**Analysis Report on Parent's Combined Educational Levels Effect on Student Academic Performance**

By Justin Grima (14248599)

**I. Abstract**

Student academic success and overall school performance have been found to have a correlation with their parent's level of education. This report is an analysis of the “Student Performance Data Set”1 (“UCI Machine Learning Repository: Student Performance Data Set,”) 2014 to **address** if a higher combined parent educational level correlates to greater student academic success. The **purpose** of this report is to **i)** classify parents' combined education levels (PCEL), **ii)** analyse students' academic performances based on their parent's combined education level and **iii)** investigate other variables that could influence students' academic performance based on their parent's combined education level. The **methodological approach** uses the program RStudio to conduct data cleaning, manipulation, representation, summarization, and visualisation using various packages such as ggplot2, tidyverse, and dplyr, on the data set named ‘Student Performance Data Set from ‘UCI Machine Learning Repository’. This CSV file tabulates “student achievement in secondary education of two Portuguese schools”1 (“UCI Machine Learning Repository: Student Performance Data Set,” 2014) in both mathematics and Portuguese subjects during 2014.

The main **findings** from this dataset show that as the PCEL increased, so did the student's final grade average for their respective PCEL groups, with two outliers. Students whose PCEL is ‘0’ (very low) outperformed students whose PCEL rankings are ‘2-4’ (low-moderate) and students whose PCEL is ‘1’ outperformed all other rankings, excluding students whose PCEL is ‘8’ (very high). Correlating these results to other variables within the data, we can observe that students whose parents' PCEL is ‘0’ received more family educational support (FES) and showed higher interest in continuing to higher education, and students whose parents' PCEL is ‘1’ received more school educational support (SES). Also, students whose PCEL were 8 ranked the highest overall in the ‘Final Grade Average’ because of consistently ranking between slightly above average and in the top percentile(majority of the time) in each variable comparison visualisation, correlating to student engagement in after-school activities, programs and allocated study time that contribute to their academic success.

From this report, we can therefore **conclude** that higher parent education levels generally correlate to higher students’ academic performance, with some consideration for possible outliers. A combination of the findings from this report and additional research could further allow advancement in the identification of factors that promotes the improvement of academic performance and, consequently, the development of effective schools and at-home educational support.

**II. Introduction**

A student's academic success can be contributed to a variety of factors: “personal factors, their interactions with others such as parents, teachers, and administrators, and lastly, the larger systems that surround the student e.g. school districts” 2 (Bertolini, Stremmel, & Thorngren, 2012). But where do the foundations of potential success start? As stated by Clearinghouse Technical Assist Team, a “child's behaviour is shaped by observation and direct learning experiences”3 (Parents’ Educational Levels Influence on Child Educational Outcomes: Rapid Literature Review, 2020) and in the beginning, a child's source of observation and direct learning experiences begins with their family, specifically parents, as they provide the foundation for a child's behaviours, social skills, habits, preferences etc. By parents modelling achievement-oriented behaviours such as completing higher education (Bachelor, Master, or PhD), as a child grows, they develop/ adopt a belief/ understanding that their family is achievement-oriented, and those values too can be expected from them.

From my own perspective, this has been solidified in my upbringing as both my parents, grandfather, uncle, and aunts have received higher education (PhD: mother, grandfather, and Master: father, uncles, and aunts) which, from a young age, has been expected from me as well. But does a parent's education level have a major contribution to their child's academic success throughout their academic career? If so, to what extent? Most of the research conducted on this topic yields favourable results to parents’ educational levels strongly influencing educational performance success and opportunities. Clearinghouse Technical Assist Team states that “the more educated the parent was when the child was 8 years old, the higher educational aspirations and attainment the child had at age 19.”3 (Parents’ Educational Levels Influence on Child Educational Outcomes: Rapid Literature Review, 2020).Within my family, my mother, grandfather, and wife are all educators and, having worked as an ESL educator for five years (almost becoming a high school educator) myself and now being a data scientist, I am curious and motivated to investigate if higher PCEL correlates to higher students’ academic performance based on the data provided.

**III. Data**

### The source of the data for this analysis was taken from the UCI Machine Learning Repository, named Student Performance which was created by Paul Cortez: Department of Information Systems, School of Engineering, University of Minho, Portugal.

### This observational study data was “collected using school reports and questionnaires.”1 (“UCI Machine Learning Repository: Student Performance Data Set,” 2014)

### The sample size (observations/number of participants) is 1044Appendix A

### There are 35 variables consisting of integers (17) and character types (18). Appendix A

### The UCI Machine Learning Repository website provides the zip file that shows pre-processing of two data sets ‘student-mat.csv’ and student-por.csv by merging them together into one data set. Appendix B

### Based on the premise of this investigation and data analysis for this report, a subset of the data frame has been created resulting in 12 variables being used and all 1044 observations. Below are the chosen variables based on relevance to the investigation:

### Variable: X, Type: Integer, Description: Student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira).

### Variable: schoolsup, Type: Character, Description: Extra educational support (binary: yes or no).

### Variable: famsup, Type: Character, Description: Family educational support (binary: yes or no).

### Variable: paid, Type: Character, Description: Extra paid classes (Math or Portuguese, binary: yes or no).

### Variable: activities, Type: Character, Description: Extra-curricular activities (binary: yes or no)

### Variable: studytime, Type: Integer, Description: Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours).

### Variable: higher, Type: Character, Description: Wants to take higher education (binary: yes or no).

### Variable: Medu, Type: Integer, Description: Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education).

### Variable: Fedu, Type: Integer, Description: Father’s education: Same as Mother’s education.

### Variable: G3, Type: Integer, Description: Final grade (numeric: from 0 to 20, output target).

### Variable: sub, Type: Character, Description: School Subject (binary: “Math” or Portuguese).

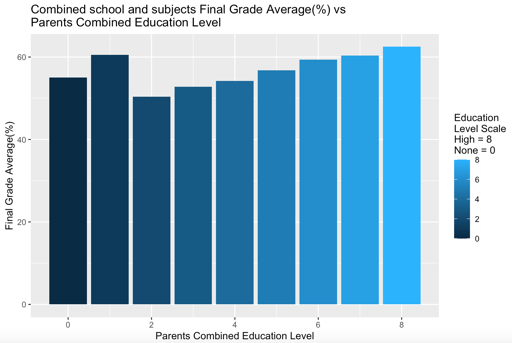
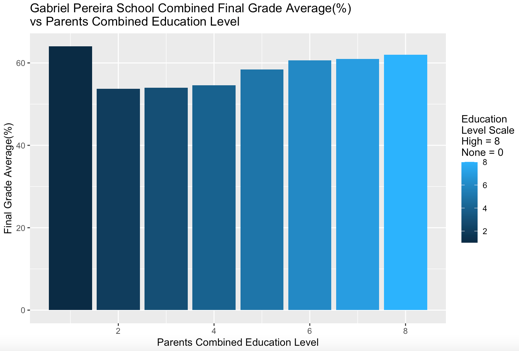
### The variables were renamed for a better description and the ‘Medu’ and ‘Fedu’ variables were combined into a new variable called ‘ParentCombinedEduRank’ (then added to the subset data frame) and ‘Medu’ and ‘Fedu’ were then removed from the subset. Appendix C

**IV. Methods**

* This report's data set was analyzed and coded using **RStudio** version 2022.7.0.548 with the release name Spotted Wakerobin. Appendix A
* ***Data representation:*** The dataset ‘student\_portmath.csv’ was imported using **read.csv()** and analyzed using the **str()** function: to determine the data set type, number of observations and variables, and variable types. The **summary ()** function provided each variable's minimum, 1st quantile, median, mean, 3rd quantile and maximum values, and **dim()** was used to verify the number of observations and variables. Appendix A The **cor( method = ‘spearman’)** function was used to determine the degree of correlation between the variable that was used in this report (the spearman correlation method was due to identified outliers. Spearman correlation ignores the magnitudes and uses solely the ranks of the values) Appendix D. The **colnames()** function was used to view the column names. Appendix E **View()** was used throughout the data analysis to review and ensure the data frame and variables created were correct.
* ***Data cleaning:*** The **rename()** function was used to chang**e** variable names for clearer variable descriptions. Appendix F **function(x) sum(is.na(x))** was created to count the number of missing values in a single column and combined with the **apply( , MARGIN = 2, FUN = Count\_miss\_val))** Appendix G  to identify the total missing values in each variable in the whole data set. No missing values were found. To determine the total number of duplicated rows, **sum(duplicated())** was used. No duplicated rows were found. Appendix G
* ***Type conversion:*** Character types were converted to factors using the **dataset[,variable] = lapply(dataset[, variable], factor)** function and **as.numeric()** then converted those factor into a numerical type. Appendix H
* ***Data subset selection and/or subsampling:*** The **subset()** function was used to create subsets with the desired variable from the original data frame and remove undesired/ no longer needed variables from the subset. Appendix I The **filter()** function was used to select rows desired rows within a variable for subsampling Appendix J
* ***Group-based data summarization:*** The **summarise(mean = )** and **group\_by ()** functions were used in calculating the ‘Final Grade(%)’ mean for each ‘ParentCombinedEduRank’ groups. Appendix K
* ***Variable selection and/or transformation:*** The **select =** was used inside the **subset()** function to select variables to be removed from the existing subset. Appendix D A new variable was created through **newvariable = variable1 + variable2**, in which two existing variables (‘MEdu and FEdu) rows were added together to create a new combined value for each row for the new variable ‘ParentCombinedEduRank’.Appendix C
* ***Exploratory visualisation using ggplot2:*** **geom\_boxplot** () was used to determine if outliers were present amongst the variables Appendix D, **ggcor()** Appendix D was used to plot the correlation between variables in the subset data, and **ggplot()** and **geom\_bar()**Appendix L were used to create visualizations between two variables.

