Applied Data Science

BE(A & B)2024-25 Even Sem

By

Prof Kahkashan Siddavatam

Module		Detailed Content	Hours
1		**Introduction to Data Science**	2
	1.1	Introduction to Data Science, Data Science Process	
	1.2	Motivation to use Data Science Techniques: Volume, Dimensions and Complexity, Data Science Tasks and Examples	
	1.3	Overview of Data Preparation, Modeling, Difference between data science and data analytics	
2		**Data Exploration**	8
	2.1	Types of data, Properties of data **Descriptive Statistics:** Univariate Exploration: Measure of Central Tendency, Measure of Spread, Symmetry, Skewness: Karl Pearson Coefficient of skewness, Bowley's Coefficient, Kurtosis Multivariate Exploration: Central Data Point, Correlation, Different forms of correlation, Karl Pearson Correlation Coefficient for bivariate distribution	

	2.2	**Inferential Statistics:** Overview of Various forms of distributions: Normal, Poisson, Test Hypothesis, Central limit theorem, Confidence Interval, Z-test, t-test, Type-I, Type-II Errors, ANOVA	
3		**Methodology and Data Visualization**	**06**
	3.1	**Methodology:** Overview of model building, Cross Validation, K-fold cross validation, leave-1 out, Bootstrapping	
	3.2	**Data Visualization** Univariate Visualization: Histogram, Quartile, Distribution Chart Multivariate Visualization: Scatter Plot, Scatter Matrix, Bubble chart, Density Chart Roadmap for Data Exploration	
	3.3	**Self-Learning Topics:** Visualizing high dimensional data: Parallel chart, Deviation chart, Andrews Curves.	
4		**Anomaly Detection**	**06**
	4.1	Outliers, Causes of Outliers, Anomaly detection techniques, Outlier Detection using Statistics	
	4.2	Outlier Detection using Distance based method, Outlier detection using density-based methods, SMOTE	

5		**Time Series Forecasting**	4
	5.1	Taxonomy of Time Series Forecasting methods, Time Series Decomposition	
	5.2	Smoothening Methods: Average method, Moving Average smoothing, Time series analysis using linear regression, ARIMA Model, Performance Evaluation: Mean Absolute Error, Root Mean Square Error, Mean Absolute Percentage Error, Mean Absolute Scaled Error	
	5.3	**Self-Learning Topics:** Evaluation parameters for Classification, regression and clustering.	
6		**Applications of Data Science**	4
		Predictive Modeling: House price prediction, Fraud Detection Clustering: Customer Segmentation Time series forecasting: Weather Forecasting Recommendation engines: Product recommendation	

Textbooks:

1	Vijay Kotu, Bala Deshpande. “Data Science Concepts and Practice”, Elsevier, M.K. Publishers.
2	Steven Skiena, “Data Science Design Manual”, Springer International Publishing AG
3	Samir Madhavan. “Mastering Python for Data Science”, PACKT Publishing
4	Dr. P. N. Arora, Sumeet Arora, S. Arora, Ameet Arora, “Comprehensive Statistical Methods”, S.Chand Publications, New Delhi.

Content

- Introduction to Data Science
- Data Science Process
- Motivation to use Data Science Techniques: Volume,Dimensions and complexity,Data Science Tasks and Examples
- Overview of Data Preparation,Modeling, Difference between data science and data analytics

Introduction to Data Science

- Data science is a field that uses a **variety of techniques to extract value** from data.
- These techniques often involve **finding patterns, connections, and relationships** within data and have wide range of applications.
- Data science is also referred as **knowledge discovery, machine learning, predictive analytics**, or data mining.
- Underlying methods of data science have been around for decades and with increasing data collection, storage and processing capabilities it is finding more **applications in diverse fields**.
- Data science process typically involves **preparing, cleaning, scrubbing and standardizing data** before running machine learning algorithms.
- These process are becoming **increasingly automated**, allowing for **more focus on interpreting results and making decisions.**

Applications of Data Science

- **Customer Segmentation:** Data science can be used to identify and group customers based on their characteristics, behaviors, and preferences. By analyzing customer data, businesses can create targeted marketing campaigns, personalize product offerings, and improve customer experience.

- **Fraud Detection**: Data science can be used to detect fraudulent activities by analyzing patterns and anomalies in financial transactions, customer behavior, and other data sources. Machine learning algorithms can be trained to detect and flag potential fraud cases, which can save businesses significant amounts of money and prevent reputational damage.

- **House Price Prediction:** Data science can be used to predict the price of a house based on various factors such as location, size, age, etc. By analyzing historical house prices and other relevant data, machine learning models can be trained to make accurate predictions of future house prices. This can be useful for both buyers and sellers in making informed decisions.

Applications of Data Science

- **Product Recommendation:** Data science can be used to recommend products to customers based on their previous purchases, browsing history, and other data sources. By analyzing customer behavior and preferences, machine learning algorithms can be trained to suggest products that are more likely to be of interest to a particular customer.

- **Stock Price Prediction**: Data science can be used to predict the future price of a stock by analyzing various data sources such as historical stock prices, news articles, and social media sentiment. By using machine learning models, traders and investors can make informed decisions about buying, selling, or holding stocks.

- **Weather Prediction**: Data science can be used to forecast the weather conditions for a specific location and time by analyzing various meteorological data sources such as temperature, humidity, and wind patterns. By using predictive models, weather forecasters can provide accurate and timely information to the public, which can be useful for planning outdoor activities, transportation, and emergency response.

Introduction to Data Science

1) Data science brings together mathematics, statistics, and computer science with the goal to generate insights and applications from data.

2) Tools and techniques that are required to understand data and retrieve useful information from data

3) Extracting Meaningful Patterns.

AI, ML and DS

- **Artificial intelligence** is about **giving machines the capability of mimicking human behavior,** particularly cognitive functions such as facial recognition, automated driving and sorting mail based on postal code.
- **Machine learning** is a subfield or tool of artificial intelligence which **providing machines with the capability of learning from data** without being explicitly programmed.
- **Data science** is a field that applies the **tools of artificial intelligence and machine learning to extract insights and value from data**. It involves a combination of data collection, cleaning, and preparation, as well as the application of machine learning algorithms to gain insights from the data.

AI, MACHINE LEARNING, AND DATA SCIENCE

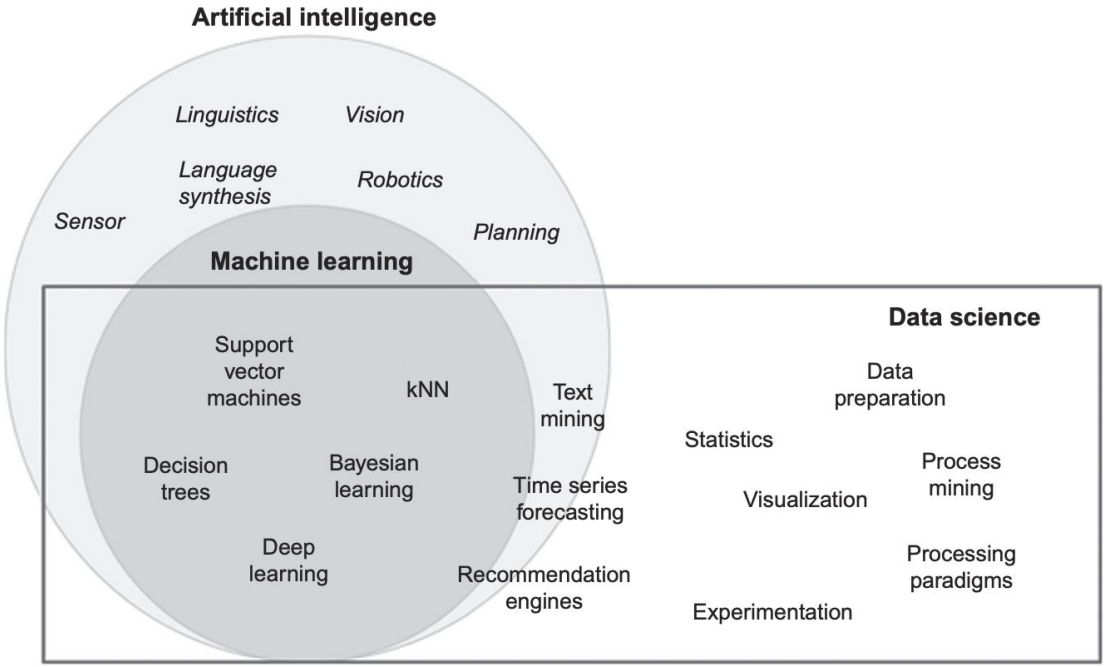


FIGURE 1.1
Artificial intelligence, machine learning, and data science.

Traditional Vs ML

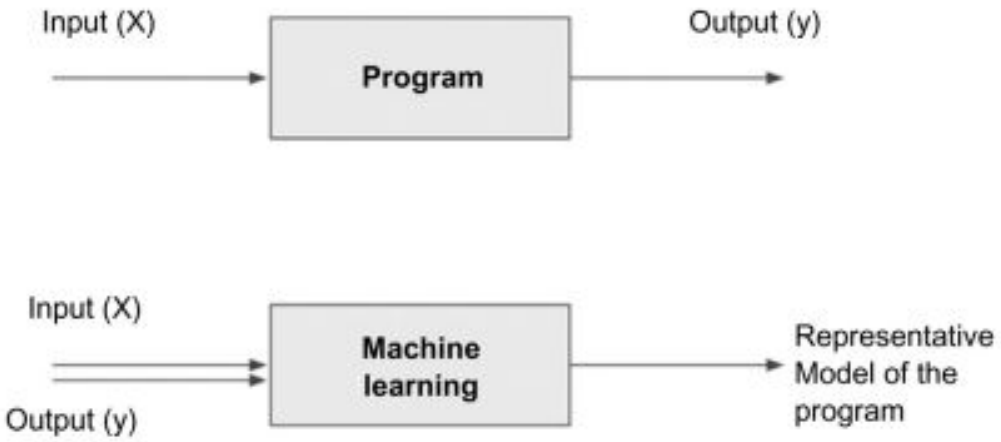


FIGURE 1.2
Traditional program and machine learning.

Data Science

- Data science is a field that **uses specialized computational methods to discover meaningful and useful structures within a dataset**

- It starts with **data,** which can range from simple arrays to complex matrices with **millions of observations and thousands of variables.**

- **Data science** is closely associated with related areas such as **database systems, data engineering, visualization, data analysis, experimentation, and business intelligence.**

- The key features and motivations of data science is to **discover useful patterns and insights** from the data and to **provide actionable knowledge for decision-making** and **business value**.

Data Science Lifecycle

Building Representative Models

- A **model in statistics** is a representation of the **relationship between variables in a dataset.**
- **Modeling** is the process of **creating a representative abstraction** from observed data.
- In data science, **a model** is used for b**oth predictive and explanatory purposes.**
- A model can **predict the value of an output variable** based on new, unseen input variables.
- A model can also be **used to understand the relationship between the output and input variables.**

Learning Algorithms

- **Data science** is the process of **discovering previously unknown patterns** in data using **automatic iterative methods**.
- **Many of these algorithms** are from the fields of **machine learning and artificial intelligence** and are based on **Bayesian probabilistic theories** and **regression analysis.**
- These iterative algorithms **automate the process of searching for an optimal solution** for a given data problem.
- Data science can be classified into tasks such as **classification, association analysis, clustering, and regression.**
- **Each task** uses **specific learning algorithms like decision trees, neural networks, k-nearest neighbors (k-NN), and k-means clustering.**

Data Science Tasks

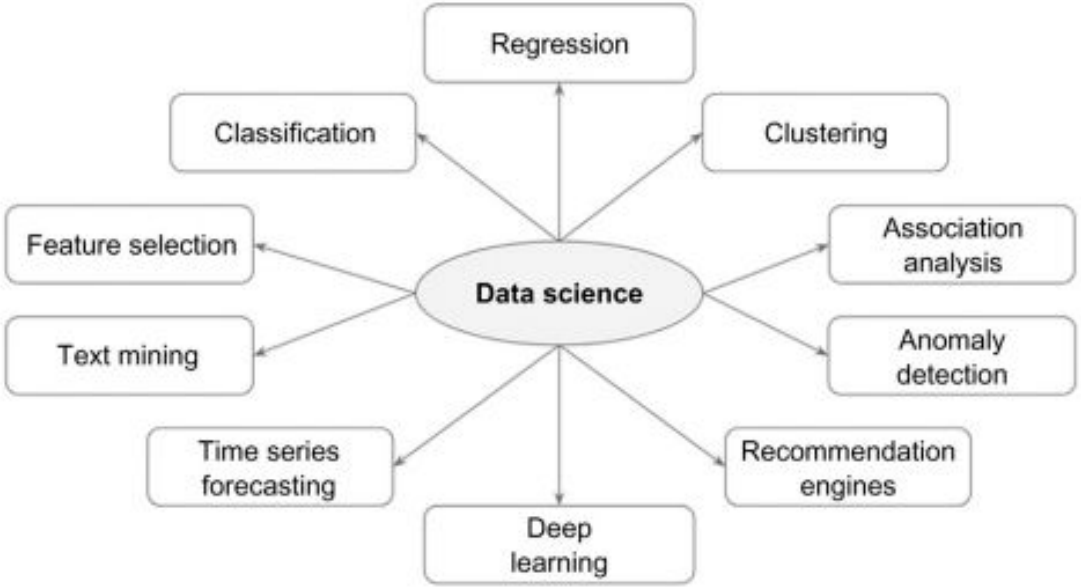


FIGURE 1.4
Data science tasks.

Data Science Tasks and Examples:

Table 1.1 Data Science Tasks and Examples

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of the predefined classes. The prediction will be based on learning from a known dataset	Decision trees, neural networks, Bayesian models, induction rules, *k*-nearest neighbors	Assigning voters into known buckets by political parties, e.g., soccer moms Bucketing new customers into one of the known customer groups
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from a known dataset	Linear regression, logistic regression	Predicting the unemployment rate for the next year Estimating insurance premium
Anomaly detection	Predict if a data point is an outlier compared to other data points in the dataset	Distance-based, density-based, LOF	Detecting fraudulent credit card transactions and network intrusion
Time series forecasting	Predict the value of the target variable for a future timeframe based on historical values	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the dataset based on inherit properties within the dataset	*k*-Means, density-based clustering (e.g., DBSCAN)	Finding customer segments in a company based on transaction, web, and customer call data
Association analysis	Identify relationships within an item set based on transaction data	FP-growth algorithm, a priori algorithm	Finding cross-selling opportunities for a retailer based on transaction purchase history
Recommendation engines	Predict the preference of an item for a user	Collaborative filtering, content-based filtering, hybrid recommenders	Finding the top recommended movies for a user

CRISP-DM (Cross-Industry Standard Process for Data Mining)

The CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely-used framework for data mining and machine learning projects. It provides a s**tructured approach for working through a project**, from understanding the problem and the data to building and deploying a model. The CRISP-DM framework consists of six main steps:

1. **Business Understanding:** Define the problem, objectives, and success criteria for the project, and understand the context of the project in the organization.
2. **Data Understanding**: Collect, review and explore the data to understand its quality, completeness, and relevance to the problem.
3. **Data Preparation**: Clean, transform, and preprocess the data so that it can be used in the model-building process.
4. **Modeling:** Apply statistical, machine learning, and data mining techniques to build and evaluate models that best solve the problem.
5. **Evaluation:** Assess the performance of the model and evaluate its effectiveness in solving the problem, and compare it with other models.
6. **Deployment**: Plan and implement the model in a production environment and monitor its performance over time.

Data Science Process

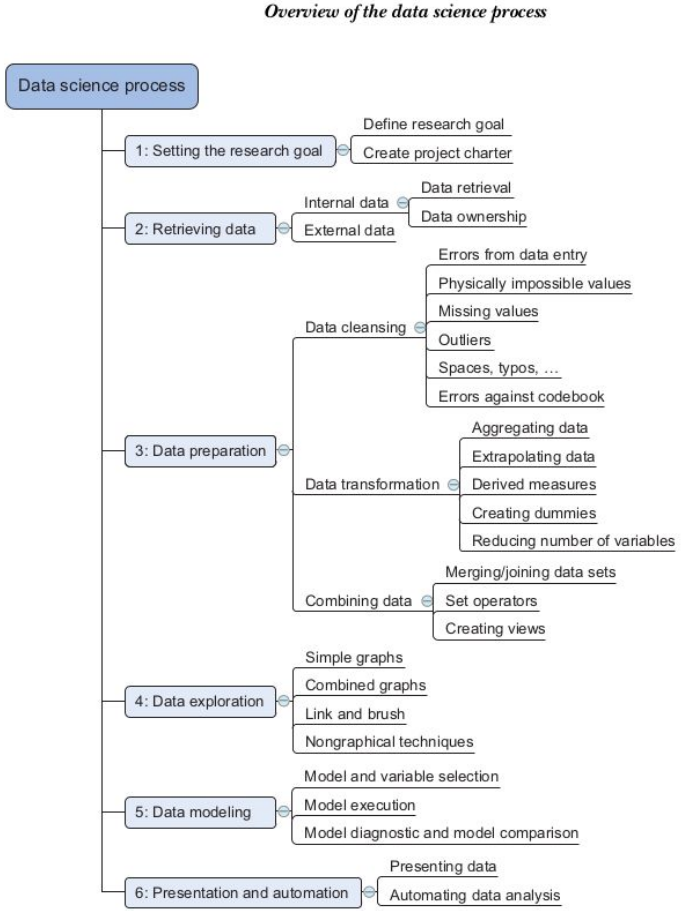


Figure 2.1 The six steps of the data science process

Step 1: Prior Knowledge

- Prior knowledge refers to **existing information about a subject** that is used to define the problem being solved in the data science process.
- It **helps to understand the business context** and **determine what data is needed** to solve the problem. E.g. Consumer Loan Business
  - **Business Understanding**
  - **Data Understanding**
- The outcome should be a clear **research goal**, a **good understanding of the context**, **well-defined deliverables,** and a **plan of action** with a timetable.

Data Understanding

Table 2.1 Dataset

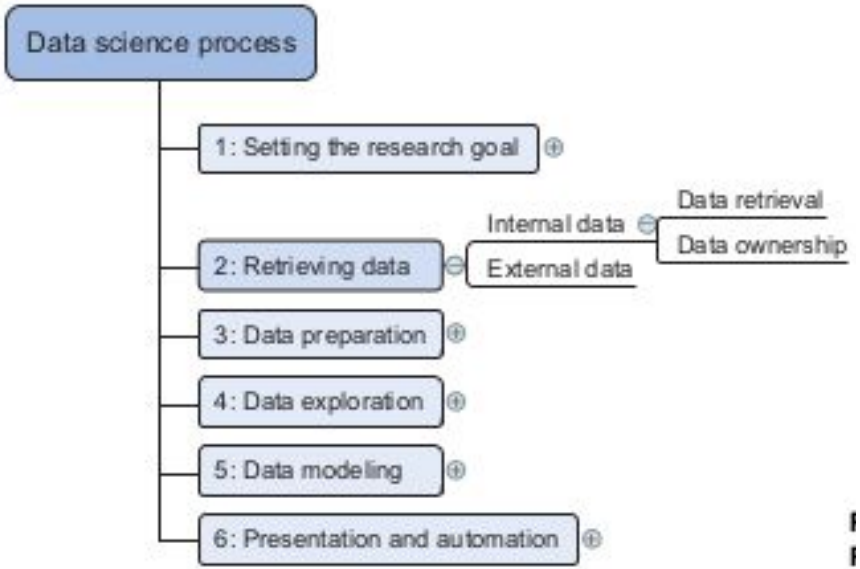
Borrower ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

Table 2.2 New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

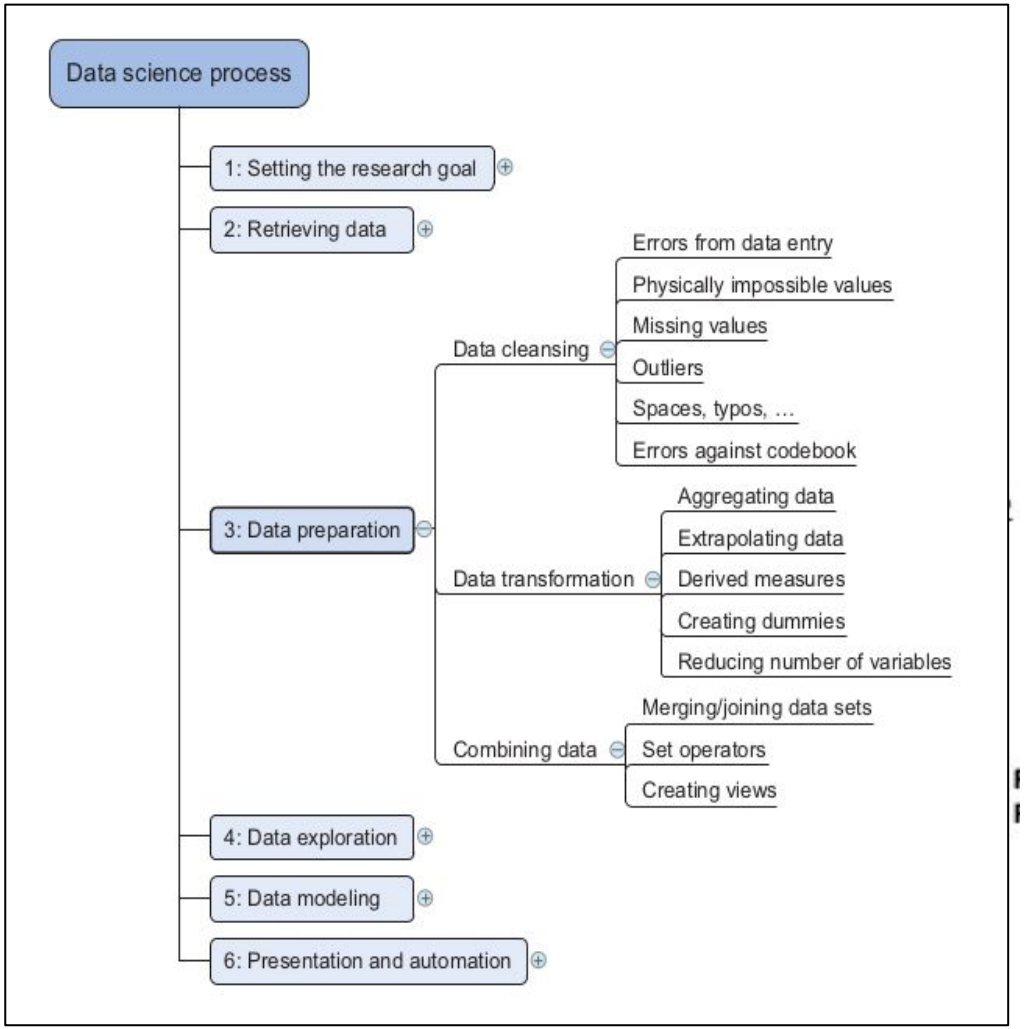
- **A dataset** (example set) is a collection of data with a defined structure.
- **A data point** (record, object or example) is a single instance in the dataset.
- **An attribute** (feature, input, dimension, variable, or predictor) is a single property of the dataset.
- **A label** (class label, output, prediction, target, or response) is **the special attribute to be predicted** based on all the input attributes. E.g. Interest rate
- **Identifiers** are special attributes that are used for locating or providing context to individual records.

Step 2: Retrieving data



Step 3: Data Preparation

- Preparing the dataset is the **most time-consuming part** of the data science process .
- It requires transforming the data into a tabular format with records in rows and attributes in columns.
- This may involve **applying functions such as pivot, type conversion, join, or transpose** to condition the data into the required structure.



Step 3 : Data Preparation
Cleansing, integrating, and transforming data.

Data Quality

- **Data quality is an ongoing concern** throughout the data collection, processing, and storage process
- **Errors in data** can impact the representativeness of the model .
- Organization uses **data alerts, cleansing, and transformation techniques** to improve and manage the quality of data
- One of the most common data quality issues is **missing attribute values** in some records
- The first step of managing missing values is to **understand the reason behind why the values are missing**, by tracking the **data lineage (provenance)** of the data source

Missing Values

- **Missing values** can be substituted with artificial data, such as the **mean, minimum or maximum value, depending on the characteristics of the attribute**
- **To build a representative model,** all the data records with **missing values or poor data quality can be ignored,** reducing the size of the dataset

- **k-nearest neighbor (k-NN)** algorithm for classification tasks are often **robust with missing values,** while **neural network models** for classification tasks **do not perform well with missing attributes,** and thus, the data preparation step is essential for developing neural network models.

Data Types and Conversion

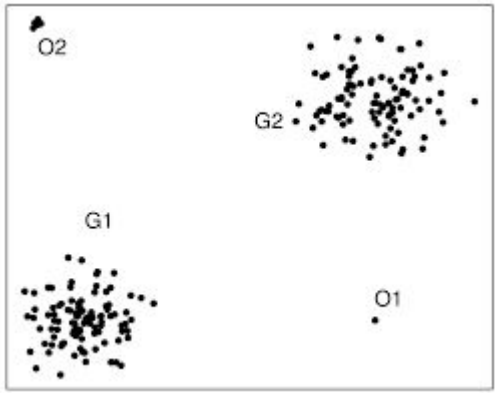
- The attributes in a dataset can be of different types such as **continuous** (interest rate), **integer numeric** (credit score), or **categorical.** For example, the credit score can be expressed as categorical values (poor, good,excellent) or numeric score.
- Different data science algorithms impose different restrictions on the attribute data types
- For example, **linear regression models** require **input attributes to be numeric**, so categorical attributes must be converted to continuous numeric attributes
- This can be done by **encoding a specific numeric score for each category value** or by using a technique called **binning,** where a **range of values are specified for each category**
- **Binning** is a process where a **range of values are grouped into a specific category. E.g. a score between 400 and 500 can be encoded as “low”**

Data Transformation

- Some data science algorithms like k-NN require input attributes to be numeric and normalized, because the **algorithm compares the values of different attributes** and **calculates distance between the data points**
- **Normalization** prevents one attribute dominating the distance results because of large values.
- For example, consider income (expressed in USD, in thousands) and credit score (in hundreds), distance calculation will always be dominated by slight variations in income
- **One solution** is to **convert the range of income and credit score to a more uniform scale from 0 to 1 by normalization**, this way a consistent comparison can be made between the two different attributes with different units.

Outliers

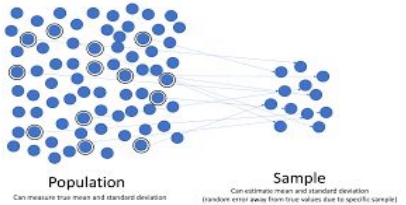
- Outliers are **anomalies in a given dataset.**
- Occurs because of **erroneous data capture (human height as 1.73 cm instead of 1.73 m).**
- **Detecting outliers :**primary purpose of some data science applications, like fraud or intrusion detection.



Feature selection

- Many datasets have hundreds of attributes.
- **Not all the attributes are equally important** or useful in predicting the target
- **Reducing the number of attributes,** without significant loss in the performance of the model, is called **feature selection.**
- It leads to a **more simplified model** and **helps to synthesize a more effective explanation of the model.**

Data Sampling



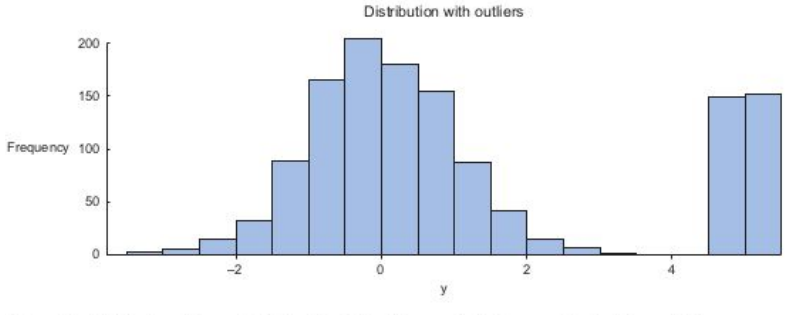
- Sampling is a process of **selecting a subset of records as a representation** of the original dataset for use in data analysis or modeling
- Sampling r**educes the amount of data** that need to be processed and speeds up the build process of the modeling
- Mostly It is **sufficient to work with samples to gain insights**, extract the information, and to build representative predictive models
- In the build process for data science applications, it is **necessary to segment the datasets into training and test samples** using simple sampling or class label specific sampling
- **Stratified sampling** is a process of **sampling where each class is equally represented** in the sample, which allows the model to focus on the difference between the patterns of each class, such as normal and outlier records
- **In classification applications**, sampling is used to **create multiple base models,** each developed using a different set of sampled training datasets
- **These base models are used to build one meta model, called the ensemble model,** where the error rate is improved when compared to that of the base models.

Data Preparation

OUTLIERS:

Outliers are anomalies in a given dataset. Outliers may occur because of correct data capture (few people with income in tens of millions) or erroneous data capture

An outlier is an observation that seems to be distant from other observations or, more specifically, one observation that follows a different logic or generative process than the other observations. The easiest way to find outliers is to use a plot or a table with the minimum and maximum values.



Combining the Data

Your data comes from several different places, and in this substep we focus on integrating these different sources. Data varies in size, type, and structure, ranging from databases and Excel files to text document

You can perform two operations to combine information from different data sets.

The first operation is joining: enriching an observation from one table with information from another table.

The second operation is appending or stacking: adding the observations of one table to those of another table.

Transformation

Transforming your data so it takes a suitable form for data modeling.

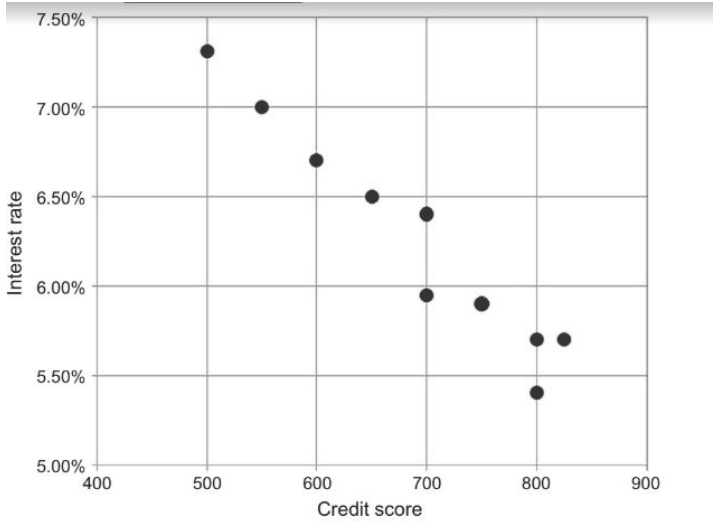
Transforming the input variables greatly simplifies the estimation problem.

Other times you might want to combine two variables into a new variable.

Sometimes you have too many variables and need to reduce the number because they don’t add new information to the model.

Step 4: Data Exploration

- EDA - It is **Exploration of the data** and gaining a better understanding of the dataset through a process called exploratory data analysis (EDA)
- EDA involves **computing descriptive statistics** and **visualizing** the data to **expose the structure, distribution, and relationships** within the dataset.
- **Descriptive statistics** such as
  - mean,
  - median,
  - mode,
  - standard deviation,
  - range provide a summary of the key characteristics of the data
- **Visual plots** offer an instant grasp of the data points, for example, a **scatterplot** can show the relationship between **credit score and loan interest rate** whereas credit score increases, interest rate decreases.



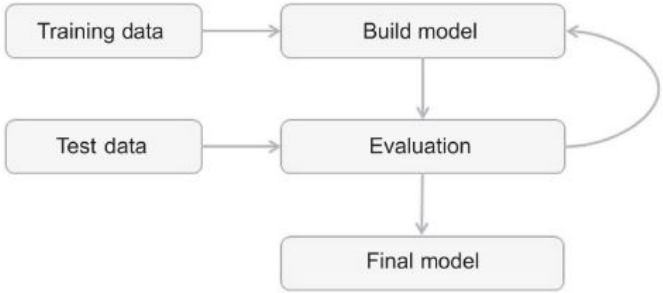
Scatter Plot

Step 4: Exploratory data analysis :It is to use graphical techniques to gain an understanding of your data and the interactions between variables.

This step helps us to discover anomalies which have been missed before, forcing you to take a step back and fix them.

Simple graphs
Combined graphs
Link and brush
Non graphical techniques(Tabulation, clustering,)

Step 5: Modeling



- **A model** is an **abstract representation** of the data and the relationships in a given dataset.
- **Modeling phase** of predictive data science consists of several steps, like **learning the model, evaluating it,** etc.
- Both predictive and descriptive models have an evaluation step.

Modeling

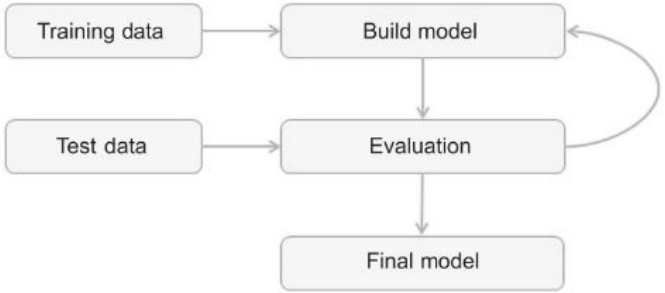


Table 2.3 Training Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*) (%)
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

Table 2.4 Test Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*)
04	700	6.40
07	750	5.90
10	825	5.70

Training and Testing Datasets

- **The modeling step** creates a representative model inferred from the data, using a training dataset with known attributes and target.
- The validity of the created model will also need to be checked with another known dataset called the **test dataset or validation dataset**
- To facilitate this process, the overall known **dataset can be split into a training dataset** and a **test dataset**.
- Standard rule of thumb of **two-thirds of the data for training** and **one-third for test dataset**

Training and Testing Datasets

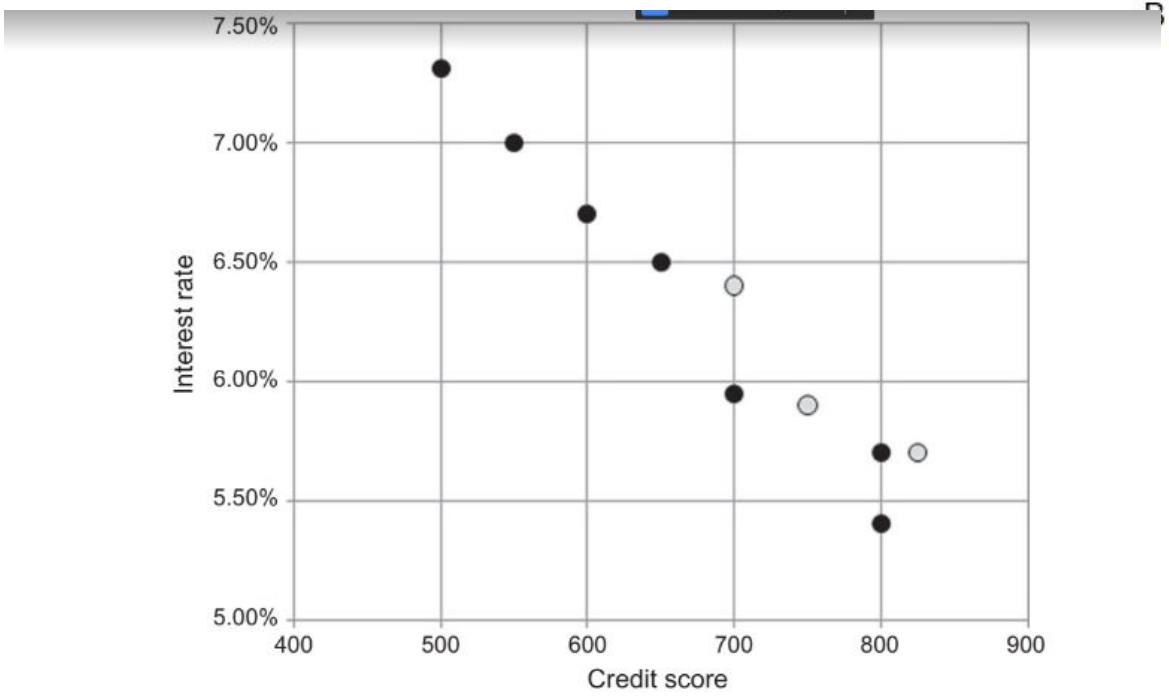


FIGURE 2.5
Scatterplot of training and test data.

Learning Algorithms

- For the interest rate dataset, the **simple linear regression** for the interest rate (y) has been calculated and using this model, the interest rate for a new borrower with a specific credit score can be calculated.
- y = a * x + b
- y is the **output** or **dependent variable**
- **x** is the **input or independent variable,**

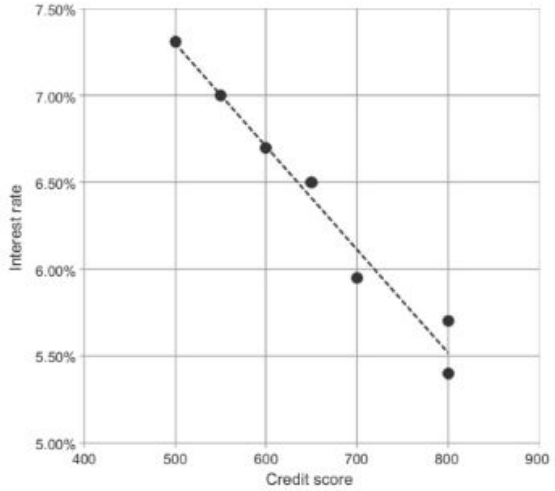


FIGURE 2.6
Regression model.

Evaluation

- As long as the error is acceptable, this model is ready for deployment and the error rate can be used to compare this model with other models developed using different algorithms.
-

Table 2.5 Evaluation of Test Dataset

Borrower	Credit Score (X)	Interest Rate (Y) (%)	Model Predicted (Y) (%)	Model Error (%)
04	700	6.40	6.11	−0.29
07	750	5.90	5.81	−0.09
10	825	5.70	5.37	−0.33

Ensemble Model

- **Ensemble modeling** is a process where **multiple diverse base models** are used **to predict an outcome**
- The motivation for using ensemble models is t**o reduce the generalization error of the prediction**
- As long as the base models are diverse and independent, the prediction error decreases when the ensemble approach is used
- The approach seeks the wisdom of crowds in making a prediction

Data Science Process-(end-to-end, multi-step, iterative process)

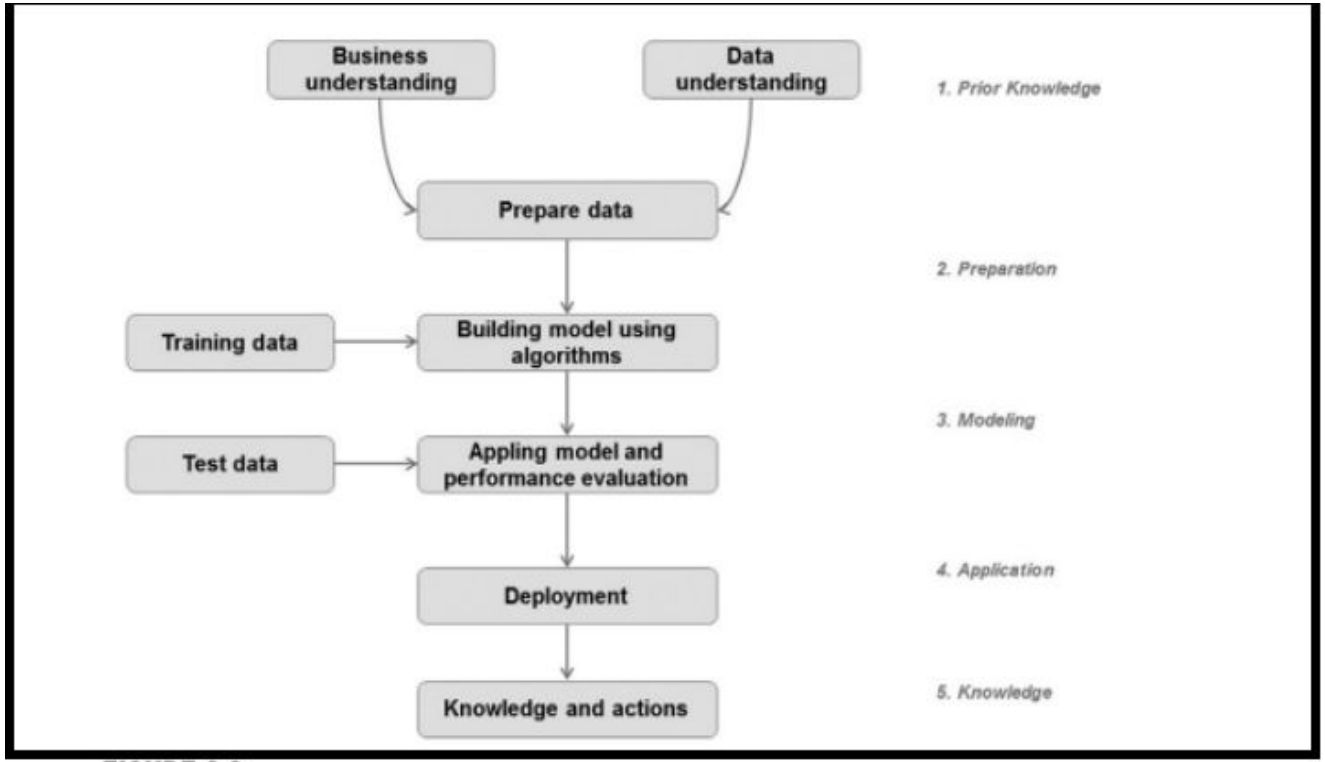


FIGURE 2.2

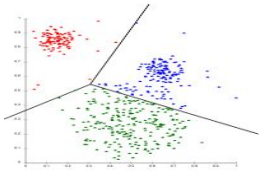
Data science process.

Data Science Process - Revision

1. **Analyzed** the business question
2. **Sourced the data** relevant to answer the question
3. **Selected a data science technique** to answer the question
4. **Picked a data science algorithm** and prepared the data to suit the algorithm
5. **Split the data** into training and test datasets
6. **Built a generalized model** from the training dataset
7. **Validated the model** against the **test dataset**

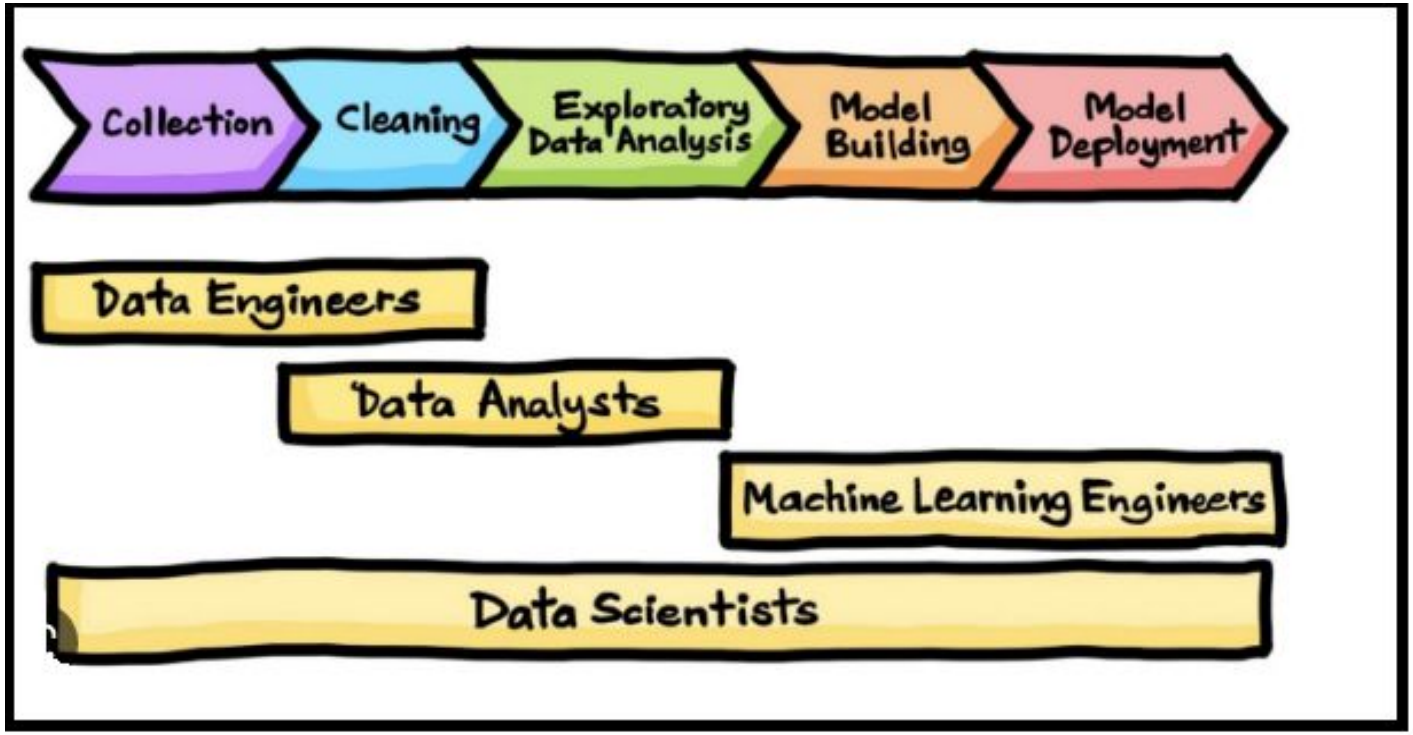
   This model can now be **used to predict the interest rate of new borrowers** by integrating it in the actual loan approval process.

Applications



1. **Fraud Detection**: Developing models to **identify patterns and anomalies** in financial transaction data that could indicate fraudulent activity.
2. **Customer Segmentation**: Analyzing customer data to identify different groups of customers with similar characteristics, which can be used to target marketing efforts.
3. **Predictive Maintenance:** Building models that use sensor data to predict when equipment is likely to fail, so that maintenance can be scheduled proactively.
4. **Image Recognition**: Developing models that can automatically identify objects, people, or scenes in images, which can be used in a variety of applications such as self-driving cars, security systems, and medical imaging.

Data Science Process



Difference Between Data Science and Data Analytics

Data Science :

- **Data science** is a broad field that includes **data analytics as well as machine learning, statistical modelling, and programming.**
- Data science goes beyond to include **prediction, causal inference, and the development** of data-driven products and systems.

Data Analytics:

- **Data analytics** is focused on the **examination, cleaning, and transformation** of data, as well as the discovery of insights and the communication of findings.
- **Data analytics** is a **subset of data science.**

Exercise - customer segmentation for Telecommunication company

1. What is the **problem or question** that needs to be addressed?
2. What **data is available** and **how will it be collected**?
3. What are the p**otential data sources** and how will they be **cleaned and preprocessed?**
4. What are the r**elevant features** or **variables** in the data?
5. How will the data be **explored and analyzed**?
6. What are the **potential models** or algorithms to be used?
7. How will the **model be evaluated and tested**?
8. How will the **model be deployed and implemented**?
9. How will the **performance of the model** be monitored and maintained?
10. How will the **insights gained from the analysis be communicated** and used to drive decision making?

Difference between data science and data analytics

Two sides of the same coin

Data Science is a combination of multiple disciplines – Mathematics, Statistics, Computer Science, Information Science, Machine Learning, and Artificial Intelligence.	On the other hand, **data analytics** is mainly concerned with Statistics, Mathematics, and Statistical Analysis.
Data science deals with explorations and new innovations.	**Data Analysis** makes use of existing resources.
Data Science mostly deals with unstructured data.	Data Analytics deals with structured data.
Data scientists, on the other hand, design and construct new processes for data modeling and production using prototypes, algorithms, predictive models, and custom analysis	Data analysts examine large data sets to identify trends, develop charts, and create visual presentations to help businesses make more strategic decisions.

Data Exploration:

— known as exploratory data analysis

—--To expose the structure of data,distribution of values,presence of extreme values,inter relationship within the dataset.

—---data exploration to study the basic characteristics of a dataset.

—-- provides a set of tools to obtain fundamental understanding of a dataset.

—--- computing statistics like mean and deviation, and plotting data as a line, bar, and scatter charts are part of data exploration techniques

Data exploration can be broadly classified into two types—

1) descriptive statistics and
2) data visualization.

Descriptive Statistics:

Definition

Statistics[1] *is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions.*

Major Branches of Statistics

1) **Descriptive Statistics**
2) **Inferential statistics**

Major branches of statistics

1. Description

Definition
The part of statistics concerned with the description and summarization of data is called descriptive statistics.

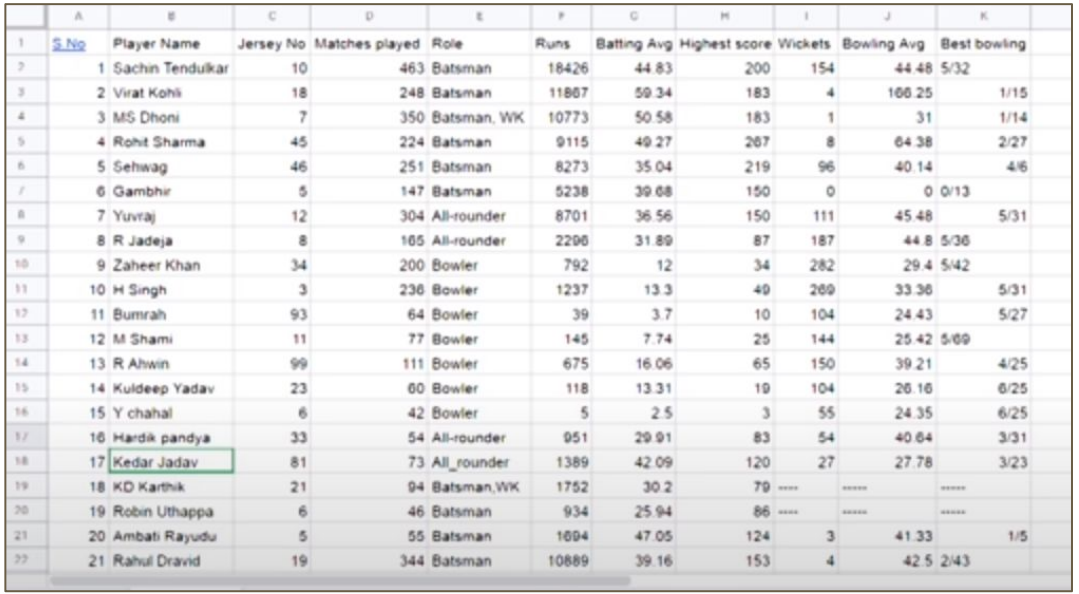
2. Inference

Definition
The part of statistics concerned with the drawing of conclusions from data is called inferential statistics.

- To be able to draw a conclusion from the data, we must take into account the possibility of chance- introduction to probability.

When would you use the descriptive or inferential statistics?

- If the purpose of the analysis[2] is to examine and explore information for its own intrinsic interest only, the study is descriptive.
- If the information is obtained from a sample of a population and the purpose of the study is to use that information to draw conclusions about the population, the study is inferential.
- A descriptive study may be performed either on a sample or on a population.
- When an inference is made about the population, based on information obtained from the sample, does the study become inferential.



S.No	Player Name	Jersey No	Matches played	Role	Runs	Batting Avg	Highest score	Wickets	Bowling Avg	Best bowling
1	Sachin Tendulkar	10	463	Batsman	18426	44.83	200	154	44.48	5/32
2	Virat Kohli	18	248	Batsman	11867	59.34	183	4	166.25	1/15
3	MS Dhoni	7	350	Batsman, WK	10773	50.58	183	1	31	1/14
4	Rohit Sharma	45	224	Batsman	9115	49.27	267	8	64.38	2/27
5	Sehwag	46	251	Batsman	8273	35.04	219	96	40.14	4/6
6	Gambhir	5	147	Batsman	5238	39.68	150	0	0	0/13
7	Yuvraj	12	304	All-rounder	8701	36.56	150	111	45.48	5/31
8	R Jadeja	8	165	All-rounder	2296	31.89	87	187	44.8	5/36
9	Zaheer Khan	34	200	Bowler	792	12	34	282	29.4	5/42
10	H Singh	3	236	Bowler	1237	13.3	49	269	33.36	5/31
11	Bumrah	93	64	Bowler	39	3.7	10	104	24.43	5/27
12	M Shami	11	77	Bowler	145	7.74	25	144	25.42	5/69
13	R Ahwin	99	111	Bowler	675	16.06	65	150	39.21	4/25
14	Kuldeep Yadav	23	60	Bowler	118	13.31	19	104	26.16	6/25
15	Y chahal	6	42	Bowler	5	2.5	3	55	24.35	6/25
16	Hardik pandya	33	54	All-rounder	951	29.91	83	54	40.64	3/31
17	Kedar Jadav	81	73	All_rounder	1389	42.09	120	27	27.78	3/23
18	KD Karthik	21	94	Batsman,WK	1752	30.2	79	----	-----	-----
19	Robin Uthappa	6	46	Batsman	934	25.94	86	----	-----	-----
20	Ambati Rayudu	5	55	Batsman	1694	47.05	124	3	41.33	1/5
21	Rahul Dravid	19	344	Batsman	10889	39.16	153	4	42.5	2/43

Who has played? Who has taken highest wickets?

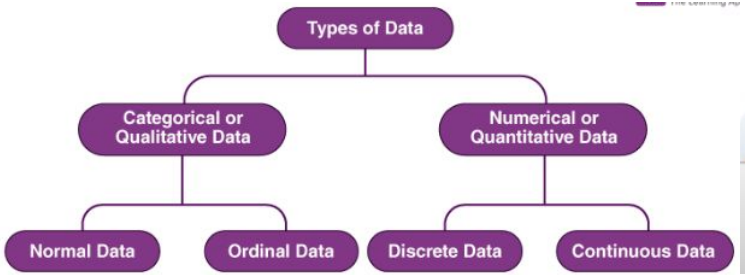
What is the total runs? The highest score etc

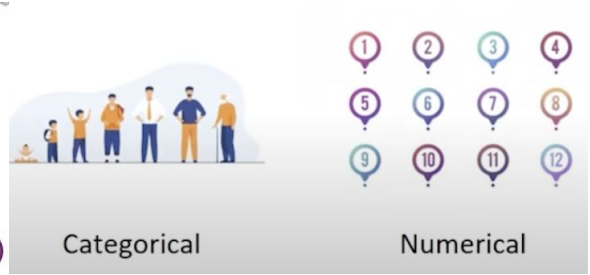
Types of data

Unstructured and structured data

- For the information in a database to be useful, we must know the context of the numbers and text it holds.
- When they are scattered about with no structure, the information is of very little use.
- Hence, we need to organize data

- Rows represent cases: for each case, same attribute is recorded
- Columns represent variables: For each variables, same type of value for each case is recorded.





Categorical : 1)Nominal (no order) 2)Ordinal(Order or rank)
Numerical: 1)Interval 2)ratio

- Categorical data
  - Also called qualitative variables.
  - Identify group membership
- Numerical data
  - Also called quantitative variables.
  - Describe numerical properties of cases
  - Have measurement units

data:

- Time series - data recorded over time
- Timeplot – graph of a time series showing values in chronological order
- Cross-sectional - data observed at the same time