You asked by a stakeholder to answer a problem with machine learning. You must explore and prepare a dataset to answer the question your stakeholder asks.

Find an open source dataset: you can use kaggle, US open data, NYC data, kalshi, etc. Any dataset so long as you can share the dataset with the class. (You may not use a proprietary dataset from work that you do not have permission to share publicly)

Define the problem you are answering: who is your stakeholder? What is the problem they have asked of you? [Ensure the problem you are answering can be addressed by the dataset you chose]

What is the target variable (what are you trying to predict)? What are the features you are going to use?

Explore the data to understand the distribution, are there missing values, are there outliers?

Use feature engineering techniques that we discussed in class or that you found in the readings, on papers, books, etc. to prepare your dataset for learning/prediction.

Discuss the techniques you chose and why you chose them, how will you handle new values, missing data, or outliers when you deploy to production and start transforming new inputs?

NOTE: You are not obligated to do any modeling for this assignment but if you would like to, or if you want to iterate on your features by evaluating them with a model you may.

The rubric on which you will be assessed on is here: <https://yu.instructure.com/courses/63465/rubrics/21352>.

Acceptance Criteria:

A written report (could be in a Jupyter Notebook or document) explaining:

Who is your stakeholder, what questions are you trying to answer, what is your target variable and what are your features

Explore the dataset and comment on patterns in the features you'll need to handle (missingness, outliers, out-of-vocabulary categories, etc.)

Discuss the feature transformations you made and what are some of the pros and cons of that transformation (sparsity, interpretability, size, etc.)

Include your code or a link to where your code is on github (git preferred)

Your code should be clean, readable, and reproducible

Use meaningful variable names and function names

Use functions where possible

Avoid magic numbers

I should be able to run your code (it's not the end of the world if I would need to change a file path in one place, but ideally it would pull from a URL)

No large chunks of commented out code

Use comments appropriately

You may use any open source programming language (Python, R, Julia, C++, pony, whatever...) but not if it requires a licensing (looking at you Matlab and SAS)

Final Note: have fun! You are free to be creative in what your "stakeholder" is or what your object is. It could be your boss, it could be a CEO you are consulting for, it could be a professor. You should be able to find a dataset that interests you and you'll have fun exploring and working with.