Assignment

Car Evaluation Dataset

The first dataset we have chosen is “Car Evaluation” where each record will consist of 6 attributes which are the car’s buying price, maintenance price, number of doors, person capacity, lug boot size and estimated safety respectively. The records also contain a classifying attribute which states the acceptability of the car.

The attributes are:

- buying: vhigh (very high), high, med, low.
- maint: vhigh, high, med, low.
- doors: 2, 3, 4, 5,more.
- persons: 2, 4, more.
- lug_boot: small, med, big.
- safety: low, med, high

The classifying attribute are unacc (unacceptable), acc (acceptable), good, vgood (very good)

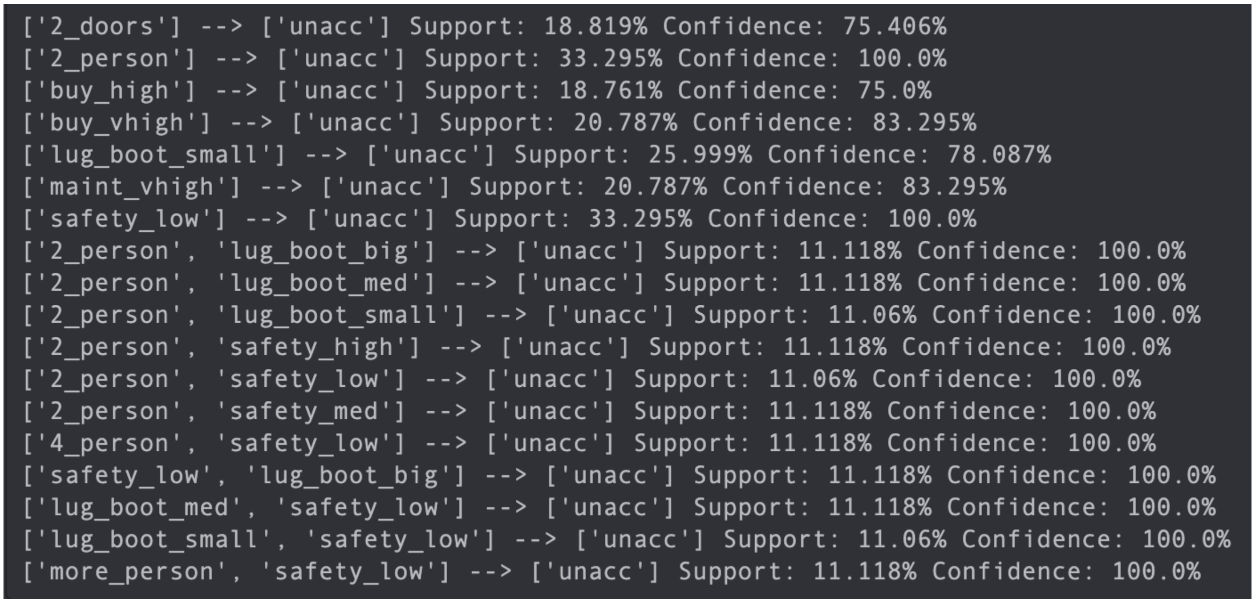
The number of instances for our data set is 1,728. The number of attributes in our dataset is 7 attributes in total with 1 being the classifying attribute.

Spotify Top 50 Songs 2021 Dataset

The second dataset we have chose is “Spotify Top 50 Songs 2021” where each record consists of 14 attributes which are popularity, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration of the song (in milliseconds) and an estimate of the song’s time signature. There are also 4 classifying attributes which are id (position of the song on list), artist name, track name and track id.

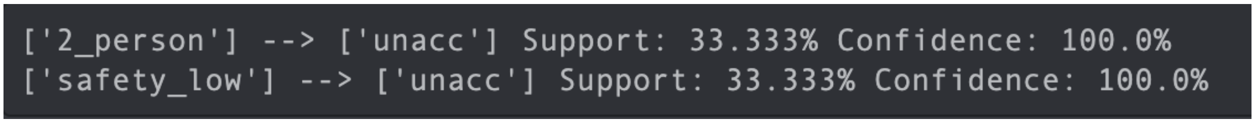
Association Analysis

The dataset used to do the association analysis was the “Car Evaluation” data. We used Apriori rules to perform the analysis and initially set the minimum support and confidence to 0.1 and 0.75 respectively. This produced the following results:



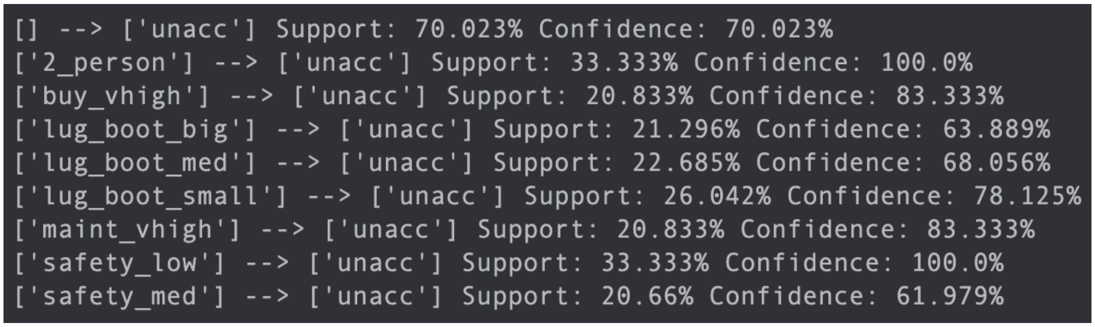
```
['2_doors'] --> ['unacc'] Support: 18.819% Confidence: 75.406%
['2_person'] --> ['unacc'] Support: 33.295% Confidence: 100.0%
['buy_high'] --> ['unacc'] Support: 18.761% Confidence: 75.0%
['buy_vhigh'] --> ['unacc'] Support: 20.787% Confidence: 83.295%
['lug_boot_small'] --> ['unacc'] Support: 25.999% Confidence: 78.087%
['maint_vhigh'] --> ['unacc'] Support: 20.787% Confidence: 83.295%
['safety_low'] --> ['unacc'] Support: 33.295% Confidence: 100.0%
['2_person', 'lug_boot_big'] --> ['unacc'] Support: 11.118% Confidence: 100.0%
['2_person', 'lug_boot_med'] --> ['unacc'] Support: 11.118% Confidence: 100.0%
['2_person', 'lug_boot_small'] --> ['unacc'] Support: 11.06% Confidence: 100.0%
['2_person', 'safety_high'] --> ['unacc'] Support: 11.118% Confidence: 100.0%
['2_person', 'safety_low'] --> ['unacc'] Support: 11.06% Confidence: 100.0%
['2_person', 'safety_med'] --> ['unacc'] Support: 11.118% Confidence: 100.0%
['4_person', 'safety_low'] --> ['unacc'] Support: 11.118% Confidence: 100.0%
['safety_low', 'lug_boot_big'] --> ['unacc'] Support: 11.118% Confidence: 100.0%
['lug_boot_med', 'safety_low'] --> ['unacc'] Support: 11.118% Confidence: 100.0%
['lug_boot_small', 'safety_low'] --> ['unacc'] Support: 11.06% Confidence: 100.0%
['more_person', 'safety_low'] --> ['unacc'] Support: 11.118% Confidence: 100.0%
```

Next, we changed the values of the minimum support and confidence and set them to 0.25 and 0.8. This produced the results below:



```
['2_person'] --> ['unacc'] Support: 33.333% Confidence: 100.0%
['safety_low'] --> ['unacc'] Support: 33.333% Confidence: 100.0%
```

Finally, we once again changed the values of the minimum support and confidence to 0.2 and 0.5 and this produced:

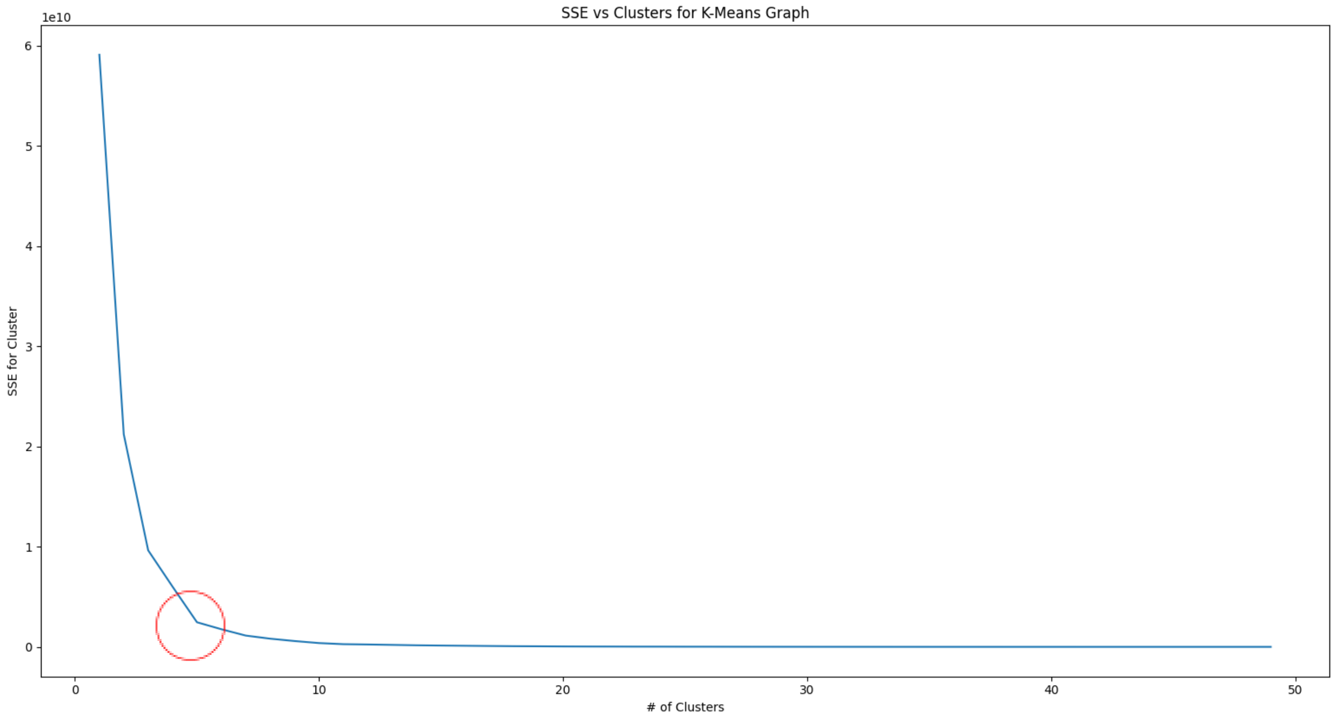


```
[] --> ['unacc'] Support: 70.023% Confidence: 70.023%
['2_person'] --> ['unacc'] Support: 33.333% Confidence: 100.0%
['buy_vhigh'] --> ['unacc'] Support: 20.833% Confidence: 83.333%
['lug_boot_big'] --> ['unacc'] Support: 21.296% Confidence: 63.889%
['lug_boot_med'] --> ['unacc'] Support: 22.685% Confidence: 68.056%
['lug_boot_small'] --> ['unacc'] Support: 26.042% Confidence: 78.125%
['maint_vhigh'] --> ['unacc'] Support: 20.833% Confidence: 83.333%
['safety_low'] --> ['unacc'] Support: 33.333% Confidence: 100.0%
['safety_med'] --> ['unacc'] Support: 20.66% Confidence: 61.979%
```

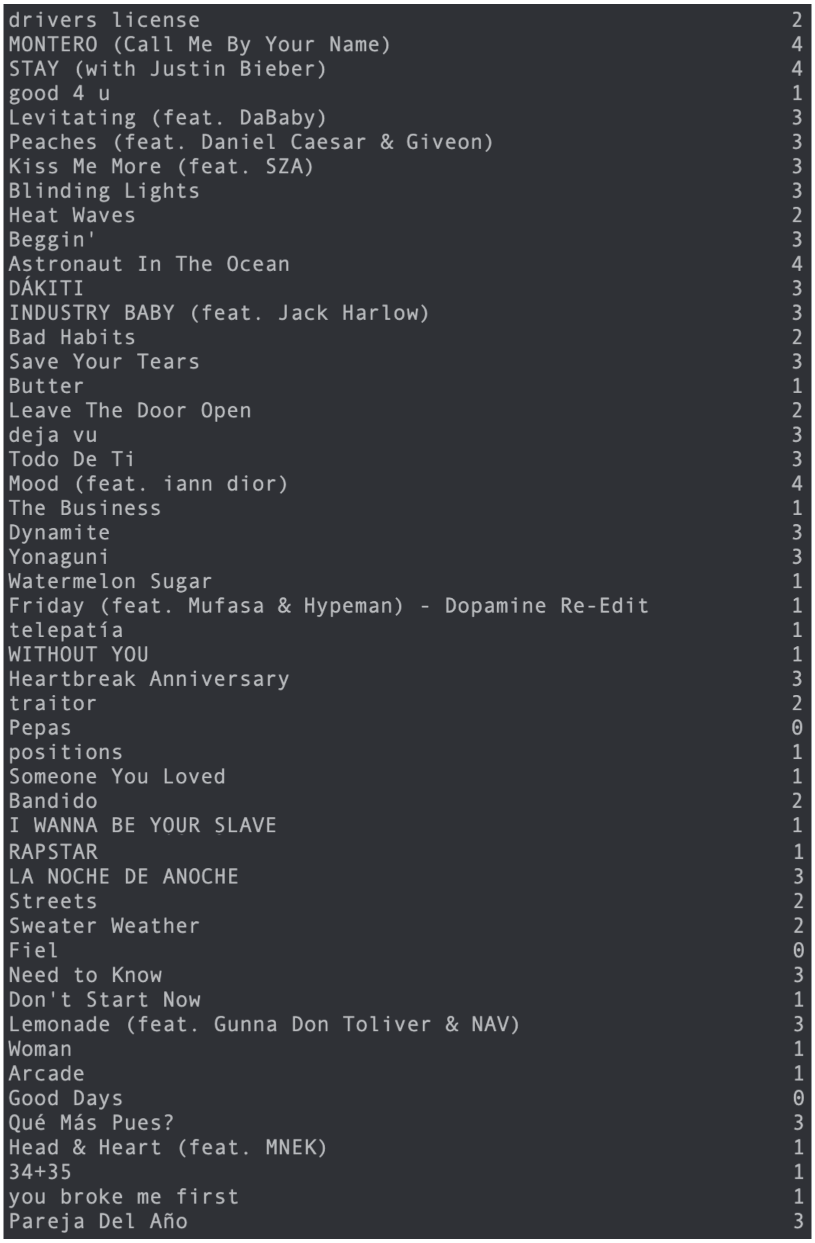
The association analysis with Apriori rules shows us which attributes relate to other attributes in the dataset. For the second set of support and confidence values, we see that relatively high support and confidence shows us that 2 person cars are unacceptable and cars with low safety are also unacceptable. However, when you decrease the parameters we see in result 3 that even medium safety is unacceptable followed by high buying and maintenance prices which also make the car unacceptable. An interesting rule we found in result 3 was that a lug boot of any size was deemed unacceptable. We believe this is the case because our dataset has significantly more instances that lead to the car being unacceptable compared to being acceptable, good and very good.

K-means Clustering

The dataset we used to do the K-means clustering is “Spotify Top 50 Songs 2021”. First, we wanted to determine the ideal K number of clusters. To do this, we used K values ranging from 1-50 and computed the SSE (sum of squared errors) for each K value. We then graphed the corresponding SSE to the K values and got the following graph:



To determine the ideal number of clusters, we analyzed the graph using the elbow method which is shown in the graph above. We found the best number of clusters to be 5. We then used the K-means algorithm with a value of 5 and got the following results:

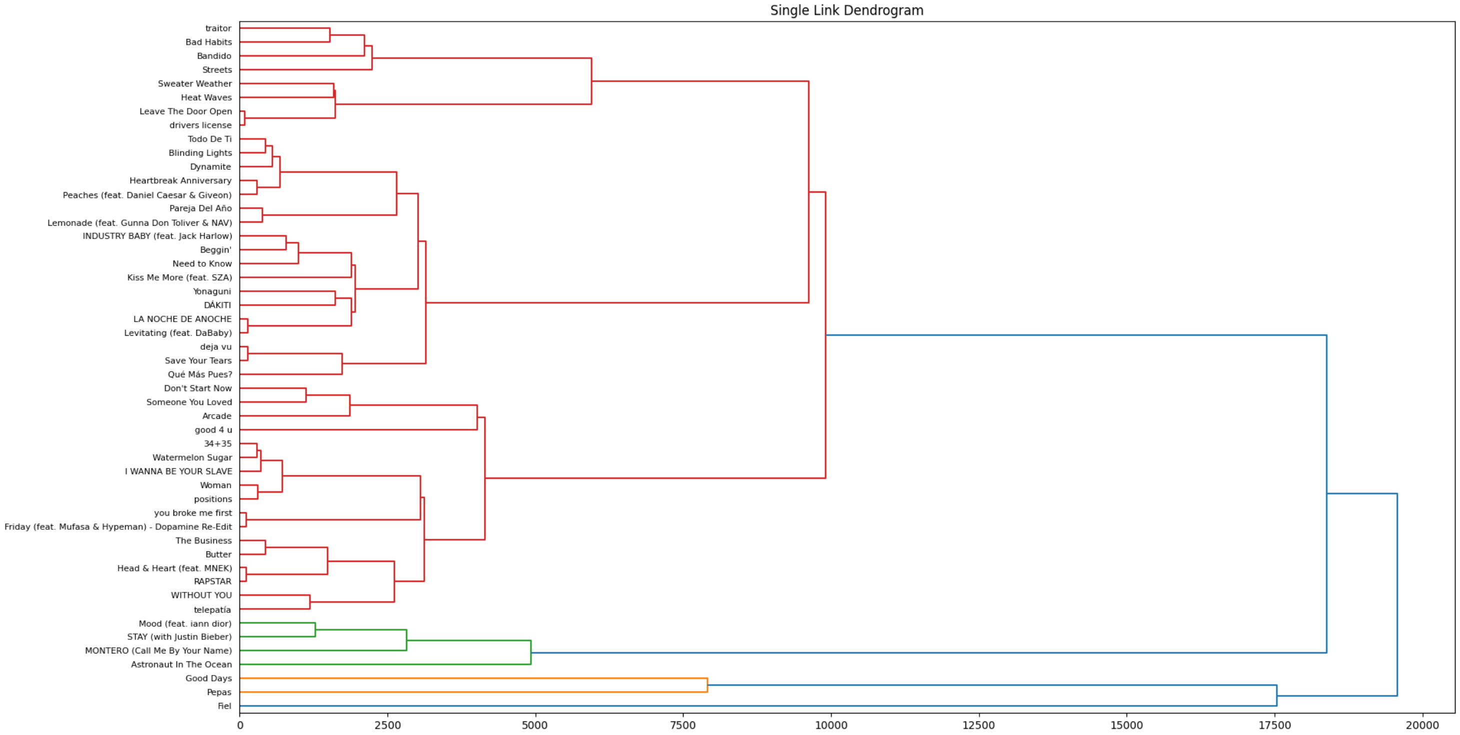


```
drivers license                                              2
MONTERO (Call Me By Your Name)                               4
STAY (with Justin Bieber)                                    4
good 4 u                                                     1
Levitating (feat. DaBaby)                                    3
Peaches (feat. Daniel Caesar & Giveon)                       3
Kiss Me More (feat. SZA)                                     3
Blinding Lights                                              3
Heat Waves                                                   2
Beggin'                                                      3
Astronaut In The Ocean                                       4
DÁKITI                                                       3
INDUSTRY BABY (feat. Jack Harlow)                            3
Bad Habits                                                   2
Save Your Tears                                              3
Butter                                                       1
Leave The Door Open                                          2
deja vu                                                      3
Todo De Ti                                                   3
Mood (feat. iann dior)                                       4
The Business                                                 1
Dynamite                                                     3
Yonaguni                                                     3
Watermelon Sugar                                             1
Friday (feat. Mufasa & Hypeman) - Dopamine Re-Edit           1
telepatía                                                    1
WITHOUT YOU                                                  1
Heartbreak Anniversary                                       3
traitor                                                      2
Pepas                                                        0
positions                                                    1
Someone You Loved                                            1
Bandido                                                      2
I WANNA BE YOUR SLAVE                                        1
RAPSTAR                                                      1
LA NOCHE DE ANOCHE                                           3
Streets                                                      2
Sweater Weather                                              2
Fiel                                                         0
Need to Know                                                 3
Don't Start Now                                              1
Lemonade (feat. Gunna Don Toliver & NAV)                     3
Woman                                                        1
Arcade                                                       1
Good Days                                                    0
Qué Más Pues?                                                3
Head & Heart (feat. MNEK)                                    1
34+35                                                        1
you broke me first                                           1
Pareja Del Año                                               3
```

