Amazon Product dataset Analysis (2023)

August 21, 2023

[1]:
```
#dataset: Amazon Products Sales Dataset 2023
```

[2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import re
import seaborn as sns
```

[3]:
```
# Assuming the delimiter is a semicolon (;)
df = pd.read_csv(r"D:\Data Science\My All Project\Amazon Product dataset Analysis (2023 )\Amazon_Products_Sales_Dataset_2023.csv")
```

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

[4]:

	name	main_category	sub_category	image	link	ratings	no_of_ratings	discount_price	actual_price	Unnamed: 0
0	Lloyd 1.5 Ton 3 Star Inverter Split Ac (5 In 1…	appliances	Air Conditioners	https://m.media-amazon.com/images/I/31UISB90sY…	https://www.amazon.in/Lloyd-Inverter-Convertib…	4.2	2,255	32,999	58,990	NaN
1	LG 1.5 Ton 5 Star AI DUAL Inverter Split AC (C…	appliances	Air Conditioners	https://m.media-amazon.com/images/I/51JFb7FctD…	https://www.amazon.in/LG-Convertible-Anti-Viru…	4.2	2,948	46,490	75,990	NaN
2	LG 1 Ton 4 Star Ai Dual Inverter Split Ac (Cop…	appliances	Air Conditioners	https://m.media-amazon.com/images/I/51JFb7FctD…	https://www.amazon.in/LG-Inverter-Convertible-…	4.2	1,206			
3	LG 1.5 Ton 3 Star AI DUAL Inverter Split AC (C…	appliances	Air Conditioners	https://m.media-amazon.com/images/I/51JFb7FctD…	https://www.amazon.in/LG-Convertible-Anti-Viru…	4	69			
4	Carrier 1.5 Ton 3 Star Inverter Split AC (Copp…	appliances	Air Conditioners	https://m.media-amazon.com/images/I/41lrtqXPiW…	https://www.amazon.in/Carrier-Inverter-Split-C…	4.1	630			

```
2         34,490       61,990          NaN
3         37,990       68,990          NaN
4         34,490       67,790          NaN
```

0.0.1 Data Pre-processing

[5]:
```
df.columns
```

[5]:
```
Index(['name', 'main_category', 'sub_category', 'image', 'link', 'ratings',
       'no_of_ratings', 'discount_price', 'actual_price', 'Unnamed: 0'],
      dtype='object')
```

[6]:
```
#drop some coloum
df.drop(['image', 'link', 'Unnamed: 0'], axis=1, inplace=True)
```

[7]:
```
df.columns #again Chek coloum
```

[7]:
```
Index(['name', 'main_category', 'sub_category', 'ratings', 'no_of_ratings',
       'discount_price', 'actual_price'],
      dtype='object')
```

[8]:
```
df.isnull().sum()
```

[8]:
```
name                  0
main_category         0
sub_category          0
ratings          332399
no_of_ratings    332399
discount_price   115687
actual_price      33968
dtype: int64
```

[9]:
```
df = df.dropna(subset=['actual_price']) #remove 'actual_price' null value
```

[10]:
```
# Define a function to extract the numeric value



def extract_numeric(text):
    match = re.search(r' ([\d,]+)', str(text))
    if match:
        numeric_value = match.group(1).replace(',', '')
        return int(numeric_value)
    return None

# Apply the function to the DataFrame column
df['discount_price'] = df['discount_price'].apply(extract_numeric)
```

```
def extract_numeric(text):
    match = re.search(r' ([\d,]+)', str(text))
    if match:
        numeric_value = match.group(1).replace(',', '')
        return int(numeric_value)
    return None

# Apply the function to the DataFrame column
df['actual_price'] = df['actual_price'].apply(extract_numeric)
```

[11]:
```
#Fill null value discount_price using median

med = df['discount_price'].median()
print(med)

df['discount_price'] = df['discount_price'].fillna(med)
```

```
694.0
```

[12]:
```
#Fill null value 'no_of_ratings' using mode

mod1=df['no_of_ratings'].mode()

mod1
df['no_of_ratings'] = df['no_of_ratings'].fillna(mod1)[1]
```

[13]:
```
#Fill null value 'ratings' using mode
mod2=df['ratings'].mode()

df['ratings'] = df['ratings'].fillna(mod2)[1]
```

[14]:
```
df.isnull().sum() #chek any null value
```

[14]:
```
name              0
main_category     0
sub_category      0
ratings           0
no_of_ratings     0
discount_price    0
actual_price      0
dtype: int64
```

[15]:
```
df.shape
```

[15]: (1014607, 7)

[16]:
```
df.dtypes
```

[16]:
```
name               object
main_category      object
sub_category       object
ratings            object
no_of_ratings      object
discount_price    float64
actual_price        int64
dtype: object
```

[17]:
```
# Convert "ratings" column to integers
df['ratings'] = df['ratings'].astype(float)

# Remove commas and convert "no_of_ratings" column to integers
df['no_of_ratings'] = df['no_of_ratings'].str.replace(',', '').astype(int)
```

[18]:
```
df.dtypes
```

[18]:
```
name               object
main_category      object
sub_category       object
ratings           float64
no_of_ratings       int32
discount_price    float64
actual_price        int64
dtype: object
```

0.0.2 EDA Analysis

[19]:
```
columns_to_correlate = ['discount_price', 'actual_price']

# Calculate the correlation coefficients
correlation_matrix = df[columns_to_correlate].corr()

# Show the output
print(correlation_matrix)
```

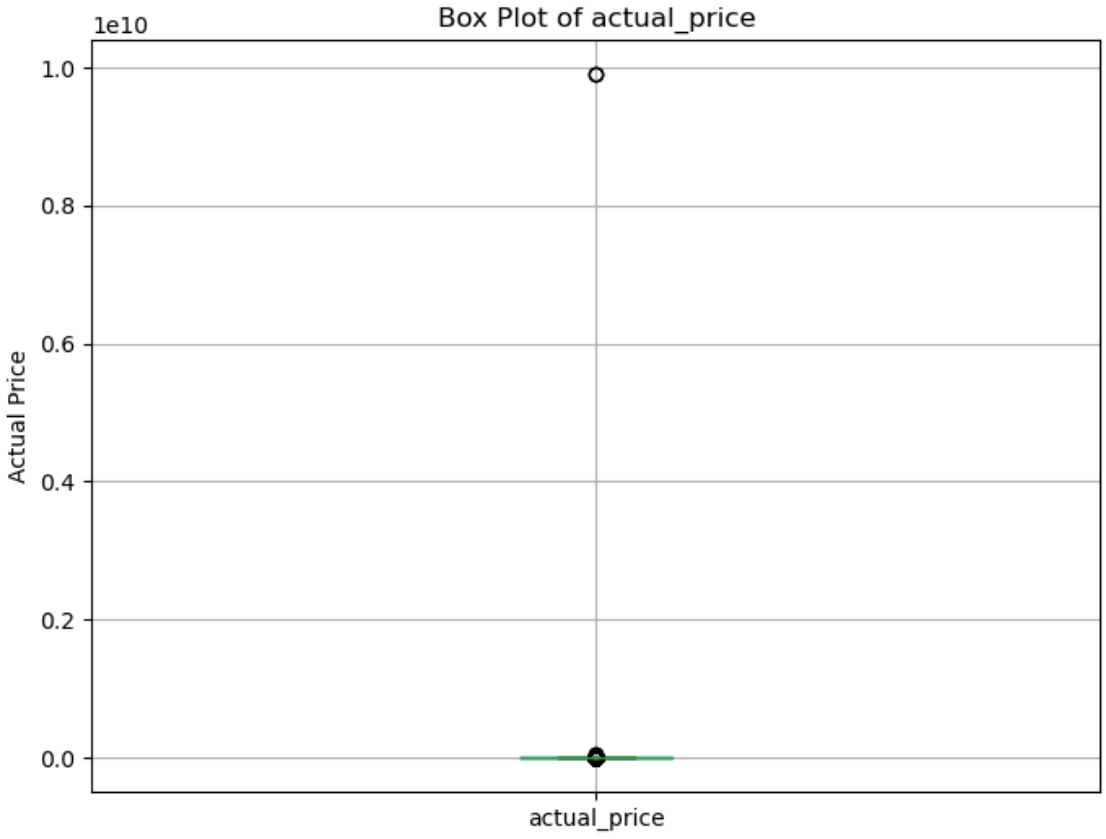
```
                discount_price  actual_price
discount_price        1.000000      0.000554
actual_price          0.000554      1.000000
```

[20]:
```
# Calculate the correlation coefficients
correlation_matrix = df[columns_to_correlate].describe()

# Show the output
```

```
print(correlation_matrix)
```

```
       discount_price  actual_price
count    1.014607e+06  1.014607e+06
mean     2.489725e+03  2.410906e+04
std      9.037429e+03  1.389982e+07
min      8.000000e+00  0.000000e+00
25%      3.990000e+02  9.900000e+02
50%      6.940000e+02  1.599000e+03
75%      1.299000e+03  3.000000e+03
max      1.249990e+06  9.900000e+09
```

[21]:
```
# Create a box plot for the "actual_price" column
plt.figure(figsize=(8, 6))  # Optional: Set the figure size
df.boxplot(column="actual_price")
plt.title("Box Plot of actual_price")
plt.ylabel("Actual Price")
plt.show()
```



[22]:
```
# Create a box plot for the "actual_price" column
plt.figure(figsize=(8, 6))  # Optional: Set the figure size
df.boxplot(column="discount_price")
plt.title("Box Plot of discount_price")
plt.ylabel("Discount Price")
plt.show()
```

[23]:
```
df.head()
```

[23]:

	name	main_category	sub_category	ratings	no_of_ratings	discount_price	actual_price
0	Lloyd 1.5 Ton 3 Star Inverter Split Ac (5 In 1…	appliances	Air Conditioners	4.2	2948	32999.0	58990
1	LG 1.5 Ton 5 Star AI DUAL Inverter Split AC (C…	appliances	Air Conditioners	4.2	2948	46490.0	75990
2	LG 1 Ton 4 Star Ai Dual Inverter Split Ac (Cop…	appliances	Air Conditioners	4.2	2948	34490.0	61990
3	LG 1.5 Ton 3 Star AI DUAL Inverter Split AC (C…	appliances					
4	Carrier 1.5 Ton 3 Star Inverter Split AC (Copp…	appliances					

```
3  Air Conditioners      4.2          2948         37990.0         68990
4  Air Conditioners      4.2          2948         34490.0         67790
```

[24]:
```
category_subcategory_counts = df.groupby('main_category')['main_category'].
 ↪value_counts().sum()
```

[25]:
```
print(category_subcategory_counts)
```

```
1014607
```

[36]:
```
category_subcategory_counts = df.groupby('main_category')['sub_category'].
 ↪value_counts().sort_values(ascending=False)

for main_category, sub_category_count in category_subcategory_counts.
 ↪groupby(level=0):
    print(f"{main_category} {sub_category_count.sum()}")
    for sub_category, count in sub_category_count.items():
        print(f"{sub_category} {count}")
```

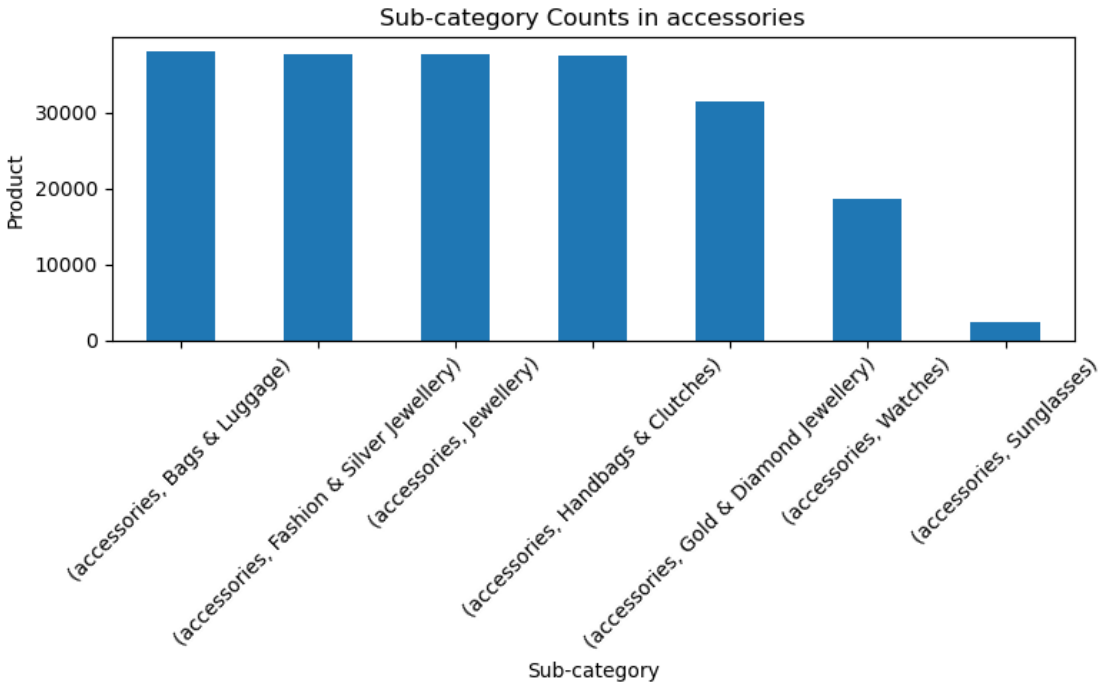
```
accessories 203662
('accessories', 'Bags & Luggage') 38000
('accessories', 'Fashion & Silver Jewellery') 37772
('accessories', 'Jewellery') 37654
('accessories', 'Handbags & Clutches') 37592
('accessories', 'Gold & Diamond Jewellery') 31486
('accessories', 'Watches') 18680
('accessories', 'Sunglasses') 2478
appliances 61499
('appliances', 'Kitchen & Home Appliances') 19006
('appliances', 'All Appliances') 18970
('appliances', 'Heating & Cooling Appliances') 18092
('appliances', 'Refrigerators') 3278
('appliances', 'Washing Machines') 1153
('appliances', 'Air Conditioners') 1000
bags & luggage 15100
('bags & luggage', 'Rucksacks') 4168
('bags & luggage', 'Backpacks') 4000
('bags & luggage', 'Suitcases & Trolley Bags') 2068
('bags & luggage', 'Wallets') 1975
('bags & luggage', 'Travel Accessories') 1802
('bags & luggage', 'Travel Duffles') 1087
beauty & health 19750
('beauty & health', 'Make-up') 5218
('beauty & health', 'Beauty & Grooming') 3830
('beauty & health', 'Diet & Nutrition') 2386
('beauty & health', 'Personal Care Appliances') 2206
('beauty & health', 'Health & Personal Care') 2196
('beauty & health', 'Household Supplies') 2154
```

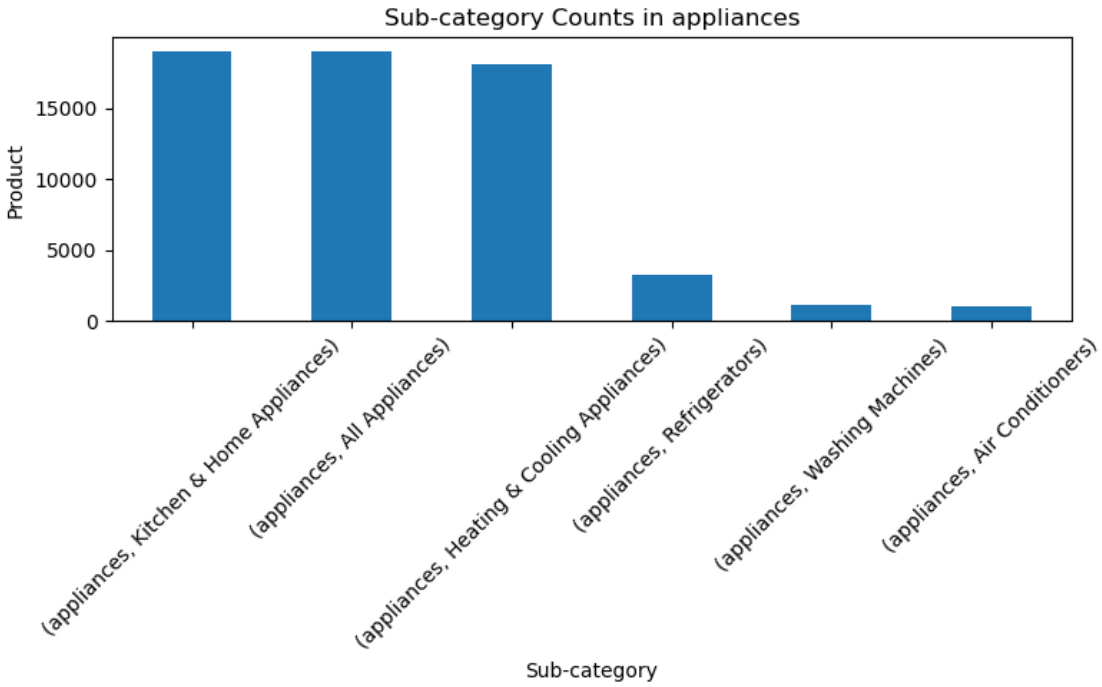
('beauty & health', 'Luxury Beauty') 1694
('beauty & health', 'Value Bazaar') 66
car & motorbike 13974
('car & motorbike', 'Car Accessories') 2782
('car & motorbike', 'All Car & Motorbike Products') 2530
('car & motorbike', 'Motorbike Accessories & Parts') 2430
('car & motorbike', 'Car Parts') 2424
('car & motorbike', 'Car Electronics') 1972
('car & motorbike', 'Car & Bike Care') 1836
grocery & gourmet foods 6564
('grocery & gourmet foods', 'Coffee, Tea & Beverages') 2568
('grocery & gourmet foods', 'Snack Foods') 2088
('grocery & gourmet foods', 'All Grocery & Gourmet Foods') 1908
home & kitchen 28946
('home & kitchen', 'Home Furnishing') 2822
('home & kitchen', 'Furniture') 2602
('home & kitchen', 'Home Décor') 2536
('home & kitchen', 'Sewing & Craft Supplies') 2518
('home & kitchen', 'Indoor Lighting') 2470
('home & kitchen', 'All Home & Kitchen') 2440
('home & kitchen', 'Home Storage') 2440
('home & kitchen', 'Bedroom Linen') 2428
('home & kitchen', 'Kitchen & Dining') 2394
('home & kitchen', 'Garden & Outdoors') 2196
('home & kitchen', 'Home Improvement') 2054
('home & kitchen', 'Kitchen Storage & Containers') 2046
home, kitchen, pets 34
('home, kitchen, pets', 'Refurbished & Open Box') 34
industrial supplies 6617
('industrial supplies', 'Lab & Scientific') 2188
('industrial supplies', 'Janitorial & Sanitation Supplies') 1786
('industrial supplies', 'Test, Measure & Inspect') 1403
('industrial supplies', 'Industrial & Scientific Supplies') 1240
kids' fashion 26412
