------------------------------Exploratory Data Analysis--

-------------------------------------Importing libraries--

In [6]:

```
import pandas as pd
import numpy as np
```

In [25]:

```
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

------------------------------------Importing Dataset---

In [4]:

```
covid = pd.read_csv("C:/Users/DELL/Desktop/Atomcamp Python/COVID_cinical.csv"
```

1. Read Dataset and Explore the dataset by checking shape, columns, see the first/last 'n' rows using head/tail.

Top rows of the dataset (n= 5,15,30)

In [80]:
```
covid.head(5)          # Top 5 Rows
```

		0001...					norm
NCT04395482	3	Lung CT Scan Analysis of SARS-CoV2 Induced Lun...	TAC-COVID19	Recruiting	No Results Available	covid19	Othe (an CO
NCT04416061	4	The Role of a Private Hospital in Hong Kong Am...	COVID-19	Active, not recruiting	No Results Available	COVID	Dia Test: 19 Dia
NCT04395924	5	Maternal-foetal Transmission of SARS-Cov-2	TMF-COVID-19	Recruiting	No Results Available	Maternal Fetal Infection Transmission|COVID-19...	Dia Diag SAR

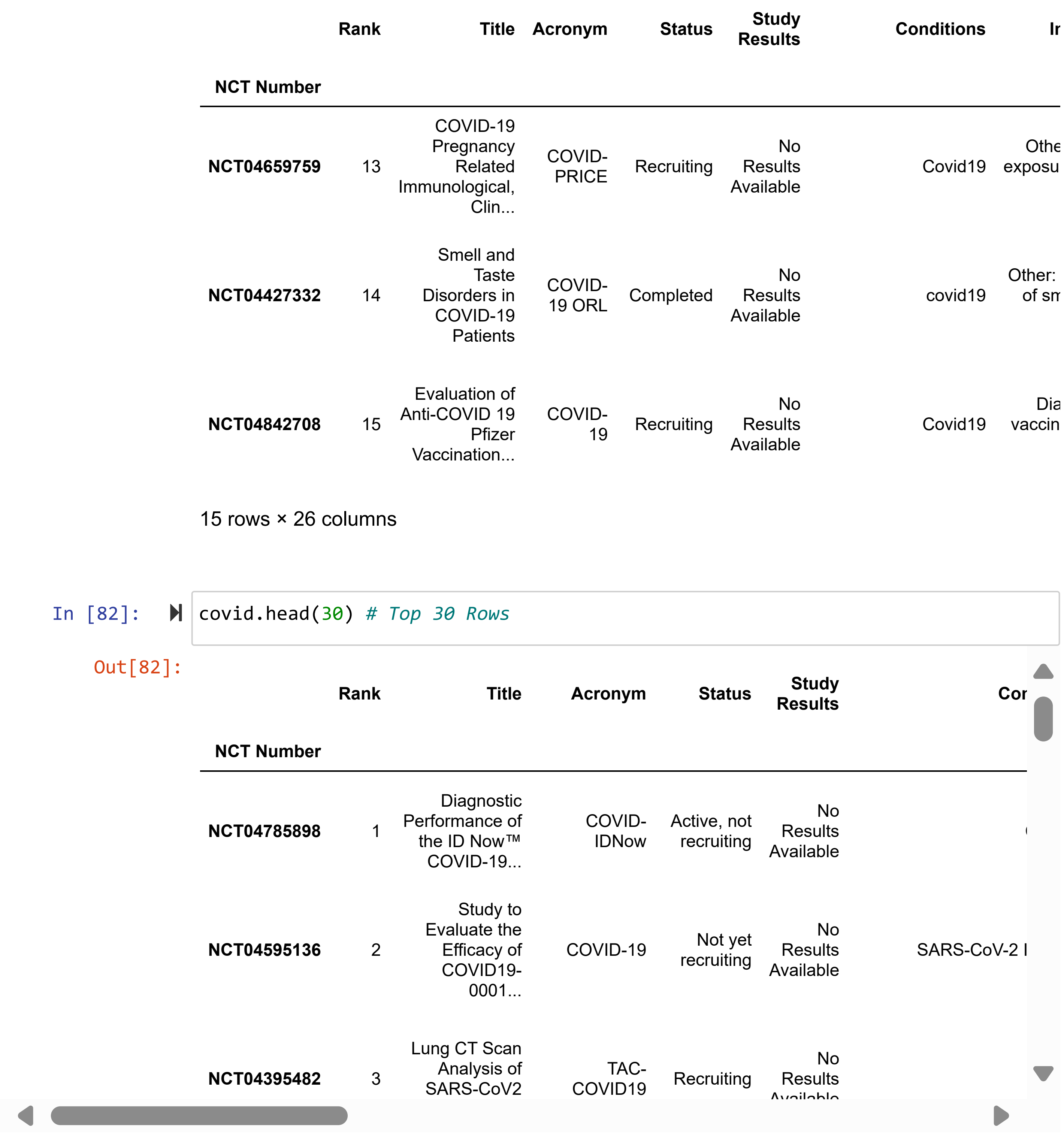
5 rows × 26 columns

In [81]:

