Section 1 (approx. 20 mins)

Read in the "./MM.data" data set: Note The data is comma separated and does not contain column labels and answer the following questions. For the following tasks it is insisted that you use Python as the language throughout.

Document your code: use scripts, functions. Make sure your code is reproducible and readable.

Exploratory Analysis and Data Visualisation

1) How many features and observations are present in the data set?
2) Identify the only categorical variable; how many levels does it contain and what are the counts for each level?
3) We want to compare the density/distributions of the first, fourth and seventh variables in the data set (V1, V4 and V7) with each variable's density plot being split by the levels of the categorical variable.
    a. Each variables density plot, should contain n-density plots, where n is the number of levels the categorical variable has.
    b. Thus V1 will have n density plots, V4 will have n-density plots and V7 should have n-density plots.
4) We're interested in the variables that have the highest and lowest correlations in our data set.
    a. Identify the two most strongly correlated variables.
    b. Identify the least correlated variables.
    c. Produce a plot to represent the "correlation matrix".

Section 2 (approx. 45 mins)

Build a classifier which has the aim of predicting the categorical variable in the data set found in section 1.

1) You are free to choose any Machine Learning algorithm to produce a classifier.
2) Use any data transformations you deem appropriate (comment why if you use any)
3) Use any target metric you see fit (comment why)
4) Quote the performance of your classifier and produce an estimate of what the classifiers out of sample error would be. [If you're taking short cuts because of time considerations please do and outline what a solution with more time would entail.]
5) Produce a ROC curve for your predictions

Section 3: Business optimisation

Conduct any required exploration and leverage appropriate transformations and machine learning techniques to understand the factors which contribute to the Lifetime impact of 12 different Facebook posts.

The input features you have concerning the posts are (category, page total likes, type, month, hour, weekday, paid). These are the features you can model against. The remaining features are all performance metrics which you are predicting.

1) Your output is a reproducible analysis of the problem at hand.
2) Classifiers which predict the target metrics of interest
3) Recommendations concerning what features future facebook posts should have in order to maximise their performance metrics.
4) A small presentation (slides) summarising your findings and recommendations.

(data attached 'dataset_Facebook' citation: Moro et al., 2016) S. Moro, P. Rita and B. Vala. Predicting social media performance metrics and evaluation.)