Lab 2: Data cleaning/preparation and visualization

Statistics 12

Some exercises based on labs by Nicolas Christou.

Submission Instructions

1. Save your final answers in PDF format.
2. Upload your PDF to the CCLE submission link
3. Submit the assignment by the deadline. NO late submissions will be accepted.

Objectives

1. Understand logical statements and subsetting
2. Reinforce knowledge on visualization techniques
3. Compute z-scores and understand the Empirical Rule

Collaboration Policy

In Lab you are encouraged to work in pairs or small groups to discuss the concepts on the assignments. However, DO NOT copy each other's work as this constitutes cheating. The work you submit must be entirely your own. If you have a question in lab, feel free to reach out to other groups or talk to the TA if you get stuck.

Intro Logical Statements/Relational Operators

Logical Expressions: Type **?Comparison** to see the R documentation on the list of all relational operators you can apply. Many logical expressions in R use these relational operators.

Try running the lines of code below that use the relational operators >, >=, <=, ==, !=. We can also use the conditions & and | to combine relational operators. Try running these examples:

```
4 > 3 # Is 4 greater than 3?
c(3,8) >= 3 # Is 3 or 8 greater than or equal to 3?
c(3,8) <= 3 # Is 3 or 8 less than or equal to 3?
c(1,4,9) == 9 # Is 1, 4, or 9 exactly equal to 9?
c(1,4,9) != 9 # Is 1, 4, or 9 not (exactly) equal to 9?
c(1,2,4,7) < 5 & c(1,2,4,7) > 1 #Which of 1,2,4,7 is less than 5 and greater than 1?
c(1,2,4,7) < 5 | c(1,2,4,7) > 1 #Which of 1,2,4,7 is less than 5 or greater than 1?
```

Notice that the output is a logical vector (i.e., uses TRUE and FALSE) that has the length of the vector on the left of the relational statement.

Applications of logical statements: calculations

We can perform certain calculations on logical vectors because R reads TRUE as 1 and FALSE as 0. Create the NCbirths object from last lab and try these examples:

```
sum(NCbirths$weight > 100) #the number of babies that weighed more than 100 ounces
mean(NCbirths$weight > 100) #the proportion of babies that weighed more than 100 ounces
mean(NCbirths$Gender == "Female") #the proportion of female babies
mean(NCbirths$Gender != "Male") #again gives the proportion of female babies
```

Applications of logical statements: subsets

We can combine logical statements with square brackets to subset data based on conditions. Examples with NCbirths:

```
fem_weights <- NCbirths$weight[NCbirths$Gender == "Female"]
```

With the line above we created a vector called fem_weights that contains the weights of all the female babies.

Good coding practices:

1. Use the pound symbol (#) often to comment on different code sections. Consider using them to label your exercise numbers and question parts, and to help describe what your code does.
2. Use good spacing. Adding a space between arguments and inside of functions makes your code easier to read. You can also skip lines for clarity
3. Create as many objects as you like to make it easier to follow. For example, consider my line above creating the fem_weights object. An alternative way to code this using best practices is below:

```
## Create an object with the baby weights from NCbirths
baby_weight <- NCbirths$weight

## Create an object with the baby genders from NCbirths
baby_gender <- NCbirths$Gender

## Create a logical vector to describe if the gender is female
is_female <- baby_gender == "Female"

## Create the vector of weights containing only females
fem_weights <- baby_weight[is_female]
```

Exercise 1

We will be working with lead and copper data obtained from the residents of Flint, Michigan from January-February, 2017. Data are reported in PPB (parts per billion, or μg/L) from each residential testing kit. Remember that "Pb" denotes lead, and "Cu" denotes copper. You can learn more about the Flint water crisis at https://en.wikipedia.org/wiki/Flint_water_crisis.

a. Download the data from CCLE and read it into R. When you read in the data, name your object "flint".

b. The EPA states a water source is especially dangerous if the lead level is 15 PPB or greater. What proportion of the locations tested were found to have dangerous lead levels?

c. Report the mean copper level for only test sites in the North region.

d. Report the mean copper level for only test sites with dangerous lead levels (at least 15 PPB).

e. What proportion of the lead observations are above zero, but also less than the dangerous level?

f. Report the mean lead and copper levels.

g. Create a box plot with a good title for the lead levels.

h. Based on what you see in part (g), does the mean seem to be a good measure of center for the data? Report a more useful statistic for this data.

Exercise 2

The data here represent life expectancies (Life) and per capita income (Income) in 1974 dollars for 101 countries in the early 1970's. The source of these data is: Leinhardt and Wasserman (1979), New York Times (September, 28, 1975, p. E-3). They also appear on Regression Analysis by Ashish Sen and Muni Srivastava. You can access these data in R using:

```
life <- read.table("http://www.stat.ucla.edu/~nchristo/statistics12/countries_life.txt",
header=TRUE)
```

a. Construct a scatterplot of Life against Income. Note: Income should be on the horizontal axis. How does income appear to affect life expectancy?

b. Construct the boxplot and histogram of Income. Are there any outliers?

c. Split the data set into two parts: One for which the Income is strictly below $1000, and one for which the Income is at least $1000. Come up with your own names for these two objects.

d. Use the data for which the Income is below $1000. Plot Life against Income and compute the correlation coefficient. *Hint: use the function cor()*

Exercises continue on the next page

Exercise 3

Use R to access the Maas river data which contains the concentration of lead and zinc in ppm at 155 locations at the banks of the Maas river in the Netherlands. Read the data in R as follows:

```
maas <- read.table("http://www.stat.ucla.edu/~nchristo/statistics12/soil.txt", header=TRUE)
```

a. Compute the summary statistics for lead and zinc using the summary() function. What is the interquartile range (IQR) for zinc?

b. Plot two histograms: one of lead and one of log(lead).

c. Plot log(lead) against log(zinc). What do you observe?

d. The level of risk for surface soil based on lead concentration in ppm is given on the table below:

Mean concentration (ppm)	Level of risk
Below 150	Lead-free
Between 150-400	Lead-safe
Above 400	Significant environmental lead hazard

Use techniques similar to last lab to give different colors and sizes to the lead concentration at these 155 locations. You do not need to use the maps package create a map of the area. Just plot the points without a map.

Exercise 4

The data for this exercise represent approximately the centers (given by longitude and latitude) of each one of the City of Los Angeles neighborhoods. See also the Los Angeles Times project on the City of Los Angeles neighborhoods at: http://projects.latimes.com/mapping-la/neighborhoods/. You can access these data at:

```
LA <- read.table("http://www.stat.ucla.edu/~nchristo/statistics12/la_data.txt", header=TRUE)
```

a. Plot the data point locations. Use good formatting for the axes and title. Then add the outline of LA County by typing:

```
map("county", "california", add = TRUE)
```

b. Do you see any relationship between income and school performance? Hint: Plot the variable Schools against the variable Income and describe what you see. Ignore the data points on the plot for which Schools = 0. Use what you learned about subsetting with logical statements to first create the objects you need for the scatter plot. Then, create the scatter plot. **Alternate methods may only receive half credit.**

c. Let's say we have a new observation of a $100,000 income. Compute the z-score for this observation using the LA Income data. What does this z-score tell us?

d. Determine the proportion of LA Income observations that are within 1SD, 2SD, and 3SD of the average. How do these results compare with the empirical rule?