```
covid.head(15)   # Top 15 rows
```

Out[81]:

NCT Number	Rank	Title	Acronym	Status	Study Results	Conditions	Ir
NCT04785898	1	Diagnostic Performance of the ID Now™ COVID-19...	COVID-IDNow	Active, not recruiting	No Results Available	Covid19	Diagn Now Sc
NCT04595136	2	Study to Evaluate the Efficacy of COVID19-0001...	COVID-19	Not yet recruiting	No Results Available	SARS-CoV-2 Infection	CO USR|
NCT04395482	3	Lung CT Scan Analysis of SARS-CoV2 Induced Lun...	TAC-COVID19	Recruiting	No Results Available	covid19	Otl sca COVI
NCT04416061	4	The Role of a Private Hospital in Hong Kong Am...	COVID-19	Active, not recruiting	No Results Available	COVID	Dia Di
NCT04395924	5	Maternal-foetal Transmission of SARS-Cov-2	TMF-COVID-19	Recruiting	No Results Available	Maternal Fetal Infection Transmission|COVID-19...	Dia Diagno C
NCT04516954	6	Convalescent Plasma for COVID-19 Patients	CPCP	Enrolling by invitation	No Results Available	COVID 19	C COVI
NCT04476940	7	COVID-19 Breastfeeding Guideline for African-A...	COVID-BF	Not yet recruiting	No Results Available	Covid19|Exclusive Breastfeeding	Behavi 19 E
NCT04634214	8	The Severity of COVID 19 in Diabetes and Non-d...	COVID19	Not yet recruiting	No Results Available	Covid19|Type2 Diabetes	
NCT04602884	9	Early Detection of COVID-19 Using Breath Analysis	COVID-19	Suspended	No Results Available	Covid19	Dia E samp
NCT04384588	10	COVID19-Convalescent Plasma for Treating Patie...	FALP-COVID	Recruiting	No Results Available	COVID-19 Infection|Cancer Patients|General Pop...	C
NCT04355897	11	CoVID-19 Plasma in Treatment of COVID-19 Patients	NaN	Recruiting	No Results Available	COVID 19	C COVI
NCT04412265	12	Frailty in Elderly Patients With COVID-19	FRA-COVID	Recruiting	No Results Available	Covid19	Ot betwe



NCT Number	Rank	Title	Acronym	Status	Study Results	Conditions	Ir
NCT04659759	13	COVID-19 Pregnancy Related Immunological, Clin...	COVID-PRICE	Recruiting	No Results Available	Covid19	Othe exposu
NCT04427332	14	Smell and Taste Disorders in COVID-19 Patients	COVID-19 ORL	Completed	No Results Available	covid19	Other: of sn
NCT04842708	15	Evaluation of Anti-COVID 19 Pfizer Vaccination...	COVID-19	Recruiting	No Results Available	Covid19	Dia vaccin

15 rows × 26 columns

```
In [82]: covid.head(30) # Top 30 Rows
```

Out[82]:

NCT Number	Rank	Title	Acronym	Status	Study Results	Cor
NCT04785898	1	Diagnostic Performance of the ID Now™ COVID-19...	COVID-IDNow	Active, not recruiting	No Results Available	
NCT04595136	2	Study to Evaluate the Efficacy of COVID19-0001...	COVID-19	Not yet recruiting	No Results Available	SARS-CoV-2 I
NCT04395482	3	Lung CT Scan Analysis of SARS-CoV2	TAC-COVID19	Recruiting	No Results Available	

Bottom rows of the dataset (n= 5,15,30)

```
In [87]:  covid.tail(5)    # Bottom 5 rows
```

Out[87]:

	Rank	Title	Acronym	Status	Study Results	Conditions	
NCT Number							
NCT04011644	5779	Mobile Health for Alcohol Use Disorders in Cli...	NaN	Recruiting	No Results Available	Alcohol Drinking|Telemedicine	Be mon
NCT04681339	5780	Antibiotic Prescription in Children Hospitaliz...	NaN	Not yet recruiting	No Results Available	Community Acquired Pneumonia in Children|Antib...	t
NCT04740229	5781	Moderate-intensity Flow-based Yoga Effects on ...	NaN	Recruiting	No Results Available	Stress|Psychological	
NCT04804917	5782	3-year Follow-up of the Mind My Mind RCT	MindMyMindFU	Recruiting	No Results Available	Emotional Problem|Anxiety Disorder of Childhoo...	
NCT04680000	5783	Chronic Pain Management In Primary Care Using ...	NaN	Not yet recruiting	No Results Available	Chronic Pain	C

5 rows × 26 columns

```
In [88]:  covid.tail(15)   # Bottom 15 rows
```

Out[88]:

	Rank	Title	Acronym	Status	Study Results	
NCT Number						
NCT04734795	5769	The Prevalence of Dysfunctional Breathing in C...	NaN	Recruiting	No Results Available	Dysfunctio
NCT04190368	5770	Team Clinic: Virtual Expansion of an Innovativ...	NaN	Not yet recruiting	No Results Available	
NCT03392883	5771	Scaling Up Science-based Mental Health Interve...	DIADA	Active, not recruiting	No Results Available	Depression|F
NCT04301518	5772	Prematurity Risk Assessment Combined With Clin...	PRIME	Recruiting	No Results Available	Prete
NCT04607902	5773	Harnessing Network Science to Personalize Scal...	NaN	Recruiting	No Results Available	
NCT04639661	5774	Predictors of Periodontal Outcomes Post-sanati...	NaN	Enrolling by invitation	No Results Available	Periodontal Disea
NCT04180709	5775	CBT to Reduce Insomnia and Improve Social Reco...	CRISP	Recruiting	No Results Available	Psychotic Dis
NCT04335643	5776	Telehealth CBT for Adolescents and Young Adult...	cSLE	Recruiting	No Results Available	Systemi
NCT04589377	5777	Mindfulness to Mitigate Psychological Threat a...	NaN	Recruiting	No Results Available	
NCT04574466	5778	Scaling-up Psychological Interventions With Sy...	NaN	Recruiting	No Results Available	Distress|PTSD|Anxiety|Depr
NCT04011644	5779	Mobile Health for Alcohol Use Disorders in Cli...	NaN	Recruiting	No Results Available	Alcoho
NCT04681339	5780	Antibiotic Prescription in Children Hospitaliz...	NaN	Not yet recruiting	No Results Available	Community

	Rank	Title	Acronym	Status	Study Results	
NCT Number						
NCT04740229	5781	Moderate-intensity Flow-based Yoga Effects on ...	NaN	Recruiting	No Results Available	
NCT04804917	5782	3-year Follow-up of the Mind My Mind RCT	MindMyMindFU	Recruiting	No Results Available	Emotional Problem|Anxiety
NCT04680000	5783	Chronic Pain Management In Primary Care Using ...	NaN	Not yet recruiting	No Results Available	

15 rows × 26 columns

In [89]:

```
covid.tail(30)  # Bottom 30 rows
```

NCT04734795	5769	The Prevalence of Dysfunctional Breathing in C...	NaN	Recruiting	No Results Available	Dysfu
NCT04190368	5770	Team Clinic: Virtual Expansion of an Innovativ...	NaN	Not yet recruiting	No Results Available	
NCT03392883	5771	Scaling Up Science-based Mental Health Interve...	DIADA	Active, not recruiting	No Results Available	Depres
NCT04301518	5772	Prematurity Risk Assessment	PRIME	Recruiting	No Results	

To get columns with their title in the dataset

In [5]:

```
covid.columns
```

Out[5]:

```
Index(['Rank', 'NCT Number', 'Title', 'Acronym', 'Status', 'Study Results',
       'Conditions', 'Interventions', 'Outcome Measures',
       'Sponsor/Collaborators', 'Gender', 'Age', 'Phases', 'Enrollment',
       'Funded Bys', 'Study Type', 'Study Designs', 'Other IDs', 'Start Dat
e',
       'Primary Completion Date', 'Completion Date', 'First Posted',
       'Results First Posted', 'Last Update Posted', 'Locations',
       'Study Documents', 'URL'],
      dtype='object')
```

To check total number of columns and rows in the dataset

In [6]:

```
covid.shape
```

Out[6]:

```
(5783, 27)
```

To get filimar with dataset, Wholestic overview of the data

In [8]:

```
covid.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5783 entries, 0 to 5782
Data columns (total 27 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Rank                     5783 non-null   int64  
 1   NCT Number               5783 non-null   object 
 2   Title                    5783 non-null   object 
 3   Acronym                  2480 non-null   object 
 4   Status                   5783 non-null   object 
 5   Study Results            5783 non-null   object 
 6   Conditions               5783 non-null   object 
 7   Interventions            4897 non-null   object 
 8   Outcome Measures         5748 non-null   object 
 9   Sponsor/Collaborators    5783 non-null   object 
 10  Gender                   5773 non-null   object 
 11  Age                      5783 non-null   object 
 12  Phases                   3322 non-null   object 
 13  Enrollment               5749 non-null   float64
 14  Funded Bys               5783 non-null   object 
 15  Study Type               5783 non-null   object 
 16  Study Designs            5748 non-null   object 
 17  Other IDs                5782 non-null   object 
 18  Start Date               5749 non-null   object 
 19  Primary Completion Date  5747 non-null   object 
 20  Completion Date          5747 non-null   object 
 21  First Posted             5783 non-null   object 
 22  Results First Posted     36 non-null     object 
 23  Last Update Posted       5783 non-null   object 
 24  Locations                5198 non-null   object 
 25  Study Documents          182 non-null    object 
 26  URL                      5783 non-null   object 
dtypes: float64(1), int64(1), object(25)
memory usage: 1.2+ MB
```

To get the summary statictics of the numerical data

In [13]:

```
covid.describe().T
```

Out[13]:

	count	mean	std	min	25%	50%	75%	max
Rank	5783.0	2892.000000	1669.552635	1.0	1446.5	2892.0	4337.5	5783.0
Enrollment	5749.0	18319.488607	404543.728784	0.0	60.0	170.0	560.0	20000000.0

Enrollment mean value is 18319.48860. The Maximum enrollment rate in 20000000.0. So, comparatively average rate in minimal.

To get the summary statictics of the non-numerical data

In [85]:

```
covid.describe(include=[object])
```

Out[85]:

	Title	Acronym	Status	Study Results	Conditions	Interventions	Outcome Measures	Sponso
count	5783	2480	5783	5783	5783	4897	5748	
unique	5775	2338	12	2	3067	4337	5687	
top	Study Assessing Vagus Nerve Stimulation in CoV...	COVID-19	Recruiting	No Results Available	COVID-19	Other: No intervention	Mortality	Assist H
freq	2	47	2805	5747	720	32	5	

4 rows × 24 columns

-------------Select all columns for the first clinical trial in the dataset.-----------------

In [5]:
```
x = covid.loc[0]
df_1 = pd.DataFrame(x)

df_1
```

Out[5]:
```
                                                                          0
Rank                                                                      1
NCT Number                                                      NCT04785898
Title                     Diagnostic Performance of the ID Now™ COVID-19...
Acronym                                                       COVID-IDNow
Status                                                Active, not recruiting
Study Results                                          No Results Available
Conditions                                                          Covid19
Interventions             Diagnostic Test: ID Now™ COVID-19 Screening Test
Outcome Measures               Evaluate the diagnostic performance of the ID ...
Sponsor/Collaborators                 Groupe Hospitalier Paris Saint Joseph
Gender                                                                  All
Age                                 18 Years and older   (Adult, Older Adult)
Phases                                                       Not Applicable
Enrollment                                                           1000.0
Funded Bys                                                            Other
Study Type                                                   Interventional
Study Designs             Allocation: N/A|Intervention Model: Single Gro...
Other IDs                                                       COVID-IDNow
Start Date                                                 November 9, 2020
Primary Completion Date                                   December 22, 2020
Completion Date                                              April 30, 2021
First Posted                                                  March 8, 2021
Results First Posted                                                    NaN
Last Update Posted                                            March 8, 2021
Locations                 Groupe Hospitalier Paris Saint-Joseph, Paris, ...
Study Documents                                                         NaN
URL                           https://ClinicalTrials.gov/show/NCT04785898
```

Setting a custom indexing

In [76]:
```
covid.set_index('NCT Number', inplace=True)
```

Retrieve the Title and Status of the clinical trial with the NCT Number 'NCT04595136'.

In [78]:

```
x = covid.loc['NCT04595136', ['Title', 'Status']]
df2 = pd.DataFrame(x)
df2
```

Out[78]:

	NCT04595136
Title	Study to Evaluate the Efficacy of COVID19-0001...
Status	Not yet recruiting

Get the Sponsor/Collaborators and Start Date for clinical trials that are Recruiting.

In [70]:

```
covid[covid['Status'] == 'Recruiting'][['Sponsor/Collaborators','Start Date']
```

Out[70]:

	Sponsor/Collaborators	Start Date
2	University of Milano Bicocca	May 7, 2020
4	Centre Hospitalier Régional d'Orléans|Centre d...	May 5, 2020
9	Fundacion Arturo Lopez Perez|Confederación de ...	April 7, 2020
10	The Christ Hospital	April 28, 2020
11	University of Milano Bicocca	April 16, 2020
...	...	...
5776	University of Pittsburgh|U.S. National Science...	October 26, 2020
5777	University of Zurich	August 25, 2020
5778	University of Wisconsin, Madison|National Inst...	March 23, 2020
5780	University of Illinois at Urbana-Champaign	February 10, 2021
5781	Mental Health Services in the Capital Region, ...	March 22, 2021

2805 rows × 2 columns

Select the first 5 rows and columns Title, Conditions, and Outcome Measures.

In [63]:

```
covid.iloc[0:5, [2,6,8]]
```

Out[63]:

	Title	Conditions	Outcome Measures
0	Diagnostic Performance of the ID Now™ COVID-19...	Covid19	Evaluate the diagnostic performance of the ID ...
1	Study to Evaluate the Efficacy of COVID19-0001...	SARS-CoV-2 Infection	Change on viral load results from baseline aft...
2	Lung CT Scan Analysis of SARS-CoV2 Induced Lun...	covid19	A qualitative analysis of parenchymal lung dam...
3	The Role of a Private Hospital in Hong Kong Am...	COVID	Proportion of asymptomatic subjects|Proportion...
4	Maternal-foetal Transmission of SARS-Cov-2	Maternal Fetal Infection Transmission|COVID-19...	COVID-19 by positive PCR in cord blood and / o...

-Find the Completion Date and URL for the last 3 clinical trials in the dataset.-

In [60]:

```
covid.iloc[-3:,[20,26]]
```

Out[60]:

	Completion Date	URL
5780	July 2021	https://ClinicalTrials.gov/show/NCT04740229
5781	December 31, 2022	https://ClinicalTrials.gov/show/NCT04804917
5782	February 2025	https://ClinicalTrials.gov/show/NCT04680000

-----Determine the missing values in the whole dataset and analyze missing values in each column.---

In [14]:

```
covid.isnull().sum()
```

Out[14]:

```
Rank                          0
NCT Number                    0
Title                         0
Acronym                    3303
Status                        0
Study Results                 0
Conditions                    0
Interventions               886
Outcome Measures             35
Sponsor/Collaborators         0
Gender                       10
Age                           0
Phases                     2461
Enrollment                   34
Funded Bys                    0
Study Type                    0
Study Designs                35
Other IDs                     1
Start Date                   34
Primary Completion Date      36
Completion Date              36
First Posted                  0
Results First Posted       5747
Last Update Posted            0
Locations                   585
Study Documents            5601
URL                           0
dtype: int64
```

-------------------------Calculate the sum of duplicate rows---------------------------------

```
In [15]:  covid.duplicated().sum()

Out[15]:  0
```

------------Solve following question by using conditional statements-----------------

Mean of the Enrollment

```
In [56]:  covid["Enrollment"].mean()

Out[56]:  18319.48860671421
```

1:How many studies have an enrollment greater than a certain threshold?

```
In [19]:  mean = covid["Enrollment"].mean()
          num_studies = (covid["Enrollment"] > mean).sum()
          print(f"Number of studies with enrollment rate higher than the mean: {num_stu

          Number of studies with enrollment rate higher than the mean: 142
```

2:How many clinical trials have 'No Results Available'?

```
In [32]:  no_rst_count = (covid['Study Results'] == 'No Results Available').sum()
          print(f'Number of Study Results: {no_rst_count}')

          Number of Study Results: 5747
```

3:How many clinical trials are in an "Completed" and "Recruiting" status?

```
In [44]:  completed_count = (covid['Status'] == "Completed").sum()
          recruiting_count = (covid['Status'] == "Recruiting").sum()
          print(f'Number of Clinical Trials (Completed): {completed_count}')
          print(f'Number of Clinical Trials (Recruiting): {recruiting_count}')

          Number of Clinical Trials (Completed): 1025
          Number of Clinical Trials (Recruiting): 2805
```

4:How many clinical trials are related to 'COVID-19'?

In [51]:

```
covid_cases = (covid['Conditions'] == 'Covid 19').sum()
print(f'Number of Covid 19 Cases: {covid_cases}')
```

```
Number of Covid 19 Cases: 9
```

5:How many clinical trials started after January 1, 2020?

In [54]:

```
strt_date = (covid['Start Date'] == 'January 1, 2020').sum()
print(f'Number of days with stating date in Jan: {strt_date}')
```

```
Number of days with stating date in Jan: 22
```

1. Distribution of Study Statuses

In [10]:

```
study_statuses = covid['Status'].value_counts()
```

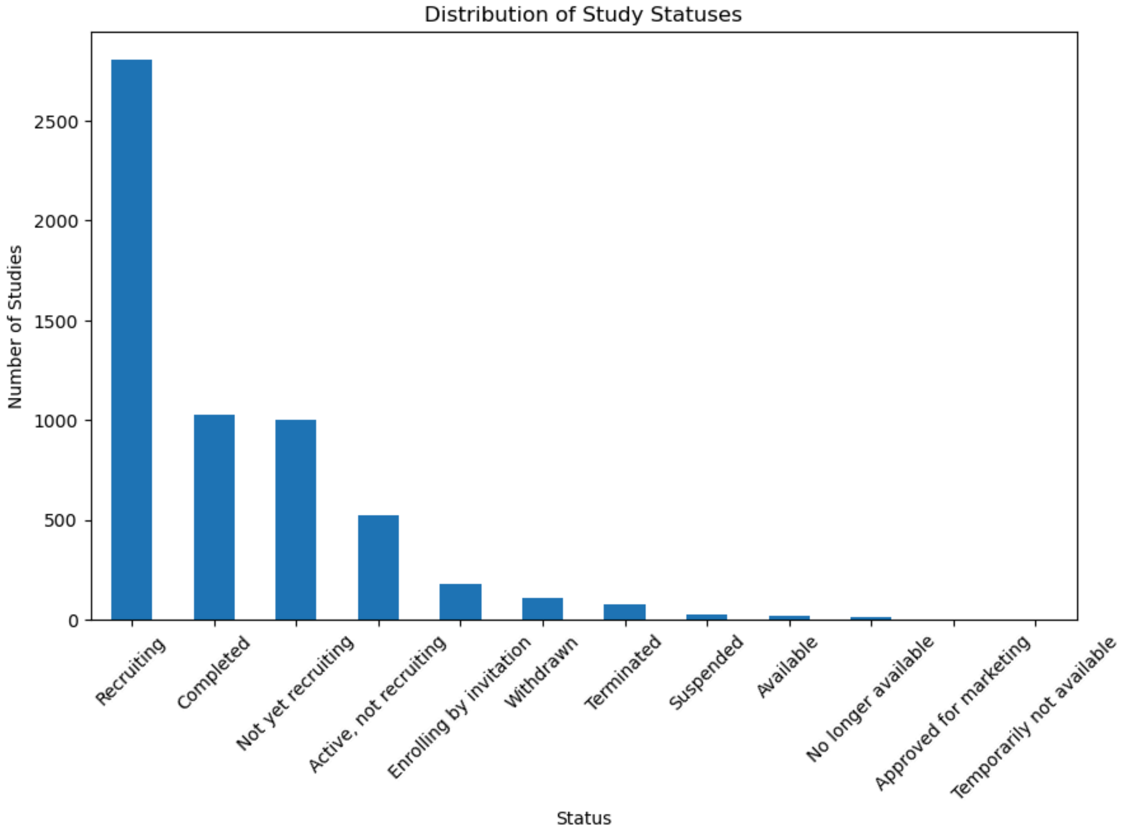
In [11]:

```
print(study_statuses)
```

```
Status
Recruiting                   2805
Completed                    1025
Not yet recruiting           1004
Active, not recruiting        526
Enrolling by invitation       181
Withdrawn                     107
Terminated                     74
Suspended                      27
Available                      19
No longer available            12
Approved for marketing          2
Temporarily not available       1
Name: count, dtype: int64
```

In [19]:

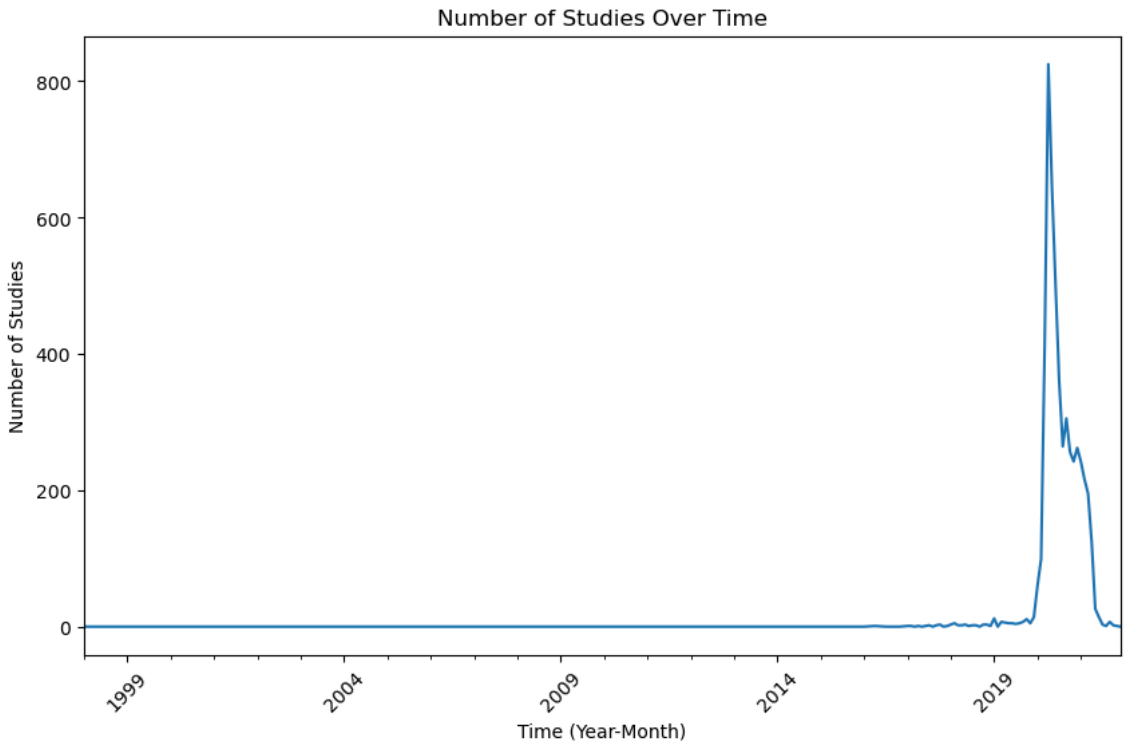
```
plt.figure(figsize=(10, 6))
study_statuses.plot(kind='bar')
plt.title('Distribution of Study Statuses')
plt.xlabel('Status')
plt.ylabel('Number of Studies')
plt.xticks(rotation=45)
plt.show()
```



2. Number of Studies Over Time

In [20]:

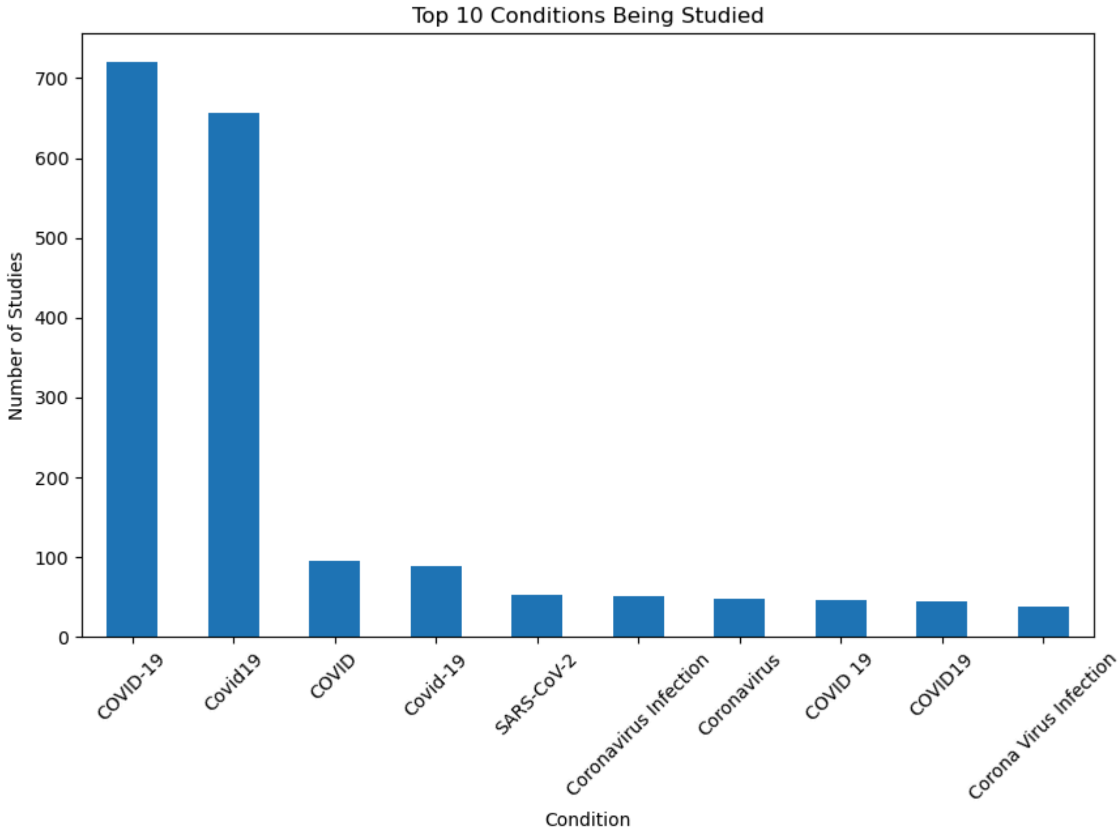
```
covid['Start Date'] = pd.to_datetime(covid['Start Date'], errors='coerce')
covid['YearMonth'] = covid['Start Date'].dt.to_period('M')
studies_over_time = covid['YearMonth'].value_counts().sort_index()
plt.figure(figsize=(10, 6))
studies_over_time.plot(kind='line')
plt.title('Number of Studies Over Time')
plt.xlabel('Time (Year-Month)')
plt.ylabel('Number of Studies')
plt.xticks(rotation=45)
plt.show()
```



3. Top Conditions Studied

In [21]:

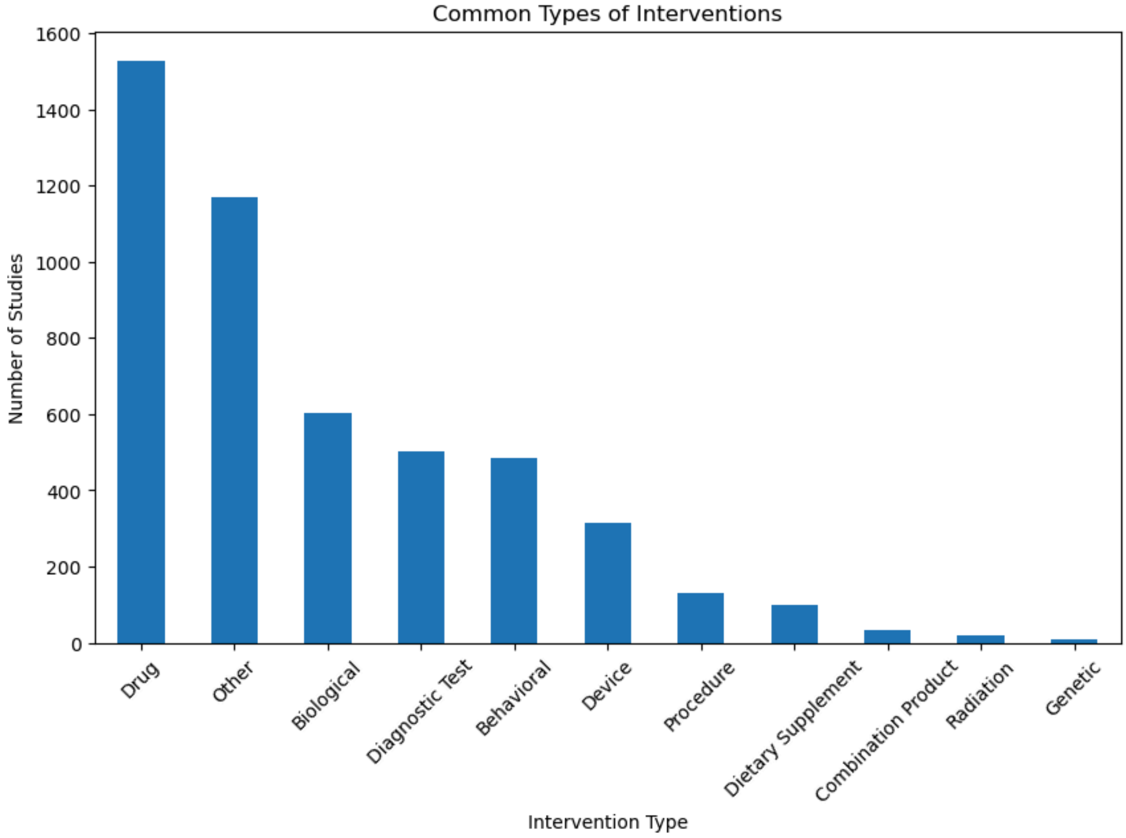
```
top_conditions = covid['Conditions'].value_counts().head(10)
plt.figure(figsize=(10, 6))
top_conditions.plot(kind='bar')
plt.title('Top 10 Conditions Being Studied')
plt.xlabel('Condition')
plt.ylabel('Number of Studies')
plt.xticks(rotation=45)
plt.show()
```



4. Intervention Types Analysis

In [23]:

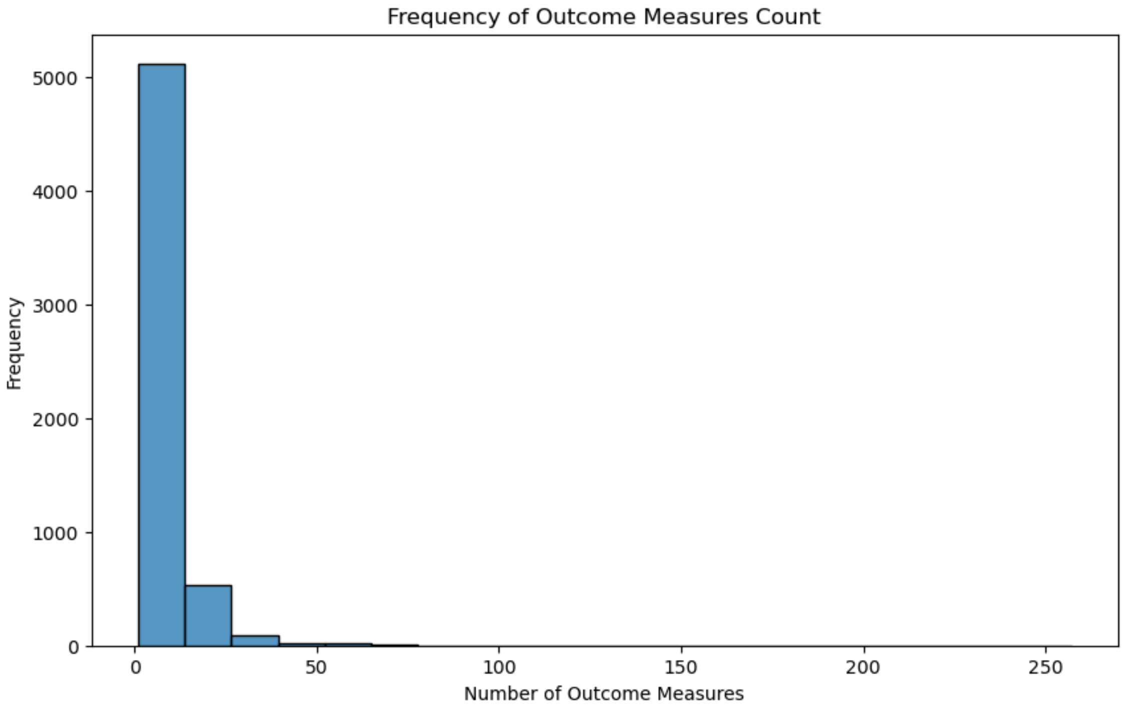
```
covid['Intervention Type'] = covid['Interventions'].str.split(':').str[0]
intervention_types = covid['Intervention Type'].value_counts()
plt.figure(figsize=(10, 6))
intervention_types.plot(kind='bar')
plt.title('Common Types of Interventions')
plt.xlabel('Intervention Type')
plt.ylabel('Number of Studies')
plt.xticks(rotation=45)
plt.show()
```



5. Outcome Measures Frequency

In [27]:

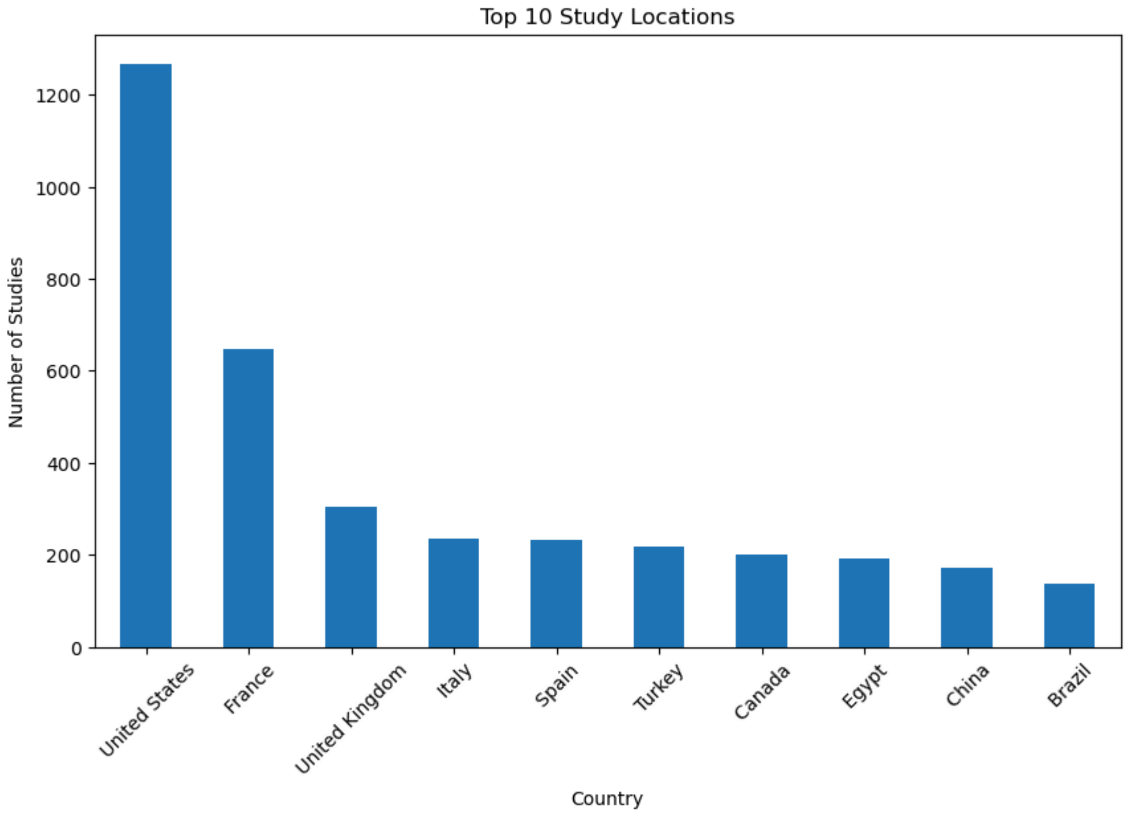
```
covid['Outcome Measures Count'] = covid['Outcome Measures'].apply(lambda x: l
plt.figure(figsize=(10, 6))
sns.histplot(covid['Outcome Measures Count'], bins=20)
plt.title('Frequency of Outcome Measures Count')
plt.xlabel('Number of Outcome Measures')
plt.ylabel('Frequency')
plt.show()
```



6. Study Locations Distribution

In [28]:

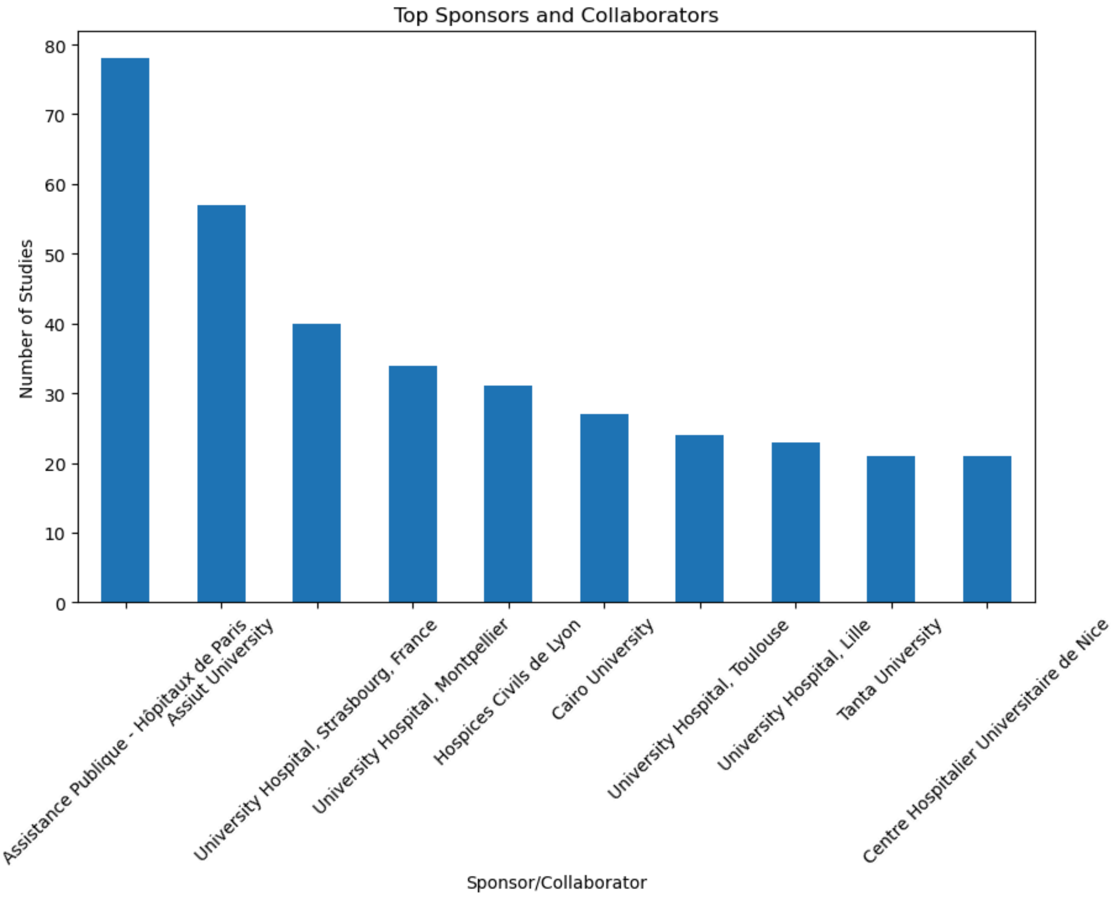
```
covid['Country'] = covid['Locations'].str.split(',').str[-1]
top_countries = covid['Country'].value_counts().head(10)
plt.figure(figsize=(10, 6))
top_countries.plot(kind='bar')
plt.title('Top 10 Study Locations')
plt.xlabel('Country')
plt.ylabel('Number of Studies')
plt.xticks(rotation=45)
plt.show()
```



7. Study Sponsors and Collaborators Analysis

In [30]:

```
top_sponsors = covid['Sponsor/Collaborators'].value_counts().head(10)
plt.figure(figsize=(10, 6))
top_sponsors.plot(kind='bar')
plt.title('Top Sponsors and Collaborators')
plt.xlabel('Sponsor/Collaborator')
plt.ylabel('Number of Studies')
plt.xticks(rotation=45)
plt.show()
```



--THE END--

In [ ]:

```

```