Analysis of Student Performance, Stress and Academic Success - Data Science Approach

Mathan Kumar Thavu Mudaliar Kanagaraj

2024-08-04

Milestone 1

INTRODUCTION

The objective of this study is to analyze and understand the factors influencing student's performance, stress and academic success. Student phase is considered to be one that builds the foundation of an individual's life. Students are the most important component of society; they determine the success of society. Understanding and mitigating major factors that causes stress in student is vital to prevent physical and mental health problem and ensure academic success.

Applications of Data Science Methods

Scientific study on contributing factors can be a major help to teachers, parents and student councilors, it can aid in planning, executing and monitoring activities to boost student performance. Understanding the root cause of stress will help evidence-based decision-making to reduce or eliminate stress directly resulting in better quality of life and greatly improving the chance of academic success. The challenge in hand ideal for applying data science methods as it involves gathering data, transforming the raw data to a superior quality that can help perform analysis, and a statistical modeling is required to gain meaningful insight.

Research Questions

1. What are the major factors that influence performance of the students? Are there factors that have influence on Mathematics and not Portuguese or vice versa?
2. What is the impact of alcohol consumption in Student performance? Is there evidence of positive impact on student that have no or low alcohol consumption? How does consuming alcohol in weekdays vs weekends impact the Student Performance?
3. What is the primary contributing to stress in Students? Is there linear relationship between one of more contributing factor and the stress level?
4. What are the top most contributing indicator (example: Sleep Quality in Physiological factor, Peer Pressure in Social Factor, etc.,) within each category and how does top contributing indicators of each category compare against top contributing indicators of other categories? Is there a clear winner that be considered the most influential category?
5. What are factors that limits academic success and what are factors that enhance academic success? Is there any factor such as age, gender, nationality that shows a clear trend in impact academic success?

Data Science Approach

Understanding the Business Problem

The objective and scope of the study will be defined in the above research question topic clearly and unambiguously.

Data Preparation

Collecting Data – Student Performance, Stress and Academic Success data have been collected from various sources for the study

Integrating and Combining Data – Combine data elements between datasets to understand the correlation between Student Performance, Stress and Academic Success

Transforming Data – Data will be corrected for spelling error, duplicate rows/columns will be removed and missing data will be addressed (if any)

Enriching Data – Supplementary data will be added wherever required (for example: Country Name for the Nationality indicator) in the Student Performance dataset

Validating Data – Datasets will be validated to check for correctness of data type, correctness of data (acceptable values based on column definition), format, and uniqueness.

Exploratory Data Analysis

Analyzing and visualizing data to understand key characteristics, uncover patterns, and identify relationships between variables in the dataset. Identify predominant traits, discover patterns, locate outliers, and identify relationships between variables.

Building Model

Build a statistical model that help understand the relationship between input variables (contributing factors such as Sleep Quality, Health of Relationship, etc.,) and the target variable (Student Performance). These models can help assess the relationships and their strength between variables that can in turn help mitigate the issues to influence better outcome

Communicating results

Illustrate the results using plots, graphs and other visualizations effectively to relevant stakeholders empowering them with insights for better decision-making.

Addressing the Problem with Proposed Approach

1. Business Problem, in this case – Analyze, understand, and communicate the factors contributing to Student Stress and the factors influencing Student Performance and Academic Success
2. Data collected such as Student Performance Dataset, Student Stress Factors Dataset, and Student Dropout and Academic Success Dataset will be thoroughly cleaned, transformed and validated to ensure the data is clean prior to analysis and modeling phase
3. Data that is a result of Data Preparation step will be analyzed to identify patterns, discover insights, and compare/contrast competing factors that influence target variables such as Performance, Stress and Academic Success/Dropout.

4. Statistical Models will be built using the Data frames built from select vectors within the datasets to establish the results between variables to show the impact of contributing factors to the target variable and compare between various variables and how they impact the outcome
5. Scatter Plots, Bar Graph, Heatmap and other result sets will be shared with stakeholders providing them insights to make decisions

Datasets:

Students Performance

The dataset created through school reports and questionnaires from secondary school contain factors influencing the student performance. The object of this dataset is to analyze and understand the negative and positive factors influencing the performance of the students across two subjects, Mathematics and Portuguese. The dataset contains student related information such as external factors(travel time, extracurricular activities) social factors (such as internet access, family relationship, romantic relationship) behavioral factors (such as alcohol consumption, absences, free time), and target variables (first period grade, second period grade and final grade)

https://archive.ics.uci.edu/dataset/320/student+performance

Cortez,Paulo. (2014). Student Performance. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.

Students Stress Factor

The dataset (gathered through a combination of online and offline survey as per author in Kaggle) contains major contributing factors related to stress in students. The objective of this dataset is to identify and understand various factors resulting in stress for students. The dataset contains contributing factors such as Psychological, Physiological, Environmental, Academic and Social factors and the target variable (stress level).

https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis/data

Student Stress Factors: A Comprehensive Analysis. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis/data

Students Dropout and Academic Success (SDAS)

The dataset created from a higher education institution contains details about student enrollment to reduce the academic dropout and failures in higher education. The dataset contains application details (such as application mode, course, daytime/evening), application details (such as marital status, nationality, previous qualification, parent's qualification), external factors (such as inflation rate, unemployment rate, GDP) and the target variable (dropout, enrolled, graduate)

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Realinho,Valentim, Vieira Martins,Mónica, Machado,Jorge, and Baptista,Luís. (2021). Predict Students' Dropout and Academic Success. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89

Packages in R

The list of R packages that are needed for the project

1. Data Visualization and Exploration
    - ggplot2
2. Data Wrangling and Transformation
    - dplyr
    - tidyr
3. Data Manipulation
    - tidyverse
4. Data Classification and Regression Testing
    - Caret

Plots and Table

1. Scatter plot to visualize relationship between input variables and target variables
    - Student Performance: Internet Access, Parents Alcohol Consumption, Relationships Vs Performance
    - Student Stress: Sleep Quality, Academic Performance, Peer Pressure, Bullying Vs Stress Level
    - Student Academic Success: Displacement, Employment Rate, Parent's Qualification, Inflation Rate Vs Academic Success
2. Tables (R Data frames) across 3 datasets based on the use case. (Select data elements and filter them as per the purpose of analysis)
3. Heatmaps to visualize the intensity of the impact caused by input variables on response variables
4. Bar Chart to perform a comparative study between various factors
    - Compare factors impacting student's academic success such as alcohol consumption, travel time, absence
    - Analyze the major contributors to Student's stress such as Academic Performance, Breathing Problem, Living Conditions

Questions for Future Steps

1. Drawing inference by combining datasets or joining data available in different datasets is possible only after exploring the individual datasets in depth
2. Goodness of fit will determine if the analysis of target variable is performed based on all predictor variables
3. Incomplete/Missing Data can be determined during further research

Milestone 2

How to import and clean my data?

Importing Data:

Datasets are imported into R program using read.csv() function. The "read.csv()" function is in "utils" packages which is automatically loaded into an R session on startup.

```
# Student Performance, Stress and Success Analysis
# install.packages("janitor")

library("dplyr")
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("ggplot2")
library("tidyr")
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(ggpubr)
library(ggsci)

stu_perf_math_ds <- read.csv("/Users/mathan/r/1A_student_math_performance.csv")
stu_stress_ds <- read.csv("/Users/mathan/r/2_stress_level.csv")
stu_success_ds <- read.csv("/Users/mathan/r/3_student_academic_success.csv")
```

Cleaning the Data:

- Used subset() function to SELECT data columns to ensure all rows selected are unique, this step will avoid duplicate entries for analysis
- Used is.na() function to check for missing data, this step will ensure analysis is performed on stable data
- Used names() function to rename the column headers to meaningful(easy to understand) column headers

