



Chapter 1 : Introduction to Data Science and Big Data		1-1 to 1-26
1.1	Introduction and Big Data Overview.....	1-1
1.1.1	The Five Vs (Characteristics) of Big Data (Data Explosion).....	1-2
1.1.2	Major Applications of Big Data.....	1-3
1.1.3	Data Formats (Types).....	1-4
1.1.4	Comparison between Data Formats.....	1-5
1.1.5	Data Collection.....	1-5
1.1.6	DIKW Pyramid.....	1-5
1.1.7	Categories of Data Analytics.....	1-6
1.1.8	Comparison between Categories of Data Analytics.....	1-7
1.2	State of the Practice in Analytics.....	1-8
1.2.1	Business Intelligence (BI) vs Data Science.....	1-8
1.3	Relationship between Data Science and Information Science.....	1-8
1.3.1	Current Analytical Architecture.....	1-9
1.3.2	Drivers (Motivation) of Big Data Analytics.....	1-10
1.3.3	Emerging Big Data Ecosystem and New Approach.....	1-11
1.3.4	Key Roles in the New Big Data Ecosystem.....	1-13
1.3.5	Key Roles for a Successful Analytics Project.....	1-13
1.4	Data Analytics Life Cycle (Data Science Life Cycle).....	1-14
1.4.1	Phase 1 : Discovery.....	1-15
1.4.2	Phase 2 : Data Preparation.....	1-15
1.4.3	Phase 3 : Model Planning.....	1-15
1.4.4	Phase 4 : Model Building.....	1-16
1.4.5	Phase 5 : Communicate Results.....	1-16
1.4.6	Phase 6 : Operationalise.....	1-17
1.5	Data Wrangling.....	1-17
1.5.1	Need for Data Wrangling.....	1-17
1.5.1(A)	Data.....	1-17
1.5.1(B)	Tasks.....	1-18
1.5.1(C)	Models.....	1-18
1.5.1(D)	Features.....	1-18
1.5.1(E)	Feature Engineering.....	1-19
1.5.1(F)	Data Engineering -vs- Feature Engineering.....	1-20
1.6	Data Wrangling Methods.....	1-21
Chapter 2 : Statistical Inference		2-1 to 2-60
2.1	Need of Statistics in Data Science and Big Data Analytics.....	2-1
2.1.1	Sampling Distributions.....	2-2
2.1.2	General Statistics.....	2-2
2.1.2(A)	Mean.....	2-2
2.1.2(B)	Median.....	2-3
2.1.2(C)	Mode.....	2-3
2.1.2(D)	Mid-range.....	2-3
2.1.2(E)	Range.....	2-3
2.1.3	Standard Deviation.....	2-4
2.1.4	Variance.....	2-4
2.1.5	Covariance.....	2-5



2.1.6	Mean Absolute Deviation	2-6
2.2	Concepts of Probability	2-6
2.2.1	Fundamental Rules of Probability	2-7
2.3	Importance of Bayesian Methods.....	2-8
2.4	Bayes' Algorithm (Theorem)	2-8
2.5	Bayes' Theorem and Concept Learning	2-10
2.6	Hypothesis Testing	2-11
2.6.1	How Hypothesis Testing Works?.....	2-11
2.6.2	Difference of Means	2-12
2.6.3	Type I and Type II Errors	2-12
2.6.3(A)	Comparison between Type I and Type II errors.....	2-14
2.6.4	Power and Sample Size	2-14
2.6.5	Pearson Correlation Coefficient (PCC)	2-15
2.6.5(A)	How to Calculate Pearson Correlation Coefficient (PCC)?.....	2-16
2.6.5(B)	How is Pearson Correlation Coefficient (PCC)used in Hypothesis Testing?.....	2-17
2.6.6	Contingency Table	2-18
2.6.7	Degrees of Freedom.....	2-19
2.6.8	Chi-Squared Tests	2-19
2.6.8(A)	Chi-square Distribution	2-20
2.6.8(B)	Chi-square Goodness of Fit Test.....	2-22
2.7	T-tests (Student's t-test)	2-32
2.7.1	The t-Distribution	2-33
2.7.2	One-tailed vs. Two-tailed Tests.....	2-36
2.7.3	The One-Sample t-Test	2-36
2.7.4	The Paired t-Test.....	2-41
2.7.5	The Two-Sample t-Test.....	2-46

Chapter 3 : Big Data Analytics Life Cycle**3-1 to 3-5**

3.1	Introduction to Big Data	1-1
3.2	Sources of Big Data	1-1
3.3	Data Analytics Lifecycle	1-5

Chapter 4 : Predictive Big Data Analytics with Python**4-1 to 4-87**

4.1	Data Quality and Remediation (Data Pre-Processing)	4-2
4.1.1	Common Data Quality Issues.....	4-2
4.1.2	Remediating (Fixing) Data Quality Issues.....	4-3
4.2	Analytics Types	4-5
4.3	Association Rules	4-5
4.3.1	Key Terms and Properties of Association Rules.....	4-6
4.3.1(A)	Itemset.....	4-6
4.3.1(B)	Support.....	4-7
4.3.1(C)	Confidence.....	4-8
4.3.1(D)	Lift.....	4-8
4.3.1(E)	Leverage	4-10
4.3.1(F)	Subsets.....	4-11
4.3.2	Evaluation of Candidate Rules	4-12
4.3.3	Apriori Algorithm.....	4-12
4.3.3(A)	How Apriori Algorithm Works	4-12



4.3.3(B)	Case Study – Transactions in Grocery Store	4-16
4.3.3(C)	Validation and Testing.....	4-20
4.3.3(D)	Diagnostics	4-20
4.4	Frequent-Pattern (FP) Growth.....	4-21
4.5	Regression Analysis.....	4-26
4.5.1	Linear Regression	4-26
4.5.1(A)	Use Cases (or Applications of) for Linear Regression.....	4-32
4.5.2	Logistic Regression	4-32
4.5.2(A)	Use Cases (or Applications of) for Logistic Regression.....	4-33
4.6	Reasons to Choose and Cautions.....	4-34
4.7	Additional Regression Models.....	4-34
4.8	Classification Models.....	4-34
4.8.1	Decision Trees	4-35
4.8.2	Key Terms and Concepts.....	4-36
4.8.2(A)	Entropy	4-36
4.8.2(B)	Information Gain	4-38
4.8.2(C)	Gain Ratio.....	4-41
4.8.2(D)	Gini Index.....	4-43
4.9	Decision Tree Algorithms.....	4-56
4.9.1	The General Algorithm	4-56
4.9.2	ID3 Algorithm	4-57
4.9.3	C4.5 Algorithm.....	4-60
4.9.4	CART Algorithm.....	4-60
4.9.5	Evaluating a Decision Tree.....	4-61
4.10	Naïve Bayes (Classification by Bayesian Belief Networks).....	4-61
4.10.1	Naïve Bayes Classifier	4-61
4.10.2	Smoothing.....	4-69
4.10.3	Advantages of Naïve Bayes Classifier.....	4-70
4.10.4	Disadvantages of Naïve Bayes Classifier	4-70
4.11	Diagnostics (Evaluation Measures) of Classifiers	4-71
4.11.1	ROC Curve.....	4-73
4.11.2	Area Under the Curve (AUC)	4-74
4.12	Additional Classification Methods	4-74
4.12.1	Bagging.....	4-74
4.12.2	Boosting.....	4-76
4.12.3	Random Forests.....	4-78
4.13	Essential Python Libraries.....	4-78

Chapter 5 : Big Data Analytics and Model Evaluation
5-1 to 5-43

5.1	Clustering.....	5-1
5.1.1	Properties of a Cluster	5-3
5.1.2	Types of Clustering.....	5-3
5.1.3	Use Cases (Applications) of Clustering.....	5-3
5.1.4	K-means	5-4
5.1.5	Determining the Number of Clusters (Elbow Plot)	5-10
5.1.6	Diagnostics (Performance Measures).....	5-12
5.1.7	Reasons to Choose and Cautions (Drawbacks / Challenges)	5-13



5.2	k-Nearest Neighbours (kNN) Classification Algorithm	5-14
5.3	Hierarchical Clustering.....	5-15
5.3.1	Dendrogram.....	5-15
5.3.2	Hierarchical Clustering Strategies (Algorithms).....	5-16
5.3.2(A)	Agglomerative Hierarchical Clustering	5-16
5.3.2(B)	Divisive Hierarchical Clustering	5-17
5.3.3	Agglomeration (Linkage) Methods.....	5-18
5.4	Time Series Analysis.....	5-19
5.4.1	Goals of Time Series Analysis.....	5-19
5.4.2	Applications of Time Series Analysis	5-20
5.4.3	Characteristics (Components) of Time Series Analysis	5-20
5.5	Modelling Time Series Data	5-22
5.5.1	Time-Domain Versus Frequency Domain Models	5-22
5.5.2	Univariate Versus Multivariate Time Series Models	5-22
5.5.3	Box-Jenkins Methodology	5-23
5.5.4	Autoregressive Integrated Moving Average (ARIMA)	5-23
5.6	Introduction to Text Analysis.....	5-24
5.6.1	Challenges in Text Analysis.....	5-24
5.6.2	Steps in Text Analysis	5-25
5.6.3	Text Pre-Processing Techniques	5-26
5.6.4	Bag-of-Words.....	5-27
5.6.5	Bag-of-n-Grams.....	5-28
5.7	Term Frequency - Inverse Document Frequency (TFIDF)	5-28
5.7.1	Term Frequency (TF).....	5-29
5.7.2	Inverse Document Frequency (IDF)	5-29
5.8	Social Network Analysis (SNA)	5-31
5.8.1	Need (Applications) of Social Network Analysis	5-32
5.9	Introduction to Business Analysis	5-33
5.9.1	Skillsset and Expertise Needed for the Business Analysis Role.....	5-34
5.9.2	Dealing with Key Stakeholders.....	5-34
5.10	Model Evaluation and Selection.....	5-35
5.10.1	Metrics for Evaluating Classifier Performance.....	5-35
5.10.1(A)	Cross-Validation	5-35
5.10.1(B)	Holdout Method	5-35
5.10.1(C)	k-Fold Cross-Validation.....	5-35
5.10.1(D)	Leave-P-Out Cross-Validation (LpOCV)	5-36
5.10.1(E)	Sub-Sampling	5-37
5.10.2	Hyperparameter Tuning Techniques.....	5-37
5.10.2(A)	What Do Hyperparameters Do?.....	5-37
5.10.2(B)	How is Hyperparameter Tuning Carried out?	5-38
5.10.2(C)	Hyperparameter Tuning Algorithms.....	5-38
5.11	Clustering and Time-series Analysis using Scikit-learn	5-40



Chapter 6 : Data Visualization and Hadoop	6-1 to 6-42
6.1 Introduction to Data Visualisation.....	6-1
6.1.1 Goals (Objectives) of Data Visualisation.....	6-2
6.2 Challenges (Difficulties) with Big Data Visualisation	6-3
6.3 Techniques for Visual Data Representations	6-4
6.3.1 Conventional Data Visualisation Tools.....	6-4
6.3.2 Types of Data Visualisation	6-4
6.3.2(A) Comparative Plots.....	6-5
6.3.2(B) Statistical Plots	6-9
6.3.2(C) Topology Plots	6-12
6.3.2(D) Spatial Plots	6-15
6.4 Data Visualisation Taxonomy	6-17
6.5 Visualizing Big Data	6-19
6.6 General Workflow of Analytics and Visualisation.....	6-21
6.7 Tools Used in Data Visualisation	6-21
6.7.1 Tableau.....	6-22
6.7.2 Microsoft Power BI.....	6-23
6.7.3 Qlik	6-23
6.7.4 ThoughtSpot	6-24
6.8 Analytical Techniques Used in Big Data Visualisation.....	6-25
6.9 Analytics for Unstructured Data	6-25
6.9.1 Use Cases for Analytics for Unstructured Data	6-25
6.9.2 Apache Hadoop	6-27
6.9.2(A) MapReduce.....	6-28
6.9.2 (B) Hadoop Distributed File System (HDFS)	6-31
6.9.2(C) YARN (Yet Another Resource Negotiator).....	6-32
6.10 The Hadoop Ecosystem.....	6-33
6.10.1 HBase	6-34
6.10.1(A) Characteristics and Features of HBase	6-34
6.10.1(B) Architecture of HBase.....	6-34
6.10.2 Pig.....	6-35
6.10.2(A) Characteristics and Features of Pig.....	6-36
6.10.2(B) Architecture of Pig.....	6-36
6.10.3 Hive.....	6-37
6.10.3(A) Characteristics and Features of Hive.....	6-37
6.10.3(B) Architecture of Hive.....	6-38
6.10.4 Mahout.....	6-39
6.10.4(A) Characteristics and Features of Mahout.....	6-39
6.11 Data Visualization using Python	6-41



Chapter 7 : Python Related Topics		7-1 to 7-73
7.1	Basic Python Programming.....	7-1
7.1.1	Installation.....	7-1
7.1.2	Interactive Python.....	7-2
7.2	Variables and Data Types.....	7-5
7.2.1	Writing Comments in Your Python Programs.....	7-7
7.2.2	Python Indentation.....	7-7
7.3	Flow Control Structures.....	7-7
7.3.1	Conditional Statements.....	7-7
7.3.2	Loops.....	7-10
7.3.2(A)	while Loop.....	7-10
7.3.2(B)	for Loop.....	7-11
7.4	Functions.....	7-12
7.4.1	Defining a Function.....	7-12
7.4.2	Python Modules (Importing a Function).....	7-12
7.4.3	Time Functions.....	7-14
7.4.4	Library Functions.....	7-15
7.5	Basic Arithmetic Programs.....	7-16
7.6	Essential Python Libraries.....	7-17
7.6.1	NumPy.....	7-17
7.6.2	Creating Matrices.....	7-19
7.6.3	Save and Load NumPy Objects.....	7-20
7.6.4	Matplotlib.....	7-20
7.6.5	Line Plot.....	7-25
7.6.5(A)	Scatter Plot.....	7-25
7.6.5(B)	Histogram.....	7-26
7.6.6	Box Plot.....	7-29
7.6.7	Seaborn.....	7-30
7.6.8	Density Plot.....	7-31
7.6.9	SciPy.....	7-33
7.6.10	Pandas.....	7-35
7.7	Linear Regression.....	7-52
7.8	Logistic Regression.....	7-54
7.9	Naïve-Bayes Classification.....	7-58
7.10	Decision Trees.....	7-64
7.11	Clustering.....	7-67