Explorartory Data Analysis

In [1]:
```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from collections import Counter
import math
import re
import os
import seaborn as sns
```

In [2]:
```
# we have give a json file which consists of all information about
# the products
# loading the data using pandas' read_json file.
data = pd.read_json('tops_fashion.json')
```

In [3]:
```
print ('Number of data points : ', data.shape[0],
'Number of features/variables:', data.shape[1])
```

```
Number of data points :  183138 Number of features/variables: 19
```

In [4]:
```
data.head(5)
```

Out[4]:

	sku	asin	product_type_name	formatted_price	author	color	brand	publisher	availability	
0	None	B016I2TS4W	SHIRT	None	None	None	FNC7C	None	None	https://www.amazon.com/revie
1	None	B01N49AI08	SHIRT	None	None	None	FIG Clothing	None	None	https://www.amazon.com/revie
2	None	B01JDPCOHO	SHIRT	None	None	None	FIG Clothing	None	None	https://www.amazon.com/revie
3	None	B01N19U5H5	SHIRT	None	None	None	Focal18	None	None	https://www.amazon.com/re
4	None	B004GSI2OS	SHIRT	$26.26	None	Onyx Black/ Stone	FeatherLite	None	Usually ships in 6-10 business days	https://www.amazon.com/revie

In [5]:
```
data.columns
```

Out[5]:
```
Index(['sku', 'asin', 'product_type_name', 'formatted_price', 'author',
       'color', 'brand', 'publisher', 'availability', 'reviews',
       'large_image_url', 'availability_type', 'small_image_url',
       'editorial_review', 'title', 'model', 'medium_image_url',
       'manufacturer', 'editorial_reivew'],
      dtype='object')
```

Of these 19 features, we will be using only 7 features in the project.

1. asin (Amazon standard identification number)
2. brand (brand to which the product belongs to)
3. color (Color information of apparel, it can contain many colors as a value ex: red and black stripes)
4. product_type_name (type of the apperal, ex: SHIRT/TSHIRT)
5. medium_image_url (url of the image)
6. title (title of the product.)
7. formatted_price (price of the product)

In [6]:
```
data = data[['asin','brand','color','product_type_name','medium_image_url','title','formatted_price']]
```

In [7]:
```
print ('Number of data points : ', data.shape[0], \
'Number of features:', data.shape[1])
data.head() # prints the top rows in the table.
```

```
Number of data points :  183138 Number of features: 7
```

Out[7]:

	asin	brand	color	product_type_name	medium_image_url	title	formatted_price
0	B016I2TS4W	FNC7C	None	SHIRT	https://images-na.ssl-images-amazon.com/images...	Minions Como Superheroes Ironman Long Sleeve R...	None
1	B01N49AI08	FIG Clothing	None	SHIRT	https://images-na.ssl-images-amazon.com/images...	FIG Clothing Womens Izo Tunic	None
2	B01JDPCOHO	FIG Clothing	None	SHIRT	https://images-na.ssl-images-amazon.com/images...	FIG Clothing Womens Won Top	None
3	B01N19U5H5	Focal18	None	SHIRT	https://images-na.ssl-images-amazon.com/images...	Focal18 Sailor Collar Bubble Sleeve Blouse Shi...	None
4	B004GSI2OS	FeatherLite	Onyx Black/ Stone	SHIRT	https://images-na.ssl-images-amazon.com/images...	Featherlite Ladies' Long Sleeve Stain Resistan...	$26.26

1. Analysis of Missing Data

1.1 Basic stats for the feature: product_type_name

In [8]:
```
print(data['product_type_name'].describe())
```

```
count     183138
unique        72
top        SHIRT
freq      167794
Name: product_type_name, dtype: object
```

We have total 72 unique type of product_type_names.

91.62% (167794/183138) of the products are shirts.

