data-analyst-internship-task3

November 5, 2025

[1]:
```
#Loading the dataset
import pandas as pd
file_path="D:\\intern\\online_retail_II.csv"
df=pd.read_csv(file_path)
print(df.head())
```

```
  Invoice StockCode                          Description  Quantity  \
0  489434     85048  15CM CHRISTMAS GLASS BALL 20 LIGHTS        12
1  489434    79323P                   PINK CHERRY LIGHTS        12
2  489434    79323W                  WHITE CHERRY LIGHTS        12
3  489434     22041         RECORD FRAME 7" SINGLE SIZE         48
4  489434     21232       STRAWBERRY CERAMIC TRINKET BOX        24

           InvoiceDate  Price  Customer ID         Country
0  2009-12-01 07:45:00   6.95      13085.0  United Kingdom
1  2009-12-01 07:45:00   6.75      13085.0  United Kingdom
2  2009-12-01 07:45:00   6.75      13085.0  United Kingdom
3  2009-12-01 07:45:00   2.10      13085.0  United Kingdom
4  2009-12-01 07:45:00   1.25      13085.0  United Kingdom
```

[2]:
```
#checking null values
df.isnull().sum()
```

[2]:
```
Invoice            0
StockCode          0
Description     4382
Quantity           0
InvoiceDate        0
Price              0
Customer ID    243007
Country            0
dtype: int64
```

[3]:
```
#deleteing duplicate data
df.drop_duplicates(inplace=True)
print(df.head())
```

```
  Invoice StockCode                          Description  Quantity  \
```

```
0  489434     85048  15CM CHRISTMAS GLASS BALL 20 LIGHTS        12
1  489434    79323P                   PINK CHERRY LIGHTS        12
2  489434    79323W                  WHITE CHERRY LIGHTS        12
3  489434     22041         RECORD FRAME 7" SINGLE SIZE         48
4  489434     21232       STRAWBERRY CERAMIC TRINKET BOX        24

           InvoiceDate  Price  Customer ID         Country
0  2009-12-01 07:45:00   6.95      13085.0  United Kingdom
1  2009-12-01 07:45:00   6.75      13085.0  United Kingdom
2  2009-12-01 07:45:00   6.75      13085.0  United Kingdom
3  2009-12-01 07:45:00   2.10      13085.0  United Kingdom
4  2009-12-01 07:45:00   1.25      13085.0  United Kingdom
```

[4]:
```
# Drop missing Customer IDs
df.dropna(subset=['Customer ID'])
```

[4]:
```
         Invoice StockCode                          Description  Quantity  \
0         489434     85048  15CM CHRISTMAS GLASS BALL 20 LIGHTS        12
1         489434    79323P                   PINK CHERRY LIGHTS        12
2         489434    79323W                  WHITE CHERRY LIGHTS        12
3         489434     22041         RECORD FRAME 7" SINGLE SIZE         48
4         489434     21232       STRAWBERRY CERAMIC TRINKET BOX        24
…           …       …                                      …       …
1067366   581587     22899         CHILDREN'S APRON DOLLY GIRL          6
1067367   581587     23254        CHILDRENS CUTLERY DOLLY GIRL          4
1067368   581587     23255      CHILDRENS CUTLERY CIRCUS PARADE         4
1067369   581587     22138        BAKING SET 9 PIECE RETROSPOT          3
1067370   581587      POST                              POSTAGE         1

                  InvoiceDate  Price  Customer ID         Country
0         2009-12-01 07:45:00   6.95      13085.0  United Kingdom
1         2009-12-01 07:45:00   6.75      13085.0  United Kingdom
2         2009-12-01 07:45:00   6.75      13085.0  United Kingdom
3         2009-12-01 07:45:00   2.10      13085.0  United Kingdom
4         2009-12-01 07:45:00   1.25      13085.0  United Kingdom
…                       …    …           …              …
1067366   2011-12-09 12:50:00   2.10      12680.0          France
1067367   2011-12-09 12:50:00   4.15      12680.0          France
1067368   2011-12-09 12:50:00   4.15      12680.0          France
1067369   2011-12-09 12:50:00   4.95      12680.0          France
1067370   2011-12-09 12:50:00  18.00      12680.0          France

[797885 rows x 8 columns]
```

[5]:
```
# Remove canceled invoices (InvoiceNo starting with 'C')
df = df[~df['Invoice'].astype(str).str.startswith('C')]
```

[6]:
```
#Total price
df['Total_Price']=df['Quantity']*df['Price']
print(df.head())
```

```
  Invoice StockCode                          Description  Quantity  \
0  489434     85048  15CM CHRISTMAS GLASS BALL 20 LIGHTS        12
1  489434    79323P                   PINK CHERRY LIGHTS        12
2  489434    79323W                  WHITE CHERRY LIGHTS        12
3  489434     22041         RECORD FRAME 7" SINGLE SIZE         48
4  489434     21232       STRAWBERRY CERAMIC TRINKET BOX        24

           InvoiceDate  Price  Customer ID         Country  Total_Price
0  2009-12-01 07:45:00   6.95      13085.0  United Kingdom         83.4
1  2009-12-01 07:45:00   6.75      13085.0  United Kingdom         81.0
2  2009-12-01 07:45:00   6.75      13085.0  United Kingdom         81.0
3  2009-12-01 07:45:00   2.10      13085.0  United Kingdom        100.8
4  2009-12-01 07:45:00   1.25      13085.0  United Kingdom         30.0
```

[7]:
```
# Convert InvoiceDate to datetime
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])
```

[8]:
```
# Extract Year and Month for analysis
df['Year'] = df['InvoiceDate'].dt.year
df['Month'] = df['InvoiceDate'].dt.month_name()
```

[9]:
```
# Basic validation
print(df.info())
print(df.describe())
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1013932 entries, 0 to 1067370
Data columns (total 11 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   Invoice      1013932 non-null  object
 1   StockCode    1013932 non-null  object
 2   Description  1009657 non-null  object
 3   Quantity     1013932 non-null  int64
 4   InvoiceDate  1013932 non-null  datetime64[ns]
 5   Price        1013932 non-null  float64
 6   Customer ID  779495 non-null   float64
 7   Country      1013932 non-null  object
 8   Total_Price  1013932 non-null  float64
 9   Year         1013932 non-null  int32
 10  Month        1013932 non-null  object
dtypes: datetime64[ns](1), float64(3), int32(1), int64(1), object(5)
memory usage: 89.0+ MB
None
```

```
           Quantity                    InvoiceDate         Price  \
count  1.013932e+06                        1013932  1.013932e+06
mean   1.073701e+01  2011-01-04 00:55:30.796325888  3.893570e+00
min   -9.600000e+03            2009-12-01 07:45:00 -5.359436e+04
25%    1.000000e+00            2010-07-05 13:56:00  1.250000e+00
50%    3.000000e+00            2010-12-09 14:09:00  2.100000e+00
75%    1.200000e+01            2011-07-27 15:16:00  4.130000e+00
max    8.099500e+04            2011-12-09 12:50:00  2.511109e+04
std    1.373870e+02                            NaN  9.492263e+01

         Customer ID   Total_Price          Year
count  779495.000000  1.013932e+06  1.013932e+06
mean    15320.262918  2.003841e+01  2.010436e+03
min     12346.000000 -5.359436e+04  2.009000e+03
25%     13971.000000  3.900000e+00  2.010000e+03
50%     15246.000000  1.000000e+01  2.010000e+03
75%     16794.000000  1.770000e+01  2.011000e+03
max     18287.000000  1.684696e+05  2.011000e+03
std      1695.722988  2.203573e+02  5.763373e-01
```

[10]:
```
# Export cleaned file
df.to_csv("D:\\intern\\Online_Retail_II_Cleaned.csv", index=False)
print("Cleaned dataset exported successfully!")
```

```
Cleaned dataset exported successfully!
```