Potential Datasets and Questions

1. **Coffee Chains Dataset:**
   https://www.kaggle.com/datasets/arjunbhaybhang/coffee-chains-dataset
   a. Chains? → area code
   Question : prediction of sales (as regards date, type of coffee, …)
   Potentially a polynomial regression to see how the different variables combination affect the sales…

2. **Coffee Quality Dataset:**
   https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi
   a. Potentially a linear regression (?)
   b. Interested in quality → kinda a subjective measure (would this be a problem?)
   c. Files for the feature descriptions: https://database.coffeeinstitute.org/files
   Question : What makes a good coffee ? (try to predict the Total.cup.points as regards the taste variables…)

3. **Student Alcohol Dataset:**
   https://www.kaggle.com/datasets/uciml/student-alcohol-consumption
   Funny dataset
   Question : Can we predict alcohol consumption as regards the other variables given.

Proposal:

Title and Authors:

An inquiry on Taste for Future Coffee World Domination 😈 (Note: This is not *temporary*)

Alexandra Ramassamy (alexandra_ramassamy@fas.harvard.edu)
Helen Zhao (hzhao@g.harvard.edu)
Katherine Hunter (katherine.hunter@g.harvard.edu)
Alyssa Chang (yujie_chang@g.harvard.edu)
Kushagra Chitkara (kushagrachitkara@g.harvard.edu)

Background and Motivation:
Undoubtedly, coffee consumption is prevalent in adults as sleep deprivation continues to be a concern for the unemployed graduate student. For the five of us, our reliance on coffee extends beyond the mere milligram intake of caffeine, and into the satisfaction gained through sensory input (on our taste buds). In order to assert dominance, our collaborative team of five motivated Master's students aim to start our own coffee chain (sometime in the *near* foreseeable future). However, we understand the coffee beans themselves are the soul to a good cup of coffee. But what else contributes to the elixir that wakes us up in the morning? This is our goal, to gain a thorough understanding of coffee beans and the features that are the best predictors of quality. Finally, we hope to create our own test dataset which will consist of coffee found around Harvard (and the SEC), and try to analyze where it ranks.

Data:

The dataset we are interested in is the Coffee Quality dataset (https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi).

It contains information on different kinds of coffee (linked to origins, taste, ratings,...). Our goal would be to study what makes a good coffee and we plan to use the different taste variables (aroma, balance, acidity,..) in the dataset and the one linked to the origin of coffee.

The dataset contains many categorical variables that will require pre-processing and missing values.

Scope:
Our project will focus on assessing coffee quality according to expert coffee tasters at CQI. Given that this is a subjective response, we are interested in the specific features that contribute to the taster's assessment of the coffee. For model interpretability, we will apply multiple linear

regression, focusing on our feature selection due to the high dimensionality of the data, we might include Propensity Matching Score as robustness check. Other methods we intend to apply are random forest regression, principal component analysis, and potentially a neural network to assess whether there are aspects of the features that contribute to the prediction that were not necessarily captured by the simpler models.

Potential issue: there are other metrics (e.g. Sweet Maria's), maybe we can compare the similarity between them and Q-score as a robustness check. But the two metrics might be very different.
(https://towardsdatascience.com/specialty-coffee-comparing-grading-methods-36777cae220f)