IntelliPaat

ASSIGNMENT 01 – **Pandas in Data Engineering**

Title: Perform Data Cleaning and Manipulation on a Sample SCM data with Pandas

Objective:

The objective of this lab is to familiarize with common data cleaning and manipulation tasks using Pandas dataframe. By the end of this lab, you should be able to:

- Load data into Pandas dataframe.
- Handle missing values.
- Filter and subset data.
- Perform basic data manipulations such as merging, grouping, and aggregating.

Duration: 2 hours

Prerequisites:

Basic understanding of Python programming language.

Familiarity with Pandas library.

Materials:

Jupyter Notebook environment with Python and Pandas installed.

Dataset: You can use the supply chain data available on Kaggle for this exercise. Make a dataset with a mix of categorical and numerical data, and make sure it has missing values for handling.

Lab Exercises:

1. Loading Data

- Load the dataset into a Pandas dataframe.
- Display the first 100 rows to understand the structure of the data.

2. Handling Missing Values

- Identify missing values in the dataset.
- Decide on a strategy to handle missing values (e.g., imputation, dropping rows/columns).
- Implement the chosen strategy and verify that missing values are handled appropriately.

3. Data Filtering and Subsetting

- Filter the data based on specific criteria (e.g., selecting rows where a certain condition is met).
- Subset the data by selecting specific columns of interest.

4. Data Manipulation

- Merge/join the dataset with another dataset on a common key.
- Group the data based on certain variables and perform aggregation functions (e.g., sum, mean) on grouped data.
- Create new columns by applying functions or operations on existing columns.

5. Visualization (Optional)

Visualize some aspects of the cleaned and manipulated data using libraries like Matplotlib or Seaborn.

6. Conclusion

- Summarize the key findings from the data cleaning and manipulation process.
- Reflect on the challenges encountered and how they were addressed.
- Discuss potential further analysis or next steps.