- Converted column headers to lower case to maintain consistency of column headers
- Used clean_names() function from "Janitor" package to cleanup column headers, this step will replace the special characters in the column headers such as "/", space, "(", ")" with underscore "_" symbol

```
# selecting student performance related columns that we will focus for analysis
perf_math_df <- subset(stu_perf_math_ds, select = c(
                                                    school,
                                                    sex,
                                                    age,
                                                    famsize,
                                                    Pstatus,
                                                    Mjob,
                                                    reason,
                                                    traveltime,
                                                    studytime,
                                                    failures,
                                                    internet,
                                                    romantic,
                                                    goout,
                                                    Dalc,
                                                    Walc,
                                                    health,
                                                    absences,
                                                    G1,
                                                    G2,
                                                    G3)
                       )


# Identify student performance data columns that have NULL (missing) values
colSums(is.na(perf_math_df)==TRUE|perf_math_df=='99'|perf_math_df=='')
```

```
##     school        sex        age    famsize    Pstatus       Mjob     reason
##          0          0          0          0          0          0          0
## traveltime  studytime   failures   internet   romantic      goout       Dalc
##          0          0          0          0          0          0          0
##       Walc     health   absences         G1         G2         G3
##          0          0          0          0          0          0
```

```
# Rename Columns student performance to "easy to understand" names
# convert Column Headers to lower case
names(perf_math_df)[names(perf_math_df) == 'school'] <- "school_type"
names(perf_math_df)[names(perf_math_df) == 'famsize'] <- "family_size"
names(perf_math_df)[names(perf_math_df) == 'Pstatus'] <- "parents_cohabit"
names(perf_math_df)[names(perf_math_df) == 'Mjob'] <- "mother_s_profession"
names(perf_math_df)[names(perf_math_df) == 'Dcalc'] <- "weekday_alcohol"
names(perf_math_df)[names(perf_math_df) == 'Wcalc'] <- "weekend_alcohol"
names(perf_math_df)[names(perf_math_df) == 'reason'] <- "reason"
names(perf_math_df)[names(perf_math_df) == 'G1'] <- "grade_1"
names(perf_math_df)[names(perf_math_df) == 'G2'] <- "grade_2"
names(perf_math_df)[names(perf_math_df) == 'G3'] <- "grade_3"

# selecting student stress columns that we will focus for analysis
```

```
stress_df <-    subset(stu_stress_ds, select = c(anxiety_level,
                                                 self_esteem,
                                                 mental_health_history,
                                                 depression,
                                                 headache,
                                                 blood_pressure,
                                                 sleep_quality,
                                                 breathing_problem,
                                                 noise_level,
                                                 living_conditions,
                                                 safety,
                                                 basic_needs,
                                                 academic_performance,
                                                 study_load,
                                                 teacher_student_relationship,
                                                 future_career_concerns,
                                                 social_support,
                                                 peer_pressure,
                                                 extracurricular_activities,
                                                 bullying,
                                                 stress_level)
                    )

# Identify student stress data columns that have NULL (missing) values
colSums(is.na(stress_df)==TRUE|stress_df=='99'|stress_df=='')
```

```
##                anxiety_level                   self_esteem
##                            0                             0
##        mental_health_history                    depression
##                            0                             0
##                     headache                blood_pressure
##                            0                             0
##                sleep_quality             breathing_problem
##                            0                             0
##                  noise_level             living_conditions
##                            0                             0
##                       safety                   basic_needs
##                            0                             0
##         academic_performance                    study_load
##                            0                             0
## teacher_student_relationship       future_career_concerns
##                            0                             0
##               social_support                 peer_pressure
##                            0                             0
##   extracurricular_activities                      bullying
##                            0                             0
##                 stress_level
##                            0
```

```
# All the columns in Student Stress Dataset have easy to understand column
# headers and column headers are already in lower case
```

```r
# Cleaning up the column header info in Student Success dataset prior to
# loading data for analysis
stu_success_ds <- clean_names(stu_success_ds)

# selecting student stress columns that we will focus for analysis
academic_success_df <-    subset(stu_success_ds,
                                 select =
                                   c( "marital_status",
                                      "daytime_evening_attendance",
                                      "nacionality",
                                      "mother_s_qualification",
                                      "father_s_qualification",
                                      "displaced",
                                      "gender",
                                      "international",
                                      "age_at_enrollment",
                                      "curricular_units_1st_sem_enrolled",
                                      "curricular_units_1st_sem_evaluations",
                                      "curricular_units_1st_sem_approved",
                                      "curricular_units_1st_sem_grade",
                                      "curricular_units_2nd_sem_enrolled",
                                      "curricular_units_2nd_sem_evaluations",
                                      "curricular_units_2nd_sem_approved",
                                      "curricular_units_2nd_sem_grade",
                                      "unemployment_rate",
                                      "inflation_rate",
                                      "gdp",
                                      "target"
                                        )
                                 )

# Identify student stress data columns that have NULL (missing) values
colSums(is.na(academic_success_df)==TRUE|academic_success_df=='99'|academic_success_df=='')
```

```
##                       marital_status           daytime_evening_attendance
##                                    0                                    0
##                           nacionality               mother_s_qualification
##                                    0                                    0
##                 father_s_qualification                           displaced
##                                    0                                    0
##                                gender                        international
##                                    0                                    0
##                    age_at_enrollment     curricular_units_1st_sem_enrolled
##                                    0                                    0
##  curricular_units_1st_sem_evaluations     curricular_units_1st_sem_approved
##                                    0                                    0
##        curricular_units_1st_sem_grade     curricular_units_2nd_sem_enrolled
##                                    0                                    0
##  curricular_units_2nd_sem_evaluations     curricular_units_2nd_sem_approved
##                                    0                                    0
##        curricular_units_2nd_sem_grade                    unemployment_rate
##                                    0                                    0
##                        inflation_rate                                  gdp
```

```
##                                     0                                     0
##                                target
##                                     0
```

```
# Rename Columns student performance to "easy to understand" names
names(academic_success_df) [names(academic_success_df) == 'nacionality'] <- "nationality"
names(academic_success_df) [names(academic_success_df) == 'curricular_units_1st_sem_enrolled'] <- "sem_
names(academic_success_df) [names(academic_success_df) == 'curricular_units_1st_sem_evaluations'] <- "s
names(academic_success_df) [names(academic_success_df) == 'curricular_units_1st_sem_approved'] <- "sem_
names(academic_success_df) [names(academic_success_df) == 'curricular_units_1st_sem_grade'] <- "sem_1_u
names(academic_success_df) [names(academic_success_df) == 'curricular_units_2nd_sem_enrolled'] <- "sem_
names(academic_success_df) [names(academic_success_df) == 'curricular_units_2nd_sem_evaluations'] <- "s
names(academic_success_df) [names(academic_success_df) == 'curricular_units_2nd_sem_approved'] <- "sem_
names(academic_success_df) [names(academic_success_df) == 'curricular_units_2nd_sem_grade'] <- "sem_2_u
names(academic_success_df) [names(academic_success_df) == 'target'] <- "academic_success"
```

What does the final data set look like?

