Name of the Programme: MSc Integrated
Course Code: IMC - 601
Title of the Course: Introduction to Data Science
Number of Credits: 6(4L-0T-2P)
Effective from AY: 2022-23

Prerequisites for the course	Statistics and Probability theory and Python Programming
Objectives	To get started with basics of Data Science and learn all aspects of Data Science in its entirety

Content Theory:	**Unit-1: Basics of Data Science:** Introduction; Typology of problems-**Data science in a big data world:** Benefits and uses of data science and big data-Facets of data-The data science process-The big data ecosystem and data science-**The data science process:** Overview of the data science process- Defining research goals and creating a project charter-Retrieving data-Cleansing, integrating, and transforming data-Exploratory data analysis-Build the models- Presenting findings and building applications on top of them. **Unit -2: Mathematics for Data science (Revision):** • Importance of linear algebra, statistics and optimization from a data science perspective; Structured thinking for solving data science problems. • Linear Algebra: Matrices and their properties (determinants, traces, rank, nullity, etc.); Eigenvalues and eigenvectors; Matrix factorizations; Inner products; Distance measures; Projections; Notion of hyperplanes; half-planes. • Probability, Statistics and Random Processes: Probability theory and axioms; Random variables; Probability distributions and density functions (univariate and multivariate); Expectations and moments; Covariance and correlation; Statistics and sampling distributions; Hypothesis testing of means, proportions, variances and correlations; Confidence (statistical) intervals; Correlation functions; White-noise process. **Unit -3: Introduction to Data Science Methods:** Linear regression as an exemplar function approximation problem; Linear classification problems. **Unit -4: Handling large data on a single computer:** • The problems you face when handling large data-General techniques for handling large volumes of data-General programming tips for dealing with large data sets-Case study 1: Predicting malicious URLs-**First steps in big data-**Distributing data storage and processing with frameworks **Unit 5: Join the NoSQL movement-Introduction to NoSQL** **Unit 6: The rise of graph databases:** • Introducing connected data and graph databases • Introducing Neo4j: a graph database **Unit 7: Data visualization to the end user:** • Data visualization options • Crossfilter, the JavaScript MapReduce library • Creating an interactive dashboard with dc.js • Dashboard development tools	6 hours 7 hours 7 hours 7 hours 7 hours 7 hours 7 hours

Content Practical:	**Suggested Lab Assignment:** **Program to understand these concepts**: Numpy Arrays objects, Creating Arrays, basic operations, Indexing, Slicing and iterating, copying arrays, shape manipulation, Identity array, eye function, Universal function **Program to understand these concepts:** Linear algebra with Numpy, eigen values and eigen vectors with Numpy **Program to understand these concepts:** Aggregation and Joining, Pandas Object: Concatenating and appending data frames, index objectsHandling Time series data using pandas **Program to understand these concepts:** Handling missing values using pandas **Program to understand these concepts:** Reading and writing the data including JSON data **Program to understand these concepts:** Web scraping using python, Combining and merging **Program to understand these concepts:** Data transformations Basic matplotlib plots, common plots used in statistical analysis in python **Program to understand these concepts:** Common plots used in statistical analysis in python Data Types **Program to understand these concepts:** Sequence generation, Vector and subscript, Random number generation Data frames and functions-Data manipulation and Data Reshaping using plyr, dplyr, reshape **Program to understand these concepts:** Parametric statistics and Non-parametric statistics- Continuous and Discrete Probability distribution using python Correlation and covariance, contingency tables- Overview of Sampling, different sampling techniques- and database connectivity2.	15 hours 5 hours 5 hours 5 hours 5 hours 5 hours 4 hours 4 hours
Pedagogy	Lectures/ Tutorials/Hands-on assignments/Self-study	

References/ Readings	1. Practical Statistics for Data Science by Peter Bruce, Andrew Bruce, Peter Gedeck, May 2017 2. Naked Statistics by Charles Wheelon, 2012 3. Business Data Science by Matt Taddy, McGraw Hill, 2019 4. Elements of statistical learning by Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie,2001 5. Python for Data Analysis by Wes McKinney, 2nd edition, 2017 6. Data Science and Big Data Analytics -EMC2 7. James Payne, “Beginning Python: Using Python 2.6 and Python 3.1” Wrox, Ist Edition, 2010. 8. Michael T. Goodrich, Roberto Tamassia, Michael H. Goldwasser, “Data Structures and Algorithms in Python”, John Wiley & sons, 2013. 9. Ivan Idris, “Python Data Analysis”, Packt Publishing Limited, 2014. 10. Wes McKinney, “Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython”, O'Reilly Media, Ist Edition, 2012. 11. Michael Heydt, “Learning Pandas - Python Data Discovery and Analysis Made Easy”, Packt Publishing Limited, 2015. 12. Jacqueline Kazil,KatharineJarmul, “Data Wrangling with Python: Tips and Tools to MakeYour Life Easier”, O'Reilly Media, Ist Edition, 2016.

	13. https://docs.scipy.org/doc/numpy-dev/reference/index.html#reference 14. http://www.python-course.eu/numpy.php
Course Outcomes	1. Understand key data science concepts. 2. Learn programming skills for data manipulation and analysis. 3. Apply data analysis techniques, including preprocessing and basic modeling. 4. Communicate data insights effectively through visualizations and presentations