Covid Analysis

Motivation

With rapid spread of Covid-19 across the globe, the lack of planning and infrastructure support to tackle the containment of such a contagious virus seems to be plunging the world into crisis mode. The goal of this project is two-fold -
1.To provide geographical insights about the virus across the USA at state and county level.
2.To leverage hospital resource utilization data to provide solutions.
The idea behind the goal is to aid optimal planning and resource allocation which play a significant role in overcoming this pandemic.

Datasets Utilized

1.NY Times COVID-19 Datasets

The New York Times has released a series of data files with cumulative counts of coronavirus cases in the United States, at the state and county level, over time. These datasets are updated on a regular basis.
Files utilised : U.S. State-Level Data (Raw CSV) | U.S. County-Level Data (Raw CSV)
The state level file provides cumulative state level Covid-19 data.
date,state,fips,cases,deaths
2020-01-21,Washington,53,1,0
The county level file provides cumulative county level Covid-19 data.
date,county,state,fips,cases,deaths
2020-01-21,Snohomish,Washington,53061,1,0

2.Definitive Healthcare: USA Hospital Beds

The dataset provides detailed information about typical bed capacity and average yearly bed utilization of hospitals in the United States. The dataset is not updated in real-time, this information is critical for understanding the impact of a high utilization event, like COVID-19.
Columns of Interest -

- *State name:* State name
- *County Name:* County Name
- *Number of Licensed Beds:* is the maximum number of beds for which a hospital holds a license to operate; however, many hospitals do not operate all the beds for which they are licensed. This number is obtained through DHC Primary Research. Licensed beds for Health Systems are equal to the total number of licensed beds of individual Hospitals within a given Health System.
- *Potential Increase in Bed Capacity:* This metric is computed by subtracting "Number of Staffed Beds from Number of Licensed beds" (Licensed Beds – Staffed Beds).

3.US County Level Census Data

Census data collected by the government utilised in the project for the purpose of population estimates at different levels of granularity. Data is current as of Release Date: March2020

Columns of Interest -

- *STNAME* : State name
- *CTYNAME* : County Name
- *SUMLEV* : Geographic summary level such that 040 = State and/or Statistical Equivalent and 050 = County and /or Statistical Equivalent
- *POPESTIMATE2019* : 7/1/2019 resident total population estimate

Business Question and Results

Geographical Insights

1. Top 10 states that are hardest hit when the cases are scaled with their population and their corresponding death rates
   Result:

state	state_population	cases	case_rates_percent	deaths	death_rates_percent
New York	38907122	321276	0.8257511311168172	19645	0.0504920410201505
New Jersey	17764380	130593	0.7351396446146727	8244	0.046407473832466996
Massachusetts	13785006	70271	0.5097640146112377	4212	0.03055493773452112
Rhode Island	2118722	9933	0.4688203549120649	355	0.016755383669967084
Connecticut	7130574	30621	0.42943246925142353	2633	0.036925498564351206
District of Columbia	1411498	5322	0.3770462303170107 3	264	0.018703533409186554
Louisiana	9297588	29996	0.32262130780585246	2042	0.02196268537603516
Delaware	1947528	5371	0.2757855086037274	187	0.009601915864624283
Illinois	25343642	65889	0.2599823655968625	2843	0.011217803660578854
Maryland	12091360	27117	0.22426757618663243	1290	0.010668775059215835

2. Top 10 counties that are hardest hit when the cases are scaled with their population and their corresponding death rates
   Result:

county	county_population	cases	case_rates_percent	deaths	death_rates_percent
Harrisonburg city	53016	525	0.9902670891806248	19	0.03583823751320356
Emporia city	5346	46	0.8604564160119716	3	0.05611672278338946
District of Columbia	705749	5322	0.7540924606340215	264	0.03740706681837311
Manassas city	41085	289	0.7034197395643179	1	0.0024339783375927956
Alexandria city	159428	983	0.616579270893444	26	0.016308302180294554
Manassas Park city	17478	96	0.5492619292825266	2	0.011442956860052637
Baltimore city	593490	2609	0.439603026167248	124	0.02089335961852769
St. Louis city	300576	1315	0.43749334610880447	78	0.025950175662727563
Galax city	6347	25	0.39388687568930203	0	0
Colonial Heights city	17370	58	0.333909038572251	6	0.03454231433506045

3. County with earliest recorded case
   Result: Snohomish

4. Counties with no new recorded cases
   Result:

county	date_of_latest_record	cases	deaths
Fergus	2020-05-04T00:00:00.000+0000	1	0
Covington city	2020-04-28T00:00:00.000+0000	1	0
Arthur	2020-04-25T00:00:00.000+0000	1	0
Loup	2020-04-25T00:00:00.000+0000	1	0
Washita	2020-04-24T00:00:00.000+0000	1	0
Brule	2020-04-17T00:00:00.000+0000	1	0
Boundary	2020-04-12T00:00:00.000+0000	1	0
Ontonagon	2020-04-07T00:00:00.000+0000	1	0
Kingsbury	2020-04-01T00:00:00.000+0000	1	0

5. Top 10 counties with longest duration of covid
   Result:

county	date_of_latest_record	date_of_earliest_record	days_of_covid
Snohomish	2020-05-05T00:00:00.000+0000	2020-01-21T00:00:00.000+0000	105
Cook	2020-05-05T00:00:00.000+0000	2020-01-24T00:00:00.000+0000	102
Orange	2020-05-05T00:00:00.000+0000	2020-01-25T00:00:00.000+0000	101
Los Angeles	2020-05-05T00:00:00.000+0000	2020-01-26T00:00:00.000+0000	100
Maricopa	2020-05-05T00:00:00.000+0000	2020-01-26T00:00:00.000+0000	100
Santa Clara	2020-05-05T00:00:00.000+0000	2020-01-31T00:00:00.000+0000	95
Suffolk	2020-05-05T00:00:00.000+0000	2020-02-01T00:00:00.000+0000	94
San Francisco	2020-05-05T00:00:00.000+0000	2020-02-02T00:00:00.000+0000	93
Dane	2020-05-05T00:00:00.000+0000	2020-02-05T00:00:00.000+0000	90
San Diego	2020-05-05T00:00:00.000+0000	2020-02-10T00:00:00.000+0000	85

6. Top 10 counties with quickest containment
   Result:

county	date_of_latest_record	date_of_earliest_record	days_of_covid
Arthur	2020-04-25T00:00:00.000+0000	2020-04-25T00:00:00.000+0000	0
Boundary	2020-04-12T00:00:00.000+0000	2020-04-12T00:00:00.000+0000	0
Kingsbury	2020-04-01T00:00:00.000+0000	2020-03-31T00:00:00.000+0000	1
Ontonagon	2020-04-07T00:00:00.000+0000	2020-04-05T00:00:00.000+0000	2
Fergus	2020-05-04T00:00:00.000+0000	2020-04-30T00:00:00.000+0000	4
Loup	2020-04-25T00:00:00.000+0000	2020-04-20T00:00:00.000+0000	5
Brule	2020-04-17T00:00:00.000+0000	2020-04-12T00:00:00.000+0000	5
Washita	2020-04-24T00:00:00.000+0000	2020-04-03T00:00:00.000+0000	21
Covington city	2020-04-28T00:00:00.000+0000	2020-04-01T00:00:00.000+0000	27

Hospital resource utilization Insights

1. Potential increase in bed capacity in counties with longest duration of covid
   Result:

county	additional_county_capacity
Dane	341
Snohomish	130
Orange	638
Los Angeles	1788
San Diego	637
San Francisco	271
Cook	3217
Santa Clara	314
Maricopa	1833
Suffolk	1037

2. Counties where number of cumulative cases has exceeded number of licensed beds available in that county
   Subset of Result:

county_name	licensed_beds	cases
Chambers	129	304
Chilton	30	60
Tallapoosa	127	303
Wilcox	30	72
Apache	162	576
Coconino	418	561
Navajo	188	869
Pinal	401	478
Cleburne	25	72
Crittenden	113	197

3. Counties with potential to increase hospital beds in a particular state which also has counties where number of cumulative cases has exceeded number of licensed beds available
   Subset of Result:

state	county_in_need	county_with_potential	requirement	additional_county_capacity
New York	Nassau	Jefferson	32332	2998
New York	Nassau	Montgomery	32332	1661
New York	Nassau	Kings	32332	1549
New York	Nassau	Richmond	32332	1505
New York	Nassau	New York	32332	1387
New York	Nassau	Wayne	32332	1155
New York	Nassau	Suffolk	32332	1037
New York	Nassau	Essex	32332	1032
New York	Nassau	Fulton	32332	943
New York	Nassau	Orleans	32332	730

References

1. https://github.com/nytimes/covid-19-data
2. https://github.com/databricks/tech-talks/tree/master/2020-04-29%20%7C%20Intro%20to%20Apache%20Spark
3. https://databricks.com/blog/2020/04/14/covid-19-datasets-now-available-on-databricks.html
4. https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals/
5. https://coronavirus-resources.esri.com/datasets/definitivehc::definitive-healthcare-usa-hospital-beds?geometry=92.988%2C-16.820%2C-117.950%2C72.123

Appendix

A1. Link to databricks notebook -

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/6384376436740792/2897288349283662/5248716443437421/latest.html

A2. Remainder of Results -

A2.1-https://docs.google.com/spreadsheets/d/1Ufw-4se1ZXHq0Vg59qEkZklQ2ZRO4gsY_t5v04Hnx5c/edit?usp=sharing

A2.2-https://docs.google.com/spreadsheets/d/13FWPXkAcvluNaBnnBil_EwmAWB0JBO4Vd2Da-vM4H30/edit?usp=sharing