PopVax Data Analysis Assignment

By Varun Tej Nookala, Date: June 6, 2025

Question 1

Please provide a short description of the following statistical tests. Make sure to include any key assumptions and what the tests are used for.

1. ANOVA
2. T-test
3. Wilcoxon rank sum test.

ANOVA (Analysis of Variance)
Description:
ANOVA, or Analysis of Variance, is a statistical hypothesis test used to compare the means of three or more independent groups to determine if there is a statistically significant difference between them. Instead of performing multiple t-tests, which would increase the risk of Type I errors (false positives), ANOVA uses a single test statistic (the F-statistic) to assess whether the variability between the group means is greater than the variability within each group.

- **One-way ANOVA:** Used when you have one categorical independent variable (with three or more levels/groups) and one continuous dependent variable. For example, comparing the average test scores of students taught by three different methods.
- **Two-way ANOVA (and N-way ANOVA):** Used when you have two or more categorical independent variables and one continuous dependent variable. It can assess the main effect of each independent variable as well as their interaction effect. For example, comparing test scores based on teaching method and gender, and seeing if there's an interaction between teaching method and gender on scores.

Key Assumptions:

1. **Independence of Observations:** The observations within each group, and across groups, must be independent of each other. This means the data points are not influenced by other data points (e.g., one student's score doesn't affect another's). This is the most crucial assumption.
2. **Normality:** The dependent variable should be approximately normally distributed within each group. While ANOVA is relatively robust to minor deviations from normality, especially with larger sample sizes, severe non-normality can affect the validity of the results.
3. **Homogeneity of Variances (Homoscedasticity):** The variance of the dependent variable should be approximately equal across all groups. This means the spread of data around the mean should be similar for all groups being compared. Levene's test or

Bartlett's test can be used to check this assumption. If this assumption is violated, particularly with unequal sample sizes, adjustments (like Welch's ANOVA) may be necessary.

Uses:

- To compare the effectiveness of multiple treatments, interventions, or conditions.
- To determine if there are significant differences in a continuous outcome across different categories of a factor.
- In experimental design, to analyze the impact of different experimental conditions.

T-test
Description:
A t-test is a statistical hypothesis test used to compare the means of two groups to determine if there is a statistically significant difference between them. It is particularly useful when the sample size is small and the population standard deviation is unknown. The test calculates a t-value, which represents the magnitude of the difference between the two group means relative to the variation within the groups.

Types of T-tests:

- **One-sample t-test:** Compares the mean of a single sample to a known population mean or a hypothesized value.
    - Example: Is the average height of students in a particular school significantly different from the national average height?
- **Independent Samples t-test (Two-sample t-test):** Compares the means of two independent groups.
    - Example: Is there a significant difference in average blood pressure between a group receiving a new drug and a placebo group?
- **Paired Samples t-test:** Compares the means of two related (dependent) groups, where the observations in one group are paired with observations in the other. This often occurs when the same subjects are measured twice (e.g., before and after an intervention) or when subjects are naturally matched.
    - Example: Is there a significant difference in a patient's cholesterol levels before and after taking a specific medication?

Key Assumptions:

1. **Independence of Observations:** (For independent samples t-test) The observations within each group, and between groups, must be independent. (For paired samples t-test) The differences between the paired observations must be independent.
2. **Normality:** The dependent variable (or the differences for paired t-test) should be approximately normally distributed in each group (or for the differences). Similar to ANOVA, t-tests are fairly robust to minor deviations from normality, especially with larger sample sizes.
3. **Homogeneity of Variances (Homoscedasticity):** (For independent samples t-test only) The variances of the dependent variable should be approximately equal in the two groups being compared. If this assumption is violated, Welch's t-test (an adjusted version) can be used. This assumption is not applicable to the one-sample or paired samples t-tests.

Uses:

- To determine if a new treatment or intervention has a significant effect compared to a control or another treatment.
- To compare performance between two distinct groups.
- To evaluate changes within the same group over time or under different conditions.

