COMP1810- DATA AND WEB ANALYTICS
PAPER EXAM STUDY NOTES

1. **Critically evaluate the terminology "Clickstream" as regard to web data analytics and describe the core method of extracting "Clickstream" from a website. [10 Marks]**

Clickstream:

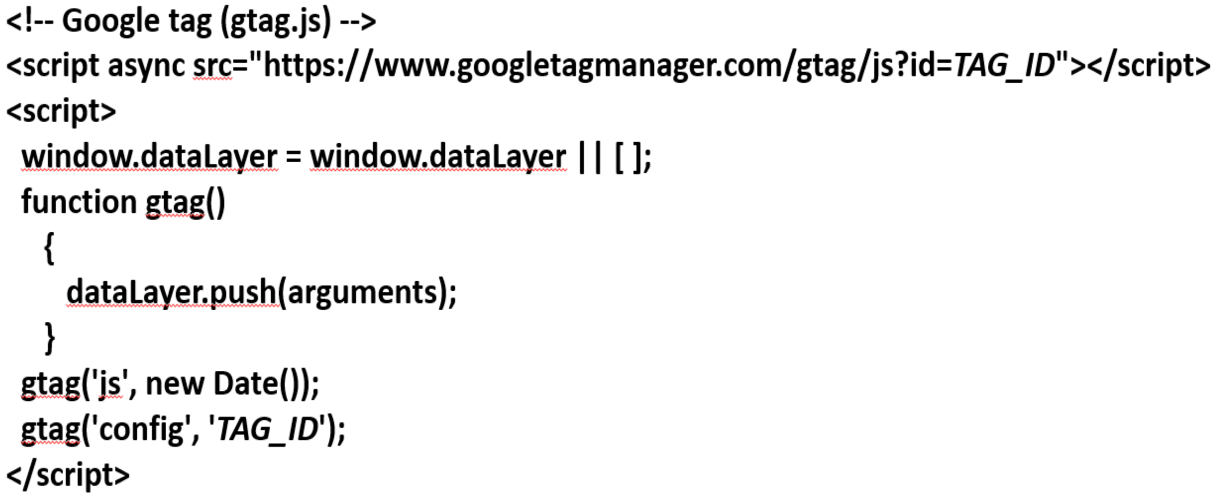
Clickstream data are a detailed log of how participants navigate through the Web site during a task. The log typically includes the pages visited, time spent on each page, how they arrived on the page, and where they went next. From an aggregate perspective, clickstream data provide tremendous insights into how easily the site is navigated, what pages are causing the greatest confusion, and what pages are critical in reaching a desired destination.

The Click Stream data helps to tell customer's story while visiting your websites. Any event that happened while visitor is on the webpage are collected by JavaScript tracker code placed on each html page. This made it possible to track all the event or how the customer engaged with the pages. The Java script could track the following events like video played, downloads, registration, purchases and many more.

Method of Extracting the Clickstream data:

Google analytics is an online platform, the user can Register and sign in. To collect the Clickstream. The user can be provided with a code (javaScript) by google analytics. This code is attached to each page of the website immediately after the head tag <head>. Each time the page is loaded the users Id is generated and all activities for example Page View", "Bounce rate", "Purchases", "Time Spent on Pages. All the these are sent into the Analytics account or any other collecting software.

The following JS code is used for tracking the activities in web pages.



```
<!-- Google tag (gtag.js) -->
<script async src="https://www.googletagmanager.com/gtag/js?id=TAG_ID"></script>
<script>
 window.dataLayer = window.dataLayer || [ ];
 function gtag()
   {
    dataLayer.push(arguments);
   }
 gtag('js', new Date());
 gtag('config', 'TAG_ID');
</script>
```

Web Scrapping Tools to extract the Clickstream from website

An automated technique known as web scraping is used to retrieve large amounts of data from websites. Usually, the internet contains unstructured data. Web scraping aids in the collection and storage of unstructured data. There are several methods for scraping websites, including the use of internet services, APIs, or custom programs. Web scraping is used to collect a large amount of data (Statistics, General Information, Temperature, etc.) from websites, which is then analysed and used to conduct surveys or R&D. Web scraping tools are designed to help businesses collect open-source web data, such as:

- Synthesized
- Cleaned
- Structured
- Processed
- Analysed by teams, and algorithms.

Web scraping can be done manually, but it is a time-consuming and resource-intensive effort, so many businesses prefer to use a tool to help automate this process.

Most common use cases for which businesses are currently employing web scraping tools:

Market Research: Industries seeking to introduce new products or enter new markets gather information on potential target audiences while also investigating successful competitor activities that can be replicated/learned from.

- ***Stock Market Data:*** Hedge funds, portfolio managers, and venture capitalists all collect financial data such as share market capacity, enterprise media articles, and growth based on employee count or geospatial data (e.g. satellite imagery on the progress of a building site or factory).
- ***Travel Aggregation:*** To better compete, online travel agencies (OTAs) gather real-time information about competing sites' vacation bundles, special offers, and flight/car rental/hotel pricing.
- ***Food Delivery Market:*** As demand for food delivery has increased over the last two years, companies are increasingly looking to collect restaurant menu data, trending cuisines via search (Chinese? Japanese? etc.), and order volume based on consumer geolocation.

- ***Collection of Search Engine Optimization (SEO) / Search Engine Results Pages (SERPs):*** Many consumer journeys start with a simple search query, propelling businesses to the top of search engine results. As a result, they gather and analyse top search results for relevant search queries and keywords in their industry to optimise their own pages and rank higher in the future.
- ***Website Testing****:* Companies that build sites/apps for different geographies or that launch new User Experiences (UX) and User Interfaces (UI) use web scraping tools to view front-end results from a consumer standpoint. This allows them to improve their Quality Assurance (QA) and load balancing.
- ***eCommerce:*** This is a highly competitive field with many value-conscious customers. Product pricing, customer reviews, Sell-Through Rates (STRs), and other data points are collected by vendors, marketplaces, and brands to optimise item listings, design, and production lines in order to capture higher conversion rates.
- ***AdTech:*** Marketing teams and agencies use web scraping tools to ensure that localised campaigns are shown to target audiences with the correct copy, images, and URLs. They also gather data on competitor ad campaigns to gain insights and optimise campaigns for higher click-through rates (CTRs).
- ***Social Media for Marketing:*** Companies use web scraping tools to gain insights into their target audience's social sentiment, to find influencers with whom they can collaborate, and to identify posts that consumers are engaging with so that they can join the narrative and generate newfound interest.

