# Student Performance Analysis and Prediction

Date of Presentation- 04/1/25
Student Name- Divij Acharya
Student's Pace Email Address- da62111n@pace.edu
Class Name: Practical Data Science
Program Name: MS in Data Science
Seidenberg School of Computer Science and Information Systems, Pace University

# Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Modeling methods
- Findings
- Business recommendations and technical next steps

# Executive summary

Problem:
The dataset contains secondary school students' demographic details, family background, academic performance, and other factors. The goal is to analyze this data to predict student performance and determine whether a student will pass or fail.

# Executive summary

Solution:

Data visualization techniques were used to explore relationships between variables (e.g., study time vs. final grades, correlation heatmap).

•A Random Forest Regressor model was trained on features like G1, G2, study time, and absences to predict final grades (G3).

•The model's performance was evaluated using Mean Squared Error and R-squared score.

•A function was created to determine if a student passes or fails based on their predicted final grade, with a threshold of 10 or above considered passing.

•The solution includes data preprocessing (scaling features) and provides an example of how to use the prediction function.

# Project plan recap

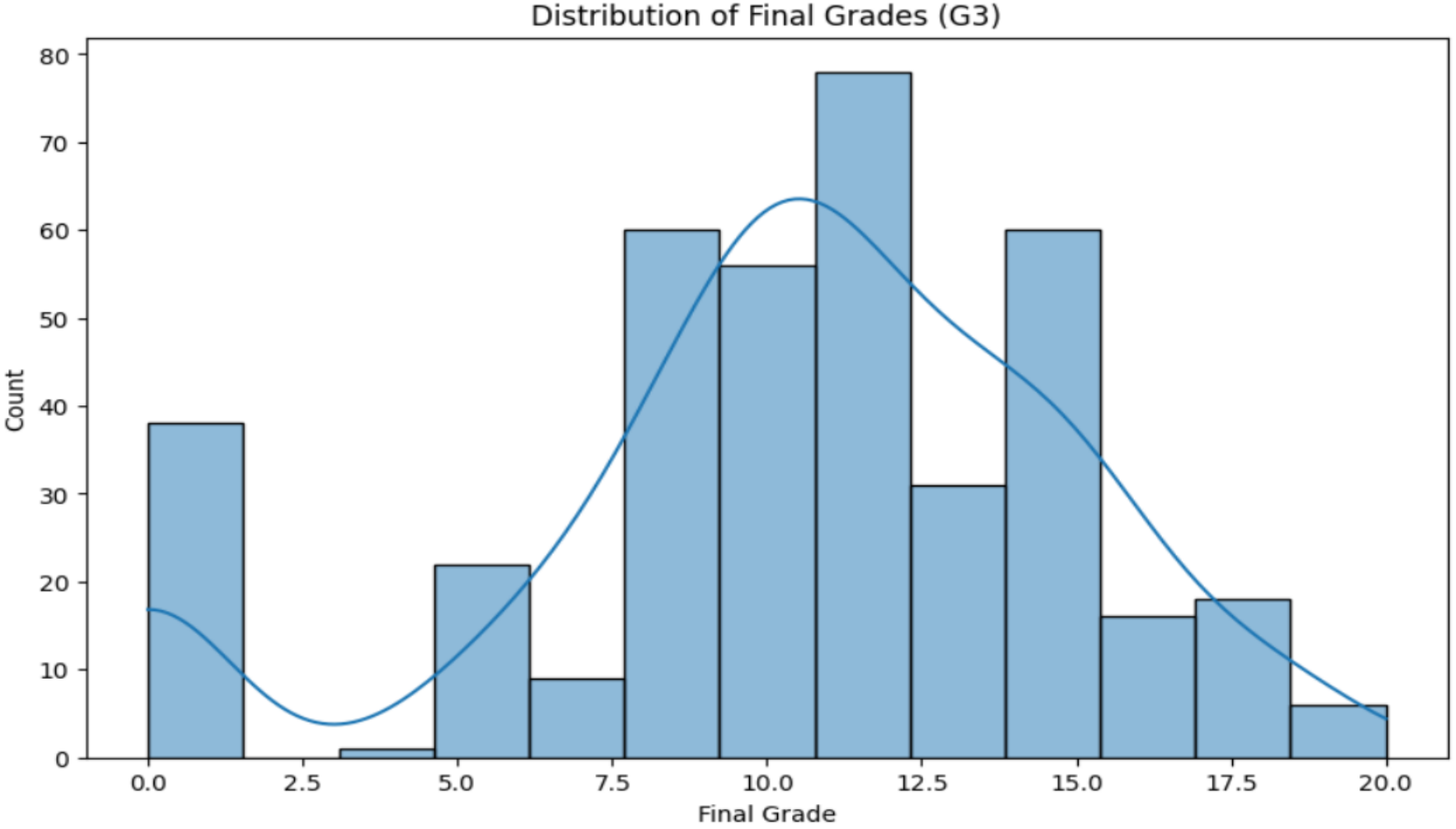
| Deliverable | Due Date | Status |
|---|---|---|
| Data & EDA | 03/25/25 | Not started / In Progress / Complete |
| Methods, Findings, and Recommendations | 04/01/25 | Not started / In Progress / Complete |
| Final Presentation | MM/DD/YY | Not started / In Progress / Complete |

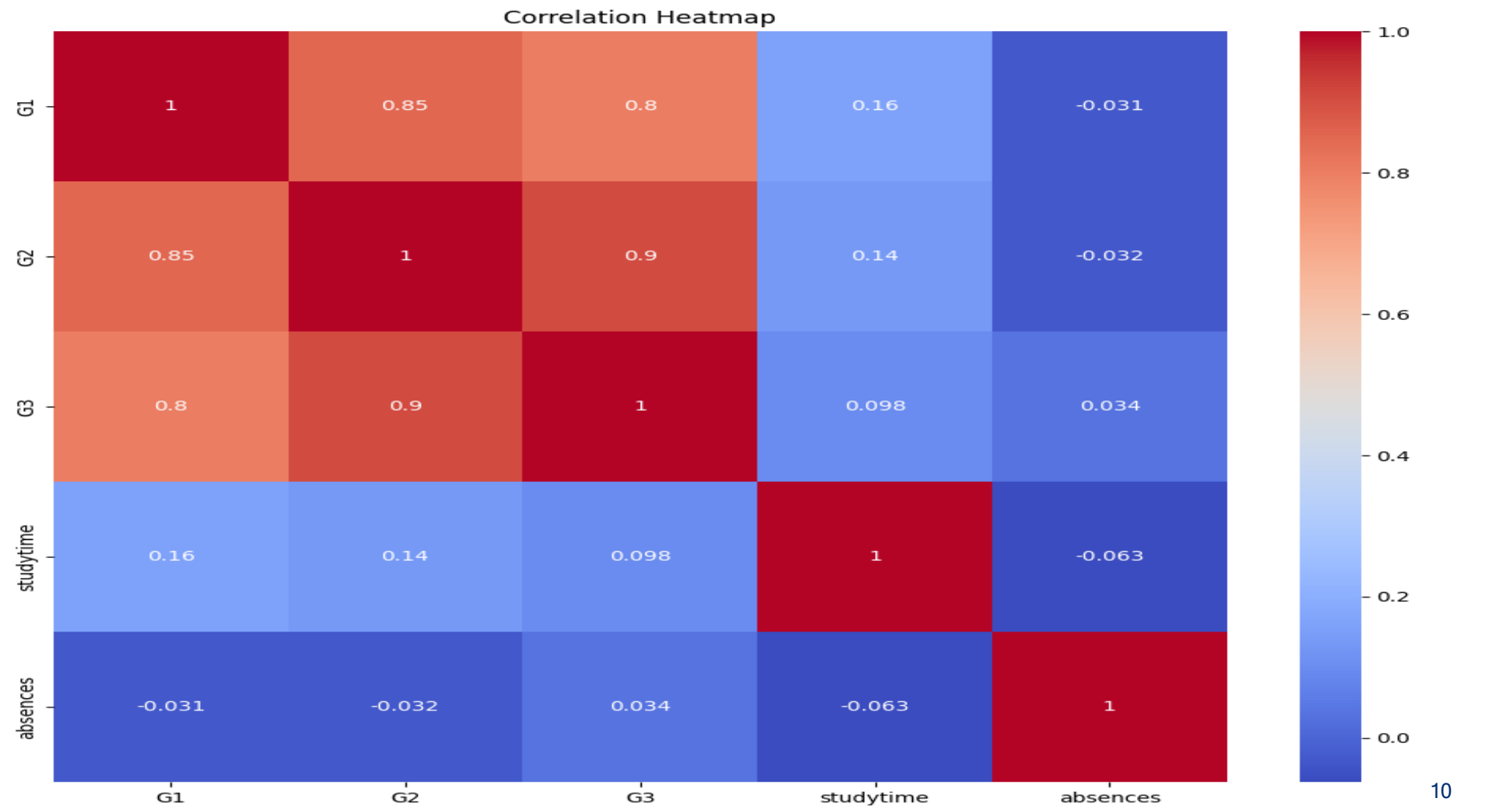
# Data

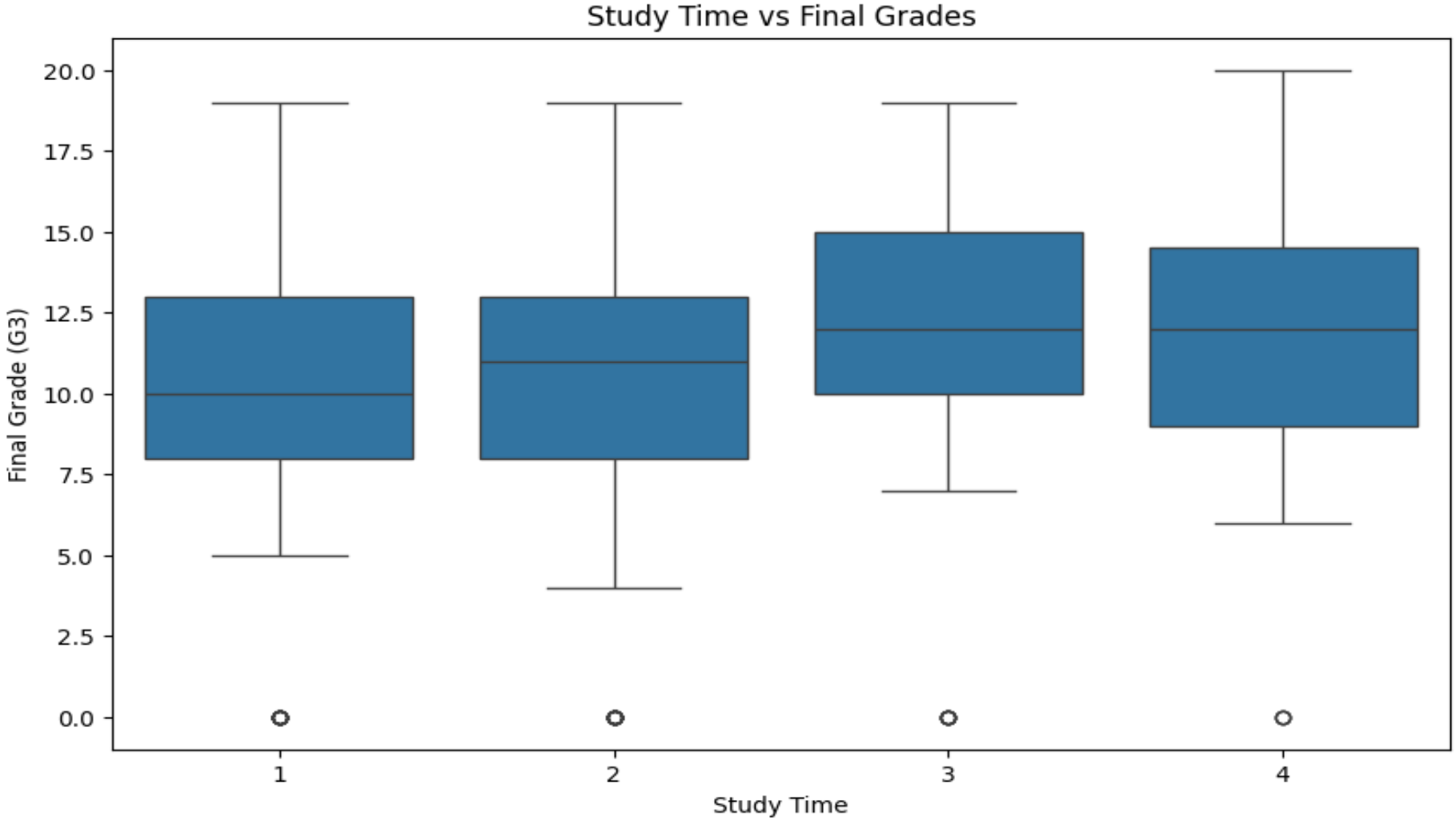
# Data

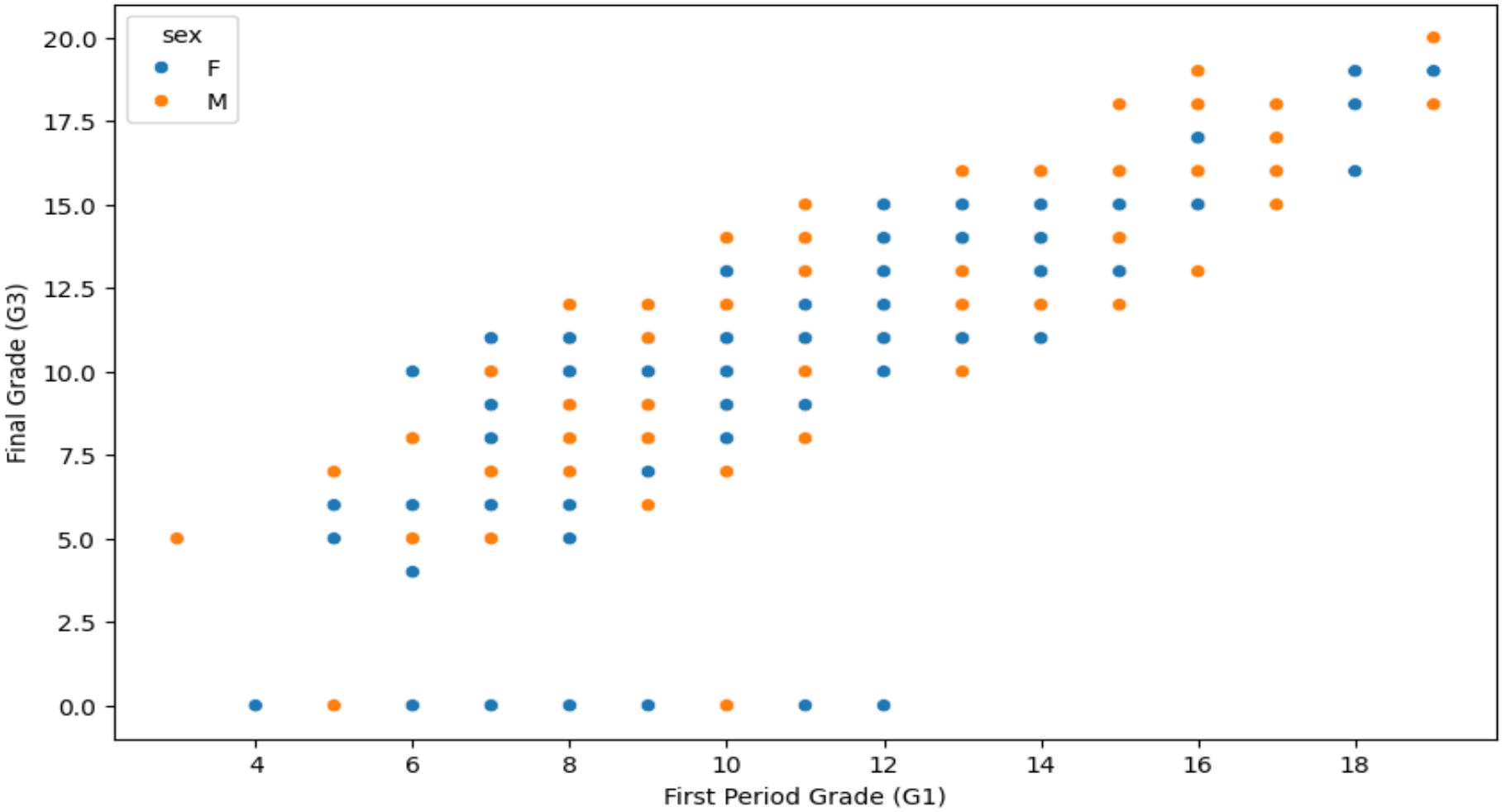
- Data source: https://archive.ics.uci.edu/dataset/320/student+performance
- Sample size: 396
- Attributes: 33
- Assumption: The dataset contains clean and complete data with no significant missing values or outliers. It is assumed that numerical features like G1, G2, study time, and absences directly correlate with the target variable (G3), which represents the final grade.

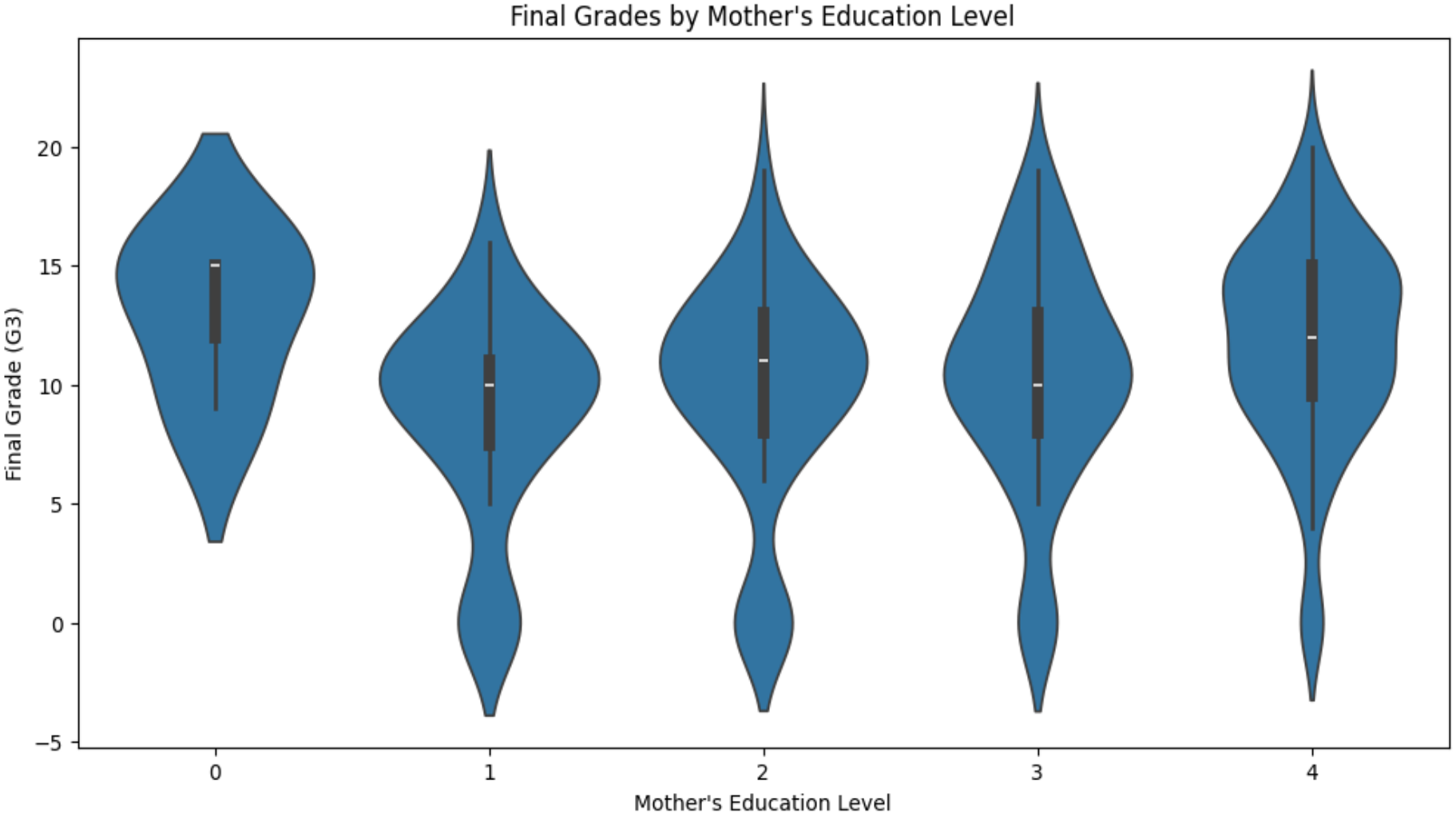
# Exploratory Data Analysis


Distribution of Final Grades (G3)
Count
Final Grade
0
10
20
30
40
50
60
70
80
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Correlation Heatmap
G1
G2
G3
studytime
absences
1
0.85
0.8
0.16
-0.031
0.85
1
0.9
0.14
-0.032
0.8
0.9
1
0.098
0.034
0.16
0.14
0.098
1
-0.063
-0.031
-0.032
0.034
-0.063
1
G1
G2
G3
studytime
absences
1.0
0.8
0.6
0.4
0.2
0.0


Study Time vs Final Grades
Final Grade (G3)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
Study Time


First Period Grade (G1) vs Final Grade (G3)
sex
F
M
Final Grade (G3)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
4
6
8
10
12
14
16
18
First Period Grade (G1)


Final Grades by Mother's Education Level
Final Grade (G3)
20
15
10
5
0
−5
0
1
2
3
4
Mother's Education Level

# Modeling Methods

# Modeling methods

Outcome Variable

- G3 (Final Grade): This variable represents the final grades of students, which is the target prediction of the model.

Features

•The model uses the following features to predict the outcome:

- G1 (First Period Grade): Grades from the first evaluation period.
- G2 (Second Period Grade): Grades from the second evaluation period.
- Study time: The Amount of study time allocated by the student.
- Absences: Number of school absences recorded.

# Modeling methods

Model Type and Rationale
Model Type: Random Forest Regressor

- Rationale:
- The Random Forest algorithm was selected for its robustness in handling non-linear relationships between features and outcomes.
- It provides insights into feature importance, which can be valuable for understanding the factors influencing student performance.
- The model is suitable for predicting continuous variables like grades while minimizing overfitting through ensemble learning.

This explanation is tailored for a non-technical audience of business stakeholders. Technical details, such as hyperparameters and theoretical aspects, are omitted but can be provided in an appendix if needed.

# Findings

# Findings

Key Findings

•Model Performance Metrics:

- Mean Squared Error (MSE): 1.92
  Indicates the average squared difference between predicted and actual grades.
- R-squared Score: 0.92
  Demonstrates that the model explains 92% of the variance in final grades (G3).

Visual Insights

•Correlation Heatmap:

Strong positive correlations were observed between G1, G2, and G3 (e.g., G2 correlates with G3 at ~0.91), validating their importance as predictors.

•Note:

Grades from earlier periods (G1 and G2) are the most significant predictors, while study time and absences contribute less to the prediction.

# Business Recommendations & Technical Next Steps

# Business Recommendations and Data Science Next Steps

Business Implications

•The model effectively identifies students likely to pass or fail based on their academic performance and behavioral factors.

•Example Prediction:

A student with G1 = 12, G2 = 14, study time = 2, and absences = 5 is predicted to "Pass" with a grade above the threshold of 10.

Recommendation

•Deploy this model to support early intervention strategies for at-risk students.

•Limitations: The model assumes consistent relationships across all students and does not account for external factors like socio-economic conditions.

# Business Recommendations and Data Science Next Steps

Technical Next Steps

Model

- Test XGBoost/LightGBM vs. RF + tune hyperparameters
- Explore neural nets for complex patterns

Features

- Add missing variables (family, behaviours)
- Create feature interactions/temporal trends

Deploy

- Deploy a minimal viable model (API + monitoring)

Data

- Expand data: motivation surveys, teaching methods

Validation

- Cross-validate yearly
- Build subject-specific models + A/B test interventions

# Thank You

# Appendix

# Project Materials

- Git Repo: <link>

# Title that connects back to the main slide in the deck

- Details
- Note: make sure that the main slide has a link to this slide because you don’t want your audience wondering what main slide this appendix slide supports