COVID-19 DATASET

In [ ]:

Objective - covid-19 data analysis using python

In [ ]:

Import some libraries

In [164…

```
import pandas as pd
import numpy as np
import requests
import json
import matplotlib  as plt
```

In [ ]:

Load data

In [165…

```
df1 = pd.read_csv('covid-19 india.csv')
df2 = pd.read_csv('covid-19 india2.csv')
df3 = pd.read_csv('covid-19 india3.csv')
df4 = pd.read_csv('covid-19 india4.csv')
df5 = pd.read_csv('covid-19 india5.csv')
```

Inspect data frame

In [166…

```
df1
```

Out[166…

	Patient Number	State Patient Number	Date Announced	Estimated Onset Date	Age Bracket	Gender	Detected City	Detected District	Dete S
0	1.0	KL-TS-P1	30/01/2020	NaN	20	F	Thrissur	Thrissur	Ke
1	2.0	KL-AL-P1	02/02/2020	NaN	NaN	NaN	Alappuzha	Alappuzha	Ke
2	3.0	KL-KS-P1	03/02/2020	NaN	NaN	NaN	Kasaragod	Kasaragod	Ke

	Patient Number	State Patient Number	Date Announced	Estimated Onset Date	Age Bracket	Gender	Detected City	Detected District	Dete S
3	4.0	DL-P1	02/03/2020	NaN	45	M	East Delhi (Mayur Vihar)	East Delhi	D
4	5.0	TS-P1	02/03/2020	NaN	24	M	Hyderabad	Hyderabad	Telang
...	...	...	...	...	...	...	...	...	
17359	NaN	NaN	16/04/2020	NaN	NaN	NaN	NaN	NaN	Maharas
17360	NaN	NaN	18/04/2020	NaN	NaN	NaN	NaN	Nagpur	Maharas
17361	NaN	NaN	18/04/2020	NaN	NaN	NaN	NaN	NaN	Maharas
17362	NaN	NaN	19/04/2020	NaN	NaN	NaN	NaN	Nagpur	Maharas
17363	NaN	NaN	19/04/2020	NaN	NaN	NaN	NaN	NaN	Maharas

17364 rows × 21 columns

check all the columns

In [167...
```
df1.columns
```

Out[167...
```
Index(['Patient Number', 'State Patient Number', 'Date Announced',
       'Estimated Onset Date', 'Age Bracket', 'Gender', 'Detected City',
       'Detected District', 'Detected State', 'State code', 'Current Status',
       'Notes', 'Contracted from which Patient (Suspected)', 'Nationality',
       'Type of transmission', 'Status Change Date', 'Source_1', 'Source_2',
       'Source_3', 'Backup Notes', 'Num Cases'],
      dtype='object')
```

In [168...
```
df2.columns
```

Out[168...
```
Index(['Patient Number', 'State Patient Number', 'Date Announced',
       'Estimated Onset Date', 'Age Bracket', 'Gender', 'Detected City',
       'Detected District', 'Detected State', 'State code', 'Current Status',
       'Notes', 'Contracted from which Patient (Suspected)', 'Nationality',
       'Type of transmission', 'Status Change Date', 'Source_1', 'Source_2',
```

```
       'Source_3', 'Backup Notes', 'Num Cases'],
      dtype='object')
```

In [169…
```
df3.columns
```

Out[169…
```
Index(['Entry_ID', 'State Patient Number', 'Date Announced', 'Age Bracket',
       'Gender', 'Detected City', 'Detected District', 'Detected State',
       'State code', 'Num Cases', 'Current Status',
       'Contracted from which Patient (Suspected)', 'Notes', 'Source_1',
       'Source_2', 'Source_3', 'Nationality', 'Type of transmission',
       'Status Change Date', 'Patient Number'],
      dtype='object')
```

In [170…
```
df4.columns
```

Out[170…
```
Index(['Entry_ID', 'State Patient Number', 'Date Announced', 'Age Bracket',
       'Gender', 'Detected City', 'Detected District', 'Detected State',
       'State code', 'Num Cases', 'Current Status',
       'Contracted from which Patient (Suspected)', 'Notes', 'Source_1',
       'Source_2', 'Source_3', 'Nationality', 'Type of transmission',
       'Status Change Date', 'Patient Number'],
      dtype='object')
```

