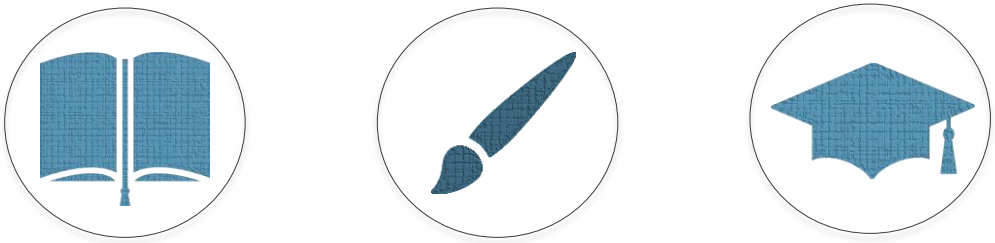


Data Science Concepts

Lesson01D–Understanding Data Attributes

Objective

After completing this lesson you will be able to:

- Understand the building blocks of statistics
- Describe the location, dispersion and shape attributes of a data through the use of sample cases

Case: Types of Data variables

Romanov, an Analytics consultant works with Credit One bank. His manager gave him a list having the name of bank's customers. Further he has been asked to pull the information from bank's database pertaining to the customer list. The information will be around the credit cards issued by the bank. He needs to define the variable types and the type of value each one of them will contain. Romanov, who has just started his professional career, doesn't has a good idea about different variable types.

Now, suppose after extracting data he approached you and asked your help in categorizing the different variables. Help Romanov in variable categorization.

Case: Types of Data variables (Data snapshot)

Sl No	Name of Customer	Customer ID	Number of Credit Cards	Age of Customer (Last Birthday)	Gender of the Customer	Marital Status of the Customer	Annual Salary (in USD)	Monthly Credit Card Usage
1	Josh	111669	5	42	F	Never Married	88,001	Low
2	Janice	146861	6	25	F	Married	592,489	Low
3	Dandre	171690	3	50	M	Divorced	272,304	Low
4	Aiden	161721	6	37	M	Married	726,593	Low
5	Celine	170359	7	50	F	Never Married	612,075	Low
6	Emilio	175646	5	41	M	Never Married	490,356	Low
7	Joaquin	180732	2	62	F	Divorced	164,732	Low
8	Justus	113136	7	26	F	Never Married	510,321	Low
9	Chaya	169254	4	24	M	Never Married	358,534	Low
10	Justyn	149771	4	35	M	Married	140,400	Low
11	Jadon	166226	7	36	M	Never Married	105,259	Low

Case: Types of Data variables

Information to be extracted by Romanov

Variable Name	Name of Customer	Customer ID	Number of Credit Cards	Age of Customer Last Birthday	Gender of Customer	Marital Status of Customer	Annual Salary	Monthly Credit Card Usage
Value Stored	?	?	?	?	?	?	?	?
Variable Type	?	?	?	?	?	?	?	?
Remarks								

Case: Types of Data variables

Information to be extracted by Romanov

Variable Name	Name of Customer	Customer ID	Number of Credit Cards	Age of Customer Last Birthday	Gender of Customer	Marital Status of Customer	Annual Salary	Monthly Credit Card Usage
Value Stored	Name of the individual customer	Unique identifier	1, 2, 3…	18, 19, 20…	Male / Female	Married / Divorced / Never Married	Amount	Low(<25%) / Medium(<50%) / High(<75%) / Very High(>75%)
Variable Type	?	?	?	?	?	?	?	?
Remarks								

Types of Data Variables

Data consists of a combination of "variables" which actually contain the values. Variables at a high level are of two types depending on the kind of values they store:

Numerical variables

Discrete

- Arises from counting. Can take only a set of particular values including negative and fractional values
- Examples: Credit score, number of credit cards owned by a person, number of states in a country, charge on electron etc.

Continuous

- Arises from measuring. Can take any value with in a specified range
- Examples: Height, Amount of money, Age etc.

Categorical variables

Binary (or Dichotomous)

- Has only two categories
- Examples: yes/no, male/female, pass/fail etc.

Nominal

- Has several unordered category
- Examples: Type of bank account, type of insurance policy etc.

Ordinal

- Has several ordered category
- Examples: questionnaire responses such as "strongly in favour / … / strongly against".

