April 2020

Shefali Emmanuel Final Data Science Project:

OBTAINING DATA

Source 1 - AWS US Covd19 DataSet https://dj2taa9i652rf.cloudfront.net

```
[/data/FinalProject]$ curl https://covid19-lake.s3.us-east-2.amazonaws.com/static-datasets/csv/state-abv/states_abv.csv > stateABV.csv
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   665  100   665    0     0   1878      0 --:--:-- --:--:-- --:--:--  1878
[/data/FinalProject]$ ls
stateABV.csv
```

Source 2 - NY Times Github
https://github.com/nytimes/covid-19-data/blob/master/us-states.csv

```
[/data/FinalProject]$ curl https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv > usStates.csv
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 64445  100 64445    0     0   184k      0 --:--:-- --:--:-- --:--:--  184k
[/data/FinalProject]$ ls
stateABV.csv  usStates.csv
```

April 2020

SCRUBBING DATA

Technique 1: Deleting Uninteresting Columns

I removed the fips column as it has no use to me from the usStates.csv file.

```
[[/data/FinalProject]$ csvcut -c date,state,cases,deaths usStates.csv > newUSStates.csv
[[/data/FinalProject]$ ls
 newUSStates.csv  stateABV.csv  usStates.csv
```

Technique 2: Join Multiple CSV Files Horizontally on State Column

```
[[/data/FinalProject]$ csvjoin -c state newUSStates.csv stateABV.csv > finalDataset.csv
```

```
[[/data/FinalProject]$ csvjoin -c state newUSStates.csv stateABV.csv |csvlook
|       date | state                |   cases | deaths | abbreviation |
| ---------- | -------------------- | ------- | ------ | ------------ |
| 2020-01-21 | Washington           |       1 |      0 | WA           |
| 2020-01-22 | Washington           |       1 |      0 | WA           |
| 2020-01-23 | Washington           |       1 |      0 | WA           |
| 2020-01-24 | Illinois             |       1 |      0 | IL           |
| 2020-01-24 | Washington           |       1 |      0 | WA           |
| 2020-01-25 | California           |       1 |      0 | CA           |
| 2020-01-25 | Illinois             |       1 |      0 | IL           |
| 2020-01-25 | Washington           |       1 |      0 | WA           |
```

Technique 3: Remove all '-' from the date column

I did this inside of a Juypter notebook.

In [25]:
```
text = open("finalDataset.csv", "r")
text = ''.join([i for i in text]) \
    .replace("-", "")
x = open("finalDataset.csv","w")
x.writelines(text)
x.close()

df.head()
```

Out[25]:

	date	state	cases	deaths	abbreviation
0	20200121	Washington	1	0	WA
1	20200122	Washington	1	0	WA
2	20200123	Washington	1	0	WA
3	20200124	Illinois	1	0	IL
4	20200124	Washington	1	0	WA

