Foundations of Data Science

R-2021 Anna University

CS3353

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

FIRST EDITION

Published & Printed in India.

ISBN: 978-81-957838-4-7

PRICE: INR 400

First Edition: October 2022

For information contact:

Email: whitecoastbooks@gmail.com

Website: http://www.whitecoastbooks.com

Dedicated

to

Our Family, Friends

&

Students

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 Beaulah

Dr.R.Reka

Ms. M.Gayathri

Ms.V.Dhanakodi

Mahendra College of Engineering
SALEM
636 106
Mahendra Salem Campus

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 Beaulah

Dr.R.Reka

Ms. M.Gayathri

Ms.V.Dhanakodi

Mahendra College of Engineering
Mahendra Salem Campus
SALEM
636 106

PRINCIPAL
Mahendra College of Engineerin
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 r2 –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.

UNIT II

Chapter 2: Describing Data

Mahendra College of Engineering • Mahendra Salem Campus • SALEM 636 106

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

UNIT III

Chapter 3: Describing Relationships

UNIT IV

Chapter 4: Python Libraries For Data Wrangling

UNIT V

Chapter 5: Data Visualization

Foundations of Data Science

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

ABOUT THE AUTHORS

Dr. H. LILLY BEAULAH
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.

ISBN 978-81-957838-4-7