Government of Ontario, School Board Achievements and Progress

Beena Kurian

2025-02-14

Load packages

```
# Load package pastecs
if(!require(pastecs)){install.packages("pastecs")}

## Loading required package: pastecs

library("pastecs")
```

Read Data and print Data

```
# Read data set
data  <- read.csv("Ontario_School_Board_Performance.txt", header=TRUE)

# Convert data frame
data  <- as.data.frame(data )

# printing head
head(data )

##     Code                          Name Language           Type       Region
## 1 B28010                   Algoma DSB  English         Public North Region
## 2 B67202 Algonquin and Lakeshore CDSB  English Roman Catholic  East Region
## 3 B66010            Avon Maitland DSB  English         Public  West Region
## 4 B66001                Bluewater DSB  English         Public  West Region
## 5 B67164 Brant Haldimand Norfolk CDSB  English Roman Catholic  West Region
## 6 B67008              Bruce-Grey CDSB  English Roman Catholic  West Region
##              City G6_EQAO G6_EQAO_S G6_EQAO_P G10_OSSLT G10_Credit_Acc
## 1 Sault Ste Marie    0.78      0.32        NA      0.72           0.69
## 2         Napanee    0.78      0.32        NA      0.86           0.84
## 3        Seaforth    0.84      0.52        NA      0.81           0.81
## 4         Chesley    0.76      0.26        NA      0.75           0.67
## 5       Brantford    0.85      0.55        NA      0.85           0.73
## 6         Hanover    0.80      0.39        NA      0.82           0.82
##   G10_Credit_Acc_P G11_Credit_Acc G11_Credit_Acc_P Y4_Grd_Rt Y4_Grd_Rt_P
## 1            -0.02           0.71            -0.08     0.719       0.002
## 2             0.00           0.87            -0.03     0.895       0.015
## 3             0.03           0.78            -0.01     0.802      -0.008
## 4            -0.01           0.71            -0.01     0.715       0.004
## 5            -0.08           0.81            -0.01     0.818       0.021
## 6             0.01           0.85             0.05     0.874       0.006
```

```
##   Y5_Grd_Rt Y5_Grd_Rt_P Enrollment   Funding
## 1     0.768      -0.026      10690 175553485
## 2     0.909      -0.016      11995 181577452
## 3     0.838      -0.022      15530 277403512
## 4     0.815      -0.007      18300 277562637
## 5     0.901       0.046      11775 157646981
## 6     0.927        0.04       4850 100621374
```

Data Transformation and Preparation

Initial Transformation

Transform variables as appropriate (e.g. selected character to factor, numeric if needed, etc.)

```
str(data )
```

```
## 'data.frame':    72 obs. of  20 variables:
##  $ Code            : chr  "B28010" "B67202" "B66010" "B66001" ...
##  $ Name            : chr  "Algoma DSB" "Algonquin and Lakeshore CDSB"
"Avon Maitland DSB" "Bluewater DSB" ...
##  $ Language        : chr  "English" "English" "English" "English" ...
##  $ Type            : chr  "Public" "Roman Catholic" "Public" "Public" ...
##  $ Region          : chr  "North Region" "East Region" "West Region" "West
Region" ...
##  $ City            : chr  "Sault Ste Marie" "Napanee" "Seaforth" "Chesley"
...
##  $ G6_EQAO         : num  0.78 0.78 0.84 0.76 0.85 0.8 0.85 0.97 0.97 0.98
...
##  $ G6_EQAO_S       : num  0.32 0.32 0.52 0.26 0.55 0.39 0.55 0.94 0.94
0.97 ...
##  $ G6_EQAO_P       : logi  NA NA NA NA NA NA ...
##  $ G10_OSSLT       : num  0.72 0.86 0.81 0.75 0.85 0.82 0.84 0.93 0.85
0.95 ...
##  $ G10_Credit_Acc  : num  0.69 0.84 0.81 0.67 0.73 0.82 0.8 0.78 0.91 0.87
...
##  $ G10_Credit_Acc_P: num  -0.02 0 0.03 -0.01 -0.08 0.01 -0.09 -0.01 -0.04
0.02 ...
##  $ G11_Credit_Acc  : num  0.71 0.87 0.78 0.71 0.81 0.85 0.84 0.84 0.92
0.91 ...
##  $ G11_Credit_Acc_P: num  -0.08 -0.03 -0.01 -0.01 -0.01 0.05 -0.06 -0.06 -
0.05 -0.02 ...
##  $ Y4_Grd_Rt       : num  0.719 0.895 0.802 0.715 0.818 0.874 0.861 0.895
0.963 0.922 ...
##  $ Y4_Grd_Rt_P     : num  0.002 0.015 -0.008 0.004 0.021 0.006 -0.021 -
0.021 0.003 -0.017 ...
##  $ Y5_Grd_Rt       : chr  "0.768" "0.909" "0.838" "0.815" ...
##  $ Y5_Grd_Rt_P     : chr  "-0.026" "-0.016" "-0.022" "-0.007" ...
##  $ Enrollment      : int  10690 11995 15530 18300 11775 4850 13715 17630
```

```
9270 16095 ...
##  $ Funding          : num  175553485 181577452 277403512 277562637
157646981 ...
```

From the structure , it is clear that, we have 72 observations and 20 variables. Among the 20 variables, 8 variables are character type and one column is logical type.

Let's keep the following columns as character type itself:

- Code
- Name
- City

Change the following character types to factor types:

- Language
- Type
- Region

Change the following type from character to numeric type:

- Y5_Grd_Rt
- Y5_Grd_Rt_P

```
# Character to factor conversion
data $Language  <- as.factor(data $Language )
data $Type  <- as.factor(data $Type )
data $Region  <- as.factor(data $Region )

# Character to numeric conversion
suppressWarnings(data $Y5_Grd_Rt  <- as.numeric(data $Y5_Grd_Rt ))
suppressWarnings(data $Y5_Grd_Rt_P  <- as.numeric(data $Y5_Grd_Rt_P ))
```

After conversion, let's review the structure of the data frame,

```
str(data )
```

```
## 'data.frame':    72 obs. of  20 variables:
##  $ Code             : chr  "B28010" "B67202" "B66010" "B66001" ...
##  $ Name             : chr  "Algoma DSB" "Algonquin and Lakeshore CDSB"
"Avon Maitland DSB" "Bluewater DSB" ...
##  $ Language         : Factor w/ 2 levels "English","French": 1 1 1 1 1 1 1
2 2 2 ...
##  $ Type             : Factor w/ 2 levels "Public","Roman Catholic": 1 2 1 1
2 2 2 1 2 2 ...
##  $ Region           : Factor w/ 5 levels "Central Region",..: 3 2 5 5 5 5 2
```

```
2 5 4 ...
##  $ City            : chr  "Sault Ste Marie" "Napanee" "Seaforth" "Chesley"
...
##  $ G6_EQAO         : num  0.78 0.78 0.84 0.76 0.85 0.8 0.85 0.97 0.97 0.98
...
##  $ G6_EQAO_S       : num  0.32 0.32 0.52 0.26 0.55 0.39 0.55 0.94 0.94
0.97 ...
##  $ G6_EQAO_P       : logi  NA NA NA NA NA NA ...
##  $ G10_OSSLT       : num  0.72 0.86 0.81 0.75 0.85 0.82 0.84 0.93 0.85
0.95 ...
##  $ G10_Credit_Acc  : num  0.69 0.84 0.81 0.67 0.73 0.82 0.8 0.78 0.91 0.87
...
##  $ G10_Credit_Acc_P: num  -0.02 0 0.03 -0.01 -0.08 0.01 -0.09 -0.01 -0.04
0.02 ...
##  $ G11_Credit_Acc  : num  0.71 0.87 0.78 0.71 0.81 0.85 0.84 0.84 0.92
0.91 ...
##  $ G11_Credit_Acc_P: num  -0.08 -0.03 -0.01 -0.01 -0.01 0.05 -0.06 -0.06 -
0.05 -0.02 ...
##  $ Y4_Grd_Rt       : num  0.719 0.895 0.802 0.715 0.818 0.874 0.861 0.895
0.963 0.922 ...
##  $ Y4_Grd_Rt_P     : num  0.002 0.015 -0.008 0.004 0.021 0.006 -0.021 -
0.021 0.003 -0.017 ...
##  $ Y5_Grd_Rt       : num  0.768 0.909 0.838 0.815 0.901 0.927 0.917 0.91
0.958 0.949 ...
##  $ Y5_Grd_Rt_P     : num  -0.026 -0.016 -0.022 -0.007 0.046 0.04 0.003 -
0.013 -0.021 -0.009 ...
##  $ Enrollment      : int  10690 11995 15530 18300 11775 4850 13715 17630
9270 16095 ...
##  $ Funding         : num  175553485 181577452 277403512 277562637
157646981 ...
```

Delete any rows of the dataframe containing more than 1 NA value.

```
# printing rows with more than one 'NA' values
data [rowSums(is.na(data )) > 1, ]
```

```
##      Code                Name Language           Type       Region
City
## 45 B29041      Northwest CDSB  English Roman Catholic North Region Fort
Frances
## 58 B29076 Superior North CDSB  English Roman Catholic North Region
Terrace Bay
##    G6_EQAO G6_EQAO_S G6_EQAO_P G10_OSSLT G10_Credit_Acc G10_Credit_Acc_P
## 45    0.83      0.48        NA        NA             NA               NA
## 58    0.68      0.00        NA        NA             NA               NA
##    G11_Credit_Acc G11_Credit_Acc_P Y4_Grd_Rt Y4_Grd_Rt_P Y5_Grd_Rt
Y5_Grd_Rt_P
## 45             NA               NA        NA          NA        NA
NA
```

```
## 58             NA                NA         NA           NA         NA
NA
##    Enrollment  Funding
## 45       1235 32762757
## 58        665  9379539
```

The rows 45 and 58 have more than one NA values. Let's delete those rows.

```
# remove rows with more than one 'NA' values
data_cleaned  <- data [rowSums(is.na(data )) <= 1, ]

# check dimension after removal
dim(data_cleaned )
```

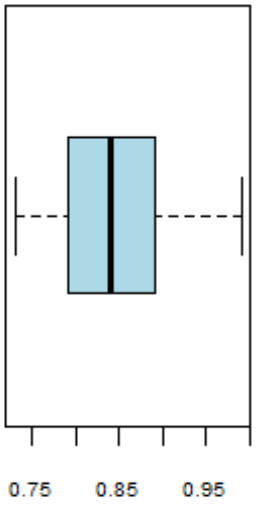
```
## [1] 70 20
```

After deletion, we have 70 rows and 20 columns in the data frame.

