PROGRAMMING FOR DATA ANALYSIS

Technology Park Malaysia

CT127-3-2-PFDA

Individual Assignment

Name: PREMSHARAAN A/L SELVA (TP056561)

Lecturer: MINNU HELEN JOSEPH

Module Code: CT127-3-2-PFDA (PROGRAMMING FOR DATA ANALYSIS)

Intake Code: APD2F2202CS(CYB)

Hand-out-Date: 28th MARCH 2022

Hand-in-Date: 9th MAY 2022

Table of Contents

Table of Contents **2**

Table of Figures **6**

1.0 Introduction **15**

2.0 Assumptions **16**

3.0 Data Importing, Cleaning, Pre-processing and Transformation **17**

3.1 Data Importing **17**

3.1.1 Installing & Importing Packages 17

3.1.2 Importing Data from CSV file 18

3.2 Data Pre-processing & Data Cleaning **20**

3.3 Data Transformation **22**

3.4 Data Exploration **25**

4.0 Questions & Analyses **27**

4.1 Question 1: How do personal relationships impact a student's grades? **27**

4.1.1 Analysis 1-1: Finding the correlation between students' romantic relationship status and their average grades. 27

4.1.2 Analysis 1-2: Finding the correlation between students' going on outings with their friends and their average grades. 34

4.1.3 Analysis 1-3: Finding the relationship between the quality of the student's family relationships and the average grades. 41

4.1.4 Analysis 1-4: Finding the relationship between the students' guardian and their average marks 46

4.1.5 Analysis 1-5: Finding the relationship between students' romantic relationship status, their study time and their average marks 51

4.2 Question 2: How does time management impact the students' grades? **54**

4.2.1 Analysis 2-1: Finding the correlation between students' study time and their average grade. 54

4.2.2 Analysis 2-2: Finding the relationship between students' free time and their average grades 59

4.2.3 Analysis 2-3: Finding the correlation between students' extra-curricular activities participation and their average grades 64

4.2.4 Analysis 2-4: -Finding the correlation between students' study time, free time and their average grades. 69

4.3 Question 3: How do resource availability and the comfort impact students' marks? 73

4.3.1 Analysis 3-1: Finding the relationship between students' joining additional paid classes and their average grades 73

4.3.2 Analysis 3-2: Finding the relationship between students' additional educational support from school and their average grades 80

4.3.3 Analysis 3-3: Finding the relationship between students' family educational support and their average marks. 87

4.3.4 Analysis 3-4: Finding the correlation between the students' travel time to the class and their average grade. 93

4.3.5 Analysis 3-5: Finding the correlation between students' Internet access and their average marks. 98

4.4 Question 4: How does students' family influence impact their marks? 104

4.4.1 Analysis 4-1: Finding the relationship between family size and students' average grades. 104

4.4.2 Analysis 4-2: Finding the correlation between parents' cohabitation status and students' average grades. 109

4.4.3 Analysis 4-3: Finding the relationship between a mother's education and a student's average mark. 112

4.4.4 Analysis 4-4: Finding the relationship between a father's education and a student's average mark. 116

4.4.5 Analysis 4-5: Finding the correlation between the quality of the mother's education level, the father's education level, and the student's average marks. 120

4.4.6 Analysis 4-6: Finding the correlation between mother's job, father's job and students' average marks. 124

4.5 Question 5: How do students' personal lives impact their marks? 128

4.5.1 Analysis 5-1: Finding the relationship between students' workday alcohol consumption and their average marks. 128

4.5.2 Analysis 5-2: Finding the relationship between students' weekend alcohol consumption and their average marks. 132

4.5.3 Analysis 5-3: Finding the correlation between students' decision to pursue higher studies and their average marks. 136

4.5.4 Analysis 5-4: Finding the relationship between students' health status and their average marks 140

4.5.5 Analysis 5-5: Finding the relationship between students' nursery school attendance and their average grade. 144

4.5.6 Analysis 5-6: Finding the relationship between students' number of absences and their average grade. 148

4.5.7 Analysis 5-7: Finding the relationship between students' past class failures and their average mark. 152

4.5.8 Analysis 5-8: Finding the relationship between students' school and their average mark. 156

4.5.9 Analysis 5-9: Finding the relationship between students' sex and their average mark. 160

4.5.10 Analysis 5-10: Finding the relationship between students' address type and their average mark. 165

5.0 Extra Features 170

5.1 Extra Feature 1: Use of Tidyverse Package **..... 170**

5.2 Extra Feature 2: Treemap **..... 171**

5.3 Extra Feature 3: Arranging Graphs Together **..... 173**

5.4 Extra Feature 4: Displaying Percentage and Counts on Pie Charts **..... 175**

5.5 Extra Feature 5: Assigning Different Colours for Bar Graphs **..... 177**

5.6 Extra Feature 6: Adding new columns **179**

6.0 Conclusion **181**

7.0 References **182**

Table of Figures

Figure 1 first 10 rows of students' data (half of the columns). 15
Figure 2 first 10 rows of students' data (rest of the columns). 15
Figure 3 shows the code to install additional packages. 17
Figure 4 shows the code to load the packages. 18
Figure 5 shows the code of importing the data from the external CSV file. 18
Figure 6 shows the **dsap_data** variable on the environment view. 19
Figure 7 shows the **dsap_data** variable executed. 19
Figure 8 shows the displayed dataset on the console. 19
Figure 9 shows the code for displaying dataset structure. 20
Figure 10 shows the data structure of the dataset. 20
Figure 11 shows the code to check null data in dataset. 21
Figure 12 shows the result of null data in dataset. 21
Figure 13 shows the code to check the duplicated data. 22
Figure 14 shows the result of duplicated data. 22
Figure 15 shows the code to transform categorical data into numeric value. 22
Figure 16 shows the output of the categorical data after executing the code. 23
Figure 17 shows the code that calculating average mark. 23
Figure 18 shows the output of calculating average mark. 23
Figure 19 shows the code that calculating average mark range. 24
Figure 20 shows the output of calculating average mark range. 24
Figure 21 shows the code that calculating overall average mean mark. 24
Figure 22 shows the output of overall average mean mark. 25
Figure 24 shows the code to view the first 10 rows from the dataset. 25
Figure 26 shows the output of the first 10 rows from the dataset. (1) 25
Figure 27 shows the output the first 10 rows from the dataset. (2) 25
Figure 24 shows the output of the first 10 rows from the dataset. (3) 26
Figure 23 shows the code to summarise all the data. 26
Figure 24 shows the output of summarisation all the data. 26
Figure 29 shows the R code used to create the data visualization figure of **Q1A1V1**. 27
Figure 30 shows the R code used to create the data visualization figure of **Q1A1V2.** 28
Figure 31 shows the R code used to create the data visualization figure of **Q1A1V3**. 29

Figure 32 shows the R code used to arrange the data visualization figures of **Q1A1V2** and **Q1A1V3** in one view. 29
Figure 33 shows the output of the **Q1A1R1** variable. 29
Figure 34 shows the bar graph of the output of Q1A1V1. 30
Figure 35 shows the output of the **Q1A1R2**. 30
Figure 36 shows of the output of **Q1A1R3**. 31
Figure 37 shows the pie charts of the output of **Q1A1V2** and **Q1A1V3**. 31
Figure 38 shows the R code used to create the data visualization figure of **Q1A2V1**..... 34
Figure 39 shows the R code used to create the data visualization figure of **Q1A2V**2..... 34
Figure 40 shows the R code used to create the data visualization figure of **Q1A2V3**..... 35
Figure 41 shows the R code used to arrange the data visualization figures of **Q1A2V2** and **Q1A2V3** in one view. 35
Figure 42 shows the output of the **Q1A2R1**. 36
Figure 43 shows the stacked bar graph of output of the **Q1A2V1**. 36
Figure 44 shows the output of the **Q1A2R2**. 37
Figure 45 shows the output of the **Q1A2R3**. 37
Figure 46 shows the pie charts of output of the **Q1A2V2** and **Q1A2V3**. 38
Figure 47 shows the R code used to create the data visualization figure of **Q1A3V1**..... 41
Figure 48 shows the R code used to create the data visualization figure of **Q1A3V2**..... 41
Figure 49 shows the R code used to create the data visualization figure of **Q1A3V3**..... 42
Figure 50 shows the R code used to arrange the data visualization figures of **Q1A3V2** and **Q1A3V3** in one view. 42
Figure 51 shows the output of the **Q1A3R1**. 42
Figure 52 shows the bar graph output of the **Q1A3V1**. 43
Figure 53 shows the output of the **Q1A3R2**. 43
Figure 54 shows the output of the **Q1A3R3**. 44
Figure 55 shows the pie charts of the output of **Q1A3V2** and **Q1A3V3**. 44
Figure 56 shows the R code used to create the data visualization figure of **Q1A4V1**..... 46
Figure 57 shows the R code used to create the data visualization figure of **Q1A4V2**..... 46
Figure 58 shows the R code used to create the data visualization figure of **Q1A4V3**..... 46
Figure 59 shows the R code used to arrange the data visualization figures of **Q1A4V2** and **Q1A4V3** in one view. 47
Figure 60 shows the output of the **Q1A4R1**. 47
Figure 61 shows the bar graph output of the **Q1A4V1**. 48

Figure 62 shows the output of the **Q1A4R2**..... 48
Figure 63 shows the output of the **Q1A4R3**..... 49
Figure 64 shows the pie charts of the output of **Q1A3V2** and **Q1A3V3**..... 49
Figure 65 shows the R code used to create the data visualization figure of **Q1A5V1**..... 51
Figure 66 shows the output of the **Q1A5R1**..... 51
Figure 67 shows the stacked bar graph output of the **Q1A5V1**..... 52
Figure 68 shows the R code used to create the data visualization figure of **Q2A1V1**..... 54
Figure 69 shows the R code used to create the data visualization figure of **Q2A1V2**..... 54
Figure 70 shows the R code used to create the data visualization figure of **Q2A1V3**..... 54
Figure 71 shows the R code used to arrange the data visualization figures of **Q2A1V2** and **Q2A1V3** in one view. 55
Figure 72 shows the output of the **Q2A1R1**..... 55
Figure 73 shows the horizontal bar graph output of the **Q2A1V1**..... 56
Figure 74 shows the output of the **Q2A1R2**..... 56
Figure 75 shows the output of the **Q2A1R3**..... 57
Figure 76 shows the point-to-point graph output of the **Q2A1V2** and **Q2A1V3**..... 57
Figure 77 shows the R code used to create the data visualization figure of **Q2A2V1** 59
Figure 78 shows the R code used to create the data visualization figure of **Q2A2V2**..... 59
Figure 79 shows the R code used to create the data visualization figure of **Q2A2V3**..... 59
Figure 80 shows the R code used to arrange the data visualization figures of **Q2A2V2** and **Q2A2V3** in one view. 59
Figure 81 shows the output of the **Q2A2R1**..... 60
Figure 82 shows the stacked bar graph output of the **Q2A2V1**..... 61
Figure 83 shows the output of the **Q2A2R2**..... 61
Figure 84 shows the output of the **Q2A2R3**..... 62
Figure 85 shows the scatter plot graph output of the **Q2A2V2** and **Q2A2V3**. 62
Figure 86 shows the R code used to create the data visualization figure of **Q2A3V1** 64
Figure 87 shows the R code used to create the data visualization figure of **Q2A3V2** 64
Figure 88 shows the R code used to create the data visualization figure of **Q2A3V3** 64
Figure 89 shows the R code used to arrange the data visualization figures of **Q2A3V2** and **Q2A3V3** in one view. 64
Figure 90 shows the output of the **Q2A3R1**..... 65
Figure 91 shows the bar graph output of the **Q2A3V1**..... 65
Figure 92 shows the output of the **Q2A3R2**..... 66

Figure 93 shows the output of the **Q2A3R3**. 66
Figure 94 shows the scatter plot graph output of the **Q2A3V2** and **Q2A3V3**. 67
Figure 95 shows the R code used to create the data visualization figure of **Q2A4V1.** 69
Figure 96 shows the R code used to create the data visualization table figure of **Q2A4R2** 69
Figure 97 shows the output of the **Q2A4R1**. 70
Figure 98 shows the bar graph output of the **Q2A4V1.** 70
Figure 99 shows the table output of the **Q2A4R2.** 71
Figure 100 shows the R code used to create the data visualization figure of **Q3A1V1.** 73
Figure 101 shows the R code used to create the data visualization figure of **Q3A1V2.** 73
Figure 102 shows the R code used to create the data visualization figure of **Q3A1V3.** 73
Figure 103 shows the R code used to arrange the data visualization figures of **Q3A1V2** and **Q3A1V3** in one view. 74
Figure 104 shows the R code used to create the data visualization figure of **Q3A1V4.** 74
Figure 105 shows the R code used to create the data visualization figure of **Q3A1V5.** 74
Figure 106 shows the R code used to arrange the data visualization figures of **Q3A1V4** and **Q3A1V5** in one view. 74
Figure 107 shows the output of the **Q3A1R1**. 75
Figure 108 shows the horizontal bar graph output of the **Q3A1V1**. 75
Figure 109 shows the output of the **Q3A1R2**. 76
Figure 110 shows the output of the **Q3A1R3**. 76
Figure 111 shows the three map charts of the output of **Q3A1V2** and **Q1A1V3**. 76
Figure 112 shows the output of the **Q3A1R4**. 77
Figure 113 shows the output of the **Q3A1R5**. 77
Figure 114 shows the pie charts of the output of **Q3A1V4** and **Q1A1V5**. 78
Figure 115 shows the R code used to create the data visualization figure of **Q3A2V1.** 80
Figure 116 shows the R code used to create the data visualization figure of **Q3A2V2.** 80
Figure 117 shows the R code used to create the data visualization figure of **Q3A2V3.** 80
Figure 100 shows the R code used to create the data visualization figure of **Q3A2V4.** 81
Figure 106 shows the R code used to arrange the data visualization figures of **Q3A2V2** **Q3A2V3** and **Q3A2V4** in one view. 81
Figure 120 shows the R code used to create the data visualization table figure of **Q3A2R5.** 81
Figure 121 shows the output of the **Q3A2R1.** 82
Figure 114 shows the scatterplot graph of the output of **Q3A2V1**. 82
Figure 123 shows the output of the **Q3A2R2.** 83

Figure 124 shows the output of the **Q3A2R2.** 83
Figure 125 shows the output of the **Q3A2R4.** 83
Figure 114 shows the treemap graph of the output of **Q3A2V2, Q3A2V3** and **Q3A2V4**..... 84
Figure 127 shows the table output of the **Q3A2R5.** 84
Figure 128 shows the R code used to create the data visualization figure of **Q3A3V1.**..... 87
Figure 129 shows the R code used to create the data visualization figure of **Q3A3V2.**..... 87
Figure 130 shows the R code used to create the data visualization figure of **Q3A3V3.**..... 87
Figure 131 shows the output of the **Q3A3R1.** 88
Figure 132 shows the horizontal bar graph output of the **Q3A3V1**..... 88
Figure 133 shows the output of the **Q3A3R2.** 89
Figure 134 shows the bar graph output of the **Q3A3V2**. 89
Figure 135 shows the output of the **Q3A3R3.** 90
Figure 136 shows the bar graph output of the **Q3A3V3.** 91
Figure 137 shows the R code used to create the data visualization figure of **Q3A4V1.**..... 93
Figure 138 shows the R code used to create the data visualization figure of **Q3A4V2.**..... 93
Figure 139 shows the R code used to create the data visualization figure of **Q3A4V3.**..... 93
Figure 106 shows the R code used to arrange the data visualization figures of **Q3A4V2** and **Q3A4V3** in one view. 93
Figure 141 shows the output of the **Q3A4R1.** 94
Figure 142 shows the stacked bar graph output of the **Q3A4V1**..... 95
Figure 143 shows the output of the **Q3A4R2.** 95
Figure 144 shows the output of the **Q3A4R3.** 96
Figure 145 shows the treemap graphs of the output of **Q3A4V2** and **Q3A4V3**. 96
Figure 146 shows the R code used to create the data visualization figure of **Q3A5V1.**..... 98
Figure 147 shows the R code used to create the data visualization figure of **Q3A5V2.**..... 98
Figure 148 shows the R code used to create the data visualization figure of **Q3A5V3.**..... 98
Figure 149 shows the R code used to arrange the data visualization figures of **Q3A5V2** and **Q3A5V3** in one view. 99
Figure 150 shows the output of the **Q3A5R1.** 99
Figure 151 shows the bar graph output of the **Q3A5V1**. 100
Figure 152 shows the output of the **Q3A5R2.** 101
Figure 153 shows the output of the **Q3A5R3.** 101
Figure 154 shows the treemap graphs of the output of **Q3A5V2** and **Q3A5V3**. 101
Figure 155 shows the R code used to create the data visualization figure of **Q4A1V1.**..... 104

Figure 156 shows the R code used to create the data visualization figure of **Q4A1V2.**..... 104
Figure 157 shows the R code used to create the data visualization figure of **Q4A1V3.**..... 104
Figure 158 shows the R code used to arrange the data visualization figures of **Q4A1V2** and **Q4A1V3** in one view. 105
Figure 159 shows the output of the **Q4A1R1.**..... 105
Figure 160 shows the horizontal bar graph output of the **Q4A1V1**..... 106
Figure 161 shows the output of the **Q4A1R2.**..... 106
Figure 162 shows the output of the **Q4A1R3.**..... 107
Figure 163 shows the stacked bar graphs of the output of **Q4A1V2** and **Q4A1V3**. 107
Figure 164 shows the R code used to create the data visualization figure of **Q4A2V1.**..... 109
Figure 161 shows the output of the **Q4A2R1.**..... 109
Figure 166 shows the horizontal bar graph output of the **Q4A2V1**..... 110
Figure 167 shows the R code used to create the data visualization figure of **Q4A3V1.**..... 112
Figure 168 shows the R code used to create the data visualization figure of **Q4A3V2.**..... 112
Figure 169 shows the output of the **Q4A3R1.**..... 113
Figure 170 shows the bar graph output of the **Q4A3V1**. 113
Figure 171 shows the output of the **Q4A3R2.**..... 114
Figure 172 shows the pie chart output of the **Q4A3V2**. 114
Figure 173 shows the R code used to create the data visualization figure of **Q4A4V1.**..... 116
Figure 174 shows the R code used to create the data visualization figure of **Q4A4V2.**..... 116
Figure 175 shows the output of the **Q4A4R1.**..... 117
Figure 176 shows the bar graph output of the **Q4A4V1**. 117
Figure 177 shows the output of the **Q4A4R2.**..... 118
Figure 178 shows the pie chart output of the **Q4A4V2**. 118
Figure 179 shows the R code used to create the data visualization figure of **Q4A5V1.**..... 120
Figure 180 shows the R code used to create the data visualization figure of **Q4A5R2.**..... 120
Figure 181 shows the output of the **Q4A5R1.**..... 121
Figure 182 shows the bar graph output of the **Q4A5V1**. 121
Figure 183 shows the table output of the **Q4A5R2**. 122
Figure 184 shows the R code used to create the data visualization figure of **Q4A6V1.**..... 124
Figure 185 shows the output of the **Q4A6R1.**..... 124
Figure 186 shows the bar graph output of the **Q4A6V1**. 125
Figure 187 shows the R code used to create the data visualization figure of **Q5A1V1.**..... 128
Figure 188 shows the R code used to create the data visualization figure of **Q5A1V2.**..... 128

Figure 189 shows the output of the **Q5A1R1.** 129
Figure 190 shows the horizontal stacked bar graph output of the **Q5A1V1**. 129
Figure 191 shows the output of the **Q5A1R1.** 130
Figure 192 shows the stacked bar graph output of the **Q5A1V2**..... 130
Figure 193 shows the R code used to create the data visualization figure of **Q5A2V1.**..... 132
Figure 194 shows the R code used to create the data visualization figure of **Q5A2V2.**..... 132
Figure 195 shows the output of the **Q5A2R1.** 133
Figure 196 shows the stacked bar graph output of the **Q5A2V1**..... 133
Figure 197 shows the output of the **Q5A2R2.** 134
Figure 198 shows the stacked bar graph output of the **Q5A2V2**..... 134
Figure 199 shows the R code used to create the data visualization figure of **Q5A3V1.**..... 136
Figure 200 shows the R code used to create the data visualization figure of **Q5A3V2.**..... 136
Figure 201 shows the output of the **Q5A3R1.** 137
Figure 202 shows the bar graph output of the **Q5A3V1**. 137
Figure 203 shows the output of the **Q5A3R2.** 138
Figure 204 shows the treemap graph output of the **Q5A3V2**..... 138
Figure 205 shows the R code used to create the data visualization figure of **Q5A4V1.**..... 140
Figure 206 shows the R code used to create the data visualization figure of **Q5A4V2.**..... 140
Figure 207 shows the output of the **Q5A4R1.** 141
Figure 208 shows the stacked bar graph output of the **Q5A4V1**..... 141
Figure 209 shows the output of the **Q5A4R2.** 142
Figure 210 shows the treemap graph output of the **Q5A4V2**..... 142
Figure 211 shows the R code used to create the data visualization figure of **Q5A5V1.**..... 144
Figure 212 shows the R code used to create the data visualization figure of **Q5A5V2.**..... 144
Figure 213 shows the output of the **Q5A5R1.** 145
Figure 214 shows the horizontal bar graph output of the **Q5A5V1**..... 145
Figure 215 shows the output of the **Q5A5R2.** 146
Figure 216 shows the pie chart output of the **Q5A5V2**. 146
Figure 217 shows the R code used to create the data visualization figure of **Q5A6V1.**..... 148
Figure 218 shows the R code used to create the data visualization figure of **Q5A6V2.**..... 148
Figure 219 shows the output of the **Q5A6R1.** 149
Figure 220 shows the horizontal bar graph output of the **Q5A6V1**..... 149
Figure 221 shows the output of the **Q5A6R1.**..... 150
Figure 222 shows the horizontal bar graph output of the **Q5A6V2**..... 150

Figure 223 shows the R code used to create the data visualization figure of **Q5A7V1.**..... 152
Figure 224 shows the R code used to create the data visualization figure of **Q5A7V2.**..... 152
Figure 225 shows the output of the **Q5A7R1.**..... 152
Figure 226 shows the stacked bar graph output of the **Q5A7V1**..... 153
Figure 227 shows the output of the **Q5A7R2.**..... 153
Figure 228 shows the stacked bar graph output of the **Q5A7V2**..... 154
Figure 229 shows the R code used to create the data visualization figure of **Q5A8V1.**..... 156
Figure 230 shows the R code used to create the data visualization figure of **Q5A8V2.**..... 156
Figure 231 shows the output of the **Q5A8R1.**..... 157
Figure 232 shows the bar graph output of the **Q5A8V1**. 157
Figure 233 shows the output of the **Q5A8R2.**..... 158
Figure 234 shows the output of the **Q5A8R2.**..... 158
Figure 235 shows the R code used to create the data visualization figure of **Q5A9V1.**..... 160
Figure 236 shows the R code used to create the data visualization figure of **Q5A9V2.**..... 160
Figure 237 shows the R code used to create the data visualization figure of **Q5A9V3.**..... 160
Figure 238 shows the R code used to arrange the data visualization figures of **Q5A9V2** and **Q5A9V3** in one view. 160
Figure 239 shows the output of the **Q5A9R1.**..... 161
Figure 240 shows the bar graph output of the **Q5A9V1**. 161
Figure 241 shows the output of the **Q5A9R2.**..... 162
Figure 242 shows the output of the **Q4A9R3.**..... 162
Figure 243 shows the stacked bar graphs of the output of **Q5A9V2** and **Q5A9V3**. 163
Figure 244 shows the R code used to create the data visualization figure of **Q5A10V1.**..... 165
Figure 245 shows the R code used to create the data visualization figure of **Q5A10V2.**..... 165
Figure 246 shows the output of the **Q5A10R1.**..... 166
Figure 247 shows the bar graph output of the **Q5A10V1**. 166
Figure 248 shows the output of the **Q5A10R2.**..... 167
Figure 249 shows the output of the **Q5A10R2.**..... 167
Figure 250 shows the code for treeemap function. 170
Figure 251 shows the code for treeemap function. 170
Figure 252 shows the code for treeemap function. 171
Figure 253 shows the example output for treeemap function..... 171
Figure 254 shows the code for arrange function..... 173
Figure 255 shows the example output for arrange function. 173

Figure 256 shows the code for calculate and display percentage and total student counts.. 175
Figure 257 shows the output of the percentage and student counts of the pie chart..... 175
Figure 258 shows the code for displaying different colours..... 177
Figure 259 shows the output of displaying different colours. 177
Figure 260 shows the code for creating new columns into the dataset. 179
Figure 261 shows the output the created columns. 179

1.0 Introduction

In this dataset delivered, the data represents the three-year final scores of degree students' marks data from Portugal, which has a total of 33 columns of the student's personal information, family backgrounds, daily routines, academic records and three-year final grades for their math tests. The dataset has details of 992 students in total. The primary purpose of this assignment is to analyse and identify how these data factors affect the students' educational performance. During this data analysis, multiple techniques will be utilized to analyse the students' data factors that affect their studies, which include data pre-processing, data exploration, data transformation, data manipulation, and data visualization. For this data analysis, it will be conducted using the R programming by making use of its numerous functions and features to draw the information.

index	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activitie
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes	no
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	yes	yes
5	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	yes	no
6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	no	yes	yes	yes
7	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no
8	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no
9	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	yes	no
10	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes	yes	yes

Figure 1 *first 10 rows of students' data (half of the columns).*

nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
yes	yes	no	no	4	3	4	1	1	3	6	5	6	6
no	yes	yes	no	5	3	3	1	1	3	4	5	5	6
yes	yes	yes	no	4	3	2	2	3	3	10	7	8	10
yes	yes	yes	yes	3	2	2	1	1	5	2	15	14	15
yes	yes	no	no	4	3	2	1	2	5	4	6	10	10
yes	yes	yes	no	5	4	2	1	2	5	10	15	15	15
yes	yes	yes	no	4	4	4	1	1	3	0	12	12	11
yes	yes	no	no	4	1	4	1	1	1	6	6	5	6
yes	yes	yes	no	4	2	2	1	1	1	0	16	18	19
yes	yes	yes	no	5	5	1	1	1	5	0	14	15	15

Figure 2 *first 10 rows of students' data (rest of the columns).*

2.0 Assumptions

From this given dataset, it is assumed that some of the data factors of students have a significant impact on their study grades. As such, a fair amount of data analyses that have been carried out in order to identify the correlation between these data attributes and how they impact the students' marks. At end of this data analysis, a conclusion will be made to understand the data attributes that are affecting the students' marks.

3.0 Data Importing, Cleaning, Pre-processing and Transformation

3.1 Data Importing

3.1.1 Installing & Importing Packages

Before beginning with the data analysis, several extra packages that are needed to be installed and loaded first in the RStudio. These additional packages will be utilized in the future for manipulating and visualizing the data from the dataset. When discussing about packages, a package contains a set of useful functions in a single unit, and it also got its own sample data and compiled code in it which will be stored in a directory called the library in the R environment. With the use of the readily available packages in R, it is very feasible for conducting the data analysis which includes the data exploration, manipulation, transformation, processing and visualization.

There was a total of 7 packages that have been installed and loaded for this data analysis, as shown in the figures above. Those packages are named tidyverse, plotrix, ggpubr, ggthemes, threemapify, grid and gridExtra. The core tidyverse package contains a collection of other function packages that will be put in use usually whenever conducting the data analysis that includes the ggplot2, dplyr, readr, tidyr, tibble and purr, stringr, and forcats. Some of the function packages that will be essentially used from this tidyverse are the ggplot2, and dplyr. The ggplot2 function from the tidyverse will be used for data visualization to plot all kinds of graphs and the dplyr is used to provide a grammar for data manipulation. Moreover, the plotrix package has a fair amount of plotting tools, such as labelling, axis, and colour scaling functions. The ggpubr packages are used to group and display multiple graphs that have been drawn into a single view. The ggthemes give additional themes, scales and geoms to plot the graph. Furthermore, the threemapify package is used to plot three maps for the data analysis. The grid is utilized to produce graphical output in form of a table to view better relationships between the result data. Finally, the gridExtra packages are used for arranging plots in a grid layout.

```
#====Installing various packages====#
install.packages("tidyverse")
install.packages("plotrix")
install.packages("ggpubr")
install.packages("ggthemes")
install.packages("treemapify")
install.packages("grid")
install.packages("gridExtra")
```

Figure 3 *shows the code to install additional packages.*

In order to import these packages, the first step is to use the "install. packages()" function to install the respective packages, as shown in figure 3 above. After successfully installing the required packages, the packages required to be loaded into our RStudio environment so that they can be used for the data analysis later on. In order to load the packages, use the "library()" function, as shown in figure 4 below. Now the packages are ready to be put into use for the data analysis.

```
#====Load the installed packages====#
library(tidyverse) #ggplot2, dplyr, readr, tidyr, tibble
library(plotrix)
library(ggpubr)
library(ggthemes)
library(treemapify)
library(grid)
library(gridExtra)
```

Figure 4 *shows the code to load the packages.*

3.1.2 Importing Data from CSV file

```
#=======Data Importing========#
#========Import the dataset========#
#Importing the dataset from CSV file while naming the object to dsap_data
#dsap stands for Degree Students Academic Performance
dsap_data = read.csv("C:\\Users\\User\\OneDrive - Asia Pacific University\\APU degree studies\\Sem1\\PFDA\\Assignment\\student.csv")
```

Figure 5 *shows the code of importing the data from the external CSV file.*

As shown in figure 5 above, to import all the data supplied from the CSV file, which in this case is the **student.csv** file to the global environment, the "read.csv()" function will be used. It will load the data file and outputs its contents on the console, but the dataset will not be stored in memory. To solve that, it has been assigned as a data frame using the variable called **dsap_data.** The variable **dsap_data** will be used throughout this entire data analysis coding.

When the importing process was completed successfully, the result will be shown in the environment tab, as shown in figure 6 below. The imported data can be viewed by simply typing and running the **dsap_data** variable, which will show all the data on the console as shown in figure 8 below.

Figure 6 *shows the* ***dsap_data*** *variable on the environment view.*

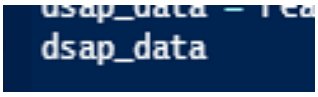
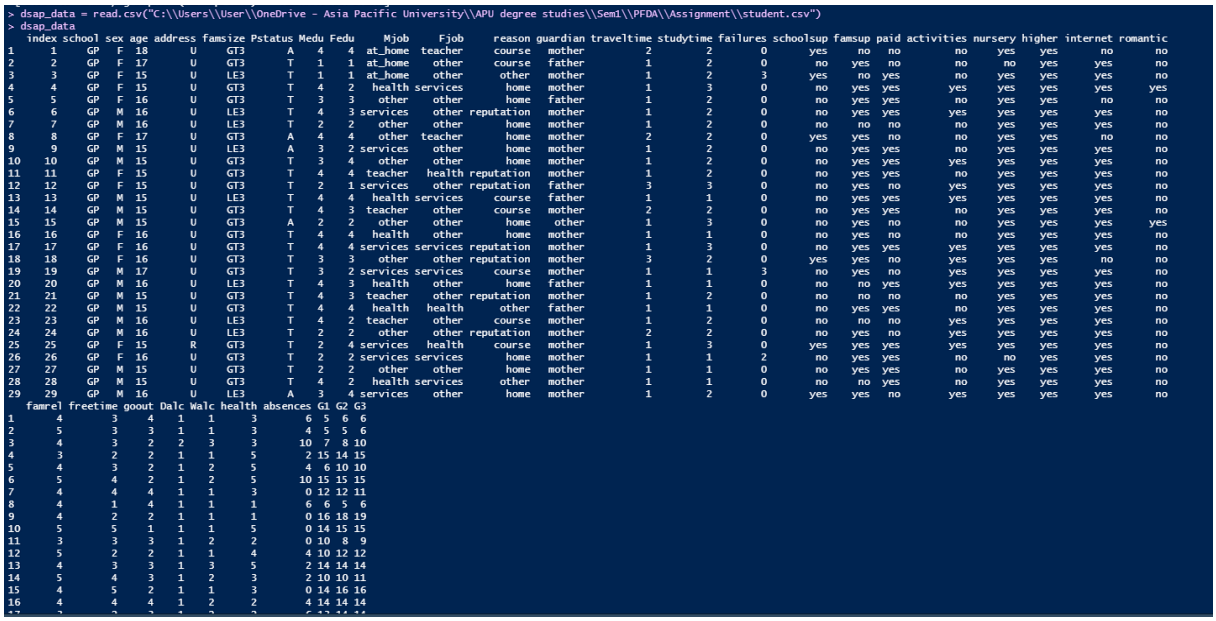


Figure 7 *shows the* ***dsap_data*** *variable executed.*



```
> dsap_data = read.csv("C:\\Users\\User\\OneDrive - Asia Pacific University\\APU degree studies\\Sem1\\PFDA\\Assignment\\student.csv")
> dsap_data
   index school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason guardian traveltime studytime failures schoolsup famsup paid activities nursery higher internet romantic
1      1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course   mother          2         2        0       yes     no   no         no     yes    yes       no       no
2      2     GP   F  17       U     GT3       T    1    1  at_home    other     course   father          1         2        0        no    yes   no         no      no    yes      yes       no
3      3     GP   F  15       U     LE3       T    1    1  at_home    other      other   mother          1         2        3       yes     no  yes         no     yes    yes      yes       no
4      4     GP   F  15       U     GT3       T    4    2   health services       home   mother          1         3        0        no    yes  yes        yes     yes    yes      yes      yes
5      5     GP   F  16       U     GT3       T    3    3    other    other       home   father          1         2        0        no    yes  yes         no     yes    yes       no       no
6      6     GP   M  16       U     LE3       T    4    3 services    other reputation   mother          1         2        0        no    yes  yes        yes     yes    yes      yes       no
7      7     GP   M  16       U     LE3       T    2    2    other    other       home   mother          1         2        0        no     no   no         no     yes    yes      yes       no
8      8     GP   F  17       U     GT3       A    4    4    other  teacher       home   mother          2         2        0       yes    yes   no         no     yes    yes       no       no
9      9     GP   M  15       U     LE3       A    3    2 services    other       home   mother          1         2        0        no    yes  yes         no     yes    yes      yes       no
10    10     GP   M  15       U     GT3       T    3    4    other    other       home   mother          1         2        0        no    yes  yes        yes     yes    yes      yes       no
11    11     GP   F  15       U     GT3       T    4    4  teacher   health reputation   mother          1         2        0        no    yes  yes         no     yes    yes      yes       no
12    12     GP   F  15       U     GT3       T    2    1 services    other reputation   father          3         3        0        no    yes   no        yes     yes    yes      yes       no
13    13     GP   M  15       U     LE3       T    4    4   health services     course   father          1         1        0        no    yes  yes        yes     yes    yes      yes       no
14    14     GP   M  15       U     GT3       T    4    3  teacher    other     course   mother          2         2        0        no    yes  yes         no     yes    yes      yes       no
15    15     GP   M  15       U     GT3       A    2    2    other    other       home    other          1         3        0        no    yes   no         no     yes    yes      yes      yes
16    16     GP   F  16       U     GT3       T    4    4   health    other       home   mother          1         1        0        no    yes   no         no     yes    yes      yes       no
17    17     GP   F  16       U     GT3       T    4    4 services services reputation   mother          1         3        0        no    yes  yes        yes     yes    yes      yes       no
18    18     GP   F  16       U     GT3       T    3    3    other    other reputation   mother          3         2        0       yes    yes   no        yes     yes    yes       no       no
19    19     GP   M  17       U     GT3       T    3    2 services services     course   mother          1         1        3        no    yes   no        yes     yes    yes      yes       no
20    20     GP   M  16       U     LE3       T    4    3   health    other       home   father          1         1        0        no     no  yes        yes     yes    yes      yes       no
21    21     GP   M  15       U     GT3       T    4    3  teacher    other reputation   mother          1         2        0        no     no   no         no     yes    yes      yes       no
22    22     GP   M  15       U     GT3       T    4    4   health   health      other   father          1         1        0        no    yes  yes        no      yes    yes      yes       no
23    23     GP   M  16       U     LE3       T    4    2  teacher    other     course   mother          1         2        0        no     no   no        yes     yes    yes      yes       no
24    24     GP   M  16       U     LE3       T    2    2    other    other reputation   mother          2         2        0        no    yes   no        yes     yes    yes      yes       no
25    25     GP   F  15       R     GT3       T    2    4 services   health     course   mother          1         3        0       yes    yes  yes        yes     yes    yes      yes       no
26    26     GP   F  16       U     GT3       T    2    2 services services       home   mother          1         1        2        no    yes  yes         no      no    yes      yes       no
27    27     GP   M  15       U     GT3       T    2    2    other    other       home   mother          1         1        0        no    yes  yes         no     yes    yes      yes       no
28    28     GP   M  15       U     GT3       T    4    2   health services      other   mother          1         1        0        no     no  yes         no     yes    yes      yes       no
29    29     GP   M  16       U     LE3       A    3    4 services    other       home   mother          1         2        0       yes    yes   no        yes     yes    yes      yes       no
   famrel freetime goout Dalc Walc health absences G1 G2 G3
1       4        3     4    1    1      3        6  5  6  6
2       5        3     3    1    1      3        4  5  5  6
3       4        3     2    2    3      3       10  7  8 10
4       3        2     2    1    1      5        2 15 14 15
5       4        3     2    1    2      5        4  6 10 10
6       5        4     2    1    2      5       10 15 15 15
7       4        4     4    1    1      3        0 12 12 11
8       4        1     4    1    1      1        6  6  5  6
9       4        2     2    1    1      1        0 16 18 19
10      5        5     1    1    1      5        0 14 15 15
11      3        3     3    1    2      2        0 10  8  9
12      5        2     2    1    1      4        4 10 12 12
13      4        3     3    1    3      5        2 14 14 14
14      5        4     3    1    2      3        2 10 10 11
15      4        5     2    1    1      3        0 14 16 16
16      4        4     4    1    2      2        4 14 14 14
```

Figure 8 *shows the displayed dataset on the console.*

3.2 Data Pre-processing & Data Cleaning

Data Pre-processing and Cleaning is a process in data analysis that is used to prepare the dataset before it can used for the analyzation.

```
#=======Data Processing=======#
#========View the dataset structure========#
glimpse(dsap_data)
```

Figure 9 *shows the code for displaying dataset structure.*

As shown in the figure above, the structure of the dataset can be viewed using the "glimpse()" function. This function displays the total number of rows and columns of the dataset and the data type of all the data of the dataset on the console, as shown in figure 10 below.

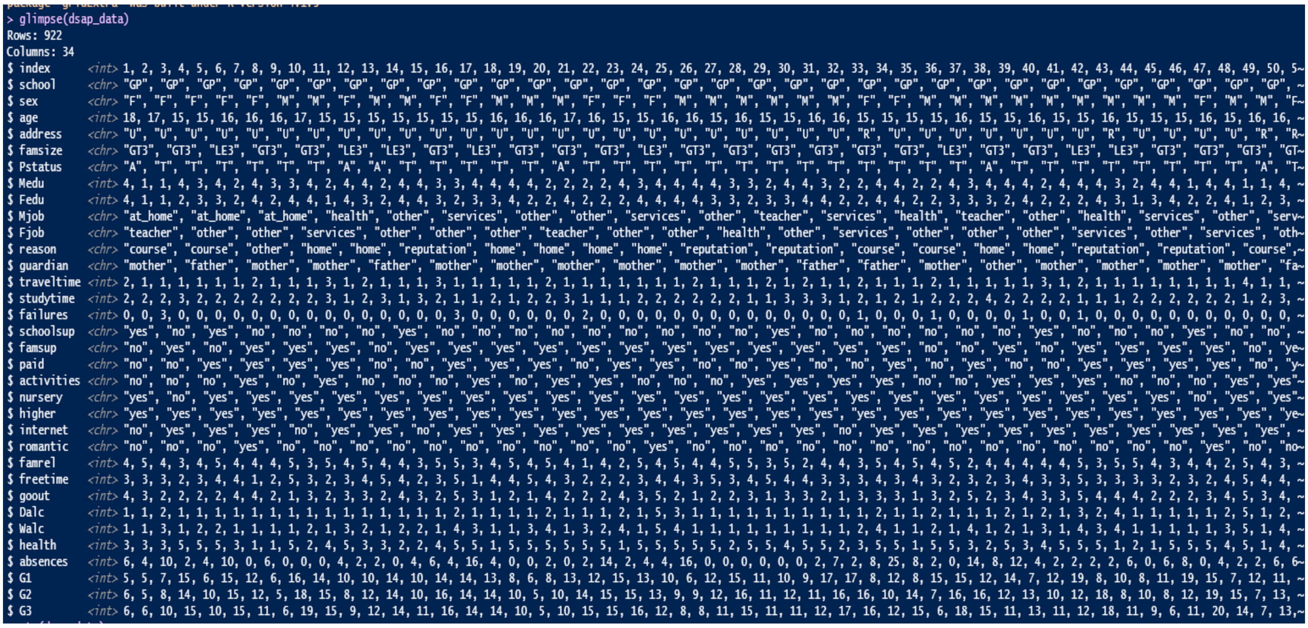


Figure 10 *shows the data structure of the dataset.*

With this data structure information, it assists to figure out any inconsistencies or missing data in the dataset. But, still, this is not a very ideal way of finding any missing data as a lot of rows need to be examined.

To check whether there is an empty value or not within each row and column in the dataset, the "sum(is.na(x))" is used and parsed into "sapply()" function as shown in figure 11 below to view the null values based on the available attributes.

```
#========Check for any null or empty values inside the dataset========#
nulldata = sapply(dsap_data,function(x) sum(is.na(x)))
nulldata
```

Figure 11 *shows the code to check null data in dataset.*

Once executing the function code, it displays the exact number of null data in each attribute from the dataset, as shown in the figure above. From the results shown, in the figure below, there isn't any null data in any of the attributes from the dataset. It is good to proceed to the next function as there isn't any null data on this dataset.

```
> #========Check for any null or empty values inside the dataset========#
> nulldata = sapply(dsap_data,function(x) sum(is.na(x)))
> nulldata
     index     school        sex        age    address    famsize    Pstatus       Medu       Fedu
         0          0          0          0          0          0          0          0          0
activities    nursery     higher   internet   romantic     famrel   freetime      goout       Dalc
         0          0          0          0          0          0          0          0          0
   Mjob       Fjob     reason   guardian traveltime  studytime   failures  schoolsup     famsup       paid
      0          0          0          0          0          0          0          0          0          0
   Walc     health   absences         G1         G2         G3
      0          0          0          0          0          0
```

Figure 12 *shows the result of null data in dataset.*

Also, it is a very good practice in data analysis to check any duplication of data in the dataset. For that, the "duplicated()" function can be used to find any duplicated rows of data in the dataset. As shown in the figures below, after executing the function code, it returns 0 on the console so it means all the data in this dataset are unique , and it doesn't contain any duplicated value.

```
#========Check for any data duplication inside the dataset
sum(duplicated(dsap_data)==TRUE)
```

Figure 13 *shows the code to check the duplicated data.*

```
> #========Check for any data duplication inside the dataset=
> sum(duplicated(dsap_data)==TRUE)
[1] 0
```

Figure 14 *shows the result of duplicated data.*

3.3 Data Transformation

After done processing the dataset, the categorical data are school additional educational support, family educational support, additional paid classes, participation in extra-curricular activities, nursery school attendance, the decision to take higher education, internet access and status of the romantic relationship are also been transformed into numerical data for efficient analysis process, where the no and yes been changed to 1 and 2 respectively. The function used to do this is the "factor()" that lists out the distinct values of the data into the "unclass()" function.

```
#========Data Transformation========#
#no - 1 || yes  - 2
dsap_data$schoolsup = unclass(factor(dsap_data$schoolsup))
dsap_data$famsup = unclass(factor(dsap_data$famsup))
dsap_data$paid = unclass(factor(dsap_data$paid))
dsap_data$activities = unclass(factor(dsap_data$activities))
dsap_data$nursery = unclass(factor(dsap_data$nursery))
dsap_data$higher = unclass(factor(dsap_data$higher))
dsap_data$internet = unclass(factor(dsap_data$internet))
dsap_data$romantic = unclass(factor(dsap_data$romantic))
```

Figure 15 *shows the code to transform categorical data into numeric value.*

schoolsup	famsup	paid	activities	nursery	higher	internet	romantic
2	1	1	1	2	2	1	1
1	2	1	1	1	2	2	1
2	1	2	1	2	2	2	1
1	2	2	2	2	2	2	2
1	2	2	1	2	2	1	1
1	2	2	2	2	2	2	1
1	1	1	1	2	2	2	1
2	2	1	1	2	2	1	1
1	2	2	1	2	2	2	1
1	2	2	2	2	2	2	1
1	2	2	1	2	2	2	1

Figure 16 *shows the output of the categorical data after executing the code.*

As there are three columns for the student's grade, new columns were created to store the average mark of these three marks, which will be feasible to analyse and compare with other attributes. The figure below shows the code to calculate the average mark and store it in a new column called **avgGradeNotRoundedOff** and **avgGrade** in the dataset.

```
#========Create a new column for the mean of the math grade for three period========#
dsap_data$avgGradeNotRoundedOff = rowMeans(subset(dsap_data, select = c(G1,G2,G3)))
dsap_data$avgGrade = round(dsap_data$avgGradeNotRoundedOff)
```

Figure 17 *shows the code that calculating average mark.*

avgGradeNotRoundedOff	avgGrade
5.666667	6
5.333333	5
8.333333	8
14.666667	15
8.666667	9
15.000000	15
11.666667	12
5.666667	6
17.666667	18
14.666667	15
9.000000	9
11.333333	11
14.000000	14

Figure 18 *shows the output of calculating average mark.*

With that, another column has also been created on the dataset column that stores the range of the students' average mark, which was calculated earlier. This column also will be useful to do the analysis and plot the graphs based on their range of average marks. The figure below shows

the code to assign the average mark to its range and store it in a new column called **avgGradeRange** in the dataset.

```
#========Assigning the average grade into a new range column========#
dsap_data = dsap_data %>% mutate(avgGradeRange = case_when(avgGrade <=5 ~ "00-05",
                                                           avgGrade <=10 ~ "06-10",
                                                           avgGrade <=15 ~ "11-15",
                                                           avgGrade <=20 ~ "16-20",))
```

Figure 19 *shows the code that calculating average mark range.*

avgGradeRange
06-10
00-05
06-10
11-15
06-10
11-15
11-15
06-10
16-20
11-15
06-10
11-15

Figure 20 *shows the output of calculating average mark range.*

The overall average mean mark has also been counted and stored into a variable called **avgMeanGrade**. This also will be useful for data analysis in the future. The figures below show on how to calculate the average mean of the mark of the students.

```
#========Average grade of all students combining together========#
avgMeanGrade = round(mean(dsap_data$avgGrade))
avgMeanGrade
```

Figure 21 *shows the code that calculating overall average mean mark.*

```
> #========Average grade of all students combining together========#
> avgMeanGrade = round(mean(dsap_data$avgGrade))
> avgMeanGrade
[1] 11
```

Figure 22 *shows the output of overall average mean mark.*

3.4 Data Exploration

After the Data Transformation process, the **dsap_data** dataset can be viewed using the "View()" function which displays the data in a table form, as shown in the figures below. Also, with the use of the "head()" function, it is possible to display the number of rows needed.

```
#View all the data from the dataset in a table form
View(head(dsap_data, 10)) #Seeing the first 10 rows of data
```

Figure 23 *shows the code to view the first 10 rows from the dataset.*

index	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	2
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	1
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	2
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	1
5	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	1
6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	1
7	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	1
8	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	2
9	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	1
10	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	1

Figure 24 *shows the output of the first 10 rows from the dataset. (1)*

famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health
1	1	1	2	2	1	1	4	3	4	1	1	3
2	1	1	1	2	2	1	5	3	3	1	1	3
1	2	1	2	2	2	1	4	3	2	2	3	3
2	2	2	2	2	2	2	3	2	2	1	1	5
2	2	1	2	2	1	1	4	3	2	1	2	5
2	2	2	2	2	2	1	5	4	2	1	2	5
1	1	1	2	2	2	1	4	4	4	1	1	3
2	1	1	2	2	1	1	4	1	4	1	1	1
2	2	1	2	2	2	1	4	2	2	1	1	1
2	2	2	2	2	2	1	5	5	1	1	1	5

Figure 25 *shows the output the first 10 rows from the dataset. (2)*

absences	G1	G2	G3	avgGradeNotRoundedOff	avgGrade	avgGradeRange
6	5	6	6	5.666667	6	06-10
4	5	5	6	5.333333	5	00-05
10	7	8	10	8.333333	8	06-10
2	15	14	15	14.666667	15	11-15
4	6	10	10	8.666667	9	06-10
10	15	15	15	15.000000	15	11-15
0	12	12	11	11.666667	12	11-15
6	6	5	6	5.666667	6	06-10
0	16	18	19	17.666667	18	16-20
0	14	15	15	14.666667	15	11-15

Figure 26 *shows the output of the first 10 rows from the dataset. (3)*

The summary for every single column from the dataset can be calculated using the "summary()" function, as shown in the figure below.

```
#=======Data Exploration========#
#========View the summary of the dataset========#
summary(dsap_data)
```

Figure 27 *shows the code to summarise all the data.*

```
> summary(dsap_data)
     index          school              sex                 age          address             famsize             Pstatus
 Min.   :  1.0   Length:922         Length:922         Min.   :15.00   Length:922         Length:922         Length:922
 1st Qu.:231.2   Class :character   Class :character   1st Qu.:16.00   Class :character   Class :character   Class :character
 Median :461.5   Mode  :character   Mode  :character   Median :17.00   Mode  :character   Mode  :character   Mode  :character
 Mean   :461.5                                         Mean   :16.74
 3rd Qu.:691.8                                         3rd Qu.:18.00
 Max.   :922.0                                         Max.   :22.00
      Medu            Fedu            Mjob               Fjob              reason            guardian          traveltime
 Min.   :0.000   Min.   :0.000   Length:922         Length:922         Length:922         Length:922         Min.   :1.000
 1st Qu.:2.000   1st Qu.:2.000   Class :character   Class :character   Class :character   Class :character   1st Qu.:1.000
 Median :3.000   Median :2.500   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.000
 Mean   :2.753   Mean   :2.536                                                                               Mean   :1.457
 3rd Qu.:4.000   3rd Qu.:3.000                                                                               3rd Qu.:2.000
 Max.   :4.000   Max.   :4.000                                                                               Max.   :4.000
   studytime        failures        schoolsup          famsup            paid           activities        nursery          higher
 Min.   :1.000   Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
 Median :2.000   Median :0.0000   Median :1.000   Median :2.000   Median :1.000   Median :1.000   Median :2.000   Median :2.000
 Mean   :2.037   Mean   :0.3319   Mean   :1.124   Mean   :1.612   Mean   :1.453   Mean   :1.498   Mean   :1.794   Mean   :1.952
 3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :4.000   Max.   :3.0000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
    internet        romantic         famrel         freetime        goout            Dalc            Walc           health
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000
 Median :2.000   Median :1.000   Median :4.000   Median :3.000   Median :3.000   Median :1.000   Median :2.000   Median :4.000
 Mean   :1.831   Mean   :1.331   Mean   :3.949   Mean   :3.252   Mean   :3.092   Mean   :1.496   Mean   :2.293   Mean   :3.565
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000
 Max.   :2.000   Max.   :2.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
    absences           G1              G2              G3        avgGradeNotRoundedOff    avgGrade     avgGradeRange
 Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00   Min.   : 1.333        Min.   : 1.00   Length:922
 1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.: 8.333        1st Qu.: 8.00   Class :character
 Median : 4.000   Median :11.00   Median :11.00   Median :11.00   Median :10.667        Median :11.00   Mode  :character
 Mean   : 5.517   Mean   :10.94   Mean   :10.77   Mean   :10.46   Mean   :10.723        Mean   :10.71
 3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00   3rd Qu.:13.333        3rd Qu.:13.00
 Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00   Max.   :19.333        Max.   :19.00
```

Figure 28 *shows the output of summarisation all the data.*

4.0 Questions & Analyses

4.1 Question 1: How do personal relationships impact a student's grades?

This Question 1 aims to analyse whether personal relationships of the students have been impacting the students' overall average marks in the three-period grades. This question will assist the school head members to determine which factor attributes or type of the students' relationships could have been the central cause of affecting the academic performance of the students. The student attributes related to this question are the romantic relationship, the student's going out on outings with their friends, the quality of student's family relationships, students' guardians, and students' average marks.

4.1.1 Analysis 1-1: Finding the correlation between students' romantic relationship status and their average grades.

In this analysis, the correlation between the students' personal relationships and the average grades, they got from their three-period grades will be visualised and analysed. First, the data will be selected and grouped that needed for the data visualization. After that, the figures will be created and arranged accordingly. A bar graph and two pie charts have been created for this analysis.

```
#=======Question 1========#
#Question 1: How do personal relationships impact a student's grades?
#Analysis 1-1
#Finding the correlation between students' romantic relationship status and their average grades.
Q1A1R1<- dsap_data %>% group_by(romantic,avgGradeRange) %>% summarise(counts = n())
Q1A1R1
Q1A1V1<- ggplot(Q1A1R1, aes(avgGradeRange, y=counts, fill = as.factor(romantic))) +
         geom_bar(stat = "identity", position=position_dodge2(), width = 0.5, color="black") +
         ggtitle("The number of Students with their Average Student Marks grouped by their Romantic Relationship Status") +
         theme(plot.title = element_text(size = 15, face = "bold")) +
         labs(fill = "Student Romantic Relationship Status", x="Average Student Marks Range", y = "Student Counts")+
         facet_wrap(~romantic, labeller = as_labeller(c(`2`="Yes", `1`="No"))) +
         geom_text(aes(label=counts, vjust=-0.3)) +
         scale_fill_manual(values = c("#FFD700", "#B8860B"),labels = c("1 - No", "2 - Yes"))
Q1A1V1
```

Figure 29 *shows the R code used to create the data visualization figure of* ***Q1A1V1****.*

The figure above, it is showing the code for the first data visualization which is the bar graph that has been created. In this code, primarily, the **Q1A1R1** variable line of code is to grep and store only the needed attributes from the dataset using the "%>%" pipe operator and together with the use of the "group_by()" function to pick and group the needed attributes. In this case, the **romantic** and **avgGradeRange** attributes have been selected and grouped using that first line of code. Moving to the second variable, which is the **Q1A1V1**, the graph is plotted using the "ggplot()" function and parsing the Q1A1R1 variable into the ggplot to initialize the selected attributes. A couple of functions that need to be considered in these lines of codes for

this **Q1A1V1** variable. The "aes()" means aesthetics, and in simple terms, it is something that you can visualize. This aesthetic will be mapped with a graphic cue and an attribute into it. The avgGradeRange has been used as the "x" axis, the total counts of the grouped attributes has been assigned to the "y" axis and the "fill" used as the romantic attribute. The next function is the "geom_bar()" which makes the graph a bar graph. Inside this function, the "stat = identity" creates the height of the bar in proportion to the number of available counts in each of the grouped data. The "position=position_dodge2()" has been included to separate the overlapping bars into different lines. The "ggtitle()" function allows giving a title for the graph while the "theme()" function is used to customize any appearance of the graph. The "labs()" function enables editing names of the "fill", "x" axis and "y" axis. The "facet_wrap()" function is used to separate the bars by their respective category based on the different data under the attribute inserted and, in this case, the **romantic** attribute used. The "geom_text()" is used to display the total number of students of the attributes on top of each bar of the graph. Lastly, the "scale_fill_manual()" has been utilized to change the colour of the bar and the labels of the legend according to the different data of the **romantic** attribute.

```
Q1A1R2 <- dsap_data %>% group_by(romantic) %>% filter(avgGrade > 15) %>%
          summarise(counts = n(), percentage= n()/length(which(dsap_data$avgGrade>15))*100)
Q1A1R2
Q1A1V2 <- ggplot(Q1A1R2, aes(x="", y =percentage, fill=as.factor(romantic))) +
          geom_col(color = "black") + coord_polar("y", start = 0) +
          theme(panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
          plot.title = element_text(size = 20, face = "bold")) +
          geom_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                    color = c("white"), position = position_stack(vjust=0.5)) +
          ggtitle("Shows the percentage of the Students that scored\n> 15 marks and their Romantic Relantionship\nStatus.") +
          labs(fill="Romantic Relationship Status")+ scale_fill_manual(values = c("#228B22","#CD5C5C"),
                                                                       labels = c("1 - No", "2 - Yes"))
Q1A1V2
```

Figure 30** shows the R code used to create the data visualization figure of **Q1A1V2.

In the figure shown above, the first variable that is **Q1A1R2** picks and stores the needed data, the same as previously explained in the **Q1A1R1** variable. But in this case, there is something different in this variable code that it filters out the **avgGrade** to get data that is more than 15 average marks using the "filter()" function and count the total data and calculate the percentage of the count data using the "summarise()" function. The second variable is **Q1A1V2** which comprises most of the same codes as the **Q1A1V1** variable. One of the important functions that differ from the previous visualization variable is the "coord_polar()" function. This function makes the shape of the graph a round pie chart. Also, additionally, the "round()" function has been implemented to round off and have a better view of the percentage values.

```
Q1A1R3 <- dsap_data %>% group_by(romantic) %>% filter(avgGrade <5) %>%
        summarise(counts = n(), percentage= n()/length(which(dsap_data$avgGrade<5))*100)
Q1A1R3
Q1A1V3 <- ggplot(Q1A1R3, aes(x="", y =percentage, fill=as.factor(romantic))) + geom_col(color = "black") +
        coord_polar("y", start = 0) +
        theme(panel.background = element_blank(),
              axis.title = element_blank(),
              axis.text = element_blank(),
              axis.line = element_blank(),
              axis.ticks= element_blank(),
        plot.title = element_text(size = 20, face = "bold")) +
        geom_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")), color = c("white"),
                  position = position_stack(vjust=0.5)) +
        ggtitle("Shows the percentage of the Students that scored\n< 5 marks and their Romantic Relantionship\nStatus.") +
        labs(fill="Romantic Relationship Status") +
        scale_fill_manual(values = c("#228B22","#CD5C5C"), labels = c("1 - No", "2 - Yes"))
Q1A1V3
```

Figure 31 *shows the R code used to create the data visualization figure of* ***Q1A1V3****.*

The code shown in the figure above is the same line of code that has been used previously for the pie graph. The significant change done to this code is where the **Q1A1R3** variable greps that average grade that is less than 5.

```
#Show combined two pies into one view
ggarrange(Q1A1V2, Q1A1V3, nrow = 2, ncol = 1)
```

Figure 32 *shows the R code used to arrange the data visualization figures of* ***Q1A1V2*** *and* ***Q1A1V3*** *in one view.*

This "ggarrange()" code shown in the figure above basically combines the variables of pie charts into a single view by parsing the graph variable and specifying the number of view rows and columns that required. In this case, the **Q1A1V2** and **Q1A1V3** graph variables have been parsed and the "nrow()" has been set to 2 and "ncol()" has been defaulting as 1, which will display both graphs in a single column view.

```
> Q1A1R1
# A tibble: 8 x 3
# Groups:   romantic [2]
  romantic avgGradeRange counts
     <int> <chr>          <int>
1        1 00-05             48
2        1 06-10            241
3        1 11-15            262
4        1 16-20             66
5        2 00-05             39
6        2 06-10            109
7        2 11-15            145
8        2 16-20             12
```

Figure 33 *shows the output of the* ***Q1A1R1*** *variable.*

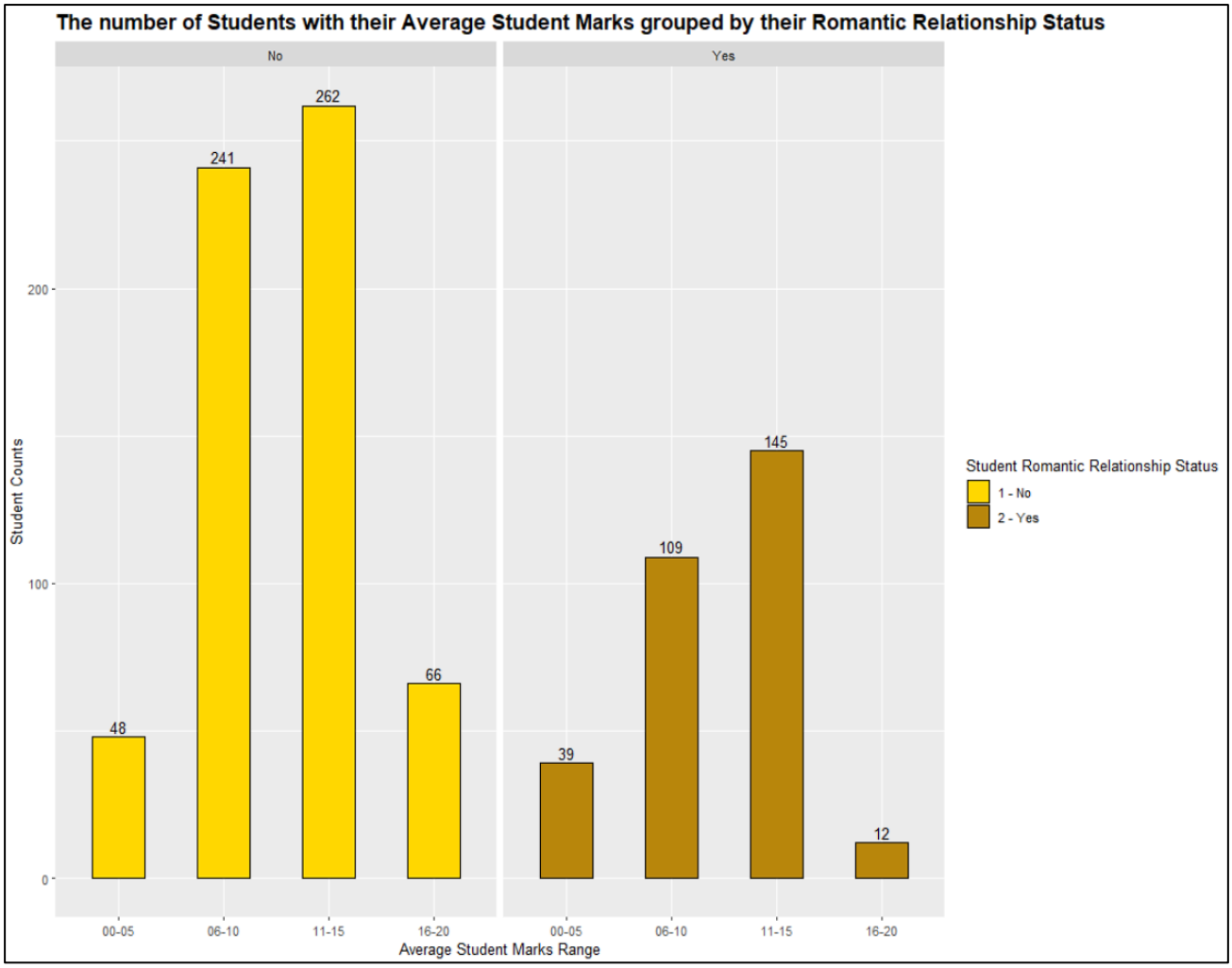


Figure 34 *shows the bar graph of the output of Q1A1V1.*

The figure 33 shows the output of the execution of the **Q1A1R1**, which shows the grouped counts of total students for each case of selected attributes that are romantic relationship status and their average grade range. The figure 34 above shows the output of the bar graph plotted after the execution of the **Q1A1V1** variable that displays the students' counts and the average grade range of students grouped by their romantic relationship status.

```
> Q1A1R2
# A tibble: 2 x 3
  romantic counts percentage
     <int>  <int>      <dbl>
1        1     66       84.6
2        2     12       15.4
> |
```

Figure 35 *shows the output of the* ***Q1A1R2****.*

```
> Q1A1R3
# A tibble: 2 x 3
  romantic counts percentage
     <int>  <int>      <dbl>
1        1     28       49.1
2        2     29       50.9
```

Figure 36 *shows of the output of* ***Q1A1R3****.*

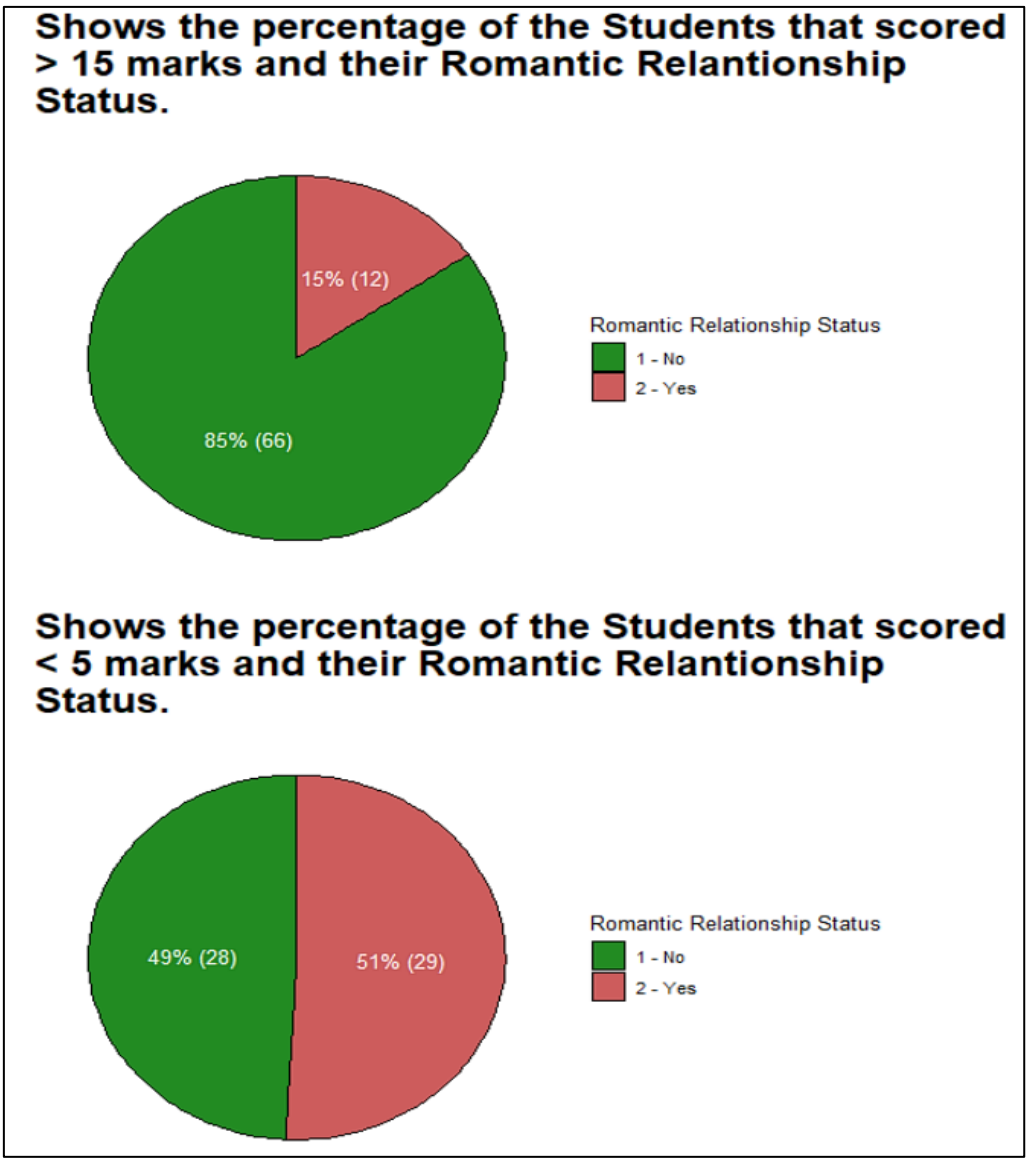


Figure 37 *shows the pie charts of the output of* ***Q1A1V2*** *and* ***Q1A1V3****.*

The figures 35 and 36 show the execution output of the **Q1A1R2** and **Q1A1R3**, which displays the calculated total student counts and percentage of the grouped relationship status. In the pie charts shown above in figure 37, it displays the percentage of students that scored more than 15 marks and less than 5 marks grouped by their romantic relationship status accordingly.

Summary for Data Findings

1) Majority of the students are not in a romantic relationship.
2) There is a total of 617 students who are not in a romantic relationship while the rest of 304 students are in a romantic relationship.
3) 85% of the students who are not in a romantic relationship scored average grades of more than 15.
4) Only 15% of the students who are in a romantic relationship scored average grades of more than 15.
5) 49% of the students who are not in a romantic relationship scored average grades of less than 5.
6) 51% of the students who are in a romantic relationship scored average grades of less than 5.

Explanation for the Data Findings.

Based on the data findings summary stated above, it can be clearly seen that the majority of the students were not been in a romantic relationship. When looking in-depth, 85%, which is a total of 66 the students who are not in a romantic relationship, while the other 15%, that is a total of only 12 of them who were in a romantic relationship have scored more than 15 average marks, which this mark viewed as an excellent average mark grade for the three-period. As can be seen, there is a big difference between these two relationship statuses in this outcome. On the other hand, when analysing the number of students that have scored an average grade of less than 5, that considered a very low average mark, 51% that is a total of 29 of them had personal relationships, while the rest of the 49%, which is a total of 28 students wasn't had one. As can be seen, there was only a little difference between these relationship statuses, which is only 2%. Hence, we can safely claim that most of the students who had excellent average scores were not been in a relationship. This could be because usually, in a relationship, it would need to endure so much commitment and time into it, which can result in the student losing their focus on their studies and reduced participation in any education-related activities. It is also possible that this romantic relationship can cause them to break down emotionally even when there is a small misunderstanding going on between their significant other where they can end up distracting all their important things including their focus on studying. According to Kasagga and Nakijoba (2020), have further supported in their research that students who have been in a romantic relationship tend to pay less attention to their classes and are unable to

manage their time which leads them to spend fewer hours studying. In conclusion, a romantic relationship of a student can affect negatively their study time and attention in class which eventually can make them not perform well in their studies and get low marks in their examination. While for the students who were not in a relationship, not the case where they can manage their studies properly without thinking and having any other thoughts than studies which can result in them getting very high grades.

4.1.2 Analysis 1-2: Finding the correlation between students' going on outings with their friends and their average grades.

In this analysis, the correlation between the students' level of going out with friends and their average grades, they got from their three-period grades will be visualised and analysed. Same as the previous, a stacked bar graph and two pie charts have been created for this analysis.

```
#Analysis 1-2
#Finding the correlation between students' going on outings with their friends and their average grades.
Q1A2R1<- dsap_data %>%  group_by(goout,avgGradeRange) %>% summarise(counts = n())
Q1A2R1
Q1A2V1<- ggplot(Q1A2R1, aes(x=avgGradeRange, y=counts, fill=as.factor(goout))) +
         geom_bar(stat="identity",width = 0.5, color="black") +
         ggtitle("The number of Students with their average score grouped by the level of them going out.")+
         labs(x="Average Student Marks Range", y = "Student Counts", fill="Students Going Out With Friends")+
         theme(plot.title = element_text(size = 15, face = "bold")) +
         scale_fill_manual(values=c("#FF6347", "#FFA500", "#DAA520", "#7CFC00", "#2E8B57", "#1E90FF"),
                           labels = c("1 - Very Low", "2 - Low", "3 - Medium", "4 - High", "5 - Very High"))+
         geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q1A2V1
```

Figure 38 *shows the R code used to create the data visualization figure of* ***Q1A2V1****.*

On figure 38 above, it shows the code for drawing the stacked bar graph with the use same code as done for the first analysis bar graph. The **goout** and **avgGradeRange** attributes have been grouped into the **Q1A2R1** variable to get the counts of total students. As to create a stacked bar graph, the "position=position_dodge()" function has been removed from the **Q1A2V1** in order to allow the overlapping bar.

```
Q1A2R2 <- dsap_data %>% group_by(goout) %>% filter(avgGrade>15) %>%
          summarise(counts = n(),percentage= n()/length(which(dsap_data$avgGrade>15))*100)

Q1A2R2
Q1A2V2 <- ggplot(Q1A2R2, aes(x="", y =percentage, fill=as.factor(goout))) + geom_col(color = "black") + coord_polar("y", start = 0) +
          theme(panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
          plot.title = element_text(size = 20, face = "bold")) +
          geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                    color = c("white"), position = position_stack(vjust=0.5)) +
          ggtitle("Shows the percentage of the Students that scored\n> 15 marks and their frequency of Going Out\nwith Friends.") +
          labs(fill="Students Going Out With Friends")+
           scale_fill_manual(values = c("#4B0082","#DA70D6","#8B4513", "#191970","#2F4F4F"),
                             labels = c("1 - Very Low", "2 - Low", "3 - Medium", "4 - High", "5 - Very High"))
Q1A2V2
```

Figure 39 *shows the R code used to create the data visualization figure of* ***Q1A2V2****.*

In figure 39, the **Q1A2R2** variable stores the grouped **goout** attribute by filtering the **avgGrade** attribute from the dataset to get the only score that is more than the 15 average scores of a student and calculating the total counts and percentage of students. The **Q1A2V2** variable is the code to draw the pie chart for the data in the **Q1A2R2** variable.

```
Q1A2R3<- dsap_data %>% group_by(goout) %>% filter(avgGrade <=5) %>%
        summarise(counts = n(), percentage= n()/length(which(dsap_data$avgGrade<=5))*100)
Q1A2R3
Q1A2V3 <- ggplot(Q1A2R3, aes(x="", y =percentage, fill=as.factor(goout))) + geom_col(color = "black") + coord_polar("y", start = 0) +
         theme(panel.background = element_blank(),
               axis.title = element_blank(),
               axis.text = element_blank(),
               axis.line = element_blank(),
               axis.ticks= element_blank(),
         plot.title = element_text(size = 20, face = "bold")) +
         geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                   color = c("white"), position = position_stack(vjust=0.5)) +
         ggtitle("Shows the percentage of the Students that scored\n<= 5 marks and their frequency of Going Out\nwith Friends.") +
         labs(fill="Students Going Out With Friends")+
         scale_fill_manual(values = c("#4B0082","#DA70D6","#8B4513", "#191970","#2F4F4F"),
                           labels = c("1 - Very Low", "2 - Low", "3 - Medium", "4 - High", "5 - Very High"))
Q1A2V3
```

Figure 40 *shows the R code used to create the data visualization figure of* ***Q1A2V3****.*

In figure 40, the **Q1A2R3** variable stores the grouped **goout** attribute by filtering the **avgGrade** attribute from the dataset to get the only score that is less than equal to the 5 average scores of a student and calculating the total counts and percentage of students. The **Q1A2V3** variable is the code to draw the pie chart for the data in the **Q1A2R3** variable.

```
ggarrange(Q1A2V2, Q1A2V3, nrow = 2, ncol = 1)
```

Figure 41 *shows the R code used to arrange the data visualization figures of* ***Q1A2V2*** *and* ***Q1A2V3*** *in one view.*

.

As shown in figure 41 above, this is code written to arrange both Q1A2V2 and Q1A2V3 variable pie charts into a single column view**.**

```
> Q1A2R1
# A tibble: 20 x 3
# Groups:   goout [5]
   goout avgGradeRange counts
   <int> <chr>          <int>
 1     1 00-05              6
 2     1 06-10             19
 3     1 11-15             27
 4     1 16-20              5
 5     2 00-05             21
 6     2 06-10             73
 7     2 11-15            112
 8     2 16-20             33
 9     3 00-05             23
10     3 06-10            111
11     3 11-15            148
12     3 16-20             25
13     4 00-05             13
14     4 06-10             98
15     4 11-15             82
16     4 16-20              7
17     5 00-05             24
18     5 06-10             49
19     5 11-15             38
20     5 16-20              8
```

Figure 42 *shows the output of the* ***Q1A2R1****.*

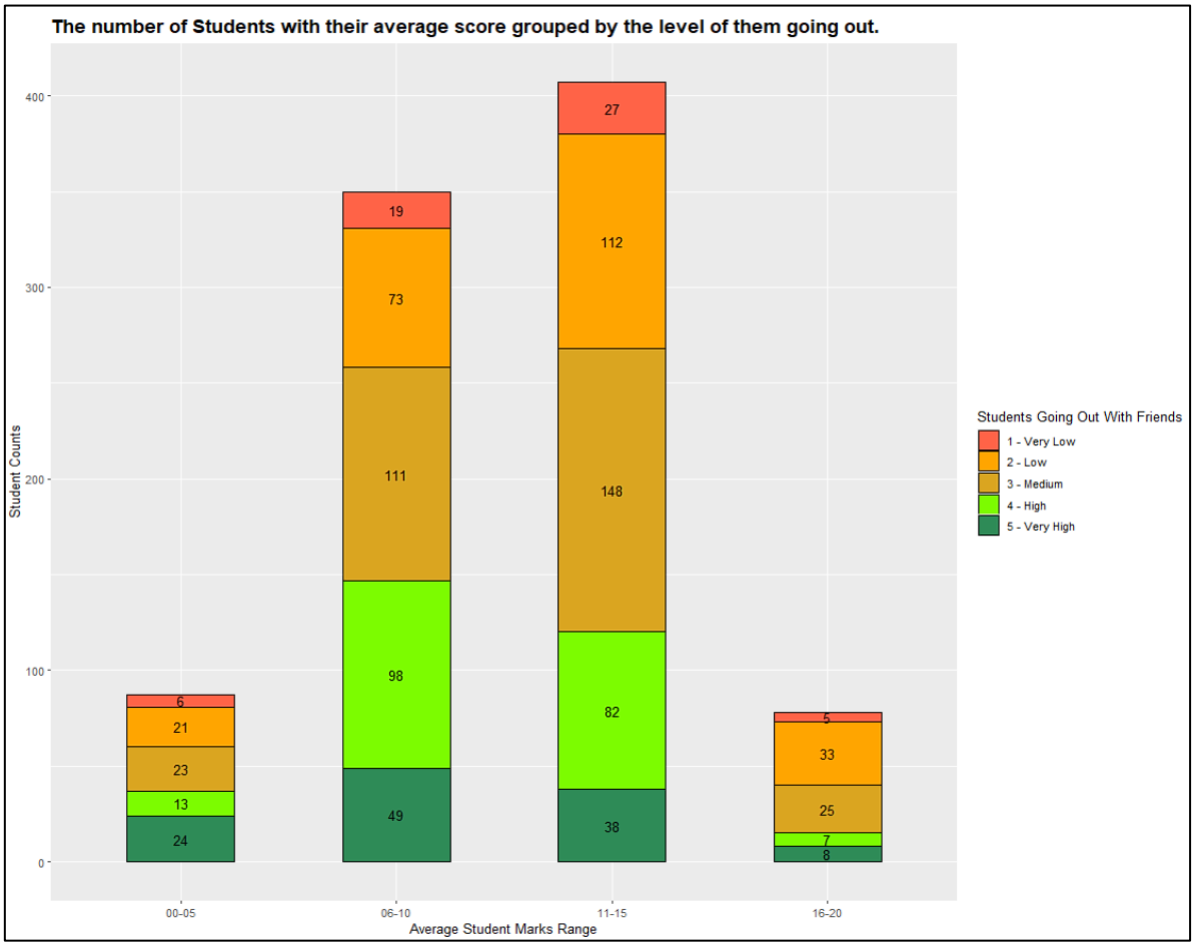


Figure 43 *shows the stacked bar graph of output of the* ***Q1A2V1****.*

The figure 42 above shows the output of the execution of the **Q1A2R1**, which shows the grouped counts of total students for each case of selected attributes that are frequency of going out with friends and their average grade range. The figure 43 above shows the output of the stacked bar graph plotted after the execution of the **Q1A2V1** variable that displays the students' counts and the average grade range of students grouped by their frequency of going out with friends.

```
> Q1A2R2
# A tibble: 5 x 3
  goout counts percentage
  <int>  <int>      <dbl>
1     1      5       6.41
2     2     33      42.3
3     3     25      32.1
4     4      7       8.97
5     5      8      10.3
```

***Figure 44** shows the output of the **Q1A2R2**.*

```
> Q1A2R3
# A tibble: 5 x 3
  goout counts percentage
  <int>  <int>      <dbl>
1     1      6       6.90
2     2     21      24.1
3     3     23      26.4
4     4     13      14.9
5     5     24      27.6
>
```

***Figure 45** shows the output of the **Q1A2R3**.*

The figures 44 and 45 show the execution output of the **Q1A2R2** and **Q1A2R3**, which displays the calculated total student counts and percentage of the grouped frequency of student going out.

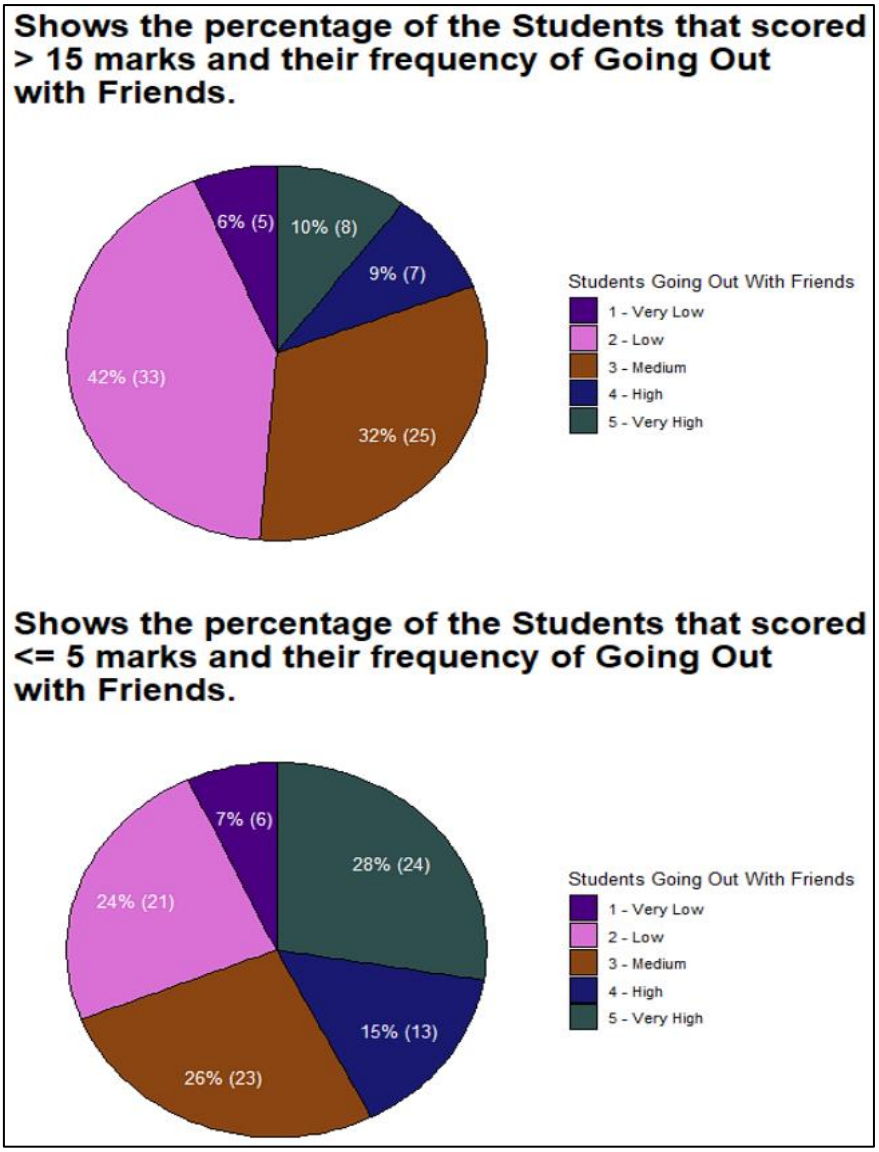


Figure 46 *shows the pie charts of output of the* ***Q1A2V2*** *and* ***Q1A2V3****.*

The figures 44 and 45 show the execution output of the **Q1A2R2** and **Q1A2R3**, which displays the calculated total student counts and percentage of the grouped frequency of students going out. In figure 46, the two pie charts that have been plotted display the percentage of students that scored more than 15 marks and less than and equal to 5 marks, grouped by their students' frequency of going out with friends accordingly.

Summary for Data Findings

1) Most numbers of the students go out with their friends on medium frequency.
2) The second most number of students go out with friends with low frequency.
3) The least number of the students go out with very low frequency with their friends.
4) About 42% of students that scored more than 15 average grades only go out with friends with low frequency.
5) About 32% of students that scored more than 15 average grades only go out with friends with medium frequency.
6) 74% of students that scored more than 15 average scores only went out with friends with less frequency.
7) About 28% of students that scored less than and equal to 5 average marks go out with friends with very high frequency.
8) About 15% of students that scored less than and equal to 5 average marks go out with friends with very high frequency.
9) 46% of students who scored less than 5 average scores went out with friends with high frequency.

Explanation for the Data Findings.

Based on the data findings summary stated above, when looking in-depth, 42% that is a total of 33 students who go out with friends a low amount times manage to get very excellent average grade marks. Another 32% is a total of 25 students who manage to get very excellent average marks by spending time with friends for a medium amount of time. Combining these two categories, it makes that about 74% of students that is 54 of them, that is the majority of students who managed to achieve great average mark are the students who fairly goes out with friends. Also, when looking in-depth at students who got very less average marks, 28% is 24 of them go out with friends with very high amount times. From this, it can be concluded that the number of times going out with friends also can impact their overall academic performance. Universally, friends have been a support system for everyone, especially in this case for the students. When going out with friends for a fair amount of time, the students can speak openly about anything that helps to reduce their worries. Together with that, by going out with their friends, they also can have fun which makes to relieve their stress. But this is not particularly always true too, because not all of them want to assist students out, and they also could be

trying to bring them down. This was further supported by Jennifer Flashman in her research that stated that friends assist and supply study resources to them while also can motivate and demotivating in terms of their studies. Going out a lot with friends can also affect their grades in a bad way where they might not get proper study time and result in them doing badly in their exams. Therefore, going out a fair amount with good friends can help students to achieve better results in their examinations.

4.1.3 Analysis 1-3: Finding the relationship between the quality of the student's family relationships and the average grades.

In this analysis, a bar graph and two pie charts have been created to find the relationship between the quality of the student's family relationship and the average grades.

```
#Analysis 1-3
#Finding the relationship between the quality of the student's family relationships and the average grades.
Q1A3R1<- dsap_data %>% group_by(famrel,avgGradeRange) %>% summarise(counts = n())
Q1A3R1
Q1A3V1<- ggplot(Q1A3R1, aes(x= avgGradeRange, y=counts, fill = as.factor(famrel))) +
         geom_bar(stat = "identity", position = position_dodge2(preserve = 'single'), width=0.9)+
         ggtitle("The number of Students with their average score grouped by their Family Relantionship quality.")+
         theme(plot.title = element_text(size = 15, face = "bold")) +
         labs(fill = "Students Family Relationship Level", x="Average Students Marks Range", y = "Student Counts")+
         geom_text(aes(label=counts), position = position_dodge2(1), vjust=-0.5) +
         scale_fill_manual(values = c("#40E0D0","#191970","#FF1493", "#DEB887","#6A5ACD"),
                           labels = c("1 - Very Bad", "2 - Bad", "3 - Okay", "4 - Good", "5 - Excellent"))
Q1A3V1
```

Figure 47 *shows the R code used to create the data visualization figure of* ***Q1A3V1****.*

```
Q1A3R2 <- dsap_data %>% group_by(famrel) %>% filter(avgGrade>15)%>%
          summarise(counts = n(), percentage= n()/length(which(dsap_data$avgGrade>15))*100)
Q1A3R2
Q1A3V2<- ggplot(Q1A3R2, aes(x="", y =percentage, fill=as.factor(famrel)))+
         geom_col(color = "black") + coord_polar("y", start = 0)+
         theme(panel.background = element_blank(),
               axis.title = element_blank(),
               axis.text = element_blank(),
               axis.line = element_blank(),
               axis.ticks= element_blank(),
         plot.title = element_text(size = 20, face = "bold")) +
         geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                   color = c("white"), position = position_stack(vjust=0.5)) +
         ggtitle("Shows the percentage of the students that scored\n> 15 marks
                 and their quality of relationship with their family.") +
         labs(fill="Students Family Relationship Level")+
         scale_fill_manual(values = c("#40E0D0","#191970","#FF1493", "#DEB887","#6A5ACD"),
                           labels = c("1 - Very Low", "2 - Low", "3 - Medium", "4 - High", "5 - Very High"))
Q1A3V2
```

Figure 48 *shows the R code used to create the data visualization figure of* ***Q1A3V2****.*

```
Q1A3R3 <- dsap_data %>% group_by(famrel) %>% filter(avgGrade<=5)%>%
          summarise(counts = n(), percentage= n()/length(which(dsap_data$avgGrade<=5))*100)
Q1A3R3
Q1A3V3<- ggplot(Q1A3R3, aes(x="", y =percentage, fill=as.factor(famrel)))+
         geom_col(color = "black") + coord_polar("y", start = 0) +
         theme(panel.background = element_blank(),
               axis.title = element_blank(),
               axis.text = element_blank(),
               axis.line = element_blank(),
               axis.ticks= element_blank(),
         plot.title = element_text(size = 20, face = "bold"))+
         geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                   color = c("white"), position = position_stack(vjust=0.5)) +
         ggtitle("Shows the percentage of the students that scored\n<= 5 marks
                 and their quality of relationship with their family.") +
         labs(fill="Students Family Relationship Level")+
         scale_fill_manual(values = c("#40E0D0","#191970","#FF1493", "#DEB887","#6A5ACD"),
                           labels = c("1 - Very Low", "2 - Low", "3 - Medium", "4 - High", "5 - Very High"))
Q1A3V3
```

Figure 49 *shows the R code used to create the data visualization figure of* ***Q1A3V3****.*

```
ggarrange(Q1A3V2, Q1A3V3, nrow = 2, ncol = 1)
```

Figure 50 *shows the R code used to arrange the data visualization figures of* ***Q1A3V2*** *and* ***Q1A3V3*** *in one view.*

As shown in the figures above, it is pretty much the same way and functions that were coded and covered in the previous analysis. It doesn't have a big difference in any functions added or changed. But one important thing is that for this analysis it was grouped based on the **famrel** and **avgGradeRange** attribute.

```
> Q1A3R1
# A tibble: 20 x 3
# Groups:   famrel [5]
   famrel avgGradeRange counts
    <int> <chr>           <int>
 1      1 00-05               3
 2      1 06-10               3
 3      1 11-15              11
 4      1 16-20               2
 5      2 00-05               6
 6      2 06-10               8
 7      2 11-15              27
 8      2 16-20               2
 9      3 00-05              15
10      3 06-10              65
11      3 11-15              67
12      3 16-20               8
13      4 00-05              31
14      4 06-10             179
15      4 11-15             207
16      4 16-20              37
17      5 00-05              32
18      5 06-10              95
19      5 11-15              95
20      5 16-20              29
```

Figure 51 *shows the output of the* ***Q1A3R1****.*

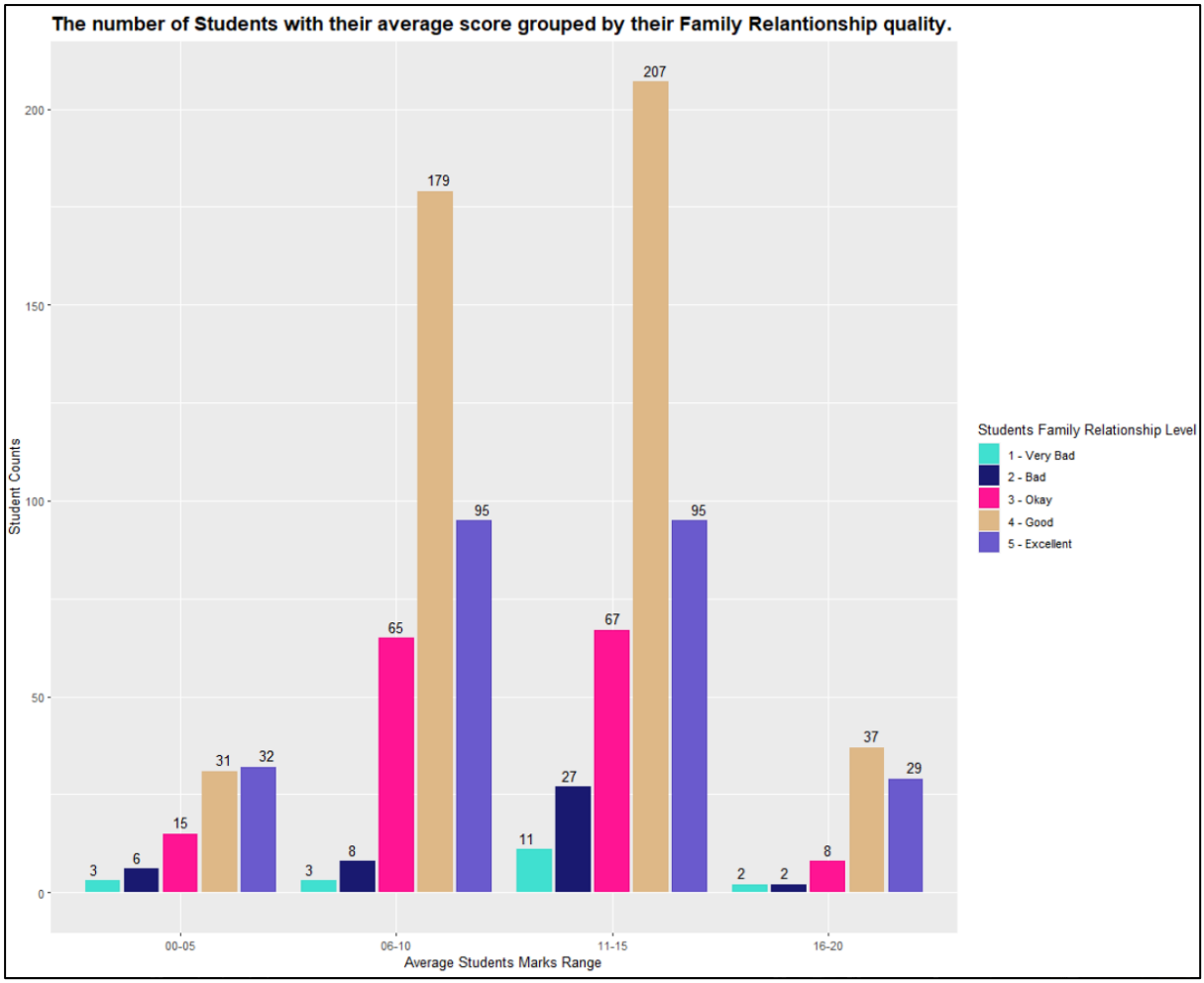


***Figure 52** shows the bar graph output of the **Q1A3V1**.*

The figure 51 above shows the output of the execution of the **Q1A3R1**, which shows the grouped counts of total students for each case of selected attributes that are the quality of their family relationship and their average grade range. The figure 52 above shows the output of the bar graph plotted after the execution of the **Q1A3V1** variable that displays the students' counts and the average grade range of students grouped by their quality of family relationship.

```
Q1A3R2
# A tibble: 5 x 3
  famrel counts percentage
   <int>  <int>      <dbl>
1      1      2       2.56
2      2      2       2.56
3      3      8      10.3
4      4     37      47.4
5      5     29      37.2
```

***Figure 53** shows the output of the **Q1A3R2**.*

```
> Q1A3R3
# A tibble: 5 x 3
  famrel counts percentage
   <int>  <int>      <dbl>
1      1      3       3.45
2      2      6       6.90
3      3     15       17.2
4      4     31       35.6
5      5     32       36.8
```

Figure 54 *shows the output of the* ***Q1A3R3****.*

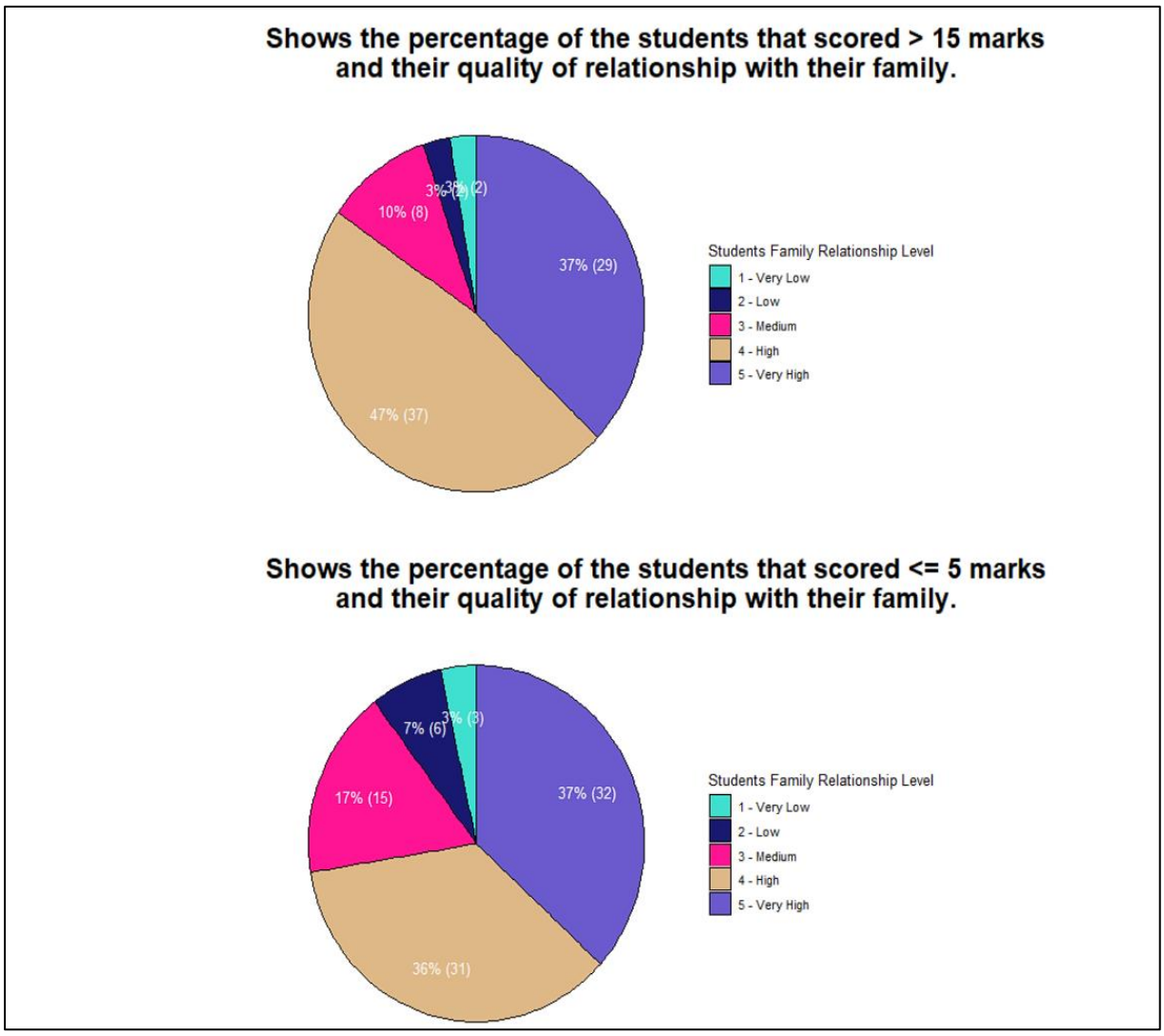


Figure 55 *shows the pie charts of the output of* ***Q1A3V2*** *and* ***Q1A3V3****.*

The figures 53 and 54 show the execution output of the **Q1A3R2** and **Q1A3R3**, which displays the calculated total student counts and percentage of the grouped quality of their family relationship. In figure 55, the two pie charts that have been plotted display the percentage of students that scored more than 15 marks and less than and equal to 5 marks, grouped by their quality of family relationship accordingly.

Summary for Data Findings

1) The majority of students had a quality 4 family relationship.
2) The second-largest portion of students had an excellent relationship with their families.
3) The least amount of students had a very bad relationship with their families.
4) 84% of students that scored an average grade of more than 15 had their family relationships above 3, which is considered a great quality relationship.
5) 73% of students that scored an average grade less than and equal to 5 had their family relationships above 3.
6) 16% of students that scored an average grade of more than 15 stated that their family relationship level was below 4, which is considered a bad quality relationship.
7) 27% of students that scored less than and equal to their average mark had a family relationship of less than 4.

Explanation for the Data Findings.

Based on the data findings summary stated above, it is pleased to see that a large portion of the students had level 4 quality relationships with their family, that were considered as good. A little percentage of overall students had a very bad relationship with their family members. Also, the majority of students who scored more than 15 and less than equal to 5 average grade had family relationships that are 3 and above. Moreover, 27%, that is a total of 24 students who scored less than and equal to 5 average marks had quality of family relationship below 4. On the other hand, 16% of the students that is 12 of them who scored more than 15 on average had family relationships below 4. As can be seen, there is an 11% contrast comparing these two statements' percentages. Thus, it can be said that a medium to a great quality family relationship can positively affect students' grades. It is because the family is also one of the support systems that assist students to keep motivated in their studies and do well in their examinations. With a good relationship with family members, students would do emotionally and mentally better, which can make them focus on the studies that eventually help them to get good grades. According to Paki et al. (2018), they further supported in their research that a family relationship is really crucial for a student's academic success. Therefore, it is true that the quality of family relationships could affect students' academic performance.

4.1.4 Analysis 1-4: Finding the relationship between the students' guardian and their average marks

For this analysis, the correlation between the students' guarding and their average grades will be visualised and analysed. A bar graph and two pie charts have been created for this analysis.

```
#Analysis 1-4
#Finding the relationship between the students' guardian and their average marks.
Q1A4R1<- dsap_data %>% group_by(guardian, avgGradeRange) %>%  summarise(counts= n())
Q1A4R1
Q1A4V1 <-  ggplot(Q1A4R1, aes(x=avgGradeRange, y = counts, fill = guardian)) +
           geom_bar(stat = "identity", width = 0.9, color="black", position = "dodge") +
           labs(fill="Guardian", x = "Average Student Marks Range", y="Student Counts") +
           ggtitle("The number of Students with their average score grouped by Guardian.") +
           theme(plot.title = element_text(size = 15, face = "bold")) +
           geom_text(aes(label=counts), position = position_dodge2(1), vjust=-0.5)
Q1A4V1
```

Figure 56 *shows the R code used to create the data visualization figure of* ***Q1A4V1****.*

```
Q1A4R2<-  dsap_data %>% group_by(guardian) %>% filter(avgGrade>=avgMeanGrade)%>%
          summarise(counts = n(), percentage= n()/length(which(dsap_data$avgGrade>=avgMeanGrade))*100)
Q1A4R2
Q1A4V2 <- ggplot(Q1A4R2, aes(x="", y =percentage, fill=guardian)) +
          geom_col(color = "black") + coord_polar("y", start = 0) +
          theme(panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
          plot.title = element_text(size = 20, face = "bold"))+
          geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                    color = c("white"), position = position_stack(vjust=0.5)) +
          ggtitle("Shows the percentage of the students that scored >=
          average mean grade and their Guardian.") +
          labs(fill="Guardian")+
          scale_fill_manual(values = c("#191970","#FF1493", "#00008B"))
Q1A4V2
```

Figure 57 *shows the R code used to create the data visualization figure of* ***Q1A4V2****.*

```
Q1A4R3<-  dsap_data %>% group_by(guardian) %>% filter(avgGrade<avgMeanGrade)%>%
          summarise(counts = n(), percentage= n()/length(which(dsap_data$avgGrade<avgMeanGrade))*100)
Q1A4R3
Q1A4V3 <- ggplot(Q1A4R3, aes(x="", y =percentage, fill=guardian))+
          geom_col(color = "black") + coord_polar("y", start = 0) +
          theme(panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
          plot.title = element_text(size = 20, face = "bold"))+
          geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                    color = c("white"), position = position_stack(vjust=0.5)) +
          ggtitle("Shows the percentage of the students that scored <
          average mean grade and their Guardian.") +
          labs(fill="Guardian")+ scale_fill_manual(values = c("#191970","#FF1493", "#00008B"))
Q1A4V3
```

Figure 58 *shows the R code used to create the data visualization figure of* ***Q1A4V3****.*

```
ggarrange(Q1A4V2, Q1A4V3, nrow = 2, ncol = 1)
```

Figure 59 *shows the R code used to arrange the data visualization figures of* ***Q1A4V2*** *and* ***Q1A4V3*** *in one view.*

As shown in the code figures above, the **guardian** and **avgGradeRange** is grouped and counted for this analysis based on their average grade. For the pie charts, the **avgGrade** has been filtered out so that it only displays the needed output for the analysis and the percentage has also been counted on the code above. Lastly, both pie charts have been arranged into a single view.

```
> Q1A4R1
# A tibble: 12 x 3
# Groups:   guardian [3]
   guardian avgGradeRange counts
   <chr>    <chr>          <int>
 1 father   00-05             16
 2 father   06-10             78
 3 father   11-15             97
 4 father   16-20             18
 5 mother   00-05             64
 6 mother   06-10            235
 7 mother   11-15            279
 8 mother   16-20             58
 9 other    00-05              7
10 other    06-10             37
11 other    11-15             31
12 other    16-20              2
>
```

Figure 60 *shows the output of the* ***Q1A4R1***.

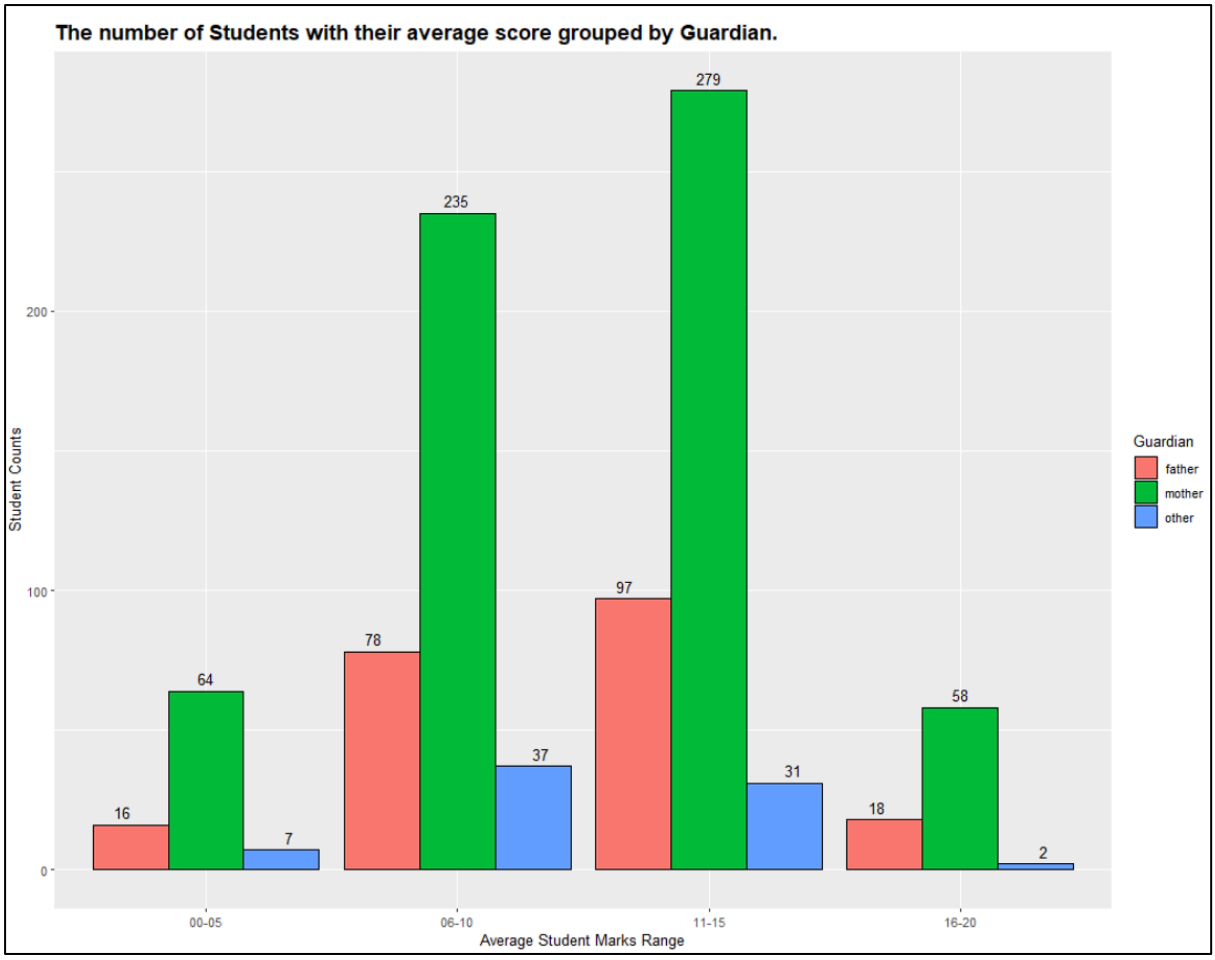


Figure 61 *shows the bar graph output of the* ***Q1A4V1****.*

The figure 60 above shows the output of the execution of the **Q1A4R1**, which shows the grouped counts of total students for each case of selected attributes that are the students' guardian and their average grade range. The figure 62 above shows the output of the bar graph plotted after the execution of the **Q1A4V1** variable that displays the students' counts and the average grade range of students grouped based on their guardians.

```
> Q1A4R2
# A tibble: 3 x 3
  guardian counts percentage
  <chr>     <int>      <dbl>
1 father      115       23.7
2 mother      337       69.5
3 other        33        6.80
```

Figure 62 *shows the output of the* ***Q1A4R2****.*

```
> Q1A4R3
# A tibble: 3 x 3
  guardian counts percentage
  <chr>     <int>      <dbl>
1 father       94       21.5
2 mother      299       68.4
3 other        44       10.1
```

Figure 63 *shows the output of the* ***Q1A4R3****.*

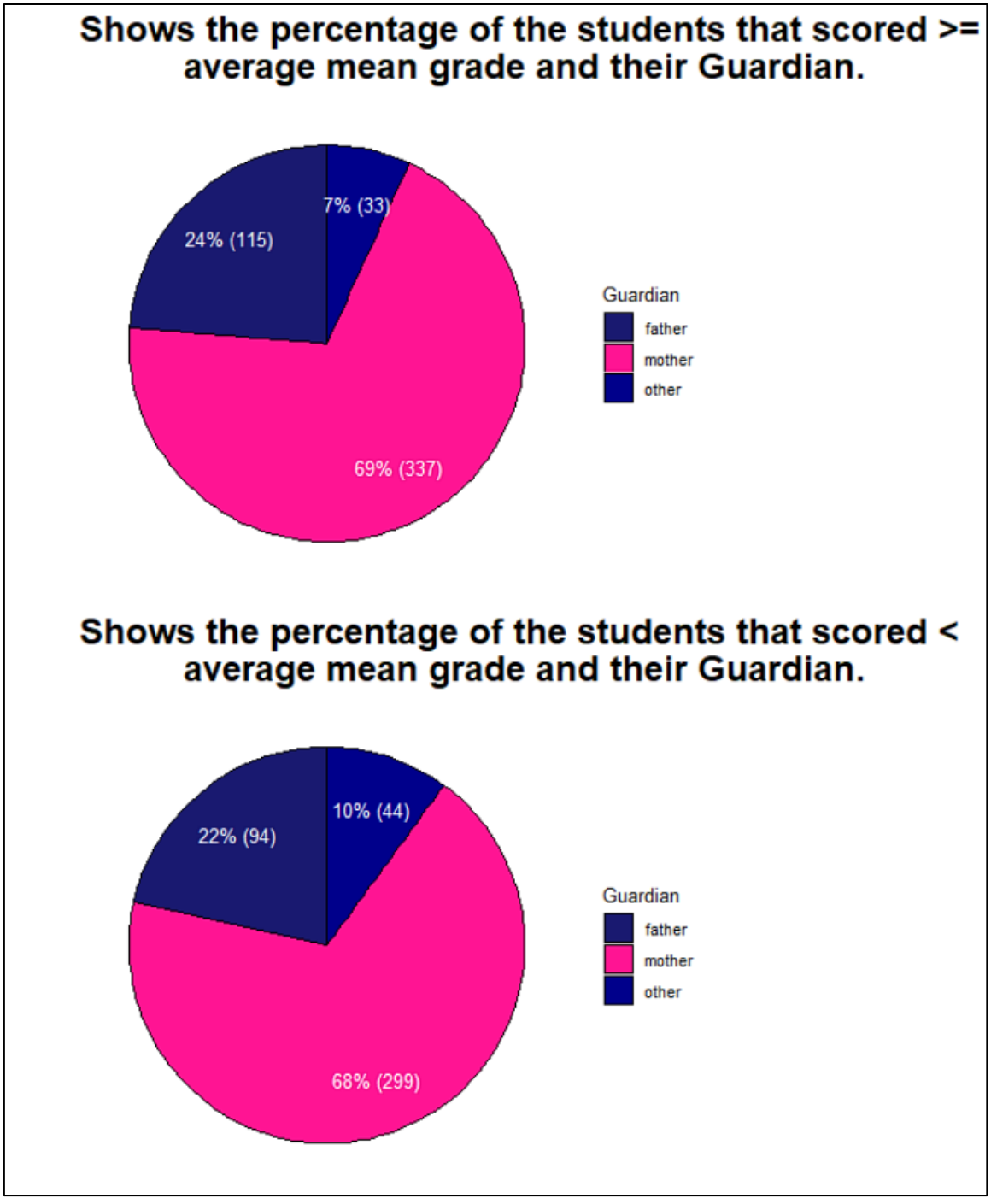


Figure 64 *shows the pie charts of the output of* ***Q1A3V2*** *and* ***Q1A3V3****.*

The figures 62 and 63 show the execution output of the **Q1A4R2** and **Q1A4R3**, which displays the calculated total student counts and percentage grouped by student guardian. In figure 64, the two pie charts that have been plotted display the percentage of students that scored more than average mean marks and less than and equal to average mean marks, grouped based on their guardians.

Summary for Data Findings

1) A huge portion of students had mothers as their guardians.
2) Second largest portion was fathers as the students' guardians.
3) Only a little amount of students had others as their guardians.
4) 69% of students that scored an average mark of more than and equal to the average mean grade had their mother as a guardian.
5) 68% of students that scored less than average mean mark had their mother as guardian.

Explanation for the Data Findings.

Based on the data findings summary stated above, when looking deeper at the graphs drawn, the largest portion of the students that is 69% which is 337 of the total of them got an overall average mark above the average mean marks. Similarly, students who scored less than the average mean grade also majority had their mothers as guardians, that are about 299 of them. This tells that most of the students' mothers are monitoring and helping the students in whatever case it is, including their study performance. The mothers play a crucial role as a guardian in assisting students with their academics. This wouldn't work out if the students don't do their part as do properly in their studies in order to achieve good grades. Overall, most students do great in academics because they tend to listen to their mothers. As this was further stated in a research done by Lara and Saracostti (2019), the involvement of a parent, especially mothers had helped students get excellent marks compared to the one who didn't involve.

4.1.5 Analysis 1-5: Finding the relationship between students' romantic relationship status, their study time and their average marks

For this analysis, the correlation between the students' guarding and their average grades will be visualised and analysed. A stacked bar graph has been created for this analysis.

```
#Analysis 1-5
#Finding the relationship between students' romantic relationship status, their study time and their average marks.
Q1A5R1 <- dsap_data %>% group_by(romantic,avgGradeRange, studytime)  %>%  summarise(counts= n())
Q1A5R1
Q1A5V1<- ggplot(Q1A5R1, aes(x=avgGradeRange, y=counts, fill = as.factor(studytime)))+
         geom_bar(stat = "identity", width = 0.5, color="black") +
         ggtitle("The number of Students with their Average Student Marks
         grouped by their Romantic Relationship Status and the total Study Time") +
         theme(plot.title = element_text(size = 15, face = "bold")) +
         labs(fill = "Student Study Time (Hours)", x="Average Student Marks Range", y = "Student Counts")+
         facet_wrap(~romantic, labeller = as_labeller(c(`1`="Relationship: No", `2`="Relationship: Yes"))) +
         geom_text(aes(label=counts), position = position_stack(vjust = 0.5)) +
         scale_fill_manual(values=rainbow(10))
Q1A5V1
```

Figure 65 *shows the R code used to create the data visualization figure of* ***Q1A5V1***.

As shown in the code figures above, the **romantic**, **studytime** and **avgGradeRange** are grouped and counted for this analysis based on their average grade range.

```
> Q1A5R1
# A tibble: 30 x 4
# Groups:   romantic, avgGradeRange [8]
   romantic avgGradeRange studytime counts
      <int> <chr>              <int>  <int>
 1        1 00-05                  1     22
 2        1 00-05                  2     21
 3        1 00-05                  3      3
 4        1 00-05                  4      2
 5        1 06-10                  1     65
 6        1 06-10                  2    131
 7        1 06-10                  3     27
 8        1 06-10                  4     18
 9        1 11-15                  1     76
10        1 11-15                  2    122
# ... with 20 more rows
```

Figure 66 *shows the output of the* ***Q1A5R1***.

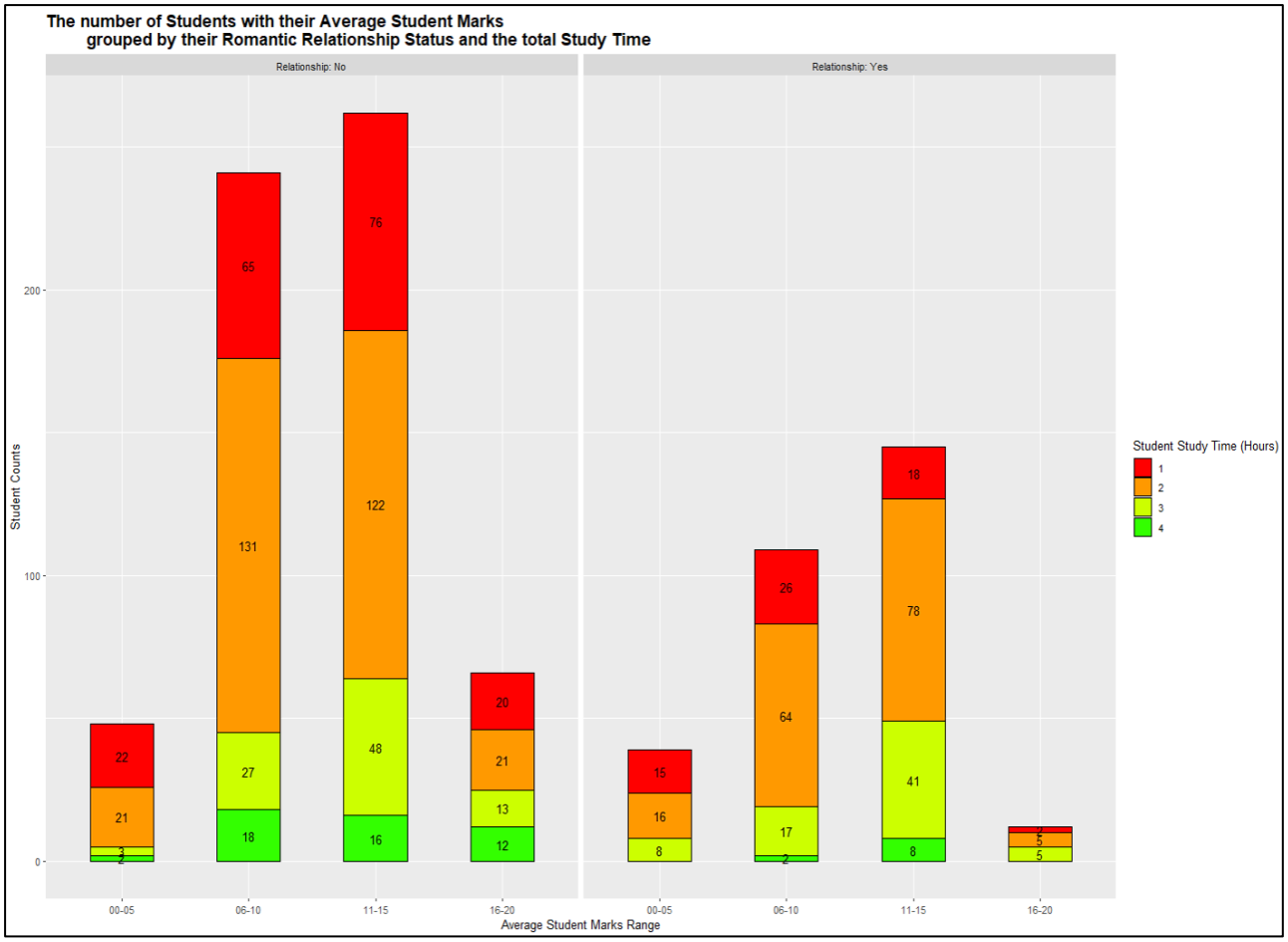


***Figure 67** shows the stacked bar graph output of the **Q1A5V1**.*

The figure 66 above shows the output of the execution of the **Q1A5R1**, which shows the grouped counts of total students for each case of selected attributes that are the students' romantic relationship status, their study time and their average grade range. The figure 67 above shows the output of the stacked bar graph plotted after the execution of the **Q1A5V1** variable that displays the students' counts and the average grade range of students grouped based on their romantic relationship status and the amount the study time.

Summary for Data Findings

1) A larger portion of students that are not in a romantic relationship studied for 4 hours as compared to those who are not in a relationship.
2) In both relationship statuses, the majority of students tend to study for 2 hours.

3) Students who scored more than 15 on the average mark that are not in a relationship spent 4 hours of study time while those who had a relationship, not a single one of them had 4 hours of study time.
4) Only a little number of students in both statuses spent 4 hours of study time.

Explanation for the Data Findings.

Based on the data findings summary stated above, it can clearly be noticed that in both romantic relationship statuses, students managed to study for two hours. Also, the total number of students who studied for 4 hours and had not had a romantic relationship is larger than the ones who had one. Moreover, students with an excellent grade that is more than 15 average grade and not in a romantic relationship had more than 4 hours of study time than the one who had a relationship that only had 3 hours at most. It can be noticed that when in a romantic relationship many students didn't get a chance to study for 4 hours. That means romantic relationships for sure affect the amount of study time of the students. As it was mentioned in the Analysis 1-1, being in a romantic relationship needs a lot of time to spend with their significant other that can take can take away students' time from their academic activities. Eventually, this does affect their academic performance can negatively. Students might experience stress completing their work before the deadline because they got spent too much time with their boyfriend or girlfriend. This was further reinforced in the research that students may spend most of their time texting which makes it harder for them to concentrate on their studies and take away a lot of their study time (Free Essays - PhDessay.com, 2017).

Conclusion

To sum this first question up, it can be confirmed that a student's personal relationship plays a vital role in impacting the success of getting excellent grades in their overall examination. All the personal relationship of a student is important as they are the one who is going to support them in their academic. Especially, the students' romantic relationship has the highest influence on affecting their average grade where the majority of students who got excellent grades are the ones who are not in a relationship as per the analysis stated. Therefore, the respective people need to advise and assist students in handling their relationships better so that it doesn't affect them mentally, emotionally and physically so they can do well in their studies.

4.2 Question 2: How does time management impact the students' grades?

The Question 2 analyses whether the way a student manages their time has been impacting their overall academic performance. This will help to know better on how these students manging their time. The student's attributes that will be covered in this question are study time, free time, and extra-curricular activities participation.

4.2.1 Analysis 2-1: Finding the correlation between students' study time and their average grade.

In the first analysis, the relationship between the students' study time and their average grades will be analysed. A horizontal bar graph and two point-to-point graphs have been created for this analysis.

```
#=======Question 2========#
#Question 2: How does time management impact the students' grades?
#Analysis 2 - 1
#Finding the relationship between the students' study time and their average marks.
Q2A1R1 <- dsap_data %>% group_by(studytime, avgGradeRange)  %>%  summarise(counts= n())
Q2A1R1
Q2A1V1<-  ggplot(Q2A1R1, aes(avgGradeRange, y=counts, fill = as.factor(studytime))) +
          geom_bar(stat = "identity", position = position_dodge2(preserve = 'single'), width=0.9) +
          ggtitle("The number of Students with their average score grouped by their Study Time.") +
          theme(plot.title = element_text(size = 15, face = "bold")) +
          labs(fill = "Student Study Time (Hours)", x="Average Students Marks Range", y = "Student Counts")+
          geom_text(aes(label=counts), position = position_dodge2(1), hjust = -0.1) +
          ylim(0,250) +
          scale_fill_manual(values = c("#40E0D0","#191970","#FF1493", "#DEB887","#6A5ACD")) +
          coord_flip() + facet_wrap(~studytime)
Q2A1V1
```

Figure 68 *shows the R code used to create the data visualization figure of* ***Q2A1V1****.*

```
Q2A1R2 <- dsap_data %>% group_by(studytime) %>% filter(avgGrade>15) %>% summarise(counts = n())
Q2A1R2
Q2A1V2 <- ggplot(Q2A1R2, aes(studytime, counts)) + geom_point(aes(color=as.factor(studytime))) +
          geom_line(aes(studytime)) +
          ggtitle("The number of Students scored > 15 average mark grouped by their Study Time.") +
          theme(plot.title = element_text(size = 15, face = "bold")) +
          labs( x="Student Study Time (Hours)", y = "Student Counts")+
          geom_text(aes(label=counts), position = position_dodge2(1), hjust = -0.5)  +
          scale_color_manual(name = "Student Study Time (Hours)",
                             values = c(`4` = "darkblue", `3` = "red", `2` = "black", `1`= "green" ))
Q2A1V2
```

Figure 69 *shows the R code used to create the data visualization figure of* ***Q2A1V2****.*

```
Q2A1R3 <- dsap_data %>% group_by(studytime) %>% filter(avgGrade<=5) %>% summarise(counts = n())
Q2A1R3
Q2A1V3 <- ggplot(Q2A1R3, aes(studytime, counts)) + geom_point(aes(color=as.factor(studytime))) +
          geom_line(aes(studytime)) +
          ggtitle("The number of Students scored <= 5 average mark grouped by their Study Time.") +
          theme(plot.title = element_text(size = 15, face = "bold")) +
          labs( x="Student Study Time (Hours)", y = "Student Counts")+
          geom_text(aes(label=counts), position = position_dodge2(1), hjust = -0.5)  +
          scale_color_manual(name = "Student Study Time (Hours)",
                             values = c(`4` = "darkblue", `3` = "red", `2` = "black", `1`= "green" ))

Q2A1V3
```

Figure 70 *shows the R code used to create the data visualization figure of* ***Q2A1V3****.*

```
ggarrange(Q2A1V2, Q2A1V3, nrow = 2, ncol = 1)
```

Figure 71 *shows the R code used to arrange the data visualization figures of* ***Q2A1V2*** *and* ***Q2A1V3*** *in one view.*

As shown in the code figures above, the **studytime** and **avgGradeRange** is grouped and counted for this analysis. To create a horizontal bar graph, the "coord_flip()" function was added to the same code used for the bar graph. This function will flip around the coordinates. For the line graphs, the **avgGrade** has been filtered out and counted the number of students based on the filtration. The "geom_point()" and "geom_line" were added to the code to create the point-to-point graph.

```
> Q2A1R1
# A tibble: 16 x 3
# Groups:   studytime [4]
   studytime avgGradeRange counts
       <int> <chr>          <int>
 1         1 00-05             37
 2         1 06-10             91
 3         1 11-15             94
 4         1 16-20             22
 5         2 00-05             37
 6         2 06-10            195
 7         2 11-15            200
 8         2 16-20             26
 9         3 00-05             11
10         3 06-10             44
11         3 11-15             89
12         3 16-20             18
13         4 00-05              2
14         4 06-10             20
15         4 11-15             24
16         4 16-20             12
```

Figure 72 *shows the output of the* ***Q2A1R1***.

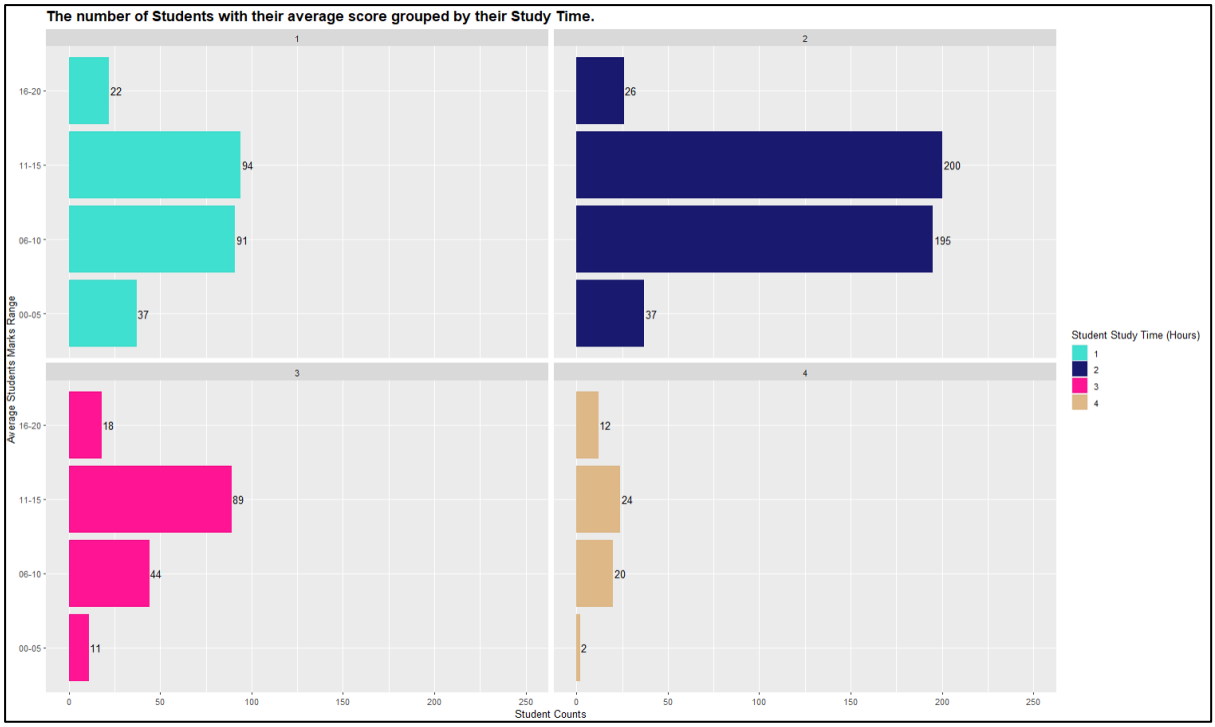


***Figure 73** shows the horizontal bar graph output of the **Q2A1V1**.*

The figure 72 above displays the output of the execution of the **Q2A1R1**, which shows the grouped counts of total students for each case of selected attributes that are the study time and their average grade range. The figure 73 above shows the outcome of the horizontal bar graph plotted after the execution of the **Q2A1V1** variable that displays the students' counts and the average grade range of students grouped based on their amount of study time.

```
> Q2A1R2
# A tibble: 4 x 2
  studytime counts
      <int>  <int>
1         1     22
2         2     26
3         3     18
4         4     12
>
```

***Figure 74** shows the output of the **Q2A1R2**.*

```
> Q2A1R3
# A tibble: 4 x 2
  studytime counts
      <int>  <int>
1         1     37
2         2     37
3         3     11
4         4      2
```

***Figure 75** shows the output of the **Q2A1R3**.*

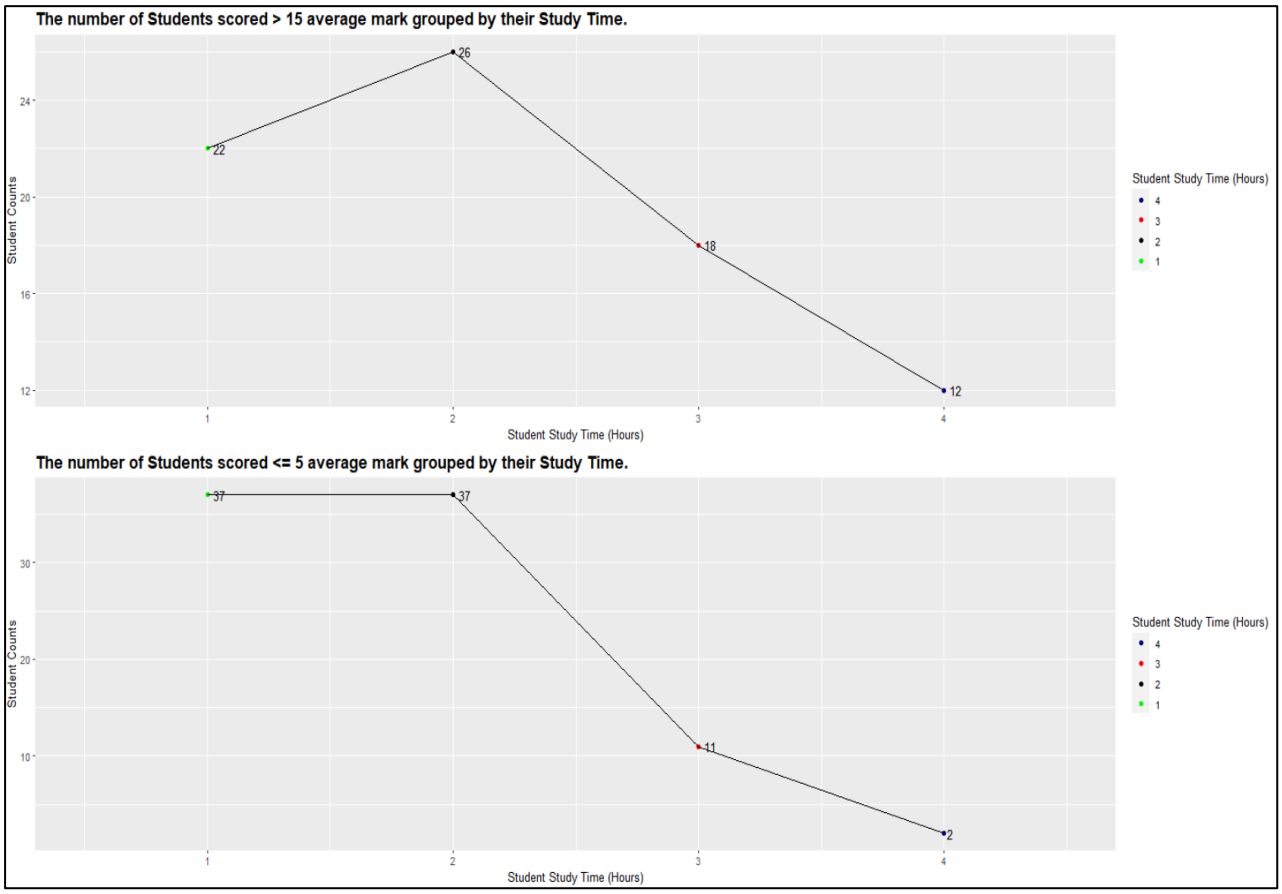


***Figure 76** shows the point-to-point graph output of the **Q2A1V2** and **Q2A1V3**.*

The figures 74 and 75 display the execution output of the **Q2A1R2** and **Q2A1R3**, which displays the calculated total student counts grouped by the amount of study time. In figure 76, the two point-to-point graph pictures the number of students that scored more than 15 average marks and less than and equal to 5 average marks grouped based on their study time.

Summary for Data Findings

1) Most number of students studied for 2 hours.
2) A very least amount of students studied for 4 hours.
3) Students who scored more than 15 on average mark had 2 hours of study time the most.
4) Students who scored less than and equal to 5 average grades had 1 to 2 hours of study time the most.

Explanation for the Data Findings.

Based on the data findings summary stated above, it portrays that many students only study for 2 hours. Moreover, very least students only had study time for 4 hours which shows they are working hard on their studies. When looking deeper, most students who scored excellent marks are the ones who studied for 2 hours. This shows that these students utilized the 2 hours efficiently and effectively to study for their examinations. Also, utilizing a few hours properly without procrastinating or any distractions can help students to revise their study resources which helps to catch up to their studies and get excellent results on their exams. While allocating more time for studies can also make them do very well in their exams if only they didn't procrastinate. This is because when they allocate the most time to studying, they can cover more topics to revise, which also enables them to study more on the difficult topics or subjects that they didn't understand very well to give a solid understanding before the exams. According to Barbarick and Ippolito (2003), it was further proven by their research that students who tend to study more hours outside of the class do better in their exam scores. Hence, it can be claimed that the longer a student studies outside the class, the chance of getting an excellent score is higher for them.

4.2.2 Analysis 2-2: Finding the relationship between students' free time and their average grades

The relationship between the students' free time and their average grades will be analysed. A stacked bar graph and two scatter plot graphs have been created for this analysis.

```
#Analysis 2-2
#Finding the relationship between students' free time and their average grades.
Q2A2R1<- dsap_data %>% group_by(freetime, avgGradeRange)  %>%  summarise(counts= n())
Q2A2R1
Q2A2V1<-  ggplot(Q2A2R1, aes(x=avgGradeRange, y=counts,fill=as.factor(freetime))) +
          geom_bar(stat="identity",width = 0.5, color="black") +
          ggtitle("The number of Students with their average score grouped by the frequency of them having Free Time (Hours).")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="Students Free Time (Hours)")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          scale_fill_manual(values=c("#FF6347", "#FFA500", "#DAA520", "#7CFC00", "#2E8B57", "#1E90FF"),labels = c("1", "2", "3", "4", "5"))+
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q2A2V1
```

Figure 77 *shows the R code used to create the data visualization figure of* ***Q2A2V1***

```
Q2A2R2<- dsap_data %>% group_by(freetime, avgGrade) %>% filter(avgGrade>15) %>% summarise(counts = n())
Q2A2R2
Q2A2V2 <- ggplot(Q2A2R2, aes(avgGrade, counts)) +
          geom_point(aes(color=as.factor(freetime)), position = position_dodge(width = 0.02)) +
          geom_line(aes(color=as.factor(freetime))) +
          ggtitle("The number of Students scored > 15 average mark grouped by their Free Time (Hours).") +
          theme(plot.title = element_text(size = 15, face = "bold")) +
          labs( x="Average Grade", y = "Student Counts")+
          geom_text(aes(label=counts), hjust = -0.6)  +
          ylim(0,20)+
          scale_color_manual(name = "Student Free Time (Hours)",
                             values = c(`5` = "yellow",`4` = "darkblue", `3` = "red", `2` = "green", `1`= "black" ))
Q2A2V2
```

Figure 78 *shows the R code used to create the data visualization figure of* ***Q2A2V2****.*

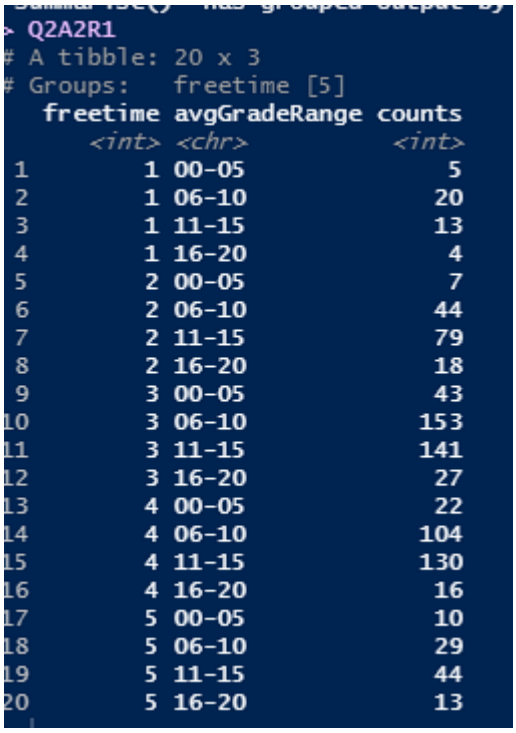
```
Q2A2R3<- dsap_data %>% group_by(freetime, avgGrade) %>% filter(avgGrade<=5) %>% summarise(counts = n())
Q2A2R3
Q2A2V3 <- ggplot(Q2A2R3, aes(avgGrade, counts)) +
          geom_point(aes(color=as.factor(freetime)), position = position_dodge(width = 0.02)) +
          geom_line(aes(color=as.factor(freetime))) +
          ggtitle("The number of Students scored <= 5 average mark grouped by their Free Time (Hours).") +
          theme(plot.title = element_text(size = 15, face = "bold")) +
          labs( x="Average Grade", y = "Student Counts")+
          geom_text(aes(label=counts), hjust = -0.5)  +
          ylim(0,20)+
          scale_color_manual(name = "Student Free Time (Hours)",
                             values = c(`5` = "yellow",`4` = "darkblue", `3` = "red", `2` = "green", `1`= "black" ))
Q2A2V3
```

Figure 79 *shows the R code used to create the data visualization figure of* ***Q2A2V3****.*

```
ggarrange(Q2A2V2, Q2A2V3, nrow = 2, ncol = 1)
```

Figure 80 *shows the R code used to arrange the data visualization figures of* ***Q2A2V2*** *and* ***Q2A2V3*** *in one view.*

As shown in the code figures above, the **freetime** and **avgGradeRange** are grouped and counted for this analysis. For the scatter plot graphs, the **avgGrade** has been filtered out and counted the number of students based on the filtration.



```
> Q2A2R1
# A tibble: 20 x 3
# Groups:   freetime [5]
   freetime avgGradeRange counts
      <int> <chr>           <int>
 1        1 00-05               5
 2        1 06-10              20
 3        1 11-15              13
 4        1 16-20               4
 5        2 00-05               7
 6        2 06-10              44
 7        2 11-15              79
 8        2 16-20              18
 9        3 00-05              43
10        3 06-10             153
11        3 11-15             141
12        3 16-20              27
13        4 00-05              22
14        4 06-10             104
15        4 11-15             130
16        4 16-20              16
17        5 00-05              10
18        5 06-10              29
19        5 11-15              44
20        5 16-20              13
```

***Figure 81** shows the output of the **Q2A2R1**.*

The figure 81 above displays the output of the execution of the **Q2A2R1**, which shows the grouped counts of total students for each case of selected attributes that are the free time and their average grade range. The figure 82 below shows the outcome of the stacked bar graph plotted after the execution of the **Q2A2V1** variable that displays the students' counts and the average grade range of students grouped based on their amount of free time.

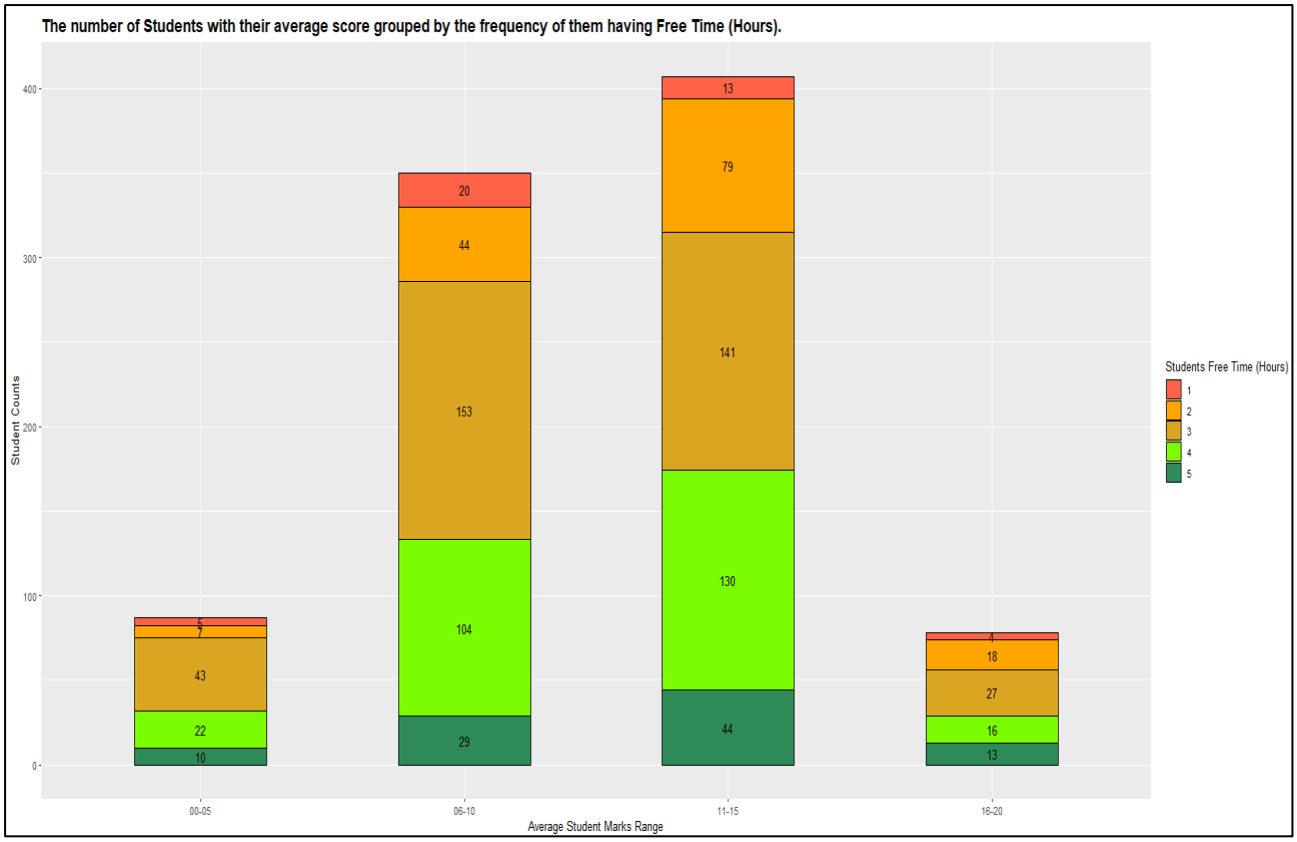


Figure 82 *shows the stacked bar graph output of the* ***Q2A2V1****.*

```
> Q2A2R2
# A tibble: 15 x 3
# Groups:   freetime [5]
   freetime avgGrade counts
      <int>    <dbl>  <int>
 1        1       18      4
 2        2       16      6
 3        2       17      4
 4        2       18      5
 5        2       19      3
 6        3       16     12
 7        3       17      8
 8        3       18      7
 9        4       16      7
10        4       17      6
11        4       19      3
12        5       16      3
13        5       17      2
14        5       18      4
15        5       19      4
>
```

Figure 83 *shows the output of the* ***Q2A2R2****.*

```
> Q2A2R3
# A tibble: 14 x 3
# Groups:   freetime [5]
   freetime avgGrade counts
      <int>    <dbl>  <int>
 1        1        4      5
 2        2        4      5
 3        2        5      2
 4        3        1      2
 5        3        2      5
 6        3        3      4
 7        3        4     19
 8        3        5     13
 9        4        2      8
10        4        3      2
11        4        4      5
12        4        5      7
13        5        2      2
14        5        5      8
```

Figure 84 *shows the output of the* ***Q2A2R3****.*

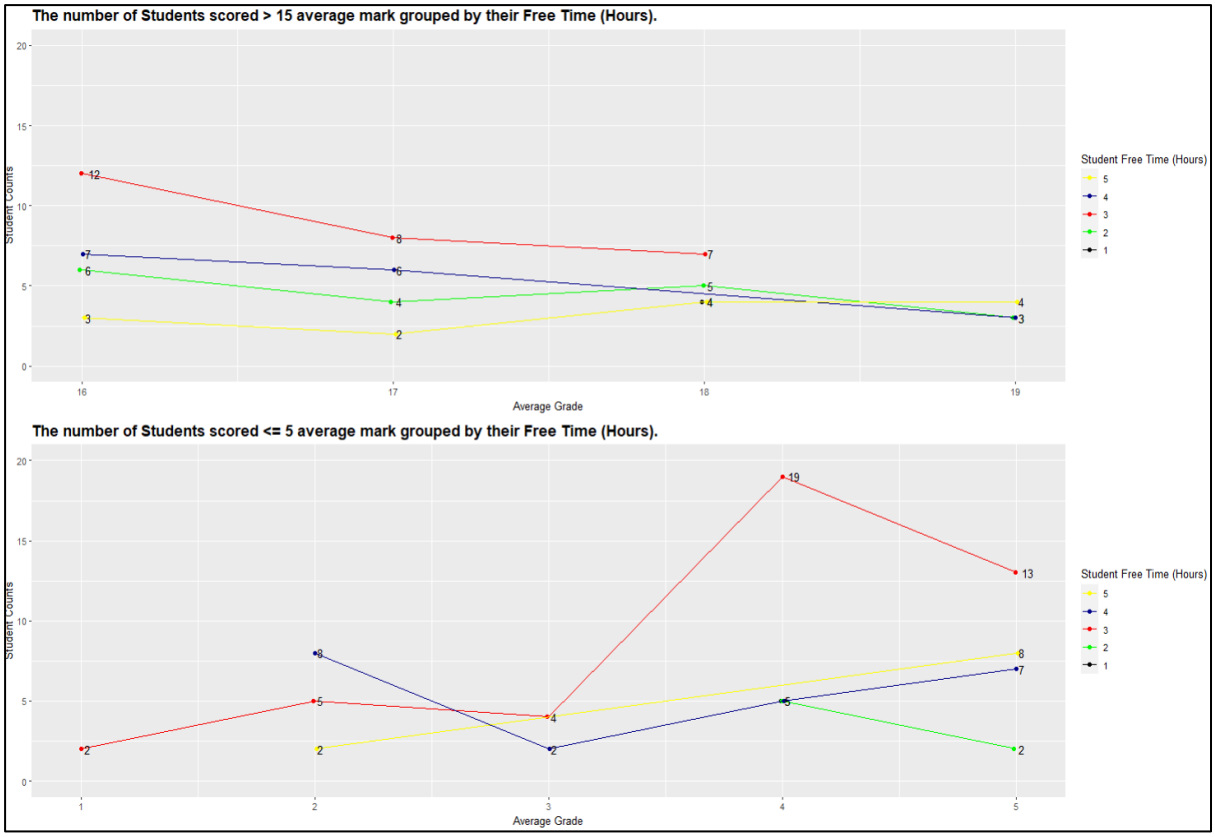


Figure 85 *shows the scatter plot graph output of the* ***Q2A2V2*** *and* ***Q2A2V3****.*

The figures 83 and 84 displays the execution output of the **Q2A2R2** and **Q2A2R3**, which displays the calculated total student counts grouped by the amount of study time. In figure 85, the two scatter plot graph pictures the number of students that scored more than 15 average marks and less than and equal to 5 average marks grouped based on their free time.

Summary for Data Findings

1) The most number of students, that is about 364 total of them had 3 hours of free time.
2) Very minimal number of students, that is about 42 of them had 1-hour free time.
3) The highest number of students who scored more than the 15 average mark had 3 hours of free time the most.
4) The least number of students who scored more than the 15 average mark had 1 hour of free time.
5) The highest number of students who scored less than and equal to the 5 average mark had 3 hours of free time the most.
6) The least number of students who scored less than and equal to 5 average mark had 1 hour of free time.

Explanation for the Data Findings.

Based on the data discoveries, it can be visualized that the majority of students had free time around 2 hours while the least number from them had 1 hour as their free time. Also, when narrowing down, the students who scored excellent grades with more than 15 average marks most of them had 3 hours free time while a very few of them had 1-hour free time. Either way, the same goes for the students who scored less than and equal to 5 average marks. However, when noticing students who got a 19 average mark which is a very great score, all of them had free time for more than 1 hour, and this tells us a lot. These students had a pretty good free time where they managed it very well to take rest and relieve their stress from studying or any study-related activities. Also, during this decent amount of free time, they can do other activities such as their hobbies and kinds of stuff which also can reduce their stress and increase their happiness overall. This can make them do their exams stress-free and it is possible to get wonderful grades when they are relaxed. On the other hand, when only got very less free time, they can't do many things other than their study-related activities, which eventually can increase their stress. According to Turner (2020), further supported that relaxed students get a lot of benefits, such as it enhances their memory and retention, increasing their concentration and productivity, and improving their overall mental health, which can impact positively on their academic grades. Hence, it is can be reasoned that students who utilise their to do relaxed activities with a fair amount of free time can relieve their stress which can improve their overall academic performance.

4.2.3 Analysis 2-3: Finding the correlation between students' extra-curricular activities participation and their average grades

The correlation between the students' extra-curricular activities participation and their average marks will be analysed. A bar graph and two stacked bar graphs have been created for this analysis.

```
#Analysis 2-3
#Finding the correlation between students' extra-curricular activities participation and their average grades.
Q2A3R1<- dsap_data %>%  group_by(activities,avgGradeRange) %>% summarise(counts = n())
Q2A3R1
Q2A3V1<-  ggplot(Q2A3R1, aes(avgGradeRange, y=counts, fill = as.factor(activities))) +
          geom_bar(stat = "identity", position=position_dodge2(), width = 0.5, color="black") +
          ggtitle("The number of Students with their Average Student Marks grouped by\ntheir Extra-curricular Activities Status") +
          theme(plot.title = element_text(size = 15, face = "bold")) +
          labs(fill = "Student Extra-curricular Activities Status",
               x="Average Student Marks Range", y = "Student Counts")+
          facet_wrap(~activities, labeller = as_labeller(c(`2`="Participated", `1`="Not Participated"))) +
          geom_text(aes(label=counts), vjust=-0.3) +
          scale_fill_manual(values = c("#800080", "#1E90FF"),labels = c("1 - No", "2 - Yes"))
Q2A3V1
```

Figure 86 *shows the R code used to create the data visualization figure of* ***Q2A3V1***

```
Q2A3R2 <- dsap_data %>% filter(activities==2, avgGrade>15) %>%
          group_by(activities, studytime, avgGrade) %>% summarise(counts = n())
Q2A3R2
Q2A3V2 <- ggplot(Q2A3R2, aes(avgGrade, y=counts, fill=as.factor(studytime))) + geom_bar(stat = "identity",width = 0.5, color="black") +
          ggtitle("Students that have scored > 15 average mark and been joined extra-curricular activites\ngrouped by their Study Time (Hours).") +
          theme(plot.title = element_text(size = 15, face = "bold")) +
          labs(fill = "Student Time (Hours)", x="Average Student Marks", y = "Student Counts")  +
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5)) +
          scale_fill_manual(values=c("#FF6347", "#FFA500", "#DAA520", "#7CFC00", "#2E8B57"))
Q2A3V2
```

Figure 87 *shows the R code used to create the data visualization figure of* ***Q2A3V2***

```
Q2A3R3 <- dsap_data %>% filter(activities==1, avgGrade>15) %>%
          group_by(activities, studytime, avgGrade) %>% summarise(counts = n())
Q2A3R3
Q2A3V3 <- ggplot(Q2A3R3, aes(avgGrade, y=counts, fill=as.factor(studytime))) + geom_bar(stat = "identity",width = 0.5, color="black") +
          ggtitle("Students that have scored > 15 average mark and not been joined \nextra-curricular activites grouped by their Study Time (Hours).") +
          theme(plot.title = element_text(size = 15, face = "bold")) +
          labs(fill = "Student Time (Hours)", x="Average Student Marks", y = "Student Counts")  +
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5)) +
          scale_fill_manual(values=c("#FF6347", "#FFA500", "#DAA520", "#7CFC00", "#2E8B57"))
Q2A3V3
```

Figure 88 *shows the R code used to create the data visualization figure of* ***Q2A3V3***

```
ggarrange(Q2A3V2, Q2A3V3, nrow = 2, ncol = 1)
```

Figure 89 *shows the R code used to arrange the data visualization figures of* ***Q2A3V2*** *and* ***Q2A3V3*** *in one view.*

As shown in the code figures above, for the bar graph, the **activities** and **avgGradeRange** are grouped and counted for this analysis. For the stacked bar graphs, the **studytime** was additionally added to the grouping. Moreover, the **activities** and **avgGrade** have been filtered out and counted the number of students based on the filtration.

```
> Q2A3R1
# A tibble: 8 x 3
# Groups:   activities [2]
  activities avgGradeRange counts
       <int> <chr>          <int>
1          1 00-05             47
2          1 06-10            174
3          1 11-15            203
4          1 16-20             39
5          2 00-05             40
6          2 06-10            176
7          2 11-15            204
8          2 16-20             39
```

Figure 90 *shows the output of the* ***Q2A3R1***.

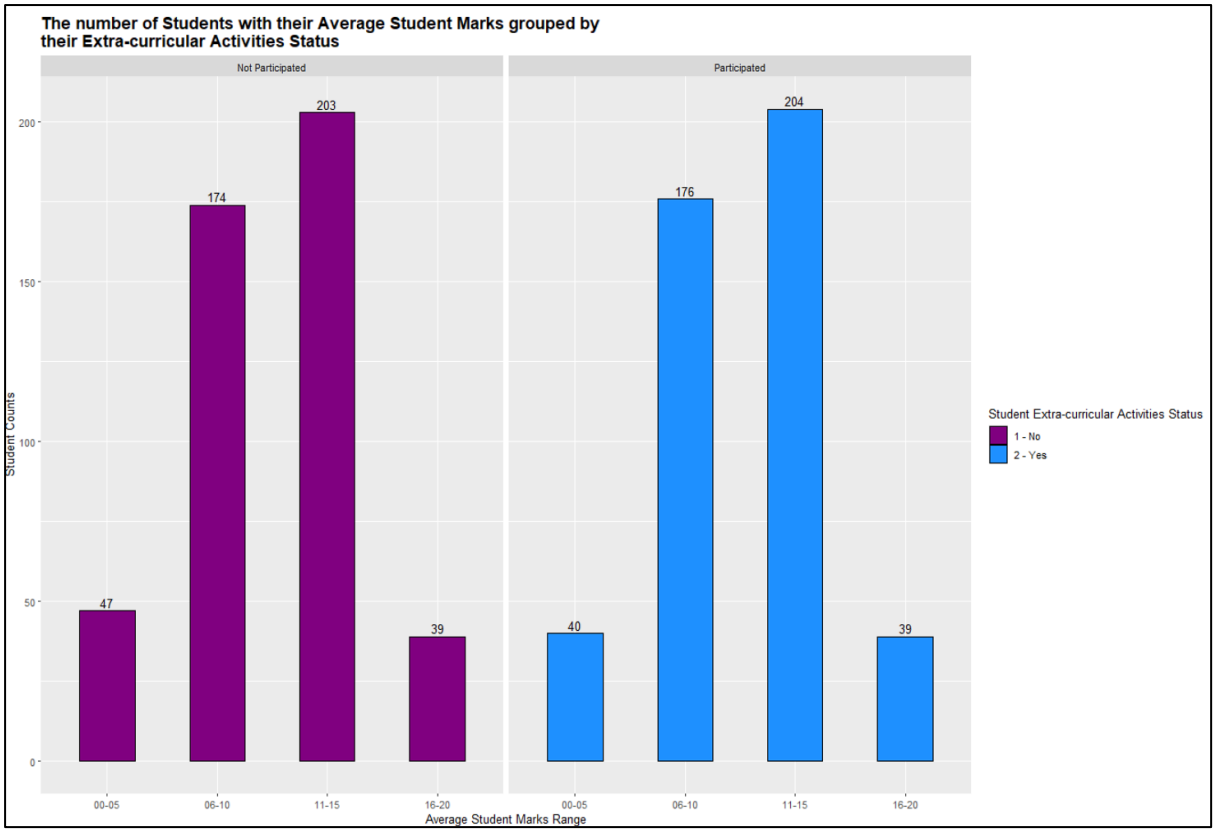


Figure 91 *shows the bar graph output of the* ***Q2A3V1***.

The figure 90 above displays the output of the execution of the **Q2A3R1**, which shows the grouped counts of total students for each case of selected attributes that are the extra-curricular activities and their average grade range. The figure 91 above shows the outcome of the bar graph plotted after the execution of the **Q2A3V1** variable that displays the students' counts and the average grade range of students grouped based on their extra-curricular activities participation.

```
> Q2A3R2
# A tibble: 11 x 4
# Groups:   activities, studytime [4]
   activities studytime avgGrade counts
        <int>     <int>    <dbl>  <int>
 1          2         1       16      2
 2          2         1       17      2
 3          2         1       18      4
 4          2         1       19      2
 5          2         2       16      5
 6          2         2       17      9
 7          2         2       18      3
 8          2         3       16      3
 9          2         3       17      4
10          2         4       17      2
11          2         4       19      3
```

***Figure 92** shows the output of the **Q2A3R2**.*

```
> Q2A3R3
# A tibble: 11 x 4
# Groups:   activities, studytime [4]
   activities studytime avgGrade counts
        <int>     <int>    <dbl>  <int>
 1          1         1       16      6
 2          1         1       18      4
 3          1         1       19      2
 4          1         2       16      4
 5          1         2       18      5
 6          1         3       16      3
 7          1         3       17      3
 8          1         3       18      2
 9          1         3       19      3
10          1         4       16      5
11          1         4       18      2
```

***Figure 93** shows the output of the **Q2A3R3**.*

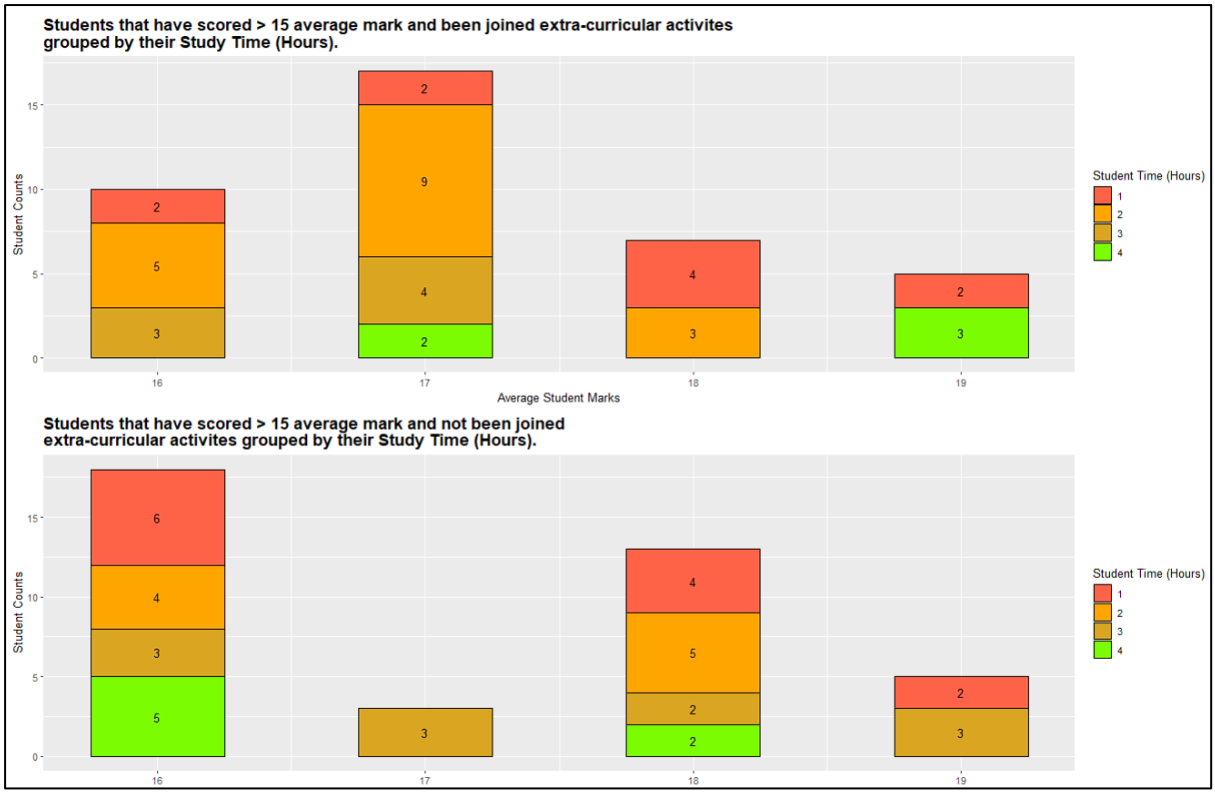


Figure 94 *shows the scatter plot graph output of the* ***Q2A3V2*** *and* ***Q2A3V3****.*

The figures 92 and 93 displays the execution output of the **Q2A3R2** and **Q2A3R3**, which displays the calculated total student counts grouped by the extra-curricular activities participation, amount of study time and average grade. In figure 94, the two stacked bar graph illustrates the number of students that scored more than 15 average marks and their status of participation in extra-curricular activities.

Summary for Data Findings

1) Only a total of 4 students is the difference between them participating and not participating in the extra-curricular activities.
2) The majority of students who got an average mark of more than 15 and participated in extra-curricular activities studied for 2 hours the most.
3) The highest number of students who scored 19 average marks and studied for 4 hours is the ones who participated in extra-curricular activities.
4) The highest number of students who scored 19 average marks and studied for 3 hours is the ones who didn't participate in extra-curricular activities.

Explanation for the Data Findings.

Based on the data discoveries, only a very minimal difference was found between the students' participation in extra-curricular activities, where the students who participated are the most with the slightest difference. As stated earlier, the largest number of students who got more than 15 average marks and joined extra-curricular activities studied for 2 hours. As the majority of the students that scored 19 average grades who participated in extra-curriculum activities studied for a total of 4 hours while those who didn't participate studied for 3 hours. It may be because the extra-curricular activities took some time from the students. Anyhow, from these data findings, it can be stated that the extra-curricular activities didn't impact the students' academic performance that much, as it doesn't have a big difference between the marks of those who participated and not participated in the activities. But still, these students who are more committed and attending extra-curricular activities would technically improve their overall academics and learn a lot of new skills. This was further supported by an article that participating in extra-curricular activities can increase brain function, improves the focus on studies and drive to do better in anything, which can assist students to score better results in their academic (Education Destination Malaysia, 2018).

4.2.4 Analysis 2-4: -Finding the correlation between students' study time, free time and their average grades.

The relationship between the students' extra-curricular activities participation and their average marks will be analysed. A bar graph and a table have been created for this analysis.

```
#Analysis 2-4
#Finding the correlation between students' study time, free time and their average grades.
Q2A4R1<- dsap_data %>%  group_by(freetime, studytime, avgGradeRange) %>%
        summarise(counts = n())
Q2A4R1

Q2A4V1<- ggplot(Q2A4R1, aes(x=avgGradeRange, y=counts, fill = as.factor(studytime))) +
        geom_bar(stat = "identity", position = position_dodge2(),width = 0.5, color="black") +
        ggtitle("The number of Students with their Average Student Marks grouped by\ntheir Study Time and Free Time") +
        theme(plot.title = element_text(size = 15, face = "bold")) +
        labs(fill = "Student Study Time (Hours)", x="Average Student Marks Range", y = "Student Counts")+
        facet_grid(studytime~freetime, labeller = labeller(.cols=label_both,.rows =label_both)) +
        geom_text(aes(label=counts), vjust=-0.3) +
        scale_fill_manual(values=c("#800080", "#F4A460", "#00BFFF", "#FFA500"))
Q2A4V1
```

Figure 95 *shows the R code used to create the data visualization figure of* ***Q2A4V1.***

```
Q2A4R2 <-  dsap_data %>% group_by(freetime, studytime) %>% filter(avgGrade>15) %>%
           summarise(counts = n())
View(Q2A4R2)
```

Figure 96 *shows the R code used to create the data visualization table figure of* ***Q2A4R2***

As shown in the code figures above, for the bar graph, the **freetime**, **studytime** and **avgGradeRange** are grouped and counted for this analysis. For the table created, it was grouped with **freetime** and **studytime** together with filtering the **avgGrade** have been filtered out and counted the number of students based on filtration.

```
Q2A4R1
# A tibble: 64 x 4
# Groups:   freetime, studytime [20]
   freetime studytime avgGradeRange counts
      <int>     <int> <chr>          <int>
 1        1         1 06-10              3
 2        1         1 11-15              2
 3        1         2 00-05              5
 4        1         2 06-10             11
 5        1         2 11-15              7
 6        1         2 16-20              2
 7        1         3 06-10              2
 8        1         3 11-15              4
 9        1         4 06-10              4
10        1         4 16-20              2
```

Figure 97 *shows the output of the* ***Q2A4R1****.*

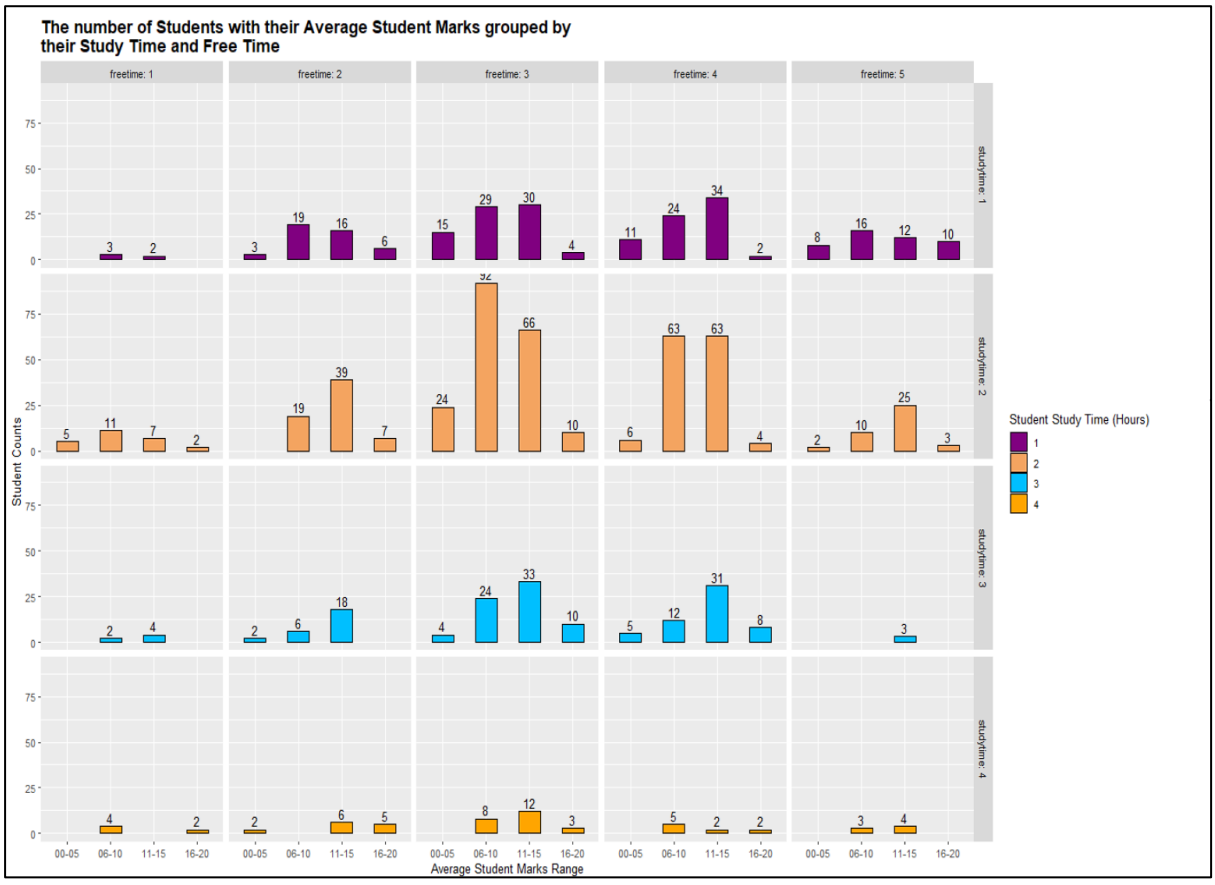


Figure 98 *shows the bar graph output of the* ***Q2A4V1.***

freetime	studytime	counts
1	2	2
1	4	2
2	1	6
2	2	7
2	4	5
3	1	4
3	2	10
3	3	10
3	4	3
4	1	2
4	2	4
4	3	8
4	4	2
5	1	10
5	2	3

Figure 99** shows the table output of the **Q2A4R2.

The figure 97 above displays the result of the execution of the **Q2A4R1**, which shows the grouped counts of total students for each case of selected attributes that are the free time, study time and their average grade range. The figure 98 above shows the outcome of the bar graph plotted after the execution of the **Q2A4V1** variable that displays the students' counts and the average grade range of students grouped based on their study time and free time. The figure shows the output of the execution of **Q2A4R2,** where it displays the counts of students that scored more than 15 average marks grouped with their free time and study time.

Summary for Data Findings

1) The largest portion of students had a reasonable amount of study time and free time.
2) The most number of students had 3 hours of study time and 2 hours free time.
3) The largest portion of students who got more than 15 average marks had a fair amount of study time and free time.
4) Three categories of groupings had the most students who scored more than 15 average marks that are the students who had (3 hours of Free Time, 2 hours of Study Time), (3 of Free Time, 3 of Study Time), and (5 hours of Free Time, 1 hours of Study Time).

Explanation for the Data Findings.

Based on the data findings, it can be stated that a big number of students have free time of 2 hours and study time of 3 hours. When analysing the students that got more than 15 average marks, the most count of students were grouped into three different categories as their student counts were the same, that was a total of 10 of them. From the three categories, it can be seen that a lot of the students managed and spent their time properly in order to balance their personal and school life. As it is seen that, the majority of these students used their time fairly in studying and spending their free time. This clearly tells us that these students knew the importance of managing their time, and they practise it very pleasingly in their both personal and study lives. From the students who scored excellent grades and spent their time fairly, on both study time and free time, it can be assumed that practising adequate time management can affect students' academic performance. According to Adams and Blair (2019), they further enforced in their research that students who manage their time effectively can help to achieve greater academic achievements. Hence, it is true that managing and spending time effectively can affect the students' overall academic performance.

Conclusion

To sum up this second question, it was observed from those analyses that a student's time management skills play a very vital role in how a student conduct things in their daily life. They need to balance out their studies and other activities to achieve great academic grades and together with that also have fun with their life. This is basically on the hands of the students on how they make them disciplined and maintained in practising good time management. Thus, the respective school representatives, parents and the students themselves should do their part to instil a solid understanding of why time management is so important to them so that all of them could do better in their academics.

4.3 Question 3: How do resource availability and the comfort impact students' marks?

This question will be analysing whether the availability of the resource and convenience to a student would impact their overall academic grades. This will assist to figure out those attributes that can assist a student to improve their grades. The student's attributes that will be covered in this question are extra paid classes, extra educational support from the school, family educational support, travel time and internet access.

4.3.1 Analysis 3-1: Finding the relationship between students' joining additional paid classes and their average grades

The relationship between the students' joining extra paid classes and their average marks will be analysed. A horizontal bar graph, two treemap graphs and two pie charts and a table have been created for this analysis.

```
#=======Question 3========#
#Question 3: How do resource availability and the comfort impact students' marks?
#Analysis 3-1
#Finding the relationship between students' joining additional paid classes and their average grades.
Q3A1R1 <- dsap_data %>% group_by(paid, avgGradeRange) %>% summarise(counts = n())
Q3A1R1
Q3A1V1 <-  ggplot(Q3A1R1, aes(avgGradeRange, y=counts, fill = as.factor(paid))) + geom_bar(stat = "identity", width=0.9) +
           ggtitle("The number of Students with their average score grouped by their joining Paid Additional Classes status.") +
           theme(plot.title = element_text(size = 15, face = "bold")) +
           labs(fill = "Student Paid Additional Classes", x="Average Students Marks Range", y = "Student Counts")+
           geom_text(aes(label=counts), position = position_dodge2(1), hjust = -0.1) + ylim(0,250) +
           scale_fill_manual(values = c("#B22222","#7CFC00"),labels = c("1 - No", "2 - Yes")) +
           coord_flip() + facet_wrap(~paid, labeller = as_labeller(c(`2`="Joined", `1`="Not Joined")))
Q3A1V1
```

Figure 100 *shows the R code used to create the data visualization figure of* ***Q3A1V1.***

```
Q3A1R2 <- dsap_data %>% group_by(paid) %>% filter(avgGrade>avgMeanGrade) %>%
          summarise(counts = n(), percentage = n()/length(which(dsap_data$avgGrade>avgMeanGrade))*100)
Q3A1R2
Q3A1V2 <- ggplot(Q3A1R2, aes(x=percentage, y="", fill = as.factor(paid), area = percentage))+ geom_treemap()+
          labs(title = "") +  theme(panel.background = element_blank(),
                                    axis.title = element_blank(),
                                    axis.text = element_blank(),
                                    axis.line = element_blank(),
                                    axis.ticks= element_blank(),
          plot.title = element_text(size = 20, face = "bold")) +
          geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")), color = c("white"), place = "centre") +
          ggtitle("Shows the percentage of the Students that scored\n> average mark and their status of attending for\npaid Additional Class.") +
          labs(fill="Student Paid Additional Classes")+ scale_fill_manual(values = c("#4B0082","#DA70D6"), labels = c("1 - No", "2 - Yes"))
Q3A1V2
```

Figure 101 *shows the R code used to create the data visualization figure of* ***Q3A1V2.***

```
Q3A1R3 <- dsap_data %>% group_by(paid) %>% filter(avgGrade<avgMeanGrade) %>%
          summarise(counts = n(), percentage = n()/length(which(dsap_data$avgGrade<avgMeanGrade))*100)
Q3A1R3
Q3A1V3 <- ggplot(Q3A1R3, aes(x=percentage, y="", fill = as.factor(paid), area = percentage)) +geom_treemap()+
          labs(title = "") +  theme(panel.background = element_blank(),
                                    axis.title = element_blank(),
                                    axis.text = element_blank(),
                                    axis.line = element_blank(),
                                    axis.ticks= element_blank(),
          plot.title = element_text(size = 20, face = "bold")) +
          geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")), color = c("white"), place = "centre") +
          ggtitle("Shows the percentage of the Students that scored\n< average mark and their status of attending for\npaid Additional Class.") +
          labs(fill="Student Paid Additional Classes")+ scale_fill_manual(values = c("#4B0082","#DA70D6"), labels = c("1 - No", "2 - Yes"))
Q3A1V3
```

Figure 102 *shows the R code used to create the data visualization figure of* ***Q3A1V3.***

```
ggarrange(Q3A1V2, Q3A1V3, nrow = 2, ncol = 1)
```

Figure 103 *shows the R code used to arrange the data visualization figures of* ***Q3A1V2*** *and* ***Q3A1V3*** *in one view.*

```
Q3A1R4 <- dsap_data %>% group_by(paid)%>%
        filter(avgGrade>15) %>% summarise(counts = n(), percentage = n()/length(which(dsap_data$avgGrade>15))*100)
Q3A1R4
Q3A1V4 <- ggplot(Q3A1R4, aes(x="", y =percentage, fill=as.factor(paid))) + geom_col(color = "black") + coord_polar("y", start = 0) +
        theme(panel.background = element_blank(),
              axis.title = element_blank(),
              axis.text = element_blank(),
              axis.line = element_blank(),
              axis.ticks= element_blank(),
        plot.title = element_text(size = 20, face = "bold")) +geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                                                                        color = c("white"), position = position_stack(vjust=0.5)) +
        ggtitle("Shows the percentage of the students that scored\n> 15 marks and their status of attending for\npaid Additional Class.") +
        labs(fill="Paid Additional Class Status")+ scale_fill_manual(values = c("#191970","#FF1493"),
                                                                    labels=c("1 - No", "2 - Yes"))

Q3A1V4
```

Figure 104 *shows the R code used to create the data visualization figure of* ***Q3A1V4.***

```
Q3A1R5 <- dsap_data %>% group_by(paid) %>% filter(avgGrade<=5) %>%
        summarise(counts = n(), percentage = n()/length(which(dsap_data$avgGrade<=5))*100)
Q3A1R5
Q3A1V5 <- ggplot(Q3A1R5, aes(x="", y =percentage, fill=as.factor(paid))) + geom_col(color = "black") + coord_polar("y", start = 0) +
        theme(panel.background = element_blank(),
              axis.title = element_blank(),
              axis.text = element_blank(),
              axis.line = element_blank(),
              axis.ticks= element_blank(),
        plot.title = element_text(size = 20, face = "bold")) +geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                                                                        color = c("white"), position = position_stack(vjust=0.5)) +
        ggtitle("Shows the percentage of the students that scored\n<= 5 marks and their status of attending for\npaid Additional Class.") +
        labs(fill="Paid Additional Class Status")+scale_fill_manual(values = c("#191970","#FF1493"),
                                                                    labels=c("1 - No", "2 - Yes"))

Q3A1V5
```

Figure 105 *shows the R code used to create the data visualization figure of* ***Q3A1V5.***

```
ggarrange(Q3A1V4, Q3A1V5, nrow = 2, ncol = 1)
```

Figure 106 *shows the R code used to arrange the data visualization figures of* ***Q3A1V4*** *and* ***Q3A1V5*** *in one view.*

As shown in the code figures above, for the horizontal bar graph, the **paid** and **avgGradeRange** are grouped and counted for this analysis. While for the treemap and pie graphs, the **paid** was grouped and filtered based on their **avgGrade** and the percentage was calculated based on the count of the students. To create a treemap graph, the "geom_treemap()" function was added to make the graph as a treemap.

```
> Q3A1R1
# A tibble: 8 x 3
# Groups:   paid [2]
   paid avgGradeRange counts
  <int> <chr>           <int>
1     1 00-05              70
2     1 06-10             179
3     1 11-15             198
4     1 16-20              57
5     2 00-05              17
6     2 06-10             171
7     2 11-15             209
8     2 16-20              21
```

Figure 107 *shows the output of the* ***Q3A1R1****.*

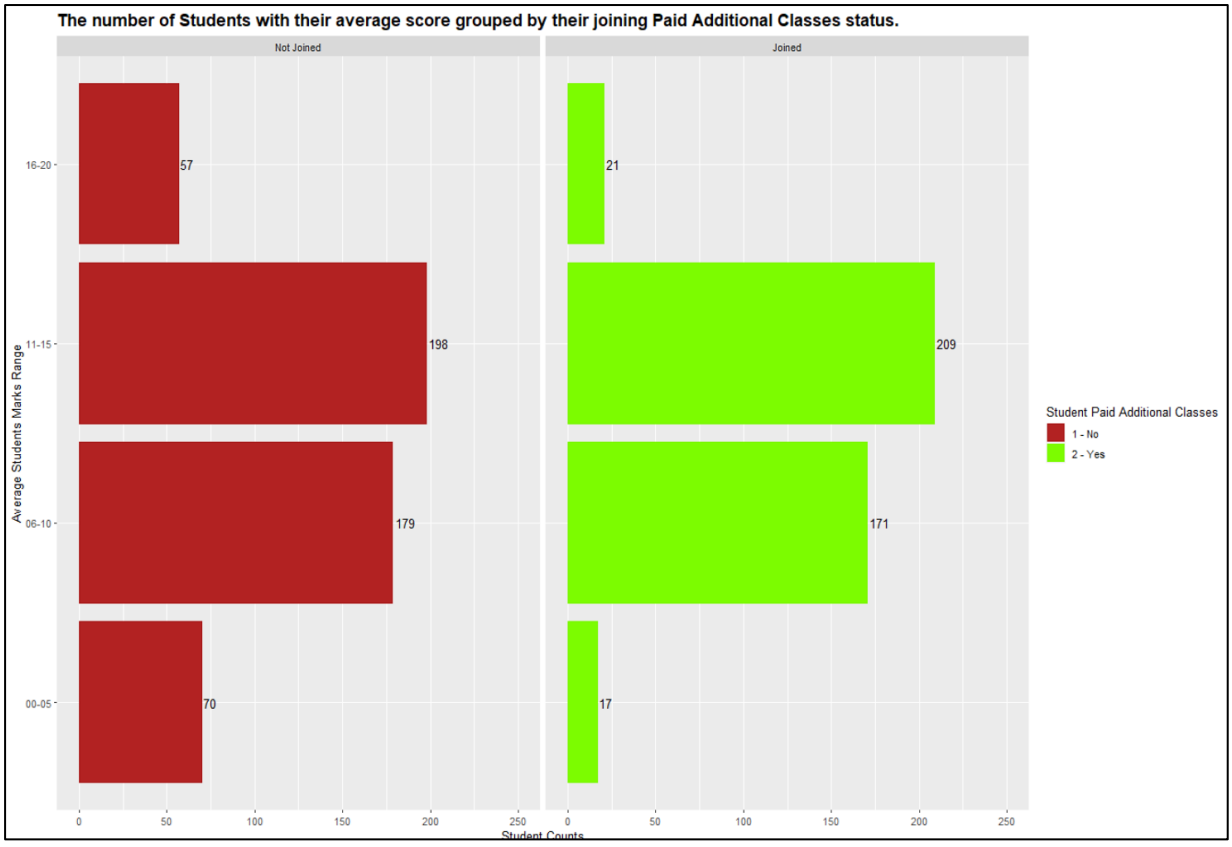


Figure 108 *shows the horizontal bar graph output of the* ***Q3A1V1****.*

The figure 107 above shows the result of the execution of the **Q3A1R1**, which shows the grouped counts of total students for each case of selected attributes that are the extra paid classes and their average grade range. The figure 108 above shows the outcome of the horizontal bar graph plotted after the execution of the **Q3A1V1** variable that displays the students' counts and the average grade range of students grouped based on their paid additional class status.

```
> Q3A1R2
# A tibble: 2 x 3
  paid counts percentage
  <int>  <int>      <dbl>
1     1    212       54.1
2     2    180       45.9
```

Figure 109 *shows the output of the* ***Q3A1R2****.*

```
> Q3A1R3
# A tibble: 2 x 3
  paid counts percentage
  <int>  <int>      <dbl>
1     1    249       57.0
2     2    188       43.0
```

Figure 110 *shows the output of the* ***Q3A1R3****.*

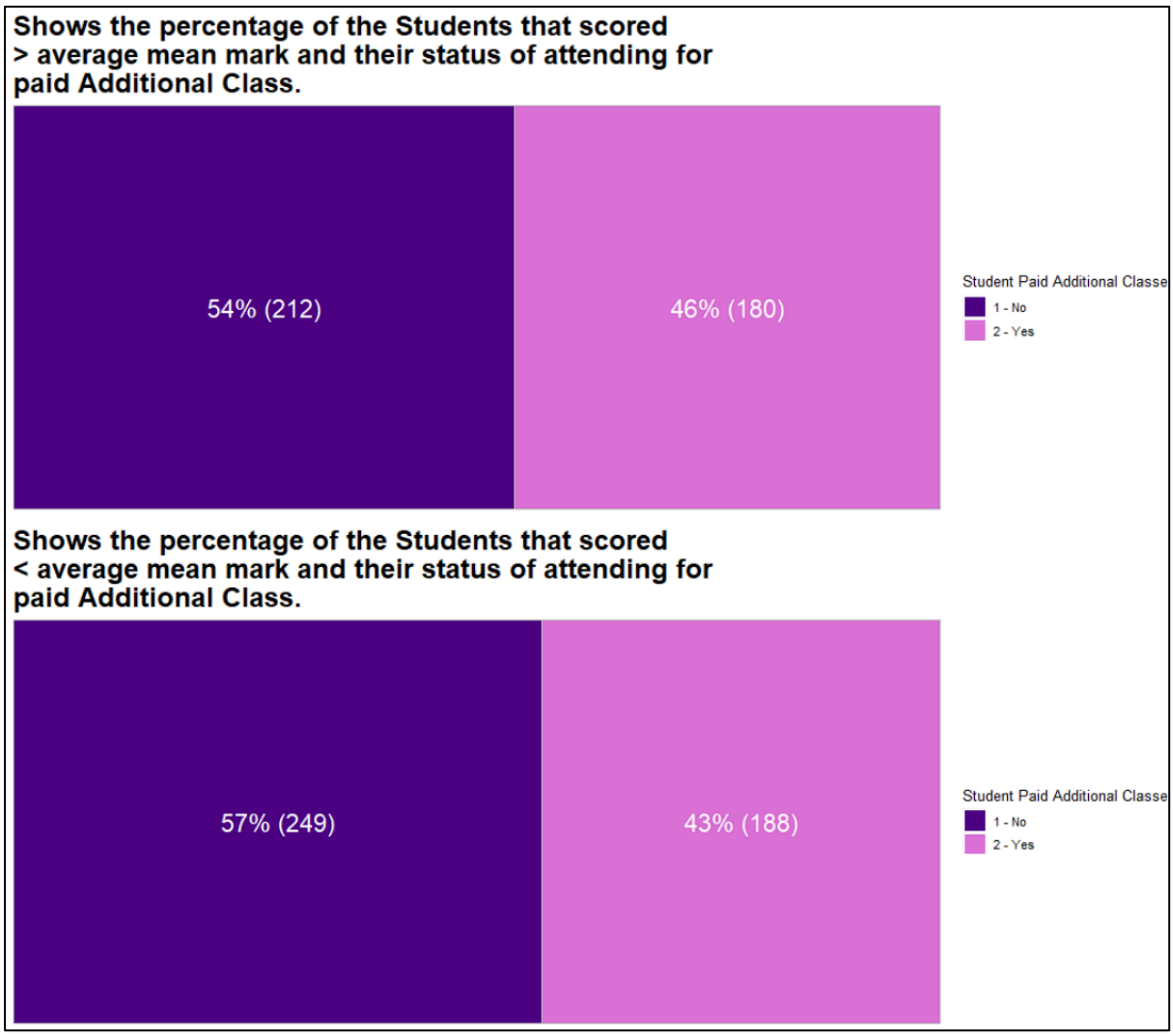


Figure 111 *shows the three map charts of the output of* ***Q3A1V2*** *and* ***Q1A1V3****.*

The figures 109, and 110 displays the execution output of the **Q3A1R2**, and **Q3A1R3**, which display the calculated total counts and percentage of students grouped by the extra paid class status. In figure 111, the two treemap graphs are displaying the percentage of students that scored more than the average mean marks and less than the average mean mark and their status of joining paid additional classes.

```
> Q3A1R4
# A tibble: 2 x 3
   paid counts percentage
  <int>  <int>      <dbl>
1     1     57       73.1
2     2     21       26.9
>
```

Figure 112 *shows the output of the* ***Q3A1R4****.*

```
> Q3A1R5
# A tibble: 2 x 3
   paid counts percentage
  <int>  <int>      <dbl>
1     1     70       80.5
2     2     17       19.5
>
```

Figure 113 *shows the output of the* ***Q3A1R5****.*

The figures 112 and 113 show the execution output of the **Q3A1R4** and **Q3A1R3**, which displays the calculated total student counts and percentage grouped by their additional paid classes status. In figure 114 below, the two pie charts that have been plotted display the percentage of students that scored more than 15 average marks and less than and equal to 5 average marks, grouped based on their paid extra class status.

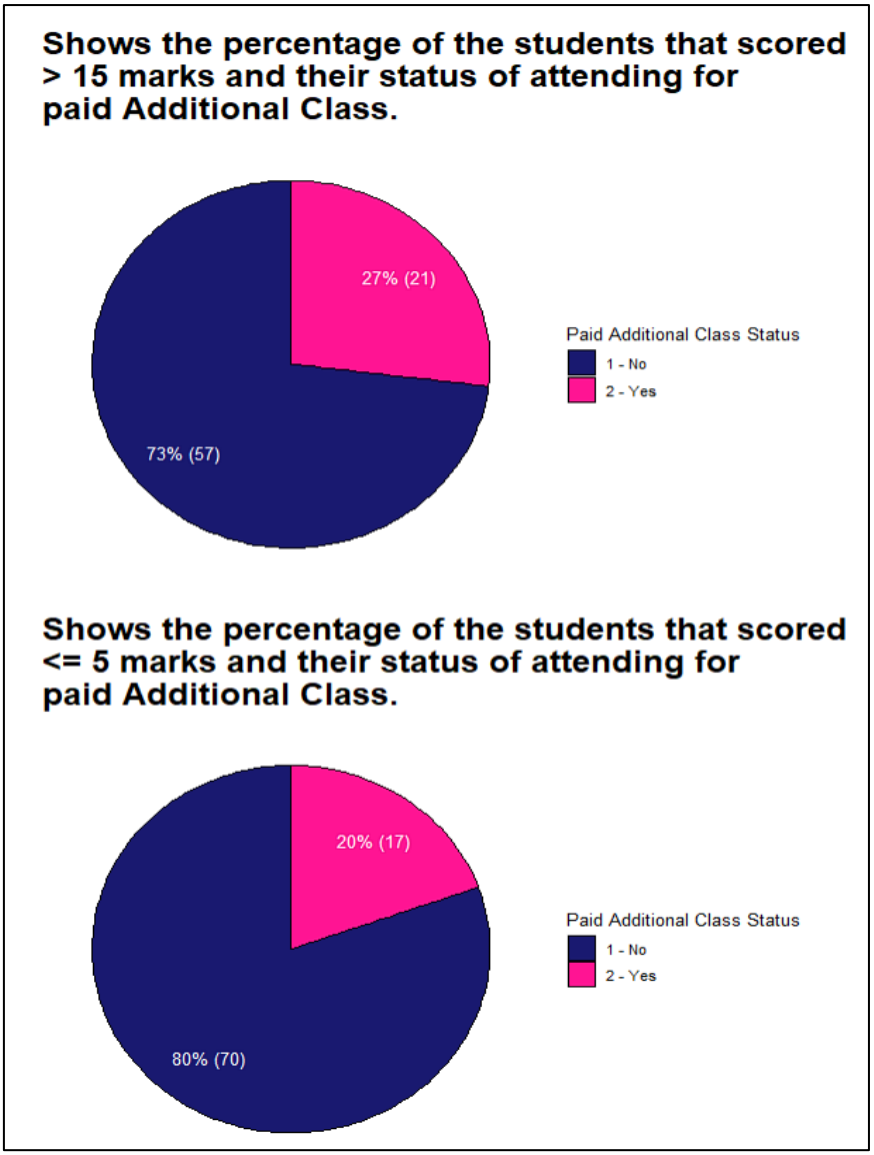


***Figure 114** shows the pie charts of the output of **Q3A1V4** and **Q1A1V5**.*

Summary for Data Findings

1. A total of 504 students had joined for extra paid classes, while the rest of the 418 students had not joined one.
2. In both cases, the largest portion of students in the average mark range of 11-15.
3. The percentage of students who scored more than the average mean mark and had not joined paid extra classes was 54%, that is about 212 of them, while the other who joined was 46%, that is 180 total of them.
4. A large percentage of students had not joined for extra paid classes who scored more than 15 and less than equal to 5 average marks.

Explanation for the Data Findings.

Based on the data analysis, it can be noticied that a larger percentage of the total students had not joined the extra paid classes. It was seen when filtering with average mean marks. Moreover, it was further can be seen when filtering the average score of more than 15 and less than equal to 5 where the majority of these students and all the excellent grades did not have much of them didn't join for the extra paid class. From this, it can be stated that extra paid classes solely may not affect the students' marks. As stated before, even when they joined the extra paid classes, without the students' effort, it won't help them achieve better grades. It is all in the students' hands on how they utilise these extra paid classes to help them with their studies.

4.3.2 Analysis 3-2: Finding the relationship between students' additional educational support from school and their average grades

The correlation between the students' extra educational support from the school and their average marks will be analysed. A scatterplot graph, three treemap graphs and a table have been created for this analysis.

```
#Analysis 3-2
#Finding the relationship between students' additional educational support from the school and their average marks.
Q3A2R1<- dsap_data %>% group_by(schoolsup, avgGradeRange) %>%  summarise(counts = n())
Q3A2R1
Q3A2V1<- ggplot(Q3A2R1, aes(x = avgGradeRange, y=counts, color= as.factor(schoolsup))) + geom_point(stat="identity",alpha = 0.8) +
         ggtitle("The number of Students with their average score grouped\nby their Extra Education Support status.") +
         theme(plot.title = element_text(size = 15, face = "bold")) +
         labs(x="Average Students Marks Range", y = "Student Counts")+
         geom_text(aes(label=counts), vjust=-1.0)  + ylim(0, 450) +
         scale_color_manual(name = "Student Additional Educational Support\n from school status",
                            values = c(`1` = "black",`2` = "red" ), labels=c("1 - No", "2 - Yes"))

Q3A2V1
```

Figure 115 *shows the R code used to create the data visualization figure of* ***Q3A2V1.***

```
Q3A2R2<-  dsap_data %>% group_by(schoolsup)%>%
          filter(G1<=5) %>% summarise(counts = n(), percentage = n()/length(which(dsap_data$G1<=5))*100)
Q3A2R2
Q3A2V2 <-  ggplot(Q3A2R2, aes(x=percentage, y="", fill = as.factor(schoolsup), area = percentage)) +
           geom_treemap()+
           labs(title = "") +  theme(panel.background = element_blank(),
                                     axis.title = element_blank(),
                                     axis.text = element_blank(),
                                     axis.line = element_blank(),
                                     axis.ticks= element_blank(),
           plot.title = element_text(size = 20, face = "bold")) +
           geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                             color = c("white"), place = "centre") +
           ggtitle("Shows the percentage of the Students that scored\n<= 5 marks in their G1,
                    grouped by\nStudent Additional Educational Support from school status") +
           labs(fill="Student Additional Educational Support\nfrom school status")+
          scale_fill_manual(values = c("#7FFF00","#008B8B"), labels = c("1 - No", "2 - Yes"))
Q3A2V2
```

Figure 116 *shows the R code used to create the data visualization figure of* ***Q3A2V2.***

```
Q3A2R3<-  dsap_data %>% group_by(schoolsup)%>%
          filter(G2<=5) %>% summarise(counts = n(), percentage = n()/length(which(dsap_data$G2<=5))*100)
Q3A2R3
Q3A2V3 <-  ggplot(Q3A2R3, aes(x=percentage, y="", fill = as.factor(schoolsup), area = percentage)) +
           geom_treemap()+
           labs(title = "") +  theme(panel.background = element_blank(),
                                     axis.title = element_blank(),
                                     axis.text = element_blank(),
                                     axis.line = element_blank(),
                                     axis.ticks= element_blank(),
           plot.title = element_text(size = 20, face = "bold")) +
           geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                             color = c("white"), place = "centre") +
           ggtitle("Shows the percentage of the Students that scored\n<= 5 marks in their G2,
                    grouped by\nStudent Additional Educational Support from school status") +
           labs(fill="Student Additional Educational Support\nfrom school status")+
          scale_fill_manual(values = c("#7FFF00","#008B8B"), labels = c("1 - No", "2 - Yes"))
Q3A2V3
```

Figure 117 *shows the R code used to create the data visualization figure of* ***Q3A2V3.***

```
Q3A2R4<-  dsap_data %>% group_by(schoolsup)%>%
          filter(G3<=5) %>% summarise(counts = n(), percentage = n()/length(which(dsap_data$G3<=5))*100)
Q3A2R4
Q3A2V4 <-  ggplot(Q3A2R4, aes(x=percentage, y="", fill = as.factor(schoolsup), area = percentage)) +
           geom_treemap()+
           labs(title = "") +  theme(panel.background = element_blank(),
                                     axis.title = element_blank(),
                                     axis.text = element_blank(),
                                     axis.line = element_blank(),
                                     axis.ticks= element_blank(),
           plot.title = element_text(size = 20, face = "bold")) +
           geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                             color = c("white"), place = "centre") +
           ggtitle("Shows the percentage of the Students that scored <= 5 marks in their G3,
                   grouped by Student Additional Educational Support from school status") +
           labs(fill="Student Additional Educational Support\nfrom school status")+
            scale_fill_manual(values = c("#7FFF00","#008B8B"), labels = c("1 - No", "2 - Yes"))
Q3A2V4
```

Figure 118 *shows the R code used to create the data visualization figure of* ***Q3A2V4.***

```
ggarrange(Q3A2V2, Q3A2V3, Q3A2V4, nrow = 3, ncol = 1)
```

Figure 119 *shows the R code used to arrange the data visualization figures of* ***Q3A2V2*** ***Q3A2V3*** *and* ***Q3A2V4*** *in one view.*

```
Q3A2R5<-  dsap_data %>% group_by(schoolsup)%>% select(G1,G2,G3) %>%
          filter(G1<=5, schoolsup == "2")
View(Q3A2R5)
```

Figure 120 *shows the R code used to create the data visualization table figure of* ***Q3A2R5.***

As shown in the code figures above, for the scatterplot graph, the **schoolsup** and **avgGradeRange** are grouped and counted for this analysis. While for the treemap graphs, the **schoolsup** was grouped and filtered based on their **G1**, **G2** and **G3** marks accordingly and the percentage was calculated based on the count of the students. For the table created, it was grouped with **schoolsup** as well as the **G1** mark and **schoolsup** status have been filtered.

```
> Q3A2R1
# A tibble: 8 x 3
# Groups:   schoolsup [2]
  schoolsup avgGradeRange counts
      <int> <chr>          <int>
1         1 00-05             85
2         1 06-10            271
3         1 11-15            376
4         1 16-20             76
5         2 00-05              2
6         2 06-10             79
7         2 11-15             31
8         2 16-20              2
```

Figure 121 *shows the output of the* ***Q3A2R1.***

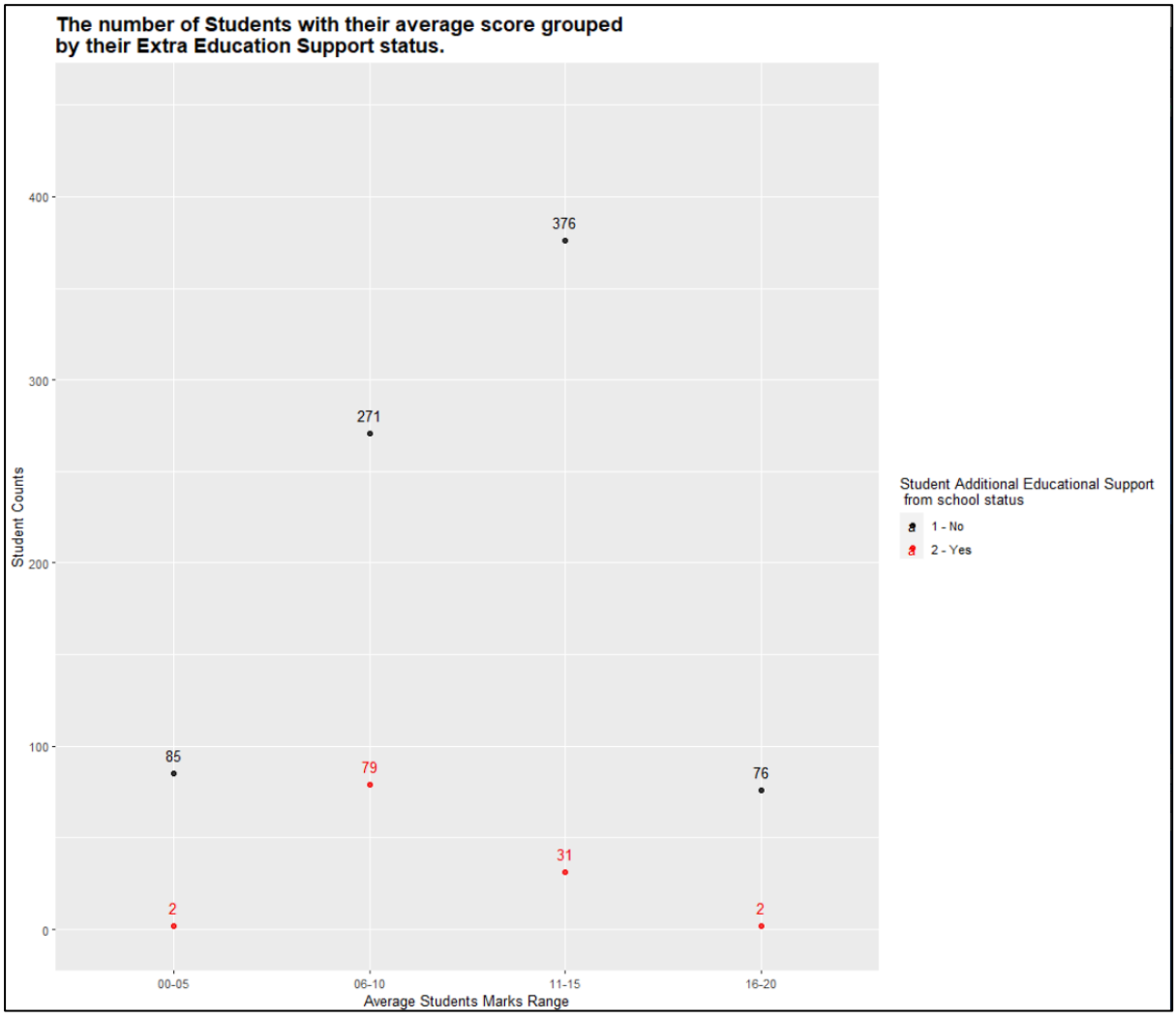


Figure 122 *shows the scatterplot graph of the output of* ***Q3A2V1****.*

The figure 121 above displays the output of the execution of the **Q3A2R1**, which shows the grouped counts of total students for each case of selected attributes that are the extra educational support and their average grade range. The figure 122 above shows the outcome of the scatterplot graph plotted after the execution of the **Q3A2V1** variable that displays the students' counts and the average grade range of students grouped based on their extra educational support joined status.

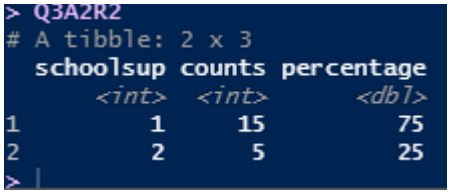


Figure 123 *shows the output of the* ***Q3A2R2.***

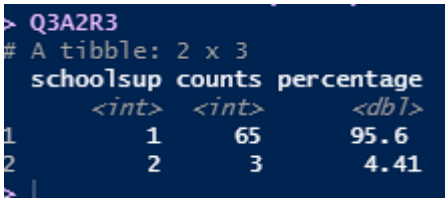


Figure 124 *shows the output of the* ***Q3A2R2.***

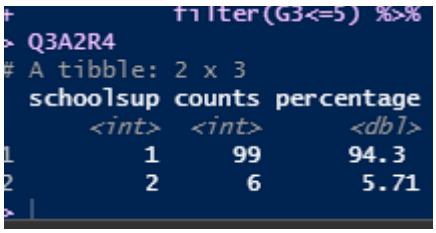


Figure 125 *shows the output of the* ***Q3A2R4.***

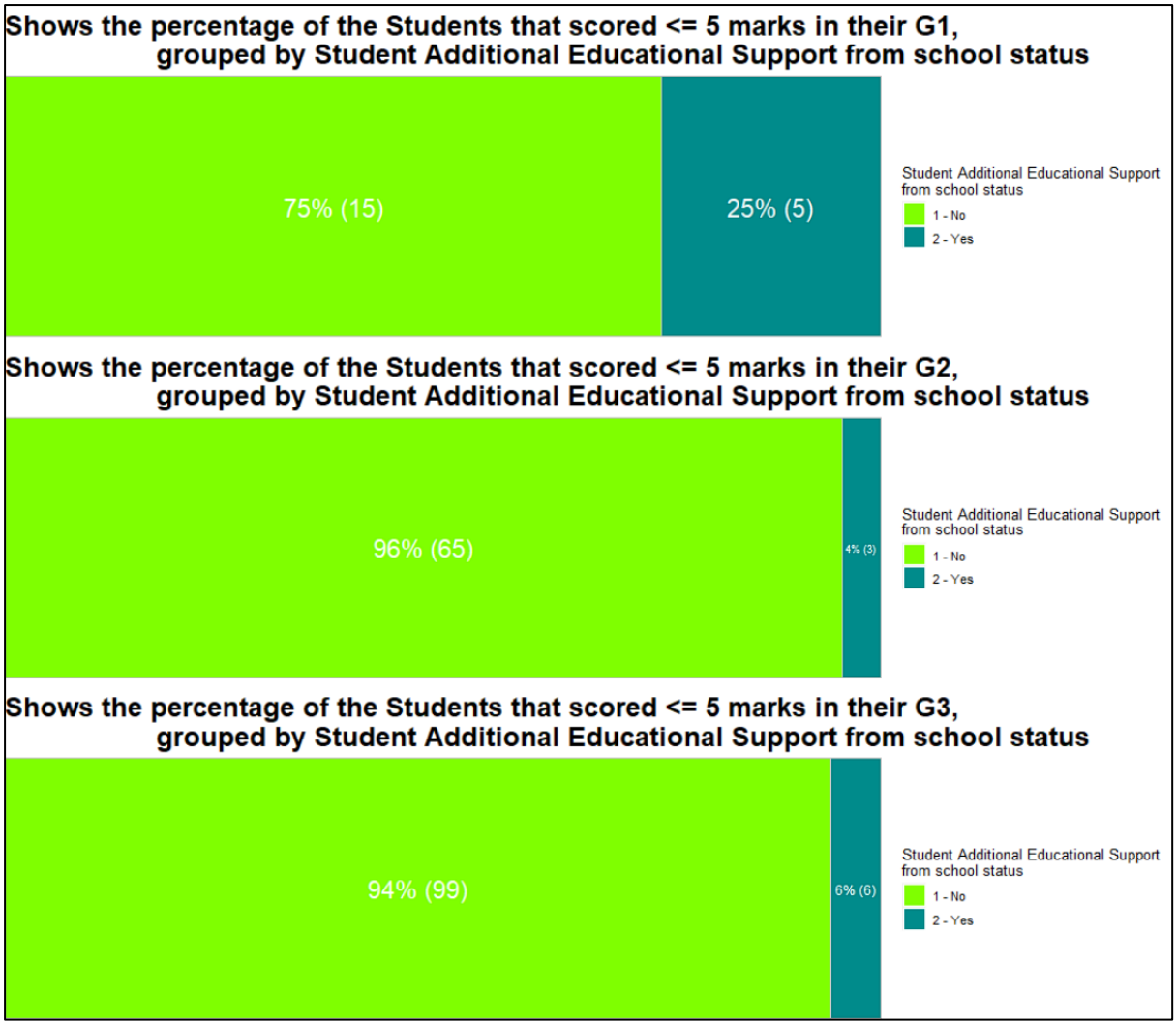


Figure 126 *shows the treemap graph of the output of* ***Q3A2V2, Q3A2V3*** *and* ***Q3A2V4****.*

	schoolsup	G1	G2	G3
1	2	5	6	6
2	2	5	9	7
3	2	5	6	6
4	2	5	9	7
5	2	5	6	6

Figure 127 *shows the table output of the* ***Q3A2R5.***

The figures 123, 124, and 125 displays the execution output of the **Q3A2R2**, **Q3A2R3** and **Q3A2R2**, which display the calculated total counts and percentage of students grouped by the extra educational support status. In figure 114, the three treemap graphs are displaying the percentage of students that scored less than and equal to 5 marks in their G1, G2 and G3 period test accordingly and their status of joining extra educational support. The table in figure 127 above shows the mark progression of students who scored less than equal to 5 marks in their G1 test who had joined the extra educational support.

Summary for Data Findings

1) The largest number of students that are 808 total of them didn't have extra educational support, while the rest of the 114 students had one.
2) 75% of students that scored less than equal to 5 average marks in G1 are the ones who didn't have extra educational support.
3) 96% of students that scored less than equal to 5 average marks in G2 are the ones who didn't have extra educational support.
4) 94% of students that scored less than equal to 5 average marks in G3 are the ones who didn't have extra educational support.
5) 25% of students that scored less than equal to 5 average marks in G1 had the extra educational support
6) 4% of students that scored less than equal to 5 average marks in G2 had the extra educational support
7) 6% of students that scored less than equal to 5 average marks in G1 had the extra educational support
8) The number of students who scored less than equal to 5 average marks in G2 that had extra educational support decreased from 25% to 4% compared to G1.
9) The number of students who scored less than equal to 5 average marks in G3 that had additional educational support increased from 4% to 6% compared to G2.

Explanation for the Data Findings.

Based on the data findings, it is absolutely clearly seen that most of the students did not have to join the extra educational support from the school. There was very a minimal amount of those who have joined compared to those who didn't. As per the result shown, the number of students for each period test who scored less than equal to 5 marks is 5 (G1), 3 (G2) and 6 (G3) of them accordingly. When seeing the 5 students who had joined the extra educational support and scored less than and equal to 5 in the G1 period, all these 5 of them have improved when comparing their marks in their G2 and G3 marks. It is not a big improvement, but at least they did better than the first-period test, which is considered a good progression. As it can be said that getting extra educational support from school can benefit students in improving their grades. This could be because during this extra educational support the students who have doubts on certain topics can use that time to ask the teachers in order clear the doubts. Also, usually, during these extra educational support sessions, the teachers also give tips on how to score certain topics in a much simpler way which really does assist the students to score better**.** Hence, extra educational support can help students to improve their grades.

4.3.3 Analysis 3-3: Finding the relationship between students' family educational support and their average marks.

The correlation between the students' educational support from family and their average marks will be analysed. A horizontal bar graph and two bar graphs have been created for this analysis.

```
#Analysis 3-3
#Finding the relationship between students' family educational support and their average marks.
Q3A3R1<- dsap_data %>% group_by(famsup,avgGradeRange) %>% summarise(counts = n())
Q3A3R1
Q3A3V1<- ggplot(Q3A3R1, aes(x=avgGradeRange, y=counts, fill=as.factor(famsup))) +
         geom_bar(stat="identity",width = 0.5, color="white") +
         ggtitle("The number of Students with their average score grouped
                  by their Family Educational Support status.")+
         labs(x="Average Student Marks Range", y = "Student Counts",
              fill="Family Educational Support status")+
         theme(plot.title = element_text(size = 15, face = "bold")) +
         scale_fill_manual(values=c("#2F4F4F", "#008080"),
         labels = c("1 - No", "2 - Yes")) + coord_flip() +
         geom_text(aes(label=counts), position = position_stack(vjust = 0.5),
                   color = "white")
Q3A3V1
```

Figure 128 *shows the R code used to create the data visualization figure of* ***Q3A3V1.***

```
Q3A3R2 <- dsap_data %>% group_by(famsup, avgGradeRange) %>%
          filter(avgGrade>avgMeanGrade) %>% summarise(counts = n())
Q3A3R2
Q3A3V2 <- ggplot(Q3A3R2, aes(x=avgGradeRange, y=counts, fill = as.factor(famsup))) +
          geom_bar(stat = "identity", position=position_dodge2(), width = 0.5, color="black") +
          ggtitle("The number of Students with their Average Student Marks > average mean marks grouped
                   by their Family Educational Suppport Status") +
          theme(plot.title = element_text(size = 15, face = "bold")) +
          labs(fill = "Family Educational Suppport Status", x="Average Student Marks Range",
               y = "Student Counts")+
          facet_wrap(~famsup, labeller = as_labeller(c(`2`="Got Support", `1`="No Support"))) +
          geom_text(aes(label=counts), vjust=-0.3) +
          scale_fill_manual(values = c("#00FFFF", "#191970"),labels = c("1 - No", "2 - Yes"))
Q3A3V2
```

Figure 129 *shows the R code used to create the data visualization figure of* ***Q3A3V2.***

```
Q3A3R3 <- dsap_data %>% group_by(famsup, avgGrade) %>%
          filter(avgGrade>15) %>% summarise(counts = n())
Q3A3R3
Q3A3V3 <-  ggplot(Q3A3R3, aes(x=avgGrade, y=counts, fill = as.factor(famsup))) +
           geom_bar(stat = "identity", position=position_dodge2(), width = 0.5, color="black") +
           ggtitle("The number of Students with their Average Student Marks > 15 marks
                            grouped by their Family Educational Suppport Status") +
           theme(plot.title = element_text(size = 15, face = "bold")) +
           labs(fill = "Family Educational Suppport Status", x="Average Student Marks Range",
                y = "Student Counts")+
           facet_wrap(~famsup, labeller = as_labeller(c(`2`="Got Support", `1`="No Support"))) +
           geom_text(aes(label=counts), vjust=-0.3) +
           scale_fill_manual(values = c("#00FFFF", "#191970"),labels = c("1 - No", "2 - Yes"))

Q3A3V3
```

Figure 130 *shows the R code used to create the data visualization figure of* ***Q3A3V3.***

As shown in the code figures above, for the horizontal bar graph, the **famsup** and **avgGradeRange** are grouped and counted for this analysis. While for the other bar graphs, the **schoolsup** was grouped and filtered based on their **avgGrade** marks accordingly.

```
> Q3A3R1
# A tibble: 8 x 3
# Groups:   famsup [2]
  famsup avgGradeRange counts
   <int> <chr>          <int>
1      1 00-05             37
2      1 06-10            120
3      1 11-15            169
4      1 16-20             32
5      2 00-05             50
6      2 06-10            230
7      2 11-15            238
8      2 16-20             46
```

Figure 131 *shows the output of the* ***Q3A3R1.***

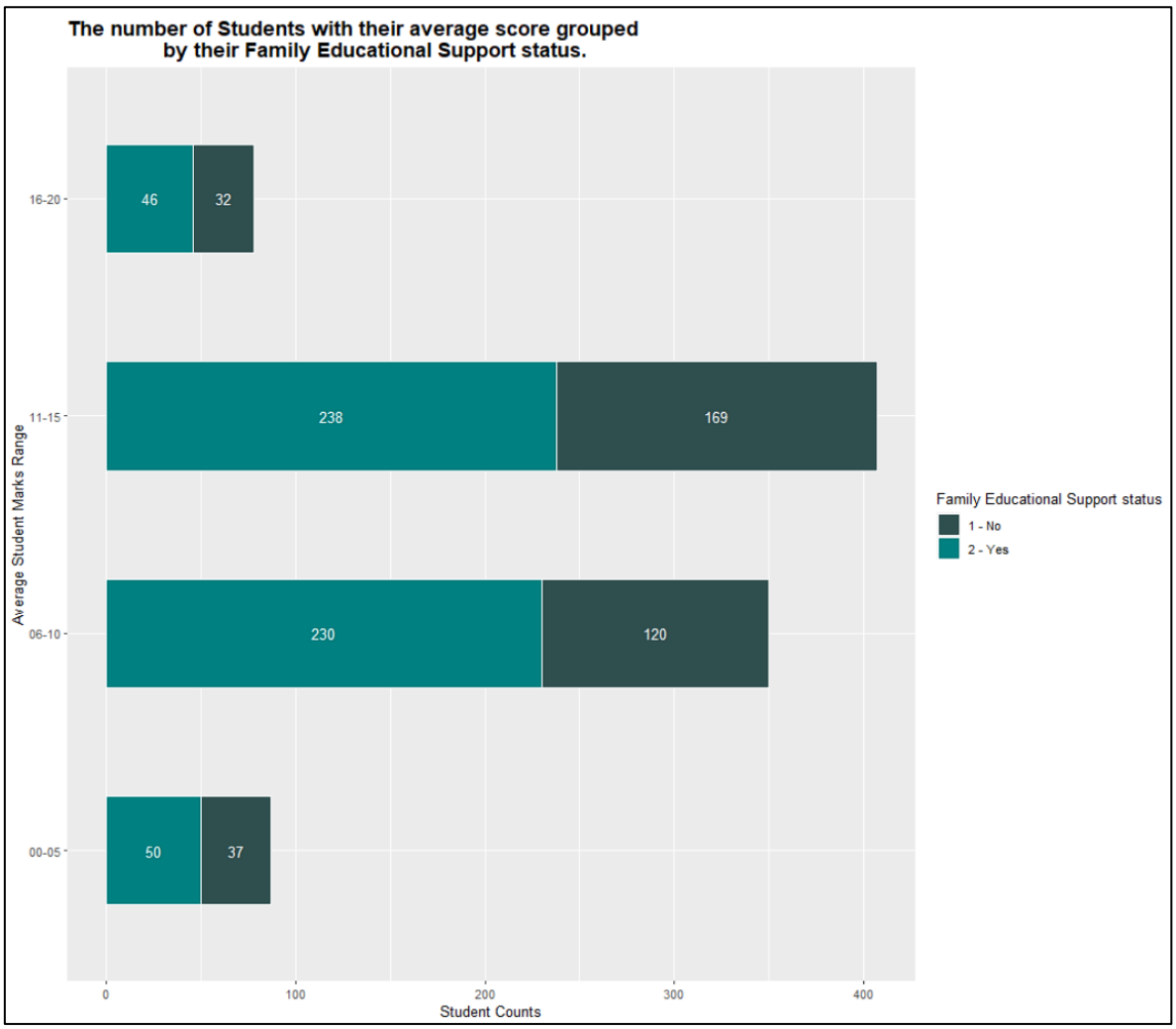


Figure 132 *shows the horizontal bar graph output of the* ***Q3A3V1****.*

The figure 131 above displays the output of the execution of the **Q3A3R1**, which shows the grouped counts of total students for each case of selected attributes that are the family educational support and their average grade range. The figure 132 above shows the outcome of the horizontal bar graph plotted after the execution of the **Q3A3V1** variable that displays the students' counts and the average grade range of students grouped based on their family educational support status.

```
> Q3A3R2
# A tibble: 4 x 3
# Groups:   famsup [2]
  famsup avgGradeRange counts
   <int> <chr>          <int>
1      1 11-15            135
2      1 16-20             32
3      2 11-15            179
4      2 16-20             46
>
```

Figure 133** shows the output of the **Q3A3R2.

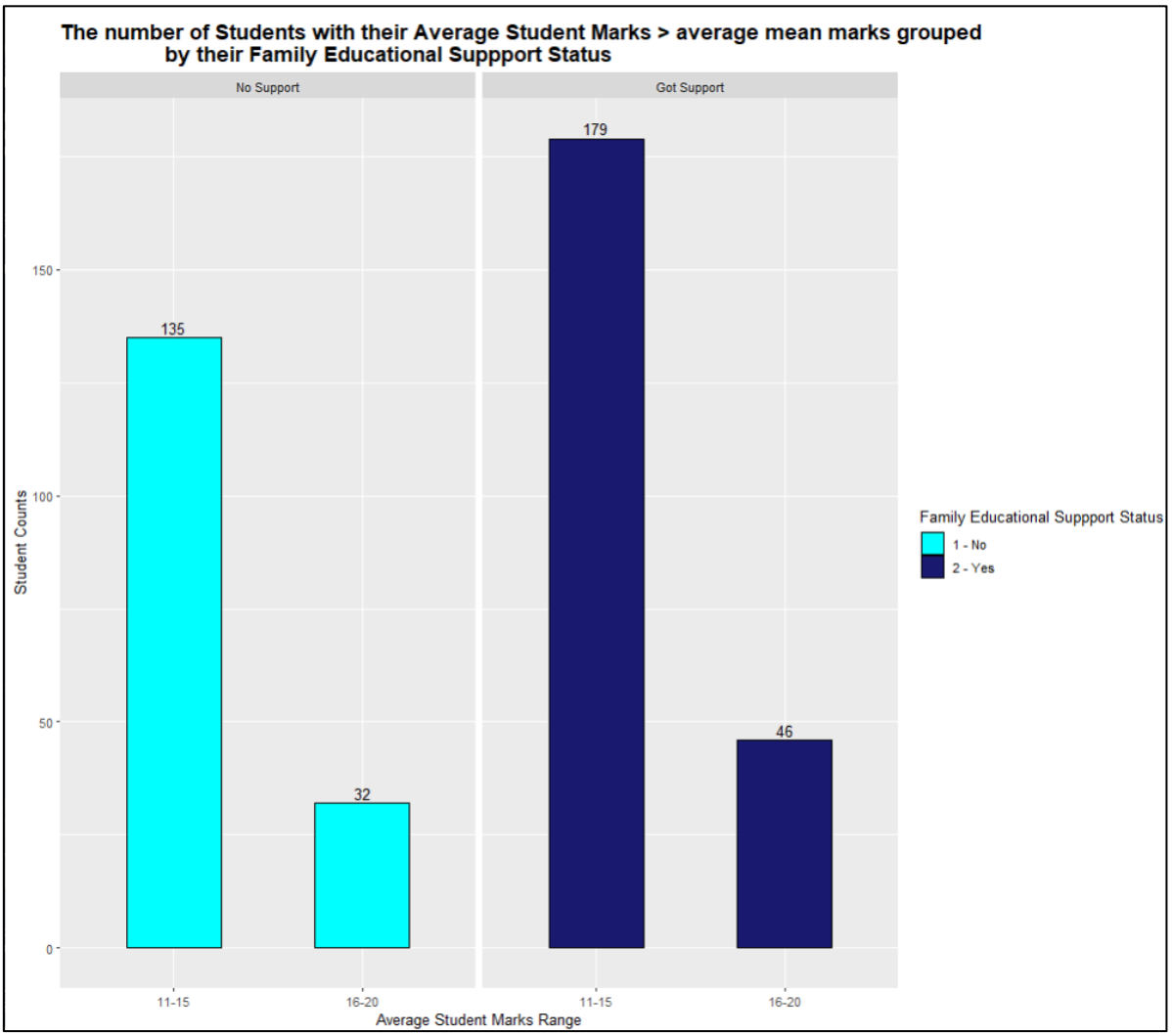


***Figure 134** shows the bar graph output of the **Q3A3V2**.*

The figure 133 above displays the output of the execution of the **Q3A3R2** which shows the grouped counts of total students for each case of selected attributes that are the family educational support and their average grade range. The figure 134 above shows the outcome of the bar graph plotted after the execution of the **Q3A3V2** variable that displays the students' counts and the average grade range of students, where it displays the counts of students that scored more than average mean marks grouped based on their family educational support status.

```
> Q3A3R3
# A tibble: 7 x 3
# Groups:   famsup [2]
  famsup avgGrade counts
   <int>    <dbl>  <int>
1      1       16     11
2      1       17      5
3      1       18      6
4      1       19     10
5      2       16     17
6      2       17     15
7      2       18     14
```

Figure 135** shows the output of the **Q3A3R3.

The figure 135 above displays the output of the execution of the **Q3A3R3** which shows the grouped counts of total students for each case of selected attributes that are the family educational support and their average grade. The figure 136 below shows the outcome of the bar graph plotted after the execution of the **Q3A3V3** variable that displays the students' counts and the average grade of students, where it displays the counts of students that got more than 15 average marks grouped based on their family educational support status.

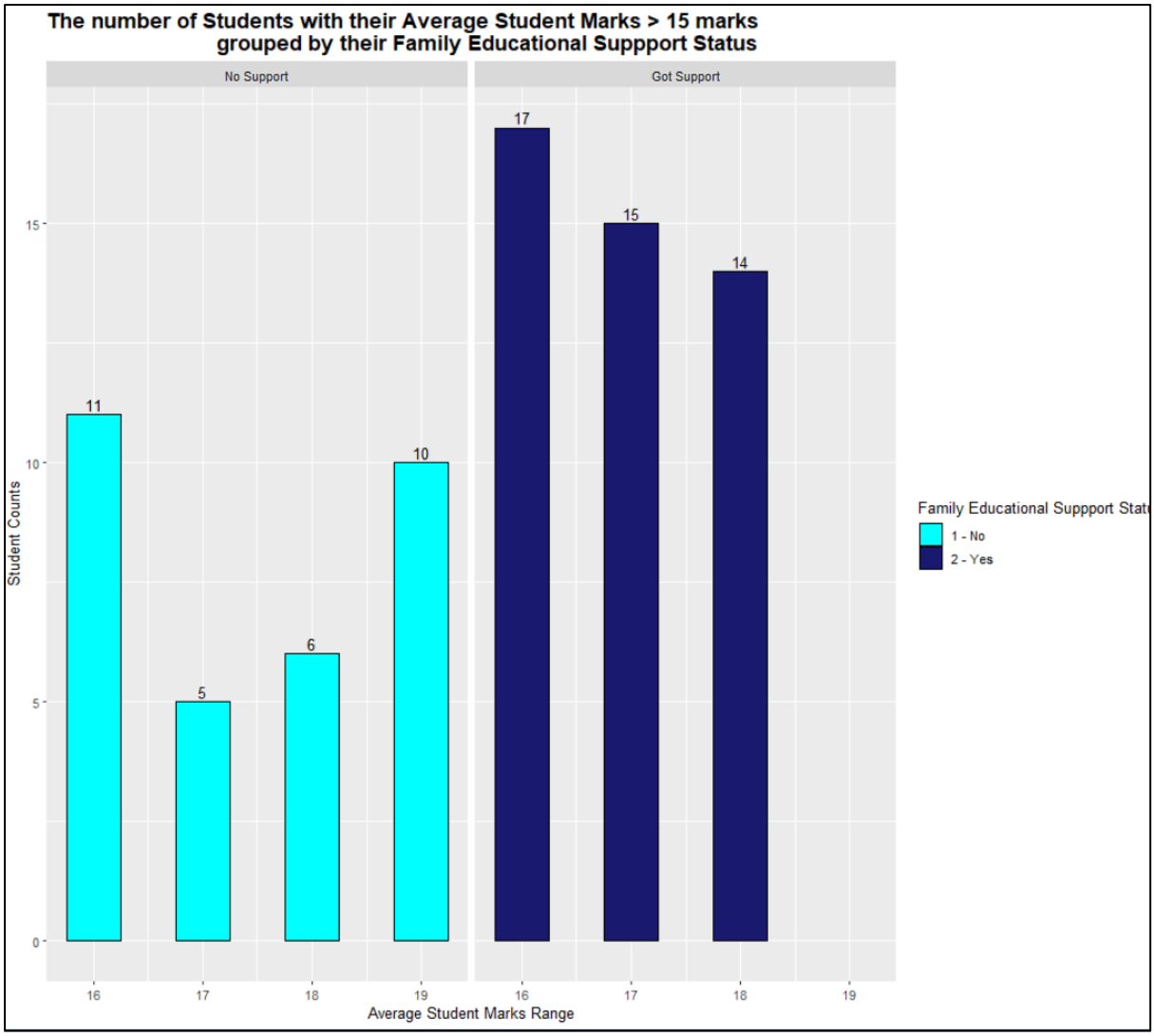


Figure 136 *shows the bar graph output of the* ***Q3A3V3.***

Summary for Data Findings

1) A large portion of students had got educational support from their families.
2) A total of 564 of the students had got educational support from their families.
3) A total of 358 of the students didn't get any education support from their families.
4) The total of students who got scored more than the average mean grade and had gotten educational support from family is 225 in total.
5) Only 167 of them scored more than the average mean grade and had not gotten any educational support from their family.
6) A total of 10 students had scored an average score of 19 who didn't get any educational support from their family

7) Not a single one of the students who scored an average mark of 19 from who had educational support from their family.

Explanation for the Data Findings.

Based on the data findings, it is confirmed, that the majority of the students got family support for their education. As can be seen the difference between those who had educational family support and those who didn't have any support who scored more than the average mean grade was quite similar to the previous statement as those who had educational support from the family is leading ahead with 58 of students. Furthermore, when analysing the students who got more than 15 average marks, the majority of them also earned educational support from their families. Despite this fact, when looking into students who scored an average mark of 19, there is none of them from the got support category, while there were a total 10 of students from the no support category. With this, it can be considered that the educational support from the family actually does impact and can boost minimally the academic performance of the students. This works maybe because when there is family support, these students would automatically attempt to put their effort to do better, which can cause them to get better results. According to Bogenschneider and Johnson (2004), they further supported in their research that, when the family members are supportive and positively involved in the students' studies, it really helps them perform well in their exams. But, still, it is also should be stated the fact that students who don't have any proper educational support also can do well in their academics if they put real effort and commitment into studying, which can relate to the students who got 19 average marks. It seems that they worked really hard to get that high average grade. Hence, it could be concluded that family educational support could not be the biggest attribute that affects the students' results based on the analysis done.

4.3.4 Analysis 3-4: Finding the correlation between the students' travel time to the class and their average grade.

The correlation between the students' travel time to school and their average marks will be analysed in this analysis. A stacked bar graph and two treemap graphs have been created for this analysis.

```
#Analysis 3-4
#Finding the relationship between the students' travel time to the class and their average grade.
Q3A4R1 <- dsap_data %>% group_by(traveltime, avgGradeRange) %>% summarise(counts = n())
Q3A4R1
Q3A4V1 <-  ggplot(Q3A4R1, aes(x=avgGradeRange,y=counts, fill=as.factor(traveltime))) +
           geom_bar(stat="identity",width = 0.5, color="black") +
           ggtitle("The number of Students with their average score grouped by their Travel Time.")+
           labs(x="Average Student Marks Range", y = "Student Counts", fill="Students Travel Time (Hours)")+
           theme(plot.title = element_text(size = 15, face = "bold")) +
           scale_fill_manual(values=c("#48D1CC", "#9370DB", "#000080","#800000"),
           labels = c("1", "2", "3", "4"))+
           geom_text(aes(label=counts), position = position_stack(vjust = 0.5), color="#F5F5F5")
Q3A4V1
```

Figure 137 *shows the R code used to create the data visualization figure of* ***Q3A4V1.***

```
Q3A4R2<- dsap_data %>% group_by(traveltime) %>% filter(avgGrade>avgMeanGrade) %>%
        summarise(counts = n(),  percentage = n()/length(which(dsap_data$avgGrade>avgMeanGrade))*100)
Q3A4R2
Q3A4V2 <-  ggplot(Q3A4R2, aes(x=percentage, y="", fill = as.factor(traveltime), area = percentage)) + geom_treemap()+
           theme(legend.justification="top",
                 panel.background = element_blank(),
                 axis.title = element_blank(),
                 axis.text = element_blank(),
                 axis.line = element_blank(),
                 axis.ticks= element_blank(),
           plot.title = element_text(size = 20, face = "bold")) +
           ggtitle("Shows the percentage of the Students that scored\n> average mark grouped by their travel time.") +
           labs(fill="Student Travel Time (Hours)")+ scale_fill_manual(values = c("#008B8B", "#191970","#4B0082","#DA70D6"),
                                                                      labels = c("1", "2", "3", "4")) +
           geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")), color = c("white"), place = "left")
Q3A4V2
```

Figure 138 *shows the R code used to create the data visualization figure of* ***Q3A4V2.***

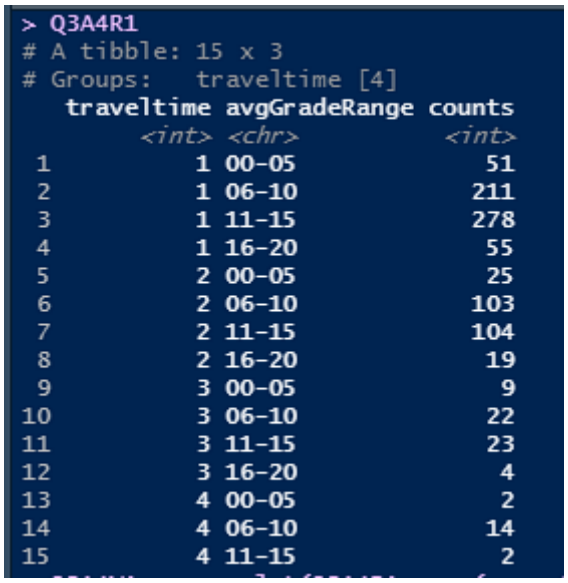
```
Q3A4R3<- dsap_data %>% group_by(studytime)%>% filter(avgGrade>avgMeanGrade & traveltime == 1) %>%
         summarise(counts = n(),  percentage = n()/length(which(dsap_data$avgGrade>avgMeanGrade & dsap_data$traveltime == 1))*100)
Q3A4R3
Q3A4V3 <- ggplot(Q3A4R3, aes(x=percentage, y="", fill = as.factor(studytime), area = percentage)) + geom_treemap()+
          theme(legend.justification="top",
                panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
          plot.title = element_text(size = 20, face = "bold")) +
          ggtitle("Shows the percentage of the Students that scored\n> average mark and their Travel Time to school were 1 hour,
                   grouped by their study time.") +
          labs(fill="Student Study Time (Hours)")+ scale_fill_manual(values = c("#008B8B", "#191970","#4B0082","#DA70D6"),
                                                                    labels = c("1", "2", "3", "4")) +
          geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")), color = c("white"), place = "left")
Q3A4V3
```

Figure 139 *shows the R code used to create the data visualization figure of* ***Q3A4V3.***

```
ggarrange(Q3A4V2, Q3A4V3, nrow=2, ncol=1)
```

Figure 140 *shows the R code used to arrange the data visualization figures of* ***Q3A4V2*** *and* ***Q3A4V3*** *in one view.*

As shown in the code figures above, for the stacked bar graph, the **traveltime** and **avgGradeRange** are grouped and counted for this analysis. For the first treemap graph, only the **traveltime** was grouped and calculated the percentage by filtering the **avgGrade.** The second treemap graph was grouped by the **studytime** and calculated the percentage by filtering the **avgGrade** and **traveltime**.



```
> Q3A4R1
# A tibble: 15 x 3
# Groups:   traveltime [4]
   traveltime avgGradeRange counts
        <int> <chr>          <int>
 1          1 00-05             51
 2          1 06-10            211
 3          1 11-15            278
 4          1 16-20             55
 5          2 00-05             25
 6          2 06-10            103
 7          2 11-15            104
 8          2 16-20             19
 9          3 00-05              9
10          3 06-10             22
11          3 11-15             23
12          3 16-20              4
13          4 00-05              2
14          4 06-10             14
15          4 11-15              2
```

Figure 141** shows the output of the **Q3A4R1.

The figure 141 above displays the result of the execution of the **Q3A4R1**, which shows the grouped counts of total students for each case of selected attributes that are the travel time and their average grade range. The figure 142 below shows the outcome of the stacked bar graph plotted after the execution of the **Q3A4V1** variable that displays the students' counts and the average grade range of students grouped based on their travel time to school.

