Student Performance Analysis Project

CID: 01862429

2024-05-14

Table of contents

0.1 Abstract

Our aim is to uncover insights from the “Student Attitude and Behavior” dataset downloadable from Kaggle https://www.kaggle.com/datasets/susanta21/student-attitude-and-behavior to better understand the factors influencing student performance and well-being.

In this detailed study, we explore the complex connections between various academic and personal factors and how they affect college students’ grades. We begin with a careful examination of the data, ensuring everything is accurate and relevant. We also perform checks to identify any unusual data points that might distort our analysis. Next, we analyze how different factors, such as study habits or family background, are interconnected to reveal underlying patterns.

Following this, we look at grade variations across different academic levels to identify trends and discrepancies. We also investigate differences in academic performance between male and female students, enhancing our understanding of gender dynamics within education.

Our study further examines the relationship between grades and various demographic and personal attributes, taking gender into account. To predict college grades effectively, we employ advanced modeling techniques like decision trees and random forests, which consider a mix of past academic performance and personal characteristics.

By comparing the predictive power of models with and without previous grades, we isolate the impact of students’ academic histories from other factors. This approach highlights the most significant predictors of academic success and offers insights for educational institutions to develop targeted support strategies.

Our thorough analysis and use of modeling techniques aim to add to the discussion in educational research by offering a clearer understanding of what affects student performance. This can guide the creation of more effective educational policies and strategies.

1 Data Introduction

This dataset comprises 235 responses from university students collected via a Google form. It encapsulates various aspects of the students' academic and personal lives, aimed at understanding behaviors and preferences that might influence their academic performance and overall well-being. The dataset includes detailed information about each student's academic background, personal habits, and socio-economic status, such as whether they have completed certification courses, their department, academic achievements at different educational levels, daily routines, and social engagements.

The data covers a range of variables:

- **Certification Course**: Indicates completion of any certification courses.
- **Gender**: The gender of the student.
- **Department**: Field of study.
- **Height (CM)** and **Weight (KG)**: Physical measurements.
- **Marks**: Academic performance in 10th grade, 12th grade, and college.
- **Hobbies**: Personal interests and activities.
- **Daily Studying Time**: Time spent studying each day.
- **Preferred Study Environment**: The environment where the student prefers to study.
- **Salary Expectation**: Expected future salary.
- **Degree Satisfaction**: Satisfaction with their chosen academic degree.
- **Career Willingness**: Willingness to pursue a career related to their degree.
- **Social Media & Video Usage**: Engagement with digital media.
- **Traveling Time**: Commute time to educational institution.
- **Stress Level**: Self-reported stress levels.
- **Financial Status**: Economic background.
- **Part-time Job**: Engagement in part-time work.

This comprehensive data collection allows for a multifaceted analysis of factors impacting student life and academic outcomes.

1.1 Data Cleaning

The data cleaning process was crucial in preparing the dataset for detailed analysis. Initially, we conducted a thorough examination of the dataset structure to understand the types and potential inconsistencies of the data. This step helped identify the need for several cleaning actions:

- **Handling Missing Values**: We checked for missing data across all variables and reported the findings. Where missing values were found, they were carefully handled either by imputation or removal, depending on their impact on the analysis.
- **Standardizing Variable Names**: To ensure consistency and ease of data manipulation, all variable names were converted to lowercase. Spaces within names were replaced with underscores, thus standardizing the names across the dataset.
- **Transforming Variables**: Key variables were transformed to better reflect their ordered nature. For instance, variables like daily studying time, social media usage, and traveling time were converted into ordered factors, which categorize their respective ranges into a structured format.
- **Deriving New Variables**: New variables such as Body Mass Index (BMI) were calculated to provide additional insights. We also explored changes in academic performance by calculating the difference and percentage change between marks obtained at different academic levels.
- **Saving the Cleaned Data**: Once cleaned, the dataset was saved in an RDS file format to maintain the integrity of the data and ensure that our results could be reproduced and further analyzed in subsequent phases of the project.

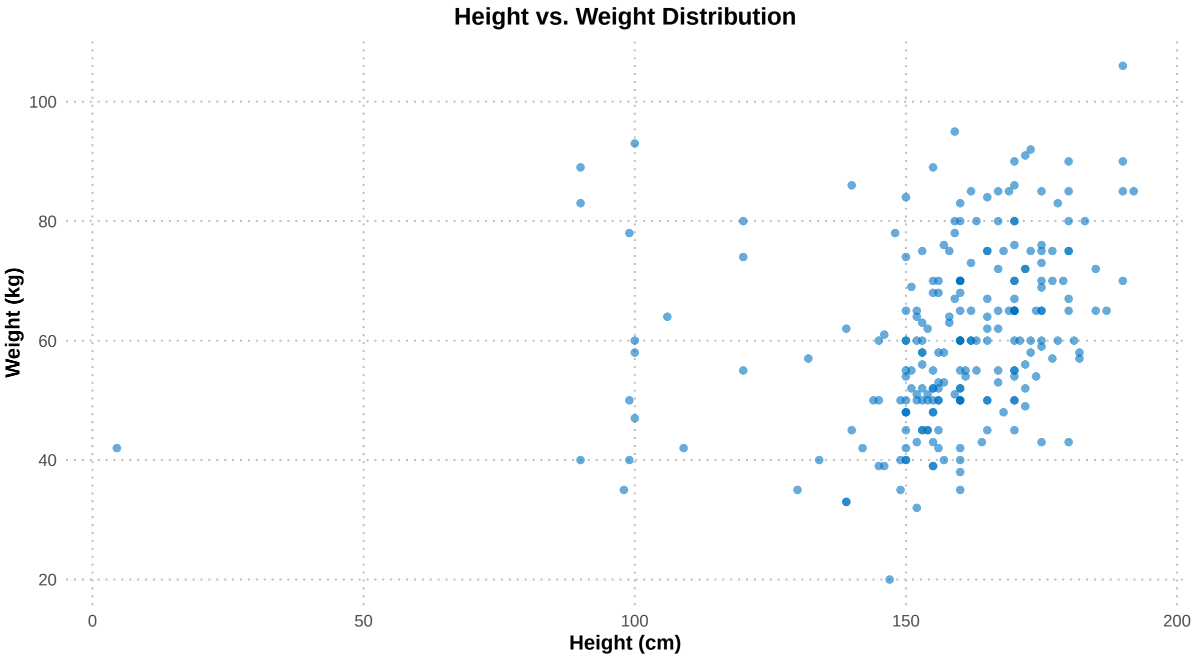
The systematic approach adopted in cleaning the dataset not only facilitated robust analysis but also adhered to best practices in data management, ensuring that the findings from this study would be reliable and valuable.

2 Outlier Analysis

In this section, we address and rectify potential outliers identified in the “Student Attitude and Behavior” dataset. Accurate outlier detection is crucial as it helps in ensuring the reliability of our statistical analyses and results.

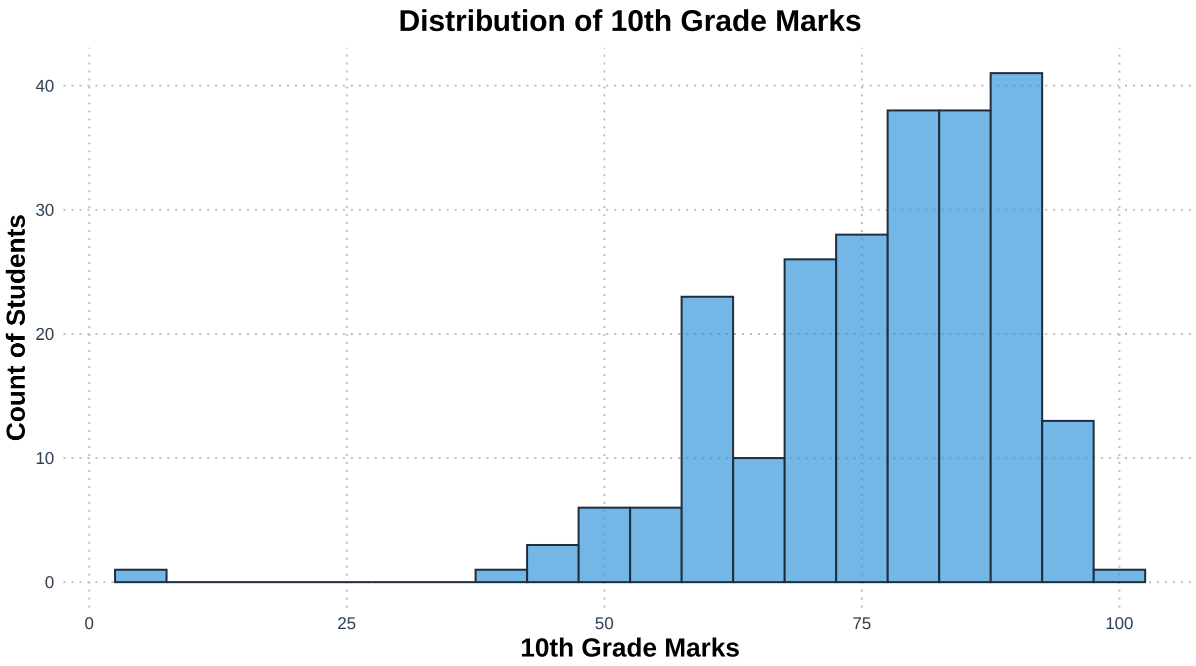
2.1 Height vs. Weight Distribution Analysis

Upon visual inspection of the Height vs. Weight Distribution plot, we identified a clearly unrealistic record where a student was reported to have a height of 4.5 cm and a weight of 42 kg. It was assumed that the height had been incorrectly entered, as other students with a similar weight had heights around 145 cm. Therefore, we corrected the height to 145 cm for this outlier.



2.2 10th Grade Marks Distribution Analysis

During our analysis of the distribution of 10th grade marks, we discovered an outlier where one student had an unusually low mark of 7.4, despite having reasonably high marks in the 12th grade and college. Given the context and comparison with peer performance, we have noticed that for both 10th and 12th grade all grades were higher than 40%, it was determined that the mark was likely incorrectly recorded. Assuming a typographical error, we corrected the 10th grade mark from 7.4 to 74.



After these corrections, we reviewed and saved the corrected dataset for further analysis. This systematic approach to identifying and rectifying outliers ensures the accuracy and credibility of our subsequent data analyses.

3 Variables Correlation

After identifying and correcting outliers in the previous sections, we proceeded to examine the interrelationships between various ordered variables within our dataset. This analysis helps us understand how different aspects of student behavior and performance are associated with each other, which can be crucial for developing targeted interventions or educational programs.

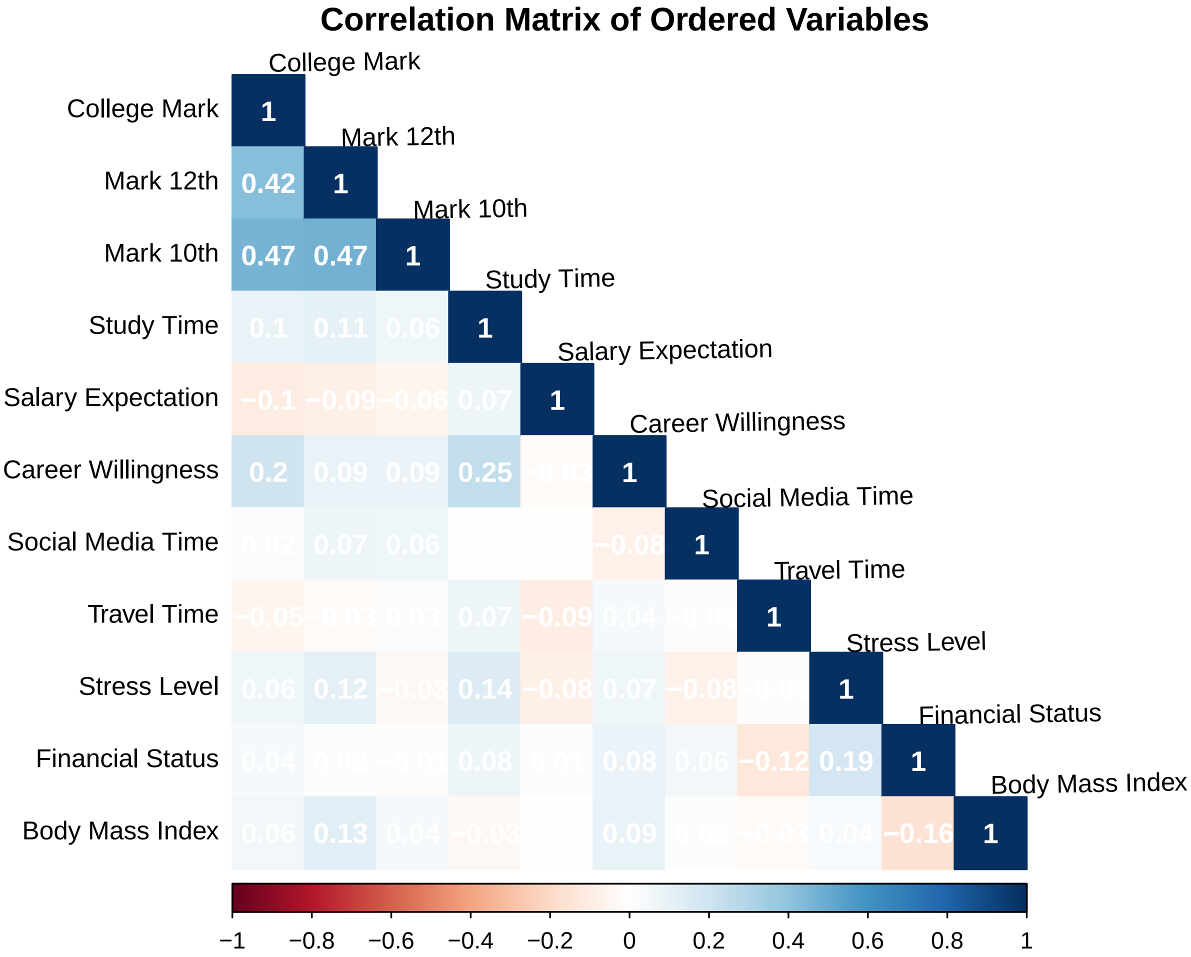
3.1 Correlation Analysis Methodology

The correlation analysis was conducted using the `corrplot` package in R, which is designed for visualizing correlation matrices. To prepare for this analysis, ordered categorical variables were first converted into numeric ranks, allowing us to calculate Pearson correlation coefficients. This transformation was crucial as it enabled the computation of linear relationships between variables that are ordinal by nature.

We selected key variables that represent different dimensions of student life and academic performance, including academic marks across different levels (college, 12th grade, and 10th grade), study habits, career aspirations, social media usage, travel times, stress levels, financial status, and body mass index (BMI). These variables were thought to encompass the primary factors that could influence a student's academic trajectory and well-being.

3.2 Correlation Matrix Results

The resulting correlation matrix is presented below:



3.2.1 Key Observations:

- **Academic Performance**: There is a moderate positive correlation between marks at different educational levels (college, 12th, and 10th), with coefficients ranging from 0.42 to 0.47. This suggests a consistent performance trend among students across different stages of their academic journey.

- **Study Time and Career Willingness**: A positive correlation (0.25) between study time and career willingness indicates that students who dedicate more time to studying tend to have a higher inclination towards pursuing a career related to their degree.
- **Social Media Time and Stress Levels**: Interestingly, social media time has a very slight negative correlation with career willingness (-0.08) and no significant correlation with stress levels, suggesting that social media usage might not be as impactful on academic or psychological outcomes as commonly perceived.
- **Travel Time and Stress Levels**: Travel time shows a positive correlation (0.19) with stress levels, highlighting that longer commutes might be contributing to higher stress among students.
- **Financial Status and BMI**: There is a notable negative correlation (-0.16) between financial status and BMI, possibly indicating socioeconomic factors influencing health and nutritional status.

3.3 Conclusion

This correlation analysis provides valuable insights into how various factors are interlinked within our dataset. Understanding these relationships helps in pinpointing areas that may require more focused research or intervention to improve student outcomes. It also aids in hypothesizing potential causal relationships, although further statistical testing and research would be required to confirm these.

4 Grade Distributions

Following the analysis of variables' correlations and outlier management, we examined the distribution of grades across different educational levels and analyzed these distributions by gender. This section helps in understanding how academic performance varies among students, and how it might be influenced by gender.

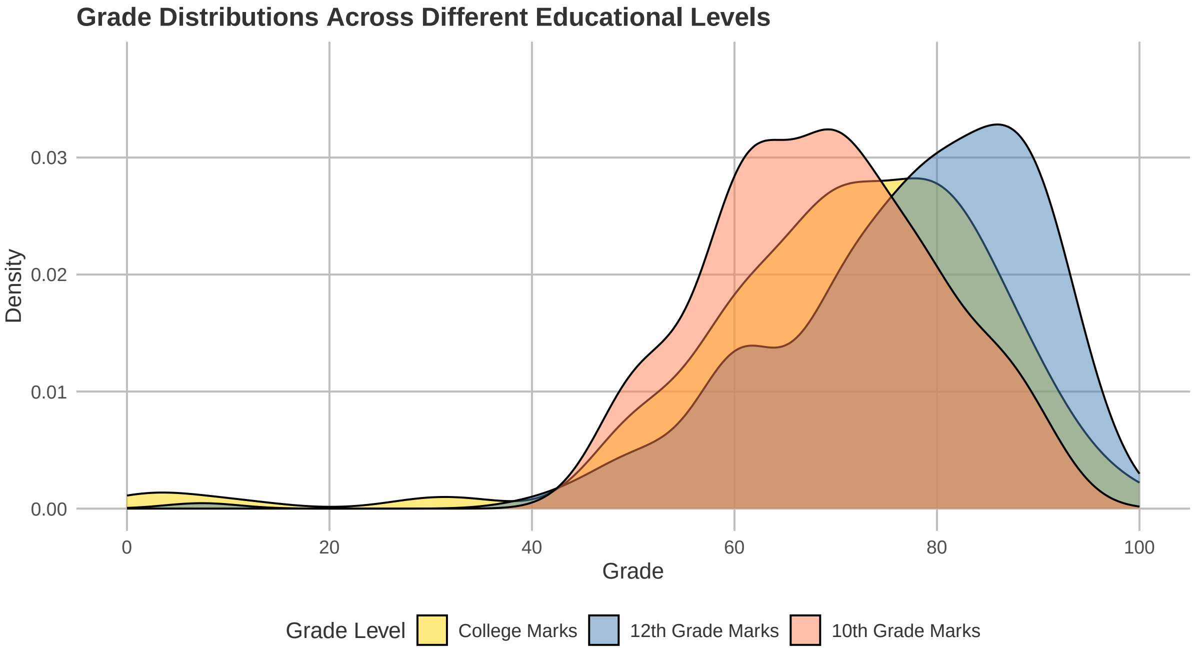
4.1 Grade Distribution Analysis Methodology

The analysis involved transforming the dataset to a long format to facilitate easier aggregation and visualization. By pivoting academic marks into a single column differentiated by educational level (10th grade, 12th grade, and college), we could efficiently generate density plots that illustrate the distribution of grades across these levels.

A custom theme was defined using `ggplot2` to ensure that all visualizations are consistent and visually appealing. This theme emphasizes clarity and readability, which are crucial for effectively communicating the findings.

4.2 Visualizing Grade Distributions

The grade distribution across different educational levels was visualized using density plots. These plots allow us to observe the frequency distribution of grades and identify patterns or anomalies such as skewness or bimodality in the data.



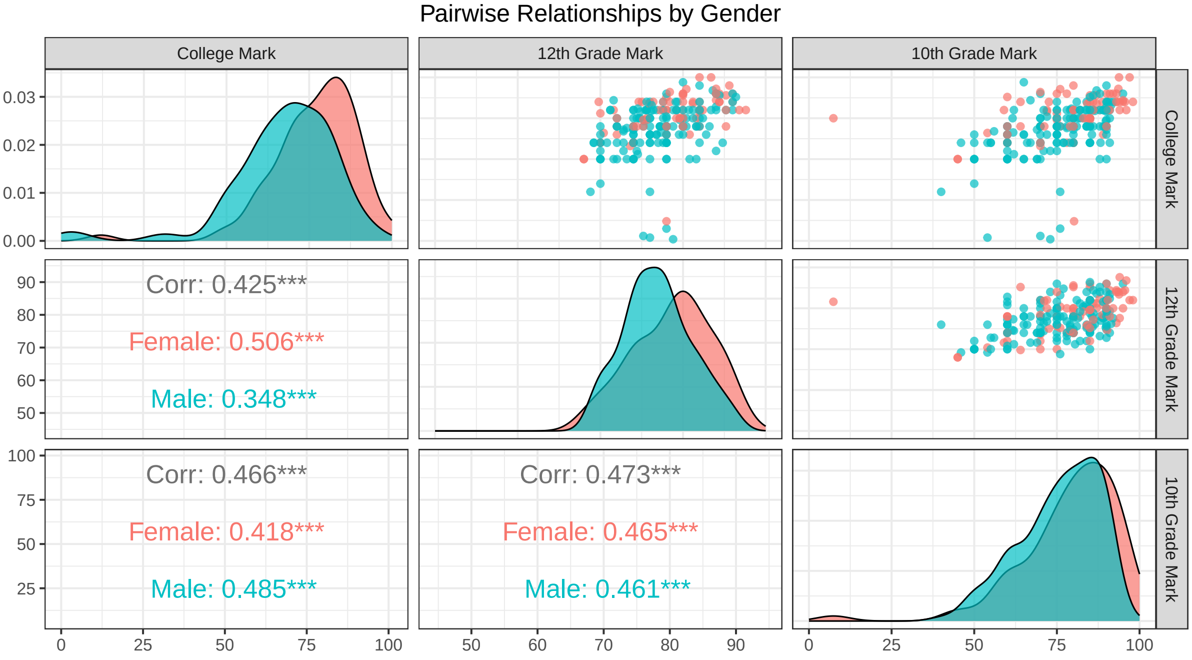
- **Consistency Across Levels**: The density plots reveal that while the distribution of marks in the 10th and 12th grades shows a moderate skew towards higher grades, college marks display a broader spread, suggesting a variation in grading criteria or student performance at this level.
- **Peaks and Tails**: Each educational level shows unique characteristics in its distribution, with 10th-grade marks generally clustering around higher values, indicating a possible leniency in scoring at lower educational levels.

4.3 Statistical Analysis Across Educational Levels

After observing the density plots which show how grades are distributed across 10th, 12th, and college levels, we conducted a Kruskal-Wallis test to statistically validate whether the differences in grade distributions across these levels are significant. The test returned a p-value of 1.123e-13, indicating strong statistical evidence that the grade distributions are indeed different across the three levels. To drill down into these differences, we performed pairwise comparisons using the Wilcoxon rank sum test. The results were telling; each pairwise comparison showed significant differences: between 10th and 12th grades (p-value = 4.4621e-14), between 10th grade and college marks (p-value = 9.3729e-07), and between 12th grade and college marks (p-value = 0.0039081). These p-values strongly support the initial visual observation that students' performance varies significantly across different educational stages.

4.4 Pairwise Relationships and Distributions by Gender

To delve deeper, we also explored how these grade distributions vary by gender. Pairwise plots were generated to visualize the relationships between different grade levels and to compare these relationships across genders.



- **Correlation Coefficients**: There is a noticeable correlation between grades at successive educational levels, with coefficients indicating moderate to strong relationships. Notably, these correlations tend to be higher among females than males, suggesting that female students may exhibit more consistency in their academic performance over time.
- **Distribution Shapes**: The plots show that the distribution of grades for female students is often tighter and more skewed towards higher grades compared to their male counterparts.

4.5 Statistical Analysis by Gender

Further statistical tests were conducted to compare grade distributions between different educational levels and to assess gender differences in academic performance:

- **Comparative Analysis**: Statistical tests confirmed the visual insights, showing significant differences in the distribution of grades between educational levels.
- **Gender Performance**: Additional testing was carried out to determine if female students consistently score higher across all educational levels. The results supported this hypothesis, indicating a statistically significant higher performance among female students in most educational settings.

Upon exploring the pairwise relationships and distributions by gender, we conducted further analyses to determine if there are statistically significant differences in grades between genders at each educational level. The normality tests showed that the data were not normally distributed, which led us to use the Mann-Whitney U test, a non-parametric test ideal for this scenario. For 10th grade marks, the test indicated a significant difference between genders (p-value = 0.014178), suggesting that gender does play a role in grade outcomes at this level. Similarly, for 12th grade and college marks, the results (p-values of 0.0026588 and 9.2326e-07, respectively) confirmed significant differences in grades between males and females.

Additionally, when testing specifically whether females scored higher than males, the results were affirmative across all educational levels with p-values of 0.0070892 for 10th grade, 0.0013294 for 12th grade, and 4.6163e-07 for college marks, providing robust evidence that female students consistently outperform their male counterparts at these stages.

4.6 Conclusion

This comprehensive analysis of grade distributions provides valuable insights into the academic performance of students across different educational levels and highlights significant gender disparities. Understanding these patterns is crucial for educational institutions aiming to address inequities and tailor interventions that support all student demographics.

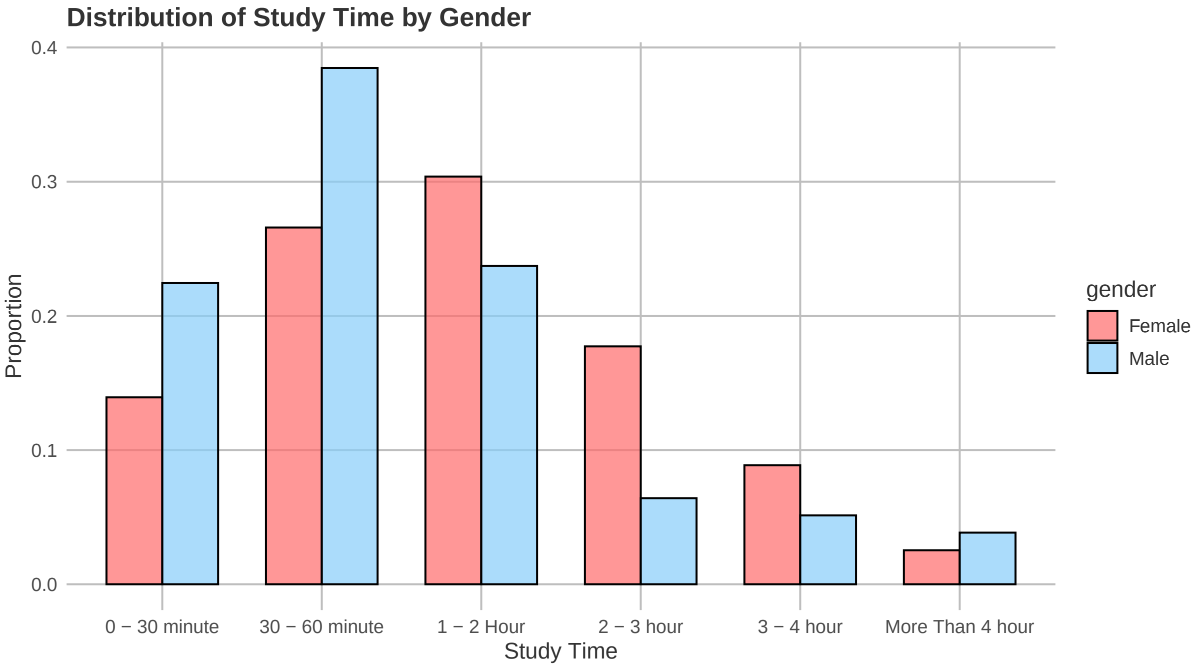
5 Variables Analysis by Gender

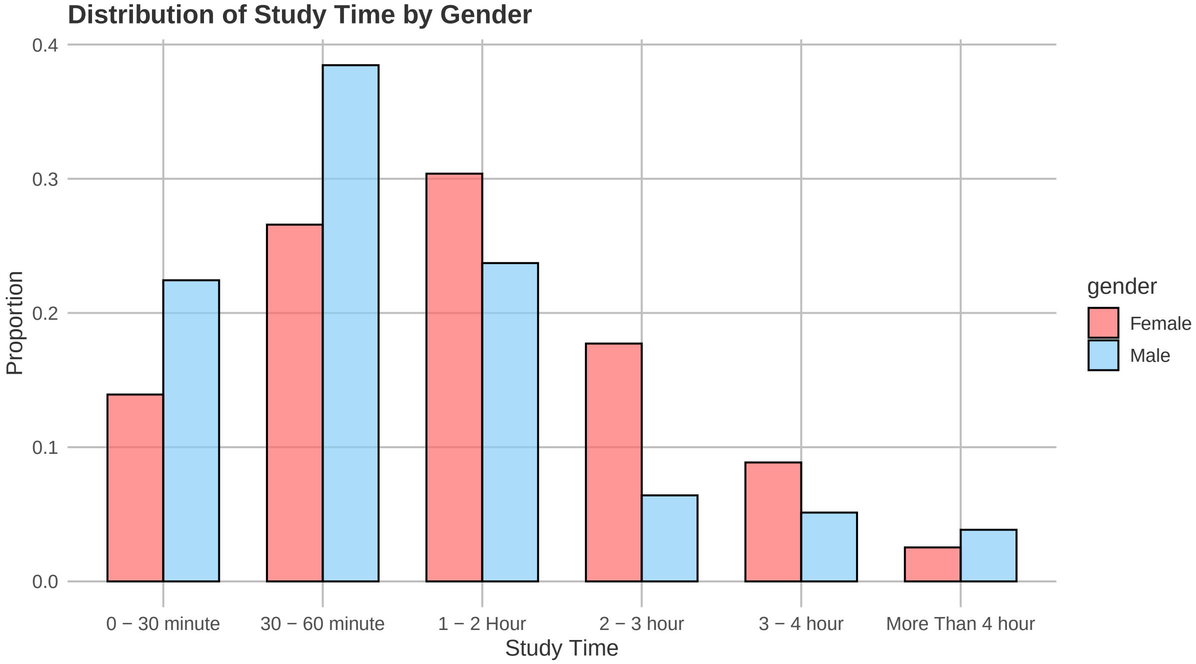
In this part of our study, we aimed to explore how various categorical variables distribute differently across genders. We specifically investigated the distributions of study time, social media time, stress levels, financial status, BMI categories, and career willingness among male and female students.

5.1 Methodology and Visual Analysis

We calculated the proportions within each category by gender to examine any potential disparities. For each category, we created bar plots to visualize these proportions, ensuring that every combination of gender and category was represented—even where there were no counts for a category in one gender. This approach helped us handle potential data sparsity effectively.

One such plot, which represents the distribution of study time by gender, highlights significant differences in how male and female students allocate time for studying. Here's the plot for **Study Time by Gender**:





This plot reveals that males tend to spend less time studying compared to females, particularly in the lower ranges of study time.

While this plot reveals visible differences in study time preferences, similar plots were generated for other variables like social media time, stress level, financial status, BMI categories, and career willingness. Each plot provided insights into the unique patterns of how different genders interact with these variables:

- **BMI Category**: The plot indicates a concerning trend where approximately 35% of students fall into overweight and obese categories. There is a statistically significant difference between genders, with a larger proportion of males being overweight or obese compared to females. This suggests gender-specific health interventions might be necessary.
- **Career Willingness**: The distributions are remarkably similar between males and females, with over 90% of students expressing at least 50% willingness to pursue a career related to their degree. This high level of career willingness across genders suggests a positive outlook towards their fields of study.
- **Travel Time**: The distribution suggests that more female students live closer to campus as their travel times are generally shorter. In contrast, the male students' travel time is less skewed, with a higher proportion traveling for more than 1.5 hours. This could influence campus engagement and needs to be considered in university planning and support services.

- **Financial Status**: The financial status across genders shows no significant differences, with most students reporting “good” financial status. However, a smaller percentage of males report “bad” financial status compared to females.
- **Social Media Time**: Both genders spend similar amounts of time on social media, with no significant differences found. The majority of students spend between 30 minutes to 2 hours on social media daily.
- **Stress Level**: Stress levels are similarly distributed across genders, with most students reporting “good” stress levels. No significant gender differences were noted here, suggesting that stress-related interventions can be uniformly applied across genders.

5.2 Statistical Analysis and Findings

To back up our visual findings with statistical evidence, we carried out Chi-squared tests for each category to test the null hypothesis that the distribution of categories is the same across genders. The Chi-squared test is particularly useful for categorical data and helps determine whether there are significant differences in the categorical distributions between two or more groups.

Three of the tested variables have resulted to have significant p-values.

- **Significant Variables**:
  - **BMI Categories**: Confirmed to be significant with a p-value of 0.0003714, suggesting that there is a disparity in the body mass distribution between males and females.
  - **Travel Time**: Showed a significant difference with a p-value of 0.02184, reflecting varying commuting behaviors.
  - **Study Time**: Highlighted significant differences in study habits with a p-value of 0.0237, suggesting that males are less likely to engage in longer study periods compared to females.

5.3 Conclusion

The analysis highlighted significant gender disparities in several key areas, including study time, BMI categories, and travel time, which are crucial for understanding student behaviors and needs. By understanding these patterns, educational institutions and policymakers can better design interventions, resources, and supports that are sensitive to the diverse needs of male and female students. These insights are vital for fostering an inclusive and supportive academic environment.

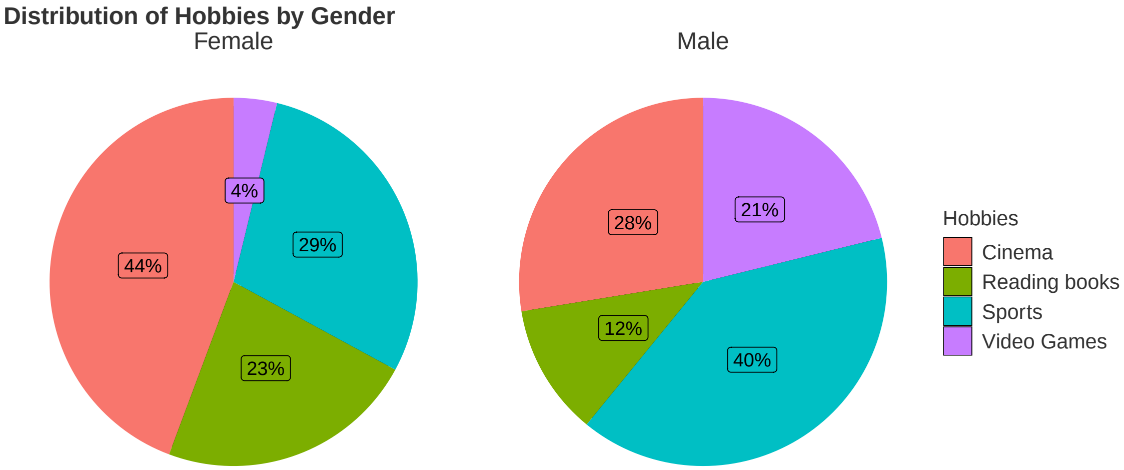
6 Categorical Variable Analysis by Gender

Our study analyzed the distribution of several categorical variables—certification, department, hobbies, like degree, and part-time job—across genders to discern any statistical differences in preferences and involvements within a student dataset.

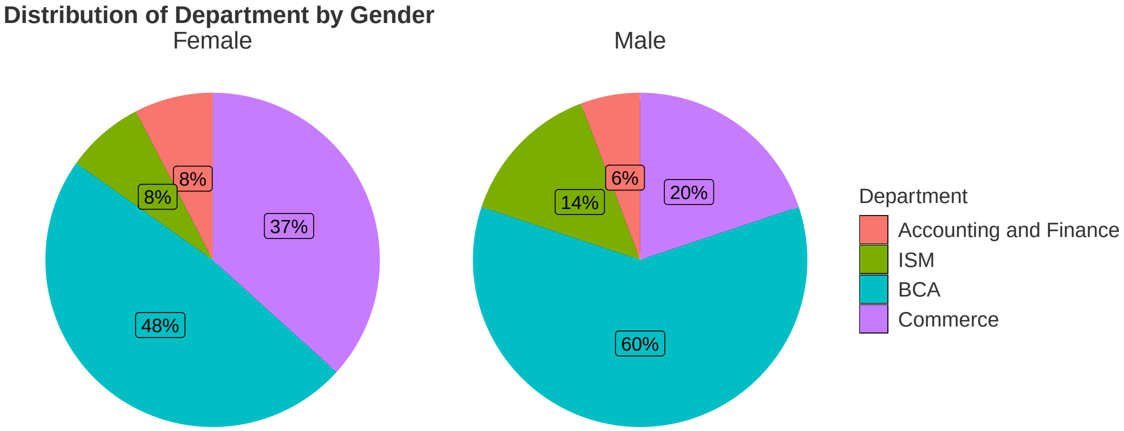
6.1 Visual Analysis and Plot Discussion

Visualizations were constructed for each category, showcasing how male and female students differ in various aspects of academic and social life. We particularly focus on "Hobbies" and "Department" as they show noteworthy patterns:

- **Hobbies by Gender**: The chart illustrates a clear divergence in leisure activities between genders. Females predominantly engage in cinema and reading, while a significant proportion of males prefer video games. This distinct variation points to gender-specific leisure preferences that could influence targeted engagement strategies.



- **Department by Gender**: This plot reveals a disparity in academic department preferences among genders. Most males are concentrated in Commerce, whereas females are more evenly spread across Accounting and Finance, ISM, and BCA. Such differences could have implications for departmental marketing and support services.



6.2 Statistical Analysis and Key Findings

We employed Chi-squared tests to validate whether the observed distributions were statistically significant:

- **Certification**: No significant difference in certification holding between genders was detected, with a p-value of 0.3256. This suggests that both genders are equally likely to pursue additional qualifications.
- **Department**: Significant differences were observed in department choices, with a p-value of 0.0251. This indicates that gender may play a role in the selection of academic fields.
- **Hobbies**: A highly significant difference in hobbies was noted, with a p-value of 0.000121, confirming gender-specific preferences in leisure activities.
- **Like Degree**: There was no significant difference in satisfaction with their degree program between genders, p-value of 0.912, suggesting similar levels of contentment across the board.
- **Part-Time Job**: The distribution of part-time job engagement showed no significant difference with a p-value of 0.05457, nearly reaching statistical significance, suggesting a trend that might warrant closer examination.

6.3 Statistical Test Outcomes

- **Significant Findings**:
  - **Department and Hobbies** showed significant differences, indicating key areas where gender influences choices and preferences.

- **Non-Significant Findings**:
  - **Certification, Like Degree, and Part-Time Job** distributions did not show significant differences, suggesting similar behaviors and choices between genders in these aspects.

6.4 Conclusion

This detailed exploration into how categorical variables distribute across genders within the student body reveals both significant and non-significant differences. Understanding these patterns allows educational institutions to better tailor their resources and support systems, ensuring that they effectively meet the diverse needs of their student population. The insights from the department and hobbies are particularly valuable for developing engagement strategies and support systems that are sensitive to gender differences, thereby fostering a more inclusive academic environment.

7 Decision Tree Modeling

7.1 Introduction to Decision Tree Analysis

Decision trees are a widely-used form of predictive modeling that are especially useful due to their interpretability and ease of understanding. They work by splitting data into branches to form a tree structure, where each decision node represents a choice between two or more paths, and each leaf node represents a predicted outcome. This makes them particularly adept at handling complex, non-linear relationships within data.

7.2 Objective of Our Analysis

In our study, we aim to understand the factors influencing student performance in college, specifically focusing on the predictive power of various student characteristics on their college marks. To do this, we will construct two decision tree models:

1. **Model Including Previous Marks**: This model considers previous academic performance (marks from 10th and 12th grades) alongside other variables.
2. **Model Excluding Previous Marks**: This model focuses solely on general characteristics like study habits, lifestyle choices, and demographic details, allowing us to explore what factors other than past academic performance might predict college success.

These models will help us identify key factors that predict college performance, which can be crucial for academic interventions and policy making.

7.3 Overview of Our Modeling Process

In our project, we employed decision trees to model and predict college marks. We used the `tidymodels` framework, specifically the `rpart` method for creating the decision trees, which allows us to visually understand what features are most influential in predicting student performance.

The dataset was initially processed to select relevant variables, such as marks from different educational stages, gender, study habits, and more. We performed a clean split of the data into training and testing sets, ensuring that both sets are representative of the overall data. This allows us to fit our model on one set of data (training) and validate its performance on unseen data (testing).

Two decision trees were constructed:

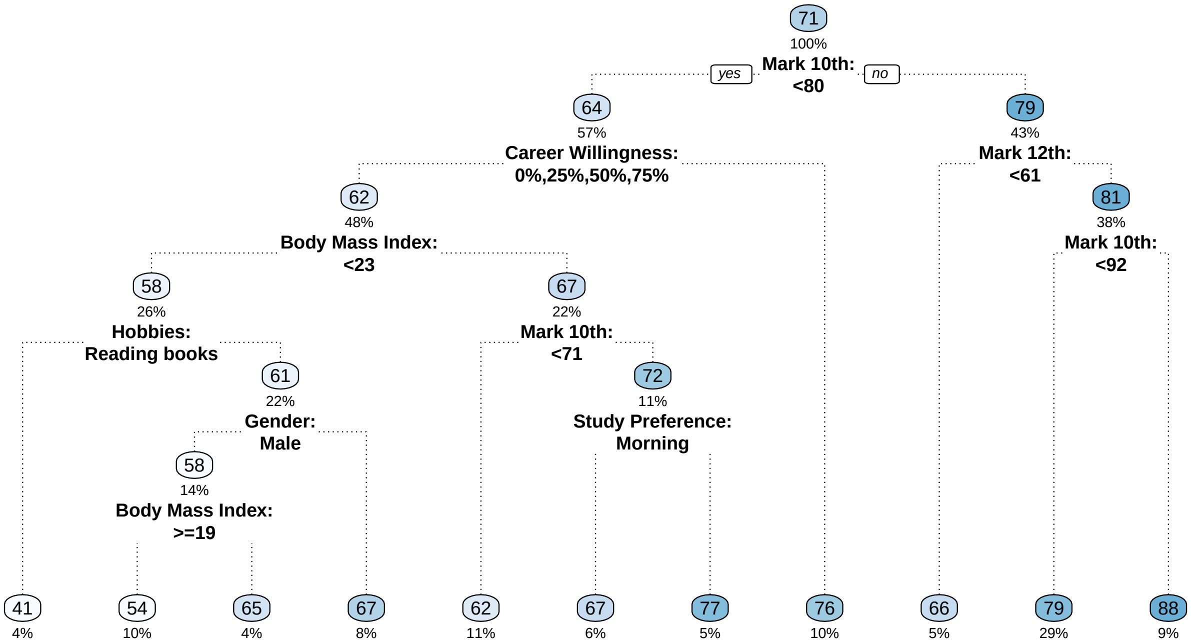
1. **Tree with Previous Grades**: This model includes marks from the 10th and 12th grades as predictors.
2. **Tree without Previous Grades**: This model excludes prior academic marks to assess the impact of other factors on college performance.

The decision trees help us visualize the decision-making process of the model, highlighting the most significant splits used to predict student marks.

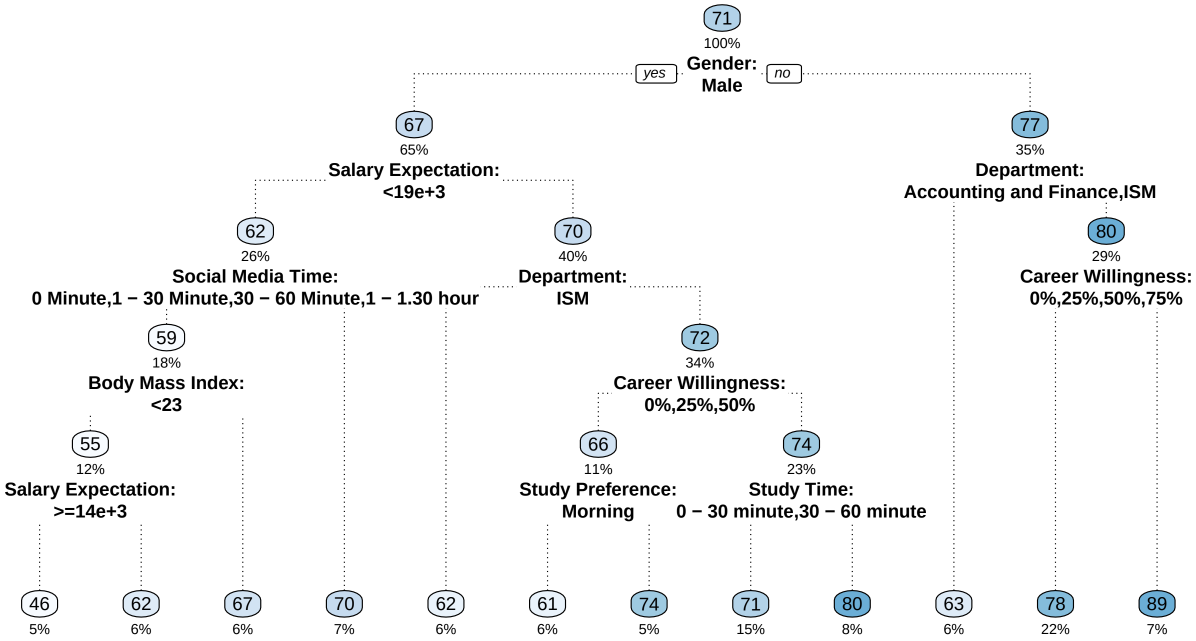
7.4 Analysis of Decision Trees and Variable Importance

7.4.0.1 Decision Tree Plots

- **Tree with Previous Grades**:
  - The primary split is on the 10th-grade marks, reflecting their strong influence on predicting college marks.
  - Subsequent splits suggest that variables like ‘Career Willingness’ and ‘Gender’ also modify predictions, indicating their secondary importance in the context where previous academic records are known.



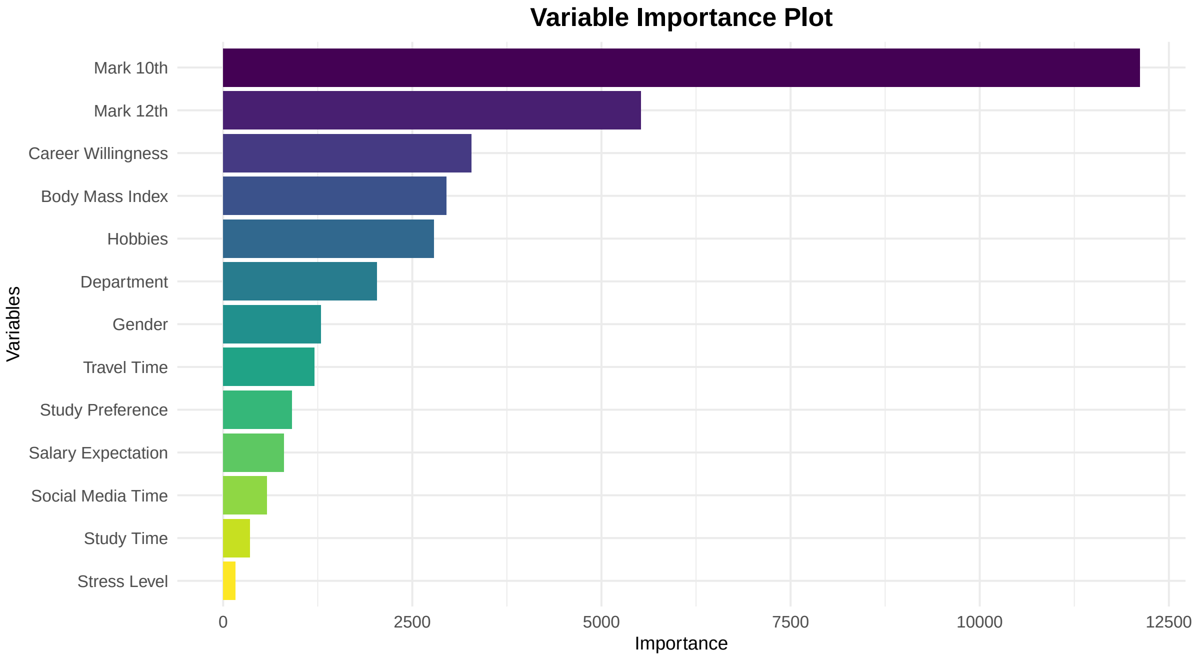
- **Tree without Previous Grades**:
  - The tree first divides on ‘Salary Expectation,’ highlighting economic expectations as a significant predictor when previous academic performance is not considered.
  - Further splits on ‘Department’ and ‘Social Media Time’ underline the impact of academic field and personal habits on college performance.



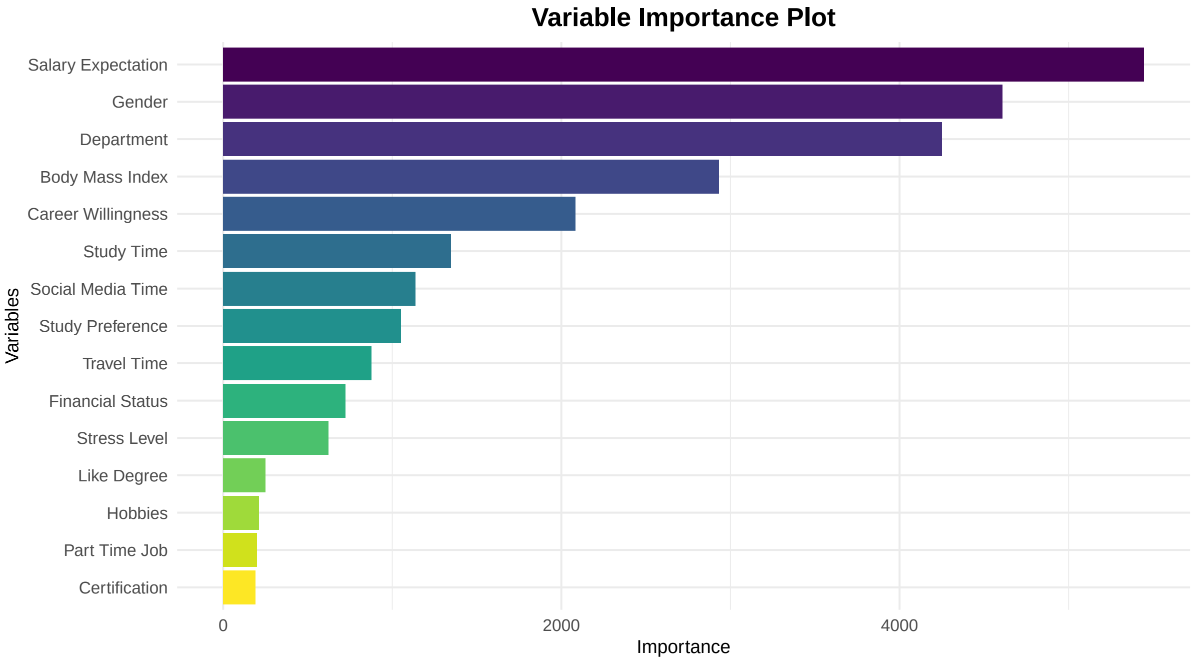
7.4.0.2 Variable Importance Plots

Variable importance in decision trees is computed based on the improvement in the model's accuracy brought by each feature at each split. Features that lead to higher gains in purity (or lower impurity) of the nodes are considered more important. These importance values help in understanding the predictive power of each feature and guide feature selection for future modeling efforts.

- **Plot for Model Including Grades**:
  - Marks from the 10th and 12th grades dominate in importance, underscoring the traditional view that past academic performance is a strong predictor of future success.
  - Other significant factors include 'Body Mass Index' and 'Salary Expectation,' suggesting a nuanced interplay between health, economic expectations, and academic outcomes.



- **Plot for Model Excluding Grades**:
  - ‘Salary Expectation’ tops the chart, followed by ‘Gender’ and ‘Department,’ pointing to socio-economic and demographic factors as critical predictors in the absence of prior grade information.
  - The prominence of ‘Career Willingness’ and ‘Social Media Time’ also indicates the relevance of motivational and lifestyle factors in student performance.



7.4.1 Model Evaluation Metrics

To assess the performance of our decision tree models, we used two common metrics: Root Mean Squared Error (RMSE) and R-squared (R^2). These metrics help quantify the accuracy of predictions and the proportion of variance in the dependent variable that is predictable from the independent variables.

- **Model with Previous Grades Metrics**:
- **RMSE**: The model achieved an RMSE of 17.36. This value represents the average deviation of the predicted college marks from the actual marks, measured in the same units as the marks themselves. The lower the RMSE, the more accurate the model is considered to be.
- **R^2**: The R^2 value for this model is approximately 0.012, indicating that about 1.2% of the variance in college marks is explained by the model. This suggests that, although the model includes historically strong predictors like 10th and 12th-grade marks, it captures only a small fraction of the factors affecting college performance.
- **Model without Previous Grades Metrics**:
  - **RMSE**: This model shows a slightly lower RMSE of 16.62, suggesting that it is slightly more accurate in predicting college marks than the model that includes previous grades.

- **R²**: The R^2 value here is approximately 0.031, which is higher than the first model. This indicates that 3.1% of the variance in college marks is explained by the model. Interestingly, this model seems to capture a bit more variability despite not having access to prior academic performance data.

The relatively low R^2 values in both models suggest that many factors influencing college performance are not captured by the variables included. This might indicate the presence of more nuanced factors not accounted for in the dataset, such as personal motivation, teaching quality, or external socio-economic factors, which are often hard to quantify.

However, the slightly better performance of the model without previous grades in terms of both RMSE and R^2 might suggest that when not overshadowed by the strong influence of past academic performance, other variables such as salary expectations, department choice, and social media habits provide valuable insights into student performance. This could be crucial for developing supportive measures that focus more on student well-being and engagement rather than solely on academic history.

7.5 Conclusion

Overall, these metrics help us understand the limitations and strengths of our models in predicting college performance. While the decision trees have provided valuable insights into the relative importance of various factors, the modest R^2 values remind us of the complexity of educational achievement and the need for broader data and more sophisticated models to fully understand and support student success in higher education.

8 Random Forest Modeling

8.1 Introduction to Random Forest

Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the average prediction of the individual trees. This method improves accuracy through the diversity of the trees and is robust against overfitting especially with a large number of trees. Random forests handle both regression and classification tasks and provide insights into feature importance.

The goal is to extend our decision tree modeling by employing Random Forests to predict college marks based on various student attributes. We build two versions of the Random Forest model:

1. **Model Including Previous Marks**: Incorporates historical academic performance to see how well past success predicts future results.

2. **Model Excluding Previous Marks**: Focuses on the influence of personal and demographic factors alone, providing insights into which traits impact academic success independent of past academic performance.

This approach allows us to compare the predictive power of a single decision tree with that of a Random Forest, offering a broader understanding of the key drivers behind student performance in college.

8.2 Model Building and Evaluation

Using the **`randomForest`** package within the **`tidymodels`** framework, we configured our Random Forest with 500 trees for robust prediction capabilities. The model fitting process involves training on our split datasets, with one model considering high school grades and the other not.

8.3 Model Performance Metrics

The performance of each model is evaluated using RMSE (Root Mean Square Error) and R-squared metrics, which provide insight into the accuracy and variance explained by the models, respectively.

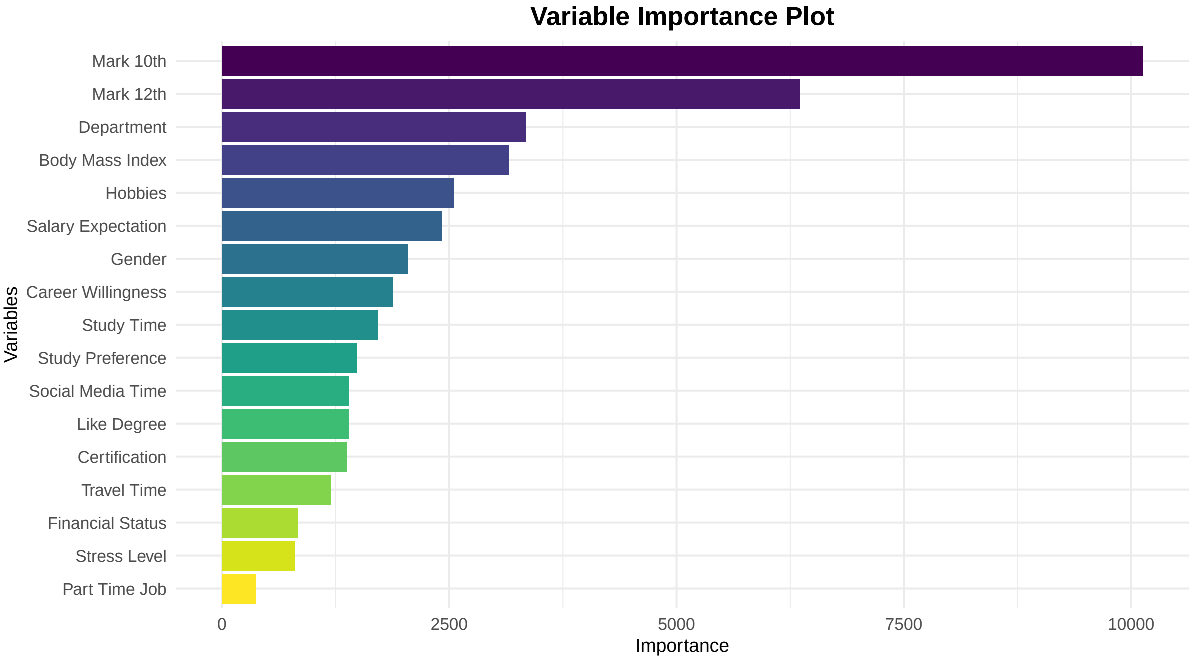
- **Model with Grades Metrics:**
  - RMSE: 14.4
  - R-squared: 0.176
- **Model without Grades Metrics:**
  - RMSE: 16.8
  - R-squared: 0.0000403

These metrics reveal that the Random Forest model including previous grades performs significantly better, suggesting that past academic performance is a strong predictor of college success.

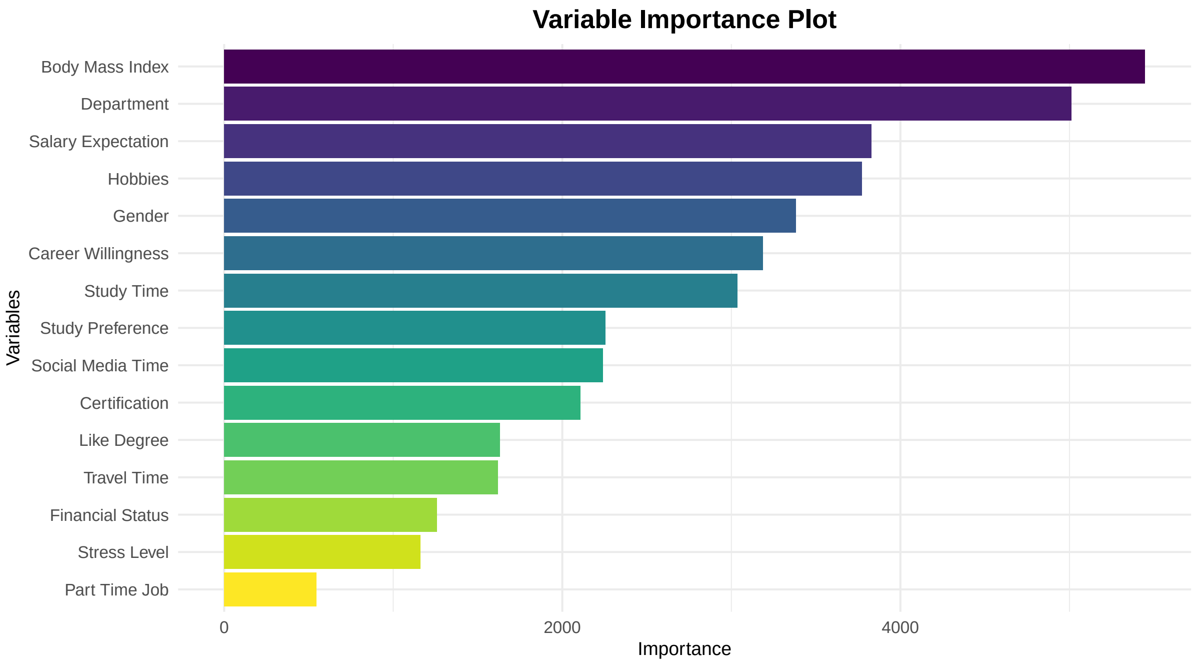
8.4 Variable Importance Plots

Variable importance plots from Random Forest models highlight which variables are most influential in predicting college marks. These plots provide a deeper insight into what factors most affect student performance, with higher values indicating greater importance.

- **Variable Importance for Model with Previous Marks**
  - **Top Influential Factors**: Previous marks (10th and 12th) dominate, followed by factors like Body Mass Index and Department, underscoring the impact of both academic history and student lifestyle.



- **Variable Importance for Model without Previous Marks**
  - **Top Influential Factors**: In the absence of grade history, Body Mass Index and Department rise in importance, along with Salary Expectations and Hobbies, indicating these factors play significant roles when previous academic performance is not considered.



8.5 Conclusion

Comparing the Random Forest models to the Decision Tree models previously discussed, Random Forest generally provides a more reliable and stable prediction due to its ensemble nature, reducing the risk of overfitting associated with single decision trees. The enhanced performance in the Random Forest model, particularly in the version including grades, also suggests better handling of complex interactions between features.

By examining both Decision Trees and Random Forests, we gain comprehensive insights into the factors influencing college success, which can guide targeted interventions to support student achievement.