Date:
P. No:

unit - 1

Data science :- Data science is kinda blended with various tools, algorithms and machine learning principles most simply it involves obtaining meaningful information or insights from structured or unstructured data and the process of analyzing, programming and business skills.

* Data collection
* Data cleaning
* Data modelling
* Data analysis
* optimization and deployment

Advantages

* Improved decision making
* cost - effective
* personalization
* innovation

Disadvantages

* Data quality
* complexity
* Bias

Date:
P. No:

Types of data -

1) structured data
2) unstructured data

1) structured data :- structured data refers to data that is organized in a specific format, typically in a table with rows & colums this format is easy for computers to process and analyze

Exp., include spreadsheets, database and csv files

Advantages of structured data

* easy to analyze
* easy to organize and store
* easy to search
* security

Disadvantages of structured data

* Limited flexibility

Date:
P. No:

* missing data
* Limited ability
* Data Integrity issues

Tools,
SQL, Excel, R, Python, Jableav

2) unstructured data :- unstructured data is the data which does not conforms to a data model and has no easily identifiable structure such that it can not be used by a computer program easily unstructured data is not organised in a predefined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database

Exp., video, Image

Disadvantages of ~~its~~ unstructured data

* limited query
* Data Integration
* Data quality

Advantages

* Flexibility
* Real time insights
* Customer understanding

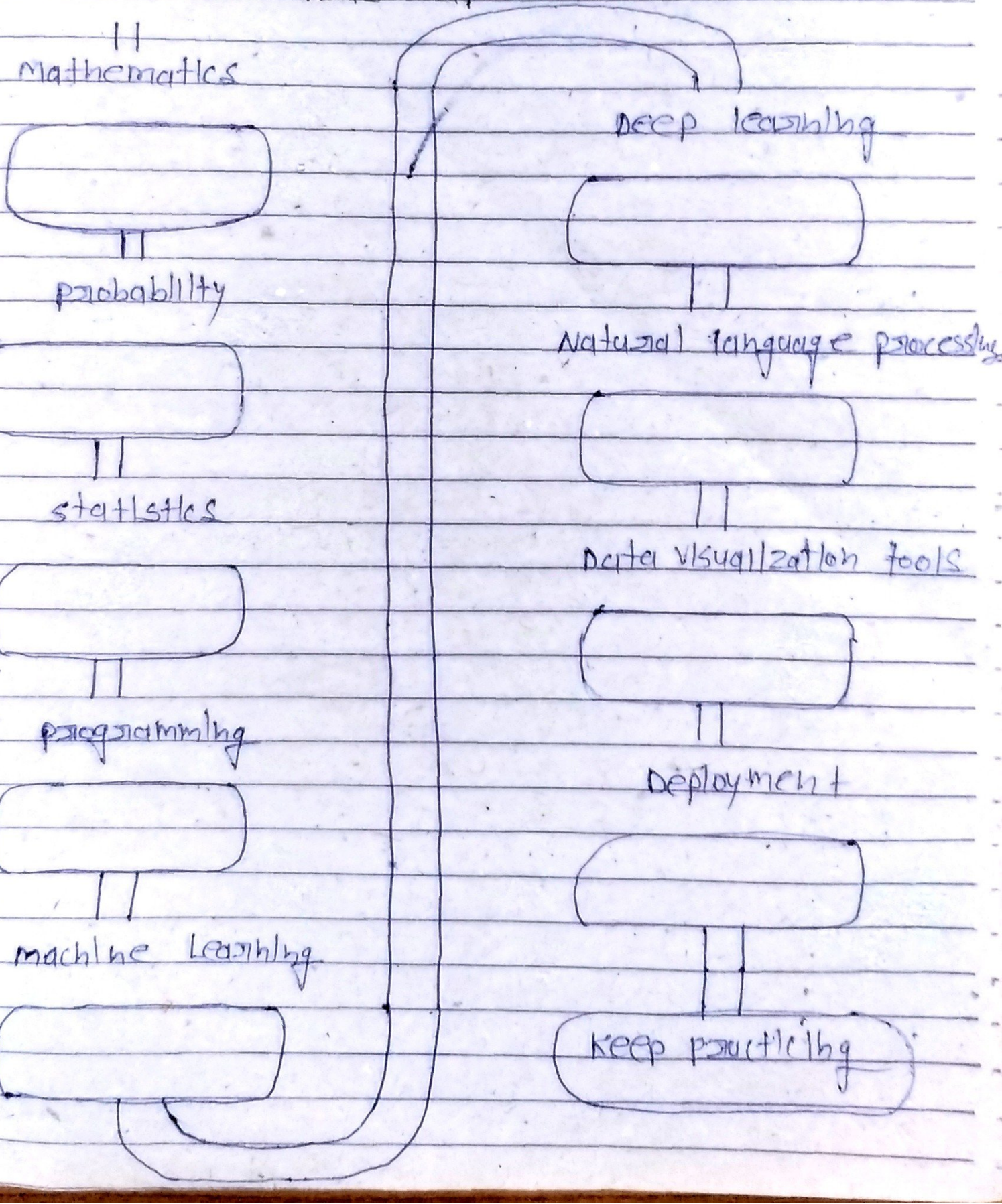
(4)

