Project Proposal

Due November 17 at 11:59pm

Ramil Mammadov(rm564),

Load Packages

Dataset 1 Student Performance

Data source: UC Irvine Machine Learning Repository (November 26, 2014), Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.

Brief description: This dataset examines student achievement in secondary education at two Portuguese schools. Its attributes include student grades, demographic, social, and school-related features, and it was collected using school reports and questionnaires. The data specifically focuses on performance in Portuguese language classes. In [Cortez and Silva, 2008], the data set was modeled under binary/five-level classification and regression tasks.

Observations: The dataset contains 649 observations and 33 columns. Each row represents a unique student and captures a range of demographic, familial, academic, and personal attributes, including demographics, family background, academic details, and personal and social aspects. These attributes provide insights into each student's background and academic progress, making the dataset useful for analyzing factors that influence academic performance.

Research question 1: How does the amount of study time relate to final grade performance (Grade_3), and does it vary based on family support?

- **Outcome Variable**: Grade 3 (Continuous)
- **Primary Independent Variables**: Study time (continuous) and Family support (nominal, with interaction)

Research question 2: Is there an association between family relationship quality and school absences?

- **Outcome Variable**: school absences (ordinal)
- **Primary Independent Variable**: family relationship (ordinal)

Load the data and provide a `glimpse()`:

```
Rows: 649 Columns: 33
-- Column specification --------------------------------------------------------
Delimiter: ";"
chr (17): school, sex, address, family_size, Parent's_status, Mother's_job, ...
dbl (16): age, Mother's_education, Father's_education, travel_time, study_ti...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Rows: 649
Columns: 33
$ school               <chr> "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "~
$ sex                  <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M",~
$ age                  <dbl> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 1~
$ address              <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U",~
$ family_size          <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", ~
$ `Parent's_status`    <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T",~
$ `Mother's_education` <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4~
$ `Father's_education` <dbl> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4~
$ `Mother's_job`       <chr> "at_home", "at_home", "at_home", "health", "other~
$ `Father's_job`       <chr> "teacher", "other", "other", "services", "other",~
$ reason               <chr> "course", "course", "other", "home", "home", "rep~
$ guardian             <chr> "mother", "father", "mother", "mother", "father",~
$ travel_time          <dbl> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1~
$ study_time           <dbl> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3~
$ failures             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ school_support       <chr> "yes", "no", "yes", "no", "no", "no", "no", "yes"~
$ family_support       <chr> "no", "yes", "no", "yes", "yes", "yes", "no", "ye~
$ extra_paid_classes   <chr> "no", "no", "no", "no", "no", "no", "no", "no", "~
$ activities           <chr> "no", "no", "no", "yes", "no", "yes", "no", "no",~
$ nursery_school       <chr> "yes", "no", "yes", "yes", "yes", "yes", "yes", "~
$ higher_school        <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", ~
$ internet_access      <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "n~
$ romantic             <chr> "no", "no", "no", "yes", "no", "no", "no", "no", ~
$ family_relationship  <dbl> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3~
$ free_time            <dbl> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2~
$ go_out               <dbl> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3~
```

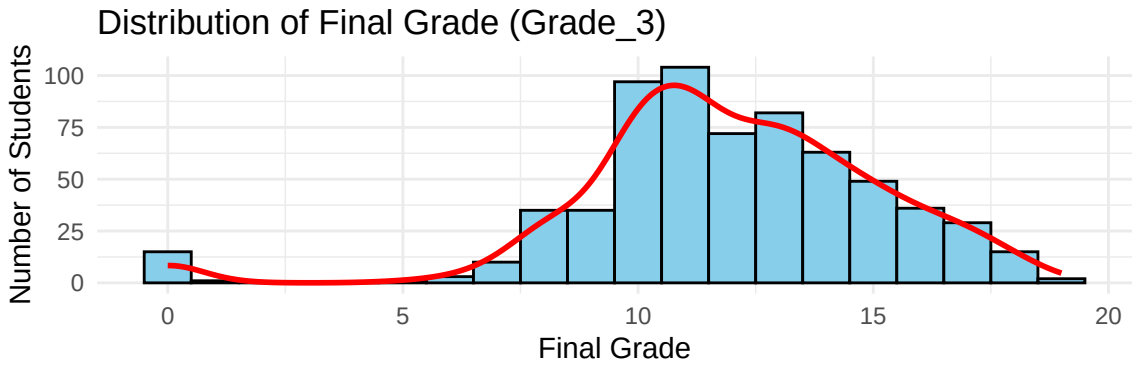
```
$ workday_alcohol      <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ weekend_alcohol      <dbl> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2~
$ healthy_status       <dbl> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2~
$ school_absences      <dbl> 4, 2, 6, 0, 0, 6, 0, 2, 0, 0, 2, 0, 0, 0, 0, 6, 1~
$ Grade_1              <dbl> 0, 9, 12, 14, 11, 12, 13, 10, 15, 12, 14, 10, 12,~
$ Grade_2              <dbl> 11, 11, 13, 14, 13, 12, 12, 13, 16, 12, 14, 12, 1~
$ Grade_3              <dbl> 11, 11, 12, 14, 13, 13, 13, 13, 17, 13, 14, 13, 1~
```

Exploratory Data Analysis (EDA)

