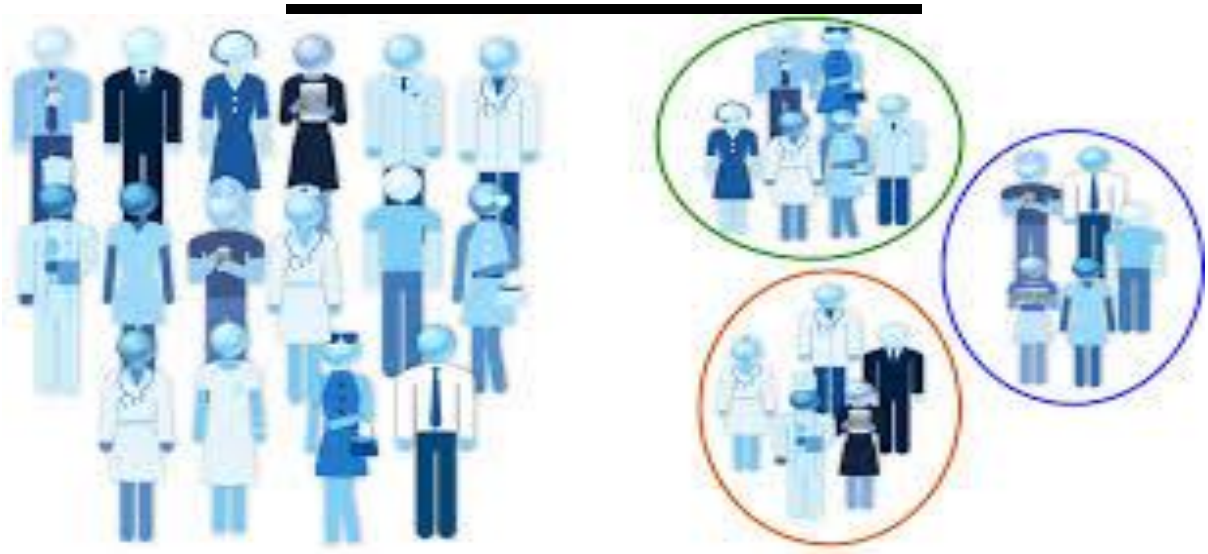
ASSIGNEMENT-3
CLUSTERING



Muhammad UMER
SID:44135149

Table of Contents

Data Wrangling

Introduction:
The data file provided for the analysis named google review data with 5456 participants and their average ratings of 24 different types of attractions in Europe. The file is firstly, checked for any missing values or messy data and then after fixing it is used for this clustering analysis. For the missing values in the data, median values of each variables calculated and later used to fill the missing values in the dataset.

Software package used for analysis:
R-studio is the software's used for the clustering analysis of the google review dataset.

Algorithms used for Clustering Analysis:
Following algorithms are used for the clustering analysis.

1. K-Means Clustering.
2. Hierarchal Clustering.

Software Outputs Results (K-Means Clustering)

Firstly, we had to deal with data in which I found missing values to be fixed
Outliers are of concern so we checked for them as well.

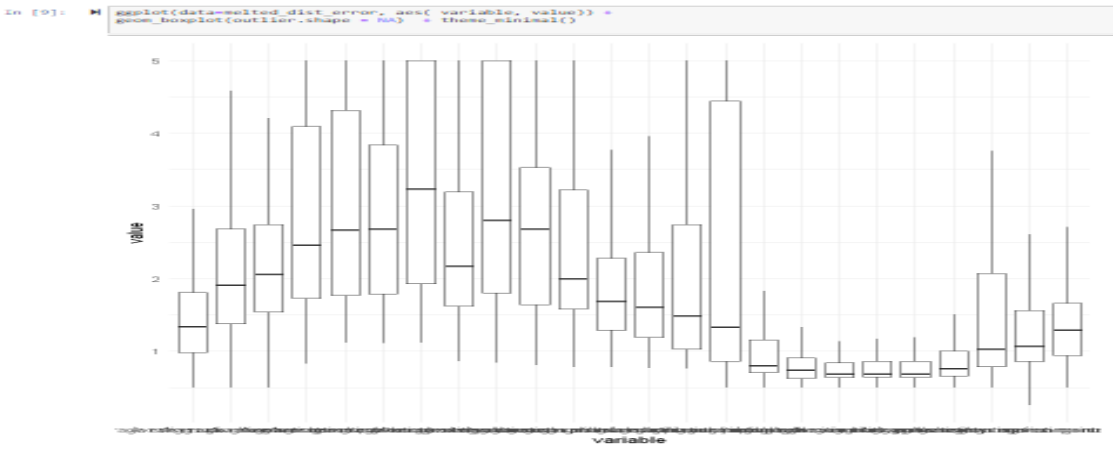
Firstly, we melted the data to be displayed:

```
melted_dist_error <- melt(data, id='User' )
```

