Agglomerative Hierarchical Clustering

Implementing Clustering on TripAdvisor Customer Rating Data

# Packages used

cluster: Used to create cluster from distance matrix

qplot: Used to create the heatmap and coloring the heatmap

# Data Description:

The data is from tripadvior in which the users have given **reviews** to different kinds of destinations. The values has been normalized to lie between 0 and 4 and are in decimals too. The data has 980 rows and 11 columns. The columns can be defined as follows.

1. User ID
2. Art Gallery
3. Dance Club
4. Juice Bars
5. Restaurants
6. Museums
7. Resorts
8. Parking
9. Beaches
10. Theaters
11. Religious Institutes

Hierarchal Clustering (summary)

Clustering is a technique of grouping same kind of elements in table together. There can be many application of clustering, but the most important once are as follows:

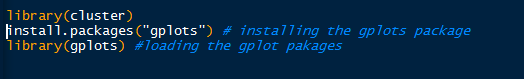
1. To find the correlation between in the data.
2. To understand the pattern in unlabeled data.

By grouping same kind of columns together with the help of clustering, we can understand the correlation between columns can take appropriate steps.

Sometimes, the data does not have labels or names given to attributes or column, in that case clustering can help us understanding the patter in distinguishing the groups so that they can be analyzed separately.

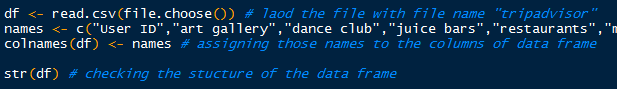
# Findings

1.

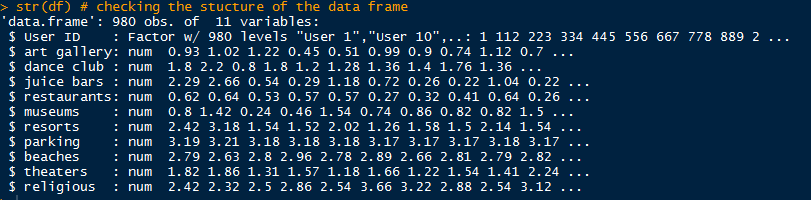


First, I have loaded appropriate libraries in order to do proper functionality. The cluster function will help creating the cluster, where as the gplots will help us create a heatmap.

2.



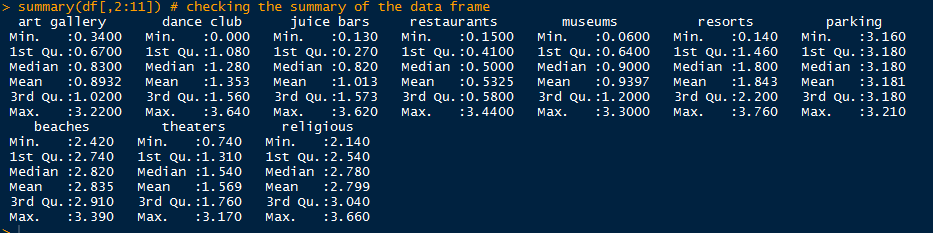
First, I am reading the file directly from the desktop as csv file, the name of the file is tripadvisor. Then, I am creating a names vector that has all the names stored in it that has to be assigned to the columns of data frame. Then, I am assigning the names to the columns.



Looks like that the structure of data frame depicts a clear picture. The first column is a factor because it is the user ID which is specific to each user. Then, all the other columns showing the reviews belong to that column are in numeric form.

3.

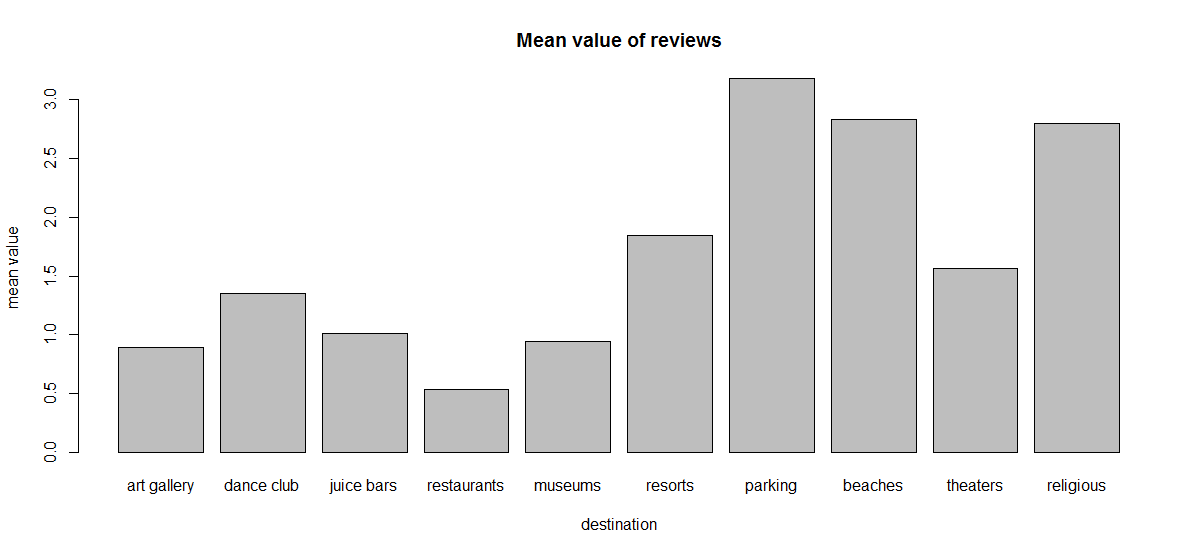




Looking at the summary of the dataset, we can see that the minimum value of reviews in each one of the destination is zero and the maximum value is below 4.

