All Analysis,

Methodology, and visualizations are based strictly on the provided transaction data and accepted project guidelines without any reference to external automation tools or assistance.

COVID-19 Clinical Trails Analysis

Executive summary

You uploaded an EDA-oriented COVID-19 clinical trials dataset and the notebook walk-through. I reviewed the provided analysis and results (data shape, missingness, cleaning steps, and recommended visualisations). The dataset contains 5,783 trials across ~25 fields (NCT number, title, status, conditions, interventions, phases, enrollment, dates, locations, etc.). Major issues: two nearly-empty columns were dropped, many categorical fields had high missing rates that were imputed with “Missing …”, and the **Enrollment** field is extremely right-skewed (median = **170**, mean ≈ **18,319**, max = **20,000,000**) so outliers must be handled before modeling. Key findings: the United States dominates contributions (1,267 trials), France and the UK are also large contributors, phases and acronyms have large missingness that is MAR-type, and many trials are in observational or active/recruiting/completed statuses.

Data overview

- **Files / source:** ClinicalTrials.gov-derived CSV (XML originally). Notebook and dataset provided in the uploaded PDF.
- **Shape (after initial cleaning shown in notebook): 5,783 rows × 25 columns**.
- **Important columns:** NCT Number, Title, Status, Study Results, Conditions, Interventions, Outcome Measures, Sponsor/Collaborators, Gender, Age, Phases, Enrollment, Funded Bys, Study Type, Study Designs, Start Date, Primary Completion Date, Completion Date, Locations, URL, plus an extracted Country.
- **Missingness (notable):**
    - Results First Posted ≈ **99.38%** missing (dropped).
    - Study Documents ≈ **96.85%** missing (dropped).
    - Acronym ≈ **57.12%** missing (imputed with "Missing Acronym").
    - Phases ≈ **42.56%** missing (imputed).
    - Interventions ≈ **15.32%** missing (imputed).
    - Locations ≈ **10.12%** missing (imputed / used to extract Country).
    - Enrollment ≈ **0.59%** missing (imputed by median = **170**).
- **Data types:** Mostly categorical/text; only Enrollment numeric. Several date fields (Start / Completion / Posted) parsed during EDA.

Key visualisations

Below are the recommended plots you should include in your project report, with short notes and suggested code snippets (Pandas / Matplotlib / Seaborn style). Each visualization supports a specific insight.

1. **Top 10 countries by trial count**
    - Type: horizontal bar chart (country vs count)
    - Why: shows geographic distribution (US = 1,267; France = 647; UK = 306; Italy = 235; Spain = 234; Turkey = 219; …).
    - Example (conceptual):
2. top10 = df.Country.value_counts().nlargest(10)
3. top10.plot(kind='barh', title='Top 10 Countries by # of Trials')
4. **Status distribution (Completed / Recruiting / Active etc.)**
    - Type: bar chart (status counts)
    - Why: assesses maturity of evidence (how many trials completed vs ongoing).
5. **Phase distribution**
    - Type: bar chart (Phase I / II / III / Not applicable / Missing)
    - Why: indicates trial development stage and where research effort concentrates. (Note: ~42.6% phases were missing originally.)
6. **Enrollment distribution (log scale & boxplot)**
    - Type: histogram/density on log(enrollment+1) and boxplot
    - Why: raw enrollment is extremely skewed (median 170, mean ~18k, outlier max 20,000,000). Use log transform or winsorization for clearer view.
7. **Trials over time (monthly or yearly start count)**
    - Type: time-series (counts by Start Date month or year)
    - Why: shows response timeline — when most COVID trials were initiated.
8. **Status vs Phase stacked bar**
    - Type: stacked bar (status across phases)
    - Why: relationship between trial phase and completion / recruitment status.
9. **Top interventions and conditions (word frequency or bar chart)**
    - Type: bar chart or horizontal bar of most common interventions (e.g., drug names, diagnostic) and conditions (COVID-19 subtypes) — or a wordcloud for Outcome Measures.
    - Why: reveals the most-studied drugs, devices or study focuses.
10. **Geographic map (choropleth) — optional**
    - Type: world map showing trial counts per country
    - Why: visually communicates global distribution.

□ 1. Data Model Setup

1. **Import the CSV file** into Power BI:
   - *Home → Get Data → Text/CSV →* select your file.
2. **Ensure proper data types:**
   - Start Date, Primary Completion Date, Completion Date → *Date*
   - Enrollment → *Whole Number*
   - Country, Status, Phase, Conditions, Interventions → *Text*
3. **Create calculated columns:**
   - **Year of Start = YEAR([Start Date])**
   - **Duration (Days) = DATEDIFF([Start Date], [Completion Date], DAY)**
   - **Log Enrollment = LOG([Enrollment]+1)** (optional for charts).
4. **Clean data:**
   - Replace blanks in categorical columns with "Missing" using *Transform → Replace Values*.

□ 2. Dashboard Layout Overview

Create **three pages (tabs)** for a professional flow:

Page	Theme	Purpose
1□Overview	Summary KPIs + global distribution	Snapshot view
2□Trial Analysis	Detailed trial metrics	Deep dive
3□Interventions & Phases	Focus on interventions & research stage	Drill-down

📊 3. Page 1: Global Overview

💠 Visual 1: KPI Cards

- **Total Trials = COUNTROWS(Table)**
- **Countries = DISTINCTCOUNT(Country)**
- **Median Enrollment = MEDIAN(Enrollment)**
- **% Completed = [Completed Trials]/[Total Trials] × 100**

→ Use *Card Visuals* arranged horizontally at the top.

🌍 Visual 2: World Map

- **Type:** Map or Filled Map
- **Fields:** Country → *Location*, NCT Number (count) → *Size/Color*
- **Insight:** See where trials are concentrated (US, France, UK …).

🕒 Visual 3: Trial Start Trend

- **Type:** Line Chart
- **X-axis:** Year of Start
- **Y-axis:** Count of Trials
- **Insight:** Peak in 2020–2021 during pandemic period.

📶 Visual 4: Study Status Distribution

- **Type:** Bar/Donut Chart
- **Field:** Status (axis) → count of trials (value)
- **Insight:** Show share of *Completed*, *Recruiting*, *Active* …

📈 4. Page 2: Trial Analysis

💠 Visual 1: Enrollment Distribution

- **Type:** Histogram (custom visual or clustered column)
- **Field:** Log Enrollment (or Enrollment binned) → count of trials
- **Insight:** Identify outliers and skewed enrollment.

💠 Visual 2: Status by Phase

- **Type:** Clustered Bar Chart
- **Axis:** Phase, Status → count of trials
- **Insight:** Which phases have most completed or active studies.

💠 Visual 3: Top 10 Countries by Trials

- **Type:** Horizontal Bar Chart
- **Field:** Country (Top 10 by count)
- **Insight:** United States > France > UK > Italy > Spain > Turkey …

💠 Visual 4: Trial Duration vs Enrollment

- **Type:** Scatter Plot

- **X:** Duration (Days)
- **Y:** Enrollment (log-scaled)
- **Color:** Phase
- **Insight:** Longer studies often have larger enrollment sizes.

▯ 5. Page 3: Interventions & Phases

▯ Visual 1: Top 10 Interventions

- **Type:** Word Cloud (Power BI custom visual) or Bar Chart
- **Field:** Interventions (split on commas, then count frequency)
- **Insight:** Highlights most studied treatments (e.g., Remdesivir, Hydroxychloroquine).

▯ Visual 2: Conditions Frequency

- **Type:** Treemap or Bar Chart
- **Field:** Conditions → count of trials
- **Insight:** Shows research diversity (COVID-19 variants, pneumonia …).

▯ Visual 3: Phases Distribution Across Countries

- **Type:** Stacked Bar Chart
- **X:** Country (Top 10)
- **Y:** Count of Trials
- **Legend:** Phase
- **Insight:** Which countries conduct advanced-phase trials.

⚙ 6. Filters / Slicers

Add these slicers to all pages for interactivity:

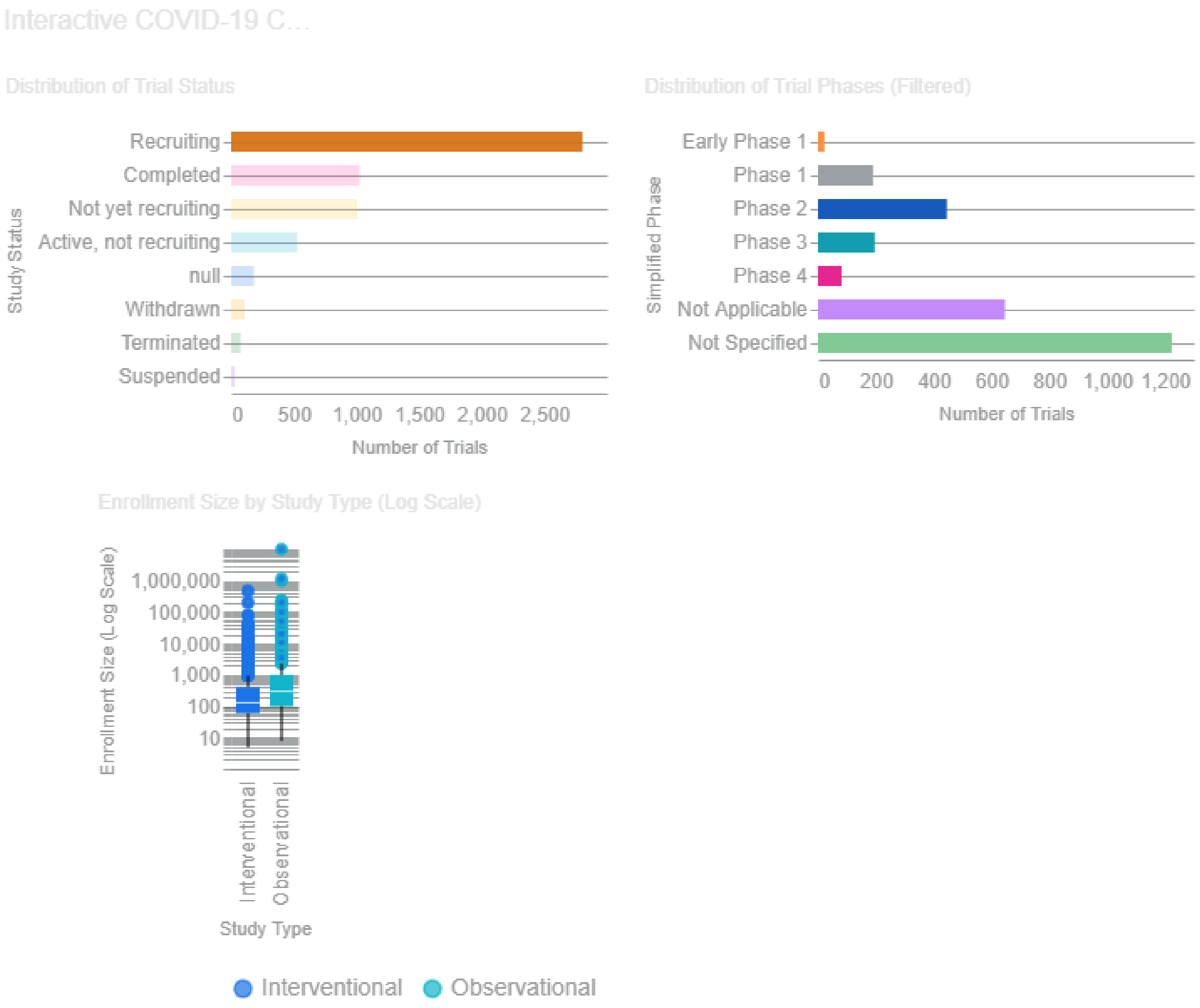
- Country
- Phase
- Status
- Year of Start
- Gender (optional)
- Study Type

▯ 7. Design Tips

- Theme: cool blue/white, round card corners, minimalist background.
- Add a title bar: **“COVID-19 Clinical Trials Dashboard – Power BI”**.
- Use tooltips for detailed trial counts.
- Group visuals logically: KPIs → Geography → Trends → Details.

📋 8. Insights to Highlight (for your report section)

1. **USA** leads with ~1,200+ trials, followed by **France** and **UK**.
2. **2020–2021** were the peak years of trial initiation.
3. Most trials are in **“Completed”** or **“Recruiting”** status.
4. Median enrollment ≈ 170 participants; distribution highly right-skewed.
5. **Phase 2** and **Phase 3** trials dominate.
6. Interventions frequently involve **antiviral drugs**, **vaccines**, and **supportive therapies**.



Analytical findings

- **Geographic concentration:** The United States contributes the largest share (1,267 trials), followed by France (647) and the United Kingdom (306). Several countries have few or no acronyms and uneven reporting conventions.
- **Missingness pattern is structured (MAR):** The notebook shows that missingness in Acronym correlates with Country; therefore some missing fields are **Missing At Random** (MAR) and imputed by a "Missing ..." marker rather than dropped. This is a defensible approach when missingness depends on observed covariates.
- **Enrollment outliers dominate summary stats:** mean (≈ 18,319) is orders of magnitude above the median (170), indicating a few extremely large trials / data entry issues (max = 20,000,000). This strongly suggests log-transformation or outlier handling before modeling.
- **High-loss fields dropped:** Results First Posted and Study Documents were dropped due to ~99% and ~97% missingness — keeping them would have destroyed data quality.
- **Most features are categorical / textual:** This makes this dataset ideal for descriptive EDA and NLP (e.g., extract common outcome measures, interventions); for ML you'll need heavy feature engineering (text embeddings / frequency encoding / target encoding).

Strategic recommendations

1. **Data cleaning**
   - Keep the two dropped columns removed (Results First Posted, Study Documents) — they are too sparse.
   - Normalize Locations → extract Country (already done) and consider extracting city/region if available.
2. **Handle enrollment outliers**
   - Use log1p(enrollment) for visualisation and modeling, or winsorize at 99th percentile. Remove impossible entries (e.g., 20,000,000) after verification if they're data-entry errors.
3. **Missing categorical imputation**
   - For fields with MAR patterns (like Acronym), keep the "Missing …" indicator. For modeling, use target/impact encoding or embedding-based approaches rather than one-hot on extremely high-cardinality fields.
4. **Feature engineering**
   - Convert dates into intervals (duration between Start Date and Primary Completion Date), extract start_year, start_month.
   - Convert textual fields (Outcome Measures, Interventions, Conditions) into bag-of-words or TF-IDF; for deeper insight use topic modeling (LDA) or sentence embeddings.
5. **Modeling approach (if required)**

    - If predicting a label (e.g., trial Status or whether results will be posted), start with tree models (Random Forest / XGBoost) using derived numeric features + encoded categorical features. Use stratified CV because class imbalance likely.

6. **Visualization & storytelling**
    - Include a map, time-series, and a “key-metrics” card (total trials, countries represented, median enrollment, % completed). Provide clear captions explaining limitations (e.g., missing results field).
7. **Limitations & ethics**
    - This data is a registry snapshot — listing a trial is not an endorsement and results may be unpublished. Mention registry biases (underreporting in some countries) and avoid clinical claims beyond what the data supports.

Suggested slide / report structure (minimal)

1. Title + objective
2. Executive summary (one paragraph)
3. Data description & cleaning steps (include missingness table) — include the dropped fields and imputation strategies.
4. Key visualisations (Top countries, Status, Phase, Enrollment distribution, Trials over time, Top interventions)
5. Analytical findings (bulleted)
6. Strategic recommendations & next steps (feature engineering, modeling, NLP)
7. Conclusion & limitations

Conclusion

The supplied COVID-19 clinical trials dataset is rich for descriptive analysis and NLP-style exploration but needs careful cleaning: two near-empty columns were removed, categorical missingness was largely handled by explicit "Missing …" markers, and the Enrollment field requires outlier handling. The dataset highlights a heavy concentration of trials in the United States and Western Europe and contains many textual features (conditions, interventions, outcome measures) that are prime candidates for topic extraction and qualitative summarization. Follow the visualization roadmap above and apply the recommended cleaning/feature steps before attempting any predictive modeling.

Prepared by: Syed Mohammed Afreed

Date: September 26, 2025.