|  |  |
| --- | --- |
| Project Title | **COVID-19 Clinical Trials EDA Pandas** |
| Tools | Python, ML, SQL, Excel |
| Domain | Data Analyst & Data scientist |
| Project Difficulties level | intermediate |

Dataset : Dataset is available in the given link. You can download it at your convenience.

[Click](https://drive.google.com/file/d/1WessMr8tdADxk04uVT-AaoP3_CxwJIYR/view?usp=sharing) [here](https://drive.google.com/file/d/1WessMr8tdADxk04uVT-AaoP3_CxwJIYR/view?usp=sharing) [to](https://drive.google.com/file/d/1WessMr8tdADxk04uVT-AaoP3_CxwJIYR/view?usp=sharing) [download](https://drive.google.com/file/d/1WessMr8tdADxk04uVT-AaoP3_CxwJIYR/view?usp=sharing) [data](https://drive.google.com/file/d/1WessMr8tdADxk04uVT-AaoP3_CxwJIYR/view?usp=sharing) [set](https://drive.google.com/file/d/1WessMr8tdADxk04uVT-AaoP3_CxwJIYR/view?usp=sharing)

# About Dataset Dataset Description

[ClinicalTrials.gov](https://www.clinicaltrials.gov/ct2/home) is a database of privately and publicly funded clinical studies conducted around the world. It is maintained by the National Institute of Health. All data is publicly available and the site provides a direct download feature which makes it super easy to use relevant data for analysis.

This dataset consists of clinical trials related to COVID 19 studies presented on the site.

The dataset consists of XML files where each XML file corresponds to one study. The filename is the NCT number which is a unique identifier of a study in the ClinicalTrials repository. Additionally, a CSV file has also been provided, which might not have as much information as contained in the XML file, but does give sufficient information.

Please refer to this notebook for details on the dataset :

<https://www.kaggle.com/parulpandey/eda-on-covid-19-clinical-trials>

# Acknowledgements

ClinicalTrials.gov is a resource provided by the U.S. National Library of Medicine.

**IMPORTANT:**

Listing a study does not mean it has been evaluated by the U.S. Federal Government.

Read our disclaimer for details.

Before participating in a study, talk to your health care provider and learn about the risks and potential benefits.

**NOTE :**

1. **this project is only for your guidance, not exactly the same you have to create. Here I am trying to show the way or idea of what steps you can follow and how your projects look. Some projects are very advanced (because it will be made with the help of flask, nlp, advance ai, advance DL and some advanced things ) which you can not understand .**
2. **You can make or analyze your project with yourself, with your idea, make it more creative from where we can get some information and understand about our business. make sure what overall things you have created all things you understand very well.**

**Example: You can get the basic idea how you can create a project from here**

Here's a step-by-step guide for performing Exploratory Data Analysis (EDA) on a COVID-19 Clinical Trials dataset using Pandas, tailored for beginners.

**Project Title:**

# Exploratory Data Analysis of COVID-19 Clinical Trials

## 1. Objective

The objective is to explore the dataset to gain insights into the characteristics of COVID-19 clinical trials, such as their status, phases, study designs, and demographics.

## 2. Importing Libraries and Loading Data

First, you'll need to import the necessary libraries and load your dataset.

import pandas as pd

# Load the dataset df = pd.read\_csv('covid\_clinical\_trials.csv') # Replace with your dataset's path

## 3. Initial Data Exploration

|  |
| --- |
| Start by exploring the basic structure and content of the dataset.  # View the first few rows of the dataset print(df.head())  # Check the columns and data types print(df.info())  # Summary statistics for numerical columns print(df.describe())  # Summary statistics for categorical columns print(df.describe(include='object'))  **4. Handling Missing Data**  Check for missing values and decide how to handle them.  # Check for missing values print(df.isnull().sum())  # Drop columns with a high percentage of missing values or fill them df = df.drop(columns=['Acronym', 'Study Documents']) # Example of dropping columns df['Results First Posted'].fillna('Unknown', inplace=True) # |

|  |
| --- |
| Example of filling missing data   1. **Univariate Analysis**   Analyze each column individually to understand the distribution and key characteristics.   * + **Status Distribution**: Analyze the status of clinical trials (e.g., Completed, Ongoing).   print(df['Status'].value\_counts())  df['Status'].value\_counts().plot(kind='bar', title='Status of Clinical Trials')   * + **Phase Distribution**: Understand the distribution of trial phases.   print(df['Phases'].value\_counts()) df['Phases'].value\_counts().plot(kind='bar', title='Distribution of Phases') ● **Age Group Analysis**: Analyze the distribution of age groups.  print(df['Age'].value\_counts())  df['Age'].value\_counts().plot(kind='bar', title='Age Group Distribution')   1. **Bivariate Analysis** |

|  |
| --- |
| Explore relationships between different variables.   * **Status vs. Phases**: Explore how trial phases are distributed across different statuses.   status\_phase = pd.crosstab(df['Status'], df['Phases']) print(status\_phase) status\_phase.plot(kind='bar', stacked=True, title='Status vs.  Phases')   * **Conditions vs. Outcome Measures**: Understand the common outcome measures for different conditions.   conditions\_outcomes = df.groupby('Conditions')['Outcome Measures'].apply(lambda x: ', '.join(x)).reset\_index() print(conditions\_outcomes)  **7. Time Series Analysis**  Analyze the trends over time, such as the number of trials started over the months.  # Convert date columns 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') |

|  |
| --- |
| # 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')   1. **Conclusion**   Summarize the findings from your EDA. For example:   * + The majority of trials are in the "Completed" phase.   + Most trials target adult populations.   + There's a steady increase in the number of trials over time.  1. **Saving Results**   You can save the processed data or specific analysis results for further use.  # Save the cleaned data  df.to\_csv('cleaned\_covid\_clinical\_trials.csv', index=False)   1. **Output and Visuals**   After running the code, you should observe:   * + Bar charts showing the distribution of trial statuses, phases, and age groups.   + A time series plot illustrating the trend of trials over time.   This project will provide a solid foundation in EDA using Pandas, with practical insights into the clinical trials landscape for COVID-19. |

**Example: You can get the basic idea how you can create a project from here**

**Sample code with output**

Import

Required

Libraries

[¶](https://www.kaggle.com/code/yahiamostafa/covid-19-clinical-trials-eda-pandas#Import-Required-Libraries)

In

[1]:

import

pandas

as

pd

import

numpy

as

np

import

seaborn

as

sns

import

matplotlib.pyplot

as

plt

Load

The

DataSet

In

[2]:

df

=

pd

.

read\_csv(

'../input/covid19-clinical-trials-dataset/COVID

clinical

trials.csv'

,

index\_col

=

0

)

Exploratory

Data

Analysis

In

[3]:

*#*

*print*

*the*

*first*

*5*

*rows*

*in*

*the*

*dataset*

df

.

head(n

=

5

)

Out[3]:

N

C

T

N

u

m

b

er

Ti

tl

e

A

c

r

o

n

y

m

S

t

a

t

u

s

S

t

u

d

y

R

e

s

u

l

t

s

Co

ndit

ion

s

In

te

rv

e

nt

io

n

s

Out

co

me

Me

asu

res

Spo

nso

r/C

olla

bor

ator

s

G

e

n

d

e

r

.

.

.

O

t

h

e

r

I

D

s

S

t

a

r

t

D

a

t

e

P

ri

m

a

r

y

C

o

m

p

l

e

ti

o

n

D

a

t

e

C

o

m

p

l

e

ti

o

n

D

a

t

e

F

i

r

s

t

P

o

s

t

e

d

R

e

s

u

l

t

s

F

i

r

s

t

P

o

s

t

e

d

L

a

s

t

U

p

d

a

t

e

P

o

s

t

e

d

L

o

c

a

ti

o

n

s

S

t

u

d

y

D

o

c

u

m

e

n

t

s

URL

R

a

n

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  | k |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | N  C  T  0 4 7 8 5 8 9 8 | D ia  g n o  st ic  P  er fo r  m  a n c e  of th e  I  D  N  o w  ™ | C  O  V  I  D  -  I  D  N  o w | A c ti v e  ,  n o  t r  e c  r  u  it i  n g | N o R e s u l t  s  A  v a  i l  a b  l  e | Co vid  19 | D ia  g n o  st ic  T e  st  : I  D  N  o w  ™  C  O  V  I  D  -  1 9 | Ev  alu ate the dia gno  stic per for  ma  nce  of the  ID  ... | Gro upe Hos  pita lier  Pari s Sai nt Jos eph | A  l l | .  .  . | C  O  V  I  D  -  I  D  N  o w | N o v e  m b e  r  9  ,  2 0 2 0 | D  e c e  m  b e  r  2  2  ,  2 0 2 0 | A p ri l  3  0  ,  2 0 2 1 | M  a  r  c h 8  ,  2 0 2 1 | N  a  N | M  a  r  c h 8  ,  2 0 2 1 | G  r   * u p e H * s   pi  t  al ie r  P a ri  s  S ai  n t-  J o s | N  a  N | https://Cli nicalTrial  s.gov/sh  ow/NCT0 4785898 |
|  | | | | | | | | | | | | | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  |  | C  O  V  I  D  -1 9.  .. |  |  |  |  | S  cr e e  ni  n g T e  st |  |  |  |  |  |  |  |  |  |  |  | e p h  ,  P a ri  s,  ..  . |  |  |
| 2 | N  C  T  0 4 5 9 5 1 3 6 | S  tu d y to  E  v  al  u  at e  th e E  ffi  c | C  O  V  I  D  -  1 9 | N o  t  y e  t r  e c  r  u  it i  n g | N  o R  e s u  l t  s  A  v a  i l  a b | SA  RS-  Co  V-2 Infe ctio  n | D  r  u g:  D  r  u g  C  O  V  I  D  1  9  - | Ch ang e on  vira  l  loa d res ults fro  m  bas  elin  e aft.. | Unit ed Me  dic al  Spe cialt ies | A  l l | .  .  . | C  O  V  I  D  1  9  -  0 0 0  1  -  U  S  R | N  o v e  m  b e  r  2  ,  2 0 2 0 | D e c e  m  b e  r  1  5  ,  2 0 2 0 | J a n u a  r  y  2  9  ,  2 0 2 1 | O  c  t  o b e  r  2  0  ,  2 0 2 0 | N  a  N | O  c  t  o b e  r  2  0  ,  2 0 2 0 | C  i  m  e  di c  al  ,  B a  rr  a n q  ui  ll | N  a  N | https://Cli nicalTrial  s.gov/sh  ow/NCT0 4595136 |
|  | | | | | | | | | | | | | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  |  | a c y of  C  O  V  I  D  1  9-  0  0  0  1.  .. |  |  | l  e |  | 0  0 0  1  -  U  S  R  |  D  r  u g: n o  r  m  al s  al in  e | . |  |  |  |  |  |  |  |  |  |  | a  ,  A  tl a n  ti c o  ,  C ol o  m  bi  a |  |  |
| 3 | N  C  T  0 | L u n g | T  A  C  - | R  e c  r | N  o R  e | cov id1  9 | O th e r: | A qua  litat  ive | Uni ver  sity of | A  l l | .  .  . | T  A  C  - | M  a y  7 | J u n e | J u n e | M  a y  2 | N  a  N | N  o v e | O  s p e | N  a  N | https://Cli nicalTrial  s.gov/sh  ow/NCT0 |
|  | | | | | | | | | | | | | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  | 4  3 9 5  4  8 2 | C  T S  c a n A n  al y  si  s of  S A  R  S  -  C  o  V  2  In d u c e d L | C  O  V  I  D  1 9 | u  it i  n g | s u  l t  s  A  v a i l  a b l  e |  | L u n g  C T s c a n a n  al y  si  s  in  C  O  V  I  D  -  1 9 p  at ie nt | ana  lysi  s of par enc hy mal lun g da  m... | Mil  ano  Bic occ a |  |  | C  O  V  I  D  1 9 | ,  2 0 2 0 | 1  5  ,  2  0  2 1 | 1  5  ,  2  0  2 1 | 0  ,  2 0  2 0 |  | m  b e  r  9  ,  2 0  2 0 | d  al  e P a p a  G io v a n  ni  X  X  II I,  B e  r  g a  m  o  ,  It  al y|  P |  | 4395482 |
|  | | | | | | | | | | | | | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  |  | u  n.  .. |  |  |  |  | s |  |  |  |  |  |  |  |  |  |  |  | ..  . |  |  |
| 4 | N  C  T  0 4 4 1 6 0 6 1 | T h e R ol e  of a P  ri  v at e H   * s   pi ta  l  in  H   * n | C  O  V  I  D  -  1 9 | A c ti v e  ,  n o  t r  e c  r  u  it i  n g | N o R e s u l t  s  A  v a  i l  a b  l  e | CO  VID | D ia  g n o  st ic  T e  st  :  C  O  V  I  D  1 9 D ia  g n o | Pro por tion of asy mpt om  atic sub  ject s|P rop  orti on..  . | Ho ng Kon g San  ator ium  & Hos  pita  l | A  l l | .  .  . | R  C  -  2 0 2  0  -  0 8 | M a y  2  5  ,  2 0 2 0 | J u  l  y  3  1  ,  2 0 2 0 | A u g u s  t  3  1  ,  2 0 2 0 | J u n e 4  ,  2 0 2 0 | N  a  N | J u n e 4  ,  2 0 2 0 | H   * n g K o n g S a n a   t   * ri u   m  & H   * s   pi  t | N  a  N | https://Cli nicalTrial  s.gov/sh  ow/NCT0 4416061 |
|  | | | | | | | | | | | | | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  |  | g K o n g A m  ... |  |  |  |  | st ic  T e  st |  |  |  |  |  |  |  |  |  |  |  | al  ,  H  o n g K o n g  ,  H o  ..  . |  |  |
| 5 | N  C  T  0 4 3 9 5 9 2 4 | M at er n  al  -f  o  et al Tr a | T  M  F  -  C  O  V  I  D  -  1 | R  e c  r  u  it i  n g | N  o R  e s u  l t  s  A  v | Mat ern al Fet al Infe ctio  n Tra ns mis | D ia  g n o  st ic  T e  st  : | CO  VID -19 by pos  itiv  e  PC R in cor | Ce  ntre  Hos  pita lier  Ré gio nal d'O  rléa ns| | F e  m  a  l  e | .  .  . | C H R  O  -  2 0 2  0  -  1 | M  a y  5  ,  2 0 2 0 | M  a y  2 0 2 1 | M  a y  2 0 2 1 | M  a y  2  0  ,  2 0 2 0 | N  a  N | J u n e 4  ,  2 0 2 0 | C H R O  rl é a n  s,  O  rl | N  a  N | https://Cli nicalTrial  s.gov/sh  ow/NCT0 4395924 |
|  | | | | | | | | | | | | | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  |  | n s  m  is si o n  of  S  A  R  S  -  C  o  v-  2 | 9 |  | a i l  a b l  e | sio n|C  OVI  D-1  9... | D ia  g n o  si  s of  S  A  R  S  -  C  o v 2 b y  R  T  -.  .. | d  blo od and  /  o... | Ce  ntre  d... |  |  | 0 |  |  |  |  |  |  | é a n  s,  F r  a n c e |  |  |
| 5 rows × 26 columns  In [4]:  *# Shape of the DataSet* | | | | | | | |  |  |  | | |  | | | | | | | | |