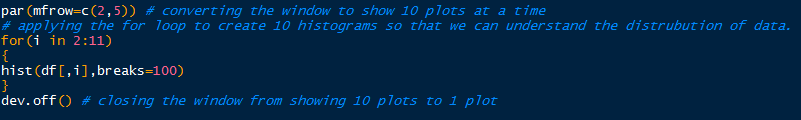
Moreover, most of the value of reviews are normally distributed because mean and median are almost equal.

To get a more clear picture, I have drawn a bar plot that is showing mean value of reviews belong to each one of the destinations. The graph is shown below.

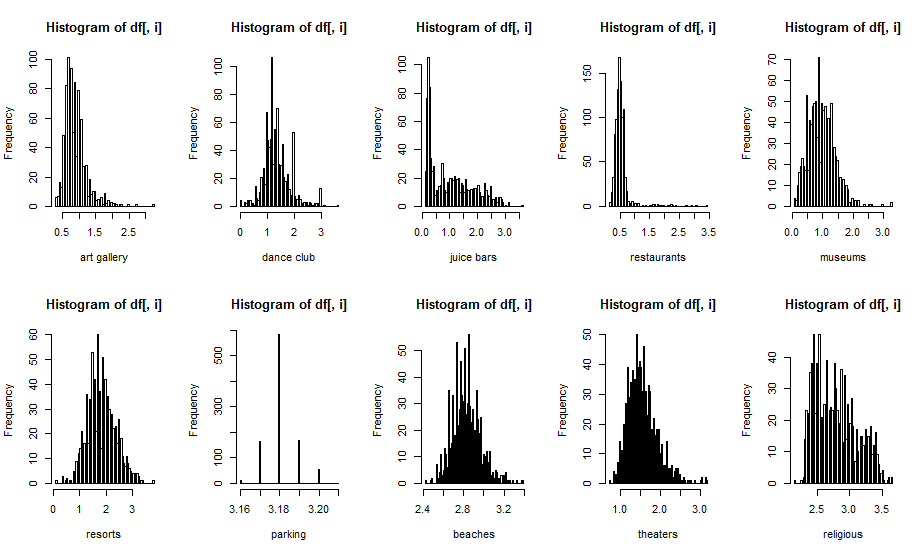


Looks like the mean value of reviews given to parking is highest. Meaning, parking is given the highest reviews on average. Moreover, it is surprising to know that restaurants are given the least ratings.

4.



In this step, I have first created a window that can hold 10 plots at a time so that we can get a big picture of the data. Then, using a for-loop to iterate through all the columns of the data frame, I have created a histogram in order to understand the distribution of data. In the end, I closed the window back to a single plot.

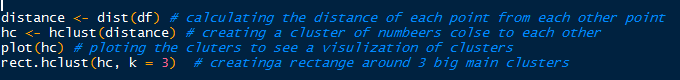


Looks like most of them are normally distributed. Except, parking juice bars and restaurants.

The juice bar and restaurants rating is rightly skewed. Meaning, people mostly have given bad rating to those places.

The parking ratings looks like it has been given on a 5 start scale that is why we have 5 straight lines.

5.

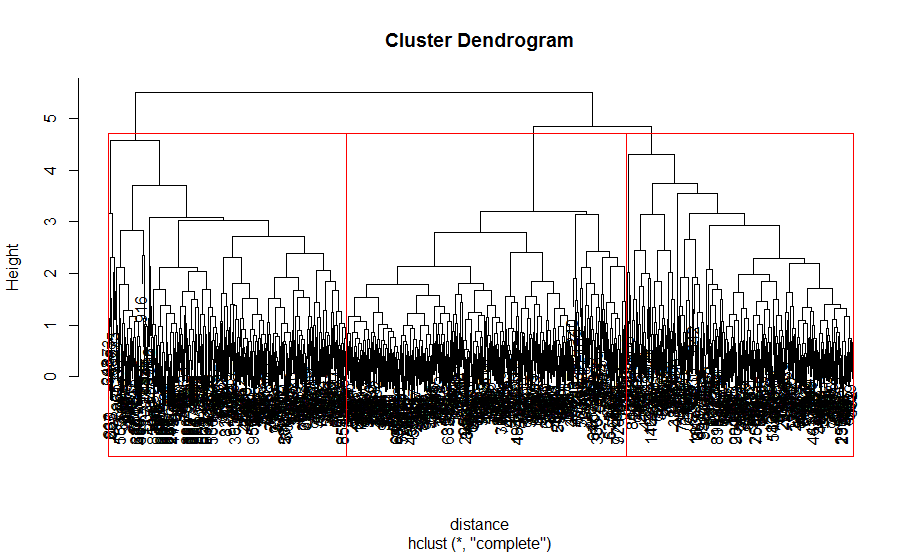


First, I am calculating distance of each point from each other point. The dist function will help us creating distance matrix that have distance calculated of each point from each other point.

Then, using hclust function and giving distance matrix as argument, we can create cluster of data points. In other words, the hclust function will automatically create groups of elements that are closer to each other.

To have a better picture of the grouping, we can create a dendrogram by using a plot command.

At last, the rect.hclust command help us create a rectangle around clusters. I have given 3 as parameter to have rectangle around 3 most big clusters.



As we can see, the red line is surrounding three section of this dendrogram. These three section are three different clusters, and each element in those cluster have same property. Meaning, each reviews that user have given has been clustered, and the reviews given in one cluster have same property because they are close to each other. We can use individual cluster to extract the user ID and can use that ID to understand the property of those user, but this is out of scope of this analysis.

6.



Using cutree command I have cut the above shown tree into three parts, each part is one cluster similar to the red rectangle showing in above graph. After cutting the tree, we can use table command to see how many user lies in each one of the cluster.

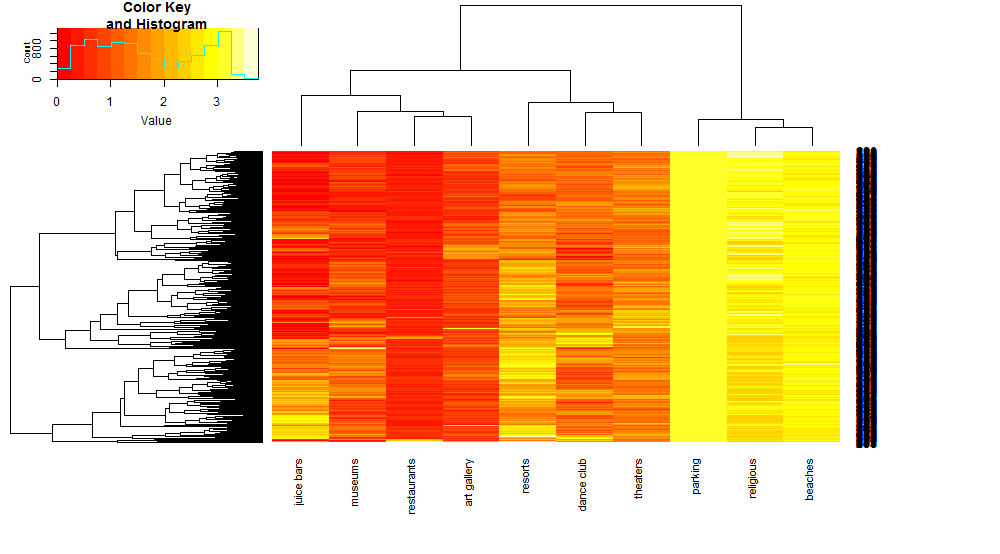


It is clear that first cluster has 314 users, the second have highest 368 user and third have lowest 298 users.

7.



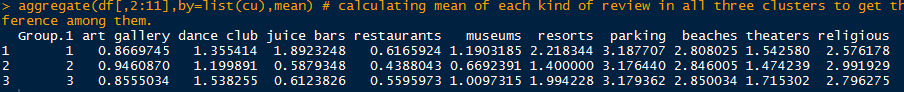
To get an overall picture of the whole analysis we have done so far, we can create a heatmap using heatmap.2 command and passing appropriate arguments. The heatmap will look like this.



It is clear by looking at the colors of heatmap that the data is divided into three main clusters, the left one which is extremen red are all the values that have low rating, the middle one is less dark and have a bit higher reviews rating, and the right most one is dark yellow which have highest ratings given by the user.

On the x-axis we can also see the name of the destinations. The parking, religious and beaches are given highest rating and hence they are clustered together. Whereas juice bars, museums and restaurants are given least ratings and hence are grouped together.

To take a step further, I have calculated the mean value of ratings belongs to each destination, separately for all three clusters. It looks something like this.



From this I observed that, in each of the three cluster people have given almost equal ratings to all of the destinations except juice bars and museums.

The second cluster has given bad reviews to museums, whereas first cluster has given very good reviews to juice bars.

References:

Anonymous. June11, 2017. *HowToDataViz.* “How to Make an R Heatmap with Annotations and Legend”. <https://www.youtube.com/watch?v=T7_j444LMZs>

Anonymous. “Travel Review Dataset”. *University of Carolina Irvin*. <https://archive.ics.uci.edu/ml/datasets/Travel+Reviews>

Davo. May 15, 2018. “Making a heatmap in R with the pheatmap package”. *DAVE TANG’S BLOG.* <https://davetang.org/muse/2018/05/15/making-a-heatmap-in-r-with-the-pheatmap-package/>