

# Data Science with Raspberry Pi



Real-Time Applications  
Using a Localized Cloud

K. Mohaideen Abdul Kadhar  
G. Anand

# **Data Science with Raspberry Pi**

**Real-Time Applications  
Using a Localized Cloud**

**K. Mohaideen Abdul Kadhar  
G. Anand**

**Apress®**

# ***Data Science with Raspberry Pi: Real-Time Applications Using a Localized Cloud***

K. Mohaideen Abdul Kadhar  
Pollachi, Tamil Nadu, India

G. Anand  
Pollachi, Tamil Nadu, India

ISBN-13 (pbk): 978-1-4842-6824-7  
<https://doi.org/10.1007/978-1-4842-6825-4>

ISBN-13 (electronic): 978-1-4842-6825-4

Copyright © 2021 by K. Mohaideen Abdul Kadhar and G. Anand

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director, Apress Media LLC: Welmoed Spahr  
Acquisitions Editor: Aaron Black  
Development Editor: Matthew Moodie  
Coordinating Editor: Jessica Vakili

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail [orders-ny@springer-sbm.com](mailto:orders-ny@springer-sbm.com), or visit [www.springeronline.com](http://www.springeronline.com). Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a Delaware corporation.

For information on translations, please e-mail [booktranslations@springernature.com](mailto:booktranslations@springernature.com); for reprint, paperback, or audio rights, please e-mail [bookpermissions@springernature.com](mailto:bookpermissions@springernature.com).

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at [www.apress.com/bulk-sales](http://www.apress.com/bulk-sales).

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at [www.apress.com/978-1-4842-6824-7](http://www.apress.com/978-1-4842-6824-7). For more detailed information, please visit [www.apress.com/source-code](http://www.apress.com/source-code).

Printed on acid-free paper

*To my wife Jashima for her support in  
writing this book.*

*—Dr. K. Mohaideen Abdul Kadhar*

*To my parents for their continuous  
encouragement in writing this book.*

*—G. Anand*

# Table of Contents

**About the Authors.....xiii**

**About the Technical Reviewer ..... xv**

**Acknowledgments ..... xvii**

**Introduction ..... xix**

**Chapter 1: Introduction to Data Science..... 1**

    Importance of Data Types in Data Science..... 3

    Data Science: An Overview ..... 4

    Data Requirements ..... 5

    Data Acquisition ..... 5

    Data Preparation ..... 5

        Data Processing..... 6

        Data Cleaning ..... 6

        Duplicates..... 6

        Human or Machine Errors..... 7

        Missing Values..... 7

        Outliers ..... 7

        Transforming the Data ..... 8

    Data Visualization..... 8

    Data Analysis ..... 9

        Modeling and Algorithms..... 9

        Report Generation/Decision-Making ..... 9

TABLE OF CONTENTS

Recent Trends in Data Science ..... 10

    Automation in Data Science ..... 10

    Artificial Intelligence–Based Data Analyst..... 10

Cloud Computing..... 11

Edge Computing..... 11

Natural Language Processing ..... 11

Why Data Science on the Raspberry Pi? ..... 12

**Chapter 2: Basics of Python Programming ..... 13**

    Why Python? ..... 14

    Python Installation ..... 14

    Python IDEs ..... 16

        PyCharm ..... 16

        Spyder ..... 16

        Jupyter Notebook ..... 17

        Python Programming with IDLE..... 17

        Python Comments ..... 20

    Python Data Types..... 21

        Numeric Data Types..... 21

        int ..... 21

        float ..... 22

        complex..... 22

        bool..... 22

        Numeric Operators ..... 23

        Sequence Data Types ..... 24

        Control Flow Statements ..... 31

    Exception Handling ..... 36

Functions .....	37
Python Libraries for Data Science.....	39
NumPy and SciPy for Scientific Computation.....	39
Scikit-Learn for Machine Learning.....	44
Pandas for Data Analysis.....	44
TensorFlow for Machine Learning .....	47
<b>Chapter 3: Introduction to the Raspberry Pi.....</b>	<b>49</b>
What Can You Do with the Raspberry Pi?.....	49
Physical Computing with the Raspberry Pi.....	50
How to Program the Raspberry Pi? .....	50
Raspberry Pi Hardware .....	50
System on a Chip.....	51
Raspberry Pi RAM.....	52
Connectivity.....	52
Setting Up the Raspberry Pi.....	53
microSD Memory Card .....	53
Installing the OS .....	53
Inserting the microSD Memory Card .....	55
Connecting a Keyboard and Mouse.....	56
Connecting a Monitor .....	56
Powering the Raspberry Pi.....	57
Raspberry Pi Enclosure .....	58
Raspberry Pi Versions .....	58
Raspberry Pi 1 .....	58
Raspberry Pi 2.....	58
Raspberry Pi 3.....	59

## TABLE OF CONTENTS

Raspberry Pi Zero (W/WH) .....	59
Raspberry Pi 4 .....	59
Recommended Raspberry Pi Version .....	59
Interfacing the Raspberry Pi with Sensors .....	60
GPIO Pins .....	60
GPIO Pinout.....	61
GPIO Outputs .....	62
Controlling GPIO Output with Python .....	62
GPIO Input Signals .....	64
Interfacing a Ultrasonic Sensor with the Raspberry Pi.....	66
Interfacing the Temperature and Humidity Sensor with the Raspberry Pi.....	68
Interfacing the Soil Moisture Sensor with the Raspberry Pi.....	71
Interfacing Cameras with the Raspberry Pi.....	72
Raspberry Pi as an Edge Device .....	75
Edge Computing in Self-Driving Cars .....	75
What Is an Edge Device? .....	76
Edge Computing with the Raspberry Pi.....	76
Raspberry Pi as a Localized Cloud.....	76
Cloud Computing .....	76
Raspberry Pi as Localized Cloud .....	77
Connecting an External Hard Drive.....	77
Connecting USB Accelerator.....	78
<b>Chapter 4: Sensors and Signals.....</b>	<b>79</b>
Signals .....	79
Analog and Digital Signals .....	80
Continuous-Time and Discrete-Time Signals.....	80



Deterministic and Nondeterministic Signals.....	81
One-Dimensional, Two-Dimensional, and Multidimensional Signals .....	81
Gathering Real-Time Data.....	82
Data Acquisition .....	82
Sensors.....	82
Analog Sensors.....	83
Digital Sensors .....	84
What Is Real-Time Data?.....	85
Real-Time Data Analytics .....	85
Getting Real-Time Distance Data from an Ultrasonic Sensor .....	85
Interfacing an Ultrasonic Sensor with the Raspberry Pi.....	86
Getting Real-Time Image Data from a Camera .....	87
Getting Real-Time Video from a Webcam .....	87
Getting Real-Time Video from Pi-cam .....	88
Data Transfer.....	88
Serial and Parallel Communication .....	88
Interfacing an Arduino with the Raspberry Pi.....	89
Data Transmission Between an Arduino and the Raspberry Pi.....	90
Time-Series Data .....	92
Time-Series Analysis and Forecasting .....	93
Memory Requirements.....	93
More Storage .....	93
More RAM.....	93
Case Study: Gathering the Real-Time Industry Data.....	94
Storing Collected Data Using Pandas .....	94
Dataframes.....	94
Saving Data as a CSV File.....	94

## TABLE OF CONTENTS

Saving as an Excel File .....	95
Reading Saved Data Files .....	95
Adding the Date and Time to the Real-Time Data .....	95
Industry Data from the Temperature and Humidity Sensor .....	96
<b>Chapter 5: Preparing the Data .....</b>	<b>99</b>
Pandas and Data Structures .....	99
Installing and Using Pandas .....	99
Pandas Data Structures .....	100
Series .....	100
DataFrame .....	104
Reading Data .....	104
Reading CSV Data .....	105
Reading Excel Data .....	105
Reading URL Data .....	106
Cleaning the Data .....	106
Handling Missing Values .....	107
Handling Outliers .....	110
Z-Score .....	113
Filtering Out Inappropriate Values .....	116
Removing Duplicates .....	118
<b>Chapter 6: Visualizing the Data .....</b>	<b>121</b>
Matplotlib Library .....	121
Scatter Plot .....	122
Line Plot .....	124
Histogram .....	127
Bar Chart .....	129

Pie Chart.....	132
Other Plots and Packages .....	134
<b>Chapter 7: Analyzing the Data .....</b>	<b>135</b>
Exploratory Data Analysis .....	135
Choosing a Dataset .....	135
Modifying the Columns in the Dataset .....	140
Statistical Analysis .....	141
Uniform Distribution .....	142
Binomial Distribution .....	144
Normal Distribution .....	146
Statistical Analysis of Boston Housing Price Dataset.....	150
<b>Chapter 8: Learning from Data .....</b>	<b>155</b>
Forecasting from Data Using Regression.....	156
Linear Regression using Scikit-Learn .....	160
Principal Component Analysis .....	162
Outlier Detection Using K-Means Clustering.....	166
<b>Chapter 9: Case Studies.....</b>	<b>171</b>
Case Study 1: Human Emotion Classification .....	171
Methodology .....	172
Dataset .....	172
Interfacing the Raspberry Pi with MindWave Mobile via Bluetooth.....	173
Data Collection Process.....	175
Features Taken from the Brain Wave Signal.....	177
Unstructured Data to Structured Dataset .....	181
Exploratory Data Analysis from the EEG Data.....	186
Classifying the Emotion Using Learning Models .....	188

TABLE OF CONTENTS

Case Study 2: Data Science for Image Data..... 191

    Exploratory Image Data Analysis ..... 195

    Preparing the Image Data for Model ..... 200

    Object Detection Using a Deep Neural Network ..... 201

Case Study 3: Industry 4.0 ..... 207

    Raspberry Pi as a Localized Cloud for Industry 4.0 ..... 208

    Collecting Data from Sensors ..... 210

    Preparing the Industry Data in the Raspberry Pi ..... 211

    Exploratory Data Analysis for the Real-Time Sensor Data..... 214

    Visualizing the Real-Time Sensor Data..... 216

    Transmitting Files or Data from the Raspberry Pi to the Computer ..... 223

**References.....229**

**Index.....233**

# About the Authors

**Dr. K. Mohaideen Abdul Kadhar** earned an undergraduate degree in electronics and communication engineering and a master of technology degree with a specialization in control and instrumentation. In 2015, he obtained his PhD in control system design using evolutionary algorithms. He has more than 14 years of experience in teaching and research. His areas of interest are evolutionary algorithms, control systems, signal processing and computer vision. Now, He is working to implement signal processing and control system concepts with Python programming on the Raspberry Pi. He has taught many courses and has delivered workshops about data science with Python programming. In addition, he has acted as a consultant for many industries in developing machine vision systems for industrial applications.

**G. Anand** obtained his bachelor of engineering degree in electronics and communication engineering in 2008 and his master of engineering degree in communication systems in 2011. He has more than nine years of teaching experience with a specialization in signal and image processing. He has taught courses and acted as a resource person in workshops related to Python programming. His current research focus is in the domain of artificial intelligence and machine learning.

# About the Technical Reviewer

**Maris Sekar** is a professional computer engineer, certified information systems auditor (ISACA), and senior data scientist (Data Science Council of America). He has a passion for using storytelling to drive better decision-making and operational efficiencies. Maris has cross-functional work experience in various domains such as risk management, data analytics, and strategy.

# Acknowledgments

First, I wish to thank the almighty Allah for giving me strength and courage in writing this book. Writing a book is more complex than I thought. We struggled many times when developing the content of this book because this book focuses not only the concepts but also on the real-time implementation details on the Raspberry Pi.

My sincere thanks to my family, especially my mom and dad. Without them, I would not have attained this level of achievement.

A very special thanks to my wife Mrs. M. Jashima Parveen for her support and love. She always set me free for writing this book. In my hard times, her support and encouragement gave me strength and courage. I could not have done it without her.

My sincere thanks to chief editor Mr. Aaron Black and book coordinator Ms. Jessica Vakkili for their enormous support. Even when some of the chapters were delayed, they gave their support in developing the contents of the book.

My heartfelt thanks to the management of Dr. Mahalingam College of Engineering and Technology, Pollachi, especially, I thank to my Head of the Department, Dr. R. Sudhakar, Professor, for his encouragement and trust in my work and knowledge.

Last but not least, special thanks to my colleague G. Anand for his support and coordination in writing the book.

# Introduction

In modern times data can be thought of as a valuable commodity like oil or gold because we can get a lot of useful information from data with the help of some scientific methods, and we can make intelligent decisions based on that information and convert it into money. *Data science* is the process of extracting knowledge/useful information from the data.

For example, IBM forecasted that the demand for skilled people in data science will increase by 28 percent in 2020. Many industries use data science concepts in different aspects of their business such as checking whether they have achieved their targets, finding the root cause of failures, etc. Recently, data science has been effectively implemented in politics to develop strategies, identify the weak regions, predict the emotions and expectations of the people, etc. Further, local governments utilize the data collected from the people of their town to devise the planning and policies for the development of the town. Data science is also successfully applied in the agricultural domain in areas like drought assessment, crops yield and remote sensing, etc. This shows that the applications related to data science concepts are emerging nowadays across multiple domains.

Most of the recent books have focused on applying data science techniques to some open and standard dataset. This book is specifically about applying data science concepts in the Raspberry Pi board. The Raspberry Pi can act as a single on board computer and can also interact with the real-time environment via sensors as most of the local servers can't do this task.

The book will start with a brief introduction to data science followed by which there will be a dedicated chapter for explaining the concepts of Python starting from the installation of the software to the various



## INTRODUCTION

data types and modules available. The next two chapters will introduce the readers to Raspberry Pi devices, their hardware description, and the setting up of the devices for gathering real-time data. The next four chapters will deal with the different operations in data science with respect to real time applications using Raspberry Pi hardware. The penultimate chapter of the book will discuss about the concepts that will enable the Raspberry Pi to learn from the data. The last chapter will have few case studies that will give the readers an idea of the range of domains where these concepts can be applied.

## CHAPTER 1

# Introduction to Data Science

Data is a collection of information in the form of words, numbers, and descriptions about the subject. Consider the following statement: “The dog has four legs, is 1.5m high, and has brown hair.” This statement has three different kinds of information (i.e., data) about the dog. The data “four” and “1.5m” is numerical data, and “brown hair” is descriptive. It is good to know the various kinds of data types to understand the data, perform effective analysis, and better extract knowledge from the data. Basically, data can be categorized into two types.

- Quantitative data
- Qualitative data

*Quantitative data* can be obtained only with the help of measurements and not through observations. This can be represented in the form of numerical values. Quantitative data can be further classified into continuous and discrete. The exact integer values are *discrete* data, whereas *continuous* data can be any value in a range. *Qualitative data* is a description of the characteristics of a subject. Usually qualitative data can be obtained from observations and cannot be measured. In other words, qualitative data may be described as categorical data, and quantitative data can be called numerical data.









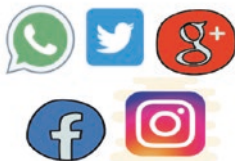

For example, in the previous statement, “brown hair” describes a characteristic of the dog and is qualitative data, whereas “four legs” and “1.5m” are the quantitative data and are categorized as discrete and continuous data, respectively.

Data can be available in structured and unstructured form. When the data is organized in a predefined data model/structure, it is called *structured data*. Structured data can be stored in a tabular format or a relational database with the help of query languages. We can also store this kind of data in an Excel file format, like the student database given in Table 1-1.

**Table 1-1.** *An Example of Structured Data*

Student Roll Number	Marks	Attendance	Batch	Sex
111401	492/500	98%	2011-2014	Male
111402	442/500	72%	2011-2014	Male
121501	465/500	82%	2012-2015	Female
121502	452/500	87%	2012-2015	Male

Most human-generated and machine-generated data are unstructured data such as emails, documents, text files, log files, text messages, images, video and audio files, messages on the Web and social media, and data from sensors. This data can be converted to a structured format only through human or machine intervention. Figure 1-1 shows the various sources of unstructured data.

   <b>Documents</b>	 <b>Log files</b>	 <b>Images</b>	 <b>Sensor data</b>
 <b>Video files</b>	 <b>Audio files</b>	 <b>Web and social media</b>	 <b>Email</b>

*Figure 1-1. Sources of unstructured data*

## Importance of Data Types in Data Science

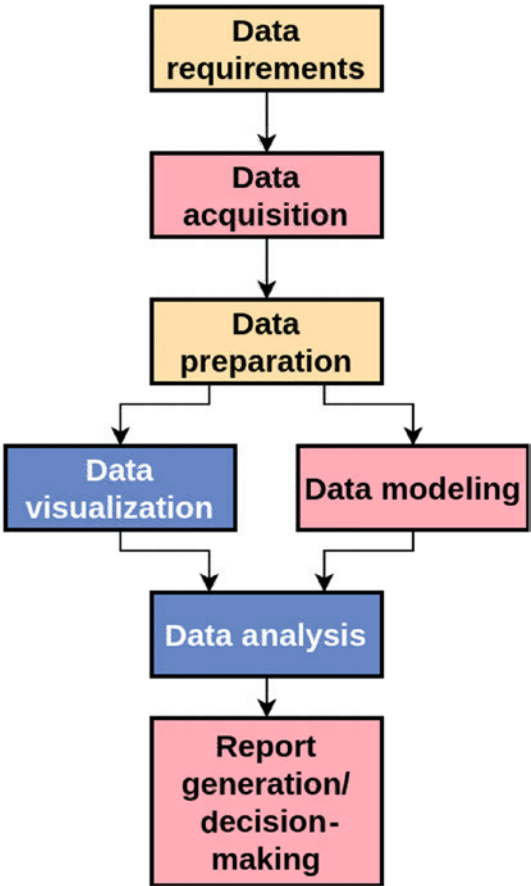
Before starting to analyze data, it is important to know about the data types so you can choose the suitable analysis methods. The analysis of continuous data is different from the analysis of categorical data; hence, using the same analysis methods for both may lead to incorrect analysis.

For example, in statistical analysis where continuous data is involved, the probability of an exact event is zero, while the result can be different for discrete data.

You can also choose the visualization tools based on the data types. For instance, continuous data is usually represented using histograms, whereas discrete data can be visualized with the help of bar charts.

# Data Science: An Overview

As discussed at the beginning of the chapter, data science is nothing but the extraction of knowledge or information from the data. Unfortunately, not all data gives useful information. It is based on the client requirements, the hypothesis, the nature of the data type, and the methods used for analysis and modeling. Therefore, a few processes are required before analyzing or modeling the data for intelligent decision-making. Figure 1-2 describes these data science processes.



*Figure 1-2. Data science process*

## Data Requirements

To develop a data science project, the data scientists first understand the problem based on the client/business requirements and then define the objectives of the problem for analysis. For example, say a client wants to analyze the emotion of people on a government policy. First, the objectives of the problem can be set as “To collect the opinion of the people about the government policy.” Then, the data scientists decide on the kind of data that can support the objective and the resources of data. For the example problem, the possible data is social media data, including text messages and opinion polls of various categories of people, with information about their education level, age, occupation, etc. Before starting the data collection, a good work plan is essential for collecting the data from various sources. Setting the objectives and work plan can reduce the time spent collecting the data and can help to prepare the report.

## Data Acquisition

There are many types of structured open data available on the internet that we call *secondary data*, because that kind of data is collected by somebody and structured into some tabular format. If the user wants to collect the data directly from a source, that is called *primary data*. Initially, the unstructured data is collected via many resources such as mobile devices, emails, sensors, cameras, direct interaction with people, video files, audio files, text messages, blogs, etc.

## Data Preparation

Data preparation is the most important part of the data science process. Preparing the data puts the data into proper form for knowledge extraction. There are three steps in the data preparation stage.

1. Data processing
2. Data cleaning
3. Data transformation

## Data Processing

This step is important as it is required to check the quality of data while we import it from various sources. This quality checking is done to ensure that the data is in the correct data type, standard format, and has no typos or errors in the variables. This step will reduce data issues when doing analysis. Moreover, in this phase, the collected unstructured data can be organized in the form of structured data for analysis and visualization.

## Data Cleaning

Once the data processing is done, cleaning the data is required as the data might still have some errors. These errors will affect the actual information present in the data. Possible errors are as follows:

- Duplicates
- Human or machine errors
- Missing values
- Outliers
- Inappropriate values

## Duplicates

In the database, some data is repeated multiple times, which results in *duplicates*. It is better to check and remove the duplicates to reduce the overhead in computation during data analysis.

## Human or Machine Errors

The data is collected from sources either by humans or by machines. Some errors are inevitable during this process due to human carelessness or machine failure. The possible solution to avoid these kinds of errors is to match the variables and values with standard ones.

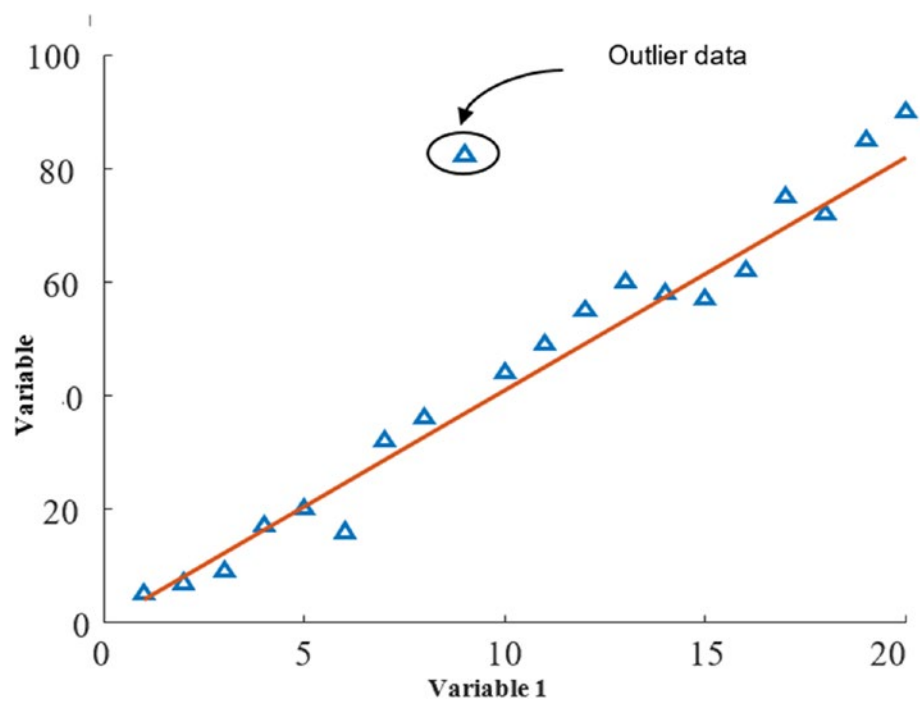
## Missing Values

While converting the unstructured data into a structured form, some rows and columns may not have any values (i.e., empty). This error will cause discontinuity in the information and make it difficult to visualize it. There are many built-in functions available in programming languages we can use to check if the data has any missing values.

## Outliers

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be because of variability in the measurement or it may indicate experimental errors; outliers are sometimes excluded from the data set. Figure 1-3 shows an example of outlier data. Outlier data can cause problems with certain types of models, which in turn will influence the decision-making.





**Figure 1-3.** *Outlier data*

# Transforming the Data

Data transformation can be done by many methods using normalization, min-max operations, correlation information, etc.

# Data Visualization

Based on the requirements of the user, the data can be analyzed with the help of visualization tools such as charts, graphs, etc. These visualization tools help people to understand the trends, variations, and deviations in a particular variable in the data set. Visualization techniques can be used as a part of exploratory data analysis.

## Data Analysis

The data can be further analyzed with the help of mathematical techniques such as statistical techniques. The improvements, deviations, and variations are determined in a numerical form. We can also generate an analysis report by combining the results of visualization tools and analysis techniques.

## Modeling and Algorithms

Today many machine learning algorithms are employed to predict useful information from raw data. For example, neural networks can be used to identify the users who are willing to donate funds to orphans based on the users' previous behavior. In this scenario, the previous behavior data of users can be collected based on their education, activities, occupation, sex, etc. The neural network can be trained with this collected data. Whenever a new user's data is fed to this model, it can predict whether the new user will give funds or not. However, the accuracy of the prediction is based on the reliability and the amount of data used while training.

There are many machine learning algorithms available such as regression techniques, support vector machine (SVM), neural networks, deep neural networks, recurrent neural networks, etc., that can be applied to data modeling. After data modeling, the model can be analyzed by giving data from new users and developing a prediction report.

## Report Generation/Decision-Making

Finally, a report can be developed based on the analysis with the help of visualization tools, mathematical or statistical techniques, and models. Such reports can be helpful in many circumstances such as forecasting the strengths and weakness of an organization, industry, government, etc.