components are able to capture most of the data from distinguished variables that might also be ready to close packed scatter plot. Although there are couple of plots which are overlapped but they can offer valuable insights and can help in providing deeper understanding of underlying trends and patterns within dataset.

2. K – means Clustering:

We applied k means clustering using KMEANS after applying PCA. K means analysis is not possible until k value is not defined, so we delve more deeper into k-means to understand our dataset and found k = 5 using silhouette score, which is basically a score that shows relation of columns between each other.

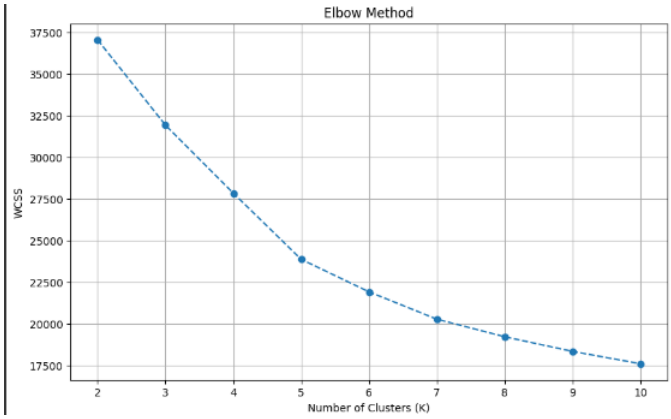

Figure 10 Elbow method.

The elbow method is heuristic method in order to determine optimal number of clusters required using Within Cluster Sum of Squares(WCSS) against number of clusters. Its a measure in which sum of squared distances between each of the data point and the centroid. Based on our chart, elbow method plot starts from 2 and end at 10, it needs to be starting from 2 as at; east 2 clusters are always required to compare them. And if we look closely, the line shows a downward bend at 5th cluster, which means 5 number of clusters would be suitable to perform k means clustering.

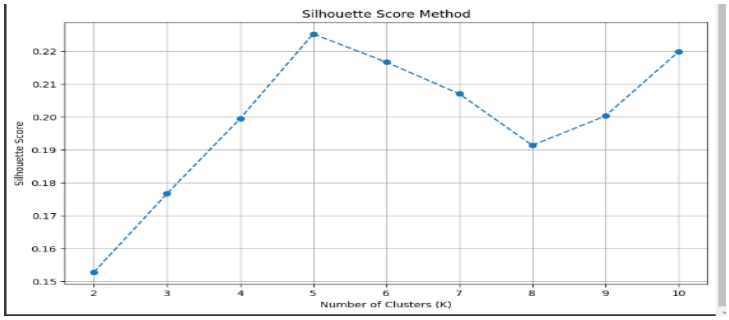

Figure 11 Silhouette Score method

The silhouette method is another method to check quality of clusters where silhouette score is put against number of clusters in order to visualise which cluster can perform well and can score more which means higher the score , more chances of that cluster to represent right amount of fatal police dataset.

After application of PCA and k means clustering to pre-processed data, we gained valuable insights into datasets which represents trends and relationship between each column helping us understanding more in depth of dataset. These techniques makes it easier for us to understand dataset more closely. Additional PCA reduced dimensionality of dataset making it more suitable for human understanding. Whilst k means clustering helped in knowing underlying trends in dataset though clustering same groups. Together these both techniques helped in providing comprehensive analysis of data as well as providing valuable insights which can help massively in decision making for law enforcement agencies.

8. Analysis of Results:

1. PCA:

For following PCA ahs been displayed with all of the principal components, in above diagram we discussed about just 2 of the principal components which were affecting majorly but fatal police shootings dataset as around 5 principal components.

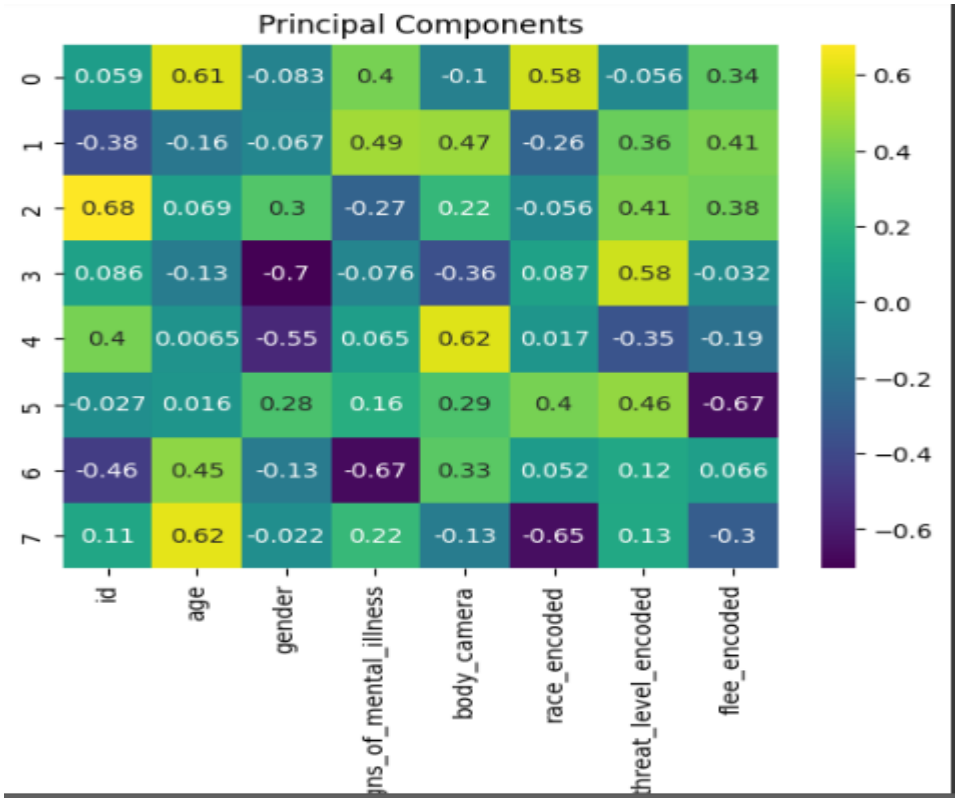

Figure 12 Visual representation of Principal Components

The heatmap visualisation interpret principal components as well as helps in understanding their relationships with original features in dataset. Principal components are shown as row and original features as column. Each cell of heatmap, presents coefficient of features for principal component. Colour scale represents positive coefficient, to negative coefficients. Features represented by large

positive or negative coefficients for principal component contribute significantly to that component. For example: PC1 has large positive coefficient for “age” feature which suggest that age plays important role in explaining variance captured by PC1. Similar coefficients can have relationships between them, either positive or negative. Furthermore, this heatmap provides explained variance ratio which provides comprehensive understanding of principal components as well as their relationship with original features.

2. Evaluating Quality of clusters (K-Means clustering):

```
Silhouette Scores: {2: 0.1528421704605312, 3: 0.17671703862796068, 4: 0.19950589557308282,
Optimal number of clusters (k): 5
Silhouette Score for optimal k: 0.22527850924821832
```

Figure 13 Cluster quality.

In the above data we have checked for the quality of clustering and the optimal number of clusters in order to make sure, 5 number so cluster that’s chosen is suitable to show the required information.

As the output shows, the optimal number of cluster are 5 and centroids are values for each clusters, these values represent mean of data points which is assigned to each cluster. By properly analysing centroids, clusters can be understood in depth. For example, if one of the clusters has high centroid value for “age” it may be trying to represent that older victims are represented or interconnected to that centroid. Silhouette score for 5 clusters shows higher score of. 0.22 which might not be an optimal but its higher than every other cluster score, proving that 5 number of clusters are good.

9. K-Means Clustering Analysis:

1. Relationship between race and threat level:

```
Race facing the highest threat level: threat_level
attack              Black
other               Asian
undetermined     Hispanic
dtype: object
Highest threat level by race: threat_level
attack          0.670654
other           0.428571
undetermined    0.058087
dtype: float64
```

Figure 14 Textual representation of threat level based on races.

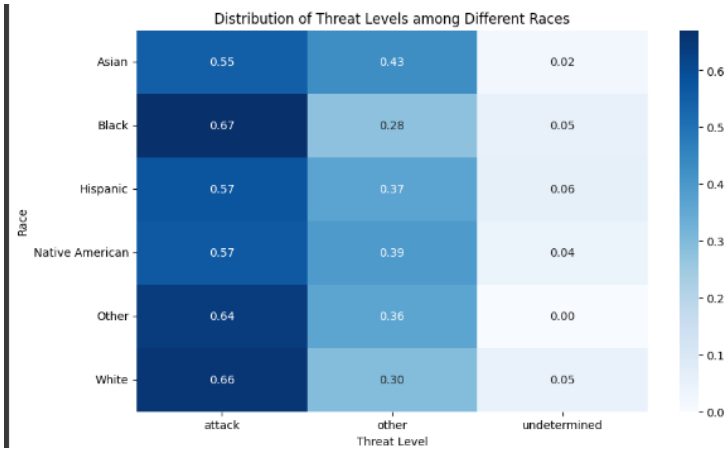

Figure 15 Representation of Threat level and races

Above heatmap and textual representation shows relationship between threats and races. By analysing these outputs, we can identify that black and Asian are at highest level. Black has attack level of around 0.67 which is pretty high. Abased on the heatmap, blue repeats higher and light blue represents lower, we have got three categories, threat level, other and determined. As the dark blue cell for black shows 0.67 which means a high proportion of black victims being assigned as “attack” threat level during their encounters with police. AS black, white, other has high difference from Asian and Hispanics, there might be some bias and if we compare attack to undetermined, it shows a pretty huge difference.

2. Relationship between Police shootings over each month:

It important to analyse the relation between how often police shootings occur based on the data we have as it might present if some duration of year is when crimes are risen specifically. S

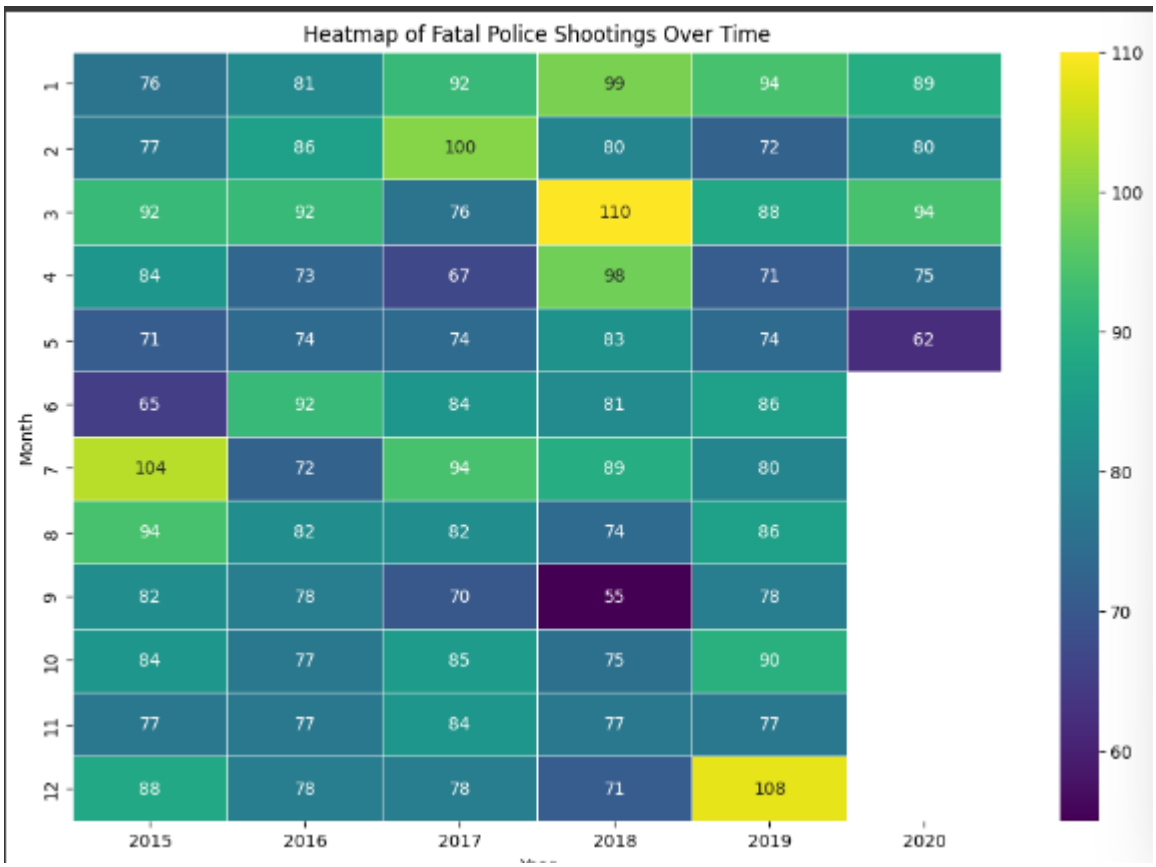

Figure 16 Fatal police shootings from 2015 to 2020

Above heatmap represents higher shooting over time as 110, which is quite a lot specially when duration is just a month. On looking closely, it can be easily determined, from 2017 to 2-019 more cells are in yellow colour showing higher police shootings. Specially in Jan 2017, 2018, 2019 shooting re almost same number, then almost every other cells has 80+ shootings ranges per month which is again quite a bit based on its monthly analysis. Although the dataset we have has date from 2015 to 2020’s may, it still shows quite a pattern that after every around 2-3 months the police shootings go up to 70-90 shootings per month.

3. Relationship between age of race and gender:

After doing analysis on races and their threat level, we can actually analyse if there's a relationship between their ages and gender as well. It can provide more in depth analysis to the demographics.

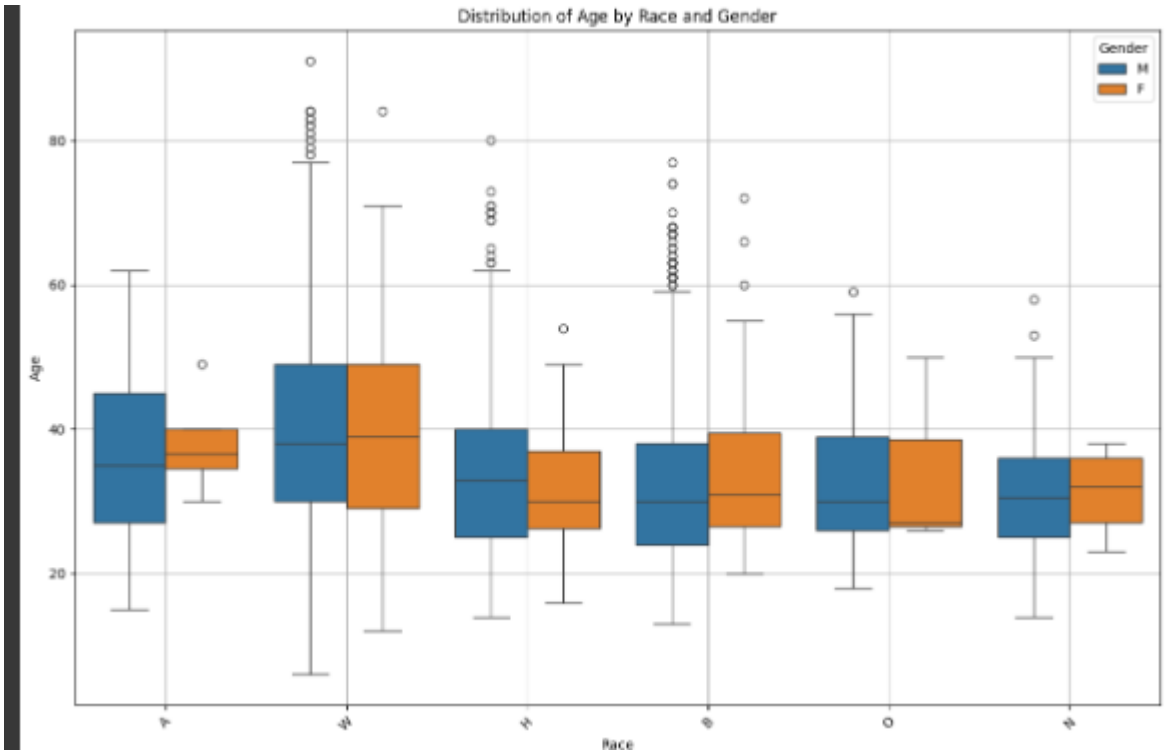

Figure 17 Box plot representing age group of genders

As in the above analysis, we have got two blue representing males and orange representing females., in this box plot, age range is show of each gender. On careful consideration, it seems like, white males and females are of almost similar age group approx. 25 to 45. If we look for Asian, its almost same as 25-45 males and around 35-40 age group of Asian females. For Hispanic, males re from 25-40 and females, 25-35. For black, other and native American its almost same as, 25-40 age range. This analysis represents white has highest age group from 25-45.

4. Analysis of mental illness:

Next analysis based on mental illness, in order of, what proportion of fatal shootings ahpend when the person showed menal illness symptoms.:

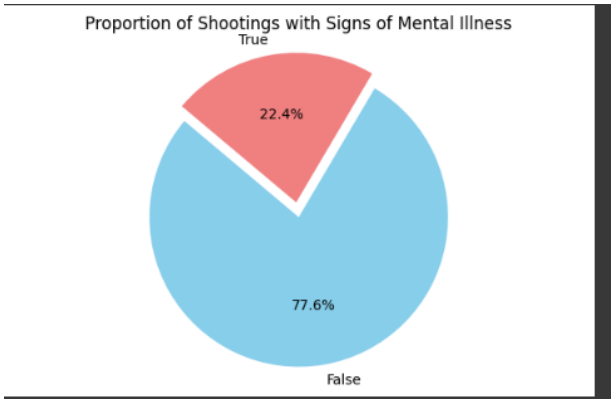

Figure 18 Mental illness person shot

This above pie chart shows even if the person showed mental illness's symptoms still they were shot. Around 22.4% people who had mental illness were shot. This si quite number but there might be scenarios when the criminal is danger to other peoples as well. So the police had to take action in order to keep public safe.

10. Evaluation:

Evaluation of PCA:

Evaluation of PCA was a crucial step on fatal police shootings dataset. In order to understand the effectiveness shown by pca's dimensionality reduction in dataset. Explained variance ratio represented proportion of variance to the original data from dataset which is captured by each principal component. Furthermore these values represented as explained variance ratio indicates as i.e.. PC1 captured 17.72% from total variance of data, PC2 captured 14.05% from variance of dataset , PC2 represented 13.15% of variance and so on.. This explained variance ratio helps in determining how many principal components are required to display variance of dataset I first place.

```
Explained Variance Ratio:
[0.1772184  0.14048909 0.13148064 0.12633318 0.12423755 0.11201187
 0.10201888 0.0862104 ]

Cumulative Variance Ratio:
[0.1772184  0.31770748 0.44918812 0.5755213  0.69975885 0.81177072
 0.9137896  1.        ]
```

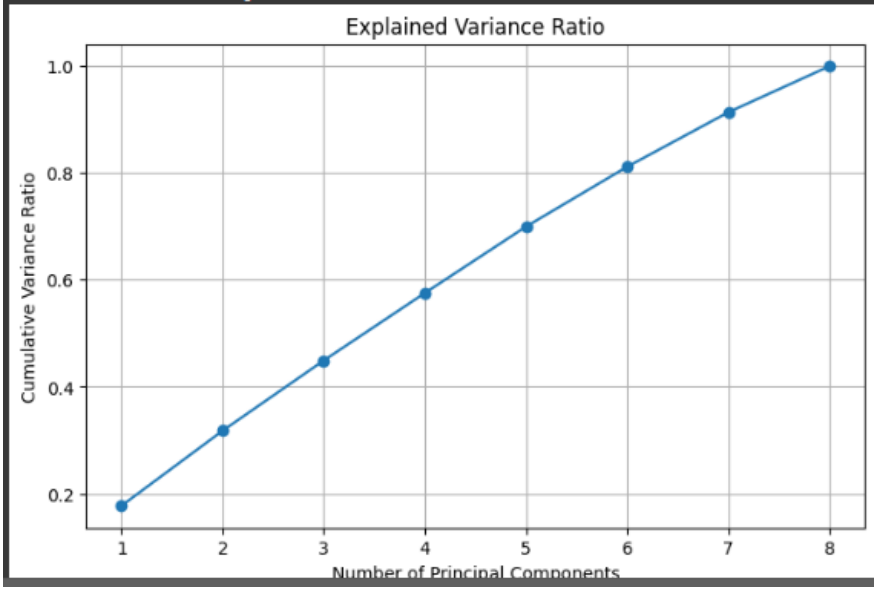

Figure 19 Explained Variance Ratio of Principal Component Analysis

Our main aim is to get as much as possible to show substantial amount of variance whilst also reducing dimensionality of dataset. Now, cumulative variance ratio represents cumulative sum of explained Ariane ratio up to each principal component. In context of PC1, it captured 17.72% of total variance, PC1 and PC2 captured 31.77% of variance and PC1, PC2, PC3all together captured 44.92% of variance and so on. Cumulative variance ratio determines how many principal components are needed to capture desired proportion of total variance.. for example, if we need to retain 80% of variance , then cumulative variance radio's first six principal components must be kept.

Above cumulative plot is helpful in determining number of principal components required. This plot would show slight curve and after some point that curve will start flattening out, ours was not much

clearer as it should have under cumulative variance ratio. But if we look closely, after curve 5 it shows a slight curve in the chart and we can say from his that 5 principal components will be suitable for visualising most of the Variance of dataset. After evaluating explained variance ratio, cumulative variance ratio and visualising explained variance curve, we can gain valuable insights which will be effective in capturing substantial amount of variance of dataset. This also helps in making decision about dimensionality reduction, features selection and further analysis of our dataset.

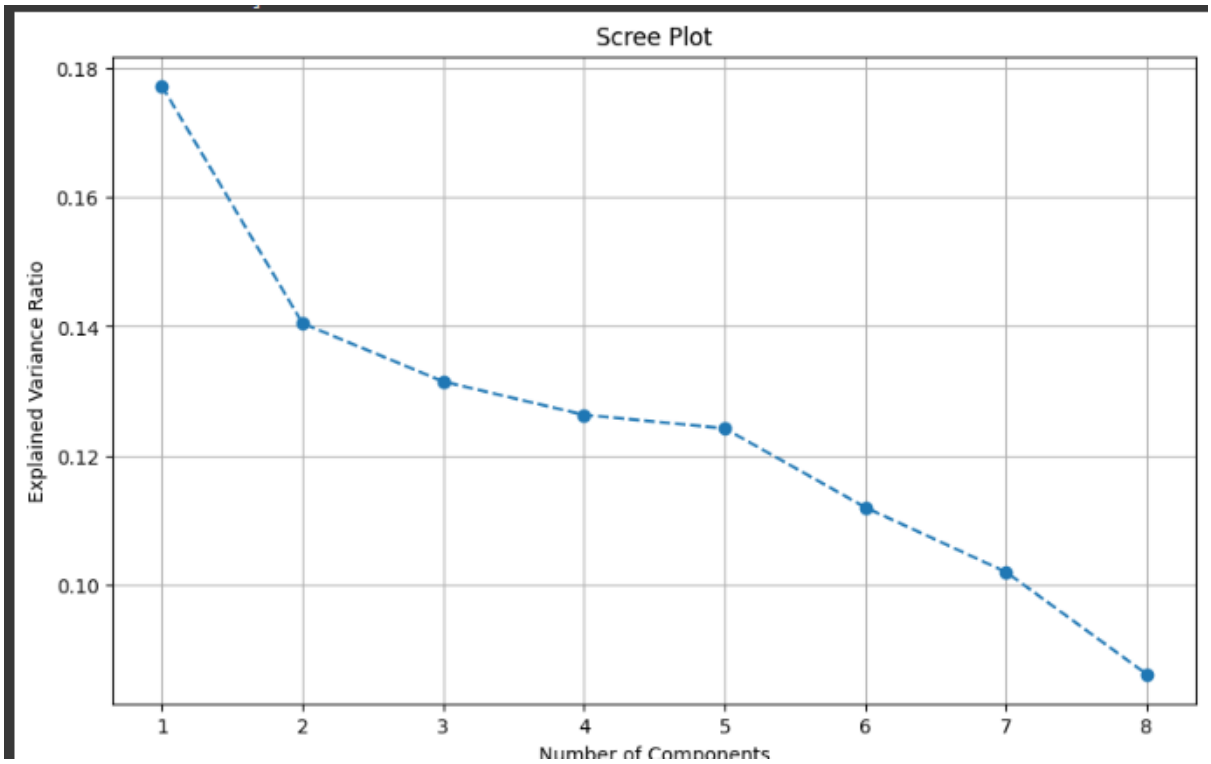

Figure 20 Scree plot

The explained variance plot showed curve but that was almost flat so just to be sure, we chose scree plot. Screen plot is graphical representation of explained variance ratio fir principal component. It would determine optimal number of principal components that can represent variance of dataset. It would create an “elbow shape” and this elbow point would help to determine how many principal components are required for our analysis. The above scree plot, shows two elbow points , one represents 2 principal components are good and another shows 5 principal components. We decided to choose 5 principal components as it would cover large range of variance also cumulative variance chart represented slight curve at around 5 so, we have chosen 5 principal components to represent our variance of original dataset. After combination of evaluation techniques, a comprehensive understanding, PCA can capture essential trends and patterns within fatal police shootings dataset.

Evaluation of K means clustering:

Evaluation of K means clustering on this dataset has been done using silhouette score So, silhouette score acts as score which measures how well data points fits into the assigned cluster as compared to another clusters. The silhouette score ranges somewhere from -1 to 1. Higher the value, more defined clusters. The evaluation was done s number of clusters which were started from 2 to 10. For each cluster same methods were applied, optimal must have higher score than rest of the clusters.

When looked for 5 it displayed result as 0.22527… So according to silhouette score, the clusters were divide into 5 and each cluster was reasonable similar to its original data value.

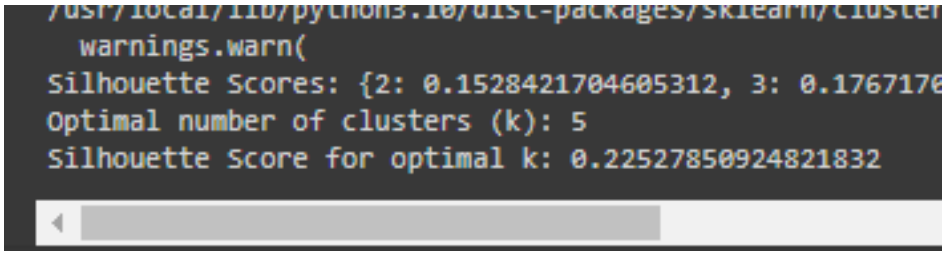

Figure 21 Silhouette score.

As the above data show 0.22 that represents as reasonably good clustering quality on fatal police shootings dataset. The silhouette score ranges from -1 to 1 where 1 shows a positive number where there's no overlapping between clusters and they are all perfectly clustered... 0 mean, data points are very close as well as overlapping clusters are visible. -1 shows, data points are likely being assigned to wrong clusters. Based on al these our score is reasonably good. Furthermore, after k means clustering's silhouette score, and analysing cluster characteristics, researchers and policy makers can understand trends or patterns deeply for the fatal police shootings dataset.

11. Conclusion:

The main goal of this analysis was to understand properly about the demographic factors which might be associated with fatal police shootings I the United States. In order to complete this, we conducted well detailed analysis on fatal police shootings dataset this dataset was from 2015 to 2020., which mean the dataset is reasonably good to perform analysis on. For beginning of our analysis, we pre-processed dataset in order to produce as recipe analysis as possible, because if pre-processing step was missed, that could have given bias based results. We realised that date has high dimensionality so, in order to lessen that dimensionality, we went with PCA, which reasonably reduced the dimensionality of our dataset. Furthermore, it revealed 5 principal components and out of them we also showed 2's scatter plot After that, analysis needs to perform and we chose k means clustering, as its suitable to perform clusters and show the relationship between 2 variables, in beginning of analysis k= 5 was the number of clusters provided by scree plot and elbow method.

After having k's value, we began analysis of relationship between race and threat level, on examination, showed us black as the highest threat level perceived and assigned during fatal police shooting encounters. Second analysis showed relationship between police shootings over each month, after analysing, approx. shooting over month by fata police ranges from 70-90 which is almost higher and so on.

Throughout these analyses, we realised several findings and trends. For example, higher threat level was for black race which also needs further investigation order to promote fair justice to everyone. Furthermore, fair justice in country would make police departments and public trust stronger, helping in laying a foundation of respectful and generous society.

12. Reference:

1. The Washington Post. (2024). Police shootings database 2015-2024: Search by race, age, department.
2. Fatal Encounters. (n.d.). Retrieved May 10, 2024
3. Ar, R., Mitra, S., & Umesh, A. C. (2018). Data Mining: Employee Turnover in a Company. IJARCCE.
4. The Washington Post. (2024). Police shootings database 2015-2024: Search by race, age, department.
5. Nix, J., Pickett, J. T., Wolfe, S. E., & Campbell, B. A. (2017). Demeanour, race, and police perceptions of procedural justice: Evidence from two randomized experiments. The Journal of Politics, 79(4), 1154-1169.
6. Legewie, J., & Fagan, J. (2019). Aggressive policing and the educational performance of minority youth. American Sociological Review, 84(2), 220-247.
7. Goff, P. A., Obermark, D., La Vigne, N., Yahner, J., & Geller, A. (2016). The science of justice: Race, arrests, and police use of force. Centre for Policing Equity
8. Goff, P. A., & Kahn, K. B. (2012). Racial bias in policing: Why we know less than we should. Social Issues and Policy Review, 6(1), 177-210.
9. Fryer, R.G. (2016). An Empirical Analysis of Racial Differences in Police Use of Force. Harvard University.
10. Bor, J., Venkataramani, A.S., Williams, D.R. and Tsai, A.C. (2018). Police killings and their spillover effects on the mental health of black Americans: a population-based, quasi-experimental study. The Lancet, 392(10144), pp.302-310.
11. Kaggle.com