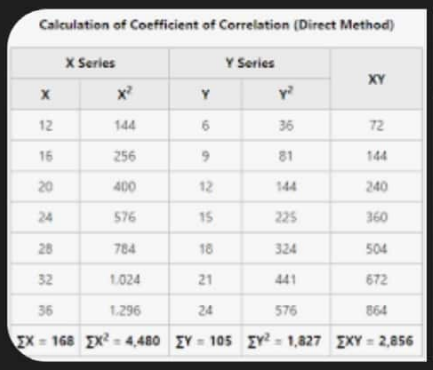
Google

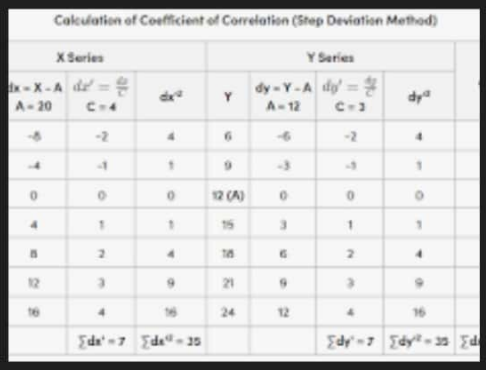
karl pearson coefficient of...

All Images Videos News Shopping Books



Calculation of Coefficient of Correlation (Direct Method)

X Series		Y Series		XY
X	X^2	Y	Y^2	
12	144	6	36	72
16	256	9	81	144
20	400	12	144	240
24	576	15	225	360
28	784	18	324	504
32	1,024	21	441	672
36	1,296	24	576	864
$\sum X = 168$	$\sum X^2 = 4,480$	$\sum Y = 105$	$\sum Y^2 = 1,827$	$\sum XY = 2,856$



Calculation of Coefficient of Correlation (Step Deviation Method)

X Series			Y Series			
$dx = X - A$, $A = 20$	$dx' = \frac{dx}{C}$, $C = 4$	dx'^2	Y	$dy = Y - A$, $A = 12$	$dy' = \frac{dy}{C}$, $C = 3$	dy'^2
-8	-2	4	6	-6	-2	4
-4	-1	1	9	-3	-1	1
0	0	0	12 (A)	0	0	0
4	1	1	15	3	1	1
8	2	4	18	6	2	4
12	3	9	21	9	3	9
16	4	16	24	12	4	16
	$\sum dx' = 7$	$\sum dx'^2 = 35$			$\sum dy' = 7$	$\sum dy'^2 = 35$

Karl Pearson's coefficient of correlation is an extensively used mathematical method in which the numerical representation is applied to measure the level of relation between linearly related variables. The coefficient of correlation is expressed by “r”. 2 Feb 2021

https://byjus.com › commerce › kar...

Karl Pearson Coefficient Of Correlation | Examples, Methods, Formula - BYJU'S

About featured snippets Feedback

People also ask

What is the Pearson correlation coefficient?

Discover Search Saved

Pearson's Coefficient Correlation

- Karl Pearson's coefficient of correlation is an extensively used mathematical method in which the numerical representation is applied to measure the level of relation between linearly related variables. The coefficient of correlation is expressed by "r".

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

Pearson correlation example

1. **When a correlation coefficient is (1),**

- that means for every increase in one variable, there is a positive increase in the other fixed proportion.
- For example, shoe sizes change according to the length of the feet and are perfect (almost) correlations.

When a correlation coefficient is (-1),

- that means for every positive increase in one variable, there is a negative decrease in the other fixed proportion.
- For example, the decrease in the temperature causes ice cream sale less/decreases
- When a correlation coefficient is (0) for every increase, that means there is no positive or negative increase, and the two variables are not related.

- **4. Possibility of Wrong Interpretation:** While interpreting the value of coefficient of correlation using this method, one has to be very careful. It is because the chances of misinterpreting the coefficient are more.

Spearman's rank correlation

- A statical measure of measuring the strength and direction of the relationship between two continuous variables.
- attributes are ranked or put in the order of their preference.
- It is denoted by the symbol "rho" (ρ)
- Can take values between -1 to +1.
- **Spearman's Correlation formula**

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

r_s = *Spearman Correlation coefficient*

d_i = *the difference in the ranks given to the two variables values fo each item of the data*

n = *total number of observation*

Advantages of Spearman's Rank Correlation:

- Easy to understand.
- It is superior for calculating qualitative observations such as the intelligence of people, physical appearance, etc.
- This method is suitable when the series gives only the order of preference and not the actual value of the variable.
- It is robust to the outliers present in the data

Disadvantages of Spearman's Rank Correlation:

- It is not applicable in the case of grouped data.
- It can handle only a limited number of observations or items.
- It only considers the ranks of the data points and ignores the actual magnitude of differences between the values of the variables.
- Converting the data into ranks for Spearman's rank correlation discards the original values of the variables and replaces them with their respective ranks. This transformation may result in a loss of information in the data, especially if the variables of the data have meaningful magnitudes or units.

ase 1: When Ranks are given

Find the correlation using Spearman's Rank Correlation

n an art competition, two judges accorded following ranks to the 10 participants.

judge X	1	2	3	4	5	6	7	8	9	10
Judge Y	6	2	9	7	1	4	8	3	10	5

judge X (R_1)	Judge Y (R_2)	$D = R_1 - R_2$	D^2
1	6	-5	25
2	2	0	0

$$r_k = 1 - \frac{6\sum D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 142}{10^3 - 10}$$

Coefficient of Correlation (rk) = 0.14

Case 1: When Ranks are given

Find the correlation using Spearman's Rank Correlation

In an art competition, two judges accorded following ranks to the 10 participants.

judge X	1	2	3	4	5	6	7	8	9	10
Judge Y	6	2	9	7	1	4	8	3	10	5

judge X (R1)	Judge Y (R2)	D = R1 – R2	D2
1	6	-5	25
2	2	0	0
3	9	-6	36
4	7	-3	9
5	1	4	16
6	4	2	4
7	8	-1	1
8	3	5	25
9	10	-1	1
10	5	5	25
N = 10			∑D2 =

$$r_k = 1 - \frac{6\sum D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 142}{10^3 - 10}$$

$$= 1 - \frac{852}{990}$$

$$= 1 - 0.860$$

$$= 0.14$$

Coefficient of Correlation (rk) = 0.14

As the rank correlation is positive and closer to 0, it means that the association between the ranks of the two judges is weaker.

Students	Maths	Science
A	35	24
B	20	35
C	49	39
D	44	48
E	30	45

Case 2: When Ranks are not given

Find the correlation using Spearman's Rank Correlation

Students	Maths	Rank	Science	Rank	d	d square
A	35	3	24	5	2	4
B	20	5	35	4	1	1
C	49	1	39	3	2	4
D	44	2	48	1	1	1
E	30	4	45	2	2	4
						14

$$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2 - 1)}$$

= 1 - (6 * 14) / 5(25 - 1)

= 0.3

The value is near 0, which means that there is a weak correlation between the two ranks.

Find the correlation using Spearman's Rank Correlation:

x	y
60	75
34	32
40	35
50	40
45	45
41	33
22	12
43	30
42	36
66	72
64	41
46	57

sr no	x	Rx	y	Ry	d	d sqare
1	60	3	75	1	2	4
2	34	11	32	10	1	1
3	40	10	35	8	2	4
4	50	4	40	6	-2	4
5	45	6	45	4	2	4
6	41	9	33	9	0	0
7	22	12	12	12	0	0
8	43	7	30	11	-4	16
9	42	8	36	7	1	1
10	66	1	72	2	-1	1
11	64	2	41	5	-3	9
12	46	5	57	3	2	4
				∑d	0	∑d2=48

n=12

rank correlation coeff 0.832168

$$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2 - 1)}$$

76	2	67	3	1	1
61	7	63	6	1	1

positive relationship between the ranks

Case 3-tie up/common rank

	x	Rx	y	Ry	d=Rx-Ry	d^2
1	48	3	13	5.5	-2.5	6.25
2	33	5	13	5.5	-0.5	0.25
3	40	4	24	1	3	9
4	9	10	6	8.5	1.5	2.25
5	16	8	15	4	4	16
6	16	8	4	10	-2	4
7	65	1	20	2	-1	1
8	24	6	9	7	-1	1
9	16	8	6	8.5	-0.5	0.25
10	57	2	19	3	-1	1
				∑d=	0	

Rank tie

in x

7,8,9=(7+8+9)/3= 8

in y

5,6=(5+6)/2= 5.5

in y

8,9=(8+9)/4= 8.5

$$r_s = 1 - \frac{6\left[\Sigma d^2 + \frac{1}{12}(3^3-3) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2)\right]}{n^3 - n}$$

$$= 1 - \frac{6\,[41 + 2 + 0.5 + 0.5]}{10^3 - 10} = 1 - \frac{6\,(44)}{990}$$

$$= 1 - \frac{264}{990} = 1 - 0.2667 = +\,0.7333.$$

Example 6.33. *From the marks obtained by 8 students in Accountancy and Statistics, compute coefficient of correlation by rank difference method.* (BU-N89

Marks in Accountancy :	60	15	20	28	12	40	80	20
Statistics :	10	40	30	50	30	20	60	30

Solution : The values are given. So ranks must be assigned.

Acct. x	Stat. y	Rx	Ry	RX-RY d	d2
60	10	2	8	–6	36
15	40	7	3	+4	16
20	30	5.5	5	+0.5	0.25
28	50	4	2	+2	4
12	30	8	5	+3	9
40	20	3	7	–4	16
80	60	1	1	0	0
20	30	5.5	5	+0.5	0.25
				$\Sigma d = 0$	$\Sigma d^2 = 81.50$

Tied Ranks

(*i*) In X series 5th & 6th.

(*ii*) In Y series 4th, 5th & 6th.

$$r_s = 1 - \frac{6\left[\Sigma d^2 + \frac{1}{12}(2^3-2) + \frac{1}{12}(3^3-3)\right]}{n^3 - n}$$

$$= 1 - \frac{6\,[81.50 + 0.5 + 2]}{8^3 - 8} = 1 - \frac{6\,(84)}{504}$$

$$= 1 - \frac{504}{504} = 1 - 1 = 0.$$

Exploratory Analysis

- The preliminary analysis of data to discover relationships between measures in the data and to gain an insight on the trends, patterns, and relationships among various entities present in the data set with the help of statistics and visualization tools is called Exploratory Data Analysis (EDA).
- Exploratory data analysis is cross-classified in two different ways where each method is either graphical or non-graphical.

What is Exploratory Data Analysis ?

In this article, we will discuss exploratory data analysis which is one of the basic and essential steps of a data science project. A data scientist involves almost 70% of his work in doing the EDA of his dataset.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables. EDA is normally carried out as a preliminary step before undertaking extra formal statistical analyses or modeling.

The Foremost Goals of EDA

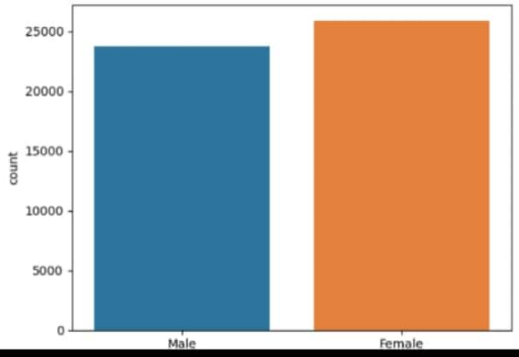
1. Data Cleaning: EDA involves examining the information for errors, lacking values, and inconsistencies. It includes techniques including records imputation, managing missing statistics, and figuring out and getting rid of outliers.

2. Descriptive Statistics: EDA utilizes precise records to recognis[illegible]nt tendency,

and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.

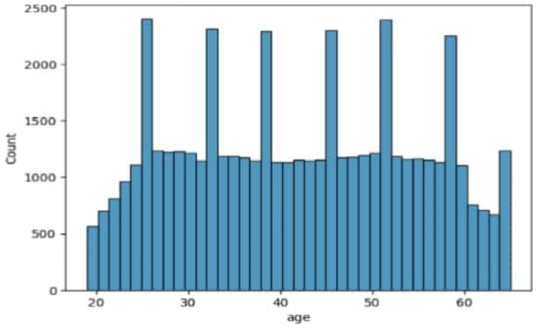
Bar Charts

- The bar graph is very convenient while comparing categories of data or different groups of data. It helps to track changes over time. It is best for visualizing discrete data.



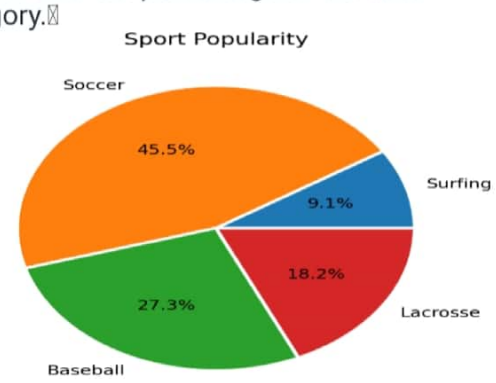
Ø Histograms

- Histograms are similar to bar charts and display the same categorical variables against the category of data. Histograms display these categories as bins which indicate the number of data points in a range. It is best for visualizing continuous data.



ØPie Chart

- A piechart helps us to visualize the percentage of the data belonging to each category.



Univariate data can be described through:

Mean, Median, Mode, Variance, Standard Deviation

Univariate data can be described through:

Mean, Median, Mode, Variance, Standard Deviation

Bivariate analysis

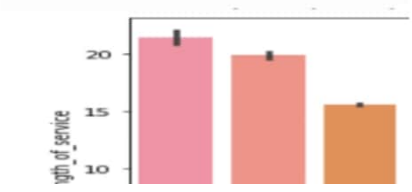
- Bivariate analysis is the simultaneous analysis of two variables. It explores the concept of the relationship between two variable whether there exists an association and the strength of this association or whether there are differences between two variables and the significance of these differences.
- The bivariate analysis involves the analysis of exactly two variables.

- **Bivariate analysis**
- The main three types we will see here are:

1. *Categorical v/s Numerical*
2. *Numerical V/s Numerical*
3. *Categorical V/s Categorical data*

Categorical v/s Numerical

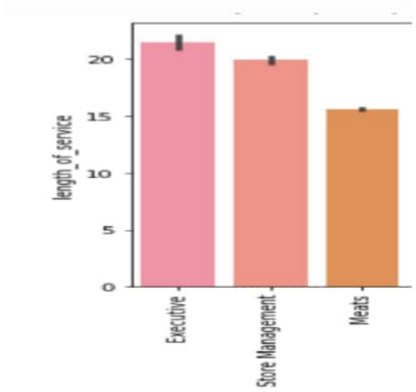
- **Department_name' vs'length_of_service'**



Black horizontal line is indicating huge differences in the length of

Categorical v/s Numerical

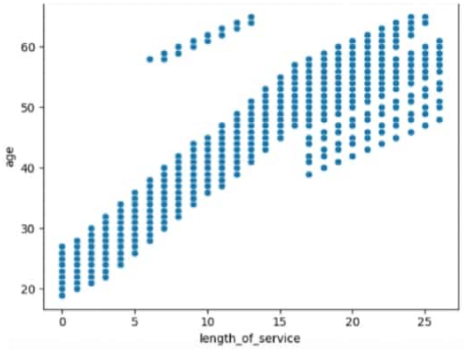
- Department_name' vs'length_of_service'



Black horizontal line is indicating huge differences in the length of service among different departments.

Numerical v/s Numerical

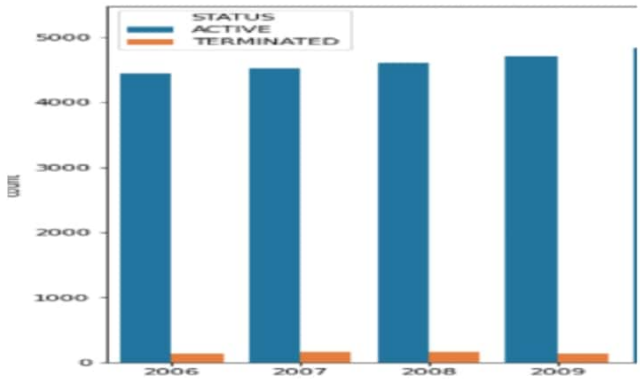
- 'length_of_service vs 'age'



younger employees have less experience in terms of their length of service.

Categorical v/s Categorical

- STATUS_YEAR vs STATUS



Importance of bivariate analysis

- Bivariate analysis helps identify trends and patterns: Bivariate analysis helps identify cause and effect relationships
- It helps researchers make predictions
- It helps inform decision-making:

- The **chi-square test** is a statistical method for identifying disparities in one or more categories between what was expected and what was observed. The test's primary premise is to assess the actual data values to see what would be expected if the null hypothesis was valid.
- Example of bivariate analysis
- Investigating the connection between education and income
- Investigating the connection between aging and blood pressure

Univariate Data Analysis	Bivariate Data Analysis
• involving a **single variable**	• involving **two variables**
• does not deal with causes or relationships	• deals with causes or relationships
• the major purpose of univariate analysis is to describe	• the major purpose of bivariate analysis is to explain
• central tendency - mean, mode, median • dispersion - range, variance, max, min, quartiles, standard deviation.	• analysis of two variables simultaneously • correlations • comparisons, relationships, causes, explanations
Sample question: How many of the students in the freshman class are female?	**Sample question:** Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

Assignment 2

Q.1 Compute Karl Pearson's coefficient of correlation from the following data

Price (`)	10	20	30	40	50	60	70
Supply (Units)	8	6	14	16	10	20	24

Q2 Find the correlation using Spearman's Rank Correlation:

x	70	80	65	78	68	65	82	65
y	13	15	12	14	13	11	16	10

Q3-Write the difference between regression and correlation

Unit IV

Hypothesis Testing

Regression

- Regression analysis is used to predicts the value of the dependent variable based on the known value of the independent variable, assuming that average mathematical relationship between two or more variables. helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.
- It predicts continuous/real values such as **temperature, age, salary, price,** etc.
- predicting the salary of an employee on the basis of **the year of experience**.

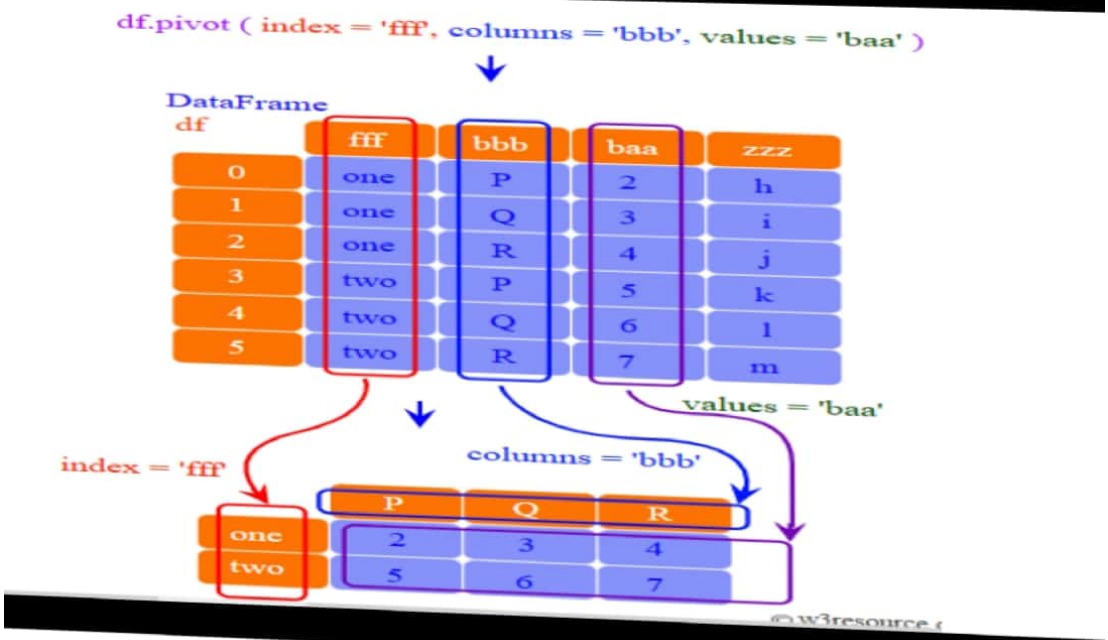
Correlation	Regression
Correlation is used to determine whether variables are related or not.	Regression is used to numerically describe how a dependent variable changes with a change in an independent variable
Correlation tries to establish a linear relationship between variables.	use predict value of another variable
Correlation uses a signed numerical value to estimate the strength of the relationship between the variables.	Regression is used to show the impact of a unit change in the independent variable on the dependent variable.
Correlation tries to establish a linear relationship between variables.	estimate an unknown variable on the basis of the known variable.
Positive relation, Negative relation, no relation	Calculation house prize

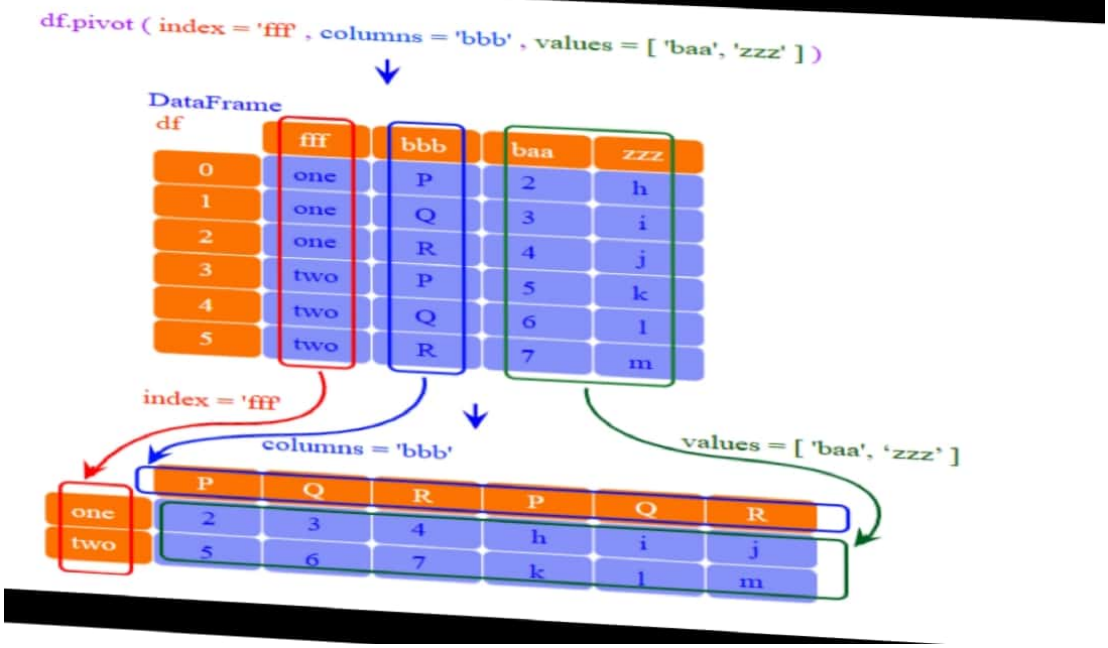
Combining and Merging Data Sets

	Semester 2	Bobby	63
Science	Semester 2	Alisa	67
	Semester 2	Bobby	89

Pivoting

- To generate easy insights into your data
- pivot table is a table of statistics that helps summarize the data of a larger table by “pivoting” that data.
- Microsoft Excel popularized the pivot table, where they’re known as PivotTables.
- Pandas gives access to creating pivot tables in Python using th e .pivot_table() function.





inaccuracy in estimating some value that is caused by only a portion of a population (i.e. *sample*) rather than the whole population

Random sampling

- It is considered one of the most popular and simple data collection methods in research fields (probability and statistics, mathematics, etc.).
- It allows for unbiased data collection
- It is also called **probability sampling**.
- The choice of observations must occur in a 'random' way

Probability Sampling Methods

- This method of sampling involves the random selection of any entity.
- each entity of such a population has an equal chance of getting selected to be part of the sample.

Type of probability Sampling

1. Simple Random Sampling
2. Systematic Sampling
3. Stratified Sampling
4. Clustered Sampling

Simple random sampling

- In this sampling method, each item in the population has an equal and likely possibility of getting selected in the sample (for example, each member in a group is marked with a specific number).
- Since the selection of item completely depends on the possibility, therefore this method is called "**Method of chance Selection**".

Simple random sampling

- In this sampling method, each item in the population has an equal and likely possibility of getting selected in the sample (for example, each member in a group is marked with a specific number).
- Since the selection of item completely depends on the possibility, therefore this method is called **"Method of chance Selection"**.
- Also, the sample size is large, and the item is selected randomly. Thus it is known as **"Representative Sampling"**.

242

Systematic Random Sampling

- Systematic sampling is the selection of specific individuals or members from an entire population.
- The selection often follows a predetermined interval (k).
- Performing quality control checks on a pen factory by selecting every 50th pen from a production line

243

Stratified Random Sampling

- n this sampling method, a population is divided into subgroups to obtain a simple random sample from each group and complete the sampling process
- (for example, number of girls in a class of 50 strength).
- These small groups are called **strata**.
- The small group is created based on a few features in the population.
- After dividing the population into smaller groups, the researcher randomly selects the sample.

244

Clustered Sampling

- Cluster sampling is similar to stratified sampling, besides the population is divided into a large number of subgroups
- After that, some of these subgroups are chosen at random and simple random samples are then gathered within these subgroups. These subgroups are known as **clusters**
- Assessing immunization coverage in a country by selecting a sample of districts and then a sample of households within each district.

244

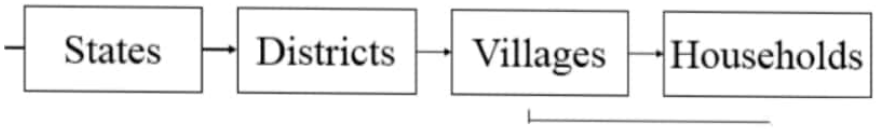
Clustered Sampling

- Cluster sampling is similar to stratified sampling, besides the population is divided into a large number of subgroups
- After that, some of these subgroups are chosen at random and simple random samples are then gathered within these subgroups. These subgroups are known as **clusters**
- Assessing immunization coverage in a country by selecting a sample of districts and then a sample of households within each district.

245

Multistage Sampling Schemes

- Multistage sampling divides large populations into stages to make the sampling process more practical
- Multistage sampling is flexible, cost effective and easy to implement. You can use as many stages as you need to reduce the sample to a workable size, with no restrictions on how you divide the groups.



246

Unit III
Regression Analysis

247

endswith()

- txt = "Hello, welcome to my world."
  x = txt.endswith(".")

- Check if position 5 to 11 ends with the phrase "my world.":
- txt = "Hello, welcome to my world."
- x = txt.endswith("my world.", 5, 11)

find()

- The find() method finds the first occurrence of the specified value.
- The find() method returns -1 if the value is not found.
- txt = "Hello, welcome to my world."
- x = txt.find("w") 7

isalnum()

- txt = "Company12"
  txt.isalnum()

isdigit()

- txt = "50800"
  txt.isdigit()

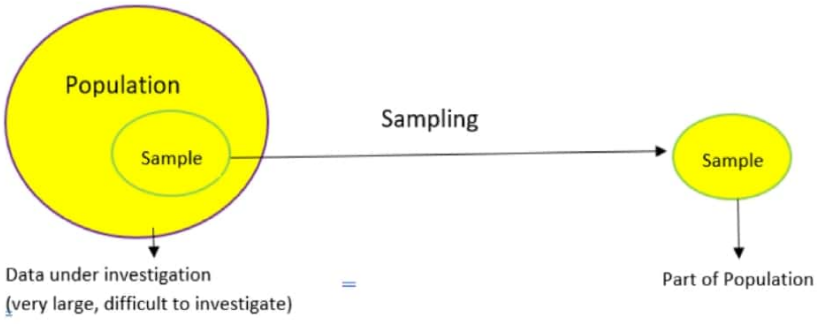
replace()

- Returns a string where a specified value is replaced with a specified value
- txt = "I like bananas"
- txt.replace("bananas", "apples")

strip()

- txt = " banana "
  x = txt.strip()

Sampling



- **Sampling** is a process in statistical analysis where researchers take a

Data Wrangling

- Data Wrangling is referred to as ***data munging***.
- It is the process of transforming and mapping data from on e "raw" data form into another format to make it more appropriate and valuable for various purposes such as analytics.
- Assure quality and useful data
- **Data wrangling typically follows a set of general steps,**
- extracting the raw data from the data source,
- "munging" the raw data (e.g., sorting) or parsing the data into predefined data structures,
- Finally depositing the resulting content into a data sink for storage and future use.

