COURSEWARE

EDF

Data Analysis Foundation Courseware

Van Haren
PUBLISHING

Data Analysis Foundation
Courseware

Colophon

Title:	Data Analysis Foundation Courseware
Authors:	Van Haren Publishing
Publisher:	Van Haren Publishing, ‘s-Hertogenbosch
ISBN Hard Copy:	978 94 018 1070 8
Edition:	First edition, first print, November, 2023
Design:	Van Haren Publishing, ‘s-Hertogenbosch
Copyright:	© Van Haren Publishing 2023

For further information about Van Haren Publishing please e-mail us at: info@vanharen.net or visit our website: www.vanharen.net

Publisher about the Courseware

The Courseware was created by experts from the industry who served as the author(s) for this publication. The input for the material is based on existing publications and the experience and expertise of the author(s). The material has been revised by trainers who also have experience working with the material. Close attention was also paid to the key learning points to ensure what needs to be mastered.

The objective of the courseware is to provide maximum support to the trainer and to the student, during his or her training. The material has a modular structure and according to the author(s) has the highest success rate should the student opt for examination. The Courseware is also accredited for this reason, wherever applicable.

In order to satisfy the requirements for accreditation the material must meet certain quality standards. The structure, the use of certain terms, diagrams and references are all part of this accreditation. Additionally, the material must be made available to each student in order to obtain full accreditation. To optimally support the trainer and the participant of the training assignments, practice exams and results are provided with the material.

Direct reference to advised literature is also regularly covered in the sheets so that students can find additional information concerning a particular topic. The decision to leave out notes pages from the Courseware was to encourage students to take notes throughout the material.

Although the courseware is complete, the possibility that the trainer deviates from the structure of the sheets or chooses to not refer to all the sheets or commands does exist. The student always has the possibility to cover these topics and go through them on their own time. It is recommended to follow the structure of the courseware and publications for maximum exam preparation.

The courseware and the recommended literature are the perfect combination to learn and understand the theory.

-- Van Haren Publishing

Other publications by Van Haren Publishing

Van Haren Publishing (VHP) specializes in titles on Best Practices, methods and standards within four domains:

- IT and IT Management
- Architecture (Enterprise and IT)
- Business Management and
- Project Management

Van Haren Publishing is also publishing on behalf of leading organizations and companies: ASLBiSL Foundation, BRMI, CA, Centre Henri Tudor, Gaming Works, IACCM, IAOP, IFDC, Innovation Value Institute, IPMA-NL, ITSqc, NAF, KNVI, PMI-NL, PON, The Open Group, The SOX Institute.

Topics are (per domain):

IT and IT Management

ABC of ICT
ASL®
CATS CM®
CMMI®
COBIT®
e-CF
ISO/IEC 20000
ISO/IEC 27001/27002
ISPL
IT4IT®
IT-CMF™
IT Service CMM
ITIL®
MOF
MSF
SABSA
SAF
SIAM™
TRIM
VeriSM™

Enterprise Architecture

ArchiMate®
GEA®
Novius Architectuur
Methode
TOGAF®

Business Management

BABOK® Guide
BiSL® and BiSL® Next
BRMBOK™
BTF
EFQM
eSCM
IACCM
ISA-95
ISO 9000/9001
OPBOK
SixSigma
SOX
SqEME®

Project Management

A4-Projectmanagement
DSDM/Atern
ICB / NCB
ISO 21500
MINCE®
M_o_R®
MSP®
P3O®
PMBOK® Guide
Praxis®
PRINCE2®

For the latest information on VHP publications, visit our website: www.vanharen.net.

COURSE DESCRIPTION

Empower Your Business Acumen with Data: Master the Art of Decision-Making!

In today's fast-paced business landscape, making swift, precise decisions is non-negotiable. Data-driven strategies have become the cornerstone of successful organizations, replacing a reliance on intuition and outdated methodologies. Whether you aim to gain a competitive edge or enhance customer service, mastering the art of data analysis for business decisions is a must-have skill.

Join our transformative course, where you'll delve into the world of data analytics, equipping yourself with the expertise to make actionable recommendations and unearth opportunities for innovative, data-driven business practices. Learn to decipher the language of business intelligence and explore the profound impact of data analytics on competitive advantage and decision-making support. This course is not just about theoretical concepts; it's a hands-on journey where you'll grasp the intricacies of probability theory, statistical inference, and forecasting, and discover how to translate data into strategic insights.

WHO SHOULD ATTEND?

- This course is tailored for professionals engaged in operations, project management, business analysis, and management roles. Whether you're a seasoned analyst, a project manager aiming for certifications like CBAP®, CCBA®, PMP®, or CAPM®, or an executive keen on enhancing your organization's analytical capabilities, this course is your gateway to transformative learning.

Key Benefits:

- **Comprehensive Learning:** Grasp essential concepts from probability theory through to data mining, empowering you to navigate the complex realm of data analysis.
- **Real-world Application:** Gain practical skills in exploring data, risk management, and forecasting, ensuring you're prepared for the challenges of contemporary business scenarios.
- **Certification Ready:** Prepare for the Effective Data Foundation (EDF) Certified Data Analysis Professional exam and earn valuable PMI® Approved PDU®s, IIBA® Approved CDU®s, or other continued education credits.
- **Expert Guidance:** Learn from seasoned professionals and gain insights into powerful reference materials, honing your decision-making prowess.
- Embark on this enriching journey and revolutionize the way you approach business decisions. Let data be your guiding light in a world driven by information and innovation. Don't just adapt; lead the change with confidence!

Enroll Now and Unleash the Power of Data in Your Decision-Making!

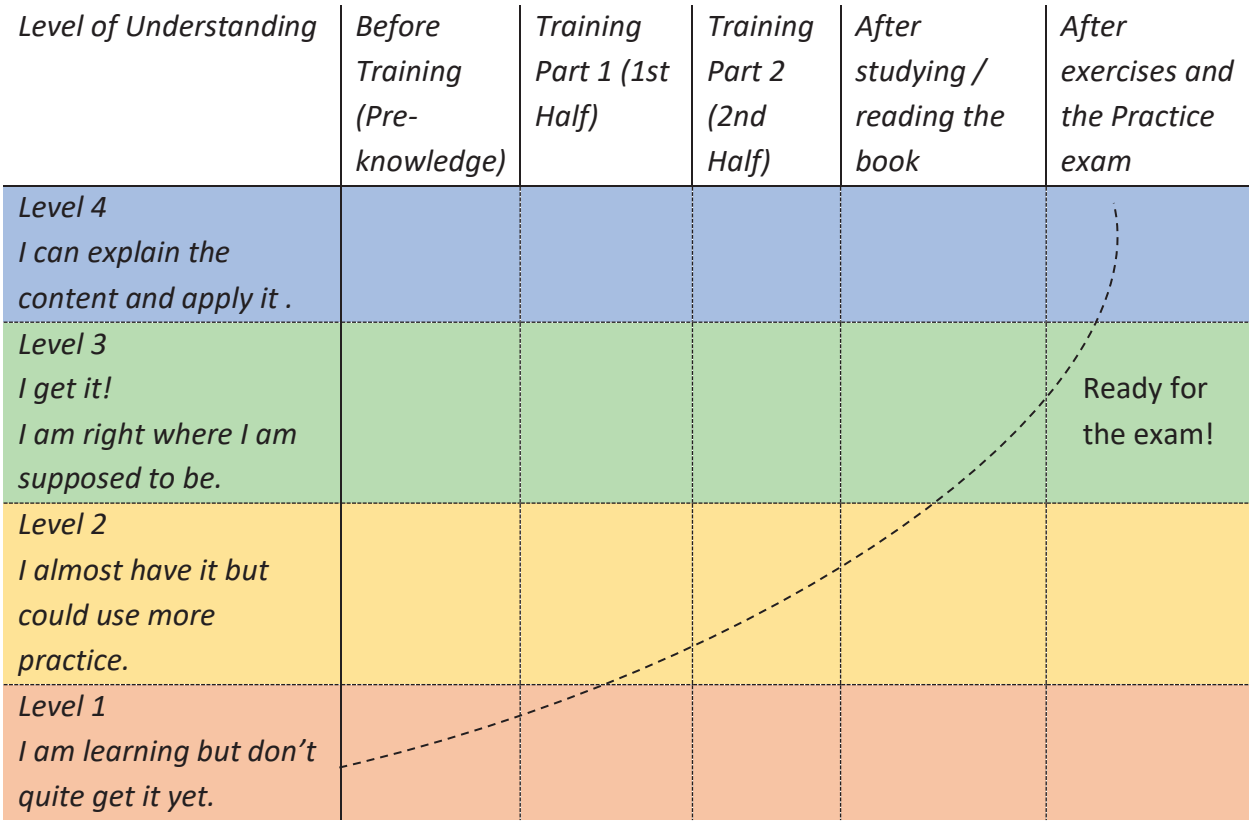
Table of content

Self-Reflection of understanding Diagram

'*What you do not measure, you cannot control*.'' – Tom Peters

Fill in this diagram to self-evaluate your understanding of the material. This is an evaluation of how well you know the material and how well you understand it. In order to pass the exam successfully you should be aiming to reach the higher end of Level 3. If you really want to become a pro, then you should be aiming for Level 4. Your overall level of understanding will naturally follow the learning curve. So, it's important to keep track of where you are at each point of the training and address any areas of difficulty.

Based on where you are within the Self-Reflection of Understanding diagram you can evaluate the progress of your own training.



(Self-Reflection of Understanding Diagram)

Write down the problem areas that you are still having difficulty with so that you can consolidate them yourself, or with your trainer. After you have had a look at these, then you should evaluate to see if you now have a better understanding of where you actually are on the learning curve.

Troubleshooting

Problem areas:	*Topic:*
Part 1	
Part 2	
You have gone through the book and studied.	
You have answered the questions and done the practice exam.	

Timetable

Section 1: Introduction to Data

- Data in practical applications
- Formats and sources of data
- The 7 Vs of big data
- Structured, semi-structured, and unstructured data

Section 2: Fundamentals of Data Analysis

- Structured, semi-structured, and unstructured data
- Application of data analysis
- The process of analyzing data
- Numerical data
- Categorical data

Section 3: Descriptive Data & Statistics

- Descriptive statistics in data analysis
- Frequency
- Measures of central tendency
- Measures of dispersion
- Data skewness and kurtosis
- Outliers
- Missing values

Section 4: Probability

- Probability overview
- Axioms of probability
- Conditional probability and Bayes' theorem

Section 5: Probability Distributions

- Discrete probability distributions
- Continuous probability distributions
- Performing distributions in Excel
- Importance of data distribution for data analysis

Section 6: Executing Data Analysis

- Covariance and correlation
- Univariate, bivariate, and multivariate analysis
- Linear regression
- Simple linear regression
- Multiple linear regression

Data Analysis Deep Dive

Course Objectives

Upon finishing this course, you will acquire the following abilities:

- Define the meaning and usage of data and its various applications
- Explain the process and stages of data analysis, including its different categories
- Apply descriptive and inferential statistics for data analysis
- Provide an overview of probability theory and its significance in data analysis
- Demonstrate proficiency in conducting data analysis utilizing Excel

Course Agenda

- Section 1: Introduction to Data
- Section 2: Fundamentals of Data Analysis
- Section 3: Descriptive Data & Statistics
- Section 4: Probability Theory
- Section 5: Probability Distributions
- Section 6: Executing Data Analysis

Section 1: Introduction to Data

Overview: Introduction to Data

Upon finishing this section, you will acquire the following abilities:

- Explain the significance of data
- Provide an overview of the various formats and sources of data
- Describe the 7Vs of big data
- Outline the differences between structured, semi-structured, and unstructured data

Section Agenda

Section 1: Introduction to Data

- Data in practical applications
- Formats and sources of data
- The 7 Vs of big data
- Structured, semi-structured, and unstructured data
- Data processing techinques

Section 2: Fundamentals of Data Analysis

Section 3: Descriptive Data & Statistics

Section 4: Probability

Section 5: Probability Distributions

Section 6: Executing Data Analysis

Subject 1: Data in Practical Applications

- Data abundance
- Data in government
- Data in entertainment
- Data in retail

Data Abundance

- Data is present in every aspect of our lives
- Over the past decade, we have witnessed an exponential surge in data generation
- Data has emerged as a vital resource for both business enterprises and governmental organizations, playing an increasingly important role in their operations

Data as a Resource

Offers valuable insights that assist, for example:

- Business executives and managers in making informed decisions
- Governments in understanding their citizen's needs and behavioral patterns
- Farmers in reducing agricultural risks and ensuring food availability

Data in Government

Government utilizes data for various purposes:

- Ensuring cities are sustainable and suitable for living
- Assessing necessary infrastructure investments to meet population needs
- Optimizing public transit schedules to efficiently serve communities

Data in Entertainment

Major media platforms such as YouTube, Spotify, and Netflix employ data for various objectives:

- Safeguarding user privacy and maintaining product security
- Enhancing the user experience
- Tailoring content and services according to customer preferences

Data in Retail

Retail stores of all sizes rely on data to:

- Determine when to restock their inventory
- Forecast customer preferences and provide personalized product recommendations on their websites

Subject 2: Formats and Sources of Data

Sources of data:

- Social media:
  - Instagram, TikTok
- eCommerce:
  - Amazon, Shopify, Squarespace
- Public systems:
  - News, criminal databases, citizen resgistry
- Enterprise applications:
  - Microsoft Dynamics 365, SAP
- Devices and sensors

Data in Finance

- Banking and insurance have for centuries been essential components of the financial services industry
- Every second, financial transactions are conducted worldwide, processing a tremendous volume of data
- Consider the immense amount of data generated from just one type of transaction, such as paying for gas

Data in Large Enterprises and Data From Devices and Sensors

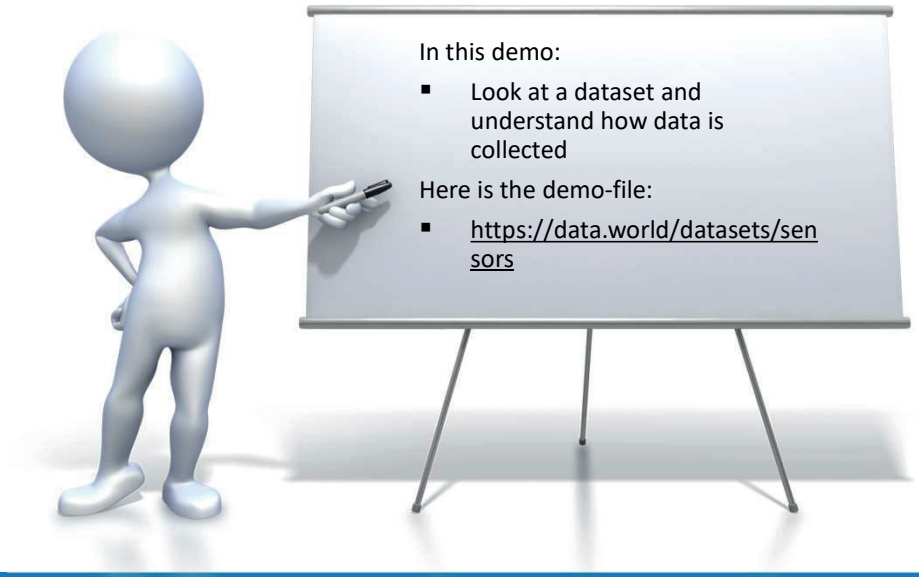
Data in large enterprises:

- Companies such as Amazon, Google, and Meta have a significant impact on economies and generate vast amounts of data daily, in diverse formats

Data collected by devices and sensors:

- Companies like Blue Origin, SpaceX, and Virgin Galactic are undertaking colossal endeavors to succeed in space travel
- To assess the performance and safety of rockets and space shuttles, they install sensors that constantly transmit telemetry data for monitoring purposes

Demo 1: Working with Sensor Datasets



Subject 3: The 7 Vs of Big Data

These seven Vs collectively capture the key aspects of big data and provide a framework for understanding its complexity and potential challenges:

- Volume
- Velocity
- Variety
- Veracity
- Value
- Variability
- Visualization

Volume

- The sheer amount or scale of data being generated and collected
- Meta, Google, Amazon, and Microsoft store a massive amount of data, approximately 1,200 petabytes (http://seedscientific.com/how-much-data-is-created-every-day/):
- 1 PB = e6 GB (1,000,000 GB)
- 1YB (yottabyte) = e15 GB (1,000,000,000,000,000 GB)

Velocity

- The speed at which data is being generated, transmitted, and processed in real-time or near real-time
- Many organizations are exploring the use of real-time data processing, which allows them to make quicker decisions and enhance the customer experience

Variety

- The diversity of data types and sources, including structured, unstructured, and semi-structured data
- Most of the data generated today is unstructured:
  - Data comes in various shapes and forms, such as the content found on social media platforms. This content can include videos, audio files, text, and photos
- Managing and analyzing rapidly changing data with diverse structures can be challenging

Veracity

- The reliability, accuracy, and trustworthiness of the data
- The accuracy of data is very important if we want to conduct trustworthy analysis based on it
- Veracity is about ensuring accuracy of data

Value

- The significance and usefulness of the data in providing insights, making informed decisions, and driving outcomes
- Having a substantial amount of accurate data, combined with effective analysis and suitable visualization techniques, can lead to valuable insights that have a significant impact on both businesses and society
- The consequences of having poor data can be enormous, with an estimated cost of $3.1 trillion dollars per year, as mentioned in this Wikibon article: (http://wikibon.com/breaking-analysis-mongodb-sends-strong-signals-despite-cautious-macro-tones/)

Variability and Visualization

Variability:

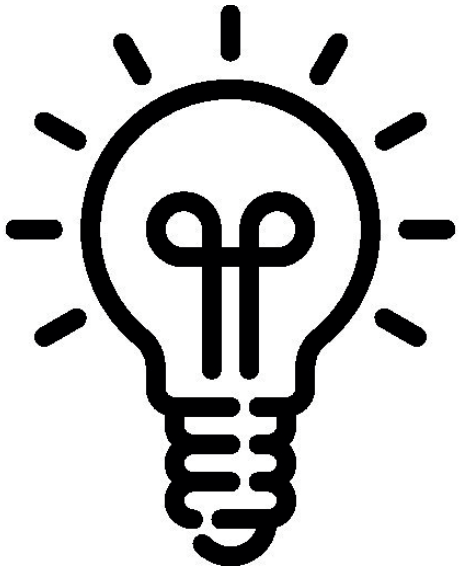
- The inconsistency or volatility of data in terms of its format, structure, or quality over time
  - For instance, when you record a video or engage in a video call using your phone's camera, it captures and processes the intensity of light

Visualization:

- The ability to effectively present and interpret data through visual representations to gain insights and communicate information

KNOWLEDGE CHECK

1. Which aspect of data is being described in this scenario where Mr. Beast's YouTube channel garners millions of views and comments daily, receives thousands of tweets, and experiences exceptional engagement on his TikTok profile?



COURSEWARE

Subject 4: Structured, Semi-structured, and Unstructured Data

Data formats:

- In order for data to be useful, it must be arranged in a specific structure
- The way data is organized can be used to classify it into three different categories:
  - Structured
  - Semi-structured
  - Unstructured

Structured Data

- When data follows a specific structure with consistent properties, it is called structured data
- Typically, structured data is presented in the form of tables
- These tables can be connected to each other using a shared property or 'key' as it is often called in relational models

Semi-structured Data

- Semi-structured data possesses a certain level of structure but does not have to strictly adhere to a fixed schema
- This type of data is commonly utilized for exchanging data across various platforms since the schema is inferred during data reading
- Examples of semi-structured data formats include JSON, XML, and HTML

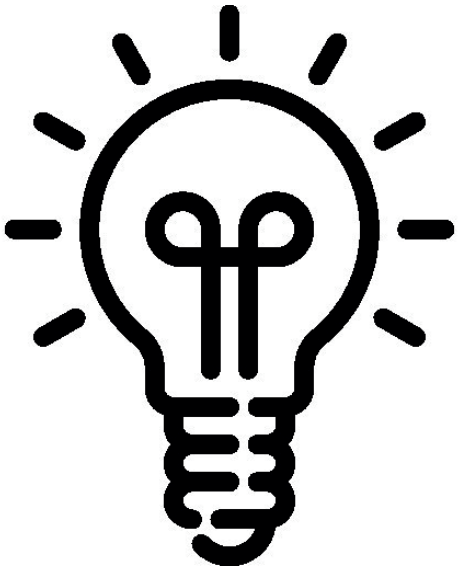
Unstructured Data

- The characteristics of sound, such as volume, pitch, and frequency, can vary due to various factors. To preserve the true nature of audio data, it is typically recorded and stored in its original format
- When data, like audio, is saved and processed without a specific structure, it is referred to as unstructured data
- Other examples of unstructured data include pictures, documents, and videos

KNOWLEDGE CHECK

1. What is the format of the data shown below?

```
<?xml version="1.0" encoding="UTF-8"?>
<note>
        <to>Tove</to>
        <from>Jani</from>
        <heading>Reminder</heading>
        <body>Don't forget me this weekend!</body>
</note>
```



COURSEWARE

Subject 5: Data Processing Techniques

- Data in its raw form may not be meaningful until it undergoes transformation to generate value
- There are two primary methods of data processing:
  - Analytical and Transactional

Analytical Data Processing

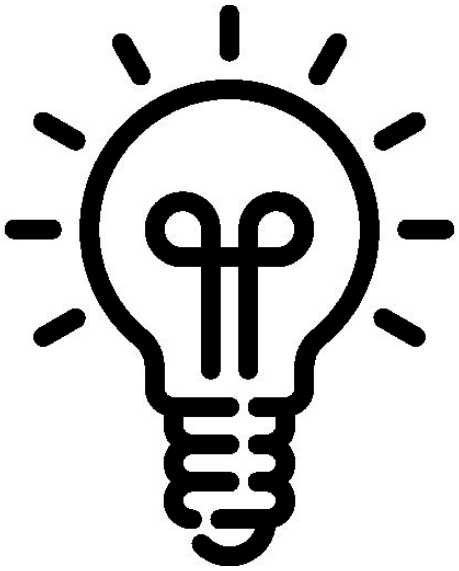
- Analytical data processing entails examining extensive volumes of historical or incoming data streams to extract insights from it
- Visualization tools can be employed to present the data in a format suitable for analysis and storytelling

Transactional Data Processing

- Transactional data processing involves recording transactions or noteworthy events for an organization
- Systems designed for this purpose are commonly known as line of business applications

KNOWLEDGE CHECK

1. What is the format of the data that is stored in its raw state?
2. When you watch YouTube videos, the platform recommends other channels for you based on your search history. What kind of data processing is YouTube using to make these recommendations?
3. Which characteristic of data ensures that it is reliable and valuable for extracting insights?



COURSEWARE

Section 2: Fundamentals of Data Analysis

Overview: Fundamentals of Data Analysis

Upon finishing this section, you will acquire the following abilities:

- Explain the significance of data
- Provide examples of how data analysis is used in different situations
- Describe the process of analyzing data, step by step
- Define what a variable is
- Describe the differences between numerical and categorical variables

Section Agenda

Section 1: Introduction to Data

Section 2: Fundamentals of Data Analysis

- Importance of data analysis
- Application of data analysis
- The process of analyzing data
- Numerical data
- Categorical data

Section 3: Descriptive Data & Statistics

Section 4: Probability

Section 5: Probability Distributions

Section 6: Executing Data Analysis

Subject 1: Importance of Data Analysis

- Data exists in various formats and arrives at varying speeds, which poses challenges in extracting valuable insights in its raw form
- Data analysis can help in:
  - Resolving complex problems efficiently, which would otherwise require substantial time and effort
  - Generating new ideas and innovations
- Reducing costs through optimized decision-making and resource allocation

Innovation and Problem-Solving With Data Analysis

Driving innovation with data analysis:

- Transportation is a major challenge and can be expensive and uncomfortable for commuters
- Zoox, a company that develops self-driving vehicles, is utilizing data analysis to assess the precision and effectiveness of their autonomous vehicles for personal transportation

Solving problems with data analysis:

- From 2020 to 2022, the world faced a severe global pandemic known as COVID-19
- Government officials relied on data insights to make decisions aimed at controlling the spread of COVID-19 and reducing its impact

Subject 2: Application of Data Analysis

Categories of analytics:

- Descriptive
- Diagnostic
- Predictive
- Prescriptive
- Cognitive

Descriptive Analytics

- Descriptive analytics is used to understand and describe past events based on available data
- It involves working with a large dataset and ensuring its accuracy through data cleaning. The goal is to provide insights and describe outcomes to interested parties

Diagnostic Analytics

Example:

- Let us consider a situation where a car manufacturer detects defects in a specific model produced on an assembly line
- To understand the cause of these defects, engineers can utilize diagnostic analytics and descriptive analytics
- They would use statistical techniques to identify relationships and patterns that explain the defects

Predictive Analytics

Example:

- An elevator company aims to predict potential failures in critical elevator components to schedule proactive maintenance
- The company seeks to save costs and ensure customer satisfaction by addressing maintenance needs before failures occur
- Predictive analytics is employed to forecast future events and anticipate when component failures are likely to happen
- By proactively scheduling maintenance, the company can avoid costly breakdowns and ensure smooth operation for their customers

Prescriptive Analytics

Example:

- The organization wants to make their revenue ten times larger in the next fiscal year
- To achieve this, they need to follow some steps that can help reach their goal
- One helpful tool is Prescriptive Analytics, which can provide answers to questions about which actions the business should take to meet its targets

Cognitive Analytics

Example:

- An e-commerce company aims to improve the number of people who open their marketing emails
- To achieve this, they can utilize cognitive analytics to figure out the best timing and reasons for customers to open an email
- Cognitive analytics will analyze historical data, including patterns of when customers opened emails and the content and subject line of those emails, to draw insights and make predictions

Subject 3: The Process of Analyzing Data

The data analysis process includes five important steps:

- Data preparation
- Data modelling
- Data visualization
- Analytics
- Artifact management

Data Preparation

- This step involves gathering and organizing the data that will be analyzed
- It may include cleaning the data, removing errors or duplicates, and structuring it in a suitable format for analysis
- Data preparation ensures that the processed data is reliable and can be trusted for analysis and decision-making