Assignment No:- 5 Date / /20

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

Problem statement:-
To execute basic statistical commands on the given dataset & explore the data to obtain useful information.

Pre-Lab:-
A basic understanding of descriptive statistics will help in executing R commands on the given 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 variable in a dataset by using the mean(DATAVAR) command, where DATAVAR is the name of the variable containing the data.
The syntax is:
mean(x, train=0, na.rm = FALSE, ...)
- x is a i/p vector.
- train is used to drop some observations from both end of the sorted vector.
- na.rm is used to remove the missing values from the i/p vector.

Pooja

Date / / 20

2) Median:-

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

The syntax is:

median(x, na.rm = FALSE)

x is the i/p vector.

3) Mode:-

The mode is the value that has highest no. of occurances in a set of data. Unlike mean & median, mode can have both numeric & char data. R does not have a standard in-build function to calculate mode. So we create a user fun to calculate mode of data set in R.

#. Mode by frequencies:-

table(mydata$Country) # gives no of occurances of each value in the vector # calculation of mode.

max(table(mydata$country)) # gives count of max. occurances of a perticular value.

names(sort(table(mydata$country))) # gives the value which has max. occurances.

4) Standard Deviation:-

Within R, standard deviations are calculated in the same way as means. The std deviation of a single variable can be computed with the sd(VAR) command, where VAR is the name of the variable whose std deviation you wish to retrieve.

Pooja

Date : / /20

The syntax is:
sd(x, na.rm=FALSE)
- x is the i/p vector.
- na.rm is used to remove the missing values from i/p vector.

5) Range:-

Minimum & maximum:-

Keeping with the patteron, a minimum can be computed on a single vectable using the min(VAR) command:
The syntax is:
min(x)
- The max, via max(VAR), operates identically.

The syntax is:-
max(x).
- x is a i/p vector.

Range can be computed on a single variable using the range (VAR) command which gives min & max value from the single variable.
The syntax is:
range(x).
- x is i/p vector.

6) Percentiles:-

6.1) Values from Percentiles (Quantiles):-

Given a data-set & desired percentile, a corresponding value can be found using the following command:
quantile(VAR, c(PROB1, PROB2, ...)).

Pooja

Date / / 20

6.2) Percentile from value (Percentile Rank):-

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

1) Count the no. of data points that are at or below the given value.
2) divide by the total no of data pts.
3) multiple by 100.

percentile rank = length(VAR[VAR<=VAL])/length(VAR)*100.

where VAR is name of the variable & VAL is the given value.

This formula makes use of the length fun in 2 variations.

The 1st, length(VAR[VAR<=VAL]), counts the no. of data points in a variable that are below the given value. Note that the "<=" operator can be replaced with other combinations of the <,>,= operators. Supposing that the fun were to be applied to diff. scenarios.

7) 5- Number summary:-

A 5-number summary is a set of 5 descriptive statistics for summarizing a continuous univariate data set. It consists of the data set's.

- minimum
- 1st quartile
- median.

Pooja

Date / / 20

3rd quartile.

max

This is a simple but very useful way of summarizing your data for several reasons.
The median gives a measure of the centre of the data.
The min & max gives the range of data.
The 1st & 3rd quartiles give a sense of the spread of the data, especially when compared to min, max & median

syntax is:-
fivenum(x)
x is i/p vection.
summary(x)

Perform the above statistical functions on the dataset given below:

NO	SEX	AGE	NOOFCHILDREN	WEIGHT	HEIGHT
1	0	47	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.

Exploring Data in R:-
The following commands are used to explore the data in R

Pooja