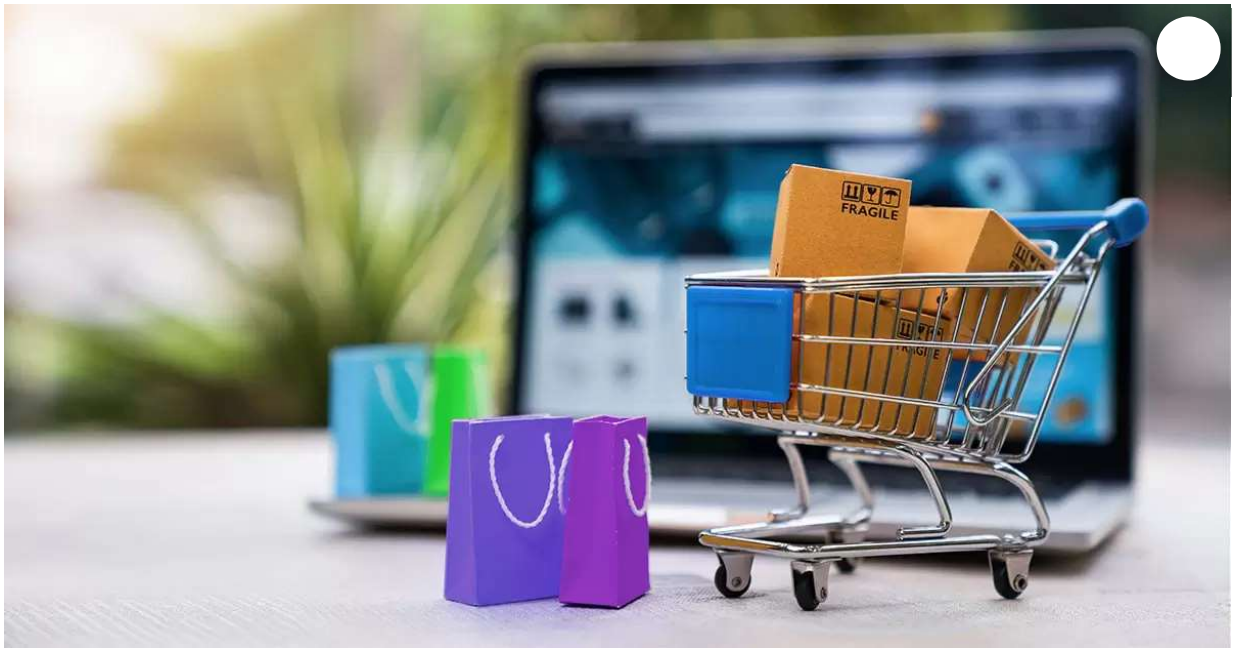
ECOMMERCE



In [2]:
```
###1.)Importing Basic Libraries.
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
print("All modules are imported.")
```

```
All modules are imported.
```

In [3]:
```
df=pd.read_csv("Downloads/Ecommerce",encoding="latin-1")
df
```

Out[3]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Countr
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	Unite Kingdor
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	Unite Kingdor
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	Unite Kingdor
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	Unite Kingdor
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	Unite Kingdor
...	...	...	...	...	...	...	...	.
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	Franc
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	Franc
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	Franc
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	Franc
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	Franc

541909 rows × 8 columns

In [4]:
```
df.head()
```

Out[4]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

In [5]:
```
df.tail()
```

Out[5]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	France

In [6]:
```
#MetaData
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
```

In [7]:
```
df.shape
```

Out[7]:
```
(541909, 8)
```

In [8]:
```
df.describe()
```

Out[8]:

	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

In [13]:
```
df["Description"].unique()
```

Out[13]:
```
array(['WHITE HANGING HEART T-LIGHT HOLDER', 'WHITE METAL LANTERN',
       'CREAM CUPID HEARTS COAT HANGER', ..., 'lost',
       'CREAM HANGING HEART T-LIGHT HOLDER',
       'PAPER CRAFT , LITTLE BIRDIE'], dtype=object)
```

