UNIFIED MENTOR
YOUR SKILL, SUCCESS & JOURNEY

Project Title	**COVID-19 Clinical Trials EDA Pandas**
Tools	Python, ML
Domain	Data Analyst & Data scientist
Project Difficulties Level	intermediate

Dataset: The Dataset is available in the given link. You can download it at your convenience.

Click here to download the data set

Dataset Description: ClinicalTrials.gov is a publicly accessible database of clinical studies worldwide, maintained by the National Institutes of Health. It offers a direct download feature, making clinical trial data easily available for analysis. This dataset includes COVID-19-related clinical trials, with each study stored as an XML file. The filename corresponds to the study's unique NCT number in the ClinicalTrials repository. A CSV file is also provided, containing key details but less information than the XML files.

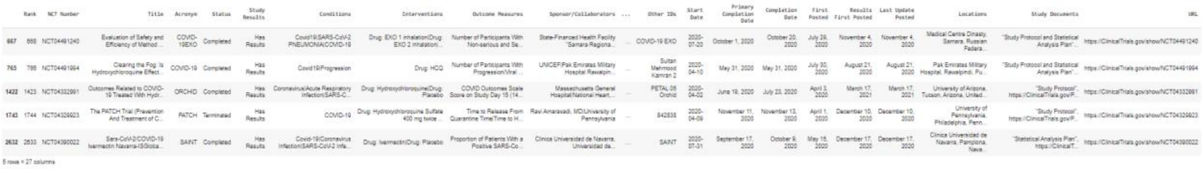
1. Importing Required Libraries:

pandas for data manipulation and analysis,numpy for numerical computations, seaborn and matplotlib.pyplot for data visualization.

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

2. Basic Data Exploration:

df.head() - Displays the first five rows of the dataset to get an initial look at the data.



	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
667	668	NCT04491240	Evaluation of Safety and Efficiency of Method ...	COVID-19EXO	Completed	Has Results	Covid19|SARS-CoV-2 PNEUMONIA|COVID-19	Drug: EXO 1 inhalation|Drug: EXO 2 inhalation|...	Number of Participants With Non-serious and Se...	State-Financed Health Facility "Samara Regiona...	...	COVID-19 EXO	2020-07-20	October 1, 2020	October 20, 2020	July 29, 2020	November 4, 2020	November 4, 2020	Medical Centre Dinasty, Samara, Russian Federa...	"Study Protocol and Statistical Analysis Plan",...	https://ClinicalTrials.gov/show/NCT04491240
765	766	NCT04491994	Clearing the Fog: Is Hydroxychloroquine Effect...	COVID-19	Completed	Has Results	Covid19|Progression	Drug: HCQ	Number of Participants With Progression|Viral ...	UNICEF|Pak Emirates Military Hospital Rawalpin...	...	Sultan Mehmood Kamran 2	2020-04-10	May 31, 2020	May 31, 2020	July 30, 2020	August 21, 2020	August 21, 2020	Pak Emirates Military Hospital, Rawalpindi, Pu...	"Study Protocol and Statistical Analysis Plan",...	https://ClinicalTrials.gov/show/NCT04491994
1422	1423	NCT04332991	Outcomes Related to COVID-19 Treated With Hydr...	ORCHID	Completed	Has Results	Coronavirus|Acute Respiratory Infection|SARS-C...	Drug: Hydroxychloroquine|Drug: Placebo	COVID Outcomes Scale Score on Study Day 15 (14...	Massachusetts General Hospital|National Heart,...	...	PETAL 05 Orchid	2020-04-02	June 19, 2020	July 23, 2020	April 3, 2020	March 17, 2021	March 17, 2021	University of Arizona, Tucson, Arizona, United...	"Study Protocol", https://ClinicalTrials.gov/P...	https://ClinicalTrials.gov/show/NCT04332991
1743	1744	NCT04329923	The PATCH Trial (Prevention And Treatment of C...	PATCH	Terminated	Has Results	COVID-19	Drug: Hydroxychloroquine Sulfate 400 mg twice ...	Time to Release From Quarantine Time|Time to H...	Ravi Amaravadi, MD|University of Pennsylvania	...	842838	2020-04-09	November 11, 2020	November 13, 2020	April 1, 2020	December 10, 2020	December 10, 2020	University of Pennsylvania, Philadelphia, Penn...	"Study Protocol", https://ClinicalTrials.gov/P...	https://ClinicalTrials.gov/show/NCT04329923
2632	2633	NCT04390022	Sars-CoV-2/COVID-19 Ivermectin Navarra-ISGloba...	SAINT	Completed	Has Results	Covid-19|Coronavirus Infection|SARS-CoV-2 Infe...	Drug: Ivermectin|Drug: Placebo	Proportion of Patients With a Positive SARS-Co...	Clinica Universidad de Navarra, Universidad de...	...	SAINT	2020-07-31	September 17, 2020	October 9, 2020	May 15, 2020	December 17, 2020	December 17, 2020	Clinica Universidad de Navarra, Pamplona, Nava...	"Statistical Analysis Plan", https://ClinicalT...	https://ClinicalTrials.gov/show/NCT04390022

