analysis

January 6, 2026

[3]:
```
import os
import pandas as pd
```

[2]:
```
df = pd.read_csv("../data/Online_Retail.csv", encoding="ISO-8859-1")
```

[5]:
```
df.head()
```

[5]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/10 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/10 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/10 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/10 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/10 8:26	3.39	17850.0	United Kingdom

[6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    541909 non-null  object 
 1   StockCode    541909 non-null  object 
 2   Description  540455 non-null  object 
 3   Quantity     541909 non-null  int64  
 4   InvoiceDate  541909 non-null  object 
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object 
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

[7]:
```
print(df.isnull().sum())
```

```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

[9]:
```
df.describe()
```

[9]:

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

[11]:
```
df.shape
```

[11]:
```
(541909, 8)
```

[12]:
```
df.isna().sum().sort_values(ascending=False)
```

[12]:
```
CustomerID     135080
Description      1454
StockCode           0
InvoiceNo           0
Quantity            0
InvoiceDate         0
UnitPrice           0
Country             0
dtype: int64
```

[13]:
```
# usuwanie rekordów bez CustomerID
df = df.dropna(subset=["CustomerID"])
```

[14]:
```
df.isna().sum().sort_values(ascending=False)
```

[14]:
```
InvoiceNo      0
StockCode      0
Description    0
```

```
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

[15]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 406829 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    406829 non-null  object
 1   StockCode    406829 non-null  object
 2   Description  406829 non-null  object
 3   Quantity     406829 non-null  int64
 4   InvoiceDate  406829 non-null  object
 5   UnitPrice    406829 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      406829 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 27.9+ MB
```

[17]:
```
print(df.duplicated().sum())
```

```
5225
```

[18]:
```
df = df.drop_duplicates()
```

[19]:
```
print(df.duplicated().sum())
```

```
0
```

[20]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 401604 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    401604 non-null  object
 1   StockCode    401604 non-null  object
 2   Description  401604 non-null  object
 3   Quantity     401604 non-null  int64
 4   InvoiceDate  401604 non-null  object
 5   UnitPrice    401604 non-null  float64
 6   CustomerID   401604 non-null  float64
```

```
 7   Country      401604 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 27.6+ MB
```

[22]:
```
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'], errors='coerce')
```

[24]:
```
df['CustomerID'] = df['CustomerID'].astype(int)
```

[25]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 401604 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    401604 non-null  object
 1   StockCode    401604 non-null  object
 2   Description  401604 non-null  object
 3   Quantity     401604 non-null  int64
 4   InvoiceDate  401604 non-null  datetime64[ns]
 5   UnitPrice    401604 non-null  float64
 6   CustomerID   401604 non-null  int64
 7   Country      401604 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(2), object(4)
memory usage: 27.6+ MB
```

[37]:
```
df.nunique()
```

[37]:
```
InvoiceNo      22190
StockCode       3684
Description     3896
Quantity         436
InvoiceDate    20460
UnitPrice        620
CustomerID      4372
Country           37
dtype: int64
```

[38]:
```
df['Quantity'].value_counts()
```

[38]:
```
Quantity
 1        69605
 12       59828
 2        57425
 6        37480
 4        32093
          …
 234          1
```

```
 1404         1
 698          1
 80995        1
-80995        1
Name: count, Length: 436, dtype: int64
```

[39]:
```
df[df['Quantity'] < 0].shape
```

[39]:
```
(8872, 8)
```

[42]:
```
df[df['Quantity'] == 0 ].shape
```

[42]:
```
(0, 8)
```

[43]:
```
df[df['Quantity'] > 0].shape
```

[43]:
```
(392732, 8)
```

[44]:
```
df = df[df['Quantity'] > 0]
```

[45]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 392732 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    392732 non-null  object
 1   StockCode    392732 non-null  object
 2   Description  392732 non-null  object
 3   Quantity     392732 non-null  int64
 4   InvoiceDate  392732 non-null  datetime64[ns]
 5   UnitPrice    392732 non-null  float64
 6   CustomerID   392732 non-null  int64
 7   Country      392732 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(2), object(4)
memory usage: 27.0+ MB
```

[54]:
```
df[df['UnitPrice'] == 0].value_counts().sum()
```

[54]:
```
np.int64(40)
```

[55]:
```
df = df[df['UnitPrice'] > 0]
```

[56]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 392692 entries, 0 to 541908
Data columns (total 8 columns):
```

```
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    392692 non-null  object
 1   StockCode    392692 non-null  object
 2   Description  392692 non-null  object
 3   Quantity     392692 non-null  int64
 4   InvoiceDate  392692 non-null  datetime64[ns]
 5   UnitPrice    392692 non-null  float64
 6   CustomerID   392692 non-null  int64
 7   Country      392692 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(2), object(4)
memory usage: 27.0+ MB
```

[57]:
```
df.head()
```

[57]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom

[58]:
```
df['Revenue'] = df['Quantity'] * df['UnitPrice']
```

[59]:
```
df.head()
```

[59]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Revenue
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34

[ ]:
```

```