Lab3

2022-09-19

**Exercise 1: Rows & Columns**

*1.1: Uploading the Test\_V\_output.csv*

The dataframe was uploaded using the files tab in r studio.

*1.2: Saving a vector with total score values*

The total score column was saved to a vector named tot\_scores. This was done by assigning the vector with a c() function while referencing the total score column in the Test\_V\_output dataframe. Test\_V\_output data frame was obtained from Dr. Casey Hansen and contains data on many molecules and their interactions.

Test\_V\_outputs <- read.csv(file = "//Users//nagendra//Desktop//Bioinformatics//Lecture//Test\_V\_output.csv" , sep = ",", header = TRUE)

tot\_scores <- c(Test\_V\_outputs$Total.Score)

*1.3: Finding the Total Score of p53 molecule*

The total score value was indexed for p53 by referencing a cell in the dataframe.

Test\_V\_outputs[2,"Total.Score"]

## [1] 440

**Exercise 2: Mean and Stddev**

*2.1: Finding the mean and stdv*

The mean of the total score column was calculated using the mean function and the tot\_scores vector

mean(tot\_scores)

## [1] 63.79919

The standard deviation of the total score column was calculated using the standard deviation function and the tot\_scores vector

sd(tot\_scores)

## [1] 55.29295

*2.2: Finding unique PMCID* The gsub function was first used to replace all junk symbols with “nothing”. Since the Paper ID column is a vector of strings, the strsplit function was then used to split the characters in the string that are seperated by commas. The unlist function was then used to turn the output from the strsplit function into a vector. The unique function was used to find unique values in the split\_ID vector that contains the ID’s. The length function then counts the number of characters in the list that is output of the unique function.

list1<- gsub("\\[","",Test\_V\_outputs$`Paper.ID`)  
list2<-gsub("\\]","",list1)  
list3<- gsub("\'","",list2)  
list4 <- gsub(" ", "", list3)  
split\_ID <- unlist(strsplit(list4, ","))  
length(unique(split\_ID))

## [1] 102

**Exercise 3: Vectorization**

*3.1: Creating my own Total scores column* A new column called Calc\_score was made to calculate the total score values of specific molecules by indexing specific score values and performing multiplication. The total score equation in Total Score = (Kind+(Evidence*Match))*Epistemic.

Test\_V\_outputs$Calc\_TotScore <- (Test\_V\_outputs$Kind.Score+(Test\_V\_outputs$Evidence.Score\*Test\_V\_outputs$Match.Score))\*Test\_V\_outputs$Epistemic.Value

**Exercise 4: Subsets**

*4.1 Deleting rows and columns*

The Cell Type, Location, and organism columns were deleted using a subset function and saved to a dataframe named Test\_V\_outputs\_short. The same dataframe was then modified further by removing all rows except the first 100 by subsetting rows. This new truncated data frame was saved as a new csv file using thee write.csv function.

Test\_V\_outputs\_short <- subset(Test\_V\_outputs, select = -c(Cell.Type, Location, Organism))  
Test\_V\_outputs\_short <- Test\_V\_outputs\_short[1:100,]  
write.csv(Test\_V\_outputs\_short, "//Users//nagendra//Desktop//Bioinformatics//Lab//lab3//Test\_V\_ouputs\_short.csv")