|  |
| --- |
| df.shape  Out[4]:  (5783, 26)  In [5]:  *# Columns in the dataset* df.columns  Out[5]:  Index(['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')

In

[6]:

*#*

*Categorical*

*Features*

df

.

select\_dtypes(include

=

'object'

)

.

columns

Out[6]:

Index(['NCT

Number',

'Title',

'Acronym',

'Status',

'Study

Results',

'Conditions',

'Interventions',

'Outcome

Measures',

'Sponsor/Collaborators',

'Gender',

'Age',

'Phases',

'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')

In

[7]:

*#*

*Neumrical*

*Features*

df

.

select\_dtypes(exclude

=

'object'

)

.

columns

|  |
| --- |
| Out[7]:  Index(['Enrollment'], dtype='object')  In [8]:  *# Detecting (Percentage) Missing Data* missing\_data = df.isnull().mean() \* 100 missing\_data  Out[8]:  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 |

dtype: float64

In [9]:

|  |  |
| --- | --- |
| 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 |

*# Visualize data without calculating* def visualize\_data(data , caption = '' , ylabel = 'Percentage of Mising Data'):

*# set figure size* sns.set(rc={'figure.figsize':(15,8.27)})

*# make ticks vertical* 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())

\

.

set\_title(caption)

*#*

*set*

*labels*

plt

.

ylabel(ylabel)

plt

.

show()

In

[10]:

visualize\_data(missing\_data

,

'Percentage

of

missing

data

in

each

feature'

)

As

shown

the

percentae

of

missing

data

in

**Results**

**First**

**Posted**

is

**99.3**

**%**

and

**Study**

**Documents**

is

**96.8**

**%**

,

so

it's

impossible

to

impute

them

without

destoying

our

dataset.

In

[11]:

*#*

*Drop*

*Study*

*Documents*

*and*

*Results*

*First*

*Posted*

df

.

drop([

'Results

First

Posted'

,

'Study

Documents'

]

,

inplace

=

True

,

axis

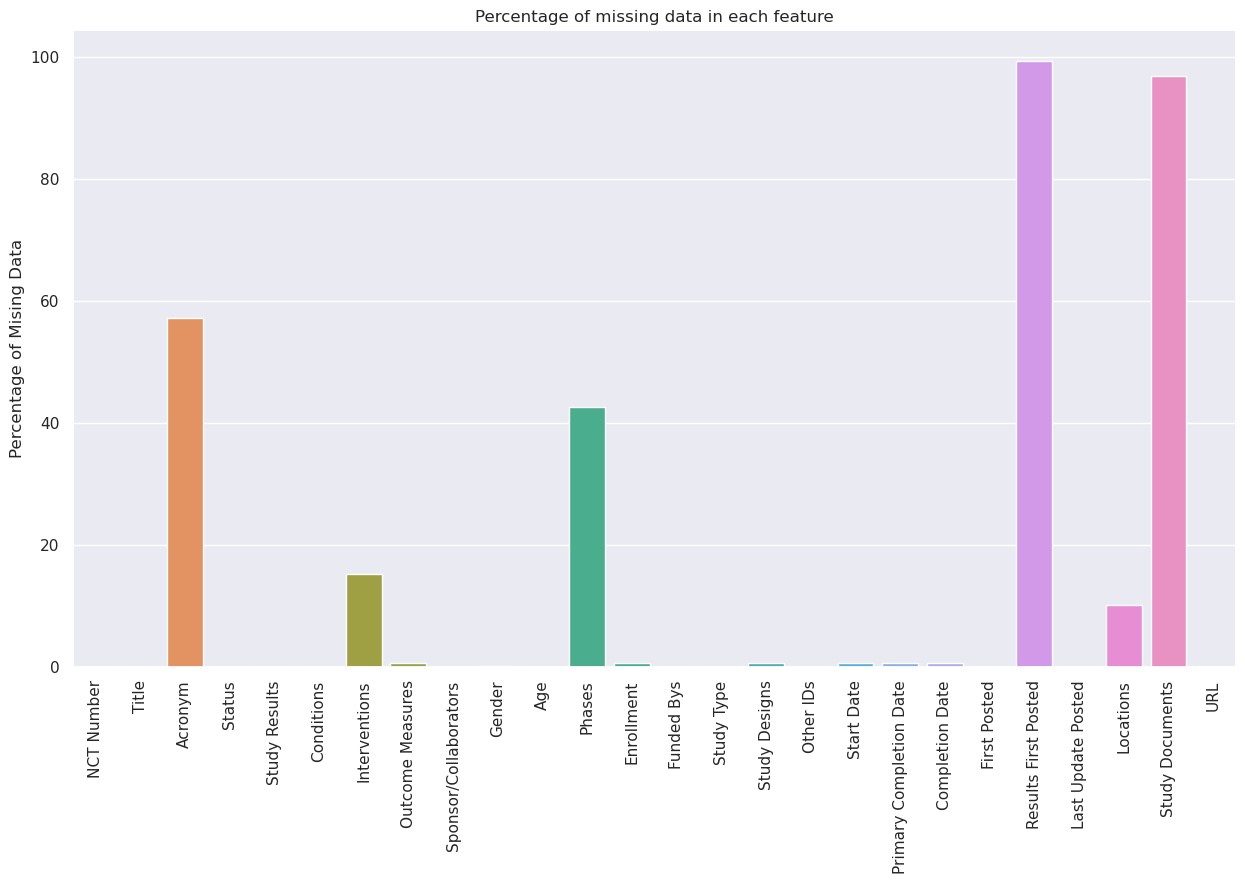
=

1

)

In

[12]:



|  |
| --- |
| *# Columns in the dataset after dropping Study Documents and Results First Posted* df.columns  Out[12]:  Index(['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'], dtype='object')  In [13]:  *# Drop Duplicate Rows* print(f"Shape before dropping duplicates data **{**df.shape**}**") df.drop\_duplicates(inplace = True) print(f"Shape after dropping duplicates data **{**df.shape**}**") |

|  |
| --- |
| Shape before dropping duplicates data (5783, 24)  Shape after dropping duplicates data (5783, 24)  There is no duplicate rows in the dataset.  In [14]:  *# Drop rows that have less than 10 non-null values* print(f"Shape before dropping Null rows **{**df.shape**}**") df.dropna(how = 'any' , axis = 0 , thresh = 10 , inplace =  True)  print(f"Shape after dropping Null rows **{**df.shape**}**")  Shape before dropping Null rows (5783, 24)  Shape after dropping Null rows (5783, 24)  There is no rows with less than 10 non-null values  In [15]:  df.isnull().mean() \* 100  Out[15]:  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 dtype: float64  In [16]: | 0.000000 |

*#*

*We*

*can*

*extract*

*a*

*new*

*feature*

*form*

*The*

*Location*

*which*

*is*

*the*

*country*

*where*

*the*

*study*

*hold*

countries

=

[

str

(

df

.

Locations

.

iloc[i])

.

split(

','

)[

-

1

]

for

i

**in**

range

(

df

.

shape[

0

])]

df[

'Country'

]

=

countries

In

[17]:

df

.

columns

Out[17]:

Index(['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')

In

[18]:

df

.

Country

.

value\_counts()[:

35

]

Out[18]:

United

States

1267

France

647

nan

585

United

Kingdom

306

Italy

235

Spain

234

Turkey

219

Canada

202

Egypt

192

China

171

Brazil

137

Germany

128

Belgium

91

Mexico

88

Switzerland

76

Russian

Federation

69

Sweden

57

Denmark

56

Israel

56

India

55

Pakistan

53

Name: Country, dtype: int64

**Now We need to clasify the missing data to one of these categories**

1. Missing Completely At Random (MCAR)

|  |  |
| --- | --- |
| Argentina | 47 |
| Netherlands | 46 |
| Norway | 38 |
| Hong Kong | 36 |
| Colombia | 33 |
| Republic of | 31 |
| Austria | 29 |
| Poland | 29 |
| Singapore | 29 |
| Saudi Arabia | 27 |
| Australia | 26 |
| Greece | 26 |
| Islamic Republic of | 23 |
| South Africa | 22 |

1. Missing At Random (MAR)
2. Not Missing At Random (NMAR)

In [19]:

*# Lets's start with Acronym*

|  |
| --- |
| print(f"Number of unique values is **{**df.Acronym.nunique()**} \n**") df.Acronym.value\_counts()  Number of unique values is 2338  Out[19]:  COVID-19 47  PROTECT 7  CORONA 6  RECOVER 5  SCOPE 5  ..  ASD 1  VICO 1  LICORNE 1  LOSVID 1  MindMyMindFU 1  Name: Acronym, Length: 2338, dtype: int64  In [20]:  *# Find the realtion between null values in Acronym and Countries* |

|  |
| --- |
| (df.Acronym.isnull().groupby(df.Country).mean().sort\_values(asc ending = False) \* 100)[:60]  Out[20]:  Country  Iraq 100.000000  Belarus 100.000000  Rwanda 100.000000  South Sudan 100.000000  Cambodia 100.000000  Bulgaria 100.000000  Cyprus 100.000000  Bosnia and Herzegovina 100.000000  Guinea-Bissau 100.000000  Dominican Republic 100.000000  Ecuador 100.000000  North Macedonia 100.000000  Bahrain 100.000000  Azerbaijan 100.000000  Uruguay 100.000000  Uzbekistan 100.000000  Kyrgyzstan 100.000000  Cape Verde 100.000000  Republic of 96.774194 |

|  |  |
| --- | --- |
| Taiwan | 93.750000 |
| Singapore | 93.103448 |
| Japan | 88.888889 |
| Kuwait | 87.500000 |
| China | 87.134503 |
| Turkey | 86.757991 |
| Ukraine | 85.714286 |
| Malaysia | 84.615385 |
| Egypt | 83.854167 |
| Hungary | 83.333333 |
| Hong Kong | 80.555556 |
| Bangladesh | 80.000000 |
| India | 80.000000 |
| Kazakhstan | 80.000000 |
| Saudi Arabia | 77.777778 |
| Puerto Rico | 76.470588 |
| Israel | 75.000000 |
| Zimbabwe | 75.000000 |
| Jordan | 72.727273 |
| Poland | 72.413793 |
| Indonesia | 71.428571 |
| United States | 69.376480 |
| Romania | 69.230769 |
| Kenya | 66.666667 |
| Nepal | 66.666667 |

|  |
| --- |
| New Zealand 66.666667  Ethiopia 66.666667  Slovakia 66.666667  Thailand 66.666667  Lebanon 66.666667 nan 66.324786  Islamic Republic of 65.217391  Russian Federation 65.217391  Chile 64.705882  Austria 62.068966  Pakistan 60.377358  Brazil 59.124088  Mexico 57.954545  Sweden 57.894737  Argentina 57.446809  Canada 55.940594  Name: Acronym, dtype: float64   * After inspecting the relation between the missing values in Acronym and Country we can conclude that there is a sort of relation between these two features, so we can say that Data is Missing At Random (MAR). * So we can Impute by Missing Category.   In [21]:  *# impute by a missing Indicator* df.Acronym = df.Acronym.fillna("Missing Acronym") |

|  |
| --- |
| In [22]:  *# Detecting (Percentage) Missing Data* df.isnull().mean() \* 100  Out[22]:  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  We can do the same for other categorical features such as Interventions , Phases , Locations and other categorical features  In [23]:  *# Impute Interventions , Phases , Locations by Missing Category*  categorical\_features = df.select\_dtypes(include = object).columns  features = categorical\_features[df[categorical\_features].isnull().mean() >  0]  for feature **in** features:  df[feature] = df[feature].fillna(f"Missing **{**feature**}**") |

|  |
| --- |
| In [24]:  *# Detecting (Percentage) Missing Data* df.isnull().mean() \* 100  Out[24]:  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  Now the Time to handle The missing data for the Enrollment  In [25]:  *# Check the skewness* df.Enrollment.skew()  Out[25]:  34.06593382031148  The value of Skewness is 34 which means that we This feature isn't normally distributed  In [26]:  *# Plotting the distribution of the enrollment* df.Enrollment.plot(kind = 'kde') |

Out[26]:

<

AxesSubplot:ylabel='Density'

>

So

We

will

impute

by

the

median

In

[27]:

*#*

*Some*

*Statstical*

*Valuse*

*for*

*the*

*Enrollment*

*Column*

min\_Value

=

df

.

