Mini-Project: Social Network Analysis

By Shivam Goel, Marcus Campbell, Dhanushka Francisku, Saranya Kanderi

Aims and goals

In this project our aims are to apply k-means clustering algorithm to social media data, to predict what common interest teenagers are likely to have if they had a specified interest. Furthermore, to this we will distinguish interests more popular with female and males, along with does the number of friends someone has affect their interests.

Processing the data

The original dataset has age ranges from 3-100, so we have excluded any data that falls outside the range 13-20.

The gender values were marked as 'F' and 'M' or missing, in the dataset we did not remove any missing values. Generating a new column labeled female and a column labeled no gender for anyone with missing gender, if they the condition is satisfied the column will have value of '1' if not both will have '0' indicating a male participant.

Running K-means

K-means is a clustering algorithm used for partitioning a set of data points into a predetermined number of clusters. Initially we are going to use 5 clusters to see our results.

Before running any tests, we produced a heatmap which indicted there should be a link between basketball and football, sex and kissed, marching and band, blonde and hair, Hollister and Abercrombie.

Cluster Sizes:

Cluster	Size
1	1038
2	601
3	4066
4	2696
5	21599

Interests Dominant Cluster:

Cluster	Interests
1	Sexy, Kissed, Music, Rock, God, Hair, Blonde, Clothes, Die, Death, Drunk, Drugs
2	Band, Marching
3	Swimming, Cheerleading, Cute, Hot, Dance, Church, Jesus, Bible, Dress, Mall, Shopping, Hollister, Abercrombie,
4	Basketball, Football, Soccer, Softball, Volleyball, Baseball, Tennis, Sports,
5	**Has no prominent interest**

As you can see cluster 5 is the largest cluster, but no interests are dominant in that cluster, whilst the other clusters apart from cluster 2 have various interests and is hard to draw a link between interests.

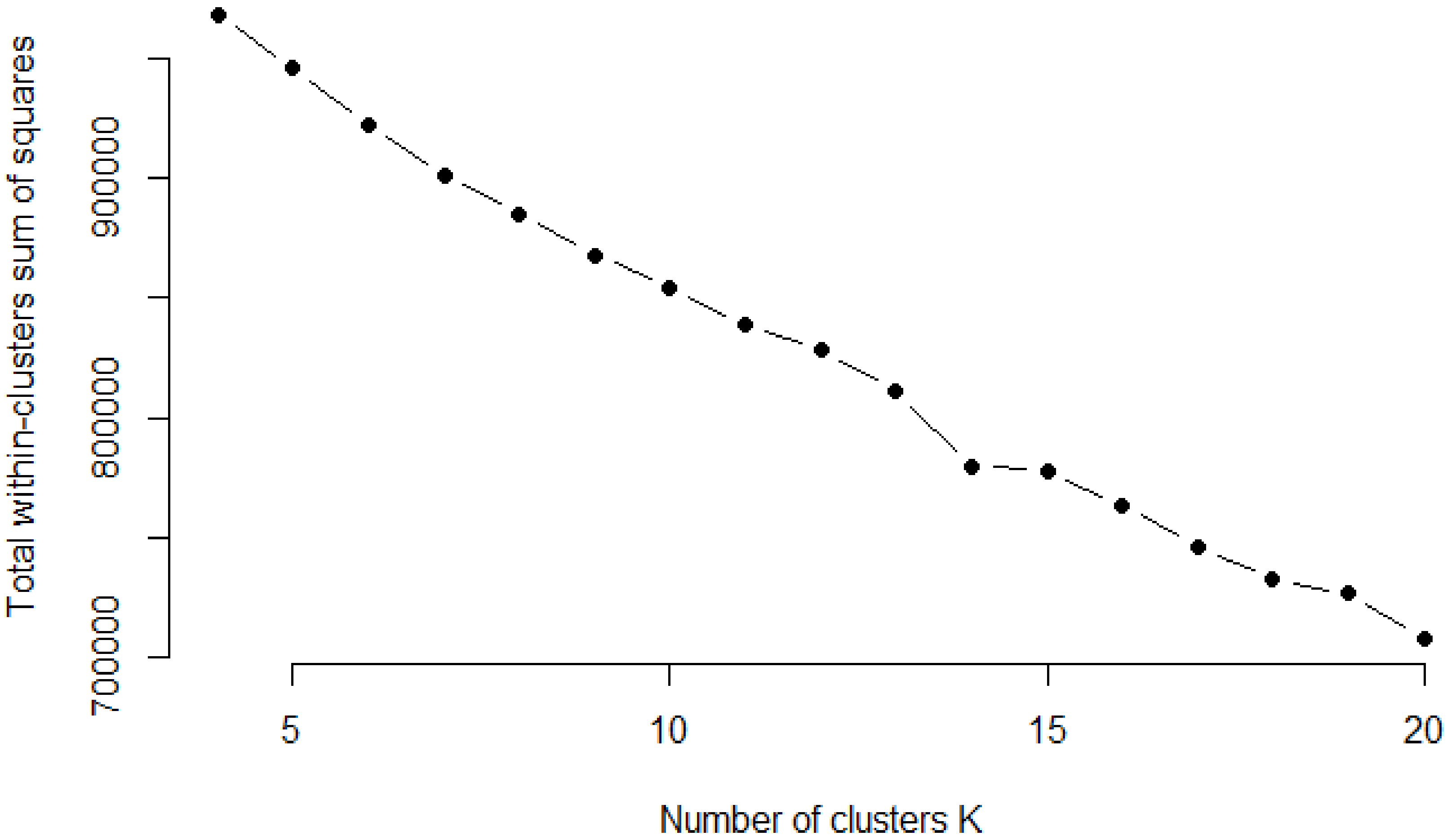
All ages for each cluster are between 17-17.4 years old, with cluster being most even split of male to female except cluster 3 which is dominantly female.

We must find optimal parameters to get establish greater links between interests, then further establish links between gender, age and friends.

Optimizing our K value

There are various methods used to calculate the optimal number of clusters. We chose to use the elbow method and silhouette method to compare outcomes as ultimately the optimal amount of cluster is subjective.

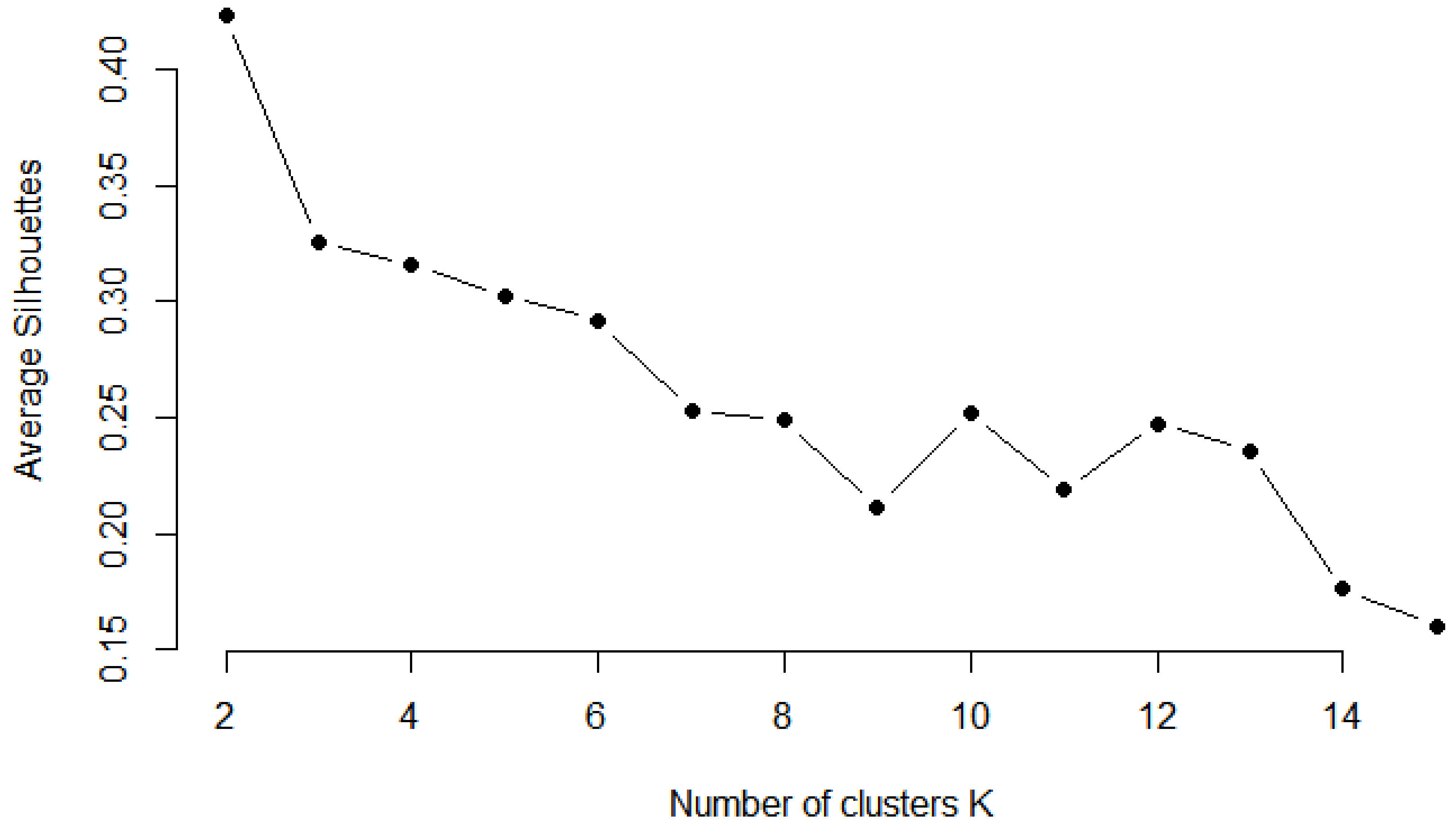
Elbow method



After apply the elbow method to our preprocessed data that has been scaled.

From the graph the elbow method indicates the optimal is between 12 or 14.

Silhouette method



This method has indicted that cluster 10 or 12 may be the most optimal.
As the elbow as suggested 12 was optimal, we will use this k value to obtain our new model.

Optimized K-means

We ran the same K-means as earlier but set clusters = 12.

Cluster sizes:

Cluster	Size	Cluster	Size
1	811	7	1124
2	634	8	36
3	3570	9	1263
4	239	10	819
5	18133	11	659
6	564	12	2148

After analyzing the cluster centers, we can attribute the following interests as being prominent to the following clusters.

Cluster	Interests	Cluster	Interests
1	Hollister, Abercrombie	7	Church
2	Cheerleading	8	God, Jesus, Bible, Blonde
3	Cute, Hot, Dance, Dress, Mall, Shopping	9	Drunk
4	Tennis	10	Die, Death
5	**Has no prominent interest**	11	Swimming, Sex, Sexy, Kissed, Music, Rock, hair, clothes, Drugs
6	Band, Marching	12	Basketball, Football, Soccer, softball, Volleyball, Baseball, Sports

The table shows each interest's dominant cluster, however there is still overlapping:

- As cluster 7 interests are also prominent in cluster 8 and vice versa.
- Music relates to cluster 6
- Cute, hot, mall and shopping are relevant in cluster 1
- Drunk is also relevant in cluster 11.

If we break down the clusters into categories, we get:

Category	Cluster	Category	Cluster
Fashion	1 and 3	Sport	4 and 12
Music	6 and 11	Cheerleading	2
Religion	7 and 8	Death	10
Partying	9 and 11		

Comparing clusters to gender, age and friends

Cluster	Female	Age	Friends
