Final Assignment - Gender Statistics and the Happiness Index

FEM11149 - Introduction to Data Science

Eleftherios Tranakos - 617019lt

2023-10-30

Introduction

The United Nations has as its fifth Sustainable Development goal the gender equality and empowerment of women. The UN strongly believes that gender equality is a pioneer of development progress. Thus, in this report, our goal is to analyze the correlation of variables from gender statistics with the happiness index. Unquestionably, the question that arises is the following:

Which factors of gender statistics influence the happiness index most, and how are they interrelated?

Data

In this report, data originates from the World Bank's Gender Statistics Database, including areas such as demography, economic opportunities, education, and health. Insights are derived as well from the World Happiness Report (WHP), which implements the World Happiness Index to evaluate life satisfaction in over 150 countries. The research relies on the exploitation of three datasets: one with gender statistics, another with happiness index data, and a third with particular data from three nations for out-of-sample prediction. These sources provide an adequate basis for exploring the relationship between gender statistics and overall well-being.

As regards the data manipulation, we proceed with the necessary process to reduce the dimension and improve the attribution of our data. The process was initiated by excluding non-available values, merging, and deleting countries either because they were twice as likely in our data set or because they did not include a lot of information for our analysis. Lastly, for the principal components analysis, we used numeric variables, while for the Lasso Penalized Regression Model, binary variables were included in addition.

Methodology

On the first part of our research, **Principal Components Analysis (PCA)** took the initiative. The PCA is a reduction of dimensionality technique which clusters correlated variables of massive data sets, based on similar characteristics, into new clusters of uncorrelated variables known as principal component (PC). The first PC absorbs the greatest variance in the data while following components gather less.

Each major component can be described mathematically as a linear combination of the initial variables.:

$$PC_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p$$

Where PC_i is the i^{th} principal component, X_j is the j^{th} variable, and a_{ij} represents the loading of the j^{th} variable on the i^{th} component.

In order to evaluate the significance of the PCA results we use the **Bootstrap** and **Permutation Test**. The former estimates the distribution of a statistic(ex.mean, variance) by sampling with replacement from the data.

- **Mean**:

$$\bar{T^*} = \frac{1}{B}\sum_{b=1}^{B} T_b^*$$

- **Variance**:

$$\mathrm{Var}(T^*) = \frac{1}{B-1}\sum_{b=1}^{B}(T_b^* - \bar{T}^*)^2$$

The latter is almost the same with the former, evaluating the PCA findings to results from random permutations of the data,including sampling without replacement, in order to determine how significant the PCA results are.

As regard the regression model, **principal components regression (PCR)** use the optimal number of principal components as predictors to forecast the dependent variable.

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{PC}_1 + \beta_2\mathbf{PC}_2 + ... + \beta_k\mathbf{PC}_k + \epsilon$$

where $k \leq p$ is the number of selected principal components and ϵ is the error term.

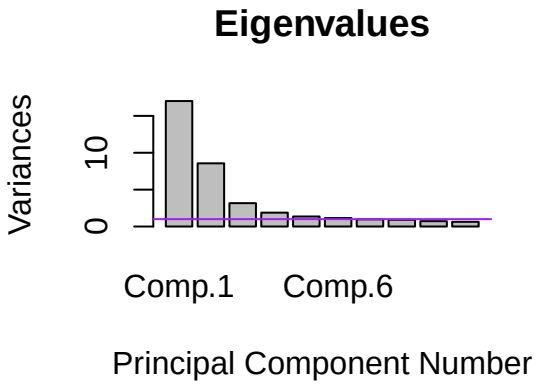
On the second part of our research, we took advantage of **Lasso** penalized regression model which applies variables selection (reducing/eliminating β for some independents variables) minimizing model complexity. The penalty term/parameter, λ, defines the weight of the penalty of each coefficient. However, we need to tune every time for the ideal value.

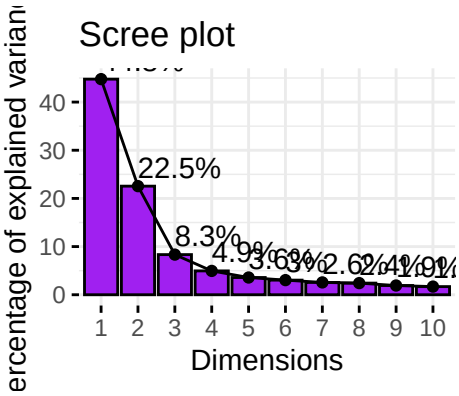
$$min(\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \lambda\sum_{i=1}^{n}|\beta_i|)$$

Parameter *lambda* tuning is accomplished via repeated cross-validation, in which the training dataset is split into n folds, in the model trained on n-1 folds and tested on the remaining one fold n times for each fold, for each *lambda*.We take the λ average accuracy metric (MSE in our analysis) for each n trials. The model is trained with parameter $\lambda = \lambda_{\min}$ to produce an optimal model, and the one with **the lowest MSE** is selected ($\lambda_{\min}$).

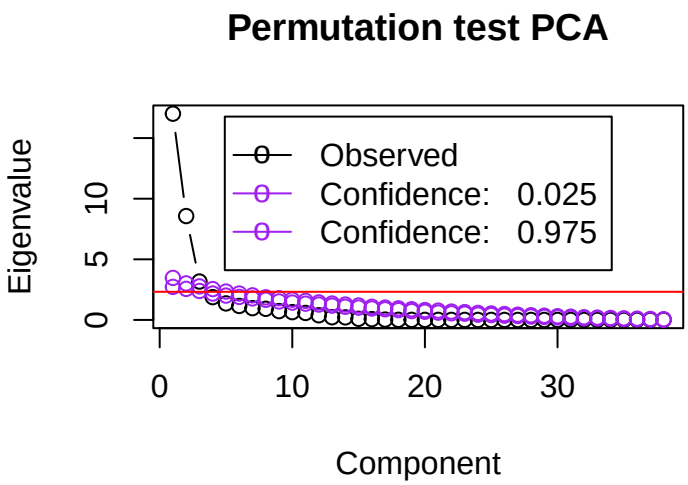
Results

Heading to the results part, applying PCA and checking the principal components at below figure, based on Kaiser's Rule, we should proceed with 5 principal components(maximum, with eigenvalues over 1). Nevertheless, staring at the elbow in the curve, on the next figure, in which the slope changes from steep to shallow, we can select the optimal number of components which is three and explains at least 70% (75.6%) of the variance in our data set. Specifically, the first component explains 44,8%, the second 22.5% and the third 8.3%.

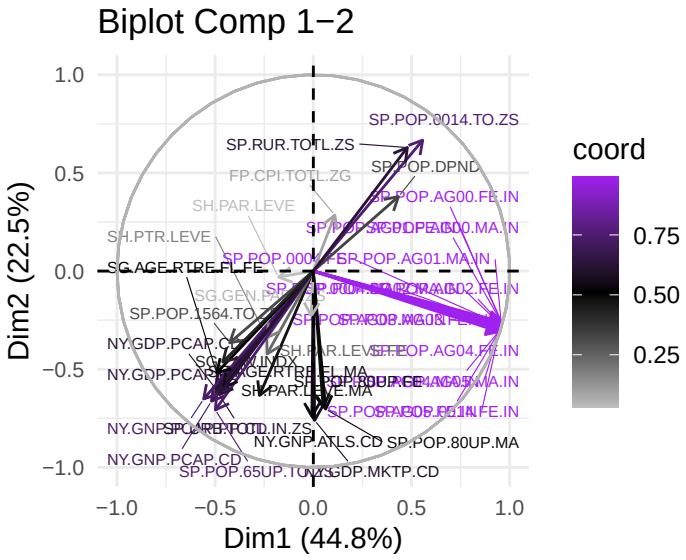




To endorse our choice, we check the importance of the principal components using the permutation test. From the figure below, we can see that the first 3 principal components are significant at a 95% confidence interval. In simple words,the later stands for the range of eigenvalues where 95% of the permuted data sets should occur within each component's eigenvalue range.



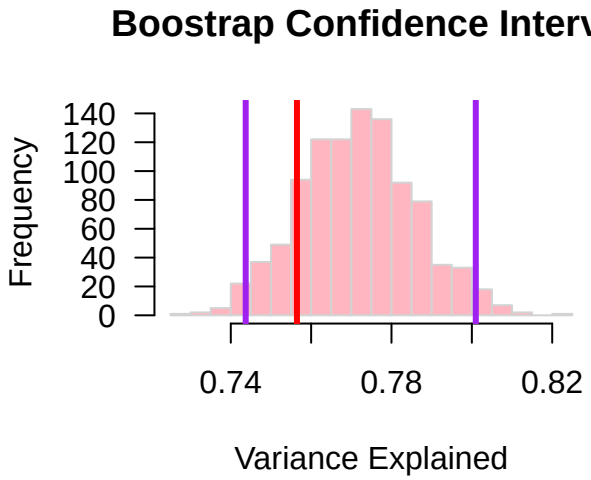
Having chosen the number of components, the below biplot, can reveal the variance explained in our data set by the first 2 components. The length of the vectors in the biplot indicates the strength of the respective variable in PCA. In our case, the purple vector which represents the variables of population from 0-14 years old (ex.SP.POP.0014.FE.IN) is positively related to component 1 in Appendix which is also negatively related to NY.GNP.PCAP.PP.CD, GNI per capita(gross national income). On the other hand, component 2 is negatively connected with NY.GDP.MKTP.CD which stands for GDP(current US$) and positively related with SP.RUR.TOTL.ZS which represents the total rural population in %.



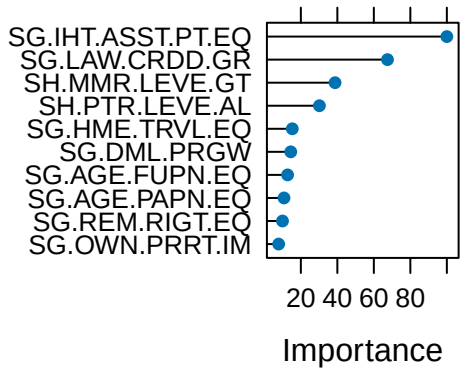
In addition, we can examine loadings in Appendix in order to see more details of each variable on the components.Component 1 can be interpreted as “Demographic” and it is expected that people for ages between 0-14 years old will have low income.Component 2 can be interpreted as “Economic Indicator” having reversed relation with people over 65 years old as the majority of the latter are retired and they do not contribute as much on the GDP. Lastly,Component 3 can be interpreted as “Socio-demographic Metrics” related with the length of paid parental leave for mother (SH.PAR.LEVE.FE) in contrast to the seats held by women in national parliaments(SG.GEN.PARL.ZS).

The below figure, shows the distribution of variance explained by three components based on bootstrap samples. The red lines indicate the 95% confidence interval meaning that the true variance that is explained by the three components occur in this range. The bell-shaped distributions

reveals that values all around the center are most probably to occur ecpecially the ones near 75,6%.



Heading further to Lasso penalized regression model,we started with 150 variables and after the test, we ended up with 33, as the most important for model predictions. As it can reduce the coefficient of some variables to zero, we were expecting to have lower number of variables.As we can see on figure below,the most important one is the "SG.IHT.ASST.PT.EQ", Sons and daughters have equal rights to inherit assets from their parents, having a strong positive effect on the dependent variable. The "SG.LAW.CRDD.GR",law prohibits discrimination in access to credit based on gender, is the second most important variable with negative coefficient, indicating that when it increases predicted value decreases.



Conclusion

To conclude, our analysis presented that the optimal number of components that explains most of the variance in our data set is three, while the first component(screeplot proof) captures the greatest variance. Furthermore, the principal component regression (pcr) performed better in contrast to that of lasso having the lowest mean square of errors(mse). As a result, pcr has more accurate predictions about the happiness score of Costa Rica,Croatia and Peru. A table of the models comparison's predictions can be found on the appendix.

In addition, in terms of predicting accurately results about happiness score in comparison with gender statistics further steps should be take.Specifically, as we saw, demographic and Economic Indicator factors (Comp1-2 respectively) contribute in a great extend to happiness score and more representative data should

be collected from now on. In simple words, there was a lack of genders diversities characteristics which is a crucial information that can be taken into account for predictin happiness score.

Regarding our analysis, the pcr with 4 components had a slight better performance in comparison with the 3 components (0.0062 difference) but we proceed with the 3 components for our predictions as we support our choice based on the permutation test.

Appendix

```
##                   Demographic Economic Indicators Socio-demographic Metrics
## SP.POP.DPND              0.11                0.13                       0.23
## SP.POP.AG00.FE.IN        0.23               -0.09                      -0.06
## SP.POP.AG00.MA.IN        0.23               -0.09                      -0.06
## SP.POP.AG01.FE.IN        0.23               -0.09                      -0.06
## SP.POP.AG01.MA.IN        0.23               -0.09                      -0.06
## SP.POP.AG02.FE.IN        0.23               -0.09                      -0.06
## SP.POP.AG02.MA.IN        0.23               -0.09                      -0.06
## SP.POP.AG03.FE.IN        0.23               -0.10                      -0.06
## SP.POP.AG03.MA.IN        0.23               -0.10                      -0.06
## SP.POP.AG04.FE.IN        0.23               -0.10                      -0.05
## SP.POP.AG04.MA.IN        0.23               -0.10                      -0.05
## SP.POP.AG05.FE.IN        0.23               -0.10                      -0.05
## SP.POP.AG05.MA.IN        0.23               -0.10                      -0.05
## NY.GDP.MKTP.CD           0.00               -0.26                       0.32
## NY.GDP.PCAP.KD          -0.12               -0.21                      -0.18
## NY.GDP.PCAP.CD          -0.12               -0.20                      -0.20
## NY.GNP.PCAP.CD          -0.13               -0.23                      -0.16
## NY.GNP.PCAP.PP.CD       -0.14               -0.22                      -0.18
## NY.GNP.ATLS.CD           0.00               -0.25                       0.35
## FP.CPI.TOTL.ZG           0.03                0.10                       0.06
## SH.PAR.LEVE.MA          -0.07               -0.22                       0.22
## SH.PAR.LEVE.FE          -0.06               -0.14                       0.19
## SH.PTR.LEVE             -0.06               -0.12                      -0.09
## SH.PAR.LEVE             -0.04               -0.01                      -0.05
## SP.POP.0004.FE           0.23               -0.09                      -0.06
## SP.POP.0004.MA           0.23               -0.09                      -0.06
## SP.POP.0014.TO.ZS        0.14                0.23                       0.09
## SP.POP.0014.FE.IN        0.23               -0.11                      -0.05
## SP.POP.1564.TO.ZS       -0.10               -0.12                      -0.23
## SP.POP.65UP.TO.ZS       -0.12               -0.24                       0.05
## SP.POP.80UP.FE           0.01               -0.24                       0.38
## SP.POP.80UP.MA           0.02               -0.24                       0.38
## SG.GEN.PARL.ZS           0.00               -0.08                      -0.23
## SG.AGE.RTRE.FL.FE       -0.12               -0.18                      -0.04
## SG.AGE.RTRE.FL.MA       -0.11               -0.20                      -0.09
## SP.RUR.TOTL.ZS           0.12                0.21                       0.10
## SP.URB.TOTL.IN.ZS       -0.12               -0.21                      -0.10
## SG.LAW.INDX             -0.11               -0.16                      -0.13
```

```
##    Countries    X.PCR  X.Lasso
## 1 Costa Rica 6.041864 5.859685
## 2    Croatia 5.917416 5.911780
## 3       Peru 5.759691 5.635888
```

Code

```
#If_require
if (!require("readxl")) install.packages("readxl")
if (!require("tidyverse")) install.packages("tidyverse")
if (!require("dplyr")) install.packages("dplyr")
if (!require("corrplot")) install.packages("corrplot")
if (!require("ggplot2")) install.packages("ggplot2")
if (!require("factoextra")) install.packages("factoextra")
if (!require("boot")) install.packages("boot")
if (!require("gridExtra")) install.packages("gridExtra")
if (!require("caret")) install.packages("caret")
if (!require("pls")) install.packages("pls")
if (!require("Metrics")) install.packages("Metrics")
if (!require("kableExtra")) install.packages("kableExtra")
if (!require("knitr")) install.packages("knitr")
if (!require("tinytex")) install.packages("tinytex")
#libraries
library(readxl)
library(dplyr)
library(corrplot)
library(ggplot2)
library(factoextra)
library(boot)
library(gridExtra)
library(caret)
library(pls)
library(Metrics)
library(tidyverse)
library(kableExtra)
library(knitr)
library(tinytex)
#read datasets
setwd("C:/Users/lefte/R/datascience")
gender_variable<-read.csv("6170191t_Gender_variable.csv")
predictions<-read.csv("Prediction.csv")
happiness_score<-read_excel("Happiness_score.xls")
#function for pca
pca_permutation_test <- function (X, nTests = 100, alpha = 0.05, center.data
                                  = TRUE, scale.data = TRUE, ...){
# Perform a permutation test for PCA
n <- nrow(X)
m <- ncol(X)
X <- scale(X, center = center.data, scale = scale.data)
if (scale.data) {a <- 1/(n - 1)} else {a <- 1}
res.X      <- prcomp(X)
eigs.X     <- res.X$sdev^2
eigs.Xperm <- matrix(0, m, nTests)
Xperm      <- matrix(0, n, m)
Xperm[, 1] <- X[, 1];
for (i in 1:nTests){
    for (j in 2:m) {
      ind <- sort(runif(n), index.return = TRUE)$ix   # Find random permutation of values 1:n
```

```
      Xperm[, j] <- X[ind, j]
    }
    res.Xperm  <- prcomp(Xperm)
    eigs.Xperm[, i] <- res.Xperm$sdev^2
  }

perc.alpha <- matrix(0, m, 2)
for (s in 1:m){
  perc.alpha[s,] <- quantile(eigs.Xperm[s,], c(alpha/2, 1 - alpha/2) )
  }
plot(1:m, eigs.X, type = "b", col = "black", main = "Permutation test PCA",
     xlab = "Component", ylab = "Eigenvalue", ...)
  lines(1:m, perc.alpha[, 1], type = "b", col="purple")
  lines(1:m, perc.alpha[, 2], type = "b", col="purple")

#String1 <- sprintf('%4.1f%% Confidence',Alpha/2)
#String2 <- sprintf('%4.1f%% Confidence',100-Alpha/2)
#legend('Data',String1,String2)
string1 <- paste("Confidence: ",formatC(alpha/2, digits=3, width=6, format="f"))
string2 <- paste("Confidence: ",formatC(1-alpha/2, digits=3, width=6, format="f"))
 legend("top", inset=.05, c("Observed", string1, string2),
  lty = c(1, 1, 1), col = c("black", "purple", "purple"), pch = c("o", "o", "o"))

}

#DATA MANIPULATION
#exclude NA values
reduced_gender_variable <-gender_variable[, colSums(is.na(gender_variable)) <
                                           nrow(gender_variable)]
#str(reduced_gender_variable)

#order datasets alphabetically based on Country Names
reduced_gender_variable<-reduced_gender_variable
[order(reduced_gender_variable$CountryName),]
happiness_score <- happiness_score[order(happiness_score$`Country name`),]

#datacleaning-countries names/merging countries/del val that are not countries
#add Hap Score of China + Hong Kong and take the average / delete Hong Kong row
happiness_score$Happiness_score[23]<- (happiness_score$Happiness_score[23] + happiness_score$Happiness_
happiness_score <- happiness_score[-48, ]
# edit the name of Venezuela from reduced_gender_variable dataset
reduced_gender_variable$CountryName[reduced_gender_variable$CountryName ==
                                      "Venezuela, RB"] <- "Venezuela"
# Korea, Rep. = South Korea / change the name in reduced_gender_variable dataset
reduced_gender_variable$CountryName[reduced_gender_variable$CountryName ==
                                      "Korea, Rep."] <- "South Korea"
reduced_gender_variable <- reduced_gender_variable[,-1]

#change the name of the column:Country name of hap score data set for merging
names(happiness_score)[names(happiness_score) == "Country name"]<-
  "CountryName"
#Merge Datasets - reduced_gender_variables / happiness score
merged_datasets<- merge(reduced_gender_variable, happiness_score, by=
```

```
                          "CountryName")

#bring hapiness score column on the 3nd position
merged_datasets <- merged_datasets[, c(names(merged_datasets)[1:2],

"Happiness_score", names(merged_datasets)[-c(1:2, which(names(merged_datasets)
                                                    == "Happiness_score"))])]
#str(merged_datasets)
#Keeping mainly values with numbers
merged_datasets <- merged_datasets[, colSums(!is.na(merged_datasets)) >
                                   colSums(is.na(merged_datasets))]

#exclude Venezu-Afghan from the data set as they includes a lot of NAs values
merged_datasets<-merged_datasets[-c(1,54),]

#replace NAs values with the average value per column
merged_datasets[] <- lapply(merged_datasets, function(x) ifelse(is.na(x),
                                                    mean(x, na.rm = TRUE), x))
lasso_dataset<-merged_datasets #creating dataset for lasso for later use

#excluding binary 0-1 numbers because differ in scale from continuous variables
# binary variables can complicate the interpretation of the components
merged_datasets <- merged_datasets[, sapply(merged_datasets, function(col)
  !all(col %in% c(0, 1)))]

####DELETE Columns with numeric values  which resembles to the ones of
categories . ex.0-25-50-75-100
merged_datasets[,c("SG.LAW.INDX.AS", "SG.LAW.INDX.EN", "SG.LAW.INDX.MR", "SG.LAW.INDX.MO","SG.LAW.INDX.

#Converting all variables into numeric apart from the first 2 columns
merged_datasets[,-(1:2)] <- lapply(merged_datasets[,-(1:2)], function(x)
  as.numeric(as.character(x)))
#str(merged_datasets)## we tried to exclude variables
#of ages 0-5, 6-14, (since are covered by the 0-14 var) but we lost a lot of performance.
##the variables 0-5 ages are quite important for our analysis
#correlation matrix
cor_matrix <- cor(merged_datasets[, sapply(merged_datasets, is.numeric)],
                  use = "pairwise.complete.obs")
#cor_matrix
# Print correlation with 'Happiness_score'
cor_matrix_hap<-cor_matrix[,"Happiness_score"]
#cor_matrix_hap
  average_score<-mean(abs(cor_matrix_hap))
average_score
#Keep variables over the average_score/average correlation
condition_check <- abs(cor_matrix_hap) > average_score
high_corr <- abs(cor_matrix_hap[condition_check])
#create new data set based on the variables that its correlation is above
#the average correlation with the happiness score
#names(high_corr)
new_data<- merged_datasets[,names(high_corr)] #does not include
#countrynames-countrycodes
#PCA
```

```r
pca_data<-new_data[,-1]
pca <- princomp(pca_data, cor=T, scores=TRUE)
#summary(pca) #3components explain more than 70%
pca_loadings_3comp<-round(pca$loadings,2)[,1:3]

colnames(pca_loadings_3comp) <- c("Demographic", "Economic Indicators",
                                  "Socio-demographic Metrics")
#pca_loadings_3comp #named our components based on what are related more
sum(pca$sdev^2)
#pca_loadings_3comp
#figure 1
# Eigenvalues Plot
plot(pca, main = "Eigenvalues", xlab = "Principal Component Number")
# Draw a line to indicate the Kaiser criterion (Eigenvalue of 1)
abline(h = 1, col = "purple", lty = 1)
legend("topright", legend = "Kaiser Criterion", col = "purple",
       lty = 1, cex = 0.8)
#figure 2
#Variance explained by the components
plot2<-fviz_eig(pca, addlabels = TRUE, ncp = 10, barfill = "purple",
                barcol = "black")


pca_permutation_test(pca_data)
abline(2.32, 0,col="red") #3
fviz_pca_var(pca, axes = c(1, 2), col.var = "coord",
                      gradient.cols = c("gray", "black", "purple"),
                      repel = TRUE, labelsize = 2, title = "Biplot Comp 1-2") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8))
pca_boot<-function(x,ind){
  res<-princomp(x[ind,],cor=TRUE)
  return(res$sdev^2)
}

#run boostrap
fit.boot<-boot(data = pca_data, statistic = pca_boot, R=1000)
#store the bootstrapped statistic(all eigenvalues)
eigenvalues.boot<-fit.boot$t
variance_explained <- eigenvalues.boot[, 1] / ncol(pca_data)
#variance explained by the number of components we chose
#head(eigenvalues.boot)
variance_explained <- rowSums(eigenvalues.boot[, 1:3]) / rowSums(eigenvalues.boot)
hist(variance_explained, xlab="Variance Explained", las=1, col="lightpink",
     main = "Boostrap Confidence Interval", breaks=20, border="lightgrey")

perc.alpha<-quantile(variance_explained,c(0.025,1 - 0.025))
abline(v=perc.alpha,col="purple", lwd=3)
abline(v=sum(pca$sdev[1:3]^2)/sum(pca$sdev^2), col="red", lwd=3)

##Lasso Regression##
#taking the data set for lasso that we created earlier #55*150
#str(lasso_dataset)
```

```r
lasso_dataset_numeric<-lasso_dataset[,-c(1:2)]
#creating test data and train data
set.seed(123)
trainIndex <- createDataPartition(lasso_dataset_numeric$Happiness_score,
                                  p = 0.8, list = FALSE)
train_data_lasso <- lasso_dataset_numeric[trainIndex, ]
test_data_lasso <- lasso_dataset_numeric[-trainIndex, ]
#CARET
fitControl <- trainControl(method = "repeatedcv", number = 10, repeats = 5,
                           verboseIter = TRUE)
#caret package
lasso_model <- train(Happiness_score ~ .,
               train_data_lasso,
               method = 'glmnet',
               tuneGrid = expand.grid(alpha = 1,
                                      lambda = seq(0.0001, 1, length =100)),
               trControl = fitControl
)
lasso_model$bestTune #best lambda
plot(lasso_model$finalModel, xvar = "lambda", label = TRUE)
best_lambda <- lasso_model$bestTune$lambda
# Get the coefficients for the best lambda
coefficients <- coef(lasso_model$finalModel, s = best_lambda)
# Print the coefficients with no zero values
non_zero_coefficients <- coefficients[coefficients[,1] != 0, , drop = FALSE]
print(non_zero_coefficients) ## var

plot(varImp(lasso_model, scale = TRUE), top = 17)

##Predictions for PCA /PCR
#spliting our dataset to training and test data / 80%-20%
set.seed(123)
new_data # 55 obs and 39 variables (including Happiness score)
turtle <- createDataPartition(new_data$Happiness_score, p = 0.8, list = FALSE)
train_data_pcr <- new_data[turtle, ]
test_data_pcr <- new_data[-turtle, ]
#Model training
train_pcr_model <- pcr(Happiness_score ~ ., data = train_data_pcr,scale =TRUE,
                       validation = "CV")
lm_model <- lm(Happiness_score ~ ., data = new_data)
summary(lm_model)
#Model testing
pcr_prediction3 <- predict(train_pcr_model, test_data_pcr, ncomp = 3)
summary(train_data_pcr)
mse_value3<-mse(actual = test_data_pcr$Happiness_score,
                predicted = as.numeric(pcr_prediction3))
#Predicting Happiness score to the predictions dataset with 3 countries
#We extract the names(var-col) that we used in pca and we put it in a new
dataset and we create a final dataset with the names and the 3 countries
prediction_names_pcr<- names(new_data[-1])
final_prediction_dataset_pcr <- (predictions[, prediction_names_pcr])
sum(is.na(final_prediction_dataset_pcr))  #no NAs
pcr_prediction_3countries <- predict(train_pcr_model,
```

```
                                    final_prediction_dataset_pcr, ncomp = 3)
#pcr_prediction_3countries

##Lasso mse + Prediction
set.seed(123)
lasso_prediction <- predict(lasso_model, test_data_lasso,
                            s = lasso_model$bestTune)
mse_lasso<-mse(actual = test_data_lasso$Happiness_score,
               predicted = as.numeric(lasso_prediction))
mse_lasso
mse_value3
#We extract the names(var-col) that we used in lasso and we put
it in a new dataset
#and we create a final dataset with the names and the 3 countries
prediction_names_lasso<- names(lasso_dataset_numeric[-1])
final_prediction_dataset_lasso <- (predictions[, prediction_names_lasso])
lasso_prediction_3countries <- predict(lasso_model,
                      final_prediction_dataset_lasso, s = lasso_model$bestTune)

# Create a data frame for happiness score for pcr - lassoo
happiness_data_predictions <- data.frame(
  Countries = c("Costa Rica", "Croatia", "Peru"),
  ` PCR` = c(6.041864,5.917416,5.759691),
  ` Lasso` = lasso_prediction_3countries
)
plot(pca, main = "Eigenvalues", xlab = "Principal Component Number")
abline(h = 1, col = "purple", lty = 1)
plot2
#permutation test - Called function
pca_permutation_test(pca_data)
abline(2.32, 0,col="red") #3
# Creating biplot for 1-2 comp
fviz_pca_var(pca, axes = c(1, 2), col.var = "coord",
                      gradient.cols = c("gray", "black", "purple"),
                      repel = TRUE, labelsize = 2, title = "Biplot Comp 1-2") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8))
#variance explained by the number of components we chose
#head(eigenvalues.boot)
variance_explained <- rowSums(eigenvalues.boot[, 1:3]) / rowSums(eigenvalues.boot)
hist(variance_explained, xlab="Variance Explained", las=1, col="lightpink",
     main = "Boostrap Confidence Interval", breaks=20, border="lightgrey")

perc.alpha<-quantile(variance_explained,c(0.025,1 - 0.025))
abline(v=perc.alpha,col="purple", lwd=3)
abline(v=sum(pca$sdev[1:3]^2)/sum(pca$sdev^2), col="red", lwd=3)
pca_loadings_3comp
happiness_data_predictions
```

References