Covid-19 Exploratory Data Analysis Project

Anisha Jain

Research questions

1. Find the Enrolment:

□ maximum

□ minimum

□ mean

□ median

□ standard deviation

□ enrolment count

2. Find the Status of different stages of clinical trials
3. Find the Phase distribution of patients
4. Which Age Group has higher exposure to Covid-19
5. Age Group distribution of reported cases
6. Count of patients Status in different Phases
7. Trials started from (2019-2025) as per available report
8. Trials Start Date
9. Trials Primary Completion Date
10. Trials Completion Date
11. Count of Trials by Date Type and Year Group (1 group= 4 years)
12. Gender Distribution of patients
13. Find the month with the highest number of cases being reported
14. Find:

□ National Clinical Trial (NCT) number

□ Shape of the data

□ Unique values in the DataFrame

Anisha Jain

```
import numpy as np
```

```
import pandas as pd
```

```
import matplotlib as mp
```

```
import matplotlib.pyplot as plt
```

LOADING CSV FILE

```
df=pd.read_csv(r"/content/COVID clinical trials (1).csv")
```

```
print(df)
```

```
      Rank   NCT Number                                              Title  \
0        1  NCT04785898  Diagnostic Performance of the ID Now™ COVID-19...
1        2  NCT04595136  Study to Evaluate the Efficacy of COVID19-0001...
2        3  NCT04395482  Lung CT Scan Analysis of SARS-CoV2 Induced Lun...
3        4  NCT04416061  The Role of a Private Hospital in Hong Kong Am...
4        5  NCT04395924         Maternal-foetal Transmission of SARS-Cov-2
...    ...          ...                                                ...
5778  5779  NCT04011644  Mobile Health for Alcohol Use Disorders in Cli...
5779  5780  NCT04681339  Antibiotic Prescription in Children Hospitaliz...
5780  5781  NCT04740229  Moderate-intensity Flow-based Yoga Effects on ...
5781  5782  NCT04804917           3-year Follow-up of the Mind My Mind RCT
5782  5783  NCT04680000  Chronic Pain Management In Primary Care Using ...

          Acronym                  Status         Study Results  \
0     COVID-IDNow  Active, not recruiting  No Results Available
1        COVID-19      Not yet recruiting  No Results Available
2     TAC-COVID19              Recruiting  No Results Available
3        COVID-19  Active, not recruiting  No Results Available
4     TMF-COVID-19             Recruiting  No Results Available
...           ...                     ...                   ...
5778          NaN              Recruiting  No Results Available
5779          NaN      Not yet recruiting  No Results Available
5780          NaN              Recruiting  No Results Available
5781  MindMyMindFU             Recruiting  No Results Available
5782          NaN      Not yet recruiting  No Results Available

                                             Conditions  \
0                                               Covid19
1                                  SARS-CoV-2 Infection
2                                               covid19
3                                                 COVID
4     Maternal Fetal Infection Transmission|COVID-19...
...                                                 ...
5778                        Alcohol Drinking|Telemedicine
5779  Community Acquired Pneumonia in Children|Antib...
5780                                 Stress|Psychological
5781  Emotional Problem|Anxiety Disorder of Childhoo...
5782                                        Chronic Pain

                                          Interventions  \
0       Diagnostic Test: ID Now™ COVID-19 Screening Test
1        Drug: Drug COVID19-0001-USR|Drug: normal saline
2     Other: Lung CT scan analysis in COVID-19 patients
3             Diagnostic Test: COVID 19 Diagnostic Test
4     Diagnostic Test: Diagnosis of SARS-Cov2 by RT-...
...                                                 ...
5778  Behavioral: A-CHESS self-monitored|Behavioral:...
5779  Other: Antibiotic treatment|Other: No antibiot...
5780                                  Behavioral: Yoga
5781                                                NaN
5782  Behavioral: Brief Cognitive Behavioral Therapy...

                                       Outcome Measures  \
0     Evaluate the diagnostic performance of the ID ...
1     Change on viral load results from baseline aft...
2     A qualitative analysis of parenchymal lung dam...
3     Proportion of asymptomatic subjects|Proportion...
4     COVID-19 by positive PCR in cord blood and / o...
```

UNDERSTANDING THE DATA

```
print(df.info)
```

```
<bound method DataFrame.info of       Rank   NCT Number                                              Title  \
0        1  NCT04785898  Diagnostic Performance of the ID Now™ COVID-19...
1        2  NCT04595136  Study to Evaluate the Efficacy of COVID19-0001...
2        3  NCT04395482  Lung CT Scan Analysis of SARS-CoV2 Induced Lun...
3        4  NCT04416061  The Role of a Private Hospital in Hong Kong Am...
4        5  NCT04395924         Maternal-foetal Transmission of SARS-Cov-2
```

```
...    ...          ...                                                ...
5778  5779  NCT04011644  Mobile Health for Alcohol Use Disorders in Cli...
5779  5780  NCT04681339  Antibiotic Prescription in Children Hospitaliz...
5780  5781  NCT04740229  Moderate-intensity Flow-based Yoga Effects on ...
5781  5782  NCT04804917           3-year Follow-up of the Mind My Mind RCT
5782  5783  NCT04680000  Chronic Pain Management In Primary Care Using ...

           Acronym                  Status         Study Results  \
0      COVID-IDNow  Active, not recruiting  No Results Available   
1         COVID-19      Not yet recruiting  No Results Available   
2      TAC-COVID19              Recruiting  No Results Available   
3         COVID-19  Active, not recruiting  No Results Available   
4     TMF-COVID-19              Recruiting  No Results Available   
...            ...                     ...                   ...   
5778           NaN              Recruiting  No Results Available   
5779           NaN      Not yet recruiting  No Results Available   
5780           NaN              Recruiting  No Results Available   
5781  MindMyMindFU              Recruiting  No Results Available   
5782           NaN      Not yet recruiting  No Results Available   

                                             Conditions  \
0                                               Covid19   
1                                  SARS-CoV-2 Infection   
2                                               covid19   
3                                                 COVID   
4     Maternal Fetal Infection Transmission|COVID-19...   
...                                                 ...   
5778                      Alcohol Drinking|Telemedicine   
5779  Community Acquired Pneumonia in Children|Antib...   
5780                               Stress|Psychological   
5781  Emotional Problem|Anxiety Disorder of Childhoo...   
5782                                       Chronic Pain   

                                          Interventions  \
0      Diagnostic Test: ID Now™ COVID-19 Screening Test   
1       Drug: Drug COVID19-0001-USR|Drug: normal saline   
2     Other: Lung CT scan analysis in COVID-19 patients   
3             Diagnostic Test: COVID 19 Diagnostic Test   
4     Diagnostic Test: Diagnosis of SARS-Cov2 by RT-...   
...                                                 ...   
5778  Behavioral: A-CHESS self-monitored|Behavioral:...   
5779  Other: Antibiotic treatment|Other: No antibiot...   
5780                                   Behavioral: Yoga   
5781                                                NaN   
5782  Behavioral: Brief Cognitive Behavioral Therapy...   

                                       Outcome Measures  \
0     Evaluate the diagnostic performance of the ID ...   
1     Change on viral load results from baseline aft...   
2     A qualitative analysis of parenchymal lung dam...   
3     Proportion of asymptomatic subjects|Proportion...   
4     COVID-19 by positive PCR in cord blood and / o...   
```

```
print(df.describe())
```

```
              Rank    Enrollment
count  5783.000000  5.749000e+03
mean   2892.000000  1.831949e+04
std    1669.552635  4.045437e+05
min       1.000000  0.000000e+00
25%    1446.500000  6.000000e+01
50%    2892.000000  1.700000e+02
75%    4337.500000  5.600000e+02
max    5783.000000  2.000000e+07
```

```
print(df.describe(include="object"))
```

```
         NCT Number                                              Title  \
count          5783                                               5783   
unique         5783                                               5775   
top     NCT04680000  Acalabrutinib Study With Best Supportive Care ...   
freq              1                                                  2   

         Acronym      Status         Study Results Conditions  \
count       2480        5783                  5783       5783   
unique      2338          12                     2       3067   
top     COVID-19  Recruiting  No Results Available   COVID-19   
freq          47        2805                  5747        720   

                Interventions Outcome Measures  \
count                    4897             5748   
unique                   4337             5687   
top     Other: No intervention        Mortality   
freq                       32                5   

                         Sponsor/Collaborators Gender  ... Other IDs  \
count                                     5783   5773  ...      5782   
unique                                    3631      3  ...      5734   
top     Assistance Publique - Hôpitaux de Paris    All  ...  COVID-19   
freq                                        78   5567  ...         6   

         Start Date Primary Completion Date    Completion Date  \
count          5749                    5747               5747   
unique          654                     877                978   
top     May 1, 2020       December 31, 2020  December 31, 2021   
freq            113                     122                179   
```

```
         First Posted Results First Posted Last Update Posted  \
count            5783                   36               5783
unique            438                   33                269
top    April 24, 2020     November 4, 2020      April 8, 2021
freq              108                    2                109

                               Locations  \
count                               5198
unique                              4255
top     Uhmontpellier, Montpellier, France
freq                                  19

                                       Study Documents  \
count                                              182
unique                                             182
top     "Statistical Analysis Plan", https://ClinicalT...
freq                                                 1

                                             URL
count                                       5783
unique                                      5783
top     https://ClinicalTrials.gov/show/NCT04680000
freq                                           1

[4 rows x 25 columns]
```

DATA CLEANING

```
print(df.head(6))
```

```
   Rank   NCT Number                                              Title  \
0     1  NCT04785898  Diagnostic Performance of the ID Now™ COVID-19...
1     2  NCT04595136  Study to Evaluate the Efficacy of COVID19-0001...
2     3  NCT04395482  Lung CT Scan Analysis of SARS-CoV2 Induced Lun...
3     4  NCT04416061  The Role of a Private Hospital in Hong Kong Am...
4     5  NCT04395924         Maternal-foetal Transmission of SARS-Cov-2
5     6  NCT04516954          Convalescent Plasma for COVID-19 Patients

        Acronym                  Status         Study Results  \
0   COVID-IDNow  Active, not recruiting  No Results Available
1      COVID-19      Not yet recruiting  No Results Available
2   TAC-COVID19              Recruiting  No Results Available
3      COVID-19  Active, not recruiting  No Results Available
4  TMF-COVID-19              Recruiting  No Results Available
5          CPCP  Enrolling by invitation  No Results Available

                                        Conditions  \
0                                           Covid19
1                               SARS-CoV-2 Infection
2                                           covid19
3                                             COVID
4  Maternal Fetal Infection Transmission|COVID-19...
5                                          COVID 19

                                     Interventions  \
0   Diagnostic Test: ID Now™ COVID-19 Screening Test
1    Drug: Drug COVID19-0001-USR|Drug: normal saline
2  Other: Lung CT scan analysis in COVID-19 patients
3           Diagnostic Test: COVID 19 Diagnostic Test
4  Diagnostic Test: Diagnosis of SARS-Cov2 by RT-...
5            Biological: Convalescent COVID 19 Plasma

                                  Outcome Measures  \
0  Evaluate the diagnostic performance of the ID ...
1  Change on viral load results from baseline aft...
2  A qualitative analysis of parenchymal lung dam...
3  Proportion of asymptomatic subjects|Proportion...
4  COVID-19 by positive PCR in cord blood and / o...
5  Evaluate the safety|Change in requirement for ...

                              Sponsor/Collaborators  ...        Other IDs  \
0             Groupe Hospitalier Paris Saint Joseph  ...      COVID-IDNow
1                      United Medical Specialties  ...  COVID19-0001-USR
2                    University of Milano Bicocca  ...      TAC-COVID19
3                    Hong Kong Sanatorium & Hospital  ...       RC-2020-08
4  Centre Hospitalier Régional d'Orléans|Centre d...  ...      CHRO-2020-10
5  Vinmec Research Institute of Stem Cell and Gen...  ...       ISC.20.11.1

         Start Date Primary Completion Date    Completion Date  \
0  November 9, 2020       December 22, 2020     April 30, 2021
1  November 2, 2020       December 15, 2020   January 29, 2021
2       May 7, 2020           June 15, 2021      June 15, 2021
3      May 25, 2020           July 31, 2020    August 31, 2020
4       May 5, 2020                May 2021           May 2021
5    August 1, 2020       November 30, 2020  December 30, 2020

       First Posted Results First Posted Last Update Posted  \
0     March 8, 2021                  NaN      March 8, 2021
```

```
print(df.isnull().sum())
```

```
Rank                          0
NCT Number                    0
Title                         0
Acronym                    3303
Status                        0
Study Results                 0
Conditions                    0
Interventions               886
Outcome Measures             35
Sponsor/Collaborators         0
Gender                       10
Age                           0
Phases                     2461
Enrollment                   34
Funded Bys                    0
Study Type                    0
Study Designs                35
Other IDs                     1
Start Date                   34
Primary Completion Date      36
Completion Date              36
First Posted                  0
Results First Posted       5747
Last Update Posted            0
Locations                   585
Study Documents            5601
URL                           0
dtype: int64
```

```
df = df.dropna(subset=["Acronym"])
```

```
df = df.dropna(subset=["Outcome Measures"])
```

```
df = df.dropna(subset=['Gender'])
```

```
df = df.dropna(subset=["Study Documents"])
```

```
df = df.dropna(subset=["Study Designs"])
```

```
df = df.dropna(subset=["Start Date"])
```

```
df = df.dropna(subset=["Primary Completion Date"])
```

```
df = df.dropna(subset=["Completion Date"])
```

```
df['Interventions']=df['Interventions'].fillna('Unknown')
```

```
df['Locations']=df['Locations'].fillna('Unknown')
```

```
df['Phases'] = df['Phases'].fillna('Unknown')
```

```
df['Results First Posted'] = df['Results First Posted'].fillna('Unknown')
```

```
print(df.isnull().sum())
```

```
Rank                       0
NCT Number                 0
Title                      0
Acronym                    0
Status                     0
Study Results              0
Conditions                 0
Interventions              0
Outcome Measures           0
Sponsor/Collaborators      0
Gender                     0
Age                        0
Phases                     0
Enrollment                 0
Funded Bys                 0
Study Type                 0
Study Designs              0
Other IDs                  0
Start Date                 0
Primary Completion Date    0
Completion Date            0
First Posted               0
Results First Posted       0
Last Update Posted         0
```

```
Locations                  0
Study Documents            0
URL                        0
dtype: int64
```

```
categorical_features = df.select_dtypes(include =object).columns
features =categorical_features[df[categorical_features].isnull().mean()>0]
print(features)
```

```
Index([], dtype='object')
```

```
for feature in features:
  df[feature] = df[feature].fillna("Missing {feature}")
df.isnull().mean() * 100
```

	0
Rank	0.0
NCT Number	0.0
Title	0.0
Acronym	0.0
Status	0.0
Study Results	0.0
Conditions	0.0
Interventions	0.0
Outcome Measures	0.0
Sponsor/Collaborators	0.0
Gender	0.0
Age	0.0
Phases	0.0
Enrollment	0.0
Funded Bys	0.0
Study Type	0.0
Study Designs	0.0
Other IDs	0.0
Start Date	0.0
Primary Completion Date	0.0
Completion Date	0.0
First Posted	0.0
Results First Posted	0.0
Last Update Posted	0.0
Locations	0.0
Study Documents	0.0
URL	0.0

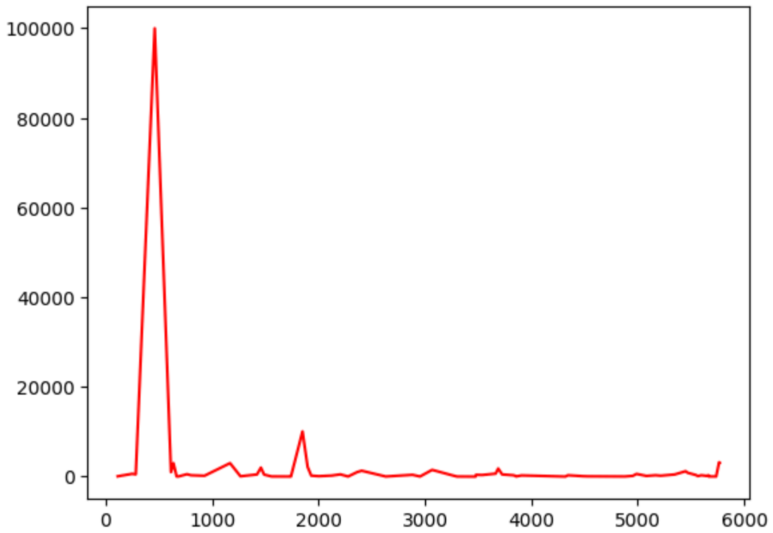
dtype: float64

```
# Check the skewness
df.Enrollment.skew()
```

```
np.float64(8.202462483511036)
```

```
# Plotting the distribution of the enrollment
df.Enrollment.plot(kind = 'line',color='red')
```

```
<Axes: >
```



```
# Using Median to impute Missing Values
df.Enrollment = df.Enrollment.fillna(median_Value)
```

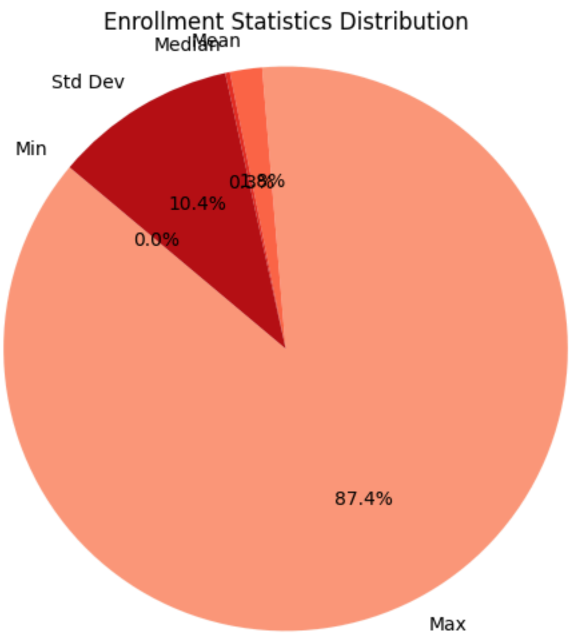
```
# Detecting (Percentage) Missing Data
df.isnull().mean() * 100
```

```
                           0
Rank                     0.0
NCT Number               0.0
Title                    0.0
Acronym                  0.0
Status                   0.0
Study Results            0.0
Conditions               0.0
Interventions            0.0
Outcome Measures         0.0
Sponsor/Collaborators    0.0
Gender                   0.0
Age                      0.0
Phases                   0.0
Enrollment               0.0
Funded Bys               0.0
Study Type               0.0
Study Designs            0.0
Other IDs                0.0
Start Date               0.0
Primary Completion Date  0.0
Completion Date          0.0
First Posted             0.0
Results First Posted     0.0
Last Update Posted       0.0
Locations                0.0
Study Documents          0.0
URL                      0.0

dtype: float64
```

ANSWER-1

```
import matplotlib.pyplot as plt

# Use your precomputed values
min_Value = df.Enrollment.min()
```

```
max_Value = df.Enrollment.max()
mean_Value = df.Enrollment.mean()
median_Value = df.Enrollment.median()
std_Value = df.Enrollment.std()

# Create a dictionary
stats = {
    'Min': min_Value,
    'Max': max_Value,
    'Mean': mean_Value,
    'Median': median_Value,
    'Std Dev': std_Value
}

# Normalize values to percentages (for better pie visuals)
total = sum(stats.values())
percentages = [val / total for val in stats.values()]

# Plot pie chart
plt.figure(figsize=(6, 6))
plt.pie(percentages,
        labels=stats.keys(),
        autopct='%1.1f%%',
        startangle=140,
        colors=plt.cm.Reds(range(50, 250, 40)))

plt.title('Enrollment Statistics Distribution')
plt.axis('equal')  # Ensures a perfect circle
plt.show()
```



DATA TRANSFORMATION

```
# Converting the date values from string type to date type
```

```
for col in ['Start Date', 'Primary Completion Date', 'Completion Date', 'First Posted', 'Results First Posted', 'Last Update Posted']:
    df[col] = df[col].astype(str).str.strip()                          # Remove spaces
    df[col] = pd.to_datetime(df[col], errors='coerce')               # Convert to datetime; invalid ones become NaT

print(df)
```

```
      Rank   NCT Number                                           Title  \
113    114  NCT04780126  Sarco-COVID Study: Measuring the Loss of Skele...
250    251  NCT04341441  Will Hydroxychloroquine Impede or Prevent COVI...
283    284  NCT04382781    Immunosuppressive Treatment in COVID-19 Patients
461    462  NCT04321811  Behavior, Environment And Treatments for Covid-19
614    615  NCT04659941  Use of BCG Vaccine as a Preventive Measure for...
...    ...          ...                                               ...
5668  5669  NCT04400682  Bioequivalence Study of Favipiravir 200 mg Fil...
5717  5718  NCT04386876  Bioequivalence Study of Lopinavir/Ritonavir 20...
5737  5738  NCT03483935  Microwave Therapy for Treatment of Precancerou...
5765  5766  NCT04429061  Reaching 90 90 90 in Adolescents in Zambia: Us...
5770  5771  NCT03392883  Scaling Up Science-based Mental Health Interve...

            Acronym                  Status          Study Results  \
113     SARCO-COVID              Recruiting  No Results Available
250    WHIP COVID-19             Terminated  No Results Available
283        SAM-COVID             Recruiting  No Results Available
```

```
461          BEAT19  Active, not recruiting  No Results Available
614          ProBCG              Recruiting  No Results Available
...             ...                     ...                   ...
5668    Favipiravir               Completed           Has Results
5717        Orvical               Completed  No Results Available
5737           MTAK               Completed           Has Results
5765         SKILLZ              Recruiting  No Results Available
5770          DIADA  Active, not recruiting  No Results Available

                                             Conditions  \
113                                  Sarcopenia|Covid19   
250   COVID-19|Coronavirus|Coronavirus Infections|SA...   
283                                  COVID-19 Infection   
461                                         Coronavirus   
614                                    COVID 19 Vaccine   
...                                                 ...   
5668                                     Bioequivalence   
5717                                     Bioequivalence   
5737         Actinic Keratoses|Precancerous Skin Lesion   
5765  HIV Infections|Pregnancy Related|STI|Mental He...   
5770                Depression|Problematic Alcohol Use   

                                          Interventions  \
113                         Other: Sarcopenia diagnosis   
250   Drug: Hydroxychloroquine - Daily Dosing|Drug: ...   
283   Drug: NO-Immunosuppressive|Drug: Immunosuppres...   
461   Other: Observation of patients with known, sus...   
614                             Biological: BCG vaccine   
...                                                 ...   
5668  Drug: FAVIRA 200 MG Film Tablet|Drug: AVIGAN 2...   
5717  Drug: Lopinavir/Ritonavir 200 mg/50 mg Film Ta...   
5737                        Other: Microwave treatment   
5765  Behavioral: SKILLZ-Girl Enhanced football curr...   
5770                                  Behavioral: Laddr   

                                       Outcome Measures  \
113   Loss of muscle mass|Prevalence of sarcopenia|N...   
250   To determine if the use of hydroxychloroquine ...   
283   Invasive ventilation or death|Ventilation|Deat...   
461   Define Natural Symptom Course|Time to Hospital...   
```

```
df['Conditions'] = df['Conditions'].str.strip().str.title()
```

```
print(df)
```

```
      Rank   NCT Number                                              Title  \
113    114  NCT04780126  Sarco-COVID Study: Measuring the Loss of Skele...   
250    251  NCT04341441  Will Hydroxychloroquine Impede or Prevent COVI...   
283    284  NCT04382781    Immunosuppressive Treatment in COVID-19 Patients   
461    462  NCT04321811  Behavior, Environment And Treatments for Covid-19   
614    615  NCT04659941  Use of BCG Vaccine as a Preventive Measure for...   
...    ...          ...                                                ...   
5668  5669  NCT04400682  Bioequivalence Study of Favipiravir 200 mg Fil...   
5717  5718  NCT04386876  Bioequivalence Study of Lopinavir/Ritonavir 20...   
5737  5738  NCT03483935  Microwave Therapy for Treatment of Precancerou...   
5765  5766  NCT04429061  Reaching 90 90 90 in Adolescents in Zambia: Us...   
5770  5771  NCT03392883  Scaling Up Science-based Mental Health Interve...   

          Acronym                  Status         Study Results  \
113   SARCO-COVID              Recruiting  No Results Available   
250  WHIP COVID-19             Terminated  No Results Available   
283     SAM-COVID              Recruiting  No Results Available   
461        BEAT19  Active, not recruiting  No Results Available   
614        ProBCG              Recruiting  No Results Available   
...           ...                     ...                   ...   
5668  Favipiravir               Completed           Has Results   
5717      Orvical               Completed  No Results Available   
5737         MTAK               Completed           Has Results   
5765       SKILLZ              Recruiting  No Results Available   
5770        DIADA  Active, not recruiting  No Results Available   

                                             Conditions  \
113                                  Sarcopenia|Covid19   
250   Covid-19|Coronavirus|Coronavirus Infections|Sa...   
283                                  Covid-19 Infection   
461                                         Coronavirus   
614                                    Covid 19 Vaccine   
...                                                 ...   
5668                                     Bioequivalence   
5717                                     Bioequivalence   
5737         Actinic Keratoses|Precancerous Skin Lesion   
5765  Hiv Infections|Pregnancy Related|Sti|Mental He...   
5770                Depression|Problematic Alcohol Use   

                                          Interventions  \
113                         Other: Sarcopenia diagnosis   
250   Drug: Hydroxychloroquine - Daily Dosing|Drug: ...   
283   Drug: NO-Immunosuppressive|Drug: Immunosuppres...   
461   Other: Observation of patients with known, sus...   
614                             Biological: BCG vaccine   
...                                                 ...   
5668  Drug: FAVIRA 200 MG Film Tablet|Drug: AVIGAN 2...   
5717  Drug: Lopinavir/Ritonavir 200 mg/50 mg Film Ta...   
5737                        Other: Microwave treatment   
```

```
5765  Behavioral: SKILLZ-Girl Enhanced football curr...
5770                                  Behavioral: Laddr

                                       Outcome Measures  \
113   Loss of muscle mass|Prevalence of sarcopenia|N...
250   To determine if the use of hydroxychloroquine ...
283   Invasive ventilation or death|Ventilation|Deat...
461   Define Natural Symptom Course|Time to Hospital...
614   Compare the cumulative incidence of SARS-CoV-2 ...
```

