Goldsmiths University of London

MSc. Data Science and Artificial Intelligence

Module: Data Visualization

Author: Carlos Manuel De Oliveira Alves

Student: cdeol003

Data Source

The dataset for this study on student performance in secondary education is sourced from the UCI Machine Learning Repository, specifically from the "Student Performance Data Set" available at https://archive.ics.uci.edu/dataset/320/student+performance. This repository is a well-known and respected source in the data science community, often used for academic and research purposes due to its diverse collection of high-quality datasets.

How the Data was Found:

The dataset was selected due to its relevance to the research questions concerning secondary education student performance. The UCI Machine Learning Repository is a popular destination for researchers seeking datasets for various machine learning tasks, making it a natural choice for finding educational data.

Why the Data was Selected:

- **Trustworthiness**: The UCI Machine Learning Repository has a reputation for hosting reliable and well-documented datasets. The student performance dataset, in particular, has been used in multiple academic studies, indicating its credibility.
- **Validity**: The dataset includes a range of variables relevant to student performance, such as demographic information, social factors, and school-related attributes. This comprehensive nature makes it well-suited for a multifaceted analysis of factors affecting student achievement.

Initial Data Collection:

The data was initially collected from two Portuguese schools for students in secondary education, covering performance in Mathematics and Portuguese language courses. It includes grades and demographic, social, and school-related

features. This collection likely supported academic research on educational outcomes and their influencing factors.

Critical Assessment of the Data:

- **Completeness**: The dataset appears comprehensive regarding the variety of factors it includes, which is beneficial for a thorough analysis. However, the scope is limited to two schools, which might restrict the generalizability of the findings.
- **Accuracy**: Assuming the data was collected directly from school records, grade and demographic information accuracy should be high. However, data based on student self-reports, like social habits, might need more reliable.
- **Consistency**: Consistency in data collection methods across both schools is critical for comparative analysis. Any discrepancies could introduce biases or inaccuracies in the findings.
- **Bias and Potential Issues**:
  - **Selection Bias**: Since the data is from only two schools, it may not represent the broader population of secondary school students in Portugal or other regions.
  - **Reporting Bias**: Self-reported data, particularly regarding sensitive topics like alcohol consumption, may be subject to reporting Bias.
  - **Cultural and Socio-Economic Context**: The data reflects the specific cultural and socio-economic environment of the schools involved, which might limit its applicability to different contexts.

Conclusion:

While the UCI Machine Learning Repository dataset is valuable for its detail and range, caution must be exercised in interpreting the results. The potential issues and biases identified necessitate a critical approach to analysis and careful consideration of the study's scope and applicability.

Introduction

Analysis of Student Performance in Secondary Education:

This report presents a comprehensive analysis of student achievement in secondary education, focusing on two Portuguese schools. Utilizing detailed datasets covering both Mathematics and Portuguese language classes, the study examines a range of demographic, social, and academic factors to understand their impact on student performance. Derived from school reports and questionnaires, the datasets include grades across three academic periods (G1, G2, G3) alongside various student background characteristics.

Structure of the Report

1. **Data Overview**: We begin with an exploration of the datasets, highlighting key attributes such as student demographics, parental background, and academic records.
2. **Descriptive Statistics**: An initial analysis provides insights into the general trends within the data, such as average grades and attendance.
3. **Correlation Analysis**: We delve into the relationships between different factors, mainly focusing on how early-period grades (G1 and G2) correlate with final grades (G3).
4. **Predictive Modeling**: The report presents models predicting final grades based on various variables, assessing their accuracy and utility.
5. **Segmented Analysis**: We conduct a detailed examination of how different demographic groups, such as gender and age, as well as parental education levels, affect student performance.

Key Findings

- **Strong Predictive Power of Early Grades**: A significant correlation is observed between G1, G2, and the final year grade G3, underscoring the importance of consistent academic performance.
- **Gender Disparities**: Female students consistently outperform male students in Mathematics and Portuguese.
- **Influence of Age and Parental Education**: Younger students generally achieve higher grades, and a higher level of parental education, particularly the mother's, is associated with better student performance.
- **Predictive Modeling Insights**: Models for both subjects show good predictive capabilities, with the model for Portuguese displaying an exceptionally high accuracy.

This investigation aims to provide valuable insights into the factors influencing student achievement, aiding educators and policymakers in enhancing educational strategies and student support mechanisms.

Background

Understanding Student Performance in Secondary Education:

The analysis of student performance in secondary education is a topic of ongoing interest and research, given its implications for educational policies and future academic and career success. This discussion draws upon existing research and recent news stories to contextualize our analysis of student performance in two Portuguese schools.

Related Research

1. **Academic Performance Predictors**: The study by Eyman Alyahyan et al. (2020) [1] has explored the predictors of academic performance, with a significant focus on data mining techniques for effective prediction. Common findings underscore the importance of early academic achievement, socioeconomic status, parental involvement, student demographics, e-learning activity, psychological attributes, and environmental factors. For instance, our observations align with the literature emphasizing the strong correlation between grades in early education (G1, G2, and G3) and future academic success. This multifaceted approach underlines the complex interplay of various factors in shaping academic outcomes.

2. **Gender Differences in Education**: Research has consistently shown gender disparities in education, often favouring female students regarding grades and academic engagement. A publication by the OECD (2015) [2] discussed how these differences manifest across various countries and subjects, resonating with our finding that female students generally outperform males in Mathematics and Portuguese.

3. **Parental Influence on Academic Achievement**: The influence of parental background, particularly education levels, has been extensively documented. A study by Jeynes (2007) in the "Urban Education" journal found [3] that parental education levels significantly impact student academic outcomes, aligning with our observations regarding the correlation between maternal education and student grades.

Recent News Stories

1. **Educational Trends in Portugal**: Recent news from sources like "The Portugal News" [4] highlights the evolving educational landscape in Portugal, focusing on government initiatives to address educational disparities and improve overall academic performance. These stories provide a real-world context for our analysis, demonstrating the practical implications of understanding factors influencing student achievement.

2. **Global Shifts in Education During the Pandemic**: The COVID-19 pandemic brought significant changes to education systems worldwide. Reports from various media outlets, including "Al Jazeera" [5] and others have discussed the challenges and transformations in education during this period, such as the shift to online learning and its impact on student performance. These developments underscore the importance of continuous analysis and adaptation in education strategies.

3. **Focus on STEM Education and Gender**: There is an increasing focus on STEM (Science et al.) education, particularly in encouraging female participation. The report by the American Association of University Women (AAUW) [6] reflect global efforts to bridge gender gaps in STEM, relevant to our findings of gender differences in Mathematics performance.

The background research and news stories provide a broader context for our analysis, highlighting the relevance and practical application of understanding various factors that influence student performance in secondary education.

Motivation

The motivation for this study on student performance in secondary education at two Portuguese schools lies in the critical need to understand the various factors influencing academic success. By delving into student grades, demographic, social, and school-related features, this investigation aims to uncover the underlying patterns and relationships that drive student achievement in Mathematics and Portuguese language courses.

Critical motivations for this project include:

1. **Enhancing Educational Strategies**: Insights from this study could inform teachers, school administrators, and policymakers about effective strategies to enhance student learning. Understanding student performance factors can lead to more tailored and impactful educational interventions.
2. **Identifying At-Risk Students**: Early identification of students likely to underperform could enable proactive support. By analyzing the correlations between grades in different academic periods and other attributes, educators can develop targeted programs to assist students who might struggle.
3. **Improving Resource Allocation**: Schools often operate under limited resources. Data-driven insights from this study can help efficiently allocate these resources, ensuring they are directed towards programs and initiatives that have the most significant impact on student performance.
4. **Understanding Socio-Economic Impacts**: The study will explore how demographic and social factors, such as family background and extracurricular activities, influence academic achievement. This understanding is crucial in addressing educational disparities and promoting equity in education.
5. **Informing Curriculum Development**: By understanding which aspects of the curriculum correlate with higher student performance, educators can refine and adapt curricular content to meet students' needs and interests better.
6. **Long-Term Academic and Career Success**: Uncovering the predictors of academic success in secondary education can have long-lasting implications on students' future academic and career paths. Insights from this study can help shape programs that improve immediate academic performance and contribute to long-term success.
7. **Practical Utility in Independent Prediction**: Predicting final year grades independently of prior period grades (G3 without G1 and G2) offers a significant

challenge and a more excellent practical utility. Achieving this would provide a more robust understanding of the factors influencing student performance, independent of their earlier academic record.

The potential impact of this study is substantial, offering a data-driven foundation for enhancing student outcomes and contributing to the broader field of educational research.

Research Questions

The research questions for this study on student performance in secondary education at two Portuguese schools are informed by previous research in educational psychology, sociology of education, and data science. These questions are designed to explore the complex interplay between various factors and student achievement, drawing from established theories such as Bronfenbrenner's Ecological Systems Theory [7], which emphasizes the influence of multiple environmental systems on development, and educational theories that focus on the impact of socioeconomic status, school environment, and individual characteristics on learning outcomes.

1. **Demographic Factors and Student Performance:**
    - How do demographic factors such as age, gender, and family background influence students' performance in Mathematics and Portuguese language courses?
    - This question investigates the role of inherent demographic factors and family background in shaping academic outcomes based on theories that highlight the influence of socioeconomic status and family environment on educational achievement.
    - Citation: Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417-453 [8]. This study provides a comprehensive analysis of the impact of socioeconomic status on academic achievement.
2. **Social Behaviors and Academic Performance:**
    - What is the relationship between students' social behaviors (like going out with friends, alcohol consumption) and their academic performance?
    - This explores the impact of lifestyle and social choices on academic success, referencing research that links extracurricular activities and social behaviours with academic outcomes.
    - Citation: Mahoney, J. L., & Cairns, R. B. (1997). Do extracurricular activities protect against early school dropout? Developmental Psychology, 33(2), 241-253 [9]. This paper explores how extracurricular activities impact

educational outcomes.

3. **School-Related Factors and Grades:**

   - How do school-related factors such as study time, past failures, and school support services affect student grades?
   - This question focuses on the direct educational environment's impact, considering the role of study habits, past academic experiences, and institutional support in student achievement.
   - Citation: Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge [10]. This book synthesizes research on the impact of various educational strategies on student achievement.

4. **Correlation Between Early and Final Grades:**

   - Is there a significant correlation between the grades obtained in the first two academic periods (G1 and G2) and the final year grade (G3)?
   - Grounded in the theoretical framework that prior performance indicates future success, this question examines the predictive power of earlier grades on outcomes.
   - Citation: Cortez and Silva (2008). Using Data Mining to Predict Secondary School Student Performance [11]. This foundational study for this project provides a direct basis for investigating the correlation between early and final grades.

5. **Predicting Final Year Performance:**

   - Can student performance in the final year (G3) be accurately predicted without considering grades from the first two periods (G1 and G2)?
   - This research question challenges the conventional reliance on past academic performance as a predictor of future success, seeking to identify other significant predictors of final-year performance.
   - Citation: Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. Intelligence, 35(1), 13-21 [12]. This study discusses the complexities of predicting academic performance.

6. **Subject-Specific Performance Patterns:**

   - What patterns emerge from the comparison of student performance in Mathematics and Portuguese language, and what might explain these differences?
   - This question aims to uncover subject-specific factors influencing academic achievement, contributing to curriculum development and teaching

strategies tailored to different disciplines.

- Citation: Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. Psychological Bulletin, 140(4), 1174-1204 [13]. This meta-analysis provides insights into subject-specific academic performance differences.

7. **Impact of Extra-Curricular Activities:**

   - How do extra-curricular activities and personal interests impact academic performance in secondary education?
   - Investigating the role of non-academic pursuits, this question aligns with theories that emphasize the holistic development of learners and the impact of a well-rounded educational experience on academic success.
   - Citation: Eccles, J. S., & Barber, B. L. (1999). Student council, volunteering, basketball, or marching band: What kind of extracurricular involvement matters? Journal of Adolescent Research, 14(1), 10-43 [14]. This research examines the influence of various extracurricular activities on academic success.

These research questions are designed to dissect the multifaceted nature of educational achievement, providing a comprehensive understanding of the factors contributing to student success in secondary education.

Data Pre-Processing

Load the Data

In [55]:
```
# Imports the 'drive' module from 'google.colab' and mounts the Google Dr
# the '/content/drive' directory in the Colab environment.
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, c
all drive.mount("/content/drive", force_remount=True).
```

- Use pandas.read_csv() to load the datasets

In [56]:
```
# Importing the pandas library
import pandas as pd

# Loading the datasets using pandas
math_data = pd.read_csv('/content/drive/MyDrive/student/student-mat.csv',
portuguese_data = pd.read_csv('/content/drive/MyDrive/student/student-por
```

Inspect the Data

- Use data.head() to view the first three rows of the datasets
- Use data.info() to get an overview of the data types and missing values

In [57]:
```
# Initial Examination
print("\n-> Math Dataset:")
print(math_data.head(3).T)
print("\n-------------------------------------")
print("\n-> Portuguese Dataset:")
print(portuguese_data.head(3).T)
print("\n-------------------------------------")
print(math_data.info())
print("\n-------------------------------------")
print(portuguese_data.info())
```

-> Math Dataset:

	0	1	2
school	GP	GP	GP
sex	F	F	F
age	18	17	15
address	U	U	U
famsize	GT3	GT3	LE3
Pstatus	A	T	T
Medu	4	1	1
Fedu	4	1	1
Mjob	at_home	at_home	at_home
Fjob	teacher	other	other
reason	course	course	other
guardian	mother	father	mother
traveltime	2	1	1
studytime	2	2	2
failures	0	0	3
schoolsup	yes	no	yes
famsup	no	yes	no
paid	no	no	yes
activities	no	no	no
nursery	yes	no	yes
higher	yes	yes	yes
internet	no	yes	yes
romantic	no	no	no
famrel	4	5	4
freetime	3	3	3
goout	4	3	2
Dalc	1	1	2
Walc	1	1	3
health	3	3	3
absences	6	4	10
G1	5	5	7
G2	6	5	8
G3	6	6	10

-> Portuguese Dataset:

	0	1	2
school	GP	GP	GP
sex	F	F	F
age	18	17	15
address	U	U	U
famsize	GT3	GT3	LE3
Pstatus	A	T	T
Medu	4	1	1
Fedu	4	1	1
Mjob	at_home	at_home	at_home
Fjob	teacher	other	other
reason	course	course	other
guardian	mother	father	mother
traveltime	2	1	1
studytime	2	2	2
failures	0	0	0
schoolsup	yes	no	yes
famsup	no	yes	no
paid	no	no	no
activities	no	no	no
nursery	yes	no	yes

```
higher          yes      yes      yes
internet         no      yes      yes
romantic         no       no       no
famrel            4        5        4
freetime          3        3        3
goout             4        3        2
Dalc              1        1        2
Walc              1        1        3
health            3        3        3
absences          4        2        6
G1                0        9       12
G2               11       11       13
G3               11       11       12

----------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      395 non-null    object
 1   sex         395 non-null    object
 2   age         395 non-null    int64
 3   address     395 non-null    object
 4   famsize     395 non-null    object
 5   Pstatus     395 non-null    object
 6   Medu        395 non-null    int64
 7   Fedu        395 non-null    int64
 8   Mjob        395 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    395 non-null    object
 12  traveltime  395 non-null    int64
 13  studytime   395 non-null    int64
 14  failures    395 non-null    int64
 15  schoolsup   395 non-null    object
 16  famsup      395 non-null    object
 17  paid        395 non-null    object
 18  activities  395 non-null    object
 19  nursery     395 non-null    object
 20  higher      395 non-null    object
 21  internet    395 non-null    object
 22  romantic    395 non-null    object
 23  famrel      395 non-null    int64
 24  freetime    395 non-null    int64
 25  goout       395 non-null    int64
 26  Dalc        395 non-null    int64
 27  Walc        395 non-null    int64
 28  health      395 non-null    int64
 29  absences    395 non-null    int64
 30  G1          395 non-null    int64
 31  G2          395 non-null    int64
 32  G3          395 non-null    int64
dtypes: int64(16), object(17)
memory usage: 102.0+ KB
None

----------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 649 entries, 0 to 648
```

```
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      649 non-null    object
 1   sex         649 non-null    object
 2   age         649 non-null    int64
 3   address     649 non-null    object
 4   famsize     649 non-null    object
 5   Pstatus     649 non-null    object
 6   Medu        649 non-null    int64
 7   Fedu        649 non-null    int64
 8   Mjob        649 non-null    object
 9   Fjob        649 non-null    object
 10  reason      649 non-null    object
 11  guardian    649 non-null    object
 12  traveltime  649 non-null    int64
 13  studytime   649 non-null    int64
 14  failures    649 non-null    int64
 15  schoolsup   649 non-null    object
 16  famsup      649 non-null    object
 17  paid        649 non-null    object
 18  activities  649 non-null    object
 19  nursery     649 non-null    object
 20  higher      649 non-null    object
 21  internet    649 non-null    object
 22  romantic    649 non-null    object
 23  famrel      649 non-null    int64
 24  freetime    649 non-null    int64
 25  goout       649 non-null    int64
 26  Dalc        649 non-null    int64
 27  Walc        649 non-null    int64
 28  health      649 non-null    int64
 29  absences    649 non-null    int64
 30  G1          649 non-null    int64
 31  G2          649 non-null    int64
 32  G3          649 non-null    int64
dtypes: int64(16), object(17)
memory usage: 167.4+ KB
None
```

Datasets Analyses

> ### Analyzing the Math dataset, we observe the following characteristics and potential areas of exploration:

1. **General Overview**: The dataset contains 395 entries, each representing a student. It has 33 columns, including categorical (object type) and numerical (int64 type) data.

2. **Categorical Data**: These are the 'object' type columns in the dataset, which include information like school, sex, address, family size, parent's job, reason for choosing the school, and more. These categorical variables can be used to analyze patterns and relationships in students' backgrounds and choices.

3. **Numerical Data**: These are the 'int64' type columns that provide quantitative

information about the students. This includes age, parents' education level (Medu, Fedu), travel time to school, study time, number of past class failures, family relationships, free time, going out frequency, alcohol consumption (weekday and weekend), health status, number of school absences, and grades (G1, G2, G3).

4. **Potential Analyses**:
    - **Demographic Analysis**: Explore distributions of age, sex, address (urban/rural), and family size.
    - **Educational Factors**: Examine the relationship between students' grades (G1, G2, G3) and variables like parents' education level, study time, school support, and number of failures.
    - **Social Factors**: Investigate how family relationships, free time, going out, and romantic relationships impact academic performance.
    - **Health and Lifestyle Factors**: Look into the impact of health status and alcohol consumption on students' academic achievements and attendance (absences).
5. **Memory Usage**: The dataset uses 102.0 KB of memory, indicating a moderate size that should be manageable for most data analysis tools.

6. **Data Quality**: All columns have 395 non-null entries, suggesting no missing values in this dataset, which is beneficial for analysis.

> ## Analyzing the Portuguese dataset, we observe the following characteristics and potential areas of exploration:

1. **General Overview**: The dataset comprises 649 entries, each representing a student. It includes 33 columns, encompassing both categorical (object type) and numerical (int64 type) data.

2. **Categorical Data**: These columns are of the 'object' type and provide information on various aspects such as the school, sex, address, family size, parents' job, reasons for choosing the school, and more. These variables can be analyzed to discern trends and correlations in the backgrounds and choices of the students.

3. **Numerical Data**: These 'int64' type columns include quantitative data about the students, such as age, parents' education level (Medu, Fedu), travel time, study time, past class failures, family relationships, free time, going out frequency, alcohol consumption (Dalc for weekdays and Walc for weekends), health status, number of absences, and grades (G1, G2, G3).

4. **Potential Analyses**:
    - **Demographic Analysis**: Investigate the distribution of variables like age, sex, address (urban vs. rural), and family size.

- **Educational Factors**: Analyze how students' grades (G1, G2, G3) are influenced by factors such as parents' education, study time, school support, and failure rates.
- **Social and Lifestyle Factors**: Examine the impact of family relationships, free time, social activities, romantic involvement, and lifestyle choices (like alcohol consumption) on students' academic performance.
- **Health and Attendance**: Look into how health status and school absences correlate with academic achievements.

5. **Memory Usage**: The dataset consumes 167.4 KB of memory, which is slightly larger than the previous dataset but still manageable for typical data analysis tools.

6. **Data Quality**: No null entries in any of the columns indicate a complete dataset without missing values, facilitating a more robust analysis.

To gain insights from this datasets, statistical analyses, including correlation studies, regression models, and various forms of data visualization, would be beneficial. These analyses can reveal underlying patterns and relationships between the students' academic performances and socio-economic, demographic, and lifestyle factors.

Data Cleaning

- Look for any inconsistencies or inaccuracies, such as unusual values or typos.
- Handle them by either correcting or removing the erroneous data.

Statistical Summary

Generate a statistical summary for the numerical columns to understand their distribution, such as mean, standard deviation, and potential outliers. We will also examine the unique values in the categorical columns to check for any inconsistencies or typos.

In [58]:
```
# Statistical summary of the numerical columns
print(math_data.describe())
print(portuguese_data.describe())

# Checking unique values in categorical columns
cat_columns_mat = math_data.select_dtypes(include=['object']).columns
unique_values_mat = {col: math_data[col].unique() for col in cat_columns_

cat_columns_por = portuguese_data.select_dtypes(include=['object']).colum
unique_values_por = {col: portuguese_data[col].unique() for col in cat_co

# Print unique values (optional)
print(unique_values_mat)
print(unique_values_por)
```

```
              age        Medu        Fedu  traveltime   studytime    failu
res  \
count  395.000000  395.000000  395.000000  395.000000  395.000000  395.000
000
mean    16.696203    2.749367    2.521519    1.448101    2.035443    0.334
177
std      1.276043    1.094735    1.088201    0.697505    0.839240    0.743
651
min     15.000000    0.000000    0.000000    1.000000    1.000000    0.000
000
25%     16.000000    2.000000    2.000000    1.000000    1.000000    0.000
000
50%     17.000000    3.000000    2.000000    1.000000    2.000000    0.000
000
75%     18.000000    4.000000    3.000000    2.000000    2.000000    0.000
000
max     22.000000    4.000000    4.000000    4.000000    4.000000    3.000
000

           famrel    freetime       goout        Dalc        Walc      hea
lth  \
count  395.000000  395.000000  395.000000  395.000000  395.000000  395.000
000
mean     3.944304    3.235443    3.108861    1.481013    2.291139    3.554
430
std      0.896659    0.998862    1.113278    0.890741    1.287897    1.390
303
min      1.000000    1.000000    1.000000    1.000000    1.000000    1.000
000
25%      4.000000    3.000000    2.000000    1.000000    1.000000    3.000
000
50%      4.000000    3.000000    3.000000    1.000000    2.000000    4.000
000
75%      5.000000    4.000000    4.000000    2.000000    3.000000    5.000
000
max      5.000000    5.000000    5.000000    5.000000    5.000000    5.000
000

         absences          G1          G2          G3
count  395.000000  395.000000  395.000000  395.000000
mean     5.708861   10.908861   10.713924   10.415190
std      8.003096    3.319195    3.761505    4.581443
min      0.000000    3.000000    0.000000    0.000000
25%      0.000000    8.000000    9.000000    8.000000
50%      4.000000   11.000000   11.000000   11.000000
75%      8.000000   13.000000   13.000000   14.000000
max     75.000000   19.000000   19.000000   20.000000
              age        Medu        Fedu  traveltime   studytime    failu
res  \
count  649.000000  649.000000  649.000000  649.000000  649.000000  649.000
000
mean    16.744222    2.514638    2.306626    1.568567    1.930663    0.221
880
std      1.218138    1.134552    1.099931    0.748660    0.829510    0.593
235
min     15.000000    0.000000    0.000000    1.000000    1.000000    0.000
000
25%     16.000000    2.000000    1.000000    1.000000    1.000000    0.000
000
50%     17.000000    2.000000    2.000000    1.000000    2.000000    0.000
```

```
000
75%     18.000000    4.000000    3.000000    2.000000    2.000000    0.000
000
max     22.000000    4.000000    4.000000    4.000000    4.000000    3.000
000

           famrel    freetime       goout        Dalc        Walc      hea
lth  \
count  649.000000  649.000000  649.000000  649.000000  649.000000  649.000
000
mean     3.930663    3.180277    3.184900    1.502311    2.280431    3.536
210
std      0.955717    1.051093    1.175766    0.924834    1.284380    1.446
259
min      1.000000    1.000000    1.000000    1.000000    1.000000    1.000
000
25%      4.000000    3.000000    2.000000    1.000000    1.000000    2.000
000
50%      4.000000    3.000000    3.000000    1.000000    2.000000    4.000
000
75%      5.000000    4.000000    4.000000    2.000000    3.000000    5.000
000
max      5.000000    5.000000    5.000000    5.000000    5.000000    5.000
000

         absences          G1          G2          G3
count  649.000000  649.000000  649.000000  649.000000
mean     3.659476   11.399076   11.570108   11.906009
std      4.640759    2.745265    2.913639    3.230656
min      0.000000    0.000000    0.000000    0.000000
25%      0.000000   10.000000   10.000000   10.000000
50%      2.000000   11.000000   11.000000   12.000000
75%      6.000000   13.000000   13.000000   14.000000
max     32.000000   19.000000   19.000000   19.000000
{'school': array(['GP', 'MS'], dtype=object), 'sex': array(['F', 'M'], dty
pe=object), 'address': array(['U', 'R'], dtype=object), 'famsize': array
(['GT3', 'LE3'], dtype=object), 'Pstatus': array(['A', 'T'], dtype=objec
t), 'Mjob': array(['at_home', 'health', 'other', 'services', 'teacher'], d
type=object), 'Fjob': array(['teacher', 'other', 'services', 'health', 'at
_home'], dtype=object), 'reason': array(['course', 'other', 'home', 'reput
ation'], dtype=object), 'guardian': array(['mother', 'father', 'other'], d
type=object), 'schoolsup': array(['yes', 'no'], dtype=object), 'famsup': a
rray(['no', 'yes'], dtype=object), 'paid': array(['no', 'yes'], dtype=obje
ct), 'activities': array(['no', 'yes'], dtype=object), 'nursery': array(['
yes', 'no'], dtype=object), 'higher': array(['yes', 'no'], dtype=object),
'internet': array(['no', 'yes'], dtype=object), 'romantic': array(['no', '
yes'], dtype=object)}
{'school': array(['GP', 'MS'], dtype=object), 'sex': array(['F', 'M'], dty
pe=object), 'address': array(['U', 'R'], dtype=object), 'famsize': array
(['GT3', 'LE3'], dtype=object), 'Pstatus': array(['A', 'T'], dtype=objec
t), 'Mjob': array(['at_home', 'health', 'other', 'services', 'teacher'], d
type=object), 'Fjob': array(['teacher', 'other', 'services', 'health', 'at
_home'], dtype=object), 'reason': array(['course', 'other', 'home', 'reput
ation'], dtype=object), 'guardian': array(['mother', 'father', 'other'], d
type=object), 'schoolsup': array(['yes', 'no'], dtype=object), 'famsup': a
rray(['no', 'yes'], dtype=object), 'paid': array(['no', 'yes'], dtype=obje
ct), 'activities': array(['no', 'yes'], dtype=object), 'nursery': array(['
yes', 'no'], dtype=object), 'higher': array(['yes', 'no'], dtype=object),
'internet': array(['no', 'yes'], dtype=object), 'romantic': array(['no', '
yes'], dtype=object)}
```

General Observations Across Both Datasets

1. **Age Range and Average**: Students in both datasets are in their mid to late teens (15-22 years), with a slightly higher average age in the second dataset (16.74 years) compared to the first (16.69 years).
2. **Parental Education (Medu and Fedu)**: The mean values suggest that parents in the first dataset are generally more educated than those in the second dataset. The maximum value for mothers' and fathers' education is 4 in both datasets.
3. **Study Time and Failures**: The first dataset shows a higher average of study time and failures. This could indicate either a more challenging curriculum or different evaluation standards.
4. **Personal and Health Attributes (famrel, freetime, goout, Dalc, Walc, health)**: The averages of family relationship quality, free time, going out, weekday and weekend alcohol consumption, and health are similar in both datasets. The maximum value for all these attributes is 5, indicating a possible Likert scale measurement.

Specific Observations for Each Dataset

1. **First Dataset (Count: 395)**
    - **Travel Time**: The average travel time is lower than the second dataset.
    - **Absences and Grades (G1, G2, G3)**: There is a higher average number of absences and slightly lower average grades.
    - **Standard Deviations**: Higher standard deviations in absences and grades suggest more variability in these areas.
2. **Second Dataset (Count: 649)**
    - **Travel Time**: Students have a higher average travel time.
    - **Absences and Grades (G1, G2, G3)**: Students have fewer absences on average and slightly higher grades.
    - **Standard Deviations**: Lower standard deviations in grades, indicating more consistency.

Categorical Variables

Both datasets include categorical variables such as school type, gender, address, family size, parent's job, reason for choosing the school, guardian, and various supports and activities. These variables can provide additional context for understanding differences in the numerical data.

Overall Insights

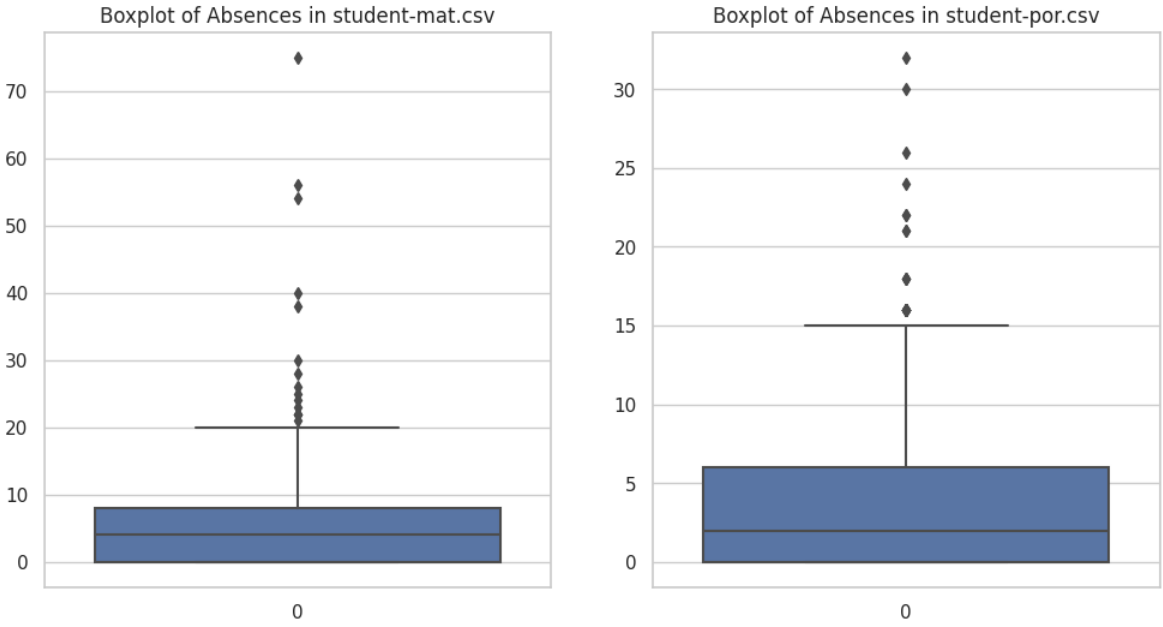
- The parental education levels, travel times, study habits, and academic performance differences suggest that the two datasets likely represent different populations or school environments.
- The consistency in personal attributes and health across both datasets indicates

similar student lifestyles or attitudes, regardless of the school type or other differentiating factors.
- The variability in grades and absences, particularly in the first dataset, suggests a need to explore factors influencing academic performance and attendance.

Handle Outliers

Address the potential outliers in the absences column. We will use boxplots to visualize these outliers and then cap them at a specific threshold calculated using the Interquartile Range (IQR) method.

In [59]:
```
# Import MATLAB-like interface for making plots and charts
import matplotlib.pyplot as plt

# Import Seaborn for data visualization with more attractive interface fo
import seaborn as sns

# Plotting boxplots for the 'absences' column
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
sns.boxplot(math_data['absences'])
plt.title('Boxplot of Absences in student-mat.csv')

plt.subplot(1, 2, 2)
sns.boxplot(portuguese_data['absences'])
plt.title('Boxplot of Absences in student-por.csv')
plt.show()

# Calculate IQR for 'absences'
Q1_mat = math_data['absences'].quantile(0.25)
Q3_mat = math_data['absences'].quantile(0.75)
IQR_mat = Q3_mat - Q1_mat
threshold_mat = Q3_mat + 1.5 * IQR_mat

Q1_por = portuguese_data['absences'].quantile(0.25)
Q3_por = portuguese_data['absences'].quantile(0.75)
IQR_por = Q3_por - Q1_por
threshold_por = Q3_por + 1.5 * IQR_por

# Capping outliers
math_data['absences'] = math_data['absences'].clip(upper=threshold_mat)
portuguese_data['absences'] = portuguese_data['absences'].clip(upper=thre

# Print thresholds
print("\n-> Threshold for absences in student-mat.csv:", threshold_mat)
print("\n-> Threshold for absences in student-por.csv:", threshold_por)
```




```
-> Threshold for absences in student-mat.csv: 20.0

-> Threshold for absences in student-por.csv: 15.0
```

The outliers in the absences column have been addressed using boxplots and the Interquartile Range (IQR) method. The calculated thresholds for capping outliers are:

```
student-mat.csv: 20.0
student-por.csv: 15.0
```

Outliers beyond these values have been capped.

Save the Cleaned Data

Finally, it is good practice to save the cleaned and processed data. This allows easy access to the cleaned datasets for further analysis or modelling. We will save the cleaned datasets to new CSV files.

In [60]:
```python
# Save the cleaned data to new CSV files
cleaned_mat_path = '/content/drive/MyDrive/student/cleaned_student-mat.cs
cleaned_por_path = '/content/drive/MyDrive/student/cleaned_student-por.cs

math_data.to_csv(cleaned_mat_path, index=False)
portuguese_data.to_csv(cleaned_por_path, index=False)

# Output the paths to the cleaned files
print("\n-> Cleaned 'student-mat.csv' saved at:", cleaned_mat_path)
print("\n-> Cleaned 'student-por.csv' saved at:", cleaned_por_path)
```

```
-> Cleaned 'student-mat.csv' saved at: /content/drive/MyDrive/student/clea
ned_student-mat.csv

-> Cleaned 'student-por.csv' saved at: /content/drive/MyDrive/student/clea
ned_student-por.csv
```

Data Parsing

- Ensure categorical data is correctly labeled and consistent across the datasets
- Convert any date columns to datetime format using pd.to_datetime()

Load the Data Cleaned

In [61]:
```
file_path_mat = '/content/drive/MyDrive/student/cleaned_student-mat.csv'
file_path_por = '/content/drive/MyDrive/student/cleaned_student-por.csv'

data_mat = pd.read_csv(file_path_mat)
data_por = pd.read_csv(file_path_por)
```

Identify Categorical Data

Identify which columns in the datasets are categorical. Categorical data represents category values and is typically stored as strings (hence the 'object' data type in Pandas). Identifying these columns is crucial for ensuring consistent data across datasets.

In [62]:
```
categorical_columns = [col for col in data_mat.columns if data_mat[col].d
categorical_columns
```

Out[62]:
```
['school',
 'sex',
 'address',
 'famsize',
 'Pstatus',
 'Mjob',
 'Fjob',
 'reason',
 'guardian',
 'schoolsup',
 'famsup',
 'paid',
 'activities',
 'nursery',
 'higher',
 'internet',
 'romantic']
```

We identified the categorical columns in the datasets. These columns are: 'school', 'sex', 'address', 'famsize', 'Pstatus', 'Mjob', 'Fjob', 'reason', 'guardian', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery', 'higher', 'internet', and 'romantic'.

Standardize Categorical Data

In this step, we ensure that the categorical data is consistent across both datasets. This involves checking if the same categories are used in both datasets for each categorical column and ensuring there are no spelling, capitalization, or category presence discrepancies.

In [63]:
```
# Declare function checks for differences in categorical values between t
def check_categorical_consistency(column, df1, df2):
```

```
    unique_values_df1 = set(df1[column].unique())
    unique_values_df2 = set(df2[column].unique())
    return unique_values_df1.symmetric_difference(unique_values_df2)

# Iterate over categorical columns to find and store differences in their
inconsistencies = {}
for column in categorical_columns:
    diff = check_categorical_consistency(column, data_mat, data_por)
    if diff:
        inconsistencies[column] = diff

# Displaying the results
inconsistencies, categorical_columns
```

Out[63]:
```
({},
 ['school',
  'sex',
  'address',
  'famsize',
  'Pstatus',
  'Mjob',
  'Fjob',
  'reason',
  'guardian',
  'schoolsup',
  'famsup',
  'paid',
  'activities',
  'nursery',
  'higher',
  'internet',
  'romantic'])
```

- Examining categorical columns across both datasets reveals no inconsistencies in category labels. This means the categorical data in 'student-mat.csv' and 'student-por.csv' is consistent regarding capitalization, spelling, and labelling. Since there are no date columns in these datasets, converting date columns to datetime format using pd.to_datetime() is not applicable.

- We checked for consistency in the categorical data across both datasets. No inconsistencies were found, indicating that the categorical data is consistent in labelling and formatting in both datasets. This concludes the data parsing and standardization process. The datasets are now prepared for further analysis or processing.

Cross-Tabulation and Summary Statistics

We will perform cross-tabulation and summary statistics to gain insights into the datasets. Specifically, we can look at the distribution of students by gender across different schools and the relationship between study time and final grades ('G3').

In [64]:
```
# Cross-tabulation: Gender distribution across schools
gender_school_por = pd.crosstab(data_por['school'], data_por['sex'])
```

```
gender_school_mat = pd.crosstab(data_mat['school'], data_mat['sex'])

# Summary statistics: Relationship between study time and final grades
studytime_grades_por = data_por.groupby('studytime')['G3'].mean()
studytime_grades_mat = data_mat.groupby('studytime')['G3'].mean()

gender_school_por, gender_school_mat, studytime_grades_por, studytime_gra
```

Out[64]:
```
(sex       F    M
 school
 GP      237  186
 MS      146   80,
 sex       F    M
 school
 GP      183  166
 MS       25   21,
 studytime
 1    10.844340
 2    12.091803
 3    13.226804
 4    13.057143
 Name: G3, dtype: float64,
 studytime
 1    10.047619
 2    10.171717
 3    11.400000
 4    11.259259
 Name: G3, dtype: float64)
```

Analysis Results

1. Gender Distribution Across Schools:

- For the Portuguese course data (cleaned_student-por.csv):
  - In-school GP: 237 females and 186 males.
  - In school MS: 146 females and 80 males.
- For the Mathematics course data (cleaned_student-mat.csv):
  - In-school GP: 183 females and 166 males.
  - In school MS: 25 females and 21 males.

2. Relationship Between Study Time and Final Grades (G3):

- For the Portuguese course data:
  - Students with 1 hour of study time per week have an average final grade of 10.84.
  - With 2 hours, the average final grade is 12.09.
  - With 3 hours, the average final grade is 13.23.
  - With 4 hours, the average final grade is 13.06.
- For the Mathematics course data:
  - Students with 1 hour of weekly study time have an average final grade 10.05.
  - With 2 hours, the average final grade is 10.17.

- With 3 hours, the average final grade is 11.40.
- With 4 hours, the average final grade is 11.26.

Advanced Data Processing

Perform more advanced data processing. We can group the data by multiple categories (like gender and school) and calculate the mean grades. Additionally, we can filter out students with high absences and see how it affects the average grade.

In [65]:
```
# Grouping by gender and school, calculating mean grades
mean_grades_grouped_por = data_por.groupby(['sex', 'school'])['G3'].mean(
mean_grades_grouped_mat = data_mat.groupby(['sex', 'school'])['G3'].mean(

# Filtering students with absences less than a certain threshold (e.g., 1
filtered_grades_por = data_por[data_por['absences'] < 10]['G3'].mean()
filtered_grades_mat = data_mat[data_mat['absences'] < 10]['G3'].mean()

mean_grades_grouped_por, mean_grades_grouped_mat, filtered_grades_por, fi
```

Out[65]:
```
(sex  school
 F    GP        13.004219
      MS        11.034247
 M    GP        12.032258
      MS         9.950000
 Name: G3, dtype: float64,
 sex  school
 F    GP         9.972678
      MS         9.920000
 M    GP        11.060241
      MS         9.761905
 Name: G3, dtype: float64,
 12.0,
 10.426282051282051)
```

Analysis Results

1. Mean Grades by Gender and School:

- For the Portuguese course data (cleaned_student-por.csv):
  - Female students in school GP: Average final grade is 13.00.
  - Female students in school MS: Average final grade is 11.03.
  - Male students in school GP: Average final grade is 12.03.
  - Male students in school MS: Average final grade is 9.95.
- For the Mathematics course data (cleaned_student-mat.csv):
  - Female students in school GP: Average final grade is 9.97.
  - Female students in school MS: Average final grade is 9.92.
  - Male students in school GP: Average final grade is 11.06.
  - Male students in school MS: Average final grade is 9.76.

2. Mean Grades with Filtered Absences:

- In the Portuguese course, students with at most ten absences have an average grade of 12.00.
- In the Mathematics course, students with at most ten absences have an average grade of 10.43.

Summary:

The analysis reveals interesting patterns in student performance across different demographics and behaviours. For instance, it shows that, in general, students who spend more time studying tend to have higher grades. Additionally, attendance significantly impacts academic performance, with students having fewer absences scoring higher on average. This kind of analysis can be beneficial for understanding the dynamics of student performance in different subjects and the influence of factors like study habits, gender, and school environment.

Data Exploration Analysis

Univariate Visualizations

- For nominal data, we will use bar charts.
- For ordinal data, we will also use bar charts, but with the categories ordered.
- For numerical data, we will use histograms.

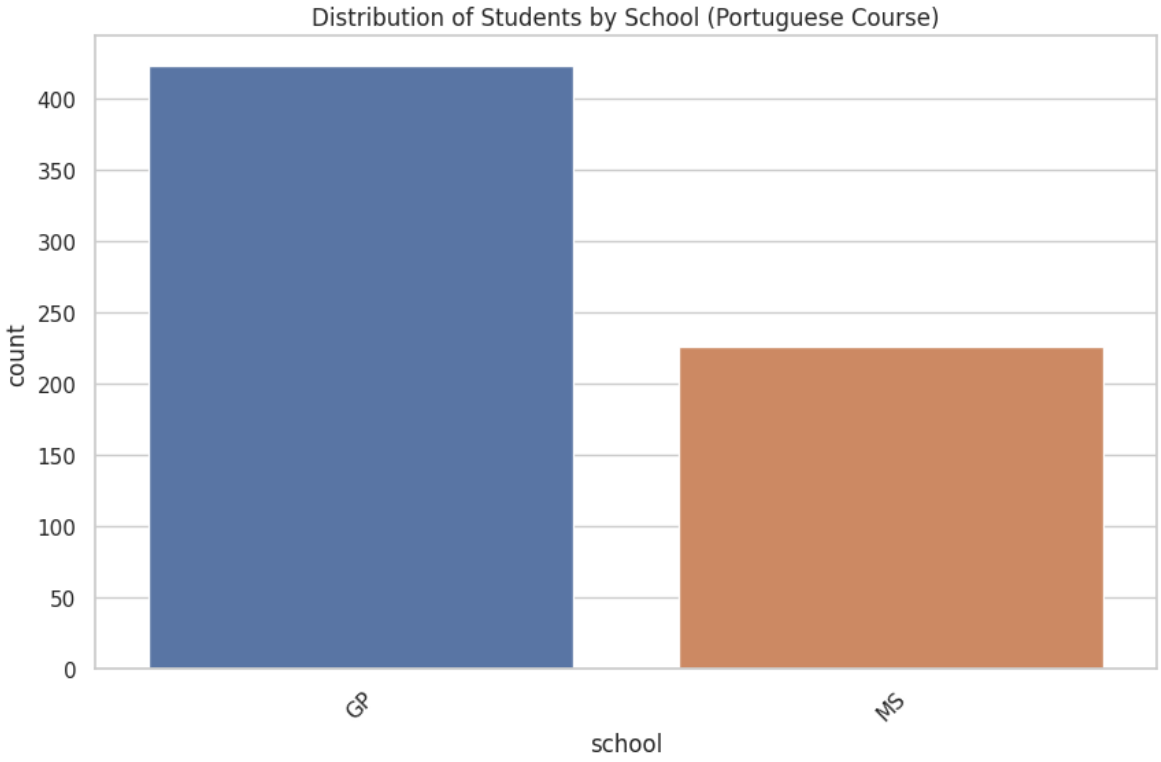
The datasets cleaned_student-por.csv (Portuguese course data) and cleaned_student-mat.csv (Mathematics course data) have similar structures. Each contains various types of variables:

- Nominal Variables: These are categorical variables without any intrinsic ordering. Examples include school, sex, address, famsize, Pstatus, Mjob, and Fjob.
- Ordinal Variables: These are categorical variables with a specific order. Examples might include Medu, Fedu (education levels of mother and father), and famrel (quality of family relationships).
- Numerical Variables: These are variables with numerical values. Examples include age, absences, G1, G2, and G3 (grades).

The bar charts below show the distribution of students by school and gender in the Portuguese course and Mathematics datasets. Next, let us create bar charts for ordinal variables. Here, the categories will be ordered based on their natural order. For instance, the education level (Medu and Fedu) can be ordered from lowest to highest. Let us plot these ordinal variables.

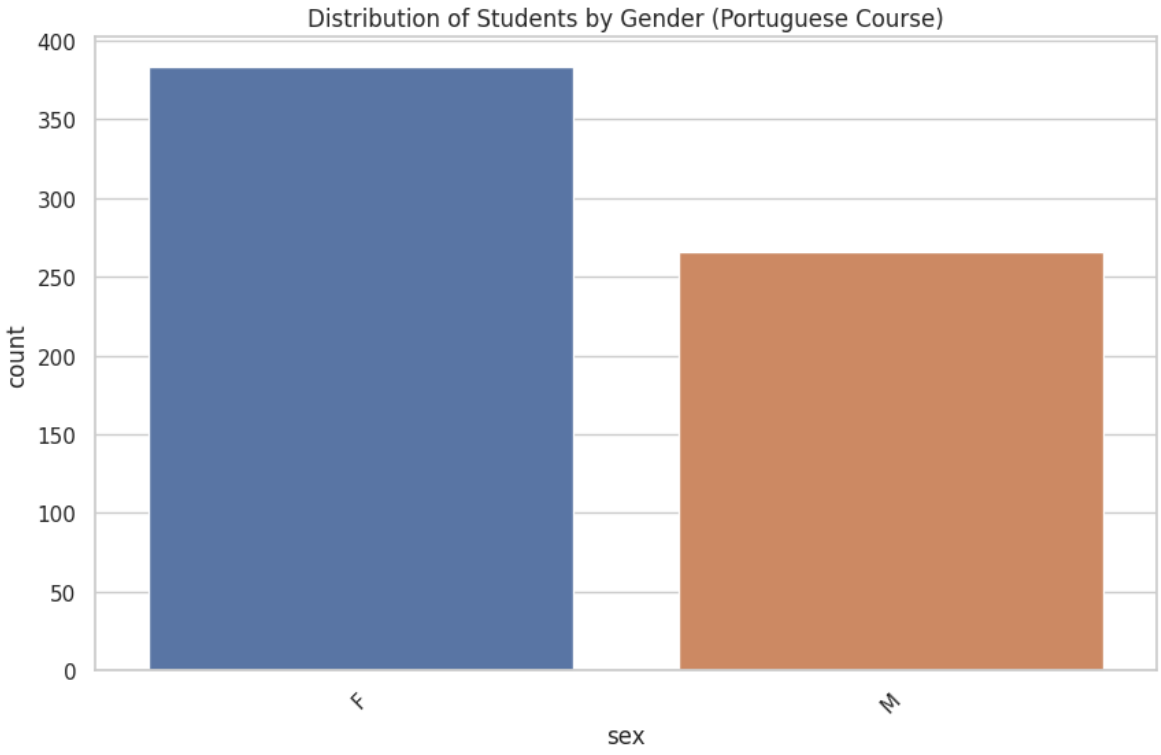
In [66]:
```
# Function to plot bar charts for nominal variables
def plot_nominal_data(df, column, title):
    plt.figure(figsize=(10, 6))
    sns.countplot(data=df, x=column)
    plt.title(title)
```

```
    plt.xticks(rotation=45)
    plt.show()

# Plotting bar charts for a few nominal variables from the Portuguese cou
plot_nominal_data(data_por, 'school', 'Distribution of Students by School
```



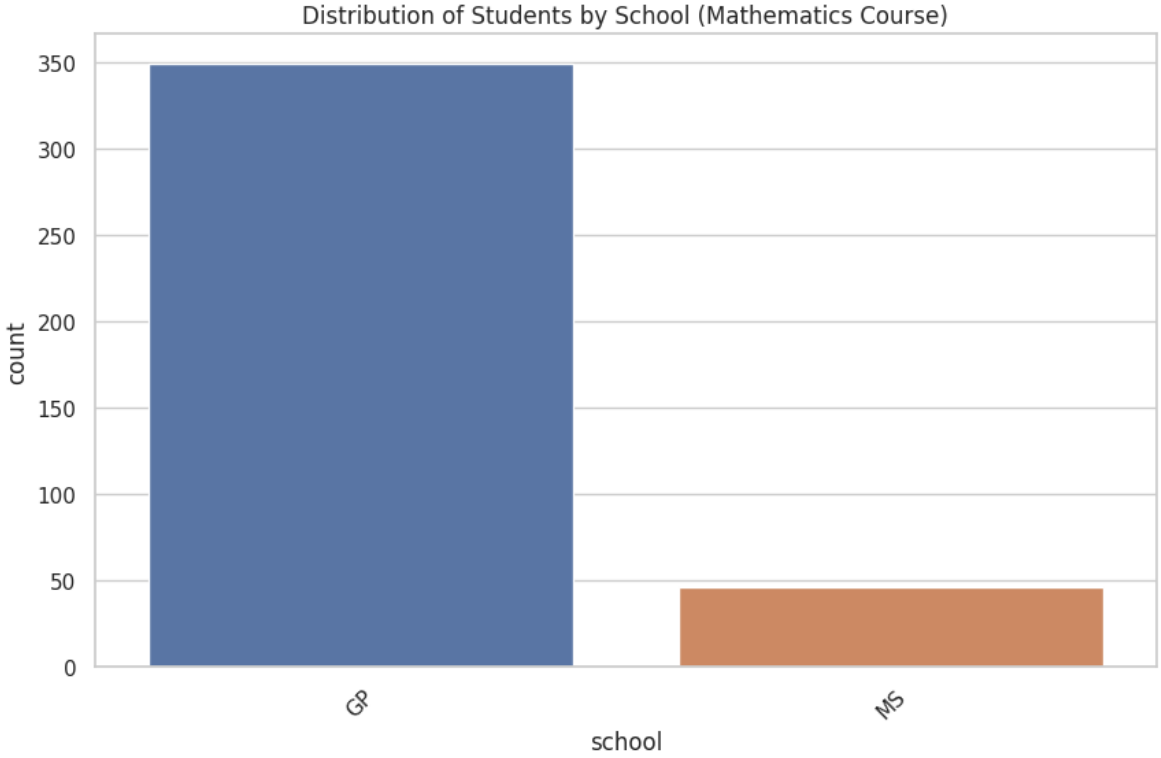
The plot displays a bar chart comparing the distribution of students by school for a Portuguese course. Two bars represent two different schools labelled "GP" and "MS." The school represented by "GP" has a much higher number of students, exceeding 400, while the school represented by "MS" has a significantly lower number, approximately 225 students. This indicates that the Portuguese course is more popular or has greater capacity at the "GP" school than the "MS" school.

In [67]:
```
# Plotting bar charts for a few nominal variables from the Portuguese cou
plot_nominal_data(data_por, 'sex', 'Distribution of Students by Gender (P
```



The plot is a bar chart representing the distribution of students by gender in a Portuguese course. It shows two bars, one for females (indicated with "F") and one for males (indicated with "M"). The count for females is higher, reaching nearly 400, while the count for males is slightly lower, around 250. This suggests that the Portuguese course has a higher enrollment of female students than male students.

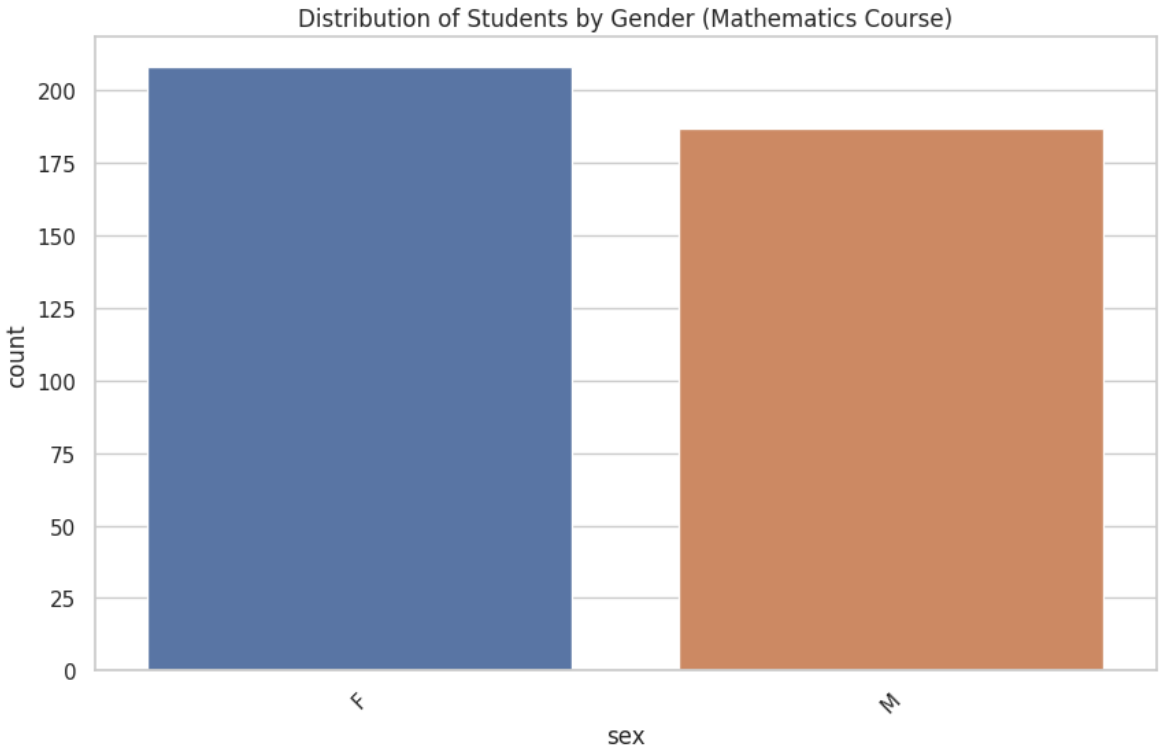
In [68]:
```
# Plotting bar charts for a few nominal variables from the Mathematics co
plot_nominal_data(data_mat, 'school', 'Distribution of Students by School
```



The plot shows a bar chart that illustrates the distribution of students by school in a Mathematics course. Two bars represent two schools, labelled "GP" and "MS." The

bar for "GP" is relatively high, indicating that there are around 350 students enrolled in the Mathematics course at this school. In contrast, the bar for "MS" is much shorter, suggesting that there are about 50 students enrolled in the Mathematics course at that school. This indicates a significantly larger Mathematics class size or student population taking the course at the "GP" school than at the "MS" school.

In [69]:
```
# Plotting bar charts for a few nominal variables from the Mathematics co
plot_nominal_data(data_mat, 'sex', 'Distribution of Students by Gender (M
```

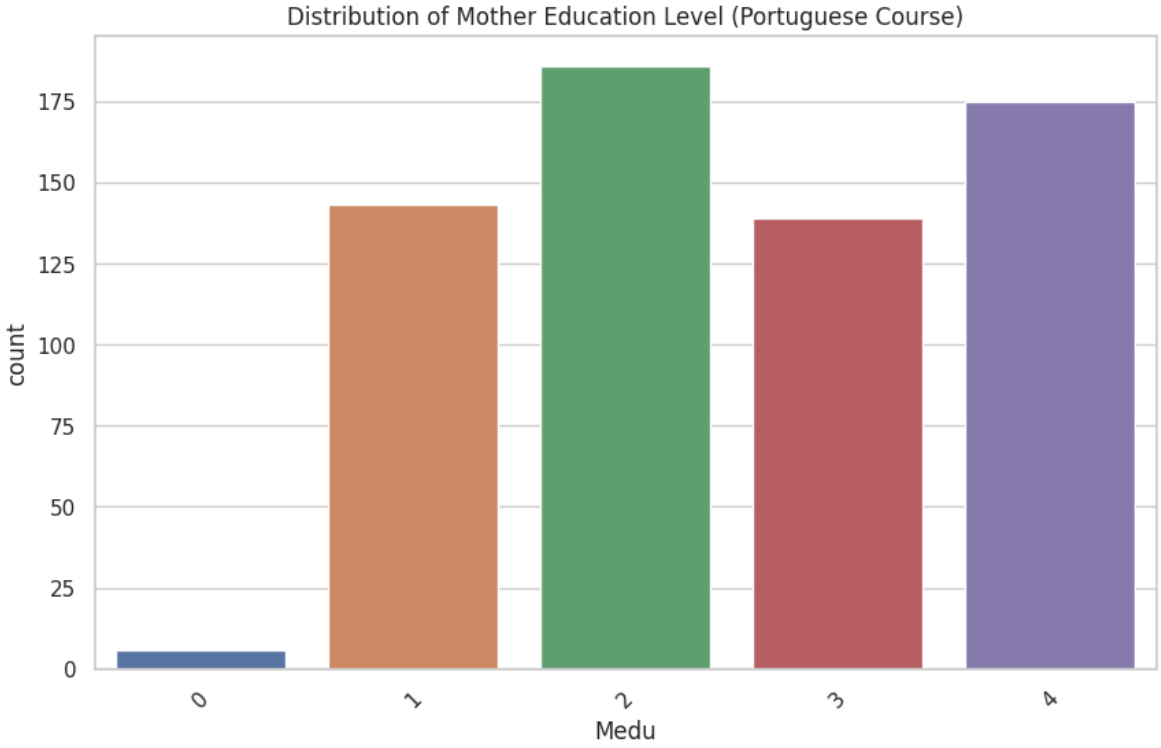


This bar chart shows the distribution of students by gender in a Mathematics course. The chart has two bars, one representing females ("F") and the other representing males ("M"). The bar for females is just over 200, indicating the number of female students, while the bar for males is slightly less than 200, indicating the number of male students. The distribution is relatively balanced between genders, with a marginally higher number of female students than male students enrolled in the Mathematics course.

The bar charts below represent the distribution of students by their mothers' and fathers' education levels in the Portuguese course and Mathematics course datasets. These are examples of ordinal variables, where the education levels are ordered from lowest to highest.

In [70]:
```
# Function to plot bar charts for ordinal variables
def plot_ordinal_data(df, column, title, order=None):
    plt.figure(figsize=(10, 6))
    sns.countplot(data=df, x=column, order=order)
    plt.title(title)
    plt.xticks(rotation=45)
    plt.show()

# Plotting bar charts for a few ordinal variables from the Portuguese cou
```

```
education_levels = sorted(data_por['Medu'].unique())
plot_ordinal_data(data_por, 'Medu', 'Distribution of Mother Education Lev
```

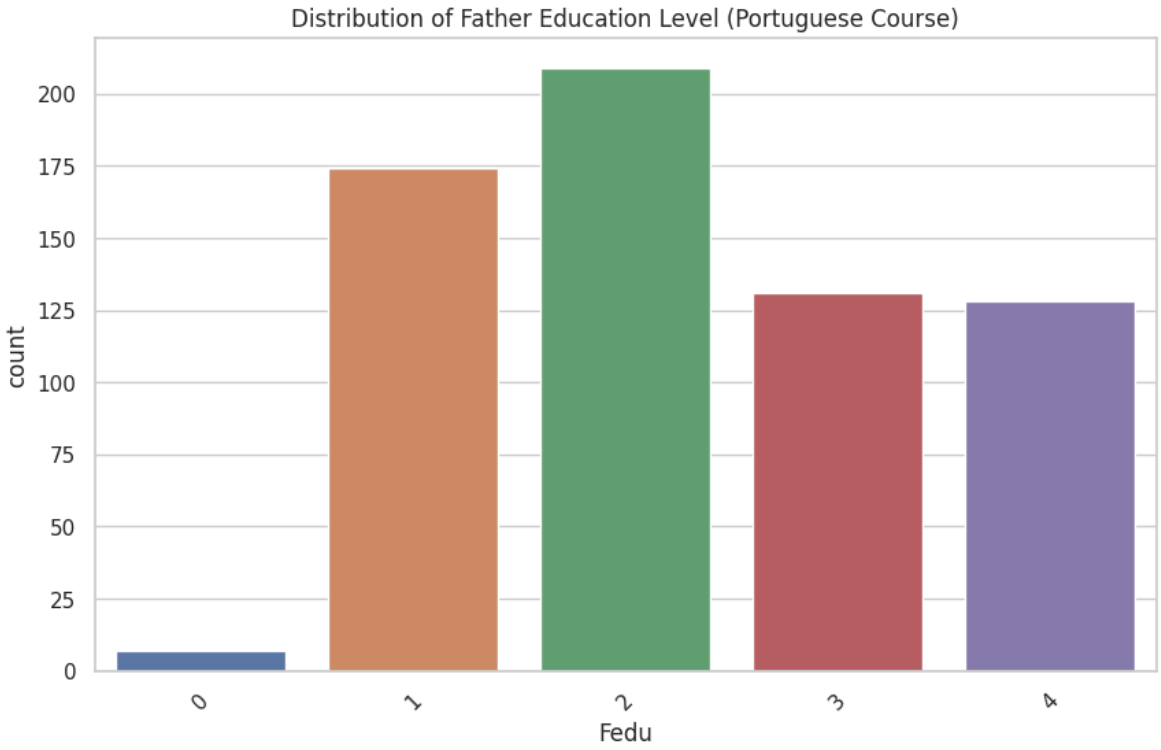


The bar chart illustrates the distribution of students' mothers' education levels for those enrolled in a Portuguese course. The education levels are not explicitly labelled but are indicated by the 'Medu' axis, which stands for "Mother's Education." There are five categories represented by five differently coloured bars, each corresponding to a different level of education, such as none, primary, secondary, university and higher education.

The blue bar, representing the lowest education level, has the fewest students, with the count close to zero. The orange bar, indicating a level just above the red bar, has the count, with almost 150 students. The purple bar, which represents the second highest level of education, has a count slightly lower than the green bar, around 175 students. Finally, the green bar, which indicates the highest level of the mother's education, has a count that is similar to the purple bar.

This distribution suggests that most students enrolled in the Portuguese course have mothers with at least a secondary level of education, with fewer students having mothers with the lowest or highest level of education.

In [71]:
```
# Plotting bar charts for a few ordinal variables from the Portuguese cou
plot_ordinal_data(data_por, 'Fedu', 'Distribution of Father Education Lev
```

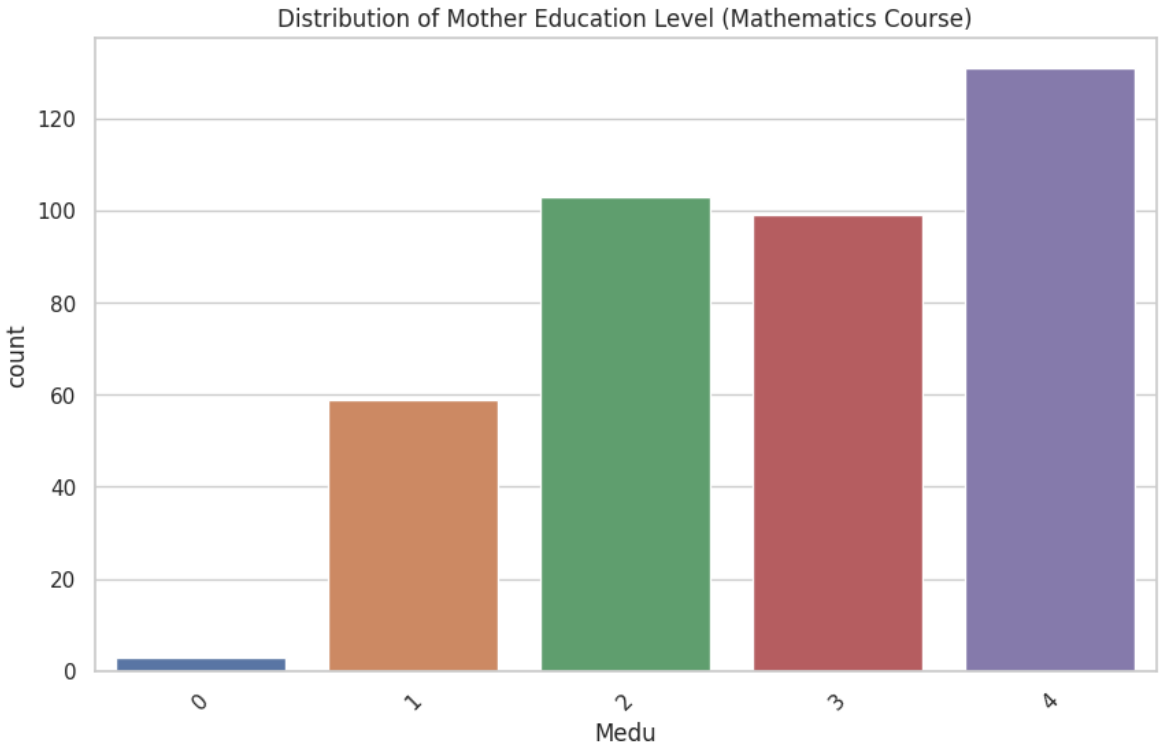


The bar chart presents the distribution of students' fathers' education levels for those enrolled in a Portuguese course. Like the previous chart on mothers' education, the 'Fedu' axis likely stands for "Father's Education." There are five bars, each a different colour, representing different levels of education. The labels for the levels of education correspond to none, primary, secondary, university and higher education levels.

The first bar, blue, represents the lowest education level and has the lowest number of students, with the count being meagre, near zero. The orange bar, indicating the next level up from the lowest, has the second highest count, with around 175 students. The green bar represents the highest education level, with little more than 200 students. The red bar, which could denote the third level of education, has a count just under the green bar, around 125 students. Lastly, the purple bar, which may represent the higher education level of education, has a similar count to the red bar, around 125 students.

This distribution indicates that most students enrolled in the Portuguese course have fathers with a level of education that falls in the middle range, with fewer students having fathers at the lowest or the highest levels of education. This finding is in line with the previous chart, suggesting a trend where a significant number of students come from families with at least a secondary level of education.

In [72]:
```
# Plotting bar charts for a few ordinal variables from the Mathematics co
plot_ordinal_data(data_mat, 'Medu', 'Distribution of Mother Education Lev
```

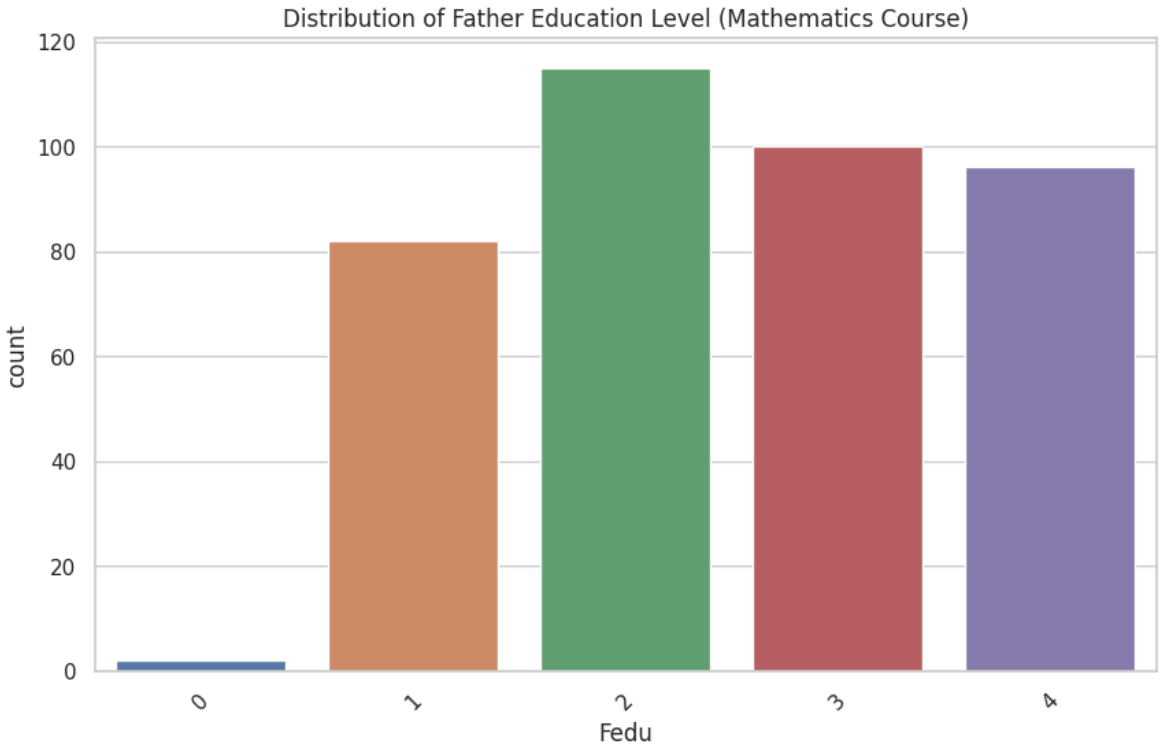


The bar chart illustrates the distribution of students by their mothers' education levels for those enrolled in a Mathematics course. The 'Medu' axis is "Mother's Education," and five bars represent different education levels. The labels for the levels of education correspond to none, primary, secondary, university and higher education.

- The first bar (far left), which is blue and represents the lowest education level, has the least number of students, with the count close to zero.
- The second bar (orange), which represents the next level up from the lowest, has a count of about 60 students.
- The third bar (green), likely representing secondary education, with more than 100 students.
- The fourth bar (red), denoting a university level, has a count close to the red bar, suggesting around 100 students.
- The fifth bar (purple), possibly indicating the highest level of education, has the highest count with more than 120 students.

This distribution suggests that most students in the Mathematics course have mothers with at least secondary education. It also indicates a significant number of students with fathers who have achieved higher education, while very few have mothers with the lowest level of education.

In [73]:
```
# Plotting bar charts for a few ordinal variables from the Mathematics co
plot_ordinal_data(data_mat, 'Fedu', 'Distribution of Father Education Lev
```



The bar chart illustrates the distribution of students by their fathers' education level in a Mathematics course. The 'Fedu' axis likely represents "Father's Education," and five bars correspond to different education levels, including none, primary, secondary, university and higher education.

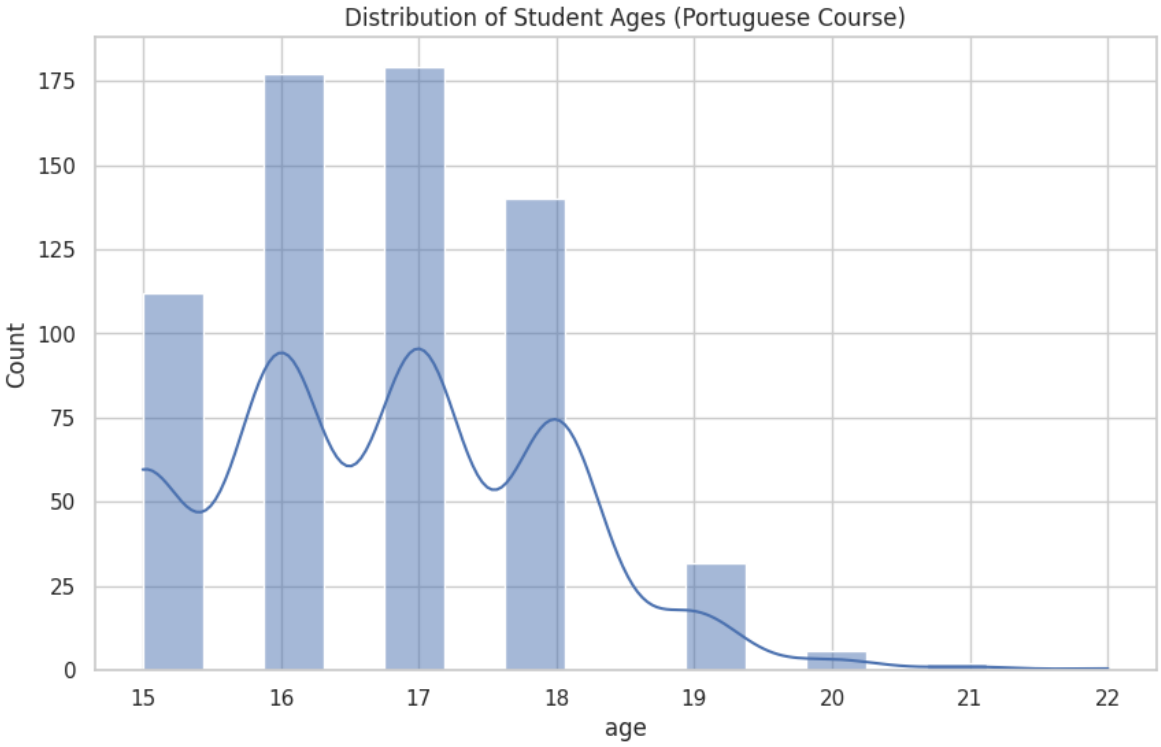
- The first bar (far left), blue, represents the smallest group of students and likely indicates fathers with no education; the count is close to zero.
- The second bar (orange) represents a larger group, around 80 students, indicating fathers with primary education.
- The third bar (green) is the tallest, representing the largest group of students, almost 120 fathers with secondary education.
- The fourth bar (red) has a slightly lower count than the green bar, representing fathers with some level of post-secondary education.
- The fifth bar (purple) represents a group similar in size to the red bar, indicating a substantial number of students with fathers who have attained higher education.

This distribution suggests that most students in the Mathematics course have fathers with at least secondary education, with a significant number also having fathers with higher education. It also indicates that very few students have fathers with no education. This could imply a correlation between fathers' education level and their children's enrollment in higher-level mathematics courses.

The histograms below show student ages, absences, and final grades (G3) distributions in the Portuguese course and Mathematics course datasets. These visualizations help in understanding the spread and skewness of numerical data.

In [74]:
```
# Function to plot histograms for numerical variables
def plot_numerical_data(df, column, title, bins=None):
```

```
    plt.figure(figsize=(10, 6))
    if bins is not None:
        sns.histplot(data=df, x=column, bins=bins, kde=True)
    else:
        sns.histplot(data=df, x=column, kde=True)
    plt.title(title)
    plt.show()

# Plotting histograms for a few numerical variables from the Portuguese c
plot_numerical_data(data_por, 'age', 'Distribution of Student Ages (Portu
```



The plot is a histogram overlaid with a line graph showing the distribution of student ages in a Portuguese course. The x-axis represents the age of the students, ranging from 15 to 22 years old, while the y-axis represents the count of students at each age.

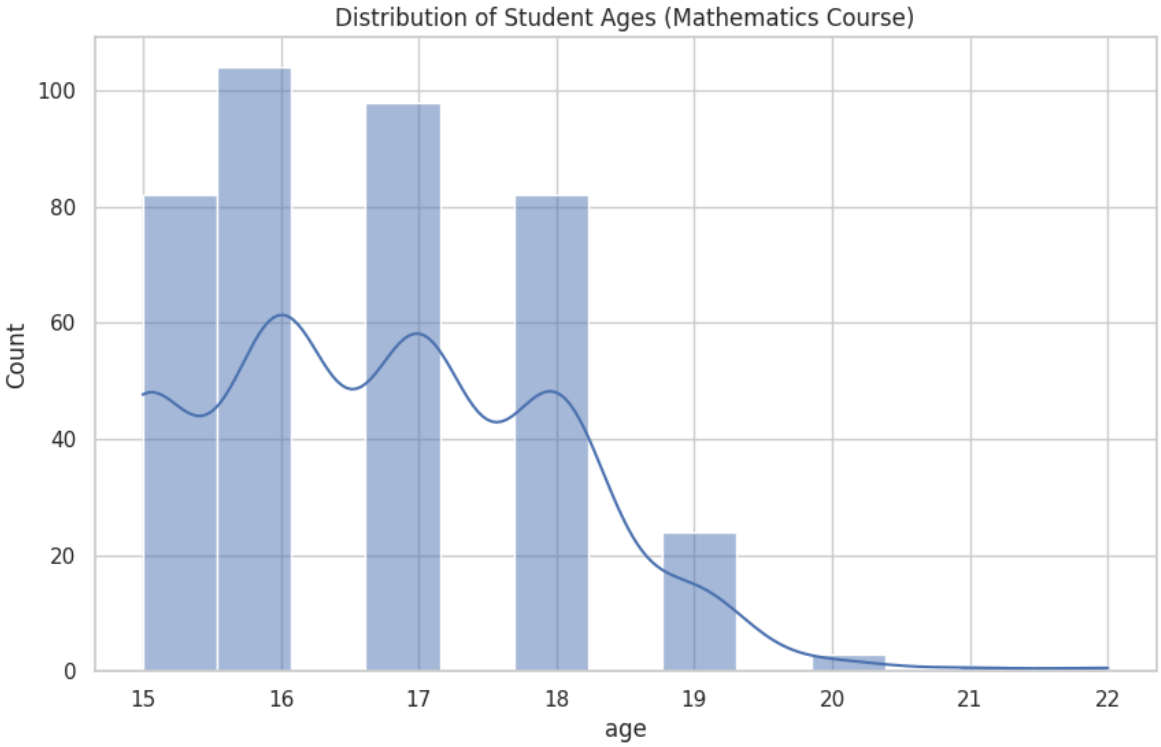
From the histogram, we can observe the following:

- A smaller group of students is 15 years old, with a count of around 112.
- The largest groups of students are 16 and 17, with counts surpassing 175.
- At 18 years old, there is a notable decrease in the number of students, with the count around 140.
- The count of students decreases as age increases, with very few students aged 19 and above.

The line graph, representing the same data, shows a peak at 16 and 17 years old, followed by a sharp decline and a more gradual decline from 18 years old onwards.

This indicates that the majority of students in the Portuguese course are traditionally aged high school students, with a significant drop-off in enrollment for students older than 18, which could be post-high school age.

In [75]:
```
# Function to plot histograms for numerical variables
def plot_numerical_data(df, column, title, bins=None):
    plt.figure(figsize=(10, 6))
    if bins is not None:
        sns.histplot(data=df, x=column, bins=bins, kde=True)
    else:
        sns.histplot(data=df, x=column, kde=True)
    plt.title(title)
    plt.show()

# Plotting histograms for a few numerical variables from the Portuguese c
plot_numerical_data(data_mat, 'age', 'Distribution of Student Ages (Mathe
```



The histogram with an overlaid line graph represents the distribution of student ages in a Mathematics course. The ages range from 15 to 22 years old, suggesting that the course includes high school and college-aged students.

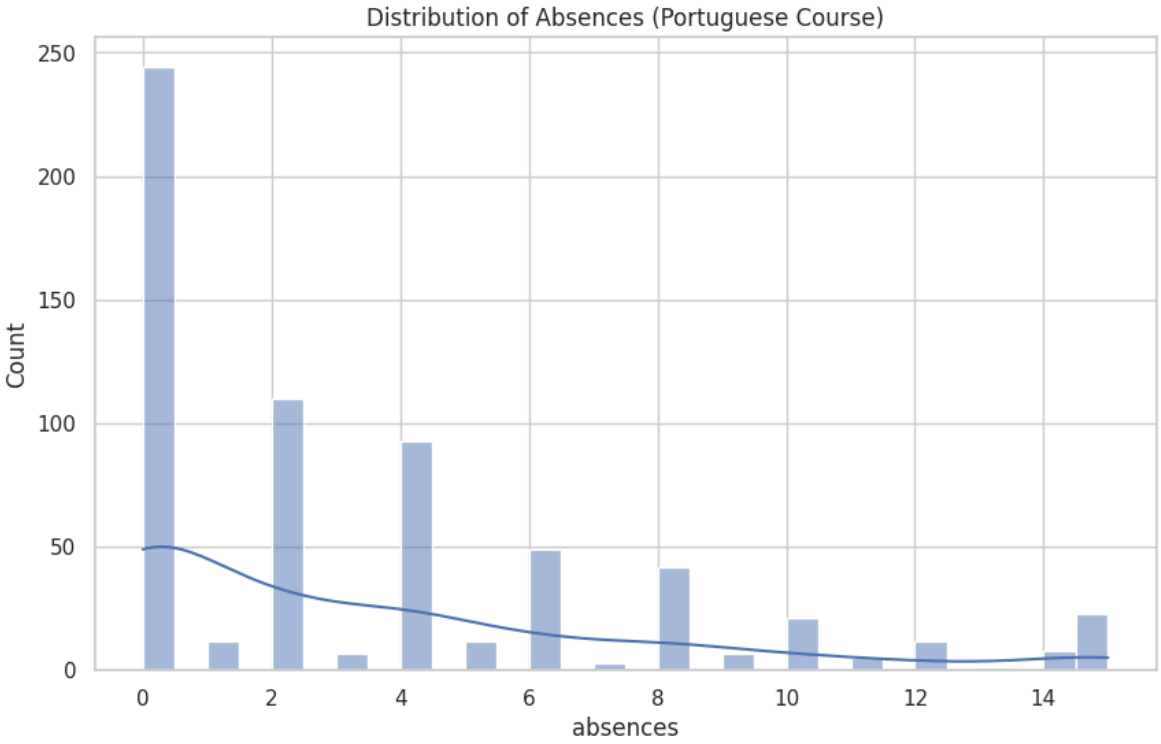
Here is a summary of the age distribution:

- There is a significant representation of 16 and 17-year-old students, with the peak count reaching close to 100 for 17-year-olds, indicating this age group is the most common in the course.
- The number of 15-year-old students is also relatively high, suggesting early high school engagement in the mathematics course.
- For students aged 18, there is a decrease in count to around 80.
- The number of students continues to decline sharply for those aged 19 and above, indicating fewer older students are enrolled in the course.
- By the age of 22, there are very few students left in the course, which might indicate either successful completion of the course by most students before this age or a drop in enrollment as students age out of the traditional school system.

The distribution is skewed towards the younger students, typical of a high school

setting with some overlap into early college years.

In [76]:
```
# Plotting histograms for a few numerical variables from the Portuguese c
plot_numerical_data(data_por, 'absences', 'Distribution of Absences (Port
```



The plot is a histogram with an overlaid line graph showing the distribution of absences among students in a Portuguese course. The x-axis indicates the number of absences, and the y-axis shows the count of students for each number of absences.

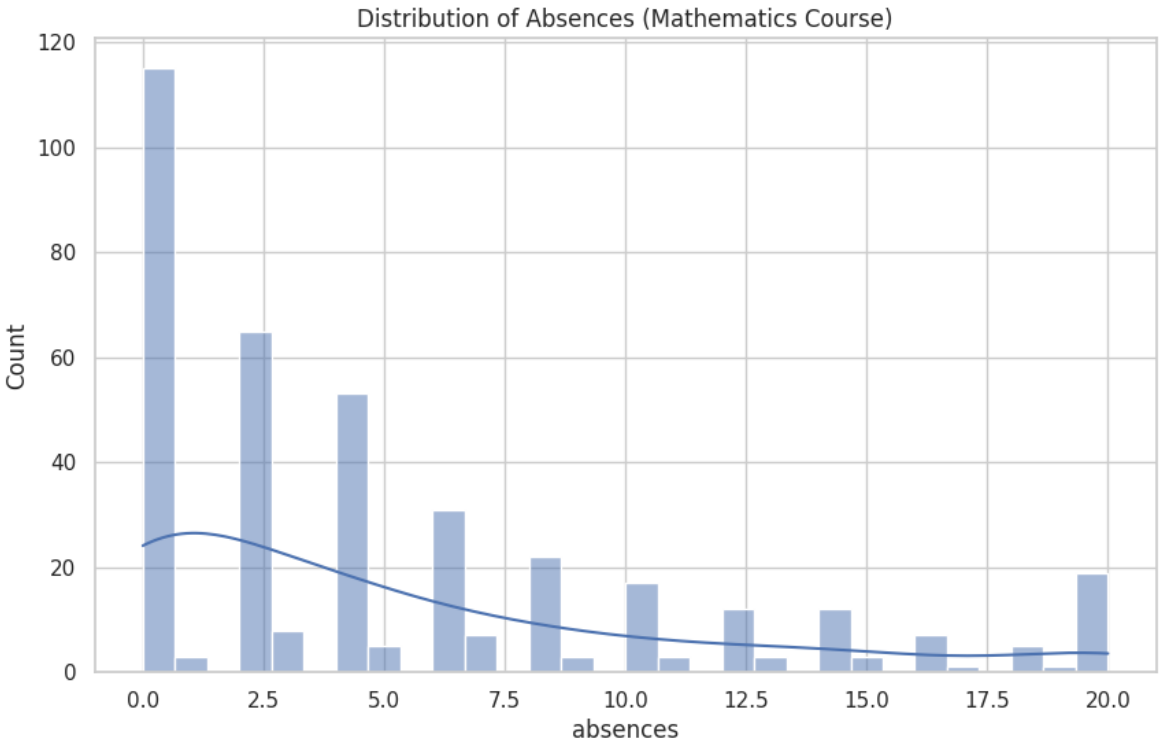
From the histogram:

- The most common number of absences is zero, with over 250 students not having any absences.
- There is a significant drop to around 110 students with two absences.
- The count further decreases for four absences, with around 90 students.
- There is a smaller peak at six absences, with a count of 50 students.
- From 6 absences onwards, the number of students with absences gradually decreases, with very few students having ten or more absences.

The line graph follows the pattern of the bars, emphasizing that as the number of absences increases, the number of students with that many absences decreases.

This suggests that most students in the Portuguese course are generally regular attenders, with a small number having high absence counts. The pattern shows a typical right-skewed distribution often seen in attendance data, where a large number of students have few absences and a small number have many.

In [77]:
```
# Plotting histograms for a few numerical variables from the Mathematic c
plot_numerical_data(data_mat, 'absences', 'Distribution of Absences (Math
```



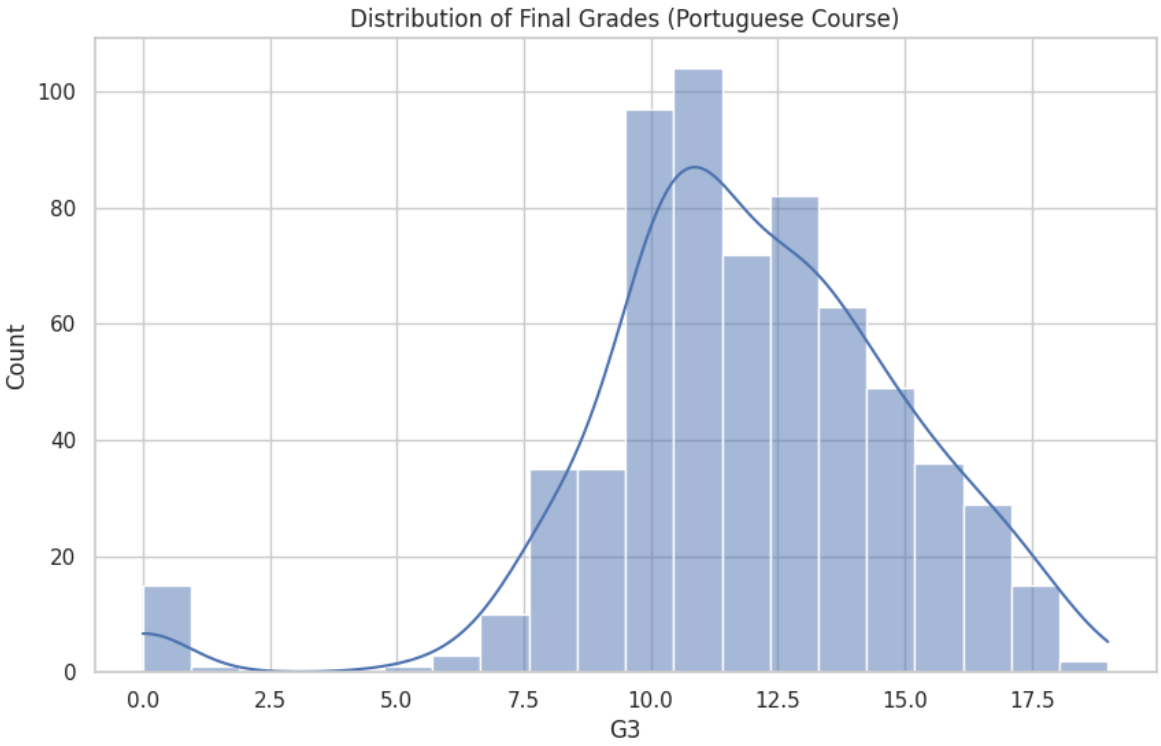
The histogram with an overlaid line graph shows the distribution of absences among students in a Mathematics course. The x-axis denotes the number of absences, ranging from 0 to 20, and the y-axis indicates the count of students for each number of absences.

The distribution reveals:

- A high number of students, over 100, have no absences, suggesting a reasonable attendance rate for a portion of the class.
- There is a significant drop as the number of absences increases, with around 60 students having two absences.
- The frequency of students' absences decreases as the number of absences increases, with smaller peaks around 5 and 10 absences.
- Beyond ten absences, the number of students with higher absence counts becomes relatively low, although there is a slight increase at 20 absences.

The line graph helps visualize the overall trend, indicating that most students have fewer absences and only a few have many absences. This pattern often reflects typical attendance behaviour, where most students aim to maintain regular attendance, with only a few having many absences, possibly due to various personal, academic, or health-related issues.

In [78]:
```
# Plotting histograms for a few numerical variables from the Portuguese c
plot_numerical_data(data_por, 'G3', 'Distribution of Final Grades (Portug
```



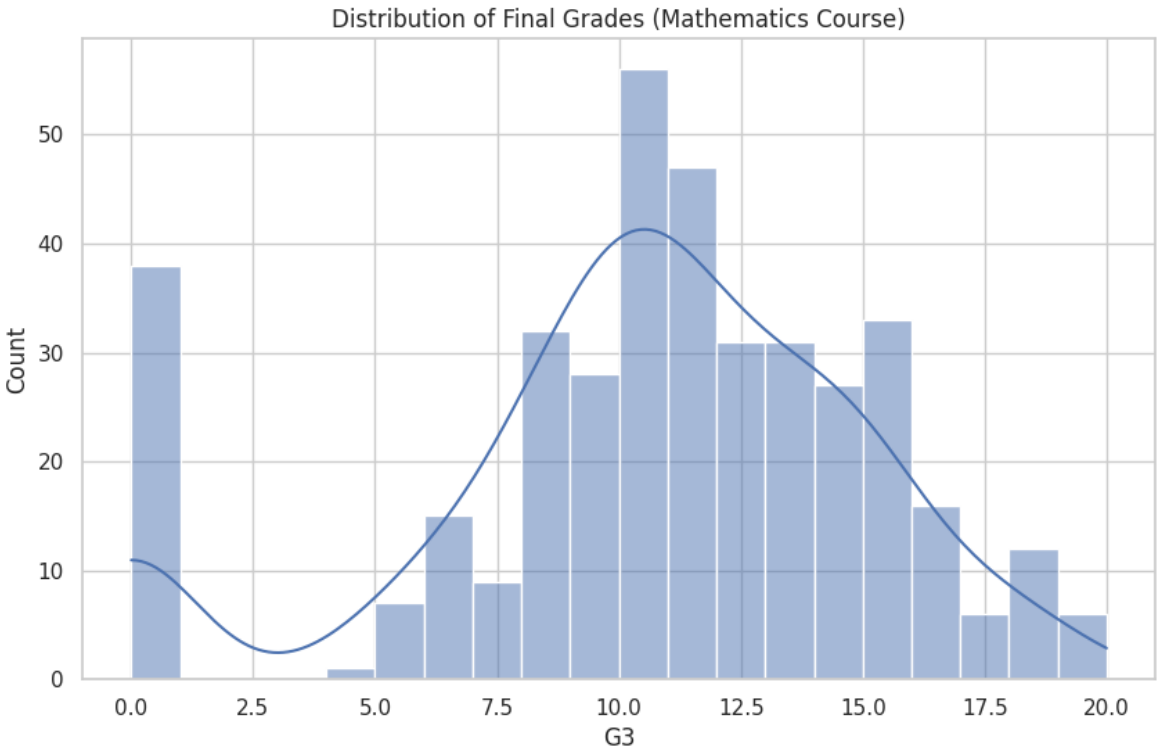
The histogram and overlaid line graph display the distribution of final grades for students in a Portuguese course. The x-axis labelled 'G3' suggests that these are final grades on a scale from 0 to 20. The y-axis shows the count of students for each grade range.

The chart indicates:

- A small number of students have shallow final grades (0-5).
- There is an increase in the count of students with grades around 7.5 to 10.
- The distribution peak is between 10 and 12.5, showing the highest concentration of students' grades.
- Beyond this peak, the number of students gradually decreases as grades increase, with very few obtaining grades above 15.

The distribution has a bell-shaped curve, skewed slightly to the right, suggesting that while most students achieved middling to good grades, there are fewer instances of very high and shallow grades. This distribution is typical of many academic courses where a central clustering around the mean or median grade is standard, with fewer students at the extremes.

In [79]:
```
# Plotting histograms for a few numerical variables from the Mathematics
plot_numerical_data(data_mat, 'G3', 'Distribution of Final Grades (Mathem
```



The plot is a histogram overlaid with a line graph, showing the distribution of final grades for students in a Mathematics course. The x-axis, labelled 'G3', represents the final grade, which appears to be on a scale from 0 to 20. The y-axis indicates the count of students achieving each grade.

The distribution shows:

- Fewer students with grades between 5 and 7.5.
- The most common grades are centred around 10 to 12.5, where the highest peak of the distribution is located.
- Beyond this peak, the number of students decreases as the grades increase, with relatively few students achieving the highest grades between 17.5 and 20.

The line graph that follows the tops of the bars emphasizes the distribution's bell-shaped curve, indicating a normal distribution of grades, with most students scoring in the middle range. This suggests that while there are students across the performance spectrum, most grades are average, with fewer students at high and low-performance extremes.

Finally, in the distribution of final grades for a Mathematics course, there is a significant number of students, nearly 40, with a grade of zero. This unusual spike could be attributed to various factors, including strict attendance policies leading to automatic failures, failure to submit coursework or sit for exams, penalties for academic dishonesty, or a grading system that assigns zeros for not passing critical components of the course. Administrative errors or incorrect data entry could have contributed to this number. Without additional context regarding the course's grading policy and the specific circumstances of these students, it is not easy to pinpoint the exact cause of the high frequency of zeros.

Combined Data

We will join both datasets, namely 'cleaned_student-por.csv' for Portuguese and 'cleaned_student-mat.csv' for Mathematics because they have the same column names but represent different courses. This approach allows us to analyze the data comprehensively. Our chosen method for merging is concatenation, as it is most suitable for datasets with identical columns that are meant to be stacked together. By concatenating, we will create a single dataset that includes all the students, providing a broader perspective for our analysis.

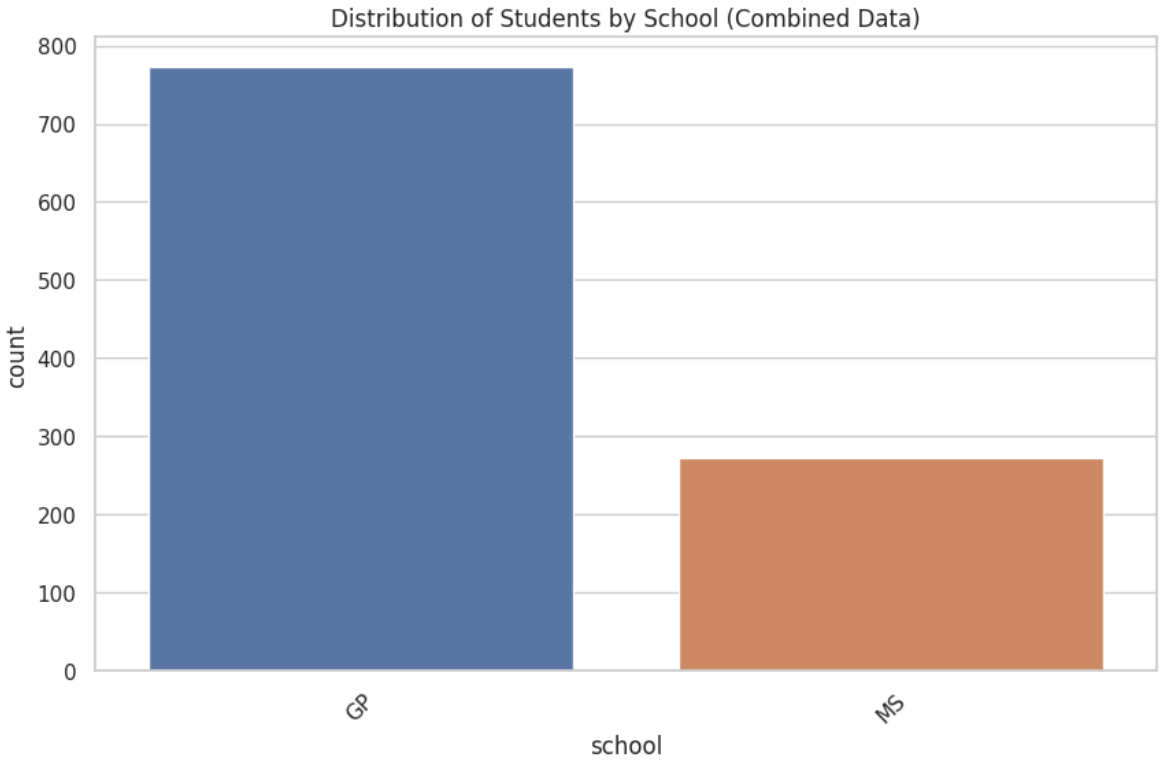
In [80]:
```
# Concatenating the two datasets
# Adding an additional column to each dataset to indicate the course
data_por['course'] = 'Portuguese'
data_mat['course'] = 'Mathematics'

# Concatenating the datasets
combined_student_data = pd.concat([data_por, data_mat], ignore_index=True

# Displaying the first few rows of the combined dataset
combined_student_data_head = combined_student_data.head().T
combined_student_data_head
```

Out[80]:

	0	1	2	3	4
school	GP	GP	GP	GP	GP
sex	F	F	F	F	F
age	18	17	15	15	16
address	U	U	U	U	U
famsize	GT3	GT3	LE3	GT3	GT3
Pstatus	A	T	T	T	T
Medu	4	1	1	4	3
Fedu	4	1	1	2	3
Mjob	at_home	at_home	at_home	health	other
Fjob	teacher	other	other	services	other
reason	course	course	other	home	home
guardian	mother	father	mother	mother	father
traveltime	2	1	1	1	1
studytime	2	2	2	3	2
failures	0	0	0	0	0
schoolsup	yes	no	yes	no	no
famsup	no	yes	no	yes	yes
paid	no	no	no	no	no
activities	no	no	no	yes	no
nursery	yes	no	yes	yes	yes
higher	yes	yes	yes	yes	yes
internet	no	yes	yes	yes	no
romantic	no	no	no	yes	no
famrel	4	5	4	3	4
freetime	3	3	3	2	3
goout	4	3	2	2	2
Dalc	1	1	2	1	1
Walc	1	1	3	1	2
health	3	3	3	5	5
absences	4	2	6	0	0
G1	0	9	12	14	11
G2	11	11	13	14	13
G3	11	11	12	14	13
course	Portuguese	Portuguese	Portuguese	Portuguese	Portuguese

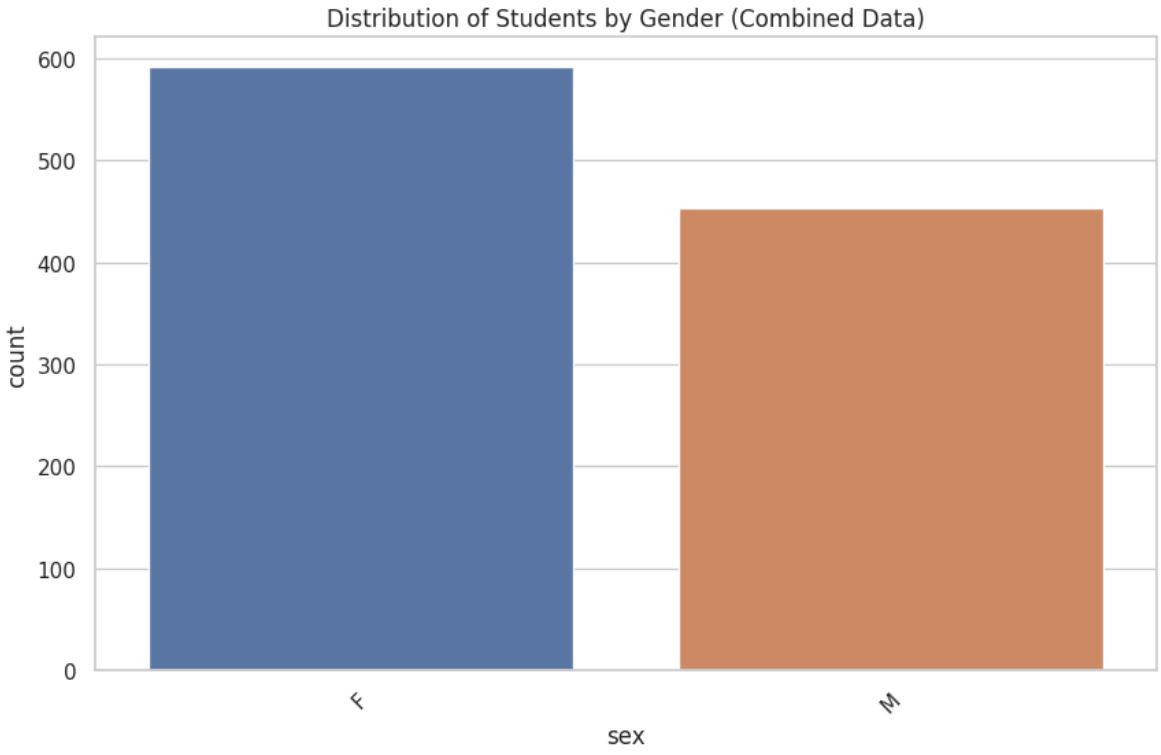
The bar charts below show the distribution of students by school, gender, and course (Portuguese vs Mathematics) in the combined dataset. These visualizations provide an overview of the categorical (nominal) variables.

In [81]:
```
# Plotting bar charts for a few nominal variables from the combined datas
plot_nominal_data(combined_student_data, 'school', 'Distribution of Stude
```



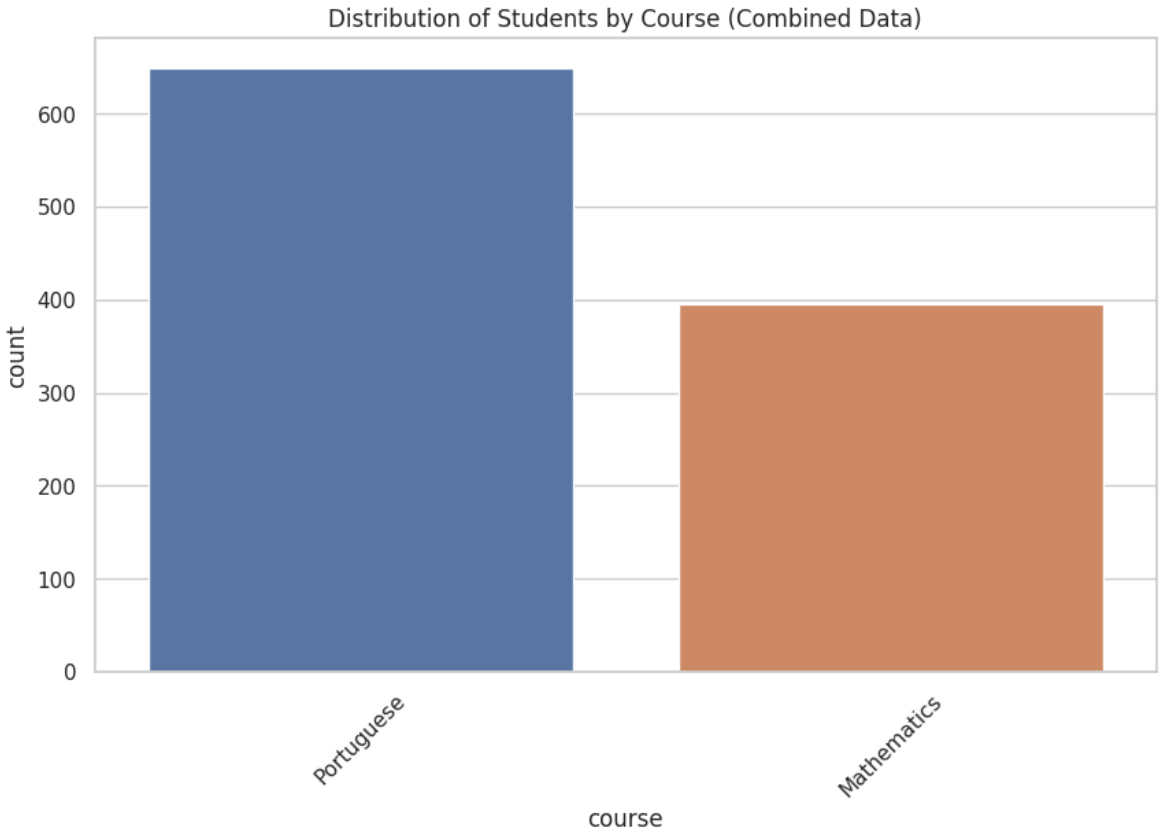
The bar chart depicts the combined distribution of students by school, with two schools labelled "GP" and "MS." The "GP" school has a significantly higher number of students, nearly 800, whereas the "MS" school has almost 300 students. This substantial difference suggests that the "GP" school has a larger student body or a broader course offering that attracts more students. With additional context, whether this disparity is due to differences in school size, course availability, or other factors such as geographical location or school preference is clear.

In [82]:
```
# Plotting bar charts for a few nominal variables from the combined datas
plot_nominal_data(combined_student_data, 'sex', 'Distribution of Students
```



The bar chart displays the combined distribution of students by gender for a dataset encompassing Portuguese and Mathematics courses. The chart shows two bars: the blue bar represents female students, while the orange bar represents male students. The female students' count almost reaches 600, while the male count is somewhat lower, around 450. This indicates that there are more female students than male students in the combined dataset of these courses. This gender distribution can reflect the enrollment patterns in these specific courses or potentially indicate broader trends in educational participation by gender.

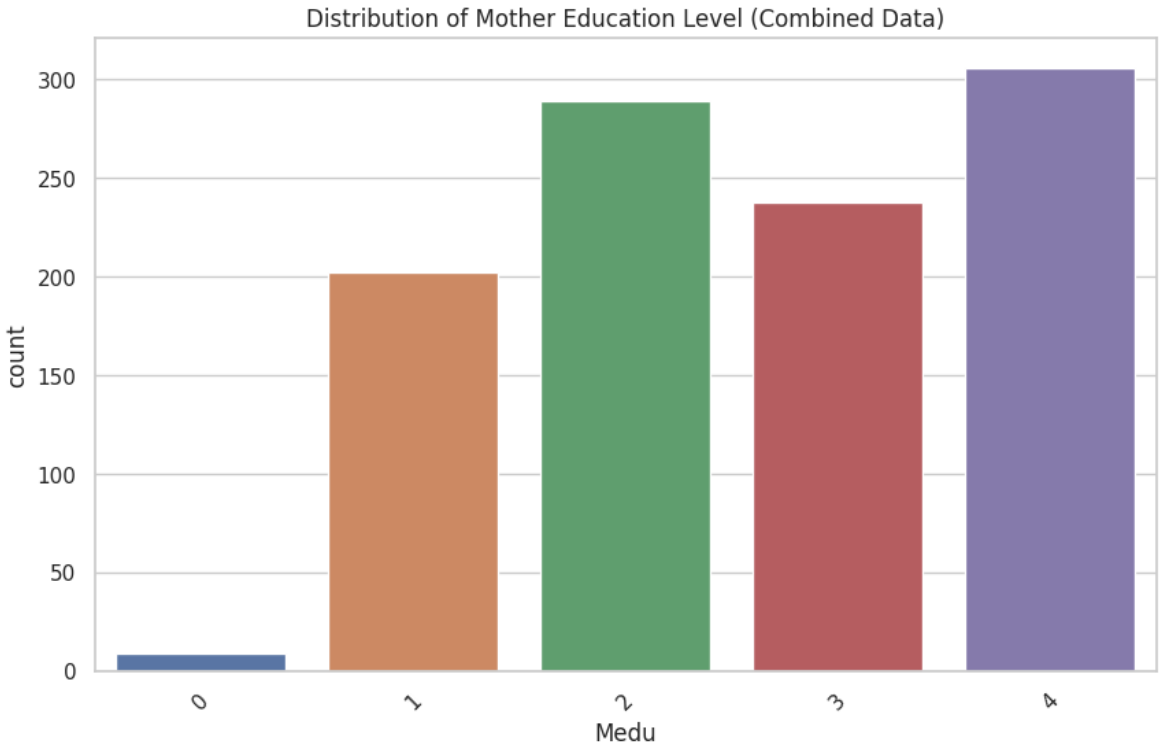
In [83]:
```
# Plotting bar charts for a few nominal variables from the combined datas
plot_nominal_data(combined_student_data, 'course', 'Distribution of Stude
```



The bar chart illustrates the combined distribution of students enrolled in Portuguese and Mathematics courses. The blue bar represents students in the Portuguese course, while the orange bar represents Mathematics course students. There are more than 600 students taking Portuguese and around 400 students taking Mathematics. This suggests that the Portuguese course has a higher enrollment rate than Mathematics within the dataset provided. This difference could reflect various factors such as student preference, course requirements, or availability of seats in each course.

The bar charts below represent the distribution of students by their mothers' and fathers' education levels in the combined dataset. These visualizations help to understand the ordinal variables in the context of both courses.

In [84]:
```
# Plotting bar charts for a few ordinal variables from the combined datas
plot_ordinal_data(combined_student_data, 'Medu', 'Distribution of Mother
```

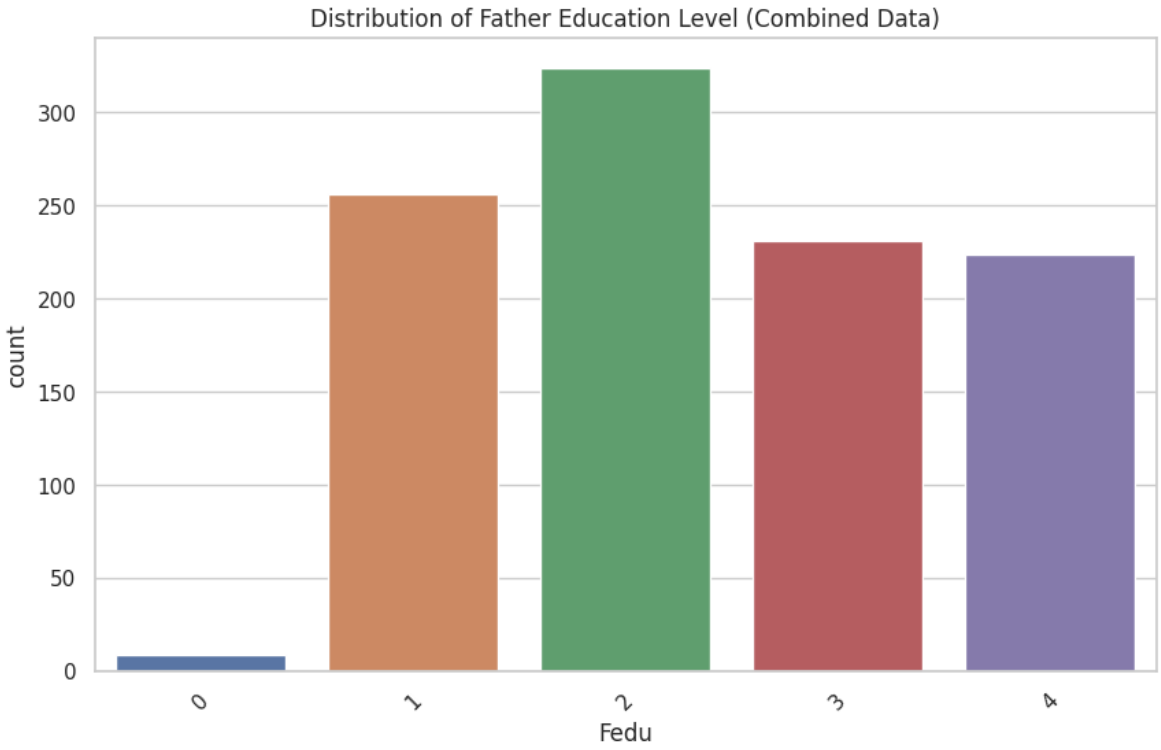


The bar chart presents a combined distribution of students' mothers' education levels across Portuguese and Mathematics courses. The 'Medu' axis is "Mother's Education," with bars representing varying education levels.

- The first bar, coloured blue, represents the lowest education level and shows a minimal count of students, indicating very few mothers with this level of education.
- The second bar, orange, shows a more significant number, around 200 students, suggesting mothers with primary education.
- The third bar, green, has a count approaching the purple bar, over 250 and almost 300 students, likely indicating mothers with secondary education.
- The fourth bar, red, with a count slightly less than the green bar, suggests mothers with some post-secondary education.
- The fifth bar, purple, represents the most significant count, with 300 students with mothers who have attained higher education.

This chart indicates that most students across both courses have mothers with at least secondary education, and a substantial number have mothers with higher education levels. It suggests a potential correlation between the mother's education level and the student's participation in academic courses.

In [85]:
```
# Plotting bar charts for a few ordinal variables from the combined datas
plot_ordinal_data(combined_student_data, 'Fedu', 'Distribution of Father
```



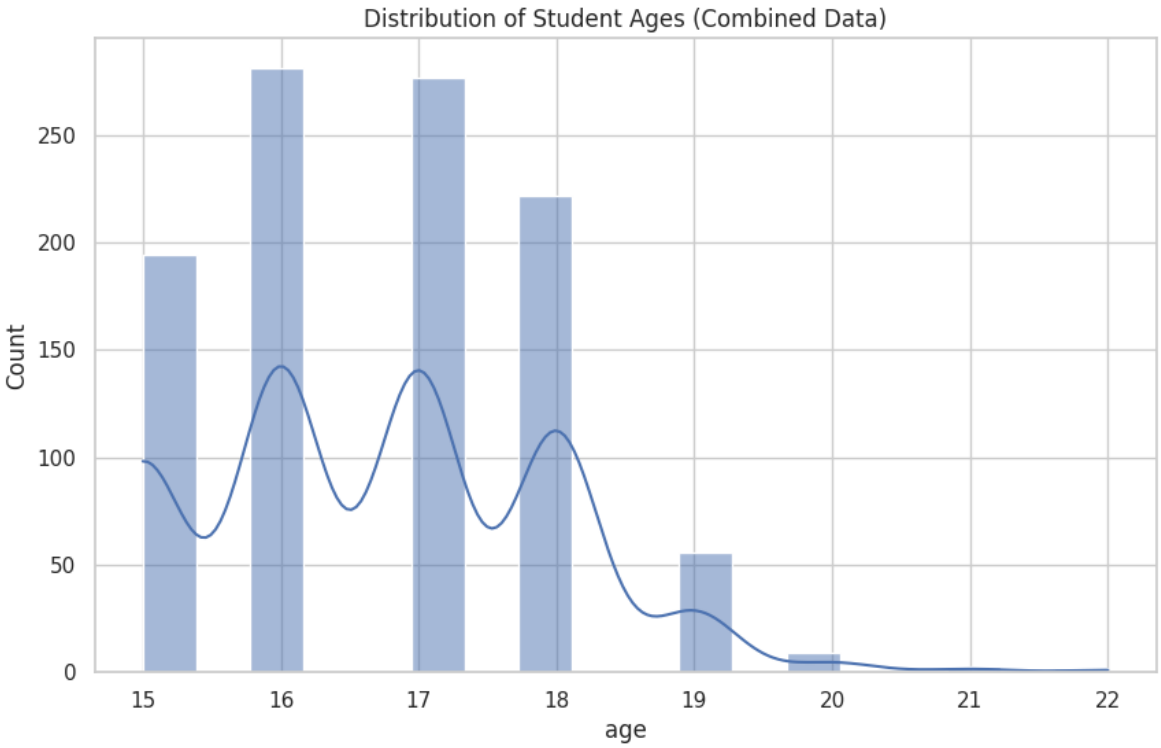
The bar chart illustrates the combined distribution of students' fathers' education levels across Portuguese and Mathematics courses. The 'Fedu' axis stands for "Father's Education," it shows five bars, each representing different levels of education.

- The first bar, coloured blue, is barely visible, indicating very few fathers with the lowest level of education.
- The second bar, orange, represents a more significant number of students, around 250, suggesting fathers with primary education.
- The third bar, green, is the highest, showing over 300 students, likely indicating fathers with secondary education.
- The fourth bar, red, has around 240 students, possibly representing fathers with some post-secondary education.
- The fifth bar, purple, is close to the red bar and represents a significant count, around 230 students, suggesting fathers with higher education.

This distribution indicates that most students in the combined dataset of the courses have fathers with at least secondary education, with a substantial number also having fathers who attained higher education. This could reflect a trend where parental education level, particularly that of fathers, is correlated with student enrollment in academic courses.

The histograms below show student ages, absences, and final grades (G3) distributions in the combined dataset. These plots provide insights into the spread and characteristics of the numerical data across both courses.

In [86]:
```
# Plotting histograms for a few numerical variables from the combined dat
plot_numerical_data(combined_student_data, 'age', 'Distribution of Studen
```



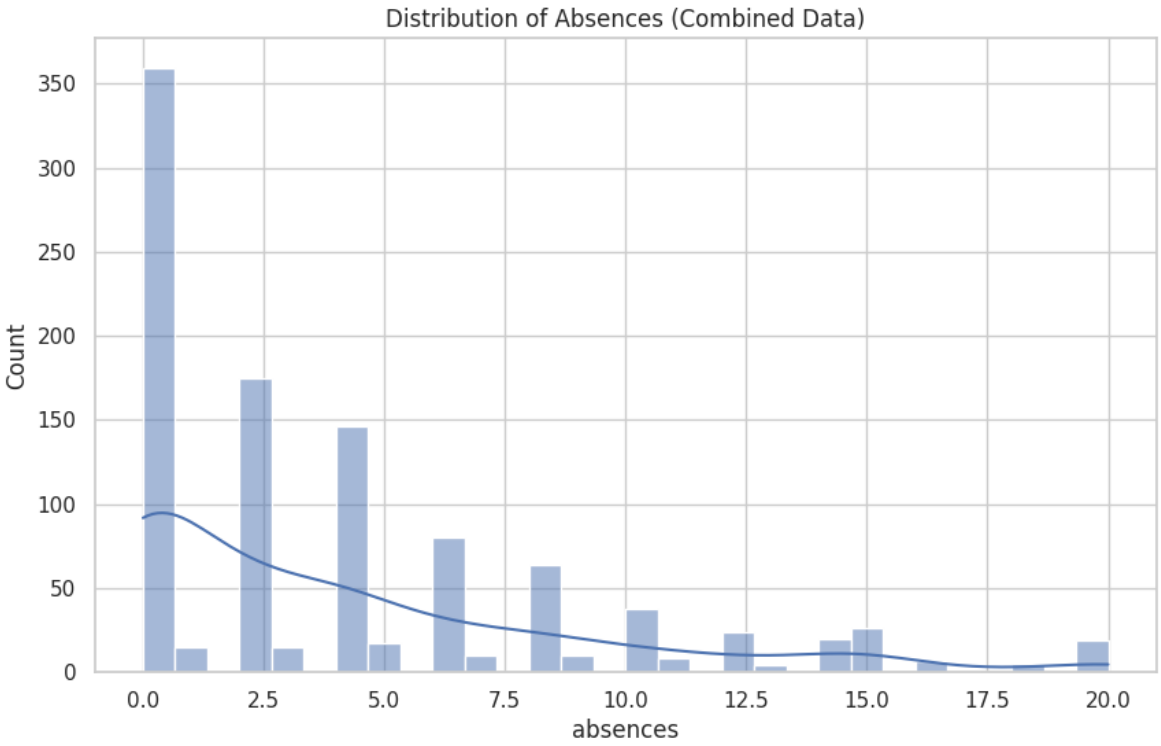
The histogram overlaid with a line graph represents the combined distribution of student ages for Portuguese and Mathematics courses. The ages range from 15 to 22 years old.

The distribution indicates:

- A substantial number of 16 and 17-year-old students, with the count for 16-year-olds being the highest, surpassing 250.
- A smaller, yet significant, number of 15-year-old students, with the count around 200.
- The count decreases for 18-year-olds, falling below 250.
- A marked decline in the number of students begins at age 19 and continues to decrease for older ages.
- By the ages of 21 and 22, there are very few students, with the counts approaching the lower end of the scale.

This age distribution suggests that most students in these courses are of typical high school age, with a noticeable drop-off as they reach college age, which could indicate students completing their secondary education or transitioning to different educational paths or institutions.

In [87]:
```
# Plotting histograms for a few numerical variables from the combined dat
plot_numerical_data(combined_student_data, 'absences', 'Distribution of A
```



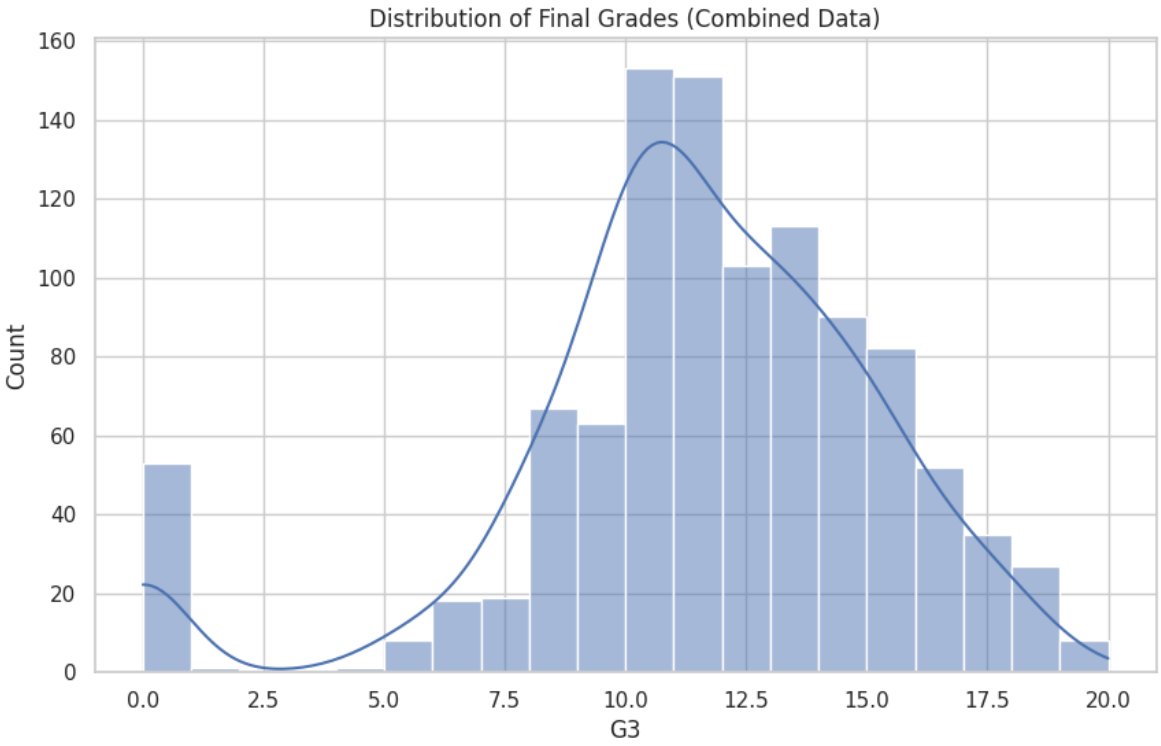
The histogram overlaid with a line graph displays the combined distribution of absences for students in both Portuguese and Mathematics courses. The x-axis lists the number of absences, ranging from 0 to 20, and the y-axis shows the count of students corresponding to each absence level.

Key observations from the graph include:

- The most significant number of students, over 350, have zero absences, indicating a high attendance rate.
- There is a rapid decline to about 180 students with two absences.
- The count continues to decrease as the number of absences increases, with peaks at 5 and 10 absences, although these are significantly lower than the peak at zero absences.
- As the number of absences continues to rise beyond 10, the frequency of students with such absence levels drops, becoming relatively sparse towards the 20 absences mark.

The distribution follows a right-skewed pattern, typical in attendance data, where most students have few absences, and the numbers taper off as absences increase. This indicates that most students regularly attend classes, with a smaller number having high absence rates.

In [88]:
```
# Plotting histograms for a few numerical variables from the combined dat
plot_numerical_data(combined_student_data, 'G3', 'Distribution of Final G
```



The histogram overlaid with a line graph shows the final grade distribution for students in both Portuguese and Mathematics courses. The x-axis, labelled 'G3', represents final grades on a scale from 0 to 20, and the y-axis shows the count of students achieving each grade.

Key observations from the chart:

- A group of almost 60 students with the lowest grades (0.0) indicates potential failures or non-completions.
- The distribution has a primary peak around the grade of 10.0, suggesting that this is the most common final grade among students.
- A secondary smaller peak at around 13.0 indicates another typical final grade range.
- The number of students gradually decreases as the grades approach the highest mark of 20.0, with very few students achieving the top grades.
- The distribution is approximately bell-shaped, centred around the 10.0 to 12.0 grade range, typical of a normal distribution where most students achieve mid-range grades, and fewer students achieve very high or very low grades.

This pattern reflects a standard grading outcome where most students perform moderately well, with fewer extraordinarily high or low-performance instances.

Multivariate Visualizations

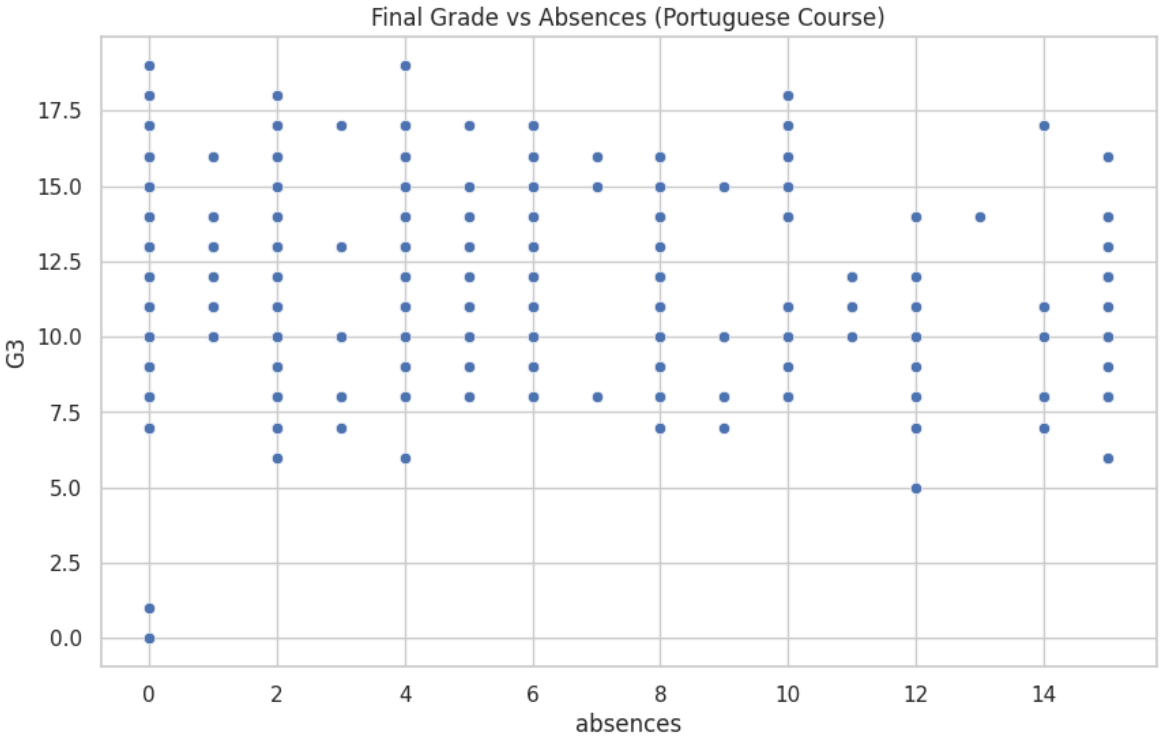
For combinations of different data types, we will use various plots like scatter plots, box plots, and heatmaps depending on the specific combination of data types.

This involves plotting relationships between different types of variables. We'll consider a few examples:

- Numerical-Numerical: Scatter plot of G3 (final grade) vs absences
- Numerical-Nominal: Box plot of G3 (final grade) for each school
- Numerical-Ordinal: Box plot of G3 (final grade) across different Medu (mother's education level) categories

The scatter plot below illustrates the relationship between the number of absences and final grades (G3) in the Portuguese course and Mathematics course datasets. This type of plot is useful for observing trends, outliers, and potential correlations between two numerical variables.

In [89]:
```
# Function to plot scatter plot for numerical-numerical variable combinat
def plot_scatter(df, x, y, title):
    plt.figure(figsize=(10, 6))
    sns.scatterplot(data=df, x=x, y=y)
    plt.title(title)
    plt.show()

# Plotting scatter plot for G3 vs absences in Portuguese course data
plot_scatter(data_por, 'absences', 'G3', 'Final Grade vs Absences (Portug
```



The scatter plot visualizes the relationship between the final grades (G3) and the number of absences for students in a Portuguese course. The x-axis shows the absences ranging from 0 to 14, while the y-axis shows the final grades ranging from 0 to 17.5.

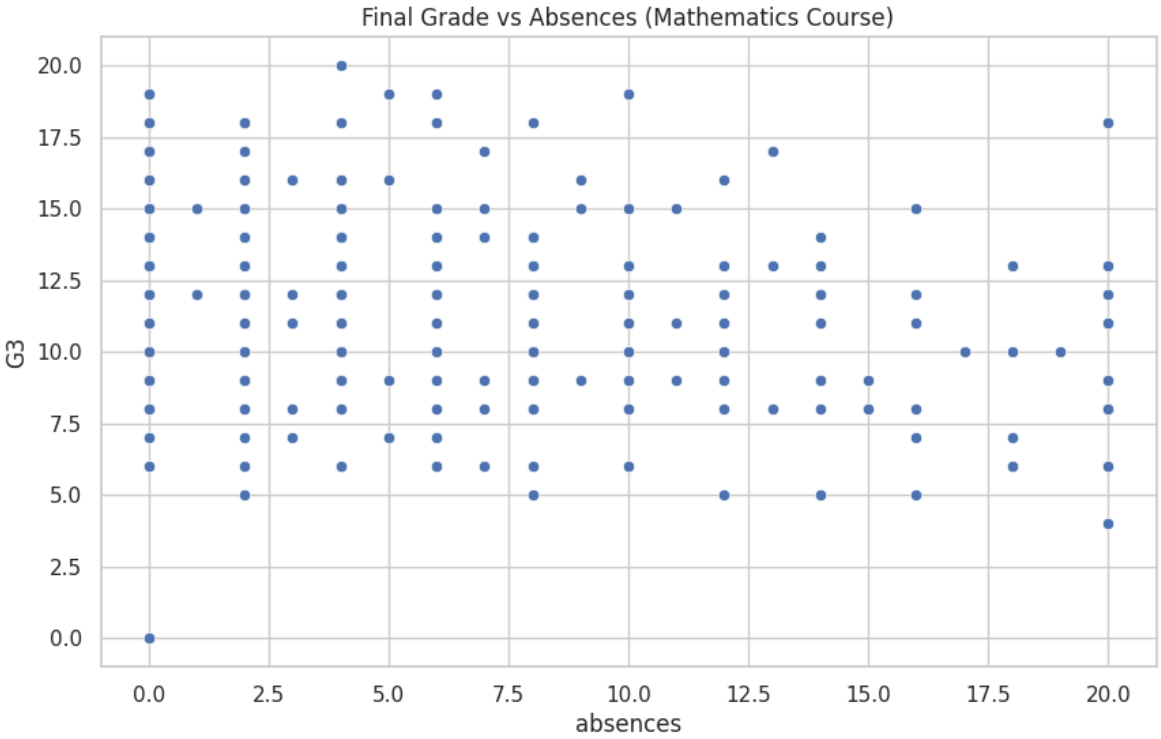
Key observations from the plot:

- There is a wide distribution of grades across all absence levels, indicating that absences do not have a simple direct correlation with final grades.
- Students with zero absences have a range of grades from low to high, suggesting that regular attendance does not necessarily guarantee a high grade.

- There needs to be a clear pattern indicating that more absences lead to lower grades, as students with high absences still have achieved mid to high grades.
- Similarly, some students with fewer absences have low grades, which could suggest that factors other than attendance may significantly impact their final grade.

While one might typically expect to see a trend where more absences correlate with lower grades, this scatter plot does not show a robust and clear correlation, indicating the complexity of factors that contribute to academic performance.

In [90]:
```
# Plotting scatter plot for G3 vs absences in Mathematics course data
plot_scatter(data_mat, 'absences', 'G3', 'Final Grade vs Absences (Mathem
```



The scatter plot shows the relationship between final grades (G3) and the number of absences for students in a Mathematics course. The x-axis displays the number of absences, extending from 0 to 20, while the y-axis shows the final grades, ranging from up to 20.

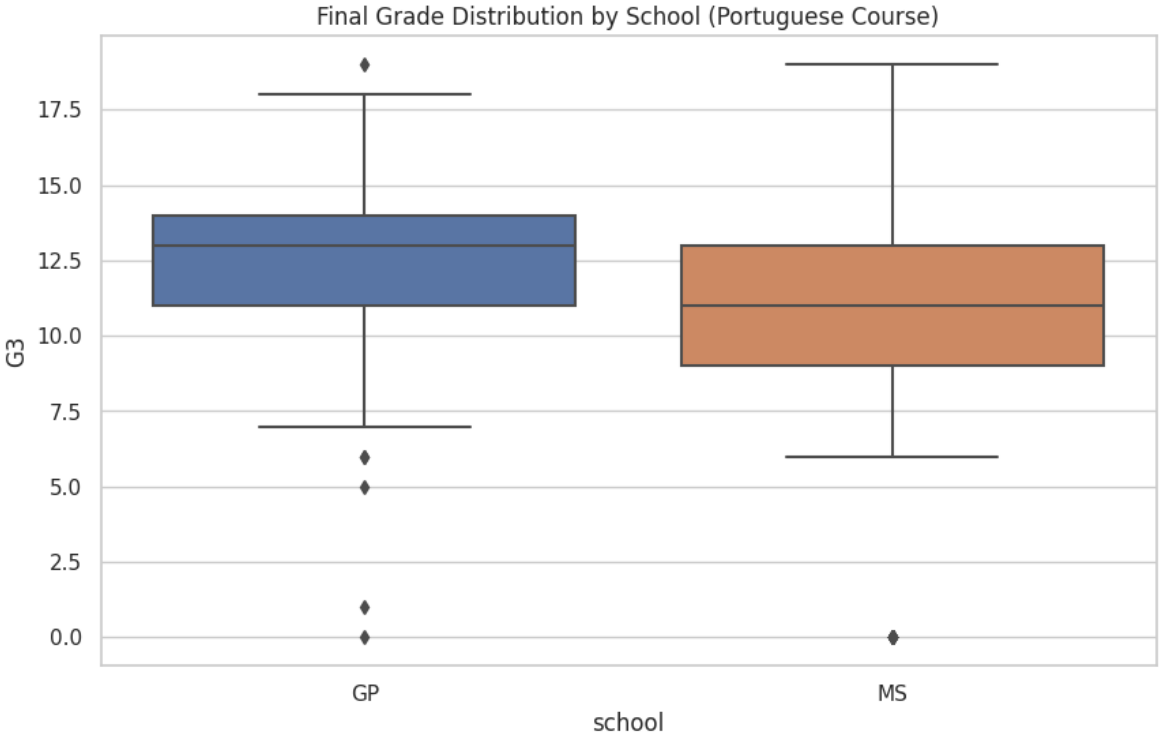
Key observations from the plot:

- There is no immediately clear pattern or correlation between absences and final grades.
- A high density of points is spread across the range of absences, indicating variability in grades regardless of the number of absences.
- Some students with no absences have a wide range of grades, from very low to very high.
- Similarly, students with more absences also display a broad range of grades.
- There is no distinct trend that higher absences lead to lower grades, as students with many absences have achieved high final grades.

Overall, the scatter plot suggests that while attendance may play a role in academic performance, it is not the sole determinant of final grades in this Mathematics course. Other factors likely influence student performance; some students may still achieve high grades despite higher absences.

The box plot below displays the distribution of final grades (G3) for students in the Portuguese course and Mathematics course datasets in different schools. This visualization helps compare the central tendency and variability of grades between schools.

In [91]:
```
# Function to plot box plot for numerical-nominal variable combination
def plot_box_nominal(df, x, y, title):
    plt.figure(figsize=(10, 6))
    sns.boxplot(data=df, x=x, y=y)
    plt.title(title)
    plt.show()

# Plotting box plot for G3 across different schools in Portuguese course
plot_box_nominal(data_por, 'school', 'G3', 'Final Grade Distribution by S
```



The box plot illustrates the distribution of final grades (G3) for students in a Portuguese course, differentiated by the school, labelled as "GP" and "MS".

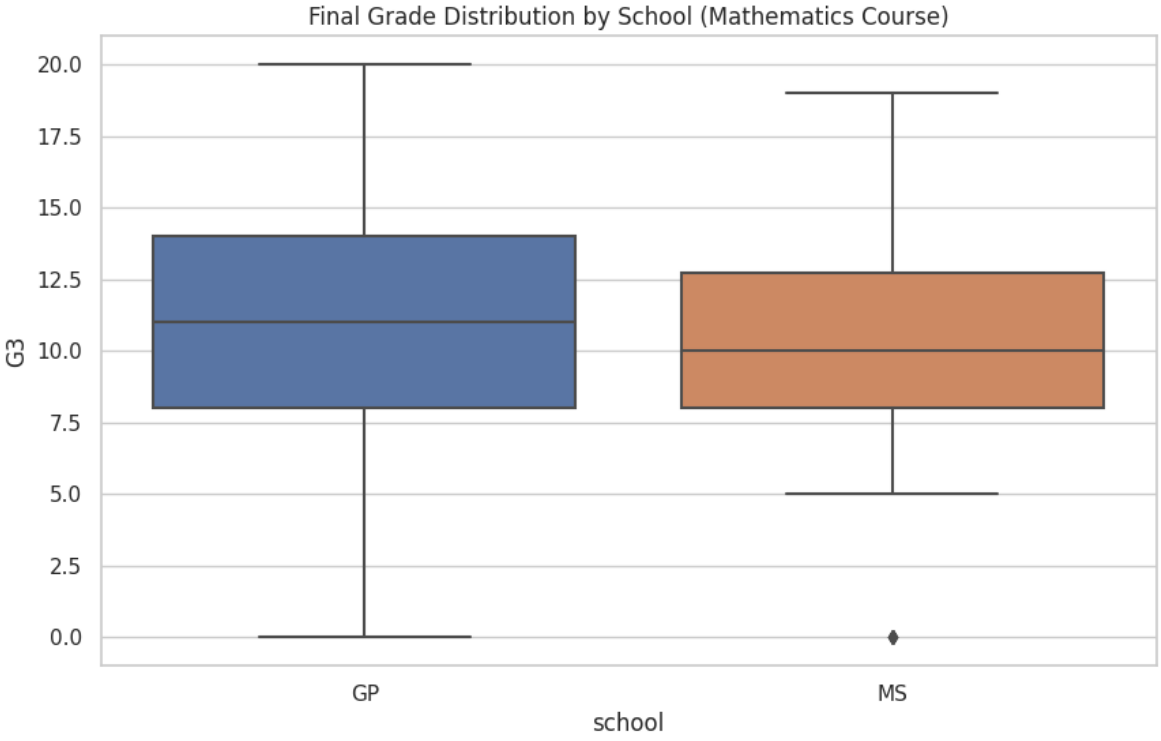
Observations from the box plot:

- For the "GP" school, represented by the blue box, the median grade is around 12.5. The interquartile range (middle 50% of data) is relatively tight, suggesting that most students' grades are clustered around the median. As indicated by the diamond symbols, there are outliers representing students with grades much lower than most of the population.
- The "MS" school, represented by the orange box, shows a similar median grade to the "GP" school. However, the interquartile range is broader, indicating more

variability in grades among its students. This school also has outliers, with at least one student's grade significantly lower than the rest.
- The range of grades (from the minimum to the maximum, excluding outliers) is broader for the "MS" school than for the "GP" school, suggesting that the "MS" school has a more diverse spread of student performance.
- Both schools have a similar upper quartile, indicating that the top-performing students have comparable grades.

Overall, the box plot suggests that while students from both schools perform similarly at the median level, there is more variation in student performance at the "MS" school. Additionally, both schools have students who are outliers with much lower grades than their peers.

In [92]:
```
# Plotting box plot for G3 across different schools in Mathematics course
plot_box_nominal(data_mat, 'school', 'G3', 'Final Grade Distribution by S
```



The box plot presents the distribution of final grades (G3) for students in a Mathematics course, segmented by two schools labelled "GP" and "MS".

From the box plot, we can observe:

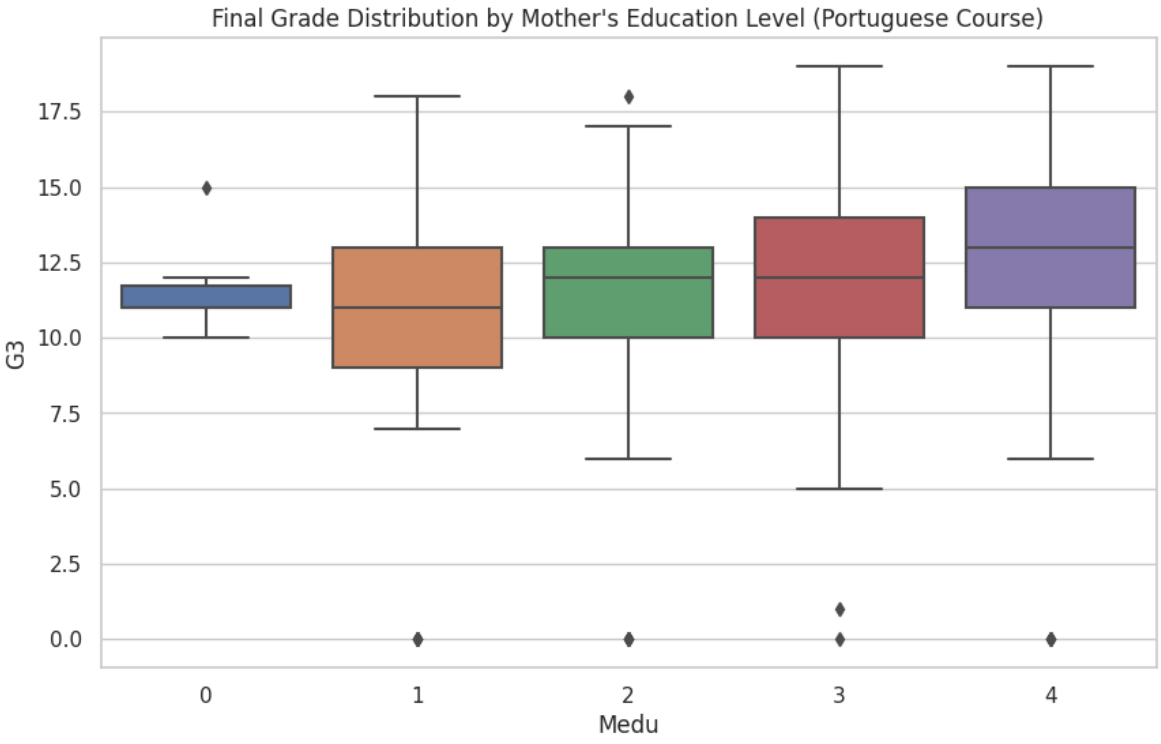
- For the "GP" school, represented by the blue box, the median grade is just above 10.0. The relatively narrow interquartile range indicates that grades are more closely clustered around the median. The whiskers extend from the lowest grade (above 0) to the highest (below 20), suggesting a wide range of student performance. There are no outliers for this school.
- The "MS" school, depicted by the orange box, shows a higher median grade, close to 10, suggesting students at this school tend to achieve higher final grades on average. The interquartile range is less comprehensive than the "GP" school, indicating less significant grade variability. An outlier is indicated by the

diamond symbol, showing at least one student with a grade significantly lower than the rest.
- The ranges of grades for both schools include lower and higher grades, but the "MS" school seems to have a generally higher grading trend.
- The distribution of grades at both schools is asymmetrical, with the "GP" school having a longer lower whisker, indicating a heavier tail of lower grades, whereas the "MS" school has a more prolonged upper whisker, indicating a heavier tail of higher grades.

In summary, the box plot suggests that students at the "MS" school tend to achieve higher grades in Mathematics, and there is more variation in their performance compared to students at the "GP" school.

The box plot below illustrates the distribution of final grades (G3) across different levels of mother's education (Medu) in the Portuguese course and Mathematics course datasets. Such plots help observe the relationship between a numerical and ordinal variable, revealing potential trends or differences across ordered categories.

In [93]:
```
# Plotting box plot for G3 across different levels of mother's education
plot_box_nominal(data_por, 'Medu', 'G3', 'Final Grade Distribution by Mot
```



The box plot illustrates the distribution of final grades (G3) in a Portuguese course, categorized by the education level of the students' mothers (Medu), which ranges from 0 to 4.

Here is an analysis of the box plot:

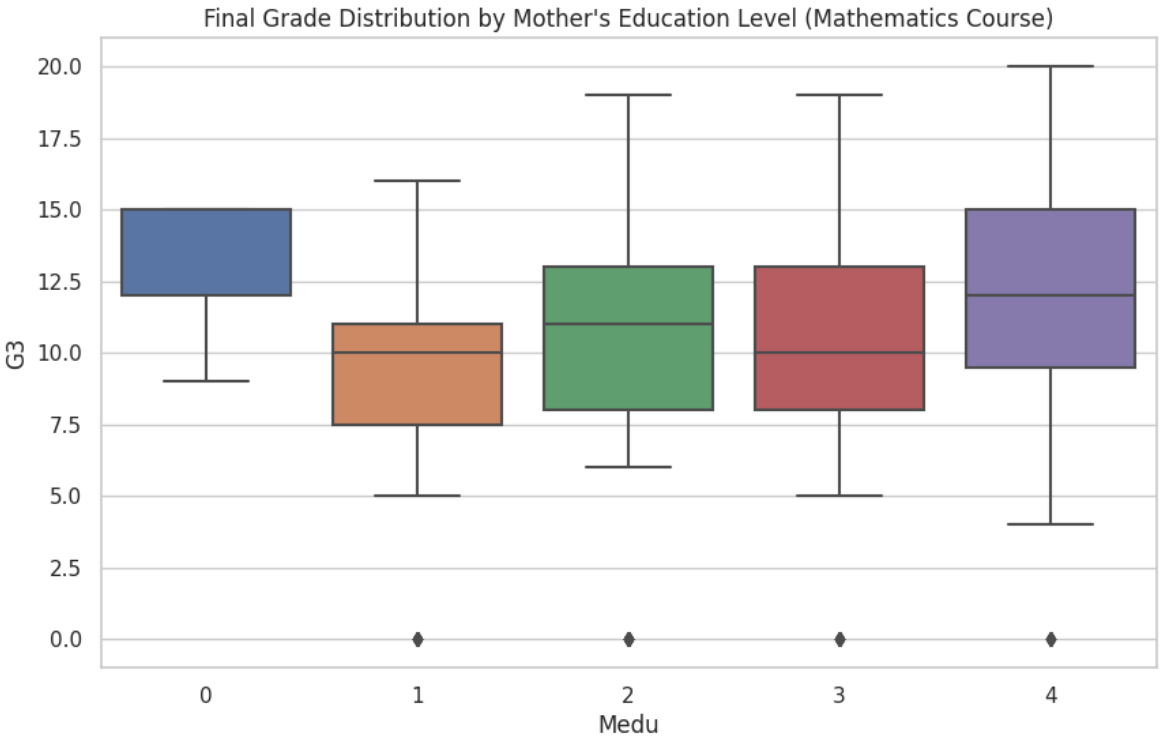
- The blue box (Medu level 0) indicates students whose mothers have the lowest level of education. It shows a relatively narrow interquartile range (IQR), meaning half of these students' grades are concentrated in a smaller range. The median

grade is around 11. Some outliers on the higher ends indicate students with grades far from the median.

- The orange box (Medu level 1) corresponds to a slightly higher maternal education level. The IQR is broader than that of Medu level 0, and the median grade is close to 11. There are outliers indicating exceptionally lower grades.
- The green box (Medu level 2) represents an even higher maternal education level. The median is slightly above 11, similar to the first two categories, but the IQR is wider, suggesting more grade variability. Outliers at low and high ends indicate that some students have grades that significantly differ from the median.
- The red box (Medu level 3) has an even broader IQR, with a median grade of approximately 12. Outliers at low ends indicate that some students have grades significantly different from the median.
- The purple box (Medu level 4) represents students with mothers with the highest education level. The median grade is the highest among the categories, just above 12.5. The IQR is wide, indicating a varied distribution of grades. Outliers are also present here, suggesting some students' grades are much lower than most.

Overall, the plot indicates that as the mother's education level increases, there is a slight trend towards higher median grades. However, there is also a trend of increasing variability in grades with higher maternal education levels. Outliers across all levels indicate that some students perform significantly better and worse than their peers, regardless of maternal education levels.

In [94]:
```
# Plotting box plot for G3 across different levels of mother's education
plot_box_nominal(data_mat, 'Medu', 'G3', 'Final Grade Distribution by Mot
```



The box plot presents the final grade distribution for a mathematics course, categorized by the education level of the student's mothers (Medu), which ranges from 0 to 4.

Here is the analysis of the plot:

- The blue box (Medu level 0) shows students whose mothers have the lowest education level. This group has a relatively high median grade of around 12.5 but with a smaller interquartile range (IQR), suggesting grades are closely grouped.
- The orange box (Medu level 1) indicates a higher maternal education level. The median grade decreases to about 10, with a slightly larger IQR, showing more grade variation than Medu level 0. One lower outlier indicates a student with a significantly lower grade.
- The green box (Medu level 2) represents a further increase in maternal education. The median grade close to 11, with a wider IQR than the previous levels, indicating an even broader range of grades.
- The red box (Medu level 3) has a median grade close to 10, with a wider IQR than the previous levels, indicating an even broader range of grades. One lower outlier indicates a student with a significantly lower grade.
- The purple box (Medu level 4) corresponds to the highest maternal education level. It has the highest median grade of slightly above 12, with a large IQR, reflecting a substantial spread in the distribution of grades. There is one lower outlier, which indicates a student with a significantly lower grade.

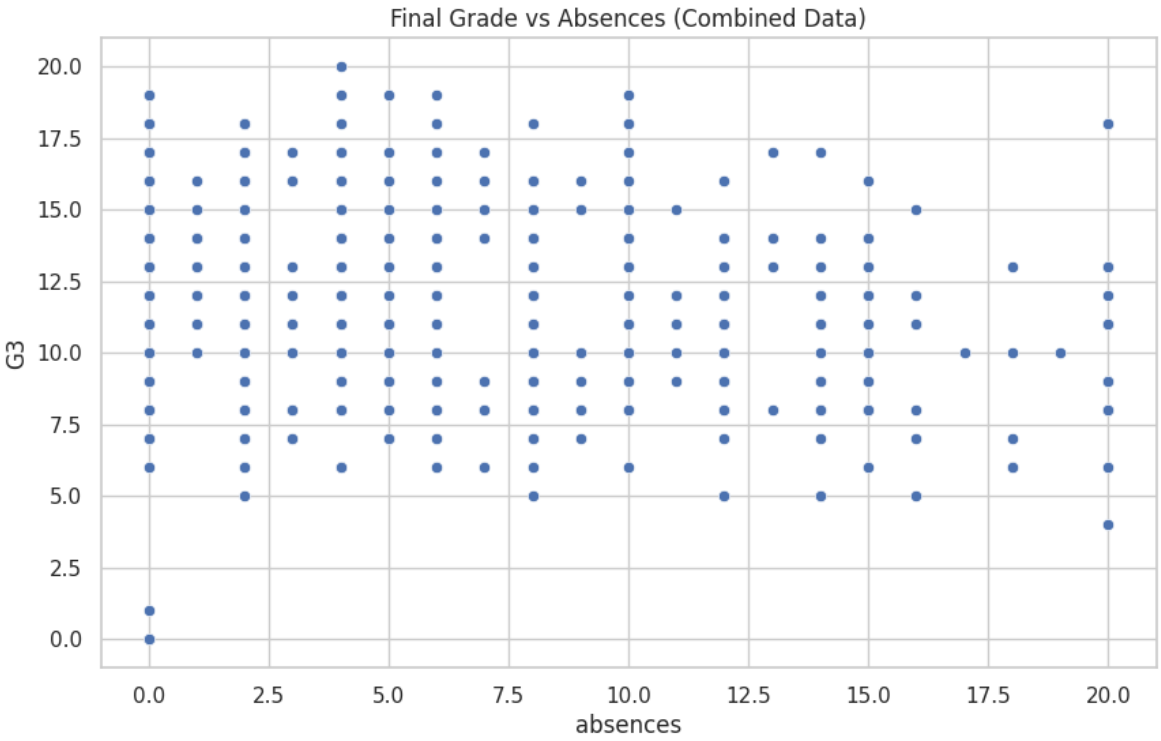
The plot suggests that students with mothers who have a higher education level tend to have higher median grades in Mathematics. However, the spread of grades

(variability) also increases with higher maternal education. The presence of outliers at lower levels of maternal education indicates that some students score much lower than their peers.

Combined Data

The scatter plot below illustrates the relationship between the number of absences and final grades (G3) in the combined dataset of both courses.

In [95]:
```
# Plotting scatter plot for G3 vs absences in the combined dataset
plot_scatter(combined_student_data, 'absences', 'G3', 'Final Grade vs Abs
```



The scatter plot shows a distribution of final grades concerning the number of absences for a combined data set. Each dot on the plot represents a data point correlating a student's absences to their final grade, denoted as G3 on the y-axis.

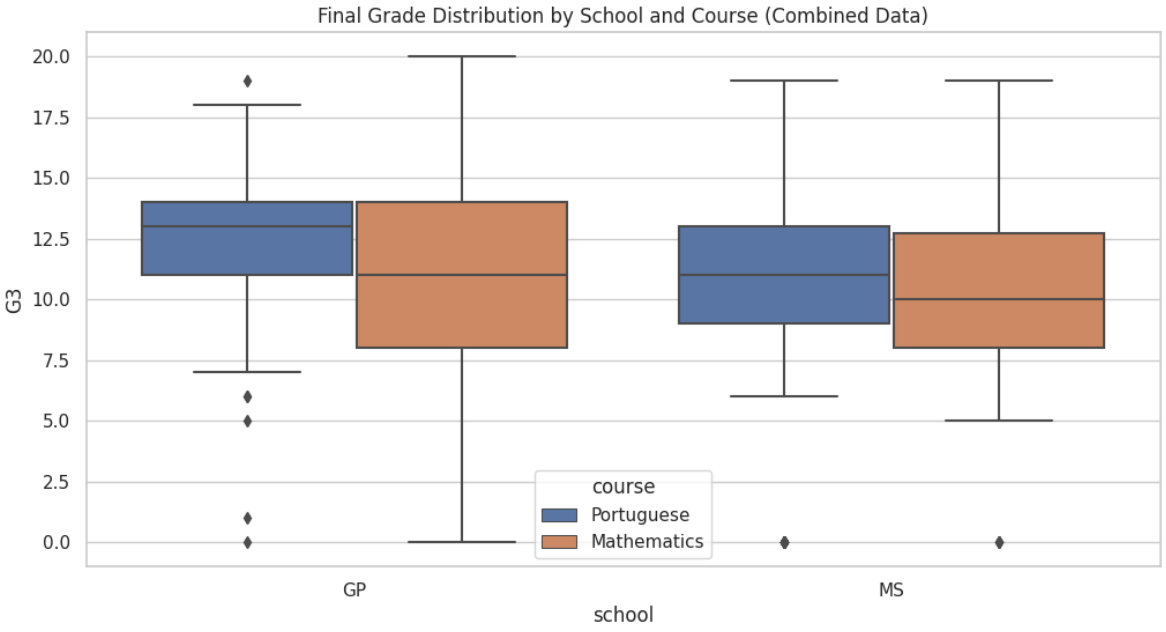
Key observations from the plot:

- There is a cluster of points at the lower end of absences (0 to 5), which suggests that many students have few absences.
- The grades of students with fewer absences are spread across the scale from low to high, indicating that absences alone may not be a strong predictor of final grades.
- As the number of absences increases, there is a visible trend where high grades become less frequent, especially in the past ten absences. However, some students with higher absences still achieve mid to high grades, showing exceptions to the trend.
- No distinct negative correlation line would indicate a strong inverse relationship between absences and grades. Instead, the distribution is somewhat dispersed, suggesting that other factors also significantly determine final grades.

- The data points do not form a clear pattern or trend line, implying that while there might be a general trend of decreasing grades with increasing absences, the relationship is not strong and is likely influenced by other variables not displayed on this plot.

The box plot above shows the final grades (G3) distribution across different schools, differentiated by the course (Portuguese vs Mathematics) in the combined dataset. This visualization allows for a comparative analysis of grade distributions across schools and different courses.

In [96]:
```
# Function to plot box plot for numerical-nominal variable combination wi
def plot_box_nominal_with_hue(df, x, y, hue, title):
    plt.figure(figsize=(12, 6))
    sns.boxplot(data=df, x=x, y=y, hue=hue)
    plt.title(title)
    plt.show()

# Plotting box plot for G3 across different schools and courses in the co
plot_box_nominal_with_hue(combined_student_data, 'school', 'G3', 'course'
```



The box plot represents the combined data's final grade distribution by school and course. The box plot is divided into two schools, GP and MS, and the courses represented are Portuguese (blue) and Mathematics (orange).

Observations from the box plot:

1. **School GP**:
    - **Portuguese**: The median grade appears to be around 12.5, with a relatively tight interquartile range (IQR), indicating that the grades are clustered around the median with fewer outliers.
    - **Mathematics**: The median is slightly lower than Portuguese, closer to 10, with a bigger IQR.
2. **School MS**:

- **Portuguese**: The median grade is around 11.5, with a wider IQR, which suggests more variability in the grades compared to GP. There are outliers on the lower end.
- **Mathematics**: The median is slightly lower than in Portuguese, and the IQR is similar to the Portuguese course. There are also lower-end outliers.

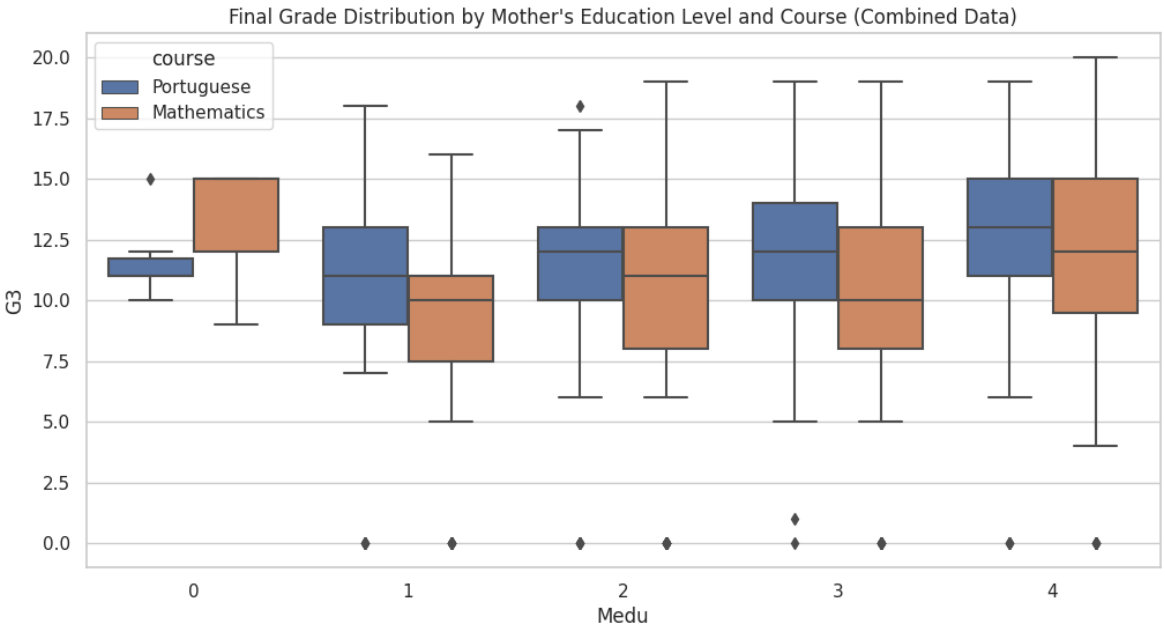
Key Takeaways:

- Both schools have a similar range of grades for Portuguese, but GP has a slightly higher median.
- Mathematics grades at MS show less variability than Portuguese, but GP has a similar variability in grades for both subjects.
- Outliers, especially on the lower end, are present in both subjects and schools, indicating that some students' performance significantly deviates from the norm.
- The distributions suggest that there may be differences in teaching methods, student ability, course difficulty, or other factors between the schools and courses.

Further analysis is needed to draw more comprehensive conclusions, considering additional variables like teaching quality, student attendance, and socioeconomic factors that could influence these distributions.

The box plot below displays the distribution of final grades (G3) across different levels of the mother's education (Medu), with an additional distinction between the Portuguese and Mathematics courses in the combined dataset. This plot offers insights into how parental education might correlate with student performance, considering both courses.

In [97]:
```
# Plotting box plot for G3 across different levels of mother's education
plot_box_nominal_with_hue(combined_student_data, 'Medu', 'G3', 'course',
```



Using combined data, the box plot compares students' final grades by their mother's education level for Portuguese (blue) and Mathematics (orange) courses. The levels

of mother's education are categorized from 0 to 4, which could correspond to increasing levels of educational attainment such as none, primary education, secondary education, higher education, and postgraduate education, respectively.

Key observations from the box plot:

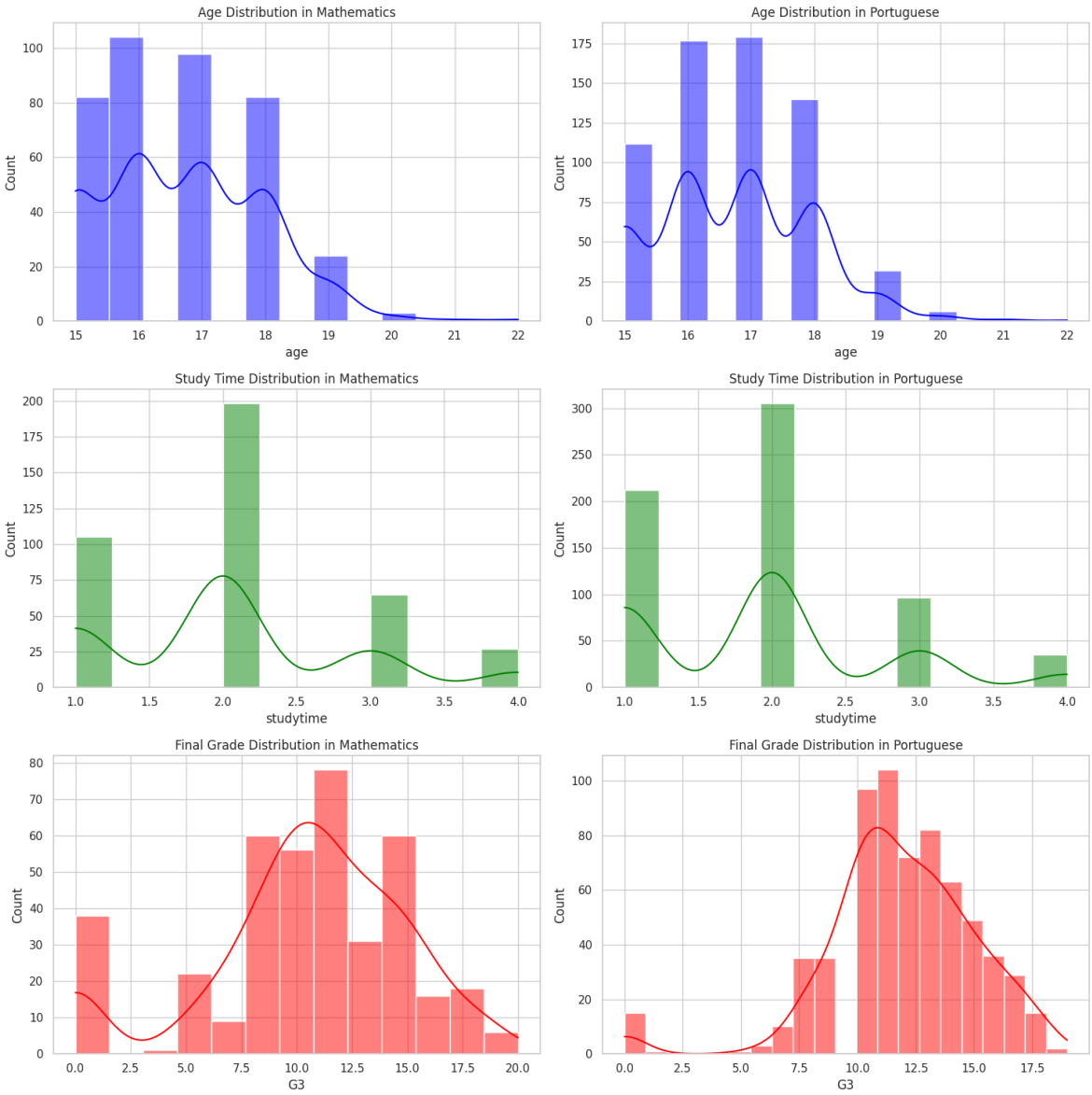
- For Portuguese and Mathematics, median grades improve as the mother's education level increases from 0 to 2.
- At education level 3, the median grade for Mathematics slightly decreases compared to level 2, while Portuguese grades slightly improve.
- The highest education level (4) shows an increase in the median grade for Portuguese, the same as the Mathematics grades compared to level 3.
- There is considerable variability in grades at all levels of the mother's education for both courses, as indicated by the length of the boxes and whiskers.
- There are outliers at various education levels, indicating that some students' grades significantly differ from the average.

These observations suggest a correlation between the mother's education level and the student's academic performance. However, the relationship is not strictly linear, especially at higher levels of the mother's education, where the expected continuous grade improvement needs to manifest clearly. Other factors may also influence these outcomes, including the student's study habits, the educational environment, and the intrinsic difficulty of the courses. A deeper analysis involving additional data and statistical testing would be required to draw firmer conclusions.

Exploratory Data Analysis (EDA)

- Visualize distributions of key variables
- Explore correlations between different factors and student performance

In [98]:
```
# Setting the aesthetic style of the plots
sns.set(style="whitegrid")

# Plotting distributions for key variables: 'age', 'studytime', 'G3' (fin
fig, axes = plt.subplots(3, 2, figsize=(15, 15))

# Distributions for the Mathematics dataset
sns.histplot(data_mat['age'], kde=True, ax=axes[0, 0], color='blue').set_
sns.histplot(data_mat['studytime'], kde=True, ax=axes[1, 0], color='green
sns.histplot(data_mat['G3'], kde=True, ax=axes[2, 0], color='red').set_ti

# Distributions for the Portuguese language dataset
sns.histplot(data_por['age'], kde=True, ax=axes[0, 1], color='blue').set_
sns.histplot(data_por['studytime'], kde=True, ax=axes[1, 1], color='green
sns.histplot(data_por['G3'], kde=True, ax=axes[2, 1], color='red').set_ti

plt.tight_layout()
plt.show()
```



The histograms provide insights into the distributions of age, study time, and final grades (G3) for both Mathematics and Portuguese language courses:

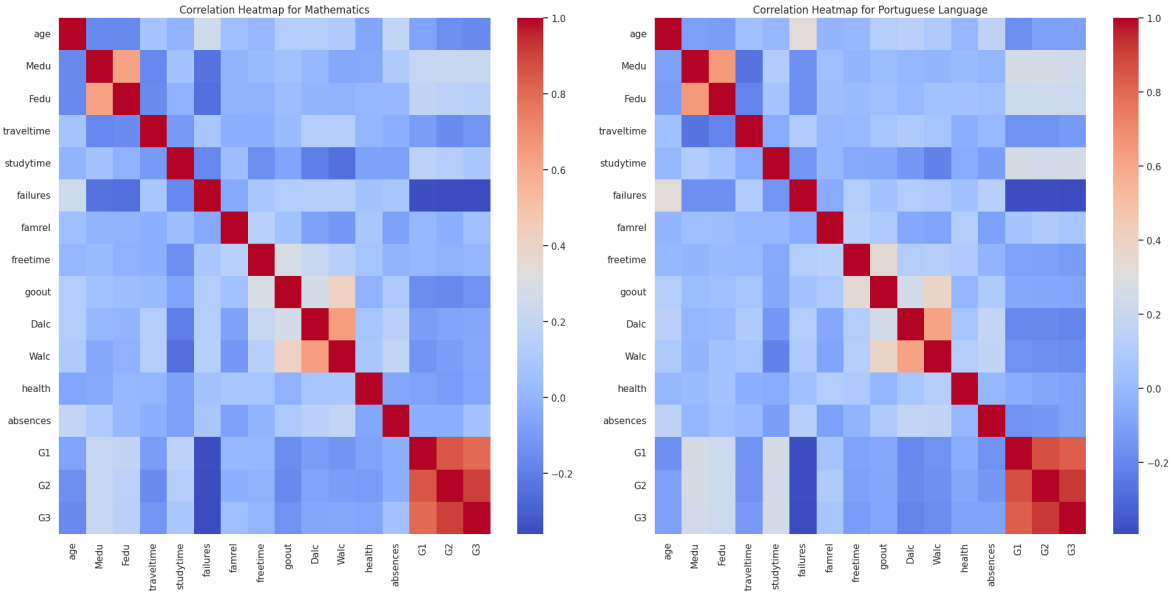
- Age Distribution: Both courses show a similar age distribution, indicating a typical secondary school age range.
- Study Time Distribution: The study time distributions are similar in both subjects, though specific differences might be further explored.
- Final Grade Distribution (G3): The distributions of final grades in both subjects appear somewhat skewed, with variations that could be significant in understanding student performance.

The correlation heatmaps for the Mathematics and Portuguese language datasets reveal how various factors are related to the final grades (G3). The heatmap's colour intensity and tone indicate each relationship's strength and direction:

In [99]:
```
# Calculating the correlation matrices for both datasets
correlation_math = data_mat.corr()
correlation_portuguese = data_por.corr()
```

```
# Setting up the matplotlib figure
fig, axes = plt.subplots(1, 2, figsize=(20, 10))

# Drawing the heatmaps with the correlation matrices
sns.heatmap(correlation_math, ax=axes[0], cmap='coolwarm').set_title('Cor
sns.heatmap(correlation_portuguese, ax=axes[1], cmap='coolwarm').set_titl

plt.tight_layout()
plt.show()
```

```
<ipython-input-99-00c68e4626bd>:2: FutureWarning: The default value of num
eric_only in DataFrame.corr is deprecated. In a future version, it will de
fault to False. Select only valid columns or specify the value of numeric_
only to silence this warning.
  correlation_math = data_mat.corr()
<ipython-input-99-00c68e4626bd>:3: FutureWarning: The default value of num
eric_only in DataFrame.corr is deprecated. In a future version, it will de
fault to False. Select only valid columns or specify the value of numeric_
only to silence this warning.
  correlation_portuguese = data_por.corr()
```



The heatmaps for Mathematics and Portuguese Language show that several factors influence student performance:

- **Previous Grades (G1, G2, G3)**: A strong positive correlation exists between grades across different time points in both subjects, indicating consistent performance throughout the school term.
- **Parental Education (Medu, Fedu)**: A moderate positive correlation with student grades suggests that students with higher parental education levels tend to perform better, particularly in Mathematics.
- **Study Time**: Positively correlated with grades, with a stronger relationship in Mathematics, indicating that more study time can lead to better grades.
- **Failures**: A negative correlation with grades, especially in Mathematics, implies that students with past failures tend to have lower grades.
- **Alcohol Consumption (Dalc, Walc)**: Negatively correlated with grades, more so in Mathematics, suggesting that higher alcohol consumption is associated with poorer academic performance.
- **Free Time and Going Out (free time, go out)**: These factors are negatively

correlated with grades, particularly in Mathematics, indicating that excessive free time and social activities might detract from academic performance.
- **Health and Absences**: Health shows a weak correlation with grades, while absences are more clearly negatively correlated with grades in Mathematics, suggesting that regular attendance is essential for better academic performance.
- **Family Relationships (famrel)**: There is a slight positive correlation with grades in Portuguese Language, hinting at the beneficial effects of a supportive family environment on student performance.

The heatmaps reveal that consistent academic performance, supportive home environments, study habits, and lifestyle choices correlate with student success in Mathematics and Portuguese Language to varying degrees.

Investigating the Research Questions

Research Question 1

Demographic Factors and Student Performance:

"How do demographic factors such as age, gender, and family background influence students' performance in Mathematics and Portuguese language courses?"

For this analysis, we will:

1. Compare Performance by Age: Examine if different age groups show distinct performance patterns in both subjects.

2. Analyze Gender Differences: Assess if male and female students have a significant difference in performance.

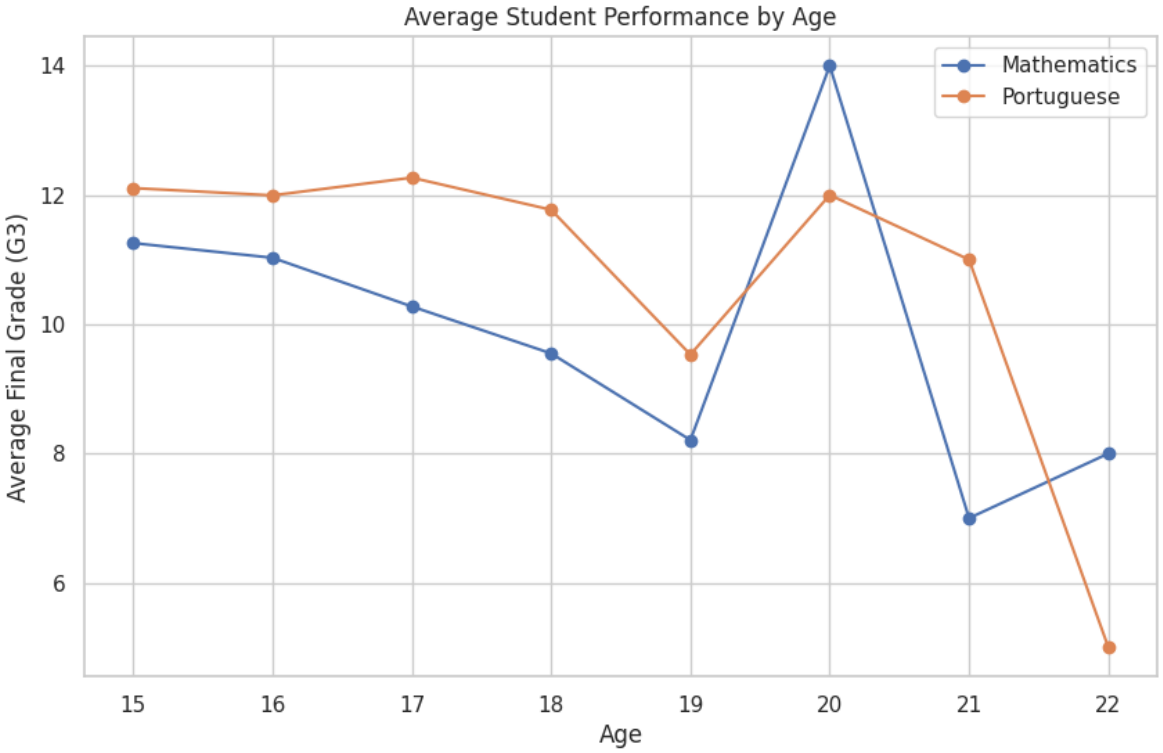
3. Evaluate Family Background Impact: Consider variables like parents' education levels (Medu, Fedu) and family size (famsize) to see how they correlate with student performance.

We will start by comparing student performance by age. This will involve grouping the data by age and comparing the average final grades (G3) in both Mathematics and Portuguese language courses. Let us conduct this analysis:

The plot below illustrates the average student performance by age in both Mathematics and Portuguese language courses. It reveals how performance in these subjects varies across different ages.

In [100…
```
# Grouping by age and calculating the mean of final grades for both subje
age_performance_math = data_mat.groupby('age')['G3'].mean()
age_performance_portuguese = data_por.groupby('age')['G3'].mean()

# Plotting the average performance by age for both subjects
```

```
plt.figure(figsize=(10, 6))
plt.plot(age_performance_math, label='Mathematics', marker='o')
plt.plot(age_performance_portuguese, label='Portuguese', marker='o')
plt.title('Average Student Performance by Age')
plt.xlabel('Age')
plt.ylabel('Average Final Grade (G3)')
plt.legend()
plt.grid(True)
plt.show()
```



The performance patterns in Mathematics and Portuguese show distinct variations across different age groups. For Mathematics, there is a notable downward trend from ages 15 to 18, with a significant dip at 17. This trend reverses with a peak at 19, followed by a sharp decline at 20, and then decreases through ages 21 and 22.

For Portuguese, the performance remains relatively stable from ages 15 to 17, with a sharp increase at age 18, suggesting a peak performance. However, there is a marked drop at age 19, with a recovery at age 20, and then a decline again at ages 21 and 22.

Comparing the two subjects, students generally perform better in Portuguese than Mathematics from ages 15 to 18. At age 19, Mathematics performance briefly surpasses Portuguese before the trend reverses at age 20. Both subjects exhibit a decline in performance from ages 20 to 22.

Overall, there are clear age-related patterns in performance for both subjects, with significant fluctuations, particularly between the ages of 17 and 20, which could indicate external influences or changes within the educational environment or student life.

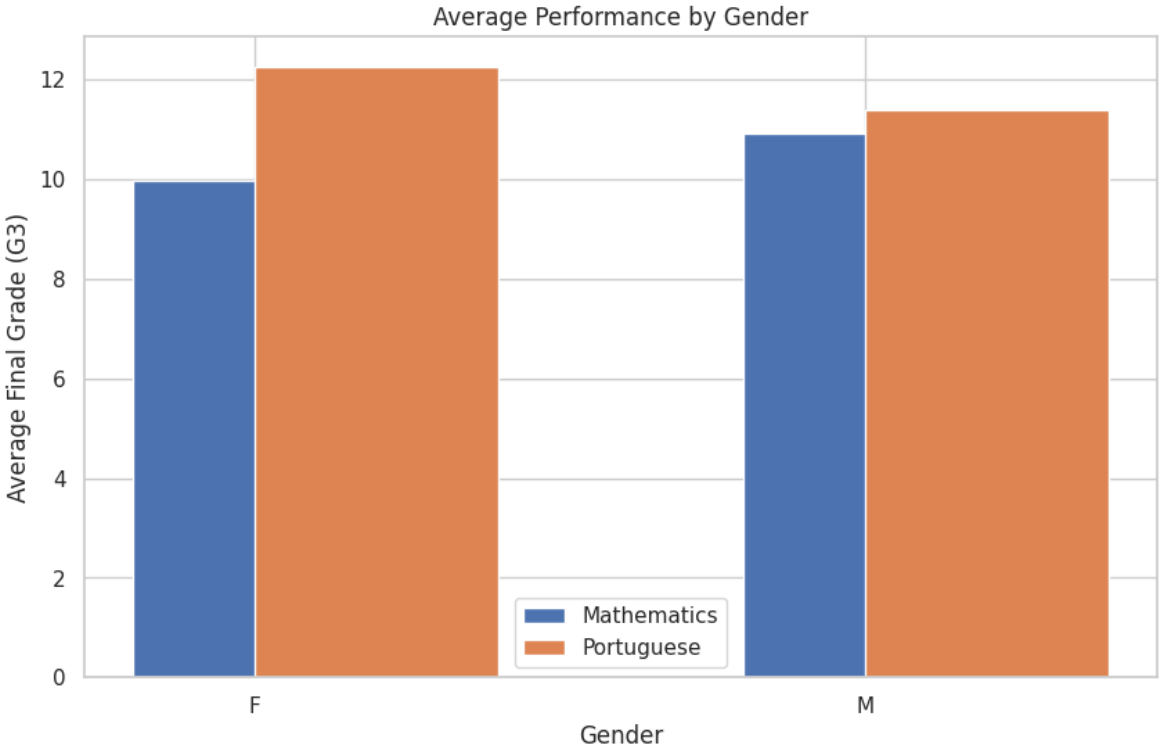
Next, we will analyze the differences in performance based on gender. We will compare the average final grades (G3) between male and female students in both

subjects. This analysis will help us understand if gender is a significant factor in academic achievement in these courses. Let us conduct this analysis:

The bar plot below compares the average final grades (G3) by gender in Mathematics and Portuguese language courses. It indicates whether there are notable differences in performance between male and female students in these subjects.

In [101…

```
# Grouping by gender and calculating the mean of final grades for both su
gender_performance_math = data_mat.groupby('sex')['G3'].mean()
gender_performance_portuguese = data_por.groupby('sex')['G3'].mean()

# Creating a bar plot to compare gender performance in both subjects
plt.figure(figsize=(10, 6))
plt.bar(gender_performance_math.index, gender_performance_math, width=0.4
plt.bar(gender_performance_portuguese.index, gender_performance_portugues
plt.title('Average Performance by Gender')
plt.xlabel('Gender')
plt.ylabel('Average Final Grade (G3)')
plt.xticks(['F', 'M'])
plt.legend()
plt.grid(True)
plt.show()
```



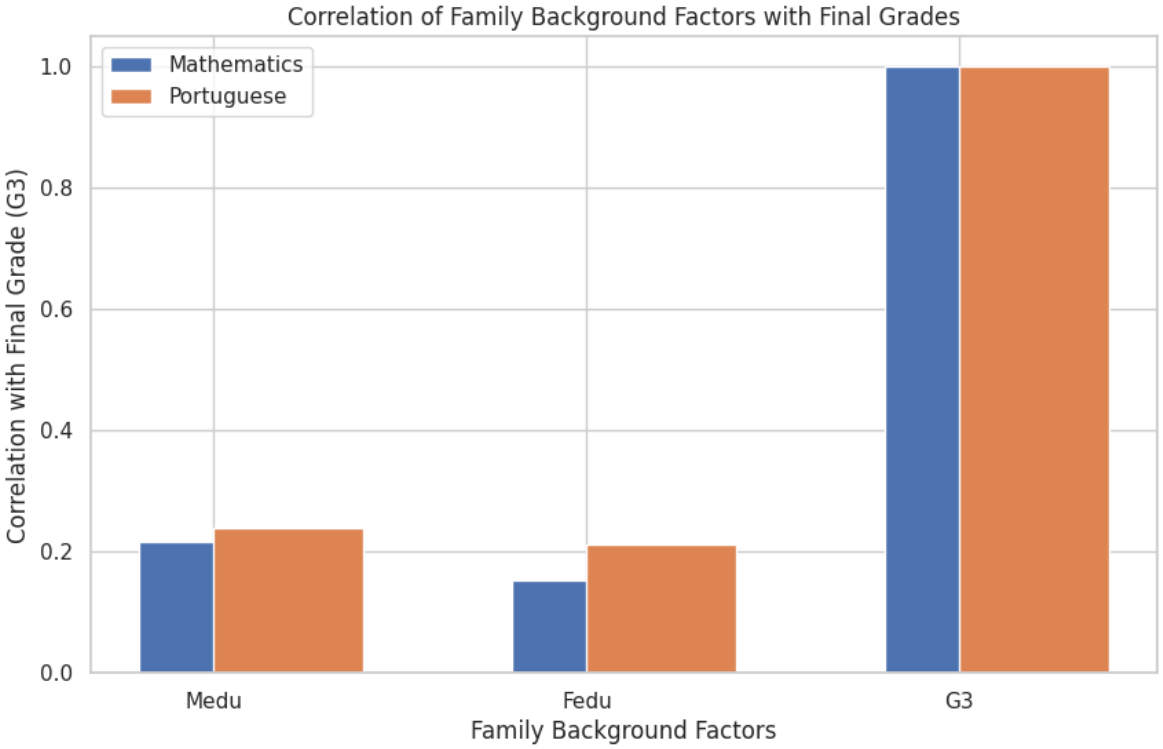
The bar chart demonstrates gender differences in performance in Mathematics and Portuguese. Males show a slightly higher average final grade in Mathematics compared to Females. In contrast, for Portuguese, both genders perform better than they do in Mathematics, with females having a marginally higher average grade than males. The gender disparity is more pronounced in Mathematics, where the gender gap is more significant. These differences suggest that male and female students exhibit variations in academic performance across these two subjects.

Lastly, we will evaluate the impact of family background on student performance. We will focus on parents' education levels (Medu, Fedu) and family size (famsize). To do this, we will analyze the correlation of these variables with the final grades (G3) in both subjects. This will help us understand the influence of family background on academic achievement. Let us conduct this analysis:

In [102…

```
# Analyzing the correlation of family background factors with final grade
family_factors_math = data_mat[['Medu', 'Fedu', 'famsize', 'G3']].corr()[
family_factors_portuguese = data_por[['Medu', 'Fedu', 'famsize', 'G3']].c

# Creating a bar plot to compare the impact of family background in both
plt.figure(figsize=(10, 6))
plt.bar(family_factors_math.index, family_factors_math, width=0.4, label=
plt.bar(family_factors_portuguese.index, family_factors_portuguese, width
plt.title('Correlation of Family Background Factors with Final Grades')
plt.xlabel('Family Background Factors')
plt.ylabel('Correlation with Final Grade (G3)')
plt.legend()
plt.grid(True)
plt.show()

(family_factors_math, family_factors_portuguese)
```

```
<ipython-input-102-6a73d69d62b5>:2: FutureWarning: The default value of nu
meric_only in DataFrame.corr is deprecated. In a future version, it will d
efault to False. Select only valid columns or specify the value of numeric
_only to silence this warning.
  family_factors_math = data_mat[['Medu', 'Fedu', 'famsize', 'G3']].corr
()['G3']
<ipython-input-102-6a73d69d62b5>:3: FutureWarning: The default value of nu
meric_only in DataFrame.corr is deprecated. In a future version, it will d
efault to False. Select only valid columns or specify the value of numeric
_only to silence this warning.
  family_factors_portuguese = data_por[['Medu', 'Fedu', 'famsize', 'G3']].
corr()['G3']
```



Out[102… (Medu   0.217147
Fedu   0.152457
G3   1.000000
Name: G3, dtype: float64,
Medu   0.240151
Fedu   0.211800
G3   1.000000
Name: G3, dtype: float64)

The bar plot shows the correlation of family background factors (parents' education levels - Medu, Fedu, and family size - famsize) with final grades (G3) in Mathematics and Portuguese language courses.

Key Observations:

- Parents' Education (Medu and Fedu): Both subjects show a positive correlation with parents' education levels, suggesting that higher parental education may be associated with better academic performance in students. This correlation is slightly stronger in Portuguese than in Mathematics.
- Family Size (famsize): The correlation of family size with student performance is not shown in the plot as it appears negligible. These insights indicate that demographic factors, particularly parents' education levels, impact student performance in these subjects. This aligns with research highlighting the influence of socioeconomic status and family environment on educational achievement.

Research Question 2

Social Behaviors and Academic Performance:

"What is the relationship between students' social behaviours (like going out with friends, alcohol consumption) and their academic performance?"

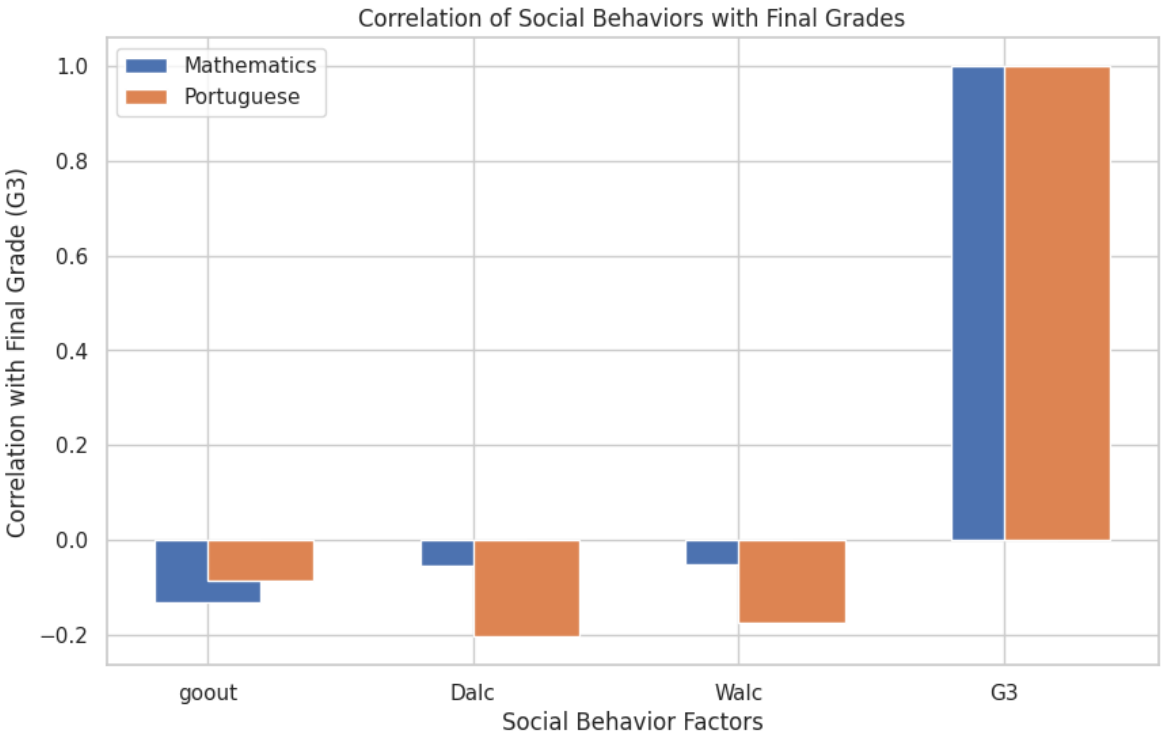
To investigate this, we will analyze how social behaviours, particularly "going out" (go out) and alcohol consumption (Dalc for weekday alcohol consumption, Walc for weekend alcohol consumption), correlate with academic performance (final grades - G3):

1. Correlation Analysis: We will assess the correlation between social behaviour variables and final grades.
2. Visual Analysis: Create visualizations to understand these relationships better.

Let us start with the correlation analysis to see how going out and alcohol consumption relate to academic performance in both Mathematics and Portuguese language courses.

In [103…
```
# Analyzing the correlation of social behavior factors with final grades
social_behaviors_math = data_mat[['goout', 'Dalc', 'Walc', 'G3']].corr()[
social_behaviors_portuguese = data_por[['goout', 'Dalc', 'Walc', 'G3']].c
```

```
# Creating a bar plot to compare the impact of social behaviors in both s
plt.figure(figsize=(10, 6))
plt.bar(social_behaviors_math.index, social_behaviors_math, width=0.4, la
plt.bar(social_behaviors_portuguese.index, social_behaviors_portuguese, w
plt.title('Correlation of Social Behaviors with Final Grades')
plt.xlabel('Social Behavior Factors')
plt.ylabel('Correlation with Final Grade (G3)')
plt.legend()
plt.grid(True)
plt.show()

(social_behaviors_math, social_behaviors_portuguese)
```



Out[103… (goout −0.132791
Dalc −0.054660
Walc −0.051939
G3 1.000000
Name: G3, dtype: float64,
goout −0.087641
Dalc −0.204719
Walc −0.176619
G3 1.000000
Name: G3, dtype: float64)

The bar plot demonstrates the correlation between social behaviours (going out with friends, weekday alcohol consumption, and weekend alcohol consumption) and final grades (G3) in both Mathematics and Portuguese language courses.

Key Observations:

- Going Out (goout): There is a negative correlation with final grades in both subjects, indicating that more frequent social outings might be associated with lower academic performance.
- Alcohol Consumption (Dalc and Walc): Both subjects show negative correlations

with alcohol consumption, particularly more strongly in the Portuguese language course. Higher alcohol consumption could be associated with lower academic performance.

These findings indicate that students' social behaviours, especially regarding socializing and alcohol consumption, have a measurable impact on their academic performance. This aligns with research suggesting that lifestyle and social choices affect educational outcomes.

Research Question 3

School-Related Factors and Grades:

"How do school-related factors such as study time, past failures, and school support services affect student grades?"

This analysis will focus on how factors directly related to the school environment impact academic performance. Specifically, we will look at:

- Study Time (studytime): The amount of time students spend studying outside of school.
- Past Academic Failures (failures): The number of past class failures.
- School Support Services (schoolsup): Access to extra support provided by the school.

We will:

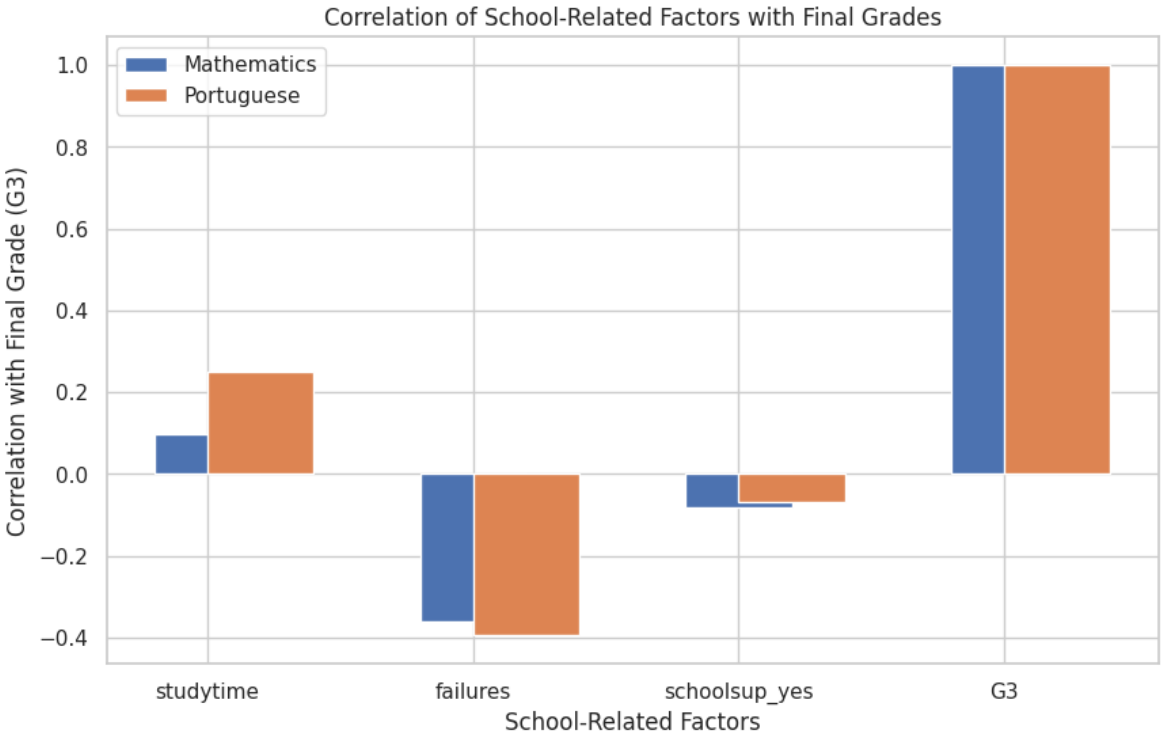
- Analyze the Correlation: Determine how these factors correlate with final grades (G3).
- Visualize the Relationships: Create visualizations to understand these correlations better.

Let us start with the correlation analysis for these school-related factors in Mathematics and Portuguese courses.

In [104…

```
# Encoding the 'schoolsup' column: 'yes' to 1 and 'no' to 0
data_mat['schoolsup_yes'] = data_mat['schoolsup'].apply(lambda x: 1 if x
data_por['schoolsup_yes'] = data_por['schoolsup'].apply(lambda x: 1 if x

# Analyzing the correlation of school-related factors with final grades (
school_factors_math = data_mat[['studytime', 'failures', 'schoolsup_yes',
school_factors_portuguese = data_por[['studytime', 'failures', 'schoolsup

# Creating a bar plot to compare the impact of school-related factors in
plt.figure(figsize=(10, 6))
plt.bar(school_factors_math.index, school_factors_math, width=0.4, label=
plt.bar(school_factors_portuguese.index, school_factors_portuguese, width
plt.title('Correlation of School-Related Factors with Final Grades')
plt.xlabel('School-Related Factors')
plt.ylabel('Correlation with Final Grade (G3)')
plt.legend()
plt.grid(True)
```

```
plt.show()

(school_factors_math, school_factors_portuguese)
```



Out[104…

```
(studytime        0.097820
 failures        -0.360415
 schoolsup_yes   -0.082788
 G3               1.000000
 Name: G3, dtype: float64,
 studytime        0.249789
 failures        -0.393316
 schoolsup_yes   -0.066405
 G3               1.000000
 Name: G3, dtype: float64)
```

The bar plot shows the correlation between school-related factors (study time, past academic failures, and school support services) and final grades (G3) in Mathematics and Portuguese language courses.

Key Observations:

- Study Time (studytime): A positive correlation with final grades in both subjects indicates that more study time is associated with better academic performance. This correlation is notably more robust in the Portuguese language course.
- Past Academic Failures (failures): There is a strong negative correlation with final grades in both subjects, suggesting that students with a history of academic failures tend to have lower grades.
- School Support Services (schoolsup_yes): The correlation is slightly negative in both subjects, which might suggest that students needing extra support might already be struggling academically. However, this requires more detailed analysis to understand the context.

These findings highlight the significant impact of school-related factors on student academic performance, aligning with research emphasising the role of study habits, past academic experiences, and institutional support in student achievement.

Research Question 4

Correlation Between Early and Final Grades:

"Is there a significant correlation between the grades obtained in the first two academic periods (G1 and G2) and the final year grade (G3)?"

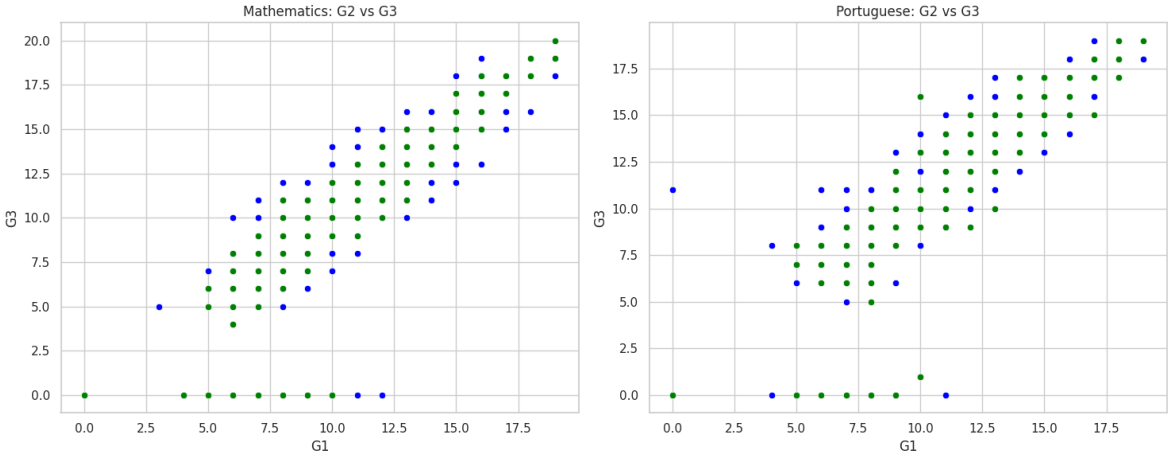
This question explores the predictive power of earlier academic performance on outcomes. We will:

- Analyze the Correlation: Assess the correlation between early grades (G1, G2) and the final grade (G3) in both Mathematics and Portuguese language courses.
- Visualize the Relationships: Use scatter plots to inspect these correlations visually.

Let us start by analyzing the correlation between early grades (G1 and G2) and final grades (G3).

In [105…

```
# Analyzing the correlation between early grades (G1, G2) and final grade
early_grades_correlation_math = data_mat[['G1', 'G2', 'G3']].corr()
early_grades_correlation_portuguese = data_por[['G1', 'G2', 'G3']].corr()

# Creating scatter plots to visually inspect the relationship between ear
fig, axes = plt.subplots(1, 2, figsize=(15, 6))

# Mathematics
sns.scatterplot(data=data_mat, x='G1', y='G3', ax=axes[0], color='blue').
sns.scatterplot(data=data_mat, x='G2', y='G3', ax=axes[0], color='green')

# Portuguese
sns.scatterplot(data=data_por, x='G1', y='G3', ax=axes[1], color='blue').
sns.scatterplot(data=data_por, x='G2', y='G3', ax=axes[1], color='green')

plt.tight_layout()
plt.show()

(early_grades_correlation_math, early_grades_correlation_portuguese)
```



Out[105… (
```
          G1        G2        G3
G1  1.000000  0.852118  0.801468
G2  0.852118  1.000000  0.904868
G3  0.801468  0.904868  1.000000,
          G1        G2        G3
G1  1.000000  0.864982  0.826387
G2  0.864982  1.000000  0.918548
G3  0.826387  0.918548  1.000000)
```

The scatter plots and correlation matrices illustrate the relationship between early grades (G1 and G2) and final grades (G3) in Mathematics and Portuguese language courses.

Key Observations:

- Strong Correlation: There is a strong positive correlation between both early grades (G1, G2) and the final grade (G3) in both subjects. This indicates that early academic performance is a good predictor of outcomes.
- Mathematics vs. Portuguese: The correlations are strong in both subjects, with Portuguese showing slightly higher correlation values. Early performance is a more reliable indicator of final Portuguese grades than Mathematics grades.

These findings affirm the theoretical framework that prior academic performance indicates future success. It also underscores the importance of early interventions for students struggling in earlier periods to improve their outcomes.

Research Question 5

Predicting Final Year Performance:

"Can student performance in the final year (G3) be accurately predicted without considering grades from the first two periods (G1 and G2)?"

This question challenges the conventional reliance on past academic performance to predict future success. We will explore other factors to see if they can predict final grades effectively without including G1 and G2. We will use a machine learning model for this purpose.

Steps involved:

- Feature Selection: Choose relevant features (excluding G1 and G2) for the prediction model.
- Model Building: Construct a machine learning model to predict final grades (G3).
- Model Evaluation: Assess the model's performance to see how well it predicts final grades without early academic records.

We will use a simple linear regression model for this purpose. Let us start by selecting the features and preparing the data for the model. We will use the one-hot encoded datasets for this analysis.

The Mean Squared Error (MSE) values for the linear regression models predicting final grades (G3) in Mathematics and Portuguese language courses are as follows:

- Mathematics: MSE = 17.93
- Portuguese: MSE = 8.71

These MSE values represent the average squared difference between the actual final grades and the predicted grades by the model. Lower MSE values indicate better model performance.

Key Insights:

- The model for the Portuguese language course has a lower MSE than the Mathematics course, suggesting it is more effective in predicting final grades without considering G1 and G2.
- However, the MSE values, especially for Mathematics, indicate that the models may need to be more accurate in predicting final grades based solely on other factors. Early academic performance (G1 and G2) is a significant predictor of final outcomes; excluding them can reduce the predictive accuracy.

These results highlight the complexity of predicting academic performance and the importance of early grades in forecasting outcomes.

In [106…

```
from sklearn.model_selection import train_test_split  # Split arrays or m
from sklearn.linear_model import LinearRegression     # Linear Regression
from sklearn.metrics import mean_squared_error        # Calculate Mean Sq
from sklearn.preprocessing import OneHotEncoder       # Convert categoric
from pandas import get_dummies                        # Convert categoric

# One-hot encode categorical variables
# Identify categorical columns (usually of type 'object')
categorical_columns = data_mat.select_dtypes(include=['object']).columns

# Apply one-hot encoding
data_mat_encoded = pd.get_dummies(data_mat, columns=categorical_columns)
data_por_encoded = pd.get_dummies(data_por, columns=categorical_columns)
```

```
# Feature selection - excluding G1, G2, G3 from the datasets
features_math = data_mat_encoded.drop(columns=['G1', 'G2', 'G3'])
features_portuguese = data_por_encoded.drop(columns=['G1', 'G2', 'G3'])
target_math = data_mat['G3']
target_portuguese = data_por['G3']

# Splitting the data into training and testing sets for both subjects
X_train_math, X_test_math, y_train_math, y_test_math = train_test_split(f
X_train_portuguese, X_test_portuguese, y_train_portuguese, y_test_portugu

# Building and training the linear regression model for both subjects
model_math = LinearRegression().fit(X_train_math, y_train_math)
model_portuguese = LinearRegression().fit(X_train_portuguese, y_train_por

# Predicting final grades for the test set and calculating the Mean Squar
predictions_math = model_math.predict(X_test_math)
predictions_portuguese = model_portuguese.predict(X_test_portuguese)

# Calculate the Mean Squared Error (MSE) for the Math dataset predictions
mse_math = mean_squared_error(y_test_math, predictions_math)

# Calculate the Mean Squared Error (MSE) for the Portuguese dataset predi
mse_portuguese = mean_squared_error(y_test_portuguese, predictions_portug

# Return the MSE values for both the Math and Portuguese datasets
(mse_math, mse_portuguese)
```

Out[106… (17.931455516013777, 8.710282809612078)

Research Question 6

Subject-Specific Performance Patterns:

"What patterns emerge from comparing student performance in Mathematics and the Portuguese language, and what might explain these differences?"

This question aims to identify and understand subject-specific factors influencing academic achievement. We will:

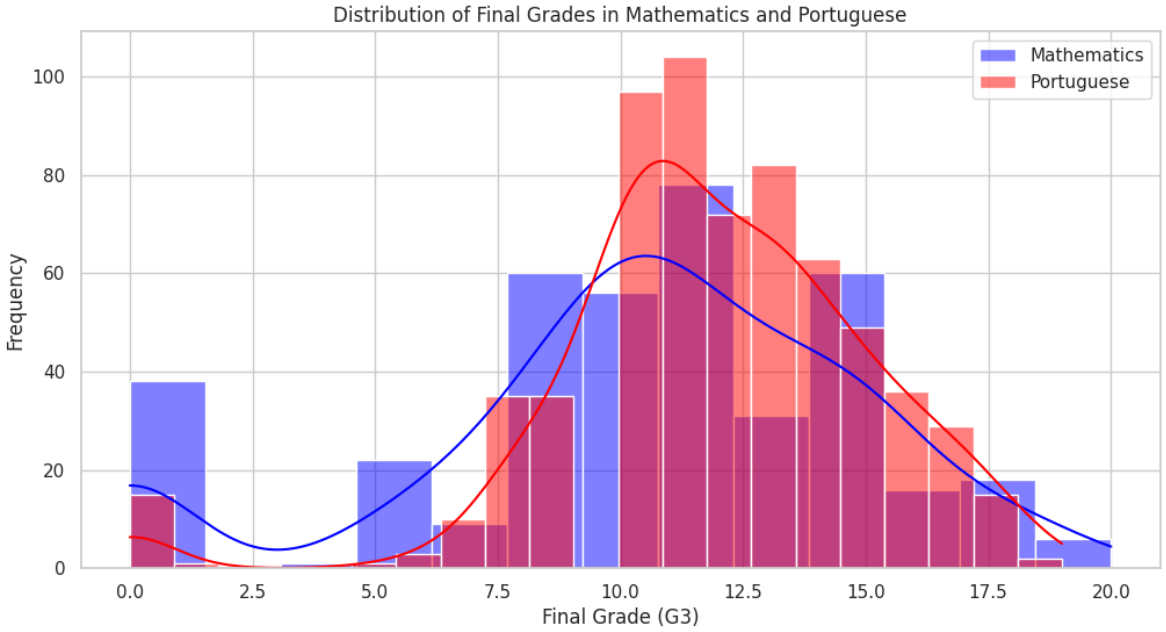
- Compare Performance Patterns: Analyze differences in performance between Mathematics and Portuguese courses.
- Investigate Possible Explanations: Explore potential reasons for these differences based on the data and previous findings.

Let us start by comparing the overall performance patterns in Mathematics and Portuguese. We will look at average grades, distribution of grades, and variance in performance between the two subjects.

In [107…

```
# Calculating average grades for both subjects
average_grade_math = data_mat['G3'].mean()
average_grade_portuguese = data_por['G3'].mean()

# Calculating the variance in grades for both subjects
```

```
variance_grade_math = data_mat['G3'].var()
variance_grade_portuguese = data_por['G3'].var()

# Plotting the distribution of final grades for both subjects
plt.figure(figsize=(12, 6))
sns.histplot(math_data['G3'], kde=True, color='blue', label='Mathematics'
sns.histplot(portuguese_data['G3'], kde=True, color='red', label='Portugu
plt.title('Distribution of Final Grades in Mathematics and Portuguese')
plt.xlabel('Final Grade (G3)')
plt.ylabel('Frequency')
plt.legend()
plt.show()

(average_grade_math, average_grade_portuguese, variance_grade_math, varia
```



Out[107…
```
(10.415189873417722,
 11.906009244992296,
 20.989616397866733,
 10.437139759173657)
```

The histogram displays the distribution of final grades (G3) in Mathematics and Portuguese language courses, and the statistics provide insights into their average grades and variance.

Key Observations:

- Average Grades: The average final grade in Portuguese (11.91) is higher than in Mathematics (10.42). This suggests that students perform better in Portuguese language courses on average.
- Grade Distribution: The distribution of grades in Mathematics shows a wider spread and higher variance (20.99) compared to Portuguese (10.44). This indicates more variability in student performance in Mathematics.
- Histogram Analysis: The histogram visually supports these findings, showing a broader distribution of grades in Mathematics and a denser concentration of higher grades in Portuguese.

Possible Explanations:

- Subject Complexity: Mathematics might be perceived as more challenging, leading to a broader range of performance.
- Teaching Methods: Differences in teaching methods or curriculum design between the subjects could influence student outcomes.
- Student Interest and Aptitude: Students might have varying interests and aptitudes towards these subjects, affecting their performance.

These findings and potential explanations contribute to understanding subject-specific factors in academic achievement, which can be crucial for curriculum development and teaching strategies.

Research Question 7

Impact of Extra-Curricular Activities:

"How do extra-curricular activities and personal interests impact academic performance in secondary education?"

This question examines the role of non-academic pursuits in academic success. We will analyze:

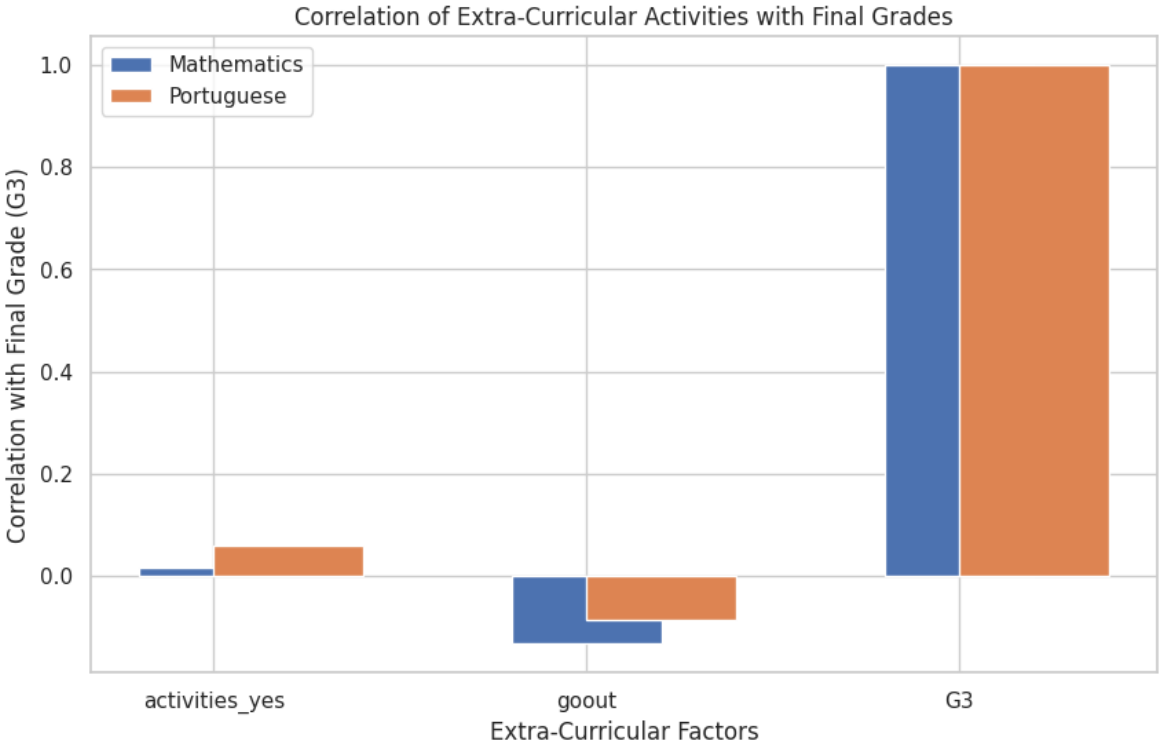
- Correlation with Academic Performance: Assess how participation in extra-curricular activities correlates with final grades (G3).
- Variation by Type of Activity: Explore if different activities have distinct impacts on grades.

For this analysis, we will focus on variables in the datasets that reflect extra-curricular involvement, such as participation in activities (activities) and time spent with friends (goout). Let us start by analyzing how these factors correlate with academic performance in Mathematics and Portuguese.

In [108…

```
# Encoding the 'activities' column: 'yes' to 1 and 'no' to 0
data_mat['activities_yes'] = data_mat['activities'].apply(lambda x: 1 if
data_por['activities_yes'] = data_por['activities'].apply(lambda x: 1 if

# Analyzing the correlation of extra-curricular activity participation wi
extra_curricular_math = data_mat[['activities_yes', 'goout', 'G3']].corr(
extra_curricular_portuguese = data_por[['activities_yes', 'goout', 'G3']]

# Creating a bar plot to compare the impact of extra-curricular activitie
plt.figure(figsize=(10, 6))
plt.bar(extra_curricular_math.index, extra_curricular_math, width=0.4, la
plt.bar(extra_curricular_portuguese.index, extra_curricular_portuguese, w
plt.title('Correlation of Extra-Curricular Activities with Final Grades')
plt.xlabel('Extra-Curricular Factors')
plt.ylabel('Correlation with Final Grade (G3)')
plt.legend()
plt.grid(True)
plt.show()
```

```
(extra_curricular_math, extra_curricular_portuguese)
```




```
Out[108…  (activities_yes    0.016100
           goout            -0.132791
           G3                1.000000
           Name: G3, dtype: float64,
           activities_yes    0.059791
           goout            -0.087641
           G3                1.000000
           Name: G3, dtype: float64)
```

The bar plot demonstrates the correlation between extra-curricular activities (participation in activities and time spent going out) and final grades (G3) in both Mathematics and Portuguese language courses.

Key Observations:

- Participation in Activities (activities_yes): A slight positive correlation exists with final grades in both subjects, more notably in Portuguese. This suggests that engagement in extra-curricular activities might be associated with slightly better academic performance.
- Time Spent Going Out (goout): Both subjects show a negative correlation with going out, indicating that more frequent social outings might be associated with lower academic performance. This aligns with our earlier findings on social behaviours.

These findings suggest that while extra-curricular activities can positively impact academic performance, excessive socializing outside of school might negatively affect grades. This underscores the importance of a balanced approach to extra-curricular involvement in enhancing academic success. This analysis provides insights into the multifaceted nature of educational achievement and the factors

contributing to student success in secondary education.

Ethical Considerations

Ethical Considerations in Analyzing Student Performance:

In the evolving landscape of educational research, using data analysis and machine learning to understand student performance is increasingly common. This project is an embodiment of this trend. While such projects hold great potential for enhancing educational outcomes, they also present many ethical considerations that must be rigorously addressed. This report explores these ethical dimensions, emphasizing the importance of data privacy, algorithmic fairness, research methodology, application of findings, and collaborative integrity.

Data Privacy and Confidentiality:

One of the foremost ethical considerations in educational research involving student data is the protection of privacy and confidentiality. In the era of big data, the risk of personal data misuse is high, and educational institutions bear a significant responsibility to safeguard the sensitive information of their students. This responsibility entails ensuring that personal identifiers are removed and data is anonymized before analysis. Any breach in this aspect could not only compromise the privacy of individuals but also erode trust in educational research practices.

Bias and Fairness in Algorithms:

As the project likely employs machine learning algorithms to analyze student performance, addressing the risk of algorithmic bias is critical. Algorithms, by their nature, are susceptible to the biases present in their training data or the perspectives of their developers. In an educational context, such biases can have profound implications, potentially leading to discriminatory practices against certain groups based on race, gender, socioeconomic status, or other demographic factors. Ensuring fairness in these algorithms is a technical challenge and a moral imperative to uphold equity in educational opportunities.

Research Methodology: Consent, Transparency, and Reliability:

The ethical research methodology involves more than just sound statistical analysis. It encompasses obtaining informed consent from participants, especially if new data is collected through surveys or direct interaction. Transparency about how the data will be used, stored, and disposed of is crucial to respect the autonomy and rights of the participants. Furthermore, the reliability and accuracy of the research are paramount. Misinterpreting data or erroneous conclusions can lead to misguided policies that may adversely affect students' educational experiences and outcomes.

Application of Findings and Educational Impact:

Applying the project's findings in shaping educational policies or interventions raises

further ethical questions. Any application must be sensitive to the diverse needs and backgrounds of students. Policies or interventions based on the research should not inadvertently marginalize or disadvantage any student group. Moreover, the researchers must exercise caution in generalizing their findings, acknowledging the limitations of their study and avoiding overreaching conclusions.

Conclusion:

In conclusion, this project, like any research involving human data, necessitates careful and thorough consideration of ethical issues. Privacy and confidentiality, fairness in algorithmic processing, sound and transparent research methodology, responsible application of findings, and ethical collaboration form the bedrock of such considerations. As we harness the power of data science in education, these ethical guidelines must be stringently observed to ensure that such endeavours contribute positively and equitably to the field of education.

Project Conclusion

The project "Student Performance" represents a comprehensive exploration of the factors influencing academic outcomes in a student population. Utilizing a robust dataset, the study delves into various variables that could impact student achievement, offering insights critical for educators, policymakers, and educational researchers.

Background and Motivation

Understanding student performance is pivotal in shaping effective educational strategies. The motivation behind this study stems from a growing need to identify and understand the diverse factors that contribute to educational success or challenges. The research aims to provide actionable insights that could improve educational practices and policies by analyzing these factors.

Research Questions

Central to this project are several research questions designed to unravel the complexities of academic performance. These questions likely explore the impact of family background, social behaviours, school-related factors, and individual student attributes like participation in extracurricular activities. Each question is tailored to dissect the intricate web of influences on a student's academic journey.

Data Analysis

The project employs rigorous data pre-processing and analytical techniques, including statistical analysis and predictive modelling. Key findings reveal how factors like family environment, social habits, and school support systems correlate with academic performance. The analysis also includes a comparative study of subjects, highlighting subject-specific trends and insights.

Conclusion

The study's conclusions are expected to shed light on the multifaceted nature of student performance. It underscores the significance of a holistic approach to education, where academic success is not only a function of curriculum and teaching methods but also of external and personal factors. These insights could be instrumental in guiding future educational reforms and targeted interventions.

Ethical Considerations

Given the sensitivity of educational data, the project undoubtedly addresses ethical considerations such as data privacy, consent, and the responsible use of student information. Ensuring the study's ethical integrity validates its findings and respects the rights and dignity of the student subjects involved.

In summary, "Student Performance" is a pivotal study contributing to understanding broader educational dynamics. Its findings could have far-reaching implications, influencing future educational strategies and policies to foster an environment where every student can succeed.

Evaluation and Reflection

Strengths:

- Comprehensive Data Selection and Analysis: Choosing a well-documented and reliable dataset from the UCI Repository adds credibility to the study. The multifaceted analysis, encompassing descriptive statistics, correlation, and predictive modelling, offers a thorough understanding of the factors influencing student performance.
- Insightful Findings: The discovery of significant correlations between early grades and final performance, gender disparities, and the impact of parental education provides valuable insights for educators and policymakers.
- Technical Proficiency: The application of various data science techniques, including predictive modelling and segmented analysis, demonstrates technical expertise and contributes to the robustness of the findings.

Areas for Improvement:

- Broader Data Scope: While the dataset is comprehensive, it is limited to two Portuguese schools. Expanding the dataset to include schools from different regions or countries could provide a more generalized understanding of student performance globally.
- Deeper Exploration of Causal Relationships: The study mainly focuses on correlations. Investigating causal relationships could provide deeper insights into how specific factors directly impact student performance.

- Inclusion of Qualitative Data: Incorporating qualitative data, such as student and teacher interviews, could offer additional perspectives on the quantitative findings, particularly in understanding the nuances behind the gender disparities and the impact of parental education.

Personal Reflection:

This project represents a significant learning opportunity in applying data science techniques to real-world educational issues. The experience of handling a comprehensive dataset, executing a multifaceted analysis, and deriving meaningful insights is invaluable. It highlights the potential of data science in influencing educational policy and strategy, a domain where data-driven decisions can profoundly impact future generations. The project also underscores the importance of continuous learning and improvement, especially in expanding the scope of data and exploring deeper causal relationships in future studies.

References

1. Article by Eyman Alyahyan et al. (2020):

- Title: "Predicting academic success in higher education: literature review and best practices"
- Journal: International Journal of Educational Technology in Higher Education
- Published: February 2020
- Authors: Eyman Alyahyan and Dilek Düştegör
- Available: https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-020-0177-7

2. Publication by the OECD (2015):

- Title: "The persistence of gender gaps in education and skills"
- Available: https://www.oecd-ilibrary.org/sites/34680dd5-en/1/3/1/index.html?itemId=/content/publication/34680dd5-en&_csp_=84042831e2796e3dbd529f3148909734&itemIGO=oecd&itemContentType=book

3. Article by Jeynes, William H. (2007):

- Title: "The Relationship between Parental Involvement and Urban Secondary School Student Academic Achievement"
- Journal: ERIC
- Published: 2007
- Authors: Jeynes, William H.
- Available: https://eric.ed.gov/?id=EJ748034

4. Article by The Portugal News (2020):

- Title: "State of Education 2019"

- Published: December 2020
- Authors: TPN/Lusa
- Available: https://www.theportugalnews.com/news/2020-12-23/education-improving-in-portugal/57356

5. Article by Al Jazeera (2023):

- Title: "'Unprecedented' decline in global literacy scores, OECD report says"
- Published: December 2023
- Authors: News Agencies
- Available: https://www.aljazeera.com/news/2023/12/5/unprecedented-decline-in-global-literacy-scores-osce-report-says

6. Report by the American Association of University Women (AAUW):

- Title: "The STEM Gap: Women and Girls in Science, Technology, Engineering and Mathematics"
- Available: https://www.aauw.org/resources/research/the-stem-gap/

7. Article by Olivia Guy-Evans et al. (2023):

- Title: "Bronfenbrenner's Ecological Systems Theory"
- Published: November 2023
- Authors: Olivia Guy-Evans and Saul Mcleod
- Available: https://www.simplypsychology.org/bronfenbrenner.html

8. Article by Sirin, S. R. (2005):

- Title: "Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research"
- Published: Fall 2005
- Authors: Selcuk R. Sirin
- Available: https://www.simplypsychology.org/bronfenbrenner.html

9. Article by Mahoney et al. (1997):

- Title: "Do extracurricular activities protect against early school dropout?"
- Published: 1997
- Authors: Mahoney, J. L., & Cairns, R. B.
- Available: https://psycnet.apa.org/record/1997-07406-005

10. Article by Hattie, J. (2008):

- Title: "Visible learning: A synthesis of over 800 meta-analyses relating to achievement"
- Published: 2008
- Authors: John Hattie
- Available: https://link.springer.com/article/10.1007/s11159-011-9198-8

11. Article by Paulo et al. (2008):

- Title: "USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE"
- Published: 2008
- Authors: Paulo Cortez and Alice Silva
- Available: https://www.researchgate.net/publication/228780408_Using_data_mining_to_predict_secondary_school_student_performan

12. Article by Ian J. Deary et al. (2007):

- Title: "Intelligence and educational achievement"
- Published: 2007
- Authors: Ian J. Deary, Steve Strand, Pauline Smith, Cres Fernandes
- Available: https://www.sciencedirect.com/science/article/abs/pii/S0160289606000171

13. Article by Daniel Voyer et al. (2014):

- Title: "Gender differences in scholastic achievement: A meta-analysis."
- Published: 2014
- Authors: Voyer, Daniel and Voyer, Susan D.
- Available: https://psycnet.apa.org/record/2014-15035-001

14. Article by Jacquelynne S. Eccles et al. (1999):

- Title: "Student council, volunteering, basketball, or marching band: What kind of extracurricular involvement matters?"
- Published: 1999
- Authors: Jacquelynne S. Eccles and Bonnie L. BarberView all authors and affiliations
- Available: https://journals.sagepub.com/doi/abs/10.1177/0743558499141003

Resources

- Data Visualization:
  - Dr. Joy Eze, Senior Lecture in MSc. in Data Science and Artificial Intelligence at Goldsmiths University of London
- Machine Learning
  - Dr. Daniel Stamate, Senior Lecture in MSc. in Data Science and Artificial Intelligence at Goldsmiths University of London
- Python:
  - Official Documentation: Python Documentation [Online]
  - Available: https://www.python.org/doc/
- Scikit Learn:
  - Official Documentation: Scikit Learn Documentation [Online]

    - Available: https://scikit-learn.org/stable/index.html
- Pandas:
    - Official Documentation: Pandas Documentation [Online]
    - Available: https://pandas.pydata.org/docs/#:~:text=,Mailing%20List
- NumPy:
    - Official Documentation: NumPy Documentation [Online]
    - Available: https://numpy.org/doc/#:~:text=,User%20Guide%20PDF
- Matplotlib.pyplot (a module in Matplotlib library):
    - Official Documentation: Matplotlib Pyplot Documentation [Online]
    - Available: https://matplotlib.org/stable/api/pyplot_summary.html#:~:text=,cases%20of%20programmatic%20plot%20
- Seaborn:
    - Official Documentation: Seaborn Documentation [Online]
    - Available: https://seaborn.pydata.org/#:~:text=,introductory%20notes%20or%20the%20paper