**Craft Beer in the U.S.**

Project 1 - Summary Document

Angie Hilliker, Bryan Wood, Stephen Walthall, Lauren Taylor

**Purpose:** This document describes our relevant data sources, data cleansing, overall analysis and major conclusions.

**Questions we found interesting:**

The common thread among our interest is the fast expansion of the beer brewing industry and whether it varies regionally across the United States. Are breweries more likely to open in proximity to a major metropolitan areas? Are their geographical preferences (i.e, the south, northeast, etc..)? Where was this rapid growth occurring and why? What demographic factors contribute to rapid growth of breweries? For example, does brewery number correlate to population size, state area, average income, etc. We were able to address some of these questions with publically available data.

**Data retrieval process/sources:**

* We located a brewery database (brewerydb.com) that contained information regarding a breweries location, state it was located in and what year it opened up. This combined with market research and publications on the expansion of brewery growth over time were used to answer the growth in the US question.
* The US government provided demographic data and taxation data that was useful in our analyses. The Alcohol and Tobacco Tax and Trade Bureau provided (1) number of operational breweries per year by state, (2) barrels of beer produced per year aggregated by state or by brewery size, and (3) how many barrels were sold in bars, bottles/cans, or by keg. 2010 population per state per year came from the 2010 census data and the federal government provides estimates of population growth for each year since. Land area for each state was also provided via the Census data. Income and poverty level data were obtained from the American Community Survey through census.gov.

**Data cleansing process:**

Utilizing brewerydb.com’s API, we were able to find a complete listing of brewery locations in the world along with data pertaining to their location (latitude and longitude) as well as state they were located in and year they were established. From there, we were was able to specifically query breweries located in the US and those pieces of information and build a dataframe from this data. Returned 9417 records. Due to limitations within the data, we identified 4086 did not include an “Established Year” field and therefore did not include them in our analysis. Our next step would have been to research other potential sources such as Google places or other databases that may have provided the established year for each of the 4086 breweries.

The federal government data was supplied as either excel, pdf or csv files, each requiring different importing methods. Data clean up included converting NaN and various symbols to zero, cutting data from US territories, making headers more readable. A github user (cphalpert) shared a useful python database on github that included state names, abbreviations, regions, and census designations. We merged this with the federal data so that we could use state name or abbreviation in graphs and could sort and group based on geographic regions. This extra database really increased the ease of using the data. We experimented with several ways to show the data, settling on the subset shown in the presentation, but more options are shown in the jupyter notebook. We found that each data set opened up new variables and possible questions, some of which couldn’t be answered with the data available. For example, massive growth in brewery number led us to ask what types of breweries were responsible for growth. As nanobreweries are responsible for the most growth, we asked what the impact on the market share is for each brewery type, since they produce at such different scales. Thinking about production led us to consider how breweries sell their product (in bars/breweries, in bottles/cans, versus by the keg). While we could break the sale type by state, we would love to compare that back to brewery type, but lack the publically available data to do so.

For the PDF conversion, there is a fantastic wrapper called tabula-py (chezou on github) that will allow you to read them as a dataframe or JSON table, or convert them into a CSV, TSV, or JSON file. An interesting issue with the cleanup was having to remove commas (an issue I wasn’t previously aware of), and making sure to mark where numbers were converted into exponentials when graphing.

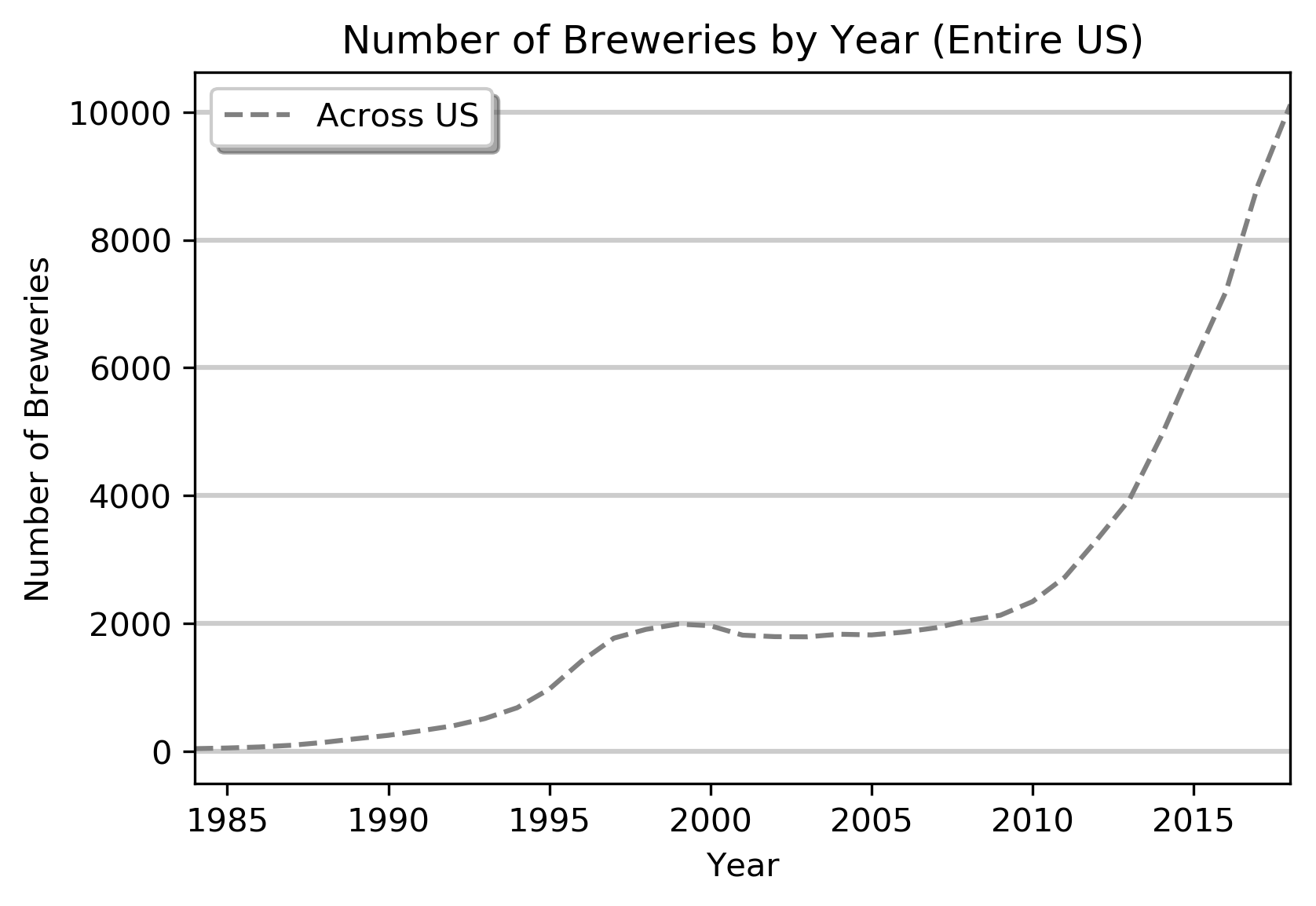
**Overall analysis:**

As we explored the data, it was clear that different trends emerged from various ways of showing the data. For example, plotting number of breweries per state over time showed a mini surge in brewery number in the late 1990’s and a rapid rate of growth since 2015. However, this growth isn’t equally distributed across the states, so plotting location of breweries better demonstrates hot spots within states. We explored the brewery count by state data, and calculated percentage change between two time periods; 1990 - 2009 and from 2009 to present. We chose these time frames as from 1990 - 2000 there really wasn’t any major growth within the industry. It was also more powerful to see how some areas of the country grew exponentially during the 10 year period between 2009 and now versus the 20 year period of 1990 - 2000. After performing the analysis for calculating percentage change, we identified that there was a slight limitation as some states had zero breweries in 1999. To get a more accurate picture of states that experienced the largest growth, we calculated the change in number of breweries between the two time periods. We reviewed the top 5 states with the largest additions of breweries and noticed that several states made the top 5 for each time period which could indicate that the industry is very strong in those areas.

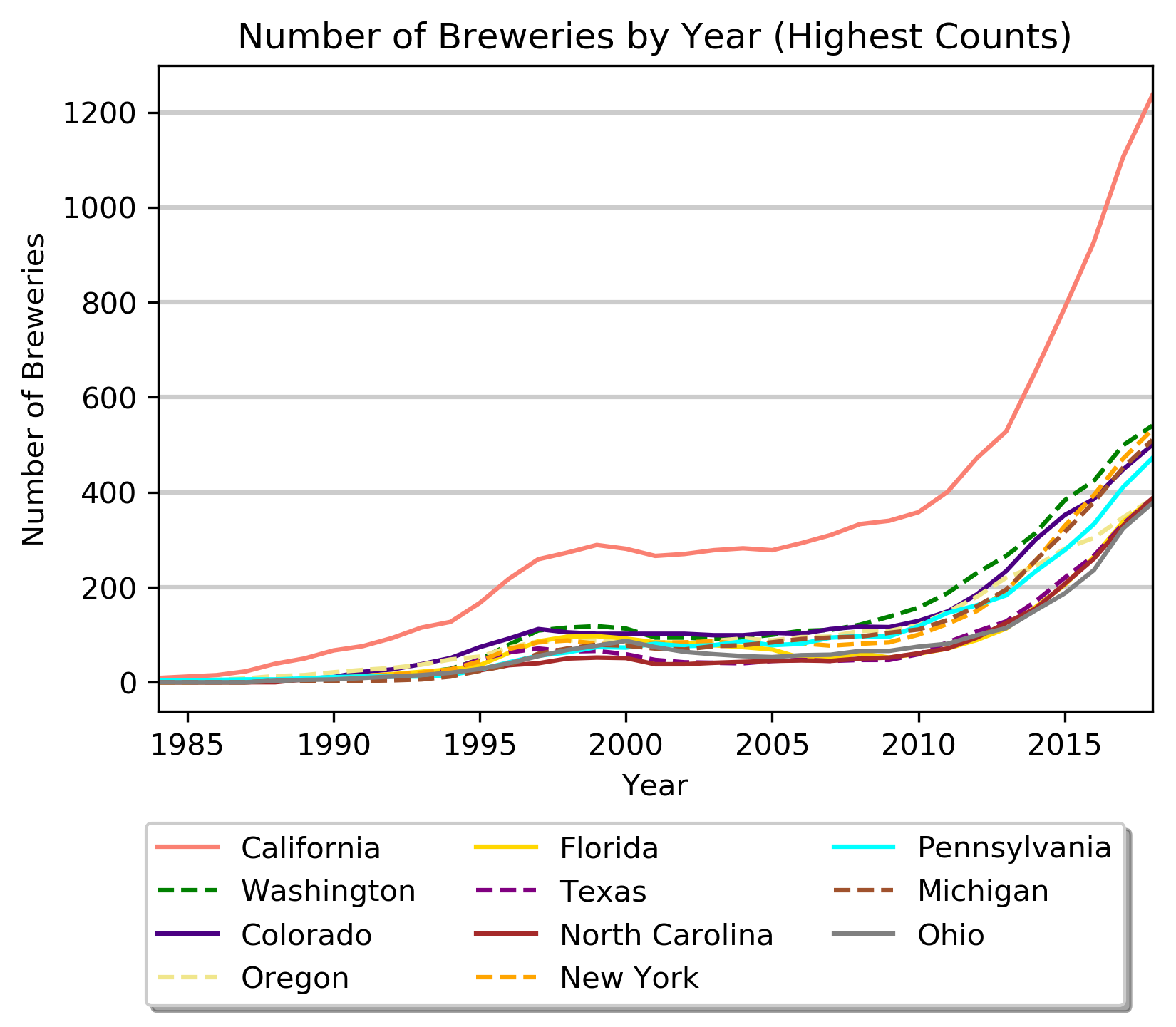
Different sources of data were more accessible than others, which made a difference in what we could analyze within the time frame of the project. For example, we found a very useful API with brewing data, but it took a significant amount of work to learn how to call the data needed. Luckily the API provided a “sandbox” of sample data to practice on, so we could avoid hitting daily call limits. The API provided dates that breweries were established and gave their location so that we could visualize where growth was happening across the country. However, learning how to use the API, pulling the data in batches at different times (to avoid triggering call limits), and mapping the data took significant time, so we mapped just a few years to show brewery growth. In contrast, federal data showing operational brewery number per year was available as a formatted table going back to 1984. This data was easier to parse and gave a 30+ year look at brewery growth, but lacked information on individual brewery openings and location. By analyzing this dataset, we could see during which years rapid brewery expansion occurred, so that we could complete the more difficult location mapping analysis on a couple of key years that illustrated the rapid growth seen over time. We leveraged different data sets successfully to make the most use of our time and observe interesting trends in the data..

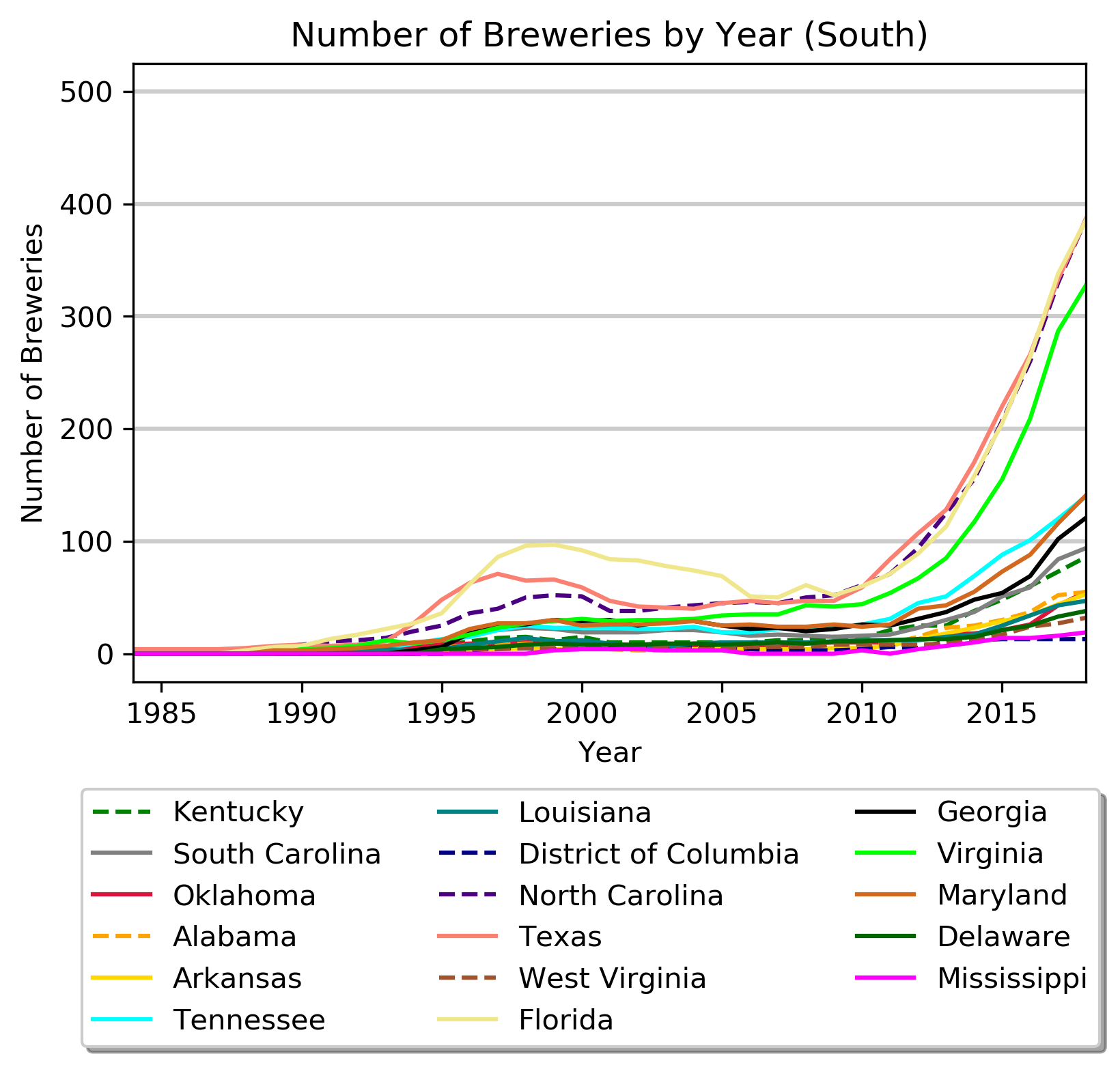
We also found that each piece of data led to more questions, which was fun, but we didn’t always have the data or time to answer these questions. For example, the brewery database included information about beer styles offered by breweries and we were hoping to use this information to analyze beer style preference by state. However, after getting familiar with the API it became clear that data set up and the call limits on the API made it impossible for us to get a representative set of national data for beer styles. We think this likely reflects real world data analysis in that you can spend significant time mining a data source before realizing its limitations given the scope of your project in terms of time or resources.

We have provided an annotated jupyter notebook file with our code. Look for the following headers to direct you to portions of the code:

* “Map brewery locations by geocoordinates” for the code to create maps of brewery location during specific years
* “Using Federal Data to Look at Brewery Growth Per State Over Time” for code making line charts of brewery growth over time
* “Comparing Growth in Brewery Number to State and Population Characteristics” for code on scatter plots and bar charts related to brewery growth by state compared to population or income
* “Growth and Production in Breweries Binned by Brewery Production Volume” stacked bar chart and pie charts of brewery types (region, micro, etc.)
* “Calculating percentage change in breweries between 1990/2009 & 2009/2019” to identify states with the biggest changes over these time periods
* “Comparing sales by type” for code that compares sales in taverns, bottle/cans, or kegs.

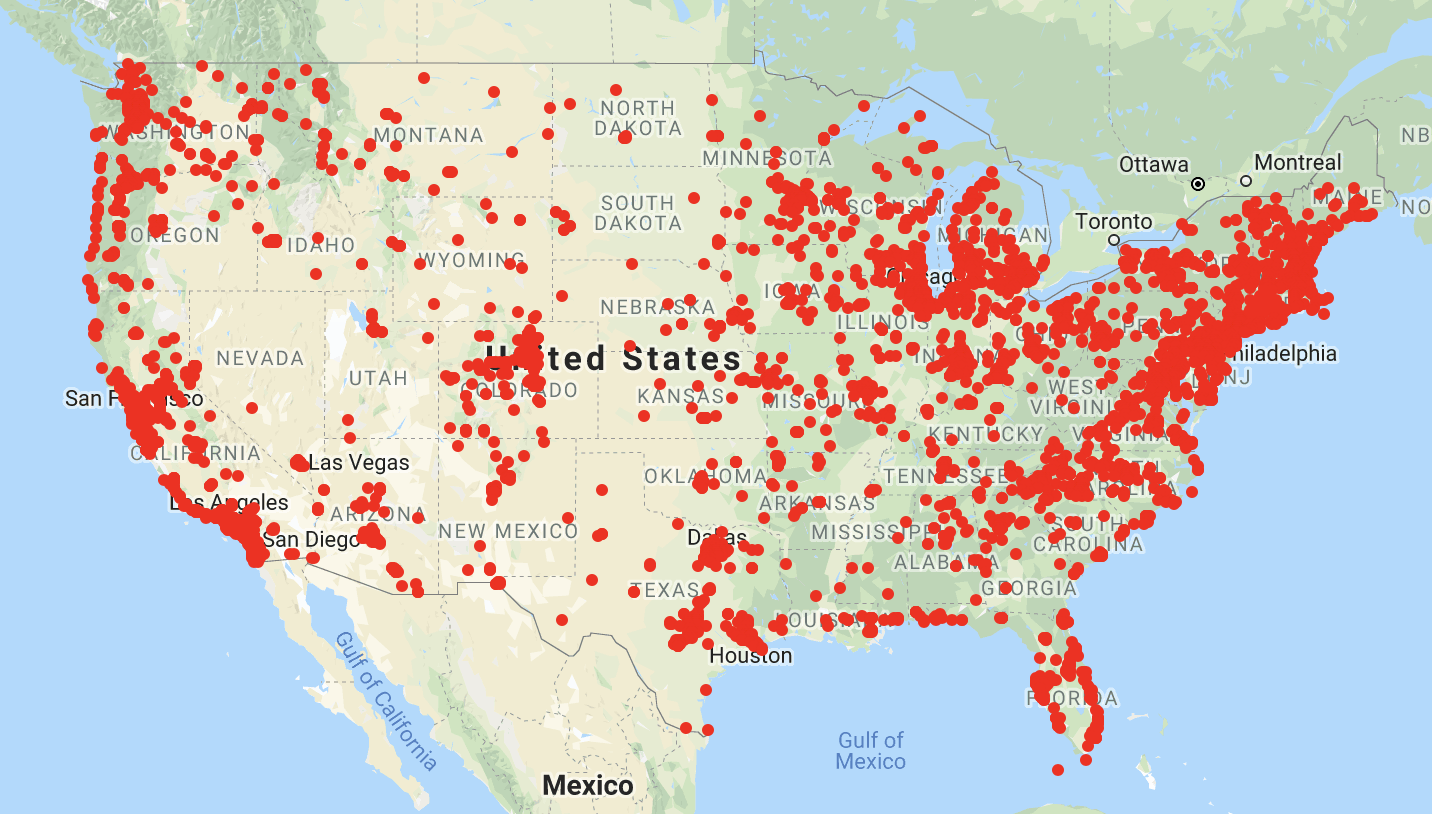
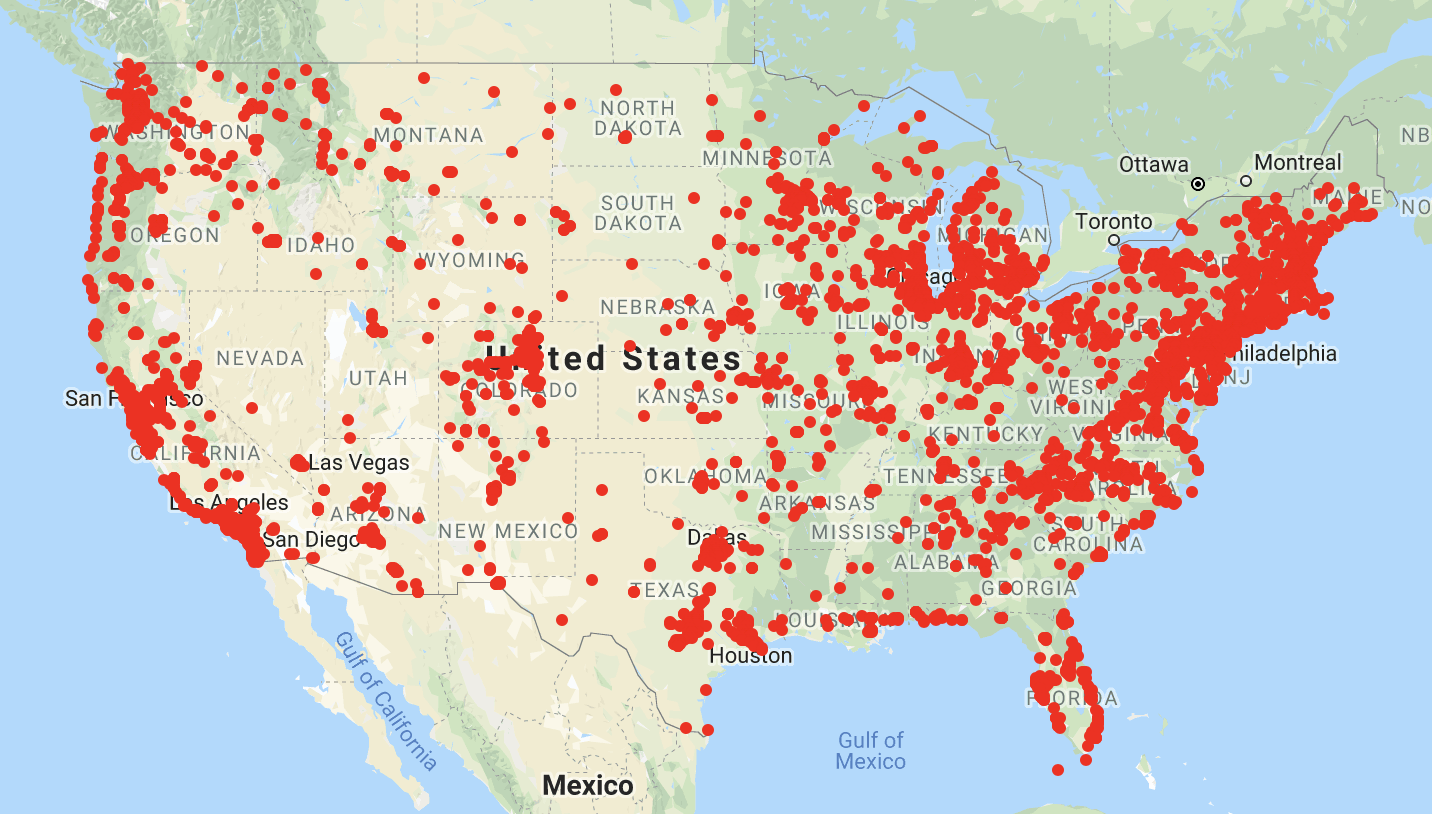
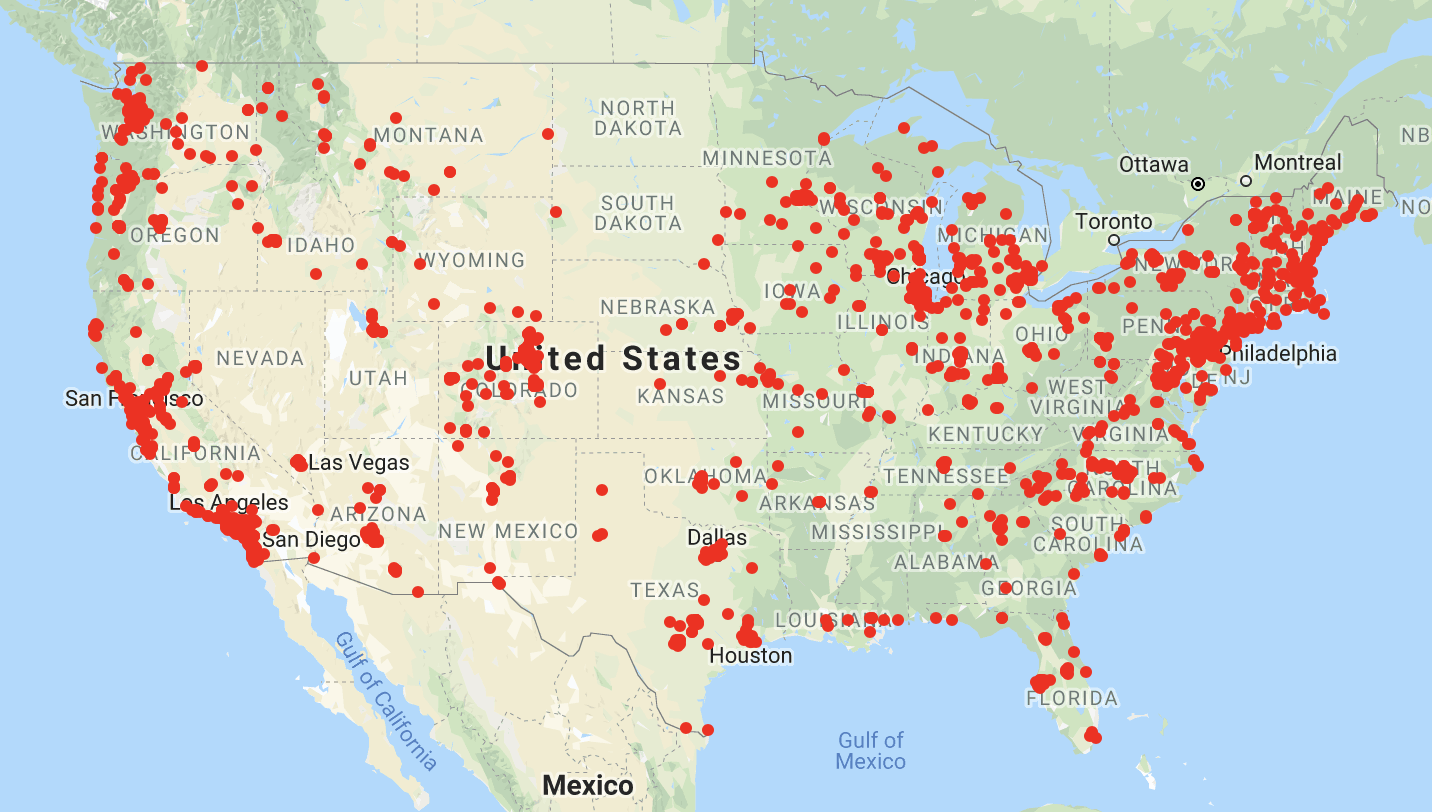
**Conclusions:**

Breweries were a rarity in the mid-1980’s with only 26 breweries in the entire United States. After a relative boom throughout the 1990’s, the US had 75 times more breweries by the turn of the century, bringing the total number of breweries to 1,964. From 2000 to 2010, the number of breweries nationwide remained relatively stable, although state to state comparisons show instability within certain states (see below). Between 2010 and 2018, there has been a second boom in breweries with a 4.3 fold increase in brewing number. Today, there are over 10,000 active breweries in the United States (Figure 1). This data comes from the Alcohol and Tobacco Tax and Trade Bureau in the US government. This data is broken down by state allowing us to analyze whether there is state to state variability in the growth of breweries. 

We grouped the states by geographic region to better visualize the nuances of state to state variation (one example is shown below). The states with the most breweries as of 2018 were grouped together for easier comparison (Figure 2). California dominates with more than twice the number of breweries compared to the next leading states. State by state analysis shows that some states suffered a significant decline in brewery number from 2000 to 2010, most notably, Florida (Figure 3) and Ohio (figure found in jupyter notebook). 

While there are small differences state to state, overall most states show a similar trend to that seen in cumulative date with the US as a whole (Figure 1), that is that brewing saw a boom in the 1990’s, was relatively stable for a decade, and has now seen rapid growth in the past 8 years with a 14-23% rise in brewery number year to year since 2011.

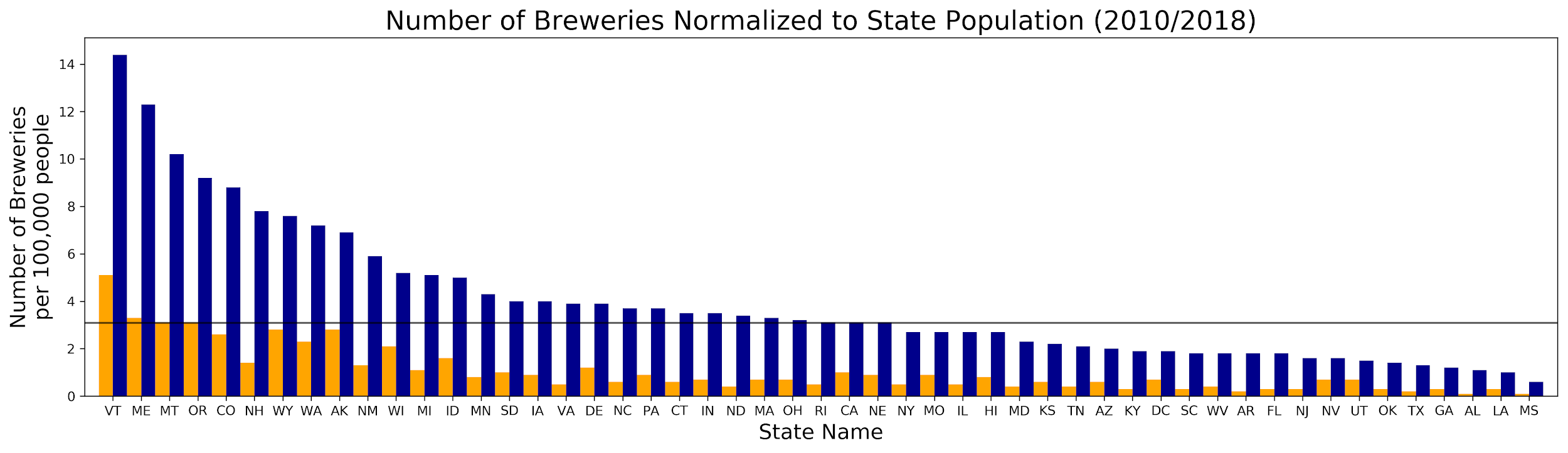
Obviously the growth in breweries is not expected to be evenly distributed among each state, as we well know from walking through Scott’s Addition, here in Richmond. To visualize whether and where the growth in breweries was concentrated in each state, we acquired information on the opening date, name, and location of each brewery in the US from the Brewery Database API (brewerydb.com) and mapped these locations for three different years. We choose to map 1990, 2009, and 2018 to reflect key points in the growth booms we observed in Figure 1. As expected, we observed rapid growth in specific areas of each state (Figure 4, below with 1990, 2009, and 2018 shown left to right).

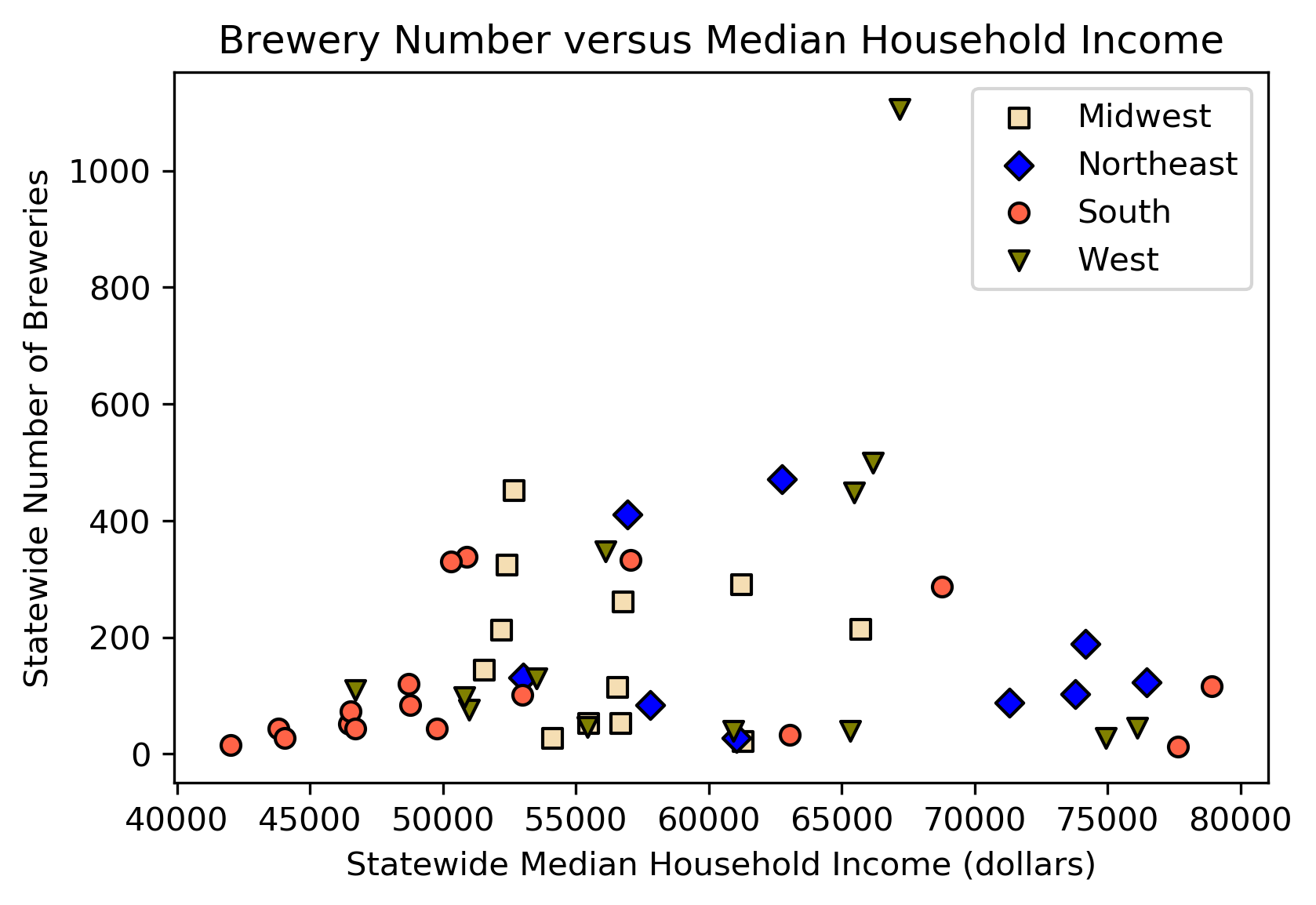


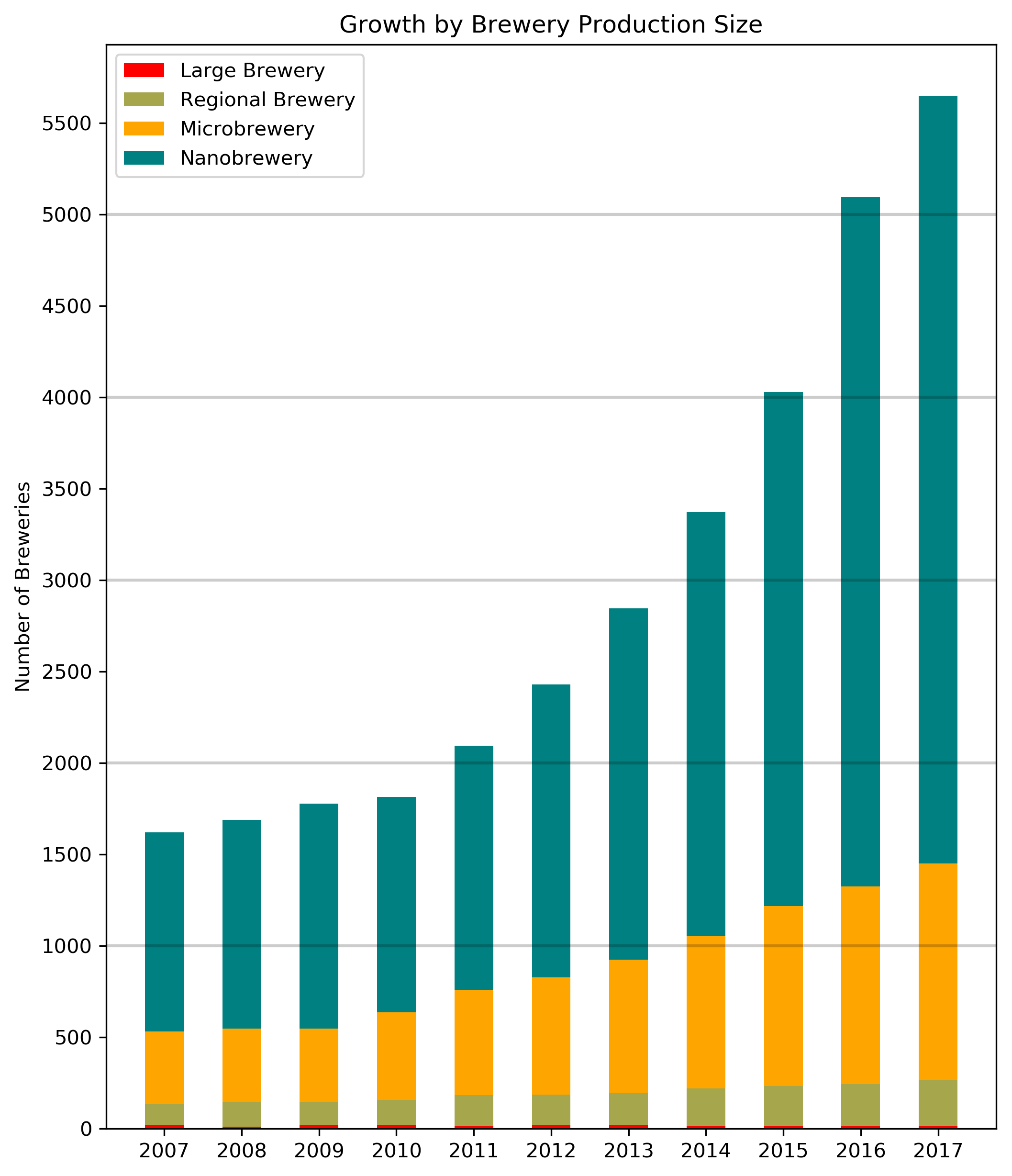
We also calculated the change in number of breweries from 1990 to 2009 and from 2009 to 2018 for each state (highlighting only the top five states) to support the map visualization by emphasizing numerically how significant the growth has been for many regions. From 1990 to 2009, we found that California gained 273 breweries, Washington gained 128, Colorado gained 105, Michigan gained 101, and Wisconsin gained 91. Then from 2009 to 2018, we observed California’s number increase by another 896 breweries, New York by 448, Michigan added 406, Washington gained 402, and Colorado increased by 384 breweries.

We were curious about what characteristics might explain the differences in brewery growth among the 50 states. One obvious question is whether the difference in brewery number is related to the population within the state. Perhaps California has more breweries, simply because there are more potential customers in state. So we gathered population estimates from the US Census. The Census office provides hard population numbers from the 2010 Census and provides estimated population for every year since the previous census. We normalized the brewery number in each state by 100,000 residents within the state. While we performed this analysis for data from 2010-2018, we are showing just 2010 and 2018 as a snapshot in this period of growth.

As shown in Figure 5 (below), California is right at the national average (as indicated by the grey line) for number of breweries per capita. Its presence as a national leader in Figure 2 is misleading when one takes into account population. In fact, Vermont leads the way in servicing its residents with the most breweries per capita. All states have shown growth in this metric since 2010, showing the popularity of beer has reached all states to some extent. But clearly, as the country averages only 3 breweries per 100,000 residents, there may be further potential for growth. Note that the tail end of the graph includes a cluster of southern states. This trend would be interesting to study. Is beer consumption lower in these states? Is there less interest in regional breweries or frequenting breweries? Or does this trend point out exciting new markets that are currently underrepresented. Finally, it is possible that state by state liquor laws slow growth of this industry.



We were interested in whether certain criteria about the state or the state’s population correlated with the number of breweries. So we compared the number of breweries to a number of characteristics, including raw population, population density, state land area, household income mean/median, and household poverty level. All of this data was available from the US Census or the American Community Survey via the Census office (data available at factfinder.census.gov). The data for many of these comparisons can be found in the jupyter notebook. Figure 6 shows one such comparison, looking at brewery number versus average household income. Each state is a data point, color coded by region of the country. Linear regression analysis showed an R squared value of 0.02 suggesting that no more than a 2% correlation with median household income, and a p-value of 0.34, suggesting even this small correlation is statistically insignificant. Since this income metric includes all households, it would be interesting to parse this data by bins of different income levels or by different types of households. For example, the Richmond Tourism board finds that millennials are the highest tourism group in Richmond and that breweries are one of the draws for that group. So, income from households of a specific age group, with or without children, etc. might correlate with brewery number. These additional associations were outside the timeframe of this study. 

We wanted to assess which type of brewery showed the most growth from 2007-2017. The Alcohol and Tobacco Tax and Trade Bureau compiles brewery number and production (in barrels) and bins that data by ten categories based on brewery size. We collapsed those categories into four bins: large breweries (at least 6 million barrels), regional breweries (15,001 to 6 million), microbreweries (1,001 to 15,000 barrels), and nanobreweries (up to 1,000 barrels). The first three designations are from the US Brewers Association. Nanobrewery is a newer term that is not strictly defined, but we limited it to 1,000 barrels or less based on the breakdown of the available data. Large breweries decreased from 17 breweries in 2007 to 15 breweries in 2017. The other categories showed growth from 2007: 115% decrease for regional breweries, 198% increase for microbreweries and a 286% increase for nanobreweries (Figure 7). Since nano- and micro-breweries dominate the growth yet produce much less beer, we wondered how much this expansive growth has actually shifted the beer production market? 

We compared the production (in barrels) from each class of brewery, comparing 2007 to 2017 (Figure 8: 2007 at left and 2017 at right). Large Breweries dominate the market, but their market share has dropped 12% from 2007. Why? Are there less large breweries? Yes, there were 15 breweries in 2017 compared to 17 in 2007. Perhaps more importantly, the average production per brewery is down 8000 barrels (data not shown). Additionally, less beer is being produced nationwide (~320K less barrels in 2017 compared to 2007. So, losing percent market share does equate to losing money for the big beer companies, given that overall barrel production is down. While there has been an increase in market share for all other categories, most of this market share has been taken by regional breweries (~10%) with microbreweries taking a1.6% and nanobreweries taking 0.3%. So the growth in nanobreweries may be increasing destinations for enjoying beer and may bring a wider array of beer choices to its local market, but the effect on the big beer industry is limited at this point. However, the breweries that expand to regional level are taking a significant bite out of the market. These regional breweries include a wide range of volume output (15,001 to 6 million barrels) and examples include Hardywood and Stone Brewing (both at the smaller end of the range). It would be interesting to repeat this analysis for smaller bins within this data set to find out which level of brewery is competing the most for market share. Many of these regional breweries started as microbreweries, showing the potential for smaller operations to expand and increase competition with the big brands over time.