In [14]:
```
df["CustomerID"].duplicated().sum()
```

Out[14]:
```
537536
```

In [19]:
```
d=df[df.duplicated()]
d
```

Out[19]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
517	536409	21866	UNION JACK FLAG LUGGAGE TAG	1	12/1/2010 11:45	1.25	17908.0	United Kingdom
527	536409	22866	HAND WARMER SCOTTY DOG DESIGN	1	12/1/2010 11:45	2.10	17908.0	United Kingdom
537	536409	22900	SET 2 TEA TOWELS I LOVE LONDON	1	12/1/2010 11:45	2.95	17908.0	United Kingdom
539	536409	22111	SCOTTIE DOG HOT WATER BOTTLE	1	12/1/2010 11:45	4.95	17908.0	United Kingdom
555	536412	22327	ROUND SNACK BOXES SET OF 4 SKULLS	1	12/1/2010 11:49	2.95	17920.0	United Kingdom
...	...	...	...	...	...	...	...	..
541675	581538	22068	BLACK PIRATE TREASURE CHEST	1	12/9/2011 11:34	0.39	14446.0	United Kingdom
541689	581538	23318	BOX OF 6 MINI VINTAGE CRACKERS	1	12/9/2011 11:34	2.49	14446.0	United Kingdom
541692	581538	22992	REVOLVER WOODEN RULER	1	12/9/2011 11:34	1.95	14446.0	United Kingdom
541699	581538	22694	WICKER STAR	1	12/9/2011 11:34	2.10	14446.0	United Kingdom
541701	581538	23343	JUMBO BAG VINTAGE CHRISTMAS	1	12/9/2011 11:34	2.08	14446.0	United Kingdom

5268 rows × 8 columns

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

Out[17]:
```
5268
```

In [22]:
```
print(d)
```

```
       InvoiceNo StockCode                          Description  Quantity  \
517       536409     21866        UNION JACK FLAG LUGGAGE TAG         1
527       536409     22866      HAND WARMER SCOTTY DOG DESIGN         1
537       536409     22900   SET 2 TEA TOWELS I LOVE LONDON           1
539       536409     22111      SCOTTIE DOG HOT WATER BOTTLE         1
555       536412     22327  ROUND SNACK BOXES SET OF 4 SKULLS         1
...          ...       ...                                 ...       ...
541675    581538     22068        BLACK PIRATE TREASURE CHEST         1
541689    581538     23318     BOX OF 6 MINI VINTAGE CRACKERS         1
541692    581538     22992            REVOLVER WOODEN RULER           1
541699    581538     22694                       WICKER STAR          1
541701    581538     23343       JUMBO BAG VINTAGE CHRISTMAS          1

            InvoiceDate  UnitPrice  CustomerID         Country
517     12/1/2010 11:45       1.25     17908.0  United Kingdom
527     12/1/2010 11:45       2.10     17908.0  United Kingdom
537     12/1/2010 11:45       2.95     17908.0  United Kingdom
539     12/1/2010 11:45       4.95     17908.0  United Kingdom
555     12/1/2010 11:49       2.95     17920.0  United Kingdom
...                 ...        ...         ...             ...
541675  12/9/2011 11:34       0.39     14446.0  United Kingdom
541689  12/9/2011 11:34       2.49     14446.0  United Kingdom
541692  12/9/2011 11:34       1.95     14446.0  United Kingdom
541699  12/9/2011 11:34       2.10     14446.0  United Kingdom
541701  12/9/2011 11:34       2.08     14446.0  United Kingdom

[5268 rows x 8 columns]
```

In [24]:
```
df.InvoiceDate.duplicated().sum()
```

Out[24]: 518649

In [25]:
```
(~df.duplicated()).sum()
```

Out[25]: 536641

In [27]:
```
df.loc[df.duplicated(),:]
```

Out[27]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
517	536409	21866	UNION JACK FLAG LUGGAGE TAG	1	12/1/2010 11:45	1.25	17908.0	United Kingdom
527	536409	22866	HAND WARMER SCOTTY DOG DESIGN	1	12/1/2010 11:45	2.10	17908.0	United Kingdom
537	536409	22900	SET 2 TEA TOWELS I LOVE LONDON	1	12/1/2010 11:45	2.95	17908.0	United Kingdom
539	536409	22111	SCOTTIE DOG HOT WATER BOTTLE	1	12/1/2010 11:45	4.95	17908.0	United Kingdom
555	536412	22327	ROUND SNACK BOXES SET OF 4 SKULLS	1	12/1/2010 11:49	2.95	17920.0	United Kingdom
...	...	...	...	...	...	...	...	..
541675	581538	22068	BLACK PIRATE TREASURE CHEST	1	12/9/2011 11:34	0.39	14446.0	United Kingdom
541689	581538	23318	BOX OF 6 MINI VINTAGE CRACKERS	1	12/9/2011 11:34	2.49	14446.0	United Kingdom
541692	581538	22992	REVOLVER WOODEN RULER	1	12/9/2011 11:34	1.95	14446.0	United Kingdom
541699	581538	22694	WICKER STAR	1	12/9/2011 11:34	2.10	14446.0	United Kingdom
541701	581538	23343	JUMBO BAG VINTAGE CHRISTMAS	1	12/9/2011 11:34	2.08	14446.0	United Kingdom