April 2020

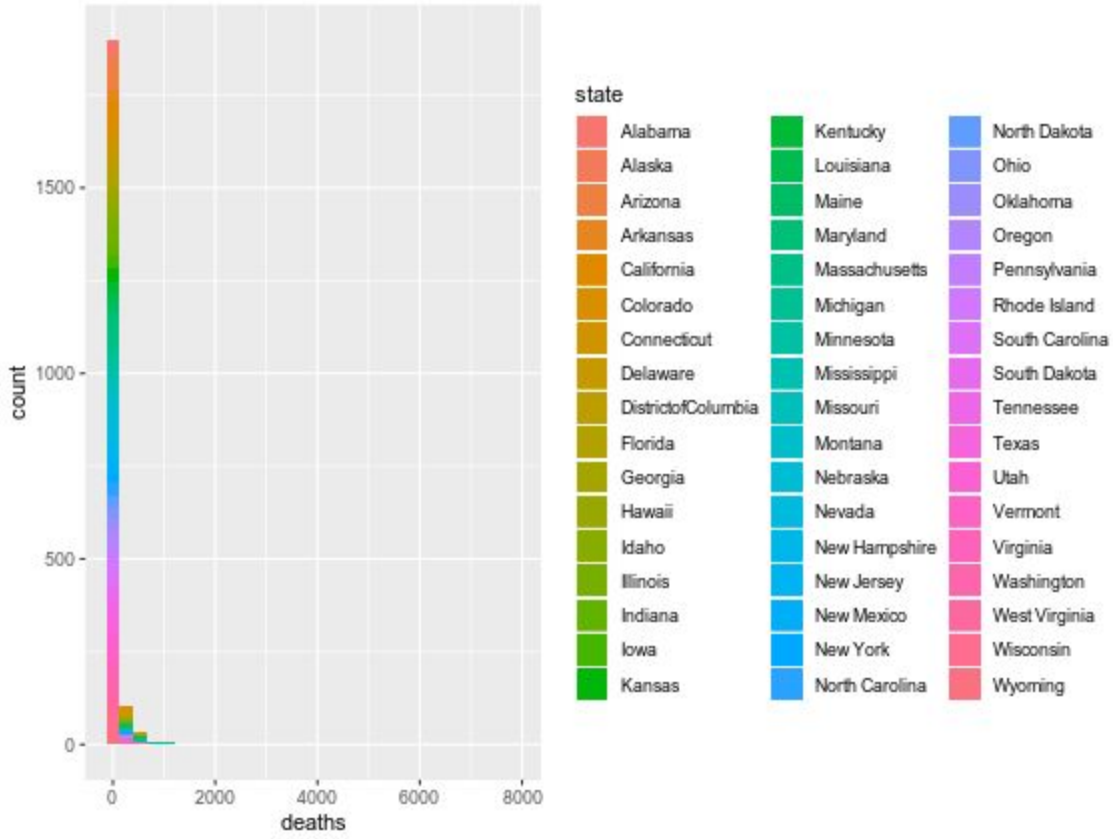
EXPLORING DATA

Technique 1: Derive statistics from the data by utilizing CSVSTAT

```
[[/data/FinalProject]$ csvstat finalDataset.csv --unique
  1. date: 81
  2. state: 51
  3. cases: 980
  4. deaths: 264
  5. abbreviation: 51
[[/data/FinalProject]$ csvstat finalDataset.csv --nulls
  1. date: False
  2. state: False
  3. cases: False
  4. deaths: False
  5. abbreviation: False
[[/data/FinalProject]$ csvstat finalDataset.csv --freq
  1. date: { "2020-03-17": 51, "2020-03-18": 51, "2020-03-19": 51, "2020-03-20": 51, "2020-03-21": 51 }
  2. state: { "Washington": 81, "Illinois": 78, "California": 77, "Arizona": 76, "Massachusetts": 70 }
  3. cases: { "1": 214, "2": 107, "3": 32, "6": 31, "4": 31 }
  4. deaths: { "0": 911, "1": 134, "2": 85, "3": 67, "5": 46 }
  5. abbreviation: { "WA": 81, "IL": 78, "CA": 77, "AZ": 76, "MA": 70 }
```

Technique 2: Create interesting visualization

```
[[/data/FinalProject]$ < finalDataset.csv Rio -ge 'g+geom_histogram(aes(deaths, fill=state))' > deathsNstates.png
 /usr/bin/Rio: line 128: warning: command substitution: ignored null byte in input
```



Please look at 'Lorenz.ipynb' that is also in this repository for more information.

