Group T- Yelp Data Analysis

Process Book

Team Members:

Mengting Wang, Naijia Wu, Zhengze Hou

Our goal is to use Yelp dataset to create visualizations which provide multilayered data about local bar business locations in a single city. We created geographical maps, interactive maps, word cloud, word pyramid, bar charts and social networks that exposes hidden commonalities and differences across local bar businesses in a single area.

Brainstorming process:

We had initial interests in many areas, like social media usage, Dating apps, Video streaming, Vintage shops, yelp data analysis, etc.. We finalized our idea and datasets with yelp data given that data is publicly available, and contains some geographic identifiers, reviews, and potential network relationship that allows us to do geographical mapping, text analysis using NLP and network visualizations.

Below are our initial thoughts on the contents in our projects:

- Yelp data (found in kaggle) https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset
    1. Local businesses and their ratings and locations
    2. Perhaps map those locations on the map
    3. Perhaps do a heat map
    4. In some regions they might have more restaurants
    5. What kind of people would want to go to which kind of restaurants
        a. Factors that affect their choices in going to which places (proximity)
        b. Text analysis and geographic analysis

Further thoughts:

After several times of discussion, we had more detailed thoughts on our datasets and types of data analysis and visualizations in our projects:

- 5 sub datasets exist within this dataset:

1. business.json : Contains business data including location data, attributes, and categories. [ggmap,heatmap on location, restaurant rating]
2. review.json : Contains full review text data including the user_id that wrote the review and the business_id the review is written for. [wordcloud, text analysis-sentiment analysis/ topic modeling]
3. user.json : User data including the user's friend mapping and all the metadata associated with the user.[for social network analysis]
4. checkin.json : Checkins on a business.
5. tips.json : Tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions.[word cloud,text analysis]

- Exploratory data analysis:
    - Regional level:
        - We are interested in creating heatmaps of all restaurant locations; the highest review count; the highest star rating restaurants; and explore what are the most popular categories of restaurants; reviews of specific categories of restaurants in various cities; cities that have most business parities, and do a deeper dive into these findings.
    - Individual level:
        - For each individual, what are the relationships between user's elite, average stars and restaurant ratings
        - Find out the critical links or highly influential people within the community
- Visualizations:
    1. Interactive Map
        a. Possibilities include the option to filter by various metadata in user tips or comments (e.g. food poisoning, rude service, speedy order fulfillment), heat mapping by comments of a particular type
    2. Word Cloud
        a. Possibilities include most common comment types within a certain area by content and sentiment, most frequently used words by business type, most frequently used words for positively rated businesses vs negatively rated businesses
    3. Bar chart

    a. Possibilities include most commented upon businesses in the area, proportion of comments that list positive vs negative comments
4. Line chart
    a. Possibilities include the change over time of comments of a certain type (e.g. do comments about food poisoning increase or decrease over time in this area?)
5. Social Network
    a. How are the relationship between Yelp elite or most influenced users who have reviewed a bar in Vegas and the relationship between bars. Is the users' relationship iterative or clustered?
    b. Are the bars in Vegas connected based on location or other factors?
    c. Help the viewers to know the most similar bar when they select one bar, and click to know this bar and its most similar bar neighborhood and address.

Finalized dataset and created roadmap of the project:

As our common interest in bar business, we decided to explore bar business in the Las Vegas area. As we know, nightclubs are the most well-known Vegas nightlife, with their elaborately designed spaces, star DJs, celebrity hosts, bottle service and all-night dancing. Hence, we were excited to do data analysis and visualization to see which kinds of bars are more popular? Why do some bars gain great reputation and higher ratings while others don't? Among the elite users who have reviewed any of bars in Vegas, what are their relationships? Is there some inside network between these celebrities in Yelp?

We spotted various star rating bars and provided interactive maps of the city and neighborhood of Las Vegas. For specific business, we did a word cloud of reviews of the bars, top 10 common words of the bars, sentiment analysis using “affin”, display of top positive reviews and bottom negative reviews of specific bars. For social network analysis, we extracted the top 10% elite users of Yelp, and visualized the networks of the users and bar business. Moreover, we also run a regression model to explore the relationship between bar ratings and review counts.

Visualizations:

- Geographical maps
- Text Analysis with word cloud, pyramid plots and bar charts
- Social Network Analysis with social network maps, regression analysis and bar charts

Roadmap:

City: Las Vegas

Business: # of review > 100

Business type: Bars (that are still opening)

Contents	Feature used	Data scale	Visualization tools
Geographical maps to explore locations of top bars	business-categories, business-name -review_count business-stars latitude longitude business-city	Region selection: {latitude, longitude,city}; Business type : {categories = 'bar'}; Business scale: {business-review_count >100}	interactive maps
Text analysis on top and tail reviewed bars	Reviews Bar	Business scale: {yelp_business-review_ count >100}	NLP text analysis Analyze business name keywords, reviews
The factors which affect elite reviews and review of bars	User_id Reviews Avgstars Fans business_id	users' fans >= 100, useful value of their reviews >= 900	regression: Y = avg stars of elite users' or bars' reviews X = Number of fans, Review_counts,
The social network between elite users and between bars and the information about most similar bars	User_id Reviews Business_id Address Neighborhood	users' fans >= 100, useful value of their reviews >= 900 Top 50 bars based on the amount of reviews	Data mining based on tidyverse and nested for loop Networkd3