Amazon_Sales

February 20, 2024

```
[ ]: !pip install numpy
     !pip install pandas
     !pip install matplotlib
     !pip install seaborn
     !pip install pandas scipy
```

Requirement already satisfied: numpy in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (1.26.3)
Requirement already satisfied: pandas in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (2.1.4)
Requirement already satisfied: numpy<2,>=1.23.2 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
pandas) (1.26.3)
Requirement already satisfied: python-dateutil>=2.8.2 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
pandas) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
pandas) (2023.4)
Requirement already satisfied: six>=1.5 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
python-dateutil>=2.8.2->pandas) (1.16.0)
Requirement already satisfied: matplotlib in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (3.8.2)
Requirement already satisfied: contourpy>=1.0.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib) (4.47.0)
Requirement already satisfied: kiwisolver>=1.3.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from

matplotlib) (1.4.5)
Requirement already satisfied: numpy<2,>=1.21 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib) (1.26.3)
Requirement already satisfied: packaging>=20.0 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib) (23.2)
Requirement already satisfied: pillow>=8 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib) (10.2.0)
Requirement already satisfied: pyparsing>=2.3.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib) (2.8.2)
Requirement already satisfied: six>=1.5 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
python-dateutil>=2.7->matplotlib) (1.16.0)
Requirement already satisfied: seaborn in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (0.13.1)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
seaborn) (1.26.3)
Requirement already satisfied: pandas>=1.2 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
seaborn) (2.1.4)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
seaborn) (3.8.2)
Requirement already satisfied: contourpy>=1.0.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (1.2.0)
Requirement already satisfied: cycler>=0.10 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (4.47.0)
Requirement already satisfied: kiwisolver>=1.3.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (1.4.5)
Requirement already satisfied: packaging>=20.0 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (23.2)
Requirement already satisfied: pillow>=8 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (10.2.0)

```
Requirement already satisfied: pyparsing>=2.3.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
pandas>=1.2->seaborn) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
pandas>=1.2->seaborn) (2023.4)
Requirement already satisfied: six>=1.5 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
python-dateutil>=2.7->matplotlib!=3.6.1,>=3.4->seaborn) (1.16.0)
Requirement already satisfied: pandas in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (2.1.4)
Requirement already satisfied: scipy in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (1.11.4)
Requirement already satisfied: numpy<2,>=1.23.2 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
pandas) (1.26.3)
Requirement already satisfied: python-dateutil>=2.8.2 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
pandas) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
pandas) (2023.4)
Requirement already satisfied: six>=1.5 in
/Users/hamidhooshmandi/anaconda3/envs/USD/lib/python3.11/site-packages (from
python-dateutil>=2.8.2->pandas) (1.16.0)
```

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('amazon.csv')
```

```python
df.columns
```

```
Index(['product_id', 'product_name', 'category', 'discounted_price',
       'actual_price', 'discount_percentage', 'rating', 'rating_count',
       'about_product', 'user_id', 'user_name', 'review_id', 'review_title',
       'review_content', 'img_link', 'product_link'],
```

```
        dtype='object')
```

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

```
[ ]: product_id            object
     product_name          object
     category              object
     discounted_price      object
     actual_price          object
     discount_percentage   object
     rating                object
     rating_count          object
     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
```

```
[ ]: print(df['category'])
```

```
     0       Computers&Accessories|Accessories&Peripherals|…
     1       Computers&Accessories|Accessories&Peripherals|…
     2       Computers&Accessories|Accessories&Peripherals|…
     3       Computers&Accessories|Accessories&Peripherals|…
     4       Computers&Accessories|Accessories&Peripherals|…
                                  …
     1460    Home&Kitchen|Kitchen&HomeAppliances|WaterPurif…
     1461    Home&Kitchen|Kitchen&HomeAppliances|SmallKitch…
     1462    Home&Kitchen|Heating,Cooling&AirQuality|RoomHe…
     1463    Home&Kitchen|Heating,Cooling&AirQuality|Fans|E…
     1464    Home&Kitchen|Kitchen&HomeAppliances|SmallKitch…
     Name: category, Length: 1465, dtype: object
```

```python
[ ]: # split category column into multiple columns on '|' delimiter.
     catsplit = df['category'].str.split('|', expand=True)


     # copy selected columns to a new dataframe df1.
     df1 = df[['product_id', 'product_name', 'category', 'discounted_price',␣
      ↪'actual_price', 'discount_percentage', 'rating', 'rating_count']].copy()
     catsplit = catsplit.rename(columns={0:'category_1', 1:'category_2', 2:
      ↪'category_3', 3:'category_4', 4:'category_5'})
```

```python
# rename and assign split categories to df1 as category_1, category_2 etc.
df1['category_1'] = catsplit['category_1']
df1['category_2'] = catsplit['category_2']
df1['category_3'] = catsplit['category_3']
df1['category_4'] = catsplit['category_4']

# drop original category column from df1.
df1.drop(columns='category', inplace=True)
df1
```

[ ]:   product_id   product_name \
0   B07JW9H4J1 Wayona Nylon Braided USB to Lightning Fast Cha…
1   B098NS6PVG Ambrane Unbreakable 60W / 3A Fast Charging 1.5…
2   B096MSW6CT Sounce Fast Phone Charging Cable & Data Sync U…
3   B08HDJ86NZ boAt Deuce USB 300 2 in 1 Type-C & Micro USB S…
4   B08CF3B7N1 Portronics Konnect L 1.2M Fast Charging 3A 8 P…
…   …   …
1460 B08L7J3T31 Noir Aqua - 5pcs PP Spun Filter + 1 Spanner | …
1461 B01M6453MB Prestige Delight PRWO Electric Rice Cooker (1 …
1462 B009P2LIL4 Bajaj Majesty RX10 2000 Watts Heat Convector R…
1463 B00J5DYCCA Havells Ventil Air DSP 230mm Exhaust Fan (Pist…
1464 B01486F4G6 Borosil Jumbo 1000-Watt Grill Sandwich Maker (…

      discounted_price actual_price discount_percentage rating rating_count \
0   399   1,099   64%   4.2   24,269
1   199   349   43%   4.0   43,994
2   199   1,899   90%   3.9   7,928
3   329   699   53%   4.2   94,363
4   154   399   61%   4.2   16,905
…   …   …   …   …   …
1460   379   919   59%   4   1,090
1461   2,280   3,045   25%   4.1   4,118
1462   2,219   3,080   28%   3.6   468
1463   1,399   1,890   26%   4   8,031
1464   2,863   3,690   22%   4.3   6,987

                 category_1   category_2 \
0   Computers&Accessories   Accessories&Peripherals
1   Computers&Accessories   Accessories&Peripherals
2   Computers&Accessories   Accessories&Peripherals
3   Computers&Accessories   Accessories&Peripherals
4   Computers&Accessories   Accessories&Peripherals
…   …   …
1460   Home&Kitchen   Kitchen&HomeAppliances
1461   Home&Kitchen   Kitchen&HomeAppliances
1462   Home&Kitchen   Heating,Cooling&AirQuality
1463   Home&Kitchen   Heating,Cooling&AirQuality

```
1464        Home&Kitchen        Kitchen&HomeAppliances


              category_3                    category_4
0         Cables&Accessories                    Cables
1         Cables&Accessories                    Cables
2         Cables&Accessories                    Cables
3         Cables&Accessories                    Cables
4         Cables&Accessories                    Cables
...                      ...                       ...
1460  WaterPurifiers&Accessories  WaterPurifierAccessories
1461      SmallKitchenAppliances         Rice&PastaCookers
1462              RoomHeaters             HeatConvectors
1463                     Fans                ExhaustFans
1464      SmallKitchenAppliances            SandwichMakers

[1465 rows x 11 columns]
```

```python
print(df1['category_1'].value_counts())
print("---------------------------------")
print(df1['category_2'].value_counts())
```

```
category_1
Electronics           526
Computers&Accessories   453
Home&Kitchen          448
OfficeProducts         31
MusicalInstruments      2
HomeImprovement         2
Toys&Games              1
Car&Motorbike           1
Health&PersonalCare     1
Name: count, dtype: int64
---------------------------------
category_2
Accessories&Peripherals              381
Kitchen&HomeAppliances               308
HomeTheater,TV&Video                 162
Mobiles&Accessories                  161
Heating,Cooling&AirQuality           116
WearableTechnology                    76
Headphones,Earbuds&Accessories        66
NetworkingDevices                     34
OfficePaperProducts                   27
ExternalDevices&DataStorage           18
Cameras&Photography                   16
HomeStorage&Organization              16
HomeAudio                             16
GeneralPurposeBatteries&BatteryChargers  14
```

```
Accessories                        14
Printers,Inks&Accessories          11
CraftMaterials                      7
Components                          5
OfficeElectronics                   4
Electrical                          2
Monitors                            2
Microphones                         2
Arts&Crafts                         1
PowerAccessories                    1
Tablets                             1
Laptops                             1
Kitchen&Dining                      1
CarAccessories                      1
HomeMedicalSupplies&Equipment       1
Name: count, dtype: int64
```
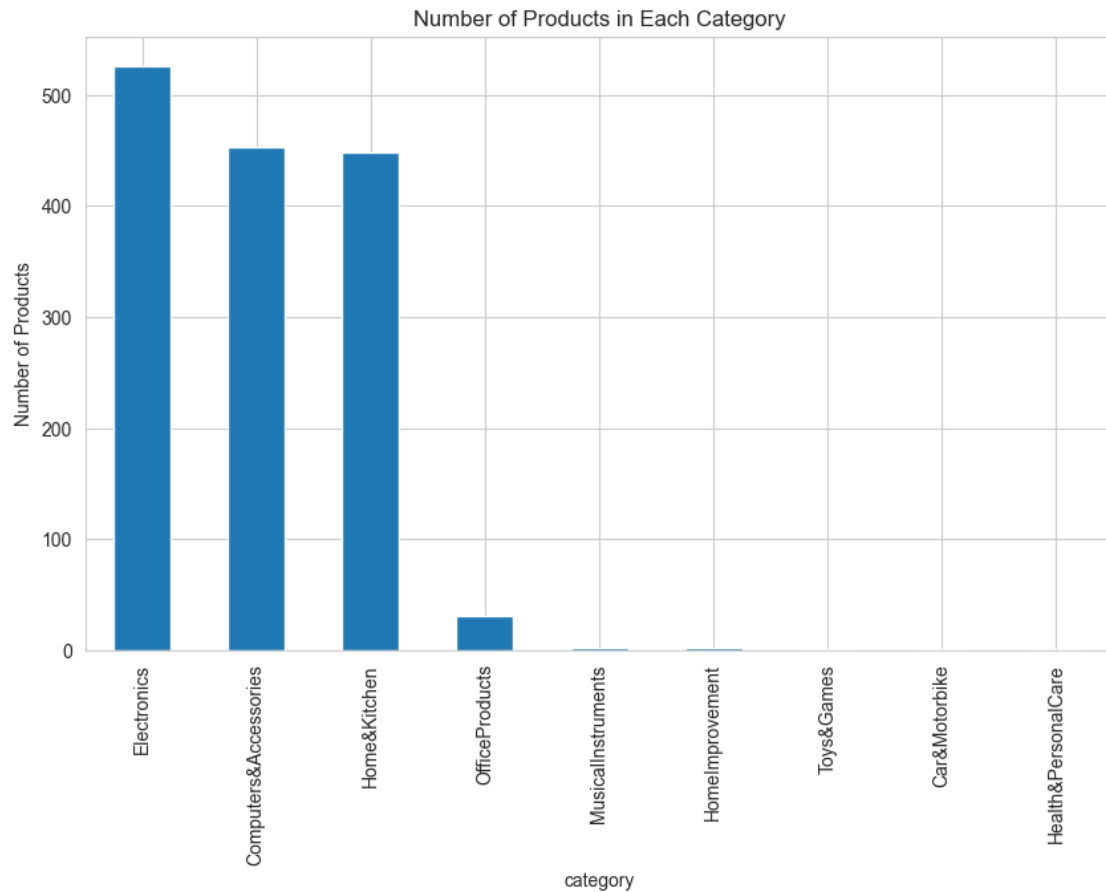
```python
import matplotlib.pyplot as plt

category_counts = df1['category_1'].value_counts()

category_counts.plot(kind='bar', figsize=(10, 6))

plt.title('Number of Products in Each Category')
plt.xlabel('category')
plt.ylabel('Number of Products')

plt.xticks(rotation='vertical')

plt.show()
```

Number of Products in Each Category


```
[ ]: df1['discount_percentage']
```

```
[ ]: 0        64%
     1        43%
     2        90%
     3        53%
     4        61%
            ...
     1460     59%
     1461     25%
     1462     28%
     1463     26%
     1464     22%
     Name: discount_percentage, Length: 1465, dtype: object
```

```
[ ]: df1['discount_percentage'] = df1['discount_percentage'].astype(str).str.
     ↪replace('%', '')
     df1['discount_percentage'] = pd.to_numeric(df1['discount_percentage'],␣
     ↪errors='coerce')
```

```
df1['discount_percentage'].mean()
```

[ ]: 47.69146757679181

[ ]:
```python
import matplotlib.pyplot as plt
import seaborn as sns

# Set the aesthetic style of the plots
sns.set_style("whitegrid")

# Create the histogram for the discount_percentage column
plt.figure(figsize=(10, 6))
sns.histplot(df1['discount_percentage'], kde=False, bins=30, color='blue')

plt.title('Distribution of Discount Percentages')
plt.xlabel('Discount Percentage')
plt.ylabel('Frequency')
plt.xticks(range(0, int(df1['discount_percentage'].max())+1, 5))

plt.show()
```
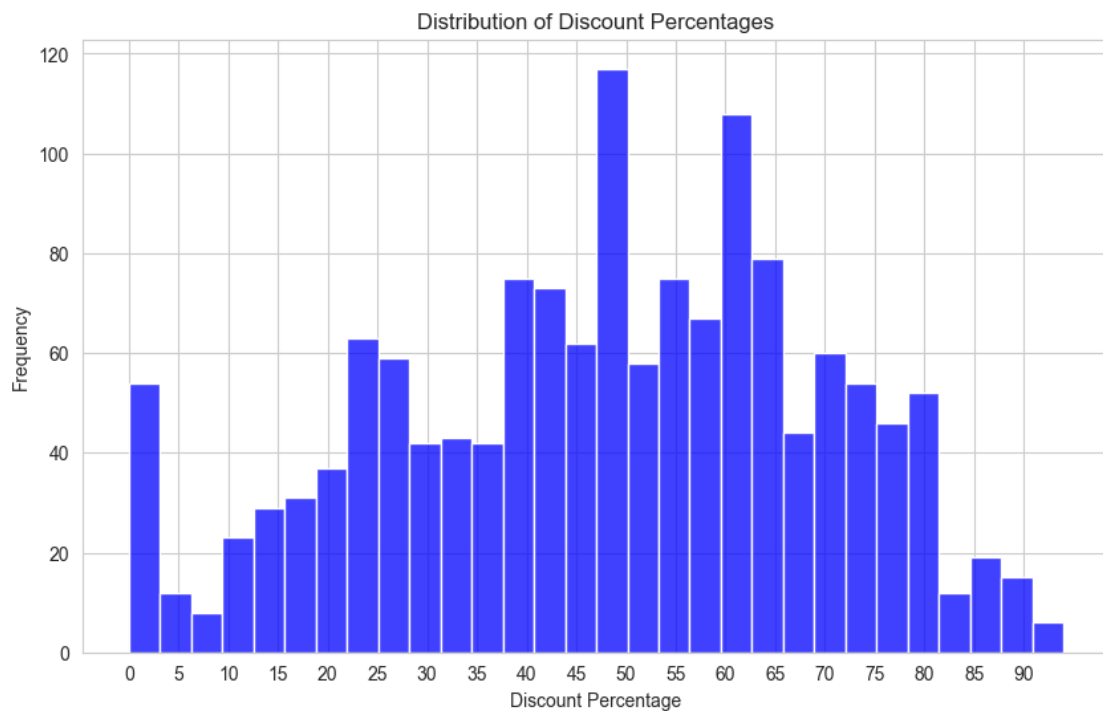

[ ]:
```python
#Finding unusual string in the rating column

df1['rating'].value_counts()
```

```
[ ]: 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
     4.8      3
     3.2      2
     2.8      2
     2.3      1
     |        1
     2        1
     3        1
     2.6      1
     2.9      1
     Name: count, dtype: int64
```

```
[ ]: df1['rating_count'] = df1['rating_count'].astype(str).str.replace(',', '')
     df1['rating_count'] = pd.to_numeric(df1['rating_count'], errors='coerce')
```

```
[ ]: df1.query('rating == "|"')
```

```
[ ]:         product_id                                         product_name  \
     1279  B08L12N5H1  Eureka Forbes car Vac 100 Watts Powerful Sucti…

           discounted_price actual_price  discount_percentage rating  rating_count  \
     1279             2,099        2,499                   16      |         992.0

                category_1              category_2                category_3  \
     1279  Home&Kitchen  Kitchen&HomeAppliances  Vacuum,Cleaning&Ironing

                  category_4
     1279  Vacuums&FloorCare
```

```python
# In Python, the float() function expects a dot (.) as the decimal separator.
df1['rating'] = df1['rating'].astype(str).str.replace(',', '')
df1['rating'] = pd.to_numeric(df1['rating'], errors='coerce')
```

```python
nan_cols = df1.columns[df1.isnull().any()].tolist()
# print(nan_cols)

nan_rows = df1[df1['rating_count'].isnull()]

df1['rating_count']
nan_rating_count = df1[df1['rating_count'].isnull()]

df1 = df1.dropna(subset=['rating_count'])
```

```python
#Changing the data type of discounted price and actual price

df1['discounted_price'] = df1['discounted_price'].map(lambda x: float(str(x).
  ↪replace(" ",'').replace(',', '')))
df1['actual_price'] = df1['actual_price'].map(lambda x: float(str(x).
  ↪replace(" ",'').replace(',', '')))
```

```
/var/folders/fy/x9y5w78j2x743gf22kp9g4cm0000gn/T/ipykernel_73223/373488213.py:3:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df1['discounted_price'] = df1['discounted_price'].map(lambda x:
float(str(x).replace(" ",'').replace(',', '')))
/var/folders/fy/x9y5w78j2x743gf22kp9g4cm0000gn/T/ipykernel_73223/373488213.py:4:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df1['actual_price'] = df1['actual_price'].map(lambda x:
float(str(x).replace(" ",'').replace(',', '')))
```

```python
df['rating_count'] = df['rating_count'].astype(str).str.replace(',', '')
df['rating_count'] = pd.to_numeric(df['rating_count'], errors='coerce')
```

```python
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import pandas as pd

df1 = df1.dropna()
```

```python
Xl = df1[['actual_price', 'rating', 'rating_count']]
yl = df1['discounted_price']

# Splitting the dataset into training and testing sets
Xl_train, Xl_test, yl_train, yl_test = train_test_split(Xl, yl,␣
 ↪random_state=21, test_size=0.3)


print("Number of rows in Xl_train: ", len(Xl_train))
print("Number of rows in Xl_test: ", len(Xl_test))
print("Number of rows in yl_train: ", len(yl_train))
print("Number of rows in yl_test: ", len(yl_test))

# Initializing and fitting the Linear Regression model
linear_model = LinearRegression()
linear_model.fit(Xl_train, yl_train)

linear_predict = linear_model.predict(Xl_test)

# Evaluating the model
linear_score = linear_model.score(Xl_test, yl_test)
print(f'Linear Regression score: {linear_score}')

print(f'Coefficients: {linear_model.coef_}')
```

```
Number of rows in Xl_train:  907
Number of rows in Xl_test:  390
Number of rows in yl_train:  907
Number of rows in yl_test:  390
Linear Regression score: 0.9516571091043098
Coefficients: [6.19394021e-01 8.28883156e+01 2.11426340e-03]
```
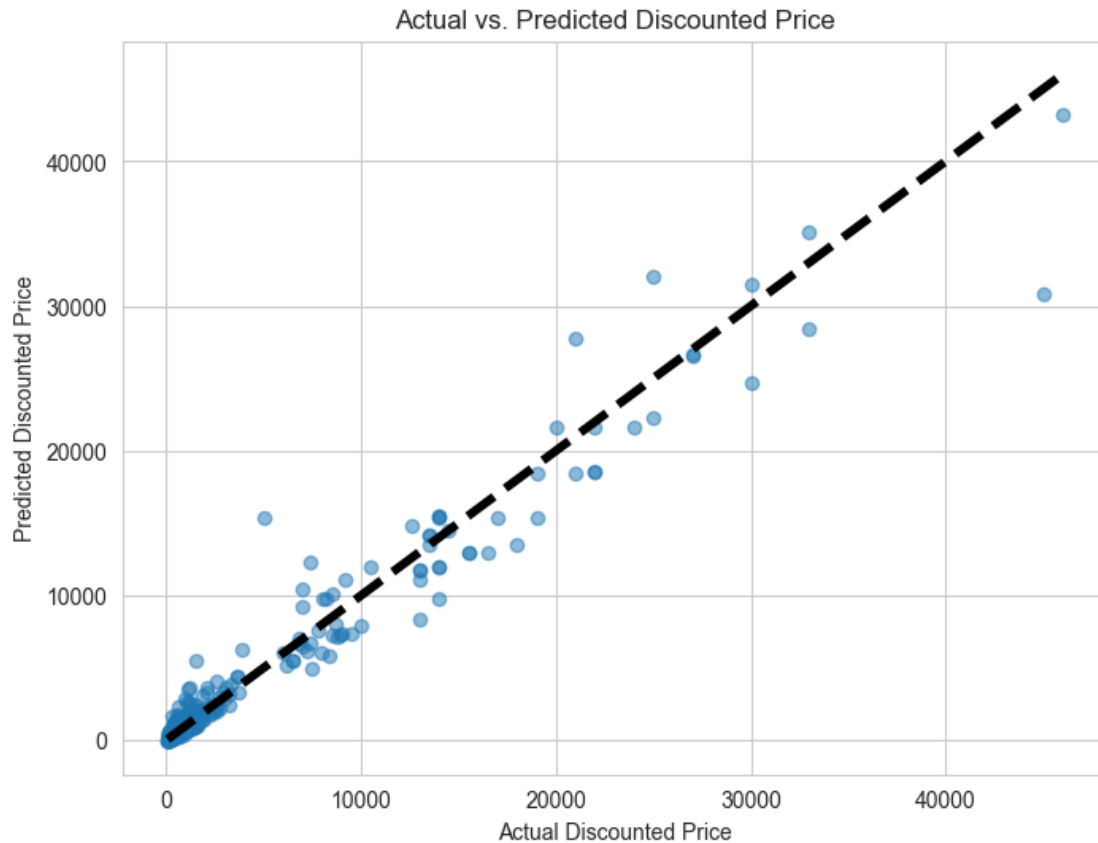
```python
[ ]: import matplotlib.pyplot as plt

plt.figure(figsize=(8, 6))
plt.scatter(yl_test, linear_predict, alpha=0.5)
plt.plot([yl_test.min(), yl_test.max()], [yl_test.min(), yl_test.max()], 'k--',␣
 ↪lw=4)
plt.xlabel('Actual Discounted Price')
plt.ylabel('Predicted Discounted Price')
plt.title('Actual vs. Predicted Discounted Price')
plt.show()
```

Actual vs. Predicted Discounted Price


```
residuals = yl_test - linear_predict

plt.figure(figsize=(8, 6))
plt.scatter(linear_predict, residuals, alpha=0.5)
plt.hlines(y=0, xmin=linear_predict.min(), xmax=linear_predict.max(),␣
  ↪colors='red', linestyles='--')
plt.xlabel('Predicted Discounted Price')
plt.ylabel('Residuals')
plt.title('Residuals of Predicted Discounted Price')
plt.show()
```

Residuals of Predicted Discounted Price