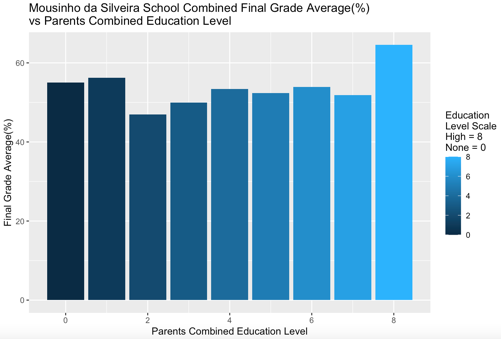
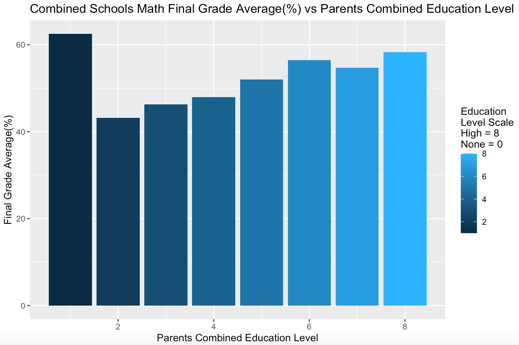
**V. Results and Discussion**

Chart, timeline, bar chart

Description automatically generated  

**Figure 1:** Correlation plot between variables. **Figure 2:** Barplot of combined school subject final **Figure 3:** Barplot of ‘GS’ school combined final grade

grade vs parents combined educational level. vs parents combined educational level.

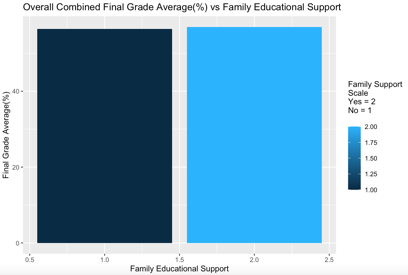
  Chart, bar chart

Description automatically generated

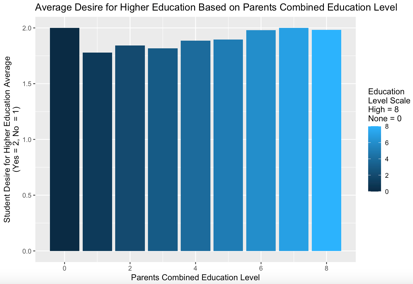
**Figure 4:** Barplot of ‘MS’ school combined final **Figure 5:** Barplot of math subject combined final **Figure 6:** Barplot of Portuguese subject combined

grade vs parents combined educational level. grade vs parents combined educational level. final grade vs parents combined educational level.

In figure 1 we can see a correlation plot of the variables used in data analysis for the purpose of this report, in which there is a high positive correlation between “Final Grade(%)” and “Higher”( 0.265759363)), and Final Grade(%)” and “ParentCombinedEduRank”( 0.234714795) compared to the other variables which indicate a strong relationship between them. In figure 2, a barplot visualisation depicts the student's final grade average for each parent's combined education level (PCEL) (variable description shows the parental education level explanation: 0 = no education background for both parents, 8 = higher education for both parents, within those parameters is a combination of PCEL using values 0-4). From our previous discussion in the introduction, we would expect the data to show lower grade averages for low students' PCEL and a gradual increase in final grade averages as students' PCEL increases. From bins 2 to 8 we can see this predicted pattern, but bins 0’ and ‘1’ are outliers. For students whose PCEL is ‘0’, we can see those students final grade average exceeds those students whose PCEL range from ‘2-4’ and students whose PCEL is ‘1’ outperform all except for the PCEL of ‘8’. Figure 2 is an overall comparison of the average of the final grades vs PCEL, and further investigation is conducted by sub-setting the data by schools and subjects to see if this trend is consistent. Figures 3-6 depict the same trend as figure 2 with some slight variation Appendix L and overall, we can conclude that students whose PCEL is ‘0’ and/ or ‘1’ received a final grade average greater than half, to most, of the other PCEL rankings in this data set. The remaining bins (3-8) in figures 2-6 generally conform to the expectations that as PCEL increases so do the student's final grade average as stated above. Students having a PCEL of 8 produced the highest final grade average except in figures 3 and 5. But the question remains, why are those students whose PCELs are 0 and 1 performing better than expected?

 Chart, bar chart

Description automatically generated Chart, bar chart, treemap chart

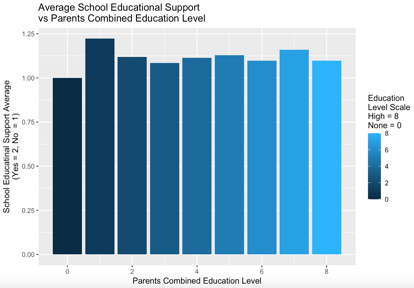
Description automatically generated 

**Figure 7:** Barplot of family educational **Figure 8:** Barplot of average family **Figure 9:** Barplot of seeking higher **Figure 10:** Barplot of desire for higher vs final grade average (%). education support vs PCEL education vs final grade average (%). education Average education vs PCEL.

In figure 7 we see the average number of students who receive family education support (FED) having a slightly higher final grade average than those who do not. Correlating this to figure 8 we can see that students whose PCEL are ‘0’ are all receiving FED (this is a binary question, where 1 = no and 2 = yes) because the average for this question is 2, which means they all answered yes, compared to others which are lower. Using this insight, we can conclude that for students, whose PCEL is ‘0’, academic success (within the parameters of this study) can be a result of family involvement in their education as “many studies have indicated strong positive correlations between parental involvement in their child’s learning and academic achievement”3 Reasons for their involvement is beyond the scope of this report. We can, however, speculate that these parents may see the value and importance of having an education as it may bring more opportunities for success, something the parents may not have had because they didn’t have an education. Other reasons may include “what parents believe they are supposed to do with respect to their children's education, their sense of efficacy for helping the child succeed in school defines how effective parents believe they can be in influencing their child's educational outcomes; and parents' perceptions of general invitations, opportunities, and demands for involvement from the school shape parents' beliefs about the school's expectations for their involvement.”4 (Reed et al., 2000).

Figure 9 depicts that those students who are seeking higher education scored a greater final grade average than those who are not. Observing figure 10 we can see that the student's PCEL of ‘0’ all want to continue to higher education; this is a binary question, where 1 = no and 2 = yes and in figure 10 we can see that for PCEL ‘0’ the average for this question is 2, which means they all answered yes for this question, compared to others which are lower. Combining the results above, we can provide reasoning for students whose PCEL is 0 having higher-than-expected final grade averages.

Chart, bar chart

Description automatically generated 

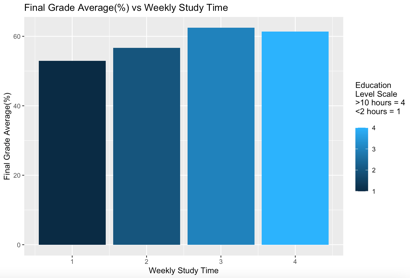
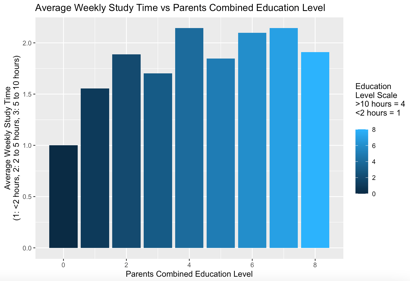
**Figure 11:** Barplot of school educational support **Figure 12:** Barplot of school educational support

vs Final Grade average(%). vs PCEL.

