## Ajay Kommineni (GR022241)

## Class Lab 2:

* Read your data in R.

data<-anscombe  
head(data)

## x1 x2 x3 x4 y1 y2 y3 y4  
## 1 10 10 10 8 8.04 9.14 7.46 6.58  
## 2 8 8 8 8 6.95 8.14 6.77 5.76  
## 3 13 13 13 8 7.58 8.74 12.74 7.71  
## 4 9 9 9 8 8.81 8.77 7.11 8.84  
## 5 11 11 11 8 8.33 9.26 7.81 8.47  
## 6 14 14 14 8 9.96 8.10 8.84 7.04

#the above command shows the first few lines of anscombe data frame

Summarize the data and comment

summary(data)

## x1 x2 x3 x4 y1   
## Min. : 4.0 Min. : 4.0 Min. : 4.0 Min. : 8 Min. : 4.260   
## 1st Qu.: 6.5 1st Qu.: 6.5 1st Qu.: 6.5 1st Qu.: 8 1st Qu.: 6.315   
## Median : 9.0 Median : 9.0 Median : 9.0 Median : 8 Median : 7.580   
## Mean : 9.0 Mean : 9.0 Mean : 9.0 Mean : 9 Mean : 7.501   
## 3rd Qu.:11.5 3rd Qu.:11.5 3rd Qu.:11.5 3rd Qu.: 8 3rd Qu.: 8.570   
## Max. :14.0 Max. :14.0 Max. :14.0 Max. :19 Max. :10.840   
## y2 y3 y4   
## Min. :3.100 Min. : 5.39 Min. : 5.250   
## 1st Qu.:6.695 1st Qu.: 6.25 1st Qu.: 6.170   
## Median :8.140 Median : 7.11 Median : 7.040   
## Mean :7.501 Mean : 7.50 Mean : 7.501   
## 3rd Qu.:8.950 3rd Qu.: 7.98 3rd Qu.: 8.190   
## Max. :9.260 Max. :12.74 Max. :12.500

# A statistical summary of each variable in the data frame   
#which includes min, median, mean, and max for numerical variables.  
#we can see a pattern in the summary all the means and medians of x are almost same   
#all medians and means of y are almost same too

1 - Mutate x3 to be multiplied by 5 and y3 to add 2 to each value.

#install.packages("dplyr") # Install dplyr if not already installed  
library(dplyr)

##   
## 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

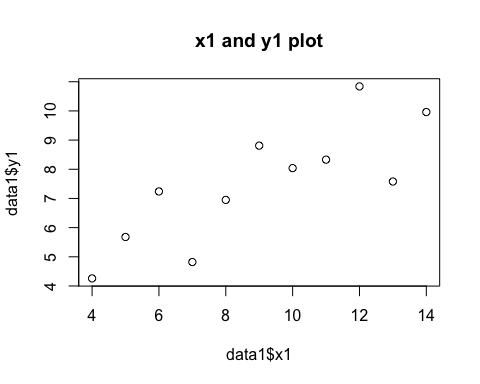
data1 <- mutate(data,  
 x3=x3\*5,  
 y3=y3+2)  
  
#by the mutate we can changes value  
#expression on x3 and y3 changes values of x3 and y3 in the feilds

data1

## x1 x2 x3 x4 y1 y2 y3 y4  
## 1 10 10 50 8 8.04 9.14 9.46 6.58  
## 2 8 8 40 8 6.95 8.14 8.77 5.76  
## 3 13 13 65 8 7.58 8.74 14.74 7.71  
## 4 9 9 45 8 8.81 8.77 9.11 8.84  
## 5 11 11 55 8 8.33 9.26 9.81 8.47  
## 6 14 14 70 8 9.96 8.10 10.84 7.04  
## 7 6 6 30 8 7.24 6.13 8.08 5.25  
## 8 4 4 20 19 4.26 3.10 7.39 12.50  
## 9 12 12 60 8 10.84 9.13 10.15 5.56  
## 10 7 7 35 8 4.82 7.26 8.42 7.91  
## 11 5 5 25 8 5.68 4.74 7.73 6.89

2 - Using the Plot() function, plot x1, y1 and comment

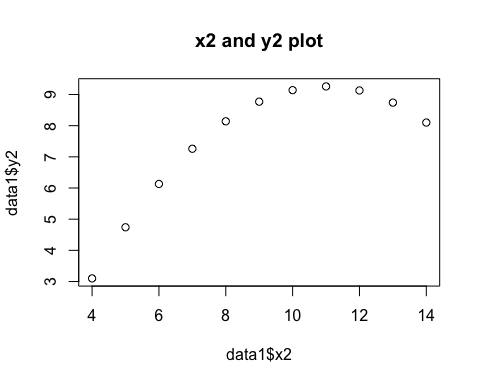
plot(data1$x1,data1$y1, main="x1 and y1 plot")



#below graph shows the values and x and y in a plot  
#seems like values of x1 and y1 and related in such way that increase x1 causes increase in y1  
#It seems like linear trend

3 - Using the Plot() function, plot x2, y2 and comment

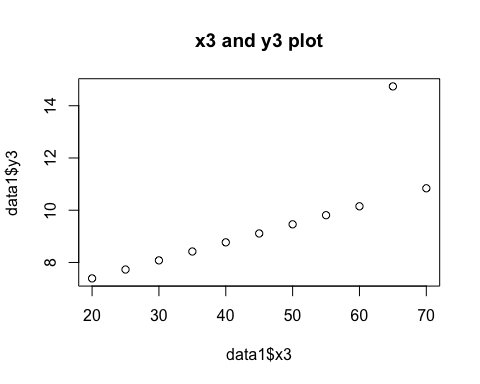
plot(data1$x2,data1$y2, main="x2 and y2 plot")



#below graph shows the values and x2 and y2 in a plot  
#graph shows a curve. It is non linear

4 - Using the Plot() function, plot x3, y3 and comment

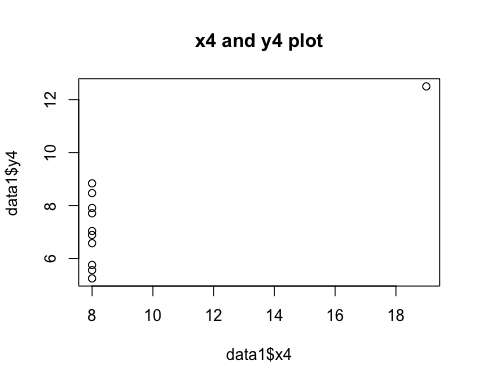
plot(data1$x3,data1$y3, main="x3 and y3 plot")



#below graph shows the values and x3 and y3 in a plot  
#One of the datasets contains a huge value that drastically affects the whole plot   
# if we draw curve we can huge change because of one set values

5 - Using the Plot() function, plot x4, y4 and comment

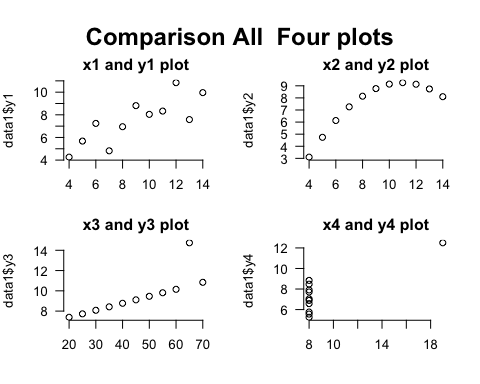
plot(data1$x4,data1$y4, main="x4 and y4 plot")



#below graph shows the values and x3 and y3 in a plot  
#the plot suggest x4 values are 8 except for one, that's why we can all values ploted are in a straight line

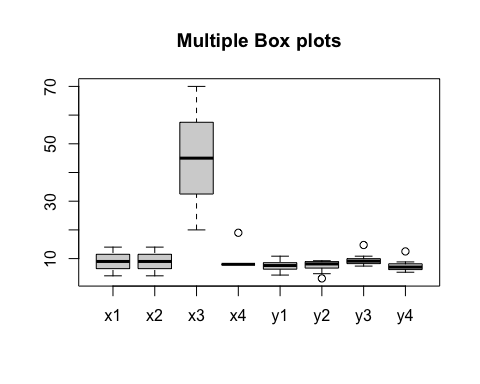
6 - Plot the last four outputs together as a single output, add one title to this output.

# Adjusting outer margins to make space for the title  
par(mfrow=c(2,2), mar=c(3,4,2,2), oma=c(1,0,3,0), las=1, bty="n")   
#oma means outer margin. It sets the outer margin so that we can title clearly  
#mar Adjusts inner margins to prevent crowding  
plot(data1$x1, data1$y1, main="x1 and y1 plot")   
plot(data1$x2, data1$y2, main="x2 and y2 plot")   
plot(data1$x3, data1$y3, main="x3 and y3 plot")   
plot(data1$x4, data1$y4, main="x4 and y4 plot")   
# main title  
mtext("Comparison All Four plots", outer=TRUE, cex=1.5, font=2)



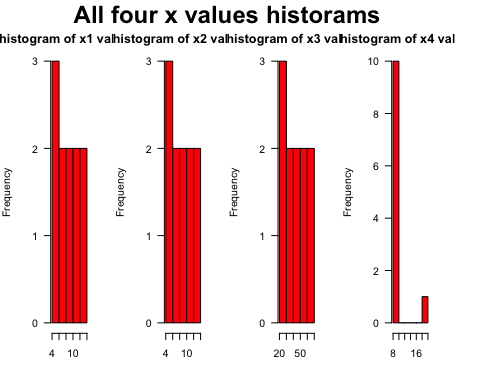
7 - Create a multiple box plot for all the x and y variables. Do this again for a Histogram and Vertical bar plot and comment on your findings.

boxplot(data1[,0:8], main='Multiple Box plots')



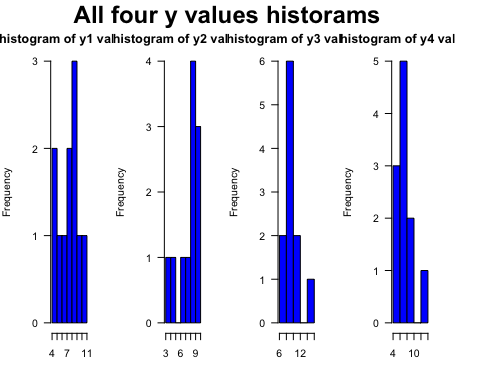
#data1[, 0:8] tries to select columns from index 0 to 8.  
#as we change of x3 values, we can see box plot in a diffrent region  
#all x values have same range of values and medians, this shows there is no values of x those are not of range   
#but y values have some variations, the below plot depicts they are few values of y those making changes in whole range

#The par(mfrow=c(1,4)) function ensures all four histograms appear in a single row.  
#oma It sets the size of the outer margins (space around all plots).  
#las Controls the orientation of axis labels.  
#bty Controls the box around the plot.  
par(mfrow=c(1,4), mar=c(3,4,2,2), oma=c(1,0,2,2), las=1, bty="n")   
hist(data1$x1, main = 'histogram of x1 values ',xlab = 'x values .', col='red')  
hist(data1$x2, main = 'histogram of x2 values ',xlab = 'x values.', col='red')  
hist(data1$x3, main = 'histogram of x3 values ',xlab = 'x values.', col='red')  
hist(data1$x4, main = 'histogram of x4 values ',xlab = 'x values.', col='red')  
mtext("All four x values historams", outer=TRUE, cex=1.5, font=2)



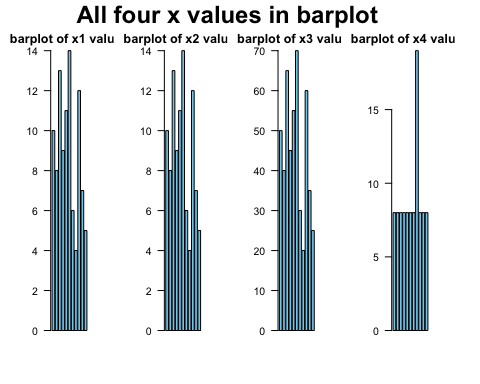
#cex for size of the text  
#font is style of the text  
  
#first three histograms appear similar because the x-values in data are nearly same across datasets.  
#only histogram of x4 there is difference beacause of a improper set of values compared to all x values

par(mfrow=c(1,4), mar=c(3,4,2,2), oma=c(1,0,2,2), las=1, bty="n")   
hist(data1$y1, main = 'histogram of y1 values ',xlab = 'y values', col='blue')  
hist(data1$y2, main = 'histogram of y2 values ',xlab = 'y values.', col='blue')  
hist(data1$y3, main = 'histogram of y3 values ',xlab = 'y values.', col='blue')  
hist(data1$y4, main = 'histogram of y4 values ',xlab = 'y values.', col='blue')  
mtext("All four y values historams", outer=TRUE, cex=1.5, font=2)



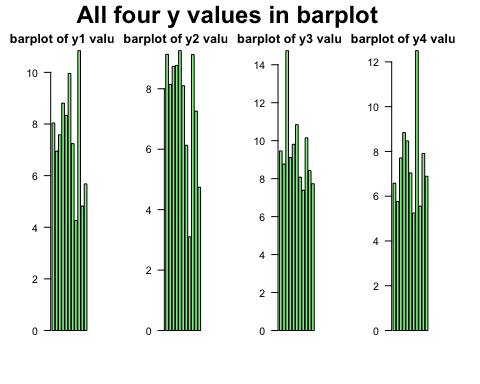
# variation is seen in the y-values compared to x-values.  
#y1 histogram semms better with a set of values , while y2 appears different but it shows increase in values of y   
#y3 and y4 likely contains an outlier, which impacts its distribution.

# Vertical bar plot  
# bar plots represent individual bars of data rather than frequency distributions.  
par(mfrow=c(1,4), mar=c(3,4,2,2), oma=c(1,0,2,2), las=1, bty="n")   
barplot(data1$x1,main='barplot of x1 values', xlab='x1 values',col='skyblue',horiz=FALSE)  
barplot(data1$x2,main='barplot of x2 values', xlab='x2 values',col='skyblue',horiz=FALSE)  
barplot(data1$x3,main='barplot of x3 values', xlab='x3 values',col='skyblue',horiz=FALSE)  
barplot(data1$x4,main='barplot of x4 values', xlab='x4 values',col='skyblue',horiz=FALSE)  
mtext("All four x values in barplot", outer=TRUE, cex=1.5, font=2)



#same like other plots x values are identical in first three data sets, even through the x3 is multiplied by 3 the graph are plots are identical  
#but in x4 all the values are similiar except one value so we can see a huge bar in the bar plot

par(mfrow=c(1,4), mar=c(3,4,2,2), oma=c(1,0,2,2), las=1, bty="n")   
barplot(data1$y1,main='barplot of y1 values', xlab='y1 values',col='lightgreen',horiz=FALSE)  
barplot(data1$y2,main='barplot of y2 values', xlab='y2 values',col='lightgreen',horiz=FALSE)  
barplot(data1$y3,main='barplot of y3 values', xlab='y3 values',col='lightgreen',horiz=FALSE)  
barplot(data1$y4,main='barplot of y4 values', xlab='y4 values',col='lightgreen',horiz=FALSE)  
mtext("All four y values in barplot", outer=TRUE, cex=1.5, font=2)



#y1 is uniformly spread so barplot seems good  
#bar plot also shows a increase in values of y2 same as like the histogram   
#in y3 and y4 all values are in same range except oneliner value which make a bar plot having a high at certain value.

Significance of Anscombe’s Quartet

* The Anscombe’s Quartet is a collection of four datasets that have nearly identical statistical properties (mean, variance, correlation, and regression line) but look very different when plotted.
* This dataset is widely used in statistics and data science education to emphasize the importance of visualizing data before drawing conclusions.

Why Do We Include It in Our Studies?

* The quartet serves as a powerful reminder that visualizing data is crucial for understanding its underlying behaviour of data. Graphs are essential for understanding data.
* The datasets have the same mean, variance, and correlation, yet they represent different patterns. This shows that depending o the numerical data can be misleading *This underscores the idea that visual representations can reveal insights that numerical data alone may obscure.* When plotted, each dataset reveals a unique structure:

1. One follows a linear pattern .
2. One shows a curve .
3. One has an outliner affecting the whole plots and histograms .
4. One has identical x-values but a different distribution.

* This proves that visualization helps in detecting patterns, trends, and anomalies. \*One of the datasets contains a single outlier that drastically affects the regression line. This highlights how outliers can distort statistical conclusions.

Incorporating Anscombe’s Quartet into studies emphasizes the critical role of data visualization in statistical analysis. It serves as a cautionary tale against the over-reliance on summary statistics and encourages a more comprehensive approach to data interpretation. By understanding the significance of the quartet, students and researchers can develop a deeper appreciation for the complexities of data analysis and the importance of visual exploration.