data-analyst-internship-task2

November 4, 2025

DATA ANALYST INTERNSHIP - TASK 2 Data Cleaning Script for Sample Superstore Dataset Author: Jagadeesh N Objective: 1. Clean and preprocess the Sample Superstore dataset. 2. Prepare it for visualization in Power BI or Tableau. 3. Ensure data consistency, accuracy, and readiness for storytelling and dashboard creation.

[1]:
```
#Loading data
import pandas as pd
file_path = r"D:\intern\Superstore.csv"
df = pd.read_csv(file_path, encoding='latin1')
print(df.head())
```

```
   Row ID        Order ID  Order Date   Ship Date       Ship Mode Customer ID  \
0       1  CA-2016-152156   11/8/2016  11/11/2016    Second Class    CG-12520
1       2  CA-2016-152156   11/8/2016  11/11/2016    Second Class    CG-12520
2       3  CA-2016-138688   6/12/2016   6/16/2016    Second Class    DV-13045
3       4  US-2015-108966  10/11/2015  10/18/2015  Standard Class    SO-20335
4       5  US-2015-108966  10/11/2015  10/18/2015  Standard Class    SO-20335

     Customer Name    Segment        Country             City  …  \
0      Claire Gute   Consumer  United States        Henderson  …
1      Claire Gute   Consumer  United States        Henderson  …
2  Darrin Van Huff  Corporate  United States      Los Angeles  …
3   Sean O'Donnell   Consumer  United States  Fort Lauderdale  …
4   Sean O'Donnell   Consumer  United States  Fort Lauderdale  …

  Postal Code  Region       Product ID         Category Sub-Category  \
0       42420   South  FUR-BO-10001798        Furniture    Bookcases
1       42420   South  FUR-CH-10000454        Furniture       Chairs
2       90036    West  OFF-LA-10000240  Office Supplies       Labels
3       33311   South  FUR-TA-10000577        Furniture       Tables
4       33311   South  OFF-ST-10000760  Office Supplies      Storage

                                        Product Name     Sales  Quantity  \
0                 Bush Somerset Collection Bookcase  261.9600         2
1  Hon Deluxe Fabric Upholstered Stacking Chairs,…  731.9400         3
2  Self-Adhesive Address Labels for Typewriters b…   14.6200         2
3     Bretford CR4500 Series Slim Rectangular Table  957.5775         5
4                   Eldon Fold 'N Roll Cart System   22.3680         2
```

```
   Discount    Profit
0      0.00   41.9136
1      0.00  219.5820
2      0.00    6.8714
3      0.45 -383.0310
4      0.20    2.5164

[5 rows x 21 columns]
```

[2]:
```
#Convert date columns
date_cols=['Order Date','Ship Date']
for col in date_cols:
    df[col]=pd.to_datetime(df[col], errors='coerce')
```

[3]:
```
#duplicate removal
duplicates =df.duplicated(subset=['Order ID', 'Product ID'])
df = df[~duplicates]
```

[4]:
```
# emove unnecessary columns
if 'Row ID' in df.columns:
    df.drop(columns=['Row ID'], inplace=True)
```

[5]:
```
# Keeping only valid discounts (0-1)
df = df[(df['Discount'] >= 0) & (df['Discount'] <= 1)]
```

[6]:
```
# Keeping rows with valid sales, profit, and quantity
df = df[(df['Sales'] >= 0) & (df['Quantity'] > 0) & (df['Profit'].notnull())]
```

[7]:
```
# Clean categorical columns
cat_cols = ['Ship Mode', 'Segment', 'Country', 'City', 'State', 'Region',␣
 ↪'Category', 'Sub-Category']
df[cat_cols] = df[cat_cols].apply(lambda x: x.str.strip())
```

[8]:
```
#Add Year-Month column
df['Order Year-Month'] = df['Order Date'].dt.to_period('M').astype(str)
```

[9]:
```
#Final Summary
print("Cleaned dataset shape:", df.shape)
print("Date range:", df['Order Date'].min(), "to", df['Order Date'].max())
print("\nSample data:\n", df.head())
```

```
Cleaned dataset shape: (9986, 21)
Date range: 2014-01-03 00:00:00 to 2017-12-30 00:00:00

Sample data:
          Order ID Order Date  Ship Date       Ship Mode Customer ID  \
```

```
0  CA-2016-152156 2016-11-08 2016-11-11    Second Class    CG-12520
1  CA-2016-152156 2016-11-08 2016-11-11    Second Class    CG-12520
2  CA-2016-138688 2016-06-12 2016-06-16    Second Class    DV-13045
3  US-2015-108966 2015-10-11 2015-10-18  Standard Class    SO-20335
4  US-2015-108966 2015-10-11 2015-10-18  Standard Class    SO-20335

     Customer Name    Segment        Country             City       State  \
0      Claire Gute   Consumer  United States        Henderson    Kentucky
1      Claire Gute   Consumer  United States        Henderson    Kentucky
2  Darrin Van Huff  Corporate  United States      Los Angeles  California
3   Sean O'Donnell   Consumer  United States  Fort Lauderdale     Florida
4   Sean O'Donnell   Consumer  United States  Fort Lauderdale     Florida

   …  Region       Product ID         Category Sub-Category  \
0  …   South  FUR-BO-10001798        Furniture    Bookcases
1  …   South  FUR-CH-10000454        Furniture       Chairs
2  …    West  OFF-LA-10000240  Office Supplies       Labels
3  …   South  FUR-TA-10000577        Furniture       Tables
4  …   South  OFF-ST-10000760  Office Supplies      Storage

                                        Product Name     Sales  Quantity  \
0                    Bush Somerset Collection Bookcase  261.9600         2
1  Hon Deluxe Fabric Upholstered Stacking Chairs,…  731.9400         3
2  Self-Adhesive Address Labels for Typewriters b…   14.6200         2
3      Bretford CR4500 Series Slim Rectangular Table  957.5775         5
4                       Eldon Fold 'N Roll Cart System   22.3680         2

   Discount    Profit  Order Year-Month
0      0.00   41.9136           2016-11
1      0.00  219.5820           2016-11
2      0.00    6.8714           2016-06
3      0.45 -383.0310           2015-10
4      0.20    2.5164           2015-10

[5 rows x 21 columns]
```

[10]:
```python
#Export Cleaned File
output_path = r"D:\intern\Cleaned_Superstore.csv"
df.to_csv(output_path, index=False)
print(f"\n File saved successfully at: {output_path}")
```

```
 File saved successfully at: D:\intern\Cleaned_Superstore.csv
```