DSC 424: Advanced Data Analysis and Regression

Assignment 03

Name: Sidhant Thakur

Student ID: 2020181

Problem 1

a)

```
> summary(p)
Importance of components:
                           PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     17.4200 6.6107 3.89966 2.37473 1.56314 1.02276 0.64873 0.25481 0.04373
Proportion of Variance  0.8158 0.1175 0.04088 0.01516 0.00657 0.00281 0.00113 0.00017 0.00001
Cumulative Proportion   0.8158 0.9333 0.97415 0.98931 0.99588 0.99869 0.99982 0.99999 1.00000
> 
```

The above table clearly shows that the first two principal components account for 90% of the total variation in the data.

b)

```
Rotation (n x k) = (9 x 9):
             PC1          PC2
Agr  0.891758406 -0.006826746
Min  0.001922618  0.092347069
Man -0.271271411  0.770269221
PS  -0.008388285  0.012015922
Con -0.049594016  0.068988571
SI  -0.191798409 -0.234416513
Fin -0.031128614 -0.130082403
SPS -0.298046310 -0.566777401
TC  -0.045364280 -0.009888386
```

Formula for PC1: 0.89Agr+0.001Min-0.27Man+ 0.008PS-0.049Con-0.19SI-0.03Fin-0.29SPS-0.04TC

PC1 has positive loadings from Agr, Min, and Negative loadings for the rest of the others.
Formula for PC2: -0.006Agr+0.092Min+0.77Man+ 0.012PS+0.068Con-0.23SI-0.13Fin-5.66SPS-0.009TC

Variables in PC2 are very near to zero, such as Agr, Min, PS, Con, Tc, and have negative loadings for Man, SI, Fin, and SPS.

For, PC1 as 89% loading is for Agriculture so, I think it’s a country where most of population is working in agriculture field.

For, PC2 as 77% loading is for manufacturing so, I think it’s a country where most of population is working in manufacturing field.

c)
For PC1, the highest value is for Turkey and lowest values is for United Kingdom.

```
> s[order(s$PC1), 1:2]
          PC1         PC2
9  -18.728675  -3.33178946
1  -17.516687  -4.92622849
21 -17.415527  10.73233092
16 -15.311975  -8.52674423
4  -14.393424   5.04749385
8  -13.900455  -9.72359023
17 -12.683839   9.77920054
7  -12.089752   2.33236877
2  -11.496688 -11.66176637
13 -10.972019  -8.85877780
3   -9.128686  -2.16828207
11  -6.837047  -3.97634061
10  -6.471418   3.35662962
6   -4.026684  -0.38889529
20  -3.246127   9.23467980
22   3.135737   4.98695108
19   4.156791   6.70685051
5    4.458174  -6.13156498
25   4.587043  -0.87197041
15   5.774973   6.15867547
14   9.403865  -0.08570061
23  13.315709   2.94482700
24  17.011336   9.12523022
12  25.427083  -1.80467718
26  34.832648   0.69274975
18  52.115644  -8.64165980
```

For Turkey, the principal component score is 52.11
For United Kingdom, the principal component score is -18.72

For PC2, the highest value is for Germany and lowest values is for Denmark.

```
> s[order(s$PC2), 1:2]
          PC1          PC2
2  -11.496688 -11.66176637
8  -13.900455  -9.72359023
13 -10.972019  -8.85877780
18  52.115644  -8.64165980
16 -15.311975  -8.52674423
5    4.458174  -6.13156498
1  -17.516687  -4.92622849
11  -6.837047  -3.97634061
9  -18.728675  -3.33178946
3   -9.128686  -2.16828207
12  25.427083  -1.80467718
25   4.587043  -0.87197041
6   -4.026684  -0.38889529
14   9.403865  -0.08570061
26  34.832648   0.69274975
7  -12.089752   2.33236877
23  13.315709   2.94482700
10  -6.471418   3.35662962
22   3.135737   4.98695108
4  -14.393424   5.04749385
15   5.774973   6.15867547
19   4.156791   6.70685051
24  17.011336   9.12523022
20  -3.246127   9.23467980
17 -12.683839   9.77920054
21 -17.415527  10.73233092
```

For Germany, the principal component score is 10.73

For Denmark, the principal component score is -11.66

d)


```
Standard deviations (1, .., p=6):
[1] 17.1006723  6.3992852  2.4740394  1.3304655  0.9427856  0.2630165

Rotation (n x k) = (6 x 6):
             PC1         PC2         PC3         PC4         PC5         PC6
Agr  0.907236925 -0.02380097  0.39349302 -0.11433257  0.08374913 -0.03784267
Man -0.283671019 -0.75775839  0.57930475 -0.02436638  0.09357491 -0.01972311
PS  -0.008639003 -0.01203248 -0.01557201  0.04005789 -0.22207900 -0.97396837
Con -0.050712144 -0.07366257 -0.24501037 -0.79415019  0.52860098 -0.14791384
SPS -0.302604481  0.64717409  0.65917094 -0.09623445  0.20357532 -0.06622625
TC  -0.047287316  0.02881811  0.12161559 -0.58719172 -0.78361268  0.15264378
> summary(p)
Importance of components:
                           PC1    PC2     PC3     PC4    PC5    PC6
Standard deviation     17.1007 6.3993 2.47404 1.33047 0.9428 0.2630
Proportion of Variance  0.8545 0.1197 0.01789 0.00517 0.0026 0.0002
Cumulative Proportion   0.8545 0.9741 0.99203 0.99720 0.9998 1.0000
```

After, removing highly uncorrelated variable fields, the new formula is

Formula for PC1: 0.90Agr-0.28Man-0.008PS-0.05Con-0.19SI-0.30SPS-0.04TC

Formula for PC2: **-**0.023Agr-0.75Man- 0.012PS-0.07Con-0.64SPS-0.02TC

Problem 2

a)

```
> head(Census2)
  ï..Population Professional Employed Government MedianHomeVal
1          2.67         5.71    69.02       30.3        148000
2          2.25         4.37    72.98       43.3        144000
3          3.12        10.27    64.94       32.0        211000
4          5.14         7.44    71.29       24.5        185000
5          5.54         9.25    74.94       31.0        223000
6          5.04         4.84    53.61       48.2        160000
> p1 = prcomp(Census2)
> print(p1)
Standard deviations (1, .., p=5):
[1] 56446.885008    10.206857     6.218887     2.246707     1.559823

Rotation (n x k) = (5 x 5):
                        PC1           PC2           PC3           PC4           PC5
ï..Population  8.537905e-07 -4.108282e-02 -7.059713e-02  4.826860e-01  8.719762e-01
Professional   3.775797e-05  7.080539e-02 -7.460074e-02 -8.714029e-01  4.796648e-01
Employed      -1.367095e-06 -5.126328e-01 -8.542663e-01 -1.524163e-02 -8.487872e-02
Government     3.004471e-05  8.546967e-01 -5.095880e-01  8.624903e-02 -4.873218e-02
MedianHomeVal  1.000000e+00 -2.901832e-05  1.701961e-05  2.987813e-05 -1.750755e-05
> summary(p1)
Importance of components:
                         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
>
```

The above picture it shows that the first principal components account for 100% of the total variation in the data.

```
> summary(Census2)
 ï..Population   Professional       Employed       Government     MedianHomeVal
 Min.   :1.360   Min.   : 0.720   Min.   :49.50   Min.   :16.30   Min.   : 93000
 1st Qu.:3.120   1st Qu.: 1.670   1st Qu.:66.42   1st Qu.:20.60   1st Qu.:130000
 Median :4.720   Median : 3.380   Median :71.30   Median :24.40   Median :149000
 Mean   :4.469   Mean   : 3.962   Mean   :71.42   Mean   :26.91   Mean   :163557
 3rd Qu.:5.760   3rd Qu.: 4.830   3rd Qu.:77.33   3rd Qu.:31.00   3rd Qu.:178000
 Max.   :9.210   Max.   :16.700   Max.   :86.54   Max.   :68.50   Max.   :364000
>
```

So, this is happening because, when I looked at the data summary, I noticed that the maximum of median home value is very high when compared to another variable. In other words, it varies more as compared to another variables.

b)

After diving the MedianHomeval by 100,000, I have following summary. Here as we can see, data very less as compared to previous one

```
> newdata=cbind(Census2,d1)
> head(newdata)
  ï..Population Professional Employed Government MedianHomeVal MedianHomeVal
1          2.67         5.71    69.02       30.3        148000          1.48
2          2.25         4.37    72.98       43.3        144000          1.44
3          3.12        10.27    64.94       32.0        211000          2.11
4          5.14         7.44    71.29       24.5        185000          1.85
5          5.54         9.25    74.94       31.0        223000          2.23
6          5.04         4.84    53.61       48.2        160000          1.60
> d2<-newdata[-5]
> View(d2)
> summary(d2)
 ï..Population    Professional      Employed       Government     MedianHomeVal
 Min.   :1.360   Min.   : 0.720   Min.   :49.50   Min.   :16.30   Min.   :0.930
 1st Qu.:3.120   1st Qu.: 1.670   1st Qu.:66.42   1st Qu.:20.60   1st Qu.:1.300
 Median :4.720   Median : 3.380   Median :71.30   Median :24.40   Median :1.490
 Mean   :4.469   Mean   : 3.962   Mean   :71.42   Mean   :26.91   Mean   :1.636
 3rd Qu.:5.760   3rd Qu.: 4.830   3rd Qu.:77.33   3rd Qu.:31.00   3rd Qu.:1.780
 Max.   :9.210   Max.   :16.700   Max.   :86.54   Max.   :68.50   Max.   :3.640
```

Applying PCA on the new dataset

```
> p2 = prcomp(d2)
> print(p2)
Standard deviations (1, .., p=5):
[1] 10.3448177  6.2985820  2.8932449  1.6934798  0.3933104

Rotation (n x k) = (5 x 5):
                       PC1         PC2         PC3         PC4          PC5
ï..Population  0.038887287 -0.07114494  0.18789258  0.97713524 -0.057699864
Professional  -0.105321969 -0.12975236 -0.96099580  0.17135181 -0.138554092
Employed       0.492363944 -0.86438807  0.04579737 -0.09104368  0.004966048
Government    -0.863069865 -0.48033178  0.15318538 -0.02968577  0.006691800
MedianHomeVal -0.009122262 -0.01474342 -0.12498114  0.08170118  0.988637470
> summary(p2)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     10.345 6.2986 2.89324 1.69348 0.39331
Proportion of Variance  0.677 0.2510 0.05295 0.01814 0.00098
Cumulative Proportion   0.677 0.9279 0.98088 0.99902 1.00000
>
```

The first principal components account for 67.7% of the total variation in the data. The second principal components account for 92.7% of the total variation in the data.
The third principal components account for 98% of the total variation in the data and the fourth principal components account for 99% of the total variation in the data and the last principal components account for 100% of the total variation in the data.
So, variation in first principal components decrease to 67.7% from the 100% i.e., when we don't divide medianvalue by 10000

c) PCA with the correlation matrix

```
> p3 = prcomp(Census2, scale=T)
> print(p3)
Standard deviations (1, .., p=5):
[1] 1.4113534 1.1694129 0.9296006 0.7314787 0.4912604

Rotation (n x k) = (5 x 5):
                     PC1        PC2         PC3        PC4        PC5
ï..Population  0.2625829 -0.4629936  0.78390268 -0.2169291  0.2347882
Professional  -0.5933541 -0.3256442 -0.16407255  0.1446471  0.7028828
Employed       0.3256978 -0.6051419 -0.22487455  0.6628689 -0.1943206
Government    -0.4792022  0.2524850  0.55070086  0.5716730 -0.2766497
MedianHomeVal -0.4932213 -0.4996473 -0.06882436 -0.4072024 -0.5801162
> summary(p3)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.4114 1.1694 0.9296 0.7315 0.49126
Proportion of Variance 0.3984 0.2735 0.1728 0.1070 0.04827
Cumulative Proportion  0.3984 0.6719 0.8447 0.9517 1.00000
```

The first principal components account for 40% of total data variation, the second principal components for 67.1 percent of total data variation, the third principal components for 84.4 percent of total data variation, the fourth principal components for 95% of total data variation, and the final principal components for 100% of total data variation.

When compared to the answer in b, the first principal component is 40% of total variation.

d)

Scaling refers to getting all of the data into the same range.

Because in the problem's data has varying scales. As a result, we employed standardization to bring them all to the same scale.

As a result, it is suitable for usage in this context.

Problem 3

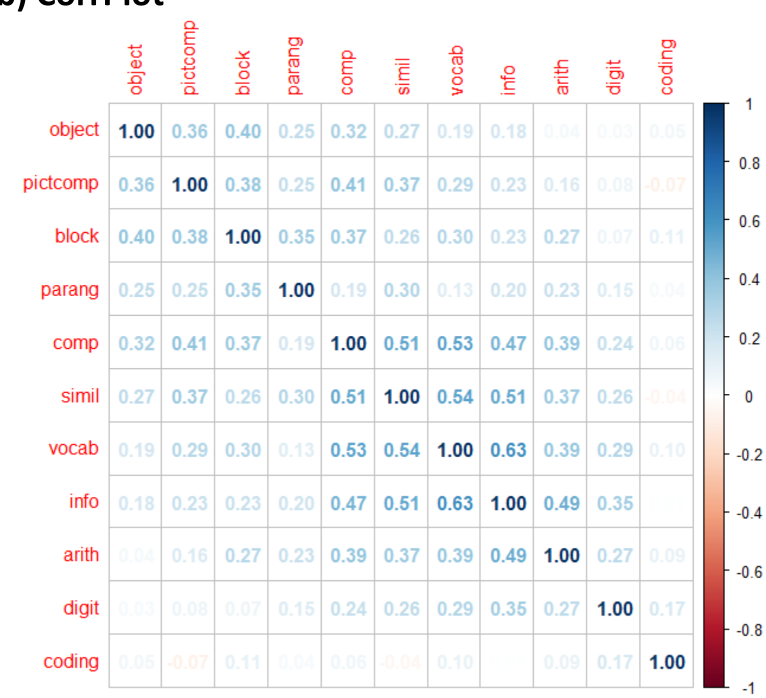
a)

```
> d<-wiscsem[,-c(1,2)]
> head(d)
  info comp arith simil vocab digit pictcomp parang block object coding
1    8    7    13     9    12     9        6     11    12      7      9
2    9    6     8     7    11    12        6      8     7     12     14
3   13   18    11    16    15     6       18      8    11     12      9
4    8   11     6    12     9     7       13      4     7     12     11
5   10    3     8     9    12     9        7      7    11      4     10
6   11    7    15    12    10    12        6     12    10      5     10
> summary(d)
      info             comp         arith           simil           vocab           digit       
 Min.   : 3.000   Min.   : 0   Min.   : 4.0   Min.   : 2.00   Min.   : 2.0   Min.   : 0.000  
 1st Qu.: 8.000   1st Qu.: 8   1st Qu.: 7.0   1st Qu.: 9.00   1st Qu.: 9.0   1st Qu.: 7.000  
 Median :10.000   Median :10   Median : 9.0   Median :11.00   Median :10.0   Median : 8.000  
 Mean   : 9.497   Mean   :10   Mean   : 9.0   Mean   :10.61   Mean   :10.7   Mean   : 8.731  
 3rd Qu.:11.500   3rd Qu.:12   3rd Qu.:10.5   3rd Qu.:12.00   3rd Qu.:12.0   3rd Qu.:11.000  
 Max.   :19.000   Max.   :18   Max.   :16.0   Max.   :18.00   Max.   :19.0   Max.   :16.000  
    pictcomp         parang          block           object         coding      
 Min.   : 2.00   Min.   : 2.00   Min.   : 2.00   Min.   : 3.0   Min.   : 0.000  
 1st Qu.: 9.00   1st Qu.: 9.00   1st Qu.: 9.00   1st Qu.: 9.0   1st Qu.: 6.000  
 Median :11.00   Median :10.00   Median :10.00   Median :11.0   Median : 9.000  
 Mean   :10.68   Mean   :10.37   Mean   :10.31   Mean   :10.9   Mean   : 8.549  
 3rd Qu.:13.00   3rd Qu.:12.00   3rd Qu.:12.00   3rd Qu.:13.0   3rd Qu.:11.000  
 Max.   :19.00   Max.   :17.00   Max.   :18.00   Max.   :19.0   Max.   :15.000  
> 
```

```r
d<-wiscsem[,(3:13)]
D <- as.data.frame(d)
head(D)
summary(D)
library(corrplot)
corrplot(cor(d),method = 'number',order='AOE') # 2-3 groups

p1 = prcomp(D,scale. = T) # scaled since all features were in same range.
print(p1)
summary(p1)
plot(p1)
abline(1, 0, col="red") # 3 groups

#########PFA###
library(psych)
p2 = principal(D, rotate="varimax", nfactors=3)

print(p2$loadings, cutoff=.4)
```

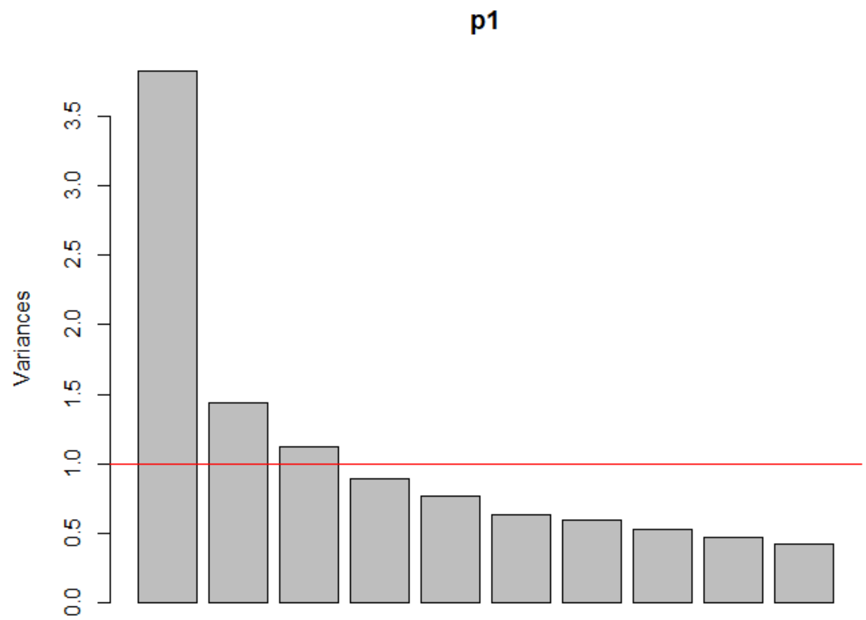
Yes, I scaled the data since all features are in the same range, so scaling will not affect these features.

b) CorrPlot



From, the corrplot, the appropriate number of factors to extract is 2 or 3.

PCA



From, the scree plot, the appropriate number of factors to extract is 3.

C)

```
> p2 = principal(D, rotate="varimax", nfactors=3)
>
> print(p2$loadings, cutoff=.4)

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
>
```

No, there aren't any variables that are likely to be single-variable factors.

We can see from the above output that RC1 has 6 loadings, RC2 has 5 loadings, and RC3 has 2 loadings, indicating that there is no single-variable factor in all three Rotated components.

d)

```
> p2 = principal(D, rotate="varimax", nfactors=3)
>
> print(p2$loadings, cutoff=.4)

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
>
```

By performing PFA, we can separate data into groups, making it easier to interpret.

For RC1, I believe that the children in this group have a good ability to think, which means that they have a good understanding of thoughts.

For RC2, I believe that the children in this group have a good understanding of the design concept, as evidenced by their ability to easily interprets good design or arrangement.

For RC3, I believe the children in this group have good logical or memorizing skills.

Furthermore, RC1 contributes 27.5 percent of the total variance, RC2 contributes 47.6 percent of the total variance, and RC3 contributes 58percent of the total variance.

e)

CFA

```
> ###############CFA###############
> fit = factanal(D, 3)
> print(fit$loadings, cutoff=.4,sort = T)

Loadings:
         Factor1 Factor2 Factor3
info      0.779
comp      0.551   0.449
arith     0.556
simil     0.620
vocab     0.721
pictcomp          0.605
block             0.714
object            0.573
digit     0.431
parang
coding

               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
```

PFA

```
> p2 = principal(D, rotate="varimax", nfactors=3)
>
> print(p2$loadings, cutoff=.4)

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
>
```

On comparing the loading of RC1 for **PFA** and **CFA**, I see that loading value in PFA is higher than CFA.
The info in PFA loading has value of 82.6%, whereas in CFA it is only 78%. Likewise for all other variables the value got reduced.

Furthermore, RC1 for PFA contributes 27.5 percent of the total variation whereas RC1 for CFA contributes to only 21.8%

Furthermore, RC2 for PFA contributes 47.6 percent of the total variation whereas RC1 for CFA contributes to only 38.2%

Furthermore, RC1 for PFA contributes 58.1 percent of the total variation whereas RC1 for CFA contributes to only 41.9%