DATA VISUALIZATION

```
# UNIVARIATE ANALYSIS-
```

```
# Status Distribution: Analyze the status of clinical trials (e.g., Completed, Ongoing).
```

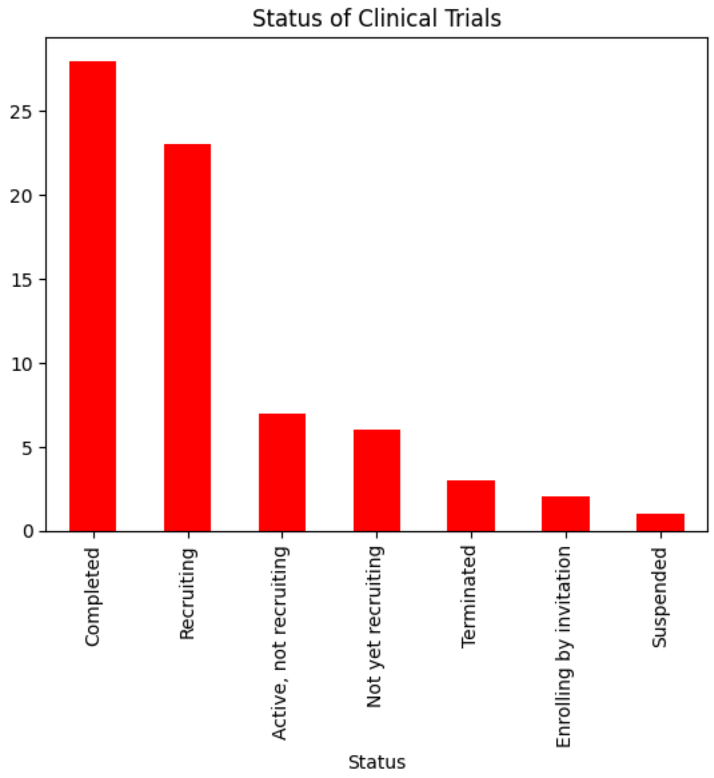
ANSWER- 2

```
print(df['Status'].value_counts())
```

```
Status
Completed                  28
Recruiting                 23
Active, not recruiting      7
Not yet recruiting          6
Terminated                  3
Enrolling by invitation     2
Suspended                   1
Name: count, dtype: int64
```

```
df['Status'].value_counts().plot(kind='bar', title='Status of Clinical Trials',color='red')
```

```
<Axes: title={'center': 'Status of Clinical Trials'}, xlabel='Status'>
```



```
#Phase Distribution: Understand the distribution of trial phases.
```

```
print(df['Phases'].value_counts())
```

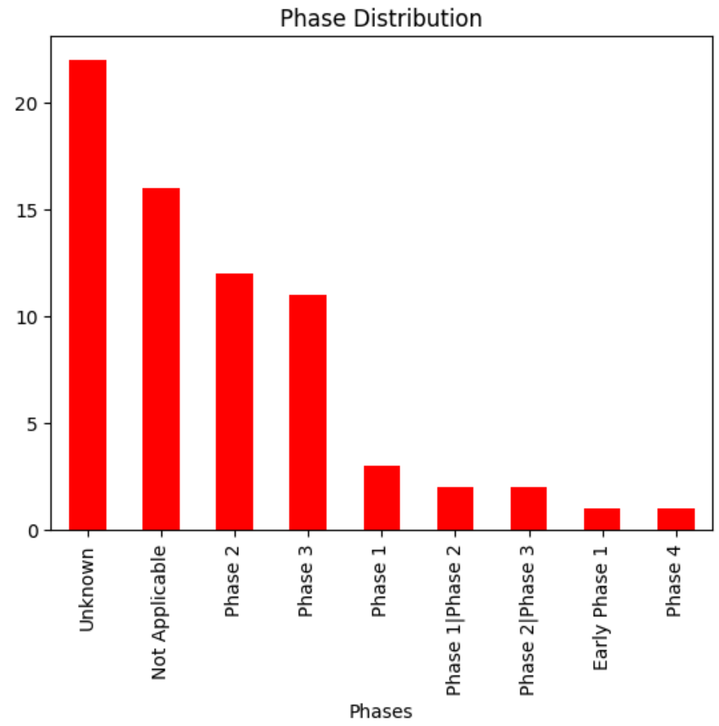
```
Phases
Unknown            22
Not Applicable     16
Phase 2            12
Phase 3            11
Phase 1             3
Phase 1|Phase 2     2
Phase 2|Phase 3     2
Early Phase 1       1
Phase 4             1
Name: count, dtype: int64
```

Anisha Jain

ANSWER- 3

```
df['Phases'].value_counts().plot(kind='bar', title='Phase Distribution',color='red')
```

```
<Axes: title={'center': 'Phase Distribution'}, xlabel='Phases'>
```



```
print(df['Age'].value_counts())
```

```
Age
18 Years and older   (Adult, Older Adult)          34
18 Years to 65 Years   (Adult, Older Adult)         4
18 Years to 75 Years   (Adult, Older Adult)         3
20 Years to 40 Years   (Adult)                      3
18 Years to 100 Years   (Adult, Older Adult)        2
18 Years to 80 Years   (Adult, Older Adult)         2
18 Years to 99 Years   (Adult, Older Adult)         2
16 Years and older   (Child, Adult, Older Adult)    2
18 Years to 59 Years   (Adult)                      2
Child, Adult, Older Adult                           2
65 Years and older   (Older Adult)                  2
18 Years to 90 Years   (Adult, Older Adult)         1
12 Years to 25 Years   (Child, Adult)               1
24 Years to 37 Years   (Adult)                      1
18 Years to 70 Years   (Adult, Older Adult)         1
up to 17 Years   (Child)                            1
5 Years and older   (Child, Adult, Older Adult)     1
24 Months to 18 Years   (Child, Adult)              1
12 Years and older   (Child, Adult, Older Adult)    1
21 Years and older   (Adult, Older Adult)           1
18 Years to 60 Years   (Adult)                      1
2 Years to 14 Years   (Child)                       1
60 Years and older   (Adult, Older Adult)           1
Name: count, dtype: int64
```

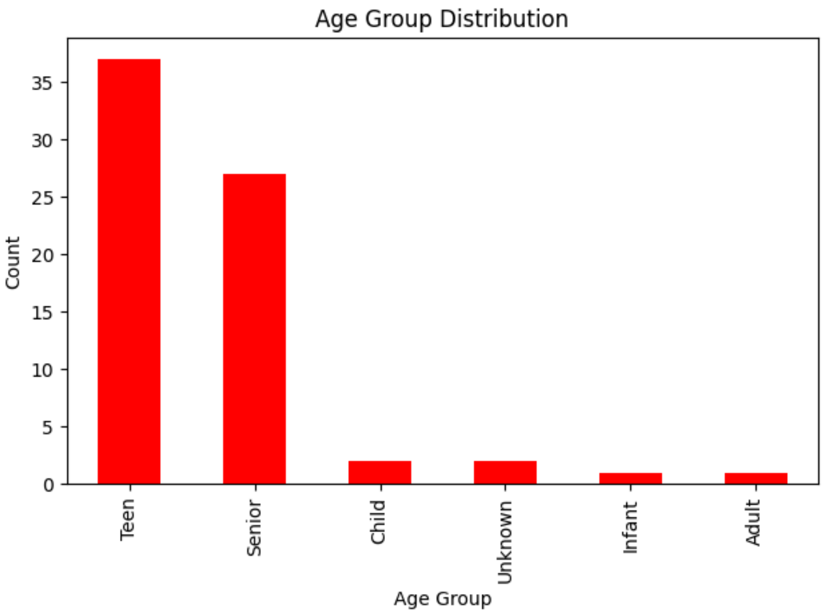
ANSWER- 4,5

```
# Values in Age column are unorganized and cluttered. For analysis, we need to group them
```

```
def age_group(age_str):
    if pd.isnull(age_str):
        return 'Unknown'
    age_str = age_str.lower().strip()

    if 'month' in age_str:
        return 'Infant'
    elif 'year' in age_str:
        digits = ''.join([c for c in age_str if c.isdigit()])
        if digits:
            age = int(digits)
            if age <= 12:
                return 'Child'
            elif age <= 19:
                return 'Teen'
            elif age <= 59:
                return 'Adult'
            else:
```

Anisha Jain

```
                return 'Senior'
        else:
            return 'Unknown'
    else:
        return 'Unknown'

# Apply to your column
df['Age_Group'] = df['Age'].apply(age_group)

# Plot grouped results
import matplotlib.pyplot as plt
df['Age_Group'].value_counts().plot(kind='bar', title='Age Group Distribution',color='red')
plt.xlabel('Age Group')
plt.ylabel('Count')
plt.tight_layout()
plt.show()
```



"" --> [c for c in age_str if c.isdigit()] This is a list comprehension.

It goes through each character c in the string age_str.

It checks if the character is a digit using c.isdigit().

If it's a digit, it includes it in the list.

--> ''.join([...]) Joins all the characters in the list into a single string.

No space or separator is added between them ('' means empty string separator).

```
# Bivariate Analysis
```

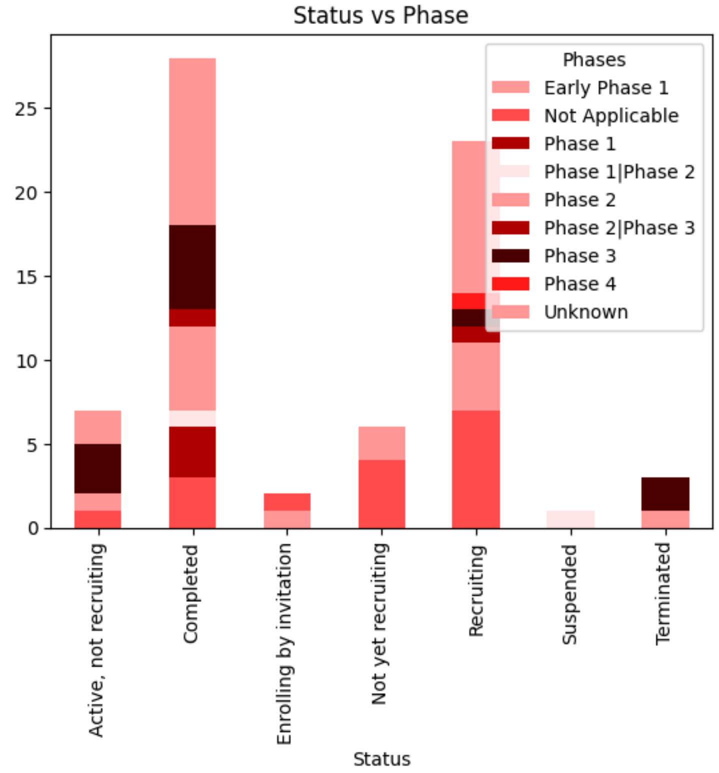
```
status_phase=pd.crosstab(df['Status'],df['Phases'])
print(status_phase)
```

```
Phases                   Early Phase 1  Not Applicable  Phase 1  \
Status
Active, not recruiting               0               1        0
Completed                            0               3        3
Enrolling by invitation              1               1        0
Not yet recruiting                   0               4        0
Recruiting                           0               7        0
Suspended                            0               0        0
Terminated                           0               0        0

Phases                   Phase 1|Phase 2  Phase 2  Phase 2|Phase 3  Phase 3  \
Status
Active, not recruiting                 0        1                0        3
Completed                              1        5                1        5
Enrolling by invitation                0        0                0        0
Not yet recruiting                     0        1                0        0
Recruiting                             0        4                1        1
Suspended                              1        0                0        0
Terminated                             0        1                0        2

Phases                   Phase 4  Unknown
Status
Active, not recruiting         0        2
Completed                      0       10
Enrolling by invitation        0        0
Not yet recruiting             0        1
Recruiting                     1        9
Suspended                      0        0
Terminated                     0        0
```

