615 Midterm Project

Jingning Yang

10/17/2019

Read the "World Value" data

```
library(readxl)
data <- read_excel("/Users/jingningyang/Desktop/F00007693-WV6_Data_United_States_2011_Excel_v20180912.x
```

Clean Data

Delete columns which has replicated values(non-information):

After we delete non-information columns, based on different data structure, we only select totally has 4 options and sounds relative to each other: I pick V4-9, V60, V62, V64 from about 298 variables(survey questions).

Convert selected survey questions from character to numeric:

```
x <- c(1:8,57,59,61) #57,59,61,,99:104
newdata[x] <- apply(newdata[x],2,function(x) as.numeric(as.factor(x)))
```

Preparation for factor analysis

Calculate correlations between selected survey questions:

```
matrix <- cor(newdata[x])
head(round(matrix,2))
```

```
##                                     V4: Important in life: Family
## V4: Important in life: Family                                1.00
## V5: Important in life: Friends                               0.38
## V6: Important in life: Leisure time                          0.20
## V7: Important in life: Politics                              0.19
## V8: Important in life: Work                                  0.19
## V9: Important in life: Religion                              0.32
##                                     V5: Important in life: Friends
## V4: Important in life: Family                                0.38
## V5: Important in life: Friends                               1.00
## V6: Important in life: Leisure time                          0.33
## V7: Important in life: Politics                              0.21
## V8: Important in life: Work                                  0.08
## V9: Important in life: Religion                              0.18
##                                     V6: Important in life: Leisure time
## V4: Important in life: Family                                      0.20
## V5: Important in life: Friends                                     0.33
## V6: Important in life: Leisure time                                1.00
## V7: Important in life: Politics                                    0.21
## V8: Important in life: Work                                        0.18
## V9: Important in life: Religion                                    0.05
##                                     V7: Important in life: Politics
## V4: Important in life: Family                                  0.19
## V5: Important in life: Friends                                 0.21
```

```
## V6: Important in life: Leisure time                          0.21
## V7: Important in life: Politics                              1.00
## V8: Important in life: Work                                  0.19
## V9: Important in life: Religion                              0.19
##                                     V8: Important in life: Work
## V4: Important in life: Family                              0.19
## V5: Important in life: Friends                             0.08
## V6: Important in life: Leisure time                        0.18
## V7: Important in life: Politics                            0.19
## V8: Important in life: Work                                1.00
## V9: Important in life: Religion                            0.16
##                                     V9: Important in life: Religion
## V4: Important in life: Family                                0.32
## V5: Important in life: Friends                               0.18
## V6: Important in life: Leisure time                          0.05
## V7: Important in life: Politics                              0.19
## V8: Important in life: Work                                  0.16
## V9: Important in life: Religion                              1.00
##                                     V10: Feeling of happiness
## V4: Important in life: Family                            0.30
## V5: Important in life: Friends                           0.25
## V6: Important in life: Leisure time                      0.21
## V7: Important in life: Politics                          0.16
## V8: Important in life: Work                              0.12
## V9: Important in life: Religion                          0.18
##                                     V11: State of health (subjective)
## V4: Important in life: Family                                 -0.01
## V5: Important in life: Friends                                 0.05
## V6: Important in life: Leisure time                            0.08
## V7: Important in life: Politics                                0.02
## V8: Important in life: Work                                    0.09
## V9: Important in life: Religion                                0.01
##                                     V60: Aims of country: first choice
## V4: Important in life: Family                                 -0.13
## V5: Important in life: Friends                                -0.07
## V6: Important in life: Leisure time                           -0.06
## V7: Important in life: Politics                               -0.11
## V8: Important in life: Work                                   -0.01
## V9: Important in life: Religion                               -0.07
##                                     V62: Aims of respondent: first choice
## V4: Important in life: Family                                    -0.05
## V5: Important in life: Friends                                   -0.01
## V6: Important in life: Leisure time                              -0.10
## V7: Important in life: Politics                                   0.09
## V8: Important in life: Work                                      -0.03
## V9: Important in life: Religion                                  -0.05
##                                     V64: Most important: first choice
## V4: Important in life: Family                                -0.08
## V5: Important in life: Friends                               -0.09
## V6: Important in life: Leisure time                          -0.04
## V7: Important in life: Politics                              -0.04
## V8: Important in life: Work                                   0.00
## V9: Important in life: Religion                              -0.02
```

```
library(psych)
cortest.bartlett(newdata[x])
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 2272.999
##
## $p.value
## [1] 0
##
## $df
## [1] 55
```

```
det(matrix)
```

```
## [1] 0.3602762
```

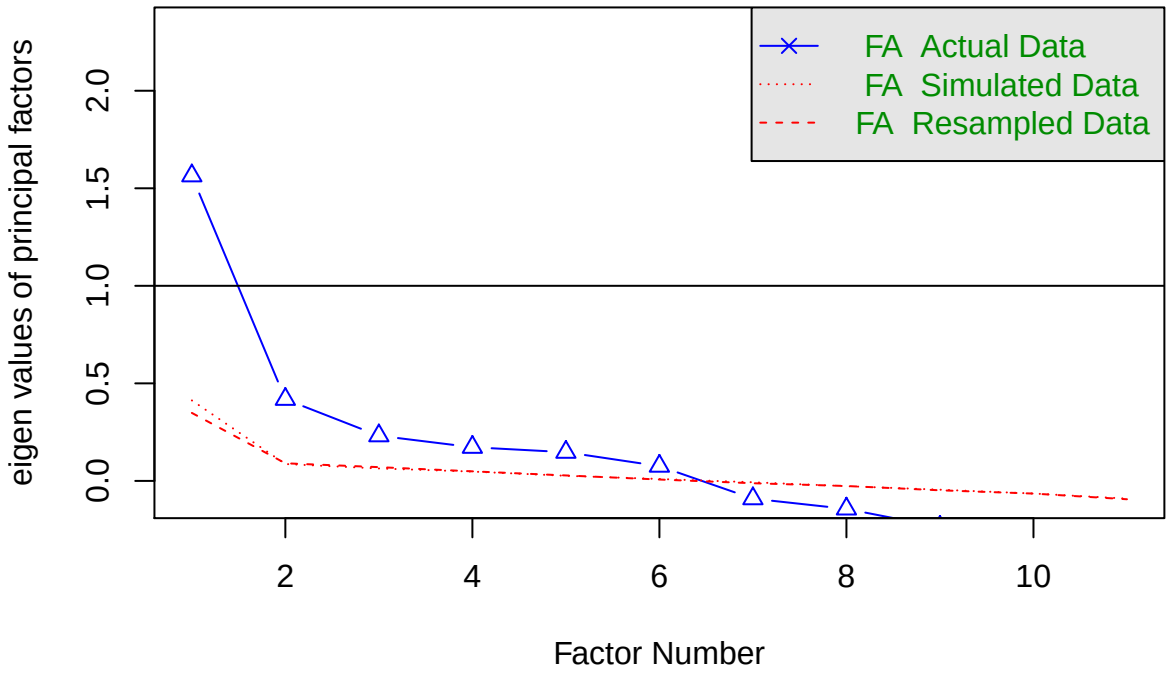
From output data, Bartlett's test is highly significant, assymptotically chisquare is 2273, and P-value of chi square smaller than 0.001, therefore factor analysis is appropriate.
And since the determinant is 0.36 larger than 0.00001, so our determinant does not seem problematic.

Since my goal is reduce number of variables(selected survey questions) in my data by extracting important one from the data, thus, I will using PCA to do factor extraction.

Factor extraction by using PCA:

Show scree Plot by using parallel analysis:

```
pc <- principal(newdata[x], nfactors = 11, rotate="none")
#using a bigger device window explicitly showing the graph:
#dev.new(width=10, height=10)
parallel <- fa.parallel(newdata[x], fm='minres', fa='fa', main = "Scree Plot")
```

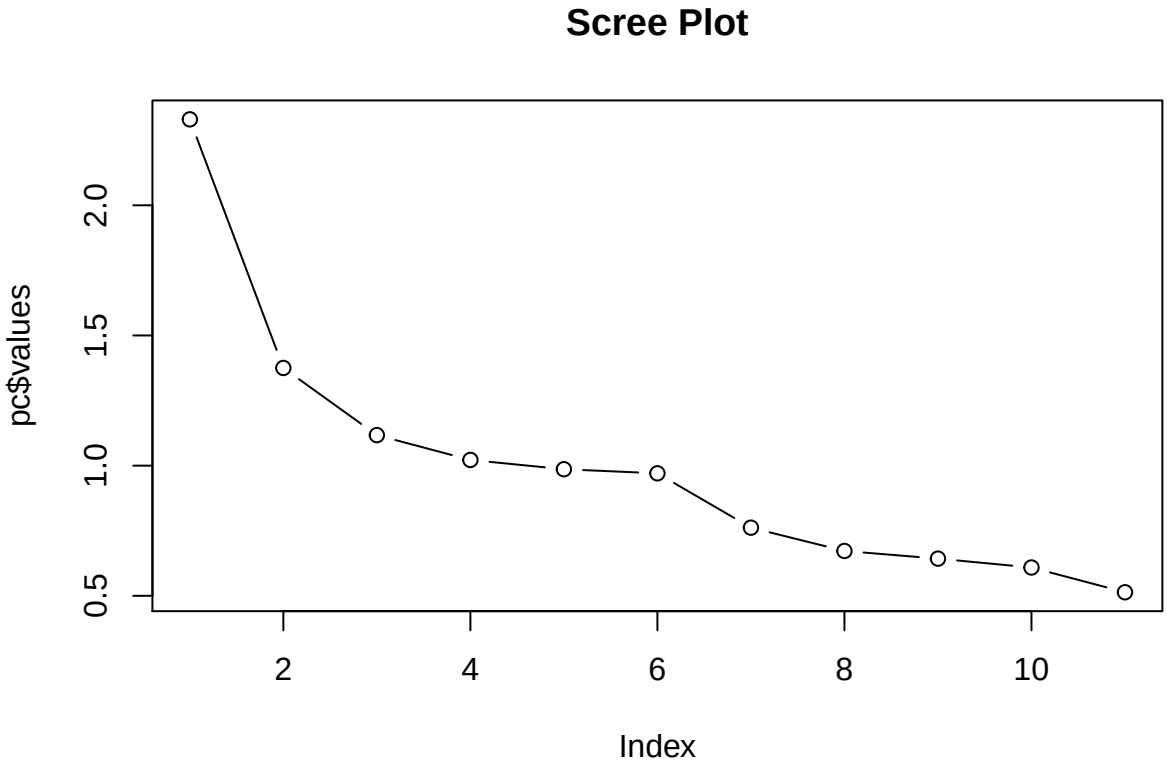



```
## Parallel analysis suggests that the number of factors =  6  and the number of components =  NA
```

From the parallel analysis, the plot shows from 2 to 6 will be great choice.

scree plot by using another way:

```
plot(pc$values, type="b", main = "Scree Plot")
```




```
#x axis: component number
#y axis:eigenvalues of principal components and factor analysis
```

Since the elbow part is about the 3rd point from the left, so the evidence from the scree plot and from the eigenvalues suggests 3 component solution may be the best. Thus, combining 2 graphs, we choose 3 as our number of factors.

Redo PCA by using 3 factors

```
pc2 <- principal(newdata[x], nfactors = 3, rotate="none")
pc2
```

```
## Principal Components Analysis
## Call: principal(r = newdata[x], nfactors = 3, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                                        PC1   PC2   PC3   h2   u2 com
## V4: Important in life: Family         0.68 -0.06  0.24 0.53 0.47 1.3
## V5: Important in life: Friends        0.64  0.00  0.08 0.42 0.58 1.0
## V6: Important in life: Leisure time   0.54  0.03 -0.08 0.30 0.70 1.1
## V7: Important in life: Politics       0.51  0.08  0.19 0.30 0.70 1.3
## V8: Important in life: Work           0.41  0.17  0.13 0.21 0.79 1.5
## V9: Important in life: Religion       0.49  0.02  0.37 0.37 0.63 1.9
## V10: Feeling of happiness             0.58  0.29 -0.38 0.57 0.43 2.3
## V11: State of health (subjective)     0.21  0.36 -0.76 0.74 0.26 1.6
## V60: Aims of country: first choice   -0.26  0.67  0.22 0.57 0.43 1.5
## V62: Aims of respondent: first choice -0.08  0.47 -0.01 0.23 0.77 1.1
## V64: Most important: first choice    -0.20  0.67  0.31 0.58 0.42 1.6
##
##                        PC1  PC2  PC3
## SS loadings           2.33 1.38 1.12
## Proportion Var        0.21 0.13 0.10
```

```
## Cumulative Var        0.21 0.34 0.44
## Proportion Explained  0.48 0.29 0.23
## Cumulative Proportion 0.48 0.77 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.11
##  with the empirical chi square  2848.59  with prob <  0
##
## Fit based upon off diagonal values = 0.51
```

Through output data, Cumulative var shows these 3 principle components explains 44% data with 11 variables.

```
#Difference between the reproduced cor matrix and the original cor matrix
residuals <- factor.residuals(matrix, pc2$loadings)
residuals <- as.matrix(residuals[upper.tri(residuals)])
large.resid <- abs(residuals) > 0.05
#proportion of the large residuals
sum(large.resid) / nrow(residuals) #0.58
```

```
## [1] 0.5818182
```

For easier to explain the output of factor extraction, we can using orthogonal rotation to decreasing noice for factors as much as possible.

Orthogonal rotation(varimaax):

```
pc3 <- principal(newdata[x], nfactors = 3, rotate = "varimax")
print.psych(pc3, cut=0.3, sort = TRUE, main="table after orthogonal rotation")
```

```
## Principal Components Analysis
## Call: principal(r = newdata[x], nfactors = 3, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                                      item   RC1   RC2   RC3   h2   u2 com
## V4: Important in life: Family           1  0.71             0.53 0.47 1.1
## V5: Important in life: Friends          2  0.62             0.42 0.58 1.2
## V9: Important in life: Religion         6  0.59             0.37 0.63 1.1
## V7: Important in life: Politics         4  0.55             0.30 0.70 1.0
## V6: Important in life: Leisure time     3  0.46             0.30 0.70 1.8
## V8: Important in life: Work             5  0.44             0.21 0.79 1.2
## V64: Most important: first choice      11        0.76       0.58 0.42 1.0
## V60: Aims of country: first choice      9        0.75       0.57 0.43 1.0
## V62: Aims of respondent: first choice  10        0.45       0.23 0.77 1.2
## V11: State of health (subjective)       8              0.86 0.74 0.26 1.0
## V10: Feeling of happiness               7  0.40        0.64 0.57 0.43 1.7
##
##                        RC1  RC2  RC3
## SS loadings           2.12 1.40 1.30
## Proportion Var        0.19 0.13 0.12
## Cumulative Var        0.19 0.32 0.44
## Proportion Explained  0.44 0.29 0.27
## Cumulative Proportion 0.44 0.73 1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 3 components are sufficient.
##
```

```
## The root mean square of the residuals (RMSR) is  0.11
##  with the empirical chi square  2848.59  with prob <  0
##  table after orthogonal rotation
## Fit based upon off diagonal values = 0.51
```

```
#cut=0.3:only loading above 0.3, otherwise correlation is not high enough, so we consider excluding the
```

According to the results and the questionnaires, We can find the questions that load highly on factor 1 are V4("Important in life:Family") with the highest loading of 0.71, and lowest loading of 0.4 is V10(Feeling of happiness). Factor 2 are mianly explained by V64(Most important:first choice) and V60(Aim of country:first choice) with loading of 0.76 and 0.75. Factor 3 are mainly explained by V11(state of health) with 0.86 and the loweest loading of 0.64 is V10(Feeling of happiness).

Based on the obsersation, we can summarize the factor 1 as hapiness people consider a lot are important in life and label factor 2 as the expectation from people to country, factor 3 as causes of healthy people ?

Since all facto1, factor2 and factor3's data are below 0.8 and over 0.3, which is encouraging.

```
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

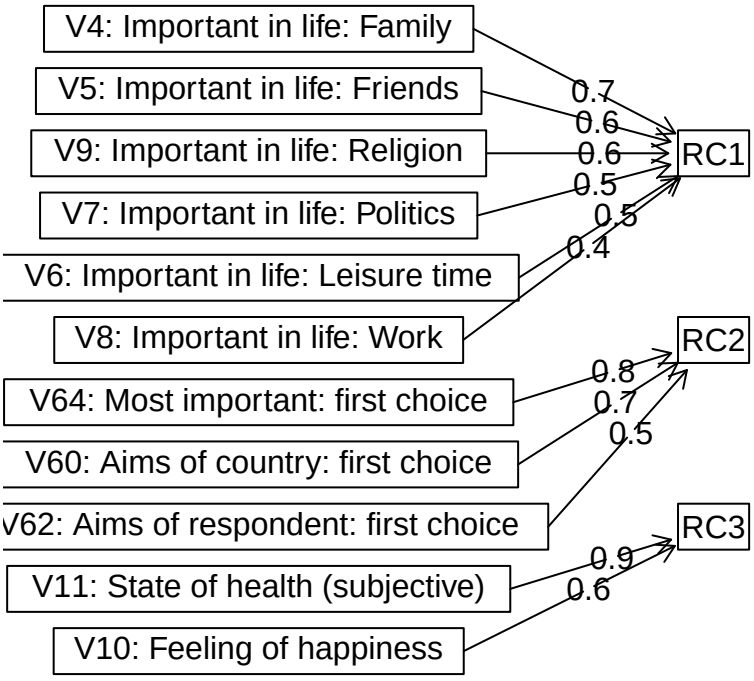
```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
library(RColorBrewer)
fa.diagram(pc3,simple=TRUE)
```

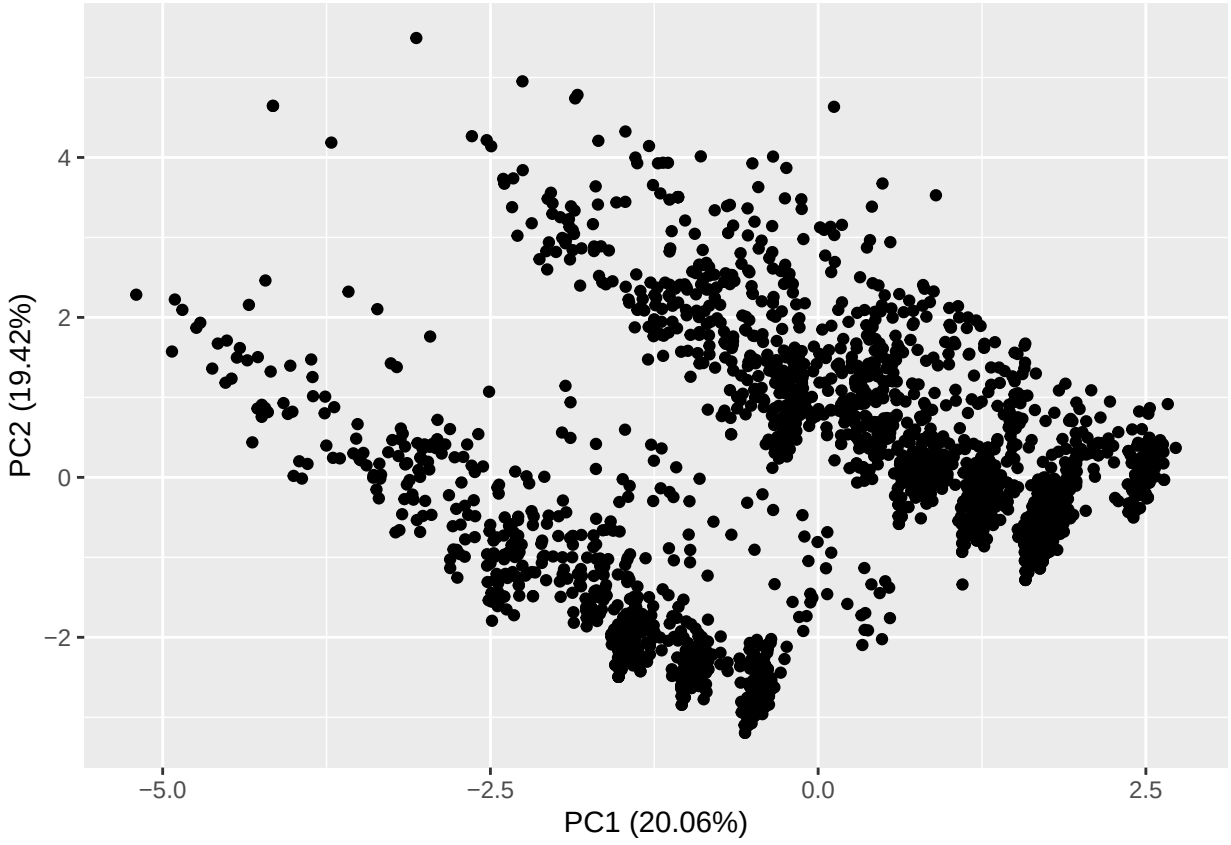
Components Analysis



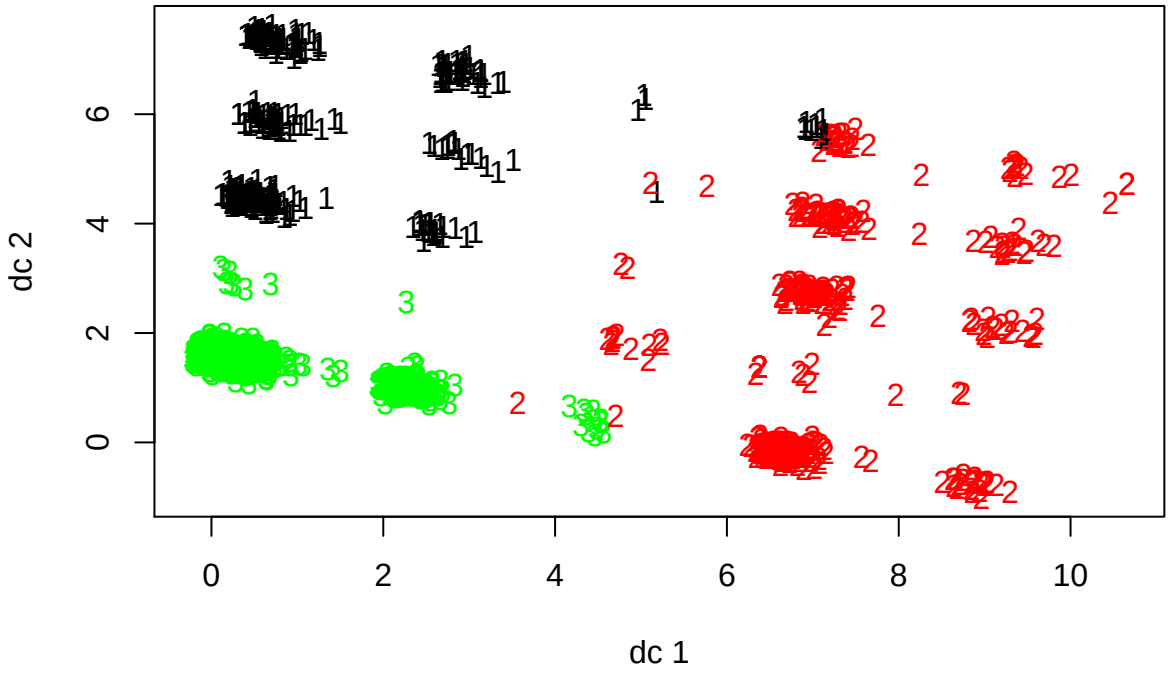

```
#dev.new(width=10, height=10)
#biplot(princomp(newdata[x]))# only plot pincipal components 1 vs 2
```

Ploting PCA

```
library(ggfortify)
autoplot(prcomp(newdata[x]), scale=0)
```




```
library(cluster)
library(fpc)
clus <- kmeans(newdata[x], centers = 3)
#dev.new(width=10, height=10)
plotcluster(newdata[x], clus$cluster)
```



```
clusplot(newdata[x], clus$cluster, color = TRUE, shade = TRUE, labels = 3, lines = 0)
```