# Student Performance Case Study (Clustering/Clasification)

* [Student Performance Case Study (Clustering/Clasification)](#X33a2b11363e9d5cac3891ac1c9a5331c3117232)
  + [Step 1: Identify business problem](#step-1-identify-business-problem)
    - [Question 1: What are the problems that we are trying to solve in this assignment?](#X7255f360c5c003d0d266fe51c87090e98489a6a)
    - [Question 2: What are the algorithms that could be used to solve each problem?](#X168e61215243e3c763bf03831ede406ff047fa9)
  + [Step 2: Identify data sources and acquire data](#Xd23e616e455f28f29db6fa7770ae6f466189032)
  + [Step 3: Process/Clean data](#step-3-processclean-data)
    - [Question 3: Find the number of missing data for each feature.](#X665ef97abf4ef1d0b191b489dc6f4446de66e7a)
    - [Question 4: What is the best way to deal with missing data (delete, calculate the average or the median, etc.)?](#X7cecd5a151d20cf340242922db2bf2398e36083)
    - [Question 5: Using an outlier detection technique (such as scatter, Z-score, or Box Plot), find if there are any outliers in the data (give these outliers).](#Xaa8266527322728464d727d889ca4a757382de8)
  + [Step 4: Perform exploratory analysis](#step-4-perform-exploratory-analysis)
    - [Question 6: Discover the relationship between going out and the grades of students (in other words, does, for example, going out so often affect the results of the student?)](#Xd493b09226be32dcbaa94c1f422bffe1e873233)
    - [Question 7: Is there a relationship between family size and the result of students?](#Xedeb74ef064f29c54d5edef68b640b3f5a805d2)
    - [Question 8: Does the quality of the family relationship affect students’ results? Explain your answer.](#X3c84bd5b00a4df968bf0cb0efda27917b6e2e0b)
    - [Question 9: Does the school travel time affect students result? Explain your answer.](#Xd1f8b59c60b4bd0c6ce8d29177640b9639f9f46)
    - [Question 10: What type of relationship does exist between the previous failures and students’ results? (in other words, if a student previously failed does this mean he/she will pass?!)](#X77dc5478c3d9d3dd8cae727bd66a09a57c1495f)
    - [Question 11: Do extra activities help to improve student performance?](#X1a696ad7220d3d506c75ddd7ac3f05e56f06fc4)
    - [Question 12: Does internet home access help to improve student performance?](#Xe57011039bd328f7ecdd20a6bbef57cc031d4aa)
    - [Question 13: What is the distribution of students who succeeded over the levels of free time?](#Xc83d1077755e1a8ade017ab83a69ac8fe16ef31)
    - [Question 14: What is the relationship between age and student result?](#X2807ee0d87eeb628b08a077d8a7ed9531018910)
    - [Question 15: What is the relationship between address and student result?](#X49b48f363725853921331d75f41f3bcfc0998cf)
    - [Question 16: What is the relationship between study time and student result?](#X90aa7fe66adeb1a5fe4545622ad7585063fcccb)
    - [Question 17: Does absence affect student performance?](#X2a76778cc871dc7478ed25878afd75cc0fb3712)
    - [Question 18: Does health affect student performance?](#X1ce86c29fb0a7ceba07147d29539b5498de9039)
    - [Question 19: Does mother and father education/job affect student result?](#X72a8103a934dfbf747c6512556150a7ad366564)
  + [Step 5: Generate the model](#step-5-generate-the-model)
    - [Question 20: Give the useful variables to perform the clustering and classification. Remove the non-useful variables if there are any.](#X90674072a5bcc80d9af67f6cba56380d97b6763)
    - [Question 21: Are there any features that we need to add to the dataset? Identify and add these features to the dataset?](#X81ec46c73ed5ef22eb78781d9e8e9b4265087d8)
    - [Question 22: Is there any data needed to be normalized? What techniques of normalization could be used?](#X05cc62575dbb2bdc5ccaae6fcaebe427b834dda)
  + [Step 6: Validate the model](#step-6-validate-the-model)
    - [part 1](#part-1)
    - [Question 23: Do the clustering based on these features with k=7. Showcase your results (plot of clusters).](#X510d98b45beca7663620a4527426d8dea9cf6d8)
    - [Question 24: Calculate the within-cluster variation.](#Xcec57d11fa41b910c963373bb35058b09b3ba5f)
    - [Question 25: Do the clustering based on these features with k=7. Showcase your results (plot of clusters).](#X4ac30b746ee135de294cce2186db6d8bc584ae5)
    - [Question 26: Calculate the within-cluster variation.](#X4320c1833c424438f4a15eda360024c6c42f447)
    - [Question 27: Do the clustering based on these features with k=7. Showcase your results (plot of clusters).](#X3bb9ad25200f74abf752892046a64b74ac364f1)
    - [Question 28: Calculate the within-cluster variation.](#Xf5abed679d9d0f0366899753f54449c2d017762)
    - [Question 29: Is this clustering better or worse than the previous ones? Explain your answer.](#X3ae19bc5ca2011f591ef46196fd96a360408603)
    - [Question 30: Find the appropriate number of clusters (k) for the student performance data.](#Xd1ade7da1bfc8ffa162dd0a050b87e731670408)
    - [part 2](#part-2)
    - [Question 31: Do the classification using KNN algorithm.](#X1e18fdccd71a579543c25eb02399ef0fa46ff63)
    - [Question 32: Assess the performance of the model for different values of K (number of neighbors) using confusion-matrix and accuracy.](#X6928a5c2f61321ed502cc32cca28641100a935c)
    - [Question 33: Find the best value of K.](#question-33-find-the-best-value-of-k)
    - [Question 34: Do the classification using SVM algorithm.](#X68d6d8a9229d8e33fb549bd3a64d47394c86b1f)
    - [Question 35: Assess the performance of the model using confusion-matrix and accuracy.](#X4e2baf9ee862435537732e73fe23d42d5d257bb)
    - [Question 36: Compare the performance of SVM to the accuracy of KNN.](#X24a1137377ea9d1935544d6e6f36632bc89dcc4)
    - [part 3](#part-3)
    - [Question 37: In order to answer this question, show the distribution of the results for each cluster. Explain the results.](#X60a35c07a3767569d5b75f4326aa6030153c2bd)
  + [Step 7: Visualize results](#step-7-visualize-results)
    - [Question 38: Create graphs to present important results concluded from the previous step. Comment on those graphs.](#X4667d529098edbd30db5fa37ae46df10d6f7538)

