JAIN DEEMED-TO-BE UNIVERSITY | SCHOOL OF COMPUTER SCIENCE AND IT

School of Computer Science & IT

Programme: BCA

INTRODUCTION TO DATA ANALYTICS (23BCAD4C01)

MODULE 3: Exploratory Data Analysis (EDA)

Dr. Ananta Charan Ojha, Professor

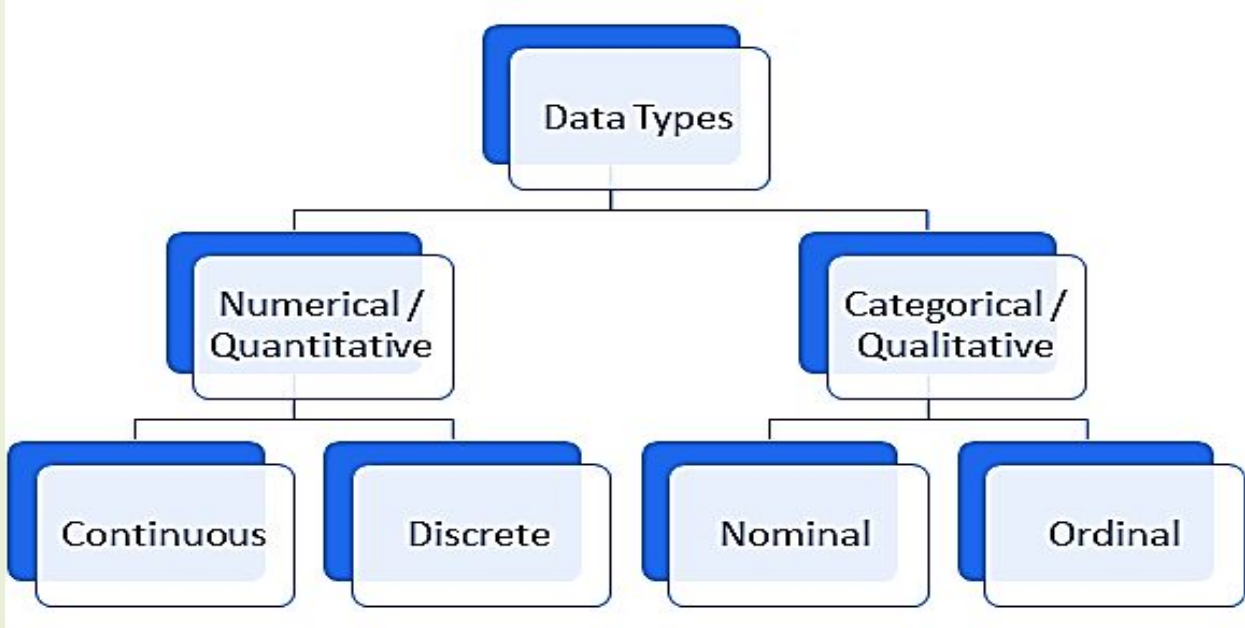
Session -1

- Data Type of Variables
- What is Exploratory Data Analysis?
- Categories of EDA

Why Data Types are important?

- Datatypes are an important concept in statistics because statistical methods can only be used with certain data types.
- You have to analyse continuous data differently than categorical data, otherwise, it would result in a wrong analysis.
- You also need to know which data type you are dealing with to choose the right visualization method. *For example,* Discrete data are represented mainly by bar graphs where as Continuous data are represented in the form of a histogram.
- Therefore, understanding different data types is a **crucial prerequisite** for doing **Exploratory Data Analysis (EDA)** in a data analytics project, since you can use certain statistical measurements only for specific data types.

Types of Data



Qualitative or Categorical Data

- Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers. These types of data are sorted by category, not by number. That's why it is also known as Categorical Data.
- Examples:
  - Gender of a person: male, female, or others.
  - A Likert scale (a psychometric scale) commonly used to gather opinion on something: strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree
  - Colours: red, white, yellow, black etc.

Nominal Data

- Nominal Data is used to label variables without any order or quantitative value. These data don't have any meaningful order; their values are distributed to distinct categories.
- Examples :
  - Hair Colour (White, Red, Brown, Black, etc.) ; Marital status (Single, Widowed, Married) ; Nationality (Indian, German, American); Gender (Male, Female, Others)

Ordinal Data

- Compared to the nominal data, ordinal data have some kind of ordering / ranking by their relative position on the scale. We can assign numbers to ordinal data but don't do any statistical manipulation on them.
- Examples *:*
  - When you ask for feedback, experience, or satisfaction on a scale of 1 to 10
  - Letter grades in the exam (A, B, C, D, etc.)
  - Ranking of peoples in a competition (First, Second, Third, etc.)
  - Economic Status of person (High, Medium, and Low)
  - Education Level (Higher, Secondary, Primary)

Quantitative or Numerical Data

- Quantitative data is countable and can be expressed in numerical values, can be used for statistical manipulation. It answers the questions like, “how much,” “how many,” and “how often.”
- Examples *:*
  - Height of a person; weight of an object; Room Temperature; Scores and Marks (Ex: 59, 80, 60, etc.); Time

Discrete Data

- The discrete data contains values that are integers / whole numbers. The discrete data are countable and have finite values; **their subdivision is not possible**.
- Examples *:*
  - Total numbers of students present in a class; Numbers of employees in a company; The total number of players who participated in a competition; Days in a week

Continuous Data

- Data that can take any value in an interval or range. Continuous data are in the form of floating-point / fractional numbers. Continuous data represents information that can be divided into smaller levels.
- Examples *:*
  - Height of a person; Speed of a vehicle; “Time-taken” to finish the work ; Market share price

- **Identify the Data Type (Nominal, Ordinal, Continuous, Discrete):**

1. CGPA of the student
2. Age of a person
3. Child, Teenager, Young, Old,
4. Boys, Girls, Men, Women
5. Very Good, Good, Bad, Very Bad
6. Rate your knowledge in a subject in a 10 point scale (0-10)
7. Morning, Noon, Afternoon, Evening, Night
8. A set of Odd numbers in 1-100
9. Number of votes in an election
10. Area of a place

What is exploratory data analysis?

- Exploratory data analysis (EDA) is the first step toward building a model. During exploratory data analysis you take a deep dive into the data. The goal isn't to cleanse the data, but you'll **still discover anomalies** you missed before, and you take a step back and fix them.
- Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.
- Exploratory data analysis (EDA) is used by data scientists to perform preliminary analysis and investigate data sets and summarize their main characteristics, often employing summary statistics, and various data visualization methods.
- It can also help determine if the statistical techniques you are considering for data analysis are appropriate.
- It helps determine how best to manipulate data sources to get the answers you need.
- EDA assists Data science professionals in various ways:-
  - Get better understanding/ insights of data
  - Uncover underlaying data patterns
  - Determine relationship among variables
  - Identify important variables
  - Detect outliers and anomalies
  - Preliminary selection of appropriate models

Categories of EDA

- Exploratory data analysis is generally cross-classified in two ways.
  - First, each method is either non-graphical or graphical.
  - And second, each method is either univariate or multivariate.
- Non-graphical methods generally involve calculation of summary statistics, while graphical methods obviously summarize the data in a diagrammatic or pictorial way.
- Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships.
- The four types of EDA are univariate non-graphical, multivariate non-graphical, univariate graphical, and multivariate graphical.

Univariate Non-Graphical EDA

Categorical data

- The characteristics of interest for a categorical variable are simply the range of values and the frequency (or relative frequency) of occurrence for each value.
- Therefore the only useful univariate non-graphical techniques for categorical variables is some form of tabulation of the frequencies, usually along with calculation of the percent of data that falls in each category.
- A simple tabulation of the frequency of each category is the best univariate non-graphical EDA for categorical data. It focuses on summarizing and understanding the distribution of categories within the variable.
- For example, if we consider categories of **specialization** in BCA at JAIN, there is a true population of students in 4th semester. If we take a random sample of 25 students, we could list the sample \ measurement of student specialization: IOT, IOT, DA, CTIS, DA, MACT, ISMA, DA, CTIS, ISMA, IOT, DA, IOT, CTIS, IOT, CTIS, MACT, MACT, ISMA, DA, ISMA, DA, MACT, ISMA, MACT
- Our EDA may look as: (Note that it is useful to have the total count (frequency) to verify that we have an observation for each category. EDA is very helpful for finding mistakes)

Statistic / Specialization	IOT	DA	CTIS	MACT	ISMA	Total
Count	**5**	**6**	**4**	**5**	**5**	**25**
Percent	**20**	**24**	**16**	**20**	**20**	**100%**

Univariate Non-Graphical EDA **for Quantitative Data**

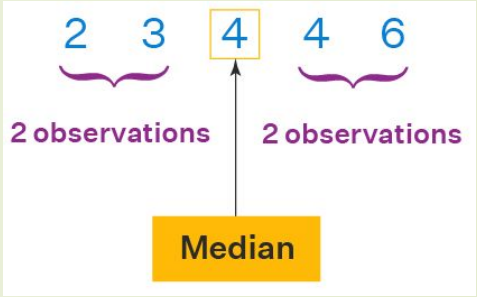
- Univariate EDA for a quantitative variable is a way to make preliminary assessments about the population distribution of the variable using the data of the observed sample.
- The characteristics of the population distribution of a quantitative variable are its center, modality, spread, shape, and outliers.
- For quantitative variables, it is worthwhile looking at the central tendency, spread/ variability, skewness, and kurtosis of the data for a particular variable from an observation.
- But for categorical variables, none of these make any sense.

Estimates of Location / Measure of Central Tendency

- Variables with measured or count data might have thousands of distinct values. It is important to estimate how the data values of a variable are distributed.
- A basic step in exploring the data is getting an estimate of each variable where most of the data is located (i.e., its central tendency).
- Location is one way in summarizing a variable (feature). Distribution of individual data values can also be assessed based on this location.

Mean

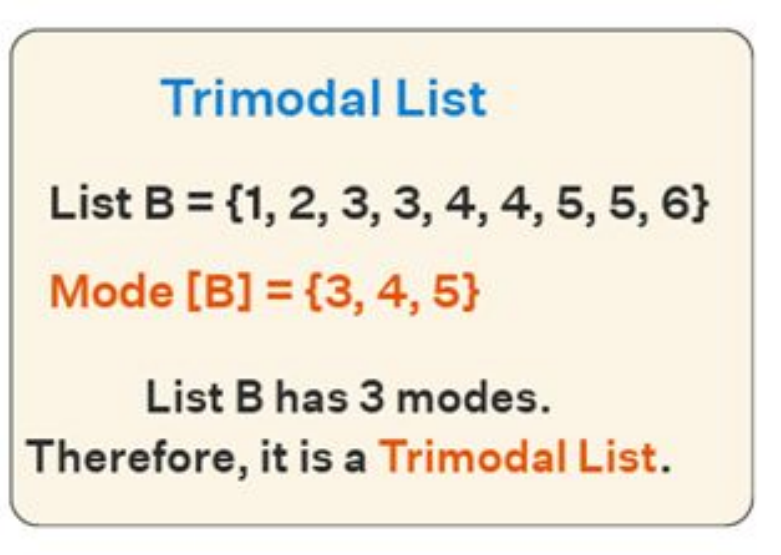
- The most basic estimate of location/ measures of central tendency is the mean, or *average* value. The mean is the sum of all the values divided by the number of values.
- Consider the following set of numbers: {3 5 1 2}. The mean is (3 + 5 + 1 + 2) / 4 = 11 / 4 = 2.75.
- Occasionally other means such as geometric, harmonic, truncated, weighted means are used as measures of centrality.



Median

- The median is another measure of central tendency. The *median* is the middle value on a sorted list of the data.
- If there are an even number of values, take the average of the two middle values.
- Compared to the mean, which uses all observations, the median depends only on the values in the center of the sorted data.
- For symmetric distributions, the mean and the median coincide.
- For unimodal skewed (asymmetric) distributions, the mean is farther in the direction of the pulled out tail of the distribution than the median is. Therefore, for many cases of skewed distributions, the median is preferred as a measure of central tendency. *Example*, scores in a test.
- 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.

- **Mode** (**M**ost **O**ften **D**ata **E**ntered)
  - An occasionally used measure of central tendency is the mode.
  - The value which appears most often in the given data i.e. the observation with the highest frequency is called a **mode** of data.
  - For discrete data, we just need to identify the observation which occurs maximum times.
  - **Mode = Observation with maximum frequency**
  - A data distribution may have 1 mode, or more than 1 mode. Depending upon the number of modes the data has, it can be called unimodal, bimodal, trimodal, or multimodal.



Measure of Central Tendency: Identifying Data Distribution

✔ Helps determine whether the data is symmetrically distributed or skewed:

- Symmetric Distribution → Mean ≈ Median.
- Right-Skewed (Positive Skew) → Mean > Median.
- Left-Skewed (Negative Skew) → Median > Mean.

❖ Dataset 1: Symmetric Distribution (Normal Distribution)

- Exam Scores: 45, 50, 55, 60, 65, 70, 75, 80, 85, 90

Calculations:

- Mean = (45 + 50 + 55 + 60 + 65 + 70 + 75 + 80 + 85 + 90) / 10 = 70
- Median = Middle value = 70
- Mode = No repeating values

Interpretation:

- Since Mean ≈ Median ≈ Mode, the data is symmetrically distributed (**normal distribution**).

❖ Dataset 2: Right-Skewed Distribution (Positive Skew)

Exam Scores: 50, 55, 60, 65, 70, 75, 80, 90, 95, 100

Calculations:

Mean = (50 + 55 + 60 + 65 + 70 + 75 + 80 + 90 + 95 + 100) / 10 = 83

Median = Middle value = 72.5

Mode = No repeating values

Interpretation:

Here, Mean (83) > Median (72.5), indicating a right-skewed (positively skewed) distribution due to the higher values.

Estimates of Variability / Spread / Dispersion

- Another way of summarizing data distribution is *variability*, also referred to as *dispersion*, or *spread* that measures whether the data values are tightly clustered or spread out.
- Variability / Spread describes how far apart data values lie from each other and from the centre of a distribution.
- Just as there are different ways to measure location (mean, median, etc.), there are also different ways to measure variability: variance,

Standard Deviation and Related Estimates

- The most widely used estimates of variability are based on the differences, or *deviations* between the estimate of location (i.e. mean) and the observed data.
- For a set of data {1, 4, 4}, the mean is 3 and the median is 4. The deviations from the mean are the differences: 1 – 3 = –2, 4 – 3 = 1 , 4 – 3 = 1. These deviations tell us how dispersed the data is around the central value.
- If you have n observations, then for any data value, xi, the corresponding deviation is (xi - ẍ), which is the signed (- for lower and + for higher) distance of the data value from the mean of all of the n data values. Note that the sum of all of the deviations of a sample is zero.

- The best-known estimates for variability are the *variance* and the *standard deviation*, which are based on squared deviations.
- The variance is an average of the squared deviations, and the standard deviation is the square root of the variance.
- The standard deviation is much easier to interpret than the variance since it is on the same scale as the original data. The standard deviation is preferred in statistics over the mean absolute deviation.

It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.

$$\text{Variance} = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

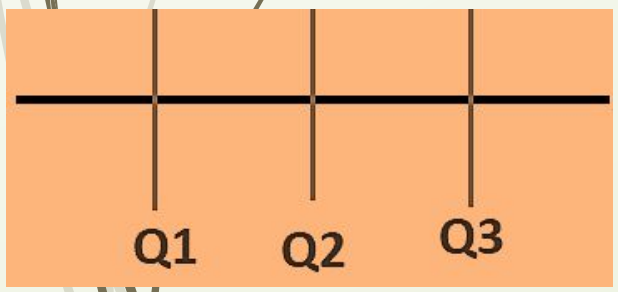
$$\text{Standard deviation} = s = \sqrt{\text{Variance}}$$

Note: The formula for the variance of observed data conventionally has n-1 in the denominator instead of n to achieve the property of “unbiasedness", which roughly means that when calculated for many different random samples from the same population, the average should match the corresponding population quantity.

$$P_x = \frac{x(n+1)}{100}$$

P_x = The value at which x percentage of data lie below that value

n = Total number of observations



Estimates Based on Percentiles / Interquartile Range

- The range is a different approach to estimating dispersion/ variability based on looking at the spread of the sorted data. It is the difference between the minimum and maximum values in an observation. But the range is extremely sensitive to outliers and not very useful as a general measure of dispersion in the data.
- To avoid the sensitivity to outliers, we can look at the range of the data after dropping values from each end. This can be achieved using Percentiles and Interquartile Range.
- **Percentile**: In a data set, the P^{th} percentile is a value such that at least P percent of the data are less than or equal to this value and at least (100 – P) percent of the data are more than equal to this value.

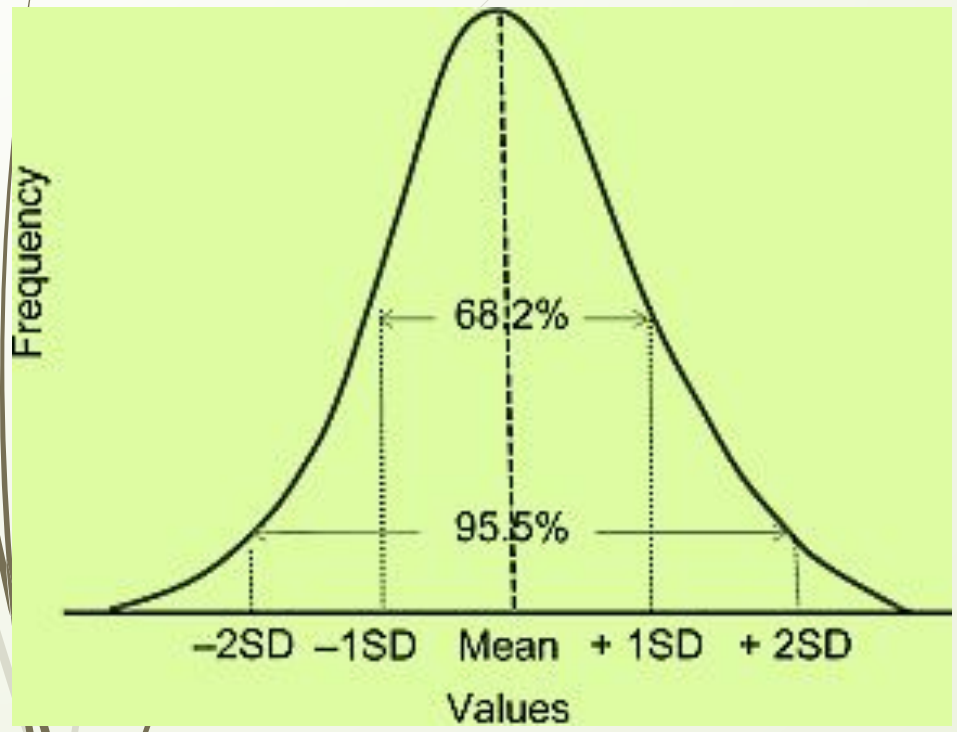
Interquartile Range (IQR)

- The IQR is a more robust measure of spread than the variance or standard deviation.
- The quartiles of a dataset are the three values which divide the observed data into 4 equal parts.
- 25^{th} percentile is 1^{st} quartile (Q_1); 50^{th} percentile is 2^{nd} quartile (Q_2); 75^{th} percentile is 3^{rd} quartile (Q_3).

Interquartile range = Upper Quartile – Lower Quartile = $Q_3 - Q_1$

- The interquartile range estimates half of the values, especially the middle half in a data set.
- **If the data are more spread out, then the IQR tends to increase, and vice versa.**

Normal Distribution



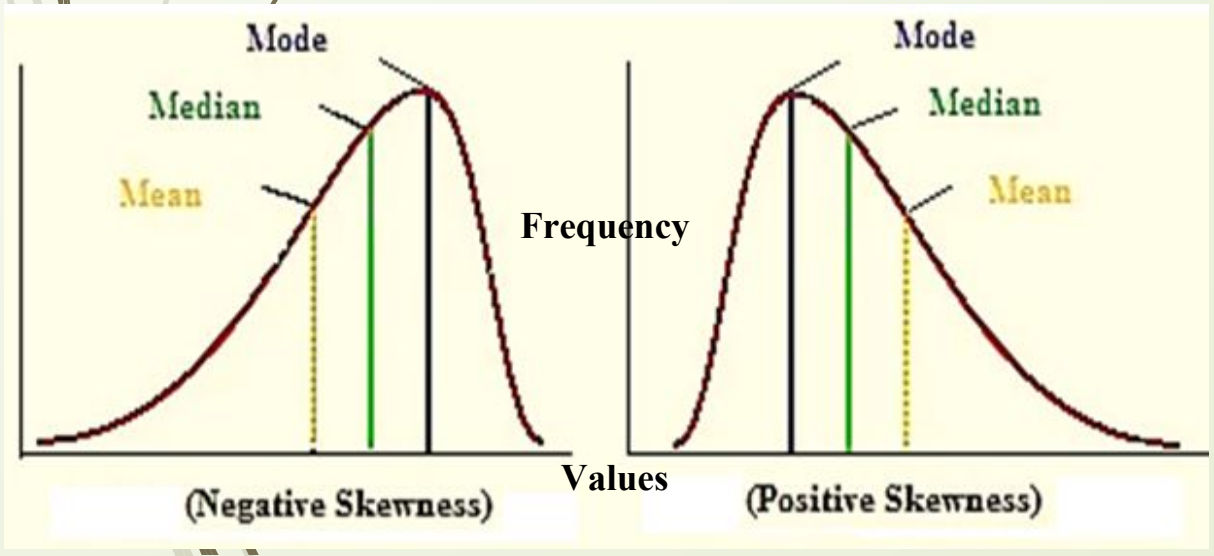
Gaussian distribution (also known as normal distribution) is a bell-shaped curve, and it is assumed that during any measurement values will follow a normal distribution with an equal number of measurements above and below the mean value.

- Mean ± 1 SD contains 68.2% of all values.
- Mean ± 2 SD contains 95.5% of all values.
- Mean ± 3 SD contains 99.7% of all values.

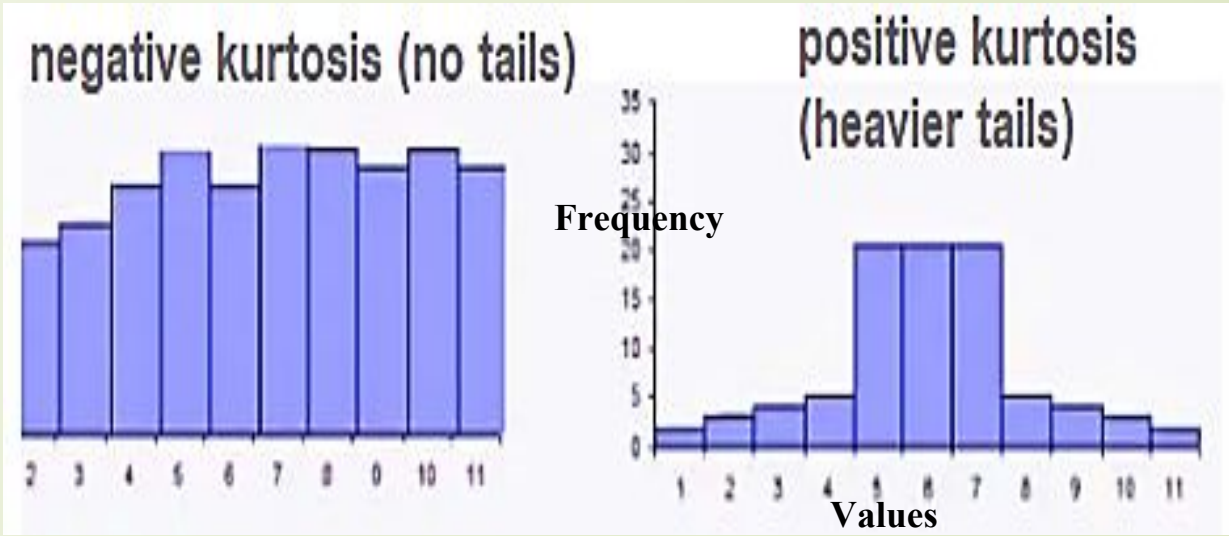
❑ Skewness and kurtosis

- Two additional useful univariate descriptors are the skewness and kurtosis of a distribution. Skewness is a measure of asymmetry. Kurtosis is a measure of “peakedness" relative to a **Gaussian shape**.
- Variance tells us about the amount of variability while skewness gives the direction of variability. Kurtosis gives a measure of flatness of distribution.

- **Positively Skewed**: The data points are more concentrated towards the left-hand side. In this distribution, **Mean > Median > Mode**.
- **Negatively Skewed**: The data points are more concentrated towards the right-hand side of the distribution. In this distribution, **Mean<Median<Mode**.



- A **positive kurtosis** tells you that you have fatter or heavy tails (i.e. a lot of data in your tails), extreme values are likely, presence of outlier is more likely.
- A **negative kurtosis** means that you have thinner or light tails (i.e. little data in your tails), so that extreme values are less likely. Presence of Outliers is less likely.



THANK YOU

Any questions…?

JGI JAIN DEEMED-TO-BE UNIVERSITY | SCHOOL OF COMPUTER SCIENCE AND IT

School of Computer Science & IT

BCA Programme

INTRODUCTION TO DATA ANALYTICS (23BCAD4C01)

MODULE 3: Exploratory Data Analysis (EDA)

Dr. Ananta Charan Ojha, Professor

Session -2

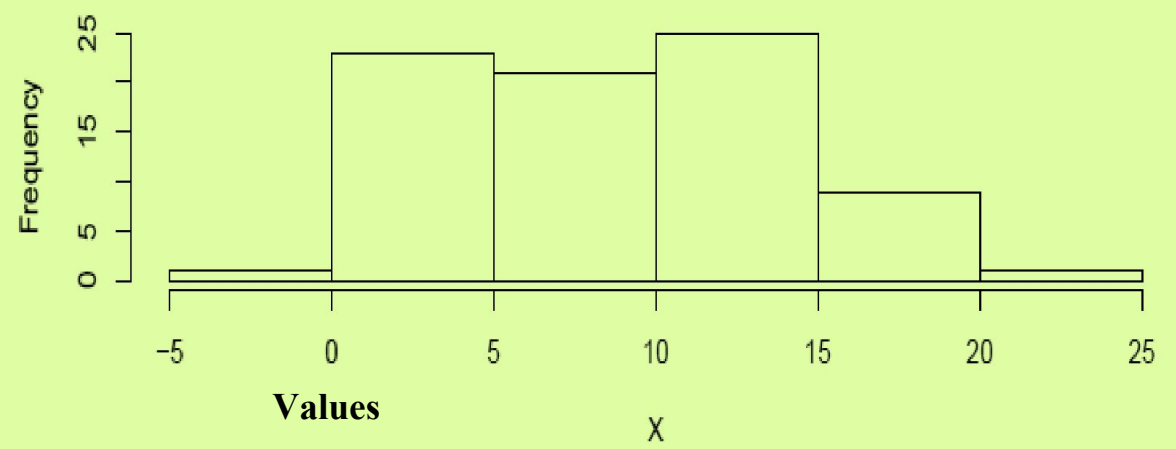
- Categories of EDA
  - Univariate Graphical EDA
  - Multivariate Graphical EDA

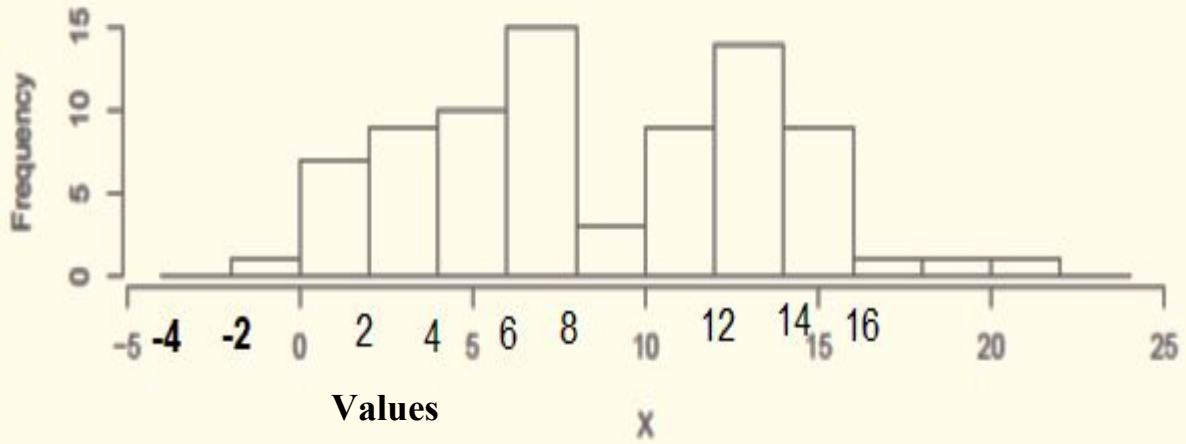
Data Visualization

- Graphical EDA uses data visualization techniques.
- Data visualization is the process of creating graphical representations of the distribution of the sample data.
- This process makes EDA easy for the viewer to interpret and draw conclusions.
- Non-graphical and graphical methods complement each other. While the non-graphical methods are quantitative and objective, they do not give a full picture of the data.
- Therefore, graphical methods, which are more qualitative and involve a degree of subjective analysis, are also required.
- There are many different types of techniques you can leverage to visualize data, and the type of data visualization technique you leverage will vary based on the type of data you're working with.
- Some important popular ones are: Histogram, Box Plots, Scattered Plot, Heat Map

Univariate Graphical EDA

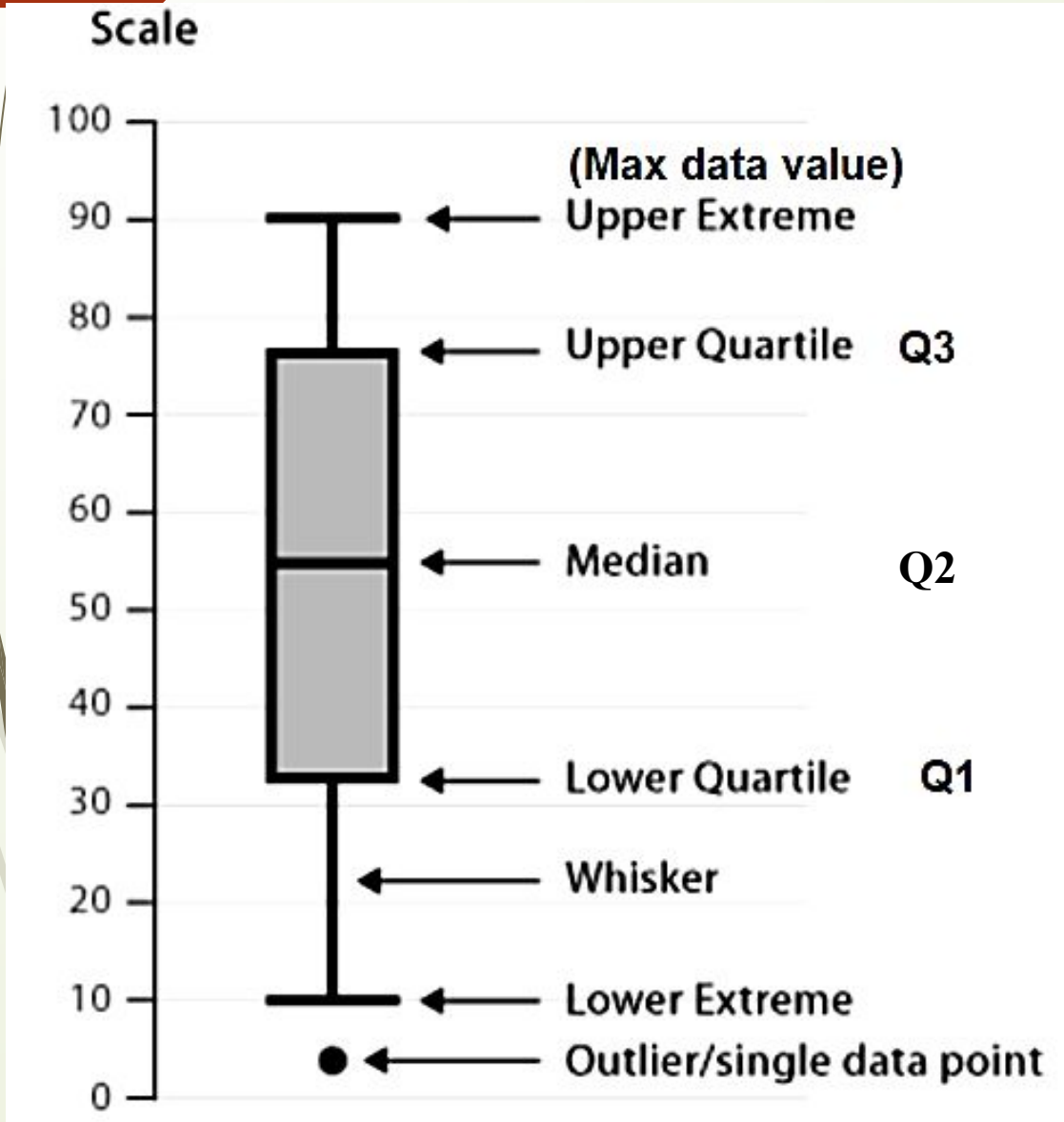
Histogram



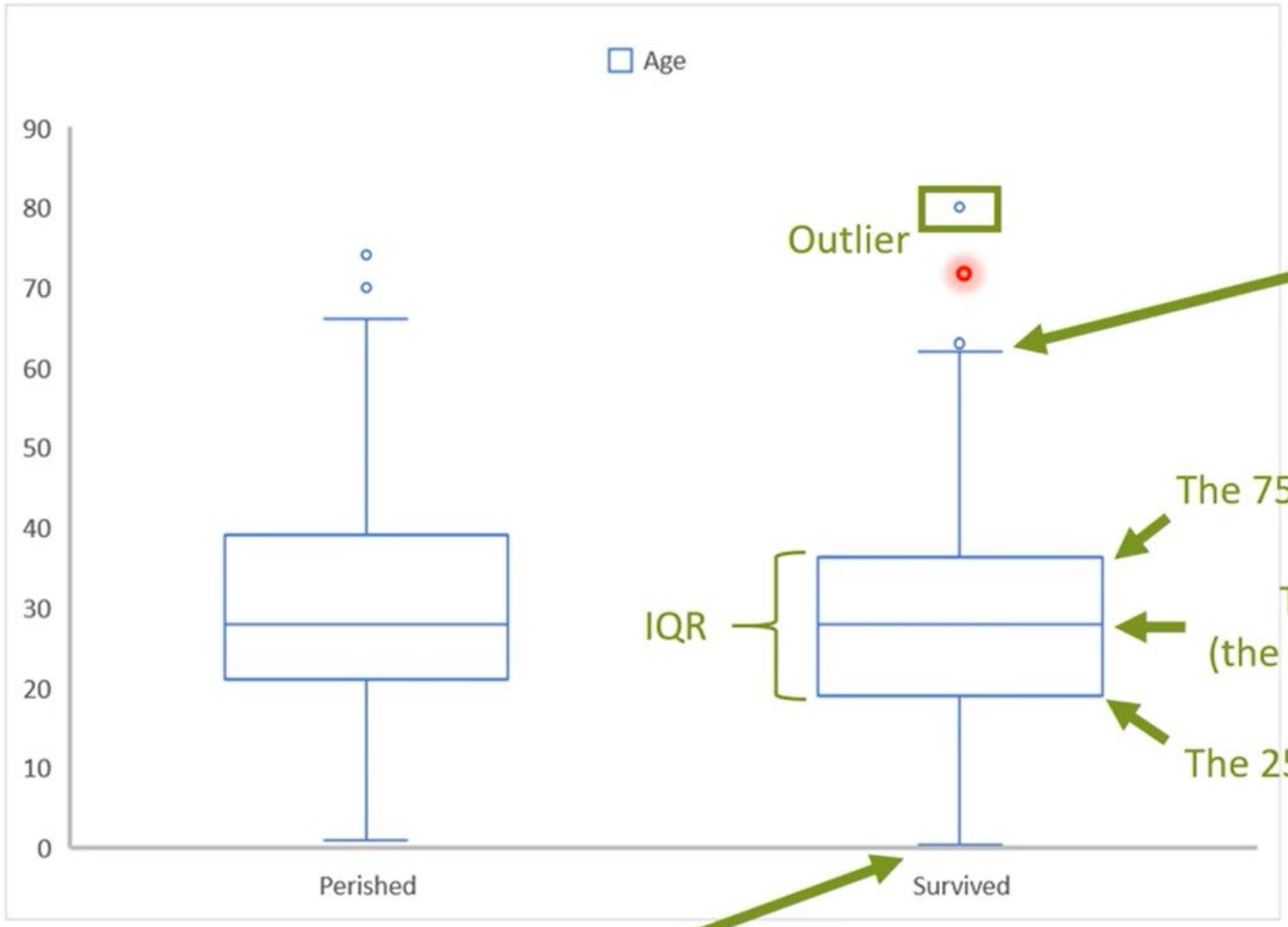


- Histogram shows us the frequency distribution of a variable.
- It partitions the spread of numeric data into parts called as "bins" and then counts the number of data points that fall into each bin.
- A histogram visualizes frequency distribution, with bins on the x-axis and data count on the y-axis.
- In general, histograms are plotted such that:
  - Empty bins are included in the graph.
  - Bins are equal width.
  - Number of bins (or, equivalently, bin size) is up to the user. Typically, between about 5 and 30 bins are chosen, depending on the amount of data and the shape of the distribution.
  - Bars are contiguous — no empty space shows between bars, unless there is an empty bin.
- Of course you need to see the histogram to know the shape of the distribution, so this may be an iterative process. It is often worthwhile to try a few different bin sizes/numbers because, there may sometimes be a different shape to the histogram when the bin size changes.
- Histograms are one of the best ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.

Boxplot



- Another very useful univariate graphical technique is the boxplot (or a box and whisker plot).
- Boxplots are very good at presenting information about the central tendency, symmetry and skew, as well as outliers. The boxplot is described here in its vertical format, which is the most common, but a horizontal format also is possible.
- A box plot provides a visual summary of data through its quartiles. First, a box is drawn from the first quartile to the third of the data set.
- A line within the box represents the median.
- "Whiskers," or lines, are then drawn extending from the box to the minimum (lower extreme) and maximum (upper extreme) values.
- Finally, outliers are represented by individual points that are in-line with the whiskers and occur outside the upper and lower extremes.



Either:
- The max data value
- 75th percentile + 1.5 * IQR

Whichever is *smaller*.

The 75th percentile

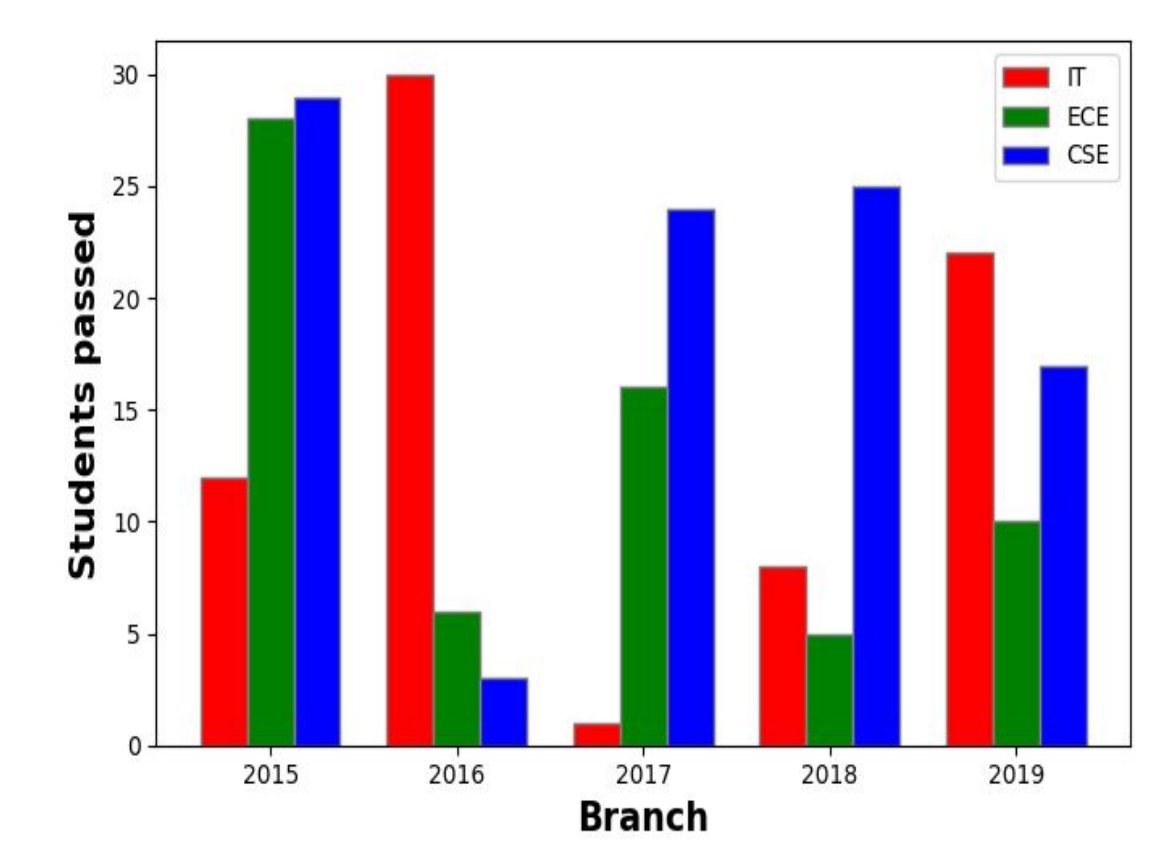
The median
(the 50th percentile)

The 25th percentile

Either:
- The min data value
- 25th percentile - 1.5 * IQR

Whichever is *larger*.

Multivariate Graphical EDA

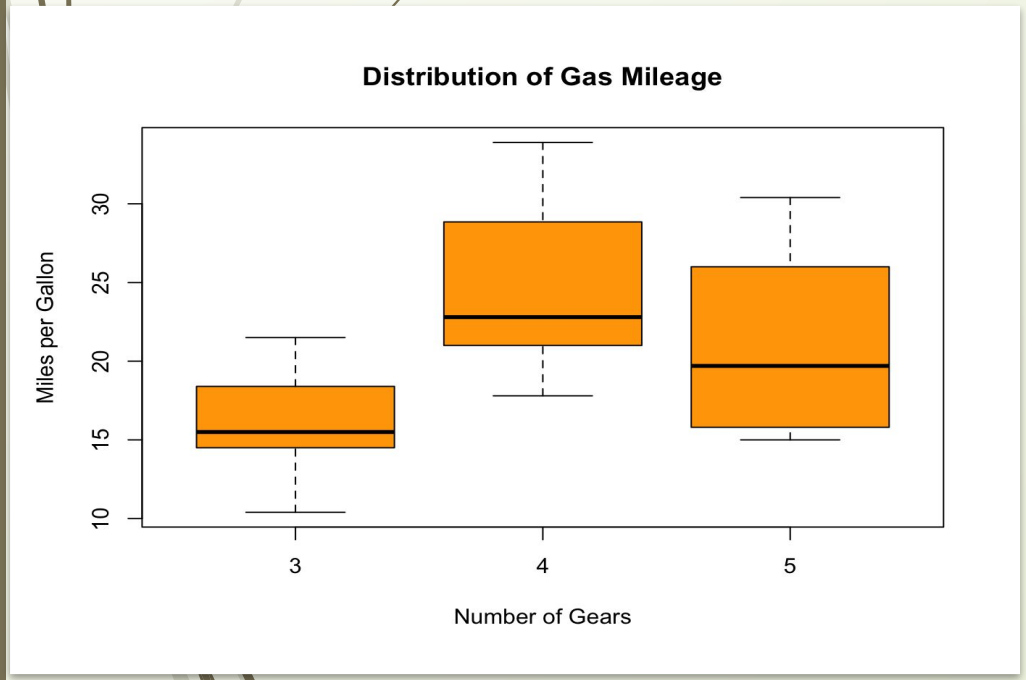


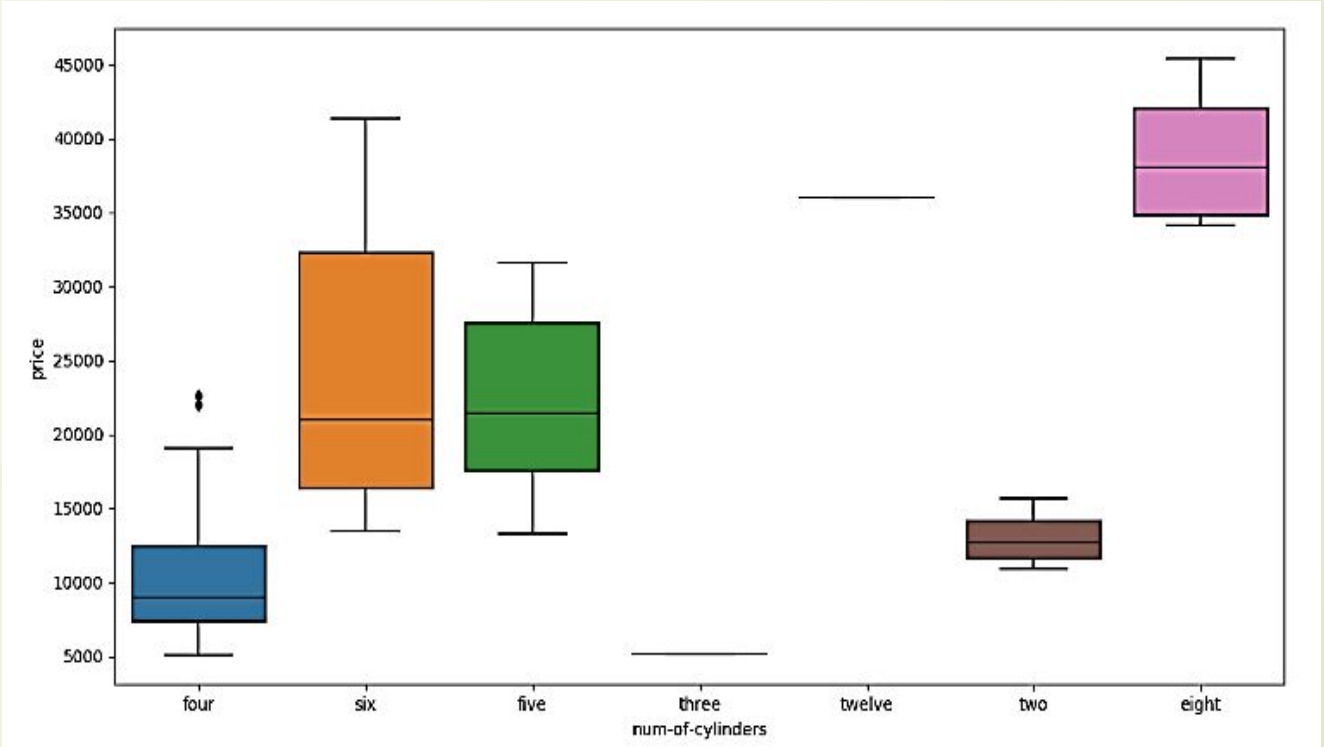
- There are a few specific techniques such as scattered plot and Heat Map available for graphical EDA of two / more categorical random variables.
- Additionally, a grouped Bar graph and side-by-side Boxplots can be used for multivariate EDA.
- In grouped bar graph, one axis of the chart shows the categories being compared, and the other axis, a measured value. The length of the bar indicates how each group measures according to the value.
- When we have one categorical (usually explanatory in x-axis) and one quantitative (usually outcome in y-axis) variables, graphical EDA usually focuses on the categorical random variable and makes plot of the quantitative variable.

Side-by-side Boxplots

In side-by-side boxplot, a categorical variable is added in the boxplots.

Side-by-side boxplots or Parallel box plots are used for multi-variate graphical EDA.

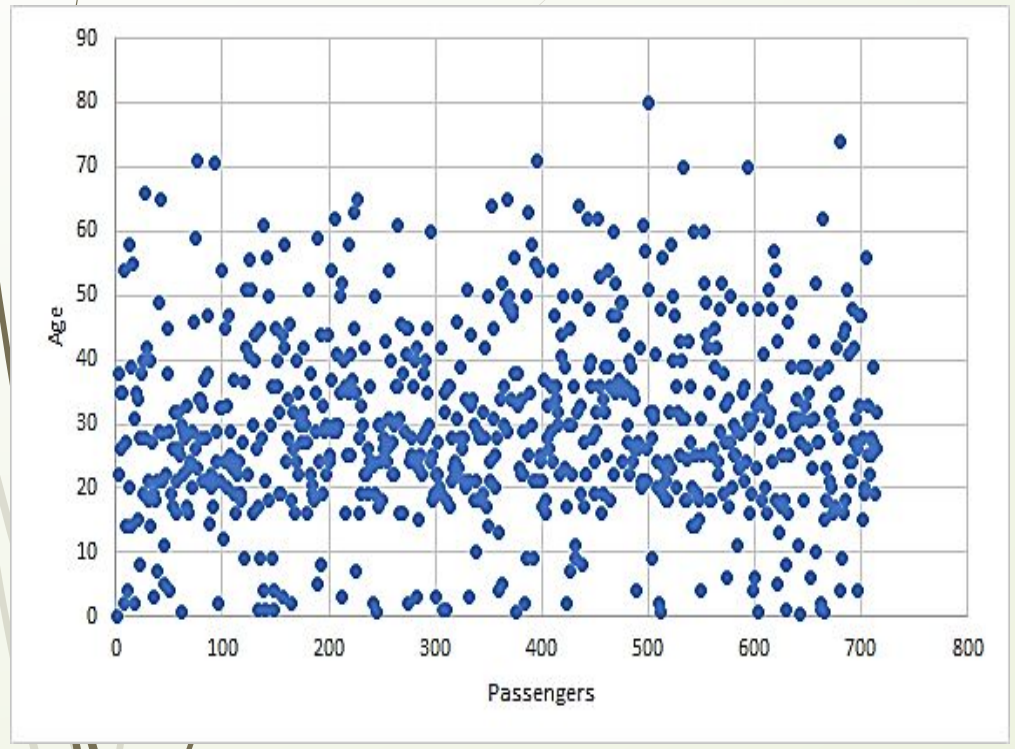




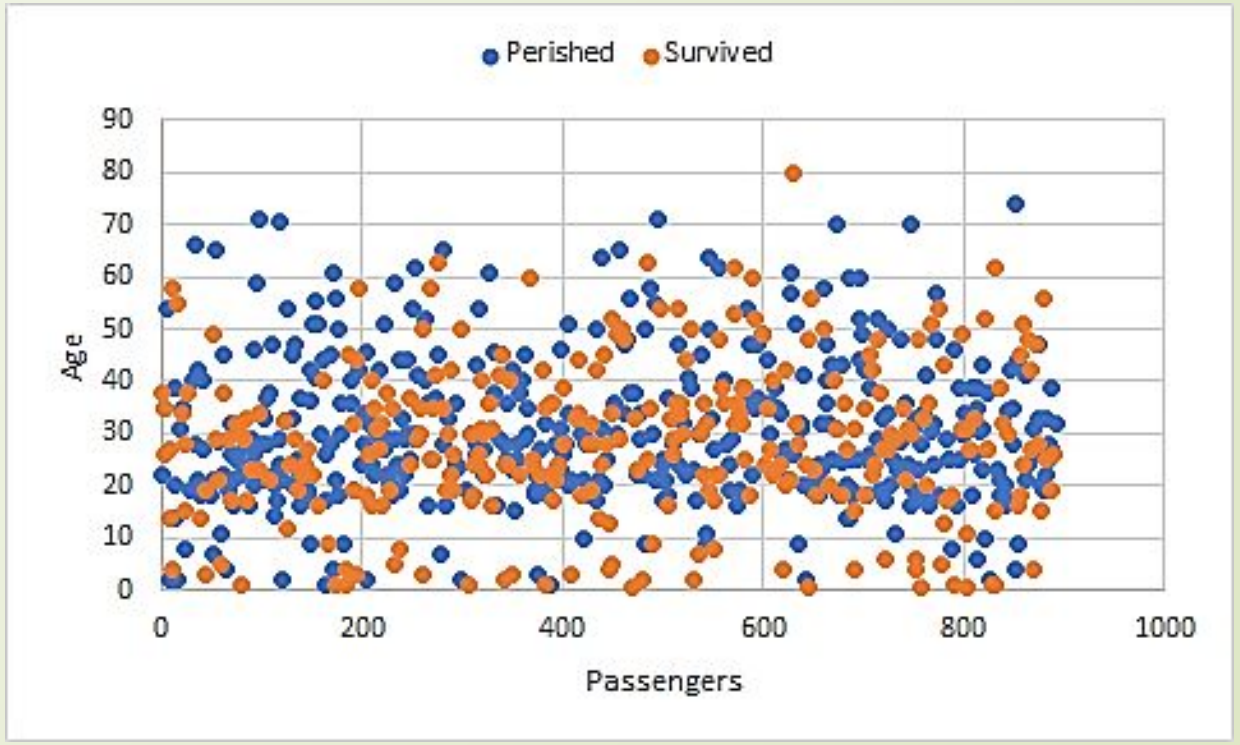
Example:

- Cars with different cylinders, we can see that the price of eight-cylinder cars lies between 35,000 to 45,000, while the price of four-cylinder cars lies between 5,000 to 19,000.
- It also tells us that the average price of five-cylinder cars and six-cylinder cars is almost same.

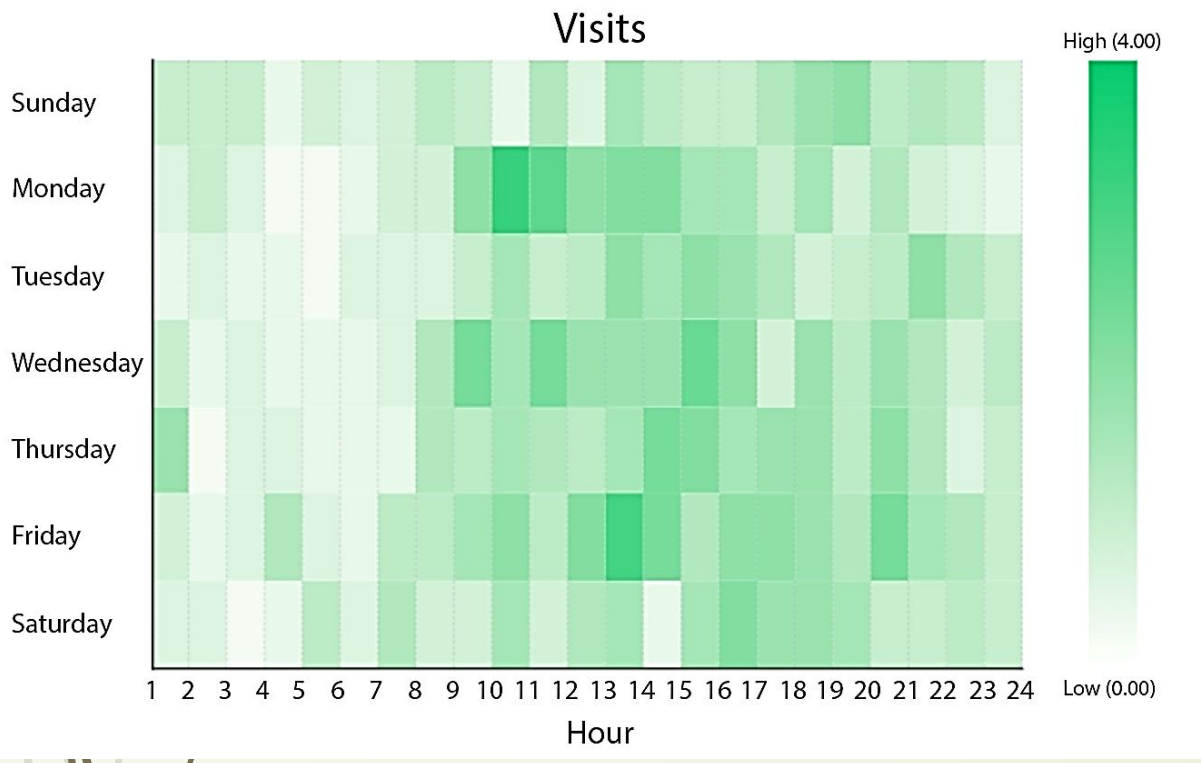
Scatter Plot



- For two **quantitative multivariate variables**, the basic graphical EDA technique is **the scatterplot** which has one variable on the x-axis, one on the y-axis and each point on the graph is a data record/ instance in the dataset.
- If one variable is explanatory and the other is outcome. Typically, the explanatory variable goes on the x-axis, and the outcome variable on the y-axis.
- Additional categorical variables can be accommodated on the scatterplot by encoding the additional information by the use of colour or symbols.



Heat Map



This Heat Map is displaying the days and times whereby a retail store is most heavily trafficked.

For example, it's clear that the store experiences an influx of traffic on Monday's beginning at 10:00 AM, which slowly dissipates throughout the day. Friday's tend to experience prolonged periods of traffic, whereas Sunday's tend to be lighter.

- Heatmaps visualize the data in a 2-dimensional format in the form of coloured maps.
- The heatmap uses colour variation to display various details and gives visual clues about the magnitude of numeric variable.
- Heatmaps can describe the density or intensity of variables, visualize patterns, variance, and even anomalies.
- Heatmaps show relationships between variables. These variables are plotted on both axes.
- We look for patterns in the cell by noticing the color change.
- It only accepts numeric data and plots it on the grid, displaying different data values by varying color intensity.

THANK YOU

Any questions…?

JGI JAIN DEEMED-TO-BE UNIVERSITY | SCHOOL OF COMPUTER SCIENCE AND IT

School of Computer Science & IT

BCA Programme

INTRODUCTION TO DATA ANALYTICS

(23BCAD4C01) MODULE 3: Exploratory Data Analysis (EDA)

Dr. Ananta Charan Ojha, Professor

Session -3

- Categories of EDA
  - Multivariate Non-Graphical EDA
    - Cross-tabulation
    - Correlation

Multivariate non-graphical EDA

- Multivariate non-graphical EDA techniques generally show the relationship between two or more variables. Crosstabulation and correlation are commonly used.

Cross-tabulation

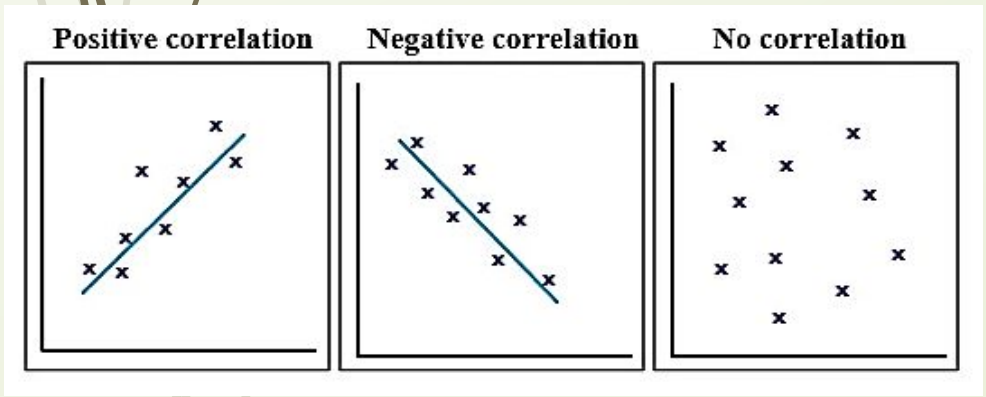
- Cross-tabulation is the basic bivariate non-graphical EDA technique used for categorical data (and quantitative data with only a few different values). It can be used for multivariate also.
- For two variables, cross-tabulation is performed by making a two-way table with column headings that match the levels of one variable and row headings that match the levels of the other variable, then filling in the counts of all data values that share a pair of levels.
- The two variables might be both explanatory, both outcome, or one of each. Depending on the goals, row percentages (which add to 100% for each row), column percentages (which add to 100% for each column) and/or cell percentages (which add to 100% over all cells) are also useful.

Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Pivot Table of Excel can be used for Cross-tabulation

Correlation

- For two quantitative variables, EDA involves examining correlation among independent variables, and between an independent variable and a dependent variable.
- When you have many quantitative variables the most common non-graphical EDA technique is to calculate all of the pairwise correlations and assemble them into a correlation matrix.
- The correlation is a measure of how much (and in what direction) one variable to change when the other changes. It measures the strength of relationship.
- 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.
- The correlation has the nice property that it is always between -1 and +1, with -1 being a perfect negative linear correlation, +1 being a perfect positive linear correlation and 0 indicating that X and Y are uncorrelated.
- The formula for the correlation is



$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

where s_x is the standard deviation of X and s_y is the standard deviation of Y .

Scatter plot is used for visualization of correlation

- In a nutshell: You should always perform appropriate EDA before further analysis of your data.
- Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn about variable distributions, and learn about relationships between variables.
- EDA is not an exact science - it is a very important art!

THANK YOU

Any questions…?

JGI JAIN DEEMED-TO-BE UNIVERSITY | SCHOOL OF COMPUTER SCIENCE AND IT

School of Computer Science & IT

BCA Programme

INTRODUCTION TO DATA ANALYTICS

(23BCAD4C01) MODULE 3: Exploratory Data Analysis (EDA)

Dr. Ananta Charan Ojha, Professor

Session -4

- Hypothesis Testing

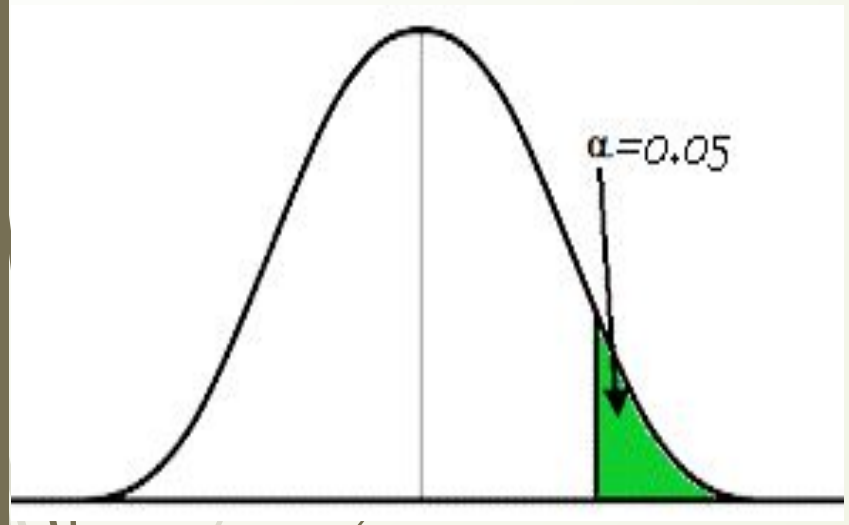
Hypothesis Testing

- Hypothesis testing is a statistical technique often used by data scientists/ analysts to validate their claims about a real-life events.
- A hypothesis is a claim about the truth related to something that exists in the world.
- For example, you may claim
  - Students studying for more hours a day gets more marks in their examination.
  - The drug A is better than the existing standard drug B.
  - The price X is more profitable than the existing price Y of the product.
- However, in order to prove the claim or reject the claim, one needs to do some experimental analysis by gathering data and evaluating the claim.
- This analysis of gathering data and evaluating claim with a goal to reject or failing to reject the hypothetical claim is termed as hypothesis testing.
- There are different techniques to test the claim and reach at the conclusion of whether the hypothesis can be used to represent the reality that exists in the world.

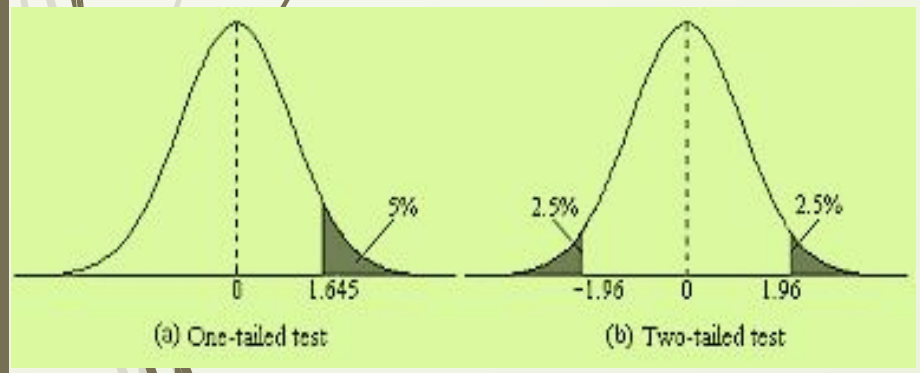
Formulation of Hypothesis

- The first step in testing a hypothesis is first defining the hypothesis. This is done by establishing both a null and alternative hypothesis.
- A null hypothesis can be thought of as a statement claiming no relationship between two measured events. It is an assumption made, which may be based on domain experience. The Null Hypothesis is denoted as H_0
- On the other hand, an alternative hypothesis contradicts the Null Hypothesis and results from the experiment that we hope to show. We want the alternative hypothesis to be true. The Alternate Hypothesis is denoted as H_1
- **Example**:
  - **Null Hypothesis** H_0 : Students studying for more hours a day **do not get** more marks in their examination.
  - **Alternate Hypothesis** H_1 : Students studying for more hours a day **get** more marks in their examination.
- In hypothesis testing, the following two are the outcomes:
  - **Reject** the Null hypothesis
  - **Fail to Reject** / **Accept** the Null hypothesis

Significance Level



Alpha = 5%, denoting significance level



- After forming our null and alternative hypotheses, we should select a significance level.
- This is the measure of the influence of the evidence that needs to be available in a sample data before rejecting the null hypothesis.
- There exists a region in the sample data where we reject the null hypothesis.
- This region is known as the critical region / alpha region. The alpha region can be one-tail or two-tail.
- The significance level (denoted as α, alpha) is the threshold probability used in hypothesis testing to determine whether to reject the null hypothesis (H_0).
- The most widely used significance level is 0.05 (5%) – meaning there is a 5% chance of rejecting a true null hypothesis.
- This means there is 95% chance that we won't reject a true null hypothesis.

Compute p-value

- The p-value is calculated from sample data and represents the probability of rejecting or accepting the H_0.
- The p-value determines whether there is enough evidence to reject the null hypothesis and retain the alternative hypothesis.
  - If p-value ≤ α, the data provides strong enough evidence to reject H_0.
  - If p-value > α, the sample data is not sufficient to reject H_0, accept H_0.
- Many tests may be used to compute the p-value. The hypothesis testing types include the t-test, z-test, ANOVA test, and chi-square test.

t-Test

Attendance	Marks
90.48	33
80.95	28
95.24	32
61.9	20
76.19	31
95.24	33
85.71	29
100	33
38.1	13
25.5	10
57.14	15
90.48	28

t-Test: Paired Two Sample for Means

	Attendance	*Marks*
Mean	74.74416667	25.41666667
Variance	579.7057174	72.99242424
Observations	12	12
Pearson Correlation	0.960068581	
Hypothesized Mean Difference	0	
df	11	
t Stat	10.64405488	
P(T<=t) one-tail	1.97438E-07	
t Critical one-tail	1.795884819	
P(T<=t) two-tail	3.94877E-07	
t Critical two-tail	2.20098516	

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\left(s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}}$$

- The t-Test is used to test the null hypothesis. Applied when Sample size is small.
- Example: Given the attendance and marks of 12 students.

$\mathbf{H_0}$: Higher attendance does not fetch good marks.

$\mathbf{H_1}$: Higher attendance fetches good marks.

Conclusion: Consider a two-tail test. If p-value ≤ α, we reject the null hypothesis.
In this case it is, 3.94877E-07 < 0.05. Therefore, we reject the null hypothesis H_0.

A null hypothesis is a claim that the data scientist / researcher hopes to reject.

THANK YOU

Any questions…?