Date
P. No.

structured data	unstructured data
1. It is based on a relational database	It is based on characte and binary data
2. It is very robust	It is less robust
3. It is easy to search	It is more difficult to search
4. structured data is less flexible	unstructured data is more flexible
5. It is pre-defined format	It is not pre defined format
6. It is using method of analysis like Regression and classification	It is using method of analysis like data mining and Data stacking
7. It is hard to scale database schema	It is more scalable

Date:
P. No:

Data science Road map



Date:
P. No:

1. Mathematics :- Math skill is very important as they help us in understanding various machine learning algorithms that play an important role in Data science.

- Part 1

linear algebra
matrix
vector calculus

- Part 2

Regression
classification
Density estimation

2. Probability :- Probability is also significant to statistics and it is considered a prerequisite for mastering machine learning.

* Introduction to probability
* 1D Random variable
* Joint probability distribution

* Discrete distribution
* Continuous distribution
* Normal distribution

Date:
P. No:

3. statistics :- The understanding of statistics is very significant as this is a part of Data analysis

* Introduction to statistics
* Hypotheses testing
* Basics of Graphs
* multiple Regression
* correlation

4. programming :- one needs to have a good grasp of programming concepts such as data structures and algorithms. The programming languages are used are python, R, Java, C++ is also useful in some places where performance is very important

- Python

Basic python
- list
- tuples
- Set
- dictionary
- function

- R

Basic R
- list
- matrix
- array
- function
- Data frame

Date:
P. No:

5. Machine learning :- machine learning is one of the most vital parts of data science and the hottest subject of research among researchers. So one at least needs to understand basic algorithms of supervised and unsupervised learning. There are multiple libraries available in Python and R for implementing these algorithms.

* Introduction
  - model validation
  - underfitting & overfitting
  - ML model
  - Data leakage
  - cross validation

6. Deep learning :- Deep learning uses TensorFlow and Keras to build and train neural networks for structured data

Date:
P. No:

* Neural Network
  - Artificial neural network
  - Convolutional neural network
  - Recurrent neural network
  - Deep Neural network
  - Binary classification

7. Natural language processing :- in NLP distinguish of learning to work with text data
   - Text classification
   - word vectors

8. Data Visualization Tools :- make great data visualizations A great way to see the power of coding
   - EXCEL VBA

   BI (Business Intelligence)
   - Tableau
   - Power BI

Date:
P. No:

9. Deployment :- The last part is doing the deployment definitely whether you are fresher.

10. Keep practicing :- so keep practing and improving your knowledge day by day. Below is a complete diagrammatical representation of the Data scientist Roadmap.

Data wrangling :- Data wrangling also known as data cleaning or data pre-processing refers to the process of transforming raw and messy data into a usable format for analysis this process typically involves several tasks such as removing or filling missing values handling duplicates dealing with outliers formatting data and merging or joining data set

process of Data wrangling -

Date:
P. No:

1. Data collection
2. Data cleaning
3. Data transformation
4. Data integration
5. Data validation
6. Data organization

1. Data collection :- Gathering raw data from various source such as database spreadsheets, text file or web application.

2. Data cleaning :- This step involves removing or filling missing values handling duplicate dealing

3. Data transformation :- This step involves formatting the data to a consistent and usable format

4. Data integration :- This step involves combining ~~the data to a consistent~~ multiple data set into a single data set often using a common field or key

5. Data validation :- this steps follows similar logic utilized in data normalization & data standardization process involving validation rules

Date:
P. No:

Data Exploratory analysis (DEA) EDA

The Exploratory data analysis is an approach that is used to analyze the data and discover trends, patterns or check assumptions in data with the help of statistical summaries and graphical representation

Types of EDA

1. Univariate analysis :- In univariate analysis we analyze or deal with only one variable at a time it does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

2. Bi-variate analysis :- This type of data involves two different variables The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship between the two variables

Date:
P. No:

3. Multivariate analysis :- when the data involves three or more variables it is categorized under multivariate.

The parts of EDA

1. Non-graphical Analysis :- In non graphical analysis we analyze data using statistical tools like mean, median or mode or skewness $\Rightarrow \frac{q_3 + q_1 - 2m_2}{q_3 - q_1}$

2. graphical Analysis :- In graphical analysis we use visualizations charts to visualize trends and patterns in the data.

Exploratory data analysis using python libraries

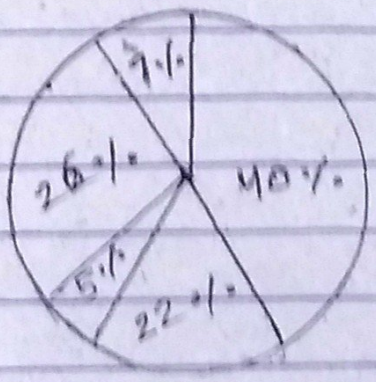
```
Import pandas as pd
Import numpy as np
```

Date:
P. No:

Graphical summaries of data

* Pie chart
* Bar Graph
* Pareto chart
* Histogram

1) Pie chart :- A pie chart is a type of graph that represents the data in the circular graph the slices of pie show the relative size of the data pie chart works the best at that time in most cases pie charts replace other graphs like the bar graph line plots, histogram etc.

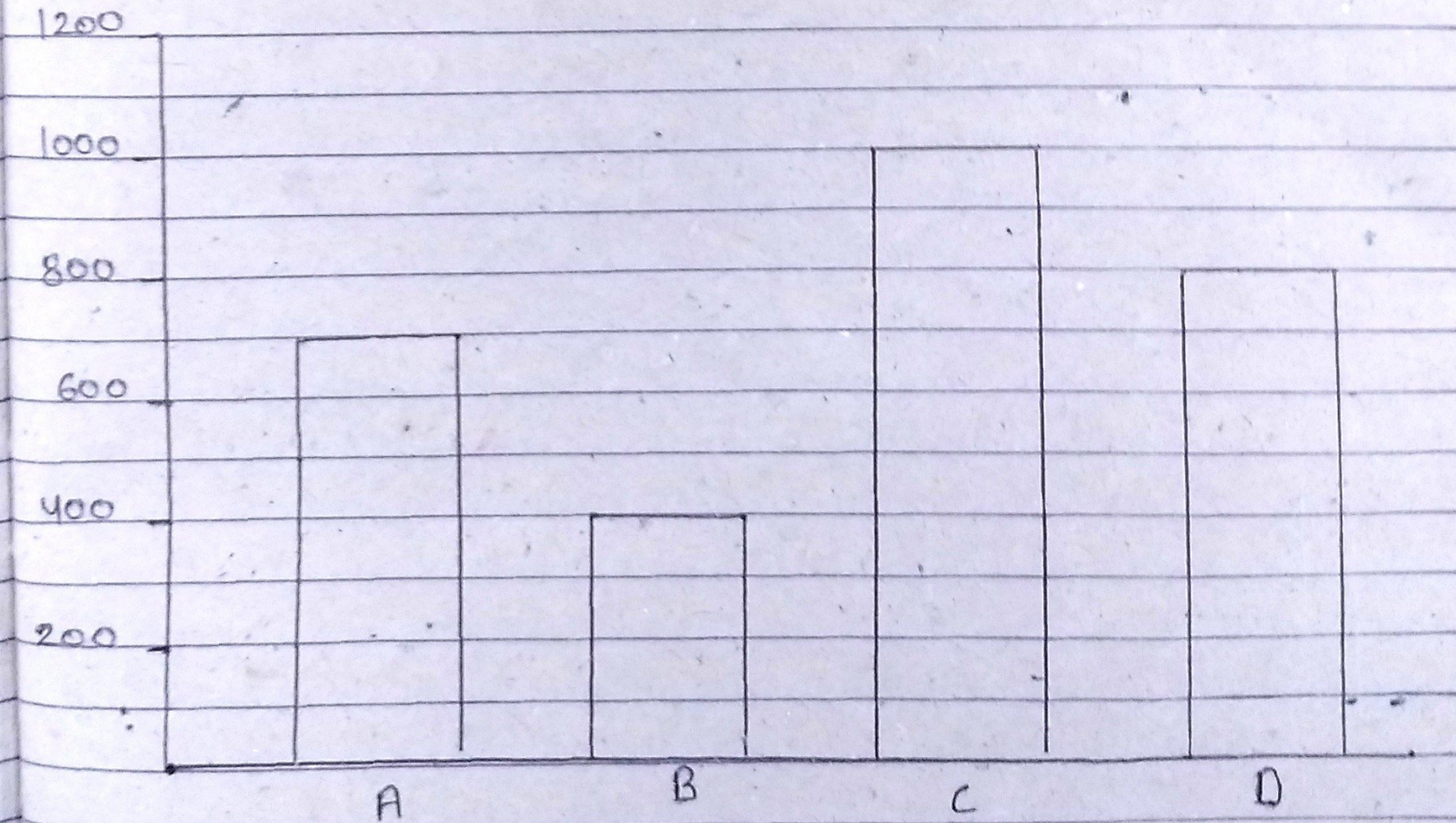


pie chart

Date:
P. No:

2) Bar Graph :- Bar Graph also known as Bar chart is a graphical representation of data that uses rectangular bars or columns to show the magnitude of value in different categories groups

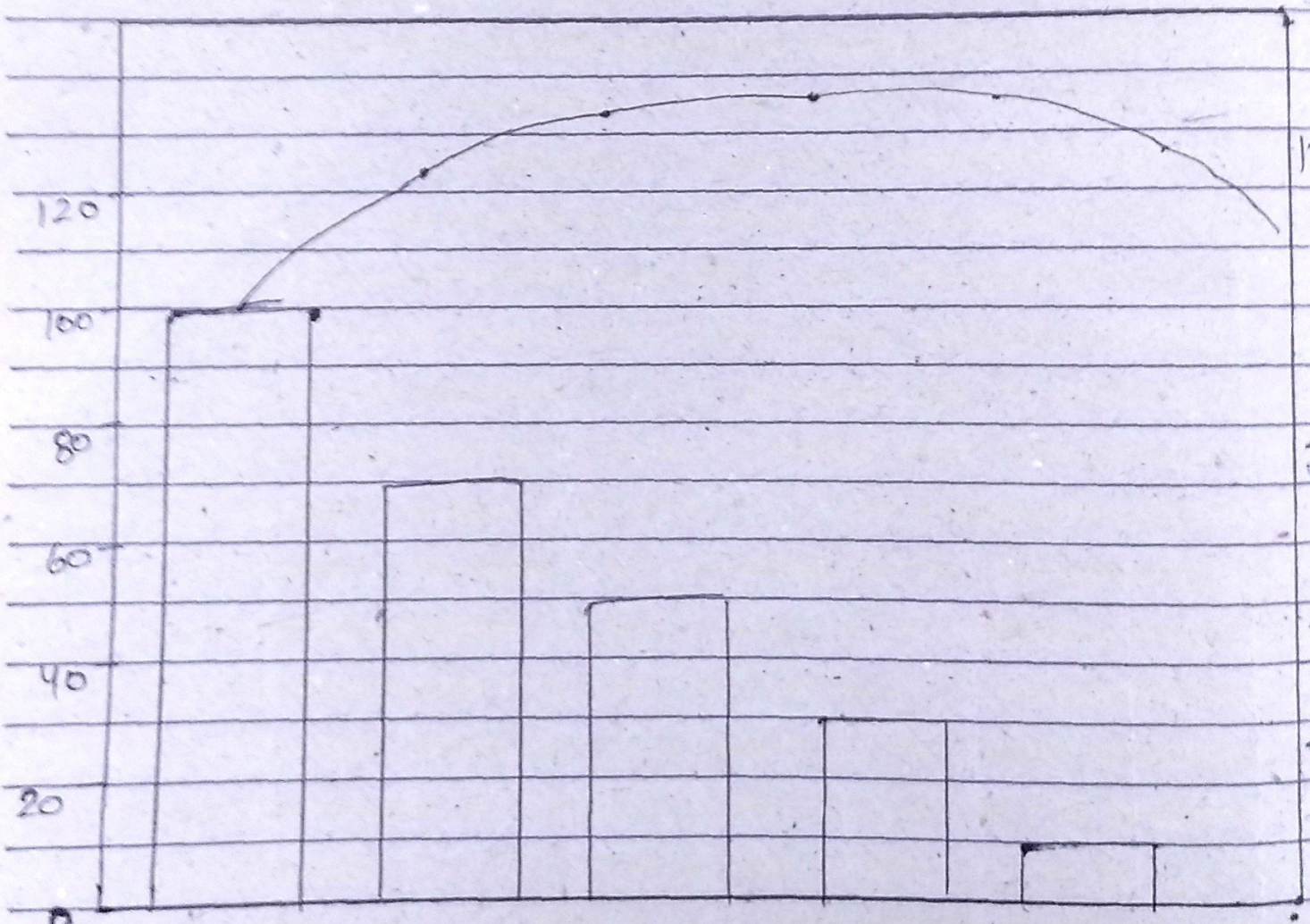
The vertical or y-axis of the chart represents the value or frequency the data while the horizontal or x-axis shows the different categories or ~~graph~~ groups being compared



Bar Chart

Date:
P. No:

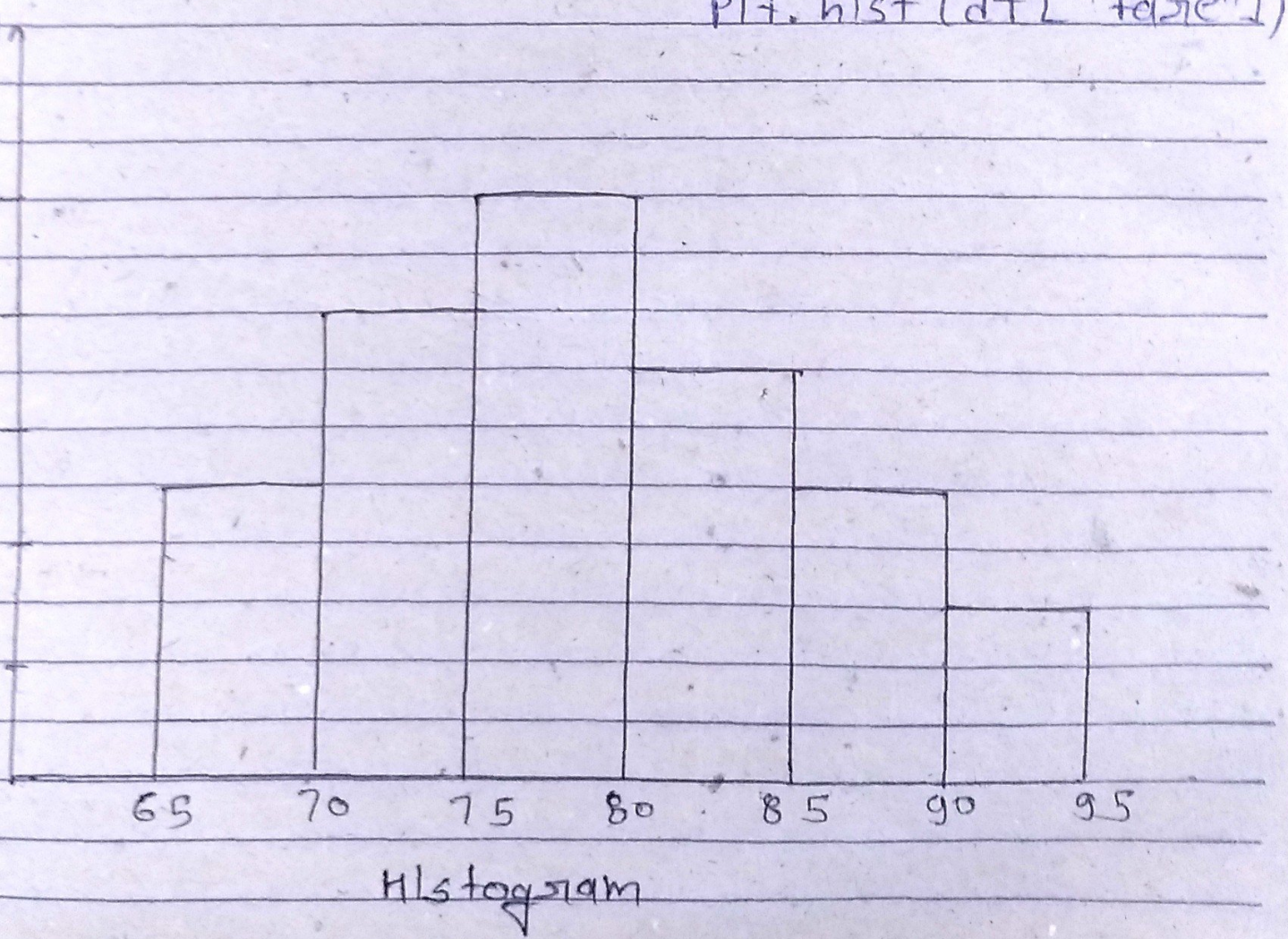
3. Pareto chart :- A pareto chart is a graphical tool used to display the relative importance of values items or categories often used in quality control & improvement it is based on the pareto principle which states that a small number of causes are responsible for a large percentage of the effects



Date:
P. No:

4. Histogram :- A histogram is a type of graph that displays the distribution of data in a set, it is a graphical representation of the frequency distribution of a continuous variable. The data is divided into a set of intervals or bins & the number of data points that fall within each bin is represented by the height of a bar.

plt.hist(df["fare"])



Histogram

Date:
P. No:

Measures of central tendency of quantitative -

1. mean
2. median
3. mode

1. mean :- The mean represents the average value of the dataset It can be calculated as the sum of all the values in the dataset divided by the number of values.

$$\text{mean } (\bar{x}) = \frac{\Sigma f(x)}{\Sigma f}$$

some other mean find the measures of central tendency

* Geometric mean
* Harmonic mean
* weighted mean

Date:
P. No:

1. Geometric mean :- The geometric mean is a type of average that is calculated by taking the nth root of the product of n numbers

$$\log GM = \frac{1}{n} \sum_{i=1}^{n} f_i \log x_i$$

2. Harmonic mean :- The harmonic mean is another type of average that is used to calculate the average of a set of numbers. It is particularly of a set useful when dealing with rates

$$H = \frac{N}{\sum f_i / x_i}$$

3. weighted mean :- The weighted mean is a type of average that is calcuted by multiplying each value by its corresponding weight,

$$W = \frac{\sum w_i x_i}{\sum w_i}$$

Date:
P. No:

2. Median :- median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order.

$$Md = l + \left(\frac{\frac{N}{2} - CF}{f} \right) \times h$$

Exp.

class	frequencies (f1)	CF
0 - 10	22	22
10 - 20	38	60
20 - 30	46	106
30 - 40	35	141
40 - 50	20	161
	$\Sigma f1 = N =$ 161	

$\frac{N}{2} = \frac{161}{2}$

$= 80.5$

median class will be 20 - 30

$l = 20$, $CF = 60$

$f = 46$, $h = 10$

$= 20 + \left(\frac{80.5 - 60}{46} \right) \times 10$

$= 20 + \left(\frac{20.5}{46} \right) \times 10$

$= 20 + 0.445 \times 10$

$= 20 + 4.456$

$= 24.45$

Ans

Date:
P. No:

3. Mode :- The mode represents the frequently occurring value in the dataset

$$M_o = l + \left[\frac{f - f_0}{2f - f_0 - f_1}\right] \times h$$

Exp.

class	frequency
0 - 2	6
2 - 4	14
4 - 6	16
6 - 8	20
8 - 10	2
10 - 12	4
12 - 14	6

The highest frequency is 20 so that the mode class 6 - 8

$l = 6$, $f = 20$, $f_0 = 16$

$f_1 = 2$, $h = 2$

$$= 6 + \left[\frac{20 - 16}{2 \times 20 - 16 - 2}\right] \times 2$$

$$= 6 + \left[\frac{4}{40 - 16 - 2}\right] \times 2$$

$$= 6 + 0.18 \times 2$$

$$= 6.363$$

Date:
P. No:

Measures of variability of quantitative Data

1. Range
2. standard deviation
3. variance

1. Range :- The range is the difference between the maximum and minimum values in a data set, it provides a quick & easy measure of the spread of the data, but it is sensitive to extreme values & outliers

$$\text{Range}(x) = \max(x) - \min(x)$$

2. standard deviation :- The standard deviation is the square root of the variance it is often used as a measure of the spread of the data in a normal distribution & is less sensitive to extreme values & outliers than the variance

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

Date:
P. No:

3. variance :- The variance is the average of the standard deviation from the mean It measure how much the data deviates from the mean, and is widely used in statistical analysis However it is sensitive to extreme values, and outliers

$$\text{variance} = \sigma^2$$

$$\sigma^2 = \frac{\Sigma f_i (x_i - \bar{x})^2}{N}$$

Date:
P. No:

Probability :- Probability is a measure of the likelihood or chance that a particular event or outcome will occur. It is a numerical value between 0 and 1

$$P(A) = \frac{n(A)}{n(S)}$$

$P(A)$ = $P(A)$ is the probability of an event A

$n(A)$ = $n(A)$ is the number of favourable outcomes

$n(S)$ = $n(S)$ is the total number of events in the sample space

Date:
P. No:

Conditional probability :- The probability of occurrence of any event A when another event B in relation to A has already occurred is known as conditional probability It is depicted by: P(A|B)

Date:
P. No:

Unit - 2

Data analysis :- Data analysis is defined as a process of cleaning, transforming and modeling data to discover useful information for business decision making

Tools :- R, python, java, SQL, SAS

1. Descriptive analysis what happended ?.
2. diagnostic analysis why happended ?.
3. ~~historical data~~ predictive analysis why did it ?
4. prescriptive analysis how to do ?,

1. Descriptive analysis :- It is the conventional form of business intelligence and data analysis It seeks to provide a summary view of facts & figures in an understandable format It prepares data for further analysis It can descripe in detail about an event that has occured in past

2. diagnostic analysis - It is a form of advanced analytics ~~to tools into a~~ which examins the data or content to answer the question why did it happened

Date:
P. No:

In a structured business environment tools for both descriptive & diagnostic analytics go parallel that means tools may be same but for the different purposes

3. Predictive :- It helps us to forcust and trends based on current events predulting the probability of an event happening in future or estimating predicting the probability of an event happening in future or estimating many different but codependent variables are analysed to predict a trend in this type of analytics

4. prescriptive analysis :- It refers to a set of techniques to indicate best course of action It suggests what decisions to take to optimize the outcome the goal of prescriptive analytics is to enable

* cost reductions
* quality improvement
* Increasing productivity