# Hypothesis-Testing Factors Affecting Students' Academic Performance

Urmika Kasi
Department of Information Science, BMS College of Engineering
Bangalore, India

Shreyas Prasad
Department of Information Science, BMS College of Engineering
Bangalore, India

## ABSTRACT
Examining the factors affecting students' academic performance is a significant aspect of consideration as it can improve teaching and learning processes. Numerous academic and non-academic facets affect student performance, such as study time, frequency of absences, recreational activities, and interpersonal relationships. Educational data mining (EDM) and learning analytics (LA) are two closely related fields that reveal useful information from educational databases to generate actionable insights. This paper investigates the aforementioned feature sets by hypothesizing the impact of these factors on student performance based on existing studies and employs a combination of LA and EDM techniques to test the hypotheses. Experimental results show that the presented hypotheses were consistent on nearly all accounts. The insights of this study can be used to bolster student performance even beyond an academic scope by educational policy improvement.

## Keywords
Educational Data Mining, Learning Analytics, Student Performance

## 1. INTRODUCTION
Students have a significant impact on a country's social and economic growth [1] and, a student's success is directly correlated with his/her academic performance. Students' performance is also the main concern of various stakeholders, including educators, administrators, recruiting agencies, and corporations. A student's performance is generally quantified into a grade or Grade Point Average (GPA) scheme based on several academic metrics set by the educational institution and is affected by many psychological, economic, social, and personal factors.

The advent of information technology has led to vast amounts of educational data stored in different formats, requiring a proper method of extracting knowledge from large repositories for better decision-making. As educational data mining is used to improve educational policymaking using data-based models, it has emerged as a vital research area to reveal presentable and applicable knowledge from large educational data repositories[2]. Learning analytics is a closely related field that deals with developing methods that lead to effective educational data sets to support the learning process[3]. By employing these both techniques, student performance can be analyzed to distinguish factors that determine a good learning environment or precondition for learning.

This research attempts to systematically hypothesize the effect of students' personal, social, academic, and economic factors on their academic performance through an extensive literature survey. A combination of educational data mining and learning analytics is used to analyze educational data to identify students' patterns of behavior and provide actionable knowledge to improve learning and learning-related processes.

This paper is organized as follows: after this introduction in Section 1, the literature survey is explained in Section 2. The hypothesis is presented in Section 3, while in Section 4 and 5, the research methodology and results are presented. Finally, the paper is concluded in Section 6.

## 2. LITERATURE SURVEY
This review is used to hypothesize factors affecting student performance based on past research in this field. It also identifies strengths in the existing literature and highlights the unique contribution that the study makes to learning analytics and educational data mining.

Several studies have been conducted to investigate the factors affecting students' final grades using learning analytics and educational data mining principles. Academic determinants such as student absences, time dedicated to studying daily, and past class failures have affected student performance. In one study [4], the researchers found that four factors are positively related to students' achievement i.e., demographic, active learning, students' attendance, and involvement in extracurricular activities. Authors of [5] have also supported this and conclude that attendance and study time allocated affects students' success. Travel time is also directly and significantly associated with GPA [6]. Students who had reliable travel times tended to have lower stress levels and frequencies of being late to class.

Studies suggest that non-academic aspects such as family, personality, social interaction, and demographic situations have also significantly influenced students' academic achievement and GPA. Authors of [7] have revealed a positive and statistically significant impact of learning facilities, communication, skills, and proper guidance from parents on students' academic performance. Several studies [5][8][9] correlated socio-economic factors and familial features such as family income, parents' age and educational backgrounds with academic excellence. Specifically, [10] has shown that family income, parents' age, and parents' education are significantly related to student performance. Utilizing Naive Bayes' model, it has been found that factors like mother's qualification and income of the family are highly correlated with the performance of the student.

More studies have shed light on the effects of recreational activities and interpersonal relationships with people within their immediate surroundings on a student's achievement. Students' social behavior includes facets such as the frequency of alcohol consumption, going out with friends, and extra-curricular activities. Researchers showed that frequent

consumption of alcohol hurts student GPA [11]. Interestingly, the effect of students' involvement in romantic relationships on GPA is unclear since it has not been systematically investigated [12].

## 3. HYPOTHESIS

The research hypotheses discussed in the following paragraphs are based on theoretical reasoning and results from previous studies, as explained in the Literature Survey section of this study. Given these considerations, we formed the following hypotheses:

1. Students who consume alcohol, have more absences and have longer travel times have a negative correlation with grades.
2. Students who have fewer failures and longer study times have a positive correlation with grades.

## 4. RESEARCH METHODOLOGY

This study demonstrates achievement as a function of school resources, student ability, student socioeconomic background, and other characteristics. For this study's purposes, sample data [13] including a wide range of attributes, including demographics, grades, academic grades, and social factors were considered. Visualization and statistical techniques were employed to generate insights and inferences

### 4.1 Cleaning and Pre-processing

The quality of data is critical for data mining algorithms to perform accurately. In data quality assessment, rows with missing data were eliminated, and null values were estimated using various interpolation methods. Type checking was done to ensure that all the related data types were consistent before encoding categorical values into numerical data. Other inconsistent data, such as capitalizations and trailing white spaces, were corrected.

### 4.2 Encoding

It is essential to convert categorical variables into numerical values as most machine learning algorithms require that input and output variables be numerical. Ordinal Encoding was employed to deal with ordinal variables. The binary values were encoded to a 1 or 0, e.g., Yes/No, Male/Female.

### 4.3 Visualization

Before visualizing attribute relationships, the data's distribution and skewness were inspected to check its correspondence with a normal distribution. A skewness test hypothesizing the sample population's distribution revealed that G1, G2, and G3 were normally distributed since they had low p-values (less than 0.05) and low Z-scores (less than 2). Box plots were mainly chosen to visualize the relationships since they are more informative regarding outliers, mean, median, maximum, and minimum values than other forms of visualization such as scatter plots.

## 5. RESULTS

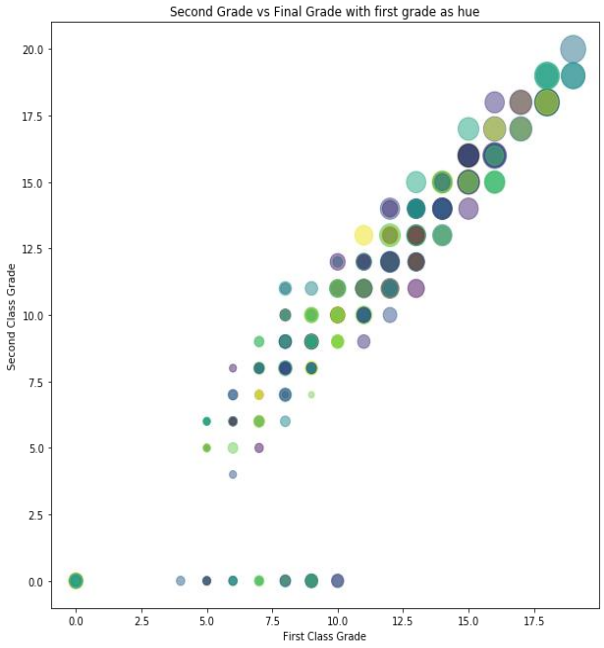


**Fig 1: Scatter plot of the three grades**

The final grade increases with an increase in second class grade. The growing size of the dots indicates an increase in first-class grade. The plot indicates a positive linear correlation between the second class grade and the first-class grade. Hence, this indicates a strong correlation between the three grades.

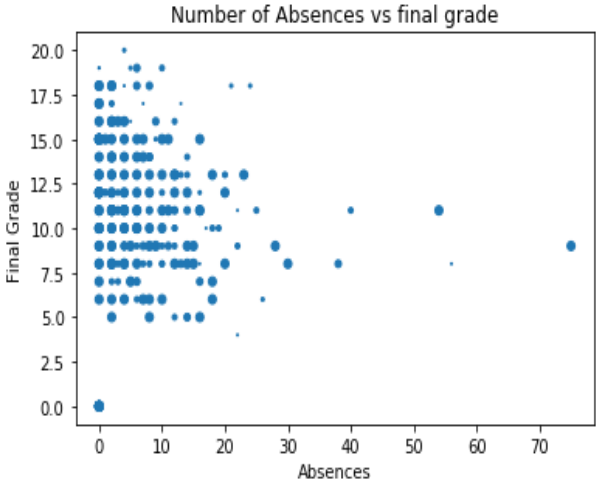


**Fig 2: Scatter plot of absences vs final grade**

The plot indicating final grade as a function of the frequency of absences tends to get more sparse with an increase in absences, indicating that students with fewer absences have better grades.

