DAPT 622 Assignment 3

Daniel Erb

March 2, 2019

Contents

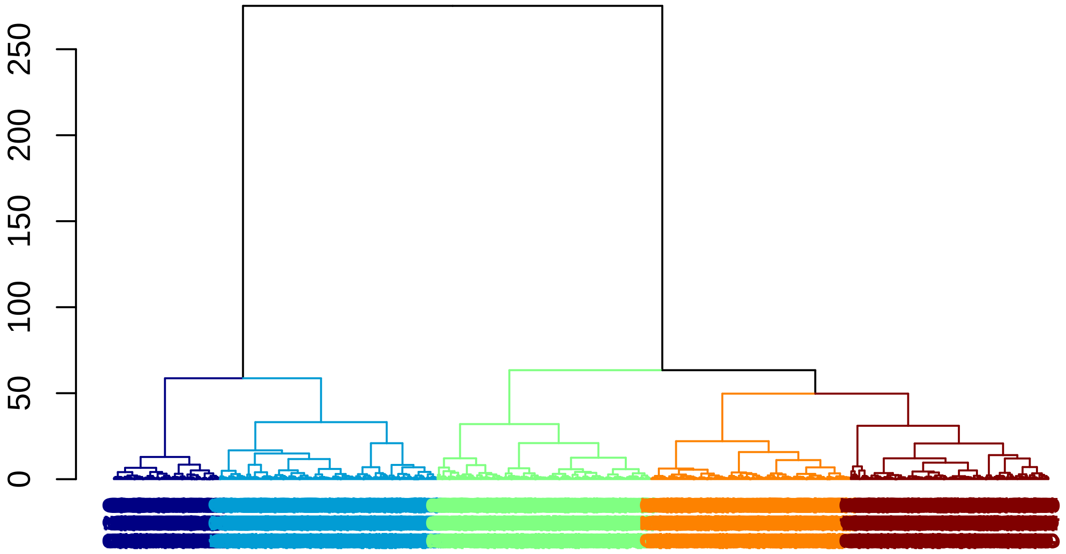
Part 1 - Cluster Analysis

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?

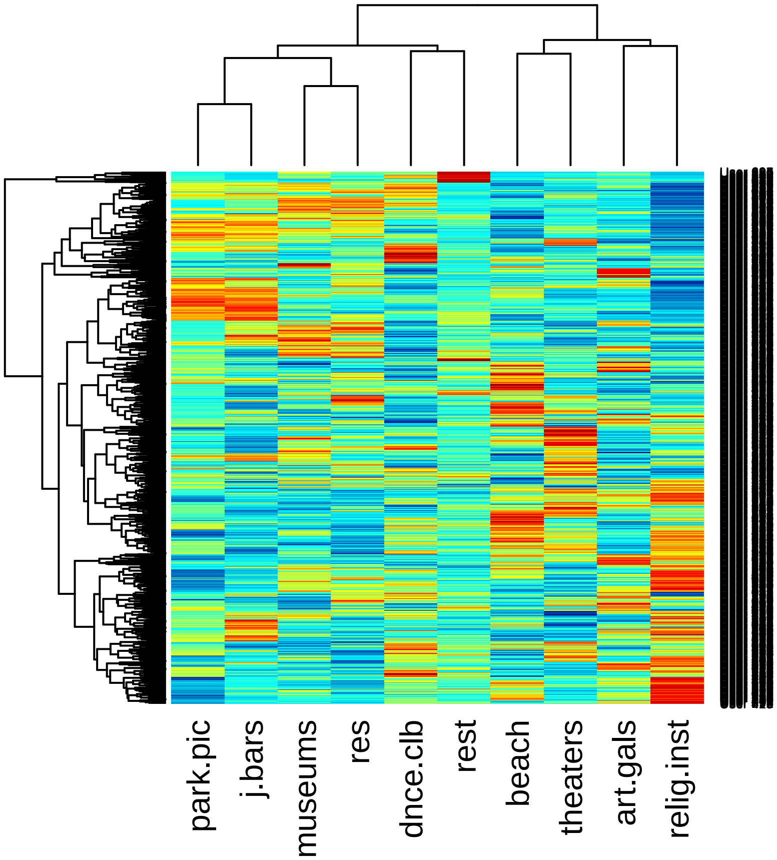
Table 1: Hierarchical Cluster Summary

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	0.78	1.30	2.38	0.55	0.99	1.93	3.19	2.80	1.51	2.56
2	1.00	1.16	1.23	0.50	0.84	1.72	3.18	2.84	1.41	2.84
3	0.84	1.51	0.37	0.64	1.10	2.09	3.18	2.86	1.71	2.82
4	0.88	1.26	0.37	0.44	0.62	1.34	3.18	2.85	1.58	3.01
5	0.95	1.61	1.65	0.60	1.41	2.54	3.19	2.80	1.62	2.54