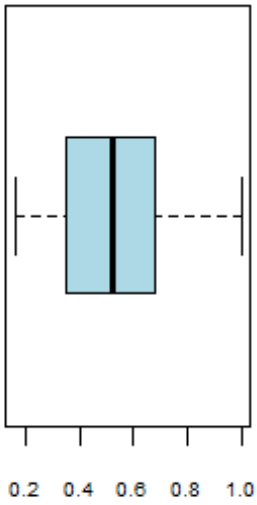
Outliers Removal

```
par(mfrow=c(1,2))
for (i in 1:ncol(data_cleaned ))
{
  if(is.numeric(data_cleaned [,i]))
  {
    boxplot(data_cleaned [i],
    main = names(data_cleaned )[i],
    horizontal = TRUE,col = "lightblue",
    pch = 2, cex.main=0.8, cex.lab=0.8, cex.axis = 0.6)
  }
}
```

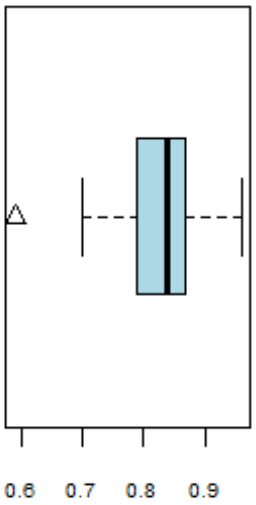
G6_EQAO



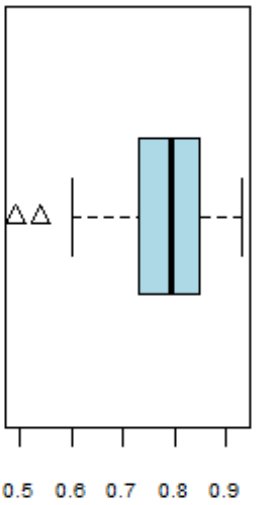
G6_EQAO_S



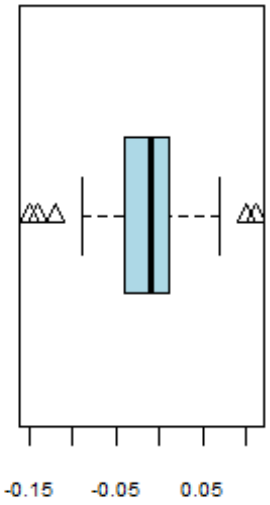
G10_OSSLT



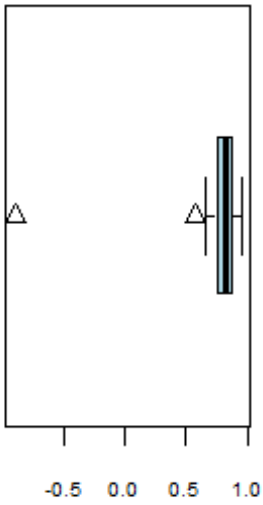
G10_Credit_Acc



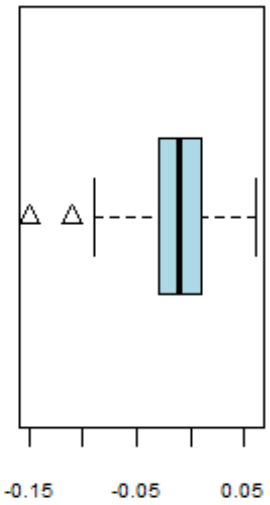
G10_Credit_Acc_P



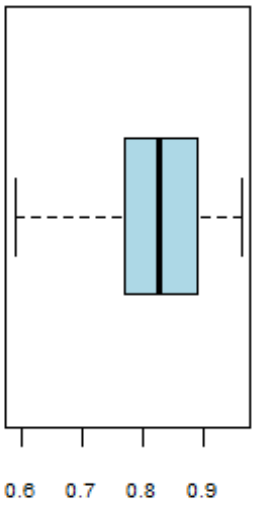
G11_Credit_Acc

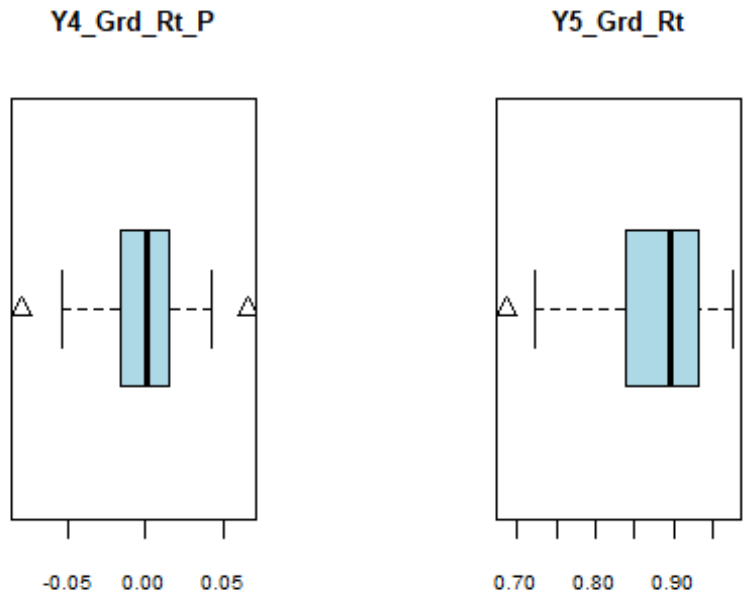


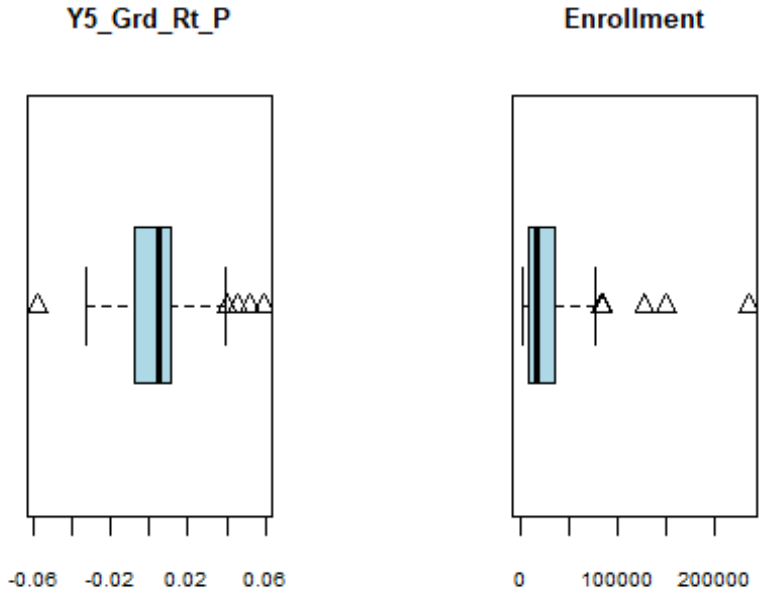
G11_Credit_Acc_P



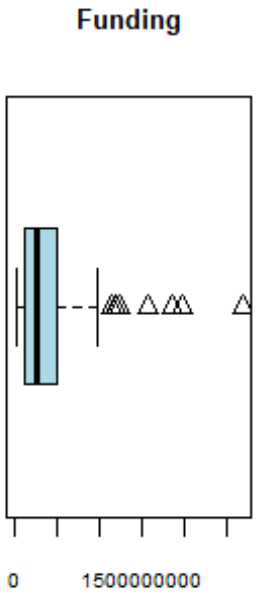
Y4_Grd_Rt








```
par(mfrow = c(1,1))
```



From Box plots, it is clear that,

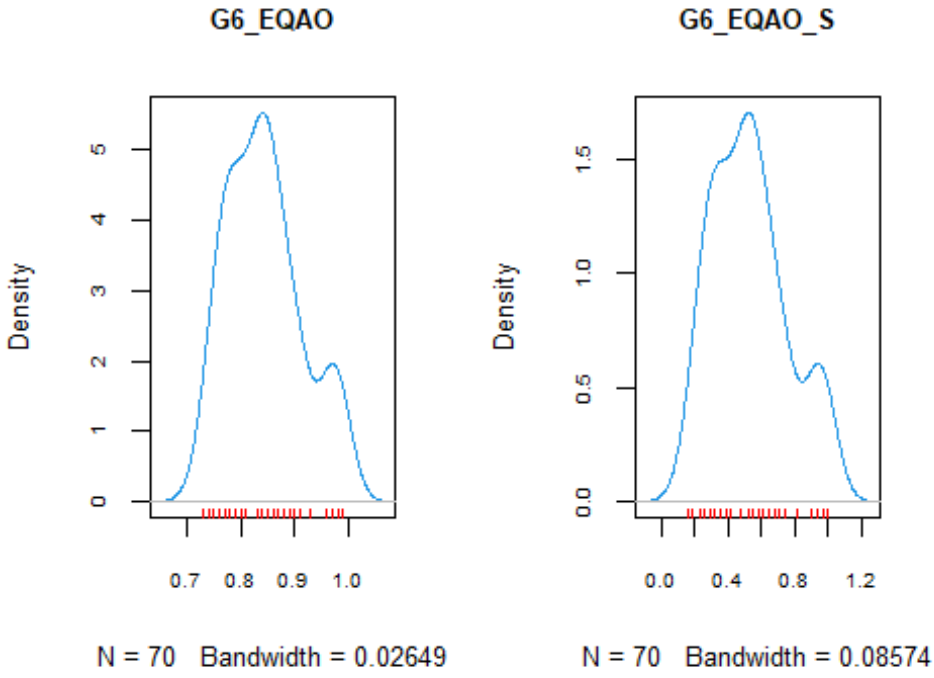
- The box plots for G6_EQAO , G6_EQAO_S , and Y4_Grd_Rt do not contain any outliers.
- All other box plots show the presence of outliers.
- Most outliers, except one in G11_Credit_Acc , can be retained as they are not significantly far from the lower or upper boundaries and could represent valid data points.
- The outlier in G11_Credit_Acc is far from the whiskers, indicating a possible data issue, as it represents negative credit accumulation which is not possible.

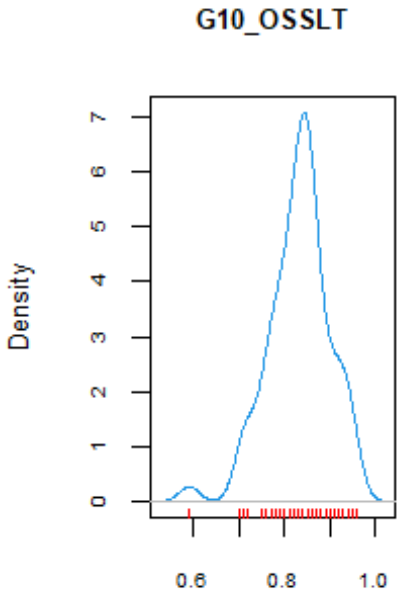
Credit accumulation can't be negative, but progress in credit accumulation can be negative if a student fails, withdraws, or earns fewer credits than expected. Thus, negative outlier in column named G11_Credit_Acc must be handled properly.

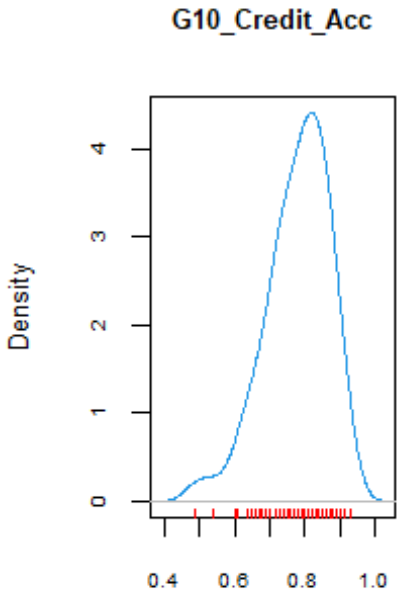
Let's also look at the density plots, to look at anomalous values if any.

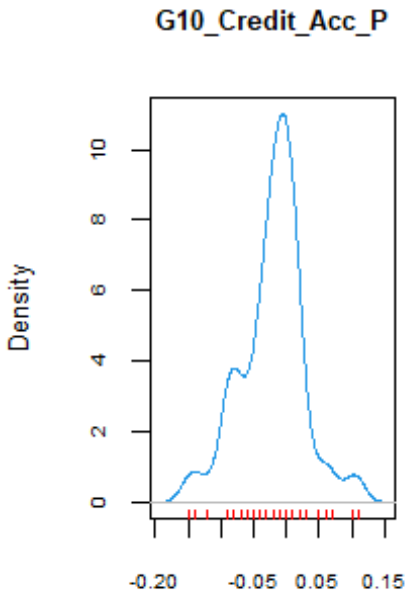
```
par(mfrow=c(1,2))
for (i in 1:ncol(data_cleaned ))
{
  if(is.numeric(data_cleaned [,i]))
  {
    plot(density(data_cleaned [[i]], na.rm = TRUE),
```

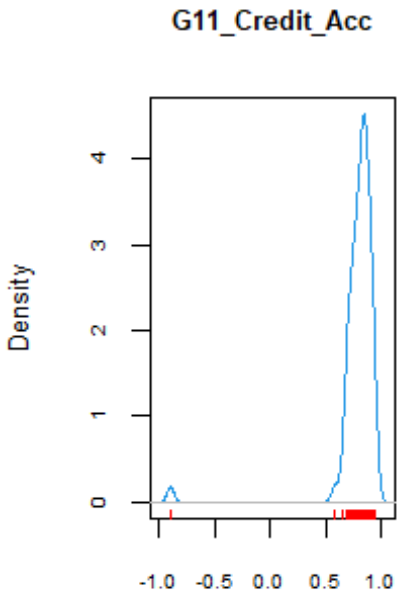
```
    main = names(data_cleaned )[i],
    pch = 10, cex.main=0.8, cex.lab=0.8, cex.axis = 0.6,col = 12)

    rug(data_cleaned [[i]], col = "red")
  }

}
```

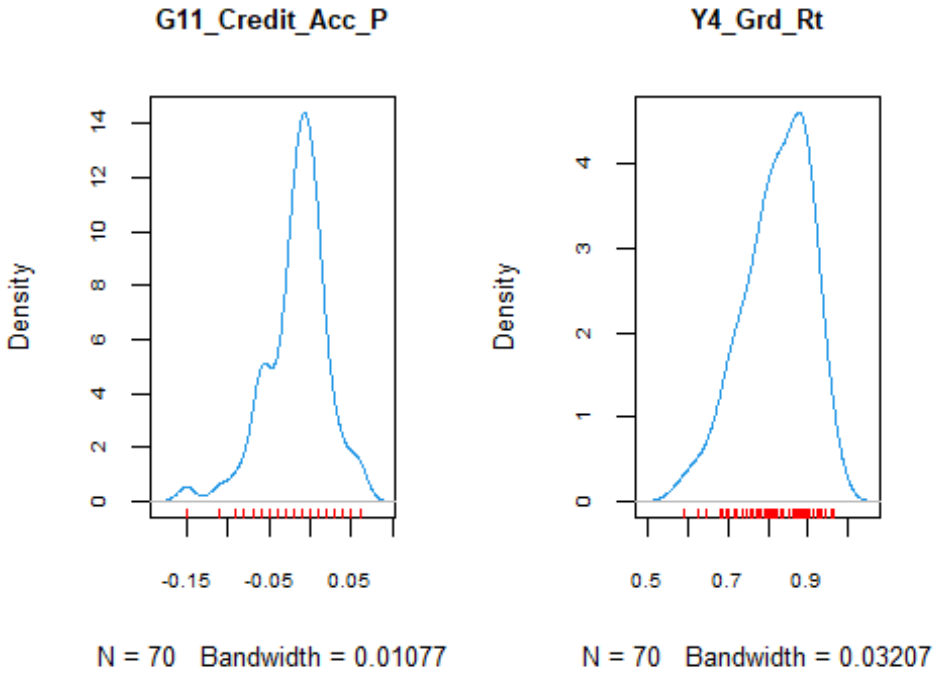


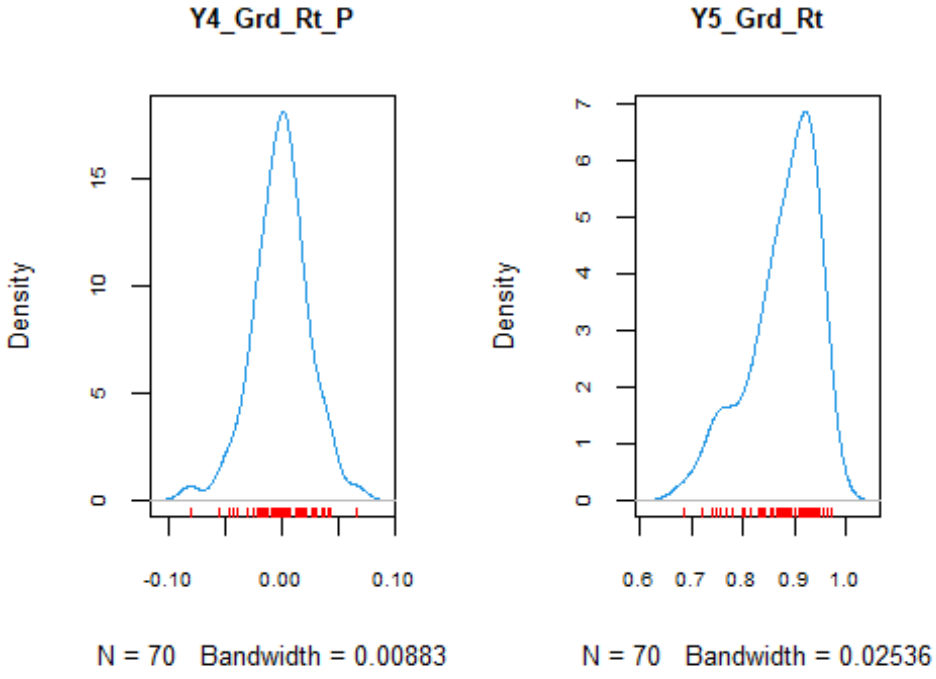

G10_OSSLT
Density
0
1
2
3
4
5
6
7
0.6
0.8
1.0
N = 70 Bandwidth = 0.02154

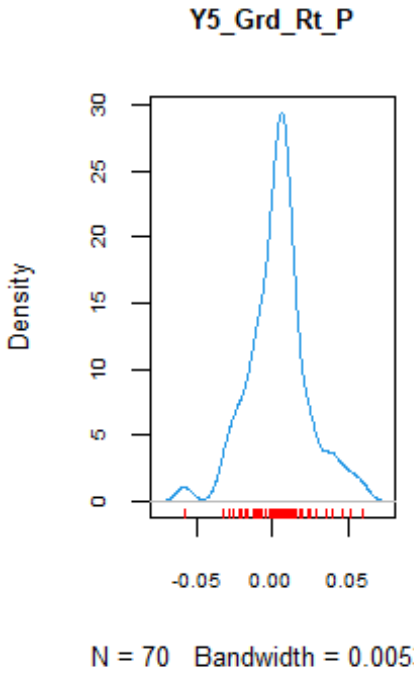

G10_Credit_Acc
Density
0
1
2
3
4
0.4
0.6
0.8
1.0
N = 70 Bandwidth = 0.03374

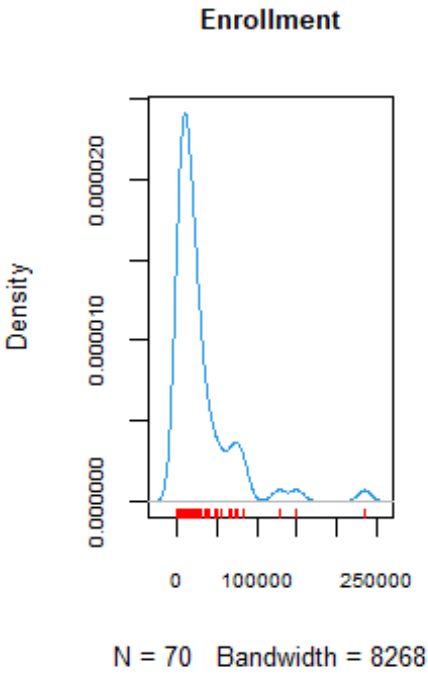

G10_Credit_Acc_P
Density
0
2
4
6
8
10
-0.20
-0.05
0.05
0.15
N = 70 Bandwidth = 0.01436


G11_Credit_Acc
Density
0
1
2
3
4
-1.0
-0.5
0.0
0.5
1.0
N = 70 Bandwidth = 0.03159

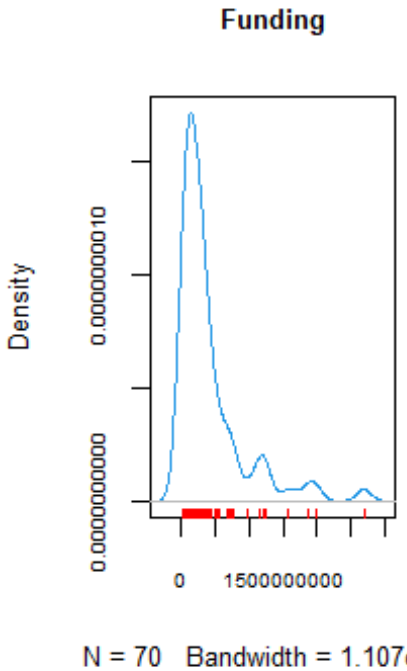

G11_Credit_Acc_P
Density
0
2
4
6
8
10
12
14
-0.15
-0.05
0.05
N = 70 Bandwidth = 0.01077
Y4_Grd_Rt
Density
0
1
2
3
4
0.5
0.7
0.9
N = 70 Bandwidth = 0.03207


Y4_Grd_Rt_P
Density
0
5
10
15
-0.10
0.00
0.10
N = 70 Bandwidth = 0.00883
Y5_Grd_Rt
Density
0
1
2
3
4
5
6
7
0.6
0.7
0.8
0.9
1.0
N = 70 Bandwidth = 0.02536






```
par(mfrow = c(1,1))
```



Analysis of Outliers in Student Credit Related Columns

By evaluating the above box plots and density plots, all the outliers are acceptable, except column named ***"G11_Credit_Acc"***. Based on the Box Plot and Density Plot of G11_Credit_Acc , it is evident that G11_Credit_Acc contains an outlier with a value below zero. Since credit accumulation cannot logically be negative, this value is likely a data entry error. To prevent this error from affecting the analysis, I am removing the row with the min value.

```
# Remove rows where G11_Credit_Acc is equal to its minimum value
data_cleaned <- data_cleaned[data_cleaned$G11_Credit_Acc !=
min(data_cleaned$G11_Credit_Acc, na.rm = TRUE), ]
```

Analysis of Outliers in Enrollment and Funding Columns

Let's check the outliers in Enrollment and Funding columns where the density plots are right skewed. Even though there are outliers in the data, I do not believe they are errors.

According to my observation, certain school types with record-high enrollments have received higher funding from the province. Also, funding allocations vary by region.

From the news article, (Link: w.cbc.ca/news/canada/london/london-area-school-boards-get-more-than-260m-to-build-6-new-schools-1.7443833) London-area school boards get more than $260M to build 6 new schools. Thus more than 260 Million Funding can not be considered as an anomaly in data. It represents actual value. In the NEWS article, they mentioned, due to record-high enrollment, they requested more funds from province to build 6 new schools. So I am keeping this outlier as it is.

The below aggregations is an evidence of my reason to keep outliers as it is.

```
top_3_funding <- data_cleaned [order(-data_cleaned $Funding ), ][1:3, ]
print(top_3_funding )
```

```
##      Code            Name Language   Type         Region        City
G6_EQAO
## 63 B66052     Toronto DSB  English Public Toronto Region     Toronto
0.83
## 72 B66095 York Region DSB  English Public Central Region      Aurora
0.88
## 48 B66125        Peel DSB  English Public Central Region Mississauga
0.83
##    G6_EQAO_S G6_EQAO_P G10_OSSLT G10_Credit_Acc G10_Credit_Acc_P
G11_Credit_Acc
## 63      0.48        NA      0.85           0.78             0.01
0.81
## 72      0.65        NA      0.93           0.87             0.00
0.90
## 48      0.48        NA      0.85           0.80             0.00
0.83
##    G11_Credit_Acc_P Y4_Grd_Rt Y4_Grd_Rt_P Y5_Grd_Rt Y5_Grd_Rt_P Enrollment
```

```
## 63             0.01     0.809       0.012     0.868       0.009     235340
## 72             0.00     0.903       0.002     0.945      -0.002     127935
## 48            -0.01     0.870      -0.008     0.921       0.006     150405
##       Funding
## 63 2695540462
## 72 1986451818
## 48 1858123930
```

```
top_3_enrollment  <- data_cleaned [order(-data_cleaned $Enrollment ), ][1:3,
]
print(top_3_enrollment )
```

```
##      Code             Name Language   Type         Region        City
G6_EQAO
## 63 B66052     Toronto DSB  English Public Toronto Region     Toronto
0.83
## 48 B66125        Peel DSB  English Public Central Region Mississauga
0.83
## 72 B66095 York Region DSB  English Public Central Region      Aurora
0.88
##    G6_EQAO_S G6_EQAO_P G10_OSSLT G10_Credit_Acc G10_Credit_Acc_P
G11_Credit_Acc
## 63      0.48        NA      0.85           0.78             0.01
0.81
## 48      0.48        NA      0.85           0.80             0.00
0.83
## 72      0.65        NA      0.93           0.87             0.00
0.90
##    G11_Credit_Acc_P Y4_Grd_Rt Y4_Grd_Rt_P Y5_Grd_Rt Y5_Grd_Rt_P Enrollment
## 63             0.01     0.809       0.012     0.868       0.009     235340
## 48            -0.01     0.870      -0.008     0.921       0.006     150405
## 72             0.00     0.903       0.002     0.945      -0.002     127935
##       Funding
## 63 2695540462
## 48 1858123930
## 72 1986451818
```

Note: Top 3 enrollments and top 3 funding school boards are same indicating more funding as they have more enrollments. Lets check region based enrollments,

```
# Compute and sort average funding by Region
avg_funding_by_region  <- aggregate(Funding  ~ Region , data = data_cleaned ,
FUN = mean, na.rm = TRUE)
print(avg_funding_by_region )
```

```
##           Region    Funding
## 1 Central Region  998329879
## 2    East Region  326404901
## 3   North Region   81267047
## 4 Toronto Region 1050225287
## 5    West Region  420222240
```

```r
# Compute and sort average enrollment by Region
avg_erollment_by_region  <- aggregate(Enrollment  ~ Region , data =
data_cleaned , FUN = mean, na.rm = TRUE)
print(avg_erollment_by_region )
```

```
##           Region Enrollment
## 1 Central Region  66745.000
## 2    East Region  26887.812
## 3   North Region   5426.842
## 4 Toronto Region  87211.250
## 5    West Region  27034.048
```

Note: Enrollments and Funding based on region is also a valid information. Region with higher enrollments received more provincial funds.

```r
# Compute average funding by Type in descending order
avg_funding_by_type  <- aggregate(Funding  ~ Type , data = data_cleaned , FUN
= mean, na.rm = TRUE)
# Display result
print(avg_funding_by_type )
```

```
##             Type   Funding
## 1         Public 576982760
## 2 Roman Catholic 252432142
```

```r
# Compute average enrollment by Type in descending order
avg_enrollment_by_type  <- aggregate(Enrollment  ~ Type , data = data_cleaned
, FUN = mean, na.rm = TRUE)
# Display result
print(avg_enrollment_by_type )
```

```
##             Type Enrollment
## 1         Public   40284.00
## 2 Roman Catholic   18842.35
```

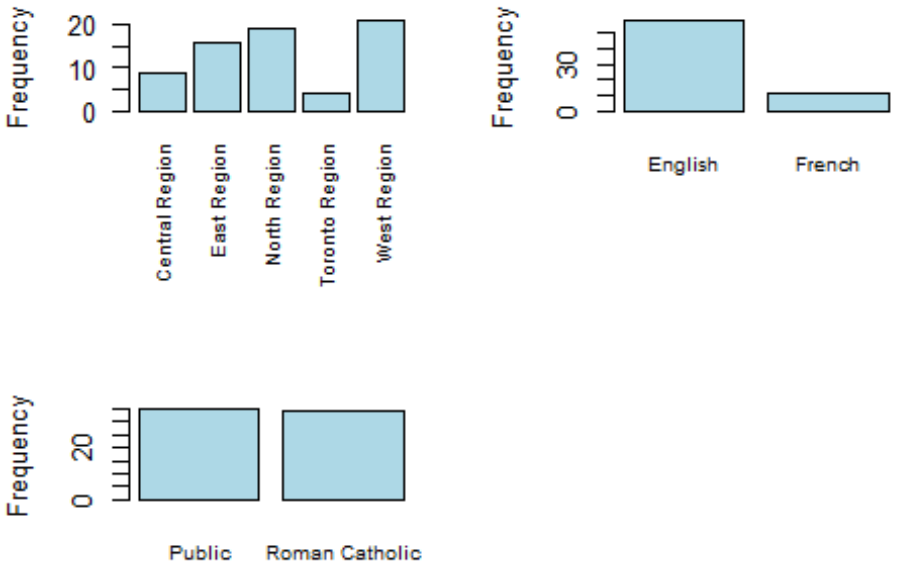
Public school boards have higher enrollments and received more funding. Based on all these analysis, I am keeping those outliers believing that they are not anomalies in data and I am keeping those values.

However, if we need to compute the average school funding at the provincial level, these high funding values can skew the mean upwards. In such cases, it is better to remove outliers to obtain a more balanced estimate of the funding per school board. Since my main goal is on analyzing credit accumulation across the province, I have decided to retain these outliers in the data set.

Check for any outliers in categorical Columns

```r
# plot box plots of categorical values
par(mfrow=c(2,2))
barplot(table(data_cleaned $Region ),col = "lightblue", cex.names=.75,
ylab="Frequency",las=2)
```

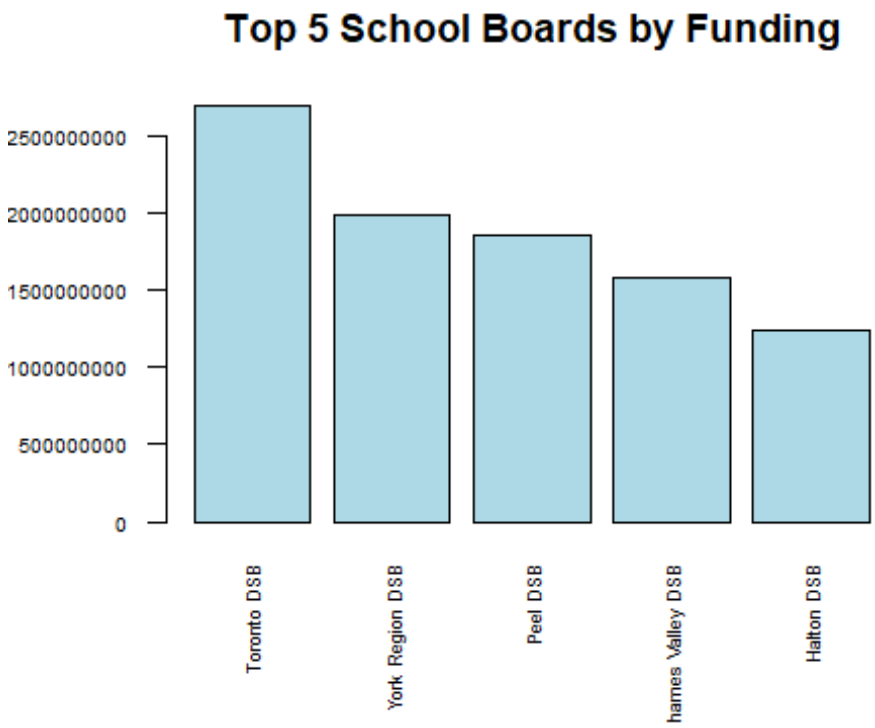
```r
barplot(table(data_cleaned $Language ),col = "lightblue", cex.names=.75,
ylab="Frequency")
barplot(table(data_cleaned $Type ),col = "lightblue", cex.names=.75,
ylab="Frequency")
par(mfrow=c(1,1))
```



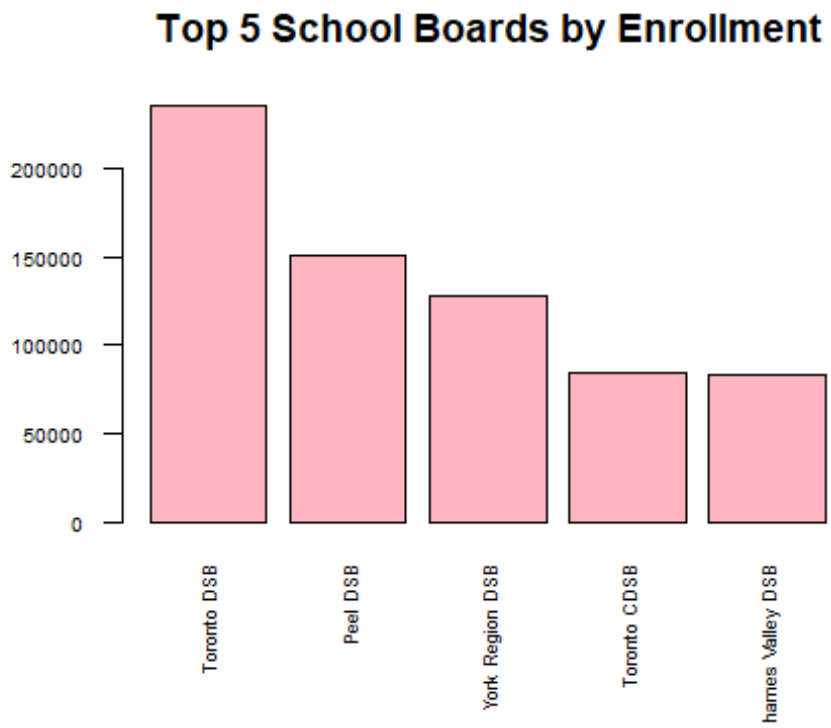
No Outliers in categorical columns.

Data Visualisations

```r
# Select top 5 school boards by funding
top_5_funding <- data_cleaned[order(-data_cleaned$Funding), ][1:5, ]

# Create bar plot
barplot(
  top_5_funding$Funding,
  names.arg = top_5_funding$Name,
  main = "Top 5 School Boards by Funding",
  xlab = "",
  ylab = "",
  col = "lightblue",
  las = 2,
  cex.names=.55,cex.lab=0.8, cex.axis = 0.6
)
```




```r
# Select top 5 school boards by Enrollment
top_5_Enrollment <- data_cleaned[order(-data_cleaned$Enrollment), ][1:5, ]

# Create bar plot
barplot(
  top_5_Enrollment$Enrollment,
  names.arg = top_5_Enrollment$Name,
  main = "Top 5 School Boards by Enrollment",
  xlab = "",
  ylab = "",
  col = "lightpink",
  las = 2,
  cex.names=.55,cex.lab=0.8, cex.axis = 0.6
)
```



Reduce Dimensionality

Drop any variables that do not contribute any useful analytical information at all.

```
colnames(data_cleaned )
```

```
##  [1] "Code"              "Name"              "Language"          "Type"
##  [5] "Region"            "City"              "G6_EQAO"           "G6_EQAO_S"
##  [9] "G6_EQAO_P"         "G10_OSSLT"         "G10_Credit_Acc"
"G10_Credit_Acc_P"
## [13] "G11_Credit_Acc"    "G11_Credit_Acc_P" "Y4_Grd_Rt"
"Y4_Grd_Rt_P"
## [17] "Y5_Grd_Rt"         "Y5_Grd_Rt_P"       "Enrollment"        "Funding"
```

Columns ‘Code’ is not useful for analytical purposes,so I am dropping that column.

```
# save a copy of data frame to compare time at th end
data_before_reduction  <- data_cleaned

# remove col 1,2,6
data_cleaned  <- data_cleaned [-c(1)]
summary(data_cleaned )
```

```
##      Name              Language              Type                Region
##  Length:69          English:58   Public         :35   Central Region: 9
##  Class :character   French :11   Roman Catholic:34   East Region   :16
##  Mode  :character                                    North Region  :19
##                                                      Toronto Region: 4
##                                                      West Region   :21
##
##      City              G6_EQAO         G6_EQAO_S      G6_EQAO_P
##  Length:69          Min.   :0.7300   Min.   :0.1600   Mode:logical
##  Class :character   1st Qu.:0.7900   1st Qu.:0.3500   NA's:69
##  Mode  :character   Median :0.8400   Median :0.5200
##                     Mean   :0.8439   Mean   :0.5297
##                     3rd Qu.:0.8800   3rd Qu.:0.6500
##                     Max.   :0.9900   Max.   :1.0000
##    G10_OSSLT      G10_Credit_Acc   G10_Credit_Acc_P   G11_Credit_Acc
##  Min.   :0.5900   Min.   :0.4900   Min.   :-0.15000   Min.   :0.5800
##  1st Qu.:0.7900   1st Qu.:0.7300   1st Qu.:-0.04000   1st Qu.:0.7600
##  Median :0.8400   Median :0.8000   Median :-0.01000   Median :0.8300
##  Mean   :0.8303   Mean   :0.7788   Mean   :-0.02014   Mean   :0.8132
##  3rd Qu.:0.8600   3rd Qu.:0.8500   3rd Qu.: 0.01000   3rd Qu.:0.8700
##  Max.   :0.9600   Max.   :0.9300   Max.   : 0.11000   Max.   :0.9500
##  G11_Credit_Acc_P     Y4_Grd_Rt       Y4_Grd_Rt_P           Y5_Grd_Rt
##  Min.   :-0.15000   Min.   :0.5880   Min.   :-0.0810000   Min.   :0.6860
##  1st Qu.:-0.03000   1st Qu.:0.7700   1st Qu.:-0.0150000   1st Qu.:0.8380
##  Median :-0.01000   Median :0.8230   Median : 0.0020000   Median :0.8940
##  Mean   :-0.01609   Mean   :0.8219   Mean   :-0.0001739   Mean   :0.8777
##  3rd Qu.: 0.01000   3rd Qu.:0.8920   3rd Qu.: 0.0150000   3rd Qu.:0.9280
##  Max.   : 0.06000   Max.   :0.9630   Max.   : 0.0660000   Max.   :0.9740
##   Y5_Grd_Rt_P          Enrollment        Funding
##  Min.   :-0.058000   Min.   :  1275   Min.   :  11395689
##  1st Qu.:-0.008000   1st Qu.:  7205   1st Qu.: 107295602
##  Median : 0.005000   Median : 16095   Median : 256366925
##  Mean   : 0.003971   Mean   : 29719   Mean   : 417059267
##  3rd Qu.: 0.011000   3rd Qu.: 35880   3rd Qu.: 507779211
##  Max.   : 0.059000   Max.   :235340   Max.   :2695540462
```

Apply the Missing Value Filter to remove appropriate columns of data.

Look at column, G6_EQAO_P

As all values are NA, I am dropping this column.

```
# remove col 8
data_cleaned  <- data_cleaned [-c(8)]
summary(data_cleaned )
```

```
##      Name              Language              Type                Region
##  Length:69          English:58   Public         :35   Central Region: 9
##  Class :character   French :11   Roman Catholic:34   East Region   :16
##  Mode  :character                                    North Region  :19
##                                                      Toronto Region: 4
```

```
##                                                  West Region   :21
##
##      City              G6_EQAO         G6_EQAO_S        G10_OSSLT
##  Length:69          Min.   :0.7300   Min.   :0.1600   Min.   :0.5900
##  Class :character   1st Qu.:0.7900   1st Qu.:0.3500   1st Qu.:0.7900
##  Mode  :character   Median :0.8400   Median :0.5200   Median :0.8400
##                     Mean   :0.8439   Mean   :0.5297   Mean   :0.8303
##                     3rd Qu.:0.8800   3rd Qu.:0.6500   3rd Qu.:0.8600
##                     Max.   :0.9900   Max.   :1.0000   Max.   :0.9600
##  G10_Credit_Acc   G10_Credit_Acc_P   G11_Credit_Acc   G11_Credit_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 :0.8000   Median :-0.01000   Median :0.8300   Median :-0.01000
##  Mean   :0.7788   Mean   :-0.02014   Mean   :0.8132   Mean   :-0.01609
##  3rd Qu.:0.8500   3rd Qu.: 0.01000   3rd Qu.:0.8700   3rd Qu.: 0.01000
##  Max.   :0.9300   Max.   : 0.11000   Max.   :0.9500   Max.   : 0.06000
##    Y4_Grd_Rt       Y4_Grd_Rt_P          Y5_Grd_Rt       Y5_Grd_Rt_P
##  Min.   :0.5880   Min.   :-0.0810000   Min.   :0.6860   Min.   :-0.058000
##  1st Qu.:0.7700   1st Qu.:-0.0150000   1st Qu.:0.8380   1st Qu.:-0.008000
##  Median :0.8230   Median : 0.0020000   Median :0.8940   Median : 0.005000
##  Mean   :0.8219   Mean   :-0.0001739   Mean   :0.8777   Mean   : 0.003971
##  3rd Qu.:0.8920   3rd Qu.: 0.0150000   3rd Qu.:0.9280   3rd Qu.: 0.011000
##  Max.   :0.9630   Max.   : 0.0660000   Max.   :0.9740   Max.   : 0.059000
##    Enrollment        Funding
##  Min.   :  1275   Min.   :  11395689
##  1st Qu.:  7205   1st Qu.: 107295602
##  Median : 16095   Median : 256366925
##  Mean   : 29719   Mean   : 417059267
##  3rd Qu.: 35880   3rd Qu.: 507779211
##  Max.   :235340   Max.   :2695540462
```

Apply the Low Variance Filter to remove appropriate columns of data.

Check for coefficient of variance value, to find low variance variables,

```
# Select only numeric columns
data_cleaned_numeric<- data_cleaned[,unlist(lapply(data_cleaned,
is.numeric))]

# Display statistics
round(stat.desc(data_cleaned_numeric),3)
```

```
##              G6_EQAO G6_EQAO_S G10_OSSLT G10_Credit_Acc G10_Credit_Acc_P
## nbr.val       69.000    69.000    69.000         69.000           69.000
## nbr.null       0.000     0.000     0.000          0.000            8.000
## nbr.na         0.000     0.000     0.000          0.000            0.000
## min            0.730     0.160     0.590          0.490           -0.150
## max            0.990     1.000     0.960          0.930            0.110
## range          0.260     0.840     0.370          0.440            0.260
## sum           58.230    36.550    57.290         53.740           -1.390
```

```
## median          0.840     0.520     0.840          0.800           -0.010
## mean            0.844     0.530     0.830          0.779           -0.020
## SE.mean         0.008     0.027     0.008          0.011            0.006
## CI.mean.0.95    0.017     0.054     0.016          0.022            0.012
## var             0.005     0.050     0.004          0.008            0.002
## std.dev         0.069     0.224     0.067          0.091            0.048
## coef.var        0.082     0.422     0.080          0.117           -2.406
##              G11_Credit_Acc G11_Credit_Acc_P Y4_Grd_Rt Y4_Grd_Rt_P
Y5_Grd_Rt
## nbr.val              69.000           69.000    69.000      69.000
69.000
## nbr.null              0.000            9.000     0.000       1.000
0.000
## nbr.na                0.000            0.000     0.000       0.000
0.000
## min                   0.580           -0.150     0.588      -0.081
0.686
## max                   0.950            0.060     0.963       0.066
0.974
## range                 0.370            0.210     0.375       0.147
0.288
## sum                  56.110           -1.110    56.712      -0.012
60.562
## median                0.830           -0.010     0.823       0.002
0.894
## mean                  0.813           -0.016     0.822       0.000
0.878
## SE.mean               0.010            0.004     0.010       0.003
0.008
## CI.mean.0.95          0.019            0.009     0.020       0.006
0.016
## var                   0.006            0.001     0.007       0.001
0.004
## std.dev               0.080            0.037     0.084       0.024
0.066
## coef.var              0.099           -2.304     0.102    -139.444
0.075
##              Y5_Grd_Rt_P     Enrollment      Funding
## nbr.val           69.000         69.000 6.900000e+01
## nbr.null           0.000          0.000 0.000000e+00
## nbr.na             0.000          0.000 0.000000e+00
## min               -0.058       1275.000 1.139569e+07
## max                0.059     235340.000 2.695540e+09
## range              0.117     234065.000 2.684145e+09
## sum                0.274    2050580.000 2.877709e+10
## median             0.005      16095.000 2.563669e+08
## mean               0.004      29718.551 4.170593e+08
## SE.mean            0.002       4656.061 6.186603e+07
## CI.mean.0.95       0.005       9291.026 1.234517e+08
## var                0.000 1495844597.869 2.640910e+17
```

```
## std.dev             0.020      38676.150 5.138978e+08
## coef.var            5.008          1.301 1.232000e+00
```

Lets look into selected columns with low coefficient of variation to check the variability in the data.

```
# Column G6_EQAO  with CV=0.082
table(data_cleaned$G6_EQAO)

##
## 0.73 0.74 0.75 0.76 0.77 0.78 0.79  0.8 0.81 0.83 0.84 0.85 0.86 0.87 0.88
0.89
##    1    1    2    5    3    5    3    4    3    3    9    7    1    2    3
2
##  0.9 0.91 0.93 0.96 0.97 0.98 0.99
##    2    3    1    1    4    3    1

# Column G10_OSSLT  with CV=0.081
table(data_cleaned$G10_OSSLT)

##
## 0.59  0.7 0.71 0.72 0.75 0.76 0.77 0.78 0.79  0.8 0.81 0.82 0.83 0.84 0.85
0.86
##    1    1    1    3    2    1    5    1    3    3    3    5    3    6    9
5
## 0.87 0.88 0.89  0.9 0.91 0.92 0.93 0.94 0.95 0.96
##    2    3    1    2    2    1    2    2    1    1

# Column Y5_Grd_Rt  with CV=0.075
table(data_cleaned$Y5_Grd_Rt)

##
## 0.686 0.722 0.741 0.749  0.75 0.755 0.768 0.769  0.78   0.8 0.805 0.815
0.833
##     1     1     1     1     1     1     1     1     1     1     1     2
1
## 0.835 0.837 0.838 0.844 0.856 0.858 0.866 0.868 0.869  0.87 0.871 0.872
0.876
##     2     1     1     1     1     1     2     2     1     1     1     1
1
##  0.88 0.883 0.886 0.892 0.894 0.896 0.901 0.909  0.91 0.914 0.916 0.917
0.92
##     1     1     1     1     1     1     1     5     1     1     1     1
1
## 0.921 0.926 0.927 0.928 0.931 0.932 0.933 0.936 0.937 0.941 0.944 0.945
0.948
##     1     1     1     2     1     1     1     2     2     1     1     1
2
## 0.949 0.958 0.963 0.974
##     2     1     1     1
```

I have checked the coefficient of variation of various columns, I didn't find a low variance column to remove. So, I am not applying this filter and going to check next high correlation filter.

Apply the High Correlation Filter to remove appropriate columns of data.

```
# Compute spearman correlation on numeric columns only
round(cor(data_cleaned[,unlist(lapply(data_cleaned
,is.numeric))],method="spearman"),3)
```

```
##                  G6_EQAO G6_EQAO_S G10_OSSLT G10_Credit_Acc
G10_Credit_Acc_P
## G6_EQAO            1.000     1.000     0.653          0.684
0.290
## G6_EQAO_S          1.000     1.000     0.653          0.684
0.290
## G10_OSSLT          0.653     0.653     1.000          0.782
0.305
## G10_Credit_Acc     0.684     0.684     0.782          1.000
0.471
## G10_Credit_Acc_P   0.290     0.290     0.305          0.471
1.000
## G11_Credit_Acc     0.756     0.756     0.760          0.912
0.332
## G11_Credit_Acc_P   0.025     0.025     0.141          0.160
0.384
## Y4_Grd_Rt          0.687     0.687     0.759          0.845
0.284
## Y4_Grd_Rt_P       -0.277    -0.277    -0.042         -0.115           -
0.117
## Y5_Grd_Rt          0.754     0.754     0.713          0.843
0.314
## Y5_Grd_Rt_P       -0.145    -0.145    -0.122         -0.256           -
0.160
## Enrollment         0.140     0.140     0.268          0.125
0.306
## Funding            0.132     0.132     0.255          0.103
0.302
##                  G11_Credit_Acc G11_Credit_Acc_P Y4_Grd_Rt Y4_Grd_Rt_P
## G6_EQAO                   0.756            0.025     0.687      -0.277
## G6_EQAO_S                 0.756            0.025     0.687      -0.277
## G10_OSSLT                 0.760            0.141     0.759      -0.042
## G10_Credit_Acc            0.912            0.160     0.845      -0.115
## G10_Credit_Acc_P          0.332            0.384     0.284      -0.117
## G11_Credit_Acc            1.000            0.177     0.918      -0.094
## G11_Credit_Acc_P          0.177            1.000     0.083       0.309
## Y4_Grd_Rt                 0.918            0.083     1.000       0.012
## Y4_Grd_Rt_P              -0.094            0.309     0.012       1.000
## Y5_Grd_Rt                 0.910            0.154     0.941      -0.124
## Y5_Grd_Rt_P              -0.185            0.284    -0.116       0.199
```

```
## Enrollment                0.164            0.210     0.271       0.254
## Funding                   0.136            0.228     0.233       0.284
##                  Y5_Grd_Rt Y5_Grd_Rt_P Enrollment Funding
## G6_EQAO              0.754      -0.145      0.140   0.132
## G6_EQAO_S            0.754      -0.145      0.140   0.132
## G10_OSSLT            0.713      -0.122      0.268   0.255
## G10_Credit_Acc       0.843      -0.256      0.125   0.103
## G10_Credit_Acc_P     0.314      -0.160      0.306   0.302
## G11_Credit_Acc       0.910      -0.185      0.164   0.136
## G11_Credit_Acc_P     0.154       0.284      0.210   0.228
## Y4_Grd_Rt            0.941      -0.116      0.271   0.233
## Y4_Grd_Rt_P         -0.124       0.199      0.254   0.284
## Y5_Grd_Rt            1.000      -0.011      0.208   0.167
## Y5_Grd_Rt_P         -0.011       1.000      0.121   0.128
## Enrollment           0.208       0.121      1.000   0.972
## Funding              0.167       0.128      0.972   1.000
```

Here, G6_EQAO and G6_EQAO_S are perfectly correlated with a positive correlation of 1.0, We don't need to keep both, I will drop the column G6_EQAO_S .

```
# removing col 7 which is G6_EQAO_S
data_cleaned<- data_cleaned[-c(7)]
head(data_cleaned,5)
```

```
##                           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
##              City G6_EQAO G10_OSSLT G10_Credit_Acc G10_Credit_Acc_P
## 1 Sault Ste Marie    0.78      0.72           0.69            -0.02
## 2         Napanee    0.78      0.86           0.84             0.00
## 3        Seaforth    0.84      0.81           0.81             0.03
## 4         Chesley    0.76      0.75           0.67            -0.01
## 5       Brantford    0.85      0.85           0.73            -0.08
##   G11_Credit_Acc G11_Credit_Acc_P Y4_Grd_Rt Y4_Grd_Rt_P Y5_Grd_Rt
Y5_Grd_Rt_P
## 1           0.71            -0.08     0.719       0.002     0.768      -
0.026
## 2           0.87            -0.03     0.895       0.015     0.909      -
0.016
## 3           0.78            -0.01     0.802      -0.008     0.838      -
0.022
## 4           0.71            -0.01     0.715       0.004     0.815      -
0.007
## 5           0.81            -0.01     0.818       0.021     0.901
0.046
##   Enrollment   Funding
## 1      10690 175553485
```

```
## 2      11995 181577452
## 3      15530 277403512
## 4      18300 277562637
## 5      11775 157646981
```

Benefits of reducing the dimensionality of this particular dataset? Be specific. For example, if it increases computational efficiency, specify how much of an improvement.

```
# Compute the time taken to process dataframe before dimensionality reduction
start1<-Sys.time()
corr_start1 <- round(cor(data_before_reduction
[,unlist(lapply(data_before_reduction,is.numeric))],method="spearman"),3)
end1  <- Sys.time()
time_before_reduction<-end1 -start1


# Compute the time taken to process dataframe after dimensionality reduction
start2  <- Sys.time()
corr_start2 <- round(cor(data_cleaned[,unlist(lapply(data_cleaned
,is.numeric))],method="spearman"),3)
end2  <- Sys.time()
time_after_reduction<- end2 -start2

# Time taken before dimensionality reduction
print(paste("Before Reduction:", round(time_before_reduction, 5)))

## [1] "Before Reduction: 0.00108"

# Time taken after dimensionality reduction
print(paste("After Reduction:", round(time_after_reduction, 5)))

## [1] "After Reduction: 0.00097"

# Time saved after dimensionality reduction
print(paste("Saved Time:", round(time_before_reduction-time_after_reduction ,
5)))

## [1] "Saved Time: 0.00011"
```

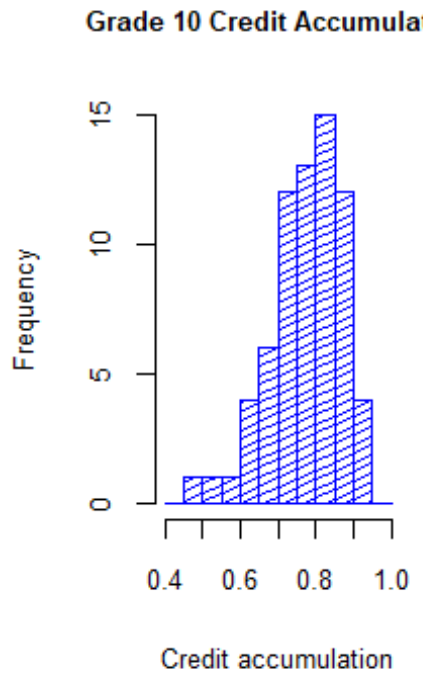
By removing non-essential columns, the time taken to process the dataset is reduced. The original computation took 0.00514 seconds, while the optimized version took 0.00487 seconds, resulting in a time savings of approximately 0.00027 seconds. This may seem small, but for larger datasets, this efficiency gain will be really make computation more efficient.

Organizing Data

Histogram and Scatter Plots

Histogram for Grade 10 Credit Accumulation

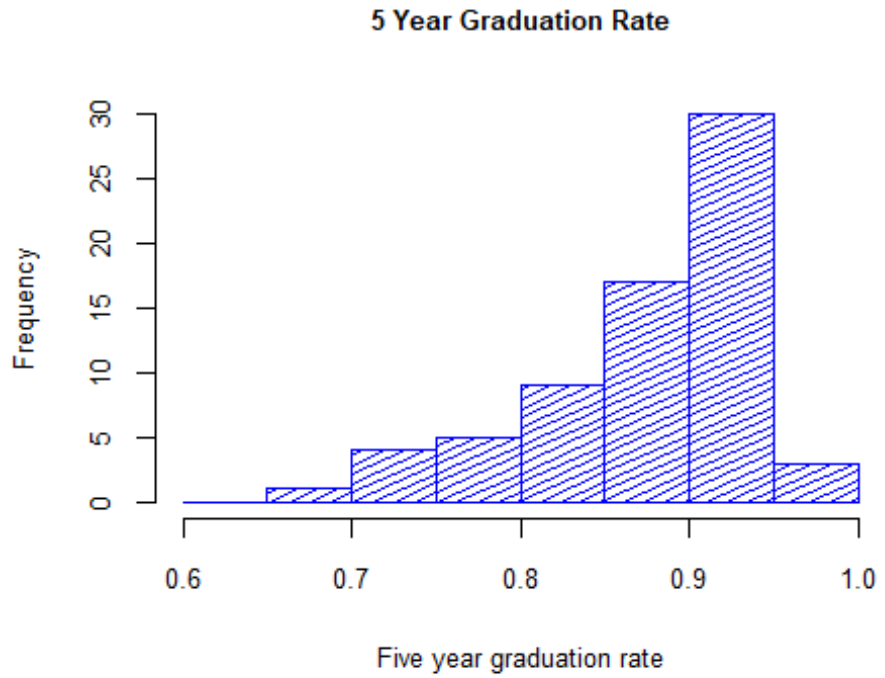
```
par(mfrow = c(1, 2))
hist(data_cleaned $G10_Credit_Acc  ,
     col="blue",
     main = "Grade 10 Credit Accumulation",
     ylab = "Frequency",
     xlab = "Credit accumulation",
     breaks = seq(0.4, 1, 0.05),
     density = 23, angle= 24,
     cex.main=0.8, cex.lab = 0.8, cex.axis = 0.8
     )
```



Histogram for Five Year Graduation Rate

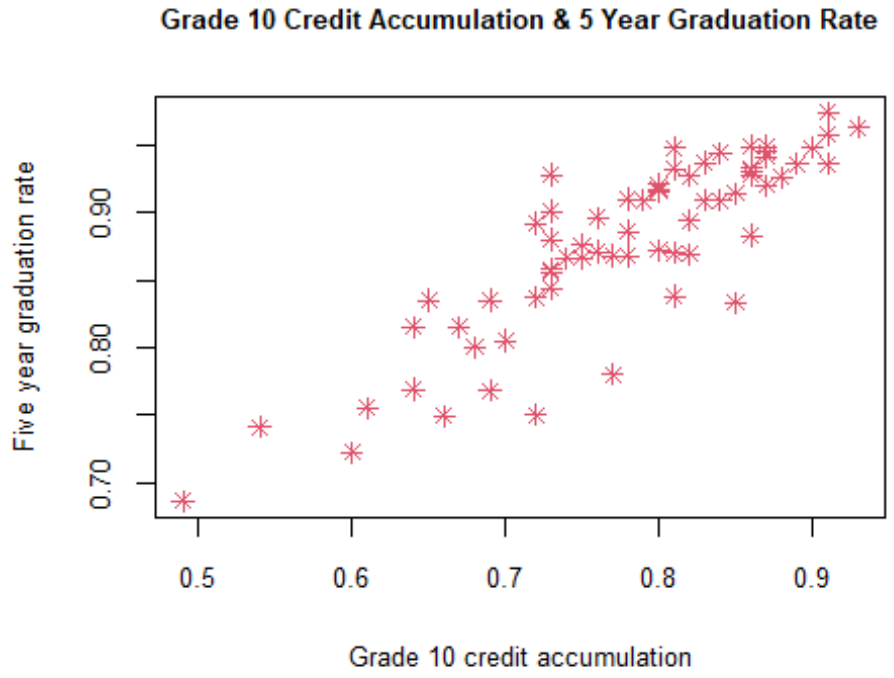
```
hist(data_cleaned $Y5_Grd_Rt  ,
     col="blue",
     main = "5 Year Graduation Rate",
     ylab = "Frequency",
     xlab = "Five year graduation rate",
     breaks = seq(0.6, 1, 0.05),
     density = 23, angle= 24,
```

```
     cex.main=0.8, cex.lab = 0.8, cex.axis = 0.8
     )
```



scatter plot showing the relationship between Grade 10 Credit Accumulation and Five Year Graduation Rate.

```
par(mfrow = c(1, 1))
plot(data_cleaned $G10_Credit_Acc , data_cleaned $Y5_Grd_Rt ,
     main = "Grade 10 Credit Accumulation & 5 Year Graduation Rate",
     xlab = "Grade 10 credit accumulation",
     ylab = "Five year graduation rate",
     col = 122,
     pch = 8,
     cex.main=0.8, cex.lab = 0.8, cex.axis = 0.8)
abline(coef = c(6,0))
```



Conclusions from chart

From the histograms, both histograms are mildly left skewed distribution.

From the scatter plot, there appears to be a strong positive correlation between Grade 10 credit accumulation rates and 5 year graduation rates. As Grade 10 credit accumulation increases, the 5 year graduation rate also tends to increase. This suggests that students who accumulate more credits by Grade 10 are more likely to graduate within five years.

###correlation coefficient between Grade 10 credit accumulation rates and 5 year graduation rates

```
# calculating spearman correlation coefficient
round(cor(data_cleaned $G10_Credit_Acc , data_cleaned $Y5_Grd_Rt , method =
"spearman"),3)
```

```
## [1] 0.843
```

Reasons for choosing the spearman correlation coefficient:

From the histogram plot,

- G10_Credit_Acc : Data in this column shows mildly left skewed distribution.
- Y5_Grd_Rt : Data in this column shows more left skewed distribution than the previous.

Pearson correlation assumes both variables are normally distributed. Since variables are skewed, Pearson is not a good choice. That is the reason behind my selection of spearman correlation test.

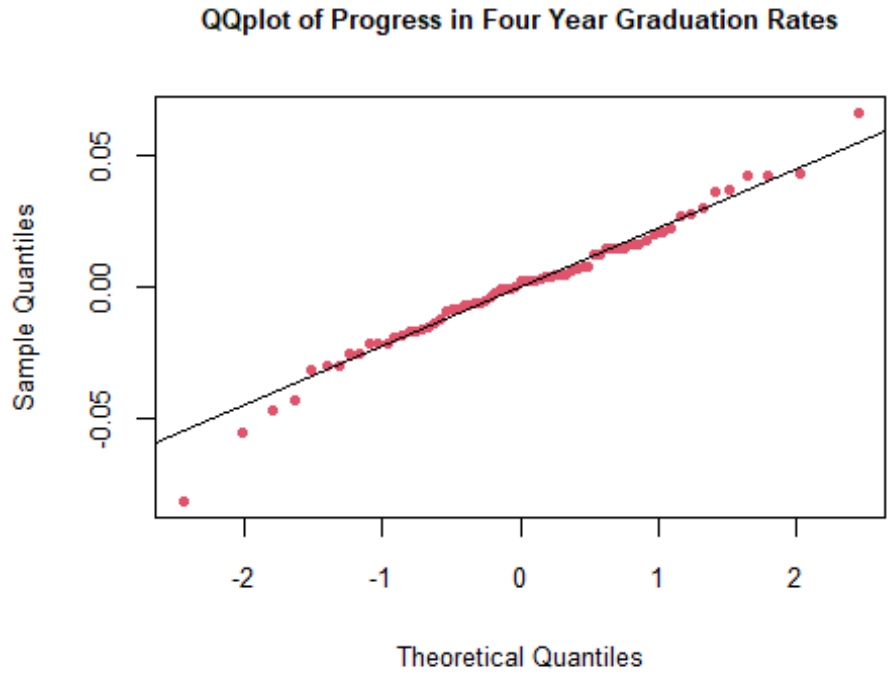
The correlation coefficient is 0.843 , which shows strong positive correlation between the variables G10_Credit_Acc and Y5_Grd_Rt .

Inference

Normality

QQ Normal plot of for Progress in Four Year Graduation Rates.

```
# Create a QQ Normal plot of for Progress in Four Year Graduation Rates.
qqnorm(data_cleaned $Y4_Grd_Rt_P ,
       main="QQplot of Progress in Four Year Graduation Rates",
       pch=20,col=26,
       cex.main=0.8,
       cex.lab = 0.8,
       cex.axis = 0.8)

qqline(data_cleaned $Y4_Grd_Rt_P )
```



statistical test for normality on Progress in Four Year Graduation Rates.

```
# Shapiro-wilk Test for normality on Progress in Four Year Graduation Rates
shapiro.test(data_cleaned $Y4_Grd_Rt_P )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_cleaned$Y4_Grd_Rt_P
## W = 0.97838, p-value = 0.2759
```

Is Progress in Four Year Graduation Rates normally distributed?

Yes, Progress in Four Year Graduation Rates are normally distributed.

From the QQ plot, It is very clear that the column Y4_Grd_Rt_P is normally distributed. However, to make sure, let's check the shapiro-wilk test results.

From Shapiro-Wilk normality test Result :

The Null Hypothesis of Shapiro-Wilk Test is that the variable is Normally distributed. As p-value = 0.2759 > 0.05, we failed to reject the Null Hypothesis. Thus, there is no strong evidence to suggest that the data is not normally distributed.

Both the QQ plot and the Shapiro-Wilk test confirm that Progress in Four-Year Graduation Rates follows a normal distribution.

Statistically Significant Differences

Compare Progress in Four Year Graduation Rates between Types of School Board in your dataset using a hypothesis test.

T-Test

```
# Run T-test ( Reason for the selection is imcluded in the coming section)
t.test(Y4_Grd_Rt_P  ~ Type , data = data_cleaned , var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Y4_Grd_Rt_P by Type
## t = 1.7232, df = 67, p-value = 0.08946
## alternative hypothesis: true difference in means between group Public and
group Roman Catholic is not equal to 0
## 95 percent confidence interval:
##  -0.001570502  0.021410838
## sample estimates:
##         mean in group Public mean in group Roman Catholic
##                  0.004714286                 -0.005205882
```

T - test, p-value = 0.08946 > 0.05, So we failed to reject the null hypothesis. This means that there is no significant difference in means of Progress in 4 yr graduation rate between Public and Roman Catholic school boards.

Reason for the choice of T-test

Selected test : T-test

First check the variable is categorical or continuous. The progress in 4 yr graduation rate is continuous,

Let's check how many groups are compared,

```
# Check for unique values in the column 4 yr graduation rate
unique(data_cleaned $Type )
```

```
## [1] Public         Roman Catholic
## Levels: Public Roman Catholic
```

We have two levels(2 groups), Public and Roman Catholic as Types of Boards.

We have two options, T-test or Wilcoxian Rank as we have 2 groups.

For T-test to use, 3 assumptions must satisfy,

1. Data Independence(satisfied, as Progress in 4 yr graduation rate of public board is independent with Roman catholic type)

2. Data is Normally distributed(satisfied, from the previous section, Progress in Four Year Graduation Rates is normally distributed as p-value = 0.2759 > 0.05, we failed to reject the null hypothesis, thus we assume it is normally distributed.)

```
# Shapiro-wilk Test for normality on Progress in Four Year Graduation Rates
shapiro.test(data_cleaned $Y4_Grd_Rt_P )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_cleaned$Y4_Grd_Rt_P
## W = 0.97838, p-value = 0.2759
```

3. Variance is unknown, but equal (satisfied from the above F-Test, p-value = 0.1462 > 0.05, failed to reject the null hypothesis. The variance of the two groups should be equal.)

```
# F- test for variance
var.test(Y4_Grd_Rt_P  ~ Type , data = data_cleaned )
```

```
##
##  F test to compare two variances
##
## data:  Y4_Grd_Rt_P by Type
```

```
## F = 0.60211, num df = 34, denom df = 33, p-value = 0.1462
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3017652 1.1970895
## sample estimates:
## ratio of variances
##          0.6021071
```

Since p-value = 0.1462 > 0.05, we failed to reject the null hypothesis. This means we do not have strong evidence to conclude that the variances are different between the groups. Therefore, we assume variances when comparing means of Y4_Grd_Rt_P between type of school boards are equal. As all three assumptions (independence, normality, and equal variance) are satisfied, I have chosen the T-test to compare the means of the two groups as in previous section.

Do we have strong evidence that Progress in Four Year Graduation Rate is different between Types of school board?

No, Since the p-value of t-test (0.08946) is greater than 0.05, we fail to reject the null hypothesis. There is no strong statistical evidence that the Progress in Four-Year Graduation Rate differs between Public and Roman Catholic schools.

Multiple Statistical Differences

Determine if Grade 6 EQAO scores vary by Language and Board Type using ANOVA (statistical) and a sequence of boxplots (graphical).

Grade 6 EQAO scores by Language and Board Type using ANOVA (statistical)

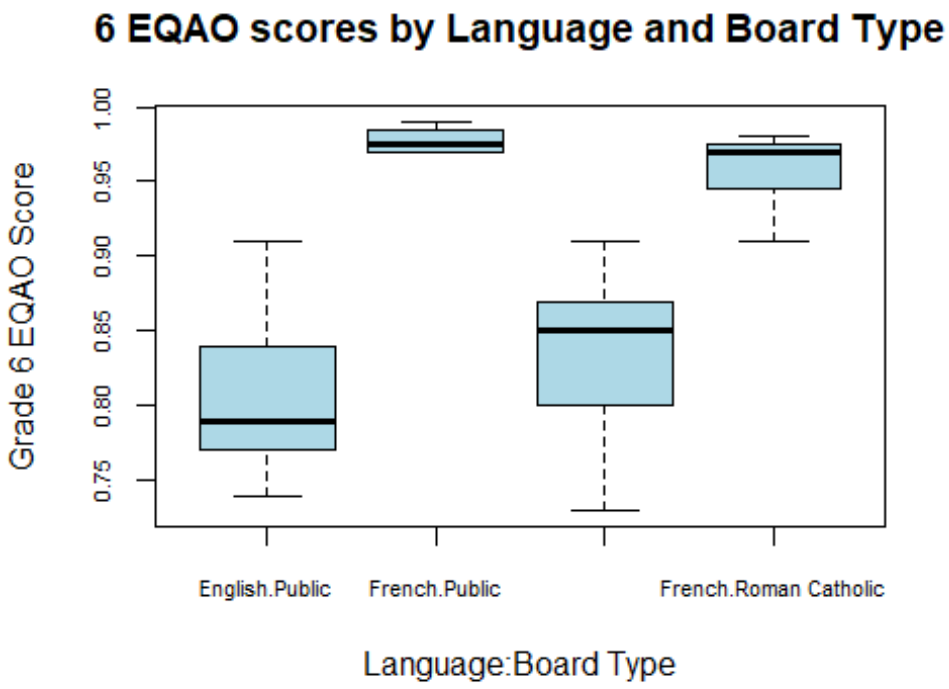
```
#Two-Way ANOVA
ANOVA_GR6_EQAO_LB  <- aov(G6_EQAO  ~ Language  + Type , data = data_cleaned )
summary(ANOVA_GR6_EQAO_LB )

##              Df  Sum Sq Mean Sq F value   Pr(>F)
## Language      1 0.19043 0.19043  98.203 1.08e-14 ***
## Type          1 0.00642 0.00642   3.313   0.0733 .
## Residuals    66 0.12799 0.00194
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Since Pr(>F) value = 1.08e-14 < 0.05, we reject the null hypothesis. We have strong evidence that Grade 6 EQAO scores vary by language.
- Since Pr(>F) value = 0.0733 > 0.05, we fail to reject the null hypothesis. There is no strong evidence to conclude that Grade 6 EQAO scores vary by board type.

- Overall, language has a much stronger impact than board type on Grade 6 EQAO scores.

Grade 6 EQAO scores by Language and Board Type using boxplots (graphical).

```
# Grade 6 EQAO scores vary by Language and Board Type
boxplot(G6_EQAO  ~ Language  + Type , data=data_cleaned ,
        main="6 EQAO scores by Language and Board Type",
        xlab= "Language:Board Type",col = "lightblue",
        ylab= "Grade 6 EQAO Score",cex.axis = 0.7)
```



From the box plot, Grade 6 EQAO vary by Language. But in case, of Grade 6 EQAO score by board types, there is no much visible variation.French-language schools scoring higher than English-language schools.Board type has a smaller impact, but Catholic schools score slightly higher than Public schools.

Determine if Grade 6 EQAO scores vary by Region using ANOVA (statistical) and a sequence of boxplots (graphical).

Grade 6 EQAO scores by Region using ANOVA (statistical)

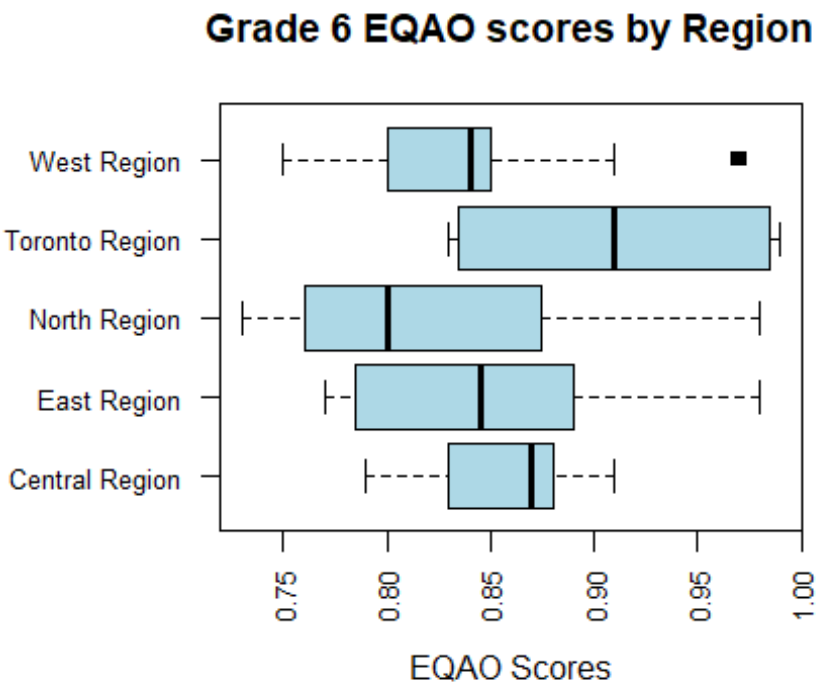
```
#One-Way ANOVA
ANOVA_GR6_EQAO_R  <- aov(G6_EQAO ~Region , data=data_cleaned )
summary(ANOVA_GR6_EQAO_R )
```

```
##              Df  Sum Sq  Mean Sq F value Pr(>F)
## Region        4 0.02648 0.006621    1.42  0.237
## Residuals    64 0.29836 0.004662
```

- Since Pr(>F) value = 0.237 > 0.05, We failed reject the null hypothesis. There is no strong evidence that Grade 6 EQAO scores differ significantly between regions.

Grade 6 EQAO scores by Region using boxplots(graphical).

```
# Box plot of Grade 6 EQAO scores  by Region
par(mar = c(5, 8, 4, 2))
boxplot(G6_EQAO  ~ Region ,
        ylab = "",
        data = data_cleaned ,
        main = "Grade 6 EQAO scores by Region",
        col = "lightblue",
        horizontal=TRUE,
        xlab = "EQAO Scores",
        pch=15,las=2,
        cex.axis = 0.8 )
```



The box plot visually represents the distribution of Grade 6 EQAO scores across different regions. There are some visible differences, but not enough to be statistically significant.

4. References

1. Marsh, D. (2025). PROG8435-L01-25W [Lecture slides].
2. Marsh, D. (2025). PROG8435-L02-25W [Lecture slides].
3. Marsh, D. (2025). PROG8435-L03-25W-Clss [Lecture slides].
4. Marsh, D. (2025). PROG8435-L04-25W [Lecture slides].
5. Marsh, D. (2025). PROG8435-L05-24F [Lecture slides].
6. Marsh, D. (2025). PROG8435-Demo-Summarize.Rmd [R Markdown file]
7. Marsh, D. (2025). PROG8435-Inference Demo.Rmd [R Markdown file]
8. Marsh, D. (2025). PROG8435_Dimensionality_Demo.Rmd [R Markdown file]
9. Marsh, D. (2025). PROG8435 Outlier Demo.Rmd [R Markdown file] 10 Marsh, D. (2025). PROG8435-ANOVA-Demo.Rmd [R Markdown file] 11.Marsh, D. (2025) Creating Graphs for All Numeric Variables [Video]. YouTube. https://www.youtube.com/watch?v=j3uhHpYtXNg
10. Marsh, D. (2025). A note on Hypothesis Testing[PDF]
11. Douglas, A., Roos, D., Mancini, F., Couto, A., & Lusseau, D. (2024). An introduction to R. Retrieved: https://intro2r.com/mult_graphs.html
12. Bhargava, I. (2025, January 28). London-area school boards get more than $260M to build 6 new schools. CBC News. https://www.cbc.ca/news/canada/london/london-area-school-boards-get-more-than-260m-to-build-6-new-schools-1.7443833
13. Government of Ontario. (2024). School board achievements and progress (2022-2023). The Education Quality and Accountability Office (EQAO). Retrieved from https://data.ontario.ca/dataset/school-board-achievements-and-progress