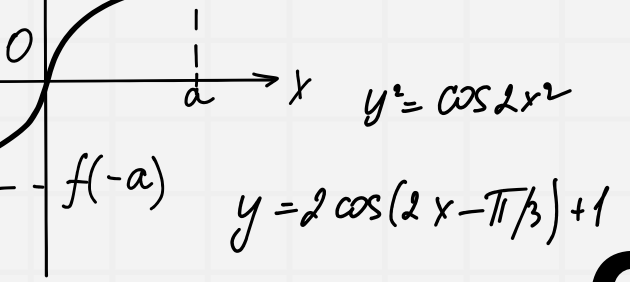
Balancing Social Variables: The Impact of Social Factors on Math Performance

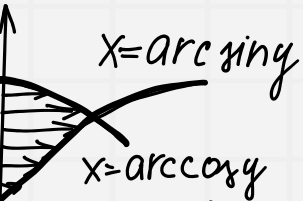
Jane Condon, Valentina Tillmann

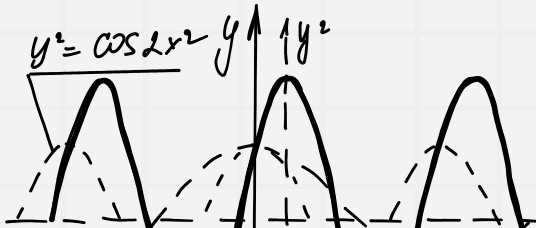


Questions of Interest

Hypothesis 1: Is Having a More Active Social Life Associated with a Better or Worse Mathematics Grade?

Hypothesis 2: What Kind of Socioeconomic and Demographic Factors Have the Strongest Effect on Mathematics Grade?



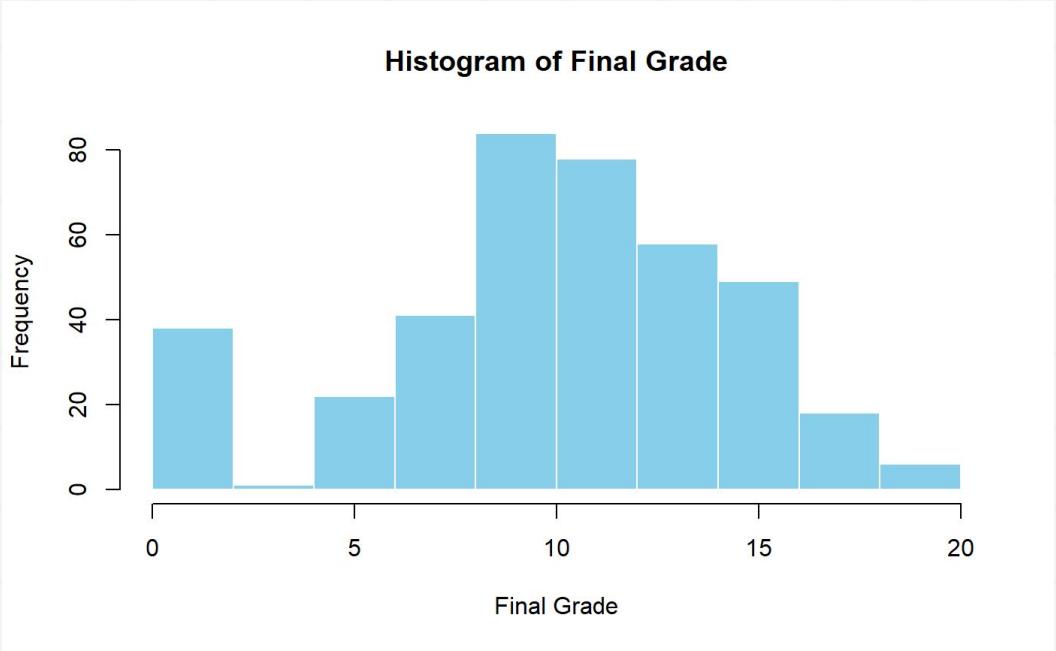


Our Data

In our analysis, we used a dataset titled "High School Student Performance & Demographics" from Kaggle.com. It contains data from two Portuguese high schools regarding student performance in Mathematics and Portuguese courses as well as social and demographic factors. It contains 34 columns and 395 rows, with the index being "Student ID." There are 21 categorical variables and 13 numerical variables.