Case: Types of Data variables (Revisited)

Information to be extracted by Romanov

Variable Name	Name of Customer	Customer ID	Number of Credit Cards	Age of Customer Last Birthday	Gender of Customer	Marital Status of Customer	Annual Salary	Monthly Credit Card Usage
Value Stored	Name of the individual customer	Unique identifier	1, 2, 3…	18, 19, 20…	Male / Female	Married / Divorced / Never Married	Amount	Low(<25%) / Medium(<50%) / High(<75%) / Very High(>75%)
Variable Type	--	--	Numerical (Discrete)	Numerical (Discrete)	Categorical (Binary)	Categorical (Nominal)	Numerical (Continuous)	Categorical (Ordinal)
Remarks	Identifier	Identifier	Arises from counting. Takes certain discrete values in a given range	Arises from counting. Takes certain discrete values in a given range	Only two categories	Several ordered category	Takes many values in a given range	Several ordered category

Case: Summarizing Data

Romanov, an Analytics consultant works with Credit One bank. His manager gave him some data around credit cards relating to number of credit cards issued to a set of customers and the credit limit of the cards. Further he has been tasked to summarize the data in a presentable form and prepare the report. Romanov, who has just started his professional career, has never played around with such kind of data, so he is clueless about the different summarizing techniques.

Now, suppose he approached you and asked your help in preparing the report. Help Romanov in summarizing the data and preparing the report.

Comments: Summarizing Data

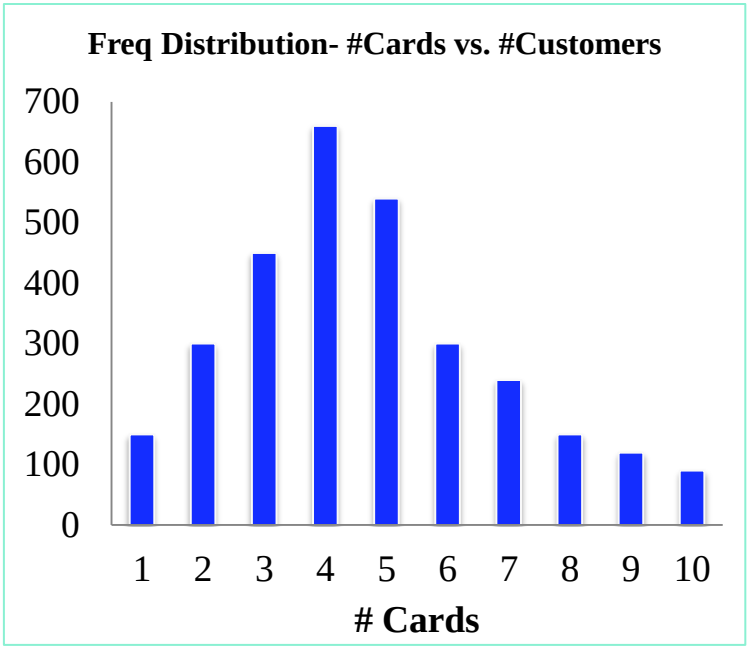
There are various ways to summarize data. Some of them are

- Frequency distribution
- Grouped frequency distribution
- Cumulative frequency distribution

