Questions:

* Can we predict an NHL goalie’s future sv% based on other statistics? If so, which stats are the most predictive of future success (in the form of sv%)
* Are goalies’ previous 5 games an indicator of success in a given game?

Data:

* <https://moneypuck.com/goalies.htm>
* Could use any year + the next year’s sv%
* Use previous 10 (5?) seasons to train model? 2011-2012 to 2021-2022?
  + Could use 2022-2023 for testing, and then predict 2023-2024?

Plan:

* Format data into one spreadsheet including all years? Add year variable?
* Build machine learning model to predict sv% for year x using stats from year (x - 1)
  + Random forest? Tree-based models can give feature importances with forest.feature\_importances\_

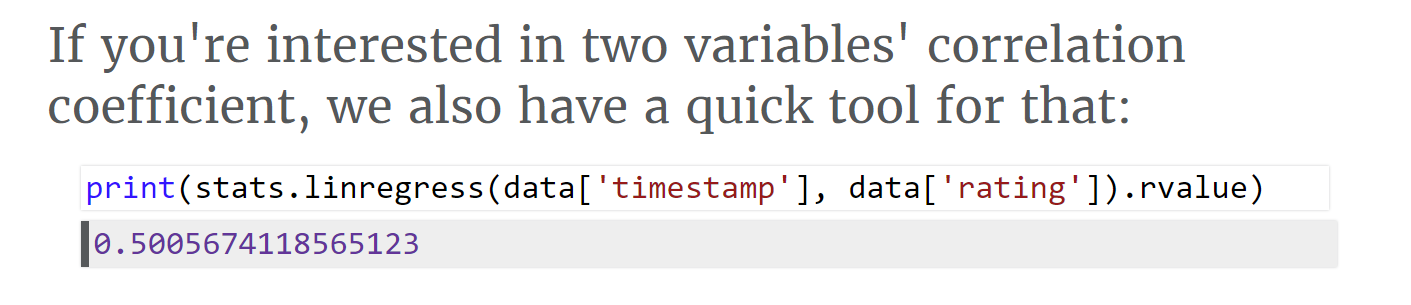
1. Combine 2017-18 to 2021-22 data into one dataset, add column for year, add column for next year’s save percentage

* Filter out goalies with less than 20 games played and rows where ‘situation’ ≠ ‘all’
* Remove NA’s (goalies who retired in year A would have no sv% recorded for year B)
* Each goalie will have a row per season where they played 20+ games (unless they didn’t play 20+ games in the following season and therefore don’t have a ‘next year sv%’)
* Remove any unnecessary columns (eg. position, team, etc.)
* Format 2022-23 data the same way, with year column at least?

1. Create random forest model

* scaler?
* boosted?
* Split training data into training/validation sets (params = everything except ‘next\_yr\_sv\_pct’, response = ‘next\_yr\_sv\_pct’)
* Tweak parameters (n\_estimators, max\_depth, min\_samples\_leaf, etc)

1. (optional) run T tests on the importance of specific variables (maybe RF does this already?)

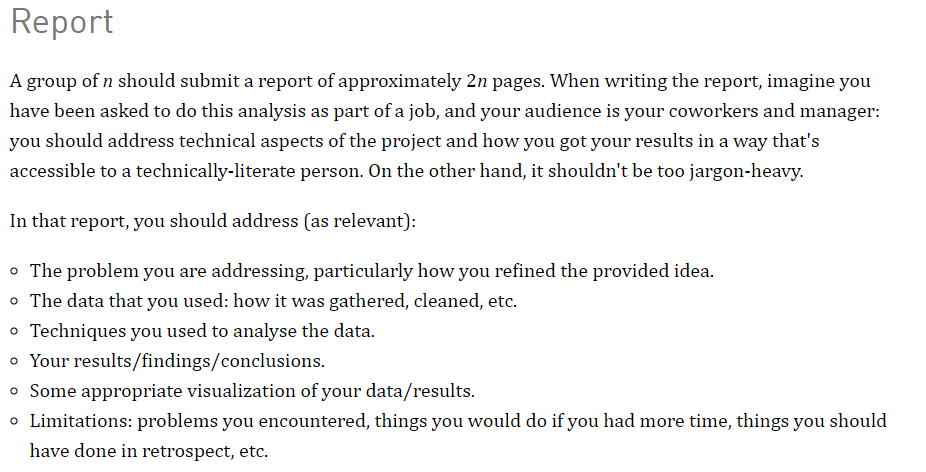
* Correlation coefficients for X’s vs Y?
* 
* Note: ONLY FOR LINEAR relationships ^^^

1. Visualize data

* Predicted vs actual save percentage for following year?
* Importance of different variables? (how feature\_importance weighted them)
  + Which stats did we find to be most important?

1. (optional) look into if the previous 5 games have a particularly strong influence on the next game
2. Write report

* ~4pg
* 12pt TNR, single spaced
* Audience is somewhat familiar with methods used but not necessarily the content itself
* (Turn below bullet points into subheadings?)



1. Write project experience summary (basically resume blurb)