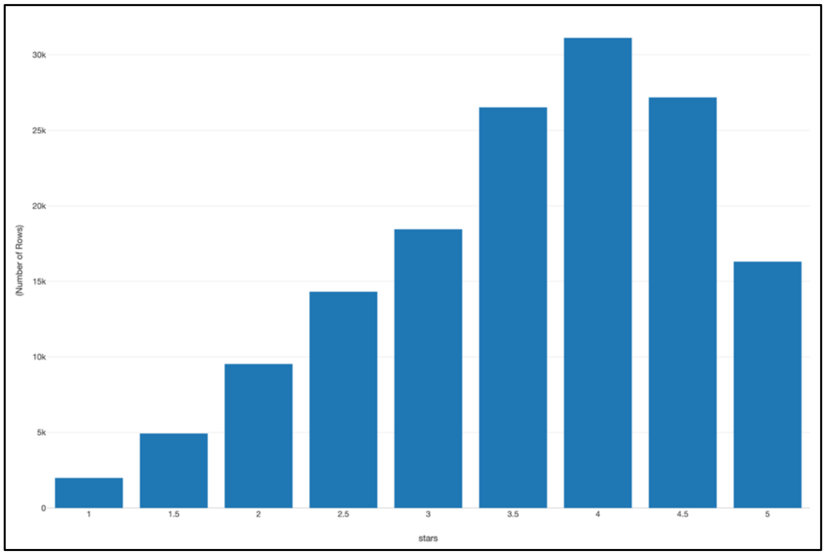
Ben Bankston, Crawford Clayton, Whitney McCormick, & David Rubin

BUS-314-01 Exploratory Group Project:

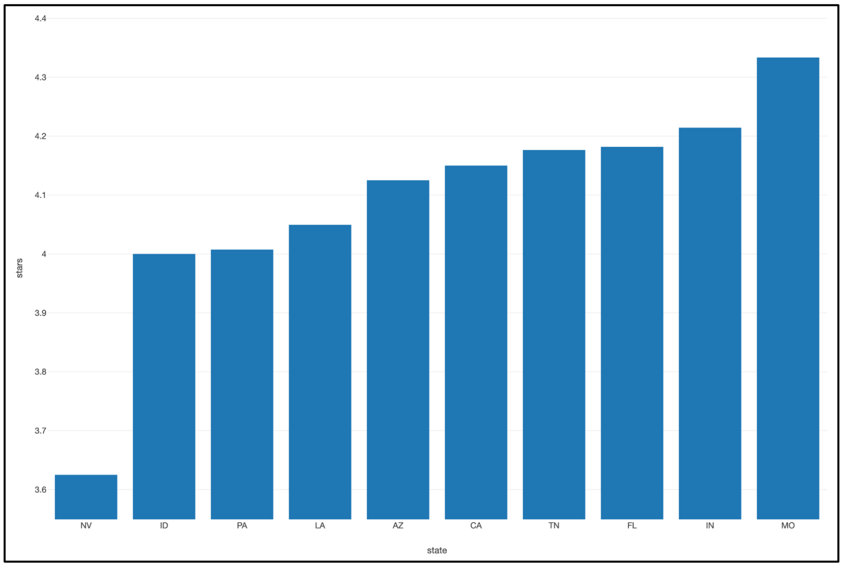
***Guiding Question*: What are the characteristics that yield high engagement and success with Yelp businesses?**

1. Number of Businesses Earning Each Star Rating



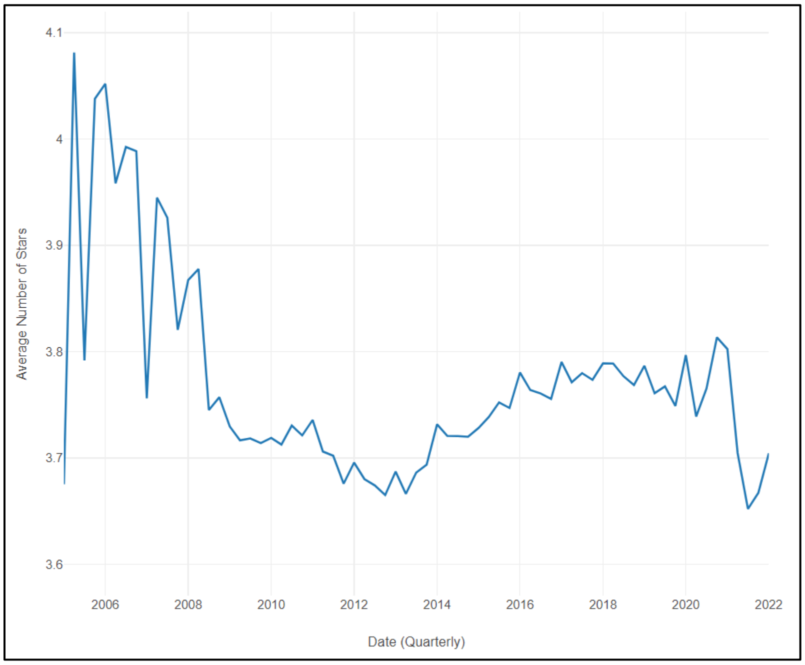
In the "business" dataset, each business has an associated average star rating between 1 and 5. About 30,000 businesses have a star rating of 4 (roughly one-fifth of the dataset). The stars are heavily leftward skewed and fall in frequency by about 5,000 reviews per half star. This shows that Yelp users are more likely to give star ratings higher on the scale and receiving a rating of 5 is about as common as a rating of 3.

2. States that Earn the Highest Average Reviews



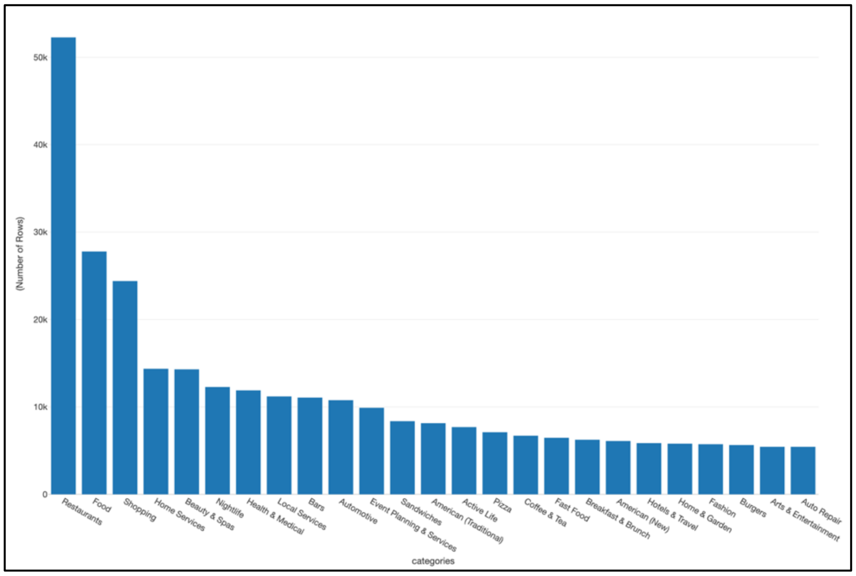
The ‘Review’ branch of the Yelp Dataset includes the reviews (from 1 to 5 stars) of the many different restaurants across the nation. We wanted to find a way to best visualize a spread of which areas are most likely to have good restaurants, and we utilized the ‘state’ variable to do that. We filtered the graph to reflect the states that have the highest average star rating, and the result reflects how highly rated the restaurants in Missouri are. One of a few possibilities for this is that businesses that register with Yelp in these states are better than what they are accustomed to, so these businesses earn higher ratings.

3. Average Star Ratings Over Time



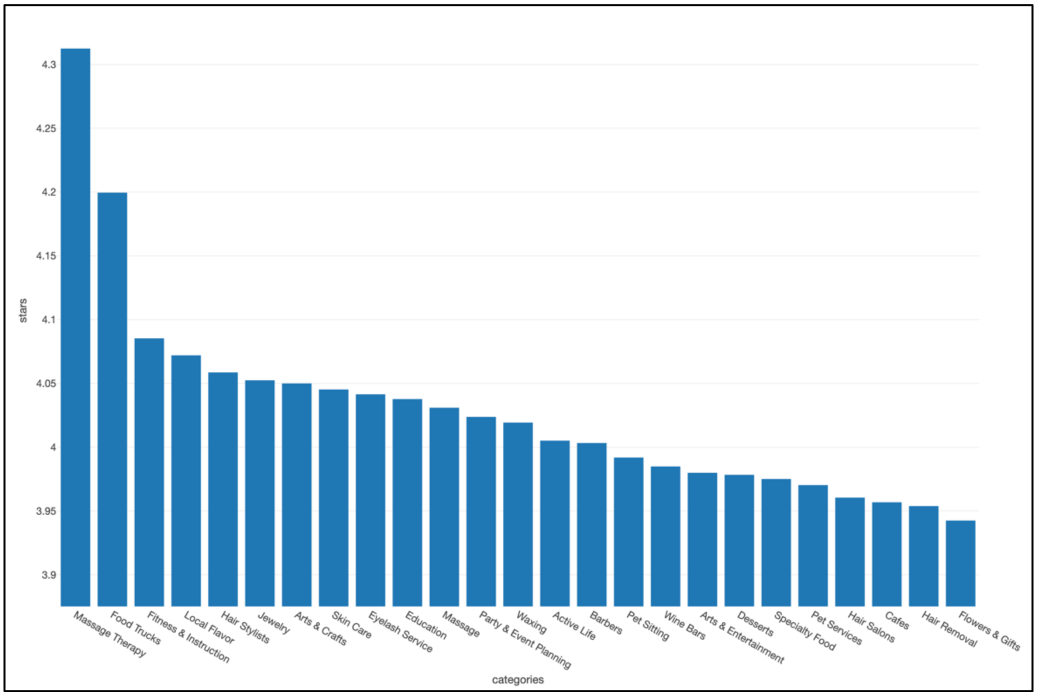
Over time, the average number of stars given to one business has decreased. The cause of this trend is unknown, but a few reasons could be 1) Yelp users are becoming more critical of restaurants or service industry businesses, 2) quality of these businesses has decreased to cause Yelp users to give them a lower rating.

4. Frequency of Business Category Labels



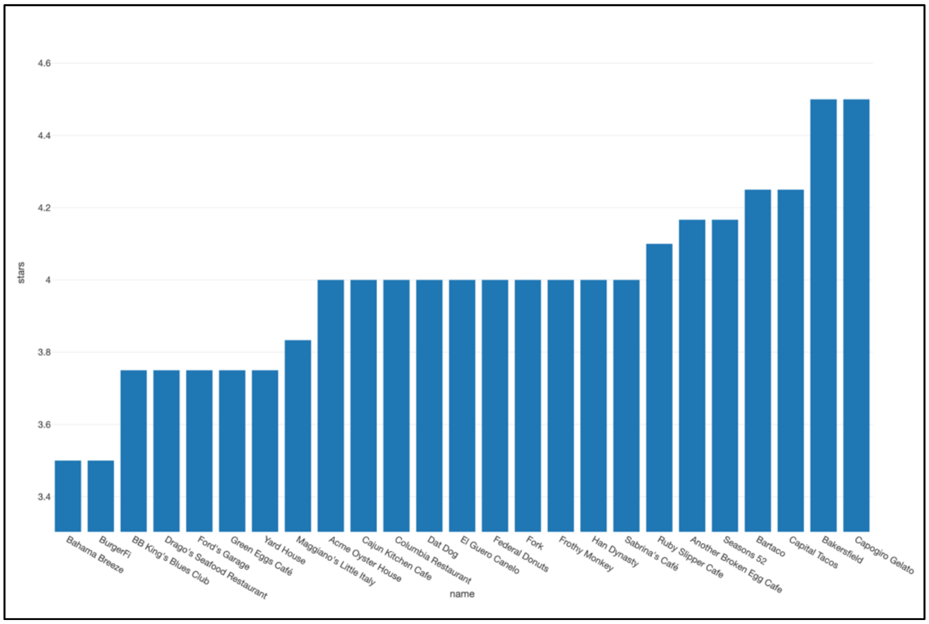
Yelp provides an extensive list of category types; businesses can choose as many of these categories as they wish to describe their business. Top categories relate directly to food (9/25) and services (5/25). The most popular category for businesses is ‘Restaurants,’ about 1/3 of businesses in the dataset describe themself as a restaurant. This makes sense because restaurants are likely the largest group for which Yelp reviews are most determinate of business.

5. **Business Categories with the Highest Ratings**



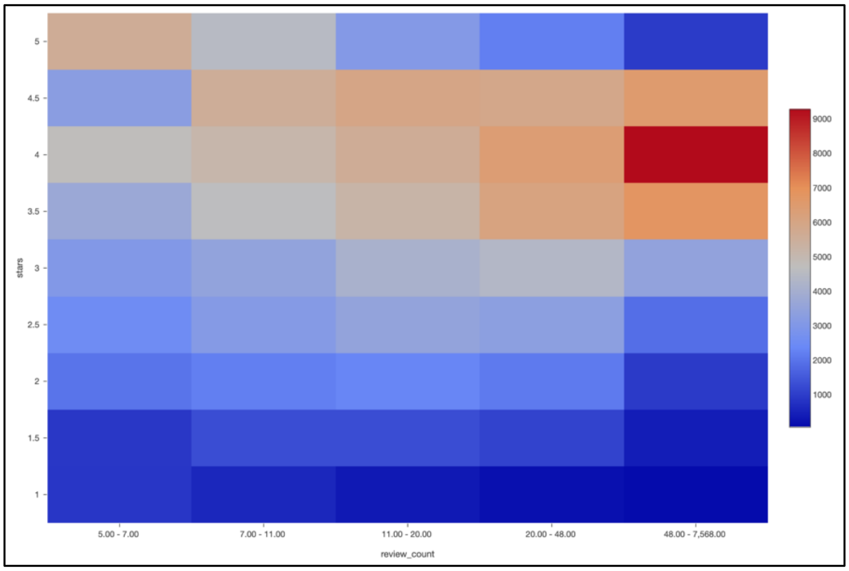
Of business categories with at least 1,000 businesses sorted under them, ‘Massage Therapy’ is the highest rated business category on average. This was determined by summarizing the separated business categories by number of rows then filtering the highest rated ones by those with at least 1,000 rows. Most of the highest rated categories deal in pleasure services (‘Massage Therapy,’ ‘Hair Stylists,’ ‘Arts & Crafts,’ ‘Party & Event Planning’); customers finish feeling physically satisfied and likely leave a positive review.

6. Businesses with the Best Yelp Reviews



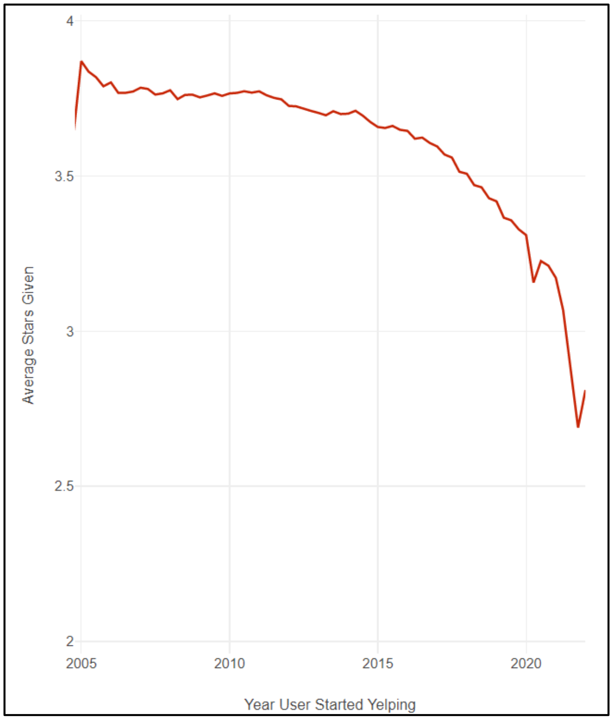
Organized by business name, these are the 25 businesses that earn the highest average star ratings (of those that have at least 500 reviews). This chart shows that most of the highest-rated businesses are restaurants or cafes, but few are rated higher than an average of 4 stars. Based on Figure 1, it makes sense that most businesses with a representative number of reviews would end up with an average rating at or below 4 stars.

