Data description

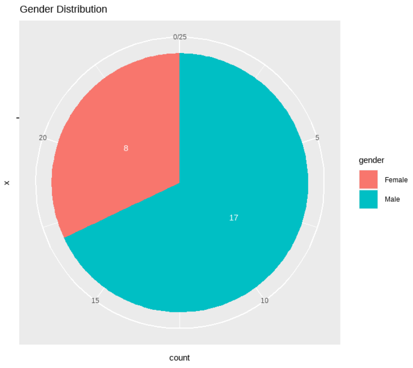
This analysis aims to examine a gradebook, an often overlooked but vital data source in education. The dataset used here is derived from simulated student data, sourced from Chapter 8 of the book “Data Science in Education Using R”. With this dataset, we will look into the interplay between different types of academic assessments (e.g. scores in classworks, homeworks, projects, formative and summative assessments, and the final grade calculated by taking the average of aforementioned types of assessments), as well as the relationship of some individual characteristics (e.g. age, race, gender, financial status, attendance) and students’ academic performance.

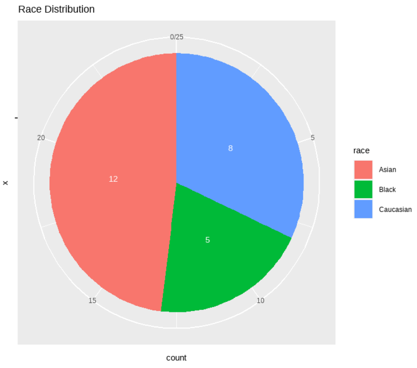
The dataset consists of 25 rows representing 25 students in a class and 59 columns, offering a comprehensive view of student performance across various assessments and demographic variables.

Data on individual characteristics

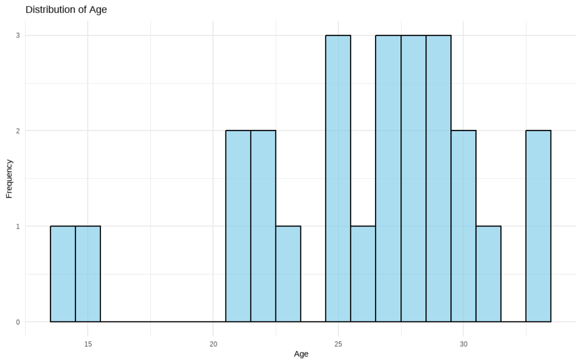
We first look into the variables that can be classified under the umbrella of “individual characteristics”. These include age, race, gender, financial status, and attendance, which comprises information on lateness and absences. These characteristics provide insights into the diverse attributes of the students within an educational setting.

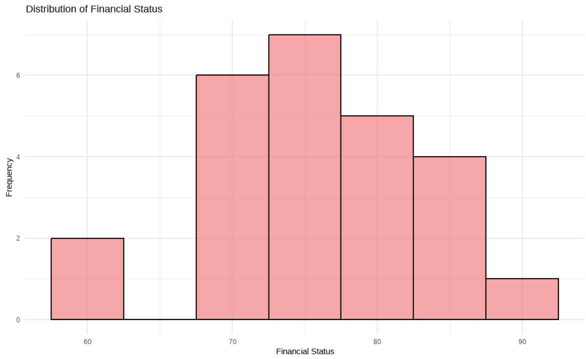
The data for gender and race is simulated using random sampling. The bar charts below depict the distribution of gender and race in the student population.



