COVID-19 Clinical Trials – Exploratory Data Analysis

OBJECTIVE

The objective of this project is to analyze global COVID-19 clinical trial data to identify patterns in trial status, phases, funding sources, geographic distribution, enrollment, and research trends.

Tools Used

Python

Pandas

Seaborn

Matplotlib

Google Colab

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

df = pd.read_csv("/content/drive/MyDrive/UNIFIED INTERNSHIP/COVID 19
CLINICAL TRIAL DATA/DATA/COVID clinical trials.csv")
df.head()

{"type":"dataframe","variable_name":"df"}

df.shape

(5783, 27)

df.columns

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.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
```

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

```
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
```

```
df.duplicated().sum()
```

```
np.int64(0)
```

```
df = df.drop_duplicates()
```

```
date_cols = [
    'Start Date',
    'Primary Completion Date',
    'Completion Date',
    'First Posted',
    'Last Update Posted'
]

for col in date_cols:
    df[col] = pd.to_datetime(df[col], errors='coerce')
```

```
df.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
```

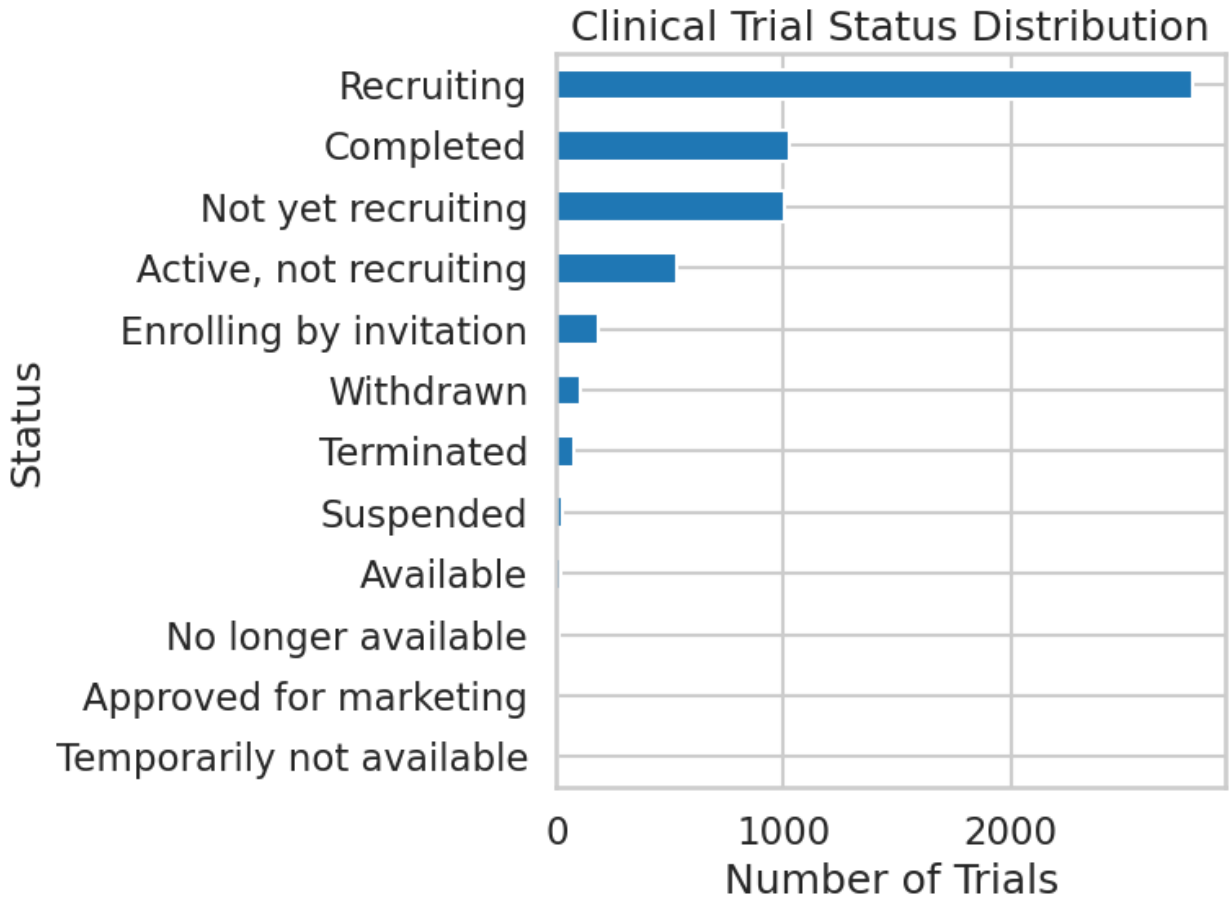
```
 9   Sponsor/Collaborators    5783 non-null   object
 10  Gender                   5773 non-null   object
 11  Age                      5783 non-null   object
 12  Phases                   3322 non-null   object
 13  Enrollment               5749 non-null   float64
 14  Funded Bys               5783 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               5263 non-null   datetime64[ns]
 19  Primary Completion Date  4321 non-null   datetime64[ns]
 20  Completion Date          4258 non-null   datetime64[ns]
 21  First Posted             5783 non-null   datetime64[ns]
 22  Results First Posted     36 non-null     object
 23  Last Update Posted       5783 non-null   datetime64[ns]
 24  Locations                5198 non-null   object
 25  Study Documents          182 non-null    object
 26  URL                      5783 non-null   object
dtypes: datetime64[ns](5), float64(1), int64(1), object(20)
memory usage: 1.2+ MB
```

Trial Status Distribution

```
status_counts = df['Status'].value_counts()
status_counts
```

```
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
```

```
status_counts.sort_values().plot(kind='barh', figsize=(8,6))
plt.title("Clinical Trial Status Distribution")
plt.xlabel("Number of Trials")
plt.tight_layout()
plt.show()
```

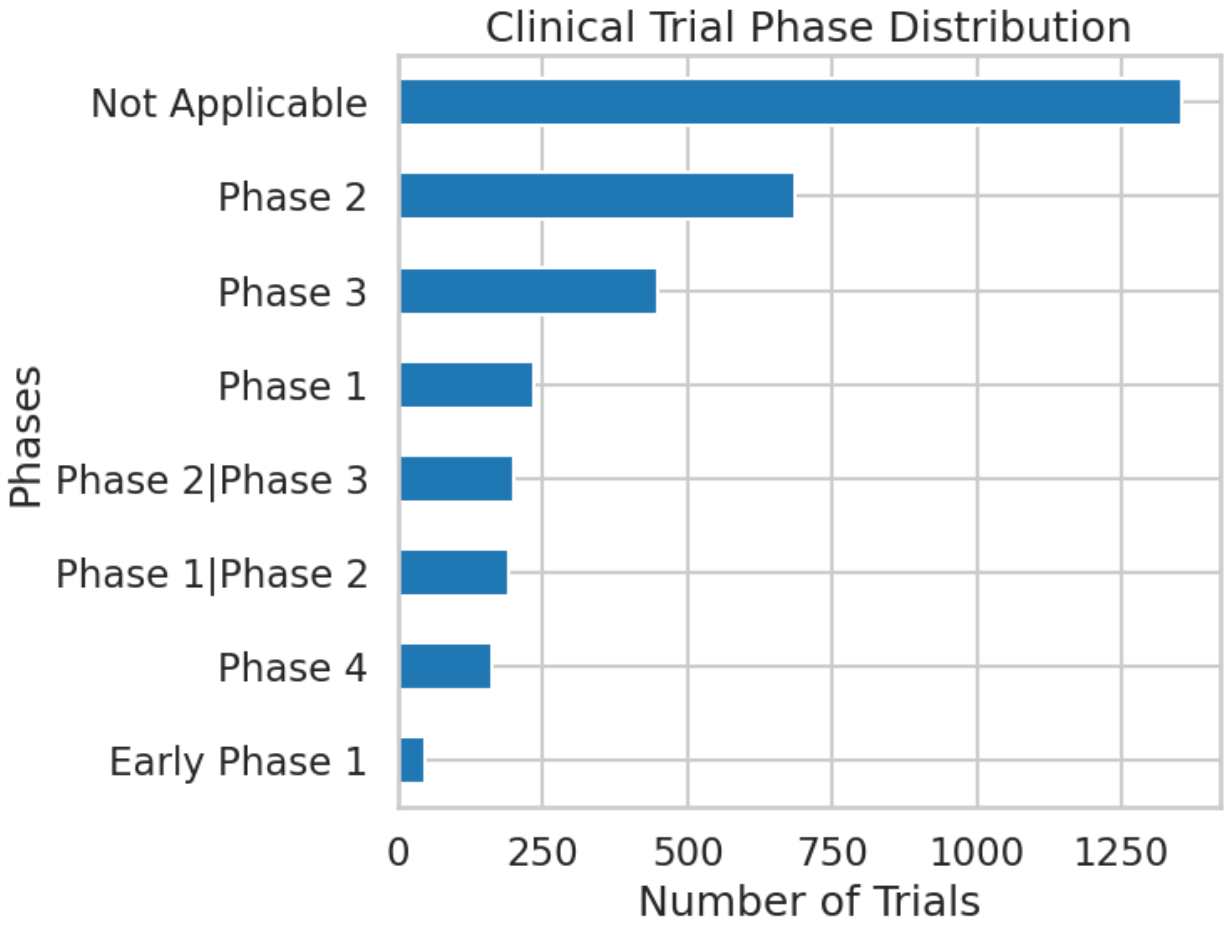


Clinical Trial Phases

```
phase_counts = df['Phases'].value_counts()
phase_counts
```

```
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
```

```
phase_counts.sort_values().plot(kind='barh', figsize=(8,6))
plt.title("Clinical Trial Phase Distribution")
plt.xlabel("Number of Trials")
plt.tight_layout()
plt.show()
```




```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly
remount, call drive.mount("/content/drive", force_remount=True).
```

!ls /content/drive/MyDrive

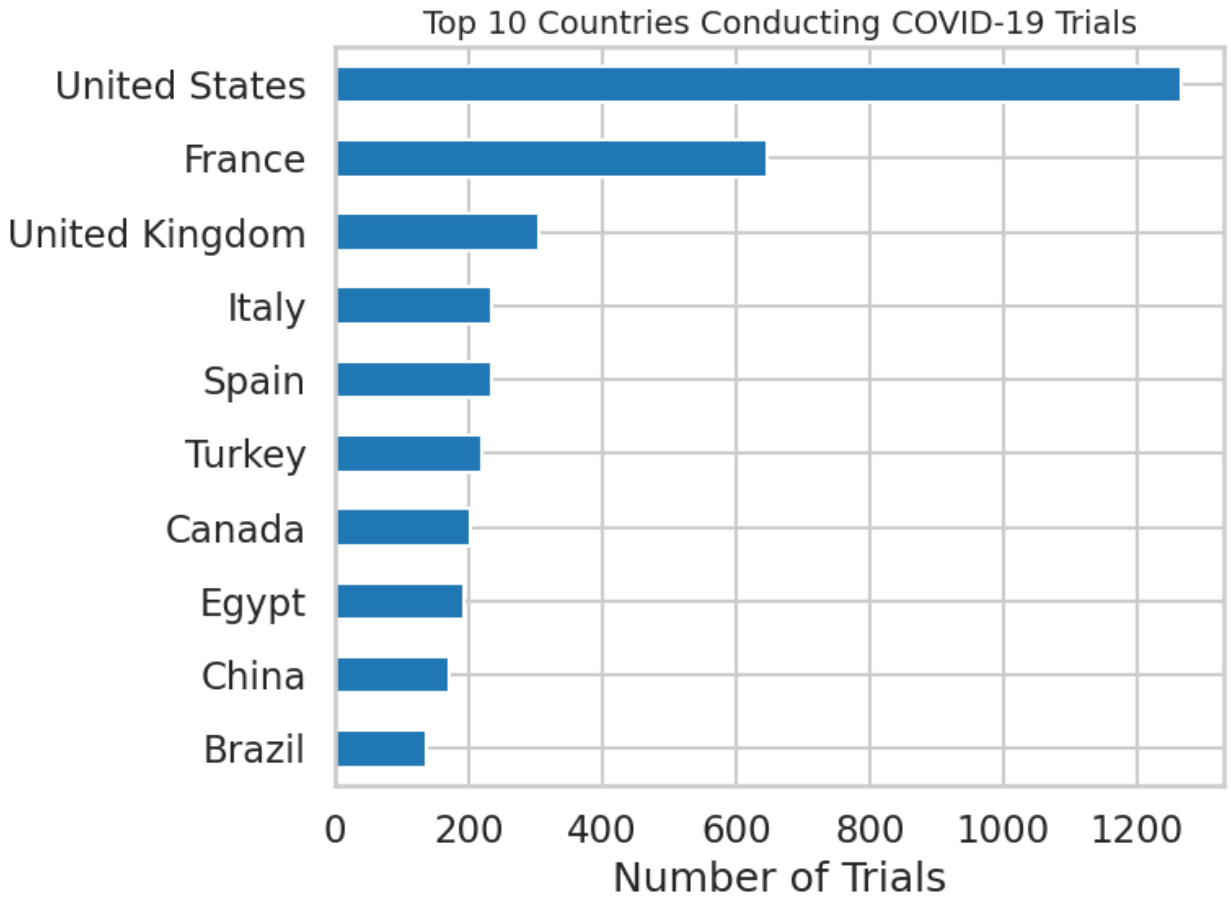
Top Countries Conducting Trials

```
top_countries = df['Locations'].value_counts().head(10)
top_countries
```

```
Locations
Uhmontpellier, Montpellier, France
19
National Institutes of Health Clinical Center, Bethesda, Maryland,
United States                                    16
CHU Amiens, Amiens, France
13
Stanford University, Stanford, California, United States
```

```
13
Massachusetts General Hospital, Boston, Massachusetts, United States
12
M D Anderson Cancer Center, Houston, Texas, United States
11
NYU Langone Health, New York, New York, United States
11
Faculty of Medicine Ain Shams University Research Institute- Clinical
Research Center, Cairo, Non-US, Egypt    11
Uh Montpellier, Montpellier, France
11
Brigham and Women's Hospital, Boston, Massachusetts, United States
11
Name: count, dtype: int64
```

```python
# Extract country (last word after comma)
df['Country'] = df['Locations'].str.split(',').str[-1].str.strip()

top_countries = df['Country'].value_counts().head(10)

plt.figure(figsize=(8,6))
top_countries.sort_values().plot(kind='barh')
plt.title("Top 10 Countries Conducting COVID-19 Trials", fontsize=14)
plt.xlabel("Number of Trials")
plt.ylabel("")
plt.tight_layout()
plt.show()
```



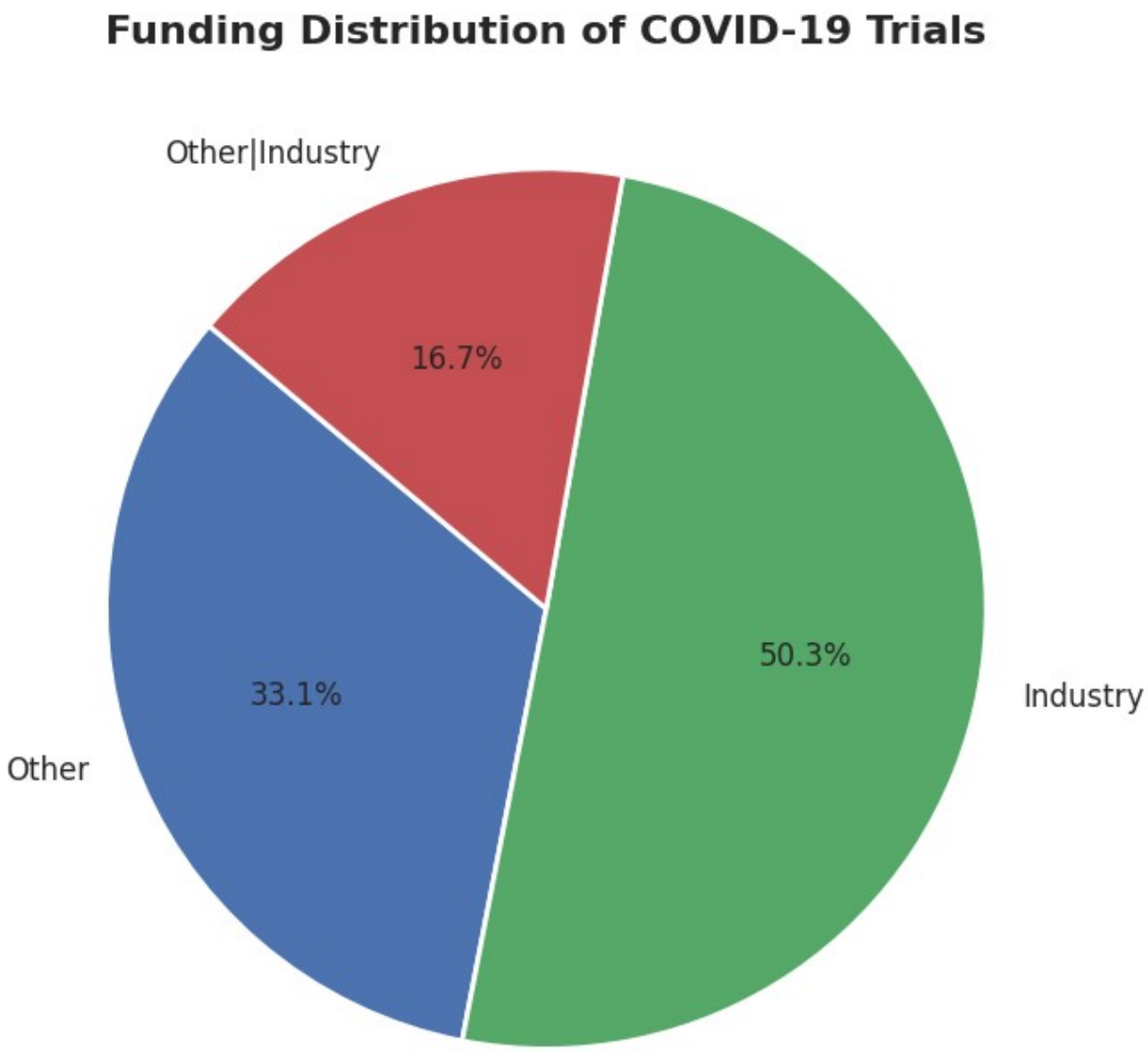
Funding Distribution

```
funding = df['Funded Bys'].value_counts()
funding
```

```
Funded Bys
Other                          4488
Industry                        651
Other|Industry                  216
Industry|Other                  190
Other|NIH                        97
NIH                              51
Other|U.S. Fed                   25
U.S. Fed                         15
Industry|U.S. Fed                10
NIH|Industry                      6
U.S. Fed|Other                    5
NIH|Other                         5
Industry|U.S. Fed|Other           3
NIH|Other|Industry                2
Industry|NIH                      2
Industry|NIH|Other                2
```

```
NIH|Other|U.S. Fed|Industry       2
Other|NIH|U.S. Fed                2
Industry|Other|NIH                2
Other|Industry|NIH                2
Other|NIH|Industry                2
Other|U.S. Fed|NIH                1
Industry|Other|U.S. Fed           1
Other|U.S. Fed|Industry           1
NIH|Industry|Other                1
Industry|U.S. Fed|NIH             1
Name: count, dtype: int64
```

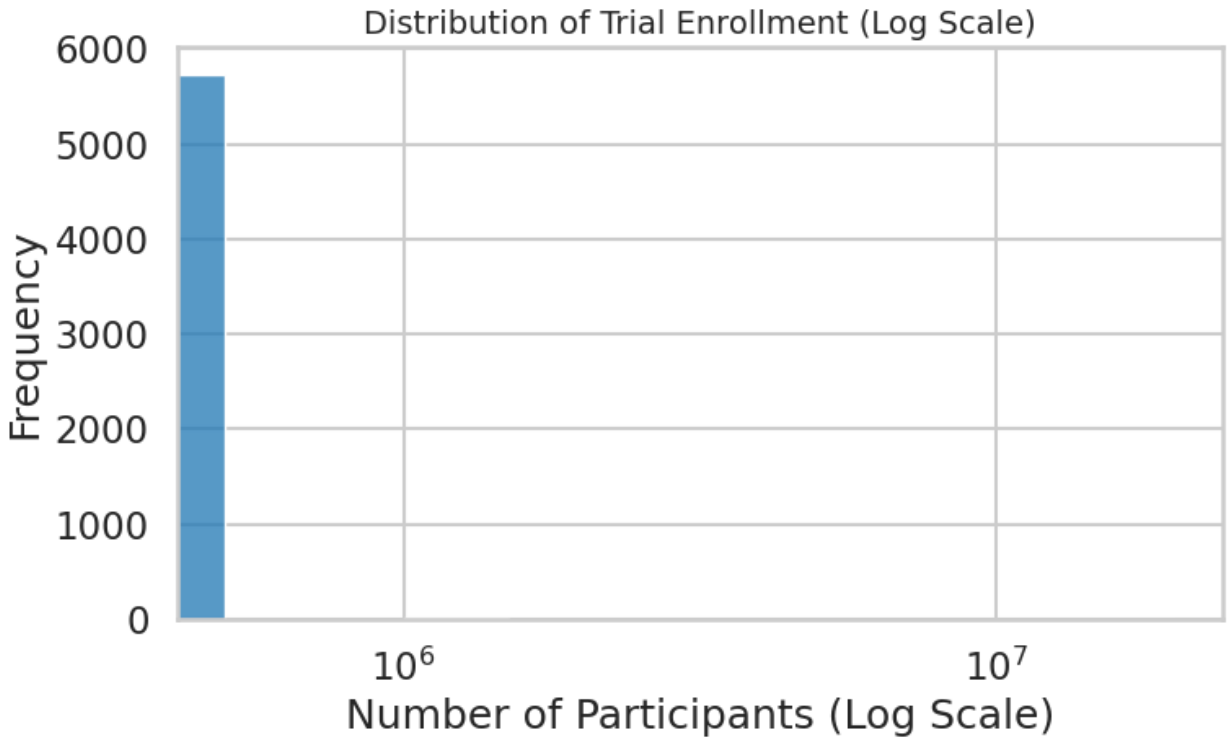
```
plt.figure(figsize=(7,7))

colors = ['#4C72B0', '#55A868', '#C44E52']  # Soft professional
palette

top_funding.plot(
    kind='pie',
    autopct='%1.1f%%',
    startangle=140,
    colors=colors,
    wedgeprops={'edgecolor':'white','linewidth':2},
    textprops={'fontsize':12}
)

plt.title("Funding Distribution of COVID-19 Trials", fontsize=16,
weight='bold')

plt.ylabel("")  # removes side label
plt.tight_layout()
plt.show()
```



Enrollment Analysis

```
plt.figure(figsize=(8,5))

sns.histplot(df['Enrollment'].dropna(), bins=40)

plt.xscale('log')   # MAGIC LINE

plt.title("Distribution of Trial Enrollment (Log Scale)", fontsize=14)
plt.xlabel("Number of Participants (Log Scale)")
plt.ylabel("Frequency")

plt.tight_layout()
```

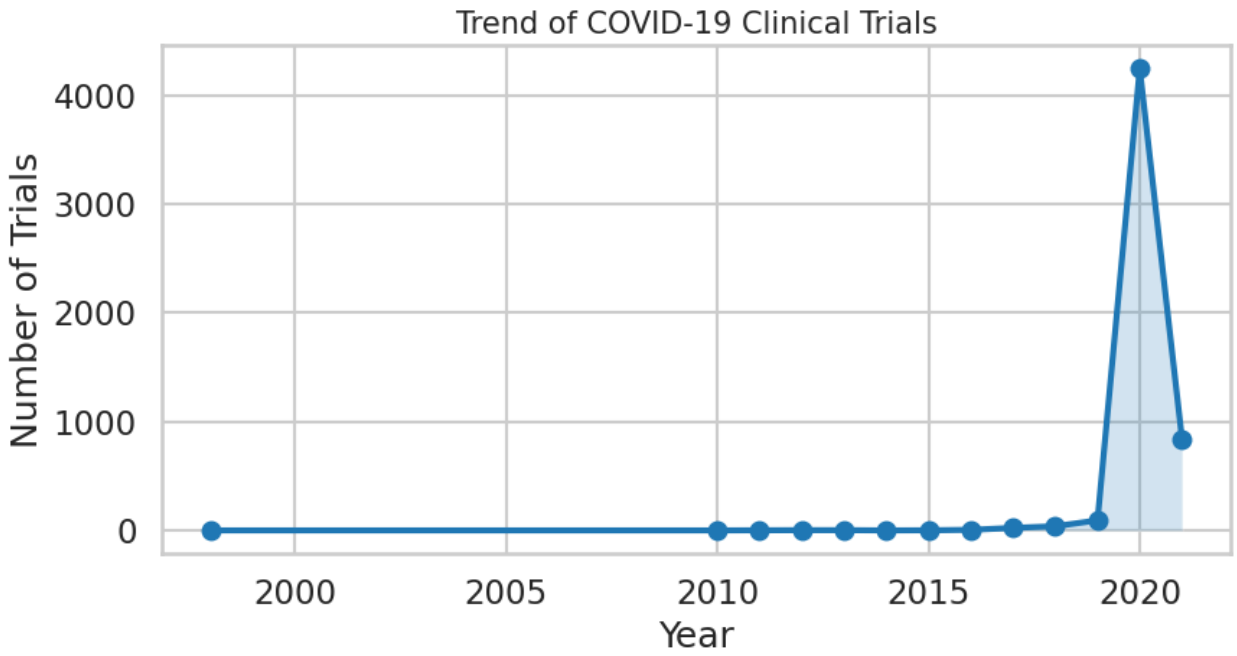
```
plt.show()
```



Clinical Trial Trends

```
sns.set_style("whitegrid")

plt.figure(figsize=(9,5))

plt.plot(
    year_counts.index,
    year_counts.values,
    marker='o',
    linewidth=3
)

plt.fill_between(
    year_counts.index,
    year_counts.values,
    alpha=0.2
)

plt.title("Trend of COVID-19 Clinical Trials", fontsize=15)
plt.xlabel("Year")
plt.ylabel("Number of Trials")
```

```
plt.tight_layout()
plt.show()
```




```
date_cols = [
    'Start Date',
    'Primary Completion Date',
    'Completion Date',
    'First Posted',
    'Last Update Posted'
]

for col in date_cols:
    df[col] = pd.to_datetime(df[col], errors='coerce')
```

Key Insights from COVID-19 Clinical Trials Analysis:

1. **Trial Activity:** A large number of trials were found to be either recruiting or completed, indicating sustained global research efforts.
2. **Trial Phases:** Most studies progressed to Phase 2 and Phase 3, suggesting rapid advancement beyond early-stage testing.
3. **Geographic Distribution:** Countries with strong research infrastructure conducted the highest number of trials, highlighting global disparities.
4. **Funding Pattern:** Government and industry funding were the primary contributors, emphasizing the importance of public-private collaboration.

5. **Enrollment Variation:** Participant numbers varied widely, reflecting both small experimental studies and large-scale clinical investigations.
6. **Research Timeline:** Clinical trial activity surged during peak pandemic years, demonstrating an urgent worldwide scientific response.

CONCLUSION

This analysis highlights the scale and speed of the global scientific response to COVID-19. The findings emphasize the importance of collaboration, funding, and research infrastructure in addressing public health emergencies.

Future work could include predictive modeling and deeper outcome-based analysis.