("kids' fashion", "Kids' Fashion") 5432
("kids' fashion", 'School Bags') 4400
("kids' fashion", 'Baby Fashion') 4304
("kids' fashion", "Kids' Shoes") 4152
("kids' fashion", "Kids' Clothing") 4100
("kids' fashion", "Kids' Watches") 4024
men's clothing 146343
("men's clothing", 'Innerwear') 37834
("men's clothing", 'Shirts') 36674
("men's clothing", 'T-shirts & Polos') 36271
("men's clothing", 'Jeans') 35564
men's shoes 109726
("men's shoes", 'Sports Shoes') 36964
("men's shoes", 'Casual Shoes') 36770

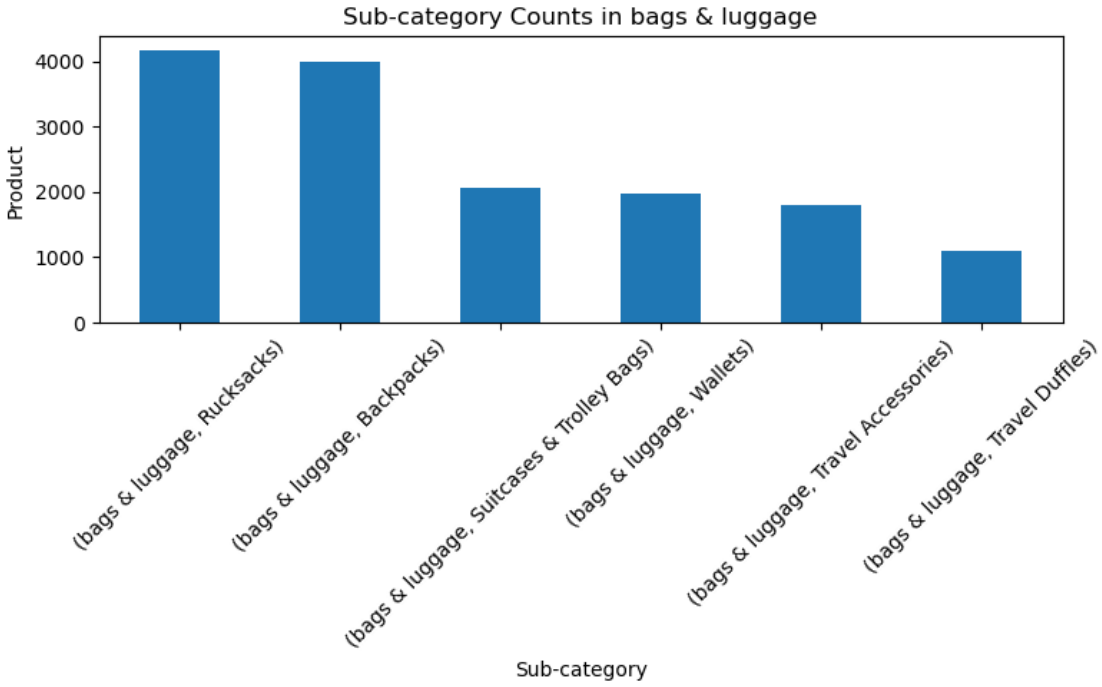
("men's shoes", 'Formal Shoes') 35992
music 2076
('music', 'Musical Instruments & Professional Audio') 2076
pet supplies 3238
('pet supplies', 'Dog supplies') 1956
('pet supplies', 'All Pet Supplies') 1282
sports & fitness 23419
('sports & fitness', 'All Sports, Fitness & Outdoors') 2440
('sports & fitness', 'Football') 2416
('sports & fitness', 'Fitness Accessories') 2414
('sports & fitness', 'Cricket') 2380
('sports & fitness', 'All Exercise & Fitness') 2338
('sports & fitness', 'Badminton') 2282
('sports & fitness', 'Cycling') 2234
('sports & fitness', 'Strength Training') 2176
('sports & fitness', 'Running') 1810
('sports & fitness', 'Camping & Hiking') 1450
('sports & fitness', 'Yoga') 1087
('sports & fitness', 'Cardio Equipment') 392
stores 60879
('stores', "Men's Fashion") 38088
('stores', 'Sportswear') 14230
('stores', 'Amazon Fashion') 4682
('stores', "Women's Fashion") 2101
('stores', 'The Designer Boutique') 1690
('stores', 'Fashion Sales & Deals') 88
toys & baby products 11264
('toys & baby products', 'Baby Bath, Skin & Grooming') 2760
('toys & baby products', 'Diapers') 2134
('toys & baby products', 'Baby Products') 2098
('toys & baby products', 'Nursing & Feeding') 2084
('toys & baby products', 'Strollers & Prams') 1118
('toys & baby products', 'Toys & Games') 908
('toys & baby products', 'STEM Toys Store') 92
('toys & baby products', 'International Toy Store') 46
('toys & baby products', 'Toys Gifting Store') 24
tv, audio & cameras 131981
('tv, audio & cameras', 'All Electronics') 19060
('tv, audio & cameras', 'Home Entertainment Systems') 18990
('tv, audio & cameras', 'Cameras') 18970
('tv, audio & cameras', 'Camera Accessories') 18940
('tv, audio & cameras', 'Speakers') 18596
('tv, audio & cameras', 'Headphones') 18544
('tv, audio & cameras', 'Security Cameras') 17446
('tv, audio & cameras', 'Televisions') 733
('tv, audio & cameras', 'Home Audio & Theater') 702
women's clothing 132609
("women's clothing", 'Clothing') 38066

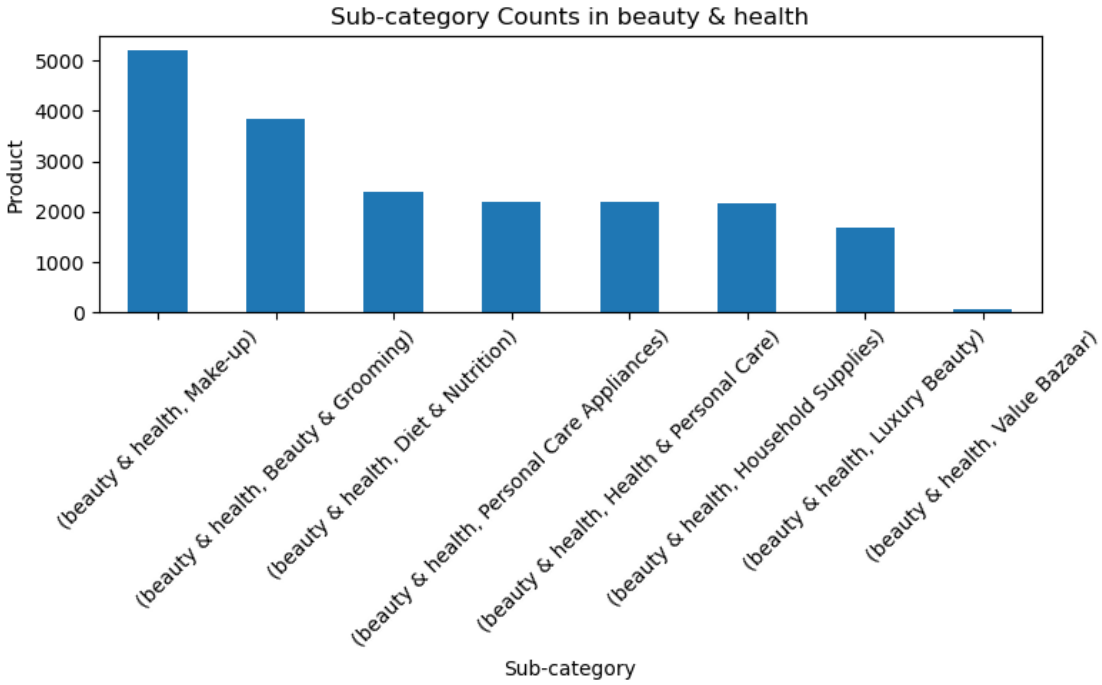
```
("women's clothing", 'Lingerie & Nightwear') 37990
("women's clothing", 'Ethnic Wear') 37538
("women's clothing", 'Western Wear') 19015
women's shoes 10514
("women's shoes", 'Fashion Sandals') 4266
("women's shoes", 'Shoes') 3678
("women's shoes", 'Ballerinas') 2570
```

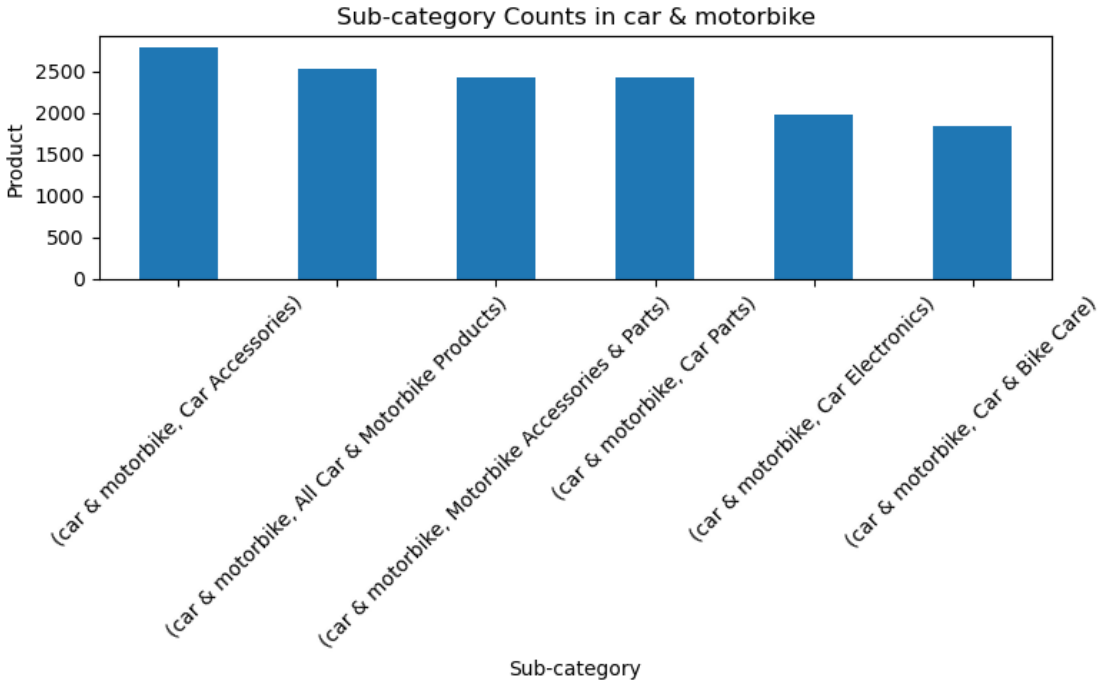
[27]:
```
category_subcategory_counts = df.groupby('main_category')['sub_category'].
 ↪value_counts()