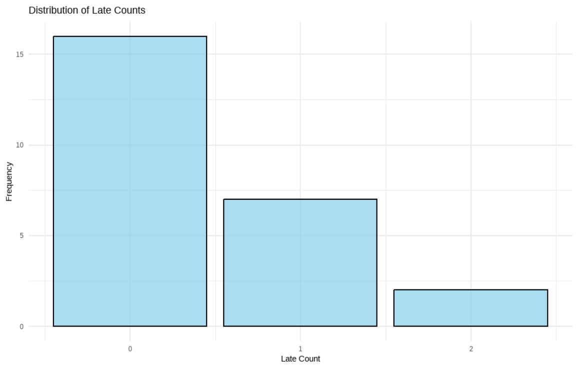


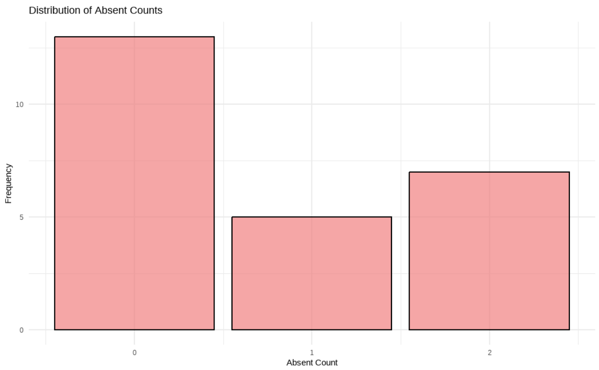
In this dataset, the “age” column is created by generating random values from a normal distribution with a mean of 25 and a standard deviation of 5. Then, the generated values are rounded to integers, resulting in a synthetic age distribution that mimics a typical mean age with some variability. Similarly, the “financial_status” column is generated by drawing random values (0-100) from a normal distribution with a mean of 75 and a standard deviation of 10. Histograms are used to provide an overview of the distribution of financial scores and age.





In the columns “absent” and “late”, 9 rows were pre-filled. To simulate new data for the remaining rows, random values are generated based on the zero-inflated negative binomial model, which has been fitted to the existing data. These values are then assigned to new columns, while keeping the original data in separate columns. Finally, the “late” and “absent” columns in the gradebook are updated with the simulated values, ensuring a blend of original and newly simulated data. To visualize the distribution of counts for “late” and “absent”, bar charts are used.

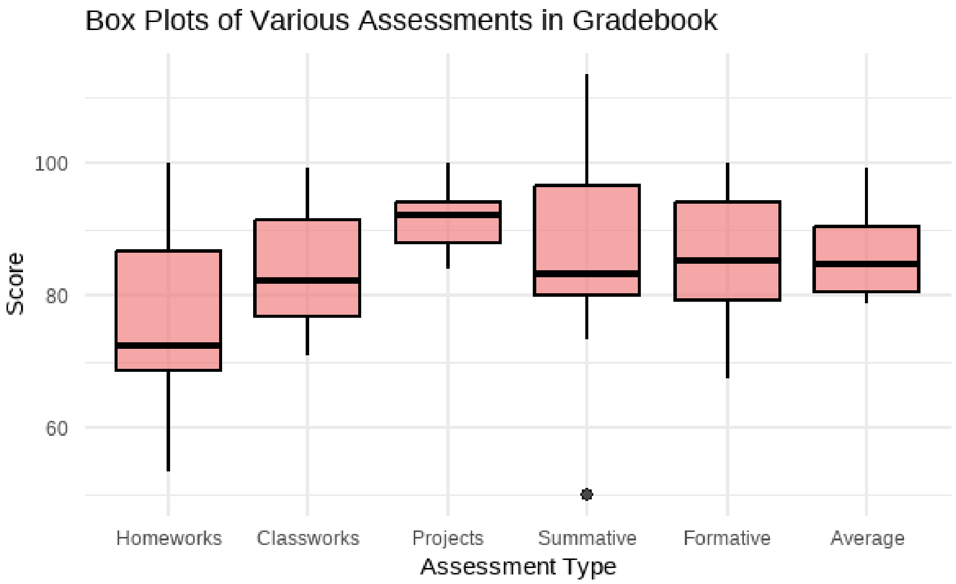




Here we can see that a majority of students are never late or absent from class.

Data on academic performance

Next, we examine variables that are related to academic outcomes. This group includes scores in various assessments such as scores in classworks, homeworks, projects, formative assessments, summative assessments and final grade which is the average of the aforementioned assessments. Box plots offer a visual summary of the distribution of these assessments.



The medians in all assessment types are rather high (above 80), except for the median in homework scores, which lies just above 70. Overall, we don’t have a great variability within the dataset because the IQR in all box plots is not too large. Only 1 outlier is present in the box plot for summative assessments.

The box plot for projects is comparatively short and its median lies above the median of other box plots, which suggest that in overall, students receive high scores in projects.

The distribution of homework scores and summative assessments is negatively skewed, which was proved by the fact that the medians are close to the lower end of the box. However, in the case of summative assessments, a short whisker pointing down suggests that the lower values in the dataset are tightly clustered. While the majority of the data is concentrated in the lower range, there are some higher values that contribute to the comparatively long whisker pointing up. In comparison, in the box plot for homework scores, whiskers on both ends are roughly of the same length, which suggest an equal spread in the lower and higher values.

For the final grades, though the majority of the data is concentrated around the median, there are some students that receive higher scores that contribute to the extended upper whisker.

Research questions

- RQ1: What is the relationship between formative and summative assessments?
- RQ2: What is the relationship between homework, classwork, and project grades?
- RQ3: What has a stronger impact on the final grade: homework, classwork, or project grades?

- RQ4: Better late or absent? What is the impact of being late vs. being absent?
- RQ5: What has a stronger impact on final grade: race or gender?
- RQ6: What has a stronger impact on final grade: age or financial status?

Methods

Method for Research Questions 1 and 2: Examining Relationships through Correlation Analysis

To address the first two research questions regarding the relationships between grades in various types of assessments, Pearson correlation coefficient is employed. Correlation analysis allows the strength and direction of associations between formative and summative assessments, as well as between homework, classwork, and project grades to be quantified. Prior to conducting the correlation analysis, assumptions that underlie this method are examined. These assumptions are listed in the book “Statistics for Research Students” by Fein et al. First, the linearity of the relationships is assessed, ensuring that the patterns between variables can be adequately represented by a straight line. Additionally, we check for normal distribution of the data using the Shapiro-Wilk normality test. Outliers, which Pearson correlation coefficient is sensitive to, are identified with box plots and addressed to enhance the robustness of our findings.

Method for Research Questions 3 to 6: Evaluating Factors’ Impact through Multiple Linear Regression

For the remaining research questions, which investigate the impact of different factors on the final grade, we employ multiple linear regression analysis. Assumptions associated with methods, which are also listed in the book “Statistics for Research Students”, are examined and validated. Similar to correlation analysis, we also need to check for linearity, but in the case of RQ5, assumption of linearity do not need to be checked because age and gender are categorical variables, and “for a categorical predictor, the linearity assumption is always met for each of the indicator functions since a straight line always fits two points exactly” (Nahhas, 2023). For RQ4, we do not check for linearity either, because though technically “late” and “absent” are count variables, they only have 4 values here, namely 0, 1, 2, and 3, so in this analysis, they are treated as categorical variables. Aside from linearity, we check for homoscedasticity with Breusch-Pagan test when the independent variables are continuous (RQ3, RQ6) and with the plot of residuals vs. fitted values when the independent variables are categorical (RQ4, RQ5). Multicollinearity is checked using VIF. Same as correlation analysis, outliers are examined using box plots, except in the case of categorical variables. Normal distribution of residuals are checked using the Shapiro-Wilk test.

Results

RQ1: What is the relationship between formative and summative assessments?

Pearson correlation is computed to assess the relationship between formative and summative assessments after the outlier in summative assessments was removed. There is a strong positive correlation between the two variables, r = .51, N = 24; the relationship is significant (p = .01). There is strong evidence that formative assessments are positively correlated with summative assessments.

RQ2: What is the relationship between homework, classwork, and project grades?

```
            classworks  homeworks   projects
classworks       1.00      0.18      0.41
homeworks        0.18      1.00      0.61
projects         0.41      0.61      1.00

n= 25



P
           classworks homeworks      projects
classworks            0.3810         0.0392
homeworks  0.3810                    0.0011
projects   0.0392     0.0011
```

A Pearson correlation coefficient matrix is used to show the relationship between homework, classwork and project grades. There is a positive correlation between each pair of variables, however, only the relationship between projects and classworks grades (r = .41, N= 25), as well as the relationship between homeworks and projects (r = .61, N = 25) are significant with p < .05. Hence, there is strong evidence for a positive correlation between projects and classworks grades, as well as a positive correlation between projects and homeworks grades.

RQ3: What has a stronger impact on the final grade: homework, classwork, or project grades?

A multiple regression is run to estimate what has a stronger impact on the final grade among homework, classwork, and project grades. This results in a significant model F = 22.44, p < .01. The individual predictors are examined further and indicate that classwork (t = 3.678, p < .01), homework (t = 2.859, p < .01), and project grades (t = 2.159, p < .05) are all significant predictors. Project grade has the largest coefficient (0.46), followed by classwork grade's coefficient (0.31) and homework grade's coefficient (0.19). This suggests that project grade has the strongest impact on the final grade, followed by classwork grade and homework grade respectively.

RQ4: Better late or absent? What is the impact of being late vs. being absent?

The multiple linear regression model is used to predict the running average based on the predictors "late" and "absent". The F-statistic compares the fit of the model with predictors to a model without predictors. The p-value of 0.4924 suggests that the model as a whole is not statistically significant at the 0.05 level. The coefficient for tardiness does not reach conventional significance levels (p = 0.369). Similarly, the coefficient for absence is not statistically significant (p = 0.333). Though the coefficient for "late" is -1.837, suggesting a negative correlation between tardiness and final grades, and the coefficient for "absent" is 1.519, indicating a positive correlation between absence and final grades, none of these relationships can be proved statistically significant. The overall model's explanatory power is limited, as indicated by the non-significant F-statistic.

RQ5: What has a stronger impact on final grade: race or gender?

The multiple linear regression model is used to examine the impact on the final grade of the predictors "race" and "gender".The F-statistic's p-value of 0.5637 suggests that the model as a whole is not statistically significant at the 0.05 level. None of the coefficients for 'race' or 'gender' reach conventional significance levels (p-values > 0.05), suggesting that the effects of these variables on the final grade are not statistically significant at the 0.05 level. The coefficient for "raceBlack" is 2.894, for “raceCaucasian” is 3.435, indicating a positive association between identifying as Black or Caucasian and the final grade. Similarly, the coefficient for 'genderMale' is -1.307, suggesting a negative association between being male and the final grade. However, none of these associations are statistically significant.

RQ6: What has a stronger impact on final grade: age or financial status?

The multiple linear regression model is applied to investigate the influence of predictors, specifically "age" and "financial_status," on the final grade. The F-statistic's p-value of 0.9158 indicates that the model as a whole is not statistically significant at the 0.05 level. Examining individual predictors, neither "age" (p-value = 0.809) nor "financial_status" (p-value = 0.726) exhibit statistically significant effects on the running average at the 0.05 significance level. The intercept is highly significant (p-value < 0.001), indicating a strong evidence that it is not zero. The coefficient for "age" is -0.066, implying a negative association between age and the final grade, although this relationship is not statistically significant. Similarly, the coefficient for "financial_status" is -0.064, also suggesting a negative association between financial status and the final grade, yet this association is not statistically significant. In conclusion, the model does not provide robust evidence for the statistical significance of the examined predictors, and their individual impacts on the running average are inconclusive.

References

Fein, E. C., Gilmour, J., Machin, T., & Hendry, L. (2022, June 16). *Section 4.2: Correlation Assumptions, interpretation, and write up*. Statistics for Research Students. https://usq.pressbooks.pub/statisticsforresearchstudents/chapter/correlation-assumptions/

Fein, E. C., Gilmour, J., Machin, T., & Hendry, L. (2022b, June 16). *Section 5.3: Multiple regression explanation, assumptions, interpretation, and write up*. Statistics for Research Students. https://usq.pressbooks.pub/statisticsforresearchstudents/chapter/multiple-regression-assumptions/

Nahhas, R. W. (2023, November 21). *Introduction to regression methods for public health using R*. 5.16 Checking the linearity assumption. https://www.bookdown.org/rwnahhas/RMPH/mlr-linearity.html