INTRODUCTION TO DATA ANALYTICS

2nd Sem, MCA

CONTENT

❑ Introduction and overview

- Data Analytics,
- Case study data analysis,
- Scope of Data analytics,
- Essential skills,
- Data sources,
- Data sets,
- Data types.

INTRODUCTION – DATA ANALYTICS

- Companies/Business/System generate vast huge volumes of data daily (log files, web servers, transactional data, and various customer-related data). Social media & user-generated data ads on to it.
  - Businesses ideally need to use all these generated data to **derive value** out of it and make **impactful business decisions**.
  - Data analytics is used to drive this purpose.
- *Data analytics is the science of analyzing raw data to make conclusions about that information.*
  - *Process of exploring and analyzing large datasets to find hidden patterns, unseen trends, discover correlations, and derive valuable insights to make business prediction to improve business speed & efficiency.*
- Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information → This information can then be used to optimize processes to increase overall business or system efficiency.
- Many of the techniques and processes of data analytics have been automated into mechanical processes with the help of **algorithms** that work over raw data for human/business consumption.

INTRODUCTION – DATA ANALYTICS

- Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.
  - Data mining is a particular data analysis technique that focuses on statistical modelling and knowledge discovery for predictive rather than purely descriptive purposes.
  - Business intelligence covers data analysis that relies heavily on aggregation, focusing mainly on business information.
- Data analysis can be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA).
  - EDA focuses on discovering new features in data.
  - CDA focuses on confirming or falsifying existing hypotheses.

INTRODUCTION – DATA ANALYTICS

- Data Analysis helps in understanding the data and provides required insights from the past to understand what happened so far.
- Data Analytics is the process of exploring data from the past to make appropriate decisions in the future by using valuable insights.
- Data analytics is primarily focused on understanding datasets and gaining insights that can be turned into actions.
- Data science is centered on building, cleaning, and organizing datasets.
  - Data scientists create and leverage algorithms, statistical models, and their own custom analyses to collect and shape raw data into something that can be more easily understood.
- Data analytics focuses on processing and performing statistical analysis of existing datasets.
  - Analysts concentrate on creating methods to capture, process, and organize data to uncover actionable insights for current problems, and establishing the best way to present this data.

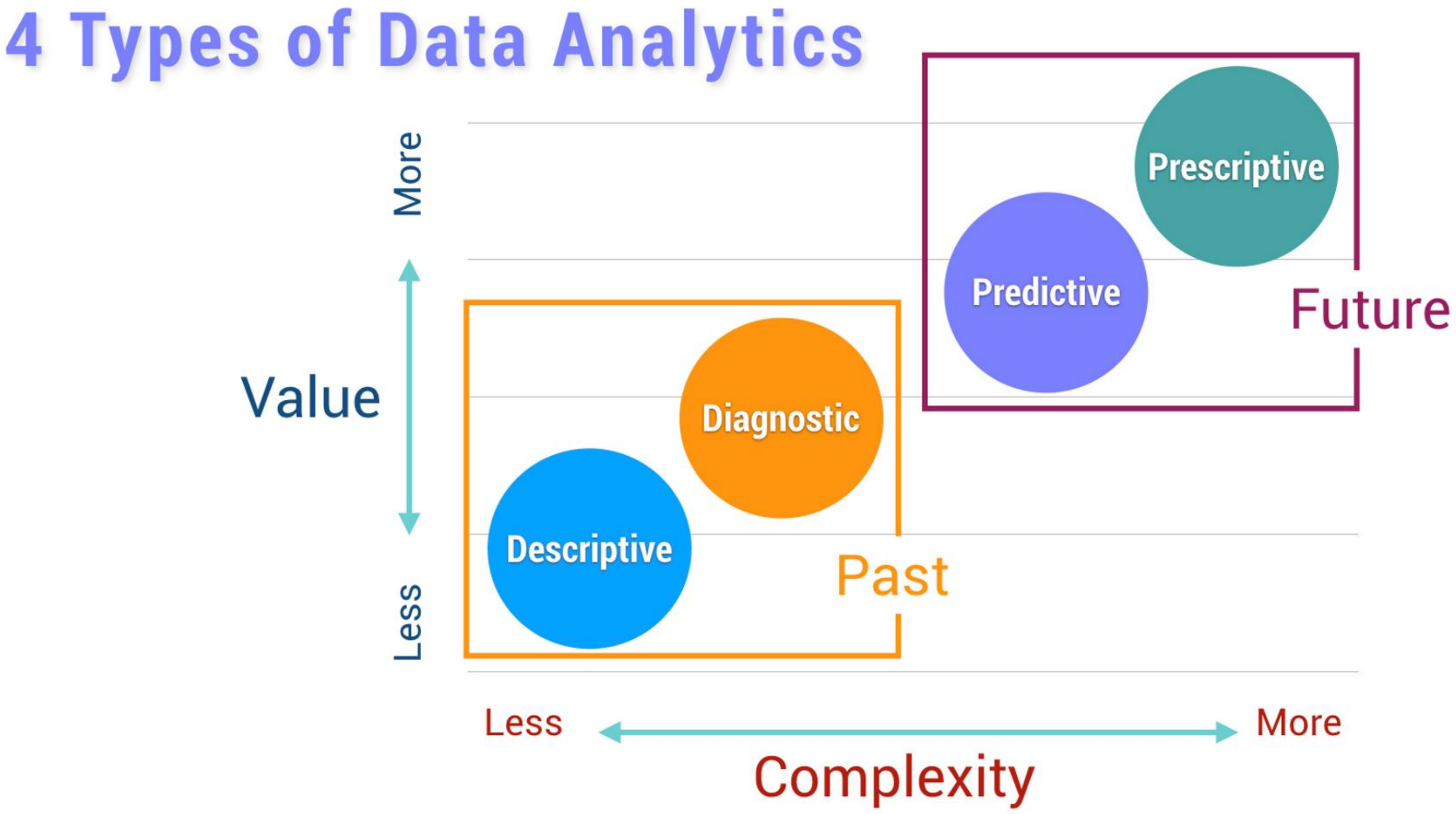
INTRODUCTION – DATA ANALYTICS

- Business analytics: Applying data analytics tools and methodologies in business setting.
  - Main goal is to extract meaningful insights from data that an organization can use to inform its strategy and, ultimately, reach its objectives.
- Business analytics usecases:
  - ***Budgeting and forecasting***: By assessing company's historical revenue, sales, and costs data alongside its goals for future growth, identify the budget and investments required to make those goals a reality.
  - ***Risk management***: By understanding likelihood of certain business risks occurring—and their associated costs—make cost-effective recommendations to help mitigate them.
  - ***Marketing and sales***: By understanding key metrics, such as lead-to-customer conversion rate, identify the number of leads their efforts must generate to fill the sales pipeline.
  - ***Product development (or research and development***)*:* By understanding how customers have reacted to product features in past, guide product development, design, and user experience in the future.

INTRODUCTION – DATA ANALYTICS

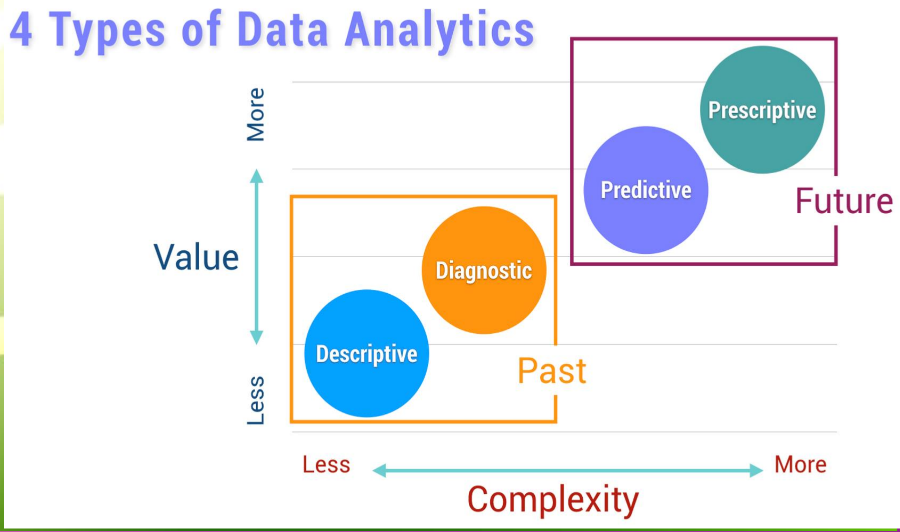
- **Data analytics** is process of cleaning, transforming, and modeling data to discover useful information for business decision-making.

- Types of Data Analytics
  - Descriptive Analytics
  - Diagnostic Analytics
  - Predictive Analytics
  - Prescriptive Analytics


4 Types of Data Analytics

INTRODUCTION – DATA ANALYTICS

- **Descriptive Analytics** answers "what happened" by summarizing past data, with help of visualization (table, graph, dashboards).
- **Diagnostic Analytics** ("why it happened") takes insights found from descriptive analytics and drills down to find causes of those outcomes (such analytics creates more connections between data and identifies patterns of behavior).
- **Predictive Analytics** attempts to answer the question "what is likely to happen" (utilizes previous data to make predictions about future outcomes).
  - Such analytics relies on statistical modeling, which requires added technology and manpower to forecast (forecasting is only an estimate; the accuracy of predictions relies on quality and detailed data).
- **Prescriptive Analytics** combines the insight from all previous analyses to determine the course of action to take to reach future targets "what/how to do".
- AI systems consume large amount of data to continuously learn & use this information to make informed decisions to communicate/put these decisions into action.



INTRODUCTION – DATA ANALYTICS

- Data Analytics is done by scrubbing the data and applying algorithmic processes to find patterns, trends, correlations, and aberrations.
  - Goal is to come up with actionable conclusions to improve business and organizational outcomes.

- A data analyst uses technical skills to analyze data and report insights.
  - Typically data analyst might use SQL skills to pull data from company database,
  - Use programming skills to analyze that data,
  - Use reporting skills for presenting the results to respective audience.

INTRODUCTION – DATA ANALYTICS

Essential skills for data analytics:

- Excellent problem solving skills: Critical Thinking.
- Solid numerical skills: Statistics, Linear Algebra, Calculus, etc.
- Microsoft Excel proficiency
- Data Cleaning, preprocessing & EDA
- Knowledge of querying languages: SQL, NoSQL, etc.
- Programming Language: R, MATLAB, Python, etc.
- Expertise in data visualization: Excel, Python, Tableau, PowerBI, etc.
- Machine Learning
- Reporting/Communication skills

INTRODUCTION - DATA

- Data collection: Process of collecting, extracting, and storing voluminous amount of data which may be in structured or unstructured form like text, video, audio, XML files, records, or other image files.
  - In data analysis, “Data collection” is the initial step before starting to analyze patterns or useful information in data.
  - The data which is to be analyzed must be collected from different valid sources.
  - Collected data is raw data; is not useful → needs cleaning & preprocessing.
- Data collection starts with asking some questions such as what type of data is to be collected and what is the source of collection.
- Most of the data collected are of two types.
  - “Qualitative data” is a group of non-numerical data such as words, sentences mostly focus on behavior and actions of the group.
  - “Quantitative data” is in numerical forms and can be calculated using different scientific tools and sampling data.

INTRODUCTION - DATA

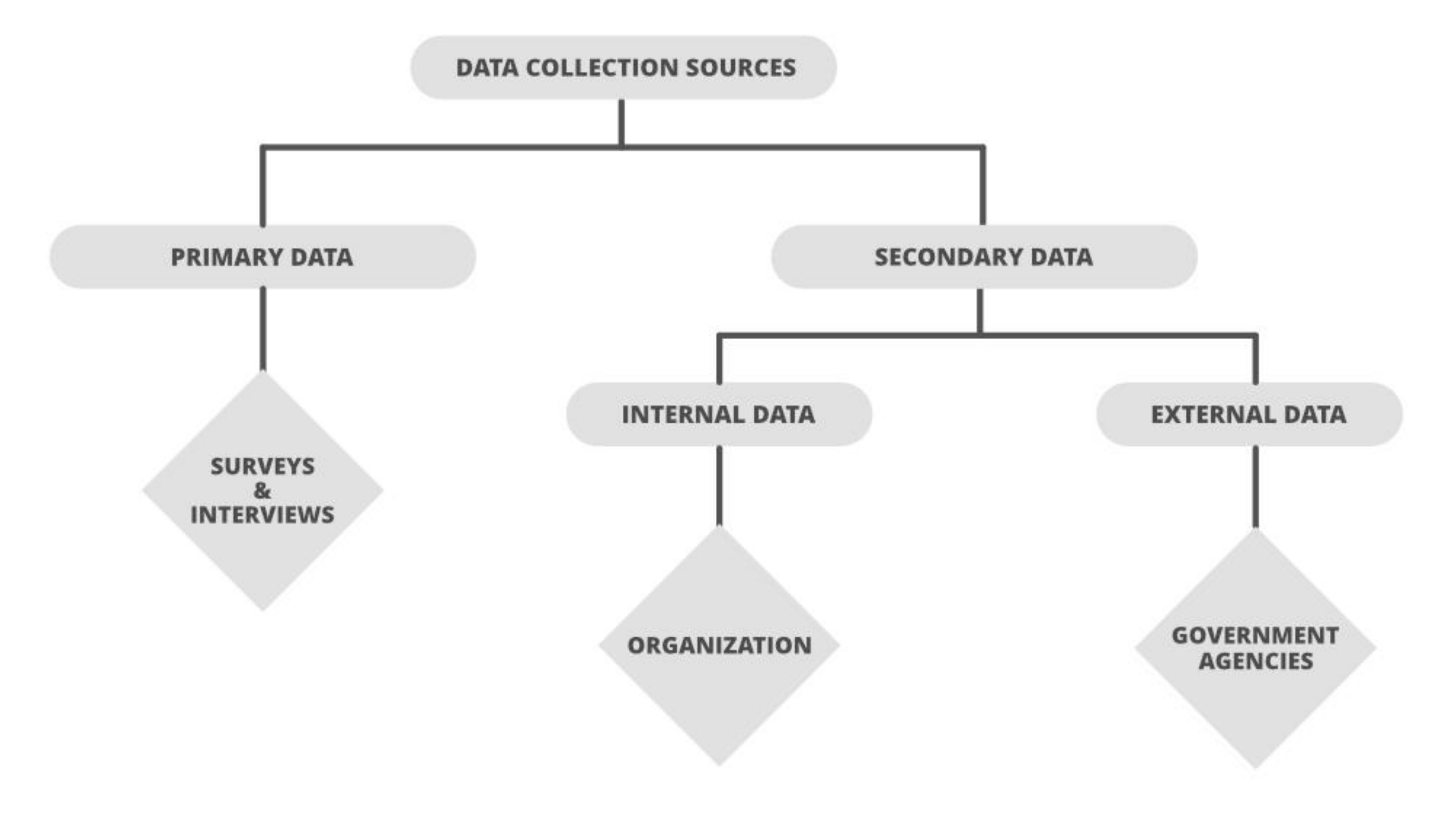
Primary data :

