Exploratory Data Analysis of Categorical Variables in Student Performance Dataset

Ken Andrea L. Bahian & Charlene P. Garridos

2024-01-30

Contents

1 R library

Libraries that are needed for the analysis and graphs.

```
# Required Libraries
library(tidyverse)
library(ggplot2)
```

2 Dataset Preparation

Load the Dataset

```
# Read the CSV file
student_performance <- read.csv("student-por.csv")

# View the data
head(student_performance)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        0       yes     no   no         no
## 4   mother          1         3        0        no    yes   no        yes
## 5   father          1         2        0        no    yes   no         no
## 6   mother          1         2        0        no    yes   no        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3
## 1        4  0 11 11
## 2        2  9 11 11
## 3        6 12 13 12
## 4        0 14 14 14
## 5        0 11 13 13
## 6        6 12 12 13
```

3 Variable Selection and Convertion

Display internal structure of the dataset.

```
# Check the structure of the variables
str(student_performance)
```

```
## 'data.frame':    649 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "no" "no" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  4 2 6 0 0 6 0 2 0 0 ...
##  $ G1        : int  0 9 12 14 11 12 13 10 15 12 ...
##  $ G2        : int  11 11 13 14 13 12 12 13 16 12 ...
##  $ G3        : int  11 11 12 14 13 13 13 13 17 13 ...
```

Identify the dataset's unique values.

```
# unique() function to provide the unique values of student_performance$studytime.
unique(student_performance$studytime)
```

```
## [1] 2 3 1 4
```

```
# unique() function to provide the unique values of student_performance$failure.
unique(student_performance$failures)
```

```
## [1] 0 3 1 2
```

```
# unique() function to provide the unique values of student_performance$famrel.
unique(student_performance$famrel)
```

```
## [1] 4 5 3 1 2
```

```
# unique() function to provide the unique values of student_performance$health.
unique(student_performance$health)
```

```
## [1] 3 5 1 2 4
```

Convert the data frame column to factors.

```
# as.factor() function to convert a vector object to a factor.
student_performance$studytime <- as.factor(student_performance$studytime)
student_performance$failures <- as.factor(student_performance$failures)
student_performance$famrel <- as.factor(student_performance$famrel)
student_performance$health <- as.factor(student_performance$health)
```

Display internal structure of the dataset

```
# Checking the internal structure again if it is now a factor.
str(student_performance)
```

```
## 'data.frame':    649 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : Factor w/ 4 levels "1","2","3","4": 2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "no" "no" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : Factor w/ 5 levels "1","2","3","4",..: 4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
```

```
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : Factor w/ 5 levels "1","2","3","4",..: 3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  4 2 6 0 0 6 0 2 0 0 ...
##  $ G1        : int  0 9 12 14 11 12 13 10 15 12 ...
##  $ G2        : int  11 11 13 14 13 12 12 13 16 12 ...
##  $ G3        : int  11 11 12 14 13 13 13 13 17 13 ...
```

4 Data Exploration

4.1 Frequency, Relative Frequency and Mode

It shows the frequency, relative frequency and the mode of each selected variables.

```
student_studytime <- student_performance$studytime

# table() function to calculate frequencies
# prop.table() function to calculate relative frequencies
freq <- table(student_studytime)
relative_freq <- prop.table(freq)

# A data frame for the tables
result <- data.frame(student_studytime = names(freq), Frequency = as.vector(freq),
                     Relative_Frequency = as.vector(relative_freq))

result
```

```
##   student_studytime Frequency Relative_Frequency
## 1                 1       212         0.32665639
## 2                 2       305         0.46995378
## 3                 3        97         0.14946071
## 4                 4        35         0.05392912
```

```
# Displays the mode of student_studytime
mode <- names(freq[freq == max(freq)])
mode
```

```
## [1] "2"
```

```
student_failures <- student_performance$failures

# table() function to calculate frequencies
# prop.table() function to calculate relative frequencies
freq <- table(student_failures)
relative_freq <- prop.table(freq)

# A data frame for the tables
result <- data.frame(student_failures = names(freq), Frequency = as.vector(freq),
                     Relative_Frequency = as.vector(relative_freq))

result
```

```
##   student_failures Frequency Relative_Frequency
## 1                0       549         0.84591680
## 2                1        70         0.10785824
## 3                2        16         0.02465331
## 4                3        14         0.02157165
```

```
# Displays the mode of student_failures
mode <- names(freq[freq == max(freq)])
mode
```

```
## [1] "0"
```

```
student_famrel <- student_performance$famrel

# table() function to calculate frequencies
# prop.table() function to calculate relative frequencies
freq <- table(student_famrel)
relative_freq <- prop.table(freq)

# A data frame for the tables
result <- data.frame(student_famrel = names(freq), Frequency = as.vector(freq),
                     Relative_Frequency =as.vector(relative_freq))

result
```

```
##   student_famrel Frequency Relative_Frequency
## 1              1        22         0.03389831
## 2              2        29         0.04468413
## 3              3       101         0.15562404
## 4              4       317         0.48844376
## 5              5       180         0.27734977
```

```
# Displays the mode of student_famrel
mode <- names(freq[freq == max(freq)])
mode
```

```
## [1] "4"
```

```
student_health <- student_performance$health

# table() function to calculate frequencies
# prop.table() function to calculate relative frequencies
freq <- table(student_health)
relative_freq <- prop.table(freq)

# A data frame for the tables
result <- data.frame(student_health = names(freq), Frequency = as.vector(freq),
                     Relative_Frequency = as.vector(relative_freq))

result
```

```
##   student_health Frequency Relative_Frequency
```

```
## 1              1        90          0.1386749
## 2              2        78          0.1201849
## 3              3       124          0.1910632
## 4              4       108          0.1664099
## 5              5       249          0.3836672
```

```r
# Displays the mode of student_health
mode <- names(freq[freq == max(freq)])
mode
```

```
## [1] "5"
```

4.2 Visualization of the Distribution

It displays the count or distribution of each selected variables of the dataset.

```r
# To arrange plot in 2X2 position
par(mfrow = c(2,2))

# plot() function for bar graphs of each selected variables
plot(student_studytime, xlab = "Weekly Study Time",
     ylab = "Count",
     main = "Distribution of Weekly Study Time",
     col = "orchid4")

plot(student_failures, xlab = "Number of Past Class Failures",
     ylab = "Count",
     main = "Distribution of Previous Failed Classes",
     col = "orchid3")

plot(student_famrel, xlab = "Quality of Family Relationships",
     ylab = "Count",
     main = "Distribution of Family Relationships Quality",
     col = "orchid2")

plot(student_health, xlab = "Health Status",
     ylab = "Count",
     main = "Distribution of Current Health Status",
     col = "orchid1")
```

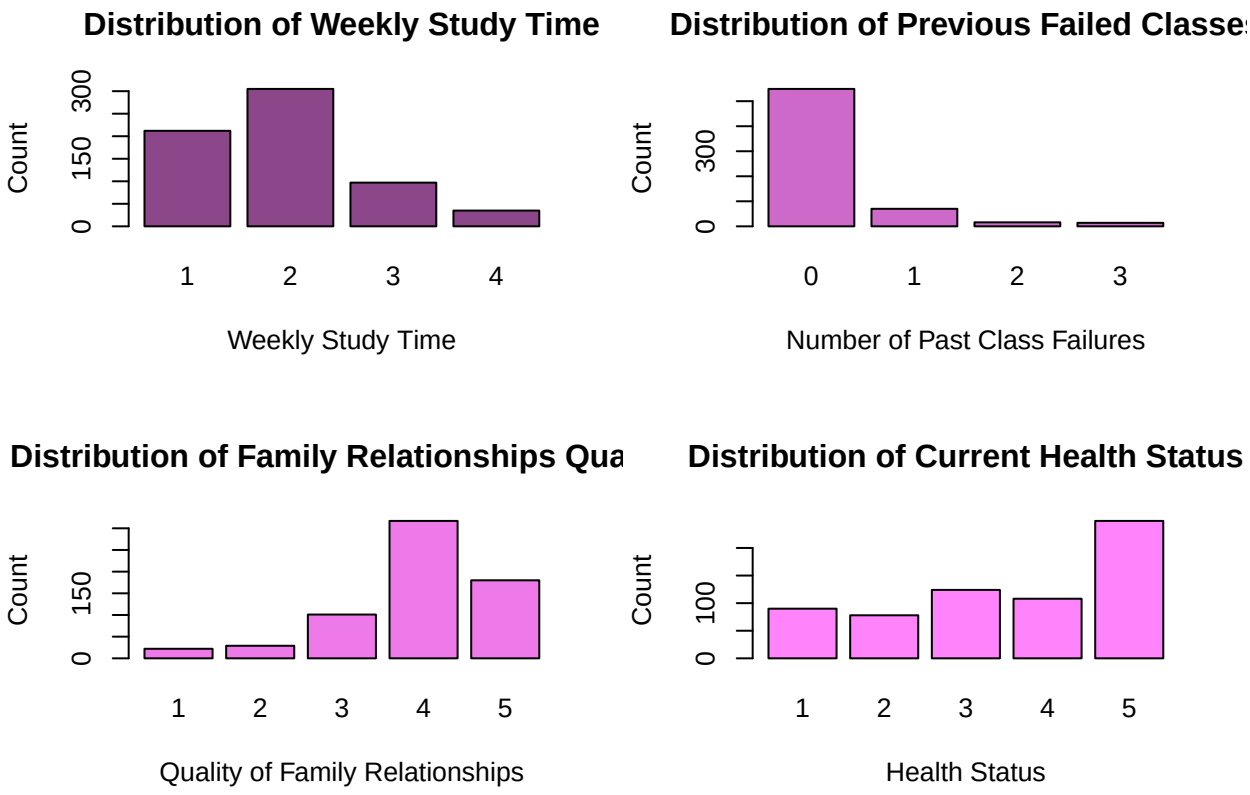

Figure 1: The Distribution of Weekly Study Time, Number of Past Class Failures, Quality of Family Relationships, and Health Status