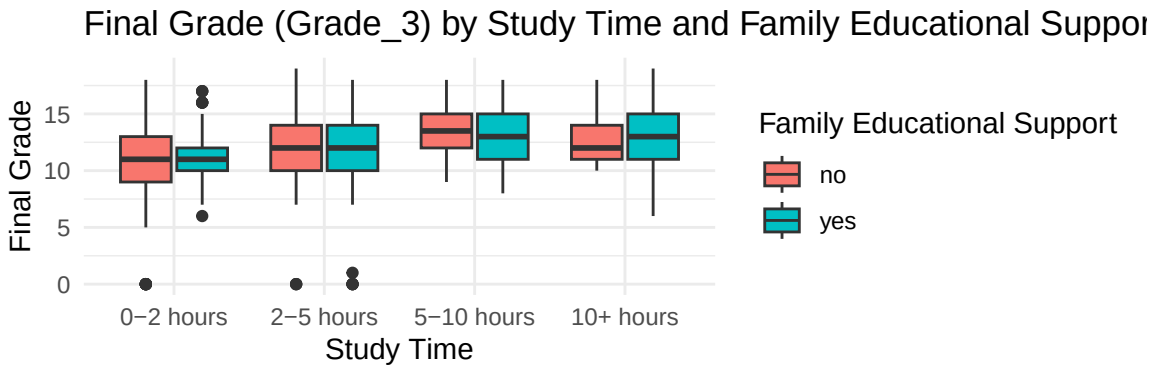
Question 1 EDA

Exploratory Plots:

1) Grade Distribution: The distribution of `Grade_3` shows a range of final grades, suggesting variations in student performance.

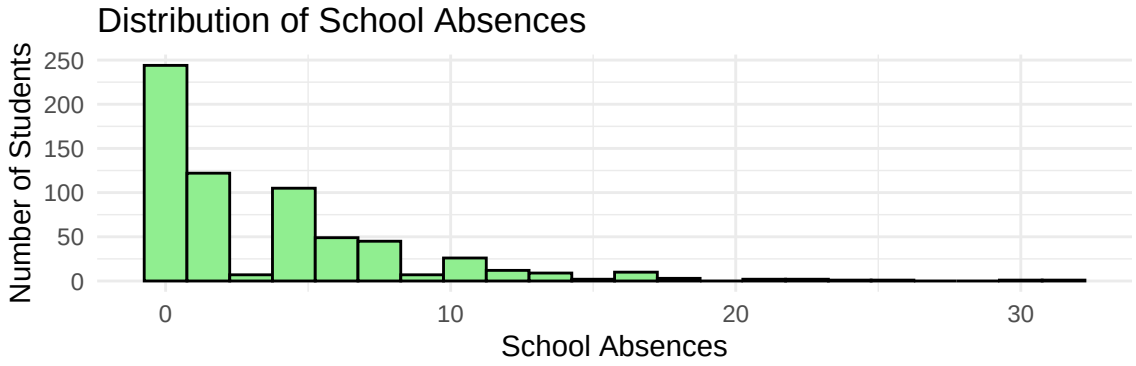


2) Grade and Study Time with Family Support: The box plot highlights the interaction between study time and family support, suggesting that the positive impact of family support is most evident when students already study for at least 5 hours. This implies that family support is effective when it supplements a student's efforts. In the lowest study time group (0-2 hours), grades are generally low and variable, regardless of family support, indicating that without a minimum level of study time, family support alone is insufficient to drive high academic achievement.

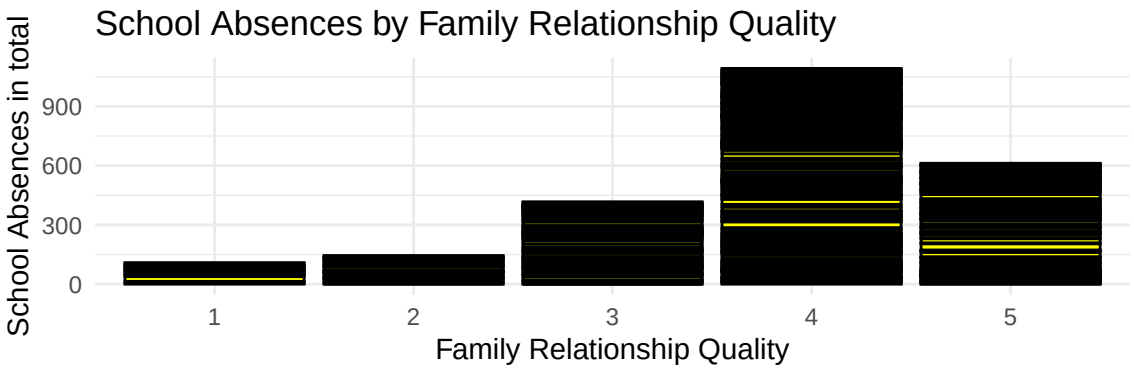


Question 2 EDA

1) School Absences: The distribution of `school_absences` is right-skewed, indicating that most students have relatively few absences.



2) School absences and Family Relationship: The bar chart demonstrates The relationship between school absences and family relationship school absences and family relationship. Family relationship quality impacts school attendance in nuanced ways. Strong family relationships (category 5) may help maintain school attendance, while moderate relationships (category 4) seem associated with higher absences, possibly due to emotional or logistical challenges at home. Another interesting factor is that according to the data students who rate their family relationships as very poor (category 1 and 2) have the fewest absences too. It's possible that these students may not feel comfortable staying home or may lack the family support to address personal issues by staying out of school.


School Absences by Family Relationship Quality
School Absences in total
900
600
300
0
1
2
3
4
5
Family Relationship Quality