Project Report

Title: Exploratory Data Analysis of COVID-19 Clinical Trials using Pandas

1. Introduction

The COVID-19 pandemic has provoked an upsurge in clinical research worldwide. ClinicalTrials.gov, operated by the U.S. National Library of Medicine, is a database of publicly and privately funded clinical studies. This project conducts **Exploratory Data Analysis (EDA)** of COVID-19 clinical trials data to extract insights about trial features, demographics, and research patterns.

2. Objectives

To play around with the dataset and see how the structure of the COVID-19 clinical trials looks.

To find missing values and deal with them accordingly.

To look at distributions of trial status, trial phase, and age groups.

To look at temporal trends of trials initiated during the pandemic.

To create a clean dataset ready for potential downstream uses like machine learning models.

3. Tools and Technologies

Programming Language: Python

Libraries: Pandas, NumPy, Matplotlib, Seaborn

Additional Tools: Excel, SQL (for exploration and queries)

Domain: Data Analysis & Data Science

Difficulty Level: Intermediate

4. Dataset Description

Source: [ClinicalTrials.gov] (https://www.clinicaltrials.gov)

Format: XML (per study) + CSV summary file

Records:

\~5,783 trials with

\~25 features

Key Attributes:

NCT Number (unique ID)

Title, Acronym

Status (Recruiting, Completed, etc.)

Conditions, Interventions, Outcome Measures

Sponsors/Collaborators

Gender, Age, Phases, Enrolment size

Study Design, Start/Completion Dates

Locations and Countries

5. Methodology

5.1 Data Loading and Inspection

Loaded dataset into Pandas Data Frame.

Checked shape: ` (5783, 26) ` columns originally.

Checked categorical and numerical features.

5.2 Handling Missing Data

Dropped columns with >95% missing values (`Results First Posted`, `Study Documents`).

Categorical columns with missing values imputed with `\"Missing <Feature>`".

Highly skewed numerical field **Enrolment** (skewness ~34) → imputed using **median.**

Location column parsed to pull out **Country**.

5.3 Univariate Analysis

Status Distribution: Most trials are "Recruiting" or "Completed."

Phases: Most trials in Phase 2 and Phase 3.

Age Groups: Mostly adults; fewer paediatric/elderly-oriented trials.

Gender: Both gender in most studies.

5.4 Bivariate Analysis

Cross-tab of **Status** by **Phases** showed trends (e.g., numerous active studies in initial phases).

Conditions correlated with **Outcome Measures** to evaluate areas of interest.

5.5 Time-Series Analysis

Translated start dates to monthly intervals.

Saw a **consistent increase in trials in mid-2020**, illustrating global research reaction.

5.6 Country-Level Analysis

Leading contributors: **United States, France, United Kingdom, Italy, Spain, Turkey, Canada, Egypt, China, Brazil.**

U.S. has made the greatest number of studies (\~1267).

6. Key Findings

1. Most COVID-19 trials are **interventional** in design.
2. Largest proportion of studies are **focused on adults**.
3. **Recruitment status** indicates a significant proportion of ongoing and active trials.
4. **Phases 2 and 3 dominate**, reflecting the types of vaccine and drug efficacy trials.
5. **United States and European countries** contribute most globally.
6. There was an **uptick in studies in 2020**, indicative of quick pandemic response.

7. Visual Insights

Bar charts: Distribution of trial statuses, phases, gender, and countries.

Time-series line plot: Trials started by month.

Geographic analysis: Contribution by country.

8. Conclusion

The investigation finds that COVID-19 clinical research has been extremely global in nature, with considerable emphasis on adult populations, Phase 2/3 drug/vaccine trials, and driven by the U.S. and European nations. Proper management of missing values was very important to prevent bias. EDA gave actionable insights into COVID-19 trial structure and trends, which can be used in informing additional statistical modelling as well as machine learning applications.

9. Future Scope

Use **machine learning** models to forecast trial outcomes using design characteristics.

Conduct **NLP-based analysis** of study descriptions for topic modelling.

Investigate **survival analysis** of completion timelines for trials.

Develop interactive dashboards for real-time tracking of clinical trials.

10. References

[ClinicalTrials.gov] (https://www.clinicaltrials.gov)

Kaggle: [COVID-19 Clinical Trials EDA] (https://www.kaggle.com/parulpandey/eda-on-covid-19-clinical-trials)

GitHub Repository: [COVID-19 EDA Analysis](https://github.com/0xpranjal/COVID-19-complete-EDA-analysis).