Interpretation

Study time: weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

Upon closer examination, the distribution of weekly study time displays a right-skewed distribution as it peaks on the left and has a longer right side. This might imply that many students spend less time studying compared to a minority who invest more hours. The graph provides insights into the varied study habits among students. The option with the highest number of responses is 2, chosen by a significant majority of 47% or 305 students in the total sample, indicating a dedication of 2 to 5 hours for studying. This suggests that most students prefer a balanced and steady approach to their studies, possibly allowing time for other activities in their lives, not solely focusing on academics.

The next highest option on the distribution of weekly study time is 1, chosen by 33% or 212 students of the total sample who allocate less than 2 hours to studying, which is notably low in terms of hours of studying. Option 3 has a lower level on the graph, with 97 students dedicating 5 to 10 hours to studying, indicating a focus on academic pursuits. Lastly, the lowest option on the graph is 4, where a group of 35 students invests more than 10 hours in studying, demonstrating a commitment to academic excellence and a potential desire to excel in their chosen field of study. In summary, the graph illustrates the diverse study hours among students, ranging from those with shorter study periods to those dedicating more extensive time to their academic pursuits.

Failures: number of past class failures (numeric: n if 1<=n<3, else 4)

The graph about past class failures shows that 85% of the total surveyed students have never failed a class, which is impressive. These 349 students are likely dedicated, hardworking, or naturally talented in their studies. Their consistent academic success is probably a result of a combination of effort and innate ability. On the other hand, options 1, 2, and 3 have low levels in the graph. Among these options, 1 has the highest count, with 70 students having experienced failure in one class. This might be due to challenges in a specific subject, temporary setbacks, or adjusting to the academic environment. A smaller number, 16 students, faced failure in two classes, indicating more complex academic challenges or difficulties in multiple subjects. Additionally, 14 students have failed three past classes, suggesting ongoing academic struggles that may need a comprehensive support plan, personalized academic assistance, or a reconsideration of their educational approach. To sum it up, most students haven't faced class failures, some had one failure, and only a few encountered at least two failures.

Famrel: quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

Upon closer inspection, the graph depicting family relationship exhibits a left-skewed distribution, meaning that its left tail is longer and it peaks towards the right. This implies that most students have a solid and healthy relationship with their families. More specifically, for the majority of students, a positive association is implied by the largest number of students, which is indicated by the peak at number 4 on the graph. Around 49% of the surveyed students, which is 317 in total, chose this rating. The fact that many students chose the second-highest rating indicates that students are generally happy and content with their family relationships. The second-largest group on the graph is represented by the number 5, with about 45% or 180 students giving the highest rating to their family relationship. This suggests that students and their families have a wonderful relationship. The next rating is 3, chosen by exactly 101 students. On the other hand, the lowest ratings on the graph are 2 and 1, with 29 students opting for 2 and only 22 choosing 1. This suggests that a small group of students may not be as happy with their family situation, possibly due to various reasons. Overall, it seems like most students have a wonderful relationship with their families.

Health: current health status (numeric: from 1 - very bad to 5 - very good)

The graph depicting how students perceive their health provides a detailed look at the various well-being experiences among students. Using a scale of 1 to 5, where 5 is the highest health status, a considerable number of 249 students or 38% of all the surveyed students gave themselves the highest rating of 5, indicating an excellent health status. These students likely prioritize well-being, follow healthy habits, and have strong immune systems.

The next highest rating is 3, with a significant gap from 5 but a smaller gap from option 4. In this category, 19% or 124 students voted that they have neither good nor bad health status. Option 4, with 17% of all surveyed students or 108 students, suggests that these students generally have good health conditions. Moving down the scale, option 1 has 90 students who rated their health as 1, indicating a very poor health condition. Possible reasons for this rating could include existing health issues, problems that developed over time, or difficulties with sleep patterns. The lowest rating is 2, chosen by 78 students who perceive their health as bad. In summary, the graph showcases diverse health perceptions among students, ranging from excellent to poor health conditions.

4.3 Contingency table

A multiple categorical contingency table is a tabular representation of the joint distribution of variables.

```
# Display the multiple categorical contingency table
table(student_failures, student_health, student_studytime, student_famrel,
      dnn = c("Failures","Health Status", "Study Time", "Family Relationship"))
```

```
## , , Study Time = 1, Family Relationship = 1
##
##         Health Status
## Failures  1  2  3  4  5
##        0  2  0  0  0  3
##        1  0  0  0  0  1
##        2  0  0  0  0  0
##        3  0  0  0  0  0
##
## , , Study Time = 2, Family Relationship = 1
##
##         Health Status
## Failures  1  2  3  4  5
##        0  4  2  0  0  2
##        1  1  0  0  0  0
##        2  0  0  0  0  0
##        3  0  0  0  0  0
##
## , , Study Time = 3, Family Relationship = 1
##
##         Health Status
## Failures  1  2  3  4  5
##        0  0  0  0  0  1
##        1  0  0  0  2  0
##        2  0  0  0  0  0
##        3  0  0  0  0  0
```

```
##
## , , Study Time = 4, Family Relationship = 1
##
##          Health Status
## Failures  1  2  3  4  5
##        0  2  0  1  0  0
##        1  0  0  1  0  0
##        2  0  0  0  0  0
##        3  0  0  0  0  0
##
## , , Study Time = 1, Family Relationship = 2
##
##          Health Status
## Failures  1  2  3  4  5
##        0  2  2  0  1  2
##        1  0  0  0  0  2
##        2  0  0  0  1  0
##        3  0  0  0  0  1
##
## , , Study Time = 2, Family Relationship = 2
##
##          Health Status
## Failures  1  2  3  4  5
##        0  2  1  5  2  1
##        1  1  0  0  2  0
##        2  1  0  0  0  0
##        3  0  0  0  0  0
##
## , , Study Time = 3, Family Relationship = 2
##
##          Health Status
## Failures  1  2  3  4  5
##        0  0  0  0  0  1
##        1  0  0  0  0  0
##        2  0  0  0  0  0
##        3  0  0  0  0  0
##
## , , Study Time = 4, Family Relationship = 2
##
##          Health Status
## Failures  1  2  3  4  5
##        0  0  0  1  0  1
##        1  0  0  0  0  0
##        2  0  0  0  0  0
##        3  0  0  0  0  0
##
## , , Study Time = 1, Family Relationship = 3
##
##          Health Status
## Failures  1  2  3  4  5
##        0  5  2  9  8 10
##        1  1  2  0  0  2
##        2  0  0  0  0  1
##        3  0  0  1  0  1
```

```
##
## , , Study Time = 2, Family Relationship = 3
##
##          Health Status
## Failures  1  2  3  4  5
##        0  5  5  9  4  8
##        1  0  0  3  0  2
##        2  0  0  0  0  2
##        3  2  0  0  1  0
##
## , , Study Time = 3, Family Relationship = 3
##
##          Health Status
## Failures  1  2  3  4  5
##        0  1  2  4  2  2
##        1  0  0  0  0  0
##        2  0  1  0  0  1
##        3  0  0  0  0  0
##
## , , Study Time = 4, Family Relationship = 3
##
##          Health Status
## Failures  1  2  3  4  5
##        0  2  1  2  0  0
##        1  0  0  0  0  0
##        2  0  0  0  0  0
##        3  0  0  0  0  0
##
## , , Study Time = 1, Family Relationship = 4
##
##          Health Status
## Failures  1  2  3  4  5
##        0  5  5  8 14 39
##        1  1  4  3  3  5
##        2  0  0  0  0  1
##        3  0  1  1  0  1
##
## , , Study Time = 2, Family Relationship = 4
##
##          Health Status
## Failures  1  2  3  4  5
##        0 23 20 26 24 49
##        1  0  2  3  1  6
##        2  0  0  2  1  2
##        3  0  0  0  0  0
##
## , , Study Time = 3, Family Relationship = 4
##
##          Health Status
## Failures  1  2  3  4  5
##        0  2  5 13 11 20
##        1  0  0  0  0  2
##        2  0  0  0  0  0
##        3  0  0  0  0  0
```

```
##
## , , Study Time = 4, Family Relationship = 4
##
##         Health Status
## Failures  1  2  3  4  5
##        0  1  2  3  4  4
##        1  0  0  0  0  0
##        2  0  0  0  0  0
##        3  0  0  0  0  0
##
## , , Study Time = 1, Family Relationship = 5
##
##         Health Status
## Failures  1  2  3  4  5
##        0  5  3  9  6 23
##        1  3  2  2  0  5
##        2  0  0  0  0  1
##        3  1  0  0  1  1
##
## , , Study Time = 2, Family Relationship = 5
##
##         Health Status
## Failures  1  2  3  4  5
##        0 13 13 10  9 27
##        1  0  2  0  0  3
##        2  0  0  0  2  0
##        3  0  0  0  0  2
##
## , , Study Time = 3, Family Relationship = 5
##
##         Health Status
## Failures  1  2  3  4  5
##        0  4  1  4  9  7
##        1  0  0  1  0  1
##        2  0  0  0  0  0
##        3  0  0  0  0  0
##
## , , Study Time = 4, Family Relationship = 5
##
##         Health Status
## Failures  1  2  3  4  5
##        0  1  0  3  0  5
##        1  0  0  0  0  1
##        2  0  0  0  0  0
##        3  0  0  0  0  0
```

Note:

- Study time: weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10
- Failures: number of past class failures (numeric: n if 1<=n<3, else 4)
- Famrel: quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- Health: current health status (numeric: from 1 - very bad to 5 - very good)

Interpretation

In the first table, it is noted that most students with no failures exhibit good health status. Some students also have zero failures despite a very bad health condition. A similar pattern is observed in the second table; however, one student failed one class. In the third table, only one student has a very bad family relationship yet was able to pass all classes, have good health, and have enough time to study. Similarly, Table 4 demonstrates fewer counts, but students with 1 failure and a better health status are noticeable. In Tables 5-8, where family relations are bad (2), most of the students pass, and there are also students who failed 1 or 2 classes. The same pattern can be observed in tables 9–10. It is also apparent that students with longer study times appear to have better health status. Moreover, in the following tables, where the family relationship of the student is better, a larger count can be observed. This implies that many students with good family relationships pass their classes. Only a few students failed some of their classes, probably due to other factors. In cases where study time is longer and family relationships are positive, instances of academic failure are markedly reduced. It suggests a potential relationship between effective study habits based on the time a student studies, familial support based on the quality of the relationship of the family, and overall student well-being.

A plot to visualize the relationship among the selected variables: class failures, health status, study time and family relationship.

```
# ggplot() function to see the relationship between the chosen variables as a bar graph.
ggplot(student_performance, aes(x = factor(student_failures),
                                fill = factor(student_health))) +
  geom_bar(position = "dodge") +
  labs(title = "Number of Failed Classes vs. Current Health Status",
       x = "Number of Failed Classes",
       y = "Number of Students"
  )
```

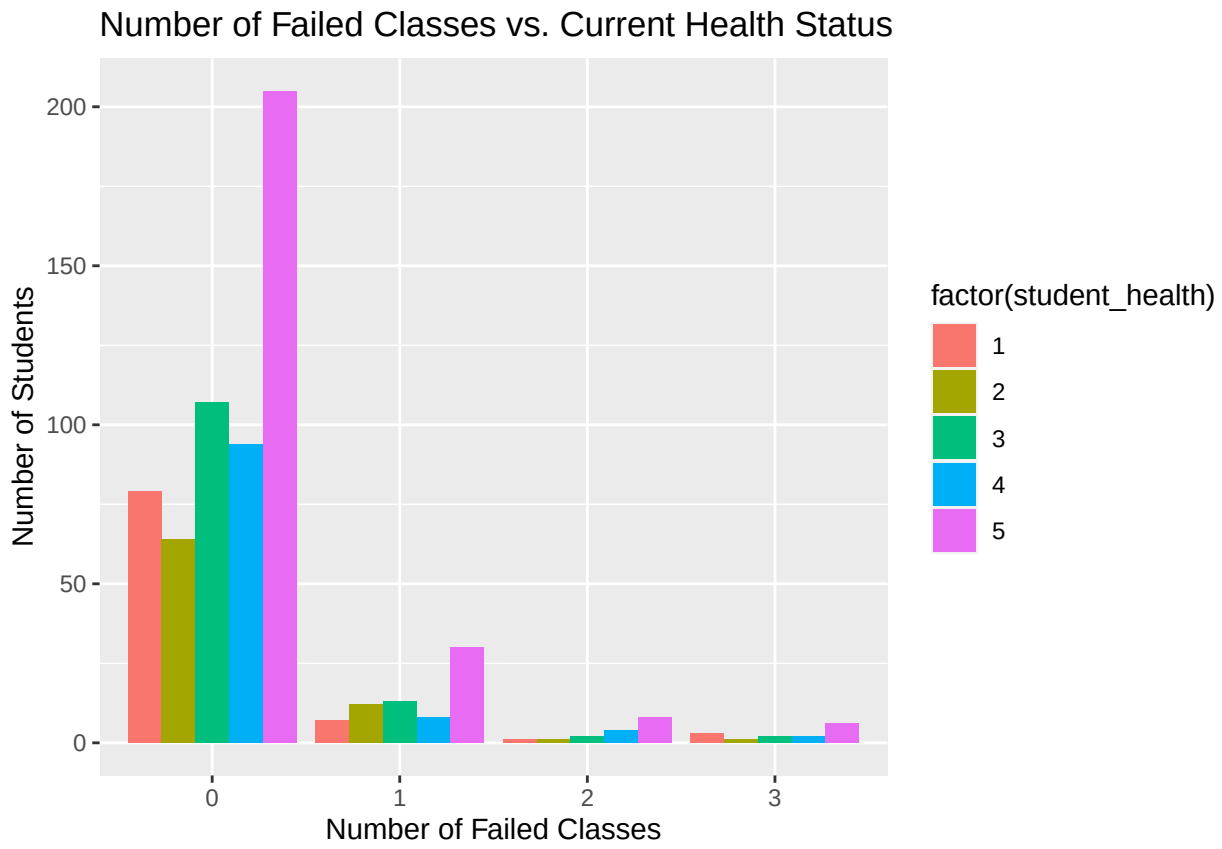

Figure 2: Distribution of Number of Failures and Current Health Status

Interpretation

Based on the observations from figure 2, it seems that many students who selected different health status options have not experienced failing a class. This suggests that having a very bad health status may not be the primary reason for a student failing a class. While it could be one factor, the graph shows that a significant number of students who rated their health as 1 (very poor) have not failed a class. Similarly, even students with excellent health status (rated as 5) have experienced 1, 2, and 3 failed classes, as indicated on the graph.

Therefore, the graph implies that having poor health is not necessarily a determinant for a student failing a class. Other factors may play a role, and the relationship between health status and academic performance is not straightforward.

```r
# ggplot() function to see the relationship between the chosen variables as a bar graph.
ggplot(student_performance, aes(x = factor(student_failures),
                                fill = factor(student_studytime))) +
  geom_bar(position = "dodge") +
  labs(title = "Number of Failed Classes vs. Weekly Study Time",
    x = "Number of Failed Classes",
    y = "Number of Students"
  )
```

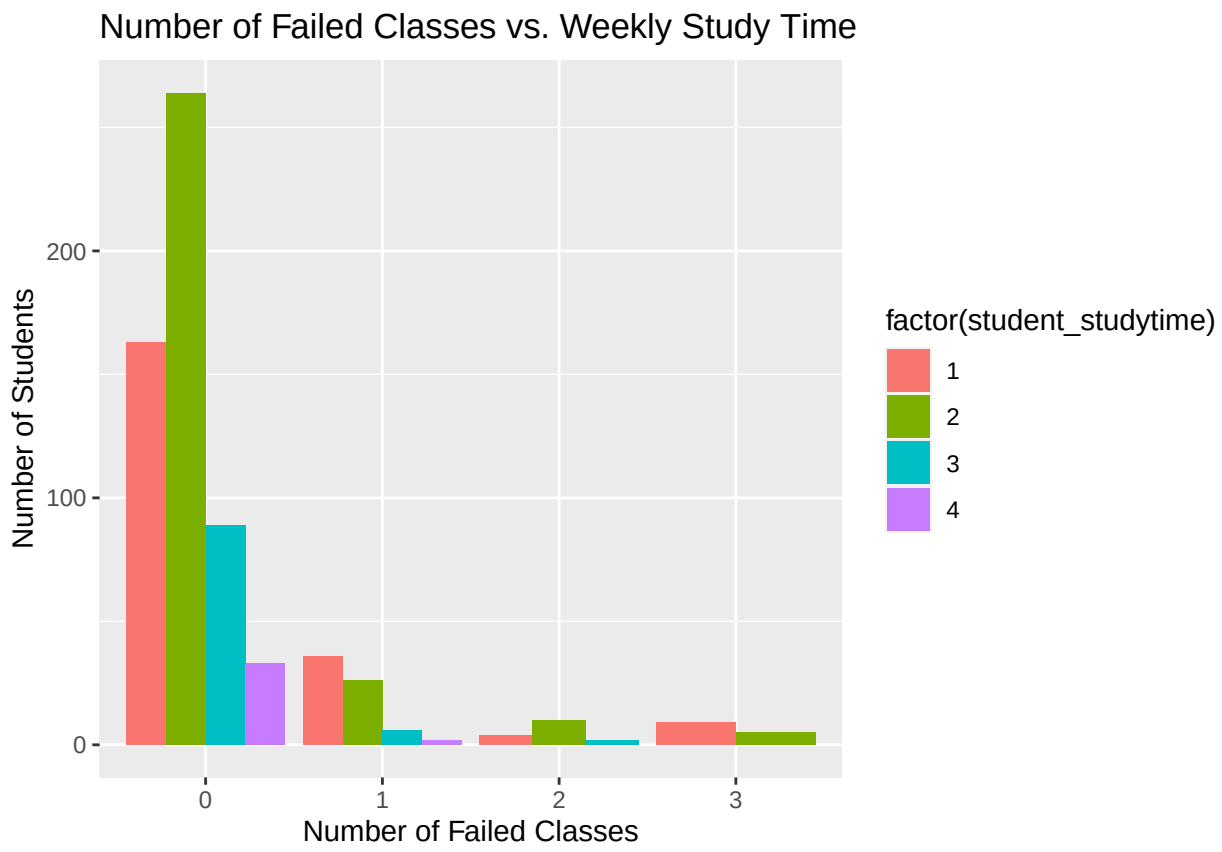

Figure 3: Distribution of Number of Failures and Weekly Study Time

Interpretation

Figure 3 indicates that the majority of students have not failed a class. Specifically, those who chose option 4 (studying for more than 10 hours per week) mostly haven't experienced failing a class. Even among the minority who did fail a class in this group, the number of failed classes is generally low, usually not exceeding 1. This suggests that the number of hours spent studying is indeed a significant factor in not having failed classes. Students who chose option 3 (studying for 5-10 hours per week) also have a low incidence of failing more than 2 times, supporting the idea that study hours contribute to academic success. There's a noticeable contrast in the data for option 2 (students studying for 2-5 hours per week) compared to the other choices. It appears that a significant number of students fall within the 2-5 hours per week study range. Although most have not failed a class, there is also a segment within this group that experienced three failures. This suggests that students dedicating 2-5 hours to studying have diverse experiences, with some achieving success in their studies and others encountering challenges. Those who chose option 1 (studying for less than 2 hours per week) mostly have not failed a class. However, it's interesting to note that among this group, there are

instances of failing 3 classes. This suggests that while some students studying less still succeed, there is a higher risk of failure among this group.

In summary, the data suggests that more study hours generally correlate with a lower likelihood of failing a class. While there are variations and exceptions, the trend indicates that dedicating more time to studying increases the chances of academic success.

```
#ggplot() function to see the relationship between the chosen variables as a bar graph.

ggplot(student_performance, aes(x = factor(student_failures),
                                fill = factor(student_famrel))) +
  geom_bar(position = "dodge") +
  labs(title = "Number of Failed Classes vs. Family Relationship",
    x = "Number of Failed Classes",
    y = "Number of Students"
  )
```

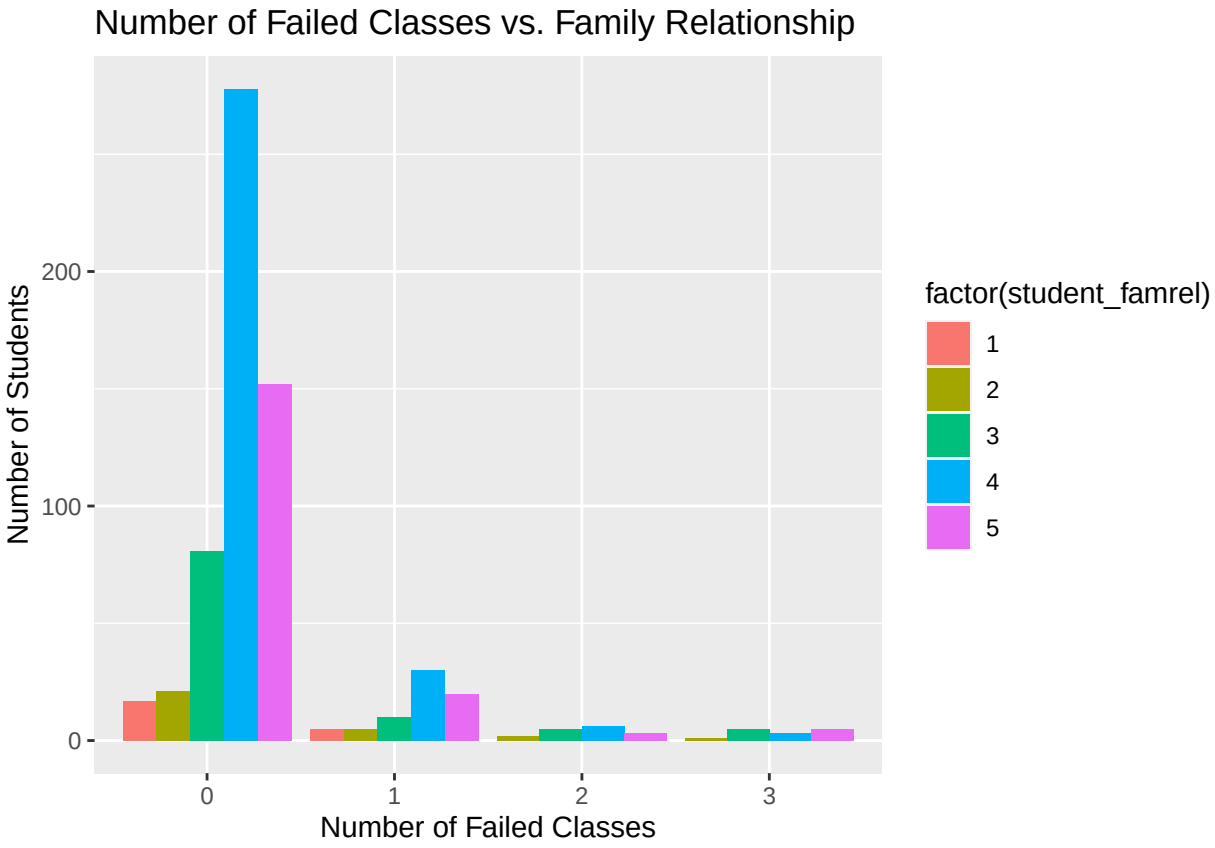

Figure 4: Distribution of Number of Failures and Family Relationship

Interpretation

Figure 4 illustrates the relationship between family dynamics and the occurrence of failed classes. Despite the low number of students who selected option 1, indicating a very poor relationship with their family, these students either have not failed a class or have only experienced one failure. This suggests that family relationship may not be a significant factor in predicting whether a student will fail a class. While it could potentially play a role, the data does not strongly support the idea, as students with a very bad family relationship have 0-1 failed classes.

Conversely, students who chose options 2, 3, 4, and 5 for their level of family relationship have experienced failed classes.While the majority have not experienced failing a class, a portion of them has encountered 2-3 failures. This implies that having a good relationship with one's family does not guarantee immunity from experiencing failed classes. In essence, the data suggests that family relationship may not have an influence on the number of failed classes a student may encounter.

5 Data Summary

Taking the summary of each variables in the dataset.

```
# summary() function to get overall summary of the data frame.
summary(student_studytime)
```

```
##   1   2   3   4
## 212 305  97  35
```

```
# summary() function to get overall summary of the data frame.
summary(student_failures)
```

```
##   0   1   2   3
## 549  70  16  14
```

```
# summary() function to get overall summary of the data frame.
summary(student_famrel)
```

```
##   1   2   3   4   5
##  22  29 101 317 180
```

```
# summary() function to get overall summary of the data frame.
summary(student_health)
```

```
##   1   2   3   4   5
##  90  78 124 108 249
```

Discussion

The provided data, along with the accompanying tables and graphs, reveals potential relationships among the selected categorical variables. Looking at the summary of each category, more than half of the students opted for categories 1 and 2, where they studied for less than 2 hours and 2-5 hours, respectively. The remaining students, less than half of the surveyed group, selected categories 3 and 4, indicating study durations of 5-10 hours and more than 10 hours, respectively. The most common category is 2, chosen by 305 students. For the category of number of failures, 549 students have never experienced a failure. On the other hand, a combined total of 100 students have encountered at least 1 failure and at most 3 failures. The dominant category, or mode, for this variable, is zero, with 549 students selecting this category. In terms of family relationships, a few students chose categories 1 and 2, suggesting a less favorable relationship with their family. A total of 101 students chose Category 3, whereas the majority, comprising 317 students, selected Category 4; additionally, 180 students opted for the highest rating, category 5, signaling a significant number of students expressing a positive and strong relationship with their family. The most frequently chosen category is 4, selected by 317 students. Regarding health, the majority of students reported good to great health. However, a significant

number also chose categories 1 and 2, indicating very bad and bad health statuses. Category 5 is the most frequent, with 249 students selecting this option.

Looking at the contingency table, it's evident that students with effective study habits tend to perform better academically. Those who dedicate more hours to studying have a lower occurrence of failed classes compared to those who study for shorter durations. Even when considering variations in health status and family relationships, students who spend more time studying are less likely to fail a subject. The data implies that longer study times play a crucial role in reducing academic failures. While health status and family relationships may contribute to the occurrence of failed subjects, the relationship observed is with the number of hours dedicated to studying.

Analyzing the graphs that potentially depict the relationship between the number of failed classes and the health of students, the weekly study time, and family relationships, several insights can be drawn. The first graph indicates that having good health doesn't necessarily mean a student has never failed a subject. Both students with great and poor health can experience failures. In the second graph, longer study times correlate with a lower likelihood of failing at least one subject, while shorter study times are associated with a higher chance of failing, with a maximum of three subjects. This suggests a significant impact of weekly study time on the number of failures. The last graph explores the potential relationship between the number of failed classes and family relationships. Students with a very bad family relationship seem to have a maximum of one failed subject, while other categories show failures, with a maximum of three subjects. This hints that family relationships may influence the occurrence of failed classes, but having a great relationship with family does not guarantee freedom from failure.

6 Conclusion

In conclusion, the comprehensive analysis of the provided data, covering study habits, academic performance, family relationships, and health statuses, has revealed useful insights into how these aspects interact among students. In particular, more than half of the surveyed students spend relatively fewer hours studying, mostly in the 2-5 hours range. Despite this, the analysis shows a connection between effective study habits, especially longer study times, and better academic performance. The most common scenario is zero failures, suggesting a pattern or trend of academic success. Family relationships play a noticeable role, with most students indicating positive connections, while health statuses show varied distributions. The analysis emphasizes that, regardless of health status and family relationships, the number of study hours significantly influences academic outcomes. Longer study times are linked to a lower chance of academic failure. The graphical representations further support the impact of study habits on academic performance, demonstrating that longer study times are associated with fewer instances of academic failure. The relationship between family dynamics and academic outcomes is clear, but the duration of study emerges as a more influential factor.

This analysis highlights the crucial role of effective study habits in shaping academic success among students, while also recognizing the nuanced influence of family relationships and health statuses. These findings offer valuable insights for educational institutions and support services to create helpful actions that take into account how these different factors interact.

7 Reference

"Student Performance in Secondary Education" dataset