untitled12

July 16, 2025

0.0.1 COVID-19 Clinical Trials EDA

[2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

[4]:
```
data = pd.read_csv('COVID clinical trials.csv')
```

0.0.2 EDA

[8]:
```
# View the first few row of the dataset
print(data.head())
```

```
   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

        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

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

                                    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…

                               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

         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

       First Posted Results First Posted Last Update Posted  \
0    March 8, 2021                  NaN      March 8, 2021
1  October 20, 2020                 NaN   October 20, 2020
2     May 20, 2020                  NaN   November 9, 2020
3     June 4, 2020                  NaN       June 4, 2020
4     May 20, 2020                  NaN       June 4, 2020

                                          Locations Study Documents  \
0  Groupe Hospitalier Paris Saint-Joseph, Paris, …             NaN
1      Cimedical, Barranquilla, Atlantico, Colombia               NaN
2  Ospedale Papa Giovanni XXIII, Bergamo, Italy|P…             NaN
3  Hong Kong Sanatorium & Hospital, Hong Kong, Ho…             NaN
4                      CHR Orléans, Orléans, France               NaN

                                        URL
0  https://ClinicalTrials.gov/show/NCT04785898
1  https://ClinicalTrials.gov/show/NCT04595136
2  https://ClinicalTrials.gov/show/NCT04395482
3  https://ClinicalTrials.gov/show/NCT04416061
4  https://ClinicalTrials.gov/show/NCT04395924

[5 rows x 27 columns]
```

[10]:
```
# Check columns and data types
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5783 entries, 0 to 5782
Data columns (total 27 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Rank                     5783 non-null   int64
 1   NCT Number               5783 non-null   object
 2   Title                    5783 non-null   object
 3   Acronym                  2480 non-null   object
 4   Status                   5783 non-null   object
 5   Study Results            5783 non-null   object
 6   Conditions               5783 non-null   object
 7   Interventions            4897 non-null   object
 8   Outcome Measures         5748 non-null   object
 9   Sponsor/Collaborators    5783 non-null   object
 10  Gender                   5773 non-null   object
 11  Age                      5783 non-null   object
 12  Phases                   3322 non-null   object
 13  Enrollment               5749 non-null   float64
 14  Funded Bys               5783 non-null   object
 15  Study Type               5783 non-null   object
 16  Study Designs            5748 non-null   object
 17  Other IDs                5782 non-null   object
 18  Start Date               5749 non-null   object
 19  Primary Completion Date  5747 non-null   object
 20  Completion Date          5747 non-null   object
 21  First Posted             5783 non-null   object
 22  Results First Posted     36 non-null     object
 23  Last Update Posted       5783 non-null   object
 24  Locations                5198 non-null   object
 25  Study Documents          182 non-null    object
 26  URL                      5783 non-null   object
dtypes: float64(1), int64(1), object(25)
memory usage: 1.2+ MB
None
```

[12]:
```
# Summary statistics for numeric columns
print(data.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
```

0.0.3 Handling Missing Data

[22]:
```
# Columns in dataset
data.columns
```

[22]:
```
Index(['Rank', 'NCT Number', 'Title', 'Acronym', 'Status', 'Study Results',
       'Conditions', 'Interventions', 'Outcome Measures',
       'Sponsor/Collaborators', 'Gender', 'Age', 'Phases', 'Enrollment',
       'Funded Bys', 'Study Type', 'Study Designs', 'Other IDs', 'Start Date',
       'Primary Completion Date', 'Completion Date', 'First Posted',
       'Results First Posted', 'Last Update Posted', 'Locations',
       'Study Documents', 'URL'],
      dtype='object')
```

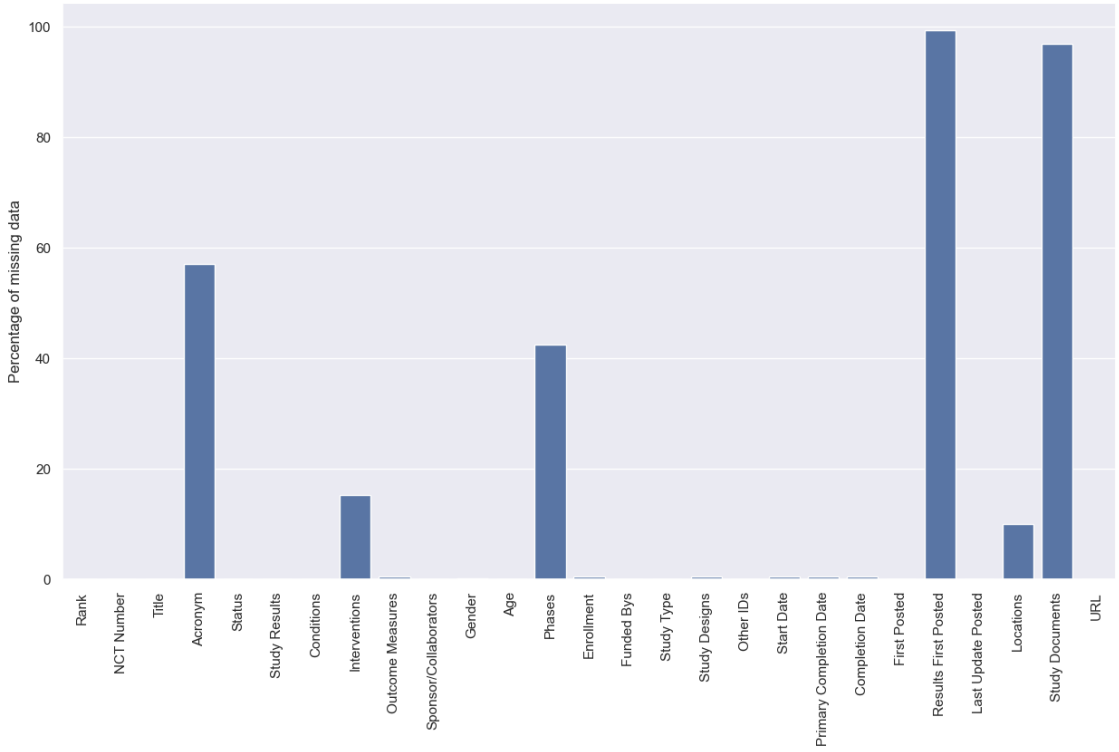
[6]:
```
# Detecting(Percentage) Missing Data
missing_data = data.isnull().mean()*100
missing_data
```

[6]:
```
Rank                        0.000000
NCT Number                  0.000000
Title                       0.000000
Acronym                    57.115684
Status                      0.000000
Study Results               0.000000
Conditions                  0.000000
Interventions              15.320768
Outcome Measures            0.605222
Sponsor/Collaborators       0.000000
Gender                      0.172921
Age                         0.000000
Phases                     42.555767
Enrollment                  0.587930
Funded Bys                  0.000000
Study Type                  0.000000
Study Designs               0.605222
Other IDs                   0.017292
Start Date                  0.587930
Primary Completion Date     0.622514
Completion Date             0.622514
First Posted                0.000000
Results First Posted       99.377486
Last Update Posted          0.000000
Locations                  10.115857
```

```
Study Documents            96.852845
URL                         0.000000
dtype: float64
```

[8]:
```
# Visualize data without calculating
def visualize_data(data,caption='',ylabel='Percentage of missing data'):
    sns.set(rc={'figure.figsize':(15,8.27)})
    plt.xticks(rotation=90)
# set title to the image and plot it or the highest 40
    fig = sns.barplot(x = data.keys()[:min(40,len(data))].tolist(),y=data.
 ↪values[:min(40,len(data))].tolist())
    plt.ylabel(ylabel)
    plt.show()
```

[10]:
```
visualize_data(missing_data,'Percentage of missing data in each feature')
```



0.0.4 As show the percentage of missing data in Result First Posted is 99.3% and

0.0.5 Study Documents is 96.8% so its impossible to impute them without destroying our dataset

[12]:
```
# Drop this columns
data.drop(['Results First Posted','Study Documents'],inplace=True,axis=1)
```

[14]:
```
# Drop the duplicate Rows
print(f"Shape before droping duplicate data{data.shape}")
data.drop_duplicates(inplace=True)
print(f"Shape after droping duplicate data {data.shape}")
```

```
Shape before droping duplicate data(5783, 25)
Shape after droping duplicate data (5783, 25)
```

[16]:
```
data.isnull().mean()*100
```

[16]:
```
Rank                        0.000000
NCT Number                  0.000000
Title                       0.000000
Acronym                    57.115684
Status                      0.000000
Study Results               0.000000
Conditions                  0.000000
Interventions              15.320768
Outcome Measures            0.605222
Sponsor/Collaborators       0.000000
Gender                      0.172921
Age                         0.000000
Phases                     42.555767
Enrollment                  0.587930
Funded Bys                  0.000000
Study Type                  0.000000
Study Designs               0.605222
Other IDs                   0.017292
Start Date                  0.587930
Primary Completion Date     0.622514
Completion Date             0.622514
First Posted                0.000000
Last Update Posted          0.000000
Locations                  10.115857
URL                         0.000000
dtype: float64
```

[18]:
```
# We can extract a new feature from the location which is the country where the␣
 ↪study hold
```

```
countries = [str(data.Locations.iloc[i]).split('.')[-1] for i in range(data.
 ↪shape[0])]
data['Country'] = countries
```

[20]:
```
data.columns
```

[20]:
```
Index(['Rank', 'NCT Number', 'Title', 'Acronym', 'Status', 'Study Results',
       'Conditions', 'Interventions', 'Outcome Measures',
       'Sponsor/Collaborators', 'Gender', 'Age', 'Phases', 'Enrollment',
       'Funded Bys', 'Study Type', 'Study Designs', 'Other IDs', 'Start Date',
       'Primary Completion Date', 'Completion Date', 'First Posted',
       'Last Update Posted', 'Locations', 'URL', 'Country'],
      dtype='object')
```

[45]:
```
data.Country.value_counts()[:35]
```

[45]:
```
Country
nan
585
Uhmontpellier, Montpellier, France
19
National Institutes of Health Clinical Center, Bethesda, Maryland, United States
16
CHU Amiens, Amiens, France
13
Stanford University, Stanford, California, United States
13
Massachusetts General Hospital, Boston, Massachusetts, United States
12
M D Anderson Cancer Center, Houston, Texas, United States
11
Faculty of Medicine Ain Shams University Research Institute- Clinical Research
Center, Cairo, Non-US, Egypt     11
NYU Langone Health, New York, New York, United States
11
Brigham and Women's Hospital, Boston, Massachusetts, United States
11
Uh Montpellier, Montpellier, France
11
CHU de Nice, Nice, France
9
Jiangsu Provincial Center for Diseases Control and Prevention, Nanjing, Jiangsu,
China                            8
Hamad Medical Corporation, Doha, Qatar
8
Assistance Publique Hôpitaux de Marseille, Marseille, France
8
```

```
University of Alabama at Birmingham, Birmingham, Alabama, United States
8
Johns Hopkins Hospital, Baltimore, Maryland, United States
7
CHU Brugmann, Brussels, Belgium
7
Mayo Clinic in Rochester, Rochester, Minnesota, United States
7
Hacettepe University, Ankara, Turkey
7
University of Michigan, Ann Arbor, Michigan, United States
7
University Health Network, Toronto, Ontario, Canada
7
University of Pennsylvania, Philadelphia, Pennsylvania, United States
7
Groupe Hospitalier Paris Saint-Joseph, Paris, France
7
University Hospital of Toulouse, Toulouse, France
6
ProgenaBiome, Ventura, California, United States
6
Mayo Clinic, Rochester, Minnesota, United States
6
Pinar Yalcin Bahat, Istanbul, İstanbul, Turkey
6
University of Miami, Miami, Florida, United States
6
Oslo University Hospital, Oslo, Norway
6
Medical University of Vienna, Vienna, Austria
6
Ankara City Hospital, Ankara, Turkey
5
University of British Columbia, Vancouver, British Columbia, Canada
5
Angers University Hospital, Angers, France
5
Kadirhan Ozdemir, İzmir, Turkey
5
Name: count, dtype: int64
```

[22]:
```
# Lets start with Acronym
print(f"Number of unique values is {data.Acronym.nunique()} \n")
data.Acronym.value_counts()
```

```
Number of unique values is 2338
```

[22]:
```
Acronym
COVID-19        47
PROTECT          7
CORONA           6
RECOVER          5
SCOPE            5
                ..
ASD              1
VICO             1
LICORNE          1
LOSVID           1
MindMyMindFU     1
Name: count, Length: 2338, dtype: int64
```

[24]:
```
# Find the realtion between null values in Acronym and Countries
(data.Acronym.isnull().groupby(data.Country).mean().
 ↪sort_values(ascending=False)*100)[:60]
```

[24]:
```
Country
 ( Site 0011), Lexington, Kentucky, United States|The Center for Pharmaceutical
Research PC ( Site 0012), Kansas City, Missouri, United States|SCRI-CCCIT GesmbH
( Site 0006), Salzburg, Austria|Medizinische Universitaet Wien ( Site 0007),
Wien, Austria|Universitair Ziekenhuis Gent ( Site 0003), Gent, Oost-Vlaanderen,
Belgium|SGS Life Science Services ( Site 0001), Antwerpen, Belgium|ATC -
Clinical Pharmacology Unit ( Site 0002), Liege, Belgium
100.0
Mansoura University, Mansoura, Select A State Or Province, Egypt
100.0
Mahatma Gandhi Mission Medical College and Hospital, Aurangabad, Maharashtra,
India|Hillel Yaffe Medical Center, Hadera, Haifa, Israel|Nazareth Hospital EMMS,
Nazareth, North, Israel|Rambam Health Care Campus, Haifa, Israel
100.0
Mahmoud S Abu-Samak, Amman, Jordan
100.0
Mahmoud Tantawy, Cairo, Egypt
100.0
Manal Hassanien, Assiut, Yes, Egypt
100.0
Manna Research, Toronto, Ontario, Canada
100.0
Mansoura Faculty of Medicine, Mansoura, Egypt
100.0
Mansoura University Hospital, Mansoura, DK, Egypt
100.0
Mansoura University Hospital, Mansoura, Dakahliya, Egypt
```

100.0
Mansoura University, Mansoura, Dakahlyia, Egypt
100.0
Mansoura university, Mansoura, Egypt
100.0
Massachusetts General Hospital, Boston, Massachusetts, United States|Brigham and Women's Hospital, Boston, Massachusetts, United States|Memorial Sloan Kettering Cancer Center, New York, New York, United States
100.0
Marcello Covino, Roma, RM, Italy
100.0
Marie-France VAILLANT, Grenoble, France
100.0
Markham Stouffville Hospital, Markham, Ontario, Canada
100.0
Marmara University School of Medicine Department of Physical Medicine and Rehabilitation, Istanbul, Turkey
100.0
Marqués de Valdecilla University Hospital, Santander, Cantabria, Spain|Puerta de Hierro University Hospital, Majadahonda, Madrid, Spain|Navarra University Hospital, Pamplona, Navarra, Spain|Barcelona Clinic University Hospital, Barcelona, Spain|Reina Sofía University Hospital, Córdoba, Spain|Ramón y Cajal University Hospital, Madrid, Spain|Fundación Jiménez Díaz University Hospital, Madrid, Spain|12 de Octubre University Hospital, Madrid, Spain|Virgen del Rocío University Hospital, Sevilla, Spain|Araba University Hospital, Vitoria, Álava, Spain 100.0
Marta Caballero, Hospitalet de Llobregat, Barcelona, Spain
100.0
Marta de la plaza, Madrid, Spain
100.0
Marwa Eid, Cairo, Egypt
100.0
Mary's Hospital, Seoul, Korea, Republic of
100.0
Mahanagar General Hospital, Dhaka (Site-1), Mugda Medical College Hospital, Dhaka (Site-2), Kurmitola General Hospital, Dhaka (Site-3), Dhaka Medical College Hospital, Dhaka (Site-4), Dhaka, Bangladesh
100.0
Maha M Farid, Cairo, Egypt
100.0
MEBO Research, Inc, Miami, Florida, United States|Kahite, Vonore, Tennessee, United States|Mary Washington Hospital Research, Fredericksburg, Virginia, United States|Mebo Research (Uk), London, England, United Kingdom
100.0
MD Mount Sinai, Baltimore, Maryland, United States|Hospital das Clinicas Ribeirao Preto, Ribeirão Preto, San Paulo, Brazil|Danish National Biobank, København, Denmark|Shonan General Hospital, Kamakura, Kanagawa, Japan|National

Center Global Health and Medicine, Shinjuku, Tokyo, Japan|Yamanashi Prefectural Central Hospital, Kōfu, Yamanashi, Japan|Unilab Group, Manila, Philippines
100.0
Linear Clinical Research - Harry Perkins Research Institute, Nedlands, Western Australia, Australia
100.0
Linear Clinical Research Ltd, Nedlands, Western Australia, Australia
100.0
Ling Liu, Nanjing, Jiangsu, China
100.0
Linköping University, Linköping, Östergötland, Sweden
100.0
Liverpool University Hospitals NHS Foundation Trust, Liverpool, United Kingdom|University Hospital Southampton NHS Foundation Trust, Southampton, United Kingdom
100.0
London Health Science Centre, London, Ontario, Canada
100.0
Los Angeles County-USC Medical Center, Los Angeles, California, United States|USC / Norris Comprehensive Cancer Center, Los Angeles, California, United States
100.0
Los Angeles Homeless Services Authority (LAHSA), Los Angeles, California, United States
100.0
Louis Mourier hospital (AP-HP), Colombes, France|Brabois Hospital (CHRU de Nancy), Vandœuvre-lès-Nancy, France
100.0
Lowell General Hospital, Lowell, Massachusetts, United States
100.0
Ludwig-Maximilians-Universität München, München, Bayern, Germany|Medizinische Hochschule Hannover, Hannover, Niedersachsen, Germany|SocraTec R&D GmbH, Erfurt, Thüringen, Germany
100.0
Luigi Sacco University Hospital, Milan, Lombardia, Italy
100.0
Lund ED, Lund, Sweden
100.0
Lund University, Lund, Sweden
100.0
Lymphoma and Leukemia Society, Rye Brook, New York, United States
100.0
M D Anderson Cancer Center, Houston, Texas, United States
100.0
MAC Clinical Research Manchester (Early Phase Unit), Neuroscience Centre of Excellence, Manchester, Greater Mancherster, United Kingdom
100.0

MAX HEALTH, Subsero Health 2055 Wood Street, Suite 100, Sarasota, Florida, United States
100.0
MD Anderson in The Woodlands, Conroe, Texas, United States|M D Anderson Cancer Center, Houston, Texas, United States|MD Anderson League City, League City, Texas, United States|MD Anderson in Sugar Land, Sugar Land, Texas, United States
100.0
Massachusetts General Hospital, Boston, Massachusetts, United States|Brigham and Women's Hospital, Boston, Massachusetts, United States|Laboratories of Cognitive Neuroscience, Boston Children's Hospital, Boston, Massachusetts, United States
100.0
Massachusetts General Hospital, Boston, Massachusetts, United States|Brigham and Women's Hospital, Boston, Massachusetts, United States|Newton-Wellesley Hospital, Newton, Massachusetts, United States
100.0
LifeFactors Zona Franca SAS, Medellín, Antioquia, Colombia
100.0
Medeniyet University, Istanbul, Turkey
100.0
Mayo Clinic in Rochester, Rochester, Minnesota, United States
100.0
Mayo Clinic, Rochester, Minnesota, United States|Hospital Clínico San Carlos, Madrid, Spain
100.0
Mayo Clinic, Rochester, Minnesota, United States|Sri Jayadeva, Bengaluru, Karnataka, India|Gregorio Marañon Hospital, Madrid, Spain|Imperial College, London, England, United Kingdom
100.0
Mays Cancer Center, UT Health San Antonio, San Antonio, Texas, United States
100.0
Mayte Serrat, Barcelona, Spain
100.0
McGill University, Montreal, Quebec, Canada
100.0
McMaster Cardio-Respiratory Research Lab, Hamilton, Ontario, Canada
100.0
MedStar Georgetown University Hospital, Washington, District of Columbia, United States|Washington Hospital Center, Washington, District of Columbia, United States|National Institutes of Health Clinical Center, Bethesda, Maryland, United States
100.0
MedStar Health Research Institute /MedStar Washington Hospital Center, Washington, District of Columbia, United States|National Institutes of Health Clinical Center, Bethesda, Maryland, United States
100.0
Medialis, Oxford, United Kingdom
100.0

```
Massachusetts General Hospital, Boston, Massachusetts, United States|King's
College London, London, United Kingdom
100.0
Name: Acronym, dtype: float64
```

0.0.6 After inspecting the relation between the missing values in Acronym and

0.0.7 Country we can conclude that there is a sort of relation between these two

0.0.8 features, so we can say that Data is Missing At Random (MAR).

0.0.9 So we can Impute by Missing Category.

[26]:
```
# impute by a missing Indicator
data.Acronym = data.Acronym.fillna("Missing Acronym")
```

[28]:
```
# Detecting (Percentage) Missing Data
data.isnull().mean() * 100
```

[28]:
```
Rank                          0.000000
NCT Number                    0.000000
Title                         0.000000
Acronym                       0.000000
Status                        0.000000
Study Results                 0.000000
Conditions                    0.000000
Interventions                15.320768
Outcome Measures              0.605222
Sponsor/Collaborators         0.000000
Gender                        0.172921
Age                           0.000000
Phases                       42.555767
Enrollment                    0.587930
Funded Bys                    0.000000
Study Type                    0.000000
Study Designs                 0.605222
Other IDs                     0.017292
Start Date                    0.587930
Primary Completion Date       0.622514
Completion Date               0.622514
First Posted                  0.000000
Last Update Posted            0.000000
Locations                    10.115857
URL                           0.000000
Country                       0.000000
dtype: float64
```

0.0.10 We can do the same for other categorical features such as Interventions , Phases ,

0.0.11 Locations and other categorical features

[30]:
```
# Impute Interventions , Phases , Locations by Missing Category

categorical_features = data.select_dtypes(include =object).columns


features =categorical_features[data[categorical_features].isnull().mean() > 0]


for feature in features:
    data[feature] = data[feature].fillna(f"Missing {feature}")
```

[32]:
```
# Detecting (Percentage) Missing Data
data.isnull().mean() * 100
```

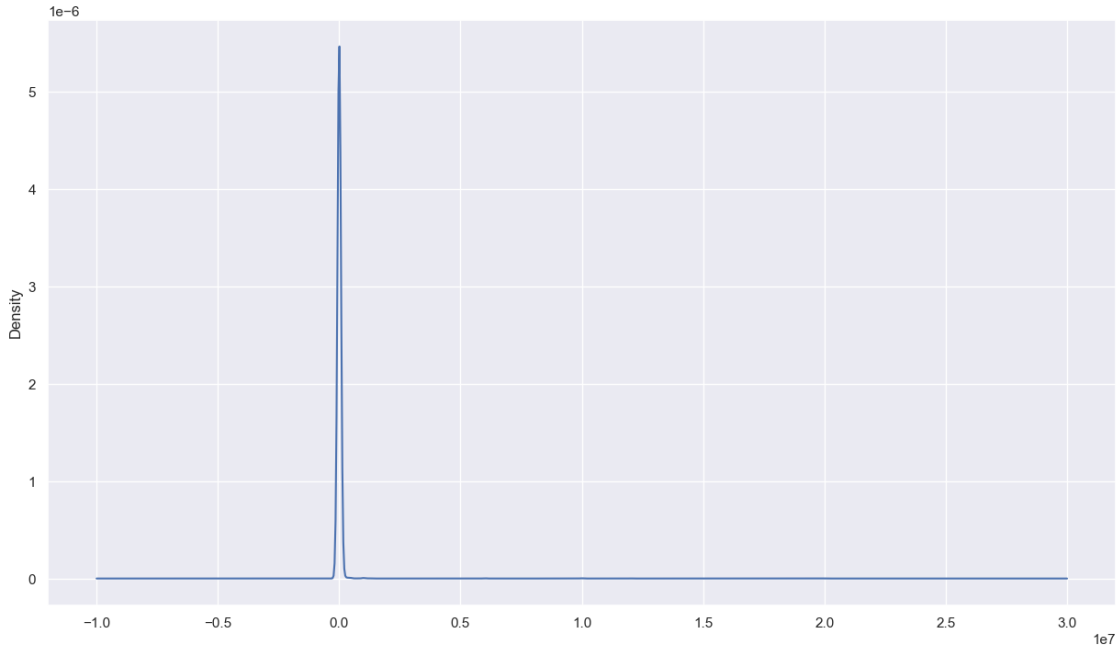
[32]:
```
Rank                        0.00000
NCT Number                  0.00000
Title                       0.00000
Acronym                     0.00000
Status                      0.00000
Study Results               0.00000
Conditions                  0.00000
Interventions               0.00000
Outcome Measures            0.00000
Sponsor/Collaborators       0.00000
Gender                      0.00000
Age                         0.00000
Phases                      0.00000
Enrollment                  0.58793
Funded Bys                  0.00000
Study Type                  0.00000
Study Designs               0.00000
Other IDs                   0.00000
Start Date                  0.00000
Primary Completion Date     0.00000
Completion Date             0.00000
First Posted                0.00000
Last Update Posted          0.00000
Locations                   0.00000
URL                         0.00000
Country                     0.00000
dtype: float64
```

[34]:
```
# Now the Time to handle The missing data for the Enrollment
# Check the skewness
data.Enrollment.skew()
# The value of Skewness is 34 which means that we This feature isn't normally␣
 ↪distributed
```

[34]:
```
34.06593382031148
```

[36]:
```
# Plotting the distribution of the enrollment
data.Enrollment.plot(kind = 'kde')

# So We will impute by the median
```

[36]:
```
<Axes: ylabel='Density'>
```



[38]:
```
# Some Statstical Valuse for the Enrollment Column

min_Value = data.Enrollment.min()
max_Value = data.Enrollment.max()
mean_Value = data.Enrollment.mean()
median_Value = data.Enrollment.median()
std_Value = data.Enrollment.std()

print(f"the min value is {min_Value} \n \
The max value is {max_Value} \n \
The mean is {mean_Value} \n \
```

```
The Median is {median_Value} \n \
Standard Devation is {std_Value}")
```

```
the min value is 0.0
 The max value is 20000000.0
 The mean is 18319.48860671421
 The Median is 170.0
 Standard Devation is 404543.7287841073
```

[42]:
```
# Using Median to impute Missing Values
data.Enrollment = data.Enrollment.fillna(median_Value)
```

[44]:
```
# Detecting (Percentage) Missing Data
data.isnull().mean() * 100
```

[44]:
```
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
Last Update Posted          0.0
Locations                   0.0
URL                         0.0
Country                     0.0
dtype: float64
```

