Analysis of the Yelp Academic Dataset

Factors that Affect the Ratings of Restaurants, Bars, Hotels, and Shopping Centers

Chelsey Legacy
Graduate Assistant
calegacy@iastate.edu

Lindong Zhou
Graduate Assistant
lindongz@iastate.edu

Evan "Pete" Walsh
Graduate Assistant
epwalsh@iastate.edu

Iowa State University
Department of Statistics
December 4, 2014

Abstract

For this project we examined the Yelp Academic Data set with information pertaining to 250 of the closest businesses to 30 large universities. Our analysis explores trends and characteristics in four different types of businesses, namely: restaurants, bars, hotels, and shopping centers. In particular, we investigate how certain factors affect the average user ratings of each establishment from Yelp reviews. Some relationships are intuitive, others are not. For restaurants we find that people will pay a premium to eat at a place that accepts reservations, yet they are not necessarily paying for better quality food or service. There are also interesting links between noise level, whether a place delivers or offers takeout, and price level with number of stars a restaurant receives. For bars, we find that there are some relationships between the attributes of a bar that will determine what type of group will venture to that bar. We also find that most bars serve American food, though other foods are well represented too. For hotels and shopping centers, we discovered that higher prices are not correlated with higher ratings, and that people's reviews are not much affected by whether the place has a parking lot nearby or free wifi.

Contents

1 Introduction

The Yelp Academic Dataset[1] provides data enthusiasts with the exciting opportunity to explore an incredible collection of information regarding the characteristics and quality of hundreds of businesses across the United States, Canada, and the UK. This paper provides some thorough insight into the data using a variety of visualization techniques.

In particular, we investigate various factors that affect the rating of restaurants, bars, hotels, and shopping centers. For restaurants, we seek to answer how whether or not an establishment accepts reservations influences the price and the overall quality of the dining experience. We also investigate how noise level effects the rating of restaurants and how this differs between places that offer live music and those that do not. We then use a pair of wordclouds to question the types of things people are most likely to talk about when they write a "tip"[2] about a really bad or really fantastic restaurant. Lastly, we investigate the types of food that are most likely to be offered as take out or delivery, and possible explanations for this.

For bars, we inspect which types of food are most commonly served and whether or not lower ratings indicate if a bar will be shut down. We also investigate the types of activities that are likely to take place at a given bar.

For hotels and shopping, we explore the question of how price level and popularity (as measured by the count of Yelp reviews) affects rating.

This paper will procede as follows: Section 2 discusses the raw data and the steps that were taken to "clean" the data into a more useful form. Section 3 examines the subset of data associated with restaurants, while sections 4, 5, and 6 examine the subsets associated with bars, hotels, and shopping, respectively. We conclude in section 7 by summarizing what we have found and offering several ideas for future studies.

2 Data

The Yelp Academic Dataset includes details and reviews on 250 of the closest businesses to 30 large universities, including the Arizona State, UNLV, the University of Edinburgh, the University of Wisconsin, and the University of Waterloo, to name a few. The raw data is in json format and contains five different types of json objects: **Business**, **Review**, **User**, **Check-in**, and **Tip**.

Each **Review** object represents an individual user-based review of a particular business. The unique encrypted business ID is given along with the date of the review, the number of stars (out of 5) that were awarded, the number and type of votes that the review received, and an optional text description provided by the user. **User** objects are unique to every person that has an active Yelp account. Each user has a name, a unique encrypted user ID, the number of votes they have cast, the average number of stars they have given, and the date they signed

---

[1] $https://www.yelp.com/academic\_dataset$

[2] In Yelp, a "tip" is a short comment or description of a business given by a Yelp user. A tip is too short to be considered a review.

up for Yelp, among other things. A **Check-in** object represents the count and time of all the registered check-ins for a particular business, and **Tip** objects represent a tip given by user for a particular business. Tips include the user's ID, the business's ID, the date, and the message that the user gave. While these objects all provide a rich source of information, for the scope of this paper we will only be examing the **Business** objects.

Business objects are unique to business ID's, and include the following information:

- the name of the business,

- the name of neighborhood in which the business is located,

- the city in which the business is located,

- the full address of the business,

- the exact latidude and longitude coordinates of the business,

- the average number of stars awarded to the business,

- the number of reviews received by the business,

- whether or not the business is still open,

- the hours that the business is open,

- the categories that the business falls under,

- and a number of different attributes which mostly concern restaurants and bars, such as whether or not smoking is allowed and the price range of the food.

To work with the data, we converted the set of all **Business** objects to a csv file in which the columns are variable names representing each aspect of a **Business** object, and each row corresponds to a unique business. Each different type of attribute was converted to its own character, numerical, or logical variable depending on what was appropriate. For example, the attribute **smoking** was converted to a logical variable where the value is "true" if the smoking is allowed at the establishment, and "false" if it is not allowed, while the attribute **price range** was converted to a numerical variable which ranges from 1 to 4.
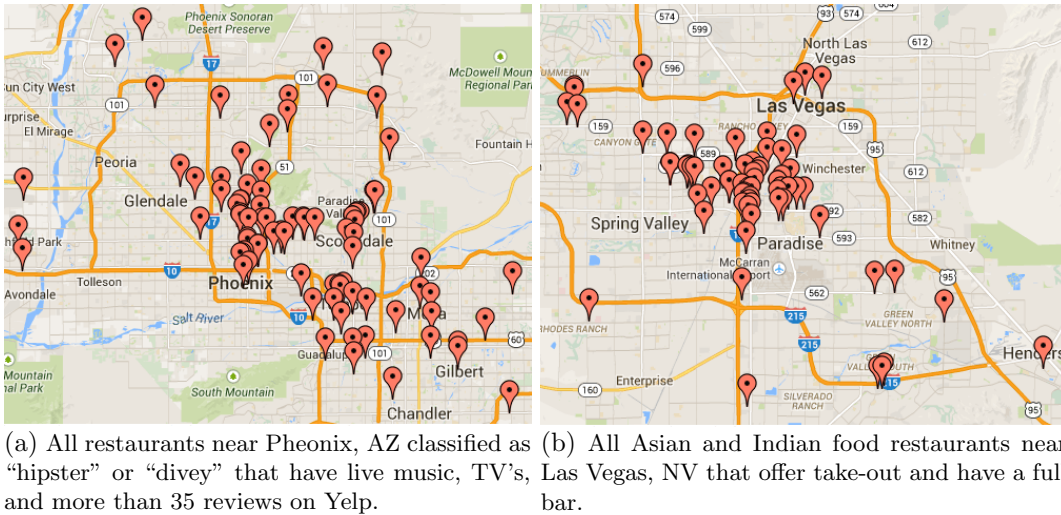
3   Restaurants

Most of the data about restaurants in the Yelp Academic Dataset are concentrated around the following cities:

- Madison, WI,

- Las Vegas, NV,

- Pheonix, AZ,

- Waterloo, Ontario, Canada,

- and Edinburgh, Scotland.

Although the choice of cities is limited, the wealth of information coming from each city is abundant. All together, our data consists of 14,303 restaurants with 45 variables describing their attributes and the collective user sentiment towards each establishment. There is a significant number of restaurants in just about any subcategory that one can imagine. For example, Figure 1a shows all of the restaurants classified as either "hipster" or "divey" near Pheonix, AZ that have televisions, offer live music, and have more than 35 reviews on Yelp. Figure 1b shows all of the Asian and Indian food restaurants near Las Vegas, NV that offer take-out and have a full bar.

Figure 1: Restaurant subsets near Pheonix, AZ and Las Vegas, NV.

(a) All restaurants near Pheonix, AZ classified as "hipster" or "divey" that have live music, TV's, and more than 35 reviews on Yelp.

(b) All Asian and Indian food restaurants near Las Vegas, NV that offer take-out and have a full bar.

3.1   The Cost of Reservations

One question we have explored with the data is whether restaurants that take reservations are more expensive and of higher quality than those that do not. Not suprisingly, it appears that one has to pay a premium, on average, to eat out at a restaurant that accepts reservations. Figure 2 shows the average price level of restaurants by type that do or do not accept reservations. For each category, the price level is higher on average for restaurants that accept reservations.

There are a couple potential explanations for this. For one, it could just be that fancier restaurants feel they need to accept reservations to attract their target demographic. Perhaps people who are willing to spend a large amount on a meal will only do so if they know they will be seated right when they arrive. Or maybe restaurants that take reservations lose some money in transition time when open tables are kept unoccupied until the party that made the reservation arrives. The higher average price of restaurants that take reservations could just be making up for this inefficiency.

Although a premium is payed to eat at restaurants that accept reservations, we find that the quality of the experience does not necessarily increase, as measured by the average stars

Figure 2: Price scale of restaurants by type and whether they accept reservations.
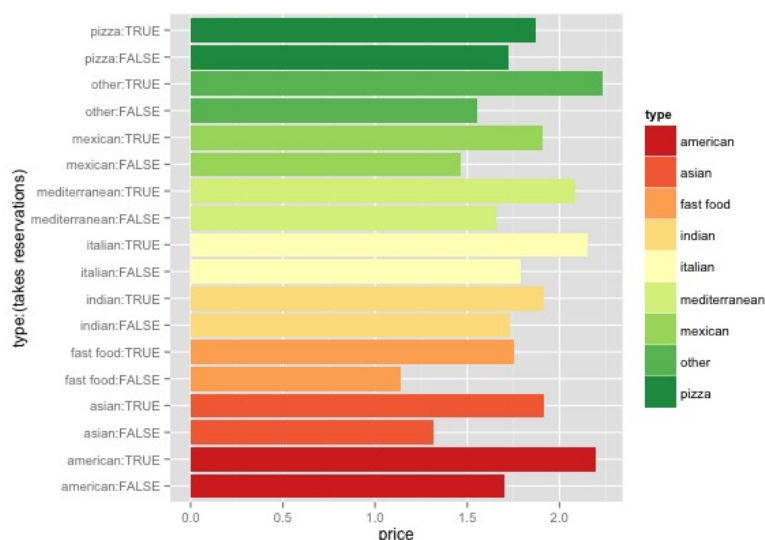
awarded by Yelp users. Figure 3 compares the average stars received by each category by whether or not reservations are accepted. While the average number of stars is slightly higher for pizza, fast food, Asian, American, and "other" style restaurants that take reservations, the average star count is actually lower for Mexican, Mediterranean, Italian, and Indian style restaurants.

The reason for this discrepancy is unclear. Maybe popular restaurants in certain categories just don't accept reservation because it would be too inefficient, and customers are fine with that. Price may also have something to do with it. Figure 4 shows the rating of restaurants by type, whether they accept reservations, and price level. For Mexican restaurants, the average rating of low priced establishments is indistinguishable between those that accept reservations and those that do not. Yet for those in the third price level, the rating is clearly higher for those that do not take reservations. The same is true for Asian style restaurants while the opposite holds for Italian and American style restaurants. This could be because people who generally like to eat at pricier Italian and American food restaurants prefer to be able to make reservations, so these restaurants that allow reservations end up with higher ratings.

3.2   Loud can be Good, but not too Loud

Another question we sought to answer was how noise pollution affects the quality of a dining experience, especially when we consider whether or not the establishment offers live music. One can imagine that people deliberately going out to eat at a restaurant with a band playing might bear a higher noise level, and thus not be as quick to penalize the restaurant for being too noisy in their review. We find that the data is consistent with this hypothesis.

For restaurants that offer live music, as shown in Figure 5, loud places have a higher average

Figure 3: Rating of restaurants by type and whether they accept reservations.
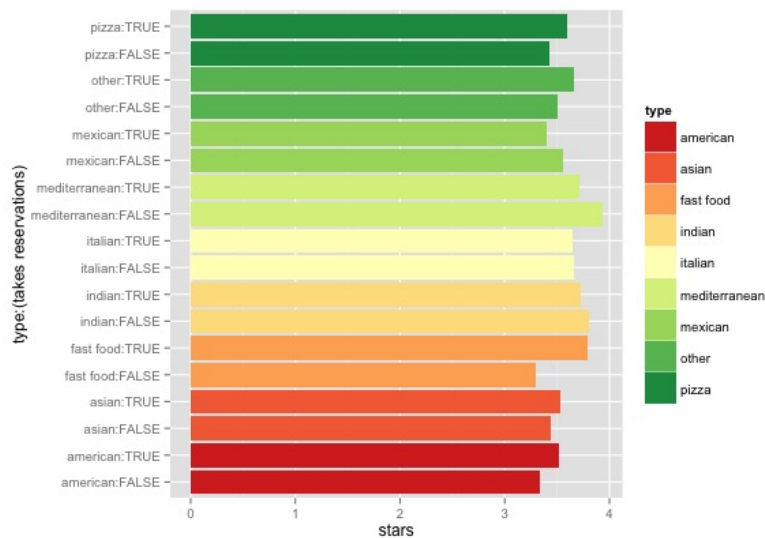
rating than places with an average noise level. Yet very loud places have the lowest average rating. For restaurants that do not offer live music, the average rating decreases with the noise level.

3.3   It's All About Service

It will come as no surprise that service is a vital part of the service industry. Friendly, competent staff can make or break the dining experience. To illustrate the validity of this statement, we examined the **Tips** objects to see which words came up the most when Yelp users were describing either a really bad or really good restaurant experience. We first matched every single restaurant tip with the corresponding restaurant, and then sorted the tips by the average number of stars that each restaurant has. We then used a simple visualization technique on each set, called a wordcloud.

A wordcloud is formed by displaying the most frequently occuring words from a set together in a random, or partly random, arrangement. The size of each word is weighted by how many times it appears in the set. A coloring scheme is often added to complement the size weighting.

We first took a random sample of 600 tips corresponding to restaurants with at least 4.5 stars, and 600 corresponding to restaurants with less than 2 stars. We then created a wordcloud for each set of tips after removing certain words like "great," "awesome," and "love" from the high rating group and words like "bad," "terrible," and "don't" from the low rating group, because those words are obvious and appeared so frequently that they "clouded" out more interesting words. There were also other words like "try," "place," and "get" that were removed from both groups for the same reasons. The resulting wordclouds are shown in Figures 6a, for the high rating group, and 6b for the low rating group.

Figure 4: Rating of restaurants by type, whether they accept reservations, and price level.
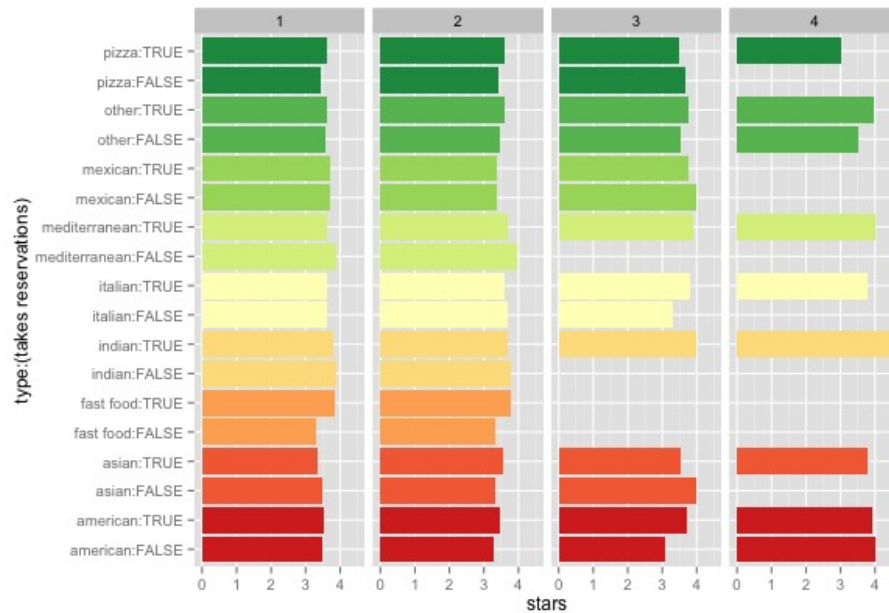

Figure 5: Rating of restaurants by their noise level and whether they host live music. The plot on the left shows the rating of restaurants by noise level that do not have live music, and the plot on the right shows the rating for restaurants with live music.
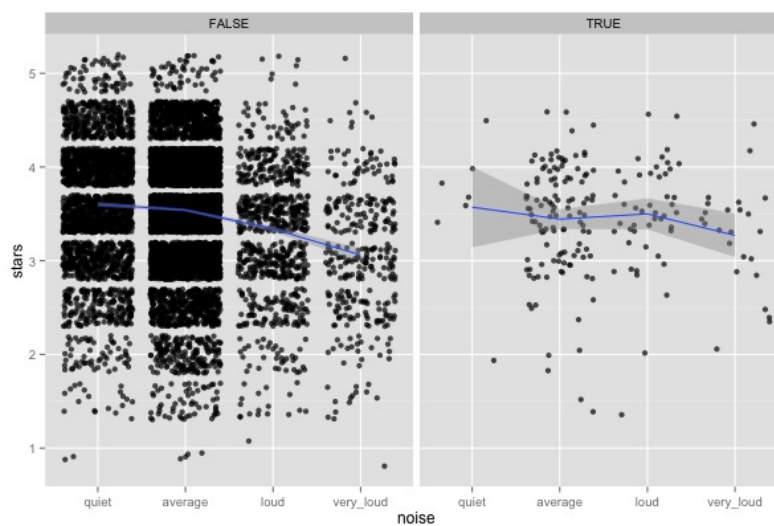
Figure 6: Wordclouds of tips for great restaurants and terrible restaurants.

(a) Tips for restaurants with atleast 4.5 stars.

(b) Tips for restaurants with less than 2 stars.

The interesting thing to notice here is that both wordclouds look extremely similar. Words like "chicken," "pizza," "time," and "order" are found in each set. Further, in both cases the word "service" was the most frequent word. Thus, it seems that people with either really good or really bad experiences tend to comment on the same types of things, particularly when it comes to service.

3.4   Delivery vs. Takeout

Figure 7: Plot of the counts of each type of food filled by if they have delivery or not.
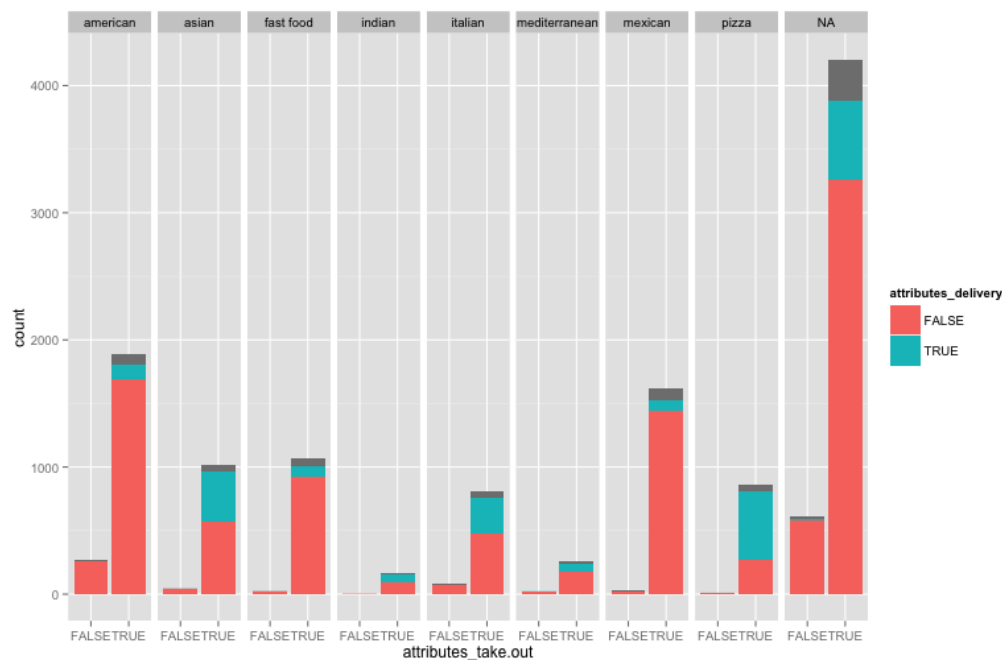
Figure 8: Plot of if a restaurant does takeout orders colored by if they do deliveries and faceted by the type of food the restaurant serves.

Now we know have gained some insight into restaurants' price and reservation style, but what about when you do not want to go eat at the restaurant? Some days there is little time for cooking, so people commonly call out for delivery or take out. But what types of food are most likely to have delivery or takeout? In other words, if we are ordering out, what is it likely we will be able to have delivered to us and what can we pick up on our way home? In Figure 7 we can see that the food types that most likely have delivery are those that serve pizza, Asian, or possibly Italian. Most likely you will not be able to find an American or fast food place that delivers. This could be due to the fact that some of these food do not stay good if they sit too long. Food like pizza and rice are fine to sit for a while, are easy to keep warm and fresh tasting, and can also be easily reheated. Thus, these are qualities that make them good options for delivery. However, American food, or fast food like burgers and fries are not good if they sit too long, and do not reheat to the quality they were when they left the restaurant. Thus, there are fewer, if any of these places that offer delivery.

We can see in more detail in Figure 8 that restaurants of all types are much more likely to offer takeout than delivery to customers. We can also see that the restaurant is more likely to offer delivery if they offer a takeout option. Restaurants of all types may be more inclined to offer delivery than takeout for a number of reasons. They could only have a limited number of people working and do not want to spare any or pay extra people to deliver the food. Delivery also means transportation costs for the business, and as discussed above it means more time for the food to decrease in quality while a driver makes several deliveries. However, takeout is a great option for restaurants, because they get to sell more food, without have to seat as many people in the restaurant. Takeout also reduces the responsibility of the quality

of the food once it arrives at its destination. If it took the customer a long time to get home and the food is not as good as it once was, this is not the restaurant's fault.

4   Bars

The Business dataset contains a lot of information that can be used to analyze the relationship between different aspects of bar culture. In order to analyze the bars available in the dataset, the data was subset to get rid of any businesses that did not have a full bar. Looking at only the businesses with a full bar gives us insight into locations that serve from a full bar such as clubs, hotels, select restaurants, lounges, bowling alleys, and other various entertainment locations. Through analyzing this data we can explore what types of food are servied at bars (if any), if lower ratings indicate a bar could shut down, and the type of activities that might take place at a given bar.

4.1   Food

Figure 9: Plot of the number of stars a restaurant received filled with the type of food served at the bar.
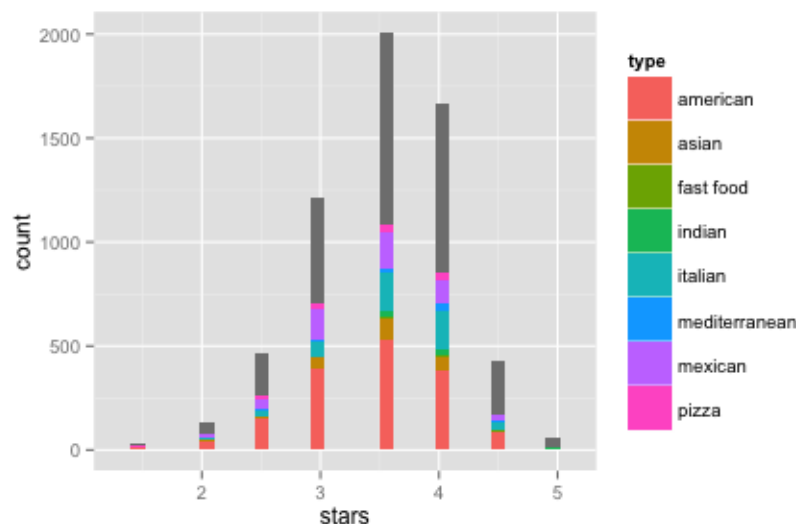

Though not all bars serve food, many bars in restaurants do which begs the question; what is the most common food served in bars? In order to investigate this question a qplot of the number of starts given to each bar filled with a count of the food types for each different rating will give us the information and more. This data revealed, unsurprisingly, that the most common food served in American bars is American food (burgers, fries, etc). Through further analysis it is determined that 50% of the bars that serve food served primarily American style food. We can see that the majority of each bar of star, no matter what the rating, is colored for American food. However, looking more closely at the graph we can see that there is

some unexpected information. From Figure 9 we can see that there are a wide number of food options for people looking for full bars. Along with American, Italian and Mexican are also abundant options making up 16% and 16.7% respectively. Another piece of information we can infer from this graphic is that along with the plot of the stars a bar receives being a normal distribution, there also is a consistency in the number of each type of food being represented at each level of stars a bar receives. Thus, we can see there is no preference for food when looking for a certain quality bar. A bar is equally likely to serve American food whether it receives 1 star or 5 stars. This could be further investigated through hypothesis testing for difference in proportions.

4.2 Stars

Figure 10: Plot of the number of stars a bar received colored by whether or not they are open or closed
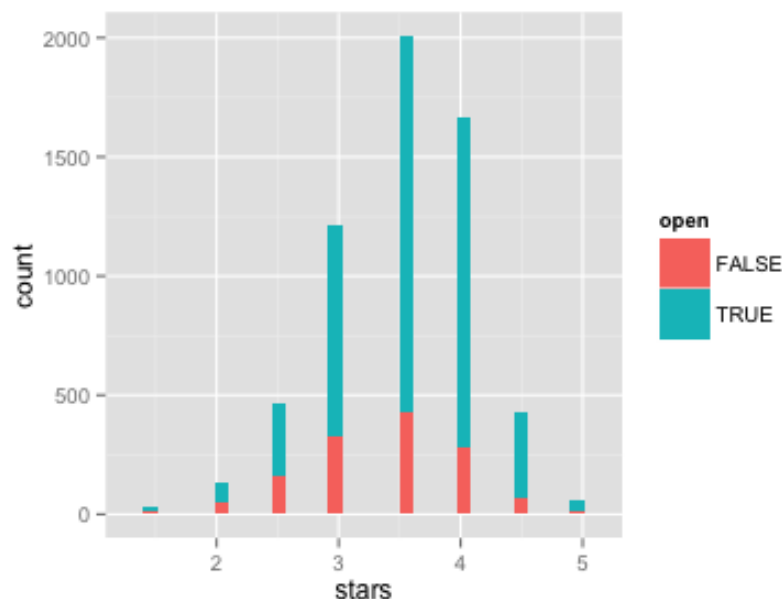
The purpose of yelp is to provide review of businesses for others in the community to take into consideration. Since this dataset provides the number stars a bar receives and the status of the bar (open or closed) as of when the data was collected. I decided to investigate the idea that possibly if the bar has lower ratings it has more of a chance of being closed. It is hard to discern any sort of relationship from the plot, except that there are more restaurants open than closed. However, if inspected closely looking at the extreme endpoints of the graph you can see it appears the percent of open restaurants at the 5 star level is higher than the percent of 1 star restaurants that are open. In order to look into this further I created several subsets of the data. I created a subset of data for 2 stars(1 star had 0 observations), 3.5 stars (which is the median of the data), and 5 stars. I then found the percent of open bars for each level. I found that the 2 stars had 64.3% open, the 3.5 stars had 78.4% and 5 stars had
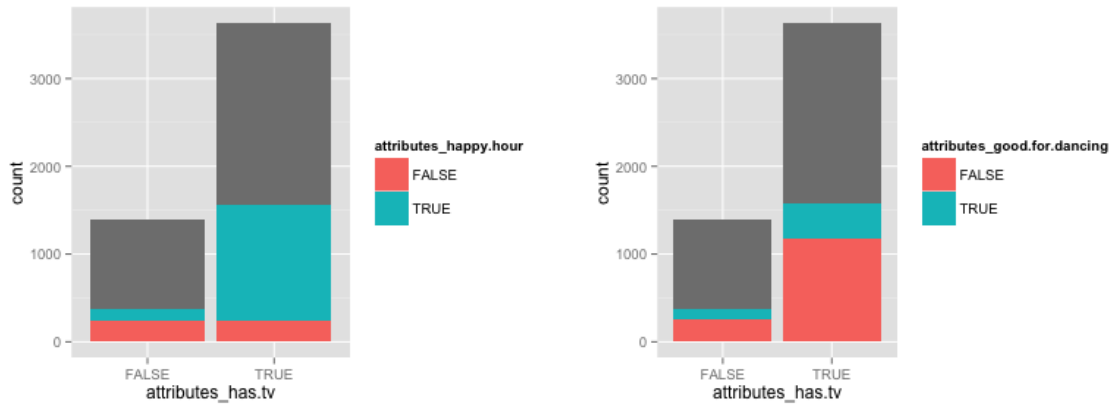
(a) Plot indicating whether or not the bar as a tv, colored by if it has happy hour specials

(b) Plot indicating whether or not the bar as a tv, colored by if it is good for dancing

Figure 11

82.1% open. As we can see the percentages of bars open increases as the number of stars the bar is rated increases. Though we can make no assumptions of causation, this is still an interesting trend tp consider for further more complex analysis.

4.3   Popular Features

Each bar in a community has its own unique features, special pricing promotions, or atmosphere. Some bars are good for dancing, watching televisions broadcasting sports games, happy hour, or some are more quiet and laid back for just talking and relaxing with friends. What are the relationships between these variables, if any? I wanted to find characteristics for each of the different atmospheres a bar may have: romantic, intimate, touristy, hipster, divey, classy, trendy, upscale, and casual. However, it quickly became clear that there are not many bars in each of these categories. Most bars are just classified as casual. Thus, there wasn't much to investigate.

Next, I thought it would be interesting to look at features of bars with televisions. Since, not all bars have televisions there must be certain characteristics for those that choose to have a television and those that do not. In Figure 11a we can see that most bars that have a tv also have happy hour specials. These are probably more casual places that people go after work to relax with coworkers and friends. We can gain even more insight into the bars with tvs when we look at Figure 11b. In this plot it is interestingly apparent that if a bar has a tv it is more likely that it is not good for dancing. This again reinforces that the bars with televisions are more of a relaxed atmosphere to hang out and chat or watch sports. Most likely a place like this would be an in inappropriate atmosphere for a group looking to go dancing. Places people go dancing are more apt to be louder places people go later at night, when happy hour specials would be over and there would be no need for televisions while dancing.

5   Hotels

We created a subset of the Yelp data to get only businesses whose categories contain Hotels. Focusing on the variables that have relatively few missing values, we included price range, wifi, location and rating in our analysis. Most of hotels in the dataset are located in the US states of Arizona, Nevada and Wisconsin as well as Edinburgh in Scotland. The dataset is helpful for us to find out whether popularity, price level, wifi and location influence the rating of a hotel.

Popularity does not necessarily mean better products and services. This statement is supported by Figure which shows the distribution of ratings by popularity. We determined the popularity of a hotel based on the total number of user reviews. Hotels that have fewer than 100 reviews are considered as least popular, whereas hotels with more than 100 and 1000 reviews have medium and high popularity respectively. Across levels of popularity, there is no significant difference in the distribution of ratings. 4-star reviews are posted on the website most frequently, followed by 5-star and 3-star reviews. Thus, switching to a better promoted hotel does not seem to affect peoples lodging experience.

Wifi does not seem to matter as well. Figure 12a shows the distribution of ratings by price range and availability of wifi. Within each price range, the distribution pattern appears roughly the same whether Wifi is free, paid or not provided. However, the graph suggests a correlation between price range and customer satisfaction. High-priced hotels receive better reviews compared to budget hotels. One possible explanation is that overnight stay in a hotel is mostly attractive to high-end customers who value quality of amenities and customer service more than just price. This observation supports the commonly held belief among travelers that you get what you pay for.

Paying more for a better service might not be true outside the United States. The notion of you get what you pay for is clearly refuted by Figure 12b which shows the distribution of ratings by price range and region. While high-priced hotels are still rated higher than budget hotels in Arizona, Nevada and Wisconsin, the ratings of hotels in Edinburgh seem to be consistent across price levels. Perhaps European customers receive equal treatment regardless of how much they pay, whereas American customers have to pay more to get better treatment. Nevertheless, we cannot justify the claim without examining more data about American and European hotels.

6   Shopping

To examine all types of shops in the dataset, we subset the data to get only businesses whose categories contain Shopping. Considering missing values, we included variables price range, parking lot, location and rating in our analysis. Like hotels, most of the shops in the dataset are located in Arizona, Nevada, Wisconsin and Edinburgh. We use the data to explore the effects of popularity, price range, parking lot and location on the rating of a retail store.

Figure again supports the statement that popularity is not associated with better products and services. We determined the popularity of a retail store using the same criteria (see

(a) Histogram of the relative frequency of ratings of hotels by price range
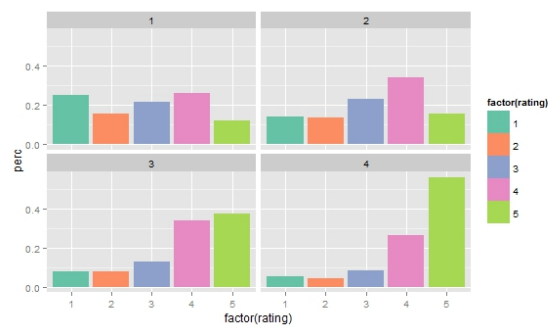
(b) Histogram of the relative frequency of ratings of hotels by price range and region
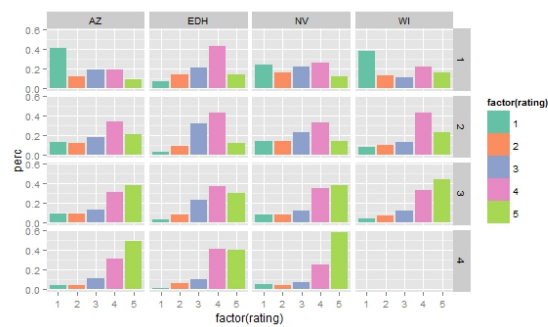

Figure 12

Section 5). The graph shows that the pattern does not change much as popularity goes up. So, customers do not seem to improve their shopping experience by joining a crowd.
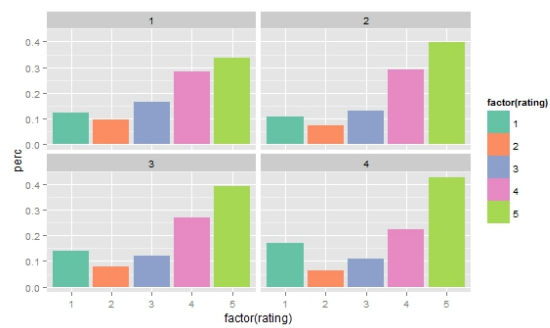
Figure 13a shows the distribution of the ratings by price range and availability of parking. Whether or not there are parking lots nearby, the stores are rated roughly the same. It is interesting to see, however, that pricey stores receive more polarized reviews than low-cost stores. In the highest price range, we find more 5-star reviews and more 1-star reviews at the same time. Customers seem to either love or hate the stores that charge them very high prices. One might argue that retail customers are more diverse than hotel customers. In other words, average people tend to focus more on price than quality. Without looking at the demographics of customers, we never know if this argument is true.

Figure 13b shows another different trend in the distribution of ratings of retail stores in Edinburgh. While pricey stores still receive more polarized reviews than low-cost stores in three states, especially Arizona, the ratings of stores in Edinburgh are again consistent across price levels. Assuming that the demographics are the same everywhere, one could possibly say that quality is better reflected by price in European stores. This implies that the quality of an American product sometimes does not justify its price. Again, we cannot verify the statement without any more information.

7 Conclusion

The Yelp Academic Dataset has provided us information on a variety of businesses including restaurants, bars, hotels, and shops. We have used the abundance of data to find key relationships between the available variables and their respective businesses. For example, we have found that reservations, price, noise level, takeout, delivery, and foremost service are key variables that provide interesting relationships for restaurant data. We have also found that features such as food type, if a restaurant has a TV, or if it is good for dancing all provide insight into what type of atmosphere we will find at a bar. When evaluating hotels and shopping we have found that popularity does not always reflect the quality of the service we will receive. This data set has provided us with many insights into possible relationships between not only these variables, but also human culture and tendencies. Though it is impossible to make solid conclusions about any of the relationships we have found based on our limited dataset, we have still come up with some interesting relationships and found even more questions about this topic that could be used for further research on the subject.

(a) Histogram of the relative frequency of ratings of shops by price range

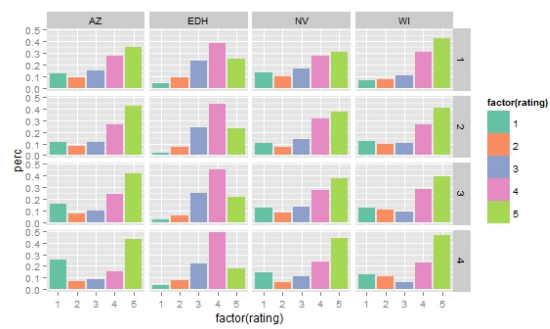(b) Histogram of the relative frequency of ratings of shops by price range and region

Figure 13