R 4 Data Science Assessment 3

# Online timed assessment

# Instructions

You have thirty minutes to complete this assessment. After thirty minutes, the system will submit your work to that point for grading.

The assessment is comprised of ten multiple-choice questions, and two short-answer questions.

The multiple-choice questions are worth 2.5 each, and the short answer questions are worth 5 each.

# Multiple-choice

*<EASY (don’t publish these subheaders), Correct answers boldened >*

1. What is the dplyr verb most often paired with summarise
   1. Select()
   2. Pull()
   3. **Group\_by()**
   4. Mutate()
2. Select the tidyverse convention of the following function
   1. readCSV()
   2. **read\_csv()**
   3. read.csv()
   4. ReadCsv
3. What type of data is the following vector?  
   vector <- c(“Ben”, Michael”, “Hadley”)
   1. Logical
   2. Numerical
   3. Datetime
   4. **String**

*<Moderate>*

1. What is the file extension of an R script, and what is the file extension of an R project? (The extensions are separated by a comma in the choices).
   1. .rcode, .rstudio
   2. .rcode, .rproj
   3. **.r, .rproj**
   4. .r, .rstudio
2. To colour the inside of a histogram, what should you map colour to?
   1. **Fill**
   2. Colour
   3. Contents
   4. Internal
3. What is wrong with this code?  
   diamonds |>  
   ggplot(mapping(x , price)) +  
   geom\_point)
   1. Brackets are incorrect
   2. Mapping is incorrect
   3. **Brackets and mapping are incorrect**
   4. Commas are incorrect

*<Harder>*

1. What is the result of the following code?  
   400 %/% 10
   1. 4
   2. 4000
   3. 400
   4. **40**
2. What is the name of the top right pane in Rstudio (assuming you haven’t reconfigured the panes)?
   1. **Source editor**
   2. Environment
   3. Console
   4. Viewer
3. Which join should you use if you only want to keep observations in dataset x with corresponding values in data set y?
   1. Anti-join
   2. **Inner-join**
   3. Left-join
   4. Right-join

*<Difficult>*

1. What is the name of the passage of time measured in human units?
   1. **Period**
   2. Interval
   3. Duration
   4. Time span

# Short-answer 1

A ) Write the code in the box below to compute the average price of the specified subset of the diamonds data set. Remove as few NAs as possible. Ensure code is formatted as if you were writing in R.

Subset:

Colour D, clarity VS2, carat greater than or equal to 1.5

**Potential solution:**

diamonds |>

select(price, color, clarity, carat) |>

drop\_na() |>

filter(

color == "D",

clarity == "VS2",

carat >= 1.5

) |>

summarise(mean\_price = mean(price))

# Short answer 2

Using the whole diamonds dataset create a new variable with the name of your choice using a case\_when statement. This variable will describe the price band of the diamonds such that those less than 5000 will be called “sub 5k”, those with 5000<price<1000 are called “betweek 5k and 10k” and those greater than 10000 and called “10k+”. Then determine how many are observations are in each group.

**Potential solution**

diamonds |>

mutate(coolness = case\_when(

price < 5000 ~ "sub 5k",

price > 5000 & price < 10000 ~ "between 5k and 10k",

price > 1000 ~ "10k+"

)) |> count(coolness)