Clustering
Data Warehousing and Data Mining
Project Phase 3

Jayitha .C - 20171401
Thapaswini Chowdary - 20161066

September 29, 2018

Abstract

In the third phase, we have attempted to study the homicide dataset by finding clusters. To do this we have used the PAM algorithm (Partitioning around Meteiods) to cluster and computed silhouette coefficients to determine how many clusters to finally select. The tools we used to perform these tasks are R and it's packages clustering(to cluster), dplyr(to find Gower distance), ggplot2(To plot sillhoute distance), readr(to read from the csv file) and finally Rtsne(to plot the cluster on a 2D surface)

1 Data Set Description

We have chosen to use the homicide data set which is what we used for the Phase 1 and Phase 2 submission as well. We chose this data set because it is easy to understand, the number of attributes aren't overwhelming but just enough to enable us to run some fun queries to yield interesting results.

s

Table 1: Homicide Dataset Post Data Cleaning

Year	Month	Day	Race	Age	Sex	City	State	Latitude	Longitude	Disposition
2007	1	1	Asian	10	Female	Albuquerque	AL	00.00	000.00	Closed by Arrest
2008	2	2	Black	20	Male	Atlanta	AZ	25.73	-122.51	Closed without Arrest
2009	3	3	Hispanic	30	S_Unknown	Baltimore	CA	25.74	-122.5	Open/ No Arrest
2010	4	.	White	40		.	CO	25.75	-122.49	
2011	5	.	Other	50		.	DC	.	.	
2012	6	.	R_Unknown	60		.	.	.	.	
2013	7	.		70		.	.	.	.	
2014	8	.		80		.	.	.	.	
2015	9	.		90		Stockton	.	45.03	.	
2016	10	29		100		Tampa	TX	45.04	-71.05	
2017	11	30		100		Tulsa	VA	45.05	-71.04	
	12	31		-500		Washington	WI	45.06	-71.02	

1.1 Data Prepossessing

To get better results we have performed the following procedures on the dataset

1.1.1 Data Cleaning

- Date - the date which was originally a composite attribute, has not been broken down to three new attributes - year, month and date. This will help us get aggregated results that are a grouping of one of the three attributes
- Latitude and Longitude - We have rounded off the latitude and longitudes upto two places after the decimal point to yield better cleaner results and get faster computation speeds
- Unknowns - we have not cleaned up the unknowns as we feel that cleaning them up by any known procedures such as replacing the unknown by mode or median might falsely affect the data. Hence the assumption we are making is that the distribution within the unknowns is the same as that of the distribution we obtain by removing all transaction with the unknown values
- Names - We have stripped the names from the dataset. We might be missing out on interesting results such as a series of death of people with the same name, but this pattern is highly unlikely.
- Unique Identifier - We have stripped this also from the dataset as the only purpose this attribute serves is to uniquely identify a record. It will not affect the final results in any way whatsoever.
- Ambiguity - To account for distances and to avoid ambiguity in the face of multiple attributes having the same domain we have modified the data set slightly as shown in Table 2
  Note: No unknown values in Date, Disposition, City or State

Table 2: Augmented Data

Attribute	Value	Augmented Value
Race	Unknown	R_Unknown
Age	Unknown	-500
Sex	Unknown	S_Unknown
Latitude	Unknown	00.00
Longitude	Unknown	000.00

1.1.2 Data Transformation

The data set we have chosen has majorly 2 kinds of attributes

- Numerical - Age, Latitude, Longitude, date and
- Nominal - Race, Sex, Disposition, city and state

Together this dataset is said to be a **mixed** dataset. While the weka took transforms Nominal to Binary and then continues to find distances using Eucledian Distance, we have attempted no such thing. Instead we have used Gower distance where the distance between categorical data is 1 if they're different and 0 if they aren't.

1.1.3 Data Sampling

The tool we used wasn't equipped to handle such a huge dataset, so we had to sample the dataset to a set that the tool could process and that still completely represented the dataset. So we sorted the datasets in the order of those attributes whose domain was the smallest and the proceeded to pick every other row. Therefore we sorted the attributes in the following order - Sex, Disposition, Race, Year, Month, Date, Age, State, City, Latitude, Longitude. Then we proceeded to take all even records. So we only operate on half the dataset.

2 Algorithms and Tools Used

We do not claim to have written the packages or created the tools that we have used in this assignment.

2.1 Distance Measure Used - Gower Distance

Distance is a numerical measurement of how far apart two objects are, i.e. a metrics used to measure proximity or similarity across individuals. The distance measure we have chosen for this assignment is **Gower Distance**. We have chosen this distance as it is convenient for mixed datasets such as ours. The formula to find Gower distance is shown in Fig1.

$$d(i,j) = \frac{1}{p}\sum_{i=1}^{p} d_{ij}^{(f)}$$

Gower distance's formula

Figure 1: Gower Distance Formula

- Gower distance is computed as the average of partial dissimilarities across individuals. Each partial dissimilarity (and thus Gower distance) ranges in [0 1].

Partial dissimilarities computation depend on the type of variable being evaluated. This implies that a particular standardization will be applied to each feature, and the distance between two individuals is the average of all feature-specific distances.

- For a numerical features, partial dissimilarity is the ratio between the absolute difference of observations and the maximum range observed from all individuals. In simple terms it is the Manhattan Distance normalized using the range. The formula is shown in Fig2.
- For a qualitative(nominal) feature, partial dissimilarity equals 1 only if observations have different values. Zero otherwise.

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f}$$

Figure 2: Partial dissimilarity computation for numerical features

Although we haven't done so in this assignment it might be interesting to see how cluster formation would take place if we used a weighted average of all the distances instead of average, thereby introducing the concept of importance to a few attributes over the others.

2.2 Clustering Algorithm - PAM

The PAM (Partitioning Around Mediods) algorithm is very similar to K-means, mostly because both are partitional algorithms, in other words, both break the dataset into groups (clusters), and both work by trying to minimize the error, but PAM works with Medoids, that are an entity of the dataset that represent the group in which it is inserted, and K-means works with Centroids, that are artificially created entity that represent its cluster.The PAM algorithm partitions the dataset of n objects into k clusters, where both the dataset and the number k is an input of the algorithm. This algorithm works with a matrix of dissimilarity, whose goal is to minimize the overall dissimilarity between the representants of each cluster and its members.
In simple words PAM is one algorithm to find a local minimum for the k-medoids problem. Maybe not the optimum, but faster than exhaustive search.Fig 3 shows the pseudo code for PAM.

2.3 Assess Consistency Within Clusters - Sillhoute Coefficient

The silhouette coefficient contrasts the average distance to elements in the same cluster with the average distance to elements in other clusters. Objects with a high silhouette value are considered well clustered, objects with a low value may be outliers. This index works well with k-Medoids clustering, and is also used to determine the optimal number of clusters.

2.4 Tools Used

Below are the details of the various tools we used to perform clustering on the dataset

- Libre Office Calc - The dataset is stored in a csv file, all data cleaning work was one using sheet operations
- readr* - The function csv.read which is used to read from the csv file is available with this R package
- dplyr* - This package was used to clean data and to compute the Gowed Distance using the daisy() function
- clustering* - This package contained all the clustering algorithms implemented in R, all out of which we chose to use PAM
- Rtsne* - Used to plot the clusters once obtained
- ggplot2* - To plot the graphs after Sillhoute coefficients computation

* - R Packages

Algorithm 1: PAM Algorithm Input: E = {e1,e2,...en} (dataset to be clustered or matrix of dissimilarity)

k (number of clusters)
metric (kind of metric to use on dissimilarity matrix)
diss (flag indicating that E is the matrix of dissimilarity or not)

Output: M = {m1,m2,...,mk} (vector of clusters medoids)

L = {l(e) | e = 1,2,...,n} (set of cluster labels of E)

foreach $m_i \in M$ **do**

$m_i \leftarrow e_j \in E$; (e.g. random selection)

end if diss ≠ true

Dissimilarity ← CalculateDissimilarityMatrix(E, metric);

else

Dissimilarity ← E;

end repeat

foreach $e_i \in E$ **do**

$l(e_i) \leftarrow$ argminDissimilarity(e_i, Dissimilarity, M);

end

changed ← false;

foreach $m_i \in M$ **do**

Mtmp ← SelectBestClusterMedoids(E, Dissimilarity, L);

end

if Mtmp ≠ M

M ← Mtmp;

changed ← true;

end

until changed = true;

Figure 3: Pseudo Code for PAM

3 Results

3.1 Sillhoute Coefficients

The Sillhoute Coefficients were computed for the number of Clusters $k = 2$ till $k = 10$ and the graph shown in Fig4 was obtained.

From the graph we can see that the optimal number of clusters to be considered to get valuable results is **3** (followed by 10 and so on).

This measure tells us that when $k = 3$ the 3 clusters are most ẗogether; i.e. the distance between elements in that cluster is small.

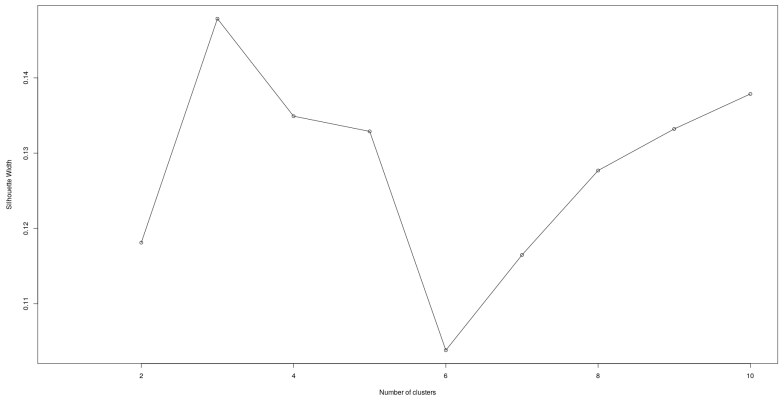

Figure 4: Sillhoute Width vs Number of Clusters

3.2 Clusters

The following 3 clusters were obtained.

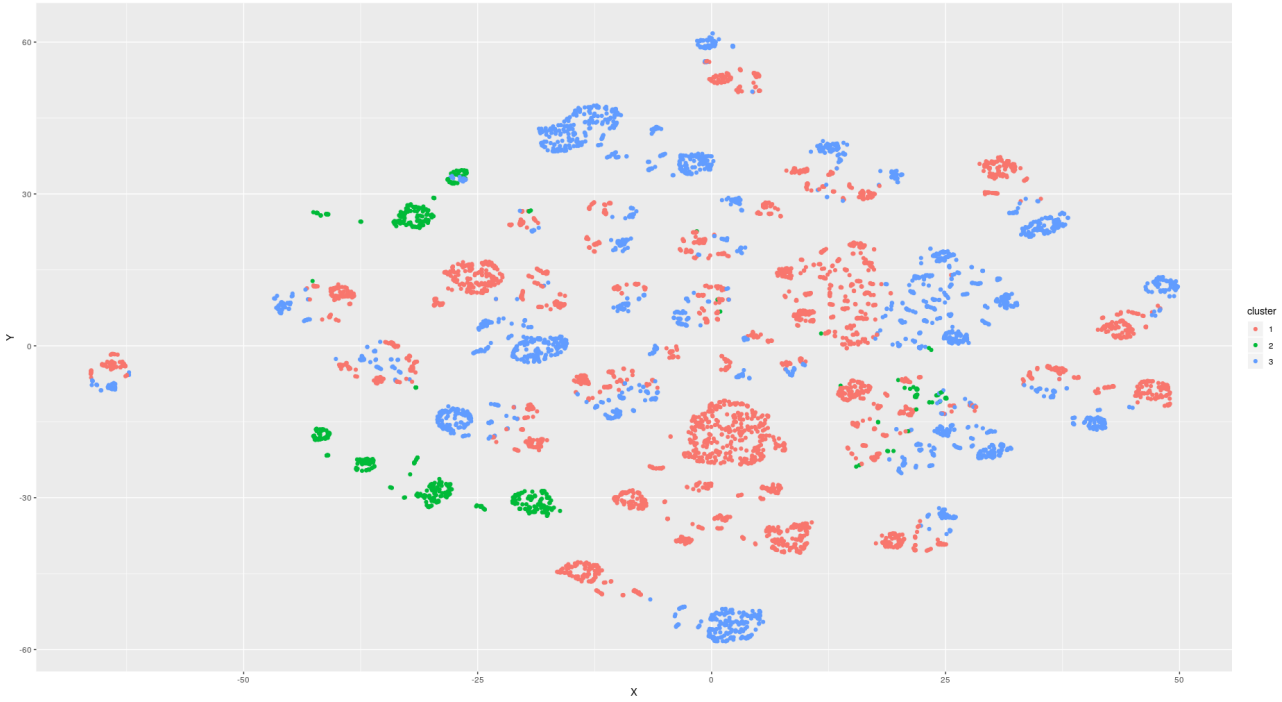

Figure 5: Cluster Visualization

Fig 6, 7 and 8 show brief descriptions of the clusters. Fig 9 and 10 show which 2 objects are closest and farthest respectively.

```
[[1]]
      Year          Month            Date             Race
 Min.   :2007   Min.   : 1.000   Min.   : 1.00   Asian    :  56
 1st Qu.:2009   1st Qu.: 4.000   1st Qu.: 8.00   Black    :2280
 Median :2012   Median : 7.000   Median :15.00   Hispanic : 503
 Mean   :2012   Mean   : 6.556   Mean   :15.16   Other    :  47
 3rd Qu.:2015   3rd Qu.: 9.000   3rd Qu.:23.00   R_Unknown:   4
 Max.   :2017   Max.   :12.000   Max.   :31.00   White    : 520

      Age                Sex                City          State
 Min.   :-500.00   Female   : 584   Chicago     : 710   IL     : 710
 1st Qu.:  21.00   Male     :2821   Philadelphia: 224   CA     : 397
 Median :  28.00   S_Unknown:   5   Houston     : 161   PA     : 263
 Mean   :  28.46                    Detroit     : 131   TX     : 259
 3rd Qu.:  39.00                    Los Angeles : 125   TN     : 176
 Max.   :  97.00                    Las Vegas   : 121   MI     : 131
                                    (Other)     :1938   (Other):1474
    Latitude       Longitude                    Disposition      cluster
 Min.   : 0.00   Min.   :-122.48   Closed by arrest     :2783   Min.   :1
 1st Qu.:35.06   1st Qu.: -95.46   Closed without arrest: 171   1st Qu.:1
 Median :38.86   Median : -87.68   Open/No arrest       : 456   Median :1
 Mean   :37.71   Mean   : -91.09                               Mean   :1
 3rd Qu.:41.76   3rd Qu.: -83.12                               3rd Qu.:1
 Max.   :45.05   Max.   :   0.00                               Max.   :1
```

Figure 6: Cluster 1 Description

```
[[2]]
      Year          Month            Date             Race
 Min.   :2007   Min.   : 1.000   Min.   : 1.00   Asian    :  5
 1st Qu.:2009   1st Qu.: 4.000   1st Qu.: 8.00   Black    :  6
 Median :2012   Median : 7.000   Median :16.00   Hispanic : 22
 Mean   :2012   Mean   : 6.618   Mean   :16.25   Other    :  4
 3rd Qu.:2015   3rd Qu.:10.000   3rd Qu.:24.00   R_Unknown:507
 Max.   :2017   Max.   :12.000   Max.   :31.00   White    : 22

      Age               Sex              City         State
 Min.   :-500.0   Female   : 50   Dallas     :197   TX     :243
 1st Qu.:-500.0   Male     :  8   Kansas City:136   MO     :136
 Median :-500.0   S_Unknown:508   Phoenix    :109   AZ     :109
 Mean   :-290.1                   Houston    : 33   FL     : 32
 3rd Qu.:  25.0                   Miami      : 32   NM     : 14
 Max.   :  86.0                   Albuquerque: 14   NV     : 11
                                  (Other)    : 45   (Other): 21
    Latitude       Longitude                    Disposition     cluster
 Min.   :25.76   Min.   :-118.42   Closed by arrest     :307   Min.   :2
 1st Qu.:32.73   1st Qu.: -98.56   Closed without arrest: 56   1st Qu.:2
 Median :33.37   Median : -96.75   Open/No arrest       :203   Median :2
 Mean   :34.00   Mean   : -98.57                               Mean   :2
 3rd Qu.:36.19   3rd Qu.: -94.57                               3rd Qu.:2
 Max.   :42.42   Max.   : -77.44                               Max.   :2
```

Figure 7: Cluster 2 Description

```
[[3]]
      Year          Month            Date             Race
 Min.   :2007   Min.   : 1.000   Min.   : 1.00   Asian    :  26
 1st Qu.:2010   1st Qu.: 4.000   1st Qu.: 9.00   Black    :1882
 Median :2013   Median : 7.000   Median :17.00   Hispanic : 339
 Mean   :2013   Mean   : 6.793   Mean   :16.36   Other    :  35
 3rd Qu.:2015   3rd Qu.:10.000   3rd Qu.:24.00   R_Unknown:  16
 Max.   :2017   Max.   :12.000   Max.   :31.00   White    : 248

      Age                Sex                City          State
 Min.   :-500.00   Female   : 267   Baltimore   : 355   CA     : 386
 1st Qu.:  23.00   Male     :2263   Detroit     : 177   MD     : 355
 Median :  29.00   S_Unknown:  16   Philadelphia: 168   TX     : 213
 Mean   :  30.13                    Houston     : 149   PA     : 201
 3rd Qu.:  39.00                    Los Angeles : 136   MI     : 177
 Max.   :  96.00                    New Orleans : 118   LA     : 135
                                    (Other)     :1443   (Other):1079
    Latitude       Longitude                    Disposition      cluster
 Min.   : 0.00   Min.   :-122.46   Closed by arrest     : 118   Min.   :3
 1st Qu.:33.94   1st Qu.: -95.53   Closed without arrest: 138   1st Qu.:3
 Median :38.65   Median : -86.02   Open/No arrest       :2290   Median :3
 Mean   :36.84   Mean   : -90.02                               Mean   :3
 3rd Qu.:39.89   3rd Qu.: -77.00                               3rd Qu.:3
 Max.   :45.05   Max.   :   0.00                               Max.   :3
```

Figure 8: Cluster 3 Description

3.2.1 Interesting Observations

- Unknowns: The clustering algorithm as somehow managed to cluster all objects that have some unknown values. This result is very interesting and useful. This cluster basically tells us that this set of objects haven't been documented well

- Disposition: We can also see an underlying theme where the clustering has occur ed based on Disposition, that is, most open/no arrest cases have been clustered into 3 and the closed by arrest ones have been clustered under 2

- Location: Although both clusters 1 & 3 have similar locations, cluster 2 is very peculiar as it seems to have clustered all objects whose city is towards the southern part of USA as shown in the map (Fig 9). This goes to show that all the states in the southern part of USA aren't very good at keeping records, this could indicate some kind of anomaly that might need to get checked out

- Black Race: From clusters 1 & 3 it is astounding to see that approximately half the black cases have been closed with arrest and the other half left open.

- Latitude and Longitude: Another peculiar fact about cluster 2 is that although it has themed itself to be the ünknowncluster, it has no objects that have unknown values of latitude and longitude, the unknown latitude and longitude objects have been clustered together with clusters 1 & 3. This could be attributed to the fact that the clusters have been separated also based on location and hence due to the nature of the numeric

values given to unknown latitudes and longitudes, this kind of clustering may have occurred

