Final Report: Analysis on Taco Business

Group 20: Yuanhao Geng, Xuelan Qian, Zhanpeng Xu, Xinyu Li

1 Introduction

We choose to analyze the data of restaurants that mainly serve tacos to help people find the key elements in running a successful taco business. Our analyses are based on two CSV files from the dataset, business and review data. Our main goal is to aid people in selecting the type of sauce and dish of tacos as well as the service and environment that can attribute to higher ratings on yelp. Also, we try to find the main attributes that are related to higher ratings of a taco restaurant.

Linear regression analysis has been done on the attributes from business data to find the attributes that can allocate a higher rating on Yelp. Sentiment analysis has been done on the users' reviews to understand people's preferences about the types of sauces and dishes, also attitudes about service and environment.

2 Data Processing and EDA

Our dataset is a subset of Yelp's businesses, reviews, tips, and user data. The business data contains information about the location, attributes, and categories of the business. The review data contains the full review text, the *user_id* that wrote the review, and the *business_id* the review is written for. We first identified all businesses with the name *Taco*, *taco*, *Tacos*, or *tacos* in the business dataset and extracted the information related to these businesses. Based on the *business_id* of the taco restaurants, we then selected all the reviews by matching the *business_id* in the review dataset.

2.1 Business Data

For the linear regression analysis on the attributes of taco restaurants, we included two large types of variables, attributes and business time. Attributes are created by selecting attributes with which the number of taco restaurants is more than 100. There are two variables in business time, one indicates whether the taco restaurant is open on weekends, and the other shows the total weekly business hours. We exclude any restaurants with one or more NA of these two features. For ordered categorical values with different n levels, we encode them into n integers. For unordered categorical values, we use one-hot encoding. For other attributes with bool values, we keep them.

Before fitting a linear regression model, we used mode to impute the missing values of each variable. We would like to contain no more than 5 variables in our regression model for the simplicity of explanation, so we choose to use all subsets regression to find the best set of variables for each model size$(1-5)$. for each of them, we calculated three metrics: Adjusted R-Squared, Mallows Cp, and BIC to select the best overall model. The final model contains 5 variables.

From Table 1, we can see that the proportion of good ratings(larger than 3) of taco restaurants with services *Outdoor Seating*, *Bike Parking*, *Restaurants TakeOut*, and *Weekends* is larger than that of taco restaurants which do not provide certain services.

2.2 Reviews Data

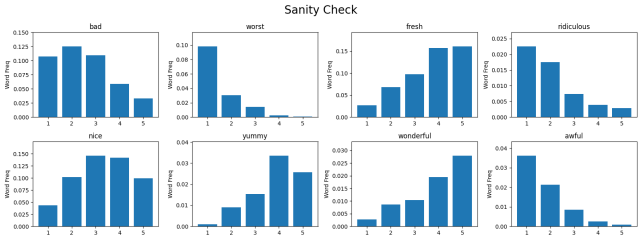
For the sentiment analysis on the reviews of taco restaurants, we included the text reviews and the rating stars. The reviews are long strings containing comments from the Yelp app

	Outdoor Seating	Bike Parking	Restaurants TakeOut	Restaurants Delivery	Weekends
0 (*Not Offering*)	0.365	0.389	0.175	0.962	0.542
1 (*Offering*)	0.886	0.720	0.654	0.471	0.963

Table 1: Proportion of Ratings Larger Than 3 of 5 Variables

user with a rating star from 1 to 5. To study the reviewer's sentiment, we first standardized the rating star per user. We computed the z-score and use the percentile of the z-score to get the new rating stars.

When we can not get all five percentile parts, for example, the user actually rate only one type of stars, we view the user doesn't have any preference so we standardized all the rating star to 3 as the middle situation. Similarly, we deal with only two types rating stars situation as dividing them into good and bad parts with 2 and 4 ratings. The other situations can be standardized similarly. We create new word features the standardized rating stars by using the above-mentioned process. We also create a list of the words of the dish and sauce taco bell's menu[1].



(a) Proportion of words by rating of reviews

From the above figures we can see that sentiment words like bad and worst are reasonably associated with the rating ($p-value \leq 0.05$ with chi-square test) As expected, positive words are positive ratings and negative words are negatively correlated with negative ratings.

3 Models and Findings

3.1 Linear Regression Analysis

$$Ratings = \alpha + \beta_1 X_{BikeParking} + \beta_2 X_{TakeOut} + \beta_3 X_{Delivery} + \beta_4 X_{OutdoorSeating} + \beta_5 X_{Weekend}$$

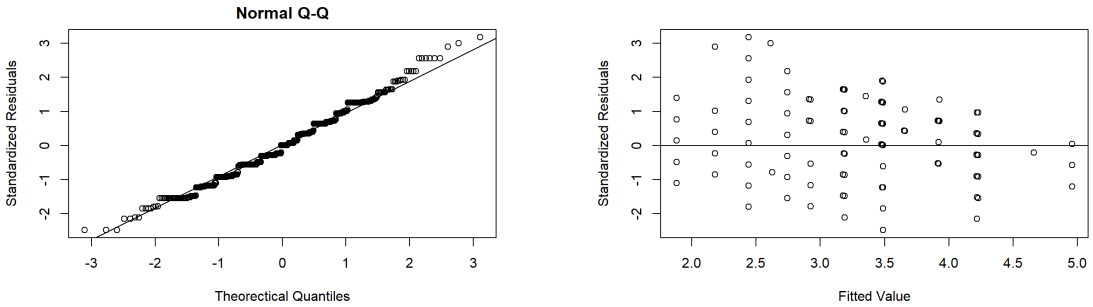
From the results from the regression model, we can see that the *F-statistic* is very large and the *p-values* of 5 variables are all basically zeros. So we rejected the null hypothesis and concluded that there is strong evidence that a relationship does exist between ratings and each of the 5 variables. Besides the diagnostics plots, which showed that our model satisfies the assumption of the linear model, we also calculated the VIF(variance inflation factor) to check if our model suffers from severe multicollinearity. The result showed that our VIF is less than $\frac{1}{1-R^2}$ which indicates that our model is fine with multicollinearity.

3.2 Sentiment Analysis

[1]https://www.tacobell.com/

	Bike Parking	Restaurants TakeOut	Restaurants Delivery	Outdoor Seating	Weekends	F-statistic
p-value	0.000131	1.09e-07	4.91e-15	<2e-16	9.28e-12	<2.2e-16
estimation	0.299	0.561	-0.731	0.745	0.739	93.95

Table 2: Regression Results



(b) Q-Q plot (c) Standardized Residuals vs Fitted values

Figure 1: Model Diagnostic

We first assigned -2 to 2 scores for 1 star to 5 star and computed average scores for each sentimental word (e.g. good, bad) among all the reviews that they appear. With these scores of the sentimental words, we considered computing the score of words from four main aspects for each restaurant: dish, sauce, service, and environment. We searched times that phrase (e.g. shrimp+good, shrimp+bad) appears in each sentence of the reviews and averaged the score by the number of reviews. Finally, with these words' scores, we could get the average scores of the four main aspects of each restaurant.

	Dish	Sauce	Service	Environment
Score	0.00420	0.00662	0.0195	0.0264

Table 3: Scores of main aspects of all taco restaurants

Table 3 shows the scores of the main aspects of all taco restaurants. The analysis shows that generally, taco restaurant has the lowest sentiment score for the *dish* and the highest for *environment*. We can then praise the *environment* and make suggestions for the *dish*.

At the same time, sentiment analysis also has some limitations. We need to prepare more words to analyze the comments comprehensively and find the scores of aspects we are interested in. However, we only used words about part of the aspects, which led to the fact that the sentiment score could not fully explain the star ratings, and there was some bias.

4 Recommendations for Businesses

Based on the results from linear regression, we can see that We recommend people who would like to run a taco business provide *Outdoor Seating*, *Bike Parking*, and *Restaurants TakeOut* services to customers and run a business on *Weekends*. Also, do not provide food delivery for the sake of getting higher ratings on yelp. The limitation of this recommendation is that the $R - Squared$ of our model is 0.46 indicating that only 46% variation of ratings can be explained by these 5 variables.

Then we mainly consider giving suggestions on the effect of these words on star ratings. We decided to use the sentiment score to measure it. In a taco business review, the higher the sentiment score of a word, the more significant the positive impact of that word on star ratings. Also, it shows that this business is doing well in this aspect and has received good reviews from most customers.

Therefore, the sentiment score also shows customers' attitudes to a certain extent. In other words, we can use the sentiment score to find out the strengths and weaknesses of the restaurant and give suggestions accordingly. For example, we can incorporate the aspect with the lowest sentiment score into the disadvantages of the restaurant and provide more targeted suggestions based on the sentiment score of the word in this area.

At the same time, we can also propose targeted solutions based on the sentiment scores of all words in this area. For example, in the analysis of a restaurant, we found the word *crowded* in *environment* had the lowest sentiment score, then we can make suggestions for *crowded*.

Due to the reasons above, we mainly propose dynamic suggestions based on sentiment analysis results. We consider using the template prepared in advance and filling in the corresponding words to complete the recommendation. Suppose for a certain restaurant, we already know that we want to praise its *service* and make recommendations for the *environment*. So the suggestions we get might be "*Your service improves your ratings a lot.*" and "*Too many complaints about the environment.*" The words "service" and "environment" in these two comments are the words we fill in according to the sentiment analysis results.

To be more specific, we want to make suggestions about the *dirty* in the *environment*, so we can get the suggestion "*Some guests are complaining about the dirty environment.*", here we fill in "dirty" and "environment."

Of course, our recommendation also has some limitations. Firstly, we can only give advice based on the four major categories (Dish Sauce Service Environment) we identified due to the number of words. Therefore, our recommendation can only tell the restaurant which of these four aspects is the best and which is the worst. However, for specific advice, we can only give them based on the sentiment score of the words we prepared.

Also, given the scope of the analysis, we may miss some of the restaurant's fatal strengths and weaknesses. For example, some merchants do well in all four areas (relatively high sentiment score) but have moderate star ratings. Our current model cannot explain this.

In addition, we are only comparing the four aspects of a single merchant, not with other merchants around them. Therefore, we can only draw a relative conclusion. After the restaurant makes improvements in this area based on our suggestions, the star ratings may stay the same because the restaurant still needs to be better compared to surrounding taco restaurants.

5 Conclusion

To sum up, our analysis mainly focuses on two datasets: business data and review data of taco restaurants. And based on the results of our analysis, we can give recommendations in two aspects.

6 Contribution

Name	Contribution
Yuanhao Geng	web-based application, data cleaning part of report and presentation
Xuelan Qian	data cleaning code, intruduction part of report and presentation
Zhanpeng Xu	sentiment analysis code, presentation and report
Xinyu Li	linear regression code, presentation and report