1	0.836	16.873	40.245
2	0.907	16.995	38.730
3	0.894	17.099	35.996
4	0.665	17.300	32.339
5	0.692	17.298	27.054
6	0.716	17.401	32.394
7	0.788	17.256	36.420
8	0.528	17.516	33.806
9	0.756	17.418	31.529
10	0.757	17.221	31.017
11	0.835	17.074	30.590
12	0.682	17.051	35.168

Drawbacks

Despite cluster 5 being the largest cluster it has no clear interest attributed to it, this could be due to a large amount of overlapping for the clusters or a potential lack of informative features to distinguish interests for each other.

Furthermore, 75% of the participants in this dataset are aged between 16-18 years meaning when averaged out the ages will likely always be around 17 years. Which can be seen in the tables on the previous slide the cluster have age range of 16.87-17.52, making it difficult to distinguish the interests of teenagers with age 16 or less and greater than 18

Lastly 73% of the participants are female, 17% male and 10% unknown, meaning the data will be skewed towards clusters being heavily female.

Findings

Overall, our findings show that females fall below the average in clusters 4,5,6,8 and 12 indicating that these are more likely male interests (sports, religion, Music). Whilst clusters 1,2,3,7,9,10 and 11 are more female dominant, these interests being around partying, death, religion, fashion, cheerleading.

Regarding age it is hard to say if it had any effect on the interests as all cluster averaged out an age of around 17, whereas those with more friends are more likely to be interested in fashion, sports and religion compared to those interested in partying and music.

Our findings show what was expected in the initial heatmap.

Parallelization

To help with running our models we used parallelization, which is the process of delegating large computation task into smaller tasks across multiple processors and other computational resources to reduce time cost.

An example of this is we ran a complex code:

- Using multiprocessing = 9010 seconds
- Using 8 cores in parallel = 8526 seconds