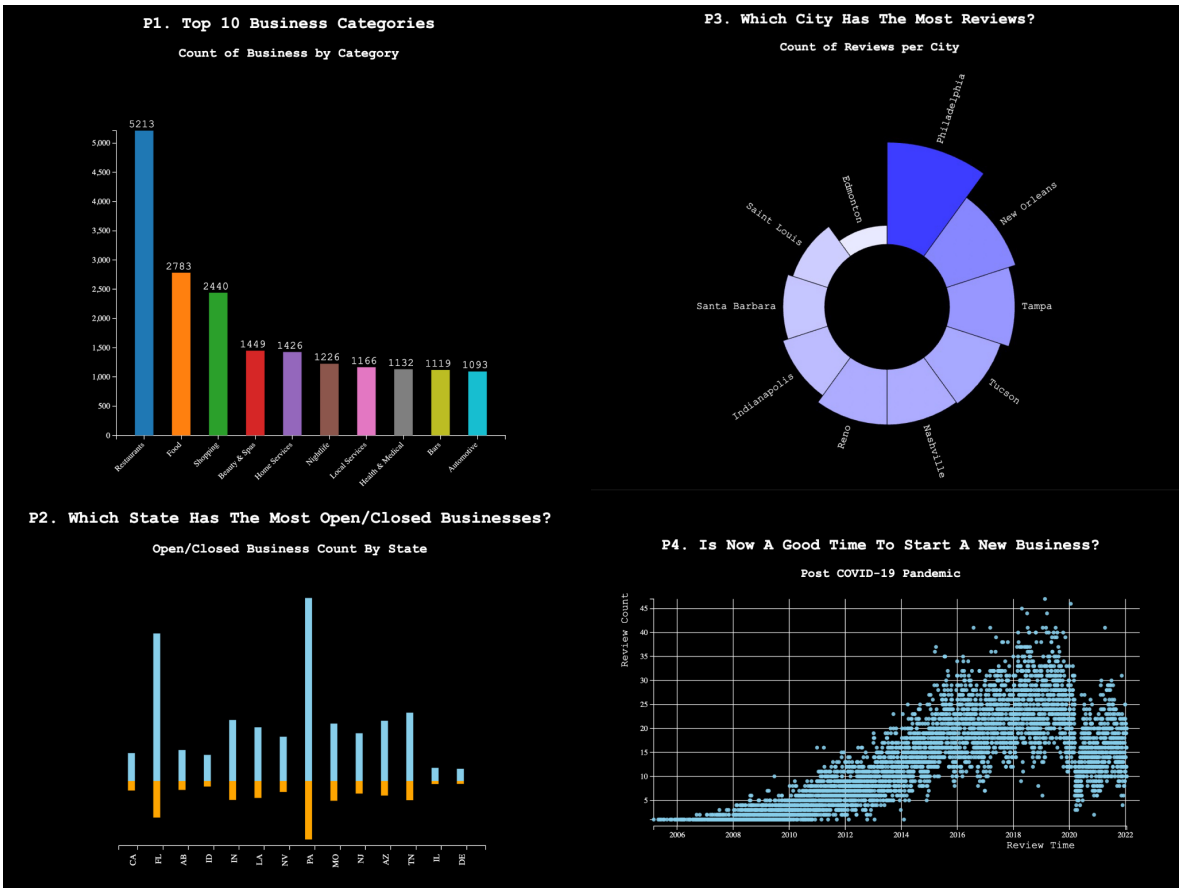
Project 1 - Final Report

Team members: Zhi Lin (zl846), Yun Zhou (yz2685), Tianchen Wang (tw537)

Visualizations Overview



Description of The Data

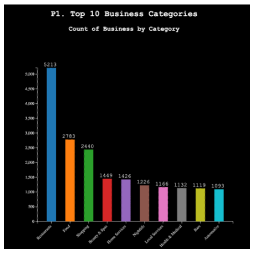
We downloaded two Yelp's business and review datasets from www.Kaggle.com, of which the original size was over 5GB with more than 7 million records. Since this was a subset of the daily Yelp data, we found out it lacked data for some major cities in the U.S., including New York City. Thus, our visualization was limited to the locations in the dataset.

Per the requirements of the project, we reduced the dataset size to be around 2MB by randomly selecting data points. To be more concrete, we preprocessed the original datasets and based our project on three subsets: **yelp_business_10percent.json**, **yelp_review_star_2percent.json**, and **yelp_review_time_1percent.json**. For the review datasets, we cleaned the original yelp_review.json dataset by deleting some attributes (e.g. reviews and locations), and only kept two of interest: "star" and "date". The "star" column was the star rating submitted by a reviewer

on Yelp, while the "date" was the review submission date. Furthermore, we created two versions of the yelp_review dataset for different graphs. First, the **yelp_review_star_2percent.json** file contained 139,806 star rating records, which was 2% of the original dataset. Second, the **yelp_review_time_1percent.json** file consists of 69,903 records with star rating and review date attributes, which was 1% of the original dataset. For the business dataset, we kept the **city**, **state**, **is_open** and **category** attributes, and randomly selected 15,035 records, or 10% of the original dataset, and saved the data in the **yelp_business_10percent** file.

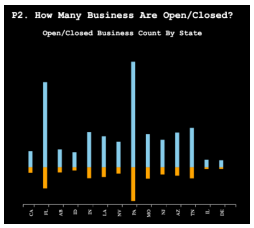
Overview of Design Rationale

Overall, our data visualizations were designed for our business goal, which was to explore the best business opportunity in the next couple years. In summary, we have leveraged bar charts and scatter plots to conduct competitive analysis and trend exploration. The following are the details of each graph:



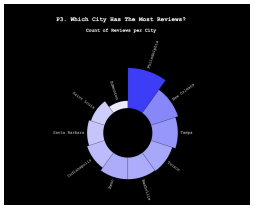
Plot 1 Bar Chart:

- Marks: Rectangles.
- Visual Channel 1: Varying the color hue of the rectangles.
- Visual Channel 2: Varying the horizontal aligned position and vertical aligned length of the rectangles.
- We chose a sorted bar chart to identify the top 10 most reviewed business categories on Yelp, because it is a kind of graph easy to compare different categories from human perception. The varying color scale is used to distinguish among various classifications of business.



Plot 2 Double Side Bar Chart:

- Marks: Rectangles
- Visual Channel 1: Varying the color hue(light blue & orange) of the rectangles.
- Visual Channel 2: Varying the horizontal aligned position, vertical aligned orientation, and vertical aligned length of the rectangles.
- We created a two directional bar chart to see the number of open/closed business units by state. Two colors represent two different status of a given business: open or closed.

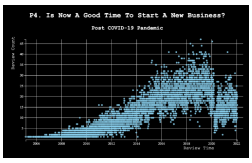


Plot 3 Circular Barplot :

- Marks: Pink circular rectangles.
- Visual Channel 1: Varying the horizontal aligned position and vertical aligned position of the rectangles.
- Visual Channel 2: Varying the circular orientation and the area of the rectangles.
- Visual Channel 3: Varying the color luminosity(from white to blue) of the rectangles.
- We used a circular bar chart to visualize the review count by city so that we can pick the most popular city on Yelp. One drawback of using a circular bar chart is that it can be challenging to

compare the area around the circle by human perception. Thus, we sorted the review count before constructing the graph and used the color luminosity to show the value difference so that one could readily compare the areas.

Plot 4 Scattorplot:

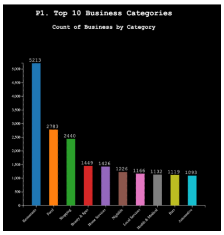


- Marks: Light blue points/dots.
- Visual Channel: Varying the horizontal aligned position and vertical aligned position of the points/dots.
- We created a scatter plot to visualize the review count over time in order to capture the trend, as well as the impact from the pandemic in 2020.

Storytelling

As investors, we would like to find out the best business opportunity in the U.S. in the next couple of years after the COVID pandemic, and the first step is to conduct market research. With 73 million active monthly users on its mobile App, Yelp.com is one of the social platforms we foray into to discover customer preferences and industry trends.

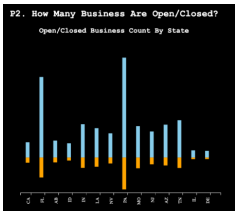
Top 10 Business Categories (Plot 1)

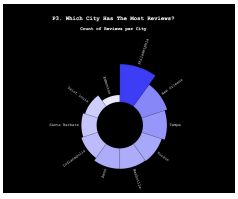


To begin with, we compared the number of businesses by each category, and found **Restaurant and any food related category such as food and bar** are the most popular business type on Yelp, which meant we might obtain the most customer behavior data in this classification. Therefore, we would like to consider proceeding with this category in our next steps.

Which State Has The Most Open/Closed Businesses? (Plot 2) Which City Has The Largest Reviews Count? (Plot 3)

Next, we need to pick a location in the U.S. to launch our restaurant, expecting rapid customer growth in the near future. As a result, we hypothesized the state with the most open businesses to be the one with the greatest potential for customer acquisition. Thus, we visualized the number of open/closed businesses by state in P2.



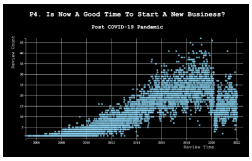


Surprisingly, the number of open businesses in the state of ***Pennsylvania***, ranked the highest, along with the number of the closed businesses, and ***Philadelphia*** (a city of Pennsylvania), ranking as the 6th largest cities in the U.S., turned out to be the city with the most reviews on Yelp, which indicates there are more active reviewers there, compared to its counterparts. We performed the sanity check, and discovered the data set itself missed data in some bigger cities, such as NYC or Los Angeles. In our future analysis, we will take this into account, but for now, we think our visualization at least provides us with some options on where to launch our business.

Interestingly, the state of **Florida** ranked the second when it comes to the number of open/closed business units, however, it was **New Orleans** (City in the state of Louisiana) that appeared to be the city with the second largest number of reviews in P3. When looking at P3, it seemed the top two cities, Philadelphia and New Orleans, are popular domestic **tourist destinations**, whereas the third most reviewed city, Tampa in Florida, is not a traditional tourist spot. This could imply that in the big tourist cities, a big chunk of Yelp users consisted of visitors who mainly relied on Yelp App to look for food, and were more likely to leave reviews after dining, compared to the local people who might use other means to find where to eat, such as word-of-mouth, friends' recommendations, and so on.

Since we utilized Yelp reviews as one of our operational benchmarks, we would like to focus on some tourist cities in the states with the most open businesses. In this case, we would like to conduct customized analysis in **Philadelphia**.

Post COVID-19 Pandemic: Is Now A Good Time To Start A New Business? (Plot 4)



Next, we would like to examine whether or not now is a good time to launch our business in the hard-hit restaurant industry after the pandemic. In P4, we analyzed the number of Yelp reviews in the past 17 years. Obviously, there was a sharp decline in the review count on Yelp in the beginning of 2020 when the pandemic started and lockdowns were imposed in some of the cities in the U.S. Fortunately, we could see a gradual recovery since then, which was a positive signal for us to enter the restaurant industry in the next 12 - 48 months.

Business Plan Conclusion

In conclusion, we could foresee a bounce-back in the restaurant industry in the near future, and we would like to conduct in-depth research on launching a new restaurant in **Philadelphia**, with the limited locations presented in the Yelp dataset. In particular, we would target travelers in Philadelphia as our main customer source, because they might be the most active user group on Yelp to generate constructive reviews for new businesses, which was crucial for us to collect customer behavioral data and adopt relevant strategies to acquire new customers in the future.

Team Contribution

Zhi Lin:

- Data Cleaning and Sampling
- Implemented data visualizations
- Wrote final report

Yun Zhou

- Implemented data visualizations
- Wrote final report

Tianchen Wang

- Implemented data visualizations
- Wrote final report

The most time consuming part in our project was to find the most efficient visualizations and implement them. We have had a couple of 1-hour brainstorming sessions to finalize the ideas, and each team member spent around 3 - 4 hours to develop the data visualizations.

References

Code from INFO 5100 lectures

The 200 largest cities in the United States by population 2022. (n.d.). Retrieved October 6, 2022, from https://worldpopulationreview.com/us-cities

Bostock, M. (2022, August 31). *D3 Gallery*. Observable. Retrieved October 6, 2022, from https://observablehq.com/@d3/gallery

Boost Medical. (2019, November 26). *11 yelp statistics for 2020 you need to know as a business owner*. Boost Medical. Retrieved October 6, 2022, from https://boostmedical.com/yelp-statistics/

Yelp, I. (2022, March 17). *Yelp dataset*. Kaggle. Retrieved October 6, 2022, from https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset?select=yelp_academic_dataset_business.json