5268 rows × 8 columns

In [28]:
```
df.duplicated(keep="first").sum()
```

Out[28]:
```
5268
```

In [41]:
```
df.drop_duplicates(inplace=True)
```

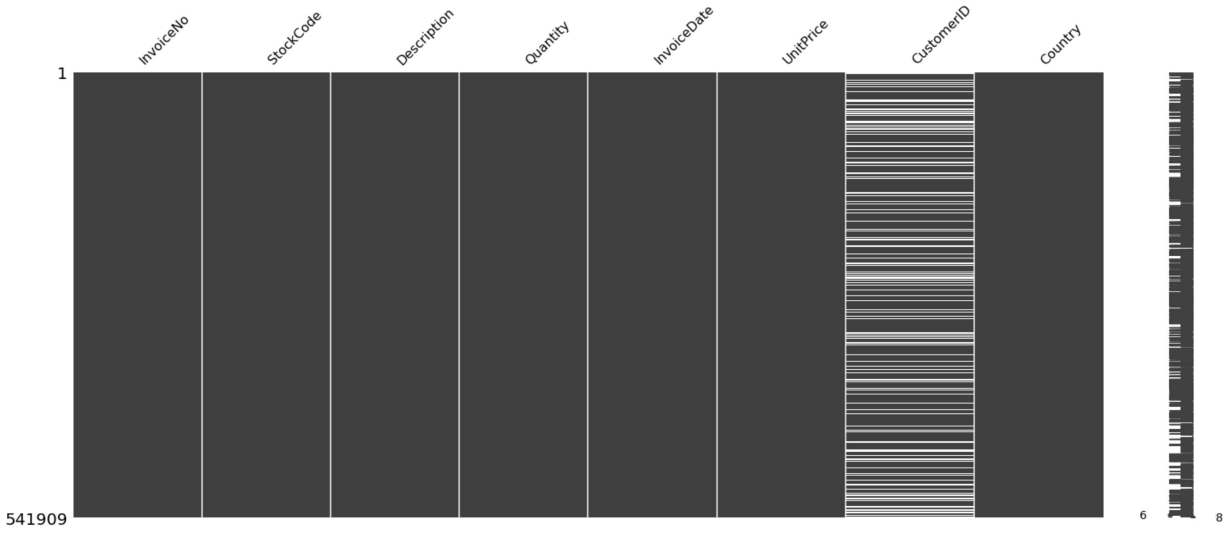
In [30]:
```
pip install missingno
```

```
Requirement already satisfied: missingno in c:\programdata\anaconda3\lib\site-p
ackages (0.4.2)Note: you may need to restart the kernel to use updated package
s.
Requirement already satisfied: scipy in c:\programdata\anaconda3\lib\site-packa
ges (from missingno) (1.5.0)
Requirement already satisfied: seaborn in c:\programdata\anaconda3\lib\site-pac
kages (from missingno) (0.10.1)
Requirement already satisfied: matplotlib in c:\programdata\anaconda3\lib\site-
packages (from missingno) (3.2.2)
Requirement already satisfied: numpy in c:\programdata\anaconda3\lib\site-packa
ges (from missingno) (1.18.5)
Requirement already satisfied: pandas>=0.22.0 in c:\programdata\anaconda3\lib\s
ite-packages (from seaborn->missingno) (1.0.5)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in c:\p
rogramdata\anaconda3\lib\site-packages (from matplotlib->missingno) (2.4.7)
Requirement already satisfied: cycler>=0.10 in c:\programdata\anaconda3\lib\sit
e-packages (from matplotlib->missingno) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\programdata\anaconda3\li
b\site-packages (from matplotlib->missingno) (1.2.0)
Requirement already satisfied: python-dateutil>=2.1 in c:\programdata\anaconda3
\lib\site-packages (from matplotlib->missingno) (2.8.1)
Requirement already satisfied: pytz>=2017.2 in c:\programdata\anaconda3\lib\sit
e-packages (from pandas>=0.22.0->seaborn->missingno) (2020.1)
Requirement already satisfied: six in c:\programdata\anaconda3\lib\site-package
s (from cycler>=0.10->matplotlib->missingno) (1.15.0)
```

In [31]:
```
import missingno as mn
```

In [32]:
```
mn.matrix(df)
```

Out[32]: <matplotlib.axes._subplots.AxesSubplot at 0x29307416c70>

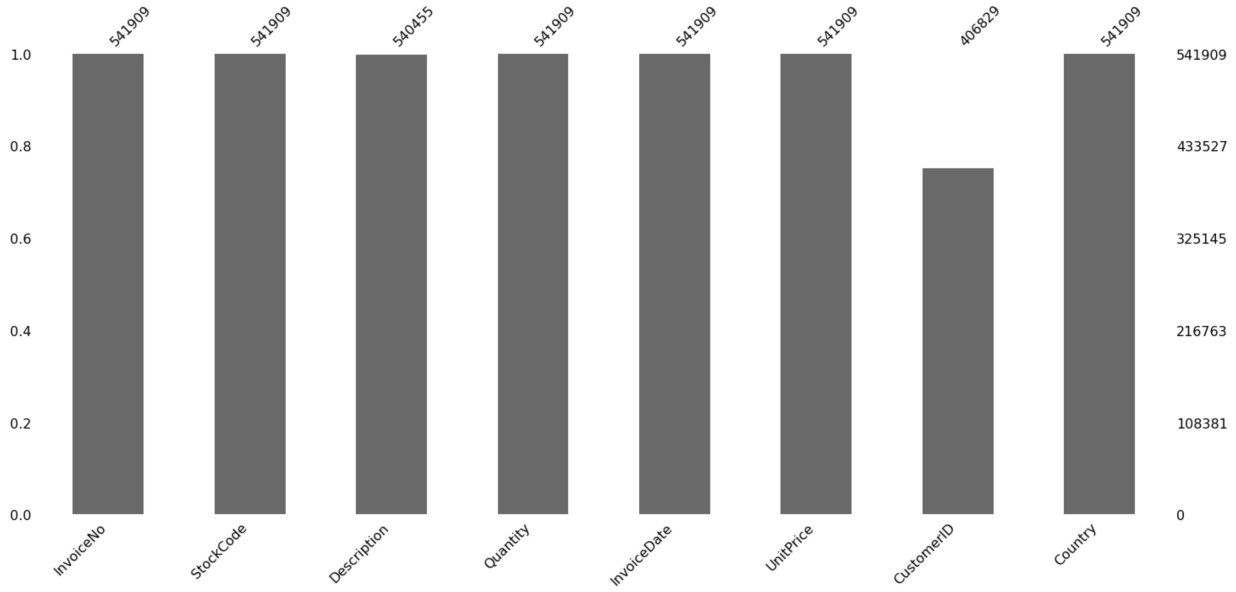


In [33]:
```
mn.heatmap(df)
```

Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x2930c719a00>

In [34]:
```
mn.bar(df)
```

Out[34]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2930cc7dfd0>
```



In [42]:
```
df.shape
```

Out[42]:
```
(536641, 8)
```

In [43]:
```
df.isnull().sum()
```

Out[43]:
```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID     135037
Country            0
dtype: int64
```

Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.

Missing At Random(MAR): The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.

Not Missing At Random(NMAR): When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

In [44]:
```
df.dropna(axis=0, how='all')
```

Out[44]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Countr
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	Unite Kingdor
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	Unite Kingdor
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	Unite Kingdor
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	Unite Kingdor
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	Unite Kingdor
...	...	...	...	...	...	...	...	.
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	Franc
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	Franc
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	Franc
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	Franc
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	Franc

536641 rows × 8 columns

In [46]:
```
df.dropna(axis=0, how='all',inplace=True)
```

In [47]:
```
df.shape
```

Out[47]:
```
(536641, 8)
```

In [49]:
```
df.to_csv("CleanEcommerce.csv",encoding = 'utf-8')
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: