Intel® AI for Manufacturing Certificate Course

Week-4 – Assignment: Data Handling Report

Name: Ayan Memon
Submission Date: 16 – 03 – 2025

1. Introduction

The dataset used in this assignment is from an e-commerce platform, containing order details, product information, and customer data. The objective of this assignment is to perform Exploratory Data Analysis (EDA), clean the dataset, preprocess features, and split the data into training and testing sets for future machine learning tasks.

2. EDA & Data Cleaning

Missing Values

- We analyzed missing values using `.info()` and `.isnull().sum()`.
- Missing values in critical columns were handled by filling them with appropriate values (mean/median) or dropping irrelevant rows.

Outlier Detection & Handling

- Box plots were used to detect outliers in numerical features.
- Extreme values were either removed or adjusted to prevent them from affecting model performance.

Data Visualization

To better understand the dataset, the following visualizations were created:

- **Histograms:** Displayed data distribution for key numerical features.
- **Scatter Plots:** Showed relationships between features.
- **Box Plots:** Identified outliers in product size, weight, and delivery times.
- **Correlation Heatmap:** Highlighted relationships between variables using `sns.heatmap(df.corr(), annot=True)`.

3. Feature Engineering & Preprocessing

- **New Features Created:**
  - `distance` (calculated between customer and seller locations)
  - `wait_time` (difference between order and delivery time)
  - `purchase_dow` (day of the week when purchase was made)
  - `purchase_month` (month of the order)
- **Encoding Categorical Variables:**
  - Label Encoding was applied to categorical features such as `geolocation_state_customer` and `geolocation_state_seller`.

4. Data Splitting

- The final dataset was divided into:
  - **80% Training Set** for model training.
  - **20% Testing Set** for evaluation.
- We used `train_test_split()` from `sklearn.model_selection` to perform this split.

5. Conclusion

The dataset was successfully cleaned and preprocessed for modeling. Missing values and outliers were handled appropriately, and useful features were engineered. The dataset is now structured and ready for machine learning applications such as delivery time prediction. The next steps involve training and evaluating predictive models using this processed dataset.

Files for Submission:

- **Report (PDF/Word file) - This document Link:**

- **Python Code (Jupyter Notebook) Link:**

- **Final Processed Dataset (CSV File) Link:**