Stat 604
Assignment 06 R

Scope:

This assignment reinforces the material covered in the lectures through R Lesson 8.

Specific Instructions for this Assignment:

This assignment will use the COVID19 data frames created and saved in the workspace of the previous assignment. If you were unable to successfully create that workspace, you may use the professor's workspace which will be made available on Canvas after the deadline for that assignment.

In addition to the console and script PDF files like you submitted in the previous assignment, this assignment will require a PDF file containing the graphics output of this assignment. (Name it with the pattern FKincheloe_HW06_graphics.pdf using your own First Initial and Last Name.)
Perform in R each of the exercises listed below. Include the required header information and a comment line in your script above the section for each step so that each is clearly identified. The steps in this assignment may require multiple lines of R code to obtain the desired results. Some of the expressions will be quite long. Please wrap these expressions across multiple lines so they will not get cut off when you convert the script file to PDF.

Prior to starting your script, execute in the console the function that will display all the graphics parameters. Locate the parameter that defines the graph margin in inches. Write down the margin values so that you can refer to them later in the assignment.

1. After the header, include housekeeping steps as you did in the previous assignments.
2. Write an expression in your script to load the workspace from the previous assignment. Show the contents of the workspace. Display a summary of the data frame containing data as of September 12.
3. On an assignment statement, use the **with** function to access the columns in the September 12 data frame and create a new column containing the death rate. Death rate is calculated as Total Deaths divided by Total Cases then multiplied by 100 so it is displayed as a number between 0 and 100. This expression will be one of the arguments in the **with** function. Write expressions to show the minimum value and maximum value of the new column.
4. Use a line of code to direct all graphic output to your PDF document. Research the available arguments for this function and set width to 11 and height to 8.5 so it will fit a normal size paper in landscape orientation. (You may want to wait until you have your graphics working correctly before you add the line to redirect to PDF so you can see the results in your R session.)
5. Create a histogram of the death rate column you created above, forcing the cells to have a width of 0.5. Start the breaks at the minimum death rate and continue to the next integer above the maximum death rate. You may hard code the start and end values when setting up your break points. (The term "hard coding" refers to entering an actual value like 50 in your program code instead of using a formula.) Create the histogram in a manner that will facilitate the addition of a distribution curve later. Label the X axis "Percent" and supply an appropriate main title for the graph.

6. Add to the graph a line that shows the normal distribution density of death rate values. Include arguments that will ensure calculations are made even when there are missing values in the data. Use a hex value to "mix" a color for the line that has a Red amount of 22, a Green amount of A0 and a Blue amount of EE.
7. Draw a vertical line at the mean death rate value. Use the second color in the R palette as the color of the line. Use a function to determine the position of the line instead of hard coding the current mean value. Include an argument to ensure the mean is calculated even if there are missing values. Draw a line at the median in the same manner except use the color name green1 to specify the line color.
8. Display in the console the names of all available R colors.
9. We want to observe the correlation between the total number of cases and the total number of deaths from each county in the September 12 data. Plot a point for each county with data using total cases for the x axis and total deaths for the y axis. Use the diamond plot character (◊). Pick an unusual name that sounds interesting to you from the list of colors as the color of your points. Any color is acceptable if the points show up well. Supply appropriate labels for the axes and an appropriate title for the graph.
10. Add a fit line to the plot.
11. Use functions to imbed text showing the date and time of creation in the upper left-hand corner of the graph area. The exact value of the y coordinate for the time stamp location is not critical if the time stamp is near the corner. You may hard code the coordinates but use 0 as the x coordinate and use an alignment value so the text starts at 0. The date and time must automatically change each time the script is run.
12. Use logic expressions as an index parameter to create a new data frame that is a subset of the Texas COVID data frame where the population of the county is not missing and is greater than 500 thousand and the value of the date column created in the previous assignment is greater than March 14, 2020. When you hard code the date value in your comparison statement, coerce it to a date so you can be sure R is comparing two values of the Date class. Include all columns in the subset. Display a summary of the new data frame. Use the tapply function to display a table showing the median number of New Cases for each county in the data frame. There should be 12 Counties displayed and the value for Bexar should be 171.
13. Increase the bottom and left margins to be one-half of an inch larger than their default values recorded at the beginning of this assignment. Create a boxplot of the number of New Cases grouped by county using the data frame of large counties created in the previous step. Supply an appropriate Y axis label and a main title for the chart. Remove the X axis label by using two quotes with nothing inside them as the value for this label. The inside of the boxes is maroon. Supply an argument that will cause the whiskers of the plot to be 4 times the interquartile range. Add the argument las=2 to cause the county names to be displayed vertically.
14. When you have finished debugging and testing your script, run it in a fresh R session. Place the answers to the questions below in comments at the bottom of your script:
    a. What was the maximum new cases on September 12?
    b. Explain why you think the death rate graphed in step 5 is or is not normally distributed.
    c. Describe the relationship with the fit line and the plotted points.
    d. What is the mean Date in the summary of the data frame of large counties?
    e. What is the median number of new cases for El Paso County?
    f. What is the name of the county and the approximate value of the outlier for the county with the largest outlier?
15. Convert the console and script to PDF and upload all three files to Canvas.