Makeup Product Analysis by Carol Ly

I am always curious to see what kind of makeup products are being widely used in the cosmetics industry. I found this cosmetic brand products dataset in Kaggle. Cosmetics has been one of my interests, which led to this analysis that I will be doing. For this assignment, I will be analyzing the makeup product dataset and want to get a better understanding of the dynamics in the cosmetic market. Here are some of the analyses I will be looking for:

1. What are the top 5 makeup products?
2. What is the average $ amount?
3. What is the maximum price?
4. What is the potential average price based on brand?
5. How many types of products are there?

The process for this analysis will be straightforward, reviewing the set of data, cleaning the data, analyzing, and providing some visualizations at the end of the analysis.

Dataset Link: https://www.kaggle.com/datasets/shivd24coder/cosmetic-brand-products-dataset/data

In [ ]:

```
#Generate file and load csv file into pandas dataframe
import pandas as pd

file = 'makeup_dataset.csv'

df = pd.read_csv(file)

#Explore Makeup Dataset
df.head()
```

Out[ ]:

	id	brand	name	price	price_sign	currency	
0	1048	colourpop	Lippie Pencil	5.0	$	CAD	https://cdn.shopify.com/s/file
1	1047	colourpop	Blotted Lip	5.5	$	CAD	https://cdn.shopify.com/s/file
2	1046	colourpop	Lippie Stix	5.5	$	CAD	https://cdn.shopify.com/s/file
3	1045	colourpop	No Filter Foundation	12.0	$	CAD	https://cdn.shopify.com/s/file
4	1044	boosh	Lipstick	26.0	$	CAD	https://cdn.shopify.com/s/fil

In [ ]:

```
#Understand the dataset based on columns, structures, and data
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 931 entries, 0 to 930
Data columns (total 19 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 931 non-null    int64  
 1   brand              919 non-null    object 
 2   name               931 non-null    object 
 3   price              917 non-null    float64
 4   price_sign         368 non-null    object 
 5   currency           368 non-null    object 
 6   image_link         931 non-null    object 
 7   product_link       931 non-null    object 
 8   website_link       931 non-null    object 
 9   description        906 non-null    object 
 10  rating             340 non-null    float64
 11  category           507 non-null    object 
 12  product_type       931 non-null    object 
 13  tag_list           931 non-null    object 
 14  created_at         931 non-null    object 
 15  updated_at         931 non-null    object 
 16  product_api_url    931 non-null    object 
 17  api_featured_image 931 non-null    object 
 18  product_colors     931 non-null    object 
dtypes: float64(2), int64(1), object(16)
memory usage: 138.3+ KB
None
```

By running the above, turns out there are 19 columns in total. Now I want to find a quick summary of the various statistics of the data.

In [ ]:

```
#Statistics Summary
print(df.describe())
```

```
                id       price      rating
count   931.000000  917.000000  340.000000
mean    531.163265   16.508593    4.319118
std     311.054915   11.028035    0.675849
min       1.000000    0.000000    1.500000
25%     263.000000    8.990000    4.000000
50%     518.000000   13.990000    4.500000
75%     814.500000   22.000000    5.000000
max    1048.000000   77.000000    5.000000
```

The next part of the process of the analysis will be to clean up the data. First I need to see how to handle the missing values or fields, remove any duplicate data, and format the data types if needed.

In [ ]:

```
#Data Cleaning - remove duplicates as needed
df.drop_duplicates(inplace=True)
```

Here is where I will be analyzing the data.

In [ ]:

```python
#Average Price
average = df['price'].mean()
print(f'Average price: ${average:.2f}')

#Max Price
maximum = df['price'].max()
print(f'Maximum price: ${maximum:.2f}')
```

```
Average price: $16.51
Maximum price: $77.00
```

In [ ]:

```python
#Analyze the average price based on brand

avg_brand = df.groupby('brand')['price'].mean()
print(avg_brand)
```

```
brand
almay                       12.661429
alva                         9.950000
anna sui                    22.000000
annabelle                    9.805455
benefit                     30.536585
boosh                       26.000000
burt's bees                  9.990000
butter london               25.480000
c'est moi                    0.000000
cargo cosmetics             29.250000
china glaze                  8.000000
clinique                    22.764674
coastal classic creation     0.000000
colourpop                    7.000000
covergirl                    9.684444
dalish                      22.000000
deciem                       6.800000
dior                        27.358108
dr. hauschka                33.916667
e.l.f.                       6.767778
essie                       10.000000
fenty                       23.200000
glossier                    25.000000
green people                 0.000000
iman                              NaN
l'oreal                     13.871957
lotus cosmetics usa          0.000000
maia's mineral galaxy        0.000000
marcelle                    14.590000
marienatie                   0.000000
maybelline                  11.138148
milani                       9.066923
mineral fusion              25.375000
misa                         9.390000
mistura                     56.490000
moov                        14.990000
nudus                        0.000000
nyx                          8.418171
orly                        10.745000
pacifica                    25.458462
penny lane organics          0.000000
physicians formula          17.213256
piggy paint                 11.990000
pure anada                  14.249375
rejuva minerals              0.000000
revlon                      13.493448
sally b's skin yummies       0.000000
salon perfect                6.990000
sante                       22.090000
sinful colours               2.990000
smashbox                    29.847826
stila                       46.247500
suncoat                     16.006667
w3llpeople                   0.000000
wet n wild                   4.306667
```

```
zorah                           25.500000
zorah biocosmetiques             0.000000
Name: price, dtype: float64
```

In [ ]:

```
#Analyze the top brand of makeup products
top_brands = df['brand'].value_counts().head(5)
print(top_brands)
```

```
brand
nyx            164
clinique        93
dior            74
maybelline      54
covergirl       54
Name: count, dtype: int64
```

In [ ]:

```
#Product Types
product_ty_count = df['product_type'].value_counts()
print(product_ty_count)
```

```
product_type
foundation     166
lipstick       154
eyeliner       148
mascara         92
eyeshadow       86
blush           78
bronzer         69
nail_polish     60
eyebrow         49
lip_liner       29
Name: count, dtype: int64
```

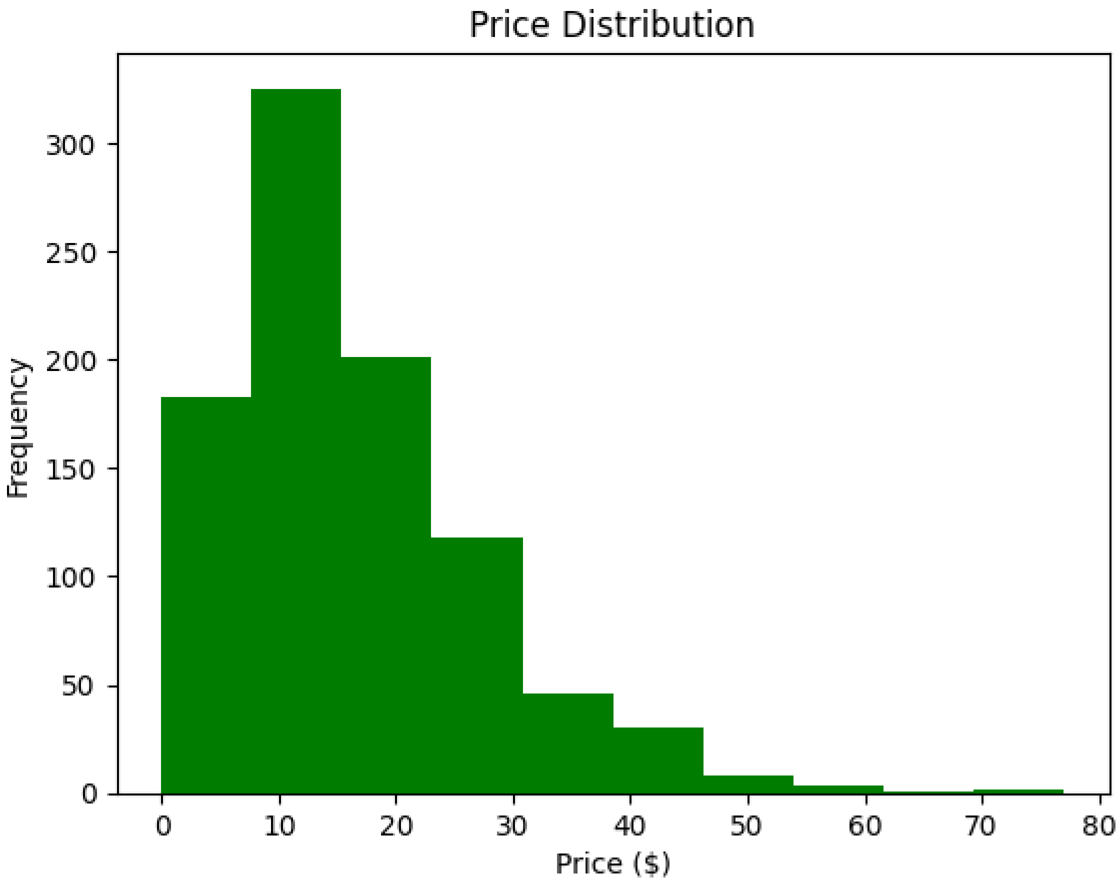
Visualizations and Outputs

In [ ]:

```
import matplotlib.pyplot as plt
```

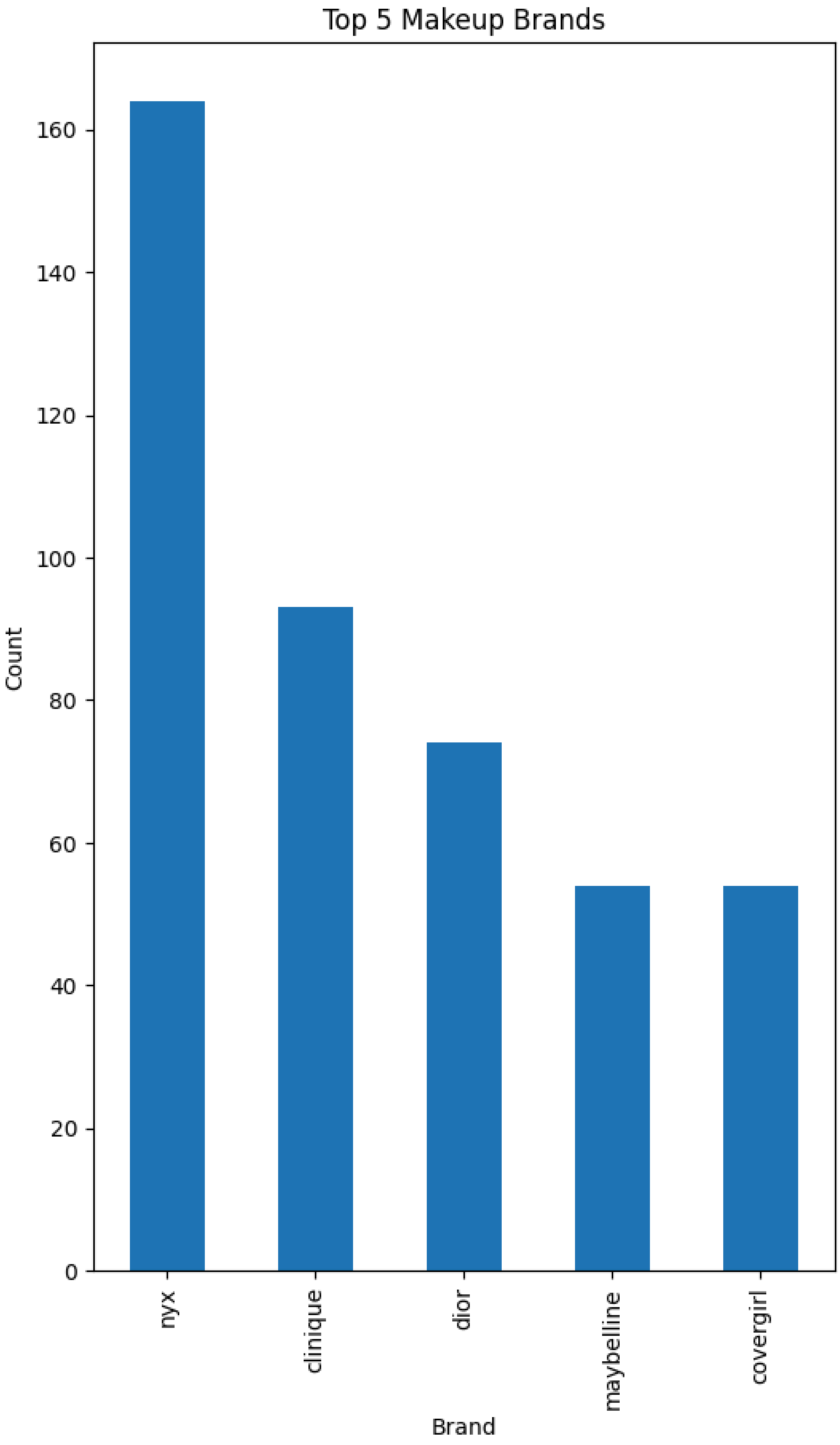
In [ ]:

```
#Create histogram that shows Price Distribution

plt.hist(df['price'], bins=10, color='green')
plt.xlabel('Price ($)')
plt.ylabel('Frequency')
plt.title('Price Distribution')
plt.show()
```



In [ ]:

```
#Create a simple bar chart to show the top 5 makeup brands

top_brands.plot(kind='bar', figsize=(6, 10))
plt.xlabel('Brand')
plt.ylabel('Count')
plt.title('Top 5 Makeup Brands')
plt.show()
```


Top 5 Makeup Brands
Count
0
20
40
60
80
100
120
140
160
nyx
clinique
dior
maybelline
covergirl
Brand

In conclusion, it was interesting to see how data is analzyed and revealed with all sorts of patterns. The analysis has answered my questions. Foundation had the most products while NYX was considered as the top makeup brand. The price distribution was in a good range, though there were some products with zero dollars. Although there could be limitations to the analysis, it was intriguing to see the price and the different product types of makeup. What could be useful to dig deeper into the data would be some customer reviews and demographics such as age group. This analysis is a good starting point to see the dynamics of the cosmetic market.