5 rows × 27 columns

df.columns - Lists all column names to understand available features.

```
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 Date',
       'Primary Completion Date', 'Completion Date', 'First Posted',
       'Results First Posted', 'Last Update Posted', 'Locations',
       'Study Documents', 'URL'],
      dtype='object')
```

df.shape - Shows the number of rows and columns.

```
(13, 27)
```

df.info() - Provides a summary of data types, non-null counts, and memory usage.

```
<class 'pandas.core.frame.DataFrame'>
Index: 13 entries, 667 to 5737
Data columns (total 27 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Rank                     13 non-null     int64 
 1   NCT Number               13 non-null     object
 2   Title                    13 non-null     object
 3   Acronym                  13 non-null     object
 4   Status                   13 non-null     object
 5   Study Results            13 non-null     object
 6   Conditions               13 non-null     object
 7   Interventions            13 non-null     object
 8   Outcome Measures         13 non-null     object
 9   Sponsor/Collaborators    13 non-null     object
 10  Gender                   13 non-null     object
 11  Age                      13 non-null     object
 12  Phases                   13 non-null     object
 13  Enrollment              13 non-null     float64
 14  Funded Bys              13 non-null     object
 15  Study Type              13 non-null     object
 16  Study Designs           13 non-null     object
 17  Other IDs               13 non-null     object
 18  Start Date              13 non-null     datetime64[ns]
 19  Primary Completion Date  13 non-null     object
 20  Completion Date         13 non-null     object
 21  First Posted            13 non-null     object
 22  Results First Posted     13 non-null     object
 23  Last Update Posted       13 non-null     object
 24  Locations               13 non-null     object
 25  Study Documents         13 non-null     object
 26  URL                     13 non-null     object
dtypes: datetime64[ns](1), float64(1), int64(1), object(24)
memory usage: 2.8+ KB
```

3. Summary Statistics :

df.describe() - Computes summary statistics for numerical columns (mean, standard deviation, min, max, etc.).

```
            Rank  Enrollment                     Start Date
count    13.000000   13.000000                             13
mean   3565.076923  131.076923  2020-02-29 01:50:46.153846272
min     668.000000    2.000000            2018-03-07 00:00:00
25%    1744.000000   24.000000            2020-04-02 00:00:00
50%    3477.000000   30.000000            2020-04-10 00:00:00
75%    5643.000000  173.000000            2020-06-19 00:00:00
max    5738.000000  540.000000            2020-07-31 00:00:00
std    1985.355445  179.489768                            NaN
```

df.describe(include='object') - Provides summary statistics for categorical columns (unique values, most frequent values, etc.).

```
        NCT Number                                        Title  \
count           13                                           13
unique          13                                           13
top     NCT04491240  Evaluation of Safety and Efficiency of Method ...
freq             1                                            1

           Acronym     Status Study Results Conditions  \
count           13         13            13         13
unique          12          2             1          9
top     Favipiravir  Completed   Has Results   COVID-19
freq             2         12            13          4

                                     Interventions  \
count                                             13
unique                                            13
top     Drug: EXO 1 inhalation|Drug: EXO 2 inhalation|...
freq                                               1

                                  Outcome Measures  \
count                                             13
unique                                            13
top     Number of Participants With Non-serious and Se...
freq                                               1

                              Sponsor/Collaborators Gender  ...  \
count                                             13     13  ...
unique                                            13      2  ...
top     State-Financed Health Facility "Samara Regiona...    All  ...
freq                                               1     11  ...

                                     Study Designs     Other IDs  \
count                                             13            13
unique                                            10            13
top     Allocation: Randomized|Intervention Model: Par...  COVID-19 EXO
freq                                               3             1

       Primary Completion Date   Completion Date   First Posted  \
count                       13                13             13
unique                      13                13             13
top            October 1, 2020  October 20, 2020  July 29, 2020
freq                         1                 1              1

       Results First Posted Last Update Posted  \
count                    13                 13
unique                   13                 13
top        November 4, 2020   November 4, 2020
freq                      1                  1

                                         Locations  \
count                                             13
unique                                            12
top     Novagenix Drug R&D Center, Akyurt, Ankara, Tur...
freq                                               2

                                   Study Documents  \
count                                             13
unique                                            13
top     "Study Protocol and Statistical Analysis Plan"...
freq                                               1

                                        URL
count                                       13
unique                                      13
```

```
top     https://ClinicalTrials.gov/show/NCT04491240
freq                                          1

[4 rows x 24 columns]
```

4. Identifying Data Types :

df.select_dtypes(include='object').columns - Identifies all categorical columns.

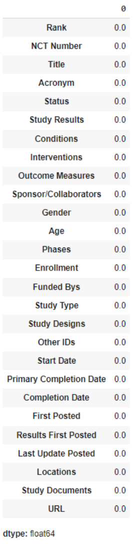
```
Index(['NCT Number', 'Title', 'Acronym', 'Status', 'Study Results',
       'Conditions', 'Interventions', 'Outcome Measures',
       'Sponsor/Collaborators', 'Gender', 'Age', 'Phases', 'Funded Bys',
       'Study Type', 'Study Designs', 'Other IDs', 'Primary Completion Date',
       'Completion Date',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

df.select_dtypes(exclude='object').columns - Identifies all numerical columns.

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

5. Checking for Missing Data: Calculates the percentage of missing values for each column to identify potential data quality issues.

```
missing_data = df.isnull().mean() * 100
missing_data
```



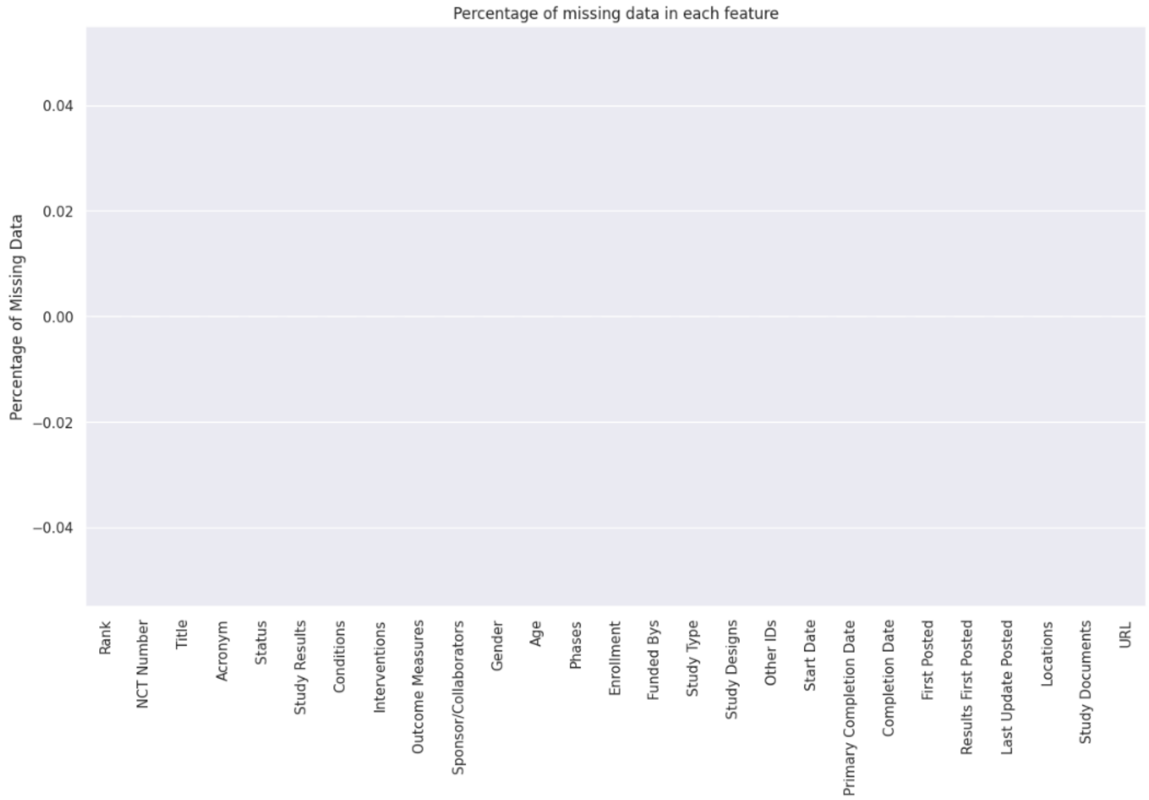
	0
Rank	0.0
NCT Number	0.0
Title	0.0
Acronym	0.0
Status	0.0
Study Results	0.0
Conditions	0.0
Interventions	0.0
Outcome Measures	0.0
Sponsor/Collaborators	0.0
Gender	0.0
Age	0.0
Phases	0.0
Enrollment	0.0
Funded Bys	0.0
Study Type	0.0
Study Designs	0.0
Other IDs	0.0
Start Date	0.0
Primary Completion Date	0.0
Completion Date	0.0
First Posted	0.0
Results First Posted	0.0
Last Update Posted	0.0
Locations	0.0
Study Documents	0.0
URL	0.0

dtype: float64

6. Visualizing Missing Data: Create a bar chart to visualize the percentage of missing data in each column, up to 40 columns.

```
def visualize_data(data , caption = '' , ylabel = 'Percentage of Missing Data'):
   sns.set(rc={'figure.figsize' : (15,8.27)}) # set figure size
   plt.xticks(rotation=90) # make ticks vertical
   fig = sns.barplot(x = data.keys()[ :min(40 , len(data))].tolist() , y = data.values[ : min(40 , len(data))].tolist()).set_title(caption) # set title to the image and plot it or the highest 40
   plt.ylabel(ylabel) # set labels
   plt.show()

visualize_data(missing_data , 'Percentage of missing data in each feature')
```



7. Checking Total Missing Values: This function prints the total number of missing values per column to identify potential issues.

```
print(df.isnull().sum())

Rank                    0
NCT Number              0
Title                   0
Acronym                 0
Status                  0
Study Results           0
Conditions              0
Interventions           0
Outcome Measures        0
```

```
Sponsor/Collaborators      0
Gender                     0
Age                        0
Phases                     0
Enrollment                 0
Funded Bys                 0
Study Type                 0
Study Designs              0
Other IDs                  0
Start Date                 0
Primary Completion Date    0
Completion Date            0
First Posted               0
Results First Posted       0
Last Update Posted         0
Locations                  0
Study Documents            0
URL                        0
dtype: int64
```

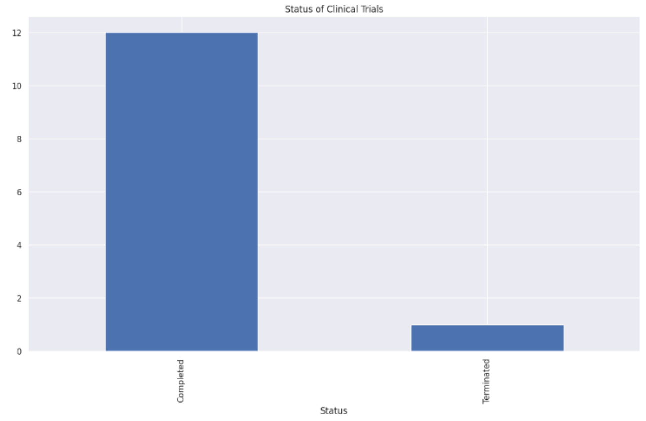
8. Dropping Unnecessary Columns: Removes the columns 'Acronym' and 'Study Documents' since they may not be useful for analysis.

```
df = df.drop(columns=['Acronym', 'Study Documents'])
```

9. Analyzing Clinical Trial Status: Counts and visualizes the distribution of different trial statuses.

```
print(df['Status'].value_counts())
df['Status'].value_counts().plot(kind='bar', title='Status of Clinical Trials')
```

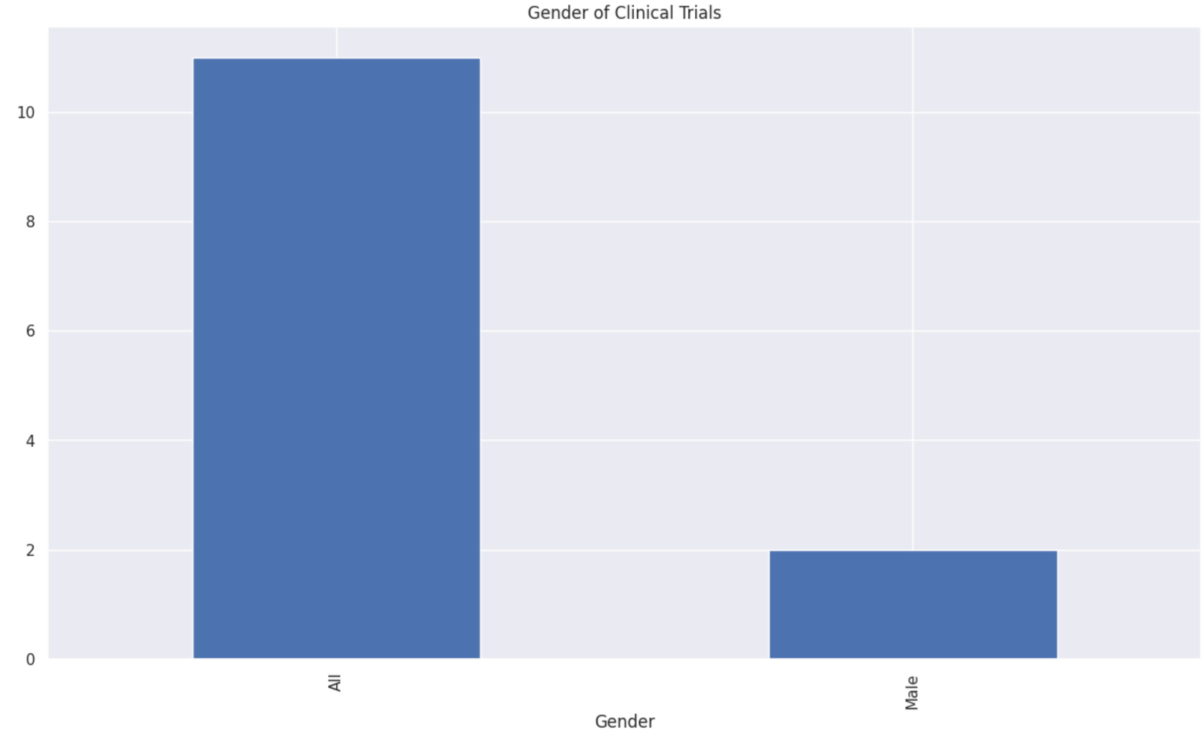
```
Status
Completed     12
Terminated     1
Name: count, dtype: int64
<Axes: title={'center': 'Status of Clinical Trials'}, xlabel='Status'>
```



10. Analyzing Gender Distribution in Trials: Displays and visualizes the number of trials based on gender distribution.

```
print(df['Gender'].value_counts())
df['Gender'].value_counts().plot(kind='bar', title='Gender of Clinical Trials')
```

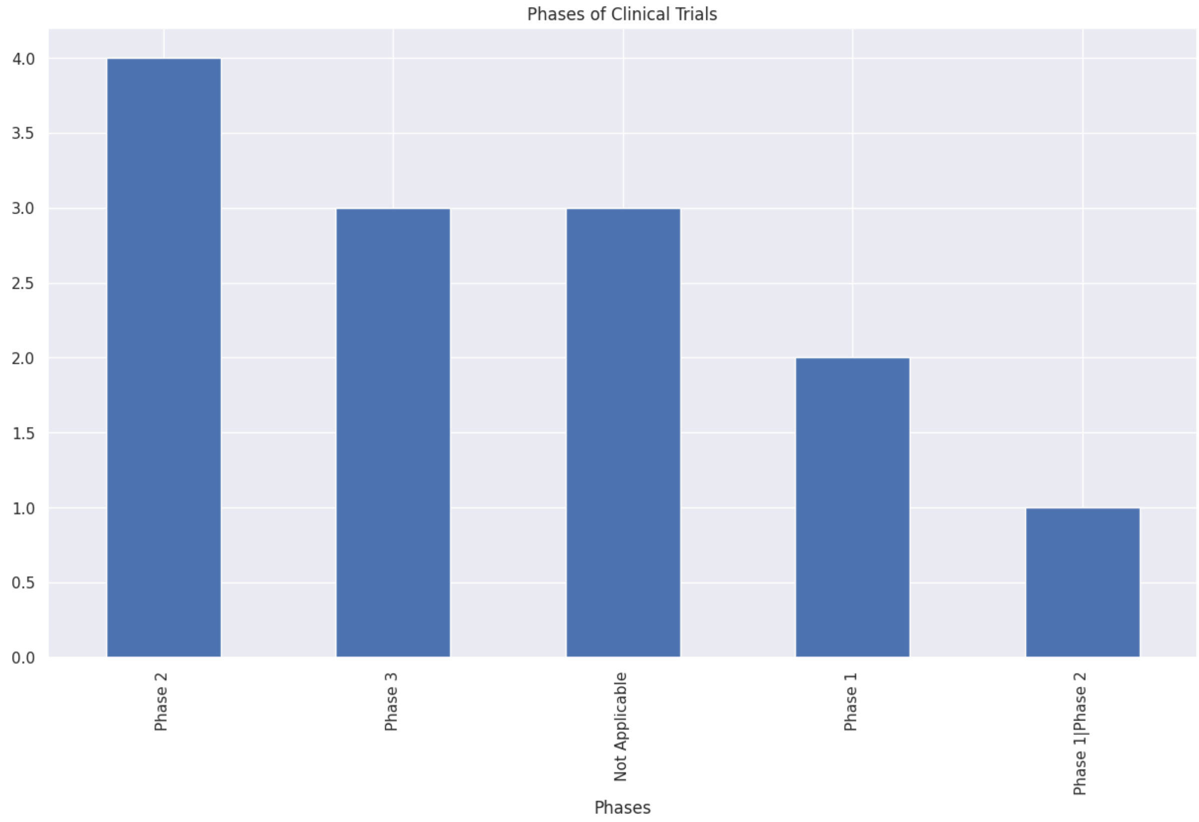
```
Gender
All     11
Male     2
Name: count, dtype: int64
<Axes: title={'center': 'Gender of Clinical Trials'}, xlabel='Gender'>
```



11. Analyzing Clinical Trial Phases: Shows the distribution of clinical trials across different phases.

```
print(df['Phases'].value_counts())
df['Phases'].value_counts().plot(kind='bar', title='Phases of Clinical Trials')
```

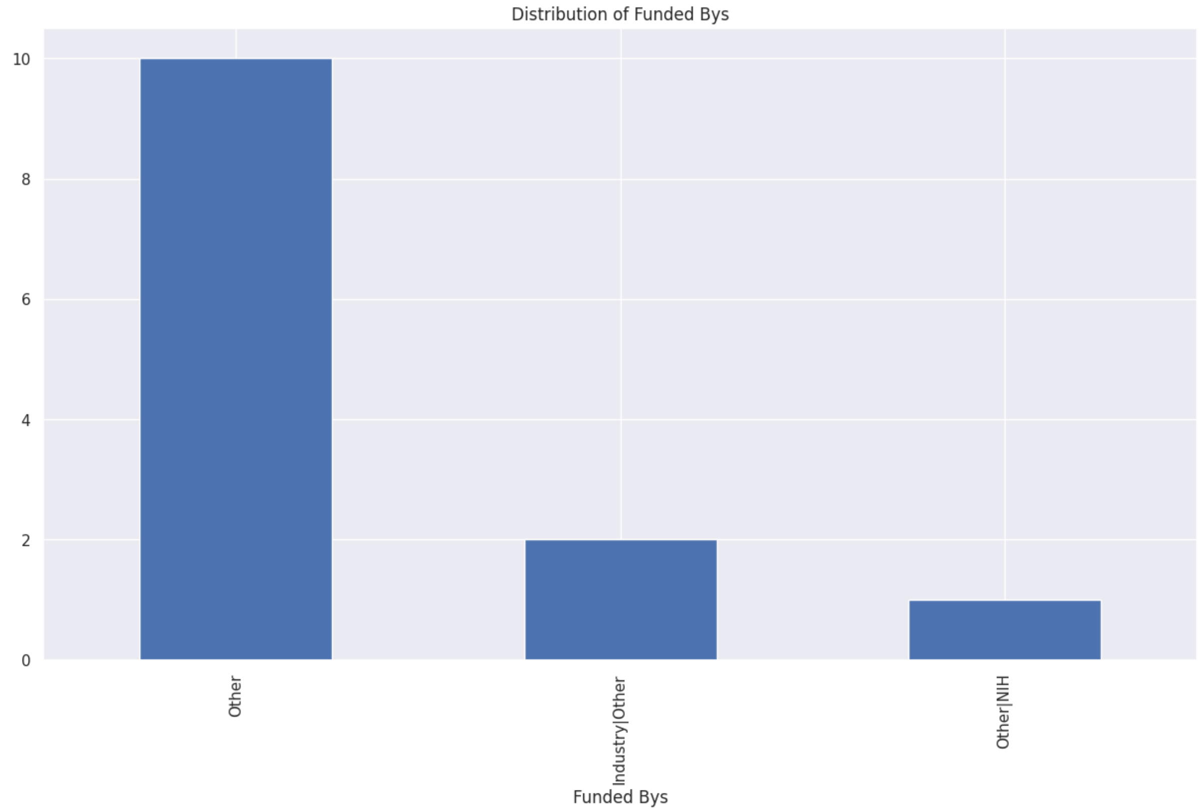
```
Phases
Phase 2            4
Phase 3            3
Not Applicable     3
Phase 1            2
Phase 1|Phase 2    1
Name: count, dtype: int64
<Axes: title={'center': 'Phases of Clinical Trials'}, xlabel='Phases'>
```



12. Analyzing Funding Sources: Identifies and visualizes the distribution of funding sources for clinical trials.

```
print(df['Funded Bys'].value_counts())
df['Funded Bys'].value_counts().plot(kind='bar', title='Distribution of Funded Bys')
```

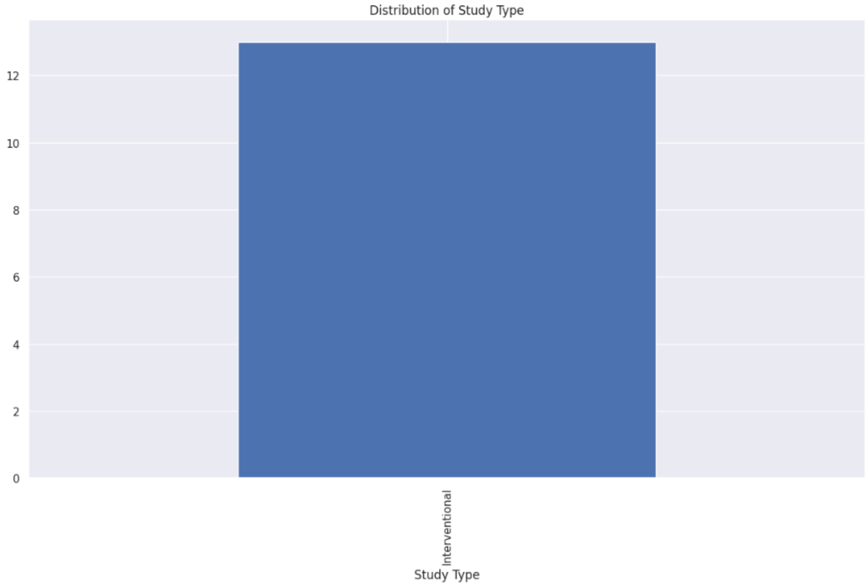
```
Funded Bys
Other            10
Industry|Other     2
Other|NIH          1
Name: count, dtype: int64
<Axes: title={'center': 'Distribution of Funded Bys'}, xlabel='Funded Bys'>
```



13. Analyzing Study Types: Displays the count and distribution of different study types in the dataset.

```
print(df['Study Type'].value_counts())
df['Study Type'].value_counts().plot(kind='bar', title='Distribution of Study Type')
```

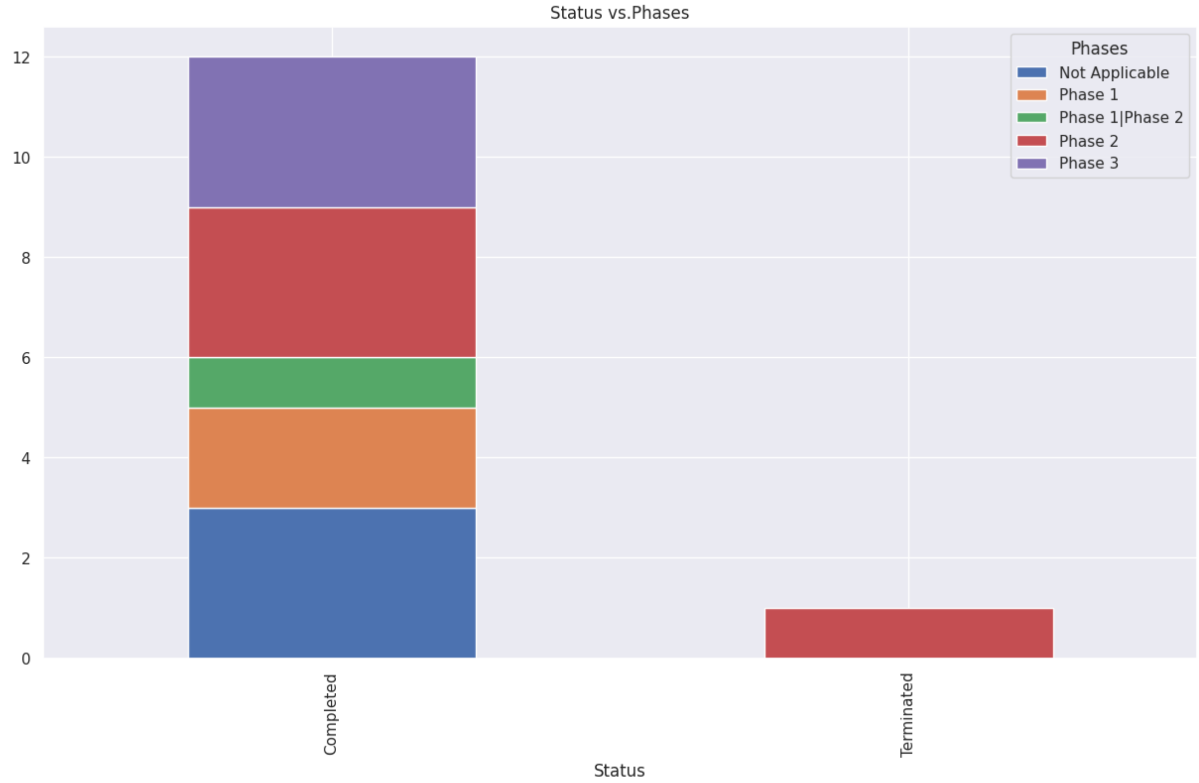
```
Study Type
Interventional    13
Name: count, dtype: int64
<Axes: title={'center': 'Distribution of Study Type'}, xlabel='Study Type'>
```



14. Analyzing Status vs. Phases Relationship: Creates a cross-tabulation of Status and Phases to analyze their relationship.Generates a stacked bar chart to visualize the trends.

```
status_phase = pd.crosstab(df['Status'], df['Phases'])
print(status_phase)
status_phase.plot(kind='bar', stacked=True, title='Status vs.Phases')
```

```
Phases      Not Applicable  Phase 1  Phase 1|Phase 2  Phase 2  Phase 3
Status
Completed                3        2                1        3        3
Terminated               0        0                0        1        0
<Axes: title={'center': 'Status vs.Phases'}, xlabel='Status'>
```

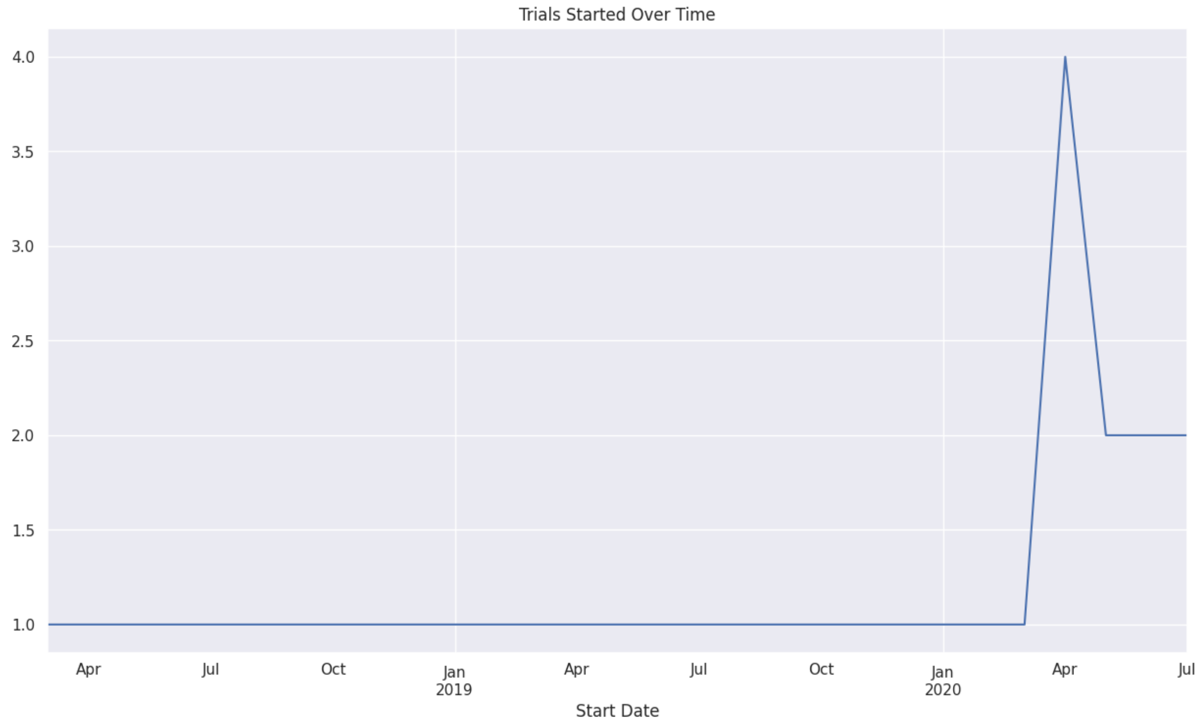


15. Converting Date Columns to Date Format: Converts Start Date and Primary Completion Date columns to datetime format for time-based analysis.

```
df['Start Date'] = pd.to_datetime(df['Start Date'], errors='coerce')
df['Primary Completion Date'] = pd.to_datetime(df['Primary Completion Date'],
errors='coerce')
```

16. Analyzing Clinical Trials Over Time: Groups clinical trials by month and visualizes their trend over time using a line chart.Helps in understanding how the number of clinical trials has changed over time.

```
df['Start Date'].dt.to_period('M').value_counts().sort_index().plot(kind='line', title='Trials Started Over Time')
```



Conclusion:

This analysis of COVID-19 clinical trials provides valuable insights into trial statuses, phases, funding sources, and study types. By identifying missing data and cleaning the dataset, we ensured reliable analysis. Visualizing trends over time helped in understanding trial progression. These findings can aid researchers and policymakers in making data-driven decisions to improve clinical trial efficiency and effectiveness.