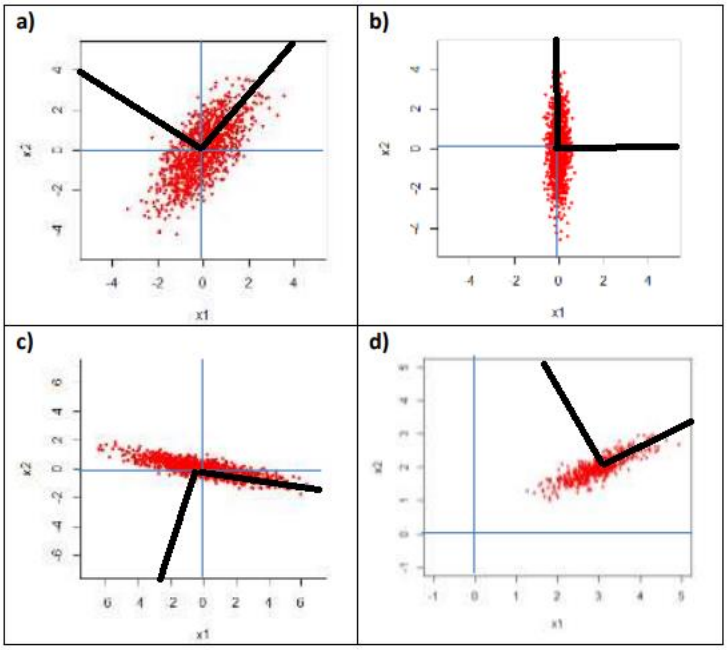
Alex Look

DSC 324

1/30/22

Assignment 3

1) Eigenvectors



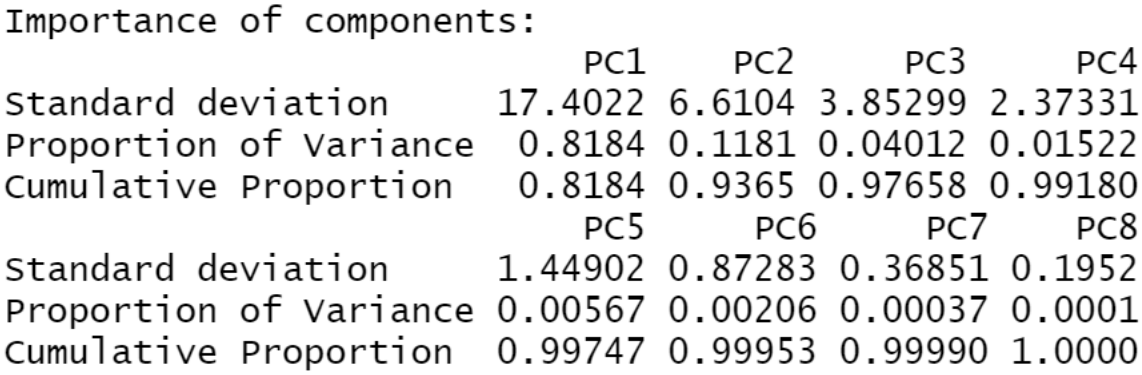
1a) V1= (4,4) V2= (-5,4)

1b) V1= (4,0) V2= (0,4)

1c) V1= (-2,6) V2= (4, -6)

1d) V1= (2,3) V2= (2,5)

2a)

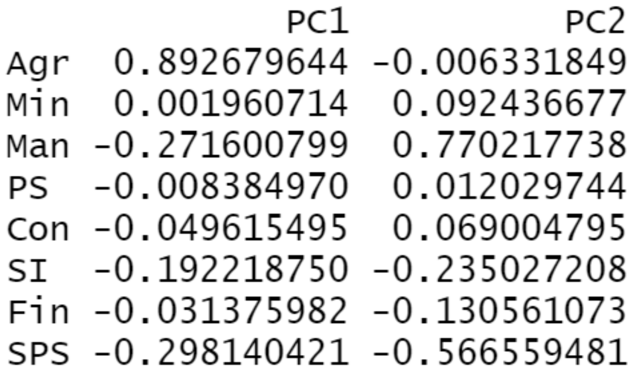


```
Importance of components:
                           PC1    PC2     PC3     PC4
Standard deviation     17.4022 6.6104 3.85299 2.37331
Proportion of Variance  0.8184 0.1181 0.04012 0.01522
Cumulative Proportion   0.8184 0.9365 0.97658 0.99180
                           PC5     PC6     PC7    PC8
Standard deviation     1.44902 0.87283 0.36851 0.1952
Proportion of Variance 0.00567 0.00206 0.00037 0.0001
Cumulative Proportion  0.99747 0.99953 0.99990 1.0000
```

