EDA

# EDA Baseline Process

Focus on a few packages

* funModeling
* Caret
* Others?

Identify 1 or 2 datasets to serve as examples

* Titanic?
* schedulingData? (in AppliedPredictiveModeling)
* Want to add in NA and NULLs to make sure they are managed  
  isnothing = function(x) {any(is.null(x)) || any(is.na(x)) || any(is.nan(x)}

Use ClubReady as initial guide

* Add lessons from DCX
* Add correlation analysis/mutlicollinearity  
  <http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software>
  + corrplot
  + Cor.test
  + rcorr from Hmisc

Miscellaneous/Random Thoughts

* Feature importance
* Factor relevel
* One hot encoding
* Unbalanced data

Document any requirements for using any functions. Example: rcorr requires the input to be a matrix (therefore need to filter out any char or factors)

Templates for R will be created in RMarkdown; Python templates will be produced in Notebooks.

* Should R be in Notebooks too for consistency? Could use Azure Notebooks for both languages.
* RStudio Notebooks support Python (and other languages). Also then provides all the output options.  
  <https://stackoverflow.com/questions/40512587/rstudio-notebook-rmarkdown-running-python-describe-not-shown>

ML Ensemble Routines

# R Code

<https://cran.r-project.org/web/packages/caretEnsemble/vignettes/caretEnsemble-intro.html>

<https://github.com/tobigithub/caret-machine-learning/blob/master/caret-cv/caret-all-cv-methods-lapply-sapply.R>

<https://amunategui.github.io/blending-models/>

<https://dzone.com/articles/a-trial-run-with-h2o-automl-automatic-machine-lear>

<https://github.com/ecpolley/SuperLearner>

<https://machinelearningmastery.com/machine-learning-ensembles-with-r/>

<https://dzone.com/articles/build-custom-ensemble-models-using-caret-in-r>

# Python

<https://www.kdnuggets.com/2017/01/current-state-automated-machine-learning.html>

<https://github.com/AxeldeRomblay/MLBox>

# Packages

* Caret
* modelR