Dependent Variable: Mathematics Final Grade

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	8.00	11.00	10.42	14.00	20.00



Independent Variables

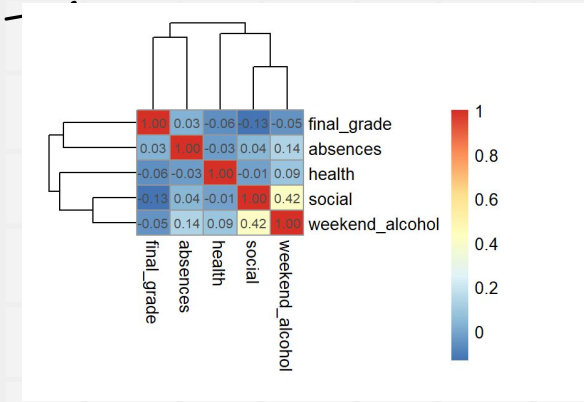
Hypothesis 1:

- **Romantic Relationship:** Whether or not a student is involved in a romantic relationship. (binary: yes or no)
- **Extracurricular Activities:** Whether or not a student is involved in extracurricular activities. (binary: yes or no)
- **Weekend Alcohol Consumption:** Student's level of alcohol consumption on weekends. (scale of 1 to 5, with 1 = very low and 5 = very high)
- **Frequency of Social Outings:** How often a student goes out with friends. (scale of 1 to 5, with 1 = very infrequently and 5 = very frequently)

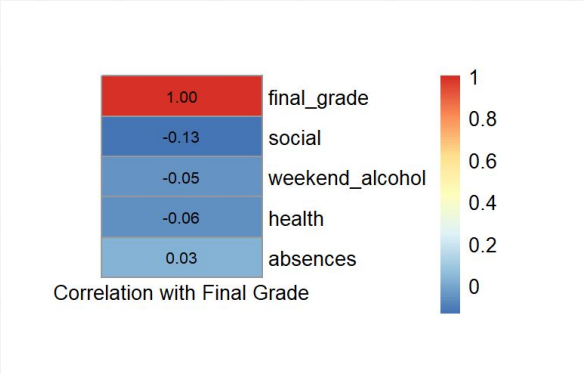
Hypothesis 2:

- **Address Type:** Type of area that a student resides in. (binary: urban or rural)
- **Family Support:** Whether or not a student's family provides them with educational support. (binary: yes or no)
- **Health:** Student's current health status. (scale of 1 to 5, with 1 = very poor and 5 = very good)
- **Internet Access:** Whether or not student has reliable access to the internet at home (binary: yes or no)
- **Mother's Education:** Student's mother's education level. (5 levels ranging from "none" to "higher education")
- **Father's Education:** Student's father's education level. (5 levels ranging from "none" to "higher education")
- **Extra Paid Classes:** Whether a student participates in extra (paid) mathematics courses outside of school. (binary: yes or no)
- **Parent Status:** Student's parents' cohabitation status. (binary: "Living together" or "Living apart")

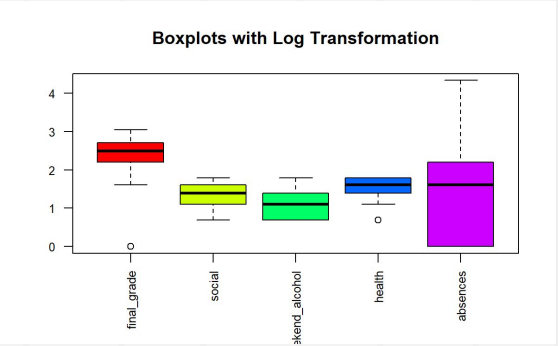
Exploratory Data Analysis: Numeric Data



Weekend alcohol consumption and social score exhibit a moderate correlation of 0.42, suggesting a meaningful but not strong relationship between these two variables. Other variables do not show significant pairwise correlations, indicating that there are no notable linear relationships among the remaining numeric variables in the dataset.

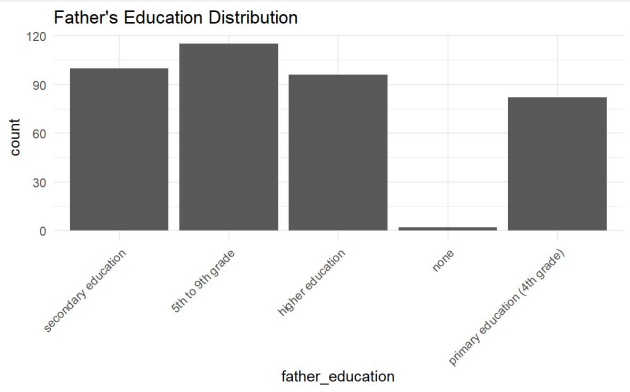


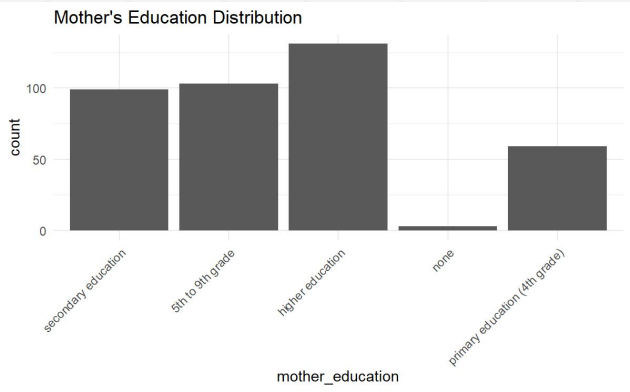
Above is the heatmap illustrating the relationship between various numeric variables in the dataset and the final grade. Among the correlations, the most notable correlation, compared to other variables, is associated with the social score, which indicates a weak negative correlation.



'final_grade' and 'social' appear approximately symmetric, while 'weekend_alcohol', 'health', and 'absences' do not. The log transformation is applied due to the severity of outliers in the 'absences' column.

Exploratory Data Analysis: Factor Variables





The nature of factor variables requires us to represent them on a different plot. Shown above are bar charts for our various two-level factor variables. Note that the counts for both categories are not significantly different for 'activities' and 'extra_paid_classes'. However, 'for address_type' the data is skewed toward 'Urban' for address type, which could mean that the two schools are located in cities. Furthermore, 'family_support' is slightly skewed towards 'yes', 'internet_access' towards 'yes', 'parent_status' toward 'Living Together', and 'romantic_relationship' towards 'no'.

There are more mothers with higher education than mothers with only high school, middle school, or primary education. There are slightly more fathers with a high school education than the other levels of education. For both mothers and fathers, there are very few that have no education whatsoever.

Hypothesis 1: Methods

Is Having a More Active Social Life Associated with a Better or Worse Mathematics Grade?

Two-Sample T-Test

We will use a two-sample Wilcoxon rank sum test to decipher the relationship between mathematics grade and involvement in a romantic relationship. We will repeat for involvement in extracurricular activities.

ANOVA Test

We will use a Kruskal-Wallis test to decipher the relationship between mathematics grade and weekend alcohol consumption. We will repeat for frequency of social outings.

Two-Sample T-Test (Wilcoxon Rank Sum Test)

Involvement in Romantic Relationship

H_0 = Mathematics grade is equally distributed across the two groups (students in a relationship vs students who are not in a relationship).
H_a = Mathematics grade is not equally distributed across the two groups (students in a relationship vs students who are not in a relationship).

```
        Wilcoxon rank sum test with continuity correction

data:  final_grade by romantic_relationship
W = 19293, p-value = 0.06953
alternative hypothesis: true location shift is not equal to 0
```

