Savitribai Phule Pune University
Fourth Year of Artificial Intelligence and Data Science (2020 Course)
417522: Data Modeling and Visualization

Teaching Scheme:	Credit	Examination Scheme:
TH: 03 Hours/Week	03	In-Sem (Paper): 30 Marks End-Sem (Paper): 70 Marks

Prerequisites Courses: Statistics (217528), Computer Graphics (210244), Database Management System (310241)

Course Objectives:

- Creating an emerging data model for the data to be stored in a database
- Conceptualized representation of Data objects
- Create associations between different data objects, and the rules
- Organize data description, data semantics, and consistency constraints of data
- Identifying data trends
- Incorporate data visualization tools and reap transformative benefits in their critical areas of operations

Course Outcomes:

After completion of the course, learners should be able to-

CO1: Summarize data analysis and visualization in the field of exploratory data science
CO2: Analyze the characteristics and requirements of data and select an appropriate data model
CO3: Describe to load, clean, transform, merge and reshape data
CO4: Design a probabilistic data modeling, interpretation, and analysis
CO5: Evaluate time series data
CO6: Integrate real world data analysis problems

Course Contents

Unit I	Introduction to Data Modelling	07 Hours

Basic probability:
Discrete and continuous random variables, independence, covariance, central limit theorem, Chebyshev inequality, diverse continuous and discrete distributions.
Statistics, Parameter Estimation, and Fitting a Distribution: Descriptive statistics, graphical statistics, method of moments, maximum likelihood estimation
Data Modeling Concepts • Understand and model subtypes and supertypes • Understand and model hierarchical data • Understand and model recursive relationships • Understand and model historical data

#Exemplar/Case Studies	Case study of sampling for any real-world problem like exit poll statistics
***Mapping of Course Outcomes for Unit I**	CO1

Unit II	Testing and Data Modeling	07 Hours

Random Numbers and Simulation: Sampling of continuous distributions, Monte Carlo methods
Hypothesis Testing: Type I and II errors, rejection regions; Z-test, T-test, F-test, Chi-Square test, Bayesian test
Stochastic Processes and Data Modeling: Markov process, Hidden Markov Models, Poisson Process, Gaussian Processes, Auto-Regressive and Moving average processes, Bayesian Network, Regression, Queuing systems

#Exemplar/Case Studies	Hypothesis Testing for examples like: Dieters lose more fat than the exercisers, New medicine testing
***Mapping of Course Outcomes for Unit II**	CO2

Unit III	**Basics of Data Visualization**	**07 Hours**

Computational Statistics and Data Visualization, Types of Data Visualization, Presentation and Exploratory Graphics, Graphics and Computing, Statistical Historiography, Scientific
Design Choices in Data Visualization, Higher-dimensional Displays and Special Structures,
Static Graphics: Complete Plots, Customization, Extensibility,
Other Issues: 3-D Plots, Speed, Output Formats, Data Handling

#Exemplar/Case Studies	Use IRIS dataset from Scikit and plot 2D-3D views of the dataset
***Mapping of Course Outcomes for Unit III**	CO3

Unit IV	**Data Visualization and Data Wrangling**	**07 Hours**

Data Wrangling: Hierarchical Indexing, Combining and Merging Data Sets Reshaping and Pivoting. Data Visualization matplotlib: Basics of matplotlib, plotting with pandas and seaborn, other python visualization tools
Data Visualization Through Their Graph Representations: Data and Graphs Graph Layout Techniques, Force-directed Techniques Multidimensional Scaling, The Pulling Under Constraints Model, Bipartite Graphs

#Exemplar/Case Studies	Use data set of your choice from Open Data Portal (https://data.gov.in/) and apply data preprocessing methods
***Mapping of Course Outcomes for Unit IV**	CO4

Unit V	**Data Aggregation and Analysis**	**07 Hours**

Data Aggregation and Group operations: Group by Mechanics, Data aggregation, General split-apply-combine, Pivot tables and cross tabulation 67 Time Series
Data Analysis: Date and Time Data Types and Tools, Time series Basics, date Ranges, Frequencies and Shifting, Time Zone Handling, Periods and Periods Arithmetic, Resampling and Frequency conversion, Moving Window Functions.

#Exemplar/Case Studies	Study and analyse Weather records/economic indicator/ patient health evolution metrics
***Mapping of Course Outcomes for Unit V**	CO5

Unit VI	**Data Analysis of Visualization and Modelling**	**07 Hours**

Reconstruction, Visualization and Analysis of Medical Images
Introduction: - PET Images, Ultrasound Images, Magnetic Resonance Images, Conclusion and Discussion, Case Study: ER/Studio, Erwin data modeler, DbSchema Pro, Archi, SQL Database Modeler, LucidChart, Pgmodeler

#Exemplar/Case Studies	Creating logical data model for l utility company to implement data modeler
***Mapping of Course Outcomes for Unit VI**	CO6

Learning Resources

Text Books:

1. Chun-houh Chen Wolfgang Härdle Antony Unwin Editors Handbook of Data Visualization, Springer
2. Visualizing Data Ben Fry Beijing , Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.
3. Fundamentals of Data Visualization - A Primer on Making Informative and Compelling Figures , Clous O.Wilke , Published by O'Reilly Media, Inc.
4. Data Visualization - A Practical Introduction by Kieran Healy
5. McKinney, W.(2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy and IPython. 2nd edition. O'Reilly Media
6. Gelman, Andrew, and Jennifer Hill. Data Analysis Using Regression and Multilevel /Hierarchical Models. 1st ed. Cambridge, UK: Cambridge University Press, 2006. ISBN: 9780521867061.
7. Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. Bayesian Data Analysis. 2nd ed. New York, NY: Chapman & Hall, 2003. ISBN: 9781584883883

Reference Books:

1. Gelman, Andrew, and Jennifer Hill. Data Analysis Using Regression and Multilevel/Hierarchical Models. 1st ed. Cambridge, UK: Cambridge University Press, 2006. ISBN: 9780521867061
2. Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. Bayesian Data Analysis. 2nd ed. New York, NY: Chapman & Hall, 2003. ISBN: 9781584883883
3. David Dietrich, Barry Hiller, "Data Science and Big Data Analytics", EMC education services, Wiley publication, 2012, ISBN0-07-120413-X
4. Trent Hauk, "Scikit-learn Cookbook", Packt Publishing, ISBN: 9781787286382
5. Chirag Shah, "A Hands-On Introduction To Data Science", Cambridge University Press, (2020), ISBN: 978-1-108-47244-9
6. S.C. Gupta, V.K. Kapoor,"Fundamentals of Mathematics Statistics (A Modern Approach) " Sultan Chand & Sons Educational Publishers, Tenth revised edition , ISBM: 81-7014-791-3
7. Medhi "Statistical Methods: An Introductory Text", Second Edition, New Age International Ltd, ISBN:8122419577

e-Resources:

1. An Introduction to Statistical Learning by Gareth James https://www.ime.unicamp.br/~dias/Intoduction%20to%20Statistical%20Learning.pdf
2. Python Data Science Handbook by Jake VanderPlas https://tanthiamhuat.files.wordpress.com/2018/04/pythondatasciencehandbook.pdf
3. Elements of Statistical Learning: data mining, inference, and prediction, 2nd Edition. (su.domains)

MOOC Courses:

1. https://www.youtube.com/watch?v=WSNqcYqByFk
2. https://www.youtube.com/watch?v=eFByJkA3ti4
3. Computer Science and Engineering - NOC:Data Science for Engineers
4. Computer Science and Engineering - NOC:Python for Data Science
5. Introduction to Data Analytics: https://nptel.ac.in/courses/110106072

The CO-PO Mapping Matrix

CO/ PO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	2	3	2	2	-	-	-	-	-	-	-	1
CO2	3	2	2	2	3	3	-	-	-	-	-	1
CO3	3	3	1	2	2	2	-	-	-	-	-	2
CO4	2	2	2	2	3	2	-	-	-	-	-	2
CO5	1	3	2	3	2	-	-	-	-	-	-	2
CO6	-	2	2	2	3	-	-	-	-	-	-	2