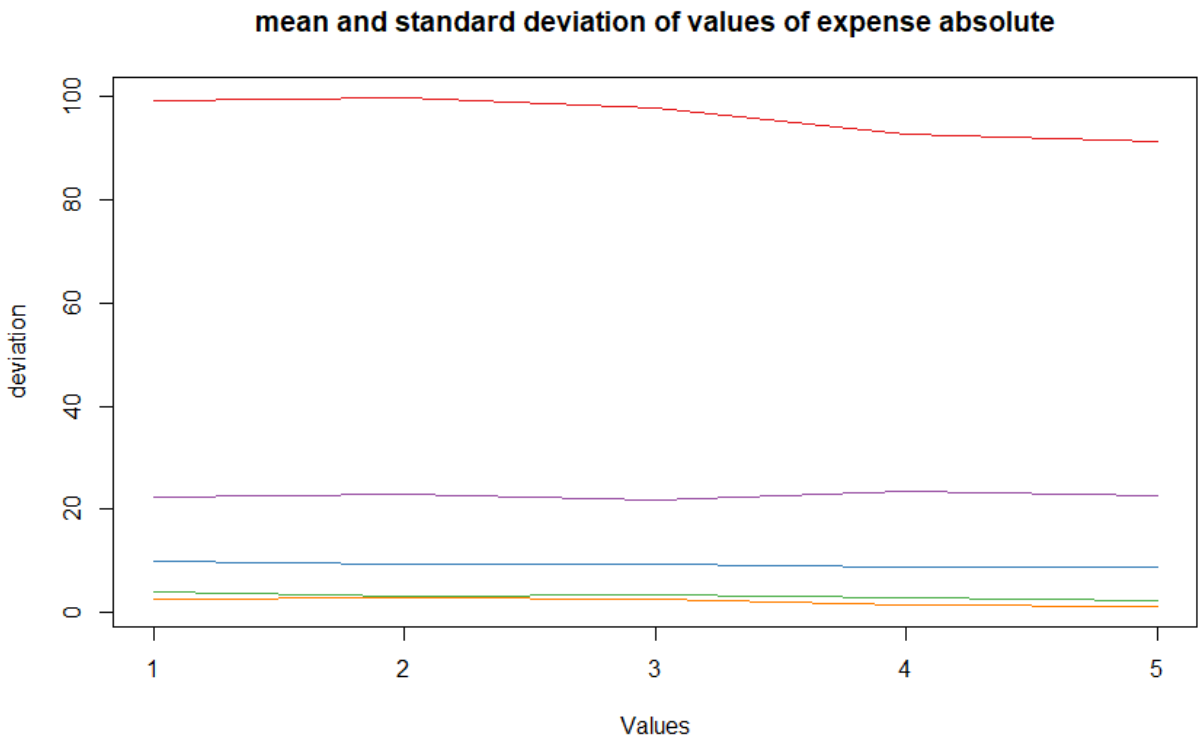
Question 3 (25 points)

For this question, we will use real-world data which came from a report to the Board of Governors in 2014. There are csv files on the learning system you can use that show revenue and expense as absolute and relative values. Using R:

a) **Show two different graphs that make the data look good**

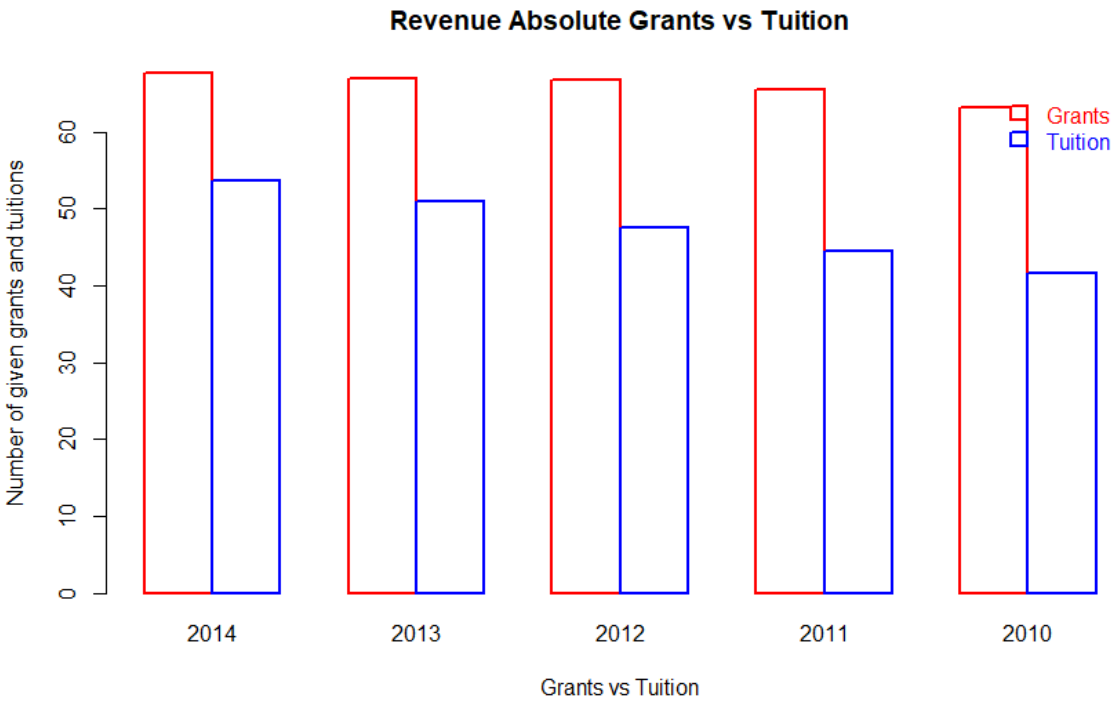
1 - Profile Plot



I think this makes data look good because it shows the proper trend without being clustered. We are able to display multiple attributes. Also, the graph is easy to read and interpret on one glance that means it makes a good use of preattentive features keeping in mind the short term memory limitation.

(Legend for some reason was not showing on R, my source code does include it. Source code : https://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/src/multivariateanalysis.html)

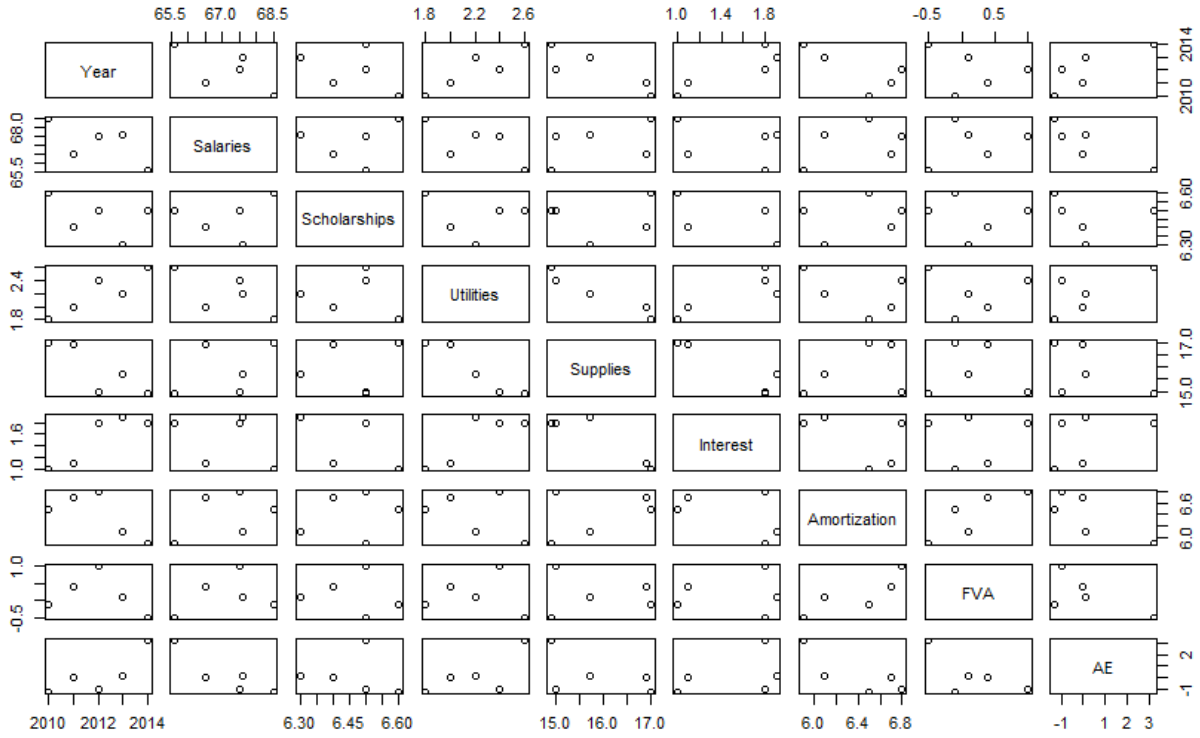
2 – *Double Column Chart*



I think this makes into the good graph group because clearly can be used for comparison of values. Also, there is a good use of colors, preattentive features and Miller's number 7 also has been kept in mind. It can take and display multiple attributes while not getting clustered and non-informative.

b) Show two different graphs that make the data look bad

1 – Scatter Plot



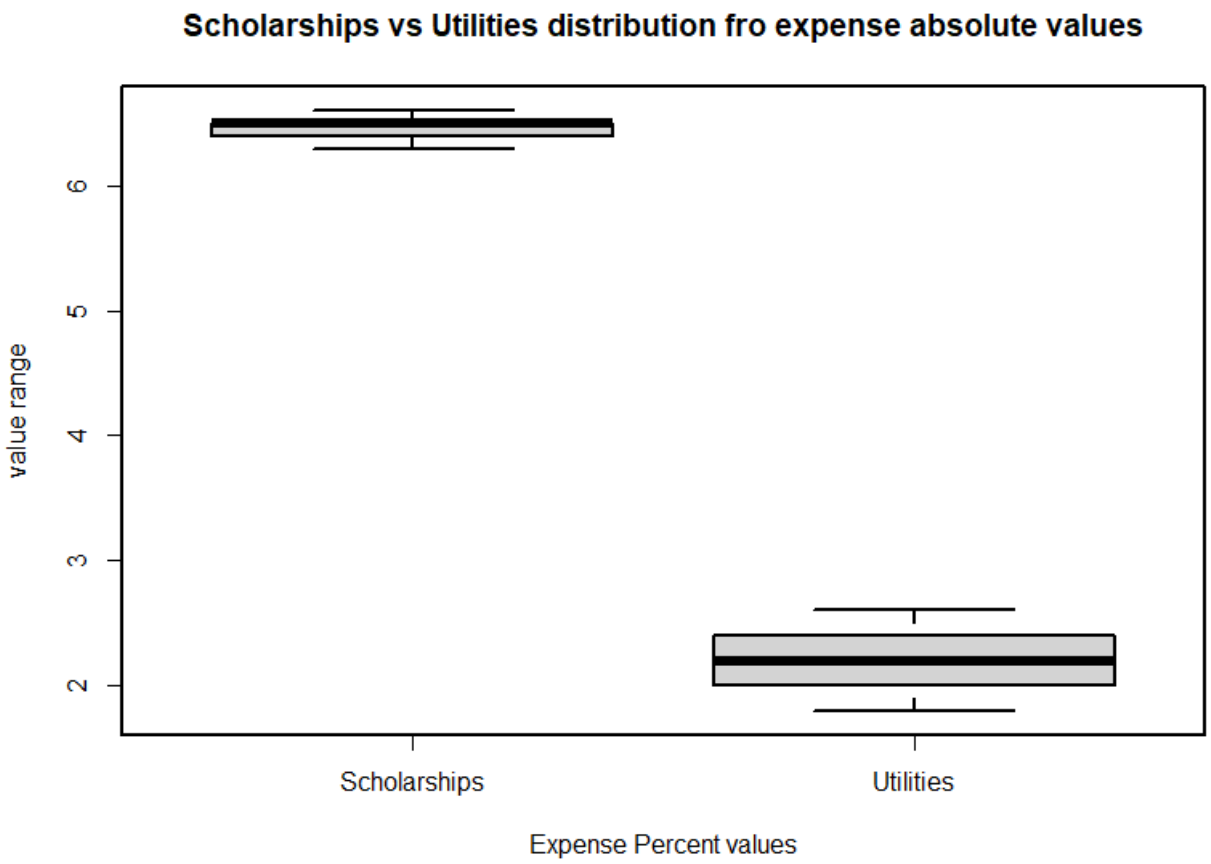
This is clearly a bad graph. No information can be taken from this graph unless very long amount of time is given. No use of colors are also another downfall.

Maybe some colors might have been a little helpful, but complete black and white graph is useless. So, this is a bad graph. The graph is very clustered.

Though scatter plots can be used for multiple attributes, but these can not be used for presentations or making a point.

This graph does not keep in mind the use of pre-attentive features. Also, the limitations of short-term memory has not been considered.

2- Boxplot



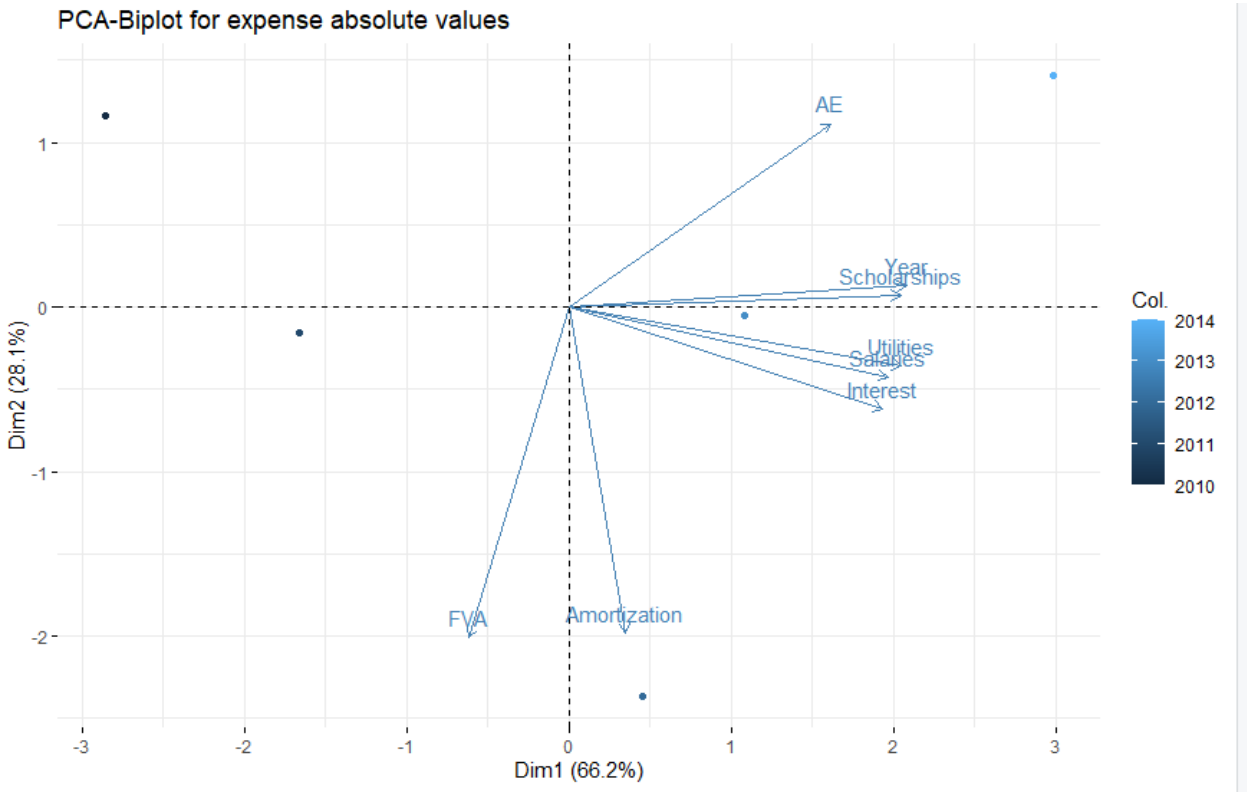
There is no doubt that this graph is a bad graph. The use of scale on y-axis is not proper, making one boxplot compressed a lot.

We can not extract a lot of information from this graph except that the range of Utilities is between 2 and 3 while for scholarships it is more than 6.

Usually box plots are a good choice of graph when we consider mean and median. But if scale is not used properly, they will be highly uninformative.

c) Show two different graphs that show things are pretty much the same over those 5 years

1 – Principle Component Analysis Biplot : PCA Biplot



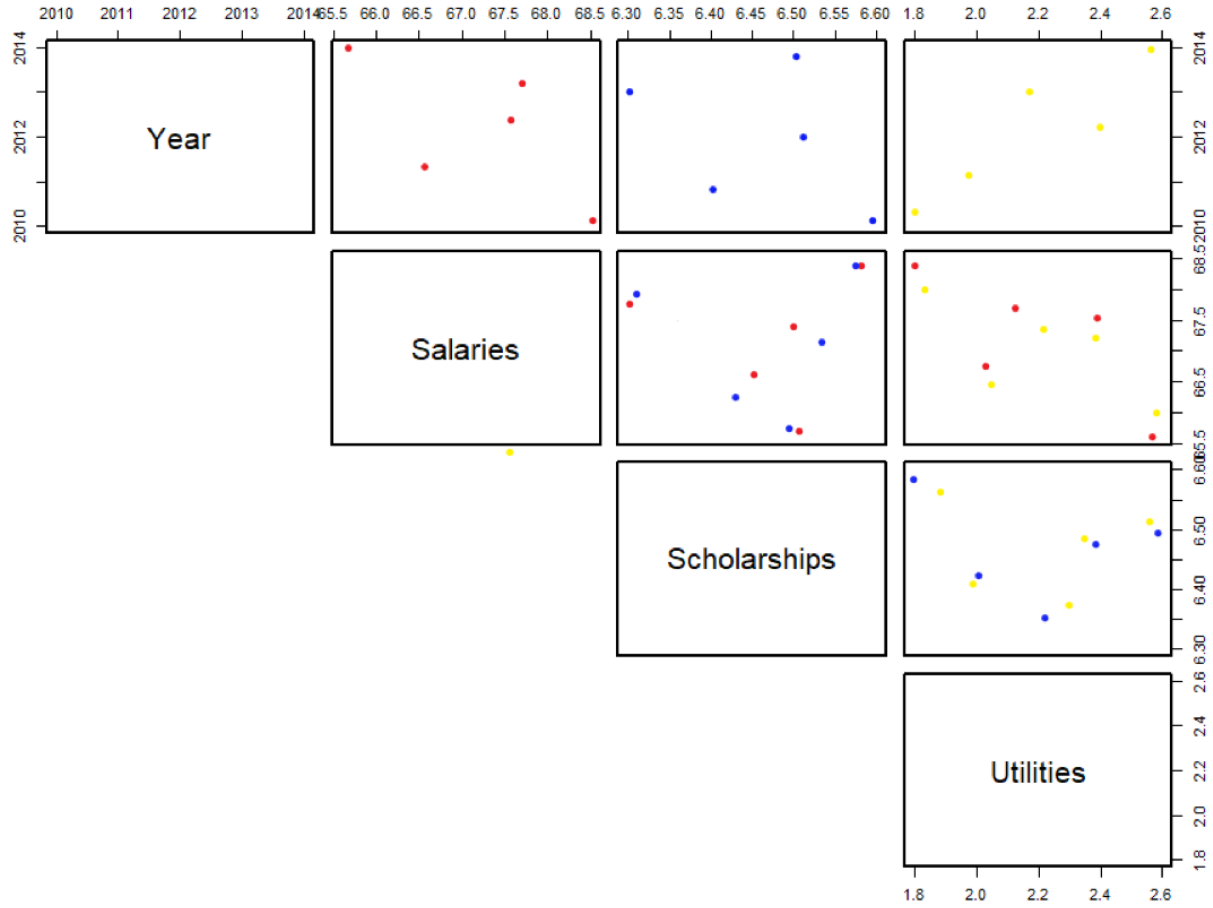
This graph shows that the variation of values over five years is same.

A PCA Biplot is a good choice to show trend over a certain period of time or certain range of changes. It is informative.

It is not clustered and can be used for multiple variables. However, too many variables or a large amount of data might make it crowded and useless.

We can use different colors as well if we wish to show various attributes for variable.

2 – Pairs



This is another graph that helps us see that the variation has been pretty much same. This graph helps comparison of two variables at a time while also displaying a lot of variables

It is a part of scatter plot, but since there have been proper use of colors and data, it is not crowded and hence informative.

R source code attached below in this same document/pdf.

R source code :

```
expabs <- read.csv("C://Users//punya//Documents//Priyam//2020 FALL//COIS-
3510H//assn2//ExpenseAbsolute.csv")
expper <- read.csv("C://Users//punya//Documents//Priyam//2020 FALL//COIS-
3510H//assn2//ExpensePercent.csv")
revabs <- read.csv("C://Users//punya//Documents//Priyam//2020 FALL//COIS-
3510H//assn2//RevenueAbsolute.csv")
revper <- read.csv("C://Users//punya//Documents//Priyam//2020 FALL//COIS-
3510H//assn2//RevenuePercent.csv")
head(expabs)
head(expper)
head(revabs)
head(revper)

##########################################

## a) 2 graphs to make data look good #######

##########################################

### 1 - Profile plot

library(RColorBrewer)
names <- c("Salary","Scholarships","Utilities","Supplies","Interest")
mylist <- list(expabs$Salaries,expabs$Scholarships,expabs$Utilities,expabs$Supplies,expabs$Interest)
makeProfilePlot(mylist,names)
makeProfilePlot <- function(mylist,names)
{
 require(RColorBrewer)
 # find out how many variables we want to include
 numvariables <- length(mylist)
 # choose 'numvariables' random colours
 colours <- brewer.pal(numvariables,"Set1")
 # find out the minimum and maximum values of the variables:
 mymin <- 1e+20
 mymax <- 1e-20
 for (i in 1:numvariables)
 {
  deviation <- mylist[[i]]
  mini <- min(deviation)
  maxi <- max(deviation)
  if (mini < mymin) { mymin <- mini }
  if (maxi > mymax) { mymax <- maxi }
 }
 # plot the variables
 for (i in 1:numvariables)  {

  deviation <- mylist[[i]]
  namei <- names[i]
```

```
   colouri <- colours[i]
   if (i == 1) { plot(main="mean and standard deviation of values of expense absolute",
                  deviation,col=colouri,type="l",ylim=c(mymin,mymax),xlab = "Values") }
   else        { points(deviation, col=colouri,type="l")                              }
   lastxval <- length(deviation)
   lastyval <- deviation[length(deviation)]
   text((lastxval-10),(lastyval),namei,col="black",cex=0.6)
  }
}

############################

### 2 - Double column charts

cols <- c('red','blue');
ylim <- c(0,max(revabs[c('Grants','Tuition')]));
par(lwd=2);
barplot(
  t(revabs[c('Grants','Tuition')]),
  beside=T,
  ylim=ylim,
  border=cols,
  col='white',
  names.arg=expabs$Year,
  main = "Revenue Absolute Grants vs Tuition",
  xlab='Grants vs Tuition',
  ylab='Number of given grants and tuitions',
  legend.text=c('Grants','Tuition'),
  args.legend=list(text.col=cols,col=cols,border=cols,bty='n')
)

#########################################

## b) 2 graphs to make data look bad #######

#########################################

### 1 – Plot
plot(expper)

#########################################

### 2 - boxPlot

boxplot(expper$Scholarships,expper$Utilities,names = c("Scholarships","Utilities"),
      main="Scholarships vs Utilities distribution fro expense absolute values", xlab="Expense Percent
values", ylab="value range")

#########################################

## c) 2 graphs to make data look same #######

#########################################
```

```
### 1 - PCA Biplot
library("factoextra")
my_data <- expabs[, -5]
res.pca <- prcomp(my_data, scale = TRUE)
fviz_pca_biplot(res.pca, col.ind = revabs$Year,
            palette = "jco", geom = "point",,
            title="PCA-Biplot for expense absolute values")

##########################################

### 2 - Pairs

my_cols <- c("yellow", "red", "blue", "green")
pairs(expper[,1:4],
    col = my_cols[expper$Year],
    lower.panel=NULL)
```