Involvement in Extracurricular Activities

H_0 = Mathematics grade is equally distributed across the two groups (students involved in activities vs students not involved in activities).
H_a = Mathematics grade is not equally distributed across the two groups (students involved in activities vs students not involved in activities).

```
        Wilcoxon rank sum test with continuity correction

data:  final_grade by activities
W = 18912, p-value = 0.6049
alternative hypothesis: true location shift is not equal to 0
```

Since our data is not normally distributed, we have opted to use **Wilcoxon Rank Sum test**, as opposed to the standard two sample t-test. The results of the tests are as follows:

- **Romantic relationship**: At $\alpha = 0.05$, we fail to reject the null hypothesis. However, at $\alpha = 0.10$, we can **reject the null hypothesis** and conclude that there may be a relationship between involvement in a romantic relationship and mathematics final grade. After further analysis, we conclude that students **involved** in a romantic relationship tend to achieve **higher** mathematics grades than those who are not.
- **Extracurricular Activities:** With a very large **p-value** of **0.6049**, we **fail to reject the null hypothesis** at both $\alpha = 0.05$ and $\alpha = 0.10$. Thus, we can conclude that there is not a significant relationship between involvement in extracurricular activities and mathematics final grade.

ANOVA Test (Kruskal Wallis Test)

Weekend Alcohol Consumption

H_0 = The median mathematics grade is equal across all levels of weekend alcohol consumption.
H_a = At least one group (weekend alcohol consumption level) has a median mathematics grade that differs from the others.

```
        Kruskal-Wallis rank sum test

data:  final_grade by factor(weekend_alcohol)
Kruskal-Wallis chi-squared = 5.453, df = 4, p-value = 0.2439
```

Frequency of Social Outings

H_0 = The median mathematics grade is equal across all levels of social outing frequency.
H_a = At least one group (level of social outing frequency) has a median mathematics grade that differs from the others.

```
        Kruskal-Wallis rank sum test

data:  final_grade by factor(social)
Kruskal-Wallis chi-squared = 14.697, df = 4, p-value = 0.005372
```

Since our data is not normally distributed, we have opted to use the **Kruskal Wallis test**, as opposed to the standard ANOVA test. The results of the tests are as follows:

- **Weekend Alcohol Consumption**: At $\alpha = 0.05$ and $\alpha = 0.10$, we **fail to reject** the null hypothesis.Thus, we can conclude that there is insufficient evidence that there is a relationship between mathematics grade and weekend alcohol consumption.
- **Frequency of Social Outings:** With a **p-value** of **0.005372**, we can **reject** the null hypothesis at $\alpha = 0.05$. Thus, we can conclude that there is a **significant relationship** between mathematics grade and frequency of social outings.
  - Our **post-hoc analysis** indicates that students who attend social outings more **frequently** achieve a **lower** mathematics grade.

Hypothesis 2: Methods

What Kind of Socioeconomic and Demographic Factors Have the Strongest Effect on Mathematics Grade?

To explore this hypothesis, we will use multiple linear regression. The regression equation will be as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \epsilon$$

Where $\mathbf{X_1}$ = Address Type, $\mathbf{X_2}$ = Family Support, $\mathbf{X_3}$ = Health, $\mathbf{X_4}$ = Internet Access, $\mathbf{X_5}$ = Mother's Education Level, $\mathbf{X_6}$ = Father's Education Level, $\mathbf{X_7}$ = Extra Paid Classes, $\mathbf{X_8}$ = Parent Status and $\mathbf{Y}$ = Mathematics Final Grade

First, we will use a standard multiple linear regression model with the original data. Next, we will apply a Box-Cox transformation to Y and repeat. Finally, we will use a normally distributed subset of the data and repeat.

Multiple Linear Regression

```
Residuals:
     Min       1Q   Median       3Q      Max
-12.1937  -2.0680   0.6614   2.9567   8.8625

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                10.5713     1.2140   8.708   <2e-16
address_typeUrban                           0.6058     0.5612   1.079   0.2811
family_supportyes                          -1.0086     0.4919  -2.050   0.0410
health                                     -0.1192     0.1655  -0.720   0.4720
internet_accessyes                          0.5154     0.6391   0.807   0.4205
mother_education5th to 9th grade           -0.5711     0.6620  -0.863   0.3889
mother_educationhigher education            1.1479     0.6509   1.764   0.0786
mother_educationnone                        2.6136     2.6492   0.987   0.3245
mother_educationprimary education (4th grade) -1.3121  0.8329  -1.575   0.1160
father_education5th to 9th grade            0.1311     0.6488   0.202   0.8400
father_educationhigher education            0.3824     0.6784   0.564   0.5733
father_educationnone                        2.3130     3.2158   0.719   0.4724
father_educationprimary education (4th grade) -0.2855  0.7724  -0.370   0.7119
extra_paid_classesyes                       0.7972     0.4830   1.650   0.0997
parent_statusLiving together               -0.5794     0.7580  -0.764   0.4451
```

```
(Intercept)                                ***
address_typeUrban
family_supportyes                          *
health
internet_accessyes
mother_education5th to 9th grade
mother_educationhigher education           .
mother_educationnone
mother_educationprimary education (4th grade)
father_education5th to 9th grade
father_educationhigher education
father_educationnone
father_educationprimary education (4th grade)
extra_paid_classesyes                      .
parent_statusLiving together
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.464 on 380 degrees of freedom
Multiple R-squared:  0.08453,    Adjusted R-squared:  0.0508
F-statistic: 2.506 on 14 and 380 DF,  p-value: 0.002004
```

- **Extra paid classes**, **mother's education**, and **family support** were the only variables to have a significant relationship with a student's final mathematics grade at a significance level of **0.10**.
- The predictive power of this model is very weak, with an adjusted **R-squared** value of **0.0508**.
- Unfortunately, this model is not entirely valid or reliable, as two of the linear regression assumptions have been violated. To address this, we will apply a Box-Cox transformation to our response variable and repeat this model.

Modified Multiple Linear Regression (Box-Cox Transformation)

```
Residuals:
     Min       1Q   Median       3Q      Max
-20.5720  -4.4922   0.6578   5.2507  17.9500

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      17.4679     2.1990   7.943 2.25e-14
address_typeUrban                                 1.1768     1.0166   1.158   0.2477
family_supportyes                                -1.8439     0.8910  -2.070   0.0392
health                                           -0.2661     0.2998  -0.887   0.3754
internet_accessyes                                1.0378     1.1576   0.897   0.3705
mother_education5th to 9th grade                 -1.0588     1.1991  -0.883   0.3778
mother_educationhigher education                  2.1622     1.1790   1.834   0.0674
mother_educationnone                              4.7572     4.7987   0.991   0.3221
mother_educationprimary education (4th grade)    -2.5235     1.5087  -1.673   0.0952
father_education5th to 9th grade                  0.4216     1.1752   0.359   0.7200
father_educationhigher education                  0.9282     1.2289   0.755   0.4505
father_educationnone                              4.3351     5.8250   0.744   0.4572
father_educationprimary education (4th grade)    -0.4116     1.3992  -0.294   0.7688
extra_paid_classesyes                             1.1397     0.8750   1.303   0.1935
parent_statusLiving together                     -0.9646     1.3731  -0.703   0.4828
```

```
(Intercept)                                      ***
address_typeUrban
family_supportyes                                *
health
internet_accessyes
mother_education5th to 9th grade
mother_educationhigher education                 .
mother_educationnone
mother_educationprimary education (4th grade)    .
father_education5th to 9th grade
father_educationhigher education
father_educationnone
father_educationprimary education (4th grade)
extra_paid_classesyes
parent_statusLiving together
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.085 on 380 degrees of freedom
Multiple R-squared:  0.08877,    Adjusted R-squared:  0.0552
F-statistic: 2.644 on 14 and 380 DF,  p-value: 0.001088
```

- In this model, the variables that have a statistically significant relationship with mathematics grade are **family support** ($\alpha = 0.05$), and **mother's education** (higher education and primary education, $\alpha = 0.10$).
- The predictive power of this model is still very weak, with an adjusted R-squared value of 0.0552.
- Unfortunately, this model is not entirely valid or reliable either, as the linearity and normality assumptions have still been violated.

Modified Multiple Linear Regression: Zeroes Removed

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.6353 -2.2431 -0.1212  2.0249  8.3484

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     11.17917    0.87560  12.767   <2e-16
address_typeUrban                                0.67487    0.41778   1.615   0.1071
family_supportyes                               -0.61332    0.36423  -1.684   0.0931
health                                          -0.18322    0.12127  -1.511   0.1318
internet_accessyes                               0.58258    0.47339   1.231   0.2193
mother_education5th to 9th grade                -0.10729    0.49046  -0.219   0.8270
mother_educationhigher education                 0.67321    0.47722   1.411   0.1592
mother_educationnone                             1.42415    1.86683   0.763   0.4461
mother_educationprimary education (4th grade)   -0.90521    0.62066  -1.458   0.1456
father_education5th to 9th grade                 0.53994    0.47494   1.137   0.2564
father_educationhigher education                 0.95396    0.49802   1.915   0.0563
father_educationnone                             1.47391    2.26515   0.651   0.5157
father_educationprimary education (4th grade)   -0.07328    0.57235  -0.128   0.8982
extra_paid_classesyes                           -0.33976    0.35521  -0.957   0.3395
parent_statusLiving together                     0.04843    0.55106   0.088   0.9300
```

```
(Intercept)                                     ***
address_typeUrban
family_supportyes                               .
health
internet_accessyes
mother_education5th to 9th grade
mother_educationhigher education
mother_educationnone
mother_educationprimary education (4th grade)
father_education5th to 9th grade
father_educationhigher education                .
father_educationnone
father_educationprimary education (4th grade)
extra_paid_classesyes
parent_statusLiving together
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.138 on 342 degrees of freedom
Multiple R-squared:  0.09208,   Adjusted R-squared:  0.05491
F-statistic: 2.477 on 14 and 342 DF,  p-value: 0.002352
```

- The mathematics final grade variable has many values of **'0'**. Since a final grade of 0 in a course generally means that a student did not complete the course or had their score voided, these scores are most likely irrelevant. Thus, we will remove all rows where final grade is equal to 0, and repeat the linear regression model.
- With zero values removed, the results slightly differ. The independent variables that have a statistically significant relationship with final mathematics grade at the **0.10** significance level are **family support** and **father's education level.**
- Fortunately, this model passes all three of the linear regression assumptions, and can be considered to be a valid and reliable model.

Effect of Missing Data: Simulating MCAR and MNAR Data

MCAR

MCAR is defined as values “Missing Completely at Random,” meaning that there is no pattern to the missing data. We simulated this by selecting 20% of all rows to contain an NA value for at least one column, and used the listwise deletion method, i.e., removing those rows, to deal with missing data.

MNAR

MNAR is defined as “Missing Not at Random,” and occurs when the likelihood of a value being missing depends on the value of the missing data itself or on unobserved factors that are not part of the dataset. We simulated this by selecting the bottom 20% of mathematics grades to be missing. Why? If a student is likely to get a failing grade in the course, they are more likely to withdraw from the course or not show up for their exams, resulting in a missing value for their final course grade. We dealt with this by using the multiple imputation method.

Retesting our Hypotheses with “MCAR” and “MNAR” Data

MCAR

Due to the way that we have dealt with our MCAR data, we are now left with a smaller sample size. A smaller sample size reduces the power of the test, i.e., we are less likely to reject the null hypothesis. Thus, we expect to see slightly higher p-values, but we do not expect to see any significant changes.

MNAR

It is near impossible to avoid bias when dealing with MNAR data, unless we use specialized methods such as pattern-mixture models or selection models to model the missingness explicitly. Due to this, we expect that our results may be significantly different.

Effects of MCAR Data On Our Results

Key Findings:

- **Romantic Relationship and Math Score**:
  - Wilcoxon test results: No significant difference in median math grades.
  - p>0.05; conclusion remains unchanged.
- **Activities and Math Score**:
  - Wilcoxon test results: No significant difference in median math grades.
  - p>0.05; lower power due to reduced sample size.
- **Weekend Alcohol Consumption and Math Score**:
  - Kruskal-Wallis test results: No significant relationship.
  - p>0.05; no post-hoc analysis needed.
- **Social Outing Frequency and Math Score**:
  - Kruskal-Wallis test results: Significant relationship (p<0.05).
  - Post-hoc analysis required for further insights.
  - Tukey test p>0.05; lower power due to reduced sample size.
- **Multiple Linear Regression**:
  - MCAR data yields slightly different results.
  - Father's education (p<0.10): Students with a father holding a bachelor's degree or higher tend to score higher in math.

Effects of MNAR Data On Our Results

Key Findings:

- **Romantic Relationship and Math Score**
  - Wilcoxon test results: No significant difference in median math grades.
  - $p > 0.05$; conclusion remains unchanged.
- **Activities and Math Score**
  - Wilcoxon test results: No significant difference in median math grades.
  - $p > 0.05$; conclusion remains unchanged.
- **Weekend Alcohol Consumption and Math Score**
  - Kruskal-Wallis test results: Significant relationship ($p < 0.05$).
  - Post-hoc Tukey test: Significant difference between weekend_alcohol = 4 and weekend_alcohol = 1.
- **Social Outing Frequency and Math Score**
  - Kruskal-Wallis test results: No significant relationship.
  - $p > 0.05$; no post-hoc analysis needed.
- **Multiple Linear Regression**
  - MNAR data yields slightly different results.
  - Significant predictors at 0.10 level:
    - Mother's education level.
    - Health.

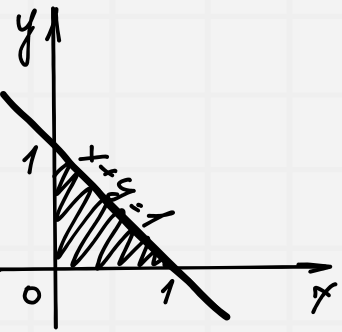
Conclusion

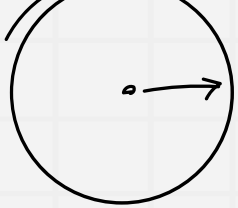
Social Factors and Math Performance

- Not much impact from individual factors (e.g., extracurricular activities, alcohol consumption, etc.). However, a few factors are significant:
- **Social Life**:
    a. Students involved in a **romantic relationship** are predicted to achieve **higher** mathematics grades compared to their peers.
    b. Students who attend social outings **very frequently** are predicted to achieve **lower** mathematics grades compared to their peers.
- **Socioeconomic Status:**
    a. Students whose **father** has completed a **Bachelor's degree** or higher are predicted to achieve **higher** mathematics grades compared to their peers.
    b. Students who receive more **family support** in the field of mathematics are predicted to achieve **lower** mathematics grades compared to their peers.
- Significant collective impact of social factors on math performance (p-value = 0.002352).

Conclusion and Future Directions

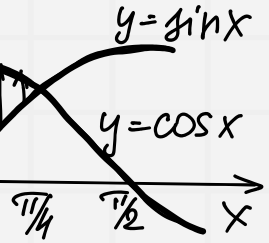
- Social factors as a whole affect math performance.
- Future research should consider geographical and cultural factors.
- Limited to Portuguese data; findings may not generalize globally.

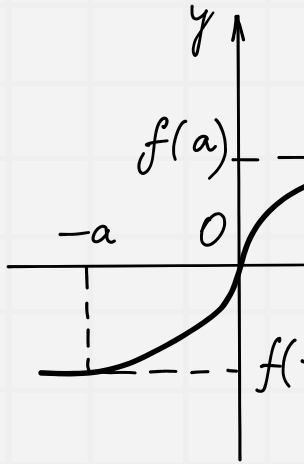




$S = 2\pi R$

Thanks!





CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon** and infographics & images by **Freepik**

Please keep this slide for attribution

$2\sqrt{y^2 - x^2}$