

**INTERIM REPORT**

**STATISTICAL DATA ANALYSIS PROJECT**

**PREPARED BY**

**SEVDE İZMİRLİ**

**2502128**

**ABSTRACT**

In this project, factors affecting the success of secondary school students were examined. Statistical methods were used in making analyses. In line with the data obtained, factors that increased or decreased student success were determined.

1. **INTRODUCTION**

It is very important to identify the factors that affect students' success in the field of education and to work on improving failures. The success of the student depends not only on his/her own efforts but also on socio-demographic, psychological and environmental variables. These variables are parental education levels, support systems, social activities and behavioral patterns such as alcohol consumption. This project aims to examine the relationship between various factors in the data set and student success. In addition, the project addresses the problem of missing data. Necessary tests have been applied regarding this problem. EDA, CDA and many statistical tests were used during the analysis.

**1.1 Data Description**

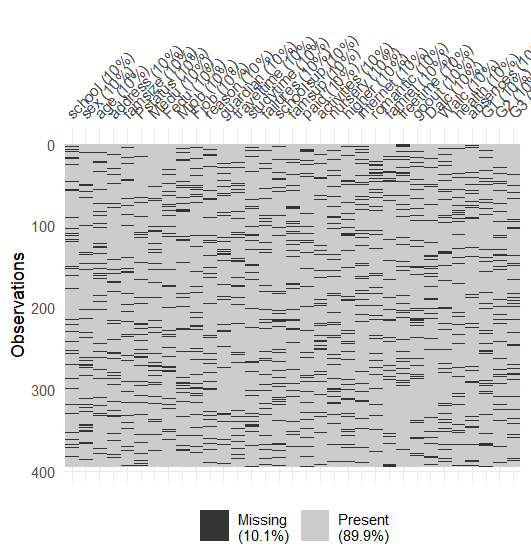
This project was conducted using a dataset containing the performance of secondary school students. Dataset contains 395 observations and 33 variables, including demographic information (such as age, sex, and address), parental background (education levels, jobs, and family structure), academic support (school and family support, paid tutoring, and higher education aspirations), lifestyle and behavioral factors (study time, absences, alcohol consumption, free time, and going out), and academic performance indicators (grades from three periods: G1, G2, and G3). Several variables are categorical (e.g., sex, schoolsup, internet), ordinal (e.g., studytime, goout), or continuous (e.g., age, absences, G3).

**1.2 Research Questions**

1. Is there a significant relationship between the education level of students' mothers and their end-of-year grade point average?
2. Is there a significant difference in academic performance categories between genders?
3. Does alcohol consumption on weekdays and weekends affect students' achievement?
4. Does the distance between home and school affect end-of-year grades?
5. Is there a difference between male and female students in terms of extracurricular activities?

**1.3 Aim of The Study**

The main objective of this project is to investigate the factors affecting students' academic performance by analyzing demographic characteristics, family background, school-related variables, and behavioral attributes through statistical tests, visualization models, and predictive modeling techniques. These patterns derived from variables such as parental education, school support, alcohol consumption, and study habits will form the basis for understanding academic achievement and risk factors. The results of this study are intended to help students make data-driven decisions to support student success.

1. **RESULTS AND FINDINGS**

To obtain a realistic data scenario and evaluate the performance of missing data handling techniques, approximately 10% missingness was added to the dataset completely at random (MCAR). In the graph, each horizontal line represents a separate observation, while each column corresponds to a variable. Grey cells represent observed (non-missing) values, while black cells represent missing values.

metin, diyagram, ekran görüntüsü, renklilik içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

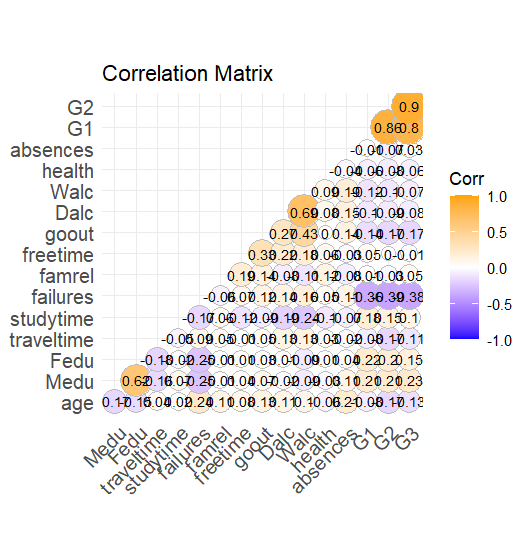
Most students have mothers with higher levels of education. In terms of school support, a large proportion of students report receiving no additional educational assistance. Most students live close to school. In terms of social behavior, students report a predominantly moderate amount of outgoing and leisure time.

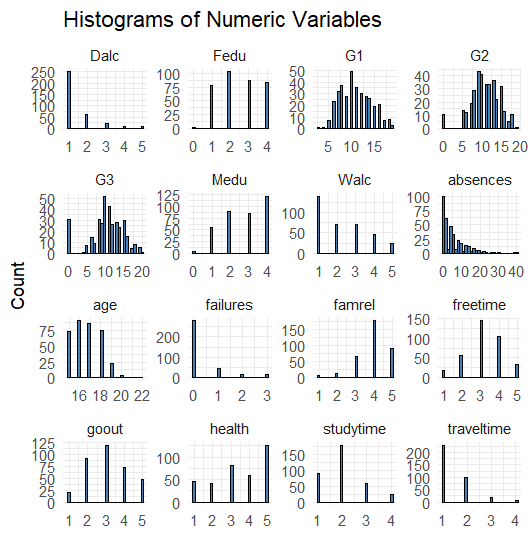
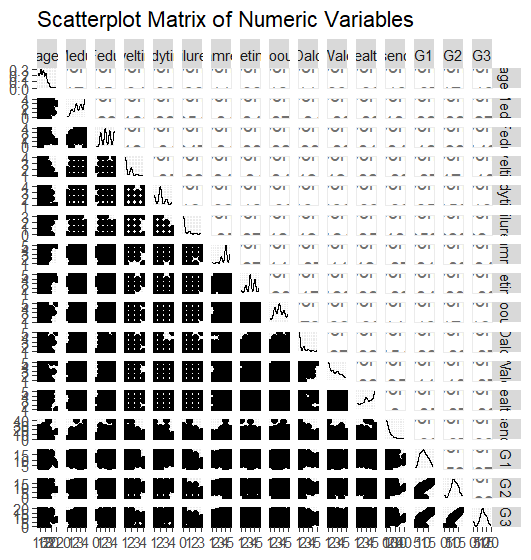
metin, diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

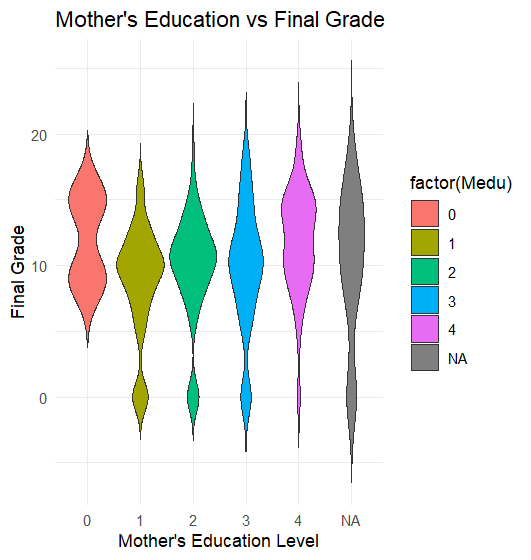
The visualizations show that most students report low alcohol consumption on both weekdays and weekends, although weekend consumption is more varied. Students generally have a moderate amount of free time after school, and the gender distribution is nearly balanced with a slight female majority.

1. **Exploratory Data Analysis**

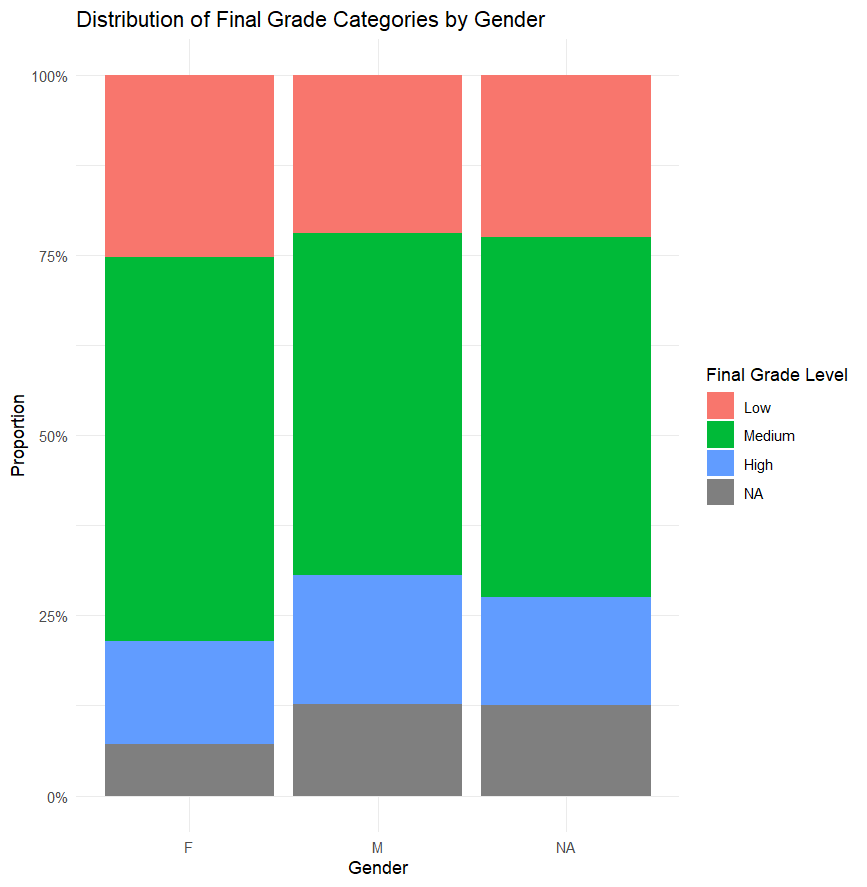
 The correlation matrix reveals the linear relationships between students' end-of-year grades (G3) and other numerical variables. The highest positive correlations are seen between G3 and G2 and G1. In addition, a significant negative correlation is observed between the number of failed courses and G3. Variables such as studytime and mother's education level (Medu) show weak but positive relationships with G3. On the other hand, weak negative correlations are found between alcohol consumption (Dalc and Walc) and G3.

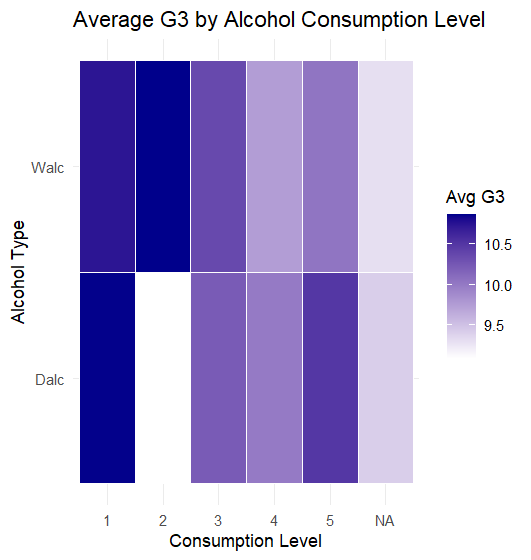
The histograms show that the majority of the numerical variables are right-skewed. While the scores for G1, G2 and G3 are close to normal distribution, variables such as absences, failures, Dalc and Walc are concentrated in low values. The scatterplot matrix clearly shows that there are strong linear relationships, especially among G1, G2 and G3. The relationships with the other variables are weak and widely distributed, indicating that there is no significant multicollinearity among the variables.

1. **Research Questions**
   1. **Is there a significant relationship between the education level of students' mothers and their end-of-year grade point average?**

 The graph shows the relationship between the mothers' level of education and the students' end-of-year grades (G3). In general, the median of the final grades of students with higher mothers' level of education is higher and their distribution is concentrated at the top. It is seen that the level of success is higher in children of mothers with education levels of 3 and 4. This situation shows that the mother's level of education may be positively related to the student's academic success.

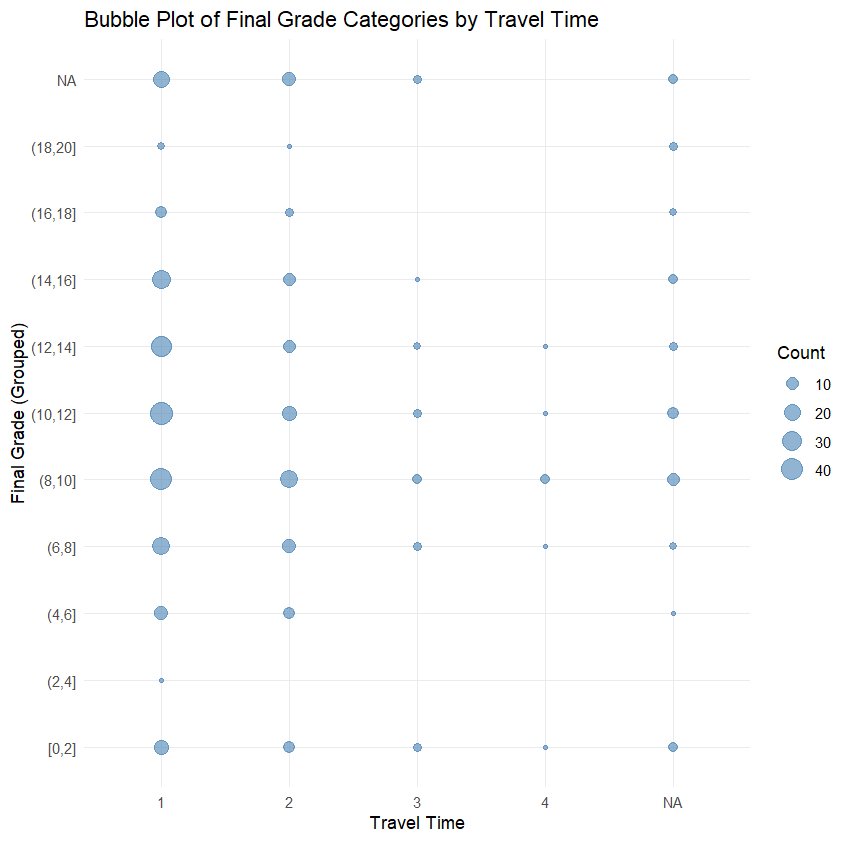
**4.2 Is there a significant difference in academic performance categories between genders?**

The graph shows the distribution of male and female students according to their end-of-year success levels in proportion. The highest rate in both genders is in the "Medium" success level. However, it is striking that the "High" success rate is higher among male students than among female students, whereas the "Low" success rate is slightly higher among female students than among males.

**4.3 Does alcohol consumption on weekdays and weekends affect students' achievement?**

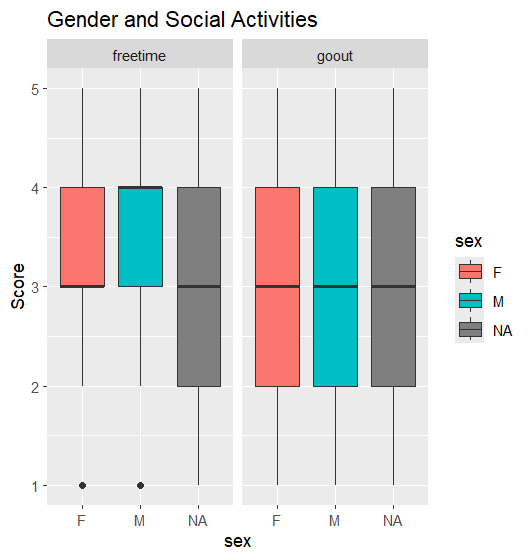
The graph shows the average final grades (G3) of students according to their alcohol consumption levels during the week (Dalc) and the weekend (Walc). The darker the shade of blue, the higher the average grade. It is clear from the graph that students have higher average achievement when alcohol consumption is at lower levels. As the level of consumption increases, the color becomes lighter, indicating lower average grades. This trend is observed for both weekdays and weekends.

**4.4) Does the distance between home and school affect end-of-year grades?**



The graph shows students’ end-of-year grades (G3) grouped into different travel time categories. The size of the bubbles indicates how many students are in each grade range. Observations show that students with the shortest travel time (category 1) tend to be concentrated in higher achievement categories. In contrast, students with longer travel times (categories 3 and 4) are less likely to be in higher achievement categories. This suggests that longer travel times may have a negative impact on academic achievement.

**4.5) Is there a difference between male and female students in terms of extracurricular activities?**



The graph shows the distribution of social activities (goout) and free time (freetime) scores of male and female students. The boxplots are quite similar for both genders; the median values ​​are almost the same. These visual findings show that there is no significant difference between male and female students in terms of leisure time and participation in social activities.

1. **MISSINGNESS MECHANISM**

**metin, ekran görüntüsü, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

The parametric test is based on normality and rejects MCAR, while the non-parametric test rejects normality and does not reject MCAR. In this case, the non-parametric test is considered more reliable, considering that the data is not normally distributed. Therefore, it is concluded that the missing values ​​in the available data "may be MCAR", that is, there is no strong evidence that the missingness structure is due to a systematic pattern.

1. **FEATURE ENGINEERING**

Several new features were designed to enrich the dataset and create more informative variables for analysis. First, a new variable called alc\_total was created by summing the alcohol consumption levels during the week (Dalc) and the weekend (Walc). Second, the variable avg\_grade was calculated as the arithmetic mean of the three semester grades (G1, G2 and G3), providing a single numerical indicator of the students' overall academic performance throughout the year. Finally, parent\_edu was defined as the average of the mother's (Medu) and father's (Fedu) education levels, aiming to capture the combined parental education effect.

1. **CONFIRMATORY DATA ANALYSIS**

For each research question, normality, variance homogeneity were tested and necessary statistical tests were applied according to the results such as Kruskal-Wallis, Wilcoxon Rank-Sum. For Question 1, Kruskal-Wallis test result found that there was a significant difference between the education level of the mothers and the end-of-year grades. For Question 2, Wilcoxon Rank-Sum test result found that there was a significant difference in the success of students who received and did not receive school support. For Question 3, Kruskal-Wallis test result found a significant difference between at least one group. It was concluded that alcohol consumption may affect success. For Question 4, Kruskal-Wallis test result showed a significant difference between travel time and success. Finally, for Question 5, Wilcoxon test result found no significant difference between social activity and gender.

1. **CROSS-VALIDATION**

To evaluate the predictive performance of the linear regression model, the dataset was randomly split into training and test sets. The model robustly explains a significant portion of the variance in the training data with an R-squared of 0.8475 and an adjusted R-squared of 0.8445. When evaluated on the test set, the model achieved a Root Mean Square Error (RMSE) of 2.2 and an R-squared of 0.79, demonstrating good generalization and predictive accuracy. These results confirm that prior performance and behavioral characteristics are strong predictors of students' final academic outcomes.

Model assumption checks were performed. The Q-Q plot created shows that the residuals tend to deviate from the normal distribution. The “residuals vs fitted” graph created shows that the points exhibit a distinct pattern, especially the concentration of extreme values ​​in the lower region, suggesting that the homoskedasticity (constant variance) assumption has been violated and that extreme values ​​may be effective in the model. As a result of multicollinearity checking, it was determined that there was no multicollinearity problem.

As a result of the comparison of training and test performance, the RMSE (Root Mean Square Error) value obtained on the training set is 1.71 and the MAE (Mean Absolute Error) value is 1.10. This shows that the model has very low errors on the training data and has successfully learned. However, the RMSE value was observed as 2.20 and the MAE value as 1.49 on the test set; this shows that the error rate increased slightly when the model was applied to the test data. This difference shows that the model did not over-learn, but its test performance was slightly lower than its training performance. Nevertheless, it can be said that the model generally gives very successful and consistent results on both training and test data.

1. **CONCLUSION**

In this study, Exploratory and Confirmatory Data Analysis (EDA and CDA) methods were used to analyze the factors affecting student success, imputation was applied with appropriate techniques for missing data and then a regression model was developed. The results obtained showed that students' end-of-year grades were significantly related to variables such as their previous grades, failure status and study time. The performance of the model was satisfactory in both training and test sets and generally exhibited a strong and consistent structure. These findings provide important clues in terms of determining the basic factors affecting educational success.