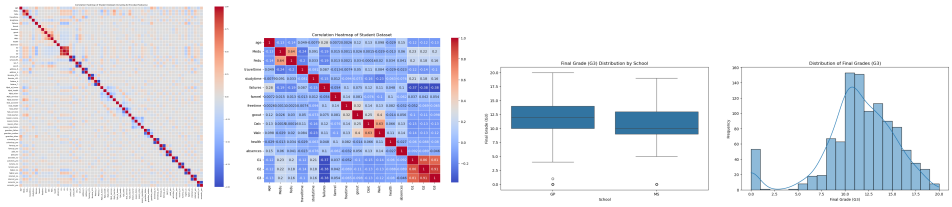
Exploratory Data Analysis



We analyzed student data from the UCI Machine Learning Repository to examine patterns in **G3** (final grades). Key observations include:

1. **Distribution of G3**:
    - The histogram shows a slightly left-skewed distribution. Most students scored between 10–15, with a peak around 12.
    - A significant number of outliers scored 0, likely representing students with multiple prior failures.
2. **Comparison by School**:
    - Boxplots revealed that students at **GP** school have slightly higher median grades than **MS**.
    - Overlapping interquartile ranges (IQRs) suggest overall similar performance, with some low-performing outliers in both schools.
3. **Correlation Heatmaps**:
    - **G3** strongly correlates with **G1** (0.81) and **G2** (0.91), reflecting the cumulative impact of earlier grades.
    - Moderate positive correlations with parental education (**Medu: 0.22**, **Fedu: 0.16**) highlight the importance of academic support at home.
    - Negative correlations with **failures (-0.38)** and weaker negative correlations with **goout (-0.10)** suggest that academic struggles and excessive social activities affect performance.

Social and Family Factors

1. **Failures vs. G3**:
    - Students with no prior failures had the highest **G3** scores. Grades dropped steeply with increasing failures.
2. **Goout vs. G3**:
    - Higher social activity levels correlated with greater grade variability. Frequent socializing often led to below-average grades.
3. **Romantic Relationships**:
    - Students in relationships had slightly lower median grades, but the impact was minimal, suggesting mild distractions.
4. **Interaction Effects**:

- Students with higher parental education and no failures consistently performed best.
- This emphasizes the compounded influence of supportive environments and academic discipline.

Key Insights

Students with strong academic histories, supportive family environments, and balanced social lives tend to excel. Targeted interventions for outliers (e.g., students with multiple failures or frequent social distractions) could significantly improve outcomes. Visualizations like heatmaps, histograms, and boxplots provide actionable insights for feature selection and modeling.

Feature Selection

Feature selection focused on variables with meaningful contributions to predicting **G3**, guided by exploratory data analysis (EDA), statistical correlations, and domain knowledge. The dataset includes 33 features across demographic, social, and academic dimensions. Key selected features include:

1. **Failures**:
   - **Definition**: Number of past academic failures.
   - **Correlation**: Strong negative (-0.38).
   - **Importance**: Captures cumulative academic struggles, making it a critical predictor.
2. **Mother’s Education (Medu)**:
   - **Definition**: Maternal education level (0–4).
   - **Correlation**: Moderate positive (0.22).
   - **Importance**: Reflects the academic support provided by an educated mother.
3. **Father’s Education (Fedu)**:
   - **Definition**: Paternal education level (0–4).
   - **Correlation**: Moderate positive (0.16).
   - **Importance**: Complements **Medu**, amplifying parental influence.
4. **Higher Education Aspiration (higher_yes)**:
   - **Definition**: Binary (plans for higher education).
   - **Correlation**: Positive (0.23).
   - **Importance**: Represents student motivation and long-term goals.
5. **Age**:
   - **Definition**: Student’s age (in years).
   - **Correlation**: Weak negative (-0.13).
   - **Importance**: Accounts for trends where older students may face challenges like repeating years or external responsibilities.
6. **Social Activity (goout)**:
   - **Definition**: Frequency of going out (1–5).

- **Correlation**: Moderate negative (-0.18).
- **Importance**: Highlights the trade-off between socialization and academic focus.

7. **Romantic Relationships (romantic_yes)**:

   - **Definition**: Binary (in a romantic relationship).
   - **Correlation**: Weak negative (-0.10).
   - **Importance**: Accounts for potential distractions in academic performance.

8. **Travel Time**:

   - **Definition**: Time taken to commute to school (1–4).
   - **Correlation**: Weak negative (-0.10).
   - **Importance**: Proxies logistical barriers, such as fatigue from long commutes.