Data Analysis with Python

Topic: Covid19 India Analysis

Project By:

Dinkal Kewlani

Index:

SR. NO.	Title	Page no.
1	Introduction	3
2	Project Aim	5
3	Data Analysis	5
4	Age Analysis	9
5	Public health facility analysis	10
6	ICMR testing labs analysis	14
7	Gender, district and state wise analysis	15
8	State wise testing details	17
9	Conclusion	18

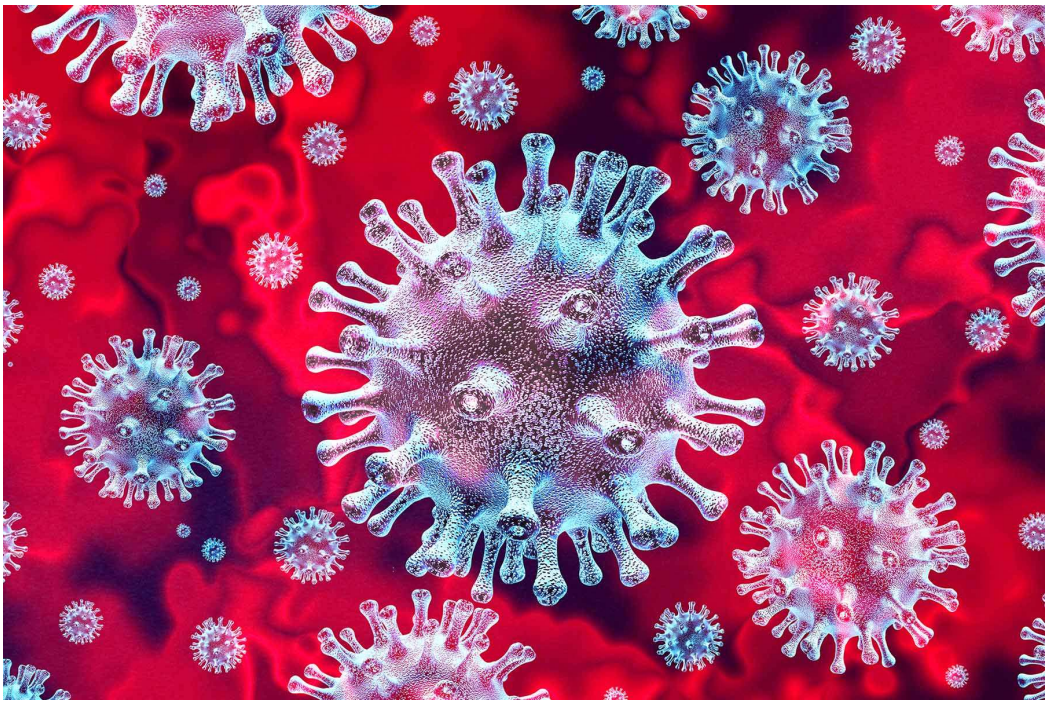
Introduction:

Coronaviruses are a large family of viruses that may cause respiratory illnesses in humans ranging from common colds to more severe conditions such as Severe Acute Respiratory Syndrome (SARS) and Middle Eastern Respiratory Syndrome (MERS).

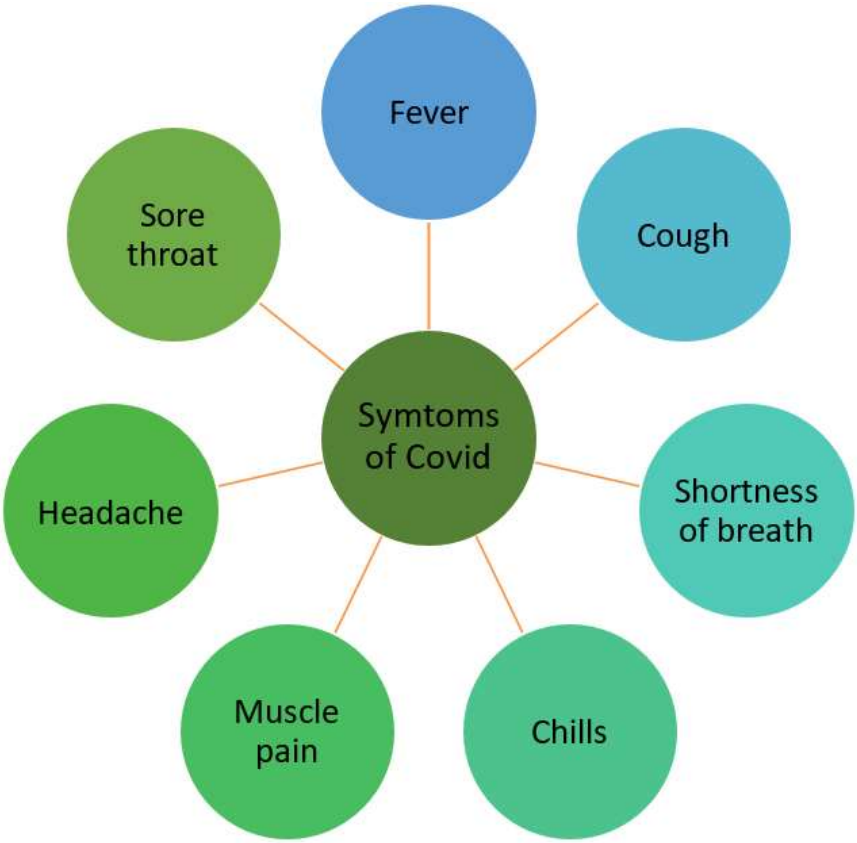
COVID-19 appeared in Wuhan, a city in China, in December 2019. Although health officials are still tracing the exact source of this new coronavirus, early hypotheses thought it may be linked to a seafood market in Wuhan, China. Some people who visited the market developed viral pneumonia caused by the new coronavirus. A study that came out on Jan. 25, 2020, notes that the individual with the first reported case became ill on Dec. 1, 2019, and had no link to the seafood market. Investigations are ongoing as to how this virus originated and spread.

COVID-19 can spread from person to person usually through close contact with an infected person or through respiratory droplets that are dispersed into the air when an infected person coughs or sneezes. It may also be possible to get the virus by touching a surface or object contaminated with the virus and then touching your mouth, nose or eyes, but it is not thought to be the main way the virus spreads. Similar to other respiratory illnesses, the symptoms of COVID-19 may include fever, cough, and shortness of breath.

Coronavirus is a large family of viruses that can infect animals or humans. In humans, several strains of viruses are known to cause respiratory infections ranging from the common cold to severe diseases such as the Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered strain is called SARS-CoV-2 strain that is causing COVID19 as it is similar to the SARS-CoV strain that had caused the SARS outbreak.



Symptoms of Covid:



Project Aim:

In this project we will analyse about COVID-19 data of India.

1. We will analyse the age group which is most affected by this virus.
2. We will analyse the number of public health facilities and number of beds.
3. We will analyse the ICMR testing labs in the country.
4. We will analyse the state wise testing details.
5. We will analyse the number of Active cases based on Gender, districts and states.

Data Analysis:

A. Datasets:

In this project we will be using 5 datasets for analysing different aspects as per the aim of our project. We have collected these datasets from various open sources like Kaggle and Github

The Pandas library is a useful tool that enables us to read various datasets into dataframe. Using this library, the csv files were converted into data frames.

1. The First Dataset is dataset tells us about how much did different age groups are affected with the COVID-19 virus in India .

```
Viewing the first dataset AgeGroupDetails to analysis which age group is most affected with corona virus.
```

```
[47]: df1.head(10)
```

t[47]:

	Sno	AgeGroup	TotalCases	Percentage
0	1	0-9	22	3.18%
1	2	10-19	27	3.90%
2	3	20-29	172	24.86%
3	4	30-39	146	21.10%
4	5	40-49	112	16.18%
5	6	50-59	77	11.13%
6	7	60-69	89	12.86%
7	8	70-79	28	4.05%
8	9	>=80	10	1.45%
9	10	Missing	9	1.30%

2. The second dataset states the number of Hospital beds in different rural and urban areas in different states of India.

```
Viewing the second dataset HospitalBedsIndia
```

```
In [48]: df2.head(7)
```

Out[48]:

	Sno	State/UT	NumPrimaryHealthCenters_HMIS	NumCommunityHealthCenters_HMIS	NumSubDistrictHospitals_HMIS	NumDistrictHospitals_HMIS
0	1	Andaman & Nicobar Islands	27	4	NaN	3
1	2	Andhra Pradesh	1417	198	31.0	20
2	3	Arunachal Pradesh	122	62	NaN	15
3	4	Assam	1007	166	14.0	33
4	5	Bihar	2007	63	33.0	43
5	6	Chandigarh	40	2	1.0	4
6	7	Chhattisgarh	813	166	12.0	32

3. The third dataset states the number of ICMR Testing Labs in India

```
In [33]: df3.head(5)
```

Out[33]:

	lab	address	pincode	city	state	type
0	ICMR-Regional Medical Research Centre, Port Blair	ICMR-Regional Medical Research Centre, Post Ba...	744103	Port Blair	Andaman and Nicobar Islands	Government Laboratory
1	Tomo Riba Institute of Health & Medical Scienc...	National Highway 52A, Old Assembly Complex, Na...	791110	Naharlagun	Arunachal Pradesh	Collection Site
2	Sri Venkateswara Institute of Medical Sciences...	Sri Venkateswara Institute of Medical Sciences...	517507	Tirupati	Andhra Pradesh	Government Laboratory
3	Rangaraya Medical College, Kakinada	Rangaraya Medical College, Kakinada Pithampura...	533001	Kakinada	Andhra Pradesh	Government Laboratory
4	Sidhartha Medical College, Vijaywada	Siddhartha Medical College, Vijayawada NH 16 S...	520008	Vijayawada	Andhra Pradesh	Government Laboratory

4. The Fourth Dataset tells the state wise testing Details in India and the details of the people getting tested.

```
In [37]: df4.head(5)
```

Out[37]:

	id	government_id	diagnosed_date	age	gender	detected_city	detected_district	detected_state	nationality	current_status	status_change_date	notes
0	0	KL-TS-P1	30/01/2020	20	F	Thrissur	Thrissur	Kerala	India	Recovered	14/02/2020	Travelled from Wuhan
1	1	KL-AL-P1	02/02/2020	NaN	NaN	Alappuzha	Alappuzha	Kerala	India	Recovered	14/02/2020	Travelled from Wuhan
2	2	KL-KS-P1	03/02/2020	NaN	NaN	Kasaragod	Kasaragod	Kerala	India	Recovered	14/02/2020	Travelled from Wuhan
3	3	DL-P1	02/03/2020	45	M	East Delhi (Mayur Vihar)	East Delhi	Delhi	India	Recovered	15/03/2020	Travelled from Austria, Italy
4	4	TS-P1	02/03/2020	24	M	Hyderabad	Hyderabad	Telangana	India	Recovered	02/03/2020	Travelled from Dubai to Bangalore on 20th Feb,...

5. The fifth dataset stated the details of the Individuals who either tested positive or negative with the novel corona virus.

```
In [49]: df5.head(10)
```

Out[49]:

	Date	State	TotalSamples	Negative	Positive
0	2020-04-17	Andaman and Nicobar Islands	1403.0	1210.0	12.0
1	2020-04-24	Andaman and Nicobar Islands	2679.0	NaN	27.0
2	2020-04-27	Andaman and Nicobar Islands	2848.0	NaN	33.0
3	2020-05-01	Andaman and Nicobar Islands	3754.0	NaN	33.0
4	2020-05-16	Andaman and Nicobar Islands	6677.0	NaN	33.0
5	2020-04-02	Andhra Pradesh	1800.0	1175.0	132.0
6	2020-04-10	Andhra Pradesh	6374.0	6009.0	365.0
7	2020-04-11	Andhra Pradesh	6958.0	6577.0	381.0
8	2020-04-12	Andhra Pradesh	6958.0	6553.0	405.0
9	2020-04-13	Andhra Pradesh	8755.0	8323.0	432.0

B. Implementation

Importing all the necessary libraries

- **Import numpy** : It can be used to perform mathematical operations on arrays such as trigonometric, statistical, and algebraic routines.
- **Import pandas :** it is used for data wrangling and analysis. It is a convenient wrapper around numpy.
- **import seaborn :** it is a visualization library based on matplotlib. It provides a high leve interface for drawing attractive and informative statistical graphics.
- **import matplotlib :** it is a plotting library which gives inline plots for quick data analysis.

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import random
        import matplotlib.colors as mcolors
```

Loading the datasets

- **read_csv() :** Data from a data file in the project directory is moved into a pandas dataframe. We can optionally specify the column names.

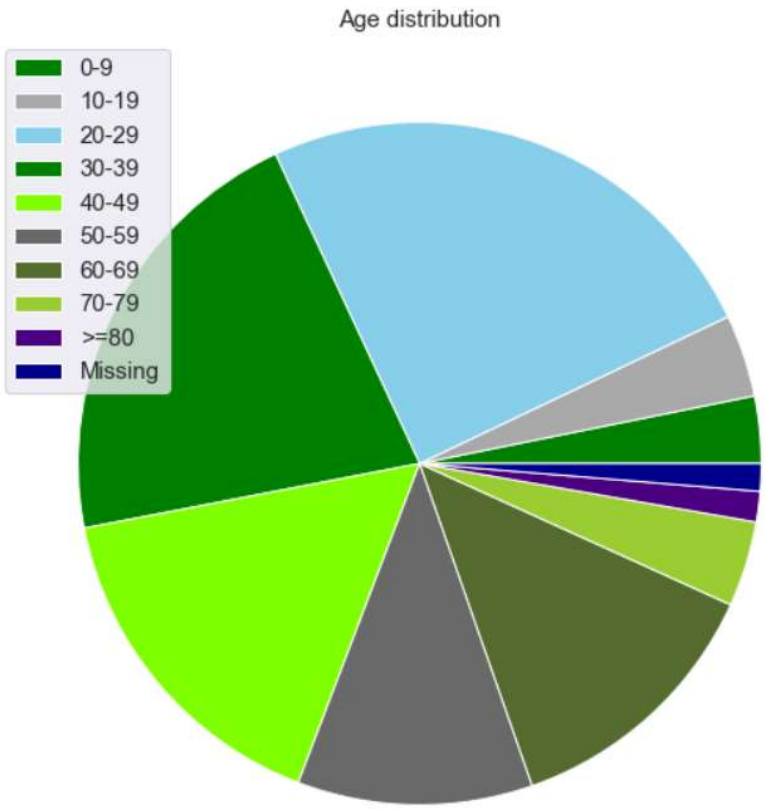
```
In [7]: df1 = pd.read_csv("AgeGroupDetails.csv")
        df2 = pd.read_csv("HospitalBedsIndia.csv")
        df3 = pd.read_csv("ICMRTestingLabs.csv")
        df5 = pd.read_csv("StatewiseTestingDetails.csv")
        df4 = pd.read_csv("IndividualDetails.csv")
```

1. Age Analysis

