Practicum

Pearl Naomi Munnangi

2023-06-27

# 1. Data Loading

# Load the downloaded datasets into separate dataframes and observe the dimensions  
gt1 <- read\_excel("G&T Results 2017-18.xlsx")  
gt2 <- read\_excel("G&T Results 2018-19.xlsx")

## New names:  
## • `` -> `...13`  
## • `` -> `...14`

dim(gt1)

## [1] 117 12

dim(gt2)

## [1] 100 14

The two files are loaded into r dataframes, there are 117 rows and 12 columsns in 2017-18 dataset and there are 100 rows and 14 columns in 2018-19 dataset

colnames(gt1)

## [1] "Timestamp" "Entering Grade Level"   
## [3] "District" "Birth Month"   
## [5] "OLSAT Verbal Score" "OLSAT Verbal Percentile"   
## [7] "NNAT Non Verbal Raw Score" "NNAT Non Verbal Percentile"  
## [9] "Overall Score" "School Preferences"   
## [11] "School Assigned" "Will you enroll there?"

colnames(gt2)

## [1] "Timestamp" "Entering Grade Level"   
## [3] "District" "Birth Month"   
## [5] "OLSAT Verbal Score" "OLSAT Verbal Percentile"   
## [7] "NNAT Non Verbal Raw Score" "NNAT Non Verbal Percentile"  
## [9] "Overall Score" "School Preferences"   
## [11] "School Assigned" "Will you enroll there?"   
## [13] "...13" "...14"

# 2. Data Combining

# Remove the the Extra columns from gt2   
gt2 <- gt2[ , -c(13:14)]  
  
# Combine the two data frames  
gt <- rbind(gt1, gt2)  
dim(gt)

## [1] 217 12

Now the dataset has 217 rows and 12 columns

str(gt)

## tibble [217 × 12] (S3: tbl\_df/tbl/data.frame)  
## $ Timestamp : POSIXct[1:217], format: "2017-04-08 06:44:01" "2017-04-07 10:40:45" ...  
## $ Entering Grade Level : chr [1:217] "1" "K" "1" "K" ...  
## $ District : chr [1:217] "6" NA NA NA ...  
## $ Birth Month : chr [1:217] "September" "August" "March" "September" ...  
## $ OLSAT Verbal Score : chr [1:217] "28/30" "25" "27" "23" ...  
## $ OLSAT Verbal Percentile : chr [1:217] "99" "99" "96" "97" ...  
## $ NNAT Non Verbal Raw Score : chr [1:217] "45/50" "39" "42" "40" ...  
## $ NNAT Non Verbal Percentile: chr [1:217] "99" "99" "99" "99" ...  
## $ Overall Score : num [1:217] 99 99 98 98 99 0 99 99 95 99 ...  
## $ School Preferences : chr [1:217] "NEST+m, TAG, Anderson, Q300" "Anderson, NEST+m" NA NA ...  
## $ School Assigned : chr [1:217] "NEST" NA NA NA ...  
## $ Will you enroll there? : chr [1:217] "YES" "Maybe" "Maybe" NA ...

# 3. Data Cleaning

#a. Missing values

# Find the missing value counts in each feature  
sapply(gt, function(x) sum(is.na(x)))

## Timestamp Entering Grade Level   
## 4 0   
## District Birth Month   
## 15 1   
## OLSAT Verbal Score OLSAT Verbal Percentile   
## 6 1   
## NNAT Non Verbal Raw Score NNAT Non Verbal Percentile   
## 6 0   
## Overall Score School Preferences   
## 0 68   
## School Assigned Will you enroll there?   
## 165 102

School preferences, School assigned and will you enroll there features are having missing values more than 10 % hence those columns/features ar excluded

# Drop the columns with high missisng values  
gt <- select(gt,-c("School Preferences","School Assigned","Will you enroll there?"))  
dim(gt)

## [1] 217 9

#Now the rows with the mising values will be ommitted  
gt <- na.omit(gt)  
dim(gt)

## [1] 193 9

Now the dataset has 193 non missing rows with 9 features

#b. Errors

# Some columns has score in x/y format. Convert that into numeric  
select\_first <- function(x) {  
 sapply(strsplit(as.character(x), "/"), `[`, 1)  
}  
  
gt$"OLSAT Verbal Score" <- as.numeric(select\_first(gt$"OLSAT Verbal Score"))

## Warning: NAs introduced by coercion

gt$"NNAT Non Verbal Raw Score" <- as.numeric(select\_first(gt$"NNAT Non Verbal Raw Score"))

## Warning: NAs introduced by coercion

# Convert other numeric variables to num from chr  
gt$"OLSAT Verbal Percentile" <- as.numeric(gt$"OLSAT Verbal Percentile")

## Warning: NAs introduced by coercion

gt$"NNAT Non Verbal Percentile" <- as.numeric(gt$"NNAT Non Verbal Percentile")

## Warning: NAs introduced by coercion

str(gt)

## tibble [193 × 9] (S3: tbl\_df/tbl/data.frame)  
## $ Timestamp : POSIXct[1:193], format: "2017-04-08 06:44:01" "2017-04-10 10:18:34" ...  
## $ Entering Grade Level : chr [1:193] "1" "K" "K" "K" ...  
## $ District : chr [1:193] "6" "22" "Anderson" "11" ...  
## $ Birth Month : chr [1:193] "September" "April" "February" "December" ...  
## $ OLSAT Verbal Score : num [1:193] 28 2 26 21 25 28 28 28 28 3 ...  
## $ OLSAT Verbal Percentile : num [1:193] 99 98 99 96 99 99 99 97 99 89 ...  
## $ NNAT Non Verbal Raw Score : num [1:193] 45 38 42 43 85 46 38 44 41 42 ...  
## $ NNAT Non Verbal Percentile: num [1:193] 99 99 99 99 99 99 99 99 99 9 ...  
## $ Overall Score : num [1:193] 99 99 99 98 99 99 99 98 99 96 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:24] 2 3 4 6 8 9 10 11 12 13 ...  
## ..- attr(\*, "names")= chr [1:24] "2" "3" "4" "6" ...

# Drop the NAs due coecion  
gt <- na.omit(gt)  
dim(gt)

## [1] 182 9

Finally the dataset has 182 rows and 9 columns

# Find the unique values in Birth months column  
unique\_months <- unique(gt$`Birth Month`)  
print(unique\_months)

## [1] "September" "April" "February" "December" "8" "May"   
## [7] "July" "11" "January" "March" "August" "November"   
## [13] "October" "2" "June" "Febrauary" "september" "12"

# there are some numeric months so replace them with month name  
month.names <- c("January", "February", "March", "April", "May", "June", "July", "August", "September", "October", "November", "December")  
  
# Replace the numeric values with month names  
for(i in 1:12){  
 gt$`Birth Month` <- replace(gt$`Birth Month`, gt$`Birth Month` == as.character(i), month.names[i])  
}  
  
# Check the result  
unique(gt$`Birth Month`)

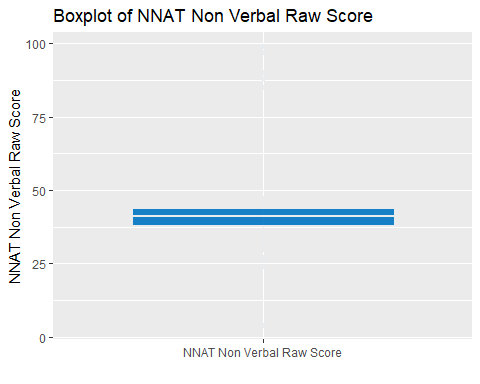
## [1] "September" "April" "February" "December" "August" "May"   
## [7] "July" "November" "January" "March" "October" "June"   
## [13] "Febrauary" "september"

# There are some instances with wrong spelleings for febrauary and september.   
#Correct the spelling of "Febrauary" and "september"  
gt$`Birth Month` <- sub("Febrauary", "February", gt$`Birth Month`)  
gt$`Birth Month` <- sub("september", "September", gt$`Birth Month`)

# 4.Data Plottng

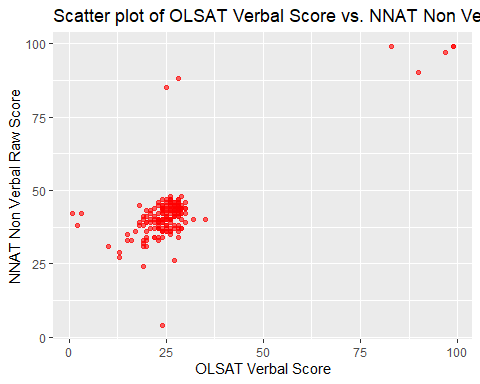
#a Box plot

# Boxplot  
ggplot(gt, aes(x = "NNAT Non Verbal Raw Score", y = `NNAT Non Verbal Raw Score`)) +  
 geom\_boxplot(fill = "#0073C2", color = "#e9ecef", alpha = 0.9) +  
 labs(title="Boxplot of NNAT Non Verbal Raw Score",  
 x="",  
 y="NNAT Non Verbal Raw Score")



#b Scatter plot

# Scatter plot  
ggplot(gt, aes(x = `OLSAT Verbal Score`, y = `NNAT Non Verbal Raw Score`)) +  
 geom\_point(alpha = 0.6, color = "red") +  
 labs(title="Scatter plot of OLSAT Verbal Score vs. NNAT Non Verbal Raw Score",  
 x="OLSAT Verbal Score",  
 y="NNAT Non Verbal Raw Score")

 #c Bar plot

ggplot(gt, aes(x = `Birth Month`)) +  
 geom\_bar(fill = "lightgreen", color = "lightblue", alpha = 0.9) +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 labs(title="Bar Chart of Birth Month Counts",  
 x="Birth Month",  
 y="Count")