When running the principal components, I found that since Country is a text variable it would have no meaning to the dataset when trying to find total variation, so it meant the dataset has 9 variables to run against instead of 10. I then ran the principal components and found that PC1 and PC2 due to their cumulative proportion got over the threshold to explain 90% of total variation. PC1=0.8184 and with PC2 it went up to 0.9365.

2b)

Before rotation



```
             PC1          PC2
Agr  0.892679644 -0.006331849
Min  0.001960714  0.092436677
Man -0.271600799  0.770217738
PS  -0.008384970  0.012029744
Con -0.049615495  0.069004795
SI  -0.192218750 -0.235027208
Fin -0.031375982 -0.130561073
SPS -0.298140421 -0.566559481
```

PC1= 0.890*Agr* + 0.001*Min* – 0.271*Man* – 0.008*PS* – 0.050*Con* – 0.190*SI* – 0.031*Fin* – 0.298*SPS*

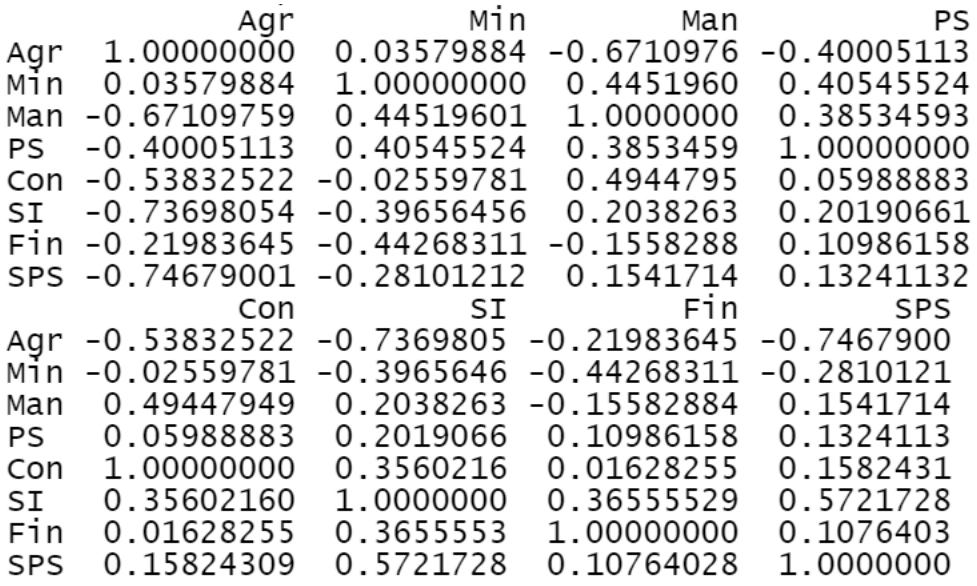
PC2= -0.006*Agr* + 0.092*Min* + 0.770*Man* + 0.0120*PS* + 0.070*Con* – 0.240*SI* – 0.130*Fin* – 0.567*SPS*

After rotation

2c) While I did not get to finish the one half of part b I can assume and predict that by rotating the data would make the ability to interpret components because it will allow for the maximum variance of a dataset. Without principal components the more it would have to explain variance.

2d) Per each principal competent for PC1 Turkey has the highest and the lowest is Yugoslavia. In PC2 the highest country is East Germany and the lowest in PC2 is Greece.

