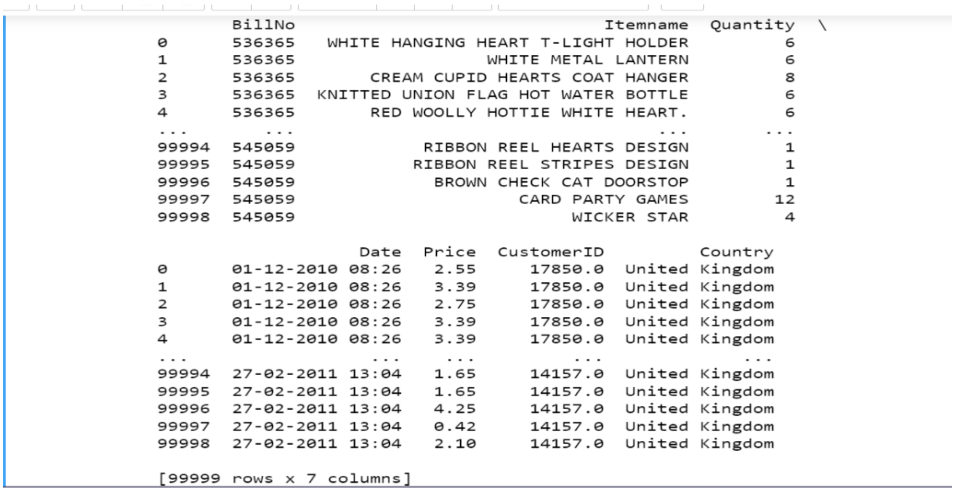
MARKET BASKET INSIGHTS

PROJ_ID: Proj_225020_Team_1

PHASE: 03

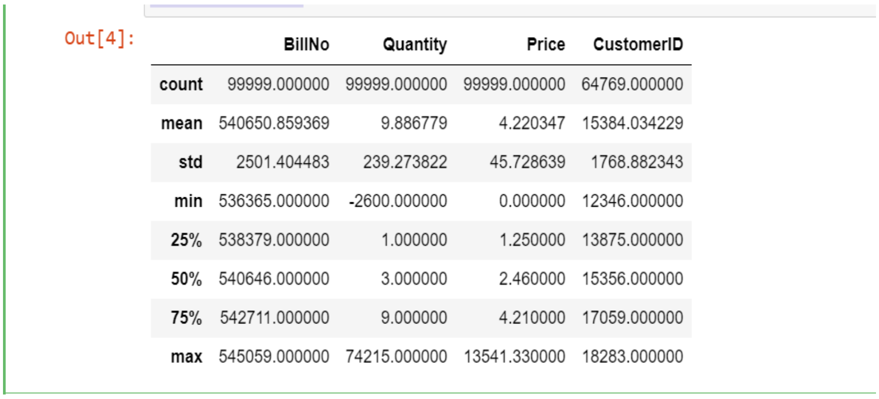
LOADING AND PREPROCESSING THE DATASET

```
import pandas as pd
data=pd.read_csv('assignment1_data.csv',on_bad_lines='skip')
print(data)
```



```
       BillNo                           Itemname  Quantity  \
0      536365   WHITE HANGING HEART T-LIGHT HOLDER         6
1      536365                  WHITE METAL LANTERN         6
2      536365       CREAM CUPID HEARTS COAT HANGER         8
3      536365  KNITTED UNION FLAG HOT WATER BOTTLE         6
4      536365       RED WOOLLY HOTTIE WHITE HEART.         6
...       ...                                  ...       ...
99994  545059                RIBBON REEL HEARTS DESIGN         1
99995  545059               RIBBON REEL STRIPES DESIGN         1
99996  545059                  BROWN CHECK CAT DOORSTOP         1
99997  545059                         CARD PARTY GAMES        12
99998  545059                              WICKER STAR         4

                   Date  Price  CustomerID         Country
0      01-12-2010 08:26   2.55     17850.0  United Kingdom
1      01-12-2010 08:26   3.39     17850.0  United Kingdom
2      01-12-2010 08:26   2.75     17850.0  United Kingdom
3      01-12-2010 08:26   3.39     17850.0  United Kingdom
4      01-12-2010 08:26   3.39     17850.0  United Kingdom
...                 ...    ...         ...             ...
99994  27-02-2011 13:04   1.65     14157.0  United Kingdom
99995  27-02-2011 13:04   1.65     14157.0  United Kingdom
99996  27-02-2011 13:04   4.25     14157.0  United Kingdom
99997  27-02-2011 13:04   0.42     14157.0  United Kingdom
99998  27-02-2011 13:04   2.10     14157.0  United Kingdom

[99999 rows x 7 columns]
```

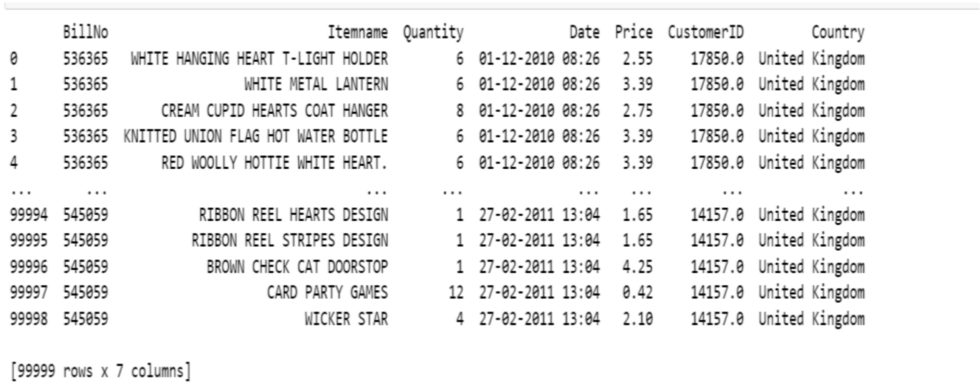
```
data.describe()
```



Out[4]:

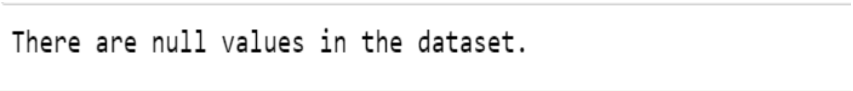
	BillNo	Quantity	Price	CustomerID
count	99999.000000	99999.000000	99999.000000	64769.000000
mean	540650.859369	9.886779	4.220347	15384.034229
std	2501.404483	239.273822	45.728639	1768.882343
min	536365.000000	-2600.000000	0.000000	12346.000000
25%	538379.000000	1.000000	1.250000	13875.000000
50%	540646.000000	3.000000	2.460000	15356.000000
75%	542711.000000	9.000000	4.210000	17059.000000
max	545059.000000	74215.000000	13541.330000	18283.000000

```
pd.set_option('display.max_columns', None)
pd.set_option('display.expand_frame_repr', False)
print(data)
```



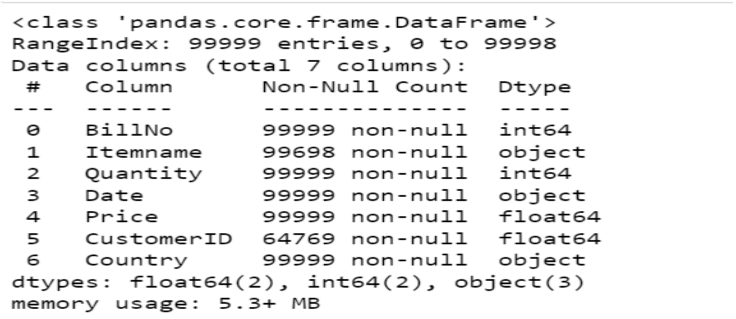
```
       BillNo                            Itemname  Quantity              Date  Price  CustomerID         Country
0      536365   WHITE HANGING HEART T-LIGHT HOLDER         6  01-12-2010 08:26   2.55     17850.0  United Kingdom
1      536365                  WHITE METAL LANTERN         6  01-12-2010 08:26   3.39     17850.0  United Kingdom
2      536365       CREAM CUPID HEARTS COAT HANGER         8  01-12-2010 08:26   2.75     17850.0  United Kingdom
3      536365  KNITTED UNION FLAG HOT WATER BOTTLE         6  01-12-2010 08:26   3.39     17850.0  United Kingdom
4      536365       RED WOOLLY HOTTIE WHITE HEART.         6  01-12-2010 08:26   3.39     17850.0  United Kingdom
...       ...                                  ...       ...               ...    ...         ...             ...
99994  545059            RIBBON REEL HEARTS DESIGN         1  27-02-2011 13:04   1.65     14157.0  United Kingdom
99995  545059           RIBBON REEL STRIPES DESIGN         1  27-02-2011 13:04   1.65     14157.0  United Kingdom
99996  545059             BROWN CHECK CAT DOORSTOP         1  27-02-2011 13:04   4.25     14157.0  United Kingdom
99997  545059                     CARD PARTY GAMES        12  27-02-2011 13:04   0.42     14157.0  United Kingdom
99998  545059                          WICKER STAR         4  27-02-2011 13:04   2.10     14157.0  United Kingdom

[99999 rows x 7 columns]
```

```
data = pd.read_csv('assignment1_data.csv')
null_values = data.isnull().any().any()
if null_values:
    print("There are null values in the dataset.")
else:
    print("There are no null values in the dataset.")
```



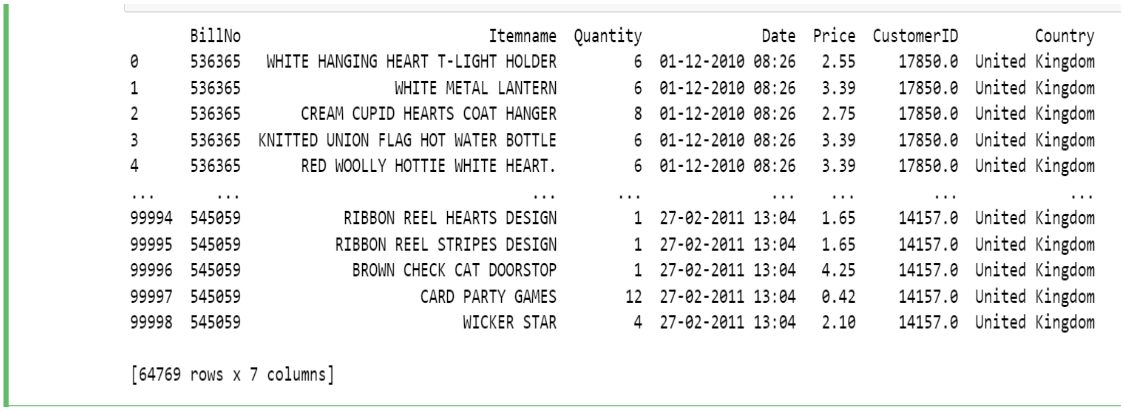
```
There are null values in the dataset.
```

```
data.info()
```



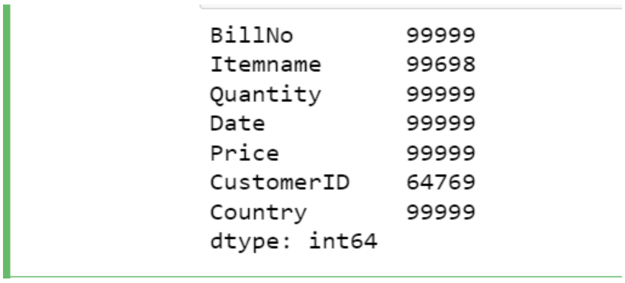
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99999 entries, 0 to 99998
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   BillNo      99999 non-null  int64  
 1   Itemname    99698 non-null  object 
 2   Quantity    99999 non-null  int64  
 3   Date        99999 non-null  object 
 4   Price       99999 non-null  float64
 5   CustomerID  64769 non-null  float64
 6   Country     99999 non-null  object 
dtypes: float64(2), int64(2), object(3)
memory usage: 5.3+ MB
```

print(data.dropna())



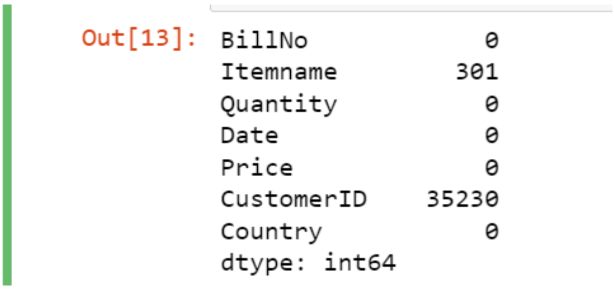
```
       BillNo                            Itemname  Quantity              Date  Price  CustomerID         Country
0      536365   WHITE HANGING HEART T-LIGHT HOLDER         6  01-12-2010 08:26   2.55     17850.0  United Kingdom
1      536365                  WHITE METAL LANTERN         6  01-12-2010 08:26   3.39     17850.0  United Kingdom
2      536365       CREAM CUPID HEARTS COAT HANGER         8  01-12-2010 08:26   2.75     17850.0  United Kingdom
3      536365  KNITTED UNION FLAG HOT WATER BOTTLE         6  01-12-2010 08:26   3.39     17850.0  United Kingdom
4      536365       RED WOOLLY HOTTIE WHITE HEART.         6  01-12-2010 08:26   3.39     17850.0  United Kingdom
...       ...                                  ...       ...               ...    ...         ...             ...
99994  545059             RIBBON REEL HEARTS DESIGN         1  27-02-2011 13:04   1.65     14157.0  United Kingdom
99995  545059            RIBBON REEL STRIPES DESIGN         1  27-02-2011 13:04   1.65     14157.0  United Kingdom
99996  545059              BROWN CHECK CAT DOORSTOP         1  27-02-2011 13:04   4.25     14157.0  United Kingdom
99997  545059                      CARD PARTY GAMES        12  27-02-2011 13:04   0.42     14157.0  United Kingdom
99998  545059                           WICKER STAR         4  27-02-2011 13:04   2.10     14157.0  United Kingdom

[64769 rows x 7 columns]
```

print(data.notnull().sum())



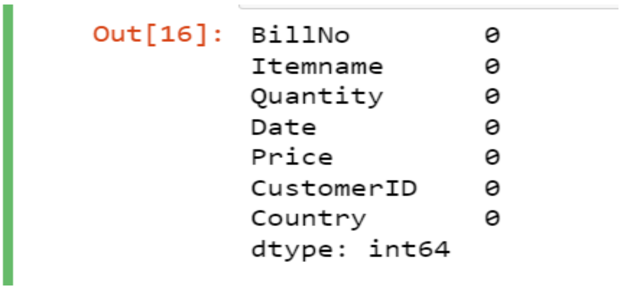
```
BillNo        99999
Itemname      99698
Quantity      99999
Date          99999
Price         99999
CustomerID    64769
Country       99999
dtype: int64
```

data.isnull().sum()



```
Out[13]: BillNo            0
         Itemname        301
         Quantity          0
         Date              0
         Price             0
         CustomerID    35230
         Country           0
         dtype: int64
```

```
data.dropna(subset=['Itemname', 'CustomerID'], inplace=True)

data.isnull().sum()
```



```
Out[16]: BillNo        0
         Itemname      0
         Quantity      0
         Date          0
         Price         0
         CustomerID    0
         Country       0
         dtype: int64
```