Name: Dhruva Rakesh B

Roll no.: 243514

Date: 25/04/2024

Advance Analytics Assignment 1

Q. What is data analysis?

Data analysis is the process of cleaning, analyzing, interpreting, and visualizing data using various techniques and business intelligence tools. Data analysis tools help you discover relevant insights that lead to smarter and more effective decision-making.

Q. What is data visualization? Explain all charts of visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Q. Write formula of mean

$$\text{Mean} = \frac{\text{Sum of All Data Points}}{\text{Number of Data Points}}$$

Q. What is advantage and limitation of mean?

Mean is easy to calculate and can be easily represented in some graphs. ADVERTISEMENT. However, it is also important to note that means is susceptible to skewness and outliers, which may not be the most reliable measure in some cases.

Q. What is advantage and limitation of median?

It is easily understandable and computable. It is well defined as an average. It is not influenced by extreme values in a data set and is also independent of the dispersion and range of data. Median can be utilized in the case of computing frequency distribution with open-ended classes.

Q. What is advantage and limitation of mode?

The mode is easy to understand and calculate. The mode is not affected by extreme values. The mode is easy to identify in a data set and in a discrete frequency distribution. The mode is useful for qualitative data. The mode can be computed in an open-ended frequency table.

Q. What is continuous variable, what is discrete variable explain with 5 examples

A discrete variable only allows a particular set of values, and in-between values are not included. If we are counting a number of things, that is a discrete value. A dice roll has a certain number of outcomes, and nothing else (we can roll a 4 or a 5, but not a 4.6). A continuous variable can be any value in a range. Usually, things that we are measuring are continuous variables, because it can be any value. The length of a car ride might be 2 hours, 2.5 hours, 2.555, and so on.

Discrete Variable Examples

- The number of workers in an office
- The number of steps you take in a day
- The number of babies born each day

Continuous Variable Examples

- The time it takes for office employees to commute to work
- The distance you walk in a day
- The weight of newborn babies

Q. What is difference between sample and population explain with 5 examples

A population is the entire group that you want to draw conclusions about. A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.

Example: Collecting data from a population

A high school administrator wants to analyze the final exam scores of all graduating seniors to see if there is a trend. Since they are only interested in applying their findings to the graduating seniors in this high school, they use the whole population dataset.

Example: Collecting data from a sample

You want to study political attitudes in young people. Your population is the 300,000 undergraduate students in the Netherlands. Because it's not practical to collect data from all of them, you use a sample of 300 undergraduate volunteers from three Dutch universities who meet your inclusion criteria. This is the group who will complete your online survey.

Q. What is descriptive, diagnostic, predictive, prescriptive data analytics?

Descriptive Analytics: is concerned with data summarization, graphs/charts, and Tables. It processes raw data into information.

Example: Mean, Median, Mode, Kurtosis, Skewness

Diagnostic: Why it is happening?

Predictive Analytics: Predictive modeling use mathematical, spreadsheet, and statistical models, and address questions such as:

- Impact of Advertisement on sales.
- Will a given process change reduce cycle time.

Prescriptive Analytics: uses optimization to identify the best alternatives to minimize or maximum objectives.

Q. Find mean, median, mode of following-

1. 89,6,90,34,65,234,8,3000,7,567,6,2,45,20

	2	
	6	
	6	
	7	
	8	
	20	
	34	
	45	
	65	
	89	
	90	
	234	
	567	
	3000	
Mean:	298.07	ROUND(AVERAGE(M2:M15),2)
Median:	39.5	MEDIAN(M2:M15)
Mode:	6	MODE.SNGL(M2:M15)

2. 5677,60,22,34,6500,23, 869,67,900,1,2,6,1,70,1

	1	
	1	
	1	
	2	
	6	
	22	
	23	
	34	
	60	
	67	
	70	
	869	
	900	
	5677	
	6500	
Mean:	948.87	ROUND(AVERAGE(Q2:Q16),2)
Median:	34	MEDIAN(Q2:Q16)
Mode:	1	MODE.SNGL(Q2:Q16)

Q. What are data analysis steps?

Decide on the objectives

Set measurement priorities

Data collection

Data cleaning

Analysis of data

Interpreting the results

Q. Explain following terms with example?

- Nominal Scale

  When numbers assigned to objects serve as labels for identification or categorization, then such numbers are in nominal scale.
  For e.g.
  - Male = 1
  - Female = 2

- Ordinal Scale

  When assigned numbers to indicate the relation between entities in terms of greater than, equal, or less than but do not state how much greater than or less than, then the scale is called ordinal scale.

  For e.g. Ranks

- Interval Scale

  When assigned numbers are such that the difference in numbers is valid but not ratios, then the scale is called interval scale.

  e.g. Temperature

- Ratio Scale

When a scale contains absolute zero, it is called ratio scale

All mathematical operations (+,-,*,/) are valid on this data

Q. What is skewness and kurtosis? How to measure skewness and kurtosis?

The histogram is an effective graphical technique for showing both the skewness and kurtosis of a data set.

Skewness measures the degree of asymmetry exhibited by the data

$$skewness = \frac{(x - \mu)^3}{\sigma^3}$$

$$kurtosis = \frac{(x - \mu)^4}{\sigma^4}$$

Kurtosis measures how peaked the histogram is.

Q. Find mean deviation and standard deviation, skewness, kurtosis of data

34,56,12,34,32,89,70,65,45,678,90,890,760,55

Mean Deviation:

Formula

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)|$$

$m(X)$ = average value of the data set

n = number of data values

x_i = data values in the set

B	C	D	E	F	G
		xi - m(x)			
	34	-173.86	173.86		
	56	-151.86	151.86		
	12	-195.86	195.86		
	34	-173.86	173.86		
	32	-175.86	175.86		
	89	-118.86	118.86		
	70	-137.86	137.86		
	65	-142.86	142.86		
	45	-162.86	162.86		
	678	470.14	470.14		
	90	-117.86	117.86		
	890	682.14	682.14		
	760	552.14	552.14		
	55	-152.86	152.86		
Average:	207.86	Sum:	3408.88	Mean Deviation:	243.49

Standard Deviation:

Formula

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

σ = population standard deviation

N = the size of the population

x_i = each value from the population

μ = the population mean

Standard Deviation: (P)	300.16	ROUND(STDEV.P(C2:C15),2)
Standard Deviation: (S)	311.49	ROUND(STDEV.S(C2:C15),2)

Skewness:

Formula

$$\tilde{\mu}_3 = \frac{\sum_i^N \left(X_i - \bar{X}\right)^3}{(N-1) * \sigma^3}$$

$\tilde{\mu}_3$ = skewness

N = number of variables in the distribution

X_i = random variable

$\bar{X}$ = mean of the distribution

σ = standard deviation

Skewness:	1.62	ROUND(SKEW(C2:C15),2)

Kurtosis:

$$skewness = \frac{(x-\mu)^3}{\sigma^3}$$

$$kurtosis = \frac{(x-\mu)^4}{\sigma^4}$$

Kurtosis:	0.91	ROUND(KURT(C2:C15),2)