1. **Initial requirements**

For this project, I have decided to conduct an analysis on behavior and performance of students based on several predicting and deciding factors. Through the research conducted in this project, I wished to explore the impact that Various factors had on determining the performance of students at a college. Through this project, I wanted to answer if the marks and expected salaries of students could be estimated based on the predictor variables (such as the daily study time of the student, certifications that the student has completed).

1. **Hypothesis**

It is crucial to understand that not all of the predicted variables have an identical effect on a student's performance overall and that various elements can have differing degrees of influence. For instance, one component might not be as effective at affecting a student's scores as another. This study focuses on these variables and investigates the best ways in which they affect the target variables as well as the connections between the predictor and the target variables. As opposed to the other predictor factors, my initial hypothesis is that "Daily Study time" and "Certifications" have a significant impact on the target variables.

Below are the key questions that the project aims at answering:

1. Does the hobby 'reading books' have a greater positive effect on college marks as opposed to the other hobbies?
2. Is there a correlation between 10th, 12th and college marks of a student?
3. Does daily study time of a student have an impact on their college marks?
4. Does the preferred time of day to study have an impact on a student’s college marks?
5. Does a student liking their degree have an impact on the student choosing their career based on the degree?
6. Does the completion of certification courses have an impact on the expected salaries of a student?
7. Does financial status have an effect on the stress levels of a student?
8. **Data Acquisition**

For the purpose of this project, a dataset has been taken from Kaggle. This dataset contains 19 attributes of over 200 students from a single university in India. Of the 19 attributes in the dataset, a subset has been taken of the most significant variables that are pertinent for the analysis.

Below are the attributes that have been taken a subset of for the analysis:

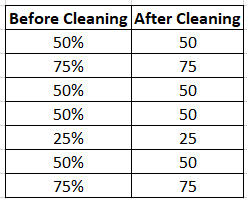
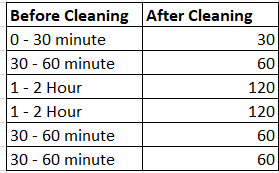
|  |
| --- |
| Certifications |
| 10th Marks |
| 12th Marks |
| College Marks |
| Hobbies |
| Daily study time |
| Prefer to study in |
| Salary expectation |
| Interest in the degree |
| Possibility of choosing  their career based on their degree |
| Stress Level |
| Financial status |

1. **Data Preparation and Cleaning**

It was necessary to modify the names of certain attributes into simpler names, for ease of analysis.

R and Excel Technologies have been used together for cleaning the data.

The attributes ‘Daily Study Time’ and ‘Career\_basedon\_degree’ have been altered as below for ease of analysis.



1. **Information Modeling**

The chief target variable for this project as per the initial hypothesis is College Marks. Hence, I have used a multiple linear regression model using the remaining variables in the dataset for predicting the college marks. This model had an R-squared value of 33.4% and an adjusted R-squared value of 27.8%.

To improve the accuracy and performance of the model, a best subset method has been used to find the most influential variables in the linear regression model, which generated a model with 7 significant variables. This model has attained a better R-squared value of 32.6% and an adjusted R-squared value of 29.6%.

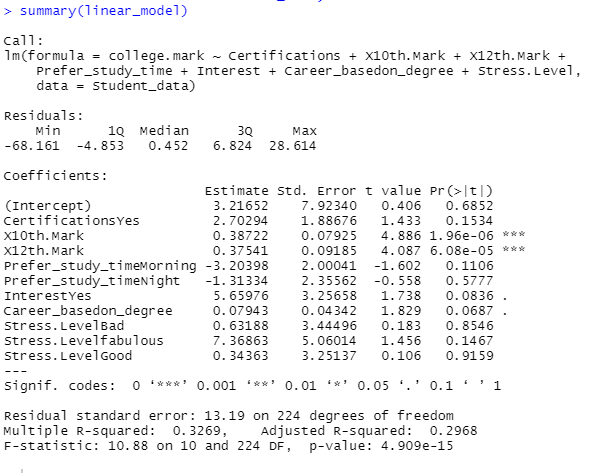


Figure 1: Summary statistics of best subset linear model

1. **Information Visualization**

RStudio has been used in this to generate the visualizations to interpret the hypothesis.

1. **Does the hobby 'reading books' have a greater positive effect on college marks as opposed to the other hobbies?**

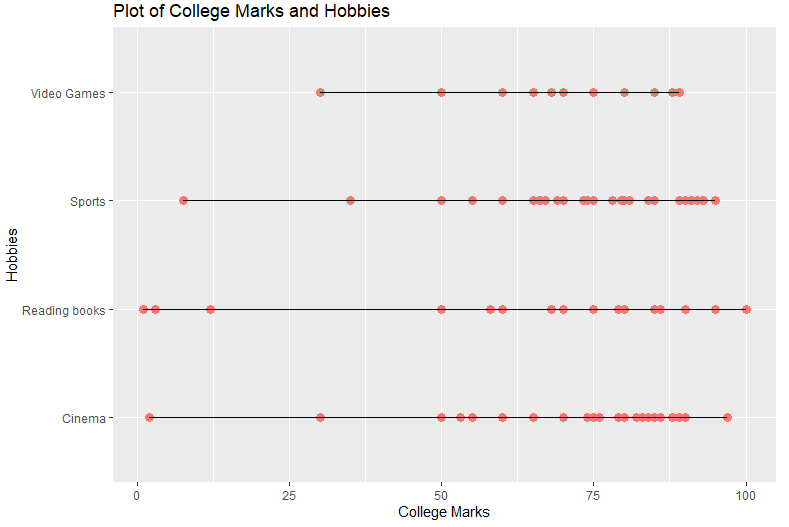


Figure 2: Plot of College Marks and Hobbies

1. **Is there a correlation between 10th, 12th and college marks of a student?**

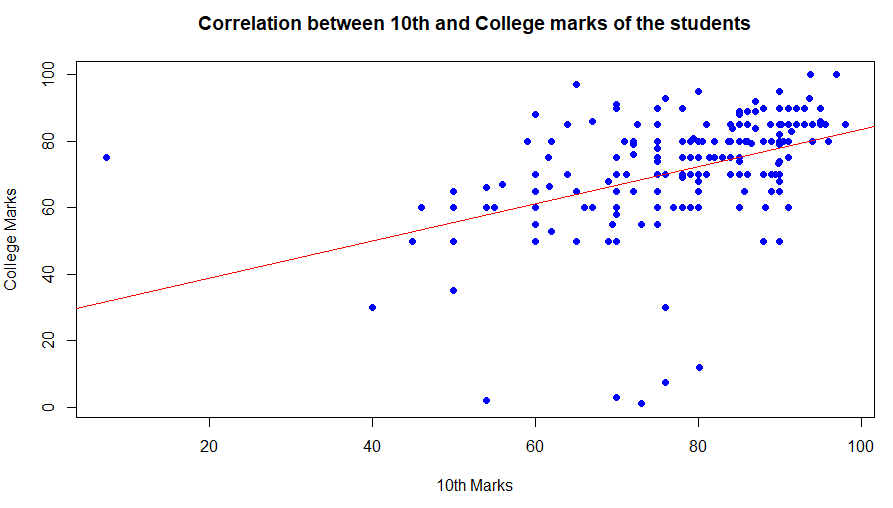


Figure 3: Correlation between 10th and College marks of the students

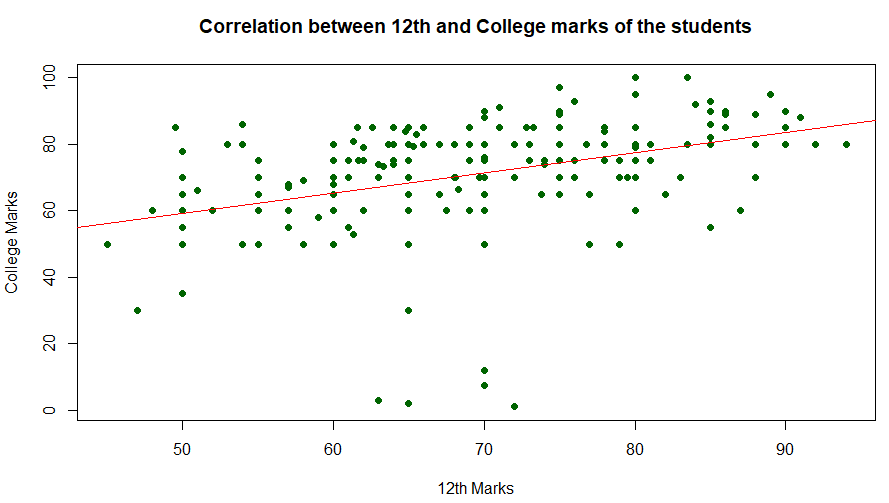


Figure 4: Correlation between 12th and College marks of the students

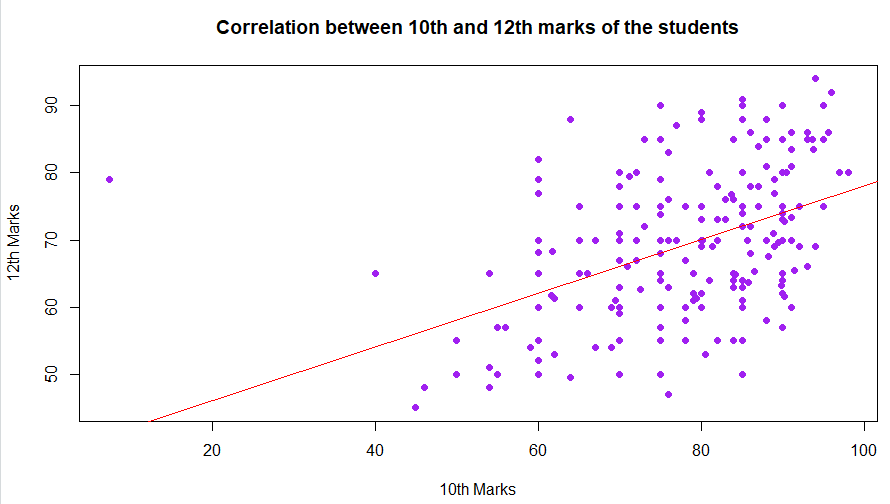


Figure 5: Correlation between 10th and 12th marks of the students

1. **Does daily study time of a student have an impact on their college marks?**

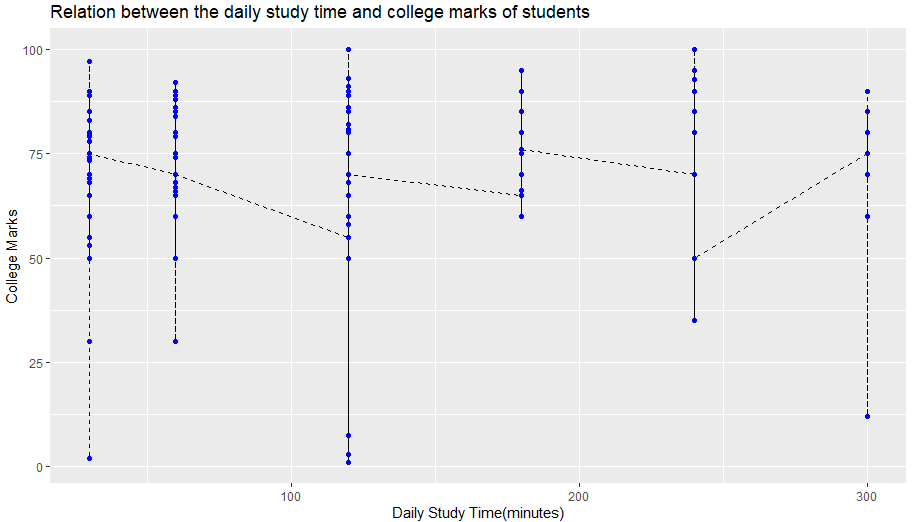


Figure 6: Relation between the daily study time and college marks of students

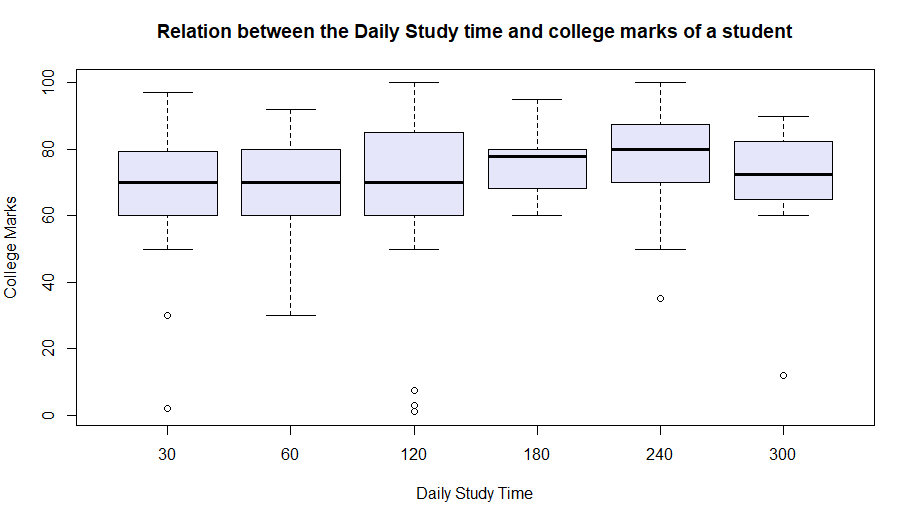


Figure 7: Relation between the Daily Study time and college marks of a student

1. **Does the preferred time of day to study have an impact on a student’s college marks?**

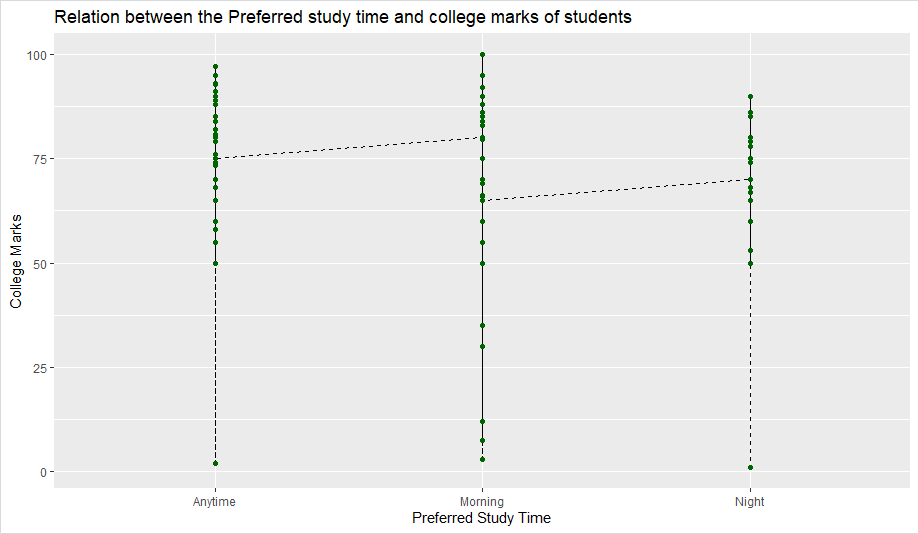


Figure 8: Relation between the Preferred study time and college marks of students

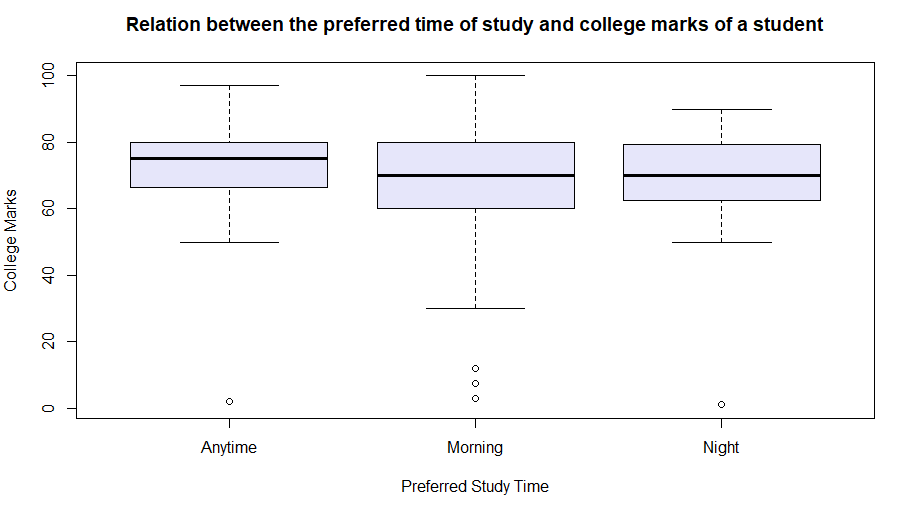


Figure 9: Relation between the preferred time of study and college marks of a student

1. **Does a student liking their degree have an impact on the student choosing their career based on the degree?**

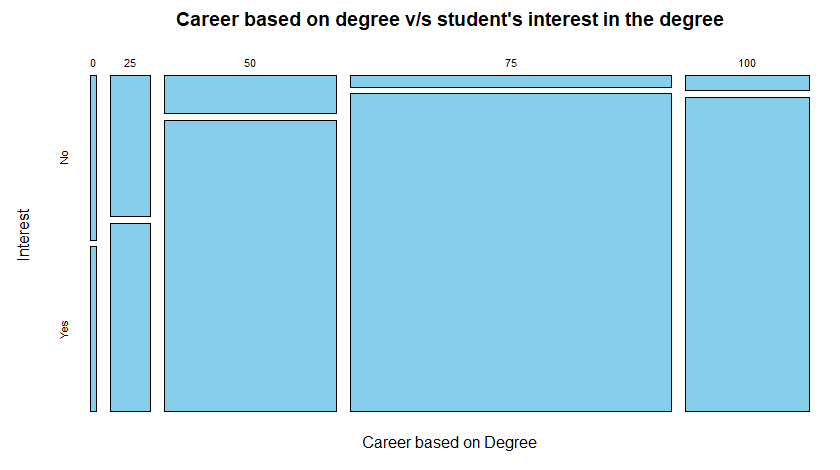


Figure 10: Career based on degree v/s student's interest in the degree

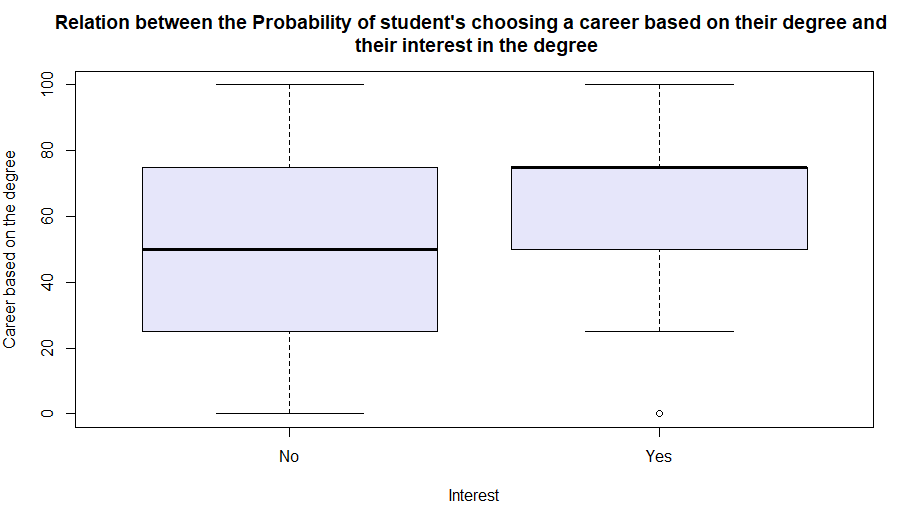


Figure 11: Relation between the Probability of student's choosing a career based on their degree and their interest in the degree

1. **Does the completion of certification courses have an impact on the expected salaries of a student?**



Figure 12: Plot of Completion of Certifications and the student's expected salary



Figure 13: Plot of Completion of Certifications and the student's expected salary (limited scale for better visualization)

1. **Does financial status have an effect on the stress levels of a student?**

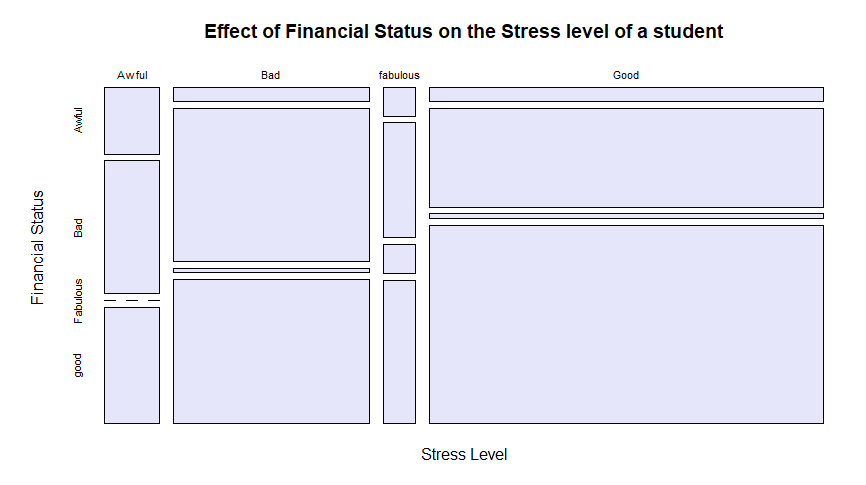


Figure 14: Effect of Financial Status on the Stress level of a student

1. **Effectiveness of Analysis / Analysis Results**

From the above visualizations answering the key questions from the initial hypothesis, it can be observed that most of the initial hypothesis has been confirmed and corroborated by the visualizations and analysis.

The below points have been proved by the visualizations:

* Although not too significant, there is a positive correlation between the hobby ‘reading books’ and a student obtaining higher marks in college.
* There appears to be a strong linear correlation between the marks that a student has obtained in their 10th, 12th and College. Meaning, if a student has obtained higher marks in one of the standards, they are bound to obtain higher marks in other standards.
* Students who study longer in a single day are more likely to score better marks than those who have lesser daily study times.
* A specific time of study does not have as much of an impact on the marks of a student as initially hypothesized.
* The interest that a student has in their degree plays a huge role in the choice that a student selects their career based on their degree. If a student is interested in their degree, there is a significant chance that the student selects their career based on their degree.
* The students who have completed Certification courses have higher expected salaries as compared to the students who have not completed any certifications.
* The financial status of a student has a major impact on the stress level that the student is under. If the financial status is good, then the student is likely to be under less stress than a student who has a poorer financial status.

The initial goals that I have set from my hypothesis have been met to a great extent by the analysis and the visualizations conducted in this study. Although there have been certain elements that have differed from what was initially hypothesized, most of the hypothesis has been proven to be true by the visualizations.

1. **Challenges**

Upon initial modeling on the dataset, I have assumed that there would be a strong linear correlation between the college marks of a student and the rest of the predictor variables in the dataset. But this has been proven to be untrue as lower accuracy than expected has been obtained from the multiple linear regression models implemented using RStudio. The initial model has obtained an accuracy of 27.8% and the improved model with 7 predictor variables has obtained an accuracy of 29.6%. These accuracies are low as per the initial hypothesis.

However, this challenge has been mitigated by the visualizations, as it was proven that there is a significant correlation between the variables and college marks and that the college marks of a student can be influenced by the target variables such as ‘Daily Study time’ and ‘Hobbies’.

1. **Conclusion**

This report could be concluded by the fact that the initial hypothesis of the project has been essentially met and proven and corroborated by the analysis and the visualizations performed, thus revealing meaningful patterns and trends in this dataset that can be applied to students elsewhere in the world to analyze their behavior and help improve their performance in college. This project re-iterates the importance of conducting structured data analysis on a dataset and the meaningful insights that can be drawn from the analysis.