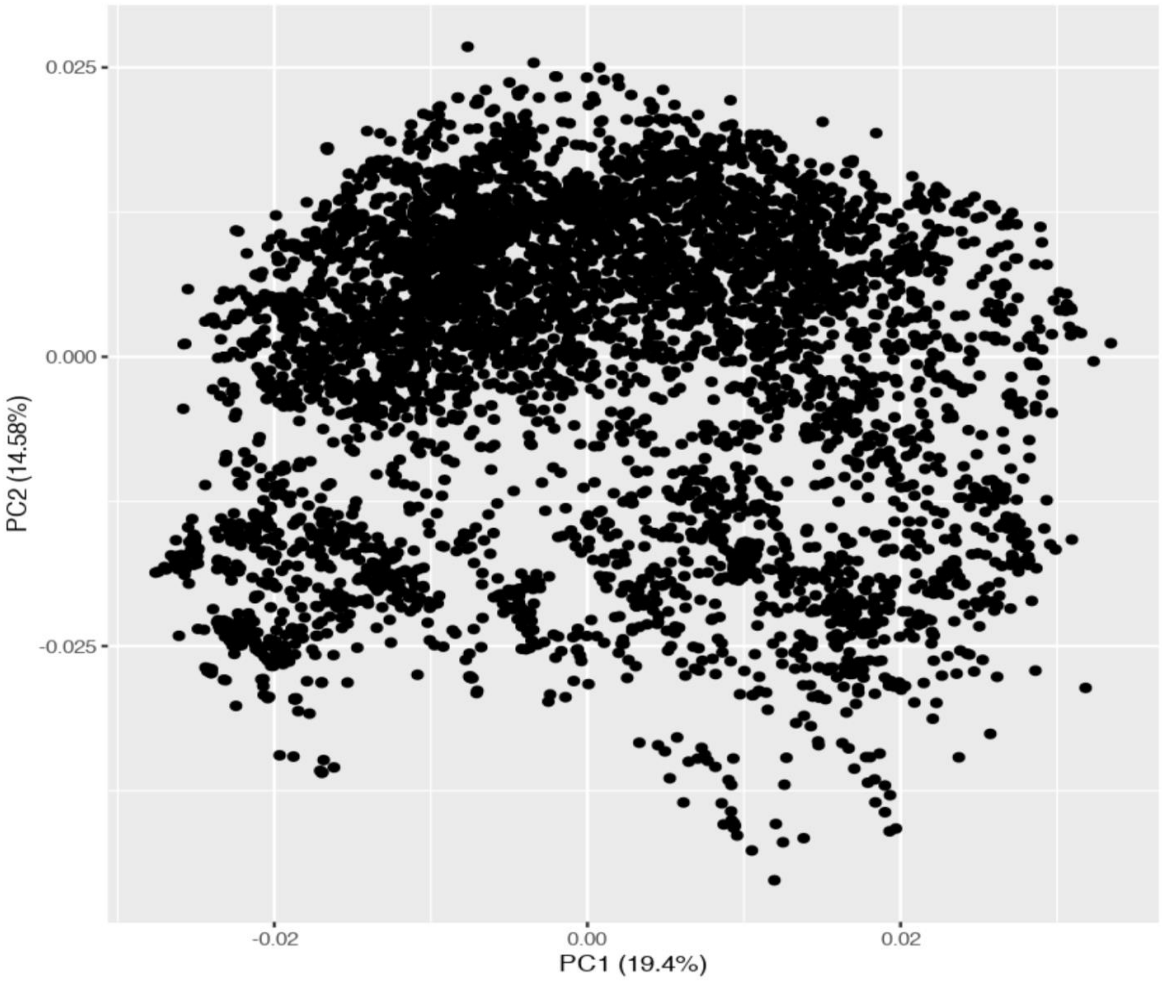
```
melted_dist_error
```

User	variable	value
User 1	Average.ratings.on.churches	1.34
User 2	Average.ratings.on.churches	1.34
User 3	Average.ratings.on.churches	1.34
User 4	Average.ratings.on.churches	1.34
User 5	Average.ratings.on.churches	1.34
User 6	Average.ratings.on.churches	1.34