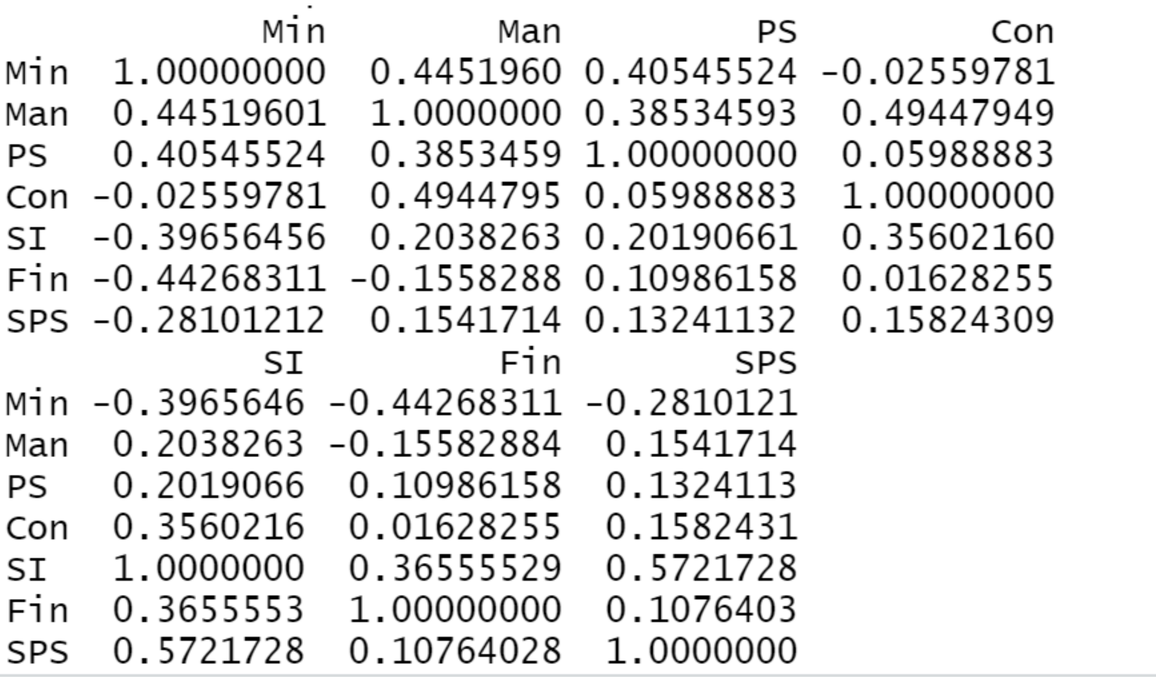
2e)



```
           Agr         Min        Man          PS
Agr  1.00000000  0.03579884 -0.6710976 -0.40005113
Min  0.03579884  1.00000000  0.4451960  0.40545524
Man -0.67109759  0.44519601  1.0000000  0.38534593
PS  -0.40005113  0.40545524  0.3853459  1.00000000
Con -0.53832522 -0.02559781  0.4944795  0.05988883
SI  -0.73698054 -0.39656456  0.2038263  0.20190661
Fin -0.21983645 -0.44268311 -0.1558288  0.10986158
SPS -0.74679001 -0.28101212  0.1541714  0.13241132
           Con         SI         Fin        SPS
Agr -0.53832522 -0.7369805 -0.21983645 -0.7467900
Min -0.02559781 -0.3965646 -0.44268311 -0.2810121
Man  0.49447949  0.2038263 -0.15582884  0.1541714
PS   0.05988883  0.2019066  0.10986158  0.1324113
Con  1.00000000  0.3560216  0.01628255  0.1582431
SI   0.35602160  1.0000000  0.36555529  0.5721728
Fin  0.01628255  0.3655553  1.00000000  0.1076403
SPS  0.15824309  0.5721728  0.10764028  1.0000000
```

When looking at all the variables and their correlation the highest correlated variable is SPS and SI at 0.5721728. I do not see any other highly correlate values very close to it the next value closest is PS and Man at 0.49447949. When thinking about threshold of 75% for correlated or uncorrelated I would say for highly correlated Man, PS, Con, SI, Fin, and SPS are all variables that have 75% correlated with the other fields. Agr is the only one with mainly 75% uncorrelated with 6 variables in the negatives.

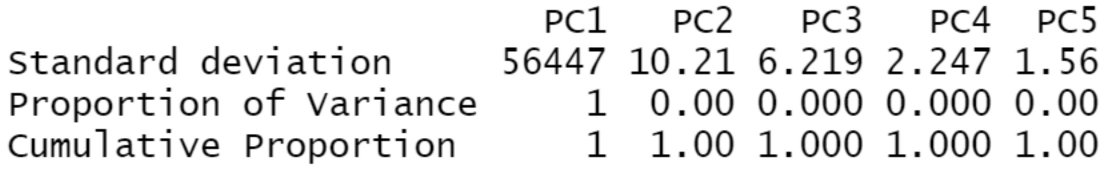
After removing Arg:



```
           Min       Man         PS         Con
Min  1.00000000 0.4451960 0.40545524 -0.02559781
Man  0.44519601 1.0000000 0.38534593  0.49447949
PS   0.40545524 0.3853459 1.00000000  0.05988883
Con -0.02559781 0.4944795 0.05988883  1.00000000
SI  -0.39656456 0.2038263 0.20190661  0.35602160
Fin -0.44268311 -0.1558288 0.10986158  0.01628255
SPS -0.28101212 0.1541714 0.13241132  0.15824309
           SI         Fin        SPS
Min -0.3965646 -0.44268311 -0.2810121
Man  0.2038263 -0.15582884  0.1541714
PS   0.2019066  0.10986158  0.1324113
Con  0.3560216  0.01628255  0.1582431
SI   1.0000000  0.36555529  0.5721728
Fin  0.3655553  1.00000000  0.1076403
SPS  0.5721728  0.10764028  1.0000000
```

After trying to remove Arg I noticed in Min while I did not see a change from the initial matrix to the updated one, I would predict that each value that is left would actually go in in correlation because when removing something uncorrelated would impact the data for the rest of the variables.

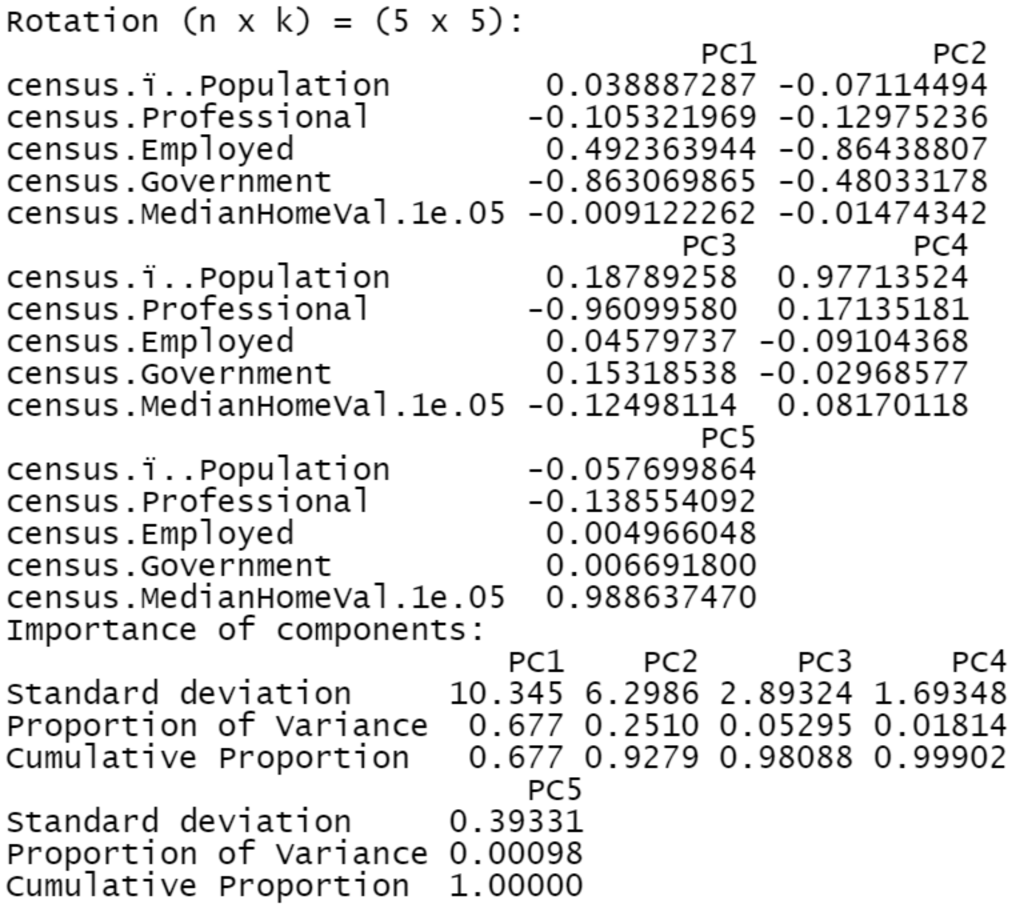
3a)



```
                          PC1   PC2   PC3   PC4  PC5
Standard deviation     56447 10.21 6.219 2.247 1.56
Proportion of Variance     1  0.00 0.000 0.000 0.00
Cumulative Proportion      1  1.00 1.000 1.000 1.00
```

When calculating the census principal component, the results show that in the first component it has a cumulative proportion and proportion of variance at 1 for both. That means just PC1 will only be needed here because it accounts for the entire dataset’s variance and that no other pc is needed. I believe this is the case because it is due to the low amount total of variables in the entire dataset that caused pc1 to have 1 in variance.

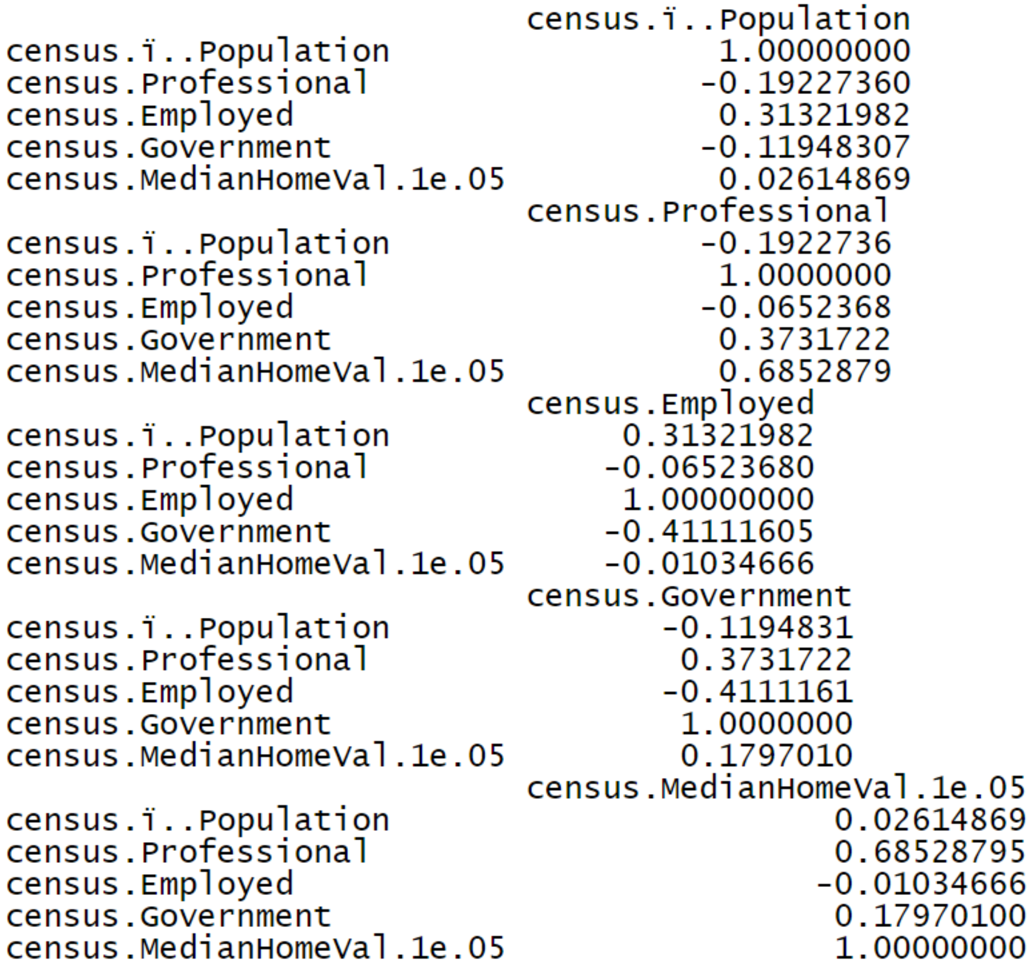
3b)



```
Rotation (n x k) = (5 x 5):
                                   PC1         PC2
census.ï..Population         0.038887287 -0.07114494
census.Professional         -0.105321969 -0.12975236
census.Employed              0.492363944 -0.86438807
census.Government           -0.863069865 -0.48033178
census.MedianHomeVal.1e.05  -0.009122262 -0.01474342
                                   PC3         PC4
census.ï..Population         0.18789258  0.97713524
census.Professional         -0.96099580  0.17135181
census.Employed              0.04579737 -0.09104368
census.Government            0.15318538 -0.02968577
census.MedianHomeVal.1e.05  -0.12498114  0.08170118
                                   PC5
census.ï..Population        -0.057699864
census.Professional         -0.138554092
census.Employed              0.004966048
census.Government            0.006691800
census.MedianHomeVal.1e.05   0.988637470
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     10.345 6.2986 2.89324 1.69348
Proportion of Variance  0.677 0.2510 0.05295 0.01814
Cumulative Proportion   0.677 0.9279 0.98088 0.99902
                           PC5
Standard deviation     0.39331
Proportion of Variance 0.00098
Cumulative Proportion  1.00000
```