- Raw/original data extracted directly from the official sources.
- Collected directly by performing techniques like questionnaires, interviews, surveys, observation & experimental method.
- Data collected must be according to the demand and requirements of the target audience on which analysis is performed; otherwise it would be a *burden in the data processing.*

Secondary data :

- Data which has already been collected and reused again for some valid purpose.
- Data is previously recorded from primary data. It has 2 types of sources; internal source and external source.
- **Internal source:** data can easily be found within organization (**market record, sales record, transactions, customer data, accounting resources**, etc). *Less cost and time consumption in obtaining internal sources.*
- **External source:** data which can't be found at internal organizations (**Government publications, news publications,, planning commission, international labor bureau, syndicate services, other non-governmental publications, Sensors data, Satellites data, Web traffic, etc)**. More cost and time consumption.

INTRODUCTION - DATA



INTRODUCTION - DATA

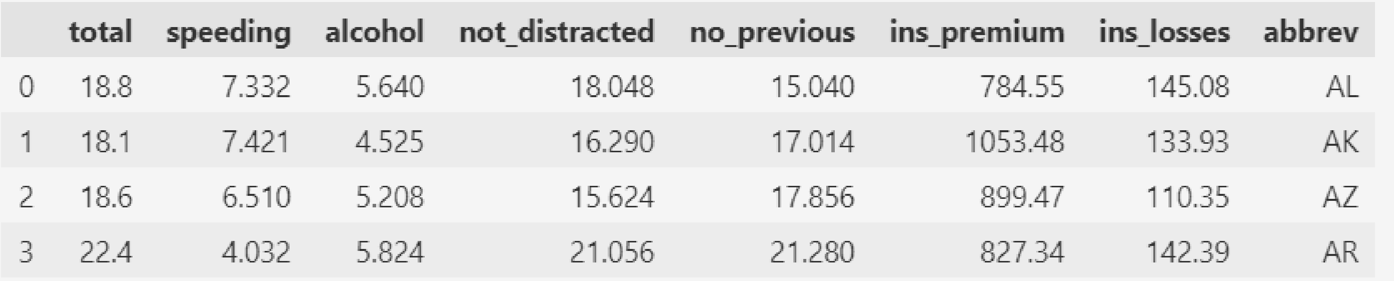
- Main data source is database, which can be located in a disk or a remote server.
- The data source for a computer program can be a data file, spreadsheet, XML file or even hard-coded data within the program.
- Companies use ETL tool to
    - collect (extract) data from their transactional databases,
    - transform them to be optimize,
    - load them into a data warehouse or other data mart.

INTRODUCTION - DATA

- A dataset is **a collection of data** (tabular data)
  - Dataset corresponds to one or more database tables,
  - every column table represents a particular variable,
  - each row corresponds to a given record of the data set.
- A data point is **a discrete unit of information**. (Any single fact is a data point.)
- Those columns that are inputs (whose value is available for analysis) are referred to as **input variables** (Independent variables).
- Those column of data that may not always be available and analyst would like to predict for new input data in the future is **output variable** (response variable / dependent variable).
- Each **feature** (or column / variables / attributes) represents **a measurable piece of data** that can be used for analysis.
  - Name, Age, Sex, Fare, and so on.
- Dimensionality in statistics refers **to how many attributes a dataset has**.

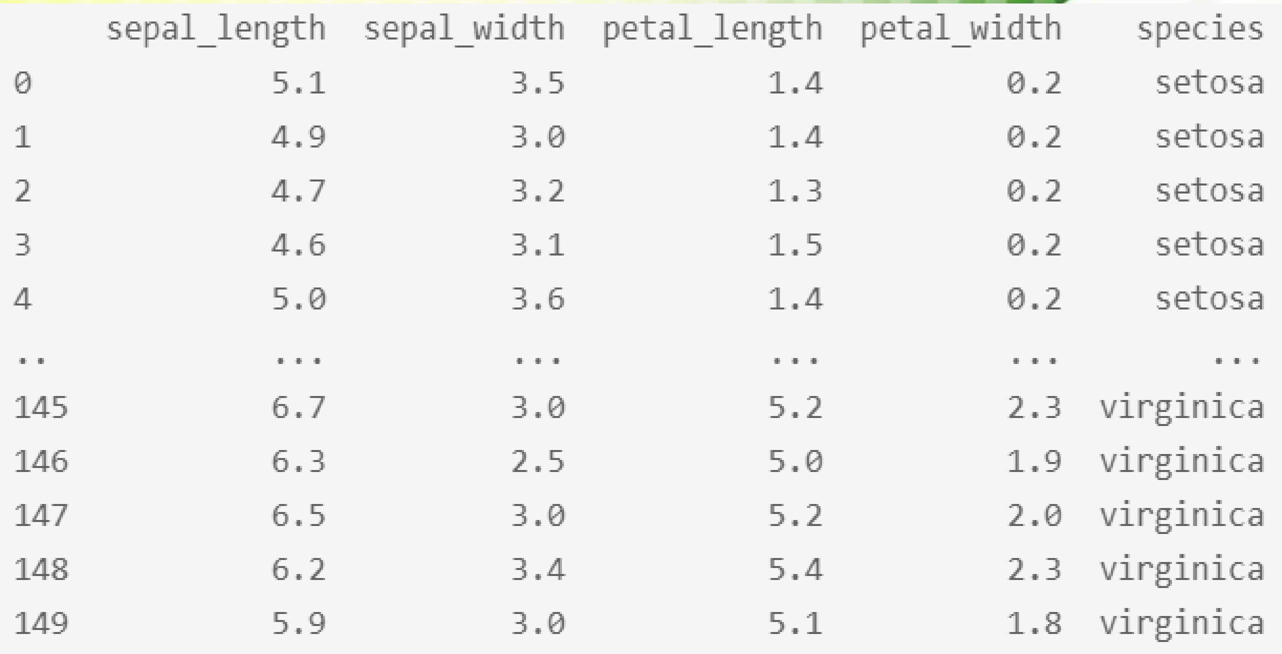
INTRODUCTION - DATA

Car Crash



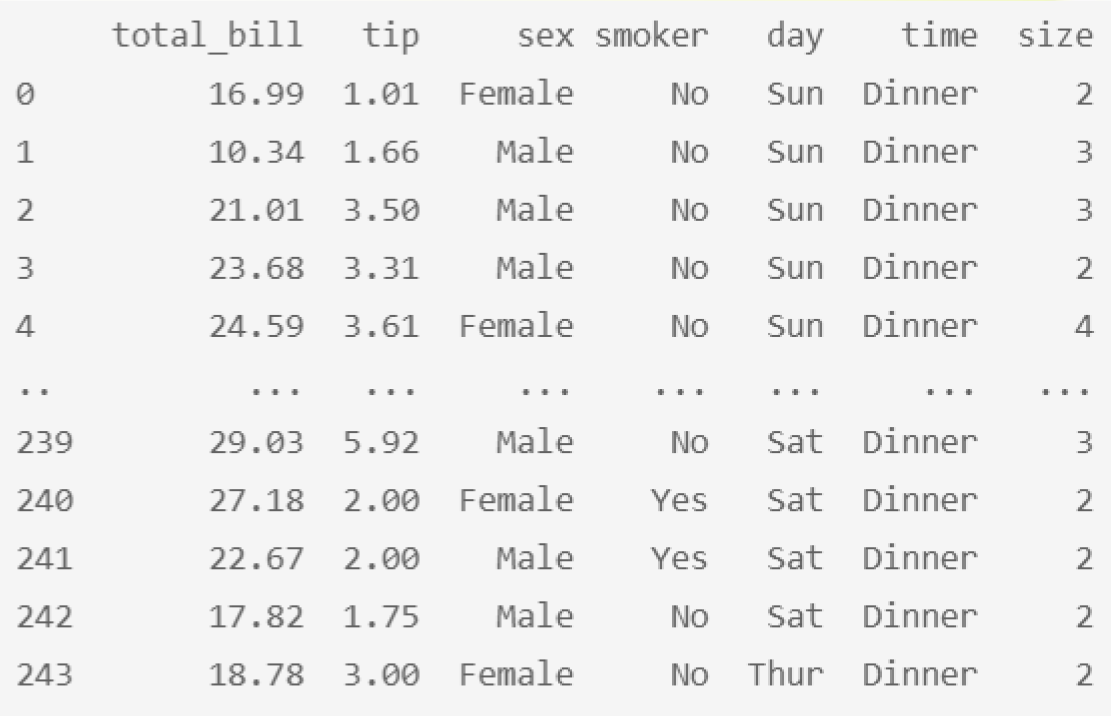
	total	speeding	alcohol	not_distracted	no_previous	ins_premium	ins_losses	abbrev
0	18.8	7.332	5.640	18.048	15.040	784.55	145.08	AL
1	18.1	7.421	4.525	16.290	17.014	1053.48	133.93	AK
2	18.6	6.510	5.208	15.624	17.856	899.47	110.35	AZ
3	22.4	4.032	5.824	21.056	21.280	827.34	142.39	AR

Iris



```
     sepal_length  sepal_width  petal_length  petal_width    species
0             5.1          3.5           1.4          0.2     setosa
1             4.9          3.0           1.4          0.2     setosa
2             4.7          3.2           1.3          0.2     setosa
3             4.6          3.1           1.5          0.2     setosa
4             5.0          3.6           1.4          0.2     setosa
..            ...          ...           ...          ...        ...
145           6.7          3.0           5.2          2.3  virginica
146           6.3          2.5           5.0          1.9  virginica
147           6.5          3.0           5.2          2.0  virginica
148           6.2          3.4           5.4          2.3  virginica
149           5.9          3.0           5.1          1.8  virginica
```

Tips



```
     total_bill   tip     sex smoker   day    time  size
0         16.99  1.01  Female     No   Sun  Dinner     2
1         10.34  1.66    Male     No   Sun  Dinner     3
2         21.01  3.50    Male     No   Sun  Dinner     3
3         23.68  3.31    Male     No   Sun  Dinner     2
4         24.59  3.61  Female     No   Sun  Dinner     4
..          ...   ...     ...    ...   ...     ...   ...
239       29.03  5.92    Male     No   Sat  Dinner     3
240       27.18  2.00  Female    Yes   Sat  Dinner     2
241       22.67  2.00    Male    Yes   Sat  Dinner     2
242       17.82  1.75    Male     No   Sat  Dinner     2
243       18.78  3.00  Female     No  Thur  Dinner     2
```

INTRODUCTION - DATA

Different types of data sets available for different types of information:

- **Numerical data sets**: data are expressed in numbers rather than natural language. (quantitative data)
- **Bivariate data sets**: A data set that has two variables.
- **Multivariate data sets**: A data set with multiple variables (three or more variables).
- **Categorical data sets**: Qualitative variable
  - Variable takes exactly two values (dichotomous variable).
  - Categorical data/variables may have more than two possible values (polytomous variables).
- **Correlation data sets:** set of values that demonstrate some relationship with each other.
  - Correlation is defined as a statistical relationship between two entities/variables.
    - Positive correlation – Two variables move in the same direction (Either both are up or both or down)
    - Negative correlation – Two variables move in opposite directions. (One variable is up and another variable is down and vice versa)
    - No or zero correlation – No relationship between two variables.

INTRODUCTION - DATA

Qualitative/Categorical Data Type

- Finite set of discrete classes; data can’t be counted/measured using numbers; divided into categories.
- **Nominal** values represent discrete units used to label variables with no order. (Male, female)
- **Ordinal** values represent discrete and ordered units. (High, medium, low)

Quantitative Data Type

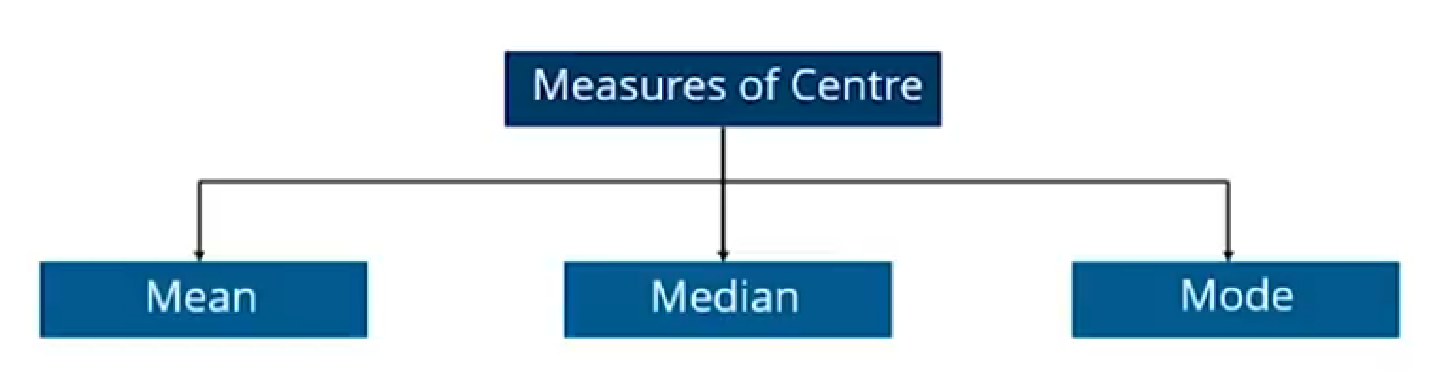
- Quantify things by considering numerical values that make it countable in nature.
- **Discrete data** has distinct and separate values, and the data variables cannot be divided into smaller parts. (number of students in class)
- **Continuous data** represents measurements on a scale or continuum and can have almost any numeric value; that could be meaningfully divided into its finer levels. (height of a person).

CONTENT

- ❑ Introduction and overview
  - Descriptive statistics -
    - Central Tendency,
    - Measures of dispersion,
    - Visualization.

INTRODUCTION - DESCRIPTIVE STATISTICS

- Descriptive statistics are used to **describe or summarize the characteristics of data set.**
- In statistics, central tendency is the descriptive summary of a data set.
  - Single value reflects center of the data distribution.
  - Does not provide information regarding individual data from dataset, rather gives a summary of the dataset.
- Measures of Central Tendency: mean, median and mode.



INTRODUCTION - DESCRIPTIVE STATISTICS

Mean: (average)

- Most popular measure of central tendency.
- Sum of all the values in the data set divided by number of values in the data set.
- For n values in a data set and they have values x_1,x_2, …,x_n; then mean is (“x bar”):

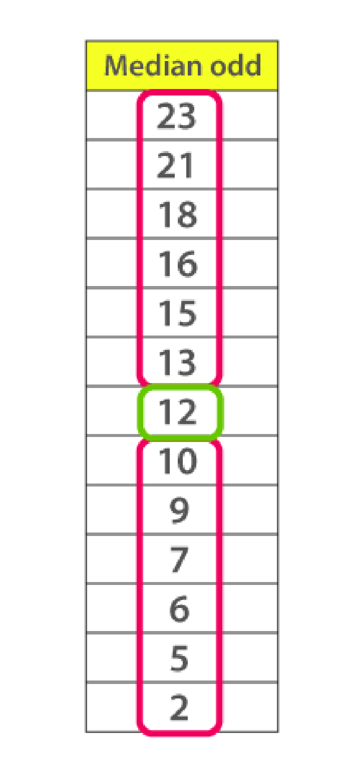
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\bar{x} = \frac{\sum x}{n}$$

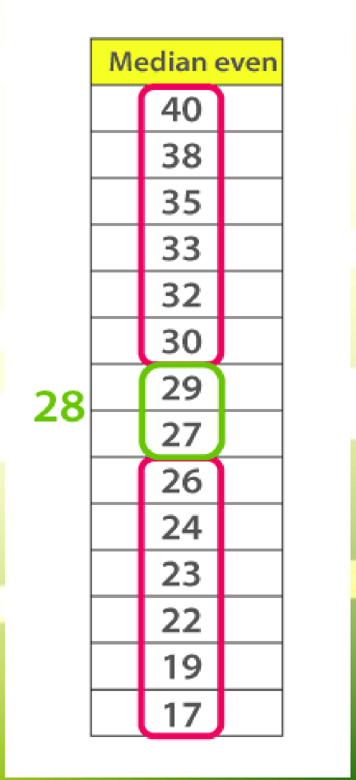
INTRODUCTION - DESCRIPTIVE STATISTICS

Median:

- Middle value of dataset (when dataset is ordered; ascending or descending).
- When dataset contains even number of values; Mean of the middle two values is median.



Median odd
23
21
18
16
15
13
12
10
9
7
6
5
2



Median even
40
38
35
33
32
30
29
27
26
24
23
22
19
17

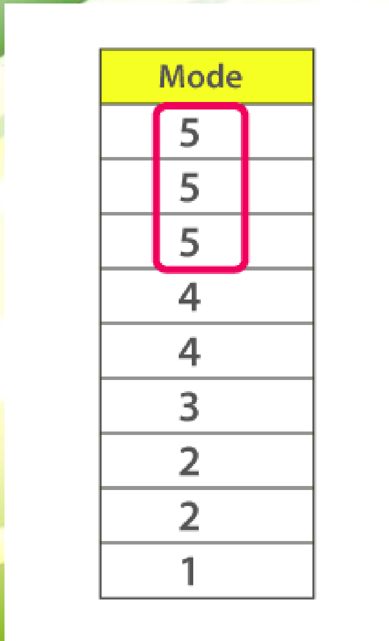
28

INTRODUCTION - DESCRIPTIVE STATISTICS

Mode:

- Represents the frequently occurring value in the dataset.
- Sometimes dataset may contain multiple modes or no mode at all.

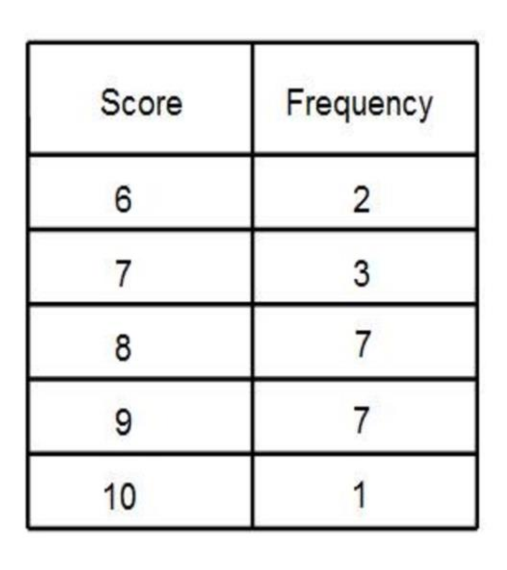
- Mode for 5, 4, 2, 3, 2, 1, 5, 4, 5



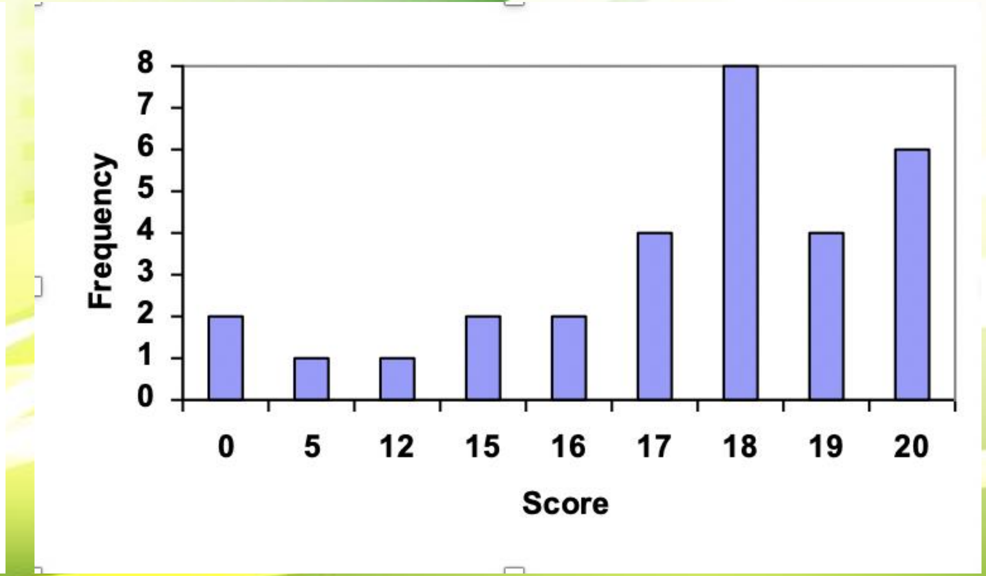
Mode
5
5
5
4
4
3
2
2
1

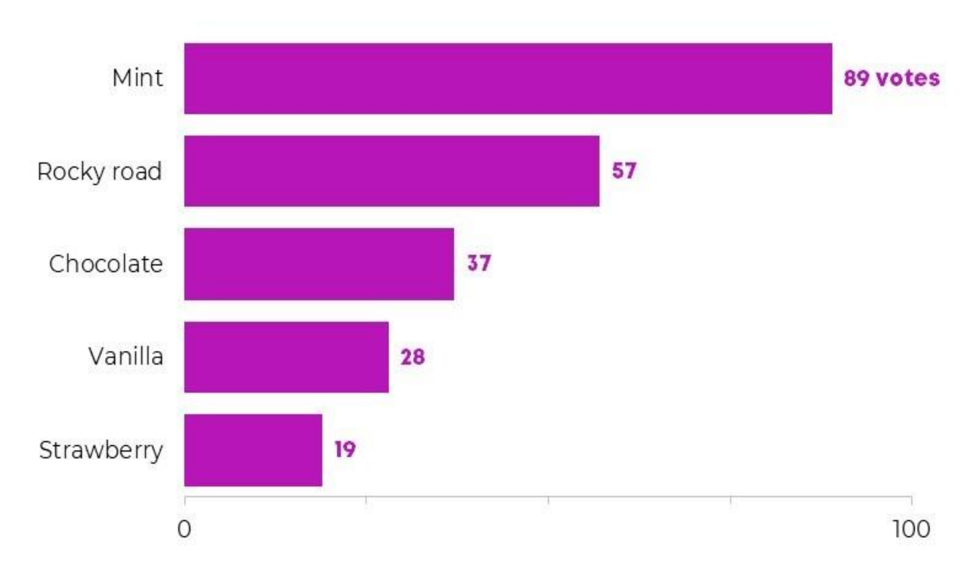
INTRODUCTION - DESCRIPTIVE STATISTICS

- **Frequency** : number of observations taking specific value.
- **Frequency table** : list of possible values and their frequencies.
- **Bar chart** consists of bars corresponding to each of the possible values, whose heights are equal to the frequencies. (**Column chart** can also be used)



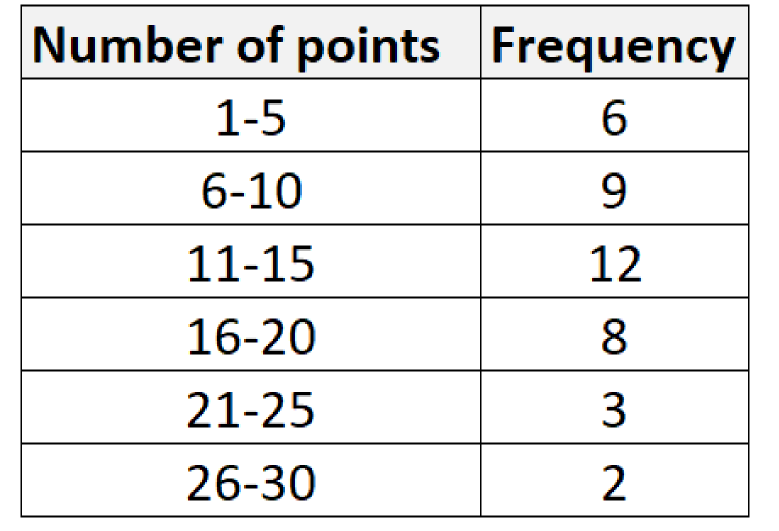
Score	Frequency
6	2
7	3
8	7
9	7
10	1



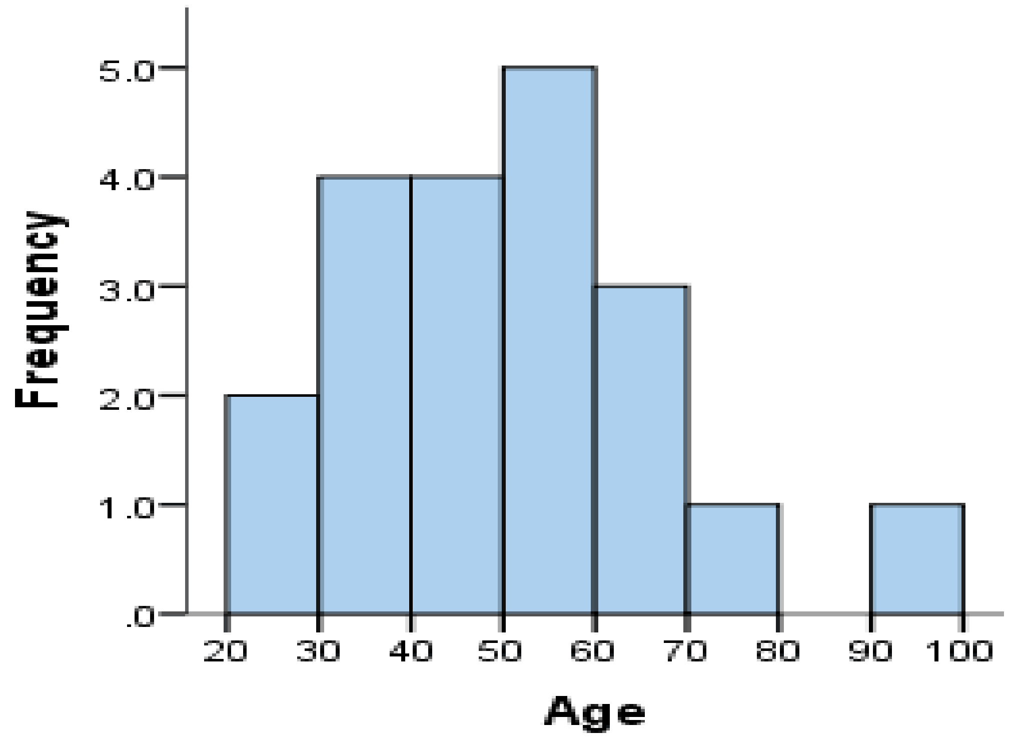


INTRODUCTION - DESCRIPTIVE STATISTICS

- When variable holds continuous values, grouping values into intervals is better than frequency of each value.
- **Histogram** plots frequency against interval.

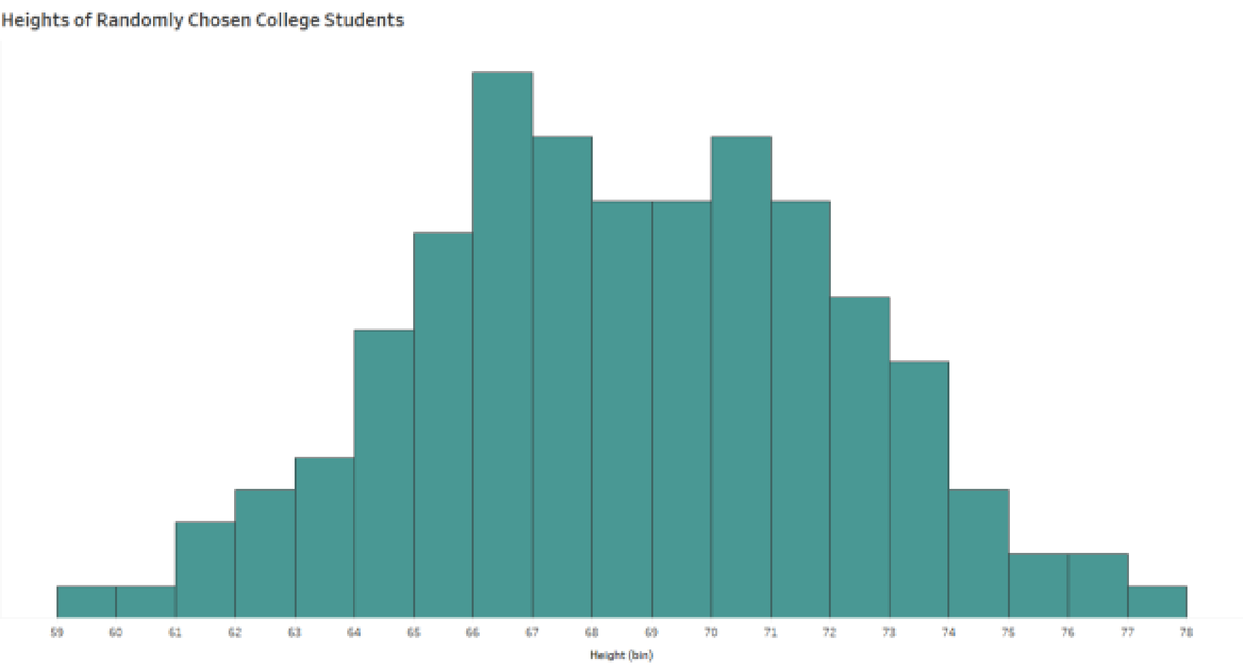


Number of points	Frequency
1-5	6
6-10	9
11-15	12
16-20	8
21-25	3
26-30	2



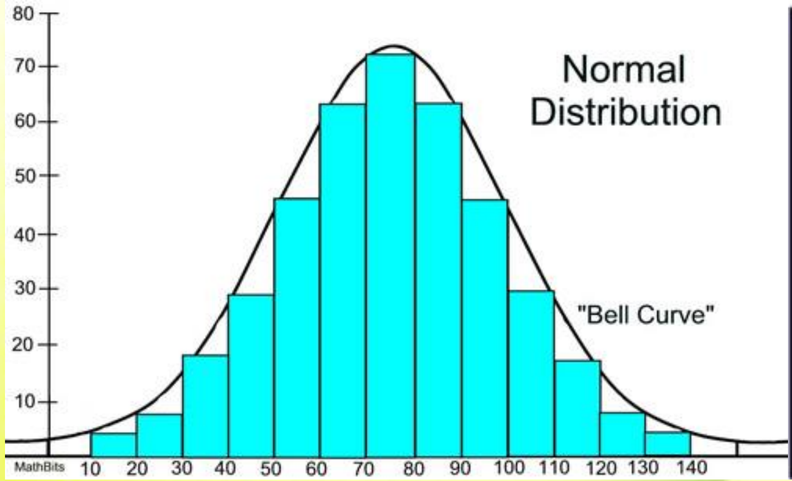
INTRODUCTION - DESCRIPTIVE STATISTICS

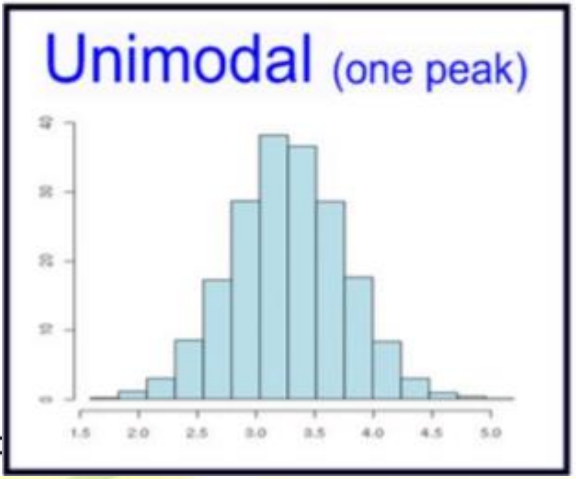
- A dataset/graph having one tall peak (maximum frequency data value) is called *unimodal.*
- *Two* peaks dataset/graph is referred to as *bimodal*.
- Multiple peaks → *multimodal*.
- A cluster of tall bars/bins is sometimes called a *modal range*.

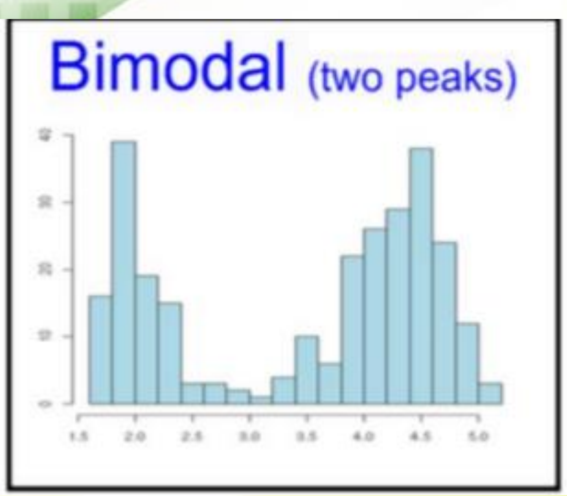


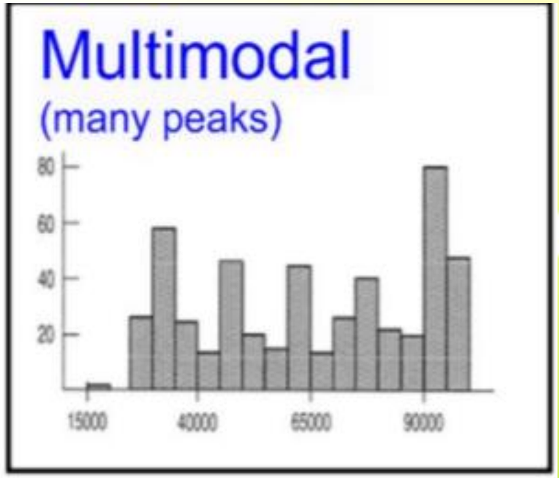
INTRODUCTION - DESCRIPTIVE STATISTICS

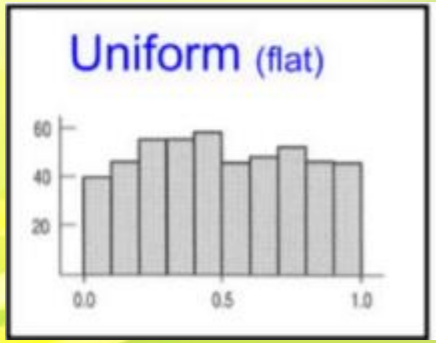
- Bar/column/histogram chart easily indicates frequency distribution.

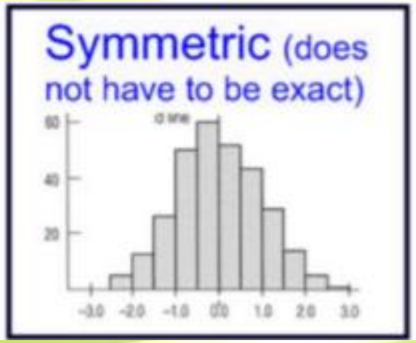


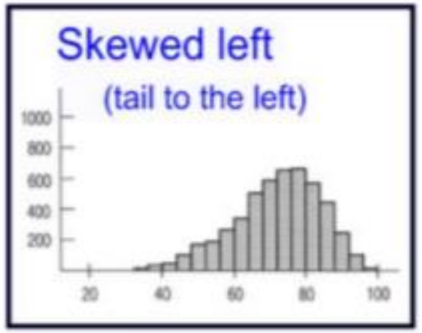


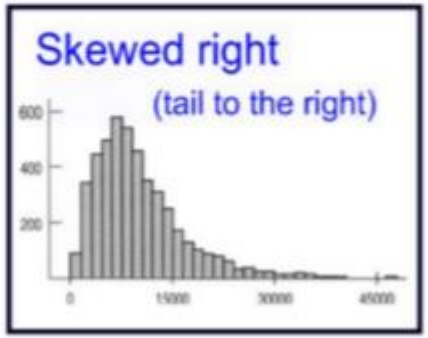






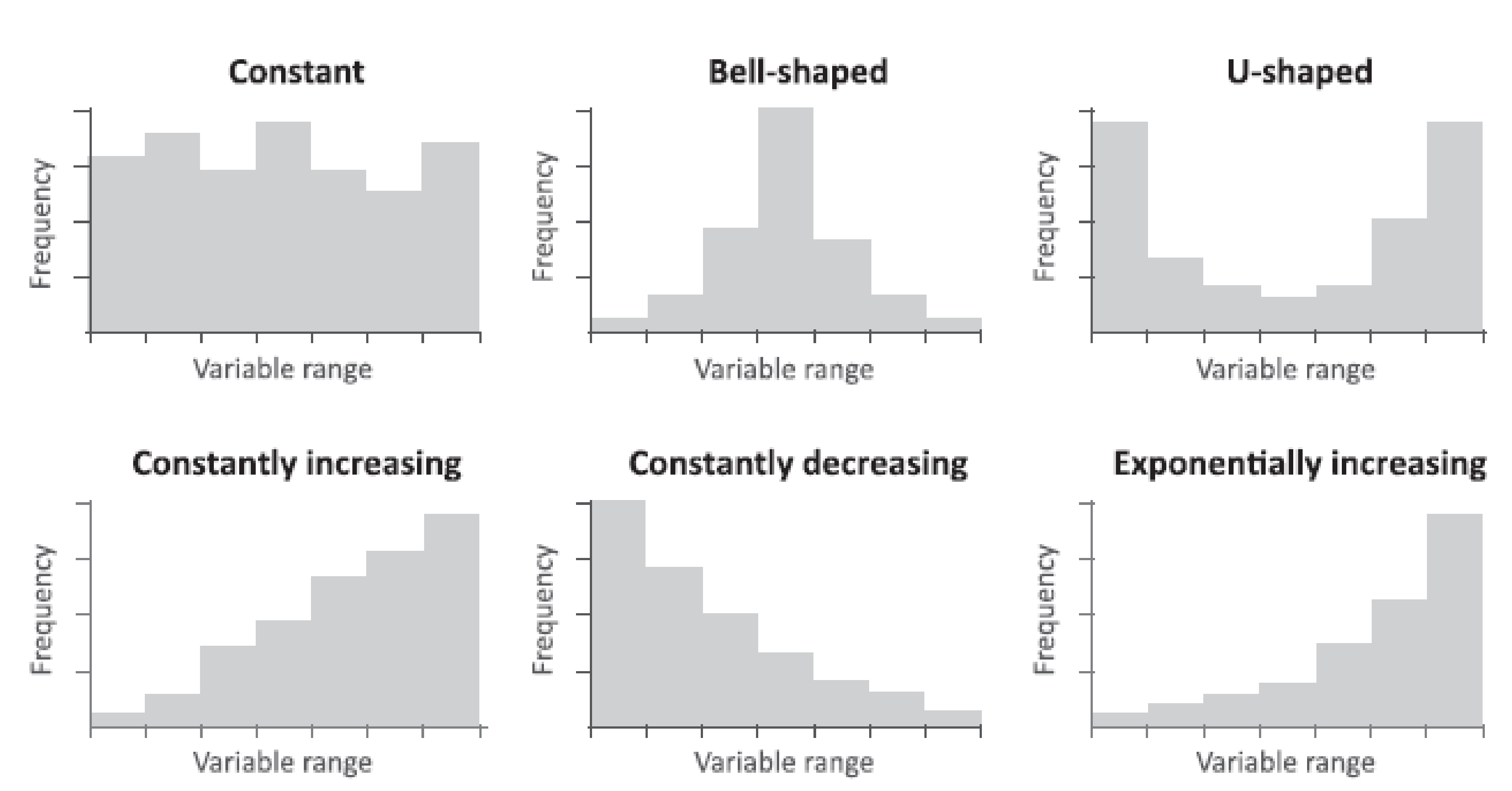






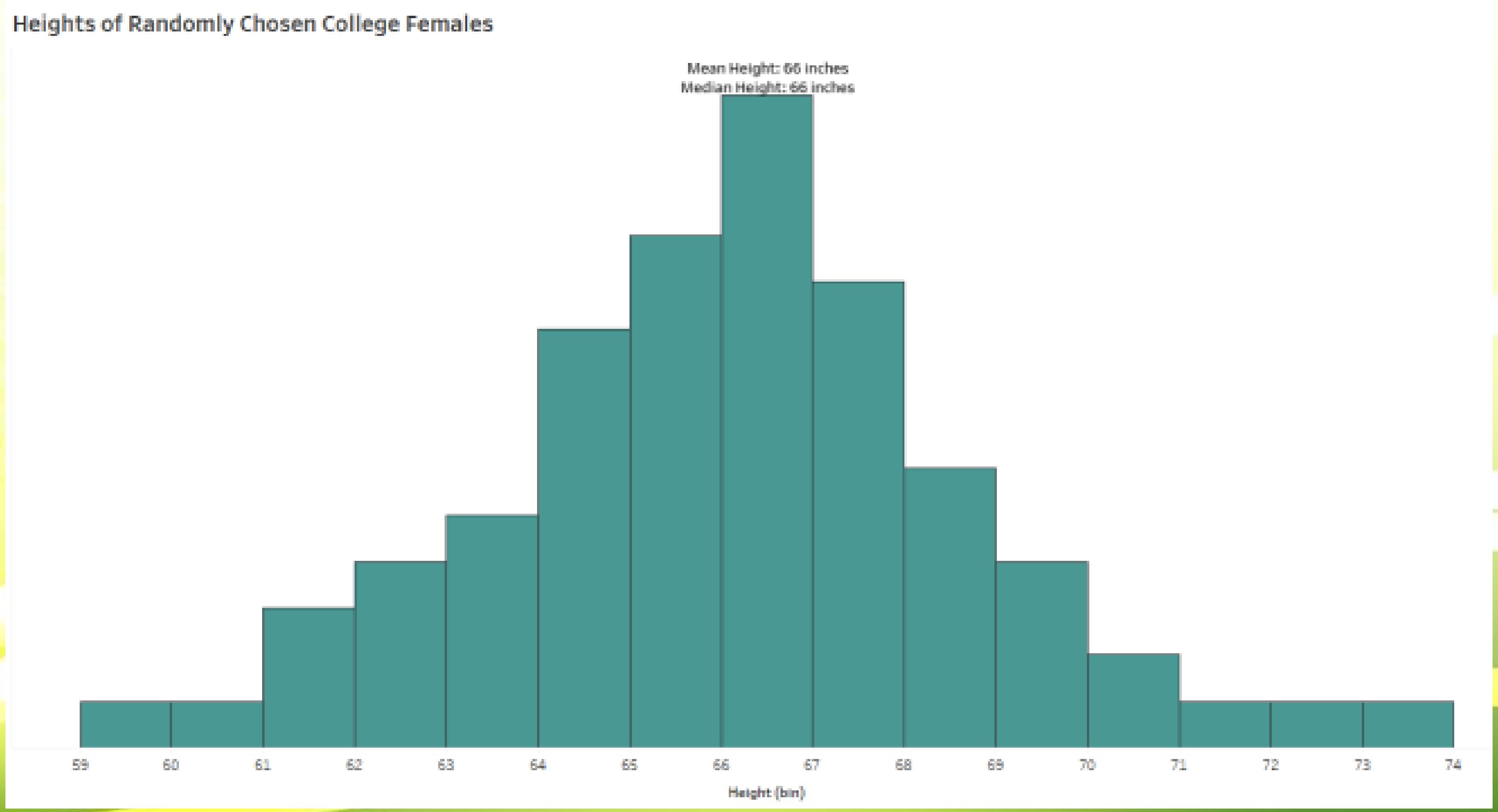
INTRODUCTION - DESCRIPTIVE STATISTICS

- Types of frequency distribution.



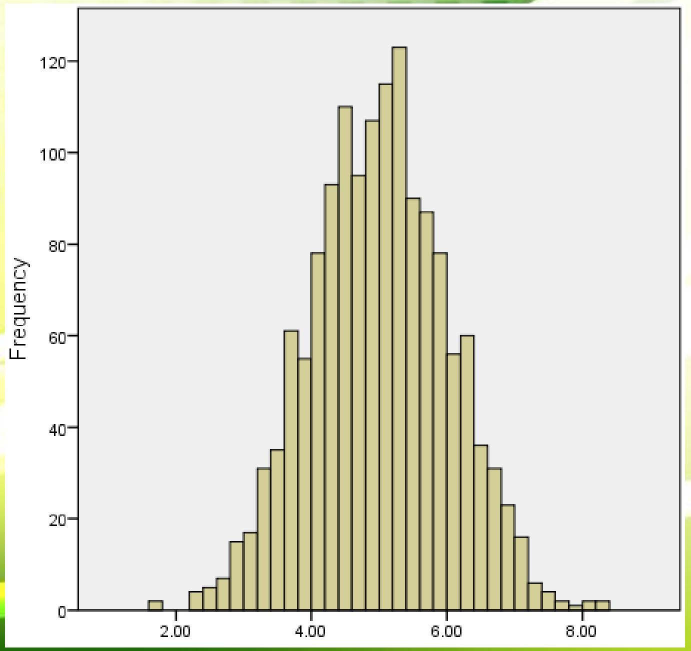
INTRODUCTION - DESCRIPTIVE STATISTICS

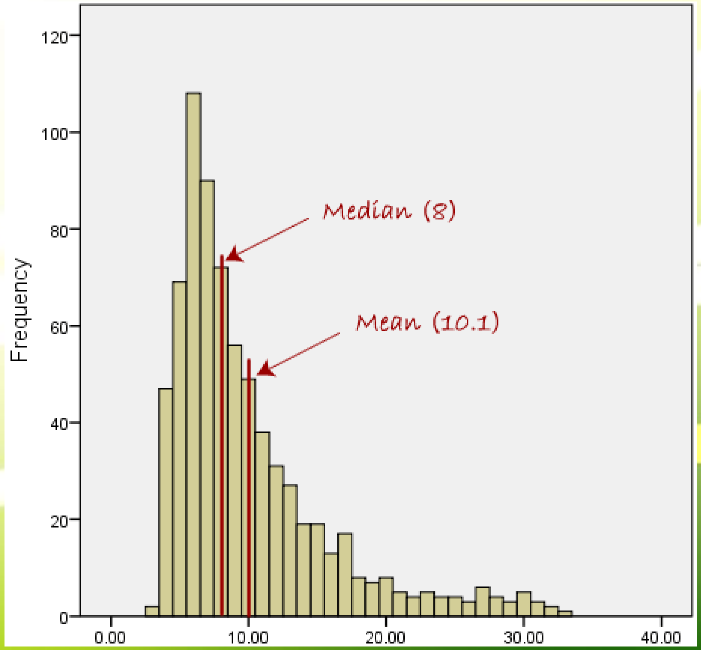
- When shape of the **distribution is symmetric and unimodal**, the mean, median, and mode are equal.



INTRODUCTION - DESCRIPTIVE STATISTICS

- For symmetrical distribution of continuous data, all three measures of central tendency hold good.
  - Mean is preferred.
- For skewed distribution, best measure of finding central tendency is median.
- For original data, both median and mode are the best choice of measuring central tendency.
- For categorical data, mode is the best choice to find central tendency.





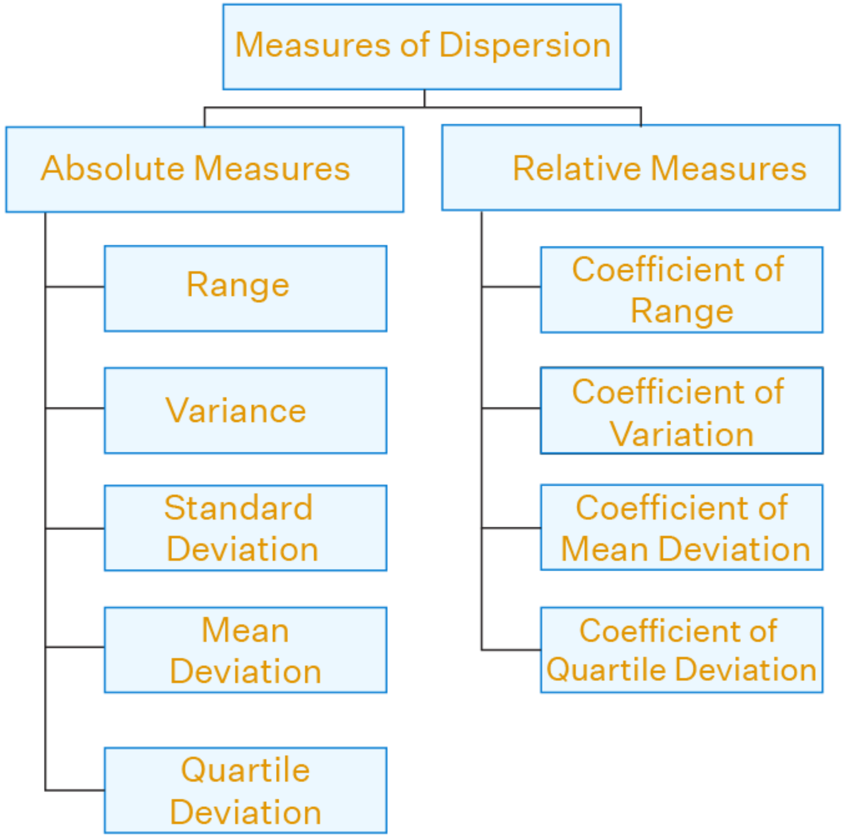
INTRODUCTION - DESCRIPTIVE STATISTICS

- Statistical dispersion means the extent/degree to which a numerical data is likely to vary about an average value (i.e. *distribution/spread of data*).
- Measures of dispersion help to interpret the variability of data.
  - How data differs from one another.
  - How much homogenous or heterogeneous the data is (*how squeezed or stretched the variable is*).
  - Always a non-negative real number.
  - Zero when all data is the same; and rises as data gets more varied.
- Two main types of dispersion methods:
  - Absolute Measure of Dispersion
  - Relative Measure of Dispersion

INTRODUCTION - DESCRIPTIVE STATISTICS

- **Absolute Measure of Dispersion** contains same unit as that of original data set.
  - Expresses variations in terms of average of deviations of observations.
  - *range, standard deviation, mean deviation, quartile deviation, etc.*
- **Relative measure of dispersion** used unit free comparison of distributions of two or more data set.
  - *coefficient of range, coefficient of mean deviation, coefficient of quartile deviation, coefficient of variation, coefficient of standard deviation, etc.*

Types of Measures of Dispersion



INTRODUCTION - DESCRIPTIVE STATISTICS

Range: difference between maximum and minimum value in data set.

- Range = $X_{max} - X_{min}$
- Good indication of how dispersed the data is.
- Other measures of variability used to discover dispersion of data from central tendency measurements.
- **Merits of Range**
  - Simplest of the measure of dispersion
  - Easy to calculate & understand
  - Independent of change of origin
- **Demerits of Range**
  - Based on two extreme observations; hence get affected by fluctuations
  - A range is not a reliable measure of dispersion
  - Dependent on change of scale

INTRODUCTION - DESCRIPTIVE STATISTICS

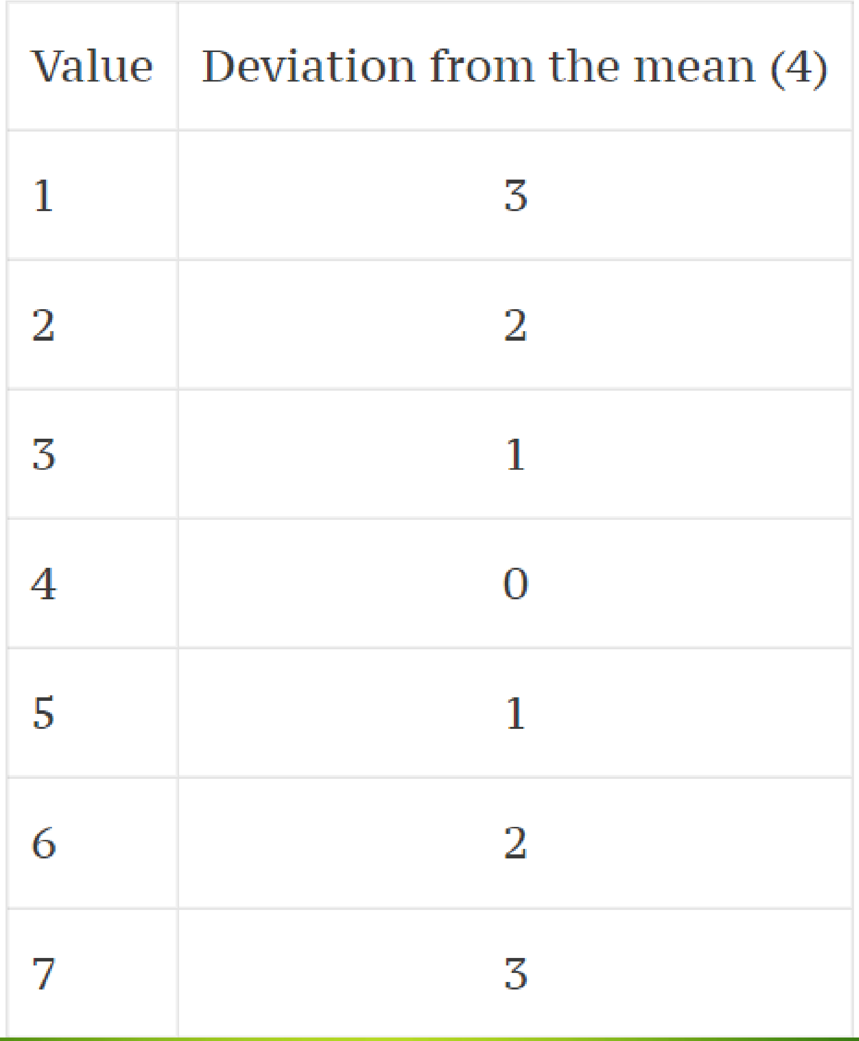
Mean Absolute Deviation (M.A.D.): average deviation of values from the mean in a sample.

1. Calculate average of the observations
2. Calculate difference of each observation from mean (deviations of all the observations).
3. Average all the deviations.

- **Merits of Mean Deviation**
  - Based on all observations
  - Provides a minimum value when the deviations are taken from the median
  - Independent of change of origin
- **Demerits of Mean Deviation**
  - Calculation is not easy and time-consuming
  - Dependent on change of scale
  - Ignorance of negative sign creates artificiality and becomes useless for further mathematical treatment

INTRODUCTION - DESCRIPTIVE STATISTICS

mean deviation = 1.714



Value	Deviation from the mean (4)
1	3
2	2
3	1
4	0
5	1
6	2
7	3

INTRODUCTION - DESCRIPTIVE STATISTICS

- **Variance**: degree of variability/spread in data set.
- Larger the spread of data, more the variance.
- Deduct mean from each data/value → square each of them → add each square → divide them by total no of values in data set. $\sigma^2 = \sum(X-\mu)^2/N$
  - *Sample A: -2, 2, 2, -2, -2, 2, -2, 2*
  - *Sample B: -1, 1, 2, -2, 2, -1, 1, -2*
  - *Mean for both these samples is '0'.*
  - *Variance for sample A is '4'*
  - *Variance for sample B is '2.5'.*

    → *Sample A has a more widespread set of data.*

INTRODUCTION - DESCRIPTIVE STATISTICS

- **Standard Deviation**: Square root of variance.

  $$\mathbf{S.D. = \sigma}$$

  $$\sigma^2 = \sum(X-\mu)^2/N$$

- A low standard deviation means values tend to be approaching the mean, while a high standard deviation indicates values are spread out across a wider range.
  - *Sample A: -2, 2, 2, -2, -2, 2, -2, 2*
  - *Sample B: -1, 1, 2, -2, 2, -1, 1, -2*
  - *Mean for both these samples is ‘0’.*
  - *Standard deviation for Sample A: 2*
  - *S.D. for Sample B: $\sqrt{2.5}$*

INTRODUCTION - DESCRIPTIVE STATISTICS

Quartiles are values that divide data set into quarters.

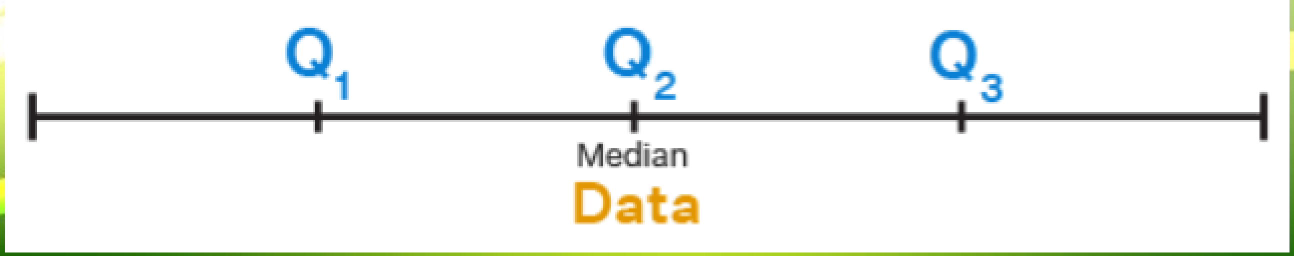
- First quartile (Q_1) : middle number between smallest number and median of data.
- Second quartile (Q_2) : median of data set.
- Third quartile (Q_3) : middle number between median and largest number.
  - 1st quartile or lower quartile separates the lowest 25% of data from the highest 75%.
  - 2nd quartile or middle quartile it divides numbers into 2 equal parts.
  - 3rd quartile or the upper quartile separates the highest 25% of data from the lowest 75%.
- **Interquartile Range (IQR)**, also called mid-spread: difference between 75th and 25th percentiles, or between upper and lower quartiles,

Q1 = [(n+1)/4]th item

Q2 = [(n+1)/2]th item

Q3 = [3(n+1)/4]th item

IQR = Q3 − Q1



INTRODUCTION - DESCRIPTIVE STATISTICS

Quartile Deviation for Grouped Data

Class Interval of Marks	45 - 50	50-55	55 - 60	60-65	65-70	70-75
Number of Students *(Frequency)*	7	5	12	11	9	6

Q_r = the r^th quartile

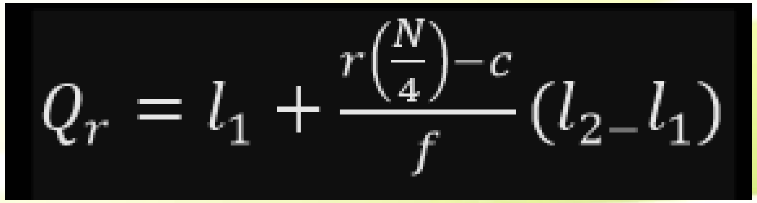
l_1 = lower limit of the quartile class

l_2 = upper limit of the quartile class

f = frequency of the quartile class

c = cumulative frequency of the class preceding the quartile class

N = Number of observations in the given data set



$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right)-c}{f}(l_{2-}l_1)$$

INTRODUCTION - DESCRIPTIVE STATISTICS

- **Quartile Deviation**: half of the distance between third and first quartile *(Semi Interquartile Range)*.

$$\textbf{Q.D.} = \tfrac{1}{2} \times (Q_3 - Q1)$$

- **Merits of Quartile Deviation**
  - All drawbacks of Range are overcome by quartile deviation
  - Uses half of the data (Ignores 50% of the data → demerit)
  - Independent of change of origin (Dependent on change of scale → demerit)
  - Best measure of dispersion for open-end classification

Height	Height
140-150	< 150
150-160	150-160
160-170	160-170
170-180	170-180
180-190	> 180

INTRODUCTION - DESCRIPTIVE STATISTICS

Find the quartile deviation for the following given data.
23, 8, 5, 16, 33, 7, 24, 5, 30, 33, 37, 30, 9, 11, 26, 32

Arrange data in ascending order.
5, 5, 7, 8, 9, 11, 16, 23, 24, 26, 30, 30, 32, 33, 33, 37

Calculation of Q1,
Q1 position = (8, 9)
$Q_1 = (8+9)/2 = 8.5$

Calculation of Q2,
Q2 position = (23, 24)
$Q_2 = (23+24)/2 = 23.5$

Calculation of Q3,
Q3 position = (30, 32)
$Q_3 = (30 + 32)/2 = 31$

Quartile Deviation $= (Q3-Q1)/2=(31-8.5)/2=22.5/2=11.25$

Quartile deviation is 11.25

Coefficient of Quartile Deviation
$= (Q3-Q1)/(Q3+Q1)=(31-8.5)/(31+8.5)$
$=22.5/39.5=0.57$

Coefficient of quartile deviation is 0.57

INTRODUCTION - DESCRIPTIVE STATISTICS

Find the quartile deviation for the following given data.
23, 8, 5, 16, 33, 7, 24, 5, 30, 33, 37, 30, 9, 11, 26, 32

Arrange data in ascending order.
5, 5, 7, 8, 9, 11, 16, 23, 24, 26, 30, 30, 32, 33, 33, 37

Calculation of Q1,
Q1 position = ¼ (16 + 1) =¼ (17) **Q1 = 4.25th Term**
Q_1 = *8+(9-8)*0.25 = 8+1*0.25 = 8.25*

Calculation of Q3,
Q3 position = ¾ (16 + 1) =¾ (17) **Q3= 12.75 Term**
Q_3 = *30+(32 - 30)*0.75 = 30+2*0.75 = 31.5*

Quartile Deviation = (Q3−Q1)/2=(31.5−8.25)/2=23.25/2=11.63

Quartile deviation is 11.63

Coefficient of Quartile Deviation
= (Q3−Q1)/(Q3+Q1)=(31.5−8.25)/(31.5+8.25)
=23.25/39.75=0.58

Coefficient of quartile deviation is 0.58

INTRODUCTION - DESCRIPTIVE STATISTICS

Find quartile deviation of the marks scored by 50 students of a class.

Class Interval of Marks	45 - 50	50-55	55 - 60	60-65	65-70	70-75
Number of Students (Frequency)	7	5	12	11	9	6

N = 50

N/4 = 50/4 = 12.5

3N/4 = 3(50/4) = 2(12.5) = 37.5

Class containing Q_1 is 55 - 60, and the class containing Q_3 is 65 - 70.

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right)-c}{f}(l_{2-}l_1)$$

Class Interval	Frequency	Cumulative Frequency
45 - 50	7	7
50 - 55	5	7 + 5 = 12
55 - 60	12	12 + 12 = 24
60 - 65	11	24 + 11 = 35
65 - 70	9	35 + 9 = 44
70 - 75	6	44 + 6 = 50

Calculation for first quartile Q_1 :

n = 1, N = 50, f = 12, c = 12, l_1 = 55, and l_2 = 60

Q1=l1+((N/4)−c)*(l2−l1)/f

Q1=55+((50/4)−12)*(60−55)/12

Q1=55+(12.5−12)*(5)/12

Q_1 = 55 +(0.5)*5/12 = 55 + 2.5/12 = 55 + 0.2083

Q_1= 55.21

Similarly, calculation for the third quartile Q_3 .

n = 3, N = 50, f = 9, c = 35, l_1 = 65, and l_2 = 70

Q_3= 65.83

Quartile Deviation = (Q3−Q1)/2=(65.83−55.21)/2=10.62/2 **= 5.31**

INTRODUCTION - DESCRIPTIVE STATISTICS

- Harry ltd. is a textile manufacturer and is working upon a reward structure. The management is in discussion to start a new initiative, but they first want to know how much their production spread is.
- The management has collected its average daily production data for the last 10 days per (average) employee.

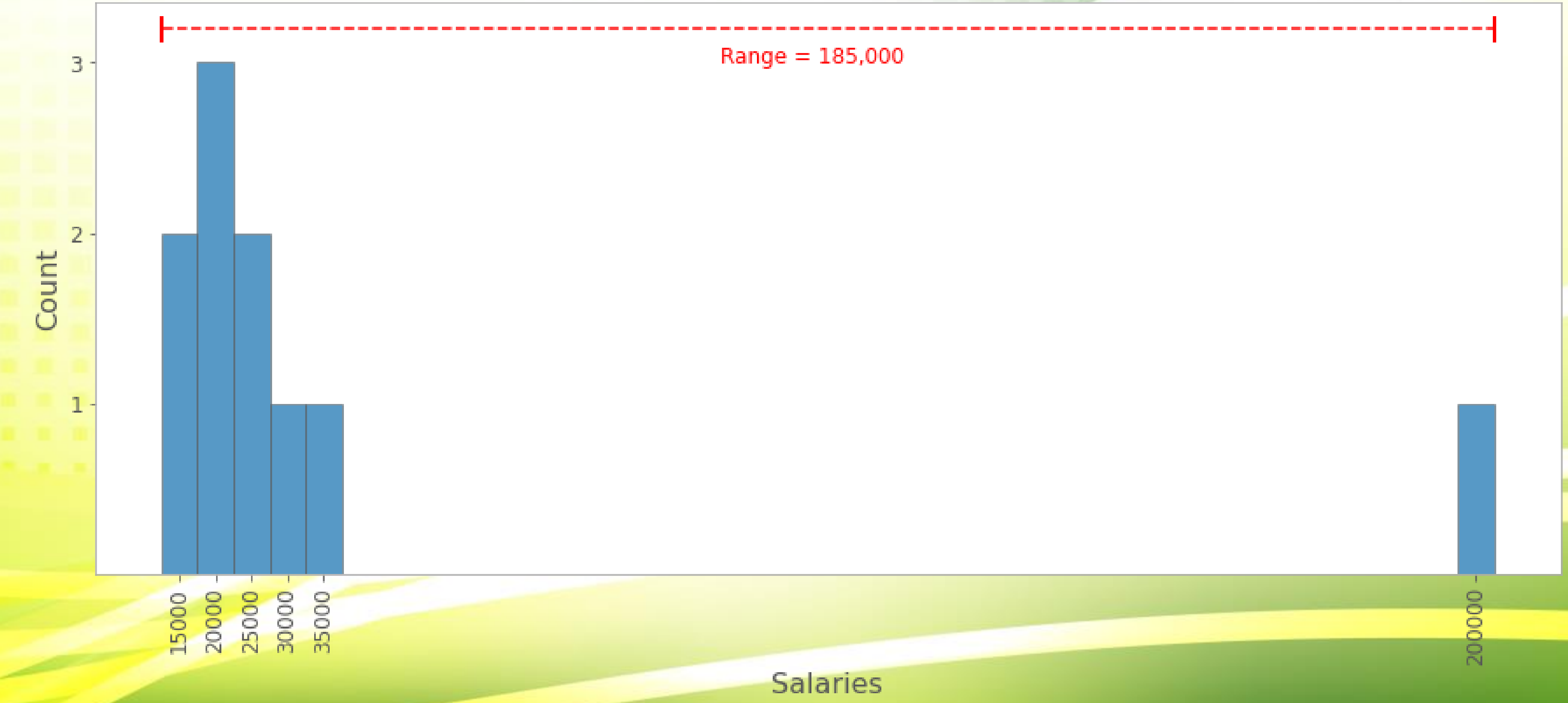
  185, 169, 188, 150, 177, 145, 140, 190, 175, 156.
- Use the Quartile formula to help management find deviation.

INTRODUCTION - DESCRIPTIVE STATISTICS

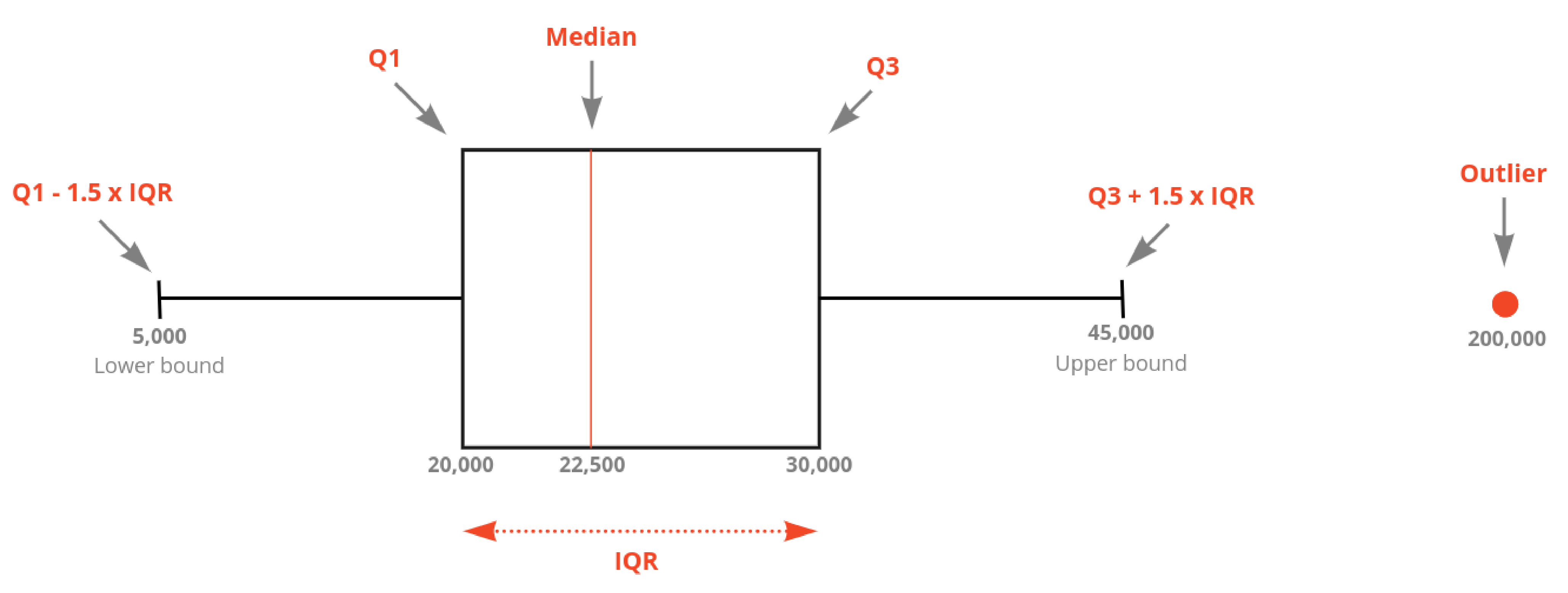
- **Co-efficient of Dispersion (C.D.)** used to compare variability of two series which differ widely in their averages.
- Dispersion coefficient is also used when two series have different measurement units.

C.D. In Terms of	Coefficient of dispersion
Range	$(X_{max} - X_{min}) / (X_{max} + X_{min})$
Quartile Deviation	$(Q3 - Q1) / (Q3 + Q1)$
Standard Deviation (S.D.)	S.D./Mean
Variation	C.V. = 100 × (S.D. / Mean)
Absolute Deviation	$\sum(X-\mu)$
Mean Absolute Deviation	M.A.D. = $\sum(X-\mu)/n$

INTRODUCTION - DESCRIPTIVE STATISTICS

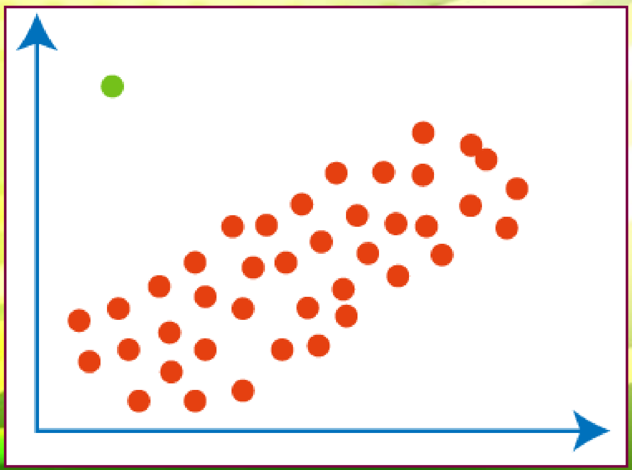


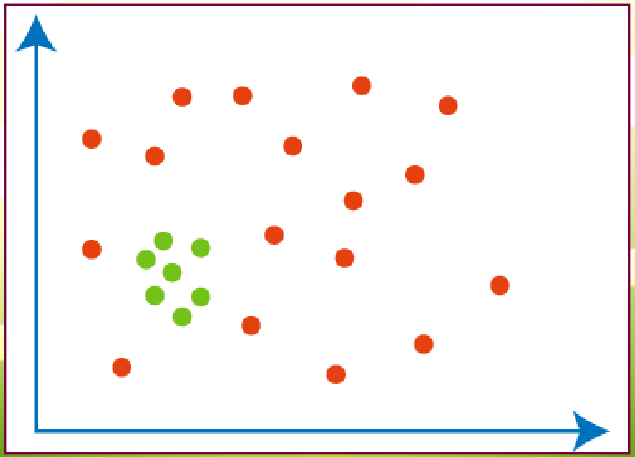
INTRODUCTION - DESCRIPTIVE STATISTICS

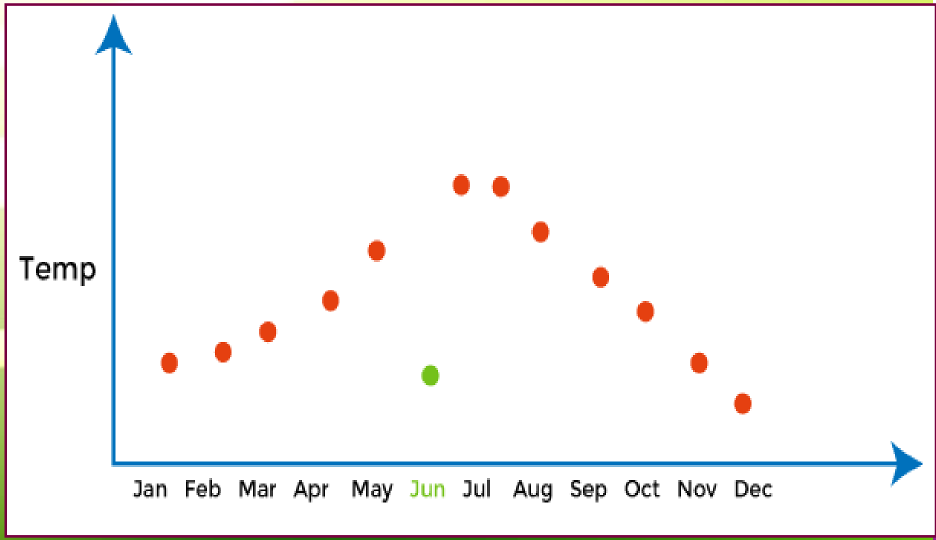


INTRODUCTION - DESCRIPTIVE STATISTICS

- An **outlier** is an observation that lies at **abnormal distance** from other values in dataset.
  - *extremely high or low data point relative to other data point.*
- **Noise**: Any unwanted error occurs in some previously measured variable.
  - Remove Noise, before finding outliers present in data set.
- Three different types Outliers:
  - Global or point outliers
  - Collective outliers
  - Contextual or conditional outliers

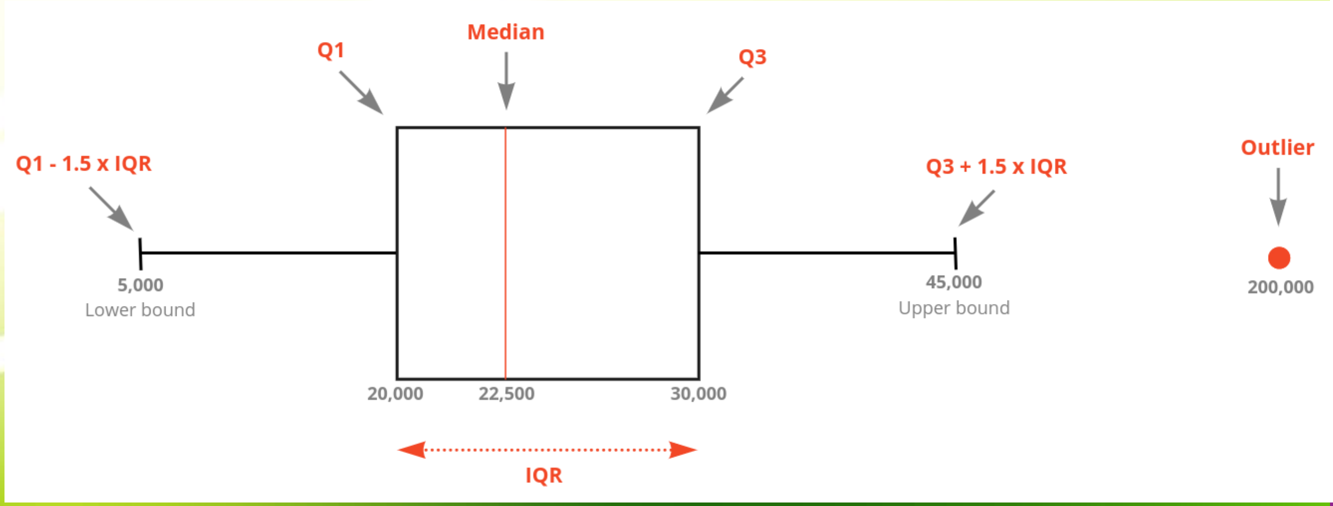






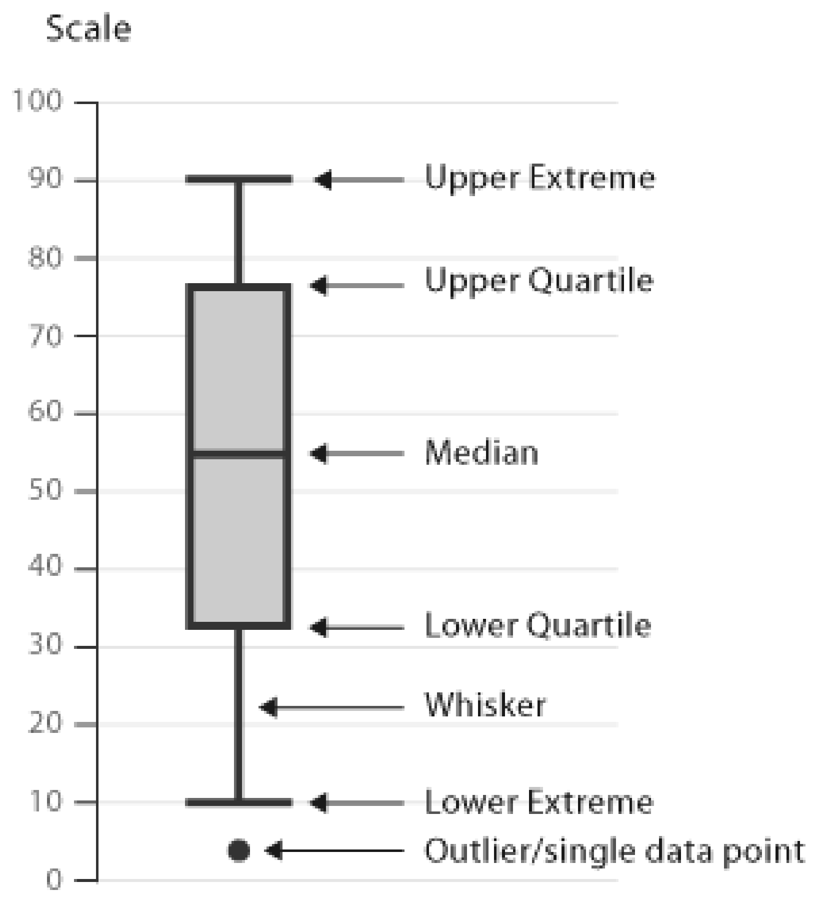
INTRODUCTION - DESCRIPTIVE STATISTICS

- Interquartile range help in **identifying outlier** (global).
  - Data point needs to fall in the 1.5 times of Interquartile range; else considered outlier.
  - Find Outlier with upper and lower bound/fence
    - lower bound/fence = Q1 - 1.5(IQR)
    - outlier < lower bound/fence
    - upper bound/fence = Q3 + 1.5(IQR)
    - outlier > upper bound/fence



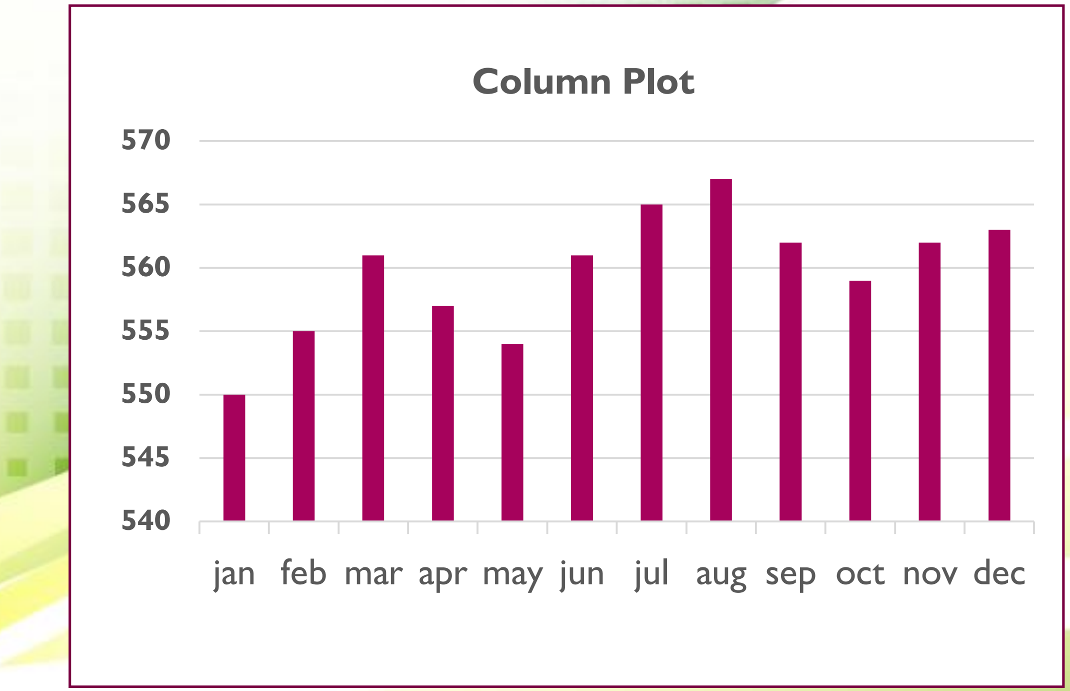
INTRODUCTION - DESCRIPTIVE STATISTICS

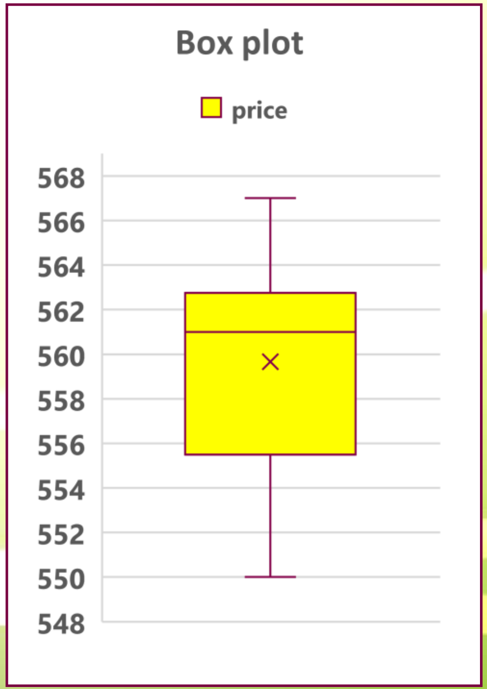
- Box plot (whisker plot) visually shows the distribution of numerical data and skewness through displaying data quartiles (percentiles) and averages.
- Shows five-number summary of a dataset :
  - minimum, 1st/lower quartile, median, 3rd/upper quartile, maximum.
- Upper/lower whiskers represent scores outside middle 50% (i.e. lower 25% and upper 25%).



INTRODUCTION - DESCRIPTIVE STATISTICS

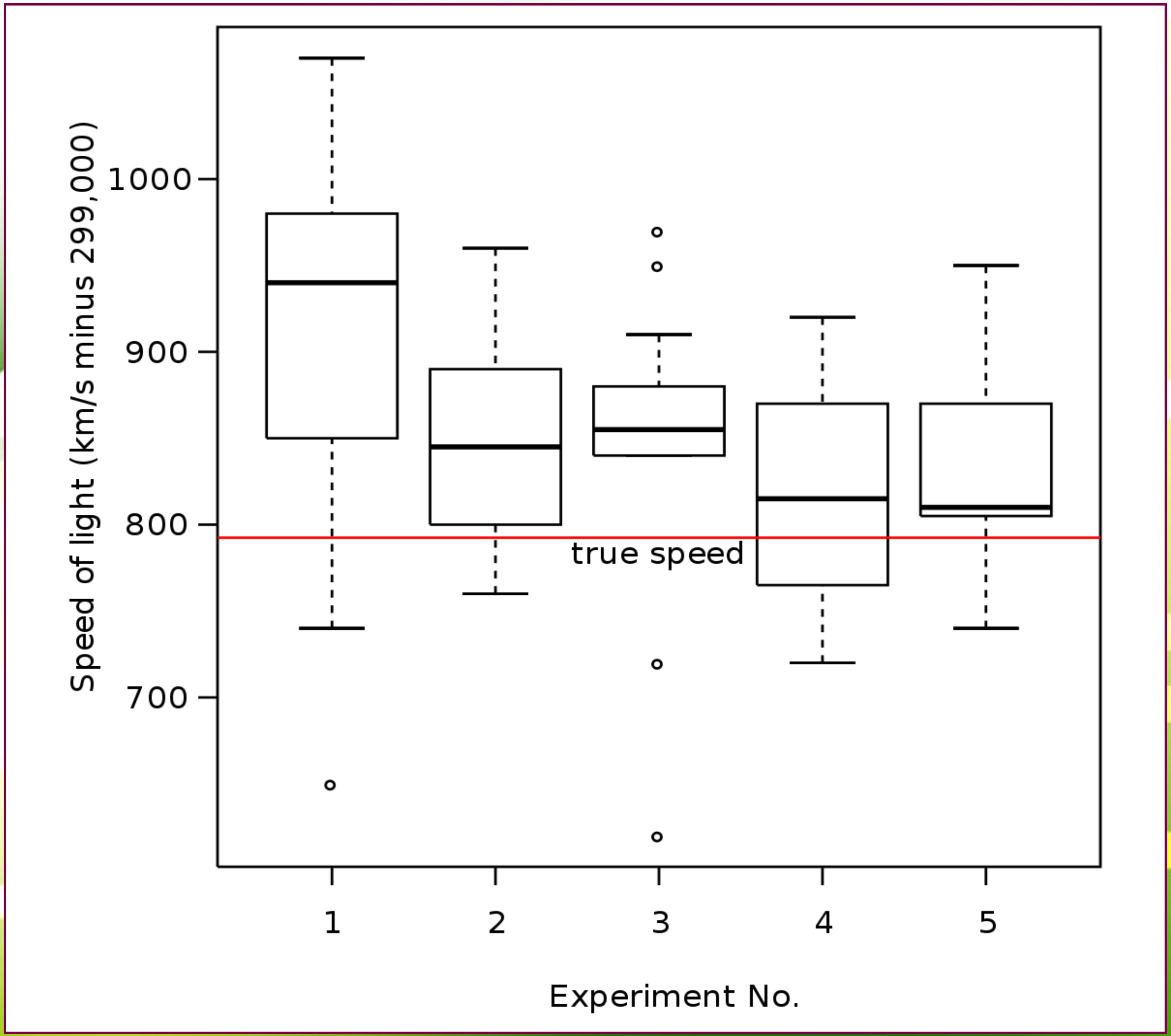
month	price
jan	550
feb	555
mar	561
apr	557
may	554
jun	561
jul	565
aug	567
sep	562
oct	559
nov	562
dec	563



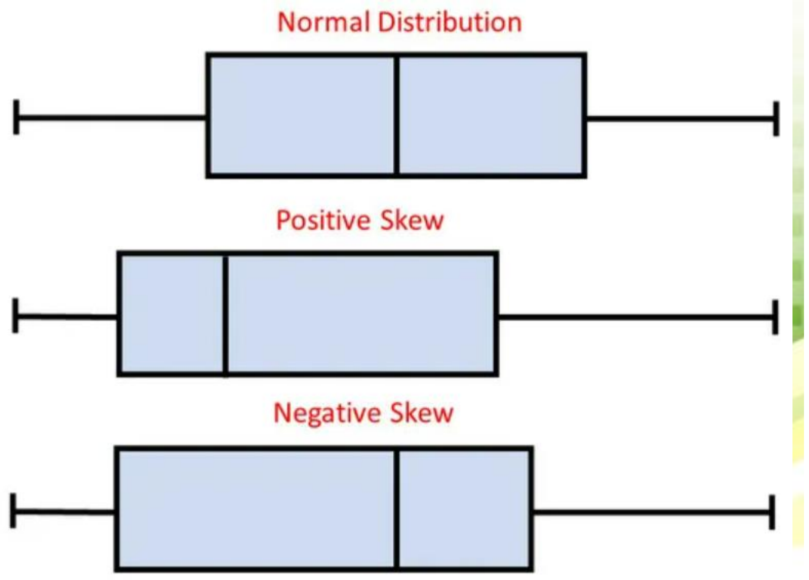


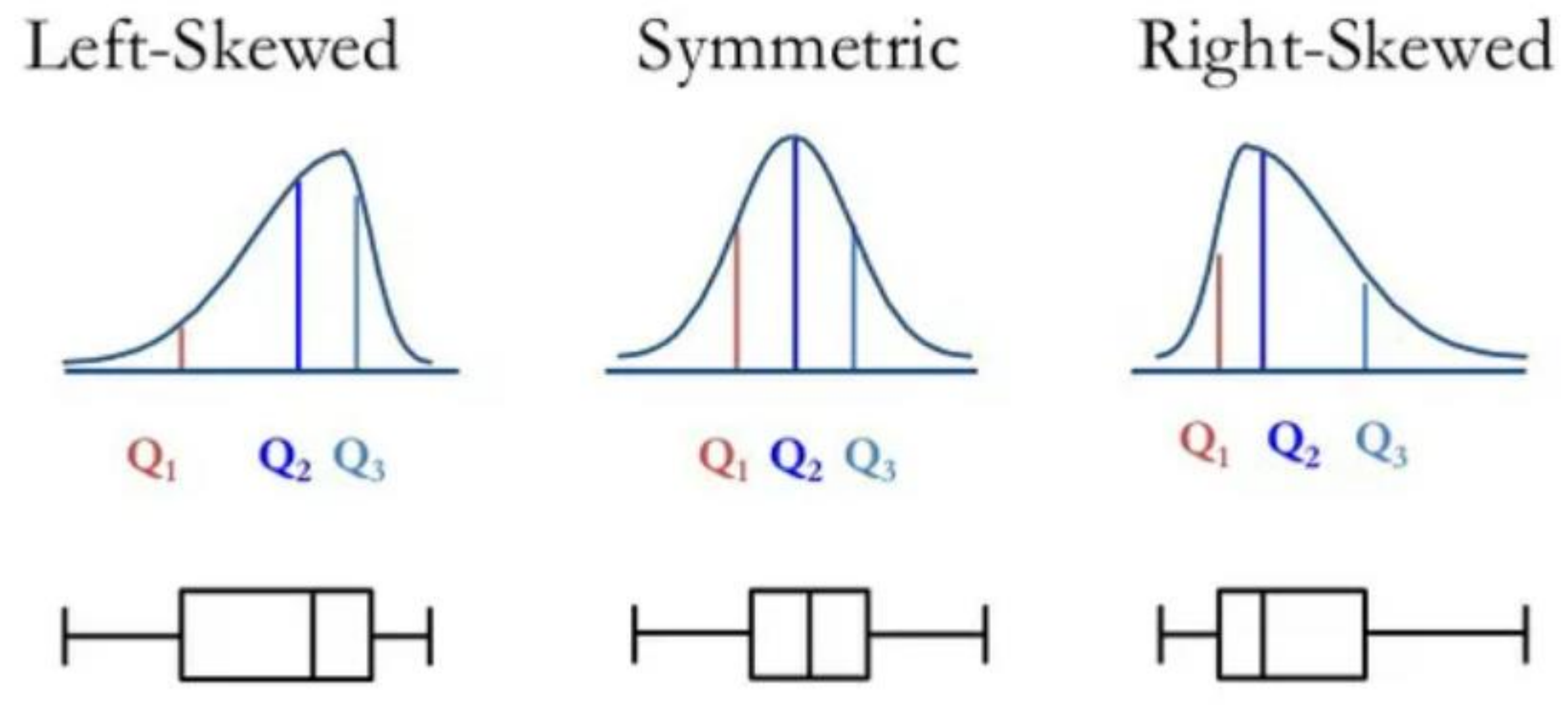
INTRODUCTION - DESCRIPTIVE STATISTICS

month	A-price	B-price	C-price
jan	550	850	650
feb	555	455	595
mar	561	661	661
apr	557	575	574
may	554	654	543
jun	561	751	631
jul	565	865	615
aug	567	697	571
sep	562	862	623
oct	559	592	692
nov	562	682	652
dec	563	763	531



INTRODUCTION - DESCRIPTIVE STATISTICS





INTRODUCTION - DESCRIPTIVE STATISTICS

- Harry ltd. is a textile manufacturer and is working upon a reward structure. The management is in discussion to start a new initiative, but they first want to know how much their production spread is.
- The management has collected its average daily production data for the last 10 days per (average) employee.

  105, 139, 158, 120, 127, 115, 140, 155, 125, 116.
- Help management find;
  - Q1, Q2, Q3, Max, Min, Range, ICR, Whisker range, Mean, Upper bound, Lower Bound.
  - whether there is any outlier in the given data?
  - What type of distribution does the data forms?