ABV- Indian Institute of Information Technology & Management Gwalior
Department of Computer Science and Engineering
MLT LAB

Assignment 1
Topic: Pandas I/O, EDA and Feature Encoding
Date: 21st Aug 2025

Dataset: Download the *Titanic* dataset from Kaggle: Titanic – Machine Learning from Disaster.

Q1. Reading, Writing, and Appending Data in Different File Formats (CSV, Excel, JSON)

Objective: Demonstrate **read**, **write**, and **append** operations across multiple file formats using Pandas.

a) **Read** the dataset (CSV) into a Pandas DataFrame.

b) **Write** the loaded DataFrame into two additional formats:

   (a) Excel (.xlsx)

   (b) JSON (.json)

c) **Read** the dataset back from each format (CSV, Excel, JSON) into new DataFrames and verify consistency (e.g., shape, column names, basic value checks).

d) **Append** a few new passenger rows to each format (CSV, Excel, JSON). Reload each file and confirm that the newly appended rows are present.

e) Briefly compare the structural differences among CSV, Excel, and JSON (e.g., schema fidelity, metadata, readability, typical use-cases).

Deliverables: Code cells or script showing all I/O operations; short comments on format differences.

Q2. Descriptive Statistics for All Columns

Objective: Compute and interpret fundamental descriptive statistics.

a) For all **numerical** columns (e.g., Age, Fare), compute summary statistics: mean, median, standard deviation, minimum, and maximum.

b) Provide short interpretations: what do these values suggest about central tendency, variability, and possible outliers?

c) For **categorical** columns (e.g., Sex, Embarked, Pclass), report frequency counts and comment on the distributions.

Deliverables: A neatly formatted table (e.g.,DataFrame.describe( ) plus custom summaries) and 4-6 bullet points of insights.

Q3. Exploratory Data Visualizations

Objective: Explore distributions, outliers, and inter-variable relationships using standard plots.

Create the following, labeling axes and adding clear titles:

a) **Histogram** for a numerical variable (e.g., Age or Fare) to visualize its distribution.

b) **Pie chart** for a categorical feature (e.g., Sex or Embarked) to show composition.

c) **Box plot** for a numerical variable (e.g., Age or Fare) to detect potential outliers.

d) **Correlation heatmap** (for numerical columns) to visualize pairwise correlations.

Deliverables: Four plots (histogram, pie chart, box plot, heatmap) with 1-2 sentences interpreting each.

Q4. Feature Encoding Techniques (Label Encoding, One-Hot Encoding, and Dummy Encoding)

Objective: Transform categorical variables appropriately for downstream modeling.

a) Apply **Label Encoding** to a suitable categorical column (e.g., Embarked or Sex). Explain what the resulting numeric labels represent.

b) Apply **One-Hot Encoding** to the same column(s). Show how k new binary columns are created for k categories. Comment on the representation and redundancy in this encoding scheme.

c) Apply **Dummy Encoding** (a variant of One-Hot Encoding). Show how $k-1$ columns are created for k categories by dropping one reference category. Explain how the dropped category is represented implicitly.

d) Compare **One-Hot Encoding** vs. **Dummy Encoding**. Discuss in which contexts each should be used (e.g., One-Hot for tree-based models, Dummy for linear regression to avoid multicollinearity).

Deliverables: Before/after snapshots of the DataFrame (selected columns), plus a short comparison note (3-5 bullet points) highlighting differences between Label, One-Hot, and Dummy Encoding.

Submission Guidelines

- Submit a single, clearly structured notebook or script with headings for Q1-Q4.
- Include generated files (CSV, Excel, JSON) demonstrating **read**, **write**, and **append**.
- Ensure plots are visible within the notebook (or attach image files with captions).
- Add brief interpretations under each result/plot.