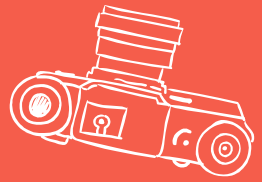
CSC 601
Data Analytics and Visualization





Structure of Syllabus

Objectives, Outcomes, Modules, Books, Question Paper Structure, Laboratory, Practical exams, Term work Submission

Prerequisite:

- ✕ Basic statistics and Maths
- ✕ Python programming

Course Objectives :

- To Introduce the concept of Data Analytics Lifecycle.
- To Develop Mathematical concepts required for advance regression.
- To Understand data modeling in time series and its process.
- To create awareness about Text analytics and its applications.
- To provide overview of Data analytics and visualization with R.
- To provide overview of Data analytics and visualization with Python.

Course Outcomes:

After successful completion of the course students will be able to:

- Comprehend basics of data analytics and visualization.
- Apply various regression models on given data set and perform prediction.
- Demonstrate advance understanding of Time series concepts and analysis of data using various time series models.
- Analyze Text data and gain insights.
- Experiment with different analytics techniques and visualization using R.
- Experiment with different analytics techniques and visualization using Python.

Chapter 1:
Introduction to Data analytics and life cycle

- **Data Analytics Lifecycle overview :** Key Roles for a Successful Analytics, Background and Overview of Data Analytics Lifecycle Project
- **Phase 1:** Discovery: Learning the Business Domain, Resources Framing the Problem, Identifying Key Stakeholders. Interviewing the Analytics Sponsor, Developing Initial Hypotheses Identifying Potential Data Sources
- **Phase 2:** Data Preparation: Preparing the Analytic Sandbox, Performing ETLT, Learning About the Data, Data Conditioning, Survey and visualize, Common Tools for the Data Preparation Phase
- **Phase 3:** Model Planning: Data Exploration and Variable Selection, Model Selection, Common Tools for the Model Planning Phase
- **Phase 4:** Model Building: Common Tools for the Model Building Phase
- **Phase 5:** Communicate Results
- **Phase 6:** Operationalize

Chapter 2: Regression Models

- **Introduction to simple Linear Regression:** The Regression Equation, Fitted value and Residuals, Least Square
- **Introduction to Multiple Linear Regression:** Assessing the Model, Cross-Validation, Model Selection and Stepwise Regression, Prediction Using Regression.
- **Logistic Regression:** Logistic Response function and logit, Logistic Regression and GLM, Generalized Linear model, Predicted values from Logistic Regression, Interpreting the coefficients and odds ratios, Linear and Logistic Regression: similarities and Differences, Assessing the models.

Chapter 3: Time Series

- Overview of Time Series Analysis Box-Jenkins Methodology,
- ARIMA Model Autocorrelation Function (ACF),
- Autoregressive Models,
- Moving Average Models,
- ARMA and ARIMA Models,
- Building and Evaluating an ARIMA Model, Reasons to Choose and Cautions

Chapter 4:
Text Analytics

- History of text mining, Roots of text mining overview of seven practices of text analytic,
- Application and use cases for Text mining: extracting meaning from unstructured text, Summarizing Text.
- Text Analysis Steps, A Text Analysis Example, Collecting Raw Text, Representing Text, Term Frequency—Inverse Document Frequency (TFIDF), Categorizing Documents by Topics, Determining Sentiments, Gaining Insights .

Chapter 5:
Data analytics and visualization with R

- Introduction to R: Data Import and Export, Attribute and Data type, Descriptive statistics.
- Exploratory Data Analysis: Visualization before analysis, Dirty Data, visualizing single variable, examining Multiple variable, Data Exploration versus presentation.

Chapter 6:
Data analytics and Visualization with Python

- Essential Data Libraries for data analytics: Pandas, NumPy, SciPy.
- Plotting and visualization with python: Introduction to Matplotlib, Basic Plotting with Matplotlib, Create Histogram, Bar Chart, Pie chart, Box Plot, violin plot using Matplotlib.
- Introduction to seaborn Library, Multiple Plots, Regression plot, regplot.

Text Books

- **Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, EMC Education services Wiley Publication**
- Data Analytics using Python: Bharati Motwani, Wiley Publications.
- **Practical Statistics for Data Scientists 50+ Essential Concepts Using R and Python, O'Reilly Publications 2nd Edition**
- Practical Text Mining and statistical Analysis for non-structured text data applications,1st edition, Grey Miner, Thomas Hill.

Reference Books

- Data Mining, Concepts and Techniques: 3rd edition, Jiawei Han, Micheline Kamber and Jian Pei
- Data Analytics using R, Bharati Motwani, Wiley Publications
- Python for Data Analysis: 3rd Edition, Wes McKinney,Publisher(s): O'Reilly Media, Inc.
- Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data by Dipanjan Sarkar

Useful Links

- http://varianceexplained.org/RData/
- https://www.kaggle.com/code/iamleonie/time-series-interpreting-acf-and-pacf
- https://www.geeksforgeeks.org/data-visualization-using-matplotlib/

Assessment

✕ Internal Assessment:

Assessment consists of two class tests of 20 marks each. The first-class test is to be conducted when approx. 40% syllabus is completed and second-class test when additional 40% syllabus is completed. Duration of each test shall be one hour.

✕ End Semester Theory Examination:

1 Question paper will consist of 6 questions, each carrying 20 marks.

2 The students need to solve a total of 4 questions.

3 Question No.1 will be compulsory and based on the entire syllabus.

4 Remaining question (Q.2 to Q.6) will be selected from all the modules.

CSL 601

Data Analytics and Visualization Lab

Prerequisite: Basic Python and R

Lab Objectives:

- To effectively use libraries for data analytics.
- To understand the use of regression Techniques in data analytics applications.
- To use time series models for prediction.
- To introduce the concept of text analytics and its applications.
- To apply suitable visualization techniques using R and Python.

Lab Outcomes:

At the end of the course, students will be able to —-

- Explore various data analytics Libraries in R and Python.
- Implement various Regression techniques for prediction.
- Build various time series models on a given data set.
- Design Text Analytics Application on a given data set.
- Implement visualization techniques to given data sets using Python/R .
- Implement visualization techniques to given data sets using industry used tools .

List of Experiments

SR. NO.	TITLE
1	Getting introduced to data analytics libraries in Python and R.
2	Simple Linear Regression in Python/R
3	Multiple Linear Regression in Python/R.
4	Time Series Analysis in Python/R.
5	Implementation of ARIMA model in Python / R.
6	Text analytics: Implementation of Spam filter/Sentiment analysis in python/R.
7	Visualization in R using Libraries.
8	Visualization in python using Matplotlib Library.
9	Visualization in python using Seaborn Library.
10	Visualization in Power BI.

Useful Links:

- https://www.geeksforgeeks.org/data-visualization-with-python
- https://www.coursera.org/specializations/data-science-python
- https://www.geeksforgeeks.org/data-visualization-in-r/
- https://towardsdatascience.com/introduction-to-arima-for-time-series-forecasting

References:

- Data Analytics using R, Bharati Motwani, Wiley Publications
- Python for Data Analysis: 3rd Edition, WesMcKinney, Publisher(s): O'Reilly Media, Inc.
- Better Data Visualizations A Guide for Scholars, Researchers, and Wonks, Jonathan Schwabish, Columbia University Press

Term Work:

- Term work should consist of 10 experiments.
- Journal must include at least 2 assignments based on Theory and Practicals
- The final certification and acceptance of term work ensures satisfactory performance of laboratory work and minimum passing marks in term work.
- Total 25 Marks (Experiments: 15-marks, Attendance Theory & Practical: 05-marks, Assignments: 05-marks)

Oral & Practical exam

- Based on the entire syllabus

Any queries