FOUNDATIONS OF DATA ANALYTICS (CSE3505)

Dr. Trilok Nath Pandey

Vellore Institute of Technology
SCOPE VIT, Chennai

triloknath.pandey@vit.ac.in

25/07/2022

Text Books

Trevor Hastie & Rob Tibshirani

An Introduction to Statistical Learning with Applications in R

Springer, 2017

Mark van der Loo & Edwin de Jonge

Learning R Studio for R Statistical Computing

Packt Publishing, 2012.

Jure Leskovek, Anand Rajaraman & Jeffrey Ullman

Mining of Massive Datasets

Cambridge University Press. 2014

Course Overview

- **Module-1: Introduction to Analytics**
  Analytics life cycle - Business Analytics - Lending Analytics- Recommendation Analytics- Healthcare Analytics- Financial Analytics - Sports Analytics.

Course Overview

- **Module-1: Introduction to Analytics**
  Analytics life cycle - Business Analytics - Lending Analytics- Recommendation Analytics- Healthcare Analytics- Financial Analytics - Sports Analytics.

Course Overview

- **Module-1: Introduction to Analytics**
  Analytics life cycle - Business Analytics - Lending Analytics- Recommendation Analytics- Healthcare Analytics- Financial Analytics - Sports Analytics.
- **Module-2: R programming Basics**
  Introduction to R, R Studio (GUI): R Windows Environment, introduction to various data types, Numeric, Character, date, data frame, array, matrix etc.

Course Overview

- **Module-1: Introduction to Analytics**
  Analytics life cycle - Business Analytics - Lending Analytics- Recommendation Analytics- Healthcare Analytics- Financial Analytics - Sports Analytics.
- **Module-2: R programming Basics**
  Introduction to R, R Studio (GUI): R Windows Environment, introduction to various data types, Numeric, Character, date, data frame, array, matrix etc.

Course Overview

- **Module-1: Introduction to Analytics**
  Analytics life cycle - Business Analytics - Lending Analytics- Recommendation Analytics- Healthcare Analytics- Financial Analytics - Sports Analytics.
- **Module-2: R programming Basics**
  Introduction to R, R Studio (GUI): R Windows Environment, introduction to various data types, Numeric, Character, date, data frame, array, matrix etc.
- **Module-3: Working with datasets and files**
  Reading Datasets, Working with different file types .txt,.csv , R studio,Files, Datasets, Extracting Datasets, Preparing datasets. Data Cleaning,Data imputation, Data conversion Analysis.

Course Overview

- **Module-4: Introduction to statistical learning and R-Programm**
  Basic statistics: mean, median, standard deviation, variance, correlation,covariance - Outliers, Combining Datasets in R, Functions and loops.Summary Statistics - Summarizing data with R - Correlation and Regression.

Course Overview

- **Module-4: Introduction to statistical learning and R-Programm**
  Basic statistics: mean, median, standard deviation, variance, correlation,covariance - Outliers, Combining Datasets in R, Functions and loops.Summary Statistics - Summarizing data with R - Correlation and Regression.

Course Overview

- **Module-4: Introduction to statistical learning and R-Programm**
  Basic statistics: mean, median, standard deviation, variance, correlation,covariance - Outliers, Combining Datasets in R, Functions and loops.Summary Statistics - Summarizing data with R - Correlation and Regression.
- **Module-5: Document Creation and Knowledge Sharing**
  Access existing documents, language standards, templates and documentation tools from their organizations knowledge base. Confirm the content and structure of the documents with appropriate people, Create documents using standard templates and agreed language standards. Review documents with appropriate people and incorporate their inputs.

Course Overview

- **Module-4: Introduction to statistical learning and R-Programm**
  Basic statistics: mean, median, standard deviation, variance, correlation,covariance - Outliers, Combining Datasets in R, Functions and loops.Summary Statistics - Summarizing data with R - Correlation and Regression.
- **Module-5: Document Creation and Knowledge Sharing**
  Access existing documents, language standards, templates and documentation tools from their organizations knowledge base. Confirm the content and structure of the documents with appropriate people, Create documents using standard templates and agreed language standards. Review documents with appropriate people and incorporate their inputs.

