STAT 405/605 Final Report

Group 3: Edward Chen, Gazi Fuad, Jeff Brover, Leah Troskot, Colin Jones

December 6, 2021

Contents

1 Introduction

1.1 Data Description

Our primary dataset comes from the 'Yelp Open Dataset' [1], a collection of several datasets that are provided by Yelp. The datasets provided include business data, review data, user data, check-in data, and tip data. These were provided in .json format which we converted into csv format to read into R. The variables captured in each dataset can be described as follows:

Business:

- business_id: a unique 22 character string representing the id assigned to each business on Yelp
- name, address, city, state, postal code: strings describing the location of the business
- latitude, longitude: floats describing the latitude and longitude of the business
- stars: float representing the star rating of the business (rounded to half stars)
- review_count: integer representing the number of reviews received by the business
- is_open: binary variable to represent if business is open or closed
- attributes: object of various attributes available at business
- categories: object of various categories of the business
- hours: object of the hours the business is open

Review:

- review_id: a unique 22 character string representing the id assigned to each review on Yelp
- user_id: a unique 22 character string representing the id assigned to each Yelp user
- business_id: a unique 22 character string representing the id assigned to each business on Yelp
- stars: integer representing the star rating given to a business
- date: string representing date in format of YYYY-MM-DD
- text: string representing the text of the actual review
- useful, funny, cool: integer representing the number of useful, funny, and cool votes received by the review

User:

- user_id: a unique 22 character string representing the id assigned to each Yelp user
- name: string representing user's first name
- review_count: integer representing the number of reviews they've written
- yelping_since: string representing when the user joined Yelp, formatted like YYYY-MM-DD
- friends: an array of user_ids representing the user's friends on Yelp
- useful, funny, cool: integers representing the number of useful, funny, and cool votes sent by the user
- elite: an array of integers, the years the user was elite
- average_stars: float representing average rating of all reviews

- compliment_[hot, more, profile, cute, list, note, plain, cool, funny, writer, photos]: integers representing the number of various compliments received by the user

Check-in:

- business_id: a unique 22 character string representing the id assigned to each business on Yelp
- date: string of a comma-separated list of timestamps for each check-in formatted YYYY-MM-DD HH:MM:SS

Tip:

- text: string representing text of the tip
- date: string representing the date the tip was written, formatted YYYY-MM-DD
- compliment_count: integer representing the number of compliments the tip received
- business_id: a unique 22 character string representing the id assigned to each business on Yelp user_id: a unique 22 character string representing the id assigned to each Yelp user

We are also using a dataset available from the Centers for Disease Control (CDC) [2], which features multiple health metrics for zip codes around the country. The health metrics captured in the dataset include lack of health insurance, arthritis, binge drinking, blood pressure, cancer, asthma, cholesterol, smoking, diabetes, kidney disease, obesity, sleeping, stroke, and many more. The variables captured in the CDC dataset include:

CDC Health:

- ZCTA5: 5 digits Zip Code Tabulation Area (ZCTA) Code
- TotalPopulation: Total population of Census 2010
- health condition_CrudePrev: Model-based estimate for crude prevalence of health condition in the zip code
- health condition_Crude95CI: Estimated confidence interval for crude prevalence of health condition in the zip code

1.2 Project Introduction

We looked to primarily perform analysis on Yelp as a whole, the businesses registered within Yelp, and on the users on Yelp and their interactions with both the app and businesses. Using the supplementary CDC health data, we also looked to study differences in prevalence of some health conditions across different areas of the United States. Combining the Yelp and CDC datasets, we attempted to find connections between food-related data available on Yelp and relevant health conditions within different metropolitan areas of the US. The locations we chose to focus on (based on data availability from Yelp) were Atlanta, Austin, Boston, Boulder, Columbus, Orlando, and Portland.

2 Data Analysis on Yelp Data

2.1 Initial Plots

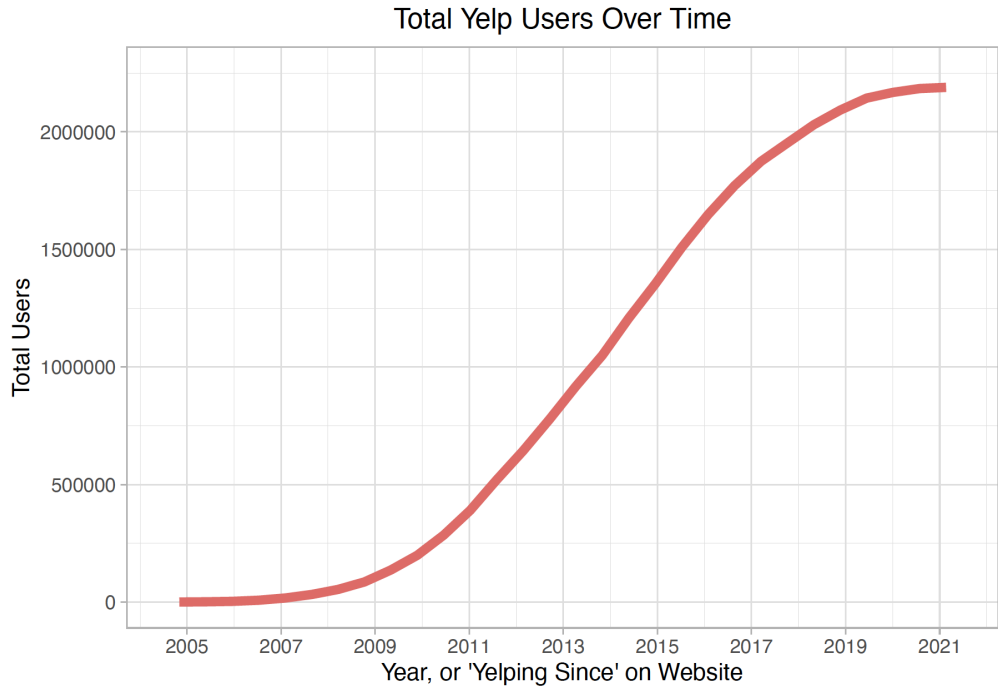


Figure 1: Overall user count on Yelp over time. From the data, we see there's a slow growth of users on the platform from 2005 to 2009. From then on, the number of users on Yelp start growing faster and faster until about 2017 from which point the growth of the platform in terms of total users starts growing slower again.

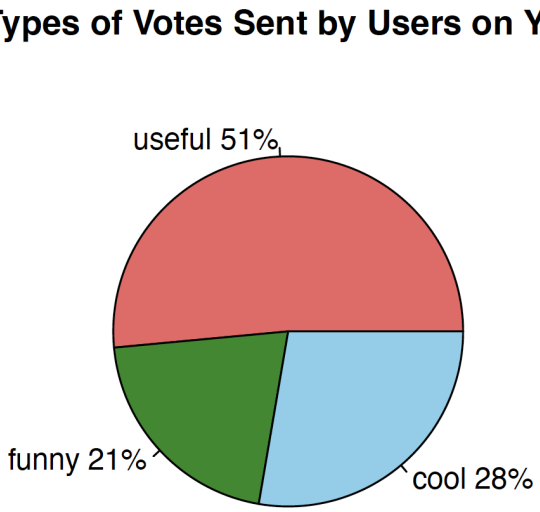


Figure 2: Pie chart of the frequency of the types of votes given to reviews within Yelp. Approximately half of all votes that users give are to indicate that a review was useful, while only about a fifth of all votes are used to signify a post was funny, with votes of cool being somewhere in between both. This might indicate the types of reviews that users are most commonly looking for when browsing through businesses on Yelp.

