# Major assignment 2: Your submission

This is your assignment template for [AnalyticsX Major assignment 2](https://courses.edx.org/courses/course-v1:AdelaideX+AnalyticsX+3T2018/courseware/e0b375053b6441a08d658133752b3531/5965d914c42540e58416a76fb1f502ba/1?activate_block_id=block-v1%3AAdelaideX%2BAnalyticsX%2B3T2018%2Btype%40vertical%2Bblock%4000cc761eea174adbb4be63d3ec1c7484). Save this document on our local machine and include all of your work within the relevant sections. Once you’ve completed all five parts of the assignment, upload the document via the submission area on the “[Submit your assignment](https://courses.edx.org/courses/course-v1:AdelaideX+AnalyticsX+3T2018/courseware/e0b375053b6441a08d658133752b3531/5965d914c42540e58416a76fb1f502ba/6?activate_block_id=block-v1%3AAdelaideX%2BAnalyticsX%2B3T2018%2Btype%40vertical%2Bblock%40cb1643bf82444588a962216431740cd7)” page at the end of Major assignment 1.

# Checklist

* Have you shown all of your working?
* Have you given all numbers to the correct number of decimal places?
* Have you included all R output and plots to support your answers where necessary?
* Have you included all of your R code – input and output?
* Have you made sure that all plots and tables each have a meaningful caption?

**Quick links:**

[Major assignment 2: Part 1](#_Major_assignment_2:)

[Major assignment 2: Part 2](#_Major_assignment_2:_1)

[Major assignment 2: Part 3](#_Major_assignment_2:_2)

[Major assignment 2: Part 4](#_Major_assignment_2:_3)

# Major assignment 2: Part 1

1. Load in the flights dataset as a Spark table [4 points]

Your input code and output code go here:

1. Create a new categorical variable long\_flight which is 1 when a flight is further than 4,000 miles. [2 points]

Your input code and output code go here:

1. Perform a principle component analysis using the variables flight time, distance, and departure time. [3 points]

Your input code and output code go here:

1. Use ggplot to visualise the results. [3 points]

Your input code and output code, including your plot, go here:

1. Use your plot to create another plot to explain the observed clustering of the data. [2 points]

Your input code and output code, including your plot, go here:

# Major assignment 2: Part 2

1. Set the seed to be 42, then create a binary variable late\_arrival which is 1 if a flight arrives more than 30 minutes late. Count the number of late arrivals. [3 points]

Your input code and output code go here:

1. Use logistic regression to predict late arrivals based on carrier, departure delay, month, and year. [4 points]

Your input code and output code go here:

1. Produce a ROC curve for your model. [4 points]

Your input and output code goes here:

1. Remove each predictor variable individually to determine which is the best single predictor of late arrivals. [3 points]

Your input and output code go here:

# Major assignment 2: Part 3

1. (a) Load the wine dataset directly from GitHub using the command provided
2. Change the Wine variable to categorical.

[1 point]

Your input code and output code go here:

1. (a) Set the random number seed to 42 to cause reproducible training.
2. Use the caret and nnet package to classify Wine based on all other attributes, using default training/structure.

[2 points]

Your input and output code goes here:

1. Plot the trained network. [1 point]

Your input and output code goes here:

1. Predict the Wine variable from the trained network and produce a classification table of predicted vs actual Wine value/category. [2 points]

Your input and output code goes here:

# Major assignment 2: Part 4

1. Initiate a H2O session [1 point]

Your input code and output code go here:

1. Load the nycflights13 data as a data frame. [1 point]

Your input and output code goes here:

1. Write the data out to a .csv file and read it back again (this solves a data format issue). [1 point]

Your input and output code goes here:

1. Create a H2O data frame from the data. [1 point]

Your input and output code goes here:

1. Create a new categorical variable in the frame called late\_arrival that is 1 if arr\_delay is greater than or equal to 30. [1 point]

Your input and output code goes here:

1. Create a data split of the H2O frame with 80% of the data being used for training (flight\_train) and 20% used for testing (flight\_test), ensuring repeatability by setting the seed for the random sampling of the split to be 42. [1 point]

Your input and output code goes here:

1. Create a predictors set *predictors* from sched\_dep\_time, dep\_delay, air\_time, and distance. [1 point]

Your input and output code goes here:

1. **Create a deep learning H20 model using the defaults except for:**
   * **the predictors being predictors,**
   * **the response variable being late\_arrival,**
   * **the training data being flight\_train, and**
   * **the results forced to be reproducible with seed 42.**

Run several times to check your results are reproducible.  
[2 points]

Your input and output code goes here:

1. Plot an ROC curve for your model using the flight\_test data. [1 point]

Your input and output code goes here:

1. Calculate the prediction accuracy produced by your model using h2o.predict on your model and test data flight\_test. [1 point]

Your input and output code goes here:

1. Calculate the prediction accuracy that would be yielded, for the test data flight\_test, by simply using dep\_delay (being larger than 30) alone as a predictor. [1 point]

Your input and output code goes here: