EDA On Flipkart Total Sales

What is the problem?

The aim is to analyze the Flipkart sales dataset to understand customer preferences, product performance, and sales trends. The analysis will focus on identifying the most popular categories, brands, pricing strategies, and customer ratings.

Why is it important to solve it?

Solving this problem will help Flipkart make decisions to improve product listings, pricing strategies, and customer satisfaction. The insights gained from this analysis could lead to increased sales, better customer response, and improved overall business performance.

Attribute information :

uniq_id: Unique identifier for each product. This can be used to identify each product in the dataset.

crawl_timestamp: Timestamp when the data was crawled.

product_url: URL of the product on Flipkart.

product_name: This column contains the name/title of the listed product.

product_category_tree: This contains the category and sub-category information of the product.

pid: Product ID.

retail_price: This is the original price of the product before any discounts.

discounted_price: This is the selling price of the product after applying discounts.

image: Image URLs of the product.

is_FK_Advantage_product: This indicates whether the product is part of the Flipkart Advantage program.

description: Detail description of the product.

product_rating: This is the rating given to the product by customers. If not available, it is marked as "No rating available".

overall_rating: This is the overall rating of the product. If not available, it is marked as "No rating available".

brand: Brand name of the product.

product_specifications: This contains detailed specifications of the product

In [1]:
```python
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```python
# loading a csv file from local host
df = pd.read_csv('flipkart_com-ecommerce_sample.csv')
```

In [3]:
```python
# taking overview of the top values of dataset
df.head()
```

Out[3]:

	uniq_id	crawl_timestamp	product_url	product_name	product_category_tree	pid	retail_price	discounted_price	image	is_FK_Advantage_product
0	c2d766ca982eca8304150849735ffef9	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	SRTEH2FF9KEDEFGF	999.0	379.0	["http://img5a.flixcart.com/image/short/u/4/a/...	False
1	7f7036a6d550aaa89d34c77bd39a5e48	2016-03-25 22:59:23 +0000	http://www.flipkart.com/fabhomedecor-fabric-do...	FabHomeDecor Fabric Double Sofa Bed	["Furniture >> Living Room Furniture >> Sofa B...	SBEEH3QGU7MFYJFY	32157.0	22646.0	["http://img6a.flixcart.com/image/sofa-bed/j/f...	False
2	f449ec65dcbc041b6ae5e6a32717d01b	2016-03-25 22:59:23 +0000	http://www.flipkart.com/aw-bellies/p/itmeh4grg...	AW Bellies	["Footwear >> Women's Footwear >> Ballerinas >...	SHOEH4GRSUBJGZXE	999.0	499.0	["http://img5a.flixcart.com/image/shoe/7/z/z/r...	False
3	0973b37acd0c664e3de26e97e5571454	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	SRTEH2F6HUZMQ6SJ	699.0	267.0	["http://img5a.flixcart.com/image/short/6/2/h/...	False
4	bc940ea42ee6bef5ac7cea3fb5cfbee7	2016-03-25 22:59:23 +0000	http://www.flipkart.com/sicons-all-purpose-arn...	Sicons All Purpose Arnica Dog Shampoo	["Pet Supplies >> Grooming >> Skin & Coat Care...	PSOEH3ZYDMSYARJ5	220.0	210.0	["http://img5a.flixcart.com/image/pet-shampoo/...	False

In [4]:
```python
df.columns          # printing column names
```

Out[4]:
```
Index(['uniq_id', 'crawl_timestamp', 'product_url', 'product_name',
       'product_category_tree', 'pid', 'retail_price', 'discounted_price',
       'image', 'is_FK_Advantage_product', 'description', 'product_rating',
       'overall_rating', 'brand', 'product_specifications'],
      dtype='object')
```

In [5]:
```python
# detailed information about dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20000 entries, 0 to 19999
Data columns (total 15 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   uniq_id                  20000 non-null  object
 1   crawl_timestamp          20000 non-null  object
 2   product_url              20000 non-null  object
 3   product_name             20000 non-null  object
 4   product_category_tree    20000 non-null  object
 5   pid                      20000 non-null  object
 6   retail_price             19922 non-null  float64
 7   discounted_price         19922 non-null  float64
 8   image                    19997 non-null  object
 9   is_FK_Advantage_product  20000 non-null  bool
 10  description              19998 non-null  object
 11  product_rating           20000 non-null  object
 12  overall_rating           20000 non-null  object
 13  brand                    14136 non-null  object
 14  product_specifications   19986 non-null  object
dtypes: bool(1), float64(2), object(12)
memory usage: 2.2+ MB
```

In [6]:
```python
df.shape
```

Out[6]: (20000, 15)