After dividing MedianHomeValue by 100,000 and rerunning the prcomp function on the new value of MedianHomeValue I ended up with PC1 going down instead of 1, now it is 0.677 for PC1. To get to total variance of 1 I would need all 5 pcs.

3c)



```
                           census.ï..Population
census.ï..Population                 1.00000000
census.Professional                 -0.19227360
census.Employed                      0.31321982
census.Government                   -0.11948307
census.MedianHomeVal.1e.05           0.02614869
                           census.Professional
census.ï..Population                -0.1922736
census.Professional                  1.0000000
census.Employed                     -0.0652368
census.Government                    0.3731722
census.MedianHomeVal.1e.05           0.6852879
                           census.Employed
census.ï..Population            0.31321982
census.Professional            -0.06523680
census.Employed                 1.00000000
census.Government              -0.41111605
census.MedianHomeVal.1e.05     -0.01034666
                           census.Government
census.ï..Population              -0.1194831
census.Professional                0.3731722
census.Employed                   -0.4111161
census.Government                  1.0000000
census.MedianHomeVal.1e.05         0.1797010
                           census.MedianHomeVal.1e.05
census.ï..Population                       0.02614869
census.Professional                        0.68528795
census.Employed                           -0.01034666
census.Government                          0.17970100
census.MedianHomeVal.1e.05                 1.00000000
```

When computing the correlation matrix using PCA and comparing my results in part b I found that for Population there were fewer negative values compared to prcomp Population in PC1. The matrix had Government and Professional as negatives compared to in b where it had Professional, Government, and MedianHomeVal as negatives. Which means they more variables become positively correlated.

Professional in the matrix has two negative values at Population and Employed against the prcomp it had all negative variables. Running the matrix meant that two variables became positively correlated. MedianHomeVal after running it on the matrix actually gained one more variable that is positively affecting it instead of two in the prcomp.

For Employed the matrix had three negative correlated variables compared to in prcomp where it had two. This meant that it got worse after running it on the matrix. Same thing for MedianHomeVal that it had gained a negatively correlated variable after running it against the matrix.

3d) When looking at the correlation matrix and its significance per each variable Population and tis highest correlated variable to it was Employed at 0.31321982 which means it is a very low correlation. For Professional the highest correlation variable was MedianHomeVal at 0.6852879 which will turn out to be the highest correlated variable out of the entire dataset. Employed's highest variable correlated was Population at 0.31321982 which is low. Government's highest variable correlation was Professional at 0.3731722 higher than both Population and Employed but still not that high. Finally, MedianHomeVal was in line with Professional variable being the highest correlated variables which means that if a variable like Profession or Government at 0.17970100 will both have correlation values in their own correlation to other variables will be higher than compared to Employed where it is a negative correlation to MedianHomeVal.

3e) A correlation matrix will be able to tell us per each variable how correlated they are to the dependent variable. Each variable will be plotted and can be referenced directly to each variable and their connection. It may be more appropriate because of the fact that there might not be enough variance to capture. Correlation matrix describes variance just to the dependent variable.