The final dataset is trimmed down to retain only key data columns that are required to study in focus. Though all the three datasets have related information, the target variable in each case is different, such as Performance, Stress Level and Academic Success. Also, the data is not gathered for same/similar set of students, these factors dictates the term for joining the dataset, which in this case is not suitable.

```
# Summary of Student Performance Dataset
summary(perf_math_df)
```

```
##  school_type            sex                 age        family_size
##  Length:395         Length:395         Min.   :15.0   Length:395
##  Class :character   Class :character   1st Qu.:16.0   Class :character
##  Mode  :character   Mode  :character   Median :17.0   Mode  :character
##                                        Mean   :16.7
##                                        3rd Qu.:18.0
##                                        Max.   :22.0
##  parents_cohabit    mother_s_profession    reason            traveltime
##  Length:395         Length:395          Length:395         Min.   :1.000
##  Class :character   Class :character    Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character    Mode  :character   Median :1.000
##                                                            Mean   :1.448
##                                                            3rd Qu.:2.000
##                                                            Max.   :4.000
##    studytime        failures        internet           romantic
##  Min.   :1.000   Min.   :0.0000   Length:395         Length:395
##  1st Qu.:1.000   1st Qu.:0.0000   Class :character   Class :character
##  Median :2.000   Median :0.0000   Mode  :character   Mode  :character
##  Mean   :2.035   Mean   :0.3342
##  3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :4.000   Max.   :3.0000
##      goout            Dalc            Walc           health
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000
##  Median :3.000   Median :1.000   Median :2.000   Median :4.000
##  Mean   :3.109   Mean   :1.481   Mean   :2.291   Mean   :3.554
```

```
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##     absences         grade_1         grade_2         grade_3
##  Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
##  Median : 4.000   Median :11.00   Median :11.00   Median :11.00
##  Mean   : 5.709   Mean   :10.91   Mean   :10.71   Mean   :10.42
##  3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00
```

```
# Summary of Student Academic Success Dataset
summary(stress_df)
```

```
##  anxiety_level    self_esteem    mental_health_history   depression
##  Min.   : 0.00   Min.   : 0.00   Min.   :0.0000        Min.   : 0.00
##  1st Qu.: 6.00   1st Qu.:11.00   1st Qu.:0.0000        1st Qu.: 6.00
##  Median :11.00   Median :19.00   Median :0.0000        Median :12.00
##  Mean   :11.06   Mean   :17.78   Mean   :0.4927        Mean   :12.56
##  3rd Qu.:16.00   3rd Qu.:26.00   3rd Qu.:1.0000        3rd Qu.:19.00
##  Max.   :21.00   Max.   :30.00   Max.   :1.0000        Max.   :27.00
##     headache     blood_pressure  sleep_quality  breathing_problem
##  Min.   :0.000   Min.   :1.000   Min.   :0.00   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:2.000
##  Median :3.000   Median :2.000   Median :2.50   Median :3.000
##  Mean   :2.508   Mean   :2.182   Mean   :2.66   Mean   :2.754
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:4.00   3rd Qu.:4.000
##  Max.   :5.000   Max.   :3.000   Max.   :5.00   Max.   :5.000
##   noise_level    living_conditions     safety       basic_needs
##  Min.   :0.000   Min.   :0.000     Min.   :0.000   Min.   :0.000
##  1st Qu.:2.000   1st Qu.:2.000     1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :2.000     Median :2.000   Median :3.000
##  Mean   :2.649   Mean   :2.518     Mean   :2.737   Mean   :2.773
##  3rd Qu.:3.000   3rd Qu.:3.000     3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000     Max.   :5.000   Max.   :5.000
##  academic_performance   study_load    teacher_student_relationship
##  Min.   :0.000        Min.   :0.000   Min.   :0.000
##  1st Qu.:2.000        1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000        Median :2.000   Median :2.000
##  Mean   :2.773        Mean   :2.622   Mean   :2.648
##  3rd Qu.:4.000        3rd Qu.:3.000   3rd Qu.:4.000
##  Max.   :5.000        Max.   :5.000   Max.   :5.000
##  future_career_concerns social_support  peer_pressure
##  Min.   :0.000          Min.   :0.000   Min.   :0.000
##  1st Qu.:1.000          1st Qu.:1.000   1st Qu.:2.000
##  Median :2.000          Median :2.000   Median :2.000
##  Mean   :2.649          Mean   :1.882   Mean   :2.735
##  3rd Qu.:4.000          3rd Qu.:3.000   3rd Qu.:4.000
##  Max.   :5.000          Max.   :3.000   Max.   :5.000
##  extracurricular_activities    bullying      stress_level
##  Min.   :0.000              Min.   :0.000   Min.   :0.0000
##  1st Qu.:2.000              1st Qu.:1.000   1st Qu.:0.0000
##  Median :2.500              Median :3.000   Median :1.0000
##  Mean   :2.767              Mean   :2.617   Mean   :0.9964
##  3rd Qu.:4.000              3rd Qu.:4.000   3rd Qu.:2.0000
```

```
##  Max.   :5.000              Max.   :5.000   Max.   :2.0000
```

```
# Summary of Student Academic Success Dataset
summary(academic_success_df)
```

```
##  marital_status  daytime_evening_attendance  nationality
##  Min.   :1.000   Min.   :0.0000             Min.   :  1.000
##  1st Qu.:1.000   1st Qu.:1.0000             1st Qu.:  1.000
##  Median :1.000   Median :1.0000             Median :  1.000
##  Mean   :1.179   Mean   :0.8908             Mean   :  1.873
##  3rd Qu.:1.000   3rd Qu.:1.0000             3rd Qu.:  1.000
##  Max.   :6.000   Max.   :1.0000             Max.   :109.000
##  mother_s_qualification father_s_qualification   displaced
##  Min.   : 1.00          Min.   : 1.00          Min.   :0.0000
##  1st Qu.: 2.00          1st Qu.: 3.00          1st Qu.:0.0000
##  Median :19.00          Median :19.00          Median :1.0000
##  Mean   :19.56          Mean   :22.28          Mean   :0.5484
##  3rd Qu.:37.00          3rd Qu.:37.00          3rd Qu.:1.0000
##  Max.   :44.00          Max.   :44.00          Max.   :1.0000
##      gender       international     age_at_enrollment sem_1_units_enrolled
##  Min.   :0.0000   Min.   :0.00000   Min.   :17.00     Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:19.00     1st Qu.: 5.000
##  Median :0.0000   Median :0.00000   Median :20.00     Median : 6.000
##  Mean   :0.3517   Mean   :0.02486   Mean   :23.27     Mean   : 6.271
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:25.00     3rd Qu.: 7.000
##  Max.   :1.0000   Max.   :1.00000   Max.   :70.00     Max.   :26.000
##  sem_1_units_evaluations sem_1_units_approved sem_1_units_grade
##  Min.   : 0.000          Min.   : 0.000       Min.   : 0.00
##  1st Qu.: 6.000          1st Qu.: 3.000       1st Qu.:11.00
##  Median : 8.000          Median : 5.000       Median :12.29
##  Mean   : 8.299          Mean   : 4.707       Mean   :10.64
##  3rd Qu.:10.000          3rd Qu.: 6.000       3rd Qu.:13.40
##  Max.   :45.000          Max.   :26.000       Max.   :18.88
##  sem_2_units_enrolled sem_2_units_evaluations sem_2_units_approved
##  Min.   : 0.000       Min.   : 0.000          Min.   : 0.000
##  1st Qu.: 5.000       1st Qu.: 6.000          1st Qu.: 2.000
##  Median : 6.000       Median : 8.000          Median : 5.000
##  Mean   : 6.232       Mean   : 8.063          Mean   : 4.436
##  3rd Qu.: 7.000       3rd Qu.:10.000          3rd Qu.: 6.000
##  Max.   :23.000       Max.   :33.000          Max.   :20.000
##  sem_2_units_grade unemployment_rate inflation_rate        gdp
##  Min.   : 0.00     Min.   : 7.60     Min.   :-0.800   Min.   :-4.060000
##  1st Qu.:10.75     1st Qu.: 9.40     1st Qu.: 0.300   1st Qu.:-1.700000
##  Median :12.20     Median :11.10     Median : 1.400   Median : 0.320000
##  Mean   :10.23     Mean   :11.57     Mean   : 1.228   Mean   : 0.001969
##  3rd Qu.:13.33     3rd Qu.:13.90     3rd Qu.: 2.600   3rd Qu.: 1.790000
##  Max.   :18.57     Max.   :16.20     Max.   : 3.700   Max.   : 3.510000
##  academic_success
##  Length:4424
##  Class :character
##  Mode  :character
##
##
##
```

Student Performance

Predictor Variable: Age, Sex, Family Size, Parents Cohabitation, Travel Time, Study Time, Failures, Internet Connection, Romantic Relationship, Week Day Alcohol Consumption, Weekend Alcohol Consumption, Absences

Response Variable: Student Final Grade (Grade 3)

Student Stress

Predictor Variable: Psychological Factors (Anxiety Level, Self Esteem, Mental Health, Depression), Physiological Factors (Headache, Blood Pressure, Sleep Quality, Breathing Problem), Environmental Factors (Noise Level, Living Conditions, Safety, Basic Needs), Academic Factors (Performance, Study Load, Teacher Student Relationship, Future Career Concerns), Social Factors (Social Support, Peer Pressure, Bullying, Extra Curricular Activities)

Response Variable: Student Stress Level

Student Academic Success

Predictor Variable: Marital Status, Age, Gender, International Student, Circular Units for Semester 1 and Semester 2 (Enrolled, Evaluated, Approved and grades), Unemployment Rate, Inflation Rate, GDP

Response Variable: Student Academic Success (Graduate / Drop-out)

What information is not self-evident?

In the Student Stress dataset, details about contributing factors such as Social, Environment, Psychological Factors are available but the information about student who took part in the survey and their background (age, gender, demography, etc.,) is not available. Student Performance Dataset and Student Academic Success Dataset are well rounded and do not appear to be missing any data elements, however, any dataset can be enhanced by adding qualitative information related to the subject of the study. For example: In the Student Performance dataset, we shall further enhance the dataset by adding year-wise performance of students and in the Student Academic Success Dataset, we shall add "Field of study that interest the student the most" field to validate if the student is pursuing the degree of his/her choice.

What are different ways you could look at this data?

Student Performance: I am planning to analyze the dataset to explore the effects of alcohol consumption in student's performance and validate there is a linear relationship (for example: the lesser the consumption the better the performance). I am planning to also look at other major contributing factors that can influence Student Performance such as Internet Connection, Romantic relationship, Absences, etc.,

Student Stress: I am planning to analyze the major contributing factors for Student Stress, such as comparing and contrasting the stress level in students that have poor sleep quality and higher level of depression. Similarly, I am planning to analyze the top most contributing factors within each category (social, environmental, psychological, etc,)

Student Academic Success: I am planning to explore the impact on academic success by external economic factors such as Unemployment Rate, GDP, Inflation and how academic performance metrics (credits enrolled, approved and grades) relates to the academic success and if there are any other personal characteristic groups (age, gender, international students) have an influence in academic success.

How do you plan to slice and dice the data?

I am planning to focus on core data elements(columns) that are most related to the subject of study and not retain unrelated data (for example: Application Number, Application Order columns are not most relevant to determine Student Academic Success and therefore have been omitted). I am also planning to apply further filtering using filter() function and grouping using group() function on dataset should the need arise to narrow the data and/or group the data at certain level to gain meaningful insights. I am also planning to use summary() function to reveal details about dataframes / models wherever applicable.

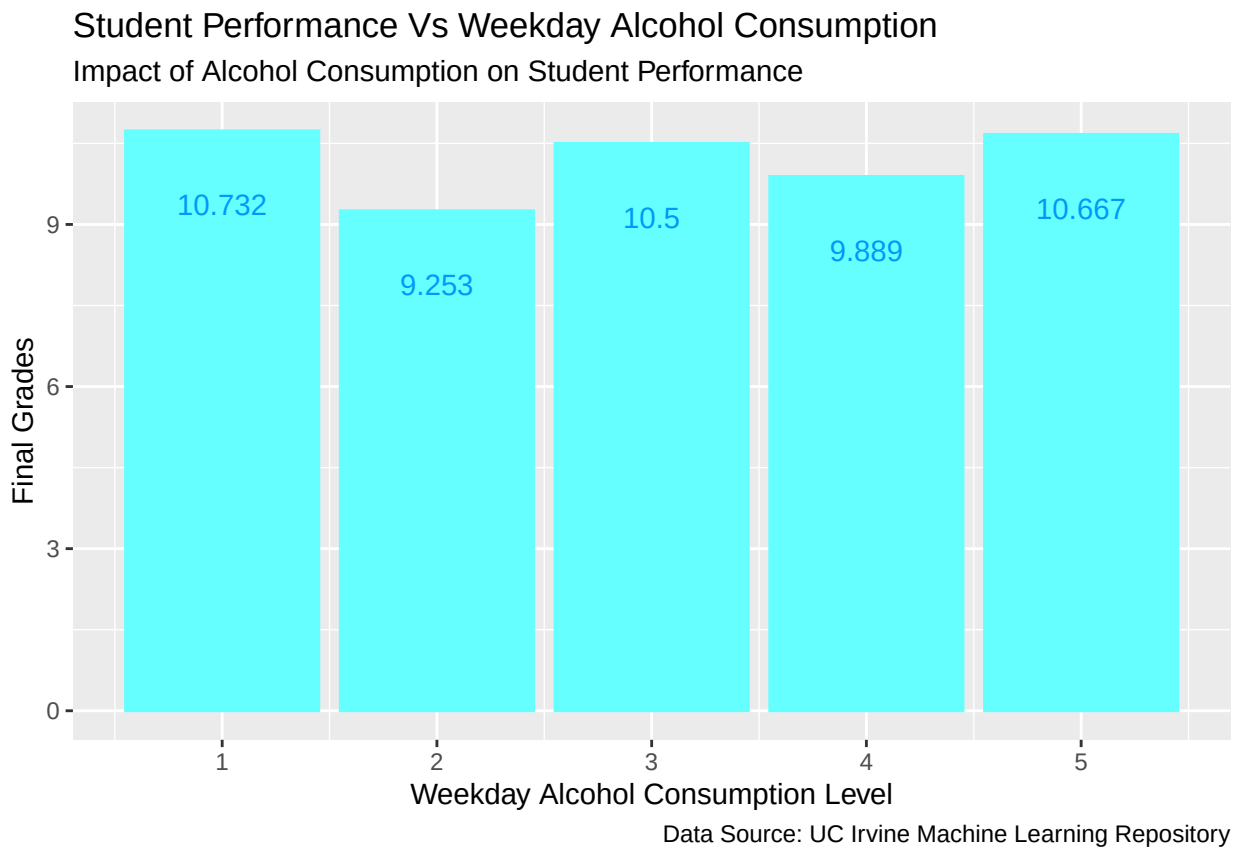
How could you summarize your data to answer key questions?

I am planning to summarize the data in two categories

1. ***Visualization:*** Plots, Bar Charts, Histograms will be created to show the relationship between predictor variables to the target variables such as Student Performance, Student Stress and Student Academic Success
2. ***Descriptive Statistics:*** Summary Statistics will be performed to show the central tendency to derive measures such as Mean, Mode and Median and to show Variance to derive measures Range, Variance, Standard Deviation, etc.,

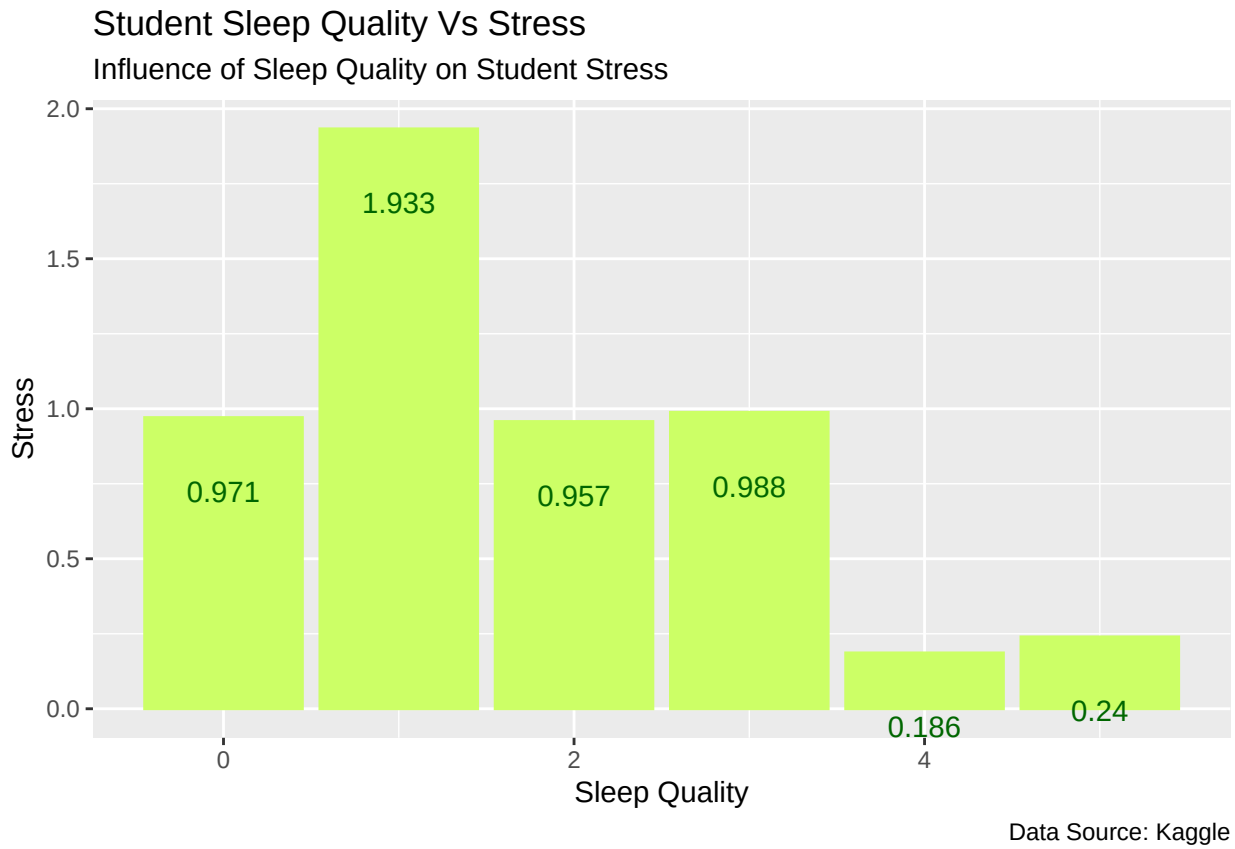
What types of plots and tables will help you to illustrate the findings to your questions?

```
# Bar chart to establish Impact of the Weekday Alcohol Consumption in Performance
dalc_perf <- perf_math_df

dalc_perf %>%
  group_by(Dalc) %>%
  summarise(mean_grade = mean(grade_3)) %>%
  ungroup() %>%
ggplot(aes(x = Dalc, y = mean_grade), length = 0.25) +
  # scale_x_discrete(expand=c(0,0)) +
 # scale_y_continuous(expand=c(0,0)) +
  geom_col(color = "#66FFFF", fill = "#66FFFF") +
   geom_text(aes(label = round(mean_grade, digits = 3)), vjust=4, size = 4, color = "#0099f9") +
  xlab("Weekday Alcohol Consumption Level") + ylab("Final Grades") +
labs(
  title = "Student Performance Vs Weekday Alcohol Consumption",
  subtitle = "Impact of Alcohol Consumption on Student Performance",
  caption = "Data Source: UC Irvine Machine Learning Repository"
)
```



