Introduction and Dataset

Introduction

Yelp is a business directory service and crowd-sourced review forum which allows users to give reviews to businesses based on their own experience. It's easy for human to understand the sentiment behind the reviews, specifically whether a review is positive or negative, even without having any context about the business. However, there is a huge amount of reviews data on yelp, it's not realistic for business owners to go through every review by themselves to improve their performance.

Our project is aimed to:

1. Provide useful, analytical insights to business owners on Yelp and, based on these insights, propose data-driven, actionable decisions to said owners in order to improve their ratings in Yelp.
2. Build a web dashboard/widget/web application that visualizes your analysis and makes it easier to understand for business owners

The Dataset

There are four datasets provided, business data, review data, tip data, and user data. We are going to focus on the analysis of attributes in the business data and the review texts in the review data of all the mexican restaurant.

For the **business data**, there are 4628 businesses who are categorized as mexican restaurant, we transform all the attributes into data frame and filter out the attributes with more than 90% of missing and impute the rest of missing values by using KNN (which use the majority value of the nearest 5 neighbors to substitute the missing value). Then we build linear model based on the clean business data, and give recommendations after conducting the χ^2 test. The clean data include the following 40 variables:

business_id	name	latitude	longitude	stars	review_count	is_open	WheelchairAccessible	RestaurantsT
romantic	intimate	classy	hipster	divey	touristy	trendy	OutdoorSeating	RestaurantsTable
upscale	casual	dessert	latenight	lunch	dinner	brunch	breakfast	RestaurantsPrice
BikeParking	GoodForKids	Caters	NoiseLevel	HasTV	WiFi	Alcohol	park	RestaurantsRese

As for **reviews data**, there are 403941 reviews which belong to mexican restaurants. We tokenize/lowercase/lemmatize the review text and extract the top 1000 most frequent words. Following that, we manually pick up 44 words that make sense for analysis, which are

taco	burrito	salsa	chips	rice	guacamole	asada	chipotle	tortilla	salad
nacho	enchilada	meat	time	chicken	fish	shrimp	beef	pork	steak
service	wait	minute	staff	waitress	waiter	atmosphere	price	breakfast	lunch
dinner	line	drink	margarita	table	fries	manager	beer	patio	bacon
vegan	quesadilla	chile	enchilada						

Then we give recommendations based on our analysis.

EDA

The details of EDA can be found in the file, "EDA_summary.pdf".

Recommendation Based on Business Attributes

For the **business data**, there are 4628 businesses who are categorized as mexican restaurant, we transform all the attributes into data frame and filter out the attributes with more than 90% of missing.

Imputation: KNN

Since there are many missing values in remaining attributes of businesses, we need to impute the missing values. We use KNN model to perform the imputation:

- Distance measure: Gower distance
- Hyper-parameter: k = 5
- Voting method: majority voting

Linear Regression

We fit a stepwise linear regression (using AIC as criterion) with the following formula:

$$rating \sim attributes$$

The following table shows the summary table of significant predictors:

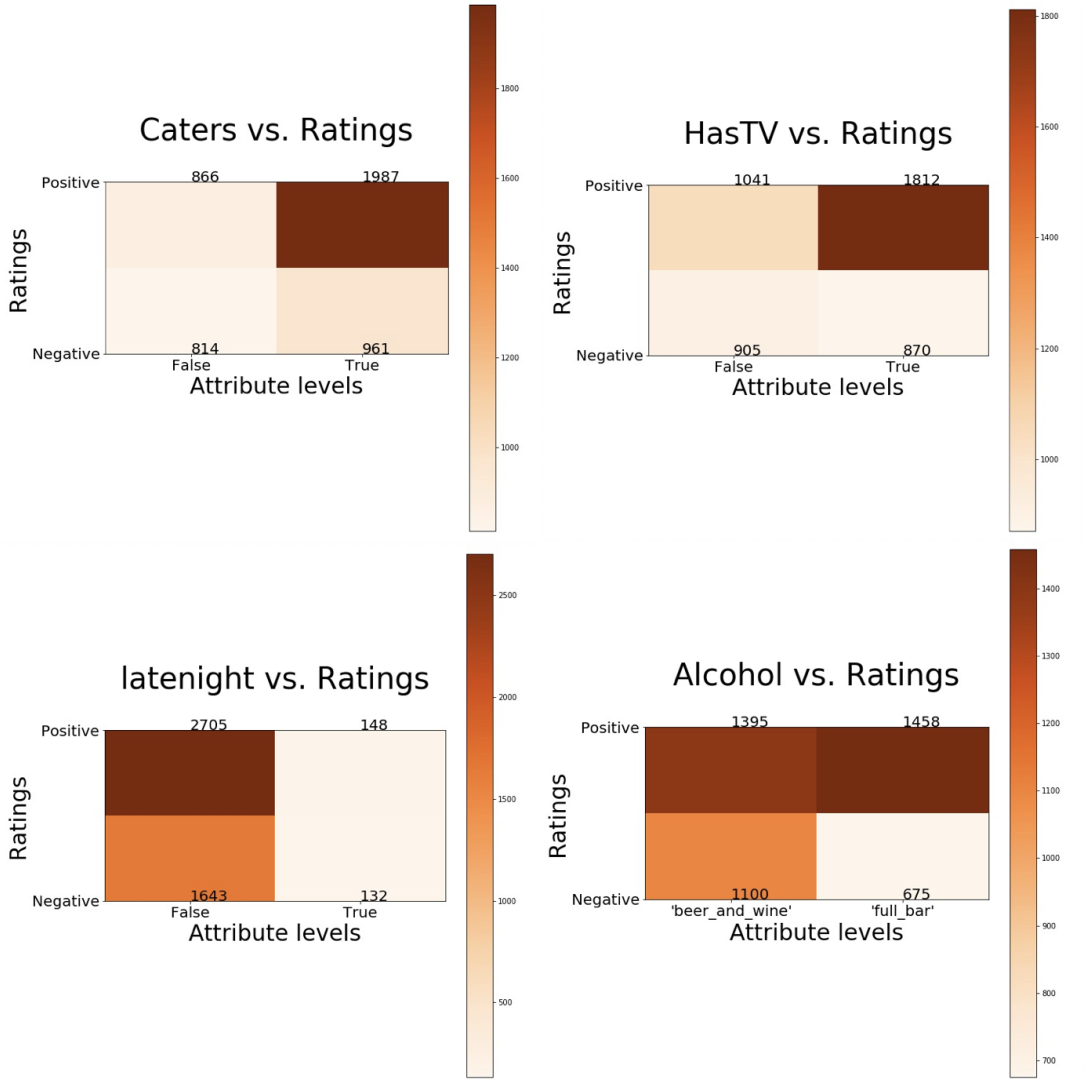
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.169	0.226	14.041	0
longitude	-0.003	0.001	-2.840	0.005
review_count	0.001	0.0001	7.414	0
intimateTrue	0.484	0.168	2.873	0.004
hipsterTrue	0.266	0.089	2.984	0.003
diveyTrue	0.218	0.063	3.431	0.001
trendyTrue	0.251	0.060	4.206	0.00003
dessertTrue	0.169	0.070	2.413	0.016
latenightTrue	-0.165	0.047	-3.488	0.0005
dinnerTrue	0.069	0.027	2.509	0.012
brunchTrue	0.243	0.045	5.351	0.00000
GoodForKidsTrue	0.125	0.044	2.821	0.005
CatersTrue	0.265	0.024	11.126	0
NoiseLevel'loud'	-0.201	0.046	-4.366	0.00001
NoiseLevel'quiet'	0.082	0.028	2.896	0.004
NoiseLevel'very_loud'	-0.225	0.080	-2.799	0.005
RestaurantsTableServiceTrue	0.209	0.033	6.366	0
RestaurantsPriceRange22	-0.234	0.030	-7.783	0
RestaurantsPriceRange23	-0.239	0.113	-2.120	0.034
OutdoorSeatingTrue	-0.057	0.024	-2.411	0.016
HasTVTrue	0.135	0.026	5.188	0.00000
WiFi'no'	0.184	0.024	7.536	0
Alcohol'full_bar'	0.089	0.029	3.023	0.003
RestaurantsDeliveryTrue	0.116	0.034	3.385	0.001
BusinessAcceptsCreditCardsTrue	-0.569	0.077	-7.394	0
WheelchairAccessibleTrue	-0.322	0.071	-4.533	0.00001
park	0.079	0.026	3.053	0.002

χ^2 test

Considering businesses with 3.5~5.0 stars as positive ones, businesses with 1.0~3.0 stars as negative ones, we performed χ^2 test for attributes cross ratings and dropped attributes that were not significant (we dropped **OutDoorSeating**, **WiFi** and **GoodForKids**):

Attribute	Chi_square_p_value	Attribute	Chi_square_p_value
intimate	6.218e-03	NoiseLevel	8.050e-05
hipster	5.350e-06	RestaurantsTableService	4.986e-30
divey	5.737e-04	RestaurantsPriceRange2	2.985e-02
trendy	2.007e-15	BikeParking	7.094e-03
casual	1.104e-14	HasTV	3.505e-22
dessert	1.794e-07	Alcohol	5.284e-18
latenight	2.235e-03	RestaurantsGoodForGroups	1.455e-02
lunch	4.136e-11	RestaurantsDelivery	2.419e-08
dinner	3.136e-25	BusinessAcceptsCreditCards	6.745e-06
brunch	2.081e-09	WheelchairAccessible	2.827e-04
RestaurantsReservations	3.953e-11	park	1.092e-16
Caters	2.056e-26		

Some contingency tables for siginificant attributes are as following:



Recommandations

Divide the significant predictors into two lists: one for positive predictors (contains the predictors with positive coefficients), the other for negative predictors (contains the predictors with negative coefficients). We will give the business owners suggestions based on such results:

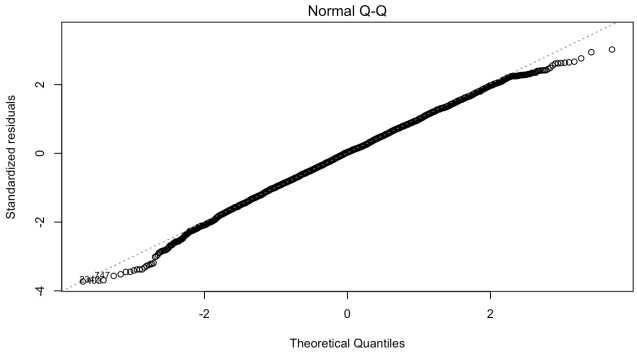
Positve	Negative
intimate	latenight
hipster	NoiseLevel
divey	RestaurantsPriceRange
trendy	BusinessAcceptsCreditCards
dessert	WheelchairAccessible
dinner	
brunch	
Caters	
RestaurantsTableService	
HasTV	
Alcohol	
RestaurantsDelivery	
park	

Take business **New Mexican Grill** (business ID: 1Dfx3zM-rW4n-31KeC8sJg) as example, we can give the owner sugestions as following:

- **Ambiance**: Please provide intimate, hipster, divey or trendy ambience. Your estimated rating will increase by 0.49 if you provide intimate ambience, increase by 0.25 if you provide hipster ambience, increase by 0.22 if you provide divey ambience and increase by 0.25 if you provide trendy ambience.
- **Dessert**: Please provide food that are good for desserts. Your estimated rating will increase by 0.16 by doing this.

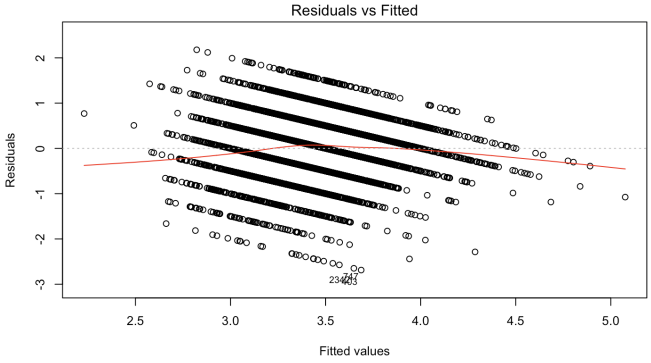
Strenghs

- From the QQ plot, we can see that the assumption of normality is well satisfied.
- Can provide business owners with interpretable and clear recommendations with statistical proof.



Weaknesses

- The coefficient of determination R^2 is not high enough.
- The residual vs fitted plot contains some patterns, which indicates that some assumptions (e.g. equal variance) may not be well satisfied.



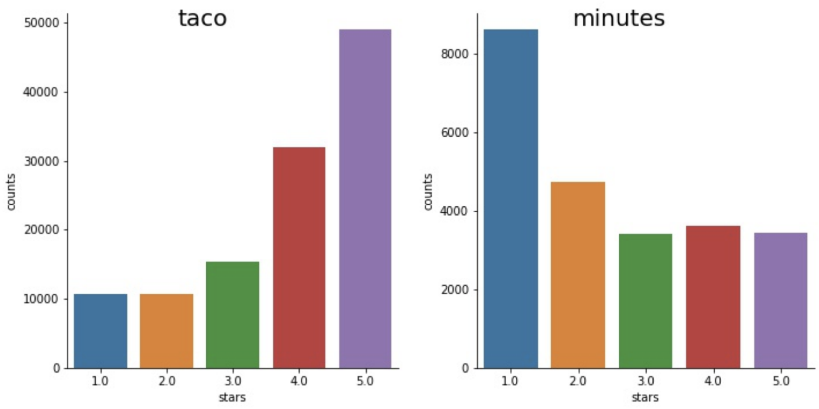
Recommendation Based on Review Texts

We tokenize/lowercase/lemmatize the 403941 review texts and extract the top 1000 most frequent words. Following that, we manually pick up 44 words which can be basis of recommendations for Mexican restaurants, and build a contingency table for each of them. For example, the contingency table for **taco** is:

	Rating Positive	Rating Negative
Taco Positive	11832	4577
Taco Negative	949	1551

The way we define the attitude in a review towards a target word is to pick up it's neighborhood, which means 6 words, and check the number of positive or negative opinion words in them. As for how to determine whether a word have a sentiment, we use a dictionary created by Liu and Hu[1](https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon). If positive opinion words around a target word is more than negative ones, we believe that the review we check contains a positive attitude towards the target word, and a negative attitude otherwise.

The reason we use χ^2 contingency test is we would like to know if there is a significant relationship between attitudes to target words and the rating of relevant reviews. After that, we can get a significant wordlist. Following is some examples of the distribution of the significant words.



We can see that **taco** is positive related to the ratings and **minutes** is negative related to the ratings. Both words show an obvious pattern which is pretty significant.

The next step is to classify the reviews of each business into two categories, positive reviews or negative reviews, and then use *TF-IDF* to extract the 100 most "important" words in these two categories. And we get positive aspects and negative aspects for each business by getting intersections of those "important" words and significant words we get earlier in the summary.

Below is an example of what recommendations we can give to the **New Mexican Grill** owners based on reviews for their businesses.

Positive parts	Negative parts
burrito	asada
pork	salsa
chile	time
salad	

Positive parts	Negative parts
chicken	

Shiny App

Shiny App address: https://1037761185qq.shinyapps.io/Tue-G7/ (https://1037761185qq.shinyapps.io/Tue-G7/)

Contributions

Han Liao and Lingfeng Zhu: Generate recommendations based on business attributes.

Qiaochu Yu and Yujie Zhang: Recommendations based on review data.

Summary and app are developed by all the members.

References

The negative/positive wordslist comes from: https://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar (https://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar)