EC6435D Foundations of Data Analytics

Pre-requisites: Fundamentals of Probability and Statistics, Computer Programming

L	T	P	C
2	0	2	3

Total hours: 26L + 26P

Course Outcomes

CO1: *Demonstrate ability to identify and integrate data of various types from a variety of sources, and make informed judgements about their use in data science research.*
CO2: *Critically evaluate the methodologies applied in data gathering, data processing and data exploration to disseminate findings using data visualization tools.*
CO3: *Apply different data science tools to create appropriate visualization of high dimensionality data, aligned to the student's area of interest.*

Module 1: (9 hours)

Introduction to Data Science: Data, knowledge and information. Structured, semi-structured, and un-structured data. Database theory for data science. Relational database, primary key, secondary key. Database normal form: First normal form, second normal form, and third normal form. SQL database for structured data: adding/deleting/modifying tables, adding/deleting/modifying rows, searching and other essential operations. Semi-structured data: XML for semi-structured data, XML syntax and parsing XML using python. Big data: Characteristics of big data, Big data models: key value model, column model, document model, graph model. High level architecture of NoSQL systems.

Module 2: (6 hours)

Data pre-processing: Introduction to Pandas in Python, Data cleaning and preparation: Duplicates, Missing data, transformation using a function/mapping, discretisation of data, errors and outliers. Data wrangling: Hierarchical indexing, combining and merging data, reshaping and pivoting. Data munging, Data cleaning. Quality of data, meta-data, Canonicalization, legal and ethical aspects.

Module 3: (11 hours)

Introduction to data visualization. Visualization plots: Bar graph and pie charts, box plots, scatter plots and bubble charts, KDE plots. Introduction to data visualization libraries in Python: matplotlib, pandas and seaborn. Data transformation: Indexing, slicing, splitting, iterating, filtering, sorting, combining and reshaping. Introduction to data transformation libraries in Python: numpy and pandas. Exploratory data analytics: Univariate analytics, bivariate analytics and multi-variate analytics. Measures of central tendency and dispersion. Data aggregation, pivot tables and correlation. Scraping online/website data, Interactive visualization plots.

References:

1. Meysman, A., Cielen, D. and Ali, M, Introducing Data Science: Big data, machine learning, and more, using Python tools, Manning Publishers, 2016.
2. C J Date. Database Design and Relational Theory: Normal Forms and All That Jazz, O'Reilly, 2012
3. Cathy Tanimura. SQL for Data Analysis: Advanced Techniques for Transforming Data into Insights, O'Reilly, 2021
4. Deborah Nolan, Duncan Temple Lang XML and Web Technologies for Data Sciences with R, Springer, 2014.
5. Andreas Meier, Michael Kaufmann. SQL & NoSQL Databases: Models, Languages, Consistency Options and Architectures for Big Data Management, Springer, 2019.
6. Andy Kirk. Data Visualisation: A Handbook for Data Driven Design, SAGE Publications Ltd, 2016.
7. Kyran Dale. Data Visualization with Python and JavaScript: Scrape, Clean, Explore & Transform Your Data, O'Reilly, 2016.
8. Abha Belorkar, Sharath Chandra Guntuku. Interactive Data Visualization with Python: Present your data as an effective and compelling story, Packt Publishing Limited, 2020