Data Wrangling Examples

- Merging several data sources into one data set for analysis
- Identifying gaps or empty cells in data and either f lling or removing them
- Deleting irrelevant or unnecessary data
- Identifying severe outliers in data and either explaining the inconsistencies or deleting them to facilitate analysis

Businesses also use data wrangling tools to

- Detect corporate fraud
- Support data security
- Ensure accurate and recurring data modeling results
- Ensure business compliance with industry standards
- Perform Customer Behavior Analysis
- Reduce time spent on preparing data for analysis
- Promptly recognize the business value of your data
- Find out data trends

Data Wrangling Process

1. **Discovery**-what your data is all about. In this first step, you get familiar with your data
2. **Organization**-ata needs to be restructured to suit the analytical model that your enterprise plans to deploy
3. **Cleaning**-involves the tackling of outliers, making corrections, or deleting bad data completely
4. **Data enrichment**:determine if you have enough data to proceed.

Data Wrangling Process

1. **Discovery-**what your data is all about. In this first step, you get familiar with your data
2. **Organization-**ata needs to be restructured to suit the analytical model that your enterprise plans to deploy
3. **Cleaning-**involves the tackling of outliers, making corrections, or deleting bad data completely
4. **Data enrichment:**determine if you have enough data to proceed.
5. **Validation:**apply validation rules to your data. apply validation rules to your data.
6. **Publishing:** Data publishing involves preparing the data for future use. This may include providing notes and documentation

Benefits of Data Wrangling

- **Data consistency:** The organizational aspect of data wrangling offers a resulting dataset that is more consistent.
- **Improved insights: Data** wrangling can provide statistical insights
- **Cost efficiency:**

Correlation Between Categorical Variables

- **Tetrachoric Correlation:** Used to calculate the correlation between binary categorical variables.
- **Polychoric Correlation:** Used to calculate the correlation between ordinal categorical variables.
- **Cramer's V:**Used to calculate the correlation between nominal categorical variables.

Tetrachoric Correlation

- To calculate the correlation between binary categorical variables
- The value for tetrachoric correlation ranges from -1 to 1
- where -1 indicates a strong negative correlation,
- 0 indicates no correlation
- 1 indicates a strong positive correlation.

- There is a weak association between gender and political party preference.

- Data transformation is the process of converting, cleansing, and structuring data into a usable format that can be analyzed to support decision making processes, and to propel the growth of an organization. Data transformation is used when data needs to be converted to match that of the destination system.

Advantages of Data Transformation

1. Improves Data Quality: Data transformation helps to improve the quality of data by removing errors, inconsistencies, and missing values.
2. Facilitates Data Integration: Data transformation enables the integration of data from multiple sources, which can improve the accuracy and completeness of the data.
3. Improves Data Analysis: Data transformation helps to prepare the data for analysis and modeling by normalizing, reducing dimensionality, and discretizing the data.
4. Increases Data Security: Data transformation can be used to mask sensitive data, or to remove sensitive information from the data, which can help to increase data security.

Disadvantages of Data Transformation

Time-consuming: Data transformation can be a time-consuming process, especially when dealing with large datasets.

Complexity: Data transformation can be a complex process, requiring specialized skills and knowledge to implement and interpret the results.

Data Loss: Data transformation can result in data loss, such as when discretizing continuous data, or when removing attributes or features from the data.

Biased transformation: Data transformation can result in bias, if the data is not properly understood or used.

High cost: Data transformation can be an expensive process, requiring significant investments in hardware, software, and personnel.

1. Smoothing

- used to remove noise from the dataset
- Data smoothing refers to a statistical approach of eliminating

mask sensitive data, or to remove sensitive information from the data, which can help to increase data security.

Disadvantages of Data Transformation

Time-consuming: Data transformation can be a time-consuming process, especially when dealing with large datasets.

Complexity: Data transformation can be a complex process, requiring specialized skills and knowledge to implement and interpret the results.

Data Loss: Data transformation can result in data loss, such as when discretizing continuous data, or when removing attributes or features from the data.

Biased transformation: Data transformation can result in bias, if the data is not properly understood or used.

High cost: Data transformation can be an expensive process, requiring significant investments in hardware, software, and personnel.

1. Smoothing

- used to remove noise from the dataset
- Data smoothing refers to a statistical approach of eliminating outliers from datasets to make the patterns more noticeable.
- It helps in predicting the patterns.
- data smoothing may eliminate the usable data points. It may lead to incorrect forecasts
- The noise is removed from the data using the techniques such as binning, regression, clustering.
- Binning method
- Regression
- Clustering

- **Regression:** This method identifies the relation among two dependent attributes so that if we have one attribute, it can be used to predict the other attribute.
- **Clustering:** This method groups similar data values and form a cluster. The values that lie outside a cluster are known as outliers

Binning method for data smoothing

- In this method, the data is first sorted and
- then the sorted values are distributed into a number of buckets

The data transformation process

- To better understand how data transformation works, data transformation process by breaking it down into four steps:

1. **Discovering**
2. **Planning**
3. **Performing**
4. **Reviewing**

1 Discovering variables in the source data

Through data discovery, you need to identify variables of interest within the source data and figure out what pre-processing actions need to be performed to facilitate the data transformation.

2 Planning the data transformation

To map the source data to its landing system, you need to determine the structure it needs to be in. In our example, we'll convert our JSON data to a tabular format of rows and columns. In addition to structure, in this step decide whether fields need to be renamed, dropped, or aggregated.

-

3. Performing the data transformation

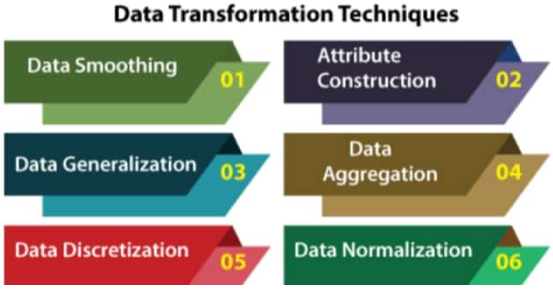
- Several tools or programming languages can be used to perform the data transformation.
- Python and SQL are popular programming languages for data transformation

4. Reviewing the data transformation

- Once the data transformation has occurred, evaluate it to make sure the results are as expected. For example, use tools to count records or verify duplicates have been removed, or that data aggregations have been performed correctly.

Data Transformation Techniques

- There are several data transformation techniques that can help structure and clean up the data before analysis or storage



Types of data analysis

- There are four main types of data analysis: **Descriptive**, **diagnostic**, **predictive**, and **prescriptive**.

Descriptive (What happened?)

- The purpose of descriptive analytics is to simply describe what has happened; it doesn't try to explain why this might have happened
- looks at what has happened in the past
- The outcome of descriptive analysis is a visual representation of the data—as a bar graph, for example, or a pie chart.
- There are two main techniques used in descriptive analytics:
  - ✓ Data aggregation
  - ✓ Data mining

Data aggregation

- The process of gathering data and presenting it in a summarized format.
- An ecommerce company collects all kinds of data relating to their customers and people who visit their website. The aggregate data, or summarized data, would provide an overview of this wider dataset—such as the average customer age, for example, or the average number of purchases made.

Data mining

- Data mining is a distinct process that turns raw data points into informative ones.
- Data mining involves finding different patterns, correlations, or anomalies within big data sets to predict outcomes or better understand the source of said data points.
- The analyst explores the data in order to uncover any patterns or trends.

Diagnostic (Why did it happen?)

- Diagnostic analytics goes deeper in order to understand why something happened.
- The main purpose is to identify and respond to **anomalies within your data**.
- For example: If your descriptive analysis shows that there was a 20% drop in sales for the month of March, you'll want to find out why. The next logical step is to perform a diagnostic analysis.
- diagnostic analytics, there are a number of different techniques such as probability theory, regression analysis, filtering, and time -series analysis

Predictive (What is likely to happen in the future ?)

- To predict what is likely to happen in the future. Based on past patterns and trends, data analysts can devise predictive models which estimate the likelihood of a future event or outcome.
- This is especially useful as it enables businesses to plan ahead.
- Use the relationship between a set of variables to make predictions; for example, you might use the correlation between seasonality and sales figures to predict when sales are likely to drop
- To predict how many takeaway orders you're likely to get on a typical Saturday night. Based on what your predictive model tells you, you might decide to get an extra delivery driver on hand.

- Predictive modeling
- Decision Analysis and optimization
- Transaction profiling

Descriptive Analytics

Descriptive analytics looks at data and analyze past event for insight as to how to approach future events. It looks at past performance and understands the performance by mining historical data to understand the cause of success or failure in the past. Almost all management reporting such as sales, marketing, operations, and finance uses this type of analysis.

The descriptive model quantifies relationships in data in a way that is often used to classify customers or prospects into groups. Unlike a predictive model that focuses on predicting the behavior of a single customer, Descriptive analytics identifies many different relationships between customer and product.

Common examples of Descriptive analytics are company reports that provide historic reviews like:

- Data Queries
- Reports
- Descriptive Statistics
- Data dashboard

For example, Prescriptive Analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demography, etc.

Diagnostic Analytics

In this analysis, we generally use historical data over other data to answer any question or for the solution of any problem. We try to find any dependency and pattern in the historical data of the particular problem.

For example, companies go for this analysis because it gives a great insight into a problem, and they also keep detailed information about their disposal otherwise data collection may turn out individual for every problem and it will be very time-consuming. Common techniques used for Diagnostic Analytics are:

- Data discovery
- Data mining
- Correlations

Usage of Data Analytics

There are some key domains and strategic planning techniques in which Data Analytics has played a vital role:

- Improved Decision-Making: If we have

Predictive Analytics

Predictive analytics turn the data into valuable, actionable information. predictive analytics uses data to determine the probable outcome of an event or a likelihood of a situation occurring. Predictive analytics holds a variety of statistical techniques from modeling, machine learning, data mining, and game theory that analyze current and historical facts to make predictions about a future event. Techniques that are used for predictive analytics are:

- Linear Regression
- Time Series Analysis and Forecasting
- Data Mining

Basic Corner Stones of Predictive Analytics

- Predictive modeling
- Decision Analysis and optimization
- Transaction profiling

Descriptive Analytics

Descriptive analytics looks at data and analyze past event for insight as to how to approach future events. It looks at past performance and understands the performance by mining historical data to understand the cause of success or failure

- Data Queries
- Reports
- Descriptive Statistics
- Data dashboard

Prescriptive Analytics

Prescriptive Analytics automatically synthesize big data, mathematical science, business rule, and machine learning to make a prediction and then suggests a decision option to take advantage of the prediction.

Prescriptive analytics goes beyond predicting future outcomes by also suggesting action benefits from the predictions and showing the decision maker the implication of each decision option. Prescriptive Analytics not only anticipates what will happen and when to happen but also why it will happen. Further, Prescriptive Analytics can suggest decision options on how to take advantage of a future opportunity or mitigate a future risk and illustrate the implication of each decision option.

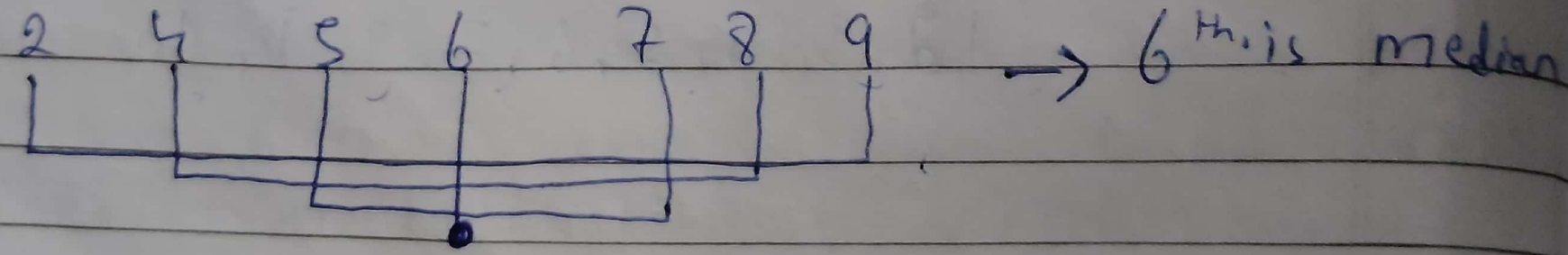
For example, Prescriptive Analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demography, etc.

cannot be determined
nor can be learned graphically
If any one of the data.

* median:

The median is the middle value of a set of data. To determine the median value in a sequence of number the numbers must first be arranged ascending order.

odd:



When the average income for a country is discussed median is most often used because it represents the middle of a group

eg:

2, 3, 4, 5, 6.6

$\frac{4+5}{2} = \frac{9}{2} = 4.5$

find median

2, 8, 6, 1, 3, 7

1, 2, 3, 6, 7, 8

$\frac{3+6}{2} = \frac{9}{2} = 4.5$

Merits of median:

- Certainty
- Simplicity
- unaffected by scheme values
- possible even if data incomplete
- graphic presentation
- Appropriate for Qualitative data.

Demerits of median:

- Arrangement of data is never necessary since the median is an average position
- median fails to be representative measure
- median is a limited representative character as it is not found based on all the terms in the series

* Mode:
- mode is a statistical concept used to describe the most frequently occuring value in a given set of data
- Simply identifies the value that appears most frequently
- To find out which clothes sizes are most popular with clients. For instance, a clothing store may utilize the mode
- 2, 4, 5, 5, 6, 7, the mode of data set is 5 since it has appeared in the set twice

Ex-4 find the mean median mode:

1) 10
10
20
30
40

Sum → 110
n → 5

∴ mean = $\frac{\text{Sum}}{n}$

= $\frac{110}{5}$ = 22

10
10 → 2
20 → 1
30 → 1
40 → 1

⇒ mode = 10
median = 20

by observation or inspection.
Everyone understands the concept of majority. Since, mode is based on this concept.

Demerits of modes:
It is not based on all the items of series.
mode is an uncertain
Affected by fluctuation of sampling:
The mode is most affected by fluctuation of sampling.

Calculate standard deviation & variance of given datasets

Juganda income (in thousands $)

	yearly income	income - mean	
a	71	9	81
b	62	0	0
c	66	4	16
d	61	-1	1
e	54	-8	64
f	67	5	25
g	55	-7	49
H	60	-2	4
mean	62		
		Total	240
		Count	8
		variance	30
		standard deviation	5.5

	name	yearly income	income - Mean	
1	a	99	37	13
2	b	14	-48	2
3	c	75	13	1
4	d	84	22	4
5	e	44	-18	3
6	f	54	-8	6
7	g	98	36	1
8	h	28	-34	1
	mean	62		

Total -	7166
Count	8
variance	895.75
Std Dev	29.9

$$\text{mean} = \frac{99+14+75+84+44+54+98+28}{8}$$

$$= 62$$

$$\text{variance} = \frac{\Sigma \text{income} - \text{mean}}{n} = \frac{7166}{8-1} = 895.75$$

$$\text{variance} = \sqrt{895.75}$$

$$= 29.9$$

find SD & variance

x	$d = x - \text{mean}$	d^2
23	-8	64
27	-4	16
28	-3	9
29	-2	4
30	-1	1
31	0	0
33	2	4
35	4	16
36	5	25
38	7	49

Mean = 31

tells you, you might decide to get an extra delivery driver on hand.

Prescriptive
(What's the best course of action?)

- to determine what should be done next.
- What steps can you take to avoid a future problem?
- prescriptive model considers all the possible decision patterns or pathways a company might take, and their likely outcomes.
- Based on all the possible scenarios and potential outcomes, the company can decide what is the best "route" or action to take.

Role of a data analyst

- **Gather data:** Analysts often collect data themselves. This could include conducting surveys, tracking visitor characteristics on a company website, or buying datasets from data collection specialists.
- **Clean data:** Raw data might contain duplicates, errors, or outliers. Cleaning the data means maintaining the quality of data in a spreadsheet or through a programming language so that your interpretations won't be wrong or skewed.

- **Model data:** Creating and designing the structures of a database. You might choose what types of data to store and collect, establish how data categories are related to each other, and work through how the data actually appears.
- **Interpret data:** Interpreting data will involve finding patterns or trends in data that could answer the question at hand.
- **Present:**by putting together visualizations like charts and graphs, writing reports, and presenting information to interested parties.

- The collection, transformation, and organization of data to draw conclusions make predictions for the future, and make informed data-driven decisions is called **Data Analysis.**
- The profession that handles data analysis is called a **Data Analyst.**

1. Step one: Defining the question

- The first step in any data analysis process is to define your objective
- Defining your objective means coming up with a hypothesis and figuring how to test it.
- you need to determine which sources of data will best help you solve it.
- “Why are we losing customers?”

16

2. Step two: Collecting the data

- A key part of this is determining which data you need. This might be quantitative (numeric) data, e.g. sales figures, or qualitative (descriptive) data, such as customer reviews.

All data fit into one of three categories:

- first-party
- second-party
- third-party data.

17

- **What is first-party data?**
- First-party data are data that you, or your company, have directly collected from customers.
- **What is second-party data?**
- . Second-party data is the first-party data of other organizations. This might be available directly from the company or through a private marketplace.
- **What is third-party data?**
- Third-party data is data that has been collected and aggregated from numerous sources by a third-party organization.
- Open data repositories and government portals are also sources of third-party data.

18

3. Step three: Cleaning the data

- Once you’ve collected your data, the next step is to get it ready for analysis. This means cleaning, or ‘scrubbing’ it, and is crucial in making sure that you’re working with high-quality data. Key data cleaning tasks include:
- **Removing major errors, duplicates, and outliers**
- Removing unwanted data points
- Bringing structure to your data
- Filling in major gaps

- Open data repositories and government portals are also sources of third-party data.

3. Step three: Cleaning the data

- Once you've collected your data, the next step is to get it ready for analysis. This means cleaning, or 'scrubbing' it, and is crucial in making sure that you're working with high-quality data. Key data cleaning tasks include:
- **Removing major errors, duplicates, and outliers**
- Removing unwanted data points
- Bringing structure to your data
- Filling in major gaps

4. Step four: Analyzing the data

- Here is where you use data analysis software and other tools to help you interpret and understand the data and arrive at conclusions

5. Step five: Sharing your results

- Data visualization is a fancy way of saying, "graphically show your information in a way that people can read and understand it ." You can use charts, graphs, maps, bullet points, or a host of other methods. Visualization helps you derive valuable insights by helping you compare datasets and observe relationships.

Data Analysis Tools

- Tableau Public
- R Programming
- Python
- Apache Spark
- SAS
- Excel
- RapidMine

Skills needed to become a data analyst.

1Mathematical skills (especially statistics and probability)

2. Knowledge of R and Python are highly used for data analysis, visualization, and building Machine Learning models

3. Thorough knowledge of database languages such as SQL

4. Experience in data visualization tools such as Tableau, Qlik, or PowerBI

Assignment - 1
D.A

* Questions:

1 What is z-score?

-> Z-Score is a statistical measurement that describes a value's relationship to the mean of a group of values.
If a value has high enough or low enough z-scores, it can be considered an outliers. As a rule of thumb values with a z-score greater than 3 or less than -3 are often determined to be outliers

z-score = (x-mean) / standard deviation

a,

Data	x-mean	Square (x-mean)	z-score = (x-mean)/std
100	-58.333	3402.778	-0.532
110	-48.333	2336.078	-0.441
110	-48.333	2336.078	-0.441
110	-48.333	2336.078	-0.441
120	-38.333	1469.418	-0.350
120	-38.333	1469.418	-0.350
130	-28.333	802.758	-0.258
140	-18.333	336.098	-0.167
140	-18.333	336.098	-0.167
150	-8.333	69.438	-0.076
170	11.667	136.118	-0.106
500	341.667	116736.338	3.121

mean = sum/n = 158.333 Std.Dev = 109.447

∴ Therefore, 500 is an outliers as its z-score is more than 3

2. Fit trend line & orginal data line using semi average method (both line):

Year	2008	2009	2010	2011	2012	2013
Income	46.17	51.65	63.81	70.19	84.91	91.64

Year	Income	Average
2008	46.17	$\frac{46.17+51.65+63.81}{3} = 53.87$
2009	51.65	

500	341.667	116736.338	3.121

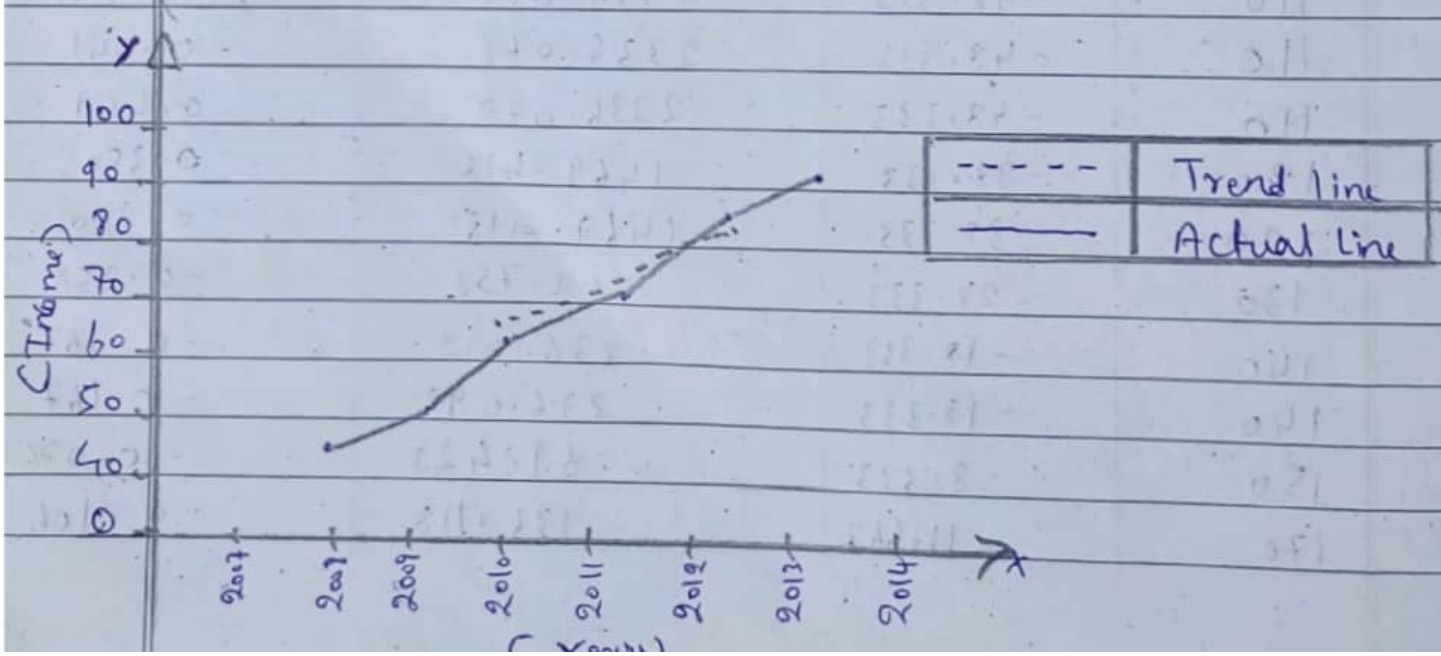
mean = sum/n = 158.333 Std. Dev = 109.447

∴ Therefore, 500 is an outliers as its Z-score is more than 3

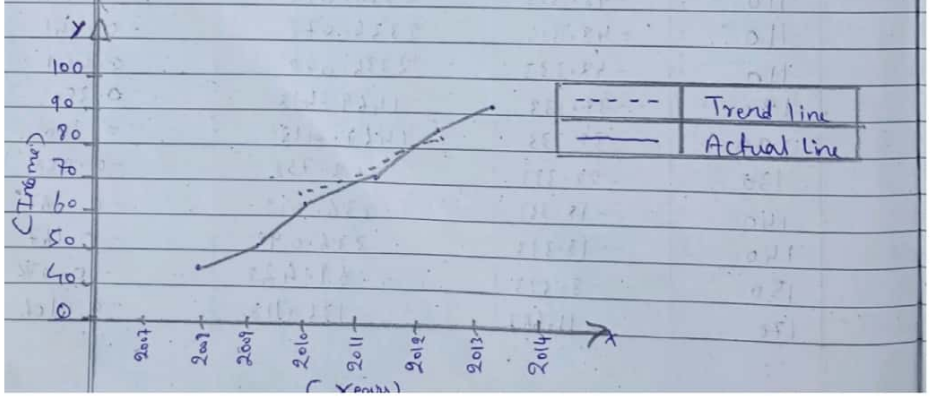
2. Fit trend line & orginal data line using semi average method (both line):

Year	2008	2009	2010	2011	2012	2013
Income	46.17	51.65	63.81	70.99	84.91	91.64

Year	Income	Average
2008	46.17	$\frac{46.17+51.65+63.81}{3} = 53.87$
2009	51.65	
2010	63.81	
2011	70.99	$\frac{70.99+84.91+91.64}{3} = 82.51$
2012	84.91	
2013	91.64	



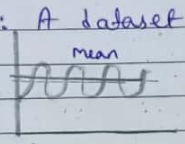
Year	Income	Average
2008	46.17	$\frac{46.17+51.65+63.81}{3} = 53.87$
2009	51.65	
2010	63.81	
2011	70.99	$\frac{70.99+84.91+91.64}{3} = 82.51$
2012	84.91	
2013	91.64	



Date Page

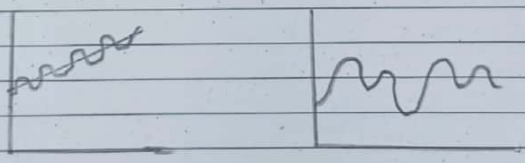
3 Explain the datatypes of the time series with diagram. There are two types of data types in time series which are disscussed below

i Stationary: A dataset should follow the below



Stationary datatypes refers to the data types whose statistical properties of a process generating a time series do not change over time. It maintains same sequence
Eg:- mean, variance.. etc..

ii Non Stationary: A dataset which change its sequence with respect to time is known as non-stationary.



Assignment: 02

1 Compute Karl pearson's Coefficient of Co-relation from the following data

price (x)	10	20	30	40	50	60	70
Supply (Units)	8	6	14	16	10	20	24

price (x)	$x = x - \bar{x}$	x^2	y	$y = y - \bar{y}$	y^2	xy
10	-30	900	8	-6	36	30x60=18
20	-20	400	6	-8	64	160
30	-10	100	14	0	0	0
40	0	0	16	2	4	0
50	10	10	10	-4	16	-40
60	20	400	20	6	36	120
70	30	900	24	10	100	300
$\Sigma x = 280$		$x = 2800$			$y^2 = 256$	$\Sigma xy = 72$

$$\bar{X} = \frac{\Sigma x}{N} = \frac{280}{7} = 40$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{720}{\sqrt{2800 \times 256}} = \frac{720}{846.64} = 0.845$$

Hence the value is near to 1 there exist positi Correlation between price (x) Supply)

Scanned with OKEN Scanner

Find the Correlation with Spearman's Rank Correlation Case.

X	70	80	65	78	68	65	82	65
y	13	15	12	14	13	11	16	10

Authorize the system to edit this file. Authorize

Edit Annotate Fill & Sign Convert All

Scanned with OKEN Scanner

Find the Correlation with Spearman's Rank Correlation Case.

X	70	80	65	78	68	65	82	65
Y	13	15	12	14	13	11	16	10

X	R: X-$\bar{x}$	y	Ry	d=Rx-Ry	d^2
70	4	13	4.5	-0.5	0.25
80	2	15	2	0	0
65	7	12	6	1	1
78	3	14	3	0	0
68	5	13	4.5	0.5	0.25
65	7	11	7	0	0
82	1	16	1	0	0
65	7	10	8	1	1

Σd^2 2.5

$$\gamma = 1 - 6\left[\Sigma d^2 + \frac{1}{12}(2^3-3) + \frac{1}{12}(3^3-2)\right] \Big/ n^3-n$$

$$\gamma = 1 - \frac{6[2.5 + 2 + 0.5)}{8}$$

$$\gamma = 1 - \frac{6(5)}{504}$$

$$= 1 - \frac{(30)}{504} \quad = 1 - 0.0592$$

$$= 0.9405$$

3 Difference b/w Regression & Correlation :

->

Regression	Correlation
It is used numerically	It is used to determine

40-60	22

Types of Frequency Distributions?

There are four types of frequency distributions that are as follows:

- Grouped Frequency Distribution
- Ungrouped Frequency Distribution
- Relative Frequency Distribution
- Cumulative Frequency Distribution

Grouped Frequency Distribution

- In Grouped Frequency Distribution observations are divided between different intervals known as class intervals and then their frequencies are counted for each class interval. This Frequency Distribution is used mostly when the data set is very large.

Class Interval	Frequency
10 – 20	5
20 – 30	8
30 – 40	12

Ungrouped Frequency Distribution

- In Ungrouped Frequency Distribution, all distinct observations are mentioned and counted individually. This Frequency Distribution is often used when the given dataset is small.

Value	Frequency
10	4
15	3
20	2
25	3
30	2

Relative Frequency Distribution

- This distribution displays the proportion or percentage of observations in each interval or class. It is useful for comparing different data sets or for analyzing the distribution of data within a set.

- Relative Frequency = Frequency of the Event/Total Number of Events

Relative Frequency Distribution

- This distribution displays the proportion or percentage of observations in each interval or class. It is useful for comparing different data sets or for analyzing the distribution of data within a set.

- Relative Frequency = Frequency of the Event/Total Number of Events

Score Range	Frequency	Relative Frequency
0-20	5	5/50 = 0.10
21-40	10	10/50 = 0.20
41-60	20	20/50 = 0.40
61-80	10	10/50 = 0.20
81-100	5	5/50 = 0.10
Total	**50**	

Cumulative Frequency Distribution

- Cumulative frequency is defined as the sum of all the frequencies in the previous values or intervals up to the current one.

Runs	Frequency
0-10	2
10-20	2
20-30	1
30-40	4

Runs	Cumulative Frequency
Less than 10	**2**
Less than 20	**4**
Less than 30	**5**

The Measure of Central Tendency
MEAN

- Mean is generally the average of a given set of numbers or data
- The average (mean) marks obtained by the students in a class.
- A cricketer's average is also an example of a mean.

Mean Formula

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

The Measure of Central Tendency
MEAN

- Mean is generally the average of a given set of numbers or data
- The average (mean) marks obtained by the students in a class.
- A cricketer's average is also an example of a mean.

Mean Formula

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

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000
Average	6250

Calculate the mean of the first 10 natural numbers.

Solution:

First 10 natural numbers = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Sum of first 10 natural numbers = (1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10)

Mean = Sum of 10 natural numbers/10

⇒ Mean = (1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10)/10

⇒ Mean = 55/10

⇒ Mean = 5.5

Merits:

1) It is easy to calculate and simple to understand for the learners.
2) Mean is based on all elements of the given data.
3) It is widely used in statistical analysis of any of the data because this is related to the mathematical operations.
4)Its value is always definite
5)Can be used for comparisons

Types of outliers

- A **univariate outlier** is an extreme value that relates to just one variable. For example, Sultan Kösen is currently the tallest man alive, with a height of 8ft, 2.8 inches (251cm). This case would be considered a univariate outlier as it's an extreme case of just one factor: height.

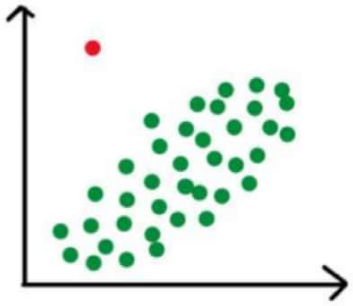
Multivariate outlier

- ITis a combination of unusual or extreme values for at least two variables.
- However, when you consider these two observations in conjunction, you have an adult who is 5ft 9 inches and weighs 110lbs—a surprising combination. That's a multivariate outlier.

- **Global outliers (otherwise known as point outliers)** are single data points that lay far from the rest of the data distribution.
- In most cases, all the outlier detection procedures are targeted to determine the global outliers.

Causes: Errors in data collection, measurement errors, or truly unusual events can result in global outliers

Example-one students marks entry is more than 100 marks as paper is only 100 marks

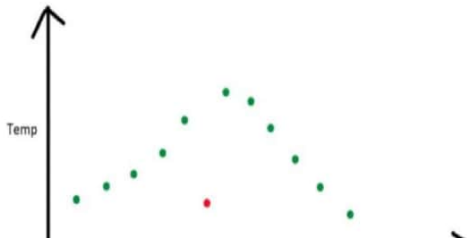


The red data point is a global outlier.

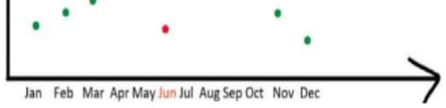
Contextual outliers (conditional outliers)

- Defines the context, e.g., time & location
- Contextual outliers may not be outliers when considered in the entire dataset, but they exhibit unusual behavior within a specific context or subgroup.

A low temperature value in June is a contextual outlier because the same value in December is not an outlier.



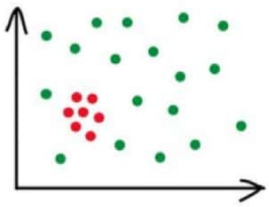
December is not an outlier.



Collective Outliers

- In a given set of data, when a group of data points deviates from the rest of the data set is called collective outliers.
- Collective outliers may not be outliers when considered individually, but as a group, they exhibit unusual behavior.
- Collective outliers can represent interesting patterns or anomalies in data that may require special attention or further investigation.

For example, if the high temperature was 40 degrees for 30 days straight then that would be suspicious.
42 degrees for few days is a perfectly average temperature that does not sound any alarm bells on its own



The red data points as a whole are collective outliers.

How can you identify outliers?

The methods commonly used to identify outliers with

- There are **four ways** to identify outliers:

1. Sorting method
2. Data visualization method
3. Interquartile range method
4. Statistical tests (z scores)

Sorting method

- You can **sort** quantitative variables from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.
- This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.

Your dataset for a pilot experiment consists of 8 values.

180	156	9	176	163	1827	166	171

You sort the values from low to high and scan for extreme values.

9	156	163	166	171	176	180	**1872**

visualizations

- You can use software to **visualize** your data with a box plot, or a box-and-whisker plot, so you can see the data distribution at a glance. This type of chart highlights minimum and maximum values (the range), the median, and the interquartile range for your data.
- Many computer programs highlight an outlier on a chart with an

visualizations

- You can use software to **visualize** your data with a box plot, or a box-and-whisker plot, so you can see the data distribution at a glance. This type of chart highlights minimum and maximum values (the range), the median, and the interquartile range for your data.
- Many computer programs highlight an outlier on a chart with an asterisk, and these will lie outside the bounds of the graph.

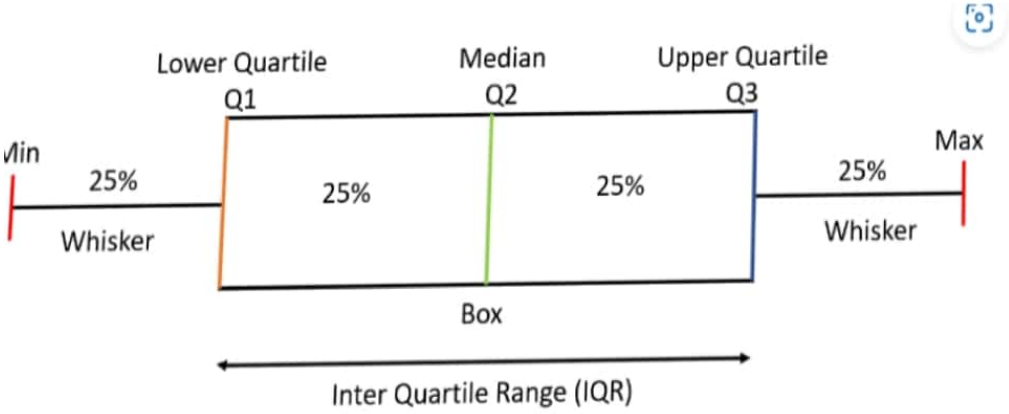
- **Box Plot:** It is a type of chart that depicts a group of numerical data through their quartiles. It is a simple way to visualize the shape of our data. It makes comparing characteristics of data between categories very easy.
- **Minimum** – It is the minimum value in the dataset excluding the outliers
- **First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.

Median (Q2) – It is the mid-point of the dataset. Half of the values lie below it and half above.

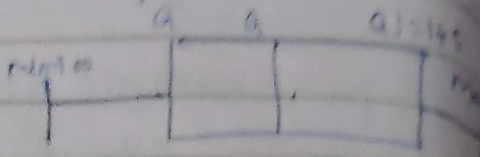
Third Quartile (Q3) – 75% of the data lies below the Third (Upper) Quartile.

Maximum – It is the maximum value in the dataset excluding the outliers.

Whiskers in a boxplot are lines that extend from the box to the minimum and maximum data values. They indicate the variability of the data outside the upper and lower quartiles, which are the ends of the box



The Outliers
which are outside this range are
The outliers = 220



data = [32, 36, 46, 47, 48, 50, 51, 600]

Q1 = 41
Q2 = 47.5
Q3 = 50.5
IQR = Q3 - Q1 = 9.5
lwrbound = q1 - (1.5 * IQR) = 26.75
uprbound = q3 + (1.5 * IQR) = 64.75
600 is Outliers.

data = [32, 36, 42, 47, 49, 50, 52, 54, 55, 56,5

median (Q2) = (50 + 52)/2 = 51
Q1 = (49 + 47)/2 = 44.5
Q3 = 54.5
IQR = Q3 - Q1 = 10
lwr bound = Q1 + (1.5 * IQR) =
uprbound = Q3 + (1.5 * IQR) =
600 is Outliers

00, 110, 110, 110, 120, 130, 130, 140, 140,
300

median (Q2) = 120
Q1 = 110
Q2 = 120

Q3 = 140

IQR = Q3 - Q1 = 140 - 110 = 30

lower Bound = Q1 - 1.5* IQR

= 110 - 1.5* 30

= 65

upper bound = Q3 + 1.5* IQR

140 + 1.5* 30

= 185

Statistical methods - Z Score:

Statistical outlier detection inv

applying statistical tests or pr

to identify extreme values.

You can convert extreme data

into Zscores that tell you h

Standard deviations away

from the mean.

If is value has a high en

or low enough Zscore it c

considered as outlier as a

of thumb values with a

greater than 3 or less th

outliers other determined a

outliers.

formula = Zscore → (x - mean) /

b) Replacing with mean:

- It is the common method of imputing missing values. However in presence of outliers, this method may lead to erroneous imputations. In such cases, median is an appropriate measure of central tendency. For some reasons, if you have to use mean values for imputation, then treat the outliers before imputations.

dataset['Height']=dataset['Height'].fillna((dataset['Height'].mean()))

Height	W
12.0	
NaN	
13.0	
15.0	
16.0	
NaN	
NaN	

Mean=(12+13+15+16)/4=14

Height
12.0
14.0
13.0
15.0
16.0
14.0
14.0

c) Replacing with Median:..

- As median is a position based measure of central tendency (middle most item), this method is not affected by presence of outliers

Weight	C
35.0	
36.0	
32.0	
NaN	
39.0	
NaN	
NaN	

Median=(32+39)/2=35.5

Weight
35.0
36.0
32.0
35.5
39.0
35.5
35.5

d) Replacing with Mode

- Replacing with mode is little bit trickier. Because unlike mean and median, mode returns a dataframe. Why? Because if there are two modal values, pandas will show both these values as modes.

For example, let us say our data set is ['A', 'A', 'B', 'C', 'C'].
Here both 'A' and 'C' are the modes as they are repeated equal number of times. Hence mode returns a dataframe containing 'A' and 'C' not a single value.

While replacing with mode, we need to use mode()[0] at the end

dataset['Country']=dataset['Country'].fillna((dataset['Country'].mode()[0]))

Country
india

Mode is US

Country
india

Seasonality

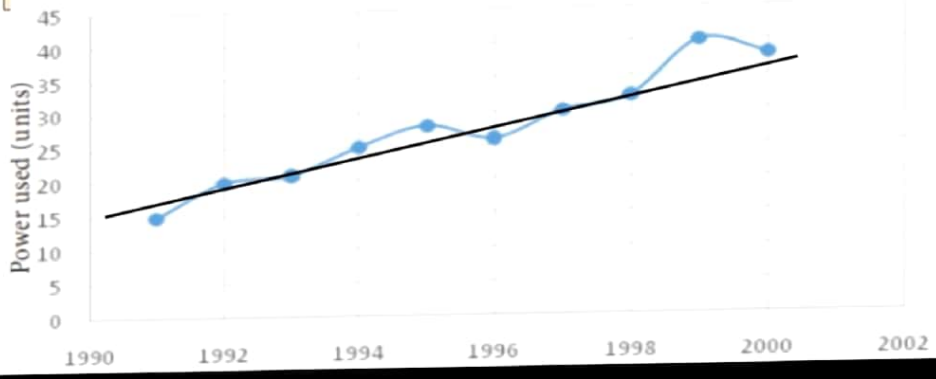
- Seasonality in time series data refers to patterns that repeat over a regular time period, such as a day, a week, a month, or a year. These patterns arise due to regular events, such as holidays, weekends, or the changing of seasons, and can be present in various types of time series data, such as sales, weather, or stock prices.
- It's important to note that time series data can have multiple types of seasonality present simultaneously,

Cyclicity

- Cyclicity in time series data refers to the repeated patterns or periodic fluctuations that occur in the data over a specific time interval. It can be due to various factors such as seasonality (daily, weekly, monthly, yearly), trends, and other underlying patterns.
- **Irregularity**: Unexpected situations/events/scenarios and spikes in a short time span.

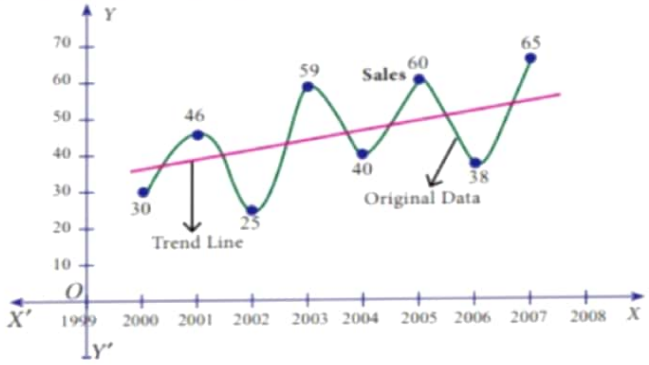
Annual power consumption per household in a certain locality was reported below.

Years	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Power used (units)	15	20	21	25	28	26	30	32	40	38

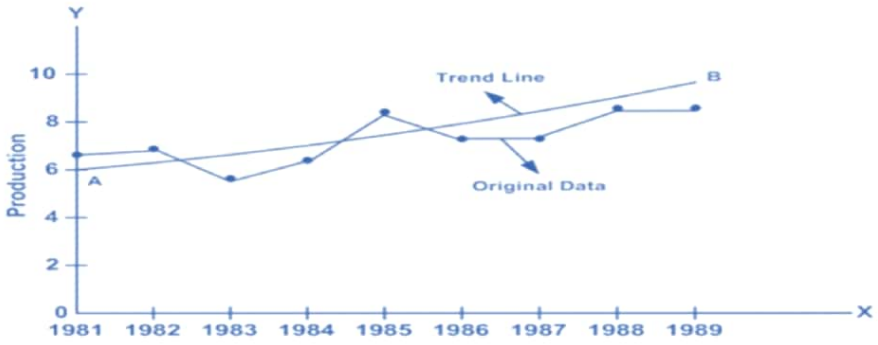


Fit a trend line by the method of freehand method for the given data.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Sales	30	46	25	59	40	60	38	65



Years	1981	1982	1983	1984	1985	1986	1987	1988	1989
Production Million Metric Tons	6.6	6.9	5.6	6.3	8.4	7.2	7.2	8.5	8.5



Semi-Average Method

- In this method, the series is divided into two equal parts and the average of each part is plotted at the mid-point of their time duration.
- (i) In case the series consists of an even number of years, the series is divisible into two halves

Semi-Average Method

- In this method, the series is divided into two equal parts and the average of each part is plotted at the mid-point of their time duration.
- (i) In case the series consists of an even number of years, the series is divisible into two halves
- ii) In case the series consists of odd number of years, it is not possible to divide the series into two equal halves. The middle year will be omitted.

Merits

- This method is very simple and easy to understand
- It does not require many calculations.

Demerits

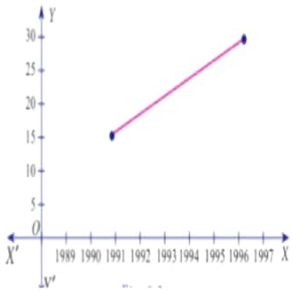
- This method is used only when the trend is linear.
- It is used for calculation of averages and they are affected by extreme

---

Fit a trend line by the method of semi-averages for the given data.

Year	1990	1991	1992	1993	1994	1995	1996	1997
Sales	15	11	20	10	15	25	35	30

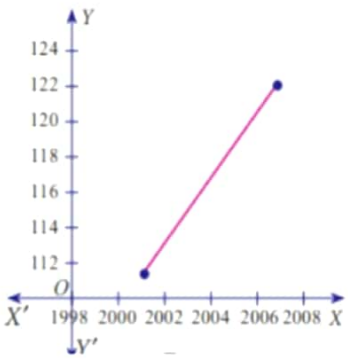
Year	Production	Average
1990	15	$\frac{15+11+20+10}{4}=14$
1991	11	
1992	20	
1993	10	
1994	15	
1995	25	$\frac{15+25+35+30}{4}=26.25$
1996	35	
1997	30	



Fit a trend line by the method of semi-averages for the given data.

Year	2000	2001	2002	2003	2004	2005	2006
Production	105	115	120	100	110	125	135

Year	Production	Average
2000	105	
2001	115	$\frac{105+115+120}{3}=113.33$
2002	120	
2003	100 (left out)	
2004	110	
2005	125	$\frac{110+125+135}{3}=123.33$
2006	135	



Year	Sales in lakh	4y T	4M A	centered	Year
1982	58				
1983	74				
1984	86	286	71.5	74.25	1984
1985	68	308	77.0	78.25	1985
1986	80	318	79.5	82.50	1986
1987	84	342	85.5	87.75	1987
1988	110	360	90	97.75	1988
1989	86	374	93.50	95.75	1989
1990	94	392	98	100	1990
1991	102	408	102	104.45	1991
1992	126	428	107	106.50	1992
1993	106	424	106		
1994	90				

	X	y	x	square of x	xy	Trend value----- y=a+bx	
1	1981	80	-3	9	-240	90+2(-3)	84
2	1982	90	-2	4	-180	90+2(-2)	86
3	1983	92	-1	1	-92	90+2(-1)	88
4	1984	83	0	0	0	90+2(-0)	90
5	1985	94	1	1	94	90+2(+1)	92
6	1986	99	2	4	198	90+2(+2)	94
7	1987	92	3	9	276	90+2(+)	96
	sum	630	0	28	56	sum	630
	n=7						
	a=sumof y/n	90	b=sum of xy/sum ofxsqare		2	y=a+bx	
		Estimation of year 1999=		90+2(6)=	102		