CMDA-3654

Homework 3

Eduardo Salvador

Due as a .pdf upload

Instructions:

Delete this Instructions section from your write-up!!

I have given you this assignment as an .Rmd (R Markdown) file.

- Change the name of the file to: `Lastname_Firstname_CMDA_3654_HW3.Rmd`, and your output should therefore match but with a `.pdf` extension.
- You need to edit the R Markdown file by adding chunks and filling them in appropriately with your code. Output will be generated automatically when you compile the document.
- You also need to add your own text before and after the chunks to explain what you are doing or to interpret the output.

Required: The final product that you turn in must be a .pdf file.

- You MUST Knit this document directly to a PDF, you are not allowed to knit to any other file type and then convert.

This assignment is to be done using Base R methods only!

The next assignment is devoted completely to plotting using ggplot2, so the use of ggplot2 is not allowed here.

Problem 1: (30 pts) Basic Summaries and Plotting with `Base R`

Install and load the `MASS` package for this problem, and load the `birthwt` data set that comes installed with `MASS`. This data set contains information on infant birth weight as well as observed risk factors. To find out more about this data set, see the help page `?birthwt`. In the following exercises, be sure to create an appropriate legend when neccesary, and label all axes and plots accordingly.

a. Provide univariate summaries for the variables in this data set.
b. Create a boxplot of birth weight (`bwt`) by `race`. Notice that the variable race is numerically coded. Make sure to assign the proper factor names when creating your plot. You should use different colors for each boxplot. Overlay a jittered stripcharts.
c. Create an overlayed density plot of birth weight given the smoking status of the mother, that is, make sure both densities are displayed onto the same plot. Use different colors and a legend.
d. Create a correlogram for all quantitative variables and comment on what you observe.
e. Make a scatterplot matrix using the `pairs()` function for all numeric variables. Color the points in the scatterplot matrix using different colors depending on race. A legend might be kinda tricky in this case, but not impossible. I'll settle for a description of which groups the colors represent.

Problem 2: (30 pts) Census Data

Turn your attention to the `adult.csv` data set..

a. Provide univariate summaries for the variables in this data set.

b. Create a bar chart displaying the counts of working class for all United States citizens.

c. Make a bivariate frequency table for the `workclass` variable as the rows and `race` as the columns. Show this table. In a second table, show the same table but with the marginal frequencies added.

d. Make a three-way frequency table using the `xtabs()` function for the `workclass`, `race`, and `sex` variable (have sex be the 3rd dimension). Then use `ftable()` to flatten the 3-D table.

e. Create a **relative frequency stacked barchart** displaying the counts of `pay` categories with respect to the `marital` category. .

Problem 3: (20 pts) The `iris` dataset

(**Note:** When we say plot **"a" vs "b"**, by default "a" is on the y-axis, and "b" is on the x-axis.)

a. Plot the Petal Width vs Petal Length with different colors and plot characters for the different classes of plants. Be sure to add a legend.

b. Plot the Sepal Width vs Sepal Length with different colors and plot characters for the different classes of plants. Be sure to add a legend.

c. What proportion of flowers have a Petal Length greater than 4, Petal widths between 1 and 2, and Sepal Widths and Lengths within 0.5 units of their median values?

d. Observing the plots in (a) and (b), if you had to distinguish between classes by using either petal dimensions or sepal dimensions, which one would you choose: petals or sepals, and why?

Problem 4: (20 pts) The `babynames` dataset

Consider the `babynames` data from assignment 1 located within the R library package of the same name..

a. Create a subset of the data with female babies named "Mary" from 1880-2014. How many observations are in this subset?

b. Create a subset of the data with female babies named "Sophia" from 1880-2014. How many observations are in this subset?

c. Construct a plot of the proportion of female babies named "Mary" from 1880-2014. On the same plot, add/overlay a plot of the proportion of female babies named "Sophia" from 1880-2014. Use different colors for "Mary" vs "Sophia" and add a legend.

d. Briefly describe your interpretation of the plot.
