Data Science for Business Professionals

A Practical Guide for Beginners

by

Probyto Data Science and Consulting Pvt. Ltd.

Table of Contents

CHAPTER 1
Data Science Overview

In ancient times land was the most important asset in the world. In the modern era, machines and factories became more important than land. In the twenty-first century, however, data will eclipse both land and machinery as the most important asset.

- Yuval Noah Harari

Data is undoubtedly the most valuable asset of our society. It captures our entire understanding of mankind, space, nature, and learnings from all our existence. In modern days, we have developed tools and methods to understand that data for path-breaking discoveries in medicine, space, machine, and so on. Now, in the 21st century, we are ready to use the technology and data to catapult mankind into the data age. It is not an exaggeration in any sense, just compare how our life is different from our grandparents' life, and you would see how technology and data have changed the way we live and work.

This book titled *Introduction to Data Science for Business - A Practical Guide for Freshers* is an attempt to cover the practical scope of data science teams in real industry set-up. The book will touch upon major areas of work, how you lead there, their significance, and some examples to start hands-on training. The book will take you to a journey and significance of the multidisciplinary nature of data science.

In this chapter, we will have a preview of the books and their content to provide an overview of data science. The chapter will build the use case of what a newcomer to data science or fresher need to be aware of before starting the data science journey.

Structure

- Evolution of data analytics
- Define data science
- Domain knowledge

- Mathematical and scientific techniques
- Tools and technology
- Data science analysis types
- Data science job roles
- ML model development process
- Data visualizations
- Result communication
- Responsible and ethical AI
- Career in data science
- Summary

Objectives

After studying this chapter, you should be able to:

- Understand the fundamentals of data science and the importance of domain knowledge.
- Identify the mathematical techniques and technology required to build any data science application.
- Recognize the various opportunities around the data science application development process.
- Understand the importance of data visualization and how it can be used in result communication.

Evolution of data analytics

Data has been collected and analyzed for very old times. The information capture and dissemination have been part of managing kingdoms, having records of lands and army. The modern use of statistics started in the 18th century with systematic ways of capturing data, and the evolution of printing the system helped in storing data.

In the context of the book, we would see the journey of data analysis in the following stages:

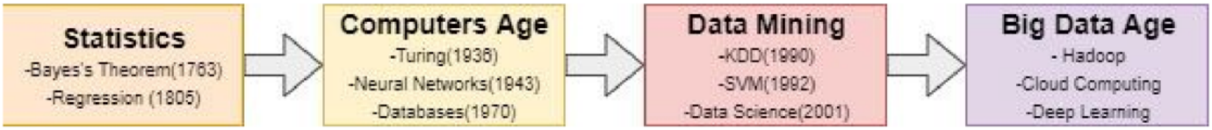


Figure 1.1: *Evolution of Data Analysis*

Statistics have been around for a long time sparingly across the world. The systematic development of statistics happened in the 18th century and was used in administrative purposes across the world. The great advancements in science during and post-industrial era set the foundation for the computer age. Starting with Alan Turing groundbreaking work in the Theory of computing and advancement in semiconductors, the computers started getting more powerful year on year.

During the mid-19th century, a lot of research work gone into understanding how a human brain learns and advancementsin the understanding structure of the brain. Early papers with Neural Networks emerged. The methodology to learn and repeat some events was entirely different from how a distribution based statistical methods explained. Databases also started getting into exclusive use by the 1980s. Digitization also started getting well recognized in the industry and government. Same time, early experiments with networking, emails, and interconnected web were emerging.

During the late 1990s, the computers were household things in the US with Microsoft having released a power operating system MS Windows 98, macOS was in the market as well. Same time enterprising computing was on rising with giants like IBM becoming the core providers of powerful servers and the internet getting its pace. This time Data mining become prominent for creating reports, analyzing customer data, and making decisions driven by basic data analysis (MS Excel was in Windows by that time). Knowledge databases, **Support Vector Machine (SVM),** SQL databases, and increased computing power market the early stages of Data Science before it exploded around 2008.

Hadoop, 2006, the distributed file storage and computation was a project started by visionary computer scientist, *Doug Cutting*, as he saw a huge wave of Big Data coming. By March 2009, Amazon had already started providing MapReduce hosting service, Elastic MapReduce. This started the era of Big Data; at the same time, GPU and cloud computing made the cost of computing cheap, leading to rapid development in deep learning and cloud infrastructure.

Today, we have a better understanding of how data science creates value for companies by making them data-driven. All the ecosystem and pre-requite to make the value of data are available now at a reasonable cost.

The data science as a discipline has grown now and had a universal appeal across the organizations and its benefits. The technology giants are shaping and directing the industry towards data-driven organizations. In *Figure 1.2*, you can see some of the biggest companies of today are based on the latest technology and data-driven decision making:

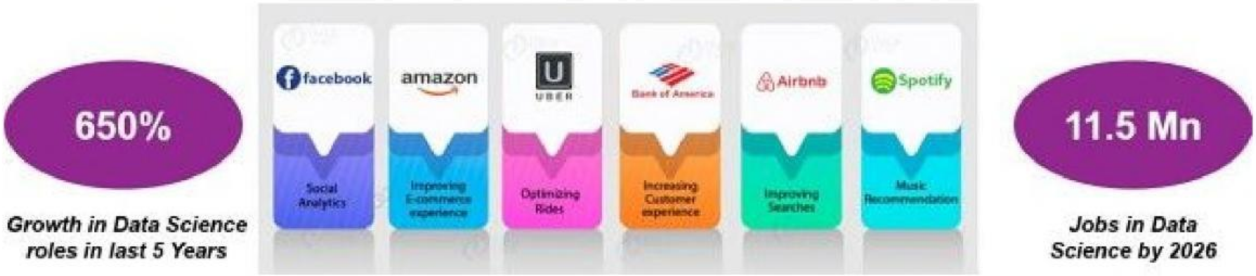

650%
Growth in Data Science roles in last 5 Years
facebook
Social Analytics
amazon
Improving E-commerce experience
UBER
Optimizing Rides
Bank of America
Increasing Customer experience
Airbnb
Improving Searches
Spotify
Music Recommendation
11.5 Mn
Jobs in Data Science by 2026

Figure 1.2: *Data Science Growth*

Some of the popular use cases in the industry are listed below for a reference. The reader must try to understand more about these use-cases and discover how data science adding direct value to the business:

- Fraud and risk detection
- Healthcare
- Internet search
- Targeted advertising
- Website recommendations
- Advanced image recognition
- Speech recognition
- Airline route planning
- Gaming
- Augmented reality
- Talent acquisition
- Credit scoring
- Price forecasting and many more.

The businesses are rapidly adopting new technologies and using AI/ML for competitive and comparative advantage. In the coming years, the adoption of data-driven features will be faster in governance and general public discourse.

Define data science

Data Science is the umbrella term that comprises the science and its application related to data. The early definition of data science was built on a combination of multiple disciplines, and moreover, it is expanding fields as data keeps adding value to multiple disciplines.

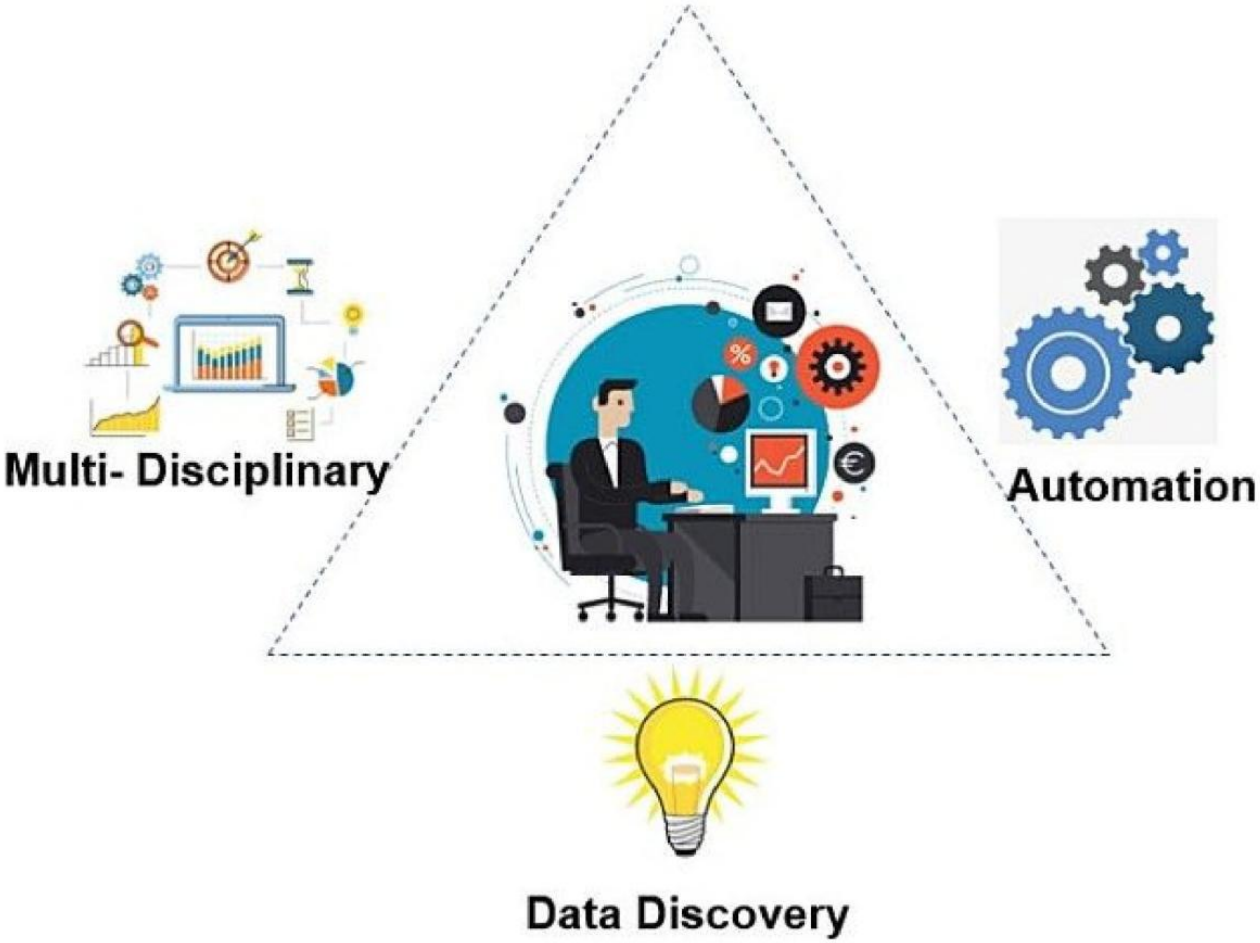


Figure 1.3: *Define Data Science*

The key features of these fields are as follows (as shown in *Figure 1.3*):

- **Multi-disciplinary:** A new discipline that combines aspects of mathematics, statistics, programming, and visualization.

- **Automation:** An automated way to analyze the enormous amount of data and extract information
- **Data discovery:** A powerful new way to make discoveries from data.

Further, data science as a discipline defines some foundational knowledge that the practitioner needs to have to be able to harness value from data. *Figure 1.4* is the Venn diagram which is a representation of such a practitioner:

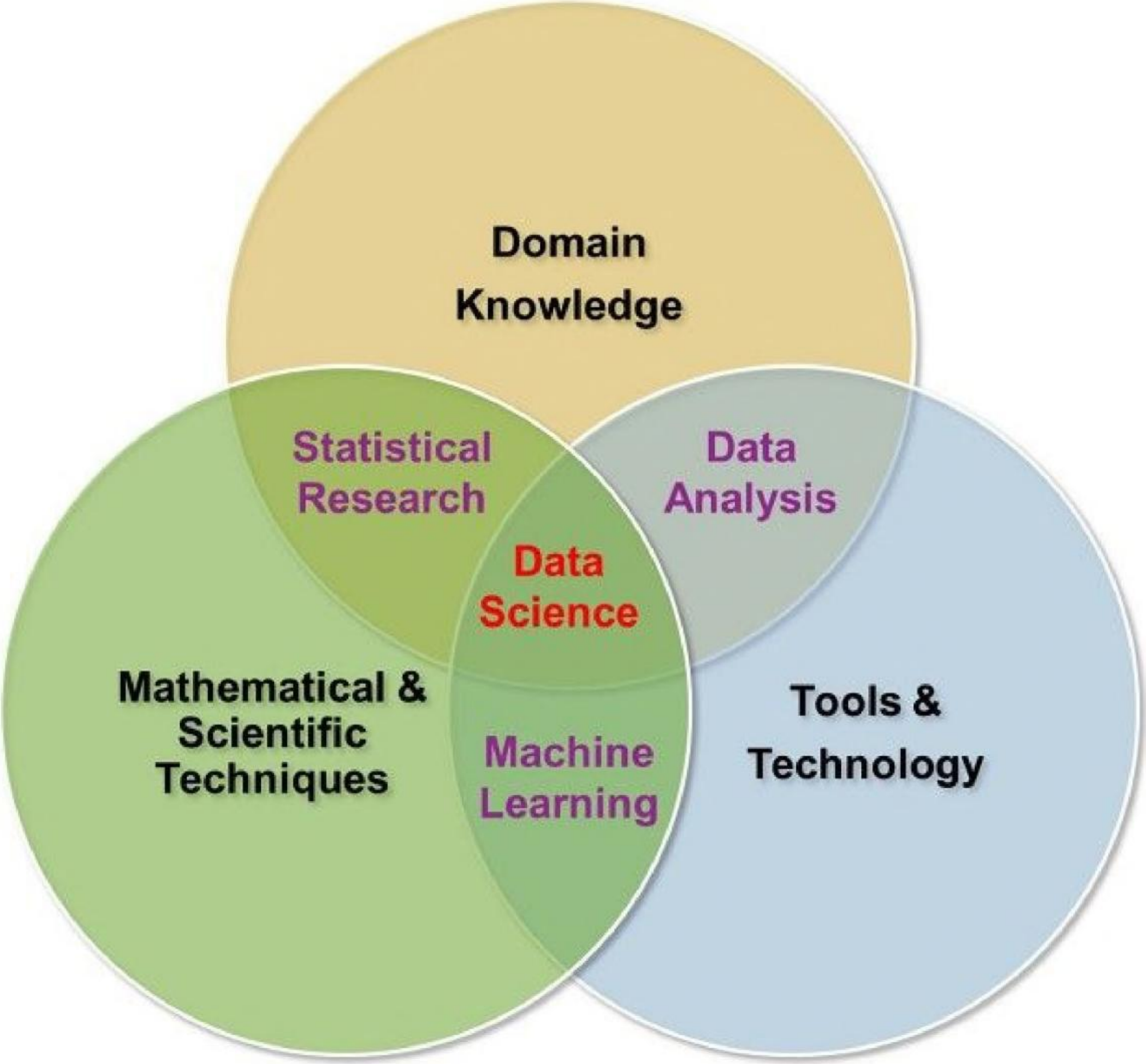


Figure 1.4: *Data Science Practitioner*

The core areas bring relevant expertise to help build data solutions;

- Domain knowledge
    - Engineering
    - Science
    - Business
    - Medicine
    - Economics
    - Finance
- Mathematical and scientific techniques
    - Linear algebra
    - Classic statistical tools like regression
    - Clustering and classification
    - Machine learning
- Tools and Technology
    - Programming
    - Operating systems
    - Analysis tools (R, SAS, Python)
    - Visualization tools (Tableau)

Note: The above list is just indicative of some very popular sub-areas for data scientists.

The different subareas within data science are not confined to traditional knowledge but also include the powerful method of experimentations and learning; he automated learning from machines is what we popularly call as *Machine Learning.* The tools and technology do the analysis at scale and provide the medium to deliver results to end-users.

Domain knowledge

Domain knowledge means the understanding of the domain from where the data problem/opportunity is originating. This is important as the data is a representation of a process or phenomenon captured by data. The data scientist is responsiblefor discovering the relationships in data in the context of that domain.

For illustration, assume we are working on a problem statement for a bank.

The problem: The bank gets of loan applications from potential borrowers. The loan officer must make a decision to issue a loan or not.

Now there are so many domain questions to be understood before starting the analysis:

- What are the terms and conditions to issue a loan?
- How different factors affect borrowing power?
- What are the chances of full recovery of loans?
- What data about the borrower can we capture as per statutory and regulations?
- What are market conditions now? How they define the relationship with loan recovery?

And so many other aspects of the banking process and its implications on the loan approval process. Economics and bank policies need to be part of the analysis. If we do not have this background, we would just crunch the data without any context and applicability of results.

Mathematical and scientific techniques

Domain knowledge will provide the context of the data and the desired results from the data science process. Mathematical and scientific techniques provide a theoretical understanding of how to quantify the behavior the business wants to investigate the data.

For Illustration, in the previous example, we had generic domain questions. For instance,

Question: What are the chances of full recovery of a loan?

The domain expert would look at application details and, based on experience, can say it's *High*. But how HIGH? He would not be able to quantify until he has some statistical technique to define and calculate HIGH. This is the area where techniques help the domain understanding gets a quantifiable number to them, so that decision boundaries can be created.

Machine learning algorithms are the techniques of learning the relationships among datasets automatically and build functional relationships for future predictions. *Figure 1.5* shows how machine learning differs from the standard computer application:

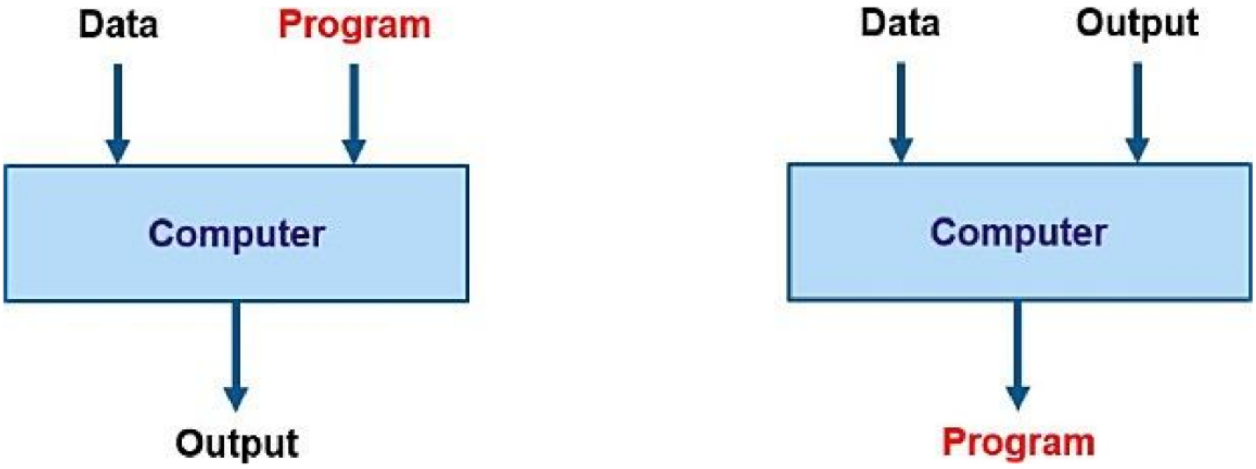


***Figure 1.5:** Traditional Rule-Based Vs. Machine Learning*

The expert systems are driven by intuition and experience of experts. In those cases, the computer is fed with input data, and a program (collection of rules/logic) and an output is generated. Referring to our previous section loan example, the loan officer will set a bar that I will only issue a loan of less than $25,000. Then the input becomes the loan application amount. The program becomes a rule that if more than $25,000, then reject or accept.

Now, this sounds a good way to start deciding the loan applications by making decisions with numbers. However, what we are missing here is the lack of evidence that $25,000 and more loan recovery are very bad. How does the loan office assume this number/logic? May be more than $25,000 loans are very profitable for the bank as they repay.

Hence, we switch to the second approach, which does not assume anything beforehand. It takes inputs (loan amount), the output (If repaid or not), and then creates a program to identify with high accuracy which bank of the loan amount is riskier and by how much. This way, the machine learning models generate the program learning from data, and then this program can be used in the traditional system to generate output for new entities.

Machine learning can be of any of three types as below:

- **Supervised learning:** We have enough historical data of input and output pairs to learn the relationship.
- **Unsupervised learning:** We do not have an output to optimize the behavior relationship but discover the internal structure of the data.

- **Re-enforcement learning:** We keep learning from feedback from the output. This allows us to have continuous learning.

Figure 1.6 depicts the type of machine learning algorithms, and each of these classes of algorithms has various types of algorithms, used for various purposes to solve different types of problems:

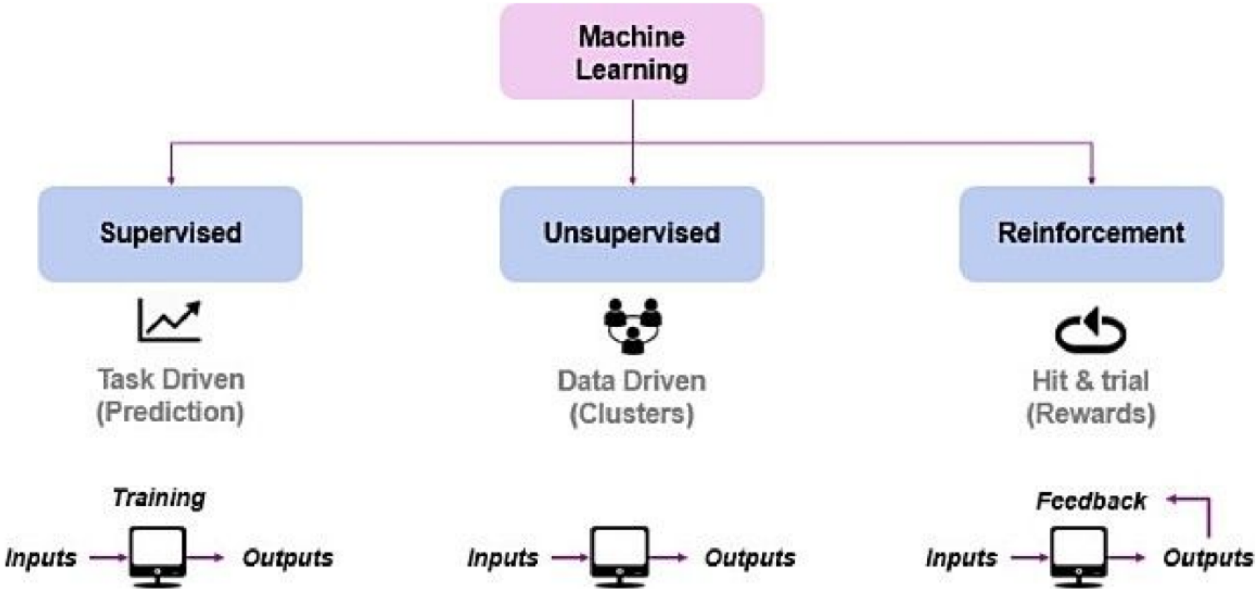


Figure 1.6: *Types of Machine Learning*

The *Machine Learning* section of the book will discuss aspects of machine learning in detail. Below table enumerates some popular algorithm of each type:

Algorithm category	Examples
Supervised algorithms	Naive Bayes Decision Trees Linear Regression Support Vector Machines (SVM) Neural Networks
Unsupervised algorithms	k-means clustering Association Rules Self-Organizing Maps Principle Component Analysis
Reinforcement learning	Q-Learning Temporal Difference (TD) Deep Adversarial Networks

***Table 1.1:** Machine learning algorithms with category and example*

All the above algorithms are now standardized for off-the-shelf use in different programming languages. R language has been the most popular among statisticians and researchers to code their algorithms, and Python is catching up fast.

To be a prudent data scientist, one has to learn the underlying concepts from mathematics as well to understand what these algorithms are actually doing. Many data science professionals ignore the solid foundation in mathematics and statistics and remain reliant on the accuracy of packages developed by others to run an algorithm. Though open-source implementations are very stable and work in most cases, however lack of knowledge of algorithms working leads to sub-optimal or wrong decision son part of models.

The *Mathematics and statistics* section of the book will have a primer to most important topics that are encountered very frequently by data science professionals in designing and implementing algorithms for business problems. The most basic concepts include:

- Linear algebra (Matrix Computations)
- Optimization
- Multivariate calculus
- Probability
- Hypothesis testing
- Distributions

And many more sub-topics. A well-read data scientist never simply trusts a package and blindly usesthe output of libraries. It is important to critically analyze the algorithms and hasa sharp eye for any limitations or deviations from the theory of the algorithms.

Tools and technology

After acquiring the knowledge of the domain and having methods to quantify behaviors, data science requires means to implement those solutions at scale. Technology is the enabler to implement the mathematical and scientific knowledge for end-use. Internet, cloud, and programming are the key to creating solutions for end-user. Tools are the means to accomplish a task; the knowledge to implement is the same implemented in different ways by different tools.

For example, you may want to have an addition done in C or Python; the way you do with these two programming languages may be different by the result will always be the same. Hence, it’s important for data science professionals to be updated with the latest tools so that they always have the most efficient and economical way to produce results.

Figure 1.7 shows the KDnuggets analytics/data science 2018 poll results for the tools the community is using to develop data solutions. Python stands out as one of the most favored tools in the community:

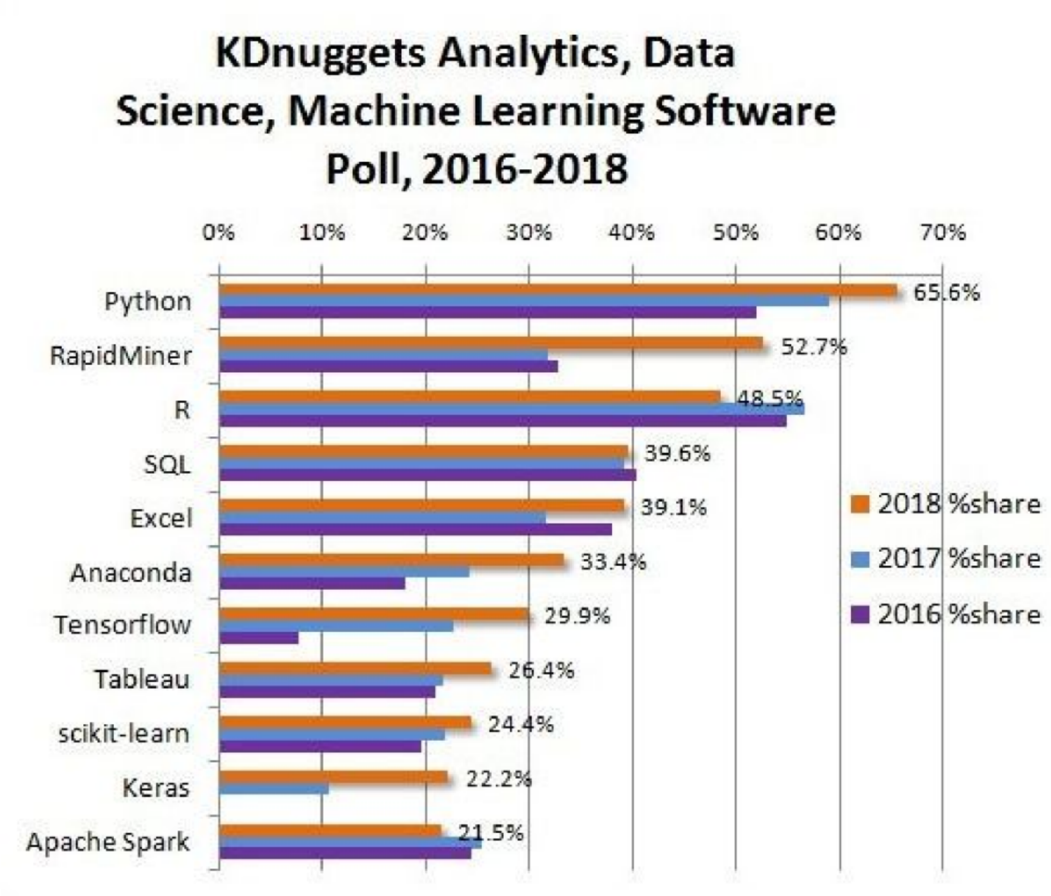


***Figure 1.7:** KDnuggets Analytics/Data Science 2018 poll results (Credits: https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html)*

The open-source nature of tools has fuelled the development of programming languages and packaged resources. Proprietary tools like Tableau and Excel are also part of the list as the enterprise uses them extensively.

Cloud computing is another set of technology which has changed the way data science is adopted, implemented, and maintained by big corporates. The whole cloud technology has made computing cheaper, on-demand, and very powerful. Now, even a small start-up can afford multi-million infrastructures by paying per hour as per need. This has changed the whole ecosystem of data-driven products. Below you see some companies which are just built on data-centric platforms and are now the biggest success stories in the corporate world:

Figure 1.8: Data Success Stories

All the above companies and solutions are hosted by public cloud vendors like Amazon Web services, Google Cloud Platform, and Azure. These platforms provide massive storage capacity and GPU computations for complex AI/ML models. Not only this, the cloud has given rise to three new types of technology service models:

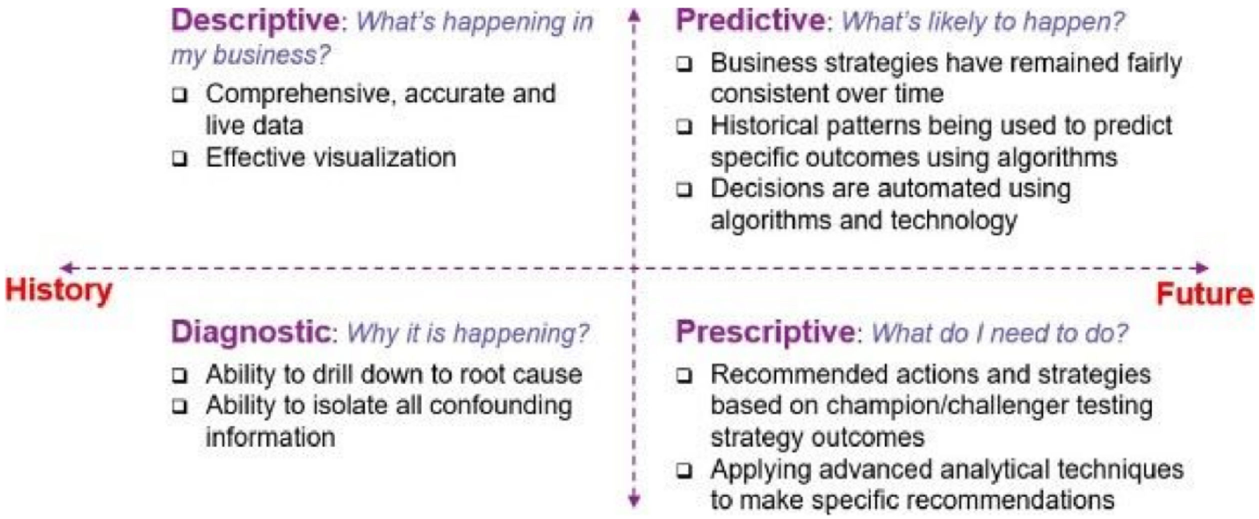
- **Software as a Service (SaaS)**
- **Platform as a Service (PaaS)**
- **Infrastructure as a Service (IaaS)**

In SaaS service, a fully hosted business application is provided on subscription bases, not only that reduced cost of the company, due to consolidation nature allows SaaS provider profits. The Silicon Valley start-up ecosystem is a true reflection of how the SaaS model allowed some of the biggest companies of our time to grow from start-ups to multi-billion-dollar companies.

Data Engineering is the new role that has emerged around extensive use of cloud technologies to build data pipelines having AI/ML results to be delivered to clients and businesses. The role of a data engineer is then to make sure that the required data by algorithms and end-user is delivered on time and with integrity. The *Data engineering* section will talk in detail about *data engineering*/pipelines, and the *Cloud computing* section of the book will talk in detail the technicalities and its benefits for data science growth.

Data science analysis types

Data Science analysis types can be divided into 4 types and can be seen as how the new tools are impacting the data analysis in more advanced methodologies and tools. *Figure 1.9* shows the types of analysis and how they differ from each other. All these types are now grown up as a separate area within the organization and also looking for having their own well-defined job roles:



Descriptive: *What's happening in my business?*

- ❑ Comprehensive, accurate and live data
- ❑ Effective visualization

Predictive: *What's likely to happen?*

- ❑ Business strategies have remained fairly consistent over time
- ❑ Historical patterns being used to predict specific outcomes using algorithms
- ❑ Decisions are automated using algorithms and technology

Diagnostic: *Why it is happening?*

- ❑ Ability to drill down to root cause
- ❑ Ability to isolate all confounding information

Prescriptive: *What do I need to do?*

- ❑ Recommended actions and strategies based on champion/challenger testing strategy outcomes
- ❑ Applying advanced analytical techniques to make specific recommendations

Figure 1.9: *Types of data science analysis*

The descriptive analysis has been historically well studied and applied in statistical methods. It reflects the empirical measurement of what is happening in the system/business under observation. The next step to that is diagnosing the happening by investigating the data by different slices and time window views to essentially answer why it happened.

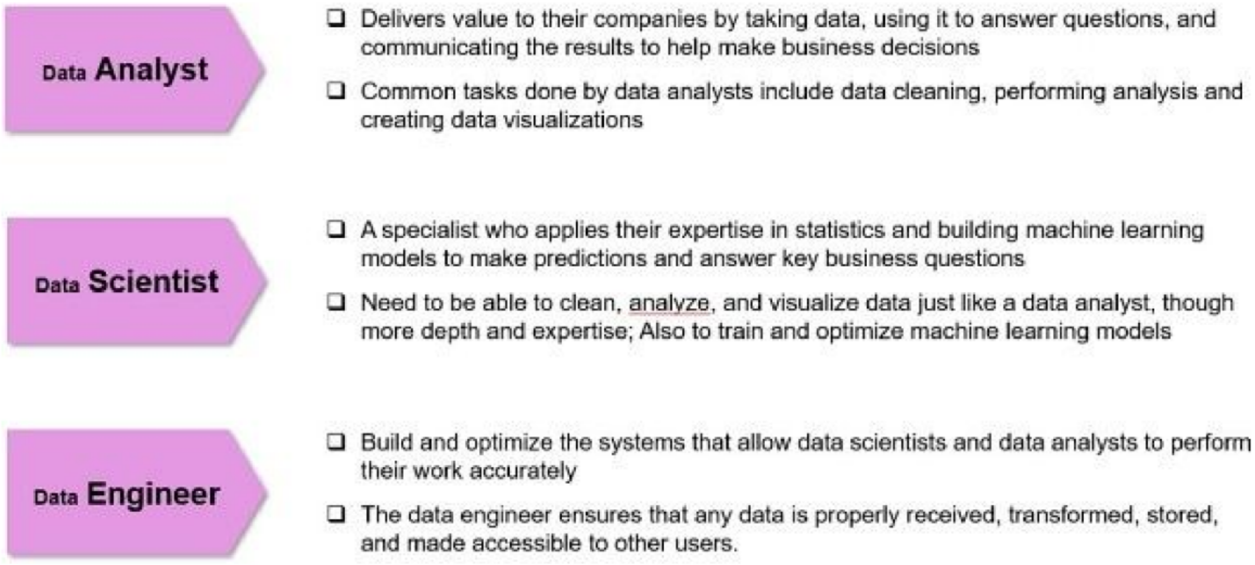
For example, my business grew by 45% in last quarter is a descriptive measure, if we also add why it grew it become diagnostics as well, that is, our new product segment grew by 90%, balancing the slowdown in an old product by 45%. Now you can observe we know what happened and why it happened. Both these analyses are very business-centric and allow the business to make quick and accurate decisions.

The modern, powerful computing environment enabled predictive and prescriptive data analysis as well. The predictive analysis helps find relationships among data points and help predict one data point if another is available, while prescriptive go one step further and recommend which variable or data point to control to get the desired results.

For example, the sales will grow by 25% next quarters as the festival season is going to start next month. This is a predictive outcome of analysis, as historical data would suggest the sales go up in the festive season. Further, the prescriptive analysis outcome will be like if we give a 10% discount, the sales will go up by 35% in the upcoming festive season. Here, we can control the discount rate, and the analysis prescribes it to be 10% for optimal gains.

Data science job roles

In the previous session, you observed the types of analysis and skillsets that have emerged to define job roles that scope the work and bring synergies and in-depth analysis capabilities in professionals. If you refer to *Figure 1.4*, which describes the three key subject areas comprising data science, you would be able to relate the job roles with the data science function. The three areas of a domain, statistical techniques, and technology give rise to three roles data analyst, data scientist, and data engineer, respectively. The roles are described in *Figure 1.10*, with their high-level responsibilities in the role:



Data **Analyst**

- ❑ Delivers value to their companies by taking data, using it to answer questions, and communicating the results to help make business decisions
- ❑ Common tasks done by data analysts include data cleaning, performing analysis and creating data visualizations

Data **Scientist**

- ❑ A specialist who applies their expertise in statistics and building machine learning models to make predictions and answer key business questions
- ❑ Need to be able to clean, analyze, and visualize data just like a data analyst, though more depth and expertise; Also to train and optimize machine learning models

Data **Engineer**

- ❑ Build and optimize the systems that allow data scientists and data analysts to perform their work accurately
- ❑ The data engineer ensures that any data is properly received, transformed, stored, and made accessible to other users.

Figure 1.10: *Data science job roles*

The job roles are essential to understand as it's not possible to keep juggling in roles in the early stage of career. All the roles are interconnected to each other; however, as a new entrant, you need to have adequate skills in one area as a major, and you can pick another area as minor. Having the knowledge, all job roles within the data science function brings high synergy in work.

ML model development process

Data science, as perceived by most of the online courses and recent public discourse, has been around how to develop accurate models for prediction. The key area within data science is focused on the development of models, that is, artificial intelligence, machine learning, and deep learning. The areas are a complete subset of the prior to them one respectively:

- **Artificial intelligence:** Programs that have the ability to learn and reason like humans.
- **Machine learning:** Programs that have the ability to learn without being explicitly programmed.
- **Deep learning:** Programs that can learn from a vast amount of complex data using artificial neural networks.

For fresher's who are starting their journey into data science, the first they need to learn is the process of developing a machine learning model and interpret them. *Figure 1.11* shows the process which has been studied and provided an indicative guideline for developing models:

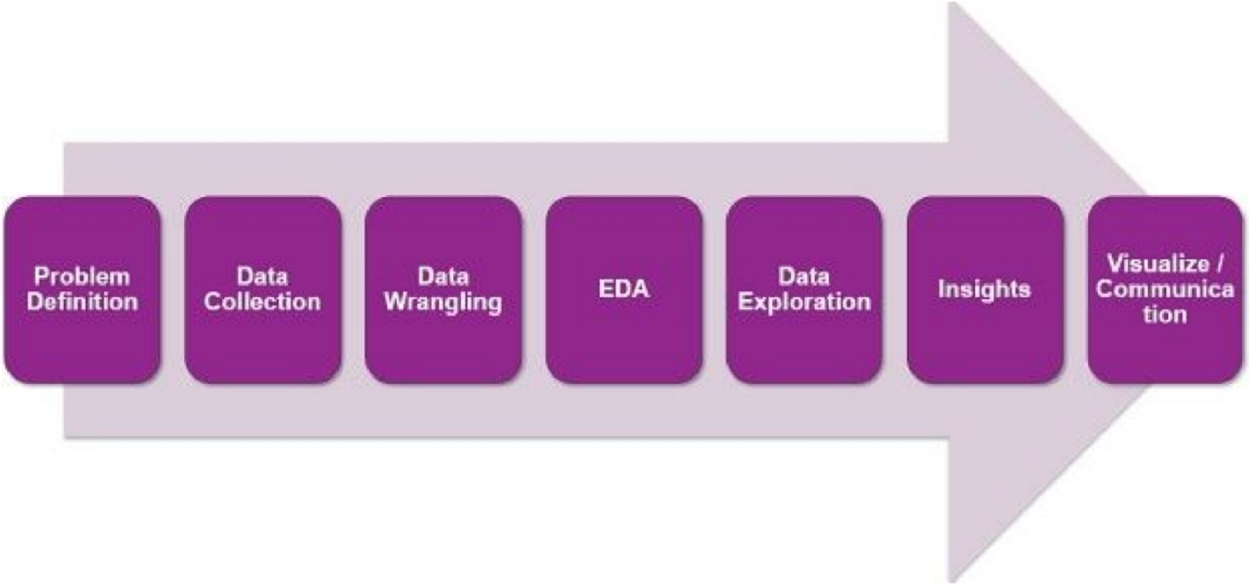


Figure 1.11: *Data science modeling process*

The steps arerising in detail in the next section of chapters on machine learning. Here we provide a basicdefinition of each step:

1. **Problem definition:** Any data analysis starts with setting up an objective that we want to achieve out of model development exercise. These objectives can be in terms of hypothesis or target results in business metrics after using the models.
2. **Data collection:** Now, the data which can help solve the problem statement is gathered through different channels and sources. Best efforts are made to have accurate and timely data for the analysis.
3. **Data wrangling:** Data wrangling has many parts to it, including cleaning the data from missing and erroneous values, removing outliers, transforming the data, feature engineering, and other steps to make data ready for empirical analysis.
4. **Exploratory Data Analysis (EDA):** EDA step is a pre-model-analysis of descriptive and diagnostic nature where we use visualizations, distributions, frequency tables, and other techniques to understand the data relationships and

make the choice of right algorithm for desired analysis.

5. **Machine Learning Algorithms:** Now, we train, test, and validate algorithms with appropriate dependent and independent variables, with appropriate techniques from the set of supervised, unsupervised, and reinforcement learning algorithms.
6. **Prediction and insights:** The algorithms will quantify all the relationships and allow us to make predictions and derive insights from the model outputs. The model results need to be transformed back to business language and presented on the same scale as of original data.
7. **Visualization and communication:** The results need to communicate back to business executives or end-user for making decisions. The results need to be communicated in simple terms while not undermining the assumptions of probability and modeling techniques.

The process is what has been observed in most of the cases, but it does not limit the data science professionals to explore new ways to bring value out of their data. Exploration and being curious is the key to develop good models that derive business.

Data visualizations

Data visualization is not an integral part of organizations and end-user applications. In today's applications, you would always start your application experience with a Dashboard and then go further into the application features. Data visualization tools, like Power BI and Tableau, have grown to the extent that they provide a self-service platform to builds visualizations to communicate data analysis for both business intelligence purpose and exploratory data analysis from machine learning. *Figure 1.12* shows the cognitive side of visualizations and how they are beneficial to users to interpret complex data and make business decisions:

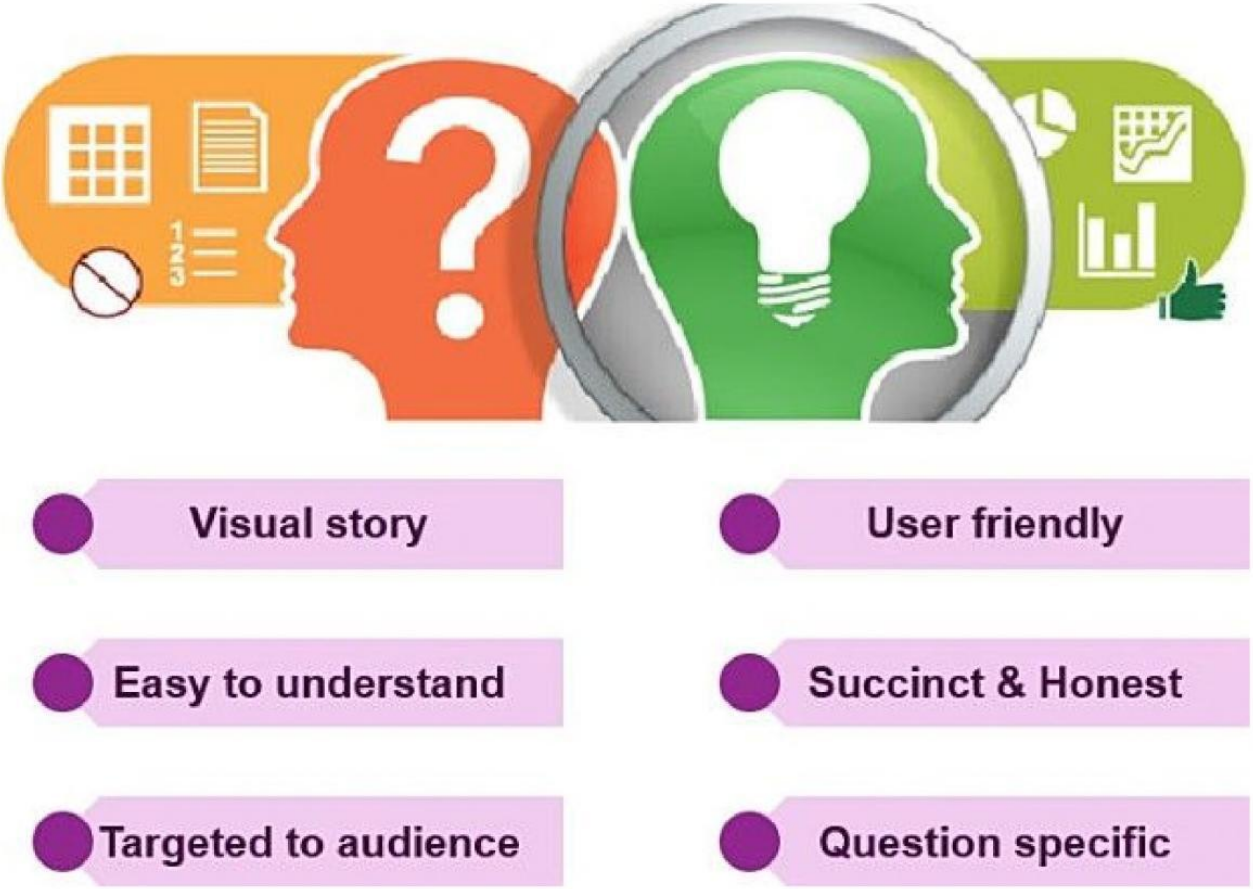
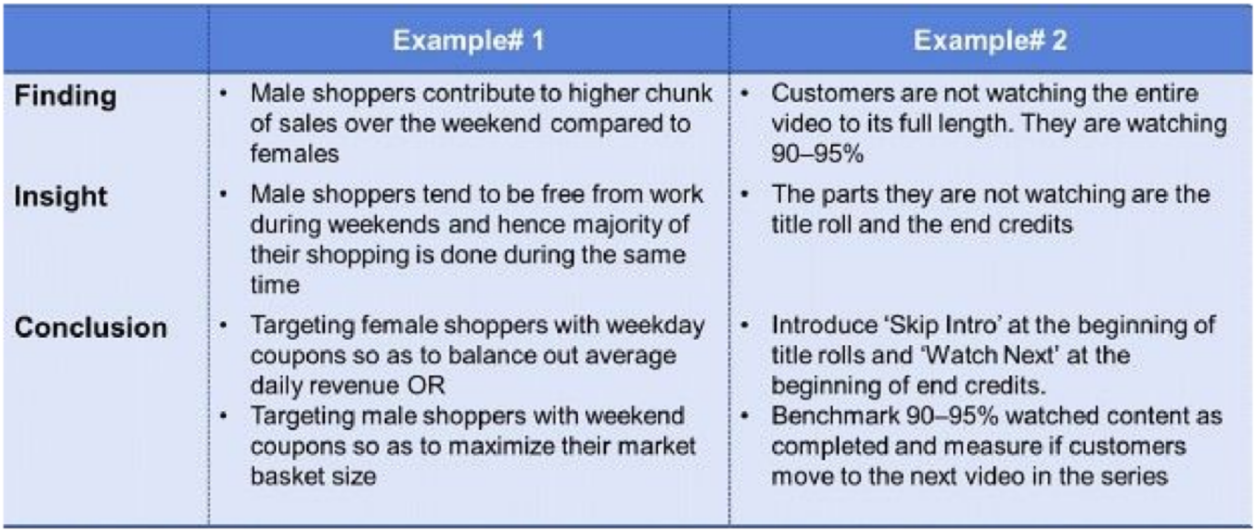


Figure 1.12: *Data visualizations*

In the section, Business intelligence, we talked in detail about how to build visualizations and how the tools play an important role in speed-up your analysis and communication of results to stakeholders.

Result communication

Result communication is both art and science. With years of experience, you develop the right balance between the two to become anefficient communicator of data science. You can assume this role be like that of a translator of language, here the data is to be communicated to business and vice versa. The choice of rights words, right visualizations, and interpretations are important for communicating the right understanding of the data. *Figure 1.13* shows two examples of effective communication and a summary of analysis which could have taken weeks and multiple tools to come to the conclusions:



	Example# 1	Example# 2
Finding	• Male shoppers contribute to higher chunk of sales over the weekend compared to females	• Customers are not watching the entire video to its full length. They are watching 90–95%
Insight	• Male shoppers tend to be free from work during weekends and hence majority of their shopping is done during the same time	• The parts they are not watching are the title roll and the end credits
Conclusion	• Targeting female shoppers with weekday coupons so as to balance out average daily revenue OR • Targeting male shoppers with weekend coupons so as to maximize their market basket size	• Introduce 'Skip Intro' at the beginning of title rolls and 'Watch Next' at the beginning of end credits. • Benchmark 90–95% watched content as completed and measure if customers move to the next video in the series

Figure 1.13: *Result communication*

It is important for data science professionals to be aware of how business communications happen and how they can be more effective to communicate their findings, and also convert their pain points into dataproblems. This role is usually done by experienced professionals, and junior resources need to learn from them.

Responsible and ethical AI

As data science and AI become more day to day affairs and start impacting how we operate our daily life and business, it becomes important to get aware of the negative aspects that AI will bring to the fore. The data science professionals need to be aware ofthe consequences of their work and the impact it can have on society. It is important to have the highest ethical practices for data science professionals. *Figure 1.14*, from ngu.eu, shows the key areas and questions pertaining to the responsible and ethical use of AI:

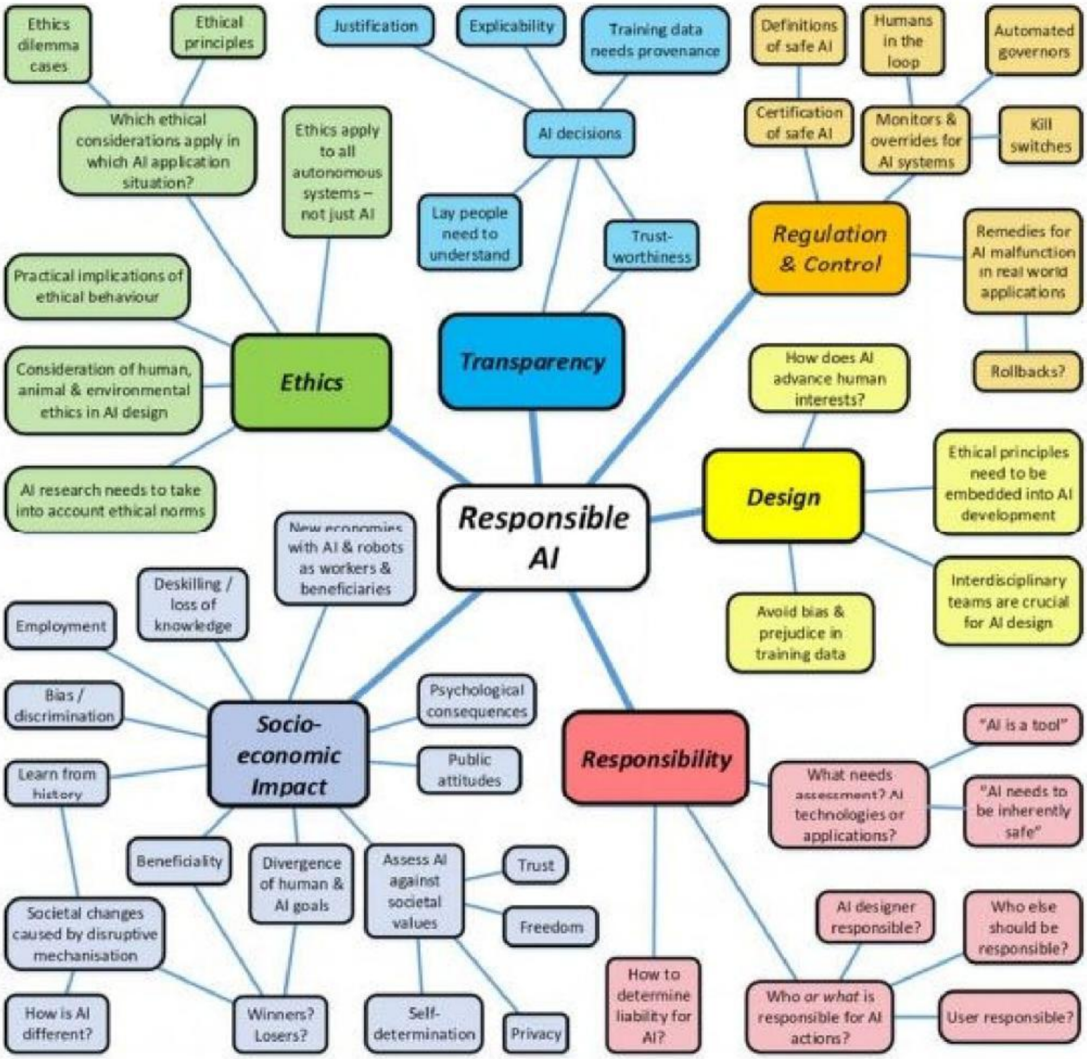

Ethics dilemma cases
Ethical principles
Justification
Explicability
Training data needs provenance
Definitions of safe AI
Humans in the loop
Automated governors
Which ethical considerations apply in which AI application situation?
Ethics apply to all autonomous systems – not just AI
AI decisions
Certification of safe AI
Monitors & overrides for AI systems
Kill switches
Lay people need to understand
Trust-worthiness
Regulation & Control
Remedies for AI malfunction in real world applications
Practical implications of ethical behaviour
Ethics
Transparency
How does AI advance human interests?
Rollbacks?
Consideration of human, animal & environmental ethics in AI design
Design
Ethical principles need to be embedded into AI development
AI research needs to take into account ethical norms
Responsible AI
New economies with AI & robots as workers & beneficiaries
Deskilling / loss of knowledge
Employment
Interdisciplinary teams are crucial for AI design
Avoid bias & prejudice in training data
Bias / discrimination
Psychological consequences
Socio-economic Impact
Public attitudes
Responsibility
"AI is a tool"
Learn from history
What needs assessment? AI technologies or applications?
"AI needs to be inherently safe"
Beneficiality
Divergence of human & AI goals
Assess AI against societal values
Trust
Freedom
Societal changes caused by disruptive mechanisation
AI designer responsible?
Who else should be responsible?
How to determine liability for AI?
How is AI different?
Winners? Losers?
Self-determination
Privacy
Who or what is responsible for AI actions?
User responsible?

Figure 1.14: *Responsible and Ethical AI (Credits: ngi.eu)*

The leading name in technology and consulting, along with governments, are leading the efforts to make AI use fair and ethical. It will have regulatory effects, like the **General Data Protection Regulation (GDPR),** Data Privacy Framework, and other self-governing structures of organizations. As a professional in the domain, one needs to pay adequate attention to their work and being fair as the models developed by them may deny a loan to a needy just because you ignored a bias in historical data for class, color, or religion. It is our duty to NOT bring our biases to systems and make it a system for the future.

Career in data science

Towards the end of the overview section, Probyto, with its years of experience, would like to give some advice to budding data scientists for a successful career. The key to success in any field is asking questions and remain a learner forever. Data science is no different. Having a strong foundation in mathematics is very helpful in a long career in data science; as technology has masked a lot of hard facts behind easy to use menu driven approached, you still need to have adequate knowledge of Mathematics. *Figure 1.15* shows the various skills required for getting into data science career:

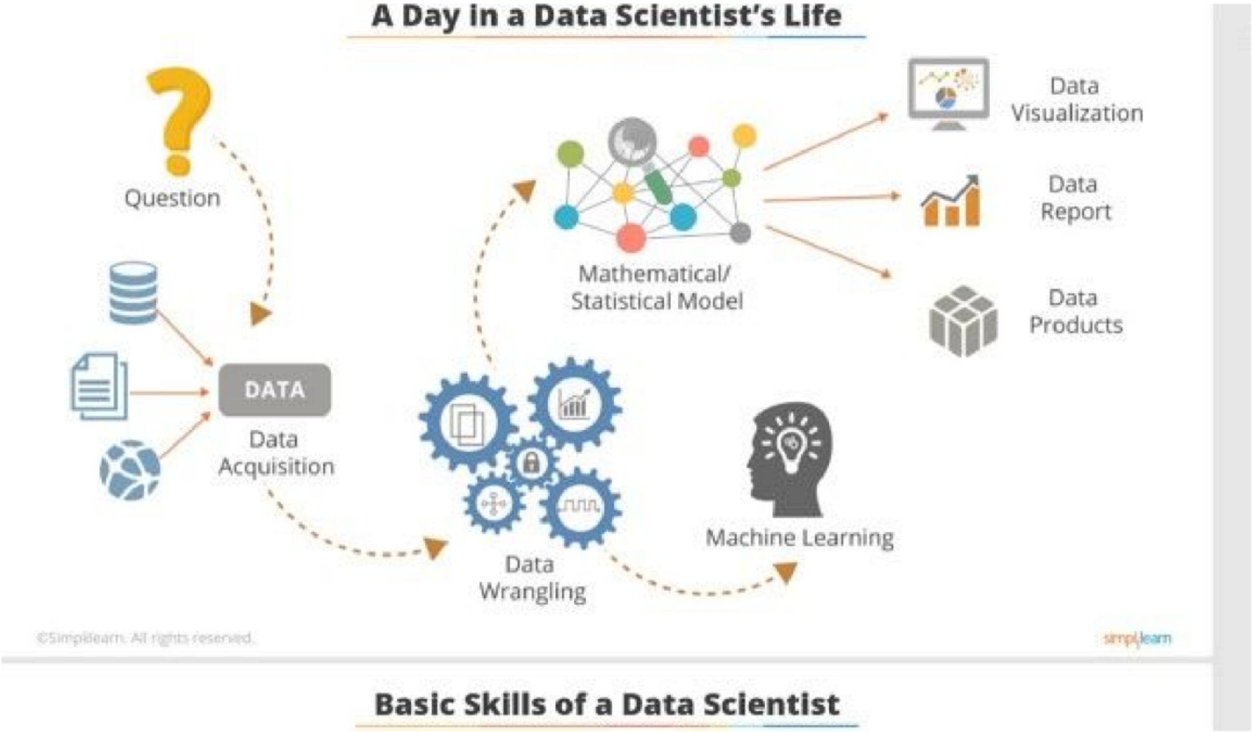

A Day in a Data Scientist's Life
Question
DATA
Data Acquisition
Data Wrangling
Mathematical/ Statistical Model
Machine Learning
Data Visualization
Data Report
Data Products
©Simplilearn. All rights reserved.
simplilearn
Basic Skills of a Data Scientist

Figure 1.15: *Key Skills Required for Data Science Career*

Some of the skills you are required to possess for a job in data science are, not exhaustive;

1. Asking the right questions
2. Understanding data structures
3. Data exploration and interpretation
4. Applying models
5. Visualization of data

In *Chapter 19: Data Science Learning Resources,* we also provide you a small list of some popular resources that help you be updated with data science all the latest happenings. A career in data science is very dynamic by ever-growing technology, research in algorithms, and new problems to solve. While its exciting career but also very challenging and require you to be up to date always.

Conclusion

The chapter sets the primer for the data science beginners by introducing multiple ideas, origin, process, and definitions in the data science domain. We talked about how the landscape of data analysis has been changing and evolved from past to today, what triggered those changes, and the development of data science. Then we defined data science as a Venn diagram having the right balance of domain knowledge, mathematics, and technology, to enable bringing value from data. The different types of analysis are then discussed to allow the reader to differentiate between multiple types of approaches to data as per need. We have then talked about different role types is data science and how they differ from each other. The ML model building process is also introduced for readers to have a preview of the next sections on machine learning. Data visualization is introduced as key to allowingthe user to interpret data faster by a visual view rather than raw data. Result communication is an important part of becoming atranslator between data and business problems. Responsible and Ethical AI is the need of the time with so much invasion of AI in our daily lives, and technology companies and government actively working to contain that. The chapter ended with some suggestions on the key skills and requirements to start a career in data science.

By reading this chapter, the reader haspreviewedthebook and its content on an overview of data science. He or she will understand what a beginner needs to acquire before stepping into the data science journey.

In the next chapter, we will discuss the mathematics concepts needed in modern data science. This will bestrong foundation to apply data science in real-world problems.

CHAPTER 2
Mathematics Essentials

Mathematics is the foundation of any modern-day discipline of science. When we look into the modern data science principles, there must be some deep mathematical behind it. Understanding and learning the fundamentals of mathematics is very important for any data scientist or junior analyst. Because they have to apply those techniques in solving problems. Mathematics foundation works like the heart of the problem-solving using data science. Other items will come in the category of just by using an API or using the new algorithms.

In order to make ameaningful prediction and recommendation to the users, we have to use algorithms very carefully. To develop a better algorithm, we must have astrong understanding ofthe mathematical principles behind the algorithms. When we have the mathematical foundation very solid, then it creates more confidence forpeers to work on it.

In this chapter, we are going to learn the very important mathematical concepts, which will be the strong foundation to apply data science in real-world problems.

Structure

- Introduction to linear algebra
- Scalars, vectors, matrices, and tensors
- Eigenvalues and eigenvectors
- Eigen decomposition and singular value decomposition
- The determinant
- Principal component analysis
- Introduction to the multivariate calculus
- Differential and integral calculus
- Partial derivatives
- The Gradient, Hessian, Jacobian, Laplacian, and Lagrangian Distribution

- The Gradient Descent algorithm
- Conclusion

Objectives

After studying this chapter, you should be able to:

- Understand the fundamentals of linear algebra and data representations.
- Apply the basic properties of matrix and vectors in data science applications.
- Derive the mathematical concepts behind data science problems.
- Build the mathematical model of an algorithm.

Introduction to linear algebra

Linear algebra is the branch of mathematics to study lines and planes, vector spaces and mappings that are required for linear transforms. Linear algebra can be called basic mathematics to understand the data, which has the intention of finding related values using linear combinations. In other words, it is the application of the problem-solving system of linear equations to find unknown or new findings from data.

Vectors and matrices are the basic notions of data. When data represented using a column-based notion, we call that as vectors, if the array is used to represent the data that has been considered as matrix format.

Example:

1. $V =$ $V = \begin{bmatrix} 3 \\ -1 \\ 9 \end{bmatrix}$ where, V is the vector which has data in column-based representation (that is, single column with three data).
2. $M = \begin{bmatrix} 2 & -1 & 4 \\ 5 & -7 & 3 \\ -9 & 8 & -6 \end{bmatrix}$ where *M* is the matrix data in array-based representation (that is, rows and columns are used to represent nine data).

Even though vectors and matrices are the languages to represent data, we have to understand some of the basic data representation with geometry space. The following section will briefly introduce those concepts.

Scalar, vectors, matrices, and tensors

Let us start exploring the fundamentals of data representation frommathematics perspective.

Scalar

A scalar is a one-dimensional representation of data. It contains only the magnitude in the form of numerical value. For example, consider the scenario to travel from one place to another place; the following are the data represented in the scalar values:

- Distance between the starting place and destination place (that is, 100 km)
- Speed of the vehicle (that is, 20 km/h)
- Traveling fair (that is, Rs 20)
- Weight of the person (that is, 60 kgs)

This example gives a brief idea about the scalar representation of data. When we represent data in higher-dimensional geometry, space will move with further additional information.

Vectors

A vector is a notion of one or more values of scalars. Already we have introduced the vector we can now explore the characteristics of vectors and how it is used to represent the data in geometry space. From the scenario of travel, we can take the example of vector data:

- Acceleration is given to the direction.
- The velocity of the vehicle traveling towards a destination.
- Displacement, i.e., traveling from one point to another point.

Characteristics of the vector:

- The vector contains magnitude and direction.
- Vector changes if either the magnitude or direction changes or both change.

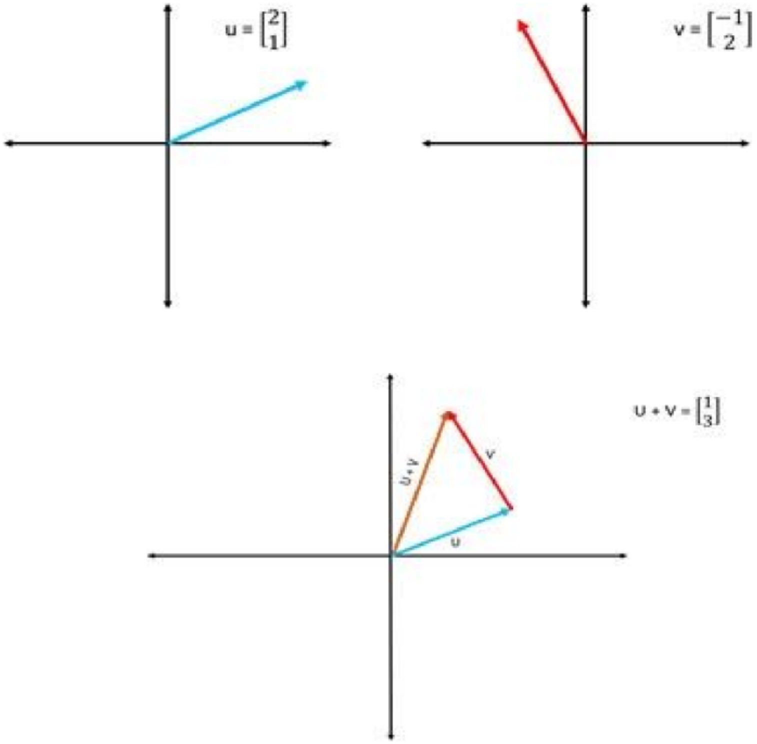

$u = \begin{bmatrix}2\\1\end{bmatrix}$
$v = \begin{bmatrix}-1\\2\end{bmatrix}$
$u + v = \begin{bmatrix}1\\3\end{bmatrix}$
u+v
v
u

Figure 2.1: *Vector Addition*

Vector plays the main role in the linear algebra because both with vectors, it is in two operations. Vector addition and scaling the vector by multiplying some scalar value. To represent those operations in geometrically refer to *Figure 2.1* and *Figure 2.2*:

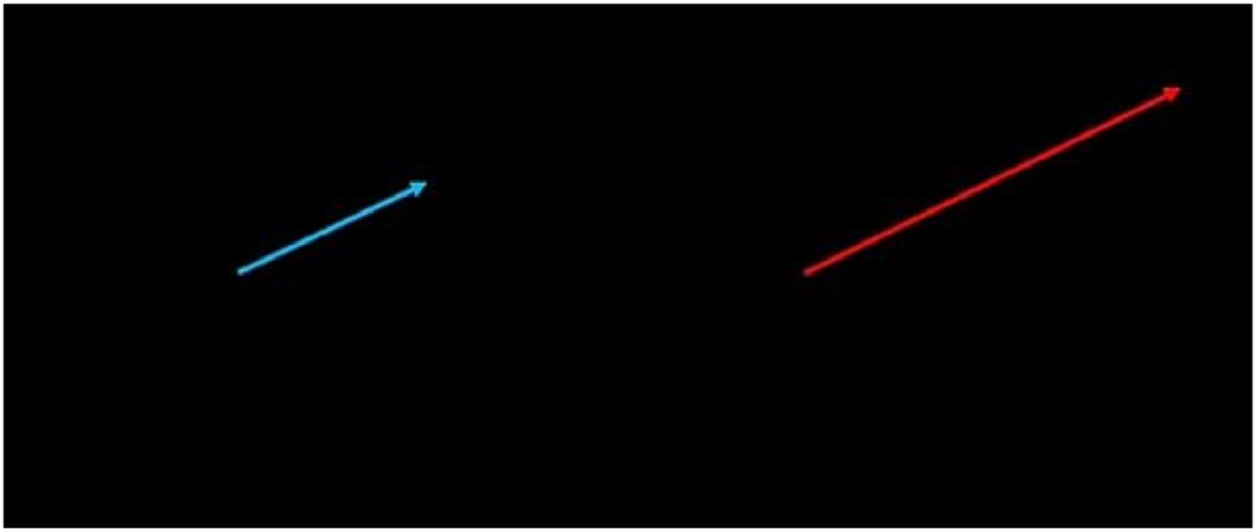


Figure 2.2: *Scalar multiplication to Vector*

Understanding the capabilities of vector operations provides the confidence in estimating the scalability of values and make use of them in data representation is an easier process.

Matrices

Data are represented in a rectangular arrangement is known as matrices. Where data is arranged in the form of rows and columns. For example, to represent the data with 2 rows and 3 columns, we will use the below notation:

$$A = \begin{bmatrix} 6 & -1 & 2 \\ 5 & -7 & 3 \end{bmatrix}$$

The representation of rows and columns is also known as the dimension of the matrix. In other words, we can say that is the size of the matrix. So, the dimension of the above matrix is *2 (rows) × 3 (colums).*

As matrices can hold the higher dimensional space in geometry, we can interpret that it is possible to hold data which comes in the form of multi-dimensional format.

A data in a matrix entry is simply a matrix element. Each element of data in a matrix is recognized by naming the row and column in which it appears.

For example, let us consider the above matrix A. The A23 is the entry of scalar data value 3 is identified by calling the second row and third column. By using this representation, we can easily access the data from the matrix. It also provides the flexibility to pick the data directly without going into the sequential way.

In general, the matrix representation is in the following format:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots\cdots\cdots\cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots\cdots\cdots\cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots\cdots\cdots\cdots & a_{3n} \\ \cdots & \cdots & \cdots & \cdots\cdots\cdots\cdots & \cdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots\cdots\cdots\cdots & a_{mn} \end{bmatrix}$$

To access the element, we have to use the representation of a_{mn}, where m represents the row value, and n represents the column value. Matrices mainly used to solve systems of equations. But first, we must learn how to represent linear systems with matrices.

A system of equations can be solved easily by representing its data in an augmented matrix. An augmented matrix in linear algebra is a matrix generated by adding columns of two different matrices, typically for the purposes of carrying out the same simple row operations on each of those matrices.

For example, consider the linear equation system:

$$x+2y+3z=5$$

$$2x+3y+z=7$$

$$x+y+z=6$$

We can represent the coefficient matrices of the system as:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 1 & 1 & 1 \end{bmatrix} \text{ and } B = \begin{bmatrix} 5 \\ 7 \\ 6 \end{bmatrix}$$

Then the augmented matrix is,

$$A \vee B = \begin{bmatrix} 1 & 2 & 3 & 5 \\ 2 & 3 & 1 & 7 \\ 1 & 1 & 1 & 6 \end{bmatrix}$$

The augment matrix will help to represent the data thatcan be solved using a combination of operation performed on matrices rows to get unknown values of coefficients in the linear equation representation.

Even though we can apply all the basic arithmetic operations on matrices, addition and multiplication play vital roles. Letus start the discussion on those operations.

The Matrix addition is quite easy and is performed input-wise. For example:

$$A = \begin{bmatrix} 2 & -1 & 4 \\ 5 & -7 & 3 \\ -9 & 8 & -6 \end{bmatrix} \text{ and } B = \begin{bmatrix} 0 & 7 & 9 \\ 3 & -6 & 3 \\ 4 & 2 & -6 \end{bmatrix} \text{ then,}$$

$$A + B = \begin{bmatrix} 2 & 6 & 13 \\ 8 & -13 & 6 \\ 5 & 10 & -12 \end{bmatrix}$$

Here, the elements from matrix *A* and *B* are mapped together, and simple addition performed to get the newvalue (that is, 2 + 0 = 2). The generated new matrix follows the same rules for all the entries.

In matrix multiplication operation, we must follow the different scenarios to do the multiplication of scalar values. When we perform the operation, we must check the dimensionality of both matrices.

If we want to perform multiplication between two matrices, we have to check that the number of columns in a matrix is equal to the number of rows in the next matrix. If it is not equal, then we can’t perform the multiplication.

For example, consider the matrices:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 0 & 2 \end{bmatrix} \text{ and } B = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} \text{ then } A.B = \begin{bmatrix} 8 & 7 \\ 20 & 15 \\ 4 & 6 \end{bmatrix}$$

Here, in matrix A, a number of columns are 2, and in matrix B number of rows is 2. Hence, we can compute the solution of AB like this.

$$A.B = \begin{bmatrix} 1*4+2*2 & 1*1+2*3 \\ 3*4+4*2 & 3*1+4*3 \\ 0*4+2*2 & 0*1+2*3 \end{bmatrix} = \begin{bmatrix} 8 & 7 \\ 20 & 15 \\ 4 & 6 \end{bmatrix}$$

If the dimension of matrices goes more than 2, then it is difficult to apply the multiplication easily for that we have to use another format to represent the data. So, handling higher dimensional data is done by tensors.

Tensors

The tensor is mathematically an N-dimensional vector, which means that an N-dimensional data can be represented by tensor. It is a multidimensional array.

Mathematically a tensor is an N-dimensional vector, which means a tensor can be used to represent N-dimensional data. A tensor is a simplification of vectors and matrices, and it can easily understand as a multidimensional array:

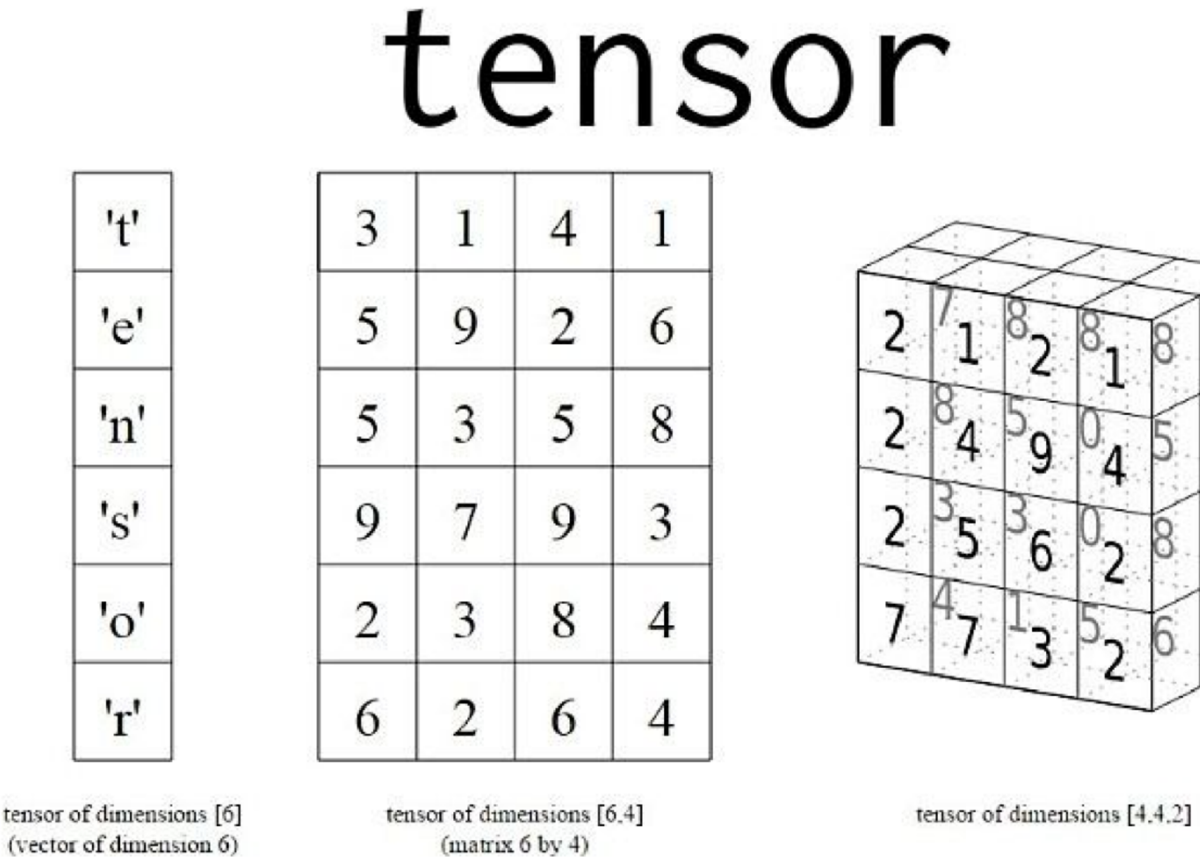


Figure 2.3: *Simplified tensor with minimum dimensions*

We can say a vector is a 1D or 1st order tensor, and a matrix is a two-dimensional or second-order tensor. Tensors have the ability to perform all types of operations with scalars, matrices, or vectors by reformulation.

The determinant

Determinants are mathematical entities that are widely used in the analysis and find the solution of systems that have the property of linear equations. The determinant value of a matrix is a special value, and it is very important that it can be calculated from a square matrix. We will discuss the calculation of determinant of a matrix with an example.

Let $A=\begin{bmatrix} 6 & 1 \\ 4 & 3 \end{bmatrix}$ be the square matrix with size *2 × 2* then, determinant of matrix A is represented using the notion $|A|$. The value of the determinant is calculated using the below method:

|A|=6*3-4*1=14

Now, we can generalize this calculation. Letus assume, $M=\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ then determinant value, $|M|=ad-bc$. Finding the determinant value for 3 × 3 matrix is a similar way only. For example:

$$G=\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 5 \\ 4 & 2 & 1 \end{bmatrix}$$

|G|=1(3*1-2*5)-2(2*1-4*5)+3(2*2-4*3)

1(3-10)-2(2-20)+3(4-12)

1(-7)-2(-18)+3(-8)

-7+36-24

5

Therefore, the determinant is $|G|=5$. In general, we can represent the calculation using the symbol:

$$\text{If, } M=\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \text{ then}$$

$$|M| = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}|$$

$$a(e * i - f * h) - b\,(d * i - f * g) + c\,(d * h - e * g)$$

$$aei + bfg + cdh - afh - bdi - ceg$$

From the calculation of determined of the matrix, we can tell whether that matrix can be inverted or not. It also consider the scalar property of a matrix.

Eigenvalues and Eigenvectors

Eigenvectors and eigenvalues are used to decrease noise in data. This impacts efficiency in the computation of the tasks. So, to improve efficiency in computationally more complex tasks, estimation of Eigenvalues and Eigenvectors plays a vital role. We can represent multidimensional data in a matrix. One eigenvalue and eigenvector are useful to capture significant information that is stored in a large matrix. Performing computations on a large matrix is a very time-consuming process. One of the key methodologies to enhance the efficiency in computationally demanding tasks is to reduce the dimensions of data after ensuring most of the important information is maintained.

Before mathematically explaining the eigenvalues, we first see the details about eigenvectors. When we apply the scalar multiplication on a vector almost, it will change the direction. In opposite to this, any vector which is not changing its direction even it is multiplied by a scalar value is known as eigenvector. Multiply an eigenvector by *A*, and the vector *Ax* is a number λ times to the original value. The basic representation is *Ax*=λ*x,* where λ is an eigenvalue of matrix *A.* This value only determines whether the *x* is expanded or shrunk or unchanged when it is multiplied by *A*. In graphical way it is shown like *Figure 2.4*:

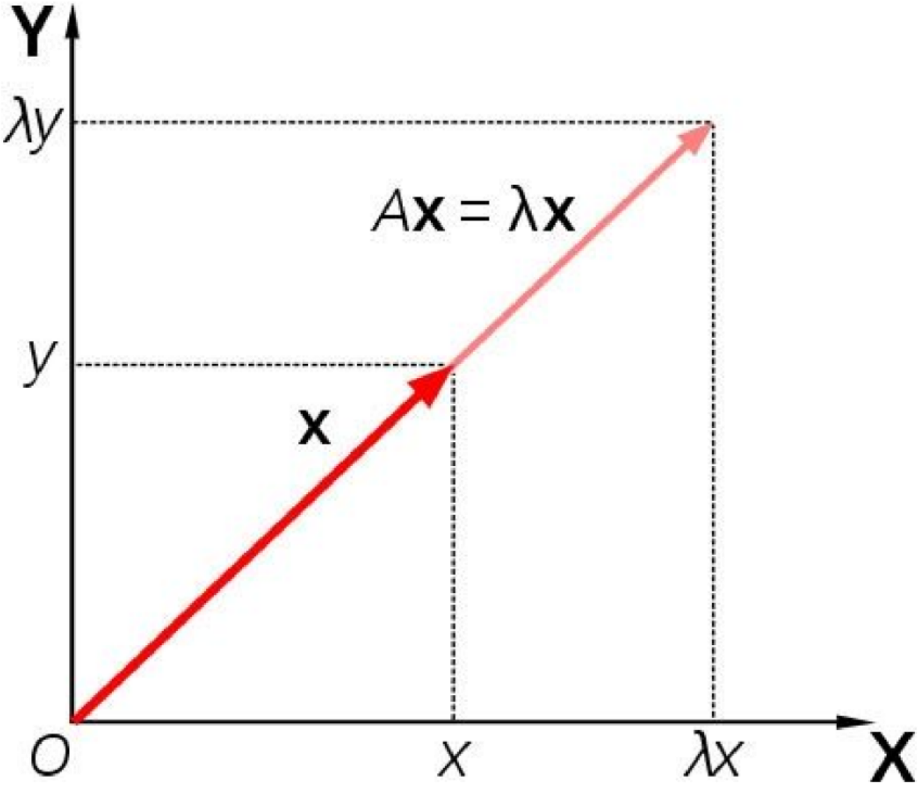


Figure 2.4: *Eigenvalue λ and eigenvector A representation*

Now consider the linear transformation of N-dimensional vectors defined by an n by n matrix A:

$$Ax = w,$$

$$\text{or} \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

Where for each row, $w_i = A_{i1} x_1 + A_{i2} x_2 + \dots + Ax_n = \Sigma(j=1)A_{ij} x_j$

If it occurs that *x* and ware scalar multiples, that is if, *Ax=w=λx*. Then v is the value of an eigenvector of the linear transformation of matrix A, and the scale factor λ is the eigenvalue corresponding to that eigenvector. We can rewrite the eigenvalue equation like, *(A-λI) x = 0*. Where I is the n by an n identity matrix, and 0 is the zero vector. The eigenvalues of *A* are values of λ that satisfy the equation, *|A-λI|=0.*

From this equation, we can get eigenvalue λ of matrix A by taking the roots of the polynomial. For example, consider the matrix, $M= \begin{bmatrix} 2 & 7 \\ -1 & -6 \end{bmatrix}$.

First, multiply λ to an identity matrix and then subtract the two matrices:

$$|M - \lambda I| = \left|\begin{bmatrix} 2 & 7 \\ -1 & -6 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right|$$

$$|M - \lambda I| = \begin{vmatrix} 2-\lambda & 7 \\ -1 & -6-\lambda \end{vmatrix} = \lambda^2 + 4\lambda - 5$$

Apply, |M - λ I|=0. Then, *$\lambda^2 + 4\lambda - 5 = 0$* by solving this equation, we can get λ values. λ=-5 and λ=1. These values are the eigenvalues of the matrix M. To find the eigenvector substitute the values of λ in the equation *(M-λI) x=0*

Substitute $\lambda=-5$, $\begin{bmatrix} 7 & 7 \\ -1 & -1 \end{bmatrix} x = 0$. By solving this linear equation, we will get the vector $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ is one of the eigenvectors for the matrix M. Using, λ = 1, we will get the vector $\begin{bmatrix} -7 \\ 1 \end{bmatrix}$ is another linearly independent eigenvector for the matrix *M*.

Eigenvalue decomposition and Singular Value Decomposition (SVD)

In linear algebra, eigenvalue decomposition or sometimes spectral decomposition is the factorization of a matrix into a canonical form, whereby the matrix is represented in terms of its eigenvalues and eigenvectors. Only diagonalizable matrices can be factorized in this way.

Let A be a square $n \times n$ matrix with n linearly independent eigenvectors q_i (where i = 1, ..., n). Then A can be factorized as,

$$A = Q \Lambda Q^{(-1)}$$

Where Q is the square $n \times n$ matrix whose ith column is the eigenvector q_i of A, and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\Lambda_{ii} = \lambda_i$. Note that only the diagonalizable matrix cab be factorized in this way.

The n eigenvectors q_i are usually normalized, but they need not be. A non-normalized set of n eigenvectors, v_i can also be used as the columns of Q. That can be implied by observing that the magnitude of the eigenvectors in Q gets canceled during the decomposition by the existence of *Q-1*.

