**Dataset 1 - Unit 1, Lecture 2 - World Health Organization**

- Arithmetical operations. Functions sqrt and abs.

- How to use R help.

- How to save variables, and how to name them.

- How to see all the variables created in the current session.

- Creating vectors. Getting an element from a vector. Creating sequencies.

- Creating data frames. Adding new variables and new observations to a existing data frame.

- Changing the working directory. Getting the current working directory.

- How to read .csv documents in R.

- How to get the structure and the summary from a data frame.

- Understanding the structure and the summary from a data frame.

- Dividing the dataset. Subset function.

- Saving .csv files from data frames in R.

- Removing unnecessary variables.

- How to access to a variable contained in a data frame.

- How to calculate the mean and standard desviation of a variable.

- How to find where the maximum and minimum values of a variable in a data frame are.

- How to create a basic plot.

- How to visualize only some of the variables contained in a data frame.

- How to create histograms and boxplots, how to interpretate them, and how to label them.

- How to use the table function.

- How to use the tapply function. Removing observations with missing values from tapply.

- How to create and save an R Script.

**Dataset 2 - Unit 1, Recitation - Nutritional Education with Data**

- Changing the working directory and reading a .csv file.

- Structure and summary functions review.

- Finding where the maximum value of a variable contained in a data frame is.

- How to use the function names.

- Creating a subset.

- How to find a name in a vector with the match function. How to find a value in a data frame.

- Creating basic plots. Labelling the axis and the plot.

- Changing the colour of a plot.

- Histograms. Creating x-axis limits in a plot. How to break the cells in an histogram.

- Boxplots.

- Creating a logical variable. Changing a logical variable (FALSE - TRUE) into numbers (0 - 1).

- Adding new variables to a data frame.

- Tables with 1 and 2 arguments and tapply function.

**Dataset 3 - Unit 1, Homework 1 - An Analytical Detective**

- Reading a .csv file.

- Getting the structure and the summary from the data frame.

- Using tables (1 and 2 arguments). Sorting values from tables.

- Converting a variable into dates (as.Date function).

- Extracting months and weekdays from a date.

- Histograms and Boxplots.

- Remmoving unnecessary categories from a variable.

**Dataset 4 - Unit 1, Homework 2 - Stock Dynamics**

- Reading a .csv file.

- Converting variables into dates.

- Getting the structure and the summary from the data. Calculating a standard desviation.

- Drawing line plots. Adding new lines to a line plot. Adding vertical lines to a plot.

- Creating y-axis limits in a plot. Selecting the data which data we want to plot in the graph.

- Calculating the mean. Using tapply function. Extracting months from dates.

**Dataset 5 - Unit 1, Homework 3 - Demographics and Employment in the US**

- Reading a .csv file. Getting the structure and the summary from the data frame.

- Creating a table (1 argument) and sorting it.

- Creating a table (2 arguments).

- Counting NA values with the is.na function. Using a tapply function with logical variables.

- Merging data frames on two variables from two different data frames.

**Dataset 6 - Unit 1, Homework 4 (O) - Internet Privacy Poll**

- Reading a .csv file. Getting the structure and the summary from the data frame.

- Creating tables (1 and 2 arguments).

- Creating a subset.

- Creating histograms.

- Adding a small amount of random noise to certain values.

- Using the tapply function.

**Dataset 7 - Unit 2, Lecture 1 - The Statistical Sommelier**

- Learning what one variable linear regression is.

- Measuring the errors. SSE, SST, RMSE, R2.

- Multiple linear regression. What "overfitting" is.

- Creating a one variable linear regression in R. Getting a summary from the model.

- Coefficients in a model. Multiple and adjusted R-squared.

- Computing SSE and SST.

- Creating multiple linear regressions.

- Understanding the model. Selecting variables: "t value", "Pr > (|t|)", the star coding scheme.

- What are correlation and multicollinearity.

- Testing correlation in R. Dealing with multicollinearity in the models.

- Training data and testing data. Out-of-sample accuracy.

- Making predictions in new data. Computing out-of-sample R2.

**Dataset 8 - Unit 2, Lecture 2 - Moneyball: The Power of Sports Analytics**

- Dividing a data frame into a training set and a testing set.

- Creating one variable linear regressions.

- Creating multivariable regressions.

- Getting summaries from the models.

- Testing correlation in R.

- Review Week 1: adding variables into a data frame, creating basic plots.

**Dataset 9 - Unit 2, Recitation - Playing Moneyball in the NBA**

- Creating a one variable linear regression. Getting a summary from the model.

- Creating multivariable regressions.

- Computing SSE and RMSE.

- Making predictions in new data. Computing out-of-sample RMSE and R2.

- Review Week 1: adding variables into a data frame, tables (2 arguments), creating basic plots.

**Dataset 10 - Unit 2, Homework 1 - Climate Change**

- Dividing a data frame into a training set and a testing set.

- Creating a multivariable linear regression. Getting a summary from the model.

- Testing correlation in R.

- Finding automatically a good compromise of model simplicity and R2. The step function.

- Making predictions in new data. Computing out-of-sample R2.

**Dataset 11 - Unit 2, Homework 2 - Reading Test Scores**

- Removing observations with missing values.

- Dealing with factor variables.

- Setting reference values in unordered factors.

- Building a linear model using all the variables in the data frame (except the dependient one).

- Getting the summary from the model.

- Getting the RMSE from the model.

- Making predictions in new data. Calculating SSE, RMSE and R2 in the testing set.

- Review Week 1: tapply function.

**Dataset 12 - Unit 2, Homework 3 - Detecting Flu Epidemics**

- Dealing with skewed variables. Using logarithms and exponential functions.

- Building a linear model with one independient variable.

- Getting the summary from the model.

- Making predictions when we use the logarithm of the dependient variable.

- Finding test in a data frame. Using the match function.

- Calculating SSE and RMSE in the testing set.

- Building a time series model with two independient variables. Lagging observations.

- Review Week 1: finding where maximum values are, histograms, basic plots.

**Dataset 13 - Unit 2, Homework 4 (O) - State Data**

- Using datasets that R has built in.

- Creating multivariable linear models and getting their summary.

- Making predictions in the training set.

- Review Week 1: basic plot, tapply, boxplot, subset, finding where minimum and maximum values are.

**Dataset 14 - Unit 2, Homework 5 (O) - Forecasting Elantra Sales**

- Splitting the data into a training set and a testing set.

- Creating a multivariable linear model.

- Converting numeric variables into factor variables. Building a model with factor variables.

- Testing correlations of more than two variables.

- Making predictions in new data. Computing out-of-sample SSE, RMSE and R2.

- Review Week 1: finding maximum values.

**Dataset 15 - Unit 3, Lecture 1 - D2Hawkeye (Modelling an Expert)**

- Categorical and binary variables.

- Logistic Regression: concept. Odds and Logit.

- Splitting randomly a dataset. Setting a seed.

- Building a two variables logistic regression model.

- Getting a summary and interpreting the model. The AIC index.

- Making predictions with logistical regressions on the training set.

- Thresold values. Classification (or confusion) matrix. True and false positives and negatives.

- Comparing with baseline models.

- Sensitivity and specificity. ROC curve. Generating ROC curves in R.

- Area under the ROC curve (AUC) and accuracy. Computing AUC in R.

- Review Week 1: table, tapply.

**Dataset 16 - Unit 3, Lecture 2 - The Framingham Heart Study**

- Splitting randomly a dataset in a classification problem. Setting a seed.

- Building a logistic model using all the variables in the frame (except the dependient one).

- Getting the summary and interpreting the model.

- Making predictions in the test set.

- Comparing with baseline models. Creating a confusion matrix in R. Calculating the accuracy.

- Calculating specificity and sensitivity.

- Understanding the importance of external validation.

- Review Week 1: table.

**Dataset 17 - Unit 3, Recitacion - Election Forecasting**

- Filling missing points (NA vaules) with average values.

- Dealing with multicollinearity.

- Building a logistic with one independient variable. Getting the summary.

- Making predictions in the training set and comparing with a smart baseline model.

- Building a two variable model.

- Making predictions in the testing set.

- Review Week 1: table, searching where data are.

- Review Week 2: splitting a data frame into a training set and a testing set. Testing correlations.

**Dataset 18 - Unit 3, Homework 1 - Popularity of Music Records**

- Removing variables from a data frame.

- Creating a logistic model with all the variables in the frame (except the independient one).

- Getting the summary of the model and interpreting it.

- Dealing with multicollinearity. Removing variables before building a model.

- Testing the model in the testing set. Making predictions.

- Calculating accuracy, sensitivity and specificity.

- Review Week 1: creating subsets, searching where data are, table, finding maximums.

- Review Week 2: dividing a data frame into a training set and a testing set. Testing correlations in R.

**Dataset 19 - Unit 3, Homework 2 - Predicting Parole Violators**

- Splitting randomly a dataset in a classification problem. Setting a seed.

- Creating a logistic model with all the variables in the frame (except the independient one).

- Getting the summary of the model and interpreting it, specially its coefficients.

- Odds and probabilities.

- Making predictions in new data. Calculating accuracy, sensitivity and specificity.

- Calculating the AUC.

- Review Week 1: table.

- Review Week 2: converting variables into factors.

**Dataset 20 - Unit 3, Homework 3 - Predicting Loan Repayment**

- Setting a random seed. Filling missing points with average values.

- Splitting randomly a dataset in a classification problem.

- Creating a logistic model with all the variables in the frame (except the independient one).

- Getting the summary of the model and interpreting it. Calculating the AUC.

- Making predictions in new data. Calculating accuracy.

- Creating a smart baseline model.

- Review Week 1: table, subset.

**Dataset 8B - Unit 3, Homework 4 (O) - Predicting World Series Chamption**

- Converting a variable into characters.

- Building logistic models, getting their summary and interpreting them.

- Review Week 1: table, subset, converting logical variables into numerical variables.

- Review Week 2: testing correlations.

**Dataset 21 - Unit 4, Lecture 1 - Predicting Supreme Court Decissions**

- Classification tree concept.

- Creating a CART model in R. The "minbucket" argument.

- Plotting and interpreting trees. Making predictions in unseen data.

- Drawing ROC curves and computing the AUC in classification tree models.

- Random forests. Implemeting them into R. Making predictions with random forests.

- Complexity parameter concept.

- Using the cross validation method in R and selecting the best cp value.

- Creating a CART model using cp instead of minbucket.

- Review Week 1: table.

- Review Week 2: converting a variable into a factor variable.

- Review Week 3: splitting randomly data into a training set and a testing set. Setting a random seed. Confusion matrices. Calculating accuracy.

**Dataset 22 - Unit 4, Lecture 2 - The D2Hawkeye Story. Health Costs**

- Penalty error concept. Penalty matrixes.

- Creating matrices in R. Converting tables into matrices.

- Multipling matrices element by element.

- Building a CART model in R in classification problems with more than 2 cattegories.

- Plotting a tree.

- Including penalty matrices in the models.

- Predicting values in unseen data.

- Review Week 1: table.

- Review Week 3: splitting randomly data into a training set and a testing set. Setting a random seed. Confusion matrices. Calculating accuracy.

**Dataset 23 - Unit 4, Recitation - Housing Data**

- Regression tree concept.

- Colouring subsets of points in a basic plot.

- Building a regression tree model. Plotting it and making predictions.

- Parameters: λ, RSS, cp. Relationships between them.

- Optimizing the tree (cross validation). Building the optimal tree and testing it.

- Review Week 1: basic plots.

- Review Week 2: creating a linear regression. Adding vertical and horizontal lines to a plot.

- Review Week 3: splitting randomly data into a training set and a testing set. Setting a random seed.

**Dataset 24 - Unit 4, Homework 1 - Why People Vote**

- Creating regression trees. Plotting them.

- Making predictions on unseen data.

- Review Week 1: table, tapply, creating a data frame.

- Review Week 3: creating a logistic regression model and making predictions with it. Classification matrices. Computing the AUC value.

**Dataset 25 - Unit 4, Homework 2 - Letter Recognition**

- Creating a classification tree, and making predictions on unseen data with it.

- Creating a classification random forest, and making predictions on unseen data with it.

- Creating a multi-classification tree, and making predictions on unseen data with it.

- Creating a multi-classification random forest, and making predictions on unseen data with it.

- Review Week 1: table

- Review Week 2: dealing with factor variables.

- Review Week 3: splitting randomly a data set into a training and a testing set. Setting a random seed. creating a logistic regression model and making predictions with it. Classification matrices.

**Dataset 26 - Unit 4, Homework 3 - Predicting Earnings from Census**

- Building classification trees, and making predictions on unseen data with it.

- Plotting trees.

- Drawing ROC curves and computing the AUC in classification tree models.

- Taking a sample from a data set.

- Creating a classification random forest, and making predictions on unseen data with it.

- Finding ways to interpret random forest models.

- Cross validation method to optimize the tree. Making predicitons.

- Review Week 3: splitting randomly a data set into a training and a testing set. Setting a random seed. creating a logistic regression model and making predictions with it. Classification matrices. Plotting ROC curves and calculating their AUC.

**Dataset 13B - Unit 4, Homework 4 (O) - State Data Revisited**

- Building regression trees. Changing their parameters and making predictions on unseen data

- Computing SSE in problems with regression trees.

- Plotting trees. Applying cross validation to optimize the tree.

- Review Week 2: Using datasets that R has built in. Creating a linear model. Evaluating it. Making predictions. Computing SSE. Getting correlations.

**Dataset 27 - Unit 5, Lecture 1 - Turning Tweets into Knowledge**

- Introduction to text analytics.

- Bag of words concept.

- How to clean up irregularities, to remove stop words and stem the words. Limitations of these approaches.

- Setting the language to default. Reading .csv documents when text analytics has to be applied

- Creating a corpus in R.

- Pre-processing the text in R: cleaning up irregularities, removing stop words and stemming them. Getting the content from any element contained in the corpus.

- Getting frequencies of each word. Finding the most popular words. Removing terms that don't appear very often from the model.

- Converting frequencies matrices in data frames. Pre-processing variable names.

- Review Week 1: Table. Creating new variables.

- Review Week 3: Splitting randomly a data set into a training set and a testing set. Setting a random seed. Computing a classification matrix and calculating accuracy. Logistic regressions.

- Review Week 4: Classification trees, and plotting them. Random forests.

**Dataset 28 - Unit 5, Recitation - Text Analytics into the Courtroom**

- Setting the language to default. Reading .csv documents when text analytics has to be applied

- Creating a corpus in R.

- Pre-processing the text in R.

- Getting frequencies of each word. Finding the most popular words. Removing terms that don't appear very often from the model.

- Converting frequencies matrices in data frames.

- Review Week 1: Table. Creating new variables.

- Review Week 3: Splitting randomly a data set into a training set and a testing set. Setting a random seed. Computing a classification matrix and calculating accuracy. Plotting the ROC curve and computing the AUC.

- Review Week 4: Classification trees, and plotting them.

**Dataset 29 - Unit 5, Homework 1 - Detecting vandalism on Wikipedia**

- Setting the language to default. Reading .csv documents when text analytics has to be applied

- Creating a corpus in R.

- Pre-processing the text in R.

- Creating frequency matrices. Removing terms that don't appear very often from the model.

- Converting frequencies matrices in data frames. Modificating variable names.

- Searching text in a data frame. Creating a variable associated with the text searched.

- Getting the row sums from a matrix.

- Review Week 1: Tables. Converting variables into factors. Combining two data frames. Adding new variables into a data frame.

- Review Week 3: Splitting randomly the data into a training and a testing set. Creating a classification matrix. Calculating accuracies.

- Review Week 4: Building a classification tree. Making predictions with it. Plotting it.

**Dataset 30 - Unit 5, Homework 2 - Automating Reviews in Medicine**

- Setting the language to default. Reading .csv documents when text analytics has to be applied

- Counting the number of characters in a text.

- Pre-processing the text in R.

- Creating frequency matrices. Removing terms that don't appear very often from the model.

- Converting frequencies matrices in data frames.

- Getting the column sums from a matrix. Pre-processing variable names.

- Review Week 1: Tables. Finding minimums/maximums. Combining two data frames. Adding new variables.

- Review Week 3: Splitting randomly the data into a training and a testing set. Creating a classification matrix. Calculating accuracies, sensitivities and specifities.

- Review Week 4: Building a classification tree and plotting it. Making predictions with it. Calculating the AUC of the model.

**Dataset 31 - Unit 5, Homework 3 + 4 (O) - Separating Spam from Ham**

- Setting the language to default. Reading .csv documents when text analytics has to be applied

- Counting the number of characters in a text.

- Pre-processing the text in R.

- Creating frequency matrices. Removing terms that don't appear very often from the model.

- Converting frequencies matrices in data frames.

- Getting the column sums from a matrix. Pre-processing variable names.

- Predicting probabilities with a random forest.

- Getting row sums.

- Review Week 1: Tables. Finding minimums/maximums. Histograms. Boxplots.

- Review Week 2: Dealing with skewed variables.

- Review Week 3: Splitting randomly the data into a training and a testing set. Creating a logistic regression model. Making predictions with it. Creating a classification matrix. Calculating accuracies. Calculating the AUC.

- Review Week 4: Building a classification tree and a random forest model. Plotting the tree. Making predictions with the tree. Calculating accuracy and the AUC of the tree. Calculating accuracy and AUC of the forest.

**Dataset 32 - Unit 6, Lecture 1 - Movies Recommendations**

- Concepts of collaborative filtering and content filtering. Concept of clustering.

- Concept of distance.

- Hierachical clustering and dendrograms.

- Getting data into R from .txt files. Naming variables. Removing variables from a data frame.

- Deleting duplicate entries from a data set.

- Computing distances in R. Doing hierachical clustering in R. Plotting a dendrogram.

- Cutting the data set into a particular number of clusters.

- Labelling the data points according to what cluster they belong.

- Searching in which cluster is a element from the data set. Searching which elements are contained into a cluster.

- Finding the cluster centroids when a lot of variables are involved.

- Review Week 1: Table. Tapply.

**Dataset 33 - Unit 6, Recitation - Segmenting Images to Create Data**

- Reading data corresponding to images into R.

- Converting a data frame into a matrix, and a matrix into a vector.

- Computing distances in R. Doing hierachical clustering in R. Plotting a dendrogram.

- Cutting the data set into a particular number of clusters.

- Visualizing the cluster cuts in the dendrogram.

- Changing the dimension of an object.

- Outputting images from a matrix in R.

- K-means clustering method concept.

- Running the K-means algorithm in R. Testing a K-means algorithm.

- Revising and comparing all the methods learnt through the course until this moment.

- Review Week 1: Tapply.

**Dataset 34 - Unit 6, Homework 1 - Document Clustering**

- Computing distances in R. Doing hierachical clustering in R. Plotting a dendrogram.

- Cutting the data set into a particular number of clusters.

- Finding maximum values in each cluster.

- Running the K-means algorithm in R.

- Review Week 1: Table.

**Dataset 35 - Unit 6, Homework 2 - Market Segmentation for Airlines**

- Normalising data, which is very useful when dealing with variables with different scales.

- Computing distances in R. Doing hierachical clustering in R. Plotting a dendrogram.

- Cutting the data set into a particular number of clusters.

- Finding where the centroids of every cluster are.

- Running the K-means algorithm in R. Finding the new centroids.

- Review Week 1: Table.

**Dataset 36 - Unit 6, Homework 3 - Predicting Stock Returns**

- Getting the means from every column.

- Removing variables from a data frame. Normalising data.

- Running the K-means algorithm in R. Testing the K-means algorithm.

- Making predictions with the cluster-then-predict method.

- Review Week 1: Table. Combining vectors.

- Review Week 2: Getting correlations.

- Review Week 3: Splitting randomly the data into a training and a testing set. Creating a logistic regression model. Making predictions with it. Creating a classification matrix. Calculating accuracies.