Training Day 9 Report

Date: 7 July 2025

Topic: GroupBy, Joins & Merges, DataFrame Manipulations, and Data Visualization using Pandas

Session Summary:
The ninth day of the training introduced advanced data manipulation operations in Pandas, focusing on grouping, joining, merging, concatenation, and visualization techniques. The session also progressed to real-world data analysis using the Titanic dataset, emphasizing missing value detection, data cleaning, and exploratory data analysis (EDA).

Key Concepts Covered:

- **1. GroupBy and Aggregation:** Learned how to group data using groupby() and perform aggregate functions such as mean, sum, and count. Resolved KeyError issues by using double brackets when grouping multiple columns. Observed insights such as mean population per continent.

- **2. Joins and Merges in Pandas:** Revisited database join concepts – inner, left, right, and outer joins. Used merge() with common keys like fellowship_id to combine tables. Discussed that outer join behaves similar to union but retains structure and fills missing values with NaN.

- **3. Concatenation:** Explored pd.concat() for vertical stacking of DataFrames, observing that it simply attaches rows without comparing keys or indexes.

- **4. Dropping Rows and Columns:** Practiced using drop() for removing rows or columns. Learned axis=0 for rows and axis=1 for columns. Understood difference between in-place modification and creating new copies.

- **5. Inserting or Updating Data:** Updated individual values using loc[] and slicing. Noted that insert() adds columns while append() is deprecated in newer versions of Pandas. Used df.loc[len(df)] = [...] to add new rows directly.

- **6. Plotting Data with Pandas:** Created visualizations using df.plot() including line, bar, and pie charts. Used subplots=True when plotting multiple numeric columns simultaneously.

- **7. Introduction to the Titanic Dataset:** Loaded the built-in Titanic dataset from Seaborn and studied its structure, key columns, and data types. Discussed redundancy and categorical encoding for better analysis.

- **8. Data Inspection & Cleaning:** Used .info() and .describe() for an overview of data. Identified missing values, computed missing counts and percentages, and created a new DataFrame summarizing these values. Sorted and filtered the data to focus on columns with null values.

- **9. Adding and Filtering Data:** Practiced adding new rows using loc[] and filtering rows based on conditions. Learned to sort values with sort_values() for better readability.

- **10. Visualization with Matplotlib:** Introduced matplotlib.pyplot as plt for advanced data visualization, discussed axes, labels, and figure customization for better presentation.

Skills Practiced:

- Data grouping and summarization
- Performing joins and merges
- Data concatenation and restructuring
- Managing missing values and null analysis
- Data cleaning and creation of derived columns
- Plotting data using Pandas and Matplotlib
- Building and understanding data frames programmatically

Instructor's Key Points:

- Always check data consistency before applying joins or merges.
- Avoid deprecated methods like append(); use loc or concat instead.
- Understand axis orientation: 0 = rows, 1 = columns.
- Choose chart types wisely for effective visualization.

Reflection & Learning Outcome:

Day 9 enhanced understanding of data manipulation and analysis using Pandas. Participants practiced grouping, joining, and merging datasets – essential skills for real-world analytics. The introduction to exploratory data analysis using the Titanic dataset provided practical exposure to missing value handling and visualization. This session effectively integrated theory and practice, strengthening overall data analysis workflow.