

Exploring the Impact of Family Dynamics, Romantic Status, and Address Type on Academic Performance: A Statistical Analysis

Course: STAT 206: Statistical Analysis

**Submitted by:
Ibraheem Nadeem**

Introduction

A student's academic performance is influenced by a mix of personal effort, family support, and social factors, with family dynamics often being one of the most crucial influences. The family serves as the primary social and emotional support system for a student, providing the foundation for their overall development and ability to navigate academic challenges. Family educational support, such as help with homework or fostering a conducive study environment, and the quality of family relationships, reflecting emotional bonds and interactions, are believed to significantly impact students' academic outcomes.

The focus of this research is to address the question: Does family educational support and family relationship quality significantly affect a student's final grades? To explore this, we leverage the Student Performance Dataset, a publicly available dataset from the UCI Machine Learning Repository. It comprises 395 observations of secondary school students in Portugal, providing detailed information on their academic performance and various socioeconomic, demographic, and behavioral factors. The dataset includes the final grades of students as the response variable, alongside explanatory variables such as family educational support, quality of family relationships, parental education, and household characteristics. By analyzing these variables, we aim to provide insights into the extent of their influence and offer practical implications for educators and families seeking to enhance academic performance.

The variables mentioned below will serve as the key focus points for this research paper, guiding our analysis and findings.

Variable Name	Description
address	Student's home address type (binary: 'U' - urban or 'R' - rural).
famsup	Family educational support (binary: 'yes' or 'no').
romantic	In a romantic relationship (binary: 'yes' or 'no').
famrel	Quality of family relationships (numeric: 1 - very bad to 5 - excellent).
G3	Final grade (numeric: 0 to 20).

Exploratory Data Analysis (EDA)

To prepare the dataset for analysis, adjustments were made to enhance clarity and visualization. The *famsup* variable, indicating family educational support, was re-coded from "yes" and "no" to "Support" and "No Support" for better interpretability. The *famrel* variable, which rates family relationship quality on a scale from 1 to 5, was categorized as a factor. This allows for clearer comparisons across different levels of family relationship quality and supports group-based statistical analyses, avoiding assumptions of a linear relationship.

As part of the exploratory analysis, we created a boxplot (Figure 1) to compare final grades (G3) based on family support (*famsup*). The boxplot reveals that the median grades for both groups are roughly similar, suggesting that family support may not significantly impact the final grades. The interquartile range (IQR) is also comparable between the two groups, indicating similar variations in student performance. However, both groups display outliers on the lower end, representing students who performed significantly below their peers.

