CWRU DSCI351-351M-451: EDA: MidTerm

Prof.:Roger French, TA:JiQi Liu

Oct. 11, 2018

Contents

1 Name:

1.0.1 Don't forget to put your name in the script, and in the filename

1.0.2 Filename Schema; 1708DSCI?51-MidTerm-YOURLASTNAME.Rmd

1.0.2.1 Where you specify 351 or 451 and fill in your last name.

1.0.3 Show your R code and put your answers at "Answer <- ?"

1.1 Question 1. The 4 Freedoms of FOSS (1 pt)

The definition of free and open-source software consists of four freedoms (freedoms 0 through 3).

Which of the following is NOT one of the freedoms that are part of the definition?

- A) The freedom to improve the program, and release your improvements to the public, so that the whole community benefits.
- B) The freedom to redistribute copies so you can help your neighbor.
- C) The freedom to study how the program works, and adapt it to your needs.
- D) The freedom to restrict access to the source code for the software.

Answer <- ?

1.2 Question 2. Variable Class (1 pt)

If I execute the expression

```
x <- 4
```

in R, what is the class of the object ‘x’?

What is the class of the object defined by the expression

```
x <- c(4, "a", TRUE)
```

?

Answer <- ?

1.3 Question 3. Row Bind (1 pt)

If I have two vectors x <- c(1,3, 5) and y <- c(3, 2, 10), what is produced by the expression rbind(x, y)?

- A) a 2 by 2 matrix
- B) a vector of length 3
- C) a matrix with two rows and three columns
- D) a vector of length 2

Answer <- ?

1.4 Question 4. Subsetting (1 pt)

Suppose I have a vector x <- c(3, 5, 1, 10, 12, 6) and I want to set all elements of this vector that are less than 6 to be equal to zero.

What R code achieves this?

1.5 Question 5. Functions (1 pt)

1.5.1 5.1) Suppose I define the following function in R

```
cube <- function(x, n) {
x^3
}
```

What is the result of running cube(3)

5.1 Answer <- ?

1.5.2 5.2) The following code will produce a warning in R.

```
x <- 1:10
if (x > 5) {
x <- 0
}
```

```
## Warning in if (x > 5) {: the condition has length > 1 and only the first
## element will be used
```

Why?

5.2 Answer <- ?

1.6 Question 6. mtcars (1 pts)

Load the 'mtcars' dataset in R with the following code

library(datasets) data(mtcars)

There will be an object names 'mtcars' in your workspace. You can find some information about the dataset by running

?mtcars

What is the absolute difference between the average horsepower of 4 cylinder cars and the average horsepower of 8 cylinder cars ?

1.7 Question 7. Graphics (1 pts)

7.1 Which of the following functions is part of the base graphics system?

- A) splom()
- B) bwplot()
- C) barchart()
- D) hist()

7.1 Answer <- ?

7.2 Which of the following functions is generally used to annotate a plot in the base graphics system?

- A) boxplot()
- B) lines()
- C) barplot()
- D) plot()

7.2 Answer <- ?

1.8 Question 8. Tidy Data Analysis of Lord of the Rings (3 pts)

If I had one thing to tell biologists learning bioinformatics, it would be "write code for humans, write data for computers". Vince Buffalo (@vsbuffalo) July 20, 2013

An important aspect of "writing data for computers" is to make your data **tidy**.

Key features of **tidy** data:

- Each column is a variable
- Each row is an observation

If you are struggling to make a figure, for example, stop and think hard about whether your data is tidy. Untidiness is a common, often overlooked cause of agony in data analysis and visualization.

I will give you a concrete example of some untidy data

```
fotr <- read.csv('1708The_Fellowship_Of_The_Ring.csv')
rotk <- read.csv('1708The_Return_Of_The_King.csv')
tt <- read.csv('1708The_Two_Towers.csv')

View(fotr)
View(rotk)
View(tt)
```

We have one table per movie. In each table, we have the total number of words spoken, by characters of different races and genders.

You could imagine finding these three tables as separate worksheets in an Excel workbook. Or hanging out in some cells on the side of a worksheet that contains the underlying data raw data. Or as tables on a webpage or in a Word document.

This data has been formatted for consumption by human eyeballs. The format makes it easy for a human to look up the number of words spoken by female elves in The Two Towers.

But this format actually makes it pretty hard for a computer to pull out such counts and, more importantly, to compute on them or graph them.

8.1 An important aspect of "writing data for computers" is to make your data **tidy**.

Two key features of **tidy** data are:

Answer <- 1.

Answer <- 2.

8.2 Look at the tables above and answer these questions:

What's the total number of words spoken by male hobbits?

Does a certain Race dominate a movie? Does the dominant Race differ across the movies?

Answer <- ?

8.3 How well does your approach scale if there were many more movies or if I provided you with updated data that includes all the Races (e.g. dwarves, orcs, etc.)?

Answer <- ?

```
lotr <- read.csv('1708lotr-tidy.csv')
View(lotr)
```

Notice that tidy data is generally taller and narrower.

It doesn't fit nicely on the page.

Certain elements get repeated alot, e.g. `Hobbit`.

For these reasons, we often instinctively resist **tidy** data as inefficient or ugly.

But, unless and until you're making the final product for a textual presentation of data, ignore your yearning to see the data in a compact form.

Now using tidyverse packages, pipes and dplyr answer the following questions

8.4 What's the total number of words spoken by male hobbits?

Answer <- ?

8.5 Does a certain race dominate a movie? Does the dominant race differ across the movies?

You'll first want tosum across gender, to obtain word counts for the different races by movie.

Answer <- ?

8.6 Now using ggplot2 lets visualize these results. We can stare hard at those numbers to answer the question. But even nicer is to depict the word counts we just computed in a barchart.

Take Home Message:

Having the data in tidy form was a key enabler for our data aggregations and visualization.

Tidy data is integral to efficient data analysis and visualization.

If you're skeptical about any of the above claims, it would be interesting to get the requested word counts, the barchart, or the insight gained from the chart without tidying or plotting the data. And imagine redoing all of that on the full dataset, which includes 3 more Races, e.g. Dwarves.