2/12/2019

George

- Looked at dataset in entirety (160GB raw)
- proposal to look at sample (1%?)
- Dataset for 2015-2018 is on BigQuery
- generate summary stats
- George will create a sample
- Action: explore sample dataset
  - Rundown of variables/types
  - distributions of variables
  - normalizations
- Where to mount the data
- Rstudio server?
- locally?
-

ACTION ITEMS

- Feature selection: Look at sample to see which variables are important
- Building new variables
- Writing exploration code so that we can plug in future data and get graphs etc.
- Everyone: Look at sample projects on Piazza
- Simeon: Look for shapefile to add neighborhood/census block
- George: Create 1% sample
- Shiv: Define exact problem we are solving
- Trace: Naive model to predict price based on other vars (track how closely this corresponds to NYC taxi fare rules)