EXTRACTING INSIGHTS FROM YELP DATA

Arpit Sidana
Hamsika Venkataramanan
Karan Puri
Kavisha Shah
Malavika Mahalingam

INTRODUCTION

Yelp is an online business directory founded in 2014. It provides crowd-sourced reviews about local businesses in cities all over the world. Users can look up reviews and ratings about the places they want to visit, and leave behind on the website their views and opinions. Websites like Yelp are fast gaining traction due to their ease of use and novel approach for facilitating interaction of business with community. Their community of consumers is growing at a fast pace, and it is fast becoming one of the most important networking platform that affects businesses all across the globe. Yelp has over 102 million reviews and 23 million unique visitors. It also has a mobile app which users can access on the go. Mining for insights from a rich data set such as this, would uncover insights that could be of immense value to Yelp and other businesses.

DATA DESCRIPTION

We obtained the official Yelp dataset from the ongoing *Yelp Dataset Challenge, 2016*. The data is broadly categorized into five major datasets. In the scope of this analysis, we will be using the following three major datasets, descriptions of which are provided as follows:

- ***Business***

This dataset contains data about each individual business, the categories to which they belong to (eg. Food, Automotive). It is noted that each business can belong to more than one distinct category. The dataset also contains average ratings for each business, their geographical locations in terms of latitude and longitude, and the number of reviews the business has received.

- ***Review***

The dataset contains Individual reviews by users for different businesses and their ratings. Each user has a unique encrypted user ID and their star ratings of a particular business ranges from 0-5. They also have user comments which is the base for sentiment analysis.

- ***Users***

This dataset contains details of each individual user, the number of reviews they have provided till date, ratings and average stars.

DESCRIPTIVE SUMMARY

The data available in Yelp for different businesses are summarized as follows.

From this graph, it can be inferred that Yelp has the most volume of data for Restaurants, followed by Shopping places and general food places.

The plots below give an idea of the user distribution across all the states for which Yelp data is available and across all years.

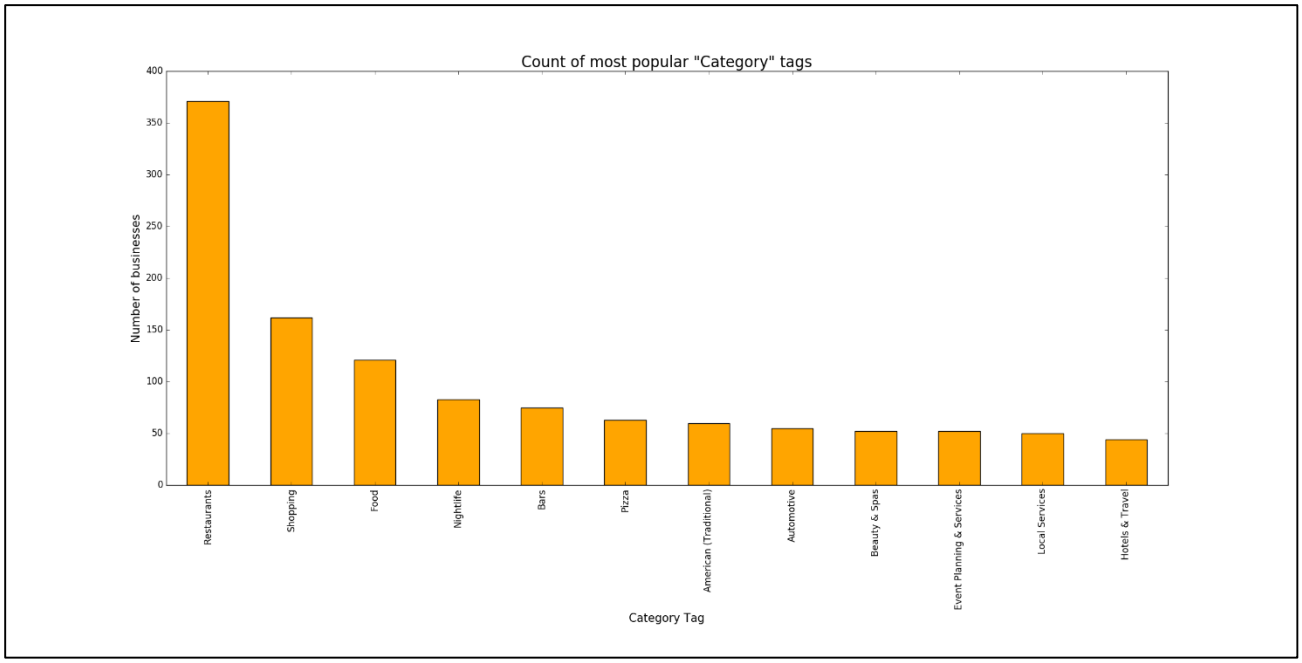

Fig 1. Most popular business category tags

We see that the number of users in Nevada and Arizona are significantly greater than the number of users in other states.

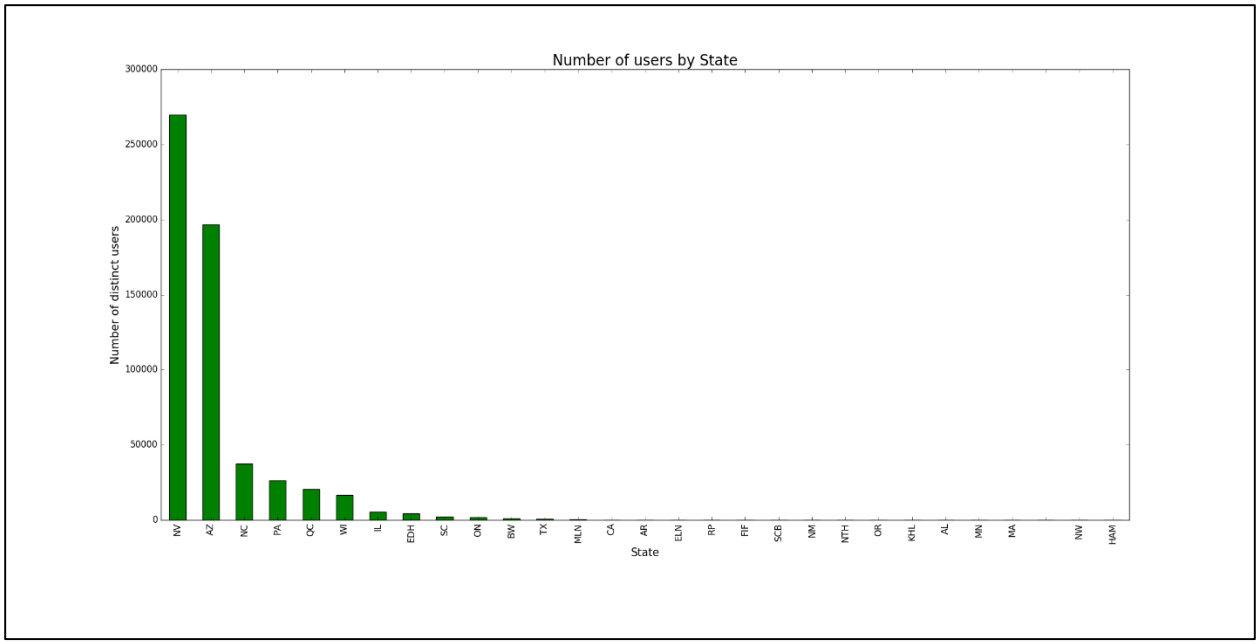

Fig 2. Users across states

It is also observed that the number of users show a consistent increase over the years, which is a testimony to the growing popularity of Yelp.

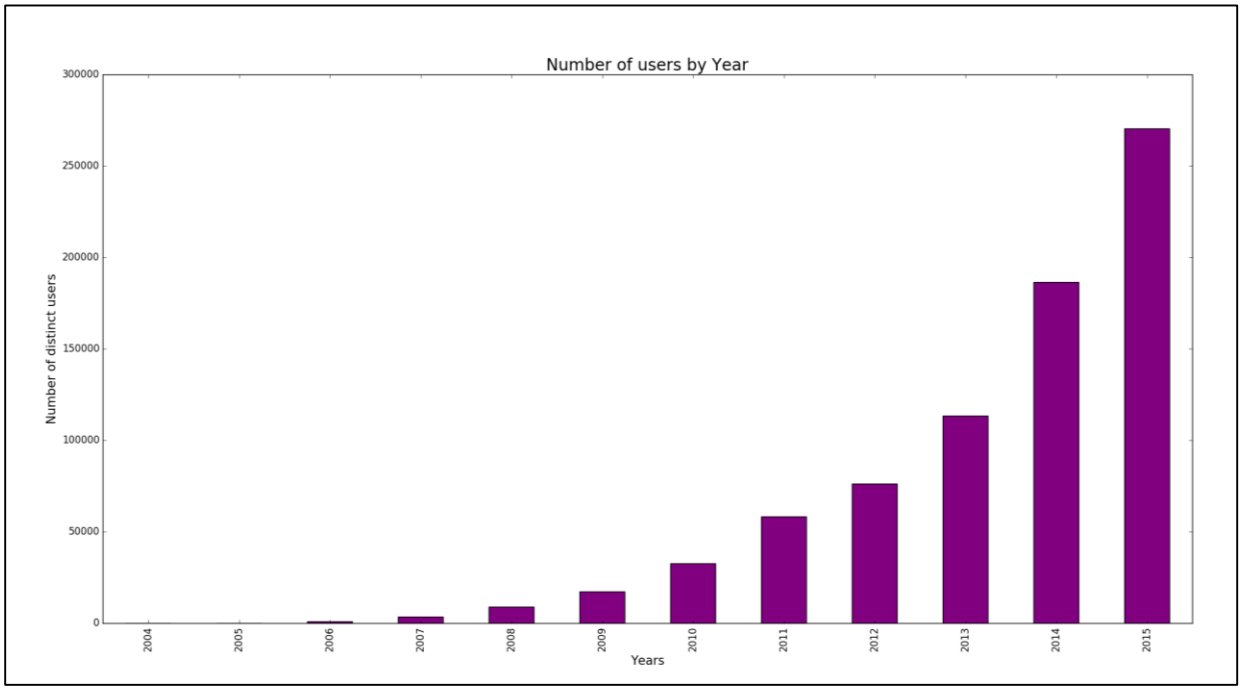

Fig 3. Users across years

RELATED WORK

The Yelp dataset has been one of the popular picks for several data scientists to work on. Some of the work that is in tandem to our work is listed below:

- ***Uncovering trends in Yelp reviews***

Berkeley School of Information student Eunkwang Joo and computer science students James Huang and Stephanie Rogers analyzed Yelp data based on restaurant reviews to predict the rating for the restaurants. They used Latent Dirichlet allocation (a generative probabilistic model for natural language processing) to detect hidden subtopics in the review texts to help them predict the ratings. They studied 50 different topics which led them to the following discoveries: ratings decrease by 0.4% of a star when customers visit during peak time, there exists a positive correlation between food quality and service and Asian restaurants were polarized with strong positive and negative reviews.

- ***Determining the authenticity of ratings***

Robert Chen used the Yelp dataset to extract names of restaurants (experimented with Indian restaurants) with genuine reviews i.e. filter the reviews to obtain the ones which are most reliable and useful for the customers. By using R, he approached the problem in 2 ways – first he gave more weight to frequent reviewers and second he generated an 'authenticity' rating by extracting the Indian names from www.indianchildnames.com to validate the ratings. The analysis led him to conclude that people of a certain ethnicity are more used to eating their cuisine and through experience alone they should be better able to distinguish quality, thus providing a more accurate rating for others to follow.

- ***Determining relationship between user rating frequency and followers***

Sung Moon from Indiana University analyzed the relationship between the number of followers at Yelp and the average rating frequency concluding that Yelp user tend to rate more frequently until some point as the number of followers' increase. By studying the review properties of average Yelp users such as well-written reviews, high quality tips, detailed personal profile and active voting record, and using statistical analysis, LDA and visualization he was able to prove stated conclusion.

- ***Predicting 'Star' in ratings from only text***

Students of University of California, Irvine conducted sentiment analysis of Yelp's ratings using review text alone by using machine learning models including Linear Regression, Support Vector Regression (with and without normalized features), and Decision Tree Regression. They found the top words to be good, food, place, great, time, back, service, menu, restaurant, ordered and chicken in this order. The final result showed Root Mean Square Error (RMSE) of 0.6 for the combination of Linear Regression with either of the top frequent words from raw data or top frequent adjectives after Part-of-Speech (POS).

ANALYSES

The aim of the project is to discover new insights from the vast business data available. We wanted to uncover hidden trends in the data by hypothesizing questions and then analyzing the data to check if the

hypotheses hold good. We narrowed down our focus into three questions of interest. These are described as follows:

1. How are the active Yelp users distributed across each state? The intent was to quantify *active users* across states that would help Yelp visualize their popularity in the US.

2. How does the performance of popular food joints (such as McDonald's) vary across each state? Businesses could gauge insights from our analyses to see the states in which it is most popular and how its performance varies across states.

3. A lot of people seem to be mentioning the weather in their reviews. Does weather on any day, at the particular geographical location have an impact of users' ratings? Looking at users' reviews, we see that it is not uncommon for users to mention weather in their comments. We glean that bad weather can create an unpleasant experience to users by causing inconvenience in parking and the like. Here, we attempt to cross an unrelated dataset with the yelp repository to check if the temperature on any given day has an effect on users' comments, which we obtain by performing a sentiment analysis on their reviews, which is a proxy for their perception of a business.

Question 1

For our first business problem we aimed to initially evaluate the total active Yelp users distributed across states in the U.S. However, since certain states are larger than others and data available might vary significantly on a state by state basis, we revamped our business problem to measure and evaluate the total percentage (%) of active users across states in the U.S. Here the term 'active' is defined by people who have been using Yelp for a long time and have posted a significant amount of reviews over time. User attributes such as 'Yelping Since', 'Number of reviews' and their state will be used to perform this analysis.

- ***Data Description***

The datasets involved to solve the business problem included business, user and review datasets. Note that business dataset contains the "business_id" and "state", the user dataset contains the "user_id" and review dataset contains both "business_id" and "user_id". This makes it imperative to include all 3 datasets to match "user_id" to "state". The variables of interest were namely, "user_id"," yelping_since"," review_count", "state" and "business_id".

We also observe that as regards U.S States data, we have data available only for the states - Alabama, Arkansas, Arizona, California, Illinois, Massachusetts, Minnesota, North Carolina, New Mexico, Oregon, Pennsylvania, South Carolina, Texas and Wisconsin.

- ***Process***

In order to access data for U.S. states corresponding to each business_id, we first used the business dataset. Only relevant metrics were retained for further analysis.

After reviewing the "yelping_since_year" values, we filtered users who had been yelping since year 2014 or earlier, as including users who joined Yelp after 2015 might not give a true picture of their activity. Next, we observed every 10^{th} percentile and deemed that users which gave 6 or more reviews (>=50^{th}

percentile) to be active. Based on these two criteria, we defined whether a user was active or not. The next step was to get the percentage of active users in each state and visualize states

- ***Results***

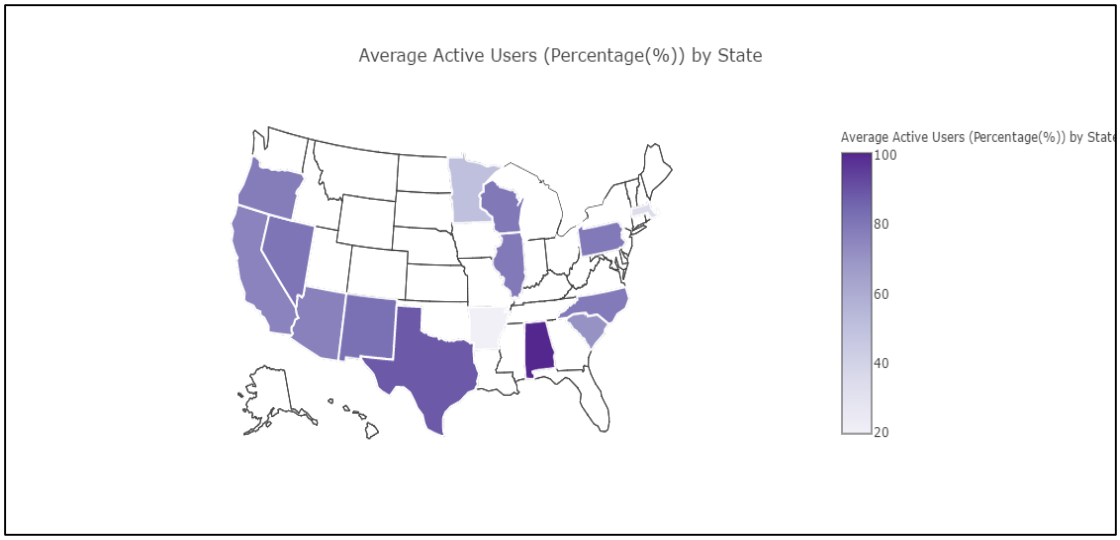

Fig 4. Active users by states (https://plot.ly/~arpit.sidana1994/10/average-active-users-percentage-by-state/)

We observe that of the data available of the aforementioned US states, Alabama records the highest number of active Yelp users with a perfect 100% active rate. This is trailed by high active percentages (>80%) at New Mexico and Texas. The lowest active percentage of Yelp users are in Arkansas (19.04%).

However, we must note that for states of Alabama, Oregon, New Mexico and Massachusetts, we lack sufficient data so the active user percentage is not a good indicator for the state. Since Alabama too records less data points, we deem the state of Texas to have the highest active user percentage.

Question 2

Based on the analysis of the data present in our dataset, we wanted to determine the performance of popular food joints based on their ratings across different states in the U.S. This information would help discover the presence as well as growth of the food joints across the nation.

- ***Data Description***

The data set involved to solve the business problem included business.json. The 'business' data set contains details of name of the restaurants, the number of stars it has received and the state in which it they are present. Note that the 'stars' are the ratings that we considered for performance of food joints and are the business ratings and not the ratings provided by the users. The variables of interest were state, name and stars which were used to determine the performance of food chains across different states.

- ***Process***

In order to determine the performance of popular food joints, we analyzed the data to identify the popular food joints. We first began by identifying food joints from our dataset by extracting 'fast food' from our 'Category' variable. We further dissected the data by extracting the fast food joints on the basis of their names, their location (states where they are present) and the ratings (stars) they have been awarded.

We grouped all the food joints on the basis of the states in which they are present and their names. This gave us an insight into the popularity of the states i.e. which state had the highest number of fast food joints.

The top five states are as follows:

Name	Number of food joints
AZ	1243
NV	812
NC	266
EDH	168
WI	77

On obtaining the above information, we defined our criteria for popularity. If the 'presence' of any food joint in a particular state is greater than 10%, then that food joint would meet the criteria of a popular food joint. Keeping this in mind, we calculated the presence of each food chain by taking a sum of the total stores of the food chain in one state (say Arizona) over the total of all food joints in Arizona. If this exceeded 10%, then the food joint was a popular one. This helped us to get popular food chains which had strong presence in different states.

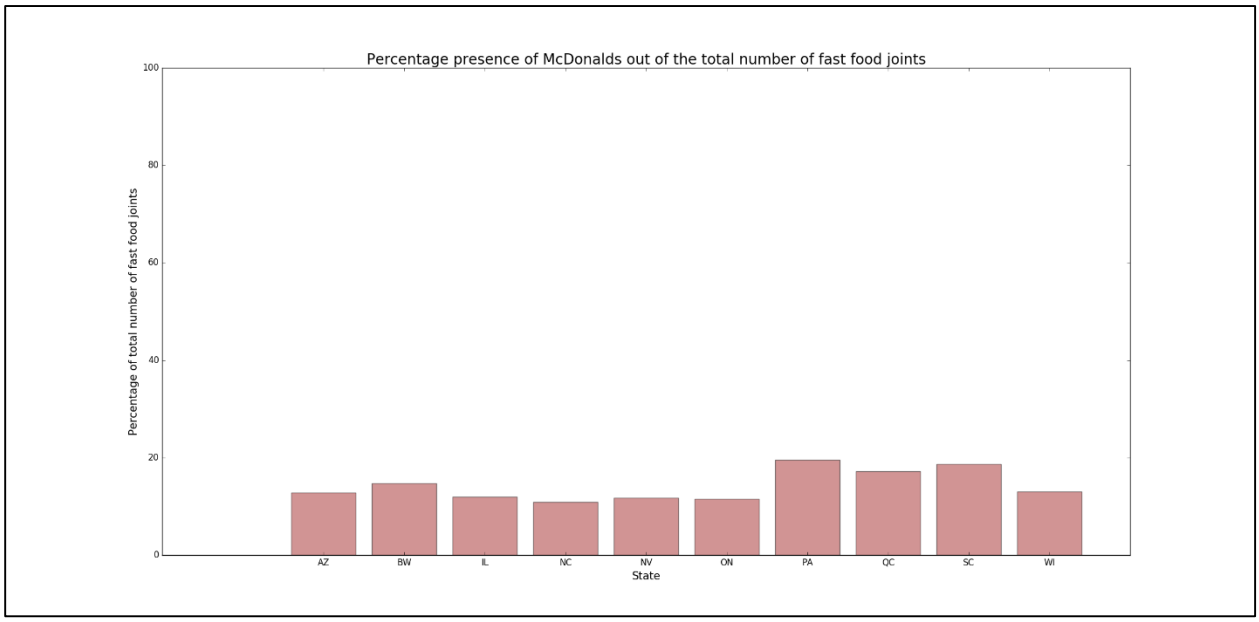

Fig 5. Presence of McDonalds across states

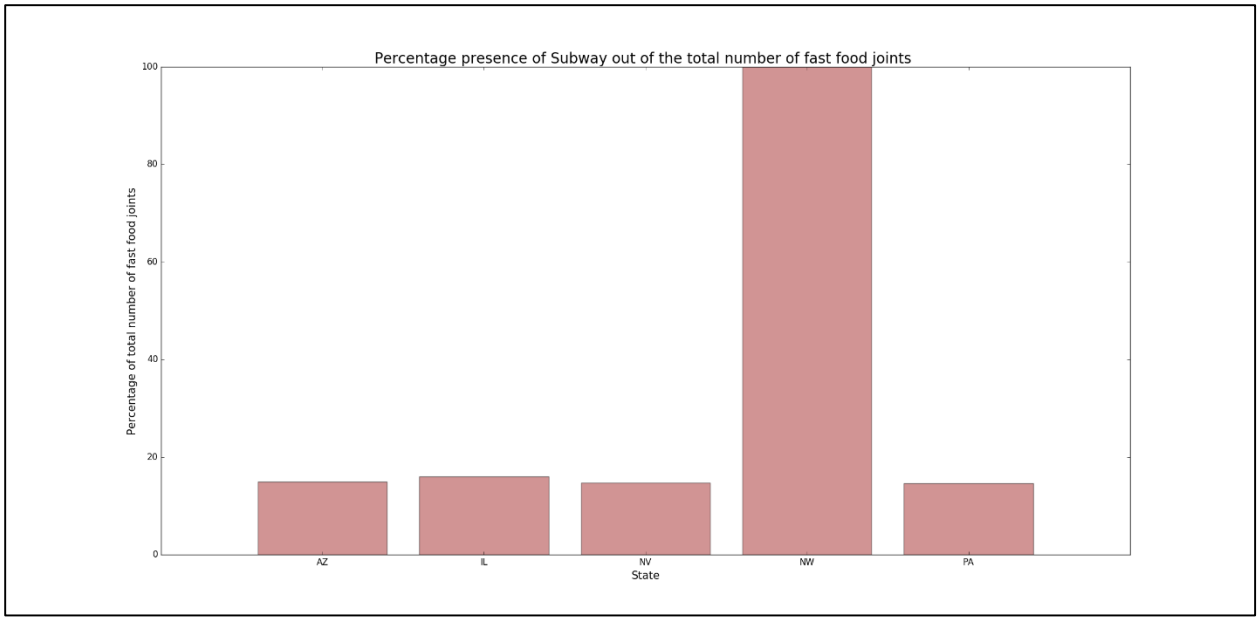

Fig 6. Presence of Subway across states

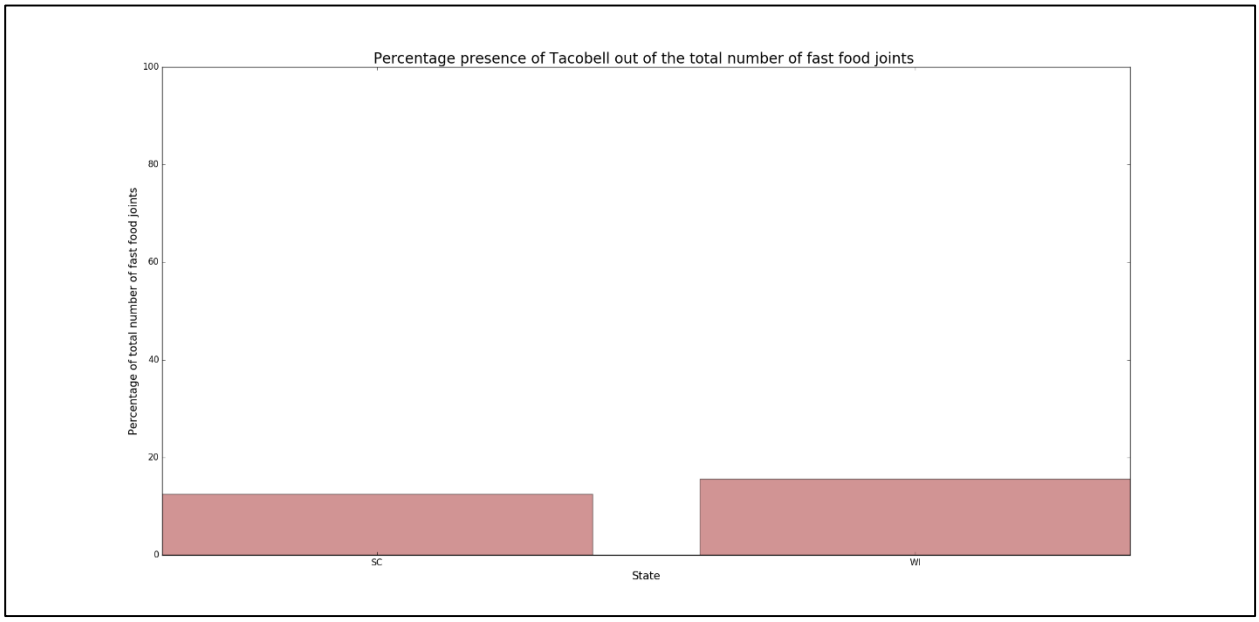

Fig 7. Presence of Tacobell across states

- ***Results***

The above graphs represent the represent the rating variations of the fast food joints across various states where the presence of each of these fast food joints is greater than 10% of all fast food joints present in that state.

Fig 5 indicates that the ratings for McDonalds is highest in Pennsylvania (PA) and South Carolina(SC). Fig 6 indicates that the ratings for Subway in NW outweigh all the states where it is present. Fig 7 shows that

Wisconsin(WI) and South Carolina(SC) both have almost similar ratings for Tacobell and are the only places where Tacobell has a presence greater than 10% suggesting its need for expansion.

Question 3

For the third business question, while reviewing the sample data for user reviews, we noticed that users seem to be mentioning weather in their reviews. We wanted to see if weather, in anyway, was affecting users' feedback for a place. We also wanted to do a sentiment analysis to check if their ratings on a range of 1-5 stars was representative of their comments.

- ***Data Description***

The data used for the purpose of analyzing this business problem is obtained from the business.json and review.json files. The review data set contains details of each individual user review, their comments and their ratings. It also contains the business for which the feedback is provided. Using this data, we obtain the latitude and longitude for each business and the corresponding historical daily temperatures are extracted using DarkSky's forecastio application. The analysis performed is for reviews for the month of January 2015 across all regions in the interest of decreasing the run time of the code.

- ***Process***

To analyze the sentiment of the users writing reviews, we followed a basic algorithm to give each review a sentiment score. The idea was to identify words that add a positive tone and the words that hint at a negative tone in the comment. To do this we used the bag of positive and negative words from an online source. Between user reviews and each of the two files, we created a list of common positive and negative words for every review that a user gave.

The sentiment score was calculated using the formula-

Sentiment score= (# of positive words – # of negative words)/Total number of words in the comment

These sentiment scores were bucketed as positive (score greater than 0), negative (score less than zero) and neutral (score equal to zero). It was important to count all the positive and negative words including repetitions:

Example 1– "The food was good. Never tasted anything like this. Good place"

If repetitions were not counted, our algorithm would give us a count of one good word and one bad word for this comment ("Never" reflects negative sentiment and "Good" reflects positive sentiment). The overall sentiment of the comment would then be *Neutral*, although it is a positive comment from the user.

Example 2– "The food was good. Great place. Good service."

In this example, the word "good" is repeated twice. It is important to count both to fully capture a user's sentiment.

From the forecastio api which enables us to look up weather anywhere in the globe, we then obtained the average daily temperatures for each day, for each latitude and longitude corresponding to each unique business id from the business table. This we did by obtaining a unique key and installed it using pip install python-forecastio. We then used forecast.hourly() to obtain hourly temperatures for each day.

From this, the average daily temperatures are calculated. This is one of the predictors in our linear regression model.

The daily temperatures are then appended to the review table, which now contains the sentiment score of each review. We also analyze whether users' ratings are consistent with their reviews. For this purpose, we create dummy variables to indicate how many stars a user has given for a particular review. These are also added as a predictor to the model, to see if it has a significant impact on the sentiment scores.

- ***Results***
  The output of our linear regression model is as follows:

OLS Regression Results

Dep. Variable:	score	R-squared:	0.228
Model:	OLS	Adj. R-squared:	0.228
Method:	Least Squares	F-statistic:	1519.
Date:	Wed, 24 Aug 2016	Prob (F-statistic):	0.00
Time:	01:49:40	Log-Likelihood:	42240.
No. Observations:	25672	AIC:	-8.447e+04
Df Residuals:	25666	BIC:	-8.442e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	0.0528	0.001	41.347	0.000	0.050	0.055
temperature	0.0001	2.4e-05	4.461	0.000	6e-05	0.000
star1	-0.0678	0.001	-79.215	0.000	-0.069	-0.066
star2	-0.0503	0.001	-45.299	0.000	-0.052	-0.048
star3	-0.0340	0.001	-34.006	0.000	-0.036	-0.032
star4	-0.0121	0.001	-15.941	0.000	-0.014	-0.011

Omnibus:	20309.593	Durbin-Watson:	1.597
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1829131.322
Skew:	3.197	Prob(JB):	0.00
Kurtosis:	43.855	Cond. No.	250.

The model which predicts a user's sentiment score based on the temperature of the day and the ratings given by the user accounts for 22.8% of the sentiment score. Keeping the ratings given by a user fixed, a degree Fahrenheit increase in daily temperatures causes an average of 0.001 increase in the user's sentiment score.

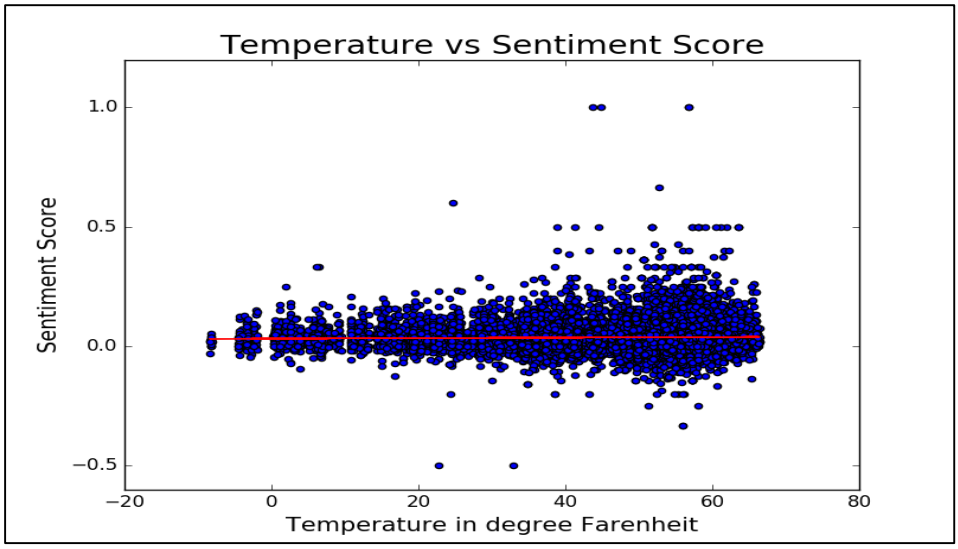

Fig 8. Relationship between temperature and sentiments

Although there is a relation between the two factors, it is not clear from this model the extent to which temperature impacts rating given that the data was analyzed only for a month across all regions. Further analyses using the entire data available will present a more accurate representation of the relationship.

The analysis of user ratings vs sentiment score proved to be more accurate.

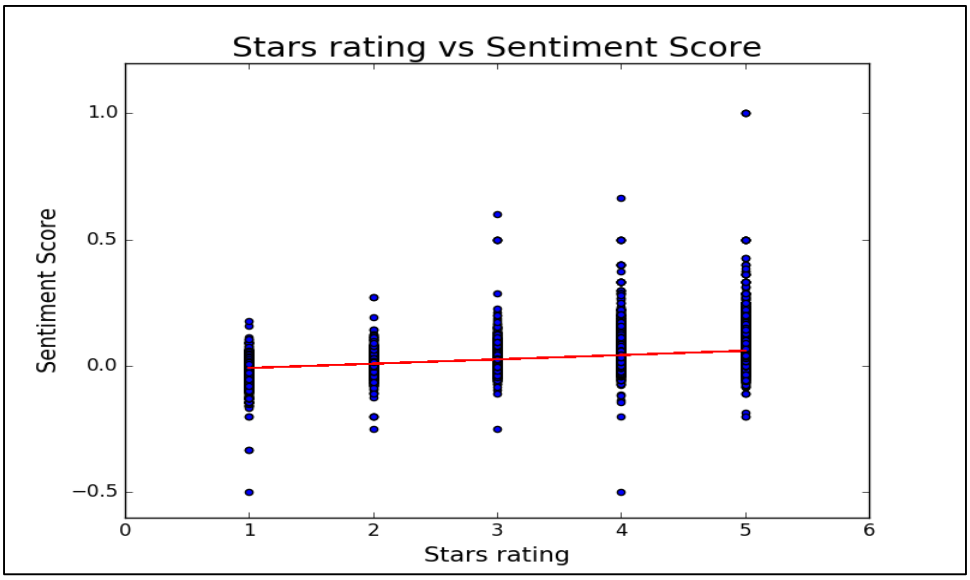

Fig 9. Relationship between star rating and sentiment score

Keeping the temperature constant, a user who gives a rating of 1 has a sentiment score of 0.0678 less than a user who gives a rating of 5, whereas a user who gives a rating of 4 has a sentiment score of just 0.0121 less than someone who gives a 5-star rating. This finding is consistent with our belief that positive comments indeed go hand in hand with higher ratings. It also validates the algorithm we used to calculate the sentiment score.

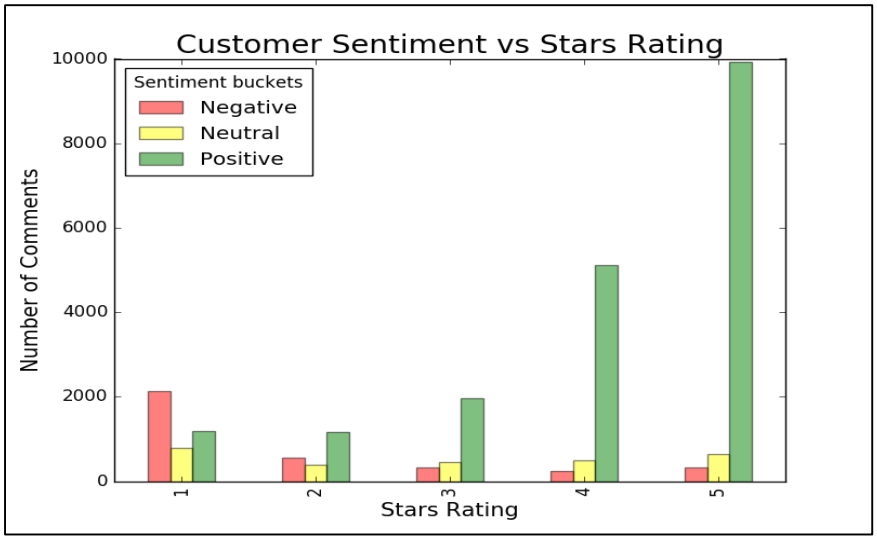

Fig 10. Relationship between customer sentiment and their stars rating

BUSINESS VALUE AND FUTURE SCOPE

The paper presents various insights from the business data available in Yelp.

The analysis is started by finding the distribution of active Yelp users across different states in the US. Based on our analysis, we can infer that if Yelp were to launch promotional or loyalty programs and needed to test waters first, the state of Texas would be a good place to start to measure and evaluate their promotional activity.

We then wanted to get an idea about how the most popular food joints perform across different states in the US. From this we glean that it would be interesting to note the effect of emergence of new food joints on existing ones. As Yelp maintains a database of all food joints where customers rate their experiences and post their reviews, the analysis of this information is particularly critical for businesses. With this information, we believe:

- The food joints can focus their marketing efforts on locations which are highly rated and most profitable
- Improve foot traffic to specific locations by placing ads
- Geo spatial targeting of customers through GPS data (‘restaurants near me’ concept)
- Gain better understanding of competition
- Gain insight on pricing
- Launch new products based on location based preferences

Lastly, from the analysis of the effect of temperature on user reviews, we see that the temperature turns out to be significant in affecting reviews, albeit in a small way. Businesses can still use this insight to be better prepared to adapt themselves to different weather conditions and thereby increase user satisfaction.

In all, the analysis gave us interesting, and sometimes, unexpected insights on the performance of different businesses across various regions. This analysis can be further extended to improve our understanding of the complex realm of user sentiments and what makes businesses successful.

REFERENCES

https://www.yelp.com/dataset_challenge

https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets

https://developer.forecast.io/

https://www.springboard.com/blog/eat-rate-love-an-exploration-of-r-yelp-and-the-search-for-good-indian-food/

http://blog.nycdatascience.com/r/project-1-exploratory-visualizations-of-yelp-academic-dataset-draft/

http://www.ischool.berkeley.edu/newsandevents/news/20131004yelpdatasetchallenge

https://arxiv.org/ftp/arxiv/papers/1401/1401.0864.pdf