April 2020

INTERPRETING DATA RESULTS

```
[/data]$ csvsql --query "SELECT * FROM finalDataset GROUP BY state ORDER BY deaths DESC" finalDataset.csv
date,state,cases,deaths,abbreviation
2020-04-10,New York,170512,7844,NY
2020-04-10,New Jersey,54588,1932,NJ
2020-04-10,Michigan,22646,1280,MI
2020-04-10,Louisiana,19253,755,LA
2020-04-10,Illinois,17887,607,IL
2020-04-10,Massachusetts,20974,599,MA
2020-04-10,California,21366,594,CA
2020-04-10,Washington,9887,483,WA
2020-04-10,Connecticut,10538,448,CT
2020-04-10,Pennsylvania,20128,435,PA
2020-04-10,Georgia,11859,425,GA
2020-04-10,Florida,17960,418,FL
2020-04-10,Indiana,6907,300,IN
2020-04-10,Colorado,6510,253,CO
2020-04-10,Texas,12288,247,TX
2020-04-10,Ohio,5878,231,OH
2020-04-10,Maryland,6968,172,MD
2020-04-10,Wisconsin,3068,131,WI
2020-04-10,Virginia,4509,121,VA
2020-04-10,Missouri,3799,105,MO
2020-04-10,Tennessee,4793,104,TN
2020-04-10,Arizona,3112,97,AZ
2020-04-10,Kentucky,1702,90,KY
2020-04-10,Oklahoma,1794,88,OK
2020-04-10,Nevada,2606,86,NV
2020-04-10,Mississippi,2469,82,MS
2020-04-10,Alabama,3008,80,AL
2020-04-10,North Carolina,3906,78,NC
2020-04-10,South Carolina,3065,72,SC
2020-04-10,Minnesota,1335,57,MN
2020-04-10,Kansas,1180,50,KS
2020-04-10,Rhode Island,2015,49,RI
2020-04-10,Oregon,1371,48,OR
2020-04-10,DistrictofColumbia,1660,38,DC
2020-04-10,Delaware,1326,32,DE
2020-04-10,Iowa,1388,31,IA
2020-04-10,Idaho,1397,25,ID
2020-04-10,Arkansas,1202,24,AR
2020-04-10,Vermont,679,24,VT
2020-04-10,New Hampshire,885,22,NH
2020-04-10,New Mexico,1091,19,NM
2020-04-10,Nebraska,679,18,NE
2020-04-10,Maine,586,17,ME
2020-04-10,Utah,2103,17,UT
2020-04-10,Hawaii,463,8,HI
2020-04-10,North Dakota,278,7,ND
2020-04-10,South Dakota,536,7,SD
2020-04-10,Montana,365,6,MT
2020-04-10,Alaska,244,5,AK
2020-04-10,West Virginia,537,5,WV
2020-04-10,Wyoming,253,0,WY
[/data]$
```

After trying various different combinations of query statements, this one gave me the most meaningful statistics. It showed the final number of cases by April 10, 2020 and sorted the list of states by which one had the most deaths. This shows me that NewYork surpassed New Jersey by 115,924 cases.

```
[[/data]$ csvsql --query "SELECT AVG(deaths),state FROM finalDataset GROUP BY state ORDER BY AVG(deaths) DESC" finalDataset.csv
AVG(deaths),state
1299.9756097560976,New York
311.63157894736844,New Jersey
257.5,Michigan
187.3030303030303,Louisiana
87.525,Georgia
76.82352941176471,Connecticut
74.17283950617283,Washington
72.46341463414635,Florida
64.9090909090909,California
64.75,Pennsylvania
51.270270270270274,Colorado
49.888888888888886,Indiana
49.18181818181818,Ohio
48.705128205128204,Illinois
48.67142857142857,Massachusetts
30.06779661016949,Texas
25.62162162162162,Maryland
23.285714285714285,Virginia
19.655172413793103,Alabama
19.542857142857144,Missouri
18.756756756756758,Nevada
18.64516129032258,Mississippi
18.22222222222222,Oklahoma
18.0,Tennessee
17.25,Kentucky
16.694444444444443,South Carolina
13.909090909090908,Wisconsin
11.512820512820513,North Carolina
10.194444444444445,Minnesota
10.13953488372093,Oregon
9.868421052631579,Arizona
9.514285714285714,Kansas
8.6,Vermont
7.6571428571428575,DistrictofColumbia
6.804878048780488,Rhode Island
6.647058823529412,Iowa
6.483870967741935,Delaware
6.354838709677419,Arkansas
6.137931034482759,Idaho
4.483870967741935,New Mexico
4.166666666666667,Maine
3.35,New Hampshire
2.739130434782609,Utah
2.3793103448275863,Montana
2.1296296296296298,Nebraska
1.9375,South Dakota
1.5161290322580645,North Dakota
1.44,West Virginia
1.4333333333333333,Alaska
1.1944444444444444,Hawaii
0.0,Wyoming
```

This screenshot showed the average death rate per state per day. New York has an average of 988 more deaths per day in comparison to the second highest death toll rate state of NJ.