Course Code	TBD		
Course Name	Data Science		
Credits	4		
Course Offered to	UG/PG		
Course Description	Data science, being an interdisciplinary subject interface with computer science, statistics, machine learning and science of data visualization. This course is designed to enable students to perform exploratory data analysis employing statistical methods and visualization tools. In this course, students will learn to manage and analyze data. They will also learn to use regression techniques to discover and interpret intrinsic patterns in data of various types. Finally, students learn to build custom softwares and web-servers for data analysis.		
Pre-requisites			
Pre-requisite (Mandatory)	Pre-requisite (Desirable)	Pre-requisite (other)	
CSE101 Intro to Programming, Any Machine Learning course (ML or SML)	Linear Algebra		
*Please insert more rows if required			
Post Conditions*(For suggestions on verbs please refer the second sheet)			
CO1	**CO2**	**CO3**	**CO4**
Students will be able to perform exploratory analysis of multivariate data and scientific data visualisation	Student will be able to conduct statistical hypothesis testing	Student will be able to use regression techniques for predictive data analytics and time series modeling	Students will build capability of real life problem solving and dealing with large data.
Weekly Lecture Plan			
Week	**Lecture Topic**	**COs Met**	**Assignment/Labs/Tutorial**

1,2	Random variable, distribution, Maximum Likelihood Estimation usingmaxLik, basic multivariate stats - matrix summarisation, Simpson's paradox, variance-covariance, correlation, canonical correlation; Data preprocessing, exploratory data analysis and high quality visualisation. Advanced scientific plots - stacked histograms for multivariate data, bi-variate scatter plots, parallel coordinate plot, table plot, mosaic plot etc.	CO1	Refresher modules will be offered by TA/PHD students on Linear Algebra and basic R commands/installations.
3	Goodness of fit - likelihood ratio test, Lagrange multiplier test, Q-Q plot, performing varity of hypothesis testings.	CO2	One of the initial assignments will be on linear algebra
4	Dimension reduction using PCA, SVD, tSNE	CO1	In class, short hands on sessions will be conducted using R/Python
5	Generalised linear models (GLM) with various link functions (eg logit). Specific focus on gamma regression	CO3	Students will require to complete 5 assignments and 1 mini project/ Kaggle challenge
6	Time series modeling using autoregressive errors (AR), moving average (MA), ARIMA - stationary and non-stationary time series data, mean stationarity, trend stationarity, statistical test for stationarity.	CO3	As part of the assignments/projects, students will be encouraged to create software modules, interactive dashbord using R/python libraries/apis
7	Survival Analysis using survfit - Kaplan Meier survival density estimation, Cox proportional hazards model	CO3	
8	Gaussian mixture model and Naive Bayes, assessment of model performance	CO3	
9	Bootstrapping and Monte Carlo methods, randomisation test.	CO1	
10, 11	Introduction to handling large data - locality sensitive hashing, sizing sketches, coreset	CO3, CO4	
12, 13	Applications - gene expression, EHR data, demand forecasting, price optimisation in retal, probability of default in banking	CO4	
*Please insert more rows if required			
Assessment Plan			
Type of Evaluation	% Contribution in Grade		
Assignment	25		
Mid-sem	15		
End-sem	35		
Project	15		
Quiz	10		

*Please insert more row for other type of Evaluation			
Resource Material			
Type	**Title**		
Textbook	[1] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011. [2] Tan, Pang-Ning. Introduction to data mining. Pearson Education India, 2007. [3] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. No. 10. New York, NY, USA:: Springer series in statistics, 2001. [4] Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014. [5] R for Data Science, by Garrett Grolemund and Hadley Wickham (2016) [6] Exploratory Data Analysis with R, by Roger D. Peng (2016) [7] An Introduction to Statistical Learning with Application in R, First Edition, by Gareth James et al. (2013) [8] Introduction to linear algebra, by Gilbert Strang		
Reference			