The given below are the box plots for each Site, it can be seen the outliers have been fixed:

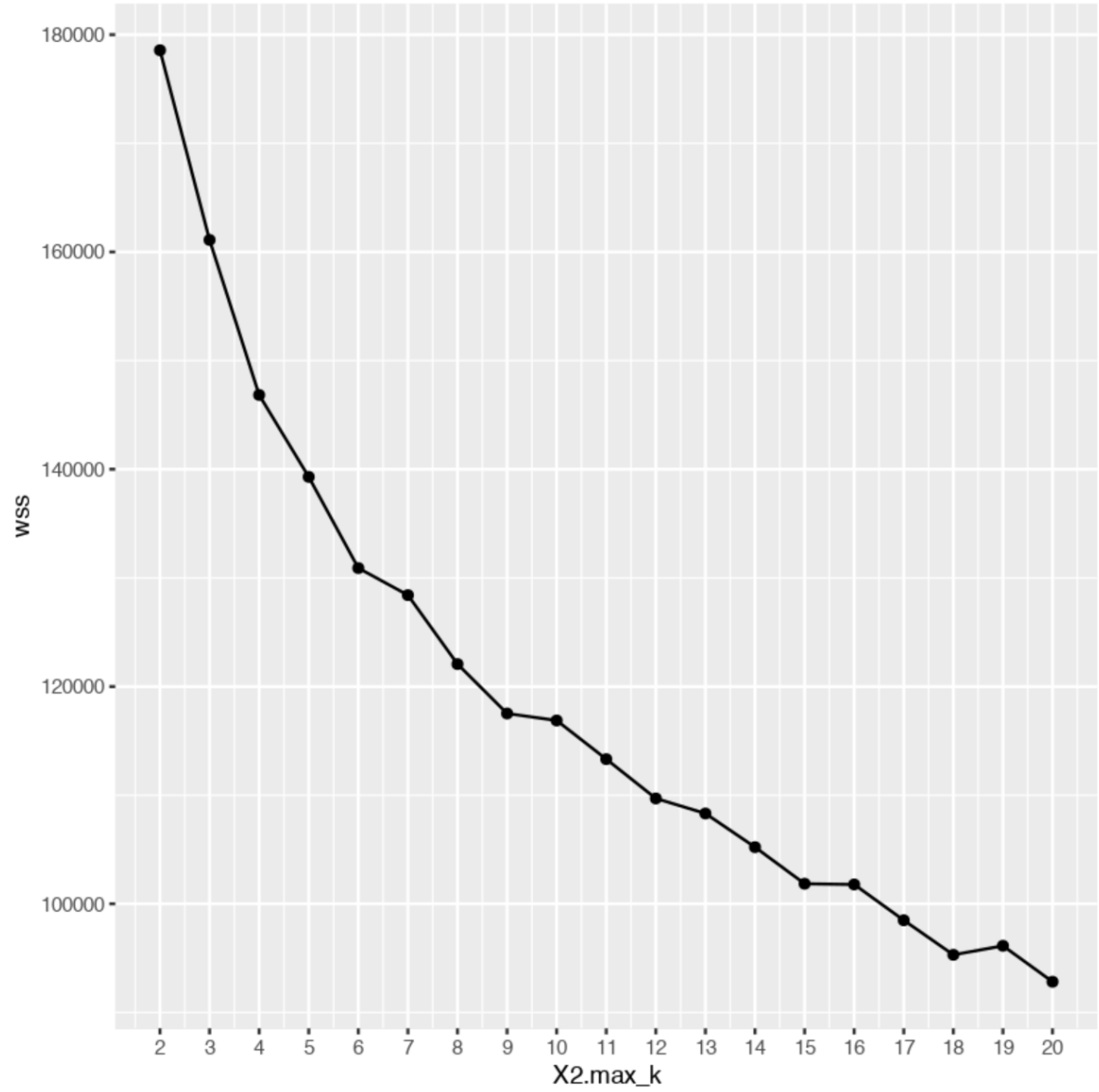


The number of clusters were to be identified since it was an unsupervised model, hence we first do a PCA on the data to reduce the components to visual form into 2 dimensions and then plot it. The plot shown above shows us the data and we can assume that there might be 3-4 clusters.



