**Assignment No. 1**

**PRN: 2019BTECS00067**

**Name: Vaibhav Kute**

**Aim:**

Design the data analysis tool (GUI) to perform the following task :

1. Data upload ( std. format like .csv, excel etc. ) and view.

2. Calculate and show the measures of central tendency for uploaded

data: mean, median, mode, midrange, variance, and standard deviation.

3. Calculate and show the dispersion of data: range, quartiles, interquartile range, and five-number summary.

4. Graphical display of the above calculated statistical description of data (provide the facility - UI form to choose different attributes from uploaded data set) :

a. Quantile plot

b. Quantile-quantile (q-q) plot

c. Histogram

d. Scatter plot

e. Boxplot

**Theory:**

**The measures of central tendency :**

In statistics, a central tendency (or measure of central tendency) is a central

or typical value for a probability distribution. It may also be called a center or location

of the distribution. Colloquially, measures of central tendency are often called

averages. The term central tendency dates from the late 1920s.

The most common measures of central tendency are the arithmetic mean, the

median, and the mode. A middle tendency can be calculated for either a finite set of

values or for a theoretical distribution, such as the normal distribution. Occasionally

authors use central tendency to denote "the tendency of quantitative data to cluster

around some central value.”

The central tendency of a distribution is typically contrasted with its dispersion or

variability; dispersion and central tendency are the often characterized properties of

distributions. Analysis may judge whether data has a strong or a weak central

tendency based on its dispersion.

The following may be applied to one-dimensional data. Depending on the

circumstances, it may be appropriate to transform the data before calculating a

central tendency. Examples are squaring the values or taking logarithms. Whether a

transformation is appropriate and what it should be, depend heavily on the data being

analyzed.

1) **Arithmetic mean or simply, mean** : the sum of all measurements divided by

the number of observations in the data set.

The arithmetic mean is the simplest and most widely used measure of

a mean, or average. It simply involves taking the sum of a group of numbers,

then dividing that sum by the count of the numbers used in the series. For

example, take the numbers 34, 44, 56, and 78. The sum is 212. The

arithmetic mean is 212 divided by four, or 53.

2) **Median** : The middle value that separates the higher half from the lower half

of the data set. The median and the mode are the only measures of central

tendency that can be used for ordinal data, in which values are ranked

relative to each other but are not measured absolutely.

Median is the middle number in a sorted list of numbers. To determine the

median value in a sequence of numbers, the numbers must first be sorted, or

arranged, in value order from lowest to highest or highest to lowest. The

median can be used to determine an approximate average, or mean, but is

not to be confused with the actual mean.

If there is an odd amount of numbers, the median value is the number that is

in the middle, with the same amount of numbers below and above.

If there is an even amount of numbers in the list, the middle pair must be

determined, added together, and divided by two to find the median value.

The median is sometimes used as opposed to the mean when there are

outliers in the sequence that might skew the average of the values. The

median of a sequence can be less affected by outliers than the mean.

Median Example

To find the median value in a list with an odd amount of numbers, one would

find the number that is in the middle with an equal amount of numbers on

either side of the median. To find the median, first arrange the numbers in

order, usually from lowest to highest.

For example, in a data set of {3, 13, 2, 34, 11, 26, 47}, the sorted order

becomes {2, 3, 11, 13, 26, 34, 47}. The median is the number in the middle {2,

3, 11, 13, 26, 34, 47}, which in this instance is 13 since there are three

numbers on either side.

To find the median value in a list with an even amount of numbers, one must

determine the middle pair, add them, and divide by two. Again, arrange the

numbers in order from lowest to highest.

For example, in a data set of {3, 13, 2, 34, 11, 17, 27, 47}, the sorted order

becomes {2, 3, 11, 13, 17, 27, 34, 47}. The median is the average of the two

numbers in the middle {2, 3, 11, 13, 17, 26 34, 47}, which in this case is fifteen

{(13 + 17) ÷ 2 = 15}.

3) **Mode** : The mode is the value that appears most frequently in a data set. A

set of data may have one mode, more than one mode, or no mode at all.

Other popular measures of central tendency include the mean, or the average

of a set, and the median, the middle value in a set.

Examples of the Mode

For example, in the following list of numbers, 16 is the mode since it appears

more times in the set than any other number:

3, 3, 6, 9, 16, 16, 16, 27, 27, 37, 48

A set of numbers can have more than one mode (this is known as bimodal if

there are two modes) if there are multiple numbers that occur with equal

frequency, and more times than the others in the set.

3, 3, 3, 9, 16, 16, 16, 27, 37, 48

In the above example, both the number 3 and the number 16 are modes as

they each occur three times and no other number occurs more often.

If no number in a set of numbers occurs more than once, that set has no

mode:

3, 6, 9, 16, 27, 37, 48

A set of numbers with two modes is bimodal, a set of numbers with three

modes is trimodal, and any set of numbers with more than one mode is

multimodal.

4) **Geometric Mean** : The nth root of the product of the data values, where there

are n of these. This measure is valid only for data that are measured

absolutely on a strictly positive scale.

5) Harmonic mean : The reciprocal of the arithmetic mean of the reciprocals of

the data values. This measure too is valid only for data that are measured

absolutely on a strictly positive scale.

6) Weighted arithmetic mean : An arithmetic mean that incorporates weighting to

certain data elements.

7) Interquartile mean : A truncated mean based on data within the interquartile

range.

8) Midrange : Midrange in layman terms is the middle of any data set or the

simply the average, mean of the data. A midrange is a statistical tool which is

also known as the measure of center in statistics. Along with the existence of

the midrange formula means, medium, average, mode, and range are also

known as the measure of central tendency. The midrange of the data set is

simply the value between the biggest value and the lowest value. In order to

find the midrange of the data set the value is then divided by 2 after summing

the lowest value present in the data set with the highest value present in the

data set.

9) Variance : The term variance refers to a statistical measurement of the spread

between numbers in a data set. More specifically, variance measures how far

each number in the set is from the mean and thus from every other number in

the set. Variance is often depicted by this symbol: σ2. It is used by both

analysts and traders to determine volatility and market security. The square

root of the variance is the standard deviation (σ), which helps determine the

consistency of an investment’s returns over a period of time.

10) Standard Deviation : A standard deviation is a statistic that measures the

dispersion of a dataset relative to its mean. The standard deviation is

calculated as the square root of variance by determining each data point's

deviation relative to the mean. If the data points are further from the mean,

there is a higher deviation within the data set; thus, the more spread out the

data, the higher the standard deviation.

11) Range : The difference between the largest value and the smallest value.

12) Quartiles : A quartile is a statistical term that describes a division of

observations into four defined intervals based on the values of the data and

how they compare to the entire set of observations.

Suppose the distribution of math scores in a class of 19 students in ascending

order is:

59, 60, 65, 65, 68, 69, 70, 72, 75, 75, 76, 77, 81, 82, 84, 87, 90, 95, 98

First, mark down the median, Q2, which in this case is the 10th value: 75.

Q1 is the central point between the smallest score and the median. In this

case, Q1 falls between the first and fifth score: 68. (Note that the median can

also be included when calculating Q1 or Q3 for an odd set of values. If we

were to include the median on either side of the middle point, then Q1 will be

the middle value between the first and 10th score, which is the average of the

fifth and sixth score—(fifth + sixth)/2 = (68 + 69)/2 = 68.5).

Q3 is the middle value between Q2 and the highest score: 84. (Or if you

include the median, Q3 = (82 + 84)/2 = 83).

Now that we have our quartiles, let’s interpret their numbers. A score of 68

(Q1) represents the first quartile and is the 25th percentile. 68 is the median

of the lower half of the score set in the available data—that is, the median of

the scores from 59 to 75.

Q1 tells us that 25% of the scores are less than 68 and 75% of the class

scores are greater. Q2 (the median) is the 50th percentile and shows that

50% of the scores are less than 75, and 50% of the scores are above 75.

Finally, Q3, the 75th percentile, reveals that 25% of the scores are greater

and 75% are less than 84.

13) Interquartile Range : The interquartile range is a measure of where the

“middle fifty” is in a data set. Where a range is a measure of where the

beginning and end are in a set, an interquartile range is a measure of where

the bulk of the values lie. That’s why it’s preferred over many other measures

of spread when reporting things like school performance or SAT scores.

The interquartile range formula is the first quartile subtracted from the third

quartile:

IQR = Q3 – Q1.

14) Five-number summary : The five number summary includes 5 items:

● The minimum.

● Q1 (the first quartile, or the 25% mark).

● The median.

● Q3 (the third quartile, or the 75% mark).

● The maximum.

The five number summary gives you a rough idea about what your data set

looks like. for example, you’ll have your lowest value (the minimum) and the

highest value (the maximum). Although it’s useful in itself, the main reason

you’ll want to find a five-number summary is to find more useful statistics, like

the interquartile range, sometimes called the middle fifty.

**Results: snapshots**

Result snapshots are included in other pdf files.

**Conclusion:**

Using the streamlit for UI and using python language this task is done.

Streamlit is an open-source app framework for Machine Learning and Data Science teams. Creates beautiful web apps.