```
Plotting the pie chart for different age group affected with COVID-19 in india
```

```
In [20]: def plot_pie_charts(x, y, title):
             c = random.choices(list(mcolors.CSS4_COLORS.values()),k = 10)
             plt.figure(figsize=(20,15))
             plt.title(title, size=20)
             plt.pie(y, colors = c)
             plt.legend(x, loc='best', fontsize=15)
             plt.show()
```

```
In [51]: plot_pie_charts(df1['AgeGroup'], df1['TotalCases'], 'Age distribution')
```

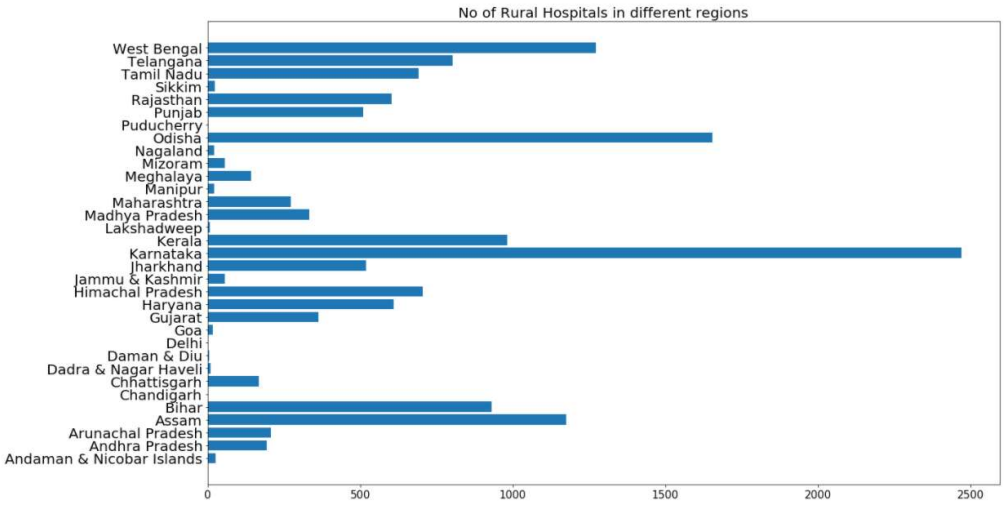


Observation: From the above graph we can observe that the age group of 20-29 is most affected by the COVID-19 in India. Further age groups of 30-39, 40-49, 50-59, 60-69 are also having a large number of cases.

2. Analysing the number of public healthcare facilities and number of hospital beds available for the patient .

```
In [22]: def plot_bar_graphs(x, y, title):
             plt.figure(figsize=(20, 12))
             plt.barh(x, y)
             plt.title(title, size=20)
             plt.xticks(size=15)
             plt.yticks(size=20)
             plt.show()
         df2 = df2.drop(df2.index[32])
```

```
In [24]: plot_bar_graphs(df2['State/UT'], df2['NumRuralHospitals_NHP18'], 'No of Rural Hospitals in different regions')
```

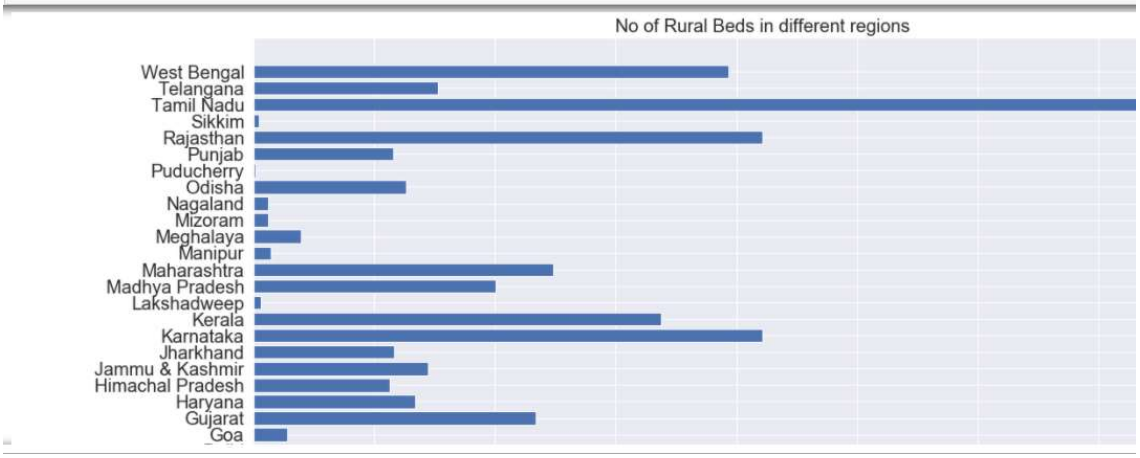


Observation: The above graph shows the plot between the states and the number of rural hospital in those states.

We can observe that:

a. Uttar Pradesh has the most number of rural hospitals that is over 4000.
b. Next Followed by Karnataka with 2500 rural hospitals is at second place.

```
plot_bar_graphs(df2['State/UT'], df2['NumRuralBeds_NHP18'], 'No of Rural Beds in different regions')
```

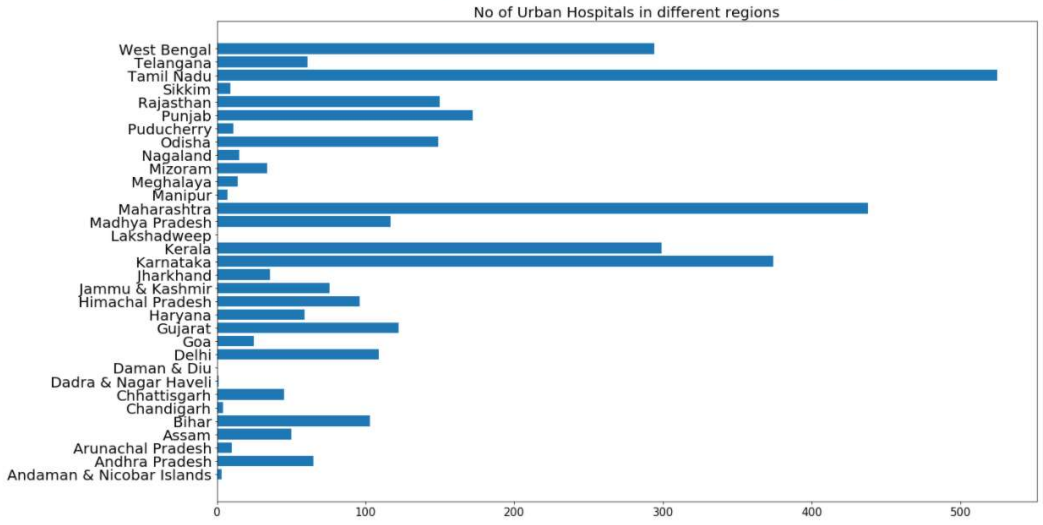


Observation: The above graph shows the plot between the states and the number of rural hospital beds in those states.

We can observe that:

a. Tamil Nadu is the state with the most beds in the rural areas with over 40000 beds.
b. At second we have Uttar Pradesh
c. At Third there are Karnataka, Rajasthan and west Bengal are having more than 20000 beds.

