Stat 6021: Homework Set 1

Download the dataset `UScovid.csv` from Canvas. The dataset was released by The New York Times and contains data on cumulative (accruing) counts of coronavirus cases and deaths in the United States, at the state and county level, over each day from Jan 21, 2020 to June 3 2021. You may read more about the data and the variable descriptions here (please note the dataset is regularly updated, we will use the file on Canvas).

1. For this question, we focus on data at the county level.

   (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`.

   (b) Calculate the case fatality rate (number of deaths divided by number of cases, and call it `death.rate`) for each county. Report the case fatality 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`.

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

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

   (e) Display the counties with the 10 highest case fatality 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?

   (f) Display the counties with the 10 highest case fatality 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.

(g) Display the number of cases, deaths, case fatality rates for the following counties:

i. Albemarle, Virginia

ii. Charlottesville city, Virginia

2. 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. For the purpose of this question, we will consider DC, Puerto Rico, Guam, Northern Mariana Islands, and the Virgin Islands, as "states" as well.

(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`.

(b) Calculate the case fatality rate (call it `state.rate`) for each state. Report the case fatality 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`.

(c) What is the case fatality rate in Virginia?

(d) What is the case fatality rate in Puerto Rico?

(e) Which states have the 10 highest case fatality rate?

(f) Which states have the 10 lowest case fatality rate?

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