In [7]:
```python
df.describe()
```

Out[7]:

	retail_price	discounted_price
count	19922.000000	19922.000000
mean	2979.206104	1973.401767
std	9009.639341	7333.586040
min	35.000000	35.000000
25%	666.000000	350.000000
50%	1040.000000	550.000000
75%	1999.000000	999.000000
max	571230.000000	571230.000000

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

Out[8]:
```
uniq_id                       0
crawl_timestamp               0
product_url                   0
product_name                  0
product_category_tree         0
pid                           0
retail_price                 78
discounted_price             78
image                         3
is_FK_Advantage_product       0
description                   2
product_rating                0
overall_rating                0
brand                      5864
product_specifications       14
dtype: int64
```

Which are the top 5 product categories based on the number of listings?

In [9]:
```python
df.product_category_tree
```

Out[9]:
```
0        ["Clothing >> Women's Clothing >> Lingerie, Sl...
1        ["Furniture >> Living Room Furniture >> Sofa B...
2        ["Footwear >> Women's Footwear >> Ballerinas >...
3        ["Clothing >> Women's Clothing >> Lingerie, Sl...
4        ["Pet Supplies >> Grooming >> Skin & Coat Care...
                               ...
19995    ["Baby Care >> Baby & Kids Gifts >> Stickers >...
19996    ["Baby Care >> Baby & Kids Gifts >> Stickers >...
19997    ["Baby Care >> Baby & Kids Gifts >> Stickers >...
19998    ["Baby Care >> Baby & Kids Gifts >> Stickers >...
19999    ["Baby Care >> Baby & Kids Gifts >> Stickers >...
Name: product_category_tree, Length: 20000, dtype: object
```

In [10]:
```python
# observing the above we need change it a little bit
# here we added new column name 'product_category' and store values in it
df['product_category'] = df['product_category_tree'].str.split('>>').str[0]
```

In [11]:
```python
df.head()
```

Out[11]:

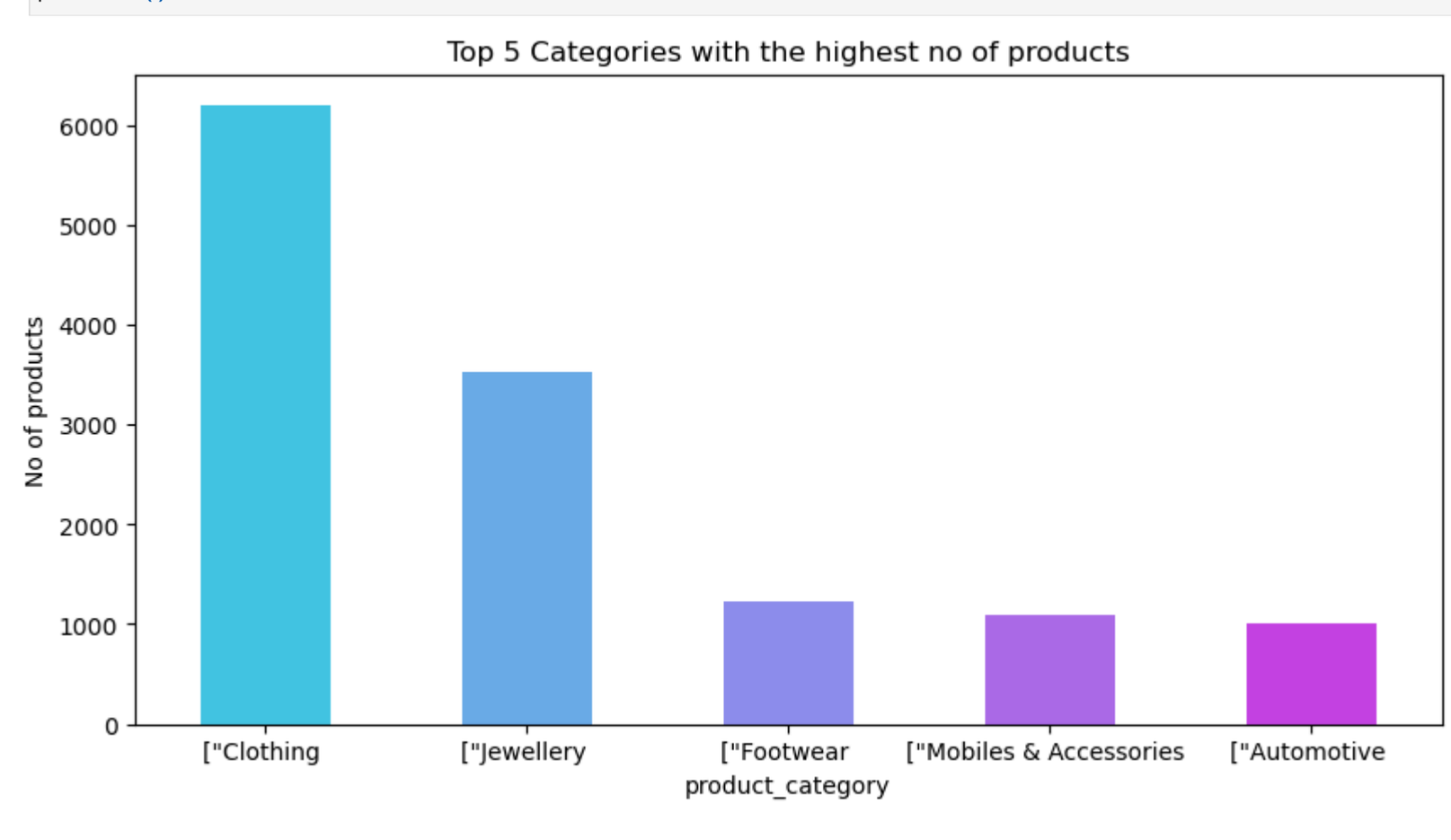
	uniq_id	crawl_timestamp	product_url	product_name	product_category_tree	pid	retail_price	discounted_price	image	is_FK_Advantage_product
0	c2d766ca982eca8304150849735ffef9	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	SRTEH2FF9KEDEFGF	999.0	379.0	["http://img5a.flixcart.com/image/short/u/4/a/...	False
1	7f7036a6d550aaa89d34c77bd39a5e48	2016-03-25 22:59:23 +0000	http://www.flipkart.com/fabhomedecor-fabric-do...	FabHomeDecor Fabric Double Sofa Bed	["Furniture >> Living Room Furniture >> Sofa B...	SBEEH3QGU7MFYJFY	32157.0	22646.0	["http://img6a.flixcart.com/image/sofa-bed/j/f...	False
2	f449ec65dcbc041b6ae5e6a32717d01b	2016-03-25 22:59:23 +0000	http://www.flipkart.com/aw-bellies/p/itmeh4grg...	AW Bellies	["Footwear >> Women's Footwear >> Ballerinas >...	SHOEH4GRSUBJGZXE	999.0	499.0	["http://img5a.flixcart.com/image/shoe/7/z/z/r...	False
3	0973b37acd0c664e3de26e97e5571454	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	SRTEH2F6HUZMQ6SJ	699.0	267.0	["http://img5a.flixcart.com/image/short/6/2/h/...	False
4	bc940ea42ee6bef5ac7cea3fb5cfbee7	2016-03-25 22:59:23 +0000	http://www.flipkart.com/sicons-all-purpose-arn...	Sicons All Purpose Arnica Dog Shampoo	["Pet Supplies >> Grooming >> Skin & Coat Care...	PSOEH3ZYDMSYARJ5	220.0	210.0	["http://img5a.flixcart.com/image/pet-shampoo/...	False

In [12]:
```python
top_product_category = df['product_category'].value_counts().head()
print(top_product_category)
```

```
product_category
["Clothing                  6198
["Jewellery                 3531
["Footwear                  1227
["Mobiles & Accessories     1099
["Automotive                1012
Name: count, dtype: int64
```

In [13]:
```python
# now its time to visualization
import matplotlib.pyplot as plt
import seaborn as sns
```

In [14]:
```python
# here we use barplot for visualization
plt.figure(figsize=(10,5))
sns.barplot(data=df, x=top_product_category.index , y=top_product_category.values, palette= 'cool' ,width =0.5 )
plt.title('Top 5 Categories with the highest no of products')
plt.ylabel('No of products')
plt.show()
```



Observation :

1. Based on this statistics, the top 5 product categories on Flipkart, ranked by the number of listings, are Clothing with 6198 listings, followed by Jewellery with 3531 listings, Footwear with 1227 listings, Mobiles & Accessories with 1099 listings, and Automotive with 1012 listings.
2. This data highlights the dominance of these categories in Flipkart's inventory, with clothing being the most prominent.

Which are the top 5 brands with the most product listings?