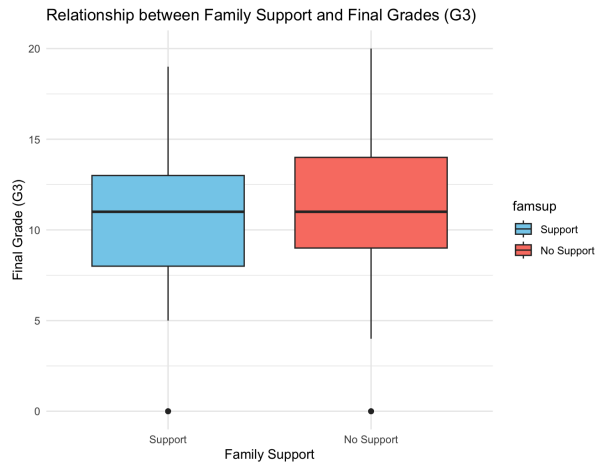


Figure 1: Boxplot Analysis of Final Grades by Family Support

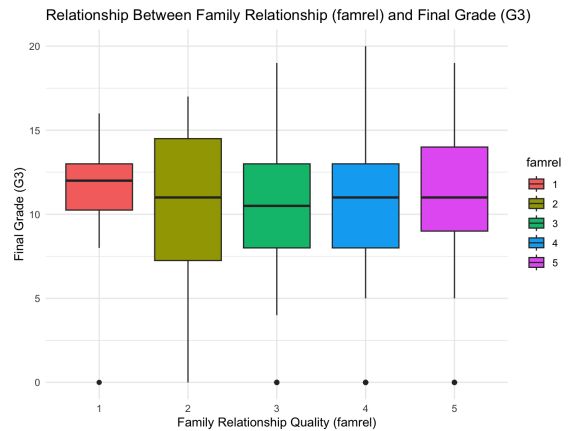


Figure 2: Boxplot Analysis of Final Grades by Family Relationship Quality

Building on the insights gained from analyzing family support, we next explored the boxplot (Figure 2) for the distribution of final grades (G3) across different levels of family relationship quality. The trend reveals a surprising curve: grades peak in Group 1, drop in Group 2, remain lower through Group 3, and then gradually rise again in Groups 4 and 5. Interestingly, students in Group 1, who reported very poor family relationships, have the highest mean grades. This challenges the common assumption that stronger family relationships correlate with better academic performance. In contrast, the median grades for Groups 2 to 5 show only minor fluctuations, remaining relatively steady. The interquartile range (IQR) is widest for levels 2 and 5, indicating greater variability in grades for these groups, whereas group 1 exhibits a narrower IQR, suggesting more uniform academic outcomes among students with very poor family relationships. Additionally, all groups display outliers near 0, indicating some very low-performing students regardless of their family relationship quality.

We next explored how family support (*famsup*) and family relationships interact to influence students' final grades (G3) by creating an interaction plot (Figure 3). The plot revealed that the effect of family relationship quality on final grades depends on family support as the two lines intersect.

The interaction plot and boxplot for family relationships show an unexpected fluctuation in grades between scores of 1 and 2. Upon examining the frequency distribution (Figure 4), it becomes clear that the data points for family relationship scores of 1 and 2 are significantly fewer compared to other levels. This limited representation likely contributes to the irregular patterns observed in the plot, making trends at these lower levels less reliable.

These findings suggest that the relationship between family relationship quality, family support and final grades is more nuanced than initially assumed. Further statistical analysis, such as ANOVA, is necessary to determine if these differences are statistically significant and to better understand the underlying factors influencing this relationship.

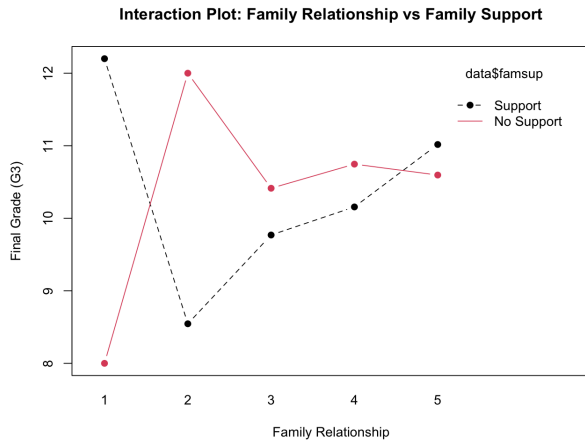


Figure 3: Interaction Plot of Family Relationship, Distribution and Frequency Analysis

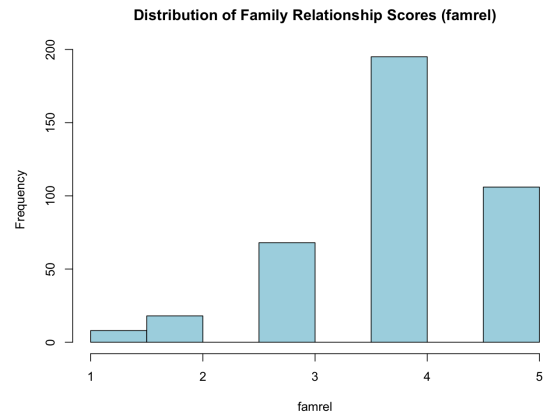


Figure 4: Family Relationship Scores: Family Support, and Final Grades

Statistical Analysis

To assess the relationships and interactions between family support, family relationship quality, and final grades (**G3**), we employed a statistical model with the following equation:

$$G3 = 10.4936 + 0.3503(NoSup) - 0.7410(Rel2) - 0.5989(Rel3) - 0.2550(Rel4) + 0.1812(Rel5)$$

where:

- NoSup : Indicator for students without family support
- Rel2 : Family relationship quality at Level 2
- Rel3 : Family relationship quality at Level 3
- Rel4 : Family relationship quality at Level 4
- Rel5 : Family relationship quality at Level 5

This equation represents a multiple linear regression model where **G3** is the response variable, and *famsup* and *famrel* serve as explanatory variables. The baseline group comprises students with family support and the lowest family relationship quality (*famrel*=1). All coefficients in the model reflect how each factor affects the predicted grade relative to this baseline group. For instance, students without family support (*NoSup*) are predicted to score 0.3503 points higher than those with support, holding all other variables constant. Similarly, the coefficients for *famrel* levels indicate the difference in predicted grades compared to the baseline level (*famrel*=1). Negative coefficients, such as -0.7410 for *Rel2*, *Rel3* and *Rel4* indicate a decrease in predicted grades relative to the baseline. Notably, *Rel5* is the only *famrel* level with a positive coefficient (+0.1812), suggesting an increase in predicted grades compared to the baseline.

To evaluate the statistical significance of these relationships, we conducted an Analysis of Variance (ANOVA) with the following hypotheses. For a two-way interaction ANOVA analyzing the effect of family support (*famsup*) and family relationship quality (*famrel*) on the final grade (**G3**), the hypotheses are as follows:

1. Family support (famsup):

Family support (famsup) has no effect on the final grade (G3):

$$H0: \mu_{Support} = \mu_{No\ Support}$$

Family support (famsup) has an effect on the final grade (G3):

$$H1: \mu_{Support} \neq \mu_{No\ Support}$$

2. Family relationship quality (famrel):

Family relationship quality (famrel) has no effect on the final grade (G3):

$$H0: \mu_{famrel1} = \mu_{famrel2} = \mu_{famrel3} = \mu_{famrel4} = \mu_{famrel5}$$

Family relationship quality (famrel) has an effect on the final grade (G3):

$$H1: \text{At least one } \mu_{famrel} \text{ differs.}$$

3. Interaction Effect (Family support and family relationship quality):

There is no interaction effect between family support and family relationship quality on the final grade (G3):

$$H0: \text{The effect of 'famrel' is the same across levels of 'famsup'.$$

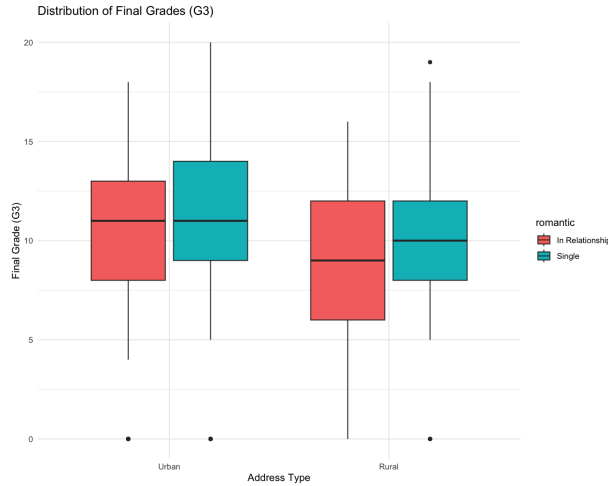
There is an interaction effect between family support and family relationship quality on the final grade (G3)

$$H1: \text{The effect of 'famrel' differs across levels of 'famsup'.$$

A significance level (α) of 0.05 was set for the analysis. The results indicated that family support did not have a statistically significant effect on final grades, with an F-value of 0.601 and a p-value greater than 0.05. Similarly, family relationships also showed no significant effect, with an F-value of 0.383 and a p-value exceeding 0.05. Additionally, the interaction between family relationship and family support was also not statistically significant, with an F-value of 1.180 and a p-value above 0.05. In summary, neither family relationship nor family support independently affects final grades, nor does their combined effect demonstrate statistical significance.

Given the lack of statistically significant effects for family support and family relationship quality on final grades (G3), it becomes essential to explore other potential predictors that might better explain the variation in students' academic performance. The next two variables we analyzed are relationship status and home address type. Relationship status captures whether a student is romantically involved, which can impact focus, time management, and emotional well-being—factors that affect academic performance. Home address type distinguishes between urban and rural environments. Urban students often have better access to educational resources, infrastructure, and extracurricular activities, whereas rural students may face challenges such as limited resources and longer commutes. To enhance clarity and improve data interpretability, the relationship variable was re-coded to "In Relationship" and "Single," and the address variable to "Urban" and "Rural."

The box plot (Figure 5) examines the distribution of final grades (G3) based on two factors: the type of address (Urban or Rural) and relationship status (In Relationship or Single). Urban students appear to have a relatively higher mean final grade compared to rural students, suggesting that address might have a significant influence on final grade. For both Urban and Rural students, the median final grades are similar across relationship statuses, suggesting that being in a relationship or single might not have a strong effect on academic performance. Among Urban students, singles have a slightly wider range of grades compared to those in relationships, with both groups showing outliers at the lower end (grade = 0). For Rural students, the medians and overall ranges for both groups are quite similar, but there is a small outlier at the lower end for students in relationships. Overall, the boxplot suggests that the type of address might influence the distribution of grades more than relationship status, but further analysis would be needed to confirm this observation.



Boxplot Analysis of Address Type and Relationship Status (Figure 5)

To further explore the combined effects of *address type* (Urban or Rural) and *relationship status* (Single or In Relationship) on academic performance, we created an interaction plot (Figure 6). Urban students consistently achieved higher mean grades than Rural students, and Singles outperform those In a Relationship across both address types. The decline in grades from Urban to Rural is more pronounced for students in a relationship, as shown by the steeper slope of the red dashed line. Notably, the two lines do not intersect, indicating that the effect of address type on final grades is consistent across romantic statuses. Overall, the boxplot and interaction plot combined suggests that while address type and romantic status independently influence final grades, there is no interaction effect between these two factors.

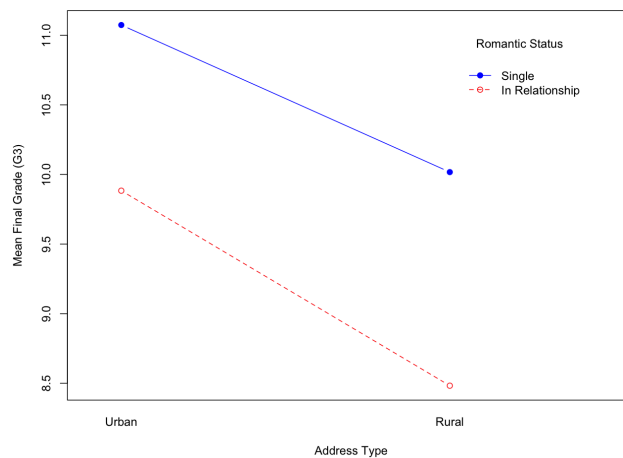


Figure 6: Interaction Plot of Romantic Status, Address Type, and Final Grades

To assess the relationships and interactions between *address type* (Urban or Rural), *romantic status* (Single or In Relationship), and final grades (*G3*), we employed the following statistical model:

$$G3 = 9.8835 - 1.4007(Rural) + 1.1900(Single) + 0.3442(Rural * Single)$$

Where:

Rural : Indicator for students living in rural areas.

Single : Indicator for students not in a romantic relationship.

Rural * Single : Interaction effect between rural address and being single.

The baseline group in this model consists of Urban students who are in a romantic relationship, with the intercept 9.88359 representing their predicted grade. Each coefficient reflects the difference in grades relative to this baseline group. Specifically, students living in rural areas score 1.4007 points lower on average compared to urban students in a relationship, holding romantic status constant. In contrast, students who are not in a romantic relationship (Single) score 1.1900 points higher on average compared to the baseline group, holding address constant. The interaction term (+0.3442) indicates that rural students who are also single experience an additional increase of 0.3442 points in their grades. This interaction suggests that being single slightly mitigates the negative impact of living in a rural area.

To evaluate the statistical significance of these relationships, we conducted an Analysis of Variance (ANOVA) analyzing the effect of *address type* (*address*) and *romantic status* (*romantic*) on the final grade (*G3*), with the following hypotheses:

1. Address type (*address*):

Address type (*address*) has no effect on final grades:

$$H0: \mu_{Urban} = \mu_{Rural}$$

Address type (*address*) has an effect on final grades:

$$H1: \mu_{Urban} \neq \mu_{Rural}$$

2. Romantic status (*romantic*):

Romantic status (*romantic*) has no effect on final grades:

$$H0: \mu_{In Relationship} = \mu_{Single}$$

Romantic status (*romantic*) has an effect on final grades:

$$H1: \mu_{In Relationship} \neq \mu_{Single}$$

3. Interaction Effect (*address type* and *romantic status*):

There is no interaction between address type and romantic status on final grades:

$$H0: \text{The effect of 'address' is independent of 'romantic'}$$

There is an interaction between address type and romantic status on final grades:

$$H1: \text{The effect of 'address' depends on 'romantic'}$$

The results showed that address type significantly influenced final grades ($F = 4.501$, $p < 0.05$), with Urban students performing better than Rural students. Romantic status also had a significant effect ($F = 6.857$, $p < 0.01$), as Single students outperformed those in relationships. However, the interaction between address type and romantic status was not significant ($F = 0.087$, $p > 0.05$), indicating that the effects of address type and romantic status on grades are independent of each other.

To further investigate the significant effects identified in the two-way factorial ANOVA, we conducted a Tukey HSD (Honestly Significant Difference) test. This post-hoc analysis was used to evaluate pairwise comparisons among the levels of *address*, *romantic status*, and their interaction (*address:romantic*) to better understand how these factors influence final grades (G3). The Tukey HSD test provides a clearer picture of which specific group differences contribute to the observed main and interaction effects.

The results revealed that Rural students have significantly lower grades than Urban students, with a mean difference of -1.16 ($p = 0.0345$). For romantic status, Single students achieved significantly higher grades than those in a Relationship, with a mean difference of 1.27 ($p = 0.0092$). Regarding interaction effects, the only statistically significant pairwise comparison was between Urban Singles and Rural students in a Relationship. Urban Singles performed significantly better, with a mean difference of 2.59 ($p = 0.0217$). No other interaction comparisons were statistically significant.

These findings suggest that both *address* and *romantic status* independently influence final grades, while interaction effects are limited to specific groups. This highlights the nuanced relationships between individual and environmental factors in shaping academic performance and emphasizes the importance of examining these effects at both individual and combined levels.

It is essential to assess whether the underlying assumptions of the model are satisfied. To check the assumption of homoscedasticity (equal variance of residuals), we created the residuals vs. fitted plot (Figure 5). The residuals are scattered evenly around the horizontal line at zero, with no clear patterns or trends, indicating that this assumption is reasonably met. While there is slight clustering of residuals near specific fitted values and one outlier (labeled 143) with a large residual, these do not significantly distort the overall trend. The Q-Q plot of residuals (Figure 6) was used to assess the assumption of normality, the plot shows that most points lie close to the diagonal line, indicating that the residuals are approximately normally distributed. Although minor deviations are observed in the tails, particularly at the lower end, the normality assumption appears to hold sufficiently for ANOVA. Additionally, the assumption of independence of observations is supported by the study design, as the data were collected from individual students, with no overlap or dependency between observations. Overall, the residual plots (Figures 4 and 5) and study design suggest that the assumptions of normality, homoscedasticity, and independence are met, making the data suitable for conducting a two-way ANOVA.

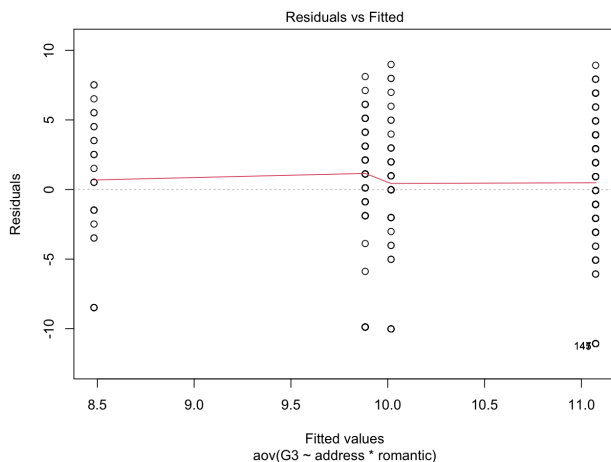


Figure 5: Residuals vs. Fitted Plot

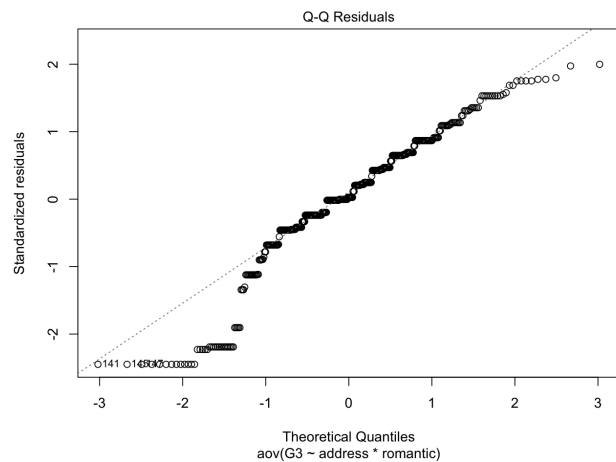


Figure 6: Q-Q Plot of Residuals

Our findings indicate that family support and family relationship quality, as measured in this dataset, do not significantly impact final grades. This suggests that broader family dynamics may have less direct influence on academic outcomes than anticipated. However, romantic status and address type emerged as significant predictors. Single students consistently outperformed those in a relationship, highlighting the importance of personal social dynamics in academic performance. Additionally, Urban students achieved higher grades than their Rural counterparts, suggesting that environmental factors such as access to resources or infrastructure may play a role. While an interaction effect between address type and romantic status was not statistically significant, pairwise comparisons revealed specific cases, such as Urban Singles outperforming Rural students In a Relationship, where combined factors influenced grades.

Future research could explore additional predictors such as study habits, peer influence, or school resources to provide a more comprehensive understanding of the determinants of academic success. Moreover, addressing limitations such as the presence of outliers or potential biases in the dataset could further refine these findings. By identifying and understanding the nuanced relationships between students' contexts and their academic outcomes, educators and policymakers can develop more targeted interventions to support student success.

References

Cortez, P., & Silva, A. M. G. (2014). Student Performance Data Set. UCI Machine Learning Repository.
<https://archive.ics.uci.edu/ml/datasets/student+performance>