Ravirajsinh Dabhi

Subject: Applied Data Science Capstone (Data section)

Title: Analysis on Covid-19 in India

Data is most essencial component of any data science project. After searching stuffs in internet I came up with one table in Wikipedia page https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_India .

en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_India

New COVID-19 cases in India by state and union territory (v·t·e) [hide]

Date (2020)	State/Union Territory																																	Cases[a]			Deaths		Source(s)
	Andaman and Nicobar Islands	Andhra Pradesh	Arunachal Pradesh	Assam	Bihar	Chandigarh	Chhattisgarh	Delhi	Goa	Gujarat	Haryana	Himachal Pradesh	Jammu and Kashmir	Jharkhand	Karnataka	Kerala	Ladakh	Madhya Pradesh	Maharashtra	Manipur	Meghalaya	Mizoram	Nagaland	Odisha	Puducherry	Punjab	Rajasthan	Tamil Nadu	Telangana	Tripura	Uttarakhand	Uttar Pradesh	West Bengal	New	Total	Diff.	New	Total	
Jan-30																1																		1	1	—			
Feb-02																1																		1	2	+100%			
Feb-03																1																		1	3	+50%			
Mar-02								1																					1					2	5	+67%			[482]
Mar-03																											1							1	6	+20%			[483]
Mar-04								1			14																1					6		22	28	+356%			[484][485]
Mar-05								1																								1		2	30	+7%			[485]
Mar-06								1																										1	31	+3%			[485]
Mar-07																	2											1						3	34	+10%			[485]
Mar-08																5																		5	39	+15%			[486][487]
Mar-09								1					1			1										1						1		5	44	+13%			[488][489]
Mar-10															4				2															6	50	+14%			[490]
Mar-11								1								8											1							10	60	+20%			[491]
Mar-12		1						2							0(1)		1		6								1					2		13	73	+22%	1	1	[492][493]
Mar-13															2	2			3													1		8	81	+11%		1	[494][495]
Mar-14								1(1)					1						12								1					1		16	97	+20%	1	2	[496]
Mar-15																3			5										2					10	107	+10%		2	
Mar-16													1		1	1	1		6					1										11	118	+10%		2	[497][498]
Mar-17								1			1				3	3	4		3(1)										2			2		19	137	+16%	1	3	
Mar-18								2			2				1	1			3						1		1					2	1	14	151	+10%		3	

This table is what actually I want. It contains day to day record of corona cases in all states of the India from first case in record in India. It has all data which is necessary to do various kind of analysis. It contains each day cases and deaths of all states, Difference from previous day recorded data,states total cases and death and whole country's total deaths and cases.

I scrap the Wikipedia web page with help of BeautifulSoup python library and convert it in pandas dataframe. Finding particular table require some trial and error. Some code in below image.

```
[4]: res = requests.get("https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_India")
     soup = BeautifulSoup(res.content,'lxml')

[5]: table = soup.find_all('table')[4]
     df= pd.read_html(str(table))

[6]: day_to_day = pd.DataFrame(df[0]).T
     day_to_day.head()
```

[6]:

		0	1	2	3	4	5	6	7	8	9	...	56	57	58	59	60	61	62
Date (2020)	**Date (2020)**	Jan-30	Feb-02	Feb-03	Mar-02	Mar-03	Mar-04	Mar-05	Mar-06	Mar-07	Mar-08	...	Apr-24	Apr-25	Apr-26	Apr-27	Apr-28	Apr-29	Total
State/Union Territory	**Andaman and Nicobar Islands**	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	4	5	6	NaN	NaN	NaN	33
	Andhra Pradesh	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	60(2)	106(2)	36	80	82	73	1332
	Arunachal Pradesh	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	1
	Assam	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1	NaN	NaN	NaN	2	NaN	38

5 rows × 66 columns

This raw data table has lot's of missing values. If there is no cases or death record for particular day that have null value.I have to convert it into zero(0). And most of data is 'object' Datatype and most of is 'string'. From string data type we can't do exploratory analysis of data.

Some of the data is in weird format like below image. Which I have to convert in appropriate format.

56(2)	82(3)	60(2)	106(2)	36	80	82	1259	31
0	0	0	0	0	0	0	1	–
0	0	1	0	0	0	2	38	1
12[d]	22	28	52	23	26	69	346	2
1	0	0	1	2	0	10	40	–
0	0	0	0	1	0	0	37	–
75	92(1)	128(2)	138(3)	111(1)	293	190	3108	54
0	0	0	0	0	0	0	7	–
206(18)	135(8)	217(9)	191(15)	256(6)	230(18)	247(11)	3548	162

There are many operation which you will find out in notebook. After this visualization of data I am using foursquare API to find top highest rated hospitals in state in which corona cases is highest.

I found out that Maharashtra state has highest cases in india and particular Mumbai city. So I use foursquare to find highest rated hospital in Mumbai.

I wil use folium and geopy python library to visualization and find location of particular address. You can see step by step explanation and code in Notebook.