CMPS 460 Machine Learning – Spring 2026

Assignment #1 – Data Preprocessing, Exploration, and Insight Discovery

Assigned: 05/02/2026 Due: **22/02/2026**

1. Assignment Objectives

The goal of this assignment is to develop strong, practical skills in data preprocessing and exploratory data analysis (EDA)—a critical foundation for any ML task.
By completing this assignment, you will learn how to:

- Assess real datasets for quality issues and limitations
- Apply systematic data cleaning and transformation techniques
- Justify preprocessing decisions using evidence and reasoning
- Extract insights through appropriate statistical measures
- Communicate findings clearly using visualizations and narrative explanations

---

2. Dataset Selection

Choose one real-world dataset from a reputable public source such as:

- Kaggle Datasets - https://www.kaggle.com/datasets
- UCI Machine Learning Repository https://archive.ics.uci.edu/ml
- World Bank Open Data https://data.worldbank.org
- Open Government Data such as https://data.gov and https://data.gov.uk
- Health Datasets: https://www.who.int/data

Your dataset must satisfy the following:

- At least 5000 records and 15–20 attributes
- Contains **real data quality issues**, such as: missing values, inconsistent formats (dates, categories, units), outliers or noise, redundant or irrelevant attributes

⚠ **Clean, "toy" datasets** with no preprocessing challenges (e.g., Iris, Titanic) are not acceptable.

---

3. Deliverable Format

- **Jupyter Notebook (.ipynb)** must include:
  - Clear **Markdown explanations**
  - Well-commented Python code
- Libraries to use: pandas, numpy, matplotlib, seaborn

---

4. Part A – Dataset Understanding & Motivation (5%)

Clearly describe:

- What the dataset represents (domain, context)
- Why it is interesting or useful
- Potential real-world or ML applications
- A link to the dataset source

---

5. Part B – Data Preprocessing & Cleaning (35%)

5.1 Data Inspection

- Data types, shape, and structure
- Initial summary statistics
- Detection of anomalies and inconsistencies

5.2 Handling Missing Data

- Identify missing values
- Justify your strategy: e.g., removal, mean/median/mode imputation, domain-specific handling, etc.

5.3 Data Cleaning

- Remove or correct invalid records
- Handle duplicates
- Standardize formats (dates, text, categories)
- Normalize or scale values (if relevant)

5.4 Outlier Detection and Treatment

- Identify outliers (visual or statistical methods)
- Decide whether to keep, transform, or remove them
- Justify your decision

5.5 Feature Preparation

- Drop irrelevant or redundant features
- Encode categorical variables
- Create derived features (optional but encouraged)

❗ **Marks are awarded for reasoning and justification**, not just code execution.

---

6. Part C – Exploratory Data Analysis (EDA)

6.1 Univariate Analysis (30%)

Select **meaningful variables only**.

- Measures of central tendency (mean, median, mode)
- Measures of variability (variance, standard deviation, IQR)
- **At least 4 suitable visualizations**

For each:

- Explain **why** the measure/chart was chosen
- Interpret the result
- Describe insights gained

6.2 Bivariate & Multivariate Analysis (25%)

Explore relationships between variables.

- Measures of correlation.
- **At least 4 visualizations**, such as: Scatter plots, Heatmaps, Pair plots and Grouped box plots.

For each:

- State the question being explored
- Interpret relationships or patterns

---

8. Grading Rubric

Component	Weight
Dataset understanding & motivation	5
Data preprocessing & cleaning (depth + justification)	35
Univariate statistical analysis + charts Include interpretations and what insights did you discover from your analysis.	30
Bi/multivariate statistical analysis + charts Include interpretations and what insights did you discover from your analysis.	30
Total	**100**

---

9. Key Evaluation Criteria

Your submission will be evaluated based on:

- ☑ Quality and realism of dataset
- ☑ Depth of preprocessing and cleaning
- ☑ Justification of decisions
- ☑ Insightfulness of analysis
- ☑ Clarity of explanations and visuals

Do not compute a statistic or plot a chart unless it provides insight.