**📌 Data Cleaning & Preprocessing Report – Stock Dataset**

**📍 Overview**

The Stock Dataset was meticulously cleaned and preprocessed to ensure accuracy, consistency, and usability for further analysis. This dataset consists of stock market-related attributes, including **date, open, close, high, low, and volume**, providing insights into market trends and price movements. The data cleaning process focused on handling duplicates, standardizing column names, converting data types, and addressing missing values.

**✅ Step 1: Loading the Dataset**

* The dataset was successfully loaded into Pandas for analysis.
* It contained **multiple rows and columns** with numerical and categorical data.
* A preliminary check revealed some missing values and data type inconsistencies that required cleaning.

**✅ Step 2: Handling Duplicates**

* Checked for duplicate rows using df.duplicated().sum().
* **Duplicate records were identified and removed**, ensuring that each stock transaction remained unique.

**✅ Step 3: Standardizing Column Names**

* Column names were converted to **lowercase** and **underscores** replaced spaces for consistency.
* Example: "Open Price" → "open\_price", "Close Price" → "close\_price".
* This improves readability and ensures uniformity when referencing columns programmatically.

**✅ Step 4: Converting Data Types**

* **Date column** was converted to datetime format for proper time-based analysis.

python

df['date'] = pd.to\_datetime(df['date'])

* **Stock prices (open, close, high, low)** were checked and ensured to be in float format.

python

df[['open\_price', 'close\_price', 'high', 'low']] = df[['open\_price', 'close\_price', 'high', 'low']].astype(float)

* **Volume column** was also converted to a numerical format where needed.

**✅ Step 5: Handling Missing Values**

* Checked for missing values using df.isnull().sum().
* **Missing values were detected in stock price columns** and handled accordingly:
  + **For numerical columns** (open\_price, close\_price, high, low, volume):
    - Missing values were filled using the **mean** of the respective column to maintain consistency in trends.

python

df['open\_price'].fillna(df['open\_price'].mean(), inplace=True)

* + **For categorical columns** (if any, such as stock name or market segment):
    - Missing values were replaced with **"Unknown"** or the mode of the column.

python

df['stock\_name'].fillna('Unknown', inplace=True)

* **For time-series data**, forward fill (ffill) was applied to maintain sequential data integrity.

python

df.fillna(method='ffill', inplace=True)

**✅ Final Outcome**

* The dataset is now **clean, structured, and ready** for financial analysis and modeling.
* It can be used for **time-series forecasting, stock trend visualization, and machine learning applications** such as predictive modeling of stock prices.
* **Potential next steps** include **Exploratory Data Analysis (EDA)**, feature engineering, and applying statistical or AI-driven market predictions.