LaughlinLab3

Ty

2022-09-19

Introduction to R - Part 2

This Markdown file contains each of the practice exercises for the in the Introduction to R - Part 2 lecture of BIO247 - Bioinformatics. Each of these exercises is used to practice using dataframes and new functions. Accompanying each exercise is an explanation of the goal of the exercise and then the solutions found to each task.

Exercise 1: Rows and Columns

This exercise was used to ensure comfort with uploading data to R studio, as well as navigating data frames through rows and columns. The data used in this exercise is the Test_V_output data uploaded to Sakai by the BIO247 professor, Dr. Casey Hansen.

Part 1: Upload Test_V_output.csv to RStudio

```
setwd("C:\\Users\\TyCam\\Desktop\\BIO247\\Lab\\Outputs")
Test_V_output <- read.csv(file = "Test_V_output.csv", sep = ",", header = TRUE)
```

The setwd() function was used to ensure that Rstudio was accessing the directory where all other BIO247 data and work is held. Then read.csv() was used to upload the Test_V-output data to Rstudio, while assigning that data to a vector, so that the data could then be used more easily going forward.

Part 2: Find the Total Score column, and save it into a vector called tot_score

```
tot_score <- Test_V_output$`Total.Score`
```

A column with name the “Total.Score” was assigned to a vector, “tot_score”, to practice accessing data from a specific column.

Part 3: Choose a row in the dataframe. Find the molecules involved in that interaction, and find its Total Score

```
Test_V_output$Target.ID[5]
```

```
## [1] "p49815"
```

```
Test_V_output$Source.ID[5]
```

```
## [1] "rsk"
```

```
tot_score[5]
```

```
## [1] 240
```

The different columns containing the different protein names and the total score were accessed, while the same row/interaction was consistently found by indexing the data frame.

Exercise 2: Mean and Stddev

This exercise was used to show that existing data can be used to then find new data. The data used in this exercise is the Test_V_output data uploaded to Sakai by the BIO247 professor, Dr. Casey Hansen.

Part 1: Find the mean and standard deviation of the Total Scores in this set

```
mean(tot_score)
```

```
## [1] 63.79919
```

```
sd(tot_score)
```

```
## [1] 55.29295
```

The mean() and sd() functions were used find the mean and standard deviation of the data within the vector, “tot_score”.

Part 2: Also find out how many unique papers were used to create this set

```
#gsub("\\]", "", Test_V_output$Paper.ID)
#gsub("\\[", "", Test_V_output$Paper.ID)
#gsub("\\'", "", Test_V_output$Paper.ID)
#gsub(" ", "", Test_V_output$Paper.ID)
split_ID <- unlist(strsplit(Test_V_output$Paper.ID, ","))
length(unique(split_ID))
```

```
## [1] 191
```

Unnecessary characters from each piece of data in the “Paper.ID” column, using the gsub() function 4 times. Each usage of gsub() substituted an unnecessary character with “ “, or nothing, effectively removing the unnecessary characters. Note that the gsub() functions were run in a chunk using include=FALSE to prevent excessive and unnecessary intermediate outputs from being displayed. To ensure that that it is still known that the calls to gsub() were there,they were placed in a the chunk with the other code, behind comments.

Exercise 3: Vectorization

This exercise was used to practice adding more data to a given data frame. The data used in this exercise is the Test_V_output data uploaded to Sakai by the BIO247 professor, Dr. Casey Hansen.

Part 1: Pretend the Total Score column doesn’t exist

No work needed to be done for this task.

Part 2: You have Evidence Scores, Kind Scores, Match Scores, Epistemic Value

```
Se <- Test_V_output$Evidence.Score
Sm <- Test_V_output$Match.Score
Sk <- Test_V_output$Kind.Score
Sb <- Test_V_output$Epistemic.Value
```

The data from each column containing scores were each added to a new vector with a more concise name, to make usage of the data easier in the future.

Part 3: I’m telling you that Total Score = (Kind+(Evidence(Match)))(Epistemic)

```
new_tot <- (Sk + (Se * Sm))*Sb
Test_V_output$Total <- new_tot
```

A new vector was made by using each of individual types of scores to recreate the tot_score column. Then, this vector was turned into a new column in the data frame.

Exercise 4: Subsets

This exercise was used to practice removing data, which is perceived as unnecessary, from a data frame. The data used in this exercise is the Test_V_output data uploaded to Sakai by the BIO247 professor, Dr. Casey Hansen.

Part 1: You decide that the locations, cell type, organism, etc. is getting in the way

```
Test_V_output2 <- subset(Test_V_output, select = -c(`Location`, `Cell.Type`, `Organism`))
```

The unnecessary columns were "removed" by creating a new vector. The subset function was used to ensure that all the data from Test_V_output was in the new data frame, except for the 3 unwanted columns.

Part 2: You also decide you only want the top 100 scored interactions (remember when we first opened the file it was sorted highest to lowest score)

```
Test_V_output3 <- Test_V_output2[1:100,]
```

Another new data frame was made, but, this time, the unwanted data was removed by indexing the previous data frame, so that only the first 100 rows were included.