High School Student Performance Project

Exploration

Objective(s)

We would like to do several explorations on students' grades and willingness for higher education.

1. How does the distribution of grades look like for each period?

2. Does educational support help students with grades?

3. Does guardian's education status and job affect the student's willingness for higher education?

Data collection and cleaning

Have an initial draft of your data cleaning appendix. Document every step that takes your raw data file(s) and turns it into the analysis-ready data set that you would submit with your final project. Include text narrative describing your data collection (downloading, scraping, surveys, etc) and any additional data curation/cleaning (merging data frames, filtering, transformations of variables, etc). Include code for data curation/cleaning, but not collection.

```r
#import libraries
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(tidyr)
library(ggplot2)
library(forcats)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v lubridate 1.9.2     v stringr   1.5.0
v purrr     1.0.2     v tibble    3.2.1
v readr     2.1.4     
```

```
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(viridis)
```

```
Loading required package: viridisLite
```

```
library(scales)
```

```

Attaching package: 'scales'

The following object is masked from 'package:viridis':

    viridis_pal

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```
library(patchwork)
```

```
#import data
mat_data <- read.csv("data/student-mat.csv", sep = ";")
por_data <- read.csv("data/student-por.csv", sep = ";")
```

```
#data joining
#add major to both dataset
mat_data <- mat_data |>
  mutate(major = "Math")
por_data <- por_data |>
  mutate(major = "Portuguese")
rm(student_data)
```

```
Warning in rm(student_data): object 'student_data' not found
```

```
# Binding the rows of the two datasets together
merged_data <- bind_rows(mat_data, por_data)
```

```
#data cleaning
student_data <- merged_data |>
#rename columns for readability
  rename(
    student_school = school,
    student_sex = sex,
    student_age = age,
    home_address_type = address,
    family_size = famsize,
    parent_cohabitation_status = Pstatus,
    mother_education = Medu,
    father_education = Fedu,
    mother_job = Mjob,
    father_job = Fjob,
    school_choice_reason = reason,
    student_guardian = guardian,
    travel_time = traveltime,
    weekly_study_time = studytime,
    past_class_failures = failures,
    extra_educational_support = schoolsup,
    family_educational_support = famsup,
    paid_extra_classes = paid,
    extra_curricular_activities = activities,
    attended_nursery_school = nursery,
    higher_education_aspiration = higher,
    internet_access_at_home = internet,
    romantic_relationship = romantic,
    family_relationship_quality = famrel,
    free_time_after_school = freetime,
    going_out_with_friends = goout,
    workday_alcohol_consumption = Dalc,
    weekend_alcohol_consumption = Walc,
    current_health_status = health,
    school_absences = absences,
    first_period_grade = G1,
    second_period_grade = G2,
    final_grade = G3
  ) |>
  #rearrrange the rows
  arrange(desc(final_grade),desc(mother_education)) |>
  mutate(id = row_number())
```

```
#data wrangling specific for grade data
student_data_grade <- student_data |>
#pivot longer to have one grade per row
  pivot_longer(cols = c(first_period_grade, second_period_grade, final_grade),
               names_to = "period",
               values_to = "score") |>
#refine the period names, convert educational support indicators to text.
  mutate(period = recode(period, 'first_period_grade' =  'first period',
                         'second_period_grade' = 'second period',
                         'final_grade' = 'final'),
         extra_educational_support = recode(extra_educational_support,
                                            "no" = "no school support",
                                            "yes" = "school support"),
         family_educational_support = recode(family_educational_support,
                                            "no" = "no family support",
                                            "yes" = "family support")) |>
#rearrange the sequence of the periods
  mutate(period = fct_relevel(period, 'first period' ,'second period', 'final'))
```

```r
#data wrangling specific for higher education data
student_data_higher <- student_data |>
  mutate(mother_education = as.character(mother_education),
         father_education = as.character(father_education),
         guardian_education = case_when(
           student_guardian == "mother" ~ mother_education,
           student_guardian == "father" ~ father_education
         )) |>
  drop_na(guardian_education) |>
  mutate(guardian_job = case_when(
           student_guardian == "mother" ~ mother_job,
           student_guardian == "father" ~ father_job
         )) |>
  drop_na(guardian_job) |>
  mutate(guardian_job = fct_relevel(guardian_job,
             'at_home', 'services' ,'other', 'teacher', 'health'),)|>
  mutate(guardian_job = recode(guardian_job, "at_home" = "at home"))
```

The two data sets are downloaded from UC Irvine machine learning repository, which contain data of student achievements in secondary education of two Portuguese schools in majors of Math and Portuguese. For data cleaning, we merge the data set of Math students and the data set of Portuguese students, arrange the data by descending order of their final grades, and rename all the variables to names with clearer description under snake case format.

Data description

There are 1044 rows and 35 columns in this student_data dataset. The rows are the data records from each student. The attributes are: student_school, student_sex, student_age, home_address_type, family_size, parent_cohabitation_status, mother_education, father_education, mother_job, father_job, school_choice_reason, student_guardian, travel_time, weekly_study_time, past_class_failures, extra_educational_support, family_educational_support, paid_extra_classes, extra_curricular_activities, attended_nursery_school, higher_education_aspiration, internet_access_at_home, romantic_relationship, family_relationship_quality, free_time_after_school, going_out_with_friends, workday_alcohol_consumption, weekend_alcohol_consumption, current_health_status, school_absences, first_period_grade, second_period_grade, final_grade, id.

This Student Performance dataset was created by Paulo Cortez for studying student achievement in two Portuguese secondary schools. This data collection aimed to examine student grades, demographics, social factors, and school-related features. This attribute selection aligns with research goals.

The data observed and recorded in this dataset were influenced by several factors:

1. Research Objectives: The data collection aligned with the research goal of studying student achievement in secondary education, leading to the inclusion of relevant data like student grades, demographics, social factors, and school-related features.
2. Educational Context: The dataset's focus on two Portuguese schools considered the unique characteristics of these schools, potentially including region-specific factors relevant to the research.
3. Data Collection Methods: Data was gathered through school reports and questionnaires, impacting the type of data collected. Questionnaires may have limited the data to survey-appropriate information.
4. Academic Subjects: The dataset is divided into separate datasets for Mathematics ("mat") and Portuguese language ("por"), reflecting the emphasis on these subjects and affecting the choice of recorded data.

The preprocessing steps described below were undertaken to ensure that the dataset is well-structured, informative, and ready for subsequent analysis.

1. Data Source: The initial datasets were obtained from the UC Irvine Machine Learning Repository, providing the foundational data for this analysis.
2. Data Merging: The dataset for Math students and Portuguese students was consolidated into a single unified dataset. This merging allowed for a comprehensive examination of student achievements across both subjects.
3. Data Sorting: The merged dataset was organized by arranging the data in descending order of students' final grades. This sorting aids in identifying the highest-performing students at the top of the dataset, facilitating further analysis and interpretation.
4. Variable Renaming: To enhance the interpretability and clarity of the dataset, all variables were systematically renamed. The new variable names were formatted in snake case, a convention that employs lowercase letters and underscores (_) to separate words within variable names. This step ensures that variable names are more descriptive and user-friendly.

These actions are integral in preparing the data for meaningful insights and facilitating ease of use in research.

Those involved in data collection, including researchers, teachers, and administrators, were likely aware of the study's objectives. The data was used for classification and regression tasks related to predicting student performance in Math and Portuguese. Researchers likely intended to gain insights into factors affecting achievement and test machine learning models for educational research.

