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
```

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           |
```

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 }
```