1 Task - Sales_Data

April 4, 2025

1 Sales Data Project

Task 1 - Data Cleaning and Preprocessing

Welcome to the Sales Data project

[5]:

```
# Import libraries and packages
import pandas as pd
import numpy as np

# Load dataste into dataframe
df = pd.read_csv(r"C:\Users\Mahadev\Downloads\sales.csv")
```

[6]:

```
# display first few rows
df.head()
```

[6]:

```
   Transaction ID        Date Customer ID  Gender  Age Product Category  \
0               1  2023-11-24     CUST001    Male   34           Beauty
1               2  2023-02-27     CUST002  Female   26         Clothing
2               3  2023-01-13     CUST003    Male   50      Electronics
3               4  2023-05-21     CUST004    Male   37         Clothing
4               5  2023-05-06     CUST005    Male   30           Beauty

   Quantity  Price per Unit  Total Amount
0         3              50           150
1         2             500          1000
2         1              30            30
3         1             500           500
4         2              50           100
```

[7]:

```
# are there any null values? are all the variable numeric?
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Transaction ID    1000 non-null   int64
 1   Date              1000 non-null   object
 2   Customer ID       1000 non-null   object
 3   Gender            1000 non-null   object
 4   Age               1000 non-null   int64
 5   Product Category  1000 non-null   object
 6   Quantity          1000 non-null   int64
 7   Price per Unit    1000 non-null   int64
 8   Total Amount      1000 non-null   int64
dtypes: int64(5), object(4)
memory usage: 70.4+ KB
```

[13]:

```
# Handling missing values
missing_values = df.isnull().sum()
missing_values
```

[13]:

```
transaction_id      0
date                0
customer_id         0
gender              0
age                 0
product_category    0
quantity            0
price_per_unit      0
total_amount        0
dtype: int64
```

[14]:

```
# Remove duplicate rows
df.duplicated().sum()

df = df.drop_duplicates()
```

[15]:

```
# Rename columns (lowercase, no spaces, using underscore)
df.columns = df.columns.str.strip().str.lower().str.replace(" ","_")
df.columns
```

[15]:

```
Index(['transaction_id', 'date', 'customer_id', 'gender', 'age',
       'product_category', 'quantity', 'price_per_unit', 'total_amount'],
      dtype='object')
```

[11]:

```
# convert Date to datetime formate
```

[12]:

```
df['date'] = pd.to_datetime(df['date'], errors='coerce')

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
```

```
---  ------            --------------  -----
 0   transaction_id    1000 non-null   int64
 1   date              1000 non-null   datetime64[ns]
 2   customer_id       1000 non-null   object
 3   gender            1000 non-null   object
 4   age               1000 non-null   int64
 5   product_category  1000 non-null   object
 6   quantity          1000 non-null   int64
 7   price_per_unit    1000 non-null   int64
 8   total_amount      1000 non-null   int64
dtypes: datetime64[ns](1), int64(5), object(3)
memory usage: 70.4+ KB
```

Summary:

Sales dataset contains 1000 rows and 9 columns. there are not missing vlues found in any columns adn no duplicate rows were found. converted date column to datetime format. all column names are now lowercase with underscores. all columns data types are appropriate and corrected where needed.