Alex Look

DSC 324

Homework 4

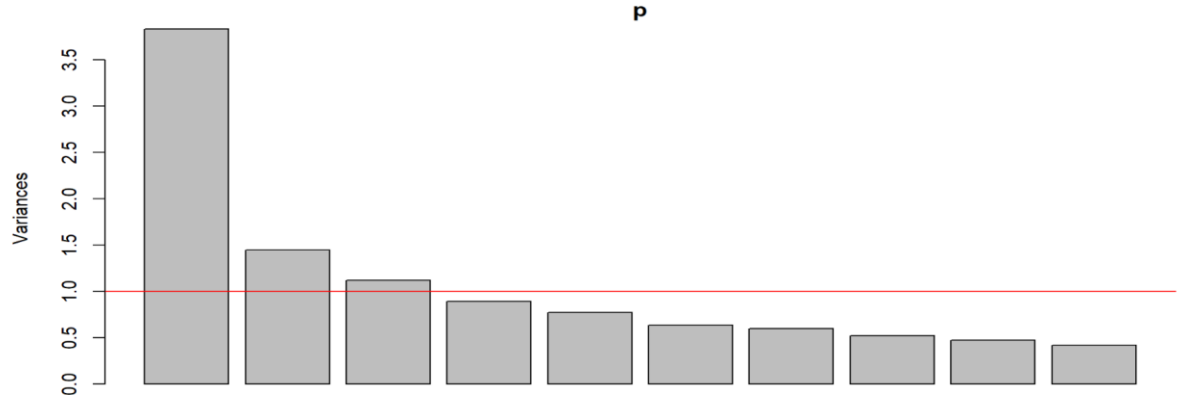
2/27/22

1a) Scaling or not for running PCA

I believe this data will not have to be scaled because the average values are normalized already. Each variable is not much different from one another. When trying to do analysis I have not run into problems when analyzing PCA.

1b)

```
Loadings:
          PC1    PC2    PC3
info       0.736
comp       0.754
arith      0.605
simil      0.744
vocab      0.744
digit      0.421 -0.470
pictcomp   0.559  0.467
parang     0.449
block      0.574  0.449
object     0.463  0.583
coding                0.864

                 PC1   PC2   PC3
SS loadings    3.829 1.442 1.116
Proportion Var 0.348 0.131 0.101
Cumulative Var 0.348 0.479 0.581
```

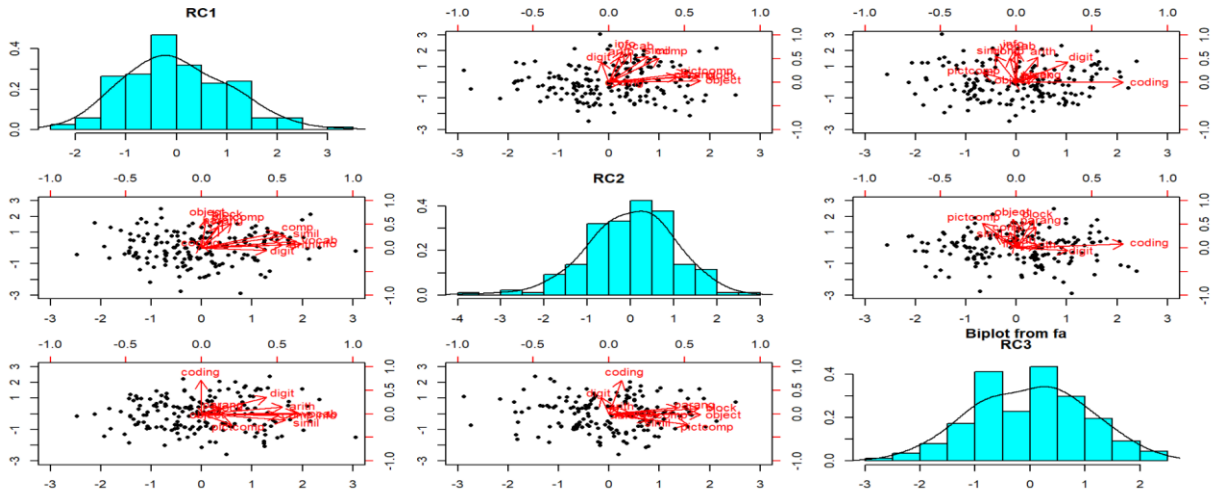


When running the output, it looks like there could be many variables that PC1 could take in about 10 from just an eye glance. In the code there is a .4 cutoff which PC1 currently is encapsulating all the

current variables except Coding. That Coding variable itself will be in PC3 which would be a single variable factor. What alarms me is that having 10 variables considered in PC1 seems like way too much when thinking about variance. Running it on a Scree plot, I can tell that there will be 3 PC's that will be enough to help capture the variance.

1c)

```
Fit based upon off diagonal values = 0.91
Loadings:
          RC1    RC2    RC3
info       0.826
comp       0.634  0.416
arith      0.669
simil      0.694
vocab      0.782
digit      0.535         0.428
pictcomp          0.649
parang            0.567
block             0.743
object            0.756
coding                   0.883

                 RC1   RC2   RC3
SS loadings    3.022 2.211 1.154
Proportion Var 0.275 0.201 0.105
Cumulative Var 0.275 0.476 0.581
```



After rotating the dataset, it moved pictcomp, parang, block, and object into PC2 meaning that PC1 now has 6 variables. Variance after the rotation went down from PC1 to 0.275, but increased PC2 0.201 and PC3 0.105 and cumulative variance stayed the same at 0.581 after PC3. The biplot helps show how for the 6 variables in PC1 how each variable is correlated to each other variable in the dataset. The histograms help show the variance it is capturing in each PC1-3.

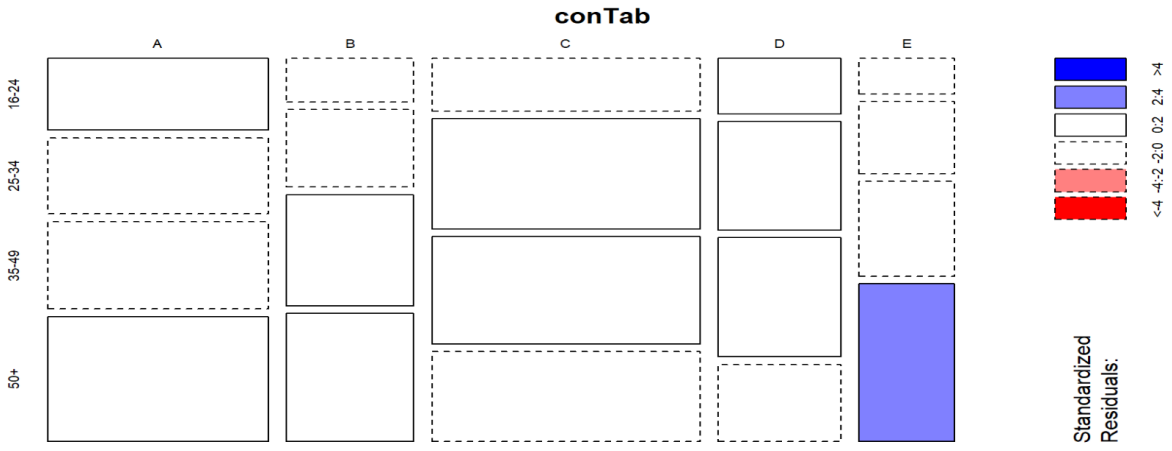
1d) What this is telling if a kid were to do well in information and vocabulary might not do well in parang or block. They excel in these subjects but struggle in the PC2 subjects. Coding was correlated with digit and pictcomp.

1e)

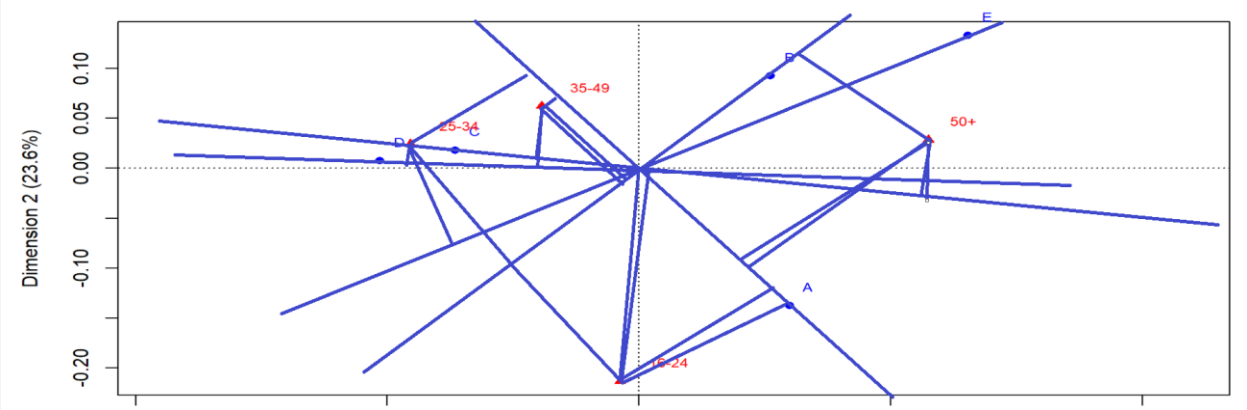
```
Loadings:
         Factor1 Factor2 Factor3
info      0.779   0.156
comp      0.551   0.449
arith     0.556   0.140   0.269
simil     0.620   0.366  -0.160
vocab     0.721   0.252
digit     0.431           0.134
pictcomp  0.202   0.605  -0.194
parang    0.154   0.392   0.135
block     0.117   0.714   0.380
object            0.573
coding                    0.290

                Factor1 Factor2 Factor3
SS loadings       2.399   1.801   0.410
Proportion Var    0.218   0.164   0.037
Cumulative Var    0.218   0.382   0.419

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 30.58 on 25 degrees of freedom.
The p-value is 0.203
```

In Factor 1 info and vocab have the highest correlation. In Factor 2 it was block then pictcomp dropped off behind it. In Factor 3 it was very low values but highest was block and lowest touched negatives in parang. When looking at the variance it is lower than form PC1 and all Factors were lower than in the PCs. Cumulative was lower in general than PCS too. The loadings in all Factors too were all lower than PCs version of SS loadings.

2a) Mosaic Plot



2b) Plot correspondence analysis. Two variables with highest/lowest correspondence?



First off apologies for this spider web of an answer just ran low on time. Would have done individual correspondence analysis with more time to look prettier and formal. The variables with the highest correspondences were C and age group 25-34. It seemed in each line that was made these two were always the closest and highest correspondence to each line. The variables with the lowest correspondences were A and age group 16-24. These I felt in most lines these were the ones most commonly that were the furthest from the lines.

2c) Age profile for the stores. What customer ages are most highly and least highly represented?

Store A- In Store A this is the most distributed age groups out of all the stores but did lean towards the older people. The lowest represented is the 16-24 age group and highest is 50+ age group.

Store B- In Store B the age profile is mainly from the 35-59 and 50+ age groups dominated the distribution, but this did lean more towards 50+ of the two. The lowest represented is the 16-24 age group and highest is 25-34 age group.

Store C- In Store C this heavily weighted in the middle age groups 25-34 and 35-49 while close this store was just ahead in the 25-34 age group. The lowest represented is the 16-24 age group and highest is 50+ age group.

Store D- In Store D this is another store with middle age groups with heavy weight in 35-34 and 35-49 but leaned of the two more towards 35-49. The lowest represented is the 16-24 age group and highest is 35-49 age group.

Store E- In Store E this store heavily is weighted on the 50+ age group while 35-49 was slightly behind. The lowest represented is the 16-24 age group and highest is 50+ age group.

3a)

```
   1 2 3 4 5 6 7
1  2 3 0 1 1 1 1
2  0 6 2 2 3 0 0
3  0 3 6 2 1 0 0
4  1 3 0 9 0 0 0
5  0 2 1 1 8 1 0
6  0 0 1 1 2 5 2
7  0 0 2 1 0 1 6
       Accuracy Prior Frequency.1 Prior Frequency.2
         0.5185            0.1111            0.1605
Prior Frequency.3 Prior Frequency.4 Prior Frequency.5
         0.1481            0.1605            0.1605
Prior Frequency.6 Prior Frequency.7
         0.1358            0.1235

Confusion Matrix
```

For one of the accuracy's, I got was 0.5185 just not sure for which one this could be for. I assume since it was the only one, I got it is for this question.

3b) Was not able to get this answer but I assume it would be got higher because once the data becomes smaller, I assume the performance would increase with lower amounts of observations.

3c) I believe a big misclassification error on bond trading would relate to if a bond company were to rate a low tiered class variable such as C but sell it off as a AAA, AA, or A. This is a problem because they are lying to potential buyers the value of this bond is higher than it is. This means investors are buying high value for a next to nothing bond. Comparing it to the opposite where if a bond's real value is higher than they put it at is not as damaging to the company and investors. The investor is buying a higher quality bond for cheaper than it should have been listed for. The company yes is losing potential revenue but is not getting complaints about how investors bought cheaper bonds. I would measure this by looking at one of the variables dealing with assets LASSLTD or LCURRAT because they will tell how much a bond might be holding with terms of if it is a profitable sell or not. There you can tell how much the bond is true value is at if it has been misclassified.