for main_category, sub_category_count in category_subcategory_counts.
 ↪groupby(level=0):
    plt.figure(figsize=(8, 5))
    sub_category_count.plot(kind='bar')
    plt.title(f"Sub-category Counts in {main_category}")
    plt.xlabel('Sub-category')
    plt.ylabel('Product')
    plt.xticks(rotation=45)
    plt.tight_layout()
    plt.show()
```

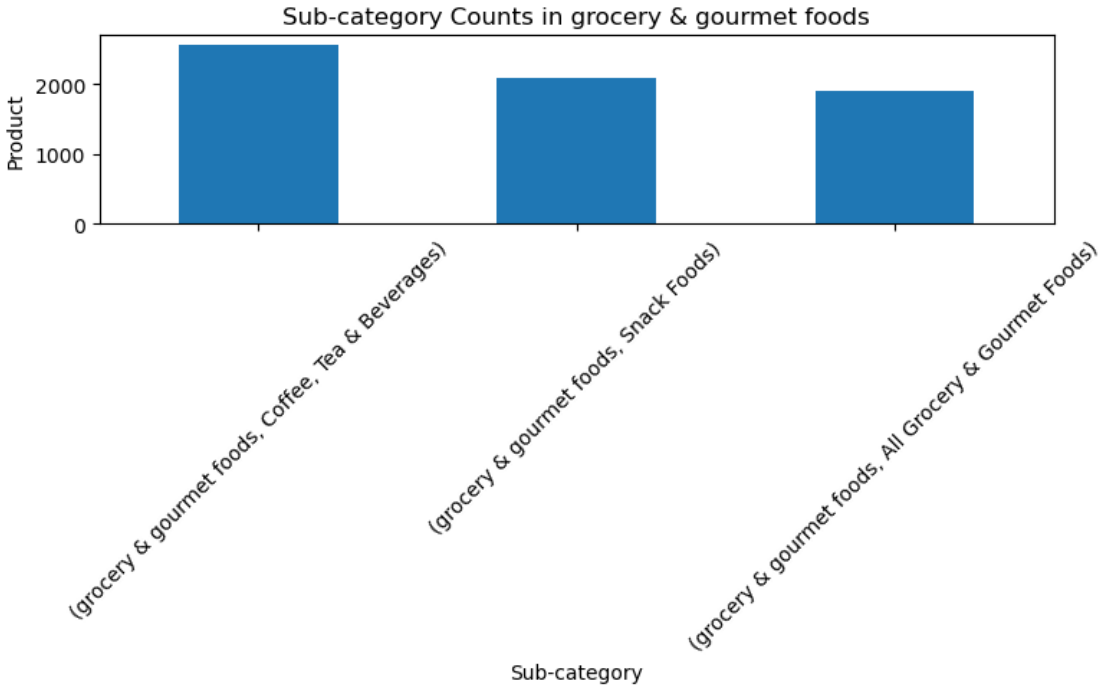


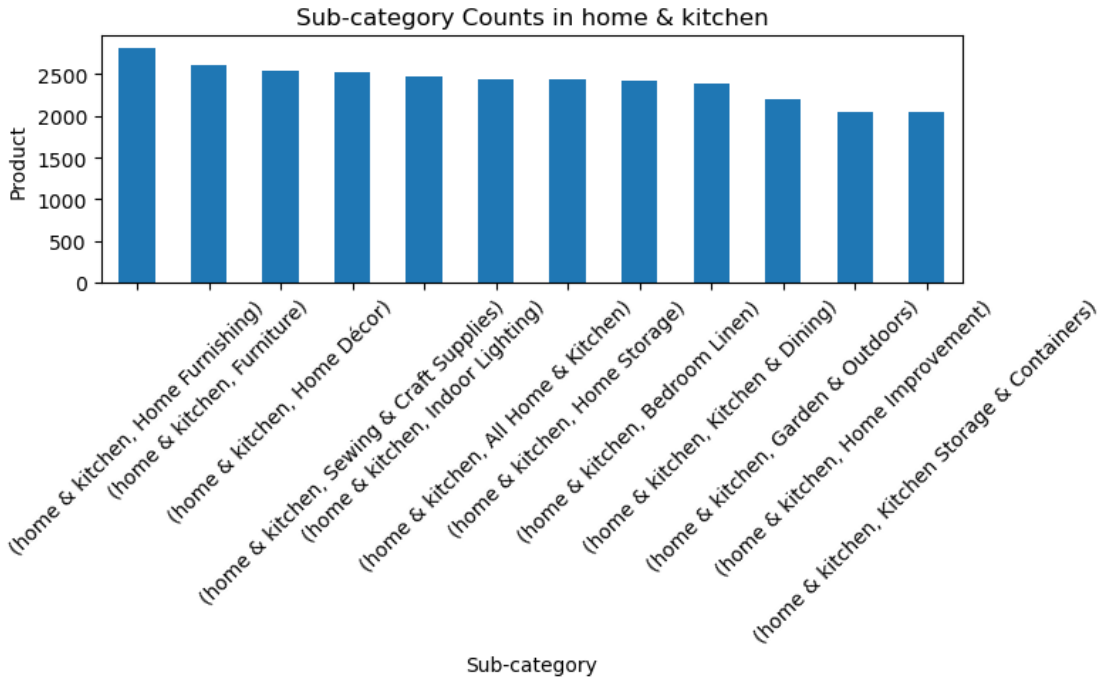

Sub-category Counts in appliances
Product
15000
10000
5000
0
(appliances, Kitchen & Home Appliances)
(appliances, All Appliances)
(appliances, Heating & Cooling Appliances)
(appliances, Refrigerators)
(appliances, Washing Machines)
(appliances, Air Conditioners)
Sub-category

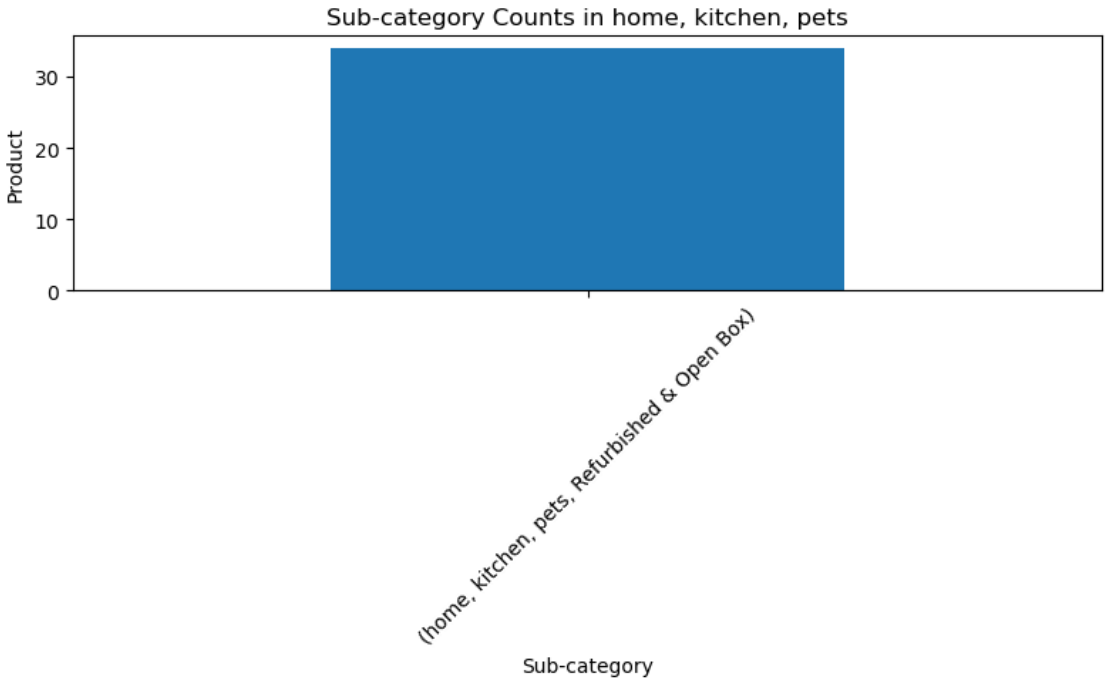

Sub-category Counts in bags & luggage
Product
4000
3000
2000
1000
0
(bags & luggage, Rucksacks)
(bags & luggage, Backpacks)
(bags & luggage, Suitcases & Trolley Bags)
(bags & luggage, Wallets)
(bags & luggage, Travel Accessories)
(bags & luggage, Travel Duffles)
Sub-category

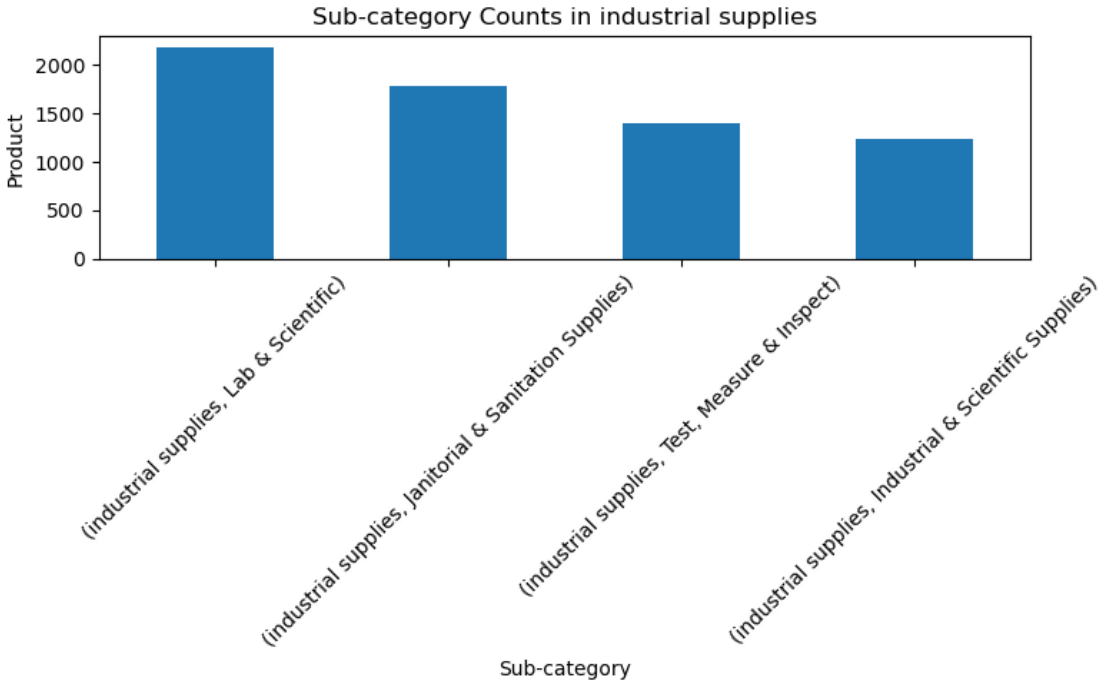

Sub-category Counts in beauty & health
Product
5000
4000
3000
2000
1000
0
(beauty & health, Make-up)
(beauty & health, Beauty & Grooming)
(beauty & health, Diet & Nutrition)
(beauty & health, Personal Care Appliances)
(beauty & health, Health & Personal Care)
(beauty & health, Household Supplies)
(beauty & health, Luxury Beauty)
(beauty & health, Value Bazaar)
Sub-category

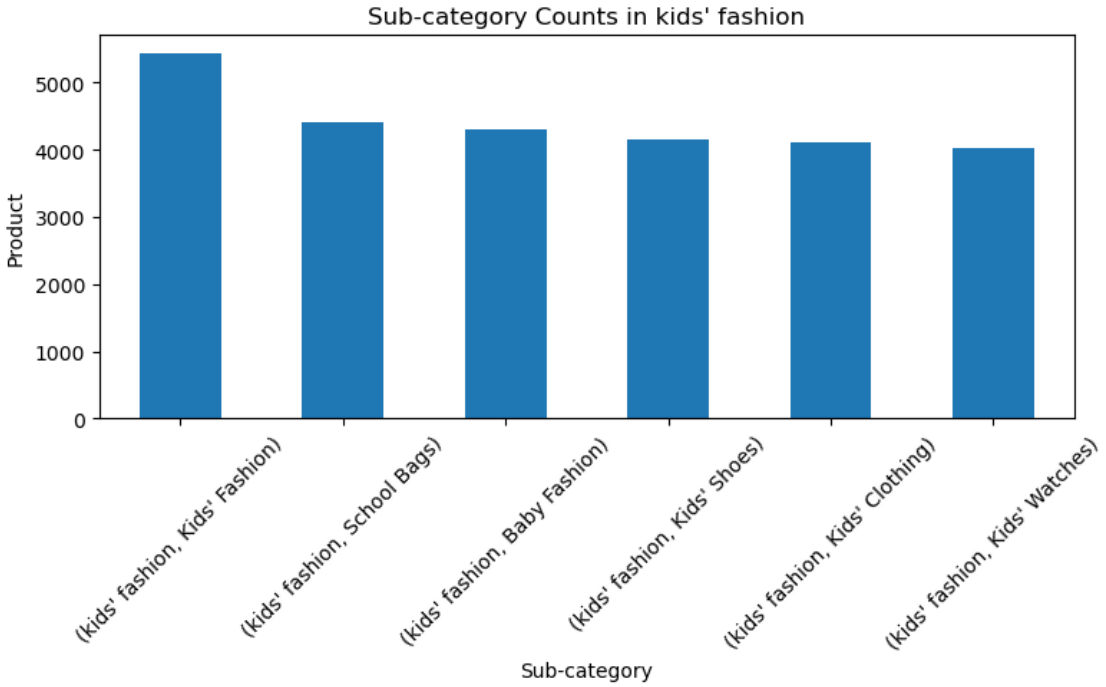

Sub-category Counts in car & motorbike
Product
2500
2000
1500
1000
500
0
(car & motorbike, Car Accessories)
(car & motorbike, All Car & Motorbike Products)
(car & motorbike, Motorbike Accessories & Parts)
(car & motorbike, Car Parts)
(car & motorbike, Car Electronics)
(car & motorbike, Car & Bike Care)
Sub-category

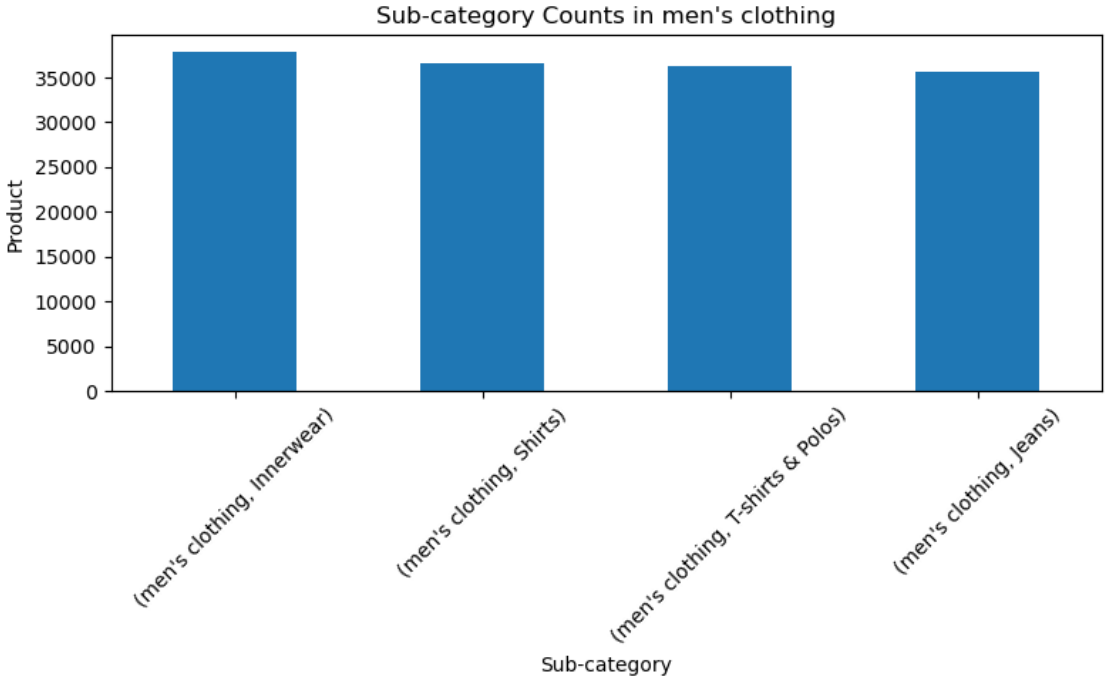

Sub-category Counts in grocery & gourmet foods
Product
2000
1000
0
(grocery & gourmet foods, Coffee, Tea & Beverages)
(grocery & gourmet foods, Snack Foods)
(grocery & gourmet foods, All Grocery & Gourmet Foods)
Sub-category

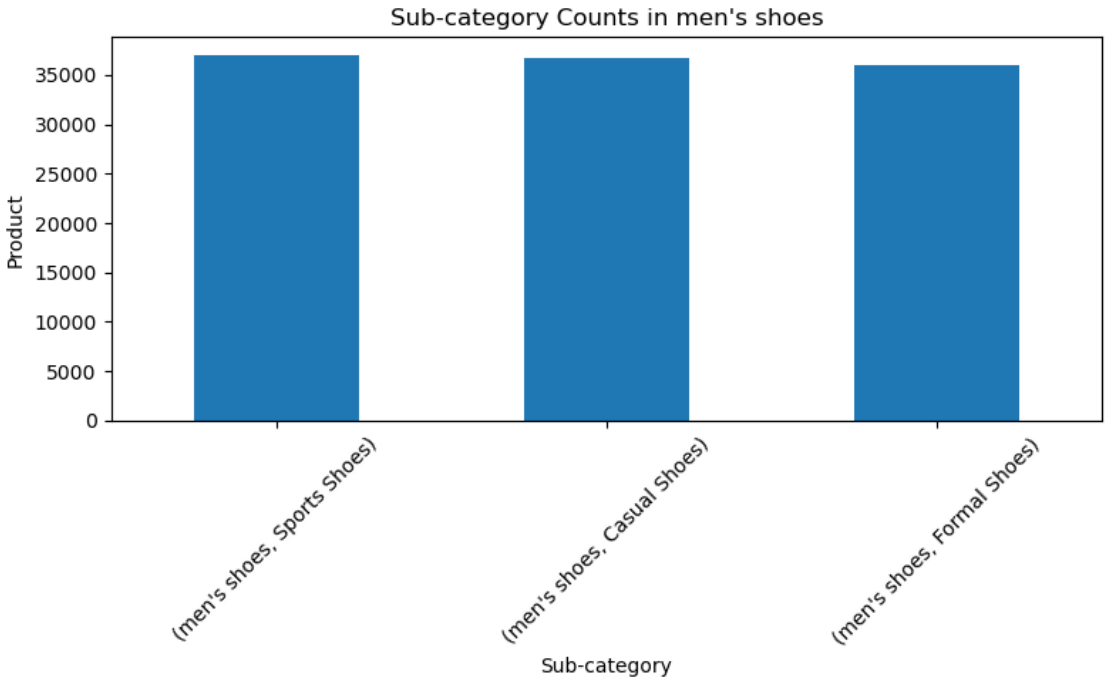

Sub-category Counts in home & kitchen
Product
2500
2000
1500
1000
500
0
(home & kitchen, Home Furnishing)
(home & kitchen, Furniture)
(home & kitchen, Home Décor)
(home & kitchen, Sewing & Craft Supplies)
(home & kitchen, Indoor Lighting)
(home & kitchen, All Home & Kitchen)
(home & kitchen, Home Storage)
(home & kitchen, Bedroom Linen)
(home & kitchen, Kitchen & Dining)
(home & kitchen, Garden & Outdoors)
(home & kitchen, Home Improvement)
(home & kitchen, Kitchen Storage & Containers)
Sub-category

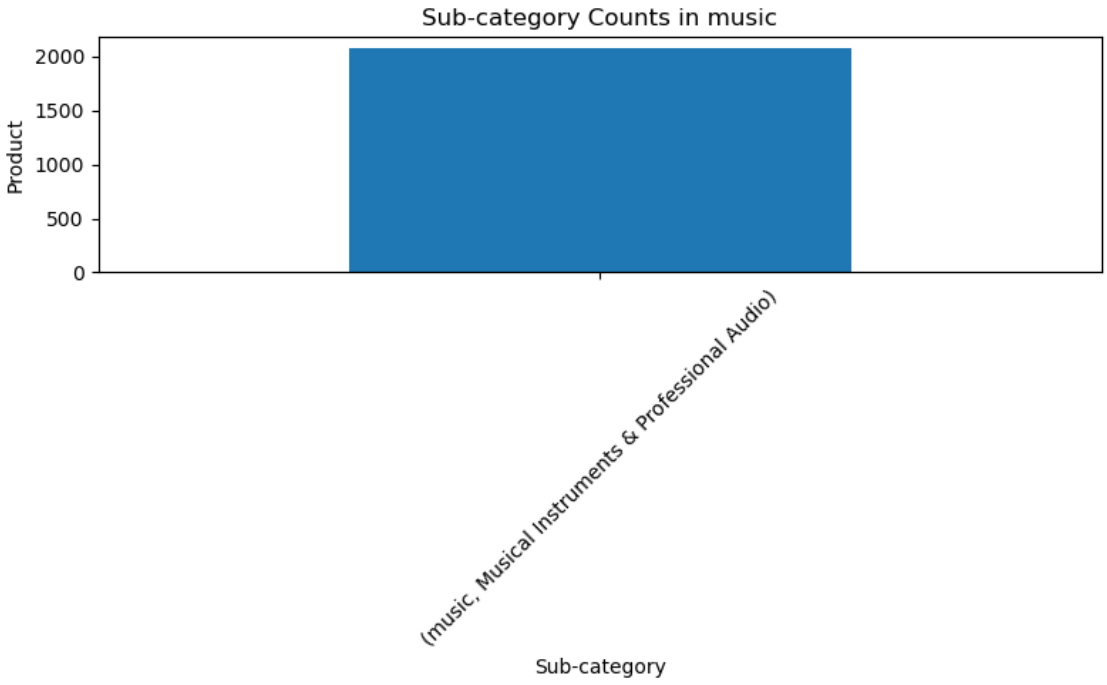

Sub-category Counts in home, kitchen, pets
Product
0
10
20
30
(home, kitchen, pets, Refurbished & Open Box)
Sub-category

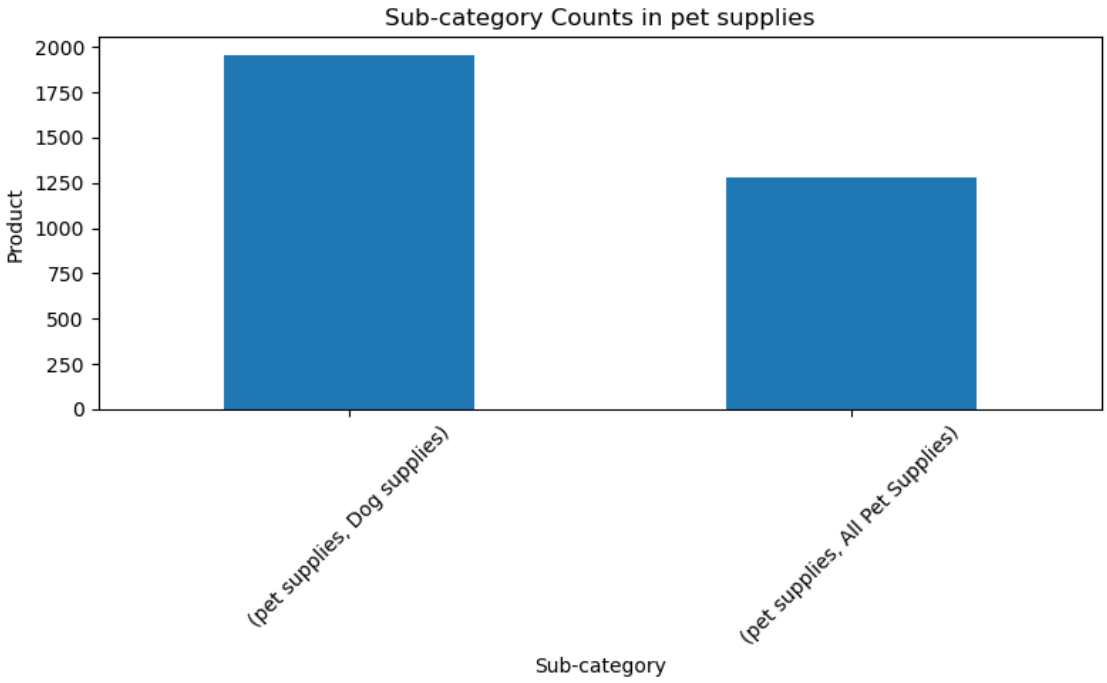

Sub-category Counts in industrial supplies
Product
0
500
1000
1500
2000
(industrial supplies, Lab & Scientific)
(industrial supplies, Janitorial & Sanitation Supplies)
(industrial supplies, Test, Measure & Inspect)
(industrial supplies, Industrial & Scientific Supplies)
Sub-category

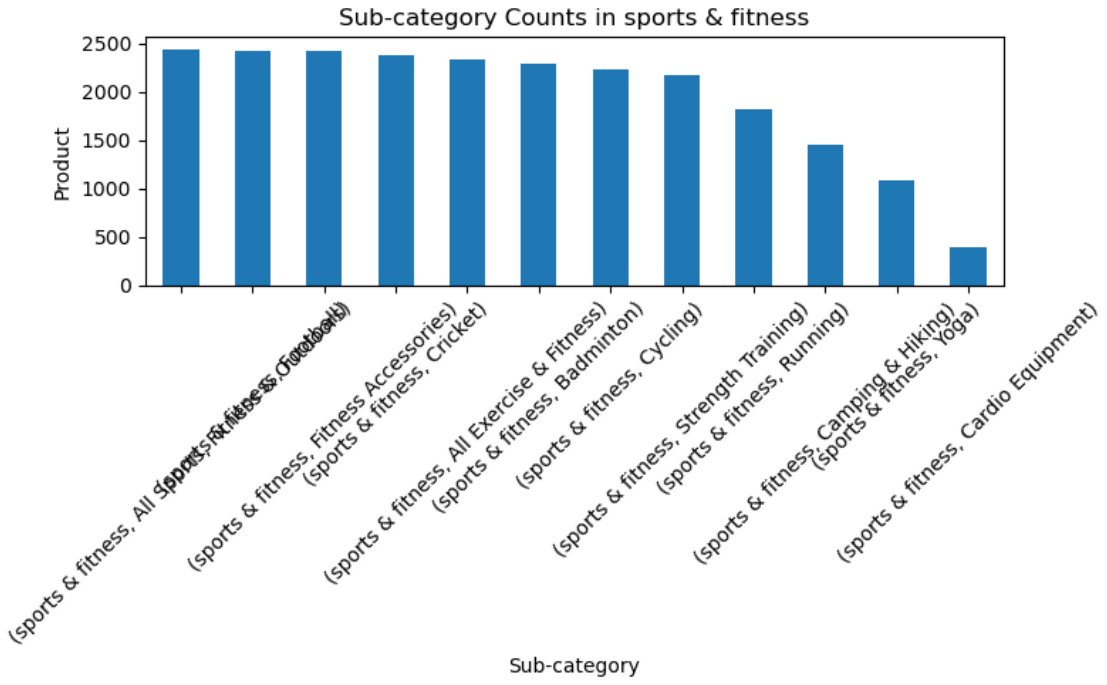

Sub-category Counts in kids' fashion
Product
0
1000
2000
3000
4000
5000
(kids' fashion, Kids' Fashion)
(kids' fashion, School Bags)
(kids' fashion, Baby Fashion)
(kids' fashion, Kids' Shoes)
(kids' fashion, Kids' Clothing)
(kids' fashion, Kids' Watches)
Sub-category

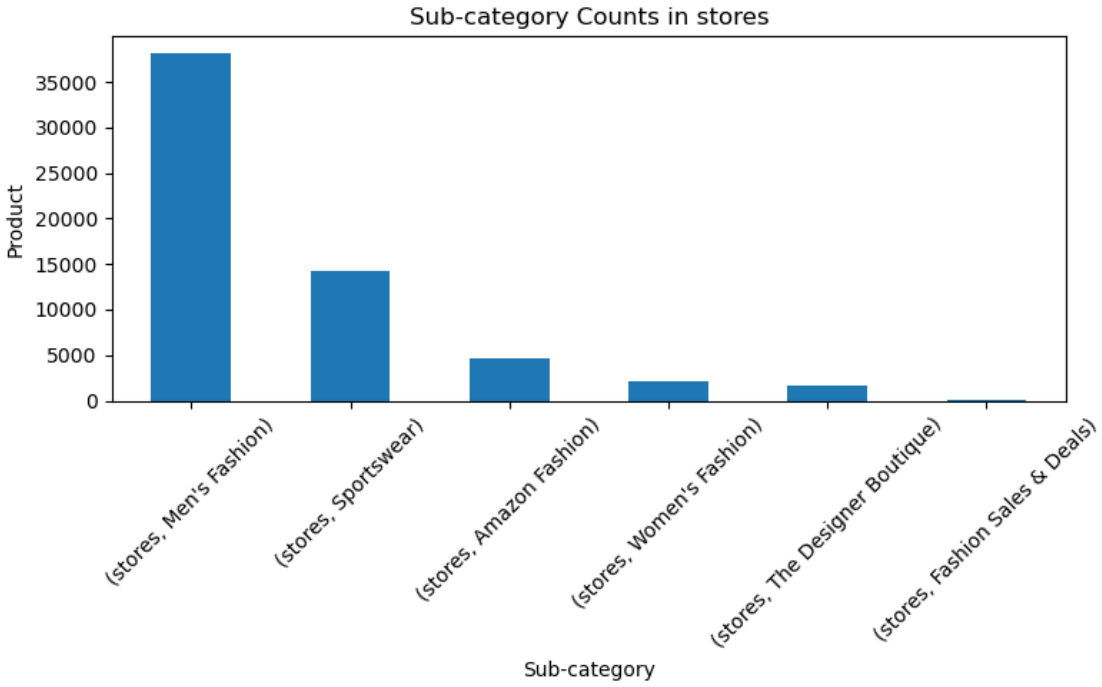

Sub-category Counts in men's clothing
Product
0
5000
10000
15000
20000
25000
30000
35000
(men's clothing, Innerwear)
(men's clothing, Shirts)
(men's clothing, T-shirts & Polos)
(men's clothing, Jeans)
Sub-category

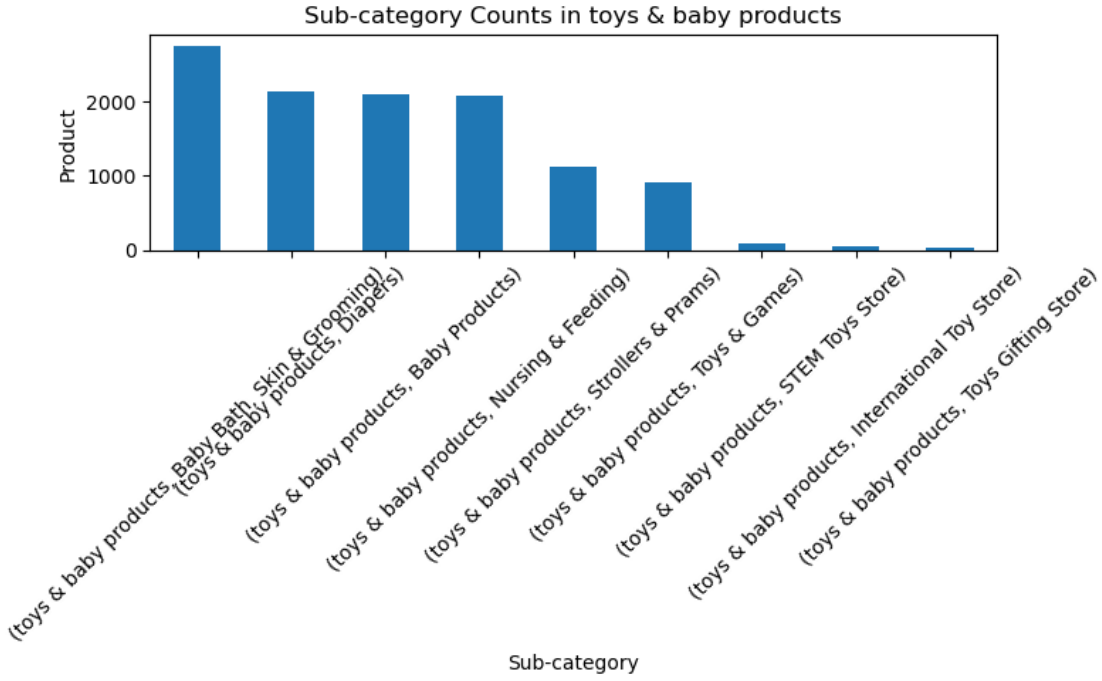

Sub-category Counts in men's shoes
Product
35000
30000
25000
20000
15000
10000
5000
0
(men's shoes, Sports Shoes)
(men's shoes, Casual Shoes)
(men's shoes, Formal Shoes)
Sub-category

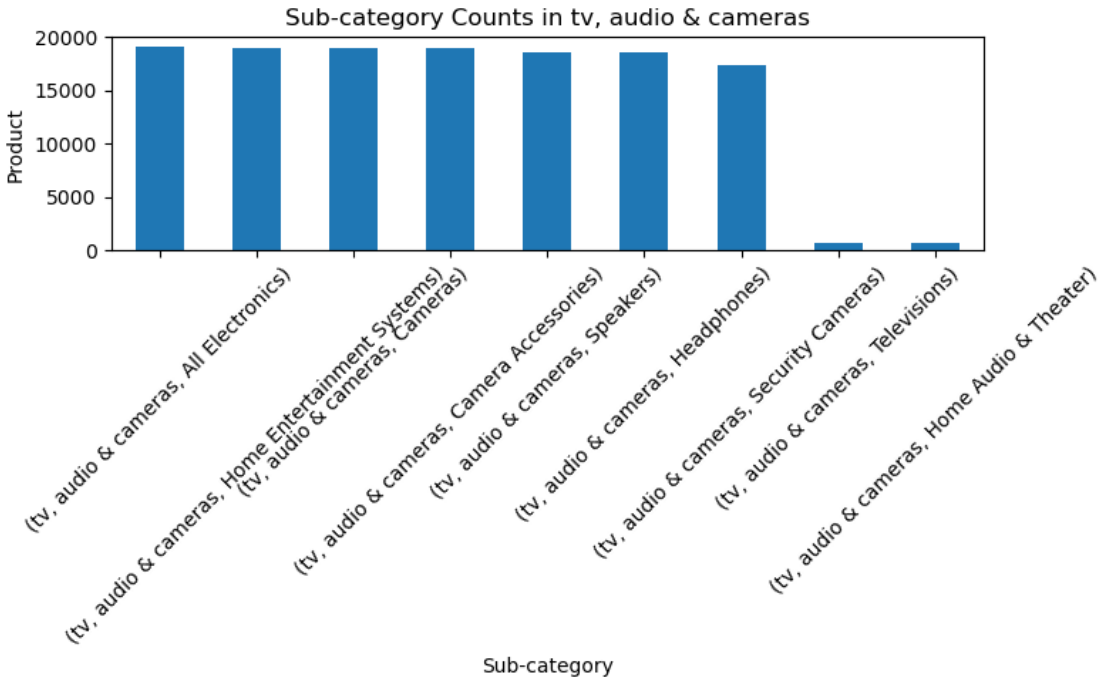

Sub-category Counts in music
Product
2000
1500
1000
500
0
(music, Musical Instruments & Professional Audio)
Sub-category

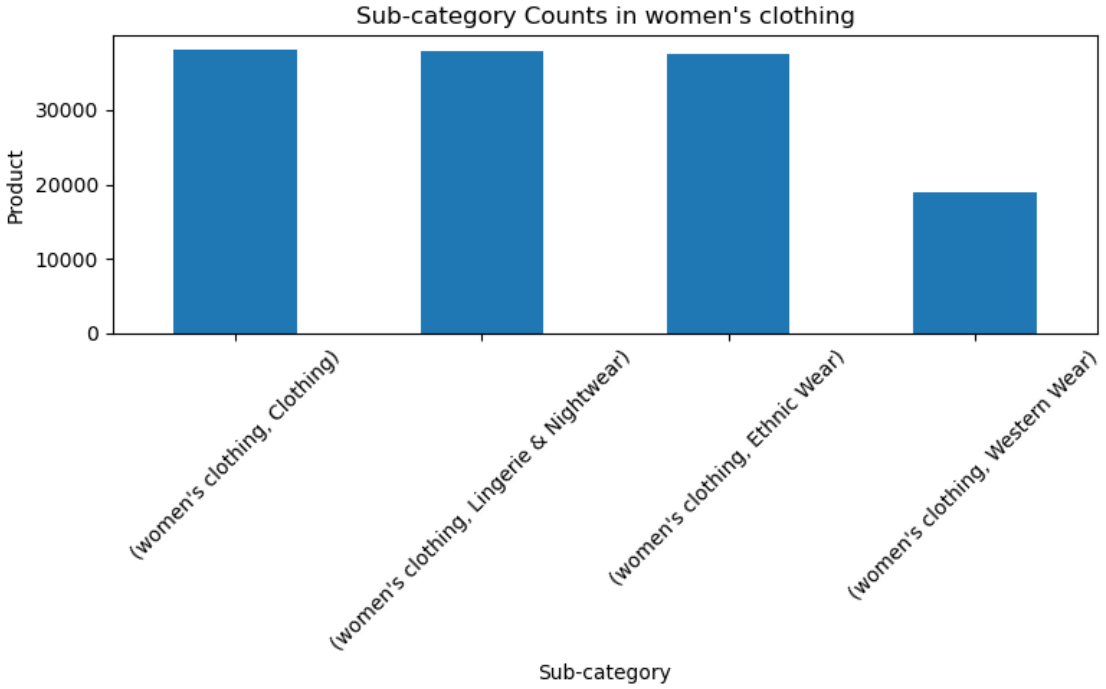

Sub-category Counts in pet supplies
Product
2000
1750
1500
1250
1000
750
500
250
0
(pet supplies, Dog supplies)
(pet supplies, All Pet Supplies)
Sub-category

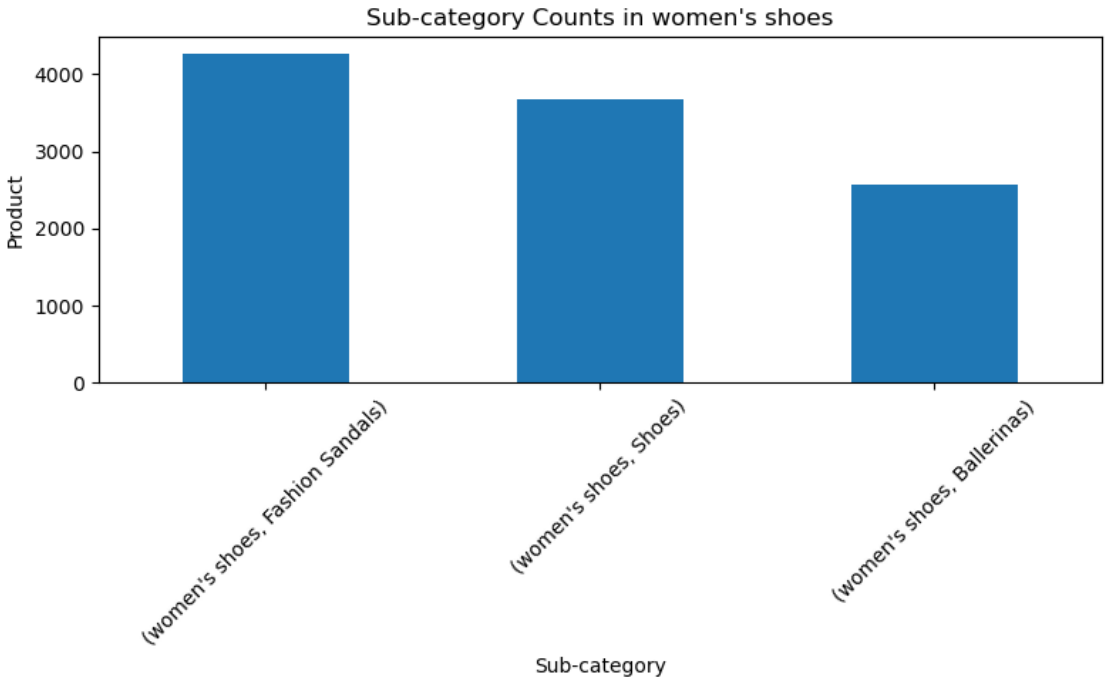

Sub-category Counts in sports & fitness
Product
2500
2000
1500
1000
500
0
(sports & fitness, Fitness Accessories)
(sports & fitness, Cricket)
(sports & fitness, All Exercise & Fitness)
(sports & fitness, Badminton)
(sports & fitness, Cycling)
(sports & fitness, Strength Training)
(sports & fitness, Running)
(sports & fitness, Camping & Hiking)
(sports & fitness, Yoga)
(sports & fitness, Cardio Equipment)
Sub-category


Sub-category Counts in stores
Product
35000
30000
25000
20000
15000
10000
5000
0
(stores, Men's Fashion)
(stores, Sportswear)
(stores, Amazon Fashion)
(stores, Women's Fashion)
(stores, The Designer Boutique)
(stores, Fashion Sales & Deals)
Sub-category


Sub-category Counts in toys & baby products
Product
2000
1000
0
(toys & baby products, Baby Bath, Skin & Grooming)
(toys & baby products, Diapers)
(toys & baby products, Baby Products)
(toys & baby products, Nursing & Feeding)
(toys & baby products, Strollers & Prams)
(toys & baby products, Toys & Games)
(toys & baby products, STEM Toys Store)
(toys & baby products, International Toy Store)
(toys & baby products, Toys Gifting Store)
Sub-category


Sub-category Counts in tv, audio & cameras
Product
20000
15000
10000
5000
0
(tv, audio & cameras, All Electronics)
(tv, audio & cameras, Home Entertainment Systems)
(tv, audio & cameras, Cameras)
(tv, audio & cameras, Camera Accessories)
(tv, audio & cameras, Speakers)
(tv, audio & cameras, Headphones)
(tv, audio & cameras, Security Cameras)
(tv, audio & cameras, Televisions)
(tv, audio & cameras, Home Audio & Theater)
Sub-category


Sub-category Counts in women's clothing
Product
30000
20000
10000
0
(women's clothing, Clothing)
(women's clothing, Lingerie & Nightwear)
(women's clothing, Ethnic Wear)
(women's clothing, Western Wear)
Sub-category


Sub-category Counts in women's shoes
Product
4000
3000
2000
1000
0
(women's shoes, Fashion Sandals)
(women's shoes, Shoes)
(women's shoes, Ballerinas)
Sub-category

[ ]: