6 December 2019

Multivariate Analysis-World Happiness (2015)

Prepared by: Bruce Archer

Dinesh Poudel

Maurice Thomas

Colin Yang

Prepared for: Dr. Alireza Sheikh-Zadeh

Multivariate Analysis

ISQS 6350 - 001

Contents

List of Figures and Tables

Introduction

The happiness report ranks world happiness by country based ordinal survey data measuring generosity, perceived freedom, trust in government, life expectancy, GDP per capita, family matters, and a happiness score of its respondents.

The World Happiness Survey has information about each country's perceived happiness among respondents. The research will identify potential factors that influence happiness. This information is helpful because it could be merged with government records on the CIA World Factbook to determine the preferred form of government: autocracy, democracy, or oligarchy. Moreover, we can identify socio-legal factors such as preferences for limits on government: constitutional (strict codified limits on government power), totalitarian (no limits on government power), or parliamentary limits of power (limits only based to the extent the current parliament allows).

This information is helpful for multinational organizations looking to identify potential investment countries with a stable economy, stable government, and healthy supply of labor. Happiness report is obtained from Kaggle.com.

Data Cleaning and Visualization

Table 1: World Happiness Data

Country	GDP	Family	Health (Life Expectancy)	Freedom	Govt.Trust	Generosity
Switzerland	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678
Iceland	1.30232	1.40223	0.94784	0.62877	0.14145	0.4363
Denmark	1.32548	1.36058	0.87464	0.64938	0.48357	0.34139
Norway	1.459	1.33095	0.88521	0.66973	0.36503	0.34699
Canada	1.32629	1.32261	0.90563	0.63297	0.32957	0.45811
Finland	1.29025	1.31826	0.88911	0.64169	0.41372	0.23351

For the purpose of our multivariate analysis, analyzed six variables that impact the overall happiness score.

Missing Values

For the purposes of this study zero values have been classified as "missing" values and have been replaced with the mean of the column. With this, our data does not have any missing values, we can now move to 'outlier' detection and removal process.

Outlier Detection

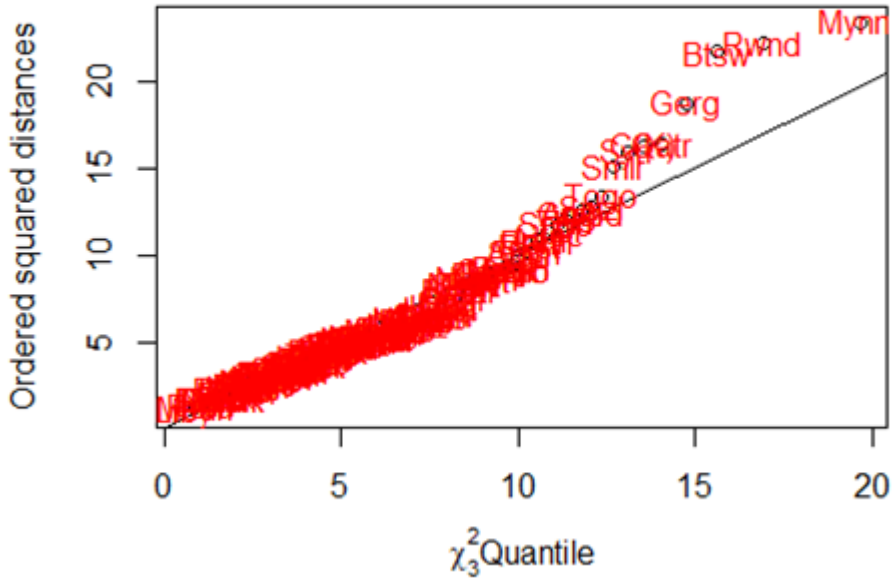

Figure 1: Multivariate Normality Plot with Outlier

We see a clear deviation from multivariate normality. However, those that are far from the line be treated as outliers. Myanmar, Botswana, Rwanda, Syria, Qatar and Somaliland are outliers and we removed these observations from the data.

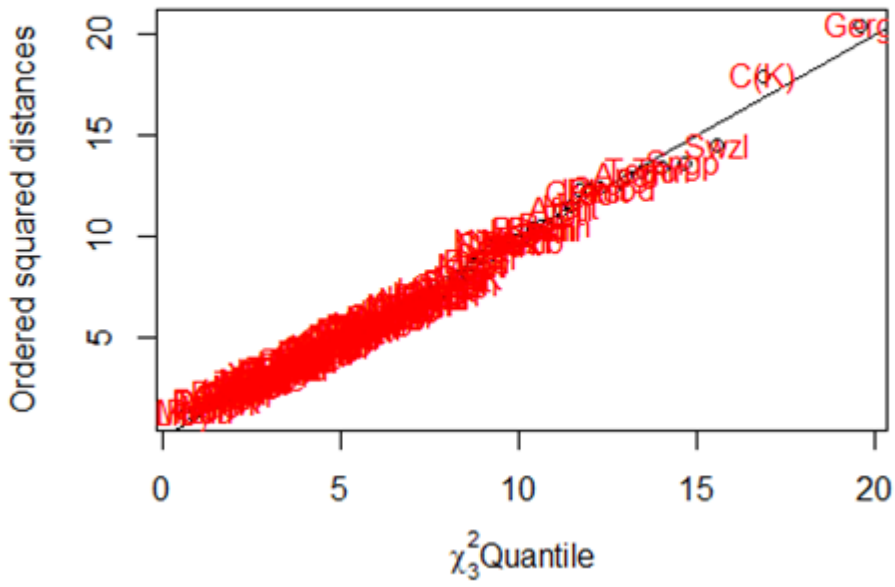

Figure 2: Multivariate Normality Plot without Outlier

After removing six outlier countries from our data and performing multivariate normality test, it is safe for us to assume our data arises from a multivariate normal distribution.

Scatterplot and Correlation Matrix

Based on scatterplot and correlation matrix, it is obvious that most of the variables are positively correlated.

The correlation between GDP and generosity as well as family and generosity is low. Likewise, the correlation between GDP and life expectancy as well as GDP and family are highly correlated to each other. The goal of dimension reduction analysis is to reduce dimension and represent these highly correlated variables with fewer new transformed variables.

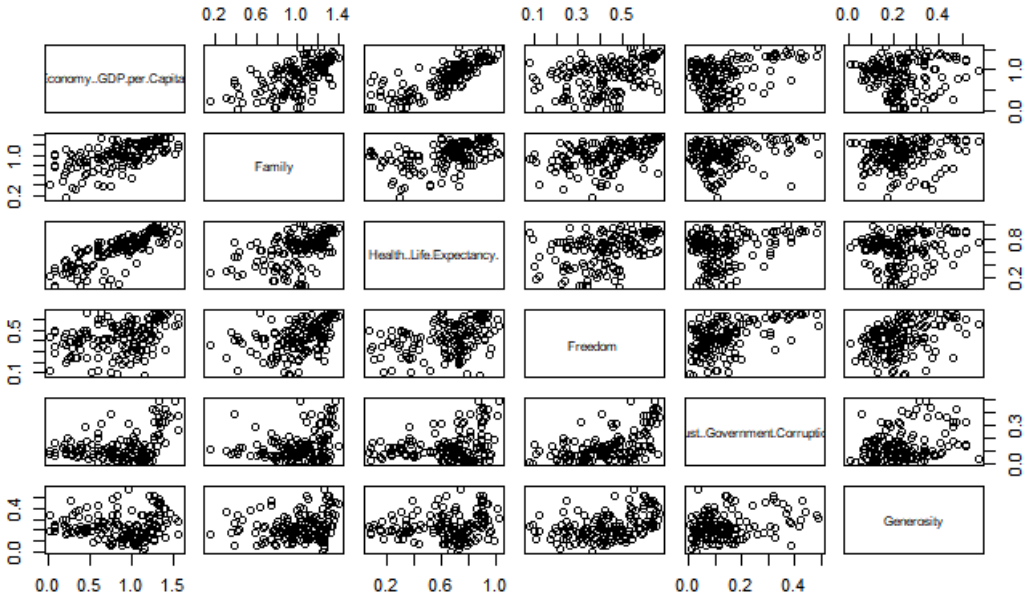

Figure 3: Scatterplot Matrix

Table 2: Correlation Matrix

	Economy..GDP.per.Capita.	Family	Health..Life.Expectancy.	Freedom	Trust..Government.Corruption.	Generosity
Economy..GDP.per.Capita.	1	0.64	0.79	0.37	0.31	
Family	0.64	1	0.51	0.47	0.21	0.09
Health..Life.Expectancy.	0.79	0.51	1	0.37	0.25	0.12
Freedom	0.37	0.47	0.37	1	0.51	0.35
Trust..Government.Corruption.	0.31	0.21	0.25	0.51	1	0.28
Generosity		0.09	0.12	0.35	0.28	1

Dimension Reduction Analysis

For our project, we performed the following dimension reduction techniques: principal component analysis, canonical correlation analysis, and multidimensional scaling. Principal Component Analysis allowed us to reduce the variables from the happiness survey into three principal components. Canonical correlation analysis produced two groups, X and Y. X represents GDP per capita, family, and life expectancy. Y represents freedom, trust in government (i.e. corruption), and generosity. Multidimensional scaling allowed us to identify countries that are similar in terms of the happiness variables.

Principle Component Analysis

"World Happiness" data is multivariate data and have some correlation in-between variables. Our goal is to find new variables or principal components that are uncorrelated to each other. We have six different variables to measure the happiness score in 158 countries. Here, we want to apply PCA to determine if the PC1 can present actual happiness of countries. As we already cleaned our data and all variables in our clean dataset are in the same direction, we can perform PCA. Our PCA output is shown on the table below.

Table 3: Principal Component Analysis

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6
Standard deviation     1.7374979 1.1019700 0.8393122 0.7295447 0.62126657 0.37961892
Proportion of Variance 0.5031498 0.2023897 0.1174075 0.0887059 0.06432869 0.02401842
Cumulative Proportion  0.5031498 0.7055395 0.8229470 0.9116529 0.97598158 1.00000000

Loadings:
                           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Economy..GDP.per.Capita.    0.484  0.382         0.246         0.743
Family                      0.440  0.206 -0.287 -0.619  0.495 -0.230
Health..Life.Expectancy.    0.466  0.333 -0.105  0.461 -0.312 -0.592
Freedom                     0.425 -0.370        -0.444 -0.682  0.130
Trust..Government.Corruption.  0.353 -0.359  0.755  0.132  0.369 -0.150
Generosity                  0.223 -0.659 -0.576  0.359  0.224
```

Looking at PCA output, the first two components account for about 70.5% of the overall variance in the data. Generally, the cut-off value for choosing PCs is 70%. Loadings show the correlation between each original variable and the new principal components. In this case, the coefficients for PC1 are positive for all six original variables. So, PC1 presents overall score (performance) in all six criteria of happiness. PC2 presents countries with score high on GDP, Family, and Health/Life Expectancy but low on Freedom, Trust in Government/Corruption, and Generosity.

Switzerland has the highest PC1 score (3.562), which is consistent because it has the happiness score in reality and the highest happiness rank in the original data. The correlation between the first principal component score and the actual happiness score is 0.84. This implies that the first principal component agrees with the scores assigned to the countries. Therefore, instead of six original variables, the reality of data can be represented with 2 new variables (Principal Components).

PCA-Biplot

Biplot is an overall scatterplot which aims to represent both the observations and variables of a matrix of multivariate data on the same plot which helps to interpret the principal component axes while looking at individuals' location.

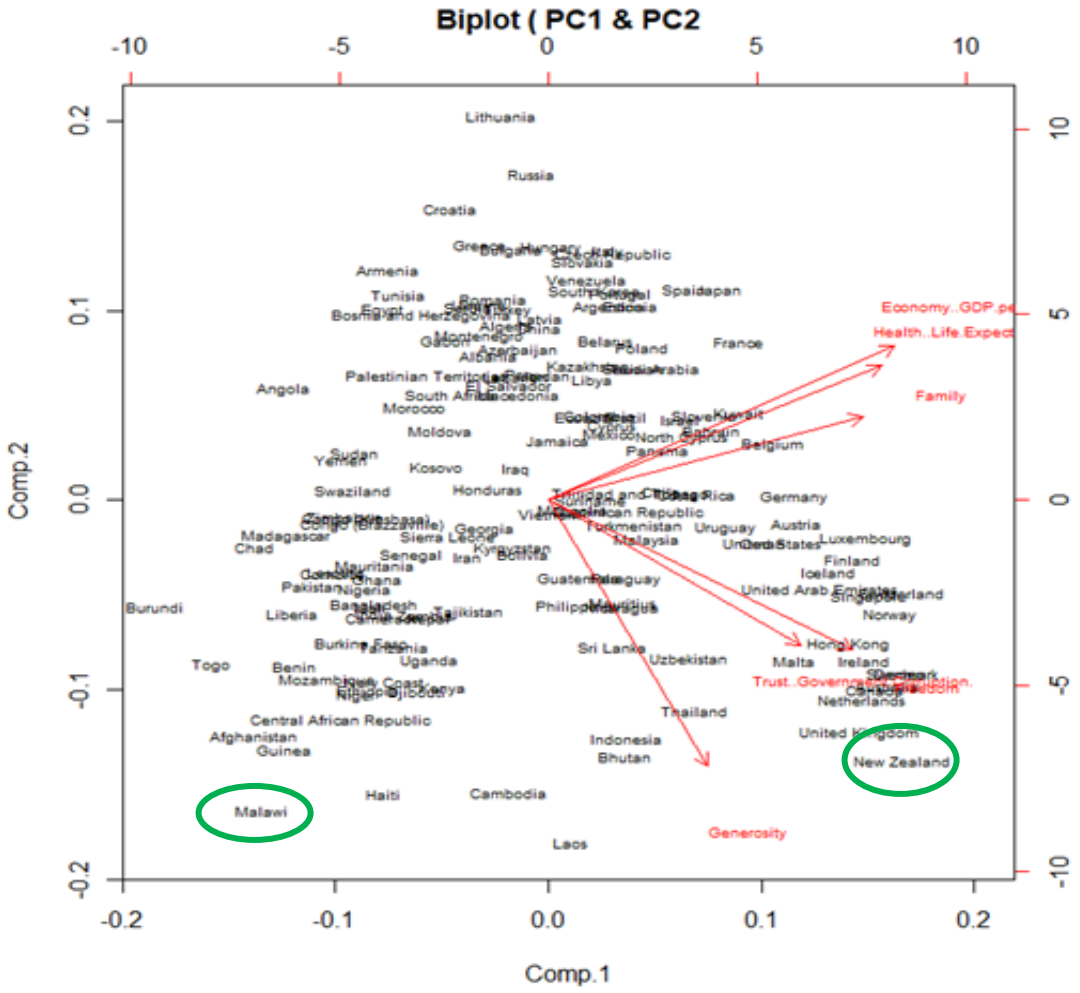

Figure 4: PCA Bi-plot

Projecting a data point onto the direction represented by an arrow gives the measurements of that variables for that data value. For example: New Zealand has high scores in all six criteria while Malawi has higher score on generosity, trust-in-government, and freedom but lower score in GDP per capita, health/life expectancy, and family.

PC2 loadings indicates the correlation between the original variables and the new PC2 score to explain the angles. GDP per capita, health/life expectancy, and family has a positive angle while freedom, generosity, and trust-in-government has a negative angle as indicated on PC2 loadings.

The biplot indicates GDP per capita, family, and health/life expectancy are highly correlated to each other. Likewise, generosity, trust-in-government/corruption, and freedom are highly correlated to each other. Exploratory factor analysis (EFA) may demonstrate that these pairs of variables may identify two latent factors. Length of the arrow shows the impact of variables on PC1 and PC2. From the plot, it looks like GDP per capita has the most impact on PC1 and PC2 scores.

Multiple Dimension Scaling

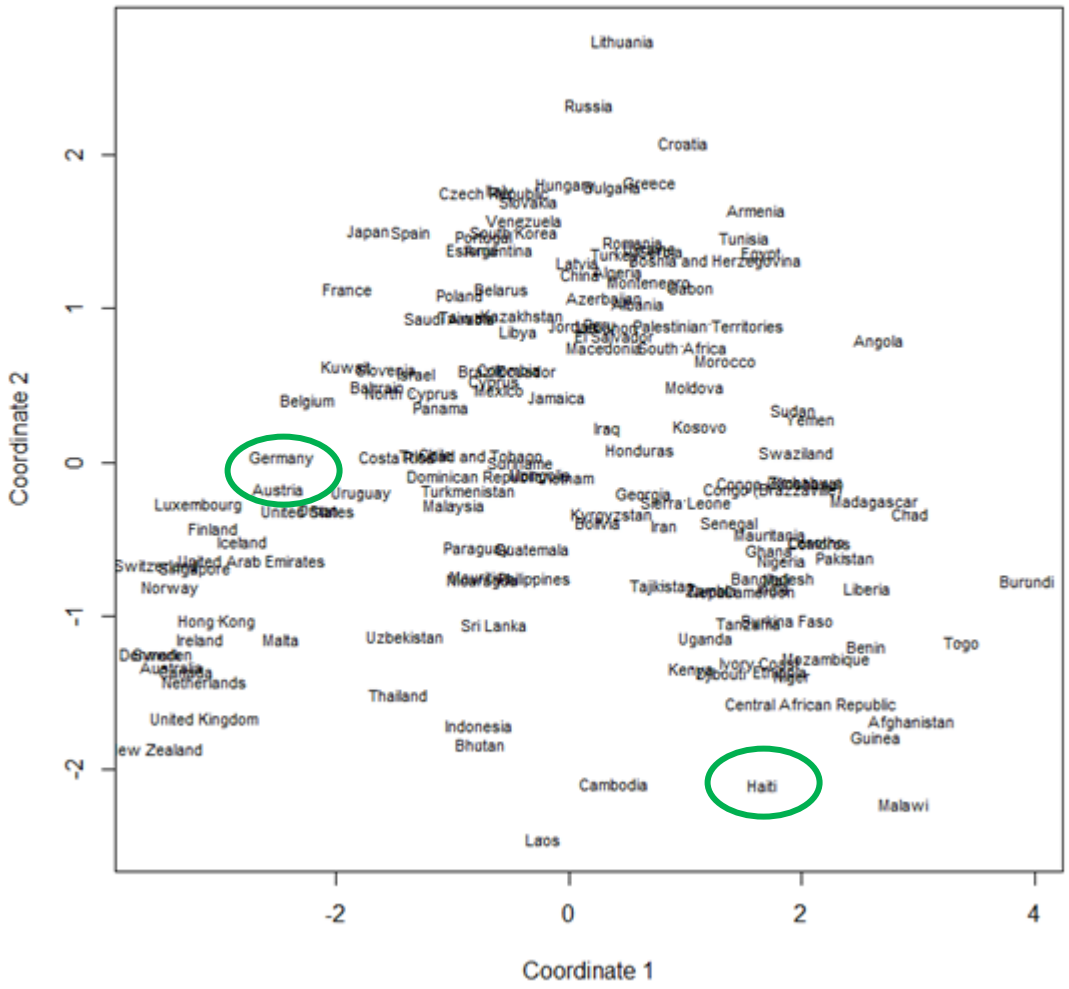

Figure 5: MDS plot for Observations

Looking at the plot for component 1 and component 2 for observations, countries with similar happiness score are in close proximity. For example, Germany & Austria have similar characteristics and closer on map and they are very far from Haiti (country with low happiness score).

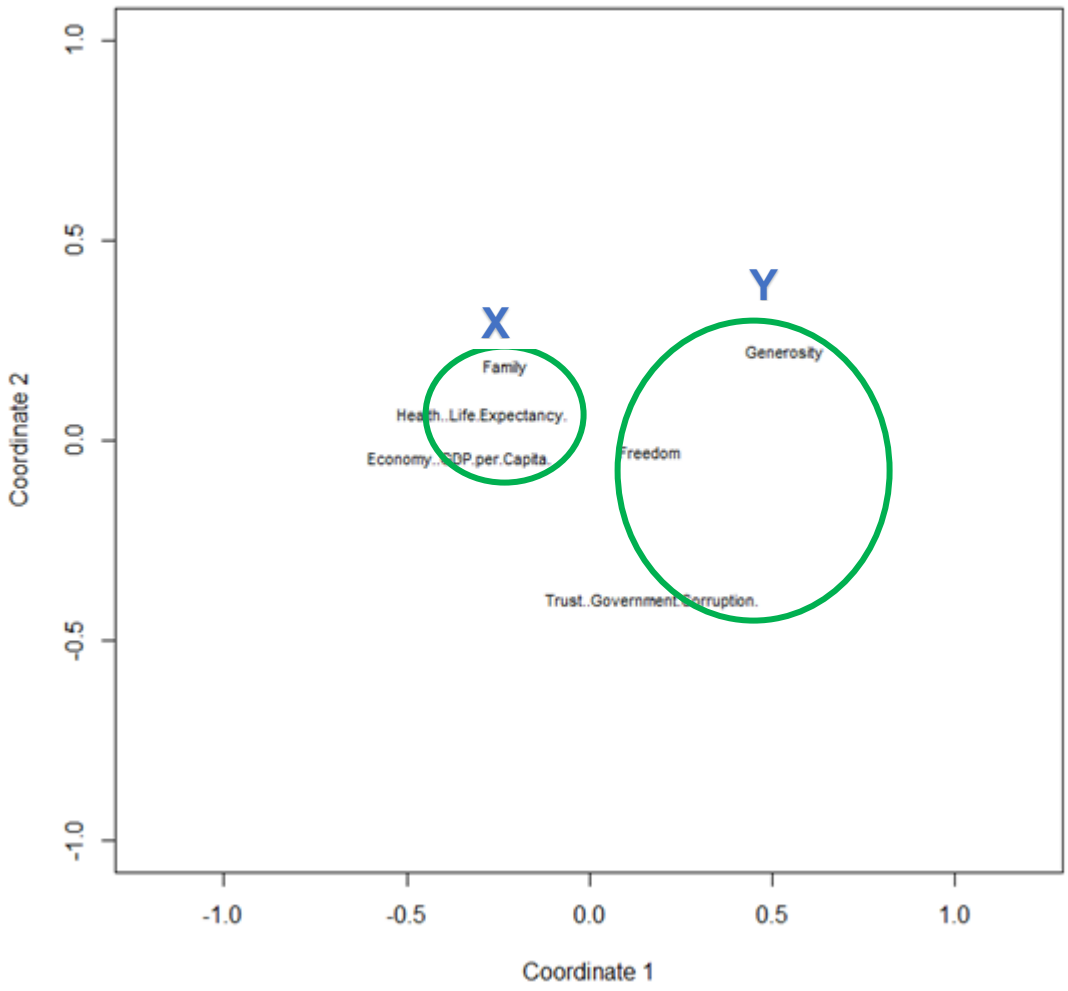

Figure 6: MDS plot for Variables

MDS can also be used for representing variables when distance matrix comes from correlation matrix. From plot above, GDP, family and Life Expectancy are closer. And, Trust on Government, Freedom and Generosity are closer. Hence, it is expected that these two pairs of variables to be on different factors on EFA.

Canonical Correlation Analysis

PCA considers interrelationships within a set of variables. On the other hand, CCA assesses the relationships between two sets of variables. CCA works by maximizing the correlation between the pairs and first pair always have the highest correlation.

Let's assume:
X represents Family, GDP and Life expectancy
Y represents Freedom, Government Trust and Generosity

Maximum possible pairs are 3, and correlation between first pair (U1, V1) is 0.509. With the X and Y coefficients found through CCA, we can create a new surrogate variable. The linear construct for the first pair is as shown:

$$U_1 = -0.004x_1 + 1x_2 + 0.670x_3$$

$$V_1 = 1y_1 + 0.148y_2 - 0.080y_3$$

As an interpretation, GDP has more dominance among the first set of variables(U1); whereas, freedom has more dominance among the second set of variables(V2). The analysis makes sense because a high GDP per capita may indicate a strong relationship to life expectancy, and strong family connections might produce overall happiness. Moreover, a country with more individual freedom might produce a high trust-in-government.

Cluster Analysis

Cluster analysis allowed us to group countries with similar happiness characteristics. We discovered the following 3 clusters: North America/Western Europe/Australia, Latin America/North Africa/Northeast Asia, and South Asia/Africa. Based on our analysis, model-based clustering was the best clustering technique because our data was multivariate normal, and we were able to attach meaning to clusters using PC1 & PC2 plot.

Hierarchical-Single

Hierarchal clustering with single linkage provided no clearly defined clusters through the dendrogram, scree plot, and the principle component plots. The scree plot indicates two clusters and from principle component plot, it appears that there is only one observation in second group.

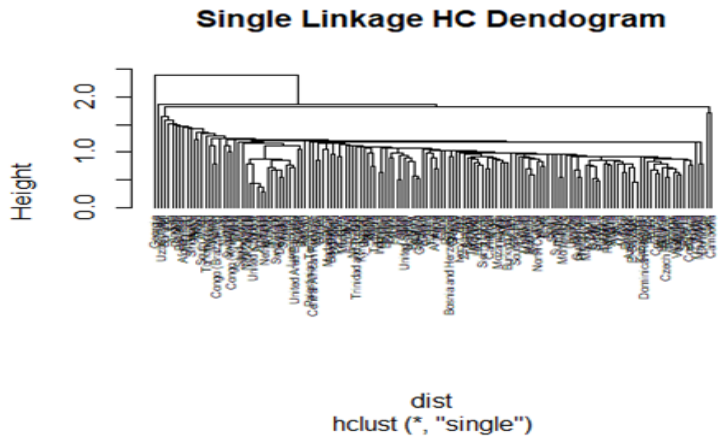

Figure 7: Dendrogram- Hierarchical (Single)

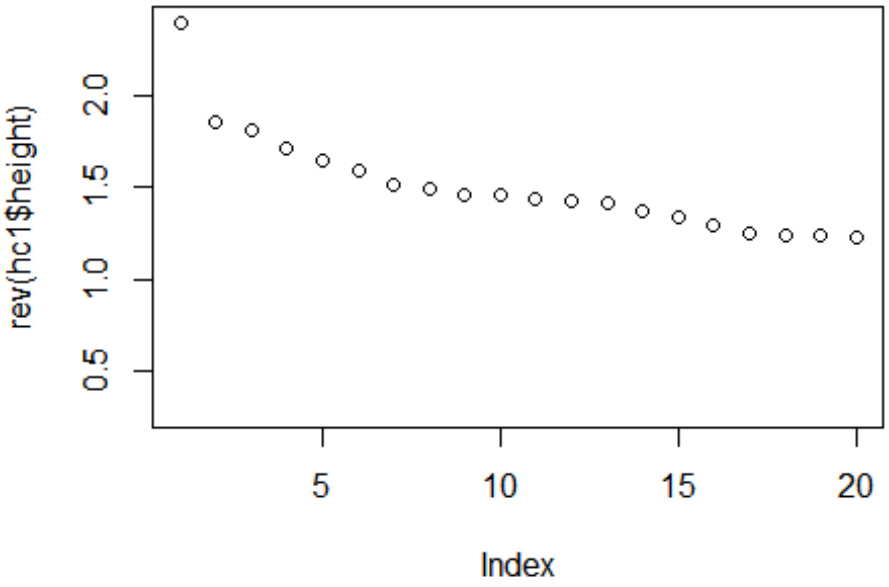

Figure 8: Scree plot- Hierarchical (Single)

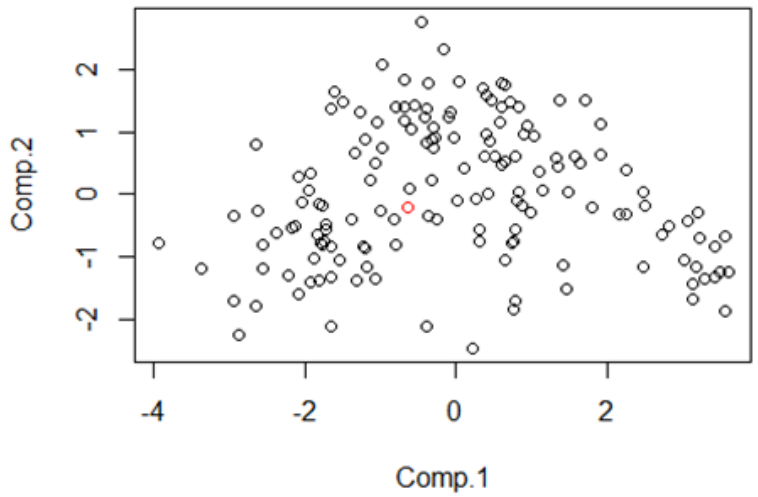

Figure 9: PC1 vs PC2 (Hierarchical Single)

Hierarchical-Complete

Hierarchical-Complete provided clearer groups than single linkage hierarchal clustering; however, the groups are still very vague. The scree plot depicts two clusters and from principle component plot, it appears that there is overlapping, and no clear meaning could be attached to PCs.

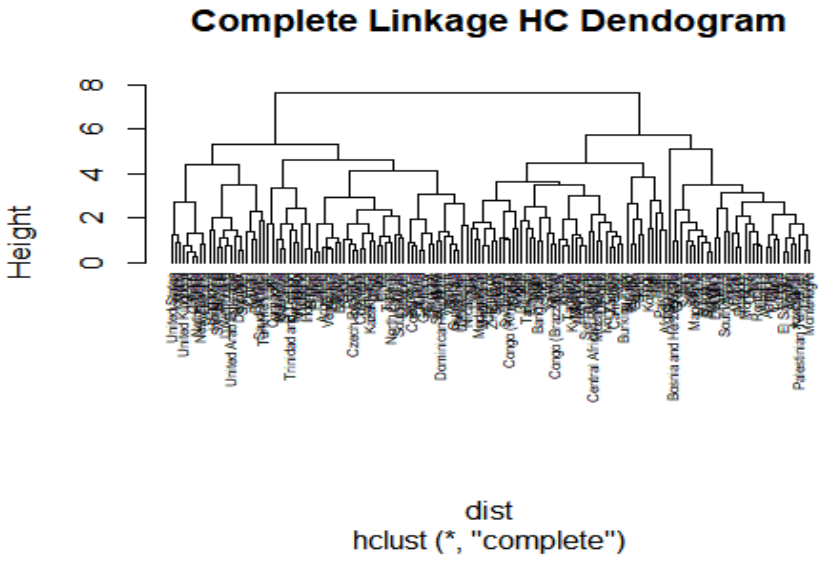

Figure 10: Dendrogram- Hierarchical (Complete)

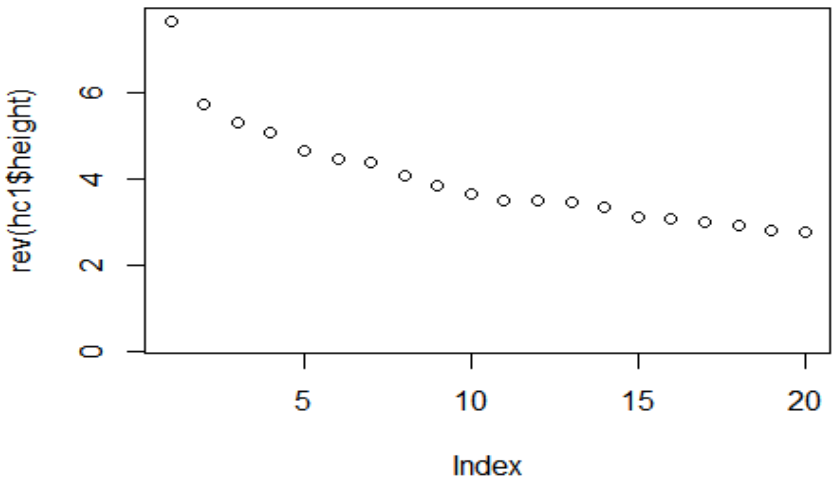

Figure 11: Scree plot- Hierarchical (Complete)

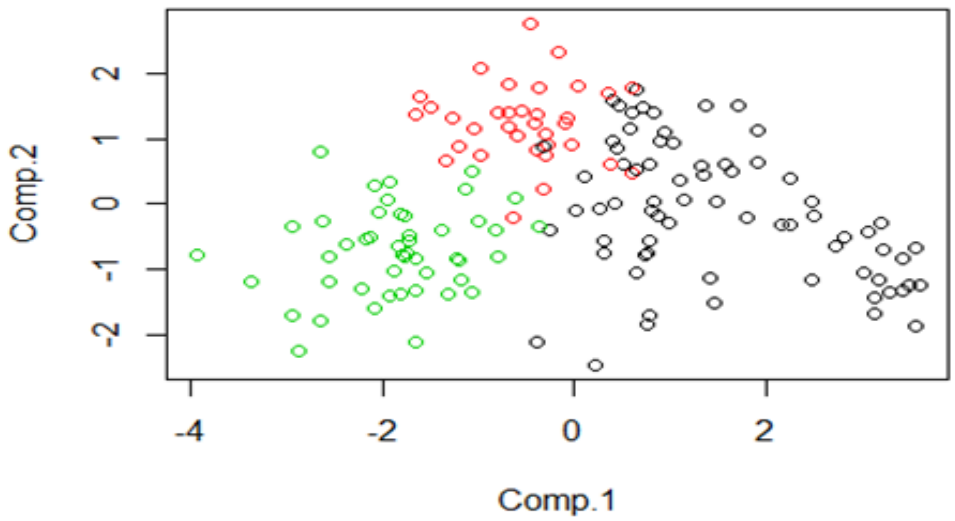

Figure 12: PC1 vs PC2 (Hierarchical Complete)

Hierarchical-Average

Hierarchal clustering with average linkage provided worse results than single linkage and complete linkage hierarchal clustering. There are no clear groups in any of the clustering techniques. The scree plot shows three clusters and from principle component plot and it appears that there is only one observation in one of the group. There is not a clear meaning could be attached to PCs.

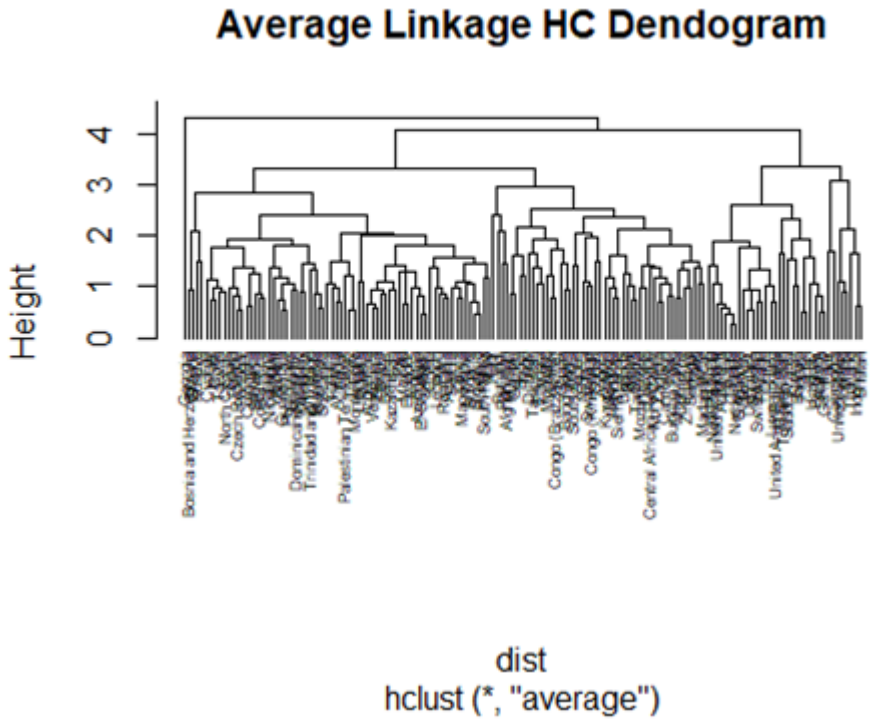

Figure 13: Dendrogram- Hierarchical (Average)

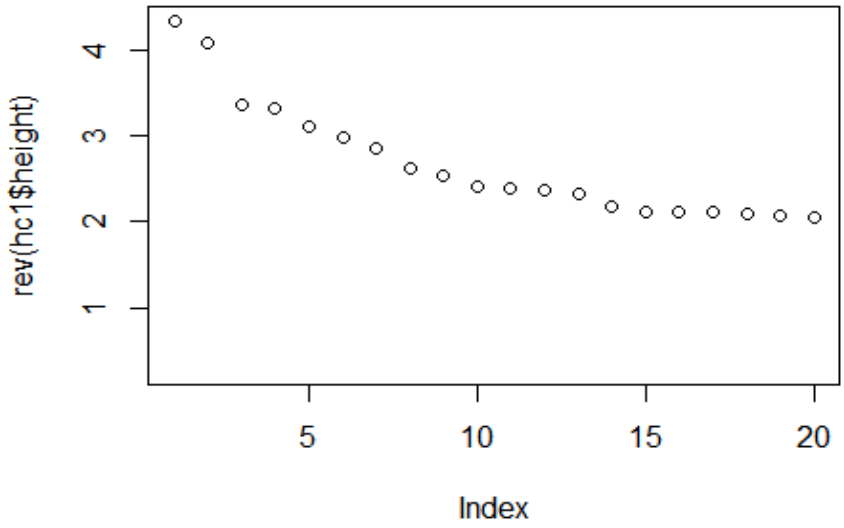

Figure 14: Scree plot- Hierarchical (Average)

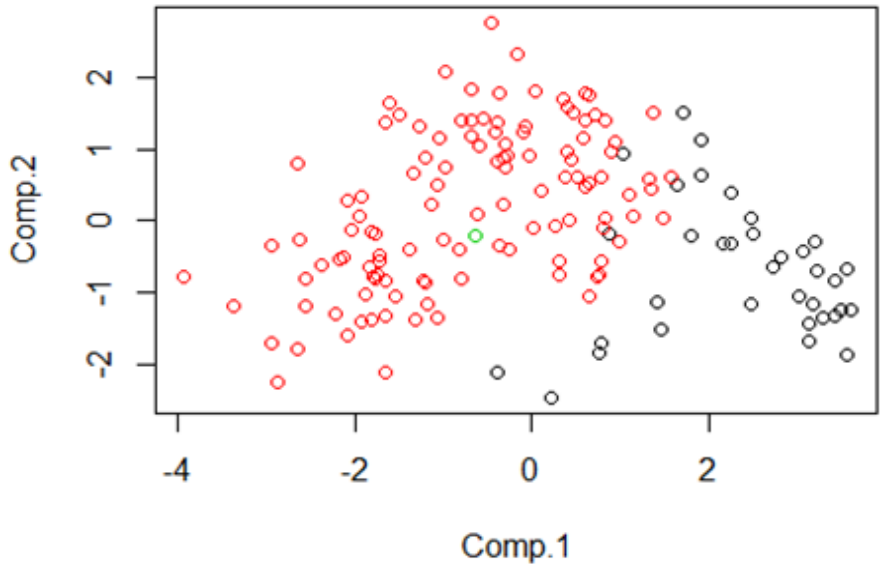

Figure 15: PC1 vs PC2 (Hierarchical Average)

K-Means

K-Means clustering provided more satisfactory than hierarchal clustering results, however, the clustering groups are not as clear as model-based clustering. The scree plot shows three clusters and from principle component plot and it appears that there is overlapping. Consequently, there is no clear meaning could be attached to PCs.

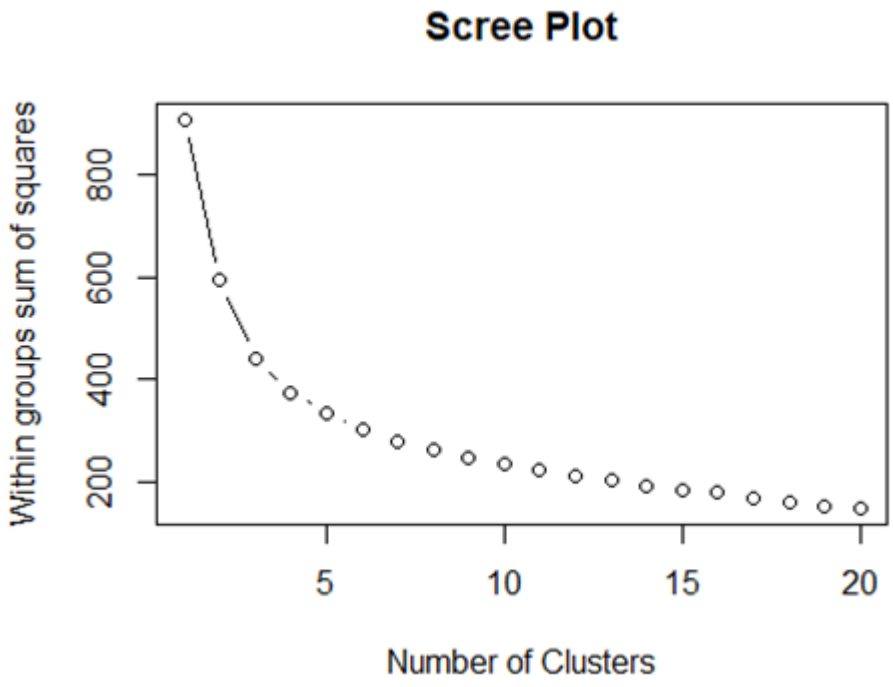

Figure 16: K-Means Scree plot

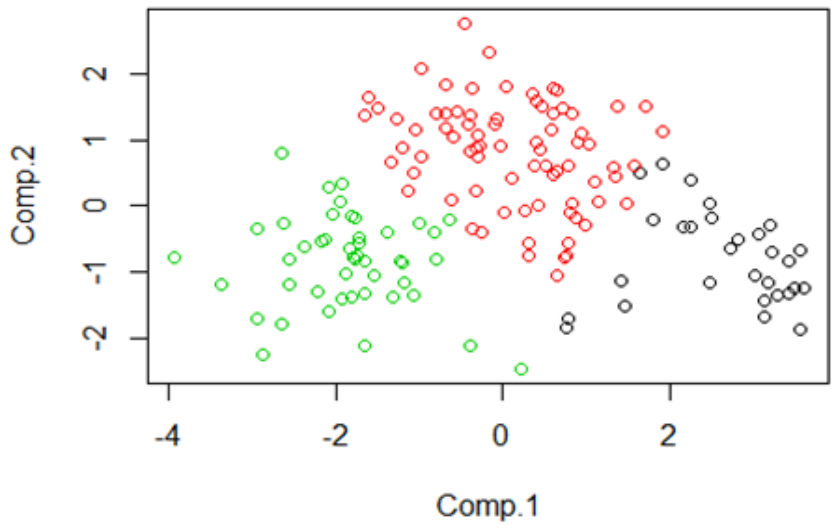

Figure 17: PC1 vs PC2 (K-Means)

Model-Based (Unsupervised)

The multivariate normal data provided the best clustering groups using model-based clustering techniques. Jordan has the most uncertainty probability. We have three groups identified by model-based clustering.

The principle component plot shows two clusters can be described with PC1; whereas, the third cluster can be described with PC2. PC1 show the overall happiness performance. PC2 represents countries which performed better in personal happiness related variables; whereas, they performed poor on societal related variables.

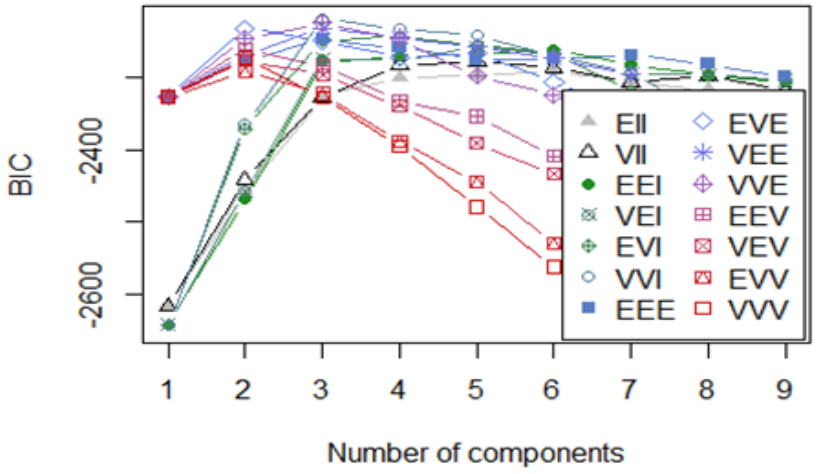

Figure 18: "BIC" plot

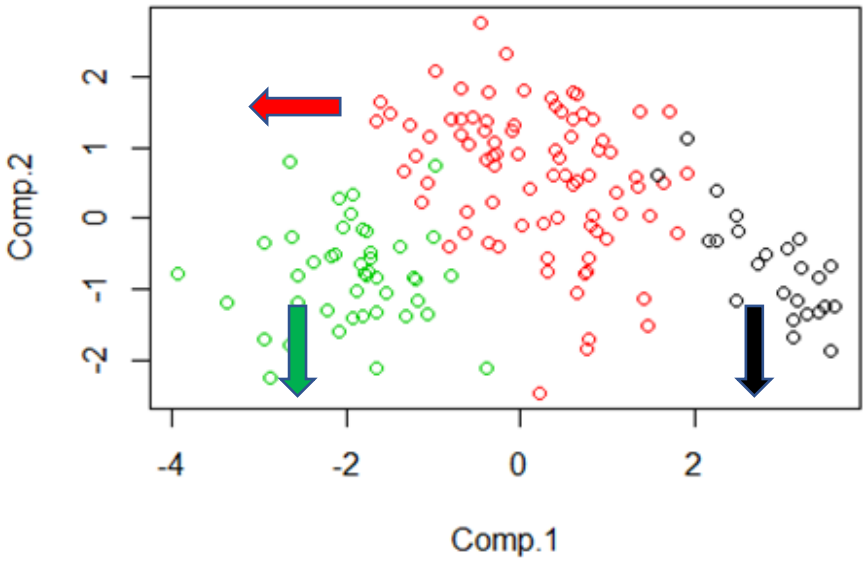

Figure 19: "PC1 vs PC2 (Model-Based)

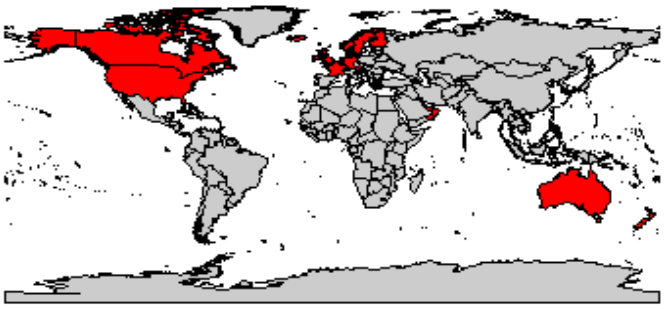

Figure 20: Cluster “1” on World Map

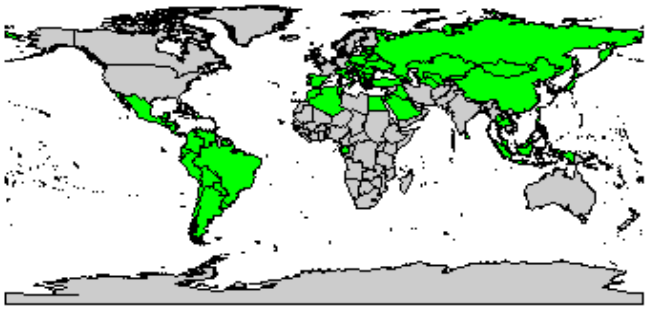

Figure 21: Cluster “2” on World Map

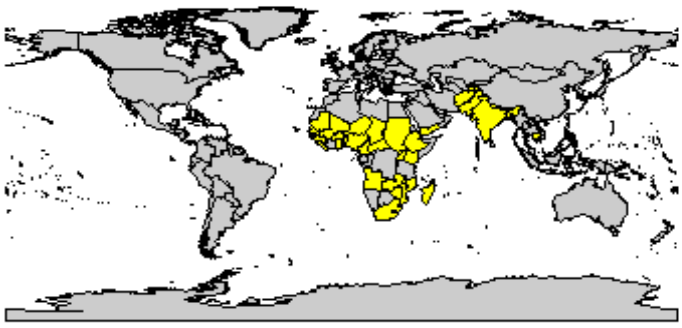

Figure 22: Cluster “3” on World Map

Model-Based Discriminant (Supervised)

Table 4: MclustDA Model Summary

```
MclustDA model summary:

 log-likelihood   n df      BIC
       444.6319 120 72 544.5643

Classes   n     % Model G
  Higher 34 28.33   VVI 2
  Low    14 11.67   XXI 1
  Medium 72 60.00   VEE 2

Training confusion matrix:
         Predicted
Class     Higher Low Medium
  Higher     23   0     11
  Low         0  11      3
  Medium      7   8     57
Classification error = 0.2417
Brier score          = 0.1631

Test confusion matrix:
         Predicted
Class     Higher Low Medium
  Higher      6   0      3
  Low         0   4      1
  Medium      1   6     11
```

Since the true cluster was not available in the given dataset, we created group labels based on the happiness score.

Happiness score	Happiness Rank
Less than 4	Low
4 – 6	Medium
6 or more	High

With random training sample of 120 observations and testing sample of 32 observations, we can classify group of any new observation with 70% to 90% accuracy.

Exploratory and Confirmatory Factor Analysis

EFA identified two potential latent variables. We believe those unobserved variables are personal/household happiness and societal happiness. Based on the EFA findings, Confirmatory Factor Analysis (CFA) confirmed that GDP per capita, family, and health/life expectancy reflect personal/household happiness. Likewise, freedom, trust-in-government, and generosity reflect societal happiness. In short, the two variables are factors inside the household and factors outside the household.

Exploratory Factor Analysis (EFA)

We performed exploratory factor analysis to identify potential unobserved/latent variables based on the following variables.

- GDP per capita
- Family
- Health/Life Expectancy
- Freedom
- Trust-in-government/Corruption
- Generosity

The RMSE is 0.0386 indicates that appropriate number of latent factors is two. As long as the RMSE is less than 0.05, it is safe to infer two unobserved factors.

Interpretations of EFA

Factor 1:

- Economy/GDP per capita
- Family
- Health/life expectancy

Factor 2:

- Freedom
- Trust-in-government/corruption
- Generosity

We can argue that factor 1 represents personal/household happiness and factor 2 represents societal happiness.

Table 5: EFA Summary

```
Loadings:
                             Factor1 Factor2
Economy..GDP.per.Capita.       0.992
Family                         0.616
Health..Life.Expectancy.       0.810
Freedom                                0.945
Trust..Government.Corruption.          0.445
Generosity                             0.432

                Factor1 Factor2
SS loadings       2.211   1.416
Proportion Var    0.368   0.236
Cumulative Var    0.368   0.604
```

Confirmatory Factor Analysis (CFA)

CFA Metrics

The SRMR, GFI, and AGFI is borderline; however, these numbers can confirm the model. The main limitation is that we have a potential ecological fallacy because we are analyzing country data as the unit of analysis to explain household happiness. A survey at the household level in each country (adjusted for population), might produce a better CFA model.

SRMR

The SRMR is 0.058. This is borderline against the arbitrary cut-off criteria of 0.05.

GFI

The GFI is 0.938. This is borderline against the arbitrary cut-off of 0.95.

AGFI

The AGFI is 0.837.

Confidence Interval

The 95% confidence interval for the correlation between society and household happiness is between 0.3489 and 0.6548. Given the sample size, we are 95% confident that there is a strong correlation between societal and household happiness.

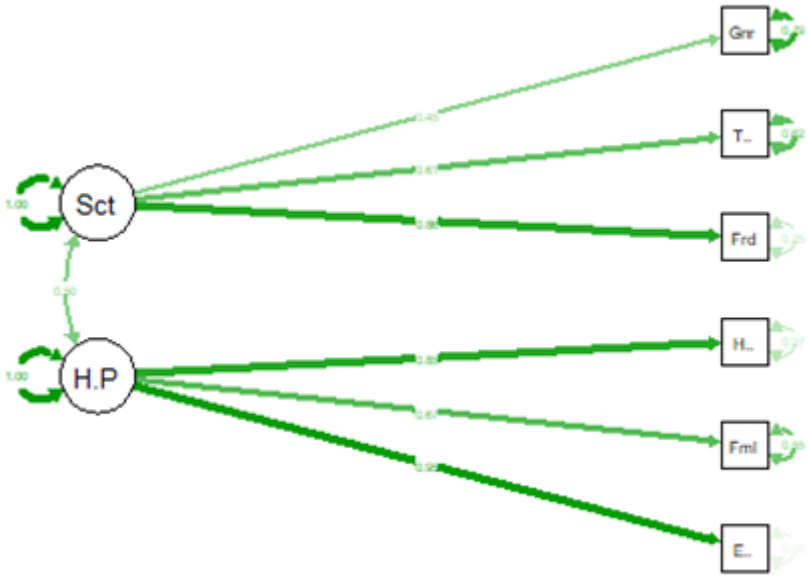

Figure 23: CFA Path Diagram

The correlation between household/personal happiness and societal happiness is 0.50. It appears that freedom and trust-in-government has more impact on societal happiness, and health and GDP per capita has more impact on household/personal happiness.

Conclusion and Recommendation

Based on our analysis, we can assume the data is a multinormal distribution. The happiness report contains two principle components that represents overall reality of our data.

Model-based clustering is the most precise clustering technique. Model-based clustering produced three groups with a 70% to 90% accuracy: North America/Western Europe/Australia, Latin America/North Africa/Northeast Asia, and South Asia/Africa.

Confirmatory Factor Analysis confirmed two latent factors that impact happiness: household/personal and societal happiness. In other words, happiness might be explained by factors inside the household and factors outside the household.

As a recommendation, further research might better explain the data to avoid an ecological fallacy to better explain individual happiness while controlling for education and environmental factors. It is very difficult to use aggregate data with countries as the unit of analysis to explain individual happiness.

Appendix

Data Cleaning & Visualization

Correlation and Missing Values

```
#import Data
world_happiness<- read.csv("C:\\Users\\Dinesh Poudel\\Desktop\\Multivariate P
roject\\2015.csv", row.names = "Country")
head(world_happiness)

happiness<- world_happiness[, c(-1,-2,-4, -11)] # Remove Region, Happiness Ra
nk, Happiness Score and Standard Error for the purpose of analysis
head(happiness)
happiness[happiness==0]<-NA

for(i in 1:ncol(happiness)){
  happiness[is.na(happiness[,i]), i] <- mean(happiness[,i], na.rm = TRUE)
}

library(corrplot)
newdata = cor(happiness[-1])
corrplot(newdata, method = "number")
```

Multivariate normality test for Outlier Detection

```
xbar <- colMeans(happiness[-1])
S <- cov(happiness[-1])
d2 <- mahalanobis(happiness[-1], xbar, S)
sd2 <-sort(d2)

quantiles <-qchisq((1:nrow(happiness[-1])-1/2)/nrow(happiness[-1]), df=ncol(h
appiness[-1]))
plot(quantiles, sd2,
     xlab=expression(paste(chi[3]^2,"Quantile")),
     ylab="Ordered squared distances", main="")
abline(a=0, b=1)
text(quantiles, sd2, abbreviate(names(sd2)), col="red", pch=0.8)

#five outlier countries matched with the data-frame
outcountry <- match(lab<-c("Myanmar", "Botswana", "Rwanda", "Syria", "Qatar",
"Somaliland region"), rownames(happiness))
clean_happiness_withscore <- happiness[-outcountry,]
clean_happiness <- clean_happiness_withscore[-1]  #drop happiness score
```

Multivariate normality test with clean data

```
xbar <- colMeans(clean_happiness)
S <- cov(clean_happiness)
```

```
d2 <- mahalanobis(clean_happiness, xbar, S)
sd2 <-sort(d2)

quantiles <-qchisq((1:nrow(clean_happiness)-1/2)/nrow(clean_happiness), df=nc
ol(clean_happiness))
plot(quantiles, sd2,
     xlab=expression(paste(chi[3]^2,"Quantile")),
     ylab="Ordered squared distances", main="")
abline(a=0, b=1)
text(quantiles, sd2, abbreviate(names(sd2)), col="red", pch=0.8)
```

Scatterplot of Clean Data

```
plot(clean_happiness) #scatterplot of clean data
```

Dimension Reduction Analysis

PCA

```
happiness_pca<- princomp(clean_happiness, cor=T)
summary(happiness_pca, loading=T)

# check correlation with the data
cor(clean_happiness$Economy..GDP.per.Capita., happiness_pca$scores[,1])

#PCA Biplot
biplot(happiness_pca, cex=0.6)
```

Multidimensional Scaling

```
s_dist <-dist(scale(clean_happiness))
mydata.mds <-cmdscale(s_dist, k=2, eig=T)
cumsum(mydata.mds$eig)/sum(mydata.mds$eig)

# MDS for observations
plot(mydata.mds$points, pch='.', xlab ="Coordinate 1", ylab="Coordinate 2")
text(mydata.mds$points, labels=rownames(clean_happiness), cex=0.7)
```

```
# MDS for variables
dist_corr<-1-cor(clean_happiness)
mydata.mds2 <-cmdscale((dist_corr), k=3, eig=T)
plot(mydata.mds2$points,xlim=c(-1.2, 1.2), ylim=c(-1,1), pch='.', xlab ="Coor
dinate 1", ylab="Coordinate 2")
text(mydata.mds2$points, labels=colnames(clean_happiness), cex=0.7)
```

Canonical Correlation Analysis(CCA)

```
X <- scale(clean_happiness[, 1:3])
Y<- scale(clean_happiness[, 4:6])

library(CCA)
cca <- cc(X, Y)
a <- cca$xcoef

U<-cca$scores$xscores
#U1 scores is the first column of xscores
head(U)

V<-cca$scores$yscores
#V1 scores is the first column of xscores
head(V)

round(cca$cor, 3)
a<-cca$xcoef
a1<-a[,1]/min(a[,1])

b<-cca$ycoef
b1<-b[,1]/min(b[,1])
```

$$U_1 = -0.004x_1 + 1x_2 + 0.670x_3$$

$$V_1 = 1y_1 + 0.148y_2 - 0.080y_3$$

Cluster Analysis

Hierchical-single

```
mydata.s <- scale(clean_happiness)
dist <- dist(mydata.s)  #distance matrix
hc1 <- hclust(dist, "single")
plot(hc1, hang=-1, cex=0.5, main ="Single Linkage HC Dendogram")  #dendogram
```

```
#scree plot
plot(rev(hc1$height), xlim=c(1,20))

ct<- cutree(hc1, 2)
table(ct)

pca <- princomp(mydata.s)
plot(pca$scores[, 1:2], col=ct)

plot(pca$scores[, 2:3], col=ct)

plot(pca$scores[, c(1,3)], col=ct)
```

Hierchical-complete

```
dist <- dist(mydata.s)  #distance matrix
hc1 <- hclust(dist, "complete")
plot(hc1, hang=-1, cex=0.5, main ="Complete Linkage HC Dendogram")#dendogram

#scree plot
plot(rev(hc1$height),xlim=c(1,20)) # 6 clusters, may be?

ct<- cutree(hc1, 3)
table(ct)

pca <- princomp(mydata.s)
plot(pca$scores[, 1:2], col=ct)

plot(pca$scores[, 2:3], col=ct)

plot(pca$scores[, c(1,3)], col=ct)
```

Hierchical-Average

```
dist <- dist(mydata.s)  #distance matrix
hc1 <- hclust(dist, "average")
plot(hc1, hang=-1, cex=0.5, main ="Average Linkage HC Dendogram") #dendogram

#scree plot
plot(rev(hc1$height),xlim=c(1,20)) # 6 clusters, may be?

ct<- cutree(hc1, 3)
table(ct)

pca <- princomp(mydata.s)
plot(pca$scores[, 1:2], col=ct)

plot(pca$scores[, 2:3], col=ct)

plot(pca$scores[, c(1,3)], col=ct)
```

K-Means

```
km <- kmeans(mydata.s, centers=3, nstart = 10)
plot.wgss = function(mydata, maxc) {
  wss = numeric(maxc)
  for (i in 1:maxc)
    wss[i] = kmeans(mydata,centers=i, nstart = 10)$tot.withinss
  plot(1:maxc, wss, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares", main="Scree Plot")
}
plot.wgss(mydata.s, 20)
```

```
pca <- princomp(mydata.s)
plot(pca$scores[, 1:2], col=km$cluster)

plot(pca$scores[, 2:3], col=km$cluster)

plot(pca$scores[, c(1,3)], col=km$cluster)
```

Model-based

```
library(mclust)
mc<-Mclust(mydata.s)
summary(mc)

plot(mc, what="BIC")

table(mc$classification)

clust.data = cbind(rownames(mydata.s), mc$classification, mc$uncertainty)
#clust.data[order(mc$uncertainty),]

plot(pca$scores[, 1:2], col=mc$classification)

plot(pca$scores[, 2:3], col=mc$classification)

plot(pca$scores[, c(1,3)], col=mc$classification)
```

```
library(maptools)
data(wrld_simpl)

cluster1=subset(rownames(mydata.s), mc$classification==1)
cluster2=subset(rownames(mydata.s), mc$classification==2)
cluster3=subset(rownames(mydata.s), mc$classification==3)
cluster=c(cluster1, cluster2,cluster3)

myCountries = wrld_simpl@data$NAME %in% cluster1
plot(wrld_simpl, col = c(gray(.80), "red")[myCountries+1])

myCountries = wrld_simpl@data$NAME %in% cluster2
plot(wrld_simpl, col = c(gray(.80), "green")[myCountries+1])

myCountries = wrld_simpl@data$NAME %in% cluster3
plot(wrld_simpl, col = c(gray(.80), "yellow")[myCountries+1])
```

Model-Based Discriminant Clustering

```
happiness1 <- clean_happiness_withscore
happiness1$Group[happiness1$Happiness.Score< 4]<-"Low"
happiness1$Group[happiness1$Happiness.Score>=4 & happiness1$Happiness.Score<6
]<-"Medium"
happiness1$Group[happiness1$Happiness.Score>=6]<-"Higher"

happiness1 =happiness1[sample(1:152, 152),]

data.train = happiness1[1:120, c(-1,-8)]
label.train = happiness1[1:120,8]
data.test = happiness1[121:152,c(-1,-8)]
label.test = happiness1[121:152,8]

DA <- MclustDA(data.train, label.train)
summary(DA, newdata = data.test, newclass = label.test)
```

EFA & CFA

EFA

```
happiness_fa <- factanal(mydata.s, factors=2, scores="regression")
happiness_fa #for large dataset p value may not describe our null hypothesis.

corHat <- happiness_fa$loadings %*% t(happiness_fa$loadings) + diag(happiness
_fa$uniquenesses)
corr <- cor(mydata.s)
rmse=sqrt(mean((corHat-corr)^2))
```

CFA

```
library(lavaan)
happiness.model <- 'Public =~ Economy..GDP.per.Capita. + Family +
                              Health..Life.Expectancy.
                    Personal=~ Freedom + Trust..Government.Corruption. +
                               Generosity'

options(fit.indices = c("GFI", "AGFI", "SRMR")) # Some fit indices
fit.cfa <- cfa(happiness.model, sample.cov= cor(mydata.s), std.lv=T, sample.n
obs = nrow(mydata.s))
options(fit.indices = c("GFI", "AGFI", "SRMR")) # Some fit indices
summary(fit.cfa, fit.measures=TRUE)
fitMeasures(fit.cfa)[c("gfi", "agfi")]

library(semPlot)

semPaths(fit.cfa, rotation=2, 'std', 'est')
```

References

Everitt, B. S., & Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. New York: Springer.

Manly, B. F. J., & A., N. A. J. (2017). *Multivariate statistical methods: a primer*. Boca Raton: Chapman & Hall/CRC.

Sustainable Development Solutions Network. (2019, November 27). World Happiness Report. Retrieved December 2, 2019, from https://www.kaggle.com/unsdsn/world-happiness.