Course Overview

- **Module-4: Introduction to statistical learning and R-Programm**
  Basic statistics: mean, median, standard deviation, variance, correlation,covariance - Outliers, Combining Datasets in R, Functions and loops.Summary Statistics - Summarizing data with R - Correlation and Regression.
- **Module-5: Document Creation and Knowledge Sharing**
  Access existing documents, language standards, templates and documentation tools from their organizations knowledge base. Confirm the content and structure of the documents with appropriate people, Create documents using standard templates and agreed language standards. Review documents with appropriate people and incorporate their inputs.

Course Overview

- **Module-6: Self and work Management**
  Establish and agree their work requirements with appropriate people - Keep their immediate work area clean and tidy - utilize their time effectively - Use resources correctly and efficiently - Treat confidential information correctly - Work in line with organizations policies and procedures - Work within the limits of their job role.

Course Overview

- **Module-6: Self and work Management**
  Establish and agree their work requirements with appropriate people - Keep their immediate work area clean and tidy - utilize their time effectively - Use resources correctly and efficiently - Treat confidential information correctly - Work in line with organizations policies and procedures - Work within the limits of their job role.

Course Overview

- **Module-6: Self and work Management**
  Establish and agree their work requirements with appropriate people - Keep their immediate work area clean and tidy - utilize their time effectively - Use resources correctly and efficiently - Treat confidential information correctly - Work in line with organizations policies and procedures - Work within the limits of their job role.
- **Module-7: Team Work and Communication**
  Communicate with colleagues clearly, concisely and accurately - Work with colleagues to integrate their work effectively with them - Pass on essential information to colleagues in line with organizational requirements - Work in ways that show respect for colleagues - carry out commitments they have made to colleagues - Let colleagues know in good time if they cannot carry out their commitments, explaining the reasons - Identify any problems they have working with colleagues and take the initiative to solve these problems.

Course Overview

- **Module-6: Self and work Management**
  Establish and agree their work requirements with appropriate people - Keep their immediate work area clean and tidy - utilize their time effectively - Use resources correctly and efficiently - Treat confidential information correctly - Work in line with organizations policies and procedures - Work within the limits of their job role.
- **Module-7: Team Work and Communication**
  Communicate with colleagues clearly, concisely and accurately - Work with colleagues to integrate their work effectively with them - Pass on essential information to colleagues in line with organizational requirements - Work in ways that show respect for colleagues - carry out commitments they have made to colleagues - Let colleagues know in good time if they cannot carry out their commitments, explaining the reasons - Identify any problems they have working with colleagues and take the initiative to solve these problems.

Course Overview

- **Module-6: Self and work Management**
  Establish and agree their work requirements with appropriate people - Keep their immediate work area clean and tidy - utilize their time effectively - Use resources correctly and efficiently - Treat confidential information correctly - Work in line with organizations policies and procedures - Work within the limits of their job role.
- **Module-7: Team Work and Communication**
  Communicate with colleagues clearly, concisely and accurately - Work with colleagues to integrate their work effectively with them - Pass on essential information to colleagues in line with organizational requirements - Work in ways that show respect for colleagues - carry out commitments they have made to colleagues - Let colleagues know in good time if they cannot carry out their commitments, explaining the reasons - Identify any problems they have working with colleagues and take the initiative to solve these problems.

Outline

- What Is Data Analytics?

Outline

- What Is Data Analytics?

Outline

- What Is Data Analytics?
- Data Analytics: Benefits

Outline

- What Is Data Analytics?
- Data Analytics: Benefits

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle
- Types of Data Analytics

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle
- Types of Data Analytics
  - Descriptive Analytics

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle
- Types of Data Analytics
  - Descriptive Analytics

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle
- Types of Data Analytics
  - Descriptive Analytics
  - Diagnostic Analytics

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle
- Types of Data Analytics
  - Descriptive Analytics
  - Diagnostic Analytics

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle
- Types of Data Analytics
  - Descriptive Analytics
  - Diagnostic Analytics
  - Predictive Analytics

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle
- Types of Data Analytics
  - Descriptive Analytics
  - Diagnostic Analytics
  - Predictive Analytics

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle
- Types of Data Analytics
  - Descriptive Analytics
  - Diagnostic Analytics
  - Predictive Analytics
  - Prescriptive Analytics

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle
- Types of Data Analytics
  - Descriptive Analytics
  - Diagnostic Analytics
  - Predictive Analytics
  - Prescriptive Analytics

Outline

- What Is Data Analytics?
- Data Analytics: Benefits
- Data Analytics: Domains
- Why Data Analytics?
- Data Analytics: Process Flow
- Data Analytics: Life Cycle
- Types of Data Analytics
  - Descriptive Analytics
  - Diagnostic Analytics
  - Predictive Analytics
  - Prescriptive Analytics
- Types of Analytics Amazon Example

Introduction

Introduction

- What Is Data Analytics?

Introduction

- What Is Data Analytics?

Introduction

- What Is Data Analytics?
  Data Analytics is the science of extracting trends, patterns, and relevant information from raw data to draw conclusions.

Data Analytics: Benefits

Data Analytics: Benefits

Cost Reduction

Effective decision making

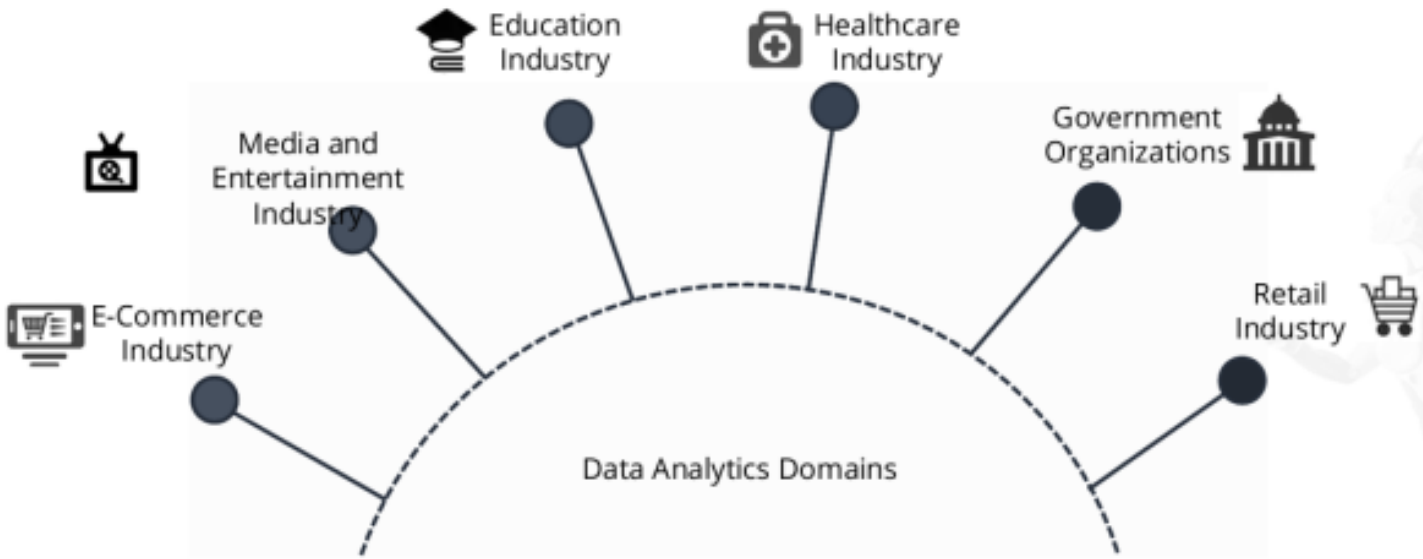
Improved efficiency

Good resource utilization

Good market insights

Data Analytics: Domains

Data Analytics: Domains



Why Data Analytics?

Why Data Analytics?

- Data analytics is important because it helps businesses optimize their performances. Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data.

Why Data Analytics?

- Data analytics is important because it helps businesses optimize their performances. Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data.

Why Data Analytics?

- Data analytics is important because it helps businesses optimize their performances. Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data.
- A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new and better products and services.

Why Data Analytics?

- Data analytics is important because it helps businesses optimize their performances. Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data.
- A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new and better products and services.

Why Data Analytics?

