The Brewery Project Report

Table of contents

1 Authors

Jessica Groven

Julia Gallwitz

Carl Klein

2 Abstract

3 Introduction

3.1 What Is The Project?

The Brewery Project takes an in depth look into the world of Breweries and Beer in the US. This team will explore features of the brewing market from beer prices to brewery locations to common demographics of successful brewery towns. We will focus on key areas of interest for a range of beer fans from the seasoned brew bros to new, potential brewers. This study will take on interesting perspectives of the beer market in hopes to identify key trends in beer production, consumption and sales.

The Brewery Project will start with a look into beer sales and pricing across the country. This will include comparison of all your basic and well known beers across the States and a closer look at local grocery store and pub prices. With a basic understanding of the market, we will begin to focus in our report to review some of the factors that may promote breweries. Sometimes beer is associated with outdoor mountain activities, so we will examine this and other conditions (such as college towns and tech hubs) that may favor the brewery scene. These styles of living may also be intertwined with population demographics, so we'll expand our exploration into additional factors such as income and age. The Brewery Project will conclude with actionable results that can guide future decisions on where to go if you want to buy beer, make beer, or drink beer.

3.2 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 - as beer lovers ourselves - are always looking for new breweries with good beers. We find it interesting how brewing communities often cluster near mountain activity regions and would like to take a closer look at beer consumption across other parts of the country. For many, breweries often foster a relaxed and welcoming community space to unwind and relax after a long day. 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 in order to maintain a diverse beer market. It is the underlying work of small companies that maintains the craft of brewing and we should do our best to support it. 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. A little extra knowledge about your target customers and their lifestyles, might just help point you to a successful brewery hotspot. Of course, if you just want to visit beer towns, this work will help you too.

3.3 Who Does The Project Affect?

Our project is relevant to both beer drinkers and beer brewers, primarily in the most populated metropolitan areas of the US. Of course as law abiding citizens, this will only apply to those who are 21+ years of age and permitted to consume alcohol.

For the bro-est of beer bros, certain localities have what's known as a "beer passport", which is a passport-like book that includes various breweries from the area. This encourages the passport holder to visit the included breweries where they will get one free beer and a stamp in their passport. With The Brewery Project, we want to take this a step further, and create a recommendation engine for beer lovers that personalizes brewery suggestions based on the user's demographics and preferences. The recommendations will be based on features such as personal characteristics, whether the user is on a budget, what activities they like to do, and their location in the US.

Similarly, we intend to create a recommendation engine for beer brewers, to suggest the best place to open a new brewery. We know that it can be tricky for any new business to enter the market, and the beer/brewery market is certainly saturated in some places. We believe that this project could assist prospective producers in determining the best location for a new brewery based on market gaps, local demographics, and local attractions correlating with beer consumption.

Ultimately, if you are 21+ and want to drink beer, or brew it for others, this project is for you!

3.4 What Has Been Done So Far?

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

What could be considered as "the world's most popular beer-rating platform" [1], UNTAPPD [2] allows the community to rate beers and breweries. They maintain a robust database which helps maintain and showcase the ratings and sustain an almost social media like presences for individuals, brewers, and venues. Previously, they had API access for private app developers to use, but at this point in time has been discontinued.

A completely open-source database project focused around brewery-related data known as the Open Brewery DB [3]. 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. They have a project page with featured projects using their database. Most apps and projects at least influenced by this project have a social media or geographical focus.

The original data source and scrapers for the Open Brewery DB project was the Brewers Association [4]. Their stated purpose is 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.

However, an industry which does have plenty of research and is potentially transferable is real estate. The article "5 Ways to Apply Data Science to Real Estate" by Nelson Lau [5] provides a breakdown of how different features in data are applied to make price predictions, perform cluster analysis, and incorporate the use of GIS.

3.5 What Can Still Be Done?

Through our research on the topic, we have explored various databases and projects relating to beers and breweries. What we have found is that the existing databases and projects primarily focus on exploratory data analysis. We located digital brewery heatmaps, databases detailing various characteristics of breweries and beers, and rating systems for breweries and beers. However, we have yet to find any existing project utilizing a predictive model in the beer and brewery space. As mentioned previously, we intend to build a predictive model for beer lovers to find new breweries, and 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, beer pricing, local attractions and nature, local schools and tech hubs, and population/census information. We will aggregate the data and implement data mining techniques to build a model that takes into account an individual's goals and personal characteristics to suggest the best brewery to visit, or the best place to enter the market as a beer producer. For example, we want to be able to recommend the best brewery for an out of state, 30 year old traveler, who is visiting Colorado to hike a specific 14er. We also want to be able to inform a prospective beer brewer in Texas on the best place to open up a microbrewery.

See [Appendix A: Introduction] for an illustrated version.

3.6 Research Questions

1. What is the frequency of breweries by US state?
2. Are breweries concentrated in any of the 4 major US regions more than others?
3. We've identified 5 types of features that could impact brewery popularity in cities. Do any of these features have redundancy/collinearity with one another?
4. How does the presence of different city features (ski resorts, national parks, tech hubs, major cities, college towns) relate to the count of breweries?
5. How can we define a brewery hotspot?
6. What are the top brewery hotspots?
7. Do brewery hotspots tend to be located near the city features we've outlined?
8. What are the average population statistics (age, race, etc.) near brewery hotspots?
9. Given the full knowledge of hotspot demographics, can we expand our models for better results?
10. Given the information that a normal business owner would have, what tools can they use to guide their location decisions?

---

4 Methods, Evaluation, and Results

4.1 Data Exploration

This section features our process for gathering, cleaning, and providing an initial exploratory analysis of the data used. See [Appendix B: Data Exploration] for a detailed process.

4.1.1 Open Brewery DB

The driving dataset of our project, this was pulled via API from the Open Brewery DB website [3].

4.1.1.1 Cleaning Process

We first reviewed the countries available, and found cleaning and slicing on the country name column were warranted.

After applying a strip and lowercase on the country column, we filtered just for the United States, and then examined the null values.

There were quite a few missing rows in **address_1** column, and even some incomplete addresses at that. With the completeness provided by **city**, **state**, and **postal_code**, we disregarded the other address columns, dropping all three.

Additionally, **state** is identical to **state_province**. We dropped **state_province**.

After removing nonessential rows and columns, we applied the strip and lowercase procedure on the remaining columns, and then reexamined the null values.

We decided to move forward with this data, and potentially massage any **longitude/latitude** data issues later. Even though there were quite a few missing values for **longitude/latitude**, keeping the columns could prove useful later. The **phone** and **website_url** columns will likely be irrelevant, and both have a high number of null values, but we retained them for the similar reason they may prove useful later.

4.1.1.2 Data Exploration

Our initial observation of the data is that we are dealing almost primarily with categorical values.

The latitudes and longitudes are *float* numeric values, however, they're more relevant from a geographical perspective than a numerical analysis perspective.

One method we could use to gather some numerical essence from the cleaned data is to analyze the categories (i.e. counts of breweries per state or per city, types of breweries per state or per city, etc.).

Let's take a look at how numerous the overall categories are.

4.1.1.3 Categorical Spread

A useful starting point in the visualizations would be to break down the spread of brewery types and find the top-5 and bottom-5 for brewery counts by state, ultimately leading into some heat maps of these spreads.

4.1.1.4 Brewery Type

The overwhelming amount of breweries are classified as *micro* breweries, and that the bottom 5 breweries have such small counts that they don't even appear on the plot. Another important observation is that *closed* breweries are somewhat significant in that they appear as a bar on the plot. *Closed* could become a factor in later models, so we decided to leave these data points in.

4.1.1.5 States with the Most and Least Breweries

From this initial glance, we can see what that the breweries per state varies wildly.

4.1.2 Top College Towns

The top 150 college towns in the United States were scraped from this article [6]. We plan on seeing if college towns are a driving factor of breweries. *Note that these are popular college towns as defined by the article and the author.*

4.1.2.1 Data Cleaning

The data was essentially *complete* due to the scraped data being contained in a table. However, a few tweaks were required. Namely, applying the lower case format and separating city and state. Additionally, after separating the city and state, we noticed the article used unconventional abbreviations for the states. Therefore, we applied a custom dictionary to show the states in full text.

4.1.2.2 Data Exploration

Perhaps the most efficient method to illustrate the concentration of top college towns by state is through a geographical representation.

There are definite similarities in the heat maps between brewery counts per state and college towns per state, let's continue digging into this.

4.1.2.3 Breweries in College Towns

Are there a significant amount of breweries in *college towns*? In this next exploration, we're looking directly at city to city matches, and not including surrounding areas. This could be something to return to given more geographical data.

We can create a new column in the dataframe scraped from the **open-brewery-db** with booleans on if the brewery lies in a college town. Out of all the breweries, **15.8%** of the total breweries are located in a college town.

When we look at our **Top and Bottom 5 States by Brewery Count**, we can examine how many of these breweries are located in college towns.

An interesting distinction from above is that there are more total breweries and breweries in college towns in the top 5 states for breweries, however in at least one of the bottom 5 states by brewery counts, more than half of the breweries are contained within a college town.

4.1.3 Top Metropolitan Cities

The top 300 metropolitan cities were scraped from an article here [7] which provided the city, rank and the current 2024 population. If large populations promote breweries, this data set will be helpful to demonstrate this relationship (or lack thereof).

4.1.3.1 Data Cleaning

The data scraped from the website had a couple of areas to be cleaned and transformed. First, this data should be accompanied with an additional set of columns for full state names, region and division to which the city belongs. In order to complete this, the State column needed to be capitalized to allow the appropriate merge. Once the adjustments to the original state column the Regions data set was merged in. This produced two columns using the same State column name. Columns were renamed to differentiate state code from state name. Then, the web scraper pulled city population including these commas which are preventing this column from being recognized as numeric. The commas were removed from the 2024 Population column, then these values were converted to integers.

4.1.3.2 Data Exploration

The Metropolitan Cities data set identifies the top 300 most populated cities in the United States. If breweries thrive in large cities, this will help us identify key examples. This is a pretty large data set though. Let's explore more about these cities and perhaps find a cutoff to scope down the data set.

4.1.3.3 Central Tendency of City Populations

Basic measurements were taken from the 300 city populations. These were the key results:

- The mean of the populations is 313522.52
- The minimum of the populations is 109513
- The median of the populations is 179505
- The maximum of the populations is 7931147
- The standard deviation of the populations is 568306.62

This data is showing a clear skew. While there is only a difference of about 70,000 people between the lower 50% of metropolitan cities, the upper 50% of populations has a difference of 7.7 million people. This is a very noteworthy distribution. Let's visualize this data to understand this pattern more fully across each region of the US.

4.1.3.4 Regional Population Distributions

Below is a Box and Whisker plot comparing the spread of city populations in each of the 4 US regions: South, Northeast, Midwest and West. In general, these regions are showing similar distribution patterns. Mainly, about 75% of major metropolitan cities in the US are less than approximately 400,000 people. There are a handful of outliers in the plot below that represent extremely large populations in comparison.

4.1.3.5 Regional City Distributions

While visualizing the spread is helpful, it is challenging to see patterns related to how many metropolitan cities each region holds. For this, we can use donut charts. It is interesting to observe from the three charts provided that the proportion of metropolitan cities each region holds is relatively constant whether you look at the full 300 cities, only the top 50 or even the top 10.

4.1.3.6 Breweries in Metropolitan Cities

The results above demonstrate that our data show consistent patterns and relatively comparable distributions at smaller scopes. To improve efficiency of our model, we may scope these cities down to the top 150 (upper 50%) of the metropolitan data set to avoid increasing the load and processing time of our models. In future work in this project, the location of metropolitan cities will be linked to the Brewery data set that has been developed above. WE know that most metropolitan cities occur in the West and the South. We'll look for any correlation with the frequency of breweries as we get into the proceeding modeling work.

4.1.4 Top Tech Hubs

26 tech hubs were identified by this site [8] that assessed tech hubs and the living conditions in each location. For the purposes of this project, only the city names were scraped to use in review of their correlation with brewery hotspots. This data set will support the ability to compare the occurrence of breweries to the presence of tech hubs.

4.1.4.1 Data Cleaning

The data scraped from the website had several needed improvements. The data was formatted as [City],[State code] when initially pulled from the web source. For our purposes this column needed to be split into two columns at the comma. It was noted that Dallas and Fort Worth, Texas were combined into one entry and also did not include a state code. After data was split and the original column was removed. The Dallas-Ft.Worth entry was removed and separate Dallas and Fort Worth entries were appended to the data set. With properly formatted City and State columns, the Regions data set was merged in based on the state code.

4.1.4.2 Data Exploration

The Tech Hubs data set identifies the 26 popular tech hubs across the United States. If breweries are more successful near places of technology and innovation, this will help us identify the correlation.

Let's review the distribution of these tech hubs across the country then map them to see the distribution.

4.1.4.3 Regional Population Distributions

In exploring this data, we find the distribution of tech hubs by region as follows: South (9), West (7), Midwest (6), and Northeast (4). We can represent this with a donut chart.

4.1.4.4 Tech Hubs Across the US Map

Understanding the spread of distribution by region is helpful, but it is also telling to review tech hubs by state. Below we can identify that Texas, California and Ohio, are the only states to contain more than two tech hubs.

4.1.4.5 Breweries in Tech Hubs

The summary above demonstrates some basic patterns in the Tech Hub data set. These patterns will be considered alongside the other features in our project to understand if Tech Hubs tend to have an increased/decreased number of breweries within the area. Further models will consider whether the breweries of the Brewery Database fall in or near the Tech Hubs that we have found.

4.1.5 Ski Resorts

A full list of currently operational US ski resorts was scraped from this source [9]. The source included the name of the ski resort, as well as the city and state the resort is located in. This data will help inform our question of whether brewery locations are correlated with outdoor recreational hotspots such as ski resorts

4.1.5.1 Data Cleaning

Pictured above is a snapshot of the initial dataset. The dataset requires a handful of manipulations to the city names in order to clean it up for analysis. As pictured above, some of the ski resorts are located near certain cities, and are designated as such (eg "near Wenatchee"). Obtaining the nearest city is suitable for our purposes, so the "near " was removed. Next, some of the city names have additional information listed after them (eg "Stampede Pass (private)" to indicate this is a private resort). Any text found in brackets was removed. Finally, certain resorts such as The Summit at Snoqualmie have multiple cities listed. Only the first city was kept as this was suitable for our purposes.

4.1.5.2 Data Exploration

Utilizing the `.info()` function, we can see that our fields include Ski Resort, City, and State. All of the values are filled in for each field and we are dealing with exclusively object data types in this dataset.

4.1.5.3 Resorts by State

We know that our top 5 states with the most breweries are California, Washington, Colorado, New York, and Michigan. The pie chart above shows the distribution of ski resorts by state as a percentage of the total. The cut off for being included in the "other" pie slice was 7 ski resorts. We can see that all of our top brewery states are listed in the pie chart as having their own slice. We also know that our bottom 5 states with the least breweries are Delaware, North Dakota, Hawaii, Mississippi, and District of Columbia. Interestingly, none of those states have their own slice in the chart above.

Next, we took a look at just the top 5 states with the most ski resorts, as pictured in the bar chart above. Notably, we find 3 out of 5 of our top brewery states also being listed in the top 5 ski resort states (New York, Michigan, and California). This indicates that there may be some association between the incidence of ski resorts and breweries.

4.1.6 National Parks

A full list of national parks in the US was scraped from this source [10]. The source included the name of the national park and the state it is located in. Both data points were captured by the scraper. This data will help inform our question of whether brewery locations are correlated with outdoor recreational hotspots such as national parks

4.1.6.1 Data Cleaning

Pictured above is a snippet of the initial dataset. There is not much cleaning necessary as the data is pretty straightforward and the source does not include any additional information/text. However we did utilize a geolocator package to add the latitude and longitude to each national park as this may be useful for future modeling.

Pictured below is a snippet of the cleaned dataset with the latitude and longitude added in. Notably, there are a handful of null values that the geolocator could not fill in. We have decided to leave those values as is while we determine whether we will be using the coordinates.

4.1.6.2 Data Exploration

Utilizing the `.info()` function, we can see that our fields include National Park, State, Latitude, and Longitude. National Park and State values are objects and all values are filled in. Latitude and Longitude values are float and we do have a handful of NaN values there. As mentioned, we are going to leave those for now while we determine whether we will be utilizing the coordinates.

4.1.6.3 National Parks by State

We know that our top 5 states with the most breweries are California, Washington, Colorado, New York, and Michigan. The pie chart above shows the distribution of national parks by state as a percentage of the total. The cut off for being included in the "other" pie slice was less than 2 national parks. We can see that 3 out of 5 of our top brewery states (California, Colorado, and Washington) are listed in the pie chart as having their own slice. We also know that our bottom 5 states with the least breweries are Delaware, North Dakota, Hawaii, Mississippi, and District of Columbia. We see that Hawaii is the only bottom 5 state to have its own pie slice.

Next, we took a look at just the top 5 states with the most ski resorts, as pictured in the bar chart above. Notably, we find 2 out of 5 of our top brewery states also being listed in the top 5 ski resort states (California and Colorado). This indicates that there may be some association between the incidence of national parks and breweries. However, we found less of our top 5 brewery states in the top 5 of national parks compared to the top 5 of ski resorts.

4.2 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. See [Appendix C: Models Implemented] for a detailed process.

4.2.1 Consolidated Data

Using datasets featured in our exploratory analysis, we merged them together in 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.

4.2.2 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.

4.2.3 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. The starting dataset can be viewed above. We implemented data preprocessing.

The PCA process was then implemented, as well as the individual variance explained by each principal component.

The principal components were then mapped against each other.

We can definitely see that there is some sort of pattern going on within our dataset. In the first plot, we see pretty distinct vertical lines which indicate clusters of data that are separated among PC1, though they appear to vary more among PC2. This may be due to our boolean variables influencing the variation of PC1. We see a somewhat similar occurrence in the PC2 and PC3 plot, however here we see clustering rather than distinct lines, which may indicate subsets of the data that share similar characteristics. Finally, we see diagonal lines in the bottom two PC plots, indicating that there may be correlations between the variables of our datasets.

4.2.4 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.

4.2.5 Defining and Classifying Brewery Hotspots

In this section, we'll create a defintion of hotspots and then create a classification model to predict how much of a hotspot a given city is (or should be).

4.2.6 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.

There are two methods that we identified for measuring top brewery hot spots.

- Breweries per capita:
  - Calculates total breweries per 1000 people living in the city
  - Benefits: Scales with population, smaller towns with many breweries can be considered hot spots
  - Challenges: Major cities with large populations may be disadvantaged. Extremely small cities with a single brewery have potential to be ranked overly high on scale.
- Total brewery count:
  - Counts total breweries per city
  - Benefits: Major cities with more breweries will rank higher. Cities with just one brewery will not be considered as a hotspot.
  - Challenges: Small towns may be discounted for only having a few breweries.

Both of these methods are ways that we can quantify brewery hot spots. When we review the outcomes of these ranking metrics, we find that the top rank of the Breweries per Capita yields many small towns that only have 1 or 2 breweries and excludes many very well known brewery locations such as Portland, Oregon (since Portland has such a large population, Breweries per capita calculated very low).

The data representing the top tier brewery hotspots calculated with the Total brewery count follows much more logically with our understanding of current brewery hot spots in the US. To avoid challenges with classifying cities with 1 or 2 breweries as hot spots, we will proceed with the analysis using the rank 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. The remainder of our analysis will use this custom ranking to assess our models 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.

4.2.7 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)

The following results were attained.

These results have been reorder by F1-Score to see the highest performing models. We can plot these results for a more visual sense of highest performing models as well.

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 [11].

4.2.8 Preparation for Modeling - Full Dataset Features

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 perform the following in preparation for modeling:

- remove `city` (unique identifiers)
- remove `state` (51 more columns don't seem necessary, especially when there is state specific data)
- remove `2021 median income` (86.4% missing values)
- remove `brewery_concentration` (a variable created from population and city brewery count)
- remove `per_capita_ranked` (a variable crated from brewery_concentration)
- remove `city_brewery_count` (basis for custom_ranked)
- remove rows where the census data is failed to be captured
- encode `region`

> Due to `city_brewery_count` being used in creating the hotspot criteria (ranked 1-6), it was removed for modeling.

After this process, our dataset is ready for modeling.

4.2.9 Classifying Brewery Hotspots - Full Dataset Features

Now that we have defined a hotspot ranking system and prepared our dataset for modeling, it's time to create some classification models!

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)

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)

Using the **sklearn** library, the default classification algorithms we'll be testing are:

- **Decision Tree**: `DecisionTreeClassifier()`
- **Logistic Regression**: `LogisticRegression()`
- **K Nearest Neighbors**: `KNeighborsClassifier()`
- **Support Vector Machine**: `VC()`
- **Naive Baye**s: `GaussianNB()`
- **Linear Discriminant Analysis**: `LinearDiscriminantAnalysis()`

4.2.9.1 Default Algorithms

After performing the default algorithms, we found that the `DecisionTreeClassifier()` performed the best across the metrics Accuracy, Precision, Recall, and F1-Score 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.*

4.2.9.2 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 the metrics Accuracy, Precision, Recall, and F1-Score 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.

4.2.9.3 Parameter Hypertuning

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

For `DecisionTreeClassifier()` (non-scaled data), we will test the following parameters:

- **criterion**: gini, entropy, log_loss
- **splitter**: best, random
- **max_depth**: None, 2, 4, 6, 8, 10
- **max_features**: None, sqrt, log2
- **class_weight**: None, balanced

From this process, we deduced that the best parameters for this model are:

- **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).

For `LogisticRegression()` (scaled data), we will test the following parameters:

- **class_weight**: None, balanced
- **max_iter**: 10000
- **multi_class**: auto, multinomial
- **penalty**: None, l2, l1, balanced
- **solver**: lbfgs, saga

From this process, we deduced that the best parameters for this model are:

- **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 with scores in Accuracy, Precision, Recall, F1-Score 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* [12]. 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.

4.2.9.4 Applying the Model

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

4.2.10 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.

4.2.10.1 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

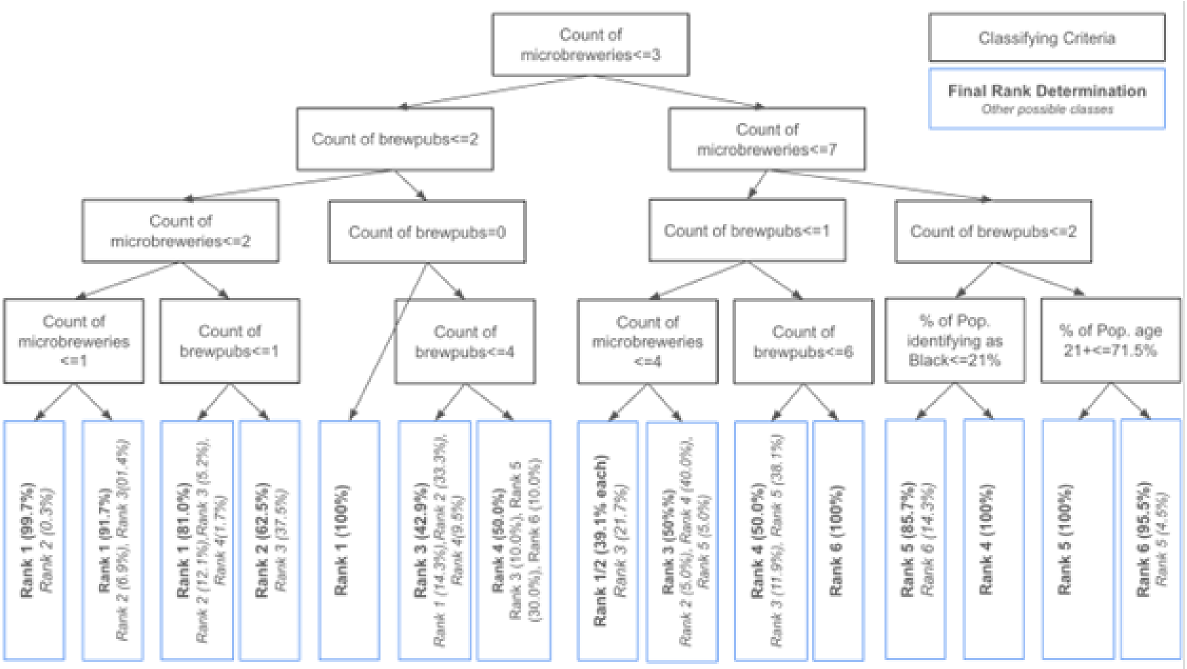
> Note that the best fitting model was saved as a pickle file, and can be found here.

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!

Conclusion and Future Work

1. Bernot K (2021) Tyranny of the tickers - how UNTAPPD ratings became craft beer's most Fickle prize. Good Beer Hunting

2. UNTAPPD (2024) UNTAPPD

3. DB OB (2024) Open brewery DB

4. Association B (2024) Brewers association

5. Lau N (2020) 5 Ways to Apply Data Science to Real Estate. Medium

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

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

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

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

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

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

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