In [171…
```
df5.columns
```

Out[171…
```
Index(['Entry_ID', 'State Patient Number', 'Date Announced', 'Age Bracket',
       'Gender', 'Detected City', 'Detected District', 'Detected State',
       'State code', 'Num Cases', 'Current Status',
       'Contracted from which Patient (Suspected)', 'Notes', 'Source_1',
       'Source_2', 'Source_3', 'Nationality', 'Type of transmission',
       'Status Change Date', 'Patient Number'],
      dtype='object')
```

obtain necessary columns

In [172…
```
df1 = df1.loc[:,['Num Cases','Date Announced','Age Bracket','Gender', 'Detected City
 'State code','Current Status']]
df2 = df2.loc[:,['Num Cases','Date Announced','Age Bracket','Gender', 'Detected City
 'State code','Current Status']]
df3 = df3.loc[:,['Num Cases','Date Announced','Age Bracket','Gender', 'Detected City
 'State code','Current Status']]
df4 = df4.loc[:,['Num Cases','Date Announced','Age Bracket','Gender', 'Detected City
 'State code','Current Status']]
df5 = df5.loc[:,['Num Cases','Date Announced','Age Bracket','Gender', 'Detected City
 'State code','Current Status']]

df1
```

Out[172…

	Num Cases	Date Announced	Age Bracket	Gender	Detected City	Detected District	Detected State	State code	Current Status
0	1	30/01/2020	20	F	Thrissur	Thrissur	Kerala	KL	Recovered
1	1	02/02/2020	NaN	NaN	Alappuzha	Alappuzha	Kerala	KL	Recovered
2	1	03/02/2020	NaN	NaN	Kasaragod	Kasaragod	Kerala	KL	Recovered
3	1	02/03/2020	45	M	East Delhi (Mayur Vihar)	East Delhi	Delhi	DL	Recovered

	Num Cases	Date Announced	Age Bracket	Gender	Detected City	Detected District	Detected State	State code	Current Status
4	1	02/03/2020	24	M	Hyderabad	Hyderabad	Telangana	TG	Recovered
...	...	...	...	...	...	...	...	...	...
17359	-2	16/04/2020	NaN	NaN	NaN	NaN	Maharashtra	MH	Hospitalized
17360	1	18/04/2020	NaN	NaN	NaN	Nagpur	Maharashtra	MH	Hospitalized
17361	-1	18/04/2020	NaN	NaN	NaN	NaN	Maharashtra	MH	Hospitalized
17362	10	19/04/2020	NaN	NaN	NaN	Nagpur	Maharashtra	MH	Hospitalized
17363	-10	19/04/2020	NaN	NaN	NaN	NaN	Maharashtra	MH	Hospitalized

17364 rows × 9 columns

Combine dataframe

In [173...]
```
df = df1.append([df2,df3,df4,df5])
df
```

Out[173...]

	Num Cases	Date Announced	Age Bracket	Gender	Detected City	Detected District	Detected State	State code	Current Status
0	1.0	30/01/2020	20	F	Thrissur	Thrissur	Kerala	KL	Recovered
1	1.0	02/02/2020	NaN	NaN	Alappuzha	Alappuzha	Kerala	KL	Recovered
2	1.0	03/02/2020	NaN	NaN	Kasaragod	Kasaragod	Kerala	KL	Recovered
3	1.0	02/03/2020	45	M	East Delhi (Mayur Vihar)	East Delhi	Delhi	DL	Recovered
4	1.0	02/03/2020	24	M	Hyderabad	Hyderabad	Telangana	TG	Recovered
...	...	...	...	...	...	...	...	...	...
20433	1.0	04/06/2020	NaN	NaN	NaN	Alappuzha	Kerala	KL	Hospitalized
20434	1.0	04/06/2020	NaN	NaN	NaN	Alappuzha	Kerala	KL	Hospitalized
20435	1.0	04/06/2020	NaN	NaN	NaN	Alappuzha	Kerala	KL	Hospitalized
20436	1.0	04/06/2020	NaN	NaN	NaN	Alappuzha	Kerala	KL	Hospitalized
20437	1.0	04/06/2020	NaN	NaN	NaN	Alappuzha	Kerala	KL	Hospitalized

