International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 1, ISSN No. 2455-2143, Pages 278-284
Published Online May 2021 in IJEAST (http://www.ijeast.com)

SELECTION OF OPTIMAL SUMMARY STATISTICS FOR DATA ANALYSIS

Vanshika Lamba
Department of Mathematics,
University Institute of Sciences,
Chandigarh University, Gharuan,
Mohali, Punjab-140413, India

Abstract: Data analysis is the core part which needs to be done over the data in order to gather its characteristics for further specifications and estimations. But achieving the goal of extracting the maximum useful characteristics is the main barrier in the path of any organization. Data analysis plays an important role in the success of organization as it helps in proper decision making. And the best decision comes out by analyzing the past information, their present scenario and future impacts. But most of the information is extracted in the numerical form from the data set collected. Therefore, we need to select some proper summary statistics for the data exploration purpose. For eg - mean, median, mode, etc. This paper focuses on the classes of summary statistics to be used for the data analysis and how important is its use in the data exploration. The paper majorly concentrates on the measure of location and a brief idea about measure of dispersion and how measure of location is related to measure of spread.

I. INTRODUCTION

Rapid development in the data collection, data processing and data storage technologies has made the organizations a center of accumulated data in large amounts. Data analysis deals with the steps which involves the gathering of data using proper applications or tools and using those tools for exploring the data in order to discover some patterns in it. By recognizing the patterns within the data the goal of discovering useful information, suggesting conclusions and supporting decision-making is achieved.

Data analysis is a mechanism of obtaining large and unstructured data from different sources and converting it into useful information that will help us in answering many questions, testing the hypothesis, some important decision-making and disproving the theories [1].

Initially, a preliminary investigation of the data is done in order to understand its characteristics, known as data exploration. The quantities that capture the specific characteristics of the huge data set of values are termed as summary statistics. Summary statistics represents the characteristics of the data set by a single number or a set of numbers. For example – Mean, Mode, Median, standard deviation, etc. are the quantities used for the purpose of data exploration. Examples of summary statistics from our daily lives are the average household expenses or the fraction of college students who complete their degree in 3 years, etc. [1, 2].

Summary Statistics and its use in data analysis:

Summary statistics summarizes the large volumes of data and provides the information about the values present in the data. So, in data analysis selecting the correct summary statistics is very important aspect and it should include only those quantities which will give us a quick and simple description of the data.

Following are the evidences to show the selection of summary statistic dependency on the data available:

1) Measures of central tendency- A measure of central tendency attempts to represent the set of available data by using only a single number that would indicate the center of the data set. These measures of central tendency are therefore, also classed as summary statistics [2].

II. IMPORTANCE IN DATA ANALYSIS:

1) Finding an average for the data set gives us a single value as the representative value for the whole data set and hence makes the further analysis of data much easier as the characteristics of the data set can be described using the single representative value.

2) Since the accumulated data is very vast, so by calculating the average all the figures of data set are converted into a single figure thus condense form of data is obtained.

3) In many situations we need to compare the different data sets, but comparing all the data sets by individual values will become so difficult. Hence, we again need a representative figure from these data sets which are obtained by the measures of central tendency. So, these averages make the comparison of data sets easy.

4) Various other techniques used in data analysis depend upon the averages or the measures of central tendency, therefore the averages are the measures of the first order.

a) Mode and the frequency: Suppose a given data set contains the unordered categorical values, so further to characterize the values we can only find the frequency with which each value occurs for a particular set of data. Nothing much can be done with such data other than finding the frequency for each value.

Let a categorical attribute, **x**, from the given data set can take values {**a_1, a_2, a_3, - - - a_i - - - , a_n**} from a set of **k** objects. Then, the frequency of value a_i is defined as,

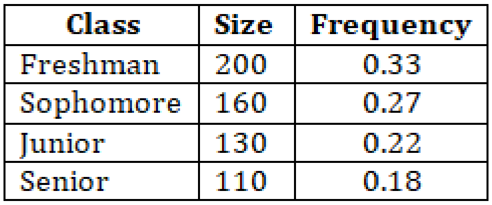
Frequency (**a_i**) = (number of objects with value **a_i**)/**k**

The mode of attribute **x** will be the value that has the maximum frequency.

