CWRU DSCI351-351M-451: EDA: MidTerm Exam

2208-DSCI-351-351m-451-MidTerm-NAME.Rmd

Prof.:Roger French, Paul Liu TA:Will Oltjen, Kristen Hernandez, Mingxuan Li

Oct. 18, 2022

Contents

0.0.0.0.1 Instructions

NAME.Rmd => …caseID.Rmd

- Don't forget to put your name in the script, and in the filename
- Filename Schema; 2208-DSCI351-351m-451-MidTerm-YOURCaseID.Rmd

Answers and Code Style

- Show your R code, using good coding style
- Explain your reasoning
- And put your answers at "ANSWER <- ?"

There are 8 Questions

- Q1: 1 pt
- Q2: 1/2 pt
- Q3: 1/2 pt
- Q4: 1 pt

- Q5: 1 pt
- Q6: 1 pt
- Q7: 2 pts
- Q8: 3 pts

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

The definition of free and open-source software (FOSS)

- 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 <- ?

0.0.0.2 Question 2. Variable Class (1/2 pt)

If I execute the expression

```
x <- 4
```

in R,

- what is the class of the object 'x'?

ANSWER <- ?

What is the class of the object

- defined by the expression `y <- c(4, "a", TRUE)`?

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

ANSWER <- ?

0.0.0.3 Question 3. Row Bind (1/2 pt)

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

- 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 <- ?

0.0.0.4 Question 4. Subsetting (1 pt)

Suppose I have a vector `d <- 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?

0.0.0.4.1 Q4.1) Using base R commands

0.0.0.4.2 Q4.2) Using tidyverse commands

0.0.0.5 Question 5. Functions (1 pt)

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

0.0.0.5.1 Q5.1) Suppose I define the following function in R What is the result of running

`cube(3)`

5.1 ANSWER <- ?

0.0.0.5.2 Q5.2) The following code will produce a warning in R. The warning may not show if run in a .Rmd file's R code block

- If so you should copy the R code below and run it in your R console.

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

Why?

5.2 ANSWER <- ?

0.0.0.6 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 ?

6. ANSWER <- ?

0.0.0.7 Question 7. Graphics (2 pts)

0.0.0.7.1 Q7.1 Make a histogram of the number of cylinders in cars in the the `mtcars` dataset using the `hist()` function. Label the x-axis 'Cylinders' and the title 'Histogram of Cylinders'.

- Make a comparable graph using ggplot2 with x-axis label 'Cylinders', y-axis label 'Frequency', and title 'Histogram of Cylinders'

0.0.0.7.2 Q7.2 This is an example of a plot generated using `subset()` and `geom_text()` in ggplot using the mtcars dataset. (1 point)

Recreate this plot using ggplot as best you can.

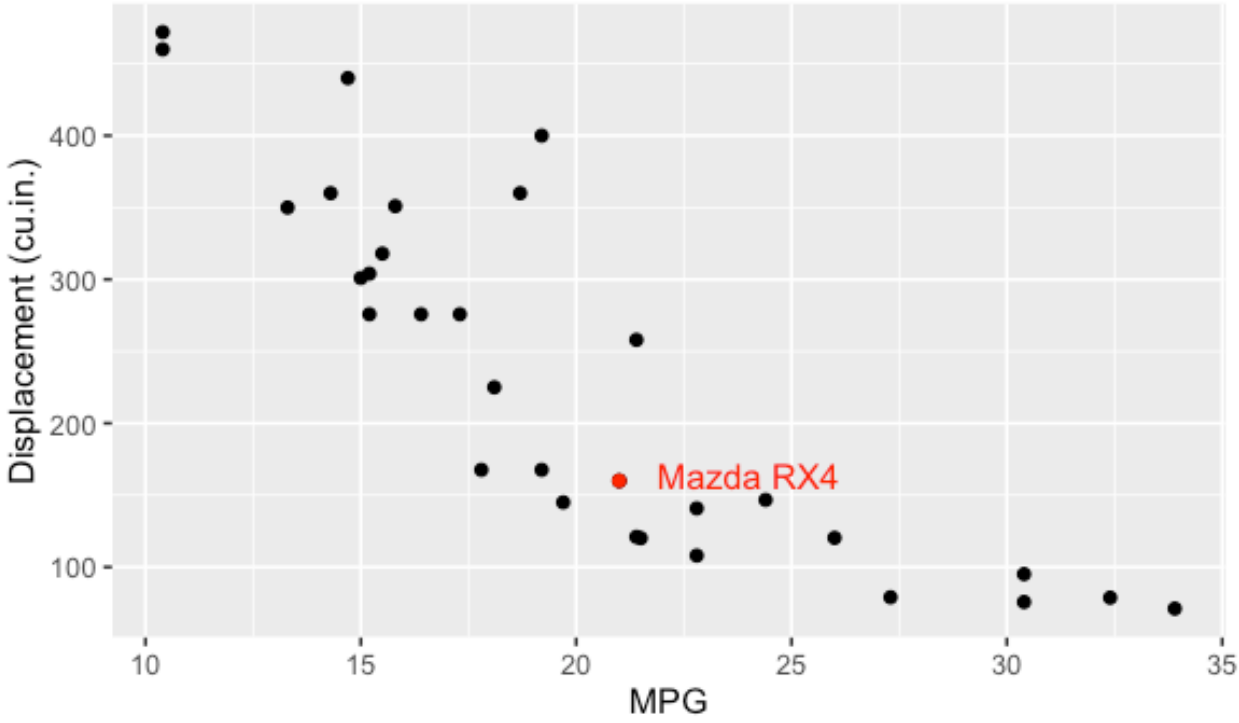

Figure 1: Scatter plot generated using mtcars with datapoint labeled.

0.0.0.8 Question 8. Analyze Lord of the Rings (LOR) (3 pts) Its often said that we should "write code for humans, write data for computers".

- 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 data visualization.

I will give you a concrete example of some untidy data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()    masks stats::lag()
```

```
fotr <- read.csv('./data/2108The_Fellowship_Of_The_Ring.csv')
rotk <- read.csv('./data/2108The_Return_Of_The_King.csv')
tt <- read.csv('./data/2108The_Two_Towers.csv')

glimpse(fotr)
```

```
## Rows: 3
## Columns: 4
## $ Film   <chr> "The Fellowship Of The Ring", "The Fellowship Of The Ring", "Th~
## $ Race   <chr> "Elf", "Hobbit", "Man"
## $ Female <int> 1229, 14, 0
## $ Male   <int> 971, 3644, 1995
```

```
glimpse(rotk)
```

```
## Rows: 3
## Columns: 4
## $ Film   <chr> "The Return Of The King", "The Return Of The King", "The Return~
## $ Race   <chr> "Elf", "Hobbit", "Man"
## $ Female <int> 183, 2, 268
## $ Male   <int> 510, 2673, 2459
```

```
glimpse(tt)
```

```
## Rows: 3
## Columns: 4
## $ Film   <chr> "The Two Towers", "The Two Towers", "The Two Towers"
## $ Race   <chr> "Elf", "Hobbit", "Man"
## $ Female <int> 331, 0, 401
## $ Male   <int> 513, 2463, 3589
```

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 raw data.
- Or as tables on a webpage or in a Word document.

This data has been formatted for consumption by human eyeballs.

- The formatmakes 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.

0.0.0.8.1 Q8.1 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.

0.0.0.8.2 Q8.2 Just looking at these tables, answer these questions: (You'll do this with code in the next part)

- What's the total number of words spoken by male hobbits in each of the three movies?

Answer <- ?

- Does a certain Race dominate a movie?

Answer <- ?

- Does the dominant Race differ across the movies?

Answer <- ?

0.0.0.8.3 Q8.3 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('./data/2108lotr-tidy.csv')
glimpse(lotr)
```

```
## Rows: 18
## Columns: 4
## $ Film   <chr> "The Fellowship Of The Ring", "The Fellowship Of The Ring", "Th~
## $ Race   <chr> "Elf", "Hobbit", "Man", "Elf", "Hobbit", "Man", "Elf", "Hobbit"~
## $ Gender <chr> "Female", "Female", "Female", "Female", "Female", "Female", "Fe~
## $ Words  <int> 1229, 14, 0, 331, 0, 401, 183, 2, 268, 971, 3644, 1995, 513, 24~
```

Notice that tidy data is generally taller and narrower.

- It doesn't fit nicely on the page.
- Certain elements get repeated a lot,
  - 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

0.0.0.8.4 Q8.4 What's the total number of words spoken by male hobbits? Answer <- ?

0.0.0.8.5 Q8.5 Does a certain race dominate a movie? Does the dominant race differ across the movies?

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

Answer <- ?

0.0.0.8.6 Q8.6 Now using ggplot2 let us visualize these results. We can stare hard at those numbers to answer the question.

- But its even nicer to depict the word counts
  - we just computed in a barchart.