ANSWER- 6

```
red_shades = ['#ff9999','#ff4d4d', '#b30000','#ffe6e6', '#ff9999','#b30000', '#4d0000','#ff1a1a']
status_phase.plot(kind='bar',stacked=True, title='Status vs Phase',color=red_shades)
```

```
<Axes: title={'center': 'Status vs Phase'}, xlabel='Status'>
```



```
print(df['Conditions'])
```

```
113                                  Sarcopenia|Covid19
250     Covid-19|Coronavirus|Coronavirus Infections|Sa...
283                                  Covid-19 Infection
461                                         Coronavirus
614                                     Covid 19 Vaccine
                              ...
5668                                      Bioequivalence
5717                                      Bioequivalence
5737            Actinic Keratoses|Precancerous Skin Lesion
5765    Hiv Infections|Pregnancy Related|Sti|Mental He...
5770                Depression|Problematic Alcohol Use
Name: Conditions, Length: 70, dtype: object
```

```
condition_outcome = df.groupby('Conditions')['Outcome Measures']\
    .apply(lambda x: ', '.join(x.dropna().astype(str)))\
    .reset_index()

print(condition_outcome)
```

```
                                           Conditions  \
0          Actinic Keratoses|Precancerous Skin Lesion
1   Acute Pancreatitis|Acute Pancreatic Necrosis|A...
2   Alcohol Consumption|Violence, Domestic|Stress,...
3                                      Bioequivalence
4   Chronic Pain|Musculoskeletal Diseases|Quality ...
5                         Community Acquired Pneumonia
6                               Corona Virus Infection
7   Corona Virus Infection|Acute Respiratory Distr...
8                                         Coronavirus
9   Coronavirus Disease 2019 (Covid-19)|Respirator...
10  Coronavirus Infection|Pneumonia, Viral|Acute R...
11  Coronavirus|Acute Respiratory Infection|Sars-C...
12                                              Covid
13                                    Covid 19 Positive
14                                     Covid 19 Vaccine
15                                   Covid, Coronavirus
16                                           Covid-19
17                                  Covid-19 Infection
18  Covid-19|Coronavirus Infection|Sars-Cov-2 Infe...
19  Covid-19|Coronavirus|Coronavirus Infections|Sa...
20                                            Covid19
21                                        Covid19|Aki
22                                       Covid19|Ards
23                     Covid19|Lung Function Decreased
24                              Covid19|Mental Health
```

```
25                                  Covid19|Pneumonia
26                                Covid19|Progression
27              Covid19|Sars-Cov-2 Pneumonia|Covid-19
28                               Covid|Ards|Pneumonia
29                                Covid|Safety Issues
30               Covid|Statin|Cardiovascular Diseases
31                 Depression|Problematic Alcohol Use
32                           Eating Behavior|Covid-19
33  Hiv Infections|Drug Use|Opioid Use|Opioid-Use ...
34  Hiv Infections|Pregnancy Related|Sti|Mental He...
35  Hydroxychloroquine|Antimalarials|Enzyme Inhibi...
36                              Hyperglycemia|Covid19
37                         Infection Control|Covid-19
38                         Loneliness|Quality Of Life
39                        Multiple Sclerosis|Covid-19
40  Pneumonia|Coronavirus Infection In 2019 (Covid...
41                       Post Intensive Care Syndrome
42  Postoperative Cognitive Dysfunction|Depressive...
43  Posttraumatic Stress Disorder|Traumatic Brain ...
44           Psychological Stress|Hemostatic Disorder
45               Respiratory Distress Syndrome, Adult
46     Respiratory Distress Syndrome, Adult|Sars-Cov2
47                Respiratory Viral Infection|Covid19
48                                 Rheumatic Diseases
49                                 Sarcopenia|Covid19
50                                    Sars-Cov 2|Ards
51                                   Sars-Cov 2|Covid
52                     Sars-Cov-2 Respiratory Failure
53  Severe Acute Respiratory Syndrome (Sars) Pneum...
54  Severe Acute Respiratory Syndrome|Ventilation ...
55                                            Suicide
56  This Is A Pilot Study Which Aims To Assess The
```

x.dropna() → removes NaN values.

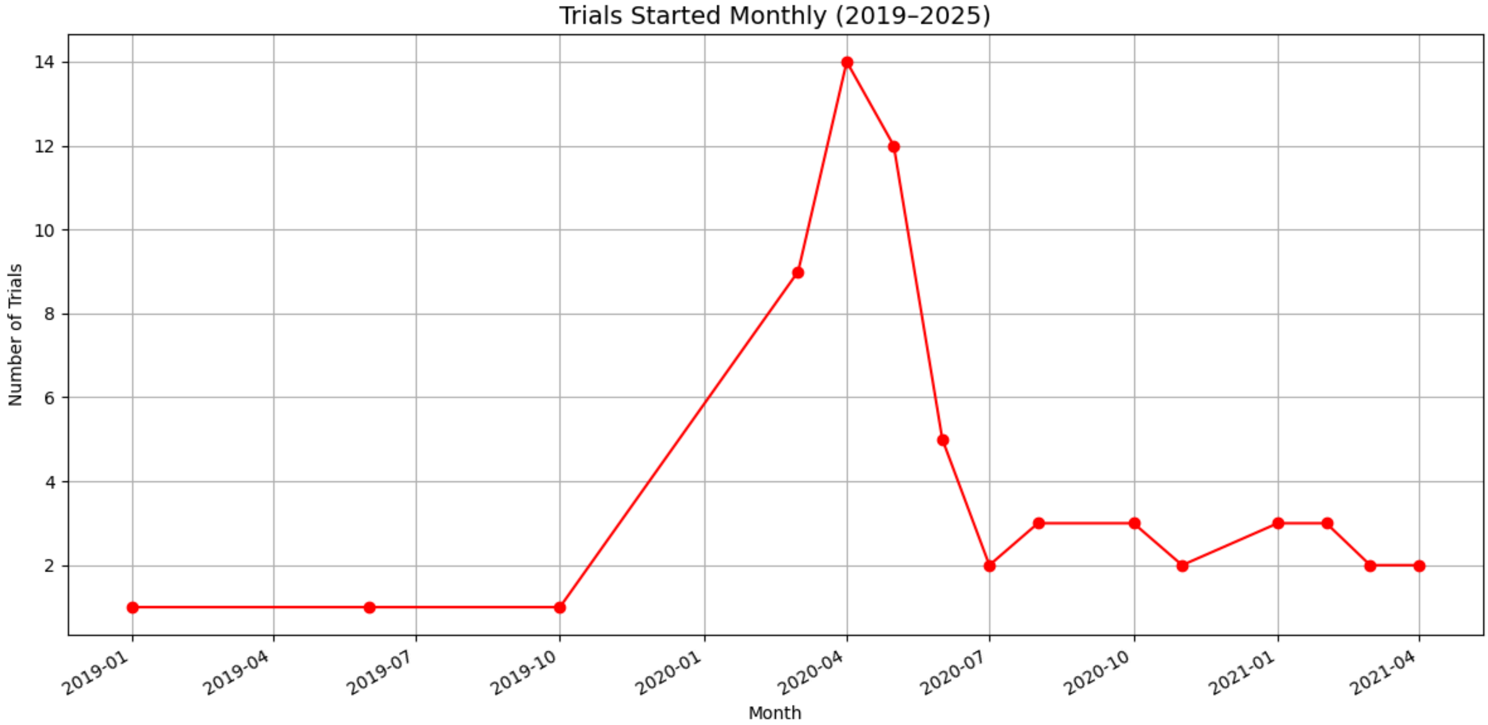
.astype(str) → converts all items to string type.

', '.join(...) → now works safely.

```
# TRIALS---
```

ANSWER-7

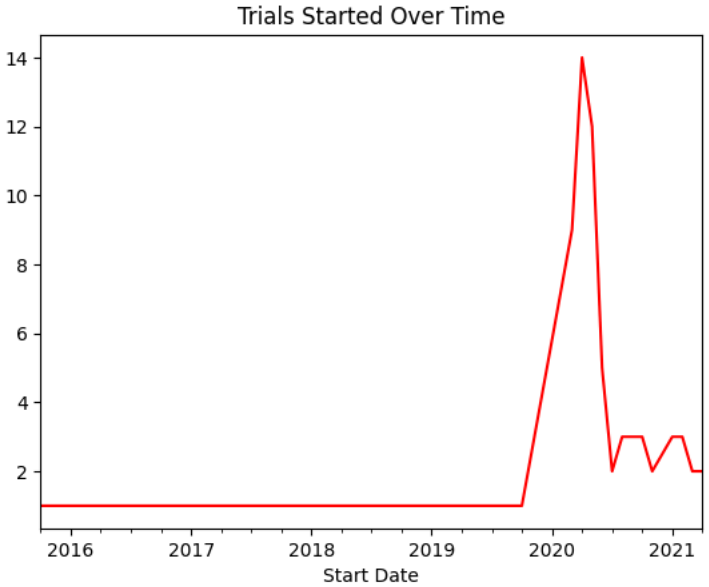
```
import pandas as pd
import matplotlib.pyplot as plt

# Convert 'Start Date' to datetime
df['Start Date'] = pd.to_datetime(df['Start Date'], errors='coerce')

# Filter dates from 2019 to 2025
df_filtered = df[(df['Start Date'].dt.year >= 2019) & (df['Start Date'].dt.year <= 2025)]

# Group by month and count
monthly_counts = df_filtered['Start Date'].dt.to_period('M').value_counts().sort_index()

# Convert PeriodIndex to datetime for plotting
monthly_counts.index = monthly_counts.index.to_timestamp()

# Plot
plt.figure(figsize=(12, 6))
monthly_counts.plot(kind='line', color='red', marker='o')
plt.title('Trials Started Monthly (2019–2025)', fontsize=14)
plt.xlabel('Month')
import pandas as pd
import matplotlib.pyplot as plt

# Convert 'Start Date' to datetime
df['Start Date'] = pd.to_datetime(df['Start Date'], errors='coerce')

# Filter dates from 2019 to 2025
df_filtered = df[(df['Start Date'].dt.year >= 2019) & (df['Start Date'].dt.year <= 2025)]

# Group by month and count
monthly_counts = df_filtered['Start Date'].dt.to_period('M').value_counts().sort_index()

# Convert PeriodIndex to datetime for plotting
monthly_counts.index = monthly_counts.index.to_timestamp()

# Plot
plt.figure(figsize=(12, 6))
monthly_counts.plot(kind='line', color='red', marker='o')
plt.title('Trials Started Monthly (2019–2025)', fontsize=14)
plt.xlabel('Month')
plt.ylabel('Number of Trials')
plt.grid(True)
plt.tight_layout()
plt.show()

plt.tight_layout()
plt.show()
```



ANSWER- 8

```
# Convert date columns to datetime
df['Start Date'] = pd.to_datetime(df['Start Date'],
errors='coerce')

# Plot the number of trials started over time
df['Start Date'].dt.to_period('M').value_counts().sort_index().plot(kind=
'line', title='Trials Started Over Time',color='red')
```

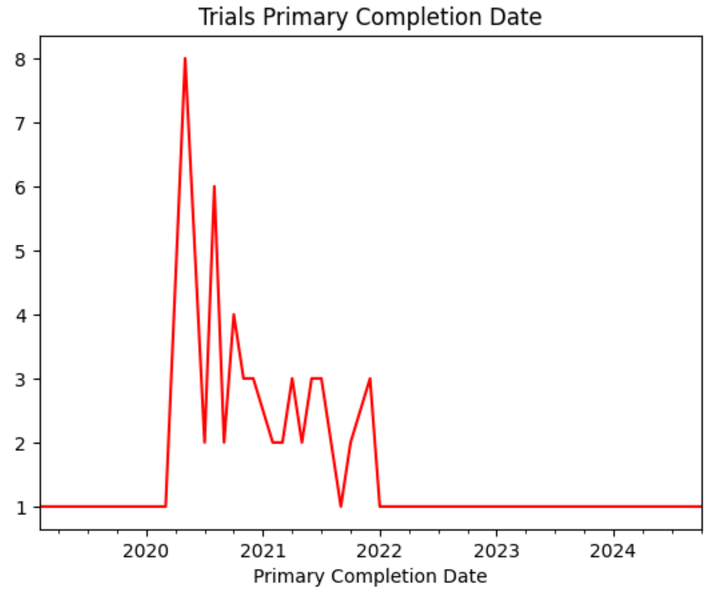
```
<Axes: title={'center': 'Trials Started Over Time'}, xlabel='Start Date'>
```



ANSWER- 9

```
df['Primary Completion Date'] = pd.to_datetime(df['Primary Completion Date'], errors='coerce')
df['Primary Completion Date'].dt.to_period('M').value_counts().sort_index().plot(kind='line', title='Trials Primary Completion Date',color='red')
```

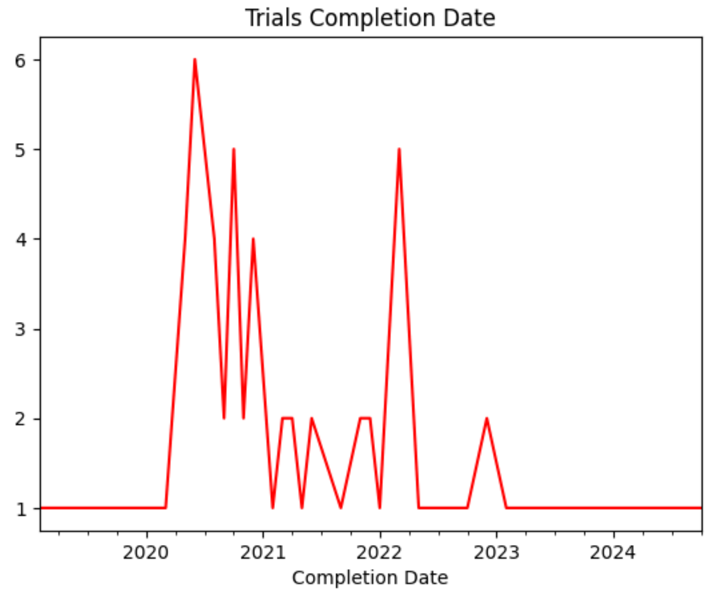
```
<Axes: title={'center': 'Trials Primary Completion Date'}, xlabel='Primary Completion Date'>
```



ANSWER- 10

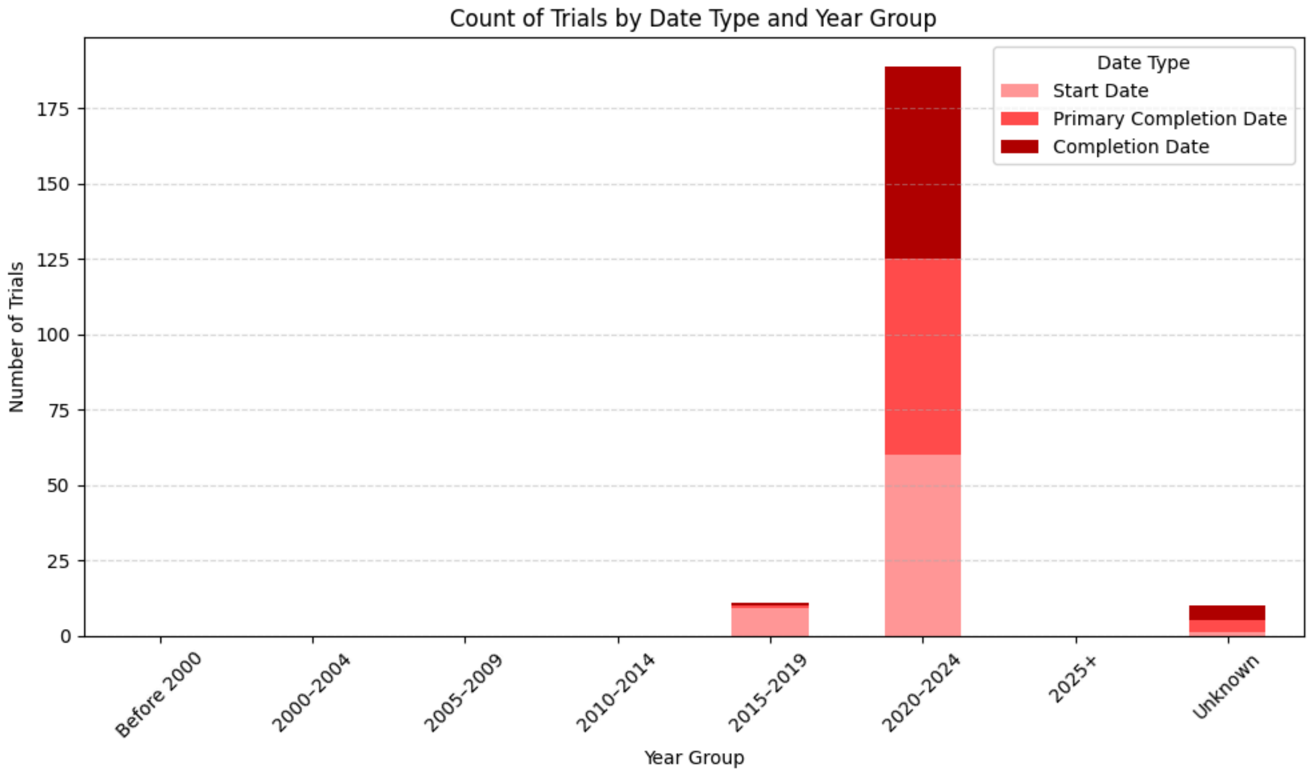
```
df['Completion Date'] = pd.to_datetime(df['Completion Date'], errors='coerce')
df['Completion Date'].dt.to_period('M').value_counts().sort_index().plot(kind='line', title='Trials Completion Date',color='red')
```

```
<Axes: title={'center': 'Trials Completion Date'}, xlabel='Completion Date'>
```



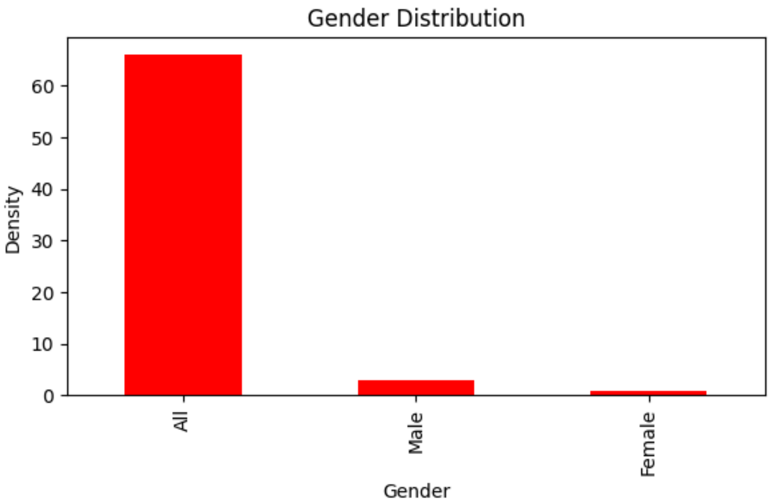
ANSWER- 11

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

# Convert to datetime
df['Start Date'] = pd.to_datetime(df['Start Date'], errors='coerce')
df['Primary Completion Date'] = pd.to_datetime(df['Primary Completion Date'], errors='coerce')
df['Completion Date'] = pd.to_datetime(df['Completion Date'], errors='coerce')

# Define function for grouping years
def year_group(date):
    if pd.isna(date):
        return 'Unknown'
    year = date.year
    if year < 2000:
        return 'Before 2000'
    elif year < 2005:
        return '2000-2004'
    elif year < 2010:
        return '2005-2009'
    elif year < 2015:
        return '2010-2014'
    elif year < 2020:
        return '2015-2019'
    elif year < 2025:
        return '2020-2024'
    else:
        return '2025+'

# Apply grouping
df['Start Group'] = df['Start Date'].apply(year_group)
df['Primary Group'] = df['Primary Completion Date'].apply(year_group)
df['Completion Group'] = df['Completion Date'].apply(year_group)

# Count each date type by year group
start_counts = df['Start Group'].value_counts()
primary_counts = df['Primary Group'].value_counts()
completion_counts = df['Completion Group'].value_counts()

# Merge counts into one DataFrame
all_years = ['Before 2000', '2000-2004', '2005-2009', '2010-2014', '2015-2019', '2020-2024', '2025+', 'Unknown']
grouped_counts = pd.DataFrame({
    'Start Date': start_counts,
    'Primary Completion Date': primary_counts,
    'Completion Date': completion_counts
}).reindex(all_years).fillna(0).astype(int)

# Plot

red_shades = ['#ff9999', '#ff4d4d', '#b30000']
grouped_counts.plot(kind='bar', stacked=True, figsize=(10, 6),color=red_shades)
plt.title('Count of Trials by Date Type and Year Group')
plt.xlabel('Year Group')
plt.ylabel('Number of Trials')
plt.legend(title='Date Type')
plt.xticks(rotation=45)
plt.tight_layout()
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.show()
```



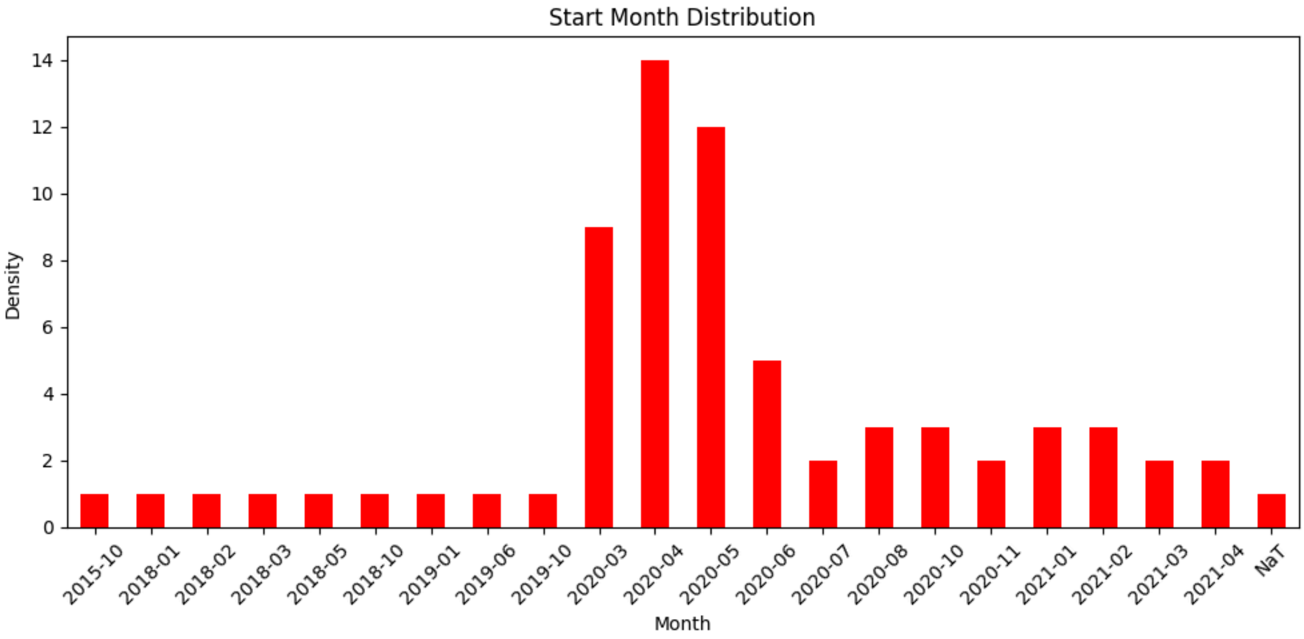
ANSWER- 12

```
import matplotlib.pyplot as plt

# Gender Visualizations
gender = df['Gender'].value_counts()

plt.figure(figsize=(6, 4))
gender.plot(kind='bar', color='red')
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Density')
plt.tight_layout()
plt.show()
```



ANSWER- 13

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

# Extract month from 'Start Date' column
start_month = pd.to_datetime(df['Start Date'], errors='coerce').dt.to_period('M').astype(str)

# Count the frequency of each start month
start_month_distribution = start_month.value_counts().sort_index()

# Plot the start month distribution
plt.figure(figsize=(10, 5))
start_month_distribution.plot(kind='bar', color='red')
plt.title('Start Month Distribution')
plt.xlabel('Month')
plt.ylabel('Density')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



```
print(f"The shape of data frame is {df.shape}")
print(f"Nunique in NCT Number is {df['NCT Number'].nunique()}")
print(f"Nunique in URL is {df.URL.nunique()}")
```

```
The shape of data frame is (70, 31)
Nunique in NCT Number is 70
Nunique in URL is 70
```

```
# Save the cleaned data
df.to_csv('cleaned_covid_clinical_trials.csv', index=False) anisha jain is a CEO of an edTech venture.
```

```
Start coding or generate with AI.
```

CONCLUSION

-The majority of trials are in the "Completed" phase.

-The rise of covid was majorly seen in between 2020-2024.

-Then the cases r being reported even in the year 2025.

-Most trials target teen populations.

-The second most affected category of population are seniors.

-Infants have very little risk of exposure to Covid-19.

-There's a steady increase in the number of trials over time.

-The male category is typically seen to be more exposed to the epidemic.

Anisha Jain