In [1]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
```

In [2]:
```
df_kaggle_22 = pd.read_csv('/kaggle/input/kaggle-survey-2022/kaggle_survey_2022_responses.csv')
df_kaggle_22.head()
```

```
/opt/conda/lib/python3.7/site-packages/IPython/core/interactiveshell.py:3552: DtypeWarning: Columns (0,208,225,255,257,260,270,271,277) have mixed types.Specify dtype option
on import or set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

Out[2]:

	Duration (in seconds)	Q2	Q3	Q4	Q5	Q6_1	Q6_2	Q6_3	Q6_4	Q6_5	...	Q44_3	Q44_4	Q44_5	Q44_6	Q44_7	Q44_8	Q44_9	Q44_10	Q44_11
0	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	Are you currently a student? (high school, uni...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...
1	121	30-34	Man	India	No	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	462	30-34	Man	Algeria	No	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	293	18-21	Man	Egypt	Yes	Coursera	edX	NaN	DataCamp	NaN	...	NaN	Kaggle (notebooks, forums, etc)	NaN	YouTube (Kaggle YouTube, Cloud AI Adventures, ...	Podcasts (Chai Time Data Science, O'Reilly Dat...	NaN	NaN	NaN	NaN
4	851	55-59	Man	France	No	Coursera	NaN	Kaggle Learn Courses	NaN	NaN	...	NaN	Kaggle (notebooks, forums, etc)	Course Forums (forums.fast.ai, Coursera forums...	NaN	NaN	Blogs (Towards Data Science, Analytics Vidhya,...	NaN	NaN	NaN

5 rows × 296 columns

In [3]:
```
df_kaggle_22.columns
```

Out[3]:
```
Index(['Duration (in seconds)', 'Q2', 'Q3', 'Q4', 'Q5', 'Q6_1', 'Q6_2', 'Q6_3',
       'Q6_4', 'Q6_5',
       ...
       'Q44_3', 'Q44_4', 'Q44_5', 'Q44_6', 'Q44_7', 'Q44_8', 'Q44_9', 'Q44_10',
       'Q44_11', 'Q44_12'],
      dtype='object', length=296)
```

In [4]:
```
df_kaggle_22.shape
```

Out[4]:
```
(23998, 296)
```

It means total participants are 23998 in kaggle survey for ML/DS 2022

In [5]:
```
df_kaggle_22.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23998 entries, 0 to 23997
Columns: 296 entries, Duration (in seconds) to Q44_12
dtypes: object(296)
memory usage: 54.2+ MB
```

we have 296 columns here form Q1 to Q44_12

In [6]:
```
df_kaggle_22.dtypes
```

Out[6]:
```
Duration (in seconds)    object
Q2                       object
Q3                       object
Q4                       object
Q5                       object
                          ...
Q44_8                    object
Q44_9                    object
Q44_10                   object
Q44_11                   object
Q44_12                   object
Length: 296, dtype: object
```

All columns datatypes are objects

Importing some usefull libraries for data anaysis and data visualization.

In [7]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
```

In [8]:
```
df_kaggle_22.columns
```

Out[8]:
```
Index(['Duration (in seconds)', 'Q2', 'Q3', 'Q4', 'Q5', 'Q6_1', 'Q6_2', 'Q6_3',
       'Q6_4', 'Q6_5',
       ...
       'Q44_3', 'Q44_4', 'Q44_5', 'Q44_6', 'Q44_7', 'Q44_8', 'Q44_9', 'Q44_10',
       'Q44_11', 'Q44_12'],
      dtype='object', length=296)
```

In [9]:
```
df_kaggle_22.rename(columns = {'Duration (in seconds)': 'Duration'}, inplace=True)
```

In [10]:
```
df_kaggle_22
```

Out[10]:

	Duration	Q2	Q3	Q4	Q5	Q6_1	Q6_2	Q6_3	Q6_4	Q6_5	...	Q44_3	Q44_4	Q44_5	Q44_6	Q44_7	Q44_8	Q44_9	Q44_10	Q
0	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	Are you currently a student? (high school, uni...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Who/what are your favorite media sources that ...	Wh
1	121	30-34	Man	India	No	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
2	462	30-34	Man	Algeria	No	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
3	293	18-21	Man	Egypt	Yes	Coursera	edX	NaN	DataCamp	NaN	...	NaN	Kaggle (notebooks, forums, etc)	NaN	YouTube (Kaggle YouTube, Cloud AI Adventures, ...	Podcasts (Chai Time Data Science, O'Reilly Dat...	NaN	NaN	NaN	
4	851	55-59	Man	France	No	Coursera	NaN	Kaggle Learn Courses	NaN	NaN	...	NaN	Kaggle (notebooks, forums, etc)	Course Forums (forums.fast.ai, Coursera forums...	NaN	NaN	Blogs (Towards Data Science, Analytics Vidhya,...	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
23993	331	22-24	Man	United States of America	Yes	NaN	NaN	NaN	NaN	NaN	...	NaN	Kaggle (notebooks, forums, etc)	NaN	YouTube (Kaggle YouTube, Cloud AI Adventures, ...	Podcasts (Chai Time Data Science, O'Reilly Dat...	NaN	Journal Publications (peer-reviewed journals, ...	NaN	
23994	330	60-69	Man	United States of America	Yes	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	YouTube (Kaggle YouTube, Cloud AI Adventures, ...	NaN	NaN	NaN	NaN	
23995	860	25-29	Man	Turkey	No	NaN	NaN	NaN	DataCamp	NaN	...	NaN	Kaggle (notebooks, forums, etc)	NaN	YouTube (Kaggle YouTube, Cloud AI Adventures, ...	NaN	NaN	NaN	NaN	
23996	597	35-39	Woman	Israel	No	NaN	NaN	Kaggle Learn Courses	NaN	NaN	...	NaN	NaN	NaN	YouTube (Kaggle YouTube, Cloud AI Adventures, ...	NaN	NaN	NaN	NaN	
23997	303	18-21	Man	India	Yes	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

23998 rows × 296 columns

In [11]:
```
df_kaggle_22['Duration'].value_counts().sum()
```

Out[11]:
```
23998
```

Q2. What is your age (# years)?

In [12]:
```
df_kaggle_22['Q2'].unique()
```

Out[12]:
```
array(['What is your age (# years)?', '30-34', '18-21', '55-59', '45-49',
       '70+', '22-24', '35-39', '40-44', '50-54', '25-29', '60-69'],
      dtype=object)
```

List of values with different range like 30-35, 18-21 and so on.

we have another method to see the uniques entries

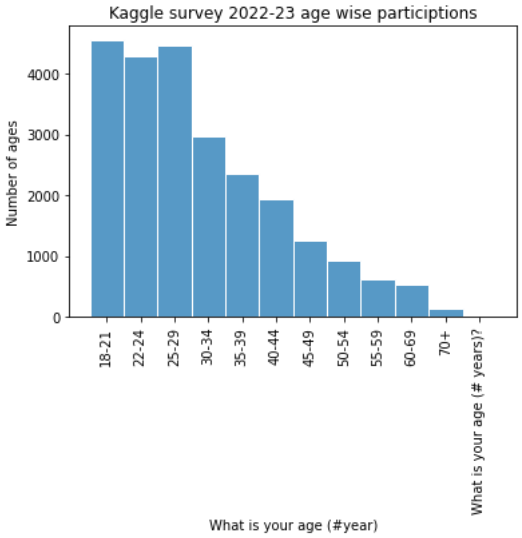
In [13]:
```
age = df_kaggle_22['Q2'].value_counts().sort_values(ascending = False)
age
```

Out[13]:
```
18-21                          4559
25-29                          4472
22-24                          4283
30-34                          2972
35-39                          2353
40-44                          1927
45-49                          1253
50-54                           914
55-59                           611
60-69                           526
70+                             127
What is your age (# years)?       1
Name: Q2, dtype: int64
```

In [14]:
```
age
```

Out[14]:
```
18-21                          4559
25-29                          4472
22-24                          4283
30-34                          2972
35-39                          2353
40-44                          1927
45-49                          1253
50-54                           914
55-59                           611
60-69                           526
70+                             127
What is your age (# years)?       1
Name: Q2, dtype: int64
```

In [15]:
```
fig, ax = plt.subplots()
sns.set_style('dark')
sns.histplot(df_kaggle_22['Q2'].sort_values(ascending=True), ax=ax)
plt.xticks(rotation=90)
ax.set(xlabel= "What is your age (#year)",
       ylabel ="Number of ages",
       title = "Kaggle survey 2022-23 age wise participtions");
```



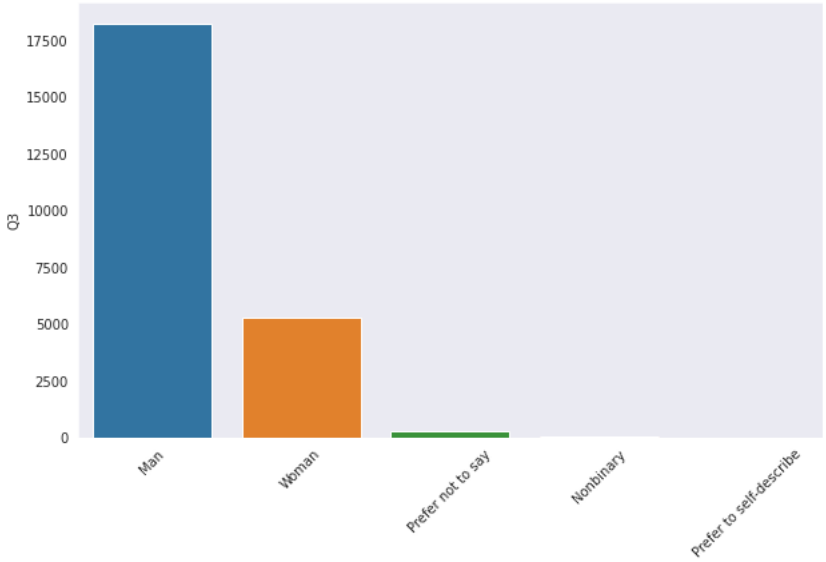
We have more participants are from age 18 to 30 age braket.

What is your gender? - Selected Choice

In [16]:
```
gender = df_kaggle_22['Q3'].value_counts()[:5]
gender
```

Out[16]:
```
Man                       18266
Woman                      5286
Prefer not to say           334
Nonbinary                    78
Prefer to self-describe      33
Name: Q3, dtype: int64
```

In [17]:
```
gender_counts = df_kaggle_22["Q3"].value_counts()[: 5]
plt.figure(figsize=(10,6))
sns.barplot(x = gender_counts.index, y = gender_counts, orient='v');
plt.xticks(rotation=45);
ax.set_ylabel("Count")
ax.set_xlabel("Gender")
ax.set_title("Gender participatent in Kaggle 2022")
plt.show()
```



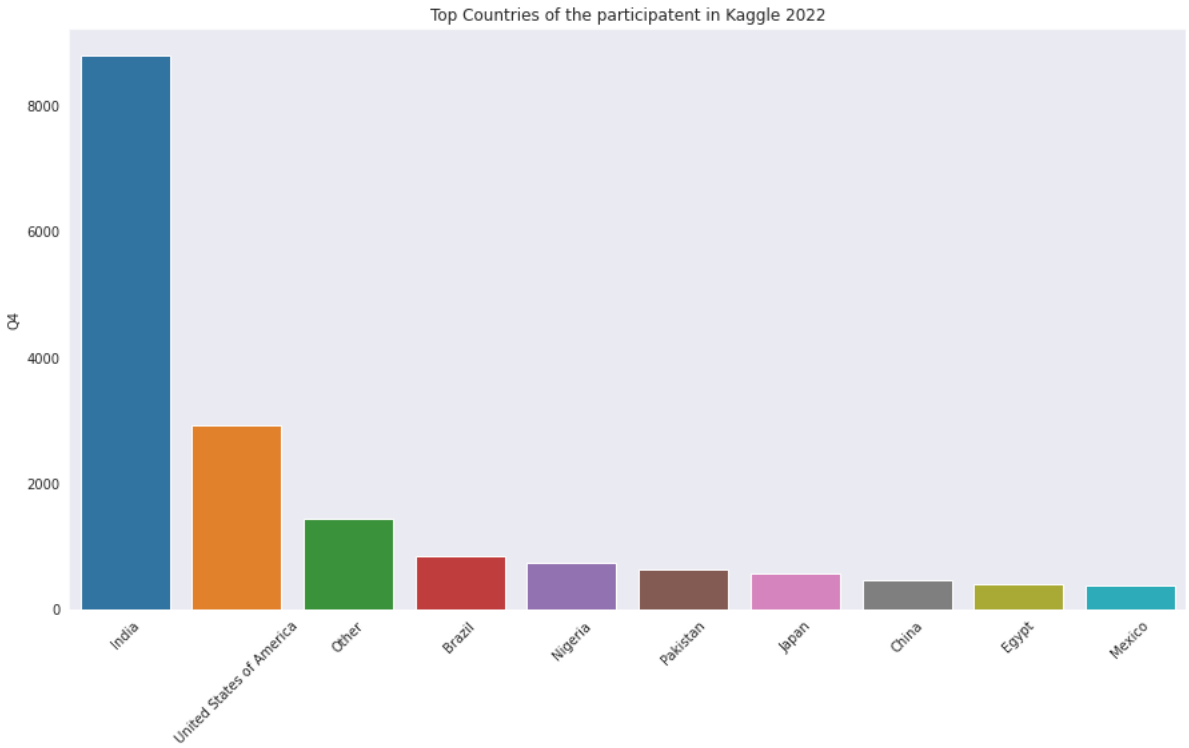
Participation of men are more in number than women, but number of women are also increasing as compare to previous.

In which country do you currently reside?

In [18]:
```
resident = df_kaggle_22['Q4'].value_counts()
resident
```

Out[18]:
```
India                                                  8792
United States of America                               2920
Other                                                  1430
Brazil                                                  833
Nigeria                                                 731
Pakistan                                                620
Japan                                                   556
China                                                   453
Egypt                                                   383
Mexico                                                  380
Indonesia                                               376
Turkey                                                  345
Russia                                                  324
South Korea                                             317
France                                                  262
United Kingdom of Great Britain and Northern Ireland    258
Canada                                                  257
Spain                                                   257
Colombia                                                256
Bangladesh                                              251
Taiwan                                                  242
Viet Nam                                                212
Argentina                                               204
Kenya                                                   201
Italy                                                   182
Morocco                                                 177
Australia                                               142
Thailand                                                132
Tunisia                                                 125
Peru                                                    121
Iran, Islamic Republic of...                            120
Chile                                                   115
Poland                                                  113
South Africa                                            109
Philippines                                             108
Netherlands                                             108
Ghana                                                   107
Israel                                                  102
Germany                                                  99
Ethiopia                                                 98
United Arab Emirates                                     94
Portugal                                                 87
Saudi Arabia                                             84
Ukraine                                                  79
Sri Lanka                                                77
Nepal                                                    75
Malaysia                                                 74
Singapore                                                68
Cameroon                                                 68
Algeria                                                  62
Hong Kong (S.A.R.)                                       58
Zimbabwe                                                 54
Ecuador                                                  54
Ireland                                                  53
Belgium                                                  51
Romania                                                  50
Czech Republic                                           49
I do not wish to disclose my location                    42
In which country do you currently reside?                 1
Name: Q4, dtype: int64
```

In [19]:
```
#Top Ten countries
top_countries = df_kaggle_22['Q4'].value_counts()[:10]
fig, ax = plt.subplots(figsize=(15,8))
# ax.set_ylim([0,20])
ax.set_ylabel("Count")
ax.set_title("Top Countries of the participatent in Kaggle 2022")
plt.xticks(rotation=45)
sns.barplot( x = top_countries.index, y = top_countries, orient='v');
plt.show()
```

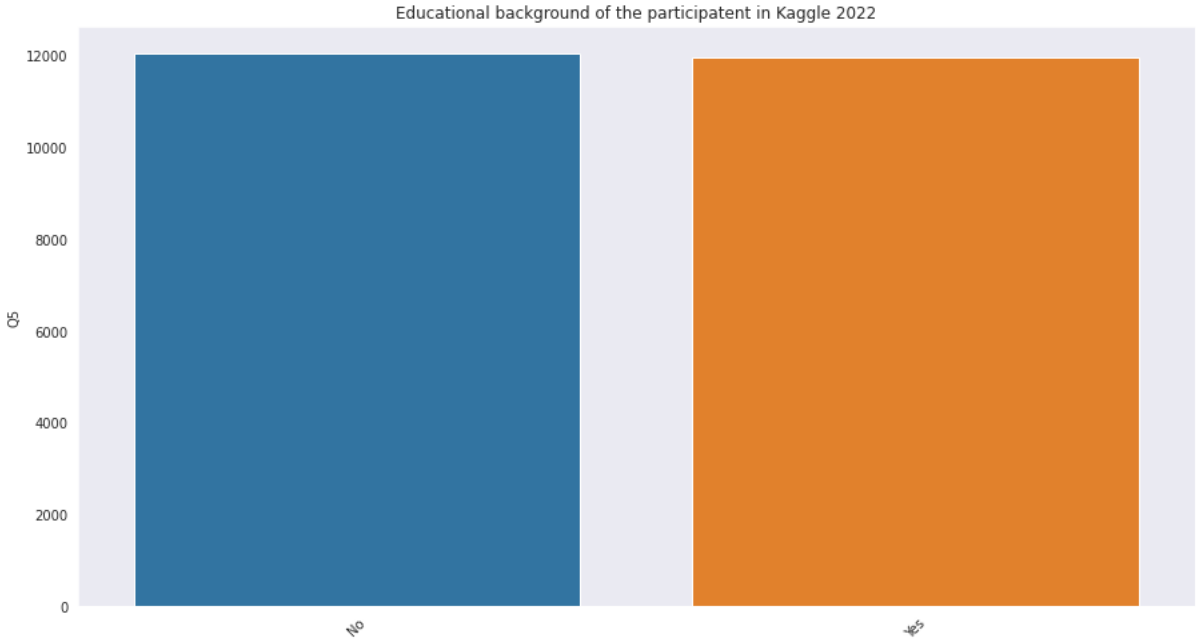


Are you currently a student? (high school, university, or graduate)

In [20]:
```
educational_background_1 = df_kaggle_22['Q5'].value_counts()
educational_background_1
```

Out[20]:
```
No                                                          12036
Yes                                                         11961
Are you currently a student? (high school, university, or graduate)      1
Name: Q5, dtype: int64
```

In [21]:
```
#Educational background
educational_background_1 = df_kaggle_22['Q5'].value_counts()[:2]
fig, ax = plt.subplots(figsize=(15,8))
# ax.set_ylim([0,20])
ax.set_ylabel("Count")
ax.set_title("Educational background of the participatent in Kaggle 2022")
plt.xticks(rotation=45)
sns.barplot( x = educational_background_1.index, y = educational_background_1)
plt.show();
```



Are you currently a student? (high school, university, or graduate). The selection was based on gruaduation or not yes mean gruaduate no mean not graduated yet.

On which platforms have you begun or completed data science courses? (Select all that apply).

- edX
- Kaggle Learn Courses
- DataCamp
- Fast.ai
- Udacity
- Udemy
- LinkedIn Learning
- Cloud-certification programs (direct from AWS, Azure, GCP, or similar)
- University Courses (resulting in a university degree)
- None
- Other This question covered overal all 6 columns

In [22]:
```
platforms_learning = df_kaggle_22[['Q6_2', 'Q6_3', 'Q6_4', 'Q6_5', 'Q6_6', 'Q6_7', 'Q6_8', 'Q6_9','Q6_10','Q6_11', 'Q6_12']]
```

In [23]:
```
platforms_learning
```

Out[23]:

	Q6_2	Q6_3	Q6_4	Q6_5	Q6_6	Q6_7	Q6_8	Q6_9	Q6_10	Q6_11	Q6_12
0	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Other
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	University Courses (resulting in a university ...	NaN	NaN
3	edX	NaN	DataCamp	NaN	Udacity	Udemy	LinkedIn Learning	NaN	University Courses (resulting in a university ...	NaN	NaN
4	NaN	Kaggle Learn Courses	NaN	NaN	NaN	Udemy	NaN	NaN	NaN	NaN	Other
...	...	...	...	...	...	...	...	...	...	...	...
23993	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	None	NaN
23994	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	University Courses (resulting in a university ...	NaN	NaN
23995	NaN	NaN	DataCamp	NaN	NaN	Udemy	NaN	NaN	NaN	NaN	NaN
23996	NaN	Kaggle Learn Courses	NaN	NaN	Udacity	NaN	NaN	NaN	University Courses (resulting in a university ...	NaN	NaN
23997	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Other

23998 rows × 11 columns

In [24]:
```
platforms_learning.dropna(axis='columns', how='all', inplace=True)
platforms_learning
```

```
/opt/conda/lib/python3.7/site-packages/pandas/util/_decorators.py:311: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  return func(*args, **kwargs)
```

Out[24]:

	Q6_2	Q6_3	Q6_4	Q6_5	Q6_6	Q6_7	Q6_8	Q6_9	Q6_10	Q6_11	Q6_12
0	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...	On which platforms have you begun or completed...
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Other
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	University Courses (resulting in a university ...	NaN	NaN
3	edX	NaN	DataCamp	NaN	Udacity	Udemy	LinkedIn Learning	NaN	University Courses (resulting in a university ...	NaN	NaN
4	NaN	Kaggle Learn Courses	NaN	NaN	NaN	Udemy	NaN	NaN	NaN	NaN	Other
...	...	...	...	...	...	...	...	...	...	...	...
23993	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	None	NaN
23994	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	University Courses (resulting in a university ...	NaN	NaN
23995	NaN	NaN	DataCamp	NaN	NaN	Udemy	NaN	NaN	NaN	NaN	NaN
23996	NaN	Kaggle Learn Courses	NaN	NaN	Udacity	NaN	NaN	NaN	University Courses (resulting in a university ...	NaN	NaN
23997	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Other

23998 rows × 11 columns

In [25]:
```
print(platforms_learning)
```

```
                                            Q6_2  \
0      On which platforms have you begun or completed...
1                                            NaN
2                                            NaN
3                                            edX
4                                            NaN
...                                          ...
23993                                        NaN
23994                                        NaN
23995                                        NaN
23996                                        NaN
23997                                        NaN

                                            Q6_3  \
0      On which platforms have you begun or completed...
1                                            NaN
2                                            NaN
3                                            NaN
4                              Kaggle Learn Courses
```

In [26]:
```
edx_participants = df_kaggle_22['Q6_2'].value_counts()[:1]
edx_participants
```

Out[26]:
```
edX    2474
Name: Q6_2, dtype: int64
```

In [27]:
```
Kaggle_Learn_Courses = df_kaggle_22['Q6_3'].value_counts()[:1]
Kaggle_Learn_Courses
```

Out[27]:
```
Kaggle Learn Courses    6628
Name: Q6_3, dtype: int64
```

In [28]:
```
DataCamp_participants = df_kaggle_22["Q6_4"].value_counts()[:1]
DataCamp_participants
```

Out[28]:
```
DataCamp    3718
Name: Q6_4, dtype: int64
```

In [29]:
```
Fast_ai = df_kaggle_22["Q6_5"].value_counts()[:1]
Fast_ai
```

Out[29]:
```
Fast.ai    944
Name: Q6_5, dtype: int64
```

In [30]:
```
Udacity_participants = df_kaggle_22["Q6_6"].value_counts()[:1]
Udacity_participants
```

Out[30]:
```
Udacity    2199
Name: Q6_6, dtype: int64
```

In [31]:
```
Udemy_participants = df_kaggle_22["Q6_7"].value_counts()[:1]
Udemy_participants
```

Out[31]:
```
Udemy    6116
Name: Q6_7, dtype: int64
```

In [32]:
```
LinkedIn_Learning = df_kaggle_22["Q6_8"].value_counts()[:1]
LinkedIn_Learning
```

Out[32]:
```
LinkedIn Learning    2766
Name: Q6_8, dtype: int64
```

Now get all together to see the famous platform of learning

In [33]:
```
print(f"""
      The Famous Platform of Learning Data Science and Machine Learning according to Kaggle Survey 2022
      _____________________________________________________________________________________________
        Edx :  {edx_participants}  ,
        Kaggle Learn Courses : {Kaggle_Learn_Courses},
        DataCamp : {DataCamp_participants},
        Fast.Ai : {Fast_ai},
        Udacity : {Udacity_participants},
        Udemy : {Udemy_participants},
        LinkedIn Learning : {LinkedIn_Learning}
        _____________________________________________________________________________________________
                    These are the famouse platform to start learning
      """
     )
```

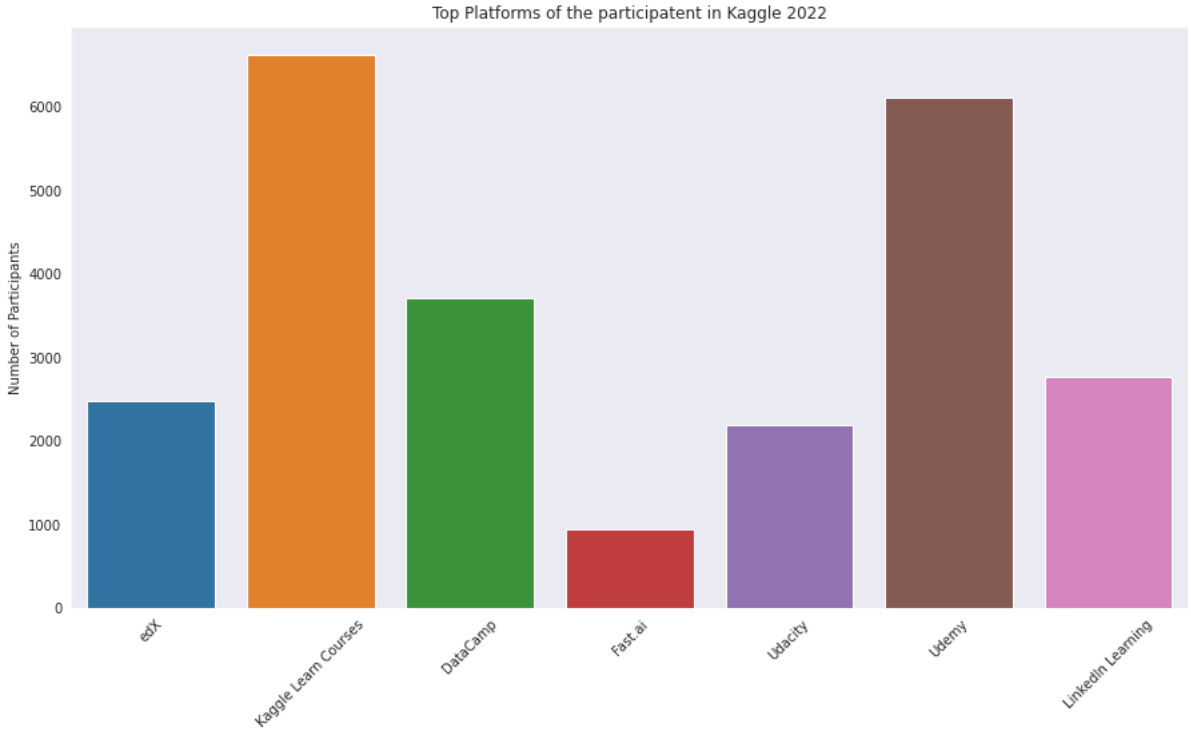
```
      The Famous Platform of Learning Data Science and Machine Learning according to Kaggle Survey 2022
      _____________________________________________________________________________________________
        Edx :  edX    2474
Name: Q6_2, dtype: int64  ,
        Kaggle Learn Courses : Kaggle Learn Courses    6628
Name: Q6_3, dtype: int64,
        DataCamp : DataCamp    3718
Name: Q6_4, dtype: int64,
        Fast.Ai : Fast.ai    944
Name: Q6_5, dtype: int64,
        Udacity : Udacity    2199
Name: Q6_6, dtype: int64,
        Udemy : Udemy    6116
Name: Q6_7, dtype: int64,
        LinkedIn Learning : LinkedIn Learning    2766
Name: Q6_8, dtype: int64
        _____________________________________________________________________________________________
                    These are the famouse platform to start learning
```

So according to this Survey Kaggle Learnin Courses platform are leading and Udemy is th second position to learn data science and machine learning.

In [34]:
```
top_platforms = pd.DataFrame([edx_participants, Kaggle_Learn_Courses, DataCamp_participants, Fast_ai, Udacity_participants, Udemy_participants, LinkedIn_Learning]).sum()
top_platforms
```

Out[34]:
```
edX                   2474.0
Kaggle Learn Courses  6628.0
DataCamp              3718.0
Fast.ai                944.0
Udacity               2199.0
Udemy                 6116.0
LinkedIn Learning     2766.0
dtype: float64
```

In [35]:
```
#Top platforms
fig, ax = plt.subplots(figsize=(15,8))
# ax.set_ylim([0,20])
ax.set_ylabel("Number of Participants")
ax.set_title("Top Platforms of the participatent in Kaggle 2022")
plt.xticks(rotation=45)
sns.barplot( x = top_platforms.index, y = top_platforms, orient='v');
plt.show()
```



In [36]:
```
top_products = df_kaggle_22[['Q7_1','Q7_2','Q7_3','Q7_4','Q7_5','Q7_6','Q7_7']]
top_products
```

Out[36]:

	Q7_1	Q7_2	Q7_3	Q7_4	Q7_5	Q7_6	Q7_7
0	What products or platforms did you find to be ...	What products or platforms did you find to be ...	What products or platforms did you find to be ...	What products or platforms did you find to be ...	What products or platforms did you find to be ...	What products or platforms did you find to be ...	What products or platforms did you find to be ...
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	University courses	NaN	NaN	NaN	Kaggle (notebooks, competitions, etc)	NaN	NaN
3	NaN	Online courses (Coursera, EdX, etc)	NaN	Video platforms (YouTube, Twitch, etc)	Kaggle (notebooks, competitions, etc)	NaN	NaN
4	NaN	Online courses (Coursera, EdX, etc)	NaN	NaN	Kaggle (notebooks, competitions, etc)	NaN	NaN
...	...	...	...	...	...	...	...
23993	University courses	NaN	NaN	Video platforms (YouTube, Twitch, etc)	Kaggle (notebooks, competitions, etc)	NaN	Other
23994	University courses	NaN	NaN	Video platforms (YouTube, Twitch, etc)	NaN	NaN	NaN
23995	NaN	Online courses (Coursera, EdX, etc)	Social media platforms (Reddit, Twitter, etc)	Video platforms (YouTube, Twitch, etc)	Kaggle (notebooks, competitions, etc)	NaN	NaN
23996	NaN	NaN	NaN	NaN	Kaggle (notebooks, competitions, etc)	NaN	NaN
23997	NaN	NaN	NaN	Video platforms (YouTube, Twitch, etc)	Kaggle (notebooks, competitions, etc)	NaN	NaN

23998 rows × 7 columns

In [37]:
```
university_courses = df_kaggle_22['Q7_1'].value_counts()[:1]
university_courses
```

Out[37]:
```
University courses    6851
Name: Q7_1, dtype: int64
```

In [38]:
```
online_courses = df_kaggle_22['Q7_2'].value_counts()[:1]
online_courses
```

Out[38]:
```
Online courses (Coursera, EdX, etc)    13714
Name: Q7_2, dtype: int64
```

In [39]:
```
social_media = df_kaggle_22['Q7_3'].value_counts()[:1]
social_media
```

Out[39]:
```
Social media platforms (Reddit, Twitter, etc)    3310
Name: Q7_3, dtype: int64
```

In [40]:
```
video_platforms = df_kaggle_22['Q7_4'].value_counts()[:1]
video_platforms
```

Out[40]:
```
Video platforms (YouTube, Twitch, etc)    12871
Name: Q7_4, dtype: int64
```

In [41]:
```
kaggle_notebook = df_kaggle_22['Q7_5'].value_counts()[:1]
kaggle_notebook
```

Out[41]:
```
Kaggle (notebooks, competitions, etc)    12700
Name: Q7_5, dtype: int64
```

What products or platforms did you find to be most helpful when you first started studying data science?

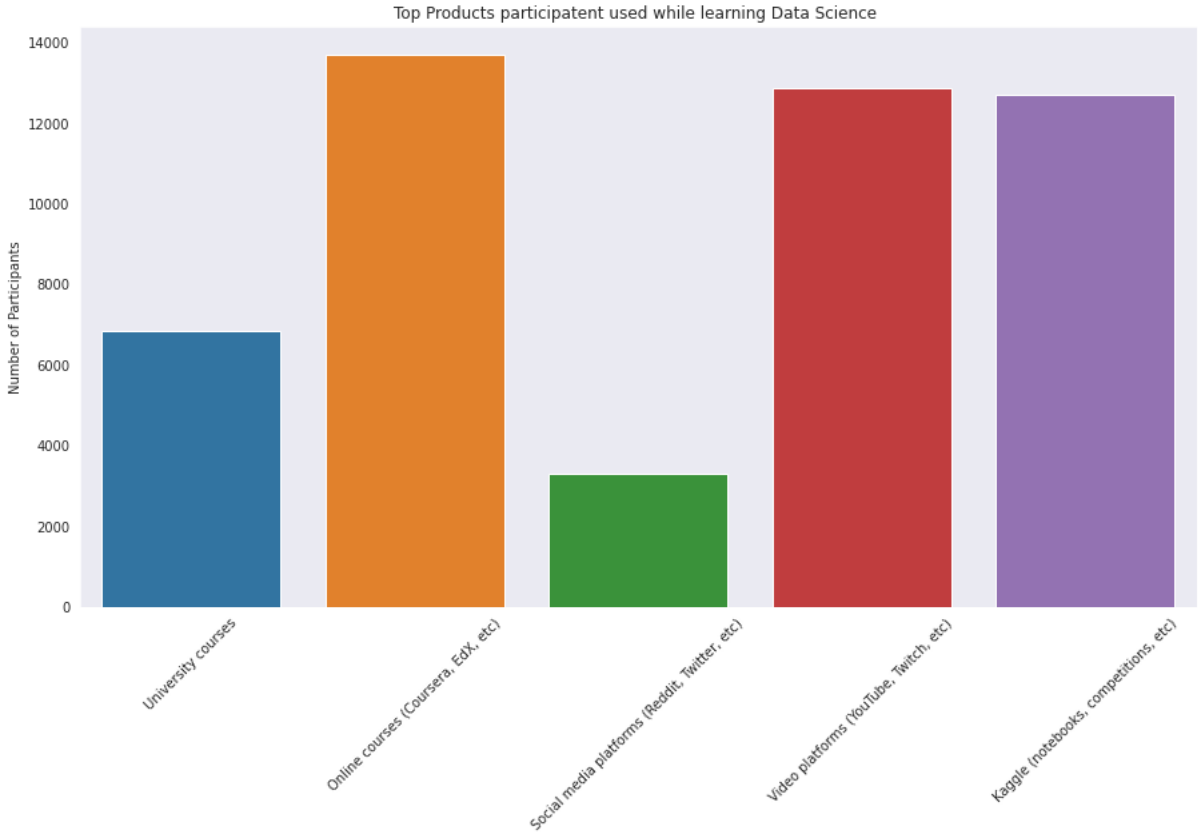
Options are:

- University courses
- Online courses (Coursera, EdX, etc)
- Social media platforms (Reddit, Twitter, etc)
- Video platforms (YouTube, Twitch, etc)
- Kaggle (notebooks, competitions, etc)

In [42]:
```
products_users = pd.DataFrame([university_courses, online_courses, social_media, video_platforms, kaggle_notebook]).sum()
products_users
```

Out[42]:
```
University courses                              6851.0
Online courses (Coursera, EdX, etc)            13714.0
Social media platforms (Reddit, Twitter, etc)   3310.0
Video platforms (YouTube, Twitch, etc)         12871.0
Kaggle (notebooks, competitions, etc)          12700.0
dtype: float64
```

In [43]:
```
#Top products
fig, ax = plt.subplots(figsize=(15,8))
# ax.set_ylim([0,20])
ax.set_ylabel("Number of Participants")
ax.set_title("Top Products participatent used while learning Data Science ")
plt.xticks(rotation=45)
sns.barplot( x = products_users.index, y = products_users, orient='v');
plt.show()
```



The above diagram shows that online courses, video platforms and kaggle notebook are leading

What programming languages do you use on a regular basis?

let see the top programming languages

In [44]:
```
python_user = df_kaggle_22['Q12_1'].value_counts()[:1]
python_user
```

Out[44]:
```
Python    18653
Name: Q12_1, dtype: int64
```

In [45]:
```
r_user = df_kaggle_22['Q12_2'].value_counts()[:1]
r_user
```

Out[45]:
```
R    4571
Name: Q12_2, dtype: int64
```

In [46]:
```
sql_user = df_kaggle_22['Q12_3'].value_counts()[:1]
sql_user
```

Out[46]:
```
SQL    9620
Name: Q12_3, dtype: int64
```

In [47]:
```
c_user = df_kaggle_22['Q12_4'].value_counts()[:1]
c_user
```

Out[47]:
```
C    3801
Name: Q12_4, dtype: int64
```

In [48]:
```
c_hash_user = df_kaggle_22['Q12_5'].value_counts()[:1]
c_hash_user
```

Out[48]:
```
C#    1473
Name: Q12_5, dtype: int64
```

In [49]:
```
c_plus_plus_user = df_kaggle_22['Q12_6'].value_counts()[:1]
c_plus_plus_user
```

Out[49]:
```
C++    4549
Name: Q12_6, dtype: int64
```

In [50]:
```
java_user = df_kaggle_22['Q12_7'].value_counts()[:1]
java_user
```

Out[50]:
```
Java    3862
Name: Q12_7, dtype: int64
```

In [51]:
```
javascript_user = df_kaggle_22['Q12_8'].value_counts()[:1]
javascript_user
```

Out[51]:
```
Javascript    3489
Name: Q12_8, dtype: int64
```

In [52]:
```
bash_user = df_kaggle_22['Q12_9'].value_counts()[:1]
bash_user
```

Out[52]:
```
Bash    1674
Name: Q12_9, dtype: int64
```

In [53]:
```
php_user = df_kaggle_22['Q12_10'].value_counts()[:1]
php_user
```

Out[53]:
```
PHP    1443
Name: Q12_10, dtype: int64
```

In [54]:
```
matlab_user = df_kaggle_22['Q12_11'].value_counts()[:1]
matlab_user
```

Out[54]:
```
MATLAB    2441
Name: Q12_11, dtype: int64
```

In [55]:
```
julia_user = df_kaggle_22['Q12_12'].value_counts()[:1]
julia_user
```

Out[55]:
```
Julia    296
Name: Q12_12, dtype: int64
```

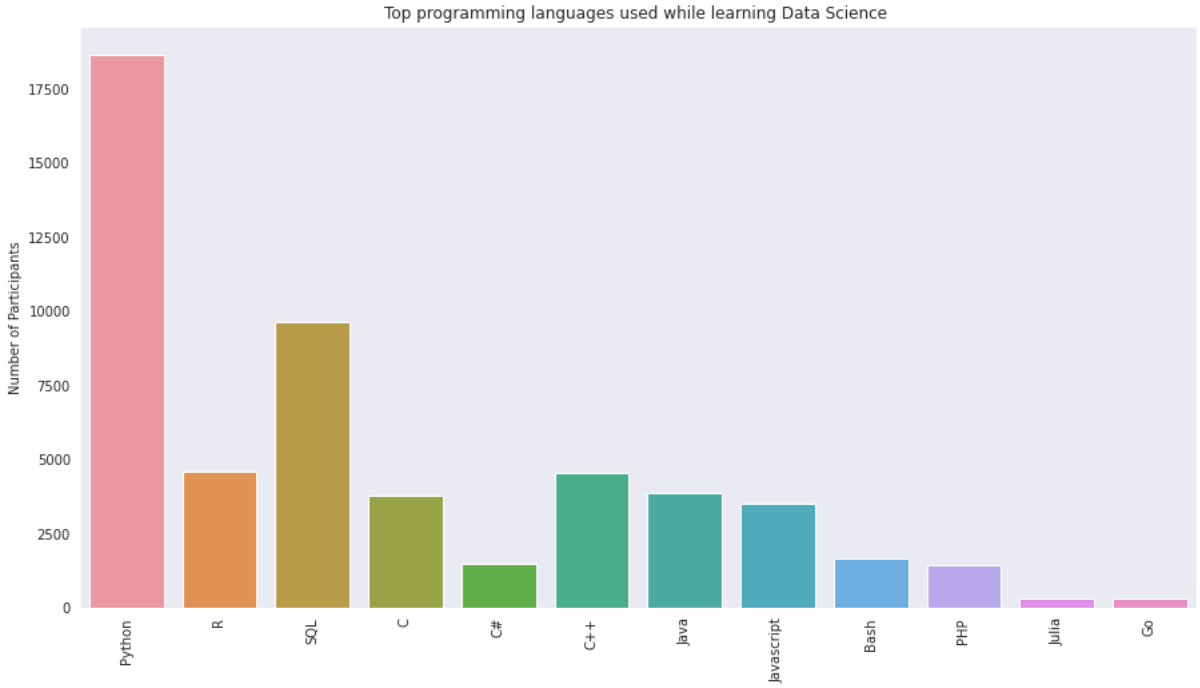
In [56]:
```
go_user = df_kaggle_22['Q12_13'].value_counts()[:1]
go_user
```

Out[56]:
```
Go    322
Name: Q12_13, dtype: int64
```

In [57]:
```
top_programming_lang = pd.DataFrame([python_user, r_user, sql_user, c_user, c_hash_user, c_plus_plus_user, java_user, javascript_user, bash_user, php_user, julia_user, go_user
top_programming_lang
```

Out[57]:
```
Python        18653.0
R              4571.0
SQL            9620.0
C              3801.0
C#             1473.0
C++            4549.0
Java           3862.0
Javascript     3489.0
Bash           1674.0
PHP            1443.0
Julia           296.0
Go              322.0
dtype: float64
```

In [58]:
```
#Top programming languages
fig, ax = plt.subplots(figsize=(15,8))
# ax.set_ylim([0,20])
ax.set_ylabel("Number of Participants")
ax.set_title("Top programming languages used while learning Data Science ")
plt.xticks(rotation=90)
sns.barplot( x = top_programming_lang.index, y = top_programming_lang, orient='v');
plt.show()
```



Python and SQL are leading the data science world!

Which of the following integrated development environments (IDE's) do you use on a regular basis?

- JupyterLab
- RStudio
- Visual Studio
- Visual Studio Code (VSCode
- PyCharm
- Sublime Text
- Jupyter Notebook

In [59]:
```
jupyterlab = df_kaggle_22['Q13_1'].value_counts()[:1]
jupyterlab
```

Out[59]:
```
JupyterLab     4887
Name: Q13_1, dtype: int64
```

In [60]:
```
rstudio = df_kaggle_22['Q13_2'].value_counts()[:1]
rstudio
```

Out[60]:
```
 RStudio     3824
Name: Q13_2, dtype: int64
```

In [61]:
```
visual_studio = df_kaggle_22['Q13_3'].value_counts()[:1]
visual_studio
```

Out[61]:
```
 Visual Studio     4416
Name: Q13_3, dtype: int64
```

In [62]:
```
visual_studio_code = df_kaggle_22['Q13_4'].value_counts()[:1]
visual_studio_code
```

Out[62]:
```
 Visual Studio Code (VSCode)     9976
Name: Q13_4, dtype: int64
```

In [63]:
```
pycharm = df_kaggle_22['Q13_5'].value_counts()[:1]
pycharm
```

Out[63]:
```
 PyCharm     6099
Name: Q13_5, dtype: int64
```

In [64]:
```
sublime_text = df_kaggle_22['Q13_8'].value_counts()[:1]
sublime_text
```

Out[64]:
```
  Sublime Text      2218
Name: Q13_8, dtype: int64
```

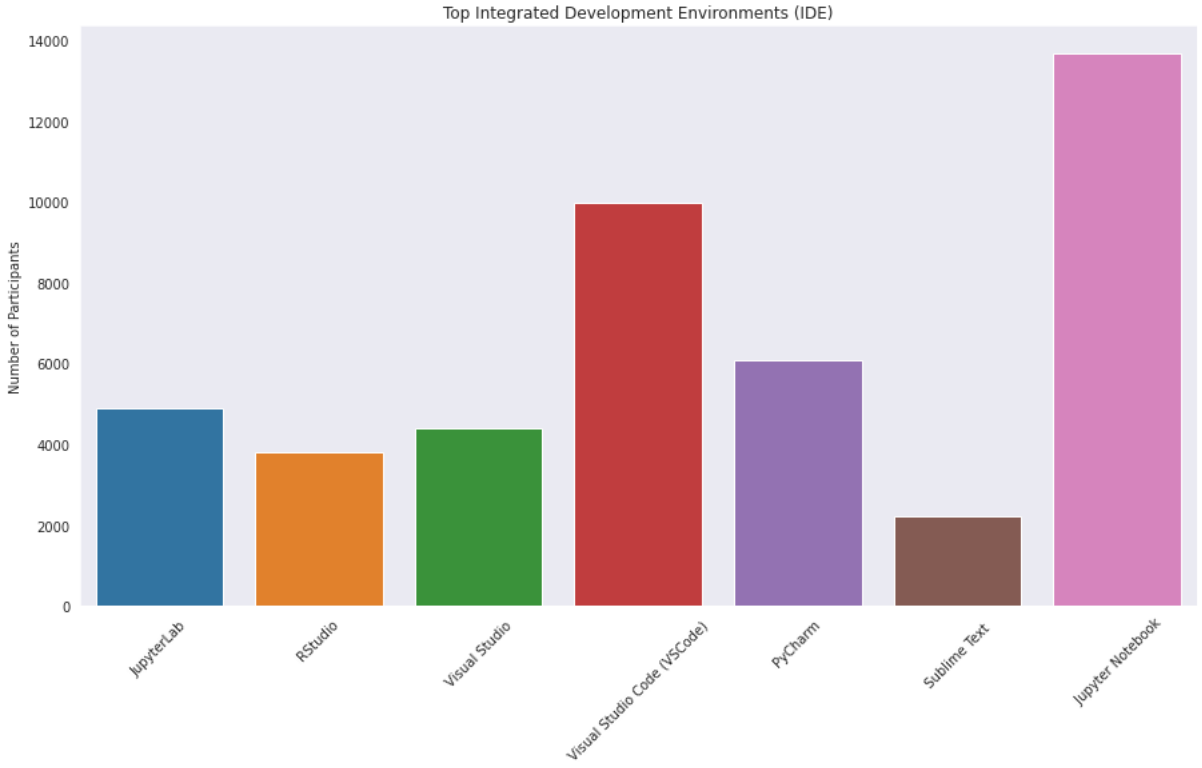
In [65]:
```
jupyter_notebook = df_kaggle_22['Q13_11'].value_counts()[:1]
jupyter_notebook
```

Out[65]:
```
 Jupyter Notebook    13684
Name: Q13_11, dtype: int64
```

In [66]:
```
ides = pd.DataFrame([jupyterlab,rstudio, visual_studio, visual_studio_code,pycharm, sublime_text, jupyter_notebook]).sum()
ides
```

Out[66]:
```
JupyterLab                      4887.0
 RStudio                        3824.0
 Visual Studio                  4416.0
 Visual Studio Code (VSCode)    9976.0
 PyCharm                        6099.0
  Sublime Text                  2218.0
 Jupyter Notebook              13684.0
dtype: float64
```

In [67]:
```
#Top Integrated Development Environments
fig, ax = plt.subplots(figsize=(15,8))
sns.set_style("ticks")
# ax.set_ylim([0,20])
ax.set_ylabel("Number of Participants")
ax.set_title("Top Integrated Development Environments (IDE)")
plt.xticks(rotation=45)
sns.barplot( x = ides.index, y = ides, orient='v');
plt.show()
```



so Jupyternotebook and VSCode is the most use IDE

For how many years have you used machine learning methods?

In [68]:
```
ml_exp = df_kaggle_22['Q16'].value_counts()
```

In [69]:
```
ml_exp
```

Out[69]:
```
Under 1 year                                                  7221
1-2 years                                                     3720
I do not use machine learning methods                         3419
2-3 years                                                     1947
5-10 years                                                    1090
3-4 years                                                     1053
4-5 years                                                      950
10-20 years                                                    483
20 or more years                                                 3
For how many years have you used machine learning methods?       1
Name: Q16, dtype: int64
```

In [70]:
```
ml_exp_data_frame = pd.DataFrame(ml_exp)[:-1]
ml = ml_exp_data_frame.rename(columns= {'Q16' : 'Total_nums'})
ml
```

Out[70]:

	Total_nums
Under 1 year	7221
1-2 years	3720
I do not use machine learning methods	3419
2-3 years	1947
5-10 years	1090
3-4 years	1053
4-5 years	950
10-20 years	483
20 or more years	3

Which of the following machine learning frameworks do you use on a regular basis?

- Scikit-learn
- TensorFlow
- Keras
- PyTorch
- fast.ai
- Xgboost
- Caret
- Jax

In [71]:
```
sckit_learn = df_kaggle_22['Q17_1'].value_counts()[:1]
sckit_learn
```

Out[71]:
```
Scikit-learn     11403
Name: Q17_1, dtype: int64
```

In [72]:
```
tensor_flow = df_kaggle_22['Q17_2'].value_counts()[:1]
tensor_flow
```

Out[72]:
```
TensorFlow     7953
Name: Q17_2, dtype: int64
```

In [73]:
```
keras = df_kaggle_22['Q17_3'].value_counts()[:1]
keras
```

Out[73]:
```
 Keras     6575
Name: Q17_3, dtype: int64
```

In [74]:
```
pytorch = df_kaggle_22['Q17_4'].value_counts()[:1]
pytorch
```

Out[74]:
```
 PyTorch     5191
Name: Q17_4, dtype: int64
```

In [75]:
```
fast_ai = df_kaggle_22['Q17_5'].value_counts()[:1]
fast_ai
```

Out[75]:
```
 Fast.ai     648
Name: Q17_5, dtype: int64
```

In [76]:
```
xgboost = df_kaggle_22['Q17_6'].value_counts()[:1]
xgboost
```

Out[76]:
```
 Xgboost     4477
Name: Q17_6, dtype: int64
```

In [77]:
```
caret = df_kaggle_22['Q17_9'].value_counts()[:1]
caret
```

Out[77]:
```
 Caret     821
Name: Q17_9, dtype: int64
```

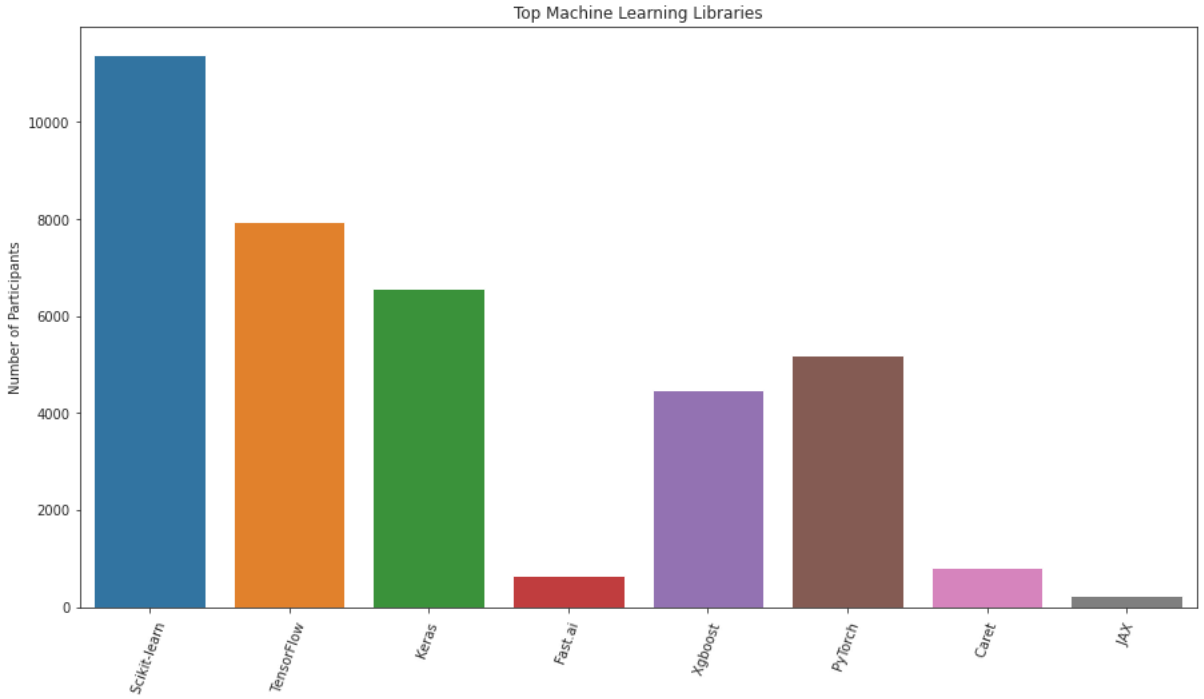
In [78]:
```
jax = df_kaggle_22['Q17_11'].value_counts()[:1]
jax
```

Out[78]:
```
 JAX     252
Name: Q17_11, dtype: int64
```

In [79]:
```
machine_learning_libraries = pd.DataFrame([sckit_learn, tensor_flow, keras,fast_ai, xgboost, pytorch, caret, jax]).sum()
machine_learning_libraries
```

Out[79]:
```
  Scikit-learn     11403.0
  TensorFlow        7953.0
 Keras              6575.0
 Fast.ai             648.0
 Xgboost            4477.0
 PyTorch            5191.0
 Caret               821.0
 JAX                 252.0
dtype: float64
```

In [80]:
```
#Top Mchine Learning Libraries
fig, ax = plt.subplots(figsize=(15,8))
# ax.set_ylim([0,20])
ax.set_ylabel("Number of Participants")
ax.set_title("Top Machine Learning Libraries")
plt.xticks(rotation=70)
sns.barplot( y = machine_learning_libraries, x = machine_learning_libraries.index, orient='v')
plt.show();
```



Scikit-learn is the most used library of Machine Learning

Which of the following ML algorithms do you use on a regular basis?

- Linear or Logistic Regression
- Decision Trees or Random Forests
- Gradient Boosting Machines (xgboost, lightgbm, etc)
- Bayesian Approaches
- Evolutionary Approaches
- Dense Neural Networks (MLPs, etc)
- Convolutional Neural Networks
- Generative Adversarial Networks

In [81]:
```
lg = df_kaggle_22['Q18_1'].value_counts()[:1]
lg
```

Out[81]:
```
Linear or Logistic Regression    11338
Name: Q18_1, dtype: int64
```

In [82]:
```
rm = df_kaggle_22['Q18_2'].value_counts()[:1]
rm
```

Out[82]:
```
Decision Trees or Random Forests    9373
Name: Q18_2, dtype: int64
```

In [83]:
```
gb = df_kaggle_22['Q18_3'].value_counts()[:1]
gb
```

Out[83]:
```
Gradient Boosting Machines (xgboost, lightgbm, etc)    5506
Name: Q18_3, dtype: int64
```

In [84]:
```
ba = df_kaggle_22['Q18_4'].value_counts()[:1]
ba
```

Out[84]:
```
Bayesian Approaches    3661
Name: Q18_4, dtype: int64
```

In [85]:
```
ea = df_kaggle_22['Q18_5'].value_counts()[:1]
ea
```

Out[85]:
```
Evolutionary Approaches    823
Name: Q18_5, dtype: int64
```

In [86]:
```
dnn = df_kaggle_22['Q18_6'].value_counts()[:1]
dnn
```

Out[86]:
```
Dense Neural Networks (MLPs, etc)    3476
Name: Q18_6, dtype: int64
```

In [87]:
```
cnn = df_kaggle_22['Q18_7'].value_counts()[:1]
cnn
```

Out[87]:
```
Convolutional Neural Networks    6006
Name: Q18_7, dtype: int64
```

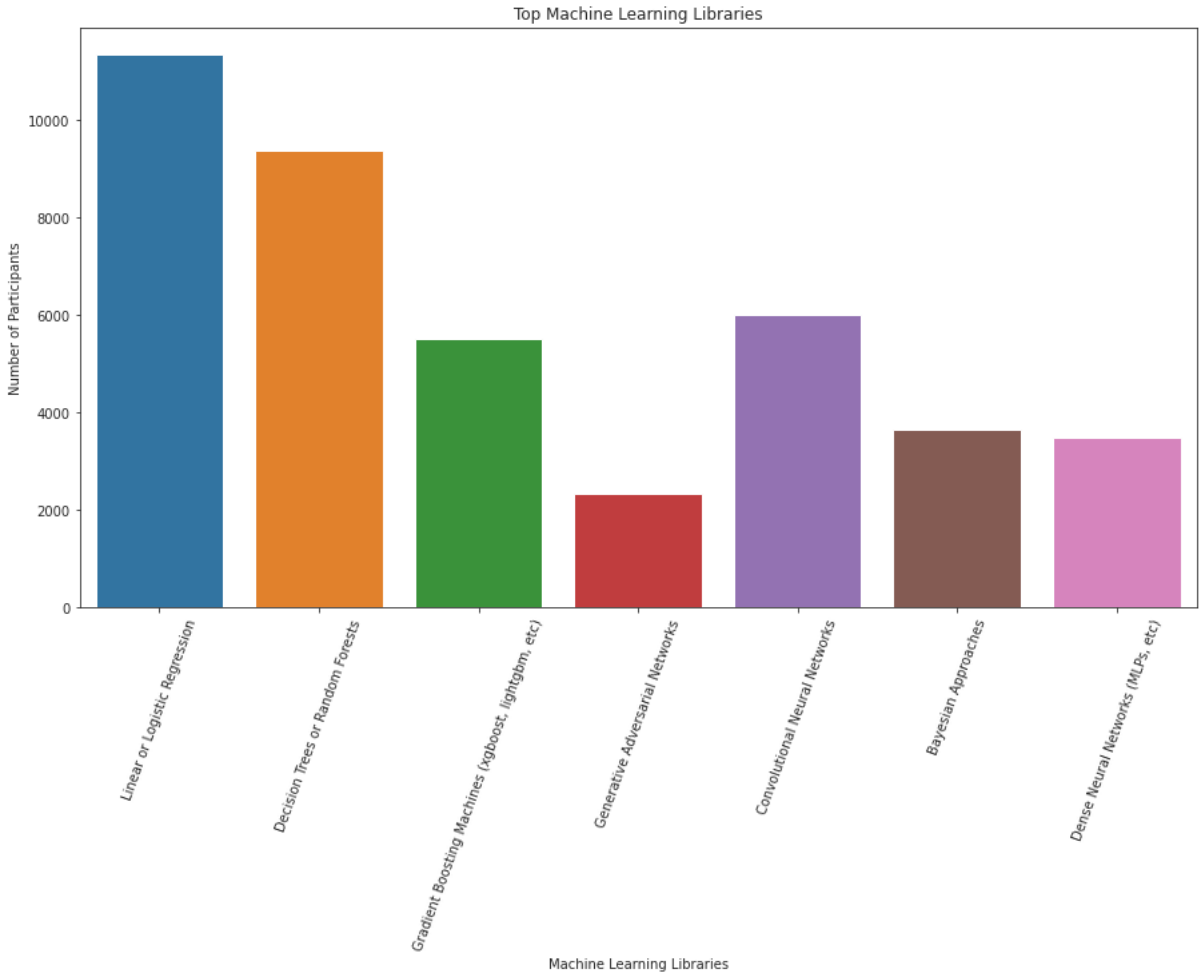
In [88]:
```
gan = df_kaggle_22['Q18_8'].value_counts()[:1]
gan
```

Out[88]:
```
Generative Adversarial Networks    1166
Name: Q18_8, dtype: int64
```

In [89]:
```
ml_algorithms = pd.DataFrame([lg,rm,gb,gan,cnn,gan,ba,dnn]).sum()
ml_algorithms
```

Out[89]:
```
Linear or Logistic Regression                        11338.0
Decision Trees or Random Forests                      9373.0
Gradient Boosting Machines (xgboost, lightgbm, etc)    5506.0
Generative Adversarial Networks                       2332.0
Convolutional Neural Networks                         6006.0
Bayesian Approaches                                   3661.0
Dense Neural Networks (MLPs, etc)                     3476.0
dtype: float64
```

In [90]:
```
#Top Mchine Learning Algorithms
fig, ax = plt.subplots(figsize=(15,8))
sns.set_style("ticks")
# ax.set_ylim([0,20])
ax.set_ylabel("Number of Participants")
ax.set_title("Top Machine Learning Libraries")
ax.set_xlabel("Machine Learning Libraries")
plt.xticks(rotation=70)
sns.barplot( y = ml_algorithms, x = ml_algorithms.index, orient='v')
plt.show();
```



Select the title most similar to your current role (or most recent title if retired):

In [91]:
```
role = df_kaggle_22["Q23"].value_counts()[:-1]
role
```

Out[91]:
```
Data Scientist                                                  1929
Data Analyst (Business, Marketing, Financial, Quantitative, etc)    1538
Currently not employed                                          1432
Software Engineer                                                980
Teacher / professor                                              833
Manager (Program, Project, Operations, Executive-level, etc)     832
Other                                                            754
Research Scientist                                               593
Machine Learning/ MLops Engineer                                 571
Engineer (non-software)                                          465
Data Engineer                                                    352
Statistician                                                     125
Data Architect                                                    95
Data Administrator                                                70
Developer Advocate                                                61
Name: Q23, dtype: int64
```

In [92]:
```
titles_in_ai_world = pd.DataFrame([role])
titles_in_ai_world
```

Out[92]:

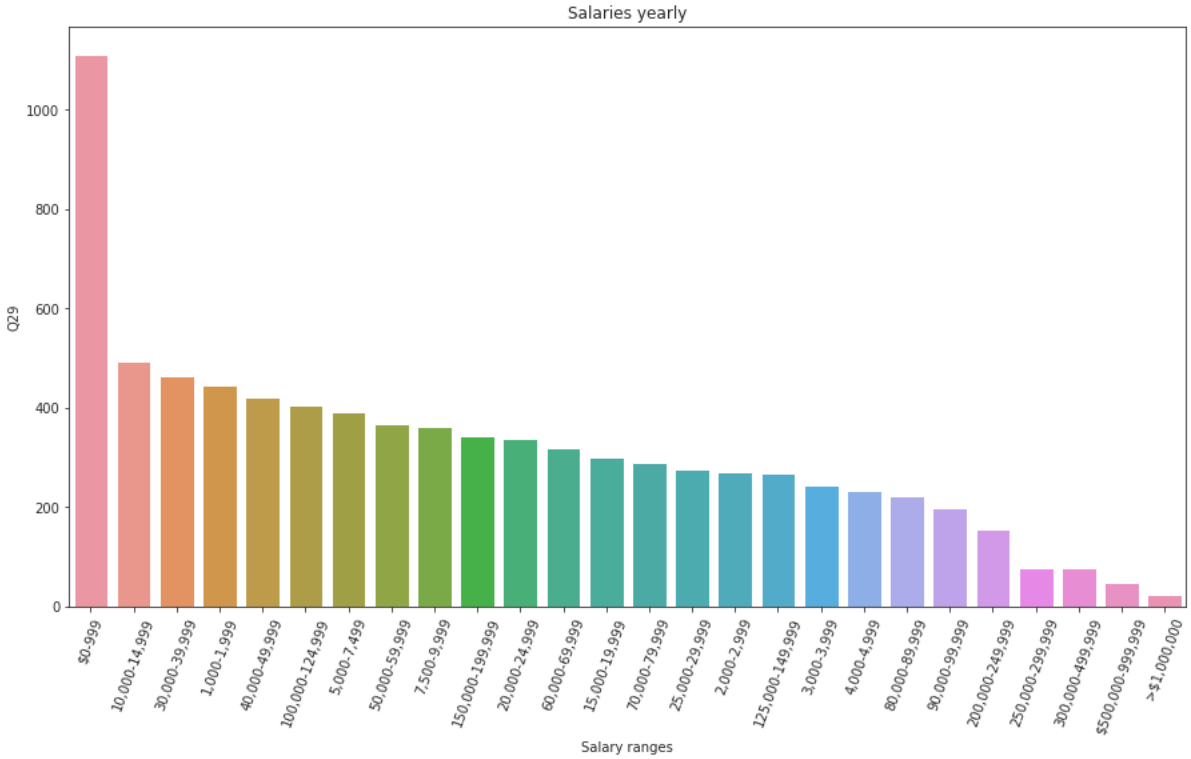
	Data Scientist	Data Analyst (Business, Marketing, Financial, Quantitative, etc)	Currently not employed	Software Engineer	Teacher / professor	Manager (Program, Project, Operations, Executive-level, etc)	Other	Research Scientist	Machine Learning/ MLops Engineer	Engineer (non-software)	Data Engineer	Statistician	Data Architect	Data Administrator	Developer Advocate
Q23	1929	1538	1432	980	833	832	754	593	571	465	352	125	95	70	61

What is your current yearly compensation (approximate $USD)?

In [93]:
```
yearly_salary = df_kaggle_22['Q29'].value_counts()[:-1]
yearly_salary
```

Out[93]:
```
$0-999              1112
10,000-14,999        493
30,000-39,999        464
1,000-1,999          444
40,000-49,999        421
100,000-124,999      404
5,000-7,499          391
50,000-59,999        366
7,500-9,999          362
150,000-199,999      342
20,000-24,999        337
60,000-69,999        318
15,000-19,999        299
70,000-79,999        289
25,000-29,999        277
2,000-2,999          271
125,000-149,999      269
3,000-3,999          244
4,000-4,999          234
80,000-89,999        222
90,000-99,999        197
200,000-249,999      155
250,000-299,999       78
300,000-499,999       76
$500,000-999,999      48
>$1,000,000           23
Name: Q29, dtype: int64
```

In [94]:
```
#Salary range
fig, ax = plt.subplots(figsize=(15,8))
sns.set_style("ticks")
# ax.set_ylim([0,20])
ax.set_ylabel("Number of Particpants")
ax.set_title("Salaries yearly")
ax.set_xlabel("Salary ranges")
plt.xticks(rotation=70)
sns.barplot( y = yearly_salary, x = yearly_salary.index, orient='v')
plt.show();
```



Which of the following cloud computing platforms do you use? (Select all that apply)

- AWS
- Azure
- GCP

Compare these three famous plateforms

In [95]:
```
com_platforms = df_kaggle_22[['Q31_1', 'Q31_2','Q31_3']]
com_platforms
```

Out[95]:

	Q31_1	Q31_2	Q31_3
0	Which of the following cloud computing platfor...	Which of the following cloud computing platfor...	Which of the following cloud computing platfor...
1	NaN	NaN	NaN
2	NaN	NaN	NaN
3	NaN	NaN	NaN
4	NaN	NaN	NaN
...	...	...	...
23993	NaN	NaN	NaN
23994	NaN	NaN	NaN
23995	NaN	Microsoft Azure	NaN
23996	NaN	NaN	NaN
23997	NaN	NaN	NaN

23998 rows × 3 columns

In [96]:
```
aws = df_kaggle_22['Q31_1'].value_counts()[:-1]
aws
```

Out[96]:
```
 Amazon Web Services (AWS)     2346
Name: Q31_1, dtype: int64
```

In [97]:
```
azure = df_kaggle_22['Q31_2'].value_counts()[:-1]
azure
```

Out[97]:
```
 Microsoft Azure     1416
Name: Q31_2, dtype: int64
```

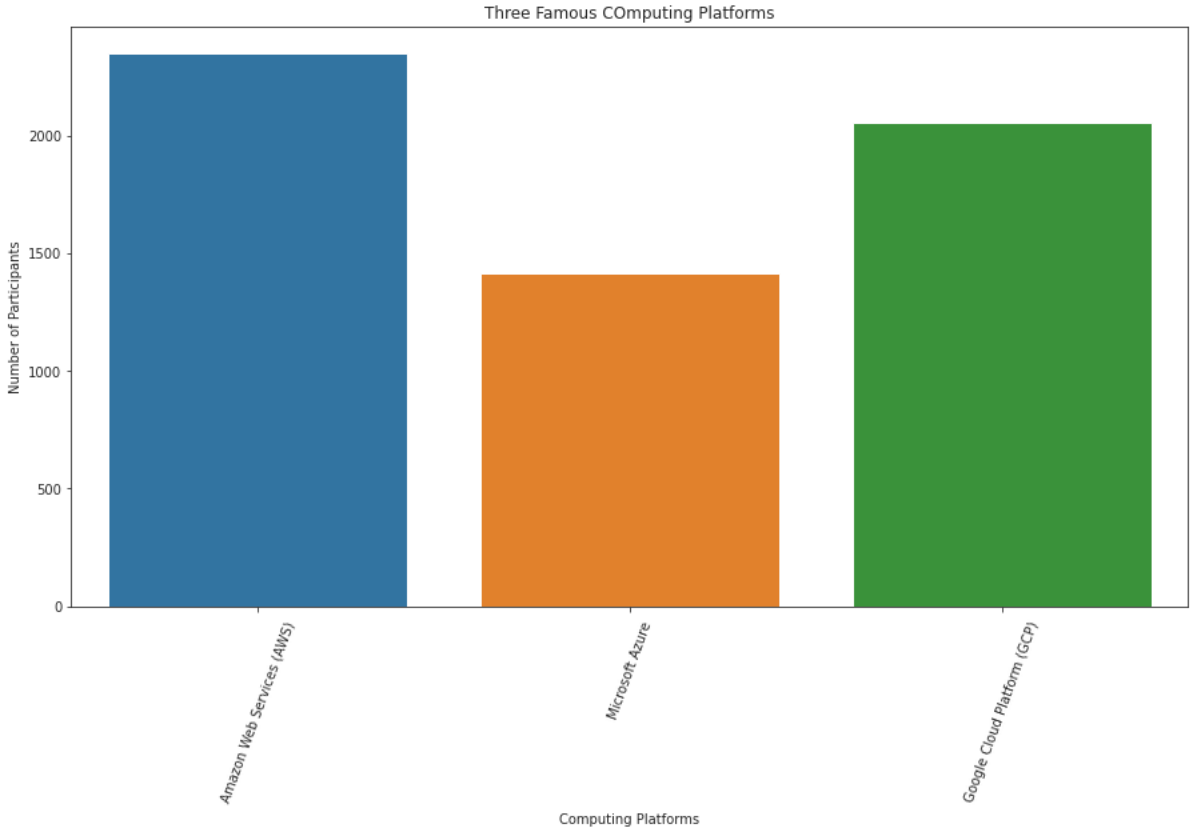
In [98]:
```
gcp = df_kaggle_22['Q31_3'].value_counts()[:-1]
gcp
```

Out[98]:
```
 Google Cloud Platform (GCP)     2056
Name: Q31_3, dtype: int64
```

In [99]:
```
com_platforms = pd.DataFrame([aws, azure, gcp]).sum()
com_platforms
```

Out[99]:
```
 Amazon Web Services (AWS)       2346.0
 Microsoft Azure                 1416.0
 Google Cloud Platform (GCP)     2056.0
dtype: float64
```

In [100]:
```
#Salary range
fig, ax = plt.subplots(figsize=(15,8))
sns.set_style("ticks")
# ax.set_ylim([0,20])
ax.set_ylabel("Number of Participants")
ax.set_title("Three Famous COmputing Platforms")
ax.set_xlabel("Computing Platforms")
plt.xticks(rotation=70)
sns.barplot( y = com_platforms, x = com_platforms.index, orient='v')
plt.show();
```



AWS is leading the computing platform.

Do you use any of the following business intelligence tools?

- Amazon QuickSight
- Microsoft Power BI
- Google Data Studio
- Tableau

In [101]:
```
aqs = df_kaggle_22['Q36_1'].value_counts()[:-1]
aqs
```

Out[101]:
```
Amazon QuickSight    224
Name: Q36_1, dtype: int64
```

In [102]:
```
pbi = df_kaggle_22['Q36_2'].value_counts()[:-1]
pbi
```

Out[102]:
```
Microsoft Power BI    1658
Name: Q36_2, dtype: int64
```

In [103]:
```
gds = df_kaggle_22['Q36_3'].value_counts()[:-1]
gds
```

Out[103]:
```
Google Data Studio    643
Name: Q36_3, dtype: int64
```

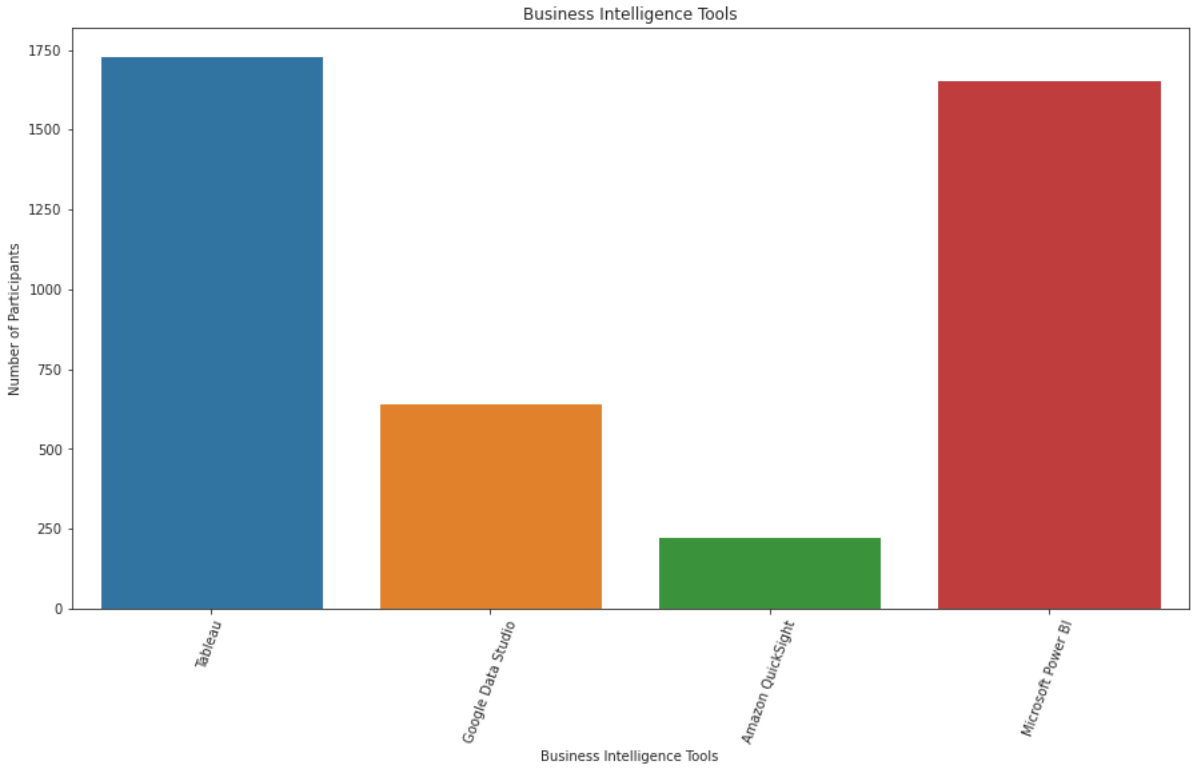
In [104]:
```
tb = df_kaggle_22['Q36_5'].value_counts()[:-1]
tb
```

Out[104]:
```
Tableau    1732
Name: Q36_5, dtype: int64
```

In [105]:
```
business_ai_tools = pd.DataFrame([tb, gds, aqs, pbi]).sum()
business_ai_tools
```

Out[105]:
```
Tableau               1732.0
Google Data Studio     643.0
Amazon QuickSight      224.0
Microsoft Power BI    1658.0
dtype: float64
```

In [106]:
```
#Salary range
fig, ax = plt.subplots(figsize=(15,8))
sns.set_style("ticks")
# ax.set_ylim([0,20])
ax.set_ylabel("Number of Particpants")
ax.set_title("Business Intelligence Tools")
ax.set_xlabel("Business Intelligence Tools")
plt.xticks(rotation=70)
sns.barplot( y = business_ai_tools, x = business_ai_tools.index, orient='v')
plt.show();
```



Tablue and Power BI is leading the business intellegent platform and visualization.

In [ ]: