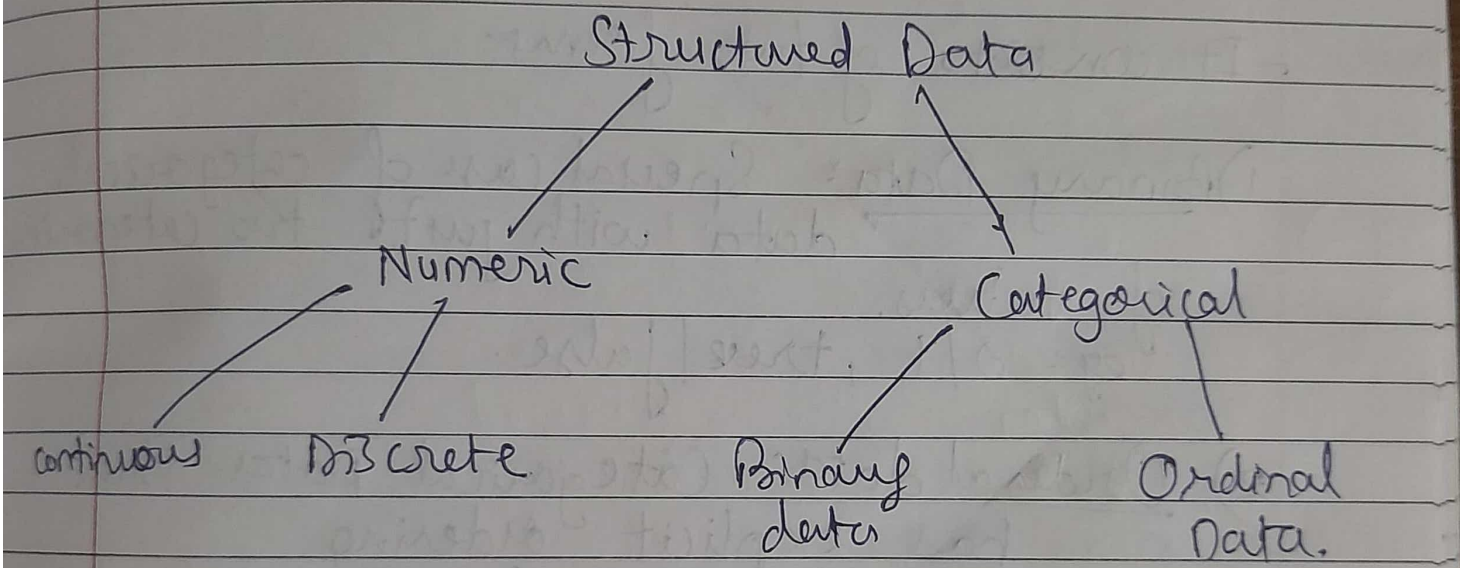
Data Science

Rainbow
PAGE:
DATE: / /

UNIT - 1

Q1) What are the elements of structured data?

→ To apply statistical concepts, unstructured raw data must be converted to structured data.



- There are two basic types of structured data: numerical & categorical.

Numerical Data: Data expressed on numeric scale.

- It consists of two forms:

1) Continuous: Such as wind speed and time duration.
   - Data that can take any value in an interval.
2) Discrete: Such as the count of occurence of an event.
   - Data that can take on only integer values.

Rainbow
PAGE:
DATE: / /

Categorical Data.

- It is the data that can take on only a specific set of values representing a set of possible categories.

eg: Type of TV screen: Plasma, LCD, LED

- It consists of two forms:

1) Binary Data: Special case of categorical data with just two categories of values.
eg: 0/1, true/false.

2) Ordinal data: Categorical data that has explicit ordering.
ex: numerical rating from 1 to 5.

(Q.2) Explain Rectangular Data with its key terms & examples.

→ - A typical frame of reference for analysis in data science is a rectangular data object, like a spreadsheet or database table.

- Rectangular data is the general term for a 2-D matrix with rows indicating records (cases) and columns indicating features (variables).

- Data doesn't always start in this form: unstructured data (eg: text) must be processed and manipulated so that it can be represented as a set of features in the rectangular data.

Key Terms for Rectangular Data.

1) **Data Frame**: Rectangular Data (like a spreadsheet) is the basic structure for statistical and ML models.

2) **Records**: A row within a table is commonly referred to as a record.

3) **Feature**: A column within a table is commonly referred to as a feature (attribute).

4) **Outcome**: - Many data science projects involve predicting an outcome ex: yes/no.
- The features are sometimes used to predict the outcome.

ex:

Category	Currency	Duration	Competitive
Game	US	5	0
Movie	US	5	0
Automotive	US	7	1
Automotive	US	7	1

- This table is a mix of measured data (duration) and categorical data (category, currency)
- The right most column is represented in form of binary data.

- An indicator variable showing whether the auction was competitive or not. This indicator variable also happens to be the outcome variable, where the scenario is to predict whether an auction is competitive or not.

Q.3) Define the various terms of estimates of location.

→

- Variables with measured or count data might have thousands of distinct values.

- A basic step in exploring your data is getting a 'typical value' for each feature (variable): an estimate of where most of the data is located. (i.e. its central tendency)

Key Terms of Estimates of Location

1) Mean: The sum of all values divided by the number of values.

2) Weighted mean: Sum of all values times a weight divided by the sum of weights.

3) Median: The value such that one half of the data lies above or below it.

4) Percentile: The value such that P percent of data lies below.

5) Weighted Median: The value such that one half of the sum of weights lies above & below the sorted data.

6) Trimmed mean: The average of all values after dropping a fixed number of extreme values.

7) Robust: Not sensitive to extreme values

8) Outlier: A data value that is very different from most of data.

Q) What do you mean by Robust estimates? Which estimates of location are robust in nature?

→ - Robust estimates of locan is not influenced by outliers (extreme case) that could skew the results.
- An outlier is any value that is very distant from the other values in the dataset.

- When outliers are the result of bad data the mean will result in poor estimate of location, while the median will still be valid.

- In any case, outliers should be identified and are usually worthy of further investigation

The estimates that are Robust in nature:

1) Median:

- Median is the middle number in sorted list of data.

- Compared to mean, which uses all observations, the median only depends on values in the center of the sorted data.

- While this might be seen as a disadvantage, there are many instances where median is better metric for location.

Put example of Billgates house in one neighbourhood

Rainbow
PAGE:
DATE: / /

2) Weighted Median: As with the median, we first sort the data,
- Instead of a middle value, such that sum of the weights is equal for the lower and upper halves of the sorted list.

- Like median, weighted median is robust to outliers.

3) Trimmed Mean: A trimmed mean is widely used to avoid the influence of outliers.
- For ex: trimming the bottom 10% of the data will provide protection against outliers in all but smallest datasets.

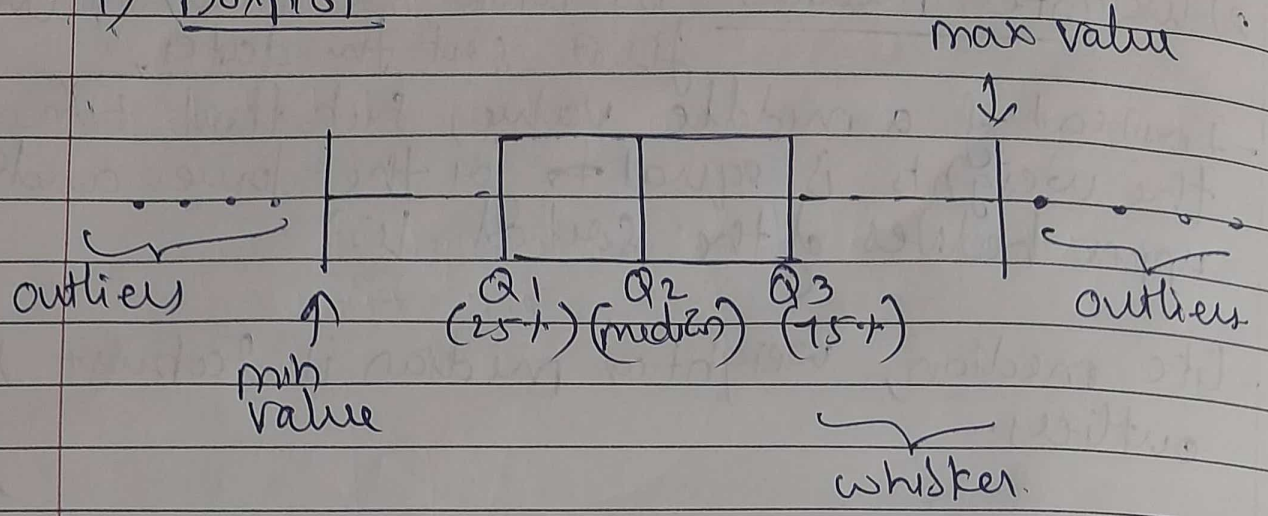
- It can be thought of as a compromise b/w median & mean.

- It is robust to extreme values, but uses more data to calculate estimate of locat^n.

Q4) Key Terms for Exploring the Data Distribution
Q explain: 1) Boxplot 3) Histogram
2) Frequency table 4) Density plot.

→ - Each of the estimates sums up the data into a single number to describe the locam of data. It is also useful to explore how the data is distributed overall.

1) Boxplot



- Introduced by Tukey as a quick way to visualize the distribution of data.
- It is based on percentiles
- It is a type of chart that depicts a group of numerical data through their ~~of~~ quantiles.
- Q1 represents 25th percentile of data
- Q2 represents the median of the data.
- Q3 represents 75th percentile of data.
- The dashed lines are reffered to as whiskers
- Any data outside the whiskers is marked by a single point. (Often considered as outliers)

Rainbow
PAGE:
DATE: / /

2) Frequency Table.

- Frequency table of a variable divides up the variable range into equally spaced segments and tells us how many values fall within each segment.

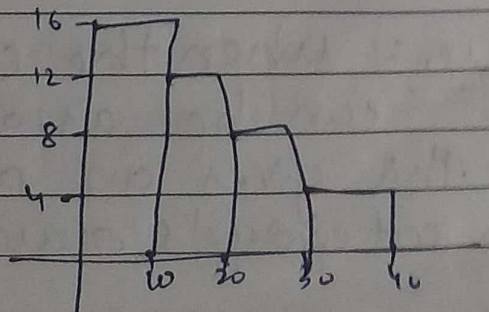
No.	Population Range (m)	Count (States)
1	0-10	12
2	10-20	10
3	20-30	8
4	30-40	6
5	40-50	4

3) Histogram

- Histogram is the way of visualizing a frequency table with bin(range) on x-axis & data count on y-axis.

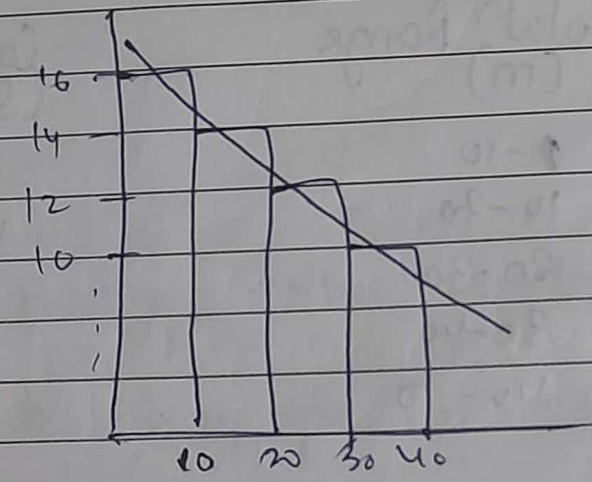
- In general, Histogram are plotted such that

1) Bins are of equal length.
2) Empty bins are included
3) Number of bins depends on user.
4) No empty space b/w bins.



4) Density Plot

- Related to Histogram is a Density Plot.
- Shows Distribution of data values in a continuous line.



- Density plot can be thought of as a Smoothed Histogram.

Q5) Explain Various Terms for binary and Categorical Data.

→ The key terms for explaining Binary & Categorical Data are

1) Mode: The most commonly occuring category or value in a dataset.

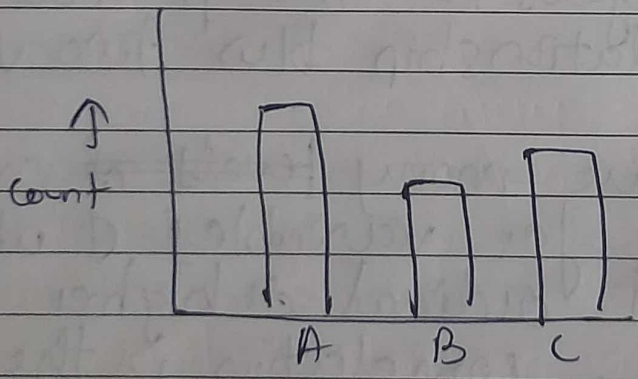
2) Expected Value: When the categories can be associated with numeric values, this gives an average value based on category's occurence.

The expected values is calculated as follows

1) Multiply each outcome by its probability of occurence.
2) Sum these values.

3) Bar Charts.

- Bar charts are a common visual tool for displaying a single categorical variable.
- Categories are listed on x-axis, frequency on y-axis.

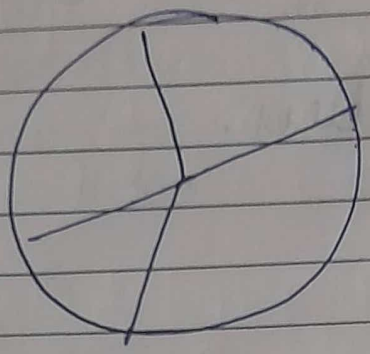


Explain this with an example

4) Pie Charts

- A pie chart is a circular graph that uses slices to represent numerical proportions.
- The arc length of each slice is proportional to the quantity it represents.

- All the slices of the pie add upto 100 percent or 360 degrees.



Explain this very much with an example.

Q)6) Exploration correlation for exploratory analysis. Explain different types of correlation with correlation coefficient

→ - Correlation refers to the process of establishing relationship b/w two variables

- While there are many ~~levels of~~ measures of association for variables which are measured at ordinal or higher level of measurement, correlation is the most commonly used approach.

- Methods of correlation summarize the relationship in a single number called Correlation Coefficient.

Correlation Coefficient

- The correlation coefficient is usually represented using the symbol 'r'

and ranges from -1 to +1.

- A correlation coefficient close to 0, either +ve or -ve, implies that there is little to no relationship b/w two variables.

- Correlation coefficient close to +1 means that positive relationship b/w two variables

- one ↑ses, another ↑ses

- Correlation coefficient close to -1 implies negative relationship b/w two variables.

- One ↑ses, other ↓ses

Correlation Matrix.

- A correlation matrix is a statistical technique that evaluates the relationship b/w two variables in a dataset

- Its a table that displays correlation coeff. of different variables.

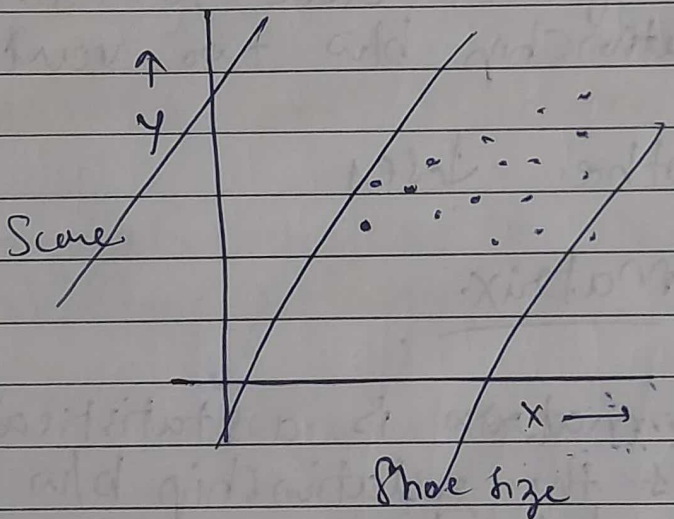
- The variables are shown on both rows & columns & the cell values are the correlation b/w them.

	A	B	C
1	1	0.5	0.5
2	0.5	1	-1
3	0	0	-0.1

Rainbow
PAGE:
DATE: / /

Scatterplot

- A scatterplot is a graph that shows the relationship b/w two numerical variables.
- The graph displays strength, direction & the form of relationship b/w the variables.
- It is created by plotting data points on ~~two~~ 2D plane. Independent variable is plotted on x-axis, dependent on y.



Q.1) Exploring two or more variables
OR
Define: a) Contingency table
b) Hexagonal binning
c) Contour plot
d) Violin plot.

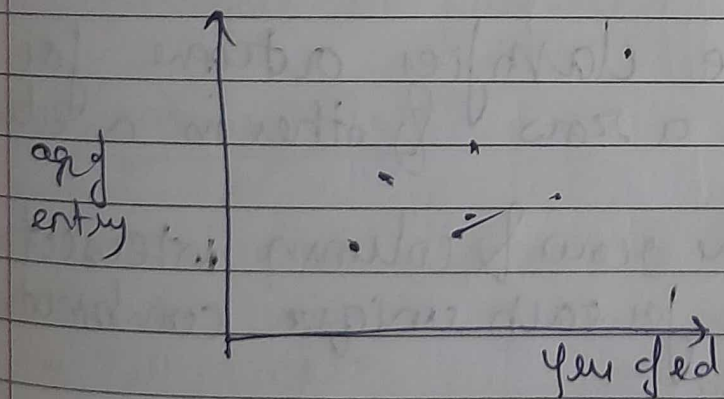
→
- The position of each point on a plot corresponds to the value of two variables for a specific observation or datapoint.

They allow us to identify patterns, trends or outlier in data

- Scatterplots can help identify any potential linear, non-linear or no-relationship b/w variables.

ex: Years of education & age of entry in labour force

No.	Yrs of ed.	Age of entry
1	10	16
2	12	17
3	14	18
4	16	19
5	18	20



(explain this briefly).

Rainbow
PAGE:
DATE: / /

Q7) Exploring two or more variables.

OR

Define: a) Contingency Table
b) Hexagonal Binning
c) Contour Plot
d) Violin Plot.

→ a) Contingency Table:

- A contingency table displays frequencies of combinations of 2 categorical variables
- Analysts also refer to contingency table as cross tabulation or two-way table.
- Contingency table classifies outcome for one variable in a row & other in a column
- And values at the row & column intersections are frequencies for each unique combinations of 2 variables

ex: Relationship b/w gender & type of computer

	PC	Mac
Male	60	40
Female	30	90
Total	90	130

Rainbow
PAGE:
DATE: / /

b) Hexagonal Binning

- It is kind of bivariate histogram useful for visualizing the structure in datasets with large numbers

- The underlying concept is simple

1) The x-y plane over set (x,y) is tesselated by a regular grid of hexagons

2) The number of points falling in each hexagon are stored in a data structure

3) Hexagons with count > 0 are plotted using color ramp or varying radius of the hexagon in proportion to the counts.

c) Contour Plot

- Contour plot (Sometimes called level plot) are a way to represent a 3-D surface into a 2D plane.

- It graphs two predictor variables x,y on y-axis and response variable z as contour

- These contours are sometimes called z slices

- This type of graph is widely used in cartography, where the contour lines indicate the elevations

d) Violin Plot

- Used to visualize distributions of numerical data of different variables
- It is similar to box plot, but gives more info about density on y-axis
- The density is mirrored & flipped over, and the resulting shape is filled in, creating a shape of violin.