This image shows the songs that are fit into clusters by their different attributes. For example, the songs “Levitating” and “Peaches” are in the same cluster because their attributes are similar in values. There are 3 songs in cluster 0, 17 songs in cluster 1, 8 songs in cluster 2, 18 songs in cluster 3 and 4 songs in cluster 4. From this we can see that the songs that dominated the top 50 list are composed of songs from clusters 1 and 3. This information is useful because it can allow you to predict if a song can make the top 50 Spotify charts based on the song's attributes and what cluster it falls under.

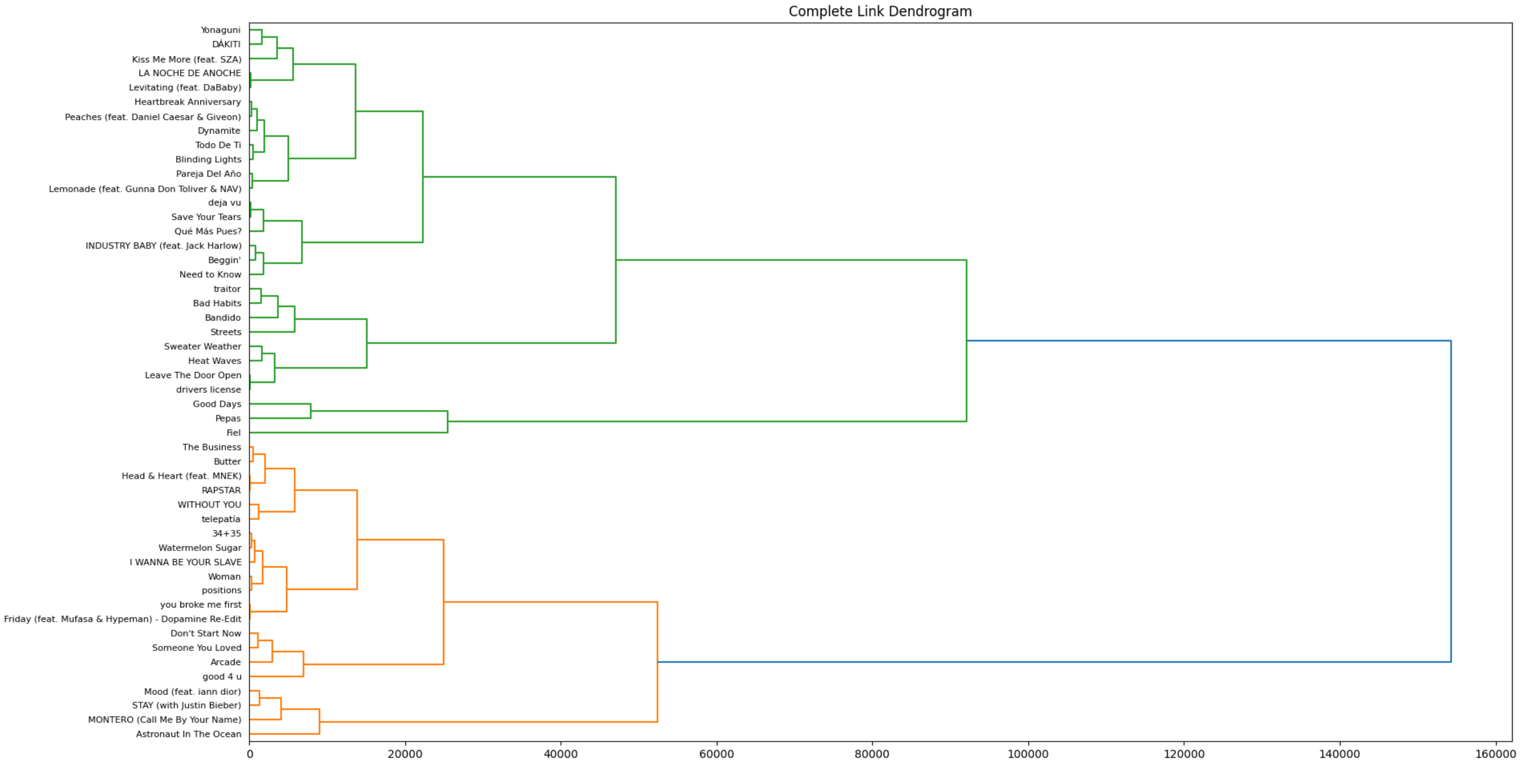
Hierarchical Clustering

For hierarchical clustering we also used the "Spotify Top 50 Songs 2021" dataset. We first applied a single link analysis and plotted the corresponding dendrogram. The dendrogram is:



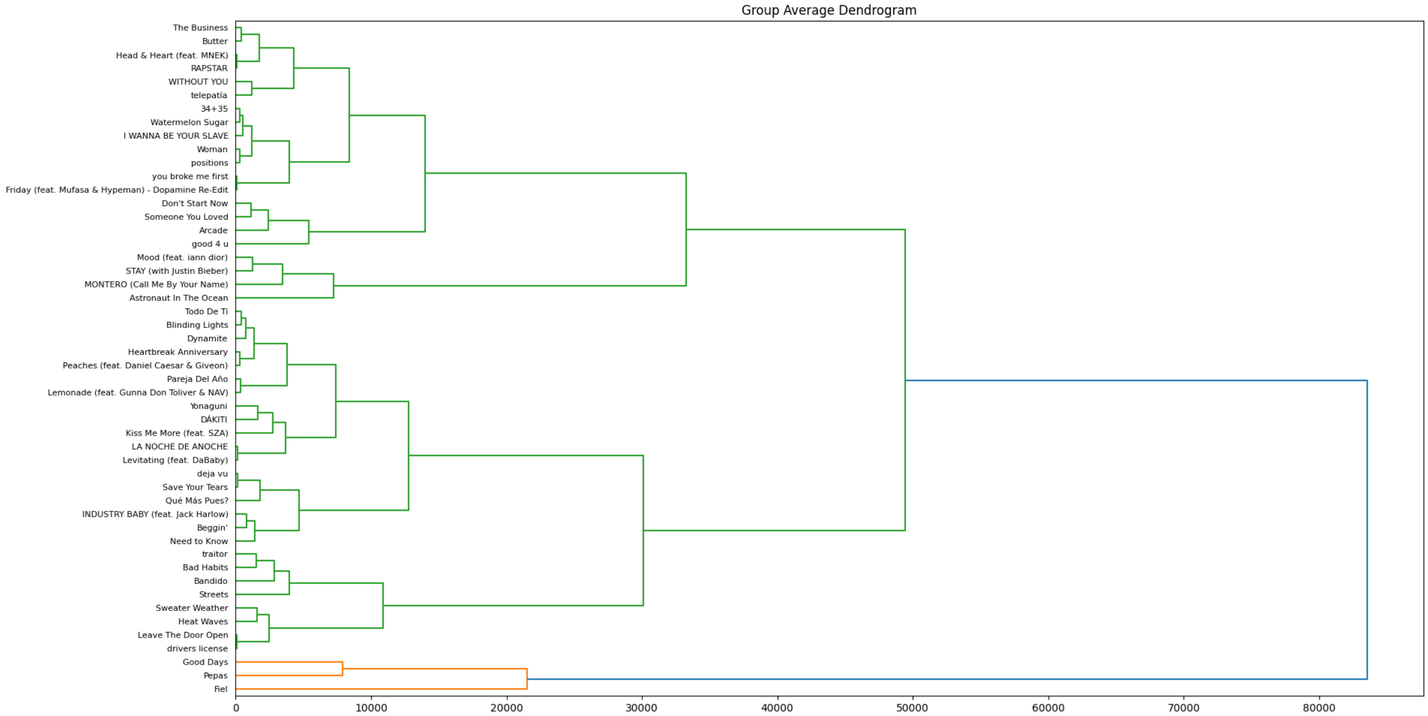
From the single link dendrogram, we see that songs with similar attributes get connected first in the hierarchy. As you go up the hierarchy tree, you see that songs with minimum distances between clusters are combined into a bigger cluster. This is repeated until all clusters are combined. From this dendrogram we see that initially there are 4 major clusters with most songs belonging in the red cluster. The single link analysis uses the optimistic approach of using the closest members to calculate the distance between clusters.

Next, we applied a complete link analysis and plotted the corresponding dendrogram. The dendrogram is show below:



For the complete link, the distance between the furthest points of the clusters are calculated and the minimum of these distances are used to merge the clusters together. Looking at the dendrogram we see that in contrast to the single link, this one initially has 2 major clusters before being merged. The complete link analysis uses the pessimistic approach where the furthest 2 points of clusters determines the distance between them.

Lastly, we applied the group average link analysis and the corresponding dendrogram is:

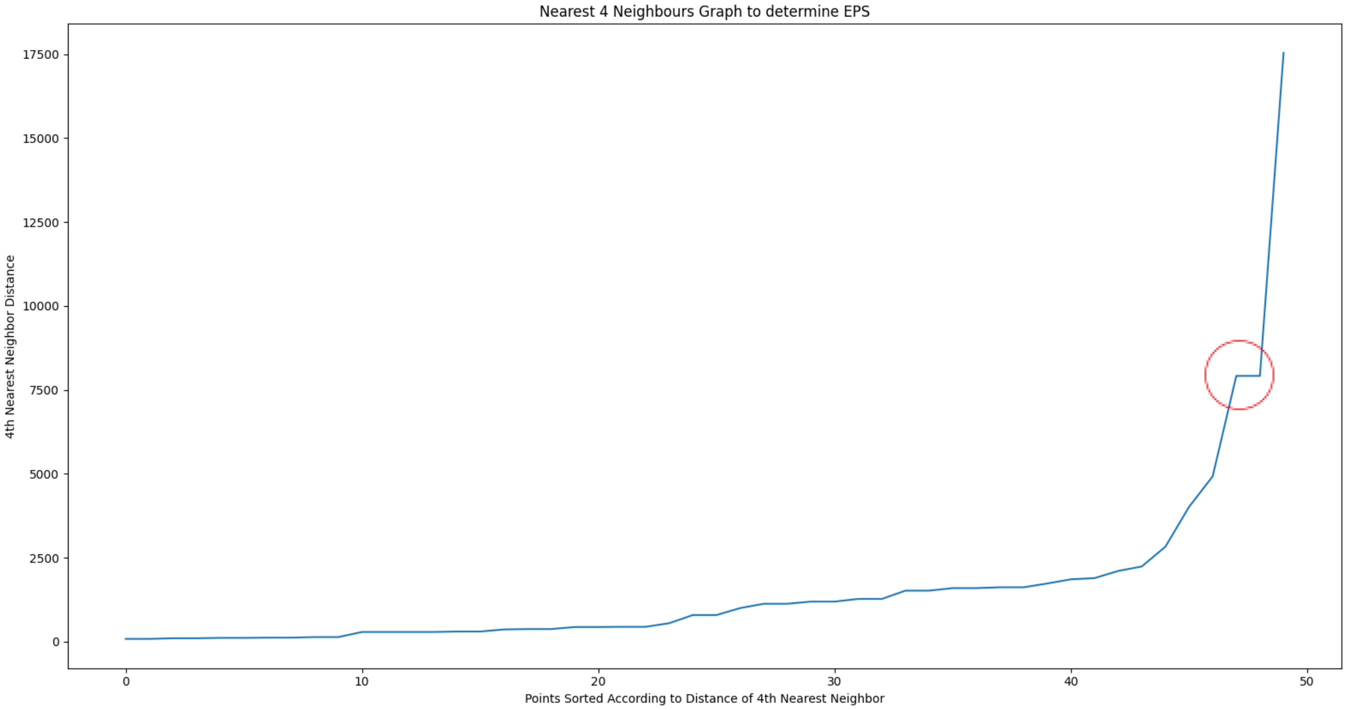


For the average group analysis, the distance between clusters is the average of the distances between members of the clusters. The minimum of these averages will determine which clusters are the closest and will be merged accordingly. Similarly to the complete link, there are 2 major clusters being merged however the size of the clusters is heavily concentrated in the green cluster. The average group analysis approach is somewhere in the middle of the single and complete link because it utilizes the distance between all members of the clusters and computers the average of these distances.

Density Based Clustering

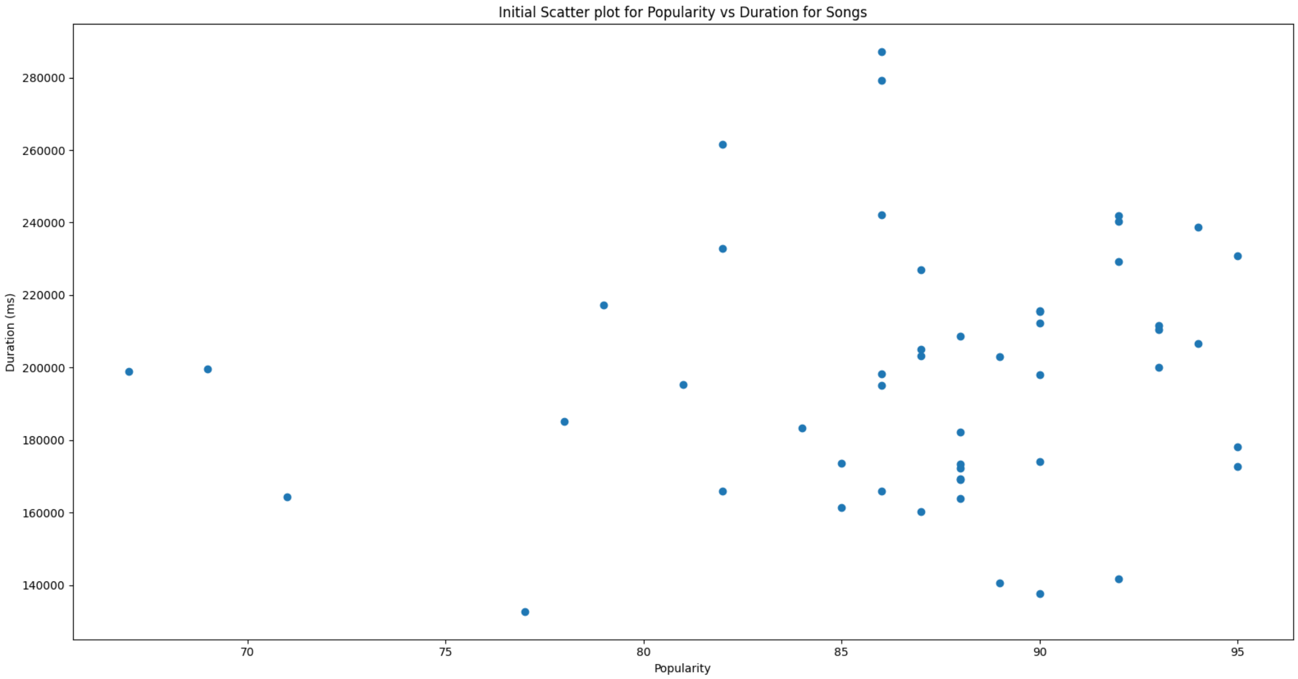
For the density based clustering, we used the “Spotify Top 50 Songs 2021”. We decided to look at the correlation between the popularity of the song and the duration of the song.

First, we applied the nearest neighbours algorithm and set the number of nearest neighbours to 4 which we decided based on the demos from class. We graphed the points sorted according to their distance of the 4th nearest neighbour to the distance which we used an online reference to do (Dbscan Algorithm Clustering in Python, 2021). The graph that was produced is:

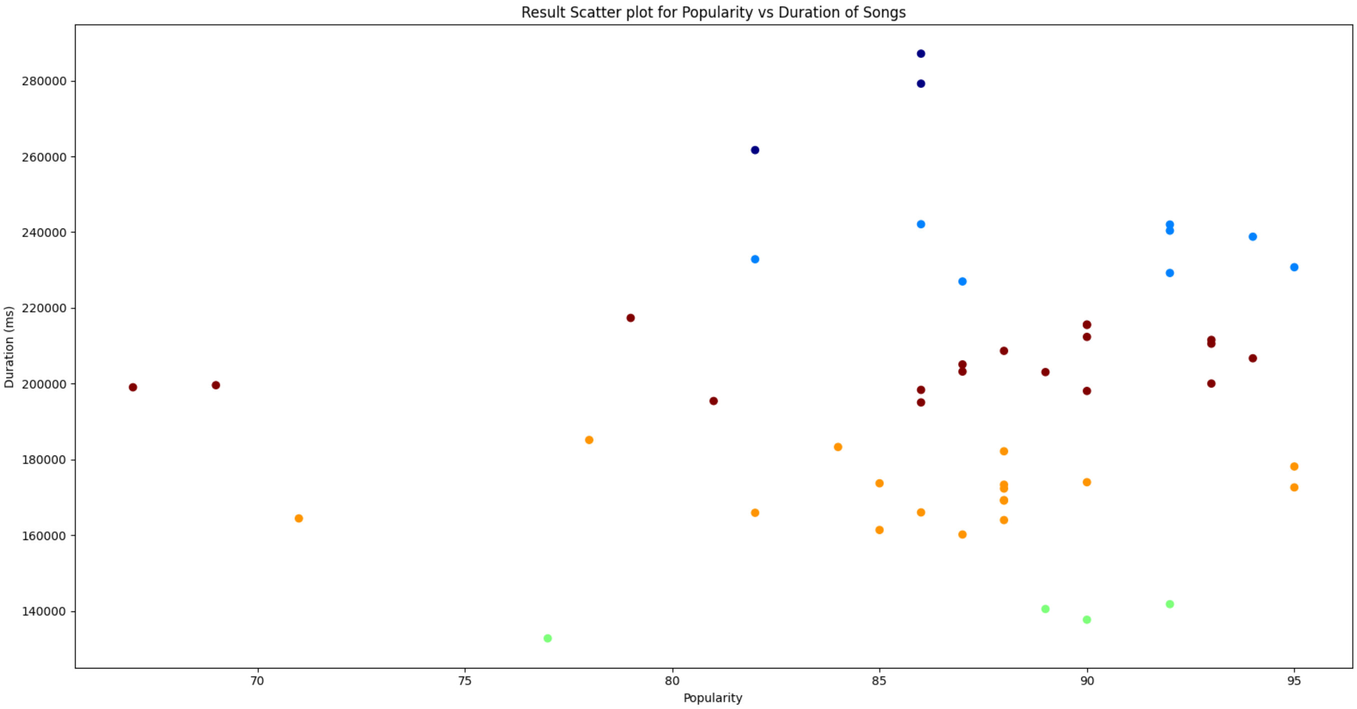


From looking at the graph, the best EPS (epsilon) is around 8000 which is where the graph starts to increase exponentially. This is also where the noise points start to increase because they have the furthest distances between their nearest neighbours so it is best to pick an EPS before this increase in noise.

Next, we plotted all the points from song popularity vs duration in a scatter plot which gave the following:



We then applied the DBSCAN algorithm using our determined EPS of 8000 and min_samples to 4. After running this algorithm, we create a new scatter plot for popularity vs song duration with each point coloured by their corresponding cluster label:



From looking at this new scatter plot, we see that there are 4 clusters excluding the noise. In comparison to the in-class demo, the graph looks emptier due to our dataset containing only 50 instances but the clusters are still easily identifiable.

Comments

All the data, association analysis and cluster analysis for the datasets we present in this report was used from “https://archive.ics.uci.edu/ml/datasets/Car+Evaluation”, “https://www.kaggle.com/datasets/equinxx/spotify-top-50-songs-in-2021” cited below. To run the python script simply place the script file in the same directory as “car.data” and “spotify_dataset.csv” and run the script. After running the code, the script will generate the 7 graphs we showed throughout this document. The various graphs and plots in this report may be hard to see in the document so we have also provided a png inside the zip file.

Data set Reference

Dbscan algorithm clustering in python. (2021). Engineering Education (EngEd) Program | Section. Retrieved April 12, 2022, from https://www.section.io/engineering-education/dbscan-clustering-in-python/

Dua, D., & Graff, C. (2017). Uci machine learning repository. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml

Spotify top 50 songs in 2021. (2021). Retrieved April 12, 2022, from https://www.kaggle.com/equinxx/spotify-top-50-songs-in-2021