[46]:
```
data.head()
```

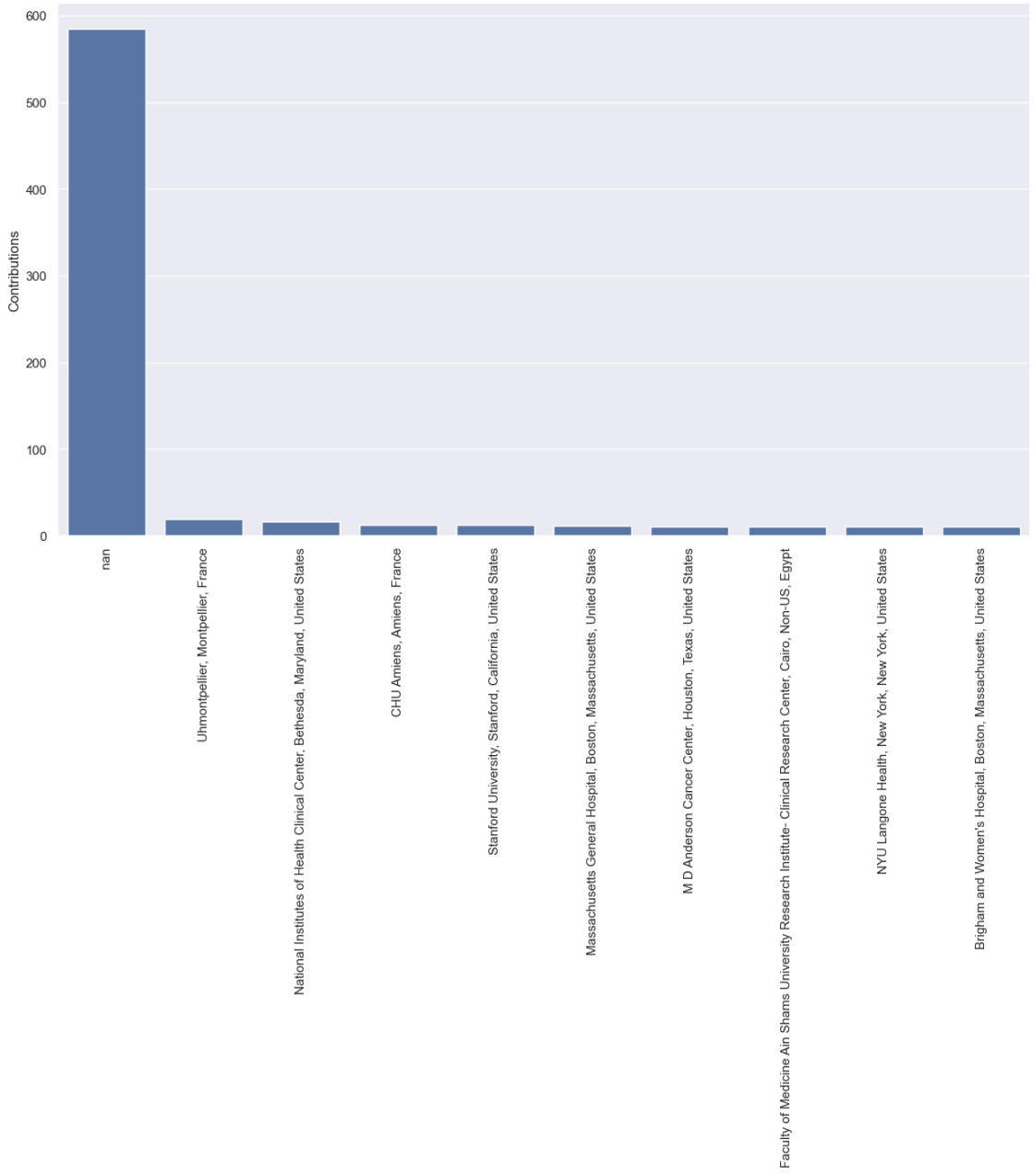
[46]:
```
    Rank   NCT Number                                              Title  \
0      1  NCT04785898  Diagnostic Performance of the ID Now  COVID-19…
```

```
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

       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

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

                                   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…

                              Sponsor/Collaborators  …  \
0             Groupe Hospitalier Paris Saint Joseph  …
1                       United Medical Specialties  …
2                    University of Milano Bicocca  …
3                  Hong Kong Sanatorium & Hospital  …
4  Centre Hospitalier Régional d'Orléans|Centre d…  …

                                   Study Designs         Other IDs  \
0  Allocation: N/A|Intervention Model: Single Gro…       COVID-IDNow
1  Allocation: Randomized|Intervention Model: Par…  COVID19-0001-USR
2  Observational Model: Cohort|Time Perspective: …       TAC-COVID19
3  Observational Model: Cohort|Time Perspective: …        RC-2020-08
4  Observational Model: Cohort|Time Perspective: …      CHRO-2020-10
```

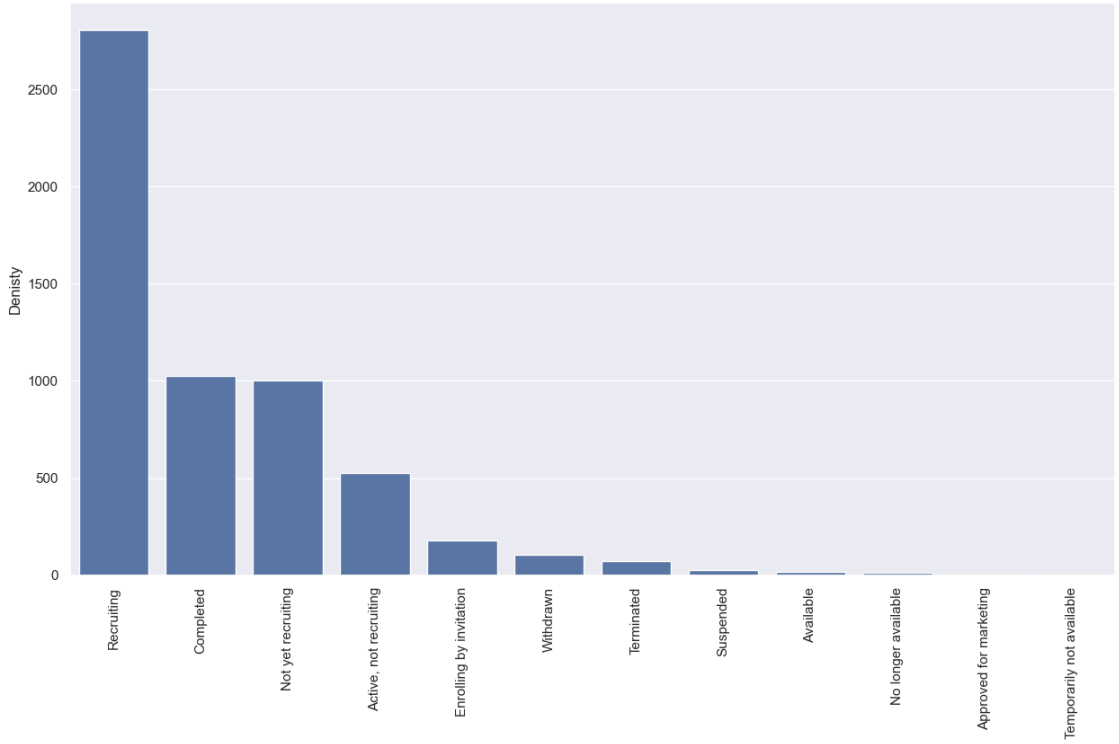
```
         Start Date  Primary Completion Date   Completion Date  \
0  November 9, 2020        December 22, 2020    April 30, 2021
1  November 2, 2020        December 15, 2020  January 29, 2021
2       May 7, 2020            June 15, 2021     June 15, 2021
3      May 25, 2020            July 31, 2020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

       First Posted Last Update Posted  \
0     March 8, 2021      March 8, 2021
1  October 20, 2020   October 20, 2020
2      May 20, 2020   November 9, 2020
3      June 4, 2020       June 4, 2020
4      May 20, 2020       June 4, 2020

                                          Locations  \
0  Groupe Hospitalier Paris Saint-Joseph, Paris, …
1       Cimedical, Barranquilla, Atlantico, Colombia
2  Ospedale Papa Giovanni XXIII, Bergamo, Italy|P…
3  Hong Kong Sanatorium & Hospital, Hong Kong, Ho…
4                      CHR Orléans, Orléans, France

                                          URL  \
0  https://ClinicalTrials.gov/show/NCT04785898
1  https://ClinicalTrials.gov/show/NCT04595136
2  https://ClinicalTrials.gov/show/NCT04395482
3  https://ClinicalTrials.gov/show/NCT04416061
4  https://ClinicalTrials.gov/show/NCT04395924

                                            Country
0  Groupe Hospitalier Paris Saint-Joseph, Paris, …
1       Cimedical, Barranquilla, Atlantico, Colombia
2  Ospedale Papa Giovanni XXIII, Bergamo, Italy|P…
3  Hong Kong Sanatorium & Hospital, Hong Kong, Ho…
4                      CHR Orléans, Orléans, France

[5 rows x 26 columns]
```

0.0.12 Data Visualization

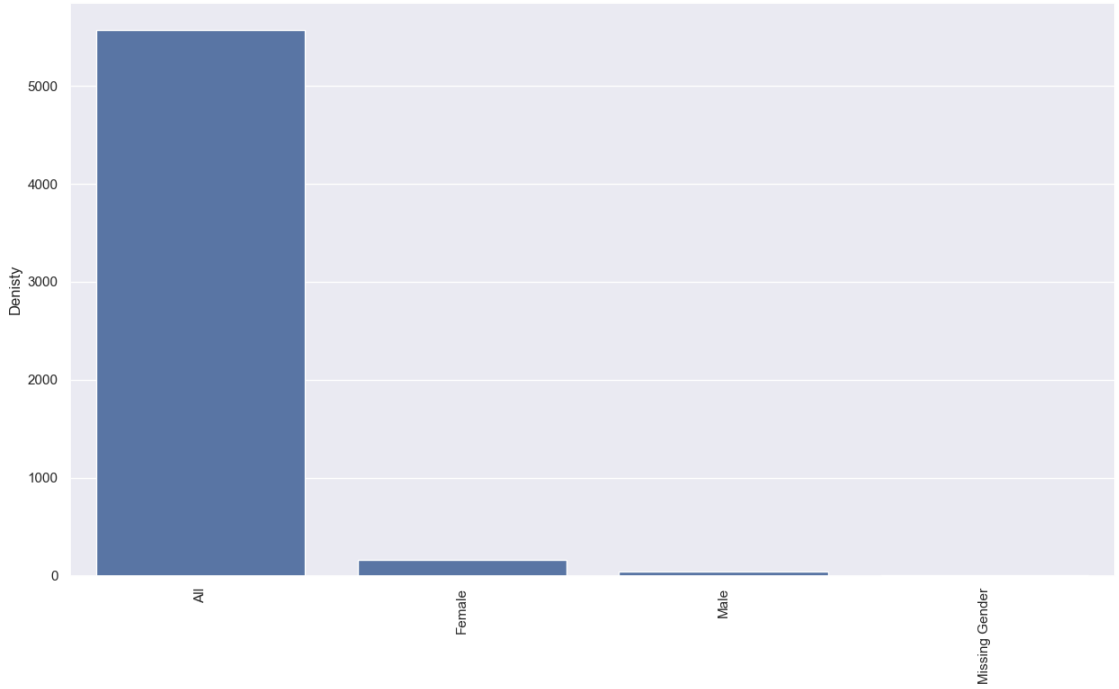
[49]:
```
# Get Countires with highest Contributiuons
top_10_Countires = data.Country.value_counts()[:10]
visualize_data(top_10_Countires , caption = 'Top 10 Countries', ylabel =␣
 ↪'Contributions')
```



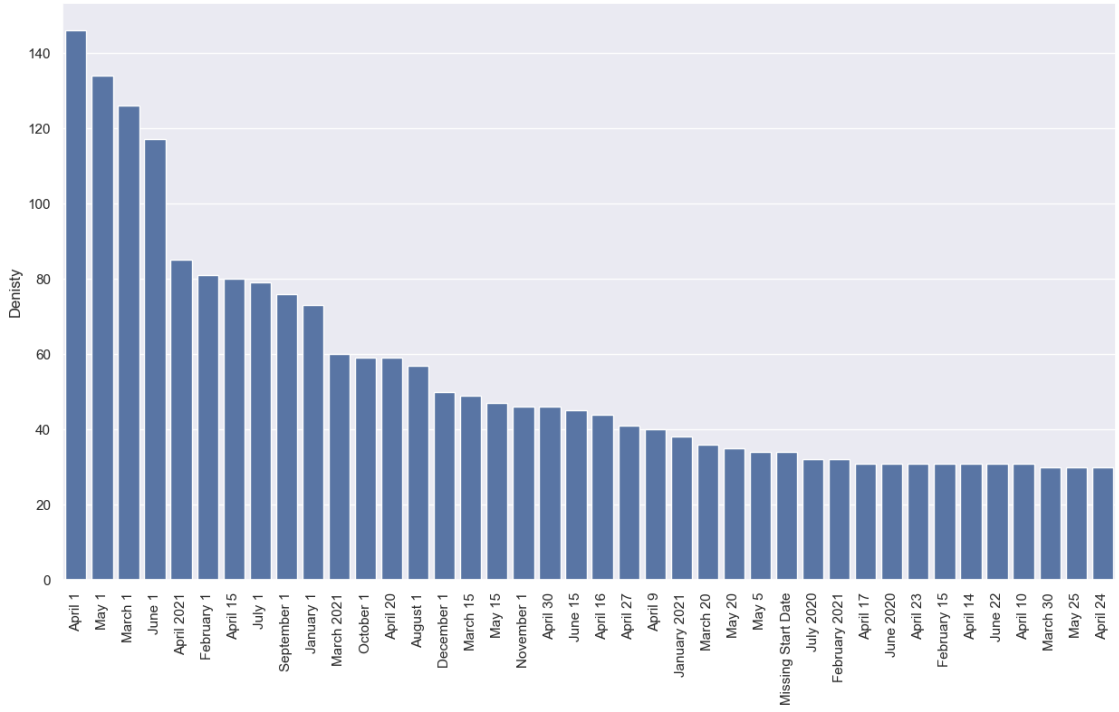
[53]:
```
# Status of the Application
status = data.Status.value_counts()

visualize_data(status , caption = 'Status of The Application' ,ylabel =␣
 ↪'Denisty')
```



