HW1 - Hr2ee

Prework

Read the data file into R and store the dataset into the object Covid.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
setwd("/Users/maxryoo/Documents/MSDS/STAT6021/hw1")
Covid = read.csv("UScovid.csv", header = TRUE)
head(Covid)
```

```
##         date    county      state  fips cases deaths
## 1 2020-01-21 Snohomish Washington 53061     1      0
## 2 2020-01-22 Snohomish Washington 53061     1      0
## 3 2020-01-23 Snohomish Washington 53061     1      0
## 4 2020-01-24      Cook   Illinois 17031     1      0
## 5 2020-01-24 Snohomish Washington 53061     1      0
## 6 2020-01-25    Orange California  6059     1      0
```

Q1

A

We are interested in the data at the most recent date, June 3 2021. Create a data frame called latest that:

- has only rows pertaining to data from June 3 2021,
- removes rows pertaining to counties that are “Unknown”,
- removes the column date and fips,
- is ordered by county and then state alphabetically

Use the head() function to display the first 6 rows of the data frame latest

- has only rows pertaining to data from June 3, 2021

```
latest = Covid
latest = latest[latest$date == '2021-06-03',]
```

- removes rows pertaining to counties that are “Unknown”,

```
latest = latest[latest$county != "Unknown",]
```

- removes the column date and fips,

```
latest$date = NULL
latest$fips = NULL
```

- is ordered by county and then state alphabetically

```
latest = latest[order(latest$county, latest$state),]
head(latest)
```

```
##            county          state cases deaths
## 1383852 Abbeville South Carolina  2599     41
## 1382557    Acadia      Louisiana  6703    195
## 1384362  Accomack       Virginia  2862     43
## 1381993       Ada          Idaho 52964    475
## 1382232     Adair           Iowa   873     32
## 1382437     Adair       Kentucky  1944     54
```

B

Calculate the death rate (call it death.rate) for each county. Report the death rate as a percent and round to two decimal places. Add death.rate as a new column to the data frame latest. Display the first 6 rows of the data frame latest.

```
latest$death.rate = round((latest$deaths / latest$cases),2)
head(latest)
```

```
##            county          state cases deaths death.rate
## 1383852 Abbeville South Carolina  2599     41       0.02
## 1382557    Acadia      Louisiana  6703    195       0.03
## 1384362  Accomack       Virginia  2862     43       0.02
## 1381993       Ada          Idaho 52964    475       0.01
## 1382232     Adair           Iowa   873     32       0.04
## 1382437     Adair       Kentucky  1944     54       0.03
```

C

Display the counties with the 10 largest number of cases. Be sure to also display the number of deaths and death rates in these counties, as well as the state the counties belong to.

```
head(latest[order(-latest$cases), c("county", "state", "deaths", "death.rate", "cases")],10)
```

```
##                 county      state deaths death.rate   cases
## 1381641    Los Angeles California  24375       0.02 1245127
## 1383311  New York City   New York  33257       0.04  949986
## 1382052           Cook   Illinois  10893       0.02  554390
## 1381539       Maricopa    Arizona  10084       0.02  551509
## 1381801     Miami-Dade    Florida   6472       0.01  501925
## 1384160         Harris      Texas   6462       0.02  401345
## 1384116         Dallas      Texas   4082       0.01  303533
## 1381655      Riverside California   4614       0.02  300879
## 1381658 San Bernardino California   4760       0.02  298599
## 1381659      San Diego California   3760       0.01  280410
```

D

Display the counties with the 10 largest number of deaths. Be sure to also display the number of cases and death rates in these counties, as well as the state the counties belong to.

```
head(latest[order(-latest$deaths), c("county", "state", "deaths", "death.rate", "cases")],10)
```

```
##                 county      state deaths death.rate   cases
## 1383311  New York City   New York  33257       0.04  949986
## 1381641    Los Angeles California  24375       0.02 1245127
## 1382052           Cook   Illinois  10893       0.02  554390
## 1381539       Maricopa    Arizona  10084       0.02  551509
## 1381801     Miami-Dade    Florida   6472       0.01  501925
## 1384160         Harris      Texas   6462       0.02  401345
## 1381652         Orange California   5070       0.02  272242
## 1382761          Wayne   Michigan   5048       0.03  164612
## 1381658 San Bernardino California   4760       0.02  298599
## 1381655      Riverside California   4614       0.02  300879
```

E

Display the counties with the 10 highest death rates. Be sure to also display the number of cases and deaths in these counties, as well as the state the counties belong to. Is there sometime you notice about these counties?