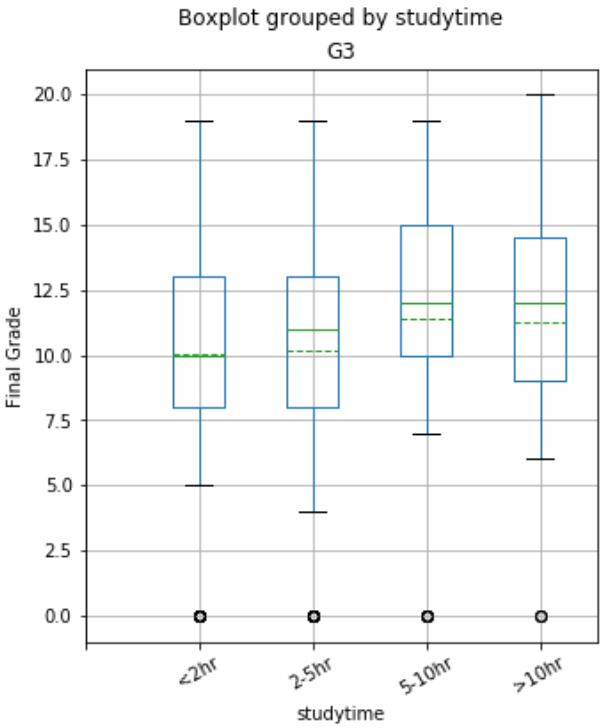


**Fig 3: Box plot of study time vs final grade**

The study time is grouped into brackets of under 2 hours, 2-5 hours, 5-10 hours, and over 10 hours. With more study time, the average increases. The maximum and minimum score increases showing a positive linear correlation. Though students who study for greater than 10 hours show higher maximum scores than those who study for 5-10 hours, the average is nearly the same.

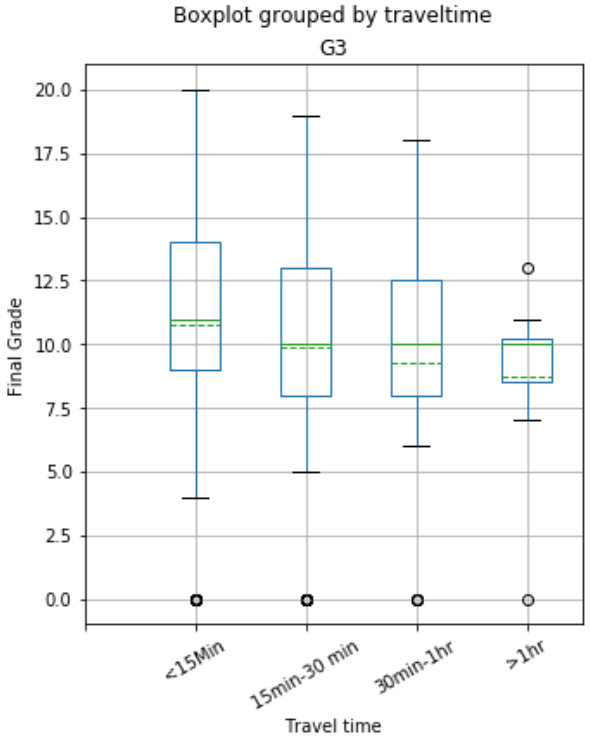


**Fig 4: Box plot of travel time vs final grade**

Students with less travel time have a higher average grade. With more travel time, the average reduces, and most of the scores are nearer to the mean. The average score reduces with an increase in the amount of travel time. Thus, travel time and final grades have a negative correlation.

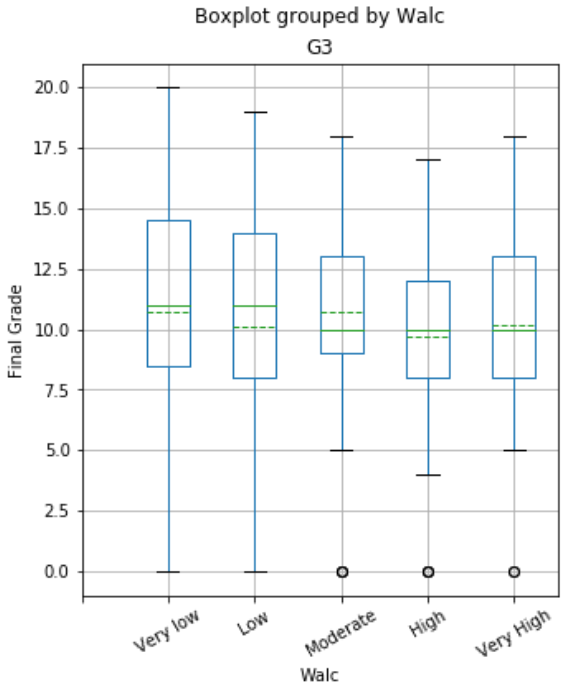


**Fig 5: Box plots of weekend alcohol vs final grade**

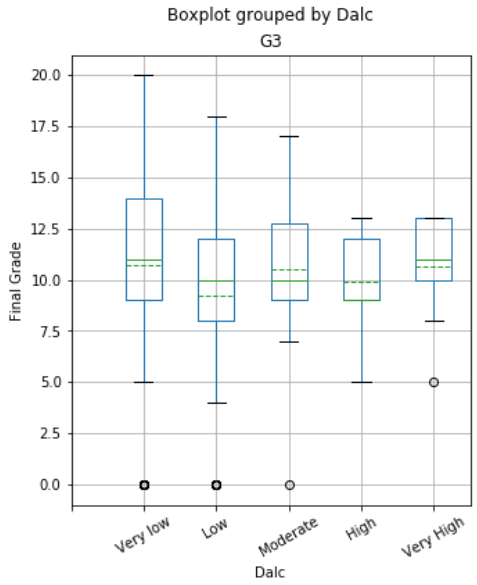


**Fig 6: Box plots of weekend alcohol vs final grade**

Students who consume more alcohol during weekdays get lower maximum scores increasingly. However, it is worth noting that the 'low' category has the lowest average, which could be attributed to more outliers on the lower side. Interestingly, those with the highest consumption have the least minimum score with zero or few outliers, unlike those with lower consumption whose scores are more widespread from the mean with more outliers. The average and maximum score decrease with increased alcohol consumption over the weekend. The deviation is not as significant as it is with

weekday consumption. With higher consumption, scores are less widespread from the mean but with more outliers on the lower end.

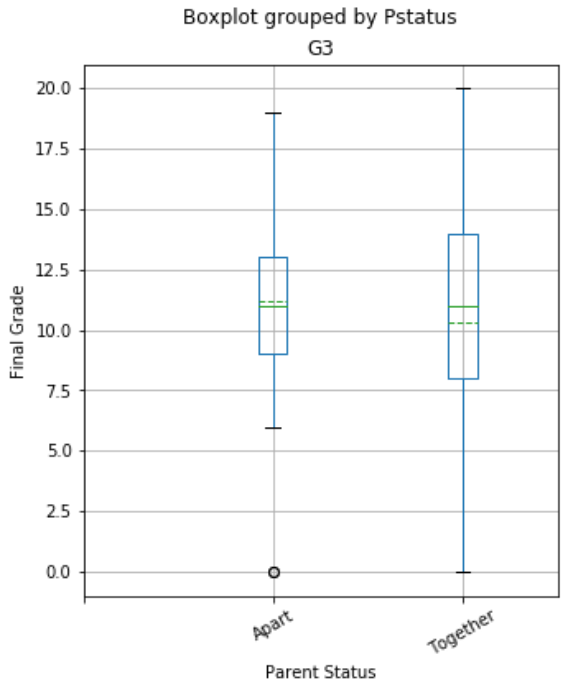


**Fig 7: Box plots of parent's status vs final grade**

The student average is seen to be lower for students whose parents are together. The scores tend to get more widespread for students whose parents live together, with a higher maximum and lower minimum score.

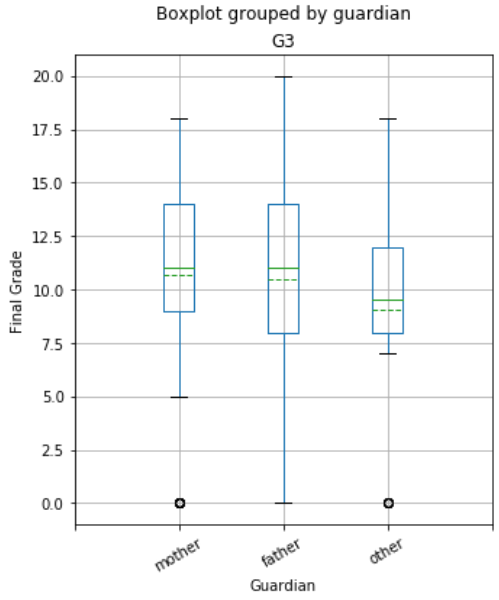


**Fig 8: Box plots of guardian vs final grade**

The box plots indicate the influence of the type of guardian on student performance. There is a larger variance of scores for students raised by their fathers than those raised by their mothers or other guardians. Students raised by their parents have a better average score than those raised by other guardians.

## 6. CONCLUSION

This study employed mining, visualization, and analytical techniques to determine the effect of various social, economic, demographic, personal, and academic factors on students' academic performance. Based on previous work in this domain, the influence of non-academic factors such as alcohol consumption and involvement extracurricular and social activities, and academic features such as frequency of absences and failures and study time on student GPA was hypothesized. To test these hypotheses, sample data of student performances and attributes affecting the same was considered. Upon processing and analyzing the sample data, this research indicated the hypotheses formed were consistent on nearly all accounts. The scatter plots and box plots constructed from the processed data confirmed the expected positive linear correlation between students' final grades and increased study time, fewer failures, and previous grades. These results also indicated the expected negative correlation between students' final grades and alcohol consumption and increased absences. This research also investigated the effects of other social and familial aspects such as romantic relationships on academic achievement that could not be hypothesized due to inadequate research in the area. Students involved in romantic relationships showed lower final scores than students who were not. Furthermore, students raised by both their parents showed higher final scores than those raised by other guardians.

By identifying the major contributors to students' grades, actions can be considered to bolster their performance effectively beyond an academic scope. This study's insights can help minimize failure ratios and improve students' overall quality of satisfaction, behavior, and learning outcomes.

## 7. LIMITATIONS AND FUTURE WORK

Sophisticated algorithms and mining techniques can be applied to the same dataset to perform visualization, classification, association, clustering, and prediction tasks. This would enhance the meaningfulness of the insights generated from the current analysis for prescriptive uses. Analytical dashboards can be constructed from this data to tailor insights to different end-users for specialized applications. This data can also be used in deep learning applications, for instance, in students' dynamic profiling based on personal characteristics. The analysis can be expanded to generalize the results by considering various datasets spanning different geographical boundaries. Since such data tends to be highly localized, the results may vary for different datasets. By considering student performance data over various age groups and courses, a better understanding of results can be obtained through comparisons.
plate.

## 8. REFERENCES

[1] Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017, April). Predicting student performance using advanced learning analytics. In Proceedings of the 26th international conference on world wide web companion (pp. 415-421).

[2] Educational Data Mining, *Wikipedia*, Retrieved from https://en.wikipedia.org/wiki/Educational_data_mining

[3] Learning Analytics, *Wikipedia*, Retrieved from https://en.wikipedia.org/wiki/Learning_analytics

[4] Ali, N., Jusof, K., Ali, S., Mokhtar, N., & Salamat, A. S. A. (2009). THE FACTORS INFLUENCING

STUDENTS'PERFORMANCE AT UNIVERSITI TEKNOLOGI MARA KEDAH, MALAYSIA. Management Science and Engineering, 3(4), 81-90.

[5] Wu, Q. (2014). Associations between travel behavior and the academic performance of university students

[6] Singh, S. P., Malik, S., & Singh, P. (2016). Research paper factors affecting academic performance of students. Indian Journal of Research, 5(4), 176-178.

[7] Okpala, C. O., Okpala, A. O., & Smith, F. E. (2001). Parental involvement, instructional expenditures, family socioeconomic attributes, and student achievement. *The Journal of Educational Research, 95(2),* 110-115.

[8] Egalite, Anna J. "How family background influences student achievement: can schools narrow the gap?." Education Next 16.2 (2016): 70-79.

[9] Devasia, Tismy, T. P. Vinushree, and Vinayak Hegde. "Prediction of students performance using Educational Data Mining." 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). IEEE, 2016.

[10] Pritchard, Mary E., and Gregory S. Wilson. "Using emotional and social factors to predict student success." Journal of college student development 44.1 (2003): 18-28.

[11] Giordano, P. C., Phelps, K. D., Manning, W. D., & Longmore, M. A. (2008). Adolescent academic achievement and romantic relationships. Social Science Research, 37(1), 37–54.

[12] *Github*, Retrieved from https://raw.githubusercontent.com/arunk13/MSDA-Assignments/master/IS607Fall2015/Assignment3/student-mat.csv