COVID-19 CLINICAL TRIALS

Exploratory Data Analysis

OBJECTIVE

Amid the COVID-19 crisis, thousands of clinical trials were launched in a global race to find treatments, vaccines, and answers.

This project leverages data science techniques to analyze these trials—identifying:

- Key trends
- Active sponsors
- Intervention types
- Study phases
- Geographical distribution

OVERVIEW

- THE DATASET CONSISTS OF 5783 ENTRIES AND 27 COLUMNS. BELOW IS A SUMMARY OF THE KEY COLUMNS AND THEIR DESCRIPTIONS:

COLUMN ANALYSIS

- **RANK, NCT NUMBER, TITLE:** METADATA FOR EACH TRIAL.
- **ACRONYM:** MISSING FOR 57% OF ROWS; CATEGORICAL.
- **STATUS:** CONTAINS 12 UNIQUE STATUSES LIKE "RECRUITING", "COMPLETED", ETC.
- **STUDY RESULTS, CONDITIONS, INTERVENTIONS, OUTCOME MEASURES:** PROVIDE DETAILED INFORMATION ABOUT EACH TRIAL.
- **SPONSOR/COLLABORATORS, GENDER, AGE:** ADDITIONAL DETAILS ABOUT PARTICIPANTS AND SPONSORS.
- **PHASES:** MISSING FOR 42% OF ROWS; DEFINES TRIAL PHASES.
- **ENROLLMENT:** NUMERIC, MISSING FOR 0.6% OF ROWS; HIGHLY SKEWED.
- **DATES:** INCLUDES START DATE, PRIMARY COMPLETION DATE, AND COMPLETION DATE.
- **LOCATIONS, STUDY DOCUMENTS:** LOCATION DATA IS PARTIALLY MISSING (10%), AND STUDY DOCUMENTS HAS SIGNIFICANT MISSINGNESS (96.8%).
- **URL:** UNIQUE FOR EVERY TRIAL.

IMPORTING AND EXPLORING THE DATA

Importing The Required Libraries and the Dataset

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

data =
pd.read_csv(r"C:\Users\91931\Downloads\Intern\Dataset\COVID
clinical trials.csv")
```

Exploring data

```
data.head()

data.info()

# Stats for all the numerical columns
data.describe()

# Stats for all the categorical columns
data.describe(include="object")
```

HANDLING MISSING DATA

Handling Missing Data

```
data.isnull().sum()

# Drop the columns with many null values and are not
required
data.drop(columns=["Acronym", "Study Documents", "Results
First Posted"], inplace=True)

data.drop_duplicates(inplace=True)
```

EXPLORATORY DATA ANALYSIS

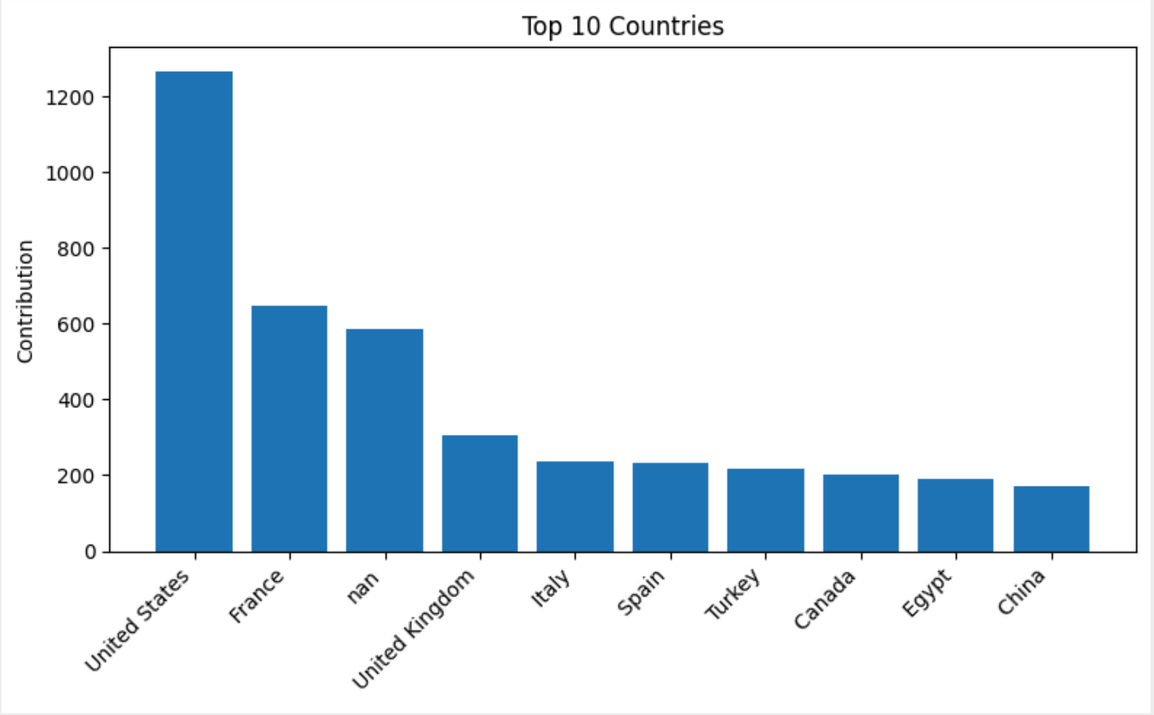
Exploratory Data Analysis

```
# We can extract a new feature from The Location which is the country
where the study hold
countries = [str(data.Locations.iloc[i]).split(",")[-1] for i in
range(data.shape[0])]
data["Country"] = countries

# Impute Interventions , Phases , Locations by Missing Category
categorical_features = data.select_dtypes(include=object).columns

features =
categorical_features[data[categorical_features].isnull().mean() > 0]

for feature in features:
    data[feature] = data[feature].fillna(f"Missing {feature}")
```

DATA VISUALIZATION

Data Visualization

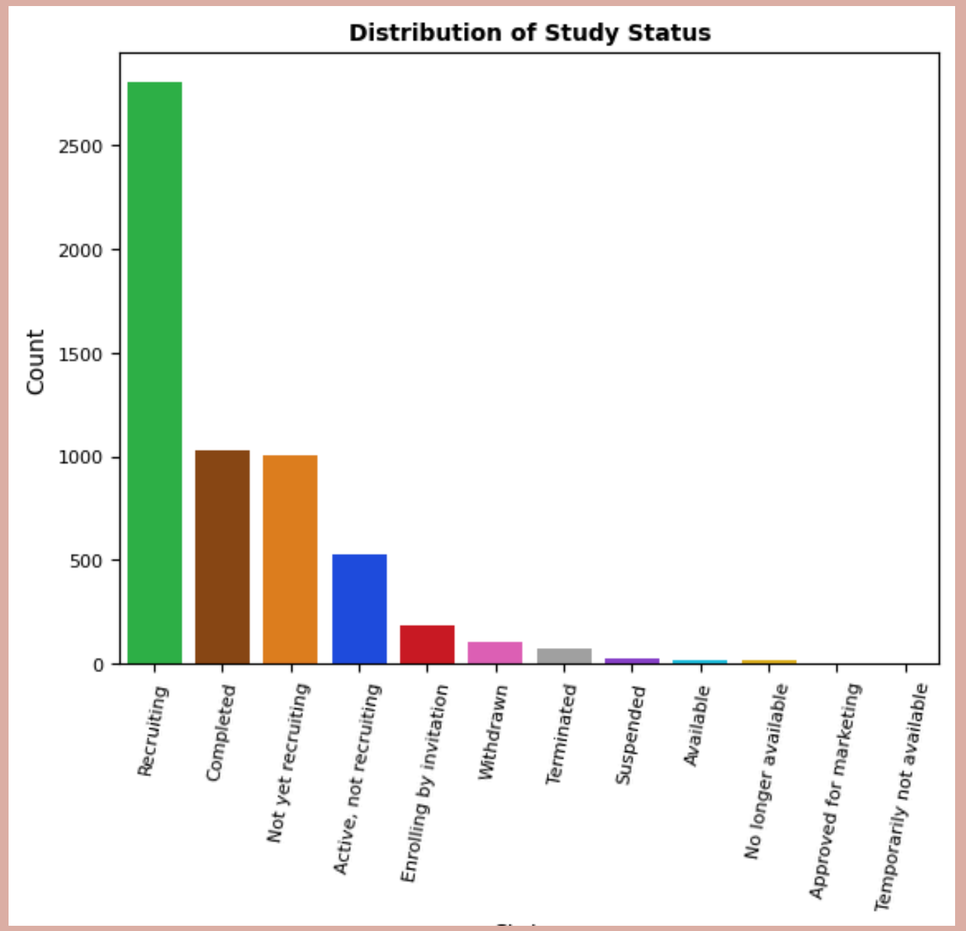
```
import seaborn as sns

top_10_Countries = data.Country.value_counts()[:10]
top_10_Countries

plt.figure(figsize=(8, 5))
plt.bar(top_10_Countries.index, top_10_Countries.values)
# plt.xlabel('Countries')
plt.ylabel("Contribution")
plt.title("Top 10 Countries")
plt.xticks(rotation=45, ha="right")
plt.tight_layout()
plt.show()
```



DISTRIBUTION OF STUDY STATUS

Data Visualization

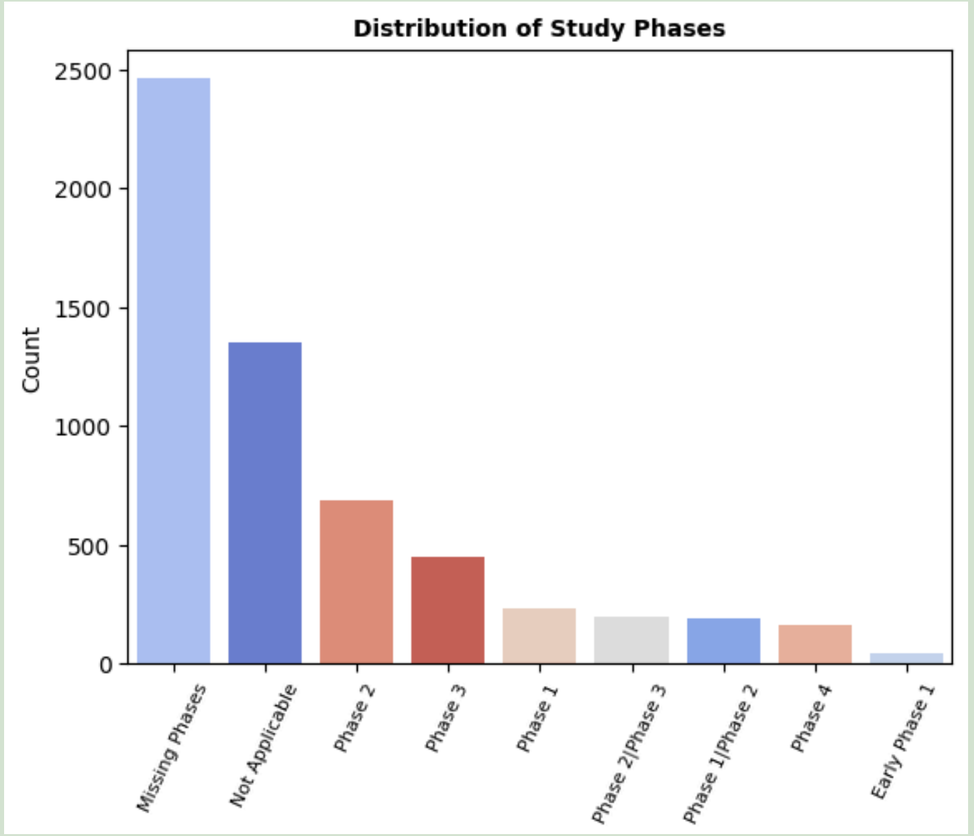
```
# Status of the Application
status = data.Status.value_counts()
status

sns.countplot(
    data=data,
    x="Status",
    order=data["Status"].value_counts().index,
    palette="bright",
    hue="Status",
)
plt.title("Distribution of Study Status", fontsize=10,
fontweight="bold")
plt.xlabel("Status", fontsize=10)
plt.ylabel("Count", fontsize=10)
plt.xticks(rotation=80, fontsize=8)
plt.yticks(fontsize=8)
plt.show()
```



DISTRIBUTION OF STUDY PHASES

Distribution of Study Phases

```
sns.countplot(
    data=data,
    x="Phases",
    order=data["Phases"].value_counts().index,
    palette="coolwarm",
    hue="Phases"
)
plt.title("Distribution of Study Phases", fontsize=10,
fontweight="bold")
plt.xlabel("Phases", fontsize=10)
plt.ylabel("Count", fontsize=10)
plt.xticks(rotation=65, fontsize=8)
plt.show()
```



CONCLUSION

- Most clinical trials are either Completed or Recruiting, reflecting a high level of research activity throughout the pandemic.
- Trials predominantly focus on adult populations and include participants of all genders, indicating inclusive study designs.
- Phase 2 and Phase 3 trials make up the majority of those with known phases, signifying significant progress toward evaluating treatment efficacy and safety.
- The United States and France emerged as major contributors to COVID-19 clinical research.
- A sharp increase in trial registrations around mid-2020 coincided with the global response to the pandemic's first wave.
- Data preprocessing included imputing missing values and dropping columns with excessive missingness; for skewed fields like Enrollment, the median was used to ensure robustness.