7. Average Star Ratings According to Number of Reviews



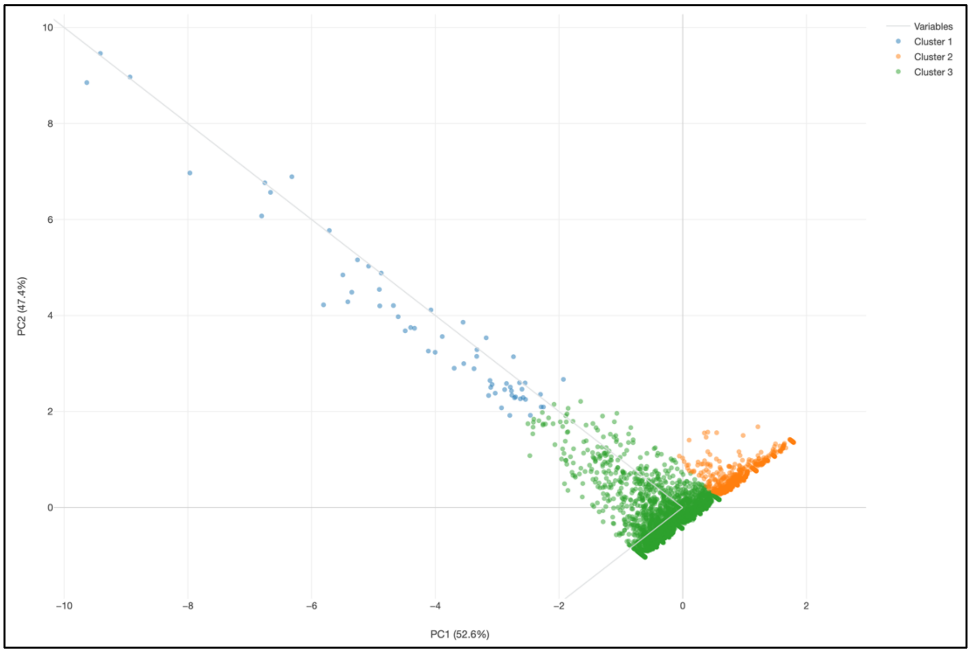
Taken from the “business” dataset, this heatmap shows the average number of stars assigned to businesses on yelp based on the number of reviews the businesses have. The ‘review_count’ variable is split based on an even number of reviews per column, stars go by halves from 1 to 5. As has been suggested throughout the data so far, 4-star averages are most common; this chart shows that this is especially true for businesses with the most reviews. Star count seems to converge and trend slightly down as businesses earn more reviews.

8. How do customers leave reviews based on when they joined yelp?



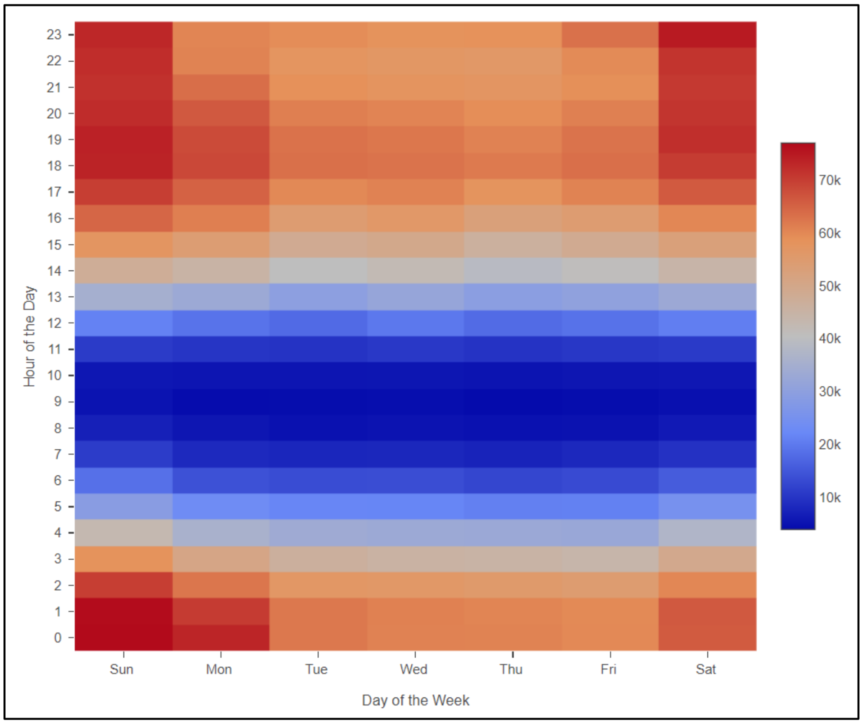
This line chart displays the average number of stars given per user based on when the user began using Yelp. The maximum average number of stars given is 3.87 stars and is thanks to the users who have been on Yelp the longest. Since then, the decrease in average stars given from year to year has steadily increased. This data suggests a couple of possibilities 1) users leave increasingly positive reviews as they are on Yelp for more time or 2) the demographic of users that joined early is more prone to leaving higher reviews.

9. Clusters Based on User Review Count and Average Review Rating



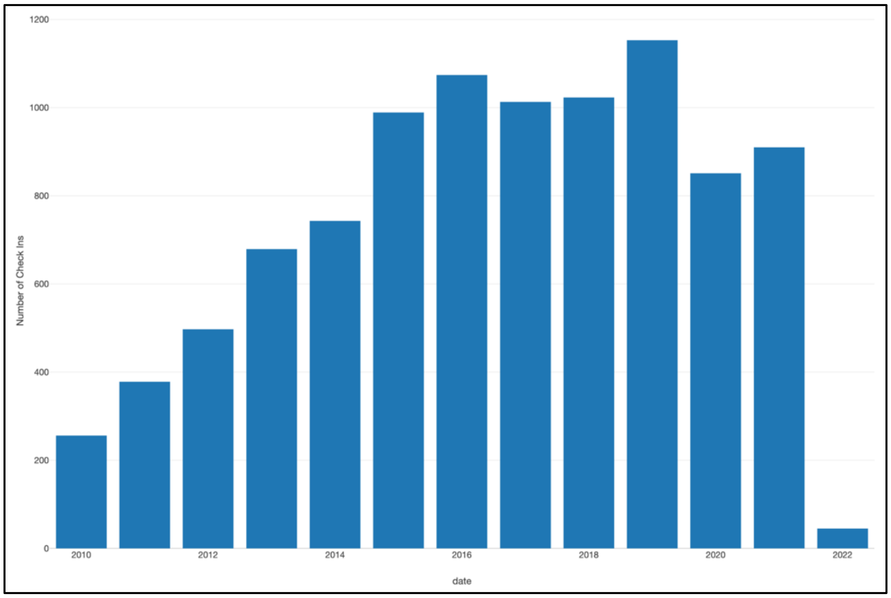
Using user review count and average number of awarded stars, this biplot demonstrates 3 clusters (the optimal number shown using the Elbow Method) of data. The largest differences to notice are between Cluster 1 and Cluster 3, which are determined primarily based on review count, and Cluster 2 and Cluster 3, which are determined primarily based on average stars. It was our expectation that generally more reviews would result in higher average ratings based on Figure 8. Instead, the associated box plot shows that none of the clusters had significantly different average star ratings despite each having significantly different numbers of reviews.

10. What days of the week and time of day are most common for leaving reviews?



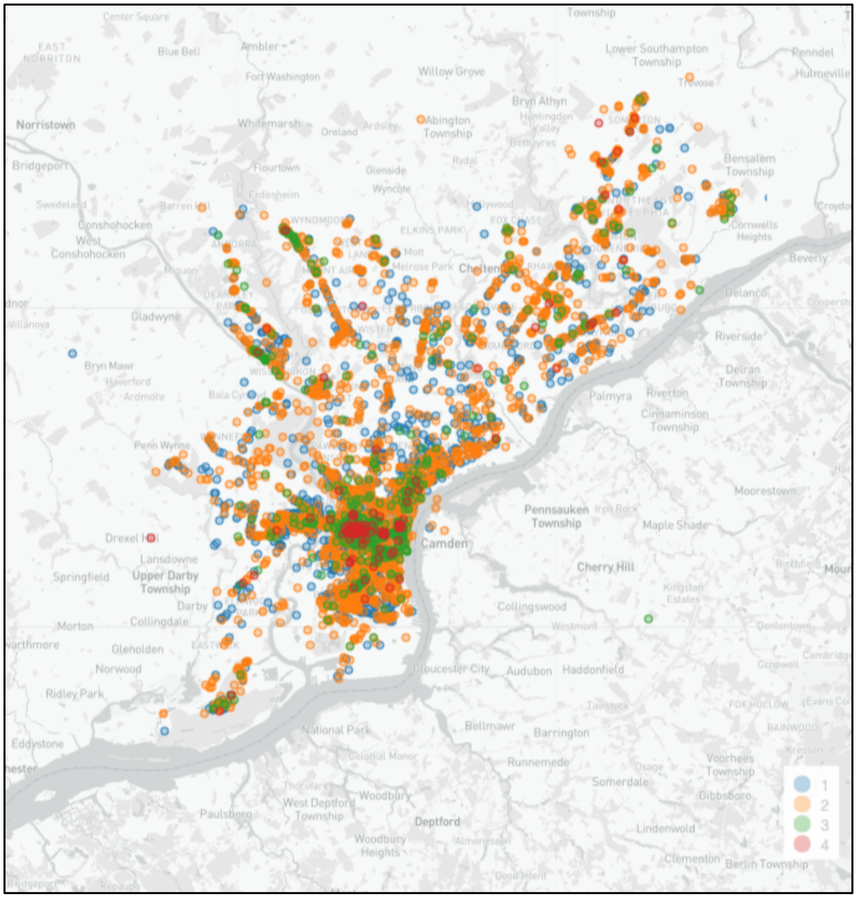
Taken from the “reviews” dataset, this heatmap shows the day and time for which the most reviews were left by users. The data suggests that reviews were most often left between the hours of 5 pm and 2 am on Saturdays and Sundays. Additionally, the least frequent times of reviews left were between 4 am and 1 pm every day of the week. One potential factor for this could be that many people are at work during the non-frequent hours, and many restaurants/bars, the most common business category in our dataset, are open during the most frequent hours.

11. Trend of check-ins over time



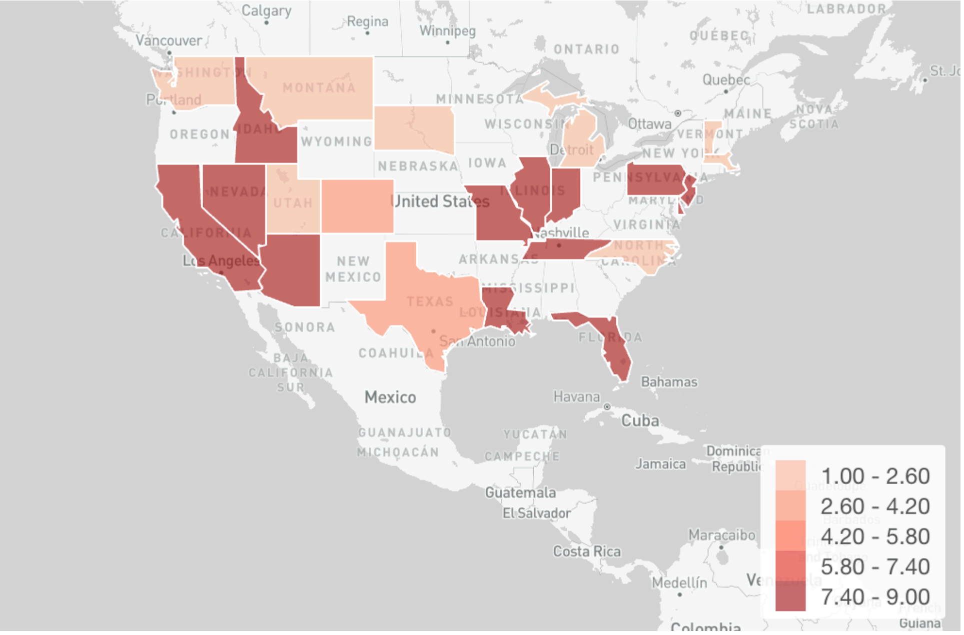
Check-ins on Yelp are “a way to keep track of the local businesses you visit and keep your friends updated with your latest comings and goings.” This graph visualizes the number of check-ins per year on Yelp. It also shows Yelp’s growth as a crowdsourcing review platform. Since 2010, when the feature was created, check-ins have grown in popularity.

12. Distribution of businesses in Philadelphia by price level



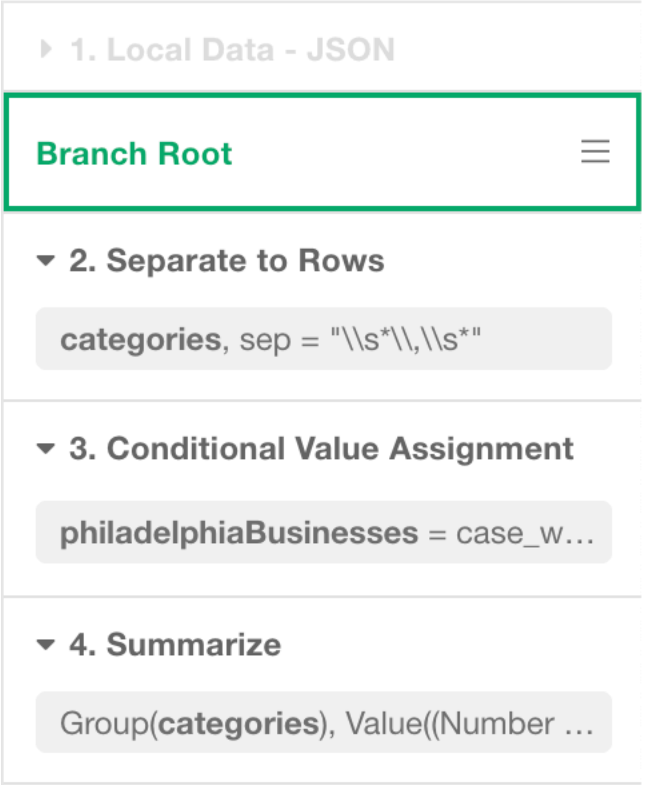
Restaurants listed as any version of the city name "Philadelphia" are sorted by price level. In the center of the city, where there are higher costs of living and higher paying jobs, there is a higher concentration of expensive restaurants. Moving away from downtown, cheaper options become more available.

13. Which states are home to the most 1-Star businesses?



The dataset’s “stars” variable was filtered to show only businesses that hold a 1-Star average review. It was then displayed in a US States Map to depict which states house the most of these businesses. The result shows a dispersal across the US with a concentration of low-ranked businesses in the Midwest and the West Coast, where states like California and Illinois have up to 9 businesses with 1-star ratings. This dispersal likely makes sense if we consider where our nation’s biggest cities lie: Florida, California, and Texas. The map was also colored red to signify the negative aspect of low business ratings, displaying dark red where they are most populous.

Data Wrangling



1. Local Data - JSON

Branch Root

2. Separate to Rows

```
categories, sep = "\\s*\\,\\s*"
```

3. Conditional Value Assignment

```
philadelphiaBusinesses = case_w…
```

4. Summarize

```
Group(categories), Value((Number …
```

We were fortunate that the Yelp dataset was largely already clean and integrated when we started with it. In our graph creation, we ran into 3 necessary data wrangling steps explained below:

1. As mentioned above, Yelp provides a list of key words that business can use to "categorize" their business. In order to analyze the most popular categories, a "Separate to Rows" by comma step was necessary. This split businesses into distinct rows for every unique category they were labeled by.
2. Not everyone knows how to correctly spell "Philadelphia." Different spelling variations were manually selected and sorted into a new column called "philadelphiaBusinesses." From there, they could be appropriately filtered and mapped by longitude.
3. Finally, we wanted to know how many times each category label was used. First grouping by category then summarizing results by number of rows per category and average stars per category allowed us to create Figure 5.

Overall Lessons Learned:

Our chosen data set reflects Yelp's user, review, business, check-in, and tip data. Our group selected this set based on our intrigue of the platform's measurements and how it uses gathered data to ultimately determine how 'good' a business is. We thought this exploration would remove the sense of subjectivity that generally exists when people determine if they like a business or not, leaving us with an objective measurement of the best businesses in our chosen categories of consideration. Moreover, in reflecting on all of our graphs, it seems as though what makes a 'good' business is not limited to a single set of characteristics. Rather, 'good' businesses are determined through a broad width of variables, and this is shown based on the variety of our data visualizations. In order to do this, we thought it was important to draw from a width of characteristics to take advantage of all of Yelp's different data avenues.

Our data visualizations provided us with a range of insights and taught us some interesting lessons. Specifically, we noticed that overall, the states that yield the highest average Yelp reviews are those that tend to have higher populations and large cities within them. We believe this to be a result of a higher concentration of businesses in areas where there are higher potential customer bases (which correlates with dense city populations), meaning that the best businesses tend to set up shop in areas where they are likely to be accessible by the most amount of people. As we were considering where highly rated businesses are generally located, there was a flip side to this consideration: specifically, the businesses that are highly rated, yet only have a couple of reviews. These were surprisingly numerous on Yelp, making up over 25% of the total businesses. This was just one of the characteristics that our group navigated through our data wrangling steps.

The largest category of businesses on Yelp was restaurants. A business that is open late, serves food, is in a big city, and caters to users that have been on Yelp the longest, maximize their chances for a consistent rating. As mentioned with Figure 9, most businesses, especially those with at least 50 reviews, will average about 4 stars with no statistically significant difference between 3 groups. Because people are people and businesses aren't perfect, 4-star businesses are likely to stay.