76861 rows × 9 columns

show date as day, month, year

In [174...]
```
date = df['Date Announced'].str.split('/',expand=True)
date.columns = ['Day','Month','Year']
date
```

Out[174...

	Day	Month	Year
0	30	01	2020
1	02	02	2020
2	03	02	2020
3	02	03	2020
4	02	03	2020
...	...	...	...
20433	04	06	2020
20434	04	06	2020
20435	04	06	2020
20436	04	06	2020
20437	04	06	2020

76861 rows × 3 columns

In [175...

```
df = pd.concat([df,date],axis=0)
df
```

Out[175...

	Num Cases	Date Announced	Age Bracket	Gender	Detected City	Detected District	Detected State	State code	Current Status	D
0	1.0	30/01/2020	20	F	Thrissur	Thrissur	Kerala	KL	Recovered	Na
1	1.0	02/02/2020	NaN	NaN	Alappuzha	Alappuzha	Kerala	KL	Recovered	Na
2	1.0	03/02/2020	NaN	NaN	Kasaragod	Kasaragod	Kerala	KL	Recovered	Na
3	1.0	02/03/2020	45	M	East Delhi (Mayur Vihar)	East Delhi	Delhi	DL	Recovered	Na
4	1.0	02/03/2020	24	M	Hyderabad	Hyderabad	Telangana	TG	Recovered	Na
...	...	...	...	...	...	...	...	...	...	
20433	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
20434	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
20435	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
20436	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
20437	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

153722 rows × 12 columns

five records from starting

In [176...

```
df.head()
```

Out[176...

	Num Cases	Date Announced	Age Bracket	Gender	Detected City	Detected District	Detected State	State code	Current Status	Day	M
0	1.0	30/01/2020	20	F	Thrissur	Thrissur	Kerala	KL	Recovered	NaN	
1	1.0	02/02/2020	NaN	NaN	Alappuzha	Alappuzha	Kerala	KL	Recovered	NaN	
2	1.0	03/02/2020	NaN	NaN	Kasaragod	Kasaragod	Kerala	KL	Recovered	NaN	
3	1.0	02/03/2020	45	M	East Delhi (Mayur Vihar)	East Delhi	Delhi	DL	Recovered	NaN	
4	1.0	02/03/2020	24	M	Hyderabad	Hyderabad	Telangana	TG	Recovered	NaN	

five records from end

In [177...

```
df.tail()
```

Out[177...

	Num Cases	Date Announced	Age Bracket	Gender	Detected City	Detected District	Detected State	State code	Current Status	Day	M
20433	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	04	
20434	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	04	
20435	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	04	
20436	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	04	
20437	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	04	

df.sort('Date Announced')

Inspect dataframe

In [178...

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 153722 entries, 0 to 20437
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Num Cases          76858 non-null  float64
 1   Date Announced     76861 non-null  object
 2   Age Bracket        30751 non-null  object
 3   Gender             33408 non-null  object
 4   Detected City      5204 non-null   object
 5   Detected District  69349 non-null  object
 6   Detected State     76853 non-null  object
 7   State code         76853 non-null  object
 8   Current Status     76859 non-null  object
 9   Day                76861 non-null  object
 10  Month              76861 non-null  object
 11  Year               76861 non-null  object
dtypes: float64(1), object(11)
memory usage: 15.2+ MB
```

Inspect null values in each column

In [179…
```
df.isnull().sum(axis=0).sort_values(ascending=False)
```

Out[179…
```
Detected City        148518
Age Bracket          122971
Gender               120314
Detected District     84373
Detected State        76869
State code            76869
Num Cases             76864
Current Status        76863
Date Announced        76861
Day                   76861
Month                 76861
Year                  76861
dtype: int64
```

null values in each row

In [180…
```
df.isnull().sum(axis=1).sort_values(ascending=False)
```

Out[180…
```
12585    10
268      10
12586    10
8261      9
8255      9
         ..
3399      3
3400      3
3401      3
3402      3
0         3
Length: 153722, dtype: int64
```

total male/female infected with coronavirus

In [184…
```
df.groupby("Gender")["Num Cases"].sum()
```

Out[184…
```
Gender
F             11756.0
M             22744.0
Non-Binary        5.0
Name: Num Cases, dtype: float64
```

which age group is infected most?

In [186…
```
df.groupby("Gender")["Num Cases"].sum().sort_values(ascending=False)
```

Out[186…
```
Gender
M             22744.0
F             11756.0
Non-Binary        5.0
Name: Num Cases, dtype: float64
```

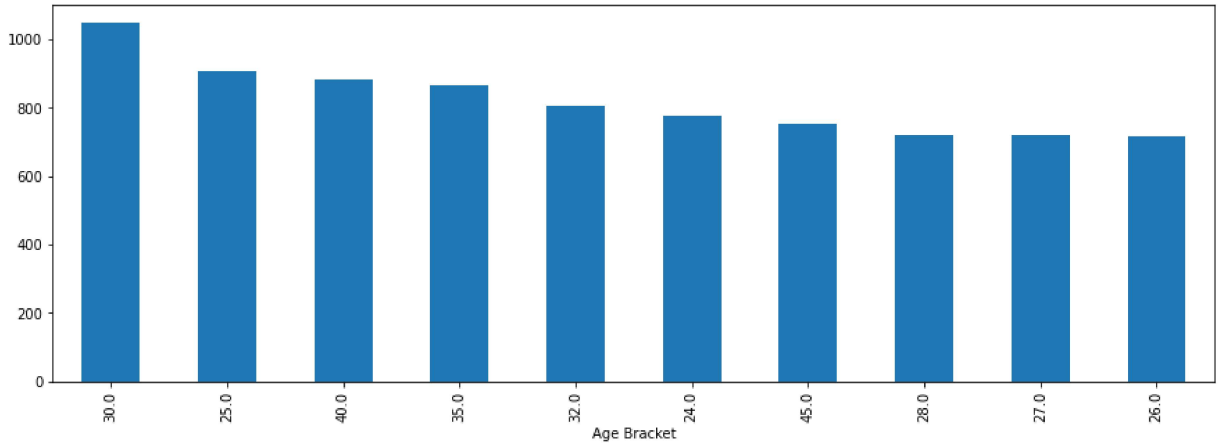
In [187…
```
m=df.groupby("Age Bracket")["Num Cases"].sum().sort_values(ascending=False).head(10)
```

```
m
```

Out[187…
```
Age Bracket
30.0    1050.0
25.0     905.0
40.0     882.0
35.0     867.0
32.0     806.0
24.0     778.0
45.0     754.0
28.0     722.0
27.0     719.0
26.0     717.0
Name: Num Cases, dtype: float64
```

In [189…
```
m.plot.bar(figsize=(15,5))
plt
```

Out[189…
```
<module 'matplotlib' from 'c:\\users\\shiv pratap singh\\appdata\\local\\programs\\python\\python39\\lib\\site-packages\\matplotlib\\__init__.py'>
```



state wise cases in india

In [191…
```
m = df[df['Current Status']=='Hospitalized'].groupby('Detected State')['Num Cases'].
m
```

Out[191…
```
Detected State
Maharashtra                 77779.0
Tamil Nadu                  27249.0
Delhi                       25000.0
Gujarat                     18604.0
Rajasthan                    9858.0
Uttar Pradesh                9228.0
Madhya Pradesh               8760.0
State Unassigned             7483.0
West Bengal                  6870.0
Bihar                        4451.0
Karnataka                    4248.0
Andhra Pradesh               4110.0
Haryana                      3269.0
Telangana                    3145.0
Jammu and Kashmir            3141.0
Odisha                       2477.0
Punjab                       2413.0
Assam                        2116.0
Kerala                       1528.0
Uttarakhand                  1152.0
Jharkhand                     827.0
Chhattisgarh                  765.0
```

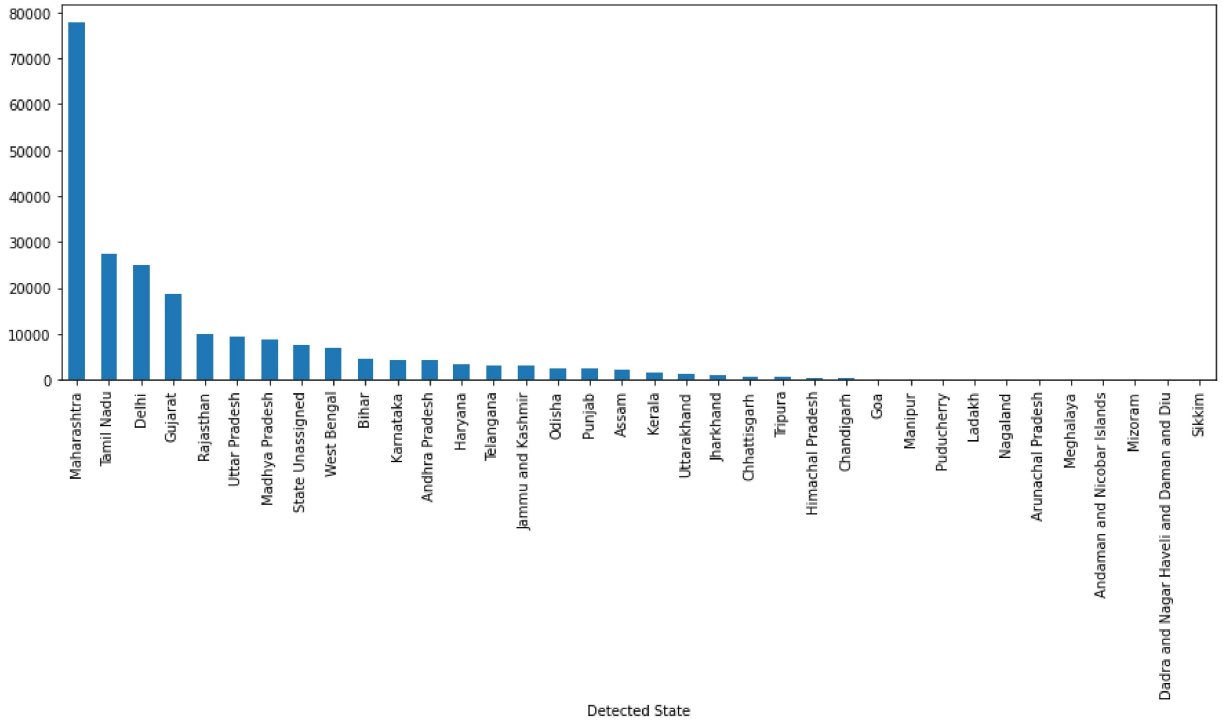
```
Tripura                                    646.0
Himachal Pradesh                           381.0
Chandigarh                                 302.0
Goa                                        166.0
Manipur                                    123.0
Puducherry                                  99.0
Ladakh                                      94.0
Nagaland                                    80.0
Arunachal Pradesh                           42.0
Meghalaya                                   33.0
Andaman and Nicobar Islands                 22.0
Mizoram                                     17.0
Dadra and Nagar Haveli and Daman and Diu    14.0
Sikkim                                       2.0
Name: Num Cases, dtype: float64
```

In [192…

```
m.plot.bar(figsize=(15,5))
plt
```

Out[192…

```
<module 'matplotlib' from 'c:\\users\\shiv pratap singh\\appdata\\local\\programs\\python\\python39\\lib\\site-packages\\matplotlib\\__init__.py'>
```



In [204…

```
df.to_csv('covid-19 Mini project.csv')
```

In [ ]: