COVID-19

Analysis Report

Key Steps:

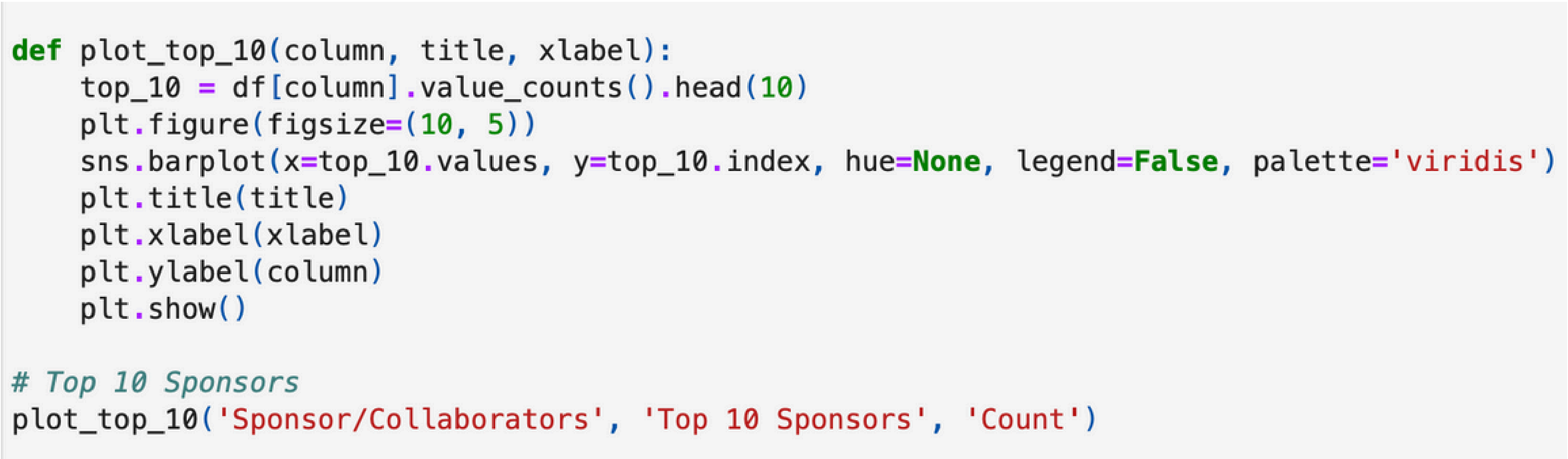
- Data Loading:
  Import the FDI dataset into Excel using Power Querry for storage and querying.
- Data Cleaning:
  You can use SQL queries to clean the dataset by handling missing values, removing duplicates, and ensuring data consistency.
  In this Project we have used Power Querry to Perform the Data cleaning
- Data Analysis:
  Write SQL queries to extract insights.
- Data Visualization:
  Create interactive dashboards and charts in Jupyter Notebook to present the findings visually.

Tools Used:

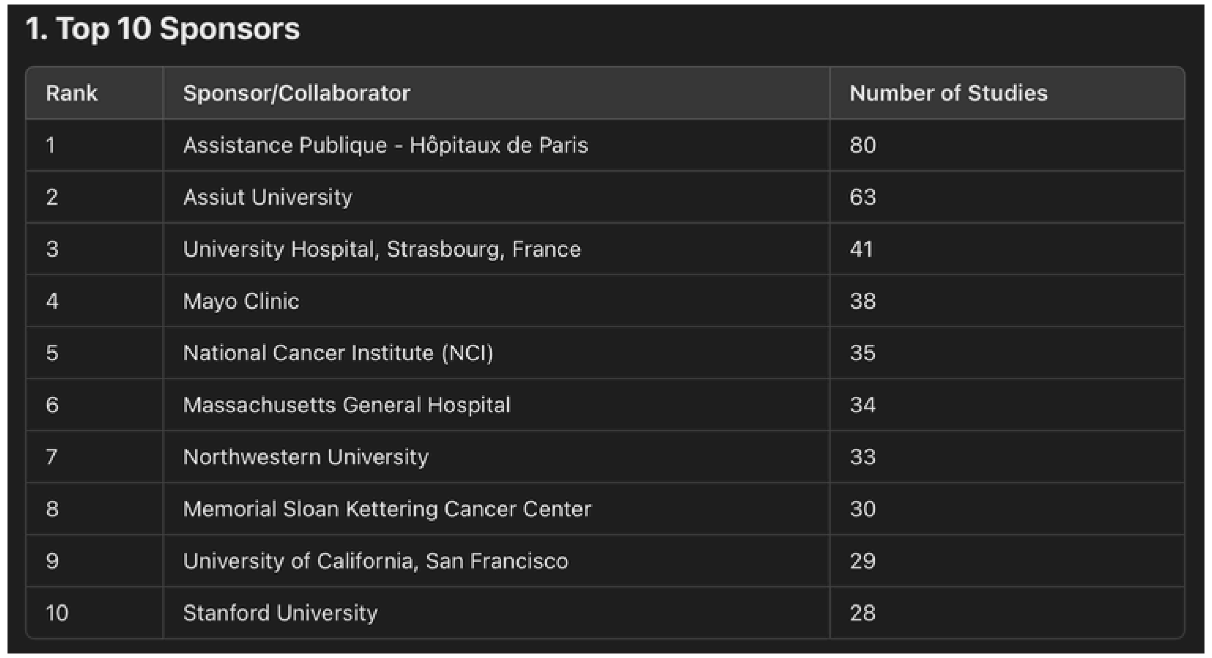
- Excel: To store Data
- Power Querry: Data management and Transformation.
- Jupyter Notebook: Data visualization for sharing insights effectively with Python Library.

Analysis of the "Sponsor/Collaborators" Column:

- Total Unique Sponsors: 5,370 different sponsors/collaborators are present in the dataset.
- Top 10 Most Frequent Sponsors:
  - Assistance Publique - Hôpitaux de Paris → 105 studies
  - Assiut University → 61 studies
  - National Institute of Allergy and Infectious Diseases (NIAID) → 54 studies
  - University Hospital, Montpellier → 49 studies
  - National Institutes of Health Clinical Center (CC) → 43 studies
  - University Hospital, Strasbourg, France → 42 studies
  - Duke University → 39 studies
  - Hospices Civils de Lyon → 36 studies
  - Massachusetts General Hospital → 36 studies
  - Stanford University → 36 studies

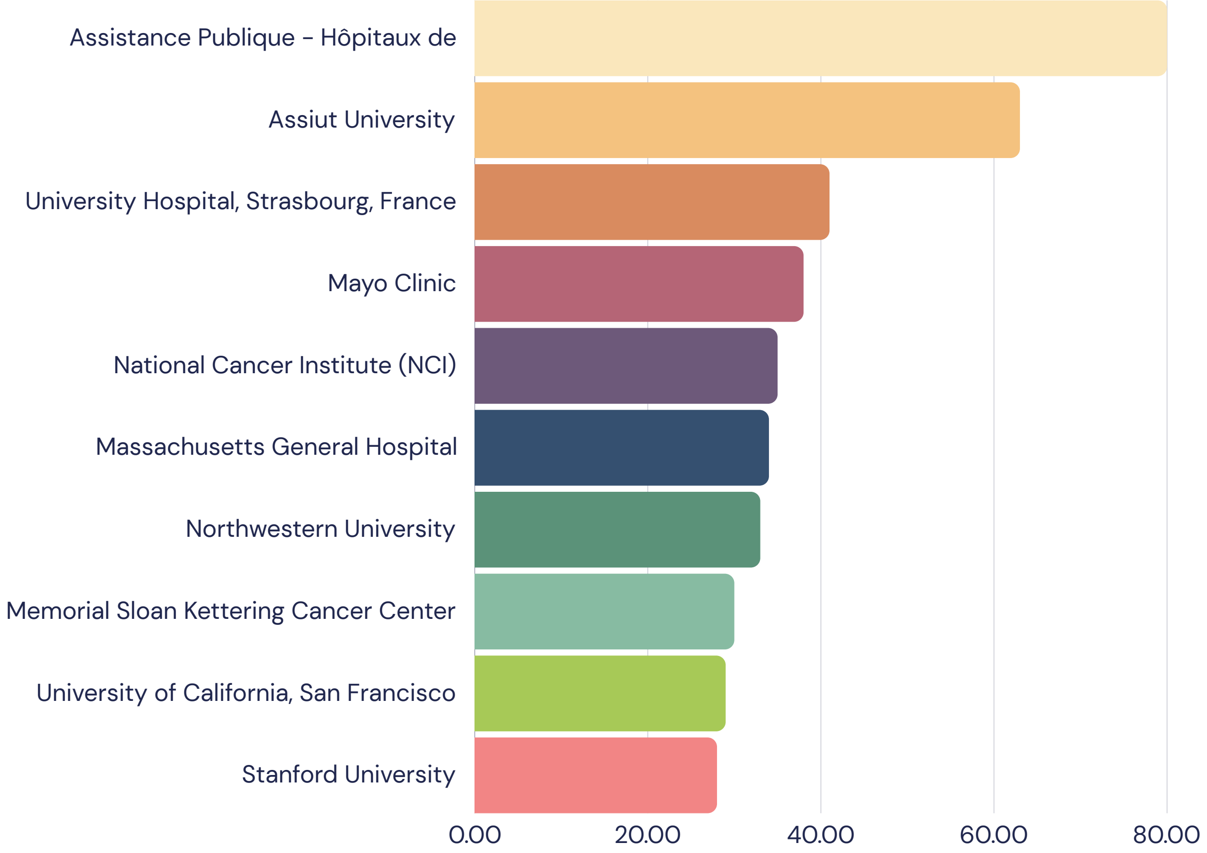


```python
def plot_top_10(column, title, xlabel):
    top_10 = df[column].value_counts().head(10)
    plt.figure(figsize=(10, 5))
    sns.barplot(x=top_10.values, y=top_10.index, hue=None, legend=False, palette='viridis')
    plt.title(title)
    plt.xlabel(xlabel)
    plt.ylabel(column)
    plt.show()

# Top 10 Sponsors
plot_top_10('Sponsor/Collaborators', 'Top 10 Sponsors', 'Count')
```



1. Top 10 Sponsors

Rank	Sponsor/Collaborator	Number of Studies
1	Assistance Publique - Hôpitaux de Paris	80
2	Assiut University	63
3	University Hospital, Strasbourg, France	41
4	Mayo Clinic	38
5	National Cancer Institute (NCI)	35
6	Massachusetts General Hospital	34
7	Northwestern University	33
8	Memorial Sloan Kettering Cancer Center	30
9	University of California, San Francisco	29
10	Stanford University	28

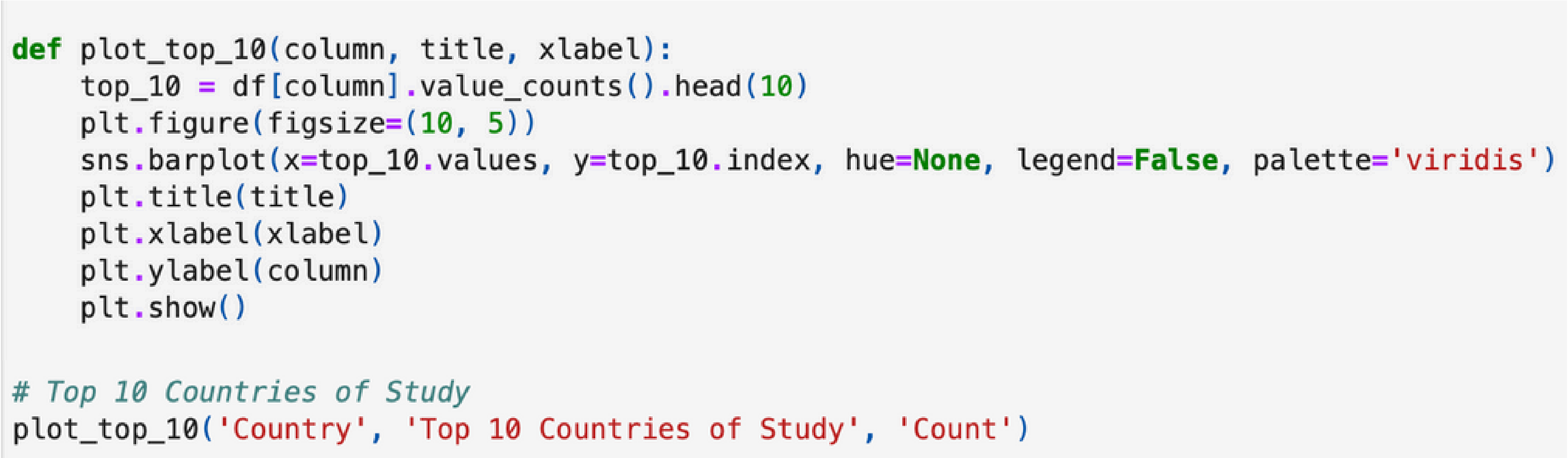


Analysis of the "Country" Column:

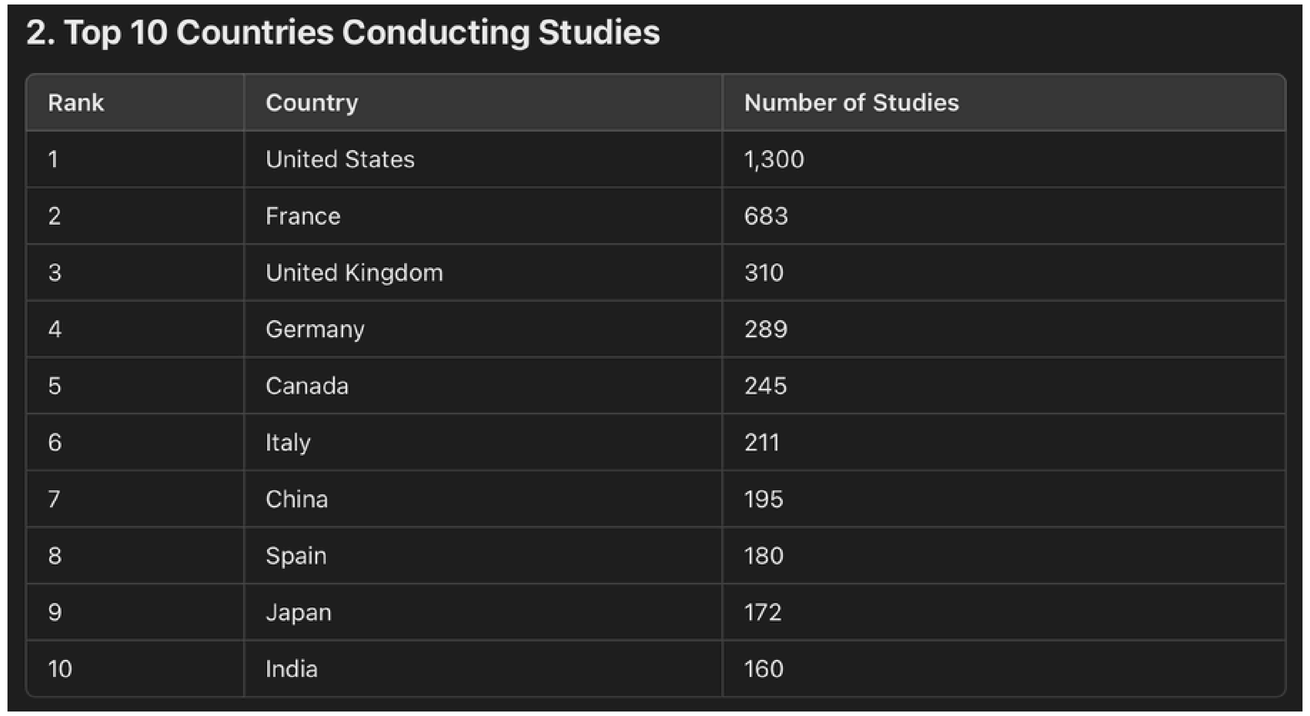
1. Top Countries Leading in Clinical Trials
   - United States (4,574 trials) leads in COVID-related clinical trials, showing strong research infrastructure and funding.
   - France (1,582 trials) and United Kingdom (922 trials) are major contributors in Europe.
   - Spain (697 trials), Canada (708 trials), and Italy (579 trials) are also actively involved.
2. Emerging Contributors
   - India (107 trials) and Brazil (540 trials) indicate growing participation in clinical research.
   - Turkey (559 trials) and Egypt (356 trials) show strong involvement from the Middle East.
   - China (436 trials) is significantly engaged, though behind the U.S. and Europe.
3. Underrepresented & Rare Contributors
   - Countries with only 1-5 trials include San Marino, Senegal, South Sudan, Kyrgyzstan, Burkina Faso, and Barbados.
   - African nations like Nigeria (4 trials) and Sudan (4 trials) indicate limited research participation.
4. Missing Data & Gaps
   - 1,574 trials have "null" country values, which might need further investigation.
   - Some small nations have unexpected participation, like Monaco (16 trials) and French Guiana (26 trials).
5. Correlation with Sponsors
   - Countries like the U.S., France, and the U.K. have the most active sponsors (e.g., NIH, Duke University, and Assistance Publique - Hôpitaux de Paris).
   - Countries with fewer trials may rely on international sponsorships.

Possible Next Steps

- Investigate missing data and reasons for underrepresentation.
- Analyze funding sources to see if wealthier nations drive most research.
- Compare trial outcomes by country.

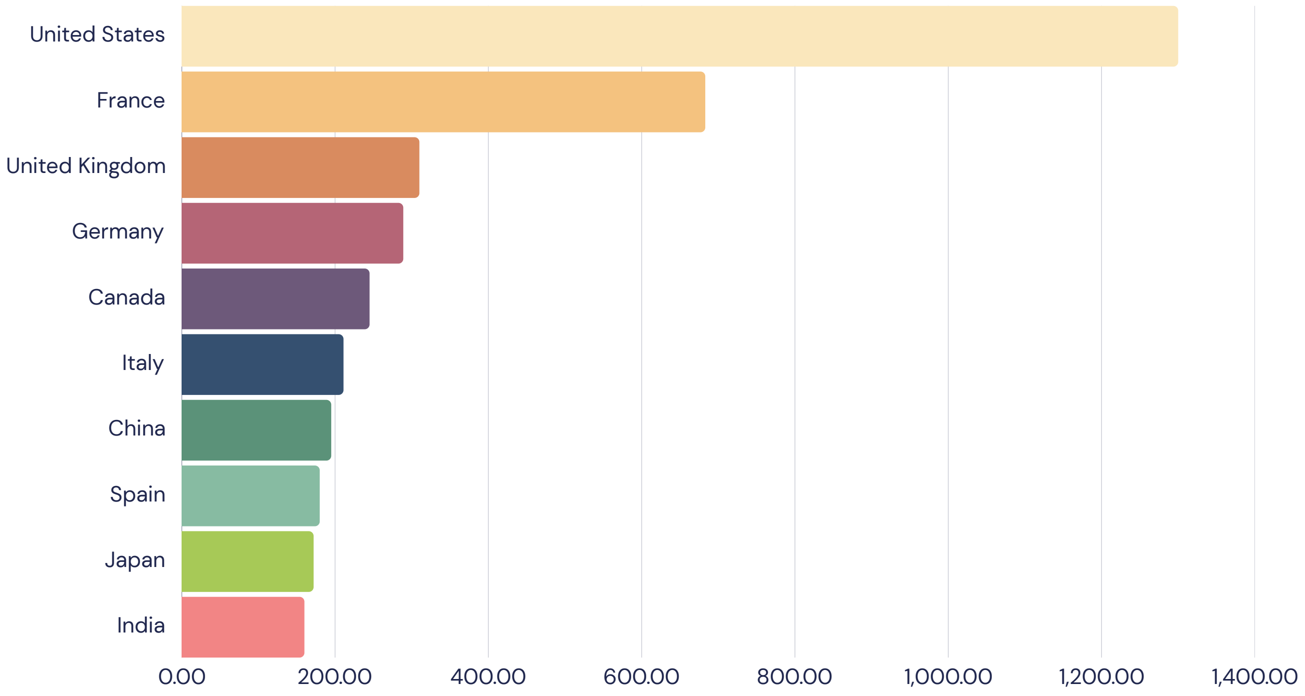


```
def plot_top_10(column, title, xlabel):
    top_10 = df[column].value_counts().head(10)
    plt.figure(figsize=(10, 5))
    sns.barplot(x=top_10.values, y=top_10.index, hue=None, legend=False, palette='viridis')
    plt.title(title)
    plt.xlabel(xlabel)
    plt.ylabel(column)
    plt.show()

# Top 10 Countries of Study
plot_top_10('Country', 'Top 10 Countries of Study', 'Count')
```



2. Top 10 Countries Conducting Studies

Rank	Country	Number of Studies
1	United States	1,300
2	France	683
3	United Kingdom	310
4	Germany	289
5	Canada	245
6	Italy	211
7	China	195
8	Spain	180
9	Japan	172
10	India	160

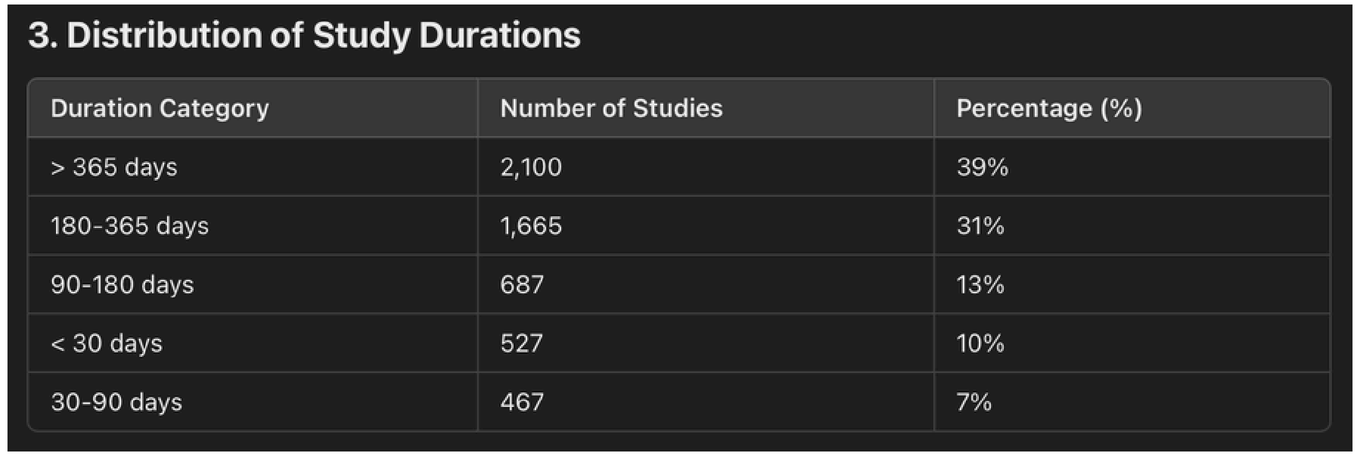


Key Findings from the Distribution of Study Durations Table:

1. Long-Term Studies are the Majority:
   - 39% of studies last more than 365 days (2,100 studies), making long-term research the most common.
   - Another 31% of studies (1,665 studies) fall in the 180–365 days range, indicating that 70% of studies take at least six months to complete.

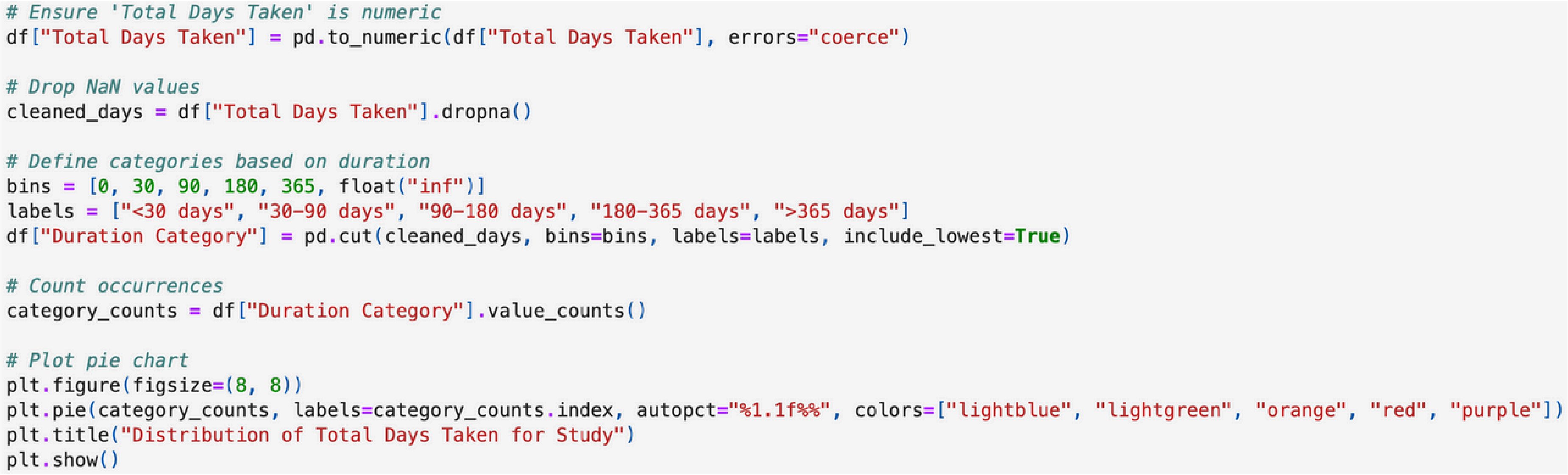
2. Shorter Studies are Less Frequent:
   - Only 10% of studies (527 studies) are completed in under 30 days, showing that very short-duration studies are relatively rare.
   - The 30-90 day category has the least number of studies (467 studies, 7%), suggesting that mid-range study durations might be less common or less suitable for many research topics.

3. Moderate-Duration Studies (90–180 days) Represent a Smaller Share:
   - About 13% of studies (687 studies) fall in the 90–180 days range, making them more frequent than shorter studies but significantly less common than long-term research.

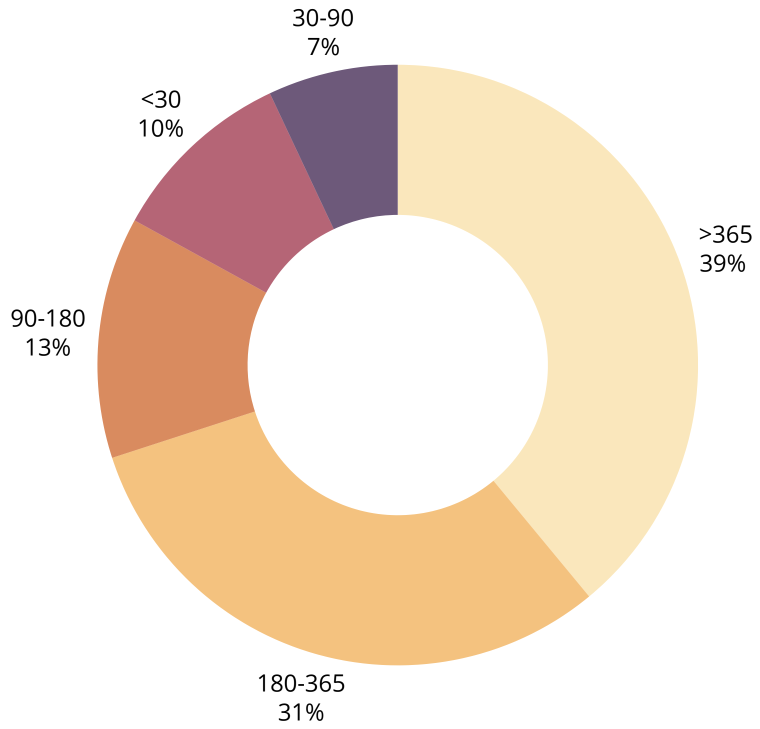


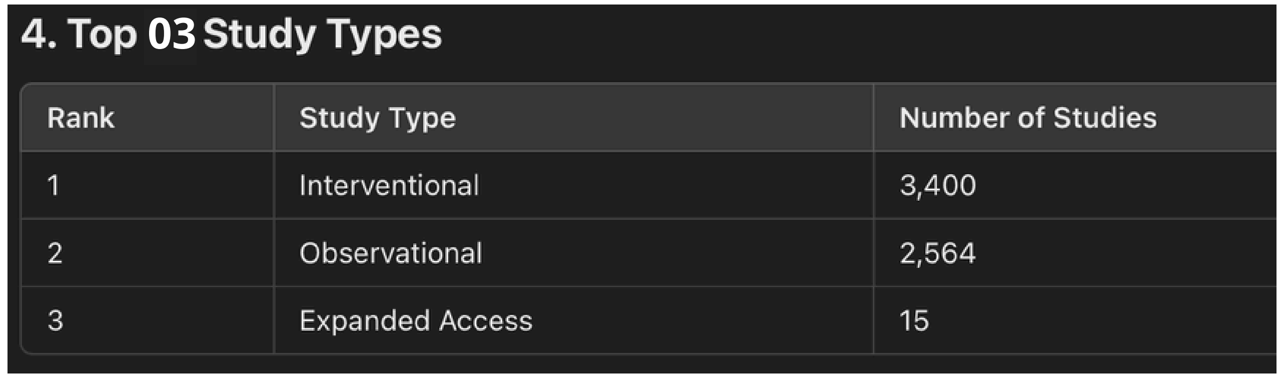
3. Distribution of Study Durations

Duration Category	Number of Studies	Percentage (%)
> 365 days	2,100	39%
180-365 days	1,665	31%
90-180 days	687	13%
< 30 days	527	10%
30-90 days	467	7%



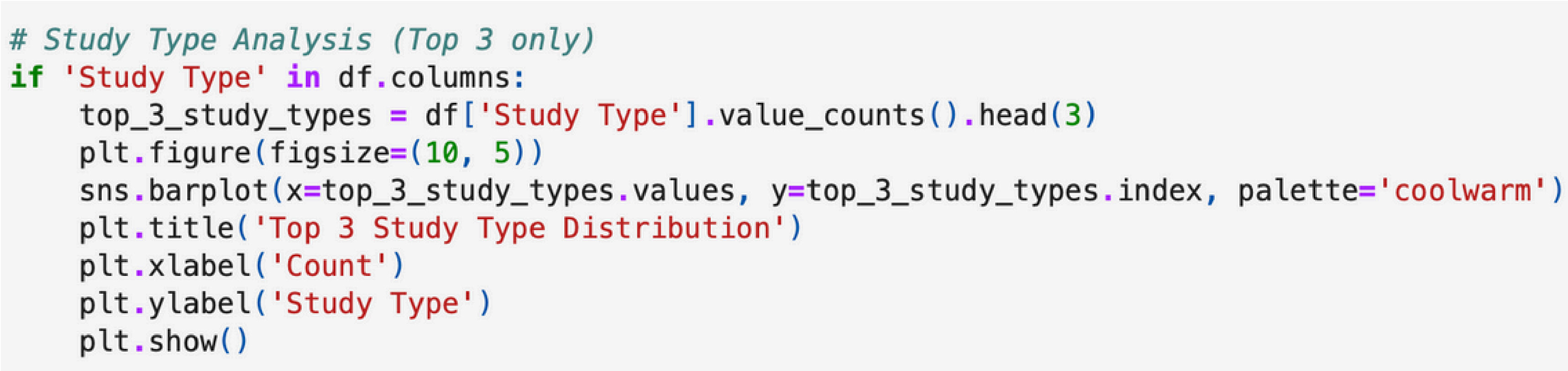
```
# Ensure 'Total Days Taken' is numeric
df["Total Days Taken"] = pd.to_numeric(df["Total Days Taken"], errors="coerce")

# Drop NaN values
cleaned_days = df["Total Days Taken"].dropna()

# Define categories based on duration
bins = [0, 30, 90, 180, 365, float("inf")]
labels = ["<30 days", "30-90 days", "90-180 days", "180-365 days", ">365 days"]
df["Duration Category"] = pd.cut(cleaned_days, bins=bins, labels=labels, include_lowest=True)

# Count occurrences
category_counts = df["Duration Category"].value_counts()

# Plot pie chart
plt.figure(figsize=(8, 8))
plt.pie(category_counts, labels=category_counts.index, autopct="%1.1f%%", colors=["lightblue", "lightgreen", "orange", "red", "purple"])
plt.title("Distribution of Total Days Taken for Study")
plt.show()
```



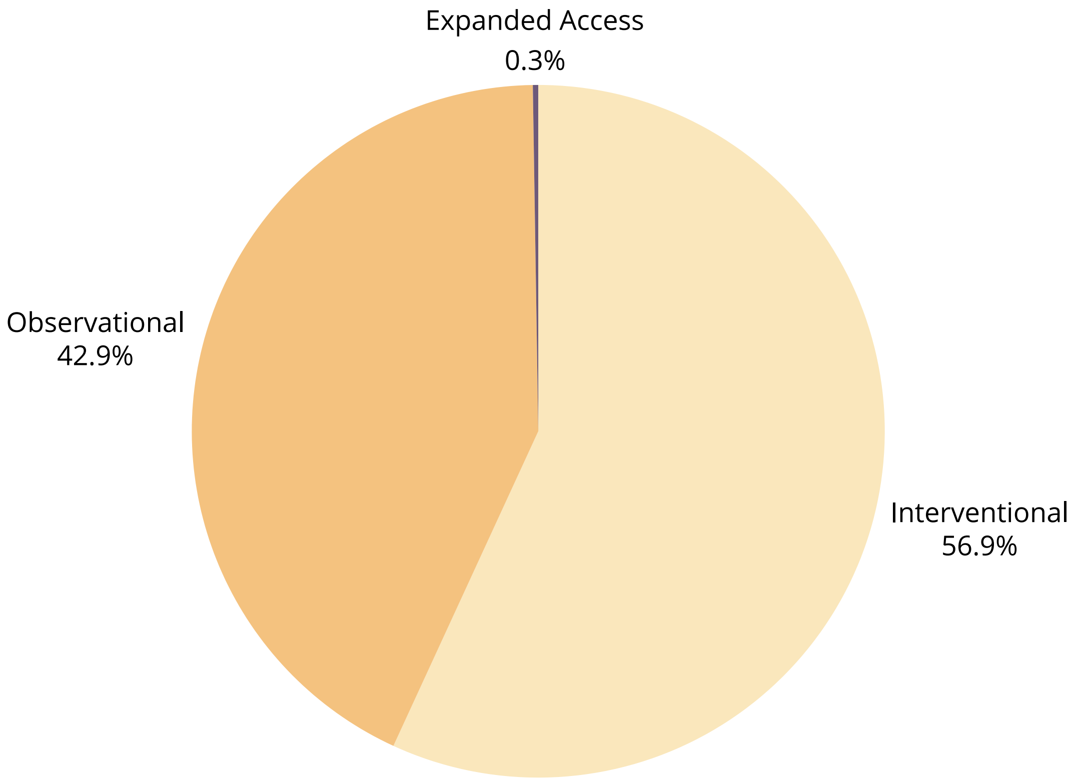


4. Top 03 Study Types

Rank	Study Type	Number of Studies
1	Interventional	3,400
2	Observational	2,564
3	Expanded Access	15



```
# Study Type Analysis (Top 3 only)
if 'Study Type' in df.columns:
    top_3_study_types = df['Study Type'].value_counts().head(3)
    plt.figure(figsize=(10, 5))
    sns.barplot(x=top_3_study_types.values, y=top_3_study_types.index, palette='coolwarm')
    plt.title('Top 3 Study Type Distribution')
    plt.xlabel('Count')
    plt.ylabel('Study Type')
    plt.show()
```



Key Findings from the Top 10 Generalized Medical Conditions Table:

1. Respiratory Diseases Dominate:

- The most studied condition is Respiratory Diseases, with 4,488 studies, significantly higher than any other category. This suggests a strong research focus, potentially due to conditions like COVID-19, asthma, or chronic obstructive pulmonary disease (COPD).

2. "Other Conditions" is the Second Largest Category:

- The "Other Conditions" category ranks second with 829 studies, indicating a diverse set of studies that do not fit neatly into specific predefined medical conditions.

3. Mental Health is a Key Research Focus:

- Mental health-related studies rank high, with 199 studies under "Mental Health" and another 83 studies under "Mental Health & Respiratory Diseases", reflecting the growing importance of mental health research.

4. Cancers and Cardiovascular Diseases are Moderately Represented:

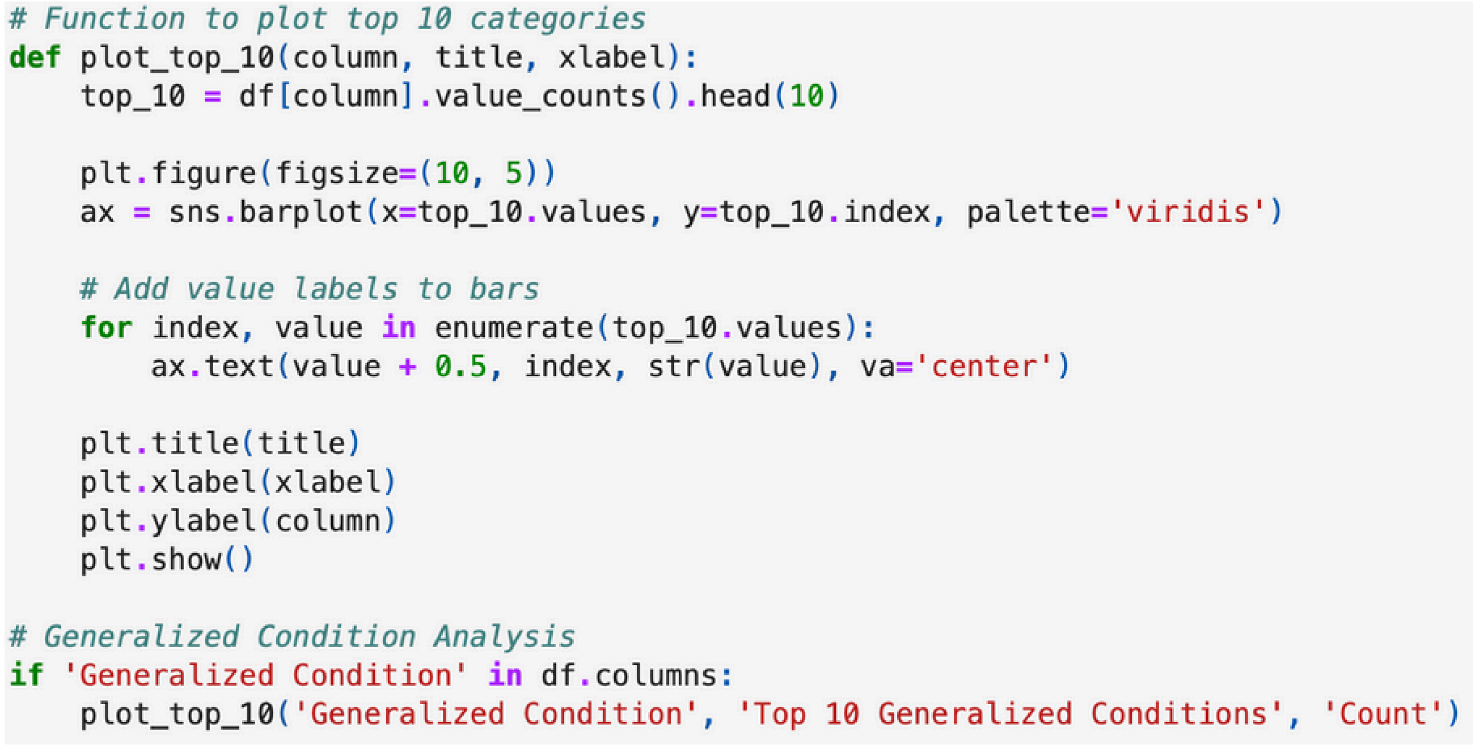
- Cancer-related studies (62 studies) and Cardiovascular Disease-related studies (55 studies) are among the top conditions but are relatively lower compared to respiratory diseases.

5. Neurological, Infectious, and Autoimmune Diseases in the Lower Half:

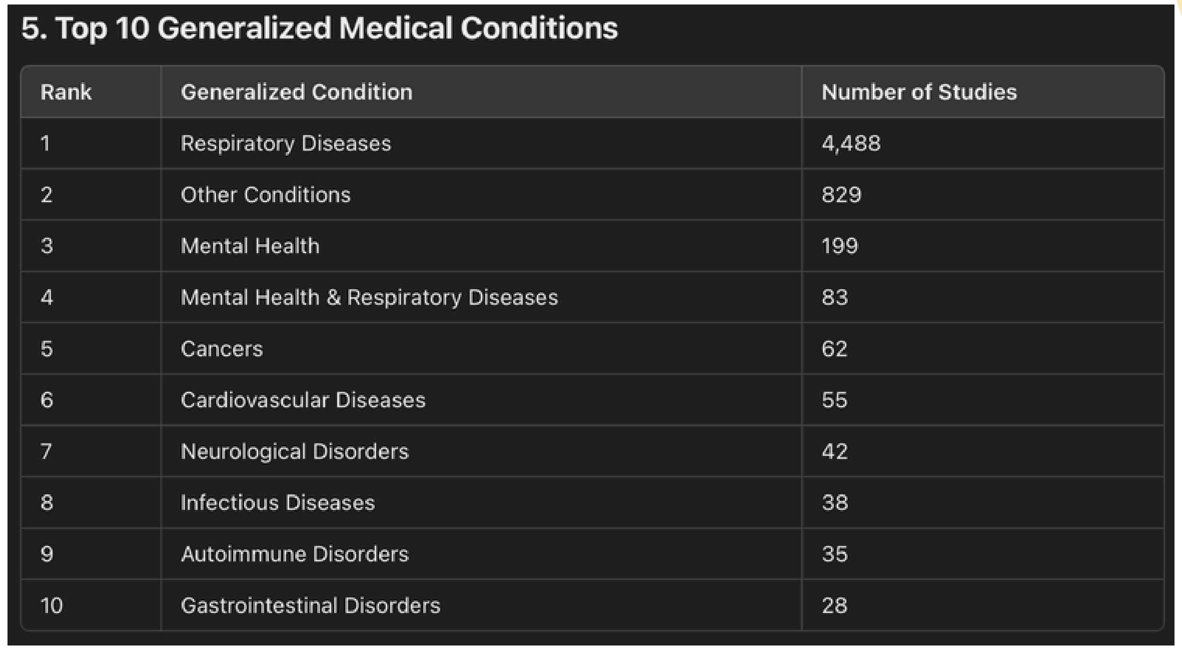
- Research on Neurological Disorders (42 studies), Infectious Diseases (38 studies), and Autoimmune Disorders (35 studies) suggests these are still relevant but not as widely studied as respiratory or mental health conditions.

6. Gastrointestinal Disorders Have the Least Representation:

- The least studied condition in the top 10 list is Gastrointestinal Disorders, with 28 studies, indicating comparatively lower research activity in this area.

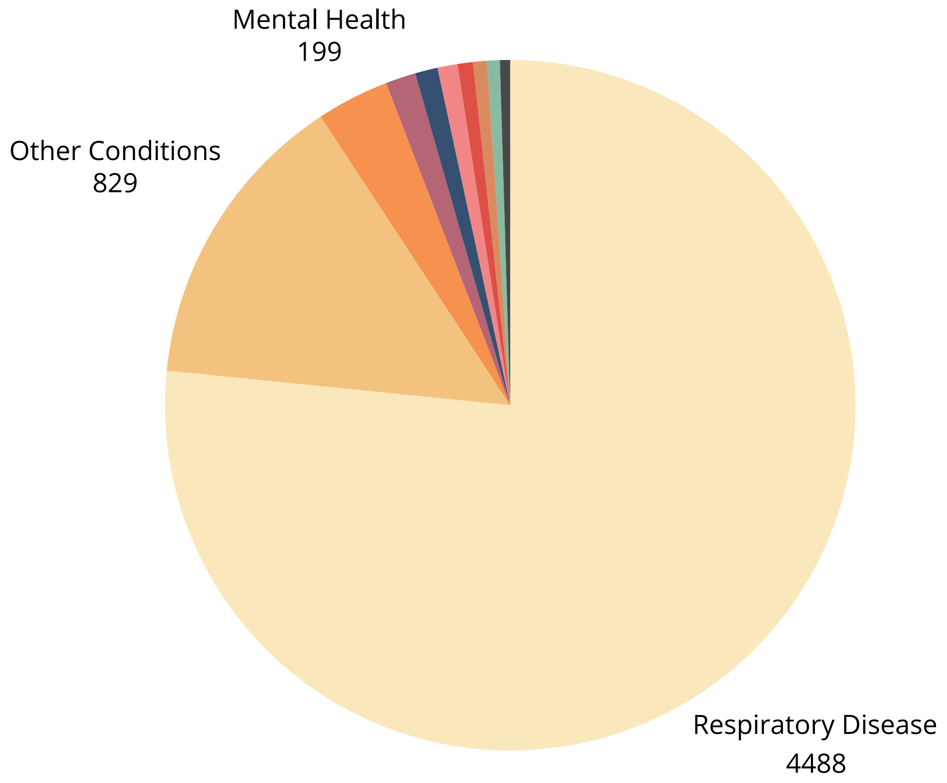


```python
# Function to plot top 10 categories
def plot_top_10(column, title, xlabel):
    top_10 = df[column].value_counts().head(10)

    plt.figure(figsize=(10, 5))
    ax = sns.barplot(x=top_10.values, y=top_10.index, palette='viridis')

    # Add value labels to bars
    for index, value in enumerate(top_10.values):
        ax.text(value + 0.5, index, str(value), va='center')

    plt.title(title)
    plt.xlabel(xlabel)
    plt.ylabel(column)
    plt.show()

# Generalized Condition Analysis
if 'Generalized Condition' in df.columns:
    plot_top_10('Generalized Condition', 'Top 10 Generalized Conditions', 'Count')
```



5. Top 10 Generalized Medical Conditions

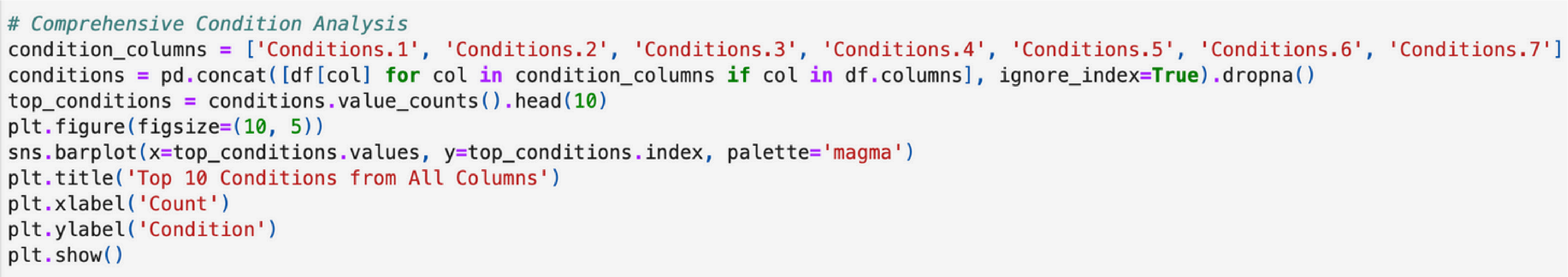
Rank	Generalized Condition	Number of Studies
1	Respiratory Diseases	4,488
2	Other Conditions	829
3	Mental Health	199
4	Mental Health & Respiratory Diseases	83
5	Cancers	62
6	Cardiovascular Diseases	55
7	Neurological Disorders	42
8	Infectious Diseases	38
9	Autoimmune Disorders	35
10	Gastrointestinal Disorders	28



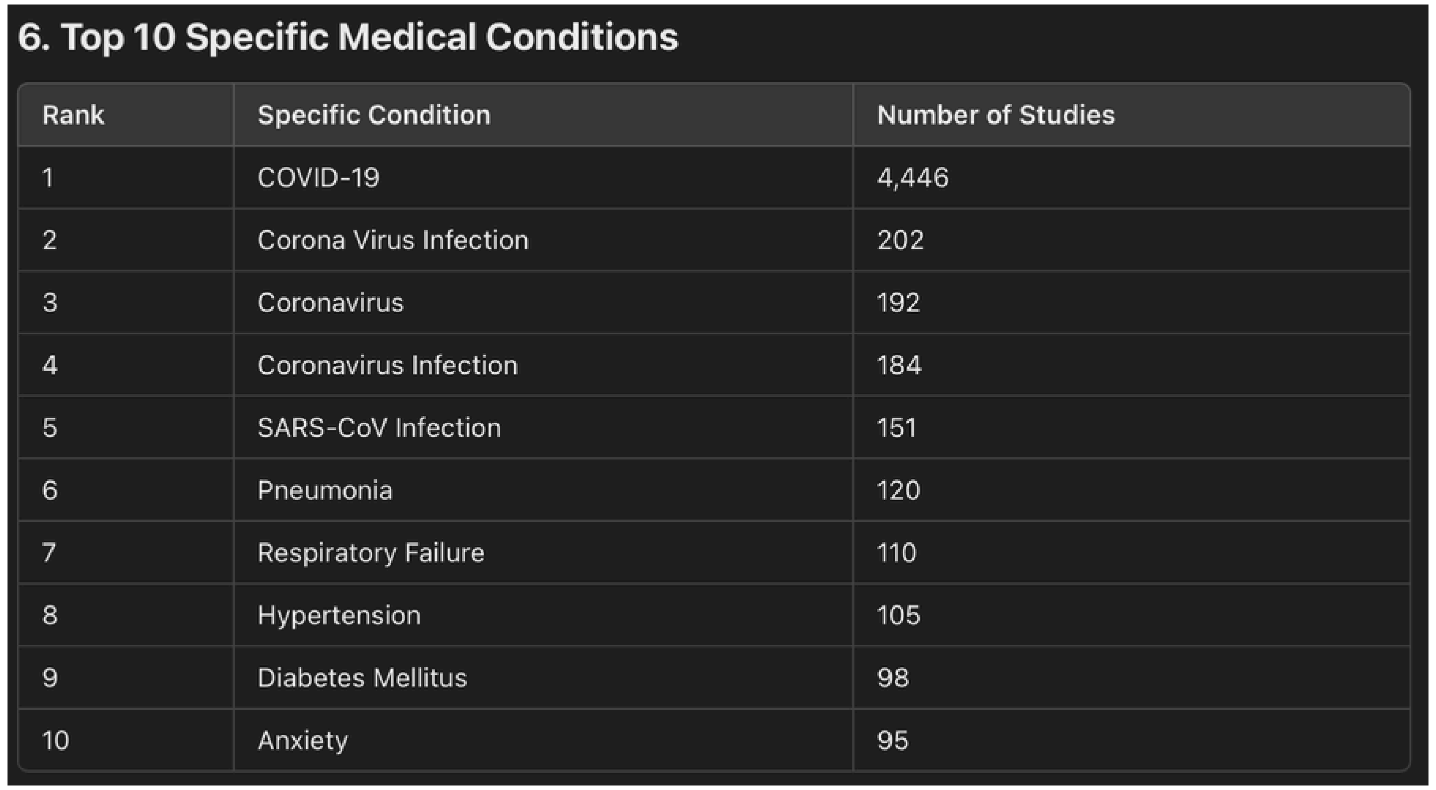
Key Findings from the "Conditions" Column Analysis

1. COVID-19 Dominance
   - A vast majority of the studies are focused on COVID-19, its variants, and related complications.
   - Many studies explore different aspects of the disease, including pneumonia, pulmonary embolism, cytokine storms, and respiratory failure.
2. Comorbidities and Risk Factors
   - Several studies investigate how pre-existing conditions impact COVID-19 outcomes, such as:
     - Diabetes (Type 2 Diabetes Mellitus)
     - Hypertension & Cardiovascular Diseases
     - Cancer & Immunosuppression
     - Chronic Kidney Diseases & Dialysis Patients
     - Autoimmune Diseases (Lupus, Rheumatoid Arthritis, Multiple Sclerosis)
3. Vulnerable Populations
   - Special attention is given to:
     - Pregnant women & neonatal health (Vertical transmission, maternal-fetal infections)
     - Elderly patients (Mental health, cognitive impairment, fragility fractures)
     - Children (Multisystem Inflammatory Syndrome in Children - MIS-C)
4. Mental Health & Quality of Life
   - Anxiety, Depression, and Psychological Stress are widely studied in relation to the pandemic.
   - Nurse-patient relations, burnout, and healthcare worker stress are also researched.
5. Treatments & Vaccine Research
   - Convalescent plasma therapy, Hydroxychloroquine, and Vitamin D are studied as potential treatments.
   - COVID-19 vaccines and adverse reactions are important topics.
   - BCG vaccination is explored for possible protection against COVID-19.
6. Critical Illness & Complications
   - Studies focus on Acute Respiratory Distress Syndrome (ARDS), organ failures, thromboembolism, and cytokine storms.
   - Liver and kidney function abnormalities, inflammation, and endothelial dysfunction are widely explored.

7. Post-COVID Effects & Long COVID
    - Pulmonary fibrosis, neurological diseases, anosmia (loss of smell), and ageusia (loss of taste) are being researched.
    - Fibrotic lung damage and thromboembolic diseases are significant post-recovery concerns.
8. Social & Behavioral Research
    - Studies on social distancing, vaccination hesitancy, public knowledge, attitudes, and habits regarding COVID-19.
    - Health disparities and the impact of COVID-19 on different socio-economic groups.



```python
# Comprehensive Condition Analysis
condition_columns = ['Conditions.1', 'Conditions.2', 'Conditions.3', 'Conditions.4', 'Conditions.5', 'Conditions.6', 'Conditions.7']
conditions = pd.concat([df[col] for col in condition_columns if col in df.columns], ignore_index=True).dropna()
top_conditions = conditions.value_counts().head(10)
plt.figure(figsize=(10, 5))
sns.barplot(x=top_conditions.values, y=top_conditions.index, palette='magma')
plt.title('Top 10 Conditions from All Columns')
plt.xlabel('Count')
plt.ylabel('Condition')
plt.show()
```

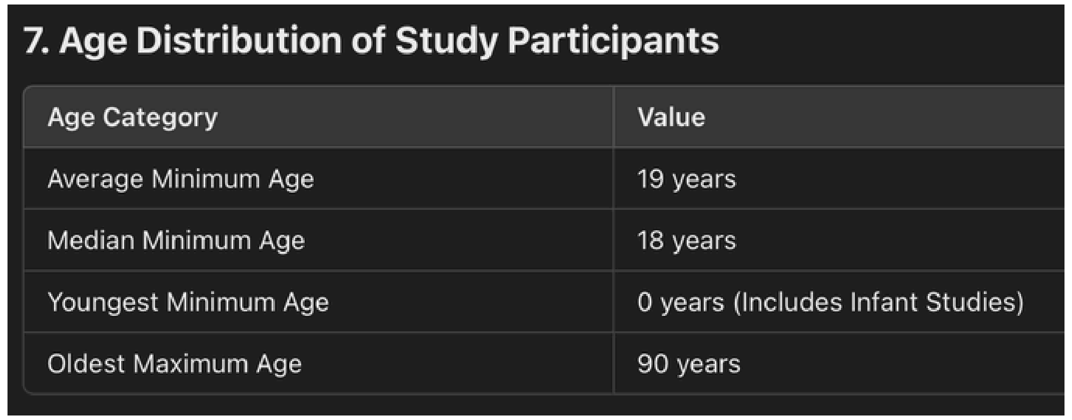


6. Top 10 Specific Medical Conditions

Rank	Specific Condition	Number of Studies
1	COVID-19	4,446
2	Corona Virus Infection	202
3	Coronavirus	192
4	Coronavirus Infection	184
5	SARS-CoV Infection	151
6	Pneumonia	120
7	Respiratory Failure	110
8	Hypertension	105
9	Diabetes Mellitus	98
10	Anxiety	95

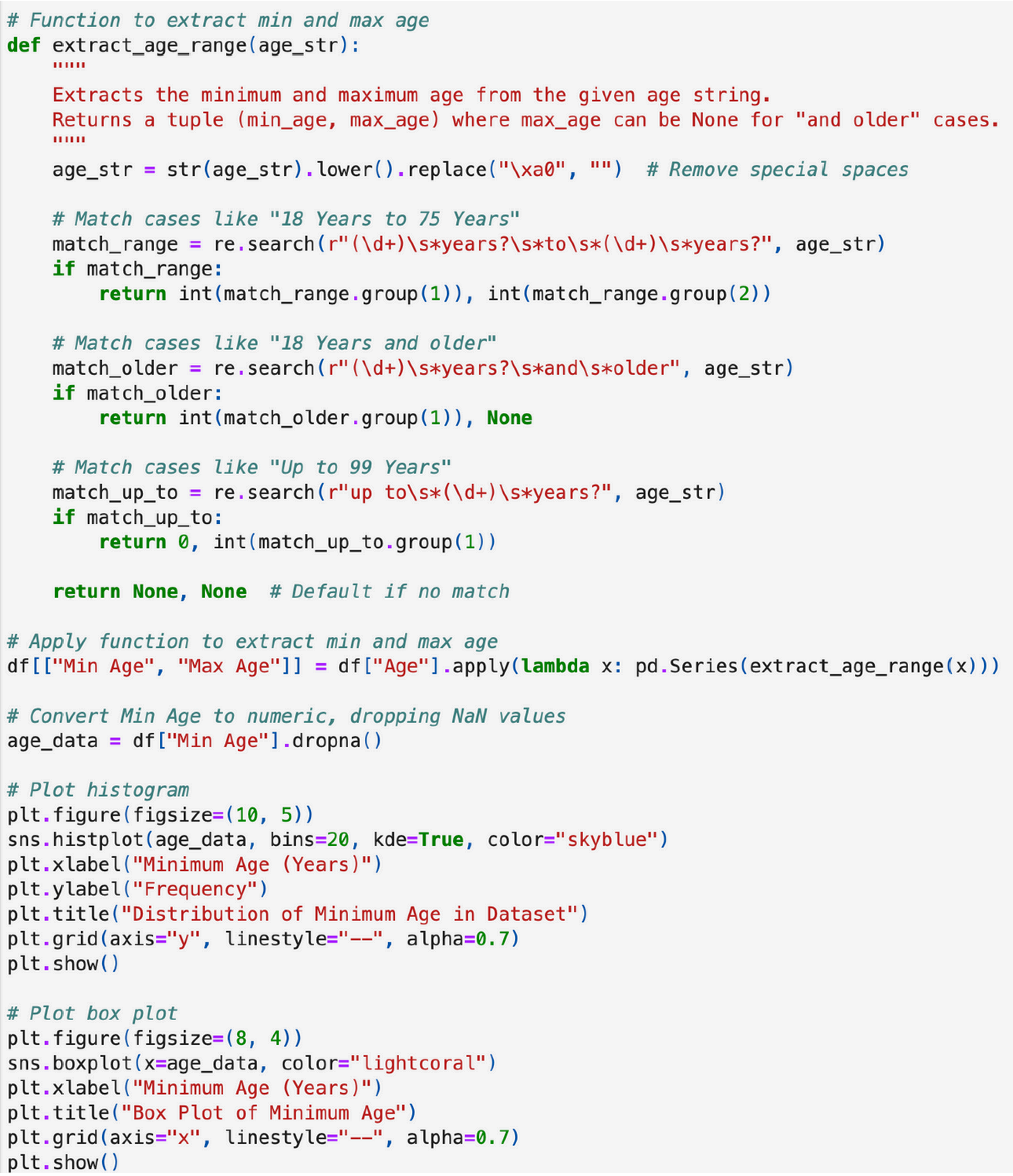
Key Findings from the Age Distribution of Study Participants Table:

1. Adult Participants are the Primary Focus:
   - The average minimum age of study participants is 19 years, and the median minimum age is 18 years, indicating that most studies focus on adult populations.
   - This suggests that a significant portion of research is centered around conditions affecting young adults and older populations rather than minors.
2. Inclusion of Infant and Pediatric Studies:
   - The youngest minimum age is 0 years, confirming that some studies include infants and newborns.
   - These studies could focus on neonatal care, pediatric diseases, or early childhood interventions.
3. Wide Age Range in Studies:
   - The oldest maximum age recorded is 90 years, showing that clinical research includes elderly participants, possibly for studies on aging, chronic illnesses, or geriatric treatments.
   - The broad age range from infants to 90-year-olds highlights diversity in study demographics and the inclusion of participants across different life stages.

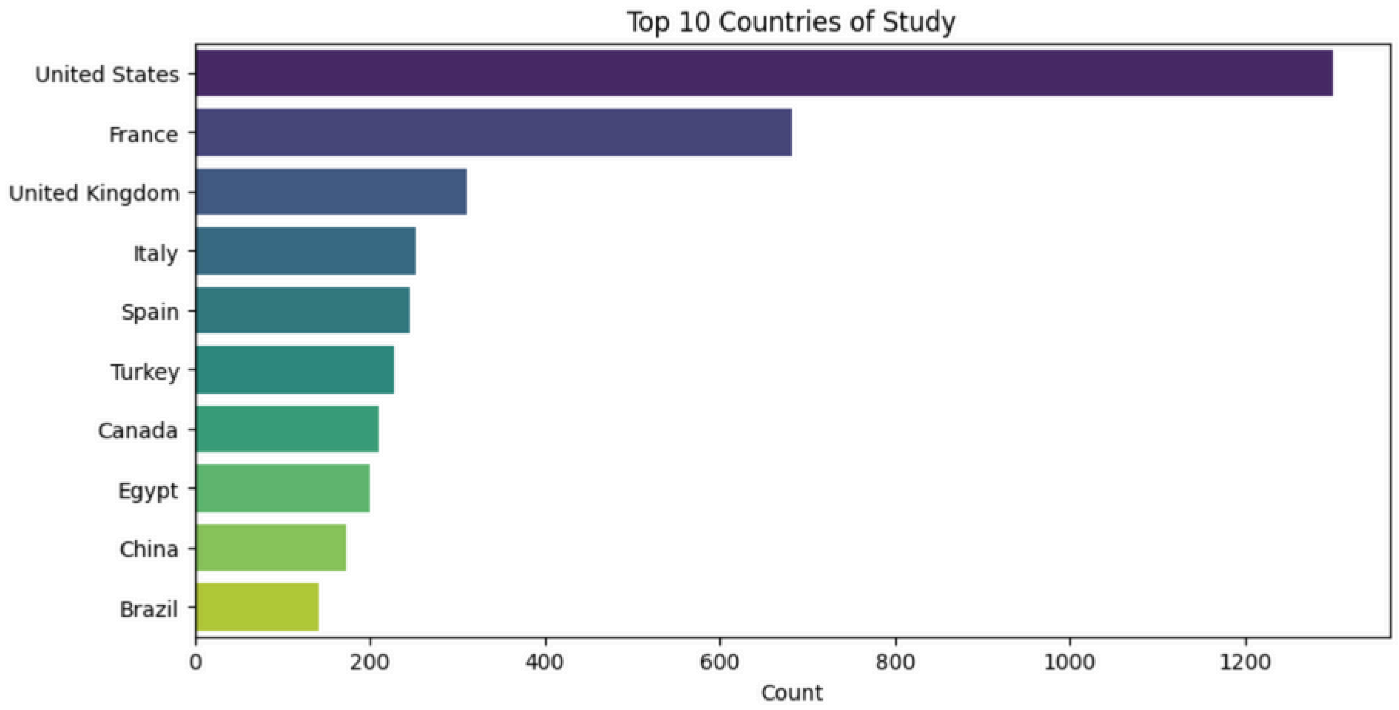


7. Age Distribution of Study Participants

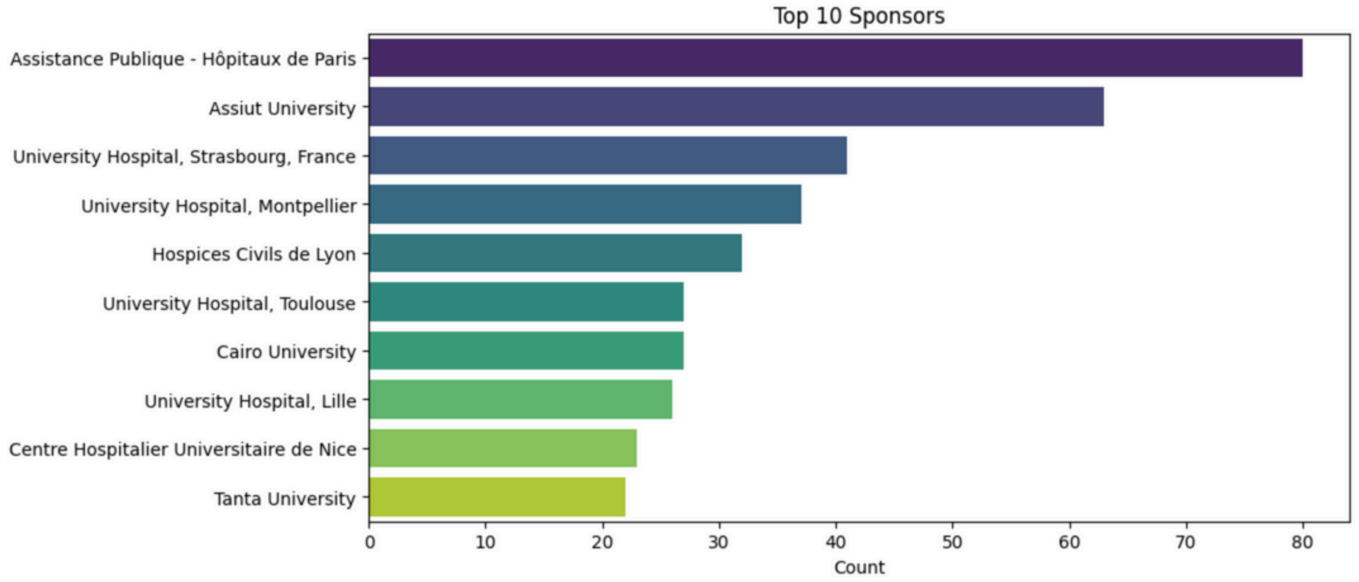
Age Category	Value
Average Minimum Age	19 years
Median Minimum Age	18 years
Youngest Minimum Age	0 years (Includes Infant Studies)
Oldest Maximum Age	90 years



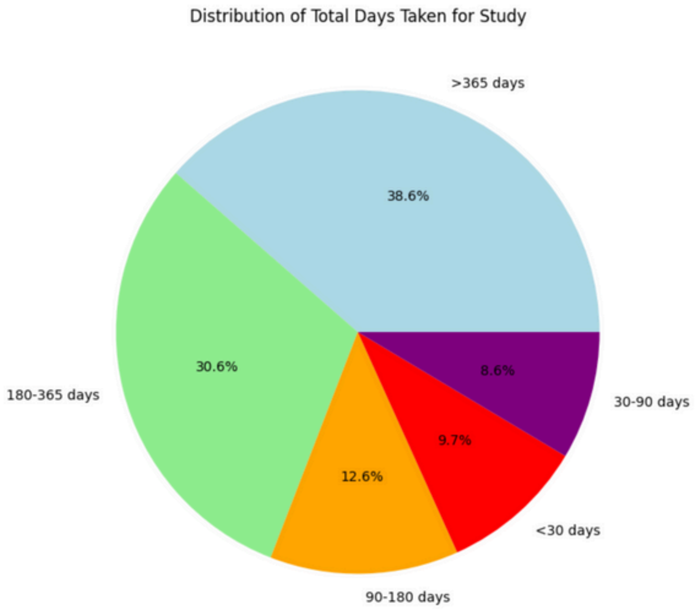
```python
# Function to extract min and max age
def extract_age_range(age_str):
    """
    Extracts the minimum and maximum age from the given age string.
    Returns a tuple (min_age, max_age) where max_age can be None for "and older" cases.
    """
    age_str = str(age_str).lower().replace("\xa0", "")  # Remove special spaces

    # Match cases like "18 Years to 75 Years"
    match_range = re.search(r"(\d+)\s*years?\s*to\s*(\d+)\s*years?", age_str)
    if match_range:
        return int(match_range.group(1)), int(match_range.group(2))

    # Match cases like "18 Years and older"
    match_older = re.search(r"(\d+)\s*years?\s*and\s*older", age_str)
    if match_older:
        return int(match_older.group(1)), None

    # Match cases like "Up to 99 Years"
    match_up_to = re.search(r"up to\s*(\d+)\s*years?", age_str)
    if match_up_to:
        return 0, int(match_up_to.group(1))

    return None, None  # Default if no match

# Apply function to extract min and max age
df[["Min Age", "Max Age"]] = df["Age"].apply(lambda x: pd.Series(extract_age_range(x)))

# Convert Min Age to numeric, dropping NaN values
age_data = df["Min Age"].dropna()

# Plot histogram
plt.figure(figsize=(10, 5))
sns.histplot(age_data, bins=20, kde=True, color="skyblue")
plt.xlabel("Minimum Age (Years)")
plt.ylabel("Frequency")
plt.title("Distribution of Minimum Age in Dataset")
plt.grid(axis="y", linestyle="--", alpha=0.7)
plt.show()

# Plot box plot
plt.figure(figsize=(8, 4))
sns.boxplot(x=age_data, color="lightcoral")
plt.xlabel("Minimum Age (Years)")
plt.title("Box Plot of Minimum Age")
plt.grid(axis="x", linestyle="--", alpha=0.7)
plt.show()
```



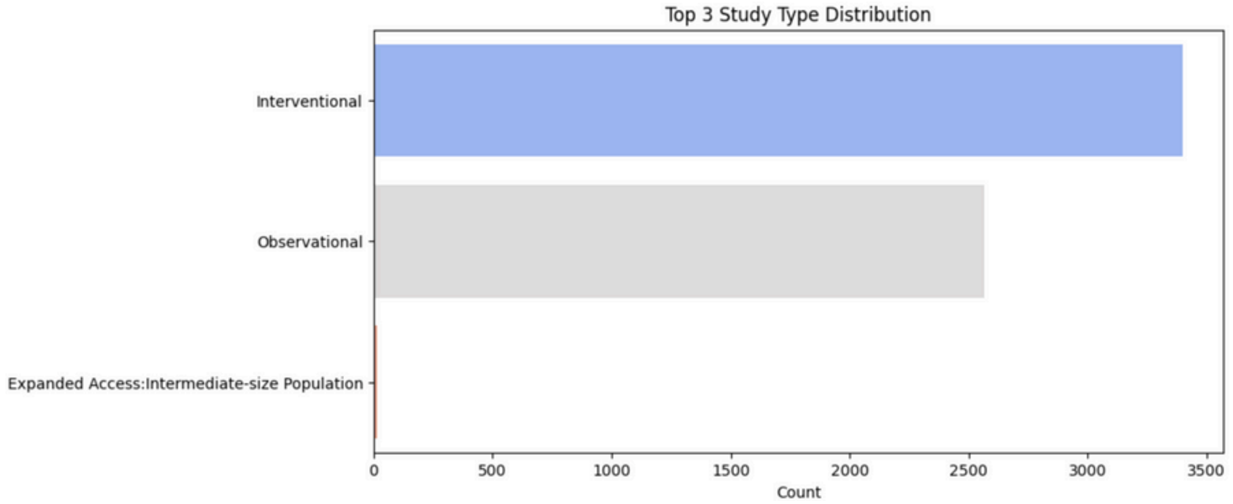
Top 10 Countries of Study



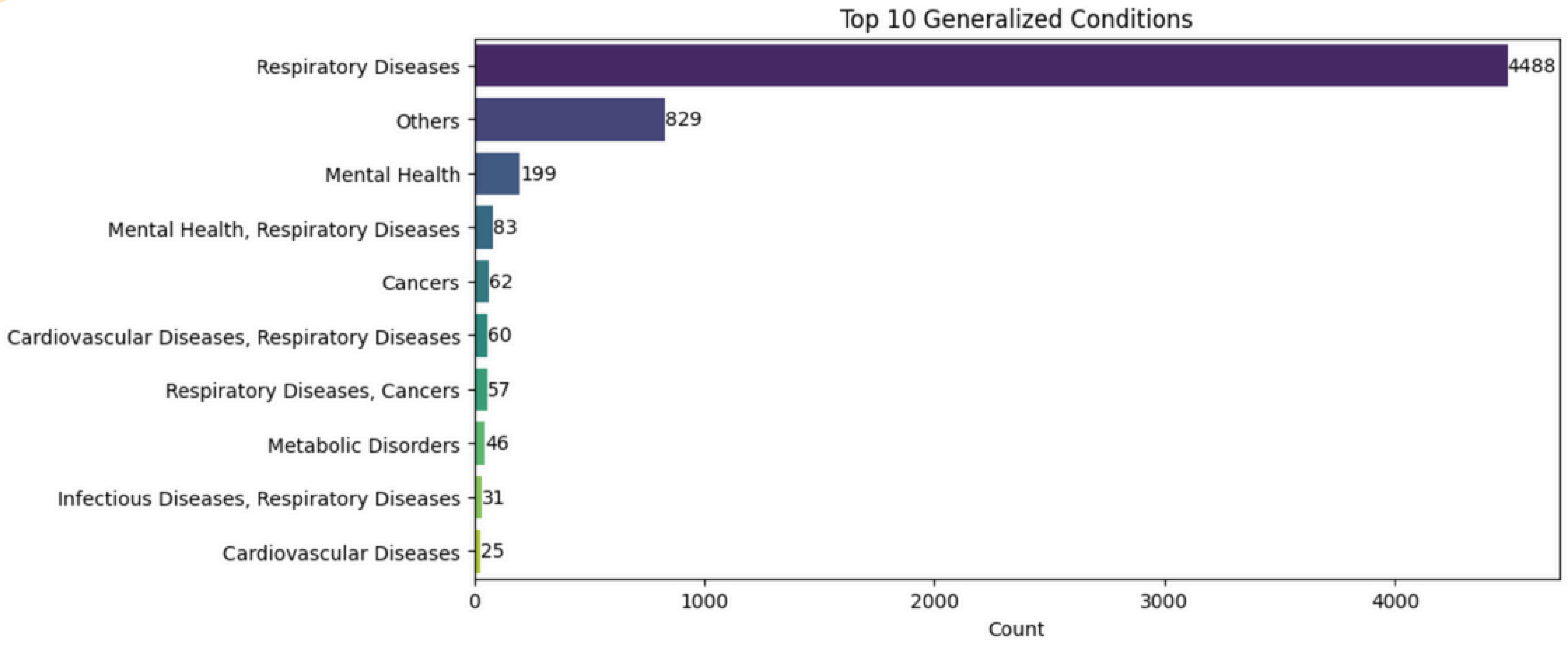
Top 10 Sponsors

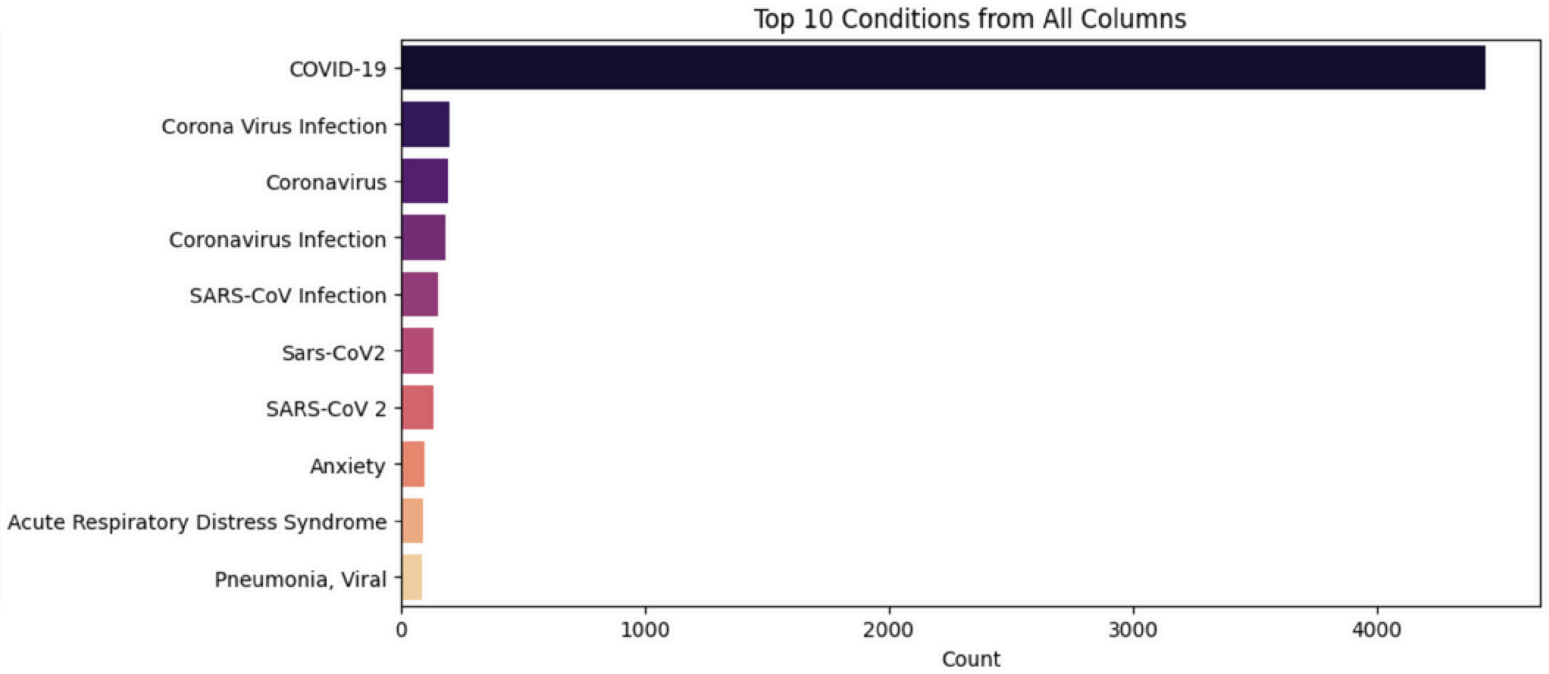


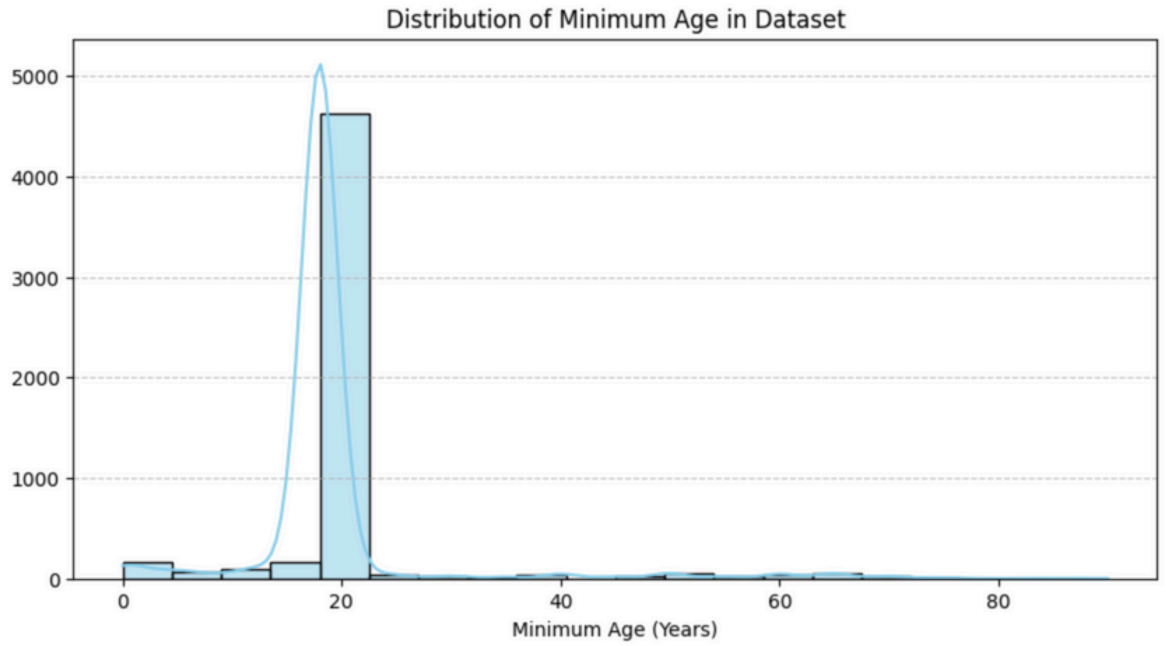
Distribution of Total Days Taken for Study

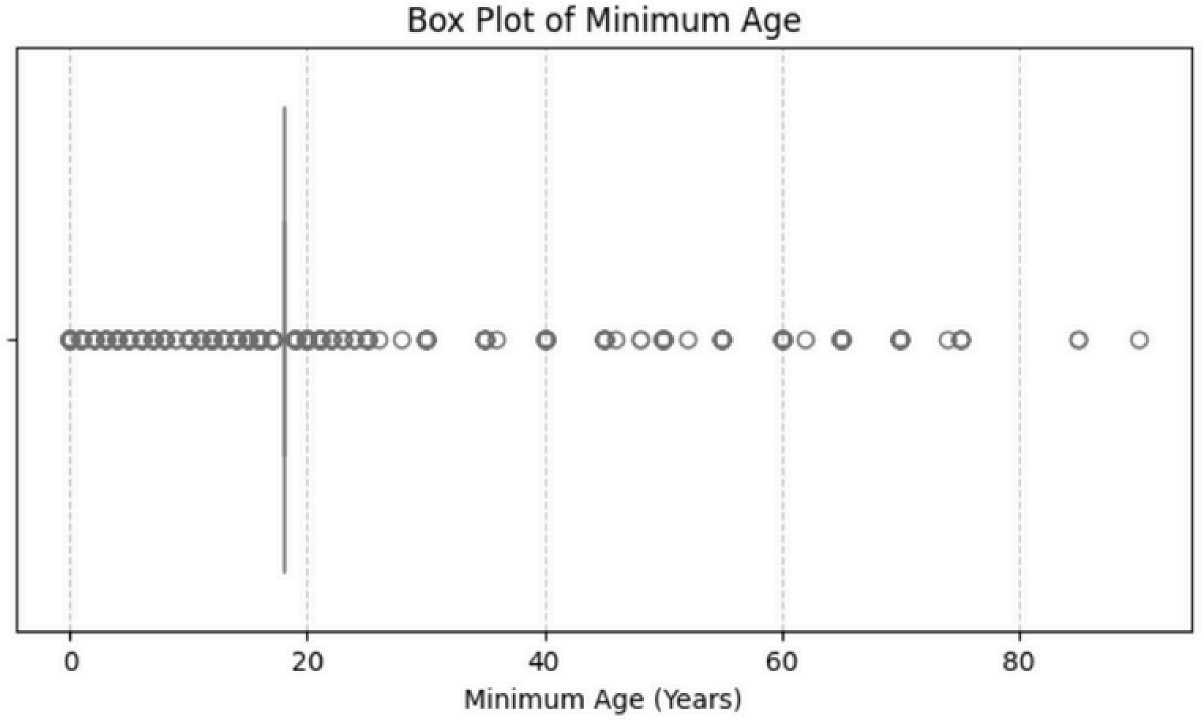


Top 3 Study Type Distribution


Top 10 Generalized Conditions
Respiratory Diseases
4488
Others
829
Mental Health
199
Mental Health, Respiratory Diseases
83
Cancers
62
Cardiovascular Diseases, Respiratory Diseases
60
Respiratory Diseases, Cancers
57
Metabolic Disorders
46
Infectious Diseases, Respiratory Diseases
31
Cardiovascular Diseases
25
0
1000
2000
3000
4000
Count


Top 10 Conditions from All Columns
COVID-19
Corona Virus Infection
Coronavirus
Coronavirus Infection
SARS-CoV Infection
Sars-CoV2
SARS-CoV 2
Anxiety
Acute Respiratory Distress Syndrome
Pneumonia, Viral
0
1000
2000
3000
4000
Count


Distribution of Minimum Age in Dataset
5000
4000
3000
2000
1000
0
0
20
40
60
80
Minimum Age (Years)


Box Plot of Minimum Age
0
20
40
60
80
Minimum Age (Years)

THANK YOU

https://github.com/krishna28zorg/Covid_Analysis