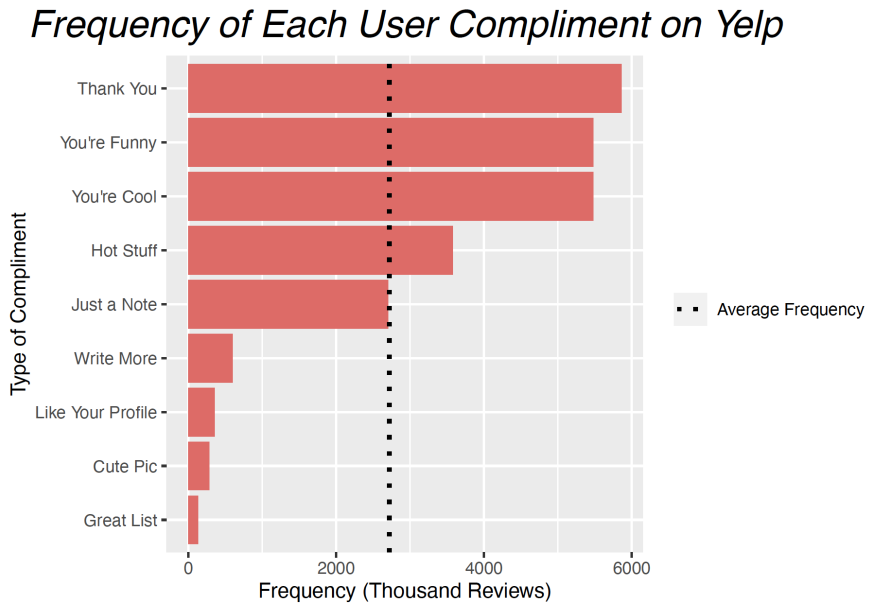


Figure 3: Frequencies of each type of user compliment on Yelp. The most common compliment types are the "Thank You", "You're Funny", and "You're Cool" compliments, which correspond to the bins "plain", "funny", and "cool" respectively which are far above the average frequency. The "Hot Stuff" ("hot") and "Just a Note" ("note) are near the average, while the rest are barely used. One possible explanation for this may be that the most commonly used compliments have been available on Yelp for a much longer time than the others.

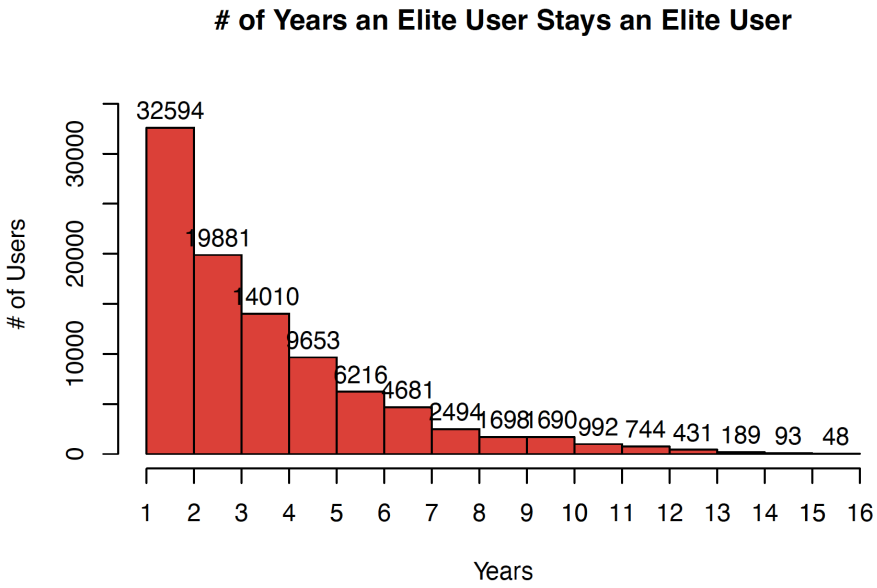


Figure 4: Histogram of the number of years a user stays "elite" on Yelp. This plot visualizes the retention rate of "elite" Yelp users for every year after the user becomes an "elite" member. It's easily shown how after the first year of membership, the number of "elite" users dramatically drops further and further each coming year. For the first 8 years, the annual retention rate ranges from somewhere between 50 to 75 percent for each year.

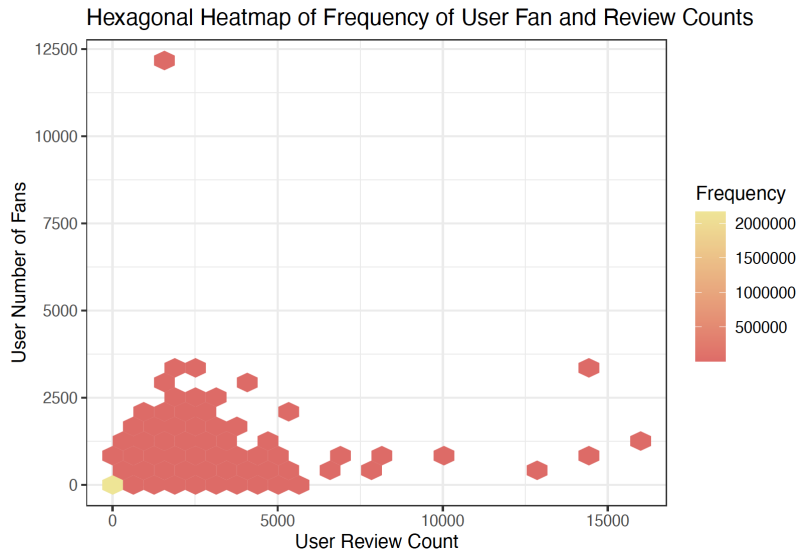


Figure 5: Heatmap plot between fans and user reviews. The overwhelming majority of Yelp users write zero comments and have zero fans. Most of the users who write comments or have some fans are in the lower range of these variables as well. However, there are some outliers with either high amounts of fans or reviews.

2.2 COVID's Impact on Yelp

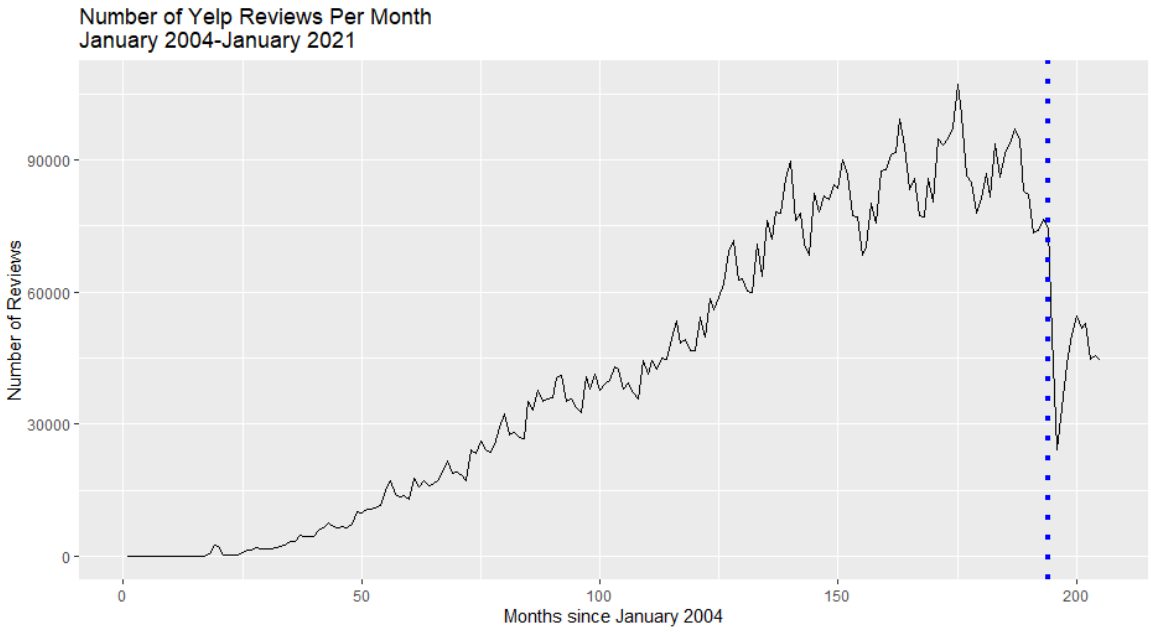


Figure 6: Line graph of the reviews per month on Yelp in the 8 metro areas analyzed since 2004. The blue dotted line signifies March 2020, when COVID-19 hit the US. In April 2020, Yelp's CEO announced that as a result of COVID-19 and lockdown related measures, Yelp searches had decreased by around 70%. This graph shows that effect, as Yelp reviews per month had steadily increased to a peak of more than 100,000 in 2019 before a sharp drop as a result of COVID.

Figure 7: Four bar plots showing how COVID affected different services on Yelp. The data is taken from the year preceding March 2020 and the year after March 2020 in the 8 metro areas analyzed. Only businesses with more than 100 reviews are included. The top left graph shows the 11 service types that had the most reviews. As shown, nightlife and bars were hit the hardest. The top right graph shows the services that had the largest magnitude change, with restaurants having the biggest change (20,000 reviews decrease). The bottom left graph shows the services with the smallest % decrease. Interestingly, the contractors category experienced an increase in reviews. Lastly, the bottom right graph shows the categories with the biggest % decreases. As expected, travel services and airport both had a more than 80% decrease.

2.3 Sentiment & Review Analysis

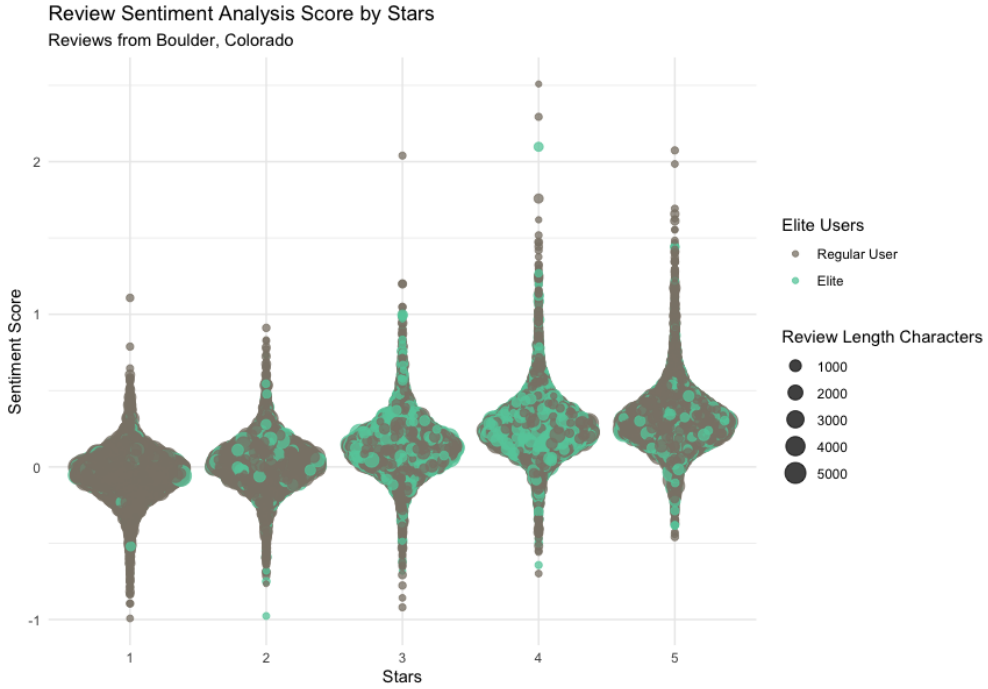
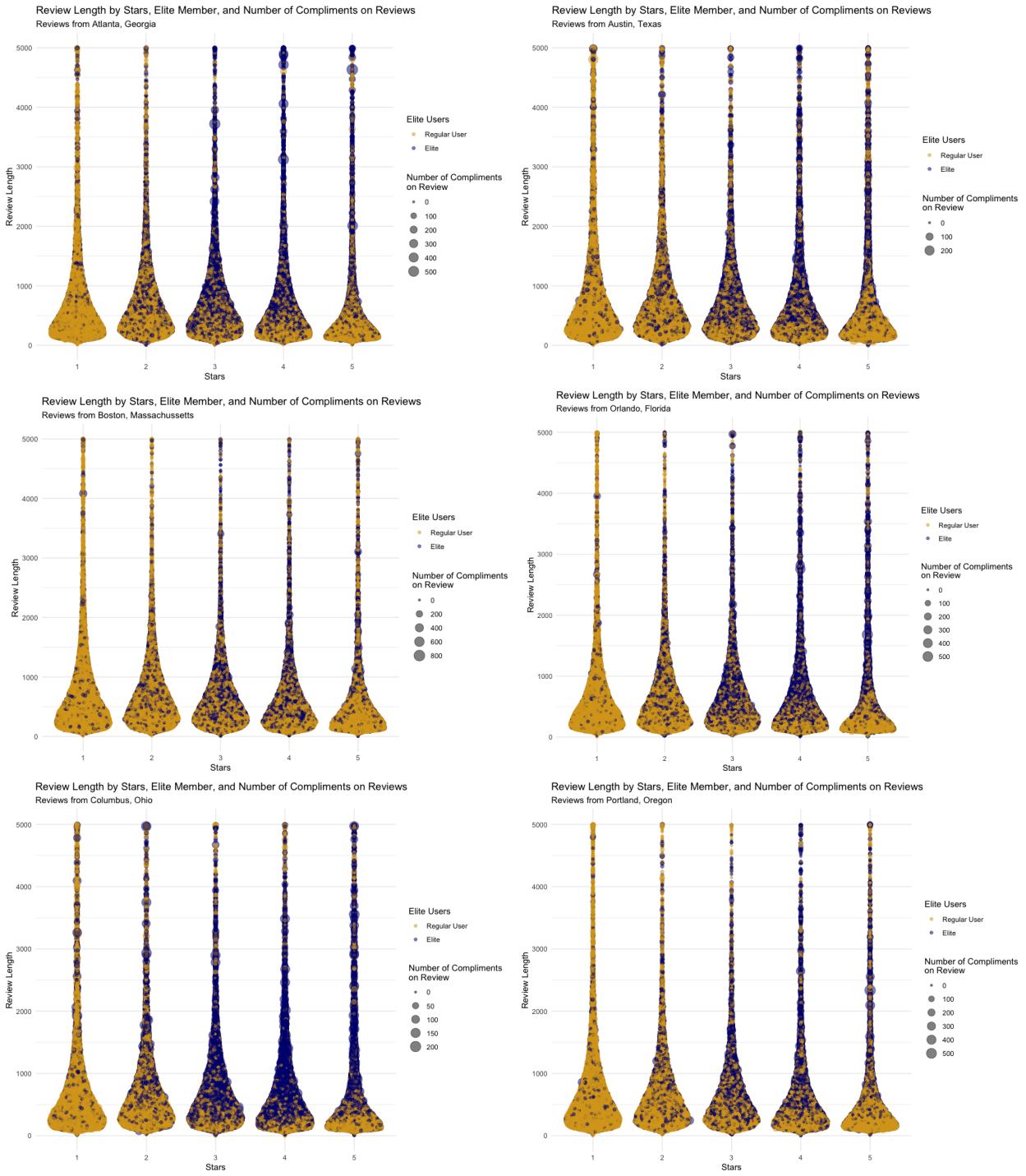


Figure 8: Sentiment analysis of reviews based on stars given in Boulder. As expected, the higher star rating given in the review, the more likely it is for the text within the review to be positively worded. One interesting thing to note though is that the sentiment results between 4 star and 5 star reviews is only marginally different, possibly indicating that people that write 4 star reviews are more likely to be thoughtful in their review, leading to a higher sentiment score. Additionally, we can see that elites (turquoise) are more heavily grouped in the 3 and 4 star categories and tend to have longer reviews (size) which may indicate that these more frequent Yelpers spend more time thinking about what to say in their review.



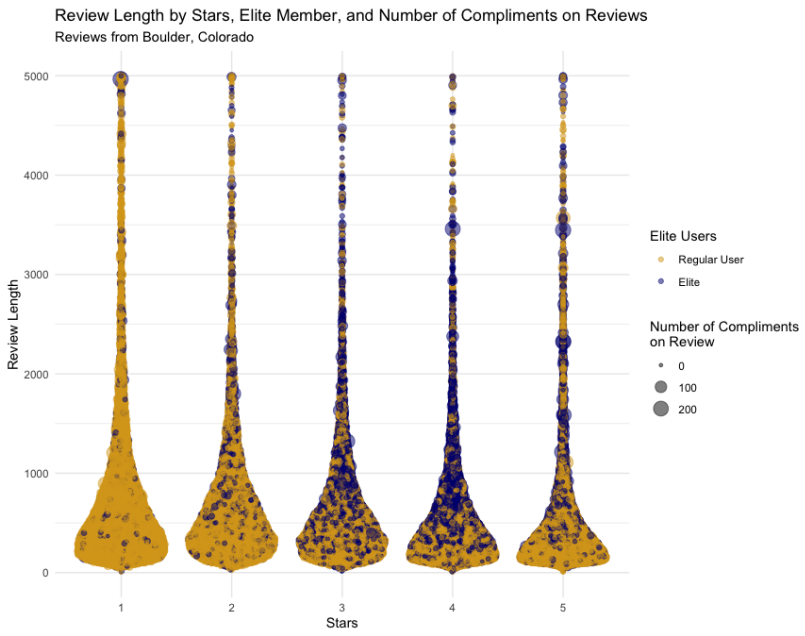
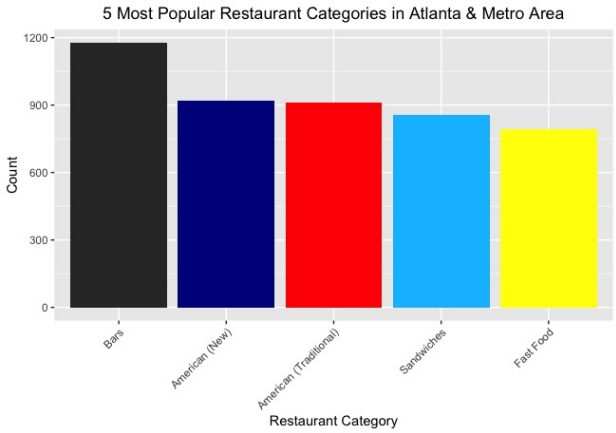


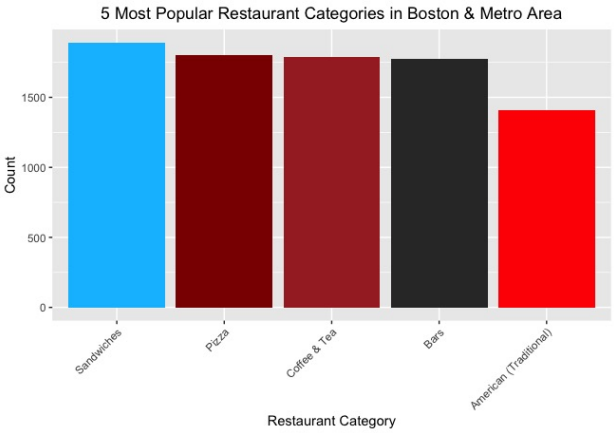
Figure 9: Above we display 7 Beeswarm Plots of the Review Length (in characters) by the star given for that review. Additionally, we display whether or not the reviewer was an 'Elite' Yelper which is displayed by the purple points. Finally, the number of compliments that the review received from other yelpers is displayed in the size of each point. We can see that in each city Elite Yelpers were more likely to give 2, 3, or 4-star reviews than non-Elite Yelpers. This could indicate that Elite yelpers are more thoughtfully considering the correct star for the review while non-Elite users are more likely to group to the extremes of 1 and 5-star reviews. Additionally, we can see that most reviews are shorter because the bulk of the points are at the bottom end of each value of stars. However, we can see a thicker vertical tail in the 1 and 2-star reviews signifying longer reviews which may indicate some longer rants produced by poor experiences. Finally, most cities seem to have similar balance of Elite to non-Elite users signified by the distribution of yellow to purple points except for Columbus which seems to have significantly more purple points in comparison to yellow. This is in contrast to Boston that seems to have relatively fewer Elite users.

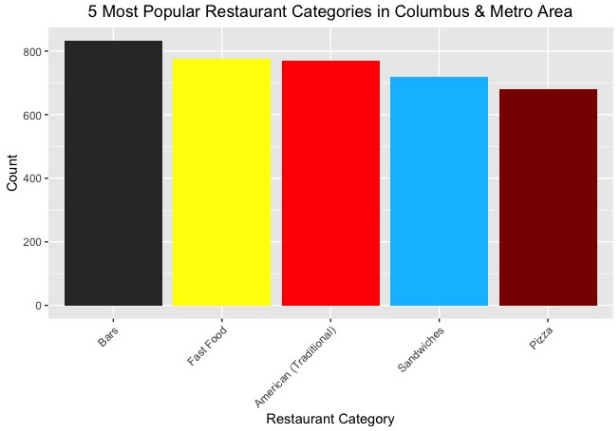
2.4 Cuisine Types by Area

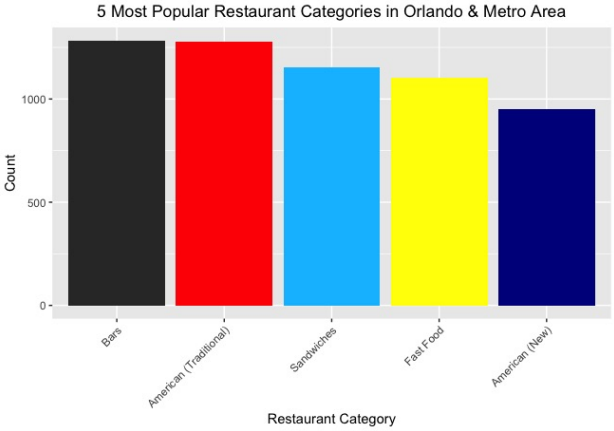
Cuisine Type	Atlanta	Austin	Boston	Boulder	Columbus	Orlando	Portland
Fast Food	3.98	3.26	1.98	2.6	5.29	4.18	2.29
Bars	12.06	10	10.04	11.11	10.35	10.42	12.57

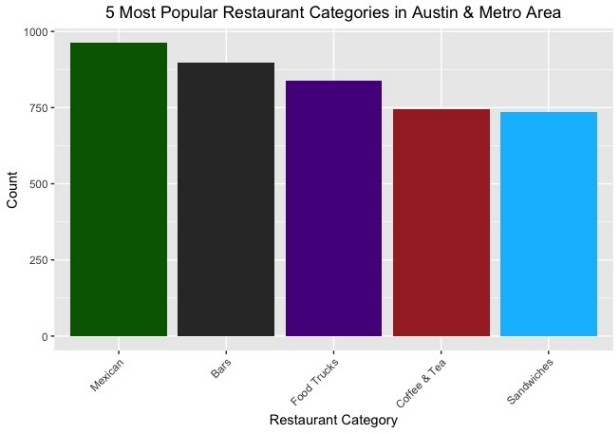
Table 1: Percentage of unhealthy cuisine types per metropolitan area. Columbus has the highest percentage of Fast Food Restaurants out of all Restaurant type businesses, and Atlanta has the highest percentage of Bars. In each metropolitan area, at least 10% of all restaurant type businesses are Bars, while the percentage of Fast Food is quite smaller.

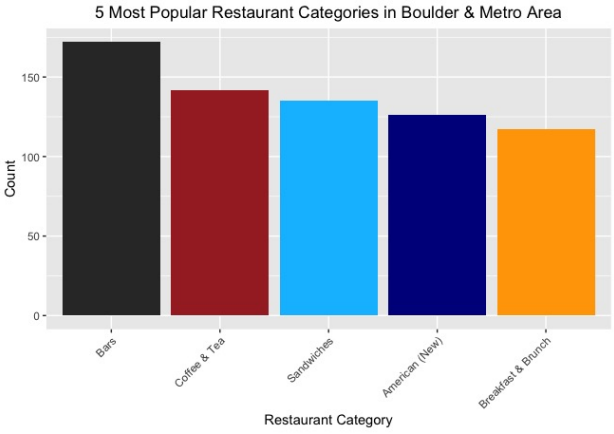

5 Most Popular Restaurant Categories in Atlanta & Metro Area
Count
1200
900
600
300
0
Bars
American (New)
American (Traditional)
Sandwiches
Fast Food
Restaurant Category


5 Most Popular Restaurant Categories in Boston & Metro Area
Count
1500
1000
500
0
Sandwiches
Pizza
Coffee & Tea
Bars
American (Traditional)
Restaurant Category


5 Most Popular Restaurant Categories in Columbus & Metro Area
Count
800
600
400
200
0
Bars
Fast Food
American (Traditional)
Sandwiches
Pizza
Restaurant Category


5 Most Popular Restaurant Categories in Orlando & Metro Area
Count
1000
500
0
Bars
American (Traditional)
Sandwiches
Fast Food
American (New)
Restaurant Category


5 Most Popular Restaurant Categories in Austin & Metro Area
Count
1000
750
500
250
0
Mexican
Bars
Food Trucks
Coffee & Tea
Sandwiches
Restaurant Category


5 Most Popular Restaurant Categories in Boulder & Metro Area
Count
150
100
50
0
Bars
Coffee & Tea
Sandwiches
American (New)
Breakfast & Brunch
Restaurant Category

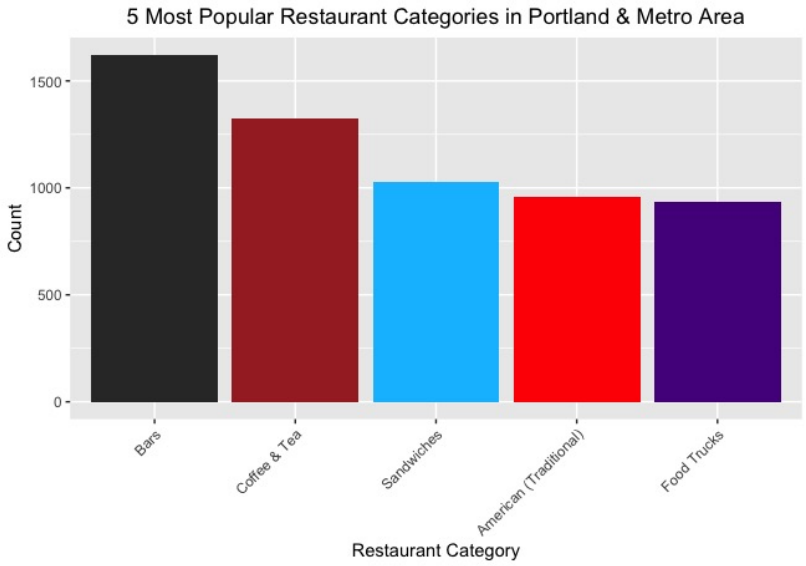


Figure 10: Restaurant category analysis, or top cuisine types, across metropolitan areas in the United States. The restaurant categories that appear in the top 5 for all of these cities are Bars and Sandwiches. Bars were the top category in 5 of the 7 metropolitan areas analyzed. American (Traditional) appears in the plots for all eastern cities and Coffee & Tea appears in the plots for all western cities. In addition, analysis showcased unique cuisine preferences for each metropolitan area, such as Mexican and Food Trucks in the Austin Metro Area.

2.5 Heatmap Analysis

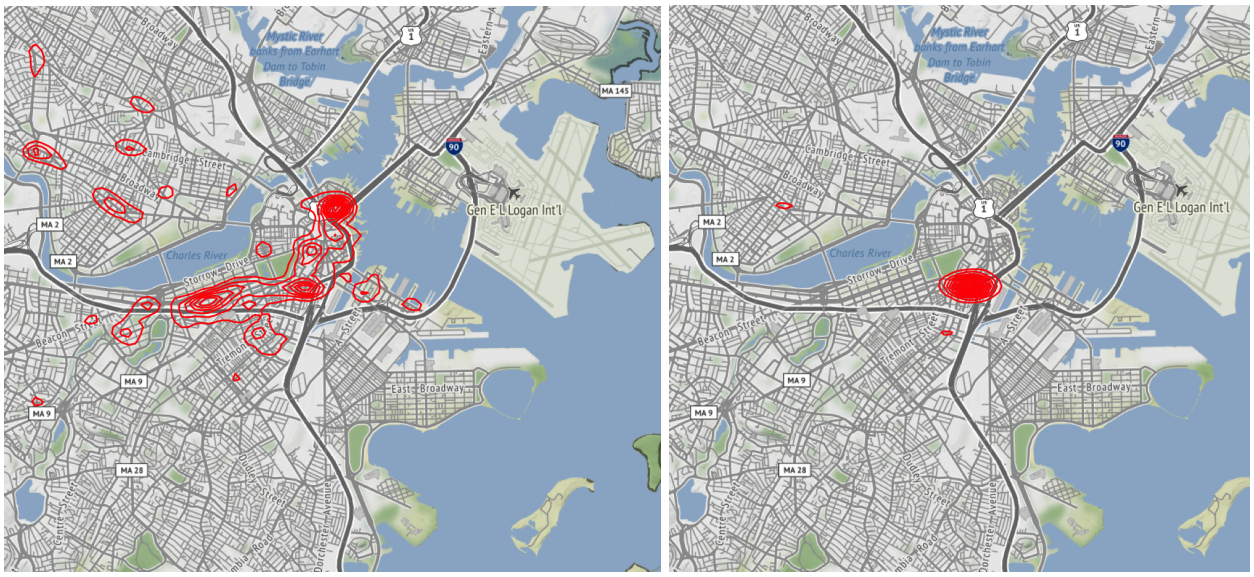


Figure 11: Heatmap of Yelp reviews for all businesses in Boston (left) and heatmap of Yelp reviews for Chinese restaurants in Boston (right). For Boston, many of the reviews for all businesses are coming from Downtown, its Chinatown, and along Newbury Street with smaller pockets in different parts of the city, including Cambridge. For Chinese restaurants in Boulder, it seems the Yelp reviews are exclusively coming from its Chinatown district.

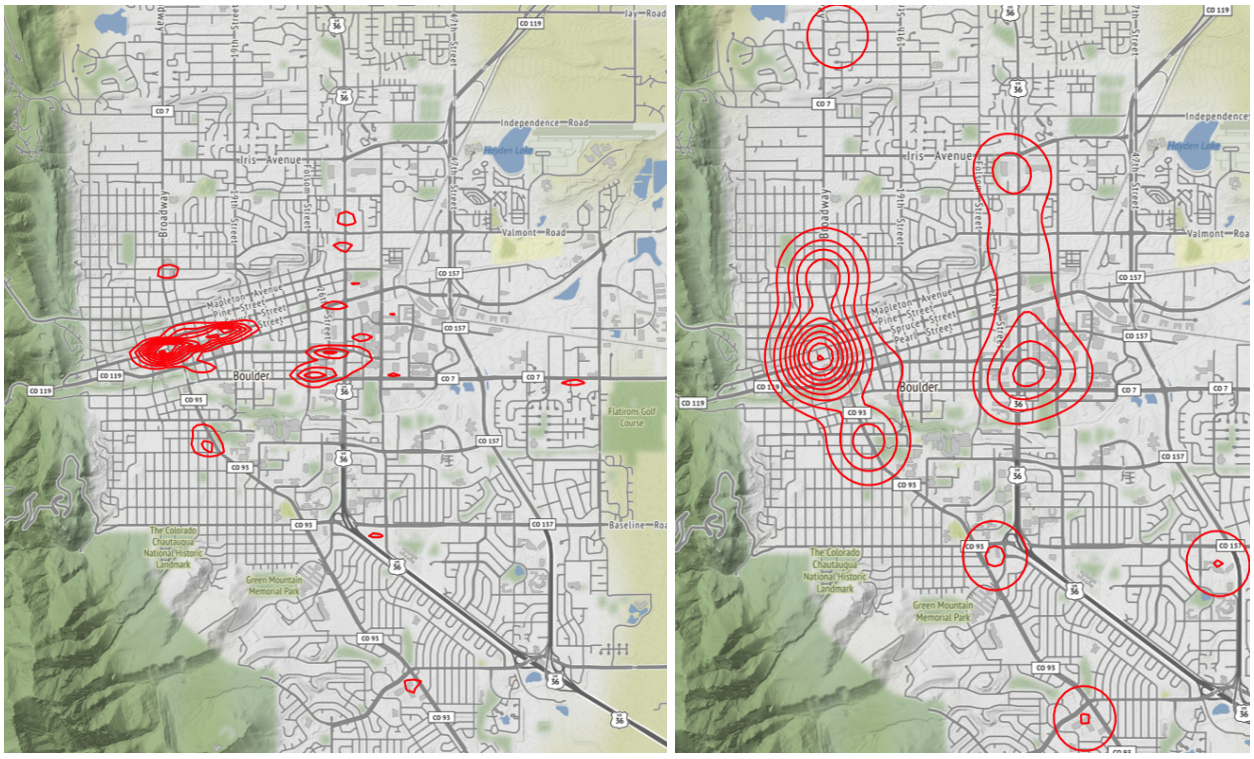


Figure 12: Heatmap of Yelp reviews for all businesses in Boulder (left) and heatmap of Yelp reviews for Chinese restaurants in Boulder (right). As expected, much of the activity is occurring in the "downtown" region of Boulder, with smaller pockets around in random parts of the city. When looking at reviews from Chinese restaurants in Boulder, it's seen how spread out is across different parts of the city which is a stark contrast to that of Boston.

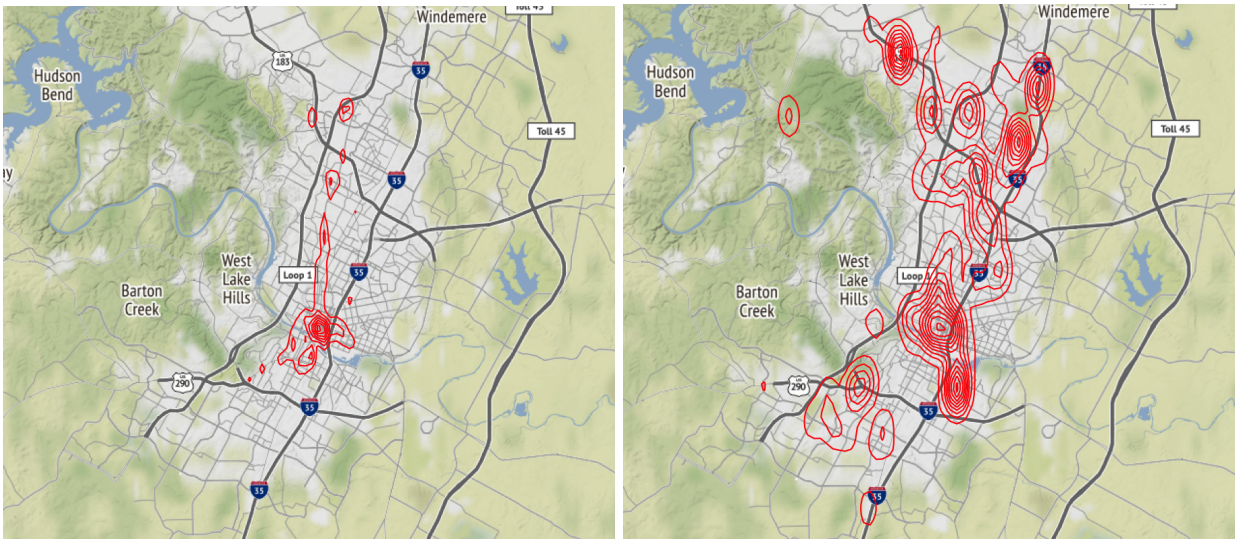


Figure 13: Heatmap of Yelp reviews for all businesses in Austin (left) and heatmap of Yelp reviews for Chinese restaurants in Austin (right). Like Boulder and Boston, Yelp reviews are most coming from Austin;s downtown area as well. In regards to Chinese restaurants in Austin, the heatmap is almost like a compromise between Boulder and Boston in that the reviews seem to be spread out, yet, there are some heavily concentrated areas within Austin and its neighboring areas.

3 Yelp & Health Data

3.1 Health Plots

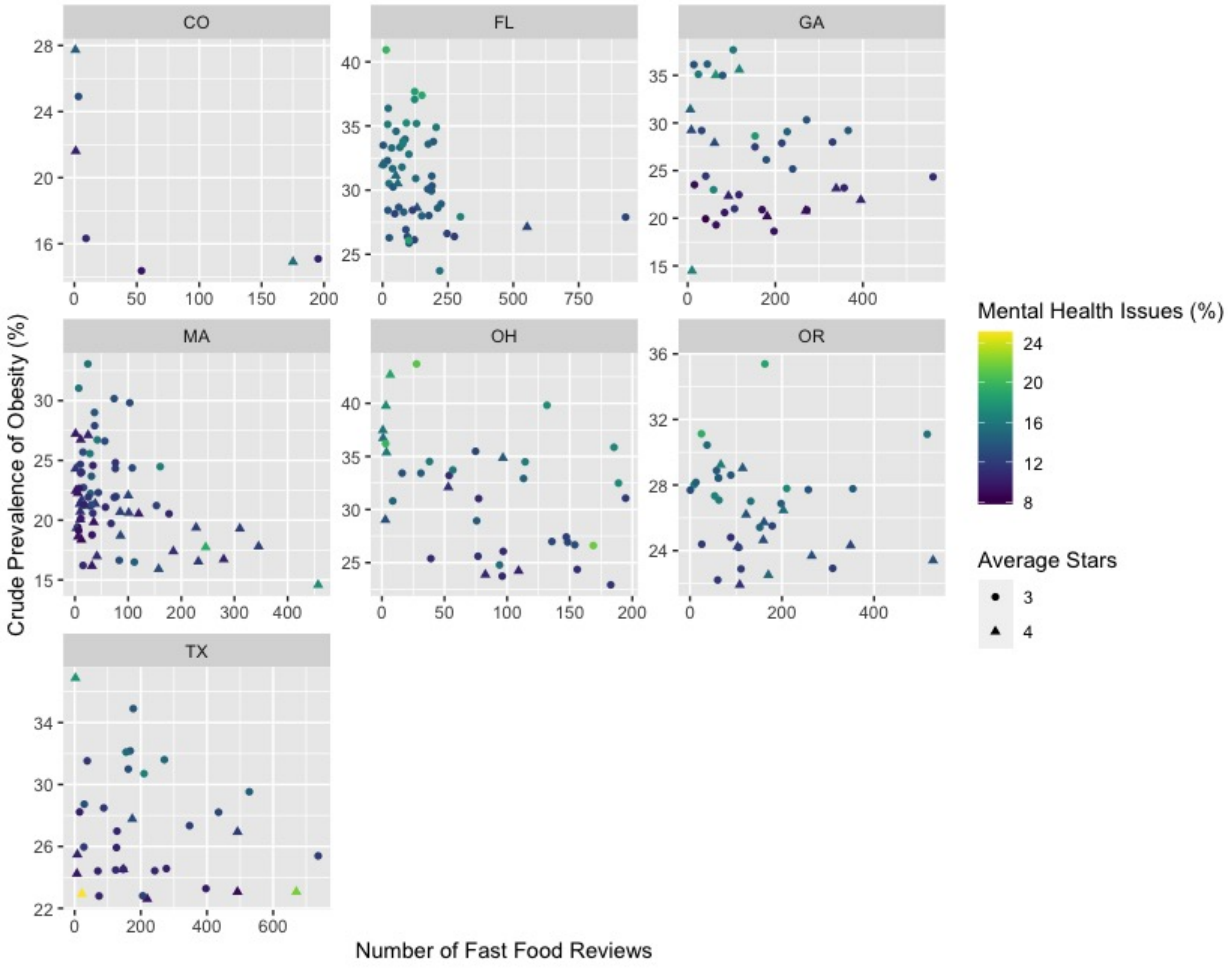


Figure 14: Number of fast food reviews vs. crude prevalence obesity (percentage). In this plot, the number of fast food reviews in our seven cities of interest, representing the relative popularity of fast food in the area, is plotted against the crude prevalence of obesity in the area as a percentage of the population for each zip code. Additionally, lighter colored points correspond to areas with more mental health issue, while the shape of the points depend on the average stars of fast food in the area. To make a distinction, while the number of fast food restaurants represents the frequency of visits, the average stars represents the popularity of visits, meaning a zip code with 4 average stars and a higher number of reviews is both well-visited and popular. This data unsurprisingly reveals an overall strong negative correlation between obesity and the amount of fast food eaten. However, taking into account the color and shape factors other patterns emerge such as the overall dis-popularity of fast food joints in terms of average stars and its relationship with mental health that varies drastically by city. For instance in Boston, there appears to be a positive correlation between fast food and mental health issues, with consistently the highest rated fast food, while in Orlando, the city with the largest heart issues, almost every zip code has lowly rated fast food places and an overall negative correlation between fast food and mental health problems.

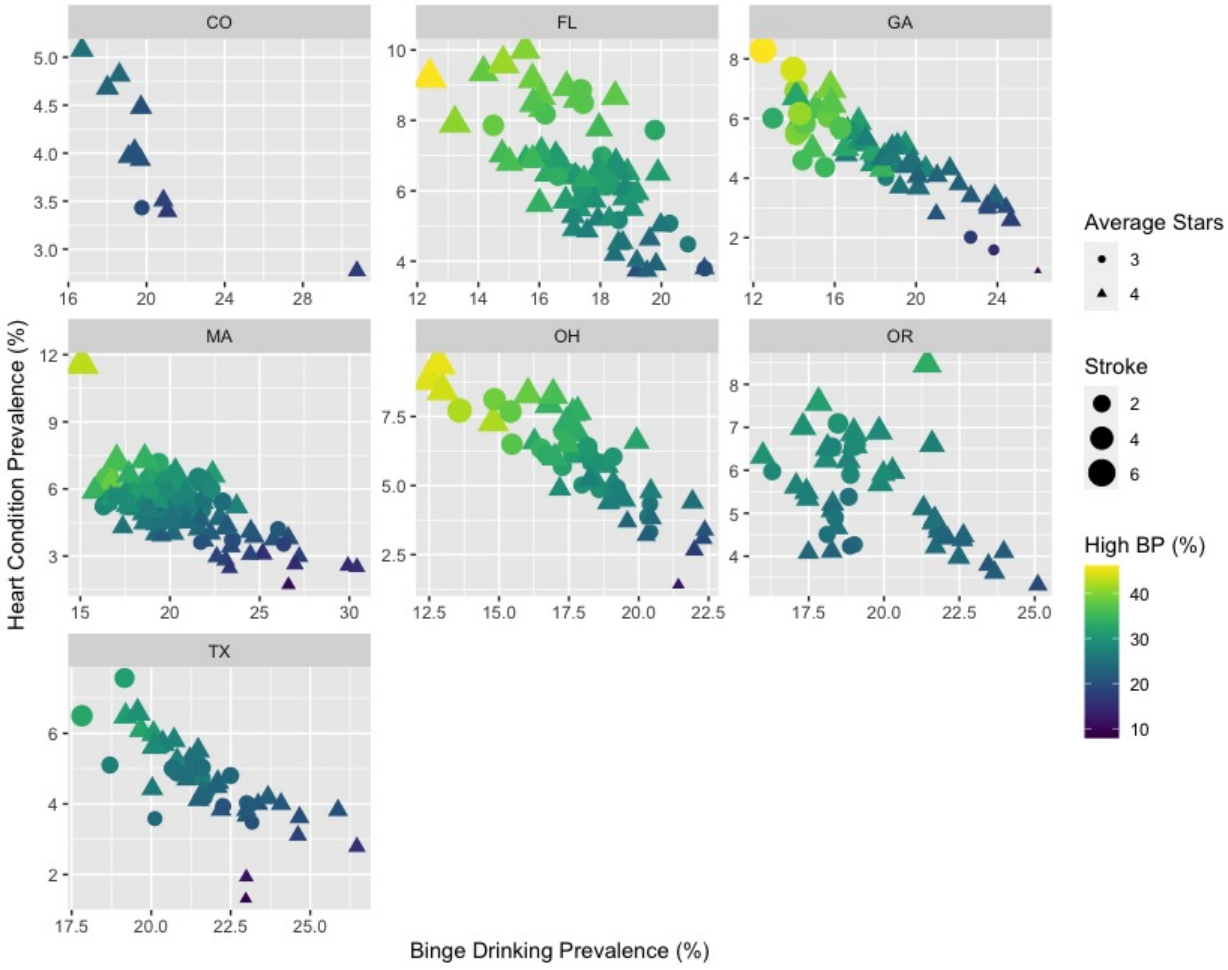


Figure 15: Binge Drinking Prevalence vs. Heart Condition Prevalence. This plot attempts to compare the relationship between four different health-related conditions, binge drinking, heart conditions, strokes, and high blood pressure, in addition to the average popularity of American restaurants in a zip code. Interestingly, the higher the binge drinking rate in a zip code, the lower the probability of there being heart conditions in the area. This surprising result is most likely because young people (who are more likely to be binge drinkers, have less health issues, visit the doctor less, and have less heart conditions) aggregate together in areas while older people (who are the opposite) also tend to aggregate together, leading to this type of trend. Beyond this trend, as expected, high blood pressure is also strongly positively correlated with heart issue and strokes, with similar reasoning for a negative correlation with binge drinking. Finally, it also appears that all health conditions, with the exception of binge drinking appear to occur less frequently in places where American restaurants, which also are likely to serve alcohol, are well-liked.

3.2 Killer Plot

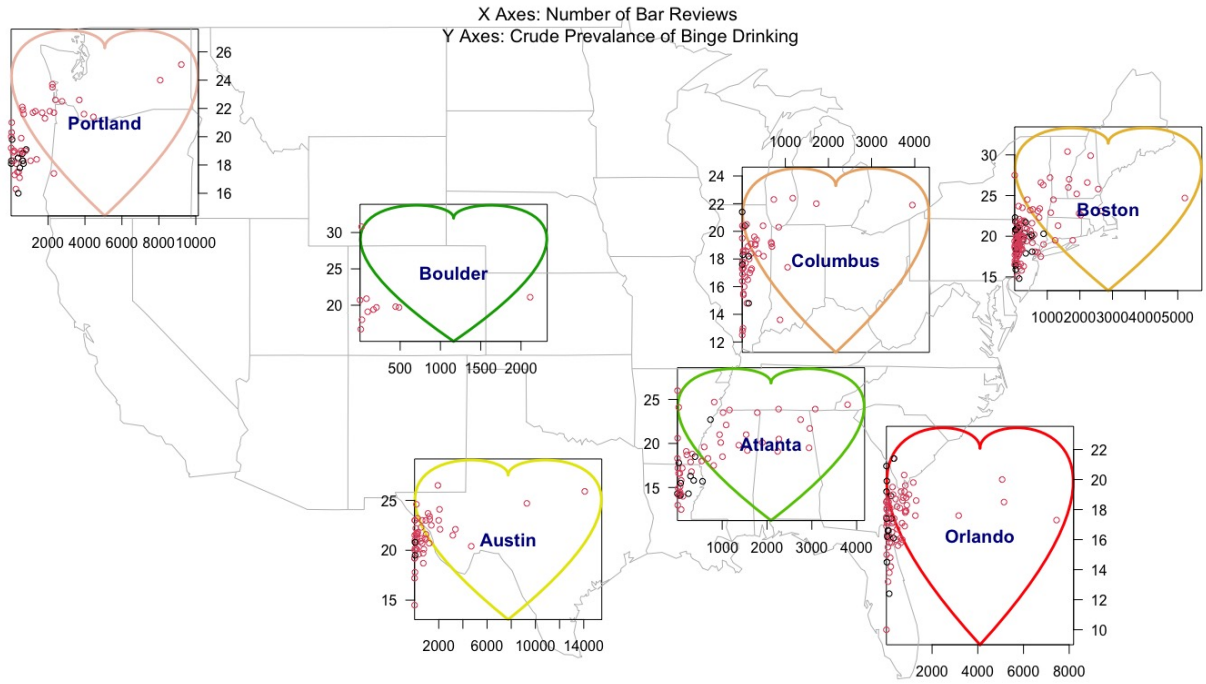


Figure 16: City Health Statistics vs. The Popularity of Bars

This "killer plot" shows seven different hearts, each approximately centered at the location of one of the seven main cities of our data set, Portland, Boulder, Austin, Atlanta, Columbus, Orlando. The height and color of the heart is representative of the crude prevalence of heart disease within each city, where a larger or more red heart represents the cities with more heart disease, such as Orlando and Portland, while the smaller greener hearts, such as Boulder and Atlanta represent those with less heart disease prevalence. Then, superimposed on each city is a graph with free scales where the x-axis represents the number of tips left on Yelp by customers at businesses classified as bars, which serves as a proxy variable for the overall popularity of bars in the area. Meanwhile, the y-axis represents the crude prevalence (percentage) of binge drinking in each city, where each data point overall represents a different zip code within the city. Red data points correspond to cities with an average of 4 out of 5 stars for bar reviews, while black data points represents 3 out of 5 stars. The results of the graph reveal a strong positive correlation between the number of bar tips left in a city and the prevalence of binge drinking across all seven cities. Additionally, for all seven cities it appears to be of the quadratic form, experiencing "diminishing returns" wherein as the number of bar tips increases, the binge drinking prevalence increases at a decreasing rate, implying that if the difference between having many bars and many many bars is much smaller than the difference between having only a couple bars to a few bars. Additionally, when adding in the variable of heart disease, it appears that more bars, and more binge drinking leads to more heart disease, but the pattern is not as strong, as for instance, Orlando, the city with the highest prevalence of heart disease, does not have the highest overall prevalence of binge drinking with Austin, the city with the median amount of heart disease, having the highest amount of binge drinking. In summary, however, this graph reveals the connection between the number of bars in the area and various health ailments associated with their popularity. Additionally, this graph can be generalized to extend to up to 50 different health conditions, as well as over 200 different classifications of restaurants.

4 Conclusion

This project resulted in the discovery of many unique insights and newfound knowledge. The ability to work with data from Yelp, an app almost all of us have used at some point, allowed us to have a deeper connection and better contextualize the results we received.

We were able to analyze Yelp's growth and its ability to both retain and engage their users to keep using their services. In addition to that, we were even able to study the COVID-19 pandemic's effect on businesses featured on Yelp, something we had seen play out live. By specifically analyzing different locations and cities within the United States, we were able to extrapolate some similarities and differences between various regions of the US. This included information such as the most popular food categories, the way reviews were written by users, and the distribution of businesses and specific cuisines in cities. Utilizing the external health data from the CDC, our plots yielded interesting results that were both expected and unexpected. This allowed us to critically think about the way our data was collected and structured to ensure that our commentary and analysis was reasonable.

In the future, our work can help identify how to outline public health recommendations related to people's diets to better improve overall health within communities around the country. Further analysis on the Yelp data could end up assisting Yelp in improving its app and services while boosting user experience. As consumers and businesses continue to navigate the ongoing pandemic that has pushed many things online, these groups will need to continue relying on services like Yelp. For Yelp itself, further work can be done to model its growth and help identify which new features can be deployed to ensure that it maintains and expands its business as much as possible.

References

[1] “Yelp Open Dataset,” Yelp Dataset. [Online]. Available: https://www.yelp.com/dataset.

[2] ”PLACES: ZCTA Data (GIS Friendly Format), 2020 release,” Centers for Disease Control (CDC). [Online]. Available: https://chronicdata.cdc.gov/500-Cities-Places/PLACES-ZCTA-Data-GIS-Friendly-Format-2020-release/kee5-23sr