E.g. – Suppose we are given the data of students of a college categorized into an attribute called class. The attribute class can take values as freshman, sophomore, junior or senior. Freshman represents the students who earned fewer than 32 semester hours(SH), sophomore who earned at least 32 SH but fewer than 64 SH, junior who earned at least 64 SH but fewer than 96 SH and Senior who earned at least 96 SH.

For analysis of such data set, the useful summary statistic can be the mode. Table 1. presents the total number of students for each value of the class attribute and their frequencies.

Table. 1



Class	Size	Frequency
Freshman	200	0.33
Sophomore	160	0.27
Junior	130	0.22
Senior	110	0.18

As indicated from the table the highest frequency (0.33) is for the value freshman and hence the mode of the class attribute is freshman. By analyzing this data and evaluation, estimations aboutthe number of dropouts can be done or anindication towards larger freshman class than the usual is also depicted [1].

Points about mode:

1) The mode is the rarely used average as it is best suited for dealing with the categorical (nominal or ordinal) data only because it do not require the data to be in a meaningful order.

2) If in a data set the frequencies of the values are equal then the concept of mode is not useful.

E.g. **-** In iris data set, the class attribute with the three values **-** the three types of flowers have the same frequency so the notion of a mode is not interesting for this data set.

3) If the data set is continuous, mode is not useful in this case also because a single value may not occur more than one time but still in some situations mode may indicate some information about the nature of values or the presence of the missing values.

E.g.**-** If the heights of 30 people is measured in mm then there are very rare chances for the height to repeat, but if it is measured in dm, then there are possibility for the repetition of the heights. So, in such type of scenario finding mode is useful only when height is measured in dm.

b) Mean and Median: When the data set is continuous, then for such data set the most widely used summary statistics are the mean and the median. Mean and median tells the location of the set of values and hence they are also known as measures of location.

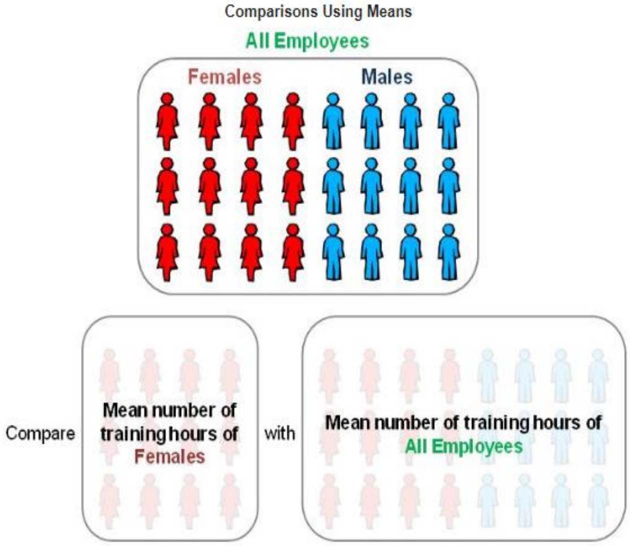
Suppose there are **m** objects and an attribute **x**, which takes the values as **b_1, b_2, - - - - , b_m.** As an eaxample we can consider the values to be the heights of **m** boxes. Let {**b_1, b_2, - - - -, b_m**} be the set of values after being sorted in ascending order. Therefore, **b_1**= min(**x**) and **b_m**= max(**x**).
The mean is calculated by adding all the values for m objects and dividing the sum by the no. of objects, i.e,
Mean = (**b_1 + b_2 - - - - + b_m**)**/m**

The median calculation do not represent any specific formula, but to summarize we can say that, if the number of values are odd then the median represents the middle value and if the number of values in the set are even the the median represents the average of the two middle values.

Thus, for 11 values – the median is **b_5 and** for 12 values – the median is ½(**b_5 + b_6**) [1, 3].

The mean serves as the yard stuck for all observations: Suppose we are interested to know that how many hours an employee spends at training in a year on an average. Therefore, since the data is continuous and symmetrical so we will find the mean training hours of the different groups of the employees. The calculated mean is used to compare with an employee annual training hours in order to judge that he also has the comparable opportunities for training as the other employees.

Eg – comparing the average annual training hours for women against all employees average annual training hours.


Comparisons Using Means
All Employees
Females
Males
Compare
Mean number of training hours of Females
with
Mean number of training hours of All Employees

source: https://stats.mom.gov.sg/SL/Pages/Mean-Uses.aspx

The calculated mean can be useful for the further analysis, for eg- Suppose we want to check that if the company increases the wage of its employees by 4% then how much would it cost to the company. This can be evaluated easily by just knowing the mean wage of the company and the no This can be done even if we only know the mean wage of the company and the number of employees without knowing the wages of each one of the employees individually. The mean is just the sum of all the values divided by the number of values, therefore, the total cost of a 5% wage increment for all workers sums up to 5% of the mean multiplied by the number of workers [4].

The median can form the basis for grouping the data- Suppose we want to divide each of the classes in a particular department into two groups in order to conduct the offline classes safely for the students so that social distancing is maintained in the class. Firstly we need to select a factor on basis of which the groups can be formed. For e.g. – We choose the roll numbers of the students which are assigned to students according to the alphabetical order of their names, as the factor of forming the groups. So, we will just sort the students in ascending order according to their roll numbers and find the median of the roll numbers and hence divide the students into two groups as the median divides the sorted data into two equal divisions.

III. POINTS ABOUT MEAN AND MEDIAN-

1) The mean is considered to be the best for use when the data is continuous as well as symmetrical, or in other words we can say that the data is normally distributed. It is generally used for the ratio and interval data . However, its use depends only on the kind of information to be analyzed.

2) Sometimes mean is used to refer the middle value of the set of values, but it is correct only for the symmetric data set. If the data is skewed then the notion of mean representing the middle value fails and hence median is the best use case for such data.

3) If we have a data set containing outliers (a data point that is different from the other values) then it is necessary to remove it or use a measure which is not affected by the presence of the outliers. So, in this case the mean is considered to be sensitive and hence the median provides the effective estimate of the middle data. In such cases when outliers are present in data set we can use mean only when the outliers are not highly influential otherwise we replace the outliers with the median value only.

E.g.- Suppose in an organization as a part of data analysis, we wish to know the average salary of the employees working in a particular department. For simple evaluation we take salaries of 10 people – Two of them earn $16000 each, three of them are paid $17000 each, and three make $18000 apiece. Another works for a stingy boss and only earns $10,000. One person (the stingy boss) gets a very generous salary of $50,000. These are the outliers present in the data set.

For finding mean firstly, we add up all the salaries, $10000 + $16000 + $16000 + $17000 + $17000 + $17000 + $18000 + $18000 + $18000 + $50000. Total = $197000, now we divide the total by number of employees (10) and get a mean salary of $19700. As we can see that the mean salary is more than the most people in this department earn.

But in such case the median salary is not very likely to be affected by the outliers. If we sort the salaries as - {$10000, $16000, $16000, $17000, $17000, $17000, $18000, $18000, $18000, $50000}. The median of this set will be the average of the two middle values, i.e.,
½($17000 + $17000) = $17000. So, this salary is more in line with the majority earnings of the employees as compared to the mean salary. Hence, the median salary is more useful for such data sets.

Missing Data – When the data is accumulated then, there are possibilities of the presence of missing data. Missing data is referred to those values which are not present in the data set but might prove to be useful if observed. Therefore, in analyzing the data handling of missing data is very important in order to extract the accurate useful information for the further analysis of data. In case, the missing data is not handled properly, we might end up with an inaccurate inference about the data which further will lead to the wrong results [5].

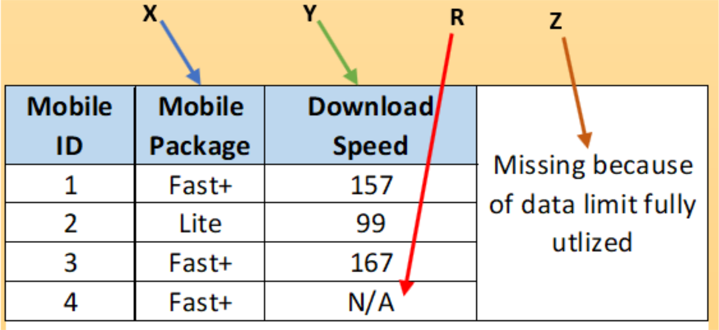
Handling missing data – While handling the missing data we can either leave the missing data or do some imputations to replace the missing values which will be a useful value in analysis. If we have a very small number of missing values, then we can drop or omit those values from our analysis of data. In statistical analysis, if the number of missing values cases is less than 5%, then we can easily drop them.

In case of multivariate (more than one variable) analysis , if the number of missing values is large, then the better option is to drop all those cases rather than imputing in order to replace them. But in case of univariate (single variable) analysis, doing the imputation is useful as it reduces the amount of bias (bias means that a small sample selected for analysis do not represent the entire data set from which it is selected) present in the data, when the values are randomly missed values [5,6].

There are three categories of missing data –

1) Missing completely at random (MCR): When the missing is not related to the observation being studied, then the data is said to be in the category of missing completely at random. For this category, the reasons of missing data are the external parameters like failure of equipment, the sample is damaged in lab, the sample is lost by the questionnaire, or the sample is not satisfactory for the analysis. Therefore, it is an ideal and unreasonable assumption. In this case, the sample is likely to represent the entire population. The advantage of such data is analysis will not be biased.

E.g. **–** Consider an example of mobile data, in collecting the samples we found that one variable of sample have a missing value but it is not due to the data set variables but due to an external reason.



Mobile ID	Mobile Package	Download Speed	Missing because of data limit fully utlized
1	Fast+	157	
2	Lite	99	
3	Fast+	167	
4	Fast+	N/A	

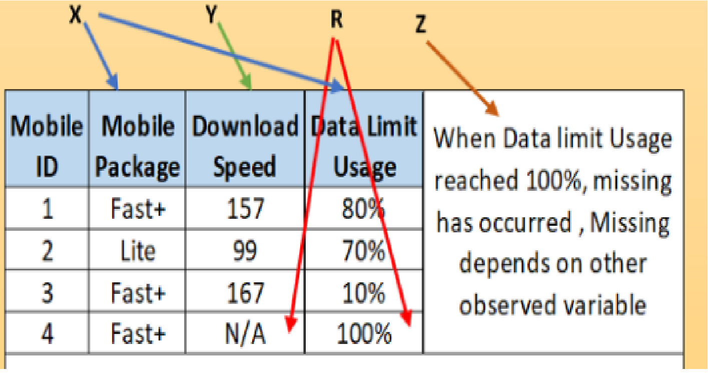
source: https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184

So, as we can see from the above fig that the reason of missing value is the use of full data limit, it has nothing related with the variables **x** – Mobile package and **y** – download speed. But it may be predicted from these variables.

2) Missing at random (MAR): When the missing data falls in this category, then in the analysis we mean that the missing data on a partly missing variable (say y) is related to some of the other variables (say x) which are completely observed variables but the missing data on partly missing variable is not related to the values of y itself. In other words, we can say that the missing data is not exactly related to the missing information.

For e.g. – Suppose a student was not able to give his examination because he was not well, so the thing that he is not well can be predicted from the data we have about his health. But this prediction is not related to the prediction which we would have made if the student had been well. Therefore, it clearly depicts that the MAR values cannot be ignored, they might have some other relations (say z).

Continuing with the example of mobile data, now we have different samples, and we found that the missing value on a partly missing variable depends on other completely observed variable.



Mobile ID	Mobile Package	Download Speed	Data Limit Usage	When Data limit Usage reached 100%, missing has occurred , Missing depends on other observed variable
1	Fast+	157	80%	
2	Lite	99	70%	
3	Fast+	167	10%	
4	Fast+	N/A	100%	

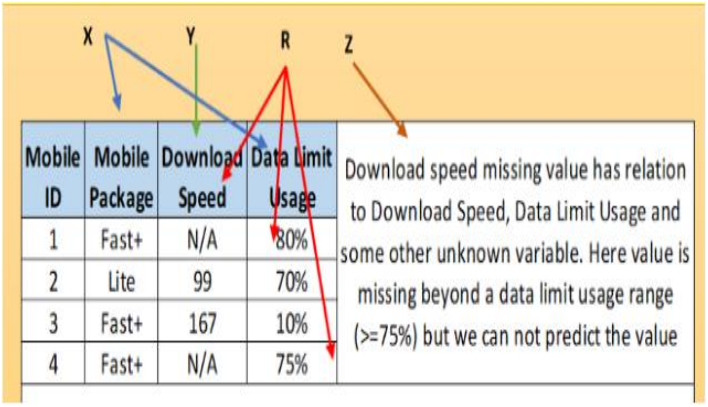
source: https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184

Looking at the above figure we can clearly infer that the missing value in partly missing varible y depends on the completely observed variable x and also it has relation with another varible z. In this case the missing value can be predicted by using the observed values.

