Log:

* 11/4/2021
  + Inspected dataset: looks like we will have high cardinality with many categorical data. Did research into what we can do to reduce the cardinality and reduce dimensionality since just doing OHE will face sparse dataset and have Curse of Dimensionality and may easily overfit our models.
  + Options:
    - Group and cluster our categorical features that have many classes
    - Continue with OHE and use strong L1/L2 regularization
    - Catboost
    - MCA/FAMD
    - Frequency thresholds
    - Grouping / clustering
  + Thinking about doing MCA/FAMD or clustering
* 11/10/21
  + Started pipeline for models with models.py and models\_helper.py
  + Did feature selection with team on: <https://docs.google.com/spreadsheets/d/1qrFCjBWOn3emAx8xtMGC3bpg0CQsx3LMaK1O56ReDO0/edit?usp=sharing>
  + Will use Random Forest to check for Feature importance
    - Then do linear regression with most important features as baseline
    - Then do RandomForest & GradBoost & AdaBoost then tune/cross validate and use best model