Homework 7

Stats 20 Lec 1

Winter 2022

General Guidelines

Please use R Markdown for your submission. Include the following files:

- Your .Rmd file.
- The compiled/knitted HTML document.

Name your .Rmd file with the convention `123456789_stats20_hw0.Rmd`, where `123456789` is replaced with your UID and `hw0` is updated to the actual homework number. Include your first and last name and UID in your exam as well. When you knit to HTML, the HTML file will inherit the same naming convention.

The knitted document should be clear, well-formatted, and contain all relevant R code, output, and explanations. R code style should follow the Tidyverse style guide: https://style.tidyverse.org/.

Any and all course material, including these homework questions, may not be posted online or shared with anyone at any time without explicit written permission by the instructor (Michael Tsiang), even after the quarter is over. Failure to comply is a breach of academic integrity.

Note: All questions on this homework should be done using only functions or syntax discussed in Chapters 1–10 of the lecture notes. No credit will be given for use of outside functions.

Basic Questions

Collaboration on basic questions must adhere to Level 0 collaboration described in the Stats 20 Collaboration Policy.

The following information is used in Questions 1 and 2.

The `births.csv` file on Bruin Learn contains data on a sample of babies born in North Carolina. We are interested in investigating the association between baby weight and the mother's smoking habit (smokers versus non-smokers).

Question 1

The objective of this question is to give practice working with data frames and constructing basic graphics.

(a)

Read the dataset into R and save it to the workspace. The categorical variables should be stored as factors. Verify that the data has been loaded correctly.

Note: Do *not* print the entire data object. It is *extremely* bad practice/style to output more than about 10 rows of a matrix or data frame.

(b)

The `Habit` variable contains the smoker status for the mother of each baby in the sample. There are several observations in the data for which there is no information about the mother's smoker status. How many observations are in this category? What is the name of the category (level) in the `Habit` variable for these observations?

(c)

Create a bar plot of the mother's smoker status in the data. Change the color of the bars from the default grey color to a different color of your choice.

(d)

The **`droplevels()`** function drops unused levels from a factor or from factors in a data frame. Extract the observations in the data for which we know the mother's smoker status (either smoker or non-smoker) and drop the unused level from the `Habit` variable. Save this subset of the data as a separate data frame in the workspace. The subset of the data should contain all of the variables from the original data.

(e)

Using the data frame from (d), create side-by-side boxplots of the `weight` variable split by the mother's smoker status.

Question 2

The objective of this question is to give practice building a plot that uses several optional arguments and low level functions, including a legend with mixed types of symbols.

Create overlapping relative frequency histograms for the distributions of `weight` split by the mother's smoker status (either smoker or non-smoker). Be sure that the plot satisifies the following criteria:

- Change the x-label and main title to be more informative.
- Use different colors for the two histograms. Change the density of the shading so that the overlap between the histograms is visible.
- Superimpose density curves of the data, matching the color of the curves to the corresponding histogram.
- Add vertical lines that show the median weights for each distribution.
- Add a legend that helps understand the various colors/components of the plot.

Based on the plot, do you think there is a significant difference between the typical weight of a baby born to a mother who smokes and the typical weight of a baby born to a mother who does not smoke?

Hint 1: For legends with mixed types of symbols (points, lines, boxes, etc.), the `pch`, `lty`, `density`, and `border` (and other) arguments use `NA` to exclude those arguments from modifying the corresponding entries in the legend (`fill` uses `0` instead of `NA`). For example, a legend with two box entries and one line entry could have arguments `density = c(20, 30, NA)`, `border = c(1, 1, NA)`, and `lty = c(NA, NA, 1)`.

Hint 2: The boxes in the legend do not need to align or be centered perfectly with the lines.

The following information is used in Questions 3 and 4.

The `diamonds` data in the `ggplot2` package contains the prices and other measurements of almost 54,000 round cut diamonds.

Question 3

The objective of this question is to give practice creating and interpreting scatterplots and visualizing categorical variables with color.

(a)

Use scatterplots to explore possible relationships between the four numeric variables `carat`, `depth`, `table`, and `price`. Based on these plots, which two variables seem to have the strongest relationship? Does this relationship appear to be linear or nonlinear?

Hint 1: Visualizing all possible relationships can be done with a single command.

Hint 2: Since the `diamonds` data contains many observations, consider using optional arguments to remove redundant plots and/or shrink points to make the plot(s) take less time to generate.

(b)

Construct a scatterplot between the two numeric variables with the strongest relationship. Put the variable with higher variability on the y-axis.

- Change the point character from the default open circle to a different symbol of your choice.
- Shrink the size of the points to between 10 and 50% of the default size so that the points are easier to distinguish.
- Set the `col` argument to the `clarity` variable (i.e., `col = clarity`) to color the points in the scatterplot according to the clarity of each diamond.
- Add a legend that explains the color coding in the plot.

Explain why the default colors of `1` to `8` are chosen.

Hint: What is the type/class of the `clarity` column?

(c)

Construct the same scatterplot from (b) again, but change the default colors to different colors of your choice. *The color of the points should still correspond to the clarity of each diamond.* Be sure to update the legend to correspond to your new color scheme.

Hint: The 657 built-in color names can be found with the command `colors()`, or view them here: http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf. You can also see hexadecimal codes for various colors and visualize different color combinations here: https://coolors.co/.

(d)

Interpret the scatterplot from (b) or (c). What does the three-way relationship you observe tell you about the diamonds in the data?

Question 4

The objective of this question is to introduce the `matplot()` function and give further practice with optional arguments and low level functions.

(a)

Compute the mean price for each color and cut combination in the `diamonds` data, and store the result. The result should be a matrix object, where the rows correspond to the levels of `color` and the columns correspond to the levels of `cut`.

Hint: This computation can be done with one command.

(b)

The **`matplot()`** function plots columns of a matrix (or plots columns of one matrix against the columns of another). Use the `matplot()` function on the output from (a) to create a line plot with a separate line for each level of `cut`.

- Distinguish each line by separate line types, line widths, and/or colors of your choice.
- Use the `xaxt = "n"` argument to suppress the tick marks and labels on the x-axis, then use the `axis()` function to set the x-axis labels to be the levels of the `color` variable.
- Add a legend that explains the differences in the lines in the plot.

(c)

Interpret the line plot from (b). Does it appear that the mean price of diamonds differs for different levels of `color`? For different levels of `cut`? Which levels tend to have higher mean prices?