## Step 1: Identify business problem

### Question 1: What are the problems that we are trying to solve in this assignment?

**Solution 1:**

### Question 2: What are the algorithms that could be used to solve each problem?

**Solution 2:**

## Step 2: Identify data sources and acquire data

**Students’ Comment:** The data provided is extracted, transformed and given to us on this assignment. This is far from real scenarios, where identifying the data sources is very crucial. Examples for this particular scenario:

1. Gabriel Pereira student database
2. Mousinho da Silveira database
3. Students school survey database
4. National Education Authority published datasets

After the identification and acquisition of data, we read it with the programming language of choice. We have selected R as our programming language with Rstudio as our IDE.

Our project structure is as follows:

Assignment: Student Performance Case Study [Project folder] (/)  
 |  
 |\_\_\_\_\_\_ assignment-1.Rproj  
 |  
 |\_\_\_\_\_\_ assignment.R  
 |  
 |\_\_\_\_\_\_ BDA-Assignment-2019.pdf  
 |  
 |\_\_\_\_\_\_ data set.csv  
 |  
 |\_\_\_\_\_\_ README.md

Our R code to read the data will be:

# Check if the project folder is our current directory  
getwd()  
  
# If this outputs a directory other than the project folder, set the working directory using `setwd()`  
  
# Read the data set  
data <- read.csv("data set.csv")

R Console Output

> # Check if the project folder is our current directory  
> getwd()  
[1] "D:/Machine Learning/Assignment: Student Performance Case Study"  
>  
> # Read the data set  
> data <- read.csv("data set.csv")

## Step 3: Process/Clean data

### Question 3: Find the number of missing data for each feature.

R code

# counting NAs of each column  
  
# We are applying the function of x defined as sum of na counts  
# to each column and returning a matrix of counts of NAs in each column  
sapply(data, function(x) sum(is.na(x)))

R Console Output

> # counting NAs of each column  
>   
> # We are applying the function of x defined as sum of na counts  
> # to each column and returning a matrix of counts of NAs in each column  
> sapply(data, function(x) sum(is.na(x)))   
 X school sex age address famsize Pstatus Medu Fedu Mjob   
 0 0 0 0 0 0 0 0 1 0   
 Fjob reason guardian traveltime studytime failures schoolsup famsup paid activities   
 0 0 0 2 0 1 0 0 0 0   
 nursery higher internet romantic famrel freetime goout health absences G1   
 0 0 0 0 1 1 0 1 0 1   
 G2 G3   
 0 1

**Solution 3:**

### Question 4: What is the best way to deal with missing data (delete, calculate the average or the median, etc.)?

**Solution 4:** There are many ways to deal with missing data, on a broad level the 3 are the most used:

1. Delete the entire record.
2. Create a seperate model to impute the missing data.
3. Statistical methods like mean, median, mode, etc.

The method to use really depends on the kind of data that is missing. Is it going to skew our results if some records were completly deleted? Or if there are lots of missing records and we assume a mean value for all those records, the deviations from the mean would be far less from the actual reality.

R Code

# Find out all data with one or more NAs  
data.incomplete <- data[!complete.cases(data), ]

In our scenario, the records with missing values is less (1.25% of the dataset)

As we are trying to predict performance, the records with no grades would be deleted. (Record 397 and 398)

For other missing records the following code should handle the missing records. We have used mean and median to impute some of the missing values as appropriate to the features.

R Code

# split data between 2 schools and handle missing values.  
require(dplyr) # loading required library for filter function  
  
data.GP <- data %>% filter(school == "GP") # seperate dataset for GP  
  
# Median value inserted in case of NAs were features are most likely highest occurance  
data.GP$Fedu[is.na(data.GP$Fedu)] <- median(data.GP$Fedu, na.rm = T) # Median father's education  
data.GP$famrel[is.na(data.GP$famrel)] <- median(data.GP$famrel, na.rm = T) # Median of family relation  
  
# Mean value inserted in case of NAs where features are most likely the average  
data.GP$traveltime[is.na(data.GP$traveltime)] <- mean(data.GP$traveltime, na.rm = T) # Mean of Travel time  
data.GP$freetime[is.na(data.GP$freetime)] <- mean(data.GP$freetime, na.rm = T) # Mean of Free time  
  
# Same is being done for MS  
data.MS <- data %>% filter(school == "MS") # seperate dataset for MS  
  
# Median value inserted in case of NAs were features are most likely highest occurance  
data.MS$Fedu[is.na(data.MS$Fedu)] <- median(data.MS$Fedu, na.rm = T) # Median father's education  
data.MS$famrel[is.na(data.MS$famrel)] <- median(data.MS$famrel, na.rm = T) # Median of family relation  
  
# Mean value inserted in case of NAs where features are most likely the average  
data.MS$traveltime[is.na(data.MS$traveltime)] <- mean(data.MS$traveltime, na.rm = T) # Mean of Travel time  
data.MS$freetime[is.na(data.MS$freetime)] <- mean(data.MS$freetime, na.rm = T) # Mean of Free time  
  
# combining the data together to get back to the original data set with a few missing values imputed.  
data <- rbind(data.GP, data.MS)  
  
# Deleting records with missing values  
data <- na.omit(data) # 2 records deleted (397 and 398)  
  
# soring based on column X (index numbers)  
data <- data[order(data$X), ]

R Console Output

> # Find out all data with one or ore NAs  
> data.incomplete <- data[!complete.cases(data), ]  
>   
> # split data between 2 schools and handle missing values.  
> require(dplyr) # loading required library for filter function  
>   
> data.GP <- data %>% filter(school == "GP") # seperate dataset for GP  
>   
> # Median value inserted in case of NAs were features are most likely highest occurance  
> data.GP$Fedu[is.na(data.GP$Fedu)] <- median(data.GP$Fedu, na.rm = T) # Median father's education  
> data.GP$famrel[is.na(data.GP$famrel)] <- median(data.GP$famrel, na.rm = T) # Median of family relation  
>   
> # Mean value inserted in case of NAs where features are most likely the average  
> data.GP$traveltime[is.na(data.GP$traveltime)] <- mean(data.GP$traveltime, na.rm = T) # Mean of Travel time  
> data.GP$freetime[is.na(data.GP$freetime)] <- mean(data.GP$freetime, na.rm = T) # Mean of Free time  
>   
> # Same is being done for MS  
> data.MS <- data %>% filter(school == "MS") # seperate dataset for MS  
>   
> # Median value inserted in case of NAs were features are most likely highest occurance  
> data.MS$Fedu[is.na(data.MS$Fedu)] <- median(data.MS$Fedu, na.rm = T) # Median father's education  
> data.MS$famrel[is.na(data.MS$famrel)] <- median(data.MS$famrel, na.rm = T) # Median of family relation  
>   
> # Mean value inserted in case of NAs where features are most likely the average  
> data.MS$traveltime[is.na(data.MS$traveltime)] <- mean(data.MS$traveltime, na.rm = T) # Mean of Travel time  
> data.MS$freetime[is.na(data.MS$freetime)] <- mean(data.MS$freetime, na.rm = T) # Mean of Free time  
>   
> # combining the data together to get back to the original data set with a few missing values imputed.  
> data <- rbind(data.GP, data.MS)  
>   
> # Deleting records with missing values  
> data <- na.omit(data) # 2 records deleted (397 and 398)  
>   
> # soring based on column X (index numbers)  
> data <- data[order(data$X), ]

### Question 5: Using an outlier detection technique (such as scatter, Z-score, or Box Plot), find if there are any outliers in the data (give these outliers).

**Solution 5:**

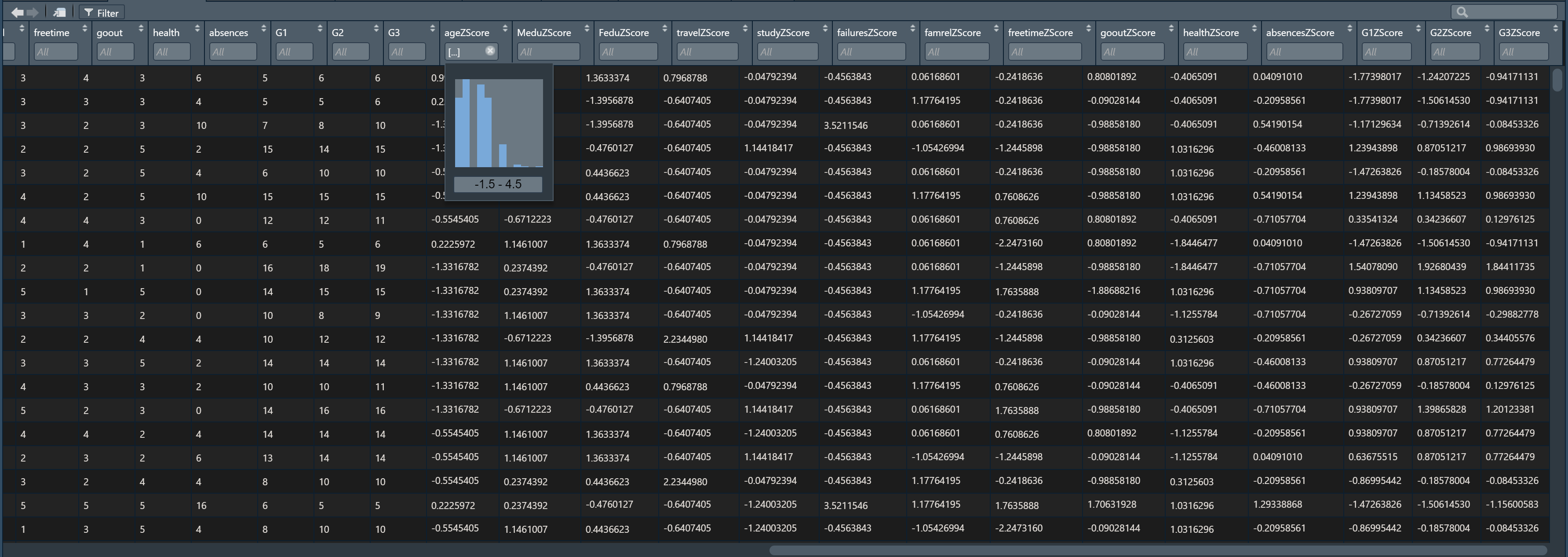
using the z-scores we could remove all the outliers by filtering out all records were the z-scores are above 3 or below -3.

Formula:

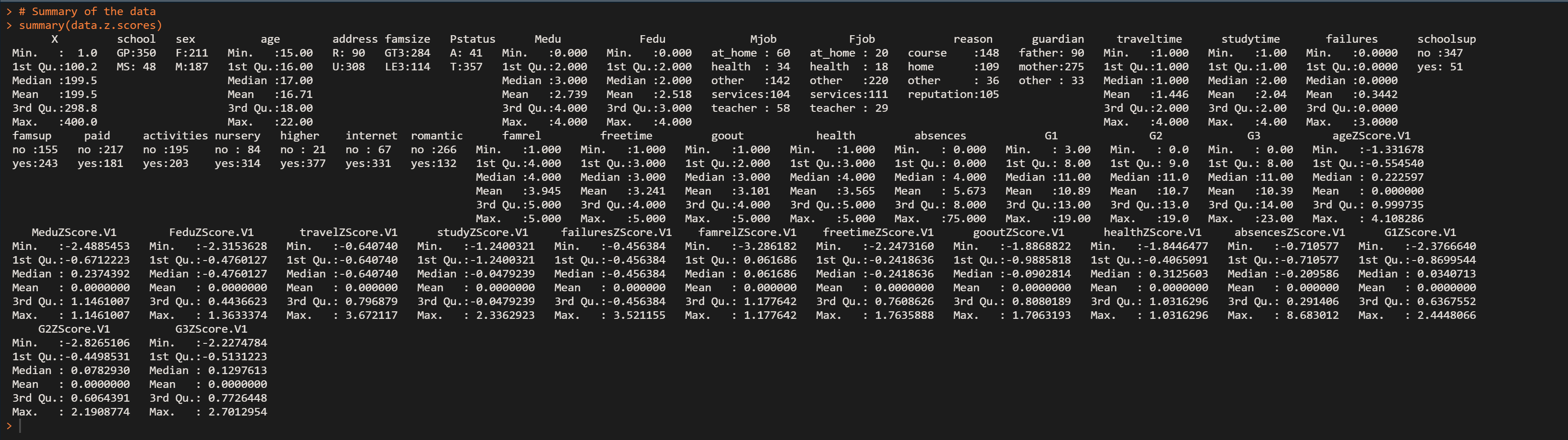
R Code

#Dectecting outlier   
# Using Z - scores  
data.z.scores <- data %>% mutate(ageZScore = scale(age),  
 MeduZScore = scale(Medu),  
 FeduZScore = scale(Fedu),  
 travelZScore = scale(traveltime),  
 studyZScore = scale(studytime),  
 failuresZScore = scale(failures),  
 famrelZScore = scale(famrel),  
 freetimeZScore = scale(freetime),  
 gooutZScore = scale(goout),  
 healthZScore = scale(health),  
 absencesZScore = scale(absences),  
 G1ZScore = scale(G1),  
 G2ZScore = scale(G2),  
 G3ZScore = scale(G3))  
  
# Summary of the data  
summary(data.z.scores)

Output:



alt text



alt text

## Step 4: Perform exploratory analysis

### Question 6: Discover the relationship between going out and the grades of students (in other words, does, for example, going out so often affect the results of the student?)

**Solution 6:**

### Question 7: Is there a relationship between family size and the result of students?

**Solution 7:**

### Question 8: Does the quality of the family relationship affect students’ results? Explain your answer.

**Solution 8:**

### Question 9: Does the school travel time affect students result? Explain your answer.

**Solution 9:**

### Question 10: What type of relationship does exist between the previous failures and students’ results? (in other words, if a student previously failed does this mean he/she will pass?!)

**Solution 10:**

### Question 11: Do extra activities help to improve student performance?

**Solution 11:**

### Question 12: Does internet home access help to improve student performance?

**Solution 12:**

### Question 13: What is the distribution of students who succeeded over the levels of free time?

**Solution 13:**

### Question 14: What is the relationship between age and student result?

**Solution 14:**

### Question 15: What is the relationship between address and student result?

**Solution 15:**

### Question 16: What is the relationship between study time and student result?

**Solution 16:**

### Question 17: Does absence affect student performance?

**Solution 17:**

### Question 18: Does health affect student performance?

**Solution 18:**

### Question 19: Does mother and father education/job affect student result?

**Solution 19:**

## Step 5: Generate the model

### Question 20: Give the useful variables to perform the clustering and classification. Remove the non-useful variables if there are any.

**Solution 20:**

### Question 21: Are there any features that we need to add to the dataset? Identify and add these features to the dataset?

**Solution 21:**

### Question 22: Is there any data needed to be normalized? What techniques of normalization could be used?

**Solution 22:**

## Step 6: Validate the model

### part 1

For features: sex, age, address, Pstatus, Medu, Fedu, Mjob, Fjob, studytime, absences.

### Question 23: Do the clustering based on these features with k=7. Showcase your results (plot of clusters).

**Solution 23:**

### Question 24: Calculate the within-cluster variation.

**Solution 24:**

For features: age, address, Fjob, traveltime, studytime, failures, and absences.

### Question 25: Do the clustering based on these features with k=7. Showcase your results (plot of clusters).

**Solution 25:**

### Question 26: Calculate the within-cluster variation.

**Solution 26:**

For features: age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob,studytime, traveltime ,failures, internet, famrel, freetime, goout,health, absences.

### Question 27: Do the clustering based on these features with k=7. Showcase your results (plot of clusters).

**Solution 27:**

### Question 28: Calculate the within-cluster variation.

**Solution 28:**

### Question 29: Is this clustering better or worse than the previous ones? Explain your answer.

**Solution 29:**

### Question 30: Find the appropriate number of clusters (k) for the student performance data.

**Solution 30:**

### part 2

features: age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, studytime, traveltime, studytime, failures, internet, famrel, freetime, goout, health, absences,G1, G2, G3, Cluster. Where Cluster is the target feature.

### Question 31: Do the classification using KNN algorithm.

**Solution 31:**

### Question 32: Assess the performance of the model for different values of K (number of neighbors) using confusion-matrix and accuracy.

**Solution 32:**

### Question 33: Find the best value of K.

**Solution 33:**

### Question 34: Do the classification using SVM algorithm.

**Solution 34:**

### Question 35: Assess the performance of the model using confusion-matrix and accuracy.

**Solution 35:**

### Question 36: Compare the performance of SVM to the accuracy of KNN.

**Solution 36:**

### part 3

### Question 37: In order to answer this question, show the distribution of the results for each cluster. Explain the results.

**Solution 37:**

## Step 7: Visualize results

### Question 38: Create graphs to present important results concluded from the previous step. Comment on those graphs.

**Solution 38:**