Statistical Report
—
Student Performance

Nhat Quang Bui - BS23DSY049

Kien Tri Le - BS23DSY041

Nhan Thanh Bui - BS23DSY047

Statistics and Probability - Prof. Suchismita

BDS 2023
SP Jain School of Global Management

December 9, 2023

Acknowledgement

We would like to express our gratitude to those who contributed to the completion of this Assignment. Special thanks to Professor Suchismita Das for valuable guidance and expertise on the technical of the research. Additionally, thanks to our members Quang, Kien, Nhan for collaborative efforts on providing resources.

Contents

1 Introduction

1.1 Report Objectives

This report examines a dataset of 100 student records, selected from a larger set, to analyze how various socio-economic and educational factors influence academic performance, particularly mean grades. Our analysis will highlight key statistics and explore correlations, aiming to provide insights into the determinants of student success.

1.2 Problem Statement

In the competitive education environment nowadays, it's essential for students, parents, and schools to understand which factors most significantly impact student performance. This dataset provides insights into how various socio-economic and personal factors correlate with students' mean grades.

Our analysis, using Microsoft Excel and relevant Python statistical library, aims to identify key determinants of academic success.

This will guide students in optimizing their educational strategies and assist institutions in recognizing and supporting students effectively.

2 Overview

This dataset is built to help students, parents, and schools to understand which factors most significantly impact student performance.

In this report, we carefully choose factors that can potentially provide novel insights, including family size, alcohol consumption, reason for choosing school, ... We also did several hypothesis testings to compare **MeanGrade** between groups. To perform that, we carry out statistical techniques such as data cleaning, central tendency assessment, t-test, ANOVA, ...

This dataset is inspired by Dev Ansodariya from Kaggle.

3 Data Cleaning

The initial step in our analysis involved cleaning the data to ensure the integrity and relevance of our dataset. This process was crucial for accurate and meaningful statistical analysis. Here's an overview of the steps taken:

- **Removing Null Values:** To maintain data quality, we removed all records with null values.
- **Refining Dataset:** We identified and removed certain columns ("sex", "age", "failures", "nursery", "higher") that were either unclear, clichéd, or not directly relevant to our core objective of understanding academic performance determinants.
- **Calculating Mean Grades:** We calculated the mean of the last three grade columns to get a consolidated measure of academic performance. This mean grade serves as a key indicator for our analysis.
- **Standardizing Responses:** We observed inconsistencies in responses (such as "yes" and "y", "no" and "n") in columns like "schoolsup", "famsup", "activities", "internet", and "romantic". To address this, we standardized these responses to "yes" and "no" for uniformity and clarity.
- **Reordering Columns:** For better readability and logical flow, we rearranged the columns in a new order to draw a line between objective and subjective factors.
- **Exporting Cleaned Data:** The cleaned and restructured data was then saved to a new file, "student_data_processed.csv", ensuring that all subsequent analysis was based on this refined dataset.

4 Data Exploration

4.1 Dataset Dictionary

school: Type of school the student attends.

address: Urban or rural area where the student lives.

famsize: Size of the student's family.

Pstatus: Parental cohabitation status.

Medu: Mother's education level.

Fedu: Father's education level.

Mjob: Mother's job.

Fjob: Father's job.

guardian: Student's guardian.

traveltime: Travel time to school.

schoolsup: Extra educational support at school.

famsup: Family educational support.

internet: Access to the Internet at home.

famrel: Quality of family relationships.

romantic: Involvement in a romantic relationship.

activities: Participation in extracurricular activities.

studytime: Weekly study time.

reason: Reason for choosing the school.

freetime: Free time after school.

goout: Going out with friends.

Dalc: Workday alcohol consumption.

Walc: Weekend alcohol consumption.

health: Current health status.

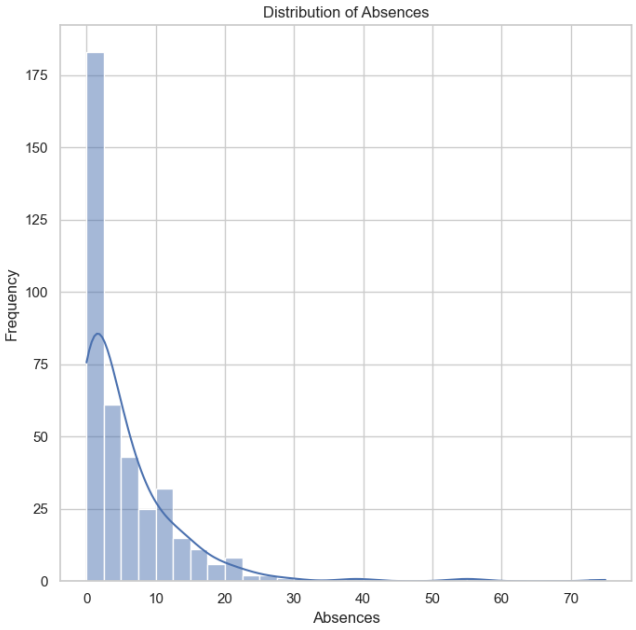
absences: Number of school absences.

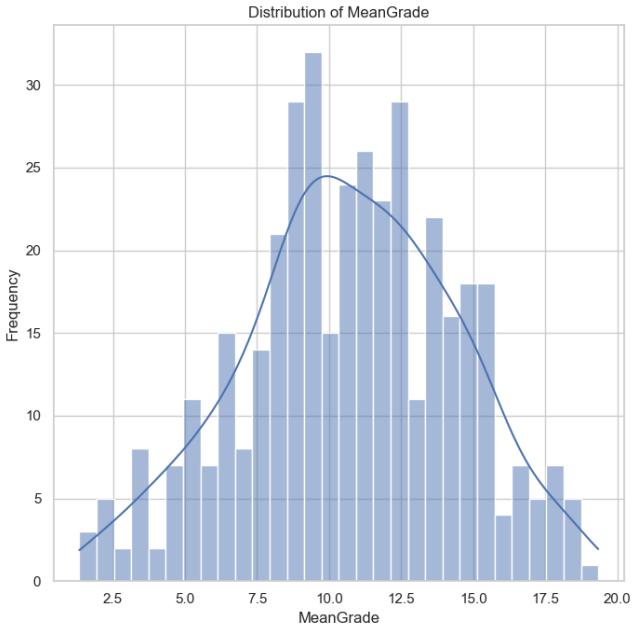
4.2 Descriptive Analysis

	count	mean	std	min	25%	50%	75%	max
Medu	395	2.75	1.09	0	2	3	4	4
Fedu	395	2.52	1.09	0	2	3	3	4
Traveltime	395	1.45	0.70	1	1	2	2	4
Famrel	395	3.94	0.90	1	4	4	5	5
Studytime	395	2.04	0.84	1	1	2	2	4
Freetime	395	3.24	1.00	1	3	4	4	5
Goout	395	3.11	1.11	1	2	3	4	5
Dalc	395	1.48	0.89	1	1	1	2	5
Walc	395	2.29	1.29	1	1	3	3	5
Health	395	3.55	1.39	3	4	5	5	5
Absences	395	5.71	8.00	0	0	4	8	75
MeanGrade	395	10.68	3.70	1.33	8.33	10.67	13.33	19.33

Table 1: Descriptive Statistics of the Dataset

5. Graphical data analysis





For absences:

- Mean: 5.71
- Median: 4.00
- Mode: 0
- Range: 75
- Variance: 64.05
- Standard Deviation: 8.00
- Interquartile Range (IQR): 8.00

For MeanGrade:

- Mean: 10.68
- Median: 10.67
- Mode: 9.00
- Range: 18.00
- Variance: 13.67

Standard Deviation: 3.70

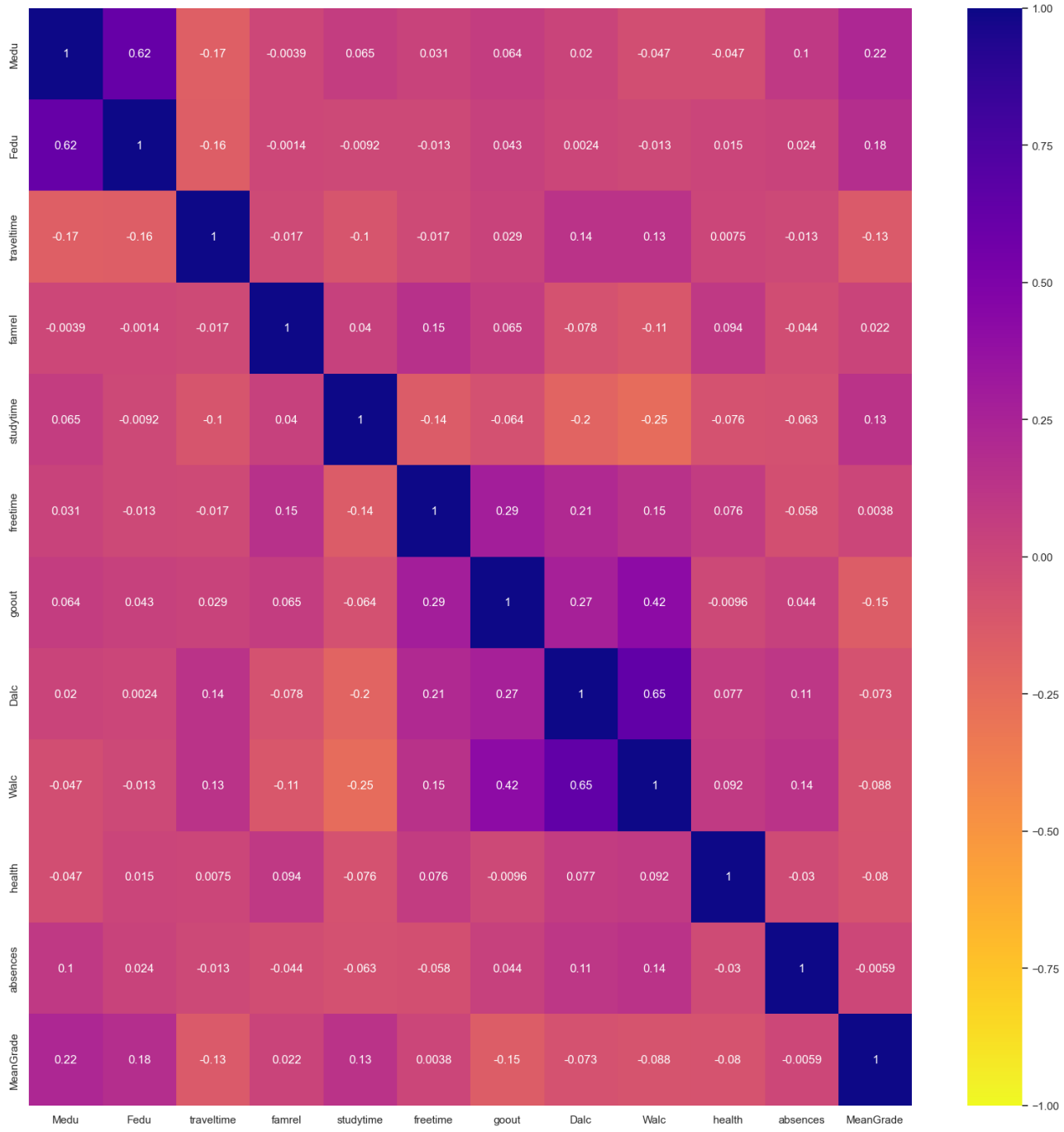
Interquartile Range (IQR): 5.00

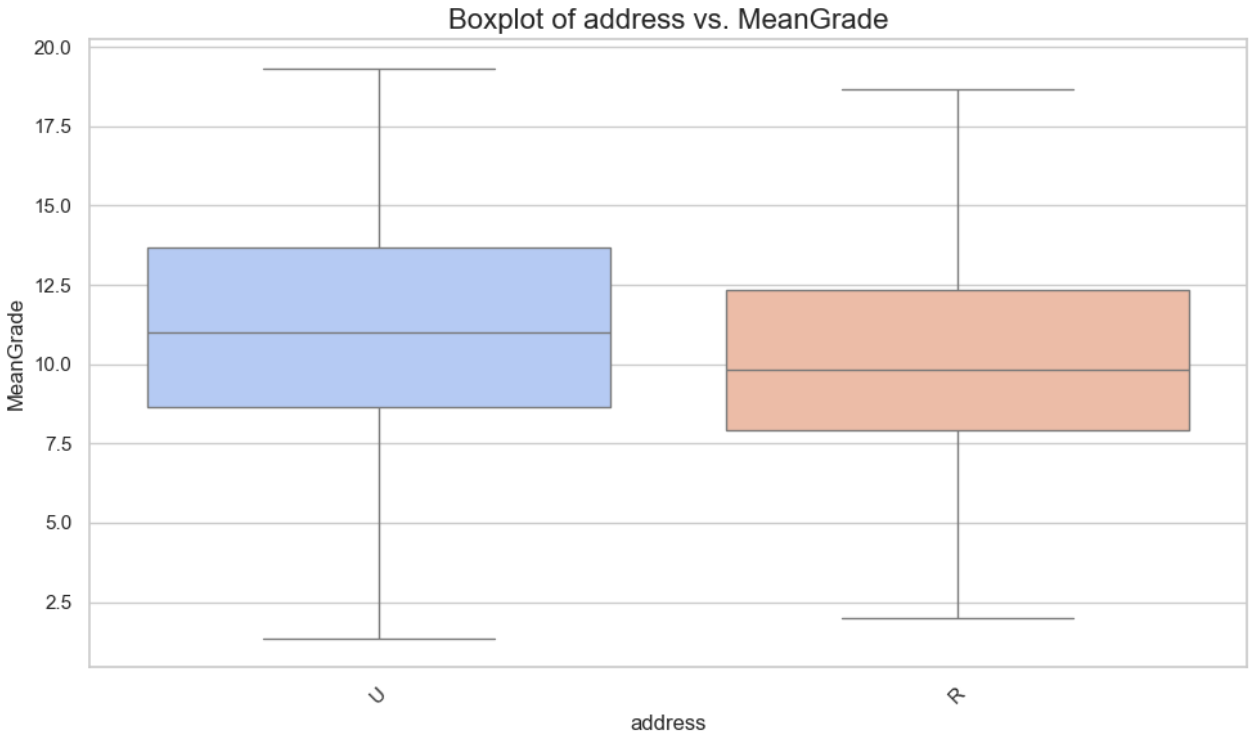
For absences, the data is heavily skewed to the right, indicating that a majority of the subjects have few to no absences. There is high variability in absences, as the range extends from 0 to over 70. However, the interquartile range is likely small, concentrated near the lower

end of the scale.

For MeanGrade, the distribution is relatively normally distributed. The variability or spread of MeanGrade is moderate, as indicated by the width of the bell curve. The standard deviation can be inferred to be moderate, allowing for a typical bell-curve distribution with most data points falling within one standard deviation of the mean.

To visualize correlation strength, we will create a heatmap. The diagonal correlation of 1 is obvious since it is the correlation of something to itself. Therefore, we are ignoring that.



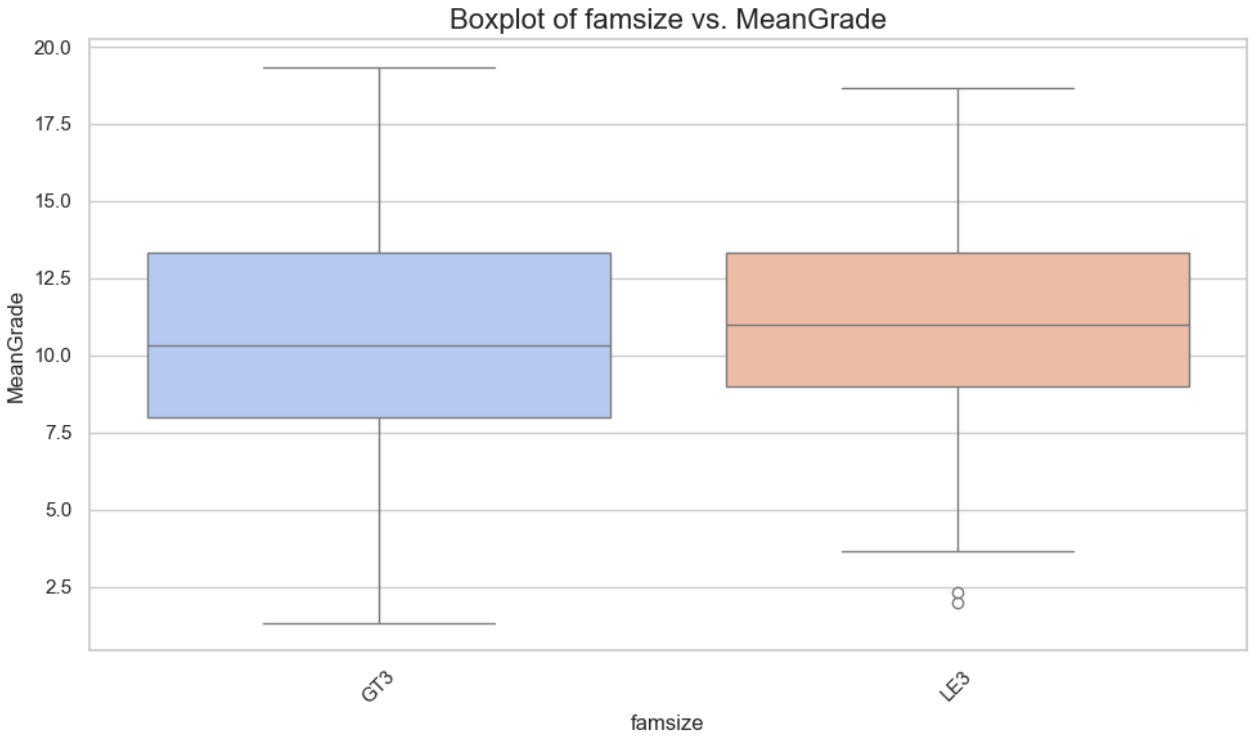


Students from urban areas (U) have a slightly higher median grade than students from rural areas (R).

The interquartile range (IQR), which represents the middle 50% of the grades, is slightly tighter for urban students, indicating less variability in their grades compared to rural students.

Both urban and rural students have a similar range of grades, as indicated by the whiskers of the boxplot, which extend from the lowest to the highest grades excluding outliers.

There appear to be no extreme outliers in either category.

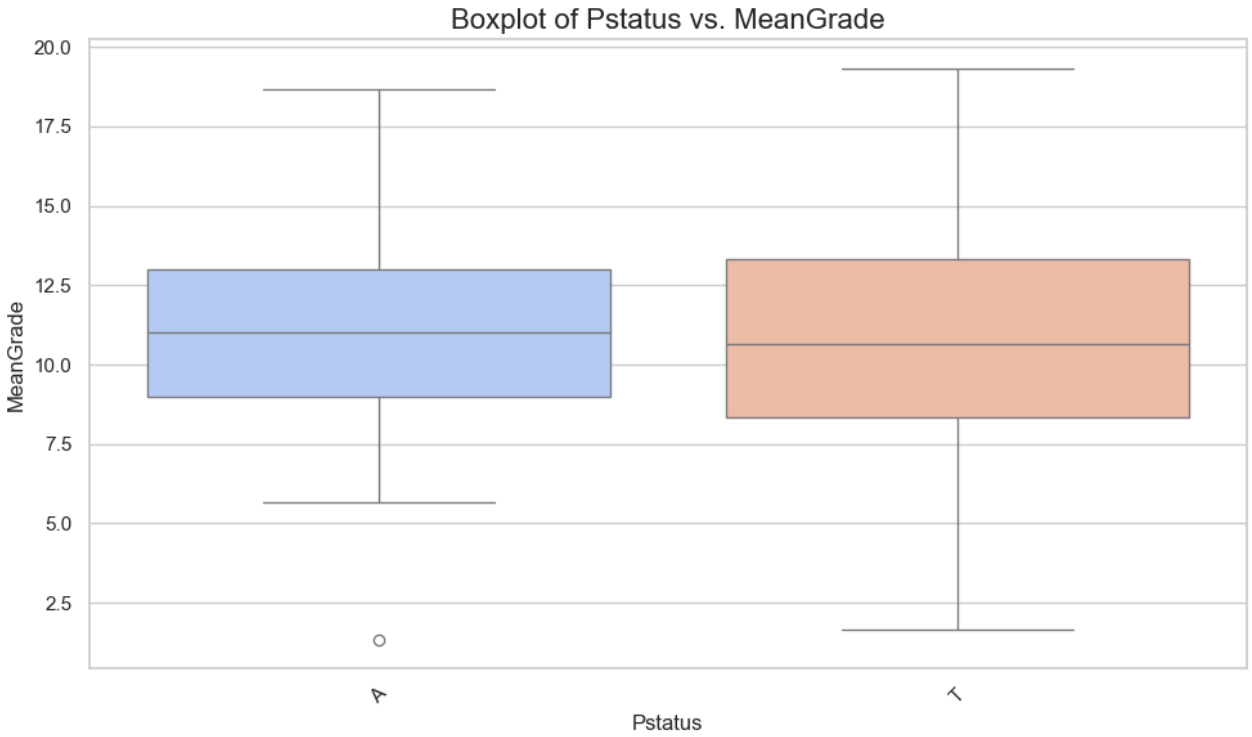


The median MeanGrade for students from larger families (GT3) is very similar to that of students from smaller families (LE3), indicating that family size has a minimal impact on the median grade.

The interquartile range for both groups is also quite similar, suggesting that the middle 50% of students in both categories have a comparable spread of grades.

Students from smaller families (LE3) appear to have a slightly wider range of grades, as indicated by the longer whiskers of the boxplot, which could suggest more variability in their academic performance.

There is an outlier in the smaller family size category (LE3), indicating an exceptionally low grade compared to the rest.

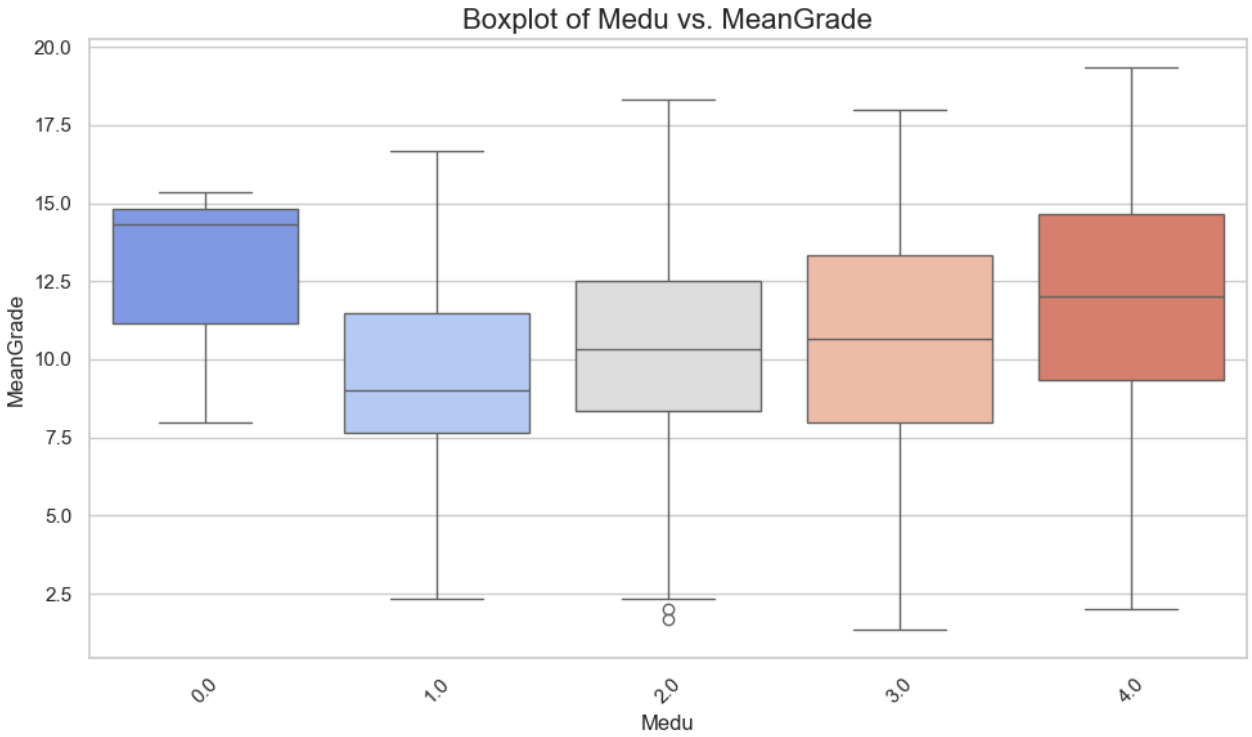


The median grade for students whose parents live together (T) is similar to that of students whose parents live apart (A), suggesting that parental cohabitation status has a limited direct impact on the MeanGrade.

The interquartile ranges are comparable, indicating a similar spread of grades among the middle 50% of students in both categories.

The range (as shown by the whiskers) is slightly wider for students whose parents live together, indicating more variability in grades among these students.

There is at least one outlier for students whose parents live apart, indicating a student with a significantly lower grade compared to the rest.



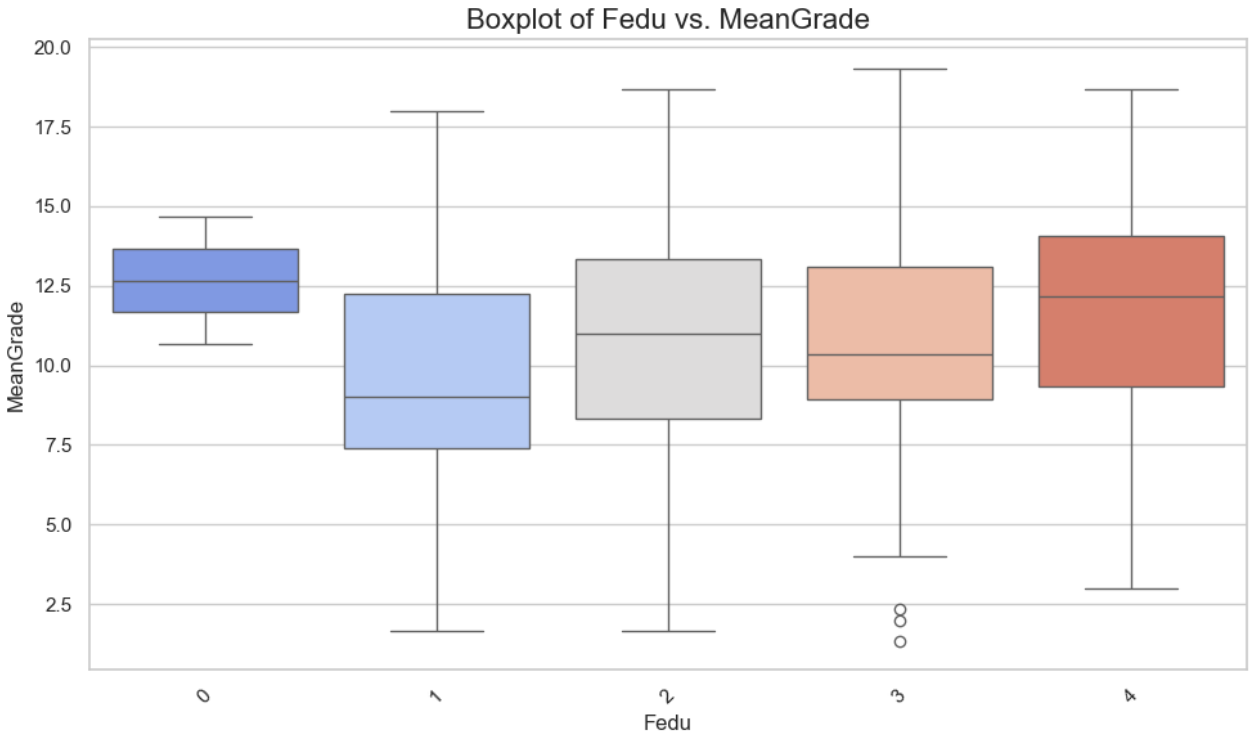
There is a general upward trend in the median MeanGrade as the education level of the mother increases.

Students with mothers who have the highest level of education (level 4) tend to have higher median grades compared to other groups.

The interquartile range (IQR) for students with mothers who have mid-level education (levels 2 and 3) is broader, indicating more variability in students' grades within these groups.

The variability in grades seems to decrease (narrower boxes) for students whose mothers have the lowest (0) and the highest (4) education levels.

There is at least one significant outlier in the group where the mother's education level is 2, indicating a student with a much lower grade than typical for that group.

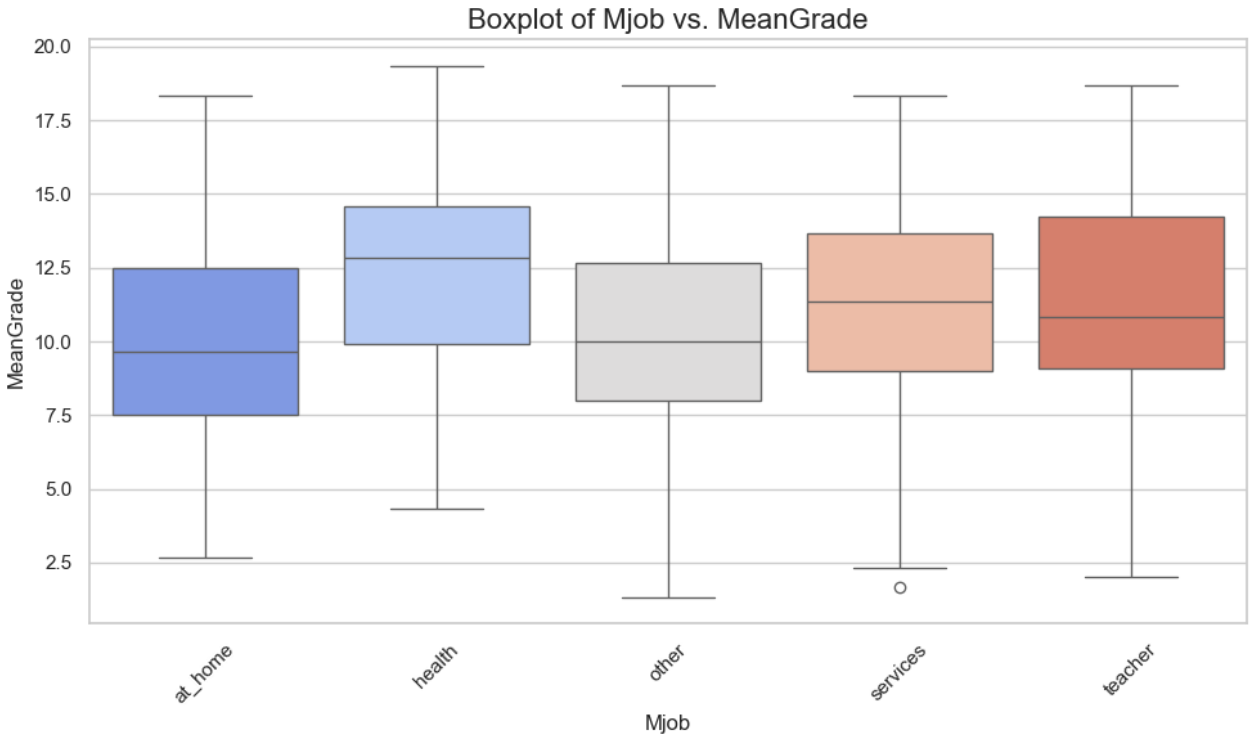


Similar to the trend observed with mother's education, there is an overall tendency for the median MeanGrade to increase with the father's education level. Students with fathers who have a higher education level (4) tend to have higher median grades.

The interquartile ranges (IQR) suggest that there is a substantial spread of grades among students at each level of father's education, with no clear pattern indicating that one level has significantly more variability than the others.

Notably, the group where the father's education level is 0 has a higher median grade than the group where the father's education level is 1, which is an interesting deviation from the overall trend.

The presence of outliers, particularly in the group where the father's education level is 3, indicates that there are students with grades significantly lower than their peers.

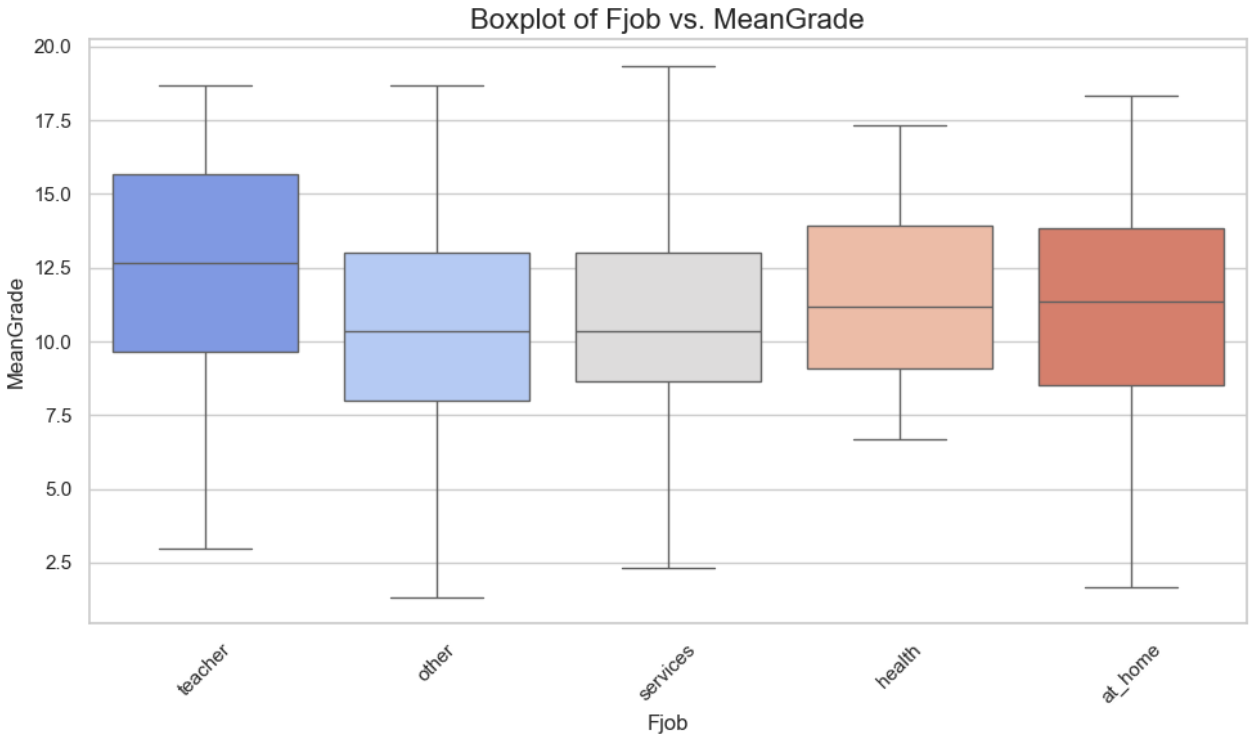


Students whose mothers are in the health or teaching sectors tend to achieve higher median grades compared to those whose mothers are in at-home roles or other services.

The interquartile ranges (IQR) for all the occupational categories indicate a broad spread of grades, suggesting a similar level of grade variation regardless of the mother's job.

While the median grades for mothers in health and teaching professions are higher, the overlap in the IQRs across different jobs suggests that the mother's occupation alone does not determine student performance.

There are outliers in the 'services' category, indicating that some students with mothers in this sector have grades that are lower than typical for their group.



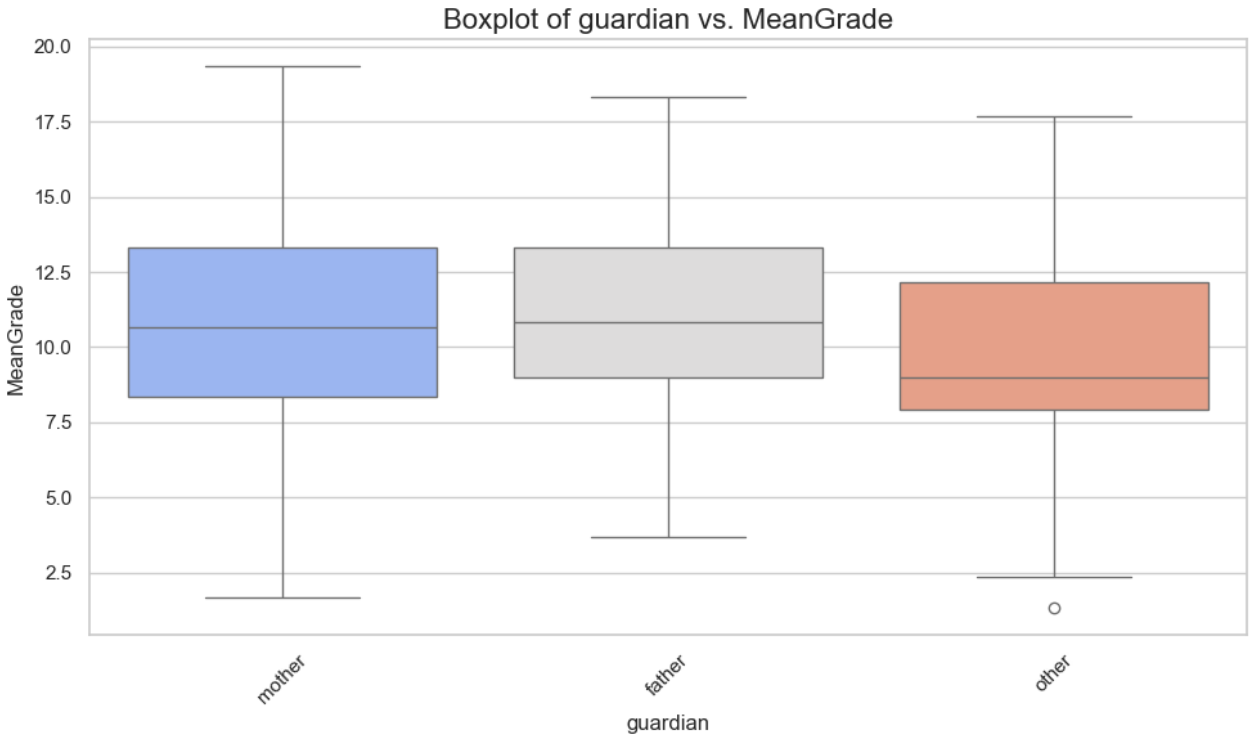
There is a variation in median grades across different father's job sectors, but the differences are not stark.

Students with fathers working as teachers tend to have a slightly higher median grade compared to other sectors.

The interquartile range is relatively consistent across the different job sectors, indicating a similar spread of grades within each group.

The at_home and other categories show a greater range in grades, as evident from the longer whiskers, suggesting more variability in student performance.

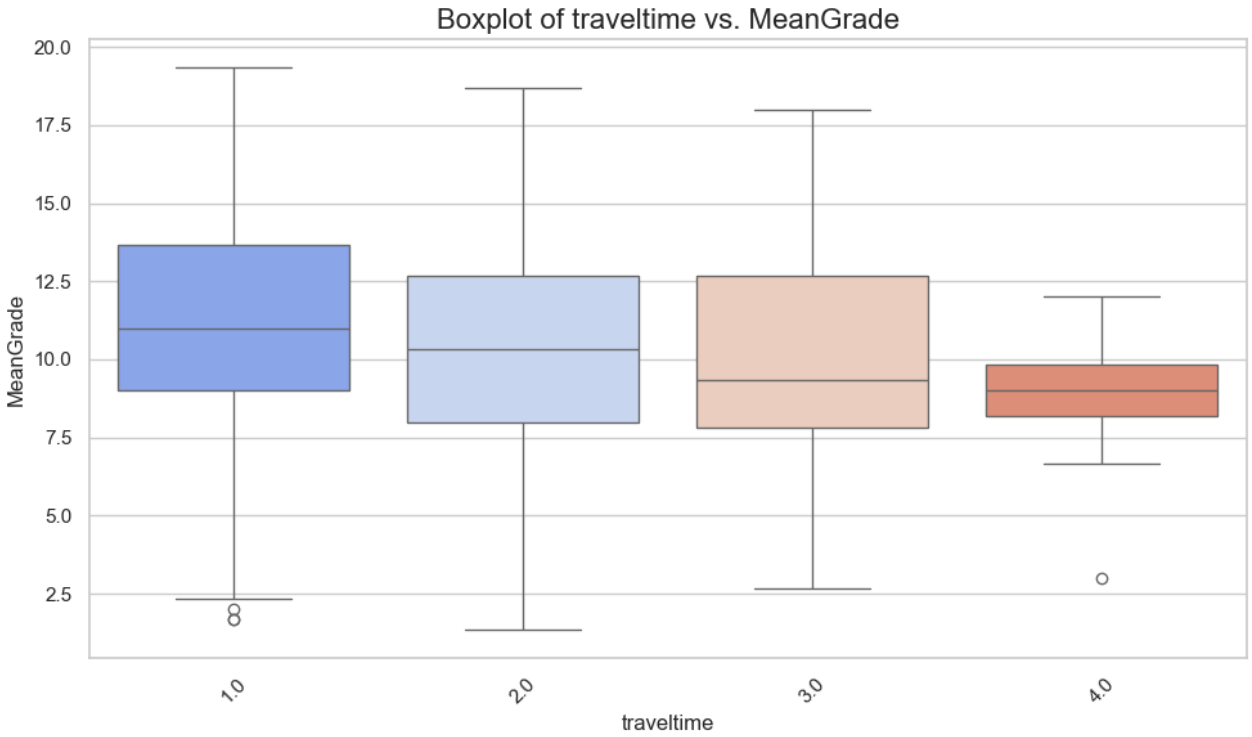
There are outliers in the services and health categories, with at least one student from each category having a significantly lower grade than their peers.



The box plot of guardian vs. MeanGrade shows that students with their mother as the guardian tend to have slightly higher median grades, while those with 'other' as their guardian have the lowest median grade.

There is some variability in grades among students with 'other' guardians, as indicated by the wider spread of data.

Additionally, outliers in the 'father' and 'other' categories suggest that a few students in these groups have significantly lower grades compared to their peers.

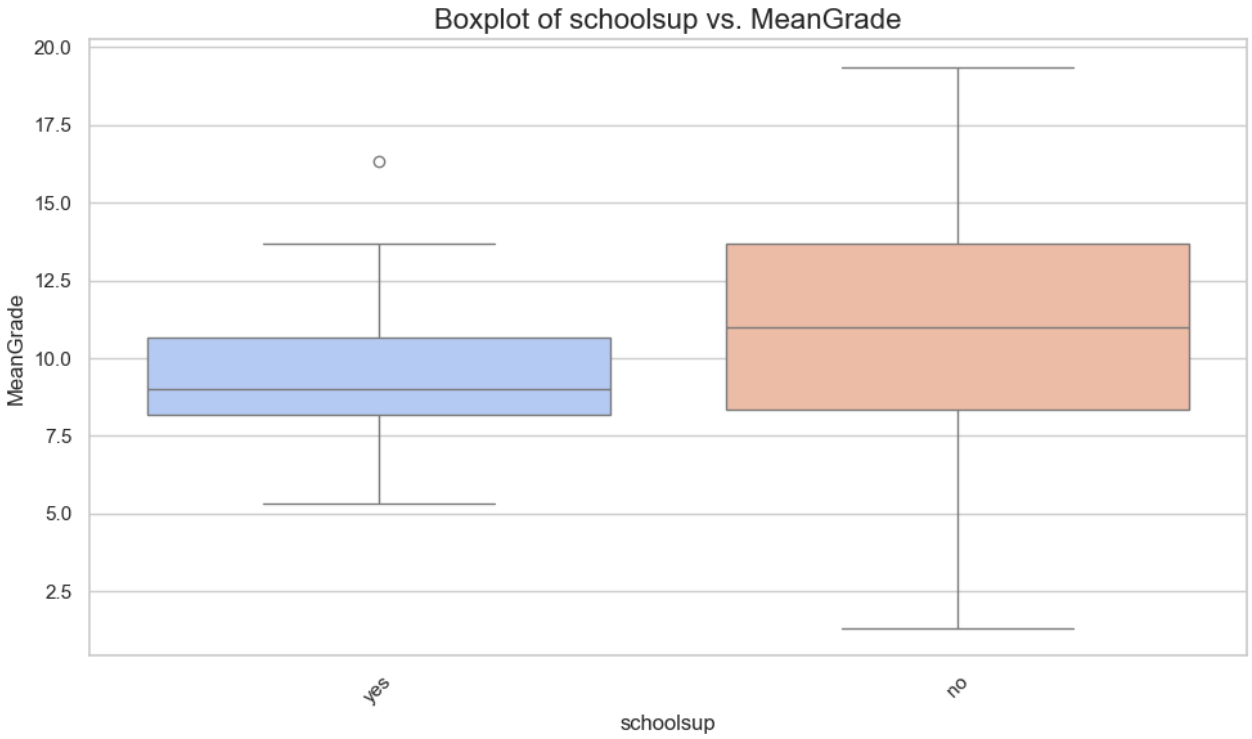


The median grades vary across the four travel time categories, though not dramatically. Students with the shortest travel time (1) display a marginally higher median grade.

The interquartile ranges are fairly consistent, suggesting a similar distribution of grades within each travel time category.

The longest travel time (4) shows a tighter grouping of grades but a lower median, indicating a modest decline in grades with increased travel time.

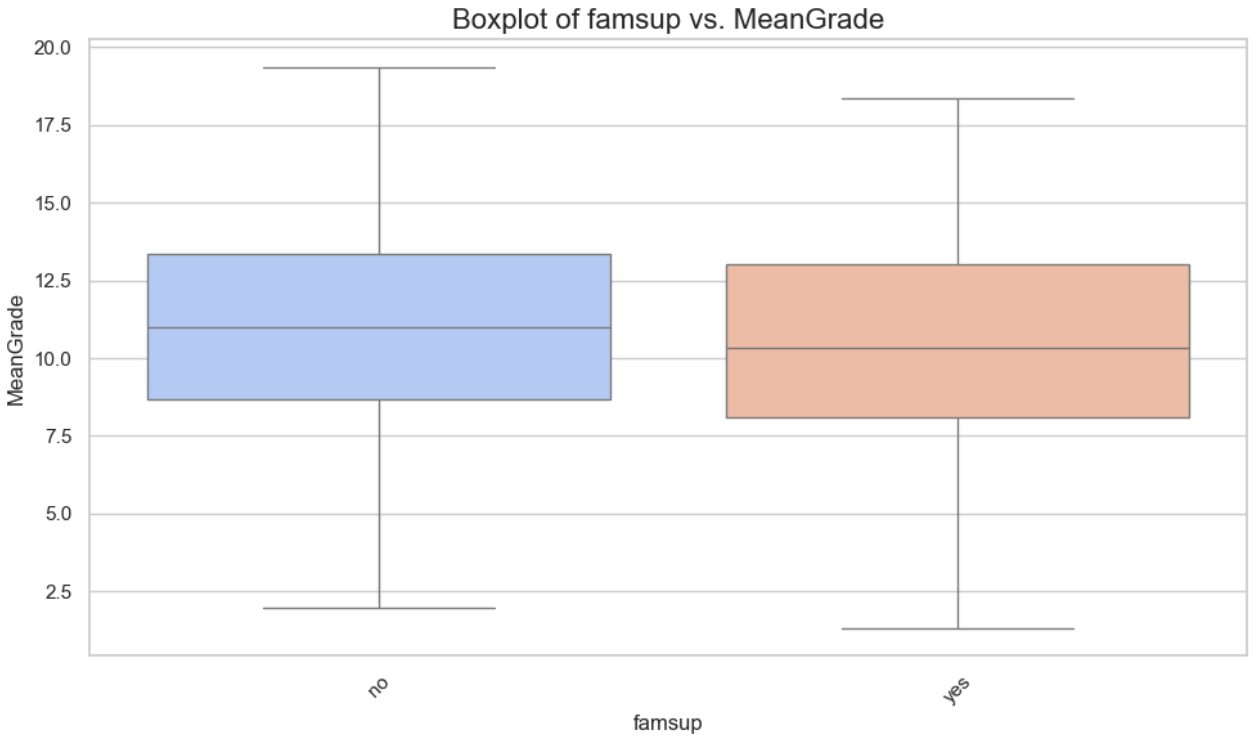
This category, along with the shortest travel time, includes outliers indicating a few students with grades significantly lower than their peers.



Students receiving school support ('yes') have a lower median grade compared to those not receiving support ('no').

The interquartile range for students with support is slightly narrower, suggesting a more consistent performance within this group. Conversely, students without support show a wider range of grades, as indicated by the broader box, and also have higher grades overall.

There is an outlier in the group receiving school support, indicating at least one student with a significantly higher grade than their peers in that category.

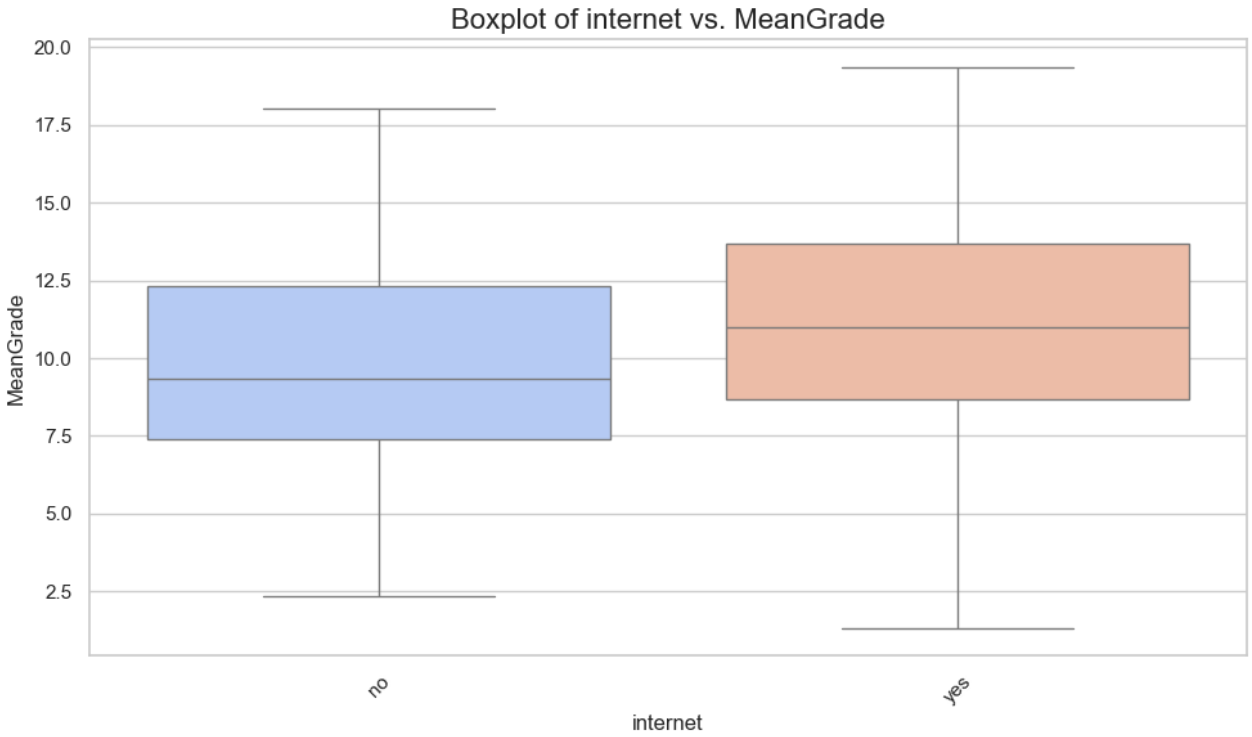


Students with family support ('yes') demonstrate a slightly higher median grade than those without ('no').

