CodeX is a German beverage company that is aiming to make its mark in the Indian market. A few months ago, they launched their energy drink in 10 cities in India.

Their Marketing team is responsible for increasing brand awareness, market share, and product development. They conducted a survey in those 10 cities and received results from 10k respondents. Peter Pandey, a marketing data analyst is tasked to convert these survey results to meaningful insights which the team can use to drive actions.

In [1]:

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import warnings as warning
```

To import data into dataframe

In [2]:

```python
df_city=pd.read_csv(r"E:\Study\Datas for projects\cold drink\Dataset\dim_cities.csv")
df_respondent=pd.read_csv(r"E:\Study\Datas for projects\cold drink\Dataset\dim_repondents.csv")
df_survey=pd.read_csv(r"E:\Study\Datas for projects\cold drink\Dataset\fact_survey_responses.csv")
```

In [3]:

```python
df_city.head(10) #This will give us top 10 rows
```

Out[3]:

	City_ID	City	Tier
0	CT111	Delhi	Tier 1
1	CT112	Mumbai	Tier 1
2	CT113	Bangalore	Tier 1
3	CT114	Chennai	Tier 1
4	CT115	Kolkata	Tier 2
5	CT116	Hyderabad	Tier 1
6	CT117	Ahmedabad	Tier 2
7	CT118	Pune	Tier 2
8	CT119	Jaipur	Tier 2
9	CT120	Lucknow	Tier 2

In [4]:

```python
df_respondent.head(10) #This will give us top 10 rows
```

Out[4]:

	Respondent_ID	Name	Age	Gender	City_ID
0	120031	Aniruddh Issac	15-18	Female	CT117
1	120032	Trisha Rout	19-30	Male	CT118
2	120033	Yuvraj Virk	15-18	Male	CT116
3	120034	Pranay Chand	31-45	Female	CT113
4	120035	Mohanlal Joshi	19-30	Female	CT120
5	120036	Zeeshan Ratta	19-30	Female	CT118
6	120037	Oorja Anne	19-30	Male	CT112
7	120038	Rhea Khanna	19-30	Male	CT116
8	120039	Zara Joshi	46-65	Male	CT116
9	120040	Sana Dhawan	19-30	Female	CT116

In [5]:

```python
df_survey.head(10) #This will give us top 10 rows
```

Out[5]:

	Response_ID	Respondent_ID	Consume_frequency	Consume_time	Consume_reason	Heard_before	Brand_perception	General_
0	103001	120031	2-3 times a week	To stay awake during work/study	Increased energy and focus	Yes	Neutral	
1	103002	120032	2-3 times a month	Throughout the day	To boost performance	No	Neutral	
2	103003	120033	Rarely	Before exercise	Increased energy and focus	No	Neutral	
3	103004	120034	2-3 times a week	To stay awake during work/study	To boost performance	No	Positive	
4	103005	120035	Daily	To stay awake during work/study	Increased energy and focus	Yes	Neutral	
5	103006	120036	Rarely	For mental alertness	To combat fatigue	Yes	Negative	
6	103007	120037	2-3 times a month	To stay awake during work/study	Increased energy and focus	No	Positive	
7	103008	120038	Rarely	Before exercise	To combat fatigue	No	Neutral	
8	103009	120039	Once a week	To stay awake during work/study	To enhance sports performance	No	Neutral	
9	103010	120040	Once a week	For mental alertness	To combat fatigue	Yes	Neutral	

10 rows × 23 columns

To check shape of data

In [6]:

```
# shape will return no of records and no of features
print('Shape of city dataframe is:',df_city.shape)
print('Shape of respondent dataframe is:',df_respondent.shape)
print('Shape of survey dataframe is:',df_survey.shape)
```

```
Shape of city dataframe is: (10, 3)
Shape of respondent dataframe is: (10000, 5)
Shape of survey dataframe is: (10000, 23)
```

Sanity check of data

In [7]:

```
df_city.info() #We observed and insured that correct data types are mentioned
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   City_ID  10 non-null     object
 1   City     10 non-null     object
 2   Tier     10 non-null     object
dtypes: object(3)
memory usage: 372.0+ bytes
```

In [8]:

```
df_respondent.info() #We observed and insured that correct data types are mentioned
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Respondent_ID  10000 non-null  int64 
 1   Name           10000 non-null  object
 2   Age            10000 non-null  object
 3   Gender         10000 non-null  object
 4   City_ID        10000 non-null  object
dtypes: int64(1), object(4)
memory usage: 390.8+ KB
```

In [9]:

```
df_survey.info() #We observed and insured that correct data types are mentioned
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 23 columns):
 #   Column                          Non-Null Count  Dtype 
---  ------                          --------------  ----- 
 0   Response_ID                     10000 non-null  int64 
 1   Respondent_ID                   10000 non-null  int64 
 2   Consume_frequency               10000 non-null  object
 3   Consume_time                    10000 non-null  object
 4   Consume_reason                  10000 non-null  object
 5   Heard_before                    10000 non-null  object
 6   Brand_perception                10000 non-null  object
 7   General_perception              10000 non-null  object
 8   Tried_before                    10000 non-null  object
 9   Taste_experience                10000 non-null  int64 
 10  Reasons_preventing_trying       10000 non-null  object
 11  Current_brands                  10000 non-null  object
 12  Reasons_for_choosing_brands     10000 non-null  object
 13  Improvements_desired            10000 non-null  object
 14  Ingredients_expected            10000 non-null  object
 15  Health_concerns                 10000 non-null  object
 16  Interest_in_natural_or_organic  10000 non-null  object
 17  Marketing_channels              10000 non-null  object
 18  Packaging_preference            10000 non-null  object
 19  Limited_edition_packaging       10000 non-null  object
 20  Price_range                     10000 non-null  object
 21  Purchase_location               10000 non-null  object
 22  Typical_consumption_situations  10000 non-null  object
dtypes: int64(3), object(20)
memory usage: 1.8+ MB
```

```
In [10]: df_city.isnull().sum() #We have no null values in our data
```

```
Out[10]: City_ID    0
         City       0
         Tier       0
         dtype: int64
```

```
In [11]: df_respondent.isnull().sum() #We have no null values in our data
```

```
Out[11]: Respondent_ID    0
         Name             0
         Age              0
         Gender           0
         City_ID          0
         dtype: int64
```

```
In [12]: df_survey.isnull().sum() #We have no null values in our data
```

```
Out[12]: Response_ID                       0
         Respondent_ID                     0
         Consume_frequency                 0
         Consume_time                      0
         Consume_reason                    0
         Heard_before                      0
         Brand_perception                  0
         General_perception                0
         Tried_before                      0
         Taste_experience                  0
         Reasons_preventing_trying         0
         Current_brands                    0
         Reasons_for_choosing_brands       0
         Improvements_desired              0
         Ingredients_expected              0
         Health_concerns                   0
         Interest_in_natural_or_organic    0
         Marketing_channels                0
         Packaging_preference              0
         Limited_edition_packaging         0
         Price_range                       0
         Purchase_location                 0
         Typical_consumption_situations    0
         dtype: int64
```

```
In [13]: print('Duplicate values in our city dataframe is:',df_city.duplicated().sum())
         print('Duplicate values in our respondent dataframe is:',df_respondent.duplicated().sum())
         print('Duplicate values in our survey dataframe is:',df_survey.duplicated().sum())
```

```
Duplicate values in our city dataframe is: 0
Duplicate values in our respondent dataframe is: 0
Duplicate values in our survey dataframe is: 0
```

Stastical summary

In [14]:
```
df_city.describe()
```

Out[14]:

	City_ID	City	Tier
count	10	10	10
unique	10	10	2
top	CT111	Delhi	Tier 1
freq	1	1	5

In [15]:
```
df_respondent.describe().T
```

Out[15]:

	count	mean	std	min	25%	50%	75%	max
Respondent_ID	10000.0	125030.5	2886.89568	120031.0	122530.75	125030.5	127530.25	130030.0

In [16]:
```
df_survey.describe().T
```

Out[16]:

	count	mean	std	min	25%	50%	75%	max
Response_ID	10000.0	108000.5000	2886.895680	103001.0	105500.75	108000.5	110500.25	113000.0
Respondent_ID	10000.0	125030.5000	2886.895680	120031.0	122530.75	125030.5	127530.25	130030.0
Taste_experience	10000.0	3.2819	1.239752	1.0	2.00	3.0	4.00	5.0

In [17]:
```
df_survey
```

Out[17]:

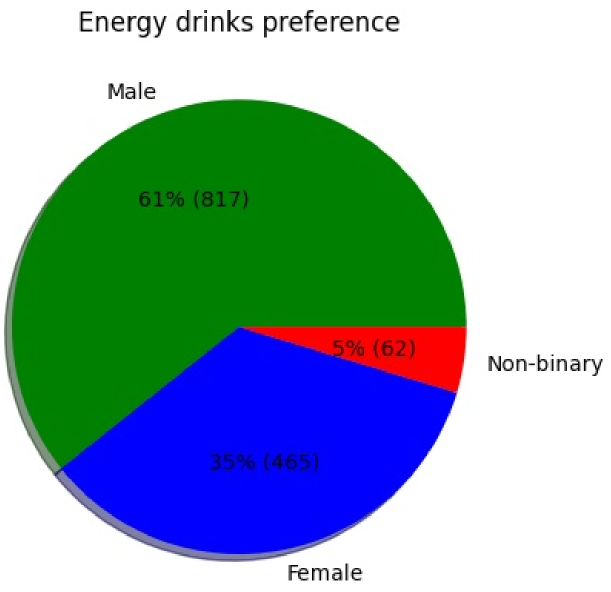
	Response_ID	Respondent_ID	Consume_frequency	Consume_time	Consume_reason	Heard_before	Brand_perception	Gene
0	103001	120031	2-3 times a week	To stay awake during work/study	Increased energy and focus	Yes	Neutral	
1	103002	120032	2-3 times a month	Throughout the day	To boost performance	No	Neutral	
2	103003	120033	Rarely	Before exercise	Increased energy and focus	No	Neutral	
3	103004	120034	2-3 times a week	To stay awake during work/study	To boost performance	No	Positive	
4	103005	120035	Daily	To stay awake during work/study	Increased energy and focus	Yes	Neutral	
...	...	...	...	...	...	...	...	
9995	112996	130026	Daily	Before exercise	To enhance sports performance	Yes	Neutral	
9996	112997	130027	Daily	To stay awake during work/study	To combat fatigue	No	Positive	
9997	112998	130028	Daily	Before exercise	Increased energy and focus	Yes	Positive	
9998	112999	130029	2-3 times a week	To stay awake during work/study	Increased energy and focus	No	Positive	
9999	113000	130030	Daily	For mental alertness	Other	Yes	Positive	

10000 rows × 23 columns

1. Demographic Insights

a. Who prefers energy drink more?

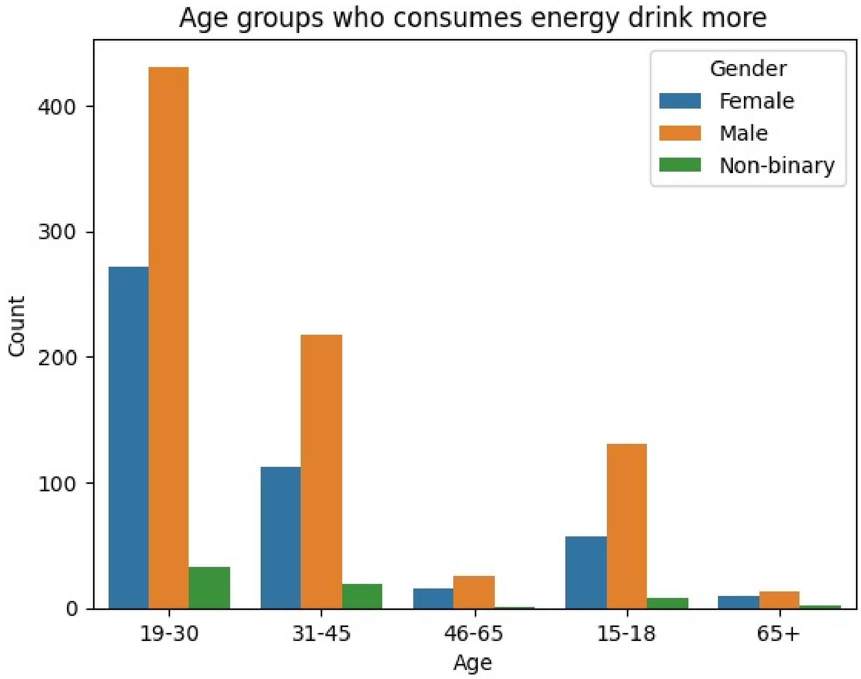
In [140…
```
daily=df_survey.groupby('Consume_frequency').get_group('Daily') #This peoples are consuming drinks daily
df1=pd.merge(daily,df_respondent,on='Respondent_ID',how='inner') #We merged to find out gender
gender_counts = df1['Gender'].value_counts()
plt.title("Energy drinks preference")
c = ["g", "b", "r"]  # Colors for the pie chart
plt.pie(gender_counts, labels=gender_counts.index, colors=c, autopct=lambda pct: f"{pct:.0f}% ({int(pct/100*gen
plt.show()
#Observation:-According to data 'Male' consumes more energy drink which is 61% followed by Female (35%)
```



b. Which age group prefers energy drinks more?

In [19]:

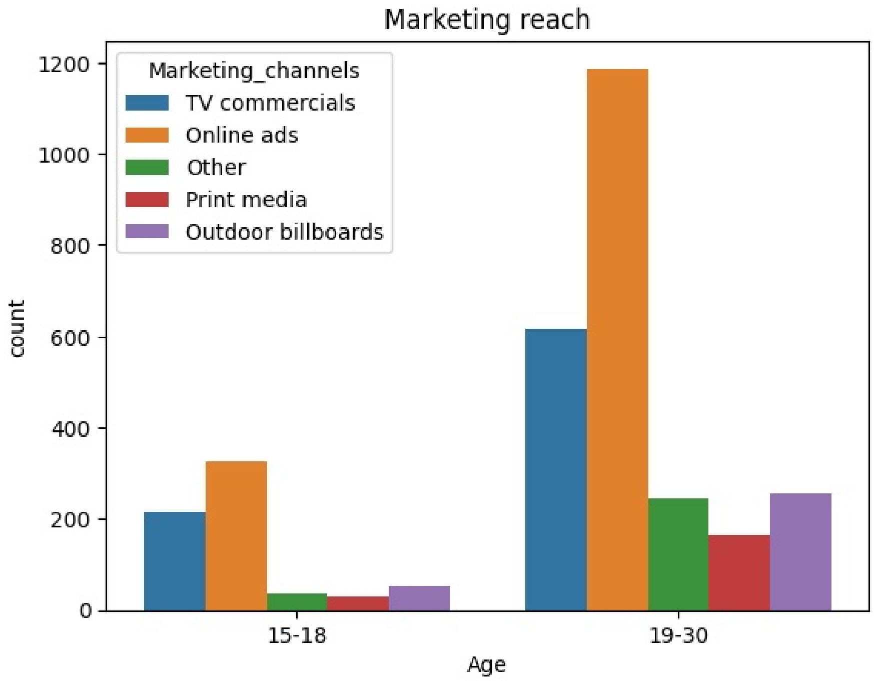
```
daily=df_survey.groupby('Consume_frequency').get_group('Daily') # Groupby is used to group data
age=df_respondent[['Respondent_ID','Age','Gender']]
df1=pd.merge(age,daily,how='inner')
sns.countplot(x='Age', hue='Gender', data=df1)
plt.title('Age groups who consumes energy drink more')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
# Observation:-Age group of 19-30 consumes more drinks additionally Males from every group have tendecny to con
```



c. Which type of marketing reaches the most Youth (15-30)?

In [141…

```
marketing=df_survey.groupby('Heard_before').get_group('Yes')
age1=df_respondent[(df_respondent['Age']=='15-18') | (df_respondent['Age']=='19-30')]
df1=pd.merge(marketing,age1,how='inner')
df1
sns.countplot(x='Age',hue='Marketing_channels',data=df1)
plt.title('Marketing reach')
plt.show()
# Observation:-As data shows Marketing teams reach is high through online ads and Tv commercials comes after it
```

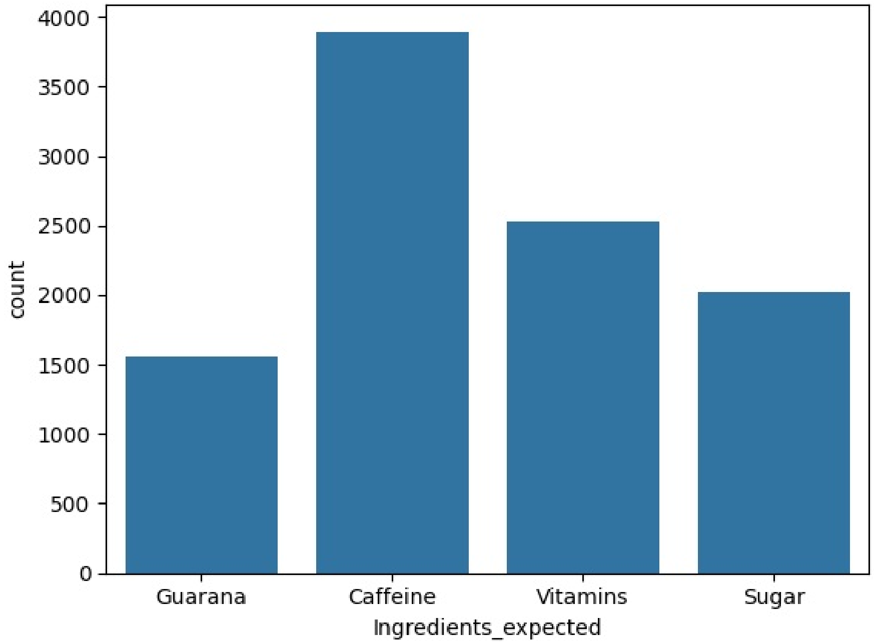


2. Consumer Preferences

a. What are the preferred ingredients of energy drinks among respondents?

In [32]:

```
df=pd.merge(df_survey,df_respondent,on='Respondent_ID')
sns.countplot(x='Ingredients_expected',data=df)
plt.show()
# Observation:-Customers demands for caffeine as most preferred ingredient.
```



b. What packaging preferences do respondents have for energy drinks?

In [46]:

```
count=df_survey['Packaging_preference'].value_counts()
print(count)
print('**'*20)
percent=df_survey['Packaging_preference'].value_counts()/df_survey.shape[0]*100
print(percent)

# Observation:- Approximately 40% of customer will prefer to have packaging as 'Compact and portable cans' and
```

```
Packaging_preference
Compact and portable cans    3984
Innovative bottle design     3047
Collectible packaging        1501
Eco-friendly design           983
Other                         485
Name: count, dtype: int64
****************************************
Packaging_preference
Compact and portable cans    39.84
Innovative bottle design     30.47
Collectible packaging        15.01
Eco-friendly design           9.83
Other                         4.85
Name: count, dtype: float64
```

3. Competition Analysis:

a. Who are the current market leaders?

In [63]:

```
market_leader=df_survey['Current_brands']
a=market_leader.value_counts()
b=market_leader.value_counts()/df_survey.shape[0]*100
print('Current market leaders:',a)
print("**"*15)
print('Current Market Leaders Share is:',b)
# Observation:- Cola-coka has maximum share of market which is 25%,Bepsi has 21% of total market share our compa
```

```
Current market leaders: Current_brands
Cola-Coka    2538
Bepsi        2112
Gangster     1854
Blue Bull    1058
CodeX         980
Sky 9         979
Others        479
Name: count, dtype: int64
******************************
Current Market Leaders Share is: Current_brands
Cola-Coka    25.38
Bepsi        21.12
Gangster     18.54
Blue Bull    10.58
CodeX         9.80
Sky 9         9.79
Others        4.79
Name: count, dtype: float64
```

b. What are the primary reasons consumers prefer those brands over ours?

In [144…

```
brand=df_survey['Reasons_for_choosing_brands']
a=brand.value_counts()
b=brand.value_counts()/df_survey.shape[0]*100
print('Current market leaders:',a)
print("**"*15)
print('Current Market Leaders Share is:',b)
# Observation:- More than 25% person choose brand beracuse of brand reputation
```

```
Current market leaders: Reasons_for_choosing_brands
Brand reputation           2652
Taste/flavor preference    2011
Availability               1910
Effectiveness              1748
Other                      1679
Name: count, dtype: int64
******************************
Current Market Leaders Share is: Reasons_for_choosing_brands
Brand reputation           26.52
Taste/flavor preference    20.11
Availability               19.10
Effectiveness              17.48
Other                      16.79
Name: count, dtype: float64
```

4. Marketing Channels and Brand Awareness:

a. Which marketing channel can be used to reach more customers?

In [65]:

```
channel=df_survey['Marketing_channels']
a=channel.value_counts()
print('Most used marketing channel is:',a)
# Observation:-Print media is list performing marketing tool as it has reached to only 8% of people.
```

```
Most used marketing channel is: Marketing_channels
Online ads            4020
TV commercials        2688
Outdoor billboards    1226
Other                 1225
Print media            841
Name: count, dtype: int64
```

5. Brand Penetration:

a. What do people think about our brand? (overall rating)

In [ ]:

```
brand=df_survey['Reasons_for_choosing_brands']
a=brand.value_counts()
b=brand.value_counts()/df_survey.shape[0]*100
print('Current market leaders:',a)
print("**"*15)
print('Current Market Leaders Share is:',b)
# Observation:- More than 25% person choose brand beracuse of brand reputation
```

b. Which cities do we need to focus more on?

In [99]:

```
df1=pd.merge(df_city,df_respondent)
count=df1['City'].value_counts().sort_values(ascending=True)
print('From below city we have very less penetration in bottom 3 city which are',count[0:3])
print('**'*15)
print(count)
# Observation:- We need to penetrate ('Lucknow','Jaipur','Delhi') as we have list preserence in this city
```

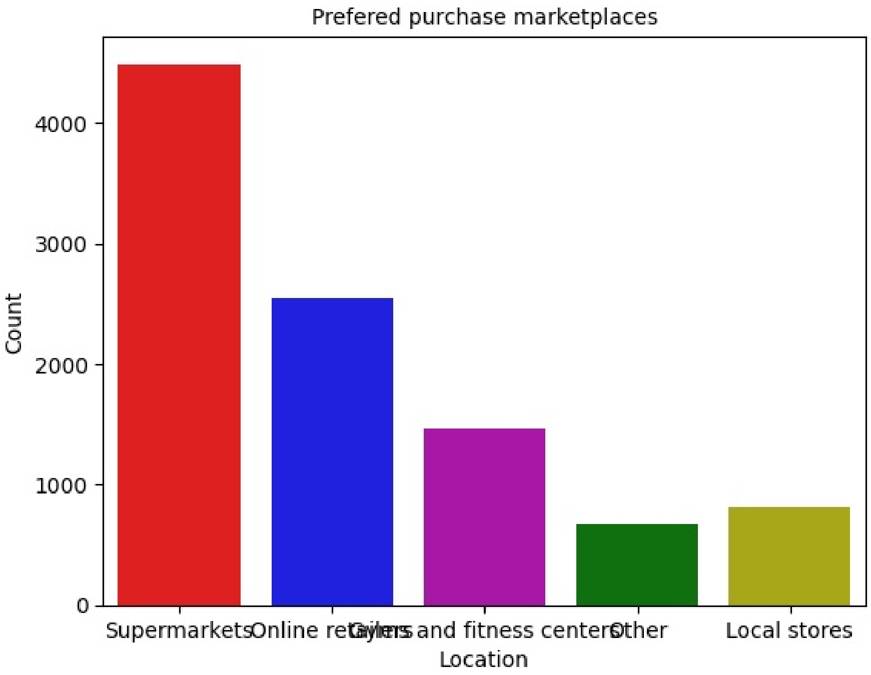
```
From below city we have very less penetration in bottom 3 city which are City
Lucknow    175
Jaipur     360
Delhi      429
Name: count, dtype: int64
******************************
City
Lucknow       175
Jaipur        360
Delhi         429
Ahmedabad     456
Kolkata       566
Pune          906
Chennai       937
Mumbai       1510
Hyderabad    1833
Bangalore    2828
Name: count, dtype: int64
```

6. Purchase Behavior:

a. Where do respondents prefer to purchase energy drinks?

In [113…

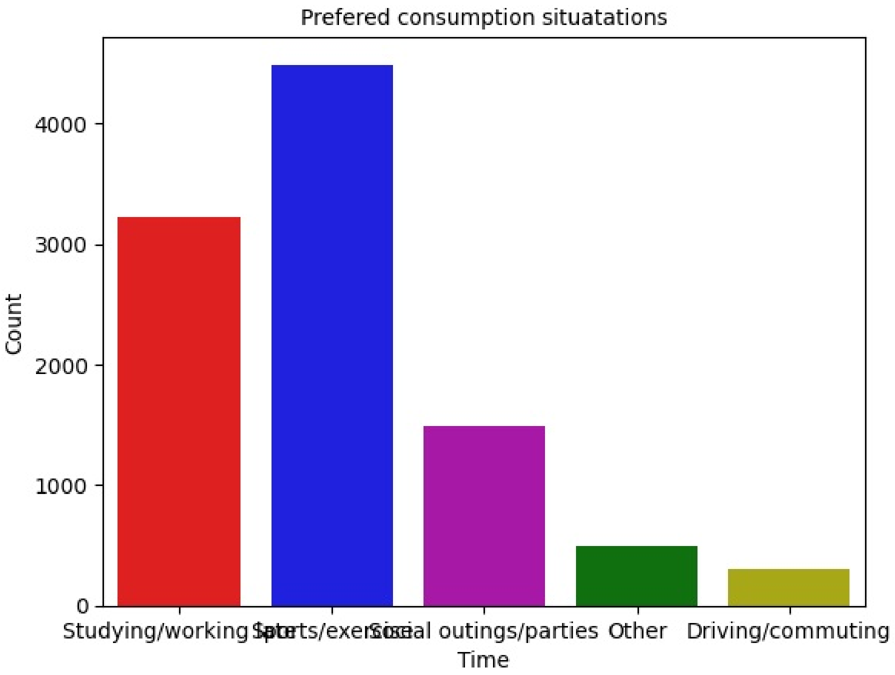
```
df=df_survey['Purchase_location']
df.value_counts()
sns.countplot(x=df)
plt.xlabel("Location",fontsize = 10)
plt.ylabel("Count",fontsize = 10)
plt.title("Prefered purchase marketplaces",fontsize = 10)
sns.countplot(x=df, palette=["r", "b", "m", "g","y"])
plt.show()
# Observation:- Most of the sales done by supermarket
```



b. What are the typical consumption situations for energy drinks among respondents?

In [121…

```
df=df_survey['Typical_consumption_situations']
df.value_counts()
sns.countplot(x=df)
plt.xlabel("Time",fontsize = 10)
plt.ylabel("Count",fontsize = 10)
plt.title("Prefered consumption situatations",fontsize = 10)
sns.countplot(x=df, palette=["r", "b", "m", "g","y"])
plt.show()
# Obervation:- Most customer prefer consume energy drinks before sports and excercise
```



c. What factors influence respondents' purchase decisions, such as price range and limited edition packaging?

In [137…

```
df=df_survey['Price_range']
a=df.value_counts()
b=df.value_counts()/df_survey.shape[0]*100
print(b)
print('*'*25)
df1=df_survey['Limited_edition_packaging']
c=df1.value_counts()
d=df1.value_counts()/df_survey.shape[0]*100
print(d)
# Observation:-i) Data shows nearly 75% of people thinks price point between 50 to 150.Nearly 11%  wants our bra
# ii) It is very hard to say that 'Limited edition packing ' has any influence on sale because approx 40% said '
```

```
Price_range
50-99        42.88
100-150      31.42
Above 150    15.61
Below 50     10.09
Name: count, dtype: float64
*************************
Limited_edition_packaging
No          40.23
Yes         39.46
Not Sure    20.31
Name: count, dtype: float64
```

Product Development

Which area of business should we focus more on our product development?

Branding- Most of the customer- Based on survey, when we did competition analysis, we got to know that most of the people prefer other brands (Cola-Coka) over ours because of brand reputation.

Availability- Company need to focus on Tier-2 citys also there is large no of chunk who has not yet heard about our Product.

Observations

1) According to data 'Male' consumes more energy drink which is 61% followed by Female (35%).

2) Age group of 19-30 consumes more drinks additionally Males from every group have tendecny to consume more energy drinks.

3) As data shows Marketing teams reach is high through online ads and Tv commercials comes after it.

4) Customers demands for caffeine as most preferred ingredient.

5) Approximately 40% of customer will prefer to have packaging as 'Compact and portable cans' and 30% customer will prefer 'Innovative bottle design.

6) Cola-coka has maximum share of market which is 25%,Bepsi has 21% of total market share our company holds 5th position with share of 9%

7) Print media is list performing marketing tool as it has reached to only 8% of people.

8) Company need to penetrate ('Lucknow','Jaipur','Delhi') as we have list preserence in this city.

9) Most of the sales done by supermarket

10) i) Data shows nearly 75% of people thinks price point between 50 to 150.Nearly 11% wants our brand to be premium category (Above150) ii) It is very hard to say that 'Limited edition packing ' has any influence on sale because approx 40% said Yes and no and 20% of people not sure so it is impossible to conclude without having additional information

```
In [ ]:
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js