

Foundations of Data Science

R-2021 Anna University

CS3353



[Signature]
Mahendra College of Engineering,
Mahendra Salem Campus,
Minnampalli, SALEM 636 106
TAMILNADU

Dr. Lilly Beaulah H
Dr. Reka R
Ms. Gayathri M
Ms. Dhanakodi V

FIRST
EDITION

Copyright © 2022 Dr. Lilly Beulah H

All rights reserved.

This book contains material protected under International and Federal Copyright Laws and Treaties.

Any unauthorized reprint or use of this material is prohibited. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system without express written permission from the author/publisher.

Published & Printed in India.

ISBN: 978-81-957838-4-7

PRICE: INR 400

First Edition: October 2022




PRINCIPAL

Mahendra College of Engineering
Mahendra Salem Campus,
Minnampalli, SALEM 636 106
TAMILNADU

For information contact:

Email: whitecoastbooks@gmail.com

Website: <http://www.whitecoastbooks.com>

Dedicated
to
Our Family, Friends
&
Students



M. J.
PRINCIPAL

Mahendra College of Engineering,
Mahendra Salem Campus,
Minnampalli, SALEM 636 106
TAMILNADU

ACKNOWLEDGEMENT

We wish to record our sincere gratitude to the Managing Director **Er.B.Maha Ajay Prasath**, Mahendra College of Engineering, for his constant encouragement and kind support in all our endeavours.

We deem it a proud privilege to extend our greatest sense of gratitude to **Dr.R.Samson Ravindran**, Executive Director, Mahendra Engineering Colleges for the inspiring guidance and valuable suggestions throughout the pursuance of this report.

We express our profound thanks to **Dr.N.Mohana Sundararaju**, Dean Academics, Mahendra College of Engineering for his great enthusiasm and inspiration which enabled us to bring this venture to fruition.

We would like to express a special note of gratitude to the fantastic editing team of **White Coast Book Publishers** in releasing this book.

Finally, this work would not have been possible without the love and support of our **colleagues, family members and friends**. We are extremely grateful to one and all.

Dr. H.Lilly Beulah

Dr.R.Reka

Ms. M.Gayathri

Ms.V.Dhanakodi




PRINCIPAL
Mahendra College of Engineering
Mahendra Salem Campus,
Minnampalli, SALEM 636 106
TAMILNADU

PREFACE

The importance of “Foundations of Data Science” is designed for students with a complete walkthrough right from the foundational groundwork required to outlining all the concepts, techniques and tools required to understand Data Science.

Data Science is an umbrella term for the non-traditional techniques and technologies that are required to collect, aggregate, process, and gain insights from enormous datasets. This book offers all the processes, methodologies, various steps like data gaining, pre-process, mining, prediction, and visualization tools for extracting insights from huge amounts of data by the use of various scientific methods, algorithms, and processes.

This book provides a stepwise approach to building solutions to data science applications right from understanding the fundamentals, performing data analytics to writing source code. We have divided this book into five chapters, where the first chapter explains the basics of data science and its benefits, data preparations and detail about the data mining & data warehousing concepts. Second chapter and third chapter discusses about Normal Distributions and standards, Regression techniques. Fourth chapter explains python libraries. Fifth chapter describes data visualization concepts.

The main aim of this book is to make the students to understand the concepts easily. This book makes the understanding of subject in a clear way and makes it more interesting.

Dr. H.Lilly Beulah

Dr.R.Reka

Ms. M.Gayathri

Ms.V.Dhanakodi




PRINCIPAL
Mahendra College of Engineering,
Mahendra Salem Campus,
Minnampalli, SALEM 636 106
TAMILNADU

FOUNDATIONS OF DATA SCIENCE - (CS3353)

SYLLABUS

UNIT I INTRODUCTION

Data Science: Benefits and uses – facets of data - Data Science Process: Overview – Defining research goals – Retrieving data – Data preparation - Exploratory Data analysis – build the model– presenting findings and building applications - Data Mining - Data Warehousing – Basic Statistical descriptions of Data. (Chapter - 1)

UNIT II DESCRIBING DATA

Types of Data - Types of Variables -Describing Data with Tables and Graphs –Describing Data with Averages - Describing Variability - Normal Distributions and Standard (z) Scores. (Chapter - 2)

UNIT III DESCRIBING RELATIONSHIPS

Correlation –Scatter plots –correlation coefficient for quantitative data –computational formula for correlation coefficient – Regression –regression line –least squares regression line – Standard error of estimate – interpretation of r^2 –multiple regression equations – regression towards the mean. (Chapter - 3)

UNIT IV PYTHON LIBRARIES FOR DATA WRANGLING

Basics of Numpy arrays –aggregations –computations on arrays –comparisons, masks, boolean logic – fancy indexing – structured arrays – Data manipulation with Pandas – data indexing and selection – operating on data – missing data – Hierarchical indexing – combining datasets – aggregation and grouping – pivot tables. (Chapter - 4)

UNIT V DATA VISUALIZATION

Importing Matplotlib – Line plots – Scatter plots – visualizing errors – density and contour plots – Histograms – legends – colors – subplots – text and annotation – customization – three dimensional plotting - Geographic Data with Basemap - Visualization with Seaborn. (Chapter - 5)



TABLE OF CONTENTS

UNIT I

Chapter 1: Introduction


	Page No.
Data Science	1
Benefits and uses	1
Facets of data	2
Data Science Process: Overview	6
Defining research goals	7
Retrieving data	9
Data preparation	11
Exploratory Data analysis	21
Build the model	25
Presenting findings and building applications	32
Data Mining	33
Data Warehousing	36
Basic Statistical descriptions of Data.	40

UNIT II

Chapter 2: Describing Data

Types of Data	47
Types of Variables	60
Describing Data with Tables and Graphs	61
Describing Data with Averages	74
Describing Variability	79




PRINCIPAL
Mahendra College of Engineering
Mahendra Salem Campus,
Minnampalli, SALEM 636 106
TAMILNADU

Normal Distributions and Standard (z) Scores	84
--	----

UNIT III

Chapter 3: Describing Relationships

Correlation	94
Scatter plots	102
Correlation coefficient for quantitative data	107
Computational formula for correlation coefficient	112
Regression	118
Regression line	120
Least squares regression line	123
Standard error of estimate	127
Interpretation of r^2	130
Multiple regression equations	132
Regression towards the mean	134

UNIT IV

Chapter 4: Python Libraries For Data Wrangling

Basics of Numpy arrays	140
Aggregations	149
Comparisons, masks, boolean logic	154
Fancy indexing	166
Structured arrays	172
Data manipulation with Pandas	180
Data indexing and selection	190
Operating on data	198
Missing data	199




PRINCIPAL
Mahendra College of Engineering
Mahendra Salem Campus,
Minnampalli, SALEM 636 106
TAMILNADU

Hierarchical indexing	203
Combining datasets	209
Aggregation and grouping	218
Pivot tables	230

UNIT V

Chapter 5: Data Visualization

Importing Matplotlib	240
Line plots	244
Scatter plots	251
Visualizing errors	255
Density and contour plots	258
Histograms	261
Legends	265
Colors	269
Subplots	275
Text and annotation	280
Three dimensional plotting	285
Geographic Data with Basemap	292
Visualization with Seaborn	303




PRINCIPAL
 Mahendra College of Engineering
 Mahendra Salem Campus,
 Minnampalli, SALEM 636 106
 TAMILNADU

Foundations of Data Science

Dr. Lilly Beulah H, Dr. Reka R
Ms. Gayathri M & Ms. Dhanakodi V

ABOUT THE AUTHORS



Dr. H. LILLY BEULAH

Professor & Head of the Department,
Department of Computer Science and Engineering,
Mahendra College of Engineering,
Salem, Tamilnadu, INDIA.



Dr. R. REKA

Associate Professor,
Department of Computer Science and Engineering,
Mahendra College of Engineering,
Salem, Tamilnadu, INDIA.



Ms. M. GAYATHRI

Assistant Professor,
Department of Computer Science and Engineering,
Mahendra College of Engineering,
Salem, Tamilnadu, INDIA.



Ms. V. DHANAKODI

Assistant Professor,
Department of Computer Science and Engineering,
Mahendra College of Engineering,
Salem, Tamilnadu, INDIA.



M. Jee
PRINCIPAL

Mahendra College of Engineering
Mahendra Salem Campus,
Minnampalli, SALEM
TAMILNADU

ISBN 978-81-957838-4-7

