BIO247 Lab 3

Whitney Domico

2022-09-19

This document details the R Exercises 1-4 found in the BIO247-Chapter8\_P2 presentation. They call for the file Test\_V\_output.csv, which is a table of relevant information about 992 proteins involved in cancer research.

## Exercise 1: Rows and Columns

The goal for this exercise is to save the Total Scores column into a vector and then find the ‘Total Score’ and molecules involved in the interaction for the user’s “favorite row,” which for these purposes will be defined as row 17.

The code below creates a vector of the total scores.

tot\_score <- c(Test\_V\_output$Total.Score)

To find the molecules involved in row 17, one must find both the ‘Target.Name’ and ‘Source.Name’.

The code below calls the ‘Target.Name’ involved in row 17.

Test\_V\_output[17,'Target.Name']

## [1] "tsc2"

The code below calls the ‘Source.Name’ involved in row 17.

Test\_V\_output[17,'Source.Name']

## [1] "ppargamma"

Together, these comprise the molecules involved. Therefore, the answer is tsc2 and ppargamma.

The final part of this exercise asks for the Total Score of row 17, which uses the code below.

Test\_V\_output[17,'Total.Score']

## [1] 240

Therefore, the Total Score for row 17 is 240.

## Exercise 2: Mean and Stddev

The goal of this exercise is to find the mean and standard deviation of the Total Scores in Test\_V\_output. The second part is to find how many unique papers were used in this set.

To find the mean of the Total Scores, the following code is used. Recall that a previous vector titled ‘tot\_score’ was created

mean(tot\_score)

## [1] 63.79919

Therefore, the mean is approximately 63.8.

To find the standard deviation of the Total Scores, the following code is used.

sd(tot\_score)

## [1] 55.29295

Therefore, the standard deviation is approximately 55.3.

The second part of this exercise is to find the amount of unique papers that were used in this set. These would be found under the ‘Paper ID’ column and would be coded as follows.

length(unique(Test\_V\_output$Paper.ID))

## [1] 157

Therefore, the number of unique papers is 157. Of course, this counts groups of papers as separate items.If we were to find the true unique number of papers, we would first have to separate the items in the vector into singular papers. The first step to this is to remove all “junk characters,” which are “[”, ”]”, ” “,”’“.

Test\_V\_output$Paper.ID <- gsub("\\'","", Test\_V\_output$Paper.ID)  
Test\_V\_output$Paper.ID <- gsub("\\[", "", Test\_V\_output$Paper.ID)  
Test\_V\_output$Paper.ID <- gsub("\\]", "", Test\_V\_output$Paper.ID)  
Test\_V\_output$Paper.ID <- gsub(" ","", Test\_V\_output$Paper.ID)

Then the individual papers need to be separated. This can be achieved by splitting them where there are commas. This will return a list (not a vector), so we also must use the ‘unlist’ function to create the new vector, SplitPaperIDs.

SplitPaperIDs <- unlist(strsplit(Test\_V\_output$Paper.ID, ","))

Now that the vector is created, the length can be found using the ‘unique’ function.

length(unique(SplitPaperIDs))

## [1] 102

Therefore, the true number of unique papers is 102.

## Exercise 3: Vectorization

The goal of this exercise is to recreate the Total Score column by calculating it using the evidence, kind, match, and epistemic scores. It is given that Total Score = (Kind+(Evidence(Match))(Epistemic).

The vector names for kind, evidence, match, and epistemic scores are as follows, respectively:

* *Test\_V\_output$Kind.Score*
* *Test\_V\_output$Evidence.Score*
* *Test\_V\_output$Match.Score*
* *Test\_V\_output$Epistemic.Value*

The new column will be named “New.Total.Score”, and is coded for as follows:

Test\_V\_output$New.Total.Score <- (Test\_V\_output$Kind.Score+(Test\_V\_output$Evidence.Score\*Test\_V\_output$Match.Score))\*Test\_V\_output$Epistemic.Value

## Exercise 4: Subsets

The goal of this exercise is to remove the columns labeled “Location”, “Cell.Type”, and “Organism”, and also to view the dataframe in a way that only includes the first 100 rows.

To remove the aforementioned columns, the following code is used:

Test\_V\_output = subset(Test\_V\_output, select = -c(Location, Cell.Type, Organism))

To view the dataframe as only the first 100 rows, the following code is used (it is represented as a store function to avoid the output of 100 lines):

store <- Test\_V\_output[1:100,]