DAPT 622 Assignment 3

Daniel Erb

March 2, 2019

Contents

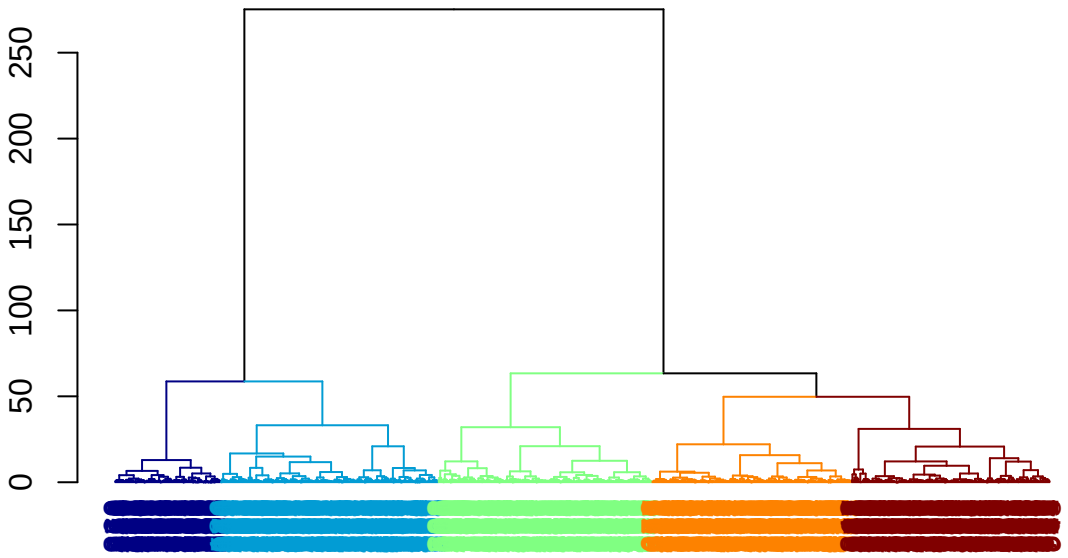
Part 1 - Cluster Analysis

Section A

Perform a heirarchical cluster analysis (via Ward's method) using all the variables except User ID.

Subsection i

Select an appropriate number of clusters. Provide a dendrogram with the clusters highlighted.



Subsection ii

Provide a table summarizing the clusters via their means. Are there any distinguishing qualities regarding the clusters?

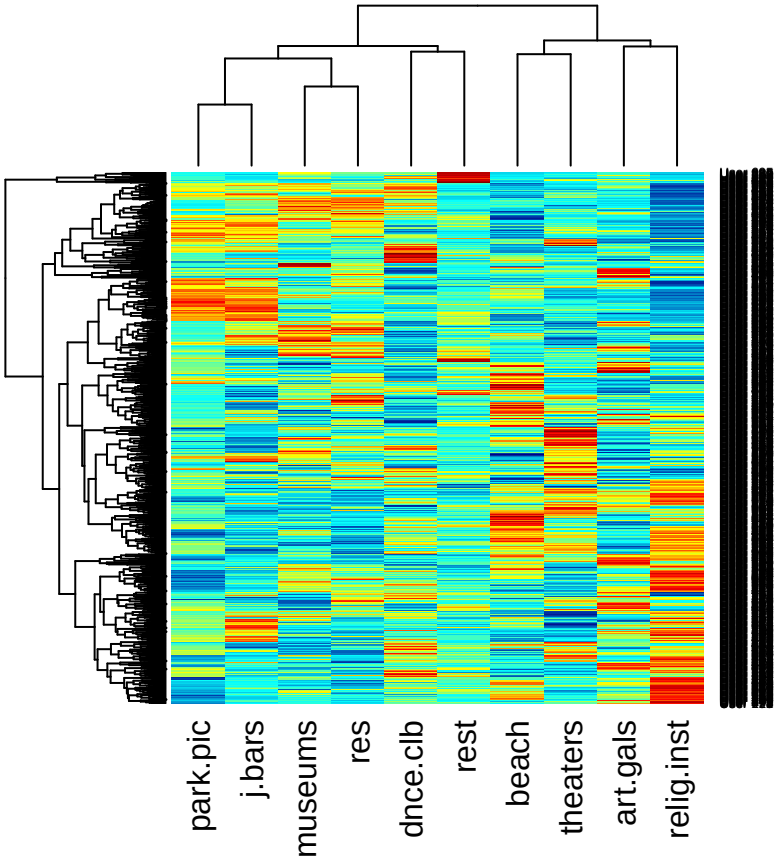
Table 1: Hierarchical Cluster Summary

Cluster	art.gals	dnce.clb	j.bars	rest	museums	res	park.pic	beach	theaters	relig.inst
1	0.94	1.53	1.56	0.56	1.32	2.45	3.19	2.82	1.63	2.57
2	1.03	1.04	0.97	0.49	0.76	1.63	3.18	2.85	1.42	2.81
3	0.87	1.36	0.42	0.41	0.68	1.34	3.17	2.84	1.48	3.14
4	0.78	1.53	0.34	0.67	0.95	1.88	3.18	2.86	1.80	2.82
5	0.78	1.26	2.36	0.55	0.97	1.91	3.19	2.80	1.47	2.56

For this dataset, 5 clusters were chosen as they seem to separate the groups into by usable chunks without being too specific as to be confusing. When looking at the average scores for each type of destination, across the individuals of cluster 1, we see that they tend to rate dance clubs, museums, and restaurants more positively than the individuals within other clusters, while only giving lower scores to religious institutions. Cluster 2, seems to favor art galaries, while disfavoring dance clubs and theaters, when compared to the other clusters. For cluster 3, individiduals within this group tend to rate juice bars, museums and restaurants lower than other clusters, while rating religious institutions higher than others. Cluster 4 tended to give particularly low ratings to juice bars. This in contrast to rating restaurants and theaters higher than other clusters. Cluster 5 had an average rating for juice bars that tended to be one or two whole ratings higher. They also tended to rate religious institutions lower than other clusters. One intersting note for the averages, both beaches and parks seemed consistently rated across all the clusters.

Subsection iii

Perform a "two-way" cluster analysis (i.e., cluster the variables) and provide the dendrogram showing the variable clusters and a heat map of the data. Which variables cluster together?



Parks and picnic areas clustered together with juice bars earlier than any other categories. The next to cluster to form was comprised of museums and restaurant ratings. These two clusters then clustered sooner than any other categories clustered wih any other categories. As we follow the tree upward, we see that beaches and theaters cluster together next, however the visual seems to give weeker evidence of this.

Section B

Now perform a k-means cluster analysis. Specify the same number of clusters as selected in part a). Display the cluster summary including the cluster means. Are different clusters produced than in part a)? Describe what you see.

Table 2: K-Means Cluster Summary

Cluster	art.gals	dnce.clb	j.bars	rest	museums	res	park.pic	beach	theaters	relig.inst
1	0.98	1.12	1.29	0.47	0.81	1.66	3.18	2.83	1.40	2.85
2	0.88	1.23	0.48	0.57	1.11	2.17	3.18	2.88	1.70	2.80
3	0.86	1.30	0.36	0.45	0.62	1.35	3.18	2.84	1.59	3.01
4	0.83	1.34	2.26	0.56	1.16	2.16	3.19	2.80	1.57	2.54
5	0.95	2.09	1.17	0.74	1.30	2.35	3.18	2.80	1.56	2.60

Table 3: First 10 Row Cluster Assignment

User.ID	hier.cluster	kmean.cluster
User 1	1	4
User 2	1	4
User 3	2	3
User 4	3	3
User 5	1	1
User 6	3	3
User 7	3	3
User 8	4	3
User 9	1	5
User 10	4	2

While many of the categories have similar values for their average ratings, the values are not the same when comparing the 5 clusters formed under the k-means and the hierarchical clustering methods. This would lead us to believe that all of the points have not been assigned to the same clusters, between these methods. We can confirm this by looking at the cluster assignments of just the first 10 rows, where user 3 and 4 are part of hierarchical clusters 2 and 3 respectively, while they are both within k-means cluster 4.

Part 2 - Correspondence Analysis

Section A

Are the variables WorkClass and education independent of each other? Perform a chi-square test of independence. Produce a mosaic plot of the data.

Table 4: Counts in Each Education Level

education	count
HS-grad	10501
Some-college	7291
Bachelors	5355
Masters	1723
Assoc-voc	1382
11th	1175
Assoc-acdm	1067
10th	933
7th-8th	646
Prof-school	576
9th	514
12th	433
Doctorate	413
5th-6th	333
1st-4th	168
Preschool	51

```
## Warning: Unknown levels in `f`: 10th, 11th, 12th, 1st-4th, 5th-6th,
## 7th-8th, 9th, Preschool, HS-grad, Assoc-acdm, Assoc-voc, Bachelors, Some-
## college, Doctorate, Masters, Prof-school
```

Table 5: Counts in Each Education Level

ed.collapsed	count
HS-grad	10501
Some-college	7291
Bachelors	5355
Masters	1723
Assoc-voc	1382
11th	1175
Assoc-acdm	1067
10th	933
7th-8th	646
Prof-school	576
9th	514
12th	433
Doctorate	413
5th-6th	333
1st-4th	168
Preschool	51

Section B

Section C