Enrollment

.

min()

max\_Value

=

df

.

Enrollment

.

max()

mean\_Value

=

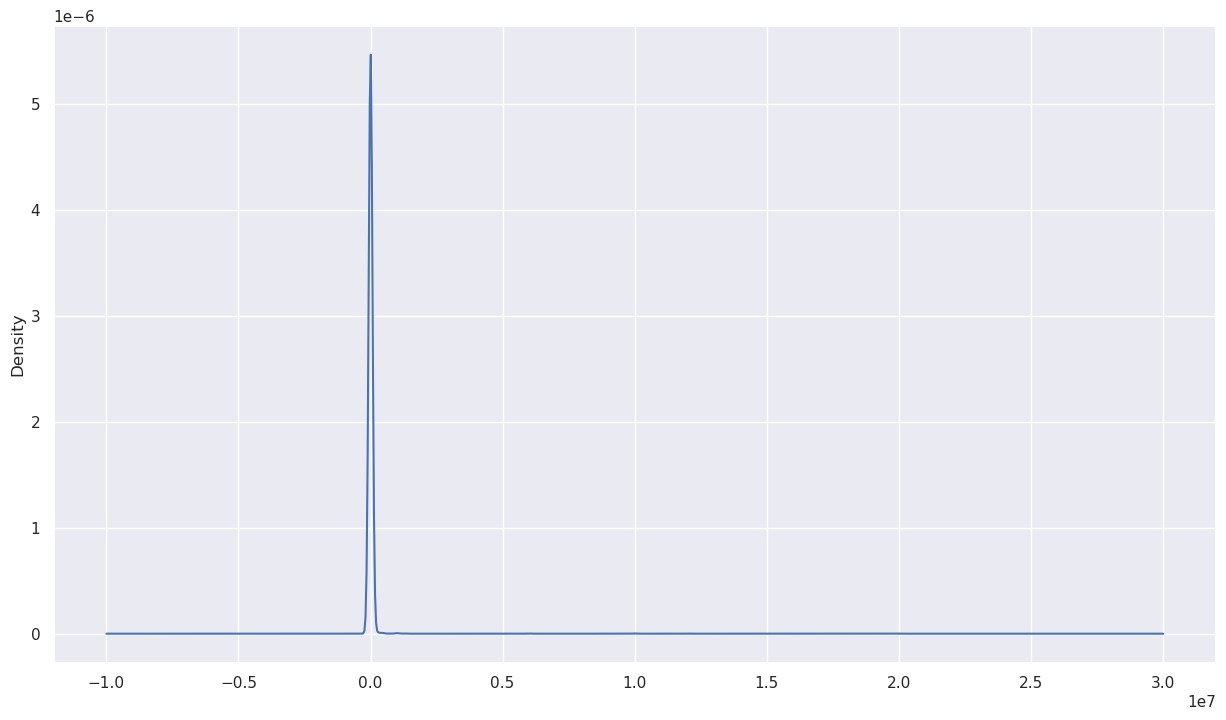
df

.

Enrollment

.

mean()



|  |
| --- |
| median\_Value = df.Enrollment.median() std\_Value = df.Enrollment.std()  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 404543.7287841079  In [28]:  *# Using Median to impute Missing Values*  df.Enrollment = df.Enrollment.fillna(median\_Value)  In [29]:  *# Detecting (Percentage) Missing Data* df.isnull().mean() \* 100 |

|  |  |
| --- | --- |
| Out[29]:  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 dtype: float64  In [30]: df.head() Out[30]: | | | | | | | |  |  | 0.0 | | | |  | | | | | | |  |  |
|  |  | N  C  T N  u  m  b  er | Ti  tl e | A  c  r  o n y  m | S  t  a  t  u s | S  t  u d y  R  e s u  l t  s | Co  ndit ion s | I  n  t  e  r  v e n  ti o n s | Ou tco me Me  as ure s | Sp ons  or/ Col lab ora tors | G  e n d e  r | .  .  . | Stu dy Des igns | O  t  h e  r  I  D  s | S  t  a  r t  D  a  t  e | P  ri  m  a  r  y  C  o  m  p  l  e  ti o n  D | C  o  m  p  l  e  ti o n D  a  t  e | F  i  r  s  t  P o s  t  e d | L a s  t  U  p d a  t  e P o s  t  e | L o c a  ti o n s | URL | o u n  t r  y |
|  | | | | | | | | | | | | | | | | | | | | | |

C

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | a  t  e |  |  | d |  |  |  |
| R  a n k |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | N  C  T  0 4 7 8 5 8 9 8 | D ia  g n o  st ic  P e rf  o  r  m  a n c e | C  O  V  I  D  -  I  D  N  o w | A  c  t i  v e  ,  n o  t r  e c  r  u  i t | N  o R  e s u  l t  s  A  v a  i l  a b  l | Co vid  19 | D ia  g n o  st ic  T e  st  : I  D  N  o w  ™ | Ev  alu ate the dia gn  osti c per for  ma  nc e  of the  ID  ... | Gro upe Ho  spit alie r  Par  is  Sai nt Jos eph | A  l l | .  .  . | Allo cati on:  N/A| Inte rven tion  Mod el:  Sin gle Gro.  .. | C  O  V  I  D  -  I  D  N  o w | N  o v e  m  b e  r  9  ,  2 0 2 0 | D  e c e  m  b e  r  2  2  ,  2 0 2 0 | A p ri l  3  0  ,  2 0 2 1 | M  a  r  c h 8  ,  2 0 2 1 | M  a  r  c h 8  ,  2 0 2 1 | G  r   * u p e H * s p   it a  li e  r  P a | https://Cl inicalTria ls.gov/sh  ow/NCT  0478589  8 | F  r  a n c e |
|  | | | | | | | | | | | | | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  |  | o f t  h e  I  D  N  o w  ™  C  O  V  I  D  -  1  9  ..  . |  | i  n g | e |  | C  O  V  I  D  -  1  9 S  c  r  e e  ni n g T e  st |  |  |  |  |  |  |  |  |  |  |  | ri s  S a i  n t-  J o s e p h  ,  P a ri  s  ,  ..  . |  |  |
| 2 | N  C  T  0 4 | S  t  u d y | C  O  V  I  D | N  o  t  y e | N  o R  e s | SA  RS  -Co  V-2  Inf | D  r  u g  : | Ch an ge on  vir | Uni ted  Me  dic al | A  l l | .  .  . | Allo cati on:  Ran dom | C  O  V  I  D | N  o v e  m | D  e c e  m | J a n u a | O  c  t  o b | O  c  t  o b | C  i  m  e d | https://Cl inicalTria ls.gov/sh  ow/NCT  0459513 | o  l  o |
|  | | | | | | | | | | | | | | | | | | | | | |

