

CONTENTS

UNIT 1: INTRODUCTION	1.1 – 1.52
1.1. Data Science Introduction	1.1
1.2. Benefits and Uses of Data Science and Big Data	1.1
1.3. Facets of data	1.2
1.3.1. Structured data	1.3
1.3.2. Unstructured data	1.3
1.3.3. Natural language	1.3
1.3.4. Machine-generated data	1.4
1.3.5. Graph-based or network data	1.5
1.3.6. Audio, Image, and Video	1.5
1.3.7. Streaming data	1.6
1.4. The Data Science Process	1.6
1.4.1. Setting the research goal	1.7
1.4.2. Retrieving data	1.7
1.4.3. Data preparation	1.7
1.4.4. Data exploration	1.7
1.4.5. Data modeling or model building	1.7
1.4.6. Presentation and automation	1.7
1.5. Overview of the Data Science Process	1.7
1.5.1. Don't be a slave to the process	1.9
1.6. Step 1: Defining research goals and creating A project charter	1.10
1.6.1. Spend time understanding the goals and context of the research	1.10
1.6.2. Create a project charter	1.10
1.7. Step 2: Retrieving data	1.11
1.7.1. Start with data stored within the company	1.11
1.7.2. Don't be afraid to shop around	1.12
1.7.3. Do data quality checks now to prevent problems later	1.12
1.8. Step 3: Cleansing, integrating, and transforming data	1.13

1.8.1.	Cleansing data.....	1.13
1.8.2.	Correct Errors as early as possible.....	1.18
1.8.3.	Combining data from different data sources.....	1.18
1.8.4.	Transforming data.....	1.21
1.9.	Step 4: Exploratory data analysis.....	1.23
1.10.	Step 5: Build the models.....	1.28
1.10.1.	Model and variable selection.....	1.29
1.10.2.	Model execution.....	1.29
1.10.3.	Model diagnostics and model comparison.....	1.33
1.11.	Step 6: Presenting findings and building applications on top of them.....	1.35
1.12.	Data Mining.....	1.35
1.12.1.	Features of Data mining.....	1.35
1.12.2.	Types of Data Mining.....	1.36
1.13.	Text Mining.....	1.38
1.13.1.	The five fundamental steps involved in text mining are:.....	1.38
1.14.	Basic Statistical Description of Data.....	1.41
1.14.1.	Measuring the Central Tendency.....	1.41
	Two Marks Questions and Answers.....	1.48
	Review Questions.....	1.52

UNIT 2: DESCRIBING DATA..... 2.1 – 2.62

2.1.	Types of Data.....	2.1
2.1.1.	Levels of Measurement.....	2.2
2.2.	Types of Variables.....	2.4
2.2.1.	Discrete And Continuous Variables.....	2.5
2.2.2.	Approximate Numbers.....	2.5
2.2.3.	Independent And Dependent Variables.....	2.5
2.2.4.	Confounding Variable.....	2.6
2.3.	Describing Data with Tables and Graphs.....	2.6
2.3.1.	Tables (Frequency Distributions).....	2.6
2.3.2.	Graphs For Quantitative Data.....	2.17

2.4. Describing Data with Averages.....	2.24
2.4.1. Mode.....	2.24
2.4.2. Median.....	2.25
2.4.3. Mean.....	2.26
2.4.4. Which Average?	2.28
2.4.5. Averages for Qualitative And Ranked Data	2.31
2.5. Describing Variability.....	2.32
2.5.1. Intuitive Approach.....	2.32
2.5.2. Range.....	2.33
2.5.3. Variance.....	2.33
2.5.4. Standard Deviation	2.34
2.5.5. Details: Standard Deviation.....	2.36
2.5.6. Degrees of Freedom (Df).....	2.42
2.5.7. Interquartile Range (IQR).....	2.43
2.5.8. Measures of Variability for Qualitative and Ranked Data.....	2.44
2.6. Normal Distributions and Standard (z) Scores	2.44
2.6.1. The Normal Curve.....	2.45
2.6.2. z Scores.....	2.47
2.6.3. Standard Normal Curve	2.48
2.6.4. Solving Normal Curve Problems.....	2.50
2.6.5. Finding Proportions	2.51
2.6.6. Finding Scores.....	2.54
2.6.7. More about z Scores	2.57
Two Marks Questions and Answers	2.58
Review Questions.....	2.62

UNIT 3: DESCRIBING RELATIONSHIPS 3.1 – 3.28

3.1. Correlation.....	3.1
3.1.1. An Intuitive Approach.....	3.1
3.2. Scatterplots	3.3
3.2.1. Construction.....	3.3
3.2.2. Positive, Negative, or Little or No Relationship?.....	3.4

3.2.3.	Strong or Weak Relationship?	3.5
3.2.4.	Perfect Relationship	3.5
3.2.5.	Curvilinear Relationship	3.5
3.3.	A Correlation Coefficient for Quantitative Data : r	3.6
3.3.1.	Key Properties of r	3.6
3.3.2.	Sign of r	3.6
3.3.3.	Numerical Value of r	3.6
3.3.4.	Interpretation of r	3.7
3.3.5.	r Is Independent of Units of Measurement	3.7
3.3.6.	Range Restrictions	3.7
3.3.7.	Verbal Descriptions	3.8
3.3.8.	Correlation Not Necessarily Cause-Effect	3.8
3.4.	Details: Computation Formula for r	3.9
3.5.	Regression	3.11
3.5.1.	Two Rough Predictions	3.11
3.6.	A Regression Line	3.12
3.6.1.	Placement of Line	3.13
3.7.	Least Squares Regression Line	3.14
3.7.1.	Least Squares Regression Equation	3.14
3.7.2.	Finding Values of b and a	3.14
3.7.3.	Key Property	3.16
3.7.4.	Solving for Y'	3.16
3.8.	Standard Error of Estimate, $S_{y x}$	3.17
3.8.1.	Finding the Standard Error of Estimate	3.17
3.8.2.	Key Property	3.17
3.8.3.	Importance of r Standard Error of Estimate	3.18
3.9.	Interpretation of r^2	3.19
3.9.1.	Repetitive Prediction of the Mean	3.19
3.9.2.	Predictive Errors	3.19
3.9.3.	Error Variability (Sum of Squares)	3.20
3.9.4.	Proportion of Predicted Variability	3.21

3.9.5.	Squared Correlation Coefficient (r^2).....	3.22
3.9.6.	r^2 Does Not Apply to Individual Scores.....	3.22
3.9.7.	Small Values of r^2	3.22
3.9.8.	r^2 Doesn't Ensure Cause-Effect	3.23
3.10.	Multiple Regression Equations	3.23
3.10.1.	Common Features.....	3.23
3.11.	Regression Toward the Mean.....	3.24
3.11.1.	Appears in Many Distributions.....	3.24
3.11.2.	The Regression Fallacy.....	3.25
3.11.3.	Avoiding the Regression Fallacy.....	3.25
Two Marks Questions and Answers		3.26
Review Questions.....		3.27

UNIT 4: PYTHON LIBRARIES FOR DATA WRANGLING . 4.1 – 4.120

4.1.	Understanding Data Types in Python	4.1
4.1.1.	A Python Integer is more than just an integer	4.2
4.1.2.	A Python List is more than just a list.....	4.3
4.1.3.	Fixed-Type Arrays in Python	4.4
4.1.4.	Creating Arrays from Python Lists.....	4.4
4.1.5.	Creating Arrays from Scratch.....	4.5
4.1.6.	NumPy Standard Data Types.....	4.6
4.2.	The Basics of NumPy Arrays	4.8
4.2.1.	NumPy Array Attributes.....	4.8
4.2.2.	Array Indexing: Accessing Single Elements	4.9
4.2.3.	Array Slicing: Accessing Subarrays	4.10
4.2.4.	Reshaping of Arrays	4.13
4.2.5.	Concatenation of arrays	4.14
4.2.6.	Splitting of arrays	4.15
4.3.	Aggregations: Min, Max, and Everything in Between.....	4.17
4.3.1.	Summing the Values in an Array.....	4.17
4.3.2.	Minimum and Maximum	4.17
4.4.	Computation on Arrays: Broadcasting.....	4.19

4.4.1.	Introducing Broadcasting.....	4.20
4.4.2.	Rules of Broadcasting.....	4.21
4.4.3.	Broadcasting in Practice	4.24
4.5.	Comparisons, Masks, and Boolean Logic	4.25
4.5.1.	Example: Counting Rainy Days	4.25
4.5.2.	Comparison Operators as ufuncs	4.27
4.5.3.	Working with Boolean Arrays.....	4.28
4.5.4.	Boolean Arrays as Masks	4.30
4.6.	Fancy Indexing	4.31
4.6.1.	Exploring Fancy Indexing	4.32
4.6.2.	Combined Indexing	4.33
4.6.3.	Example: Selecting Random Points.....	4.34
4.6.4.	Modifying Values with Fancy Indexing.....	4.35
4.6.5.	Example: Binning Data.....	4.36
4.7.	Structured Data: NumPy's Structured Arrays	4.38
4.7.1.	Creating Structured Arrays.....	4.39
4.7.2.	More Advanced Compound Types	4.40
4.7.3.	RecordArrays: Structured Arrays with a Twist	4.40
4.8.	Data Manipulation with Pandas	4.41
4.8.1.	Installing and Using Pandas.....	4.41
4.8.2.	Introducing Pandas Objects	4.41
4.9.	Data Indexing and Selection.....	4.50
4.9.1.	Data Selection in Series	4.50
4.9.2.	Data Selection in DataFrame	4.53
4.10.	Operating on Data in Pandas	4.58
4.10.1.	Ufuncs: Index Preservation.....	4.58
4.10.2.	UFuncs: Index Alignment.....	4.59
4.10.3.	Ufuncs: Operations Between DataFrame and Series	4.62
4.11.	Handling Missing Data	4.63
4.11.1.	Trade-Offs in Missing Data Conventions.....	4.63
4.11.2.	Missing Data in Pandas	4.64

4.11.3. Operating on Null Values	4.67
4.12. Hierarchical Indexing	4.71
4.12.1. A Multiply Indexed Series.....	4.71
4.12.2. Methods of MultiIndex Creation	4.74
4.12.3. Indexing and Slicing a MultiIndex	4.78
4.12.4. Rearranging Multi-Indices.....	4.81
4.12.5. Data Aggregations on Multi-Indices	4.84
4.13. Combining Datasets: Concat and Append	4.85
4.13.1. Recall: Concatenation of NumPy Arrays.....	4.86
4.13.2. Simple Concatenation with pd.concat	4.86
4.14. Combining Datasets: Merge and Join	4.90
4.14.1. Relational Algebra.....	4.90
4.14.2. Categories of Joins.....	4.91
4.14.3. Specification of the Merge Key	4.93
4.14.4. Specifying Set Arithmetic for Joins.....	4.97
4.14.5. Overlapping Column Names; The suffixes Keyword.....	4.98
4.15. Aggregation and Grouping.....	4.99
4.15.1. Planets Data	4.99
4.15.2. Simple Aggregation in Pandas.....	4.99
4.15.3. GroupBy: Split, Apply, Combine	4.102
4.16. Pivot Tables	4.111
4.16.1. Motivating Pivot Tables	4.111
4.16.2. Pivot Tables by Hand.....	4.112
4.16.3. Pivot Table Syntax.....	4.113
Two Marks Questions and Answers	4.115
Review Questions.....	4.120

UNIT 5: DATA VISUALIZATION 5.1-5.96

5.1. General Matplotlib Tips	5.1
5.1.1. Importing matplotlib.....	5.1
5.1.2. Setting Styles	5.1

5.1.3.	Show() or No show()? How to Display Your Plots	5.1
5.1.4.	Saving Figures to File	5.3
5.2.	Two Interfaces for the Price of One	5.4
5.2.1.	MATLAB-style interface	5.4
5.2.2.	Object-oriented interface	5.5
5.3.	Simple Line Plots	5.6
5.3.1.	Adjusting the Plot: Line Colors and Styles	5.8
5.4.	Simple Scatter Plots	5.14
5.4.1.	Scatter Plots with plt.plot	5.15
5.4.2.	Scatter Plots with plt.scatter	5.17
5.4.3.	plot Versus scatter: A Note on Efficiency	5.19
5.5.	Visualizing Errors	5.19
5.5.1.	Basic Errorbars	5.20
5.5.2.	Continuous Errors	5.21
5.6.	Density and Contour Plots	5.23
5.6.1.	Visualizing a Three-Dimensional Function	5.23
5.7.	Histograms, Binnings, and Density	5.27
5.7.1.	Two-Dimensional Histograms and Binnings	5.29
5.8.	Customizing Plot Legends	5.31
5.8.1.	Choosing Elements for the Legend	5.32
5.8.2.	Legend for Size of Points	5.33
5.8.3.	Multiple Legends	5.35
5.9.	Customizing Colorbars	5.36
5.9.1.	Customizing Colorbars	5.36
5.10.	Multiple Subplots	5.41
5.10.1.	plt.axes: Subplots by Hand	5.41
5.10.2.	plt.subplot: Simple Grids of Subplots	5.42
5.10.3.	plt.subplots: The Whole Grid in One Go	5.43
5.10.4.	plt.GridSpec: More Complicated Arrangements	5.45
5.11.	Text and Annotation	5.47
5.11.1.	Example: Effect of Holidays on US Births	5.47

5.11.2. Transforms and Text Position	5.49
5.11.3. Arrows and Annotation	5.51
5.12. Customizing Ticks.....	5.54
5.12.1. Major and Minor Ticks.....	5.55
5.12.2. Hiding Ticks or Labels	5.56
5.12.3. Reducing or Increasing the Number of Ticks.....	5.57
5.12.4. Fancy Tick Formats.....	5.59
5.12.5. Summary of Formatters and Locators	5.61
5.13. Customizing Matplotlib: Configurations and Stylesheets.....	5.62
5.13.1. Plot Customization by Hand.....	5.62
5.13.2. Changing the Defaults: rcParams	5.64
5.13.3. Stylesheets	5.65
5.14. Three-Dimensional Plotting in Matplotlib.....	5.68
5.14.1. Three-Dimensional Points and Lines	5.69
5.14.2. Three-Dimensional Contour Plots.....	5.70
5.14.3. Wireframes and Surface Plots	5.71
5.14.4. Surface Triangulations.....	5.72
5.15. Geographic Data with Basemap	5.74
5.15.1. Map Projections.....	5.76
5.15.2. Drawing a Map Background	5.79
5.15.3. Plotting Data on Maps	5.81
5.15.4. Example: California Cities	5.82
5.15.5. Example: Surface Temperature Data.....	5.83
5.16. Visualization with Seaborn	5.85
5.16.1. Seaborn Versus Matplotlib.....	5.85
5.16.2. Exploring Seaborn Plots.....	5.87
Two Marks Questions and Answers.....	5.93
Review Questions	5.96

DATA SCIENCE LABORATORY L.1 – L.93

List of Experiments – Details L.1

Solved Anna University Question Papers S.Q.1 – S.Q.16

Model Anna University Question Papers..... M.Q.1 – M.Q.8