Table 3: First 10 Row Cluster Assignment

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

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

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

Table 5: Consolidated Education Levels

ed.collapsed	count
HS-grad	10501
Some-college	7291
Bachelors	5355
Less than High School	4253
Post Bachelors	2712

ed.collapsed	count
Associates	2449

Section A

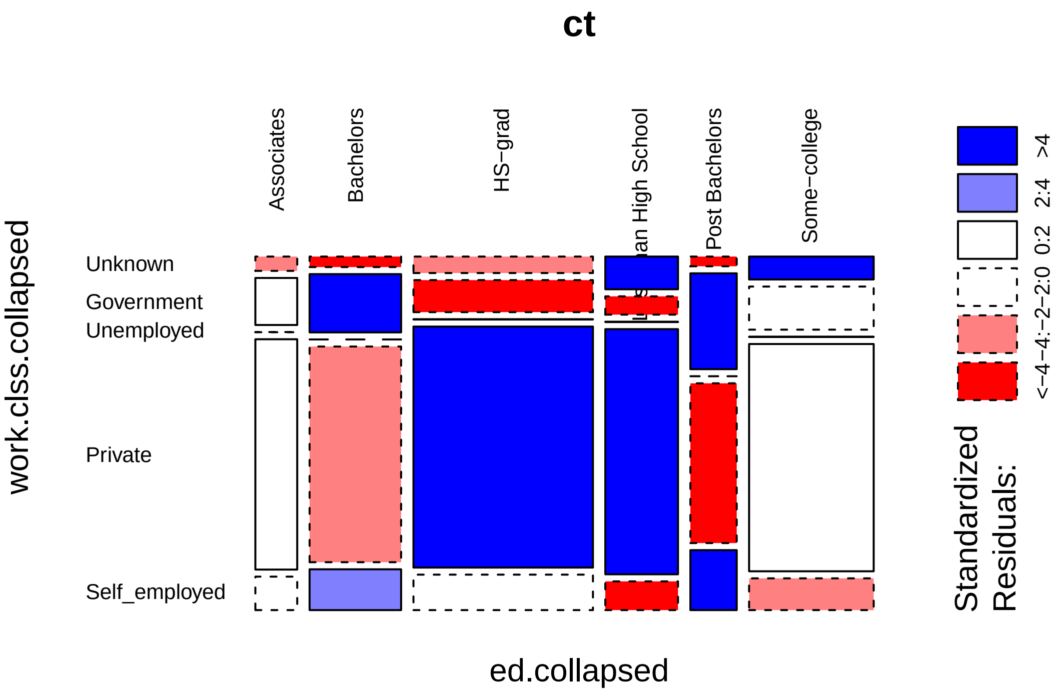
Are the variables WorkClass and education independent of each other? Perform a chi-square test of independence. Produce a mosaic plot of the data.

Table 6: Cross-tabulation of Education and Work Class

	Unknown	Government	Unemployed	Private	Self_employed
Associates	108	354	1	1734	252
Bachelors	173	959	0	3551	672
HS-grad	532	1034	10	7780	1145
Less than High School	428	239	5	3205	376
Post Bachelors	81	799	0	1332	500
Some-college	514	966	5	5094	712

```
## Warning in chisq.test(ct): Chi-squared approximation may be incorrect
```

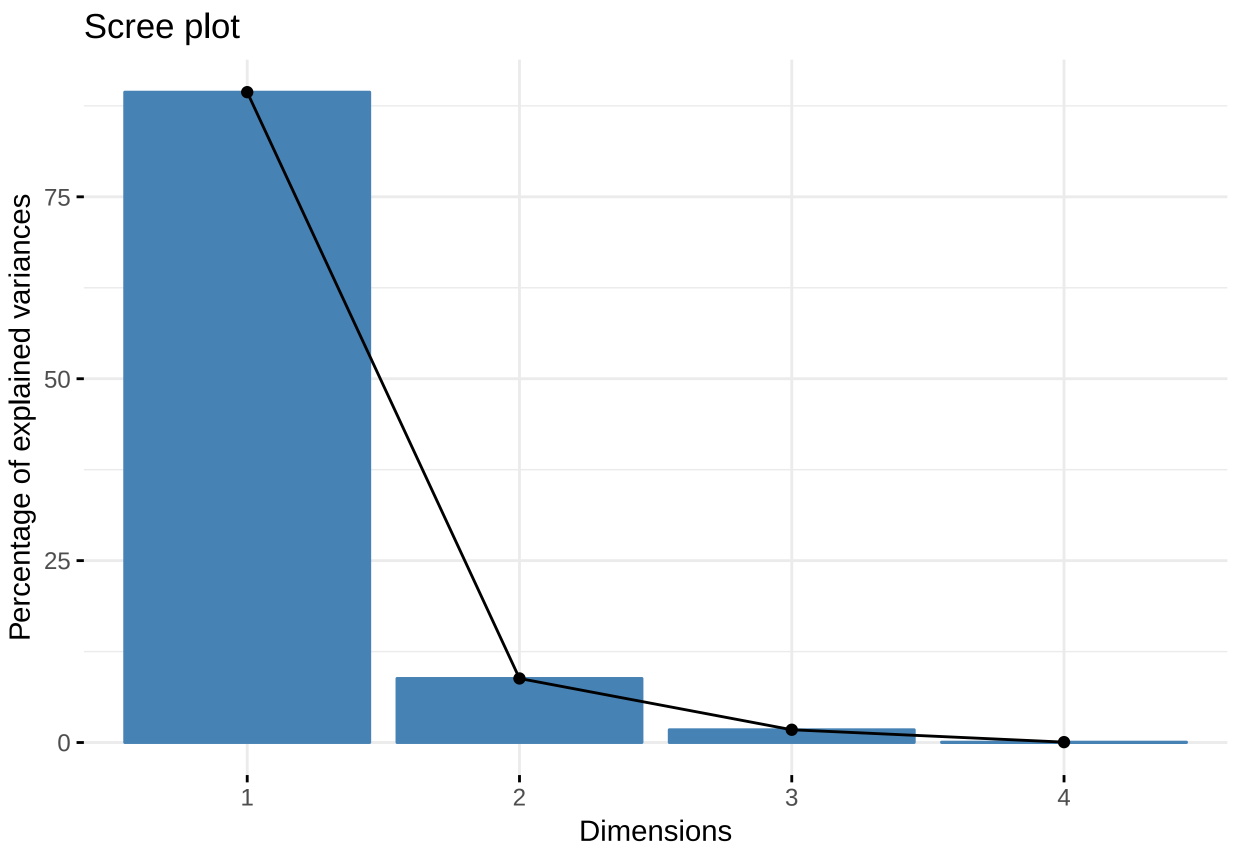
```
##
##  Pearson's Chi-squared test
##
## data:  ct
## X-squared = 1577.7, df = 20, p-value < 2.2e-16
```

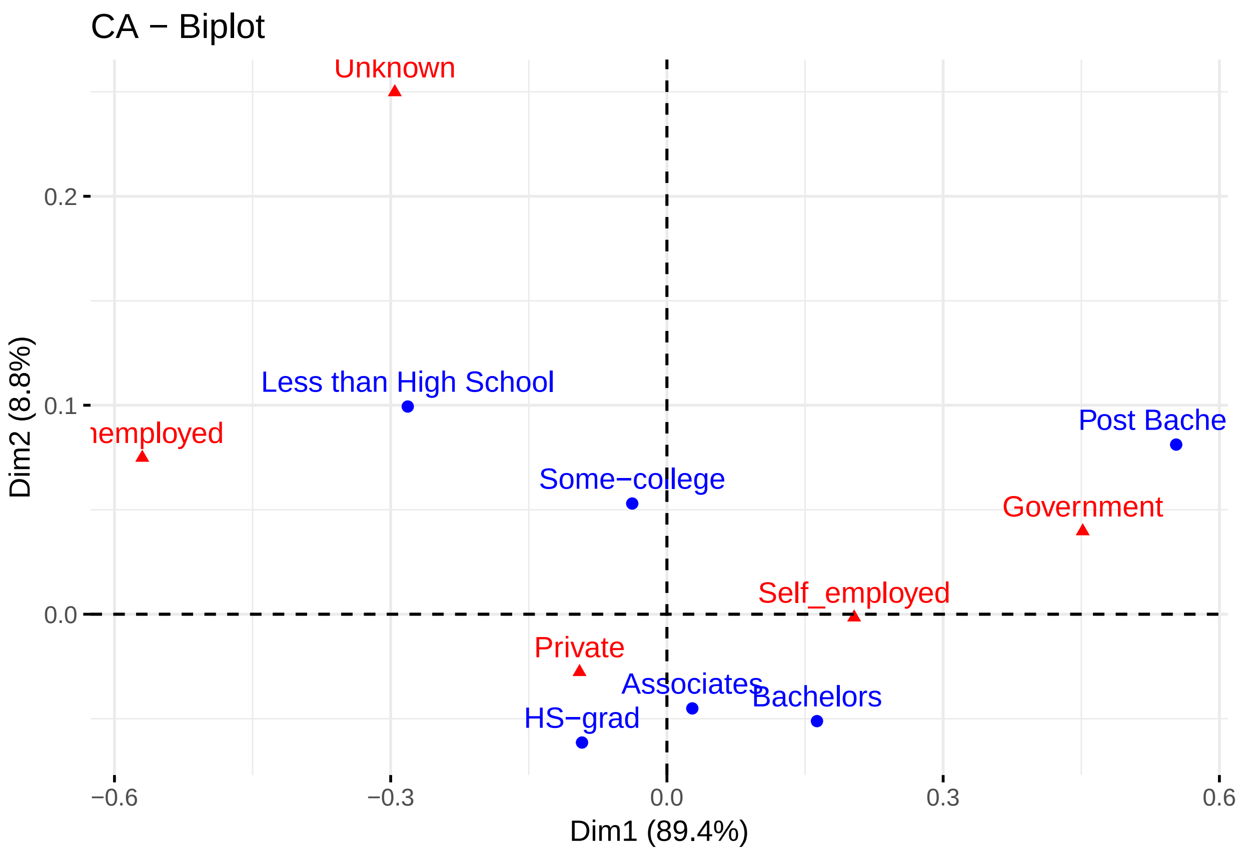


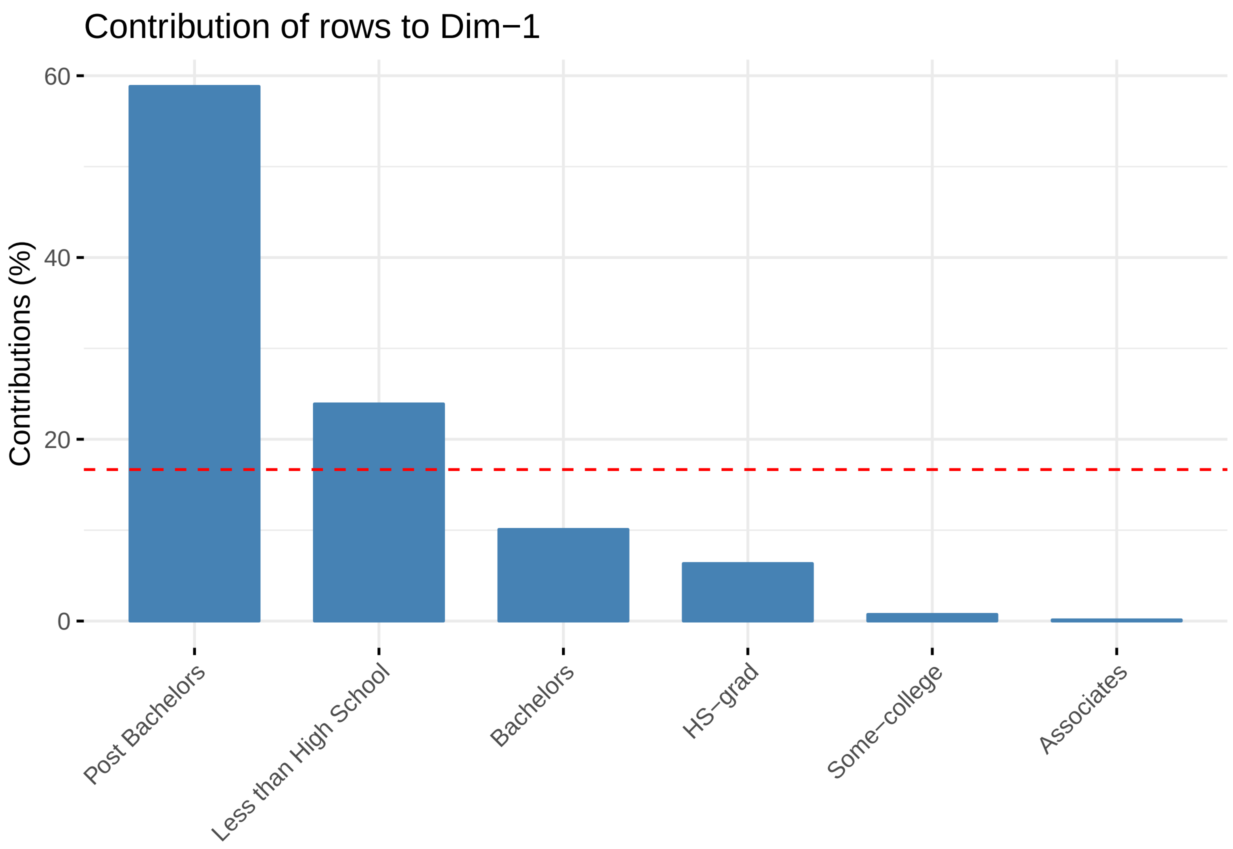
Section B

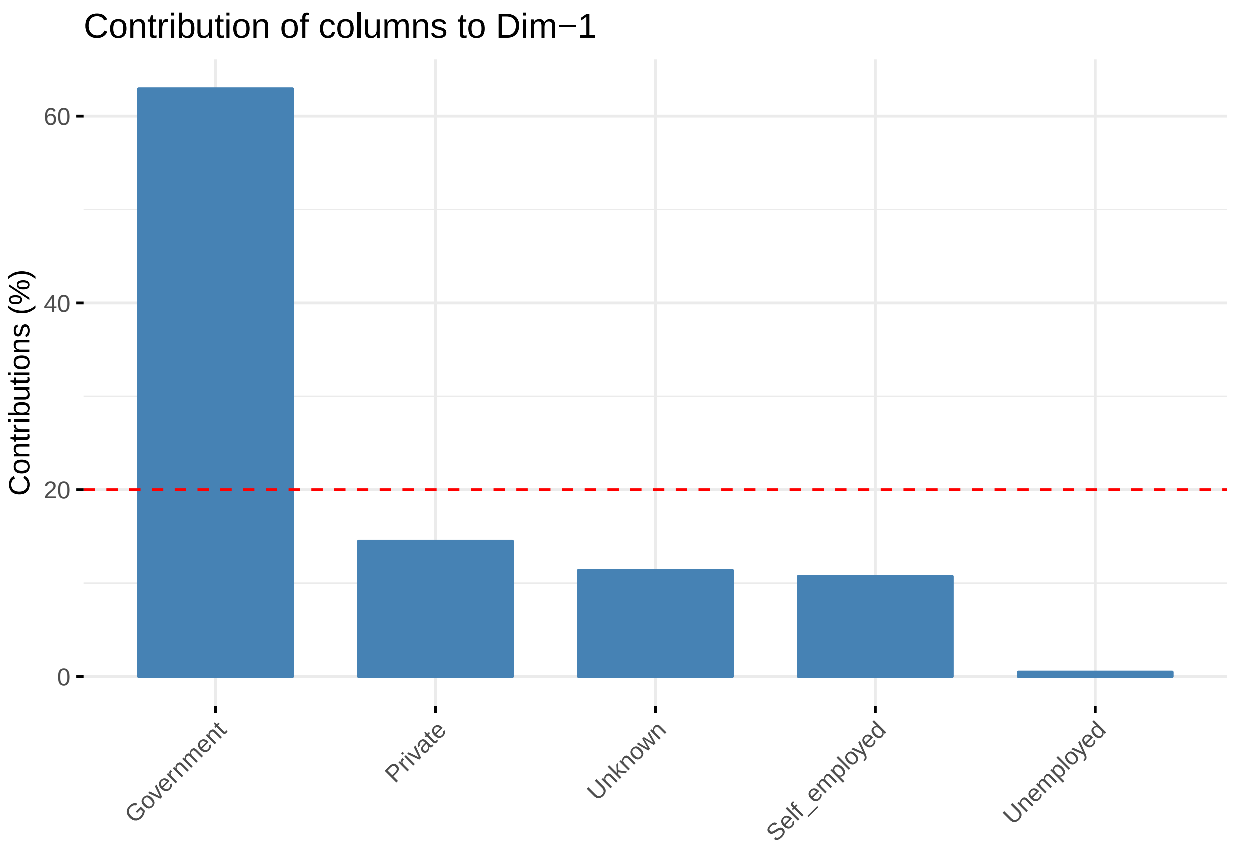
Perform a correspondence analysis on WorkClass and education. How many dimensions should be retained? Using the first two dimensions, construct a visualization of the correspondence analysis. Interpret what you see.

```
##
## Call:
## CA(X = ct, graph = FALSE)
##
## The chi square of independence between the two variables is equal to 1577.719 (p-value =  8.399116e-3
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4
## Variance               0.043   0.004   0.001   0.000
## % of var.             89.383   8.805   1.756   0.056
## Cumulative % of var.  89.383  98.188  99.944 100.000
##
## Rows
##                         Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr
## Associates          |     0.347 |  0.028  0.132  0.165 | -0.045  3.580
## Bachelors           |     4.866 |  0.163 10.083  0.897 | -0.051 10.087
## HS-grad             |     4.139 | -0.092  6.331  0.663 | -0.061 28.499
## Less than High School |    11.717 | -0.281 23.893  0.883 |  0.099 30.241
## Post Bachelors      |    26.162 |  0.553 58.827  0.974 |  0.081 12.864
## Some-college        |     1.223 | -0.038  0.734  0.260 |  0.053 14.729
##                        cos2    Dim.3    ctr   cos2
## Associates            0.440 | -0.043 16.120  0.395 |
## Bachelors             0.088 | -0.018  6.461  0.011 |
## HS-grad               0.294 |  0.023 20.909  0.043 |
## Less than High School  0.110 |  0.024  8.639  0.006 |
## Post Bachelors        0.021 |  0.040 15.871  0.005 |
## Some-college          0.514 | -0.035 32.000  0.223 |
##
## Columns
##                         Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr
## Unknown             |     8.445 | -0.296 11.372  0.583 |  0.250 82.483
## Government          |    27.611 |  0.452 62.927  0.987 |  0.040  4.939
## Unemployed          |     0.271 | -0.570  0.483  0.772 |  0.075  0.085
## Private             |     6.839 | -0.095 14.496  0.918 | -0.028 12.487
## Self_employed       |     5.287 |  0.203 10.722  0.878 | -0.002  0.006
##                        cos2    Dim.3    ctr   cos2
## Unknown               0.417 |  0.004  0.103  0.000 |
## Government            0.008 | -0.033 17.266  0.005 |
## Unemployed            0.013 |  0.223  3.760  0.118 |
## Private               0.078 | -0.006  3.313  0.004 |
## Self_employed         0.000 |  0.076 75.558  0.122 |
```


Scree plot
Percentage of explained variances
0
25
50
75
1
2
3
4
Dimensions


CA − Biplot
Unknown
Less than High School
ıemployed
Post Bache
Some−college
Government
Self_employed
Private
Associates
Bachelors
HS−grad
0.2
0.1
0.0
−0.6
−0.3
0.0
0.3
0.6
Dim1 (89.4%)
Dim2 (8.8%)


Contribution of rows to Dim−1
Contributions (%)
60
40
20
0
Post Bachelors
Less than High School
Bachelors
HS−grad
Some−college
Associates



Section C

Using all of the categorical variables in this data set, perform a multiple correspondence analysis. Summarize your results and interpret.