C

m

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  | 5  9 5  1  3 6 | t  o E  v  al  u a  t  e  t  h e E  ffi c a c y o  f  C  O  V  I  D  1  9  -  0 | -  1 9 | t r  e c  r  u  i t i  n g | u  l t  s  A  v a  i l  a b  l  e | ecti  on | D  r  u g  C  O  V  I  D  1  9  -  0  0  0  1  -  U  S  R  |  D  r  u g  :  n o  r | al loa d res ults fro  m  ba  seli  ne aft.  .. | Sp eci alti  es |  |  | ized |Inte rven tion  Mod el:  Par.  .. | 1  9  -  0  0  0  1  -  U  S  R | b e  r  2  ,  2 0  2 0 | b e  r  1  5  ,  2  0 2 0 | r  y  2  9  ,  2 0  2 1 | e  r  2  0  ,  2 0  2 0 | e  r  2  0  ,  2 0  2 0 | ic  a  l,  B a  rr  a n q u  ill  a  ,  A tl a n  ti c o  ,  C  o  l  o  m  b  i | 6 | b  i  a |
|  | | | | | | | | | | | | | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  |  | 0  0  1  ..  . |  |  |  |  | m al s  al in  e |  |  |  |  |  |  |  |  |  |  |  | a |  |  |
| 3 | N  C  T  0 4 3 9 5 4 8 2 | L u n g  C  T S  c a n A n  al y  si  s o  f  S A | T  A  C  -  C  O  V  I  D  1 9 | R  e c  r  u  i t i  n g | N  o R  e s u  l t  s  A  v a  i l  a b  l  e | cov id1  9 | O t  h e r:  L u n g  C T s c a n a n  al y  si | A qu  alit ativ e an  aly sis of par en chy ma  l  lun g da  m..  . | Uni ver  sity of Mil  ano  Bic occ a | A  l l | .  .  . | Obs erva tion al  Mod el:  Coh  ort|  Tim e Per spe  ctiv e: ... | T  A  C  -  C  O  V  I  D  1 9 | M a y  7  ,  2 0 2 0 | J u n e 1  5  ,  2 0 2 1 | J u n e 1  5  ,  2 0 2 1 | M  a y  2  0  ,  2 0 2 0 | N  o v e  m  b e  r  9  ,  2 0 2 0 | O s p e d a  l  e P a p a  G  i  o v a n n  i | https://Cl inicalTria ls.gov/sh  ow/NCT  0439548  2 | a n  a  r  i  n o |
|  | | | | | | | | | | | | | | | | | | | | | |

S

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  |  | R  S  -  C o  V  2  I  n d u c e d L u n  ..  . |  |  |  |  | s  in  C  O  V  I  D  -  1 9 p a  ti e n  ts |  |  |  |  |  |  |  |  |  |  |  | X  X  II  I,  B e  r  g a  m  o  , It a  ly  |  P  ..  . |  |  |
| 4 | N  C  T  0 4 4 1 | T h e R ol  e o | C  O  V  I  D  -  1 | A  c  t i  v e  , | N  o R  e s u  l | CO  VI  D | D ia  g n o  st ic | Pro por  tio  n  of asy mp | Ho ng Ko ng Sa  nat oriu | A  l l | .  .  . | Obs erva tion al  Mod el:  Coh | R  C  -  2 0 2 0 | M  a y  2  5  ,  2 | J u  l  y  3  1  , | A u g u s  t  3 | J u n e 4  ,  2 | J u n e 4  ,  2 | H  o n g K o n | https://Cl inicalTria ls.gov/sh  ow/NCT  0441606  1 | * n g * n |
|  | | | | | | | | | | | | | | | | | | | | | |

M H

K

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  | 6  0 6 1 | f  a P  ri  v a  t  e H   * s   pi  t  al in  H   * n g K o n g A m   ..  . | 9 | n o  t r  e c  r  u  i t i  n g | t  s  A  v a  i l  a b  l  e |  | T e  st  :  C  O  V  I  D  1 9 D ia g n o  st ic  T e  st | to  ma  tic su bje cts|  Pro por  tio  n... | m  &  Ho  spit al |  |  | ort|  Tim e Per spe  ctiv e: ... | -  0 8 | 0  2 0 | 2  0 2 0 | 1  ,  2  0  2 0 | 0  2 0 | 0  2 0 | g S a n a  t   * ri u   m  & H   * s p   it a  l,  H   * n g K o n g   ,  H  o |  | g |
|  | | | | | | | | | | | | | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | ..  . |  |  |
| 5 | N  C  T  0  4 3 9 5 9 2 4 | M a t  e  r  n  al  -f  o e  t  al  T  r  a n s  m  is si  o n o  f | T  M  F  -  C  O  V  I  D  -  1 9 | R  e c  r  u  i t i  n g | N  o R  e s u  l t  s  A  v a  i l  a b  l  e | Ma ter nal Fet al Inf ecti on Tra ns mis sio n|C  OV  ID19.  .. | D ia g n o  st ic  T e  st  :  D ia  g n o  si  s o  f  S A  R  S | CO  VI  D19 by po  siti  ve  PC R in cor d  blo od an d /  o... | Ce ntr e Ho  spit alie r  Ré gio nal d'O  rléa ns|  Ce  ntr  e  d... | F e  m  a  l  e | .  .  . | Obs erva tion al  Mod el:  Coh  ort|  Tim e Per spe  ctiv e: ... | C H R  O  -  2 0 2  0  -  1 0 | M a y  5  ,  2 0 2 0 | M a y  2 0 2 1 | M a y  2 0 2 1 | M  a y  2  0  ,  2 0 2 0 | J u n e 4  ,  2 0 2 0 | C  H  R O rl é a n s  ,  O  rl é a n s  ,  F  r  a n c e | https://Cl inicalTria ls.gov/sh  ow/NCT  0439592  4 | F  r  a n c e |
|  | | | | | | | | | | | | | | | | | | | | | |

S

A

R

S

-

C

o

v

-

2

-

C

o

v

2

b

y

R

T

-.

..

5

rows

×

25

columns

Data

Visualizations

In

[31]:

*#*

*Get*

*Countires*

*with*

*highest*

*Contributiuons*

top\_10\_Countires

=

df

.

Country

.

value\_counts()[:

10

]

visualize\_data(top\_10\_Countires

,

caption

=

'Top

10

Countries'

,

ylabel

=

'Contributions'

)

In

[32]:

*#*

*Status*

*of*

*the*

*Application*

status

=

df

.