3) Missing not at random (MNAR): If the missing data do not falls under the categories of MCAR or MAR then it is said be in the category of missing not at random. In this case the reason of missing data is exactly the data which is missing. We can say that the missing data on a partly missing variable (say y) is related to the missing data itself (y) and it can have relations with other variables (say x and z) as well.

For eg – If a student did not take the German Proficiency test because of his poor german language skill or someone did not opted for singing competition because he is not good in that skill.

Continuing with the example of mobile data and observing some other samples we found that the missing value is dependent on the missing variable itself and other variables as well.



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	N/A	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	75%

Download speed missing value has relation to Download Speed, Data Limit Usage and some other unknown variable. Here value is missing beyond a data limit usage range (>=75%) but we can not predict the value

source: https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184

So, it is obvious from the above figure that the missing value depends on variable y itself and on variable x and z also. In such case the prediction of missing data gets difficult.

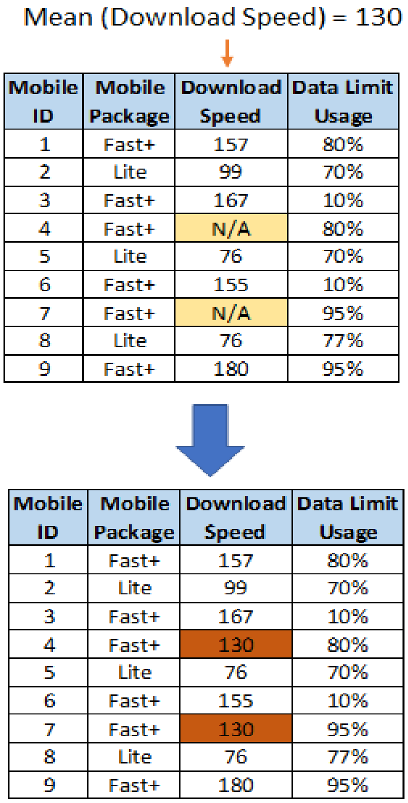
Using summary statistics in imputing the missing values:
The goal of imputation technique is to replace the missing values by the estimated result obtained on applying summary statistics – mean, median and the mode.

The Mean - This method is useful only when the variables have random missing data and it is symmetrical. Then we replace the missing value with the mean of that variable. The advantage in doing so is the mean of the sample remains the same even after the replacement of the missing values.

The Median - In case the variable have missing values with greater inequality then the mean substitution may lead to inconsistent bias and hence we can use the median for the replacement of missing values.

The Mode - We can also replace the missing values with the most frequent value, as it is a very likely occurrence.

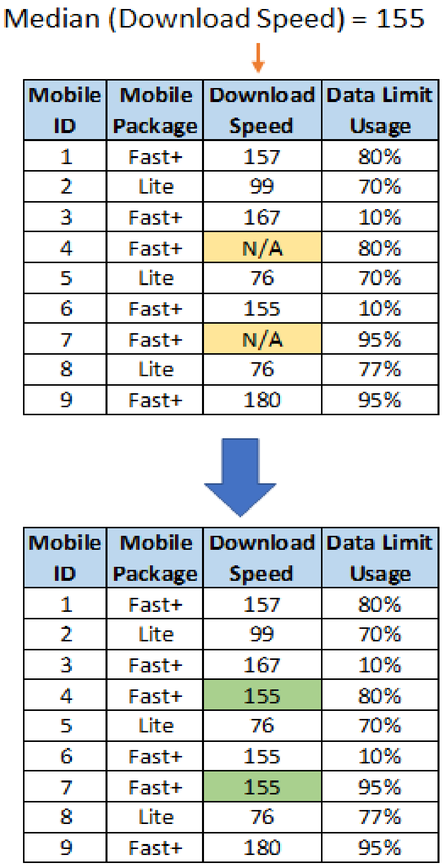
Following three figures shows the imputation technique to replace the missing values with the summary statistics measures and hence retaining all the data for the analysis purpose [6, 7].



Mean (Download Speed) = 130

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

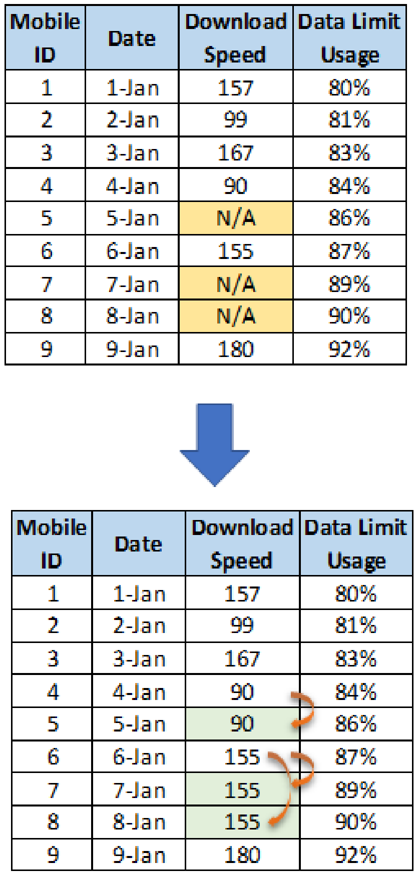
Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	130	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	130	95%
8	Lite	76	77%
9	Fast+	180	95%



Median (Download Speed) = 155

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	155	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	155	95%
8	Lite	76	77%
9	Fast+	180	95%



Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	N/A	86%
6	6-Jan	155	87%
7	7-Jan	N/A	89%
8	8-Jan	N/A	90%
9	9-Jan	180	92%

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	90	86%
6	6-Jan	155	87%
7	7-Jan	155	89%
8	8-Jan	155	90%
9	9-Jan	180	92%

source: https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184

2) Measure of Spread: A measure of spread is also known as measure of dispersion and it is used to measure the variation in the sample or the complete data set. Range and variance are the commonly used summary statistics for continuous data in order to measure the spread of a set of values. The relationship of measure of spread with the measure of central tendency makes it easier to understand how well the measures of central tendency like mean represents the entire data set so that further analysis can be continued. The mean is much more representative when the spread of values in the data set is small as compared to when the spread of values in the data set is large because the large spread means that there are large inequalities among the values. Variation or the dispersion of data is defined as the extent upto which the numerical data has a tendency to spread. Dispersion also helps us in identifying the heterogeneity (more scattered) and the homogeneity (less dispersion) in the data sets [8].

a) Range - To summarize the data using range, we need to sort the data in ascending order and then the difference between the lowest value and largest value gives the range of the data set. It just gives the idea about how widely the observations are spread out but it cannot tell anything about how the other data points lie.

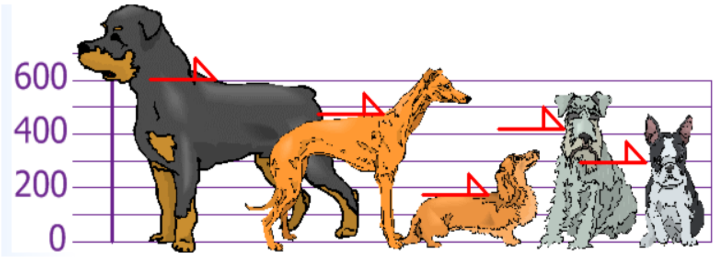
Range = Max. value – Min. value

E.g. – Suppose, we have the data set of the age of the students in the class and we wish to infer the range in which the age lies from that data set. First, we need to sort it in ascending order and then find the difference between the lowest and largest value [8, 9].
A small sorted sample of data set is - {19,19,19,20,20,21,21,22,22,22,23,23,23,23,24}

Max. value = 24 and Min. value = 19
Range = 24 – 19 = 5.

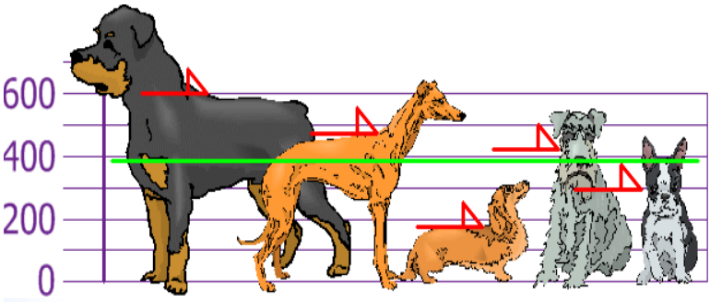
b) Variance – The range is sensitive to outliers and hence, we need some other measure like standard deviation. The square of standard deviation (s.d) is known as the variance of the data set or standard deviation is known as the square root of the variance. Variance is calculated as follows:
1) Find the mean of data points
2) Subtract the mean from each data point and square the result
3) Find the average of the squares calculated in step 2.

Eg – Suppose we want to categorize the dogs into their respective breeds based upon their heights, so as we need to find the variability in their heights. The best measure for such data will be standard deviation.
There are 5 dogs with the heights(at shoulders) in mm – {300, 430, 170, 470, 600 }



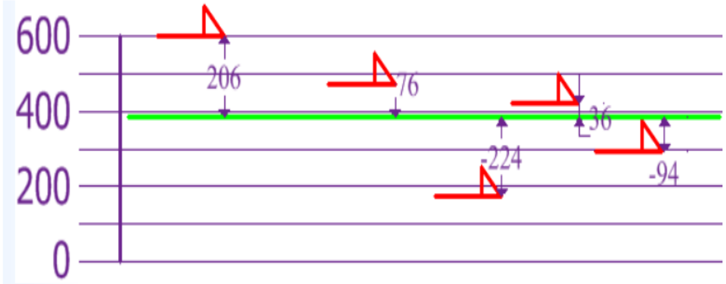
source: https://www.mathsisfun.com/data/standard-deviation.html

In order to categorize them, we will find the standard deviation –
1) Mean = $^1/_5$ (300 + 430 + 170 + 470 + 600) = 394.
The green line in the following figure depicts the average height to be 394mm.



source: https://www.mathsisfun.com/data/standard-deviation.html

2) Find the deviation of data points from the mean.



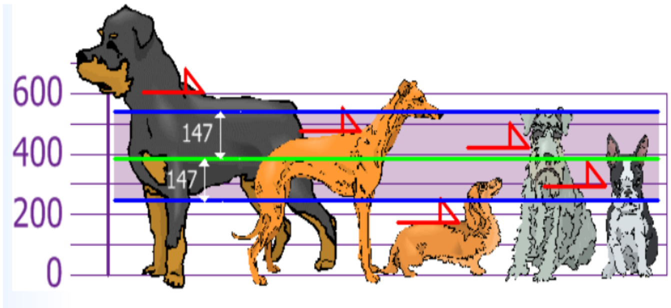
source: https://www.mathsisfun.com/data/standard-deviation.html

3) Calculating variance and standard deviation –

Variance = $^{1}/_{5}(206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2) = 21704$

Standard deviation = $\sqrt{21704} = 147.32 \approx 147$mm

So, using this summary statistic we can infer which heights are within one s.d of mean. The section enclosed with the blue line in the following figure represents this information. Hence, it is useful in estimating which dogs are of normal height, extra small height or extra large height.



source: https://www.mathsisfun.com/data/standard-deviation.html

Hence, we conclude that the normal dogs lies in the enclosed section, Rottweiler are extra large and Dachshunds are extra small dogs [10].

IV. CONCLUSION

Analyzing data is an important step which is going to set up the conditions for the future processing of the data and the required modeling for the predictions. There can be many best measures of summary statistics with respect to the data analyzed, but there is no technique to be considered as the only best measure. The reason behind this is that the data accumulated always contains different types of data, it may be nominal, continuous, symmetrical, or the data contains outliers, or missing data is present, etc. So, using a technique whether the mean, median or mode, or some other measure of dispersion, is dependent on the type of data available. Another factor responsible for the selection of a summary statistics is the information that we will be extracting from the data. Also measure of dispersion such as range and variance or standard deviation tells us about how well the measure of location is representing the data set or a standard way in which the data can be compared. In case the data set is normally distributed then the summary statistics - mean, median and mode have the same value.

V. REFERENCES

1) Tan Pang-Ning, Steinbach Michael, Kumar Vipin, 2006 “Introduction to Data Mining”, First edition, Pearson Education ISBN 0-321-42052-7
2) Gupta S.C., Kapoor V.K., 2014 “Fundamental of Mathematical Statistics”, 7th edition, Sultan Chand & Sons, New Delhi, ISBN-13:978-8180545283
3)https://towardsdatascience.com/descriptive-statistics-f2beeaf7a8df
4)https://stats.mom.gov.sg/SL/Pages/Mean-Uses.aspx
5)https://www.statisticssolutions.com/missing-values-in-data/
6)https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/
7)https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184
8)https://www.statisticshowto.com/measures-of-spread/
9)https://statistics.laerd.com/statistical-guides/measures-of-spread-range-quartiles.php
10)https://www.mathsisfun.com/data/standard-deviation.html