In [9]:
```
print(data['product_type_name'].unique())
```

```
['SHIRT' 'SWEATER' 'APPAREL' 'OUTDOOR_RECREATION_PRODUCT'
 'BOOKS_1973_AND_LATER' 'PANTS' 'HAT' 'SPORTING_GOODS' 'DRESS' 'UNDERWEAR'
 'SKIRT' 'OUTERWEAR' 'BRA' 'ACCESSORY' 'ART_SUPPLIES' 'SLEEPWEAR'
 'ORCA_SHIRT' 'HANDBAG' 'PET_SUPPLIES' 'SHOES' 'KITCHEN' 'ADULT_COSTUME'
 'HOME_BED_AND_BATH' 'MISC_OTHER' 'BLAZER' 'HEALTH_PERSONAL_CARE'
 'TOYS_AND_GAMES' 'SWIMWEAR' 'CONSUMER_ELECTRONICS' 'SHORTS' 'HOME'
 'AUTO_PART' 'OFFICE_PRODUCTS' 'ETHNIC_WEAR' 'BEAUTY'
 'INSTRUMENT_PARTS_AND_ACCESSORIES' 'POWERSPORTS_PROTECTIVE_GEAR' 'SHIRTS'
 'ABIS_APPAREL' 'AUTO_ACCESSORY' 'NONAPPARELMISC' 'TOOLS' 'BABY_PRODUCT'
 'SOCKSHOSIERY' 'POWERSPORTS_RIDING_SHIRT' 'EYEWEAR' 'SUIT'
 'OUTDOOR_LIVING' 'POWERSPORTS_RIDING_JACKET' 'HARDWARE' 'SAFETY_SUPPLY'
 'ABIS_DVD' 'VIDEO_DVD' 'GOLF_CLUB' 'MUSIC_POPULAR_VINYL'
 'HOME_FURNITURE_AND_DECOR' 'TABLET_COMPUTER' 'GUILD_ACCESSORIES'
 'ABIS_SPORTS' 'ART_AND_CRAFT_SUPPLY' 'BAG' 'MECHANICAL_COMPONENTS'
 'SOUND_AND_RECORDING_EQUIPMENT' 'COMPUTER_COMPONENT' 'JEWELRY'
 'BUILDING_MATERIAL' 'LUGGAGE' 'BABY_COSTUME' 'POWERSPORTS_VEHICLE_PART'
 'PROFESSIONAL_HEALTHCARE' 'SEEDS_AND_PLANTS' 'WIRELESS_ACCESSORY']
```

In [10]:
```
# find the 10 most frequent product_type_names.
product_type_count = Counter(list(data['product_type_name']))
product_type_count.most_common(10)
```

Out[10]:
```
[('SHIRT', 167794),
 ('APPAREL', 3549),
 ('BOOKS_1973_AND_LATER', 3336),
 ('DRESS', 1584),
 ('SPORTING_GOODS', 1281),
 ('SWEATER', 837),
 ('OUTERWEAR', 796),
 ('OUTDOOR_RECREATION_PRODUCT', 729),
 ('ACCESSORY', 636),
 ('UNDERWEAR', 425)]
```

1.2 Basic stats for the feature: brand

In [11]:
```
# there are 10577 unique brands
print(data['brand'].describe())
# 183138 - 182987 = 151 missing values.
```

```
count     182987
unique     10577
top         Zago
freq         223
Name: brand, dtype: object
```

In [12]:
```
brand_count = Counter(list(data['brand']))
brand_count.most_common(10)
```

Out[12]:
```
[('Zago', 223),
 ('XQS', 222),
 ('Yayun', 215),
 ('YUNY', 198),
 ('XiaoTianXin-women clothes', 193),
 ('Generic', 192),
 ('Boohoo', 190),
 ('Alion', 188),
 ('Abetteric', 187),
 ('TheMogan', 187)]
```

1.2 Basic stats for the feature: Color

In [13]:
```
print(data['color'].describe())
# we have 7380 unique colors
# 7.2% of products are black in color
# 64956 of 183138 products have brand information. That's approx 35.4%.
```

```
count     64956
unique     7380
top       Black
freq      13207
Name: color, dtype: object
```

In [14]:
```
color_count = Counter(list(data['color']))
color_count.most_common(10)
```

Out[14]:
```
[(None, 118182),
 ('Black', 13207),
 ('White', 8616),
 ('Blue', 3570),
 ('Red', 2289),
 ('Pink', 1842),
 ('Grey', 1499),
 ('*', 1388),
 ('Green', 1258),
 ('Multi', 1203)]
```

1.3 Basic stats for the feature: formatted_price

In [15]:
```
print(data['formatted_price'].describe())
# Only 28,395 (15.5% of whole data) products with price information
```

```
count      28395
unique      3135
top       $19.99
freq         945
Name: formatted_price, dtype: object
```

In [16]:
```
price_count = Counter(list(data['formatted_price']))
price_count.most_common(10)
```

Out[16]:
```
[(None, 154743),
 ('$19.99', 945),
 ('$9.99', 749),
 ('$9.50', 601),
 ('$14.99', 472),
 ('$7.50', 463),
 ('$24.99', 414),
 ('$29.99', 370),
 ('$8.99', 343),
 ('$9.01', 336)]
```

1.3 Basic stats for the feature: title

In [17]:
```
print(data['title'].describe())
# All of the products have a title.
# Titles are fairly descriptive of what the product is.
```

```
count                                             183138
unique                                            175985
top       Nakoda Cotton Self Print Straight Kurti For Women
freq                                                  77
Name: title, dtype: object
```

In [18]:
```
data.to_pickle('pickels/180k_apparel_data')
```