Checkpoint 1: Exploratory Data Analysis
CPSC 4300/6300 Applied Data Science

Michael Joseph Ellis

September 27, 2025

Background, Objective, and Framing

The dataset *Students Performance Dataset* provides records for 2,392 synthetic high-school students with demographics, study habits, parental involvement, extracurricular activities, and academic outcomes. Our objective for the semester project is to build a predictive model that estimates a student's final grade category, `GradeClass`, using the provided features. This frames the problem as a supervised multi-class classification task. We will ultimately evaluate tree-based gradient boosting models (e.g., XGBoost) and compare against strong baselines.

Rationale. `GradeClass` is an ordinal categorical outcome derived from GPA. Predicting `GradeClass` is useful in practice (early warning and support targeting) and aligns with the project's emphasis on interpretable, actionable features (study time, absences, parental support, etc.).

Dataset Summary

Source and unit of analysis. Data are downloaded from Kaggle (Rabie El Kharoua, CC BY 4.0). Each row corresponds to a unique student (`StudentID`). The file used is `Student_performance_data_.csv`. The class labels are coded as: 0=A, 1=B, 2=C, 3=D, 4=F.

Size. The dataset contains 2,392 observations (all unique) and 15 columns (including the target). All students are ages 15–18.

Key variables. Demographics (Age, Gender, Ethnicity, ParentalEducation), study habits (StudyTimeWeekly, Absences, Tutoring), parental involvement (ParentalSupport), extracurriculars (Extracurricular, Sports, Music, Volunteering), performance (GPA), and target (GradeClass).

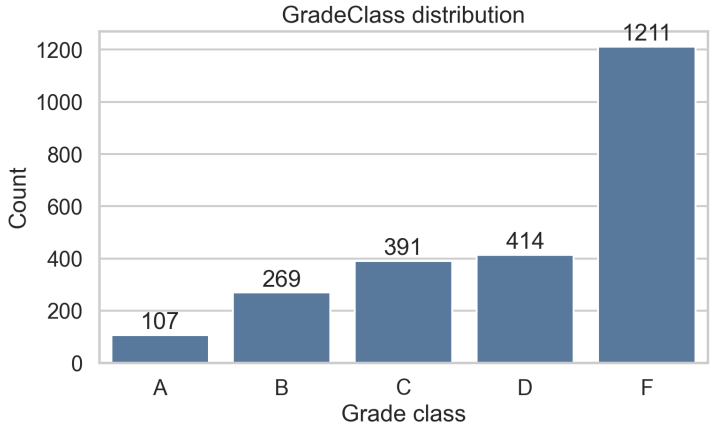

Figure 1: Distribution of GradeClass (A–F).

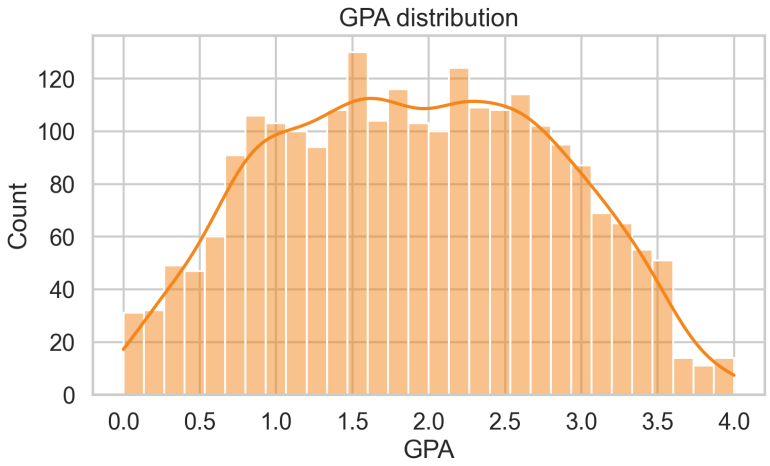

Figure 2: GPA distribution.

Data Cleaning Summary

We verified that `StudentID` is unique (2,392 unique IDs for 2,392 rows) and computed per-column missingness. The exported audit file (`figures/missingness.csv`) shows *zero* missing values across all columns, so no rows were dropped and no imputations were applied. Basic range checks matched the data card: Age 15–18; StudyTimeWeekly 0–20 hours; Absences 0–30; and GPA 0–4. Categorical fields (e.g., Gender, Ethnicity, ParentalEducation, Tutoring, ParentalSupport, Extracurricular, Sports, Music, Volunteering) are currently integer-coded and were left as-is for EDA; for modeling we will treat them as categorical (e.g., one-hot encoding).

We did not remove potential outliers at this stage. Distributions show expected tails (e.g., higher absences and lower GPA for some students) that are plausible for the domain. Tree-based boosting methods are robust to such values; we will revisit outlier influence during model diagnostics.

Outcome Variable

Definition. `GradeClass` is derived from GPA with bins: A (≥ 3.5), B ($[3.0, 3.5)$), C ($[2.5, 3.0)$), D ($[2.0, 2.5)$), and F (< 2.0). We visualize the distribution above and the underlying GPA shape to understand separability across bins.

Initial Feature Exploration

We inspected several hypothesized drivers of performance.

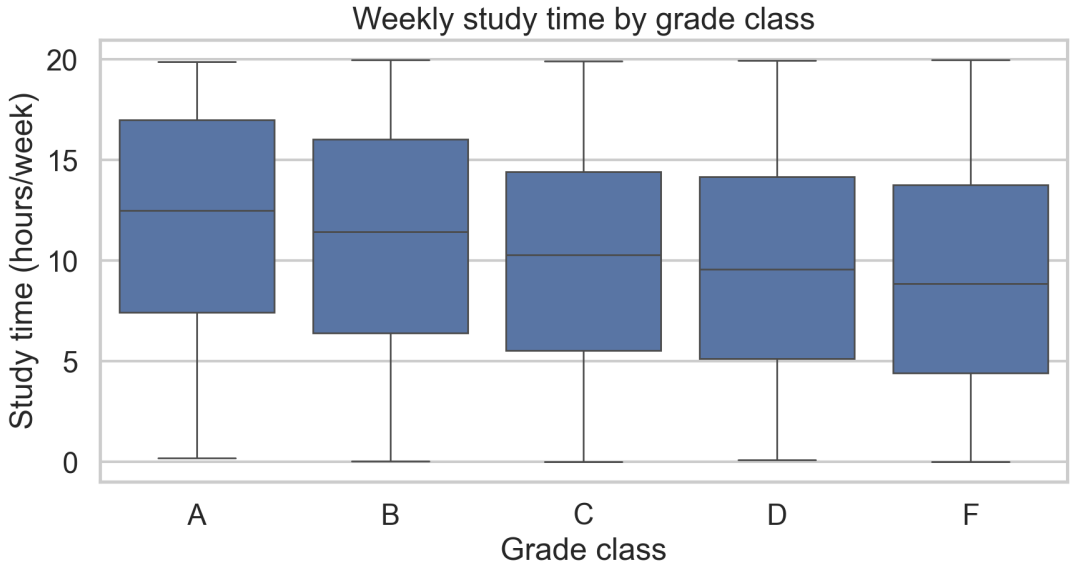

Figure 3: Weekly study time vs. grade class. Higher study time is associated with stronger letter grades.

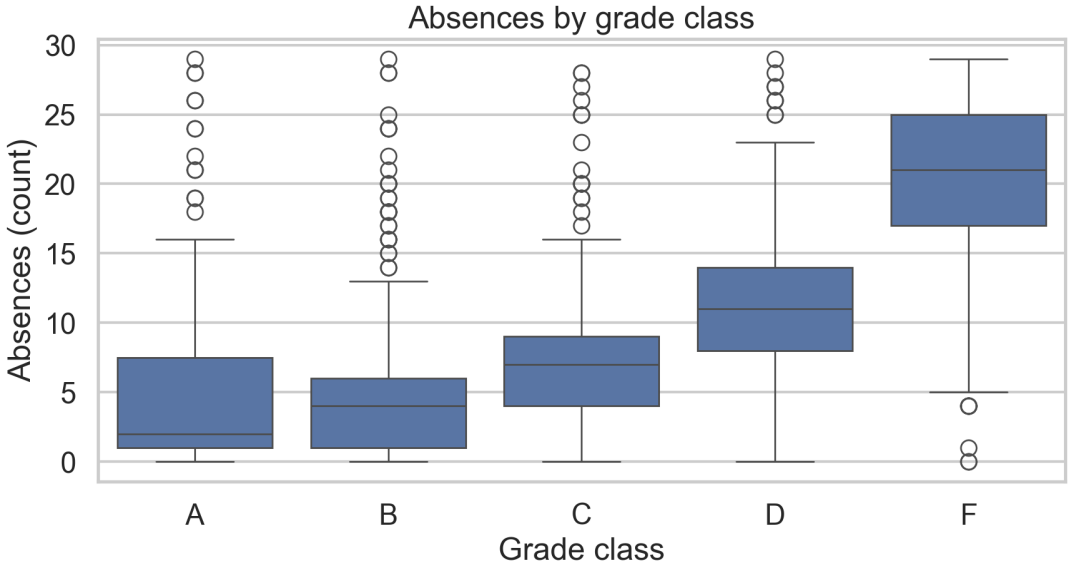

Figure 4: Absences vs. grade class. More absences correlate with lower grades.

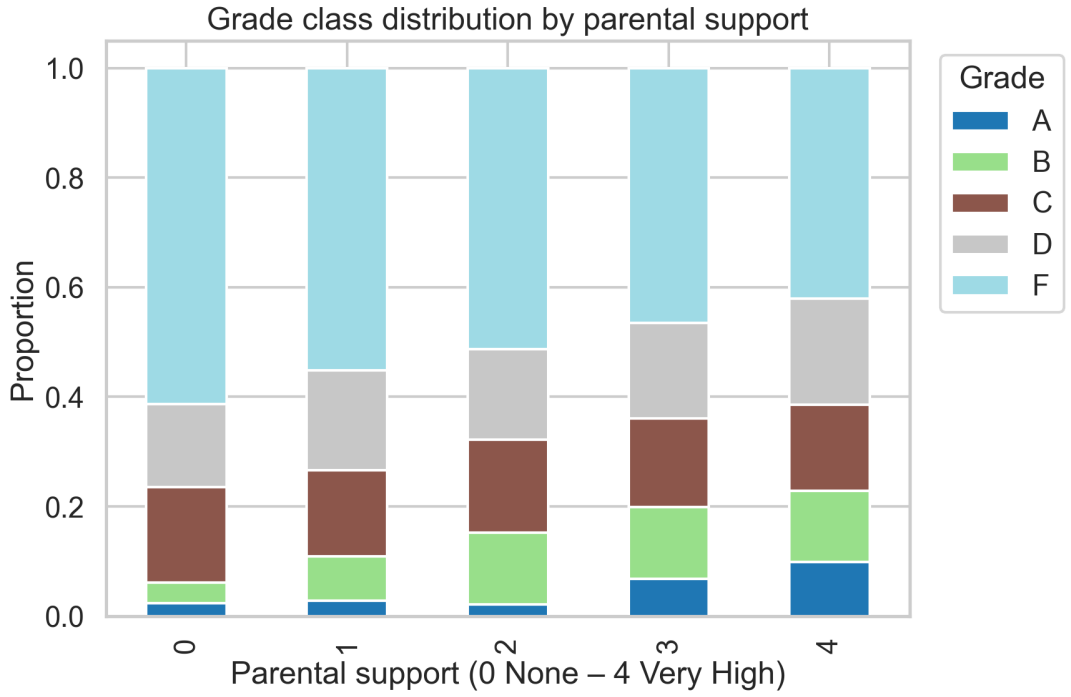

Figure 5: Grade-class composition by parental support level. Greater support shifts mass toward A/B.

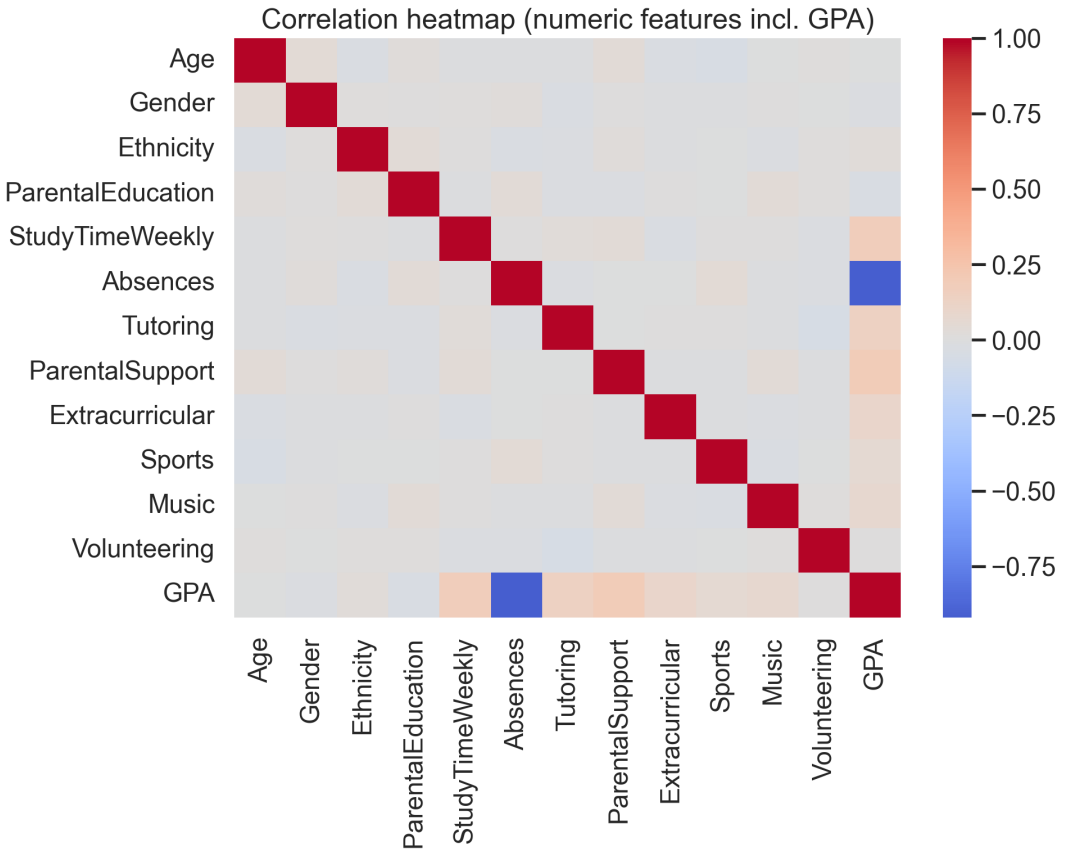

Figure 6: Correlation heatmap among numeric variables (including GPA).

Key Predictors for Modeling

We compared each candidate predictor to the response using distribution plots and simple summaries. Three features consistently showed the clearest relationships with performance:

- **Absences**: Strong negative association with performance. Spearman correlation with GPA is ≈ -0.93. Median absences rise monotonically across grade bins: A=2, B=4, C=7, D=11, F=21. The boxplot in Figure 4 shows a sharp shift toward higher absence counts for lower grades.
- **StudyTimeWeekly**: Positive association with performance. Spearman with GPA is ≈ 0.17. Median weekly study time decreases from A (≈ 12.5 h/wk) to F (≈ 8.8 h/wk). While the effect is weaker than absences, it is consistent and practically interpretable for interventions.
- **ParentalSupport**: Positive association with performance. Spearman with GPA is ≈ 0.18. The share of A/B among students with high support (levels 3–4) is about 20.7%, compared to 9.4% among those with low support (levels 0–1), reflecting a meaningful shift in grade mix (see stacked bars).

Selection for this assignment. In line with the rubric (choose one or two predictors), we select **Absences** and **StudyTimeWeekly** as the most important initial predictors:

1. *Absences* has the strongest empirical signal and a clear causal story (missed classes reduce learning opportunity).

2. *StudyTimeWeekly* is actionable (students can adjust effort) and complements Absences by capturing productive engagement outside class.

ParentalSupport will be included in subsequent modeling as an additional contextual predictor but we prioritize the two above for Checkpoint 1.

Modeling Plan (Preview)

For Checkpoint 2, we will implement an XGBoost multi-class classifier (softmax objective) with a stratified train/validation split and macro-averaged F1 as the primary metric. Categorical features coded as integers will be treated as ordered or one-hot encoded as appropriate; we will compare both encodings. We will include simple preprocessing (imputation and scaling where needed) inside a scikit-learn pipeline to keep evaluation clean and reproducible.

Reproducibility

All figures in this report are generated by `data analysis/reader.py` and saved to `Checkpoint 1/figures/`. Dependencies are listed in `requirements.txt`.
The complete code, CSV, and rendered figures are available in the public repository:

- GitHub (this project path): `https://github.com/Michael-Joseph-Ellis/disgusting-undergrad-busy-wor tree/main/CPSC%20Coursework/CPSC4300/Semester-Project`
- Data analysis code (modular): `CPSC Coursework/CPSC4300/Semester-Project/data analysis/eda/` (`main.py`, `plots.py`, `loader.py`, `stats.py`, `paths.py`);
- Figures generated by EDA: `CPSC Coursework/CPSC4300/Semester-Project/Checkpoint 1/figures/`