B.Sc. (Data Science)
Discipline Specific Course (DSC)
Semester III
BSDB32303T: Data Preparation

Total Marks: 100
External Marks: 70
Internal Marks: 30
Credits: 4
Pass Percentage: 40%

Objective

This course will make student familiar with the methods of data gathering and preparing the data for processing. Student will explore the sampling and data processing techniques.

Section A

Unit I: Data collection and management: Introduction, Sources of data, Data collection and APIs, Exploring and fixing data, Data storage and management, using multiple data sources, Data Quality, Addressing Data Quality Issues.

Unit II: Data Collection: Experiments and surveys, Collection of Primary and Secondary Data, selection of appropriate method for data collection, case study method. Introduction to Data Preparation process, Some problems in preparation process, Missing values and Outliers, types of Analysis.

Unit III Data Preparation: Data formats, parsing and transformation, Scalability and real-time issues. Data Cleaning: Consistency checking, Heterogeneous and missing data, Data Transformation and segmentation

Unit IV: Research Design: Meaning, Need for Research Design, Features of a Good Design, Important Concepts relating to Research Design, Different Research Designs. cluster analysis: Introduction, distance measures Clustering algorithms, agglomerative clustering.

Section B

Unit V: Sampling distributions: Non - central chi - square, t and F distributions and their properties - Distributions of quadratic forms under normality -independence of quadratic form and a linear form - Cochran's theorem.

Unit VI: Simple random sampling (WR and WOR) - Use of Random number Table, Stratified Random Sampling: Properties, Systematic Sampling: Estimation of the Mean and Variance – Comparison of Simple, Stratified and Systematic Sampling

Unit VII Testing f Hypothesis: Hypothesis, Basic Concepts Concerning Testing the Hypotheses, Test Statistic and Critical region, critical value and Decision Rule, Procedure for Hypothesis Testing, Hypothesis Testing for – Means, Proportions, variance, difference of two mean, difference of two proportions, difference of two variances; P-Value approach, power of test,

Unit VIII: Concepts of time series – Components of time series – Additive and multiplicative models for the analysis of time series - Measurement of trend by (i) Graphic method, (ii) Semi Average method, (iii) Method of Curve Fitting by principle of least squares, (iv)Method of Moving Averages

Suggested Readings

1. C. R. Kothari – Research Methodology Methods and Techniques - New Age International Publishers, 3rd Edition, 2014.

2. Garg, B.L., Karadia, R., Agarwal, F. and Agarwal, An introduction to Research Methodology, RBSA Publishers. 2002
3. Gupta, S.C. and Kapoor, V.K.: Fundamentals of Applied Statistics, Sultan Chand & Co., 11th ed., 2001
4. William G. Cochran.: Sampling Techniques, John Wiley Sons, 3rd edition, 1977