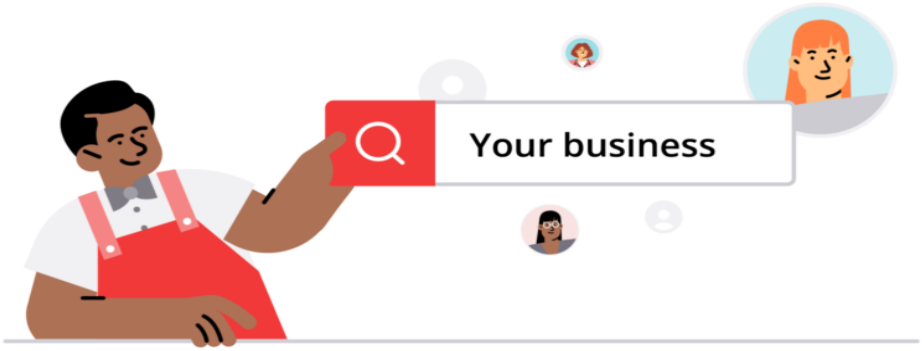
A project report on

Analysis of Yelp Dataset



BDA : Semester Project

By

Swanand Barsawade
Masters in computer science,
San Diego State University

Red ID: 825876097
Email ID: sbarsawade0801@sdsu.edu

INDEX

ABSTRACT:

Yelp is a `crowd-sourced review` forum, as well as an American multinational corporation headquartered in San Francisco, California. It develops, hosts and markets Yelp.com and the Yelp mobile app, which publishes `crowd-sourced reviews` about local businesses, as well as the online reservation service Yelp Reservations. The company also trains small businesses in how to respond to reviews, hosts social events for reviewers, and provides data about businesses, including health inspection scores.

A Harvard Business School study published in 2011 found that each "star" in a Yelp rating affected the business owner's sales by `5–9 percent`. A 2012 study by two Berkeley economists found that an increase from 3.5 to 4 stars on Yelp resulted in a `19 percent` increase in the chances of the restaurant being booked during peak hours. A 2014 survey of 300 small business owners done by Yodle found that `78 percent` were concerned about negative reviews. Also, `43 percent` of respondents said they felt online reviews were unfair, because there is no verification that the review is written by a legitimate customer.

Yelp has released part of their data which offers a chance for people to conduct research or analysis and discover what insights lie hidden in their data. I tried to analyze this data and find all the parameters that were necessary for a restaurant to run successfully and in case they were trying to expand, what insights will help them to decide the location for a new restaurant. I used Jupyter notebook to do this analysis along with the integration of Google maps API. I have applied machine learning in this project using logistic regression and sentiment analysis to extract sense out of the Yelp dataset.

ACKNOWLEDGEMENT:

I would like to convey my heartfelt gratitude to Prof. Roger Whitney for his tremendous support and assistance in the completion of my project. Your useful advice and suggestions were really very helpful to me during this project's completion. I learned that I should not use any technology just for its sheer application but understand the limitations and proper applications of each and every technology. Prof. Roger has taught me more than I could ever give him credit for here and I am eternally grateful to him.

Objective:

Customers' buying patterns and decisions are heavily influenced by social networking sites. According to a 2017 Forbes survey, consumer social analysis is the second most valuable use case for big data analytics. Yelp is a social networking site that allows customers to provide reviews and rate businesses. According to Harvard Business School, a one-star rise in a restaurant's rating results in a 5 percent to 9 percent boost in sales earnings.

I tried to undertake analysis on this data and submit the findings. This project's purpose is to use Jupyter notebook for data visualization and to derive actionable and scalable insights from the data. I have used these insights to give recommendations about ideal locations for a new restaurant.

Questions I aimed to answer:

1. How does the star distribution look like for all the businesses?
2. Which are popular states with lots of businesses?
3. Which city has the maximum number of reviews?
4. Which month has maximum number of reviews?
5. Which state has maximum number of restaurants?
6. What is the relation between stars and reviews?
7. What is the ideal location for a new restaurant?

Dataset Information:

Yelp dataset: The Yelp dataset is a subset of Yelp's businesses and reviews data that has been made publicly available for use for personal, educational, and academic purposes.

It consists of 4 tables: Business, Review, User and Checkin

Business entity:

Contains different businesses and their attributes. The unique key is business_id.
Source file type: Json
No. of columns: 15 and number of rows: 15,707

Schema:

Sn	Columns Name	Data Type	Description
1	business_id	string	unique string business id
2	name	string	business's name
3	neighborhood	string	Neighborhood of business
4	address	string	full address of the business
5	city	string	the city
6	state	string	state code
7	postal code	string	postal code
8	latitude	float	latitude
9	longitude	float	longitude
10	stars	float	star rating, rounded to half-stars
11	review_count	integer	number of reviews
12	is_open	integer	0 or 1 for closed or open
13	attributes	string	Features of the business like BYOB, Parking, etc.
14	categories	string	Categories like fast food, cafe, restaurant, etc.
15	hours	string	Day and Time the business is open

Review Entity:

Each review in this dataset has a unique review_id along with the user_id of the user who has given a review, a business_id of the restaurant, the date and time, context, and the ratings/reactions it has received by other users who have seen the review.
There are 9 columns and initially 100000 rows in the Review table. Downloaded file type: Json

Schema:

Sn	Columns Name	Data Type	Description
1	review_id	string	22 character unique review id
2	user_id	string	22 character user id, maps to the user in user.json
3	business_id	string	22 character business id, maps to business in business.json
4	stars	integer	star rating given by user
5	useful	integer	number of useful votes received
6	funny	integer	number of funny votes received
7	cool	integer	number of cool votes received
9	date	datetime	Date when review was made
10	created_dt	timestamp	Creation date-time of the table in BigQuery
11	last_updated_dt	timestamp	Table last modified
12	text	String	The entire review of business

Error Handling:

Missing Values:

The first step was to check for missing/null values in certain important columns like business_id, review_id, user_id, etc. in business and review entities. The missing values were removed from these datasets.

Garbage Values:

Garbage values like -66666, -2359,etc. were found in the business entity which were removed.

Datatype Conversion:

The reviews had punctuation marks which needed to be removed and reviews were now in simple text format which made them easier to analyze.

Derived Attribute:

Day, time, hours, and other date fields in the review entity were not all in datetime format as some of them were strings. All of them needed to be converted to analyze them quickly while merging two data frames.

Deleting Redundant Columns for EAD:

'Useful', 'Funny' and 'Cool' columns had to be removed for the purpose of exploratory data analysis to understand certain data visualizations like star rating distribution but were later, these columns were used for filtering reviews as positive, negative, and neutral reviews.

PHASES OF DATA VISUALISATION:

1) Reviews:

Analysis was done to filter reviews into positive, negative, and neutral based on the text data in the reviews columns of review entity.

1.1) Stars Assessment:

When I counted the number of reviews for different stars, it is clear to see that most people `42.7%` tended to give 5-star or 4-star ratings `23.9%`, which means that for about `66.6%` people were at least satisfied with their businesses. And then, the third highest rating star was 1-star with `13.3%`. Of course, this data makes more sense because most customers only write reviews when they have very positive or negative experiences. Writing reviews takes a decent amount of effort, so unless a customer is highly motivated or so disappointed, it's unlikely that they write anything at all (barring external factors like incentives). NBC news published an article which stated that people nowadays tend to give either perfect reviews like 5-star and 4-star or offer awful ratings such as 1-star. Sometimes ratings can't reveal the information matters to people because of fake or misleading reviews that can bump star ratings up or down.

1.2) Reviews Assessment:

More votes among 'useful', 'funny' and 'cool' usually imply that users were more agree with those reviews. Like `useful`, people felt that those `1-star` reviews were more helpful, or furthermore, more reliable. People thought extreme `low-star` reviews can reveal and offer more information about that business. While for `high-star` reviews, customers seemingly held a skeptical attitude, especially comparing with `1-star` review.

1.3) Reviews Sentiment Analysis:

I have used the VADER lexicon to analyze the sentiment of users' reviews. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media which is great for our usage. The VADER lexicon gives the sentiment of individual words. I converted text to lowercase the text since the lexicon is also lowercase. And then, replaced the punctuation with a space since it may fail to match words. The basic idea is:

i. For each review's text, find the sentiment of each word.
ii. Calculate the sentiment of each review's text by taking the sum of the sentiments of its words.

When I plotted the polarity of each review by their received stars, I could see that nearly it makes much sense. For high star reviews they usually had higher polarity and vice versa. But when comparing 4-star with 5-star reviews, their polarities were approximately equal, average polarity of 4-star review was even a little higher than 5-star's. Mostly it was because an outlier with polarity `159` which was maximal in all reviews only gave a 4-star review, and the second highest polarity `145.8` even gave a worse 3-star review.

1.4) Reviews Keywords Analysis:

I tried to find if there some certain words that often led to `high-star` reviews. So, I built a dataframe to find out top high-star words. In this part, I focused on businesses whose category contains 'Restaurant' and at least have '500' reviews. I could clearly see that most highly positive words are adjectives like `gorgeous`, `exceptional`, `outstanding`, `incredible`, `delicious`, `phenomenal`, some are adverbs which can demonstrate emotions like `absolutely`, `truly`, `perfectly`, and some exclamations such as `omg`, `wow` and so on were always present. In contrast, words that always lead to negative stars were also obvious like `upset`, `worse`, `refund`, etc.

1.5) Classifier:

It turns out that those keywords can be used reversely to predict whether the star of review is high or low. This is a classification problem, so I used logistic regression to make a classifier. First thing I needed to do was to give each review a label (0 for low, 1 for high) to represent its star, for those with 4-5 review star I counted them as high-star reviews, and on the other hand for those with 1-2 stars, I understood them as low-star reviews, and I took out all the 3-star reviews.

I used the data available for both training models and testing these models. So, I splitted the training data into separate training and test datasets. Then I used this test data to evaluate the model once it was done training. In order to take the text of an email and predict whether the text is ham or spam. This is a classification problem, so I used logistic regression to make a classifier. Recall that to train a logistic regression model I need a numeric feature matrix '*Phi*' and corresponding binary labels '*Y*'. Unfortunately, our data contains text, not numbers. To address this, I created numeric features derived from the email text and use those features for logistic regression. I created a function called `words_in_texts` that takes in a list of `words` and a pandas Series of email `texts`. It outputs a 2-dimensional NumPy array containing one row for each email text. The row should contain either a 0 or a 1 for each word in the list: 0 if the word doesn't appear in the text and 1 if the word does.

2) Business:

Businesses tend to appear in clusters, like restaurants. In this phase, I have devised a way to group together restaurants that are close to each other in Las Vegas.

2.1) K-means:

The k-means algorithm is a method for discovering the centers of clusters. It is called an unsupervised learning method because the algorithm is not told what the correct clusters are; it must infer clusters from the data. The k-means algorithm finds k centroids within a dataset that each correspond to a cluster of inputs. To do so, k-means begins by choosing k centroids at random, then alternates between the following two steps:

i. Group the restaurants into clusters, where each cluster contains all restaurants that are closest to the same centroid.
ii. Compute a new centroid (average position) for each new cluster.

2.2) Categories:

"What was the most popular category for all the restaurants in Las Vegas?" "Which category had the highest `average-star'?" To answer these questions, I first filtered out restaurants with `review-count` less than 50.
After plotting graphs, I could see that for Las Vegas and Phoenix, their top10 restaurant categories were exactly same, just a little bit different in proportion sorting. Restaurants about `Nightlife`, `Bars`, with `Sandwiches` and `Pizza` were always popular, and for exotic restaurants people were interested in `Mexican` and `Italian` food.

As for Toronto, its top10 restaurant categories were nearly identical with them, people still like restaurants about `Nightlife` and `Bars`. Only different part was that `American(new)` restaurants changed to `Canadian(new)` and I understood that people in Toronto were more interested in `Japanese` food than `Mexican` food, which resulted in more `Sushi Bars`.

What about `average-star`? Which category had the highest `average-star`? Were they same for different cities? To answer these questions, I analyzed restaurants in Las Vegas and found out that Vegan, Specialty foods and Korean were among the top categories there. Specialty Foods are foods that are typically considered as "unique and high-value food items made in small quantities from high-quality ingredients".

But when I compared using average star criteria from the previous section, I found out that only `Vegan` always had top5 average-star, and others were `Specialty Food`, `Greek` and `Coffee & Tea` which had two cities in top5 average-star. If I extend top5 to top10 average-star of restaurant categories, I found that it still only had `Vegan` in Top5 highest average-star for all three cities.

2.3) Vegan restaurants:

After running the regression model on business data entity for vegan restaurants, I figured out that the graph was a positive linear function. Hence it is very clear that the percentage of vegan restaurants was increasing every year. Moreover, if applied the average-star criteria to data for vegan restaurants, they were found to be well received and reputable.

2.4) Attributes:

I tried to analyze the 'attributes' columns in business entity to find out if there was any connection between attributes and star ratings. When I compared the 'WiFi' attribute with star ratings, it was very clear free WiFi has the potential to get higher stars and star ratings declined when the WiFi service was paid.

When I analyzed attributes like noise level at a restaurant and parking available for the restaurant, quiet and restaurants with roadside parking had higher star ratings.

2.5) Finding ideal location for a new restaurant:

I took the inputs for the zip code in which a new restaurant is trying to establish its presence and cuisine of the new restaurant. I then analyzed the datasets for by taking into consideration as much as insights as possible from the previous data analysis to find locations of the restaurants that were doing well at the requested locations. It depends on the business strategy of the new restaurant if it wants to directly compete with the already existing restaurants or wants to open in a nearby location to do good business. All the locations were plotted on a heatmap so that it's easier for business to make decision about expansion and eventually increase their profits.

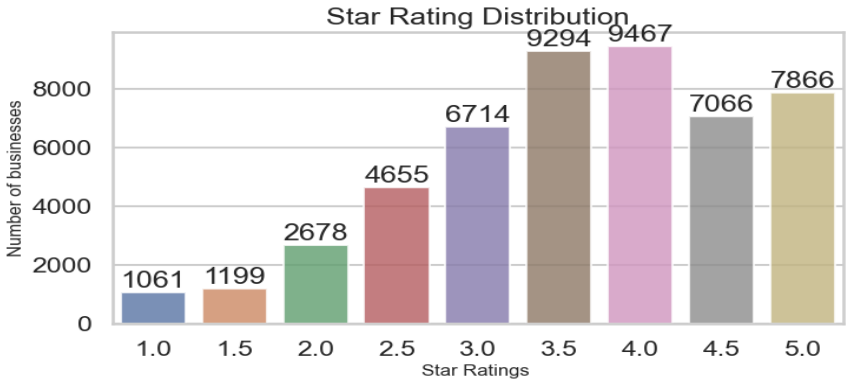
Conclusion:

The analysis of yelp dataset gives us some important insights like:

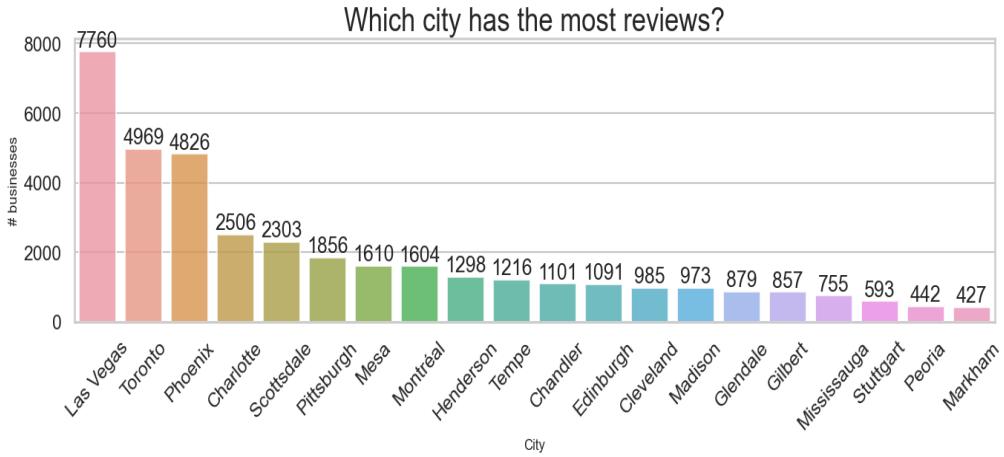
- Being a popular state on yelp simply means that it has maximum number of businesses. To maximize profits, one should focus on their business type and gather information with that perspective. For example, if the business type is restaurants, one should focus on cities like 'Las Vegas' as it has lots of restaurants.
- Maximum number revies means maximum exposure to customer problems and the factors that make customers happy.
- In the month with maximum reviews, restaurants should focus on their service more as that will affect their reputation drastically.
- Trying to compete in favorite restaurant categories is difficult, but it also gives you more customer base as a greater number of people prefer those categories.
- Restaurant owners should focus on getting the trending type of reviews for example 'cool' in this dataset as it will get the restaurant good ratings and eventually be beneficial for the business in the long term.
- Attributes of the restaurant like roadside parking, free WIFI, quiet restaurant, etc. play a vital role in getting good reviews and make it easy restaurant to be well received.
- Area of the restaurant make a huge difference for people's choices of food. For example, Sushi might be favorite category in Japan, but it might not be the same case in Canada or US. Hence restaurant should focus on universally accepted restaurants like vegan restaurants.
- Ideal location for restaurant might be considered and a good recommendation but ultimately, it depends on the business strategy the restaurant wants to opt for. It might be a good idea for some restaurants to open a branch at a location where many similar restaurants exist, but for others, it might be good if they open the branch at a farther location.
- One cannot rely completely on Yelp dataset, as it is skewed in nature. For example, New York City has its own local apps and websites, so it does not rely much on Yelp for finding good restaurants.
- No information about the traffic of population at a location is given in the Yelp dataset which means we will have to consider the populations dataset as well to find more accurate recommendations.
- No information about the median income is given in the Yelp dataset as well which means we will have to take purchasing power of people at different locations into consideration as that parameter will help to give precise recommendations.

Output Screenshots:

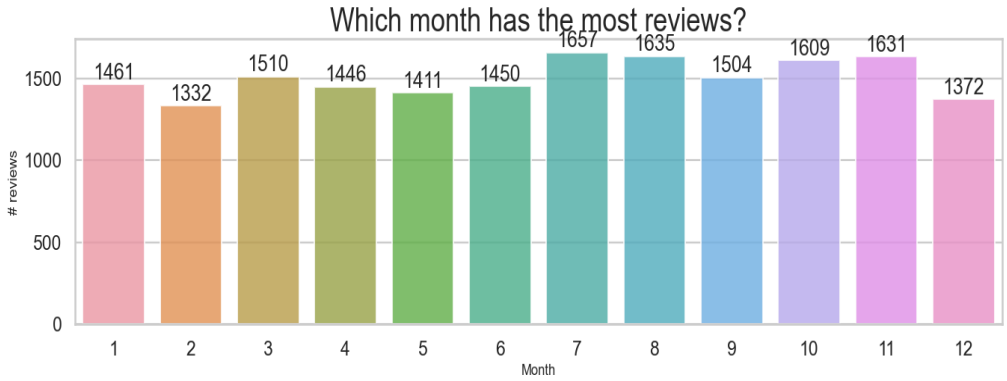
1) Distribution of star ratings across businesses.



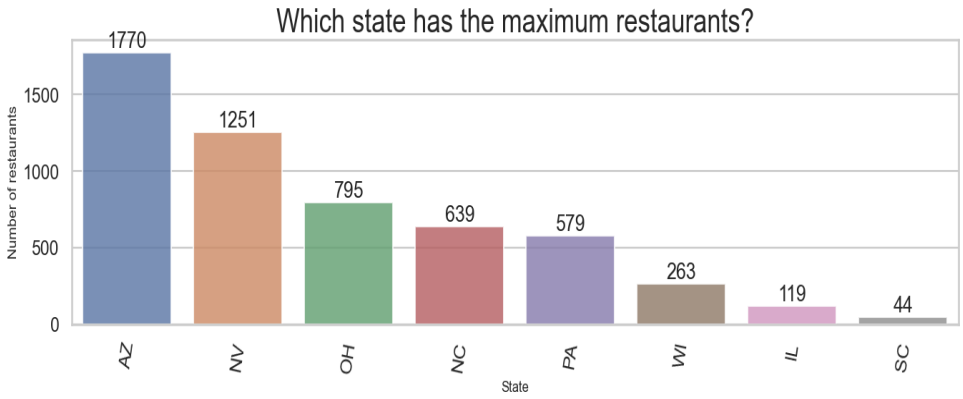
2) City with maximum reviews.



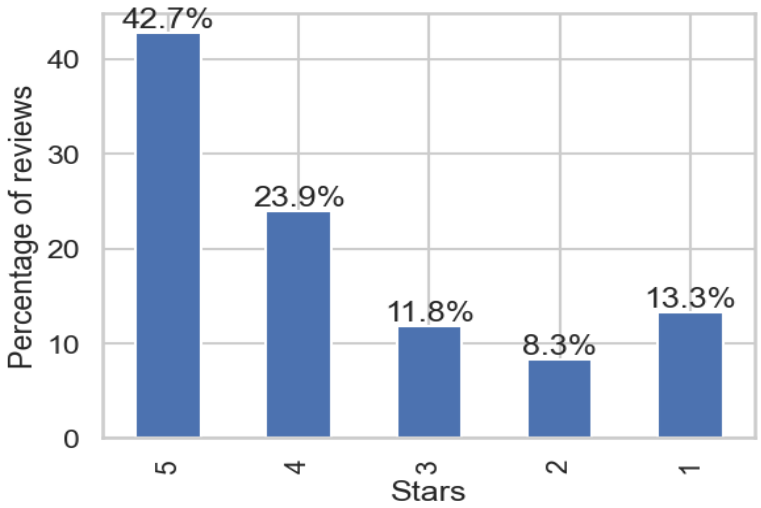
3) Month wise reviews distribution.



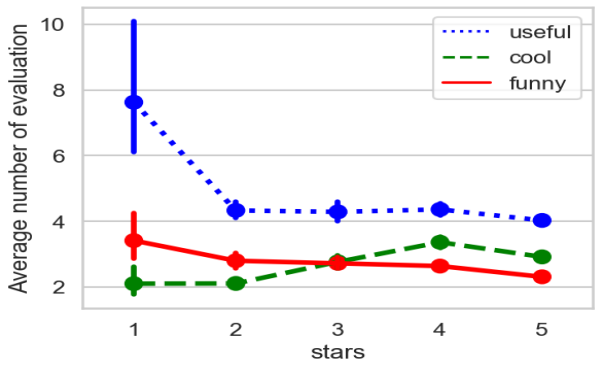
4) State wise number of restaurants.



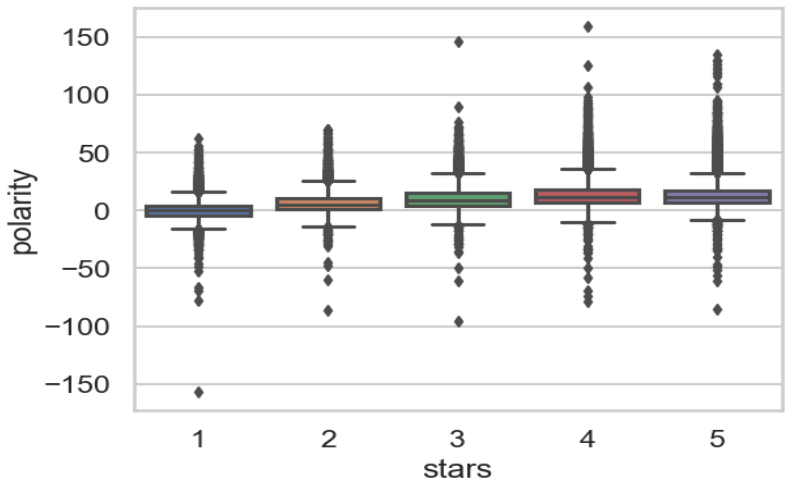
5) Stars Vs Reviews comparison.



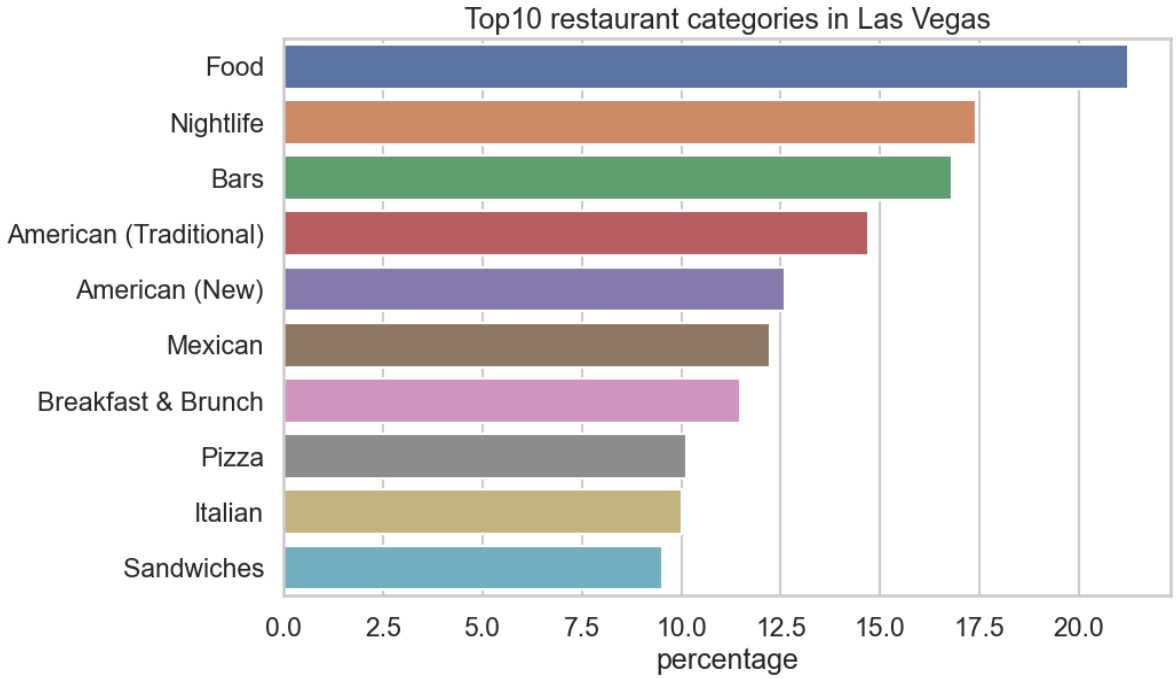
6) Evaluation of Stars Vs Useful, Funny and Cool type of reviews.



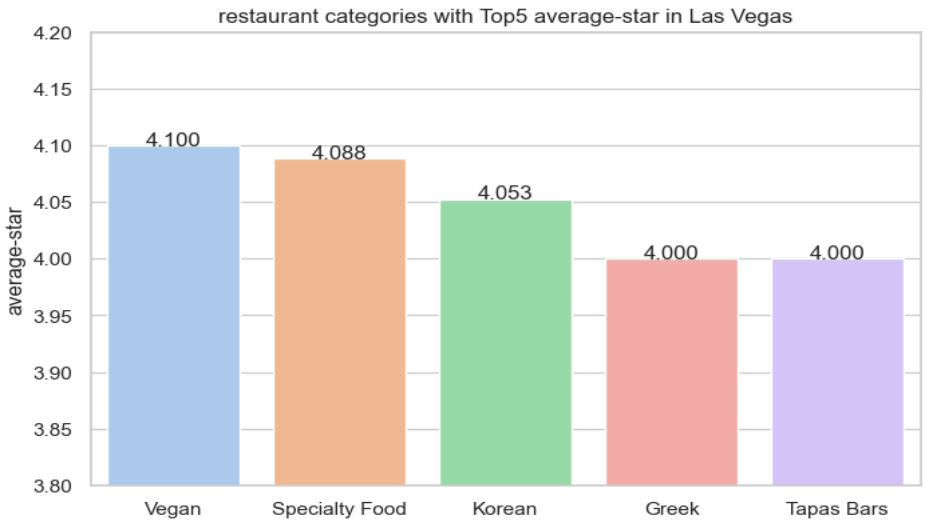
7) Evaluating polarity of star ratings.



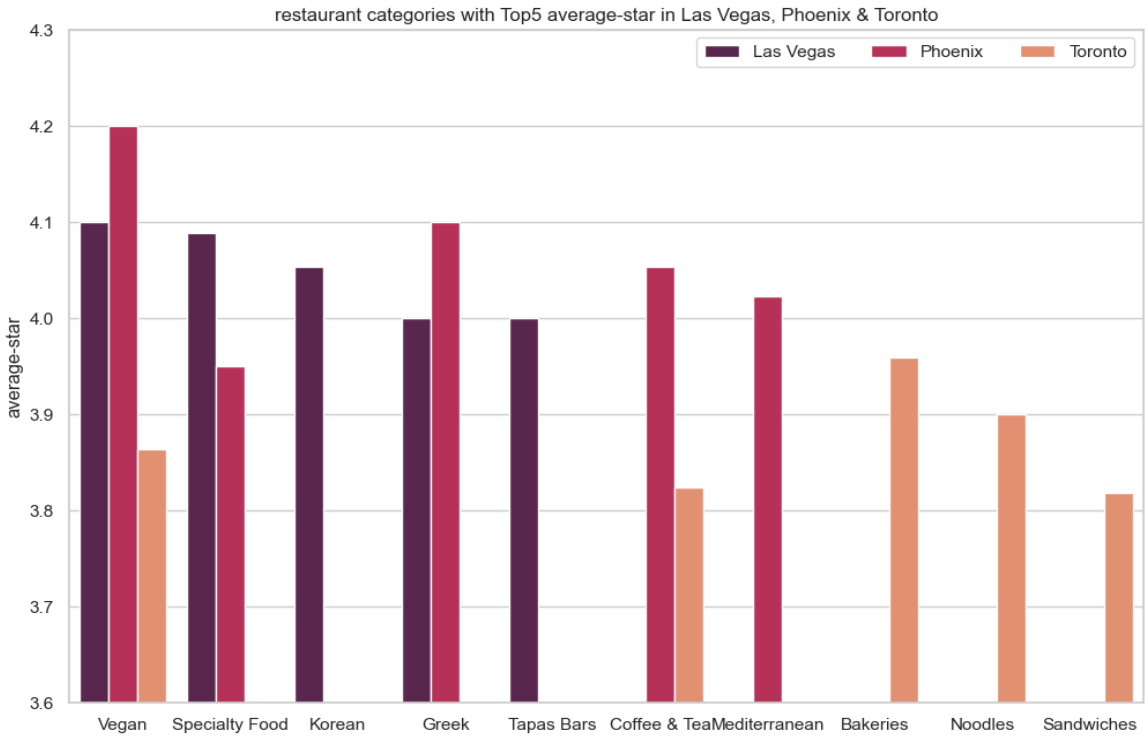
8) Top categories of restaurants in Las Vegas.



9) Top5 average-star rating categories of restaurants in Las Vegas.



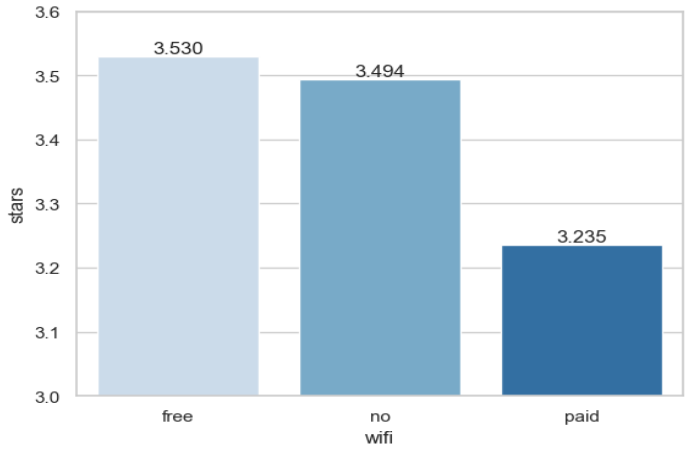
10) Comparing the average star rating for other cities. (Phoenix and Toronto)



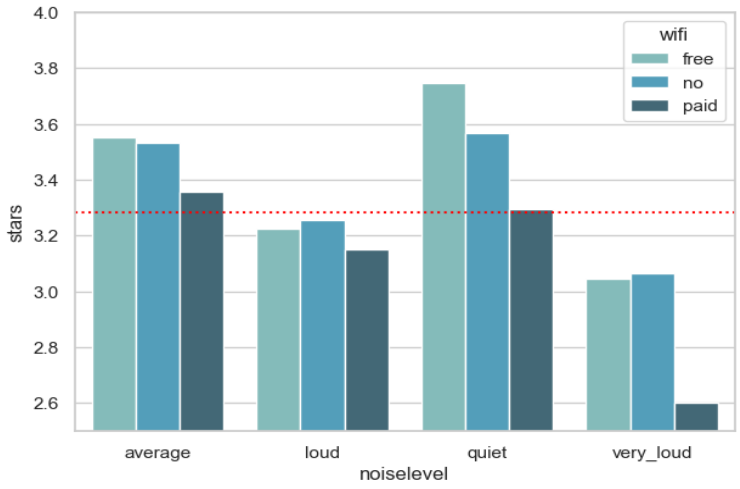
11) Trend of Vegan Restaurants over the past decade.

12) Attribute Analysis

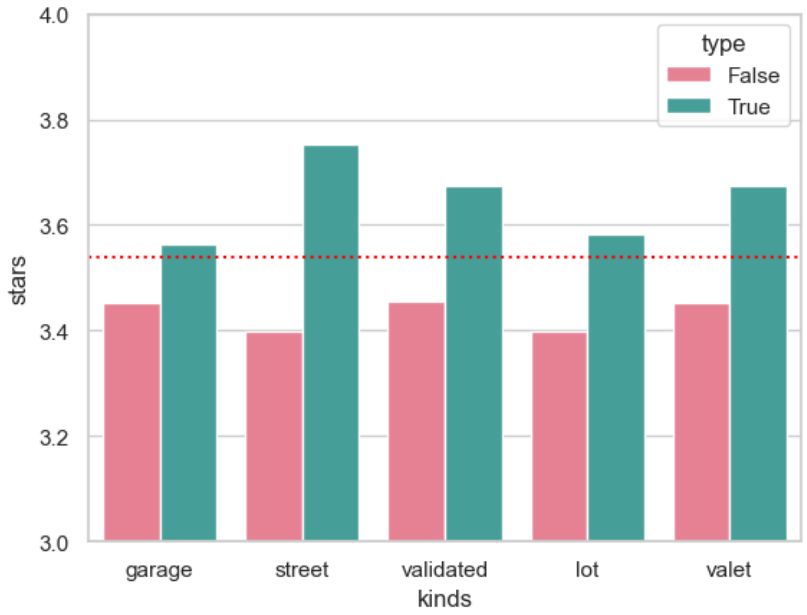
A. WIFI attribute



B. Noise level attribute



C. Noise level attribute



Future Scope:

An additional dataset can be used like the dataset from the American Community Survey (5 year estimate). The subject table S1903 which shows the median income can be take along with the locations data because that will help us predict the purchasing power of people in restaurant's neighborhood. It will be easier to predict the ideal location for a new restaurant with this added parameter, but a data pipeline will be needed for this implementation as the data gets updated every now and then.

We can also consider datasets about density of population to consider the traffic of people at a location when considering ideal locations for a new restaurant. For example, less footfall in a sparsely populated area might not be a critical problem hence we can consider those locations as well.

References:

1. https://www.yelp.com/dataset
2. http://snap.stanford.edu/class/cs224w-2015/projects_2015/Getting_Recommendation_on_Starting_New_Businesses_Based_on_Yelp_Data.pdf
3. https://towardsdatascience.com/load-yelp-reviews-or-other-huge-json-files-with-ease-ad804c2f1537
4. https://interviewnoodle.com/yelp-proximate-places-search-system-architecture-ed699952d859
5. https://cs229.stanford.edu/proj2013/SawantPai-YelpFoodRecommendationSystem.pdf
6. https://towardsdatascience.com/yelp-restaurant-recommendation-system-capstone-project-264fe7a7dea1
7. https://researchcode.com/code/1617756168/analysis-of-yelp-reviews/
8. https://krex.k-state.edu/handle/2097/38237