Assignment 1 (c)

Date : / /20

Title :- Basic statistical Commands on the dataset using R data exploration.

Problem statement :- To execute basic stastical Commands on the given dataset and explore the data to obtain useful information.

Pre-Lab :- A basic understanding of descriptive stastics will help in executing R command on dataset.

Theory :-

Statistics Commands in R :-

1. Mean :-

In R, a mean can be calculated on an isolated variable. Alternatively, a mean can be calculated for each of the variables in a dataset by using the name (DATAVAR) Command where DATAVAR is the name of variable containing the data. The syntax is

mean(x, trim-0, na.rm-FALSE, ...)

- x is input vector
- trim is used to drop some observation from both end of sorted vector.
- na.rm is used to remove the missing values from the input vector.

2. Median :-

The middle most value in a data series is called the median.

The syntax is
median (x, na.rm = FALSE)

Following is the description of parameter used-

- x is the input vector.
- na.rm is used to remove the missing values from the input vector.

3. Mode:-

The mode is value that has highest number of occurances in a set of data. Unlike mean & median mode can have both numeric & character data. R does not have a standard in-buit function to calculate mode of a data set in R. This function take the vector as input and gives the mode value as output.

We can calculate mode by making use of frequency of the dataset. Commands related to calculating mode by frequencies is as follows:

```
#Mode by frequencies
  table (mydata$country)#gives no of character of
   occurance of each values in vector
   # calculation of mode
 max (table (mydata$country)) # gives count of maximum
  occurance of a particular value.
 names (sort (table (mydata$country)))
 #gives value which has maximum occurances.
```

4. Standard Deviation-

Within R, standard deviation are calculated in same way as means. The standard deviation of a single variable can be computed with sd (VAR) command, where VAR is name of variable whose standard deviation you wish to retrieve. similarly a standard deviation can be calculated for each of the variable in a dataset by using sd (DATAVAR). is name of variable containing the data.

The syntax is :

ad (x, na.rm = FALSE)

Following is the description of parameters -

- x is the input vector.
- na.rm is used to remove the missing values from input vectors.

5 Range-

Minimum & Maximum

Keeping with the pattern a minimum canbe found on single-variable using the min (VAR) command.

The syntax is :

min(x).

Following is the description of parameters used-

- x is input vector

The maximum, via max(VAR), operates identically,

The syntax is .

max (x)

Following is the description of parameters used -

- x is the input vector.

However, in contrast to the mean & standard deviation functions, min(DATAVAR) or max(DATAVAR) or dataset not form each individual. Therefore, it is recomonded that minimums & maximum be calculated on individual variables, rather than entire datasets.

Range can be computed on single variable using the range (VAR) command which gives minimum & maximum value from the single variable.

The syntax is -

range (x)

Following is the description of parameter used-

- x is the input vector.

6. Percentiles -

6.1 values from percentiles (quantiles) :-

Given a dataset + a desired percentile, a corresponding value can be found using the following command

quantile (VAR, c(PROB1, PROB2, ----))

where, VAR refers to variable name and PROBL, PROB 2, etc. relate to probability values. The probabilities must between 0 & 1, therefore making then equivalent to decimal version of desired percentiles (i.e. 50% = 0.5)

6.2 Percentiles From values (Percentile Rank):-

In the opposite situation, where a percentile rank corresponding to a given value is needed one has to devise a custom method. To begin, consider the steps involved in calculating a percentile Rank

1. count the number of data points that are at or below the given below.
2. divide the total number of data points.
3. multiply by 100.

From the preceding steps, the formula for calculating a percentile rank can be derieved:
percentile rank = length (VAR [VAR <= VAL]) / len(VAR) *10

where VAR is the name of variable & VAL is the given value. This formula make use of length in two variables.
The first, length (VAR [VAR <= VAL]). counts the number of data points in a variable that are below given value. Note that the "<=" operator can be replaced with other combination of the <, > & = operators, supposing that the function were to be applied to different scenarios.
The second, length (VAR), counts the total number of data points in the variable. Together, they accomplish steps one & two of the percentile rank computation process.

## 7 : 5- Number Summary :-

A 5-Number summary is a set of descriptive statistics for summarizing a continuos univariate data set. It consist of data set's

- minimum
- 1st quartile
- median

- 3rd quartile
- maximum

This is a simple but very useful way of summarizing your data for several reasons.

- the median gives a measure of centre of data.
- the minimum & maximum give the range of data.
- the 1st & 3rd quartiles gives a sense of spread of the data, espesially when compared to the minimum, maximum and median.

The syntax is

fivenum(x)

- x is the input vector

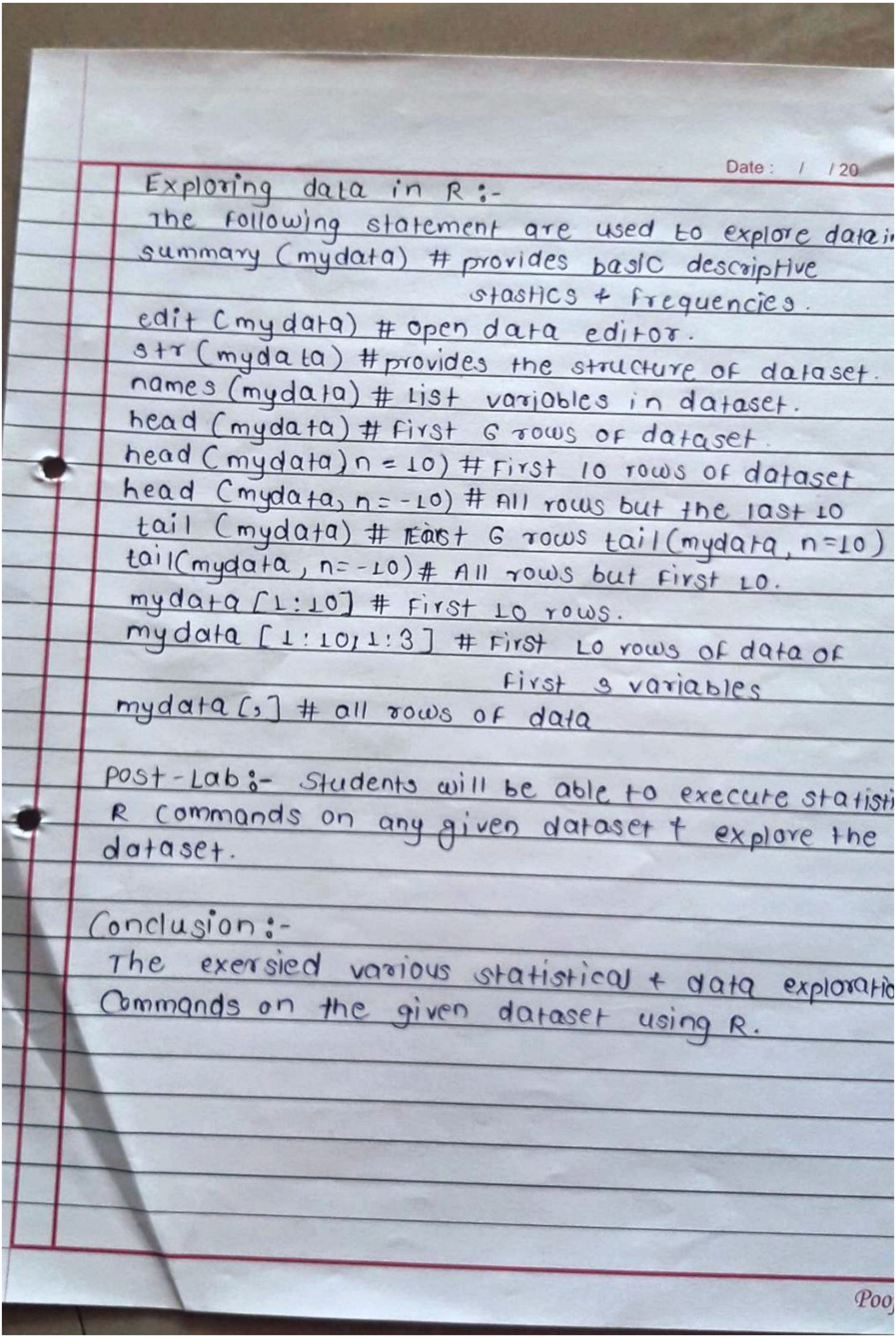
summary(x)

Following is the description of parameters used-

- x is the input vector.

perform the above statistical functions on dataset

| NO | SEX | AGE | NOOFCHILDREN | WEIGHT | HEIGHT |
|---|---|---|---|---|---|
| 1 | 0 | 57 | 1 | 65 | 158 |
| 2 | 1 | 70 | 3 | 100 | 175 |
| 3 | 0 | 45 | 0 | 71 | 162 |
| 4 | 0 | 38 | 2 | 58 | 164 |
| 5 | 0 | 25 | 1 | 81 | 170 |
| 6 | 1 | 50 | 4 | 68 | 172 |
| 7 | 1 | 61 | 0 | 85 | 179 |



Date : / /20

## Exploring data in R :-

The following statement are used to explore data i

summary (mydata) # provides basic descriptive stastics & frequencies.
edit (mydata) # open data editor.
str (mydata) # provides the structure of dataset.
names (mydata) # List variobles in dataset.
head (mydata) # First 6 rows of dataset.
head (mydata) n = 10) # First 10 rows of dataset
head (mydata, n = -10) # All rows but the last 10
tail (mydata) # Last 6 rows tail(mydata, n=10)
tail(mydata, n= -10) # All rows but First 10.
mydata [1:10] # First 10 rows.
mydata [1:10, 1:3] # First 10 rows of data of first 3 variables
mydata [,] # all rows of data

Post-Lab :- Students will be able to execute statisti R commands on any given dataset & explore the dataset.

## Conclusion :-

The exersied various statistical & data exploratio Commands on the given dataset using R.

Poo