# Learning Objectives: Data Science & Big Data Analytics 2019

## Working in R

- Be able to use R to perform mathematical calculations.
- Understand the different data types used by R.
- Know how to work with vectors (creating, indexing)
- Be able to use R to perform element-wise calculations to vectors/columns, (e.g. adding columns together, averaging columns, scaling columns).
- Be able to write a for loop in R e.g. to print a simple sequence 2,4,6,8 or perform a certain task 100 times and store the results in a data frame or vector.
- Understand how to use tests > < == != and combine them using & | to test for a given condition on a value or column of values
- Be able to use the ifelse function in R on a column to generate a new column based on the column values.
- Be able to write simple R functions that take some input arguments and return a calculated result.
- Be able to generate random number sets e.g. from a uniform or normal distribution, and set a random seed appropriately.

# Learning Objectives: Data Science & Big Data Analytics 2019

**Working with datasets in R**

- Be able to load data in R using a variety of formats (csv, xls, stata, Rdata, Rdat)
- make selections of rows/columns
- delete rows /columns
- filter rows based on conditions (and sets of conditions)
- use column data to create new columns e.g. based on a transformation of units or a formula involving other columns
- convert columns between types (e.g. between character/numerical/factor)
- be able to use cut to convert a numerical column into a factor column
- be able to rename rows and columns
- be able to change the level labels for factor columns
- Be able to sort a dataframe by column values
- Be able to randomly sample a dataframe (e.g. to create training and test subsets, bootstrapped datasets, or to randomise row order)
- Be able to normalise columns (by shifting and scaling) e.g. so that it has mean = 0 and standard deviation = 1.
- be able to use the dplyr group_by and summarise commands
- be able to use apply to apply a function to the rows or columns of a dataframe
- be able to build a dataframes (e.g by joining columns, adding columns to an existing dataframe, or by adding new rows to an existing dataframe)

## Learning Objectives: Data Science & Big Data Analytics 2019

**Exploring datasets in R**

- Be able to find and interpret mean, median of a column
- Be able to find and interpret standard deviation and variance of a column
- Be able to find and interpret range and quantiles (e.g. e.g. to find 95% interval) of a column
- Be able to inspect for outliers visually and find them in the dataset and take appropriate action.
- Be able to find the correlation between data columns
- Be able to make and interpret scatter plots
- Be able to make and interpret histograms (with defined cut/bin positions)
- Be able to make and interpret boxplots
- Be able to style plots: set x and y axis limits, set x and y axis labels, set title, set point marker types, set line styles, and set point colour based on a criteria or factor class.
- Be able to style plots: set x and y axis labels, set title
- Be able to style plots: set point marker types and set line styles,
- Be able to style plots: set point colour based on a criteria or factor class.
- Be able to use the table command to build a frequency table.
- Be able to add annotations to plots: add text to a plot
- Be able to add annotations to plots: add vertical / horizontal lines at given values
- Be able to view a smoothed trend line onto scatter plot using scatter.smooth
- Be able to takes appropriate steps to identify where columns have missing (NA) values, and take appropriate measures (removal or exclusion from an analysis).

# Learning Objectives: Data Science & Big Data Analytics 2019

## Hypothesis testing in R

- Understand the assumptions made in performing a t-test, how to carry it out in R and interpret and explain the results (t-value, p-value, confidence interval).
- Be able to carry out a t-test on two groups of values to find evidence for a difference in the means, or test a hypothesis that one mean is greater/less than the other.
- Be able to carry out a t-test on two columns to find evidence to test the hypothesis that the columns are correlated.

## Bivariate Regression in R

- Understand and be able to explain the principle of least-squares regression.
- Be able to carry out simple linear regression in R to make a linear model that predicts a response based on a single independent variable, and plot the line of best fit onto a scatter plot of the data.
- Be able to display and understand the results of the fit, and the associated measures returned.
- Be able to recall and explain the assumptions of linear regression, use R to explore if these assumptions are met, and carry out suggested steps to identify and manage these issues (e.g. to perform a suggested transformation, or remove outliers).
- Be able to explore how simple variable transformations improve/do not improve linear fitting.
- Be able to make predictions using the fit (both for the training data, or a new dataset)
- Be able to access the residuals, and calculate values of RSS, MSE, and RMSE.

## Learning Objectives: Data Science & Big Data Analytics 2019

### Multivariate Regression in R

- Understand how to perform a multivariate fit in R.
- Understand how to interpret the fit coefficients in a multivariate fit, and their associated p-values.
- Understand how to interpret the f-test value returned from a multivariate fit.
- Understand how multivariate fits can include categorical variables, the use of dummy variables, and the interpretation of the resulting fit coefficients and their associated p-values.
- Be able to include interaction terms in a multivariate fit, interpret the resulting coefficients, and test for their significance.
- Be able to plot the result of multivariate fits appropriately (e.g. plots of the fit for simple models: 1 numerical predictor and 1 factor predictor (with/without interaction); plot of actual vs predicted values for more complex models; plot residuals vs fitted values)
- Understand the problem arising with highly correlated predictors, and be able to detect and suggest suitable steps to deal with these issues.

### Model selection

- Be able to interpret fit results to evaluate the significance of the different predictors used, and suggest potential model improvements.
- Understand how to perform an ANOVA test on a multivariate model to compare the performance to a model containing only a subset of the predictors, and interpret the result.
- Be able to use the stepAIC function to perform stepwise model selection to optimise a model, understanding the arguments that can be used.
- Be able to use the leaps function to test all possible predictor subsets and interpret the results and plots produced.
- Be able to use and optimise multivariate fits. using Ridge Regression and Lasso methods to constrain fit coefficients.

# Learning Objectives: Data Science & Big Data Analytics 2019

## Cross Validation

- Understand why the performance of a model on its training data may differ significantly in comparison to a test / validation / new data.
- Be able to describe the approaches of validation set (aka hold out) and be able to carry it out in R.
- Be able to describe the approaches of LOOCV and be able to carry it out in R using suitable functions
- Be able to describe the approaches of k-fold validation and be able to carry it out in R using suitable functions.
- Understand how to interpret the results of validation testing to select optimum models.
- Be able to describe the relative advantages/disadvantages of validation set (aka hold out), LOOCV, k-fold validation methods.
- Understand and be able to carry out the method of resampling to generate bootstrapped datasets that can be used when evaluating model performance.
- Understand how the process of bootstrapping can be used to investigate the distribution of fit parameter results.
- Be able to carry out a bootstrap analysis to investigate e.g. fit performance or fit coefficients
- Be able to use a for loop to carry out an optimisation analysis using a cross validation method, plot the results.

## Classification

- Be able to use R to calculate simple probabilities associated with sampling a dataset.
- Understand how the logistic function can be fitted to data to predict the probabilities associated with a binomial (0 or 1) variable in relation to the values of a set of predictors. Know why it is used in preference to a linear function.
- Be able to carry out a logistic regression fit in R and interpret the results.
- Understand how to display the results of a classification model using table to display the confusion matrix comparing predictions to true values.
- Be able to use the results of a classification model to make predictions, and measure the performance according to the misclassification rate.
- Be able to describe the assumptions which Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) make to build a classification model.
- Be able to use R to perform a LDA (Linear Discriminant Analysis) fit for classification.
- Be able to use R to perform a QDA (Quadratic Discriminant Analysis) fit for classification.
- Be able to apply a suitable methods of cross validation to measure performance of different classification methods and optimise models (e.g. to compare different fitting models).

## K- Nearest Neighbours algorithm

- Be able to describe the principle of the K-nearest neighbours algorithm for classification and regression problems.
- Understand KNN requires consideration of the scale of each variable used as a predictor, and how to apply normalisation when appropriate.
- Be able to explore how the performance of the KNN method varies with K for a particular dataset and select an optimal K value using a cross validation method.

## Learning Objectives:  Data Science & Big Data Analytics 2019

**Comparisons of Fitting methods**

- Understand that models can be used for both inference and prediction.
- Be able to carry out suitable cross validation testing to measure the performance of different fit methods and evaluate the results.
- Be able to describe the different fitting methods considered in terms of their assumptions, and relative usefulness for making predictions/inferences.

**Working with R code files and notebooks**

- Be able to create and run code in R scriptfiles (.R extension)
- Be able to write a report in R Markdown as a notebook, that consists of code sections and appropriately styled text (headers, inline code, font styles).
- Know how to work with code that is contained in an R Notebook in RStudio.
- Be able to knit the notebook into HTML formatted report.