The interquartile range for both groups is similar, indicating a consistent spread of grades among students regardless of family support.

Both categories exhibit a wide range of grades, with the 'no' group showing a slightly more condensed range, which suggests a tighter cluster of grades around the median.

There are no visible outliers, implying that all students' grades fall within a close range of the upper and lower quartiles in both categories.

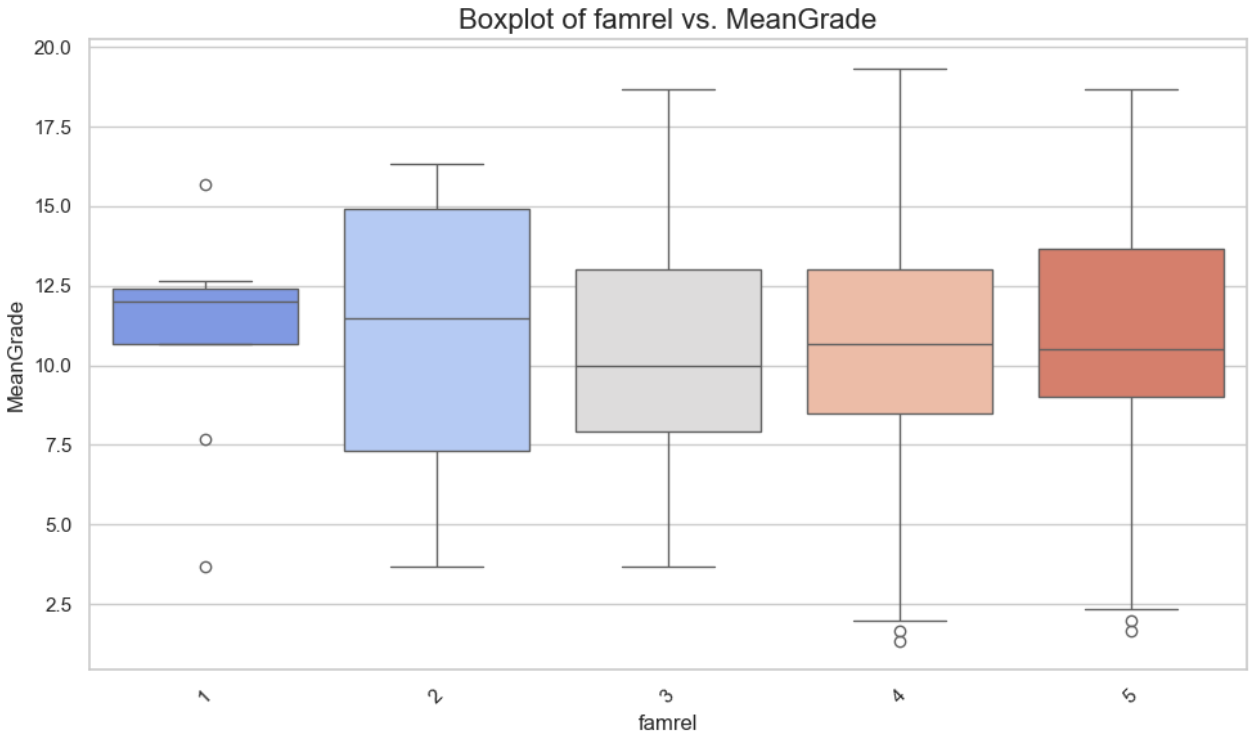


Students with internet access ('yes') have a higher median grade than those without ('no').

The interquartile range for both groups is comparable, suggesting a similar distribution of grades within each category. Students without the internet show a slightly narrower interquartile range, indicating a more consistent grade performance among this group.

The range of grades, represented by the whiskers, is broader for students with internet access, pointing to a greater variability in their grades.

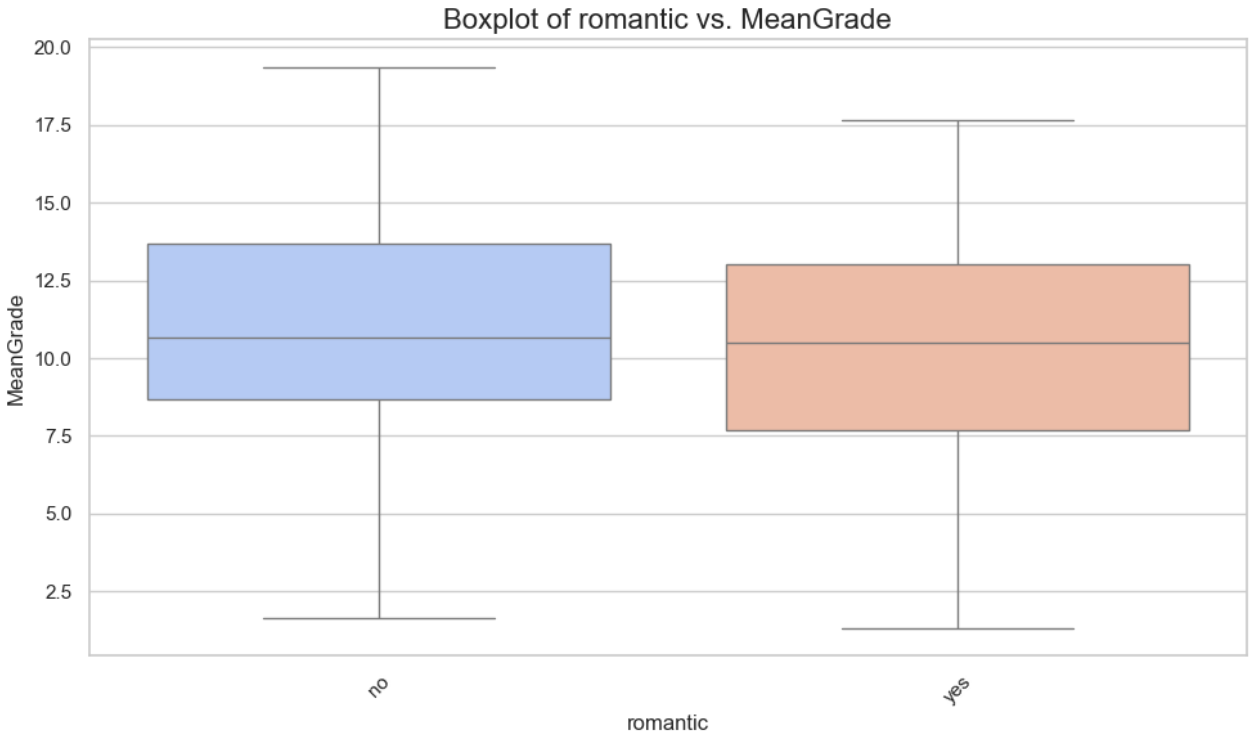
No outliers are present, which means all students' grades fall within the expected range based on the quartiles for both groups.



Grades vary across different levels of family relationships, with no clear upward or downward trend associated with better family relationships.

Students with the lowest 'famrel' rating (1) have a high median grade but also the presence of outliers, suggesting inconsistencies in their academic performance. The second category (2) shows a wide range of grades, while the middle category (3) has a more moderate range. Categories 4 and 5, which may indicate better family relationships, do not show a significant difference in median grades compared to lower ratings.

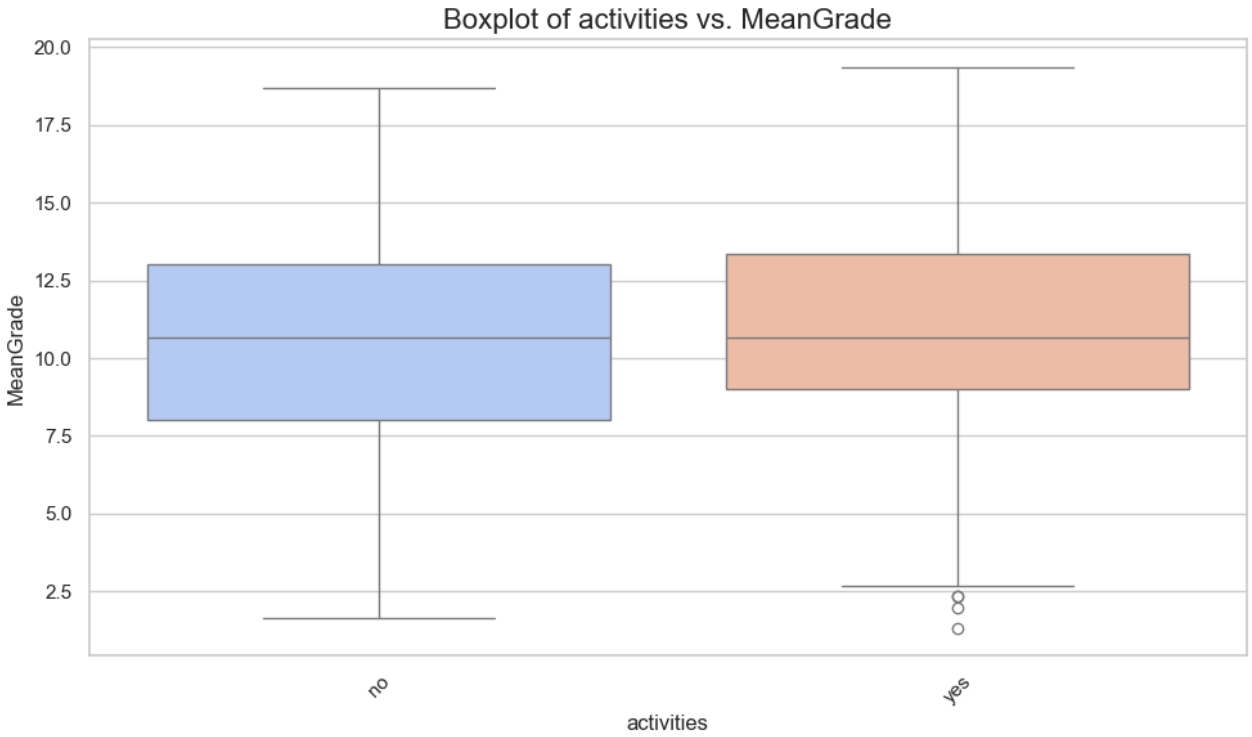
Outliers are present in categories 1, 2, and 4, indicating individual variations that deviate from the common trend.



The boxplot shows that students not involved in romantic relationships ('no') have a higher median grade compared to those who are ('yes').

The interquartile range for both groups is similar, indicating a comparable spread of grades among both romantically involved and non-involved students. The range, as denoted by the whiskers, is slightly more extended for those in a romantic relationship, suggesting a broader variety of grades.

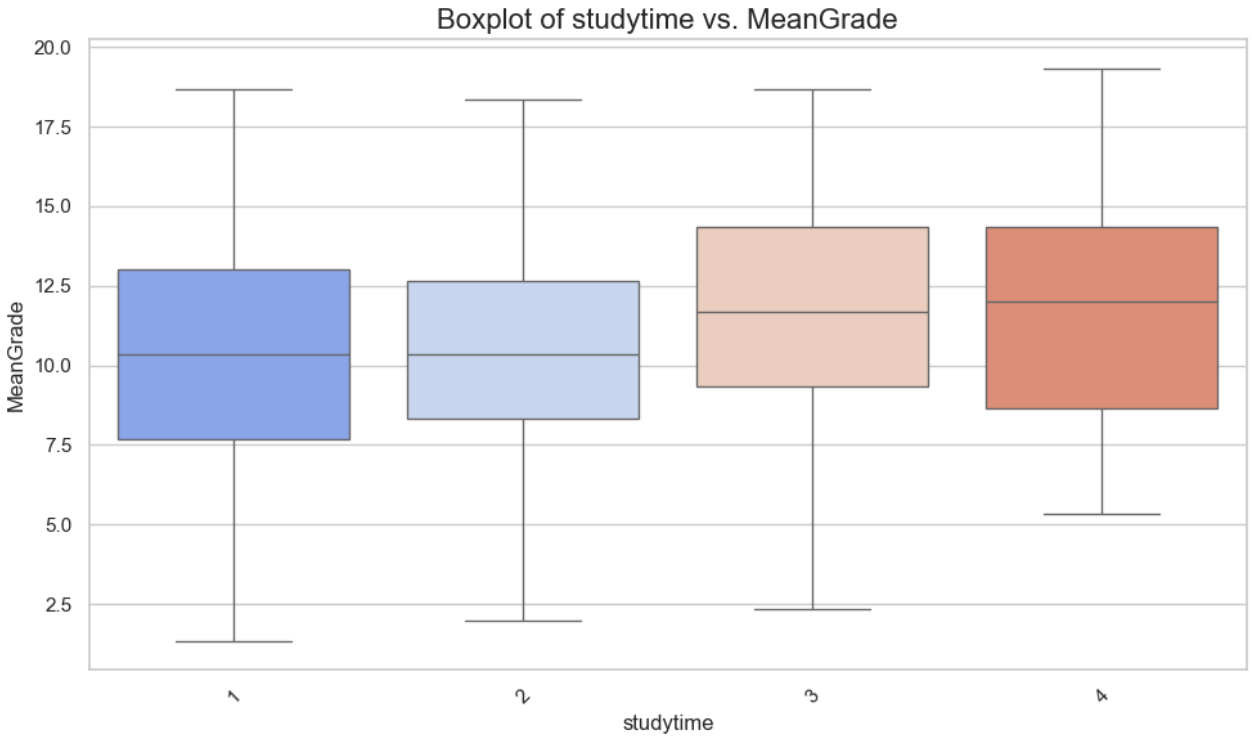
There are no visible outliers, which implies a relative consistency in grades within the expected range for both categories. This could suggest that romantic involvement might have a slight association with lower median grades.



The box plot comparison indicates that students who participate in activities ('yes') have a slightly lower median grade than those who do not ('no').

The interquartile range for both groups is comparable, suggesting similar grade variability within each group. Notably, the 'yes' category has a longer lower whisker and outliers, pointing to a few students with significantly lower grades than their peers.

This could imply that while involvement in activities does not substantially affect the median grade, it may be associated with a few cases of notably lower academic performance.

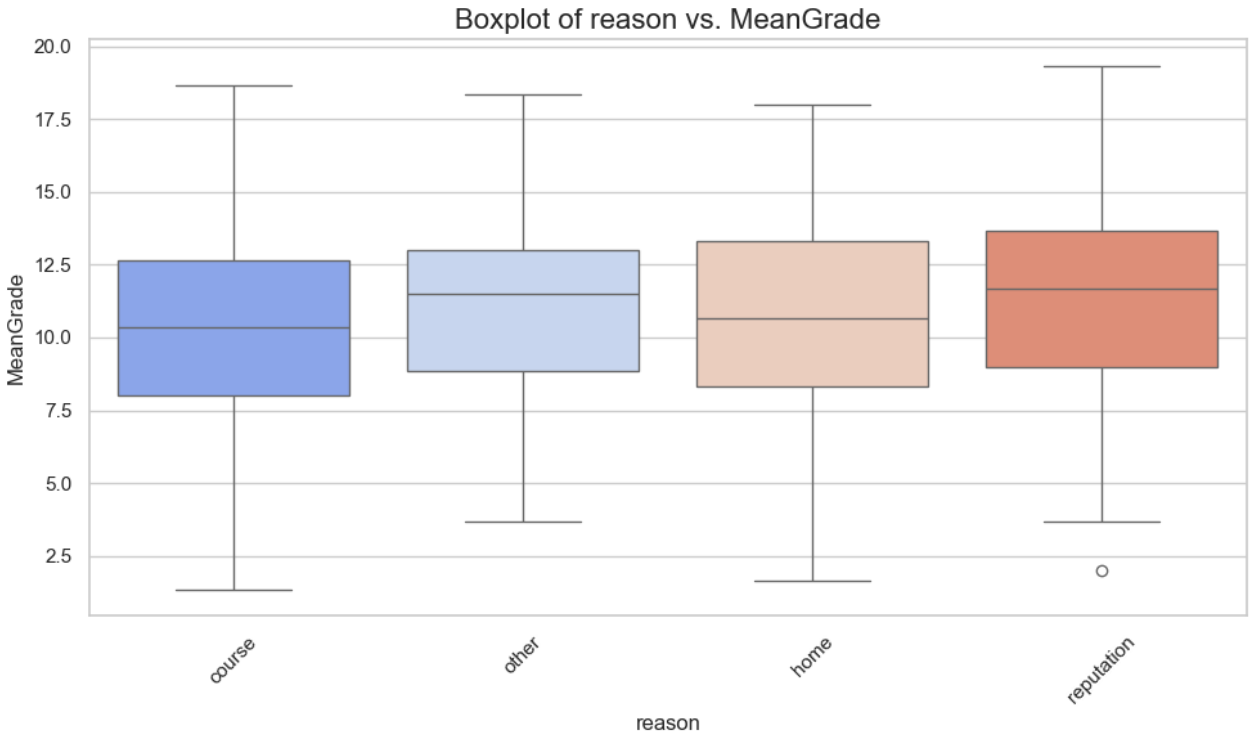


The box plot reveals that as study time increases from 1 to 4, there's a general trend of increasing median grades.

The first category (1) has the lowest median grade, and each subsequent category shows a progressive increase in median grade. The interquartile ranges are similar across categories 1 to 3, suggesting a consistent spread of grades regardless of study time.

However, the highest study time category (4) has a narrower interquartile range, indicating a more compact distribution of grades.

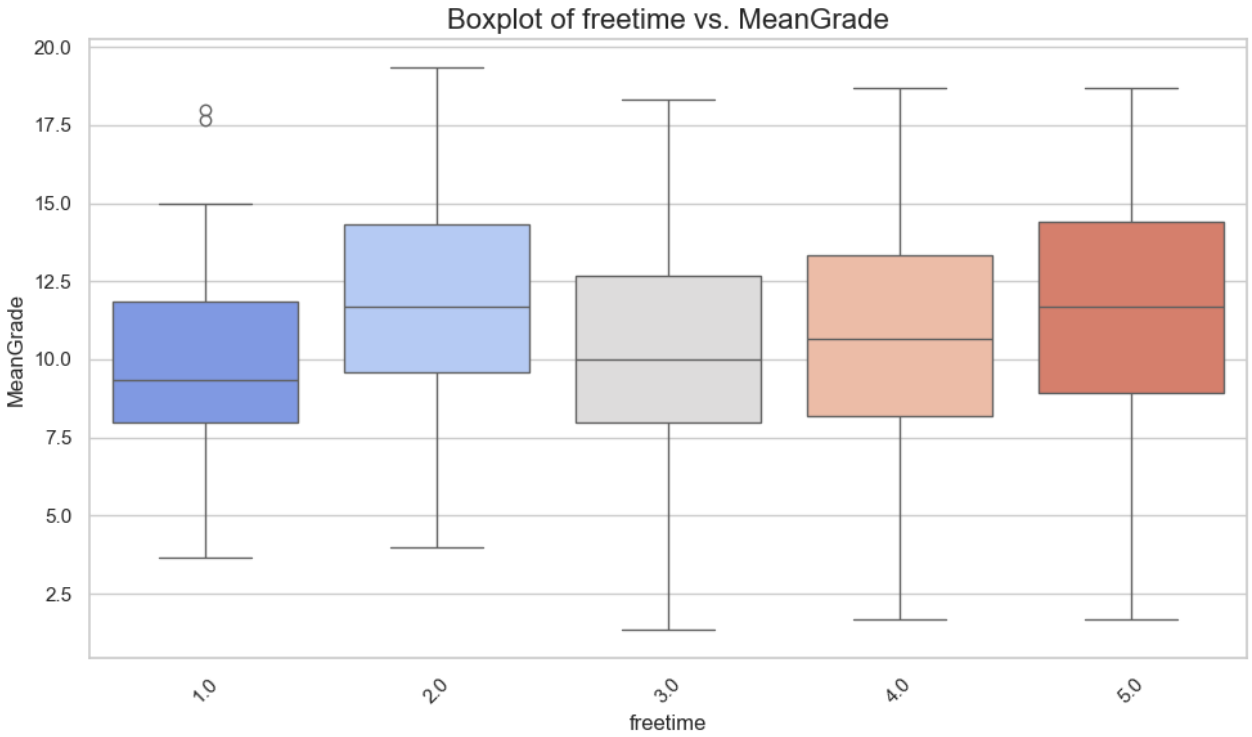
The data points are spread out for all categories, as seen in the long whiskers, with no significant outliers, suggesting that additional study time could be correlated with a modest improvement in grades.



The box plot suggests that the reason for attending school, categorized as 'course', 'other', 'home', and 'reputation', does not significantly affect the median grades, which are relatively similar across all categories.

The interquartile ranges are also quite consistent, indicating a comparable spread of grades within each reason category. The 'course' and 'reputation' reasons show slightly more variability in grades, as evidenced by the longer whiskers, compared to 'home' and 'other'.

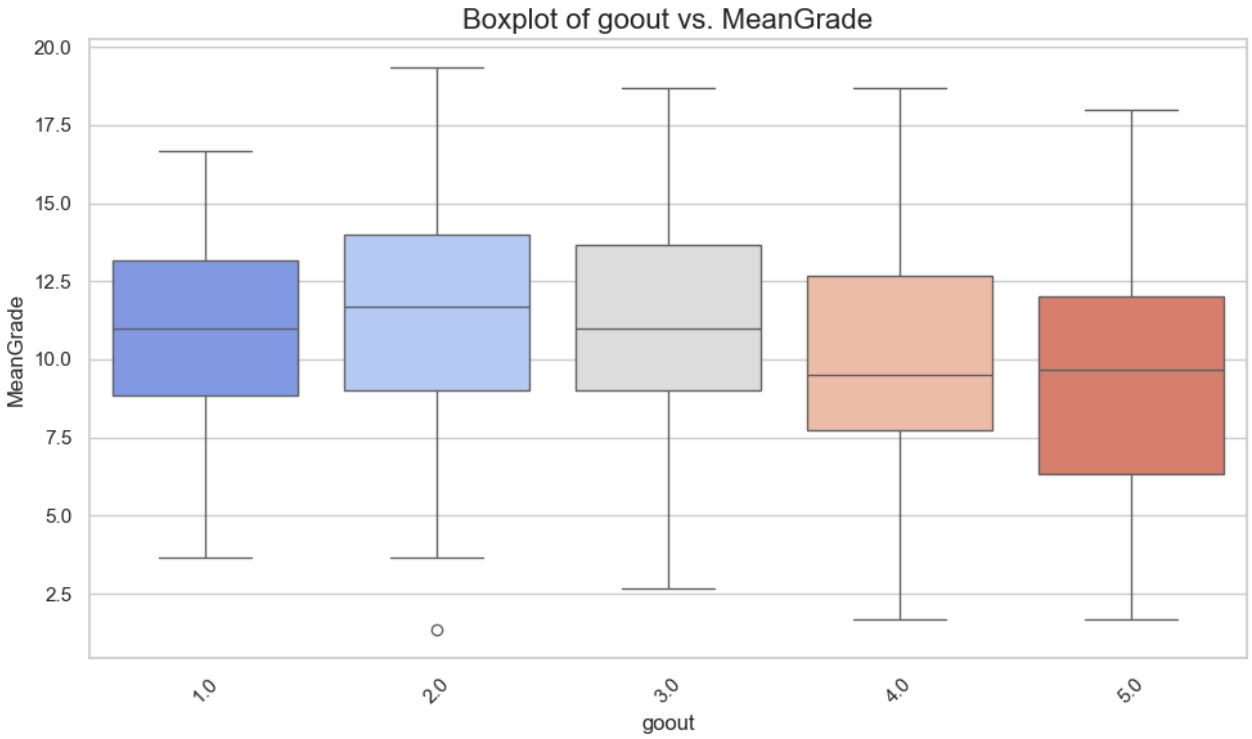
There are no apparent outliers, suggesting that extreme grades are not common within any of the reason groups. This indicates that the motivation for attending school, as defined by these categories, might have a minimal direct impact on mean grades.



The boxplot illustrates that the median grades remain fairly consistent across different levels of 'freetime', ranging from 1 to 5.

There is a slight decrease in median grade as freetime increases, with category 1 having the highest median and category 5 the lowest.

The interquartile ranges and whiskers indicate a similar degree of grade variability across all levels of freetime, although category 1 shows an outlier with a grade significantly higher than the rest. This suggests that while more freetime may be associated with a slight decrease in median grades, the overall effect is minimal, and students' performance varies similarly regardless of the amount of free time they have.

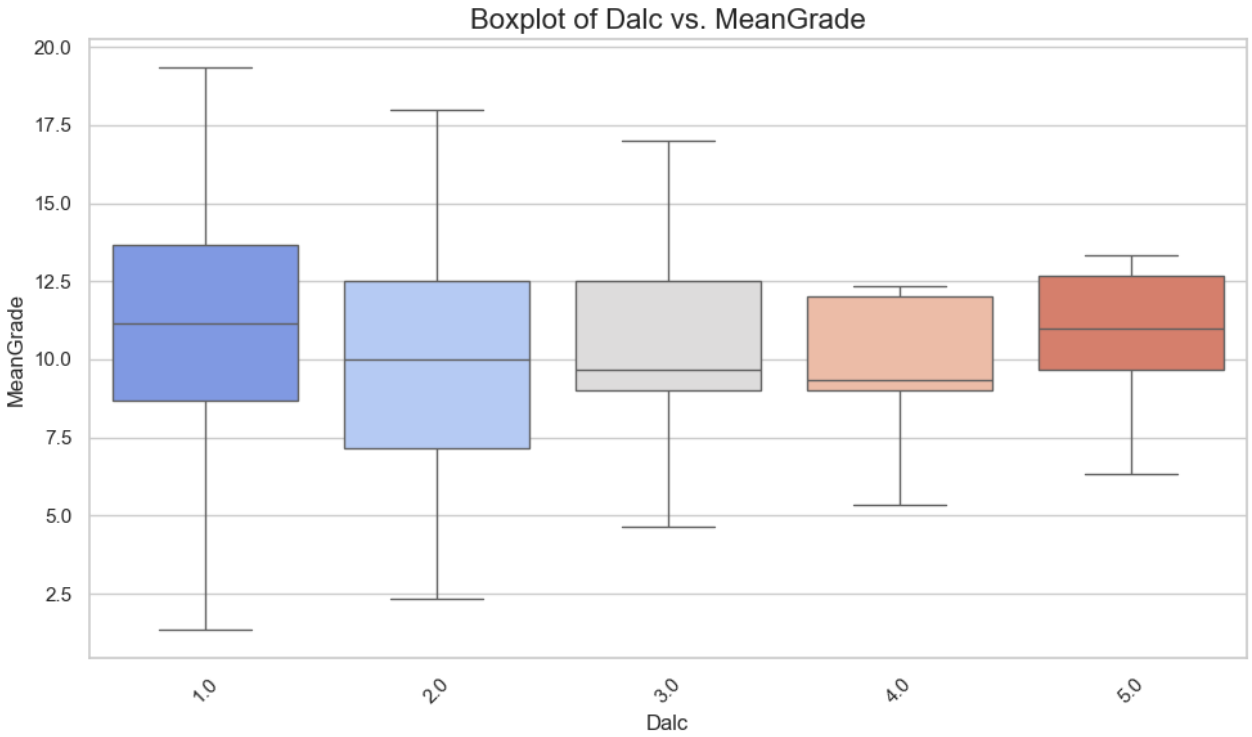


The box plot indicates that median grades across varying frequencies of 'goout' (going out with friends) are relatively consistent, with no clear trend as the frequency increases from 1 to 5.

The interquartile ranges are similar for all categories, suggesting a comparable distribution of grades among students, regardless of how often they go out.

The length of the whiskers points to a moderate variability in grades for each level of 'goout', but there is one notable outlier in category 2, where a student has a significantly lower grade.

Overall, the frequency of going out does not appear to have a marked impact on median grades.



The box plot shows that as 'Dalc' (presumably daily alcohol consumption) increases from 1 to 5, there isn't a straightforward trend in median grades. The median grades are relatively stable across categories 1 to 4, with a slight decrease in category 5.

The interquartile ranges remain consistent for categories 1, 2, and 3, suggesting a similar spread of grades among these students.

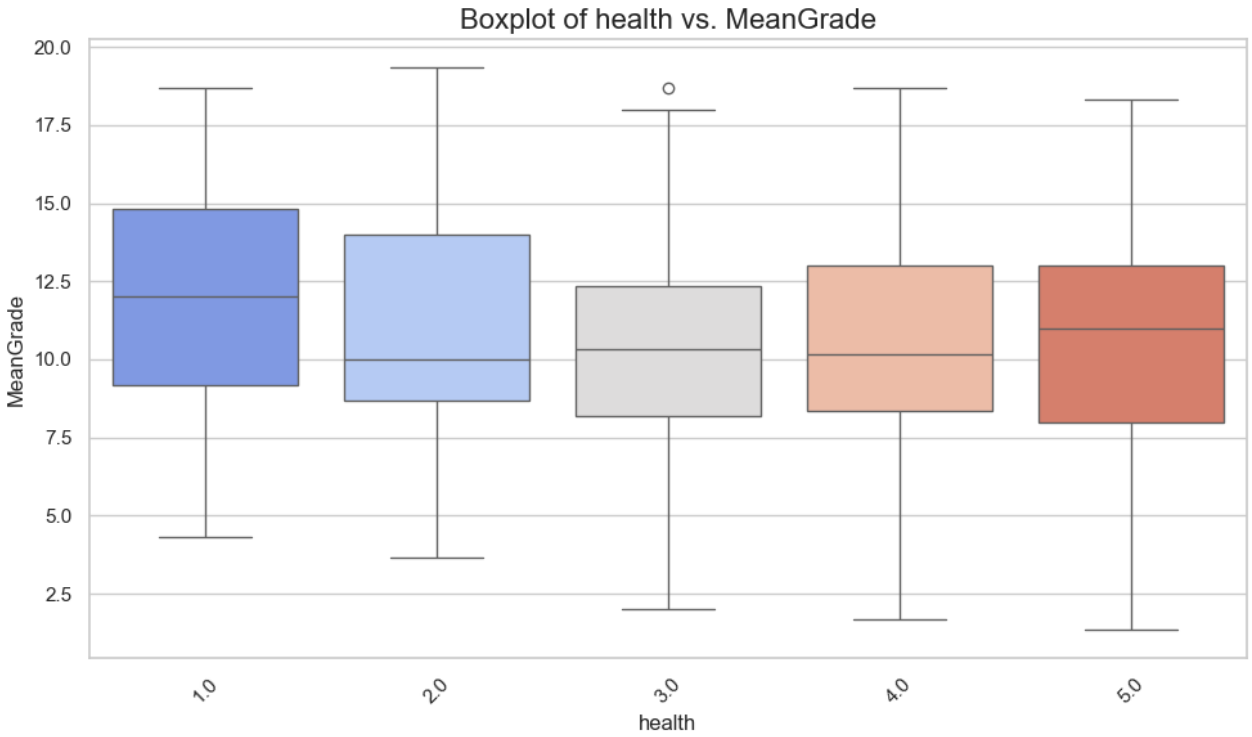
However, the range widens in categories 4 and 5, indicating greater variability in grades as daily alcohol consumption increases. The distribution of grades within each category, indicated by the whiskers, shows moderate variability with no significant outliers, suggesting that daily alcohol consumption has a minimal to moderate influence on academic performance.

The box plot comparing 'Walc' (likely weekend alcohol consumption) with 'MeanGrade' reveals that median grades do not show a significant variation across the levels of alcohol consumption.

The interquartile ranges are relatively consistent, indicating that the spread of grades within each category of alcohol consumption remains similar.

Categories 1 through 4 show a moderate range of grades, but category 5 displays a slightly wider spread, suggesting more variability in grades at the highest level of weekend consumption.

Outliers are present in categories 1 and 4, where some students have notably lower grades than their peers. Overall, weekend alcohol consumption does not appear to have a strong correlation with median grades.

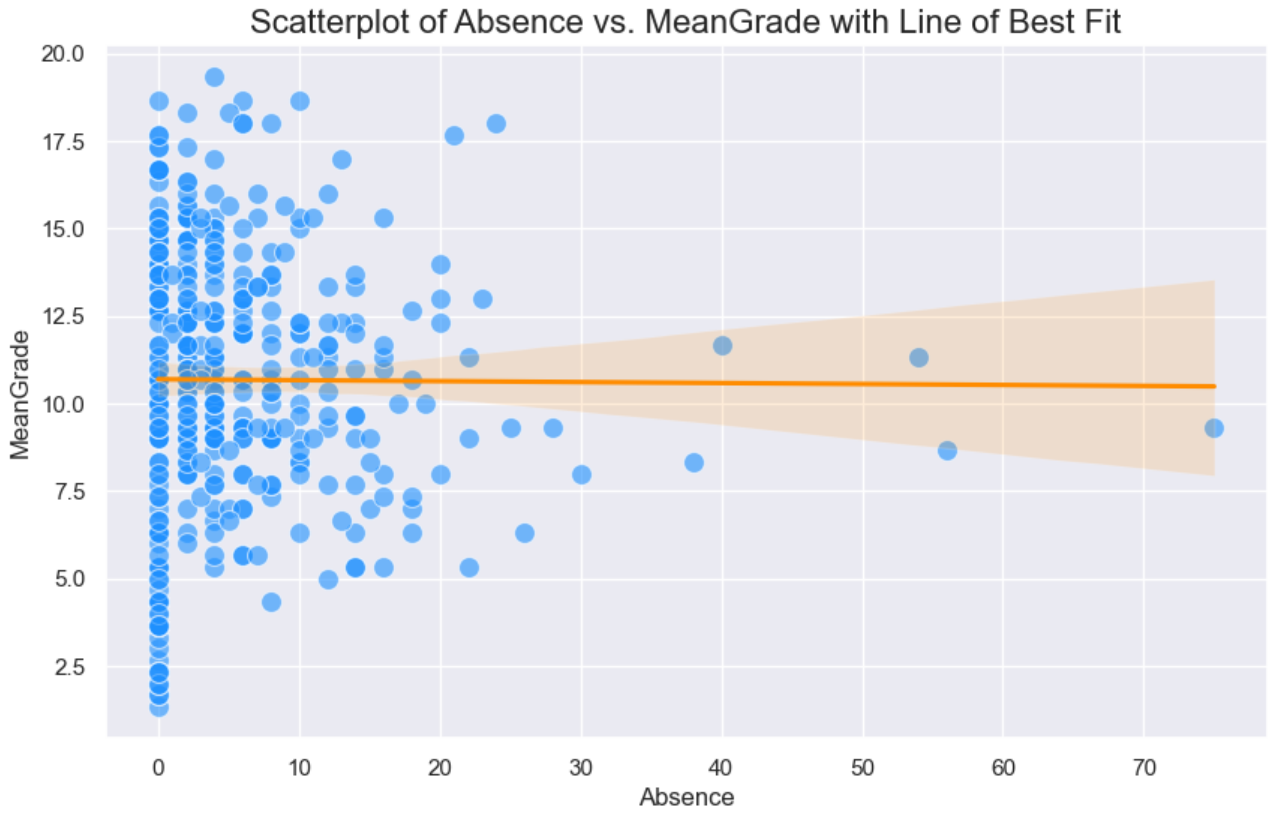


The box plot suggests that students' reported health status, rated from 1 (poor) to 5 (excellent), does not exhibit a clear pattern in relation to their MeanGrade.

The median grades across different health categories remain relatively consistent, with no significant trend indicating that better self-reported health correlates with higher grades.

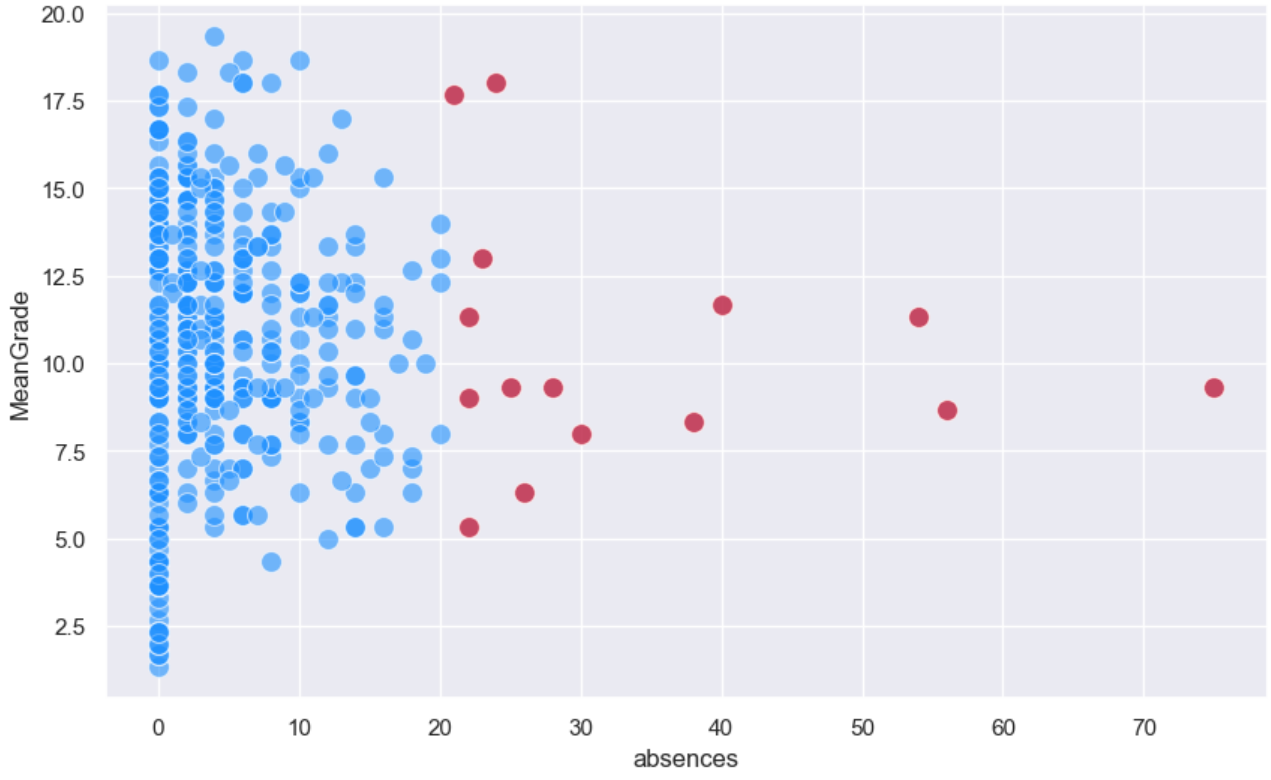
The interquartile ranges are similar for all health categories, pointing to a comparable spread of grades regardless of health status.

There is an outlier in the category representing the poorest health, indicating a student with a significantly higher grade compared to peers in the same category. Overall, health status as represented here shows no strong association with academic performance.

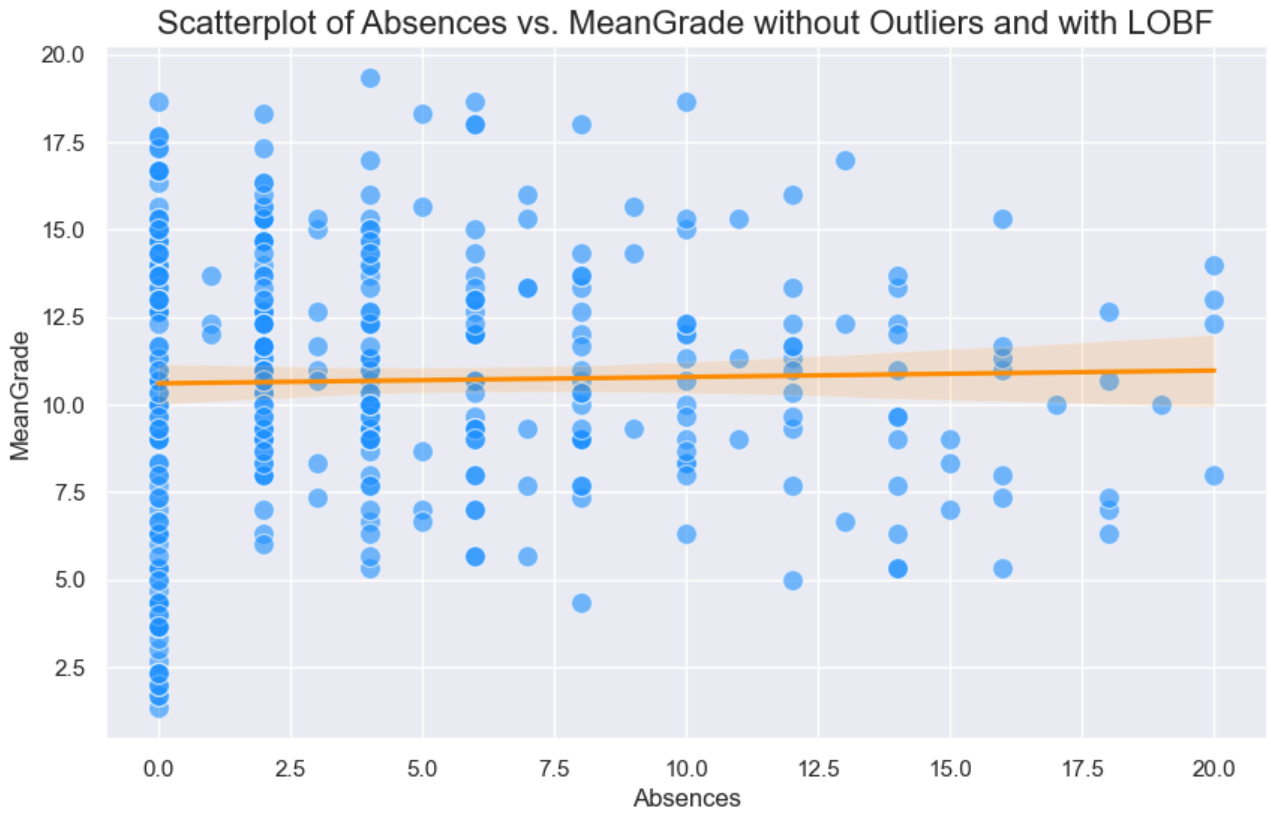


The correlation coefficient between absences and MeanGrade is approximately -0.006. This indicates a very weak negative correlation between the number of absences and the mean grade of students.

We thought this could be because of outliers, since it is very counter-intuitive that the grade is not affected by the number of absences. So, we try identifying outliers:



And draw a separate graph without them:



But the result does not change much. So, we have to conclude that absence indeed does not affect grade.

5 Hypothesis Testing

T-Test for Mean of *MeanGrade*

Test Details:

- Null Hypothesis (H_0): The mean of *MeanGrade* is 10.
- Alternative Hypothesis (H_1): The mean of *MeanGrade* is not 10.
- Significance Level (α): 0.05.

Based on the provided data, we have:

- Sample mean (*MeanGrade*): $\bar{x} \approx 10.68$
- Population mean (hypothetical): $\mu_0 = 10$
- Sample standard deviation (*MeanGrade*): $s \approx 3.70$
- Sample size (n): 395

The calculated t-statistic is approximately 3.65. With the degree of freedom of 395 (sample size) - 1 = 394, we got the p-value of approximately 0.0003.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

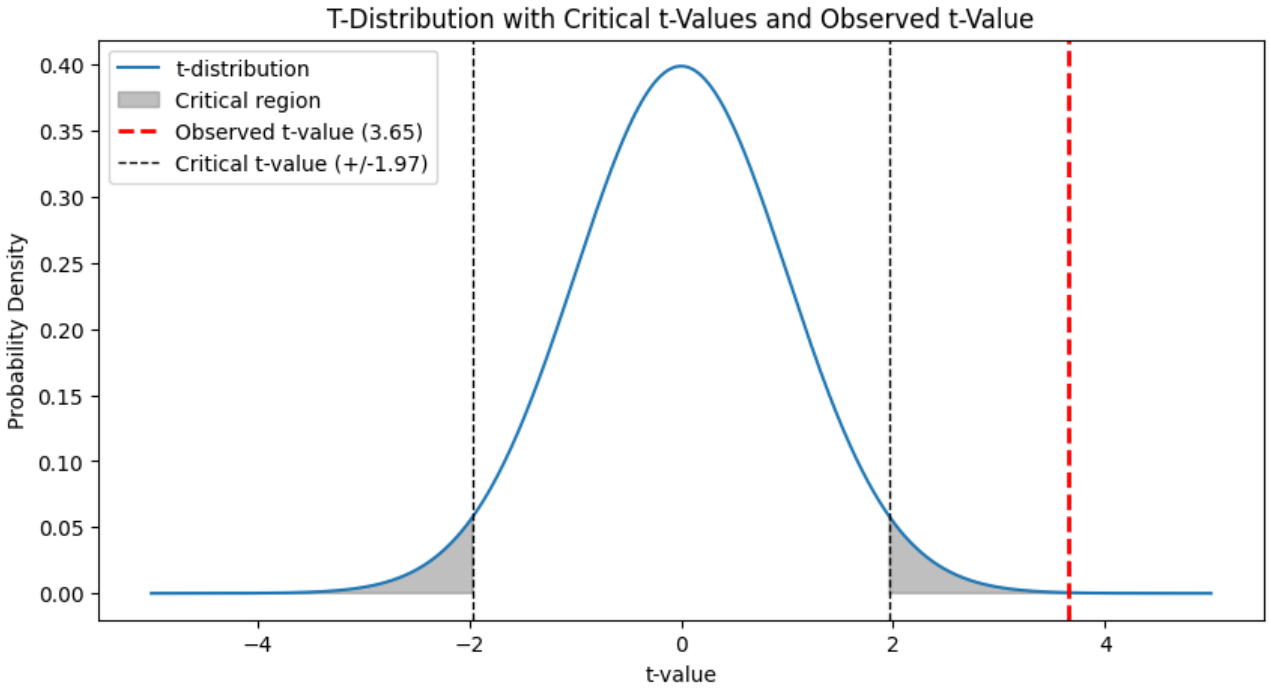


Figure 1: T-Distribution with Critical t-Values and Observed t-Value

Since the p-value (0.0003) is less than the significance level (0.05), we reject the null hypothesis. This suggests that the actual mean grade of students is significantly higher than the hypothesized mean of 10.

Two-sample T-test comparing the mean

Test Details:

- Null Hypothesis (H_0): There is no difference in mean grades (μ) between students with less than or equal to 7 absences and those with more than 7 absences ($\mu_{\leq 7} = \mu_{>7}$).
- Alternative Hypothesis (H_1): There is a difference in mean grades (μ) between students with less than or equal to 7 absences and those with more than 7 absences ($\mu_{\leq 7} \neq \mu_{>7}$).

Significance Level (α): 0.05.
For students with less than or equal to 7 absences:

- Mean = 10.726,
- Standard deviation = 3.91,
- Sample size = 287.

For students with more than 7 absences:

- Mean = 10.556,
- Standard deviation = 3.068,
- Sample size = 108.

Use this formula:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 0.455$$

P-value = 0.650 $\Rightarrow$ we do not reject the null hypothesis
This suggests that there is no statistically significant difference in the mean grades between the two groups of students divided by the number of absences.

The Analysis of Variance (ANOVA) to compare performance between schools

Null Hypothesis (H_0): The mean grades of students are the same across all schools.

Alternative Hypothesis (H_1): At least one school's mean grade is different from the others.

Sum of Squares for Treatment (SSTr)

$$SSTr = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = 63.584$$

Degree of freedom for treatment

$$k - 1 = 3$$

Mean Square for Treatment (MSTr)

$$MSTr = \frac{SSTr}{k-1} = 21.195$$

Sum of Squares for Error (SSE)

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 = 5320.908$$

Degree of freedom for Error

$$k(n-1) = 391$$

Mean Square for Error (MSE)

$$MSE = \frac{SSE}{k(n-1)} = 13.608$$

Computed F

$$f = \frac{MSTr}{MSE} = 1.557$$

$\Rightarrow$ p-value $= 0.199 > 0.05$

$\Rightarrow$ we do not reject the null hypothesis

$\Rightarrow$ there are no statistically significant differences in the mean grades among the different schools.

6 Conclusion

Detailed Summary of Key Findings

- **Influence of Socio-Economic and Educational Factors:**
  - *Parental Education*: A clear positive correlation was observed between the education level of parents (especially mothers) and student grades. This suggests the pivotal role of parental educational background in shaping student academic success.
  - *Occupational Impact*: Certain parental occupations, like teaching and health, were associated with higher student grades, highlighting the potential influence of parental career on educational outcomes.
  - *Internet Access and Study Time*: Consistent internet access and increased study time positively impacted student grades, emphasizing the importance of resources and dedicated study habits.

In-Depth Insights on Student Performance

- **Alcohol Consumption**: The study revealed a nuanced relationship between alcohol consumption and academic performance, with no straightforward negative impact on grades, contrary to common expectations.
- **Travel Time and Family Size**: Surprisingly, factors like travel time to school and family size showed minimal correlation with student grades, suggesting these factors might be less critical than traditionally thought.

Comprehensive Statistical Analysis

- **Hypothesis Testing**: Advanced statistical tests, including T-tests and ANOVA, were applied. For instance, the two-sample T-test comparing mean grades based on absences showed no significant difference, challenging the conventional belief that more absences lead to lower grades.
- **ANOVA for School Performance Comparison**: The ANOVA analysis indicated no significant differences in mean grades among different schools, suggesting that school choice might not be as influential as individual student factors.

Implications for Educational Stakeholders

- **Actionable Strategies for Students**: The analysis offers students concrete areas to focus on for potential academic improvement, such as leveraging parental support and optimizing study habits.
- **Institutional Policy Recommendations**: Educational institutions could benefit from these insights by developing targeted support systems, especially focusing on internet accessibility and study resources.

Concluding Remarks

- **Redefining Educational Success Factors**: This study challenges traditional views on what influences academic success, advocating for a more holistic understanding of student performance.

- **Commitment to Evolving Educational Practices**: It underscores the need for continuous research and adaptation in educational methods to meet the diverse needs of students in a rapidly changing world.