****

**PROJECT REPORT:**

**ASSOCIATION MINNING AND CLUSTERING**

**PROJECT REPORT BY:**

**TUSHAR JINDAL**

**ROLL NUMBER: 18CSU218**

**INDEX**

|  |  |  |
| --- | --- | --- |
| S.NO | TOPIC | PAGE NUMBER |
| 1 | Apriori- Objective | 3-4 |
| 2 | Data Set Description | 5-6 |
| 3 | Rule Mining Process | 6 |
| 4 | Resulting Rules | 7 |
| 5 | Result | 8 |
| 6 | K-means Clustering- Objective | 8 |
| 7 | Process | 9 |
| 8 | K-Means Result | 10-11 |
| 9 | Conclusion | 11 |

**APRIORI ALGORITHM**

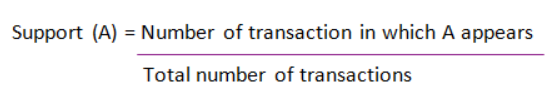
**1. OBJECTIVE**

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an itemset occurs in a transaction. Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

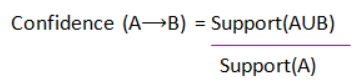
**Apriori**is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database

Important Terms:

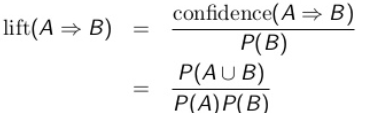
1. Minimum support: Support is an indication of how frequently the itemset appears in the dataset.



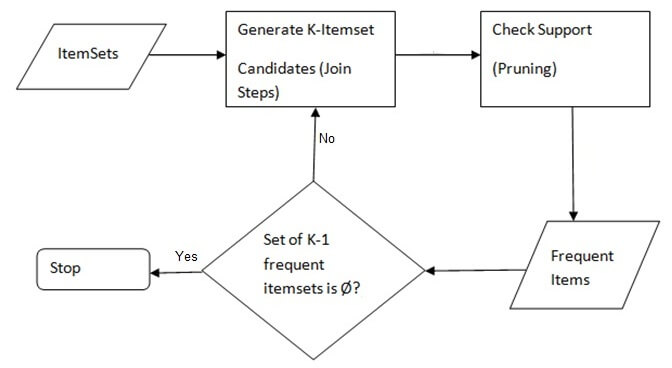
1. Confidence: confidence constraint is applied to these frequent itemsets in order to form rules.



1. Lift: This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is.

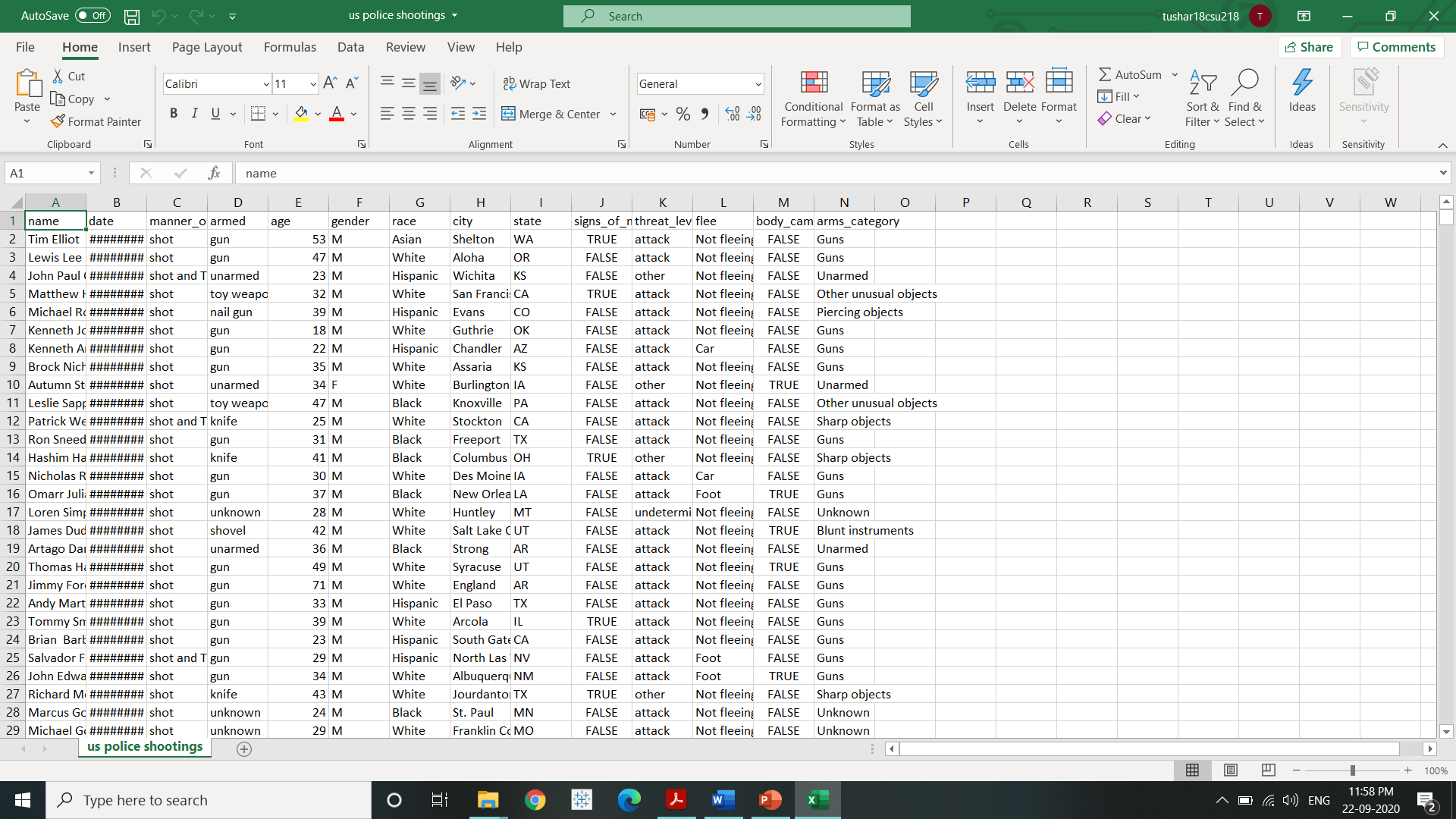


**Flow Diagram for Apriori algorithm:**



1. **DATA SET DESCRIPTION**

The data set I have chosen for the project is the **US police shooting dataset.** The reason for choosing this dataset was to address the recent growing concerns about police shooting people of a certain community without required permissions. This gave rise to a lot of social concerns about the growing racism shown by the US police. I have tried to find patterns in the shooting incidents and tried to verify if this is true or not.



My Dataset contains 4895 rows and 14 columns.

**Data pre-processing and reasons:**

1. **Dropping the unnecessary columns**: I have Dropped the columns 'name', 'date', 'body\_camera', 'signs\_of\_mental\_illness', 'city' and 'state' as values of name and date were not frequent and columns like 'body\_camera', 'signs\_of\_mental\_illness', 'city' and 'state' were not required according to the objective of my project.
2. **Checking of missing values:** I checked if my dataset contained any missing values by which I found that there were no missing values in my dataset.
3. **RULE MINING PROCESS**

**Parameter setting:**

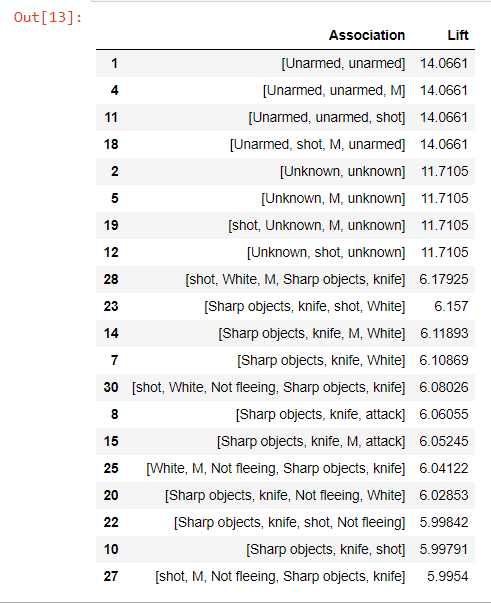
1. Minimum length: we want at least 3 items to be associated. No point in having a single or two items in the result.
2. Minimum lift: Minimum of 3 (less than that is too low)
3. Minimum support: 0.05 (Randomly taken)
4. Minimum confidence: At least 60%

**Choice of algorithm:**

I have applied apriori to find the frequent patterns in the shooting incidents and also applied K-means Clustering to try to verify if both are showing same results.

**4. RESULTING RULES**

1. General description: The reason for choosing this dataset was to address the recent growing concerns about police shooting people of a certain community without required permissions. This gave rise to a lot of social concerns about the growing racism shown by the US police. I have tried to find patterns in the shooting incidents and tried to verify if this is true or not.
2. Number of Rules: 33
3. Selection Of rules for client: Top 20 rules having highest lift.



**5. RESULTS**

By the study, police does not seem to target any particular community therefore showing no sign of racism. Many people shot were unarmed or very few of the suspects had any kind of sharp objects like a knife.

**K-MEANS CLUSTERING**

K-means is a simple unsupervised machine learning algorithm that groups data into a specified number (k) of clusters.

**Objective:**

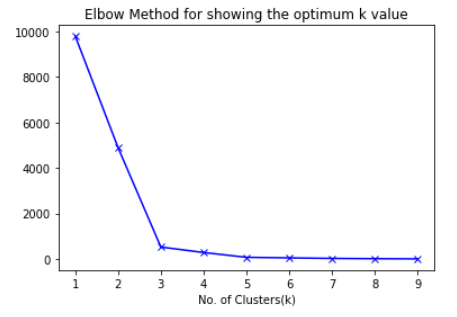
I have made clusters using attributes “manner of death” and “race” to verify the results of apriori algorithm.

**Data pre-processing:**

1. Dropping the columns: I have made clusters using attributes “manner of death” and “race” because I wanted to check if any particular race is being targeted showing the sign of racism and compare it with the apriori algorithm.
2. Checking for missing values: My dataset does not contain any missing value.
3. Encoding: I have encoded attributes “manner of death” and “race” in order to convert them from categorical to numerical values so that clustering can be applied to them.
4. Scaling data: It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

**K-MEANS CLUSTERING PROCESS:**

1. Elbow method: Implements the elbow method for determining the optimal number of clusters. Choose a number of clusters that covers most of the variance.



Since we are getting elbow at k=3 therefore number of clusters for this project is 3.

1. Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select ‘c’ cluster centers. (Elbow method)

2) Calculate the distance between each data point and cluster centers.

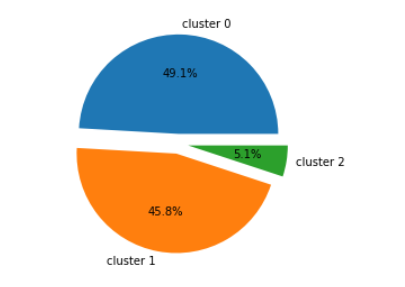
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using new centers

5) Repeat the same steps till the centers of clusters do not change.

**K-MEANS CLUSTERING RESULTS**

1. After performing clustering on the data, the data was divided into 3 clusters. Here is the pie chart visualising clusters telling what percentage of the data has been assigned to that cluster.

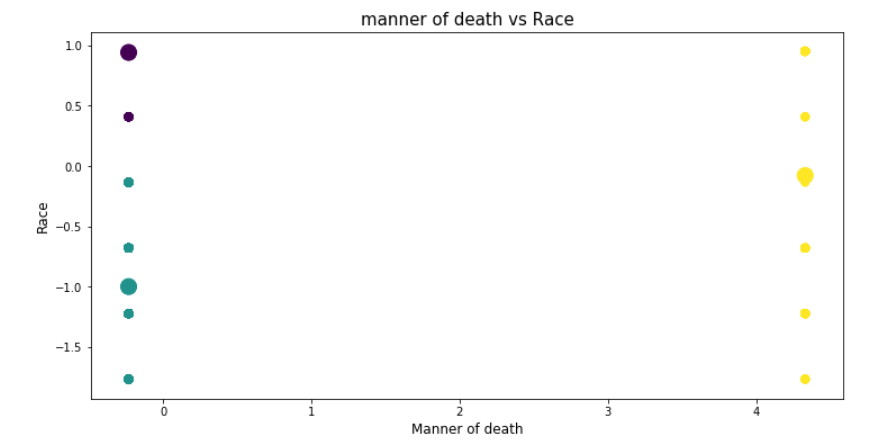


1. cluster 0 consists of 49.1% of data

2. cluster 1 consists of 45.8% of data

3. cluster 2 consists of 5.1% of data

1. Plotting Scatter plot to show the three clusters with their respective centres.



**All three clusters were of mixed “race” which shows that there was no selected targeting of people of particular community.**

**CONCLUSION**

Both Apriori and k-means clustering algorithm verified that that was no racism shown by US police. They are not targeting any particular community.