```
In [27]: plot_bar_graphs(df2['State/UT'], df2['NumUrbanHospitals_NHP18'], 'No of Urban Hospitals in different regions')
```

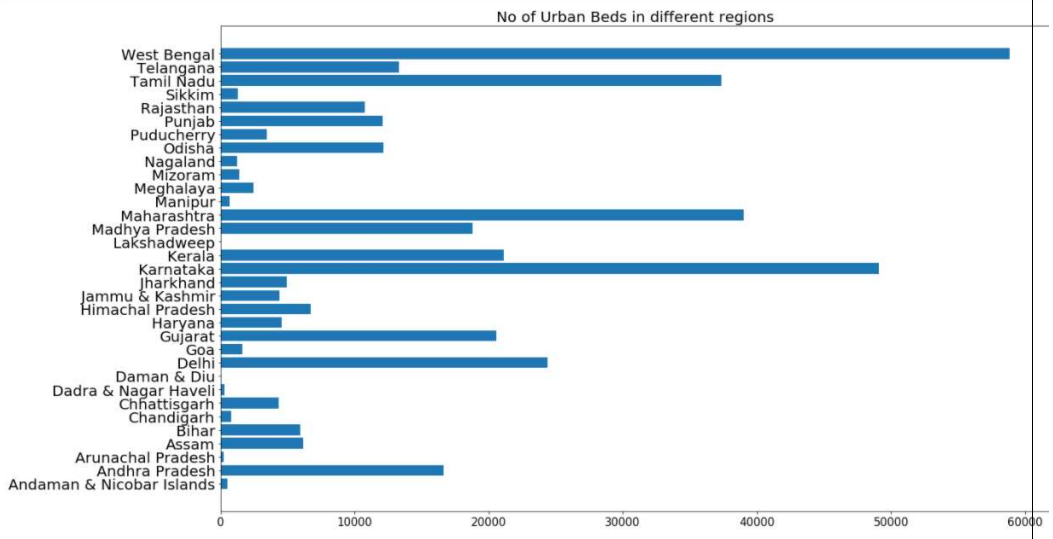


Observation: The above graph shows the plot between the states and the number of urban hospitals in those states.

We can observe that:

a. Tamil Nadu has most number of urban hospitals that is over 500.
b. Maharashtra is having over 400 urban hospitals.
c. Karnataka is having nearly 400 urban hospitals.

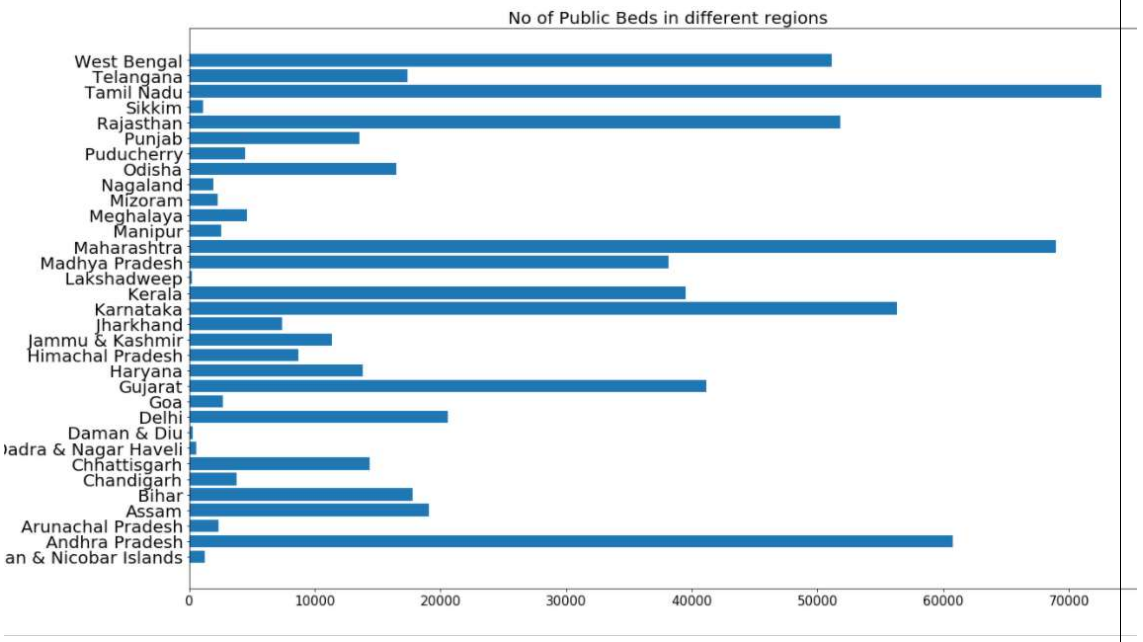
```
In [28]: plot_bar_graphs(df2['State/UT'], df2['NumUrbanBeds_NHP18'], 'No of Urban Beds in different regions')
```



Observation: The above graph shows the plot between the states and the number of urban hospital beds in those states.

We can observe that:

a. West Bengal is having the almost 60000 beds stands at first position.
b. Karnataka in second place with almost 50000 beds.
c. At third there are Uttar Pradesh, Tamil Nadu and Maharashtra with over 38000 beds.



Observation: The above graph shows the plot between the states and the number of public beds in those states.

We can observe that:

a. Tamil Nadu has most number of beds with over 70000 of them.
b. At second there are Maharashtra and Andhra Pradesh with over 60000 beds.

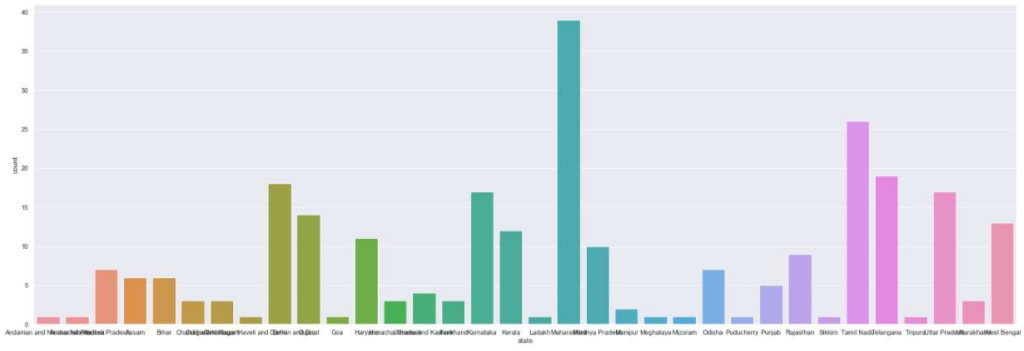
3. Analysing the number of ICMR testing labs in India.

```
In [34]: df3['state'].value_counts()
```

```
Out[34]: Maharashtra                          39
         Tamil Nadu                           26
         Telangana                            19
         Delhi                                18
         Uttar Pradesh                        17
         Karnataka                            17
         Gujarat                              14
         West Bengal                          13
         Kerala                               12
         Haryana                              11
         Madhya Pradesh                       10
         Rajasthan                             9
         Odisha                                7
         Andhra Pradesh                        7
         Assam                                 6
         Bihar                                 6
         Punjab                                5
         Jammu and Kashmir                     4
         Himachal Pradesh                      3
         Uttarakhand                           3
         Chandigarh                            3
         Jharkhand                             3
         Chhattisgarh                          3
         Manipur                               2
         Meghalaya                             1
         Mizoram                               1
         Sikkim                                1
         Andaman and Nicobar Islands           1
         Arunachal Pradesh                     1
         Goa                                   1
         Tripura                               1
         Dadra and Nagar Haveli and Daman and Diu  1
         Ladakh                                1
         Puducherry                            1
         Name: state, dtype: int64
```

```
In [35]: sns.set(rc={'figure.figsize':(30,10)})
         sns.countplot(x = "state", data = df3)
```

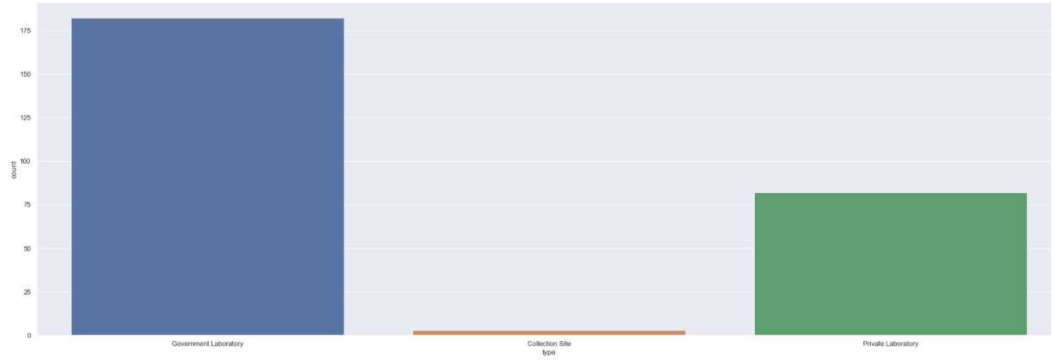
```
Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0x21fb5a55b08>
```



Observation:

a. From the above graph we can observe that Maharashtra has the most number of ICMR testing labs that is 39.
b. Tamil Nadu is at second position with 26 labs
c. Telangana has 19 labs.

```
In [36]: sns.countplot(x = "type", data = df3)
Out[36]: <matplotlib.axes._subplots.AxesSubplot at 0x21fb5901a08>
```



Observation:

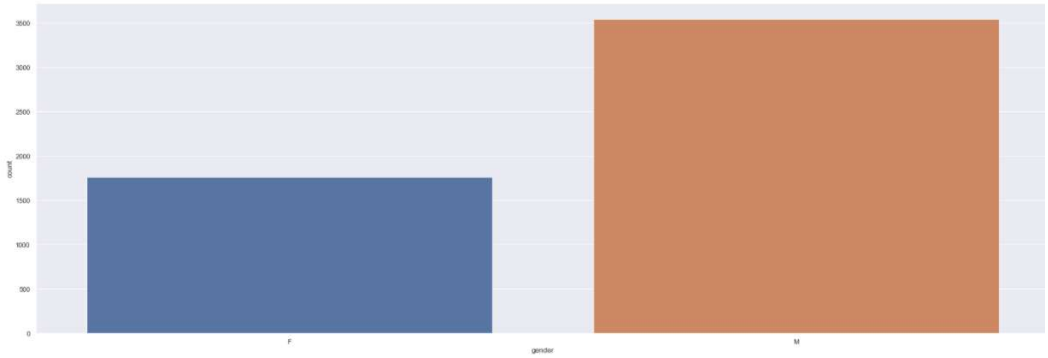
From the above graph we can observe that around 180 of the ICMR testing labs in India are Government labs and around 80 are Private labs.

4. Analysing based on Gender, districts, States and Number of Active cases.

```
In [38]: df4['gender'].isna().sum()
Out[38]: 22577

In [39]: df4['gender'].value_counts()
Out[39]: M    3547
         F    1766
         Name: gender, dtype: int64
```

```
In [40]: sns.countplot(x = "gender", data = df4)
Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x21fb5ccecc8>
```

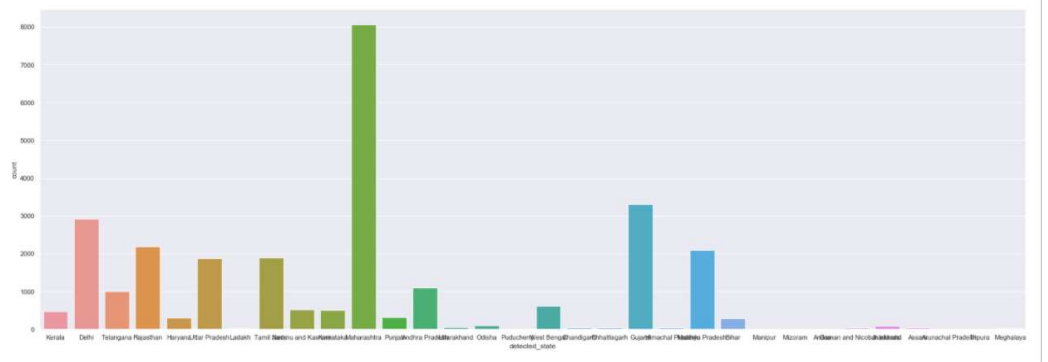


Observation: From the above graph we can observe that Males are more likely to get affect by the virus. The ratio of males getting affected is almost 2x times than females.

```
In [41]: df4['detected_district'].value_counts()
Out[41]: Mumbai              2687
         Ahmedabad           2181
         Indore              1036
         Jaipur               808
         Pune                 680
                              ...
         Badgam                 1
         Jalaun                 1
         Rajsamand              1
         Sri Muktsar Sahib      1
         North East Delhi       1
         Name: detected_district, Length: 449, dtype: int64
```

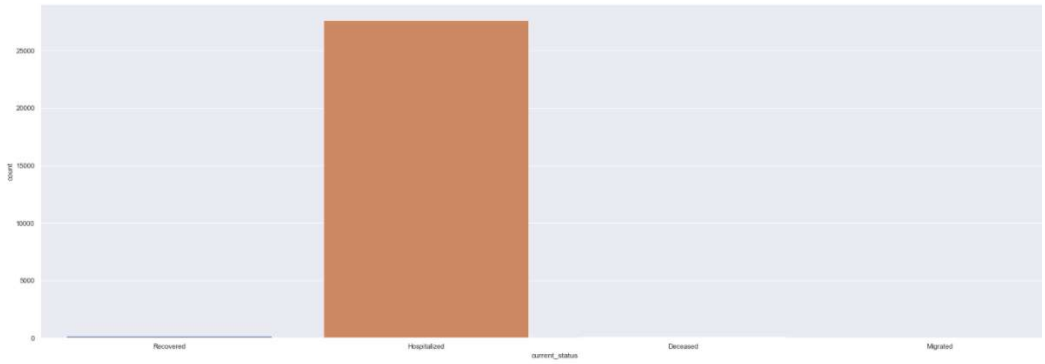
Observation: From the above graph we can observe that majority of the cases have been seen in Mumbai when seen district wise.

```
In [42]: sns.countplot(x = "detected_state", data = df4)
Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x21fb5a2d088>
```



Observation: From the above graph we can observe that most cases have been observed in the state of Maharashtra when seen in state wise analysis.

```
In [43]: sns.countplot(x = "current_status", data = df4)
Out[43]: <matplotlib.axes._subplots.AxesSubplot at 0x21fb59e0508>
```



Observation: From the above graph we can observe that most of the cases are still active and only a small percentage has recovered deceased or migrated.

5. Analysing based on state wise testing details.

```
In [45]: df5_sort = df5.sort_values(by = 'TotalSamples', ascending=False).head()
         df5_sort.head()
```

Out[45]:

	Date	State	TotalSamples	Negative	Positive
1068	2020-05-20	Tamil Nadu	360068.0	3346311.0	13191.0
1067	2020-05-19	Tamil Nadu	348174.0	334839.0	12448.0
1066	2020-05-18	Tamil Nadu	337841.0	325546.0	11760.0
1065	2020-05-17	Tamil Nadu	326720.0	315019.0	11224.0
1064	2020-05-16	Tamil Nadu	313639.0	302523.0	10585.0

Observation: From the above graph we can observe that Tamil Nadu has the highest number of samples being tested.

In [46]:
```
df5_sort1 = df5.sort_values(by = 'TotalSamples', ascending=True).head()
df5_sort1.head()
```

Out[46]:

	Date	State	TotalSamples	Negative	Positive
765	2020-04-07	Mizoram	58.0	0.0	1.0
764	2020-04-06	Mizoram	58.0	0.0	1.0
804	2020-04-06	Nagaland	60.0	47.0	0.0
806	2020-04-11	Nagaland	70.0	70.0	0.0
805	2020-04-10	Nagaland	70.0	69.0	0.0

Observation: From the above graph we can observe that Mizoram and Nagaland have the lowest number of tested samples recorded in one day.

C. Conclusion:

We have observed some very useful information from this project regarding the COVID-19 spread in India. We have analysed the datasets according to the aim of our project and successfully found some useful information that can be further explored and used to create models for predicting different aspects related to COVID-19 spread in India.

D. Project Link:

https://colab.research.google.com/drive/1iZVWR0x9N-PqS3LoYjm2pg8YoMmdGwFz