Education Analytics Project: Data Cleaning & Analysis Assignment

# Work Order: Attendance and Achievement Correlation Analysis

From: Director of Education Analytics

To: Junior Data Analyst (You)

Due: End of Week

## Background

One of our partner districts is concerned about how student attendance impacts academic performance. They’ve provided us with a de-identified dataset containing attendance records and test scores from the last two academic years. They want actionable insights to guide interventions and resource allocation.

## Objective

Deliver an analysis that answers:  
1. How strongly is attendance correlated with academic performance?  
2. Which student groups (by grade level or subject area) show the greatest performance drop when attendance declines?  
3. What actionable recommendations can we make to the district based on the data?

## Tasks

1. Data Preparation  
- Import the dataset into your environment.  
- Clean the data: check for missing values, duplicates, or incorrect entries.  
- Create summary statistics for both attendance and test scores.  
  
2. Exploratory Analysis  
- Generate scatterplots showing attendance % vs. test scores.  
- Calculate correlation coefficients overall and by grade level.  
- Identify attendance thresholds (e.g., <85%) where performance noticeably declines.  
  
3. Deeper Cuts  
- Compare attendance-performance correlation across subjects.  
- Segment students into attendance tiers (High, Moderate, Low) and compare average scores.  
  
4. Deliverable  
- Write a 2–3 page report that includes:  
 • Executive summary (1–2 paragraphs, non-technical).  
 • Key findings with 2–3 charts/graphs.  
 • A short recommendations section.  
- Submit your cleaned dataset and analysis notebook alongside the report.

## Notes

- Use Python (pandas, matplotlib/seaborn) or R (tidyverse/ggplot2) for analysis.  
- Assume your audience is non-technical administrators.  
- You’re not expected to build a predictive model—just strong descriptive analysis.

# Data Cleaning Plan

Project: Attendance & Student Performance Dataset

Analyst: [Your Name]

Date: [Insert Date]

Tooling: Python 3.13, Jupyter Notebook, Windows CMD environment

## 1. Objective

Prepare the raw student dataset for analysis by ensuring it is complete, consistent, and reliable. The goal is to transform the raw CSV into a cleaned dataset that is ready for exploratory and statistical analysis of the relationship between attendance (absences) and academic performance (grades).

## 2. Scope of Work

Step 1: Initial Data Intake  
- Load the dataset into Python using pandas.  
- Inspect shape (rows × columns), file size, and preview records.  
  
Step 2: Schema Review  
- Generate schema summary (df.info()) to confirm data types.  
- Categorize fields into demographic, academic, attendance, lifestyle.  
- Ensure numeric fields are not stored as strings.  
  
Step 3: Completeness Check  
- Use df.isnull().sum() to identify missing values.  
- Scan for placeholders (e.g., '?', 'unknown', 'NA').  
- Document columns with missingness patterns.  
  
Step 4: Duplicates and Integrity  
- Check for duplicate rows (df.duplicated()).  
- Confirm student records are unique (no repeated IDs, if available).  
  
Step 5: Value Range Validation  
- Numeric fields: confirm grades fall in [0, 20], absences >= 0, age reasonable.  
- Categorical fields: validate unique values, standardize inconsistent strings.  
  
Step 6: Outlier Detection  
- Visualize distributions with histograms/boxplots.  
- Flag and review extreme values (e.g., absences >80).  
- Decide to remove, cap, or retain with justification.  
  
Step 7: Standardization & Encoding  
- Map categorical fields (Yes/No → 1/0).  
- One-hot encode multi-category fields.  
- Ensure encoding consistency across dataset.  
  
Step 8: Derived Fields (if useful)  
- Create average progress score (G1 + G2 / 2).  
- Convert absences to approximate attendance %.  
- Group attendance into tiers (High, Medium, Low).  
  
Step 9: Final Validation  
- Re-run df.describe() and df.info() after cleaning.  
- Ensure no unexpected nulls remain.  
- Verify row counts before vs. after cleaning.  
- Export cleaned dataset (student\_clean\_v1.csv).  
  
Step 10: Documentation  
- Record cleaning steps in Jupyter markdown.  
- Maintain reproducibility by saving notebook.  
- Version datasets clearly (v1, v2).

## 3. Deliverables

- Cleaned dataset (student\_clean\_v1.csv)  
- Jupyter notebook (data\_cleaning.ipynb)  
- One-page cleaning summary (this plan + final notes)

## 4. Assumptions / Risks

- Dataset is de-identified and suitable for analysis.  
- Extreme absence values may be entry errors or true; will document decision.  
- Encoding of categorical variables must remain interpretable for stakeholders.