Status

.

value\_counts()

visualize\_data(status

,

caption

=

'Status

of

The

Application'

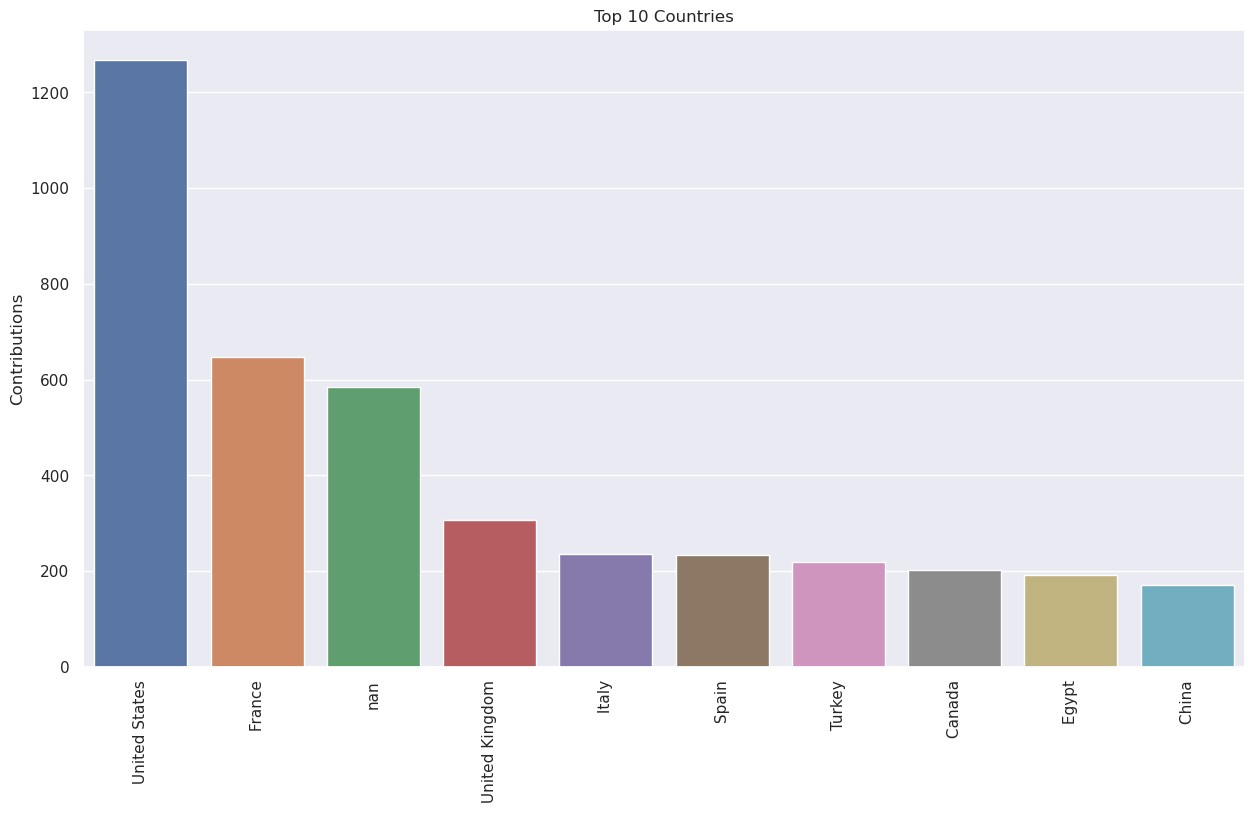
,

ylabel

=

'Denisty'

)



In

[33]:

*#*

*Gender*

*Visualiztions*

gender

=

df

.

Gender

.

value\_counts()

visualize\_data(gender

,

caption

=

'Gender

Distribution'

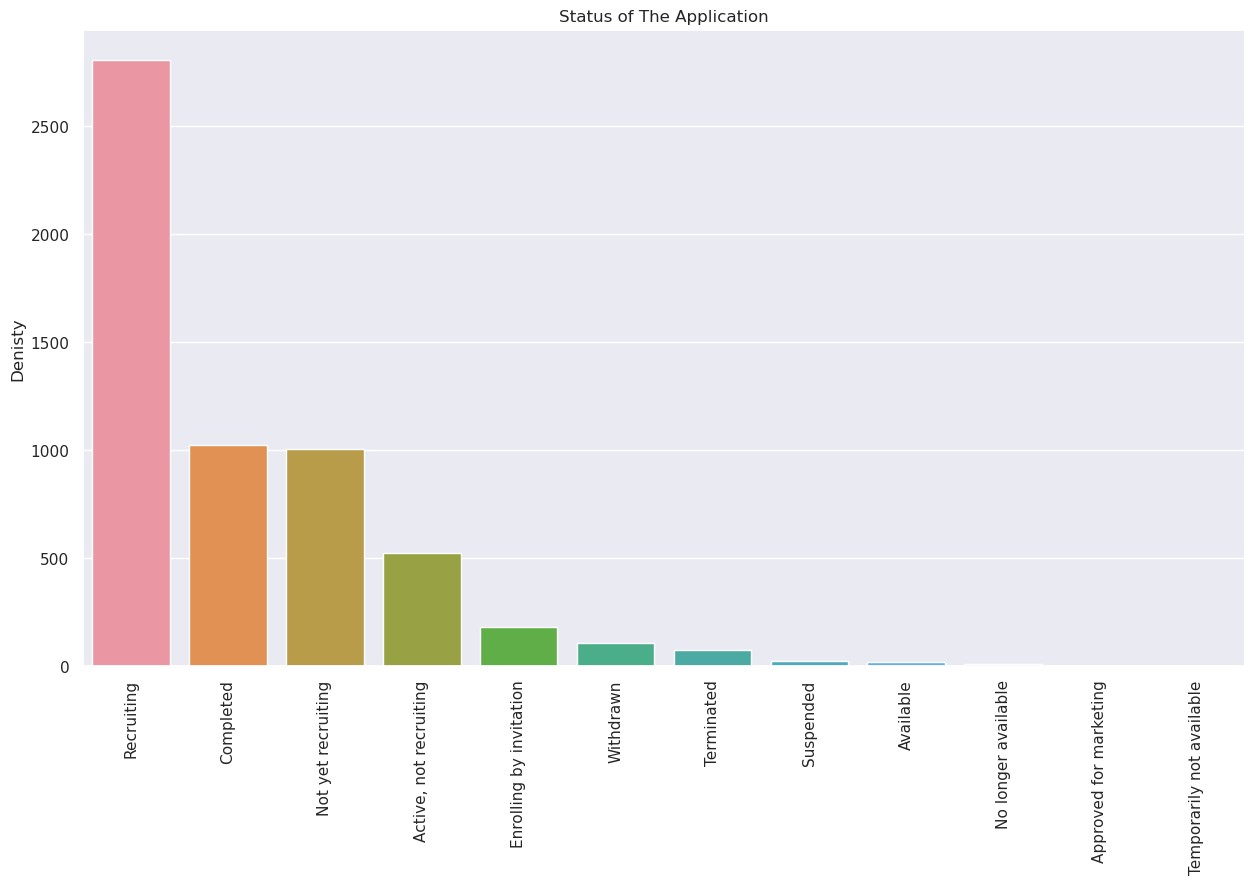
,

ylabel

=

'Denisty'

)



In

[34]:

*#*

*Which*

*month*

*has*

*the*

*highest*

*start*

start\_month

=

pd

.

Series([

str

(

df

[

'Start

Date'

]

.

iloc[i])

.

split(

'

'

)[

0

]

for

i

**in**

range

(

df

.

shape[

0

])])

start\_month\_Distribution

=

start\_month

.

value\_counts()

visualize\_data(start\_month\_Distribution

,

caption

=

'Start

Month

Distribution'

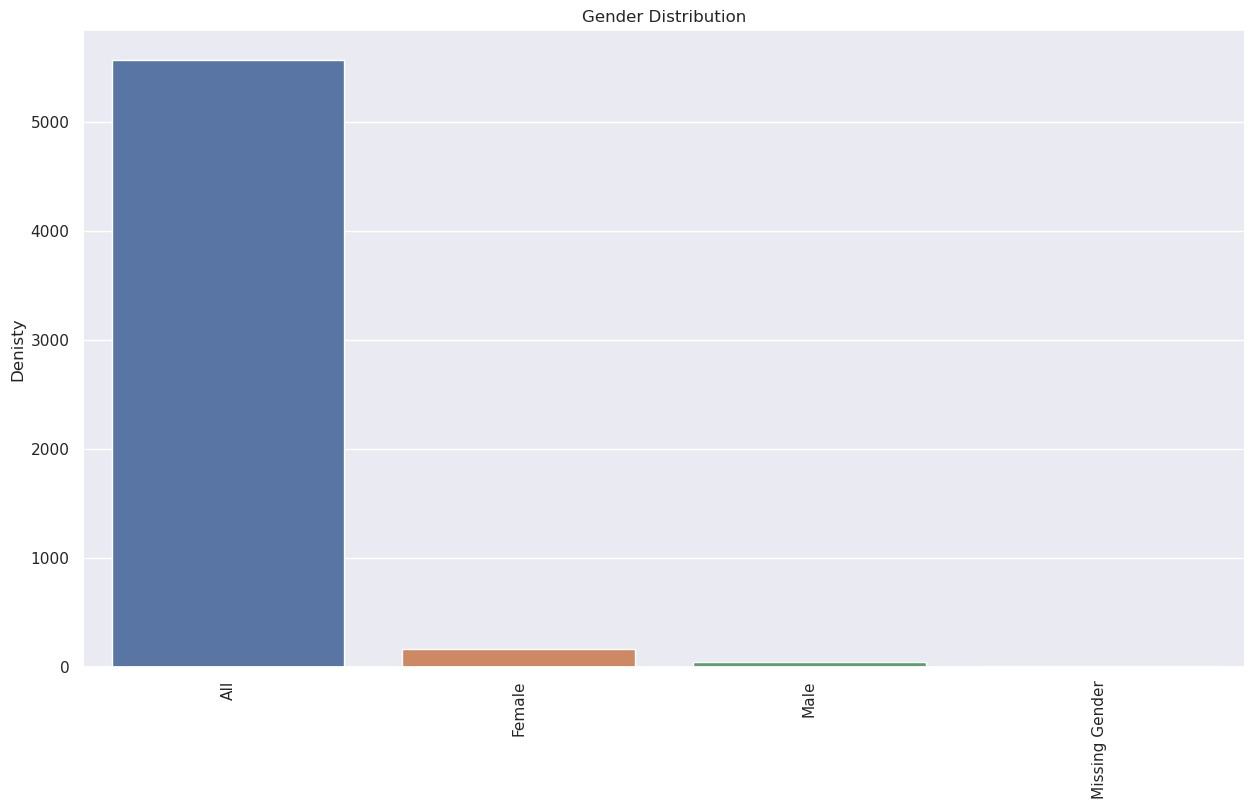
,

ylabel

=

'Denisty'

)



In

[35]:

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

(5783

,

25)

Nunique

in

NCT

Number

is

5783

Nunique

in

URL

is

5783

So

If

We

are

going

to

apply

a

(

Machine

Learning)

ML

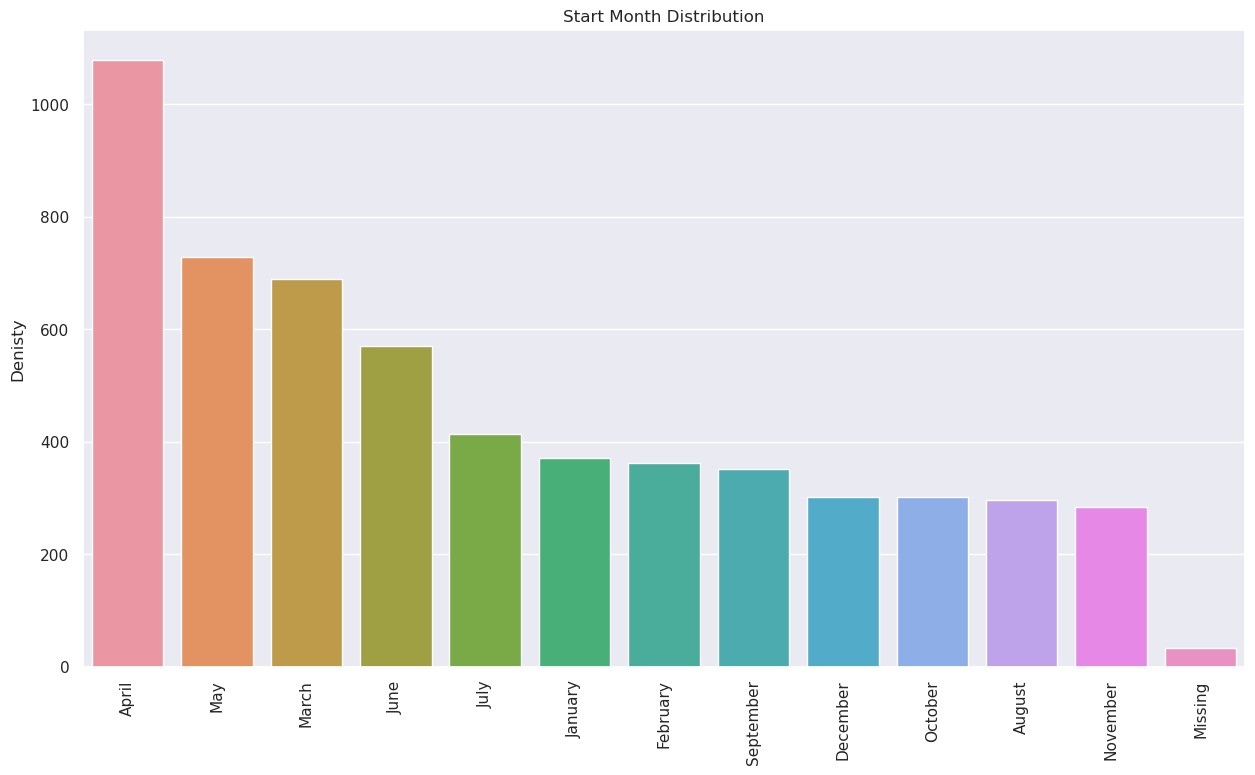
model

we

can

drop

NCT



Number and URL because there is an index already which is Rank. To reduce the number of categorical Features, Specially because they will need to be doecoded inorder to be used in a ML Model.

[Reference](https://github.com/0xpranjal/COVID-19-complete-EDA-analysis) [link](https://github.com/0xpranjal/COVID-19-complete-EDA-analysis)