Assignment 1: Exploratory Data Analysis with 'R'

Sreejith Ajithkumar Nair

2025-01-25

Part 1

Section 1

You are working for the Ministry of Education Ontario. The following statement is made by your manager. Secondary student (High School) educational results are worse than they were 10 years ago. Use MS CoPilot to try to transform this in to a question that can be answered with data analytics. Include the prompt and the response in your answer.

Answer:

My prompt to MS CoPilot:

I work for Ministry of Education Ontario. The following statement is made by my manager.

"Secondary student (High School) educational results are worse than they were 10 years ago."

Use above statement to transform it in to a question that can be answered with data analytics.

MS CoPilot Response:

How have the educational results of secondary (high school) students in Ontario changed over the past 10 years, based on standardized test scores and graduation rates?

Now, without using MS CoPilot, based on the examples and discussion in Week 1, transform the original statement in to a question that can be answered with data analytics. Make sure you discuss the logic and reasoning you use to transform it and what questions you might ask.

Answer:

My Response:

1. What are the changes in graduation rates in secondary schools in Ontario over the past decade?

2. What are the trends in average grades by subject in secondary schools in Ontario over the past 10 years?
3. How have curriculum changes impacted educational results over the past decade?
4. What was the impact of the COVID-19 pandemic on secondary school educational results in Ontario?
5. What are the trends in secondary schools education results in different regions of Ontario? Which regions had the most significant decline in results over the past decade?
6. Are there considerable differences in education results in public and private schools in Ontario?
7. Which interventions could help reduce the student drop out rates in secondary schools in Ontario?
8. What would be the impact of this continuing trend over the next decade?
9. How can changes in teacher-student ratio affect this trend in the upcoming years?
10. What strategies can help to increase student engagement in secondary schools?

Discuss how your answer is different from the Gen AI answer and why you think your answer is better.

Answer:

MS Copilot generated only one question in response to the given prompt. The question tried to explore certain aspects of the declining trend, i.e., standardized test scores and graduation rates and focused only on descriptive analytics, giving a limited view of the issue. In contrast, the questions I asked combine descriptive, exploratory, prescriptive, and predictive analytics. My response tries to uncover the underlying cause of the decline by focusing on key educational metrics, subject-specific performance trends, geographical trends, policy changes and recent events. In addition to this, it includes questions about potential solutions and strategies to improve education, as well as planning for the future. This multidimensional approach takes a deeper dive into the issue while providing avenues for proactively addressing it, which is helpful for strategic planning and decision-making.

Section 2

Consider the following three arrays of data. Each array is data for one web site. The numbers in the array represent the number of unique visitors to each site in a day (for example, Site A had 203 visitors on the first day, 213 on the second and so on).

Site A: (203 213 213 217 221 223 223 176 213 193 162)

Site B: (202 227 198 198 223 210 199 201 214 235 218)

Site C: (117 117 110 129 101 148 129 101 115 92 108)

```
#defining data for each site as vector
Site_A_SN <- c(203, 213, 213, 217, 221, 223, 223, 176, 213, 193, 162)
Site_B_SN <- c(202, 227, 198, 198, 223, 210, 199, 201, 214, 235, 218)
Site_C_SN <- c(117, 117, 110, 129, 101, 148, 129, 101, 115, 92, 108)
```

Based on the data provided, and using the skills learned in this class, answer the following questions. Make sure to provide evidence for your answers

a) Which Site has the most visits on a typical day?

```
#Calculate mean for each Site
mean_A_SN <- mean(Site_A_SN)
mean_B_SN <- mean(Site_B_SN)
mean_C_SN <- mean(Site_C_SN)

print(paste("Mean of the Site A:", mean_A_SN)) # Source: (GeeksforGeeks,
Printing output of an R program)

## [1] "Mean of the Site A: 205.181818181818"

print(paste("Mean of the Site B:", mean_B_SN))

## [1] "Mean of the Site B: 211.363636363636"

print(paste("Mean of the Site C:", mean_C_SN))

## [1] "Mean of the Site C: 115.181818181818"
```

Answer: Based on the calculated mean, Site B has the most visits on a typical day.

b)Which Site has the least consistent usage?

```
#Calculate standard deviation for each Site
sd_A_SN <- sd(Site_A_SN)
sd_B_SN <- sd(Site_B_SN)
sd_C_SN <- sd(Site_C_SN)

#Calculate coefficient of variation (CV) for each site
cv_A_SN <- round(sd_A_SN/mean_A_SN,3)
```

```r
cv_B_SN <- round(sd_B_SN/mean_B_SN,3)
cv_C_SN <- round(sd_C_SN/mean_C_SN,3)

print(paste("CV of the Site A:", cv_A_SN))
```

```
## [1] "CV of the Site A: 0.098"
```

```r
print(paste("CV of the Site B:", cv_B_SN))
```

```
## [1] "CV of the Site B: 0.062"
```

```r
print(paste("CV of the Site C:", cv_C_SN))
```

```
## [1] "CV of the Site C: 0.137"
```

Answer: Site C has the highest coefficient of variation, meaning it has the most variation relative to its mean and has the least consistent usage.

Part 2

The dataset is a summary of information about School Boards in Ontario which have been gathered by the provincial goverment. The following tasks will seek to describe and explore some of this data.

Section 1 : Basic Manipulation

1. Read in the text file and change to a data frame.

2. Append your initials to all variables in the data frame.

```r
data_SN <- read.csv("PROG8435-25W-Assign-01.txt", header=TRUE, sep=",")
data_SN <- as.data.frame(data_SN)

#display first few rows
head(data_SN)
```

```
##                            Name Language           Type       Region
## 1                   Algoma DSB  English         Public North Region
## 2 Algonquin and Lakeshore CDSB  English Roman Catholic  East Region
## 3            Avon Maitland DSB  English         Public  West Region
## 4                Bluewater DSB  English         Public  West Region
## 5 Brant Haldimand Norfolk CDSB  English Roman Catholic  West Region
## 6              Bruce-Grey CDSB  English Roman Catholic  West Region
##              City G6_EQAO G10_OSSLT G10_Cr_Acc G10_Cr_Acc_P G11_Cr_Acc
## 1 Sault Ste Marie    0.78      0.72       0.69        -0.02       0.71
## 2         Napanee    0.78      0.86       0.84         0.00       0.87
## 3        Seaforth    0.84      0.81       0.81         0.03       0.78
## 4         Chesley    0.76      0.75       0.67        -0.01       0.71
## 5       Brantford    0.85      0.85       0.73        -0.08       0.81
```

```
## 6         Hanover    0.80      0.82       0.82         0.01       0.85
##   G11_Cr_Acc_P G4_Grad_Rate G4_Grad_Rate_P G5_Grad_Rate G5_Grad_Rate_P
## 1        -0.08        0.719          0.002        0.768         -0.026
## 2        -0.03        0.895          0.015        0.909         -0.016
## 3        -0.01        0.802         -0.008        0.838         -0.022
## 4        -0.01        0.715          0.004        0.815         -0.007
## 5        -0.01        0.818          0.021        0.901          0.046
## 6         0.05        0.874          0.006        0.927          0.040
```

3. Change the appropriate character variables to factor variables and make sure all other variables are the proper format.

```
data_SN$Name <- as.factor(data_SN$Name)
data_SN$Language <- as.factor(data_SN$Language)
data_SN$Type <- as.factor(data_SN$Type)
data_SN$Region <- as.factor(data_SN$Region)
data_SN$City <- as.factor(data_SN$City)
data_SN$G6_EQAO <- as.numeric(data_SN$G6_EQAO)

## Warning: NAs introduced by coercion
```

4. Determine if there are any missing values. If there are, delete the row(s) which contains them and specify which row(s) you deleted.

```
# Summarize the dataset and provide statistics for each column
summary(data_SN)

##                              Name       Language               Type
##  Algoma DSB                    : 1   English:59   Public         :36
##  Algonquin and Lakeshore CDSB: 1   French :12   Roman Catholic:35
##  Avon Maitland DSB             : 1
##  Bluewater DSB                 : 1
##  Brant Haldimand Norfolk CDSB: 1
##  Bruce-Grey CDSB               : 1
##  (Other)                       :65
##             Region             City       G6_EQAO         G10_OSSLT
##  Central Region: 9   North Bay  : 4   Min.   :0.7300   Min.   :0.3800
##  East Region   :17   Sudbury    : 4   1st Qu.:0.7900   1st Qu.:0.7900
##  North Region  :20   Thunder Bay: 3   Median :0.8400   Median :0.8400
##  Toronto Region: 4   Timmins    : 3   Mean   :0.8446   Mean   :0.8254
##  West Region   :21   Toronto    : 3   3rd Qu.:0.8875   3rd Qu.:0.8650
##                      Windsor    : 3   Max.   :0.9900   Max.   :0.9600
##                      (Other)    :51   NA's   :1
##    G10_Cr_Acc      G10_Cr_Acc_P        G11_Cr_Acc      G11_Cr_Acc_P
##  Min.   :0.4900   Min.   :-0.15000   Min.   :0.5800   Min.   :-0.15000
##  1st Qu.:0.7300   1st Qu.:-0.04000   1st Qu.:0.7600   1st Qu.:-0.03000
##  Median :0.7950   Median :-0.01000   Median :0.8300   Median :-0.01000
##  Mean   :0.7789   Mean   :-0.01606   Mean   :0.8144   Mean   :-0.01586
##  3rd Qu.:0.8475   3rd Qu.: 0.01000   3rd Qu.:0.8700   3rd Qu.: 0.00750
##  Max.   :0.9300   Max.   : 0.30000   Max.   :0.9500   Max.   : 0.06000
##  NA's   :1                           NA's   :1        NA's   :1
```

```
##   G4_Grad_Rate    G4_Grad_Rate_P       G5_Grad_Rate    G5_Grad_Rate_P
##  Min.   :0.5880   Min.   :-0.081000   Min.   :0.6860   Min.   :-0.058000
##  1st Qu.:0.7740   1st Qu.:-0.015500   1st Qu.:0.8395   1st Qu.:-0.007500
##  Median :0.8310   Median : 0.002000   Median :0.8950   Median : 0.005000
##  Mean   :0.8233   Mean   :-0.000662   Mean   :0.8786   Mean   : 0.004324
##  3rd Qu.:0.8905   3rd Qu.: 0.015000   3rd Qu.:0.9303   3rd Qu.: 0.011500
##  Max.   :0.9630   Max.   : 0.066000   Max.   :0.9740   Max.   : 0.059000
##                                       NA's   :1
```

```
#sum of missing values in dataset
print(paste("Number of missing values:",sum(is.na(data_SN))))
```

```
## [1] "Number of missing values: 5"
```

```
#row containing missing value
data_SN[rowSums(is.na(data_SN))>0,]
```

```
##                   Name Language   Type      Region      City G6_EQAO
G10_OSSLT
## 53 Renfrew County ESB  English Public East Region  Pembroke      NA
0.38
##    G10_Cr_Acc G10_Cr_Acc_P G11_Cr_Acc G11_Cr_Acc_P G4_Grad_Rate
G4_Grad_Rate_P
## 53         NA          0.3         NA           NA         0.87
0.004
##    G5_Grad_Rate G5_Grad_Rate_P
## 53           NA          0.019
```

```
#delete the missing value
cleaned_data_SN <-na.omit(data_SN)
```

```
#checking for missing values
print(paste("Number of missing values:",sum(is.na(cleaned_data_SN))))
```

```
## [1] "Number of missing values: 0"
```

5. What are the dimensions of the dataset (rows and columns)?

```
rows_SN <- sum(rowSums(!is.na(cleaned_data_SN))>0)
col_SN <- sum(colSums(!is.na(cleaned_data_SN))>0)

print(paste("Number of Rows:",rows_SN))
```

```
## [1] "Number of Rows: 70"
```

```
print(paste("Number of Columns:",col_SN))
```

```
## [1] "Number of Columns: 15"
```

Section 2 : Summarizing Data

1. Means and Standard Deviations

a. Calculate the mean and standard deviation for the four year graduation rate.

```
G4_mean_SN <- round(mean(cleaned_data_SN$G4_Grad_Rate),3)
G4_sd_SN <- round(sd(cleaned_data_SN$G4_Grad_Rate),3)

print(paste("Mean for the four year graduation rate:",G4_mean_SN))

## [1] "Mean for the four year graduation rate: 0.823"

print(paste("Standard Deviation for the four year graduation
rate:",G4_sd_SN))

## [1] "Standard Deviation for the four year graduation rate: 0.083"
```

b. Use the results above to calculate the coefficient of variation

```
G4_cv_SN <- round(G4_sd_SN/G4_mean_SN,3)
print(paste("Coefficient of variation for the four year graduation
rate:",G4_cv_SN))

## [1] "Coefficient of variation for the four year graduation rate: 0.101"
```

c. Calculate the mean and standard deviation for the five year graduation rate.

```
G5_mean_SN <- round(mean(cleaned_data_SN$G5_Grad_Rate),3)
G5_sd_SN <- round(sd(cleaned_data_SN$G5_Grad_Rate),3)

print(paste("Mean for the five year graduation rate:",G5_mean_SN))

## [1] "Mean for the five year graduation rate: 0.879"

print(paste("Standard Deviation for the five year graduation
rate:",G5_sd_SN))

## [1] "Standard Deviation for the five year graduation rate: 0.066"
```

d. Calculate the coefficient of variation.

```
G5_cv_SN <- round(G5_sd_SN/G5_mean_SN,3)
print(paste("Coefficient of variation for the five year graduation
rate:",G5_cv_SN))

## [1] "Coefficient of variation for the five year graduation rate: 0.075"
```

e. Does the four or five year graduation rate have more variation?

Based on the coefficient of variation calculated above, **four year graduation rate** has more variation.

2. Calculate the 11th percentile of the number of grade 11 credit accumulation.

```
G11_SN <- quantile(cleaned_data_SN$G11_Cr_Acc,0.11)
print(paste("11th percentile of the number of Grade 11 credit
accumulation:",G11_SN))

## [1] "11th percentile of the number of Grade 11 credit accumulation: 0.71"
```

Section 3: Organizing Data

1. Summary Table

a.Create a table showing the Grade 6 EQAO score by region. This should be rounded to three decimal places.

```
table_G6_Region_SN <- table(cleaned_data_SN$Region,cleaned_data_SN$G6_EQAO)
table_G6_Region_SN <- round(prop.table(table_G6_Region_SN),3)
#display the table Grade 6 EQAO score by region
table_G6_Region_SN

##
##                   0.73  0.74  0.75  0.76  0.77  0.78  0.79   0.8  0.81
0.83
##   Central Region 0.000 0.000 0.000 0.000 0.000 0.000 0.014 0.000 0.014
0.029
##   East Region    0.000 0.000 0.000 0.000 0.014 0.043 0.014 0.014 0.014
0.000
##   North Region   0.014 0.014 0.014 0.057 0.014 0.014 0.000 0.029 0.000
0.000
##   Toronto Region 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
0.014
##   West Region    0.000 0.000 0.014 0.014 0.014 0.014 0.014 0.014 0.014
0.000
##
##                   0.84  0.85  0.86  0.87  0.88  0.89   0.9  0.91  0.93
0.96
##   Central Region 0.000 0.000 0.000 0.014 0.029 0.000 0.014 0.014 0.000
0.000
##   East Region    0.014 0.029 0.000 0.014 0.000 0.029 0.000 0.000 0.000
0.014
##   North Region   0.043 0.000 0.000 0.000 0.000 0.014 0.000 0.014 0.014
0.000
##   Toronto Region 0.014 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
0.000
##   West Region    0.057 0.071 0.014 0.000 0.014 0.000 0.014 0.014 0.000
0.000
##
##                   0.97  0.98  0.99
##   Central Region 0.000 0.000 0.000
##   East Region    0.014 0.014 0.000
```

```
##   North Region   0.029 0.014 0.000
##   Toronto Region 0.000 0.014 0.014
##   West Region    0.014 0.000 0.000
```

b.Which region has, on average, the highest EQAO scores?

```
table_region_SN <- aggregate(cleaned_data_SN$G6_EQAO,
by=list(cleaned_data_SN$Region), FUN=mean, na.rm=TRUE)
colnames(table_region_SN) <- c("Region", "Mean_G6_EQAO")

table_region_SN$Mean_G6_EQAO <- round(table_region_SN$Mean_G6_EQAO,3)

#display table Average EQAO scores by Region
table_region_SN
```

```
##           Region Mean_G6_EQAO
## 1 Central Region        0.856
## 2    East Region        0.851
## 3   North Region        0.829
## 4 Toronto Region        0.910
## 5    West Region        0.838
```

From the table **Toronto Region** has the highest EQAO scores.

2. Cross Tabulation

a.Create a table counting language by region.

```
table_language_region_SN <- table(cleaned_data_SN$Language,
cleaned_data_SN$Region)

#display table counting language by region
table_language_region_SN
```

```
##
##           Central Region East Region North Region Toronto Region West
Region
##   English              9          13           14              2
20
##   French               0           3            6              2
1
```

b.Change the table to show the language percentage across regions. That is, how different languages are distributed across all regions. This should be rounded to three decimal places.

```
table_language_region_per_SN <- round(prop.table(table_language_region_SN,1)
*100, 3)
table_language_region_per_SN
```

```
##
##           Central Region East Region North Region Toronto Region West
Region
```

```
##   English         15.517      22.414       24.138          3.448
34.483
##   French           0.000      25.000       50.000         16.667
8.333
```

c.Which region has the most French language Boards (percentage and number)?

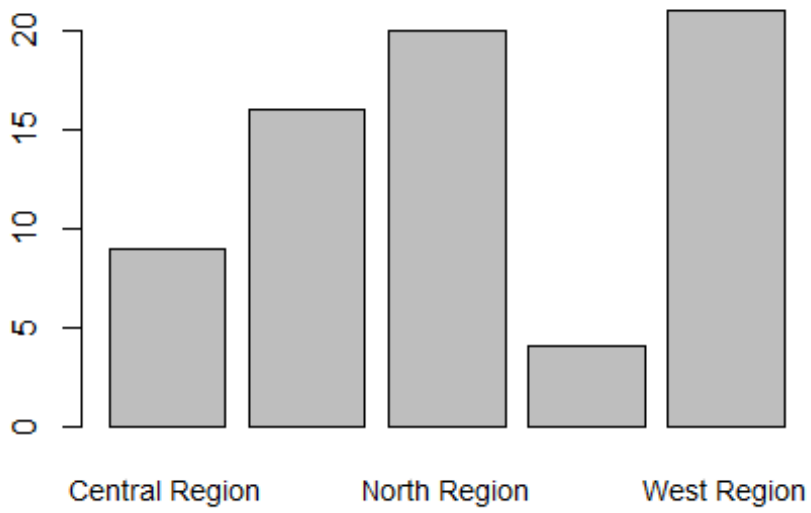
```
table_language_region_SN

##
##           Central Region East Region North Region Toronto Region West
Region
##   English              9          13           14              2
20
##   French               0           3            6              2
1

table_language_region_per_SN

##
##           Central Region East Region North Region Toronto Region West
Region
##   English         15.517      22.414       24.138          3.448
34.483
##   French           0.000      25.000       50.000         16.667
8.333
```

From the two tables above, **North Region has the highest number(6) and percentage(50%) of French boards**

3. Bar Plot

a. Create a column plot of number of public boards by region.

```
public_boards_SN <- colSums(table_language_region_SN)
barplot(public_boards_SN, cex.names = 0.9 )
```




```
#cex.names to adjust the size of labels in x-axis
```

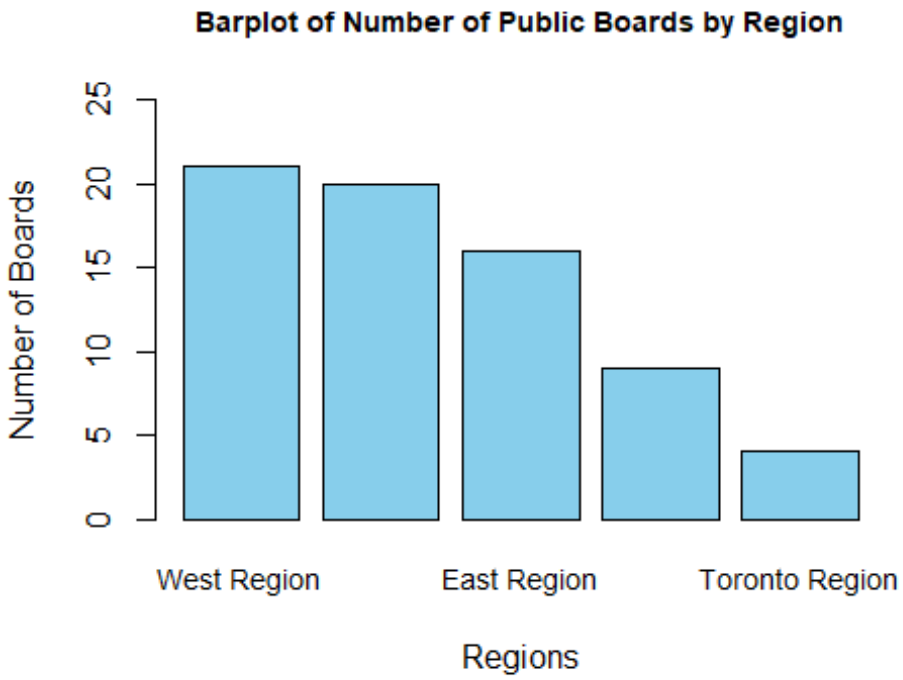
b. The plot should be:

i. Rank ordered by highest to lowest count of public boards.

ii. Properly labeled (title, x-axis, etc)

iii. The bars should have a different colour than the one shown in class.

```
public_boards_SN <- public_boards_SN[order(public_boards_SN, decreasing =
TRUE)]
barplot(public_boards_SN,
        main = "Barplot of Number of Public Boards by Region", cex.main =
0.9,
        xlab = 'Regions', ylab='Number of Boards',
        cex.names = 0.9, col = 'skyblue' , ylim = c(0,25))
```




```
#cex.names reduce the size of labels in x-axis
```

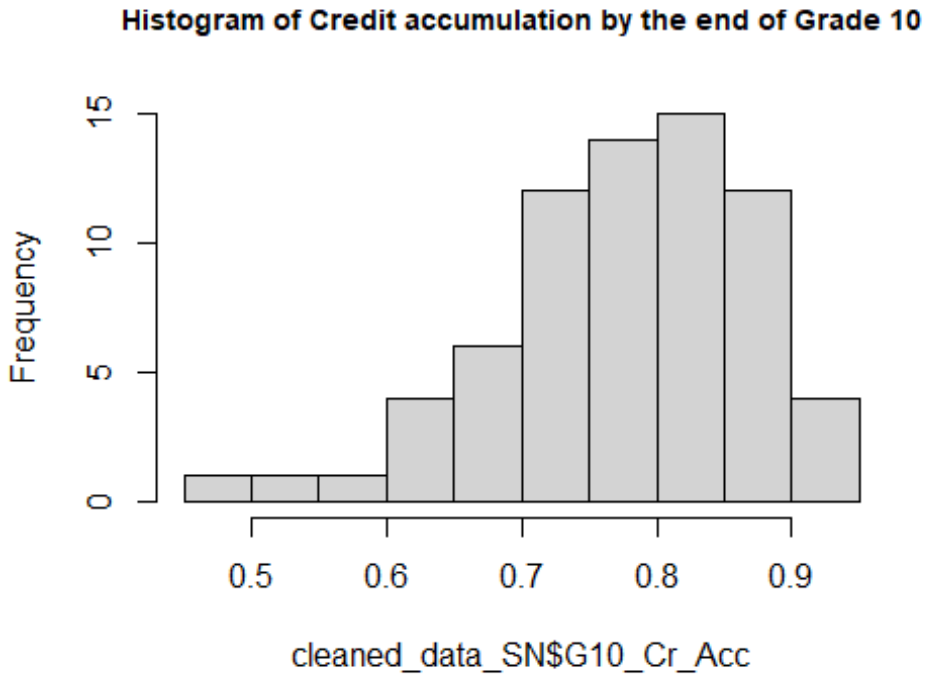
c.Based on the bar plot, (approximately) how many of public boards are in the Eastern region?

Based on the bar plot, there are approximately **16** public boards in the Eastern Region.

4. Histogram

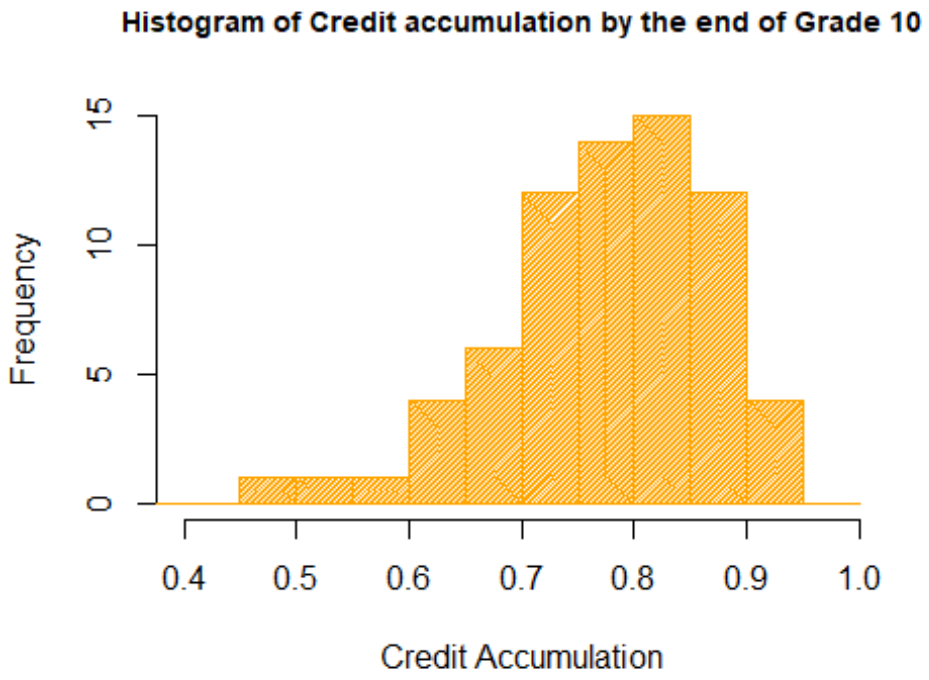
a. Create a histogram of credit accumulation to the end of grade 10.

```
hist(cleaned_data_SN$G10_Cr_Acc, main = "Histogram of Credit accumulation by
the end of Grade 10",cex.main = 0.9)
```



b. The plot should be properly labeled and a unique colour and have bins .05 wide (i.e. 0 to 0.05, 0.05 to 0.10, etc.)

```
hist_SN <-hist(cleaned_data_SN$G10_Cr_Acc, main = "Histogram of Credit
accumulation by the end of Grade 10", cex.main = 0.9,
     xlab = 'Credit Accumulation',
     breaks = seq(0,1,0.05),
     col = 'orange', density = 90, angle = 45,
     xlim = c(0.4,1))
```



c. Which range of credit accumulation to end of grade 10 is the most common?

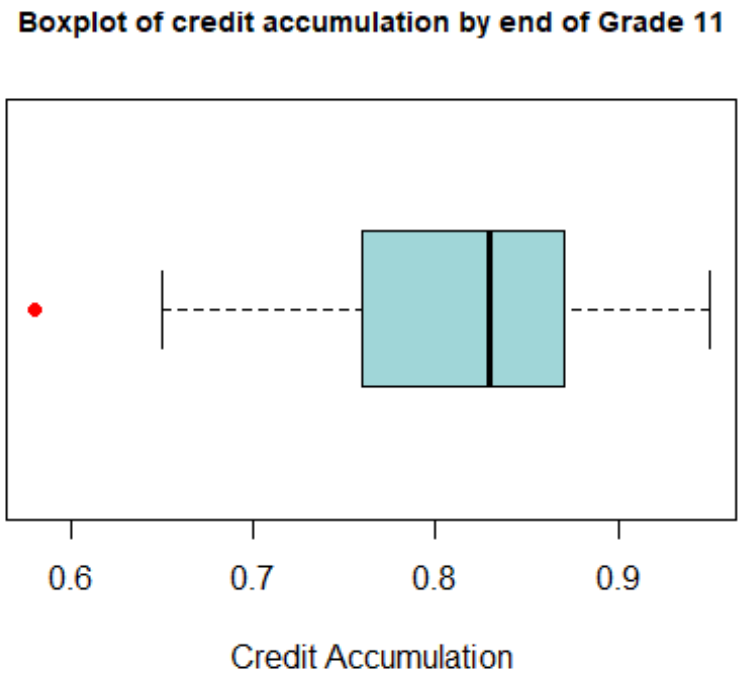
The peak in the histogram represents the most common credit accumulation by end of Grade 10 which lies in the range of 0.8-0.85 in this case.

5. Box plot

a. Create a horizontal box plot of credit accumulation to end of grade 11.

b. The plot should be properly labeled and a unique colour.

```
boxplot(cleaned_data_SN$G11_Cr_Acc,
        main = "Boxplot of credit accumulation by end of Grade 11", cex.main
= 0.9,
        horizontal = TRUE, col ='#A0D6D8', xlab='Credit Accumulation', pch
=16,
        outcol = "red" )
```




```
#outcol changes color of outliers. Source :(TutorialsPoint, How to change the color of outliers)
```

c. Based on the box plot, approximately how many Boards have credit accumulation to the end of grade 11 of more than ~ 0.83? Is there anything else noteable about the chart.

```
median( cleaned_data_SN$G11_Cr_Acc)
```

```
## [1] 0.83
```

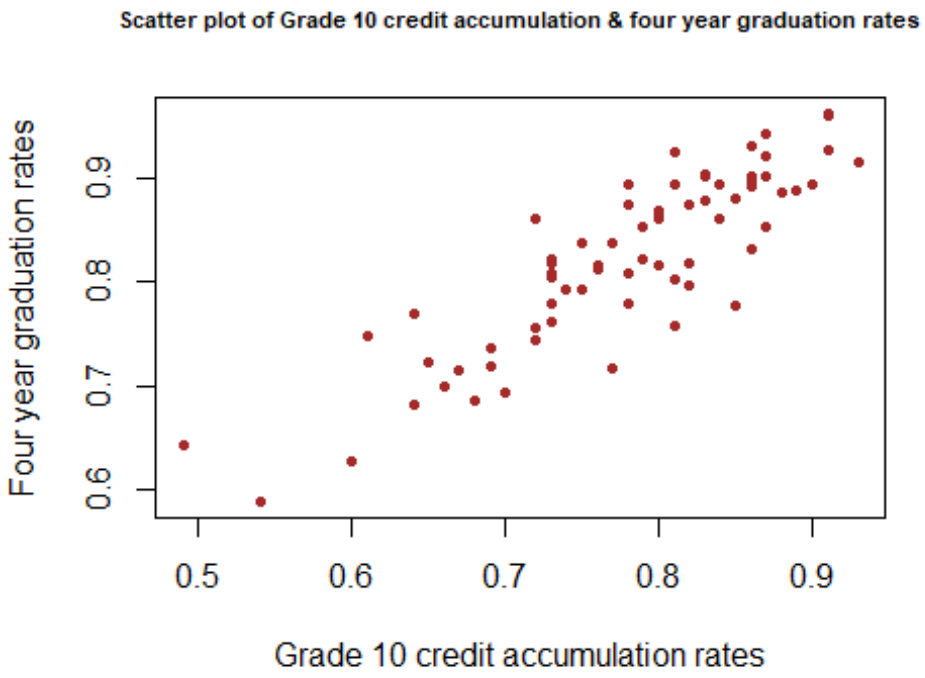
Median is 0.83, which indicates that 50 % of total Boards have credit accumulation of more than 0.83 by the end of Grade 11. The box plot is left skewed, i.e., most of the value are larger than the mean of the data. The outlier lies on the lower end of the data.

6. Scatter Plot

a. Create a scatter plot comparing Grade 10 credit accumulation rates (horizontal axis) with four year graduation rates.

b. The plot should be properly labeled with a marker type different than the one demonstrated in class.

```
plot(cleaned_data_SN$G10_Cr_Acc, cleaned_data_SN$G4_Grad_Rate, main =
"Scatter plot of Grade 10 credit accumulation & four year graduation rates",
cex.main = 0.7,
     xlab= "Grade 10 credit accumulation rates",
     ylab = 'Four year graduation rates',
     col = 'brown',
     pch = 20)
```



c. Does there appear to be an association between Grade 10 credit accumulation rates and four year graduation rates? Diuscuss.

From the Scatter plot, it appears that there is a positive correlation between Grade 10 accumulation rates and four year graduation rates. It is visible that as Grade 10 accumulation rates increase, four year graduation rates also increase. The plot indicates a linear upward trend.

References

1. GeeksforGeeks. Printing output of an R program. GeeksforGeeks. Retrieved from https://www.geeksforgeeks.org/printing-output-of-an-r-program/

2. TutorialsPoint. How to change the color of outliers in base R boxplot. TutorialsPoint. Retrieved from https://www.tutorialspoint.com/how-to-change-the-color-of-outliers-in-base-r-boxplot