|  |  |
| --- | --- |
| **Covid Assignment** | |
|  |  |
| General Issues:   1. Please include both R-code and output in your answers. 2. Please keep you output to the minimal but sufficient to answer the questions 3. Please shrink your plots without compromising their intelligibility 4. Please include commonly used R functions such as “select,” “rename,” “recode” etc. even when it is not part of the question but necessary for you to obtain answers. 5. There is no need to elaborate on social, economic, political, or cultural implication of your findings. Please keep your answers short and relevant. | |

This project is based on the data set named “causes\_of\_death\_19\_20.csv.” It contains information on the causes of death in 2019 and 2020 (up to August 15th, 2020) in the U.S. by the week and the state. However, the “jurisdiction of occurrence” variable contains a few categories that are not states; ex. New York City, DC, and the United States. Moreover, there are two variables that measure the number of covid-19 deaths as indicated in the table below. We will focus on the one named “COVID-19 (U071, Multiple Cause of Death).”

Here is a description of the variables included:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Jurisdiction of Occurrence | Jurisdiction of Occurrence |
| MMWR Year | MMWR Year (MMWR stands for Morbidity and Mortality Weekly Report from CDC) |
| MMWR Week | MMWR Week |
| Week Ending Date | Week Ending Date |
| All Cause | All Cause |
| Natural Cause | Natural Cause |
| Septicemia (A40-A41) | Septicemia (A40-A41) |
| Malignant neoplasms (C00-C97) | Malignant neoplasms (C00-C97) |
| Diabetes mellitus (E10-E14) | Diabetes mellitus (E10-E14) |
| Alzheimer disease (G30) | Alzheimer disease (G30) |
| Influenza and pneumonia (J09-J18) | Influenza and pneumonia (J09-J18) |
| Chronic lower respiratory diseases (J40-J47) | Chronic lower respiratory diseases (J40-J47) |
| Other diseases of respiratory system (J00-J06,J30-J39,J67,J70-J98) | Other diseases of respiratory system (J00-J06,J30-J39,J67,J70-J98) |
| Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27) | Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27) |
| Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99) | Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99) |
| Diseases of heart (I00-I09,I11,I13,I20-I51) | Diseases of heart (I00-I09,I11,I13,I20-I51) |
| Cerebrovascular diseases (I60-I69) | Cerebrovascular diseases (I60-I69) |
| COVID-19 (U071, Multiple Cause of Death) | COVID-19 (U071, Multiple Cause of Death) |
| COVID-19 (U071, Underlying Cause of Death) | COVID-19 (U071, Underlying Cause of Death) |

1. Create a data frame in R to include only information on the state, year, week, week ending day, deaths of all causes, and the variable “COVID-19 (U071, Multiple Cause of Death). (2 points)
2. Rename variables and recode some variables as you see fit. (2 points) (Hint: Most of the variables have long names)
3. Convert the “week ending date” variable into a “Date” data type. Check if you have succeeded in the conversion. (1 points) (Hint: the format you need to indicate as an argument in one of the variations of “mm-dd-yyyy” is not the format you intend to have but the format in the original data).
4. (5 points)
5. Create a subset that contains only information of 2020 and includes only variables of the state, the week, and "covid-19 Multiple Cause of Death."
6. Then take a further step to only select the states (city) of Arizona, California, Florida, New Jersey, New York, New York City, and Texas. (NYC and DC are included as non-state jurisdiction)
7. As a last step, "spread" the data frame by week (making "week" into columns) on the values of the variable "covid-19 Multiple Cause of Death ".
8. At[[1]](#footnote-1) which week did covid-19 deaths started to occur in most of the 6 states and NYC?
9. (6 points)
10. Use the data frame from step 4 (the one that contains only 2020 information) to plot the number of covid-19 deaths (Multiple Cause of Death) by the week; and add a dimension of "state" by color.
11. Which state (or city) saw deaths peak the earliest? At which week?
12. What state(s) (or city) saw a second wave of death? During which weeks?
13. (7 points)
14. (Use the data frame from step 5) Use the "summarise" function in "dplyr" to create a data set that contains only “the total number of death of covid (with multiple causes)” and the state (city) name. Then "merge" the new data frame with "population\_by\_state\_2019.csv" (available on Canvas) to combine the death toll with information on state population. You should end up with a data frame containing 3 variables: state, total covid death, and state population.
15. Obtain a bar-chart to see which state has the highest covid death toll.
16. Calculate a new variable of “percentage of covid death in the total population” and save it to the new (small) data frame (Hint: you can use either the mutate function or simple base R function).
17. Please plot the percentage you obtain in the previous step by the state. Which state has the highest percentage of covid death in the total population?
18. (7 points)
19. Formulate a data frame for 2020 that contains only
20. the state,
21. covid death (Multiple Cause of Death),
22. and deaths of all causes (basically the number of all people who died in the first 33 weeks in 2020)

This data frame should contain 3 variables but all cases.

1. Then use the "summarise" function to create a smaller data frame with aggregated information that contains
2. total covid deaths and
3. total deaths of all causes
4. Next, reduce the data frame to include only the 6 states above and NYC, as we did in the previous questions.
5. Then create a variable: the percentage of covid deaths in total deaths; plot this percentage by state. Which state (city) has the highest percentage of covid deaths of all deaths?

1. Weeks are points in the data set with no duration. [↑](#footnote-ref-1)