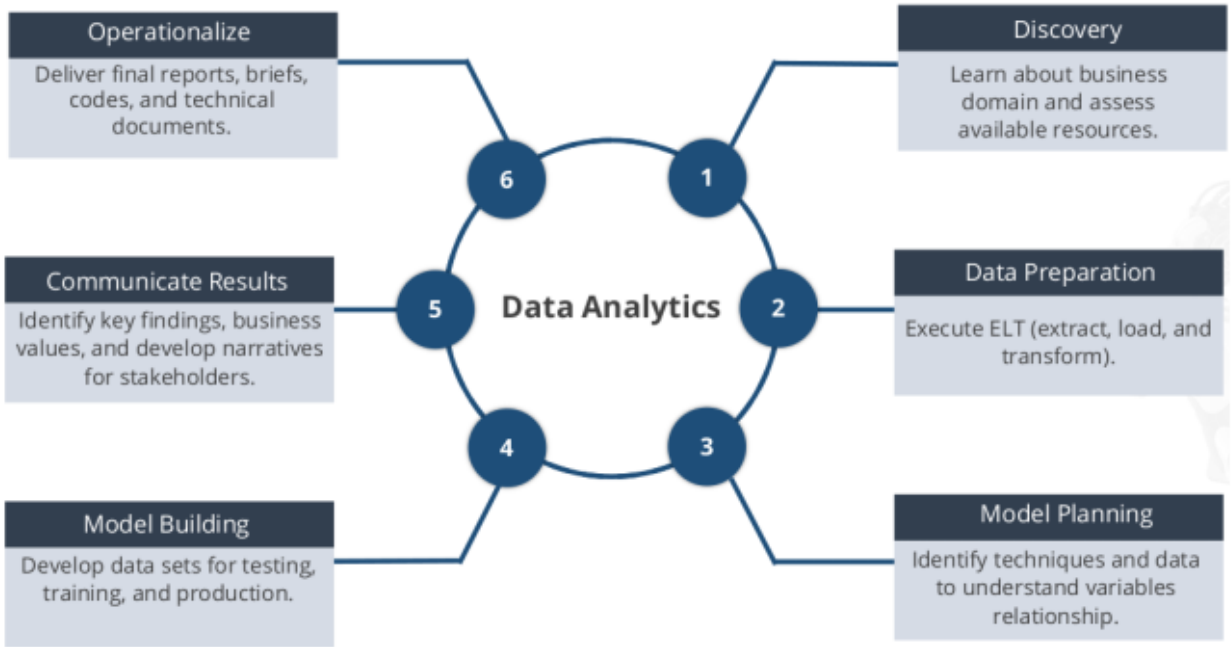
- Data analytics is important because it helps businesses optimize their performances. Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data.
- A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new and better products and services.

Data Analytics: Process Flow

Data Analytics: Process Flow

Data Analytics Life Cycle

Data Analytics Life Cycle



Phases of Data Analytics

Phases of Data Analytics

- **Phase 1 – Discovery**

Phases of Data Analytics

- **Phase 1 – Discovery**

Phases of Data Analytics

- **Phase 1 – Discovery**
  the team learns the business domain from the past.Any similar projects has been attempted in the past from which they can learn Assess available resources – people, technology, time, data

Phases of Data Analytics

- **Phase 1 – Discovery**
  the team learns the business domain from the past.Any similar projects has been attempted in the past from which they can learn Assess available resources – people, technology, time, data
- **Activities**

Phases of Data Analytics

- **Phase 1 – Discovery**
  the team learns the business domain from the past.Any similar projects has been attempted in the past from which they can learn Assess available resources – people, technology, time, data
- **Activities**

Phases of Data Analytics

- **Phase 1 – Discovery**
  the team learns the business domain from the past.Any similar projects has been attempted in the past from which they can learn Assess available resources – people, technology, time, data
- **Activities**
  Framing the business problem as an analytics challenge that can be addressed in subsequent phases Formulating initial hypotheses (IHs) to test and begin learning the data

Phases of Data Analytics

- **Phase 1 – Discovery**
  the team learns the business domain from the past.Any similar projects has been attempted in the past from which they can learn Assess available resources – people, technology, time, data
- **Activities**
  Framing the business problem as an analytics challenge that can be addressed in subsequent phases Formulating initial hypotheses (IHs) to test and begin learning the data

Phases of Data Analytics

Phases of Data Analytics

- **Phase 2 – Data Preparation**

Phases of Data Analytics

- **Phase 2 – Data Preparation**

Phases of Data Analytics

- **Phase 2 – Data Preparation**
  the team prepares the analytic sandbox to perform analytics on the data. The most labor-intensive step (nearly 50% of the time) Execute ETL/ELT (E- Extract, L- Load, T-Transform)

Phases of Data Analytics

- **Phase 2 – Data Preparation**
  the team prepares the analytic sandbox to perform analytics on the data. The most labor-intensive step (nearly 50% of the time) Execute ETL/ELT (E- Extract, L- Load, T-Transform)
- **Activities**

Phases of Data Analytics

- **Phase 2 – Data Preparation**
  the team prepares the analytic sandbox to perform analytics on the data. The most labor-intensive step (nearly 50% of the time) Execute ETL/ELT (E- Extract, L- Load, T-Transform)
- **Activities**

Phases of Data Analytics

- **Phase 2 – Data Preparation**
  the team prepares the analytic sandbox to perform analytics on the data. The most labor-intensive step (nearly 50% of the time) Execute ETL/ELT (E- Extract, L- Load, T-Transform)
- **Activities**
  Data Pre-processing – Cleaning, Normalizing the data etc.

Phases of Data Analytics

- **Phase 2 – Data Preparation**
  the team prepares the analytic sandbox to perform analytics on the data. The most labor-intensive step (nearly 50% of the time) Execute ETL/ELT (E- Extract, L- Load, T-Transform)
- **Activities**
  Data Pre-processing – Cleaning, Normalizing the data etc.

Phases of Data Analytics

Phases of Data Analytics

- **Phase 3 – Model Planning**
  The team determines the methods, techniques, and work flow to be followed in the subsequent model building phase

Phases of Data Analytics

- **Phase 3 – Model Planning**
  The team determines the methods, techniques, and work flow to be followed in the subsequent model building phase

Phases of Data Analytics

- **Phase 3 – Model Planning**
  The team determines the methods, techniques, and work flow to be followed in the subsequent model building phase
- **Activities**

Phases of Data Analytics

- **Phase 3 – Model Planning**
  The team determines the methods, techniques, and work flow to be followed in the subsequent model building phase
- **Activities**

Phases of Data Analytics

- **Phase 3 – Model Planning**
  The team determines the methods, techniques, and work flow to be followed in the subsequent model building phase
- **Activities**
  Exploratory data analysis to understand the trend, patterns, associations/relationship between variables Selects key variables and most important models

Phases of Data Analytics

- **Phase 3 – Model Planning**
  The team determines the methods, techniques, and work flow to be followed in the subsequent model building phase
- **Activities**
  Exploratory data analysis to understand the trend, patterns, associations/relationship between variables Selects key variables and most important models

Phases of Data Analytics

Phases of Data Analytics

- **Phase 4 – Model Building**

Phases of Data Analytics

- **Phase 4 – Model Building**

Phases of Data Analytics

- **Phase 4 – Model Building**
  The team develops datasets for testing, training, and production. Build and Execute the model based on model planning Considers whether its existing tools will suffice for running the models, or if it will need a more robust environment (fast hardware and parallel processing)

Phases of Data Analytics

- **Phase 4 – Model Building**
  The team develops datasets for testing, training, and production. Build and Execute the model based on model planning Considers whether its existing tools will suffice for running the models, or if it will need a more robust environment (fast hardware and parallel processing)

Phases of Data Analytics

Phases of Data Analytics

- **Phase 5 – Communicate Results**

Phases of Data Analytics

- **Phase 5 – Communicate Results**

Phases of Data Analytics

- **Phase 5 – Communicate Results**
  The team determines along with all stakeholders if the results of the project are a success or a failure based on the criteria developed in Phase 1.

Phases of Data Analytics

- **Phase 5 – Communicate Results**
  The team determines along with all stakeholders if the results of the project are a success or a failure based on the criteria developed in Phase 1.
- **Activities**

Phases of Data Analytics

- **Phase 5 – Communicate Results**
  The team determines along with all stakeholders if the results of the project are a success or a failure based on the criteria developed in Phase 1.
- **Activities**

Phases of Data Analytics

- **Phase 5 – Communicate Results**
  The team determines along with all stakeholders if the results of the project are a success or a failure based on the criteria developed in Phase 1.
- **Activities**
  Identify key findings, Quantify the business value, Develop a narrative to summarize, Convey findings to stakeholders.

Phases of Data Analytics

- **Phase 5 – Communicate Results**
  The team determines along with all stakeholders if the results of the project are a success or a failure based on the criteria developed in Phase 1.
- **Activities**
  Identify key findings, Quantify the business value, Develop a narrative to summarize, Convey findings to stakeholders.

Phases of Data Analytics

Phases of Data Analytics

- **Phase 6 – Operationalize**

Phases of Data Analytics

- **Phase 6 – Operationalize**

Phases of Data Analytics

- **Phase 6 – Operationalize**
  the team communicates the results more broadly

Phases of Data Analytics

- **Phase 6 – Operationalize**
  the team communicates the results more broadly
- **Activities**

Phases of Data Analytics

- **Phase 6 – Operationalize**
  the team communicates the results more broadly
- **Activities**

Phases of Data Analytics

- **Phase 6 – Operationalize**
  the team communicates the results more broadly
- **Activities**
  run a pilot project to implement the models in a production environment, Assess the benefits, delivers final report, briefings, code, and technical documents, Model execution in the production environment

Phases of Data Analytics

- **Phase 6 – Operationalize**
  the team communicates the results more broadly
- **Activities**
  run a pilot project to implement the models in a production environment, Assess the benefits, delivers final report, briefings, code, and technical documents, Model execution in the production environment

Types of Data Analytics

Types of Data Analytics

- The four main types of analytics based on the workflow and requirements of data analytics.

Types of Data Analytics

- The four main types of analytics based on the workflow and requirements of data analytics.

Types of Data Analytics

- The four main types of analytics based on the workflow and requirements of data analytics.

Descriptive Analytics

- Descriptive analytics is designed to access information about the past.
- It focuses on the summarized view of facts.

- It is the conventional form of analytics.
- Its purpose is to summarize the findings.

Descriptive Analytics Techniques

Descriptive Analytics Techniques

- Data aggregation is the process of gathering and expressing information in a summarized form.

Descriptive Analytics Techniques

- Data aggregation is the process of gathering and expressing information in a summarized form.

Descriptive Analytics Techniques

- Data aggregation is the process of gathering and expressing information in a summarized form.
- Tools used for data aggregation include MS Excel, MATLAB, SPSS,and STATA.

Descriptive Analytics Techniques

- Data aggregation is the process of gathering and expressing information in a summarized form.
- Tools used for data aggregation include MS Excel, MATLAB, SPSS,and STATA.

Descriptive Analytics Techniques

- Data aggregation is the process of gathering and expressing information in a summarized form.
- Tools used for data aggregation include MS Excel, MATLAB, SPSS,and STATA.
- Company report is an example of descriptive analytics.

Descriptive Analytics Techniques

- Data aggregation is the process of gathering and expressing information in a summarized form.
- Tools used for data aggregation include MS Excel, MATLAB, SPSS,and STATA.
- Company report is an example of descriptive analytics.

Descriptive Analytics Techniques

- Data aggregation is the process of gathering and expressing information in a summarized form.
- Tools used for data aggregation include MS Excel, MATLAB, SPSS,and STATA.
- Company report is an example of descriptive analytics.

Diagnostic Analytics

- Diagnostic analytics helps you identify why something happened in the past.
- It takes a deeper look at data to understand the root cause of events.
- It has a limited ability to provide actionable insights.
- It provides an understanding of causal relationships and sequences.

Diagnostic Analytics Techniques

Predictive Analytics

Predictive analytics is used in:

- Predicting future outcomes in terms of probability of an event to occur
- Analyzing sentiments where all opinions posted on social media are collected to predict a person's sentiment
- Identifying target audience for a promotional campaign
- Forecasting weather, plan-failure prediction, and travel products recommender system

A predictive model is built on the preliminary descriptive analytics stage.

Predictive Analytics Tools

Machine learning algorithms such as random forests, SVM and statistics.

Trained data scientists and machine learning experts building these models

Popular tools for predictive analytics: Python, R and RapidMiner.

Prescriptive Analytics

Types of Analytics: Amazon Example

Diagnostic Analytics

- Amazon's revenue increased in the West Coast during the past one year
- Increased spending on sales training

Predictive Analytics

- Purchase factors: price, time, weather, and festive seasons
- Predicted 10–12 percent increase in revenue

Types of Analytics: Amazon Example

- Spent $20M in different sales training the previous year

- Sales trainings fetched good ROI
- Implemented a suitable optimization plan to maximize revenue

THE END