The Brewery Project Report

CU Boulder - Spring 2024

Jessica Groven - Julia Gallwitz - Carl Klein

Table of contents

1 Abstract

Breweries have become popular socializing spaces due to their tasty craft beers, cozy environments, and fun activities like trivia and board games. Yet the frequency of breweries varies greatly from location to location, with some cities appearing to be "hotspots" for breweries and other cities seeming to arbitrarily lack a notable frequency of breweries. The goal of The Brewery Project is to assess the various features that contribute to brewery frequency in US cities. In order to evaluate this, we gathered the most comprehensive data available on brewery locations in the US, as well as data on outdoor attractions, city classification features, and census data. We built various models with the goal of classifying brewery hotspots, observing the impact of various features on brewery frequency, and determining the most influential features on brewery frequency. We found that a city being a tech hub had a significant influence on a city's brewery frequency, as well as the age and race demographics of the city, and finally the types of existing breweries (e.g. microbrewery, brewpub). These findings can be used to inform potential future brewery operators on the best new locations to open up a brewery.

---

2 Introduction

2.1 What Is The Project?

The Brewery Project takes an in-depth look into the world of Breweries in the US. This team will explore features of brewery locations and the common demographics of successful brewing towns.

The Brewery Project will start with data collection on records of breweries in the US and towns with specific features that we hypothesize to have a higher brewery count. From our own experience, we believe that city features that may be linked to more support of the brewing industry could include major cities, tech hubs, college towns, and proximity to outdoor recreation such as ski resorts or National Parks. We will begin to review if any of the factors promote breweries. After an analysis to see if these features do link to higher brewery count in cities, we will apply a ranking system to classify cities on a scale from "No Hotspot" to "Hotspot". We aim to classify each of our cities into tiers of hotspot using these basic city features. These features may also be intertwined with population demographics, so we will then expand our exploration into additional population factors. The Brewery Project will conclude with actionable results that can guide future decisions on where potential or current business owners may want to set up shop next.

2.2 Why Does The Project Matter?

The Brewery Project gives insight into where brewery hotspots are and how they are associated with communities and activities.We hope to identify key areas that have taken strongly to the Brewery life as this may point us in the direction of identifying new towns that could support similar businesses.

The composition of breweries (micro to large commercial) and the kinds of populations/areas that support local breweries are important to study. This information can ultimately inform business decisions in the brewing industry if you are a new or experienced brewery looking for your next potential location.

2.3 Related Work

There are several websites, databases, and applications associated with individual beer brands, breweries, and venues.

- UNTAPPD [1] allows the community to rate beers and breweries. They maintain a robust database which helps showcase brewery ratings. Previously, they had API access to use, but this has been discontinued.
- Open Brewery DB [2] is a completely open-source database focused around brewery-related data. They maintain API access, a GitHub page with international brewery data available for download, and even a Discord for this community to discuss their projects surrounding this open-source data.
- The Brewers Association [3] aims to "promote and protect American craft brewers, their beers, and the community of brewing enthusiasts." Their website features detailed exploratory analysis into the growth and other aspects of the brewing industry.

In summary, many of the current applications have been initial data compilations of craft beer and breweries, and apps focused on a social media or geographical recommendation focus.

2.4 What Can Still Be Done?

Through our research on the topic, we have explored various databases and projects relating to beers and breweries. We found that the existing databases and projects primarily focus on exploratory data analysis. However, we have yet to find any existing project utilizing a predictive model in the beer and brewery space. We intend to build a predictive model for prospective beer producers to find spaces to enter the market.

To build this model, we intend to utilize the various available datasets related to brewery locations and information, local natural attractions, schools, tech hubs, and population/census information. We will aggregate the data and implement data mining techniques to build a model to advise potential business owners on their future locations.

---

3 Methods, Evaluation, and Results

3.1 Data Collection

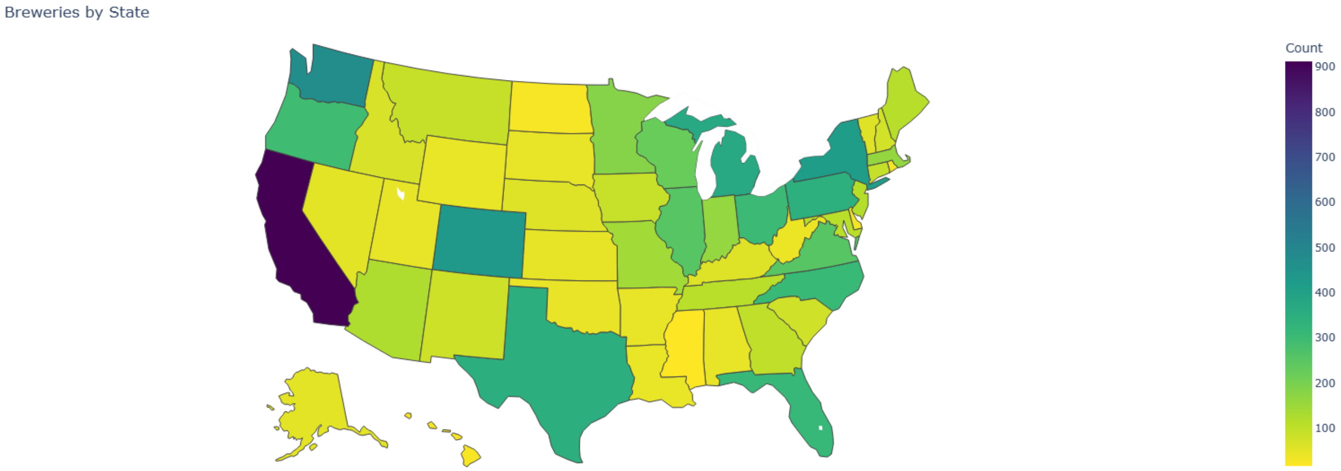
Data for The Brewery Project was collected from multiple web sources to capture information surrounding documented US breweries, cities fitting our key features (college towns, tech hubs, major cities, proximity to outdoor recreation), and population census data. The following sources were used in this endeavor:

1. Open Brewery DB [2]: An API database of US breweries and locations
2. Top College Towns [4]: A list of the top 150 college towns in the US
3. Major Cities [5]: A list of US cities with highest population
4. Tech Hubs [6]: A list of US tech hubs and key characteristics
5. National Parks [7]: A list of all US national parks, their yearly visitor count, and location
6. Ski resorts [8]: A list of all US ski resorts and their location
7. Census data [9]: Data collected from the 2020 US Decennial Census

These data sources were cleaned and transformed into a master data set used for the remainder of the project.

3.2 Data Exploration

With this aggregated dataset, we explored different features we thought might be associated with our research goals, using different types of visualizations to test these hypotheses. One of these visualizations was a heatmap of the number of breweries per state in the US.



For a detailed exploratory analysis, please see this page on our website.

3.3 Models Implemented

Here we consolidate our data and results from our exploratory analyses to answer our research questions and ultimately provide for multiple facets within the brewery community. For a detailed process, please see this page on our website.

3.3.1 Consolidated Data

We merged together datasets featured in our exploratory analysis into a format conducive to begin building models.

Most columns are boolean or numeric based already, but some columns were kept in categorical format for additional exploratory analysis.

Additionally, some entries contain missing values that we've kept in effort to prevent information loss. We'll test different subsets of columns and rows to produce the best models.

The consolidation process can be found here.

3.3.2 Addressing Frequency

Through our exploratory analysis, we identified 5 different city features that may impact the frequency of breweries. The features we identified were the number of ski resorts in the state that the brewery is located in, number of national parks in the state that the brewery is located in, and whether the location is considered a tech hub, a college town, or a major city. We represented tech hubs, college towns, and major cities with Boolean values where a 1 meant the location of the brewery did meet that characteristic and a 0 meant the location did not meet that characteristic.

3.3.3 Principal Component Analysis (PCA)

First, we began by implementing a PCA analysis to check the relationships between our variables and identify any redundancy in our dataset.

After some preprocessing, The PCA process was then implemented, resulting in the individual variance explained by each principal component.

The principal components were then mapped against each other.

Our analysis showed that there is some sort of pattern going on within our dataset. Our findings indicate that there is clustering within the data and subsets of data that share similar characteristics.

3.3.4 Linear Regression

To analyze our model further, we began with implementing a linear regression of the 5 identified city features mentioned above. Using **scikit-learn's** linear regression feature, a regression model was identified with the following coefficients to predict city brewery count.

We can see that being a **tech hub** was associated with the greatest increase in brewery count. Holding all features constant, being a tech city is associated with an increase of about 27 breweries on average. The next most influential feature appears to be whether or not the city was considered a **major city**. We see that being a major city is associated with an increase of about 6 breweries on average. Next, we see that being a **college town** is associated with an increase of almost 3 breweries on average. The relationships between the number of ski resorts in a state and breweries, and the number of national parks in the state and breweries are positive but seem to be quite small.

Ultimately, this gave us some insight into how our 5 selected features relate to the number of breweries in a city, but the details of the model were minimal. To dive a bit deeper into the details, we implemented linear regression using **statsmodels** OLS module.

Interestingly, our R^2 isn't very high at 0.459 and two of our coefficients are insignificant at the $\alpha = 0.05$ level. Our **MSPE** for the full model was calculated to be 17.183. We implemented backwards selection to see if the model would improve. The first step is to remove `state_national_park_count` as that feature has the highest p-value.

Our R^2 for this reduced model remains at 0.459 and the **MSPE** was calculated to be 17.199. However we are still seeing that one of our coefficients, `ski_resort_count`, is statistically insignificant. The next step in the backward selection process is to remove that feature.

Here we see that all of our features are statistically significant, but our R^2 for the final reduced model has decreased slightly to 0.458 and the **MSPE** for this model was calculated to be 17.283.

Overall, we see that our model captures around 45.8% of the variability in the `city_brewery_count`. It appears that some of the city features we identified as impacting brewery count are likely to be influential, but do not capture the entire story.

Namely, out of our 5 starting features, **brewery count per city** is most influenced by **tech hubs**, **major cities**, and **college towns**.

Note that the best fitting model was saved as a pickle file, and can be found here.

Next, we will look at other model types to explore how we can improve our analyses.

3.3.5 Defining a Hotspot

Next, we will investigate the power of the city features and population in identifying brewery hotspots. For our training and testing, we need to identify a measure to quantify these hot spots.

After considering a few methods, this analysis proceeds using a rank system based on total brewery count by city with one notable exception: all cities with less than 3 breweries will be classified a 1 on the 1-6 Brewery Hotspot ranking *(1:not a hotspot, 6:major hotspot)*. The remainder of our analysis will use this system to assess our model's ability to classify towns into the proper ranks. We hope to be able to clearly distinguish cities with a high rank of 6, indicating a sure brewery hotspot.

3.3.6 Preparation for Modeling

Returning back to our initial dataset, we'll perform classification using more features to see if we can create a better performing model.

We will want to remove the following features in preparation for modeling: **city**, **state**, **2021 median income**, **brewery__concentration**, **per__capita__ranked**, **city__brewery__count**, and rows with incomplete census data. We will also encode the **region** into 3 boolean columns.

After this process, our dataset is ready for modeling.

3.3.7 Simple Classification

Now that we have established a metric for quantifying how much of a hotspot each city is, we can proceed to implementing models that can train on this data. We start by asking if common city features (college towns, tech hubs, major cities, national parks, and ski resorts) and population can help predict if a town is high or low on the brewery hotspot scale *(1:low - 6:high)*.

To start our analysis, we will reduce our master data set to only the necessary columns. Using this data, we implemented preliminary modeling using the following techniques: **Decision Trees**, **Logistic Regression**, **K Nearest Neighbors**, **SVM**, **Naive Bayes**, **Linear Discriminant Analysis**.

This work can be reviewed here. An initial unrefined model of each of these was used and the final results were compared.

The metrics we'll be reporting on are:

- **Accuracy**: the ratio of correct predictions to all predictions

- **Precision**: the ratio of true positives to the total number of positives (a measure of exactness)
- **Recall**: the ratio of true positives to the number of total correct predictions (a measure of completeness)
- **F1-Score**: the *harmonic mean* between precision and recall (a balanced combination, or equal weight, of both precision and recall)

Naive Bayes, **SVM**, and **Linear Discriminant Analysis** were the top performers in the initial testing based on their F1-Score. We followed these tests with more detailed hypertuning to maximize their performance. We utilized the `GridSearchCV` function to run each model under a series of potential parameters settings.

The best parameters for the **Naive Bayes** classification were:

- **var_smoothing**: 1e-06

The best parameters for the **Support Vector Machine (SVM)** classification were:

- **class_weight**: None
- **degree**: 2
- **kernel**: rbf

The best parameters for the **Linear Discriminant Analysis** classification were:

- **shrinkage**: 0.5
- **solver**: lsqr

The parameters of each best performing model were used in final refined models and yielded these results using the test data.

The **Naive Bayes** classification model performed the best out of the front runners based on the F1-Score, although each of these models are performing quite similarly in terms of accuracy, precision and recall.

The results here show that our models can generally calculate accurate city ranking in about 85-86% of cases. This is decent, but opens up a conversation on using more of our available data to classify these cities.

> Note that the best fitting model was saved as a pickle file, and can be found here.

In the next section, we will expand our models to use additional data [9].

3.3.8 Classifying Brewery Hotspots - Full Dataset Features

The complete classification modeling process can be found here.

Our modeling process will consist of:

- Performing default algorithms, reporting metrics
- Applying scaling, performing default algorithms, reporting metrics
- Running hypertuning on the top performing default algorithm(s)

Using the **sklearn** library, we tested the previously mentioned default models.

3.3.8.1 Default Algorithms

After performing the default algorithms, we found that the `DecisionTreeClassifier()` performed the best across all of our metrics for a non-scaled data single-run. Several of the other classifiers also performed well. *Note that even though this was supposed to be strictly a default algorithm section,* `LogisticRegression()` *required a change to the* `max_iter` *parameter to even run.*

3.3.8.2 Default Algorithms with Scaled Data

Next, we'll perform the same process with the default algorithms, this time using scaled data applied with `StandardScaler()`.

After performing the default algorithms, we found that the `LogisticRegression()` performed the best across all of our metrics for a scaled data single-run. Several of the other classifiers also performed well. *Note that even though this was supposed to be strictly a default algorithm section,* `LogisticRegression()` *required a change to the* `max_iter` *parameter to even run.*

To summarize, `DecisionTreeClassifier()` performed the best for a non-scaled data single-run and `LogisticRegression()` performed the best for a scaled data single-run.

3.3.8.3 Parameter Hypertuning

Using the **sklearn** library, we can use `GridSearchCV()` to test a multitude of different combinations for algorithms for our best default models.

We deduced the best parameters for `DecisionTreeClassifier()`:

- **criterion**: entropy
- **splitter**: best
- **max_depth**: 6
- **max_features**: None
- **class_weight**: None

Namely, using *entropy* as the criterion and *max_depth* of 6 resulted in the best best model for the `DecisionTreeClassifier()` (non-scaled data).

We deduced the best parameters for `LogisticRegression()`:

- **class_weight**: None
- **max_iter**: 10000
- **multi_class**: auto
- **penalty**: l1
- **solver**: saga

Namely, using *max_iter* of 10000, *penalty* of l1, and *solver* as saga resulted in the best best model for the `LogisticRegression()` (scaled data).

Overall, `LogisticRegression()` with scaled data performs the best across all Accuracy, Precision, Recall, and F1-Score with scores in the mid-90s.

One final note is that the overall scores for the *best* `DecisionTreeClassifier()` are lower than the originally run model. This is due to how `GridSearchCV()` tests models, which is through cross validation. The default (which we had ran the models through) uses 5-fold cross validation. This means that the *initial data is randomly partitioned into 5 mutually exclusive subsets (folds), each of approximately equal size, and then training and testing are performed 5 times* [10]. This helps in assuring a better model than just using a single split. Although not produced during the timeframe of this project, with how the scores for the `DecisionTreeClassifier()` changed, and how much of an increase `LogisticRegression()` received, it may be worth utilizing `GridSearchCV()` across more of the models at a later time.

Note that the best fitting model was saved as a pickle file, and can be found here.

3.3.9 Making Decisions: Is that a Good Place for a New Brewery?

Using the full data set for an expanded classification model has proven to strengthen our insights greatly. We know that having all this data (current brewery counts/types, city features, understanding of population) can therefore help us classify cities within our ranked levels of hotspots. It is Business 101 that you should do your research and collect this information if you are planning to start a business anywhere.

3.3.9.1 Creating the Decision Tree

We wanted to provide a clear and helpful resource for any current or potential business owners looking for their next brewery location. To do this, we've constructed a decision tree that will help to put market research to good use. We utilized `GridSearchCV` again to model our decision tree using different combinations of parameters. Keeping in mind that this model should be *succinct* and *usable* to the human eye, the **max_depth** of the model was capped at 4 to maintain a manageable tree size. Limiting this parameter also *sacrifices* some of the *accuracy* of the model, so we keep careful watch on the performance metrics with this reduced depth.

The best performing **decision tree** used the following parameters:

- **class_weight**: None
- **criterion**: gini
- **max_depth**: 4
- **max_features**: None
- **splitter**: best

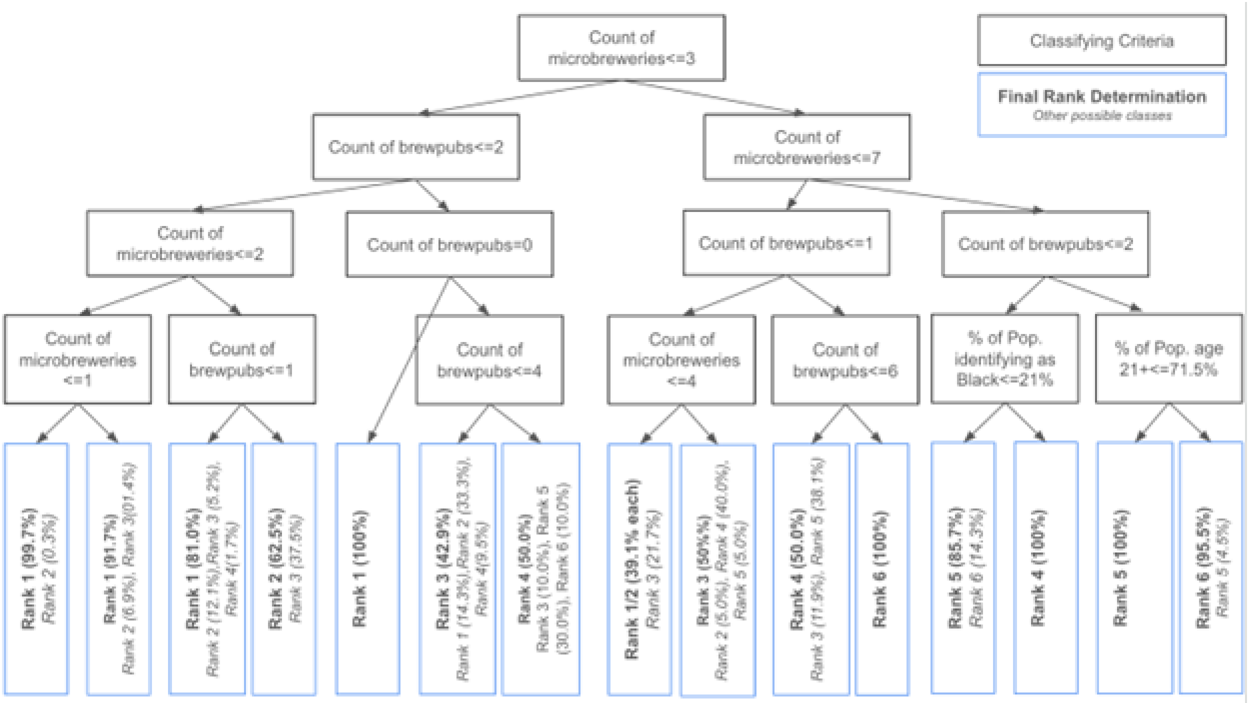
Note that the best fitting model was saved as a pickle file, and can be found here.

This model allowed us to achieve strong performance at a very *realistic/approachable* scale.

This model requires users to attain 4 key pieces of information to make their decisions:

- **microbreweries**
- **brewpubs**
- **race**
- **age**

They will be able to see the classification of their prospective location and whether that is a known brewery hotspot*(6)*, on the rise*(4-5)*, or lower on the ranking scale.


Count of microbreweries<=3
Classifying Criteria
Final Rank Determination
Other possible classes
Count of brewpubs<=2
Count of microbreweries<=7
Count of microbreweries<=2
Count of brewpubs=0
Count of brewpubs<=1
Count of brewpubs<=2
Count of microbreweries <=1
Count of brewpubs<=1
Count of brewpubs<=4
Count of microbreweries <=4
Count of brewpubs<=6
% of Pop. identifying as Black<=21%
% of Pop. age 21+<=71.5%
Rank 1 (99.7%) Rank 2 (0.3%)
Rank 1 (91.7%) Rank 2 (6.9%), Rank 3(01.4%)
Rank 1 (81.0%) Rank 2 (12.1%),Rank 3 (5.2%), Rank 4(1.7%)
Rank 2 (62.5%) Rank 3 (37.5%)
Rank 1 (100%)
Rank 3 (42.9%) Rank 1 (14.3%),Rank 2 (33.3%), Rank 4(9.5%)
Rank 4 (50.0%) Rank 3 (10.0%), Rank 5 (30.0%), Rank 6 (10.0%)
Rank 1/2 (39.1% each) Rank 3 (21.7%)
Rank 3 (50%%) Rank 2 (5.0%), Rank 4 (40.0%), Rank 5 (5.0%)
Rank 4 (50.0%) Rank 3 (11.9%), Rank 5 (38.1%)
Rank 6 (100%)
Rank 5 (85.7%) Rank 6 (14.3%)
Rank 4 (100%)
Rank 5 (100%)
Rank 6 (95.5%) Rank 5 (4.5%)

This decision tree can help to inform brewery owners about the city they are thinking of opening a new location in. Cities ranking 4, 5, or 6 on the scale have demonstrated success and interest in the brewery scene and these may be good spots to invest in. Keep in mind though that the Rank 6 cities may also have more brewery competition!

---

4 Conclusion and Future Work

The Brewery Project takes an in-depth look at breweries in the US to determine what exactly makes a city a so-called brewery hotspot. As brewery lovers ourselves, we wanted to analyze the different features that contributed to the frequency of breweries in a given area and see if we could build a model to predict the frequency of breweries in a city given its specific features. We were also interested in building a recommendation model for potential brewery owners to assist them in determining a location for a future brewery. Initially, we theorized that certain nearby recreational areas such as national parks and ski resorts would influence the frequency of breweries in nearby cities. We also theorized that city features such as whether the city is considered a college town, major city, or tech hub and the demographics of the city would influence brewery frequency.

After building our dataset, we utilized data preprocessing and exploratory data analysis to visualize our data and see if we could identify any patterns or interesting features. This aspect of the project can be viewed at the 'Data Exploration' tab of our website. Next we trained and fine-tuned various models to assess our research questions. Our processes for building the models as well as the model results can be viewed at the 'Models Implemented' tab of our website.

The data for this project is compiled from various online resources. Our brewery data comes directly from an online database. City features that were close to national parks/ski resorts, considered major cities, tech hubs or college towns were web scraped from online articles. Lastly, city population data was pulled from the US

Census' public data base. Each of these data were combined in order to gain understanding about the key features that predict the number of breweries a city may have and tell us more about what classifies a brewery hotspot in the US. Initially, we focused on a limited data set with brewery information and city features. Then, we expanded our data set according to the results of each subsequent model.

As part of our model implementation process, we first executed PCA to check the relationships between our variables and identify any redundancy in our dataset. We noticed some pretty strong patterns in the PCA plots indicating that there was some clustering within the data as well as subsets of data that shared similar characteristics.

Next, we implemented linear regression on 5 selected city features- count of ski resorts in the state, count of national parks in the state, and whether the city was considered a tech hub, college town, or major city. This analysis helped us determine which of those features had a significant impact on the brewery frequency. Our model found that being a tech hub had the most significant impact on brewery frequency, where being a tech hub was associated with an increase of around 27 breweries on average. However, our best linear regression model indicated that the city features we identified as potentially impacting brewery count are likely to be influential, but do not capture the entire story.

Having an understanding of how these city features predict the count of breweries in each city, we then applied a ranking based on total brewery count to each city. These ranks (1-6) were then used to train models to classify brewery hotspots based on the basic city features and total population, using a few different methods. Our results found that the best model used a Naive Bayes Classification method. With this model, we were able to classify cities into brewery hotspot ranks reasonably well based on city characteristics and overall population.

In an attempt to create a model with better performance in classifying cities into our hotspot ranking system, we decided to use all available features. These included the number of breweries in a state, the type of breweries in a city, information on nearby attractions such as colleges, metropolitan areas, national parks, and ski resorts, as well as census data and region designations. We only excluded variables we knew would cause collinearity issues. After testing across different data transformations and model parameters, we were ultimately able to create a model with better performance than the prior model with reduced features.

This provides evidence that given more information about the demographics and existing brewery scene of a city does result in a better hotspot ranking classification. Although the number of features were notably greater than in the reduced model, they're not too elusive for an invested business owner to find, or even estimate, given they were curious about placing a brewery in a given city.

We can conclude that our models performed better when information about the current breweries in the city and details about the population were incorporated. While this is significant, it may be unrealistic for a standard business owner to use such a complex and coded model. On that note, we produced a more user-friendly decision tree that is easily readable and actionable for business owners. This tree works with basic market research of existing breweries and local demographics to sort prospective business locations into tier 1-6 hotspots. Ranks 4 and 5 hotspots are strong contenders for brewery locations. Rank 6 is a demonstrated top tier hotspot meaning the city is supporting many breweries, although this may also mean more competition.

Even with decent performance in both the minimal feature and expanded feature models, we had several limitations due to availability of data that resulted in us using a more macro approach than we had previously expected. The brewery data itself contained information such as zip code, latitude and longitude that did not make the final modeling dataset as features. Cities, especially those on the larger end, have distinguishable pockets of demographics, industries, and customers. Even being able to discern our data at the granular level of zip codes would provide improved insight for potential business owners and brewery goers.

Part of this issue stemmed from aggregating the data between all of our sources in a way useful for modeling. An extension of this project would be to gather data in a manner conducive to providing finer details about different locales and communities within a city.

Another potential shortcoming of our model is the hotspot ranking system. Although we were overall satisfied with our method of creating hotspots through using the number of breweries in a city, we recognize that it wasn't a perfect system. Another extension of this project could be refining the definition and ranking system of a brewery hotspot. This could be in line with using our preexisting algorithm on the more granular level of zip codes or locales mentioned above, including features such as metrics based on brewery ratings, or even a more complex ranking system. We were limited to cities we knew had breweries, so we weren't able to include cities without breweries in our models.

To continue with the idea of granularity improving recommendations for multiple facets within the brewing community, further research could include data associated with pricing and information about ease of import and export to certain areas.

Limitations aside, we have created models which are usable for brewer and consumer to make informed decisions with. Even given limited information about a city, someone setting up shop or deciding where to grab a beer can dependably rely on this research.

References

1. UNTAPPD (2024) UNTAPPD

2. DB OB (2024) Open brewery DB

3. Association B (2024) Brewers association

4. Michelle Delgado DrFO (2023) The 150 best college towns

5. Review WP (2024) The 200 largest cities in the united states by population 2024

6. Sweeney M (2022) Top tech hubs in the US. ZDNET

7. STAFF (2021) U.s. National parks by state. Outside Interactive, Inc

8. Wikipedia (2024) List of ski areas and resorts in the united states

9. Bureau UC County Population by Characteristics: 2020-2022 — census.gov

10. Han J, Pei J, Tong H (2023) Data mining: Concepts and techniques, 4th ed. Morgan Kaufmann