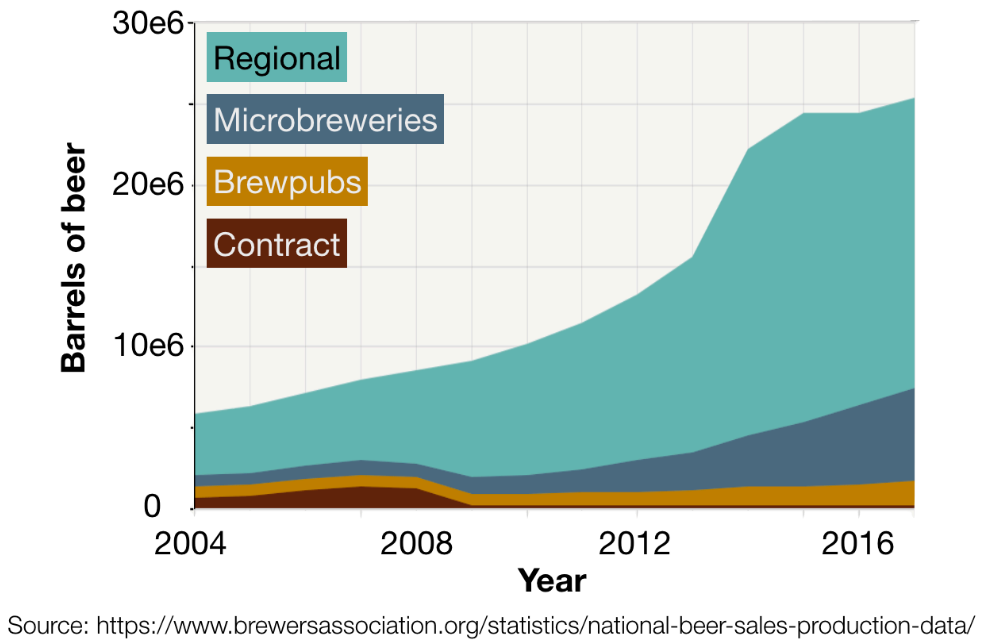
COURSERA / IBM

DATA SCIENCE CAPSTONE

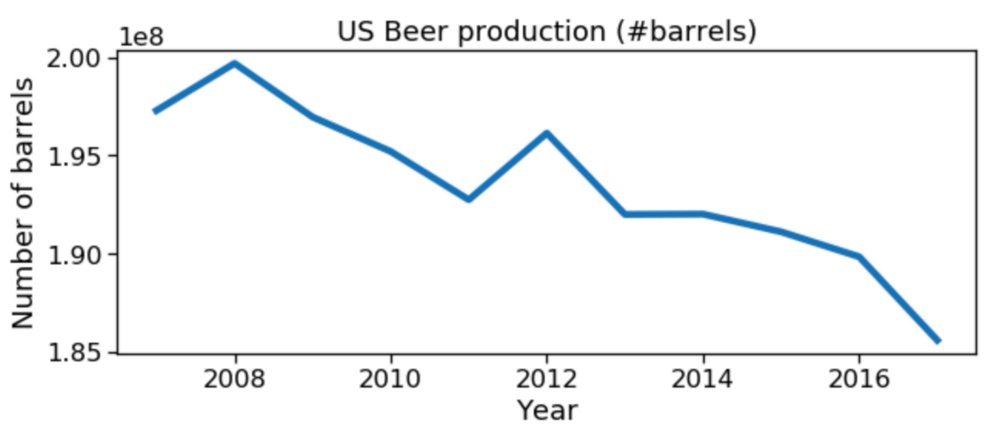
THE BREWERY PROJECT

THE BEER INDUSTRY IS CHANGING

- **Craft breweries are on the rise.**
- According to the Brewers Association website, craft beer:
  - continues to increase in terms of production and market share
  - despite a small overall decrease in beer production
  - now accounts for nearly a quarter of the US beer market



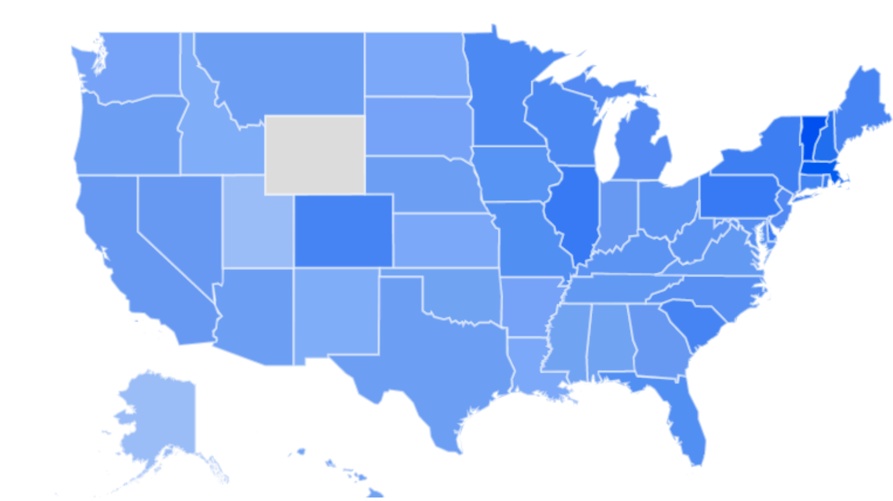
Source: https://www.brewersassociation.org/statistics/national-beer-sales-production-data/



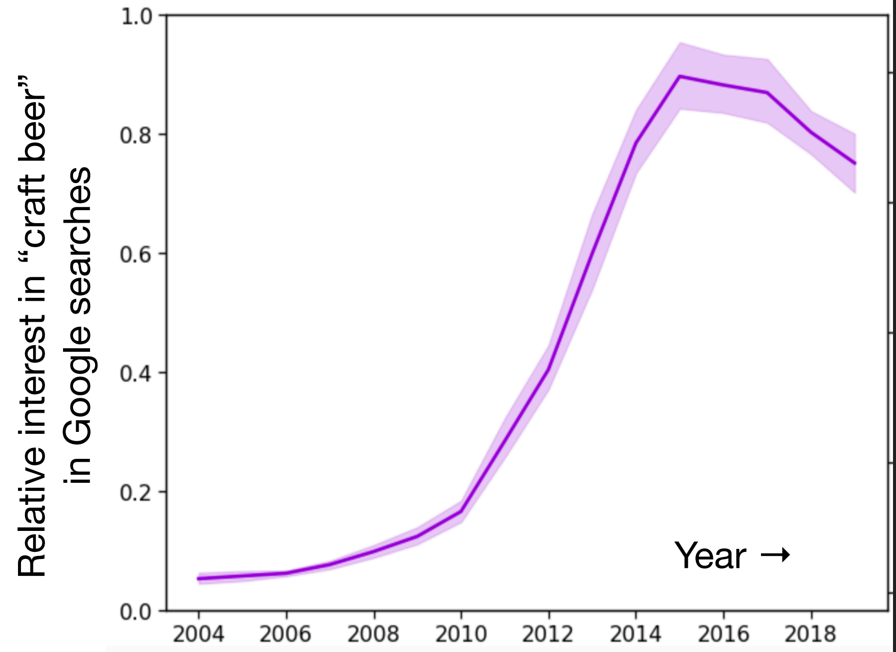
Source: brewersassociation.com, https://www.ttb.gov/beer/

THE BEER INDUSTRY IS CHANGING

- **People are generally more interested in craft beer:**
- Google searches for "craft beer"
  - rose rapidly between 2011 and 2015
  - is more localized in the Eastern parts of the US



State-wise relative interest in "craft beer" in Google searches between 2004-2018,



Source: trends.google.com

TARGET AUDIENCE

- Investors and beer enthusiasts looking to be part of the changing US beer landscape

THE CHALLENGE

- How to help potential stakeholders determine the LOCATION of a new brewey and WHAT to brew there.

APPROACH

- Determine which beers are being most appreciated by beer enthusiasts.
- Extract various features of these well-appreciated beers
- Use machine learning (ML) to identify the relation between various extracted features and how much a beer is appreciated by enthusiasts
- Make recommendations based on observed ML model parameters.

OPERATIONALIZING THE APPROACH: DEPENDENT VARIABLE

- Collect Best Beers data from BeerAdvocate.com
    - **Beer score**
        - The dependent variable, beer score is derived from weighted user ratings, using a Bayesian model*
        - This beer score will be the dependent variable to be predicted from the other variables.
    - Beer name and brewery name

***Source**: https://www.beeradvocate.com/community/threads/top-rated-beers-explained.587593/

OPERATIONALIZING THE APPROACH: PREDICTORS

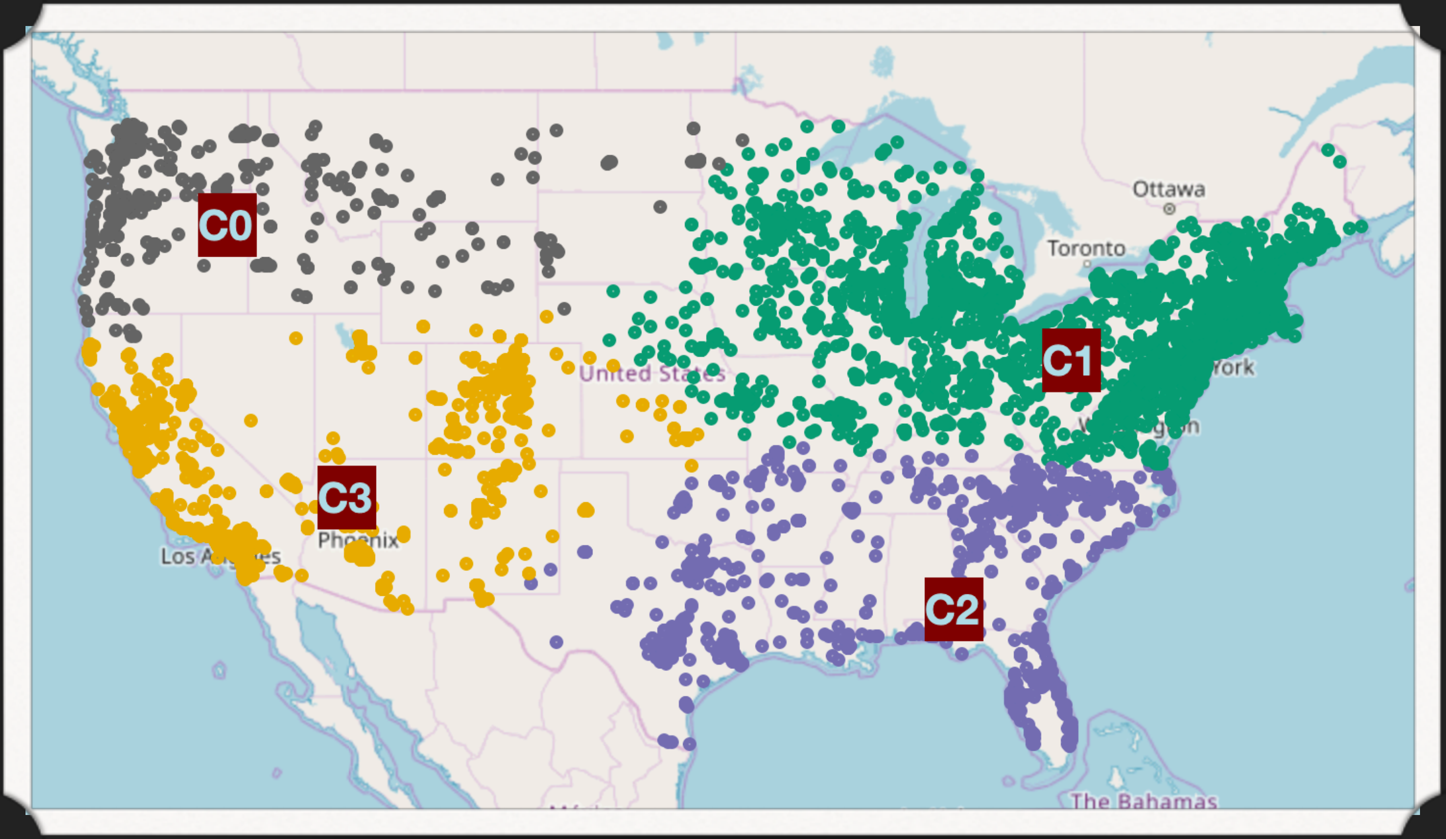
- Predictors are the properties of the beers and the breweries producing these beers that might contribute to a beer's score.
- Predictors from BeerAdvocate.com
  - **Style of beer:** e.g., Belgian Saison, New England IPA, etc.
  - **ABV%:** Alcohol content of the Beer by Volume, expressed as a percentage.
  - **Number of ratings:** how many users contributed to the score

OPERATIONALIZING THE APPROACH: PREDICTORS

- Other predictors
  - **Local brewery density:** a count of all other breweries within a 10Km radius of the brewery producing the particular beer, gathered using the Foursquare API*.
  - **Geographical region:** operationalized as cluster membership, where the clusters are derived from an automated classification and prediction algorithm based on latitude/longitude coordinates of US breweries.

***Source**: https://developer.foursquare.com/

GEOGRAPHICAL CLUSTERS



- **Geographical regions:** ~6,900 US breweries retrieved from brewersassociation.com were clustered based on their latitude, longitude coordinates using KMeans clustering. Cluster labels for each region are positioned at the mean coordinates for that region.

***Source**: https://www.beeradvocate.com/community/threads/top-rated-beers-explained.587593/

MACHINE LEARNING: WHAT PREDICTS A HIGH BEER SCORE?

- Linear regression models used to determine how the various variables contribute towards an observed beer score.
- General model:

$$Score \propto ABV + Num\_ratings + Num\_nearby\_breweries + Geo\_cluster + Style$$

  - i.e., predicting the score as a function of ABV, number of ratings, Number of nearby breweries, geographical cluster, and style of beer

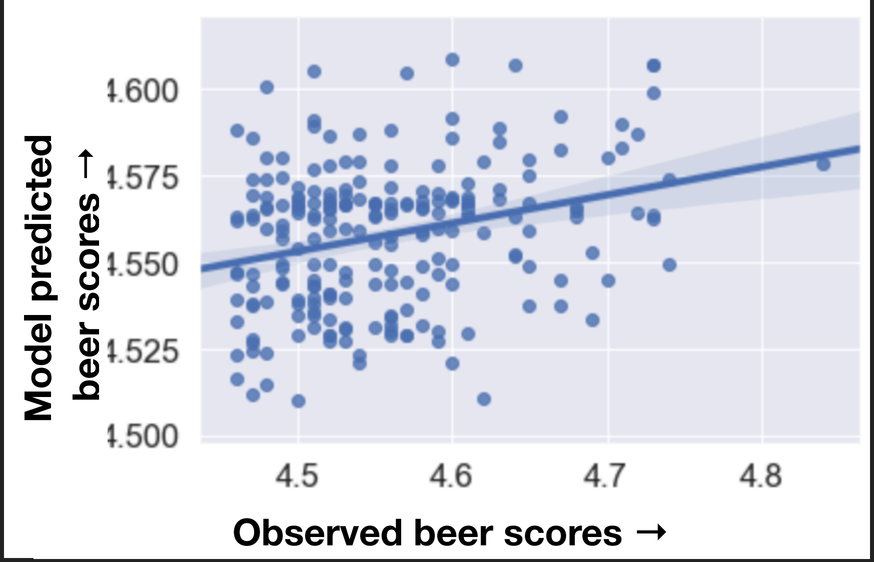
***Source**: https://developer.foursquare.com/

OVERALL MODEL FIT

- An Ordinary Least Squares model showed the best fit of the data:

- Root-mean-square error of 0.0694

- Significant correlation between observed and predicted beer scores
  - Pearson's R = 0.285, *P*=1.9e-5



- These findings indicate that the model was partially successful at explaining what makes a beer score better

HOW DO THE DIFFERENT PREDICTORS PERFORM?

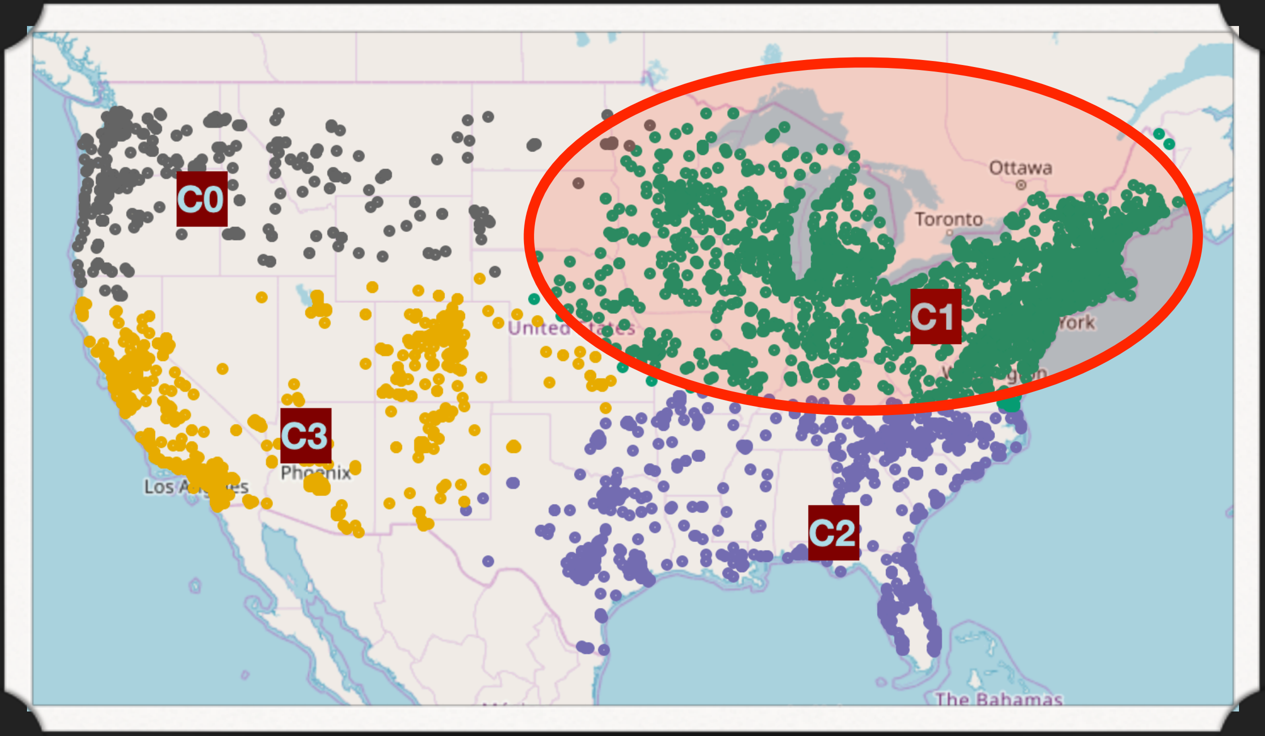
- The coefficients of the linear regression indicate the following:
- **Beer characteristics**
  - the alcohol content of beer is not a major factor
  - instead, specific styles, such as American Imperial Stouts, Belgian Saisons, American Wild Ales, and New England India Pale Ales are positively correlated with high scores
  - other styles, in particular Russian Imperial Stouts, ae associatd with lower scores

HOW DO THE DIFFERENT PREDICTORS PERFORM? (CONTD..)

- **Location characteristics:**
    - belonging to geographic cluster 1, centered around North-east US, carries the most weight in predicting a higher beer score.
    - number of other nearby breweries has a very small, negative effect on beer scores.

WHERE SHOULD A STAKEHOLDER OPEN A NEW BREWERY?

- **Somewhere in North-Eastern USA!**
  - New England, New York, Pennsylvania, the Great Lakes



WHAT KIND OF BEER SHOULD THEY BREW?

- **An American Imperial Stout!**
  - E.g., Hunahpu's Imperial Stout - Double Barrel Aged (Cigar City Brewing, Florida)

Image credit: https://www.flickr.com/photos/adamjackson/12969271133