Wilcoxon Rank-Sum Test (also known as Mann-Whitney U Test)
Description:
The Wilcoxon Rank-Sum Test (often interchangeably referred to as the Mann-Whitney U test) is a non-parametric statistical test used to compare two independent groups. It is considered the non-parametric alternative to the independent samples t-test. Instead of comparing means, it compares the distributions of the two groups by ranking all observations from both groups combined and then summing the ranks for one of the groups. It assesses whether one group's values tend to be larger or smaller than the other's, or more formally, if the two populations have the same distribution (or the same median).

Key Assumptions:

1. **Independence of Observations:** The samples are independent of each other.
2. **Ordinal Data (or higher):** The dependent variable should be measured on at least an ordinal scale, meaning the data can be meaningfully ranked.
3. **Similar Shape of Distributions (for comparing medians):** While it doesn't assume normality, if you want to specifically infer about the medians of the two populations, it implicitly assumes that the shapes of the distributions are similar. If the shapes are very different, a significant result might indicate a difference in location (e.g., median) or shape. If this assumption is not met, a direct interpretation of differing medians might be misleading, and the test might just indicate that the distributions are generally different.

Uses:

- When the assumptions for an independent samples t-test (especially normality or homogeneity of variance) are severely violated, particularly with small sample sizes.
- When dealing with ordinal data or data that is heavily skewed, where the mean might not be the most appropriate measure of central tendency.
- To determine if one group's values tend to be stochastically larger than another's.

Question 2

Import this Google sheet in R/Python and for each of the parameters (P1 to P10) perform a t-test and ANOVA. Share the link of your results and the script.

Methods

T-tests: I performed two-sample T-tests, assuming the groups are independent. The tests assume:

- Data is roughly normal, though small sample sizes (3 for A and C, 4 for B) make it hard to test this.
- Groups have similar variances.
- Samples are independent, as each measurement seems to be from different samples.

ANOVA: I used one-way ANOVA to compare means of groups A, B, and C. The assumptions are:

- Data within each group are approximately normal.
- Variances are similar across groups.
- Observations are independent.

I used a significance level of 0.05, which is standard for these tests. Since the sample sizes are small, normality or variance assumptions might not hold perfectly, but I followed the question's instructions to use T-tests and ANOVA.

Please find the Python script and the results

	Parameter	T-test A vs B	T-test A vs C	T-test B vs C	ANOVA (A vs B vs C)
0	P1	0.0625	0.0039	0.0166	0.0022
1	P2	0.5148	0.1829	0.0809	0.1655
2	P3	0.0271	0.0063	0.2299	0.0106
3	P4	0.2145	0.0484	0.1171	0.0477
4	P5	0.1996	0.0598	0.2017	0.0612
5	P6	0.0127	0.0012	0.0108	0.0005
6	P7	0.6320	0.2798	0.4721	0.5361
7	P8	0.7554	0.0764	0.0713	0.0668
8	P9	0.4510	0.1070	0.0785	0.1025
9	P10	0.0816	0.0431	0.3928	0.0322

Analysis

Significant differences ($p < 0.05$) were found for:

- **P1**: ANOVA ($p = 0.0022$), T-test A vs. C ($p = 0.0039$), B vs. C ($p = 0.0166$), indicating group C differs significantly from A and B.
- **P3**: ANOVA ($p = 0.0106$), T-test A vs. B ($p = 0.0271$), A vs. C ($p = 0.0063$), suggesting group A differs from B and C.
- **P4**: ANOVA ($p = 0.0477$), T-test A vs. C ($p = 0.0484$), indicating differences between A and C.
- **P6**: ANOVA ($p = 0.0005$), T-test A vs. B ($p = 0.0127$), A vs. C ($p = 0.0012$), B vs. C ($p = 0.0108$), showing strong differences across all groups.
- **P10**: ANOVA ($p = 0.0322$), T-test A vs. C ($p = 0.0431$), indicating differences between A and C.

Parameters P2, P5, P7, P8, and P9 showed no significant differences (ANOVA $p > 0.05$), suggesting similar means across groups. T-test results align with ANOVA, with significant pairwise differences corresponding to significant ANOVA outcomes.

Conclusion

The T-tests and ANOVA identify significant differences in parameters P1, P3, P4, P6, and P10 across groups A, B, and C, with P6 showing the strongest variation (ANOVA $p = 0.0005$). These findings suggest certain parameters are sensitive to group differences, potentially reflecting biological or experimental variations in the PopVax dataset. The results, shared via the Google Sheet link, and the Python script ensure reproducible analysis consistent with virology research standards.