The decomposition can be derived from the fundamental property of eigenvectors.

$$Av = \lambda v$$

$$AQ = Q\Lambda$$

$$A = Q \Lambda Q^{(-1)}$$

For example, consider, $\begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix}$may be decomposed into a diagonal matrix through the multiplication of a non-singular matrix B= $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

Then, $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix}$

For some real diagonal matrix $\begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix}$ multiplying both sides of the equation on the left by B:

$$\begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & y \end{bmatrix}$$

The above equation can be decomposed into two simultaneous equations:

$$\begin{bmatrix}1 & 0\\ 1 & 3\end{bmatrix}\begin{bmatrix}a\\ c\end{bmatrix} = x\begin{bmatrix}a\\ c\end{bmatrix}$$

$$\begin{bmatrix}1 & 0\\ 1 & 3\end{bmatrix}\begin{bmatrix}b\\ d\end{bmatrix} = y\begin{bmatrix}b\\ d\end{bmatrix}$$

Letting, $\vec{a} = \begin{bmatrix}a\\ c\end{bmatrix}, \vec{b} = \begin{bmatrix}b\\ d\end{bmatrix}$ this gives us two vector equations:

$$A\vec{a} = x\vec{a}$$

$$A\vec{b} = y\vec{b}$$

And can be represented by a single vector equation involving two solutions as eigenvalues:

$$Au = \lambda u$$

Where λ represents the two eigenvalues x and y. From eigenvalue equation we can write:

$$(A - \lambda I)\, u = 0$$

By solving the equation:

$$(1-\lambda)(3-\lambda)=0$$

$$\begin{bmatrix}1 & 0\\ 1 & 3\end{bmatrix}\begin{bmatrix}a\\ c\end{bmatrix} = 1\begin{bmatrix}a\\ c\end{bmatrix}$$

$$\begin{bmatrix}1 & 0\\ 1 & 3\end{bmatrix}\begin{bmatrix}b\\ d\end{bmatrix} = 3\begin{bmatrix}b\\ d\end{bmatrix}$$

$$a = -2c \text{ and } b = 0$$

Thus, the matrix B required for the Eigen decomposition of A is

$$B = \begin{bmatrix}-2c & 0\\ c & d\end{bmatrix}$$

That is:

$$\begin{bmatrix} -2c & 0 \\ c & d \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} -2c & 0 \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$$

Singular value decomposition

In short, we can call SVD for Singular-Value Decomposition. The reduction of a matrix into its integral parts to make subsequent matrix is called as SVD. Simply, we can call this as a matrix decomposition method to split a matrix. It makes the matrix calculation as a simpler process.

For the case of simplicity, we will focus on the SVD for real-valued matrices and ignore the case for complex numbers.

$$A = U.Sigma.V^{T}$$

Where *A* is the real m x n matrix that we wish to decompose, U is an m x m matrix, Sigma (often represented by the uppercase Greek letter Sigma) is an m x n diagonal matrix, and *V^T* is the transpose of an *n x n* matrix where *T* is a superscript.

The diagonal values in the Sigma matrix are known as the singular values of the original matrix *A*. The columns of the *U* matrix are called the left-singular vectors of A, and the columns of *V* are called the right-singular vectors of *A*.

The SVD is calculated via iterative numerical methods. We will not go into the details of these methods. Every rectangular matrix has singular value decomposition, although the resulting matrices may contain complex numbers, and the limitations of floating-point arithmetic may cause some matrices to fail to decompose neatly.

The common problem is that the response matrix is singular or close to singular, so it has no well-defined inverse. Of the various algorithms that have been developed to deal with this problem, **singular value decomposition (SVD)** has emerged as the most popular. Any matrix can be represented with SVD as follows:

$$M = \sum_{k=1}^{n} \vec{u}_k \, w_k \vec{v}_k^{T}$$

Where *vk* is a set of orthonormal steering magnet vectors, *uk* is a corresponding set of orthonormal BPM vectors, and *wk* are the singular values of the matrix *M*.

Given the SVD of a matrix, the matrix inverse is:

$$M^{-1} = \sum_{k=1}^{n} \vec{v}_k \frac{1}{w_k} \vec{u}_k^T$$

It follows from the orthonormality of the two vector sets. It is immediately apparent from the singular value decomposition if the response matrix is singular one or more of singular values, *wk*, are zero. Physically, a zero wk implies that there is some combination of steering magnet changes, *vk*, which gives no measurable change in orbit. The orbit shift from this *vk* is zero at all the BPMs. Removing the terms with zero wk from the sum in the above equation produces a pseudo inverse for orbit correction, which generates no changes in the steering magnet strengths along the corresponding eigenvectors *vk*.

Principal component analysis

Principal component analysis (PCA) is a technique that is useful for the compression and classification of data. The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of variables that nonetheless retains most of the sample's information. By information, we mean the variation present in the sample, given by the correlations between the original variables. The new variables, called **principal components (PCs),** are uncorrelated and are ordered by the fraction of the total information each retains.

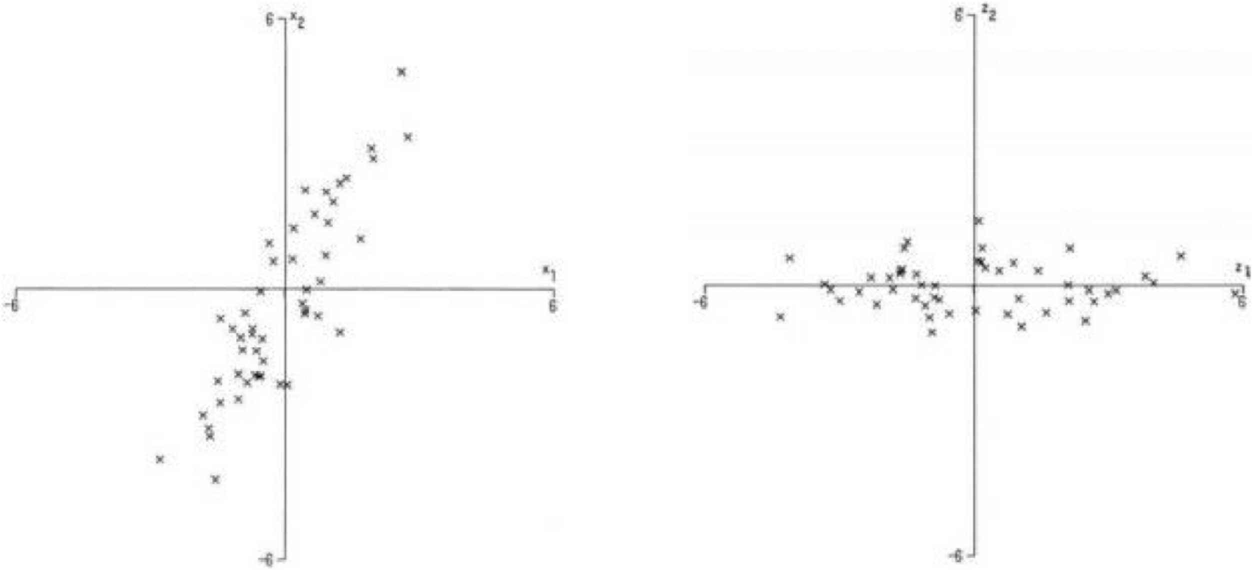


Figure 2.5: *Plot of n observations with two variables and a plot of the same wrt their principal axes.*

PCA is mathematically defined as an orthogonal linear transformation (meaning it rotates and scales) that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

Consider a sample space of p random variables where n observations are made. $x = (x_1,x_2,x_3,...x_p)$. When we apply a linear transformation to the observations along with the first principal component z_1 we will have the below equation:

$$z_1 = a_1^T x_1 = \sum_{j=1}^{p} a_{j1}\, x_j$$

Where the vector $a_1 = (a_{11}, a_{21}, a_{31}, ... a_{p1})$

z_1 will be our first principal component only when $var\,[z_1]$ is maximum. Likewise, when the k^{th} principal component has to be calculated, the above equation can be transformed as below:

$$z_k = a_k^T x_k = \sum_{j=1}^{p} a_{jk}\, x_j$$

Where the vector $a_k = (a_{1k}, a_{2k}, a_{3k},...a_{pk})$ must be chosen such that $var\,[z_k]$ is the maximum subject to $a_1^T\, a_1 = 1$:

$$var\,[z_k] = a_k^T\, \Sigma a_k$$

Where Σ is the covariance of the variables(x_1, x_2, x_3,...x_p). The problem seems to be a constraint optimization problem. Applying LaGrange multipliers where λ is the LaGrange multiplier.

$$a_k^T \Sigma a_k - \lambda_k)$$

Optimization involves differentiating and equating to 0 when we have to find a maximum or minimum of a function. Applying differentiation:

$$\Sigma a_k - \lambda a_k = 0 \textit{ can also be written as } \Sigma a_k = \lambda a_k$$

Therefore referring to previous topics of eigenvectors we can say for Σ as a covariance matrix a_k will be the kth eigenvectors and λ_k will be the kth eigenvalue.

Multivariate calculus

Multivariate Calculus (also known as multivariable calculus) is the extension of calculus in one variable to calculus with functions of several variables: the differentiation and integration of functions involving multiple variables, rather than just one:

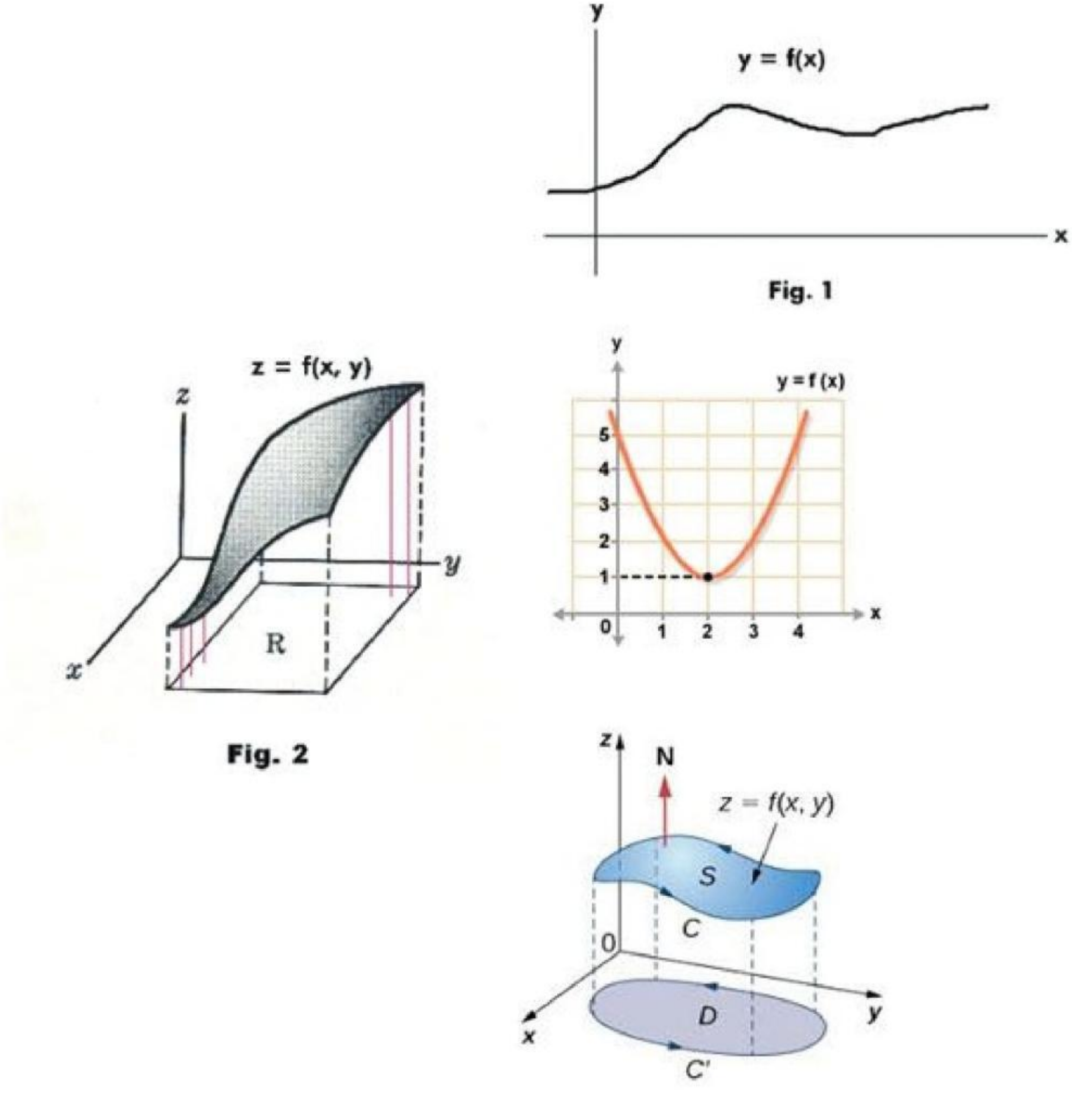


Figure 2.6: *Multivariate Calculus (Credit: https://www.toppr.com/bytes/multivariable-calculus/)*

Calculus is a set of tools for analyzing the relationship between functions and their inputs. In Multivariate Calculus, we can take a function with multiple inputs and determine the influence of each of them separately.

Why is Multivariate Calculus important in data science?

In data science, we try to find the inputs which enable a function to best match the data. The slope or descent describes the rate of change off the output with respect to an input. Determining the influence of each input on the output is also one of the critical tasks. All this requires a solid understanding of Multivariate Calculus.

What is a function?

A function is a connection between input and output. In that, the notation of f(x) is a function of the variable x. This relationship can be seen as the growth over therun of how the change in one variable affects the relationship in another variable.

Differential Calculus

The gradient of a variable in a function is the rise over run against the other variables. *Figure 2.7* shows the gradient representation using run and rise in a plot. In the normal *x, y* coordinated plot, the rise over run is computed using two points plotted on the graph:

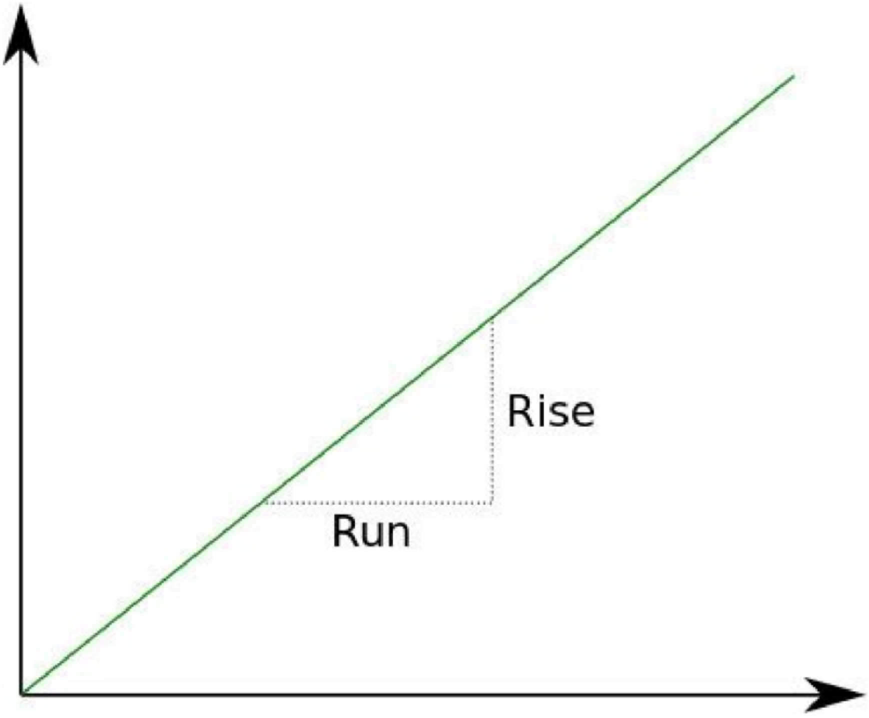


Figure 2.7: *Gradient is equal to Rise over Run*

The representation for the derivative or the gradient can also be shown as such:

$$\frac{df}{dx} = f'(x) = \lim_{dx \to 0} \frac{f(x + dx) - f(x)}{dx}$$

If we apply the gradient for a non-linear function, it changes based on the value of the variable change. Considering only the two data points to calculate gradient will produce inaccurate gradient value.

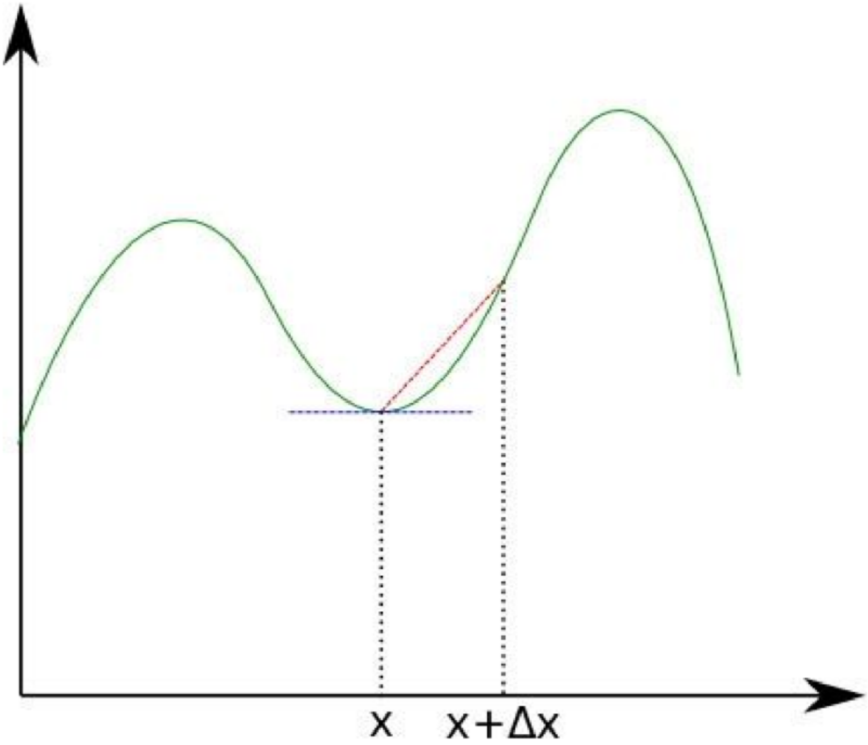


Figure 2.8: *As dx tends to 0, the gradient becomes more accurate*

Let us take *Figure 2.8* as an example. Considering two points *x* and *x + dx*, we can get an evaluation of the gradient at x. However, if we were to change the value of the second point, where *dx* tends to *0*, the gradient we calculate will be more accurate. Hence, we should aim to get a *dx* where it is infinitely small yet not 0. Using the notation given above, let's try an example where *f(x)=3x+2.*

$$f'(x) = \lim_{dx \to 0} \frac{3(x + dx) + 2 - (3x + 2)}{dx}$$

$$\lim_{dx \to 0} \frac{3x + 3dx + 2 - (3x + 2)}{dx}$$

$$\lim_{dx \to 0} \frac{3dx}{dx} = 3$$

The above example shows how we can use the gradient calculations between the two points.

Sum rule

The derivatives of the sum of two functions are the sum of the derivative of each function on its own. This rule extends indefinitely.

$$\frac{d}{dx}\big(f(x)+g(x)\big)=\frac{df(x)}{dx}+\frac{dg(x)}{dx}$$

Power rule

The derivative of a function with respect to a variable with an exponent is given by taking the value of the exponent and multiplying the function with it. Then deduct one from the value of the exponent.

$$Forf(x)=ax^{b}$$

$$f^{\wedge\prime}\ (x)=abx^{(b\text{-}1)}$$

The above equations represent how we can use the power rule in representing the function.

Special cases

The *f(x)* = *1/x* function is a special one which includes a discontinuity:

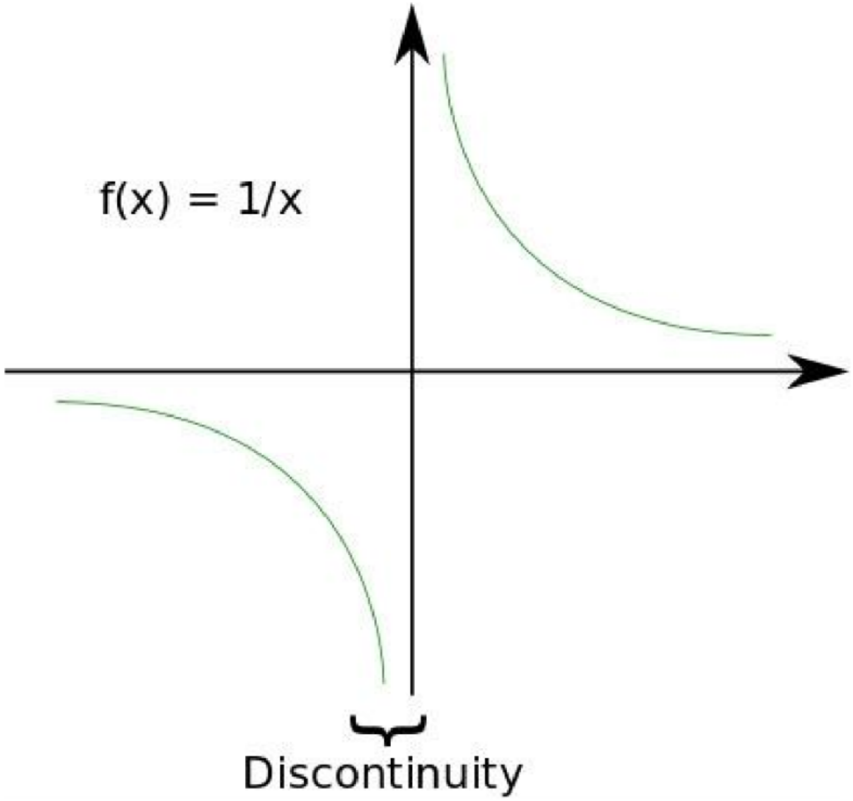


Figure 2.9: *Discontinuity in the f(x) = 1/x graph*

Figure 2.10 show that this function also has a negative gradient for all values of *x*, excluding *x = 0* which is undefined:

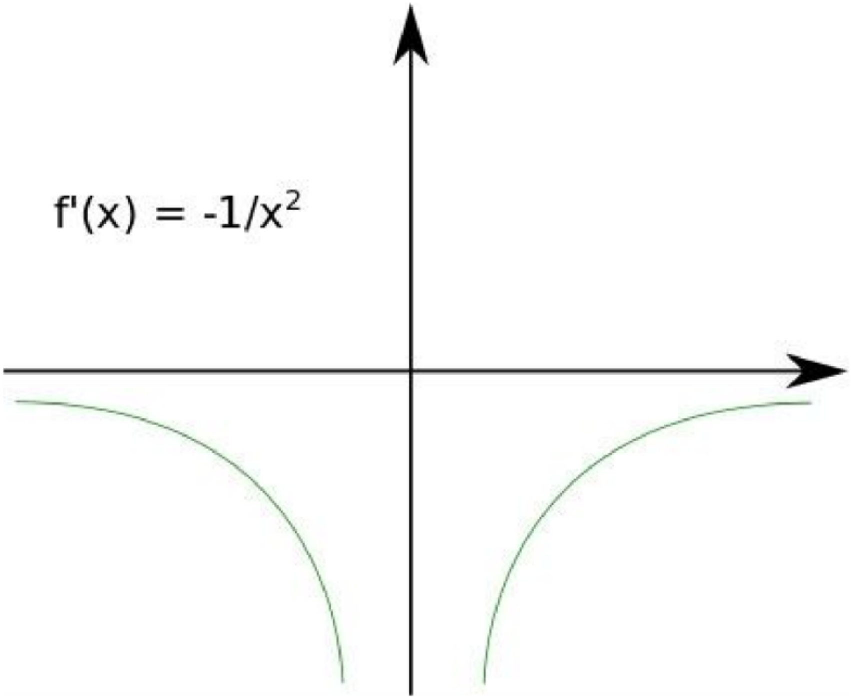


Figure 2.10: *Discontinuity in the $f'(x) = -1/x^2$ graph*

Another special case:

$$f(x)=e^x$$

This has the special property in which $f(x) = f'(x)$, the function is equal to its derivative. Where $f(x) = f'(x)$ holds, its values are either positive, negative, or *0*. Taking the negative of the above equation gives a negative example while taking $f(x) = 0$ gives the *0* example. *Figure 2.11* shows the plot of the above function:

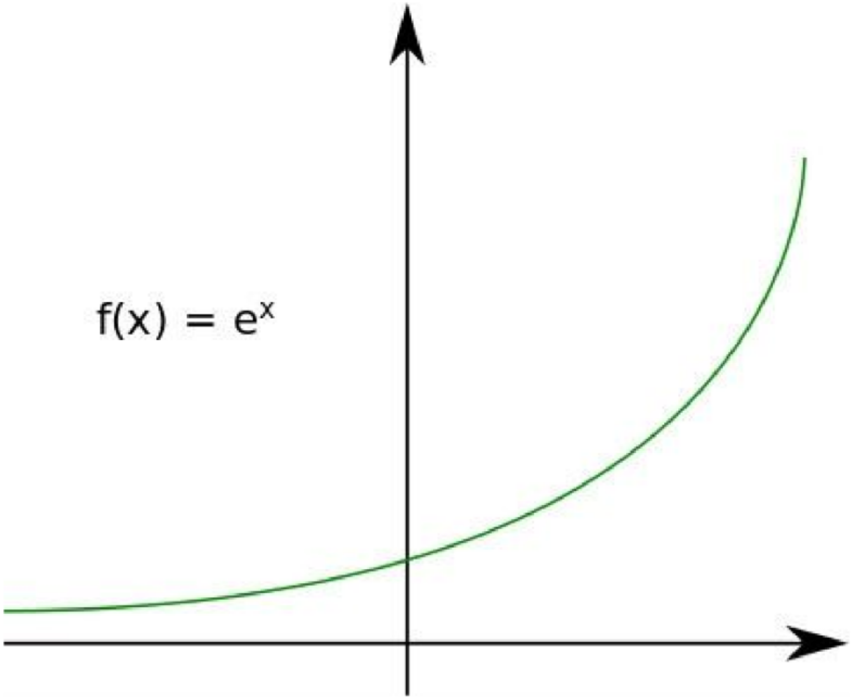


Figure 2.11: *f(x)=e^x*

These representations are the few special cases in differential calculus but not always occurs in data representation.

Trigonometric functions

The two trigonometric functions that we are focusing on will be the sine and cosine functions. *Figure 2.12* represents the notation sign(x):

Sine:f(x)=sin(x)

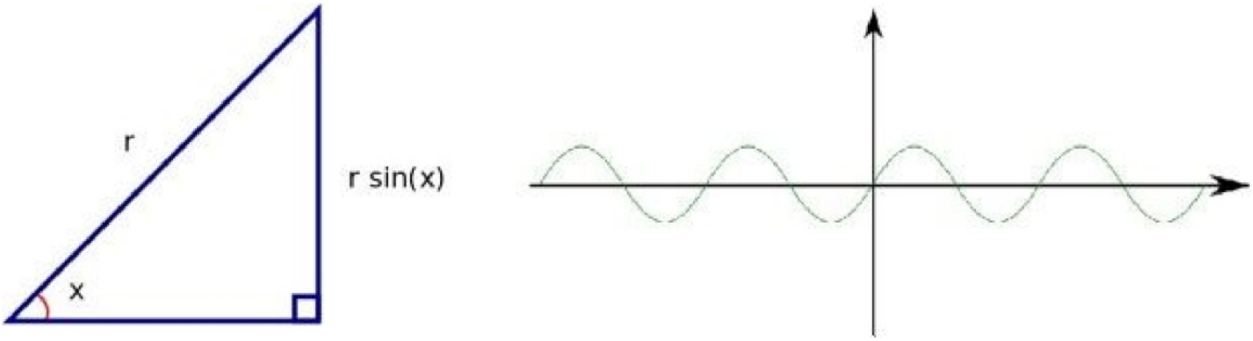


Figure 2.12: *Sine triangle and sine graph*

The derivative of the sine function is the cosine function which represented in *Figure 2.13*:

Cosine:f(x)=cosin(x)

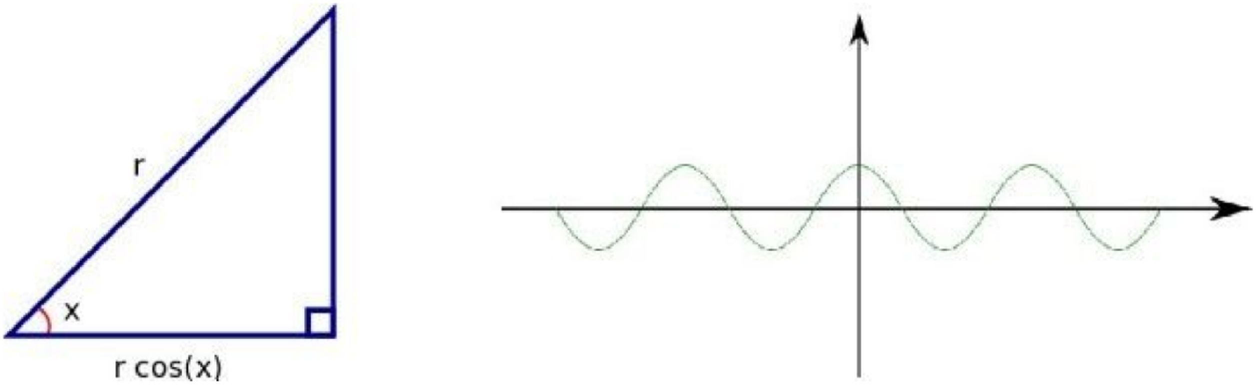


Figure 2.13: *Cosine triangle and cosine graph*

The derivative of the cosine function is the negative sine function. The two functions form a derivative loop that returns to the start every 4th derivation. *Figure 2.14* shows the sine cosine derivative loop:

$$\begin{tikzcd}
\sin(x) \arrow[r] & \cos(x) \arrow[d] \\
-\cos(x) \arrow[u] & -\sin(x) \arrow[l]
\end{tikzcd}$$

Figure 2.14: *Sine Cosine derivative loop*

Understanding the functions of sin and cosine will provide support when we plot the data in a graph then interpret the information from the data.

Product rule

Product rule defines identifying a derivative when two functions are in the product:

$$A(x) = f(x)g(x)$$

$$A'(x) = f'(x)g'(x)$$

$$\lim_{dx\to 0} \frac{dA(x)}{dx} = \lim_{dx\to 0} \frac{f(x)\big(g(x+dx)-g(x)\big)+g(x)\big(f(x+dx)-f(x)\big)}{dx}$$

$$\lim_{dx\to 0} \frac{f(x)\big(g(x+dx)-g(x)\big)+g(x)\big(f(x+dx)-f(x)\big)}{dx}$$

$$\lim_{dx\to 0} \frac{f(x)\big(g(x+dx)-g(x)\big)}{dx}+\frac{g(x)\big(f(x+dx)-f(x)\big)}{dx}$$

$$\lim_{dx\to 0} f(x)g'(x)+g(x)f'(x)$$

Chain rule

In general, a function has an *inside function* and an *outside function*. We can say the chain rule identifies derivatives as the *inside function* and the *outside function*. We can differentiate the outside function leaving the inside function alone and multiply all of this by the derivative of the inside function. This is represented in the below equation:

$$If h = h(p) \wedge p = p(m),$$

$$Then \frac{dh}{dm} = \frac{\frac{dh}{dp} * dp}{dm}$$

The application of chain rule is very helpful when we try to find derivatives in function has the looping property.

Quotient rule

This rule can be derived from the product rule:

$$ConsiderafunctionA(x) = \frac{f(x)}{g(x)}$$

$$ThenA'(x) = \frac{g(x) * f'(x) - f(x)g'(x)}{g(x)^2}$$

This rule can be used when we want to estimate the function fractional value.

Multiple variables

In a function, there are independent variables, dependent variables, and possibly constants as well. Consider *Figure 2.15* shows the example of plot car's speed time to time basis. Here, time is the variable of independent in nature. Where the speed is a dependent variable. The relationship between these two variables is such that at any given speed, there can be time periods matching that particular speed, while at each time period, there is only one speed:

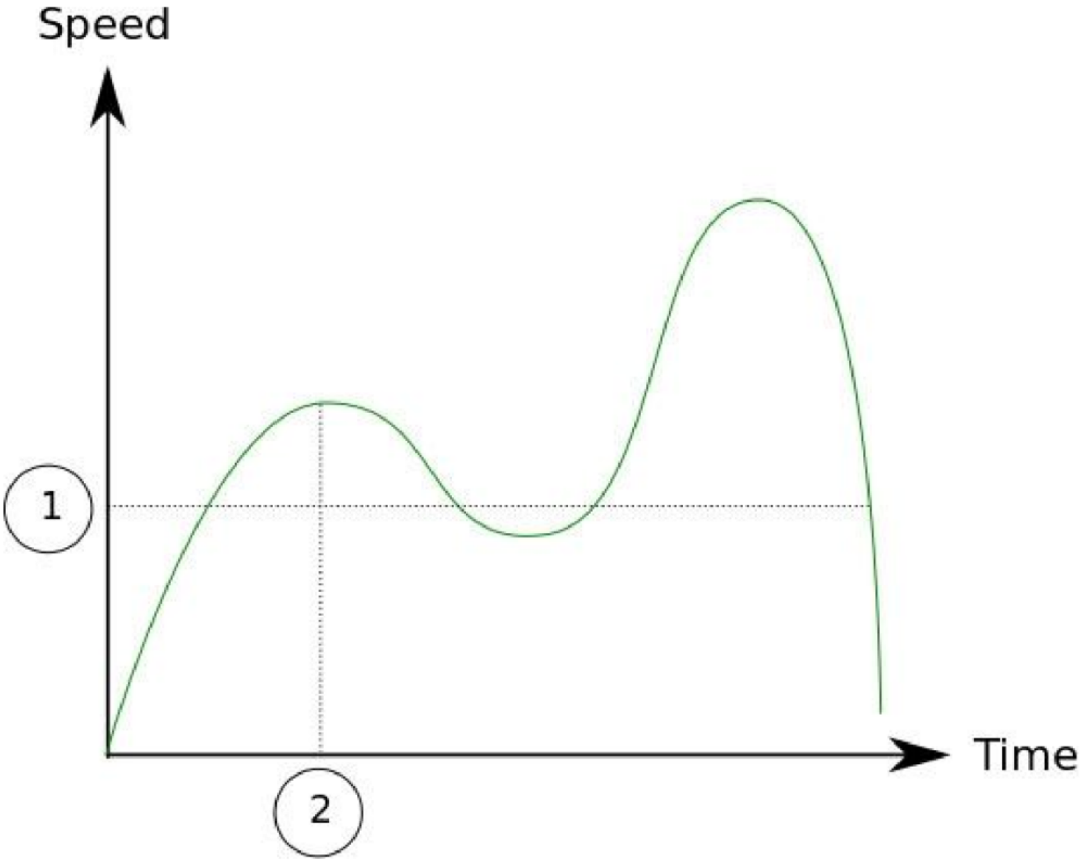


Figure 2.15: *Speed vs. Time*

As such, the speed of the car is dependent on the time period mentioned. The parameter's valuesa function that may vary in their variable types based on the context in which the function is used.

Partial differentiation

Partial differentiation is differentiating with respect to each variable in turn. For each partial differentiation, regard the other variables as constants in the differentiation context.

Total derivative

The total derivative is the derivative with respect to a specific variable where the function is dependent on the variable not only directly but also on other variables that are dependent on that specific variable.

Integral calculus

Integration is a way of adding slices to find the whole. Integration can be used to find areas, volumes, central points, and many useful things. But it is easiest to start with finding the *area under the curve of a functionlike this:*

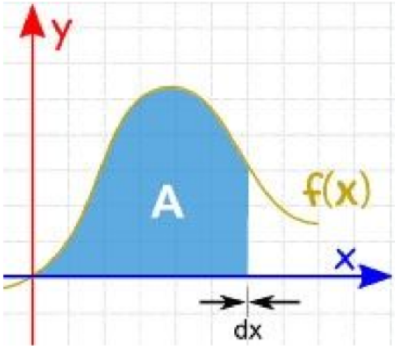
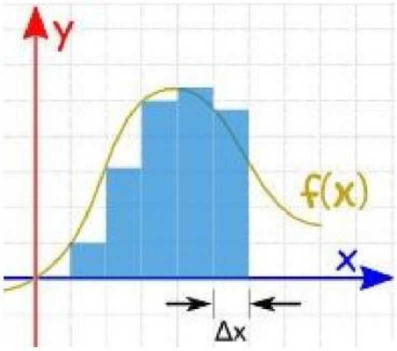
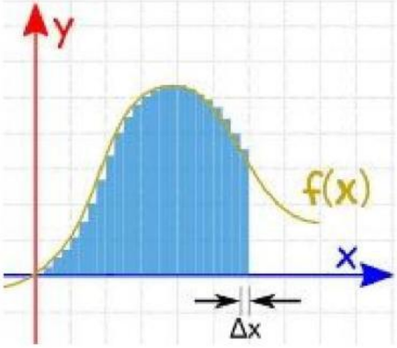
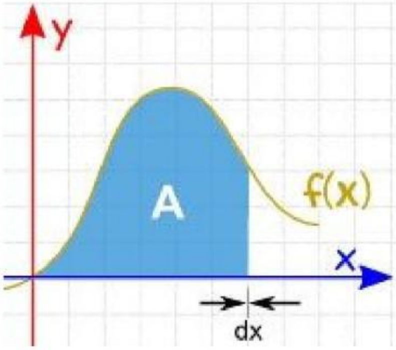


Figure 2.16: *What is the area undery = f(x)?*

Slices

We could calculate the function at a few points and add up slices of width Δx like this (but the answer won't be very	

accurate) as shown in *Figure 2.17*:	 ***Figure 2.17:** shows adding slices of width Δx:*
We can make Δ x a lot smaller and *add up many small slices* (the answer is getting better) as shown in *Figure 2.18*:	 ***Figure 2.18:** shows adding smaller slices of width Δx.*
And as the slices approach zero in width, the answer approaches the *true answer*, as shown in *Figure 2.19*. We now write *dx* to mean the *Δx* slices are approaching zero in width.	 ***Figure 2.19:** shows slices of zero width.*

These are the different representations of slices in multiple variable estimations.

Definite vs.indefinite integrals

We have been doing *Indefinite Integrals* so far, which is *If f(x)* is an anti-derivative of *f(x)* then the most general anti-derivative of f(x) is called an indefinite integral and denoted:

$$\int f(x)dx=F(x)+c$$

Where *c* is any constant.

Figure 2.20 shows the representation of finite integral and infinite integral values. A Definite Integral has actual values to calculate between (they are put at the bottom and top of the *S*):

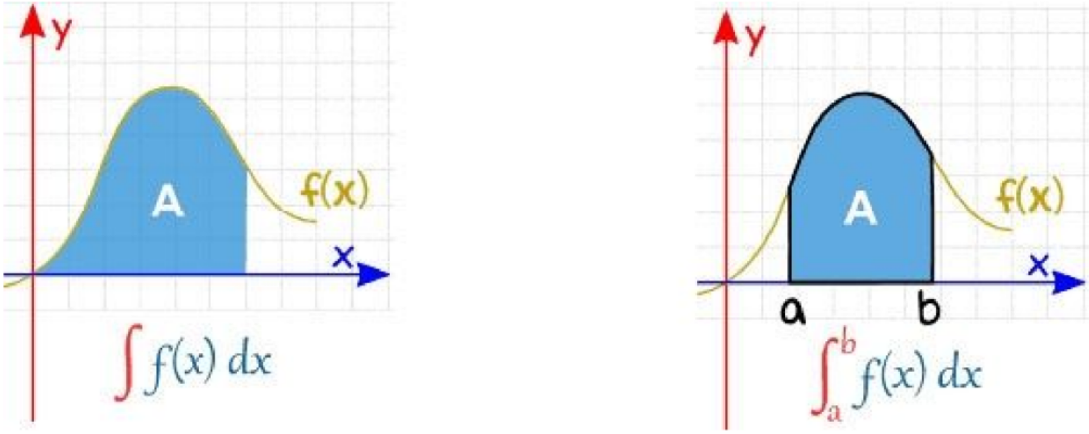


Figure 2.20: *Indefinite Integral & finite Integral*

The definite integral of *f(x)* from a to b is:

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\Delta x$$

Given a function *f(x)* that is continuous on theinterval *[a,b]*,we divide the interval into n subintervals of equal width *(Δx)* and from each interval choose a point, *xi.*

The Gradient

To calculate a derivative of a function which is dependent on more than one variable or multiple variables, a Gradient takes its place. A gradient is calculated using Partial Derivatives. Also, another major difference between the Gradient and aderivativeis that a Gradient of a function produces a Vector Field.

The Jacobian

The Jacobian of a set of functions is a matrix of partial derivatives of the functions. If you have just one function instead of a set of functions, the Jacobian is the gradient of the function.

The Hessian

The rate of change of the function is simply described as Hessian. It is positive definite and helps us to check if point x is a local maxima, local minima, or a saddle point. The function attains an isolated local minimum and an isolated local maximum at *x*, if the Hessian is positive definite negative definite at x, respectively. X is a saddle-point for function, if the Hessian has both positive and negative eigenvalues.

The Lagrange multipliers

The Lagrange multiplier technique lets you find the maximum or minimum of a multivariable function *f(x,y,…)* when there is some constraint on the input values you are allowed to use.

Laplace interpolation

Consider a (two dimensional) data matrix with some values missing, and you want to fill the holes by interpolated values. One particularly simple but at the same time powerful method is Laplace interpolation. For each missing value, take the average over the 4 surrounding values, that is, of the entries above, below, left and right. The Laplace interpolation replaces these values by interpolated ones and writes them back to the input matrix sample2.

Optimization

In ML, the focus is onlearning from data. A cost function is a measure of how wrong the model is in terms of its ability to estimate the relationship between X and Y.

The objective of an ML model, therefore, is to find parameters, weights, or a structure thatminimizesthe cost function. The cost function can be estimated by iteratively running the model to compare the estimated predictions of y by the model against the known values ofy.

We will use the Mean Squared Error function to calculate the cost. At first, we will find the difference between the actual y and predicted y value *(y = mx + c)*, for a given *x*. Then we will square this difference. Finally, we will calculate the mean of the squares for every value in X.

$$Y = mX + c$$

$$E = \frac{1}{n}\sum_{i=0}^{n}(y_i - \acute{y}_i)^2$$

$$E = \frac{1}{n}\sum_{i=0}^{n}\big(y_i - (mx_i + c)\big)^2$$

The Gradient Descent algorithm

Gradient Descent is an iterative optimization algorithm to find the minimum of a function. Here that function is our *Loss Function.*

Let’s try applying gradient descent tomandcand approach it step-by-step:

1. Initially let *m = 0* and *c = 0*. Let L be our learning rate. This controls how much the value ofmchanges with each step. L could be a small value like *0.0001* for good accuracy.
2. Calculate the partial derivative of the loss function with respect to m, and plug in the current values of *x, y, m* and *c* in it to obtain the derivative value *D*. D_m is the value of the partial derivative with respect to *m*.

$$D_m = \frac{1}{n}\sum_{i=0}^{n} 2\big(y_i - (mx_i + c)\big)(-x_i)$$

$$D_m = \frac{-2}{n}\sum_{i=0}^{n} x_i(y_i - \acute{y}_i)$$

Similarly let’s find the partial derivative with respect to *c*, *Dc:*

$$D_c = \frac{-2}{n}\sum_{i=0}^{n}(y_i - \acute{y}_i)$$

Below is a Python code showing the Gradient Descent Algorithm:

```
m = 0
c = 0
L = 0.0001# The learning Rate
iterations = 1000 # The number of iterations to perform
gradient descent
n = float(len(X)) # Number of elements in X # Performing
Gradient Descent
for i in range(iterations):
 Y_pred = m*X + c # The current predicted value of Y
 D_m = (-2/n) * sum( X * (Y - Y_pred)) # Derivative w.r.t m
 D_c = (-2/n) * sum( Y - Y_pred) # Derivative w.r.t c
 m = m - L * D_m # Update m
 c = c - L * D_c # Update c
print (m, c)
```

Figure 2.21 shows the red straight line as the linear regression line fitting:

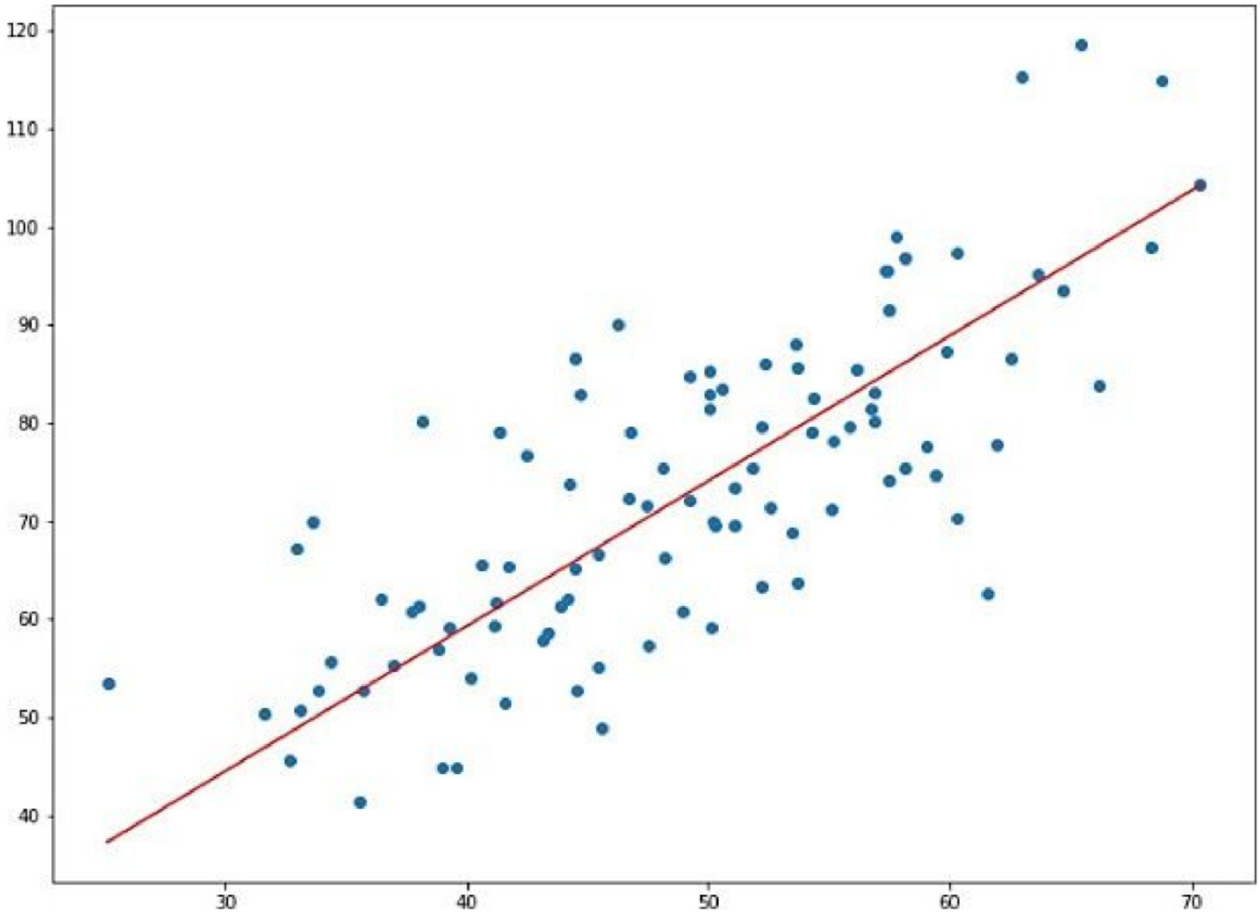


Figure 2.21: Linear regression line fitting

The equation: *Y_pred = mX + c* will give the best fit linear equation for the data set as shown as the red straight line in the *Figure 2.21*.

Conclusion

In this chapter, we have learned the essential mathematics for any data science applications. This is not limited to learn the concepts behind the algorithms. The sections are created starting from basic mathematical notation data to the advanced level mathematical concepts applied in the data science applications.

By reading this chapter, the reader will have better learning to approach any data science applications from the mathematical perspective. In the next chapter, we will start exploring statistical analysis and the techniques involved in the data analysis. We will discuss the basic statistical concepts that a data science practitioner needs to understand.

CHAPTER 3

Statistics Essentials

If we consider data science as an art, then statistics is the key to perform the operations on it. If we see from the perspective of high-level concepts, we can say statistics is the application of mathematics to perform technical based analysis on data. Simply by using the bar chart, we infer some high-level information from the data, but if we use statistics, we can get deeper into the data and find much more information towards the objectives of the analysis. Statistics provide solid proof about our data, not just random estimation.

Using statistics, we can go further in deep and more fine-grained essential insights into how exactly the given data is structured and based on that structure how we can optimally apply other data science techniques to get even more information. In the previous chapter, we studied the essentials of mathematics. In this chapter, we're going to look at basic statistics concepts that data scientists need to know and how they can be applied most effectively.

Structure

- Introduction to probability and statistics
- Descriptive statistics
- Conditional probability
- Random variable
- Inferential statistics
- Conclusion

Objectives

After studying this chapter, you should be able to:

- Understand the fundamentals of probability.
- Understand the importance of descriptive statistics and apply related

techniques.

- Apply the conditional probability calculations in a real-world problem.
- Understand the usage of random variables and inferential statistics.

Introduction to probability and statistics

Meteor showers are rare, but the probability of them occurring can be calculated. (credit: *Navicore/Flickr*). Probability is the chance that something will happen and how likely it is that some event will happen.

Probability of an event happening P(E) = Number of ways it can happen n(E)/ Total number of outcomes n(T)

We can say the probability is the measure of the chance of an event occurrence. The probability is measured with a scale of 0 to 1. Where 0 means impossibility and 1 is a certainty. In different aspects of our day to day activities, randomness and uncertainty happen. To understand those uncertainties, understanding the probability is very helpful. It enables us to create an estimationof what is going to come next. Also, we can say that probability learning works based on the informed assessment with the pattern of information collected earlier. In general, data science uses information based on the statistical properties. It must forecast or perform some analysis on the data to find some trends in it. Probability distribution has a major impact on statistical information formation. That makes that knowing probability and its application is very important to work with data science applications.

Statistics is a mathematical tool for collecting, analyzing, interpreting, and presenting the information. Statistics is used to deal with complicated issues in the actual globe so that data scientists and analysts can search for relevant trends and data modifications. In easy words, statistics can be used by making mathematical computations on it to obtain significant ideas from the information. In order to analyze raw information and create a statistical model and predict the outcome, different statistical functions, and algorithms are applied. Statistics has an impact on all areas of life, but to name a few are the stock market, insurance, weather, life sciences, retail, and education.

Descriptive statistics

In descriptive statistics, analysis of data is done so that information can be described, revealed, or summarized meaningfully so that anyone who sees it can detect specific patterns. The first stage in your statistical analysis, when looking at information, is to determine whether the dataset is a population or a sample.

Figure 3.1 shows that collecting the interest items based on the study gives the population and the notation of it is capital letter *N:*

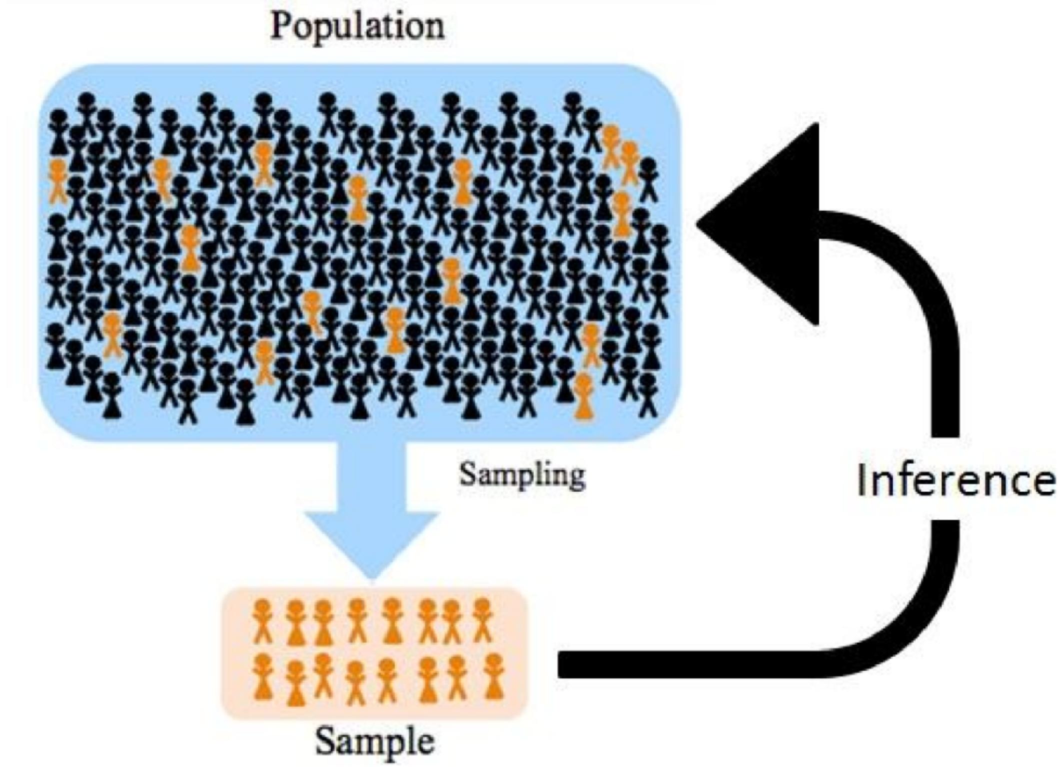


Figure 3.1: *Population, Sample, and Inference*

From that parameters are the calculated values in the analysis of the population. On other hand, the population subset is known as a sample. It is denoted by using the small letter n.

It is very difficult to evaluate and define the populations in real life. Studying the entire population and it influences. This process is inefficient because of time-consuming process as well; its cost is too high. Errors will easily acquire when we use the entire population as it is.

A sample is opposite to the population, meaning that a sample will consider only part of a population to evaluate, and this makes the process easier because it reduces the size of data. It directly implies the assessment time of a sample is very less expensive, and the occurrences of mistakes are very less. If we choose a random size sample, then that must act as a representative for the entire population. This must provide the flexibility to anybody to deduct the population from the sample.

Figure 3.2 shows that there are different varieties of data. In a dataset, data may be in the form of numerical or categorical values. Categorical data represented by groups or categories in which they belong. For example, ages, names, automotive brands are coming under the categorical data. Numerical data further divided into two categories discrete numbers and continuous numbers. Let us see in brief about the types of numerical data:

- **Discrete:** This type of data can only take certain values. Only a set of values to which you have access are defined. Age, number of vehicles on a road, number of fingers, for instance.
- **Continuous:** This type of data may, without limitations, take any true or fractional value between certain ranges. For example, weight, the balance of a bank account, purchase value, examination grade.

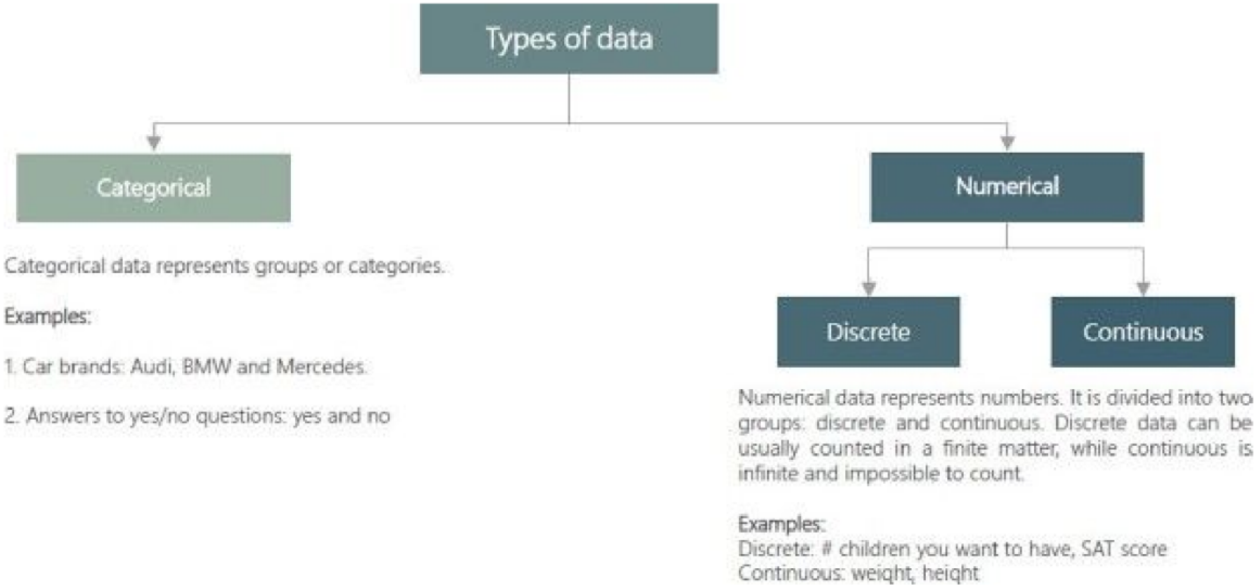

Types of data
Categorical
Numerical
Discrete
Continuous
Categorical data represents groups or categories.
Examples:
1. Car brands: Audi, BMW and Mercedes.
2. Answers to yes/no questions: yes and no
Numerical data represents numbers. It is divided into two groups: discrete and continuous. Discrete data can be usually counted in a finite matter, while continuous is infinite and impossible to count.
Examples:
Discrete: # children you want to have, SAT score
Continuous: weight, height

Figure 3.2: *Types of Data*

Measurement of data can be done in two levels: qualitative and quantitative.

- **Qualitative data:** This type of data measurement characterizes data but doesn't measure the attributes in data. It can be further divided into two groups: nominal and ordinal.
  - **Nominal:** Nominal data are not numbers, and it can't be placed in any

order.

Example: Gender (for example,male, female, others)

- **Ordinal:** Ordinal data consist of groups and categories in strict order

  Example: Grades (for example,good, satisfy, bad)

- **Quantitative data:** It measures attributes in the data. It can be divided into two groups: interval and ratio.
  - **Interval:** It is represented by numbers, without having a true zero. In this case, the zero value is meaningless.
  - **Ratio:** It is represented by numbers and has a true zero.

Based on the context, we can take the quantitative data as an interval or ratio. For instance, consider the temperature value. When we say, 00 Celsius or 00 Fahrenheit is not having any significance, since it is not really zero.

It depends on the context in which the quantitative data is considered as an interval or ratio. Think of the temperature, for instance. There is no significance saying 0º Celsius or 0º Fahrenheit since it is not the real zero. The value of absolute zero temperature is -273.15 ºC and -459.67 ºF. So, the temperature must, therefore, be considered as interval data in this instance, since the zero value is irrelevant.

But if the temperature is measured in Kelvin is analyzed, the value of absolute zero is 0º Kelvin, so now the value of the temperature is ratio since it is a true zero.

The measure of central tendency

The measure of central tendency relates to the concept that there is a number that summarizes the whole set best. Mean, median, and mode are the most popular.

Mean

It is the most reliable measure of central tendency for hypothesizing a single sample population. In the case of a population value, μ symbol is used, whereas the sample means denoted byx *(Refer Figure 3.3):*

Figure 3.3: *Mean value calculation*

Mean is calculated by adding all the components and then divide the sum value by the number of components. This is the most commonly used measure of central tendency of data. But it can get affected by the outlier's data easily. It is not enough to come to a conclusion sometimes because of the excess amount of outliers' presence. The below equation is representing the mean of x value with n components:

$$\acute{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Median

In an orderly ascending dataset, the median of the data set is the center value, also known as the 50th percentile. It is generally a good idea to calculate the median to prevent the error in the mean by outliers.

Below are the two equations for finding the median in the sets of data with an odd and even number of values:

1,3,3,6,7,8,9 Median = 6

1,2,3,4,5,6,8,9 Median = (4+5) ÷ 2 = 4.5

Mode

The mode gives us the most frequent value. It can be used for both numerical and categorical variables. If no single value appears more than once, then we can say that there is no mode. Rather than independently, the measures should be used together. In addition, these measures all appear at the same midline point in a normal distribution. The mean, mode, and medium are all the same!

Measures of variability

The measure of variability relates to the concept of evaluating the dispersion in our data by the mean value. Range, interquartile, variance, and standard deviation are the most common measures of variability.

Range

The range shows the difference between the largest and the smallest points in your data.

12,24,41,51,67,67,85,99

Range is 99-12 = 87

Variance

The variance and standard deviation are the most challenging methods to measure the dispersion of the data from the mean value of the dataset:

$$\text{Population Variance: } \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - X_{avg})^2}{n}$$

$$\text{Sample Variance: } s^2 = \frac{\sum_{i=1}^{n} (X_i - X_{avg})^2}{n-1}$$

The variance is calculated by measuring the difference between every data point and the mean and by squaring that value and adding all available data points. Lastly, the sum is divided by a total number of data points; thus the variance is calculated.

There are two main purposes for squaring the differences:

- Dispersion is a positive value because we square the subtraction to ensure that the negative values are not present and that they are not canceled.

- The effect of huge differences is amplified.

While calculating variance, squaring changes the unit of measurement from the original data. To nullify this problem, the standard deviation is computed, which is in the original unit.

Covariance

Covariance measures the changes in the two variables, x, and y, together:

$$\sigma(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x-\mu_x)(y-\mu_y)$$

- When two variables are the same, the variance is the covariance.

$$\sigma(x, x)=\sigma^2 (x)$$

- The variables are uncorrelated when the covariance value is 0
- Moreover $\sigma(x,y)=\sigma(y,x)$
- In an n-dimensional dataset, the covariance matrix, Σ, computes all possible pairs of dimensions: $\Sigma = \begin{bmatrix} \sigma(X_1,X_1) & \sigma(X_1,X_2) & \dots & \sigma(X_1,X_n) \\ \sigma(X_2,X_1) & \sigma(X_2,X_2) & \dots & \sigma(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(X_n,X_1) & \sigma(X_n,X_2) & \dots & \sigma(X_n,X_n) \end{bmatrix}$ where Xi refers to the i-th element of all the vectors in the feature space, the covariance matrix is a **symmetric matrix.**
- From a set of vectors, if we subtract the mean value from each vector is known as Mean Centering. You can form n mean-centered vectors into a matrix, Z, where each row of the matrix will match one of the vectors. The covariance matrix is then directly proportional to the transpose of Z multiplied by Z.
- $\Sigma Z^T Z$

Standard Deviation

Standard deviation is usually far more significant than a variance. It is the preferred variation metric, as it can be interpreted directly. Standard deviation is the square root of our variance.

$$\text{Population Standard Deviation: } \sqrt{\frac{\sum_{i=1}^{n}(X_i - X_{avg})^2}{n}}$$

$$\text{Sample Variance: } \sqrt{\frac{\sum_{i=1}^{n}(X_i - X_{avg})^2}{n-1}}$$

The best use of standard deviation is when data is in a unimodal shape (Refer *Figure 3.4*). In one standard deviation away from the mean, approximately 34% of data points are distributed in a normal distribution. Thus, we have 68.2% of data points arranged one standard deviation away from the mean since the normal distribution is symmetrical. Between the two standard deviations aside from the mean, approximately 95% of points are allocated, whereas, under three standard deviations, it is around 99.7%. Using Z-score, you can verify the total standard deviations below or above the mean.

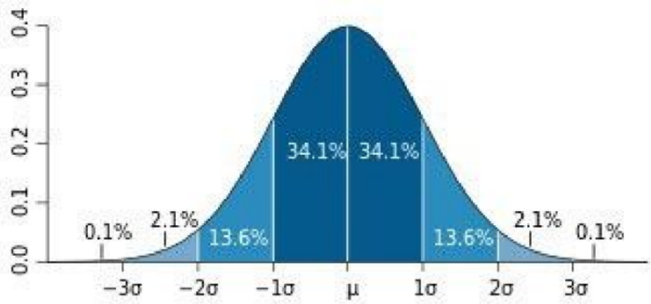


Figure 3.4: *Standard Deviation calculation*

This is how the measure of variability has been calculated by using standard deviation calculation. Let us move to the measure of asymmetry.

Measure of asymmetry

To measure the asymmetry in data, we can use two methods: Modality and Skewness. Let will start exploring the concepts.

Modality

The number of peaks presented by data allows for determining the Modality of a distribution. *Figure 3.5* shows different types of models. Generally, the majority of distribution is unimodal, and it has one value occurring frequently. There are two frequently occurring values in a bimodal. In uniform modal, the data distributed uniformly. In multi-modal data, more than two frequently occurring values will be there.

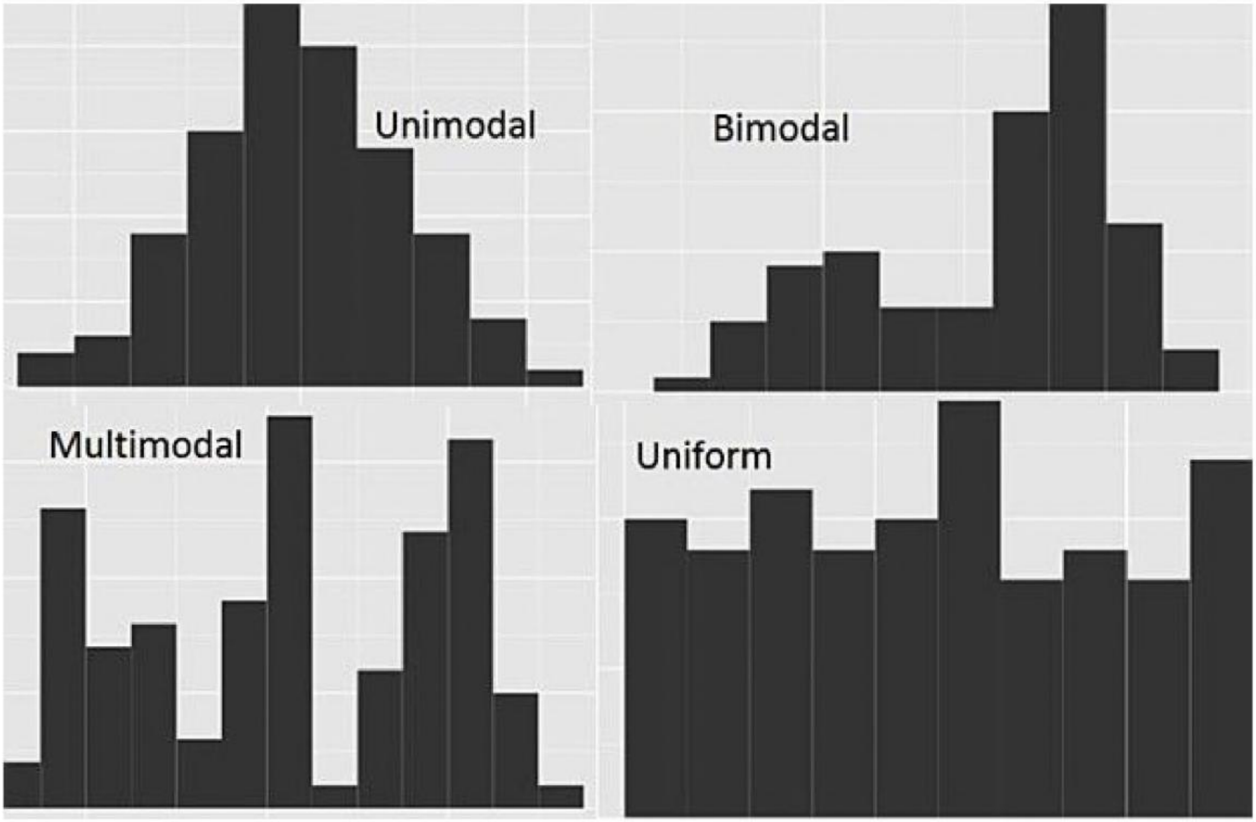


Figure 3.5: Types of Modal

Skewness

Skewness is one of the tools to measure asymmetry in data distribution. To view the where the data clustered more skewness helps to visualize it. Another important use of skewness is it can be used to capture the outliers in data. Based on the position of mean, mode, and median calculation, we find the skewness. *Figure 3.6* shows the different position of those values and how the asymmetry is calculated based on that. If the data has median value between the mean and mode, then it is called positive skew. In other words, we say outliers more towards the right side of the distribution. On the contrary, a median value higher than the mean value, then that is called as left-skewed. If the mean, median, and mode are at the same point, then that distribution is called **symmetrical distribution**.

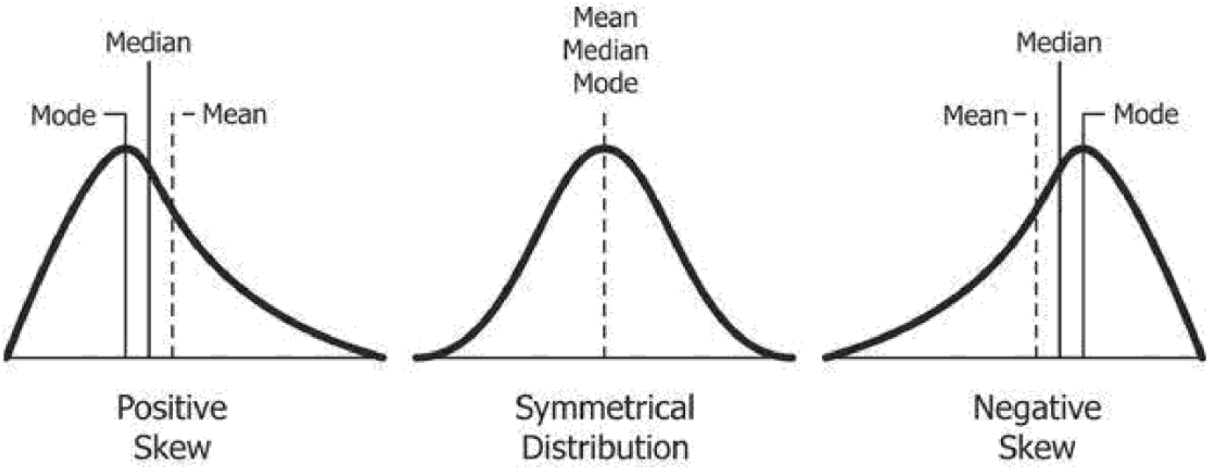


Figure 3.6: *Measurement of asymmetry using Skewness*

The connection between probability theory and central tendency measures are the measures of asymmetry. It helps to acquire more insights into the data we deal with.

Populations and samples

If the dataset contains entire data values, then that is called as populations. If we choose some random data from the population then that is called as samples. *Figure 3.7* shows that from the perspective of statistics, populations are parameters and samples are used to do statistical analysis:

- **Population:** It is the set of all possible states of a random variable. The size of the population may be either infinite or finite.
- **Sample:** It is a subset of the population. Normally, when the population is big enough to analyze the entire set, we use samples.

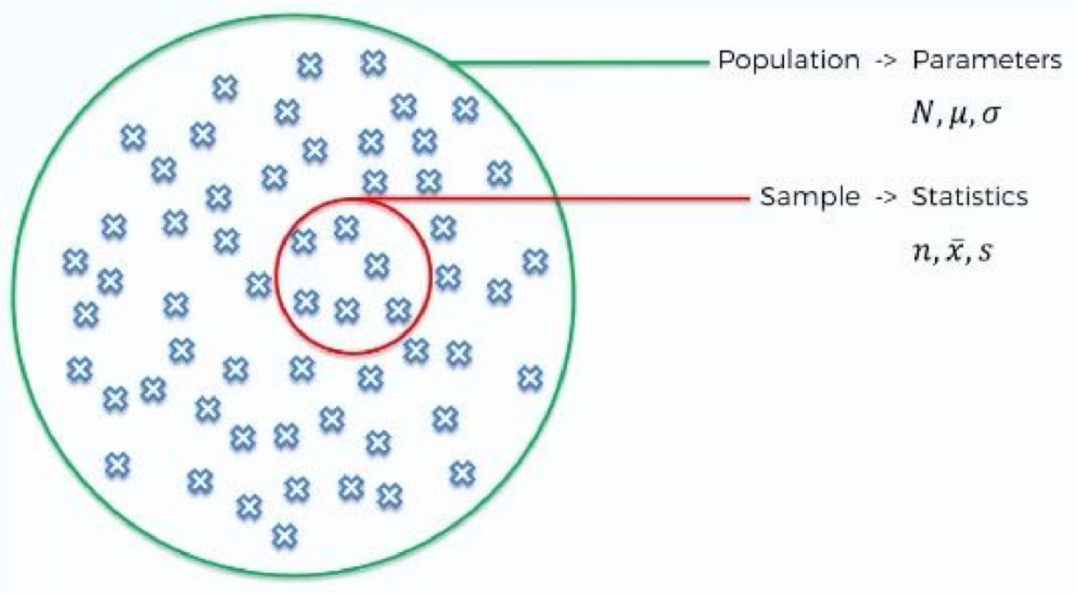


Figure 3.7: *Population and Sample*

Appropriate selection of populations and samples will provide more clarity on the processing of the data. That also helps to get more information about the data with less computing time and power.

Central Limit Theorem

It is a powerful and most crucial theorem of Mathematics. It states that *the sampling distribution will look like a normal distribution regardless of the population you are analyzing.*

Sampling distribution

As discussed earlier, to estimate the parameters of a population, we take a sample. But it's not the only way of extracting the exact estimates of the parameters. *Figure 3.8* shows the sampling distribution from a population. We can also take multiple samples from the population. For instance, we will calculate the average for every sample. So, at the end of the day, we have several mean estimate values, and we can then visualize them on a chart. This will be called the sampling distribution of the sample mean.

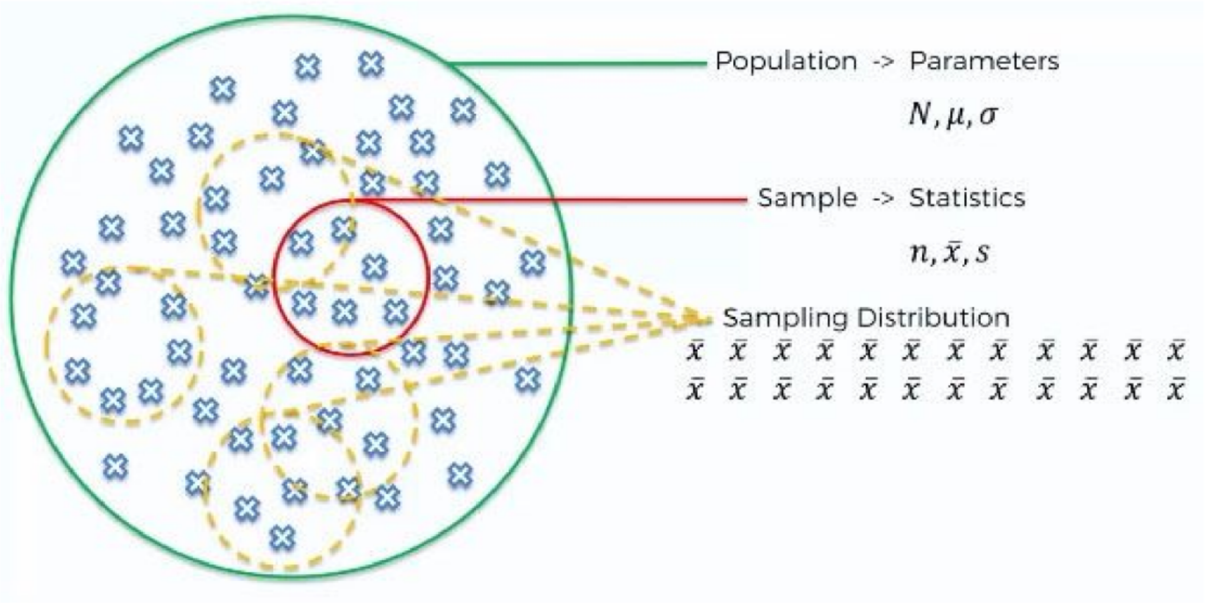

Population -> Parameters
N, μ, σ
Sample -> Statistics
$n, \bar{x}, s$
Sampling Distribution
$\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$
$\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$

Figure 3.8: *Sampling Distribution*

Identifying the sample distribution on the sample means it is very important to understand the data in a better way. The visualization of sample distribution is more useful to investigate the data distribution easily.

Conditional probability

Conditional probability can be characterized as a measure of the probability of an event provided that some other event has occurred.

$$P(B|A) = \frac{P(A \wedge B)}{P(A)}$$

The probability of an event B if event A is equivalent to the probability of event A and event B divided by the probability of event A. Most of the data science techniques depend on Bayes' Theorem. It is a formula that describes how to update the probability of hypotheses if the evidence is provided:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Given a new set of attributes, it is possible to create a learner using the Bayes' theorem that calculates the probability of the variable belonging to some other class:

$$posterior = \frac{prior \times likelihood}{evidence}$$

In a case, where A represents a hypothesis Hand B represents some observed evidence E, the equation can be written as:

$$P(H|E) = \frac{P(E|H)}{P(E)} P(H)$$

It co-relates the probability of the hypothesis before acquiring evidence P(H), to the probability of the hypothesis after receiving the evidence, P(H/E). Thus,P(H)is called theprior probability, while P(H|E)is called theposterior probability and $\frac{P(E|H)}{P(E)}$, is called thelikelihood ratio. Hence, Bayes' theorem can be rephrased as *the posterior probability equals the prior probability times the likelihood ratio.*

Random variables

A random variable is a collection of probable values of a random experiment, or it can be characterized as a variable whose probable values are the product of a random experiment. There can be discrete or continuous random variables. Within a range, continuous random variables can take any value, but discrete random variables can only take certain values.

A discrete random variable can be defined as a variable that can only take into consideration a limited amount of specific values such as 0, 1, 2, 3, 4, ... A random variable must be discrete if it is able to take only a finite number of distinct values. For example, we can take counting the number of candidates in the examination hall, the number of students in a school, the number of patients in a doctor's chamber, the number of faulty light bulbs in a set of eight are coming under the distinct random variables.

A discrete random variable’s probability distribution is a list of probabilities associatedwith every possible value. Let us suppose kdifferent values are taken by a random variable X, with the probability that X = xi, which can be defined as P (X = xi) = pi. The probabilities pi must meet the following requirements:

1. *0≤ pi≤1 for each i*
2. *p1+ p2+ ... + pk= 1*

For all the discrete and continuous random variables, there must be a cumulative distribution function. For every value of x, the probability that the random variable X being less than and equal to *x* is given by this function. The cumulative distribution function can be determined by summing up all the probabilities for a discrete random variable.

Acontinuous random variableis one which takes an infinite set of possible values. They are basically measurements; for example: height, weight, the amount of salt in toothpaste, the time required to walk 1 kilometer. A continuous variable is defined over anintervalof values and is expressed by thearea under a curve. As the random variable may take an infinite number of values, hence the probability of occurrence of value is found to be 0.

Let us suppose that over an interval of real numbers, all values are taken by a random variable X. Then, the probability that X is in the set of outcomes *A*, *P(A)*, is defined to be the area above *A* and under a curve. The curve representing a function *p(x)*, must meet the below points:

1. There are no negative values in the curve *(p(x)≥0 for all x)*.
2. The total area under the curve is 1.

Understanding the random variables is very important to use them effectively in the analysis of data based on inferential statistics. Let us start reading more about the inferential statistics in the next section.

Inferential statistics

In Inferential statistics, a random sample of data from a population is used for the purpose of describing the population and predicting it. Inferential statistics are basically intended to make inferences on populations based on the samples taken from data. This gives us a conclusion that descriptive statistics describe the data, and inferential statistics enables you to estimate from that data.

In this chapter, we will be analyzing the below concepts.

Probability distributions

- Normal
- Binomial
- Poisson

- Geometric
- Exponential

What is a probability distribution?

A probability distribution is a mathematical function that can be interpreted to be the probability that various possible effects happen during an experiment. Also, you can say that it is a function showing the probable values and how often they occur. It is the rule that determines the relation of the values with each other. *Figure 3.9* describes associations between different distributions. Most of them follow Bernoulli distribution:

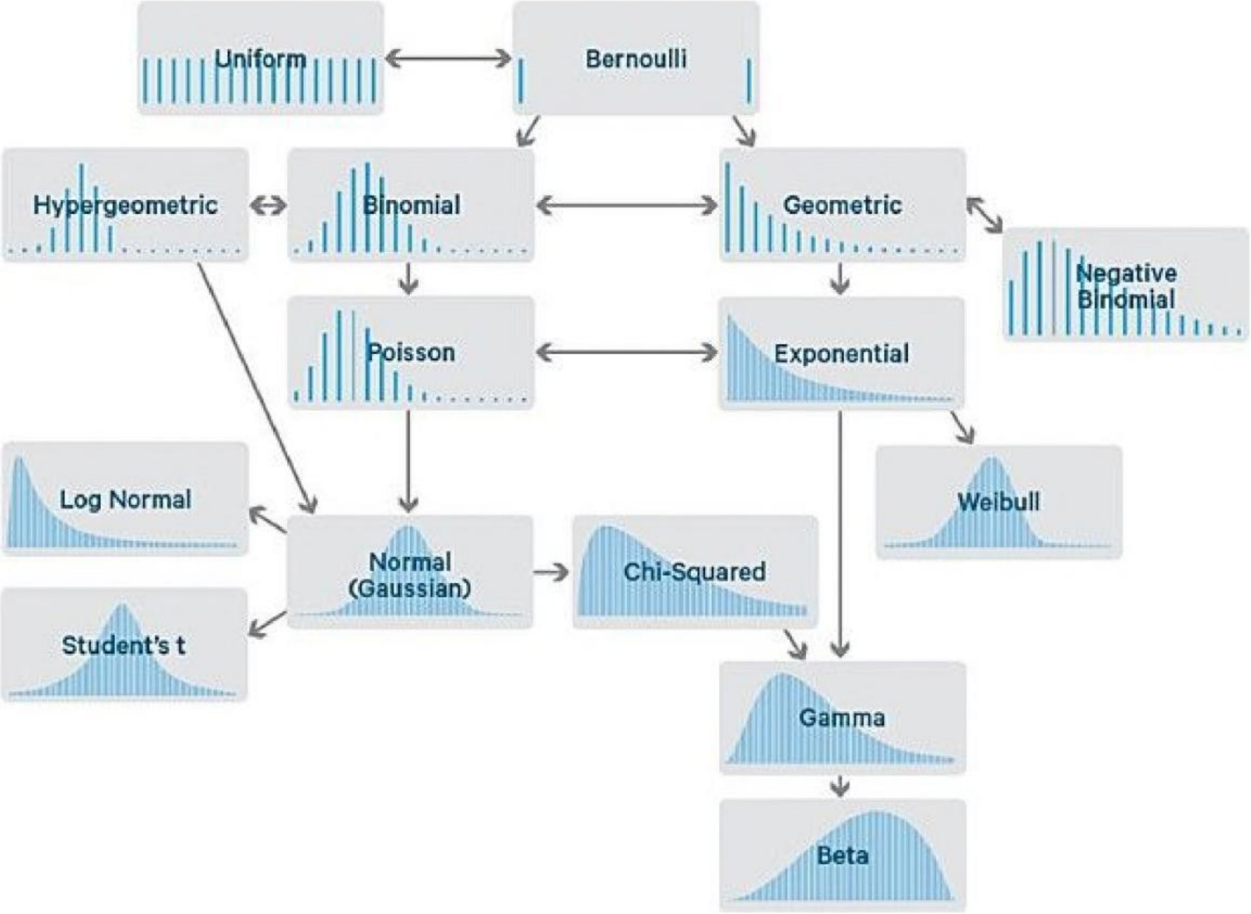


Figure 3.9: *Types Probability Distribution*

Let us begin with the most widely used distribution, normal distribution.

Normal distribution

The most common continuous variable probability distribution is the normal distribution, which is also known as Gaussian distribution or the Bell curve. *Figure 3.10* shows the normal distribution with its function. It is represented by the below notation:

$$N(\mu,\sigma^2)$$

Where *N* means normal, ~ represents the distribution, μis the mean, and σ^2 is the variance.

Normal distribution does not have any skewness as it is symmetrical, and its median, mean, and mode are of the same value.

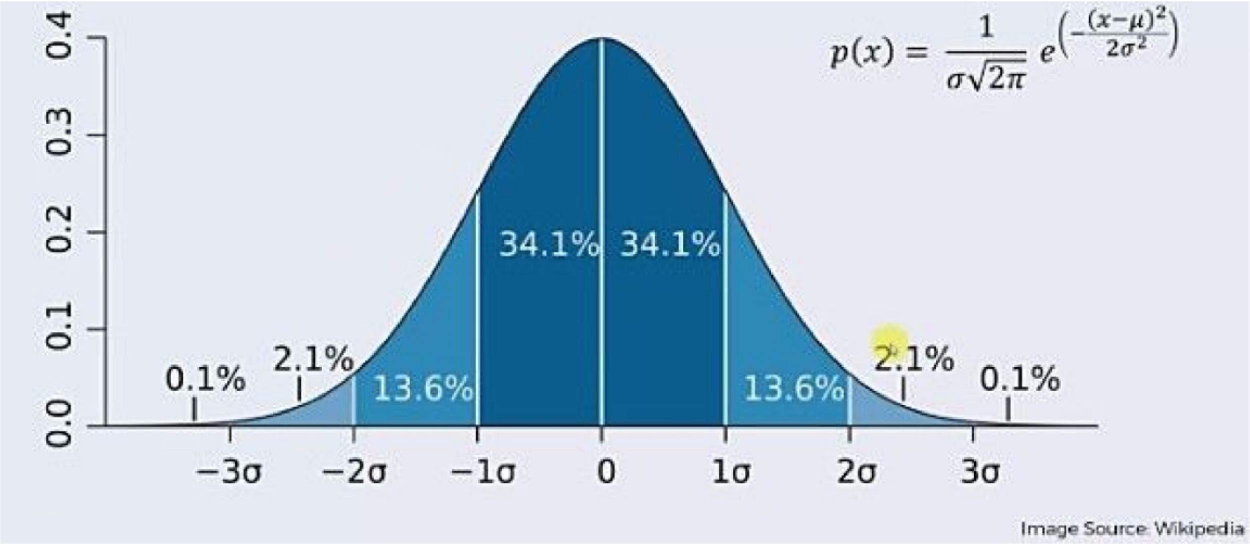


Image Source: Wikipedia

***Figure 3.10:** Normal Distribution*

For example, using Python's `scipy.stats module's rvs()` method, let us generate 10000 random variables. The loc parameter defines the mean and scale defines the standard deviation of the distribution.

Let us start to generate some random numbers and see how the normal distribution is applied by code:

```
import pandas as pd
import seaborn as sns
from scipy import stats
from scipy.stats import norm, binom
```