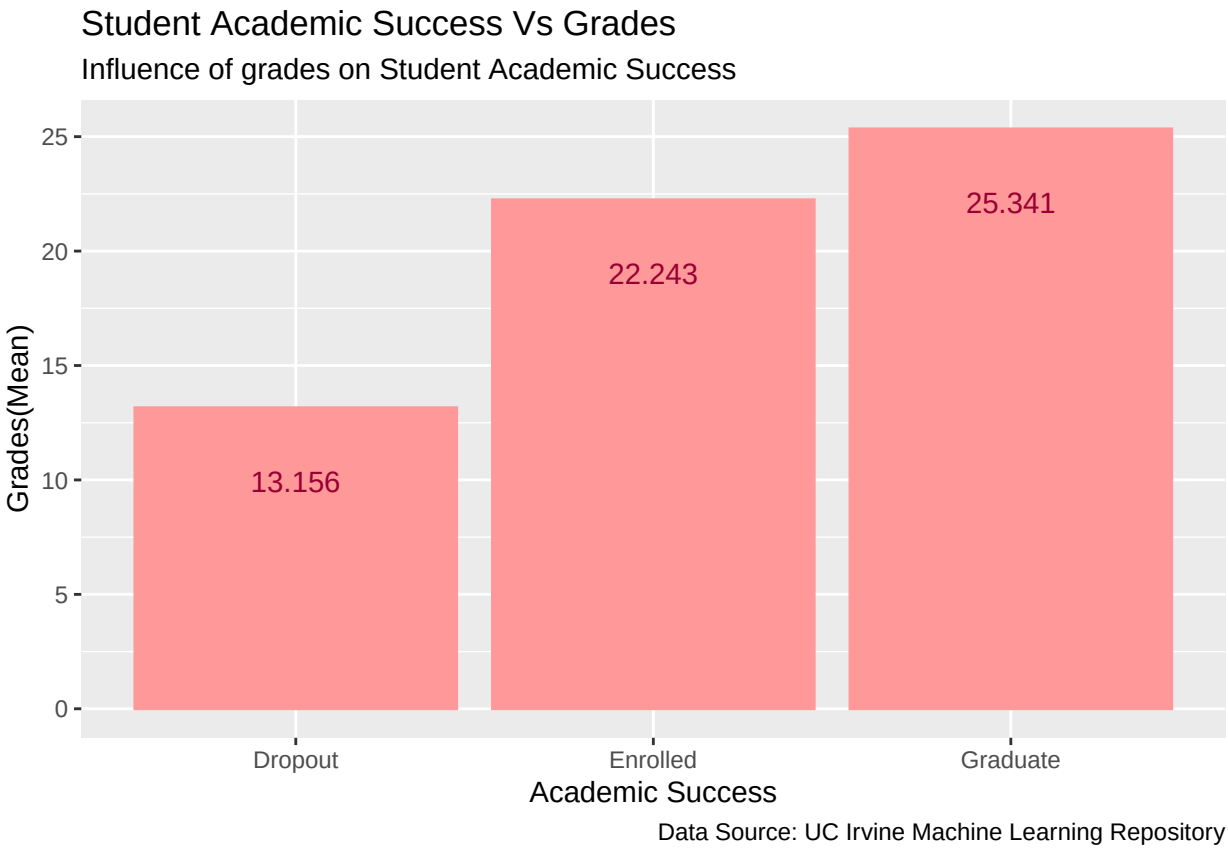
Observation: While it is clearly evident that the student with lowest level of Alcohol Consumption have scored highest among the students, the correlation between Alcohol Consumption Level on Weekday and Final Grade is not strong as per the bar chart

```
# Bar chart to establish Influence of the Sleep Quality in Student Stress

sleep_impact <- stress_df

sleep_impact %>%
  group_by(sleep_quality) %>%
  summarise(mean_stress = mean(stress_level)) %>%
  ungroup() %>%
  ggplot(aes(x = sleep_quality, y = mean_stress), length = 0.25) +
  # scale_x_discrete(expand=c(0,0)) +
  # scale_y_continuous(expand=c(0,0)) +
  geom_col(color = "#CCFF66", fill = "#CCFF66") +
  geom_text(aes(label = round(mean_stress, digits = 3)), vjust=4, size = 4, color = "#006600") +
  xlab("Sleep Quality") + ylab("Stress") +
  labs(
    title = "Student Sleep Quality Vs Stress",
    subtitle = "Influence of Sleep Quality on Student Stress",
    caption = "Data Source: Kaggle"
  )
```



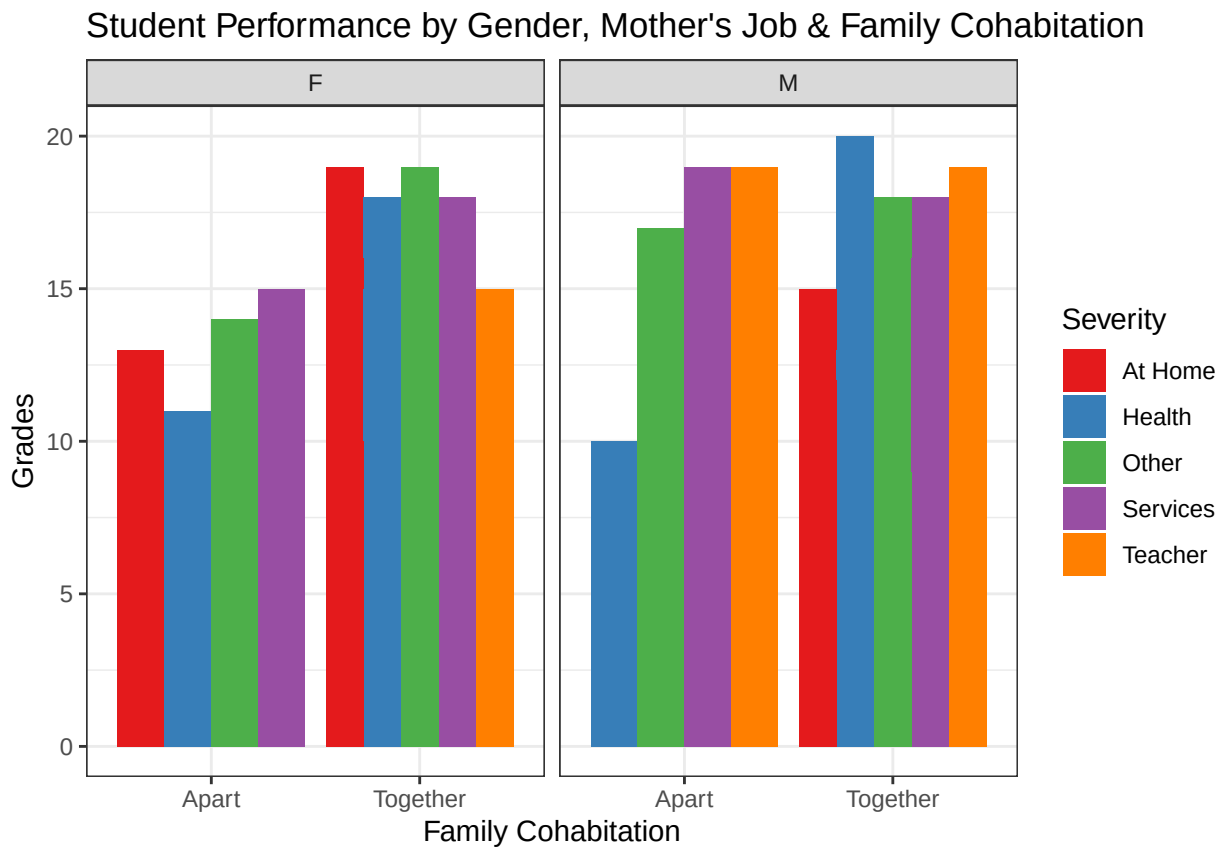
Observation: As per Bar Chart, it is clear that the student with the highest sleep quality experience less stress, but the relationship between sleep quality and the stress level is not linear (for example: students with sleep quality level 0 experience lower stress when compared to students with sleep quality level 1), this could be because of other contributing factors

```
# Bar chart to establish Impact of the Grades in Student Academic Success

grade_impact <- academic_success_df

grade_impact %>%
  group_by(academic_success) %>%
  summarise(mean_grades = mean(sem_1_units_grade+sem_2_units_grade)) %>%
  ungroup() %>%
  ggplot(aes(x = academic_success, y = mean_grades), width = 0.25) +
  # scale_x_discrete(expand=c(0,0)) +
  # scale_y_continuous(expand=c(0,0)) +
   geom_col(color = "#FF9999", fill = "#FF9999") +
   geom_text(aes(label = round(mean_grades, digits = 3)), vjust=4, size = 4, color = "#990033") +
  xlab("Academic Success") + ylab("Grades(Mean)") +
  labs(
    title = "Student Academic Success Vs Grades",
    subtitle = "Influence of grades on Student Academic Success",
    caption = "Data Source: UC Irvine Machine Learning Repository"
  )
```



Observation: As per Bar Chart, it is clear that the student with highest grades have highest chances of success. The relationship between Grades and Academic Success have strong correlation. The higher the Grades the better the Academic Success

```
# Multiple Category Bar Chart to analyze the role of mother and family
# Cohabitation in Student Performance

ggplot(perf_math_df, aes(parents_cohabit, grade_3)) +
  geom_col(aes(fill = mother_s_profession), position = "dodge") +
theme_bw() +
  facet_wrap(~sex) +
  labs(x = "Family Cohabitation", y = "Grades", fill = "Severity") +
  ggtitle("Student Performance by Gender, Mother's Job & Family Cohabitation") +
  scale_x_discrete(labels = c("Apart", "Together")) +
  scale_fill_manual(values = c("at_home" = "#E41A1C", "health" = "#377EB8",
                               "other" = "#4DAF4A", "services" = "#984EA3",
                               "teacher" = "#FF7F00"),
                    labels = c("At Home", "Health",
                               "Other", "Services","Teacher"))
```

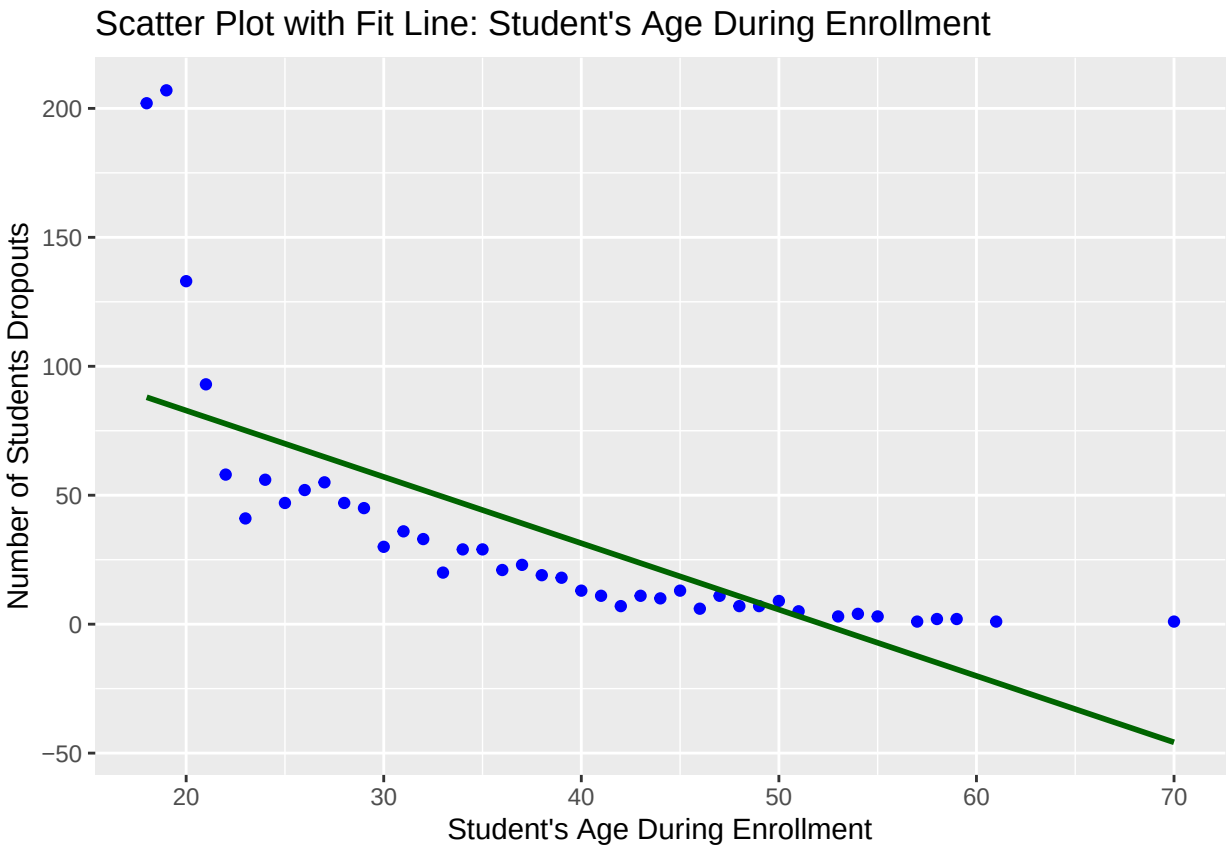


Observation: As per multiple category bar chart, there is evident of moderate correlation between genders and student performance when measured against family cohabitation and mother's profession. Categories of Mother's profession plays a similar role across genders when the family cohabit and when they are not.

It is important to note that there is no representation of certain groups For example: There is no data for families that do not cohabit and have female students with mother's profession is teacher.

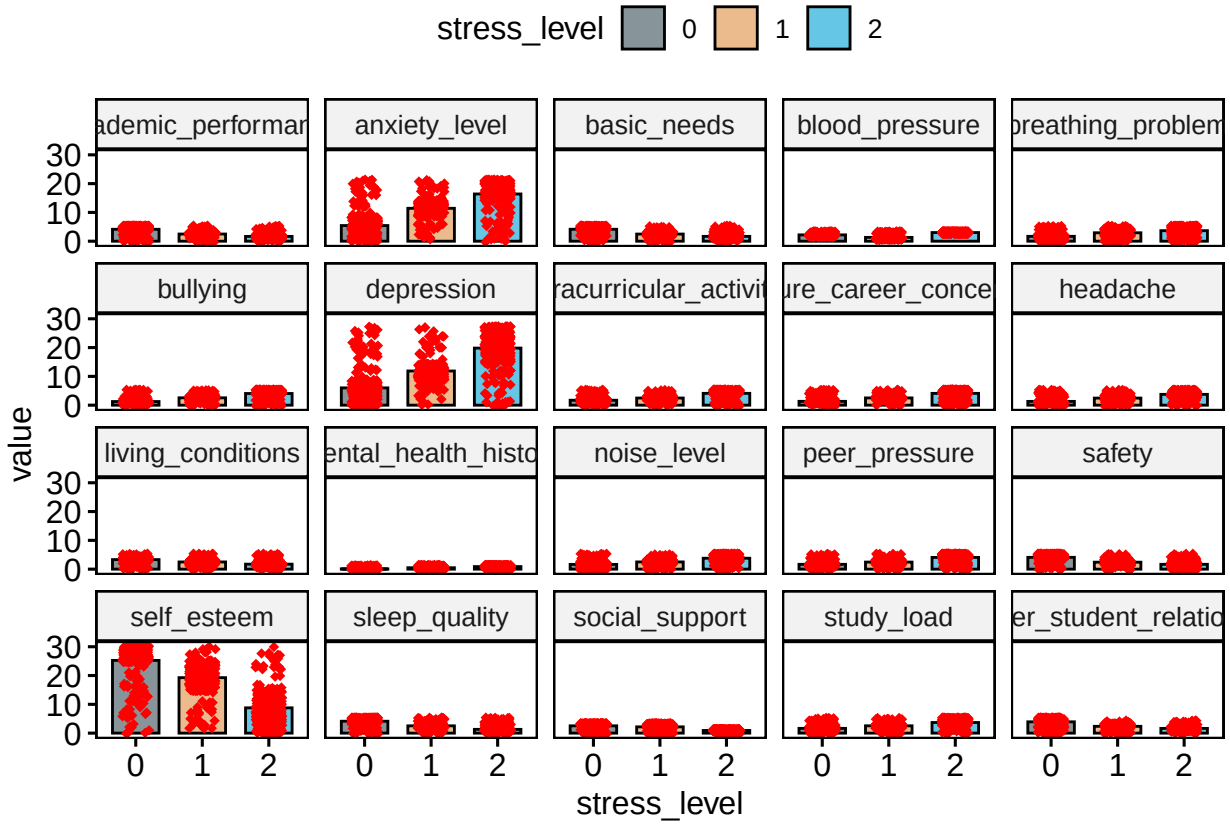
```
academic_success_df  %>%
  filter(academic_success == "Dropout") %>%
 group_by(age_at_enrollment)  %>%
  summarise(Count = n(), .groups = "drop") %>%
  ggplot(aes(x = age_at_enrollment, y = Count)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "darkgreen") +
  labs(x = "Student's Age During Enrollment", y = "Number of Students Dropouts") +
  ggtitle("Scatter Plot with Fit Line: Student's Age During Enrollment")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Observation: As per the scatter plot, it is evident that students who are younger are the highest number of students who dropout and students who are at age 50 or more, the "Dropout" rate is nearly zero

```
# Compare and Contrast the contributing factors to Stress Level

stress_df$stress_level <- as.factor(stress_df$stress_level)

# Plot the bar chart
stress_df%>%
  pivot_longer(-stress_level) %>%
  ggbarplot(x = "stress_level",
            y = "value",
            facet.by = "name",
            add = c("mean_se", "jitter"),
            add.params = list(color = "red",shape = "diamond"),
            color = "black",
            fill = "stress_level",
            alpha = .6,
            palette = "jama",
            position = position_dodge(.5)) +
  theme(legend.position="top")
```



Observation: ggbarplot helps compare all the contributing factors to Student Stress at the same time. It is evident that contributing factors such as depression, self esteem have linear relationship (example: the lower the depression, the lower the stress, similarly the higher the self esteem, the lower the stress is linear though inverse). It is surprising to see the week correlation between factors like bullying, headache, safety to stress level

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I am planning to use machine learning technique by developing one or more predictive model to illustrate the impact/influence of major contributing factors for Student Performance, Stress and Academic Success. Considering the target variable of the datasets they are categorical (ordinal in two cases), Logistic Regression Model appears to be a suitable method for this research. I am planning to estimate the probability of each factor (for example: Mother's Profession in Student Performance, Sleep Quality in Student Stress, Semester 1 Grade in Academic Success, etc.,) and understand the goodness of the fit and the variability.

Questions for Future Steps

1. The optimal model for the Student Performance Analysis, Student Stress Analysis and Student Academic Success is yet to be determined. Though Logistic Regression Model appear to be best fit, it will be known only after fitting the model to the datasets.

Milestone 3

INTRODUCTION

The primary objective of this project is to employ the data science methods to analyze the major factors that contributes to Student Performance, Student Stress and Student Academic Success. The student community are the most important part of human society. A successful society is one that acknowledges and nurture its student community. By rigorously analyzing the datasets related to student performance, stress and academic success, we can get valuable insights into major factors that contributes to student stress and make improvements to enhance student performance and subsequently their academic success.

The problem statement you addressed

The problem that I have attempted to address includes

• Identify the existence and strength of correlation between student performance and the factors influencing performance such as Alcohol Consumption, Family Background and romantic relationships

• Establish if there is correlation between Student Stress and external factors such as social factors, physiological factors

• Identify the correlation presence and strength between Student Grades and Student Academic Success

How you addressed this problem statement

- Plotted the Student Performance data in bar chart to understand the influence of weekday alcohol consumption in student performance
- Plotted the Student Stress Data in bar chart to understand the impact of sleep quality in stress level in students
- Plotted the Student Academic Success data to understand the influence of Grades in Academic success
- Compared the Family Cohabitation Status (Parents Living Together or Apart) influence the Student Performance between Male and Female Students
- Plotted the Student Age information across the line on a Scatter Plot to understand students under a specific age range are more likely to Dropout
- Plotted all the contributing reasons across external factors (travel time, extracurricular activities) social factors (such as internet access, family relationship, romantic relationship) behavioral factors (such as alcohol consumption, absences, free time) to the Stress Level on a side-by-side bar chart to compare and contrast the impact/influence of various factors towards Stress Level
- Create Logistic Regression models to gauge the correlation between contributing factors(Predictor variables) and target variables(Performance/Stress/Academic Success)

Logistic Regression Model

```
# The binary variable "high_stress" built based on stress_level column is set
# to 1 if the stress level is 2, otherwise it is set to 0

stress_df$high_stress <- ifelse(as.numeric(as.character(stress_df$stress_level)) >= 2, 1, 0)

model_df <- subset(stress_df, select=
                     c(anxiety_level,self_esteem,mental_health_history,depression,
                       headache,blood_pressure,sleep_quality,breathing_problem,
```

```
                      noise_level,living_conditions,safety,basic_needs,
                      academic_performance,study_load,teacher_student_relationship,future_career_concer
                      social_support,peer_pressure,extracurricular_activities,bullying,
                      high_stress))

set.seed(123)
train_indices <- sample(1:nrow(model_df), 0.8 * nrow(model_df))
train_data <- model_df[train_indices, ]
test_data <- model_df[-train_indices, ]
#Create a logistic regression model
logistic_model <- glm(high_stress ~ ., data = train_data, family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
#Summary of the logistic regression model
summary(logistic_model)
```

```
##
## Call:
## glm(formula = high_stress ~ ., family = "binomial", data = train_data)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -5.890e+01  2.740e+03  -0.021  0.98285
## anxiety_level               -4.288e-05  2.961e-02  -0.001  0.99884
## self_esteem                 -5.778e-02  2.239e-02  -2.581  0.00984 **
## mental_health_history        3.206e-01  3.951e-01   0.811  0.41718
## depression                   1.762e-02  2.337e-02   0.754  0.45081
## headache                     2.705e-01  1.156e-01   2.340  0.01927 *
## blood_pressure               1.866e+01  9.132e+02   0.020  0.98370
## sleep_quality               -2.332e-01  1.092e-01  -2.136  0.03271 *
## breathing_problem            1.200e-01  1.069e-01   1.123  0.26153
## noise_level                  3.243e-01  1.069e-01   3.034  0.00242 **
## living_conditions           -1.393e-02  1.156e-01  -0.120  0.90411
## safety                      -8.342e-02  1.190e-01  -0.701  0.48334
## basic_needs                 -8.408e-02  1.154e-01  -0.729  0.46620
## academic_performance        -1.623e-01  1.137e-01  -1.427  0.15349
## study_load                   1.232e-01  1.057e-01   1.166  0.24362
## teacher_student_relationship  2.969e-01  1.433e-01   2.073  0.03820 *
## future_career_concerns       8.944e-02  1.106e-01   0.809  0.41879
## social_support               8.168e-01  3.845e-01   2.125  0.03362 *
## peer_pressure                6.378e-02  1.160e-01   0.550  0.58235
## extracurricular_activities   1.493e-01  1.087e-01   1.373  0.16961
## bullying                     2.377e-01  1.095e-01   2.170  0.02998 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1133.2  on 879  degrees of freedom
## Residual deviance:  234.6  on 859  degrees of freedom
## AIC: 276.6
##
```

```
## Number of Fisher Scoring iterations: 20
```

Observation:

1. Based on the Summary of the model, it is clearly evident that "Noise Level" has maximum z value and lowest P value and therefore is the highest contributor to the Stress level
2. The coefficient for "Sleep Quality", "-2.332e-01", suggests that sleep quality id associated with a decrease in the log-odds of the target variable, compared other contributing factors for stress level

```
# The binary variable "high_performance" built based on stress_level column is set
# to 1 if the Final Grade(grade_3) is greater than or equal to 15, otherwise
# it is set to 0

perf_math_df$high_performance <- ifelse(perf_math_df$grade_3 >= 15, 1, 0)

pf_model_df <- subset(perf_math_df, select=
                      c(family_size,parents_cohabit,mother_s_profession,reason,
                        traveltime,studytime,failures,
                        internet,romantic,goout,
                        Dalc,Walc,health,absences,
                        high_performance))

set.seed(123)
train_indices <- sample(1:nrow(pf_model_df), 0.8 * nrow(pf_model_df))
train_data <- pf_model_df[train_indices, ]
test_data <- pf_model_df[-train_indices, ]
#Create a logistic regression model
logistic_model <- glm(high_performance ~ ., data = train_data, family = "binomial")
#Summary of the logistic regression model
summary(logistic_model)
```