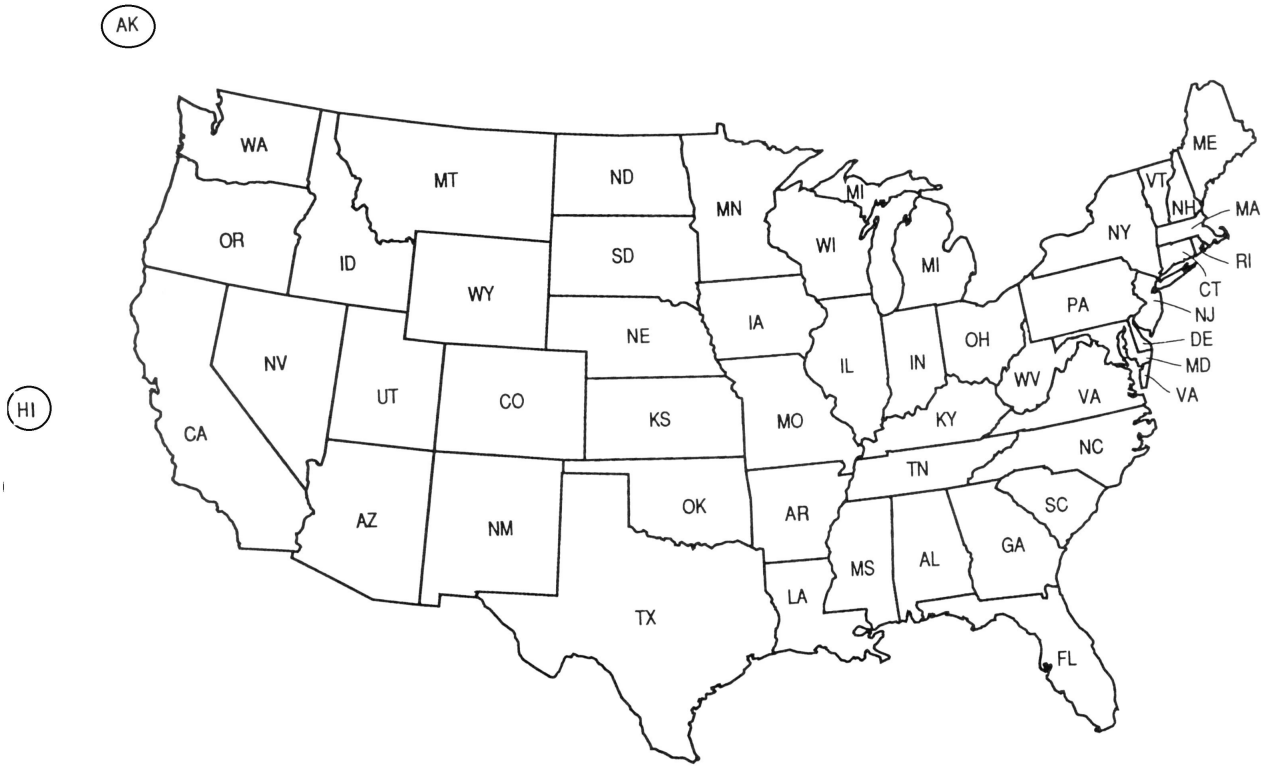

Figure 9: States Map of USA

3.3 Distance Characterization

According to the distance function we've chosen the two c̈loseständ the two f̈arthestöbjects are as shown in Fig 10 & 11 respectively

	Year	Month	Date	Race	Age	Sex	City	State	Latitude	Longitude	Disposition
6301	2017	10	1	R_Unknown	36	S_Unknown	Las Vegas	NV	36.1	-115.18	Closed without arrest
6300	2017	10	1	R_Unknown	33	S_Unknown	Las Vegas	NV	36.1	-115.18	Closed without arrest

Figure 10: Most Similar Objects

	Year	Month	Date	Race	Age	Sex	City	State	Latitude	Longitude	Disposition
6446	2014	12	31	R_Unknown	-500	S_Unknown	Phoenix	AZ	33.45	-112.2	Open/No arrest
976	2007	4	3	Black	31	Male	St. Louis	MO	0.00	0.0	Closed by arrest

Figure 11: Most Dissimilar Objects

From the above result it is clear that in some sense clustering looks at an unknown record as some sort of outlier (not to clusters itself but to the dataset as the distance between any record and an unknown record is high) and that the record with the the black male is typical in some sense. This isn't technically correct, but it does indeed capture the essence of the dataset.

4 Conclusion

In conclusion, we have managed to use a clustering algorithm on our mixed dataset to yield very interesting results out of which the separation of unknown values from all the other was the most resounding observation. We also gained some insights into the data set but we proceed to conclude that clustering wasn't as useful as other techniques such as association rule mining in deriving new information, even if it did in some sense try to segregate data. We might be able to produce some more results if maybe we used a weighted distance function.