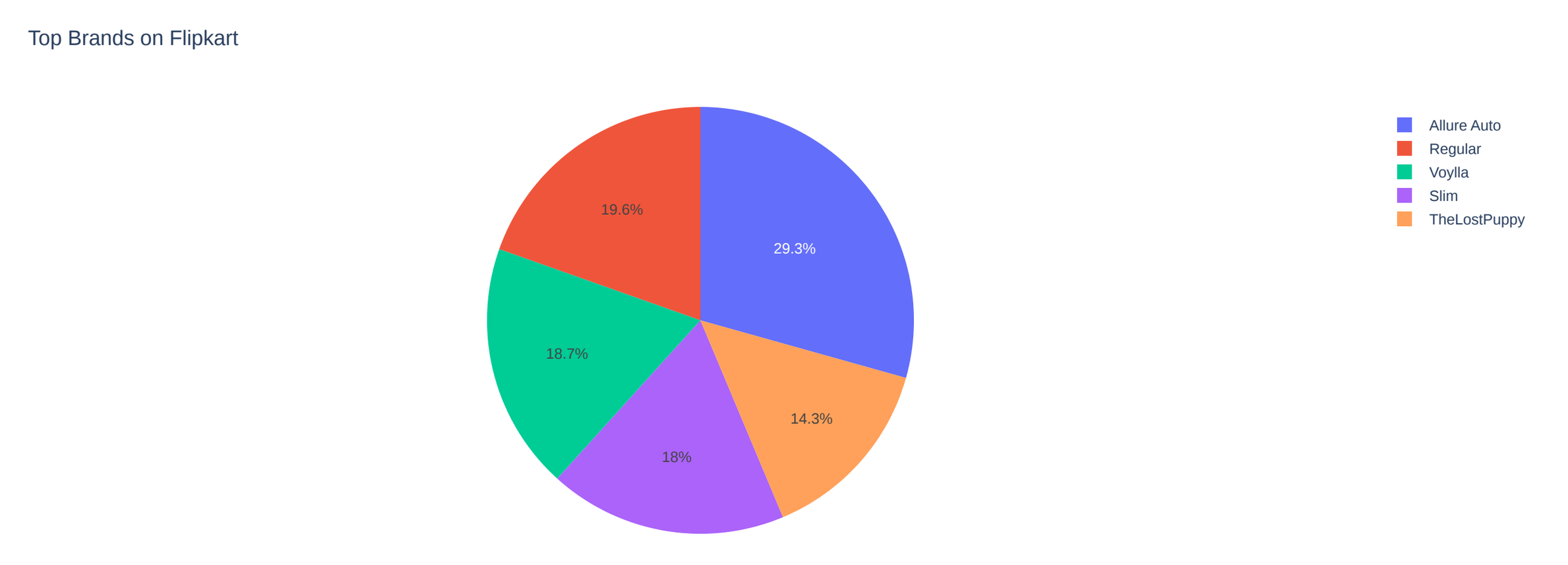
In [15]:
```python
df['brand'].value_counts().head().reset_index()        # this are the top 5 brands with maximum product listed
```

Out[15]:

	brand	count
0	Allure Auto	469
1	Regular	313
2	Voylla	299
3	Slim	288
4	TheLostPuppy	229

In [16]:
```python
# here we use pie chart for visualization
import plotly.express as px
```

In [17]:
```python
brand_counts = df['brand'].value_counts().head()
fig = px.pie(brand_counts, values=brand_counts.values, names=brand_counts.index, title='Top Brands on Flipkart')
fig.show()
```



Observation

1. The analysis reveals that Allure Auto secures top position among all the brands on flipkart with 469 product that is around 29.3% of the total listed products.
2. Following closely behind is Regular with 313product that is 19.6% of total listed products.
3. Voylla as the third most prominent brand with 299 (18.7%) products after that fourth is Slime with 288 (18%) products and last is The Lost Puppy with 229 (14.3%) products listed in flipkart.

What is the average discount offered by Flipkart on all products?

In [18]:
```python
df['discount_percentage'] = ((df['retail_price'] - df['discounted_price']) / df['retail_price'])*100
avg_discount = df['discount_percentage'].mean().round(decimals=2)
print("The average discount offered by Flipkart on all products is:", avg_discount, "%")
```

```
The average discount offered by Flipkart on all products is: 40.52 %
```

How many products have customer ratings?

In [19]:
```python
rating = df[df['product_rating'] != 'No rating available'].shape[0]
print (rating , 'no of products have actual customer ratings ')
```

```
1849 no of products have actual customer ratings
```

What percentage of products are part of the Flipkart Advantage program?

In [20]:
```python
df['is_FK_Advantage_product'].value_counts()
```

Out[20]:
```
is_FK_Advantage_product
False    19215
True       785
Name: count, dtype: int64
```

In [21]:
```python
# Calculate the percentage of True values present in the 'is_FK_Advantage_product' column using the mean() method
FK_Advantage_products = df['is_FK_Advantage_product'].mean() * 100
print(FK_Advantage_products,"% of products are part of the Flipkart Advantage program.")
```

```
3.925 % of products are part of the Flipkart Advantage program.
```

Which are the top 3 most expensive products listed on Flipkart?

In [22]:
```python
z = df['retail_price'].nlargest(3).reset_index()
print(z)
```

```
   index  retail_price
0    116      571230.0
1  11631      250500.0
2  11507      217500.0
```

Which brands have the highest average product rating?

In [23]:
```python
# Replace the 'No rating abailable' values from 'product_rating' with 0
df['product_rating'] = df['product_rating'].replace('No rating available', '0')
```

In [24]:
```python
# Change the datatype of the column to float
df['product_rating'] = df['product_rating'].astype('float')
df.product_rating.dtype
```

Out[24]: dtype('float64')

In [25]:
```python
# calculateing mean of each brand and sorting them in ascending order to find final brand rating
df.groupby('brand')['product_rating'].mean().sort_values(ascending=False).reset_index()
```

Out[25]:

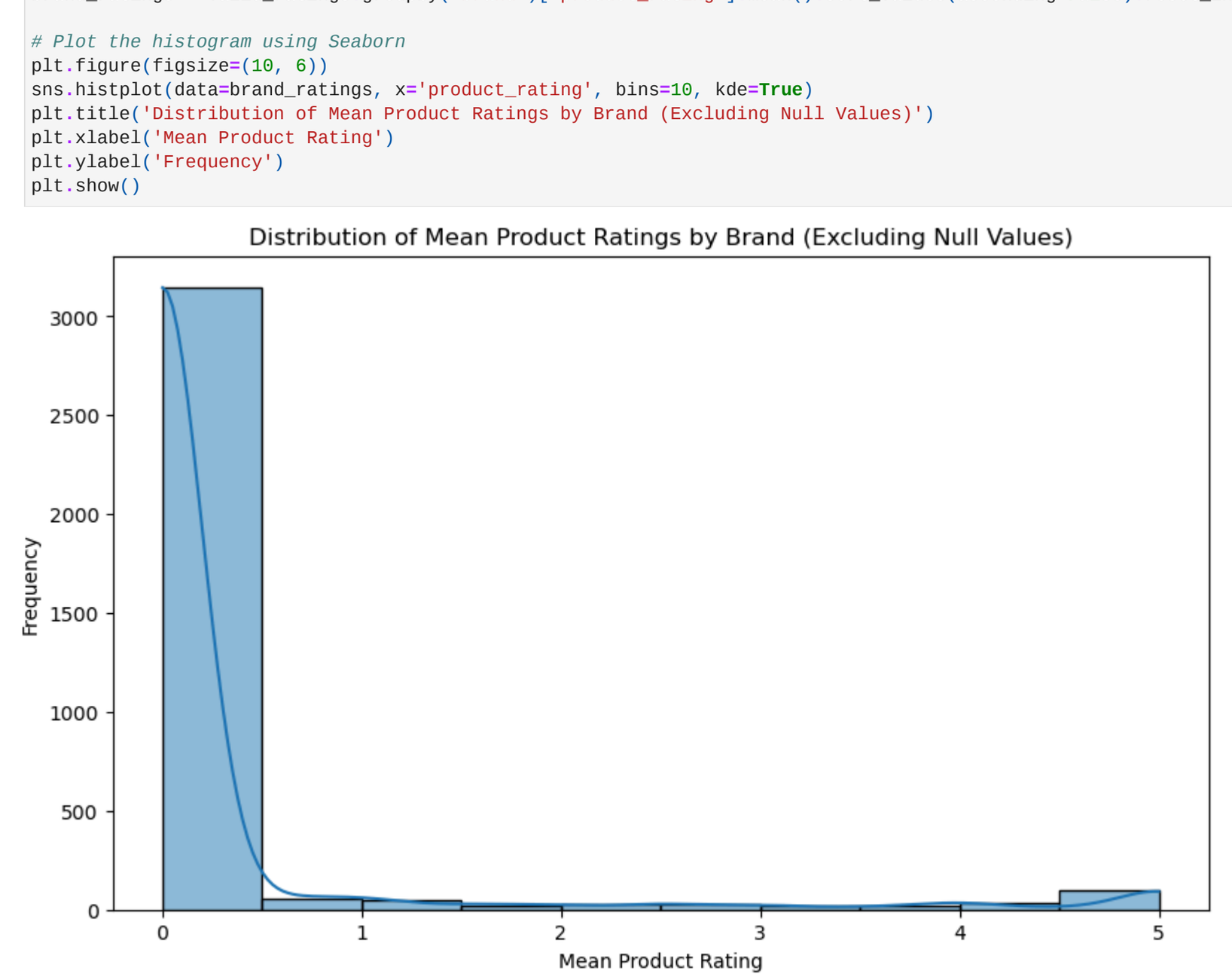
	brand	product_rating
0	Jewels Guru	5.0
1	ASIAN	5.0
2	Iball	5.0
3	Bond Beatz	5.0
4	METMO	5.0
...	...	...
3494	Home Creations	0.0
3495	Home Decor Line	0.0
3496	Home Delight	0.0
3497	Home Fashion Gallery	0.0
3498	Tarkan	0.0

3499 rows × 2 columns

In [26]:
```python
# Exclude null values from the product_rating column
valid_ratings = df.dropna(subset=['product_rating'])      #https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.dropna.html

# Calculate the mean product rating for each brand and sort the values in descending order
brand_ratings = valid_ratings.groupby('brand')['product_rating'].mean().sort_values(ascending=False).reset_index()

# Plot the histogram using Seaborn
plt.figure(figsize=(10, 6))
sns.histplot(data=brand_ratings, x='product_rating', bins=10, kde=True)
plt.title('Distribution of Mean Product Ratings by Brand (Excluding Null Values)')
plt.xlabel('Mean Product Rating')
plt.ylabel('Frequency')
plt.show()
```



Observation

1. Among the all brands, the highest average rating is 5.0. Additionally, only few brands achieved a rating of 4.8, while some brands received ratings of 4.5 and 4.6, respectively.
2. Lots of products also receive zero rating.
3. we can not make any decision after looking at the above graph.

Which product category has the highest average retail price?

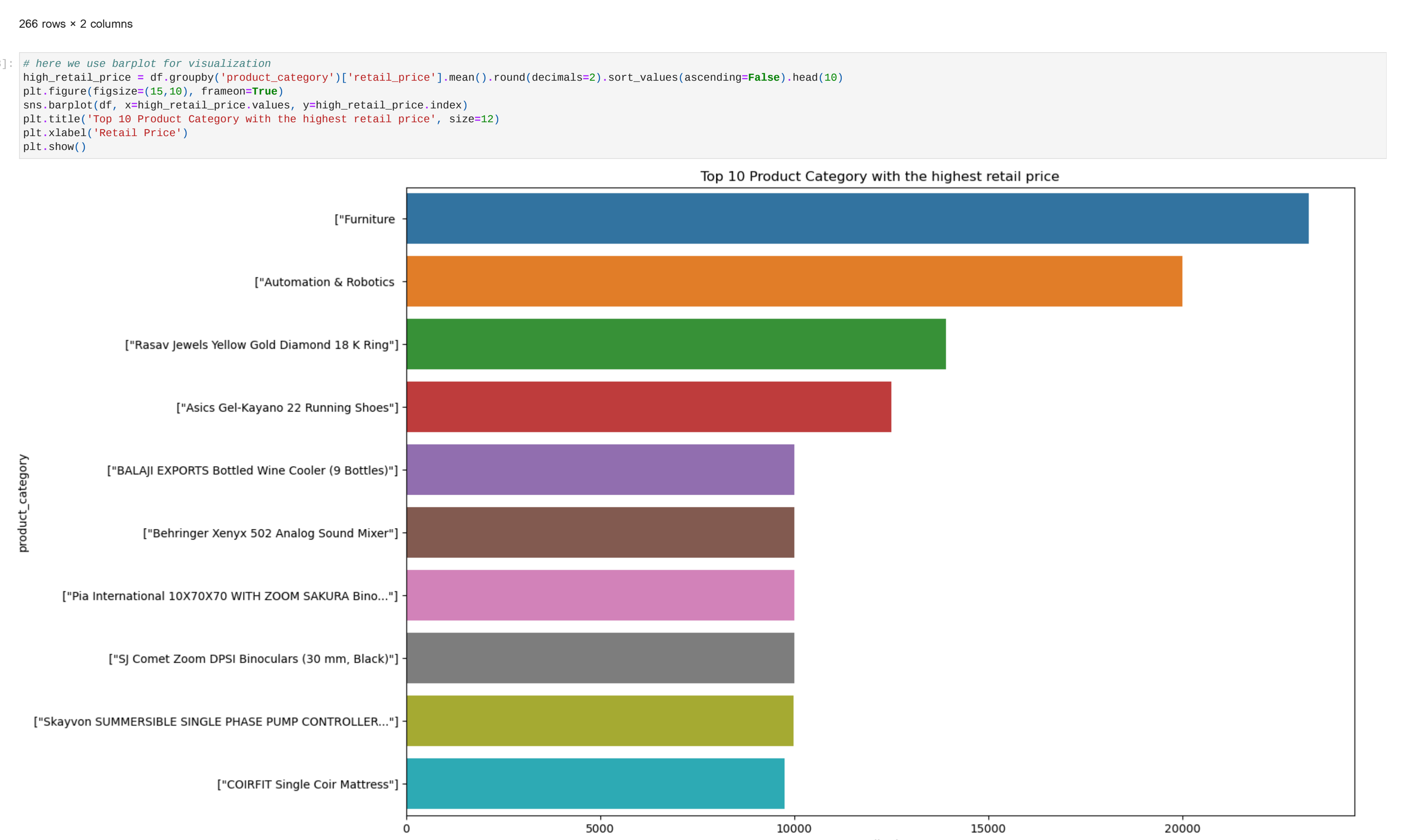
In [27]:
```python
# Calculate average retail price per category, sort categories by average retail price (highest first)
df.groupby('product_category')['retail_price'].mean().round(decimals=2).sort_values(ascending=False).reset_index()
```

Out[27]:

	product_category	retail_price
0	["Furniture	23262.97
1	["Automation & Robotics	19999.00
2	["Rasav Jewels Yellow Gold Diamond 18 K Ring"]	13903.00
3	["Asics Gel-Kayano 22 Running Shoes"]	12499.00
4	["BALAJI EXPORTS Bottled Wine Cooler (9 Bottle...	10000.00
...	...	...
261	["Siemens 5SL Betagard 5SL MCB (1)"]	197.00
262	["SUPERMOD Men's Brief"]	139.00
263	["Disney Printed Baby Boy's Hooded Grey T-Shirt"]	NaN
264	["INKT INKT A5 Wiro Notebook A5 Notebook Ring ...	NaN
265	["fourgee Slim Fit Boy's Black Jeans"]	NaN

266 rows × 2 columns

In [28]:
```python
# here we use barplot for visualization
high_retail_price = df.groupby('product_category')['retail_price'].mean().round(decimals=2).sort_values(ascending=False).head(10)
plt.figure(figsize=(15,10), frameon=True)
sns.barplot(df, x=high_retail_price.values, y=high_retail_price.index)
plt.title('Top 10 Product Category with the highest retail price', size=12)
plt.xlabel('Retail Price')
plt.show()
```



Observation :

1. The product category with the highest average retail price falls within the range of 9,000 to 20,000 INR.
2. This category includes items such as Automation & Robotics, premium jewelry pieces, high-end running shoes, and specialized optical equipment.
3. These findings suggest that products requiring advanced technology, intricate craftsmanship, or specialized features tend to command higher retail prices on Flipkart.

Which products have the longest and shortest descriptions?

In [29]:
```python
max_length = df['description'].str.len().nlargest()
min_length = df['description'].str.len().nsmallest()
print ('Index of the longest description is', max_length )
print ('####################################')
print ('Index of the shortest descrtion is', min_length )
```

```
Index of the longest description is 439      5309.0
457      5300.0
1481     4692.0
18752    4494.0
18267    4467.0
Name: description, dtype: float64
####################################
Index of the shortest descrtion is 10952    74.0
10768    79.0
19241    79.0
10797    83.0
10809    84.0
Name: description, dtype: float64
```

Observation :

1. The longest description belongs to products with index numbers 439 and 457, with a length of 5309.0 characters.
2. Multiple products share the title of having the shortest descriptions. Products with index no 10952, 10768, 19241, 10797, and 10809 all have descriptions that are only 74 to 84 characters long.

In which month was the data mostly crawled?

In [30]:
```python
# as we seen data type of crawl_timestamp is object so first we need to change it
df['crawl_timestamp'] = pd.to_datetime(df['crawl_timestamp'])          #https://pandas.pydata.org/pandas-docs/version/1.5/reference/api/pandas.to_datetime.html
df.crawl_timestamp.dtype
```

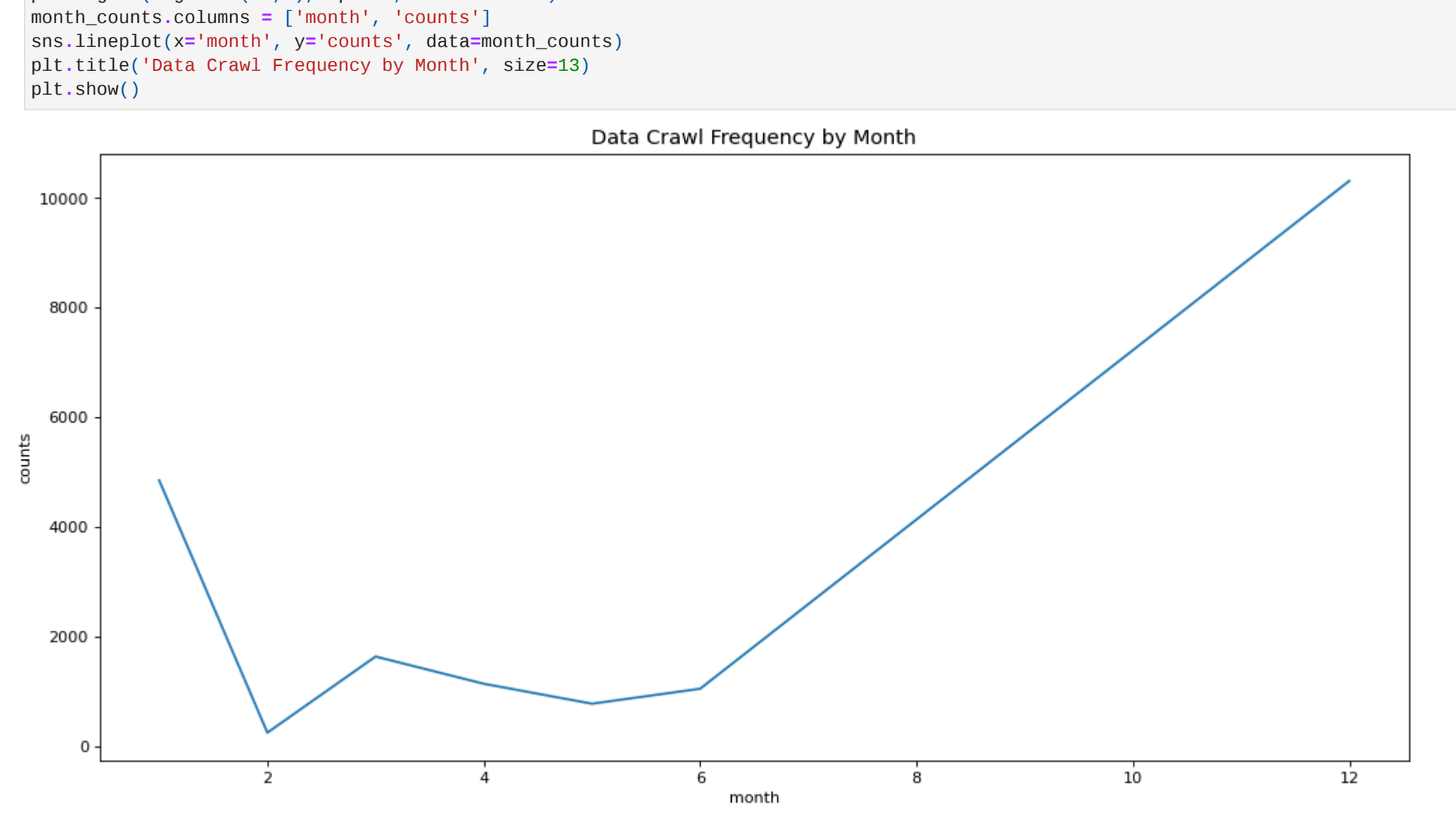
Out[30]: datetime64[ns, UTC]

In [31]:
```python
df['crawl_timestamp'].dt.month.value_counts().reset_index()
```

Out[31]:

	crawl_timestamp	count
0	12	10315
1	1	4850
2	3	1634
3	4	1137
4	6	1046
5	5	773
6	2	245

In [32]:
```python
# here we use lineplot for visualization
month_counts = df['crawl_timestamp'].dt.month.value_counts().reset_index()
plt.figure(figsize=(15,7), dpi=80, frameon=True)
month_counts.columns = ['month', 'counts']
sns.lineplot(x='month', y='counts', data=month_counts)
plt.title('Data Crawl Frequency by Month', size=13)
plt.show()
```



Observation :

1. The data indicates that the crawling activity peaked in December, with 10,315 instances recorded, suggesting that December was the month when the dataset was most extensively crawled.
2. The crawling activity was comparatively lower in other months, with January having the next highest count of 4,850 instances.

Conclusion

> **Top Brands**: **Allure Auto** leads with 469 products (29.3%), followed by **Regular** (313 products, 19.6%), **Voylla** (299 products, 18.7%), **Slime** (288 products, 18%), and **The Lost Puppy** (229 products, 14.3%).

> **Top Product Categories**: The data indicates a diverse range of product categories on Flipkart, with **clothing** has the highest number of listings (6,198), followed by **Jewellery** (3,531), **Footwear** (1,227), **Mobiles & Accessories** (1,099), and **Automotive** (1,012).

Despite a wide range of product categories, only a small **0.600%** are part of the Flipkart Advantage program.

Products with higher retail prices typically involve advanced technology or specialized features.

Crawling activity peaked in December with 10,210 instances recorded, indicating extensive dataset collection during that month. January had the next highest count of 4,930 instances, showing comparatively lower crawling activity in other months.