```
### CODE:1 Clinical Trials EDA Pandas ###

# Use libraries to ingest the dataset in pair complex clinical trials
# Analyze participation trends, study phases, geographic distribution, and
# demographics
```

```
### 2. Operating Libraries and Loading Data ###

import pandas as pd

df = pd.read_csv('ZYTOVIG_Clinical_Trials.csv')
df.head()
```

	Rank	NCT Number	Title	Acronym	Status	Study Results	Conditions	Interventions	Outcome Measures	Sponsor/Collaborators	...	Other IDs	Start Date	Comp. Date	Primary Completion Date	Enrollment	First Posted	Results First Posted	Last Update Posted
0	1	NCT04567843	Diagnostic Performance of Ag-Ab COVID-19 Test	COVID-Effman	Active, not recruiting	No Results Available	Covid19	Diagnostic Test: ID Ag-Ab COVID-19 Screening Test	Evaluate the diagnostic performance of the ID ...	Grupo Hospitalar Dom Aguirr Joaquim	...	COVID-Effman	November 9, 2020	Dec 31, 2021	NaN	1000	2020-09-29	NaN	2021-09-29
1	2	NCT04580746	Study to Evaluate the Safety of COVID-19 Vaccine	COVID-19	Not yet recruiting	No Results Available	SARS-CoV-2 Infection	Drug: Drug COVE19 Dose 1 Drug: Drug COVE19 Dose 2	Change in oral total results from baseline compared with ...	United Medical Specialties	...	COVE19-0001-USR	November 2, 2020	Dec 31, 2021	Dec 31, 2021	30	2020-10-07	NaN	2020-10-07
2	3	NCT04583995	Long CT Scan Analysis in COVID-19 Infected Lun...	TAC-COVID19	Recruiting	No Results Available	covid19	Other: Long CT scan analysis in COVID-19 patients	A qualitative analysis of parenchymal lung dam...	University of Milano Bicocca	...	TAC-COVID19	May 17, 2020	Jun 30, 2020	NaN	100	2020-10-08	NaN	2020-10-08
3	4	NCT04587583	The Role of Chest Imaging in Hospital in Hong Kong Pu...	COVID-19	Active, not recruiting	No Results Available	COVID	Diagnostic Test: COVID-19 Diagnostic Test	Proportion of asymptomatic subjects(Proportion...	Hong Kong Sanatorium & Hospital	...	NC-2020-08	May 29, 2020	NaN	NaN	2000	2020-10-13	NaN	2020-10-13
4	5	NCT04590388	Molecular Epidemiology of SARS-Cov-2	TMP-COVID	Not yet recruiting	No Results Available	Maternal Fetal Medicine	Diagnostic Test: Diagnosis of SARS-Cov2 by RT-...	COVID-19 by polymerase PCR in cord blood and / o...	Centre Hospitalier Régional d'Orléans(Centre d...	...	CHRO-2020-10	May 5, 2020	May 31, 2020	NaN	100	2020-10-14	NaN	2020-10-14

```
### CODE:2 Initial Data Exploration ###

df.head()
```

	Rank	NCT Number	Title	Acronym	Status	Study Results	Conditions	Interventions	Outcome Measures	Sponsor/Collaborators	...	Other IDs	Start Date	Comp. Date	Primary Completion Date	Enrollment	First Posted	Results First Posted	Last Update Posted
0	1	NCT04567843	Diagnostic Performance of Ag-Ab COVID-19 Test	COVID-Effman	Active, not recruiting	No Results Available	Covid19	Diagnostic Test: ID Ag-Ab COVID-19 Screening Test	Evaluate the diagnostic performance of the ID ...	Grupo Hospitalar Dom Aguirr Joaquim	...	COVID-Effman	November 9, 2020	April 30, 2021	March 31, 2021	1000	2020-09-29	NaN	March 4, 2021
1	2	NCT04580746	Study to Evaluate the Safety of COVID-19 Vaccine	COVID-19	Not yet recruiting	No Results Available	SARS-CoV-2 Infection	Drug: Drug COVE19 Dose 1 Drug: Drug COVE19 Dose 2	Change in oral total results from baseline compared with ...	United Medical Specialties	...	COVE19-0001-USR	November 2, 2020	December 31, 2020	December 31, 2021	30	2020-10-07	NaN	October 8, 2020
2	3	NCT04583995	Long CT Scan Analysis in COVID-19 Infected Lun...	TAC-COVID19	Recruiting	No Results Available	covid19	Other: Long CT scan analysis in COVID-19 patients	A qualitative analysis of parenchymal lung dam...	University of Milano Bicocca	...	TAC-COVID19	May 17, 2020	June 30, 2020	June 15, 2020	100	2020-10-08	NaN	November 9, 2020
3	4	NCT04587583	The Role of Chest Imaging in Hospital in Hong Kong Pu...	COVID-19	Active, not recruiting	No Results Available	COVID	Diagnostic Test: COVID-19 Diagnostic Test	Proportion of asymptomatic subjects(Proportion...	Hong Kong Sanatorium & Hospital	...	NC-2020-08	May 29, 2020	NaN	NaN	2000	2020-10-13	NaN	2020-10-13
4	5	NCT04590388	Molecular Epidemiology of SARS-Cov-2	TMP-COVID	Not yet recruiting	No Results Available	Maternal Fetal Medicine Transmission|COVID	Diagnostic Test: Diagnosis of SARS-Cov2 by RT-...	COVID-19 by polymerase PCR in cord blood and / o...	Centre Hospitalier Régional d'Orléans(Centre d...	...	CHRO-2020-10	May 5, 2020	May 31, 2020	NaN	100	2020-10-14	NaN	2020-10-14

```
df.describe()
```

	Rank	Enrollment
count	5793.000000	5.160000e+03
mean	2892.000000	7.431984e+04
std	1674.015245	4.304547e+05
min	1.000000	4.000000e+00
25%	946.500000	6.000000e+01
50%	2892.000000	1.500000e+02
75%	4837.500000	5.600000e+02
max	5793.000000	2.000000e+07

```
# General statistics for categorical columns
print(df.describe(include=['object']))
```

```
       NCT Number                                              Title  \
count        5793                                               5793
unique       5793  Molecular Study Assessing Vapor Nerve Stimulation in COP...
freq            1                                                  1

       Acronym  Status  Study Results Conditions  \
count     5793    5793           5793       5793
unique    1157       7              2        104
freq       147    2600           4500       1000

                                                 Interventions  Outcome Measures  \
count                                                     5793              5793
unique                                                    4607              1348
freq                                                       100                 1
top     Other: No intervention|Other: No intervention...  Mortality

       Sponsor/Collaborators Gender  ... Other IDs  \
count                   5793   5793  ...     5793
unique                  4557      3  ...     5793
freq                     100   5600  ...        1
top     Bernstein Psychiatry - Ohpiatan Dr Mehdi  All  ...  COVID-19

       Start Date Primary Completion Date  Completion Date  \
count        5793                    5793             5793
unique       1344                     966             1000
freq            1                       9               9
top     May 15, 2020  December 31, 2020  December 31, 2020
top            NaN                     NaN              NaN

       First Posted Results First Posted Last Update Posted
count          5793                 0               5793
unique         1344                 0                966
top     April 30, 2020               NaN  April 9, 2020
freq             18                 NaN             100

                                                 Location  \
count                                                5793
unique                                               4557
freq                                                   19
top     Chemoprophylaxis, Hospitalization, Mortality...
freq                                                   19

                                                   Study Documents  \
count                                                         5793
unique                                                        5793
freq                                                             1
top     "Materialised analysis Plan", https://clinical...  ...
freq                                                             1

                                                       [illegible]
count                                                          5793
unique                                                         5793
freq                                                              1
top     https://clinicaltrials.gov/show/NCT04567843  ...
freq                                                              1

[4 rows x 25 columns]
```

```
### 3. Handling Missing Data ###

# Check for missing values
print(df.isnull().sum())
```

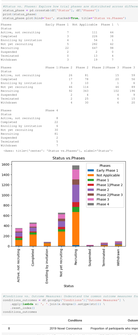
```
Rank                     0
NCT Number               0
Title                    0
Acronym                1157
Status                   0
Study Results            0
Conditions               0
Interventions            0
Outcome Measures         0
Sponsor/Collaborators    0
Gender                   3
Age                      0
Phases                 1051
Enrollment             633
Funded Bys             2599
Study Documents        4557
Other IDs                0
Start Date               0
Primary Completion Date  5
Completion Date          9
First Posted             0
Results First Posted     0
Last Update Posted       0
Locations             1401
Study First Submitted    0
Study First Posted       0
Last Update Posted       0
Results First Posted     0
Last Update Posted       0
dtype: int64
```

```
# Drop columns with a high percentage of missing values or fill them
df['Phase'] = df['Phase'].fillna('Unknown', inplace=True)
```

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

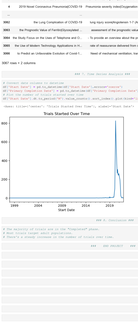
```
### 4. Descriptive Analysis ###

# Count distributions by phase the status of clinical trials.
plt.rcParams['font.family'] = 'sans-serif'
sns.countplot(data=df, x='Status', palette='Set2', hue='Status', legend=False)
plt.title('Status of Clinical Trials')
```

```
<Axes:xlabel='Status', ylabel='count'>
```




```
# Phase distributions: Analyze the distribution of trial phases.
plt.rcParams['font.family'] = 'sans-serif'
sns.countplot(data=df, x='Phase', palette='Set3', hue='Phase', legend=False)
plt.title('Distribution of Phases')
```

```
<Axes:xlabel='Phase', ylabel='count'>
Text(0.5, 1.0, 'Distribution of Phases')
```




```
# Age Group Analysis: Analyze the distribution of age groups
plt.rcParams['font.family'] = 'sans-serif'
df['Age'] = df['Age'].replace({'Child (birth-17)': 'Children', 'Adult (18-64)': 'Adult', 'Senior (65+)': 'Senior'})
sns.countplot(data=df, x='Age', palette='Set1', hue='Age', legend=False)
plt.title('Age Group Distribution')
```

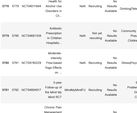
```
Age
Adult (18-64)                 2895
Child (birth-17)              1464
Senior (65+)                  1432
Adult (18-64)|Child (birth-17)   1
Adult (18-64)|Senior (65+)       1
Child (birth-17)|Senior (65+)    1
Name: count, dtype: int64
<Axes:xlabel='Age', ylabel='count'>
Text(0.5, 1.0, 'Age Group Distribution')
```



```
### 5. Bivariate Analysis ###

# Relate the Phases Explore how trial phases and distribution across different statuses.
plt.rcParams['font.family'] = 'sans-serif'
pivot_table = pd.crosstab(df['Phase'], df['Status'])
sns.heatmap(pivot_table, annot=True, fmt='d', cmap='Blues', linewidths=0.5)
```

```
Status                 Early Phase 1  Not Applicable  Phase 1  \
Phase
Active, not recruiting             1             217       30
Completed                          1             719       98
Enrolling by invitation            0               1        0
Not yet recruiting                 0              97       17
Recruiting                         0             347       40
Suspended                          0               1        0
Withdrawn                          0               1        0
Unknown                            0               1        0

Status                 Phase 1/Phase 2  Phase 2  Phase 2/Phase 3  Phase 3  \
Phase
Active, not recruiting               5       19               15       19
Completed                            3       75               35       32
Enrolling by invitation              0        0                0        0
Not yet recruiting                   1        8                2        2
Recruiting                           9      165               112      174
Suspended                            0        0                 0        1
Withdrawn                            1        0                 0        0
Unknown                              0        1                 0        0

Status                 Phase 4
Phase
Active, not recruiting       8
Completed                   13
Enrolling by invitation      0
Not yet recruiting           0
Recruiting                  41
Suspended                    0
Withdrawn                    0
Unknown                      0
```

```
<Axes:xlabel='Status', ylabel='Phase'>
```




```
# Additional vs. Outcome Measures Explore the common outcome measures for different conditions.
condition_counts = df.groupby('Conditions')['Outcome Measures'].\
    .value_counts()
top_conditions = condition_counts.\
    .groupby('Conditions').head(1)
```

	Conditions	Outcome Measures
0	2019 Novel Coronavirus	Proportion of participants who improve by at l...
1	2019 Novel Coronavirus Infection	Inter-observer COVID-19Severe of Participants with...
2	2019 Novel Coronavirus Infection|COVID-19 Vac...	Number of participants with treatment emergent...
3	2019 Novel Coronavirus Pneumonia	Clinical Victory|Time|Complete Novel Intelli Ca...
4	2019 Novel Coronavirus Pneumonia|COVID-19	Pneumonia severity index|Oxygenation index (Pa...
...	...	...
1962	Ear Lung Complication of COVID-19	Lung injury score|Hypoxemia 1-7 (Avg 1-7)...|H...
1963	Ibn Ghazala Value of Ventilo|Suspended	assessment of the prognostic value of Ventilo...
1964	Ibn Ghazala Value of the Index of Respiratory and C...	To provide an overview about the early and co...
1965	Ibn Zoar of Modern Technology Applications in A...	rate of reassurance delivered from doctors by ...
1966	Ivy Predict an Unfavorable Evolution of Covid-1...	Need of mechanical ventilation, transfer to an ...

```
1967 rows × 2 columns
```

```
### 7. Time Series Analysis ###

# Convert date columns to datetime
df['Start Date'] = pd.to_datetime(df['Start Date'],errors='coerce')
df['Completion Date'] = pd.to_datetime(df['Completion Date'],errors='coerce')
# Plot the number of trials over time.
df['Start Date'].dropna().resample('M').count().plot(kind='line', color='teal')
plt.title('Trials Started Over Time')
```

```
<Axes:xlabel='Trials Started Over Time', ylabel='Start Date'>
```



```
### 8. Conclusion ###

# The majority of trials are in the "Completed" phase.
# Most trials are conducted in Phase 3.
# There's a steady increase in the number of trials over time.

### END CONCLUSION ###
```