2. Provide a detailed Review of the components of the Clickstream. [40 Marks]

The following are the components and the constraints of the Clickstream Data Analysis.

1. **Dimension.**

   A descriptive attribute or characteristic of data. Browser, Landing Page and Campaign are all examples of default dimensions in Analytics. A dimension is a descriptive attribute or characteristic of an object that can be given different values. For example, a geographic location could have dimensions called Latitude, Longitude, or City Name. Values for the City Name dimension could be San Francisco, Berlin, or Singapore. Browser, Exit Page, Screens, and Session Duration are all examples of dimensions that appear by default in Analytics. Dimensions appear in all of your reports, though you might see different ones depending on the specific report. Use them to help organize, segment, and analyse your data.

2. **Metric.**

   The Examples of clickstream data metrics include (1) number of page views, (2) pattern of websites visited, including most frequent exit page and prior website, (3) length of stay on the website, (4) dates and times of visits, (5) number of registrations filled out per 100 visitors, (6) number of abandoned registrations, (7) demographics of registered visitors, (8) number of customers with shopping carts, and (9) number of abandoned shopping carts.

3. **Bounce rate.**
   This is the rate at which customer enters a web page and leave immediately without doing anything. It is a significant indication that all is not well with the web page. Why are customer's getting to the page and just leaving without engaging.

4. **Pageview.**
   The Pageview illustrates the most popular page that users view and this will enable the owners to optimise the particular pages, and questions like; what made such pages popular? Can such information be used to improve other pages? So that other pages could become popular also. How can these popular pages be commercialised to generate incomes.

5. **Conversion Rates**
   Another key benefit of analyzing clickstream data is the ability to optimize conversion rates. By tracking user behavior through the conversion funnel, businesses can identify pain points and optimize the user journey to drive more conversions. This can help increase revenue and maximize the return on investment for digital marketing efforts.

6. **Goal**
   Goal and KPI are synonymous both are used to assess the performance of the website or pages. While KPI is a metric of measurement, the goal is threshold metric value. If KPI is downloading a movie. The Goal 200 people downloading movies per day. If KPI is people registering on your website. The Goal 100 people registering on website per day.

7. **KPI**
   KPI stands for **Key Performance Indicators.** This a term used to assess the performance of any metric. What metric should be used to measure the performance of websites? KPI are metrics set by the owners of the business. For example, it may be a particular value of metric eg downloading a movie in a day or people registering on your website. It could also be a collection of metrics that must be watched to know how the website is performing. KPI is a specifics target of measurement that shows that the website is performing or not.

3. Critical review the terms with an appropriate example “Null hypothesis” with an appropriate example. [12Marks]

Definition of Null Hypothesis:

The **null hypothesis** is a kind of hypothesis which explains the population parameter whose purpose is to test the validity of the given experimental data. This hypothesis is either rejected or not rejected based on the viability of the given population or sample. In other words, the null hypothesis is a hypothesis in which the sample observations result from the chance. It is said to be a statement in which the surveyors want to examine the data. It is denoted by H_0.

In statistics, the null hypothesis is usually denoted by **letter H with subscript ‘0’ (zero), such that H_0. It is pronounced as H-null or H-zero or H-nought.**

When the experiment is carried out to test the Hypothesis. Based on the result the possible outcome of the null hypothesis is either.

reject null hypothesis

fail to reject null hypothesis

We don’t used the word accept in output of Hypothesis. Accepting it mean that the result is specifically correct which are not the case.

Null Hypothesis Formula

Here, the hypothesis test formulas are given below for reference.

The formula for the null hypothesis is:

$$\mathbf{H_0: p = p_0}$$

The null hypothesis says there is no correlation between the measured event (the dependent variable) and the independent variable. We don’t have to believe that the null hypothesis is true to test it. On the contrast, you will possibly assume that there is a connection between a set of variables (dependent and independent).

Null Hypothesis Examples

Ex-1 If a medicine reduces the risk of cardiac stroke, then the null hypothesis should be “the medicine does not reduce the chance of cardiac stroke”. This testing can be performed by the administration of a drug to a certain group of people in a controlled way. If the survey shows that there is a significant change in the people, then the hypothesis is rejected.

Ex-2 Do teenagers are using mobile phones more than grown-ups to access the internet?

Ans: Age has no limit on using mobile phones to access the internet.

Ex-3 Does having an apple daily will not cause fever?

Ans: Having apple daily does not assure of not having fever but increases the immunity to fight against such diseases.

Ex-4 Do the children better in doing mathematical calculations than grown-ups?

Ans: Age has no effect on Mathematical skills.

In many common applications, the choice of the null hypothesis is not automated, but the testing and calculations may be automated. Also, the choice of the null hypothesis is completely based on previous experiences and inconsistent advice. The choice can be more complicated and based on the variety of applications and the diversity of the objectives.

4. Critically review the term with an appropriate example "Alternative hypothesis" with an appropriate example. [12Marks]

Definition of Alternate Hypothesis:

Alternative hypothesis defines there is a statistically important relationship between two variables. Whereas null hypothesis states there is no statistical relationship between the two variables.

The alternative hypothesis is a statement used in statistical inference experiment. It is contradictory to the null hypothesis and **denoted by H_a or H_1**. We can also say that it is simply an alternative to the null. In hypothesis testing, an alternative theory is a statement which a researcher is testing.

This statement is true from the researcher's point of view and ultimately proves to reject the null to replace it with an alternative assumption. In this hypothesis, the difference between two or more variables is predicted by the researchers, such that the pattern of data observed in the test is not due to chance.

Types of Alternate Hypothesis

Basically, there are three types of the alternative hypothesis, they are;

Left-Tailed: Here, it is expected that the sample proportion (π) is less than a specified value which is denoted by π_0, such that;

$$H_1 : \pi < \pi_0$$

Right-Tailed: It represents that the sample proportion (π) is greater than some value, denoted by π_0.

$$H_1 : \pi > \pi_0$$

Two-Tailed: According to this hypothesis, the sample proportion (denoted by π) is not equal to a specific value which is represented by π_0.

$$H_1 : \pi \neq \pi_0$$

Note: The null hypothesis for all the three alternative hypotheses, would be $H_1 : \pi = \pi_0$.

Examples of Alternate Hypothesis:

Ex-1 To check the water quality of a river for one year, the researchers are doing the observation. As per the null hypothesis, there is no change in water quality in the first half of the year as compared to the second half. But in the alternative hypothesis, the quality of water is poor in the second half when observed.

Ex-2: Rohan will win less than Rs.100000 in lucky draw.

Ex-3: The company staff believed that the bread mixer make bread whose weight is 10g. But the customers think that the weight of the bread is below 10g , represent this as Null and alternative Hypothesis

Ho: $\mu = 10g$ This is the null Hypothesis.

Ha: $\mu \neq 10g$ This is the alternative Hypothesis. (This is the one that should be tested) mathematically opposite of other.

Ex-4: The advertisement campaign may or may not be effective in increasing visitors to website.

Ho: In null (no difference) form:

There are **no significant differences** between the number of visitors to the website before and after the advertisement campaigned.

Ha: In alternative (“difference”) form:

There are **significant differences** between the number of visitors to the website before and after the advertisement campaigned.

Both Ho and Ha above are correct. Let Assume that the average visitor to the website weekly is 200 before the advertisement.

The Ho and Ha could also be written as

Ho: $\mu = 200$

Ha: $\mu \neq$ or > 200

Hence it will be helpful to know what use to be the visitor’s numbers before.

5. **With the suitable diagram, provide a detailed review of the term correlation and their types. [10 Marks]**

Definition of Correlation:

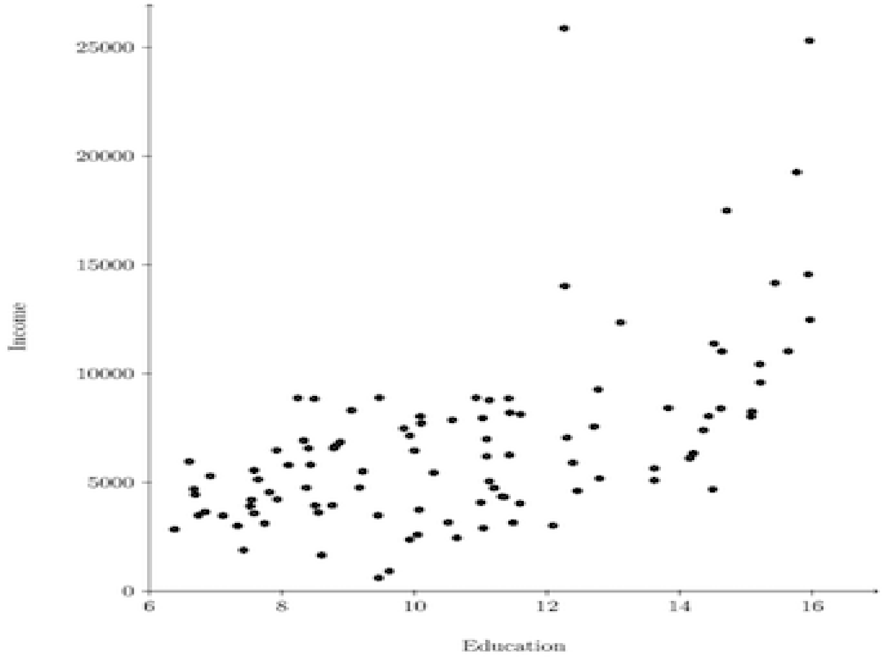
Correlation Analysis explains relationship between variables. It is concerned majorly with Correlation and Covariance.

Covariance and Correlation practically measure the same relationship between two random variables, but with different approach.

While Covariance is a measure of the relationship between two random variables and from infinite -∞ to +∞ values, correlation measure from –1 to + 1. Therefore, correlation values are standardized, hence make more sense.

Correlation refers to a process for establishing the relationships between two variables.

Correlation is used to describe how data sets are related to one another. Correlation can be seen when two sets of data are graphed on a **scatter plot**, which is a graph with an X and Y axis and dots representing the data points.

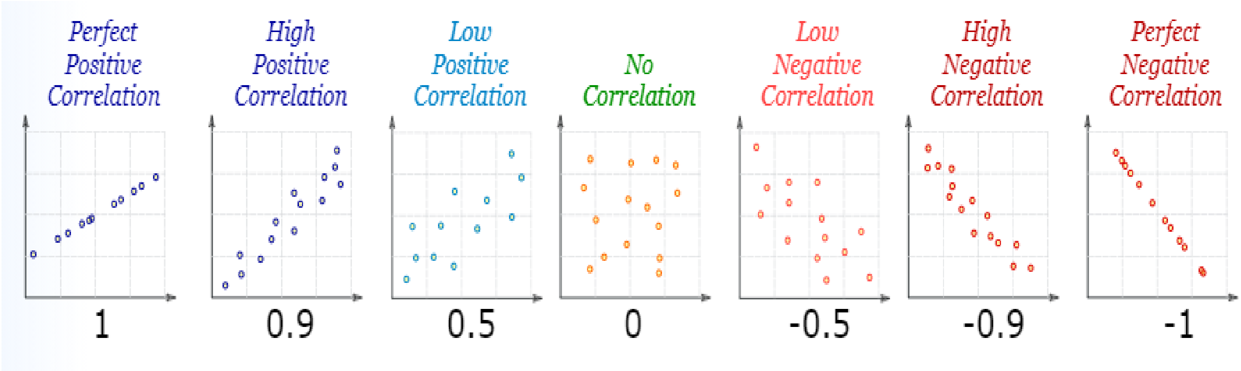


Types of Correlation

The scatter plot explains the correlation between the two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables

- **Positive Correlation** – when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.

- **Negative Correlation** – when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.
- **No Correlation** – when there is no linear dependence or no relation between the two variables.



When the temperature of the day is high more ice cream is sold. This is **positive correlation** since both are moving in the same direction (High Temperature, More ice cream). Two variables may also **correlate negatively**. Both moving in opposite direction. An example of negative correlation is the amount of money you spend for heating vs the temperature of the day. That is heat amount get higher as temperature get lower. Correlation is measured between +1 to –1. Scattered point graph like and 10 are used to show correlation between two variables.

Positive and negative is not the only way to describe correlation; correlation can also be described by its strength. Data sets can also have perfect correlation, strong correlation, or weak correlation. The closer the data points are together and the more they form a straight line, the stronger the correlation. If the data points form a perfect straight line, the data sets are said to have perfect positive or negative correlation depending on which way the line is going (up and right = positive, down and right = negative).

Calculating Correlations:

Pearson Correlation Coefficient Formula

The most common formula is the Pearson Correlation coefficient used for linear dependency between the data sets. The value of the coefficient lies between -1 to +1. When the coefficient comes down to zero, then the data is considered as not related. While, if we get the value of +1, then the data are positively correlated, and -1 has a negative correlation.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Importance of Correlation:

Correlation is one of the techniques for attributes selections in machine learning. Most datasets are made up of many variables, these variables are correlated with each other and the target labels.

6. Provide a detailed explanation and justification of the techniques to replace a missing value in a dataset. [10 marks]

Data Understanding is very important because of the following reasons

- Provides insight into the type of error in the datasets.
- Knowing the types of errors would help in deciding how to correct them.
- Provide insight into the initial and deeper analysis to use on the dataset.

Data Profiling

In estimating data quality, is a whole lot of processes called data profiling. profiling is the process of evaluating the data sources, checking the data types, contents to assess the quality, identify potential errors, in other to propose the techniques for quality improvement.

Factor that Affects data quality:

- Data quality dimensions.
- Data errors.

Correcting Data Error and cleaning

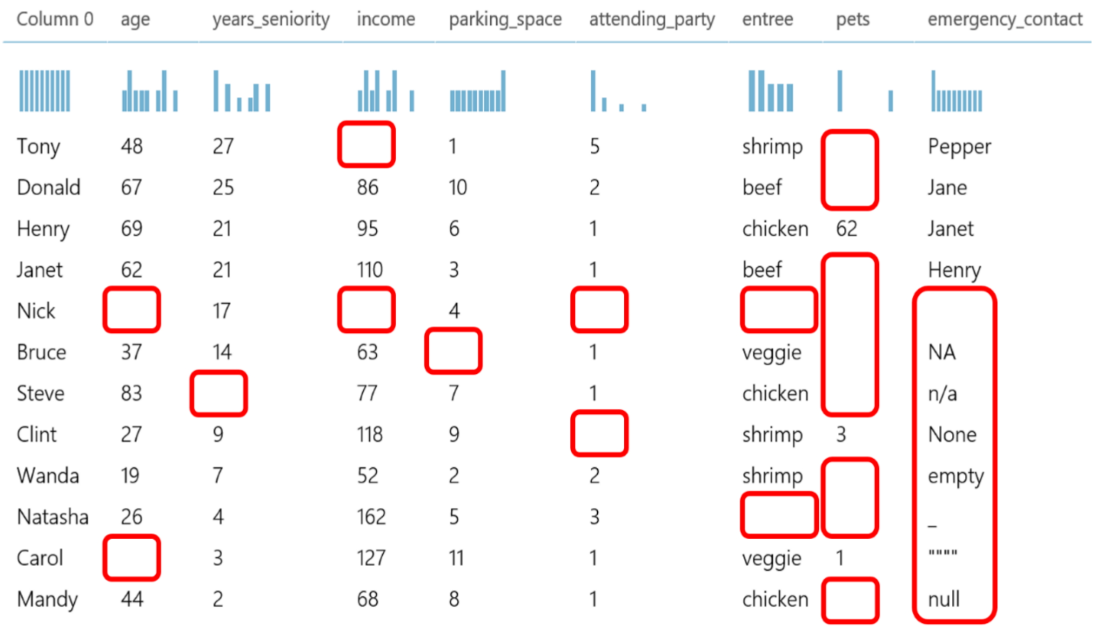
All data collected are usually "Dirty" or has "Errors":

"dirty data" is a term used in describing the different states of the rawness of data that could impact the ability to extract information from the data set. The **dirty data must be clean** by the process of detecting, correcting or removing noise or errors in the data set.

To put it in perspective, what makes the data dirty? Dirty data are regarded as having the following issues listed below among many others.

Incomplete or Missing data

If any position were a data item should be, have been left blank, nothing is written or some odd values or character in the place to indicate a missing value.



Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Duplicate data: mistakenly repeating row in a table more than once.

Incorrect data type: this is when wrong data types were used. For example if the age of a person is 36 years, an error was made by inputting the alphabet "wy" in place of 36 due to typographic error.

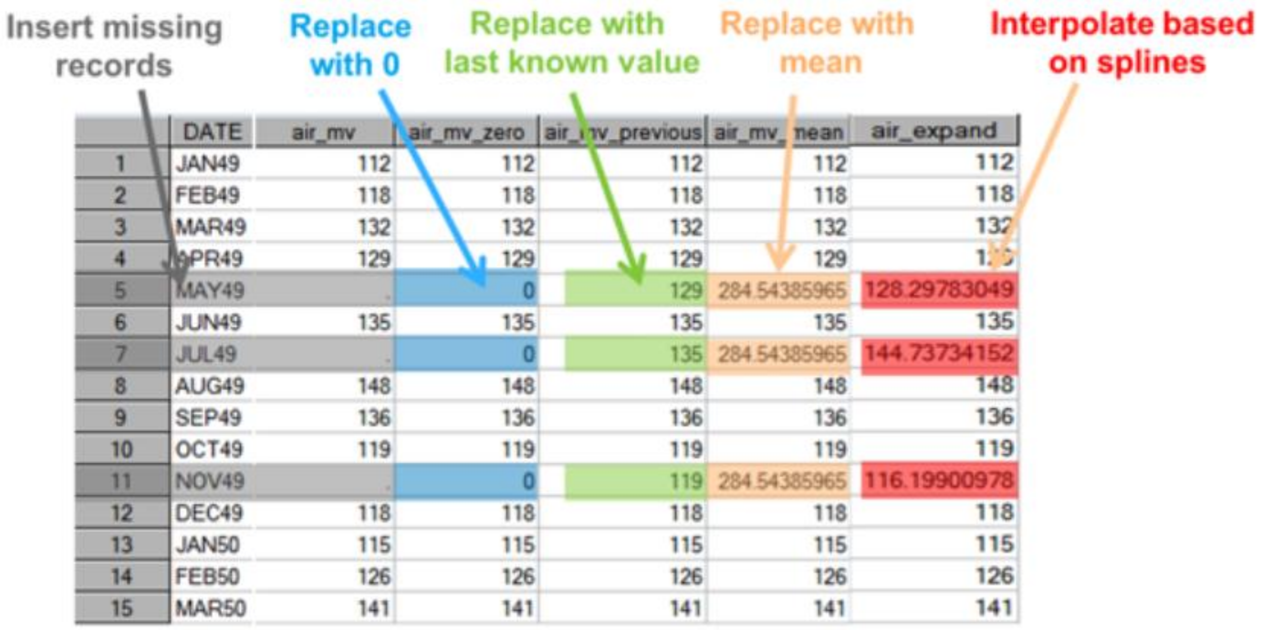
Inaccurate data/Noise/ Outliers: the data item input is not correct. For example, if the right value for age is 36 years, but 360 has been written.

Correcting Incomplete/ Missing data

- How are missing values represented in the datasets?
- Is the space for the missing values left blank or other characters are used to replace it?

This is the reason why first Stage (Business and Data understanding is very necessary) If the variable is a column of age or BMI and is represented by 0, that is an indications of a missing value because no person age or Body Mass Index is 0. Missing values are replaced using; mean, median mode or default values or even using some uncommon techniques depending on context and justification. Below are some of the methods of replacing missing values.

The average or mean is the most common techniques of replacing missing values. But it may not be the right techniques always.


Insert missing records
Replace with 0
Replace with last known value
Replace with mean
Interpolate based on splines

	DATE	air_mv	air_mv_zero	air_mv_previous	air_mv_mean	air_expand
1	JAN49	112	112	112	112	112
2	FEB49	118	118	118	118	118
3	MAR49	132	132	132	132	132
4	APR49	129	129	129	129	129
5	MAY49	.	0	129	284.54385965	128.29783049
6	JUN49	135	135	135	135	135
7	JUL49	.	0	135	284.54385965	144.73734152
8	AUG49	148	148	148	148	148
9	SEP49	136	136	136	136	136
10	OCT49	119	119	119	119	119
11	NOV49	.	0	119	284.54385965	116.19900978
12	DEC49	118	118	118	118	118
13	JAN50	115	115	115	115	115
14	FEB50	126	126	126	126	126
15	MAR50	141	141	141	141	141

Before deciding to replace the missing value by the mean, the Skewness should be checked. If the Skewness is > or < than 1, then the distribution is either Skewed to the right or left. Therefore, it is more accurate to replace the missing value by the median instead of the mean.

7. What is "Regression" in terms of data analytic provide one example of a situation where "Regression" could be used. [10 marks]

Regression Analysis:

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.

It provides the values of the dependent variable from the value of an independent variable. The main use of regression analysis is to determine the strength of predictors, forecast an effect, a trend, etc.

The prediction is **continuous value**, it has a range from minimum to maximum, it cannot be group into categories, hence is called supervised learning Regression.

Regression is used to predict an output (Y)called dependent variables using inputs (X) called independent variables.

Example, we want to predict the Sales of Ice Cream based on the Temperature of the day.

Sales (Y)is the Output which in turn called as dependent variables.

Temperature (X)is input which in turn called as independent variables.

Sales (Y)of Ticket based on the TV advertisement.

Sales is the Output which in turn called as dependent variables.

TV advertisement (X) is input which in turn called as independent variables.

Linear Regression:

The measure of the extent of the relationship between two variables is shown by the **correlation coefficient**. The range of this coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data for two variables.

A linear regression line equation is written in the form of:

$$Y = mx + b$$

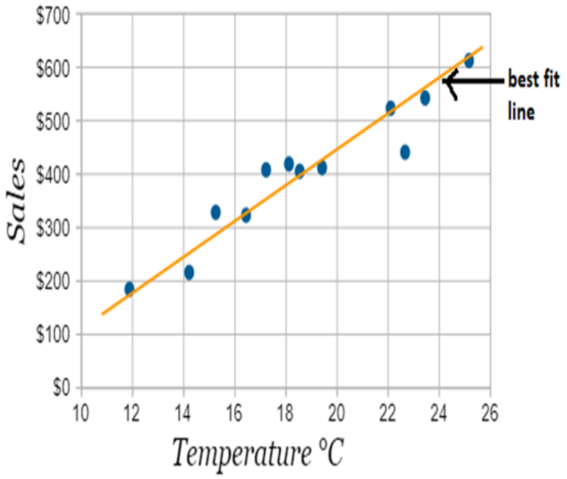
where X is the independent variable and plotted along the x-axis

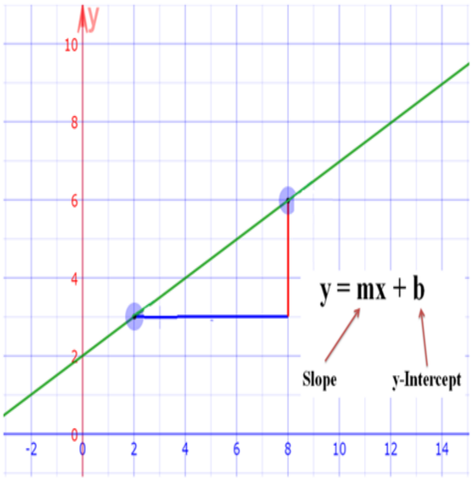
Y is the dependent variable and plotted along the y-axis

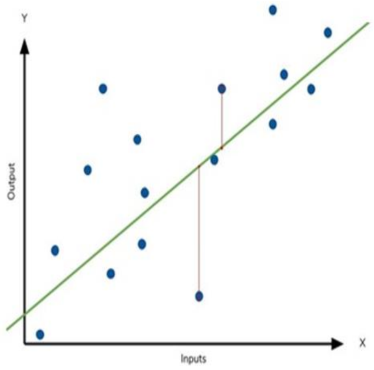
The slope of the line is m, and b is the intercept (the value of y when x = 0).

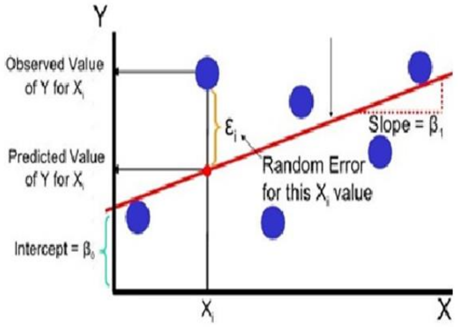
Univariate Linear Regression

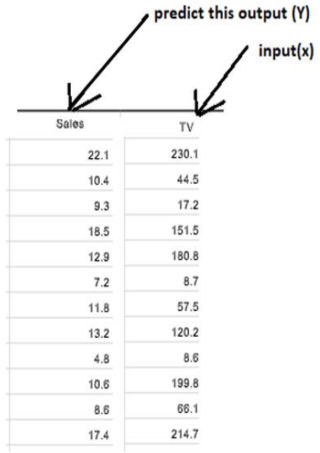
The example is called Univariate Linear Regression because a single variables(Unit) which is TV advertisement(x) is used to predict the number of Sales (Y).








Y
Observed Value of Y for X_i
ε_i
Slope = β_1
Predicted Value of Y for X_i
Random Error for this X_i value
Intercept = β_0
X_i
X


predict this output (Y)
input(x)

Sales	TV
22.1	230.1
10.4	44.5
9.3	17.2
18.5	151.5
12.9	180.8
7.2	8.7
11.8	57.5
13.2	120.2
4.8	8.6
10.6	199.8
8.6	66.1
17.4	214.7

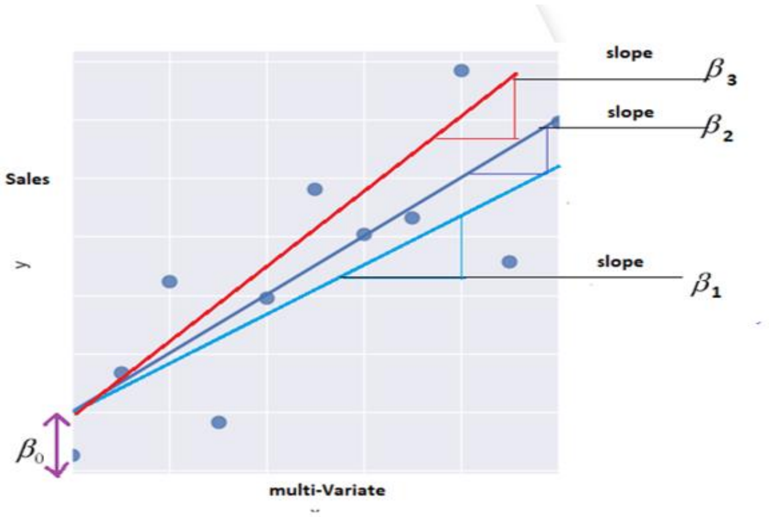
$$y = b + mx + e$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

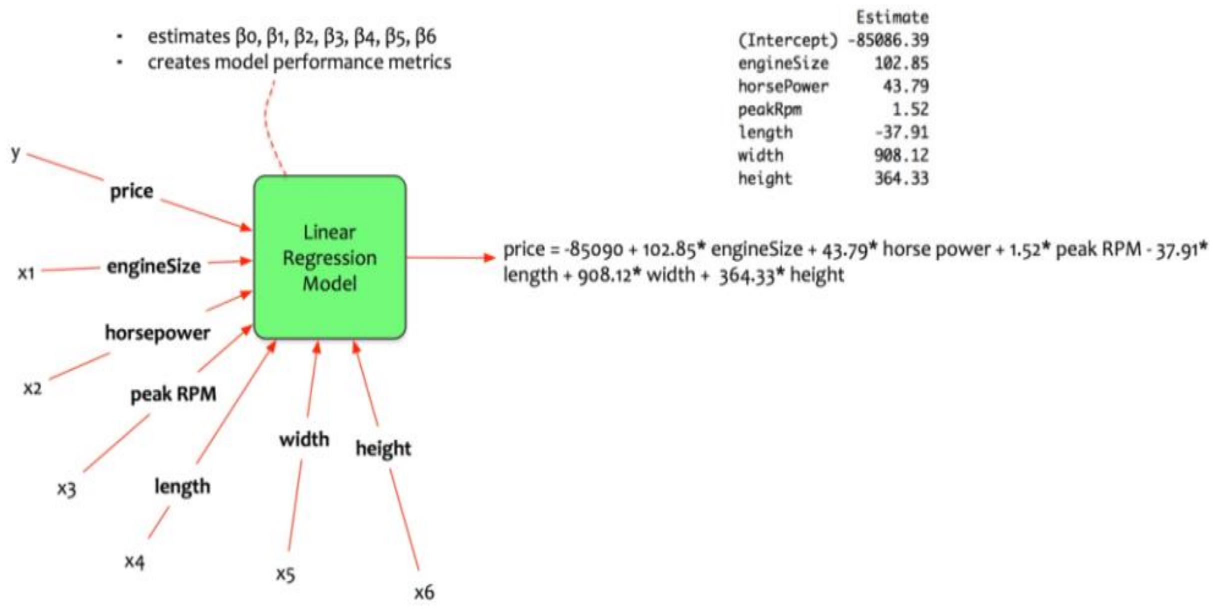
X -> Explanatory independent variables that are used to predict or associate with Y axis
b -> Y intercept
m -> Slope of the explanatory variables
e -> Regression Residual Error term

Multivariate Linear Regression

It is applicable when more than one input (x1,x2, ….x3) are used to predict an Output (Y). For example;
Predict the Sale (Y) based on three inputs;
TV advertisement **(x1),**
Radio advertisement (x2)
Newspaper advertisement (x3)



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_{nk} + \varepsilon_n$$


• estimates β0, β1, β2, β3, β4, β5, β6
• creates model performance metrics
y
price
x1
engineSize
horsepower
x2
peak RPM
x3
length
x4
width
x5
height
x6
Linear Regression Model
price = -85090 + 102.85* engineSize + 43.79* horse power + 1.52* peak RPM - 37.91* length + 908.12* width + 364.33* height

```
               Estimate
(Intercept) -85086.39
engineSize     102.85
horsePower      43.79
peakRpm          1.52
length         -37.91
width          908.12
height         364.33
```

To evaluate the performance of Regression model is totally different for that of classification. In regression, all the evaluation is trying to justify how well the plotted points is closer to the actual points.

8. What is "Classification" in terms of data analytic provide one example of a situation where "Classification" could be used. [10 Marks]

Classification:

The method of arranging data into homogeneous classes according to the common features present in the data is known as classification.

A planned data analysis system makes the fundamental data easy to find and recover. This can be of particular interest for legal discovery, risk management, and compliance. Written methods and sets of guidelines for data classification should determine what levels and measures the company will use to organize data and define the roles of employees within the business regarding input stewardship. Once a data -classification scheme has been designed, the security standards that stipulate proper approaching practices for each division and the storage criteria that determines the data's lifecycle demands should be discussed. This classification belongs to the group of Machine Learning (ML) were what is being predicted is known, for example

- Predicting the outcome of election (win or lose) (**Classification**)
- Predicting if a patent has cancer or not. (**Classification**)
- Predicting if a patient has diabetes or not. (**Classification**)
- Predicting if a customer is going to churn or stay. (**Classification**)

Binary Classification:

It is also **classification** because the prediction put them into groups. When it is two groups is called **Binary Classification.** This is a type of Machine learning (ML) where what is being predicted (output class label or target) are known and a categories or group. For example, consider a cross section of Wisconsin Breast cancer dataset.

It has nine attributes and one output Class.

•The algorithm will teach the computer system to learn to predict, who will have cancer (predict 1) or not (predict 0).

•The emphasis here, is that we know what is being predicted; predict 1 or 0

Multiclass Classification:

More than two groups are called M**ulti-class Classification.** It has a variety of applications. It can be used to identify animals from images and sort them into categories. Cybersecurity companies can use multiclass classification to sort incoming emails as spam or not. It can also be used when analyzing an individual's mood into more than positive or negative. Instead, it will use multiple categories such as, happy, sad, depressed, excited, etc. There is no limit to the amount of classes used in multiclass classification.

Example: An analogy is to consider a study of a population consisting of 1000 patients, and if 900 patients out of 1000 have no disease, a model that predicts all 1000 as not having the disease would still appear to be 90% accurate, even if the remaining 100 patients have the disease, and they were not identified. Therefore, Accuracy has failed or rather is not enough to estimate the performance of classification models due to imbalance nature of data sets.

The research into techniques for handling imbalance dataset or reducing the negative effects of the imbalance is an active area of research.

Confusion Matrix:

If Accuracy has failed, then we can measure the performance of classification modelling through a matrix called as confusion matrix.

A **confusion matrix**, also known as an error matrix is a table used to visualized classification performance.

- True positives (TP): The algorithm predicted positive, and the correct answer is positive; (correctly predicted);
- True negatives (TN): The algorithm predicted negatives, and correct answer is negatives (correctly predicted);
- False positives (FP): The algorithm predicted Positive, but the correct answer is negative (incorrectly predicted); and
- False negatives (FN): The algorithm predicted negatives, but the correct answer is positives (incorrectly predicted).

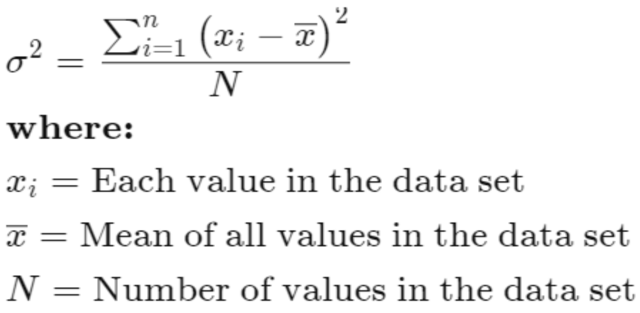
9. Provide a detail explanation of the term variance [10 marks].

Variance:

- Variance is a measurement of the spread between numbers in a data set.
- It measures the degree of dispersion of data around the sample's mean.
- Investors use variance to see how much risk an investment carries and whether it will be profitable.
- Variance is also used in finance to compare the relative performance of each asset in a portfolio to achieve the best asset allocation.
- The square root of the variance is the standard deviation.

In statistics, variance measures variability from the average or mean. It is calculated by taking the differences between each number in the data set and the mean, then squaring the differences to make them positive, and finally dividing the sum of the squares by the number of values in the data set.

Variance is calculated by using the following formula:



$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{N}$$

where:

$x_i =$ Each value in the data set

$\overline{x} =$ Mean of all values in the data set

$N =$ Number of values in the data set

Example:

- Find the mean, standard deviation and variance for the following data: 6, 7,10, 12, 13, 4, 8, 12.

Solution:

Given data: 6, 7,10, 12, 13, 4, 8, 12

Finding Mean:

We know that mean is the ratio of the sum of observations to the total number of observations.

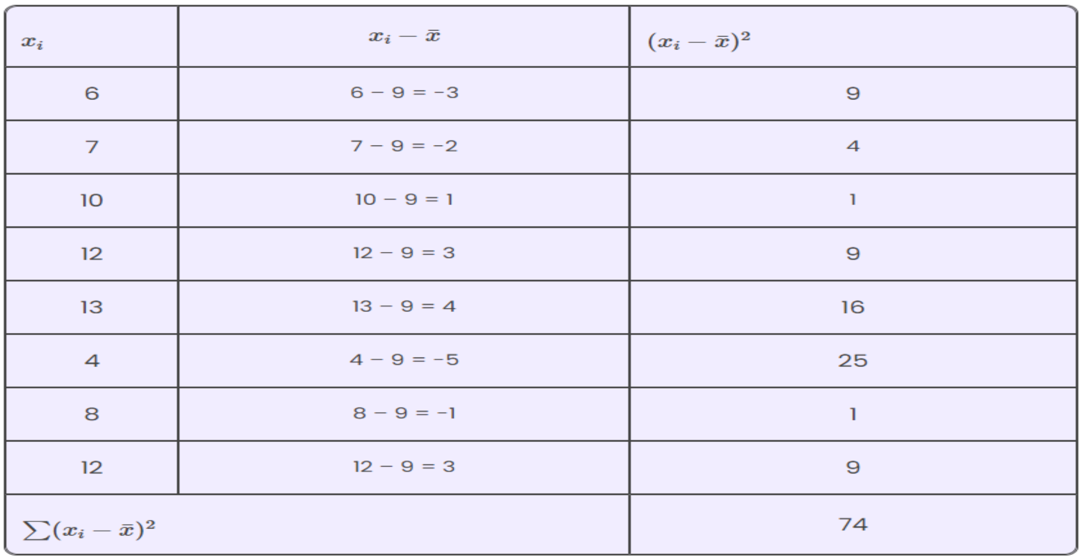
Mean = Sum of observations / Total number of observations.

Mean = (6+7+10+12+13+4+8+12)/8

Mean = 72/8

Mean = 9.

Calculating Variance



x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
6	6 − 9 = -3	9
7	7 − 9 = -2	4
10	10 − 9 = 1	1
12	12 − 9 = 3	9
13	13 − 9 = 4	16
4	4 − 9 = -5	25
8	8 − 9 = -1	1
12	12 − 9 = 3	9
$\sum(x_i - \bar{x})^2$		74

Hence, Variance = 74/8

Variance = 9.25

Finding Standard Deviation:

We know that variance is the square of standard deviation. Hence, the standard deviation can be found by taking the square root of variance.

- Therefore, standard deviation = √variance
- Standard deviation = √(9.25) = 3.041.
- Hence, the mean, variance and standard deviation of the given data are 9, 9.25, 3.041 respectively.

10.Explain positively skew and reflect on the mean, median and mode. [10 marks]. and

Explain negatively skew and reflect on the mean, median and mode. [10 marks].

Skewness:

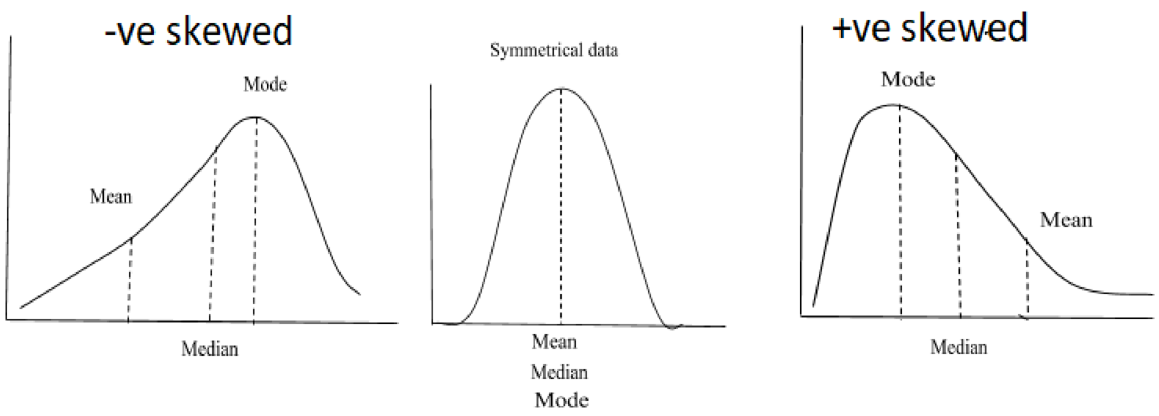
Skewness is the degree of asymmetry observed in a probability distribution. When data points on a bell curve are not distributed symmetrically to the left and right sides of the median, the bell curve is skewed. Distributions can be positive and right-skewed, or negative and left-skewed. A normal distribution exhibits zero skewness.

Types of Skews

Negative, or left-skewed refers to a longer or fatter tail on the left side of the distribution, while positive, or right-skewed, refers to a longer or fatter tail on the right. These two skews show the direction or weight of the distribution.

The three probability distributions below are right-skewed to an increasing degree. The mean of positively skewed data will be greater than the median. In a left-skewed distribution, the mean of negatively skewed data will be less than the median.

A **right-skewed or (+ve) positive distribution** means its tail is more pronounced on the right side than on the left. Since the distribution is positive, the assumption is that its value is positive. As such, most of the values end up left of the mean. This means that the most extreme values are on the right side.



Negative or (-ve) left-skewed means the tail is more pronounced on the left rather than the right. Most values are found on the right side of the mean in negative skewness. As such, the most extreme values are found further to the left.

Implications of Skewness:

This is the extent to which a dataset is bias, in the example above supposing for some reasons the height or blood pressure of male in a city are very tall or high much more than normal because of some genetic abnormally or some defects, then the distributions will not be symmetrical or bell shape. Rather it will skew to either left –vely skewed or to the right or +vely skewed.

The skewness is usually due to the presence of bias in the dataset. If it is a test score , the normal data should show few with very high marks and few with very low marks while majority should be around the mean in approximately 68, 95 and 99 % (1,2 and 3) from the standard deviations.
But when the distribution is skewed either left or right. Then students pass very well or failed too much.
If the people in a city has a high or low pressure than normal, then there must be something wrong.

If the death rate is above normal distribution, then some thing is wrong.
Skewness is one of the test for normality in real life situations.
Skewness are measure from -1 to + 1
Normal distributed data has Skewness = 0
Skewed data is
-1 >x <1