In figure 11 we can see that students who do not receive school educational support have a final grade average greater than those who do receive educational school support. “Educational Support provides opportunities for all students to gain access to, and fully participate in learning, in an inclusive environment.” 5 (“Educational Support | Townsville Grammar School,” n.d.). This includes things such as providing clothes, books and supplies, transportation assistance, tutoring, academic advising etc. With that said there are other variables to consider such as extracurricular activities (figures 13 and 14), and weekly study time (figures 15 and 16), within this data set that also contributes to higher grade achievement, which may contribute to the results in figure 11. Students may not need the school's educational support because of their family and living situations. Some students may not receive school educational support but, instead, they may receive academic support elsewhere. To specifically investigate why students whose PCEL are ‘1’ performed better than those whose PCELs are ‘0’ and ‘2-7’ we can refer to figure 12. Figure 12’s data shows that students whose PCEL is ‘1’ are receiving the most school educational support. In this instance, receiving this support could be correlated to why students whose PCEL is ‘1’ performed the second highest in the overall final grade average based on the data.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated  

**Figure 13:** Barplot of extra-curricular **Figure 14:** Barplot of average student **Figure 15:** Barplot of weekly study time **Figure 16:** Barplot of average weekly

activity participation vs final grade activity engagement vs PECL. vs Final Grade average(%). study time vs PECL.

average(%).

Excluding PCEL ‘0’ and ‘1’, as PCEL increases so do the averages for those measure variables on the y-axis (with some variation). These visualisations can then correlate to the expected pattern seen in figure 2, with PCEL ‘8’ ranking the highest because of consistently ranked slightly above average and in the top percentile in each variable comparison visualisation. With that said, within the data set, there are other variables that can further contribute to the correlation of students' final grade average based on their PCEL such as extra-curricular activities (Figure 13) and weekly study time (figure 15). This is mentioned to point out other factors that contribute to the overall results in figure 2, which correlates to the report's investigation into the question mentioned in the introduction (Figure 1 shows some of those variables' correlation strengths to the PCEL and can be used for further investigation). Aside from the variable discussed in this report, several other variables within the data set contribute to how a student succeeds academically such as students' family dynamics, living situations and social life that fall beyond the scope of this report's investigative analysis into the purposed question.

**VI. Conclusion**

In this analytical report, we investigated if higher parent combined education level (PCEL) correlated to higher student academic performance based on the Student Performance Data Set obtained from UCI Machine Learning Repository which tabulates the student achievement in secondary education of two Portuguese schools in both mathematics and Portuguese subjects. The report compared the overall (combining both subjects and schools) final grade average and found that as PCEL increased so did the final grade average with two exceptions. Students whose PCEL were ‘0’ had the sixth-highest average and students whose PCEL was ‘1’ had the second-highest average out of 9. Further investigation found that students whose PCEL was ‘0’ ranked higher than anticipated due to receiving family educational support as well as wanting to seek higher education which we can contribute to their final grade average ranking. Students whose PCEL was ‘1’ ranked highest for receiving school educational support which too can contribute to their final grade average ranking being higher than anticipated.

Within this data analysis’s variable correlation visualisations, for PCEL 2-8 we can see that as PCEL increases so do the average for those measured variables on the y-axis (with some variation) which contributes to the overall expected pattern seen in figure 2. Students whose PCEL was ‘8’ consistently ranked slightly above average and in the top percentile, which can correlate to students participating in a variety of beneficial activities, programs, and study habits to better their education, success, and final grade average. The intention of this study was not to generalize the results and findings to other populations; but rather, to further the understanding of students' success in school based on the variables and results provided. Despite the obtained results, we can go further to deepen the research carried out by incorporating new study variables and further investing existing variables that were not discussed in this report due to the nature of the report's question/ investigation, which may also have effects on students' success in school. Additional research could further allow advancement in the identification of factors that promotes the improvement of academic performance and, consequently, the development of effective schools and at-home educational support.

**VII. References**

1. UCI Machine Learning Repository: Student Performance Data Set. (2014). Retrieved from Uci.edu website: <https://archive.ics.uci.edu/ml/datasets/Student+Performance> on October 17th, 2022.
2. Bertolini, K., Stremmel, A., & Thorngren, J. (2012). *Human Sciences Department of Teaching, Learning and Leadership*. Retrieved from <https://files.eric.ed.gov/fulltext/ED568687.pdf> on October 18th 2022.
3. Clearinghouse for Military Family Readiness at Penn State. (2020). *Parents’ Educational Levels Influence on Child Educational Outcomes: Rapid Literature Review*. Accessed from: <https://militaryfamilies.psu.edu/wp-content/uploads/2020/01/Parents-Educational-Levels-Influence-on-Child-Educational-Outcomes.20Jan06.final_.pdf> on October 17th, 2022.
4. Reed, R. P., Jones, K. P., Walker, J. M., & Hoover-Dempsey, K. V. (2000, April 1). *Parents’ Motivations for Involvement in Children’s Education: Testing a Theoretical Model.* [Review of *Parents’ Motivations for Involvement in Children’s Education: Testing a Theoretical Model.*]. Retrieved from <https://files.eric.ed.gov/fulltext/ED444109.pdf> on October 17th, 2022.
5. Educational Support | Townsville Grammar School. (n.d.). Retrieved from <https://www.tgs.qld.edu.au/school/academic/secondary-school-curriculum/educational-support/#:~:text=Educational%20Support%20provides%20opportunities%20for> on October 17th, 2022.

**VIII. Appendices**

**Appendix A:** Checking RStudio version and examples of data representation in RStudio.

```{r}

RStudio.Version() *#Check RStudio type.*

#Load packages

install.packages('groupdata2')

library(reshape)

library(dplyr)

library(tidyr)

library(tidyverse)

library(cluster)

library(ggplot2)

library(GGally)

library(groupdata2)

*#Load document.*

student\_portmaths\_df= read.csv('student\_portmaths.csv') *#load students\_portmath.csv into R.*

str(student\_portmaths\_df) *#View structure of student: multivariate dataset (integer character).*

summary(student\_portmaths\_df) *#View summary of data frame.*

dim(student\_portmaths\_df) *#View dimension of data frame.*

```

**Appendix B:** Pre-processing zipfile provided top merge data sets in RStudio.

This script was provided in the zip file as pre-processing to merge the data sets/ csv files.

```{r}

d1=read.table("student-mat.csv",sep=";",header=TRUE)

d2=read.table("student-por.csv",sep=";",header=TRUE)

d3=merge(d1,d2,by=c("school","sex","age","address","famsize","Pstatus","Medu","Fedu","Mjob","Fjob","reason","nursery","internet"))

print(nrow(d3)) # 382 students

```

**Appendix C:** Creating a new variable by adding two variables values together and removing used ‘MEdu’ and ‘Fedu’ from the data set in RStudio. Variable Selection and/or transformation.

```{r}

*#Compute the sum of mother and father education rankings into one 'ParentCombinedEduRank' variable (9. Create a new variable through existing variable transformation).*

ParentCombinedEduLevel = student\_portmaths\_df\_subset$MEdu + student\_portmaths\_df\_subset$FEdu

*#Add variable to data subset.*

student\_portmaths\_df\_subset = student\_portmaths\_df\_subset %>% add\_column(ParentCombinedEduLevel)

*#Remove 'MEdu' and 'FEdu' as they are not needed because we have combined them into one variable (7. Data subset selection)*

student\_portmaths\_df\_subset = subset(student\_portmaths\_df\_subset, select = -c(MEdu, FEdu))

```

**Appendix D:** Removing variable from the subset, plot boxplot for outliers, conduct spearman correlation, plot ggcorr plot in RStudio. Data representation, Exploratory Visualisation using ggplot2, Data selection and/or subsampling.

Data Subsampling and Spearman correlation check between vairables in data set.

```{r}

*#Create a new data frame without `Student` in it as it is not needed for determining variable outliers because it serves as only identification and holds no real value (7. Data subset selection).*

subset\_missing\_student = subset(student\_portmaths\_df\_subset, select = -(Student))

boxplot(subset\_missing\_student) *#Use boxplot to determine if outliers in variables. Outliers are present in the data set, therefore methods proceeding will be chosen based on robustness and ability to mitigate outliers.*

*#Spearman correlation between variables (1.Data Representation)*

student\_portmaths\_corr= as.matrix(cor(student\_portmaths\_df\_subset, method = 'spearman'))

View(student\_portmaths\_corr) *#View correlation matrix.*

ggcorr(student\_portmaths\_corr) *#Plot correlation matrix.*

```

**Appendix E:** View column names in RStudio. Data representation.

```{r}

colnames(student\_portmaths\_df\_subset) *#Identify columns. In the original data set, there is no column name for 'Students'. "X" has been automatically added as the column name for students' numbers by R.*

```

**Appendix F:** Rename variable names and view changes in RStudio. Data cleaning

```{r}

*#Change column "X" name to "Student" and clean up other column names for organization and a better understanding of columns. (3. Data Cleaning).*

student\_portmaths\_df\_subset = rename(student\_portmaths\_df\_subset , c("X" = "Student", "school"="School", "schoolsup" = "SchoolSup", "famsup" = "FamSup", "paid" = "Paid", "activities" = "Activities", "studytime" = "StudyTime", "higher" = "Higher", "Medu" = "MEdu", "Fedu" = "FEdu", "G3" = "Final Grade(%)", "sub" = "Subject"))

colnames(student\_portmaths\_df\_subset) *#View column names to ensure the right changes have been made.*

```

**Appendix G:** Convert grading from marks from 0-20 to percentages. Identify missing values and duplicates in RStudio. Data cleaning

```{r}

*#Convert final grades to a percentage score. Scores are originally from 0 - 20. For better understanding, change it to percentage. (9. Variable Transformation).*

student\_portmaths\_df\_subset$`Final Grade(%)` = ((student\_portmaths\_df\_subset$`Final Grade(%)`/20)\*100)

View(student\_portmaths\_df\_subset) *#Check data set with column name change and variable transformation.*

*#Identify Missing variables (3. Data Cleaning)*

Count\_miss\_val = function(x) sum(is.na(x*)) #Create function to count the number of missing values in a single data column (variable)*

student\_miss\_vals = apply(student\_portmaths\_df\_subset, MARGIN = 2, FUN = Count\_miss\_val) *#look for NA in data set*

student\_miss\_vals *#Call function: 0 missing values for all variables in the data set.*

*#Identify if there are any duplicate rows (3.Data Cleaning*)

sum(duplicated(student\_portmaths\_df\_subset)) *# [1] 0. Zero duplicates found.*

```

**Appendix H:** Convert character types to factors and then to numeric types in RStudio. Type conversion.

```{r}

*#Convert character type variables to factors (4. Type conversion)*

convert\_to\_factors = c("School", "SchoolSup", "FamSup", "Paid", "Activities", "Higher","Subject") *#Create variable to hold columns that will be converted from characters to factors.*

student\_portmaths\_df\_subset[,convert\_to\_factors] = lapply(student\_portmaths\_df\_subset[,convert\_to\_factors], factor) *#Convert chosen character type variables to factors.*

*#Convert factors to numeric variables (4. Type conversion).*

student\_portmaths\_df\_subset$School = as.numeric(student\_portmaths\_df\_subset$School)

student\_portmaths\_df\_subset$SchoolSup = as.numeric(student\_portmaths\_df\_subset$SchoolSup)

student\_portmaths\_df\_subset$FamSup = as.numeric(student\_portmaths\_df\_subset$FamSup)

student\_portmaths\_df\_subset$Paid = as.numeric(student\_portmaths\_df\_subset$Paid)

student\_portmaths\_df\_subset$Activities = as.numeric(student\_portmaths\_df\_subset$Activities)

student\_portmaths\_df\_subset$Higher = as.numeric(student\_portmaths\_df\_subset$Higher)

student\_portmaths\_df\_subset$Subject = as.numeric(student\_portmaths\_df\_subset$Subject)

```

**Appendix I:** Create a subset from the original data set in RStudio. Data subsampling.

```{r}

*#Basis of the question for the report some variables not being needed. I have chosen the appropriate variables (based on metadata) to conduct the data analysis for the report and to also reduce 'noise' in the dataset.*

*#Create a subset of desired variables from the original data for report data analysis (7.Data subsampling).*

student\_portmaths\_df\_subset = subset(student\_portmaths\_df1, select = c(X,school, schoolsup, famsup, paid, activities, studytime, higher, Medu, Fedu, G3, sub))

```

**Appendix J:** Examples of filtering used for visualisations in RStudio. Data subsampling.

```{r}

*# Data Subset Selection (Gabriel Pereira School).*

PComEdu\_GP\_School = filter(student\_portmaths\_df\_subset, School == '1')

*# Data Subset Selection (Mousinho da Silveira School).*

PComEdu\_MS\_School = filter(student\_portmaths\_df\_subset, School == '2')

*#Data Subset Selection (Math Subject).*

PComEdu\_Math\_Subject = filter(student\_portmaths\_df\_subset, Subject == '1')

*#Data Subset Selection (Math Subject).*

PComEdu\_Portugese\_Subject = filter(student\_portmaths\_df\_subset, Subject == '2')

