Project Title

Exploratory Data Analysis of COVID-19 Clinical Trials

1.Objective

The objective of this project is to perform an in-depth Exploratory Data Analysis (EDA) on the COVID-19 Clinical Trials dataset to uncover key trends, patterns, and insights related to global clinical research conducted during the COVID-19 pandemic.

2.Importing Libraries

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

3.Initial Data Exploration

In [7]:
```
data=pd.read_csv('COVID clinical trials.csv')
```

In [9]:
```
data
```

Out[9]:

	Rank	NCT Number	Title	Acronym	Status	Study Results	Conditions	Interventions	Outcome Measures	Sponsor/Collaborators	...	Other IDs	Start Date	Pr Comp
0	1	NCT04785898	Diagnostic Performance of the ID Now™ COVID-19...	COVID-IDNow	Active, not recruiting	No Results Available	Covid19	Diagnostic Test: ID Now™ COVID-19 Screening Test	Evaluate the diagnostic performance of the ID ...	Groupe Hospitalier Paris Saint Joseph	...	COVID-IDNow	November 9, 2020	Dece 22,
1	2	NCT04595136	Study to Evaluate the Efficacy of COVID19-0001...	COVID-19	Not yet recruiting	No Results Available	SARS-CoV-2 Infection	Drug: Drug COVID19-0001-USR|Drug: normal saline	Change on viral load results from baseline aft...	United Medical Specialties	...	COVID19-0001-USR	November 2, 2020	Dece 15,
2	3	NCT04395482	Lung CT Scan Analysis of SARS-CoV2 Induced Lun...	TAC-COVID19	Recruiting	No Results Available	covid19	Other: Lung CT scan analysis in COVID-19 patients	A qualitative analysis of parenchymal lung dam...	University of Milano Bicocca	...	TAC-COVID19	May 7, 2020	Jun
3	4	NCT04416061	The Role of a Private Hospital in Hong Kong Am...	COVID-19	Active, not recruiting	No Results Available	COVID	Diagnostic Test: COVID 19 Diagnostic Test	Proportion of asymptomatic subjects|Proportion...	Hong Kong Sanatorium & Hospital	...	RC-2020-08	May 25, 2020	Ju
4	5	NCT04395924	Maternal-foetal Transmission of SARS-Cov-2	TMF-COVID-19	Recruiting	No Results Available	Maternal Fetal Infection Transmission|COVID-19...	Diagnostic Test: Diagnosis of SARS-Cov2 by RT-...	COVID-19 by positive PCR in cord blood and / o...	Centre Hospitalier Régional d'Orléans|Centre d...	...	CHRO-2020-10	May 5, 2020	May
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5778	5779	NCT04011644	Mobile Health for Alcohol Use Disorders in Cli...	NaN	Recruiting	No Results Available	Alcohol Drinking|Telemedicine	Behavioral: A-CHESS self-monitored|Behavioral:...	Number of risky drinking days|Number of patien...	University of Wisconsin, Madison|National Inst...	...	2019-0337|R01AA024150|A532007|SMPH/FAMILY MED/...	March 23, 2020	A
5779	5780	NCT04681339	Antibiotic Prescription in Children Hospitaliz...	NaN	Not yet recruiting	No Results Available	Community Acquired Pneumonia in Children|Antib...	Other: Antibiotic treatment|Other: No antibiot...	Antibiotic treatment rates in hospitalized chi...	ARCIM Institute Academic Research in Complemen...	...	PKA-03	April 2021	Nov
5780	5781	NCT04740229	Moderate-intensity Flow-based Yoga Effects on ...	NaN	Recruiting	No Results Available	Stress|Psychological	Behavioral: Yoga	Perceived Stress|Task switching paradigm|Digit...	University of Illinois at Urbana-Champaign	...	21584	February 10, 2021	July
5781	5782	NCT04804917	3-year Follow-up of the Mind My Mind RCT	MindMyMindFU	Recruiting	No Results Available	Emotional Problem|Anxiety Disorder of Childhoo...	NaN	The child's impact of mental health problems r...	Mental Health Services in the Capital Region, ...	...	MHSCRDenmark, F-61502-03-1	March 22, 2021	M
5782	5783	NCT04680000	Chronic Pain Management In Primary Care Using ...	NaN	Not yet recruiting	No Results Available	Chronic Pain	Behavioral: Brief Cognitive Behavioral Therapy...	Defense and Veterans Pain Rating Scale (DVPRS)...	The University of Texas Health Science Center ...	...	HSC20200520H	February 2021	Fel

5783 rows × 27 columns

In [11]:
```
data.head()
```

Out[11]:

	Rank	NCT Number	Title	Acronym	Status	Study Results	Conditions	Interventions	Outcome Measures	Sponsor/Collaborators	...	Other IDs	Start Date	Primary Completion Date	Completion Date	First Posted	Results First Posted	Last Update Posted
0	1	NCT04785898	Diagnostic Performance of the ID Now™ COVID-19...	COVID-IDNow	Active, not recruiting	No Results Available	Covid19	Diagnostic Test: ID Now™ COVID-19 Screening Test	Evaluate the diagnostic performance of the ID ...	Groupe Hospitalier Paris Saint Joseph	...	COVID-IDNow	November 9, 2020	December 22, 2020	April 30, 2021	March 8, 2021	NaN	March 8, 2021
1	2	NCT04595136	Study to Evaluate the Efficacy of COVID19-0001...	COVID-19	Not yet recruiting	No Results Available	SARS-CoV-2 Infection	Drug: Drug COVID19-0001-USR|Drug: normal saline	Change on viral load results from baseline aft...	United Medical Specialties	...	COVID19-0001-USR	November 2, 2020	December 15, 2020	January 29, 2021	October 20, 2020	NaN	October 20, 2020
2	3	NCT04395482	Lung CT Scan Analysis of SARS-CoV2 Induced Lun...	TAC-COVID19	Recruiting	No Results Available	covid19	Other: Lung CT scan analysis in COVID-19 patients	A qualitative analysis of parenchymal lung dam...	University of Milano Bicocca	...	TAC-COVID19	May 7, 2020	June 15, 2021	June 15, 2021	May 20, 2020	NaN	November 9, 2020
3	4	NCT04416061	The Role of a Private Hospital in Hong Kong Am...	COVID-19	Active, not recruiting	No Results Available	COVID	Diagnostic Test: COVID 19 Diagnostic Test	Proportion of asymptomatic subjects|Proportion...	Hong Kong Sanatorium & Hospital	...	RC-2020-08	May 25, 2020	July 31, 2020	August 31, 2020	June 4, 2020	NaN	June 4, 2020
4	5	NCT04395924	Maternal-foetal Transmission of SARS-Cov-2	TMF-COVID-19	Recruiting	No Results Available	Maternal Fetal Infection Transmission|COVID-19...	Diagnostic Test: Diagnosis of SARS-Cov2 by RT-...	COVID-19 by positive PCR in cord blood and / o...	Centre Hospitalier Régional d'Orléans|Centre d...	...	CHRO-2020-10	May 5, 2020	May 2021	May 2021	May 20, 2020	NaN	June 4, 2020

5 rows × 27 columns

In [13]:
```
data.describe()
```

Out[13]:

	Rank	Enrollment
count	5783.000000	5.749000e+03
mean	2892.000000	1.831949e+04
std	1669.552635	4.045437e+05
min	1.000000	0.000000e+00
25%	1446.500000	6.000000e+01
50%	2892.000000	1.700000e+02
75%	4337.500000	5.600000e+02
max	5783.000000	2.000000e+07

In [15]:
```
data.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
```

4.Handling missing values

In [17]:
```
data.isnull().sum()
```

Out[17]:
```
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
```

In [21]:
```
data.isnull()
```

Out[21]:

	Rank	NCT Number	Title	Acronym	Status	Study Results	Conditions	Interventions	Outcome Measures	Sponsor/Collaborators	...	Other IDs	Start Date	Primary Completion Date	Completion Date	First Posted	Results First Posted	Last Update Posted	Locations	Study Documents	URL
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	True	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	True	False
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5778	False	False	False	True	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	True	False
5779	False	False	False	True	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	True	False
5780	False	False	False	True	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	True	False
5781	False	False	False	False	False	False	False	True	False	False	...	False	False	False	False	False	True	False	False	True	False
5782	False	False	False	True	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	True	False

5783 rows × 27 columns

In [24]:
```
# Check missing values in the dataset
missing_values = data.isnull().sum().sort_values(ascending=False)
missing_percentage = (missing_values / len(data)) * 100
```

In [26]:
```
# Combine into a DataFrame for easier inspection
missing_data = pd.DataFrame({
    'Missing Values': missing_values,
    'Percentage': missing_percentage
})
```

In [28]:
```
# Display columns with missing values only
missing_data = missing_data[missing_data['Missing Values'] > 0]
```

In [30]:
```
missing_data
```

Out[30]:

	Missing Values	Percentage
Results First Posted	5747	99.377486
Study Documents	5601	96.852845
Acronym	3303	57.115684
Phases	2461	42.555767
Interventions	886	15.320768
Locations	585	10.115857
Completion Date	36	0.622514
Primary Completion Date	36	0.622514
Study Designs	35	0.605222
Outcome Measures	35	0.605222
Enrollment	34	0.587930
Start Date	34	0.587930
Gender	10	0.172921
Other IDs	1	0.017292

In [32]:
```
# Drop columns with excessive missing values
data_cleaned = data.drop(columns=['Results First Posted', 'Study Documents'])
```

In [34]:
```
# Fill missing values
fill_values = {
    'Acronym': 'Not Provided',
    'Phases': 'Not Specified',
    'Interventions': 'Not Provided',
    'Locations': 'Unknown'
}
```

In [36]:
```
# Fill specified columns
data_cleaned.fillna(value=fill_values, inplace=True)
```

In [44]:
```
# Fill remaining missing values:
# - For categorical: fill with 'Unknown'
# - For numerical: fill with median
for col in data_cleaned.columns:
    if data_cleaned[col].isnull().sum() > 0:
        if data_cleaned[col].dtype == 'object':
            data_cleaned[col].fillna('Unknown', inplace=True)
        else:
            data_cleaned[col].fillna(data_cleaned[col].median(), inplace=True)
            import warnings
warnings.filterwarnings('ignore')
```

In [50]:
```
# Check if any missing values remain
missing_after_cleaning = data_cleaned.isnull().sum().sum()
```

In [52]:
```
missing_after_cleaning
```

Out[52]: 0

5.Univariate Analysis

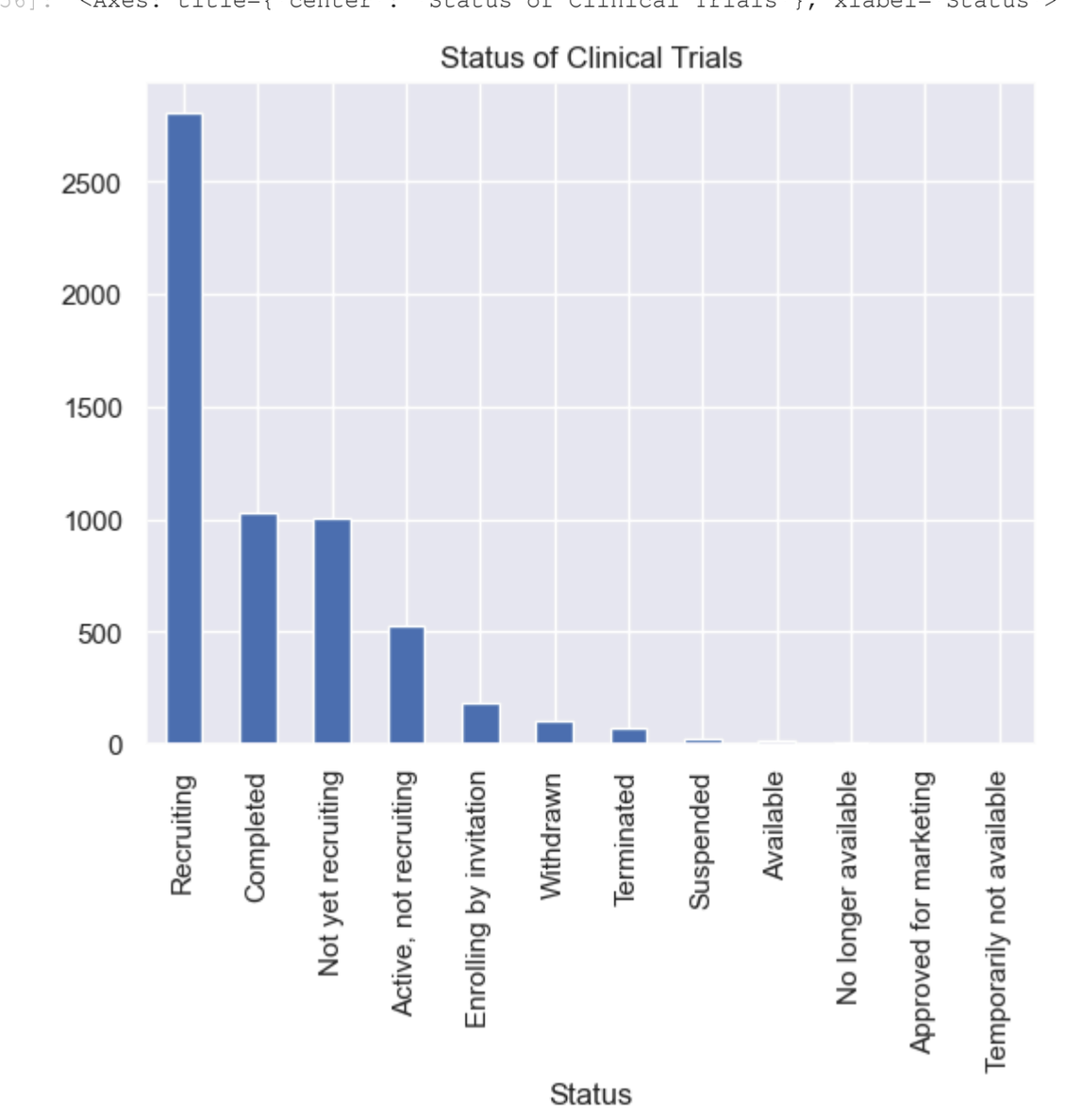
In [54]:
```
data['Status'].value_counts()
```

Out[54]:
```
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 [ ]:
```
# 1.Status Distribution
```

In [56]:
```
data['Status'].value_counts().plot(kind='bar', title='Status of Clinical Trials')
```

Out[56]: <Axes: title={'center': 'Status of Clinical Trials'}, xlabel='Status'>



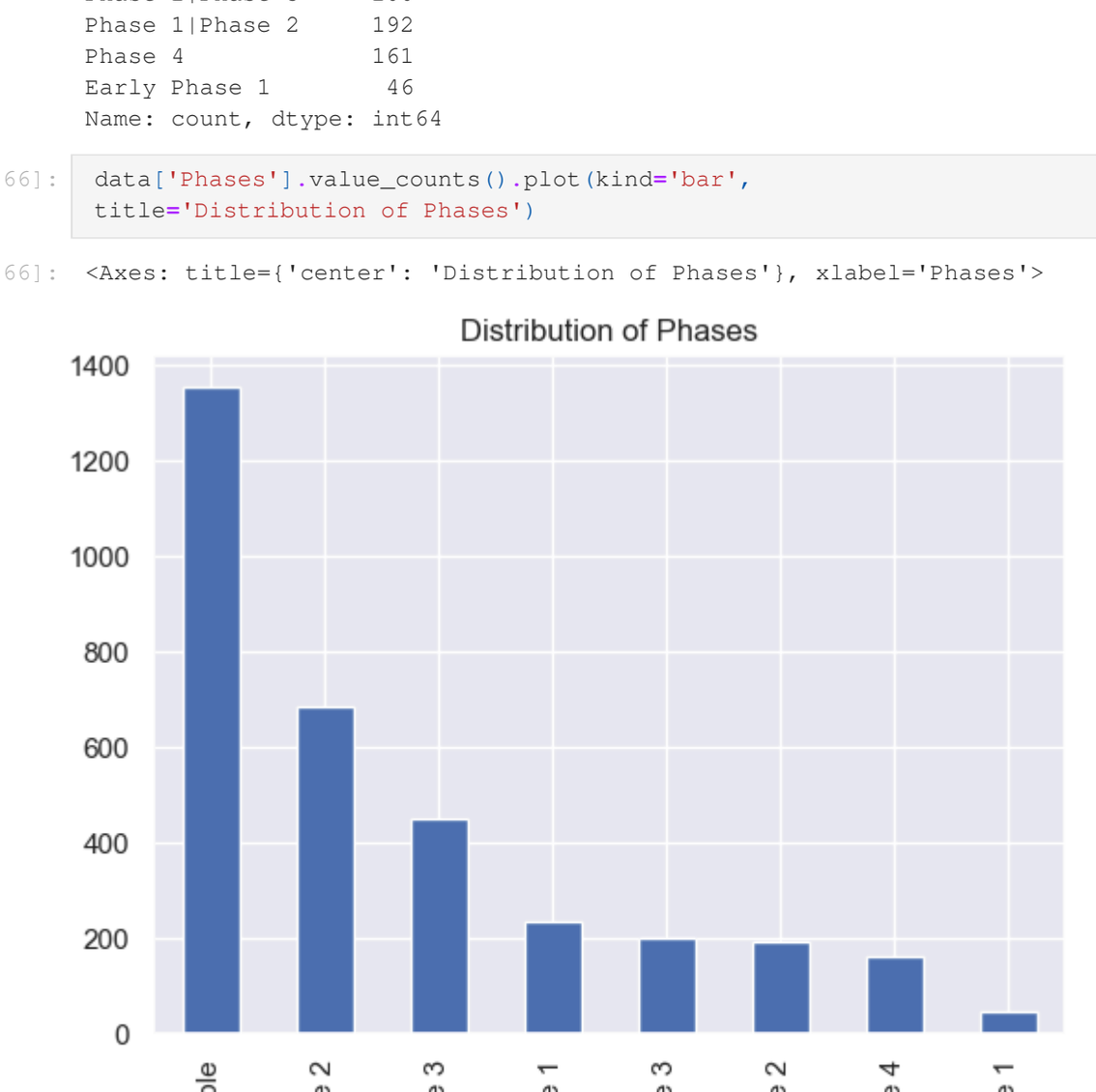
In [62]:
```
# 2.PhasecDistribution
```

In [64]:
```
data['Phases'].value_counts()
```

Out[64]:
```
Phases
Not Applicable     1354
Phase 2             685
Phase 3             450
Phase 1             234
Phase 2|Phase 3     200
Phase 1|Phase 2     192
Phase 4             161
Early Phase 1        46
Name: count, dtype: int64
```

In [66]:
```
data['Phases'].value_counts().plot(kind='bar',
title='Distribution of Phases')
```

Out[66]: <Axes: title={'center': 'Distribution of Phases'}, xlabel='Phases'>



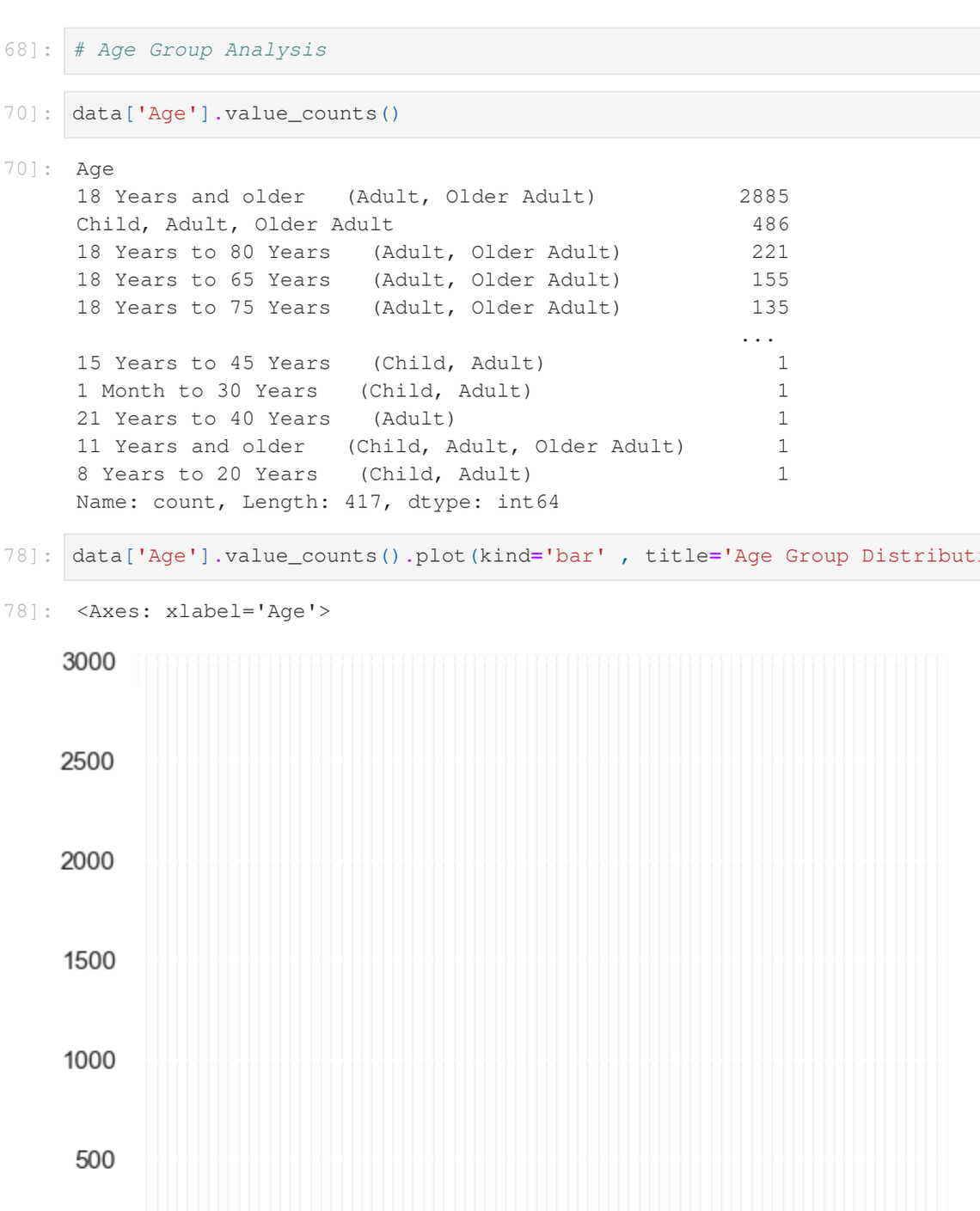
In [68]:
```
# Age Group Analysis
```

In [70]:
```
data['Age'].value_counts()
```

Out[70]:
```
Age
18 Years and older   (Adult, Older Adult)            2885
Child, Adult, Older Adult                             486
18 Years to 80 Years   (Adult, Older Adult)           221
18 Years to 65 Years   (Adult, Older Adult)           155
18 Years to 75 Years   (Adult, Older Adult)           135
                                                      ...
15 Years to 45 Years   (Child, Adult)                   1
1 Month to 30 Years   (Child, Adult)                    1
21 Years to 40 Years   (Adult)                          1
11 Years and older   (Child, Adult, Older Adult)        1
8 Years to 20 Years   (Child, Adult)                    1
Name: count, Length: 417, dtype: int64
```

In [78]:
```
data['Age'].value_counts().plot(kind='bar' , title='Age Group Distribution')
```

Out[78]: <Axes: xlabel='Age'>



6. Bivariate Analysis

In [81]:
```
#  Status vs Phases
```

In [83]:
```
status_phase = pd.crosstab(data['Status'], data['Phases'])
```

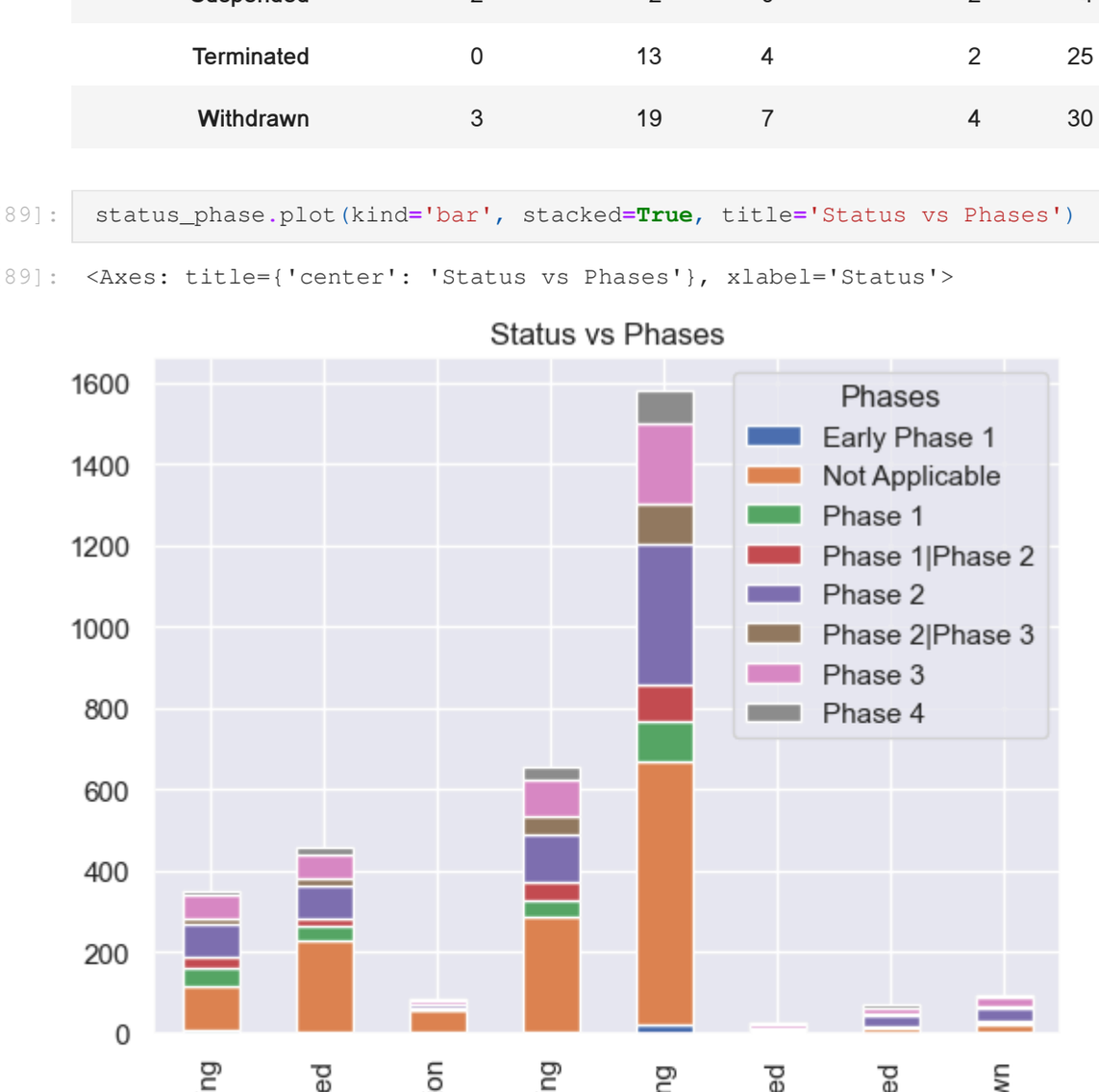
In [85]:
```
status_phase
```

Out[85]:

Phases / Status	Early Phase 1	Not Applicable	Phase 1	Phase 1|Phase 2	Phase 2	Phase 2|Phase 3	Phase 3	Phase 4
Active, not recruiting	7	111	44	26	81	15	59	8
Completed	3	226	38	17	78	20	56	22
Enrolling by invitation	4	54	1	3	10	1	6	6
Not yet recruiting	5	282	42	46	114	46	89	30
Recruiting	22	647	98	92	343	102	196	81
Suspended	2	2	0	2	4	4	9	2
Terminated	0	13	4	2	25	6	15	5
Withdrawn	3	19	7	4	30	6	20	7

In [89]:
```
status_phase.plot(kind='bar', stacked=True, title='Status vs Phases')
```

Out[89]: <Axes: title={'center': 'Status vs Phases'}, xlabel='Status'>



In [91]:
```
#  Conditions vs. Outcome Measures
```

In [123…
```
conditions_outcomes = data.groupby('Conditions')['Outcome Measures'].apply(lambda x: ', ').reset_index()
```

In [125…
```
conditions_outcomes
```

Out[125…

	Conditions	Outcome Measures
0	2019 Novel Coronavirus	,
1	2019 Novel Coronavirus Infection	,
2	2019 Novel Coronavirus Infection|COVID-19 Viru...	,
3	2019 Novel Coronavirus Pneumonia	,
4	2019 Novel Coronavirus Pneumonia|COVID-19	,
...	...	...
3062	the Lung Complication of COVID-19	,
3063	the Prognostic Value of Ferritin|Glycosylated ...	,
3064	the Study Focus on the Uses of Telephone and O...	,
3065	the Use of Modern Technology Applications in H...	,
3066	to Predict an Unfavorable Evolution of Covid-1...	,

3067 rows × 2 columns

7.Time Series Analysis

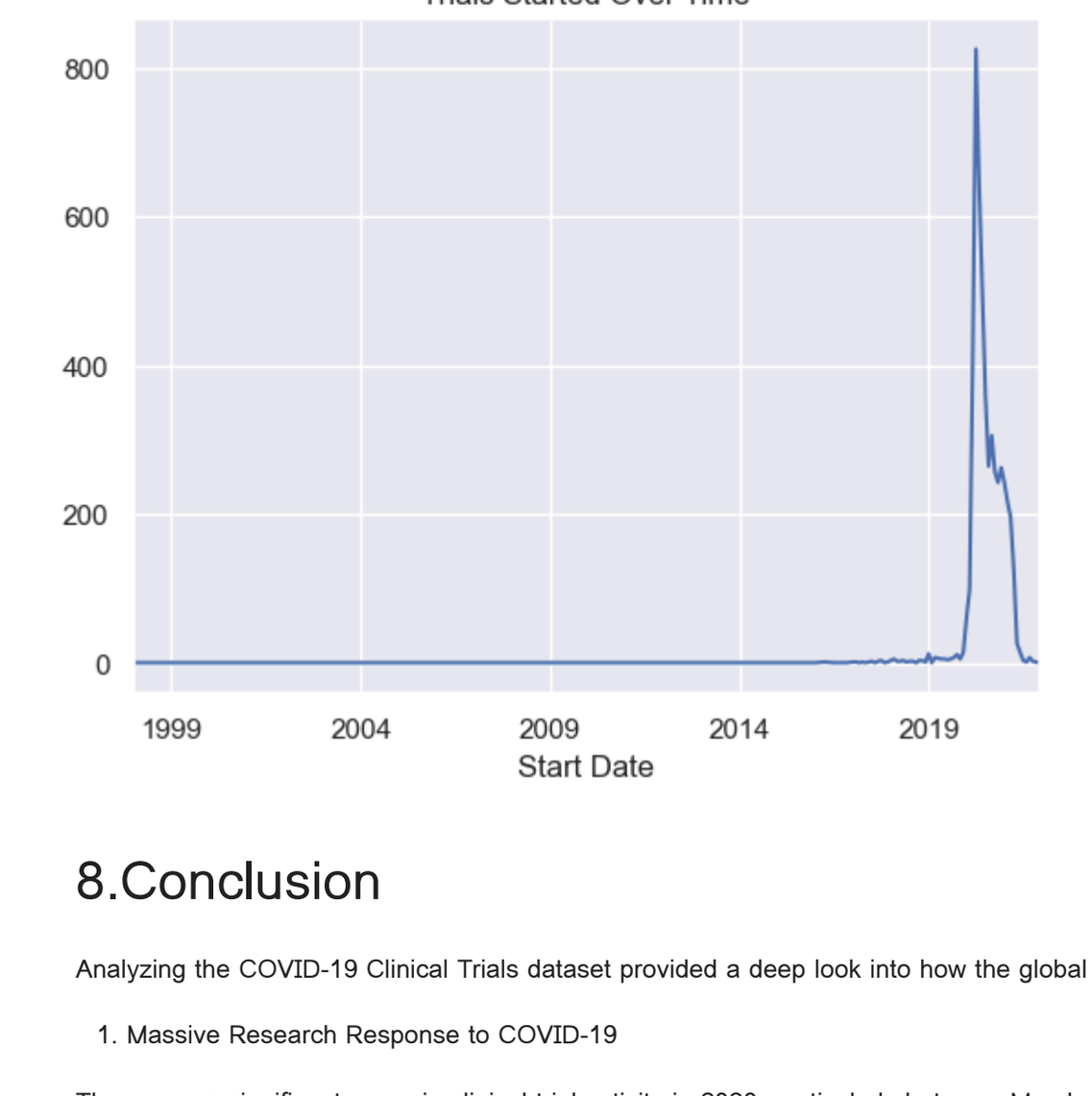
In [130…
```
# Convert date columns to datetime
data['Start Date'] = pd.to_datetime(data['Start Date'], errors='coerce')
```

In [142…
```
# Convert date columns to datetime
data['Start Date'] = pd.to_datetime(data['Start Date'], errors='coerce')
```

In [144…
```
data['Primary Completion Date'] = pd.to_datetime(data['Primary Completion Date'], errors='coerce')
```

In [148…
```
# Plot the number of trials started over time
data['Start Date'].dt.to_period('M').value_counts().sort_index().plot(kind='line', title='Trials Started Over Time')
```

Out[148… <Axes: title={'center': 'Trials Started Over Time'}, xlabel='Start Date'>



8.Conclusion

Analyzing the COVID-19 Clinical Trials dataset provided a deep look into how the global research community responded during the pandemic. Here are the key takeaways from the analysis:

1. Massive Research Response to COVID-19

There was a significant surge in clinical trial activity in 2020, particularly between March and July. This reflects the urgency with which medical and scientific institutions worldwide began testing treatments, vaccines, and diagnostics in response to the rapid spread of the virus.

2. Majority of Trials Were Interventional

Most of the trials in the dataset were interventional studies, meaning they involved actively testing drugs, devices, or procedures. Observational studies were also present, but in smaller numbers.

3. Widespread Missing Results

More than 99% of trials did not post any results in the dataset. This indicates either a lag in publication, incomplete studies, or issues with result reporting. It highlights a major gap in data transparency and public access to trial outcomes.

4. Data Quality Issues

The dataset had missing or inconsistently filled fields in critical columns like:

Phases (missing in ~42% of trials)

Interventions

Locations

Study Documents

This required careful cleaning and suggests the need for more standardized data reporting, especially during emergencies.

5. Uneven Distribution Across Sponsors and Locations

A few large institutions sponsored a large number of trials, while others were more scattered. Some trials didn’t list any location, and many were concentrated in the US, Europe, and parts of Asia.

6. Broad Range of Enrollment Sizes

Enrollment numbers varied widely—from small-scale studies with fewer than 10 participants to large trials involving tens of thousands. This shows both early-stage exploratory research and large, pivotal trials were conducted simultaneously.

In [ ]: