df <- read.csv("datasetHA.csv")

Data Pre-processing

head(df)

## Reference.Area Time.Period Sex Age.group Units.of.measurement  
## 1 Albania 2007 Not applicable Not applicable Number  
## 2 Albania 2008 Not applicable Not applicable Number  
## 3 Algeria 2001 Not applicable Not applicable Number  
## 4 Algeria 2002 Not applicable Not applicable Number  
## 5 Algeria 2003 Not applicable Not applicable Number  
## 6 Algeria 2004 Not applicable Not applicable Number  
## Observation.Value  
## 1 845500000  
## 2 1665500000  
## 3 9734253000  
## 4 16571247000  
## 5 10306455000  
## 6 10058086000

keeps <- c("Reference.Area", "Time.Period","Observation.Value")  
df <- df[keeps]  
head(df)

## Reference.Area Time.Period Observation.Value  
## 1 Albania 2007 845500000  
## 2 Albania 2008 1665500000  
## 3 Algeria 2001 9734253000  
## 4 Algeria 2002 16571247000  
## 5 Algeria 2003 10306455000  
## 6 Algeria 2004 10058086000

sum(is.null(df))

## [1] 0

Normalizing the observation values.

library(caret)

## Warning: package 'caret' was built under R version 4.1.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.1.3

## Loading required package: lattice

process <- preProcess(as.data.frame(df$Observation.Value), method = c("range"))  
scale <- predict(process, as.data.frame(df$Observation.Value))  
df <- cbind(df, scale)  
colnames(df) <- c("ReferenceArea", "TimePeriod", "ObservationValue", "Scale")  
head(df)

## ReferenceArea TimePeriod ObservationValue Scale  
## 1 Albania 2007 845500000 1.325951e-05  
## 2 Albania 2008 1665500000 2.612545e-05  
## 3 Algeria 2001 9734253000 1.527257e-04  
## 4 Algeria 2002 16571247000 2.599993e-04  
## 5 Algeria 2003 10306455000 1.617036e-04  
## 6 Algeria 2004 10058086000 1.578067e-04

Data Splitting

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.3

##   
## Attaching package: 'dplyr'

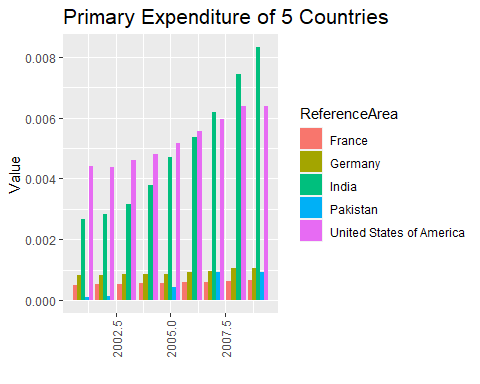
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

df %>% filter(ReferenceArea == "India" ) -> a1  
df %>% filter(ReferenceArea == "France" ) -> a2  
df %>% filter(ReferenceArea == "Germany" ) -> a3  
df %>% filter(ReferenceArea == "Pakistan" ) -> a4  
df %>% filter(ReferenceArea == "United States of America" ) -> a5  
  
train <- rbind(a1, a2, a3, a4, a5)  
  
train %>% filter(TimePeriod > 2000 & TimePeriod < 2010) -> train

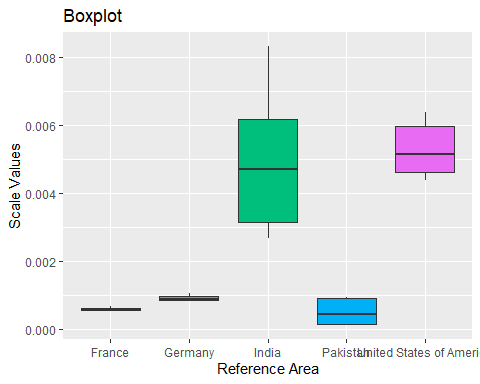
Bar Plot

ggplot(train, aes(fill=ReferenceArea, y=Scale, x=TimePeriod))+geom\_bar(position="dodge", stat="identity")+theme(axis.text.x =element\_text(angle=90,vjust=0.5),plot.title=element\_text(color="black",size=16))+labs(x="",y="Value",title="Primary Expenditure of 5 Countries")



Box Plot

ggplot(train,aes(ReferenceArea, Scale,fill=ReferenceArea))+geom\_boxplot(outlier.color="red",outlier.shape=4,outlier.size = 4)+theme(legend.position = "dodge")+ggtitle("Boxplot")+xlab("Reference Area")+ylab("Scale Values")



Meta Data

str(train)

## 'data.frame': 41 obs. of 4 variables:  
## $ ReferenceArea : chr "India" "India" "India" "India" ...  
## $ TimePeriod : int 2001 2002 2003 2004 2005 2006 2007 2008 2009 2001 ...  
## $ ObservationValue: num 1.70e+11 1.81e+11 2.01e+11 2.41e+11 2.99e+11 ...  
## $ Scale : num 0.00267 0.00284 0.00315 0.00378 0.0047 ...

summary(train)

## ReferenceArea TimePeriod ObservationValue Scale   
## Length:41 Min. :2001 Min. :7.018e+09 Min. :0.0001101   
## Class :character 1st Qu.:2003 1st Qu.:4.107e+10 1st Qu.:0.0006443   
## Mode :character Median :2005 Median :6.148e+10 Median :0.0009647   
## Mean :2005 Mean :1.681e+11 Mean :0.0026370   
## 3rd Qu.:2007 3rd Qu.:2.993e+11 3rd Qu.:0.0046965   
## Max. :2009 Max. :5.304e+11 Max. :0.0083223

Hypothesis Testing for Year 2010

df %>% filter(TimePeriod == 2010) -> Hypotrain  
meanObs <- mean(df$ObservationValue)  
cat("Mean of Population Data: ",meanObs)

## Mean of Population Data: 700900369239

stdDev <- sd(df$ObservationValue)  
cat("\n\nStandard Deviation of Population Data: ", stdDev)

##   
##   
## Standard Deviation of Population Data: 4.279967e+12

Samean <- mean(Hypotrain$ObservationValue)  
cat("\n\nMean of Sample Data: ",Samean)

##   
##   
## Mean of Sample Data: 893658701821

Pval<- pnorm(Samean,meanObs,stdDev)  
cat("\n\nP-Value: ",Pval)

##   
##   
## P-Value: 0.5179612

if(Pval < 0.05){  
 cat("\n\nReject null Hypothesis for 0.05")  
}else{  
 cat("\n\nDo not Reject null Hypothesis for 0.05")  
}

##   
##   
## Do not Reject null Hypothesis for 0.05

if(Pval < 0.01){  
 cat("\n\nReject null Hypothesis for 0.01")  
}else{  
 cat("\n\nDo not Reject null Hypothesis for 0.01")  
}

##   
##   
## Do not Reject null Hypothesis for 0.01

Principle Components

head(train)

## ReferenceArea TimePeriod ObservationValue Scale  
## 1 India 2001 170381500000 0.002673310  
## 2 India 2002 180881600000 0.002838059  
## 3 India 2003 200863400000 0.003151577  
## 4 India 2004 241172400000 0.003784032  
## 5 India 2005 299325800000 0.004696470  
## 6 India 2006 342383900000 0.005372059

prcomtrain <- train[, 2:3]  
  
head(prcomtrain)

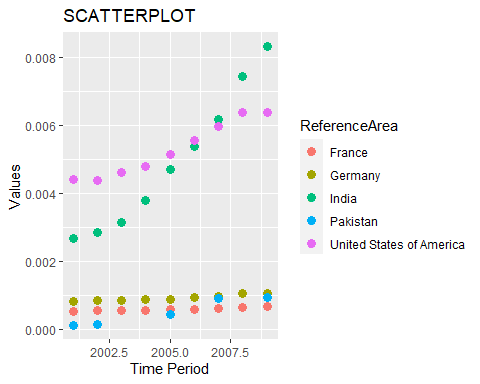
## TimePeriod ObservationValue  
## 1 2001 170381500000  
## 2 2002 180881600000  
## 3 2003 200863400000  
## 4 2004 241172400000  
## 5 2005 299325800000  
## 6 2006 342383900000

prcom <- prcomp(prcomtrain, scale. = TRUE)  
summary(prcom)

## Importance of components:  
## PC1 PC2  
## Standard deviation 1.1292 0.8514  
## Proportion of Variance 0.6376 0.3624  
## Cumulative Proportion 0.6376 1.0000

Cor-relation

library(ggplot2)  
ggplot(train,aes(TimePeriod,Scale,col=ReferenceArea))+geom\_point(size=3)+geom\_text(aes(label=""),size=3,nudge\_x=0.050)+labs(title="SCATTERPLOT",x="Time Period",y="Values")



cor(train$TimePeriod, train$Scale)

## [1] 0.2751963

Indi <- cor(a1$TimePeriod, a1$Scale)  
Frn <- cor(a2$TimePeriod, a2$Scale)  
Usa <- cor(a5$TimePeriod, a5$Scale)  
Ger <- cor(a3$TimePeriod, a3$Scale)  
Pak <- cor(a4$TimePeriod, a4$Scale)  
  
fdata <- data.frame(C\_name = rep(c('India', 'France', 'Germany', 'Pakistan', 'USA')),  
 Cor\_Val = rep(c(Indi, Frn, Ger, Pak, Usa)))  
fdata

## C\_name Cor\_Val  
## 1 India 0.9465078  
## 2 France 0.9958850  
## 3 Germany 0.9845807  
## 4 Pakistan 0.9594195  
## 5 USA 0.9931212

Regression

l1 <- lm(Scale~TimePeriod, a1)  
summary(l1)

##   
## Call:  
## lm(formula = Scale ~ TimePeriod, data = a1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.255e-03 -8.884e-04 -3.359e-05 6.942e-04 2.086e-03   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.223e+00 1.118e-01 -10.93 3.06e-08 \*\*\*  
## TimePeriod 6.127e-04 5.583e-05 10.97 2.92e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.001029 on 14 degrees of freedom  
## Multiple R-squared: 0.8959, Adjusted R-squared: 0.8884   
## F-statistic: 120.5 on 1 and 14 DF, p-value: 2.916e-08

#Predicted Values  
prd <- predict(l1)  
prd

## 1 2 3 4 5 6   
## 0.0025684612 0.0031811719 0.0037938827 0.0044065935 0.0050193042 0.0056320150   
## 7 8 9 10 11 12   
## 0.0062447257 0.0068574365 0.0074701472 0.0080828580 0.0086955688 0.0093082795   
## 13 14 15 16   
## 0.0001176182 0.0007303289 0.0013430397 0.0019557504

#Confirming the Predicted Values  
head(a1)

## ReferenceArea TimePeriod ObservationValue Scale  
## 1 India 2000 161987800000 0.002541611  
## 2 India 2001 170381500000 0.002673310  
## 3 India 2002 180881600000 0.002838059  
## 4 India 2003 200863400000 0.003151577  
## 5 India 2004 241172400000 0.003784032  
## 6 India 2005 299325800000 0.004696470

#Error Values  
er <- a1$Scale - prd  
er

## 1 2 3 4 5   
## -2.684977e-05 -5.078619e-04 -9.558241e-04 -1.255017e-03 -1.235272e-03   
## 6 7 8 9 10   
## -9.355454e-04 -8.726668e-04 -6.695838e-04 -4.032260e-05 2.394116e-04   
## 11 12 13 14 15   
## 1.040727e-03 2.085993e-03 1.280937e-03 9.346030e-04 6.140168e-04   
## 16   
## 3.032557e-04

#Scatter plot for India with Regression Line  
ggplot(a1, aes(TimePeriod, Scale))+geom\_point()+geom\_smooth(method = "lm", formula = y~x, col="red", se=F)

