HW1

Dmitry Mikhaylow

9/7/2021

Loading and storing COVID data in R:

```
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
```

1. Country level analysis

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:

```
latest <- subset(Covid, date == "2021-06-03")
dim(latest)
```

```
## [1] 3247    6
```

- removes rows pertaining to counties that are “Unknown”

```
latest <- subset(latest, latest$county!='Unknown')
dim(latest)
```

```
## [1] 3221    6
```

- removes the column `date` and `fips`

```
latest <- subset(latest, select=-c(date, fips))
head(latest)
```

```
##          county   state cases deaths
## 1381437 Autauga Alabama  7172    111
## 1381438 Baldwin Alabama 21684    312
## 1381439 Barbour Alabama  2343     59
## 1381440    Bibb Alabama  2665     64
## 1381441  Blount Alabama  6894    139
## 1381442 Bullock Alabama  1236     42
```

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) * 100, 2)
head(latest)
```

```
##          county   state cases deaths death.rate
## 1381437 Autauga Alabama  7172    111       1.55
## 1381438 Baldwin Alabama 21684    312       1.44
## 1381439 Barbour Alabama  2343     59       2.52
## 1381440    Bibb Alabama  2665     64       2.40
## 1381441  Blount Alabama  6894    139       2.02
## 1381442 Bullock Alabama  1236     42       3.40
```

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, decreasing = TRUE),], 10)
```

```
##                county      state   cases deaths death.rate
## 1381641    Los Angeles California 1245127  24375       1.96
## 1383311  New York City   New York  949986  33257       3.50
## 1382052           Cook   Illinois  554390  10893       1.96
## 1381539       Maricopa    Arizona  551509  10084       1.83
## 1381801     Miami-Dade    Florida  501925   6472       1.29
## 1384160         Harris      Texas  401345   6462       1.61
## 1384116         Dallas      Texas  303533   4082       1.34
## 1381655      Riverside California  300879   4614       1.53
## 1381658 San Bernardino California  298599   4760       1.59
## 1381659      San Diego California  280410   3760       1.34
```

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, decreasing = TRUE),], 10)
```

```
##                county      state   cases deaths death.rate
## 1383311  New York City   New York  949986  33257       3.50
## 1381641    Los Angeles California 1245127  24375       1.96
## 1382052           Cook   Illinois  554390  10893       1.96
## 1381539       Maricopa    Arizona  551509  10084       1.83
## 1381801     Miami-Dade    Florida  501925   6472       1.29
## 1384160         Harris      Texas  401345   6462       1.61
## 1381652         Orange California  272242   5070       1.86
## 1382761          Wayne   Michigan  164612   5048       3.07
## 1381658 San Bernardino California  298599   4760       1.59
## 1381655      Riverside California  300879   4614       1.53
```

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?

Note: death rates are high but the actual number of cases is relatively low:

```
head(latest[order(latest$death.rate, decreasing = TRUE),], 10)
```

```
##               county      state cases deaths death.rate
## 1383143        Grant   Nebraska    41      4       9.76
## 1384261       Sabine      Texas   524     45       8.59
## 1383084    Petroleum    Montana    12      1       8.33
## 1383261      Harding New Mexico    12      1       8.33
## 1384137        Foard      Texas   124     10       8.06
## 1381896      Hancock    Georgia   928     68       7.33
## 1381888     Glascock    Georgia   269     19       7.06
## 1384232       Motley      Texas   116      8       6.90
## 1381847      Candler    Georgia   978     67       6.85
## 1384283 Throckmorton      Texas    73      5       6.85
```

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.

```
head(latest[order(c(latest$cases, latest$death.rate), decreasing = TRUE),], 10)
```

```
##                county      state   cases deaths death.rate
## 1381641    Los Angeles California 1245127  24375       1.96
## 1383311  New York City   New York  949986  33257       3.50
## 1382052           Cook   Illinois  554390  10893       1.96
## 1381539       Maricopa    Arizona  551509  10084       1.83
## 1381801     Miami-Dade    Florida  501925   6472       1.29
## 1384160         Harris      Texas  401345   6462       1.61
## 1384116         Dallas      Texas  303533   4082       1.34
## 1381655      Riverside California  300879   4614       1.53
## 1381658 San Bernardino California  298599   4760       1.59
## 1381659      San Diego California  280410   3760       1.34
```

Another way to do the same:

```
head(latest[order(-latest$cases, latest$death.rate), ], 10)
```

```
##                 county      state   cases deaths death.rate
## 1381641    Los Angeles California 1245127  24375       1.96
## 1383311  New York City   New York  949986  33257       3.50
## 1382052           Cook   Illinois  554390  10893       1.96
## 1381539       Maricopa    Arizona  551509  10084       1.83
## 1381801     Miami-Dade    Florida  501925   6472       1.29
## 1384160         Harris      Texas  401345   6462       1.61
## 1384116         Dallas      Texas  303533   4082       1.34
## 1381655      Riverside California  300879   4614       1.53
## 1381658 San Bernardino California  298599   4760       1.59
## 1381659      San Diego California  280410   3760       1.34
```

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

- Albemarle, Virginia

```
latest[which(latest$county=="Albemarle"), c(3, 4, 5)]
```

```
##         cases deaths death.rate
## 1384363  5801     83       1.43
```

- Charlottesville city, Virginia

```
latest[which(latest$county=="Charlottesville city"), c(3, 4, 5)]
```

```
##         cases deaths death.rate
## 1384385  4014     57       1.42
```

2. State level analysis

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

```
state.level <- subset(Covid, date == "2021-06-03")
state.level <- subset(state.level, select=-c(date, county, fips))
```

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

```
dim(state.level)
```

```
## [1] 3247    3
```

- is ordered alphabetically by name of the state
- Display the first 6 rows of the data frame state.level.

```
state.level <- aggregate(cbind(cases, deaths)~state, data=state.level, FUN=sum)
head(state.level)
```

```
##        state   cases deaths
## 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) * 100, 2)
head(state.level)
```

```
##        state   cases deaths state_rate
## 1    Alabama  545028  11188       2.05
## 2     Alaska   69826    352       0.50
## 3    Arizona  882691  17653       2.00
## 4   Arkansas  341889   5842       1.71
## 5 California 3793055  63345       1.67
## 6   Colorado  547961   6746       1.23
```

c. What is the death rate in Virginia?

```
state.level[which(state.level$state=="Virginia"), c(1,4)]
```

```
##       state state_rate
## 51 Virginia       1.66
```

d. What is the death rate in Puerto Rico?

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

```
##          state cases deaths state_rate
## 42 Puerto Rico  5589   2512      44.95
```

Note: PR has extrememly high state.rate due to low number of reported cases.

e. Which states have the 10 highest death rates?

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

```
##                   state   cases deaths state_rate
## 42          Puerto Rico    5589   2512      44.95
## 32           New Jersey 1017044  26253       2.58
## 23        Massachusetts  707523  17893       2.53
## 34             New York 2102003  52811       2.51
## 7           Connecticut  347748   8245       2.37
## 9  District of Columbia   49041   1136       2.32
## 26          Mississippi  318048   7324       2.30
## 41         Pennsylvania 1208879  27349       2.26
## 20            Louisiana  472617  10605       2.24
## 33           New Mexico  203330   4275       2.10
```

f. Which states have the 10 lowest death rates?

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

```
##                       state  cases deaths state_rate
## 2                    Alaska  69826    352       0.50
## 48                     Utah 406895   2308       0.57
## 50           Virgin Islands   3512     28       0.80
## 49                  Vermont  24240    255       1.05
## 29                 Nebraska 223517   2385       1.07
## 14                    Idaho 192704   2103       1.09
## 37 Northern Mariana Islands    183      2       1.09
## 54                Wisconsin 675152   7923       1.17
## 55                  Wyoming  60543    720       1.19
## 6                  Colorado 547961   6746       1.23
```

g. Export this dataset as a .csv file named stateCovid.csv.

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