Summarizing Data–Frequency distribution

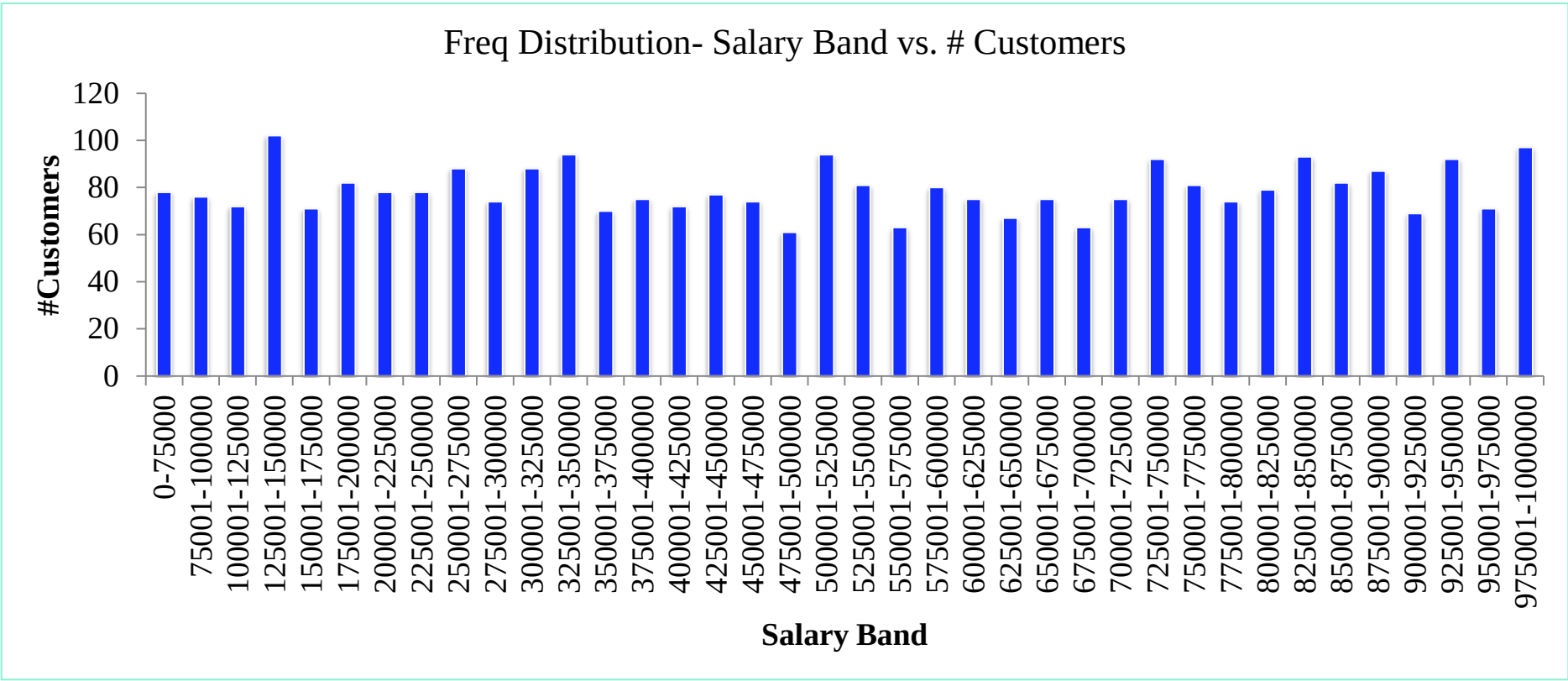
- A technique to summarize discrete data
- A simple process which involves counting of distinct discrete values
- The representation can be either tabular or graphical
- Example: Number of credit cards owned in a sample of 3000 individuals

Number of Credit Cards	# Customers
1	150
2	300
3	450
4	660
5	540
6	300
7	240
8	150
9	120
10	90



Summarizing Data–Grouped Frequency Distribution

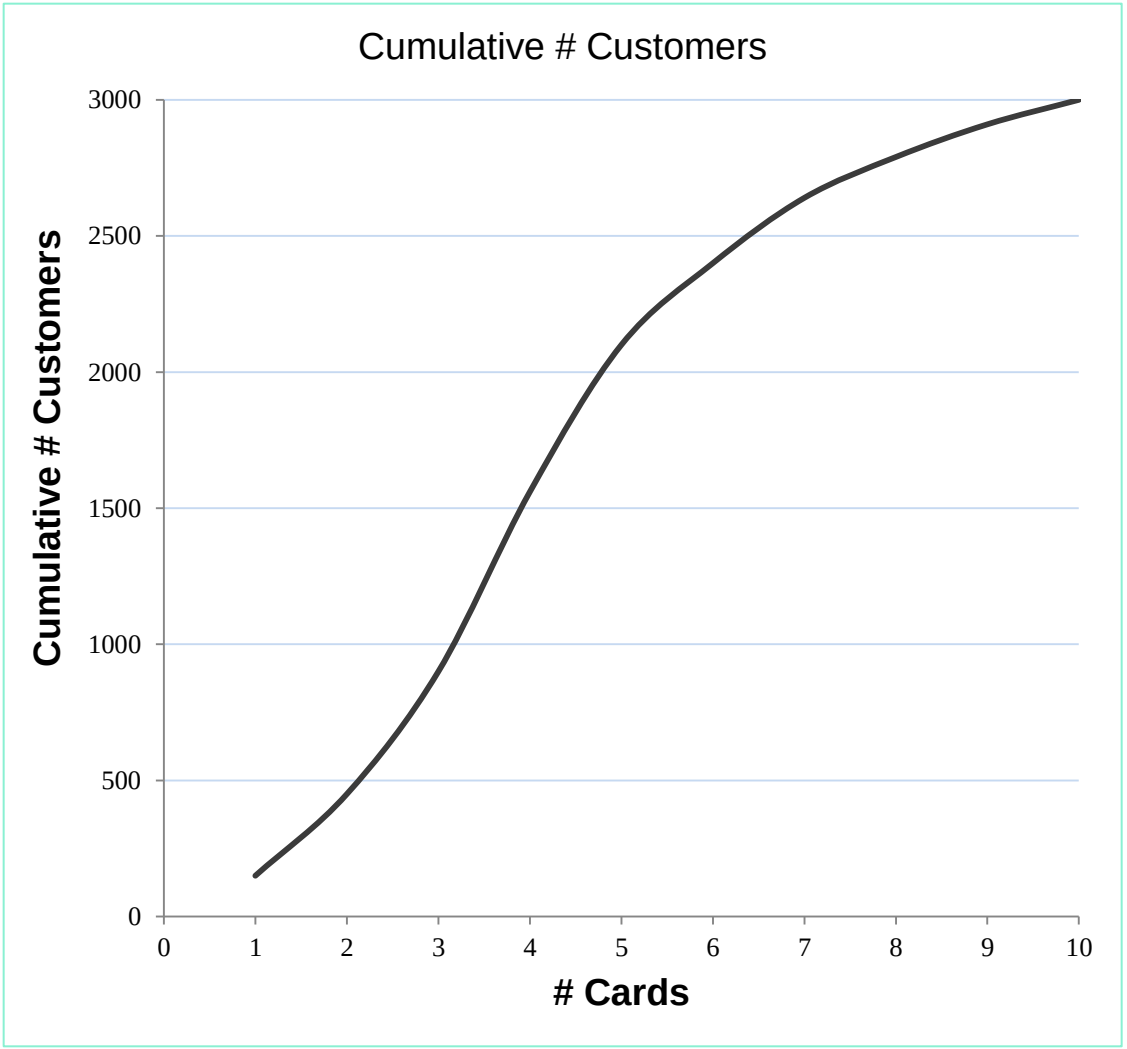
- A technique to summarize continuous data or discrete data having large number of observations and an extended range
- A simple process which involves counting of values falling under the different intervals (grouped)
- Example: Number of customers falling under different Salary groups



Summarizing Data–Cumulative Frequency Distribution

- Cumulative frequencies are obtained by accumulating the frequencies to give the total number of observations up to and including the value or group in question.
- Example: Cumulative number of cards in the sample of 3000 individuals

Number of Credit Cards Up to	Cumulative # Customers
1	150
2	450
3	900
4	1560
5	2100
6	2400
7	2640
8	2790
9	2910
10	3000



Case: Measure of Central Tendency/Location

After Romanov presented the summarized data to his manager at Credit One, he was asked to produce the various measures of Central Tendency of the Credit Card data.

Now, Romanov being unaware of the term "central tendency" again approached you and asked your help in calculating the central tendency of the data in question. Help Romanov in carrying out his task.

Measure of Central Tendency/Location

- There are a number of different quantities, which can be used to estimate the central point of a sample.
- These are called measures of central tendency or measures of location.
- Just different ways of calculating the "average" value of dataset.

Three ways to summarize the central tendency

- Mean
- Median
- Mode

Measure of Central Tendency/Location

- Mean or average of a list of values is given by:

Mean = Sum of values/Count of values

- Median is that value which splits list of numbers into two equal halves. Median of a list of numbers is calculated after sorting the numbers in increasing/ascending order:

Count of numbers is odd: Median is the middle value
Count of numbers is even: Median is the sum of two middle values divided by 2

- Mode is the value in the list of numbers which occurs most frequently. For ease, sort the value in increasing/ascending order:

Count the value which occurs most number of times.

Measure of Central Tendency/Location

The measure of central tendency for the customer's age:

Sl No	Name of Customer	Age of Customer (Last Birthday)	Mean	Median	Mode
1	Josh	42	39	37	50
2	Janice	25			
3	Dandre	50			
4	Aiden	37			
5	Celine	50			
6	Emilio	41			
7	Joaquin	62			
8	Justus	26			
9	Chaya	24			
10	Justyn	35			
11	Jadon	36			

Case: Measure of Spread

After Romanov presented the summarized data along with "measures of Central tendency" to his manager at Credit One, he was further asked to add the various measures of spread to the report.

Now, Romanov being unaware of the term "measures of spread" again approached you and asked for your help. Help Romanov in carrying out his task.

Measure of Spread

- The central tendency of a data set is usually the main feature of interest. But another feature of interest is the spread (or variability or dispersion or scatter).
- Spread determines how widely scattered the data is about the mean (or other measure of location).

Three ways to summarize the spread are:

- Variance and Standard Deviation
- The Range
- The Inter quartile range

Measure of Spread

- Standard deviation is a measure to show how far on average the observations are from the mean.
- The range is a measure to show the spread as a difference between the largest and smallest observations in the data set.

Range of a dataset = (Max value in dataset –Min value in dataset)

- Interquartile range is a measure of spread and is calculated based on the quartiles of a data set.
  - Quartile divides the data set into 4 quarters and is denoted by Q1, Q2 and Q3.
  - Interquartile range is given by Q3-Q1.
  - Use Quartile function in excel to compute the quartile values.

- Standard deviation is the most commonly used metric for measure of spread.
- Range is a poor measure of spread as it relies on the extreme values.
- Interquartile range is similar to range but is not affected by extreme values.

Measure of Spread

The measure of dispersion for the customer's annual salary (Mean salary is USD 369, 188):

Sl No (N)	Name of Customer	A= Annual Salary (in USD)	Deviation D=(A-Mean)	E=Square(D)	Variance V= Sum(E)/N	Standard Deviation= SQRT(V)	Range (Min-Max)	IQR = Q3-Q1
1	Josh	88,001	-281,187	79,065,924,469	47,596,985,579	218,167	638,592	398,839
2	Janice	592,489	223,301	49,863,499,002				
3	Dandre	272,304	-96,884	9,386,438,995				
4	Aiden	726,593	357,405	127,738,593,956				
5	Celine	612,075	242,887	58,994,271,414				
6	Emilio	490,356	121,168	14,681,772,346				
7	Joaquin	164,732	-204,456	41,802,107,241				
8	Justus	510,321	141,133	19,918,626,331				
9	Chaya	358,534	-10,654	113,499,968				
10	Justyn	140,400	-228,788	52,343,782,553				
11	Jadon	105,259	-263,929	69,658,325,093				

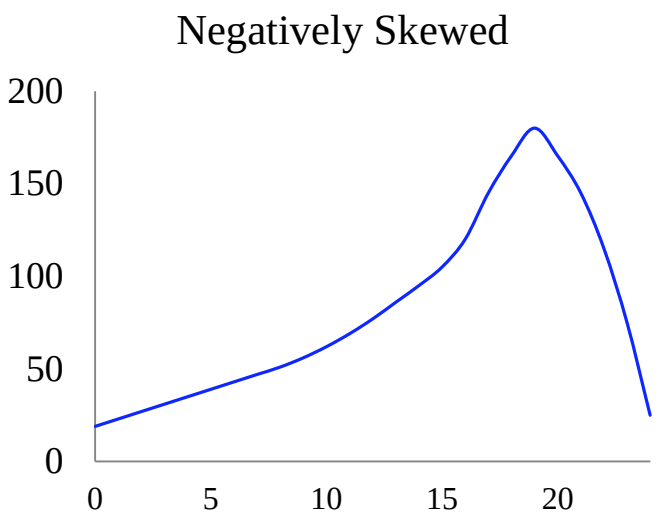
Case: Symmetry and skewness of data

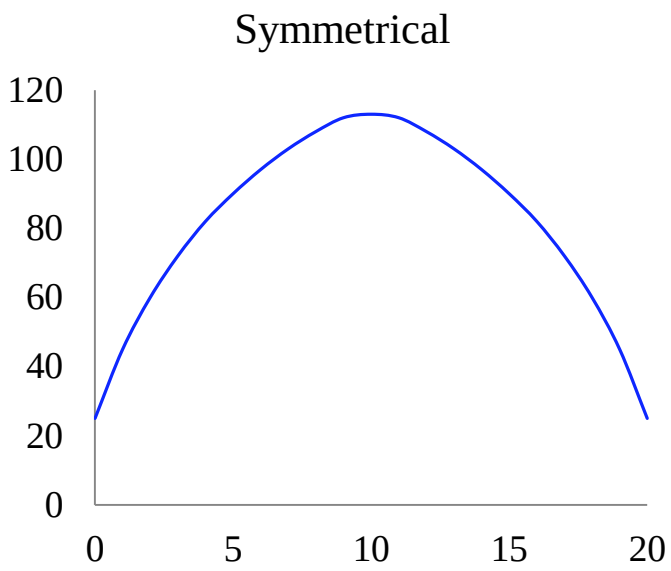
Romanov got appreciations after he presented the summarized data along with "measures of Central tendency" and "measure of spread" to his manager at Credit One. But, he was further asked to create an illustration around symmetry and skewness of data. Following that carry out the analysis of credit card data

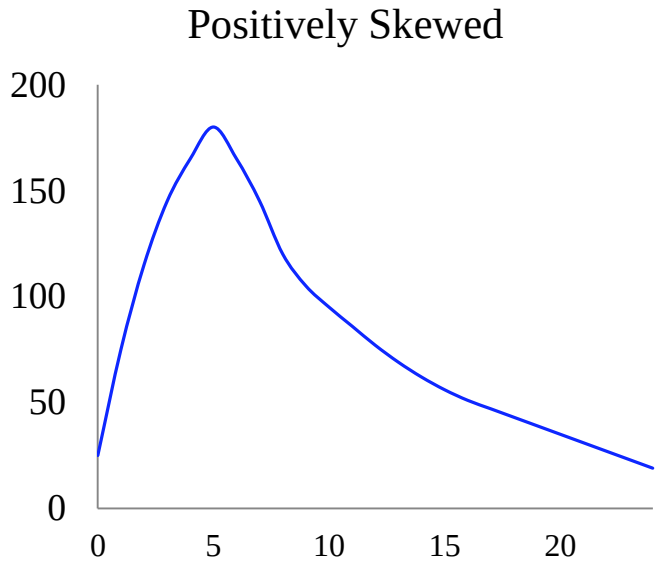
Now, Romanov being unaware of the term "symmetry and skewness" again approached you and asked for your help. In return he promised to gift you a bottle of Champagne. Help Romanov in carrying out his task.

Symmetry and skewness

- Symmetry and skewness deals with the shape of the distribution of a data set.
- The approximate shape of a distribution can be determined by looking at a histogram.
- Density plot is a better representation to analyze the shape of the distribution







Symmetry and Skewness

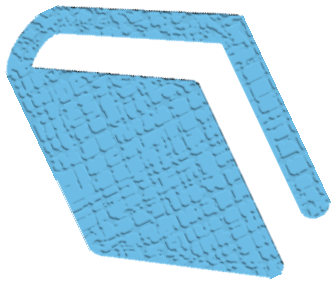
The measure of shape for the customer's age:

Sl No	Name of Customer	Age of Customer (Last Birthday)	Mean	Median	Mode
1	Josh	42	39	37	50
2	Janice	25			
3	Dandre	50			
4	Aiden	37			
5	Celine	50			
6	Emilio	41			
7	Joaquin	62			
8	Justus	26			
9	Chaya	24			
10	Justyn	35			
11	Jadon	36			

Symmetrical: Mean = Median = Mode
Positively Skewed: Mean > Median > Mode
Negatively Skewed: Mean < Median < Mode

Summary

Summary of the topics covered in this lesson:



- Data behavior is explained through location, spread and shape or distribution of the data.
- Mean, Median and Mode are the three attributes which explains the location or central tendency of the data.
- Standard deviation is a measure to understand the spread of the data. This is the most commonly used attribute apart from range.
- Shape of the data is explained by histogram plot. However, histogram may be misleading in understanding the shape/distribution of data. Density plot is a better representation to understand data distribution.

QUIZ TIME

Quiz Question 1

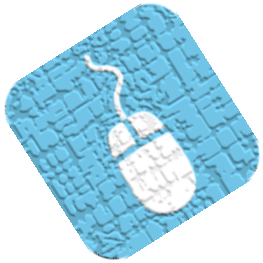
Quiz 1	What are the attributes to understand the central tendency of data? *Select all that apply.*

a. Mean

b. Variance

c. Median

d. Standard deviation



Quiz Question 1

Quiz 1	What are the attributes to understand the central tendency of data? *Select all that apply.*

a. Mean

b. Variance

c. Median

d. Standard deviation

Correct answer is:

a & c

Mean and Median are the two attributes to understand central tendency. The other attribute is Mode.

End of Lesson01D–Understanding Data Attributes