The second approach for the same is making a function to run for K 1 to 20 and find the optimum value of K from the graph of K. i.e. total sum of squared from the centres hence our evaluation criteria for the K-means models as well.

After plotting the graph we observed that optimum k value is 9 as a diminishing return is observed from the above plot. Now we choose to see the model values and cluster sizes at 9 and draw inference from the same.These are the cluster sizes for the decided 9 clusters.



After observing the market segments for the users as per the valuation of a particular sites. We have observed the cluster centres and the maximum values of them in the chosen 9 clusters.

Average.ratings.on.beaches	Average.ratings.on.parks	Average.ratings.on.theatres	Average.ratings.on.museums
3.456849	3.219562	2.599365	2.337155
1.731430	2.090789	2.171652	2.264094
1.760441	2.378236	3.026887	3.950778
1.940971	1.932114	1.918400	1.900171
3.411947	4.215061	4.564515	3.759727
1.608792	1.558517	1.565466	1.585445
2.640236	2.591863	2.348090	1.787995
2.132119	2.073582	1.979164	1.865254
2.622648	2.645441	3.043039	3.487554

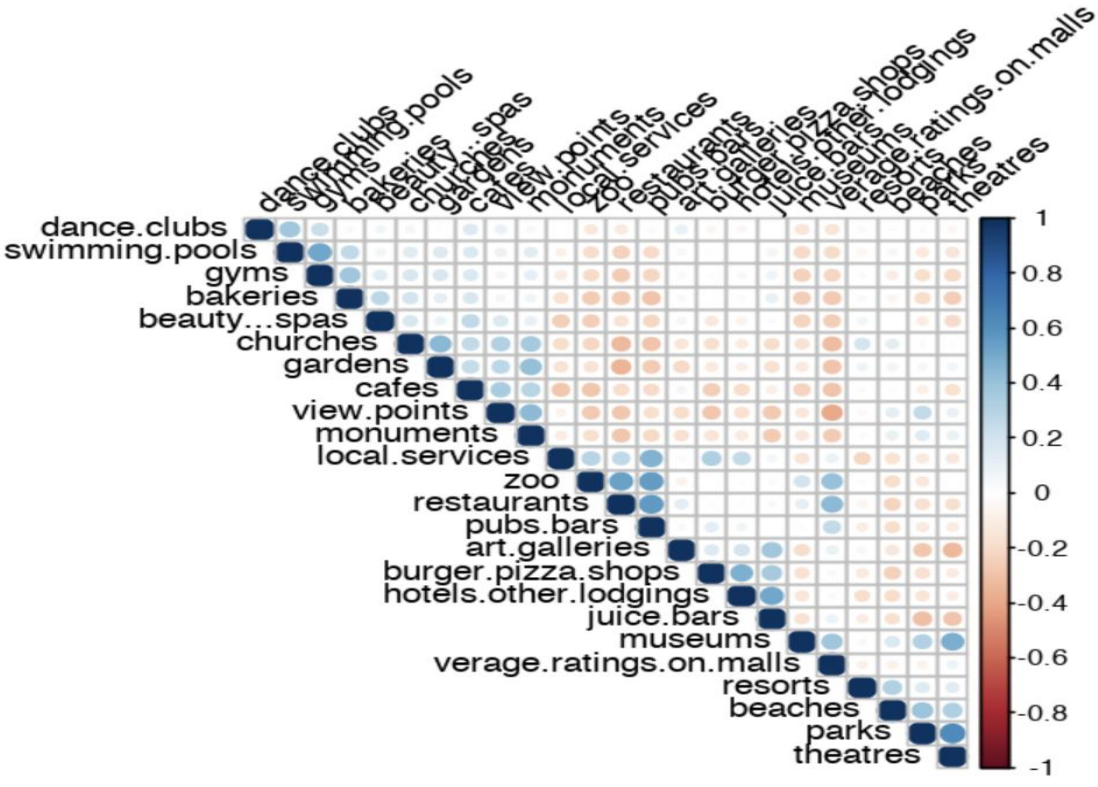
```
pc_cluster_2$size
```

457 811 771 175 1320 472 424 335 691

In the image displayed above we can see the cluster centres values of each cluster for given site.

Observation:

Average rating on theatre and parks column has highest value for 5th cluster and from image 2 its evident that theatres and parks have a high impact on users (as the given value of cluster centres is high) as we can see that the highest number of users fall in the 5th cluster which sums up to 1320.

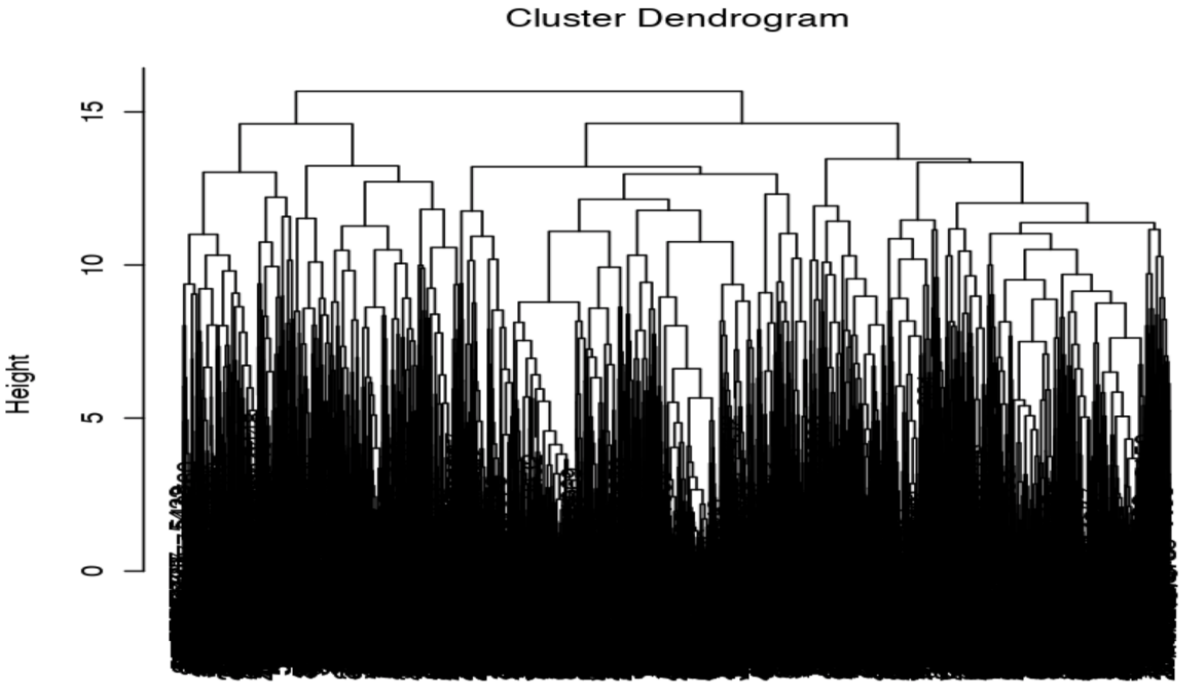


From the given correlation matrix, we identified a high correlation in several Sites with others for example:
Churches are highly correlated with gardens, cafes, viewpoints monuments from which it can be inferred that users who visit garden tends to visit, café viewpoints, monuments so these all might fall in a same cluster.

Software Output Results (Hierarchal Clustering)

We tried to make clusters using hierarchical technique

As we can see the hierarchical clustering dendrogram isn't giving appropriate results and further parameter tuning is required or an alternate approach where PCA components are fed to the hierarchical clusters for interpretation and further we will get principal components on which we will do hierarchical clustering.



It is observed from the clustering analysis that K-Means performed better than hierarchal clustering at number of clusters equal to 9 which has been decided from the graph of K-Means VS Mean Square. Hence, K-Means is an unsupervised learning method performing better than hierarchal clustering.