# Data Set 1 R code and Analysis

1. Created a folder on desktop called “ISQA8086\_RCode\_Assignment” and
2. Downloaded Set 1 data from GitHub repository, saving data to newly created folder from step 1.
3. Opened R studio
4. Set working directory to newly created folder  
   setwd("C:/Users/cale/Desktop/ISQA8086\_RCode\_Assignment")
5. Created data frame in R  
   set1data <- read.csv("Set 1.csv",stringsAsFactors = FALSE)
   1. This did not work because data is file type .xlsx
   2. Converted Set 1 into .CSV
   3. Re ran step 5
6. Wanted to understand more about the data  
   str(set1data)
7. Looked at data types and realized that some values were datatype of INT but should be a Factor, because they are Identity columns and not intended for math.

set1data$County <- factor(set1data$County)

set1data$District <- factor(set1data$District)

set1data$School <- factor(set1data$School)

set1data$DataYears <- factor(set1data$DataYears)

set1data$Grade\_Code <- factor(set1data$Grade\_Code)

1. Re-ran step 6, to make sure that factors were created

str(set1data)

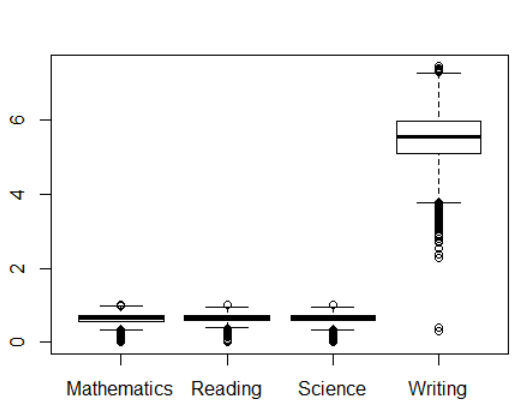
# Start of Analysis

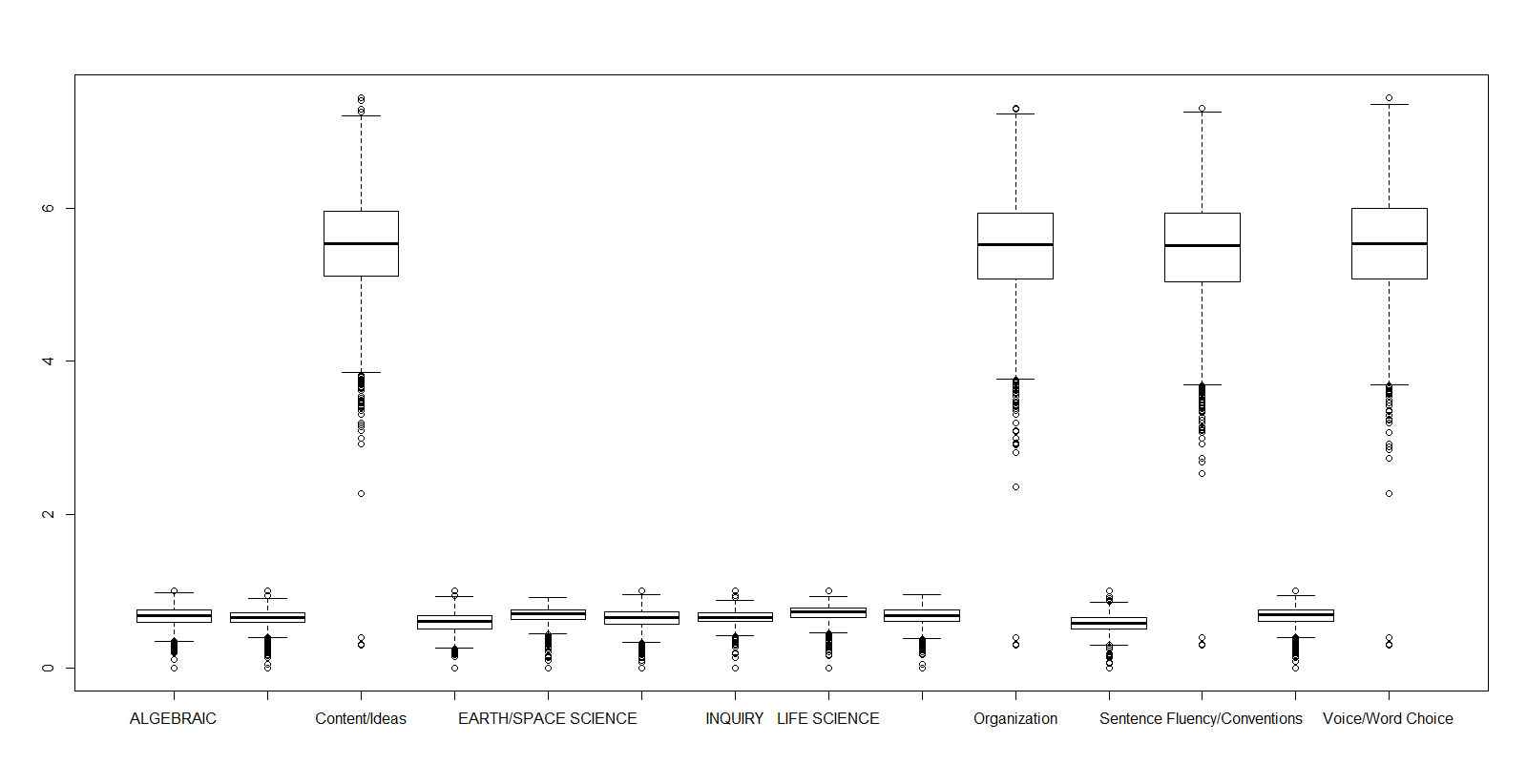
1. Of the dataset there are only 2 values we could perform statistics on.
   1. Standard.Correct.Percent
   2. Student.Count
2. Ran initial Descriptive Statistics  
   summary(set1data$Standard.Correct.Percent)  
    Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 0.610 0.710 1.956 4.520 7.440

1. Noticed that Mean was 1.956. Was under the impression test scores were between 0-1. Need to perform analysis as to why results are not as expected.
2. After a few attempts found the culprit. Writing test scores do not follow the same pattern the other exam types have.

boxplot(set1data$Standard.Correct.Percent~set1data$Subject)



1. ‘Standard’ is a subtype of ‘Subject’, wanted to verify if all scores are in Writing are follow a different scale  
   boxplot(set1data$Standard.Correct.Percent~set1data$Standard)  
   suspicion confirmed ‘Content/Ideas’, Organization, Sentence Fluency, Word Choice all belong to the Writing subject. We will need to treat the Writing subject differently from the other Subjects.
2. Split the data sets by ‘Subject’  
   subjectlist <- split(set1data, set1data$Subject)
3. Identified the Subject was split correctly  
   summary(subjectlist)  
    Length Class Mode

Mathematics 15 data.frame list

Reading 15 data.frame list

Science 15 data.frame list

Writing 15 data.frame list

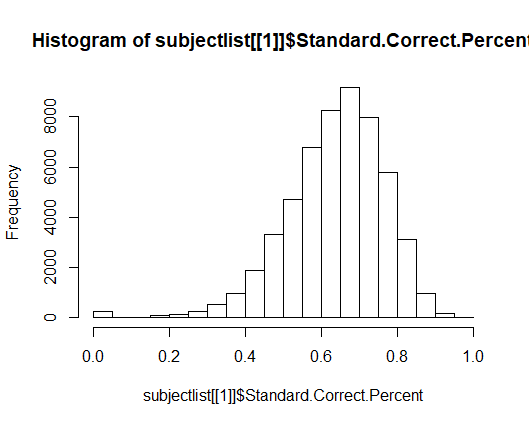
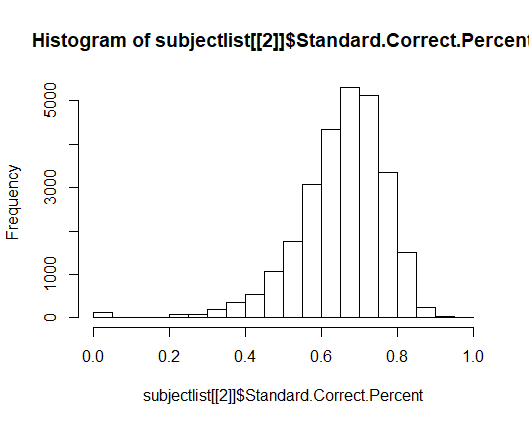
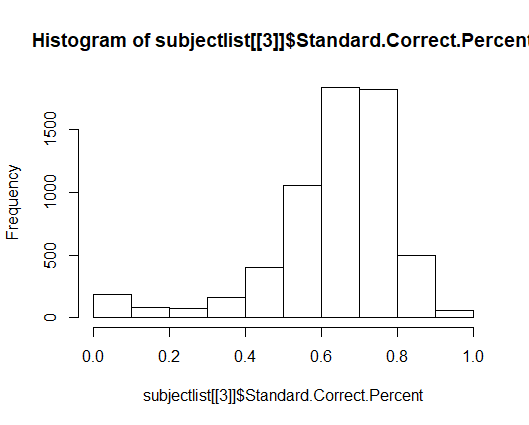
1. Ran Summary Descriptive Statistics on each Data frame  
   [1]= Mathematics

[2]= Reading

[3]= Science

[4]= Writing

|  |
| --- |
| summary(subjectlist[[1]]$Standard.Correct.Percent) |
| Min. 1st Qu. Median Mean 3rd Qu. Max. |
| 0.0000 0.5700 0.6500 0.6409 0.7300 1.0000 |
| summary(subjectlist[[2]]$Standard.Correct.Percent) |
| Min. 1st Qu. Median Mean 3rd Qu. Max. |
| 0.0000 0.6000 0.6700 0.6582 0.7400 1.0000 |
| summary(subjectlist[[3]]$Standard.Correct.Percent) |
| Min. 1st Qu. Median Mean 3rd Qu. Max. |
| 0.0000 0.5800 0.6700 0.6349 0.7400 1.0000 |
| summary(subjectlist[[4]]$Standard.Correct.Percent) |
| Min. 1st Qu. Median Mean 3rd Qu. Max. |
| 0.300 5.080 5.530 5.492 5.960 7.440 |

1. In order to perform graphical analysis need to attach the data frame  
   attach(subjectlist)
2. Build histograms of ‘Standard.Correct.Percent’ for each data frame  
   hist(subjectlist[[1]]$Standard.Correct.Percent)  
     
   hist(subjectlist[[2]]$Standard.Correct.Percent)  
     
   hist(subjectlist[[3]]$Standard.Correct.Percent)  
     
   hist(subjectlist[[4]]$Standard.Correct.Percent)  
   