```
##
## Call:
## glm(formula = high_performance ~ ., family = "binomial", data = train_data)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.64834    1.20865  -0.536   0.5917
## family_sizeLE3              0.22798    0.35322   0.645   0.5187
## parents_cohabitT           -0.41470    0.55836  -0.743   0.4577
## mother_s_professionhealth   1.36436    0.72678   1.877   0.0605 .
## mother_s_professionother    0.17256    0.64980   0.266   0.7906
## mother_s_professionservices 1.26710    0.63543   1.994   0.0461 *
## mother_s_professionteacher  0.77339    0.67646   1.143   0.2529
## reasonhome                 -0.08533    0.42163  -0.202   0.8396
## reasonother                 0.21580    0.58668   0.368   0.7130
## reasonreputation           -0.27035    0.42888  -0.630   0.5285
## traveltime                 -0.18915    0.29841  -0.634   0.5262
## studytime                   0.05256    0.20326   0.259   0.7960
## failures                   -1.32621    0.57595  -2.303   0.0213 *
## internetyes                 0.75026    0.59803   1.255   0.2096
## romanticyes                -0.36829    0.37282  -0.988   0.3232
## goout                       0.04779    0.16641   0.287   0.7740
```

```
## Dalc                         -0.50966    0.33795  -1.508   0.1315
## Walc                         -0.03080    0.19396  -0.159   0.8738
## health                       -0.12482    0.11482  -1.087   0.2770
## absences                     -0.07781    0.03639  -2.138   0.0325 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 301.29  on 315  degrees of freedom
## Residual deviance: 250.50  on 296  degrees of freedom
## AIC: 290.5
##
## Number of Fisher Scoring iterations: 6
```

Observation:

1. Based on the Summary of the model, it is clearly evident that mother's profession, especially, "Health" has maximum z value and one of the lowest P value and therefore is the highest contributor to the Stress level
2. The coefficient for the variable "Failures", "-1.32621", suggests that variable failure is associated with a decrease in the log-odds of Student Performance, compared to other contributing factors
3. Mother's Profession: The coefficients for different values in Mother's Profession represent the effects of those conditions on Student Performance. For example, if Mother's Profession is "Services"Other" that has a coefficient of 0.17256, indicating a small increase in the log-odds of the response variable compared to other Professions of Mothers

```r
# The binary variable "academic_result" built based on stress_level column is set
# to 1 if the Academic Success(Target) is "Graduated", otherwise it is set to 0

academic_success_df$academic_result <- ifelse(academic_success_df$academic_success == 'Graduate', 1, 0)

pf_model_df <- subset(academic_success_df, select=
                        c(daytime_evening_attendance,nationality,
                          mother_s_qualification,father_s_qualification,
                          displaced,gender,international,age_at_enrollment,
                          sem_1_units_grade,sem_2_units_grade,
                          unemployment_rate,inflation_rate,gdp,
                          academic_result))

set.seed(123)
train_indices <- sample(1:nrow(pf_model_df), 0.8 * nrow(pf_model_df))
train_data <- pf_model_df[train_indices, ]
test_data <- pf_model_df[-train_indices, ]
#Create a logistic regression model
logistic_model <- glm(academic_result ~ ., data = train_data, family = "binomial")
#Summary of the logistic regression model
summary(logistic_model)
```

```
##
## Call:
## glm(formula = academic_result ~ ., family = "binomial", data = train_data)
##
```

```
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.168820   0.356944  -6.076 1.23e-09 ***
## daytime_evening_attendance -0.294962   0.148368  -1.988  0.04681 *
## nationality                 -0.012892   0.009756  -1.321  0.18636
## mother_s_qualification       0.001447   0.003050   0.474  0.63525
## father_s_qualification       0.003935   0.003022   1.302  0.19284
## displaced                    0.155755   0.085393   1.824  0.06815 .
## gender                      -0.580788   0.084171  -6.900 5.20e-12 ***
## international                0.339375   0.421774   0.805  0.42103
## age_at_enrollment           -0.043313   0.006628  -6.535 6.35e-11 ***
## sem_1_units_grade            0.034851   0.019907   1.751  0.08000 .
## sem_2_units_grade            0.228809   0.019440  11.770  < 2e-16 ***
## unemployment_rate            0.045689   0.015896   2.874  0.00405 **
## inflation_rate               0.002797   0.028504   0.098  0.92183
## gdp                          0.020036   0.018518   1.082  0.27928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4905.9  on 3538  degrees of freedom
## Residual deviance: 3863.8  on 3525  degrees of freedom
## AIC: 3891.8
##
## Number of Fisher Scoring iterations: 5
```

Observation:

1. Based on the Summary of the model, it is clearly evident that
   semester 2 grade has maximum z value and therefore is the highest contributor to the Academic Success
2. The coefficient for "gender", -0.580788 is the lowest. It suggests that "gender" variable is associated with a decrease in the log-odds of Student Academic Success, compared to other contributing factors

Summary of Analysis

• As per the Student Performance Analysis, the students with lowest level of Alcohol Consumption have scored highest among the students, though the correlation between Alcohol Consumption Level on Weekday and Final Grade is not strong

• As per the Student Stress Analysis, the he student with the highest sleep quality experience less stress, but the relationship between sleep quality and the stress level is not linear

• As per the Student Academic Success Analysis, the he students with highest grades have highest chances of success. The relationship between Grades and Academic Success have strong correlation. The higher the Grades the better the Academic Success

• As per multiple category bar chart, there is evident of moderate correlation between genders and student performance when measured against family cohabitation and mother's profession. Categories of Mother's profession plays a similar role across genders when the family cohabit and when they are not cohabiting

• As per the scatter plot, it is evident that students who are younger are the highest number of students who dropout and students who are at age 50 or more, the "Dropout" rate is nearly zero

• With the help of ggbarplot, we were able to compare all the contributing factors to Student Stress at the same time. It is evident that contributing factors such as depression, self-esteem have linear relationship, it is also important to note the week correlation between factors like bullying, headache, safety to stress level

• As per the Logistic Regression model built for the Student Stress Analysis

• "Noise Level" has maximum z value and lowest P value and therefore is the highest contributor to the Stress level

• The coefficient for "Sleep Quality", " -2.332e-01", suggests that sleep quality id associated with a decrease in the log-odds of the target variable, compared other contributing factors for stress level

• As per the Logistic Regression model built for the Student Performance Analysis

• Mother's profession, especially, "Health" has maximum z value and one of the lowest P-value and therefore is the highest contributor to the Stress level

• The coefficient for the variable "Failures", "-1.32621", suggests that variable failure is associated with a decrease in the log-odds of Student Performance, compared to other contributing factors

• As per the Logistic Regression model built for the Student Academic Success Analysis

• Semester 2 grade has maximum z value and therefore is the highest contributor to the Academic Success

• The coefficient for "gender", -0.580788 is the lowest. It suggests that "Gender" variable is associated with a decrease in the log-odds of Student Academic Success, compared to other contributing factors

Implications

The research should be conducted on larger dataset, if possible, Student Performance, Student Stress and Student Academic Success survey should be conducted on same set of students to cross join the datasets and directly map the Performance data to Stress and subsequently Student Academic success. Result of study from Stress Data should be thoroughly be thoroughly understood to help reduce the stress level in students. Similarly, factors contributing to Student dropout should be reduces or eliminated which factors contributing to academic success should be boosted.

Limitations

The datasets are individual datasets, while it helps understand the contributing factors for Performance, Stress, and Academic Success, they are not from same subjects (in this case, students). The datasets are collected from specific geographies though the datasets have data about international students. This will limit the results of the studies to specific to certain demography and cannot be applied to global students' community. While the dataset has captured the most critical aspects of student's profile, there are missing data for specific scenarios (for example: There are no female students representing the group where Mother's profession is Teacher and the family is apart) that limits the results to be applied across gender.

Concluding Remarks

To summarize, research related to performance, stress and academic results is a very important for the society, enabling the students to be a value asset to the community and help them achieve greater success in their careers and in life. Bar Charts, Scatter Plots, comparison charts help understanding the pattern leading to stress level, which can be used to identify and mitigate the issues and reduce the overall stress level of students which will enhance their performance, eventually guiding them towards Academic Success. Logistic Regression Models are built to analyze the contributing factors which leads to academic success of students and imitating the same on student who are showing signs of struggle will boost their chances of success.