```

**Appendix K:** Examples of grouping variables and summarising using mean in RStudio. Group-based data summarization.

```{r}

# Calculate 'Final Grade (%)' Mean for each group in the 'ParentCombinedEduRank' variable, for barplot visualisation (8.Group-based data summarisation).

student\_PComEduRank\_grade\_means = student\_portmaths\_df\_subset %>% group\_by(`ParentCombinedEduLevel`) %>% summarise(Grade\_mean =mean(`Final Grade(%)`))

# Gabriel Pereira School students calculated 'Final Grade (%)' mean for each group in the 'ParentCombinedEduRank' variable, for barplot visualisation (8. Group-based data summarization).

GPStudent\_PComEduRank\_grade\_means = PComEdu\_GP\_School %>% group\_by(`ParentCombinedEduLevel`) %>% summarise(Grade\_mean =mean(`Final Grade(%)`))

```

Appendix L: Example of boxplots used in the report in RStudio. Exploratory visualisation using plot 2.

```{r}

*#Bar PLot (10. Exploratory visualisation using ggplot2).* *All schools*

*#PComEduRank vs Final Grade Average*

ggplot(student\_PComEduRank\_grade\_means) + geom\_bar(mapping = aes(x = `ParentCombinedEduLevel`, y = `Grade\_mean`, fill = `ParentCombinedEduLevel`), stat = 'identity') + xlab("Parents Combined Education Level") + ylab("Final Grade Average(%)") + ggtitle("Combined school and subjects Final Grade Average(%) vs\nParents Combined Education Level") + labs(fill = "Education\nLevel Scale\nHigh = 8\nNone = 0")

*#Barplot (10. Exploratory visualisation using ggplot2). GP School*

*#PComEdu vs Final Grade Average*

ggplot(GPStudent\_PComEduRank\_grade\_means) + geom\_bar(mapping = aes(x = `ParentCombinedEduLevel`, y = `Grade\_mean`, fill = `ParentCombinedEduLevel`), stat = 'identity') + xlab("Parents Combined Education Level") + ylab("Final Grade Average(%)") + ggtitle("Gabriel Pereira School Combined Final Grade Average(%)\nvs Parents Combined Education Level") + labs(fill = "Education\nLevel Scale\nHigh = 8\nNone = 0")

*#Barplot (10. Exploratory visualisation using ggplot2). MS School*

*#PComEdu vs Final Grade Average*

ggplot(MSStudent\_PComEduRank\_grade\_means) + geom\_bar(mapping = aes(x = `ParentCombinedEduLevel`, y = `Grade\_mean`, fill = `ParentCombinedEduLevel`), stat = 'identity') + xlab("Parents Combined Education Level") + ylab("Final Grade Average(%)") + ggtitle("Mousinho da Silveira School Combined Final Grade Average(%)\nvs Parents Combined Education Level") + labs(fill = "Education\nLevel Scale\nHigh = 8\nNone = 0")

*#Barplot (10. Exploratory visualisation using ggplot2). Math subject*

*# PComEdu vs Final Grade Average*

ggplot(Math\_PComEduRank\_grade\_means) + geom\_bar(mapping = aes(x = `ParentCombinedEduLevel`, y = `Grade\_mean`, fill = `ParentCombinedEduLevel`), stat = 'identity') + xlab("Parents Combined Education Level") + ylab("Final Grade Average(%)") + ggtitle("Combined Schools Math Final Grade Average(%) vs Parents Combined Education Level") + labs(fill = "Education\nLevel Scale\nHigh = 8\nNone = 0")

*#Barplot (10. Exploratory visualisation using ggplot2). Portugese*

*#PComEdu vs Final Grade Average*

ggplot(Portugese\_PComEduRank\_grade\_means) + geom\_bar(mapping = aes(x = `ParentCombinedEduLevel`, y = `Grade\_mean`, fill = `ParentCombinedEduLevel`), stat = 'identity') + xlab("Parents Combined Education Level") + ylab("Final Grade Average(%)") + ggtitle("Combined Schools Portugese Final Grade Average(%) vs\nParents Combined Education Level") + labs(fill = "Education\nLevel Scale\nHigh = 8\nNone = 0")

```