SELF LEARNING
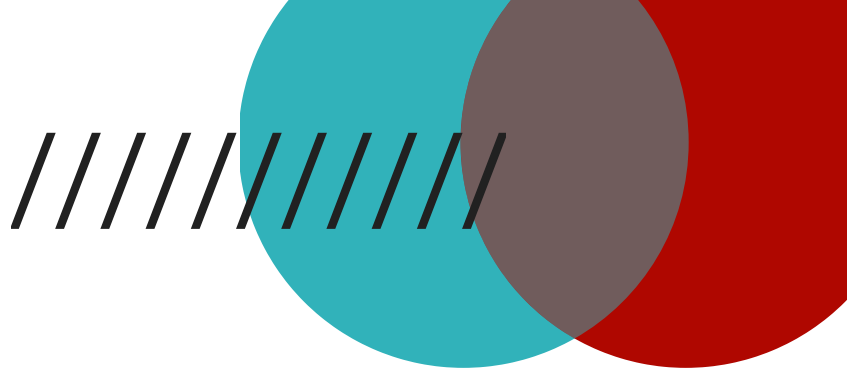
# DATA SCIENCE

IN 31 DAYS

////////////

SPECIAL EDITION

## THE RESEARCH NEST

Empowering Humanity With Exclusive Insights

## COMPILED BY

Aditya Vivek Thota

# PREFACE

With vast amounts of data being generated in recent times, there is an ever increasing need for professionals who can make any valuable sense out of it: the data scientists. Today, with humongous amounts of resources available online, self-learning is not beyond scope anymore.

This unique booklet is intended to enable individuals for the same by providing the best curated resources to learn and implement practical projects.

## THE PROJECT

The booklet is split into 4 major parts with each one laying emphasis on certain fundamental aspects of data science. The book further focuses on practicing data science using Python.

## NOTE FROM THE EDITOR

- Please do mention the credits when you share this booklet elsewhere.
- For any feedback/errata you can mail us at the.research.nest@gmail.com

## WHAT TO EXPECT?

**01** EXPLANATORY ARTICLES

**02** HANDS-ON TUTORIALS
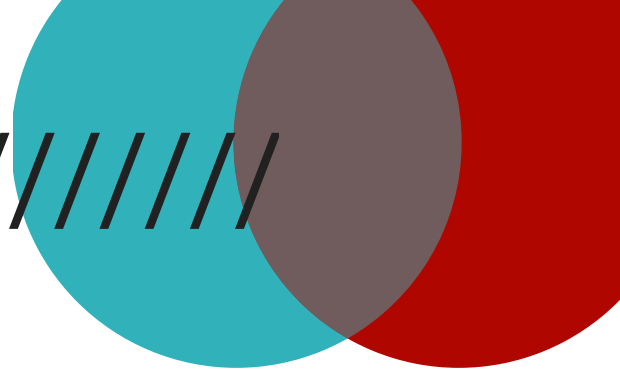
**03** PRACTICAL INSIGHTS

DATA PREPARATION

# PART ONE

"Data really powers everything that we do"

**- JEFF WEINER, CEO, LINKEDIN**

# FINDING YOUR DATA

The first step is all about identifying what domain you want to work in and finding the relevant dataset. Data science starts with data collection after all. Choose a dataset in your domain of interest, download the same and get ready for some action!

Below are some links, where you can find public datasets in different sectors:
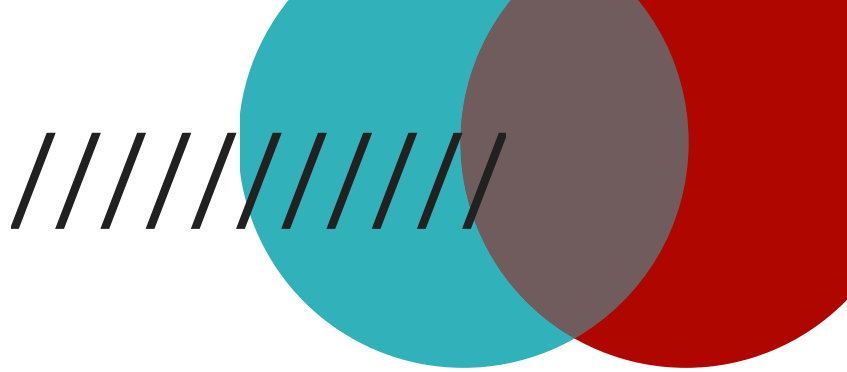
## REFERENCE LINKS TO OBTAIN DATASETS

**01** **KAGGLE**

**02** **UC IRVINE MACHINE LEARNING REPOSITORY**

**03** **A COMPILATION OF ALL PUBLIC DATASETS ON GITHUB**

## WHAT CAN YOU DO WITH YOUR DATASET?

Once you have your dataset ready, there are broadly (but not limited to) three kinds of applications you can build using the same. These include prediction, classification, or recommendation.

Apart from that, you can try to find hidden patterns in the data. Have a good look at your dataset and the variables in it. Identify what kind of analysis it can be used for and finalize the problem to tackle.

Is it classification, regression, or clustering based? If your dataset appears inconclusive to any of the above-mentioned categories, as a beginner we would recommend you to change your dataset and find a more relevant one.

# SUBJECTS AND PRE-REQUSITES

Here is a comprehensive compilation of learning resources you may need on your journey en-route to becoming a data scientist.
While you may not need to know all of them in detail to get started. Having a general idea of these topics can prove to be extremely useful.

## LINKS TO QUICKLY LEARN SOME KEY CONCEPTS

**01** FIVE BASIC STATISTICS CONCEPTS DATA SCIENTISTS NEED TO KNOW

**02** BASICS OF PROBABILITY FOR DATA SCIENCE

**03** A COMPREHENSIVE GUIDE TO LINEAR ALGEBRA FOR DATA SCIENTISTS

**04** CALCULUS IN DATA SCIENCE

# DATA PRE-PROCESSING

Before one can start analyzing the dataset, one needs to make some modifications to make it a bit more programming friendly. Here are some standard approaches used. Try implementing these techniques as per relevance for your chosen dataset.

## TUTORIALS OF VARIOUS PRE-PROCESSING APPROACHES

**01** HANDLING MISSING VALUES

**02** DEALING WITH CATEGORICAL DATA

**03** NORMALIZATION OF DATA
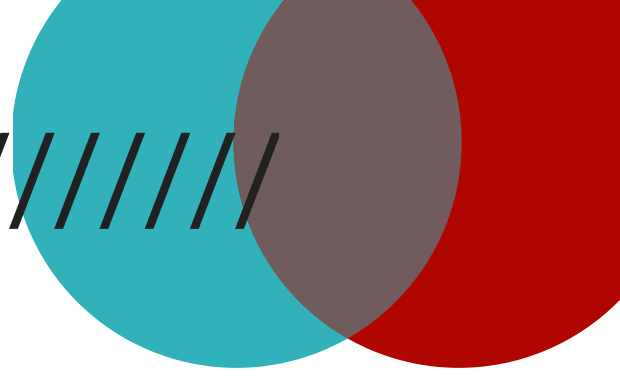
**04** DATA PRE-PROCESSING SUMMARY

EXPLORATORY DATA ANALYSIS

# PART TWO

"In God we trust. All others must bring data."

**- W. EDWARDS DEMING, STATISTICIAN**

## PERFORMING EDA

Once we have a detailed and clean dataset in hand, we can do various statistical analyses and visualizations to better understand our data.

Wikipedia has an entire page dedicated to EDA. You can refer the same to get the overview of what it is all about.

## LINKS TO SOME USEFUL RESOURCES

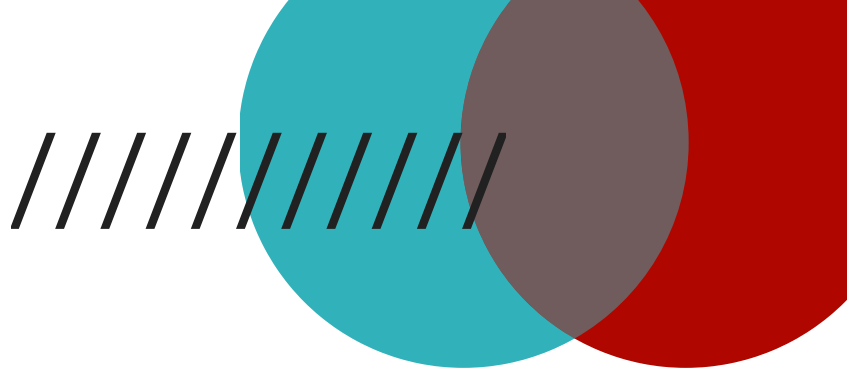**01** COMPREHENSIVE GUIDE TO DATA EXPLORATION

**02** VARIOUS EDA TECHNIQUES

**03** HANDS-ON KAGGLE TUTORIAL FOR EDA USING PYTHON

**04** WIKIPEDIA PAGE ON EDA

There are several libraries available in Python for performing EDA. You can easily find one based on your requirement and proceed further.
Once the data is thoroughly analyzed, we can proceed to the next step of building some predictive models using different techniques and ultimately formulate a tangible application with practical significance.

## TO LEARN MORE ABOUT THE STATISTICS BEHIND HYPOTHESIS TESTING, VISIT THESE LINKS:

**01** LECTURE SLIDES ON HYPOTHESIS TESTING

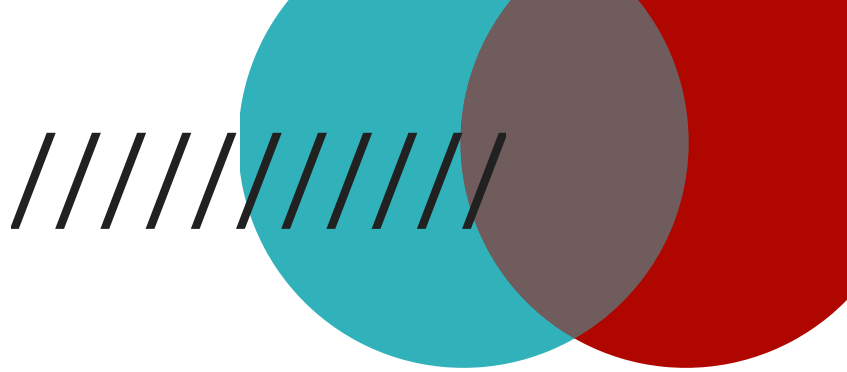**02** YOUR GUIDE TO MASTER HYPOTHESIS TESTING IN STATISTICS

# PART THREE

"Not everything that can be counted counts, and not everything that counts can be counted."

**- ALBERT EINSTEIN, PHYSICIST**

You have a clean dataset ready and doing an exploratory data analysis should give a very clear picture of what we can do with the dataset. Choosing the right model for the situation can be challenging for a beginner.

Based on your understanding, you can finalize to use 2-3 methods and get ready to build your model.

Here are two useful articles exploring basic machine learning algorithms for data science and the scenarios in which they are preferred.

**01**    TOP 10 MACHINE LEARNING ALGORITHMS

**02**    CHOOSING THE RIGHT ALGORITHM FOR YOUR DATASET

BUILDING YOUR MODELS

# PART FOUR

"The goal is to turn data into information, and information into insight."

**- CARLY FIORINA, FORMER CEO, HP**

# ESSENTIAL MACHINE LEARNING

With the dataset prepared and problem statements formulated, the stage is all set to build and train your models using various ML methods.

Here are some must-read resources for any aspiring data scientist summarizing almost everything you need to know.

## USEFUL REFERENCE LINKS

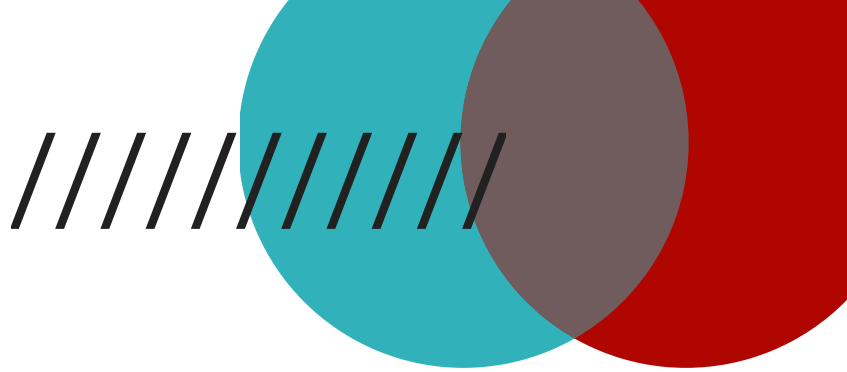**01**    HOW TO APPROACH (ALMOST) ANY MACHINE LEARNING PROBLEM?

**02**    IMPLEMENTATION OF DIFFERENT MACHINE LEARNING ALGORITHMS

**03**    THE ULTIMATE KAGGLE TUTORIAL FOR DATA SCIENCE

**04**    THE ULTIMATE KAGGLE TUTORIAL FOR MACHINE LEARNING

## ADDITIONAL HANDS-ON TUTORIALS

The following tutorials are for those interested in further exploring the practical applications of machine learning.
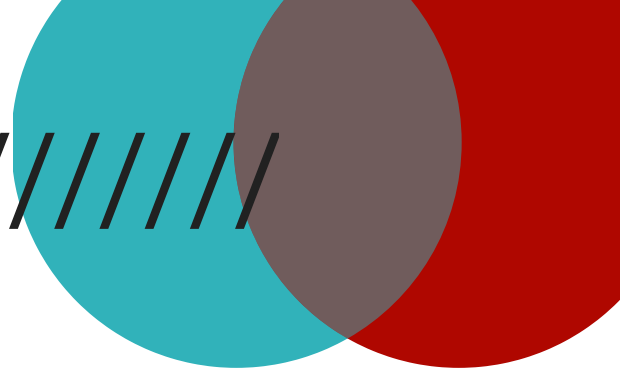
**01**    PREDICTING THE PRICE OF A HOUSE

**02**    SIGN LANGUAGE RECOGNITION USING HAND GESTURES

**03**    TEXT EMOTION DETECTION USING NATURAL LANGUAGE PROCESSING

# END NOTES

This compilation is a effort of The Research Nest and is associated with the e-learning social media campaign, The December Data Festival, 2018.

We would love to hear your feedback and suggestions for improvement. Do drop us a mail at the.research.nest@gmail.com.

Hope you found this useful. To support and stay updated with more such initiatives, please do follow Research Nest on their social media handles.

(Click to follow)