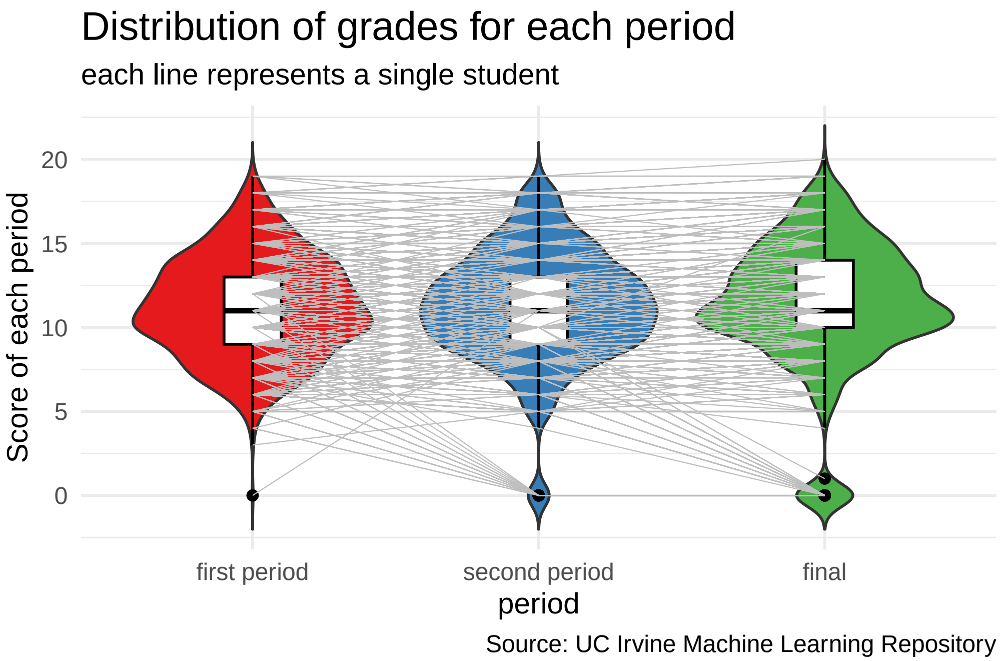
Data limitations

1. Although there are 35 variables after initial cleaning, many of the variables are stored in a binary form, and some are numeric but just several integers, so it will be hard to do linear visualizations
2. Although we combined two dataframes, the amount of observations is still not high, may be less accurate if using machine learning in the future.
3. The limitation in data collection brings only data from two Portuguese schools, making it not especially convincing if we try to generalize to all students in secondary education.

Exploratory data analysis

1. How does the distribution of grades look like for each period?

```r
#first visualization
student_data_grade |>
#draw the graph
  ggplot(mapping = aes(x = period, y= score, fill = period)) +
  geom_violin(trim = FALSE, scale = "count") +
  geom_boxplot(width = 0.2, fill = "white",
               color = "black", outlier.color = "black") +
  geom_line(aes(group = id), color = 'gray', linewidth = 0.2) +
#change graph style
  labs(
    x = "period",
    y = "Score of each period",
    title = "Distribution of grades for each period",
    subtitle = "each line represents a single student",
    caption = "Source: UC Irvine Machine Learning Repository"
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    axis.text = element_text(size = 9),
    plot.title = element_text(size = 15),
    plot.margin = margin(5, 20, 5, 5)
  ) +
  scale_fill_brewer(palette = "Set1")
```


Distribution of grades for each period

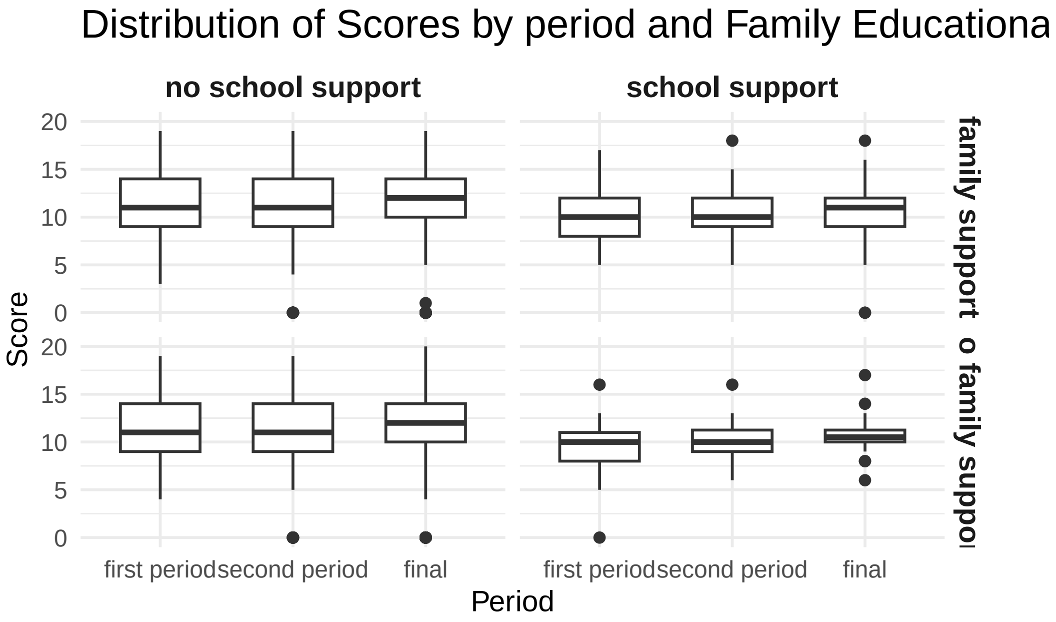
each line represents a single student

Source: UC Irvine Machine Learning Repository

In this graph, we can see that the final period has the most amount of outliers, and the first period has the least. For the boxplot for each period, even though the final period's values of Q1 and Q3 are different from the first and second periods, they three have similar median value. From the individual lines we can generally see both grade increase and decrease. However, it is especially hard to get away from a grade of zero.

2. Does educational support help students with grades?

```
#second visualization
student_data_grade |>
#draw the graph
  ggplot(mapping = aes(x = period, y= score)) +
  geom_boxplot(position = "dodge", width = 0.6) +
  facet_grid(vars(family_educational_support), vars(extra_educational_support)) +
#change graph style
  theme_minimal() +
   theme(
    legend.position = "bottom",
    axis.text = element_text(size = 9),
    plot.title = element_text(size = 15),
    plot.margin = margin(5, 20, 5, 5),
    strip.text = element_text(size = 11, face = "bold"),
  ) +
  scale_fill_brewer(palette = "Pastel1") +
  labs(title = "Distribution of Scores by period and Family Educational Support",
       x = "Period",
       y = "Score",
       fill = "Extra Educational Support",
       caption = "Source: UC Irvine Machine Learning Repository"
       )
```



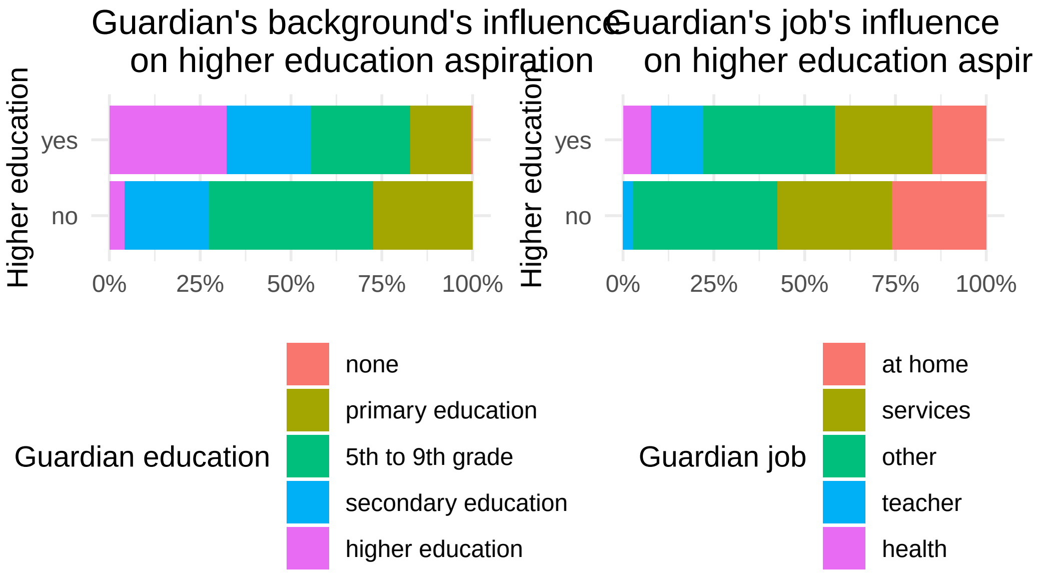
Source: UC Irvine Machine Learning Repository

In this graph, we can see that the medians of each boxplot are very similar, which indicates that with or without school or family support has little impact on students' median scores. Most boxplot has little outliers. But only the boxplot representing the students who only have school support has the most amount of outliers.

3. Does the guardian's education status and job affect the student's willingness for higher education?

```
#third visualization
# graph according to guardian education
graph_edu <- student_data_higher |>
#draw the graph
  ggplot(aes(fill = guardian_education,
             y = higher_education_aspiration)) +
  geom_bar(position = "fill") +
#change graph style
  theme_minimal() +
  scale_x_continuous(labels = label_percent()) +
  theme(legend.position = "bottom") +
  guides(fill = guide_legend(nrow = 5)) +
  labs(
    x = NULL,
    y = "Higher education",
    title = "Guardian's background's influence
    on higher education aspiration"
  ) +
  scale_fill_discrete("Guardian education",
                      labels=c("none", "primary education", "5th to 9th grade",
                               "secondary education", "higher education"))
# graph according to guardian job
graph_job <- student_data_higher |>
#draw the graph
  ggplot(aes(fill = guardian_job,
             y = higher_education_aspiration)) +
  geom_bar(position = "fill") +
#change graph style
  theme_minimal() +
  scale_x_continuous(labels = label_percent()) +
  theme(legend.position = "bottom") +
  guides(fill = guide_legend(nrow = 5)) +
  labs(
    x = NULL,
    y = "Higher education",
    title = "Guardian's job's influence
    on higher education aspiration",
    caption = "Source: UC Irvine Machine Learning Repository"
  ) +
  scale_fill_discrete("Guardian job")

#combine two graphs
(graph_edu | graph_job)
```



Source: UC Irvine Machine Learning Repository

In this graph, we can see that the student guardian's education background is positively related to higher education aspiration. We can see more than 25% of student's who wants higher education have guardian from higher education, and more than 50% have at least secondary education. These stats for students not wanting higher education is about 5% and 30%. We can see a similar effect in terms of jobs, as students with guardians in teacher and health industry are much more likely to want higher education.

Questions for reviewers

1. Do you think it is reasonable to use grades as a whole in some cases, though they belong to two different subjects?
2. Do you think it is possible to do machine learning with this dataset?