LAB SHEET-2

Data Manipulation - 1

For this lab sheet, we will be using the "mtcars" dataset which is preloaded in R

Reading and Writing Data from CSV files -

data <- read.csv(filename.csv)
write.csv(data,filename.csv)

So !! We will be using the dataset "mtcars" for the rest of the lab it is one of the preloaded datasets in R studio and can be easily accessed using the variable mtcars.

Subsetting and Sorting -

Subsetting -

Data[row vector,column vactor]

If you want to keep the row vector or the column vector empty it will choose all the rows or columns in the subset respectively.

E.g

Suppose you want to get a subset of the dataset “mtcars” extracting only the first 10 rows with only mpg, cyl columns.

```
data <- mtcars[1:10,c("mpg","cyl")]
View(data)
```

Subsetting using logical vectors -

Data can also be subsetted using a logical vector in the place of rows or columns and that part of data is kept which evaluates to True for the condition.

E.g

Suppose you want to subset the dataset such that the mpg of the cars is greater than the mean mpg of all the cars?

```
print(mean(mtcars$mpg))
data <- mtcars[mtcars$mpg > mean(mtcars$mpg),]
View(data)
```

SORTING -

To sort a data frame in R, use the order() function. By default, sorting is ASCENDING. Prepend the sorting variable by a minus sign to indicate DESCENDING order.

E.g

Sort mtcars on the basis of HP in descending order. Then sort on the basis of MPG in ascending order.

```
data <- mtcars[order(-mtcars$hp),]
data <- mtcars[order(mtcars$mpg),]
```

SUMMARY -

The **summary()** function is used to summarise a data set or a vector normally, the information displayed normally depends on the type of vector or data set

```r
# Summary

summary(mtcars$mpg)
summary(mtcars$gear)

x <- as.character(mtcars$mpg)
summary(x)

y <- x=="21"
summary(c(y))
```

72:1 (Top Level) R Script

Console Terminal Jobs

R 4.1.1 · C:/Users/Aryan/Desktop/

```
> summary(mtcars$mpg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.40   15.43   19.20   20.09   22.80   33.90
> summary(mtcars$gear)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   3.000   4.000   3.688   4.000   5.000
>
> x <- as.character(mtcars$mpg)
> summary(x)
   Length     Class      Mode
       32 character character
>
> y <- x=="21"
> summary(c(y))
   Mode   FALSE    TRUE
logical      30       2
>
```

QUANTILES -

quantiles(vector1 , prob = vector2)
This function is to find the quantiles for a particular numerical vector and if vector 2 is not passed it takes the default value of 0,25,50,75 percent but vector 2 can be used to find custom quantiles.

E.g

Find 35 percent quantile of all the horse powers of mtcars

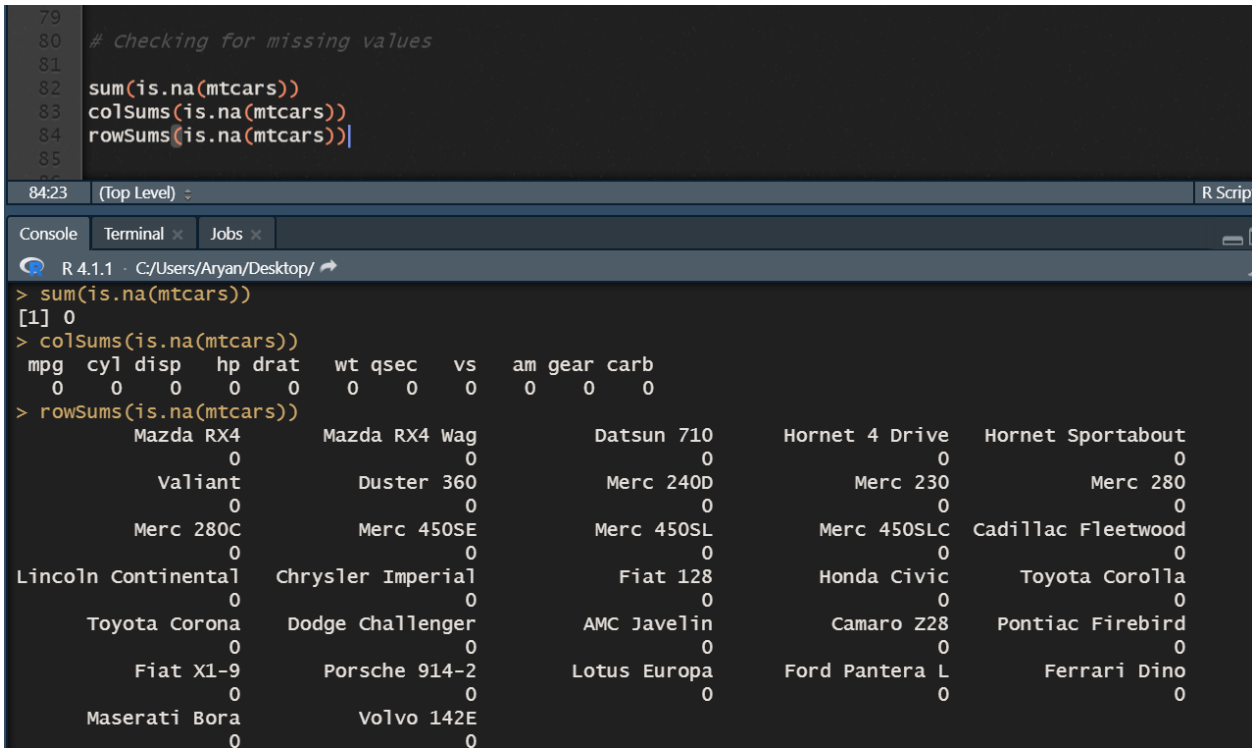
```
quantile(mtcars$hp , 0.35)
```

Checking for NA's -

is.NA(vector) <- returns True for every element which is na else returns false .

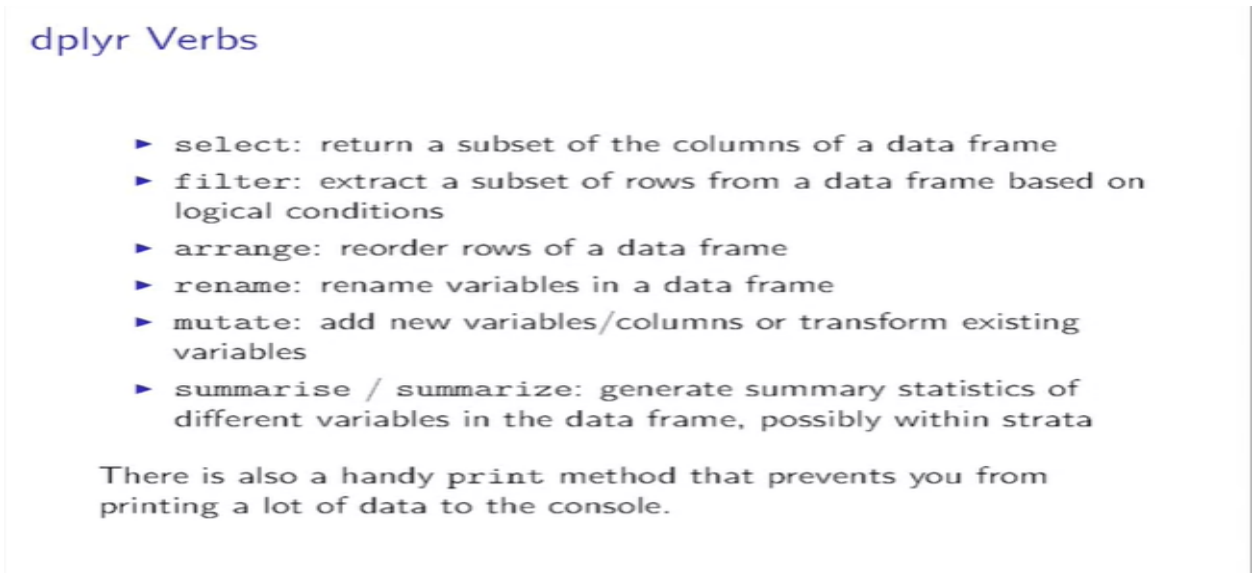
colSums(data) <- return Sum of columns in the dataset

rowSums(data) <- return Sum of rows in the dataset



Dplyr Package

Before starting load the dplyr package by using library("dplyr")



Practice Questions

Q1) Subset the dataset mtcars by taking every odd alternate row and(1,3,5,.....) and every even column(2,4..) then find the sum of the column for each column

Q2) Subset the dataset such that it contains only data of cars with 4 gears and then find the top 3 cars with the max mpg

Q3)Create a new dataset from mtcars having the mean value of mpg , max value of hp , median of wt for each value of gear

Q4)Find the 30 percent quartile of disp for each gear using dplyr package

Q5) Take the first 10 rows and the last 10 rows of the dataset and sort them in descending order of hp , then take the first 16 rows of the dataset . try to find the mean and median of disp for each gear and then use mutate to find the sum of both the values in a new column .

Q6) The first dataset contains the following details of students-

1) Id - id of students unique to students
2) Age
3) Gender

The second dataset contains the following details -

1) Id
2) Marks1
3) Marks2

Find the total marks of the top 10 students with the most marks and their age and gender and also mean, median of total marks, and age for both genders in the whole class ??

Code for datasets - copy and paste to get the data

```
set.seed(0)
data1 <- data.frame(id = 1:100 , age =
as.integer(rnorm(100,30,10)) , sex =
sample(c("Male","Female"),100,replace = T))

data2 <- data.frame(id = 1:100 , marks1 =
as.integer(rnorm(100,40,10)) , marks2 =
as.integer(rnorm(100,50,30)))
```