```
head(latest[order(-latest$death.rate), c("county", "state", "deaths", "death.rate", "cases")],10)
```

```
##            county      state deaths death.rate cases
## 1383143     Grant   Nebraska      4       0.10    41
## 1384261    Sabine      Texas     45       0.09   524
## 1384137     Foard      Texas     10       0.08   124
## 1383261   Harding New Mexico      1       0.08    12
## 1383084 Petroleum    Montana      1       0.08    12
## 1384076    Borden      Texas      2       0.07    30
## 1381847   Candler    Georgia     67       0.07   978
## 1381888  Glascock    Georgia     19       0.07   269
## 1381896   Hancock    Georgia     68       0.07   928
## 1384232    Motley      Texas      8       0.07   116
```

An intersting finding for these counties was that the number of cases is not that high when compared to the other counties. Take for Instance the cases for L.A (1245127) compared to the highest death rate Grant, NE (41). Death Rate is by porportion and not the count. There can be a very high count of covid cases in an area that may not neccesarily mean that it has a high death rate. There could be some other factors that are contributing to deathrate. A possible hypothesis could be the vacine rate. It seems like all the counties shown above are part of "red" states, which in the news there have been lots of talks about low vaccination rates. This could be a contributing factor that could be later tested, does vacination rate have a big role in death rate.

F

Display the counties with the 10 highest death rates among counties with at least 100,000 cases. Be sure to also display the number of cases and deaths in these counties, as well as the state the counties belong to.

```
atleast = latest[latest$cases >= 100000, ]
head(atleast[order(-atleast$death.rate),  c("county", "state", "deaths", "death.rate", "cases")],10)
```

```
##                county        state deaths death.rate  cases
## 1383311 New York City     New York  33257       0.04 949986
## 1383229        Bergen   New Jersey   2868       0.03 104301
```

```
## 1382672     Middlesex Massachusetts   3761       0.03 134980
## 1382761         Wayne      Michigan   5048       0.03 164612
## 1383701     Allegheny  Pennsylvania   1985       0.02 101411
## 1384074         Bexar         Texas   3577       0.02 224096
## 1383202         Clark        Nevada   4419       0.02 252137
## 1382052          Cook      Illinois  10893       0.02 554390
## 1383514      Cuyahoga          Ohio   2183       0.02 115137
## 1384129       El Paso         Texas   2719       0.02 136182
```

G

Display the number of cases, deaths, death rate for the following counties:

- Albemarle, Virginia

```
latest[(latest$state == "Virginia") & (latest$county == "Albemarle"),]
```

```
##            county    state cases deaths death.rate
## 1384363 Albemarle Virginia  5801     83       0.01
```

- Charlottesville City, Virginia

```
latest[(latest$state == "Virginia") & (latest$county == "Charlottesville city"),]
```

```
##                       county    state cases deaths death.rate
## 1384385 Charlottesville city Virginia  4014     57       0.01
```

Q2

For this question, we focus on data at the state level. Note that the dataset has data on the 50 states ,plus DC, Puerto Rico, Guam, Northern Mariana Islands, and the Virgin Islands.

A

We are interested in the data at the most recent date, June 3 2021. Create a data frame called state.level that:

- has 55 rows: 1 for each state, DC, and territory
- has 3 columns: name of the state, number of cases, number of deaths
- is ordered alphabetically by name of the state

Display the first 6 rows of the data frame state.level.

1st Step is to filter the data for dates after 2020-06-02 2nd Step is to drop the counties, date, fips since they are irrelavent 3rd Step is to group by and add the data since the columns are all numeric 4th Step is to order alphabetically by the name of the state

```
# Step 1
state.level = Covid
state.level = state.level[state.level$date == '2021-06-03',]

# Step 2
state.level$county = NULL
state.level$fips = NULL
state.level$date = NULL

# Step 3
## Removed NA due to quality issues.
state.level = state.level %>% group_by(state) %>% summarise(across(everything(), sum,  na.rm=TRUE))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
state.level
```

```
## # A tibble: 55 x 3
##    state                cases deaths
##    <chr>                <int>  <int>
##  1 Alabama             545028  11188
##  2 Alaska               69826    352
##  3 Arizona             882691  17653
##  4 Arkansas            341889   5842
##  5 California         3793055  63345
##  6 Colorado            547961   6746
##  7 Connecticut         347748   8245
##  8 Delaware            108957   1668
##  9 District of Columbia 49041   1136
## 10 Florida            2329859  36972
## # ... with 45 more rows
```

```
# Step 4
state.level = state.level[order(state.level$state),]

# Show top 6
head(state.level)
```

```
## # A tibble: 6 x 3
##   state        cases deaths
##   <chr>        <int>  <int>
## 1 Alabama     545028  11188
## 2 Alaska       69826    352
## 3 Arizona     882691  17653
## 4 Arkansas    341889   5842
## 5 California 3793055  63345
## 6 Colorado    547961   6746
```

B

Calculate the death rate (call it state.rate) for each state. Report the death rate as a percent and round to two decimal places. Add state.rate as a new column to the data frame state.level. Display the first 6 rows of the data frame state.level.

```
state.level$state.rate = round((state.level$deaths / state.level$cases),2)
head(state.level)
```

```
## # A tibble: 6 x 4
##   state        cases deaths state.rate
##   <chr>        <int>  <int>      <dbl>
## 1 Alabama     545028  11188       0.02
## 2 Alaska       69826    352       0.01
## 3 Arizona     882691  17653       0.02
## 4 Arkansas    341889   5842       0.02
## 5 California 3793055  63345       0.02
## 6 Colorado    547961   6746       0.01
```

C

What is the death rate in Virginia?

```
state.level[state.level$state == "Virginia", c("state","state.rate")]
```

```
## # A tibble: 1 x 2
##   state    state.rate
##   <chr>         <dbl>
## 1 Virginia       0.02
```

D

What is the death rate in Puerto Rico?

```
state.level[state.level$state == "Puerto Rico", c("state","state.rate")]
```

```
## # A tibble: 1 x 2
##   state       state.rate
##   <chr>            <dbl>
## 1 Puerto Rico       0.01
```

The above output when investigated throroughly had some NA attributes for Deaths while the cases were >0. This is some odd data collection and should be brought back to the data collection team if possible, but the above calcualtions were done with removing NA deaths.

E

Which states have the 10 highest death rates?

```
head(state.level[order(-state.level$state.rate),],10)
```

```
## # A tibble: 10 x 4
##    state                  cases deaths state.rate
##    <chr>                  <int>  <int>      <dbl>
##  1 Massachusetts         707523  17893       0.03
##  2 New Jersey           1017044  26253       0.03
##  3 New York             2102003  52811       0.03
##  4 Alabama               545028  11188       0.02
##  5 Arizona               882691  17653       0.02
##  6 Arkansas              341889   5842       0.02
##  7 California           3793055  63345       0.02
##  8 Connecticut           347748   8245       0.02
##  9 Delaware              108957   1668       0.02
## 10 District of Columbia   49041   1136       0.02
```

F

Which states have the 10 lowest death rates.

```
head(state.level[order(state.level$state.rate),],10)
```

```
## # A tibble: 10 x 4
##    state            cases deaths state.rate
##    <chr>            <int>  <int>      <dbl>
##  1 Alaska           69826    352       0.01
##  2 Colorado        547961   6746       0.01
##  3 Hawaii           35152    498       0.01
##  4 Idaho           192704   2103       0.01
##  5 Maine            67986    837       0.01
##  6 Minnesota       601974   7530       0.01
```

```
##  7 Montana         112236   1627       0.01
##  8 Nebraska        223517   2385       0.01
##  9 New Hampshire    98840   1354       0.01
## 10 North Carolina 1004699  13147       0.01
```

G

Export this dataset as a .csv file named stateCovid.csv. We will be using this file for the next homework.

```
write.csv(state.level, file="stateCovid", row.names = FALSE)
```