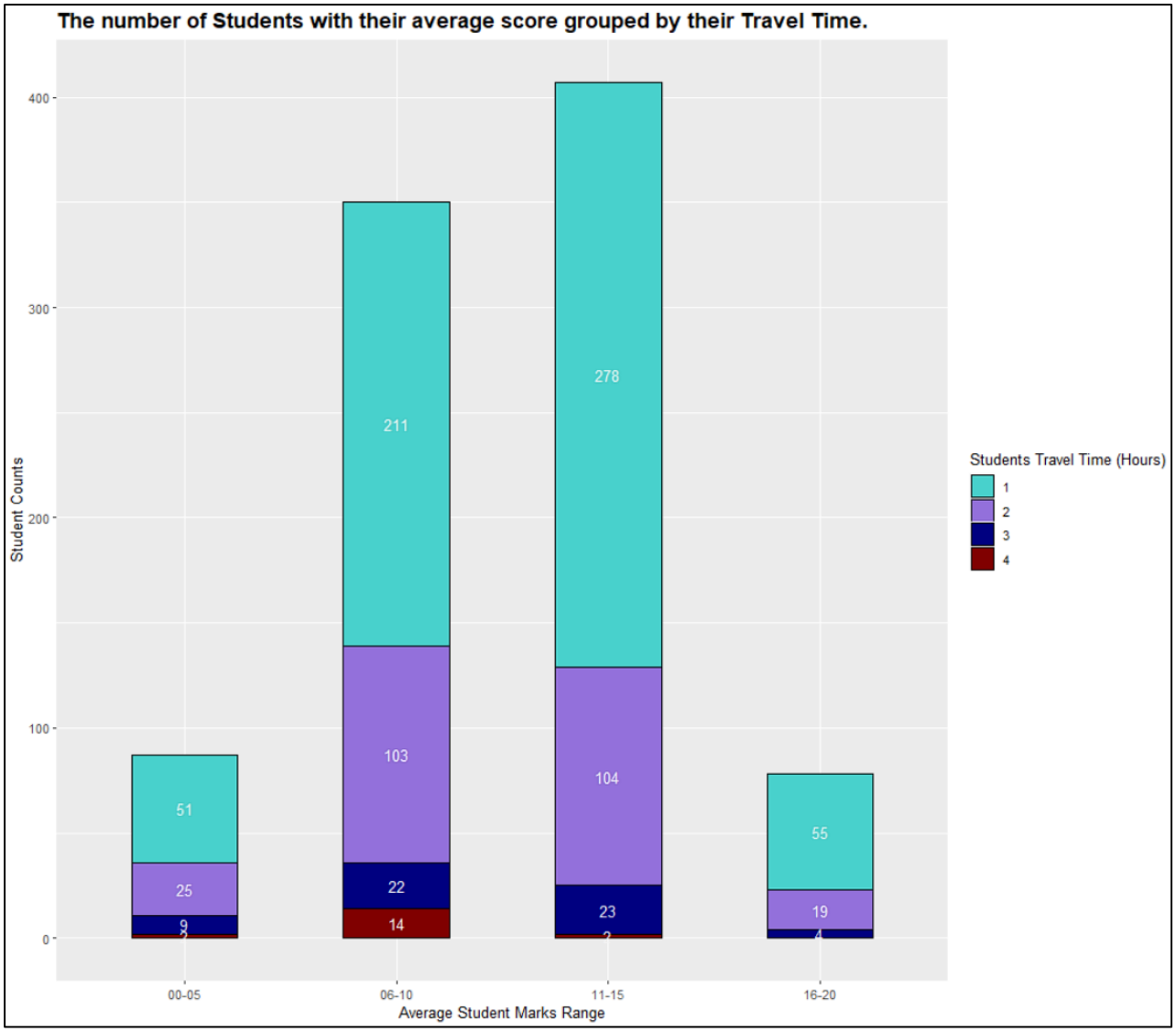


Figure 142 *shows the stacked bar graph output of the* ***Q3A4V1****.*

```
Q3A4R2
A tibble: 4 x 3
traveltime counts percentage
     <int>  <int>      <dbl>
         1    270       68.9
         2     99       25.3
         3     21       5.36
         4      2      0.510
```

Figure 143 *shows the output of the* ***Q3A4R2.***

```
> Q3A4R3
# A tibble: 4 x 3
  studytime counts percentage
      <int>  <int>      <dbl>
1         1     71       26.3
2         2    116       43.0
3         3     63       23.3
4         4     20       7.41
>
```

Figure 144 *shows the output of the* ***Q3A4R3.***

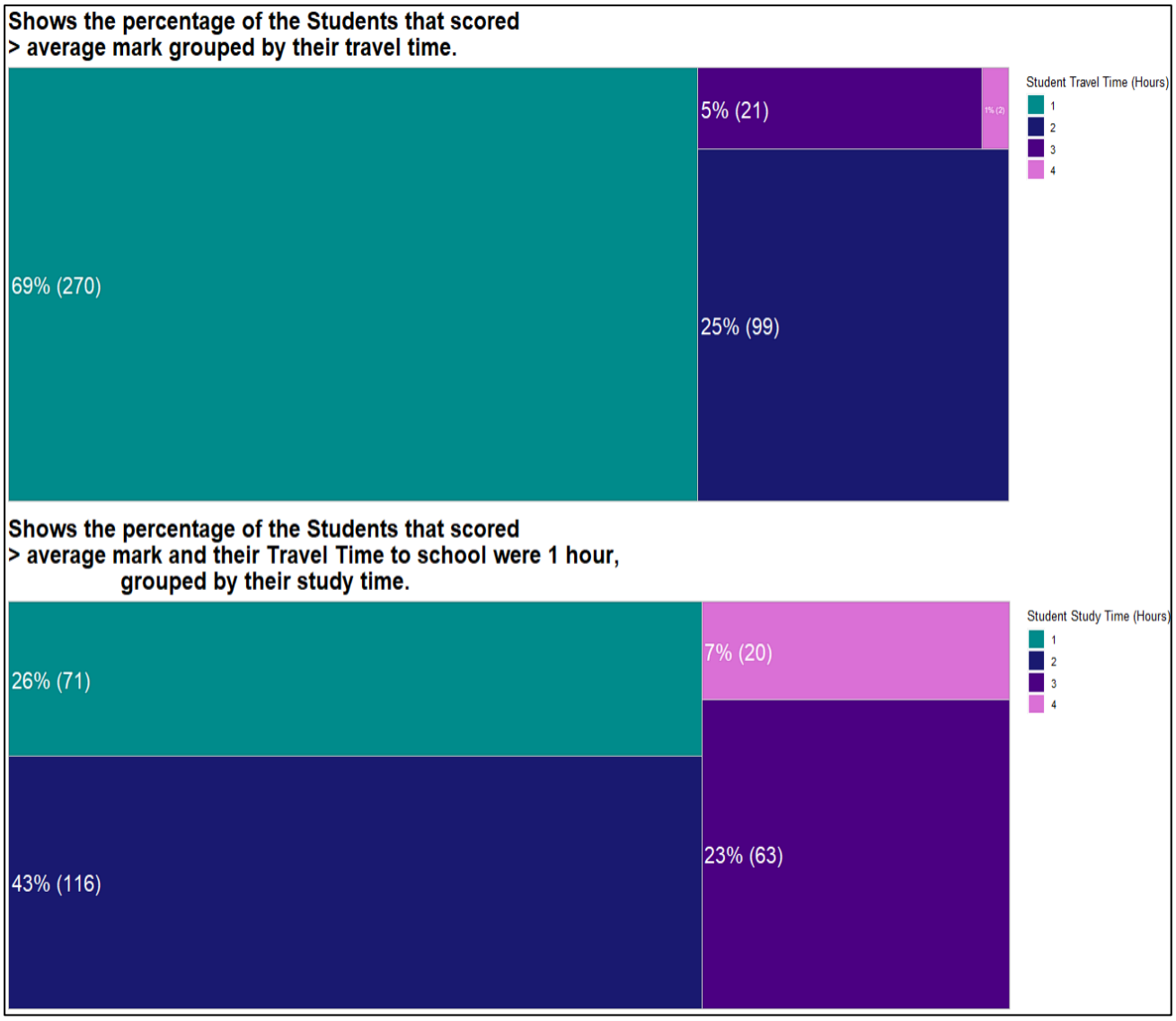


Figure 145 *shows the treemap graphs of the output of* ***Q3A4V2*** *and* ***Q3A4V3****.*

The figures 143 and 144 above shows the execution output of the **Q3A4R2** and **Q3A4R3**. In figure 145, the first treemap graph that has been drawn displays the calculated total student counts and the percentage of them who scored more than the average mean marks grouped with

their travel time. The second treemap graph picturing the calculated total student counts and percentage of them who scored more than average mean grade and their travel time was 1 hour grouped with their study time.

Summary for Data Findings

1) The large portion of the students' travel time to school was 1 hour.
2) There is not a single student who had 4 hours travel time in the 16-20 average mark range.
3) In all average mark ranges, most of the students had 1-hour travel time to their school.
4) 69% of students who got more than the average mean grade had 1-hour travel time to school.
5) A large percentage of students that scored more than the average mean grade who had 1-hour travel time to had more than

Explanation for the Data Findings.

Based on the data findings, it clearly can be viewed that a large number of the students had to travel for 1 hour to reach their school. There were none of those students who scored more than 15 average marks had a travel time of 4 hours to school. Moreover, 69% that is a total of 270 students who scored more than the average mean score had only 1 hour of travel time to reach their school while the others had more than 1 hour travel time. When further analysed, it can be noticed that a big number of the students who got more than the average mean grade and their travel time was only 1 hour had studied for more than 1 hour. From this, it can be claimed that there is a high chance that travel time to school is influencing the students' academic performance. It can because when the travelling time to the school is quite long, it is possible that it can consume out the students' daily activities that include their time for studying. Normally, without a fair amount of study time, the students won't be able to excel on their exams because they didn't get enough study time to cover all the important topics that would be asked. According to Wu (2014), in her thesis, it was further supported that long travelling hours can shorten the students' study time which can end up them getting poor marks in their examinations. Hence, it can be guaranteed that the travel time has a big influence on the students' academic performance where the students with short travel time will manage to do better.

4.3.5 Analysis 3-5: Finding the correlation between students' Internet access and their average marks.

The correlation between the students' Internet access status and their average marks will be analysed in this analysis. A bar graph and two treemap graphs have been created for this analysis.

```
#Analysis 3-5
#Finding the correlation between students' internet access and their average marks.
Q3A5R1 <- dsap_data %>% group_by(internet, avgGradeRange) %>% summarise(counts = n())
Q3A5R1

Q3A5V1 <-  ggplot(Q3A5R1, aes(x=avgGradeRange, y = counts, fill = as.factor(internet))) +
           geom_bar(stat = "identity", width = 1, color="black", position = position_dodge2()) +
           labs(fill="Student Internet access", x = "Average Student Marks Range", y="Student Counts") +
           ggtitle("The number of Students with their average score grouped by Internet Access status.") +
           theme(plot.title = element_text(size = 15, face = "bold")) +
           geom_text(aes(label=counts), position = position_dodge2(1), vjust=-0.5) +
           scale_fill_manual(values=c("#48D1CC", "#9370DB"),labels = c("1 - No", "2 - Yes"))
Q3A5V1
```

Figure 146 *shows the R code used to create the data visualization figure of* ***Q3A5V1.***

```
Q3A5R2<- dsap_data %>% group_by(internet) %>% filter(avgGrade>15) %>%
        summarise(counts = n(),  percentage = n()/length(which(dsap_data$avgGrade>15))*100)
Q3A5R2
Q3A5V2 <- ggplot(Q3A5R2, aes(x=percentage, y="", fill = as.factor(internet), area = percentage)) +
          geom_treemap()+
          theme(legend.justification="top",
                panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
          plot.title = element_text(size = 20, face = "bold")) +
          ggtitle("Shows the percentage of the Students that scored\n> 15 average mark grouped by the Internet Access status.") +
          labs(fill="Internet Access")+ scale_fill_manual(values = c("#008B8B", "#191970","#4B0082","#DA70D6"),
                                                         labels = c("1 - No", "2 - Yes")) +
          geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                            color = c("white"), place = "left")
Q3A5V2
```

Figure 147 *shows the R code used to create the data visualization figure of* ***Q3A5V2.***

```
Q3A5R3<- dsap_data %>% group_by(internet) %>% filter(avgGrade<=5) %>%
        summarise(counts = n(),  percentage = n()/length(which(dsap_data$avgGrade<=5))*100)
Q3A5R3
Q3A5V3 <- ggplot(Q3A5R3, aes(x=percentage, y="", fill = as.factor(internet), area = percentage)) + geom_treemap()+
          theme(legend.justification="top",
                panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
          plot.title = element_text(size = 20, face = "bold")) +
          ggtitle("Shows the percentage of the Students that scored\n<= 5 average mark grouped by the Internet Access status.") +
          labs(fill="Internet Access")+ scale_fill_manual(values = c("#008B8B", "#191970","#4B0082","#DA70D6"),
                                                         labels = c("1 - No", "2 - Yes")) +
          geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                            color = c("white"), place = "left")
Q3A5V3
```

Figure 148 *shows the R code used to create the data visualization figure of* ***Q3A5V3.***

```
ggarrange(Q3A5V2, Q3A5V3, nrow=2, ncol=1)
```

Figure 149 *shows the R code used to arrange the data visualization figures of* ***Q3A5V2*** *and* ***Q3A5V3*** *in one view.*

As shown in the code figures above, for the bar graph, the **internet** and **avgGradeRange** are grouped and counted for this analysis. For both treemap graphs, only the **internet** was grouped and calculated the students' percentage by filtering the **avgGrade.**

```
> Q3A5R1
# A tibble: 8 x 3
# Groups:   internet [2]
  internet avgGradeRange counts
     <int> <chr>           <int>
1        1 00-05              17
2        1 06-10              73
3        1 11-15              58
4        1 16-20               8
5        2 00-05              70
6        2 06-10             277
7        2 11-15             349
8        2 16-20              70
```

Figure 150 *shows the output of the* ***Q3A5R1.***

The figure 150 above displays the result of the execution of the **Q3A5R1**, which shows the grouped counts of total students for each case of selected attributes that are the Internet access and their average grade range. The figure 151 below shows the outcome of the bar graph plotted after the execution of the **Q3A5V1** variable that displays the students' counts and the average grade range of students grouped based on their Internet access status.

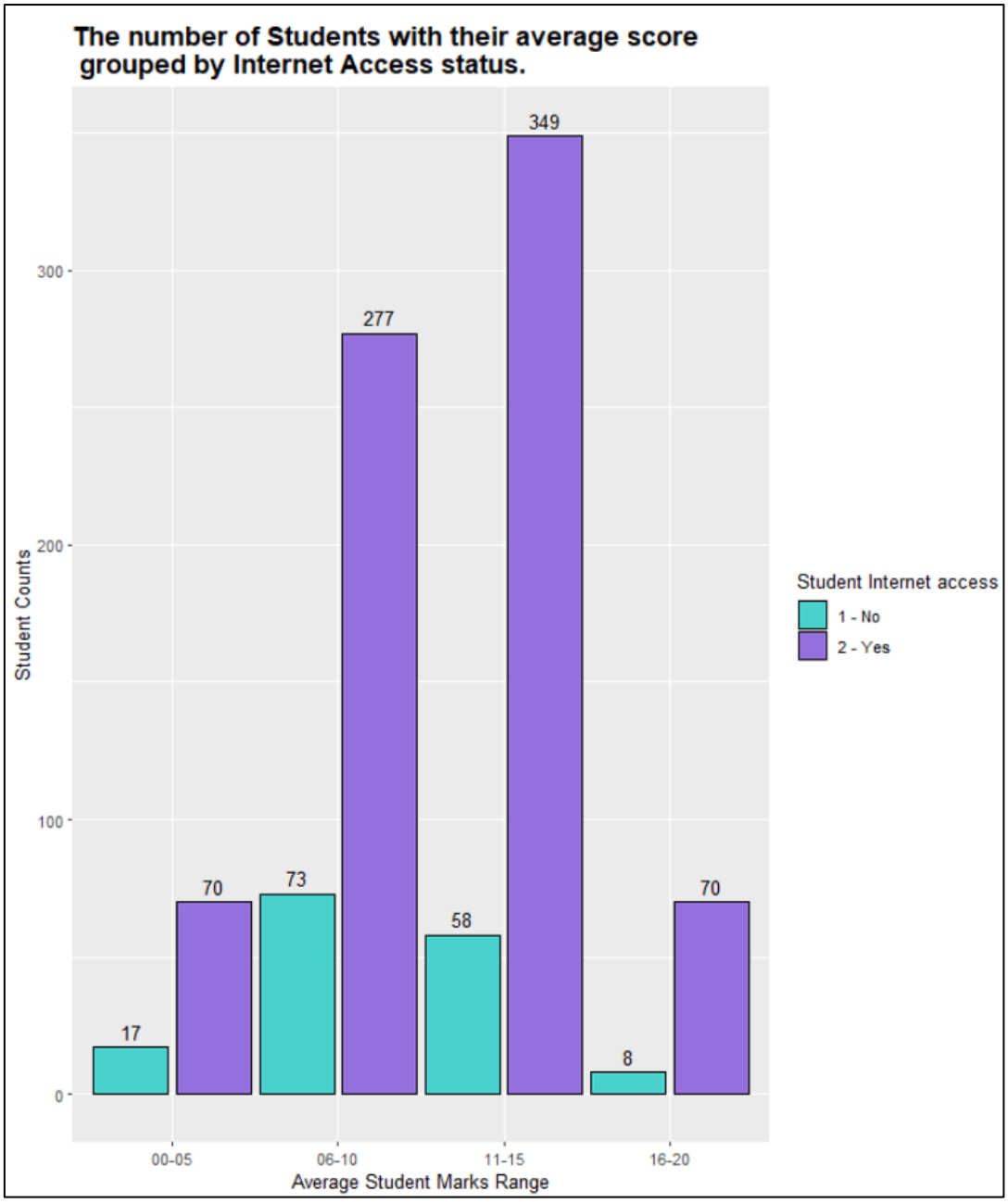


Figure 151 *shows the bar graph output of the* ***Q3A5V1****.*

The figures 152 and 153 below show the execution output of the **Q3A5R2** and **Q3A5R3**. In figure 154, the treemap graphs that has been drawn displays the calculated total student counts and the percentage of them who scored more than 15 average marks and less than equal to 5 average mark grouped with their Internet access status.

```
> Q3A5R2
# A tibble: 2 x 3
  internet counts percentage
     <int>  <int>      <dbl>
1        1      8       10.3
2        2     70       89.7
```

Figure 152 *shows the output of the* ***Q3A5R2.***

```
> Q3A5R3
# A tibble: 2 x 3
  internet counts percentage
     <int>  <int>      <dbl>
1        1     17       19.5
2        2     70       80.5
```

Figure 153 *shows the output of the* ***Q3A5R3.***

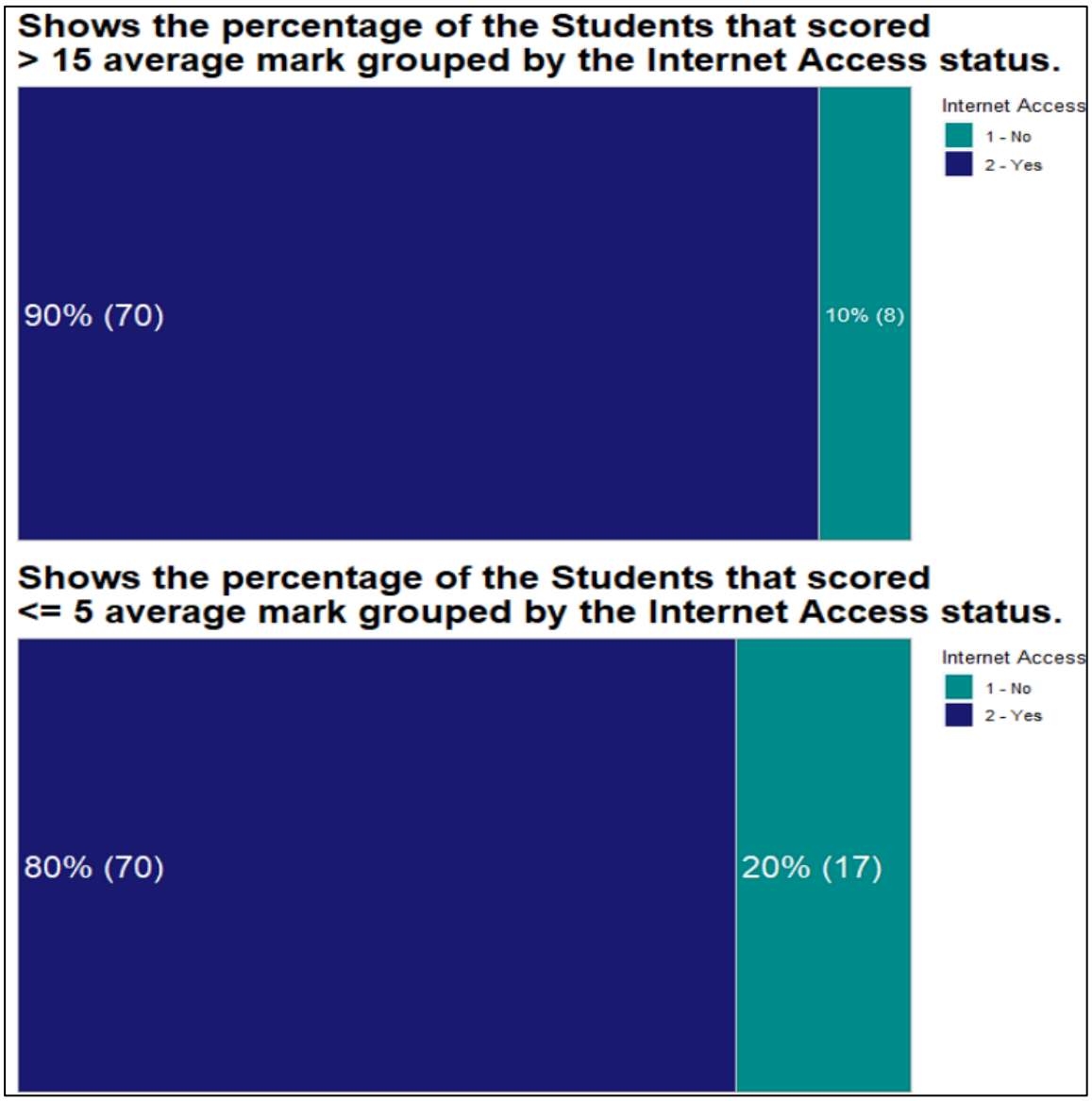


Figure 154 *shows the treemap graphs of the output of* ***Q3A5V2*** *and* ***Q3A5V3****.*

Summary for Data Findings

1) The largest number of students that is total of 766 of them had internet access for each average mark range.
2) Only a total of 156 of students didn't have internet access.
3) The 11-15 average mark range had the greatest number of students who had internet access.
4) The 06-10 average mark range had the greatest number of students who don't have internet access.
5) 90% of students who scored more than 15 average mark had internet access, while the other 10% of them didn't have one.
6) 80% of students who scored lesser than equal to 5 average mark had internet access, while the other 20% didn't have one.

Explanation for the Data Findings.

Based on the data findings, it can be noticed that the majority of these students had internet access while a little number of them doesn't have one. Even though, the most percentage of students who scored more than 15 average mark and less than equal to 5 average mark is the one with internet access. This tells us quite a lot that internet access has produced a quite influence on the student's grades. Only when Internet access has been put in use in a good manner, it certainly can help students to get better academic grades. It is because Internet access has tremendous benefits that can be done especially since it allows students the capability to find relevant study materials and any information related to studies in just seconds. However, on the flip side, it can be also one of the factors that affect badly the students' academic performance if it wasn't utilised in a good way. This is because excessive usage of the Internet can cause Internet addiction where they tend to waste a lot of time and do nothing, which eventually they will not study for anything ended up getting bad grades. According to Empton et. al.(2021), they further supported in their research that quality access to the Internet definitely shapes the students' academic performance. Hence, it can be assured that Internet usage has quite an essential influence on students' academic achievements.

Conclusion

As a conclusion for this third question, based on the observed analyses, it can be reasoned that a student's resource availability and comfort also play an essential role in influencing academic performance despite the that the results could vary based on the attributes. Especially, it was found that students' travel time is one of the most influential attributes that impacted their academic grades. Overall, the availability of resources and comfort should be accessible to all the students as it has to be fair for everyone which can assist them to utilize it as a beneficial way to do well in their academics.

4.4 Question 4: How does students' family influence impact their marks?

This fourth question will be analysing whether the family influence of a student would affect their overall academic grades. The student's attributes that will be covered in this question are family size, parents' cohabitation status, mother's education, father's education, mother's occupation and father's occupation.

4.4.1 Analysis 4-1: Finding the relationship between family size and students' average grades.

The correlation between the students' family size and their average marks will be analysed in this analysis. A horizontal bar graph and two stacked bar graphs have been created for this analysis.

```
#=======Question 4========#
#Question 4: How does students' family influence impact their marks?
#Analysis 4 - 1
#Finding the relationship between family size and students' average grades.
Q4A1R1 <- dsap_data %>% group_by(famsize, avgGradeRange) %>% summarise(counts = n())
Q4A1R1

Q4A1V1 <-  ggplot(Q4A1R1, aes(x=avgGradeRange, y = counts, fill=as.factor(famsize))) +
           geom_bar(stat="identity",width = 0.5, color="white") +
           ggtitle("The number of Students with their average score grouped by their Family Size.")+
           labs(x="Average Student Marks Range", y = "Student Counts", fill="Student Family Size")+
           theme(plot.title = element_text(size = 15, face = "bold")) +
           scale_fill_manual(values=c("#BC8F8F", "#483D8B"))+ coord_flip() +
           facet_wrap(~famsize, labeller = as_labeller(c(`GT3` = "GT3 - Greater Than 3", `LE3`= "LE3 - Less Than 3"))) +
           geom_text(aes(label=counts),  position = position_stack(vjust = 0.5), color = "white")
Q4A1V1
```

Figure 155 *shows the R code used to create the data visualization figure of* ***Q4A1V1.***

```
Q4A1R2<- dsap_data %>% group_by(famsize, avgGrade) %>%
         filter(avgGrade>15) %>% summarise(counts = n())
Q4A1R2
Q4A1V2 <- ggplot(Q4A1R2, aes(x=avgGrade, y= counts, fill=as.factor(famsize))) +
          geom_bar(stat="identity",width = 0.5, color="black") +
          ggtitle("The number of Students with their average score > 15 grouped by their Family Size.")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="Students Family Size")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          scale_fill_manual(values=c("#FF6347", "#FFA500"))+
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q4A1V2
```

Figure 156 *shows the R code used to create the data visualization figure of* ***Q4A1V2.***

```
Q4A1R3<- dsap_data %>% group_by(famsize, avgGrade) %>%
        filter(avgGrade<=5) %>% summarise(counts = n())
Q4A1R3
Q4A1V3 <- ggplot(Q4A1R3, aes(x=avgGrade, y= counts, fill=as.factor(famsize))) +
          geom_bar(stat="identity",width = 0.5, color="black") +
          ggtitle("The number of Students with their average score <= 5 grouped by their Family Size.")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="Students Family Size")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          scale_fill_manual(values=c("#FF6347", "#FFA500"))+
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q4A1V3
```

Figure 157 *shows the R code used to create the data visualization figure of* ***Q4A1V3.***

```
ggarrange(Q4A1V2, Q4A1V3, nrow=2, ncol=1)
```

Figure 158 *shows the R code used to arrange the data visualization figures of* ***Q4A1V2*** *and* ***Q4A1V3*** *in one view.*

As shown in the code figures above, for all the graphs, the **famsize** and **avgGradeRange** are grouped and counted for this analysis. For both stacked bar graphs, it was filtered with the **avgGrade.**

```
> Q4A1R1
# A tibble: 8 x 3
# Groups:   famsize [2]
  famsize avgGradeRange counts
  <chr>   <chr>          <int>
1 GT3     00-05             74
2 GT3     06-10            251
3 GT3     11-15            278
4 GT3     16-20             51
5 LE3     00-05             13
6 LE3     06-10             99
7 LE3     11-15            129
8 LE3     16-20             27
```

Figure 159 *shows the output of the* ***Q4A1R1.***

The figure 159 above displays the result of the execution of the **Q4A1R1**, which shows the grouped counts of total students for each case of selected attributes that are the family size and their average grade range. The figure 160 below shows the outcome of the horizontal bar graph plotted after the execution of the **Q4A1V1** variable that displays the students' counts and the average grade range of students grouped based on their family size.

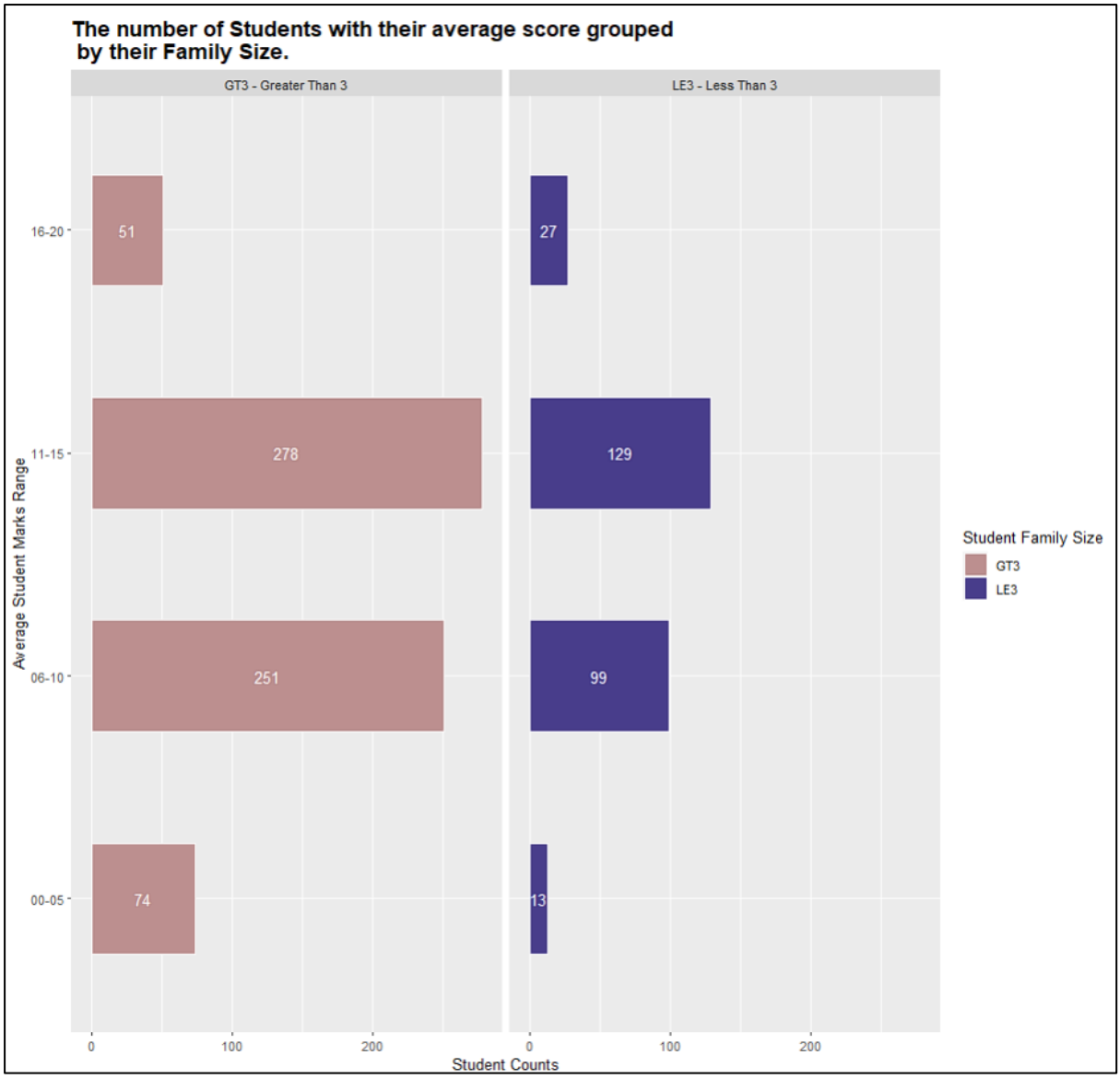


Figure 160 *shows the horizontal bar graph output of the* ***Q4A1V1****.*

```
> Q4A1R2
# A tibble: 8 x 3
# Groups:   famsize [2]
  famsize avgGrade counts
  <chr>      <dbl>  <int>
1 GT3           16     18
2 GT3           17     15
3 GT3           18     15
4 GT3           19      3
5 LE3           16     10
6 LE3           17      5
7 LE3           18      5
8 LE3           19      7
```

Figure 161 *shows the output of the* ***Q4A1R2.***

```
> Q4A1R3
# A tibble: 8 x 3
# Groups:   famsize [2]
  famsize avgGrade counts
  <chr>      <dbl>  <int>
1 GT3            1      2
2 GT3            2     11
3 GT3            3      6
4 GT3            4     28
5 GT3            5     27
6 LE3            2      4
7 LE3            4      6
8 LE3            5      3
```

Figure 162 *shows the output of the* ***Q4A1R3.***

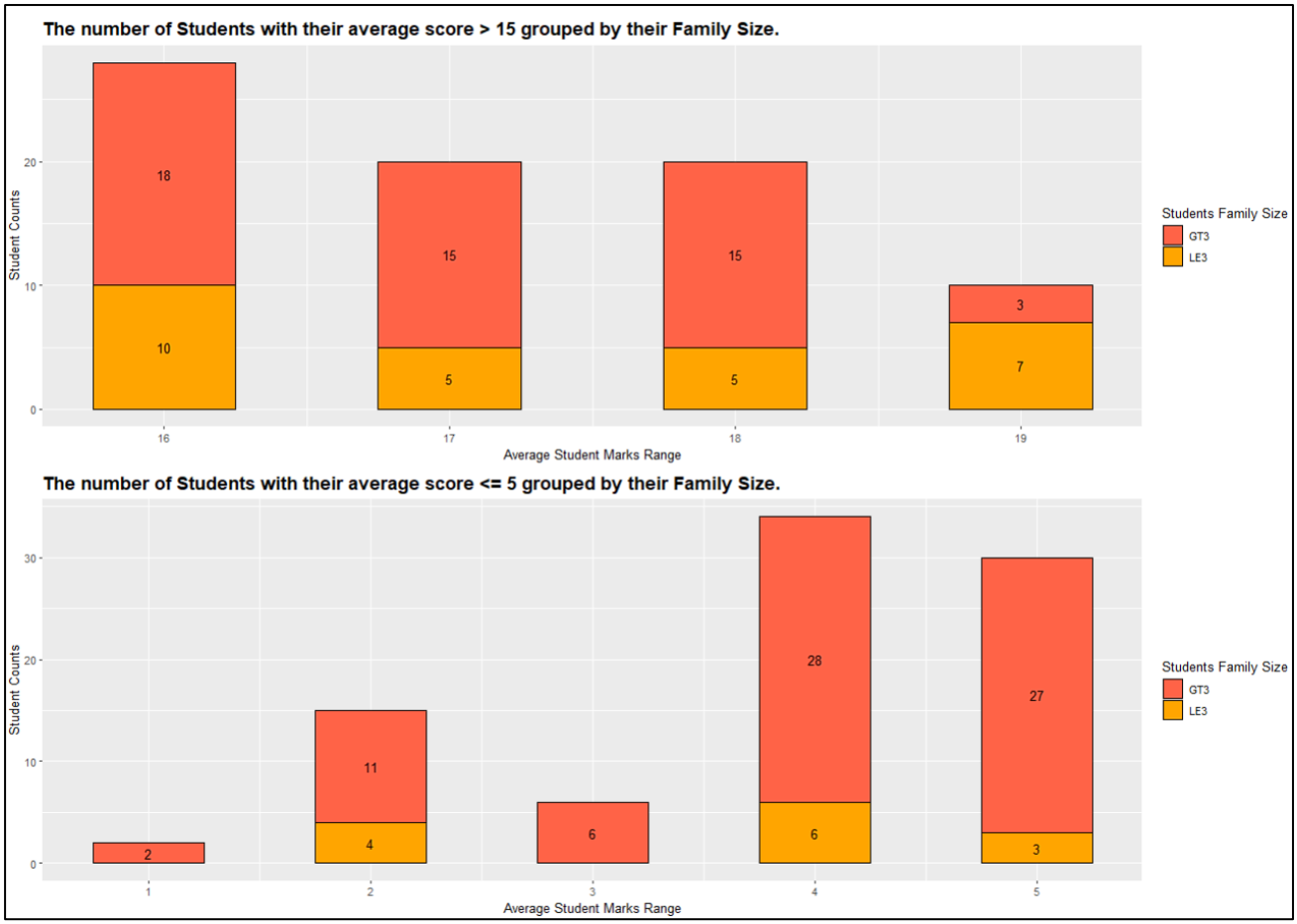


Figure 163 *shows the stacked bar graphs of the output of* ***Q4A1V2*** *and* ***Q4A1V3****.*

The figures 161 and 162 displays the execution output of the **Q4A1R2** and **Q4A1R3**, which displays the calculated total student counts grouped by the family size and average grade. In figure 163, the two stacked bar graph illustrates the number of students that scored more than 15 average marks and less than equal to 5 average marks with their family size.

Summary for Data Findings

1) A majority number of students, that is a total of 654, have a family size greater than 3.
2) The other 268 total of students has a family size less than equal to 3.
3) The very least students who had size of the family less than equal to 3 are in the average mark range of 00-05.
4) The majority of students who scored 19 average marks had a size of the family of less than equal to 3.

Explanation for the Data Findings.

Based on the data findings, it was noticed that a big numeral of students has families with a family size greater than 3 when compared to those who had a size of the family less than equal to 3. The majority number of students who scored an excellent average grade of 19 had a family size that is less than equal to 3. With this, it can state that there is a slight chance of family size influencing students' grades. When the family size is small, these students' parents can effortlessly monitor their child's academic progress better because they don't have to split much of their attention among their other children. Moreover, with smaller family sizes, it will be easy for these parents' to give educational support to their children where they can assist them with full involvement. It was further supported that small family size will most often accelerate positive impact on the students' academic performance (Ella et al., 2015). It was also mentioned that a large family makes the parents unable to show proper attention to their children, which results in the students' getting bad results in their exams. Hence, it is true that having a small family size can positively impact the students' marks.

4.4.2 Analysis 4-2: Finding the correlation between parents' cohabitation status and students' average grades.

The correlation between the parents' cohabitation status and their average marks will be analysed in this analysis. A bar graph has been created for this analysis.

```
#Analysis 4 - 2
#Finding the correlation between parents' cohabitation status and students' average grades.
Q4A2R1 <- dsap_data %>% group_by(Pstatus, avgGradeRange) %>% summarise(counts = n())
Q4A2R1
Q4A2R1 <- ggplot(Q4A2R1, aes(x=avgGradeRange, y = counts, fill=as.factor(Pstatus))) +
          geom_bar(stat="identity",width = 0.5, color="white") +
          ggtitle("The number of Students with their average score grouped by their Parents' Cohabitation Status")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="Student Parents' Cohabitation Status")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          scale_fill_manual(values=c("#4169E1", "#F4A460"))+
          facet_wrap(~Pstatus, labeller = as_labeller(c(`A`= "A - Living Apart",
                                                        `T`= "T - Living Together"))) +
          geom_text(aes(label=counts),  position = position_stack(vjust = 0.5), color = "black")
Q4A2R1
```

Figure 164 *shows the R code used to create the data visualization figure of* ***Q4A2V1.***

As shown in the code figure above, for the bar graph, the **Pstatus** and **avgGradeRange** are grouped and counted for this analysis.

```
> Q4A2R1
# A tibble: 8 x 3
# Groups:   Pstatus [2]
  Pstatus avgGradeRange counts
  <chr>   <chr>          <int>
1 A       00-05              2
2 A       06-10             36
3 A       11-15             48
4 A       16-20             11
5 T       00-05             85
6 T       06-10            314
7 T       11-15            359
8 T       16-20             67
>
```

Figure 165 *shows the output of the* ***Q4A2R1.***

The figure 161 above displays the result of the execution of the **Q4A2R1**, which shows the grouped counts of total students for each case of selected attributes that are the parents' cohabitation status and their average grade range. The figure 162 below shows the outcome of the bar graph plotted after the execution of the **Q4A2V1** variable that displays the students' counts and the average grade range of students grouped based on their parents' cohabitation status.

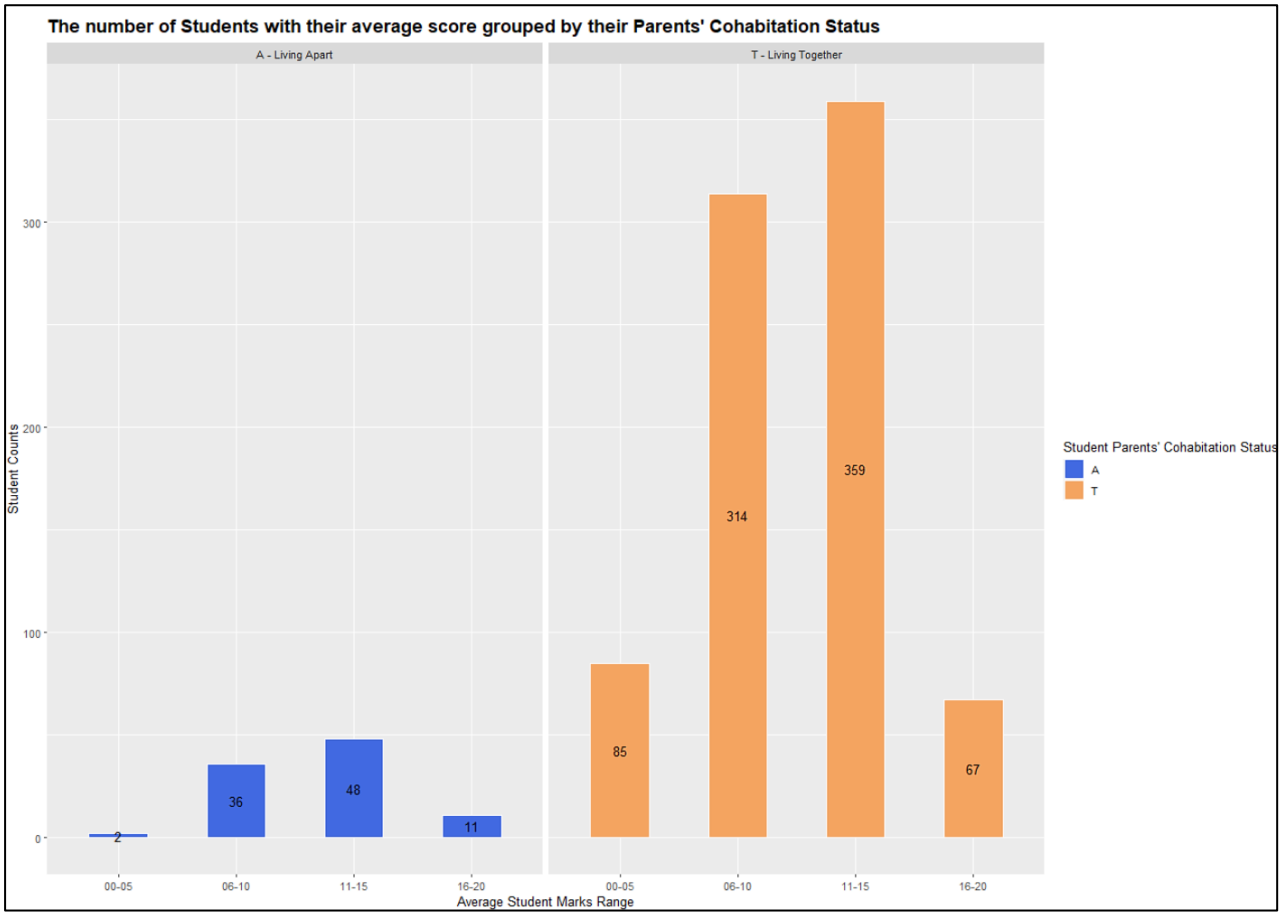


Figure 166 *shows the horizontal bar graph output of the* ***Q4A2V1****.*

Summary for Data Findings

1) The majority of the parent's cohabitation status of these students, that is 825 total of them, were living together.
2) Only a total of 97 students' parent's cohabitation status were living apart.
3) The 00-05 average mark range has the least amount of students that is only 2 of them whose parents' cohabitation status was living apart.

Explanation for the Data Findings.

Based on the data findings, it can be noticed that most of these students' parents were living together with them. Only very few parents were living apart from the students. As there weren't enough numbers of students to compare with, it can be assumed that there is not much influence of the cohabitation status on the academic performance of the students. Moreover, students with parents living apart were noticed in all average mark ranges. Hence, it is not going to be a false analysis that the cohabitation status of parents does not really impact the students' academic performance much as compared to other attributes based on this dataset.

4.4.3 Analysis 4-3: Finding the relationship between a mother's education and a student's average mark.

The correlation between the mother's education level and their average marks will be analysed in this analysis. A bar graph and a pie chart have been drawn for this analysis.

```
#Analysis 4 - 3
#Finding the relationship between a mother's education and a student's average mark.
Q4A3R1 <- dsap_data %>% group_by(Medu, avgGradeRange) %>% summarise(counts = n())
Q4A3R1
Q4A3V1 <-ggplot(Q4A3R1, aes(x=avgGradeRange, y = counts, fill=as.factor(Medu))) +
          geom_bar(stat="identity",width = 0.5, color="black") +
         ggtitle("The number of Students with their average score grouped by their Mother's Education Level")+
         labs(x="Average Student Marks Range", y = "Student Counts", fill="Mother's Education Level")+
         theme(plot.title = element_text(size = 15, face = "bold")) +
         scale_fill_manual(values=c("#483D8B", "#1E90FF","#2F4F4F", "#4B0082","#DA70D6"),
                           labels=c(c("0 - None",
                                      "1 - Primary Education (4th grade)",
                                      "2 - 5th to 9th grade", "3 - secondary education",
                                      "4 - higher education")))+
         facet_wrap(~Medu, labeller = as_labeller(c(`0`= "0 - None",
                                                    `1`= "1 - Primary Education (4th grade)",
                                                    `2`= "2 - 5th to 9th grade",
                                                    `3`="3 - secondary education",
                                                    `4`=" 4 - higher education"))) +
         geom_text(aes(label=counts),  position = position_stack(vjust = 0.5), color = "white")

Q4A3V1
```

Figure 167 *shows the R code used to create the data visualization figure of* ***Q4A3V1.***

```
Q4A3R2<- dsap_data %>% group_by(Medu) %>% filter(avgGrade>15) %>%
         summarise(counts = n(),  percentage = n()/length(which(dsap_data$avgGrade>15))*100)
Q4A3R2
Q4A3V2 <-ggplot(Q4A3R2, aes(x="", y =percentage, fill=as.factor(Medu))) + geom_col(color = "white") +
         coord_polar("y", start = 0) +
          theme(panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
                plot.title = element_text(size = 20, face = "bold")) +
         geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                   color = c("white"), position = position_stack(vjust=0.5)) +
         ggtitle("Shows the percentage of the Students that scored\n> 15 average marks and their Mother's Education Level") +
         labs(fill="Mother's Education Level")+
         scale_fill_manual(values = c("#483D8B", "#1E90FF","#2F4F4F", "#4B0082","#DA70D6"),
                    labels = c("1 - Primary Education (4th grade)",
                               "2 - 5th to 9th grade",
                               "3 - Secondary education",
                               "4 - Higher education"))
Q4A3V2
```

Figure 168 *shows the R code used to create the data visualization figure of* ***Q4A3V2.***

As shown in the code figure above, for the bar graph, the **Medu** and **avgGradeRange** are grouped and counted for this analysis. For the pie chart, only the **Medu** was grouped and calculated the students' percentage by filtering the **avgGrade.**

```
> Q4A3R1
# A tibble: 18 x 3
# Groups:   Medu [5]
    Medu avgGradeRange counts
   <int> <chr>          <int>
 1     0 06-10              2
 2     0 11-15              5
 3     1 00-05             25
 4     1 06-10             70
 5     1 11-15             42
 6     1 16-20              3
 7     2 00-05             22
 8     2 06-10            102
 9     2 11-15            101
10     2 16-20             10
11     3 00-05             27
12     3 06-10             82
13     3 11-15            103
14     3 16-20             20
15     4 00-05             13
16     4 06-10             94
17     4 11-15            156
18     4 16-20             45
>
```

Figure 169 *shows the output of the* ***Q4A3R1.***

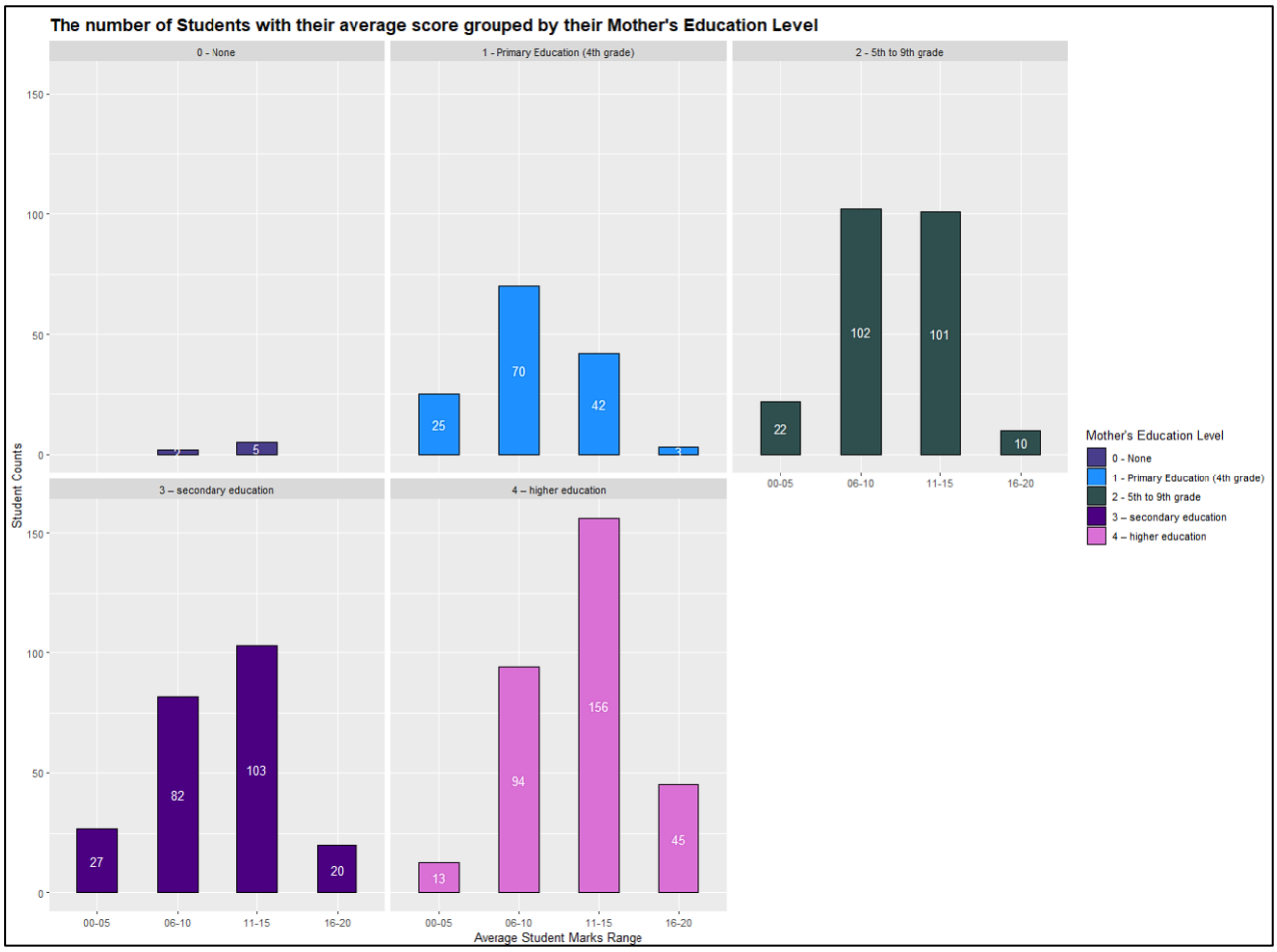


Figure 170 *shows the bar graph output of the* ***Q4A3V1****.*

The figure 169 above displays the output of the execution of the **Q4A3R1**, which shows the grouped counts of total students for each case of selected attributes that are the mother's education level and their average grade range. The figure 170 above shows the outcome of the bar graph plotted after the execution of the **Q4A3V1** variable that displays the students' counts and the average grade range of students grouped based on their mother's education level.

```
> Q4A3R2
# A tibble: 4 x 3
   Medu counts percentage
  <int>  <int>      <dbl>
1     1      3       3.85
2     2     10      12.8
3     3     20      25.6
4     4     45      57.7
>
```

Figure 171** shows the output of the **Q4A3R2.

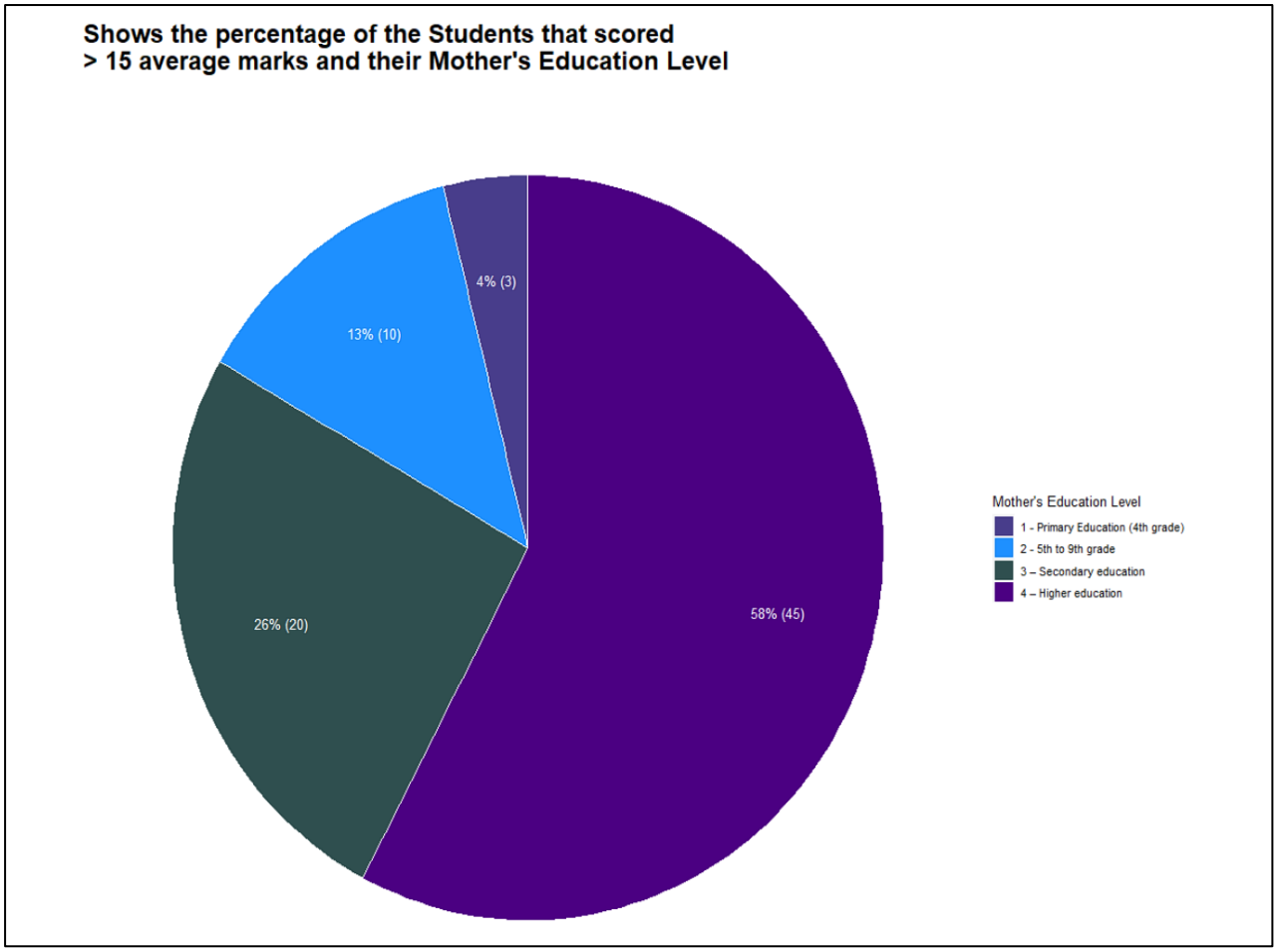


***Figure 172** shows the pie chart output of the **Q4A3V2**.*

The figure 171 above shows the execution output of the **Q4A3R2**. In figure 172, the pie chart that has been drawn displays the calculated total student counts and the percentage of them who scored more than 15 average marks grouped with their mother's education level.

Summary for Data Findings

1) A large percentage of students had mothers who studied for higher education (Level 4).
2) Very few students that are total of 7 students' mother educational level was none (Level 0).
3) 58% of students who scored more than 15 average marks had mothers who studied for higher education.
4) There are none of the students who scored more than 15 average marks had mothers without education.

Explanation for the Data Findings.

Based on the data findings, it was noticed that the largest quantity of students' mothers had studied for higher education. The least number of students had mothers without education When looking deeper, it was seen the most percentage of students who scored more than 15 average marks had mothers that studied for higher education. It could be because these mothers who had studied till higher education would have known the importance of education in their children's lives and they will prioritise it in order to make sure that their children excel in it. Furthermore, with higher education, these mothers could have also taught their children and prepared them for their examinations. According to Ghafoor and Kauser (2015), they further supported in their research that educated mothers tend to guide their children and prepare them in a better way for their future sake. Hence, it can be concluded that mothers who studied till higher education would positively influence the students' overall academic accomplishments.

