Stat 604
Assignment 05 R

Scope:

This assignment reinforces the concepts covered in Lessons R01 through R06. NOTE: You may need to combine several functions and/or expressions together in a single command to accomplish some of the objectives of this assignment.

Specific Instructions for this Assignment:

There is a file named **HW5 Data Files.zip** included on the module with this assignment in Canvas. Download the **HW5 Data Files.zip** file from Canvas and extract it contents to an easily accessible folder on your computer where you stored the workspace downloaded with the previous assignment. You MUST extract these two files in order to avoid problems with R being able to successfully read them. The **csv** file is a large raw text file that contains daily statistics on COVID-19 cases and deaths from around the world. The **txt** file is also a delimited file containing geographic and population data on the locations that report Covid statistics. The cardinal rule for data analysis is "Know thy data." Open the **COVID Activity.csv** file with your favorite text editing software (it may be too large for Windows Notepad) and look carefully at its contents so you can get an idea of what information is available and how clean it is, etc. before analyzing it with R. For Windows users, Notepad++ is a more robust editor than the Notepad program that is delivered with Windows. You can download Notepad++ for free from cnet.com. If you are using a Mac you can get Text Wrangler from the App Store for similar functionality. Your computer will probably want to open the file with Excel. If you open it in Excel DO NOT save the file when you exit or you risk altering the file so that it does not produce the results you want in R. It is OK to look at it in Excel, but Excel will mask some of the structure of the file from you, so be sure you also look at the file in a text editor. Use the same technique to become familiar with the **txt** file.

Perform in R each of the exercises listed below. Include a comment line in your script above the section for each step so that each is clearly identified.

1. Prepare the header and perform housekeeping steps as described in HW 4. You will NOT be using the sink function to save the output of this assignment.
2. Import the **COVID Activity.csv** file into an R data frame using the appropriate function. DO NOT include code to display the data frame upon creation as it will likely overload the console due to the amount of data.
    a. Show the structure of the new data frame.
    b. Some of the columns have very long names that could be shortened without any negative consequences. However, the column order has not always been consistent in the download of this data so we need to make the changes using a value replacement You can use the names function to access the column names as a vector that you can manipulate as you would any other vector. (Remember you are not actually changing anything unless you use an assignment statement.) Change the columns shown in the table below:

From	**To**
POSITIVE_CASES_COUNT	TOTAL_CASES
DEATH_NEW_COUNT	NEW_DEATHS

POSITIVE_NEW_CASES_COUNT	NEW_CASES
DEATH_COUNT	TOTAL_DEATHS

   c. Display the first 10 rows and all columns of the modified data frame.
3. Create a new data frame that is a subset of the data frame created from the CSV file. The subset will contain only rows for the state of Texas. Use a list of column numbers in your subscript so the new data frame contains only the following columns in the order shown: COUNTY_NAME, REPORT_DATE, NEW_CASES, TOTAL_CASES, NEW_DEATHS, TOTAL_DEATHS. Display in the console the structure of the new data frame.
4. Write an expression to import the **txt** file into a data frame. You may spread the expression across multiple lines in your script so it does not get cut off when you convert the script to pdf if you will insert your breaks between elements of the expression or function.
   a. Display the structure of the new data frame.
   b. Change the name of the column that contains population data to POPULATION to be more concise.
   c. Display the structure again showing the modifications.
   d. Display the first 10 rows of the modified data frame.
5. Create a new data frame by combining the "Texas" data frame with the "population" data frame that you created in the previous step. When the "population" data frame is referenced in your expression to combine the data frames, use expressions for the rows and columns so that only rows from Texas are selected and only the COUNTY_NAME and POPULATION columns. Include non-matches in the resulting data frame. The new data frame should have 153,255 rows.
   a. Display a summary of the new data frame.
   b. Display the first 50 rows of the new data frame.
6. Execute a function that will make the columns of the data frame available to R directly by column name to simplify coding in the modifications described below:
   a. Use a function to convert REPORT_DATE to an actual R date value and assign it to a new column in the data frame. Display a summary of the new date column. Note: You cannot refer to this column only by name because it did not exist when you executed the function to make the columns available.
   b. The COVID activity statistics are contained in four columns whose names were changed as instructed earlier in the assignment. Create four new columns in the data frame that represent each of the statistics as a percentage of the population of that county. This is done by dividing the original column by the POPULATION column. Include PCT in the names of your new columns to differentiate them from the originals. Leave the percentage values in their raw format of a value between 0 and 1. You will notice that some of the percentages are so small they are displayed in exponential notation.
   c. Display the structure of the updated data frame and its first 20 rows.
   d. Execute a function so that the column names of the data frame are no longer available in the R search path
7. Create and display a new data frame that is a subset of the data frame created in the previous step. Use a logical test to subset the rows to only those where the REPORT_DATE is the last available and POPULATION is not missing. Determine the last date value based on the summary of the Date column from the previous step. Hard code this value into your expression. Display the structure of the new data frame.
8. Use the colSums function to display the statewide totals of each of the columns containing the original Covid count statistics. Use the apply function to make the same calculation. Include an argument on your functions so that you will get a total even if there are missing values for some counties.

9. Using the last data frame created, display a list of County names, TOTAL_CASES, POPULATION, and percent of TOTAL_CASES, listed from the highest percentage to the lowest.
10. Display all data for counties whose names contain the letter V, ignoring case.
11. Display the contents of the workspace.
12. Remove everything from the workspace except the data frame created beginning in step 5 above and the data frame created in step 7. Display the contents of the workspace again.
13. Save the workspace in case we want to use it in the next assignment. Name it HW05.RData. You may save it initially using the R GUI but your script must contain code to save the workspace in case you submit the script again.
14. After you have debugged your program and successfully executed it in a new R session, use the information in your console to answer the questions below in comment lines at the bottom of your script:
    a. How many observations were loaded from the CSV file?
    b. How many observations and variables are in the data frame loaded from the **txt** file?
    c. What is one possible explanation for the minimum value of NEW_CASES shown in the summary from step 5a and what is your reaction to this value as an analyst?
    d. Explain the difference in the summaries of the two date columns. What are the minimum and maximum dates in the data frame?
    e. What is the total number of COVID cases and deaths in the state of Texas on the last date reported?
    f. What is the name and population of the county with the lowest percentage of cases as of the last date reported?
15. Convert your script and console to PDF and submit them to Canvas.