

Data Visualization with Python

*Exploring Matplotlib, Seaborn, and
Bokeh for Interactive Visualizations*

Dr. Pooja



www.bpbonline.com

Copyright © 2023 BPB Online

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor BPB Online or its dealers and distributors, will be held liable for any damages caused or alleged to have been caused directly or indirectly by this book.

BPB Online has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, BPB Online cannot guarantee the accuracy of this information.

First published: 2023

Published by BPB Online

WeWork

119 Marylebone Road

London NW1 5PU

UK | UAE | INDIA | SINGAPORE

ISBN 978-93-55515-384

www.bpbonline.com

Dedicated to

This book is dedicated to my beloved parents, who have been my guiding lights. Your unconditional love, unwavering support, and boundless encouragement have shaped me into who I am today.

I would also like to dedicate this book to my wonderful daughters. You are the inspiration behind my pursuit of knowledge and the driving force behind my aspirations. Your curiosity, enthusiasm, and zest for learning constantly remind me of the joy that lies in discovery. May this book serve as a tribute to the love and joy you bring into my life.

I offer my deepest gratitude and love to my parents and my daughters. Your presence and influence have enriched my journey, and I dedicate this book to you with immense appreciation and affection.

About the Author

Dr. Pooja is an accomplished individual with almost two decades of experience in imparting education and making significant contributions in the field of Computer Science and Engineering. With a strong background in education, she has dedicated her career to sharing knowledge and inspiring others through her expertise. She has delivered numerous hands-on training sessions to students/learners/industry personnel on Artificial Intelligence and Machine learning using Python. Furthermore, she has worked with NITTTR, Chandigarh, towards the co-creation and delivery of courses for the national online education portal “Swayam portal” (NPTEL) on “Smart grid analytics” implementing a “Machine Learning module.”

Throughout her professional journey, she has published over 90 publications, which include national and international journal/conference papers and book chapters, including IEEE, Springer conferences, and Scopus Indexed Journals. The extensive body of work reflects their commitment to advancing knowledge and making valuable contributions to the field of Artificial Intelligence/Machine Learning and Deep Learning. Her expertise and insights have likely been sought after by peers, students, and fellow researchers alike.

About the Reviewers

- ❖ **Arun Kumar** is a Lead Data Scientist at SkyBridge Infotech with a Bachelor of Engineering degree. With a strong background in Fintech, Automotive, and Banking, he brings diverse industry experience to his role. He is currently dedicated to the fields of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL).

As an accomplished professional, Arun's expertise lies in leveraging data-driven insights to drive business growth and innovation. With a focus on AI, ML, and DL, he possesses in-depth knowledge of advanced analytics techniques and their practical applications. Arun's proficiency in these areas enables him to analyze complex datasets, develop predictive models, and extract valuable insights.

Arun's experience working in different domains equips him with a comprehensive understanding of industry-specific challenges and opportunities. This knowledge allows him to evaluate technical content effectively, making him an ideal candidate for a technical reviewer role. His attention to detail, analytical mindset, and commitment to accuracy ensure the quality of technical materials.

Beyond his professional endeavors, he likes to play volleyball and marathon running. You can reach Arun Kumar at [<https://www.linkedin.com/in/arunkumar040891>].

- ❖ **Sakil Ansari** is a versatile data scientist specializing in Artificial Intelligence (AI). He holds a bachelor's degree in Computer Science and Engineering from Jawaharlal Nehru Technological University Hyderabad, and a master's degree in Machine Learning from the same institution. Ansari has conducted research at esteemed institutions like the Indian Institute of Technology Madras (IITM) and the Indian Institute of Science (IISc) Bangalore. With twelve research papers published in international journals, he actively contributes to the scholarly community and has participated in numerous AI conferences. Ansari authored the book "Introduction to Natural Language Processing - A Practical Guide for Beginners" and has a passion for open-source initiatives, creating machine learning libraries for public use.

He is an experienced corporate trainer in data science and has a broad range of research interests, including natural language processing, speech processing, neural networks, and music information retrieval. With expertise in analytics problem-solving across various industries, such as Manufacturing, Retail, Finance, Entertainment, Sports, Automotive, and Healthcare, Ansari is adept at designing customized solutions based on advanced analytics and user requirements. He remains committed to leveraging data to create a better future for all.

Acknowledgement

There are several individuals and organizations I would like to express my gratitude for their unwavering support throughout the process of writing this book. Their encouragement and assistance have been invaluable, and I am truly thankful for their contributions.

First and foremost, I would like to extend my deepest appreciation to my family. Their continuous support and belief in my writing endeavors have been instrumental in completing this book. Without their unwavering encouragement, I would not have been able to accomplish this feat.

I am also grateful to the course and the companies that provided support during my learning journey of Data Visualization. I would also like to thank MTF for providing me with the opportunity to deliver Data Visualization sessions, which compelled me to excel more in this area and initiate writing a book on it.

To all those who provided hidden support and assistance, I offer my heartfelt thanks. Your behind-the-scenes contributions have not gone unnoticed, and I am truly grateful for your help.

I would like to express my sincere acknowledgment to Arun Kumar and Sakil Ansari for their kind technical scrutiny of this book. Their expertise and valuable insights have greatly enhanced the quality of the content.

Furthermore, I would like to extend my gratitude to the team at BPB Publications. Their unwavering support and flexibility in allowing me ample time to complete the book and publish it is truly appreciated. Their understanding and cooperation have been instrumental in making this book a reality.

I extend my heartfelt thanks to all the individuals and organizations mentioned above for their continuous support, guidance, and belief in this project. It is through their contributions that this book has come to fruition.

Preface

In today's data-driven world, understanding and interpreting data is becoming increasingly crucial. Whether you are a researcher, a business professional, or simply someone interested in uncovering insights, the power of data visualization cannot be overstated. This book serves as a comprehensive guide to the art and science of data visualization, equipping you with the knowledge and tools to effectively communicate complex information through visually compelling representations.

The chapters in this book cover various aspects of data visualization, providing a structured approach to learning and applying these techniques. Here is a brief overview of what each chapter explores:

Chapter 1: Understanding Data- In this chapter, we lay the foundation by exploring the fundamentals of data. We discuss data types, sources, and formats and introduce key concepts such as variables, observations, and data structures. Understanding data is essential for effective visualization, as it allows us to identify relevant information and prepare it for visualization.

Chapter 2: Data Visualization - Importance- Building upon the understanding of data, we delve into the reasons why data visualization plays a vital role in understanding and communicating data, highlighting its ability to reveal patterns, trends, and correlations that might otherwise remain hidden.

Chapter 3: Data Visualization Use Cases- We examine a range of practical use cases for data visualization. From business analytics to scientific research, data visualization finds application in diverse fields, enabling us to make informed decisions, identify outliers, and communicate findings effectively.

Chapter 4: Data Visualization Tools and Techniques- To help you embark on your data visualization journey, we provide an overview of essential tools and techniques. We explore popular data visualization tools like various types of charts and plots, offering insights into their features and capabilities.

Chapter 5: Data Visualization with Matplotlib- In this chapter, we focus specifically on Matplotlib, a powerful library for creating static, animated, and interactive visualizations in Python. We cover its various plotting functions and customization options to create visually appealing visualizations.

Chapter 6: Data Visualization with Seaborn- In this chapter, we explore Seaborn, a high-level data visualization library built on top of Matplotlib. We discuss its specialized functions for statistical plotting and how it simplifies the creation of aesthetically pleasing visualizations.

Chapter 7: Data Visualization with Bokeh- In this chapter, we dive into Bokeh, a Python library for interactive data visualization. We explore its interactive plotting capabilities, including interactivity with web browsers and creating dynamic, interactive dashboards.

Chapter 8: Exploratory Data Analysis- We delve into the realm of exploratory data analysis, a crucial step in understanding and gaining insights from raw data. Through interactive visualizations and statistical techniques, you will learn how to uncover patterns, identify outliers, and much more that drive further analysis.

Throughout this book, we strive to strike a balance between theoretical concepts and practical examples. We provide clear explanations, step-by-step tutorials, and real-world case studies to help you grasp the principles and apply them to your own data visualization projects.

Whether you are a beginner or an experienced practitioner, this book aims to expand your understanding and proficiency in the art of data visualization. We invite you to embark on this journey with us as we explore the captivating world of visualizing data and unlocking its transformative power.

Happy visualizing!

Code Bundle and Coloured Images

Please follow the link to download the *Code Bundle* and the *Coloured Images* of the book:

<https://rebrand.ly/bvc0kdo>

The code bundle for the book is also hosted on GitHub at **<https://github.com/bpbpublications/Data-Visualization-with-Python>**. In case there's an update to the code, it will be updated on the existing GitHub repository.

We have code bundles from our rich catalogue of books and videos available at **<https://github.com/bpbpublications>**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Did you know that BPB offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.bpbonline.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at :

business@bpbonline.com for more details.

At www.bpbonline.com, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on BPB books and eBooks.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at business@bpbonline.com with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit www.bpbonline.com. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit www.bpbonline.com.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

<https://discord.bpbonline.com>



Table of Contents

1. Understanding Data.....	1
Structure.....	1
Objectives.....	2
What is Data?	2
<i>Categories of data</i>	3
<i>Data attributes</i>	9
<i>The Purpose of data</i>	11
Data collection.....	13
Data processing.....	15
Conclusion.....	17
Points to remember	17
Questions	18
2. Data Visualization – Importance.....	21
Structure.....	22
Objectives.....	22
Introduction.....	22
Data visualization.....	23
<i>Factors that influence data visualization choices</i>	23
<i>Content</i>	23
<i>Audience</i>	24
<i>Context</i>	24
<i>Purpose</i>	24
<i>Dynamics</i>	24
Significance of visualization.....	24
<i>Data visualization identifies data trends</i>	25
<i>Data visualization tends to give the data a perspective</i>	25
<i>Data visualization provides the correct context for the data</i>	25

<i>Time is saved by data visualization</i>	25
<i>A data story is told through data visualization</i>	26
Benefits of data visualization.....	26
<i>Fast comprehension of information</i>	26
<i>Correlations in relationships</i>	26
<i>Trends over time</i>	27
<i>Frequency</i>	27
<i>Examining the market</i>	27
<i>Risk and reward</i>	27
<i>Responding to the market</i>	27
<i>Simple data sharing</i>	28
Data visualization trends.....	28
<i>Video visualization</i>	28
<i>Data beyond visuals</i>	29
<i>Data democratization</i>	29
<i>Data visualization is more social</i>	29
<i>Real-time visualization</i>	30
<i>Data storytelling</i>	30
<i>Data journalism is becoming more popular</i>	30
<i>Mobile and social data visualization</i>	31
<i>Artificial intelligence and machine learning</i>	31
<i>Mobile-friendly data</i>	32
Examples of current trends in data visualization and business intelligence.....	32
<i>Augmented analytics</i>	32
<i>Plug and play analytics solutions</i>	32
<i>Data Quality Management</i>	33
<i>Data governance</i>	33
<i>Integrated capabilities</i>	33
<i>NLP Driven Analytical Queries</i>	33
<i>Data-driven culture</i>	34

<i>Clear and concise data visualization</i>	34
<i>Collaborative business intelligence</i>	34
<i>Mobile BI</i>	34
Conclusion.....	35
Points to remember	35
Questions	36
3. Data Visualization Use Cases	37
Structure.....	37
Objectives.....	38
Introduction.....	38
<i>Importance of use cases</i>	39
Various domains where data visualization is implemented	39
<i>National Geographic</i>	39
<i>Improvado use case examples</i>	40
<i>Grafana</i>	43
19E.....	43
<i>Wufoo (Infinity-Box)</i>	44
<i>Payback</i>	44
<i>Associated Press</i>	45
<i>IDS</i>	45
<i>Lifetime Brands</i>	46
<i>Informatica</i>	46
<i>QualiSystems</i>	47
<i>How different industries and sectors utilize data visualization</i>	48
Possible reasons to use data visualization	51
Domain-wise use-case.....	53
<i>Other common data visualization use cases</i>	54
Conclusion.....	55
Points to remember	55
Questions	56

4. Data Visualization Tools and Techniques	57
Structure.....	57
Objectives.....	58
Introduction.....	58
Different types of visualizations.....	58
<i>Bar chart</i>	59
<i>Pie chart</i>	59
<i>Donut chart</i>	60
<i>Half donut chart and gauge chart</i>	61
<i>Nested pie chart</i>	61
<i>Line chart</i>	62
<i>Scatter plot</i>	63
<i>Cone chart</i>	64
<i>Pyramid chart</i>	65
<i>Funnel chart</i>	66
<i>Radar triangle</i>	66
<i>Radar Polygon</i>	67
<i>Polar chart</i>	68
<i>Area chart</i>	69
<i>Tree chart</i>	69
<i>Flowchart</i>	70
<i>Table</i>	71
<i>Geospatial map and choropleth map</i>	71
<i>Percentage bar</i>	72
<i>Radial wheel</i>	72
<i>Concentric Circles</i>	73
<i>Gantt chart</i>	74
<i>Network diagram</i>	75
<i>Timeline</i>	76
<i>Venn Diagram</i>	77
<i>Histogram</i>	77

<i>Mind map</i>	78
<i>Dichotomous key</i>	79
<i>PERT chart</i>	80
<i>Box plot</i>	80
<i>Heatmap</i>	81
Most popular data- visualization tools and techniques.....	82
<i>Tableau</i>	82
<i>QlikView</i>	83
<i>Microsoft Power BI</i>	83
<i>Data wrapper</i>	83
<i>Plotly</i>	84
<i>Sisense</i>	84
<i>Excel</i>	85
<i>Zoho Analytics</i>	85
Other available tools and techniques.....	85
<i>Infogram</i>	85
<i>Fusion charts</i>	86
<i>D3.js</i>	86
<i>High charts</i>	87
<i>Fine Report</i>	87
Use-case of commonly used tools	88
<i>Used tools on Facebook</i>	88
<i>Used tools in Google</i>	88
<i>Used tools on Twitter</i>	88
<i>Used tools on Netflix</i>	89
<i>Used tools by Intel</i>	89
<i>Used tools by retails companies</i>	90
Conclusion.....	90
Points to remember	90
Questions	91

5. Data Visualization with Matplotlib.....	93
Structure.....	94
Objectives.....	94
Introduction.....	94
Importing Matplotlib and its documentation.....	95
Understanding figure and subplots.....	96
Basic plots	104
<i>Line plot</i>	105
<i>Bar chart</i>	107
<i>Histogram</i>	108
<i>Scatter plot</i>	110
<i>Pie chart</i>	112
<i>Area plot</i>	114
<i>Boxplot</i>	115
Advanced plots.....	117
Saving the plots.....	121
Features of Matplotlib library	124
Conclusion.....	124
Points to remember	125
Questions	125
Annexure Installing Python.....	126
Annexure Pandas.....	127
Annexure Numpy.....	131
6. Data Visualization with Seaborn	135
Structure.....	135
Objectives.....	136
Introduction.....	136
Importing seaborn and its documentation	137
<i>Seaborn datasets</i>	139
Understanding figure, axes, subplots, and palette	140

<i>Figure, axes, and subplots</i>	140
<i>Palettes</i>	144
Categorical plots.....	146
<i>Bar plot</i>	147
<i>Count plot</i>	150
<i>Box plot</i>	152
<i>Violin plot</i>	154
<i>Strip plot</i>	157
<i>Swarm plot</i>	159
<i>Cat plot</i>	161
<i>FacetGrid</i>	164
Other plots.....	166
<i>Scatter plot</i>	166
<i>Line plot</i>	169
<i>Heatmap</i>	172
<i>Joint plot</i>	174
<i>Pair plot</i>	177
<i>Distribution plot</i>	182
<i>Regression plot</i>	185
Saving the plot.....	186
Conclusion.....	187
Points to remember.....	188
Questions.....	189
7. Data Visualization with Bokeh	191
Structure.....	191
Objectives.....	192
Introduction.....	192
Importing Bokeh and its documentation.....	194
Understanding figures, glyphs, axes, and palettes.....	194
<i>Figure</i>	194

<i>Glyphs</i>	196
<i>Axes</i>	199
<i>Changing the axis range</i>	200
<i>Adjusting the tick marks and labels</i>	200
<i>Changing the axis line color and width</i>	200
<i>Formatting tick labels</i>	200
Palettes and colours.....	200
<i>Palettes</i>	201
<i>Turbo</i>	201
<i>Colors</i>	202
Creating plots.....	203
<i>Scatter plot</i>	203
<i>Line plot</i>	205
<i>Bar plot</i>	206
<i>HeatMap</i>	208
<i>Histogram</i>	210
<i>Patch plot</i>	211
<i>Area plots</i>	213
<i>Stacked bar plot</i>	215
Creating interactive plots	217
Creating multiple plots.....	219
Saving the plot	226
Conclusion.....	228
Points to remember	228
Questions	230
8. Exploratory Data Analysis.....	231
Structure.....	231
Objectives.....	232
Introduction.....	232
Types of EDA.....	232

<i>Univariate analysis</i>	233
<i>Bivariate analysis</i>	234
<i>Multivariate analysis</i>	235
Steps to perform EDA.....	237
<i>Statistical analysis</i>	238
<i>Data cleaning</i>	242
<i>Handling missing values</i>	243
<i>Handling outliers</i>	244
<i>Missing inconsistent values</i>	248
<i>Data visualization</i>	249
<i>Visualizations for Univariate EDA</i>	250
<i>Visualizations for Bivariate EDA</i>	251
<i>Visualizations for Multivariate EDA</i>	252
<i>Feature engineering</i>	253
<i>Handling categorical values</i>	255
<i>Feature scaling</i>	257
<i>Standardization example on 'titanic_survival' dataset</i>	258
<i>Min-Max Scaling example on 'titanic_survival' dataset</i>	260
<i>Normalization example on 'titanic_survival' dataset</i>	261
<i>Correlation analysis</i>	262
EDA using dataset in Python.....	265
<i>Example 1</i>	267
Conclusion.....	271
Points to remember.....	271
Questions.....	273
Index	275-284

CHAPTER 1

Understanding Data

“Data really powers everything that we do.”

— Jeff Weiner

In this chapter, you will get familiar with “Data.” The chapter will provide an understanding of what Data is and what are the ways to collect it. The chapter also presents different ways of categorizing data with suitable examples of each type. More importantly, we will discuss how data can be analyzed and how it can be used properly.

The chapter presents an introduction to data and various categories into which it can be segregated. For a better understanding of it, the attributes of data are also elaborated. For any application, data cannot be used directly. Data preprocessing is quite an essential stage. The chapter provides necessary techniques and ways to preprocess the data as well.

Structure

In this chapter, we will discuss the following topics:

- What is Data?
 - Categories of Data
 - Data attributes

- Purpose of Data
 - Data Collection
 - Data Processing

Objectives

The foremost objective of this chapter is to make you familiar with data so that you have a basic understanding of what exactly you are working on in the upcoming chapter while visualizing the data. After studying this unit, you should be able to understand and analyze the data as various categories and data attributes, and further, you will be able to collect and prepare it well for further model building.

What is Data?

In our daily routines, we come across various important instances that can be termed as data. Let us assume you are on a stroll, and you meet someone. The conversation may start like this...*Hi! I am David. What is your name?* See here, “David” is important information and is a fact that the other person is referred to as David. Here, “David” is data, and if you are supposed to create a program/application which can fetch the names of the user, “David” will be considered as a string type data.

Let us take another example. You went to buy bread. The conversation might be:

A: "Do you have bread?"

B: Yes, how much do you want?

A: Please pack 2. How much do I have to pay?

B: It will cost you 100.50 INR.

See here we have a piece of essential information viz the number of packs required and the cost to be paid. The quantity is “2,” and the amount is “100.50”. Again, if we want our machine/system to calculate the cost, this type of data will be considered as integer and float type, respectively.

Many times we fill up some kind of a form for, say, customer support by providing some information. Or you go for your medical check-up, and at the reception, you would be required to fill in the basic information about yourself. The form may consist of yes-no questions where you will ‘Tick’ mark the correct option. Actually, the data is being collected through this form. Here, the data may be in the form of a symbol.

The weather forecast on your mobile screen is another example of data; this data is processed data coming from the meteorological department after analyzing the historical data.

Our routines are full of data. The data captured by your smartwatch on your body parameters, the messages you type, the photographs you upload on social media, and so on. So, we can conclude that data is a collection of numbers, floating points, strings, and symbols that represents some value or situation. Data is information that can be used and translated into a form that is effective and efficient for processing. Data is facts and statistics collected together for reference or analysis. We rely on data mostly to make decisions or analyze a situation.

Note: Data is information that can be used and translated into a form that is effective and efficient for processing.

Categories of data

Two broad categories in which data can be classified on the basis of their format are:

Structured Data

Structured data is data that is ordered and may be recorded in a certain way. Structured data is presented in an organized manner. Structured data is often kept in a computer in a tabular (rows and columns) format, with each column representing distinct data for a specific parameter known as an attribute/ characteristic/ variable and each row representing data of observation for multiple attributes. The data in the excel sheets, data pulled from finance teams, sales data, and CRM data are all structured data. Being in a pre-defined format, it is always easy to search for an element/ data item from the whole dataset. Please refer to the following figure:

1	Customer ID	Customer Name	Address	Phone number	Gender	Occupation	Passport Number	Social Security Number	Email ID	Remarks
2	C10-001	David	34, enclave 2323, cityroad, district, india	9841519646	Male	Teacher	AB983923	SA44277	cool.david@net.ru	
3	C10-002	Adam	856, enclave 36427, cityroad, district, india	9932262265	Male	Engineer	AB983924	SB36904	mightyadam4562@gmail.com	
4	C10-003	Ruth	34, enclave 2323, cityroad, district, india	9856193020	Female	Teaching Assistant	AB983925	SS28287	girl.Power1212@yahoo.com	
5	C10-004	Yuka	856, enclave 36427, cityroad, district, india	9924649579	Male	Self-Employed	AB983926	SJ35486	cool.david@net.ru	
6	C10-005	Robert	34, enclave 2323, cityroad, district, india	9992849407	Male	Mechanic	AB983927	SJ33500	mightyadam4562@gmail.com	
7	C10-006	Umar	856, enclave 36427, cityroad, district, india	9990874070	Male	Baker	AB983928	SN25758	girl.Power1212@yahoo.com	
8	C10-007	Otabek	34, enclave 2323, cityroad, district, india	9854468619	Male	Researcher	AB983929	SN44898	cool.david@net.ru	
9	C10-008	Ibrat	856, enclave 36427, cityroad, district, india	9877088676	Male	Self-Employed	AB983930	SN35993	mightyadam4562@gmail.com	

Figure 1.1: Structured data (the data is generated synthetically)

Unstructured data

Unstructured data is information that lacks a predefined data model or is not arranged in a predefined way. Unstructured data is often text-heavy, although it may also include data such as dates, figures, and facts. For instance, consider the data available on a webpage. It consists mostly of text; however, multimedia content is also available, viz. images, audio, video, and so on.

Consider the social media content; it includes text, emojis, special characters, GIFs, and so on. Social media content also falls under unstructured data.

Considering the data captured in the healthcare sector, the content written by the physician in the slip recommendation/notes is unstructured in nature. The data captured in the form of imaging is also unstructured in nature.

Thus, we can say that data, which are not in the traditional row and column structure, are unstructured in nature. It is always tedious to work on unstructured data due to the lack of any predefined format or schema. Further, unstructured data consumes more storage.

Note: Structured data is presented in an organized predefined manner like row-column format. Unstructured data lacks predefined format and is not organized.

Data can also be represented in the following categories:-

- Qualitative and Quantitative Data.
- Continuous and Discrete Data.
- Primary and Secondary Data.

Qualitative and Quantitative Data

Qualitative data are measurements of 'types,' and they can be represented by a name, symbol, or number code. Data concerning categorical variables constitute qualitative data (for example, what type). Qualitative data results from information, which has been classified.

Quantitative data are numerical variables' values (for example, how many; how much; or how often). Quantitative data occurs when the measurement of data is possible on a scale. Quantitative data can also be discrete or continuous data varying on the elements being used and observed.

Refer to *table 1.1* for better understanding. Here 'age' and 'total marks' are numeric variables containing quantitative data values (numeric values), while 'Fail/Pass status' and 'Gender' are categorical variables holding qualitative values.

Data Instance	Quantitative data		Qualitative data	
	Age	Total Marks obtained	Fail/Pass status	Gender
Student				
David	25	82	Pass	Male
Ruth	22	80	Pass	Female
Adam	23	40	Fail	Male
Luka	25	42	Pass	Male

Table 1.1: Qualitative - Quantitative data

Some numeric variable examples:

- "How many siblings do you have?"
- "How much do you earn?"
- "How many days do you work?"
- "How much is the area of your house?"
- "How often do you visit your aunt?"
- "How many employees are above 40?"

In the *table 1.2* below students have been categorized according to the age group bracket they fall in. Students falling or belonging to the same age group are grouped or huddled up together. These groupings are based on the age numbers of students, meaning the data is Numerical and thus referred to as Quantitative data.

Age	No. of Students
18 years and under	6
19-21 years	85
22-25 years	115
26-30 years	100
30 years and above	74

Table 1.2: Number of students in an age group

The *table 1.3* shows the data of the different specific times that people tend and usually wake up. What is being observed or taken under consideration here is the time that these people usually wake up.

Wake up Time	No. of people
5AM-6AM	35
7AM-9AM	50
10AM-1130AM	60
12PM-130PM	20
2PM-3PM	15

Table 1.3: Number of persons as per wake-up time

Some categorical variable examples-

- “Are you a student?”
- “In which country were you born?”
- “What is the occupation of your father?”
- “Will they play today?” (Yes/No form)
- “Which category does this flower belong to?”
- “Is it a dog or a cat?”

Flower features sepal length, sepal width, petal length, petal width (cm)	Category
5.1, 3.5, 1.4, 0.2	Iris Setosa
7.0, 3.2, 4.7, 1.4	Iris Versicolour
6.5, 3.2, 5.1, 2.0	Iris Virginica

Table 1.4: Categories of flowers based on their characteristics (Iris dataset)

Note: Quantitative data is the value of a numeric variable. Qualitative data is the value of categorical variable

Continuous and discrete data

Continuous data is data that can take any value. It appears as a sequence of values. Height, weight, temperature and length are all examples of continuous data. It represents the information that could be meaningfully divided into its finer levels. It can be measured on a scale or continuum and can have almost any numeric value. This type of data is referred to as Continuous data.

Discrete data is a type of data that includes whole, concrete numbers or categorical variables with specific and fixed data values determined by counting. Discrete data on the other hand may be shown in gaps in scale, with no real values to be found.

For example, the number of students in a class is an example of discrete data since we can count whole individuals but can't count like 2.5, 3.75, kids. In simple words, discrete data can take only certain values and the data variables cannot be divided into smaller parts. It has a limited number of possible values for example days of the month.

Name	Gender	Age	Survived
Allen, Miss. Elisabeth Walton	female	29	1
Allison, Master. Hudson Trevor	male	0.9167	1
Allison, Miss. Helen Loraine	female	2	0
Allison, Mr. Hudson Joshua Creighton	male	30	0
Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25	0
Anderson, Mr. Harry	male	48	1
Andrews, Miss. Kornelia Theodosia	female	63	1

Table 1.5: Titanic survival data instances

In the *table 1.5* above instances of ‘Titanic Survival’ dataset are pulled to elaborate on continuous and discrete data. Here the categorical data variable ‘Gender’ is discrete in nature because it has only two values (Countable) viz. male and female. Similarly the data values of ‘Survived’ are also discrete in nature (0 or 1) while the data in ‘age’ is continuous in nature and we can also subdivide the values into categories like adult, infant, senior citizens and so on.

Characteristics of Continuous data

- Continuous elements are not counted, but are measurable.
- Continuous data values can be categorized and further subdivided into smaller pieces with additional meaning.
- It is usually graphically displayed by histograms.
- Continuous data is first and foremost present and gives a better sense of variation.

Some continuous data examples:

- The weight of people.
- The height of footballers.
- The waking up time of people.
- Speed of cars.
- Weight of trucks.
- The height of children.

- House prices.
- Temperature

Characteristics of Discrete data

- Discrete data can be counted and is usually counted in whole numbers.
- Discrete data cannot be measured at all.
- Discrete data values and elements cannot be subdivided into smaller pieces.
- It is usually graphically displayed by a Bar Graph.
- Binary attributes are a special case of discrete attributes where the count of discrete values is always two (0/1, False/True).
- Discrete data may also be ordinal or nominal data.
- It may be ordinal data meaning when the values fit into one of many categories and there is an order or rank to the values.
- It may be nominal data meaning when the values fit into one or many categories, especially where there is not any order between the values.

Some discrete data examples:

- The number of students admitted to a College.
- The number of people attending a Seminar.
- The number of Football teams participating in a Tournament.
- The number of cars in a Car Dealership.
- The number of staff working in a company.
- The number of patients admitted to a hospital.
- The number of teachers working in a school.

Note: Continuous data is data that can take any value. It appears as a sequence of values. Discrete data is countable/fixed values. It can take only certain values

Primary and secondary data

Primary data is data that is collected by people or on behalf of the person who is going to make use of the data. We can say, it is the data collected for the first time. *For example*, if you contact children's parents and ask them about the educational qualifications of their children concerning their performance this also grants or gives them Primary data. Whereas Secondary data is data used by a person or by people other than the people whom it was intended for. We can say secondary data is the data that have already been collected by some other person

Characteristics of Primary Data

- Usually collected for the first time.
- Original and more reliable than most types of data.
- It is first-hand information gathered and collected usually by an Investigator or Surveyor.

Characteristics of Secondary Data

- It is second-hand information collected, gathered and reported.
- Usually obtained from already published or unpublished sources.
- Useful tips for using Secondary data.
- How should the data be collected and processed?
- Accuracy of the data.
- How far the data can and should be summarized.
- Comparing the data with other tabulations.
- How to interpret the data?

Note: Primary data is data that is collected for the first time. Secondary data is the data that have already been collected by some other person

Data attributes

Data is a collection of data objects and their attributes. In a particular dataset, we get the features and instances with feature values. These features are actually the attributes of the data while available instances are data objects. These instances possess values for all or some attributes.

Attributes				
	name	gender	age	survived
Objects	Allen, Miss. Elisabeth Walton	female	29	1
	Allison, Master. Hudson Trevor	male	0.9167	1
	Allison, Miss. Helen Loraine	female	2	0
	Allison, Mr. Hudson Joshua Creighton	male	30	0
	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25	0
	Anderson, Mr. Harry	male	48	1
	Andrews, Miss. Kornelia Theodosia	female	63	1

Figure 1.2: Data Attributes and Objects

We can understand that an attribute is a property or characteristic of a data object. For instance, if we have to create a dataset of persons, eye color, hair color, height, weight, face shape can be considered as attributes. An attribute can also be referred to as variable, field, characteristic, dimension, or feature. and attribute values are numbers, symbols, values assigned to an attribute for a particular object. An object can be referred to as a point, instance, record, entry, and sample.

Note: Data Attribute is the property or characteristic of a data / data object.

Majorly the attributes can have the following type, [NOIR]:-

Qualitative (Categorical) data

- Nominal (N)
- Ordinal (O)

Quantitative (Numeric) data

- Interval (I)
- Ratio (R)

Nominal

A nominal attribute is used to name, label, or categorize certain measurements or features. It accepts qualitative values representing several categories, yet these categories are not intrinsically ordered. Although numbers can be used to code nominal variables, the order is arbitrary and arithmetic operations cannot be performed on the numbers.

A nominal variable is the simplest of all measurement variables and is one of two types of categorical variables. A person's phone number, national identification number, postal code, and other personal information are examples. A nominal value can be classified into two or more groups. Gender, for example, is a nominal variable that may accept the values as male/female or M/F. A nominal variable is qualitative, which implies that numbers are solely employed to categorize or identify items in this context. The number on the back of a player's shirt, for example, is used to indicate the position he or she is playing. They can also take numerical values. These quantitative values, however, lack numeric features. That is, they cannot be used for mathematical operations. They only possess the property of distinctness (equal or not equal).

Ordinal

The ordinal attribute value provides sufficient information to order the objects. They are built upon nominal scales by assigning numbers to objects to reflect a rank or ordering on an attribute. Also, there is no standard ordering in the ordinal