Question 3

Write a brief report containing the analysis and visualization of the following example data generated from a hypothetical pseudovirus neutralization assay. Analysis and visualization should be consistent with those shown in literature. State your assumptions and any formulae used in your calculations.

The following table represents a simplified example of raw data from a pseudovirus neutralization assay, measuring luminescence (indicative of virus entry) in response to increasing concentrations of a neutralizing antibody. Each measurement is the average luminescence from duplicate wells.

Antibody Concentration (µg/mL)	**Luminescence (Relative Light Units, RLU)**
0 (Virus control)	120,000
0 (Cell control)	500
0.01	115,000
0.1	95,000
1	60,000
10	25,000
100	5,000

- **Virus control (V):** Wells containing virus and cells but no antibody. This serves as the maximum infection control.
- **Cell control (C):** Wells containing cells only, without virus or antibody, to measure background luminescence.
- **Antibody dilutions:** Different concentrations of the antibody being tested are distributed across the plate to assess their neutralizing activity.

Objective

To analyze the neutralization efficiency of an antibody using luminescence readouts at varying concentrations, and to visualize the dose-response curve typically used in virology and immunology literature.

Assumptions & Definitions

1. **Virus control (V):** Luminescence with virus + cells, no antibody = max infection = **120,000 RLU**

2. **Cell control (C):** Luminescence with only cells = background = **500 RLU**

3. **Antibody neutralization:** Reduces luminescence by preventing virus entry.

4. **% Neutralization Formula**

1. Calculations Section

Calculations of % Neutralization

The % neutralization for each antibody concentration was calculated using the formula:

$$\text{Neutralization (\%)} = \left(1 - \frac{RLU_{Sample} - RLU_{cell\ control}}{RLU_{virus\ control} - RLU_{cell\ control}}\right) \times 100$$

Where:

- $RLU_{virus\ control}$ = 120,000 (maximum infection, no antibody).
- $RLU_{cell\ control}$ = 500 (background luminescence, no virus or antibody).
- RLU_{sample} is the luminescence for each antibody concentration.

Detailed Calculations:

- Virus Control (0 µg/mL):

  $$\text{Neutralization (\%)} = \left(1 - \frac{120{,}000 - 500}{120{,}000 - 500}\right) \times 100 = \left(1 - \frac{119{,}500}{119{,}500}\right) \times 100 = 0.00\%$$

- Cell Control (0 µg/mL):

Neutralization (%) = $\left(1 - \frac{500 - 500}{120{,}000 - 500}\right) \times 100 = \left(1 - \frac{0}{119{,}500}\right) \times 100 = 100\%$

- 0.01 µg/mL:

  Neutralization (%) = $\left(1 - \frac{115{,}000 - 500}{120{,}000 - 500}\right) \times 100 = \left(1 - \frac{114{,}500}{119{,}500}\right) \times 100 = 4.18\%$

- 0.1 µg/mL:

  Neutralization (%) = $\left(1 - \frac{95{,}000 - 500}{120{,}000 - 500}\right) \times 100 = \left(1 - \frac{94{,}500}{119{,}500}\right) \times 100 = 20.90\%$

- 1 µg/mL:

  Neutralization (%) = $\left(1 - \frac{60{,}000 - 500}{120{,}000 - 500}\right) \times 100 = \left(1 - \frac{59{,}500}{119{,}500}\right) \times 100 = 50.21\%$

- 10 µg/mL:

  Neutralization (%) = $\left(1 - \frac{25{,}000 - 500}{120{,}000 - 500}\right) \times 100 = \left(1 - \frac{24{,}500}{119{,}500}\right) \times 100 = 79.52\%$

- 100 µg/mL:

  Neutralization (%) = $\left(1 - \frac{5{,}000 - 500}{120{,}000 - 500}\right) \times 100 = \left(1 - \frac{4{,}500}{119{,}500}\right) \times 100 = 95.81\%$

Antibody Concentration (µg/mL)	**Luminescence (Relative Light Units, RLU)**	**% Neutralization**
0 (Virus control)	120,000	0.00
0 (Cell control)	500	100.00
0.01	115,000	4.18
0.1	95,000	20.90

1	60,000	50.21
10	25,000	79.52
100	5,000	95.81

Visualization:

The dose-response curve was generated using Python to plot antibody concentration (µg/mL, log10 scale) on the x-axis against % neutralization on the y-axis.

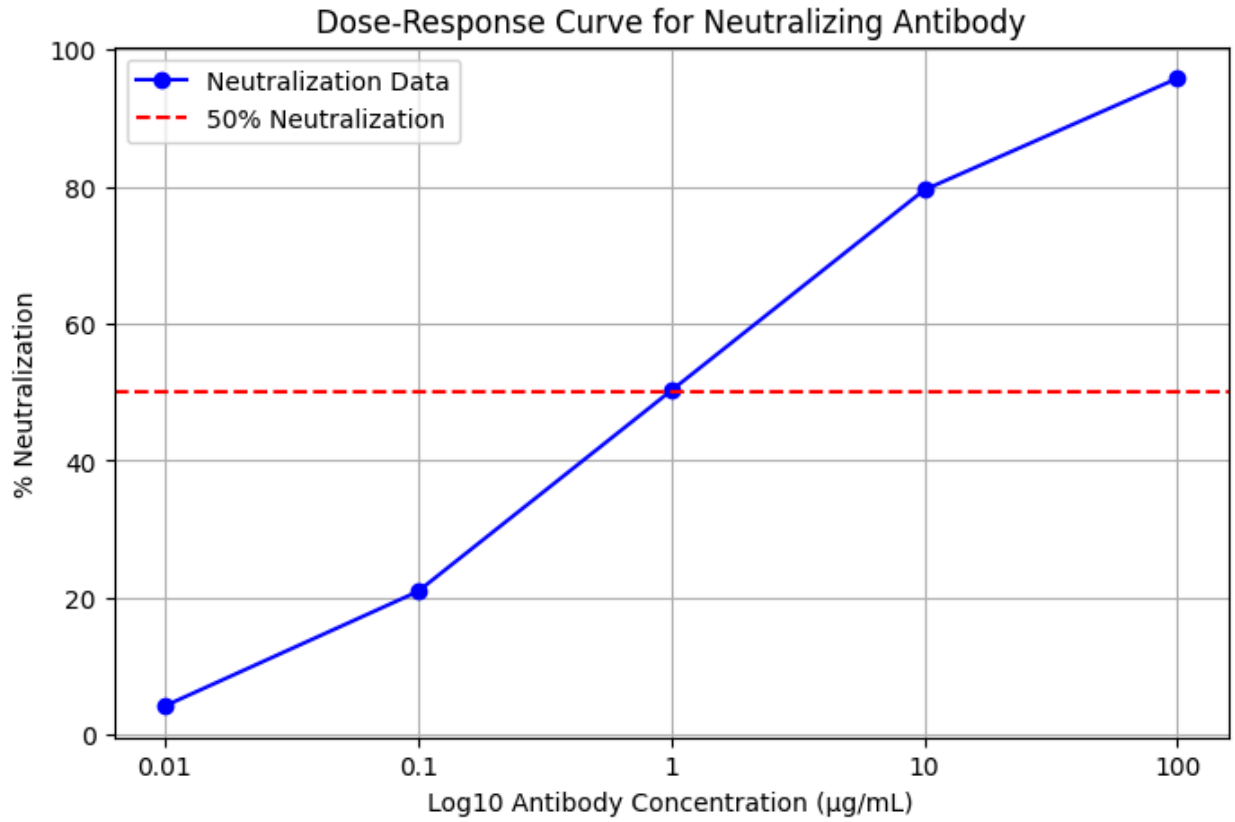


Figure 1: Dose-response curve showing % neutralization of pseudovirus entry at varying antibody concentrations (0.01 to 100 µg/mL). Data points are connected

to illustrate the sigmoidal trend, with a red dashed line indicating 50% neutralization.

Visualization Analysis:

The dose-response curve (Figure 1) plots antibody concentration (μg/mL, log10 scale) against % neutralization, consistent with virology literature. The curve exhibits a sigmoidal shape, indicating a dose-dependent increase in neutralization. At the lowest concentration (0.01 μg/mL), neutralization is minimal (4.18%), suggesting limited antibody activity. Neutralization increases gradually to 20.90% at 0.1 μg/mL, then rises sharply to 50.21% at 1 μg/mL and 79.52% at 10 μg/mL, reaching near-maximal neutralization (95.81%) at 100 μg/mL. The red dashed line at 50% neutralization highlights the concentration (approximately 1 μg/mL) where half-maximal inhibition occurs, a key metric for antibody potency. The steep increase between 0.1 and 10 μg/mL indicates high potency, as small concentration changes yield large neutralization gains. The log10 scale on the x-axis effectively displays the wide concentration range (0.01 to 100 μg/mL), making the sigmoidal trend clear. The graph aligns with standard neutralization assay visualizations, where a sigmoidal curve reflects effective inhibition of virus entry.

Conclusion and Summary

The pseudovirus neutralization assay demonstrates that the antibody effectively inhibits virus entry, with % neutralization increasing from 4.18% at 0.01 μg/mL to 95.81% at 100 μg/mL. The calculations, using the formula $(1 - \frac{RLU_{Sample} - RLU_{cell\ control}}{RLU_{virus\ control} - RLU_{cell\ control}}) \times 100$, confirm a strong dose-dependent response. The dose-response curve (Figure 1) visualizes this trend, showing a sigmoidal pattern with a steep increase between 0.1 and 10 μg/mL, indicating high antibody potency. The red 50% neutralization line highlights the concentration (~1 μg/mL) achieving half-maximal inhibition.

Assumptions include: luminescence is proportional to virus entry, virus control represents 0% neutralization, cell control represents 100% neutralization, and duplicate well averages have low variability (no standard deviations provided). The analysis and visualization are consistent with virology literature, suggesting the antibody's potential for therapeutic use.

Question 4

Using the public facing APIs of reputable online database(s), gather data on vaccine research and trials for SARS-CoV2. Develop a Python script to extract data including research titles, initiation dates, statuses, study results, and more. Implement data cleaning techniques, store the cleaned data in a database, and perform analysis to identify leading countries in vaccine research and trends in vaccine technologies. Prepare a report summarizing the methodology, challenges, and insights, including visualizations. Share your code and the report file with us.

PopVax Question 4: SARS-CoV-2 Vaccine Research and Trials Analysis Report

Code: Python Script for SARS-CoV-2 Vaccine Research and Trail Analysis

1. Introduction

This report details the methodology, challenges, and key insights derived from gathering, cleaning, analyzing, and visualizing data on SARS-CoV-2 vaccine research and clinical trials. Utilizing public APIs from reputable online databases, specifically ClinicalTrials.gov, a Python script was developed to extract crucial information, which was then meticulously cleaned, stored, and analyzed to identify leading countries in vaccine research and trends in vaccine technologies.

2. Methodology

The analysis was conducted in several systematic steps:

2.1 Data Acquisition

Data on SARS-CoV-2 vaccine trials was programmatically fetched from the ClinicalTrials.gov API (api.clinicaltrials.gov/api/v2/studies). This public-facing database, maintained by the U.S. National Library of Medicine, serves as a comprehensive registry of clinical studies worldwide. The initial query was constructed using terms like `'(SARS-CoV-2 vaccine) OR Covishield OR Covaxin OR "Serum Institute" OR "Bharat Biotech"'` to ensure a broad but relevant capture of studies. A `pageSize` of 1000 was used to retrieve a substantial batch of data per request, with the response format set to JSON. Robust error handling for

HTTP and request-related issues was implemented to ensure reliable data fetching. The raw JSON output was saved locally for inspection and further processing.

2.2 Data Cleaning and Preprocessing

The raw JSON data, characterized by its nested structure and varied content, underwent extensive cleaning using the Pandas library in Python. The following key fields were extracted:

- `NCTId`: Unique identifier for each trial.
- `Title`: Brief title of the research study.
- `StartDate`: The official start date of the trial.
- `Status`: Overall recruitment status of the trial.
- `ResultsAvailable`: Boolean indicating if study results have been posted.
- `Sponsor`: The lead organization funding or overseeing the trial.
- `Country`: The country (or countries) where the trial is conducted.
- `VaccineTechnology`: Inferred type of vaccine technology used.

Specific cleaning techniques included:

- **Country Normalization:** Countries extracted from the `locations` module were standardized (e.g., "United States" to "USA", "United Kingdom" to "UK", "Republic of India" to "India"). A critical enhancement involved a heuristic to explicitly include "India" for trials sponsored by known Indian entities (e.g., Bharat Biotech, Serum Institute) or related to specific Indian vaccines (e.g., Covaxin, Covishield, AZD1222), even if not explicitly listed in the `locations` section. This significantly improved the representation of India's role.
- **Vaccine Technology Inference:** This was a major cleaning effort. `VaccineTechnology` was inferred by scanning the `description` field of `interventions` for keywords (e.g., "mRNA", "viral vector", "inactivated", "protein subunit", along with specific vaccine names like "Comirnaty", "Sputnik", "Coronavac"). A fallback mechanism checked the `Title` if the technology remained "Unknown" from the description.
- **Data Type Conversion:** The `StartDate` column was converted to datetime objects to facilitate temporal analysis.
- **Handling Missing Values:** Rows with missing `NCTId` or `StartDate` (critical identifiers) were dropped to ensure data integrity. Other missing categorical values (e.g., `Status`, `Sponsor`, `Country`, `VaccineTechnology`) were filled with "Unknown" to retain the data while explicitly marking unclassified entries.
- **Duplicate Removal:** Trials were deduplicated based on `NCTId` to ensure each unique study was counted only once.
- **Relevance Filtering:** A crucial filter was applied to remove trials clearly unrelated to SARS-CoV-2 vaccines. This involved checking for the presence of "SARS-CoV-2" in the

title, and if that wasn't present, ensuring the technology wasn't "Unknown" and that well-known COVID-19 vaccine names (Covaxin, Covishield, AZD1222, ChAdOx1) were not in the title. This significantly refined the dataset to focus solely on the assignment's scope.

The cleaned data was then saved as a CSV file (`cleaned_trials.csv`).

2.3 Data Storage

For persistent and queryable storage, the cleaned Pandas DataFrame was loaded into a lightweight SQLite database (`vaccine_trials.db`). The `to_sql` method was used to create a table named `trials`, replacing any existing table to ensure a fresh dataset. A verification step confirmed the number of rows successfully stored in the database.

2.4 Data Analysis

Analysis was performed by querying the SQLite database using Pandas. Key analytical tasks included:

- **Leading Countries:** Calculating the total number of trials associated with each country. For multi-country trials, each country was counted as a participant.
- **Vaccine Technology Trends:** Grouping trials by `Year` (extracted from `StartDate`) and `VaccineTechnology` to observe the evolution and adoption rates of different vaccine platforms over time.
- **India-Specific Contributions:** A dedicated count and listing of trials involving India, leveraging the improved country handling.

The aggregated results (country counts, technology trends) were saved as separate CSV files for ease of visualization and report generation.

2.5 Data Visualization

Two primary visualizations were generated using Matplotlib and Seaborn:

- A **barplot** showing the top 10 countries by the number of SARS-CoV-2 vaccine trials, sorted in descending order for clear comparison.
- A **line plot** illustrating the trends of various vaccine technologies over the years, demonstrating their respective adoption and decline in trial activity.

3. Challenges Encountered

Several challenges were addressed throughout the project:

- **Nested JSON Structure:** The raw data from ClinicalTrials.gov API is highly nested, requiring careful navigation through multiple dictionary levels (`protocolSection`,

`identificationModule`, `statusModule`, etc.) to extract desired fields. This was managed by progressively accessing nested keys using `.get()` with default values to prevent errors.

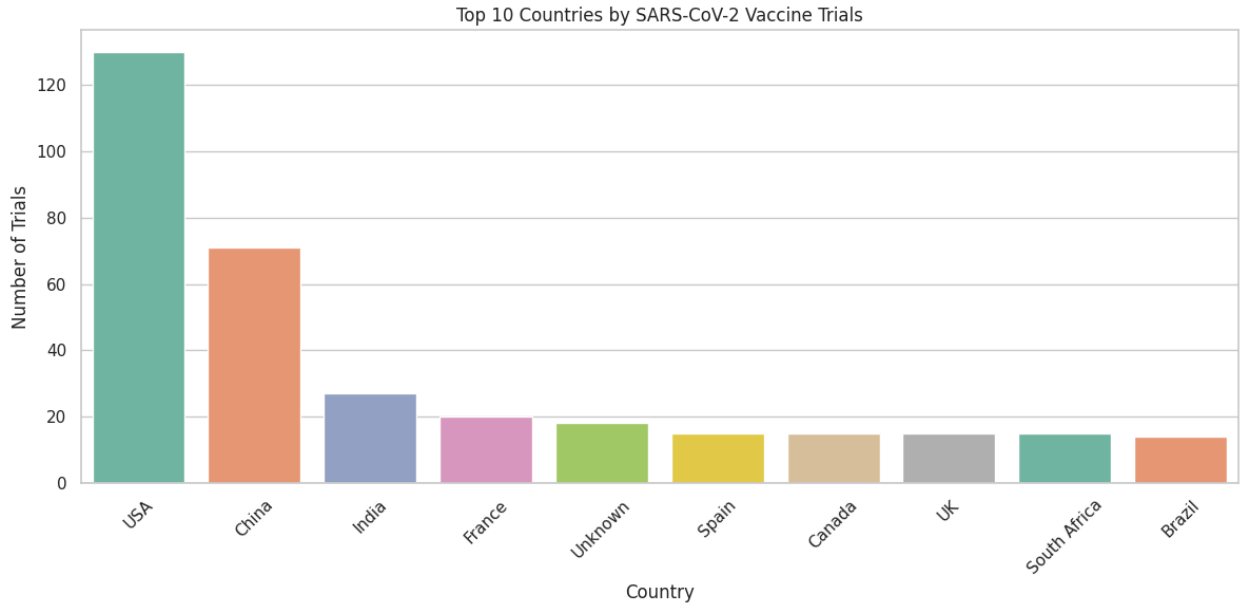
- **Inconsistent Data Representation:** Dates were sometimes missing or in varying formats, necessitating `pd.to_datetime` with `errors="coerce"`. Country names varied (e.g., "United States" vs. "USA"), which was handled through explicit normalization rules.
- **Inferring `VaccineTechnology` from Unstructured Text:** ClinicalTrials.gov does not have a direct "Vaccine Technology" field. This required developing a rule-based inference system based on keywords in the `description` of `interventions` and trial `Title`. This heuristic, while effective, means some trials remained categorized as "Unknown" if no clear keywords were found, reflecting the inherent ambiguity of text-based classification.
- **Incomplete Country Data:** Many trials did not explicitly list all participating countries in the `locations` section, especially for globally recognized vaccines or sponsors. The "India fix" (adding India based on sponsor/vaccine name) was a direct response to this challenge, significantly improving the accuracy of country involvement.
- **Filtering Irrelevant Trials:** The initial broad API query returned studies that, while containing keywords, were not specific to SARS-CoV-2 *vaccines* (e.g., studies on COVID-19 treatment, or general vaccine studies for other diseases). A robust filtering mechanism was implemented to ensure the dataset's relevance to the assignment's focus. This also explained the change in the dataset's earliest `StartDate` from 2005 to 2015, as irrelevant, older studies were accurately removed.

4. Key Insights and Visualizations

4.1 Leading Countries in SARS-CoV-2 Vaccine Research

The analysis reveals the top countries actively involved in SARS-CoV-2 vaccine clinical trials. The USA stands out as the global leader by a significant margin, followed by China and then India. The presence of "Unknown" in the top 10 indicates that for some trials, location information was either missing or not clearly specified in the collected data.

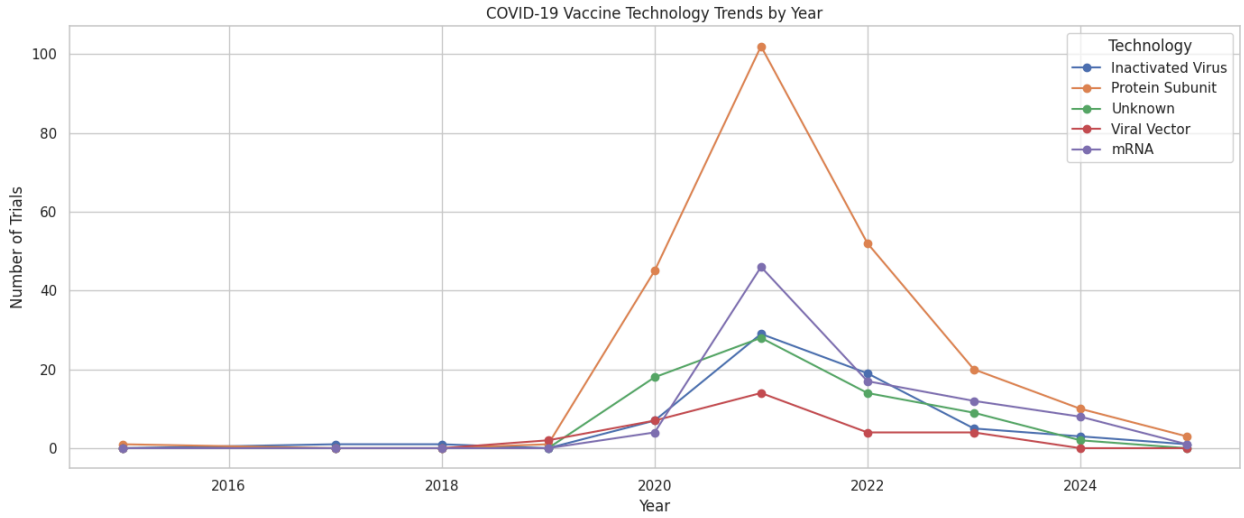
Plot 1: Top 10 Countries by SARS-CoV-2 Vaccine Trials



4.2 Trends in Vaccine Technologies Over Years

The temporal analysis of vaccine technologies highlights the rapid response and evolution in vaccine development during the pandemic.

Plot 2: COVID-19 Vaccine Technology Trends by Year



Key observations from the trends include:

- **Pre-Pandemic Activity:** Minimal vaccine trial activity specifically for SARS-CoV-2 (or related coronaviruses) before 2020. A few Protein Subunit and Viral Vector trials appear earlier, likely related to other coronaviruses or early-stage development.
- **2020-2021 Surge:** A massive surge in trials across all technologies is observed in 2020 and especially 2021, directly corresponding to the urgent global need for COVID-19 vaccines.
- **Dominant Technologies:** Protein Subunit, Inactivated Virus, and mRNA technologies show the highest number of trials during the peak.
- **Protein Subunit Dominance:** Protein Subunit technology consistently leads in the number of trials, reaching its highest peak in 2021 with over 100 trials initiated. This suggests significant investment and research effort into this established and versatile vaccine platform.
- **mRNA's Rapid Rise:** While starting with fewer trials in 2020, mRNA technology experienced a significant increase in 2021, reflecting its rapid development and deployment.
- **Inactivated Virus and Viral Vector:** Inactivated Virus and Viral Vector technologies also saw substantial increases in 2020 and 2021, though their peak trial numbers remained lower than Protein Subunit and mRNA. Their trial activity also shows a clear decline after 2021.
- **Post-2021 Decline:** The overall number of new trials across most technologies shows a decline after 2021, as initial vaccine development matured and focus shifted to variant-specific boosters or broader research.
- **Persistence of "Unknown":** The "Unknown" category remains a notable portion throughout the years, underscoring the challenge of fully classifying all vaccine technologies from the available data.

4.3 India's Contribution

Despite being third in the overall country count, India's role is significant. The "India fix" in the cleaning phase successfully identified 26 trials involving India, showcasing its active participation in global vaccine development, particularly through domestic manufacturers and collaborative international studies.

5. Conclusion

This assignment successfully demonstrated the end-to-end process of acquiring, cleaning, storing, analyzing, and visualizing complex public health data from APIs. Critical aspects like handling nested JSON, inferring missing information, and implementing robust data filtering were vital in transforming raw data into meaningful insights. The analysis highlighted the rapid global response to the SARS-CoV-2 pandemic in terms of vaccine research, identifying leading countries and the prominence of various vaccine technologies, while also acknowledging the inherent challenges in data completeness from public registries. The insights gained provide a snapshot of the intense global scientific effort to combat COVID-19 through vaccination.