Basic.R

Xiru

Mon Feb 19 22:27:17 2018

```r
# Load the data
PB2010 <- read.csv('./Data/BP Apprehensions 2010.csv')
PB2017 <- read.csv('./Data/PB Apprehensions 2017.csv')
monthly <- read.csv('./Data/PB monthly summaries.csv')

# Clean the data (the goal is to eliminate the extra row and column of the dataset, and also to change
PB2017 <- PB2017[,-14]
PB2017 <- PB2017[-10,]
PB2017 <- cbind(Sector=PB2010$Sector,as.data.frame(sapply(PB2017[2:ncol(PB2017)],function(x){as.integer

# Save cleaned data
saveRDS(PB2010,file="./Data/PB Apprehensions 2010.rds")
saveRDS(PB2017,file='./Data/PB Apprehensions 2017.rds')
saveRDS(monthly,file='./Data/PB monthly summaries.rds')



# PART A plots

# Compare by month

# write a function that would produce barplots by month
# PB1 and PB2 should be dataframes, and the function is designed to let PB2010 = PB1 and PB2017 = PB2
# n should be a string that indicates the month of the graph


month_graph <- function(PB1,PB2,n){
    month <- c('October','November','December','January','Feburary','March','April','May','June','July'
               'September')
    time <- which(month==n)
    par(mar=c(5, 4, 4, 7), xpd=TRUE) # adjustments for the barplot
    barplot(as.matrix(rbind(PB1[,time+1],PB2[,time+1])),col=c('red','orange'),xlab='Month',ylab='Number
            main=paste('Total Number of Apprehensions in', n,sep = ' '), names.arg = PB1[,1],
            beside = TRUE,las=2,cex.names = 0.5)
    legend("topright", inset=c(-0.15,0),
           legend = c('2010','2017'), cex=0.7,
           fill = c("red", "orange"))
}


# Compare by sector

# write a function that would produce barplots by sector
# PB1 and PB2 should be dataframes, and the functions is designed to take PB2010 as PB1 and PB2017 as P
# n should be a string that indicates the sector of the graph

sector_graph <- function(PB1,PB2,n){
```

```
    sector <- as.character(PB2010$Sector)
    s <- which(sector==n)
    par(mar=c(5, 4, 4, 7), xpd=TRUE)
    barplot(as.matrix(rbind(PB1[s,2:13],PB2[s,2:13])),col=c('green','blue'),ylab='Number of Apprehensio
            main=paste('Total Number of Apprehensions in',n,sep = ' '),cex.names=0.9,
            beside = TRUE, las=2)
    legend("topright",inset=c(-0.15,0),
           legend = c('2010','2017'), cex=0.7,
           fill = c("green", "blue"))
}


# PART B t-test by sector

# Sector with most apprehensions in 2010
PB2010$Total <- apply(PB2010[,-1],1,sum) # calculate total apprehensions for each sector in 2010
most_2010 <- PB2010[PB2010$Total==max(PB2010$Total),2:13] # pull out values of the sector with most app
most_2010_s <- as.character(PB2010$Sector[PB2010$Total==max(PB2010$Total)]) # gives the name of the sec

# Sector with most apprehensions in 2017
PB2017$Total <- apply(PB2017[,-1],1,sum) # calculate total apprehensions for each sector in 2017
most_2017 <- PB2017[PB2017$Total==max(PB2017$Total),2:13] # pull out values of the sector with most app
most_2017_s <- as.character(PB2017$Sector[PB2017$Total==max(PB2017$Total)]) # gives the name of the sec

# Combine most apprehensions in 2010 and 2017, with 2010 data in the first column and 2017 data in the
most <- t(rbind(most_2010,most_2017))
most <- apply(most,2,function(x) as.numeric(as.character(x)))

# a variance test first
# the result of the test is that two samples have same variance

var.test(most[,1],most[,2])
```

```
##
##  F test to compare two variances
##
## data:  most[, 1] and most[, 2]
## F = 0.93205, num df = 11, denom df = 11, p-value = 0.9092
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2683166 3.2376648
## sample estimates:
## ratio of variances
##           0.932051
```

```
# then the t-test to find out if the mean level of apprehension is higher in 2010 than in 2017
# with a significance level of 0.05, we have to reject the null hypothesis and conclude that the mean l
t.test(most[,1],most[,2],var.equal = TRUE,alternative = 'greater')
```

```
##
##  Two Sample t-test
##
## data:  most[, 1] and most[, 2]
## t = 1.9547, df = 22, p-value = 0.03172
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  756.0001      Inf
## sample estimates:
## mean of x mean of y
##   17683.5   11463.5
```

```r
# PART C t-test for three month period

# Three months periods

# write a function three_period that would return a data frame with the total number of apprehensions i
# 'object' should be a data frame, particularly, should be either PB2010 or PB2017

three_period <- function(object) {
  period <- data.frame() # an empty data frame first
  for (i in (1:nrow(object))) {
    for (j in (1:(ncol(object)-3))){
      period[i,j] <- object[i,j+1]+object[i,j+2]+object[i,j+3]
    }
  }
  period <- rbind(period,colSums(period))
  colnames(period) <- c('Oct-Dec','Nov-Jan','Dec-Feb','Jan-Mar','Feb-Apr','Mar-May','Apr-Jun',
                        'May-Jul','Jun-Aug','Jul-Sep')
  rownames(period) <- c(PB2017$Sector,'Total')
  period
}

t_month_2010 <- three_period(PB2010[,1:13]) # calculate the number of apprehensions in three-month peri
t_month_2017 <- three_period(PB2017[,1:13]) # calculate the number of apprehensions in three-month peri

# Find the maximum apprehension level for each year and return the time period
max_2010 <- t_month_2010[,t_month_2010[length(t_month_2010),]==max(t_month_2010[length(t_month_2010),])]
max_2017 <- t_month_2017[,t_month_2017[length(t_month_2017),]==max(t_month_2017[length(t_month_2017),])]

max_2010_period <- colnames(t_month_2010)[t_month_2010[length(t_month_2010),]==max(t_month_2010[length(t
max_2017_period <- colnames(t_month_2017)[t_month_2017[length(t_month_2017),]==max(t_month_2017[length(t

# Combine these two samples together under a new data frame
max_three <- as.data.frame(cbind(max_2010,max_2017))

# Perform a variance test first
# the result of the variance test is that two samples having the same variance

var.test(max_three$max_2010,max_three$max_2017)
```

```
##
##  F test to compare two variances
##
## data:  max_three$max_2010 and max_three$max_2017
## F = 1.43, num df = 9, denom df = 9, p-value = 0.6027
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3551856 5.7570716
## sample estimates:
```

```
## ratio of variances
##           1.429975
```

```r
# then the t-test
# with a significance level of 0.05, we can conclude that there is no difference in means of these two
t.test(max_three$max_2010,max_three$max_2017,var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  max_three$max_2010 and max_three$max_2017
## t = 0.25287, df = 18, p-value = 0.8032
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -39459.94  50258.74
## sample estimates:
## mean of x mean of y
##   32728.6   27329.2
```

```r
# PART D time series data

# order the dataset so that values in the column 'year' is in ascending order
monthly <- monthly[order(monthly$year),]

# yearly average apprehension
monthly$mean <- apply(monthly[,-1],1,mean)

ts <- ts(as.vector(t(monthly[,2:13])),start = c(2000,1),frequency=12) # create a time series object
# note that we manually adjusted the beginning date of the time series data so that it follows the fisc

# write a function for the time series plot, which would return not only the basic time series plot, bu
tsplot <- function(ts) {
  ts.plot(ts, gpars=list(xlab="Fiscal Year", ylab="Apprehensions", lty=c(1:3)),col='blue',main='Monthly
  label <- as.character(seq(from=2000,to=2017))
  for (i in 1:18) {
    segments(monthly$year[i],monthly$mean[i],monthly$year[i]+1,monthly$mean[i],col='red')
    text(x=monthly$year[i]+0.9,y=monthly$mean[i],pos=4,labels=label[i],col='red',cex=0.5,font=2)
  }
  abline(v=2000:2018,col='grey',lty=3)
  legend('topright',lty=1,col='red',legend = 'Average Apprehensions for fiscal year 20xx',cex = 0.75)
}
```