[55]:
```
# Gender Visualiztions
gender = data.Gender.value_counts()
visualize_data(gender , caption = 'Gender Distribution' ,ylabel = 'Denisty')
```



[63]:
```
# Which month has the highest start
start_month = pd.Series([ str(data['Start Date'].iloc[i]).split(',')[0] for i␣
 ↪in range (data.shape[0])])


start_month_Distribution = start_month.value_counts()

visualize_data(start_month_Distribution , caption = 'Start Month Distribution'␣
 ↪, ylabel = 'Denisty')
```



[65]:
```
print(f"The shape of data frame is {data.shape}")
print(f"Nunique in NCT Number is {data['NCT Number'].nunique()}")
print(f"Nunique in URL is {data.URL.nunique()}")
```

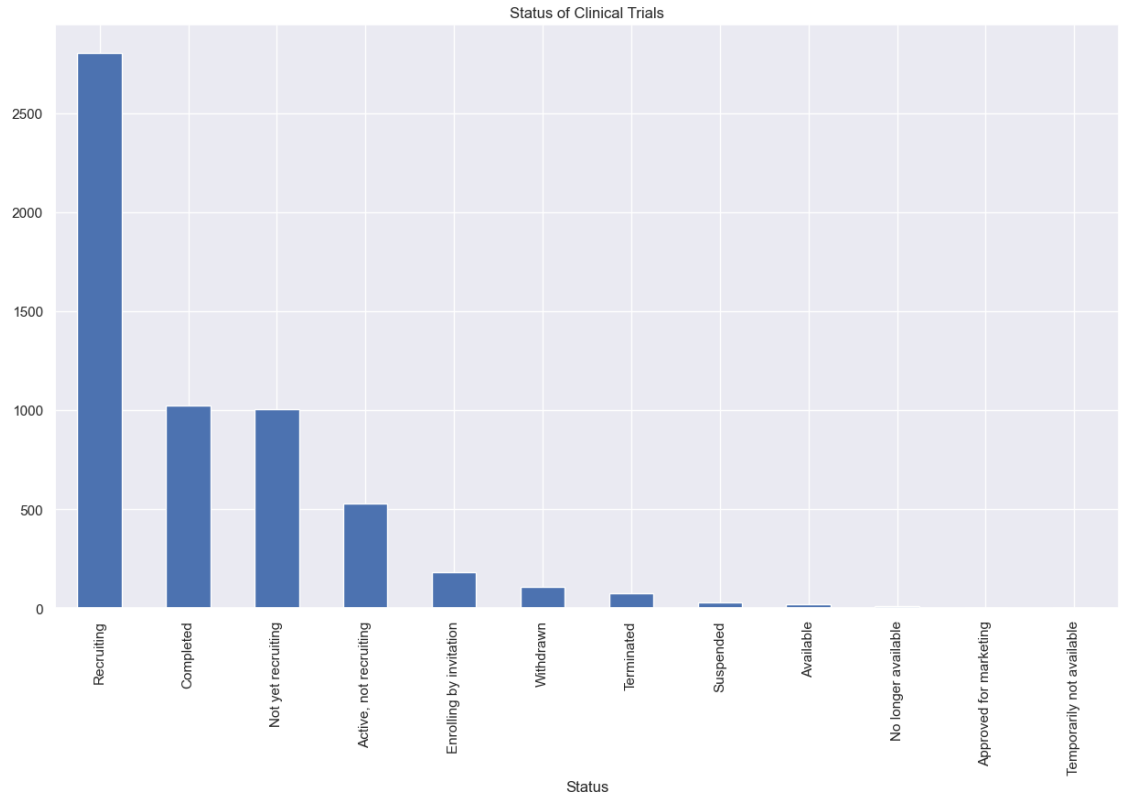
```
The shape of data frame is (5783, 26)
Nunique in NCT Number is 5783
Nunique in URL is 5783
```

0.0.13 Univariate Analysis

[68]:
```
# Status Distribution: Analyze the status of clinical trials
print(data['Status'].value_counts())
data['Status'].value_counts().plot(kind='bar', title='Status of Clinical␣
 ↪Trials')
```

```
Status
Recruiting                  2805
Completed                   1025
Not yet recruiting          1004
Active, not recruiting       526
Enrolling by invitation      181
Withdrawn                    107
Terminated                    74
Suspended                     27
Available                     19
No longer available           12
Approved for marketing         2
Temporarily not available      1
Name: count, dtype: int64
```

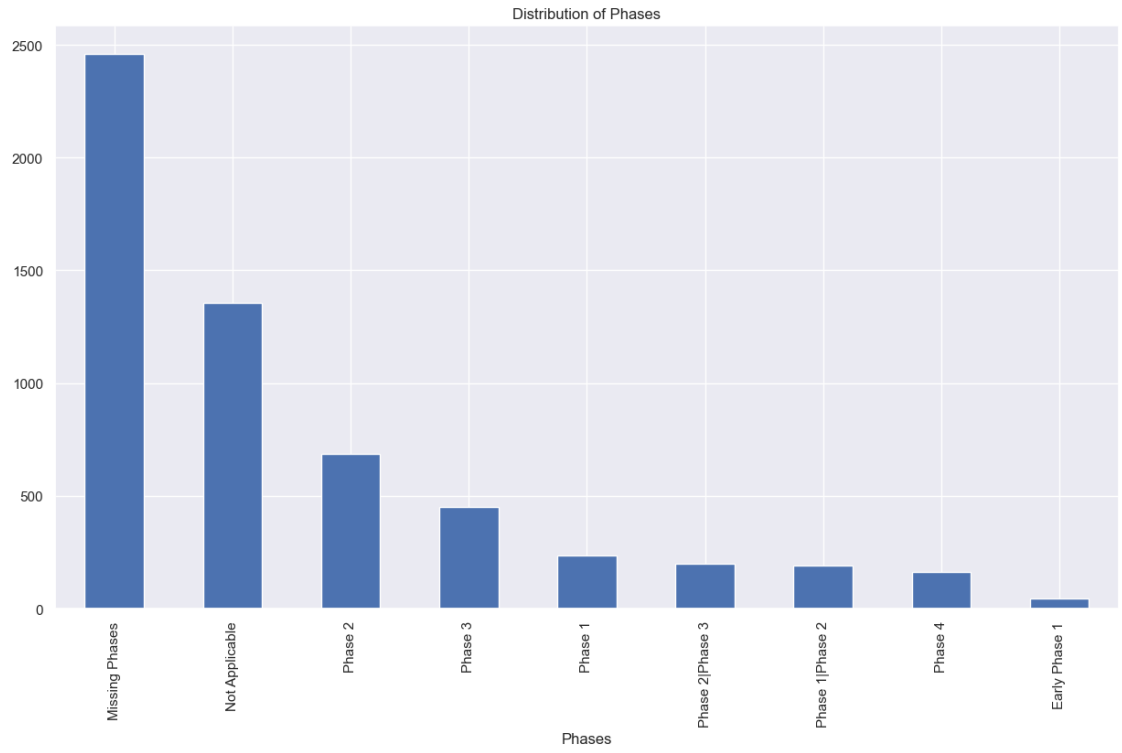
[68]: <Axes: title={'center': 'Status of Clinical Trials'}, xlabel='Status'>



[70]:
```
# Phase Distribution: Understand the distribution of trial phases.
print(data['Phases'].value_counts())
data['Phases'].value_counts().plot(kind='bar',title='Distribution of Phases')
```

```
Phases
Missing Phases     2461
Not Applicable     1354
Phase 2             685
Phase 3             450
Phase 1             234
Phase 2|Phase 3     200
Phase 1|Phase 2     192
Phase 4             161
Early Phase 1        46
Name: count, dtype: int64
```

[70]: <Axes: title={'center': 'Distribution of Phases'}, xlabel='Phases'>



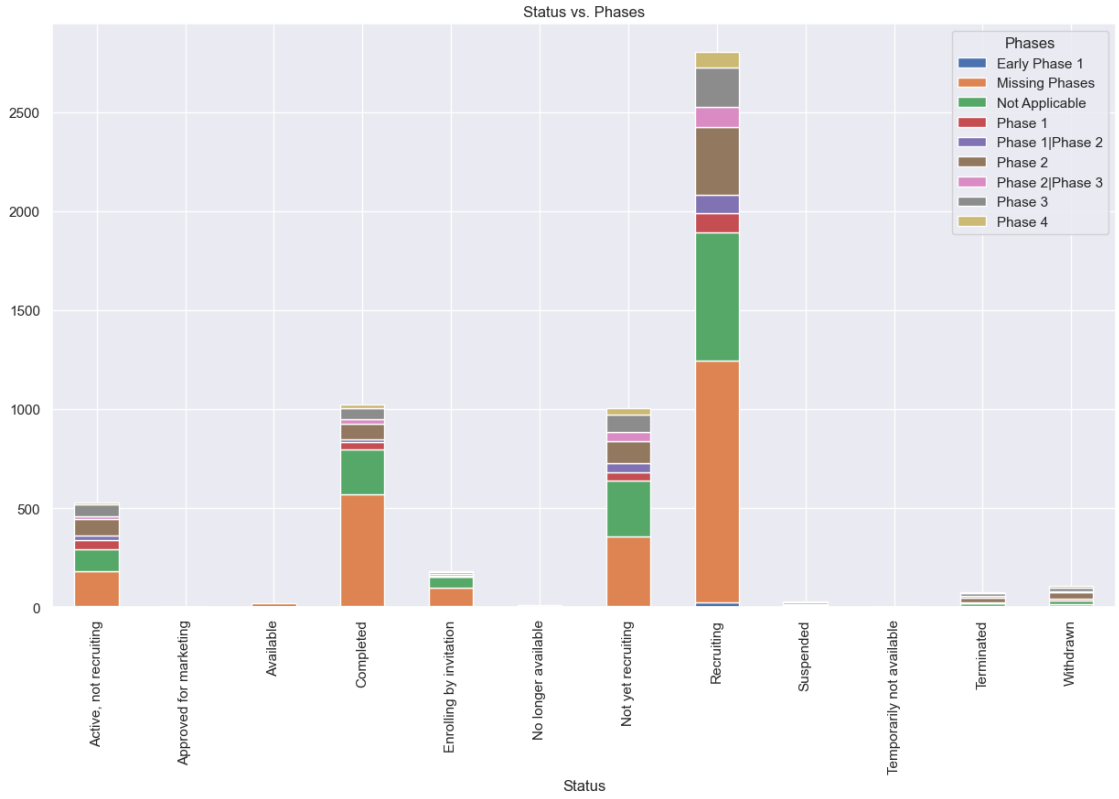
0.0.14 Bivariate Analysis

[75]:
```
#Status vs. Phases: Explore how trial phases are distributed across different statues

status_phase = pd.crosstab(data['Status'], data['Phases'])
print(status_phase)
status_phase.plot(kind='bar', stacked=True, title='Status vs. Phases')
```

```
Phases                      Early Phase 1  Missing Phases  Not Applicable  \
Status
Active, not recruiting                  7             175             111
Approved for marketing                  0               2               0
Available                               0              19               0
Completed                               3             565             226
Enrolling by invitation                 4              96              54
No longer available                     0              12               0
Not yet recruiting                      5             350             282
Recruiting                             22            1224             647
Suspended                               2               2               2
Temporarily not available               0               1               0
Terminated                              0               4              13
Withdrawn                               3              11              19

Phases                      Phase 1  Phase 1|Phase 2  Phase 2  Phase 2|Phase 3  \
Status
Active, not recruiting           44               26       81               15
Approved for marketing            0                0        0                0
Available                         0                0        0                0
Completed                        38               17       78               20
Enrolling by invitation           1                3       10                1
No longer available               0                0        0                0
Not yet recruiting               42               46      114               46
Recruiting                       98               92      343              102
Suspended                         0                2        4                4
Temporarily not available         0                0        0                0
Terminated                        4                2       25                6
Withdrawn                         7                4       30                6

Phases                      Phase 3  Phase 4
Status
Active, not recruiting           59        8
Approved for marketing            0        0
Available                         0        0
Completed                        56       22
Enrolling by invitation           6        6
No longer available               0        0
Not yet recruiting               89       30
```

```
Recruiting                      196        81
Suspended                         9         2
Temporarily not available         0         0
Terminated                       15         5
Withdrawn                        20         7
```

[75]: <Axes: title={'center': 'Status vs. Phases'}, xlabel='Status'>



[77]:
```
#Conditions vs. Outcome Measures: Understand the common outcome measures for␣
 ↪different conditions.

conditions_outcomes = data.groupby('Conditions')['Outcome Measures'].
 ↪apply(lambda x: ', '.join(x)).reset_index()

print(conditions_outcomes)
```

```
                                           Conditions  \
0                             2019 Novel Coronavirus
1                       2019 Novel Coronavirus Infection
2       2019 Novel Coronavirus Infection|COVID-19 Viru…
3                       2019 Novel Coronavirus Pneumonia
4              2019 Novel Coronavirus Pneumonia|COVID-19
```

```
…                                                     …
3062                    the Lung Complication of COVID-19
3063  the Prognostic Value of Ferritin|Glycosylated …
3064  the Study Focus on the Uses of Telephone and O…
3065  the Use of Modern Technology Applications in H…
3066  to Predict an Unfavorable Evolution of Covid-1…

                                          Outcome Measures
0     Proportion of participants who improve by at l…
1     new-onset COVID-19|Number of Participants with…
2     Number of participants with treatment emergent…
3     Clinical recovery time|Complete fever time|Cou…
4     Pneumonia severity index|Oxygenation index (Pa…
…                                                     …
3062  lung injury score|Angiotensin 1-7 (Ang 1-7) ch…
3063  assessment of the prognostic value of ferritin…
3064  - To provide an overview about the pros and co…
3065  rate of reassurance delivered from doctors to …
3066  Need of mechanical ventilation, transfer to an…

[3067 rows x 2 columns]
```

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]: