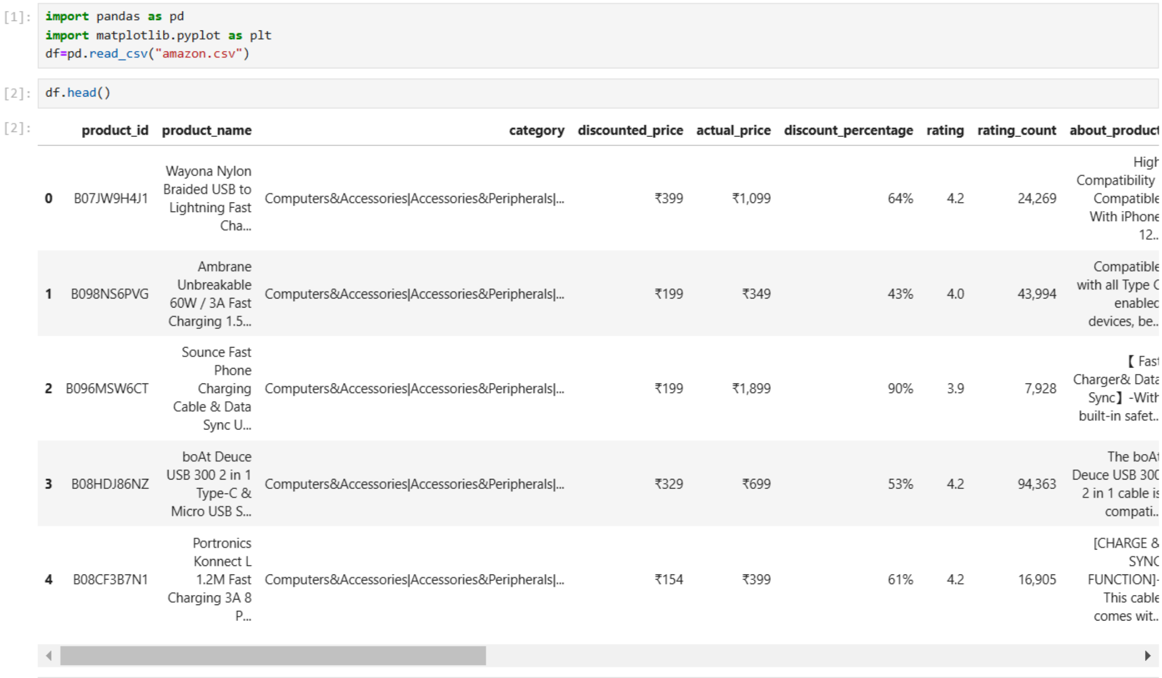
TASK-1: Data Cleaning and Pre-processing.

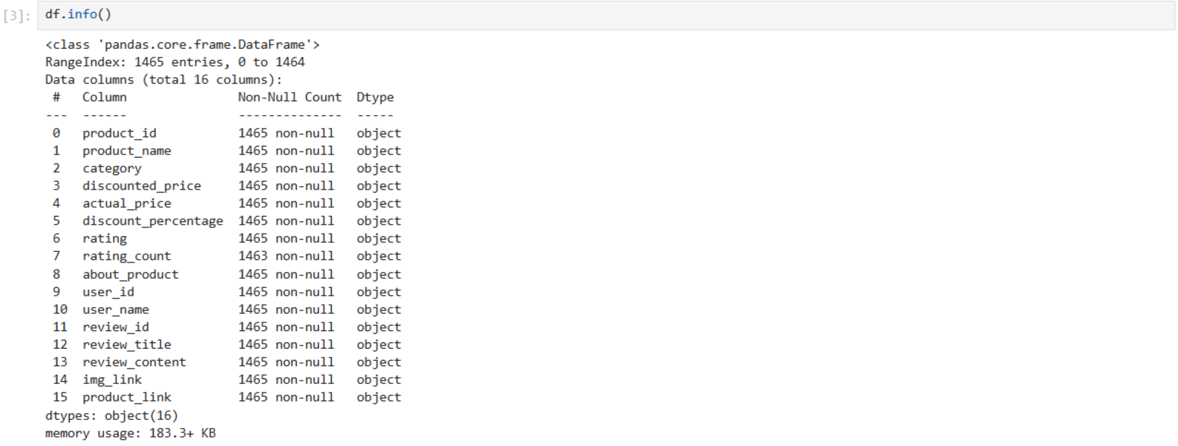


```
[1]: import pandas as pd
     import matplotlib.pyplot as plt
     df=pd.read_csv("amazon.csv")
```

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

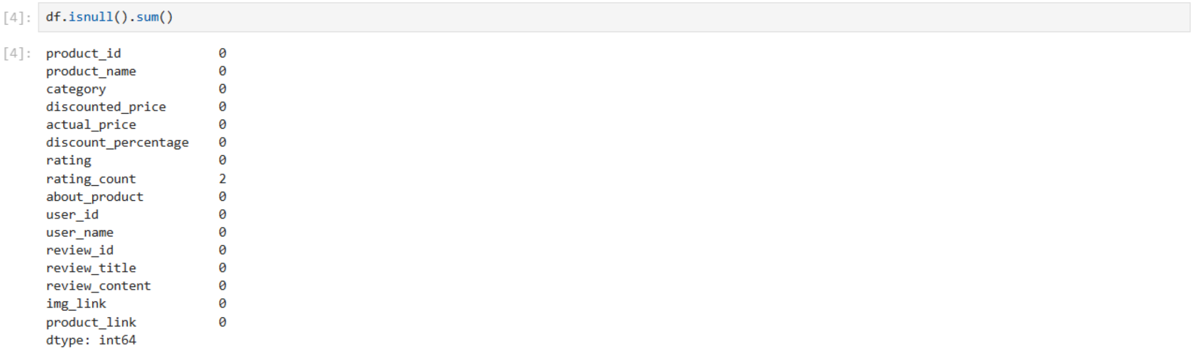
[2]:

	product_id	product_name	category	discounted_price	actual_price	discount_percentage	rating	rating_count	about_product
0	B07JW9H4J1	Wayona Nylon Braided USB to Lightning Fast Cha...	Computers&Accessories|Accessories&Peripherals|...	₹399	₹1,099	64%	4.2	24,269	High Compatibility Compatible With iPhone 12..
1	B098NS6PVG	Ambrane Unbreakable 60W / 3A Fast Charging 1.5...	Computers&Accessories|Accessories&Peripherals|...	₹199	₹349	43%	4.0	43,994	Compatible with all Type C enabled devices, be..
2	B096MSW6CT	Sounce Fast Phone Charging Cable & Data Sync U...	Computers&Accessories|Accessories&Peripherals|...	₹199	₹1,899	90%	3.9	7,928	【 Fast Charger& Data Sync】-With built-in safet..
3	B08HDJ86NZ	boAt Deuce USB 300 2 in 1 Type-C & Micro USB S...	Computers&Accessories|Accessories&Peripherals|...	₹329	₹699	53%	4.2	94,363	The boAt Deuce USB 300 2 in 1 cable is compati..
4	B08CF3B7N1	Portronics Konnect L 1.2M Fast Charging 3A 8 P...	Computers&Accessories|Accessories&Peripherals|...	₹154	₹399	61%	4.2	16,905	[CHARGE & SYNC FUNCTION]- This cable comes wit..



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

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1465 entries, 0 to 1464
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   product_id           1465 non-null   object
 1   product_name         1465 non-null   object
 2   category             1465 non-null   object
 3   discounted_price     1465 non-null   object
 4   actual_price         1465 non-null   object
 5   discount_percentage  1465 non-null   object
 6   rating               1465 non-null   object
 7   rating_count         1463 non-null   object
 8   about_product        1465 non-null   object
 9   user_id              1465 non-null   object
 10  user_name            1465 non-null   object
 11  review_id            1465 non-null   object
 12  review_title         1465 non-null   object
 13  review_content       1465 non-null   object
 14  img_link             1465 non-null   object
 15  product_link         1465 non-null   object
dtypes: object(16)
memory usage: 183.3+ KB
```



```
[4]: df.isnull().sum()

[4]: product_id             0
     product_name           0
     category               0
     discounted_price       0
     actual_price           0
     discount_percentage    0
     rating                 0
     rating_count           2
     about_product          0
     user_id                0
     user_name              0
     review_id              0
     review_title           0
     review_content         0
     img_link               0
     product_link           0
     dtype: int64
```

```
dtype: int64
```

```
[7]: df.drop_duplicates()
```

[7]:

	product_id	product_name	category	discounted_price	actual_price	discount_percentage	rating	rating_count	about_p
0	B07JW9H4J1	Wayona Nylon Braided USB to Lightning Fast Cha...	Computers&Accessories|Accessories&Peripherals|...	₹399	₹1,099	64%	4.2	24,269	Compa Con With
1	B098NS6PVG	Ambrane Unbreakable 60W / 3A Fast Charging 1.5...	Computers&Accessories|Accessories&Peripherals|...	₹199	₹349	43%	4.0	43,994	Con with al devic
2	B096MSW6CT	Sounce Fast Phone Charging Cable & Data Sync U...	Computers&Accessories|Accessories&Peripherals|...	₹199	₹1,899	90%	3.9	7,928	Charge Sync built-i
3	B08HDJ86NZ	boAt Deuce USB 300 2 in 1 Type-C & Micro USB S...	Computers&Accessories|Accessories&Peripherals|...	₹329	₹699	53%	4.2	94,363	T Deuce U 2 in 1 co

```
[8]: df['discounted_price'] = df['discounted_price'].str.replace('₹','')
     df['discounted_price'] = df['discounted_price'].str.replace(',','')
     df['discounted_price'] = df['discounted_price'].astype('float64')
```

```
[9]: df['actual_price'] = df['actual_price'].str.replace('₹','')
     df['actual_price'] = df['actual_price'].str.replace(',','')
     df['actual_price'] = df['actual_price'].astype('float64')
```

```
[10]: df['discount_percentage'] = df['discount_percentage'].str.replace('%','').astype('float64')
      df['discount_percentage'] = df['discount_percentage']/100
```

```
[12]: df['rating'].value_counts()
```

```
[12]: rating
      4.1    244
      4.3    230
      4.2    228
      4.0    129
      3.9    123
      4.4    123
      3.8     86
      4.5     75
      4       52
      3.7     42
      3.6     35
      3.5     26
      4.6     17
      3.3     16
      3.4     10
      4.7      6
      3.1      4
      5.0      3
      3.0      3
```

```
[14]: df['rating'] = df['rating'].str.replace('|','3.9').astype('float64')
```

```
[15]: df['rating_count'] = df['rating_count'].str.replace(',','').astype('float64')
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1465 entries, 0 to 1464
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   product_id           1465 non-null   object 
 1   product_name         1465 non-null   object 
 2   category             1465 non-null   object 
 3   discounted_price     1465 non-null   float64
 4   actual_price         1465 non-null   float64
 5   discount_percentage  1465 non-null   float64
 6   rating               1465 non-null   float64
 7   rating_count         1463 non-null   float64
 8   about_product        1465 non-null   object 
 9   user_id              1465 non-null   object 
 10  user_name            1465 non-null   object 
 11  review_id            1465 non-null   object 
 12  review_title         1465 non-null   object 
 13  review_content       1465 non-null   object 
 14  img_link             1465 non-null   object 
 15  product_link         1465 non-null   object 
dtypes: float64(5), object(11)
memory usage: 183.3+ KB
```

```
[17]: df.isnull().sum()
```

```
[17]: product_id             0
      product_name           0
      category               0
      discounted_price       0
      actual_price           0
      discount_percentage    0
      rating                 0
      rating_count           2
      about_product          0
      user_id                0
      user_name              0
      review_id              0
      review_title           0
      review_content         0
      img_link               0
      product_link           0
      dtype: int64
```

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

[18]:

	discounted_price	actual_price	discount_percentage	rating	rating_count
count	1465.000000	1465.000000	1465.000000	1465.000000	1463.000000
mean	3125.310874	5444.990635	0.476915	4.096451	18295.541353
std	6944.304394	10874.826864	0.216359	0.291620	42753.864952
min	39.000000	39.000000	0.000000	2.000000	2.000000
25%	325.000000	800.000000	0.320000	4.000000	1186.000000
50%	799.000000	1650.000000	0.500000	4.100000	5179.000000
75%	1999.000000	4295.000000	0.630000	4.300000	17336.500000
max	77990.000000	139900.000000	0.940000	5.000000	426973.000000

```
[19]: round(df.isnull().sum() / len(df) * 100, 2).sort_values(ascending=False)
```

```
[19]: rating_count           0.14
      product_id             0.00
      product_name           0.00
      category               0.00
      discounted_price       0.00
      actual_price           0.00
      discount_percentage    0.00
      rating                 0.00
      about_product          0.00
      user_id                0.00
      user_name              0.00
      review_id              0.00
      review_title           0.00
      review_content         0.00
      img_link               0.00
      product_link           0.00
      dtype: float64
```

```
[20]: df[df['rating_count'].isnull()].head(5)
```

[20]:

ame	category	discounted_price	actual_price	discount_percentage	rating	rating_count	about_product	
rand 65W ging ide...	Computers&Accessories|Accessories&Peripherals|...	199.0	999.0	0.80	3.0	NaN	USB C to C Cable: This cable has type C connec...	AE7CFHY23VAJT2FI4N
JSB- ning 3FT, ple...	Computers&Accessories|Accessories&Peripherals|...	249.0	999.0	0.75	5.0	NaN	💎 [The Fastest Charge] - This iPhone USB C cabl...	AGJC5O5H5BBXWUV7WI

```
[21]: df['rating_count']=df.rating_count.fillna(value=df['rating_count'].median())
```

```
[22]: df.isnull().sum()
```

```
[22]: product_id             0
      product_name           0
      category               0
      discounted_price       0
      actual_price           0
      discount_percentage    0
      rating                 0
      rating_count           0
      about_product          0
      user_id                0
      user_name              0
      review_id              0
      review_title           0
      review_content         0
      img_link               0
      product_link           0
      dtype: int64
```

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

```
[24]: 0
```

```
[28]: df.dtypes
```

```
[28]: product_id              object
      product_name            object
      category                object
      discounted_price       float64
      actual_price           float64
      discount_percentage    float64
      rating                 float64
      rating_count           float64
      about_product           object
      user_id                 object
      user_name               object
      review_id               object
      review_title            object
      review_content          object
      img_link                object
      product_link            object
      dtype: object
```

```
[ ]:
```