STUDENT PERFORMANCE ANALYSIS

Abstract

This study investigates the relationships between various student attributes and course performance. Paulo Cortez, a data analyst at the University of Minho in Portugal, gathered this data. An examination of student data finds substantial correlations between their final grade (G3) and a variety of parameters. Notably, G2 and G1 (past grades) have a strong positive link with G3, emphasizing the necessity of continuous performance throughout the academic year. Also, the amount of time students devote to studying (study-time) has a positive effect on their grades. In contrast, characteristics such as the number of previous class failures and age correlate with the final grade as well. These findings highlight the complexities of academic achievement and the impact of both internal and external influences on student outcomes.

Introduction

The achievements of an educational institution's students are frequently used to measure its success. With growing competition in the educational field worldwide, there is an urgent need to comprehend the factors that influence student success. Joseph Johnson, owner of GP & MS School, wishes to determine these elements. Therefore, he has commissioned data scientist Patrick Chaccour to look into a dataset of the students' information in order to uncover patterns and connections that may influence their grades. Understanding these factors allows him to make informed decisions to improve the learning environment and provide targeted support.

Key Business Questions

- ➢ Does the school in which a student attends have an impact on their grades?
- ➢ Does gender affect student grades?
- ➢ Does a student's living environment affect their grades?
- ➢ Does the quantity of free time have an effect on a student’s final grade?

Hypotheses

H_0 → There are no grade differences between students from various schools.

H_1 → Students from different schools have significantly varying grades.

H_0 → There are no grade differences between male and female students.

H_1 → Male and female students achieve significantly different grades.

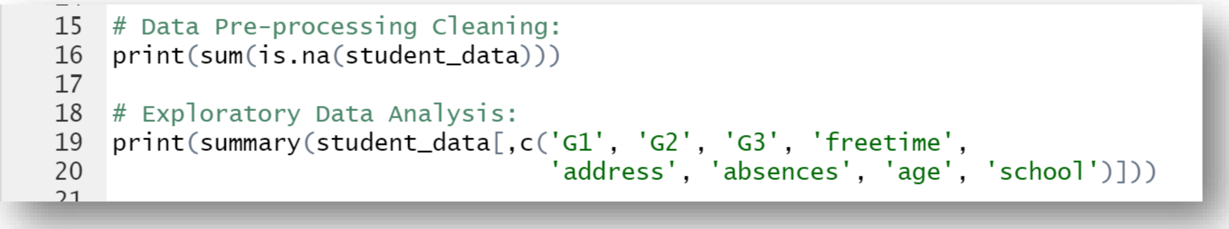
H_0 → There are no grade differences between students from urban and rural locations.

H_1 → There is a huge gap in grades between urban and rural kids.

H_0 →The amount of free time students have, has no effect on their grades.

H_1 → The quantity of free time students have, has a significant impact on their grades.

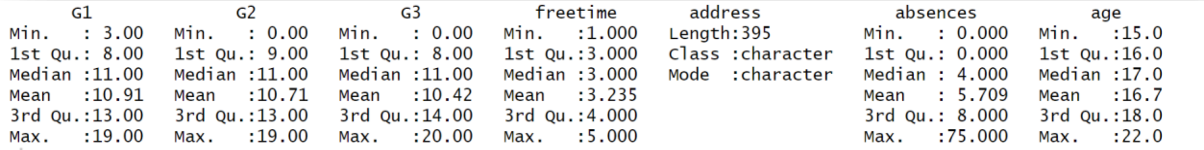
Data Preparation, Sampling, and Cleaning:



```r
# Data Pre-processing Cleaning:
print(sum(is.na(student_data)))

# Exploratory Data Analysis:
print(summary(student_data[,c('G1', 'G2', 'G3', 'freetime',
                              'address', 'absences', 'age', 'school')]))
```

According to the results in my R code, we see that there are no missing values in the dataset.

Descriptive Statistics Observations:



```
      G1              G2              G3            freetime        address              absences            age      
 Min.   : 3.00   Min.   : 0.00   Min.   : 0.00   Min.   :1.000   Length:395         Min.   : 0.000   Min.   :15.0  
 1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:3.000   Class :character   1st Qu.: 0.000   1st Qu.:16.0  
 Median :11.00   Median :11.00   Median :11.00   Median :3.000   Mode  :character   Median : 4.000   Median :17.0  
 Mean   :10.91   Mean   :10.71   Mean   :10.42   Mean   :3.235                      Mean   : 5.709   Mean   :16.7  
 3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00   3rd Qu.:4.000                      3rd Qu.: 8.000   3rd Qu.:18.0  
 Max.   :19.00   Max.   :19.00   Max.   :20.00   Max.   :5.000                      Max.   :75.000   Max.   :22.0  
```

Screenshot extracted from the printout of my R code.

Age: Students' ages range from 15 to 22, with most students being around 16-18 years old.

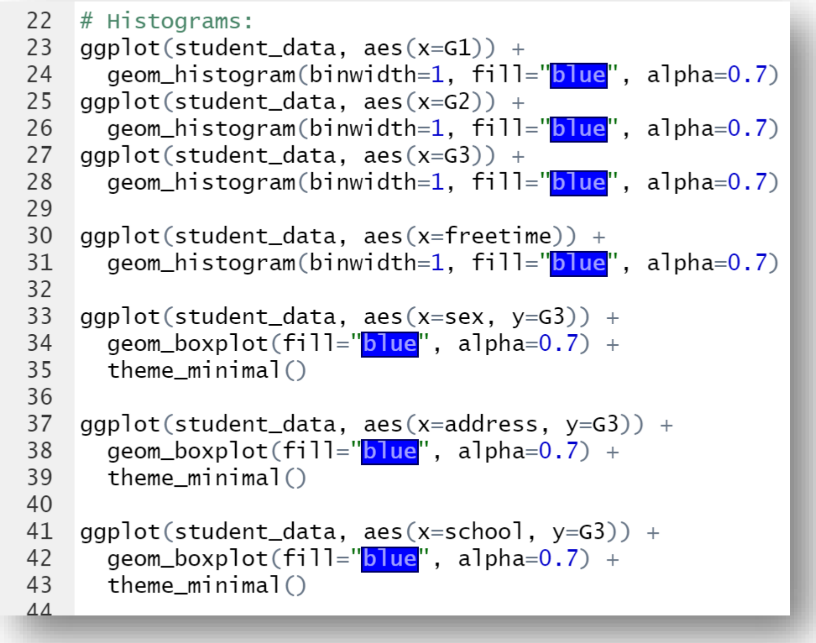
Absences: On average, students have about 5.7 absences, but the maximum number of absences is 75, which could be considered an outlier.

Grades (G1, G2, G3): The grades range from 0 to 20, with an average grade of around 10-11 for all three periods.

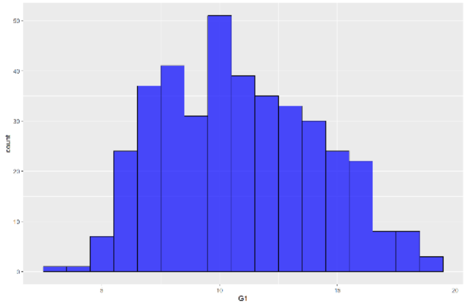
Free time: Students often have a substantial amount of free time, as seen by a mean score of around 3.235 on a scale of 1 to 5.

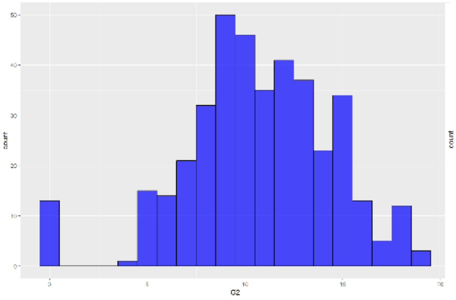
Exploratory Data Analysis:

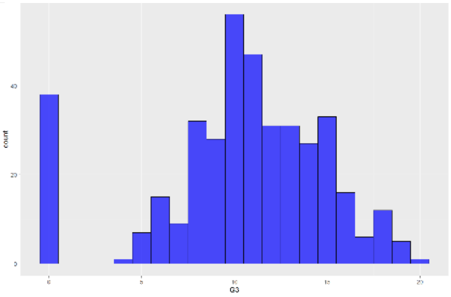
Statistical Visualization:

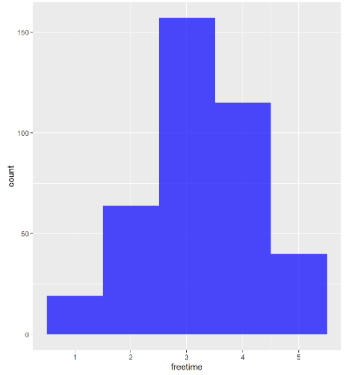


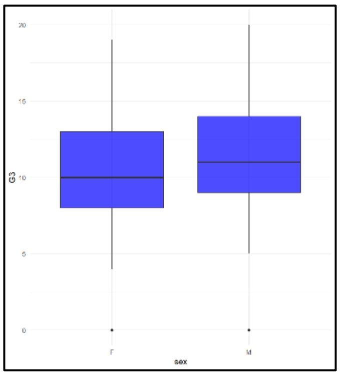
```r
# Histograms:
ggplot(student_data, aes(x=G1)) +
  geom_histogram(binwidth=1, fill="blue", alpha=0.7)
ggplot(student_data, aes(x=G2)) +
  geom_histogram(binwidth=1, fill="blue", alpha=0.7)
ggplot(student_data, aes(x=G3)) +
  geom_histogram(binwidth=1, fill="blue", alpha=0.7)

ggplot(student_data, aes(x=freetime)) +
  geom_histogram(binwidth=1, fill="blue", alpha=0.7)

ggplot(student_data, aes(x=sex, y=G3)) +
  geom_boxplot(fill="blue", alpha=0.7) +
  theme_minimal()

ggplot(student_data, aes(x=address, y=G3)) +
  geom_boxplot(fill="blue", alpha=0.7) +
  theme_minimal()

ggplot(student_data, aes(x=school, y=G3)) +
  geom_boxplot(fill="blue", alpha=0.7) +
  theme_minimal()
```

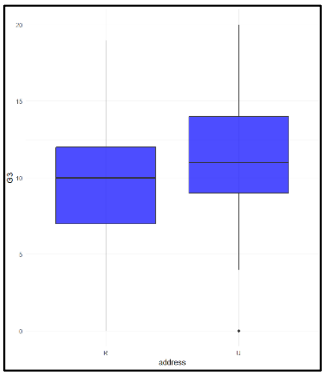


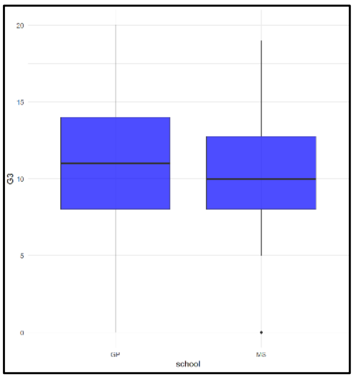












Exploratory Data Analysis:

Grades: G1, G2, and G3 grade distributions are somewhat similar. The majority of students score between 8 and 15, with a significant count receiving a grade of 0 in G3.

Free time: Most students have a reasonable amount of free time, with 3 and 4 being the most prevalent values.

G3 grades by sex: According to the boxplot, male students had a somewhat broader range of grades in G3 than female students. Both genders appear to have similar median scores.

G3 grades by Address: According to the boxplot, students who live in urban environments tend to score higher than students from rural environments.

G3 grades by School: Students from both GP and MS schools seem to score similarly. As we can see in the visualization the majority score around the 11 mark.

Data Pre-processing, Sampling, and Cleaning:

Based on our preliminary examination, the data appears to be relatively clean. There are no missing data, and the descriptive statistics did not disclose any significant outliers.

Inferential Statistical Test:

Given our hypothesis, the ANOVA test would be appropriate because we are comparing means across many groups. Before we perform the test, we must verify its assumptions: normality, variance homogeneity, and observation independence.

First Hypothesis

Step 1: Using the Shapiro-Wilk test for each group in our first hypothesis we will test the normality of G3.

Results:

- GP school –

Test Statistic: 0.9271

p-value: 5.11 x 10^{-12}

- MS school –

Test Statistic: 0.9318

p-value: 9.8 x 10^{-3}

Given standard significance levels, both schools' p-values are less than 0.05. That means, the 'G3' grades do not follow a normal distribution.

Step 2: While using the Bartlett's test, we will check the assumption of variance homogeneity.

Results:

Test Statistic: 0.5810

p-value: 0.4459

The p-value is greater than 0.05. This implies that we lack sufficient data to reject the null hypothesis that variances are identical across groups. Consequently, the assumption of variance homogeneity is met.

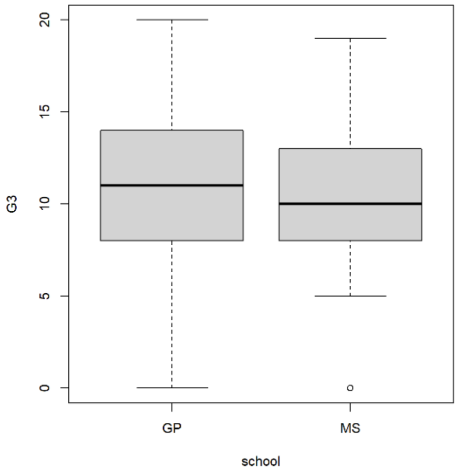
Step 3: Conducting the ANOVA test, while keeping tabs on the limitations

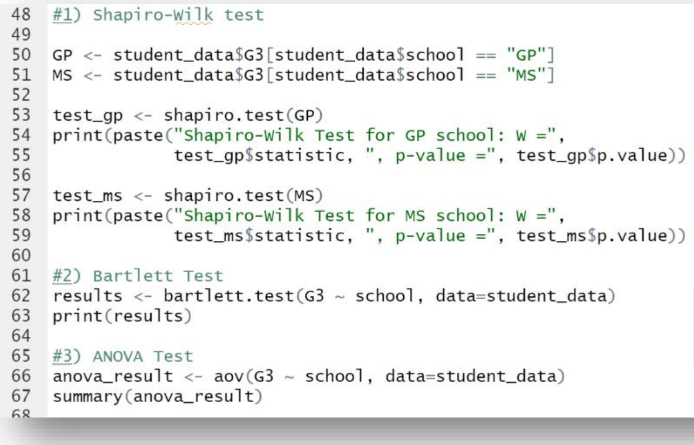
Results:

F-statistic: 0.7980

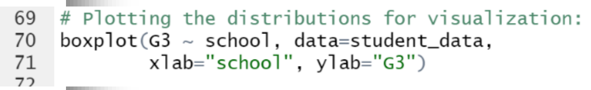
p-value: 0.3722

The p-value is greater than 0.05. This implies that we lack sufficient data to reject the null hypothesis. Consequently, there is no substantial difference in 'G3' grades between students from the GP and MS schools.





```
#1) Shapiro-Wilk test

GP <- student_data$G3[student_data$school == "GP"]
MS <- student_data$G3[student_data$school == "MS"]

test_gp <- shapiro.test(GP)
print(paste("Shapiro-Wilk Test for GP school: W =",
            test_gp$statistic, ", p-value =", test_gp$p.value))

test_ms <- shapiro.test(MS)
print(paste("Shapiro-Wilk Test for MS school: W =",
            test_ms$statistic, ", p-value =", test_ms$p.value))

#2) Bartlett Test
results <- bartlett.test(G3 ~ school, data=student_data)
print(results)

#3) ANOVA Test
anova_result <- aov(G3 ~ school, data=student_data)
summary(anova_result)
```



```
# Plotting the distributions for visualization:
boxplot(G3 ~ school, data=student_data,
        xlab="school", ylab="G3")
```

Second Hypothesis

Step 1: Shapiro-Wilk

Results:

- Male Students –

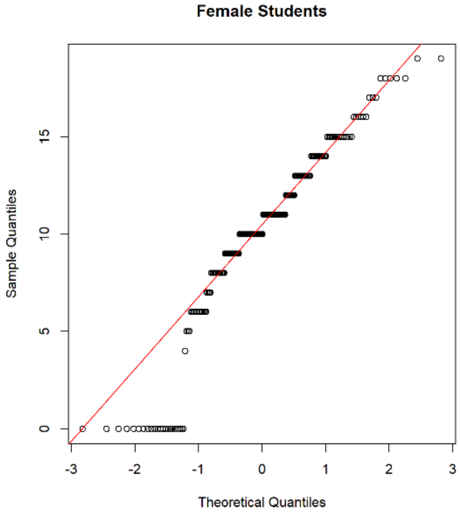
Test Statistic: 0.9308

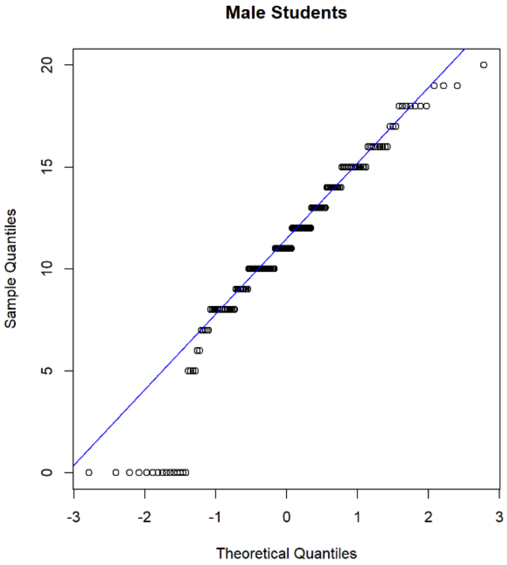
p-value: 9.0 x 10^{-8}

- Female Students –

Test Statistic: 0.9242

p-value: 7.2 x 10^{-9}





Given standard significance levels, both genders p-values are less than 0.05. That means, the 'G3' grades do not follow a normal distribution.

Step 2: Bartlett test

Results:

Test Statistic: 0.1516

p-value: 0.697

The p-value is greater than 0.05. This implies that we lack sufficient data to reject the null hypothesis that variances are identical across groups. Consequently, the assumption of variance homogeneity is met.

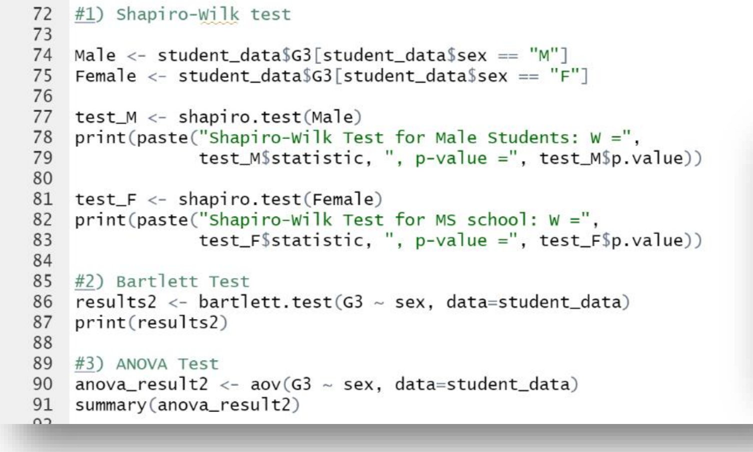
Step 3: ANOVA test

Results:

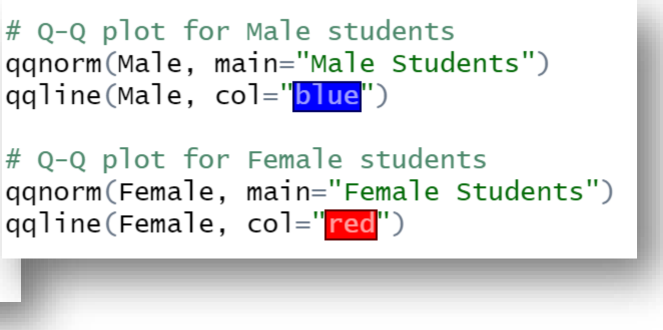
F-statistic: 4.252

p-value: 0.0399

The p-value is less than 0.05. As a result, there is statistically significant evidence that at least one group's mean differs from the others. In other words, gender does have an effect on the final grade.

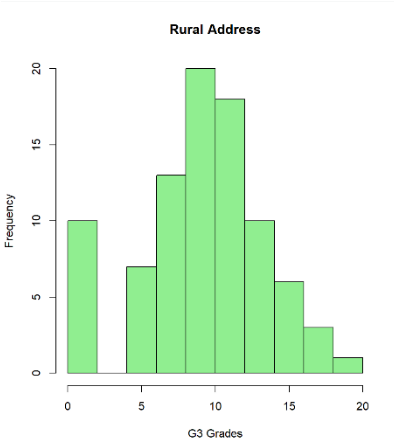


```
#1) Shapiro-Wilk test

Male <- student_data$G3[student_data$sex == "M"]
Female <- student_data$G3[student_data$sex == "F"]

test_M <- shapiro.test(Male)
print(paste("Shapiro-Wilk Test for Male Students: W =",
            test_M$statistic, ", p-value =", test_M$p.value))

test_F <- shapiro.test(Female)
print(paste("Shapiro-Wilk Test for MS school: W =",
            test_F$statistic, ", p-value =", test_F$p.value))

#2) Bartlett Test
results2 <- bartlett.test(G3 ~ sex, data=student_data)
print(results2)

#3) ANOVA Test
anova_result2 <- aov(G3 ~ sex, data=student_data)
summary(anova_result2)
```



```
# Q-Q plot for Male students
qqnorm(Male, main="Male Students")
qqline(Male, col="blue")

# Q-Q plot for Female students
qqnorm(Female, main="Female Students")
qqline(Female, col="red")
```

Third Hypothesis



Step 1: Shapiro-Wilk

Results:

- Urban Students –

Test Statistic: 0.9243

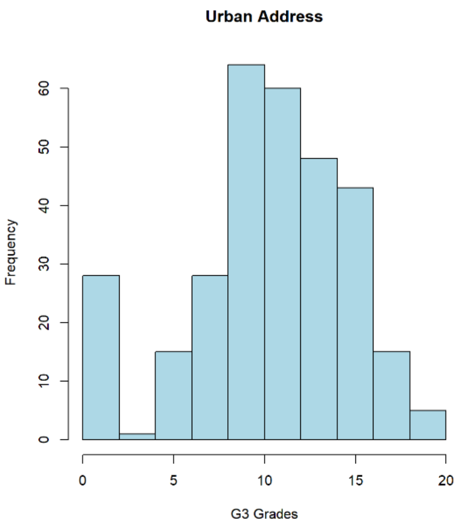
p-value: 2.3 x 10^{-11}

- Rural Students –

Test Statistic: 0.9328

p-value: 2.0 x 10^{-4}

Given standard significance levels, both schools' p-values are less than 0.05. That means, the 'G3' grades do not follow a normal distribution.



Step 2: Bartlett's test

Results:

Test Statistic: 0.000311

p-value: 0.9859

The p-value is greater than 0.05. This implies that we lack sufficient data to reject the null hypothesis that variances are identical across groups. Consequently, the assumption of variance homogeneity is met.

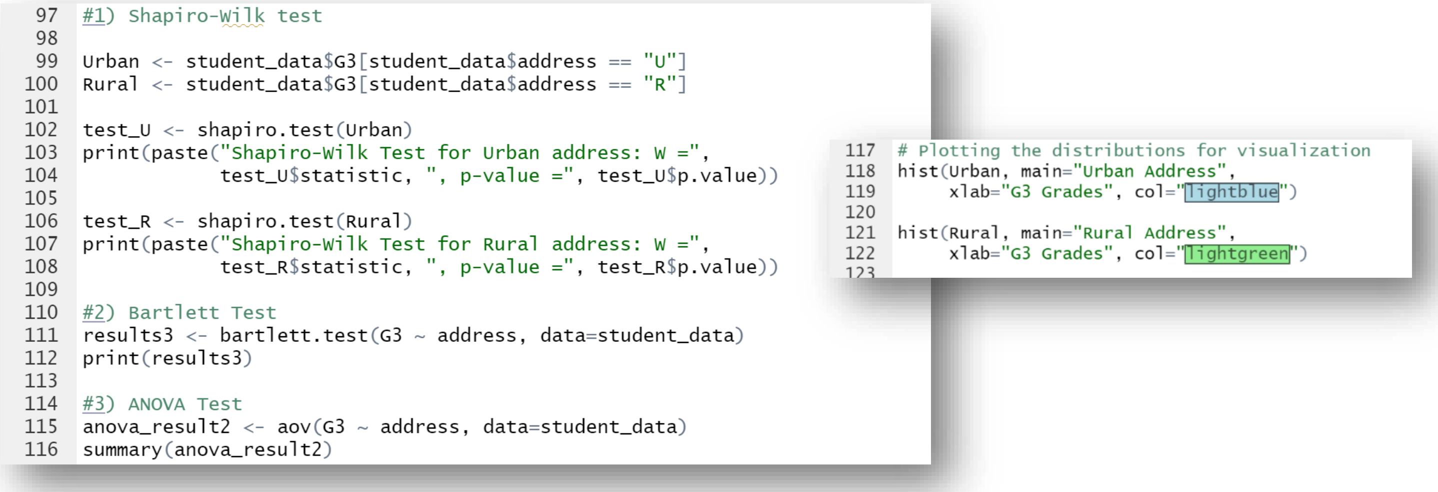
Step 3: ANOVA test

Results:

F-statistic: 4.445

p-value: 0.0356

The p-value is less than 0.05. As a result, there is statistically significant evidence that at least one group's mean differs from the others. In other words, the address in which a student lives in, does have an effect on the final grade.



```
#1) Shapiro-Wilk test

Urban <- student_data$G3[student_data$address == "U"]
Rural <- student_data$G3[student_data$address == "R"]

test_U <- shapiro.test(Urban)
print(paste("Shapiro-Wilk Test for Urban address: W =",
            test_U$statistic, ", p-value =", test_U$p.value))

test_R <- shapiro.test(Rural)
print(paste("Shapiro-Wilk Test for Rural address: W =",
            test_R$statistic, ", p-value =", test_R$p.value))

#2) Bartlett Test
results3 <- bartlett.test(G3 ~ address, data=student_data)
print(results3)

#3) ANOVA Test
anova_result2 <- aov(G3 ~ address, data=student_data)
summary(anova_result2)
```

```
# Plotting the distributions for visualization
hist(Urban, main="Urban Address",
     xlab="G3 Grades", col="lightblue")

hist(Rural, main="Rural Address",
     xlab="G3 Grades", col="lightgreen")
```

Fourth Hypothesis

Note: Unique Values in the "freetime" column are [1, 2, 3, 4, 5]

Step 1: Shapiro-Wilk

Results:

	Test Statistic	p-value
1	0.9261	1.5 x 10^-1
2	0.9183	4.2 x 10^ -4
3	0.9088	2.4 x 10^-8
4	0.9353	3.0 x 10^-5
5	0.9618	1.9 x 10^-1

Tests 2, 3, and 4 show that their individual datasets are not normally distributed. Tests 1 and 5, on the other hand, do not provide enough evidence to indicate a deviation from normalcy for respective datasets. This suggests that the data from Tests 1 and 5 are roughly normally distributed.

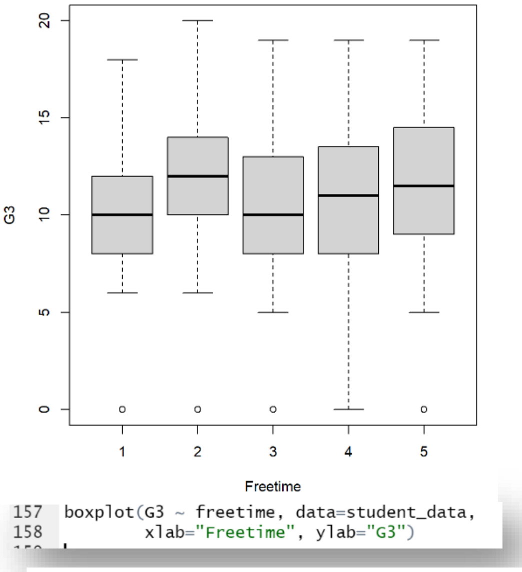
Step 2: Bartlett's test

Results:

Test Statistic: 2.176

p-value: 0.7034

The p-value is greater than 0.05. This implies that we lack sufficient data to reject the null hypothesis that variances are identical across groups. Consequently, the assumption of variance homogeneity is met.



```
boxplot(G3 ~ freetime, data=student_data,
          xlab="Freetime", ylab="G3")
```

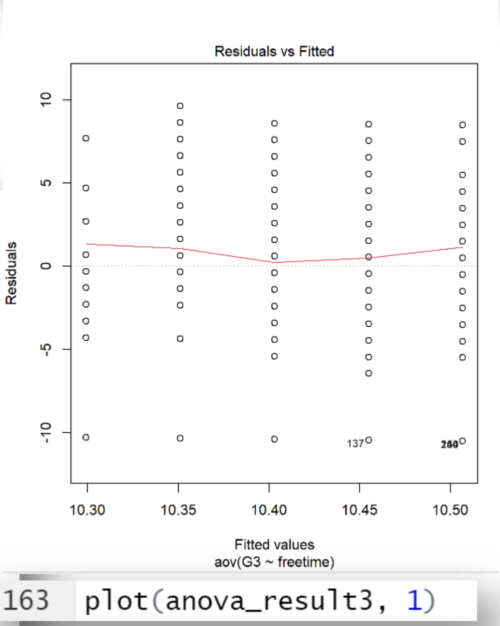
Step 3: ANOVA test

Results:

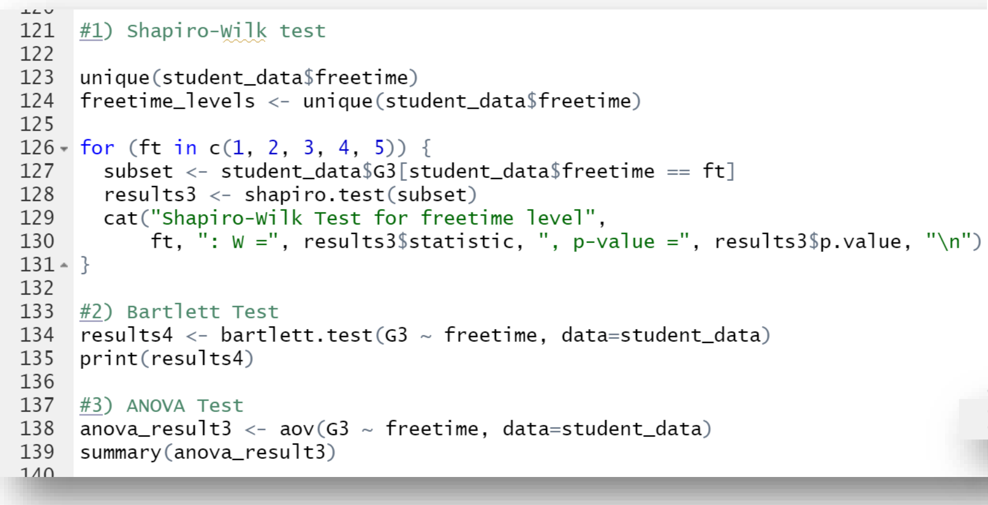
F-statistic: 0.05

p-value: 0.823

The p-value is greater than 0.05. This implies that we lack sufficient data to reject the null hypothesis. Consequently, there is no substantial difference in 'G3' grades between students from the GP and MS schools.



```
plot(anova_result3, 1)
```



```
#1) Shapiro-Wilk test

unique(student_data$freetime)
freetime_levels <- unique(student_data$freetime)

for (ft in c(1, 2, 3, 4, 5)) {
  subset <- student_data$G3[student_data$freetime == ft]
  results3 <- shapiro.test(subset)
  cat("Shapiro-Wilk Test for freetime level",
      ft, ": W =", results3$statistic, ", p-value =", results3$p.value, "\n")
}

#2) Bartlett Test
results4 <- bartlett.test(G3 ~ freetime, data=student_data)
print(results4)

#3) ANOVA Test
anova_result3 <- aov(G3 ~ freetime, data=student_data)
summary(anova_result3)
```

Implications for the Business Problem:

My findings show that a student's school (between GP and MS) and amount of free time have no significant impact on their final grade ('G3'). As a result, interventions or recommendations aimed at improving student performance should not be based just on those criteria but should also consider other factors. Gender and the student's living environment, on the other hand, have an impact on their grade. However, because gender genes are unchangeable, it is not an important component to consider while attempting to improve a student's performance. In conclusion, a student's living environment is the only significant factor that I was able to deduce.

Limitations:

1. According to one of ANOVA's assumptions, my study found that the 'G3' grades do not have a normal distribution. This could affect the accuracy and dependability of our ANOVA results. A non-parametric technique could be investigated in future analyses.

2. Factors from the dataset that we not tested may have an impact on student grades. And should be looked into thoroughly.

3. External factors that are not captured in the dataset may influence student grades.

Future Work:

1. Investigate other variables and their effects on student grades.

2. Expand the study to include a larger student sample.

3. Forecast utilizing advanced modeling techniques such as regression or machine learning.

LINKS

Kaggle Dataset:

https://www.kaggle.com/datasets/dipam7/student-grade-prediction

Github repository:

https://github.com/patrickchaccour/Assignements