

UNIT - 1
Exploratory Data Analysis

SYLLABUS –

Elements of Structured Data, Rectangular Data, Estimates of Location, Estimates of Variability, Exploring the Data Distribution, Exploring Binary and Categorical Data, Correlation, Exploring two and more Variables.

What are the Elements of Structured Data?

- Data comes from many sources: sensor measurements, events, text, images, and videos. The Internet of Things (IoT) is spewing out streams of information.
- Much of this data is unstructured: images are a collection of pixels, with each pixel containing RGB (red, green, blue) color information.
- Texts are sequences of words and nonword characters, often organized by sections, subsections, and so on.
- Clickstreams are sequences of actions by a user interacting with an app or a web page. In fact, a major challenge of data science is to harness this torrent of raw data into actionable information.
- The unstructured raw data must be processed and manipulated into a structured form. One of the commonest forms of structured data is a table with rows and columns as data might emerge from a relational database or be collected for a study.
- There are two basic types of structured data: **numeric and categorical**.
 - i. Numeric data comes in two forms: continuous, such as wind speed or time duration, and discrete, such as the count of the occurrence of an event.
 - ii. Categorical data takes only a fixed set of values, such as a type of TV screen (plasma, LCD, LED, etc.) or a state name (Alabama, Alaska, etc.). Binary data is an important special case of categorical data that takes on only one of two values, such as 0/1, yes/no, or true/false. Another useful type of categorical data is ordinal data in which the categories are ordered; an example of this is a numerical rating (1, 2, 3, 4, or 5).
- For the purposes of data analysis and predictive modelling, the data type is important to help determine the type of visual display, data analysis, or statistical model.
- In fact, data science software, such as R and Python, uses these data types to improve computational performance. More important, the data type for a variable determines how software will handle computations for that variable.
- After all, categories are merely a collection of text (or numeric) values, and the underlying database automatically handles the internal representation. However, explicit identification of data as categorical, as distinct from text, does offer some advantages:
 - i. Knowing that data is categorical can act as signal telling software how statistical procedures, such as producing a chart or fitting a model, should behave.
 - ii. Storage and indexing can be optimized (as in a relational database).

- iii. The possible values a given categorical variable can take are enforced in the software (like an enum).
- Data is typically classified in software by type.
- Data types include num
- eric (continuous, discrete) and categorical (binary, ordinal).
- Data typing in software acts as a signal to the software on how to process the data.

Data Types

Numeric

Data that are expressed on a numeric scale.

Continuous

Data that can take on any value in an interval. (*Synonyms*: interval, float, numeric)

Discrete

Data that can take on only integer values, such as counts. (*Synonyms*: integer, count)

Categorical

Data that can take on only a specific set of values representing a set of possible categories. (*Synonyms*: enums, enumerated, factors, nominal)

Binary

A special case of categorical data with just two categories of values, e.g., 0/1, true/false. (*Synonyms*: dichotomous, logical, indicator, boolean)

Ordinal

Categorical data that has an explicit ordering. (*Synonym*: ordered factor)

What is Rectangular Data?

- The typical frame of reference for an analysis in data science is a rectangular data object, like a **spreadsheet or database table**.
- Rectangular data is the general term for a **two-dimensional matrix** with **rows** indicating records (cases) and **columns** indicating features (variables); data frame is the specific format in R and Python.
- The data doesn't always start in this form: unstructured data (e.g., text) must be processed and manipulated so that it can be represented as a set of features in the rectangular data.
- Data in relational databases must be extracted and put into a single table for most data analysis and modelling tasks.
- In Table 1-1, there is a mix of measured or counted data (e.g., duration and price) and categorical data (e.g., category and currency).
- As mentioned, a special form of categorical variable is a binary (yes/no or 0/1) variable, seen in the rightmost column in Table 1-1— an indicator variable showing whether an auction was competitive (had multiple bidders) or not.

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

Key Terms for Rectangular Data

Data frame

Rectangular data (like a spreadsheet) is the basic data structure for statistical and machine learning models.

Feature

A column within a table is commonly referred to as a *feature*.

Synonyms

attribute, input, predictor, variable

Outcome

Many data science projects involve predicting an *outcome*—often a yes/no outcome (in Table 1-1, it is “auction was competitive or not”). The *features* are sometimes used to predict the *outcome* in an experiment or a study.

Synonyms

dependent variable, response, target, output

Records

A row within a table is commonly referred to as a *record*.

Synonyms

case, example, instance, observation, pattern, sample

Category	currency	sellerRating	Duration	endDay	ClosePrice	OpenPrice	Competitive?
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	1
Automotive	US	3115	7	Tue	0.01	0.01	1

Table-1-1. A typical data frame format

What is Data Frames and Indexes?

- Traditional database tables have one or more columns designated as an index, essentially a row number.
- This can vastly improve the efficiency of certain database queries.
- Index is a basic object that stores axis labels for all pandas objects.
- DataFrame is a two-dimensional data structure, immutable, heterogeneous tabular data structure with labeled axis rows, and columns. pandas DataFrame consists of three components principal, data, rows, and columns. In DataFrame the row labels are called index.

- In Python, with the pandas library, the basic rectangular data structure is a DataFrame object.
- By default, an automatic integer index is created for a DataFrame based on the order of the rows.
- In pandas, it is also possible to set multilevel/hierarchical indexes to improve the efficiency of certain operations.
- In R, the basic rectangular data structure is a data.frame object. A data.frame also has an implicit integer index based on the row order.
- The native R data.frame does not support user-specified or multilevel indexes, though a custom key can be created through the row.names attribute.
- To overcome this deficiency, two new packages are gaining widespread use: data.table and dplyr. Both support multilevel indexes and offer significant speedups in working with a data.frame.
- Example

Create DataFrame with Index

Create pandas DataFrame from List

```
import pandas as pd
technologies = [ ["Spark",20000, "30days"],
                  ["pandas",20000, "40days"],
                  ]
df=pd.DataFrame(technologies)
```

Add Column & Row Labels to the DataFrame

```
column_names=["Courses","Fee","Duration"]
row_label=["a","b"]
df=pd.DataFrame(technologies,columns=column_names,index=row_label)
print(df)
```

Output -

	Courses	Fee	Duration
a	Spark	20000	30days
b	pandas	20000	40days

Explain about Nonrectangular Data Structures?

- There are other data structures besides rectangular data. Time series data records successive measurements of the same variable.
- It is the raw material for statistical forecasting methods, and it is also a key component of the data produced by devices—the Internet of Things.
- Spatial data structures, which are used in mapping and location analytics, are more complex and varied than rectangular data structures.
- Non-Rectangular data does not have a defined structure. Json, XML, Time series records, spatial data, graph data are few of the examples. There are multiple techniques involved in processing non-rectangular data and make it 'model ready', the

techniques ranges from conversion to rectangular format to signal processing on raw data based on the problem statement.

- In the object representation, the focus of the data is an object (e.g., a house) and its spatial coordinates. The field view, by contrast, focuses on small units of space and the value of a relevant metric (pixel brightness, for example).
- Graph (or network) data structures are used to represent physical, social, and abstract relationships.
- For example, a graph of a social network, such as Facebook or LinkedIn, may represent connections between people on the network.
- Distribution hubs connected by roads are an example of a physical network. Graph structures are useful for certain types of problems, such as network optimization and recommender systems.
- Each of these data types has its specialized methodology in data science. The focus of this book is on rectangular data, the fundamental building block of predictive modeling.
- The basic data structure in data science is a rectangular matrix in which rows are records and columns are variables (features).

Define the various terms of Estimates of Location or Measures of Location?

Ans –

- 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).
- A fundamental task in many statistical analyses is to estimate a location parameter for the distribution; i.e., to find a typical or central value that best describes the data.
- There is an immensity of variables when it comes to data, and variables with measured or count data might have millions of values, and that is the reason why we need to locate most of our data and explore it with a "central value".
- Measures of location describe the central tendency of the data. They include the mean, median and mode.
- Various Terms for Estimates of Location

Mean- The sum of all values divided by the number of values.

Synonym- average

Weighted mean- The sum of all values times a weight divided by the sum of the weights.

Synonym weighted average

Median- The value such that one-half of the data lies above and below.

Synonym- 50th percentile

Percentile- The value such that P percent of the data lies below.

Synonym- quantile

Weighted median- The value such that one-half of the sum of the weights lies above and below the sorted data.

Trimmed mean- The average of all values after dropping a fixed number of extreme values.

Synonym truncated mean

Robust- Not sensitive to extreme values.

Synonym- resistant

Outlier- A data value that is very different from most of the data.

Synonym- extreme value

Mean

The most basic estimate of location is the mean, or average value. The mean is the sum of all values divided by the number of values. Consider the following set of numbers: {3 5 1 2}. The mean is $(3 + 5 + 1 + 2) / 4 = 11 / 4 = 2.75$. You will encounter the symbol \bar{x} (pronounced "x-bar") being used to represent the mean of a sample from a population. The formula to compute the mean for a set of n values x_1, x_2, \dots, x_n is:

$$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

N (or n) refers to the total number of records or observations.

Trimmed Mean

A variation of the mean is a trimmed mean, which you calculate by dropping a fixed number of sorted values at each end and then taking an average of the remaining values. Representing the sorted values by x_1, x_2, \dots, x_n where x_1 is the smallest value and x_n the largest, the formula to compute the trimmed mean with p smallest and largest values omitted is:

$$\text{Trimmed mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}$$

Weighted Mean

Weighted mean, which you calculate by multiplying each data value x_i by a user-specified weight w_i and dividing their sum by the sum of the weights. The formula for a weighted mean is:

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

There are two main motivations for using a weighted mean:

- Some values are intrinsically more variable than others, and highly variable observations are given a lower weight. For example, if we are taking the average from multiple sensors and one of the sensors is less accurate, then we might downweight the data from that sensor.

- The data collected does not equally represent the different groups that we are interested in measuring. For example, because of the way an online experiment was conducted, we may not have a set of data that accurately reflects all groups in the user base. To correct that, we can give a higher weight to the values from the groups that were underrepresented.

Median

- The median is defined as the middle point of the ordered data. It is estimated by first ordering the data from smallest to largest, and then counting upwards for half the observations.
- The estimate of the median is either the observation at the centre of the ordering in the case of an odd number of observations, or the simple average of the middle two observations if the total number of observations is even.
- More specifically, if there are an odd number of observations, it is the $[(n+1)/2]$ th observation, and if there are an even number of observations, it is the average of the $[n/2]$ th and the $[(n/2) + 1]$ th observations.

Outliers

- The median is referred to as a robust estimate of location since it is not influenced by outliers (extreme cases) that could skew the results.
- An outlier is any value that is very distant from the other values in a data set. The exact definition of an outlier is somewhat subjective, although certain conventions are used in various data summaries and plots.
- When outliers are the result of bad data, the mean will result in a poor estimate of location, while the median will still be valid.

The basic metric for location is the mean, but it can be sensitive to extreme values (outlier). Other metrics (median, trimmed mean) are less sensitive to outliers and unusual distributions and hence are more robust.

How Variability Metrics can be used to measures whether the data values are tightly clustered or spread out?

Or

Explain Measures of Dispersion or Variability?

Ans –

- Location is just one dimension in summarizing a feature. A second dimension, variability, also referred to as dispersion, measures whether the data values are tightly clustered or spread out.
- At the heart of statistics lies variability: measuring it, reducing it, distinguishing random from real variability, identifying the various sources of real variability, and making decisions in the presence of it.
- Various Terms for Variability Metrics are

Deviations- The difference between the observed values and the estimate of location.
Synonyms errors, residuals

Variance- The sum of squared deviations from the mean divided by $n - 1$ where n is the number of data values.
Synonym mean-squared-error

Standard deviation- The square root of the variance.

Mean absolute deviation- The mean of the absolute values of the deviations from the mean.

Median absolute deviation from the median- The median of the absolute values of the deviations from the median.

Range- The difference between the largest and the smallest value in a data set.

Order statistics- Metrics based on the data values sorted from smallest to biggest.
Synonym ranks

Percentile- The value such that P percent of the values take on this value or less and $(100-P)$ percent take on this value or more.
Synonym quantile

Interquartile range- The difference between the 75th percentile and the 25th percentile.
Synonym IQR

- Dispersion of data used to understand the distribution of data.
- It helps to understand the variation of data and provides a piece of information about the distribution data.
- Range, IQR, Variance, and Standard Deviation are the methods used to understand the distribution data.
- *Dispersion of data helps to identify outliers in a given dataset.*

Explain Exploring the Data Distribution?

Or

Explain Boxplot, Frequency Table, Histogram and Density Plot.

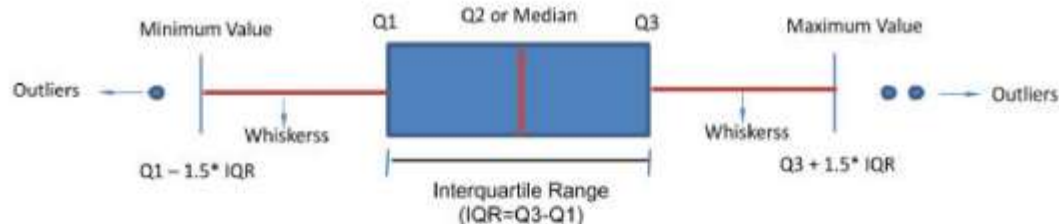
Ans –

Boxplot

- A single box which gives you a visual idea about 5 components in a dataset. It is also known as box and whiskers plot or simply box plot.
- It is useful for describing measures of central tendencies and measures of dispersion in a dataset.
- Boxplots can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped and if and how your data is skewed.

Box Plot represents the following points in a dataset.

1. Minimum Value ($Q1 - 1.5 \times IQR$)
2. First Quartile ($Q1$ or 25th Percentile)
3. Second Quartile ($Q2$ or 50th Percentile)
4. Third Quartile ($Q3$ or 75th Percentile)
5. Maximum Value ($Q3 + 1.5 \times IQR$)



Along with the above 5 components, Boxplot also gives us below information:

- **Outliers:** Points lying beyond the minimum and maximum values are outliers
- **Interquartile range:** It is $Q3 - Q1$. It is the spread or range of the middle 50% of the data.
- **Whiskers:** From Minimum Value to $Q1$ is the first 25% of data

From $Q3$ to Maximum value is the last 25% of the data

Example –

Suppose you have a dataset of runs scored by a batsman in his 12 matches. You arrange the dataset in descending order. Divide the dataset into 4 equal parts. Now how to find out 3 percentiles? 25th percentile, 50th percentile and 75th percentile of the boxplot. Percentile is the number below which a given percentage falls or you can also apply the below formula to find out the percentile.

Position of the number, for given percentile

$$(P_n) = \text{Percentile } (N+1)/100$$

Where N = No. of items in the dataset

Priyadarshini Bhagwati College of Engineering, Nagpur
Department of Computer Science & Engineering
B. TECH 6th Semester - CSE **Subject – Elective III- Data Science**
Subject Notes By – Prof. D.V. Jamthe

- If the above result comes in float (in decimals) then, take the mean of 2 numbers of P_n and $P_{(n+1)}$.
- If the result comes in integer, then take the value of P_n

In our example $N = 12$

Position number for 25th percentile = $25(12+1)/100 = 3.25$

The result is a decimal number; we will take mean of 3rd and 4th number

Position number for 50th percentile = $50(12+1)/100 = 6.5$

The result is a decimal number; we will take mean of 6th and 7th number

Position number for 75th percentile = $75(12+1)/100 = 9.75$

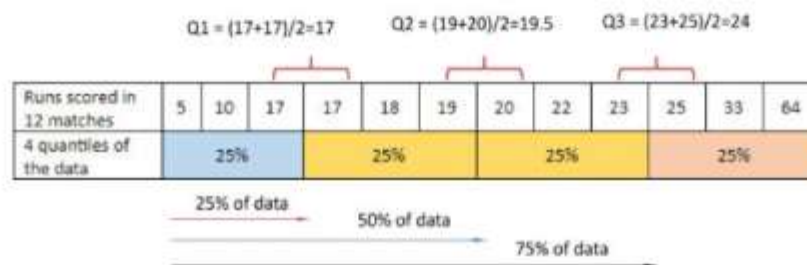
The result is a decimal number; we will take mean of 9th and 10th number

Let us visualize it:

Q1 or 25th percentile = Number below which 25% is falling

Q2 or 50th percentile = Number below which 50% is falling

Q3 or 75th percentile = Number below which 75% is falling



Q1=17

Q2=19.5

Q3=24

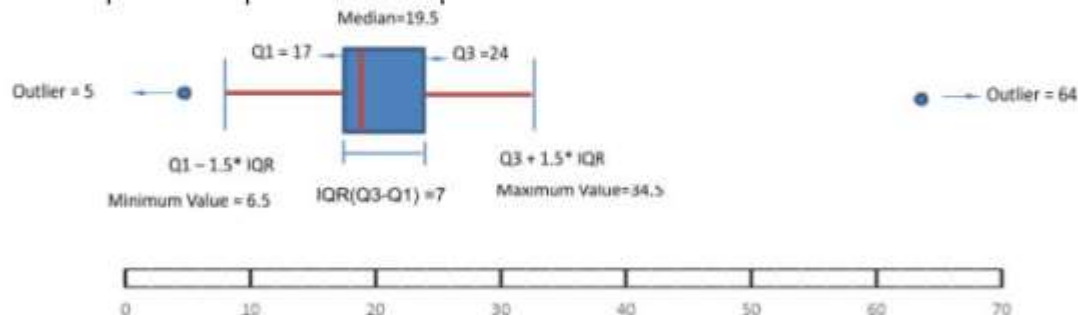
Interquartile range (IQR) = $Q3 - Q1 = 24 - 17 = 7$

Minimum Value = $Q1 - 1.5 * IQR = 17 - 1.5 * 7 = 6.5$

Maximum Value = $Q3 + 1.5 * IQR = 24 + 1.5 * 7 = 34.5$

Outliers of the dataset = 5 & 64

Let us impose all the points of the boxplot on a number line:



- Boxplot also tells us about the distribution and symmetry of the data. The above example shows the data is right skewed. Interquartile range shows us that the middle 50% of the data lies between 17 runs to 24 runs. Whiskers of the box plot cover approximately 99.65% of the data.

WHEN TO USE A BOXPLOT

- For some distributions/data sets, you will find that you need more information than the measures of central tendency (median, mean and mode).
- You need to have information on the variability or dispersion of the data. A boxplot is a graph that gives you a good indication of how the values in the data are spread out.
- Although boxplots may seem primitive in comparison to a histogram or density plot, they have the advantage of taking up less space, which is useful when comparing distributions between many groups or data sets.

Frequency Table

- **Frequency** is visual display which is used to show data. Each type of visual tool has advantages and the best type of plot or graph depends on the situation.
- Indeed, sometimes it is a matter of preference as many different graphs could be used to illustrate the same data.
- A **frequency table** of a variable divides up the variable range into equally spaced segments and tells us how many values fall within each segment.
- **Frequency** is a measure of how often something occurs. A **frequency table** is used to measure and visually show how often a data value occurs. A **frequency table** shows the number of times a value occurs in each category or interval.

For an example; a teacher is preparing for parent conferences. In order to provide parents with the most information possible about their children, he wants to organize the grades of the class so that they can compare the grades to the rest of the class.

The math percentages have been calculated and his students earned the following grades:

88, 86, 92, 65, 72, 75, 81, 84, 85, 93, 99, 50, 78, 80, 86, 76, 74, 95, 81, 87, 90, 72, 76, 61, 85, 84, 78, 83.

Grades are determined by percent where

0-59% is an F,
 60-69% is a D,
 70-79% is a C,
 80-89% is a B, and
 90-100% is an A.

These values make the most logical intervals. **Intervals** are always chosen depending on the range of the data. He will make a frequency table to illustrate the information. First, for each student who scored in the given range, he puts an X. Sometimes, frequency tables use X's and other times, they can use lines for tally marks.

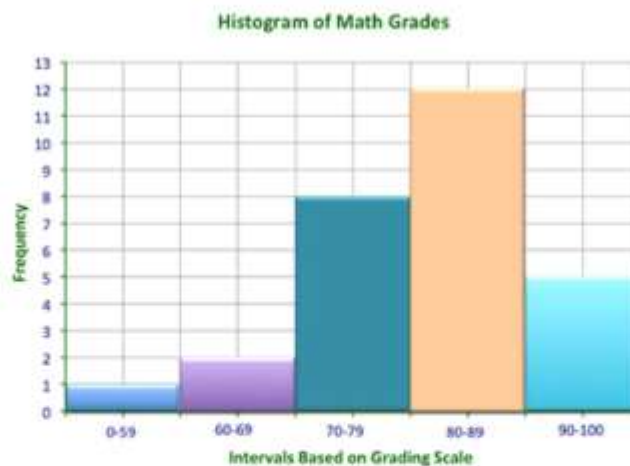
Interval	Tally	Frequency
90 - 100	XXXXX	5
80 - 89	XXXXXXXXXXXX	12
70 - 79	XXXXXXXX	8
60 - 69	XX	2
0 - 59	X	1

Frequency Tables of Grades

- This tally is useful in the sense that it communicates to parents how many students in the class scored in the A range, B range, etc.
- It would not be as important for the parents to see the individual scores of each student as it would be to see the total number of students in each interval.
- That way, if their child earned a B, then they would know that the child falls in a category that most other students scored in.
- If a child earned a D, for example, it would indicate that they are below the general level of the other students and might need additional help.

Histogram

- A histogram is a way to visualize a frequency table, with bins on the x-axis and the data count on the y-axis.
- Histograms are plotted such that:
 - Empty bins are included in the graph.
 - Bins are of equal width.
 - The number of bins (or, equivalently, bin size) is up to the user.
 - Bars are contiguous—no empty space shows between bars, unless there is an empty bin.
- A histogram is the most commonly used graph to show frequency distributions. It is a graph that uses bars to show the number of values in each category or interval. The bars of a histogram always touch.
- A **histogram** is similar to a bar graph in that it uses columns to illustrate data on x- and y-axes.
- In a histogram, you can use the same intervals as you did for the frequency table. The bars in the histogram will have no space between them.



- The histogram shows the same information as the frequency table does.
- However, the histogram is a type of graph, meaning that it is visual representation.
- The bars on the histogram are interpreted more easily by size than numerical data.

Example 1 - Create a histogram of the mass of geodes found at a volcanic site. Scientists measured 24 geodes in kilograms and got the following data:

0.8, 0.9, 1.1, 1.1, 1.2, 1.5, 1.5, 1.6, 1.7, 1.7, 1.7, 1.9, 2.0, 2.3, 5.3, 6.8, 7.5, 9.6, 10.5, 11.2, 12.0, 17.6, 23.9, and 26.8.

Solution - First, let's think about intervals.

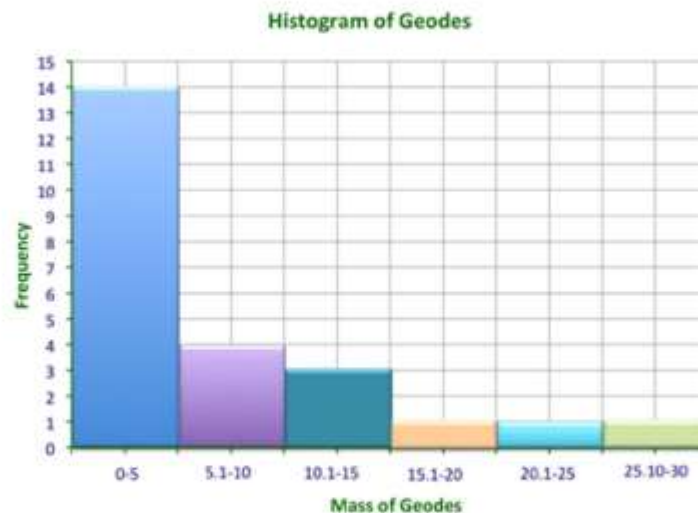
The minimum item is 0.8 kg and the maximum is 26.8. To get a good idea of the data, you could use intervals that encompass perhaps 4 kg intervals, 5 kg intervals, or 6 kg intervals.

Let's try intervals of 5 kg.

Begin with a frequency table.

Interval	Tally	Frequency
0 - 5	 	14
5.1 - 10		4
10.1 - 15		3
15.1 - 20		1
20.1 - 25		1
25.1 - 30		1

Next, create a histogram for this data.



Example 2 -

The coach is comparing the heights of his team, the Markwell Cougars to the rival team, the Sampson Hawks. You need to create a double histogram for the data. The data collected is:

Markwell Cougars:

170, 172, 175, 176, 176, 176, 178, 181, 182, 183, 183, 183, 185, 185, 187, 188, 188, 189, 190, 195

Sampson Hawks:

169, 175, 176, 176, 178, 179, 180, 183, 183, 186, 186, 186, 187, 187, 187, 187, 187, 188, 190, 191, 192

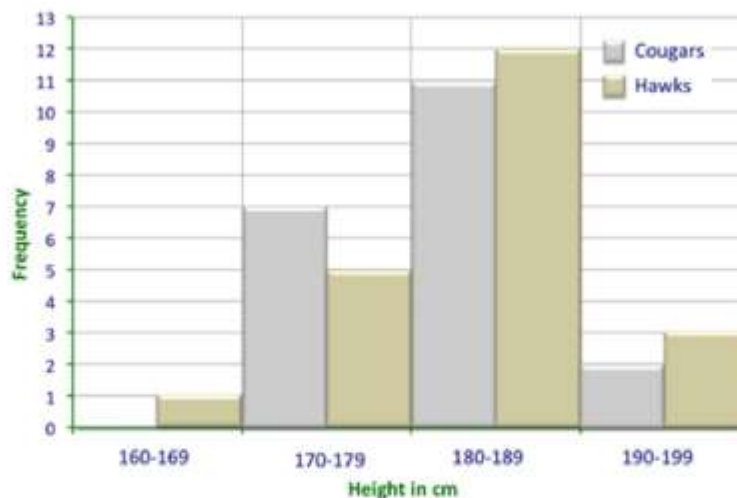
Solution -

First, create a frequency table for the data. Since you are comparing two sets of data, you have to put both data sets into the frequency table.

	Markwell Cougars		Sampson Hawks	
Interval	Tally	Frequency	Tally	Frequency
160 - 169				1
170 - 179		7		5
180 - 189		11		12
190 - 199		2		3

Next, use this data to create a histogram that compares the data.

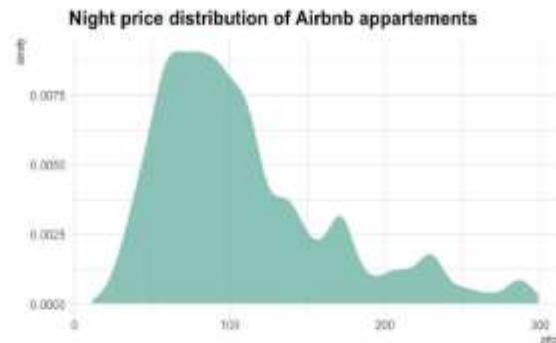
Histogram of Track Heights



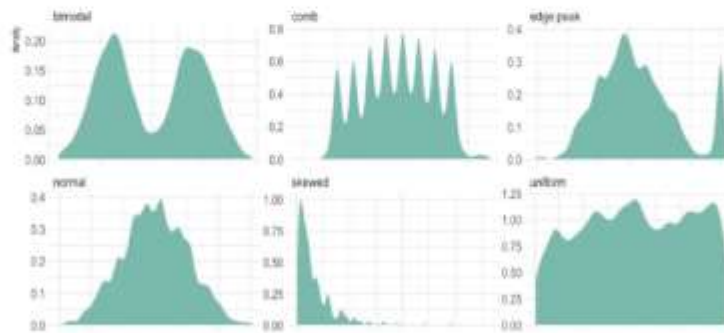
You can see from the histogram that both teams have more players in the 180 - 189 interval. However, while the Cougars have more players in the 170 - 179 interval, the Hawks have slightly more in the taller interval. The Hawks have a slight height advantage.

Density Plot

- A density plot is a smoothed version of a histogram; it requires a function to estimate a plot based on the data.
- Related to the histogram is a density plot, which shows the distribution of data values as a continuous line.
- A density plot can be thought of as a smoothed histogram, although it is typically computed directly from the data through a kernel density estimate.
- The **histogram** is the graphical representation that organizes a group of data points into the specified range. Creating the histogram provides the Visual representation of data distribution. By using a histogram we can represent a large amount of data, and its frequency.
- **Density Plot** is the continuous and smoothed version of the Histogram estimated from the data. It is estimated through Kernel Density Estimation.
- In this method Kernel (continuous curve) is drawn at every individual data point and then all these curves are added together to make a single smoothed density estimation. Histogram fails when we want to compare the data distribution of a single variable over the multiple categories at that time Density Plot is useful for visualizing the data.
- It is a smoothed version of the histogram and is used in the same concept. Here is an example showing the distribution of the night price of Rbnb appartements in the south of France.



- Density plots are used to study the distribution of one or a few variables. Checking the distribution of your variables one by one is probably the first task you should do when you get a new dataset.
- It delivers a good quantity of information. Several distribution shapes exist, here is an illustration of the 6 most common ones:



Explain Exploring Binary and Categorical Data?

Or

Explain various terms of Binary and Categorical data.

Ans –

- For categorical data, simple proportions or percentages tell the story of the data. Key Terms for Exploring Categorical Data are
 - 1) **Mode** - The most commonly occurring category or value in a data set.
 - 2) **Expected value** - When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence.
 - 3) **Bar charts** - The frequency or proportion for each category plotted as bars.
 - 4) **Pie charts** - The frequency or proportion for each category plotted as wedges in a pie.

Mode

- The mode of a distribution with a discrete random variable is the value of the term that occurs the most often.
- It is not uncommon for a distribution with a discrete random variable to have more than one mode, especially if there are not many terms.
- This happens when two or more terms occur with equal frequency, and more often than any of the others.
- A distribution with two modes is called bimodal. A distribution with three modes is called trimodal.
- The mode of a distribution with a continuous random variable is the maximum value of the function. As with discrete distributions, there may be more than one mode.

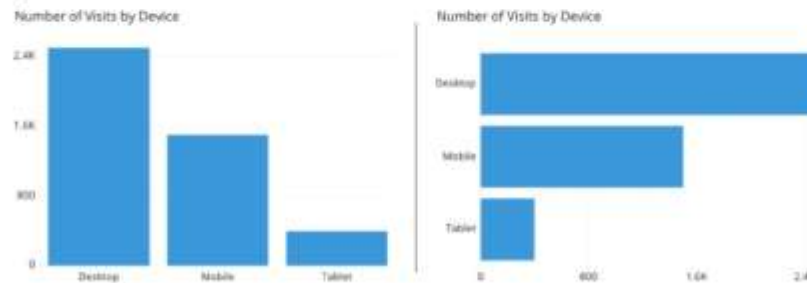
Expected value

- A special type of categorical data is data in which the categories represent or can be mapped to discrete values on the same scale.
- The expected value is calculated as follows:
 1. Multiply each outcome by its probability of occurrence.
 2. Sum these values.
- The expected value is really a form of weighted mean: it adds the ideas of future expectations and probability weights, often based on subjective judgment.
- Expected value is a fundamental concept in business valuation and capital budgeting for example, the expected value of five years of profits from a new acquisition, or the expected cost savings from new patient management software at a clinic.

Bar Graphs or Chart

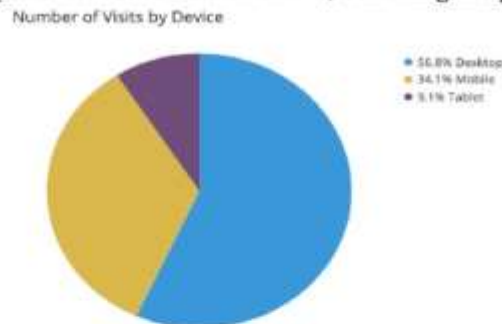
- A bar graph or chart refers to a chart that plots data, quantities, or numeric values using bars.
- These graphs **usually represent categorical** data and consist of two axes. One axis consists of bars representing different categories, while the other axis represents discrete values.
- The number of bars on a bar graph depends on the number of data categories. For example, if there are seven categories, the bar graph will have seven bars.
- The length of the bars demonstrates the numeric values of the category.

- Bar graphs are usually used to compare different variables or show changes in data over time.
- A bar graph uses rectangular bars that are either horizontal or vertical to represent data. The length of each bar relates to the measurement in the data.
- The bars that are used are the same width, but the height of each bar varies depending on the data.
- In that way, it is easy to see which piece of data ranks higher, has more volume, or has more of whatever was measured as well as which sets of data (or bars) have less.
- Example



Pie Chart

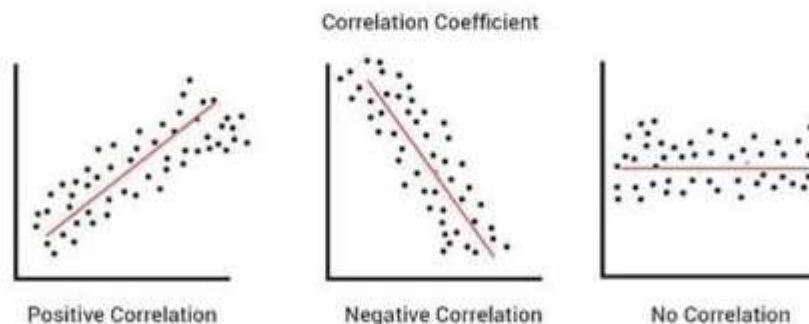
- A pie chart shows how some total amount is divided among distinct categories as a circle (the namesake pie) divided into radial slices.
- Each category is associated with a single slice whose size corresponds with the category's proportion of the total.
- The figure below plots the same data as above, but using the pie chart form instead.



- A pie chart consists of a circular graph divided into different pieces/slices, where **each slice represents a category**. The size of each slice demonstrates the proportion of the whole each category signifies.
- These charts essentially depict how a total amount is divided between different variables or categories.
- Pie charts are a great choice when you want to plot data to show the percentages of a whole.
- When creating a pie chart, it's essential to consider the order of the slices to ensure viewers can understand it quickly. It's best to arrange slices from largest to smallest. However, if the variables have a specific ordering, you should follow that.

Explain Correlation for Exploratory Analysis?

- Correlation refers to the statistical relationship between the two entities. It measures the extent to which two variables are linearly related.
- A correlation coefficient is a descriptive statistic that summarizes the data and helps you compare results between sample data. It is unit-free, which means that you can compare the coefficients directly.
- For example, the height and weight of a person are related, and taller people tend to be heavier than shorter people.
- Exploratory data analysis in many modelling projects (whether in data science or in research) involves examining correlation among predictors, and between predictors and a target variable.
- Variables X and Y (each with measured data) are said to be positively correlated if high values of X go with high values of Y, and low values of X go with low values of Y.
- If high values of X go with low values of Y, and vice versa, the variables are negatively correlated.
- Various key terms for Correlation are
 - i. **Correlation coefficient** - A metric that measures the extent to which numeric variables are associated with one another (ranges from -1 to +1).
 - ii. **Correlation matrix** - A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables.
 - iii. **Scatterplot** - A plot in which the x-axis is the value of one variable, and the y-axis the value of another.
- You can apply correlation to a variety of data sets. In some cases, you may be able to predict how things will relate, while in others, the relation will come as a complete surprise. It's important to remember that just because something is correlated doesn't mean its causal.
- There are three types of correlation:
 1. **Positive Correlation:** A positive correlation means that this linear relationship is positive, and the two variables increase or decrease in the same direction.
 2. **Negative Correlation:** A negative correlation is just the opposite. The relationship line has a negative slope, and the variables change in opposite directions, i.e., one variable decreases while the other increases.
 3. **No Correlation:** No correlation simply means that the variables behave very differently and thus, have no linear relationship.



Correlation coefficients give you the measure of the strength of the linear relationship between two variables.

The letter r denotes the value, and it ranges between -1 and $+1$

If $r < 0$, it implies negative correlation

If $r > 0$, it implies positive correlation

If $r = 0$, it implies no correlation

Calculating the correlation coefficient takes time; therefore, data is entered into a calculator, computer, or statistics program to calculate the correlation coefficient.

Types of Correlation Coefficient

There are mainly two types of correlation coefficients.

Pearson's Product Moment Correlation

- The Pearson correlation coefficient is defined in statistics as the measurement of the strength of the relationship between two variables and their association. It is denoted by r .
- The correlation coefficient can be calculated by using the below formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = Coefficient of correlation

\bar{x} = Mean of x -variable

\bar{y} = Mean of y -variable

$x_i y_i$ = Samples of variable x, y .

Spearman's Rank Correlation

- Spearman's rank correlation measures the strength and direction of association between two ranked variables.
- It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function.

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

ρ = Spearman rank correlation

d_i = Difference between the ranks of corresponding variables

n = Number of Observations

Define Scatterplot with example?

- The standard way to visualize the relationship between two measured data variables is with a scatterplot. The x-axis represents one variable and the y-axis another, and each point on the graph is a record.
- **Scatter plots** are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a **Cartesian system**.
- The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis. These plots are often called **scatter graphs** or **scatter diagrams**.
- A scatter plot is also called a scatter chart, scattergram, or scatter plot, XY graph. The scatter diagram graphs numerical data pairs, with one variable on each axis, show their relationship.
- Scatter plots are used in either of the following situations.
 - i. When we have paired numerical data.
 - ii. When there are multiple values of the dependent variable for a unique value of an independent variable.
 - iii. In determining the relationship between variables in some scenarios, such as identifying potential root causes of problems, checking whether two products that appears to be related both occur with the exact cause and so on.
- A scatter plot is a diagram where each value in the data set is represented by a dot.

Scatter plot Example

Draw a scatter plot for the given data that shows the number of games played and scores obtained in each instance.

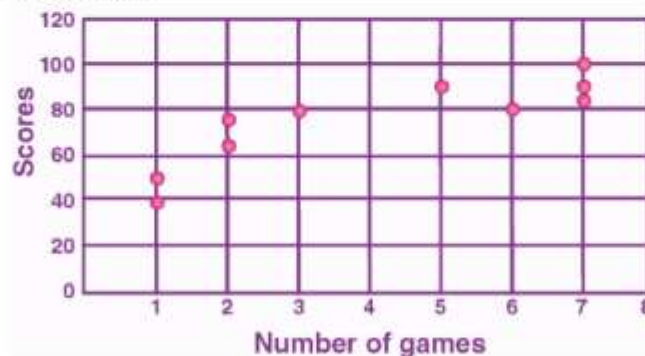
No. of games	3	5	2	6	7	1	2	7	1	7
Scores	80	90	75	80	90	50	65	85	40	100

Solution:

X-axis or horizontal axis: Number of games

Y-axis or vertical axis: Scores

Now, the scatter graph will be:



Explain Exploring Two or More Variables?

Or

Explain Exploring Multivariate Analysis?

Or

Explain following terms in details?


- a. Contingency table
- b. Hexagonal binning
- c. Contour plot
- d. Violin plot

Ans –

- Familiar estimators like mean and variance look at variables one at a time (univariate analysis).
- Correlation analysis is an important method that compares two variables (bivariate analysis) which estimates and plots, and at more than two variables (***multivariate analysis***).

Contingency table (Two Categorical Variables)

- Estimations like mean, median, standard deviation, and variance are very much useful in case of the univariate data analysis. But in the case of bivariate analysis (comparing two variables) correlation comes into play.
- **Contingency Table** is one of the techniques for exploring two or even more variables. It is basically a tally of counts between two or more categorical variables.
- A contingency table is a table that displays data for one variable in rows and data for another variable in columns.
- By arranging the data in rows and columns, the relationship between the two variables can easily be detected by evaluating the table cells where the two data sets overlap. Statistical analysis, such as a chi-square test, can be performed to determine association or dependency.
- The table shows how closely associated the variables are with one another.
- Contingency tables are commonly used in engineering, surveys, and scientific research. Some elements are common to most contingency tables: columns (containing primary data points), sample size, and percentages.



Count		Gender		Total
		Men	Women	
College major	Humanities	4	10	14
	Natural Sciences	11	10	21
	Social Sciences	8	14	22
Total		23	34	57

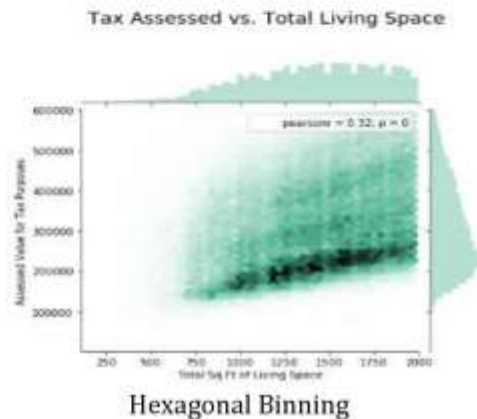
Figure 1: Contingency table example

- In the contingency table in Figure 1, the rows and columns draw associations between college majors and gender. The columns contain gender data, and the rows contain college major data. Further testing can be done on the cells contained within the contingency table to determine association.

Hexagonal binning

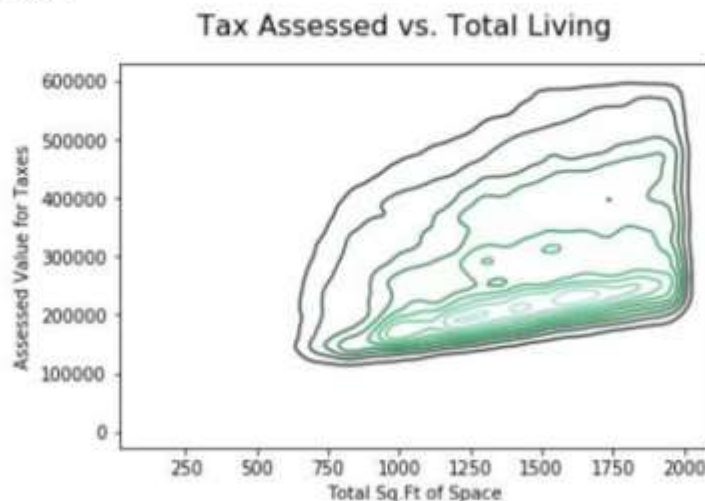
- **Hexagonal binning** is a plot of two numeric variables with the records binned into hexagons.
- The below is a hexagon binning plot of the relationship between the Total Living Space versus the tax-assessed value for homes.
- Rather than plotting points, records are grouped into hexagonal bins and color indicating the number of records in that bin.
- Example;

	TaxAssessedValue	SqFtTotLiving	ZipCode
0	NaN	1730	98117.0
1	206000.0	1870	98002.0
2	303000.0	1530	98166.0
3	361000.0	2000	98108.0
4	459000.0	3150	98108.0



Contour Plot

- A contour plot is a curve along which the function of two variables has a constant value.
- It is a plane section of the three-dimensional graph of the function $f(x, y)$ parallel to the x, y plane.
- A contour line joins points of equal elevation (height) above a given level. A contour map is a map is illustrated in the code below.
- The contour interval of a contour map is the difference in elevation between successive contour lines.



Violin plot

- A violin plot is a hybrid of a box plot and a kernel density plot, which shows peaks in the data.
- It is used to visualize the distribution of numerical data. Unlike a box plot that can only show summary statistics, violin plots depict summary statistics and the density of each variable.
- Violin plots have many of the same summary statistics as box plots:
 - i. the white dot represents the median
 - ii. the thick gray bar in the center represents the interquartile range
 - iii. the thin gray line represents the rest of the distribution, except for points that are determined to be "outliers" using a method that is a function of the interquartile range.
- **Violin Plot** is a method to visualize the distribution of numerical data of different variables.
- It is similar to Box Plot but with a rotated plot on each side, giving more information about the density estimate on the y-axis.
- The density is mirrored and flipped over and the resulting shape is filled in, creating an image resembling a violin.
- The advantage of a violin plot is that it can show nuances in the distribution that aren't perceptible in a boxplot. On the other hand, the boxplot more clearly shows the outliers in the data.
- Violin Plots hold more information than the box plots, they are less popular. Because of their unpopularity, their meaning can be harder to grasp for many readers not familiar with the violin plot representation.