4.4.4 Analysis 4-4: Finding the relationship between a father's education and a student's average mark.

The relationship between the education level of the father and their average marks will be analysed in this analysis. A bar graph and a pie chart have been drawn for this analysis.

```
#Analysis 4 - 4
#Finding the relationship between father's education and students' average mark.
Q4A4R1 <- dsap_data %>% group_by(Fedu, avgGradeRange) %>% summarise(counts = n())
Q4A4R1
Q4A4V1 <-ggplot(Q4A4R1, aes(x=avgGradeRange, y = counts, fill=as.factor(Fedu))) +
        geom_bar(stat="identity",width = 0.5, color="black") +
        ggtitle("The number of Students with their average score grouped by their Father's Education Level")+
        labs(x="Average Student Marks Range", y = "Student Counts", fill="Father's Education Level")+
        theme(plot.title = element_text(size = 15, face = "bold")) +
        scale_fill_manual(values=c("#483D8B", "#1E90FF","#2F4F4F", "#4B0082","#DA70D6"),
                          labels=c(c("0 - None",
                                     "1 - Primary Education (4th grade)",
                                     "2 - 5th to 9th grade",
                                     "3 - secondary education",
                                     " 4 - higher education")))+
        facet_wrap(~Fedu, labeller = as_labeller(c(`0`= "0 - None",
                                                   `1`= "1 - Primary Education (4th grade)",
                                                   `2`= "2 - 5th to 9th grade",
                                                   `3`="3 - secondary education",
                                                   `4`=" 4 - higher education"))) +
        geom_text(aes(label=counts),  position = position_stack(vjust = 0.5), color = "white")
Q4A4V1
```

Figure 173 *shows the R code used to create the data visualization figure of* ***Q4A4V1.***

```
Q4A4R2<- dsap_data %>% group_by(Fedu) %>% filter(avgGrade>15) %>%
        summarise(counts = n(),  percentage = n()/length(which(dsap_data$avgGrade>15))*100)
Q4A4R2
Q4A4V2 <-ggplot(Q4A4R2, aes(x="", y =percentage, fill=as.factor(Fedu))) + geom_col(color = "white") + coord_polar("y", start = 0) +
         theme(panel.background = element_blank(),
               axis.title = element_blank(),
               axis.text = element_blank(),
               axis.line = element_blank(),
               axis.ticks= element_blank(),
               plot.title = element_text(size = 20, face = "bold")) +
         geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                   color = c("white"), position = position_stack(vjust=0.5)) +
         ggtitle("Shows the percentage of the Students that scored\n> 15 average marks and their Father's Education Level") +
         labs(fill="Father's Education Level")+
         scale_fill_manual(values = c("#483D8B", "#1E90FF","#2F4F4F", "#4B0082","#DA70D6"),
                   labels = c("1 - Primary Education (4th grade)",
                              "2 - 5th to 9th grade",
                              "3 - Secondary education",
                              "4 - Higher education"))
Q4A4V2
```

Figure 174 *shows the R code used to create the data visualization figure of* ***Q4A4V2.***

As shown in the code figure above, for the bar graph, the **Fedu** and **avgGradeRange** are grouped and counted for this analysis. For the pie chart, only the **Fedu** was grouped and calculated the students' percentage by filtering the **avgGrade.**

```
> Q4A4R1
# A tibble: 17 x 3
# Groups:   Fedu [5]
    Fedu avgGradeRange counts
   <int> <chr>          <int>
 1     0 11-15              4
 2     1 00-05             28
 3     1 06-10             92
 4     1 11-15             58
 5     1 16-20              9
 6     2 00-05             27
 7     2 06-10             99
 8     2 11-15            129
 9     2 16-20             15
10     3 00-05             18
11     3 06-10            102
12     3 11-15             88
13     3 16-20             25
14     4 00-05             14
15     4 06-10             57
16     4 11-15            128
17     4 16-20             29
```

Figure 175 *shows the output of the* ***Q4A4R1.***

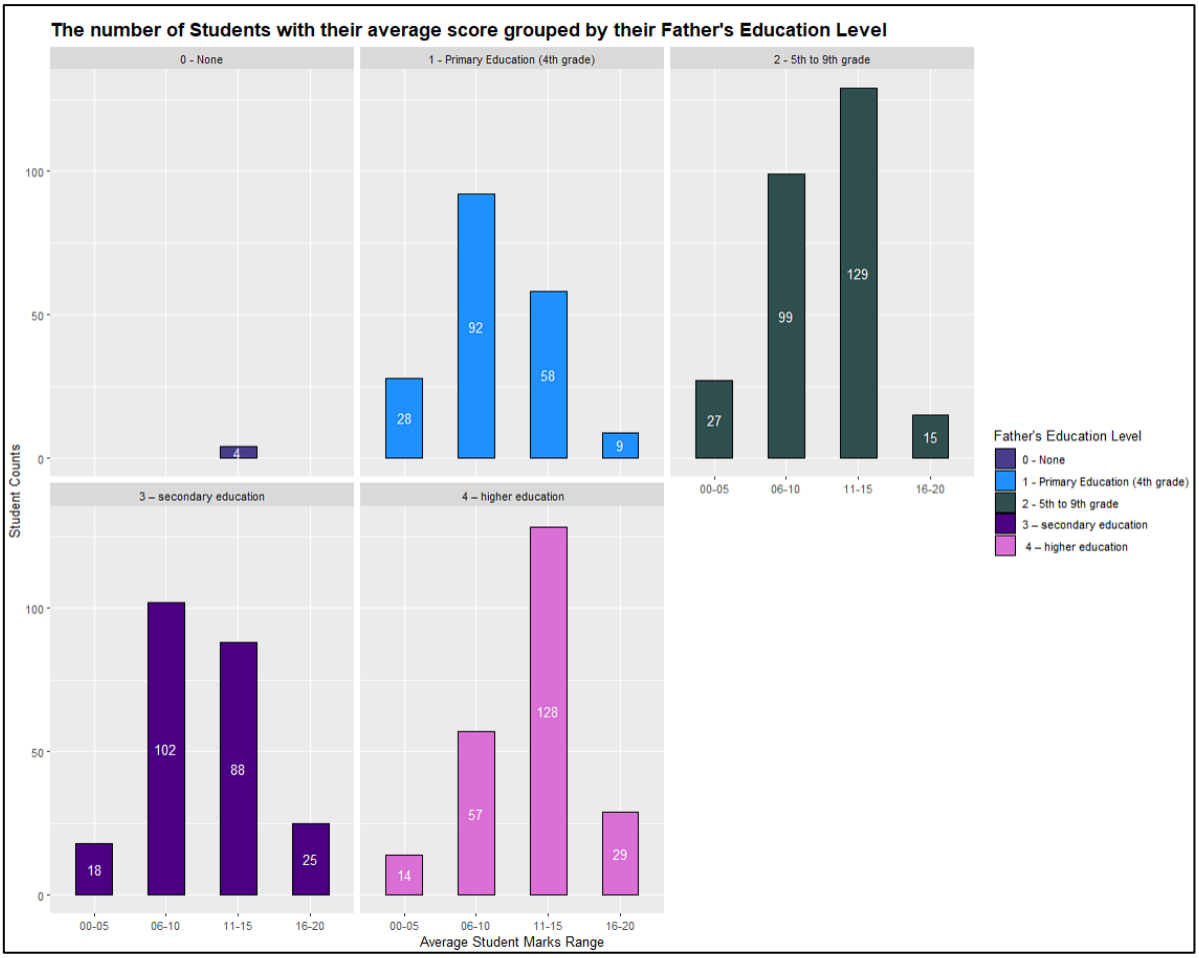


Figure 176 *shows the bar graph output of the* ***Q4A4V1****.*

The figure 175 above displays the output of the execution of the **Q4A4R1**, which shows the grouped counts of total students for each case of selected attributes that are the father's education level and their average grade range. The figure 176 above shows the outcome of the bar graph plotted after the execution of the **Q4A4V1** variable that displays the students' counts and the average grade range of students grouped based on their father's education level.

```
> Q4A4R2
# A tibble: 4 x 3
   Fedu counts percentage
  <int>  <int>      <dbl>
1     1      9       11.5
2     2     15       19.2
3     3     25       32.1
4     4     29       37.2
```

Figure 177** shows the output of the **Q4A4R2.

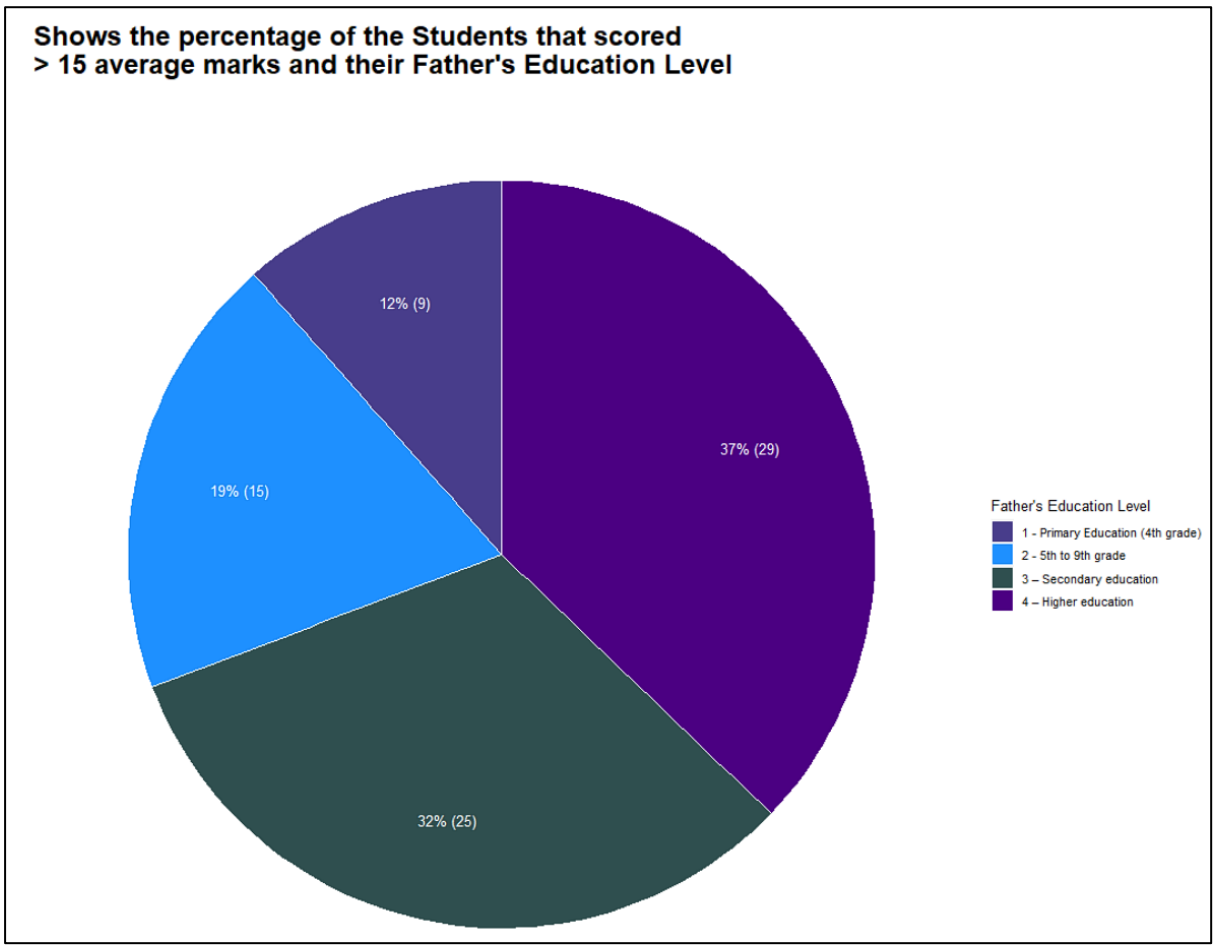


***Figure 178** shows the pie chart output of the **Q4A4V2**.*

The figure 177 above shows the execution output of the **Q4A4R2**. In figure 178, the pie chart that has been drawn displays the calculated total student counts and the percentage of them who scored more than 15 average marks grouped with their father's education level.

Summary for Data Findings

1) A large number of students had fathers who studied for 5th to 9th grade (Level 2).
2) Only a total of 4 students had fathers without education (Level 0).
3) 37% of the students who scored more than 15 average marks had fathers with higher education (Level 4).
4) 0% of the students who scored more than 15 average marks had fathers with no education

Explanation for the Data Findings.

Based on the data findings, it can be noticed that the most number of students who had fathers studied from 5th to 9th grade. The smallest portion of students with only 4 of them had fathers without any education. Moreover, looking deeper into the data findings, it was found that the majority percentage of students who scored more than 15 average marks had fathers with higher education levels. Also, not a single one of the students who scored more than 15 average marks had fathers with no education. From this, it can be stared that a fathers play a quite a role in impacting the students. It is basically the same as the mothers' education level. When the father has studied a fair education level, they have to be known how important the academic it is for their child's life. These father with high education could have helped their child in whatever study-related activities in order to make their child do their greatest. According to Idris et al. (2020), they additionally supported in their research that the fathers with high education also tend to contribute positively to their children's academic achievements When have studied for high-level education, they most probably had a very good profession which they can provide all the required study-related facilities for their children. This will also help the students a lot to achieve better results. Hence, it is true that fathers' education level also slightly affects the student's academic performance.

4.4.5 Analysis 4-5: Finding the correlation between the quality of the mother's education level, the father's education level, and the student's average marks.

The correlation between the quality of the mother's and father's education level and their average marks will be analysed in this analysis. A bar graph and a table have been drawn for this analysis.

```
#Analysis 4 - 5
#Finding the relationship between the quality of the mother's education level, the father's education level, and the student's average marks.
Q4A5R1<- dsap_data %>%  group_by(Medu,Fedu,avgGradeRange) %>% summarise(counts = n())
Q4A4R1
Q4A5R1
Q4A5V1<- ggplot(Q4A5R1, aes(x=avgGradeRange, y=counts ,fill = as.factor(Medu))) +
         geom_bar(stat = "identity", position = position_dodge2(),width = 0.5, color="black") +
         ggtitle("The number of Students with their Average Student Marks grouped by\ntheir Mother's and Father's Education Level") +
         theme(plot.title = element_text(size = 15, face = "bold")) +
          labs(fill = "Mothers's Education Level", x="Average Student Marks Range", y = "Student Counts")+
         facet_grid(Medu~Fedu, labeller =labeller(.rows=c(`0`= "Medu: 0 - None",
                                                          `1`= "Medu: 1 - Primary Education (4th grade)",
                                                          `2`= "Medu: 2 - 5th to 9th grade",
                                                          `3`= "Medu: 3 - Secondary Education",
                                                          `4`= "Medu: 4 - Higher Education"),
                                                  .cols=c(`0`= "Fedu: 0 - None",
                                                          `1`= "Fedu: 1 - Primary Education (4th grade)",
                                                          `2`= "Fedu: 2 - 5th to 9th grade",
                                                          `3`= "Fedu: 3 - Secondary Education",
                                                          `4`= "Fedu: 4 - Higher Education"))) +
         geom_text(aes(label=counts), vjust=-0.3) + ylim(0, 80) +
         scale_fill_manual(values=c("#483D8B", "#1E90FF","#2F4F4F", "#4B0082","#DA70D6"),
                           labels = c("0 - None",
                                      "1 - Primary Education (4th grade)",
                                      "2 - 5th to 9th grade",
                                      "3 - Secondary education",
                                      "4 - Higher education"))
Q4A5V1
```

Figure 179 *shows the R code used to create the data visualization figure of* ***Q4A5V1.***

```
Q4A5R2<- dsap_data %>% group_by(Fedu, Medu) %>%
         filter(avgGrade> 15) %>% summarise(counts = n())
View(Q4A5R2)
```

Figure 180 *shows the R code used to create the data visualization figure of* ***Q4A5R2.***

As shown in the code figure above, for the bar graph, the **Medu**, **Fedu** and **avgGradeRange** are grouped and counted for this analysis. For the table, only the **Fedu** and **Medu** were grouped and counted the total number of students by filtering the **avgGrade.**

```
> Q4A5R1
# A tibble: 57 x 4
# Groups:   Medu, Fedu [20]
    Medu  Fedu avgGradeRange counts
   <int> <int> <chr>          <int>
 1     0     1 06-10              2
 2     0     2 11-15              5
 3     1     0 11-15              2
 4     1     1 00-05             18
 5     1     1 06-10             49
 6     1     1 11-15             19
 7     1     1 16-20              3
 8     1     2 00-05              5
 9     1     2 06-10             10
10     1     2 11-15             18
# ... with 47 more rows
```

Figure 181 *shows the output of the* ***Q4A5R1.***

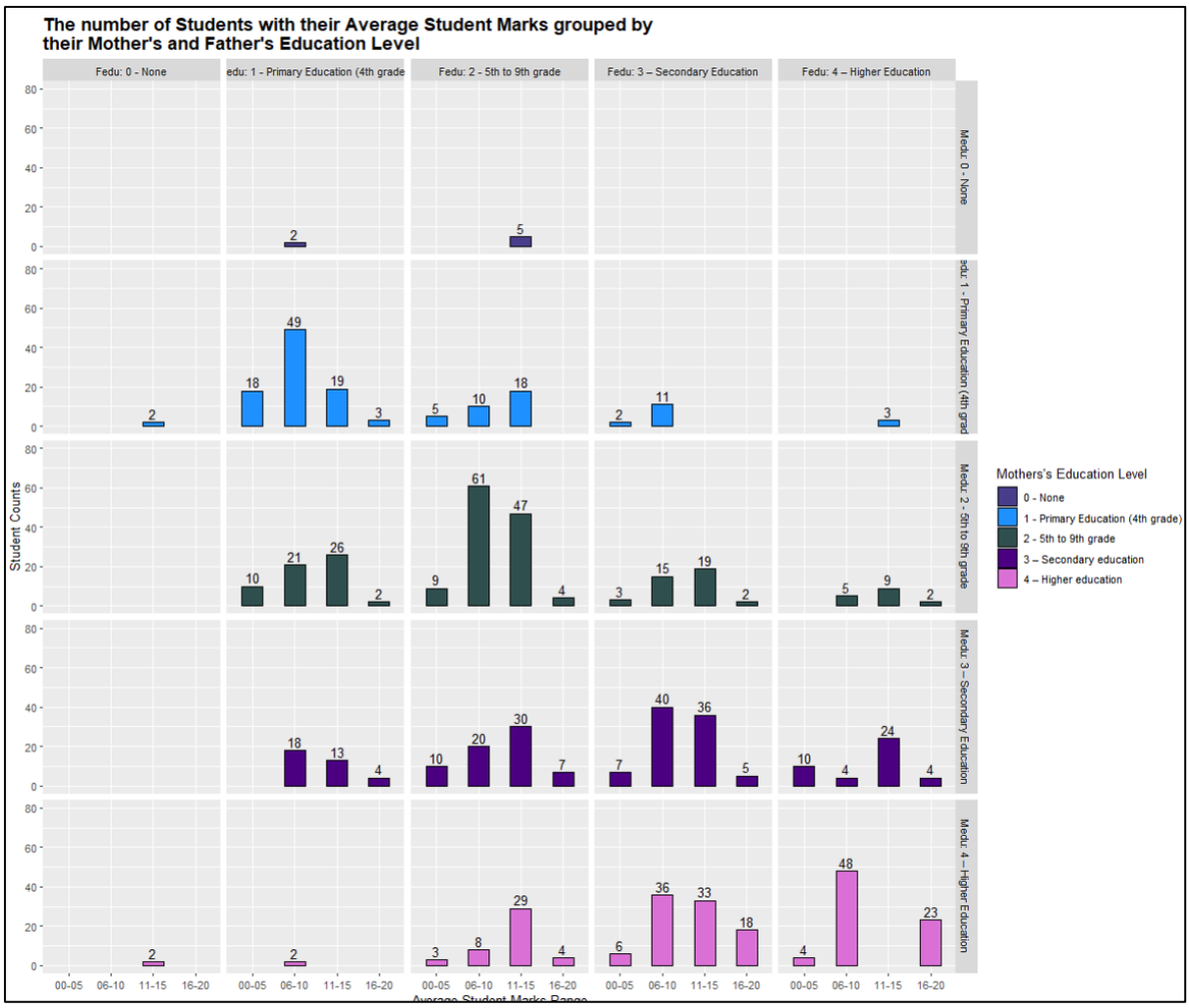


Figure 182 *shows the bar graph output of the* ***Q4A5V1****.*

The figure 181 above displays the output of the execution of the **Q4A5R1**, which shows the grouped counts of total students for each case of selected attributes that are the father's education level, mother's education level, and their average grade range. The figure 182 above

shows the outcome of the bar graph plotted after the execution of the **Q4A5V1** variable that displays the students' counts and the average grade range of students grouped based on both parents' education levels.

.

	Fedu	Medu	counts
1	1	1	3
2	1	2	2
3	1	3	4
4	2	2	4
5	2	3	7
6	2	4	4
7	3	2	2
8	3	3	5
9	3	4	18
10	4	2	2
11	4	3	4
12	4	4	23

***Figure 183** shows the table output of the **Q4A5R2**.*

The figure 182 above shows the execution result of the **Q4A5R2** that displays the total students count of them who scored more than 15 average marks grouped based on their father's and mother's education level.

Summary for Data Findings

1) The largest number of students had parents who studied till higher education (Level 4).
2) There are no students from the categories of father and mother with no education, the father with secondary education and mother with no education, the father with higher education and mother with no education, the father with no education and mother with 5th to 9th-grade education and the father with no education and mother with secondary education.
3) Most of the students that are 24 of them who scored more than 15 average marks had both parents with higher education.
4) The second most of the students are 18 of them who scored more than 15 average marks had fathers with secondary education and mothers with higher education.

Explanation for the Data Findings.

Based on the data findings, it can be witnessed that most number of students had parents who studied till higher education. When narrowing down those who got more than 15 average marks, it was also been noticed that most students' fathers and mothers had studied for higher education. Yet, the second-highest number of students who scored more than 15 average marks had their fathers who had studied for secondary education and mothers who had studied for higher education. From this, it can be expressed that both parents' education plays quite an important role in impacting the students' academic performance. As fairly educated parents knew the significance of education, they desire their children to do the greatest in it. Well, they will also make sure that they fully assist their children in terms of any necessities for their education. Also, when both parents had high education, they probably would have made shifts in educating their children where whenever one of them is busy with their work, the other one will help the children. This was further supported that highly educated parents would have spent more time actively with their children in order to expand their children's skills and talents (Clearinghouse Technical Assistance Team, 2020). From the literature review, it was also proved that parents' educational level has a high chance of impacting the students' academic performance. Hence, it can be confirmed that parents' education level is another important attribute that making quite an influence on students' marks.

4.4.6 Analysis 4-6: Finding the correlation between mother's job, father's job and students' average marks.

The correlation between the mother's and father's job and their average marks will be analysed in this analysis. A bar graph and a table have been drawn for this analysis.

```
#Analysis 4 - 6
#Finding the correlation between mother's job, father's job and students' average marks.
Q4A6R1<- dsap_data %>%  group_by(Mjob,Fjob,avgGradeRange) %>% summarise(counts = n())
Q4A6R1
Q4A6V1<- ggplot(Q4A6R1, aes(x=avgGradeRange, y=counts, fill = as.factor(Fjob))) +
         geom_bar(stat = "identity", position = position_dodge2(),width = 0.5, color="black") +
         ggtitle("The number of Students with their Average Student Marks grouped by\ntheir Mother's and Father's Job") +
         theme(plot.title = element_text(size = 15, face = "bold")) +
         labs(fill = "Father's Job", x="Average Student Marks Range", y = "Student Counts")+
         facet_grid(Fjob~Mjob, labeller = labeller(.rows= c(`teacher`  = "Fjob - Teacher",
                                                            `health`  = "Fjob - Health - care related",
                                                            `services`= "Fjob - Civil Services (e.g. administrative or police)",
                                                            `at_home` = "Fjob - At Home",
                                                            `other`   = "Fjob - Other"),
                                                   .cols = c(`teacher` = "Mjob - Teacher",
                                                             `health`  = "Mjob - Health - care related",
                                                             `services`= "Mjob - Civil Services (e.g. administrative or police)",
                                                             `at_home` = "Mjob - At Home",
                                                             `other`   = "Mjob - Other"))) +
         geom_text(aes(label=counts), vjust=-0.3) + ylim(0, 80) +
         scale_fill_manual(values=c("#483D8B", "#1E90FF","#2F4F4F", "#4B0082","#DA70D6"),
         labels = c("At Home",
                    "Health - care related",
                    "Other",
                    "Civil Services (e.g. administrative or police)",
                    "Teacher"))
Q4A6V1
```

Figure 184 *shows the R code used to create the data visualization figure of* ***Q4A6V1.***

As shown in the code figure above, for the bar graph, the **Mjob**, **Fjob** and **avgGradeRange** are grouped and counted for this analysis.

```
> Q4A6R1
# A tibble: 70 x 4
# Groups:   Mjob, Fjob [24]
   Mjob    Fjob     avgGradeRange counts
   <chr>   <chr>    <chr>          <int>
 1 at_home at_home  06-10              5
 2 at_home at_home  11-15             10
 3 at_home at_home  16-20              2
 4 at_home health   06-10              2
 5 at_home health   11-15              2
 6 at_home other    00-05             13
 7 at_home other    06-10             38
 8 at_home other    11-15             24
 9 at_home other    16-20              2
10 at_home services 00-05              2
# ... with 60 more rows
```

Figure 185 *shows the output of the* ***Q4A6R1.***

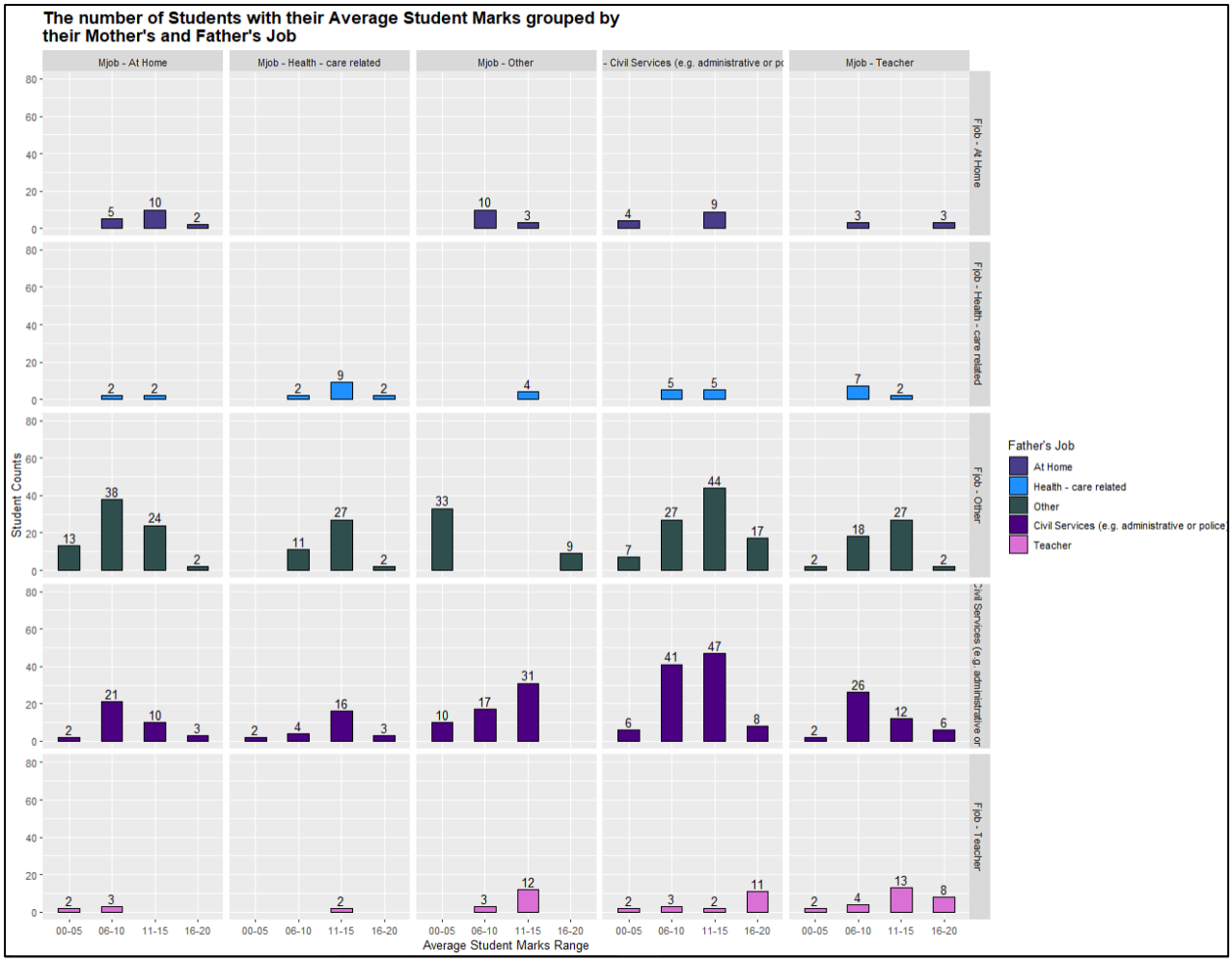


Figure 186 *shows the bar graph output of the* ***Q4A6V1****.*

The figure 185 above displays the output of the execution of the **Q4A6R1**, which shows the grouped counts of total students for each case of selected attributes that are the mother's job, father's job and their average grade range. The figure 186 above shows the outcome of the bar graph plotted after the execution of the **Q4A6V1** variable that displays the students' counts and the average grade range of students grouped based on both parents' jobs.

Summary for Data Findings

1) The largest portion of students had parents who both worked other jobs.
2) There are no students from the categories of fathers staying at home and mothers working in the health-related category.
3) The majority of students that is total of 17 of them who scored more than 15 average marks had fathers working in other jobs and mothers working in civil services.
4) The highest number of students, which is a total of 33 students who scored less than equal to 5 average marks are both parents working in other jobs.

Explanation for the Data Findings.

Based on the data findings, it was caught that most of the students' parents work in other jobs as compared to other categories. After looking in-depth, it was noticed that a big number of students who scored more than 15 average marks had fathers working in other jobs and mothers working in services related jobs. Also, the highest number of students who got scored less than equal 5 average marks had both fathers and mothers working in other jobs. It can be guessed that the other jobs might require more additional engagement and time to get the income. This would have caused the parents not to show proper attention and guidance to their children which would provoke them to get very low average marks. While not the case for those who scored very good average marks, because their parents who work jobs such as services that require less time and engagement would have shown full attention to their children and assisted them in every possible way in order to make them do the best in their studies. When the parents spend quality time with their children, the children would be mentally happy, and they will have a mindset to do the best in their academics for their parents. This was further supported that when these students don't get much parental supervision, they tend to participate in risky behaviours which could harm their academic performance badly (Heinrich, 2014). Hence, it can be concluded that students that spend quality time with them would affect their children's academic performance positively.

Conclusion

As a conclusion for this fourth question, based on the observed analyses, it can be reasoned that family attributes play quite an essential role in influencing the students' education performance. This could be because they are the ones who the students can primarily seek for help. With that help, these students would be highly motivated and try their best to do well in their academics. Especially, their education level would be one of the most important attributes that would help students if both of their students had studied a fair amount of education. Overall, the family need to continuously do their part in assisting and supporting these students to help them better in their academics.

4.5 Question 5: How do students' personal lives impact their marks?

This fifth question will be analysing whether the personal lives of a student would affect their overall academic grades. The student's attributes that will be covered in this question are workday alcohol consumption, weekend alcohol consumption, decision to pursue higher studies, health status and nursery school attendance.

4.5.1 Analysis 5-1: Finding the relationship between students' workday alcohol consumption and their average marks.

The relationship between the students' workday alcohol consumption and their average marks will be analysed in this analysis. A horizontal stacked bar graph and a stacked bar graph have been drawn for this analysis.

```
#=======Question 5========#
#Question 5: How do students' personal lives impact their marks?
#Analysis 5 - 1
#Finding the relationship between students' workday alcohol consumption and their average marks.
Q5A1R1<- dsap_data %>% group_by(Dalc,avgGradeRange) %>% summarise(counts = n())
Q5A1R1
Q5A1V1<- ggplot(Q5A1R1, aes(x=avgGradeRange, y=counts, fill=as.factor(Dalc))) +
         geom_bar(stat="identity",width = 0.5, color="black") +
         ggtitle("The number of Students with their average score grouped by Workday Alcohol Consumption.")+
         labs(x="Average Student Marks Range", y = "Student Counts", fill="Workday Alcohol Consumption")+
         theme(plot.title = element_text(size = 15, face = "bold")) + coord_flip()+
         scale_fill_manual(values=c("#8B008B", "#00CED1", "#1E90FF", "#7CFC00", "#CD853F"),
                           labels = c("1 - Very Low",
                                      "2 - Low",
                                      "3 - Medium",
                                      "4 - High",
                                      "5 - Very High"))+
         geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A1V1
```

Figure 187 *shows the R code used to create the data visualization figure of* ***Q5A1V1.***

```
Q5A1R2<- dsap_data %>% group_by(Dalc, avgGrade) %>% filter(avgGrade>15) %>% summarise(counts = n())
Q5A1R2
Q5A1V2<- ggplot(Q5A1R2, aes(x=avgGrade, y=counts, fill=as.factor(Dalc))) +
         geom_bar(stat="identity",width = 0.5, color="black") +
         ggtitle("The number of Students with their average score > 15 grouped by Workday Alcohol Consumption.")+
         labs(x="Average Student Marks Range", y = "Student Counts", fill="Workday Alcohol Consumption")+
         theme(plot.title = element_text(size = 15, face = "bold")) +
         scale_fill_manual(values=c("#8B008B", "#00CED1", "#1E90FF", "#7CFC00", "#CD853F"),
                           labels = c("1 - Very Low",
                                      "2 - Low",
                                      "3 - Medium",
                                      "4 - High",
                                      "5 - Very High"))+
         geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A1V2
```

Figure 188 *shows the R code used to create the data visualization figure of* ***Q5A1V2.***

As shown in the code figures above, for both the graphs, the **Dalc** and **avgGradeRange** are grouped and counted for this analysis. Additionally, for the stacked bar graph, the **avgGrade** was filtered.

```
> Q5A1R1
# A tibble: 17 x 3
# Groups:   Dalc [5]
    Dalc avgGradeRange counts
   <int> <chr>           <int>
 1     1 00-05              60
 2     1 06-10             224
 3     1 11-15             287
 4     1 16-20              66
 5     2 00-05              22
 6     2 06-10              75
 7     2 11-15              71
 8     2 16-20              10
 9     3 00-05               2
10     3 06-10              33
11     3 11-15              27
12     3 16-20               2
13     4 00-05               3
14     4 06-10               9
15     4 11-15               9
16     5 06-10               9
17     5 11-15              13
```

Figure 189 *shows the output of the* ***Q5A1R1.***

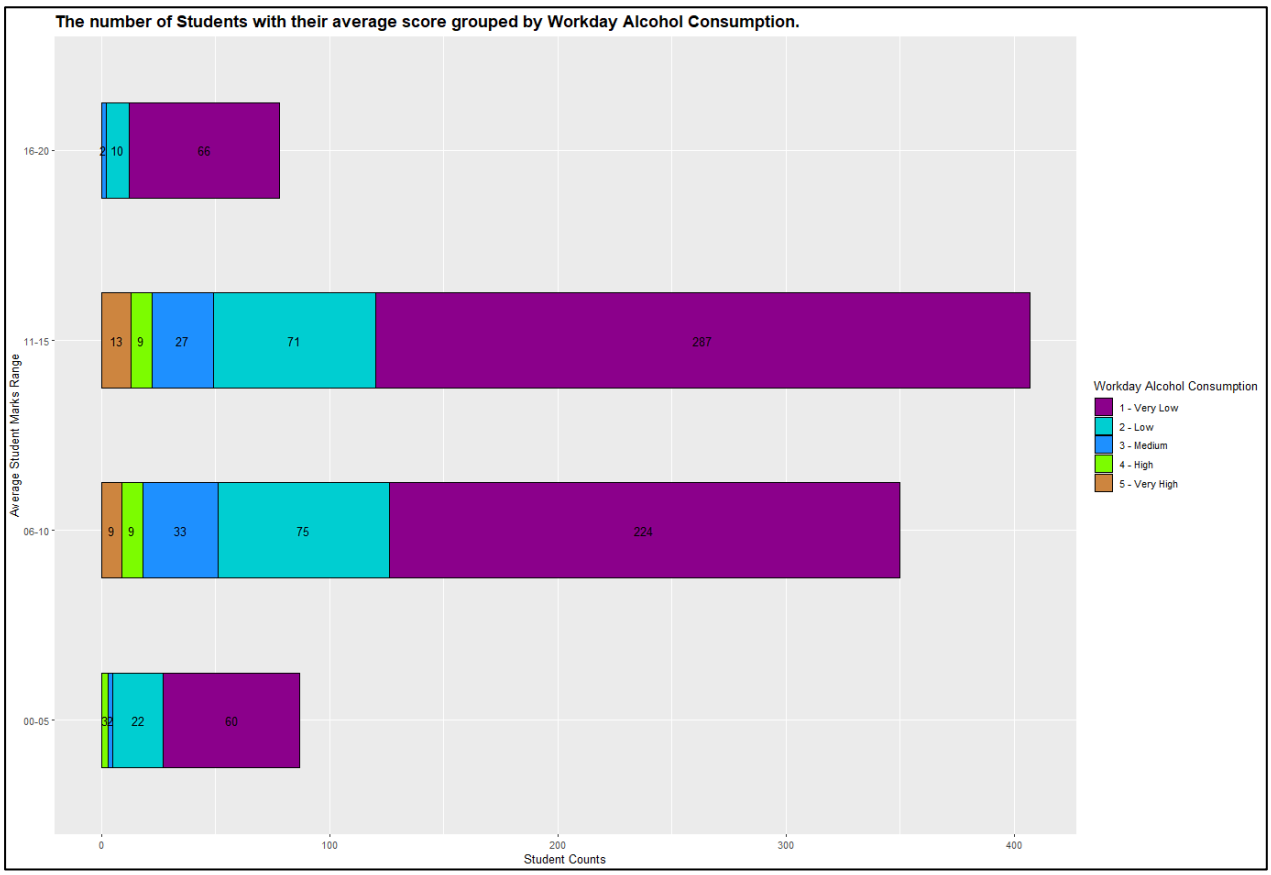


Figure 190 *shows the horizontal stacked bar graph output of the* ***Q5A1V1****.*

The figure 189 above displays the output of the execution of the **Q5A1R1**, which shows the grouped counts of total students for each case of selected attributes that are the workday alcohol consumption level and their average grade range. The figure 190 above shows the outcome of the horizontally stacked bar graph plotted after the execution of the **Q5A1V1** variable that displays the students' counts and the average grade range of students grouped based on their workday alcohol consumption level.

```
> Q5A1R2
# A tibble: 8 x 3
# Groups:   Dalc [3]
   Dalc avgGrade counts
  <int>    <dbl>  <int>
1     1       16     24
2     1       17     16
3     1       18     16
4     1       19     10
5     2       16      4
6     2       17      2
7     2       18      4
8     3       17      2
```

Figure 191** shows the output of the **Q5A1R1.

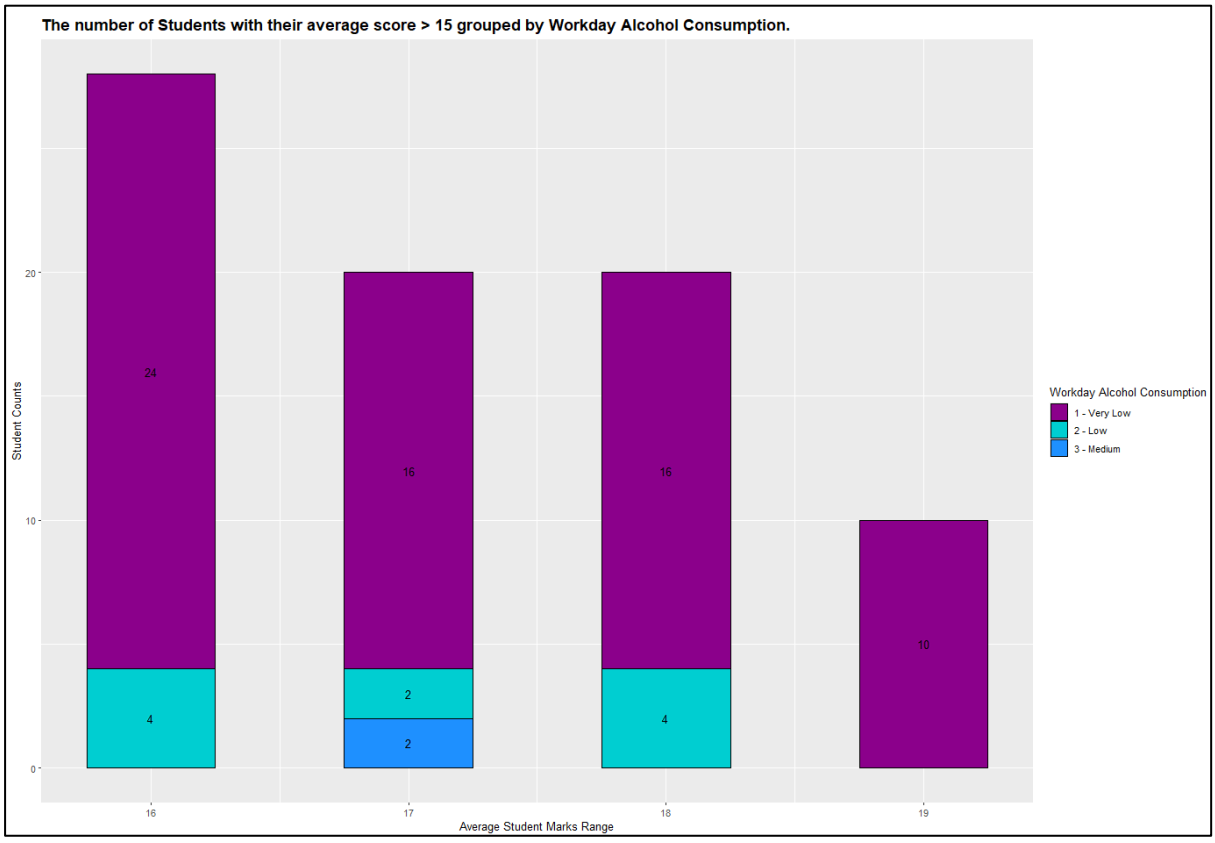


***Figure 192** shows the stacked bar graph output of the **Q5A1V2**.*

The figures 191 displays the execution output of the **Q5A1R2**, which displays the calculated total student counts grouped by the workday alcohol consumption and average grade. In figure 192, the stacked bar graph illustrates the number of students that scored more than 15 average marks with their workday alcohol consumption level.

Summary for Data Findings

1) The largest number of students had very low (Level 1) workday alcohol consumption.
2) The least number of students had very high (Level 5) workday alcohol consumption.
3) There is no single student who scored above 15 average marks had more than level 3 workday alcohol consumption.
4) Not a single student who scored the average mark of 19 had a workday alcohol consumption level of more than 1.

Explanation for the Data Findings.

Based on the data findings, it was seen that majority of students had very low workday alcohol consumption levels. It same goes for students who scored more than 15 average marks. Moreover, looking in-depth at those who scored 19 average marks, all of them also had very low workday alcohol consumption levels. It could be because these students could have understood the various negative effects of daily drinking on their academic performance and made the choice of not drinking alcohol during the workday. After all, consuming alcohol during workdays would cause them to lose focus and couldn't concentrate during their classes, which can make them lose a lot of knowledge. It was further supported in research that high weekday alcohol consumption has a significant negative effect on the students' grades (Analytics Vidhya, 2020). Hence, it is true that a low level of alcohol consumption during the workday will positively impact their results.

4.5.2 Analysis 5-2: Finding the relationship between students' weekend alcohol consumption and their average marks.

The correlation between the students' weekend alcohol consumption and their average marks will be analysed in this analysis. A two stacked bar graphs have been created for this analysis.

```
#Analysis 5 - 2
#Finding the relationship between students' weekend alcohol consumption and their average marks.
Q5A2R1<- dsap_data %>%  group_by(Walc,avgGradeRange) %>% summarise(counts = n())
Q5A2R1
Q5A2V1<-  ggplot(Q5A2R1, aes(x=avgGradeRange, y=counts, fill=as.factor(Walc))) +
          geom_bar(stat="identity",width = 0.5, color="black") +
          ggtitle("The number of Students with their average score grouped by Weekend Alcohol Consumption.")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="Weekend Alcohol Consumption")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          scale_fill_manual(values=c("#8B008B", "#00CED1", "#1E90FF", "#7CFC00", "#CD853F"),
                            labels = c("1 - Very Low",
                                       "2 - Low",
                                       "3 - Medium",
                                       "4 - High",
                                       "5 - Very High"))+
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A2V1
```

Figure 193 *shows the R code used to create the data visualization figure of* ***Q5A2V1.***

```
Q5A2R2<- dsap_data %>% group_by(Walc, avgGrade) %>% filter(avgGrade>15) %>% summarise(counts = n())
Q5A2R2
Q5A2V2<- ggplot(Q5A2R2, aes(x=avgGrade, y=counts, fill=as.factor(Walc))) +
          geom_bar(stat="identity",width = 0.5, color="black") +
          ggtitle("The number of Students with their average score > 15 grouped by Weekend Alcohol Consumption.")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="Weekend Alcohol Consumption")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          scale_fill_manual(values=c("#8B008B", "#00CED1", "#1E90FF", "#7CFC00", "#CD853F"),
                            labels = c("1 - Very Low",
                                       "2 - Low",
                                       "3 - Medium",
                                       "4 - High",
                                       "5 - Very High"))+
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A2V2
```

Figure 194 *shows the R code used to create the data visualization figure of* ***Q5A2V2.***

As shown in the code figures above, for both the graphs, the **Walc** and **avgGradeRange** are grouped and counted for this analysis. Additionally, for the second stacked bar graph, the **avgGrade** was filtered.

```
> Q5A2R1
# A tibble: 20 x 3
# Groups:   Walc [5]
    Walc avgGradeRange counts
   <int> <chr>          <int>
 1     1 00-05             36
 2     1 06-10            118
 3     1 11-15            143
 4     1 16-20             49
 5     2 00-05             22
 6     2 06-10             66
 7     2 11-15            100
 8     2 16-20             13
 9     3 00-05             14
10     3 06-10             85
11     3 11-15             88
12     3 16-20             10
13     4 00-05             11
14     4 06-10             51
15     4 11-15             51
16     4 16-20              2
17     5 00-05              4
18     5 06-10             30
19     5 11-15             25
20     5 16-20              4
>
```

Figure 195 *shows the output of the* ***Q5A2R1.***

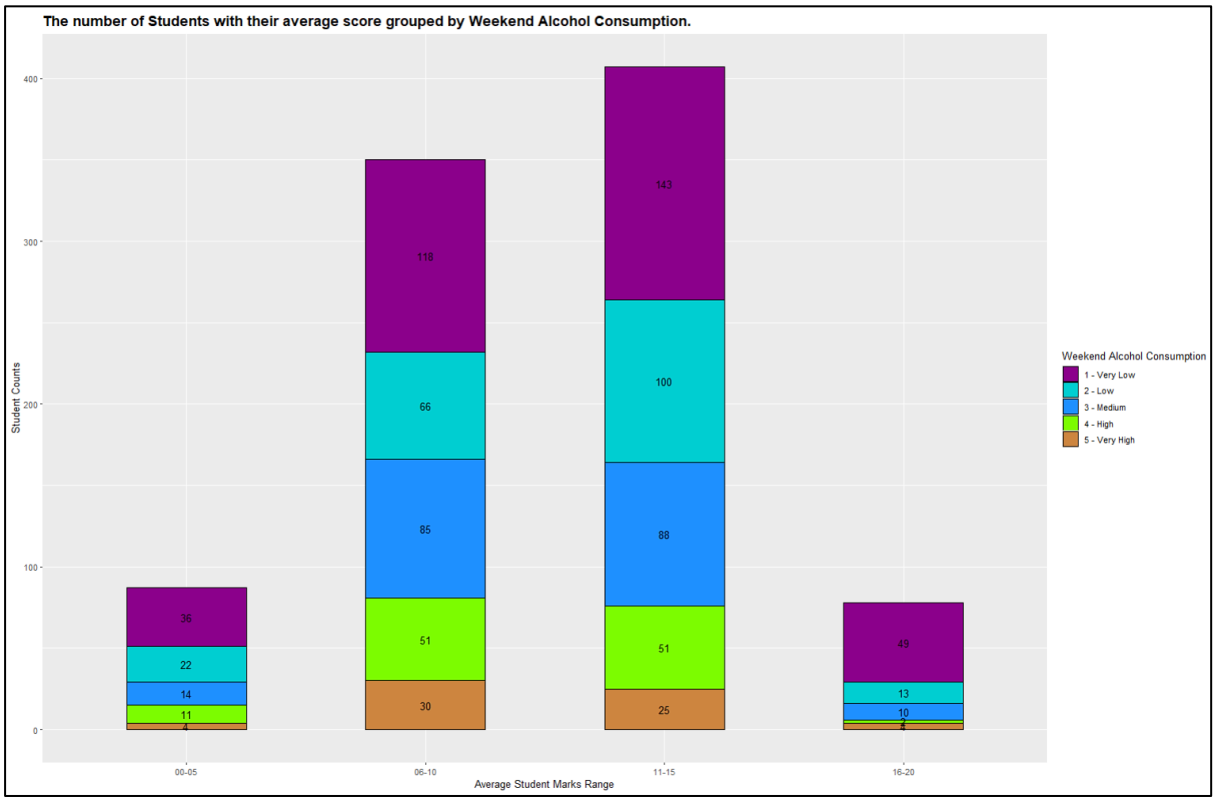


Figure 196 *shows the stacked bar graph output of the* ***Q5A2V1****.*

The figure 195 above displays the output of the execution of the **Q5A2R1**, which shows the grouped counts of total students for each case of selected attributes that are the weekend alcohol consumption level and their average grade range. The figure 196 above shows the outcome of the stacked bar graph plotted after the execution of the **Q5A2V1** variable that displays the students' counts and the average grade range of students grouped based on their weekend alcohol consumption level.

```
Q5A2R2
A tibble: 13 x 3
Groups:   Walc [5]
   Walc avgGrade counts
  <int>    <dbl>  <int>
1     1       16     15
2     1       17     12
3     1       18     12
4     1       19     10
5     2       16      9
6     2       17      2
7     2       18      2
8     3       16      2
9     3       17      4
10    3       18      4
11    4       17      2
12    5       16      2
13    5       18      2
```

Figure 197** shows the output of the **Q5A2R2.

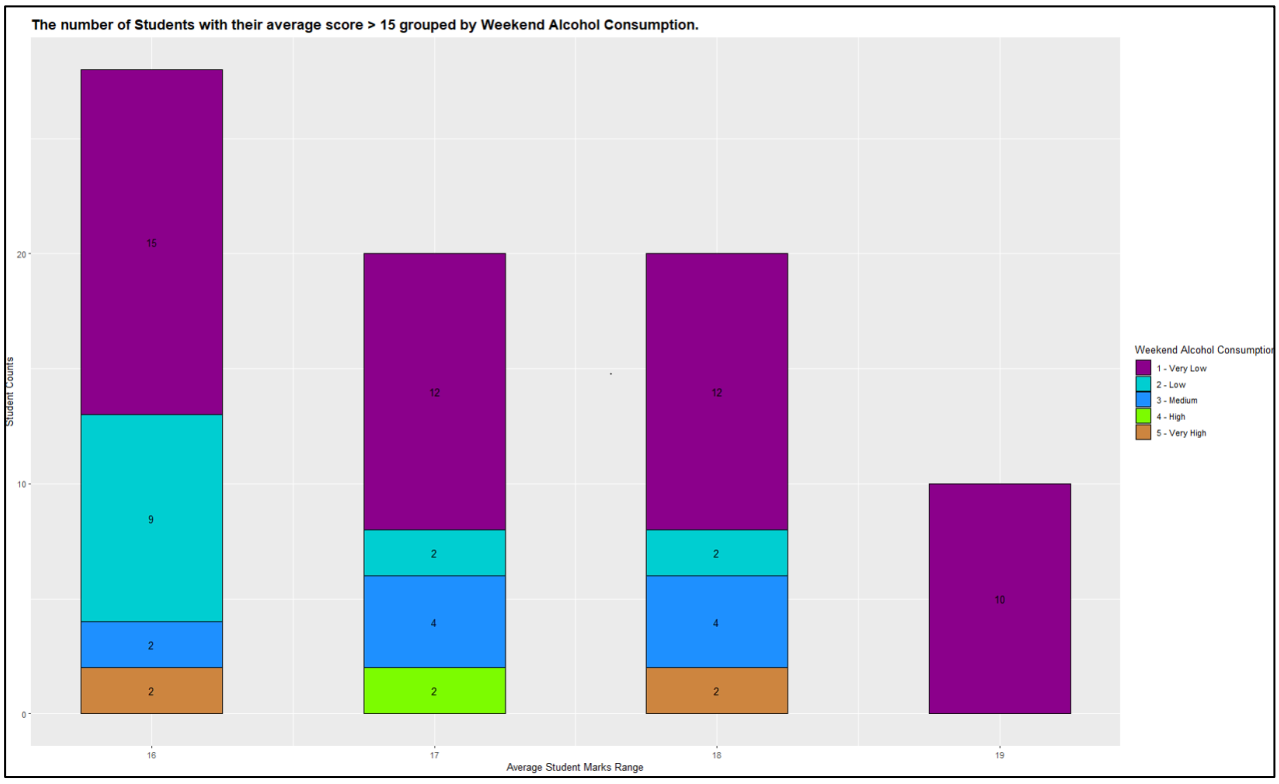


***Figure 198** shows the stacked bar graph output of the **Q5A2V2**.*

The figures 197 displays the execution output of the **Q5A2R2**, which displays the calculated total student counts grouped by the weekend alcohol consumption and average grade. In figure 198, the stacked bar graph illustrates the number of students that scored more than 15 average marks with their weekend alcohol consumption level.

Summary for Data Findings

1) The largest number of students had very low (Level 1) weekend alcohol consumption.
2) The least number of students had very high (Level 5) weekend alcohol consumption.
3) There is a very minimal number of students who scored more than 15 average marks had weekend alcohol consumption high (Level 4) and very high (Level 5).
4) All the students who scored an average mark of 19 had very low (Level 1) weekend alcohol consumption levels

Explanation for the Data Findings.

Based on the data findings, same as the previous analysis, the most number of students that are 342 total of them, had very low alcohol consumption during the weekend. While the very least students that are 63 total of them, had consumed very high alcohol during the weekend. When seeing students who scored more than 15 average marks, the majority of them had very low weekend alcohol consumption. When looking into deeper students who scored 19 as their average mark, all the of them had also very low weekend alcohol consumption levels. The results are basically identical to the previous analysis. It can be stated that these students who scored very excellent scores have known that a high level of alcohol consumption during the weekend could have negatively affected their academics which could cause them to lose focus and concentration during accomplishing any study-related stuff. It was further stated high alcohol consumption during the weekend can affect the students' brains as they could lose the memory of the things they learnt in classes (Placzek, 2020). Hence, it is proven that alcohol consumption during the weekend plays quite a role in impacting the students' grades, as students with a low level of alcohol drinking would do better in their academics.

4.5.3 Analysis 5-3: Finding the correlation between students' decision to pursue higher studies and their average marks.

The correlation between the students' decision to pursue higher studies and their average marks will be analysed in this analysis. A stacked bar graph and a treemap graph have been drawn for this analysis.

```
#Analysis 5 - 3
#Finding the correlation between students' decision to pursue higher studies and their average marks.
Q5A3R1<- dsap_data %>% group_by(higher, avgGradeRange) %>% summarise(counts = n())
Q5A3R1

Q5A3V1<-  ggplot(Q5A3R1, aes(avgGradeRange, y=counts, fill = as.factor(higher))) +
          geom_bar(stat = "identity", position = position_dodge2(preserve = 'single'), width=0.9) +
          ggtitle("The number of Students with their average score grouped by their decision of pursuing for Higher Studies.") +
          theme(plot.title = element_text(size = 15, face = "bold")) +
          labs(fill = "Students Higher Studies Decided Status", x="Average Students Marks Range", y = "Student Counts")+
          geom_text(aes(label=counts), position = position_dodge2(1), vjust=-0.5) +
          scale_fill_manual(values = c("#191970","#FF1493"),labels = c("1 - No", "2 - Yes"))
Q5A3V1
```

Figure 199 *shows the R code used to create the data visualization figure of* ***Q5A3V1.***

```
Q5A3R2<- dsap_data %>% group_by(higher) %>% filter(avgGrade>15) %>%
         summarise(counts = n(),  percentage = n()/length(which(dsap_data$avgGrade>15))*100)
Q5A3R2

Q5A3V2 <- ggplot(Q5A3R2, aes(x=percentage, y="", fill = as.factor(higher), area = percentage)) + geom_treemap()+
          theme(legend.justification="top",
                panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
          plot.title = element_text(size = 20, face = "bold")) +
          ggtitle("Shows the percentage of the Students that scored\n> 15 average mark grouped
                  by their decision of pursuing for Higher Studies.") +
          labs(fill="Students Higher Studies Decided Status")+
          scale_fill_manual(values = c("#4B0082","#DA70D6"), labels = c("2 - Yes")) +
          geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                            color = c("white"), place = "left")
Q5A3V2
```

Figure 200 *shows the R code used to create the data visualization figure of* ***Q5A3V2.***

As shown in the code figures above, for the bar graph, the **higher** and **avgGradeRange** are grouped and counted for this analysis. For the treemap graph, only the **higher** was grouped and calculated the students' percentage by filtering the **avgGrade.**

```
> Q5A3R1
# A tibble: 7 x 3
# Groups:   higher [2]
  higher avgGradeRange counts
   <int> <chr>          <int>
1      1 00-05             12
2      1 06-10             22
3      1 11-15             10
4      2 00-05             75
5      2 06-10            328
6      2 11-15            397
7      2 16-20             78
>
```

Figure 201 *shows the output of the* ***Q5A3R1.***

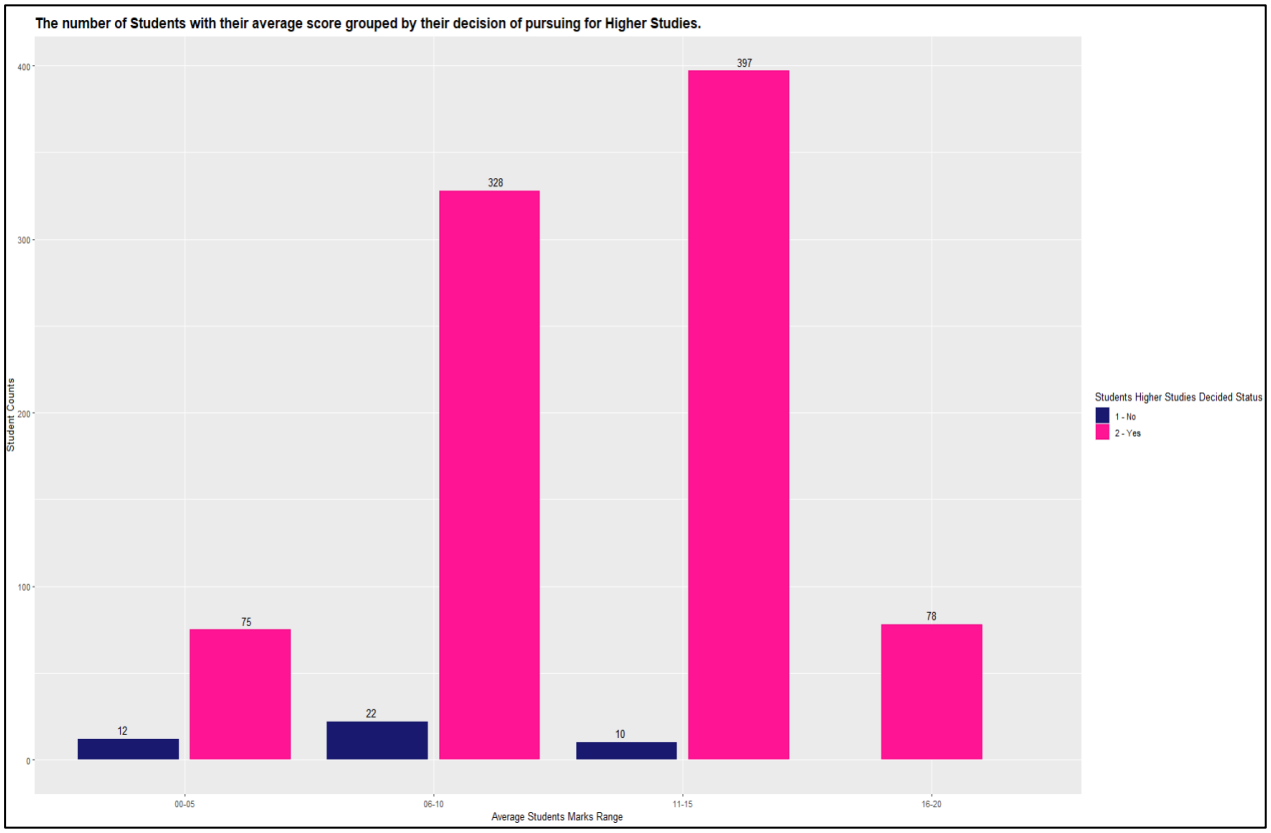


Figure 202 *shows the bar graph output of the* ***Q5A3V1****.*

The figure 201 above displays the output of the execution of the **Q5A3R1**, which shows the grouped counts of total students for each case of selected attributes that are the decision of taking higher education and their average grade range. The figure 202 above shows the result of the bar graph plotted after the execution of the **Q5A3V1** variable that displays the students' counts and the average grade range of students grouped based on their decision on taking higher education

```
> Q5A3R2
# A tibble: 1 x 3
  higher counts percentage
   <int>  <int>      <dbl>
1      2     78        100
```

Figure 203 *shows the output of the* ***Q5A3R2.***

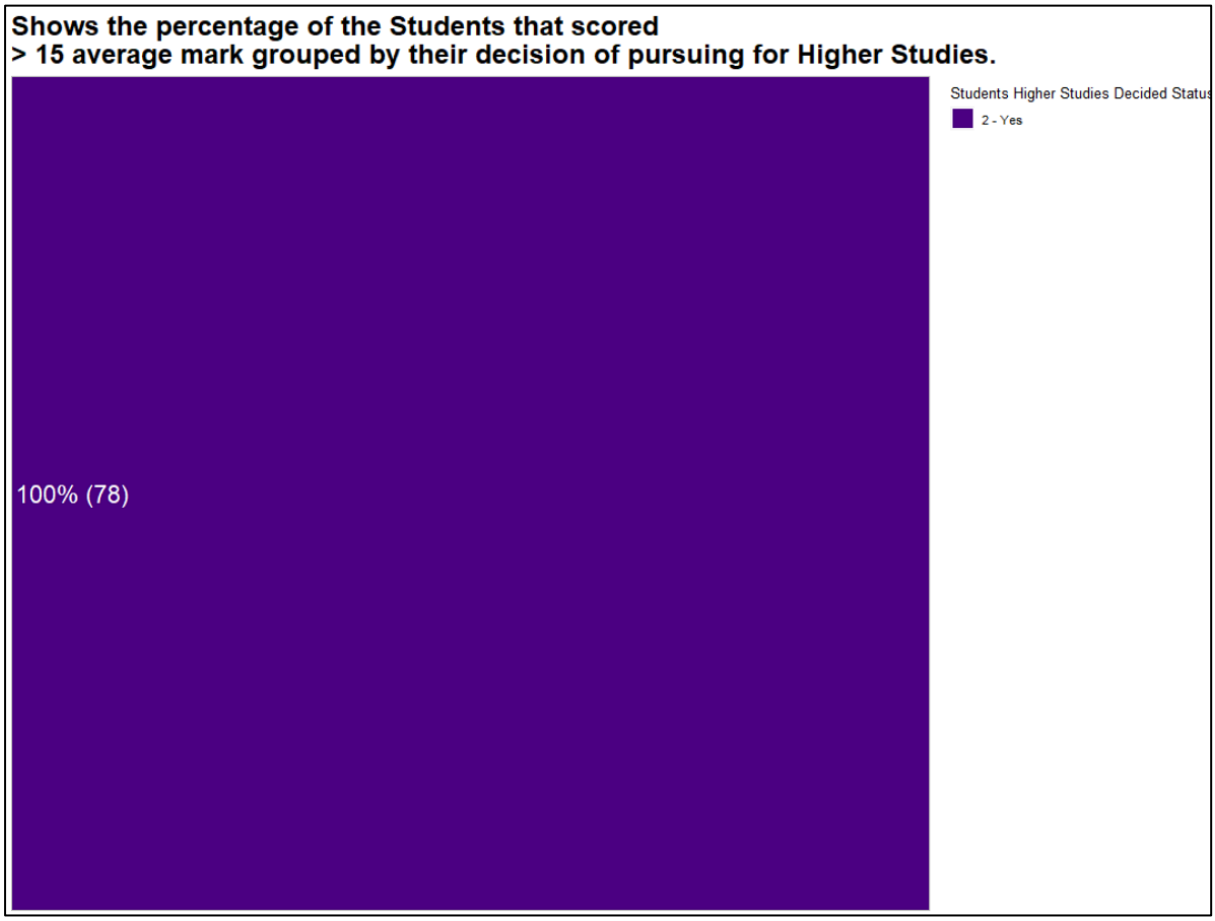


Figure 204 *shows the treemap graph output of the* ***Q5A3V2****.*

The figure 203 above shows the execution output of the **Q5A3R2**. In figure 204, the treemap graph that has been drawn displays the calculated total student counts and the percentage of them who scored more than 15 average marks grouped based on their decision of pursuing higher education.

Summary for Data Findings

1) The majority of students that is 878 of them have decided to pursue higher education.
2) Very least of students that is only 44 of them had decided to not pursue higher education.
3) 100% of students who scored more than 15 average marks had a desire to pursue higher education.

Explanation for the Data Findings.

Based on the data findings, it was found that a very large amount of students had chosen to pursue higher education in the future. While very few of them have determined not to pursue it in the future. When looking deeper, all of the students who scored more than 15 average marks have the desire to pursue higher education. From this, it can be assumed that these students who scored a very excellent average grade and were intended to pursue higher education took their academics very seriously, and they would have felt that it was their responsibility to put the effort and do well in their current academics so that they could pursue their desired higher education. According to Mati et al. (2016), it was further supported in their research that students who make proper decisions that are related to their lives tend to excel in their academics. Hence, it is proven that students who made the decision to pursue higher education have a higher possibility of them having great academic performance.

4.5.4 Analysis 5-4: Finding the relationship between students' health status and their average marks

The relationship between the students' health status and their average marks will be analysed in this analysis. A stacked bar graph and a treemap graph have been drawn for this analysis.

```
#Analysis 5 - 4
#Finding the relationship between students' health status and their average marks
Q5A4R1<- dsap_data %>% group_by(health,avgGradeRange) %>% summarise(counts = n())
Q5A4R1
Q5A4V1<-  ggplot(Q5A4R1, aes(x=avgGradeRange, y=counts, fill=as.factor(health))) +
          geom_bar(stat="identity",width = 0.5, color="black") +
          ggtitle("The number of Students with their average score grouped by Health Status.")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="Student Health Status")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          scale_fill_manual(values=c("#FFC0CB", "#FFE4C4", "#87CEFA","#F0E68C","#00FF7F"),
                            labels = c("1 - Very Low",
                                       "2 - Low",
                                       "3 - Okay",
                                       "4 - Good",
                                       "5 - Very Good"))+
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A4V1
```

Figure 205 *shows the R code used to create the data visualization figure of* ***Q5A4V1.***

```
Q5A4R2<- dsap_data %>% group_by(health) %>% filter(avgGrade>avgMeanGrade) %>%
          summarise(counts = n(),  percentage = n()/length(which(dsap_data$avgGrade>avgMeanGrade))*100)
Q5A4R2
Q5A4V2 <- ggplot(Q5A4R2, aes(x=percentage, y="", fill = as.factor(health), area = percentage)) + geom_treemap()+
          theme(legend.justification="top",
                panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank()
                ,axis.line = element_blank(),
                axis.ticks= element_blank(),
                plot.title = element_text(size = 20, face = "bold")) +
          ggtitle("Shows the percentage of the Students that scored\n> average mean mark grouped by their Health status.") +
          labs(fill="Health Status")+
          scale_fill_manual(values = c("#FFC0CB", "#FFE4C4", "#87CEFA","#F0E68C","#00FF7F"),
                            labels = c("1 - Very Low",
                                       "2 - Low",
                                       "3 - Okay",
                                       "4 - Good",
                                       "5 - Very Good")) +
          geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                            color = c("black"), place = "left")
Q5A4V2
```

Figure 206 *shows the R code used to create the data visualization figure of* ***Q5A4V2.***

As shown in the code figures above, for the stacked bar graph, the **health** and **avgGradeRange** are grouped and counted for this analysis. For the treemap graph, only the **health** was grouped and calculated the students' percentage by filtering the **avgGrade.**

```
Q5A4R1
A tibble: 20 x 3
Groups:   health [5]
 health avgGradeRange counts
  <int> <chr>          <int>
      1 00-05             10
      1 06-10             32
      1 11-15             49
      1 16-20             17
      2 00-05             10
      2 06-10             49
      2 11-15             37
      2 16-20             11
      3 00-05             20
      3 06-10             87
      3 11-15             94
      3 16-20             10
      4 00-05             13
      4 06-10             67
      4 11-15             51
      4 16-20             17
      5 00-05             34
      5 06-10            115
      5 11-15            176
      5 16-20             23
```

Figure 207 *shows the output of the* ***Q5A4R1.***

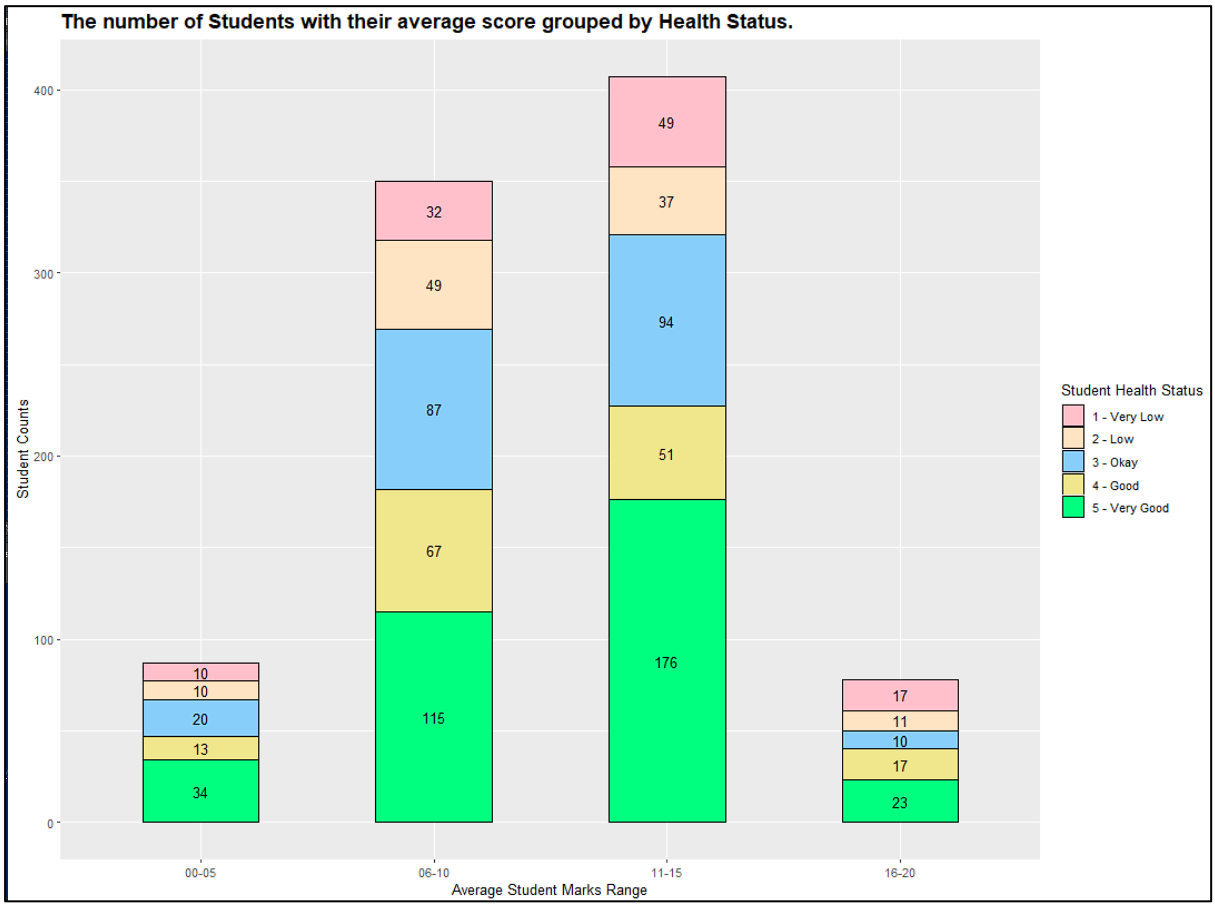


Figure 208 *shows the stacked bar graph output of the* ***Q5A4V1****.*

The figure 207 above displays the output of the execution of the **Q5A4R1**, which shows the grouped counts of total students for each case of selected attributes that are their health status and their average grade range. The figure 208 above shows the outcome of the stacked bar graph plotted after the execution of the **Q5A4V1** variable that displays the students' counts and the average grade range of students grouped based on health status level.

```
> Q5A4R2
# A tibble: 5 x 3
  health counts percentage
   <int>  <int>      <dbl>
1      1     58       14.8
2      2     48       12.2
3      3     74       18.9
4      4     51       13.0
5      5    161       41.1
```

Figure 209** shows the output of the **Q5A4R2.

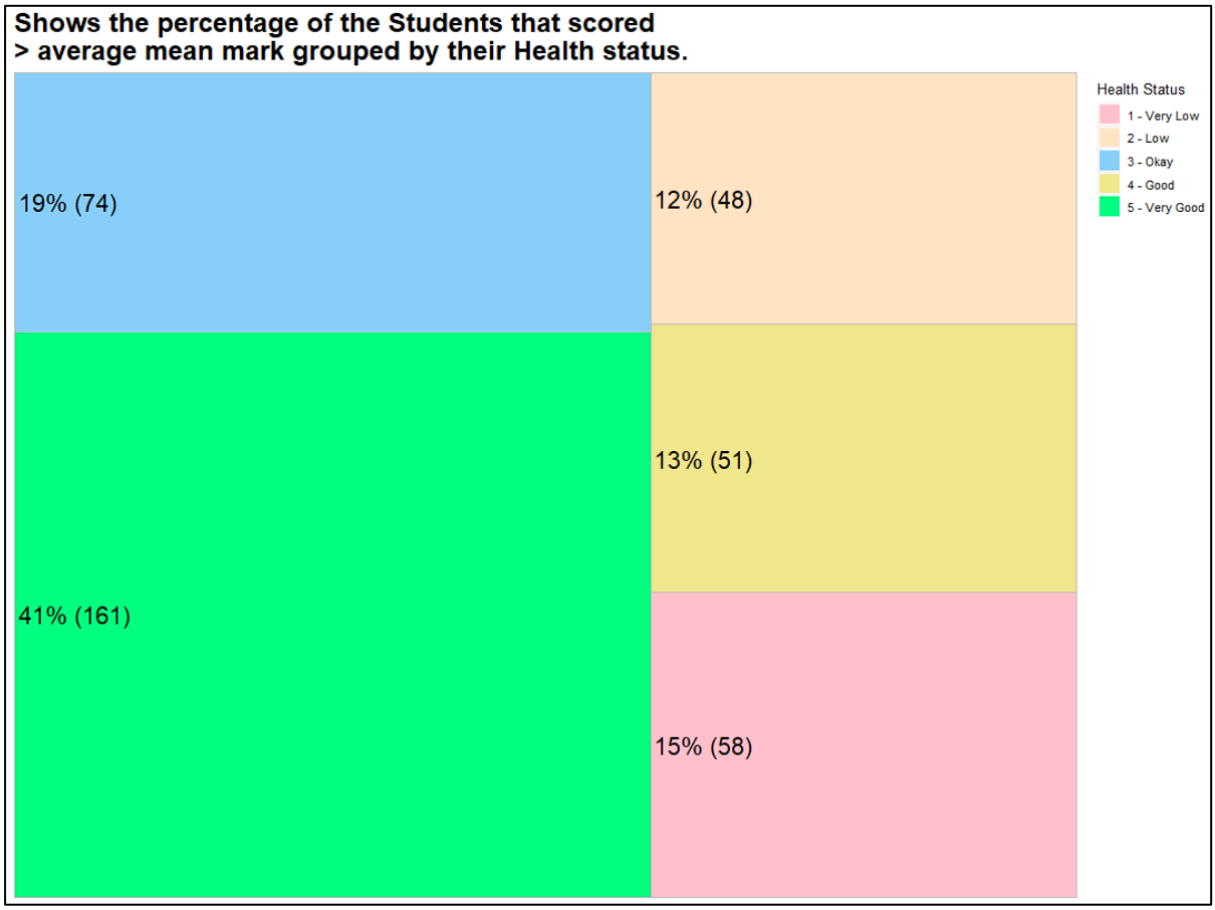


***Figure 210** shows the treemap graph output of the **Q5A4V2**.*

The figure 209 above shows the execution output of the **Q5A4R2**. In figure 210, the treemap graph that has been drawn displays the calculated total student counts and the percentage of them who scored more than the average mean mark grouped based on the health status level.

Summary for Data Findings

1) The majority of students that is 348 of them have had a very good (Level 5) health status.
2) A little number of students that is 107 total of them had low (Level 2) health status.
3) The largest number of students who scored more than the average mean grade had very good (Level 5) health status.

Explanation for the Data Findings.

Based on the data findings, it was clearly seen that most students had a very good health status. While a very minimal amount of students had low health status. Furthermore, students who scored more than the average mean grade also had most of them with very good health status. From this, it can be expressed good health status had a positive influence on affecting students' academic performance. With good health status, they would be physically felt active and better with that they can carry out all the academic-related tasks efficiently. Moreover, good health also would allow them to fully concentrate and focus during their classes, and they wouldn't possibly miss anything from the class. On the other hand, as a negative effect of bad health, it would lead the students to be less focused and attentive during the classes, and also increases their absences which causes them to lose a lot of knowledge. According to Matingwina (2018), it was further proven that bad health would cause students to confront education problems like insufficient retention, lack of concentration, high chances of failing their exams and dropping out of school. Hence, it is proven that health status plays an essential role in impacting a student's grade and students with good health whose most likely going to do better in their academics.

4.5.5 Analysis 5-5: Finding the relationship between students' nursery school attendance and their average grade.

The correlation between the students' nursery attendance and their average marks will be analysed in this analysis. A horizontal bar graph and a pie chart have been created for this analysis.

```
#Analysis 5 - 5
#Finding the relationship between students' nursery school attendance and their average grade.
Q5A5R1 <- dsap_data %>% group_by(nursery, avgGradeRange) %>% summarise(counts = n())
Q5A5R1

Q5A5V1 <-   ggplot(Q5A5R1, aes(x=avgGradeRange, y = counts, fill=as.factor(nursery))) +
            geom_bar(stat="identity",width = 0.5, color="black") +
            ggtitle("The number of Students with their average score grouped by their status of attending Nursery School.")+
            labs(x="Average Student Marks Range", y = "Student Counts", fill="Nursery School")+
            theme(plot.title = element_text(size = 15, face = "bold")) +
            scale_fill_manual(values=c("#6495ED", "#D8BFD8"), labels = c("1 - No", "2 - Yes"))+ coord_flip() +
            facet_wrap(~nursery, labeller = as_labeller(c(`1` = "Not Attended",
                                                          `2`= "2 - Attended"))) +
            geom_text(aes(label=counts),  position = position_stack(vjust = 0.5), color = "black")
Q5A5V1
```

Figure 211 *shows the R code used to create the data visualization figure of* ***Q5A5V1.***

```
Q5A5R2<- dsap_data %>% group_by(nursery) %>% filter(avgGrade>avgMeanGrade) %>%
        summarise(counts = n(),  percentage = n()/length(which(dsap_data$avgGrade>avgMeanGrade))*100)
 Q5A5R2
 Q5A5V2 <-ggplot(Q5A5R2, aes(x="", y =percentage, fill=as.factor(nursery))) + geom_col(color = "black") +
          coord_polar("y", start = 0) +
          theme(panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
                plot.title = element_text(size = 20, face = "bold")) +
           geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                     color = c("black"), position = position_stack(vjust=0.5)) +
          ggtitle("Shows the percentage of the Students that scored\n> average mean mark grouped by their status of atteding Nursery School.") +
          labs(fill="Nursery School Status")+
          scale_fill_manual(values = c("#F0E68C","#00FF7F"),
                            labels =  c("1 - No", "2 - Yes"))

 Q5A5V2
```

Figure 212 *shows the R code used to create the data visualization figure of* ***Q5A5V2.***

As shown in the code figures above, for the horizontal bar graph, the **nursery** and **avgGradeRange** are grouped and counted for this analysis. For the pie chart, only the **health** was grouped and calculated the students' percentage by filtering the **avgGrade.**

```
> Q5A5R1
# A tibble: 8 x 3
# Groups:   nursery [2]
  nursery avgGradeRange counts
    <int> <chr>          <int>
1       1 00-05             23
2       1 06-10             64
3       1 11-15             97
4       1 16-20              6
5       2 00-05             64
6       2 06-10            286
7       2 11-15            310
8       2 16-20             72
>
```

Figure 213 *shows the output of the* ***Q5A5R1.***

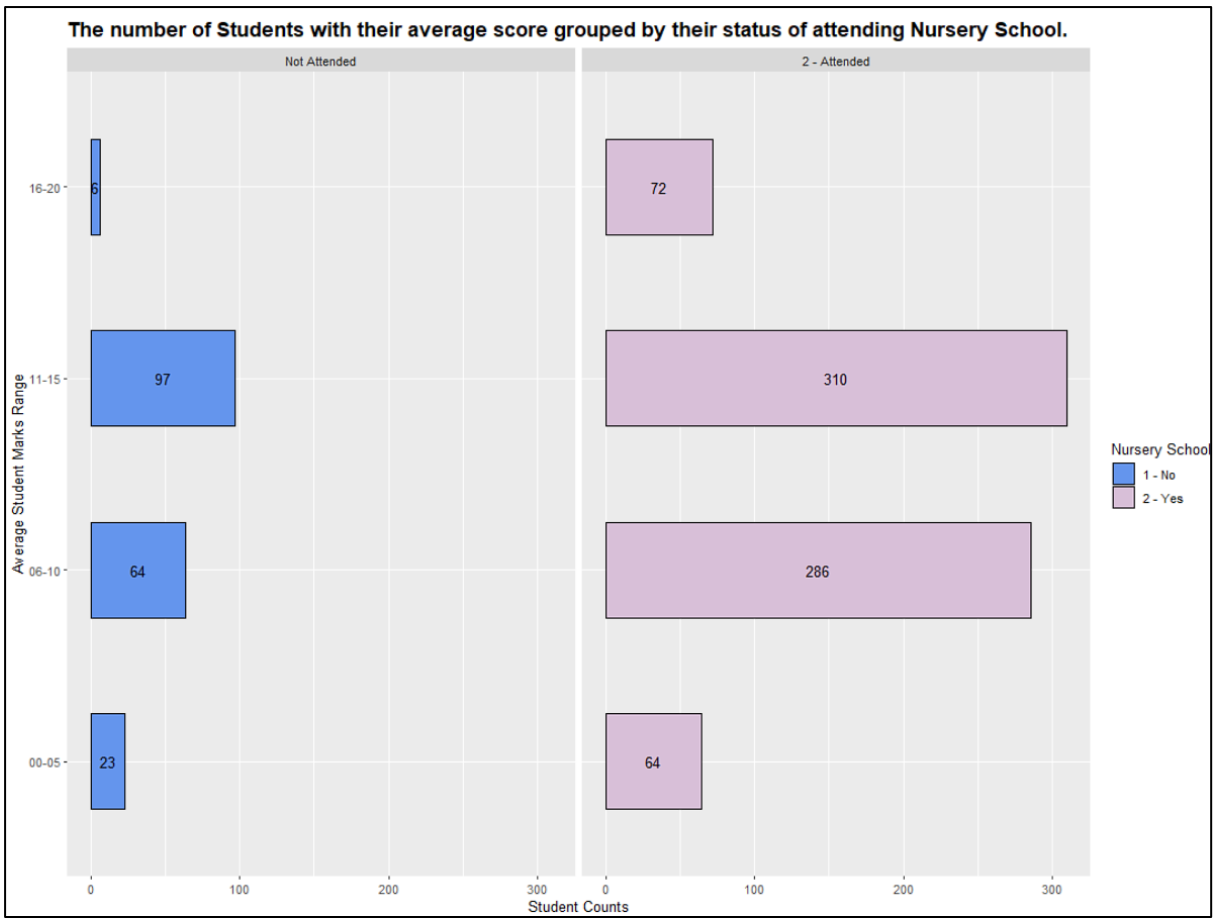


Figure 214 *shows the horizontal bar graph output of the* ***Q5A5V1****.*

The figure 213 above displays the output of the execution of the **Q5A5R1**, which shows the grouped counts of total students for each case of selected attributes that are their nursery school attendance and their average grade range. The figure 214 above shows the outcome of the horizontal bar graph plotted after the execution of the **Q5A5V1** variable that displays the students' counts and the average grade range of students grouped based on the status nursery school attendance status.

```
> Q5A5R2
# A tibble: 2 x 3
  nursery counts percentage
    <int>  <int>      <dbl>
1       1     78       19.9
2       2    314       80.1
```

Figure 215 *shows the output of the* ***Q5A5R2.***

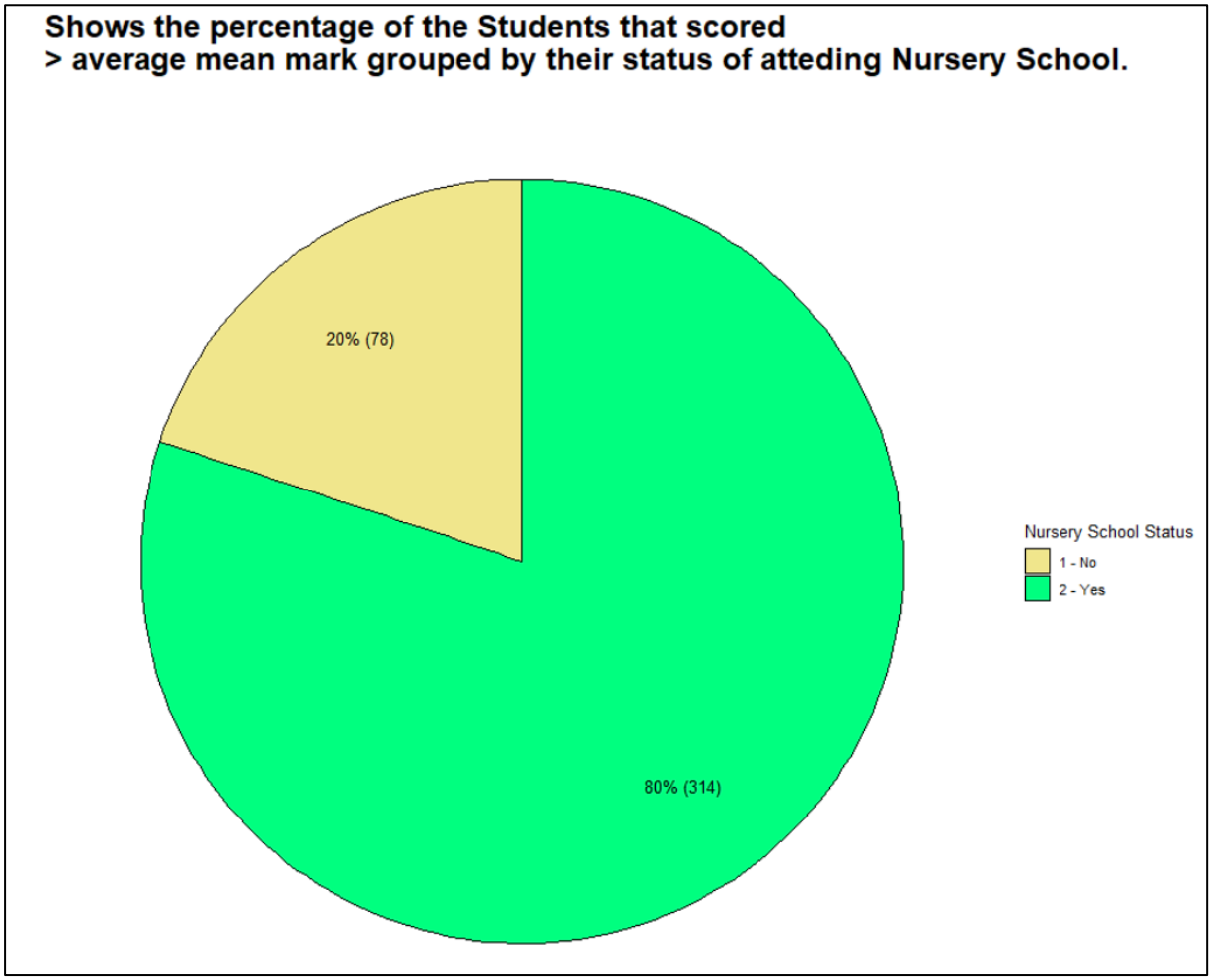


Figure 216 *shows the pie chart output of the* ***Q5A5V2****.*

The figure 215 above shows the execution output of the **Q5A5R2**. In figure 216, the pie chart that has been created displays the calculated total student counts and the percentage of them who scored more than the average mean mark grouped based on the status nursery school attendance status.

Summary for Data Findings

1) The majority of students that is 632 of them have attended nursery school.
2) A small number of students that is 190 total of them had not attended the nursery school.
3) 80% that is a total of 314 students who scored more than the average mean mark had attended nursery school

Explanation for the Data Findings.

Based on the data findings, it was noticed that most numbers of students had attended nursery school. While very least amount of these students didn't attend the nursery school. When looking in-depth, students who scored more than the average mean mark had most of those who attended nursery school. From this, it is safe to say that nursery school also do quite an influence in helping students to get good grades. As nursery school become the first step of a student's educational journey, it promotes the students' social life and emotional blossoming and they will be prepared for the rest of their lives. These students who have experienced nursery school would take more interest in their studies. Furthermore, they also will be more responsible for any education-related stuff. With prior knowledge from the nursery school, these students would also easily understand the learning resources very easily and quickly. According to Bibi and Ali (n.d.), in their research, further supported that these students who attended nursery school would be more confident and ask a lot of study related questions during the class right away which can clear all their doubts in their studies. This would help them do better in their academics. Hence, attendance at nursery school contributes a small part to influencing the students' academic performance where students who attended nursery school have a higher likelihood of doing better in their academics.

4.5.6 Analysis 5-6: Finding the relationship between students' number of absences and their average grade.

The correlation between the students' number of absences and their average marks will be analysed in this analysis. A scatterplot graph and a treemap have been created for this analysis.

```
#Analysis 6 -Finding the relationship between students' number of school absences and their average grade.
Q5A6R1 <- dsap_data %>% group_by(absences, avgGradeRange) %>% summarise(counts = n())
Q5A6R1

Q5A6V1 <- ggplot(Q5A6R1, aes(absences, counts)) + geom_point(aes(color=as.factor(avgGradeRange))) +
          ggtitle("The number of Students absences grouped by their average grade range.") +
          theme(plot.title = element_text(size = 15, face = "bold")) +
          labs( x="Number of Absences", y = "Student Counts") +
          xlim(0,80) +
          scale_color_manual(name = "Student Average Mark Range",
                             values = c(`00-05` = "darkblue",
                                        `06-10` = "red",
                                        `11-15` = "yellow",
                                        `16-20`= "green"))
Q5A6V1
```

Figure 217 *shows the R code used to create the data visualization figure of* ***Q5A6V1.***

```
Q5A6R2<- dsap_data %>% group_by(absences) %>% filter(avgGrade>15) %>%
         summarise(counts = n(),  percentage = n()/length(which(dsap_data$avgGrade>15))*100)
Q5A6R2

Q5A6V2 <-  ggplot(Q5A6R2, aes(x=percentage, y="", fill = as.factor(absences), area = percentage)) + geom_treemap()+
          theme(legend.justification="top",
                panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
                 plot.title = element_text(size = 20, face = "bold")) +
          ggtitle("Shows the percentage of the Students that scored\n> 15 average marks grouped by the number of absences.") +
          labs(fill="Total Absences")+
          geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                            color = c("black"), place = "left")
Q5A6V2
```

Figure 218 *shows the R code used to create the data visualization figure of* ***Q5A6V2.***

As shown in the code figures above, for the scatterplot graph, the **absences** and **avgGradeRange** are grouped and counted for this analysis. For the pie chart, only the **absences** was grouped and calculated the students' percentage by filtering the **avgGrade.**

```
> Q5A6R1
# A tibble: 69 x 3
# Groups:   absences [34]
   absences avgGradeRange counts
      <int> <chr>          <int>
 1        0 00-05             69
 2        0 06-10             73
 3        0 11-15            102
 4        0 16-20             27
 5        1 11-15              7
 6        2 06-10             52
 7        2 11-15             82
 8        2 16-20             16
 9        3 06-10              5
10        3 11-15             17
# ... with 59 more rows
```

Figure 219 *shows the output of the* ***Q5A6R1.***

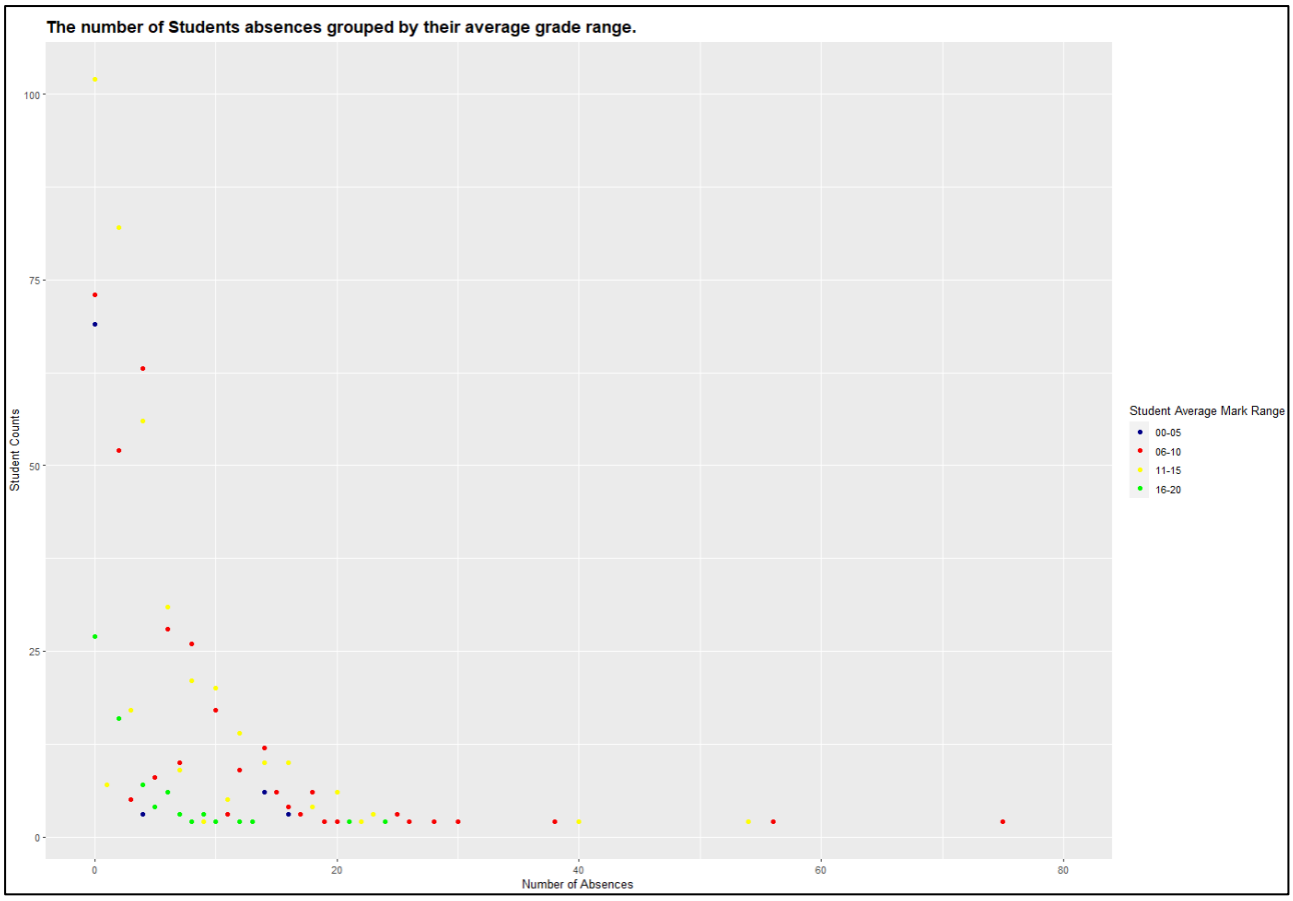


Figure 220 *shows the horizontal bar graph output of the* ***Q5A6V1****.*

The figure 219 above displays the output of the execution of the **Q5A6R1**, which shows the grouped counts of total students for each case of selected attributes that are their number of absences and average grade range. The figure 220 above shows the outcome of the scatterplot plotted after the execution of the **Q5A6V1** variable that displays the students' counts and the average grade range of students grouped based on the number absences.

```
> Q5A6R2
# A tibble: 13 x 3
   absences counts percentage
      <int>  <int>      <dbl>
 1        0     27       34.6
 2        2     16       20.5
 3        4      7       8.97
 4        5      4       5.13
 5        6      6       7.69
 6        7      3       3.85
 7        8      2       2.56
 8        9      3       3.85
 9       10      2       2.56
10       12      2       2.56
11       13      2       2.56
12       21      2       2.56
13       24      2       2.56
```

Figure 221 *shows the output of the* ***Q5A6R1.***\

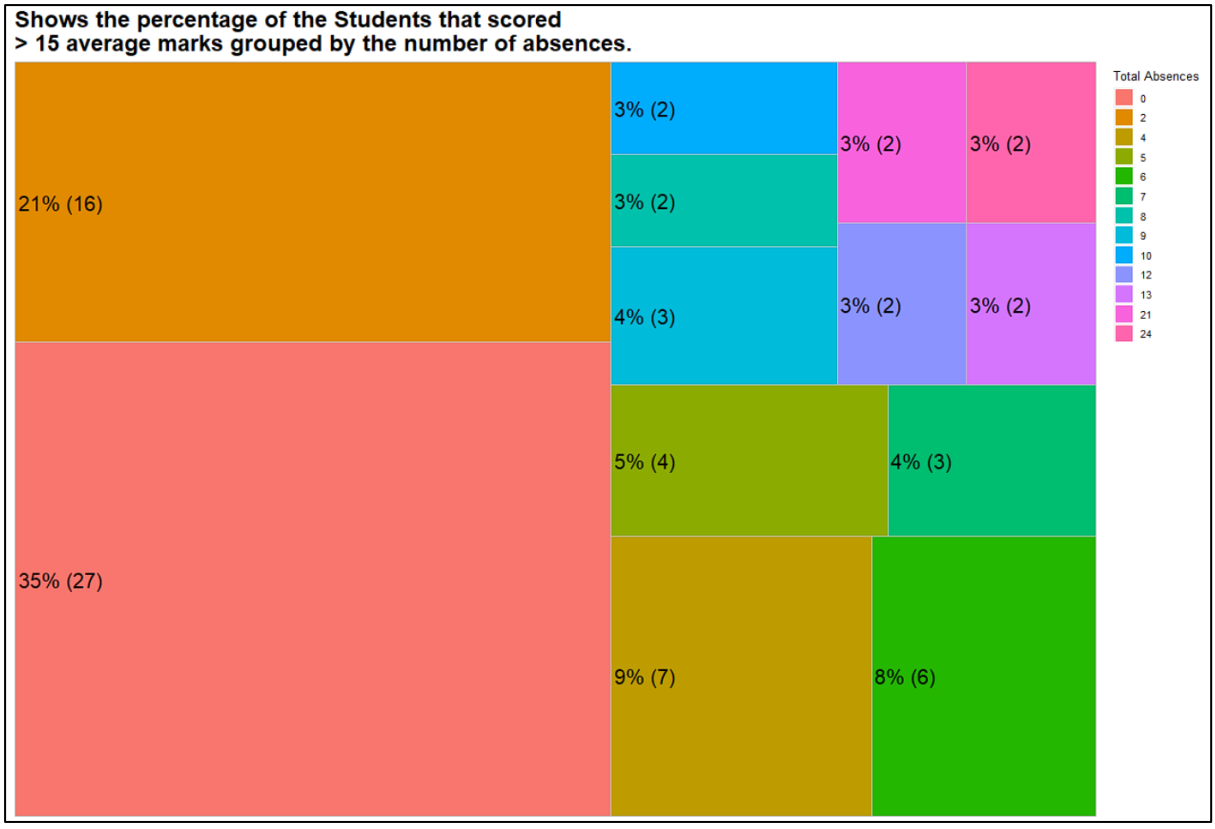


Figure 222 *shows the horizontal bar graph output of the* ***Q5A6V2****.*

The figure 221 above shows the execution output of the **Q5A6R2**. In figure 222, the treemap graph that has been created displays the calculated total student counts and the percentage of them who scored more than 15 average marks grouped based on the number of absences.

Summary for Data Findings

1) The majority number of students had 0 absences.
2) Students who got scored in the range of 06-10 average marks had the most absences.
3) The majority of students that is total of 35% of them, who scored more than 15 average marks had 0 absences.

Explanation for the Data Findings.

Based on the data findings, it was noticed that most numbers of students had 0 total absences. When looking deeper, the students who scored average marks in the range of 06-10 average marks had the big number of total absences. The biggest portion of students who scored more than 15 average marks had 0 absences. Moreover, there are no more than 24 absences from the students who scored more than 15 average marks. From this, it can be stated that low absences can influence positively the academic performance of the students. When students are absent too much excused or not, it can cause them to lose a lot in terms of the knowledge they could get from the classes they missed, eventually, this will decline their academic performance. But this is not the case for those who have never been absent that much, they would be on right track in their learning and can gain every single bit of knowledge, which let them do well in their studies. While excessive absence also possibly grows social anxiety in the students as they would feel shy and scared to talk to their peers. Hence, it is proven that students' absent attributes also play a key role in influencing students' grades whereas students with less number of absent do well in their academics.

4.5.7 Analysis 5-7: Finding the relationship between students' past class failures and their average mark.

The correlation between the students' number of past failure and their average marks will be analysed in this analysis. Two stacked bar graphs have been created for this analysis.

```
#Analysis 7 - Finding the relationship between students' past class failures and their average mark.
Q5A7R1 <- dsap_data %>% group_by(failures, avgGradeRange) %>% summarise(counts = n())
Q5A7R1
Q5A7V1 <- ggplot(Q5A7R1, aes(x=avgGradeRange, y=counts, fill=as.factor(failures))) +
          geom_bar(stat="identity",width = 0.5, color="black") +
          ggtitle("The number of Students with their average score grouped by Past Failures.")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="Student Total Past Failures")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          scale_fill_manual(values=c("#FFC0CB", "#FFE4C4", "#87CEFA", "#F0E68C"),
                            labels = c("1", "2", "3", "4"))+
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A7V1
```

Figure 223 *shows the R code used to create the data visualization figure of* ***Q5A7V1.***

```
Q5A7R2<- dsap_data %>% group_by(failures, avgGrade) %>% filter(avgGrade>15) %>% summarise(counts = n())
Q5A7R2
Q5A7V2<- ggplot(Q5A7R2, aes(x=avgGrade, y=counts, fill=as.factor(failures))) +
          geom_bar(stat="identity",width = 0.5, color="black") +
          ggtitle("The number of Students with their average score > 15 grouped by Past Failures.")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="Student Total Past Failures")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          scale_fill_manual(values=c("#FFC0CB", "#FFE4C4", "#87CEFA", "#F0E68C"),
                            labels = c("1", "2", "3", "4"))+
         geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A7V2
```

Figure 224 *shows the R code used to create the data visualization figure of* ***Q5A7V2.***

As shown in the code figures above, for the both stacked bar graph, the **failures** and **avgGradeRange** are grouped and counted for this analysis. Additionally, for the second stacked bar graph, the **avgGrade** was filtered.

```
> Q5A7R1
# A tibble: 13 x 3
# Groups:   failures [4]
   failures avgGradeRange counts
      <int> <chr>           <int>
 1        0 00-05              39
 2        0 06-10             252
 3        0 11-15             361
 4        0 16-20              76
 5        1 00-05              23
 6        1 06-10              53
 7        1 11-15              41
 8        1 16-20               2
 9        2 00-05              12
10        2 06-10              21
11        2 11-15               5
12        3 00-05              13
13        3 06-10              24
```

Figure 225 *shows the output of the* ***Q5A7R1.***

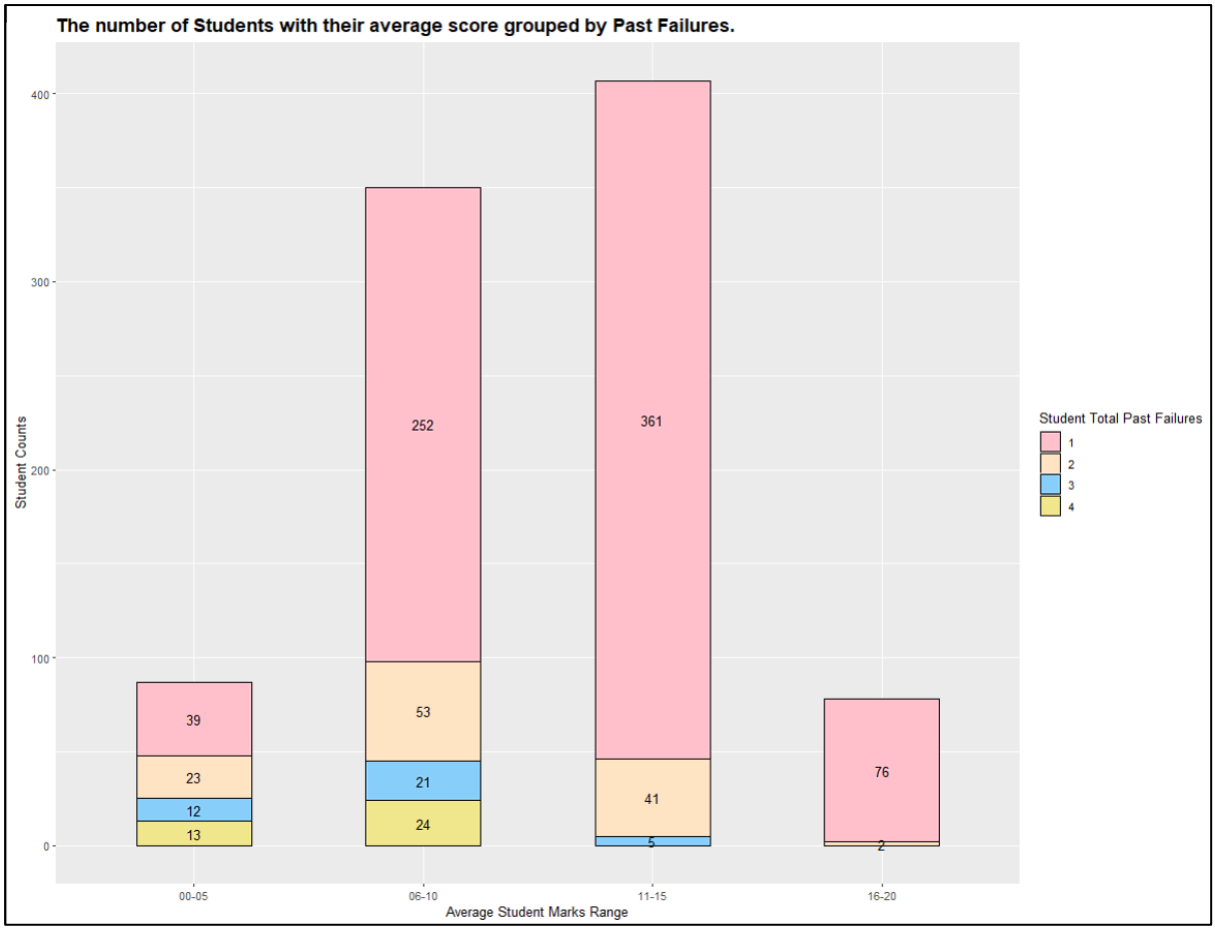


Figure 226 *shows the stacked bar graph output of the* ***Q5A7V1****.*

The figure 225 above displays the output of the execution of the **Q5A7R1**, which shows the grouped counts of total students for each case of selected attributes that are the total number of past failures and their average grade range. The figure 226 above shows the outcome of the stacked bar graph plotted after the execution of the **Q5A7V1** variable that displays the students' counts and the average grade range of students grouped based on their number of past failures.

```
> Q5A7R2
# A tibble: 5 x 3
# Groups:   failures [2]
  failures avgGrade counts
     <int>    <dbl>  <int>
1        0       16     28
2        0       17     20
3        0       18     18
4        0       19     10
5        1       18      2
```

Figure 227 *shows the output of the* ***Q5A7R2.***

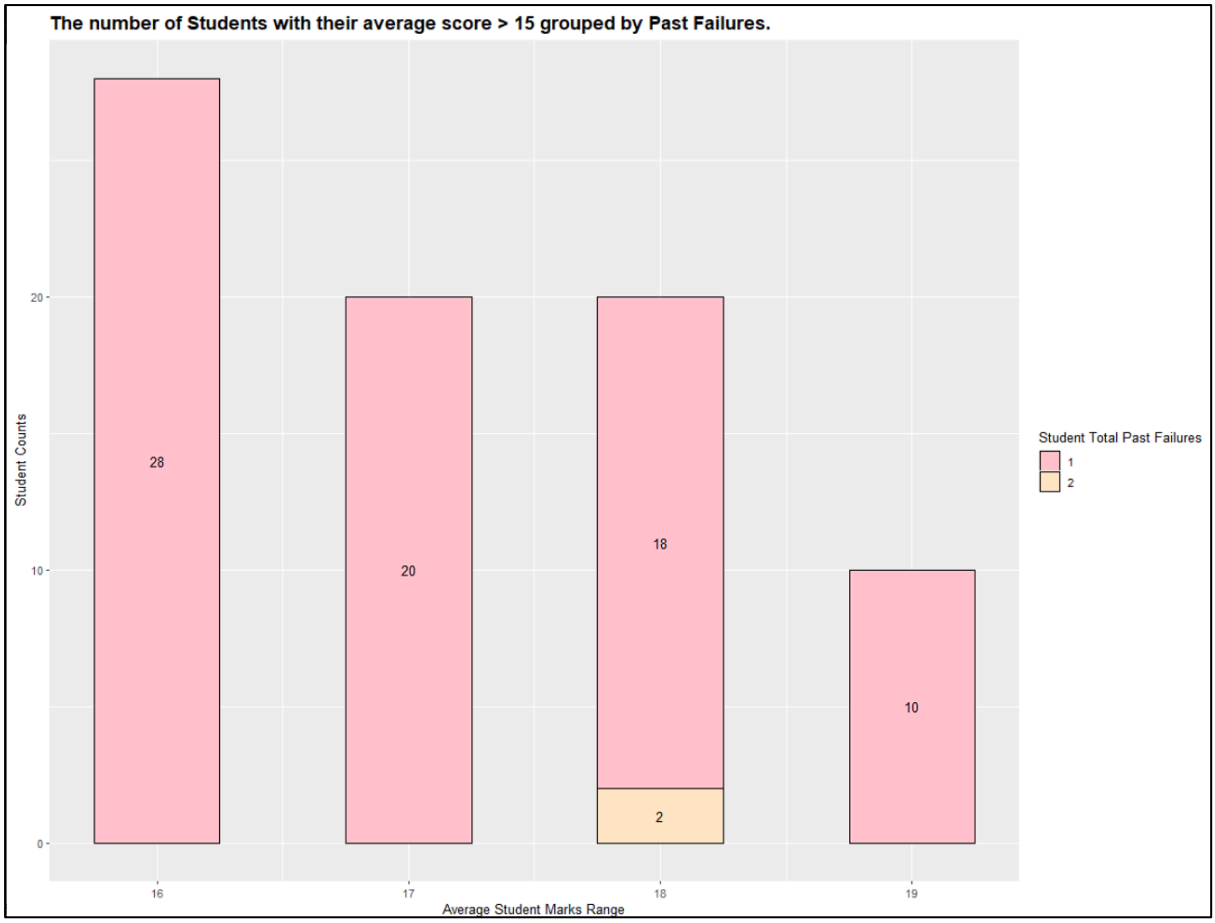


Figure 228 *shows the stacked bar graph output of the* ***Q5A7V2****.*

The figures 227 displays the execution output of the **Q5A7R2**, which displays the calculated total student counts grouped by the student total past failures and average grade. In figure 228, the stacked bar graph illustrates the number of students that scored more than 15 average marks with their number of past failures.

Summary for Data Findings

1) The majority number of students had a total of 1 past failure the most.
2) The very least number of students had past failures of 3.
3) Most of the students who scored more than 15 on average mark had 1 past failure.
4) Very least that is only a total of 2 students from those who scored more than 15 average marks had past failures of 2.
5) Not a single student who scored more than 15 average marks had more than 2 past failures.

Explanation for the Data Findings.

Based on the data findings, it was noticed that the largest number of students had a 1 past failure. While 3 past failures are the least found from students. When looking in-depth, a lot of students who scored more than 15 average marks had only a single failure. Surprisingly, there wasn't a single one of the students who scored more than 15 average marks and had total failures above 2. From this, it can be stated that past failures do influence the students' grades. The main problem with this could be that some students could have lost motivation and are emotionally impacted, such as they could have felt sadness, anger, embarrassment, anxiety, and shame, from these multiple past failures. With those uncomfortable feelings, they could have decided to not put a lot of effort anymore and thought that it is just a waste of effort putting into their studies as they faced too many failures. When the failure is minimal, some students who really want to excel in their studies will put in their full commitment and effort in order to do their best. Continued commitment and effort in anything will help the person to become better, especially, in this case, it would support students to do better in their studies for sure. Hence, it is true that past failures have also quite an influence on the academic performance of the students.

4.5.8 Analysis 5-8: Finding the relationship between students' school and their average mark.

The relationship between the students' type of school and their average marks will be analysed in this analysis. A bar graph and a stacked bar graphs have been created for this analysis.

```
#Analysis 8 - Finding the relationship between students' school and their average mark.
Q5A8R1 <- dsap_data %>% group_by(school, avgGradeRange) %>% summarise(counts = n())
Q5A8R1
Q5A8V1 <- ggplot(Q5A8R1, aes(x=avgGradeRange, y=counts, fill=as.factor(school))) +
          geom_bar(stat="identity",width = 0.5, color="black") +
          ggtitle("The number of Students with their average score grouped by their School.")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="School")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          facet_wrap(~school, labeller = as_labeller(c(`GP`="Gabriel Pereira",
                                                       `MS`="Mousinho da Silveira"))) +
          scale_fill_manual(values=c("#87CEFA", "#F0E68C"),
                            labels = c("Gabriel Pereira", "Mousinho da Silveira"))+
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A8V1
```

Figure 229 *shows the R code used to create the data visualization figure of* ***Q5A8V1.***

```
Q5A8R2<- dsap_data %>% group_by(school, avgGrade) %>% filter(avgGrade>15) %>% summarise(counts = n())
Q5A8R2
Q5A8V2<-ggplot(Q5A8R2, aes(x=avgGrade, y=counts, fill=as.factor(school))) +
        geom_bar(stat="identity",width = 0.5, color="black") +
        ggtitle("The number of Students with their average score > 15 average mark grouped by Past Failures.")+
        labs(x="Average Student Marks Range", y = "Student Counts", fill="School")+
        theme(plot.title = element_text(size = 15, face = "bold")) +
        scale_fill_manual(values=c("#87CEFA", "#F0E68C"),
                          labels = c("Gabriel Pereira", "Mousinho da Silveira"))+
        geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A8V2
```

Figure 230 *shows the R code used to create the data visualization figure of* ***Q5A8V2.***

As shown in the code figures above, for the both bar graph, the **school** and **avgGradeRange** are grouped and counted for this analysis. Additionally, for the second stacked bar graph, the **avgGrade** was filtered.

```
> Q5A8R1
# A tibble: 8 x 3
# Groups:   school [2]
  school avgGradeRange counts
  <chr>  <chr>          <int>
1 GP     00-05             70
2 GP     06-10            287
3 GP     11-15            326
4 GP     16-20             66
5 MS     00-05             17
6 MS     06-10             63
7 MS     11-15             81
8 MS     16-20             12
```

Figure 231 *shows the output of the* ***Q5A8R1.***

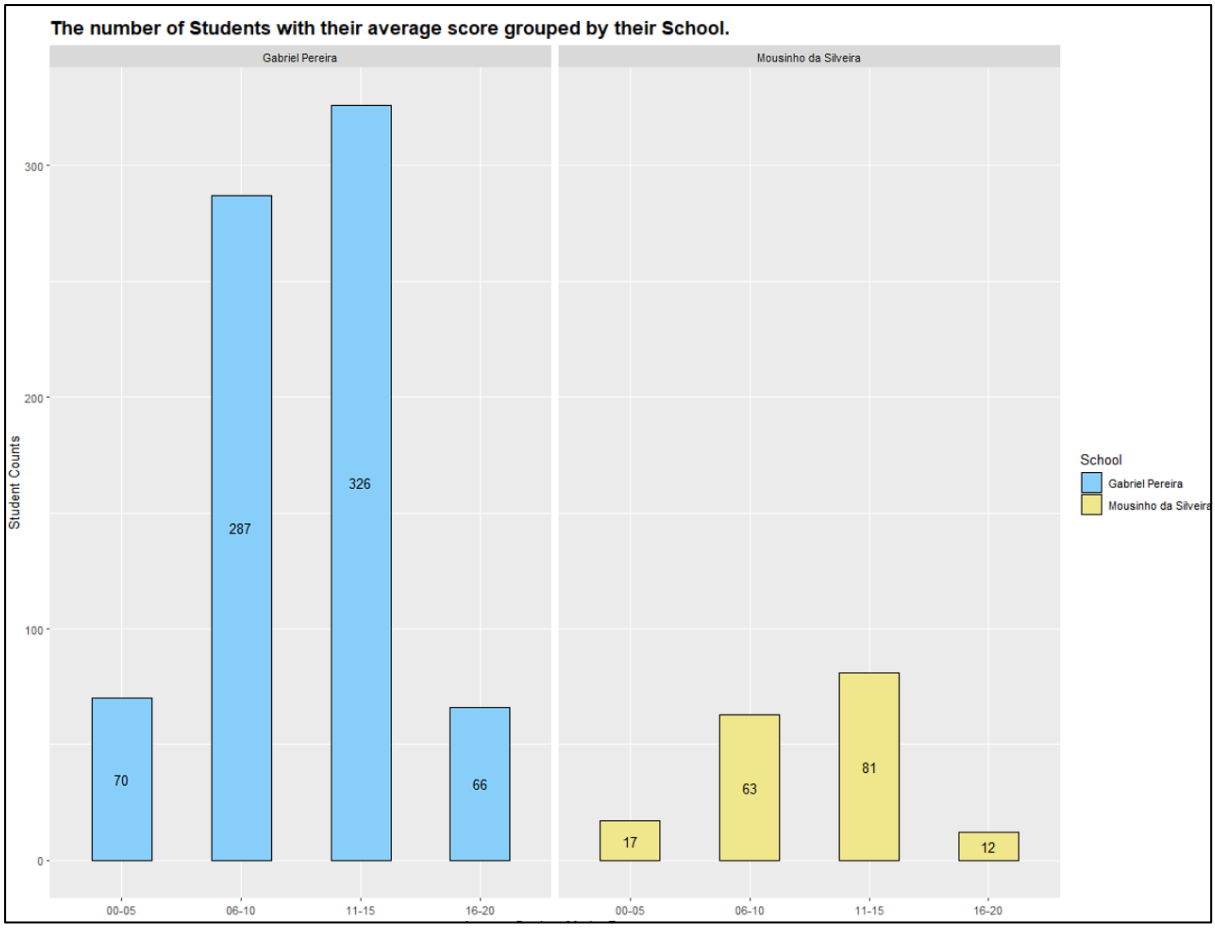


Figure 232 *shows the bar graph output of the* ***Q5A8V1****.*

The figure 231 above displays the result of the execution of the **Q5A8R1**, which shows the grouped counts of total students for each case of selected attributes that are the type of school and their average grade range. The figure 232 above shows the outcome of the bar graph plotted after the execution of the **Q5A8V1** variable that displays the students' counts and the average grade range of students grouped based on their schools.

```
> Q5A8R2
# A tibble: 8 x 3
# Groups:   school [2]
  school avgGrade counts
  <chr>     <dbl>  <int>
1 GP           16     24
2 GP           17     17
3 GP           18     18
4 GP           19      7
5 MS           16      4
6 MS           17      3
7 MS           18      2
8 MS           19      3
```

Figure 233 *shows the output of the* ***Q5A8R2.***

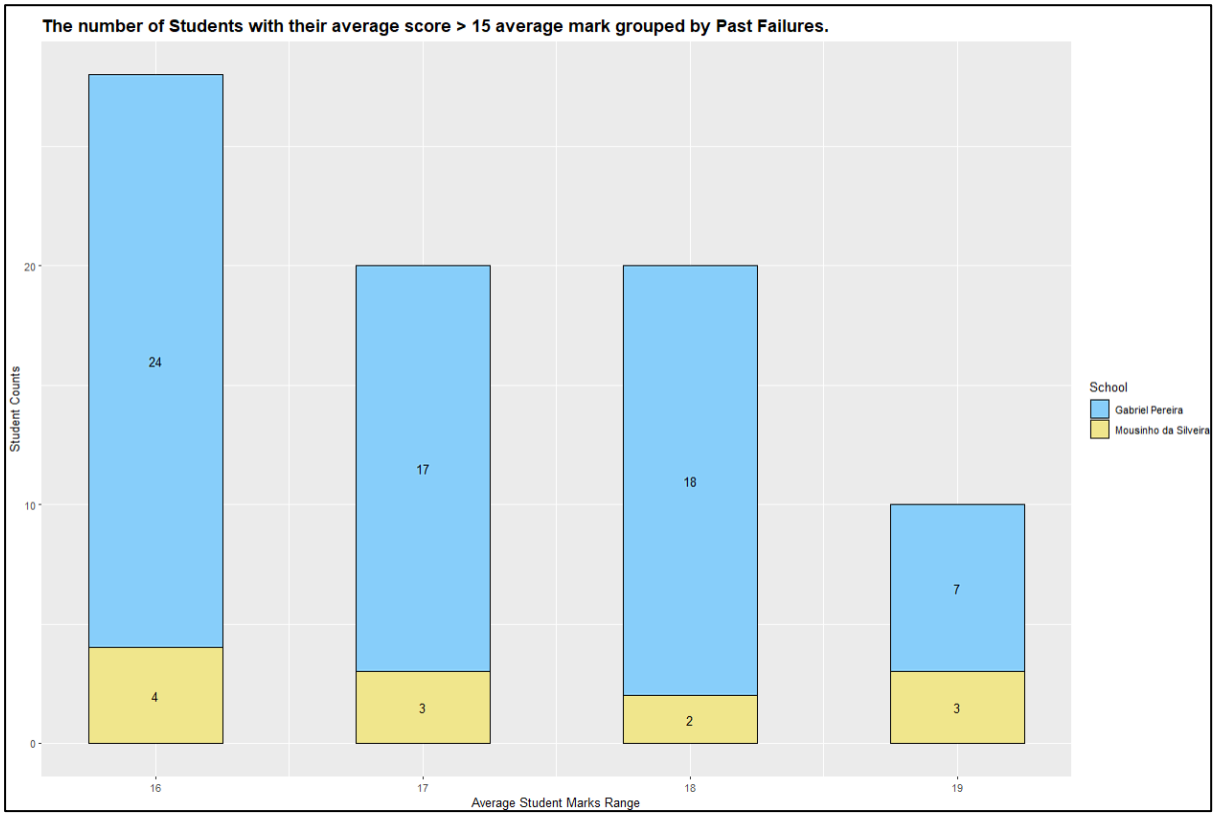


Figure 234 *shows the output of the* ***Q5A8R2.***

The figures 233 displays the execution output of the **Q5A7R2**, which displays the calculated total student counts grouped by the type of school and average grade. In figure 234, the stacked bar graph illustrates the number of students that scored more than 15 average marks with their school type.

Summary for Data Findings

1) The majority number of students are from the Gabriel Pereira school.
2) The very least number of students are from the Moushinho da Silveira school.
3) Students from both schools had a few numbers of them who scored more than 15 average marks.

Explanation for the Data Findings.

Based on the data findings, it was noticed most of the students in this dataset are studying in the Gabriel Pereira school while a very small portion of them studying in Moushinho da Silveira school. As it is quite unfair to compare these schools as the population of students was significantly different. But, when looking more in-depth, it can be found there were few students from both who had scored very excellent scores that are more than 15 marks which means that these from both schools doing great in their academics. From this, it can be said that school is not really affecting the students' academic that much. It doesn't matter which school they are studying in but how committed they are into studying. The students who scored more than 15 average marks would have studied hard in order to excel in their academics. Without hard work, not everything will be in our favour most of the time. Every school does their part accordingly to help these students to do well in terms of their academic performance but in the end, it is just the students who need to put the effort to learn. Hence, it is safe to say different schools don't influence much academic performance.

4.5.9 Analysis 5-9: Finding the relationship between students' sex and their average mark.

The correlation between the students' sex and their average marks will be analysed in this analysis. A bar graph and two stacked bar graphs have been created for this analysis.

```
#Analysis 9 - Finding the relationship between students' sex and their average mark.
Q5A9R1 <- dsap_data %>% group_by(sex, avgGradeRange) %>% summarise(counts = n())
Q5A9R1
Q5A9V1 <- ggplot(Q5A9R1, aes(x=avgGradeRange, y=counts, fill=as.factor(sex))) +
          geom_bar(stat="identity",width = 0.5, color="black") +
          ggtitle("The number of Students with their average score grouped by their Sex.")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="Sex")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          facet_wrap(~sex, labeller = as_labeller(c(`F`="Female",
                                                    `M`="Male"))) +
          scale_fill_manual(values=c("#F0E68C","#00FF7F"),
                            labels = c("Female", "Male"))+
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A9V1
```

Figure 235 *shows the R code used to create the data visualization figure of* ***Q5A9V1.***

```
Q5A9R2<- dsap_data %>% group_by(sex, avgGrade) %>% filter(avgGrade>15) %>% summarise(counts = n())
Q5A9R2
Q5A9V2<-ggplot(Q5A9R2, aes(x=avgGrade, y=counts, fill=as.factor(sex))) +
        geom_bar(stat="identity",width = 0.5, color="black") +
        ggtitle("The number of Students with their average score > 15 average mark grouped by their Sex.")+
        labs(x="Average Student Marks Range", y = "Student Counts", fill="Sex")+
        theme(plot.title = element_text(size = 15, face = "bold")) +
        scale_fill_manual(values=c("#F0E68C","#00FF7F"),
                          labels = c("Female", "Male"))+
        geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A9V2
```

Figure 236 *shows the R code used to create the data visualization figure of* ***Q5A9V2.***

```
Q5A9R3<- dsap_data %>% group_by(sex, avgGrade) %>% filter(avgGrade<=5) %>% summarise(counts = n())
Q5A9R3
Q5A9V3<-ggplot(Q5A9R3, aes(x=avgGrade, y=counts, fill=as.factor(sex))) +
        geom_bar(stat="identity",width = 0.5, color="black") +
        ggtitle("The number of Students with their average score <= 5 average mark grouped by their Sex.")+
        labs(x="Average Student Marks Range", y = "Student Counts", fill="Sex")+
        theme(plot.title = element_text(size = 15, face = "bold")) +
        scale_fill_manual(values=c("#F0E68C","#00FF7F"),
                          labels = c("Female", "Male"))+
        geom_text(aes(label=counts), position = position_stack(vjust = 0.5))
Q5A9V3
```

Figure 237 *shows the R code used to create the data visualization figure of* ***Q5A9V3.***

```
ggarrange(Q5A9V2, Q5A9V3, nrow=2, ncol=1)
```

Figure 238 *shows the R code used to arrange the data visualization figures of* ***Q5A9V2*** *and* ***Q5A9V3*** *in one view.*

As shown in the code figures above, for all the graphs, the **sex** and **avgGradeRange** are grouped and counted for this analysis. For both stacked bar graphs, it was filtered with the **avgGrade.**

```
`summarise()` has grouped outp
> Q5A9R1
# A tibble: 8 x 3
# Groups:   sex [2]
  sex   avgGradeRange counts
  <chr> <chr>           <int>
1 F     00-05              50
2 F     06-10             206
3 F     11-15             196
4 F     16-20              33
5 M     00-05              37
6 M     06-10             144
7 M     11-15             211
8 M     16-20              45
```

Figure 239 *shows the output of the* ***Q5A9R1.***

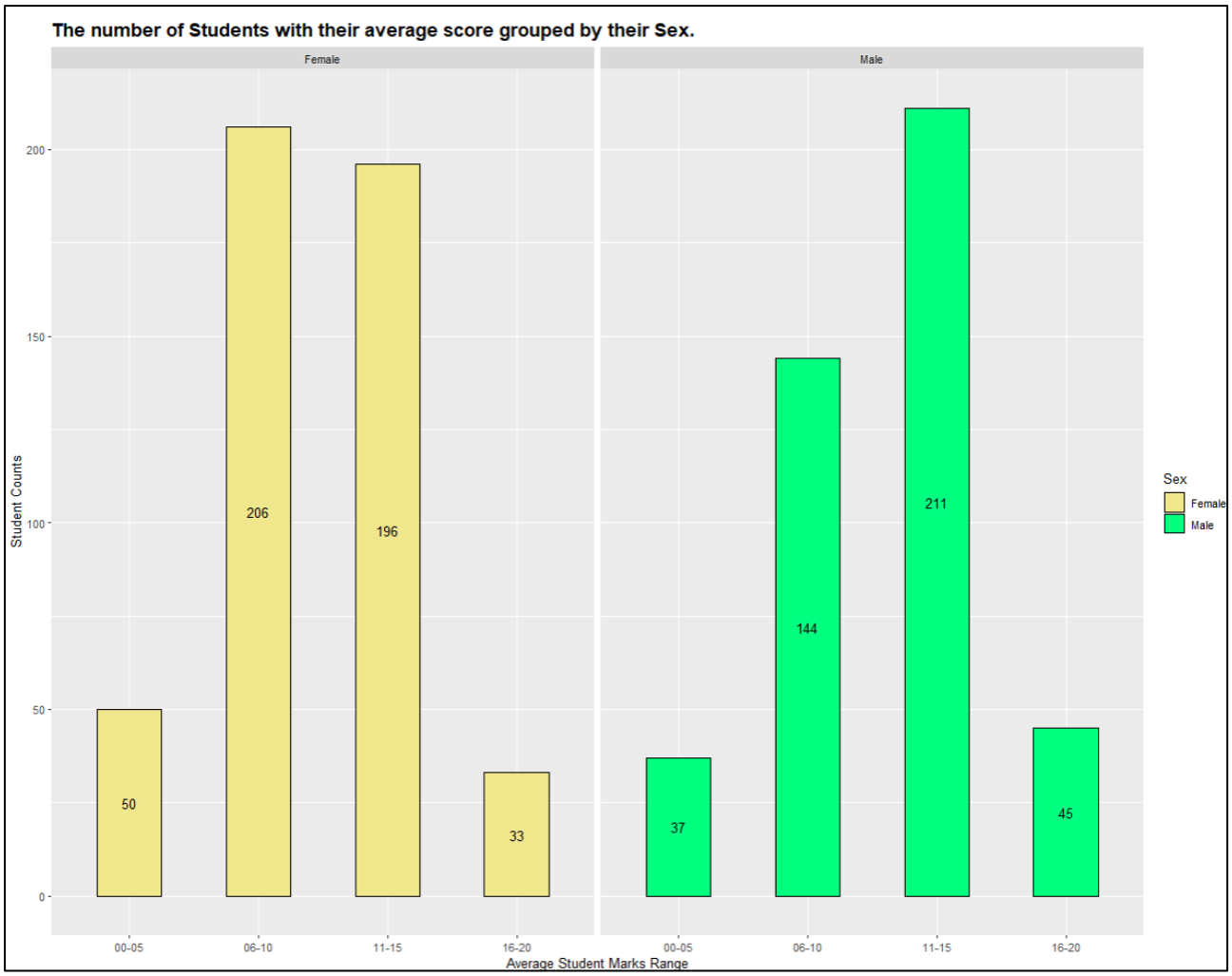


Figure 240 *shows the bar graph output of the* ***Q5A9V1****.*

The figure 239 above displays the result of the execution of the **Q5A9R1**, which shows the grouped counts of total students for each case of selected attributes that are the gender and their average grade range. The figure 240 above shows the outcome of the bar graph plotted after the execution of the **Q5A9V1** variable that displays the students' counts and the average grade range of students grouped based on their gender.

```
> Q5A9R2
# A tibble: 8 x 3
# Groups:   sex [2]
  sex   avgGrade counts
  <chr>    <dbl>  <int>
1 F           16     13
2 F           17      9
3 F           18      8
4 F           19      3
5 M           16     15
6 M           17     11
7 M           18     12
8 M           19      7
>
```

Figure 241 *shows the output of the* ***Q5A9R2.***

```
> Q5A9R3
# A tibble: 9 x 3
# Groups:   sex [2]
  sex   avgGrade counts
  <chr>    <dbl>  <int>
1 F           1       2
2 F           2       7
3 F           3       2
4 F           4      19
5 F           5      20
6 M           2       8
7 M           3       4
8 M           4      15
9 M           5      10
```

Figure 242 *shows the output of the* ***Q4A9R3.***

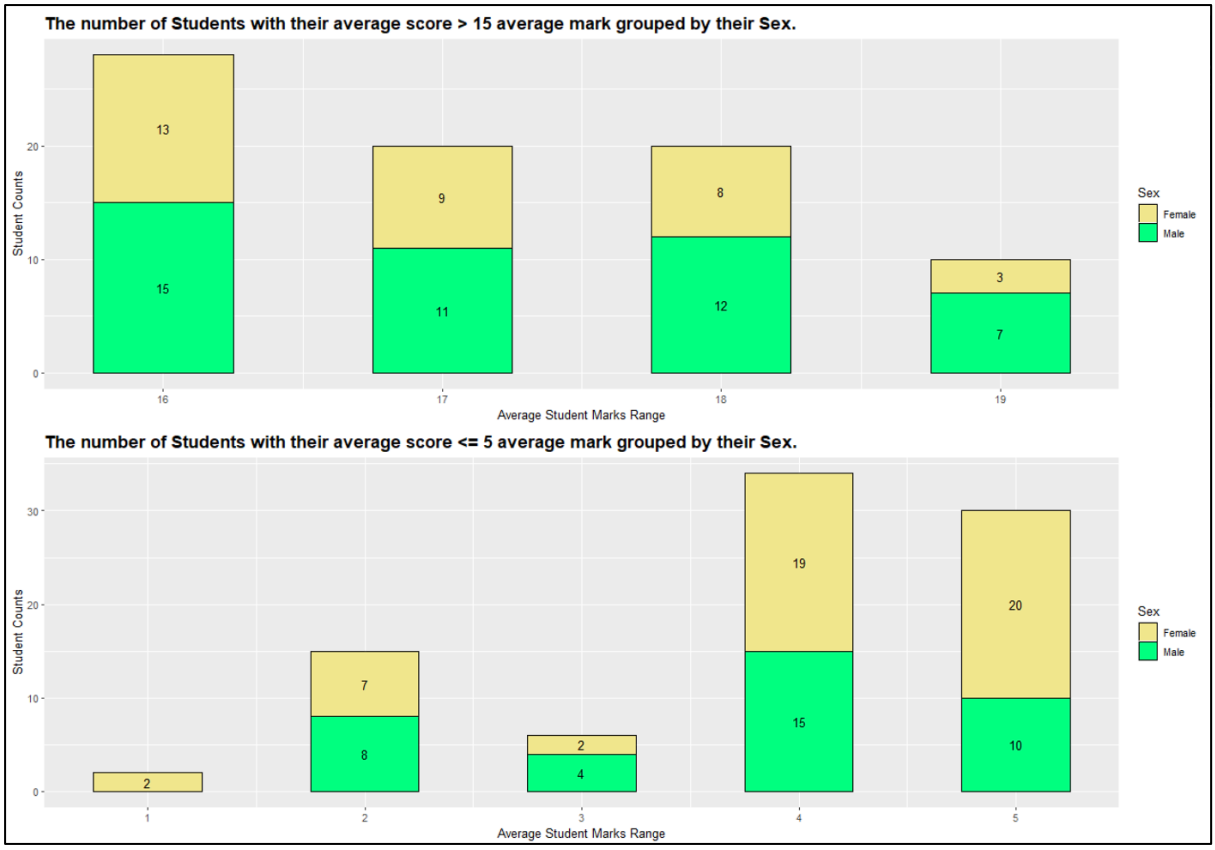


Figure 243 *shows the stacked bar graphs of the output of* ***Q5A9V2*** *and* ***Q5A9V3****.*

The figures 241 and 242 displays the execution output of the **Q5A9R2** and **Q5A9R3**, which displays the calculated total student counts grouped by the gender and average grade. In figure 243, the two stacked bar graph illustrates the number of students that scored more than 15 average marks and less than equal to 5 average marks with their gender.

Summary for Data Findings

1) There are a total of 485 female students and 437 male students.
2) A very large number of male students have scored more than 15 on average score compared to the female students.
3) A very large number of female students have scored less than equal to 5 average scores compared to the male students.

Explanation for the Data Findings.

Based on the data findings, it can be viewed that there was a slight difference between students' gender, where it was just 48 students. When examining deeper, there is a significant difference between students' gender who scored more than 15 average marks. The male students have dominated compared to the female students in getting excellent grades. While looking into those who scored less than equal to 5 average mark, there it can be seen that there was quite a big number of female students who scored compared to male students. From this, it can be claimed that students' sex also does affect students' academic performance solely based on this dataset. It seems that male students have outperformed female students in their overall academic performance. This could be because of various reasons maybe they had better motivation, which helps them to do better. It is hard to pinpoint only a single reason for this result. Hence, it is proven that students' gender also could do a part in influencing academic performance.

4.5.10 Analysis 5-10: Finding the relationship between students' address type and their average mark.

The relationship between the students' address and their average marks will be analysed in this analysis. A bar graph and a stacked bar graphs have been created for this analysis.

```
#Analysis 10 - Finding the relationship between students' address type and their average mark.
Q5A10R1 <- dsap_data %>% group_by(address, avgGradeRange) %>% summarise(counts = n())
Q5A10R1
Q5A10V1 <-  ggplot(Q5A10R1, aes(x=avgGradeRange, y=counts, fill=as.factor(address))) +
            geom_bar(stat="identity",width = 0.5, color="black") +
            ggtitle("The number of Students with their average score grouped by their Address.")+
            labs(x="Average Student Marks Range", y = "Student Counts", fill="Address")+
            theme(plot.title = element_text(size = 15, face = "bold")) +
            facet_wrap(~address, labeller = as_labeller(c(`U`="Urban",
                                                          `R`="Rural"))) +
            scale_fill_manual(values=c("#191970","#FF1493"),
                              labels = c("Rural", "Urban"))+
            geom_text(aes(label=counts), position = position_stack(vjust = 0.5), color="white")
Q5A10V1
```

Figure 244 *shows the R code used to create the data visualization figure of* ***Q5A10V1.***

```
Q5A10R2<- dsap_data %>% group_by(address, avgGrade) %>% filter(avgGrade>15) %>% summarise(counts = n())
Q5A10R2
Q5A10V2<- ggplot(Q5A10R2, aes(x=avgGrade, y=counts, fill=as.factor(address))) +
          geom_bar(stat="identity",width = 0.5, color="white") +
          ggtitle("The number of Students with their average score > 15 average mark grouped by their Address.")+
          labs(x="Average Student Marks Range", y = "Student Counts", fill="Address")+
          theme(plot.title = element_text(size = 15, face = "bold")) +
          scale_fill_manual(values=c("#191970","#FF1493"),
                            labels = c("Rural", "Urban"))+
          geom_text(aes(label=counts), position = position_stack(vjust = 0.5), color="white")
Q5A10V2
```

Figure 245 *shows the R code used to create the data visualization figure of* ***Q5A10V2.***

As shown in the code figures above, for the both bar graph, the **address** and **avgGradeRange** are grouped and counted for this analysis. Additionally, for the second stacked bar graph, the **avgGrade** was filtered.

```
`summarise()` has grouped output
> Q5A10R1
# A tibble: 8 x 3
# Groups:   address [2]
  address avgGradeRange counts
  <chr>   <chr>          <int>
1 R       00-05             29
2 R       06-10             90
3 R       11-15             79
4 R       16-20             19
5 U       00-05             58
6 U       06-10            260
7 U       11-15            328
8 U       16-20             59
```

Figure 246 *shows the output of the* ***Q5A10R1.***

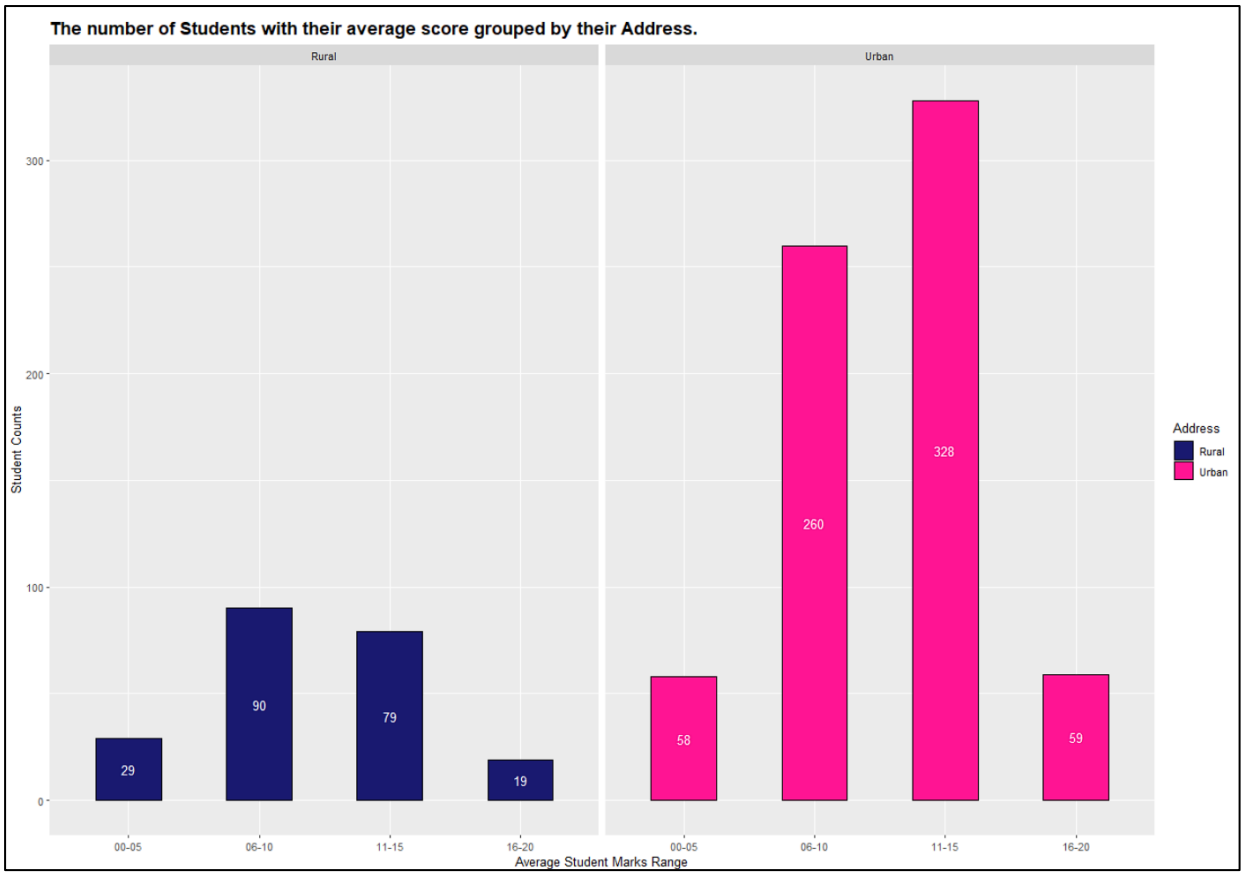


Figure 247 *shows the bar graph output of the* ***Q5A10V1****.*

The figure 246 above displays the result of the execution of the **Q5A10R1**, which shows the grouped counts of total students for each case of selected attributes that are the address of the student and their average grade range. The figure 247 above shows the outcome of the bar graph plotted after the execution of the **Q5A10V1** variable that displays the students' counts and the average grade range of students grouped based on their address.

```
> Q5A10R2
# A tibble: 8 x 3
# Groups:   address [2]
  address avgGrade counts
  <chr>      <dbl>  <int>
1 R             16     10
2 R             17      2
3 R             18      4
4 R             19      3
5 U             16     18
6 U             17     18
7 U             18     16
8 U             19      7
```

Figure 248 *shows the output of the* ***Q5A10R2.***

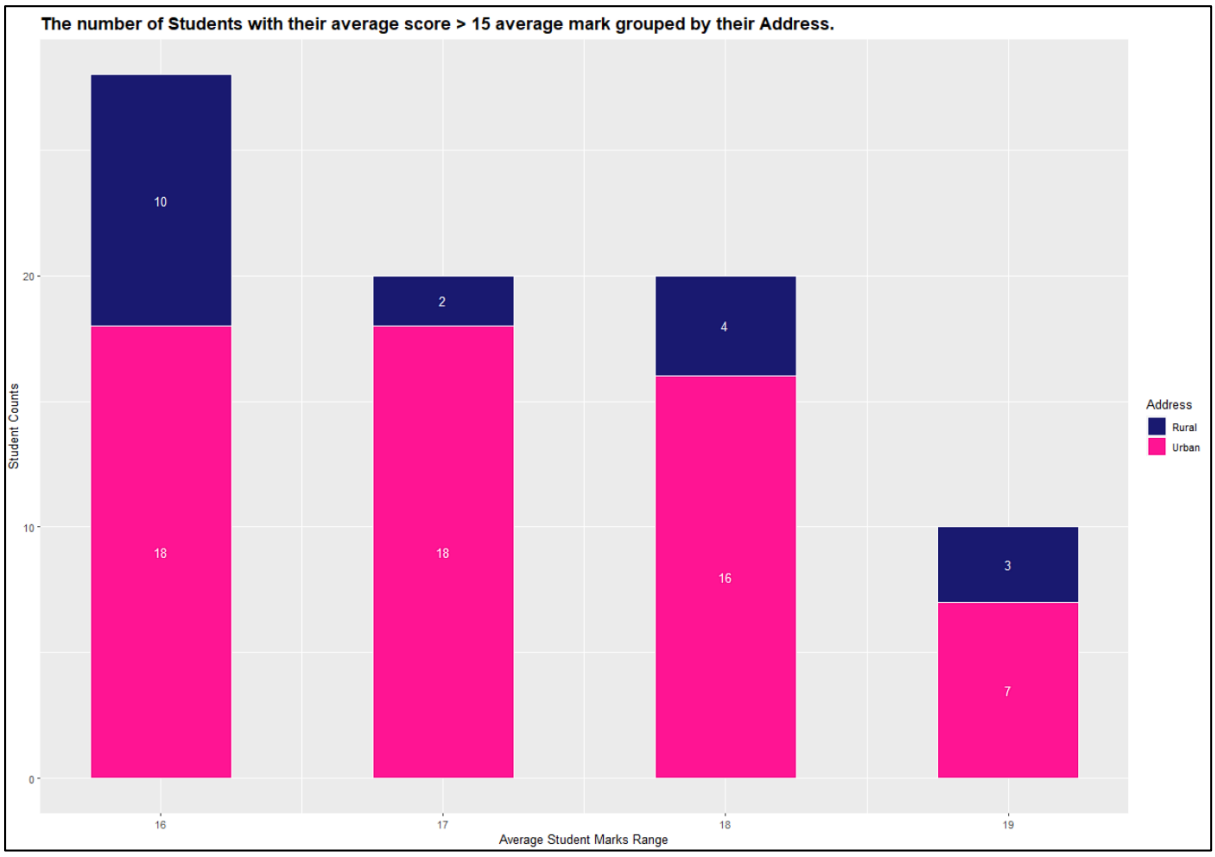


Figure 249 *shows the output of the* ***Q5A10R2.***

The figures 248 displays the execution output of the **Q5A10R2**, which displays the calculated total student counts grouped by the address and average grade. In figure 249, the stacked bar graph illustrates the number of students that scored more than 15 average marks with their address.

Summary for Data Findings

1) There are a total of 705 students living in the urban area and 217 students living in the rural area.
2) A very large number of students who live in urban areas have scored more than 15 on average score compared to the students who live in the rural area.

Explanation for the Data Findings.

Based on the data findings, it can be seen that the majority of the students are living in urban areas. While looking in-depth into those who scored more than 15 average marks, it can be clearly seen that a lot of them are living in the urban area, but that doesn't change the fact some students live in rural areas who got average marks more than 15. From this, it can be stated, that the address attribute of the students also can fairly influence the students' marks. As one of the reasons why the students who live in rural areas will not frequently get good results is the lack of access to quality resource material. Normally, living in rural areas is much harder than living in urban areas as they wouldn't get much availability of resources. But, this is not the case for those who are living in urban areas where they easily can find quality resource materials to study for the examinations, and overall they do better in their academics. In urban areas, anything is easily accessible by the students so it assists them a lot in terms of for example going for tuition classes. Hence, it is proven that students' address also plays a little bit of influence in impacting their academic performance

Conclusion

As a conclusion for this fifth question, based on the observed analyses, it can be reasoned that the students' personal life attributes, including all the important decisions they make, can highly influence their academic performance. Even though these students have the freedom to make any decision on their own in their life, but it is highly advised that they should obtain proper guidance and suggestions from their respective people like their families and teachers before making any important decisions in their lives. It is because some attributes like getting a higher education and excessive alcohol consumption would not just affect their grades but will also affect their entire life later on. Overall, the students have complete control over their personal lives, and in the end, they are the ones that going to make the final decision on anything related to them, so before proceeding with the decisions they made, they should know how that certain decision would affect them personally in their lives as well as their academics.

5.0 Extra Features

There were a few additional features were added in order to conduct better data analysis.

5.1 Extra Feature 1: Use of Tidyverse Package

The first additional feature is utilizing the tidyverse package.

```
#====Installing various packages====#
install.packages("tidyverse")
```

Figure 250 *shows the code for treeemap function.*

```
#====Load the installed packages====#
library(tidyverse) #ggplot2, dplyr, readr, tidyr, tibble
```

Figure 251 *shows the code for treeemap function.*

In the figures above, it shows on how to install and import the tidyverse package into our RStudio. The Tidyverse is a package of comprises other useful packages that are required for any data analysis. Instead of traditionally installing and importing certain essential packages like ggplot2, dplyr and more, installing and loading just the tidyverse makes the work faster and easier with just single command. Moreover, the 8 packages that are bundled up in the tidyverse package are the packages which are used for almost all of the everyday analysis.

5.2 Extra Feature 2: Treemap

The next additional feature is implementing the Treemap graphs. The treemap graph was created using the "ggplot" and "treemapify" packages.

```
Q3A4R2<- dsap_data %>% group_by(traveltime) %>% filter(avgGrade>avgMeanGrade) %>%
        summarise(counts = n(),  percentage = n()/length(which(dsap_data$avgGrade>avgMeanGrade))*100)
Q3A4R2
Q3A4V2 <-  ggplot(Q3A4R2, aes(x=percentage, y="", fill = as.factor(traveltime), area = percentage)) + geom_treemap()+
           theme(legend.justification="top",
                 panel.background = element_blank(),
                 axis.title = element_blank(),
                 axis.text = element_blank(),
                 axis.line = element_blank(),
                 axis.ticks= element_blank(),
           plot.title = element_text(size = 20, face = "bold")) +
           ggtitle("Shows the percentage of the Students that scored\n> average mark grouped by their travel time.") +
           labs(fill="Student Travel Time (Hours)")+ scale_fill_manual(values = c("#008B8B", "#191970","#4B0082","#DA70D6"),
                                                                      labels = c("1", "2", "3", "4")) +
           geom_treemap_text(aes(label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                             color = c("white"), place = "left")
Q3A4V2
```

Figure 252 *shows the code for treeemap function.*

As the figure above shown, in order to produce a treemap using the "ggplot()", the "geom_treemap()" function should be included. This function is vital because it is one that will produce the structure of the graph into a treemap. Also, the "geom_treemap_text()" function was added in order to display the exact output of the percentage and count of the students on each of the treemap portions.

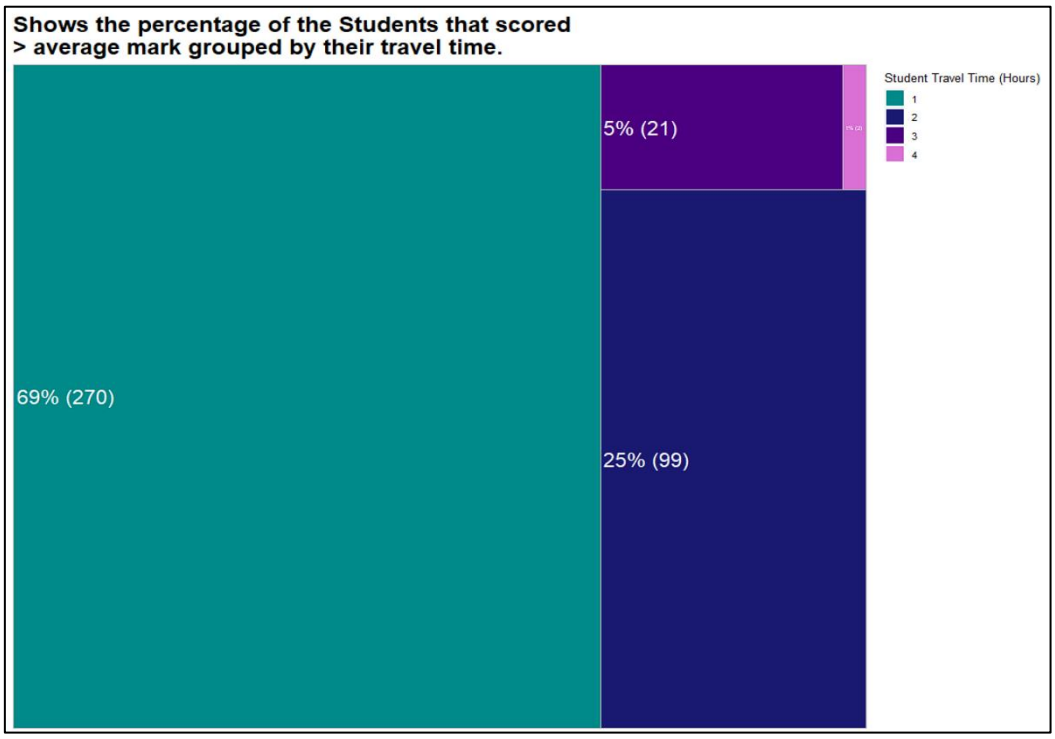


Figure 253 *shows the example output for treeemap function.*

As shown in the figure above, it is one example of treemap graph output. Treemap was been utilized for certain analyses because it helps to visualize the relative size of the data categories, which enables us to observe the data output quickly. Also, different colours were used for different data categories which makes them easily recognizable and analysable.

5.3 Extra Feature 3: Arranging Graphs Together

The third additional feature was the implementation of the "ggarrange()" function using the "ggpubr" package.

```
#Show combined two pies into one view
ggarrange(Q1A1V2, Q1A1V3, nrow = 2, ncol = 1)
```

Figure 254 *shows the code for arrange function.*

The figure above displays on to code of the "ggarrange()" function. The graph variables that wanted to combine together should be parsed in first. Then, the "nrow" and "ncol" will be altered according to how they want the graph to be arranged. In this case, the "nrow" was altered to 2 and "ncol" to 1 in order to display both pie charts in a single column view.

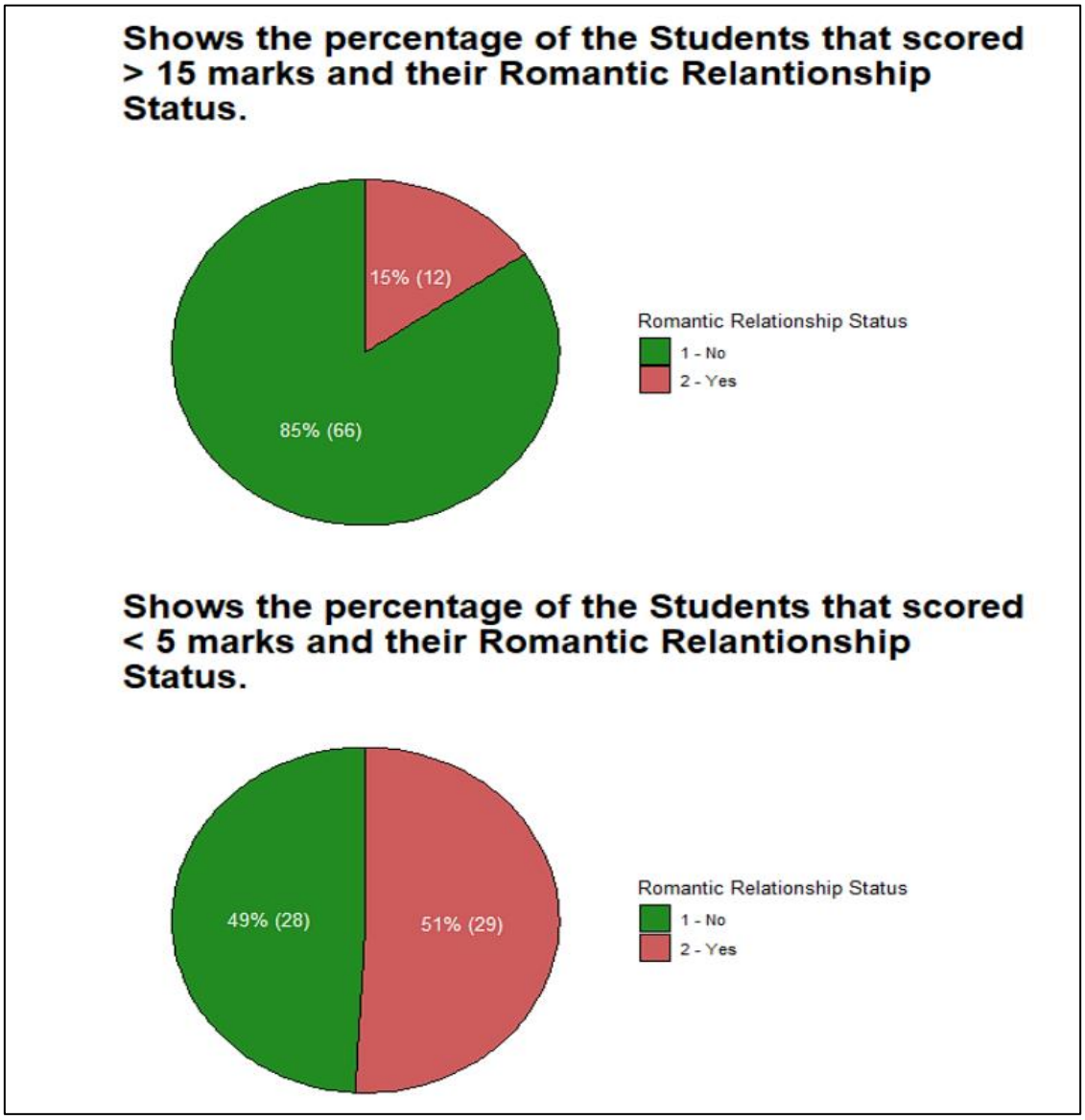


Figure 255 *shows the example output for arrange function.*

The figure above shows the arranged pie charts in a single column view. This function was helpful in grouping together two graphs into a single view, which makes the observation easier to view the differences between the related data for the same analysis in order to make a better analysis.

5.4 Extra Feature 4: Displaying Percentage and Counts on Pie Charts

The next additional feature was the implementation of the percentage and student count on each portion of the pie chart.

```
Q1A2R2 <- dsap_data %>% group_by(goout) %>% filter(avgGrade>15) %>%
          summarise(counts = n(), percentage= n()/length(which(dsap_data$avgGrade>15))*100)

Q1A2R2
Q1A2V2 <- ggplot(Q1A2R2, aes(x="", y =percentage, fill=as.factor(goout))) + geom_col(color = "black") + coord_polar("y", start = 0) +
          theme(panel.background = element_blank(),
                axis.title = element_blank(),
                axis.text = element_blank(),
                axis.line = element_blank(),
                axis.ticks= element_blank(),
          plot.title = element_text(size = 20, face = "bold")) +
          geom_text(aes(x=1.2, label = paste0(round(percentage),"%",sep=" ","(",counts,")")),
                    color = c("white"), position = position_stack(vjust=0.5)) +
          ggtitle("Shows the percentage of the Students that scored\n> 15 marks and their frequency of Going Out\nwith Friends.") +
          labs(fill="Students Going Out With Friends")+
          scale_fill_manual(values = c("#4B0082","#DA70D6","#8B4513", "#191970","#2F4F4F"),
                            labels = c("1 - Very Low", "2 - Low", "3 - Medium", "4 - High", "5 - Very High"))
Q1A2V2
```

Figure 256 *shows the code for calculate and display percentage and total student counts.*

As the code above shows, first, the percentage was calculated and stored in the percentage variable. Then, with the use of the "geom_text() " function, the calculated percentage and student counts were displayed on the pie charts.

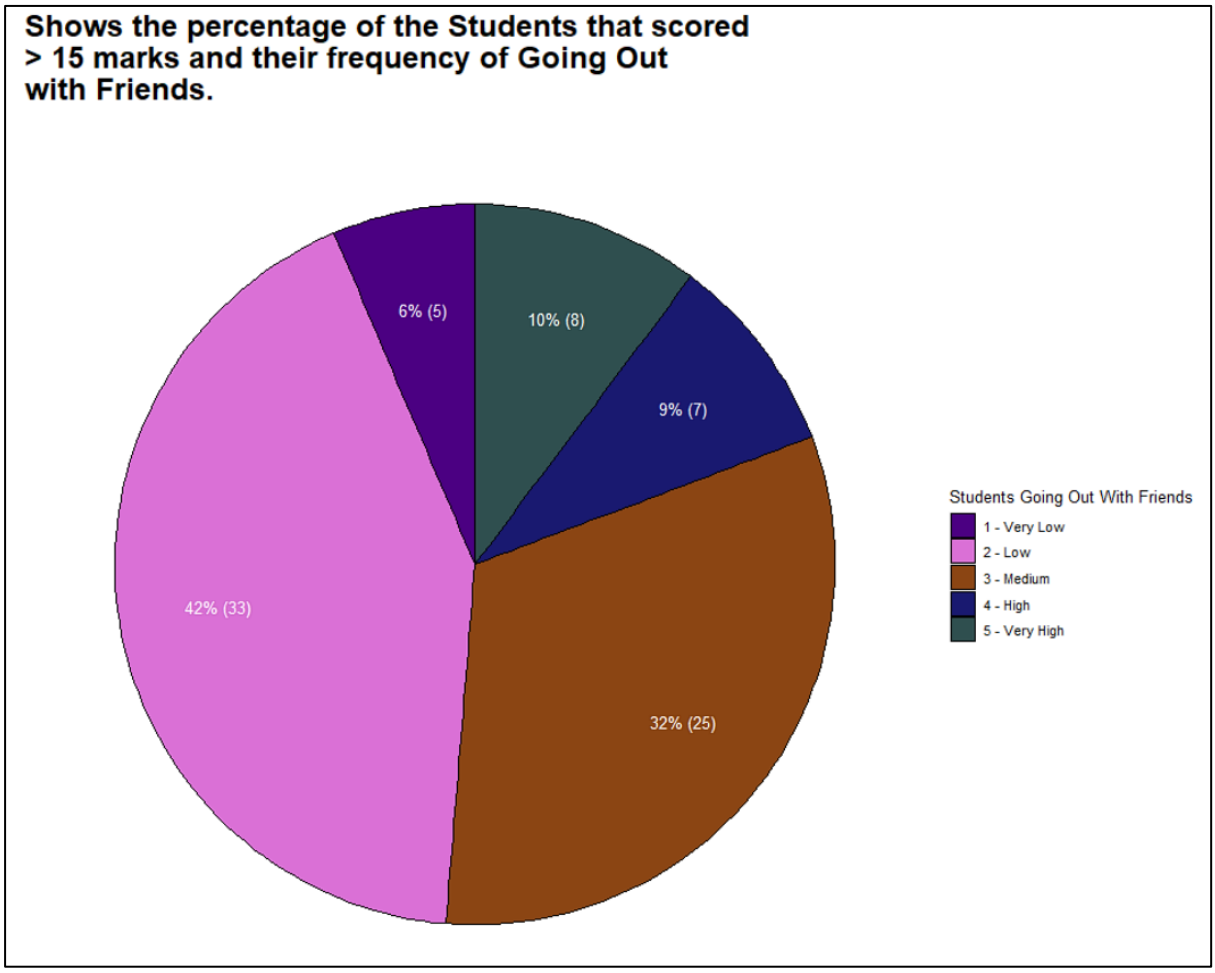


Figure 257 *shows the output of the percentage and student counts of the pie chart.*

The figure above shows the displayed percentages and counts for each portion of the pie charts. This additional feature is useful in making observations efficiently on the different category values of the data. It will be easy for us to make a better analysis with the percentage and total counts shown on the pie chart.

5.5 Extra Feature 5: Assigning Different Colours for Bar Graphs

The next extra feature that is added to our data analysis is assigning different colours based on their categorical data values.

```
Q1A3R1<- dsap_data %>% group_by(famrel,avgGradeRange) %>% summarise(counts = n())
Q1A3R1
Q1A3V1<- ggplot(Q1A3R1, aes(x= avgGradeRange, y=counts, fill = as.factor(famrel))) +
         geom_bar(stat = "identity", position = position_dodge2(preserve = 'single'), width=0.9)+
         ggtitle("The number of Students with their average score grouped by their Family Relantionship quality.")+
         theme(plot.title = element_text(size = 15, face = "bold")) +
         labs(fill = "Students Family Relationship Level", x="Average Students Marks Range", y = "Student Counts")+
         geom_text(aes(label=counts), position = position_dodge2(1), vjust=-0.5) +
         scale_fill_manual(values = c("#40E0D0","#191970","#FF1493", "#DEB887","#6A5ACD"),
                           labels = c("1 - Very Bad", "2 - Bad", "3 - Okay", "4 - Good", "5 - Excellent"))
Q1A3V1
```

Figure 258 *shows the code for displaying different colours.*

The figure above displays on how to assign different colours for each of the categorical values. For this, the selected colours in the "c()" function was assigned to the "values" variable and parsed into the "scale_fill_manual()" function, which displays the data outputs based on the assigned colours.

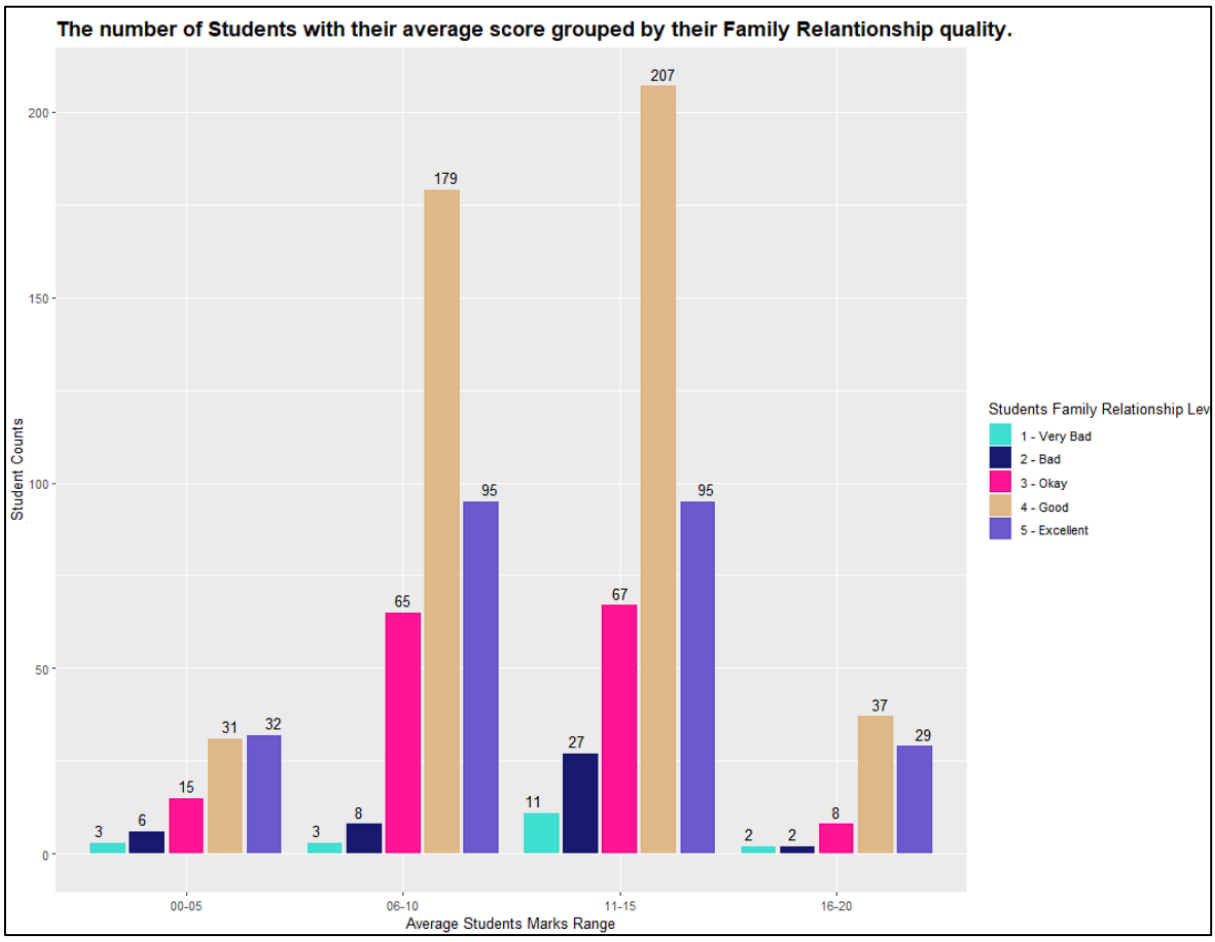


Figure 259 *shows the output of displaying different colours.*

The figure above shows the output of different colours based on their categorical value. This is also another useful feature that helps to make observations over different categorical data easily. As it shows colours based on their category, it is also easier to picture and produce better analysis.

5.6 Extra Feature 6: Adding new columns

Moreover, adding new columns to the memory also can be considered an extra feature. There were three new columns were added that are **avgGradeNotRoundedOff**, **avgGrade** and **avgGradeRange**.

The **avgGradeNotRoundedOff** consists of each student's calculated average grade combining the **G1**, **G2**, and **G3** grades without rounding off. The **avgGrade** column holds the output after rounding off the values from the **avgGradeNotRoundedOff** column. The **avgGradeRange** consists of the assigned grade range based on their **avgGrade** value.

```
#========Create a new column for the mean of the math grade for three period========#
dsap_data$avgGradeNotRoundedOff = rowMeans(subset(dsap_data, select = c(G1,G2,G3)))
dsap_data$avgGrade = round(dsap_data$avgGradeNotRoundedOff)


#========Assigning the average grade into a new range column========#
dsap_data = dsap_data %>% mutate(avgGradeRange = case_when(avgGrade <=5 ~ "00-05",
                                                           avgGrade <=10 ~ "06-10",
                                                           avgGrade <=15 ~ "11-15",
                                                           avgGrade <=20 ~ "16-20",))
```

Figure 260 *shows the code for creating new columns into the dataset.*

The code figure displays show on how to create, calculate and assign the values to the new column. Two ways were used to create these columns. The first way is using the "$" operator and the second way is using the "mutate()" function.

avgGradeNotRoundedOff	avgGrade	avgGradeRange
5.666667	6	06-10
5.333333	5	00-05
8.333333	8	06-10
14.666667	15	11-15
8.666667	9	06-10
15.000000	15	11-15
11.666667	12	11-15
5.666667	6	06-10
17.666667	18	16-20
14.666667	15	11-15

Figure 261 *shows the output the created columns.*

The figure above shows the output of newly created columns in the dataset, which are temporarily stored in the memory. This feature is beneficial as it allows us to efficiently compare the average marks especially the average grade range helps to make categorised comparisons very easily and effectively.

6.0 Conclusion

In conclusion, after conducting a total of 5 questions and 30 analyses, it was discovered that various attributes related to the students within the dataset play a very vital role in influencing their academic performance. To name a few of them that influence academic performance hugely are students' relationship status, study time, family educational support, workday and weekend alcohol consumption level, travel time to school, decision to pursue higher studies, health status and absences. Honestly, every other attribute of the students also actually does at least a small influence on their academics. The respective parties have to make sure to check on these attributes, which affect the students and help them in order to solve their issues on that. Hope that the results got from this data analysis would help to make a change to improve the student's overall academic performance. Especially, from this data analysis, the students would be able to realise those attributes that impacted them and would allow them to fix it where eventually they can do better in their academics. As said before, without students' involvement in this, it wouldn't change anything because they are one can change anything about themselves easily.

7.0 References

1) Kasagga, U., & Nakijoba, S. (2020). Effect of Romantic Relationship on Undergraduate Students' Academic Performance: A Case Study of Islamic University in Uganda. *Islamic University Multidisciplinary Journal. 7(2),* 200-208. Retrieved 2022, from https://www.iuiu.ac.ug/journaladmin/iumj/ArticleFiles/77866.pdf

2) Rezaei-Dehaghani, A., Keshvari, M., & Paki, S. (2018). The Relationship between Family Functioning and Academic Achievement in Female High School Students of Isfahan, Iran, in 2013-2014. *Iranian journal of nursing and midwifery research*, *23*(3), 183–187. Retrieved 2022, from https://doi.org/10.4103/ijnmr.IJNMR_87_17

3) Lara, L., & Saracostti, M. (2019, June 27). *Effect of parental involvement on children's Academic Achievement in Chile*. Frontiers. Retrieved 2022, from https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01464/full

4) Effects of Romantic Relationships on Academic Performance and Family Relationship. (2022, April 20). Retrieved 2022, from https://phdessay.com/effects-romantic-relationships-academic-performance-family-relationship/

5) Bogenschneider , K., & Johnson, C. (2004, February). Family involvement in education: How important is it? What can legislators do? *A policymaker's guide to school finance: Approaches to use and questions to ask*, 19-29. Retrieved 2022, from https://www.purdue.edu/hhs/hdfs/fii/wp-content/uploads/2015/06/fia_brchapter_20c02.pdf

6) Turner, M. (2020, November 24). *Stress can seem unavoidable to students up against it, but relaxation can play a key role in improving study and changing your life...* We Heart. Retrieved May 8, 2022, from https://www.we-heart.com/2020/04/14/how-relaxation-can-improve-study-change-your-life/

7) Barbarick, K. A., & Ippolito, J. A. (2003). *32 • J. Nat. Resour. Life Sci. Educ., Vol. 32, 2003 Does the Number of Hours Studied Affect Exam Performance?* Agronomy. Retrieved 2022, from https://www.agronomy.org/files/jnrlse/issues/2003/e02-14.pdf

8) Hampton, K. N., Robertson, C. T., Fernandez, L., Shin, I., & Bauer, J. M. (2021, October). *How variation in internet access, digital skills, and media use are related to rural student outcomes: GPA, SAT, and educational aspirations*. Telematics and Informatics. Retrieved 2022, from https://www.sciencedirect.com/science/article/pii/S0736585321001052

9) Education Destination Malaysia. (2018, June 29). *5 benefits of extracurricular activities*. Education Destination Malaysia. Retrieved 2022, from https://educationdestinationmalaysia.com/blogs/5-benefits-of-extracurricular-activities

10) *Tidyverse packages*. Tidyverse. (n.d.). Retrieved May 2022, from https://www.tidyverse.org/packages/

11) Adams, R. V., & Blair, E. (2019, January 18). *Impact of Time Management Behaviors on Undergraduate Engineering Students' Performance*. Sage Journals. Retrieved 2022, from https://journals.sagepub.com/doi/full/10.1177/2158244018824506

12) Flashman J. (2012). Academic Achievement and Its Impact on Friend Dynamics. *Sociology of education*, *85*(1), 61–80. Retrieved 2022, from https://doi.org/10.1177/0038040711417014

13) Ella, R. E., Odok, A. O., & Ella, G. E. (2015, November). Influence of Family Size and Family Type on Academic Performance of Students in Government in Calabar Municipality, Cross River State, Nigeria. *International Journal of Humanities Social Sciences and Education (IJHSSE)*. *11(2)*, 108-114. Retrieved 2022, from https://www.arcjournals.org/pdfs/ijhsse/v2-i11/11.pdf

14) Awan, A. G., & kauser, D. (2015). Impact of Educated Mother on Academic Achievement of her Children: A Case Study of District Lodhran- Pakistan. *Journal of Literature, Languages and Linguistics.* Retrieved 2022, from https://core.ac.uk/download/pdf/234693048.pdf

15) Idris, M., Ahmad, N., & Hussain , S. (2020, December). *Relationship between Parents' Education and their children's Academic Achievement*. Research Gate.

Retrieved 2022, from https://www.researchgate.net/publication/348138126_Relationship_between_Parents'_Education_and_their_children's_Academic_Achievement

16) Clearinghouse for Military Family Readiness at Penn State. (2020, January 6). *Parents' educational levels influence on child educational outcomes: Rapid literature.* Retrieved 2022, from https://militaryfamilies.psu.edu/wp-content/uploads/2020/01/Parents-Educational-Levels-Influence-on-Child-Educational-Outcomes.20Jan06.final_.pdf

17) Heinrich, C. J. (2014). *Parents' Employment and Children's Wellbeing*. Retrieved 2022, from https://files.eric.ed.gov/fulltext/EJ1029033.pdf

18) Analytics Vidhya. (2020, January 21). *Effect of alcohol use on academic performance of school students*. Medium. Retrieved 2022, from https://medium.com/analytics-vidhya/effect-of-alcohol-use-on-academic-performance-of-school-students-c9ed44dafbba

19) Placzek, D. (2020, April 22). *How weekend binge drinking can affect academic performance*. CampusWell. Retrieved 2022, from https://www.campuswell.com/the-academic-hangover/

20) Mati, A., Gatumu, J. C., & Chandi, J. R. (n.d.). Students' Involvement in Decision Making and Their Academic Performance in Embu West Sub-County of Kenya. *Universal Journal of Educational Research, 4(10)*. Retrieved 2022, from https://files.eric.ed.gov/fulltext/EJ1116351.pdf

21) Matingwina, T. (2018, September 19). *Health, Academic Achievement and School-Based Interventions*. IntechOpen. Retrieved 2022, from https://www.intechopen.com/chapters/62994

22) Bibi, W., & Ali, A. (n.d.). *The Impact of Pre-school Education on the Academic Achievements of Primary School Students*. Retrieved May 8, 2022, from https://qurtuba.edu.pk/thedialogue/The%20Dialogue/7_2/Dialogue_April_June2012_152-159.pdf