Yelp Business Exploration

Danlu Liu
Computer Science
University of Missouri
Columbia, Missouri
dltb9@mail.missouri.edu

Dan Huang
Computer Science
University of Missouri
Columbia, Missouri
dh4tx@mail.missouri.edu

Yu Hou
Computer Engineering
University of Missouri
Columbia, Missouri
yhtwc@mail.missouri.edu

Abstract— **In this paper, we use Yelp data to analyze the number and popularity of different business in the United States. Some details of the restaurants are visualized to give more insight. Uses can choose the restaurants based on their own preferences. The visualization and website models can help user to filter out the results suit for them. In the paper, we use the JavaScript and D3.js library to create an interactive view. Also, we use the Tableau to build the detailed view for decision making.**

keywords—**Yelp, data visualization, D3.js, Tableau**

I. Introduction

Gourmet food fans should be very familiar with Yelp. This food commentary website is recognized by the world as "food strategy provider". It is one of the best examples of using public experience and comment, meanwhile, many evaluation data and ranking mechanisms also provide convenience for people when they choose restaurants.

This paper uses Yelp dataset [1] to grasp useful information and provide a concise and clear vision for decision makers. Combined with data visualization tool and SQL data manipulation language, we identify the impact attributes and features related to restaurants. Besides, we also extract valuable information from user evaluation.

Two visualization tools are applied to show the overall and detailed information in the dataset, compare with using words to describe something, an image seems more vivid obviously.

The remaining part of this paper is organized as follows: First, we introduce the dataset information; Second, we talk about the related work; Third, we describe the visualization design and our approach; Fourth, we illustrate our visualization work and give an evaluation of it; Finally, we discuss the results by providing and conclude our work.

II. Dataset

The data source comes from Yelp 2017 Dataset Challenge, which provides 6 tables, business, review, user, check-in, tip and photo table. In this paper, we describe the data more about the company level, thus we use business data as our dataset.

Business data lists the name, location, business hours, cuisine categories, average star rating, evaluation quantity and other related factors, such as noise level, reservation policy. In addition, it contains 156639 business records. Most of these records are in the United States but the dataset only contains several states' information and we use these records in our analysis. Bar chart and line chart can display the total number and tendency of the dataset. Those charts are useful tools in the data visualization area and we also apply these tools into our work.

III. Related Work

The restaurant rankings in Yelp are recognized globally, and it is one of the best examples of the use of public experience and comments.

There are several papers uses Yelp dataset to retrieve relationship information [2] or build a recommendation system [3]. However, there is not much work shows the dataset in an obvious way. There are lots of charts or visualization methods exist to help people understand the dataset in an easy sand direct way. Pie chart is a circular statistic graph to illustrate the numerical portion. TreeMap can display the hidden hieratical relationship of the dataset [4].

IV. Visualization Design & Approach

4.1 Different tools

In the data visualization area, there are huge number of tools. Some tools are flexible and can be used in lots of scenarios. However, it is difficult for beginners to use. Some tools are easy to use, but it only provides several visualization methods and it lacks the flexibility. D3.js is broadly used in different areas. It is flexible to combine with the web applications. Also, it provides tons of examples and suitable for many cases. However, the D3.js is based on JavaScript and it is not easy for BI analyst or other non-technicians to program. Tableau is a data visualization tool that can help anyone to understand the data. In our paper, we use tableau to do the more complex data visualization and decision making work. We provide an interactive way to help user to explore the data.

4.2 Different aspects of exploration

In our exploration, we want to display multiple aspects of the dataset. Firstly, we give the business distribution in different States and then analyze its popularity based on the stars of that company. Secondly, we analyze the delivery situation and top K restaurants' style based on different criterion. Thirdly, we create a comprehensive chart with filters to help user determine which restaurant is more prefer.

V. Implementation & Evaluation

5.1 Business Distribution and Popularity

In Yelp data, it provides the category description of the business. However, one company can have lots of tags to describe itself. One example of the categories is in Figure 1.

"categories": ["Food", "Soul Food", "Convenience Stores", "Restaurants"],

Figure 1. Example of a record's categories.

We want to dig the information of the restaurants and the shop. Thus, we extract the records contain the "Restaurants" or "Shopping". For these two groups (Restaurants and Shopping), we extract all the category keywords respectively and then calculate the frequency of these words. Some word frequency examples are shown in Figure 2.

word	frequency		word	frequency
Restaurants	51614		Shopping	25131
Food	19766		Services	8345
Bars	12680		Home	8336
American	10008		Stores	7513
Nightlife	6969		Fashion	6420
Sandwiches	5864		Garden	4584
Fast	5792		Clothing	4196
Traditional	5737		Beauty	3579
Pizza	5652		Shops	2899
New	5181		Local	2729
Italian	4411		Food	2634
Burgers	4236		Repair	2415
Breakfast	4122		Women	2355
Brunch	4103		Arts	2334
Mexican	3935		Spas	2074
Chinese	3775		Furniture	1876
Cafes	2893		Goods	1858
Tea	2631		Accessories	1739
Services	2564		Flowers	1704
Coffee	2404		Gifts	1699
Event	2327		Supply	1666
Chicken	2224		Sporting	1662
Japanese	2187		Cosmetics	1651

Figure 2. Word Frequency. Left side is for Restaurant and Right side is for Shopping.

We use these keywords and its frequency to show the word cloud [5]. Users can find the subcategory information of Restaurants and Shopping. To provide a direct perception. We create a word cloud based on these keywords using D3.js. Figure 3 shows an example of Restaurant's word cloud. It also gives us the insight of how to split Restaurants and Shopping into interesting subgroups. The keywords in Red in Figure 2 are the insight for our subgroups.

Figure 3. Word Cloud of Restaurant.

After split the data into subgroups, we use United States map to show the number of restaurants, the deeper color means more restaurants/shops on that state [6]. The bar chart shows the distributions of different restaurants in that state. The pie chart shows the popularity of different restaurants' style. The user can click different states the see the details of that state. Also, they can choose bar chart to see the popularity of a restaurant's style which they are interested in. All those charts are connected and interactive and they are shown in the Figure 4.
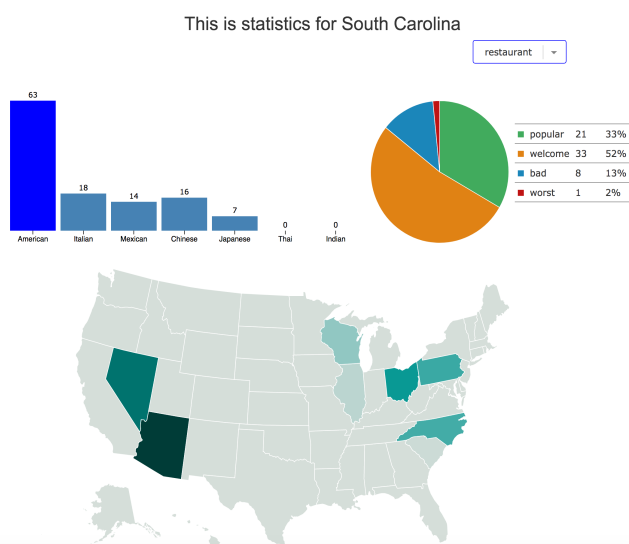
Figure 4. United States restaurant distribution and popularity of different styles.

Given the overview of restaurants, we explore some certain attributes by using D3.js. Based on the exploration of data, user can tell the distribution clearly, as well as the connection between them. We use four attributes as an example. As shown in Figure 5, we connect every two attributes as a group. On the left hand of each chart represents the global distribution, if we hover the mouse pointer over the bar, the corresponding individual distributions will be highlighted, which is illustrated in Figure 6. In this way, users can understand both global trend and the search for a specific target with a certain condition.

Figure 5. The visualization of attributes for exploring the global trends and connection between different attributes

The sunburst chart is used to visualize hierarchical data, depicted by concentric circles. The circle in the center represents the top 10 number of record, with the hierarchy moving outward from the center. A segment of the inner circle bears a hierarchical relationship to those

segments of the outer circle which lie within the angular sweep of the parent segment. See Figure 7.
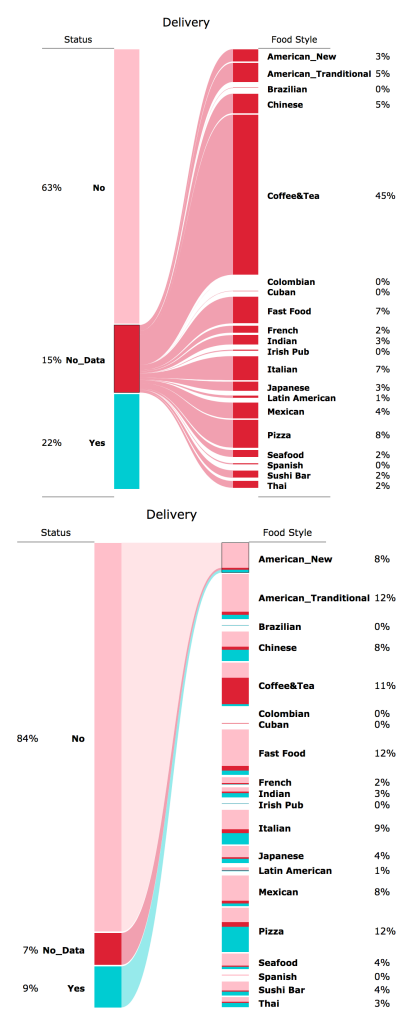

Figure 6. The interactively exploration between global distribution and individual distributions
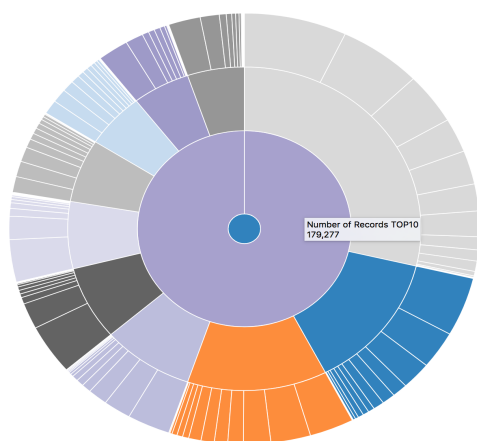

Figure 7. Sunburst chart

Tableau is a good way to create some beautiful interactive charts. Like Figure 8, we use the shade of the color in a tree map to indicate the star of each category, as well show it in a bar chart. Also, we use a pie chart with different color to show the total review number for each category. When move the mouse on a part in the Tableau, it will come up the relationship between those charts and the message we would like to bring to our readers will appear.
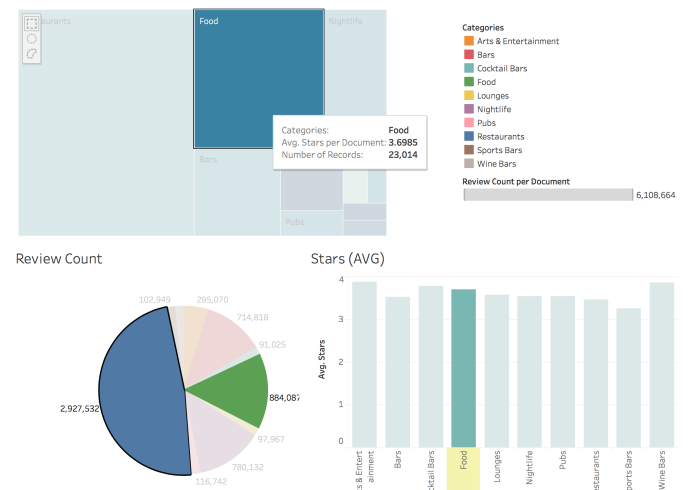

Figure 8. Tableau example to show related information

We try to use a variety of visualization method to implement our data. In Figure 9, we use a bar chart, a line chart and a scatter plot. There's two features we want to analysis, the Noise Level and Is Open. This part give people more detailed information for each single restaurant. Different color shows 4 different noise levels, and the depth of the blue in the bottom chart shows the number of open restaurants. Deeper, more opens.

Besides, it is also necessary if we can put several related information in one picture, that will convenient for readers to retrieve information. As shown in Figure 10, we put the name, and some service customers may concern a lot when they choose a restaurant like Wi-Fi, smoking, outdoor seats on the left side of the picture, and there can be more if you like, not limited in these four. Also for each restaurant, there's the price range and score on the right side, the score is still show by the depth of blue.
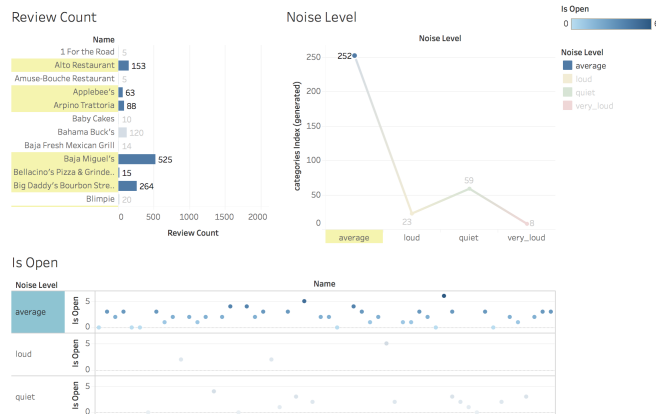
Figure 9. Variety of visualization method

Figure 10. Six details in one picture

VI. Results & Conclusions

In view of the diversity of data types, a variety of visual analysis methods are implemented. Our exploration not limited in restaurant, but extend to other fields. Through the analysis of these images, we found that it is more vivid to explain or show some information by images than words. We analyzed several features, we combined some of them, found out the connection between them.

Data analysis plus visualization, it can be imagined that one day, every consumer's behavior at the store will be recorded, analyzed, and real-time recommendation with their preferences can be promoted by correlation.

References

[1]. Yelp dataset challenge, 2017.
https://www.yelp.com/dataset/challenge

[2]. Yu, X., Ren, X., Sun, Y., Sturt, B., Khandelwal, U., Gu, Q., Norick, B. & Han, J. Recommendation in heterogeneous information networks with implicit user feedback. In Proceedings of the 7th ACM conference on Recommender systems, pages 347-350. ACM, 2013.

[3]. Feng, He, and Xueming Qian. "Recommendation via user's personality and social contextual." Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.

[4]. Cleveland, William S. (1985). The Elements of Graphing Data. Pacific Grove, CA: Wadsworth & Advanced Book Program. ISBN 0-534-03730-5.

[5]. Jason Davies. (2015, Sep 24). d3-cloud.
https://github.com/jasondavies/d3-cloud

[6]. D3 Data-Driven Documents. https://d3js.org