1. Amazon Sales Data

Description

This dataset contains information on **1K+** Amazon products, including their ratings, reviews, and other details.

Features

- **product_id**: Unique identifier for each product
- **product_name**: Name of the product
- **category**: Category of the product
- **discounted_price**: Discounted price of the product
- **actual_price**: Actual price of the product
- **discount_percentage**: Percentage of discount for the product
- **rating**: Rating of the product (1-5)
- **rating_count**: Number of people who voted for the Amazon rating
- **about_product**: Description of the product
- **user_id**: ID of the user who wrote the review
- **user_name**: Name of the user who wrote the review
- **review_id**: Unique identifier for the review
- **review_title**: Short review title
- **review_content**: Full content of the review
- **img_link**: Image link of the product
- **product_link**: Official website link for the product

Source

Amazon Sales Data

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [ ]:
```
df1=pd.read_csv("./amazon.csv")
df1
```

Out[ ]:

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

1465 rows × 16 columns

Q1. What is the average rating for each product category?

In [ ]:
```
df1.columns
```

Out[ ]:
```
Index(['product_id', 'product_name', 'category', 'discounted_price',
       'actual_price', 'discount_percentage', 'rating', 'rating_count',
       'about_product', 'user_id', 'user_name', 'review_id', 'review_titl
e',
       'review_content', 'img_link', 'product_link'],
      dtype='object')
```

In [ ]:
```
df1["rating"] = df1["rating"].replace('|', float('nan')).astype("float64")
```

In [ ]:
```
df1["rating"]
```

Out[ ]:
```
0       4.2
1       4.0
2       3.9
3       4.2
4       4.2
       ...
1460    4.0
1461    4.1
1462    3.6
1463    4.0
1464    4.3
Name: rating, Length: 1465, dtype: float64
```

Average rating for each category:

In [ ]:
```
average_rating=df1.groupby("category")["rating"].mean()
```

In [ ]:
```
average_rating.round(2)
```

Out[ ]:
```
category
Car&Motorbike|CarAccessories|InteriorAccessories|AirPurifiers&Ionizers
3.80
Computers&Accessories|Accessories&Peripherals|Adapters|USBtoUSBAdapters
4.15
Computers&Accessories|Accessories&Peripherals|Audio&VideoAccessories|PCHead
sets                                                    3.50
Computers&Accessories|Accessories&Peripherals|Audio&VideoAccessories|PCMicr
ophones                                                 3.60
Computers&Accessories|Accessories&Peripherals|Audio&VideoAccessories|PCSpea
kers                                                    4.05

...
OfficeProducts|OfficePaperProducts|Paper|Stationery|Pens,Pencils&WritingSup
plies|Pens&Refills|GelInkRollerballPens        4.25
OfficeProducts|OfficePaperProducts|Paper|Stationery|Pens,Pencils&WritingSup
plies|Pens&Refills|LiquidInkRollerballPens     4.15
OfficeProducts|OfficePaperProducts|Paper|Stationery|Pens,Pencils&WritingSup
plies|Pens&Refills|RetractableBallpointPens    4.30
OfficeProducts|OfficePaperProducts|Paper|Stationery|Pens,Pencils&WritingSup
plies|Pens&Refills|StickBallpointPens          4.13
Toys&Games|Arts&Crafts|Drawing&PaintingSupplies|ColouringPens&Markers
4.30
Name: rating, Length: 211, dtype: float64
```

In [ ]:
```
print("Overall average rating:",average_rating.mean())
```

```
Overall average rating: 4.129893157821581
```

In [ ]:
```
df1.columns
```

Out[ ]:
```
Index(['product_id', 'product_name', 'category', 'discounted_price',
       'actual_price', 'discount_percentage', 'rating', 'rating_count',
       'about_product', 'user_id', 'user_name', 'review_id', 'review_titl
e',
       'review_content', 'img_link', 'product_link'],
      dtype='object')
```

Q2 What are the top rating_count products by category?

In [ ]:
```
df1["rating_count"].fillna(0, inplace=True)
```

In [ ]:
```
df1["rating_count"].replace(",", "", regex=True, inplace=True)
#df1["rating"] = df1["rating"]
```

In [ ]:
```
df1["rating_count"].astype("int64")
```

Out[ ]:
```
0       24269
1       43994
2        7928
3       94363
4       16905
        ...
1460     1090
1461     4118
1462      468
1463     8031
1464     6987
Name: rating_count, Length: 1465, dtype: int64
```

In [ ]:
```
df1["rating_count"] = df1["rating_count"].astype(int)
df_2 = df1[df1["rating_count"].notnull()]
```

In [ ]:
```
df_2
```

Out[ ]:

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

1465 rows × 16 columns

Top 10 rating_count products by rating_counts:

In [ ]:
```
df_2[["category","product_name","rating_count"]].nlargest(10,"rating_count")
```

Out[ ]:

	category	product_name	rating_cou
12	Electronics|HomeTheater,TV&Video|Accessories|C...	AmazonBasics Flexible Premium HDMI Cable (Blac...	4269
47	Electronics|HomeTheater,TV&Video|Accessories|C...	Amazon Basics High-Speed HDMI Cable, 6 Feet - ...	4269
65	Electronics|HomeTheater,TV&Video|Accessories|C...	Amazon Basics High-Speed HDMI Cable, 6 Feet (2...	4269
684	Electronics|HomeTheater,TV&Video|Accessories|C...	AmazonBasics Flexible Premium HDMI Cable (Blac...	4269
352	Electronics|Headphones,Earbuds&Accessories|Hea...	boAt Bassheads 100 in Ear Wired Earphones with...	3637
400	Electronics|Headphones,Earbuds&Accessories|Hea...	boAt Bassheads 100 in Ear Wired Earphones with...	3637
584	Electronics|Headphones,Earbuds&Accessories|Hea...	boAt BassHeads 100 in-Ear Wired Headphones wit...	3637
370	Electronics|Mobiles&Accessories|Smartphones&Ba...	Redmi 9 Activ (Carbon Black, 4GB RAM, 64GB Sto...	3138
371	Electronics|Mobiles&Accessories|Smartphones&Ba...	Redmi 9A Sport (Coral Green, 2GB RAM, 32GB Sto...	3138
473	Electronics|Mobiles&Accessories|Smartphones&Ba...	Redmi 9A Sport (Carbon Black, 2GB RAM, 32GB St...	3138

In [ ]:
```
df_2[["category","product_name","rating_count"]].sort_values(by="rating_cour
```

Out[ ]:

	category	product_name	rating_
12	Electronics|HomeTheater,TV&Video|Accessories|C...	AmazonBasics Flexible Premium HDMI Cable (Blac...	4
65	Electronics|HomeTheater,TV&Video|Accessories|C...	Amazon Basics High-Speed HDMI Cable, 6 Feet (2...	4
47	Electronics|HomeTheater,TV&Video|Accessories|C...	Amazon Basics High-Speed HDMI Cable, 6 Feet - ...	4
684	Electronics|HomeTheater,TV&Video|Accessories|C...	AmazonBasics Flexible Premium HDMI Cable (Blac...	4
400	Electronics|Headphones,Earbuds&Accessories|Hea...	boAt Bassheads 100 in Ear Wired Earphones with...	3
...	...	...	
1344	Home&Kitchen|Heating,Cooling&AirQuality|RoomHe...	Longway Blaze 2 Rod Quartz Room Heater (White,...	
1309	Home&Kitchen|Heating,Cooling&AirQuality|RoomHe...	Khaitan ORFin Fan heater for Home and kitchen-...	
1459	Home&Kitchen|Kitchen&HomeAppliances|Vacuum,Cle...	NGI Store 2 Pieces Pet Hair Removers for Your ...	
324	Computers&Accessories|Accessories&Peripherals|...	REDTECH USB-C to Lightning Cable 3.3FT, [Apple...	
282	Computers&Accessories|Accessories&Peripherals|...	Amazon Brand - Solimo 65W Fast Charging Braide...	

1465 rows × 3 columns

In [ ]:

```
#top 10 rating_count products by category:
top_10_products = df_2.groupby("category").apply(lambda x: x.nlargest(10, "r
top_10_products = top_10_products[["category", "product_name", "rating_count
top_10_products.sort_values(by="rating_count",ascending=False)[0:20]
```

Out[ ]:

	category	product_name	rating_cou
288	Electronics|HomeTheater,TV&Video|Accessories|C...	Amazon Basics High-Speed HDMI Cable, 6 Feet - ...	4269
287	Electronics|HomeTheater,TV&Video|Accessories|C...	AmazonBasics Flexible Premium HDMI Cable (Blac...	4269
289	Electronics|HomeTheater,TV&Video|Accessories|C...	Amazon Basics High-Speed HDMI Cable, 6 Feet (2...	4269
290	Electronics|HomeTheater,TV&Video|Accessories|C...	AmazonBasics Flexible Premium HDMI Cable (Blac...	4269
249	Electronics|Headphones,Earbuds&Accessories|Hea...	boAt Bassheads 100 in Ear Wired Earphones with...	3637
250	Electronics|Headphones,Earbuds&Accessories|Hea...	boAt Bassheads 100 in Ear Wired Earphones with...	3637
251	Electronics|Headphones,Earbuds&Accessories|Hea...	boAt BassHeads 100 in-Ear Wired Headphones wit...	3637
426	Electronics|Mobiles&Accessories|Smartphones&Ba...	Redmi 9 Activ (Carbon Black, 4GB RAM, 64GB Sto...	3138
427	Electronics|Mobiles&Accessories|Smartphones&Ba...	Redmi 9A Sport (Coral Green, 2GB RAM, 32GB Sto...	3138
429	Electronics|Mobiles&Accessories|Smartphones&Ba...	Redmi 9A Sport (Coral Green, 3GB RAM, 32GB Sto...	3138
428	Electronics|Mobiles&Accessories|Smartphones&Ba...	Redmi 9A Sport (Carbon Black, 2GB RAM, 32GB St...	3138

	category	product_name	rating_cou
252	Electronics|Headphones,Earbuds&Accessories|Hea...	boAt Bassheads 225 in Ear Wired Earphones with...	2731
548	Home&Kitchen|Kitchen&Dining|KitchenTools|Manua...	Pigeon Polypropylene Mini Handy and Compact Ch...	2705
153	Computers&Accessories|ExternalDevices&DataStor...	SanDisk Cruzer Blade 32GB USB Flash Drive	2531
204	Electronics|Accessories|MemoryCards|MicroSD	SanDisk Extreme SD UHS I 64GB Card for 4K Vide...	2050
253	Electronics|Headphones,Earbuds&Accessories|Hea...	JBL C100SI Wired In Ear Headphones with Mic, J...	1925
254	Electronics|Headphones,Earbuds&Accessories|Hea...	JBL C100SI Wired In Ear Headphones with Mic, J...	1925
255	Electronics|Headphones,Earbuds&Accessories|Hea...	JBL C100SI Wired In Ear Headphones with Mic, J...	1925
154	Computers&Accessories|ExternalDevices&DataStor...	SanDisk Ultra Dual 64 GB USB 3.0 OTG Pen Drive...	1891
256	Electronics|Headphones,Earbuds&Accessories|Hea...	boAt Airdopes 121v2 in-Ear True Wireless Earbu...	1809

Q3 What is the distribution of discounted prices vs. actual prices?

In [ ]:
```
df1.columns
```

Out[ ]:
```
Index(['product_id', 'product_name', 'category', 'discounted_price',
       'actual_price', 'discount_percentage', 'rating', 'rating_count',
       'about_product', 'user_id', 'user_name', 'review_id', 'review_titl
e',
       'review_content', 'img_link', 'product_link'],
      dtype='object')
```

In [ ]:
```
df1[["discounted_price","actual_price"]]
```

Out[ ]:

	discounted_price	actual_price
0	₹399	₹1,099
1	₹199	₹349
2	₹199	₹1,899
3	₹329	₹699
4	₹154	₹399
...	...	...
1460	₹379	₹919
1461	₹2,280	₹3,045
1462	₹2,219	₹3,080
1463	₹1,399	₹1,890
1464	₹2,863	₹3,690

1465 rows × 2 columns

In [ ]:
```
DP=df1["discounted_price"].str.replace("₹","").str.replace(",","").astype("f
DP
```

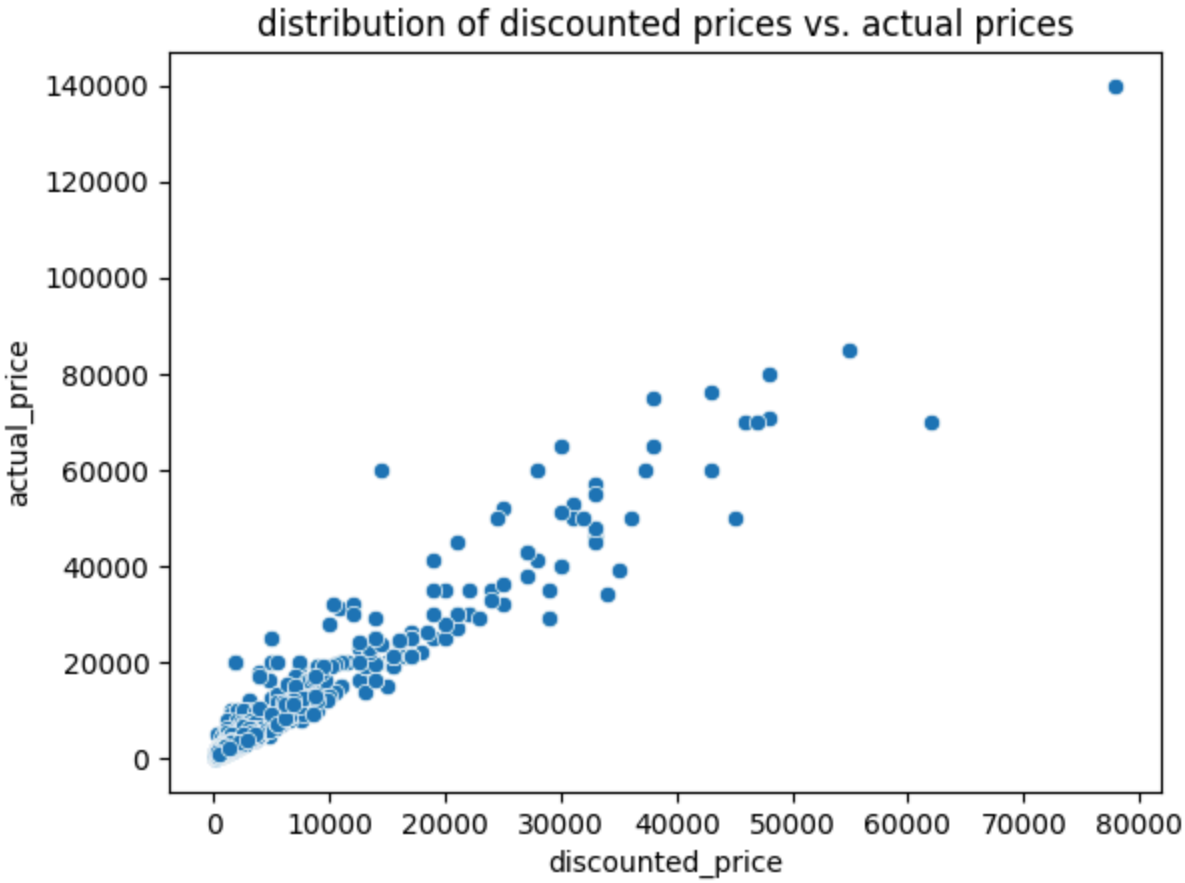
Out[ ]:
```
0        399.0
1        199.0
2        199.0
3        329.0
4        154.0
         ...
1460     379.0
1461    2280.0
1462    2219.0
1463    1399.0
1464    2863.0
Name: discounted_price, Length: 1465, dtype: float64
```

In [ ]:
```
AP=df1["actual_price"].str.replace("₹","").str.replace(",","").astype("float
AP
```

Out[ ]:
```
0       1099.0
1        349.0
2       1899.0
3        699.0
4        399.0
         ...
1460     919.0
1461    3045.0
1462    3080.0
1463    1890.0
1464    3690.0
Name: actual_price, Length: 1465, dtype: float64
```

In [ ]:
```
sns.scatterplot(x=DP,y=AP,palette="deep")
plt.title("distribution of discounted prices vs. actual prices")
plt.show()
```

```
<ipython-input-23-638b52f7ab08>:1: UserWarning: Ignoring `palette` because n
o `hue` variable has been assigned.
  sns.scatterplot(x=DP,y=AP,palette="deep")
```



Insight: As we can clearly see through the Scatterplot, as the actual price increases,Discounted price increases simultaneously. More precisely actual price and discounted price are linearly related.

Q4. How does the average discount percentage vary across categories?

In [ ]:
```
df1.columns
```

Out[ ]:
```
Index(['product_id', 'product_name', 'category', 'discounted_price',
       'actual_price', 'discount_percentage', 'rating', 'rating_count',
       'about_product', 'user_id', 'user_name', 'review_id', 'review_titl
e',
       'review_content', 'img_link', 'product_link'],
      dtype='object')
```

In [ ]:
```
df1['discount_percentage'] = df1['discount_percentage'].str.replace("%", "")
average_discount_by_category = df1.groupby("category")["discount_percentage"
```

In [ ]:
```
average_discount_by_category.sort_values(ascending=False)
```

Out[ ]:
```
category
Electronics|Mobiles&Accessories|MobileAccessories|Décor|PhoneCharms
90.0
Computers&Accessories|Accessories&Peripherals|Cables&Accessories|CableConne
ctionProtectors                            90.0
Electronics|Headphones,Earbuds&Accessories|Earpads
90.0
Electronics|Headphones,Earbuds&Accessories|Adapters
88.0
Computers&Accessories|Accessories&Peripherals|Keyboards,Mice&InputDevices|K
eyboard&MiceAccessories|DustCovers       87.5

...
OfficeProducts|OfficeElectronics|Calculators|Basic
0.0
Home&Kitchen|Kitchen&HomeAppliances|SmallKitchenAppliances|SmallAppliancePa
rts&Accessories|StandMixerAccessories     0.0
Electronics|HomeAudio|MediaStreamingDevices|StreamingClients
0.0
Electronics|Cameras&Photography|Accessories|Film
0.0
Toys&Games|Arts&Crafts|Drawing&PaintingSupplies|ColouringPens&Markers
0.0
Name: discount_percentage, Length: 211, dtype: float64
```

In [ ]:
```
d_mean=df1["discount_percentage"].mean()
d_mean
```

Out[ ]:
```
47.69146757679181
```

In [ ]:
```
more_than_average = (average_discount_by_category >= average_discount_by_cat
less_than_average= (average_discount_by_category < average_discount_by_categ
print("more_than_average:",more_than_average)
print("less_than_average:",less_than_average)
```

```
more_than_average: 115
less_than_average: 96
```

Insight:Hence as per the study, there are 115 out of 211 categories which are having "average discount percentage" greater than or equal to overall mean viz: 47.69, while on the other hand, there are 96 categories out of 211, which are having less "average discount percentage" than 47.69 (overall mean).

Q5. What are the most popular product names?

In [ ]:
```
df1.columns
```

Out[ ]:
```
Index(['product_id', 'product_name', 'category', 'discounted_price',
       'actual_price', 'discount_percentage', 'rating', 'rating_count',
       'about_product', 'user_id', 'user_name', 'review_id', 'review_titl
e',
       'review_content', 'img_link', 'product_link'],
      dtype='object')
```

In [ ]:
```
df_5=df1[["product_name","rating","rating_count"]]
```

In [ ]:
```
df_5.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1465 entries, 0 to 1464
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   product_name  1465 non-null   object
 1   rating        1464 non-null   float64
 2   rating_count  1465 non-null   int64 
dtypes: float64(1), int64(1), object(1)
memory usage: 34.5+ KB
```

In [ ]:
```
#As the most popular product also have highest rating count
Most_popular_products=df_5.sort_values(by="rating_count",ascending=False)
```

Finding top 20 popular product

In [ ]:
```
Most_popular_products.head(20)
```

Out[ ]:

	product_name	rating	rating_count
12	AmazonBasics Flexible Premium HDMI Cable (Blac...	4.4	426973
65	Amazon Basics High-Speed HDMI Cable, 6 Feet (2...	4.4	426973
47	Amazon Basics High-Speed HDMI Cable, 6 Feet - ...	4.4	426973
684	AmazonBasics Flexible Premium HDMI Cable (Blac...	4.4	426972
400	boAt Bassheads 100 in Ear Wired Earphones with...	4.1	363713
352	boAt Bassheads 100 in Ear Wired Earphones with...	4.1	363713
584	boAt BassHeads 100 in-Ear Wired Headphones wit...	4.1	363711
370	Redmi 9 Activ (Carbon Black, 4GB RAM, 64GB Sto...	4.1	313836
371	Redmi 9A Sport (Coral Green, 2GB RAM, 32GB Sto...	4.1	313836
473	Redmi 9A Sport (Carbon Black, 2GB RAM, 32GB St...	4.1	313832
566	Redmi 9A Sport (Coral Green, 3GB RAM, 32GB Sto...	4.1	313832
760	boAt Bassheads 225 in Ear Wired Earphones with...	4.1	273189
1028	Pigeon Polypropylene Mini Handy and Compact Ch...	4.1	270563
588	SanDisk Cruzer Blade 32GB USB Flash Drive	4.3	253105
864	SanDisk Extreme SD UHS I 64GB Card for 4K Vide...	4.5	205052
347	JBL C100SI Wired In Ear Headphones with Mic, J...	4.1	192590
479	JBL C100SI Wired In Ear Headphones with Mic, J...	4.1	192589
598	JBL C100SI Wired In Ear Headphones with Mic, J...	4.1	192587
718	SanDisk Ultra Dual 64 GB USB 3.0 OTG Pen Drive...	4.3	189104
591	boAt Airdopes 121v2 in-Ear True Wireless Earbu...	3.8	180998

Q6. What are the most popular product keywords?

In [ ]:
```
df1.columns
```

Out[ ]:
```
Index(['product_id', 'product_name', 'category', 'discounted_price',
       'actual_price', 'discount_percentage', 'rating', 'rating_count',
       'about_product', 'user_id', 'user_name', 'review_id', 'review_titl
e',
       'review_content', 'img_link', 'product_link'],
      dtype='object')
```

In [ ]:
```
# Assuming your DataFrame is named 'amazon_sales_data'
keywords = df1["product_name"].str.cat(df1["about_product"],sep=" ",na_rep="
#Flatten the list
word_list=[word for sublist in keywords for word in sublist]
#Count the frequency of words
word_count=pd.Series(word_list)
word_count.value_counts()
```

```
#Top 20 most popular product keywords used are:
top_keywords=word_count.value_counts().head(20)
print(top_keywords)
```

```
and      5590
the      4342
to       3995
with     3837
for      3226
of       2200
a        1905
|        1888
your     1880
is       1481
in       1372
usb      1346
&        1326
you      1256
cable    1026
or        983
on        980
it        956
-         900
can       891
Name: count, dtype: int64
```

Q7. What are the most popular product reviews?

In [ ]:
```
df1.columns
```

Out[ ]:
```
Index(['product_id', 'product_name', 'category', 'discounted_price',
       'actual_price', 'discount_percentage', 'rating', 'rating_count',
       'about_product', 'user_id', 'user_name', 'review_id', 'review_titl
e',
       'review_content', 'img_link', 'product_link'],
      dtype='object')
```

In [ ]:
```
popular_reviews = df1.sort_values(by='rating_count', ascending=False)[['proc
popular_reviews
```

Out[ ]:

	product_name	about_product	rating_count	review_title	review_content
12	AmazonBasics Flexible Premium HDMI Cable (Blac...	Flexible, lightweight HDMI cable for connectin...	426973	It's quite good and value for money,Works well...	I am using it for 14 days now The experience ..
65	Amazon Basics High-Speed HDMI Cable, 6 Feet (2...	HDMI A Male to A Male Cable: Supports Ethernet...	426973	It's quite good and value for money,Works well...	I am using it for 14 days now The experience ..
47	Amazon Basics High-Speed HDMI Cable, 6 Feet - ...	Please select appropriate display resolution &...	426973	It's quite good and value for money,Works well...	I am using it for 14 days now The experience ..
684	AmazonBasics Flexible Premium HDMI Cable (Blac...	Flexible, lightweight HDMI cable for connectin...	426972	It's quite good and value for money,Works well...	I am using it for 14 days now The experience ..
400	boAt Bassheads 100 in Ear Wired Earphones with...	The perfect way to add some style and stand ou...	363713	Best value for money,HEAD PHONE POUCH NOT RECE...	The sound quality of this earphone are really ..

Q8. What is the correlation between discounted_price and rating?

In [ ]:
```
df1.head(1)
```

Out[ ]:

	product_id	product_name	category	d
0	B07JW9H4J1	Wayona Nylon Braided USB to Lightning Fast Cha...	Computers&Accessories|Accessories&Peripherals|...	

In [ ]:
```
import warnings
df_8=df1[["discounted_price","rating"]]
df_8["discounted_price"]=df_8["discounted_price"].str.replace("₹","").str.re
df_8
```

```
<ipython-input-41-d9cf8363ce57>: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
  df_8["discounted_price"]=df_8["discounted_price"].str.replace("₹","").str.replace(",","").astype(float)
```

Out[ ]:

	discounted_price	rating
0	399.0	4.2
1	199.0	4.0
2	199.0	3.9
3	329.0	4.2
4	154.0	4.2
...	...	...
1460	379.0	4.0
1461	2280.0	4.1
1462	2219.0	3.6
1463	1399.0	4.0
1464	2863.0	4.3

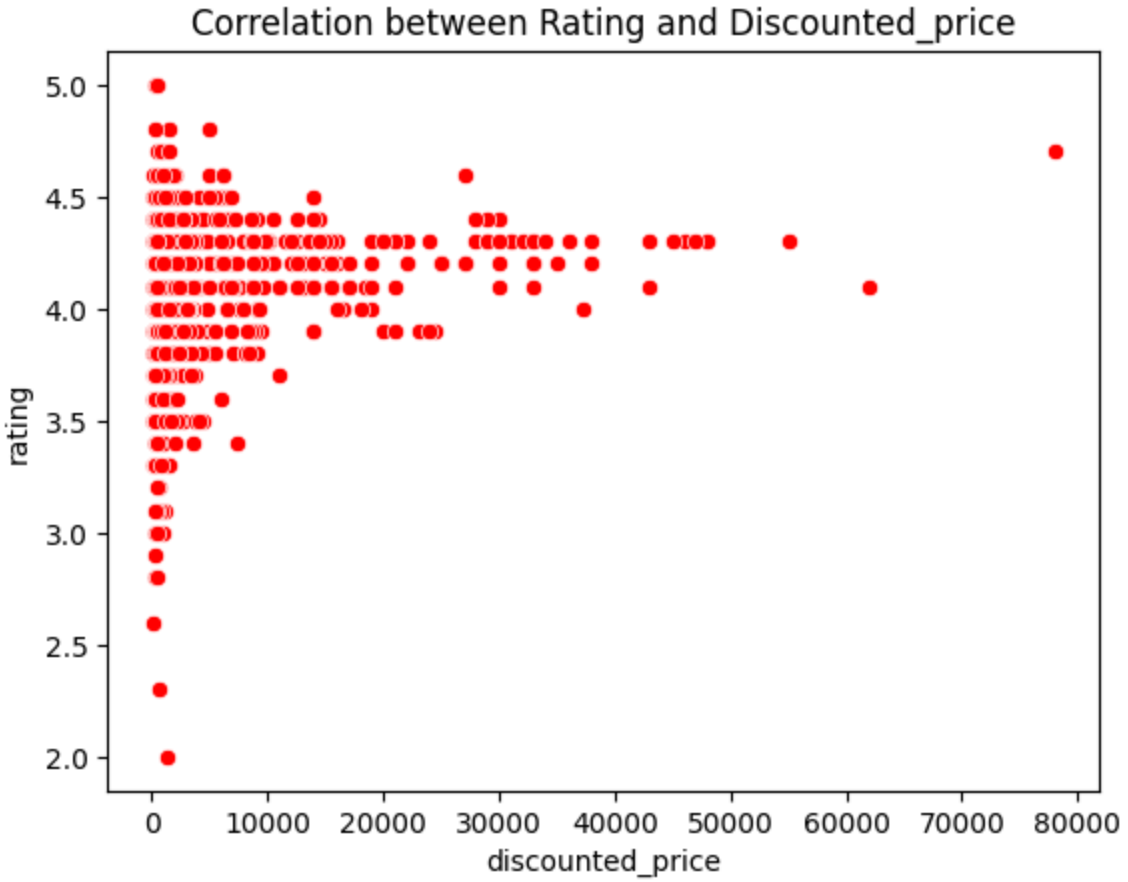
1465 rows × 2 columns

In [ ]:
```
df_8.corr()
```

Out[ ]:

	discounted_price	rating
discounted_price	1.000000	0.120337
rating	0.120337	1.000000

In [ ]:
```
sns.scatterplot(x=df_8["discounted_price"],y=df_8["rating"],color="r")
plt.title("Correlation between Rating and Discounted_price")
plt.show()
```



Insight and Conclusion: As per the above studies suggest that Rating and Discounted_price has very weak positive correlation of 0.12, which suggest that as the rating increases the increase in discounted_price also happens slightly, But a the correlation is not so strong we should keep other factors in mind too. There may be other factors such as product quality, brand reputation, or market demand that influence both the rating and the discounted price of a product. So, it's always a good idea to consider multiple factors before drawing any conclusions.

Q9. What are the Top 5 categories based on the highest ratings?

In [ ]:
```
df1.columns
```

Out[ ]:
```
Index(['product_id', 'product_name', 'category', 'discounted_price',
       'actual_price', 'discount_percentage', 'rating', 'rating_count',
       'about_product', 'user_id', 'user_name', 'review_id', 'review_titl
e',
       'review_content', 'img_link', 'product_link'],
      dtype='object')
```

In [ ]:
```
df_9=df1[["category","rating","rating_count"]]
df_9
```

Out[ ]:

	category	rating	rating_count
0	Computers&Accessories|Accessories&Peripherals|...	4.2	24269
1	Computers&Accessories|Accessories&Peripherals|...	4.0	43994
2	Computers&Accessories|Accessories&Peripherals|...	3.9	7928
3	Computers&Accessories|Accessories&Peripherals|...	4.2	94363
4	Computers&Accessories|Accessories&Peripherals|...	4.2	16905
...	...	...	...
1460	Home&Kitchen|Kitchen&HomeAppliances|WaterPurif...	4.0	1090
1461	Home&Kitchen|Kitchen&HomeAppliances|SmallKitch...	4.1	4118
1462	Home&Kitchen|Heating,Cooling&AirQuality|RoomHe...	3.6	468
1463	Home&Kitchen|Heating,Cooling&AirQuality|Fans|E...	4.0	8031
1464	Home&Kitchen|Kitchen&HomeAppliances|SmallKitch...	4.3	6987

1465 rows × 3 columns

In [ ]:
```
df_9.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1465 entries, 0 to 1464
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   category      1465 non-null   object 
 1   rating        1464 non-null   float64
 2   rating_count  1465 non-null   int64  
dtypes: float64(1), int64(1), object(1)
memory usage: 34.5+ KB
```

Top 5 category based on rating:

In [ ]:
```
df_9.sort_values(by="rating",ascending=False).head(5)
```

Out[ ]:

	category	rating	rating_count
324	Computers&Accessories|Accessories&Peripherals|...	5.0	0
174	Computers&Accessories|Accessories&Peripherals|...	5.0	5
775	Computers&Accessories|Accessories&Peripherals|...	5.0	23
1145	Home&Kitchen|Heating,Cooling&AirQuality|WaterH...	4.8	53803
1201	Home&Kitchen|Kitchen&HomeAppliances|SmallKitch...	4.8	28

Top 5 category based on rating_count:

In [ ]:
```
df_9.sort_values(by="rating_count",ascending=False).head(5)
```

Out[ ]:

	category	rating	rating_count
12	Electronics|HomeTheater,TV&Video|Accessories|C...	4.4	426973
65	Electronics|HomeTheater,TV&Video|Accessories|C...	4.4	426973
47	Electronics|HomeTheater,TV&Video|Accessories|C...	4.4	426973
684	Electronics|HomeTheater,TV&Video|Accessories|C...	4.4	426972
400	Electronics|Headphones,Earbuds&Accessories|Hea...	4.1	363713

Q10.Identify any potential areas for improvement or optimization based on the data analysis.

Answer - Based on the data analysis, some potential areas for improvement or optimization could be:

1. Improving the product's rating by addressing any issues mentioned in the reviews.
2. Increasing the number of rating counts to provide a more accurate representation of customer satisfaction.
3. Enhancing the product descriptions to better highlight its features and benefits.
4. Adjusting the pricing strategy to maximize the discount percentage and attract more customers.

2. Spotify Data: Popular Hip-hop Artists and Tracks

Description

The dataset titled **"Spotify Data: Popular Hip-hop Artists and Tracks"** provides a curated collection of approximately 500 entries showcasing the vibrant realm of hip-hop music. These entries meticulously compile the most celebrated hip-hop tracks and artists, reflecting their significant influence on the genre's landscape. Each entry highlights not only the popularity and musical composition of the tracks but also the creative prowess of the artists and their profound impact on global listeners.

Application in Data Science

This dataset serves as a valuable resource for various data science explorations:

- **Trend Analysis**: Analyze the popularity dynamics of hit hip-hop tracks over recent years.
- **Network Analysis**: Explore collaborative patterns among top artists and uncover insights into the genre's evolving collaborative landscape.
- **Predictive Modeling**: Develop models to forecast track popularity based on diverse features, offering insights for artists, producers, and marketers.

Features (Column Descriptors)

- **Artist**: The name of the artist, providing direct attribution to the creative mind behind the track.
- **Track Name**: The title of the track, encapsulating its identity and essence.
- **Popularity**: A numeric score reflecting the track's reception and appeal among Spotify listeners.
- **Duration (ms)**: The track's length in milliseconds, detailing the temporal extent of the musical experience.
- **Track ID**: A unique identifier within Spotify's ecosystem, enabling direct access to the track for further exploration.

Source

Spotify Data: Popular Hip-hop Artists and Tracks

In [2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [4]:
```
df2=pd.read_csv("./spotify.csv")
df2
```

Out[4]:

	Artist	Track Name	Popularity	Duration (ms)	Track ID
0	Drake	Rich Baby Daddy (feat. Sexyy Red & SZA)	92	319191	1yeB8MUNeLo9Ek1UEpsyz6
1	Drake	One Dance	91	173986	1zi7xx7UVEFkmKfv06H8x0
2	Drake	IDGAF (feat. Yeat)	90	260111	2YSzYUF3jWqb9YP9VXmpjE
3	Drake	First Person Shooter (feat. J. Cole)	88	247444	7aqfrAY2p9BUSiupwk3svU
4	Drake	Jimmy Cooks (feat. 21 Savage)	88	218364	3F5CgOj3wFlRv51JsHbxhe
...	...	...	...	...	...
435	French Montana	Splash Brothers	44	221863	3fBsEOnzwtlkpS0LxXAZhN
436	Fat Joe	All The Way Up (feat. Infared)	64	191900	7EzwtgfW7khBrpvaNPtMoT
437	A$AP Ferg	Work REMIX (feat. A$AP Rocky, French Montana, ...	69	283693	7xVLFuuYdAvcTfcP3IG3dS
438	Diddy	Another One Of Me (feat. 21 Savage)	65	220408	4hGmQboiou09EwhcTWa0H6
439	Rick Ross	Stay Schemin	68	267720	0nq6sfr8z1R5KJ4XUk396e

440 rows × 5 columns

In [6]:
```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Artist         440 non-null    object
 1   Track Name     440 non-null    object
 2   Popularity     440 non-null    int64 
 3   Duration (ms)  440 non-null    int64 
 4   Track ID       440 non-null    object
dtypes: int64(2), object(3)
memory usage: 17.3+ KB
```

In [8]:
```
df2.head()
```

Out[8]:

	Artist	Track Name	Popularity	Duration (ms)	Track ID
0	Drake	Rich Baby Daddy (feat. Sexyy Red & SZA)	92	319191	1yeB8MUNeLo9Ek1UEpsyz6
1	Drake	One Dance	91	173986	1zi7xx7UVEFkmKfv06H8x0
2	Drake	IDGAF (feat. Yeat)	90	260111	2YSzYUF3jWqb9YP9VXmpjE
3	Drake	First Person Shooter (feat. J. Cole)	88	247444	7aqfrAY2p9BUSiupwk3svU
4	Drake	Jimmy Cooks (feat. 21 Savage)	88	218364	3F5CgOj3wFlRv51JsHbxhe

1.Identify the top 5 popular artists based on the mean popularity of their tracks.Show the mean popularity of tracks for the top 5 popular artists Using BarPlot.

In [12]:
```
df2.columns
```

Out[12]:
```
Index(['Artist', 'Track Name', 'Popularity', 'Duration (ms)', 'Track ID'],
dtype='object')
```

In [14]:
```
df2["Artist"].value_counts()
```

Out[14]:
```
Artist
Drake             20
Travis Scott      12
21 Savage         11
¥$                11
Lil Nas X         11
                  ..
Arizona Zervas     1
Fivio Foreign      1
Pressa             1
David Guetta       1
Diddy              1
Name: count, Length: 115, dtype: int64
```

In [32]:
```
df_1=df2[["Artist","Popularity"]].groupby("Artist").mean().round(2)
Top_Popular_Artist=df_1.reset_index().sort_values(by="Popularity",ascending=
Top_Popular_Artist
```

Out[32]:

	Artist	Popularity
113	cassö	92.00
104	Trueno	89.00
24	David Guetta	87.00
103	Travis Scott	87.00
114	¥$	86.09
...	...	...
89	RAYE	55.00
107	Wyclef Jean	54.50
7	Arizona Zervas	54.00
52	Justin Bieber	49.00
85	Pressa	29.00

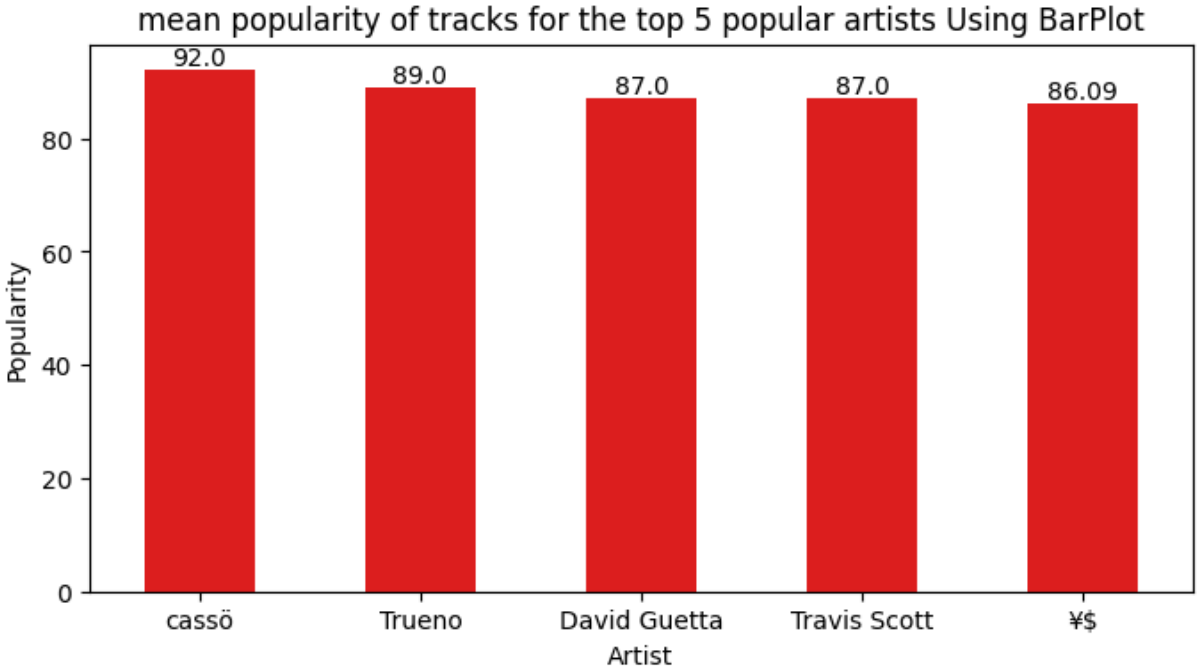
115 rows × 2 columns

In [33]:
```
top_5_Popular_Artist=Top_Popular_Artist.head(5)
top_5_Popular_Artist
```

Out[33]:

	Artist	Popularity
113	cassö	92.00
104	Trueno	89.00
24	David Guetta	87.00
103	Travis Scott	87.00
114	¥$	86.09

In [42]:
```
#Plot a bar
plt.figure(figsize=(8,4))
sns.barplot(x=top_5_Popular_Artist["Artist"],y=top_5_Popular_Artist["Popular
for i, count in enumerate(top_5_Popular_Artist["Popularity"]):
  plt.text(i,count,str(count),ha="center",va="bottom")
plt.title("mean popularity of tracks for the top 5 popular artists Using Bar
plt.show()
```



2. Determine the top 5 popular songs based on their popularity ratings.Display the popularity ratings of the top 5 popular songs using BarPlot.

In [44]:
```
df2.columns
```

Out[44]:
```
Index(['Artist', 'Track Name', 'Popularity', 'Duration (ms)', 'Track ID'],
dtype='object')
```

In [47]:
```
df_2=df2[["Track Name","Popularity"]]
df_2
```

Out[47]:

	Track Name	Popularity
0	Rich Baby Daddy (feat. Sexyy Red & SZA)	92
1	One Dance	91
2	IDGAF (feat. Yeat)	90
3	First Person Shooter (feat. J. Cole)	88
4	Jimmy Cooks (feat. 21 Savage)	88
...	...	...
435	Splash Brothers	44
436	All The Way Up (feat. Infared)	64
437	Work REMIX (feat. A$AP Rocky, French Montana, ...	69
438	Another One Of Me (feat. 21 Savage)	65
439	Stay Schemin	68

440 rows × 2 columns

In [56]:
```
Top_popular_songs=df_2.sort_values(by="Popularity",ascending=False).drop_dup
Top_popular_songs
```

Out[56]:

	Track Name	Popularity
40	Lovin On Me	97
280	CARNIVAL	96
70	redrum	96
30	FE!N (feat. Playboi Carti)	93
0	Rich Baby Daddy (feat. Sexyy Red & SZA)	92
...	...	...
407	911 (feat. Mary J. Blige)	48
225	On Me - Remix	47
413	Splash Brothers	44
231	Intentions	35
207	Attachments (feat. Coi Leray)	29

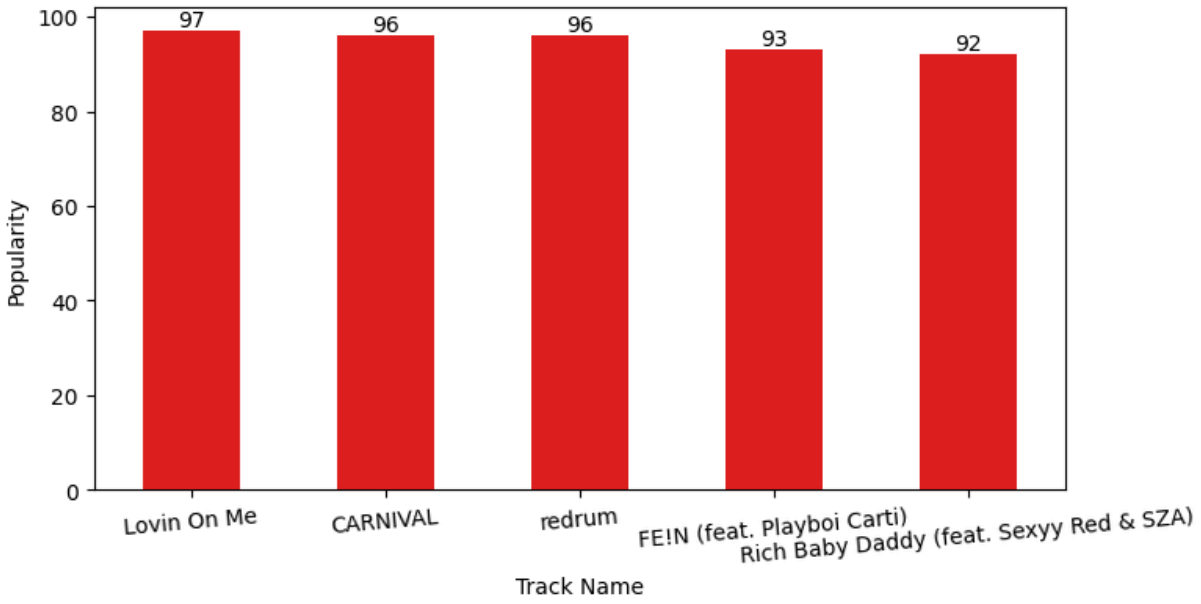
413 rows × 2 columns

In [57]:
```
Top_5_popular_songs=Top_popular_songs.head(5)
Top_5_popular_songs
```

Out[57]:

	Track Name	Popularity
40	Lovin On Me	97
280	CARNIVAL	96
70	redrum	96
30	FE!N (feat. Playboi Carti)	93
0	Rich Baby Daddy (feat. Sexyy Red & SZA)	92

In [69]:
```
plt.figure(figsize=(8,4))
sns.barplot(x=Top_5_popular_songs["Track Name"],y=Top_5_popular_songs["Popul
plt.xticks(rotation=5)
for i,count in enumerate(Top_5_popular_songs["Popularity"]):
  plt.text(i,count,str(count),va="bottom",ha="center")
plt.show()
```



3.Find the top 5 trending genres based on the mean popularity of tracks within each genre.Visualize the mean popularity of tracks for the top 5 trending genres.

In [73]:
```
mean_popularity_by_artist = df2.groupby('Artist')['Popularity'].mean().reset
mean_popularity_by_artist
```

Out[73]:

	Artist	Popularity
0	*NSYNC	67.000000
1	2 Chainz	72.000000
2	21 Savage	84.181818
3	A Boogie Wit da Hoodie	80.000000
4	A$AP Ferg	69.000000
...	...	...
110	Young Nudy	67.000000
111	Young Thug	73.750000
112	benny blanco	72.000000
113	cassö	92.000000
114	¥$	86.090909

115 rows × 2 columns

In [116…

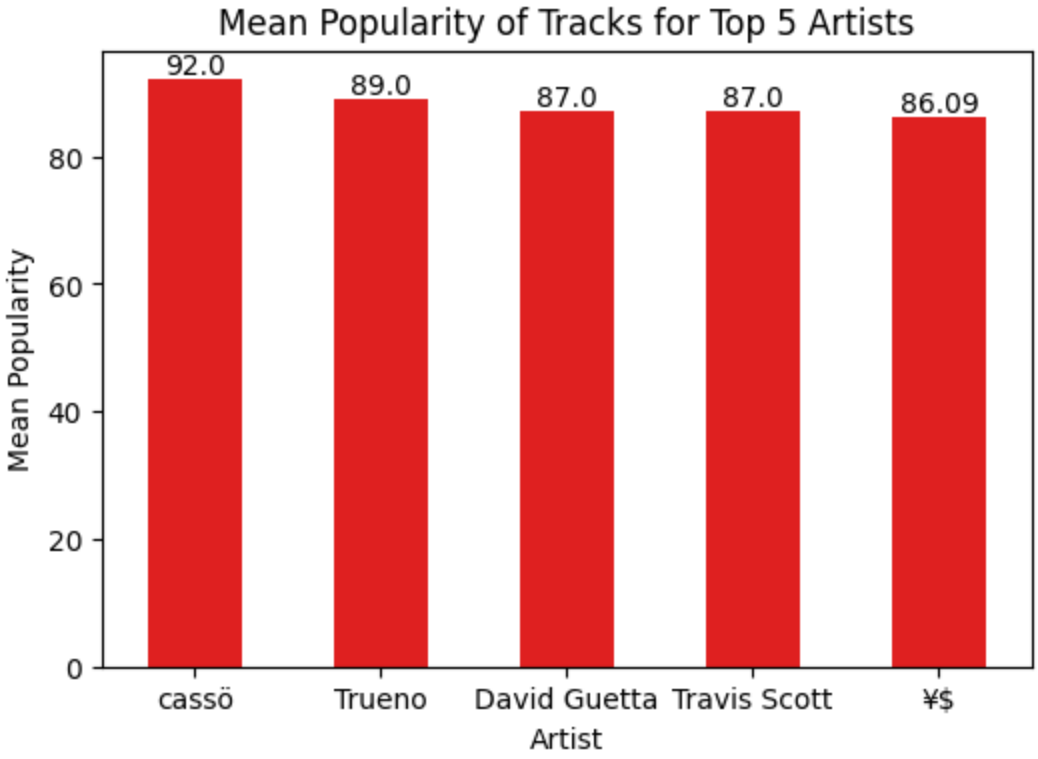
```
top_5_artists = mean_popularity_by_artist.nlargest(5, 'Popularity').round(2)
top_5_artists
```

Out[116…

	Artist	Popularity
113	cassö	92.00
104	Trueno	89.00
24	David Guetta	87.00
103	Travis Scott	87.00
114	¥$	86.09

In [118…

```
#Visualize the mean popularity of tracks for the top 5 artists
plt.figure(figsize=(6, 4))
sns.barplot(x='Artist', y='Popularity', data=top_5_artists, color='r',width=
plt.xlabel('Artist')
plt.ylabel('Mean Popularity')
plt.title('Mean Popularity of Tracks for Top 5 Artists')
for i,count in enumerate(top_5_artists["Popularity"]):
  plt.text(i,count,str(count),ha="center",va="bottom")
plt.show()
```



4.Identify the top 5 longest songs among the tracks of the top 5 popular artists.Represent the duration of the top 5 longest songs among the tracks of the top 5 popular artists using BarPlot.

In [81]:
```
df2.columns
```

Out[81]:
```
Index(['Artist', 'Track Name', 'Popularity', 'Duration (ms)', 'Track ID'],
dtype='object')
```

In [100…
```
mean_popularity_by_artist = df2.groupby('Artist')['Popularity'].mean().reset
mean_popularity_by_artist
```

Out[100…

	Artist	Popularity
0	*NSYNC	67.000000
1	2 Chainz	72.000000
2	21 Savage	84.181818
3	A Boogie Wit da Hoodie	80.000000
4	A$AP Ferg	69.000000
...	...	...
110	Young Nudy	67.000000
111	Young Thug	73.750000
112	benny blanco	72.000000
113	cassö	92.000000
114	¥$	86.090909

115 rows × 2 columns

In [110…

```
top_5_artists = mean_popularity_by_artist.nlargest(5, 'Popularity')
top_5_artist_tracks =df2[df2["Artist"].isin(top_5_artists['Artist'])].drop_c
top_5_artist_tracks
```

Out[110…

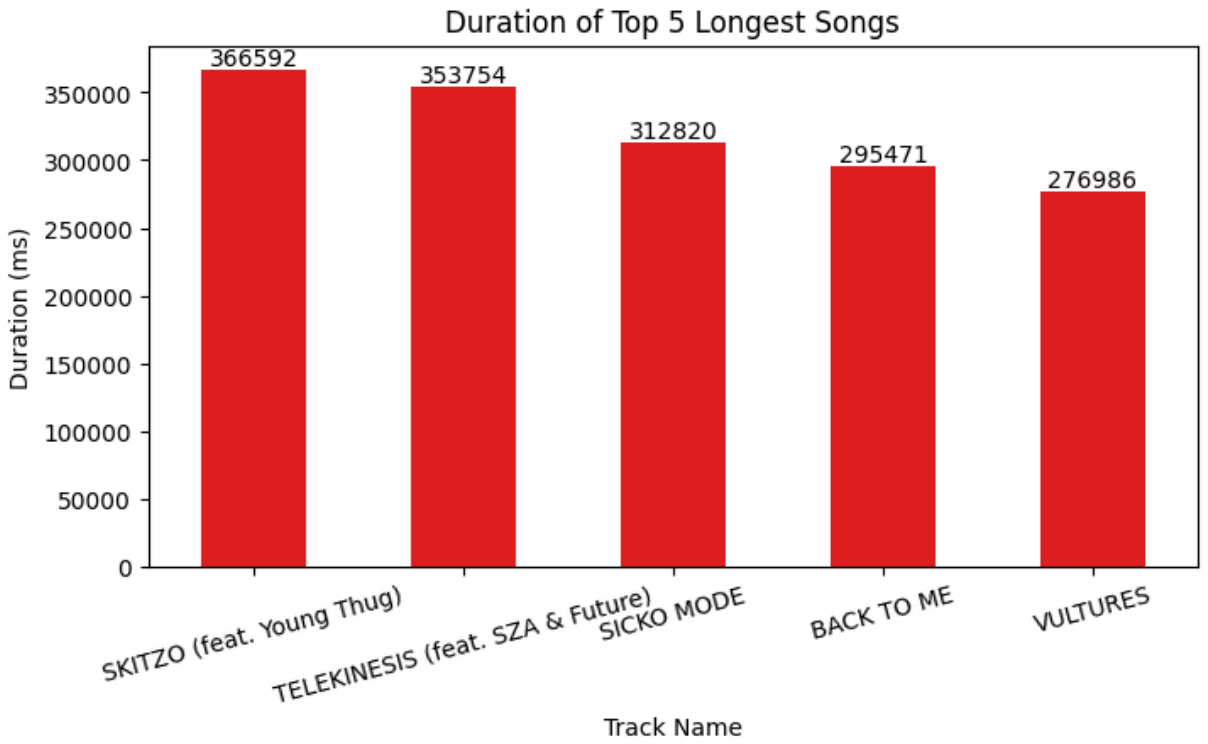
	Artist	Track Name	Popularity	Duration (ms)	Track ID
7	Travis Scott	MELTDOWN (feat. Drake)	86	246133	67nepsnrcZkowTxMWigSbb
30	Travis Scott	FE!N (feat. Playboi Carti)	93	191700	42VsgItocQwOQC3XWZ8JNA
31	Travis Scott	I KNOW ?	92	211582	6wsqVwoiVH2kde4k4KKAFU
32	Travis Scott	MY EYES	91	251249	4kjI1gwQZRKNDkw1nI475M
33	Travis Scott	goosebumps	89	243836	6gBFPUFcJLzWGx4lenP6h2
37	Travis Scott	SICKO MODE	87	312820	2xLMifQCjDGFmkHkpNLD9h
38	Travis Scott	TELEKINESIS (feat. SZA & Future)	86	353754	1i9lZvlaDdWDPyXEE95aiq
140	cassö	Prada	92	132359	59NraMJsLaMCVtwXTSia8i
173	Travis Scott	SKITZO (feat. Young Thug)	78	366592	0bkV1iQHSxBaksUqgEkcbc
200	David Guetta	Baby Don't Hurt Me	87	140017	3BKD1PwArikchz2Zrlp1qi
215	Travis Scott	CIRCUS MAXIMUS (feat. The Weeknd & Swae Lee)	77	258842	4GL9GMX9t7Qkprvf1YighZ
241	Trueno	Mamichula - con Nicki Nicole	89	219201	0TUW9faHNaBmi89wsYGp9y
260	¥$	CARNIVAL	96	264324	3w0w2T288dec0mgeZZqoNN
261	¥$	BURN	89	111458	04CyMEHliadfQWMUJb1w99
262	¥$	FUK SUMN	88	209577	5tEaVciE2GnR28aN6W9cLS
263	¥$	BACK TO ME	86	295471	1icgLGTpX2fQXKRe4D7w2b
264	¥$	STARS	84	115238	347AQK5Lyhn6RvB8tBGYxt
265	¥$	DO IT	83	225000	2iGvsJuc2mC4mDVOVMNAP6
266	¥$	TALKING	81	185110	1eaqMiiUn2P7MnqJK4XeK0
267	¥$	PAID	82	195117	2y4ZR0BUAVePljHSsZyIgj
268	¥$	PAPERWORK	82	145785	2yyO7EKRr7c3txi4xCXUFk
269	¥$	VULTURES	80	276986	3SIRBp4RRQ2AO5H4NO7xfq

In [112…
```
#Sorting the tracks by duration and selecting the top 5 longest songs
top_5_longest_songs = top_5_artist_tracks.nlargest(5, 'Duration (ms)')
top_5_longest_songs
```

Out[112…

	Artist	Track Name	Popularity	Duration (ms)	Track ID
173	Travis Scott	SKITZO (feat. Young Thug)	78	366592	0bkV1iQHSxBaksUqgEkcbc
38	Travis Scott	TELEKINESIS (feat. SZA & Future)	86	353754	1i9lZvlaDdWDPyXEE95aiq
37	Travis Scott	SICKO MODE	87	312820	2xLMifQCjDGFmkHkpNLD9h
263	¥$	BACK TO ME	86	295471	1icgLGTpX2fQXKRe4D7w2b
269	¥$	VULTURES	80	276986	3SIRBp4RRQ2AO5H4NO7xfq

In [120…
```
#Visualizing the duration of the top 5 longest songs using a bar plot
plt.figure(figsize=(8, 4))
sns.barplot(x='Track Name', y='Duration (ms)', data=top_5_longest_songs, col
plt.title('Duration of Top 5 Longest Songs')
plt.xticks(rotation=15)
for i,count in enumerate(top_5_longest_songs["Duration (ms)"]):
  plt.text(i,count,str(count),ha="center",va="bottom")
plt.show()
```



5.Determine the top 5 most danceable songs based on their danceability scores.Illustrate the danceability

scores of the top 5 most danceable songs using PieChart.

In [124…
```
df_5=pd.read_csv("/content/top2018.csv")
```

In [127…
```
df_5.head(3)
```

Out[127…

	id	name	artists	danceability	energy	key
0	6DCZcSspjsKoFjzjrWoCd	God's Plan	Drake	0.754	0.449	7.0
1	3ee8Jmje8o58CHK66QrVC	SAD!	XXXTENTACION	0.740	0.613	8.0
2	0e7ipj03S05BNilyu5bRz	rockstar (feat. 21 Savage)	Post Malone	0.587	0.535	5.0

In [126…
```
df_5.columns
```

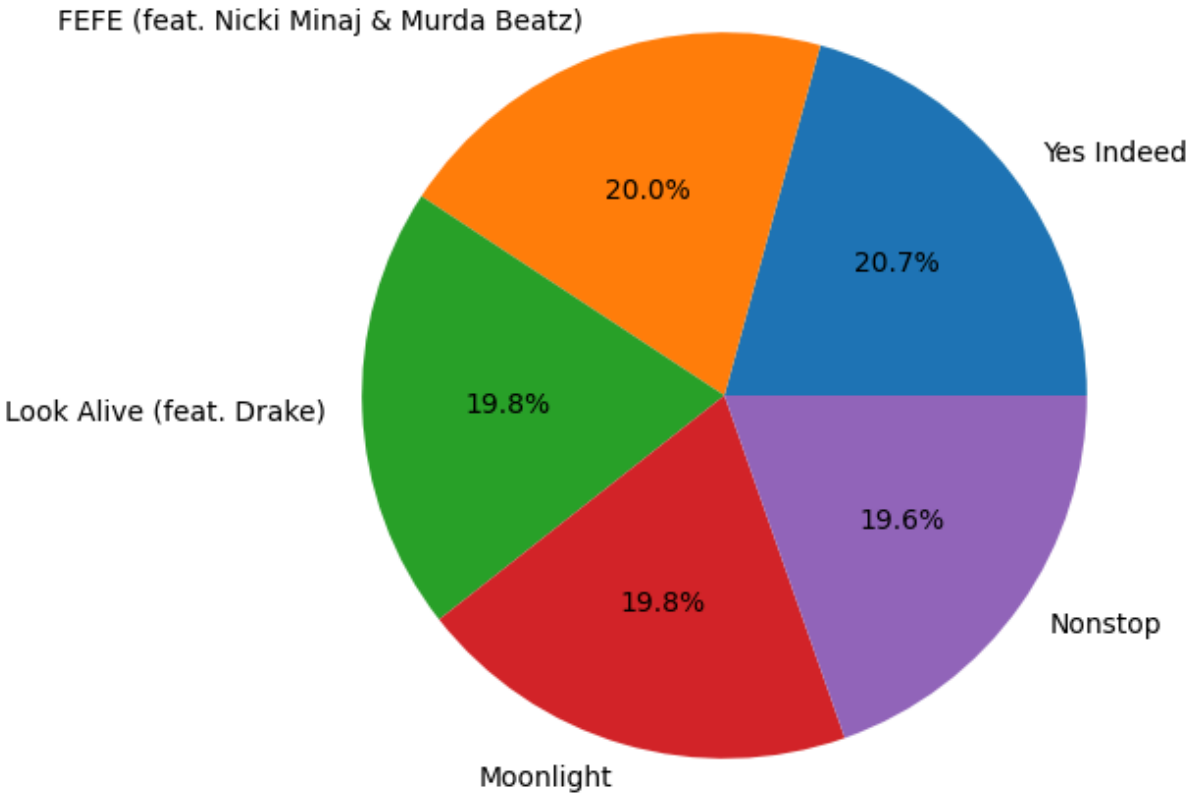
Out[126…
```
Index(['id', 'name', 'artists', 'danceability', 'energy', 'key', 'loudnes
s',
       'mode', 'speechiness', 'acousticness', 'instrumentalness', 'livenes
s',
       'valence', 'tempo', 'duration_ms', 'time_signature'],
      dtype='object')
```

In [133…
```
#top_danceable_song=df_5[["name","danceability"]].sort_values(by="danceabili
top_danceable_song=df_5[["name","danceability"]].nlargest(5,"danceability")
top_5_danceable_song=top_danceable_song.head(5)
top_5_danceable_song
```

Out[133…

	name	danceability
91	Yes Indeed	0.964
55	FEFE (feat. Nicki Minaj & Murda Beatz)	0.931
19	Look Alive (feat. Drake)	0.922
18	Moonlight	0.921
61	Nonstop	0.912

In [141…
```
#Plot piechart
plt.figure(figsize=(6,6))
plt.pie(top_5_danceable_song['danceability'],labels=top_5_danceable_song['na
plt.show()
```



In [ ]: