Exploratory Data Analysis

Module 1 : Data Sourcing

Module Introduction:-

Welcome to a dynamic module in the field of machine learning on Exploratory Data Analysis, also known as EDA.

Prerequisites

Before proceeding with this module, you are expected to have a good grasp over the different Python libraries such as Pandas, NumPy, Matplotlib and Seaborn. The modules for these libraries are provided as part of the optional preparatory content.

In this module

You will learn how to explore a data set end-to-end, i.e., how to extract the maximum insights from a data set and how to make useful business decisions based on those insights.

As you move ahead in this module, you will learn about the different steps involved in exploratory data analysis and also understand how to infer useful and actionable insights from a given data set. EDA is arguably the most important and revelatory step in any kind of data analysis.

Let’s quickly go through the module flow with Anand.

By now, you have a fair understanding of EDA and the following broad topics that will be covered in this module:

- Data sourcing
- Data cleaning
- Univariate analysis
- Bivariate and multivariate analysis

In order to understand the practical aspects of EDA, you will be working on a case study using the **‘Bank Telemarketing Campaign’** data set implemented in Python.

In this session

You will learn about various data sources and also learn how to source data from public and private sources. Data sourcing is the very first step of any data analysis activity. You will focus on public data sets, as they are open to use and fetch. Here, you will be introduced to certain useful websites and techniques such as web scraping, which are used to obtain data from websites.

Introduction to EDA:-

Now, let’s discuss what EDA actually means.

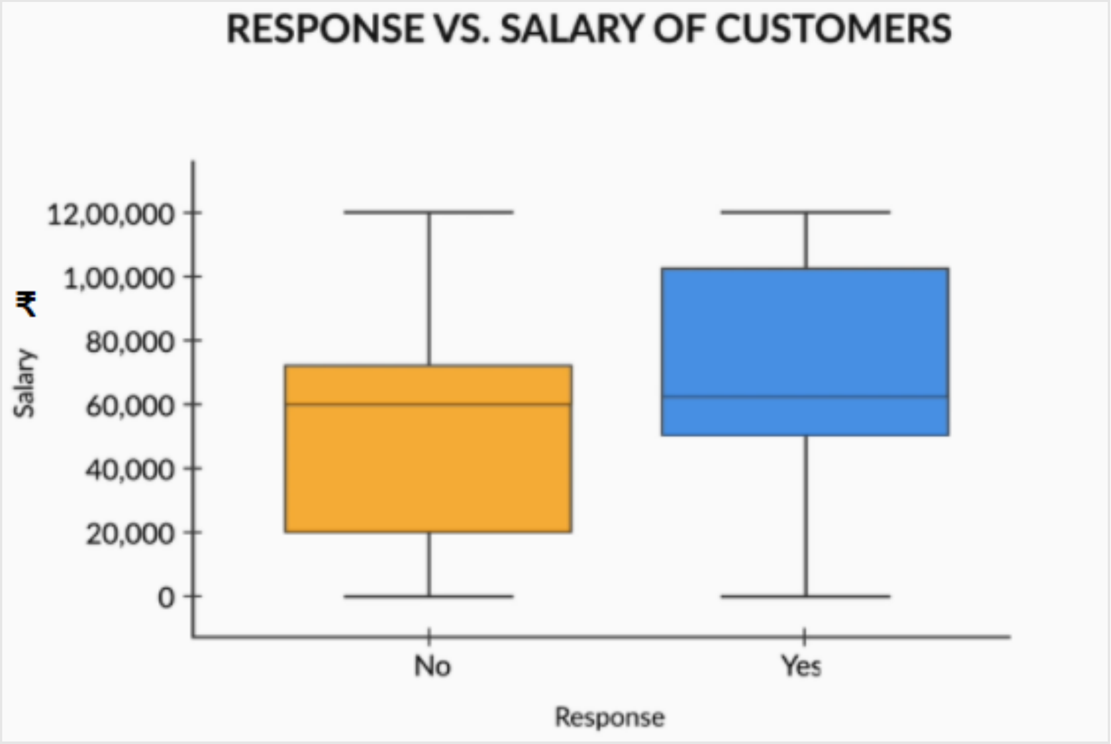
Exploratory data analysis uses data visualisation techniques to draw inferences and obtain insights from them. However, EDA is much more than plotting graphs or visualising data, it is more about understanding and studying the given data in detail. Visualisation of data into plots/graphs can be termed one of the tools in the EDA process.

EDA also involves the preparation of data sets for analysis by removing irregularities in the data so that these irregularities do not affect further steps in the process of data analysis and machine learning model building.

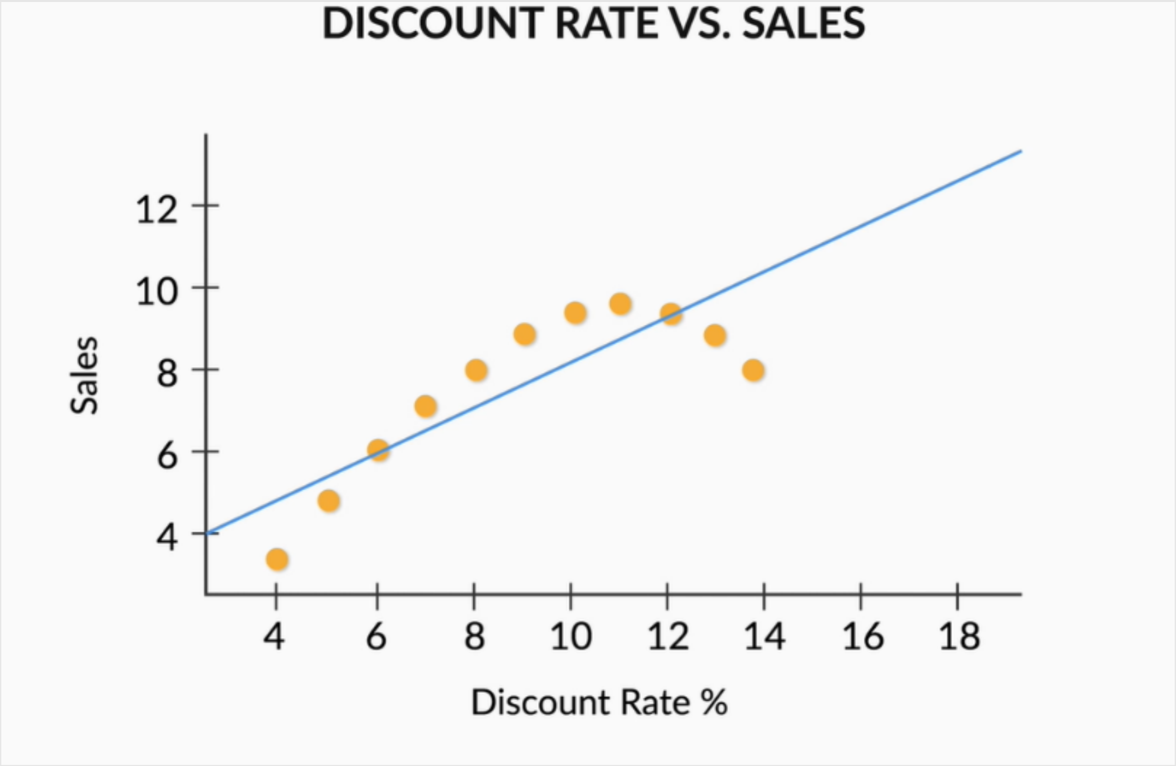
You have gone through the utility of EDA, which is as follows:

- Maximise the insight in the data set
- Detect outliers and anomalies
- Test underlying assumptions

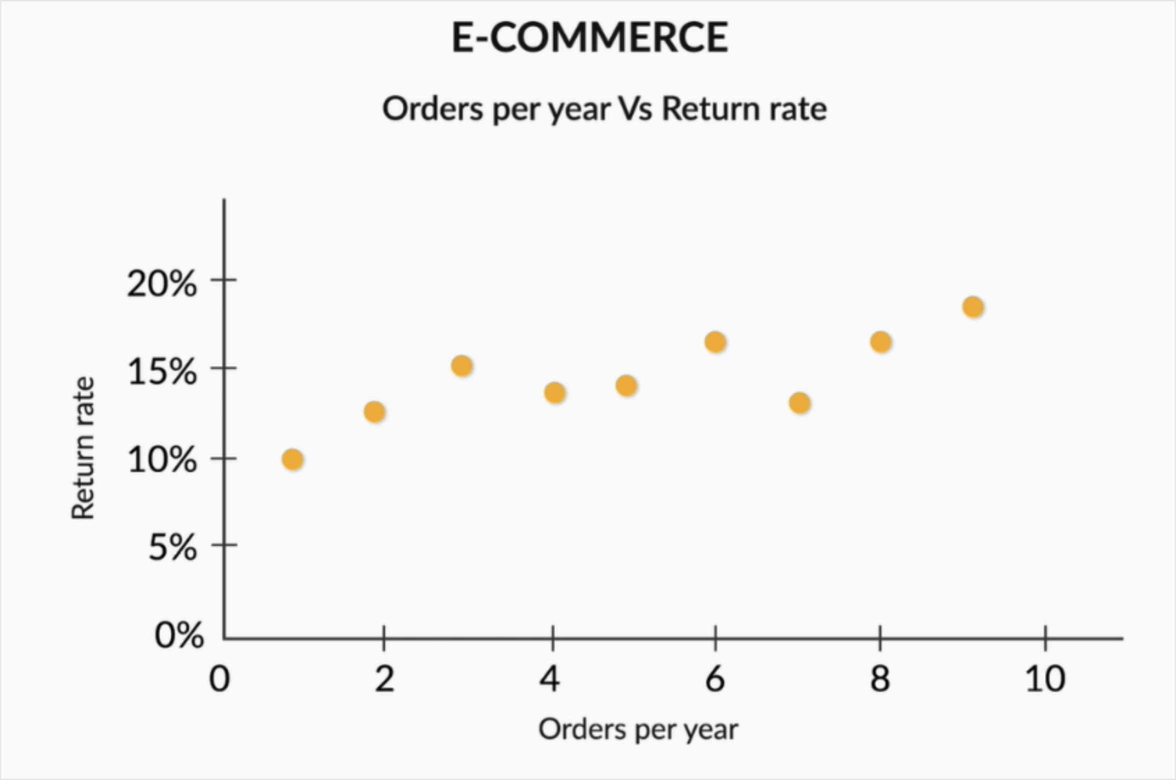
You also saw how the box plot of the **banking data set** gives a clear idea that more positive responses came from people with higher salaries because 50% of the data with a 'yes' response lies in the higher salary region. This is despite the fact that people with positive as well as negative responses have almost the same median values.



It is generally believed that higher discount means more sales. However, from the **sales** example covered previously, you understood that after a certain level of discount, sales actually start dropping. One of the possible inferences that we can draw from this is that customers may believe that a very high discount implies a compromise of quality.



Also, through the e-commerce example, you must have understood that frequent buyers have more returns frequency.



So, now you have understood that EDA is an important exercise before proceeding further with a data set. It does not involve merely finding irregularities in the data, such as missing values or outliers; it is a combination of fixing the data set for useful purposes and then deriving maximum insights from that data, by either plotting graphs or using statistical parameters.

An important takeaway from this is that EDA should be the first step in any data science / machine learning activity. Based on the results of EDA, companies also make business decisions, which can have repercussions later. Hence, we observe the following:

- If not performed properly, EDA can hamper the further steps in the machine learning model building process.
- If done well, it may improve the efficacy of all we do in the next steps.

In the next video, you will get an idea of how EDA has evolved and what kind of work has been done before in this field.

Let's listen to a brief history of EDA:-

In 1977, John W. Tukey wrote a book on EDA and developed box plots, which are also called Tukey's box plots. Since then, many books have been written in the field. You may refer to this link to get the resources of the evolution of EDA.

Public and Private Data:-

Broadly, data sources can be of two types:

- **Private data**
- **Public data**

Private data: As the name suggests, it is private and belongs to an organisation, and there are certain security and privacy concerns attached to it. It is used for the companies' internal analysis purposes in order to gain business and growth insights. Some examples of such organisational private data are **telecom data, retail data,** and **banking** and **medical data**.

Public data: This is the data that is available for public use and is offered by many sites such as government websites and public agencies for the purpose of research. Accessing this data does not require any special permission or approval, hence the name. Also, there are many programming techniques that are used to fetch public data through code, which you will learn about later in the module.

Let's listen to Anand to understand more about data sources types.

To summarise, data is of two types, public and private, and each comes with its set of disadvantages. Public data is not always relevant or useful, and private data is not always easily available.

Private Data:-

A large number of organisations seek to leverage data analytics to make crucial business decisions. As organisations become customer-centric, they utilise insights from data to enhance customer experience while also optimising their daily processes.

So, private data has some security and privacy issues and is not publicly available. You learnt about the use of data in the banking, telecom and human resources sectors.

- **Banking data:** Banks use data to make credit-related decisions. This data is highly sensitive, as it contains customer transaction details, account details, etc. Security of such data is of topmost importance. Banks can use such data to predict which customer is likely to take a loan in the near future or which customers are interested in investing in term deposits, etc. With this help of such data, banks can also identify which customers are likely to default on their loans.
- **Telecom data:** Telecom companies use data to optimise plans for customers and predict customer churn. Telecom data can be used to optimise the coverage area based on the customers' calls data and their call performances.
- **HR data:** HR data analytics helps identify and predict employee

behaviour.

you learnt the applications of data in the media and retail industries.

- **Retail data:** Retail data analytics helps drive decisions such as product purchasing, pricing and stocking.
- **Media data:** The media industry uses data extensively to target viewers. Advertisers use data to identify the best avenues for targeting customers, while journalists use data visualisation to gather relevant information.

Public Data:-

Now, let's learn about public data sources and the techniques to extract data from them. Public data is available on various platforms on the internet. Such platforms can be any open websites such as government websites or any online learning websites, some of which we will discuss here.

So, now you have an idea of how public platforms can be a good source of data. Let's quickly discuss some interesting platforms that are helpful to explore data analytics and machine learning fields.

- **Kaggle:** It is a subsidiary of Google LLC. It is an online community of data scientists and machine learning engineers, where you can find and publish data sets and explore your own developed solutions on an open web-based environment. Kaggle also organises several machine learning competitions online. Here is the link to the Kaggle website.
- **UCI Repository of Machine Learning:** It is an online community of data science and machine learning engineers. As the name suggests, it is a repository of data sets that are openly available. You can find interesting case studies to explore data analytics and machine learning. Here is the link to UCI Repository of Machine Learning website.

Given below are the links to some public data sets. You may explore these open sources to get the data.
GitHub: **Awesome public data sets**, **Github data sets**
Open government data set: **Open Government data**

In the next segment, you will learn about web scraping and understand how you can fetch data from websites using code.

Web Scraping-I:-

The IMDB website provides movie-related information such as **release date, runtime duration, cast, genre, ratings,** etc. Now, consider a specific webpage on the IMDB website that lists the top-rated movies along with information about them. It lists the top 50 rated movies, the number of votes, etc. Now, what if you want to perform a deeper analysis to answer certain questions as follows?

- Which director has the highest number of movies in the top 50 rated movies?
- Which genre has the highest rating among the top 50 rated movies?
- What is the gross expenditure of the lowest-rated movies as compared with the highest-rated ones?

One of the approaches to answering these kinds of questions is manual, which involves checking the information manually and entering it in a spreadsheet. Does this not seem to be a tedious and mundane task? This is where the technique of web scraping comes into the picture. It eases the task of obtaining and processing data with the help of a structured format. This would help you perform deeper analyses and answer the aforementioned questions.

So, web scraping helps to fetch such information from websites. You will learn about web scraping in three parts:

- Need for and application of web scraping
- The basics of an HTML page
- Python libraries and codes for web scraping

So, let's listen to Rahim to understand how useful web scraping is.

Note:
Please note that web scraping is not always legal for all websites. Certain websites provide access to others to scrape their data from the web page. Another important aspect is that if the content is copyrighted (such as video, pictures and articles), it is not illegal to scrape it, but it is illegal to republish it. Also, you cannot scrape a website just to build a duplicate competing site; it is acceptable to scrape data as long as you are using it to create something new.

The Basics of an HTML Page

The basic requirement of web scraping is the web page that we are going to scrape. All web pages are written in HTML. So, you can perform web scraping using Python only after you understand the basic structure of an HTML page.

So, let's get into the basics of how an HTML page looks and what its tags are.

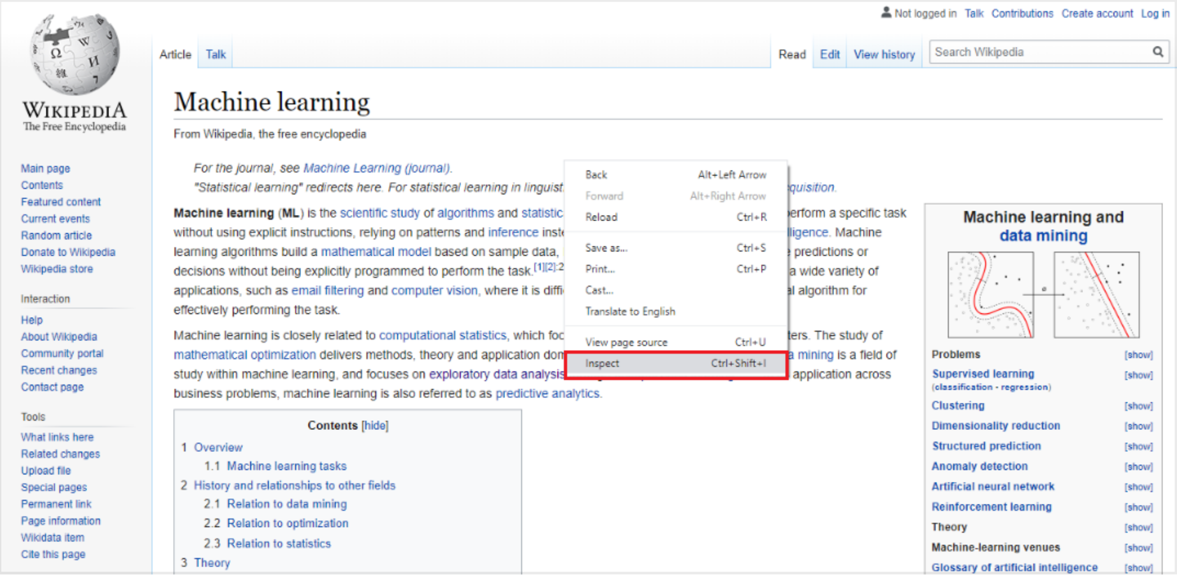
Note: Here, we are not going to have an end-to-end discussion on HTML codes; you will gain an understanding of only those concepts that are useful for fetching data from the web page in the scraping process.

HTML stands for 'Hypertext Markup Language'. It is used for creating an electronic document to display it on the world wide web. Each page that you see on the internet is written in HTML.

You can learn the basics of HTML using the Wikipedia page on 'Machine learning' which is shown below in the form of snapshot.

You can check the HTML code of any web page by following the instructions provided below:

Open web page -> Right-click -> Inspect



Machine learning

From Wikipedia, the free encyclopedia

For the journal, see Machine Learning (journal).
"Statistical learning" redirects here. For statistical learning in linguist quisition.

Machine learning (**ML**) is the scientific study of algorithms and statistic erform a specific task without using explicit instructions, relying on patterns and inference inste ligence. Machine learning algorithms build a mathematical model based on sample data, predictions or decisions without being explicitly programmed to perform the task.[1][2]:2 a wide variety of applications, such as email filtering and computer vision, where it is diffi l algorithm for effectively performing the task.

Machine learning is closely related to computational statistics, which foc ters. The study of mathematical optimization delivers methods, theory and application don a mining is a field of study within machine learning, and focuses on exploratory data analysis application across business problems, machine learning is also referred to as predictive analytics.

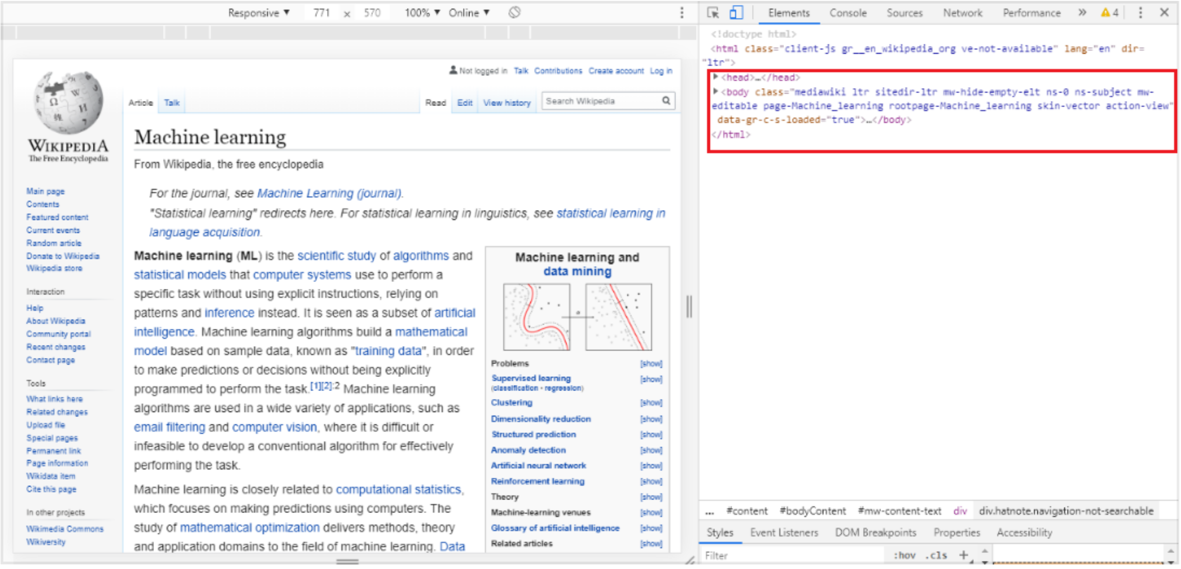
Contents [hide]

1 Overview
  1.1 Machine learning tasks
2 History and relationships to other fields
  2.1 Relation to data mining
  2.2 Relation to optimization
  2.3 Relation to statistics
3 Theory

Machine learning and data mining

Problems	[show]
Supervised learning (classification • regression)	[show]
Clustering	[show]
Dimensionality reduction	[show]
Structured prediction	[show]
Anomaly detection	[show]
Artificial neural network	[show]
Reinforcement learning	[show]
Theory	[show]
Machine-learning venues	[show]
Glossary of artificial intelligence	[show]

Once you click on 'inspect', you will see the HTML code of this particular page on the right side of the screen as shown in the screenshot provided below.



Machine learning

From Wikipedia, the free encyclopedia

For the journal, see Machine Learning (journal).
"Statistical learning" redirects here. For statistical learning in linguistics, see statistical learning in language acquisition.

Machine learning (**ML**) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.[1][2]:2 Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data

```
<!doctype html>
<html class="client-js gr__en_wikipedia_org ve-not-available" lang="en" dir=
"ltr">
 ▶<head>…</head>
 ▶<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-
editable page-Machine_learning rootpage-Machine_learning skin-vector action-view"
  data-gr-c-s-loaded="true">…</body>
</html>
```

HTML code has a tree-like hierarchical structure, or nested structure, which contains a Head and a Body. The web page that you see on screen is due to the 'body', which contains most of the important codes for the web page.

An HTML page broadly consists of two basic elements:

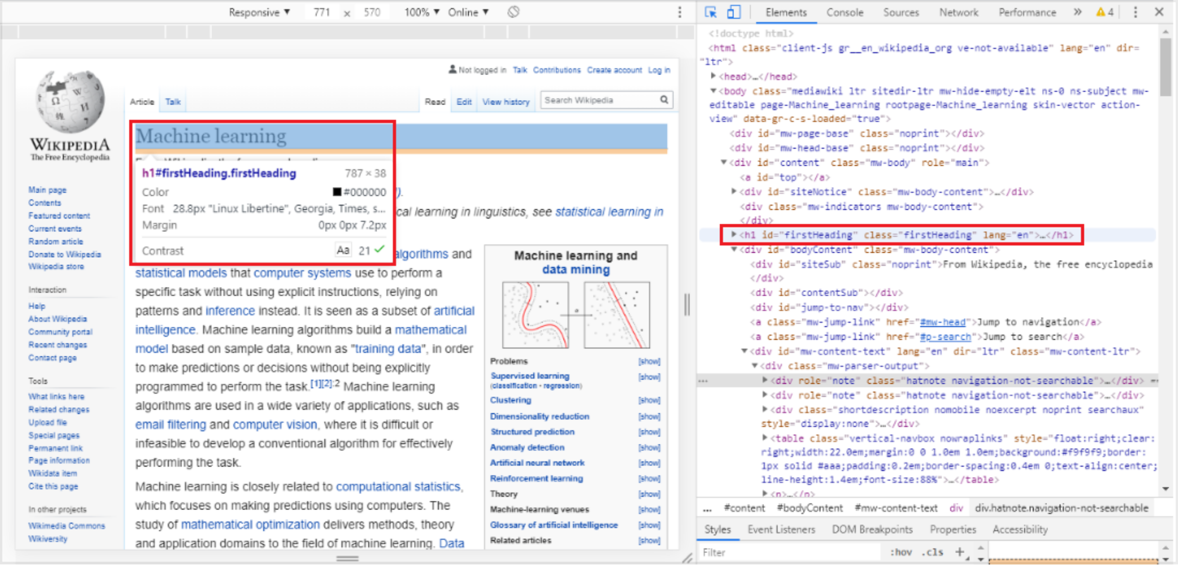
- **Attributes**: These are used to describe the characteristics of an element. They majorly contain the **class, id** and **href**. These are like objects that are created to define the different segments of a web page.
- **Tags**: A tag is a way to represent an HTML element. Tags majorly contain **h (heading), p (paragraph), a (hyperlink)** and **div**.

Let’s briefly go through the attributes one by one.

- **Class**: The HTML class attribute is used to specify a single or multiple class names for an HTML element.
- **Id**: This attribute is used to provide a specific ID to an element.
- **href**: This attribute is used to provide any web page link that is embedded in the text on the HTML page.

A group of elements may have the same attributes but will have different tags. Let’s go through the tags using the Wikipedia page examples to understand the concept better.

- **Heading:** It is represented by ‘h’ in HTML code. It is used to place the headings of sections on a web page.



Responsive ▾ 771 × 570 100% ▾ Online ▾
Elements Console Sources Network Performance » ⚠4
Not logged in Talk Contributions Create account Log in
Article Talk Read Edit View history Search Wikipedia
WIKIPEDIA
The Free Encyclopedia
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page
Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page
In other projects
Wikimedia Commons
Wikiversity
Machine learning and data mining
Problems [show]
Supervised learning (classification • regression) [show]
Clustering [show]
Dimensionality reduction [show]
Structured prediction [show]
Anomaly detection [show]
Artificial neural network [show]
Reinforcement learning [show]
Theory [show]
Machine-learning venues [show]
Glossary of artificial intelligence [show]
Related articles [show]
Styles Event Listeners DOM Breakpoints Properties Accessibility
Filter :hov .cls +

Machine learning

h1#firstHeading.firstHeading 787 × 38
Color ■#000000
Font 28.8px "Linux Libertine", Georgia, Times, s...
Margin 0px 0px 7.2px
Contrast Aa 21 ✓

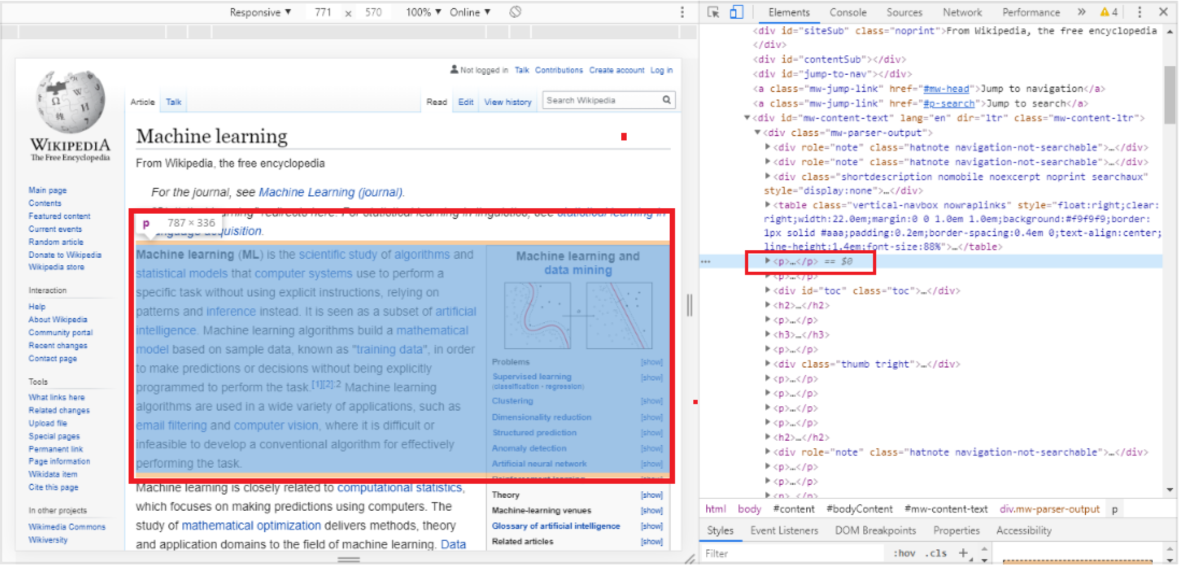
...cal learning in linguistics, see statistical learning in

...algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.[1][2]:2 Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data

```
<!doctype html>
<html class="client-js gr__en_wikipedia_org ve-not-available" lang="en" dir=
"ltr">
 ▶<head>…</head>
 ▼<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-
 editable page-Machine_learning rootpage-Machine_learning skin-vector action-
 view" data-gr-c-s-loaded="true">
     <div id="mw-page-base" class="noprint"></div>
     <div id="mw-head-base" class="noprint"></div>
   ▼<div id="content" class="mw-body" role="main">
      <a id="top"></a>
     ▶<div id="siteNotice" class="mw-body-content">…</div>
      <div class="mw-indicators mw-body-content">
      </div>
     ▶<h1 id="firstHeading" class="firstHeading" lang="en">…</h1>
     ▼<div id="bodyContent" class="mw-body-content">
        <div id="siteSub" class="noprint">From Wikipedia, the free encyclopedia
        </div>
        <div id="contentSub"></div>
        <div id="jump-to-nav"></div>
        <a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
        <a class="mw-jump-link" href="#p-search">Jump to search</a>
       ▼<div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr">
         ▼<div class="mw-parser-output">
...         ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
           ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
           ▶<div class="shortdescription nomobile noexcerpt noprint searchaux"
           style="display:none">…</div>
           ▶<table class="vertical-navbox nowraplinks" style="float:right;clear:
           right;width:22.0em;margin:0 0 1.0em 1.0em;background:#f9f9f9;border:
           1px solid #aaa;padding:0.2em;border-spacing:0.4em 0;text-align:center;
           line-height:1.4em;font-size:88%">…</table>
           ▶<p>…</p>
... #content #bodyContent #mw-content-text div div.hatnote.navigation-not-searchable
```

- **Paragraph:** It is represented by 'p' in HTML code. It is used to place a paragraph on the web page.



Responsive ▾ 771 × 570 100% ▾ Online ▾
Elements Console Sources Network Performance » ⚠4
Not logged in Talk Contributions Create account Log in
Article Talk Read Edit View history Search Wikipedia
WIKIPEDIA
The Free Encyclopedia
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page
Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page
In other projects
Wikimedia Commons
Wikiversity
Machine learning and data mining
Problems [show]
Supervised learning (classification • regression) [show]
Clustering [show]
Dimensionality reduction [show]
Structured prediction [show]
Anomaly detection [show]
Artificial neural network [show]
Theory [show]
Machine-learning venues [show]
Glossary of artificial intelligence [show]
Related articles [show]
Styles Event Listeners DOM Breakpoints Properties Accessibility
Filter :hov .cls +

Machine learning

From Wikipedia, the free encyclopedia

For the journal, see Machine Learning (journal).

p 787 × 336

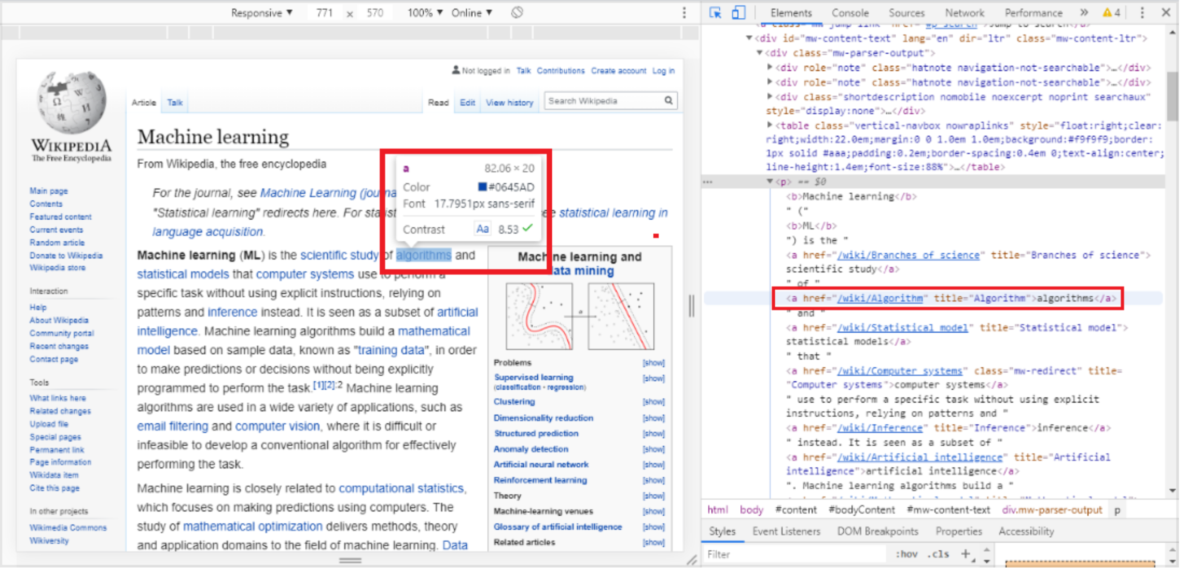
...language acquisition.

Machine learning (**ML**) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.[1][2]:2 Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data

```
        <div id="siteSub" class="noprint">From Wikipedia, the free encyclopedia
        </div>
        <div id="contentSub"></div>
        <div id="jump-to-nav"></div>
        <a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
        <a class="mw-jump-link" href="#p-search">Jump to search</a>
       ▼<div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr">
         ▼<div class="mw-parser-output">
           ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
           ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
           ▶<div class="shortdescription nomobile noexcerpt noprint searchaux"
           style="display:none">…</div>
           ▶<table class="vertical-navbox nowraplinks" style="float:right;clear:
           right;width:22.0em;margin:0 0 1.0em 1.0em;background:#f9f9f9;border:
           1px solid #aaa;padding:0.2em;border-spacing:0.4em 0;text-align:center;
           line-height:1.4em;font-size:88%">…</table>
...         ▶<p>…</p> == $0
           ▶<p>…</p>
           ▶<div id="toc" class="toc">…</div>
           ▶<h2>…</h2>
           ▶<p>…</p>
           ▶<h3>…</h3>
           ▶<p>…</p>
           ▶<div class="thumb tright">…</div>
           ▶<p>…</p>
           ▶<p>…</p>
           ▶<p>…</p>
           ▶<p>…</p>
           ▶<h2>…</h2>
           ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
           ▶<p>…</p>
           ▶<p>…</p>
html body #content #bodyContent #mw-content-text div.mw-parser-output p
```

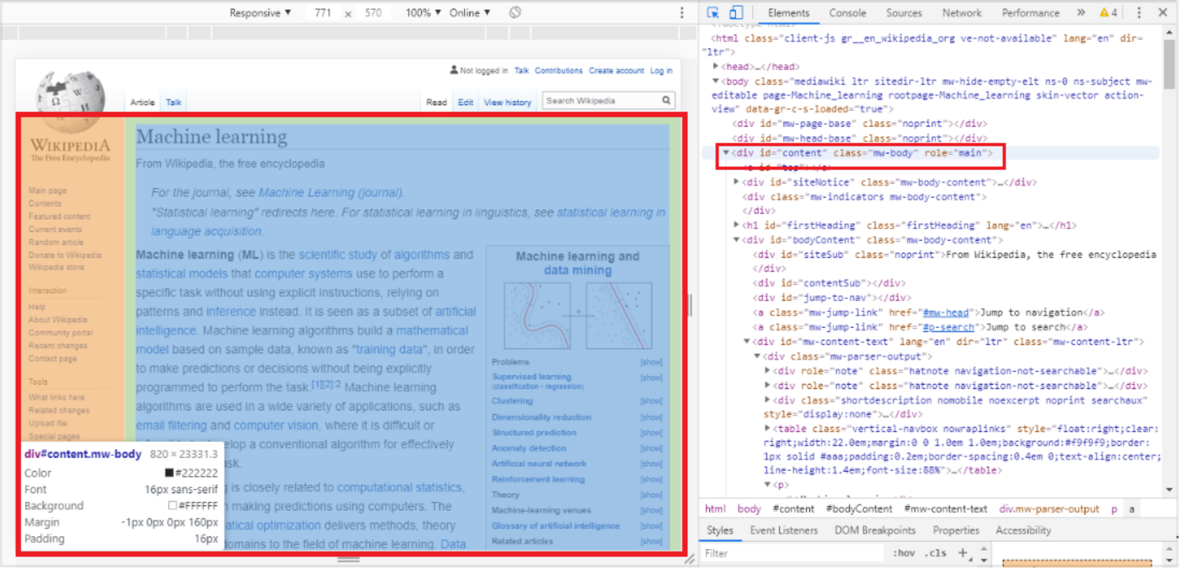
- **Hyperlink:** It is represented by 'a' in HTML code. It is used to provide a link to any other web page on the present web page.



```
▼<div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr">
  ▼<div class="mw-parser-output">
    ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
    ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
    ▶<div class="shortdescription nomobile noexcerpt noprint searchaux"
    style="display:none">…</div>
    ▶<table class="vertical-navbox nowraplinks" style="float:right;clear:
    right;width:22.0em;margin:0 0 1.0em 1.0em;background:#f9f9f9;border:
    1px solid #aaa;padding:0.2em;border-spacing:0.4em 0;text-align:center;
    line-height:1.4em;font-size:88%">…</table>
    ▼<p> == $0
        <b>Machine learning</b>
        " ("
        <b>ML</b>
        ") is the "
        <a href="/wiki/Branches_of_science" title="Branches of science">
        scientific study</a>
        " of "
        <a href="/wiki/Algorithm" title="Algorithm">algorithms</a>
        " and "
        <a href="/wiki/Statistical_model" title="Statistical model">
        statistical models</a>
        " that "
        <a href="/wiki/Computer_systems" class="mw-redirect" title=
        "Computer systems">computer systems</a>
        " use to perform a specific task without using explicit
        instructions, relying on patterns and "
        <a href="/wiki/Inference" title="Inference">inference</a>
        " instead. It is seen as a subset of "
        <a href="/wiki/Artificial_intelligence" title="Artificial
        intelligence">artificial intelligence</a>
        ". Machine learning algorithms build a "
```

Responsive ▾ 771 × 570 100% ▾ Online ▾
Elements Console Sources Network Performance
a 82.06 × 20
Color #0645AD
Font 17.7951px sans-serif
Contrast Aa 8.53 ✓
html body #content #bodyContent #mw-content-text div.mw-parser-output p
Styles Event Listeners DOM Breakpoints Properties Accessibility
Filter :hov .cls +

- **Div:** It is used to structure the HTML page. It is a nested structure that contains other HTML elements. The main purpose of the div tag is to promote encapsulation.



```
<html class="client-js gr__en_wikipedia_org ve-not-available" lang="en" dir=
"ltr">
 ▶<head>…</head>
 ▼<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-
 editable page-Machine_learning rootpage-Machine_learning skin-vector action-
 view" data-gr-c-s-loaded="true">
    <div id="mw-page-base" class="noprint"></div>
    <div id="mw-head-base" class="noprint"></div>
  ▼<div id="content" class="mw-body" role="main">
    ▶<div id="siteNotice" class="mw-body-content">…</div>
     <div class="mw-indicators mw-body-content">
     </div>
    ▶<h1 id="firstHeading" class="firstHeading" lang="en">…</h1>
    ▼<div id="bodyContent" class="mw-body-content">
       <div id="siteSub" class="noprint">From Wikipedia, the free encyclopedia
       </div>
       <div id="contentSub"></div>
       <div id="jump-to-nav"></div>
       <a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
       <a class="mw-jump-link" href="#p-search">Jump to search</a>
     ▼<div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr">
       ▼<div class="mw-parser-output">
         ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
         ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
         ▶<div class="shortdescription nomobile noexcerpt noprint searchaux"
         style="display:none">…</div>
         ▶<table class="vertical-navbox nowraplinks" style="float:right;clear:
         right;width:22.0em;margin:0 0 1.0em 1.0em;background:#f9f9f9;border:
         1px solid #aaa;padding:0.2em;border-spacing:0.4em 0;text-align:center;
         line-height:1.4em;font-size:88%">…</table>
         ▼<p>
```

Responsive ▾ 771 × 570 100% ▾ Online ▾
Elements Console Sources Network Performance
div#content.mw-body 820 × 23331.3
Color #222222
Font 16px sans-serif
Background #FFFFFF
Margin -1px 0px 0px 160px
Padding 16px
html body #content #bodyContent #mw-content-text div.mw-parser-output p a
Styles Event Listeners DOM Breakpoints Properties Accessibility
Filter :hov .cls +

- **Span:** This tag is used for grouping and applying styles to inline elements.

This is the basic information required to understand the HTML page structure. We will not be covering HTML codes in depth.

Web Scraping-II:-

So, you have a very basic understanding of what an HTML page looks like. Now, let's come to the application part, ie, how you can fetch the website data using Python.

Let’s understand the HTML page of an IMDB web page from Rahim in the next video.

You took a look at the web page of the top 50 IMDB movies, which contains movie names, rating, votes, director details and cast in a specific container-like structure. Now, you will learn how to code in order to fetch data from the web page in Python.

Here is a summary of the major takeaways from the video provided above:

- **request library**: It is a Python library that is used to read the web page data from the URL of the corresponding page.
- **BeautifulSoup**: It is a Python package that helps in parsing and extracting data from HTML and XML files.
- The web scraping process can be divided into four major parts:

1. **Reading**: For HTML page read and upload
2. **Parsing**: For beautifying the HTML code in an understandable format
3. **Extraction**: For extraction of data from the web page
4. **Transformation**: For converting the information into the required format, e.g., CSV

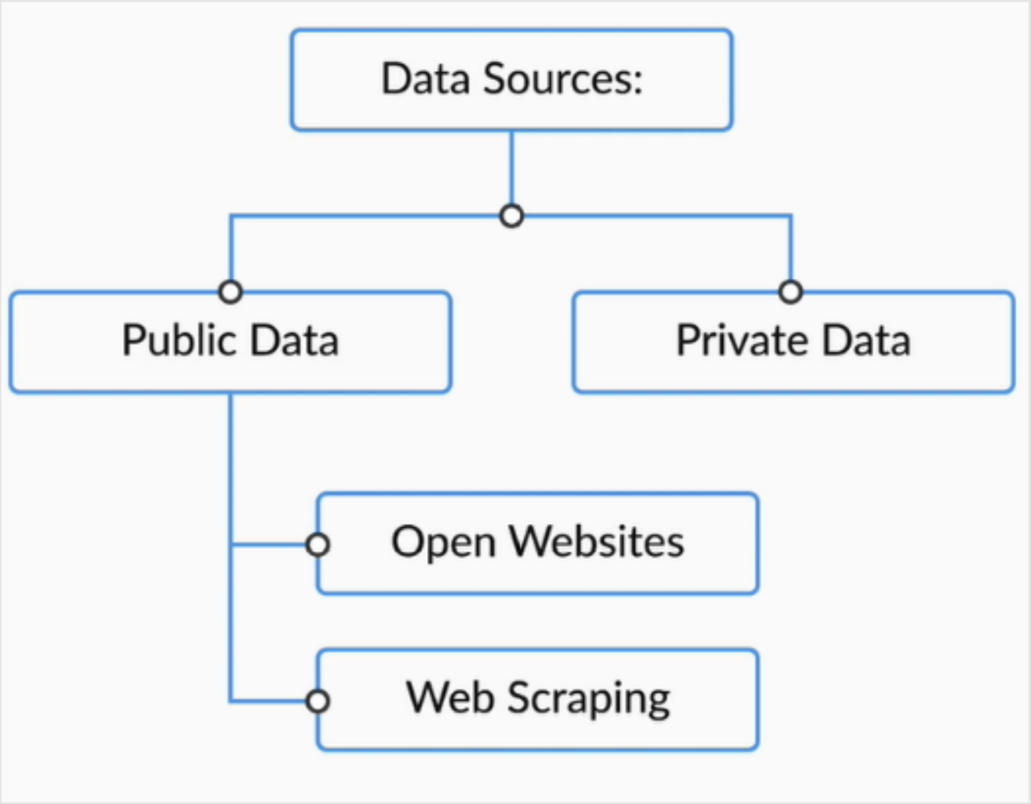
You are provided with the well-commented Jupyter Notebook that was covered in the video. This is just for your reference, and it is a basic web scraping example. There are many other techniques and concepts in web scraping, but they are out of the scope of this module. However, you have been given an idea of the process of web scraping with a basic understanding of HTML.

You have gone through the top 50 movies page on IMDB's website and seen the scraping process using Python. This page contains the information of movies such as: name, rating, votes, runtime, genre, director details, actors and plot of the movie. As explained earlier, if you want to fetch information from the web page into a CSV file, then you need to look into its HTML code to get an idea about tags and attributes.

Summary:-

Let’s summarise our understanding of data sourcing, about which you learnt in this session.
There are two main types of data:

- **Public data:** This is the data that is made publicly available for the purpose of research and learning.
- **Private data**: This is organisational data, and organisations have some security and privacy concerns. Company approvals are needed to access such data. It is useful for internal policymaking and business strategy building for an organisation.



Given below are the links to some public datasets. You may explore these open sources to get the data:
GitHub: Awesome public data sets, Github data sets
Open government data set: Open Government data.

Kaggle website link: Kaggle Website
UCI repository of machine learning: UCI machine learning data set repository.

Apart from the learning of public and private data sources, you learnt about a data fetching technique, i.e., **web scraping**, which is very useful to fetch data directly from webpages. It is useful in many applications like e-commerce price comparison, real estate, share market, etc.

Web scraping majorly involves 4 steps:

- **HTML loading and reading**: It includes the loading of the HTML page into Python. The library which is used here to request for the HTML page is the "**request**" library.
- **HTML parsing**: This step involves the process of presentation of HTML code into a readable format. One of the important classes of Python called "**BeautifulSoup**" is used here to parse the data.
- **Data extraction**: This step involves the extraction of data from the web page using HTML elements like tags and attributes.
- **Transformation into required format:** Once you have the data, you can save it into your required format, like CSV.

Data sourcing is the very first step of EDA, and after getting the data into the required file types, we need to clean it up. In the next session, you will learn the end-to-end process of cleaning a data set with the help of a practical case study.

Module 2 : Data Cleaning

Introduction:-

Welcome to the next step in the process of EDA called 'Data Cleaning'.

In the last session, you learnt about the data sourcing techniques. Once you source the data, it is essential to get rid of the irregularities in the data and fix it to improve its quality.

One can encounter different kinds of issues in a dataset. Irregularities may appear in the form of **missing values, anomalies/outliers, incorrect format** and **inconsistent spelling**, etc. These irregularities may propagate further and affect the assumptions and analysis based on that dataset and may hamper the further process of machine learning model building. Hence, data cleaning is a very important step in EDA.

In this session

In this session, you will learn the process of data cleaning using a case study on **'Bank Marketing Campaign Dataset'**. Though data cleaning is often done in a somewhat haphazard manner, and it is difficult to define a 'single structured process', you will study data cleaning through the following steps:

1. Identifying the data types
2. Fixing the rows and columns
3. Imputing/removing missing values
4. Handling outliers
5. Standardising the values
6. Fixing invalid values
7. Filtering the data

Before going any further, it is important for you to get familiar with the problem statement that you are going to solve in this module to understand the EDA practically.

Problem statement
The bank provides financial services/products such as savings accounts, current accounts, debit cards, etc. to its customers. In order to increase its overall revenue, the bank conducts various marketing campaigns for its financial products such as credit cards, term deposits, loans, etc. These

campaigns are intended for the **bank’s existing customers.** However, the marketing campaigns need to be cost-efficient so that the bank not only increases their overall revenues but also the total profit. You need to apply your knowledge of EDA on the given dataset to analyse the patterns and provide inferences/solutions for the future marketing campaigns.

A bank conducted a telemarketing campaign for one of its financial products called ‘Term Deposits’ to help foster long-term relationships with existing customers. The dataset contains information about all the customers who were contacted during a particular year to open term deposit accounts with the bank.

What is a term deposit?
Term deposits, also called fixed deposits, are the cash investments made for a specific time period ranging from 1 month to 5 years for predetermined fixed interest rates. The fixed interest rates offered for term deposits are higher than the regular interest rates for savings accounts. The customers receive the total amount (investment plus the interest) at the end of the maturity period. Also, the money can only be withdrawn at the end of the maturity period. Withdrawing money before that will result in penalty charges, and the customer will not receive any interest returns.

Important Note:
To enhance the learning outcome, you are expected code along with the instructor as you watch the videos. So, please pace yourself accordingly. To assist you, you are provided with a structured and blank Python notebook to code. This is a **must-do** task for you to answer certain in-segment questions, as it serves the purpose of practice. Also, the final notebook will act as a reference for you in the future as well.
Please do not expect a complete solution notebook attached at the end of this module.

Data Types:-

Now, let’s talk about a very important aspect of any dataset, i.e., data types. In a particular dataset, you have multiple types of variables with different kinds of data types such as integers, string, floats, etc.

For data analysis, you will use the following libraries through the entire module, which you must have already covered in prep content:

- **Pandas**: It is a library to deal with dataframes in Python. Pandas is an acronym derived from panel data. It is solely used for data analysis purposes in Python.
- **NumPy**: This library is used for performing numerical operations on a dataset.

Now, let’s go through the bank marketing dataset along with Rahim and try to find out the data types that are present in it.

In general, any given data set is expected to have different types of data. Following are some examples with their data types.

Example	Variable Type	Data Type
Height, weight, age, temperature	Numerical variable	Int, float
Size of clothes, months, type of jobs, blood group.	Categorical variable	Object
Grades in exam, education level, months, integer ratings	Ordinal categorical type	Object, int, float
Date, time, timestamp	Date and time variable	Date and time

Fixing the Rows and Columns:-

You learnt about some of the issues with raw data and understood the need for data cleaning. Now, let's listen to Anand to understand different cases in fixing the columns and rows of a given dataset.

Now let’s summarise what you learnt with the help of the checklists below. Make sure you correctly identify these issues and resolve them before moving on to the next stage of data cleaning.

Checklist for fixing rows:

- Delete summary rows: Total and Subtotal rows
- Delete incorrect rows: Header row and footer row
- Delete extra rows: Column number, indicators, blank rows, page number

Checklist for fixing columns:

- if needed, merge columns for creating unique identifiers, for example, merge the columns State and City into the column Full Address.
- Split columns to get more data: Split the Address column to get State and City columns to analyse each separately.
- Add column names: Add column names if missing.
- Rename columns consistently: Abbreviations, encoded columns.
- Delete columns: Delete unnecessary columns.
- Align misaligned columns: The data set may have shifted columns, which you need to align correctly.

You have seen in the above video that both heading rows have been deleted, as they have no use in our analysis. It is very important to note here that if you find anything irregular at the very glance of the data set then it is very essential to get rid of that at the very first process.

Now you have learnt to fix the following columns:

- **Customerid**: It has been dropped, as it has no specific use in the analysis.
- **Jobedu**: It has been separated to extract job and education. Job and education have to be analysed separately. You will understand in further sessions how education and job play a very important role in determining the customer segment who will respond positively to term deposits.
- **Month**: The month name will be extracted in the further segments based on the missing values imputation analysis.

Impute/Remove Missing Values:-

You learnt how to fix columns and rows, and applied those learnings to the bank marketing dataset. Now, you will learn what missing values are and how they should be treated. Before working on the dataset, let’s listen to Anand as he explains the different methods to fix missing values in a dataset.

The most important takeaway from this lecture is: good methods add information, bad methods exaggerate information. In case you can add information from reliable external sources, you should use it to replace missing values. But often, it is better to let missing values be and continue with the analysis rather than extrapolate the available information.

Let’s summarise the takeaways from the above video:

- **Set values as missing values:** Identify values that indicate missing data, for example, treat blank strings "NA", "XX", "999", etc., as missing.
- **Adding is good, exaggerating is bad:** You should try to get information from reliable external sources as much as possible, but if you can’t, then it is better to retain missing values rather than exaggerating the existing rows/columns.
- **Delete rows and columns:** Rows can be deleted if the number of missing values is insignificant, as this would not impact the overall analysis results. Columns can be removed if the missing values are significant in number.

- **Fill partial missing values using business judgement:** Such values include missing time zones, century, etc. These values can be identified easily.

Following is a list of the major takeaways from the video.

Types of missing values:

- **MCAR**: It stands for Missing completely at random. The reason behind the missing value is not dependent on any other features.
- **MAR**: It stands for Missing at random. The reason behind the missing value may be associated with some other features.
- **MNAR**: It stands for Missing not at random. There is a specific reason behind the missing value.

Now, let’s apply all these concepts to the bank marketing campaign data set to tackle the issue of missing values in the age and month columns.

There are various ways to deal with missing values. Either you can drop the entries that are missing if you find that the percentage of missing values in a column is very small, or you can impute the missing values with some other values. Let’s look into the various ways to impute the missing values.

Imputation on categorical/numeric columns:

1. **Categorical column:**

- Impute the most popular category.
- Imputation can be done using logistic regression techniques.

2. **Numerical column:**

- Impute the missing value with mean/median/mode.
- The other methods to impute the missing values involve the use of interpolation, linear regression. These methods are useful for continuous numerical variables.

In this video, you will go through the analysis of the 'pdays' variable to deal with its missing values.

The major takeaway from the above video is that missing values do not always have to be null. So, now you must have a clear understanding of how to treat missing values in a dataset.

- Sometimes, it is good to just drop the missing values because they are missing completely at random.

- Sometimes, it is good to impute them with another value, maybe mean/median/mode, because they are not missing at random and have to be incorporated for further analysis.

You have gone through with the bank telemarketing data set. There is a 'response' variable which is basically the target variable of the data set. You have learnt about the missing values and the process to treat them. Based on your understanding of codes and process on missing values, answer the following questions.

Handling Outliers:-

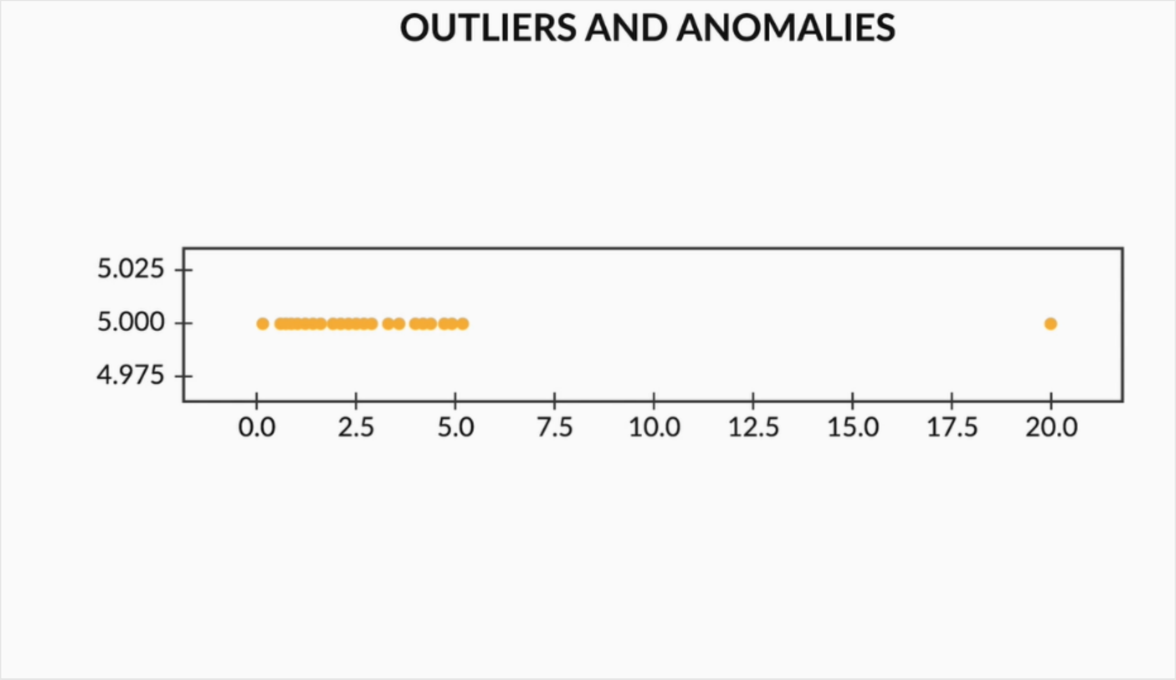
You have learnt what missing values are and how to treat them. Now, let’s move to the next concept of data cleaning, which is outliers.

The definition of outliers is as follows:
Outliers are values that are much beyond or far from the next nearest data points.

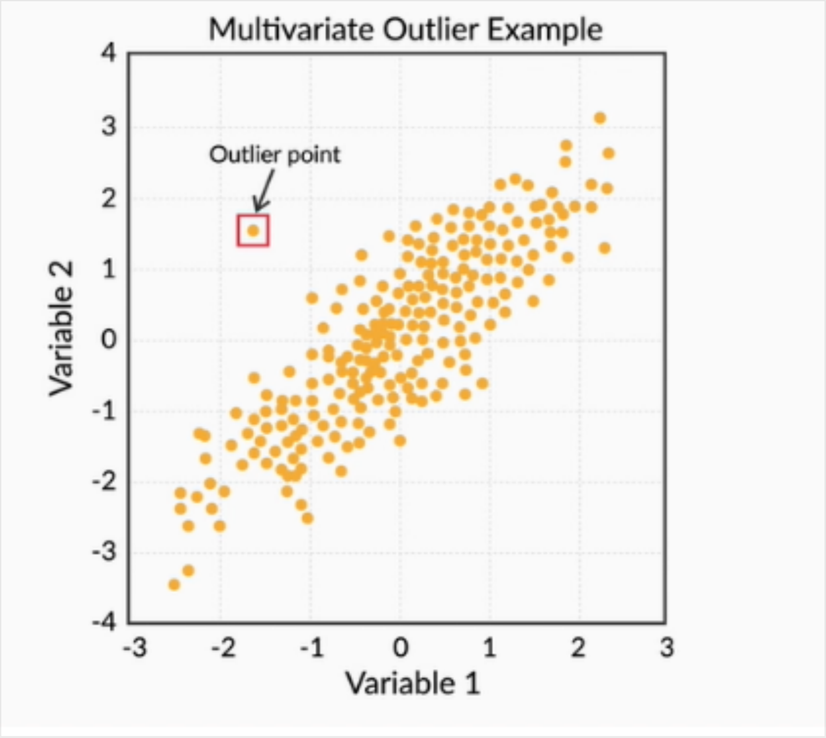
In this video, Rahim will help you understand the concept of outliers.

You learnt that there are two types of outliers. These are:

- **Univariate outliers:** Univariate outliers are those data points in a variable whose values lie beyond the range of expected values. You can get a better understanding of univariate outliers from the image below. Here, almost all the points lie between 0 and 5.0, and one point is extremely far away (at 20.0) from the normal norms of this data set.



- **Multivariate outliers:** While plotting data, some values of one variable may not lie beyond the expected range, but when you plot the data with some other variable, these values may lie far from the expected value. These are called multivariate outliers. You can refer to the image below to get a better understanding of multivariate outliers.



Now, let’s proceed to the next video, where you will learn about the reasons behind the appearance of outliers in data and how to

treat them.

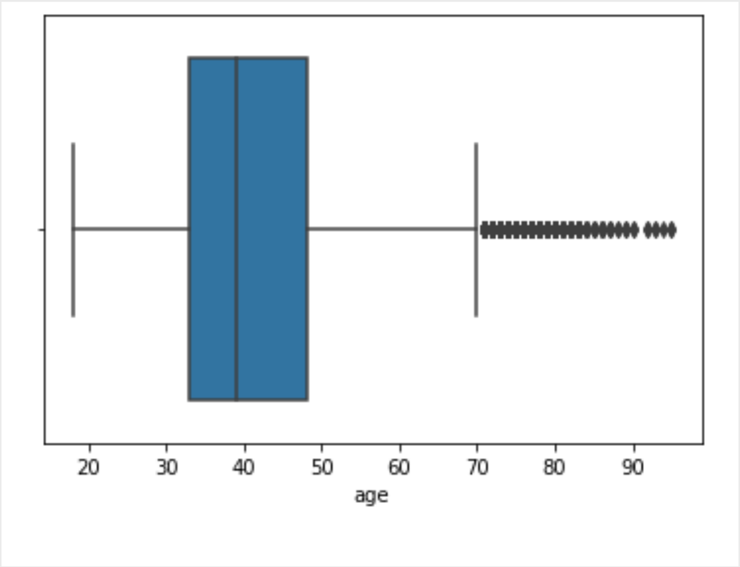
From the video, you must have understood that outliers should be treated before investigating data and drawing insights from a dataset.

Now, the major approaches to the treatment of outliers can include:

- Imputation
- Deletion of outliers
- Binning of values
- Capping the outliers

In the process of handling missing values and outliers of different columns, you are already performing univariate analysis. You will learn more about it in further sessions. In this video, you will learn how to implement all your learning on the bank marketing dataset.

So, in the above video, you have seen that the age variable has outliers, but these can be treated as the normal values of age because any person can be over 70 or 80 years of age. Also, the 70-90 age group is sparsely populated and participate in opening the term deposit account, which is why these set of people fall out of the box plot but they are not outliers and can be considered as normal values.



Let's listen to Rahim as he explains the variable 'balance'.

An important aspect that has been covered in this video is **quantiles**. Sometimes, it is beneficial if you look into the quantiles instead of the box plot, mean or median. Quantile may give you a fair idea about the outliers. If there is a huge difference between the maximum value and the 95th or 99th quantiles, then there are outliers in the data set.

In the next segment, you will learn about the standardisation process in EDA.

Standardising Values:-

You learnt different techniques to handle outliers and also implemented the same in the 'bank marketing' dataset. Now, you will learn the next important aspect, which is to standardise values in a dataset.

In this video, Anand will explain how to standardise quantitative values in a dataset.

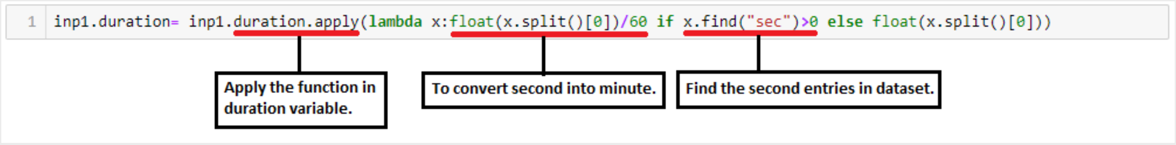
Scaling ensures that the values in a dataset have a common scale; this makes it easy to perform data analysis. Let's take a data set that contains the grades of students studying in different universities. Some of the universities assign grades on a scale of 4, whereas the others assign grades on a scale of 10. Hence, you cannot assume that a GPA of 3 on a scale of 4 is equal to a GPA of 3 on a scale of 10, even though they are the same quantitatively. Thus, for the purpose of analysis, these values need to be brought to a common scale, such as the percentage scale.

Now, let’s summarise what you learnt so far about standardising the variables in a dataset. Given below is a list of the points that we covered. You could use this as a checklist for future data cleaning exercises:

- **Standardise units:** Ensure all observations under one variable are expressed in a common and consistent unit, e.g., convert lbs to kg, miles/hr to km/hr, etc.
- **Scale values if required:** Make sure all the observations under one variable have a common scale.
- **Standardise precision** for a better presentation of data, e.g., change 4.5312341 kg to 4.53 kg.

Now that you have learnt how to standardise the numeric values in a data set, let's proceed to learn how to standardise text values, which is an equally important aspect of data analysis.

In the videos, you saw the application of standardisation with a real-life example of the ‘duration’ variable in the 'bank marketing' data set. The duration variable has data in minutes as well as in seconds, which has to be converted into minute only. You can understand the entire code to convert the 'duration' variable into minutes in the image below.



```
inp1.duration= inp1.duration.apply(lambda x:float(x.split()[0])/60 if x.find("sec")>0 else float(x.split()[0]))
```

Apply the function in duration variable.
To convert second into minute.
Find the second entries in dataset.

(**Note**: Please open image in new tab to view zoomed image)

In the next segment, you will learn how to fix invalid values in a data set.

Fixing Invalid Values and Filter Data:-

In the previous segments, you learnt the concepts to deal with different kinds of irregularities in a data set. You also went through bank marketing dataset, where you saw the practical aspect of all the concepts covered. Datasets also have some other irregularities, which you need to get rid of. Though our bank marketing dataset does not have these kinds of irregularities, it is essential to deal with these as well.

Let's watch the next video to gain more insight into fixing invalid values in a data set.

If your data set has invalid values, and if you do not know which accurate values could replace the invalid values, then it is recommended that you treat these values as missing. For example, if the Contacts columns in a data set contain a string 'tr8ml', then it is recommended that you remove the invalid value and treat it as a missing value.

Now, let's summarise what you learnt about fixing invalid values in a data set. Given below is a list of points that we covered. You could use this as a checklist for future data cleaning exercises:

- **Encode unicode properly**: In case the data is being read as junk characters, try to change the encoding, for example, use CP1252 instead of UTF-8.
- **Convert incorrect data types**: Change the incorrect data types to the correct data types for ease of analysis. For example, if numeric values are stored as strings, then it would not be possible to calculate metrics such as mean, median, etc. Some of the common data type corrections include changing a string to a number ("12,300" to "12300"), a string to a date ("2013-Aug" to "2013/08"), a number to a string ("PIN Code 110001" to "110001"), etc.
- **Correct the values that lie beyond the range**: If some values lie beyond the logical range, for example, temperature less than -273° C (0° K), then you would need to correct those values as required. A close look at the data set would help you determine whether there is

scope for correction or the value needs to be removed.

- **Correct the values not belonging in the list**: Remove the values that do not belong to a list. For example, in a data set of blood groups of individuals, strings 'E' or 'F' are invalid values and can be removed.
- **Fix incorrect structure**: Values that don't follow a defined structure can be removed from a data set. For example, in a data set containing the pin codes of Indian cities, a pin code of 12 digits would be an invalid value and would need to be removed. Similarly, a phone number of 12 digits would be an invalid value.
- **Validate internal rules**: Internal rules, if present, should be correct and consistent. For example, the date of a product's delivery should definitely come after the date of purchase.

After you have fixed the missing values, standardised the existing values and corrected the invalid values in a data set, you would arrive at the last stage of data cleaning. Although you have a largely accurate dataset by now, you may not need all of it for your analysis. It is important for you to understand what you need in order to draw insights from the data, and then choose relevant parts of the dataset for your analysis. Thus, you need to filter the data in order to get what you need for your analysis.

Let's watch Anand as he takes us through the various steps of data filtering in the next video.

Now, let's summarise what you learnt about filtering data. Given below is a list of the points we covered. You could use this as a checklist for future data cleaning exercises:

- **Deduplicate data:** Remove identical rows and the rows in which some columns are identical.
- **Filter rows**: Filter rows by segment and date period to obtain only rows relevant to the analysis.
- **Filter columns**: Filter columns relevant to the analysis.
- **Aggregate data**: Group by the required keys and aggregate the rest.

Summary:-

Having completed this session, you must be clear about the various irregularities that can be present in a data set. They can be unfixed rows/columns, missing values, outliers or may even be in the form of un-standard/un-scaled data, etc.

Let's summarise the steps in Data Cleaning:

- **Fixing the rows and columns:** You need to remove the irrelevant columns and heading lines from the dataset. The irrelevant columns or rows are those that are of absolutely no use for analysis on the data set. Like in the Bank Marketing Dataset, the headers and customer ID columns are of absolutely no use.

- **Remove/impute the missing values:** There are different types of missing values in the dataset. Based on their type and origin, you need to take a decision regarding whether they can be removed if their percentage is too low, or whether they can be considered as a separate category. There is an important possibility where you need to impute missing values with some other value. While doing imputation, one should be very careful because it should not add any wrong information into the dataset. The imputation can be done using mean, median, mode or using quantile analysis.
- **Outlier handling:** Outliers are those points which are beyond the normal trend. There are two types of outliers:
  1. **Univariate**
  2. **Multivariate**

An important aspect that has been covered is that outliers should not always be treated as anomalies in the dataset. You can understand this using the Bank Marketing Dataset itself, where age has outliers, but these high values of age are as relevant as other values.

- **Standardising values:** Sometimes, there are many entries in the dataset which are not in the correct format. Like you have seen in the Bank Marketing dataset itself, the duration of the call is in seconds and minutes. It has to be in the same format. The other standardisation involves the unit and precision standardisation.
- **Fixing invalid values:** Sometimes, there are some values in the dataset that are invalid, maybe in the form of their unit, range, data type, format, etc. It is essential to deal with these types of irregularities before processing the dataset.
- **Filter data:** Sometimes, filtering out certain details can help you get a clearer picture of the dataset.

It is very important to get rid of such irregularities to be able to analyse a dataset. Otherwise, it may hamper further analysis of the dataset, either while building a machine learning model or in EDA itself.

Now that you are done with the process of data cleaning, the next important step is data analysis. This is covered in the following

two sessions:

- Univariate analysis
- Bivariate/multivariate analysis.

Module 3 : Univariate Analysis

Introduction to Univariate Analysis:-

Now that you have a clear idea about the process of cleaning a dataset, the next step in Exploratory Data Analysis (EDA) is to learn univariate analysis. It deals with analysing a single column/variable. Let's hear our expert Rahim, as he explains the concept of univariate analysis.
The major takeaway from the video above is that univariate analysis is nothing new to us; you have performed this step on numerical variables while handling missing values and outliers.

In this session

You will learn about univariate analysis, which broadly is of the following four types:

- Categorical unordered univariate analysis
- Categorical ordered univariate analysis
- Statistics on the numerical variable
- Numerical variable univariate analysis

Categorical Unordered Univariate Analysis:-

Univariate analysis involves the analysis of a single variable at a time. The concept of univariate analysis is divided into **ordered** and **unordered** category of variables. In this segment, you will learn how to conduct univariate analysis on **categorical unordered variables.**

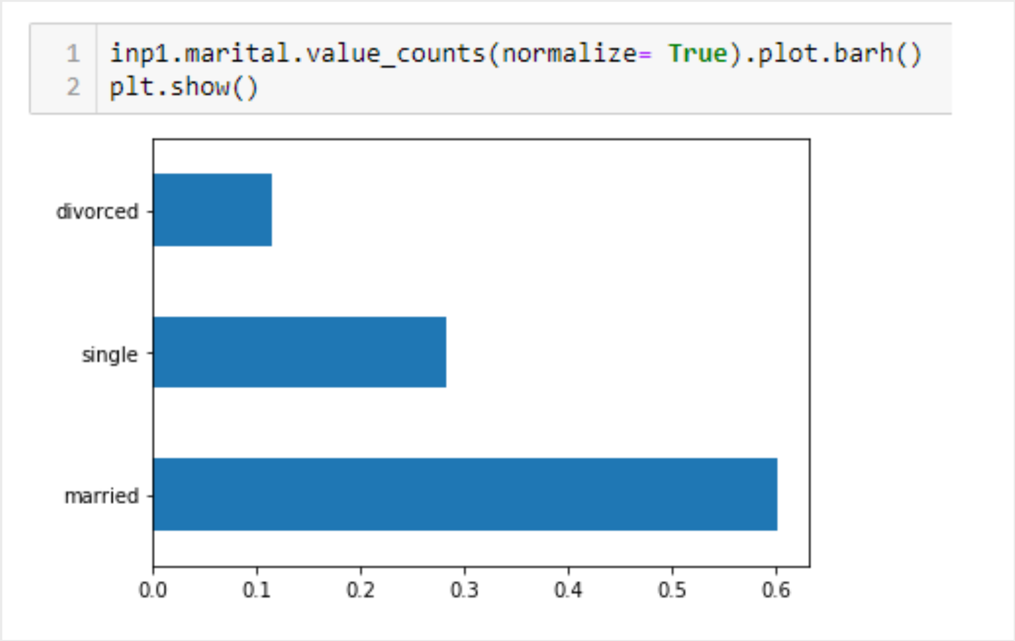
Unordered data is the type of data that does not have any measurable terms (measurable terms can be like high-low, more-less, fail-pass, etc.) Example:

- The type of loan taken by an individual (home loan, personal loan, auto loan, etc.) does not have any ordered notion. They are just different types of loans.
- Departments of an organisation — Sales, Marketing, HR — are different departments in an organization, with no measurable attribute attached to any term.

Unordered variables also called **Nominal** variables.

An unordered variable, primarily, is a categorical variable that has no defined order. Let's consider the example of the **job** column in the Bank Marketing dataset. There, '**job**' is divided into many sub-categories like technician, blue-collar, services, management, etc. There is no weight or measure given to any value in the '**job**' column.

You can see from the above video that married people have been contacted the most by the bank. This can be visualised in Python using the following graph.



```
inp1.marital.value_counts(normalize= True).plot.barh()
plt.show()
```

In bivariate analysis, when variables like marital status, job and education will be plotted with response variables, then you will be in a position to decide which categories in respective columns have the highest chances of a positive reply.

In the next segment, you will learn about categorical ordered univariate analysis.

Categorical Ordered Univariate Analysis:-

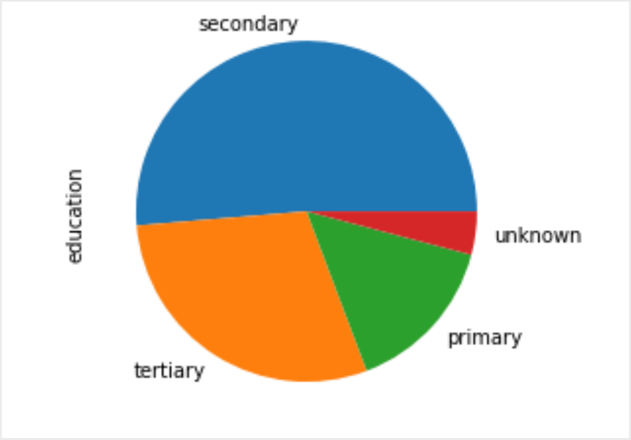
Ordered variables are those variables that follow a natural rank of order. Some examples of categorical ordered variables from the Bank Marketing dataset are:

- Age group: <30, 30-40, 40-50 and so on
- Month: Jan, Feb, Mar, etc.
- Education: primary, secondary and so on

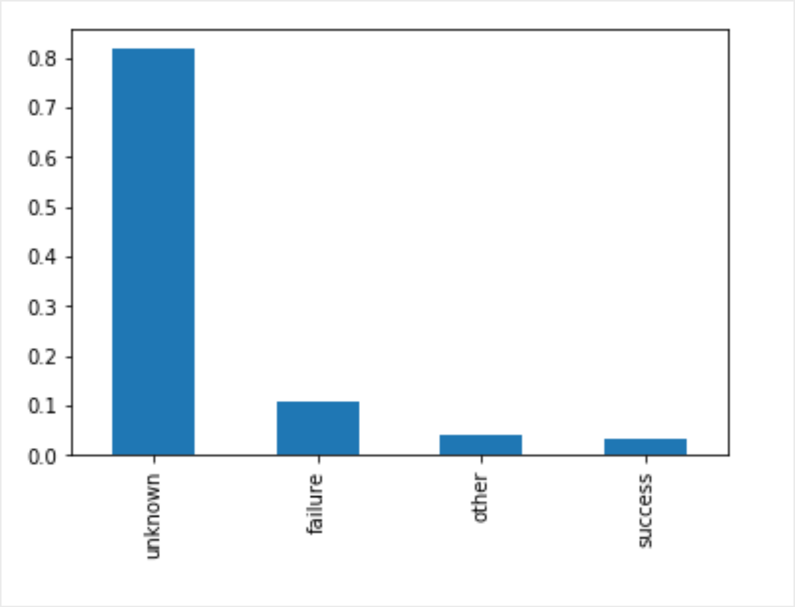
There are other ordered variables in the Bank Marketing data set as well. Let's perform order univariate analysis on that dataset. At the very beginning of this video, you will see the job variable bar graph, which you have already covered in the previous segments.

Let’s summarise the major takeaways from the above video:

- You have seen that *education*, *poutcome* and *response* are the ordered categorical variables.
- The bank has primarily contacted those customers who have completed their secondary education. You can observe that in the pie chart below:



- For the majority of the customers, the previous campaign has not been conducted. Refer to the bar graph below to understand more about the '*poutcome'* variable. As you can see, 'unknown' has the major share within the '*poutcome*' variable.



Transition of a numerical variable into an ordered categorical variable

Let’s consider a very interesting example of your school life. Suppose you have a dataset containing the marks of all the students in the 'Science' subject, and you are one of the students in that group. These marks can be considered as categorical if you divide the total marks into different categories like High, Medium, Average, Below Average, Poor. From this analysis, you can determine your ranking in the class and also find out how many students got more marks than you and how far away your score is from the **mean** or the **average** score.

The important thing to note here is that your marks are a numerical variable, which you have then categorised into 'high marks' and 'low marks'. This is an approach that you will need to adopt in the future, and you will learn more about this approach in the next segment on numerical variable analysis.

In the next video, you will understand the basics of statistics and its applications in real-life examples.

Statistics on Numerical Features:-

You have seen how to conduct univariate analysis on categorical variables. Now, let's look at quantitative or numeric variables.

Numeric variables can be continuous like height, temperature, weight, etc. Numerical variables can also be discrete like the number of items bought by a customer in a store, the number of people in a city, the number of 'heads' you get when flipping three coins.

In this segment, our expert Anand will take you through various statistical metrics such as mean, median, mode and standard deviation.

Let’s now learn how to analyse quantitative variables.

Mean and median are single values that broadly give a representation of the entire data. As Anand states clearly, it is very important to understand when to use these metrics to avoid inaccurate analysis.

While '**mean**' gives an average of all the values, the '**median**' gives a typical value that can be used to represent the entire group. As a simple rule of thumb, always question someone if they use 'mean' since 'median' is primarily a better measure of ‘representativeness’.

Let’s now look at some other descriptive statistics such as mode, interquartile distance, standard deviation, etc.

Both standard deviation and interquartile difference are used to represent the spread of the data.

The interquartile difference is a much better metric than standard deviation if there are **outliers** in the data because the standard deviation will be influenced by outliers, while the interquartile difference will simply ignore them.

You also saw how box plots are used to understand the spread of data.

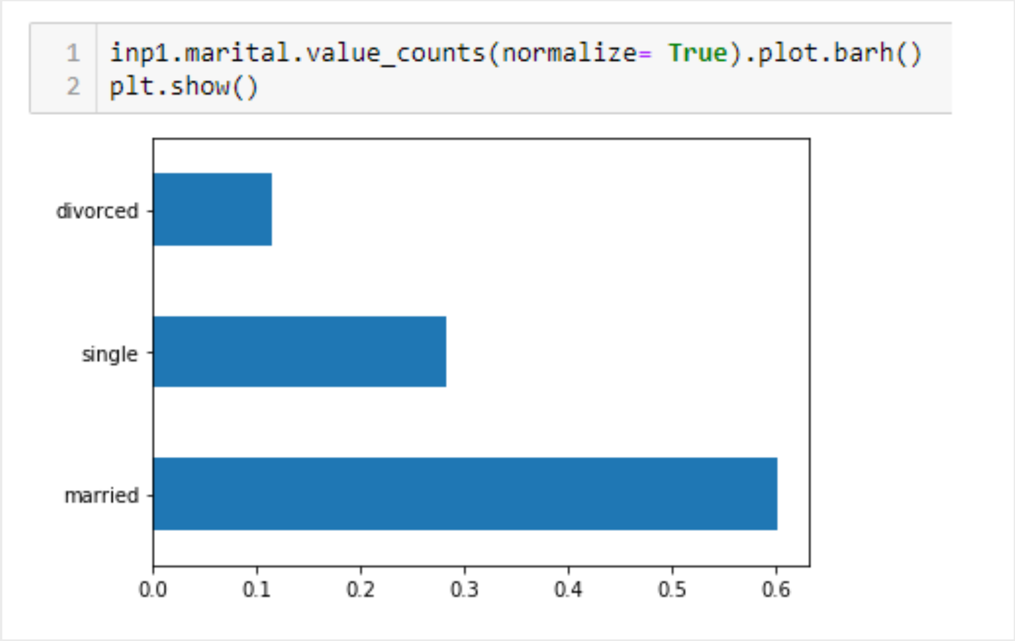
Summary:-

Univariate analysis is the analysis of a single variable at a time. This particular variable can be ordered or unordered, or it may be a numerical variable. So, based on the types of variables, the whole understanding of univariate analysis is divided into the following parts:

- **Categorical unordered univariate analysis**: Unordered variables are those variables that do not contain any notion of ordering, for example, increasing or decreasing order. These are just various types of any category. The examples can be job types, marital status, blood groups, etc.
- **Categorical ordered univariate analysis**: Ordered variables are those that have some kind of ordering in them, like high-low, fail-success, yes or no. Examples can be education level, salary group like high or low, gradings in any exam, etc.
- **Numerical variable univariate analysis**: Numerical variables can be classified into continuous and discrete type. To analyse numerical variables, you need to have an understanding of statistic metrics like mean, median, mode, quantiles, and box plots, etc. It is important to understand that numerical variable univariate analysis is nothing but what we have done earlier, i.e., the treatment of missing values and handling outliers. The crux of univariate analysis lies in the single variable analysis, which is covered in the process of cleaning the dataset.
- **The transition of a numerical variable into a categorical variable:** This is an important aspect that you need to think about before performing univariate analysis. Sometimes, it is essential to just convert numerical variables into categorical ones, through a process which is called 'binning'.

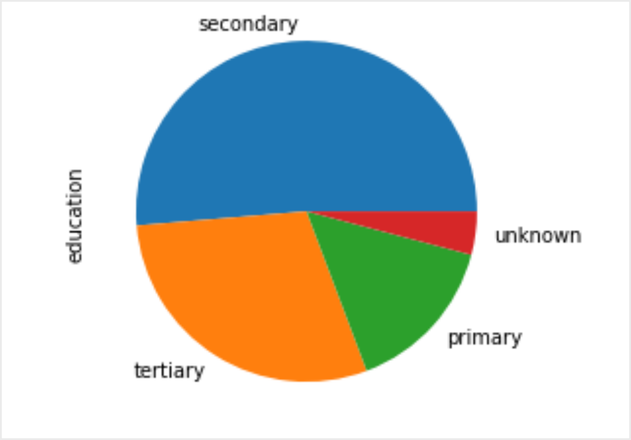
Let's summarise univariate analysis on the Bank Marketing Campaign dataset.

- You have seen that there is a variable called "**marital**" in the Bank Marketing dataset. This is **categorical unordered variable**. You have seen that the bank has contacted mostly married people, as can be seen in the image below.



```
inp1.marital.value_counts(normalize= True).plot.barh()
plt.show()
```

- There is a variable called "**education**" in the Bank Marketing dataset. This is a **categorical ordered variable** because there is ordering of education levels, like primary, secondary and tertiary education. You have seen that the bank has mostly contacted people who have completed secondary education, as can be seen in the image below.



- You have already performed univariate analysis on numerical variables in the process of missing values treatment and handling outliers. You have seen that there are no outliers in the "**age**" variable, as the values of age like 80 or 90 are also genuine values. There are higher values in 'balance' and 'salary' variables, which can be treated as outliers. Hence, it can be avoided while performing the analysis.

Hence, univariate analysis is nothing but an analysis of one particular variable at a time. It is important to look at each and every variable and perform analysis on it.

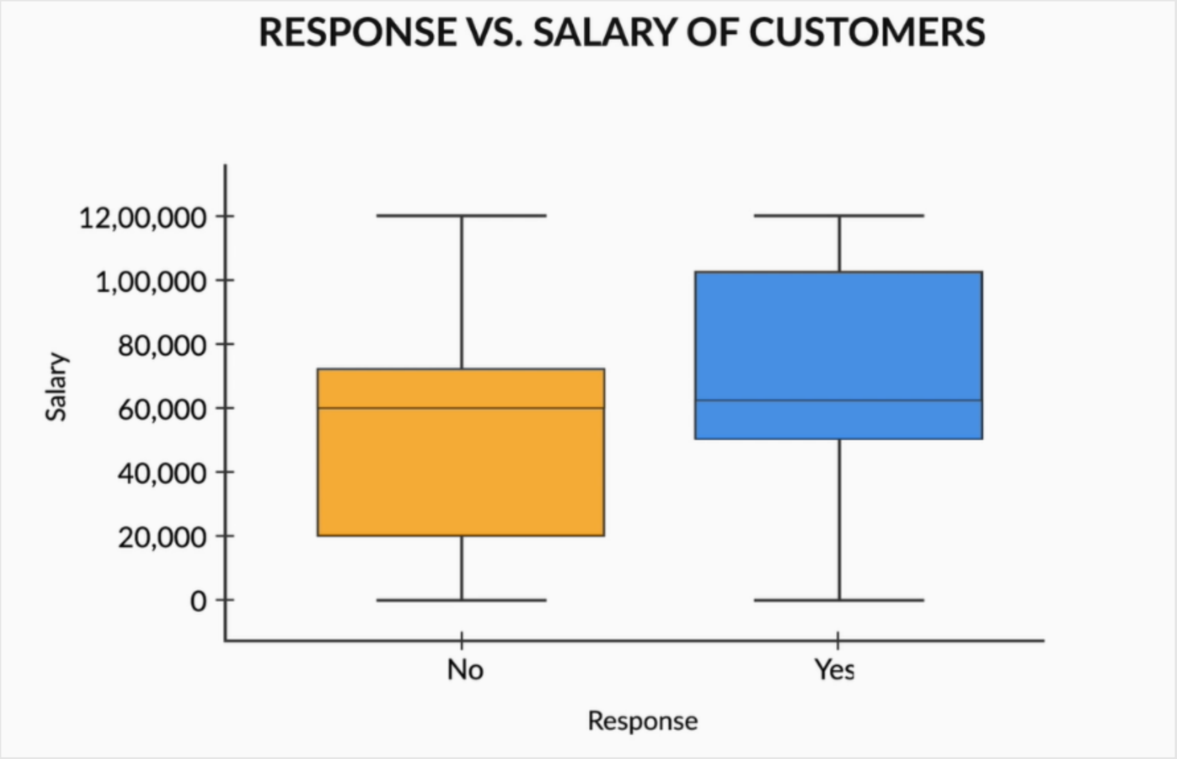
Module 4 : Bivariate and Multivariate Analysis

Introduction:-

Welcome to the session on **'Bivariate and Multivariate Analysis'**.

So far, you have learnt how to preprocess and clean data, and then you performed a univariate analysis on the bank marketing dataset. A univariate analysis includes the analysis of individual categorical variables like job, education, response, marital status, etc. It also explains the concept of outliers, and mean, median or mode of numerical variables such as salary, balance and age.

Now, consider the following graph, which plots the 'Response' and the 'Salary' columns from the bank marketing dataset. You have already plotted this graph in the previous segments. Although the median , maximum and minimum values are the same, customers with a higher salary range show interest in opening term deposit accounts with the bank. This is nothing but a bivariate analysis, that is, the analysis of two variables/columns in the data set.



In this session

In this session, you will learn how to analyse two or more variables at a time. You will also observe and draw better inferences on the types of customers that are showing interest in opening term deposit accounts with the bank. This will give you better insights and understanding of how to conduct effective marketing campaigns in the future.

You have already learnt about the types of variables, that is, categorical and numeric variables, in the previous sessions. This session has been divided into the following topics based on the different types of variables:

- Analysis Between Two Numeric Variables
- Analysis Between Numeric and Categorical Variables
- Correlation Versus Causation
- Analysis Between Two Categorical Variables
- Multivariate Analysis

Numeric - Numeric Analysis:-

In this segment, you will learn how to analyse two numerical variables using the Bank marketing dataset. Now, there are multiple tools to analyse numerical variables. In the next video, you will learn about the different tools and plots that are helpful for extracting insights using numerical variables from a data set.

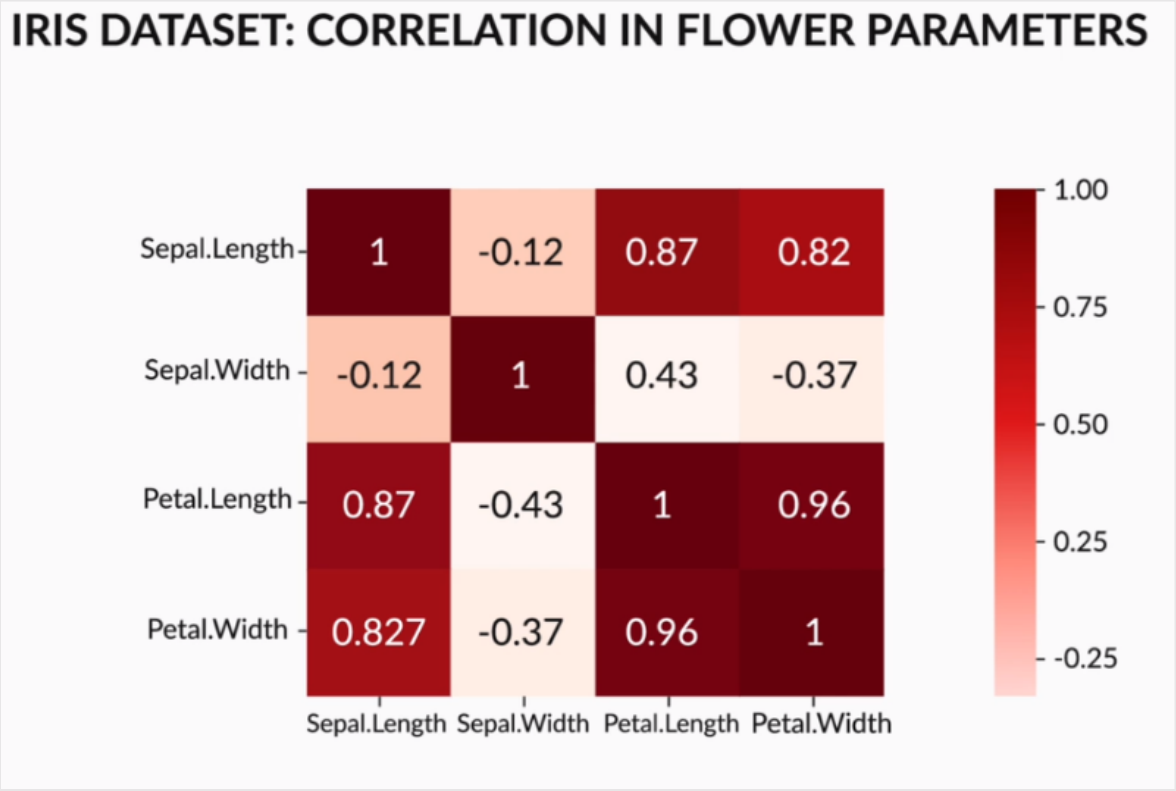
Note: In the video, at 3:08 mins, the value of the correlation between Petal Length and Sepal Width is -0.43 and not 0.43 at both the places of the correlation matrix.

One very important concept that has been covered in the video above is that of correlation coefficient. Now, correlation coefficient depicts only a linear relationship between numerical variables. It does not depict any other relationship between variables. A zero correlation does not imply that there is no relation between variables; it merely indicates that there will no linear relationship between them. Also, there can be a negative or positive correlation between variables. A negative correlation means that if the value of one variable increases, the value of another decreases, whereas it is the opposite for a positive correlation.

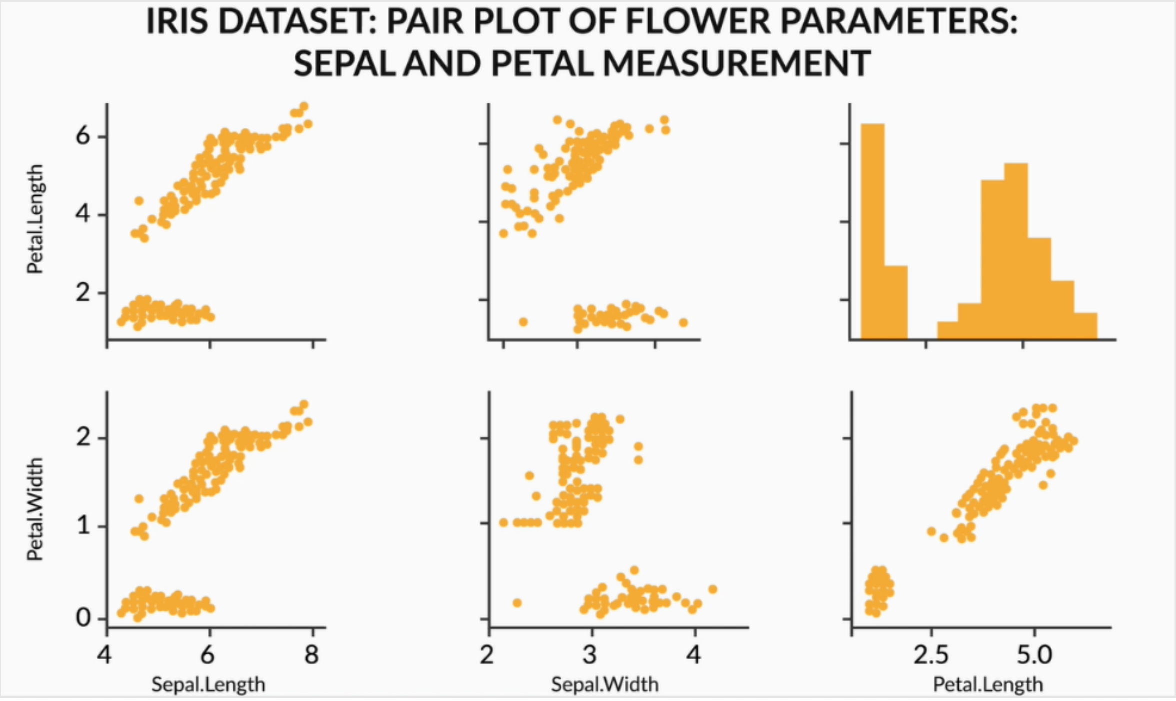
Now, the higher the coefficient of correlation between numerical variables, the higher the linear relation between them.

From the **correlation matrix** below, you can observe that petal length has a high correlation with sepal length, with a correlation coefficient of 0.87. Also, there is a very high correlation coefficient of 0.96 between petal width and petal length.

Note: The value of the correlation between Petal Length and Sepal Width is -0.43 and not 0.43 at both the places of the below correlation matrix.

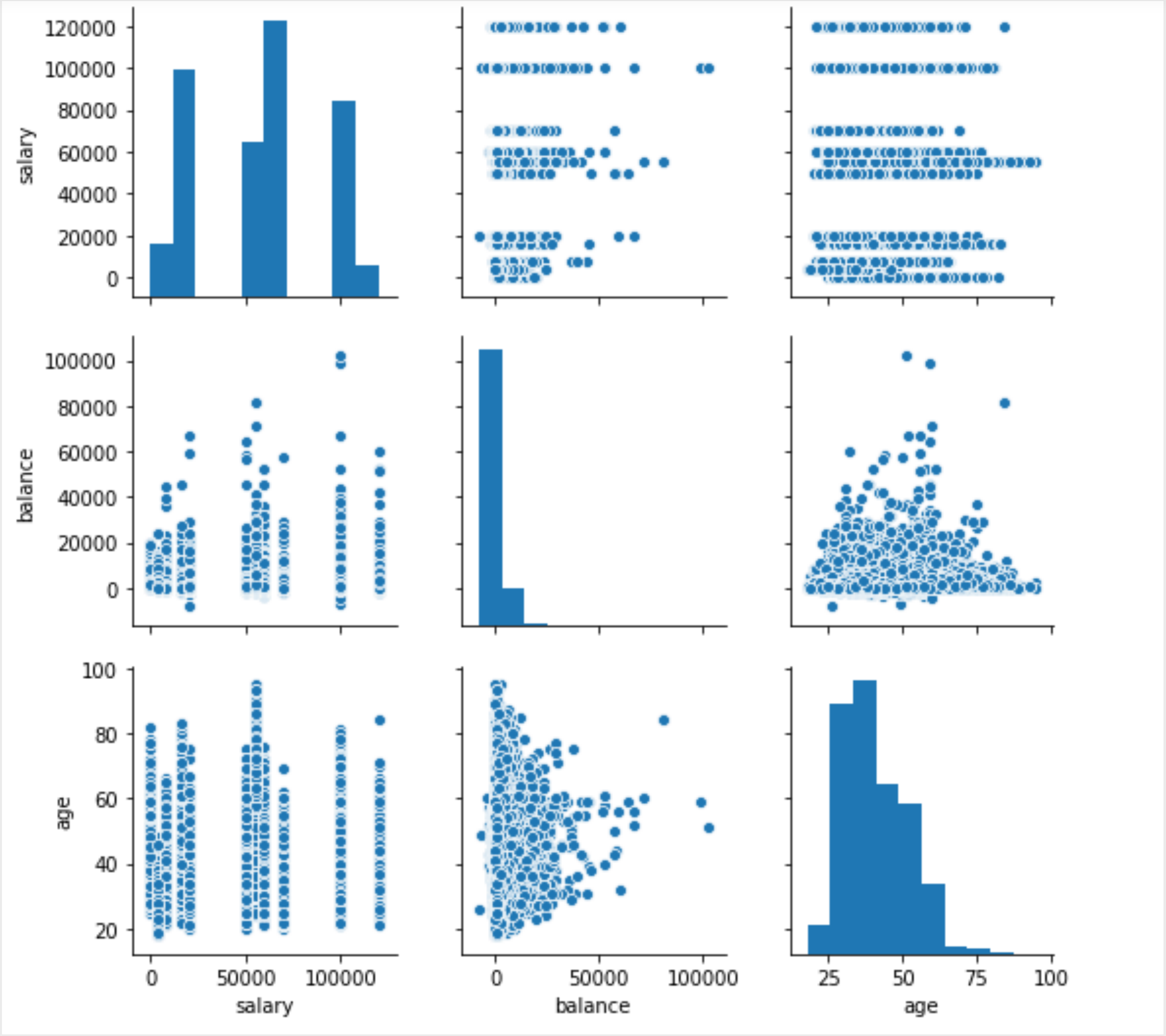


However, the correlation matrix has its own limitations where you cannot see the exact distribution of a variable with another numeric variable. To solve this problem, we use **pair plots**. Pair plots are scatter plots of all numeric variables in a data set. It shows the exact variation of one variable with respect to the others. You can observe how one variable is varying with respect to another in the image below.



Now, in the following video, Rahim will explain how to perform a numeric bivariate analysis using the bank marketing dataset.

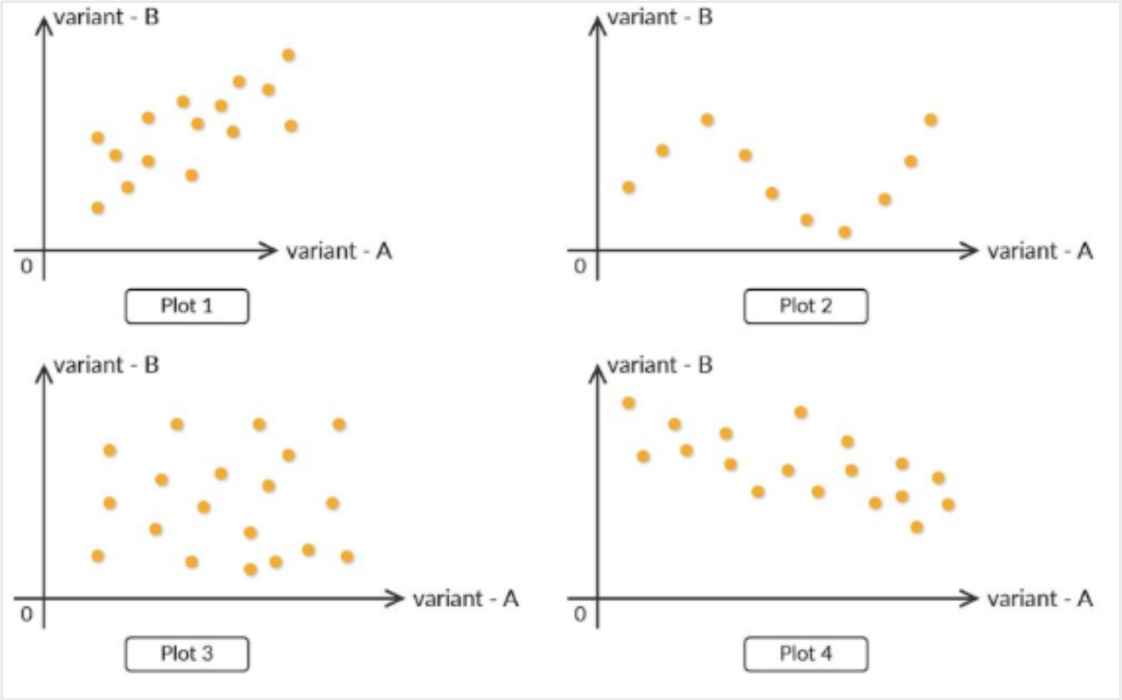
So, in the video, you saw how a pair plot can help you determine that there is no correlation between the 'age', 'balance' and 'salary' variables. Now, refer to the image below and observe how there is no correlation between these variables.



A high correlation coefficient does not imply that there will be a correlation with another numeric variable every time because there can be no causation between them. There may be cases where you will see a high correlation coefficient between two variables but there is no relation between them. You will understand this in detail in the next segment that how correlation is related to the causation.

Comprehension: Correlation

Consider the following four scatter plots of two variables A and B.



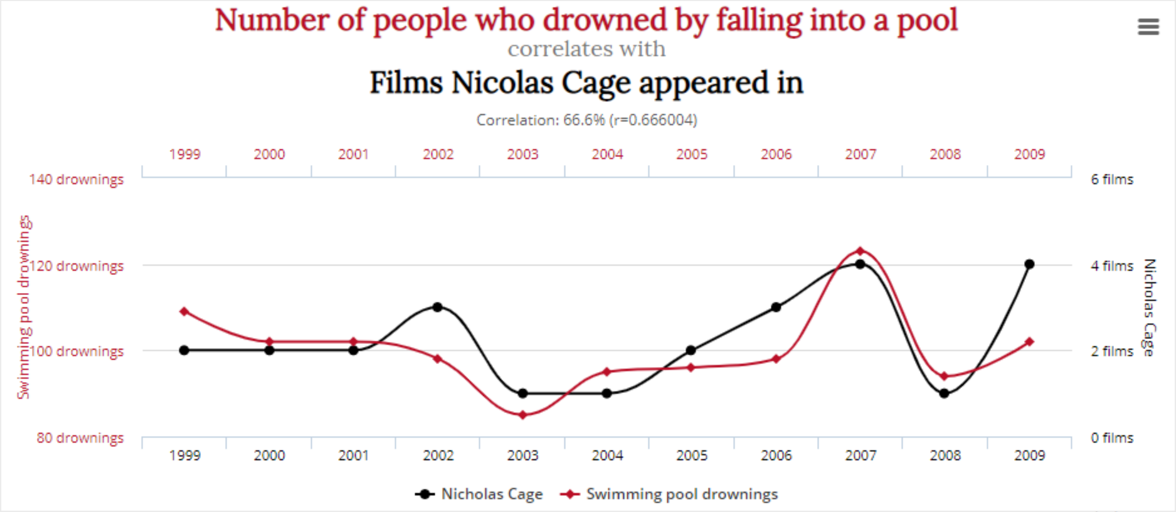
Based on your learning in this segment, answer the following questions.

Correlation vs Causation:-

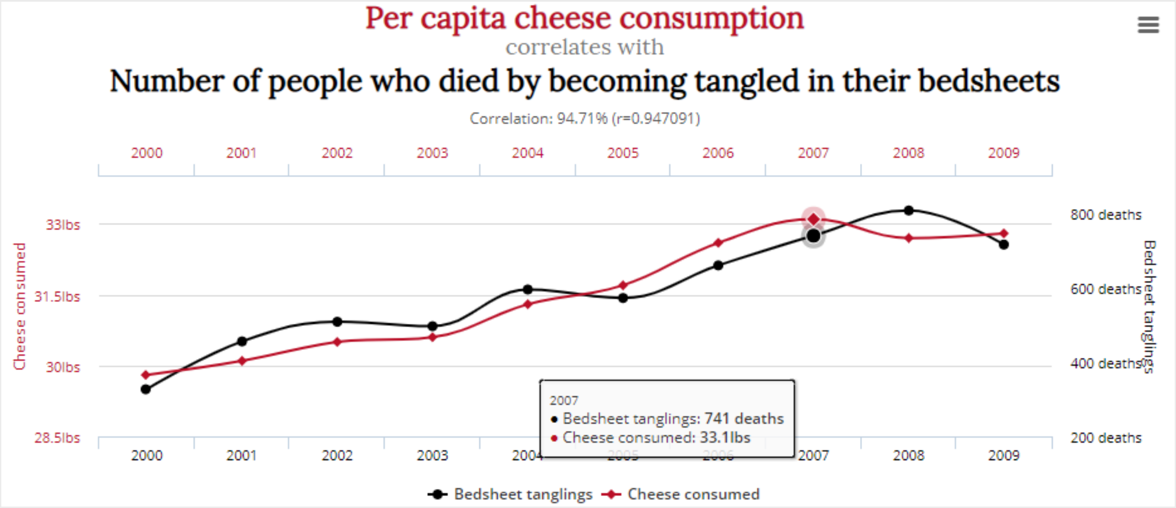
In the EDA exercise, it is very important to note that although some numerical variables can sometimes be highly correlated to each other, there may not be a cause of any relationship between them.

Let's first listen to Rahim in the next video and try to get a holistic picture of correlation among variables using some compelling examples.

So, the major takeaway from the video is that correlation does not imply causation. In the video, you saw that the number of people who drowned by falling into a pool is not related to movies starring Nicolas Cage. However, if you observe the plot below, you will notice that there is a very high correlation between them, as both the plots follow almost the same path.



Now, in the example below, it is quite obvious that the per capita cheese consumption has no relation with people dying from being tangled in bed sheets, although the plot shows a high relation between them.



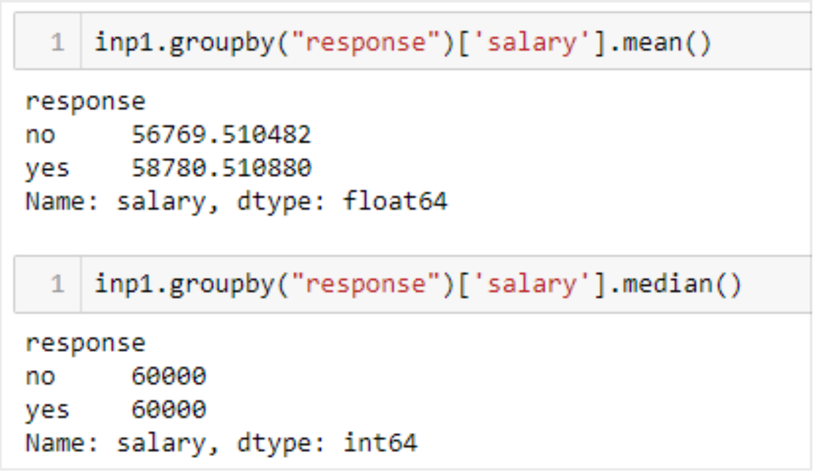
For more such compelling examples, where causation and correlation are not related to each other, you can refer to this link.

In this way, you have now a clearer idea that how causation is different from correlation. In the next segment, you will learn about bivariate analysis using numerical and categorical variables.

Numerical - Categorical Analysis:-

Previously, you learnt about the bivariate analysis of numerical variables. In this segment, you will learn about the associations between numerical and categorical variables. You will learn how to apply this analysis on the same bank marketing dataset.

So, in the video, you saw how the salary variable is varying with respect to the response variable. Their mean and median are the same, as shown in the image below.



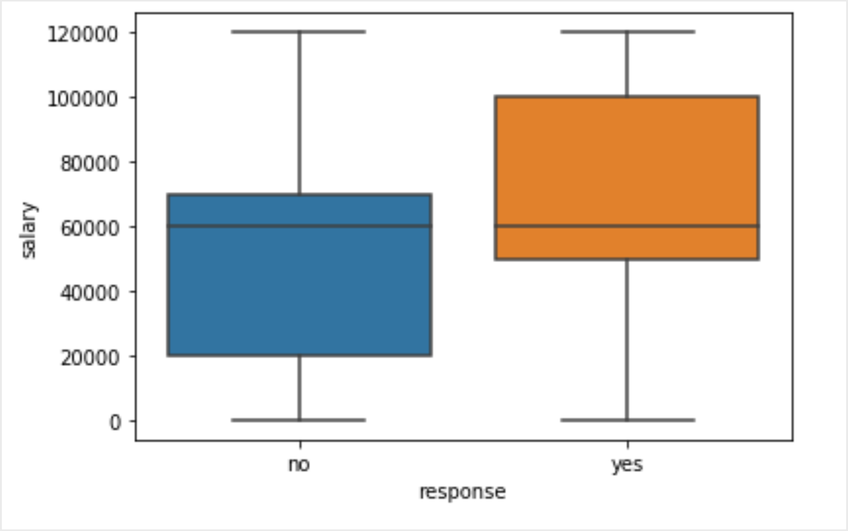
```
inp1.groupby("response")['salary'].mean()
```

```
response
no     56769.510482
yes    58780.510880
Name: salary, dtype: float64
```

```
inp1.groupby("response")['salary'].median()
```

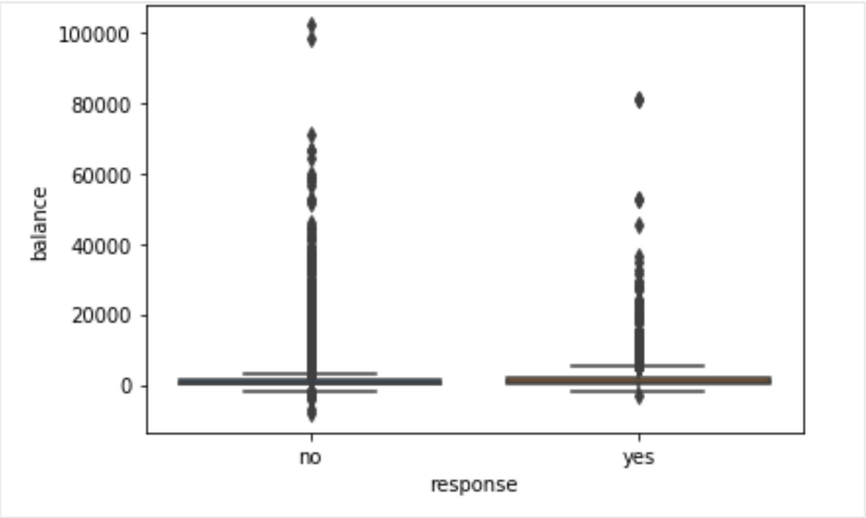
```
response
no     60000
yes    60000
Name: salary, dtype: int64
```

However, a very different picture emerges when you plot a boxplot. The interquartile range for customers who gave a positive response is on the higher salary side. This is actually true, because people who have higher salaries are more likely to invest in term deposits.



Now, in the next video, we will take a look at a different variable in the bank marketing dataset.

In the video, you observed that after the balance versus response graph is plotted, it does not make any sense at first glance. Sometimes only a boxplot is not sufficient to draw insights, because of a high concentration of data and or because of higher values in the data set, for example, the balance variable.



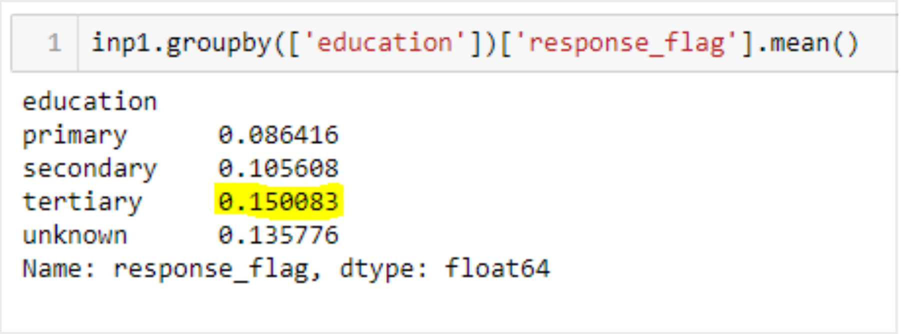
In such cases, it is a good practice to analyse the data using mean, median or quartiles. In the video, you saw that the mean and median values of the balance variable are higher for customers who gave a positive response, which is again true, because people who have higher balance in their bank accounts are more likely to invest in term deposits.

In the next segment, you will get an idea about categorical versus categorical variable analysis.

Categorical - Categorical Analysis:-

In this segment, you will learn about the associations between two categorical variables in a bivariate analysis. Statistical analysis is essential for numerical variables, and it includes different metrics like mean, median, mode, quantiles and boxplots. Here, you will learn how to analyse categorical variables using graphs and charts, and derive maximum insights from them.

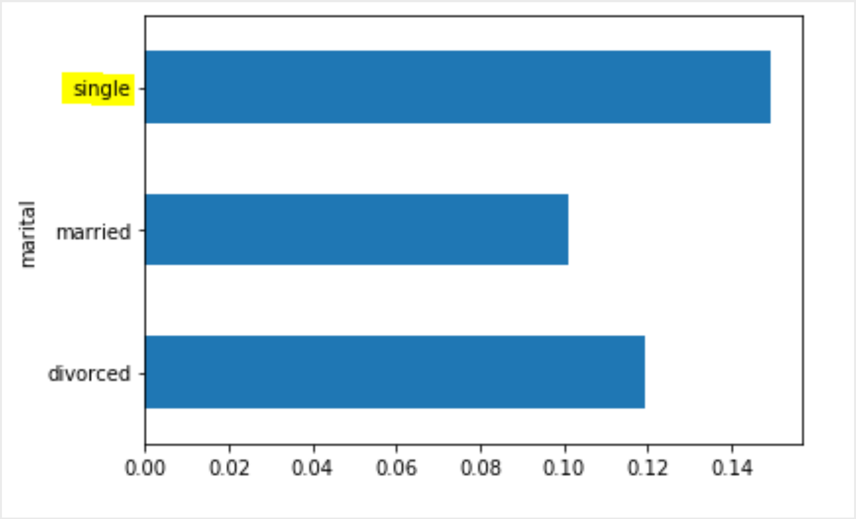
In the video, you saw that the positive response of customers to opening a term deposit with the bank increases with the education level. From this, you can infer that the bank should contact people with higher education levels to effectively increase the positive response for opening a term deposit.



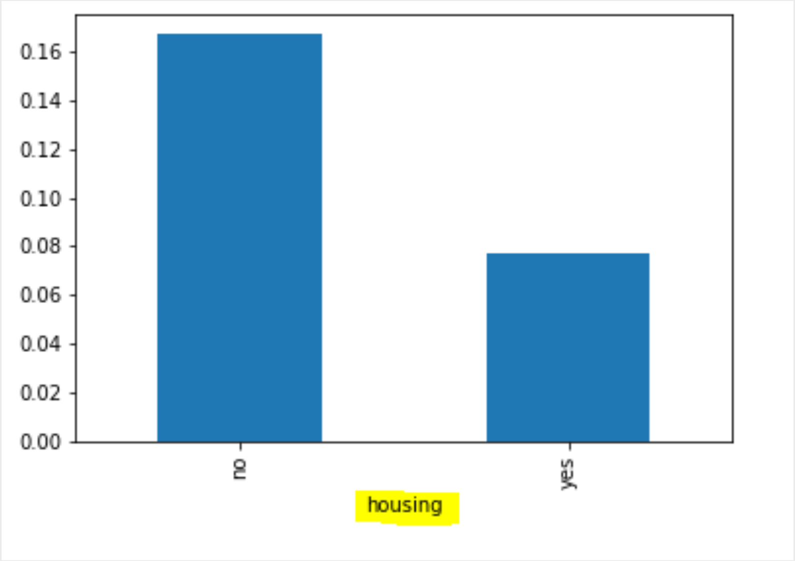
```
inp1.groupby(['education'])['response_flag'].mean()
```

```
education
primary      0.086416
secondary    0.105608
tertiary     0.150083
unknown      0.135776
Name: response_flag, dtype: float64
```

Also, based on marital status analysis, you can infer that single individuals have a higher positive response rate. This could be due to various reasons: One reason could be that compared with other categories of customers, single individuals have available income to deposit in long-term savings accounts (term deposit). Hence, the campaign should target single customers.

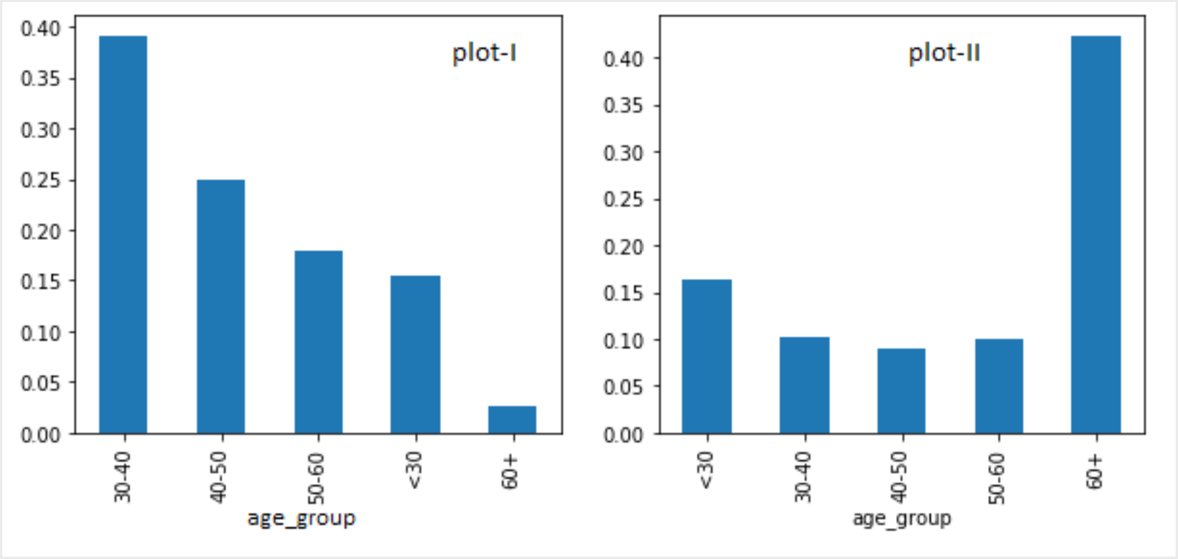


Another very interesting inference is that people who have not purchased any housing or personal loan are more likely to open a term deposit account with the bank. This is true, probably because people who have already availed loans may not have the necessary funds to invest in a term deposit.



Now, let's study the association between the age variable and response rate in the next video.

So, age group analysis showed that people in the age group of 60+ or <30 are more likely to respond positively. It may be true for older people, since they want to invest through more secure investment methods such as term deposits to have a secure old age.



From the image above, you can observe that the bank has mostly contacted people in the age group of 30-50, and have made much less contact with people in the age group of 60+ (plot-I), although the chances of getting a positive response are higher from the people who are in the age group of 50+ or 60+ (as shown in plot-II). This is a very important insight that one can draw from this data set, i.e., the bank should target the people in the 50+ age group.

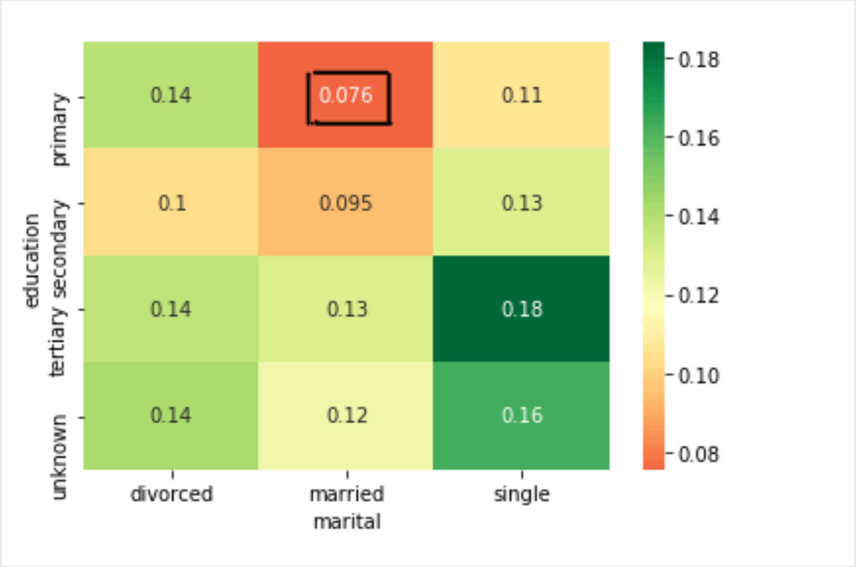
Multivariate Analysis:-

In this segment, we will discuss the next part of EDA, i.e., multivariate analysis.

So far, you have learnt how two variables can be visualised based on their type, for example, numerical, categorical, etc. Now, let's analyse two variables simultaneously. One of the key features of multivariate analysis is that it gives you a very precise idea about the various elements, since you are now combining multiple variables to visualise the data set. You will learn about this in more detail in the forthcoming videos.

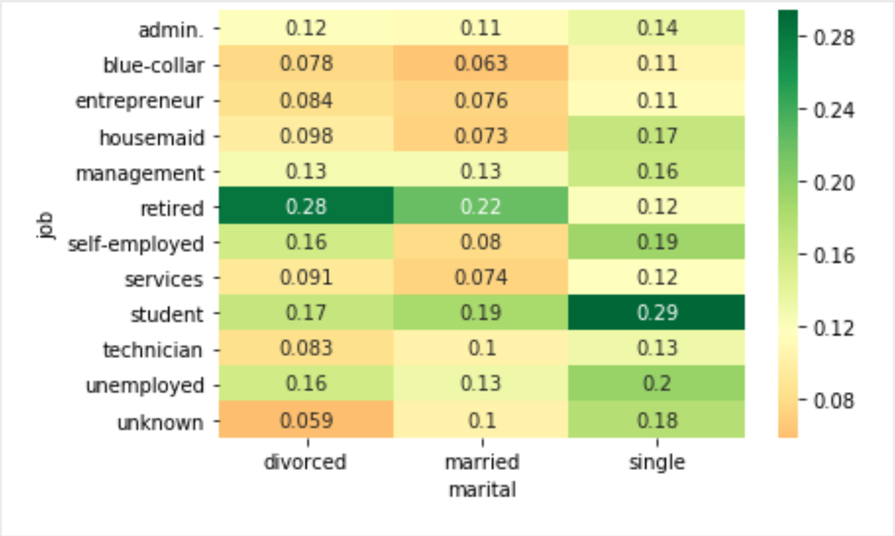
First, let's listen to Rahim as he explains his inferences from the bank marketing dataset in the next video.

In the video above, you saw that our expert performed a three-variable analysis between education, marital status and response. You can see that people who are married and who have completed just their primary education are least likely to give a positive response on term deposits. This can be explained by the fact that people educated only up to the primary level are not aware of the benefits of term investments. Also, married individuals need money to fulfil their daily needs, and they require cash-on-hand to buy the daily essentials; hence, they won't prefer investing in term deposits.



In the next video, you will see how job and marital status are varying with respect to the response variable.

In the video, you saw that the combinations of married with blue-collar, entrepreneur and housemaid are least likely to go for term deposits. The highest rate of positive response came from students with single marital status. The bank should, therefore, consider these aspects before taking any decision.



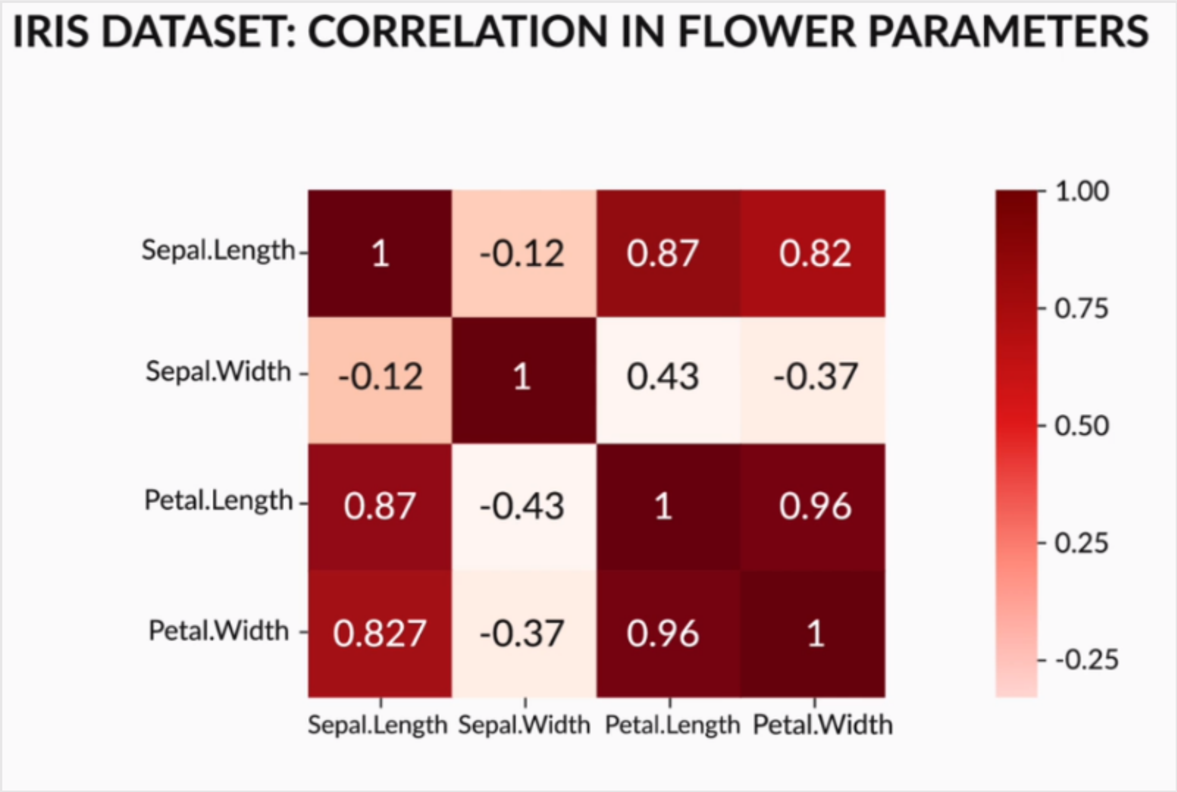
Having gone through all these examples, you must have a clear idea about the EDA process and the various steps involved in it.

Summary:-

In this session, you learnt about the various types of bivariate

and multivariate analyses. These include the following:

- **Analysis between two numerical variables:** The most important thing to remember is that **correlation** and **scatter plots** are the best methods to perform an analysis on numerical variables. Correlation coefficient indicates how much two numerical variables are correlated linearly. And scatter plots offer the exact visualisation between the numerical variables.

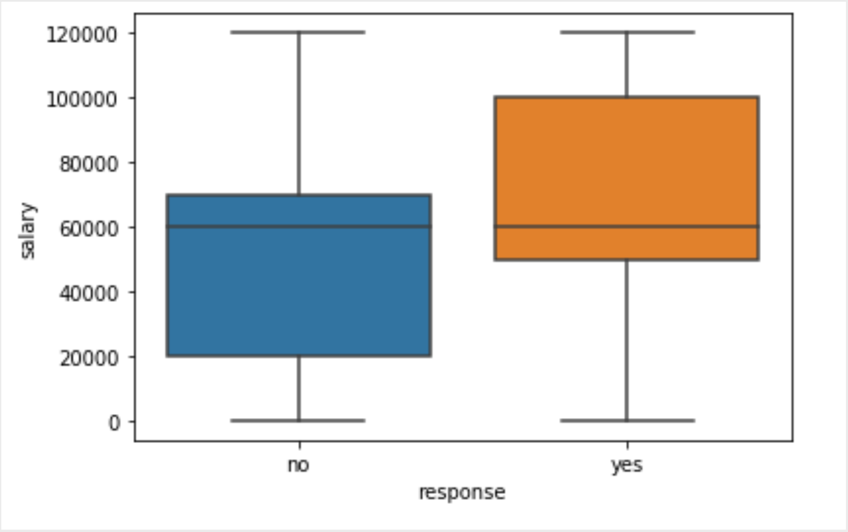


IRIS DATASET: CORRELATION IN FLOWER PARAMETERS

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1	-0.12	0.87	0.82
Sepal.Width	-0.12	1	0.43	-0.37
Petal.Length	0.87	-0.43	1	0.96
Petal.Width	0.827	-0.37	0.96	1

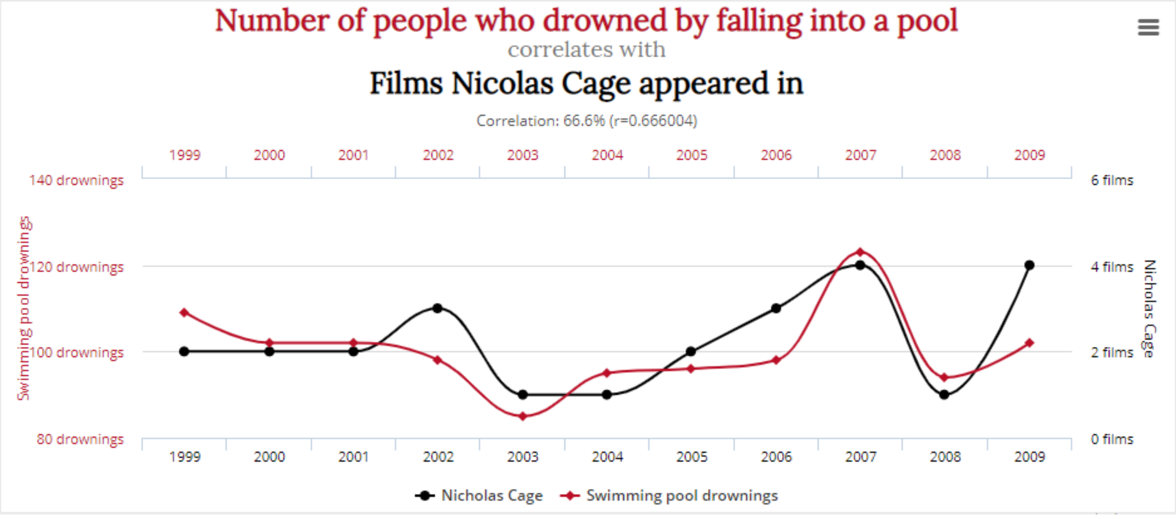
As you can observe in the correlation matrix above, among all the combinations in the data set, there is a high correlation between petal length and sepal length, and petal width and petal length.

- **Analysis between numerical and categorical variables:** This gives an idea about the variation of a particular numerical variable with respect to different categories of a categorical variable. **Boxplot** is the best way to look at a numerical variable with respect to a categorical variable. However, boxplots may sometimes not be useful because of the huge difference between the maximum and minimum values in the data set, or due to the higher concentration of data in the numerical variable. Another approach could be to look into the mean/median or quartiles, which are a more efficient way to deal with a numerical variable when combined with a categorical variable. Take a look at the example shown below.

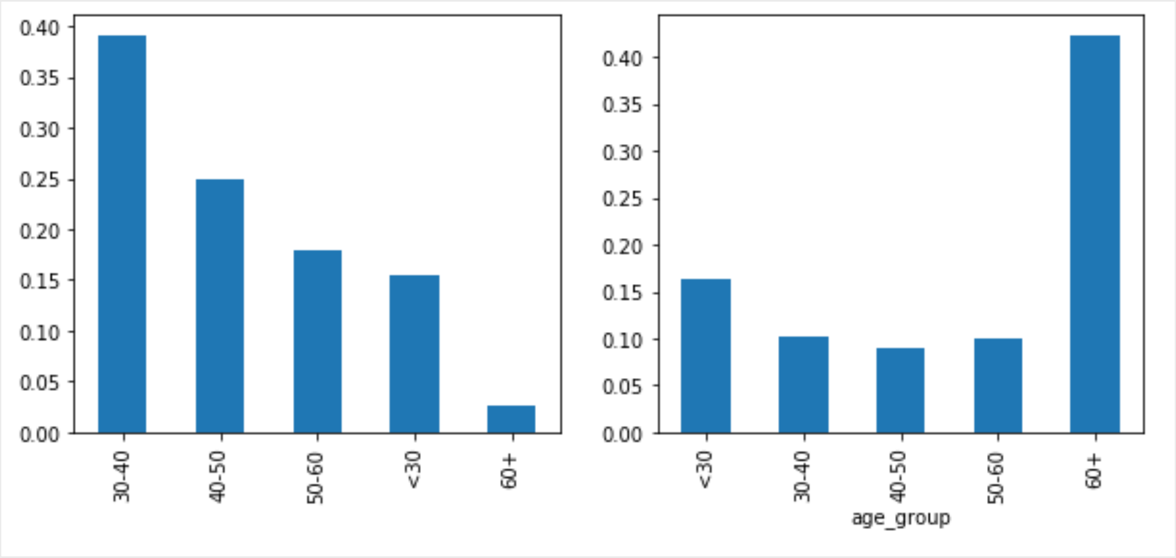


As you can see in the box plot (already explained in the bank marketing dataset) above, customers with a higher salary range are more likely to give a positive response.

- **Correlation vs causation:** This is a very important concept of data anaylsis, which states that correlation is not always related to causation. Although there may be a very high correlation between variables, there may be no causation at all.

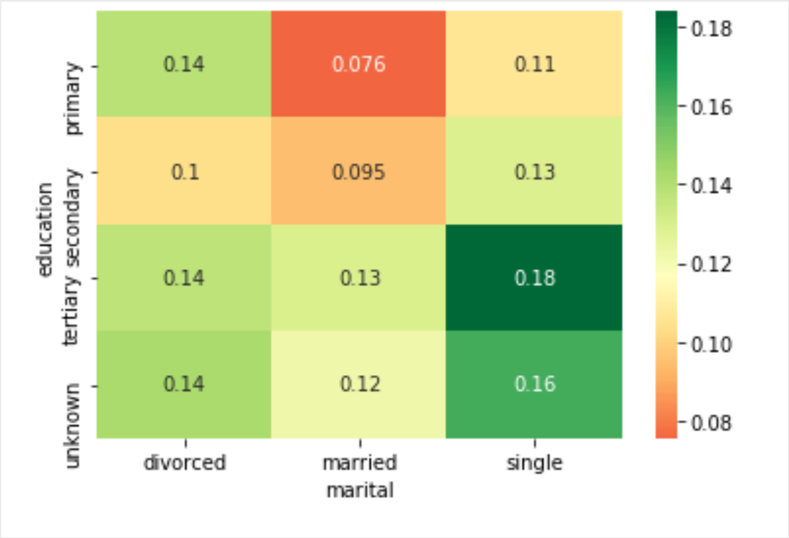


- **Analysis between two categorical variables:** A **bar graph** is the best approach to analysing two categorical variables.



One of the interesting examples, also covered in the bank marketing dataset, is that the bank has mostly contacted people in the age group of 30-50, although people in the age group of 60+ gave more positive responses among all the age groups. This is a very important inference that the bank can draw, i.e., it should contact more individuals in the age group of 60+.

- **Multivariate analysis:** Multivariate analysis yields very specific information about a data set. It basically involves the analysis of more than two variables at a time. For instance, **heat maps** are the best way to look at three variables at a time. In multivariate analysis, it is essential to look into the data by grouping the variables and infer decisions from it.



As you have seen already in the bank marketing case study, single people with tertiary education are more likely to give a positive response to term deposit. And married individuals and those who have completed up to primary education are least

likely to give a positive response.

In the next segment, Rahim will summarise the module on Exploratory Data Analysis.

Module Summary:-

Exploratory Data Analysis (EDA) helps a data analyst to look beyond the data. It is a never-ending process—the more you explore the data, the more the insights you draw from it. As a data analyst, almost 80% of your time will be spent understanding data and solving various business problems through EDA. If you understand EDA properly, that will be half the battle won.

Now, one thing that you should keep in mind is that EDA is far more than plain visualisation. It is an end-to-end process to analyse a data set and prepare it for model-building.

In this module, you have learnt about the four most crucial steps in any kind of data analysis. These steps include the following:

- **Gather data for analysis:** In the data sourcing part, you learnt about the various sources of data. There are majorly two types of data sources, namely, **public data** and **private data**. Private data is associated with some security and privacy concerns, whereas public data is freely available to use without any restrictions on access or usage. There are many websites that provide access public data set available. You have also learnt about the basics of web scraping—a process to fetch the data from a web page directly.
- **Preparation and cleaning of data:** In the cleaning process, the main objective is to remove irregularities from a data set. There are many ways to clean data, but the two most important approaches that you learnt as part of the cleaning step are **treatment of missing values** and **outlier handling**.

Now, there are many ways to deal with missing values, for example, removing an entire column or rows with missing values; however, you need to keep in mind that it should not hamper the data with loss of information. The other method to deal with missing values is to just impute them with other values such as mean, median, mode or quantiles. The third method is to treat the missing values as a separate category; this is the safest method to deal with missing values.

Next, you learnt about the different methods for analysing variables. These methods include the following:

- **Univariate analysis:** Univariate analysis involves the analysis of a single variable at a time. Now, there are multiple types of variables,

such as categorical ordered and unordered variables, and numerical variables. A univariate analysis gives insights about a single variable and how it varies, and what the counts of each and every category in it are.
- **Bivariate and multivariate analysis:** Bivariate/multivariate analysis involves analysing two or more variables at the same time. These analyses yield very specific insights about a data set. You can infer various findings through bivariate analysis.

Now, let's listen to Rahim as he summarises the key learnings in the next video.

Module 5 : EDA Interview Practice

Introduction:-

Welcome to the session on EDA: Interview Practice!

Till now, you would have gone through the fundamental EDA techniques that you employ when working with datasets. From understanding how to clean datasets to analysing the numeric and categorical variables in univariate/ bivariate analyses, you now have ample practice on how to perform EDA in a variety of situations. Now, as an added bonus, we have compiled a list of frequently asked interview questions in EDA for your benefit. Make sure you go through this content to understand the various dos and don'ts of interview preparation.

In this session

You will get to hear from Anjali Rajvanshi from the upGrad Content Team, who will be demonstrating some crucial strategies during interview preparation and how to answer the questions. She'll also be discussing the sample answers for some of the commonly asked interview questions.

Overview:-

Before we begin with the interview questions, here's the basic structure of each video that you're about to consume

- Every video starts with a question related to EDA
- The SME explains the thought process and methodology required for answering the question adequately. This will get you thinking in the right direction.
- You are asked to give an answer for the same question to the best of your abilities.
- After that, the expert gives a sample answer for the same question. This will help you identify your strengths and weaknesses in the conceptual understanding of the topic and help you get prepared for

the actual interview process.

The questions have been curated for the following topics in EDA

- Missing Values and their treatment
- Data Visualisation
- Correlation between multiple variables
- Handling Outliers
- Overview of EDA steps

Make sure you revise all the above concepts once again to attain maximum benefit out of this session. Here are some additional links that you can also use as a reference for giving examples while giving answers:

- Missing Values
- Outliers
- Correlation
- Correlation vs Causation
- Establishing cause and effect

Now let's go ahead and start with our first question, which deals with the overview of how you perform EDA.

Missing Values:-

This segment deals with interview questions related to missing values. The first question deals with the occurrence of missing values. Please take a look at the video given below

The next video deals with the various steps involved in handling missing values

Correlation and Causation:-

In this segment, you'll get to know some commonly asked questions from the topics of correlation and causation.

The next question deals with the idea of whether "correlation implies causation".

In the next question, you'll be discussing how to find causation between variables.

Data Visualisation:-

In this segment, you'll be tackling some questions on data visualisation.

The next question is very interesting since it involves visualising multiple dimensions using only a 2D plot. Take a look.

Other topics:-

In this segment, you will be discussing about handling outliers and choosing between IQR and standard deviation

In the next video, you'll be discussing on when to choose the IQR and when to choose the standard deviation for explaining the variability in the data.

With that, we come to the end of the session. Make sure that you review these questions from time to time to be prepared for any interviews you're taking in the future. Good Luck!