Data Science: Sem VI								
Course Code	Course Name	Teaching Scheme (Contact Hours)			Credits Assigned			
		Theory	Practical	Tutorial	Theory	Practical	Tutorial	Total
HDSC601	Statistical Learning for Data Science	04	--	--	04	--	--	04

Course Code	Course Name	Examination Scheme							
		Theory Marks				Exam Duration	Term Work	Practical and Oral	Total
		Internal Assessment			End Sem. Exam.				
		Test1	Test2	Avg.					
HDSC601	Statistical Learning for Data Science	20	20	20	80	03	--	--	100

Course Prerequisites:	
1	Engineering Mathematics, Probability and Statistics
Course Objectives:	
1	To understand basic statistical foundations for roles of Data Scientist.
2	To develop problem-solving skills.
3	To infer about the population parameters using sample data and perform hypothesis testing.
4	To understand importance and techniques of predicting a relationship between data and determine the goodness of model fit.
Course Outcomes:	
After successful completion of the course, the student will be able to:	
1	Develop various visualizations of the data in hand.
2	Analyze a real-world problem and solve it with the knowledge gained from sampling and probability distributions.
3	Analyze large data sets and perform data analysis to extract meaningful insights.
4	Develop and test a hypothesis about the population parameters to draw meaningful conclusions.
5	Fit a regression model to data and use it for prediction.

Module No.		Topics	Hours.
1.0		**Introduction**	08
	1.1	**Data and Statistics**: Elements, Variables, and Observations, Scales of Measurement, Categorical and Quantitative Data, Cross-Sectional and Time Series Data, Descriptive Statistics, Statistical Inference, Descriptive Statistics: Tabular and Graphical Summarizing Categorical Data, Summarizing Quantitative Data, Cross Tabulations and Scatter Diagram.	
	1.2	**Descriptive Statistics: Numerical Measures**: Measures of Location, Measures of Variability, Measures of Distribution Shape, Relative Location, and Detecting Outliers, Box Plot, Measures of Association Between Two Variables	

2.0		Probability	08
	2.1	**Probability :** Experiments, Counting Rules, and Assigning Probabilities, Events and Their Probabilities, Complement of an Event, Addition Law Independent Events, Multiplication Law, Baye's theorem	
	2.2	**Discrete Probability Distributions** Random Variables, Discrete Probability Distributions, Expected Value and Variance, Binomial Probability Distribution, Poisson Probability Distribution	
	2.3	**Continuous Probability Distributions:** Uniform Probability Distribution, Normal Curve, Standard Normal Probability Distribution, Computing Probabilities for Any Normal Probability Distribution	
3.0		**Sampling and Sampling Distributions**	05
	3.1	Sampling from a Finite Population, Sampling from an Infinite Population, Other Sampling Methods, Stratified Random Sampling, Cluster Sampling, Systematic Sampling, Convenience Sampling, Judgment Sampling	
	3.2	Interval Estimation: Population Mean: Known, Population Mean: Unknown, Determining the Sample Size, Population Proportion	
4.0		**Hypothesis Tests**	05
	4.1	Developing Null and Alternative Hypotheses, Type I and Type II Errors, Population Mean: Known Population Mean: Unknown Inference About Means and Proportions with Two Populations-Inferences About Population Variances, Inferences About a Population Variance, Inferences About Two Population Variances	
	4.2	Tests of Goodness of Fit and Independence, Goodness of Fit Test: A Multinomial Population, Test of Independence	
5.0		**Regression**	08
	5.1	**Simple Linear Regression:** Simple Linear Regression Model, Regression Model and Regression Equation, Estimated Regression Equation, Least Squares Method, Coefficient of Determination, Correlation Coefficient, Model Assumptions, testing for Significance, Using the Estimated Regression Equation for Estimation and Prediction Residual Analysis: Validating Model Assumptions, Residual Analysis: Outliers and Influential Observations	
	5.2	**Multiple Regression:** Multiple Regression Model, Least Squares Method, Multiple Coefficient of Determination, Model Assumptions, Testing for Significance, Categorical Independent Variables, Residual Analysis	
6.0		**Time Series Analysis and Forecasting**	05
	6.1	Time Series Patterns, Forecast Accuracy, Moving Averages and Exponential Smoothing, Trend Projection, Seasonality and Trend and Time Series Decomposition	
	6.2	**Nonparametric Methods** Sign Test, Wilcoxon Signed-Rank Test, Mann-Whitney-Wilcoxon Test, Kruskal-Wallis Test, Rank Correlation	
		Total	48

Text Books:

1 https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf
2 Data Science from Scratch, FIRST PRINCIPLES WITH PYTHON, O'Reilly, Joel Grus,
3 Data Science from Scratch (oreillystatic.com)
4 Practical Time Series Analysis, Prediction with statistics and Machine Learning, O'Reilly, Aileen Nielsen [DOWNLOAD] O'Reilly Practical Time Series Analysis PDF (lunaticai.com)
5 R for data science: Import, Tidy, Transform, Visualize, And Model Data, O'Reilly , Garrett Grolemund, Hadley Wickham
6 Python for Data Analysis, 2nd Edition, O'Reilly Media, Wes McKinney.
7 https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf

References:

1 Data Science for Dummies Paperback, Wiley Publications, Lillian Pierson
2 Storytelling with Data: A Data Visualization, Guide for Business Professionals, Wiley Publications, Cole Nussbaumer Knaflic
3 Probability and Statistics for Engineering and the Sciences, Cengage Publications Jay L. Devore.

Assessment:

Internal Assessment: (20)

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

End Semester Theory Examination: (80)

1 Question paper will comprise of **total 06** questions, each carrying **20 marks**.
2 **Question No: 01** will be **compulsory** and based on the entire syllabus wherein 4 to 5 sub-questions will be asked.
3 Remaining questions will be mixed in nature and randomly selected from all the modules.
4 Weightage of each module will be proportional to number of respective lecture hours as mentioned in the syllabus.
5 **Total 04 questions** need to be solved.