BUSINESS ANALYTICS PROJECT

Analysis of factors influencing Yelp ratings of the NYC restaurants

GROUP: 11

Aleksandr Kniaginin

Elizaveta Demyanenko

Qitian Ma

Weiting Ye

Contents

1 Introduction

In the digital age, high Yelp rating of a restaurant becomes a competitive advantage in attracting new visitors. A variety of factors might influence ratings that visitors assign to the restaurants, and the knowledge of the scale and the direction of these effects might prove useful to the managers who are interested in getting outstanding ratings compared to their competitors. Some of these factors can be scraped straight from the Yelp pages, whereas such characteristics as overall quality of service and food are not that easy to get and must be extracted from the Yelp reviews. In our project we mainly focus on investigating whether violation of sanitary standards exists among these factors.

1.1 Data

In our project we use the following data:

- Dataset from New York City Department of Health and Mental Hygiene which contains information on hygiene violations of the NYC restaurants

	CAMIS	DBA	BORO	BUILDING	STREET	ZIPCODE	PHONE	CUISINE DESCRIPTION	INSPECTION DATE	ACTION	VIOLATION CODE	VIOLATION DESCRIPTION
0	41127209	BACI & ABBRACCI	BROOKLYN	204	GRAND STREET	11211.0	7185996599	Italian	03/09/2018	Violations were cited in the following area(s).	04H	Raw, cooked or prepared food is adulterated, c...
1	50002502	LE REVE	MANHATTAN	125	E 54 ST	10022.0	2127597777	Middle Eastern	02/24/2017	Violations were cited in the following area(s).	02G	Cold food item held above 41Â° F (smoked fish ...
2	50038792	BEST BURGER PALACE	BROOKLYN	986	ATLANTIC AVE	11238.0	7183983130	American	05/03/2017	Violations were cited in the following area(s).	06D	Food contact surface not properly washed, rins...
3	50035589	DAE SONG CHINESE RESAURANT	QUEENS	4332	CORPORAL KENNEDY ST	11361.0	7182292279	Chinese	04/09/2018	Violations were cited in the following area(s).	06C	Food not protected from potential source of co...
4	50018594	THE SPOTTED OWL	MANHATTAN	211	AVENUE A	10009.0	2123889844	American	02/08/2016	Violations were cited in the following area(s).	04L	Evidence of mice or live mice present in facil...

Figure 1: Sanitary violations data

- Reviews for these restaurants from Yelp.com

For each restaurant we scrape at most 200 reviews which are transformed into sentiments and then averaged to get the sentiment score. Both datasets have undergone some preprocessing

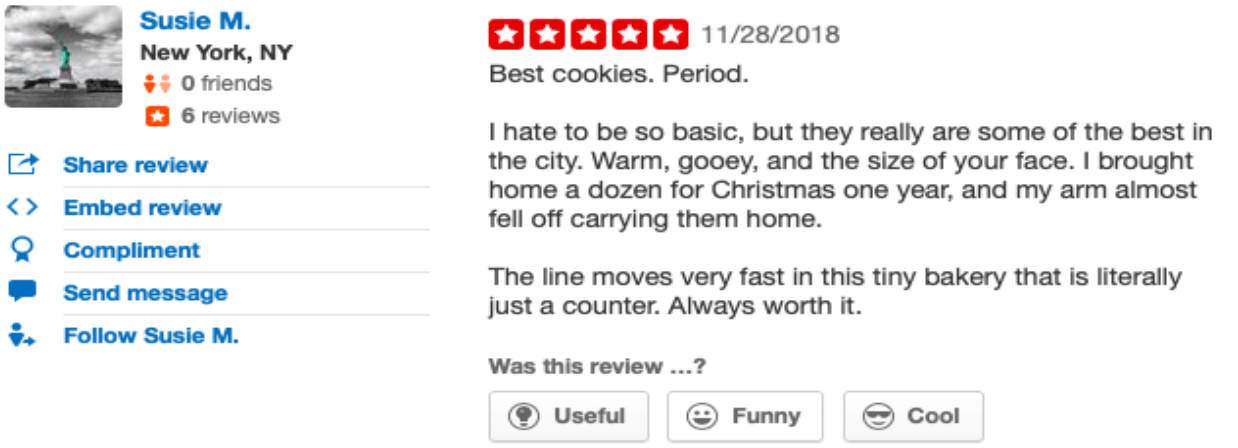

Figure 2: Example of Yelp review

such as deletion of observations with too many NA values or feature selection. To reduce the amount of data for scraping and make our data more homogenous, we constrain ourselves to the restaurants with American cuisine in Manhattan. Therefore it restricts our findings only to this type of restaurants. Our preparation of the data consists of the following main steps:

- Cleaning and reducing the dataset from New York City Department of Health and Mental Hygiene
- Scraping the reviews and descriptive variables of the chosen restaurants from Yelp.com
- Concatenation of the obtained datasets

We assume that customers use Yelp rating in order to choose the restaurant among already predetermined cohort (among expensive restaurants, cafes, canteens etc.). It means that in order to make a good managerial decision owner/manager of the restaurant should care only about how to stand out among its direct competitors.

2 Theory behind the analysis

In our project, we apply three Natural Language Processing (NLP) techniques.

2.1 Levenshtein distance

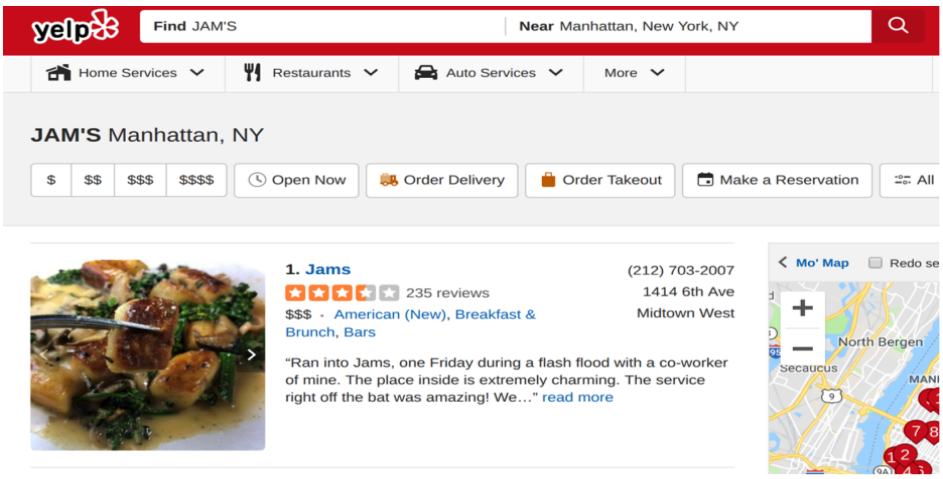

Figure 3: Actual name of the restaurant and its name on Yelp page

Levenshtein Distance is a measure of distance between two strings by insertion, deletion and substitution. This measure is implemented in Python package FuzzyWuzzy and normalized by the length of two strings. We use Levenshtein Distance to fuzzy-match the precise name of a restaurant (in government dataset) and its name on Yelp during the web scraping process. The exact process is:

- Fetch the search page
- Scan through the search page and fetch the restaurant with the largest similarity index After viewing the fetched data, we decided to drop observations with similarity index below 60

2.2 Sentiment analysis

Sentiment Analysis involves two major families: 1. machine learning 2 lexicon. Here we apply lexicon-based techniques – VADER, due to its specialization in social media context and our lack of labeled training data. VADER is implemented in Python package NLTK. We use VADER to compute the average sentiment score of a restaurant's reviews for later regression analysis.

2.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation is a method to extract topics from documents. In the given text corpus, it detects main topics which we can further interpret through the most relevant words for each topic. Here are the clusters that we got:

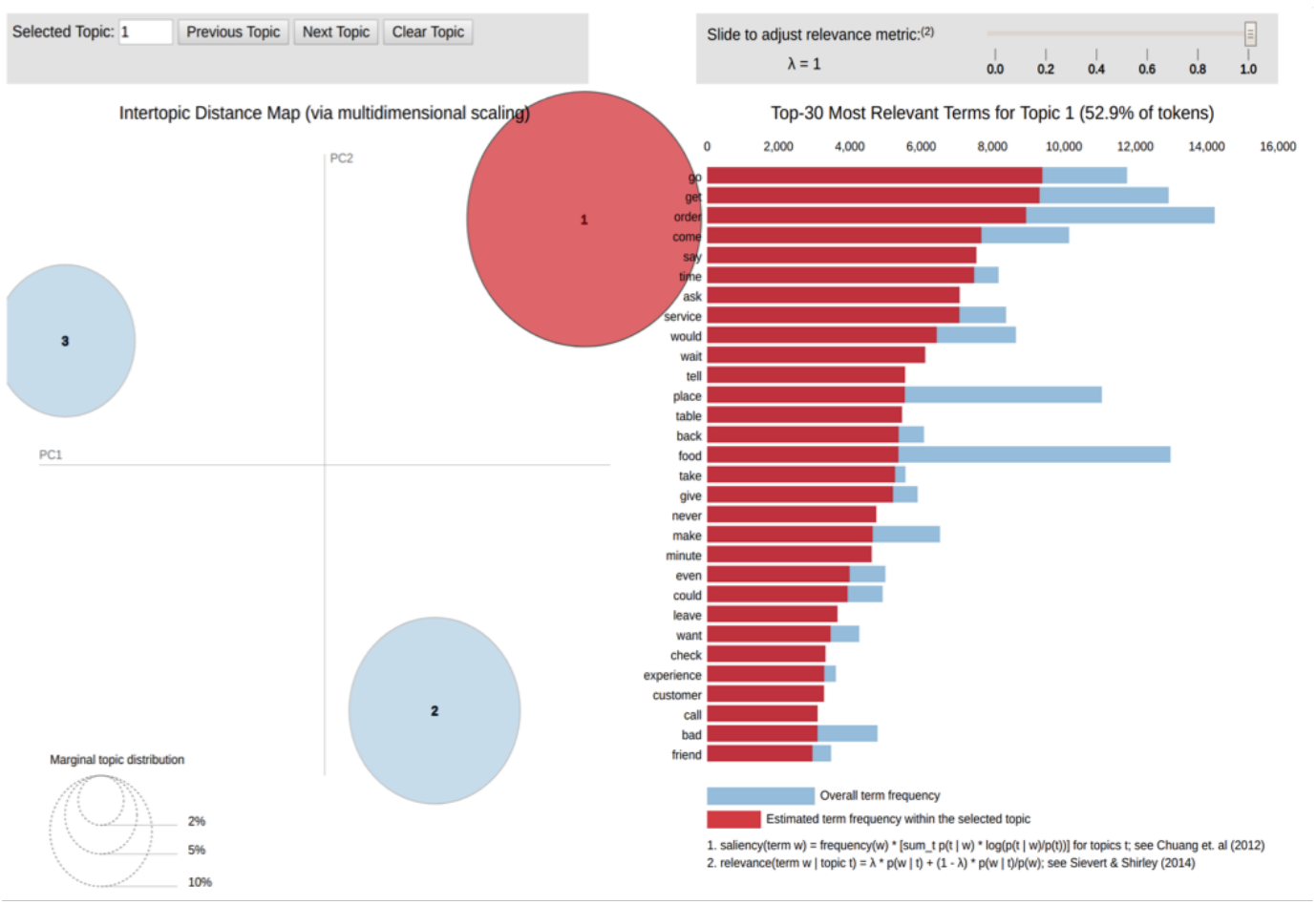

Figure 4: LDA Service

From the pictures we can see that visitors first of all pay attention to the service, and then to the food. Hygiene is a rarely mentioned topic in reviews compared to the previous two. Therefore, customers are unlikely to be aware of potential violations.

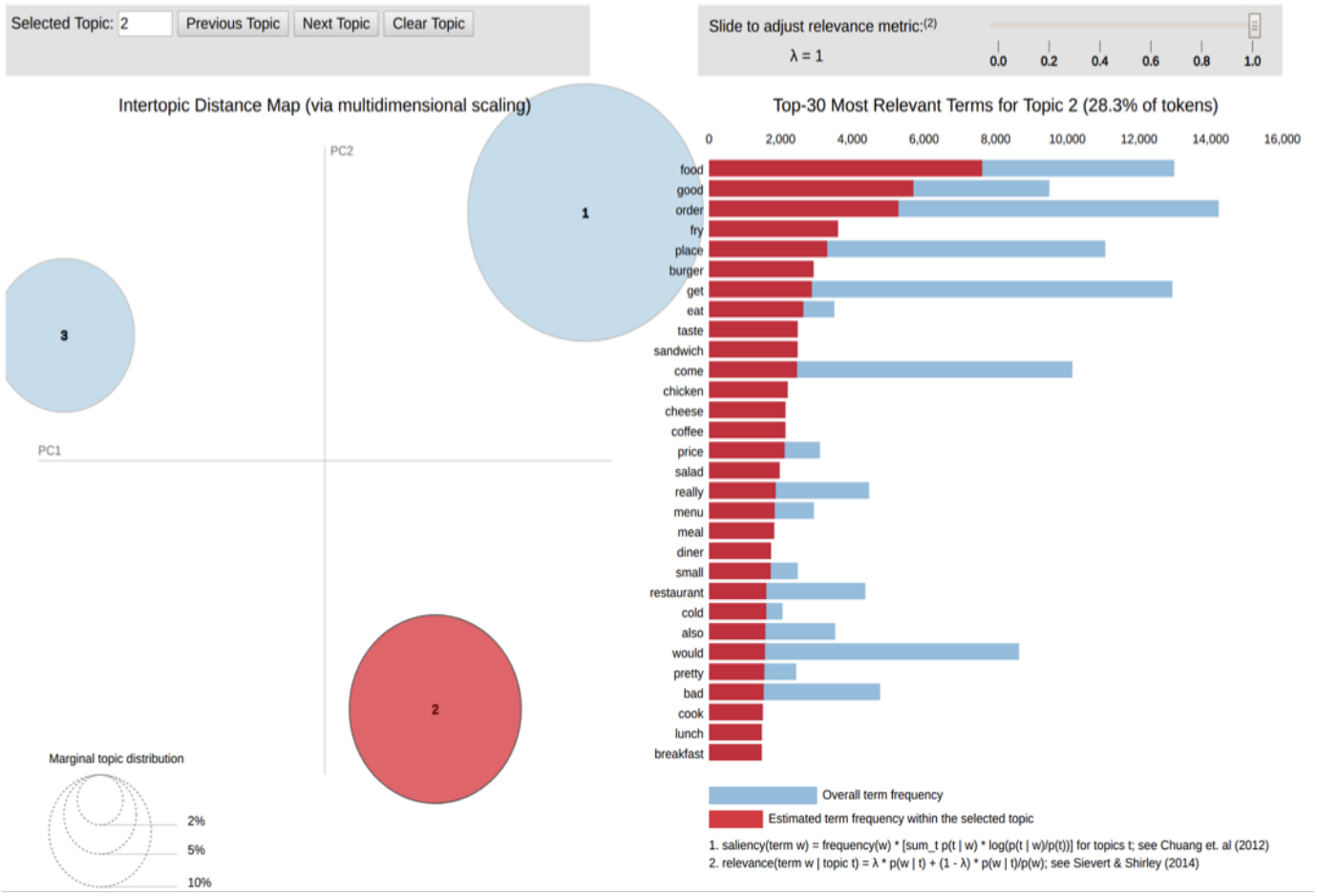

Figure 5: LDA Food

2.4 Ordered logit

As we stated in the introduction, the absolute value of the Yelp rating is not of importance for the restaurants: what we are interested in is the value of the rating compared to the values of the ratings of competitors. Therefore, we should account for the different restaurant markets and try to explain the variation within these groups of restaurants i.e. separate common and individual effects. We already have a rather homogenous sample: all the places belong to the same geographical market and type of cuisine. However, we still have a variating price range which means that not all of these restaurants actually compete with each other. To address this, we set price range as the panel variable, thus accounting for individual effects of the ratings in the different price categories.

If we run the regression of overall rating on violation score, we will most definitely face the omitted variables issue that results in the endogeneity problem. Hence, we add controls, among which the most important is the sentiment score which is a proxy for the quality of food and service compared with prices. Other independent variables are categorical and

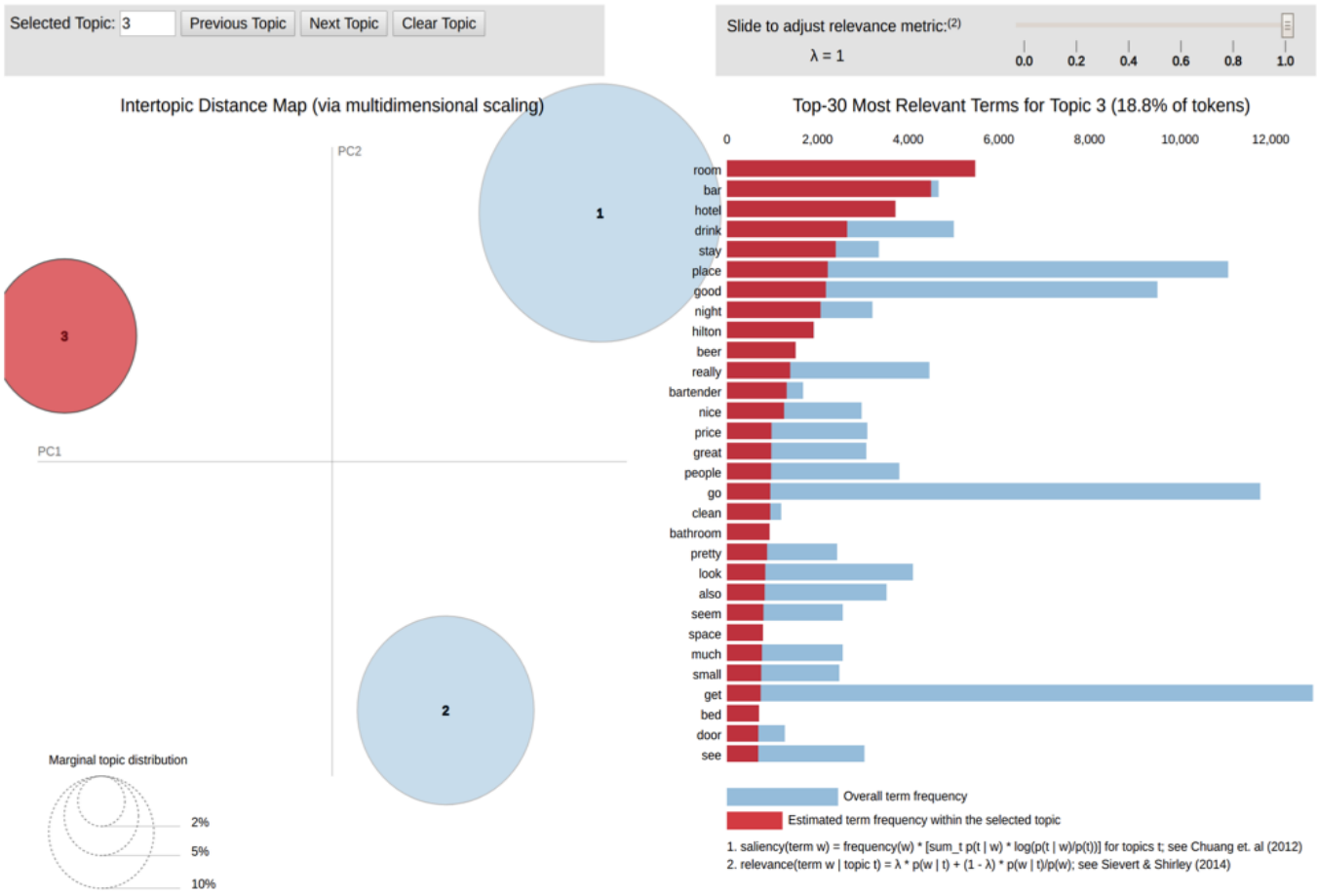

Figure 6: LDA Facilities

include Alcohol (No, Beer & Wine Only, Full Bar), Provision of Catering, Delivery, Parking, Bike Parking, Takeout, Presence of TV, Outdoor Seating, Taking Reservations, Presence of Wi-Fi, Noise Level (Quiet, Average, Loud, Very Loud).

Our dependent variable is Overall Rating, which takes values from the set 2, 2.5, 3, 3.5, 4, 4.5, 5. Since values of the dependent variable are discrete and ordered, we use ordered logit model. We chose random effects specification, because ordered logit with fixed effects needs a high number of assumptions and has no implementation in Stata. After running the xtologit Stata command with standard errors adjusted by the price range we obtain following results (figure 7) :

(Std. Err. adjusted for 3 clusters in prices2)

overall_rating	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
sentiment_score	21.50517	2.832027	7.59	0.000	15.9545	27.05585
violation score	.0039728	.0168465	0.24	0.814	-.0290458	.0369915
alcohol2						
Beer & Wine Only	-.0739764	.2500878	-0.30	0.767	-.5641394	.4161867
Full Bar	-.8236885	.2077385	-3.97	0.000	-1.230848	-.4165286
cater2						
Yes	-.3729586	.0364323	-10.24	0.000	-.4443645	-.3015526
delivery2						
Yes	-.0052882	.1292056	-0.04	0.967	-.2585265	.2479501
has tv2						
Yes	-.29883	.0355385	-8.41	0.000	-.3684842	-.2291758
outdoor seating2						
Yes	-.211635	.1084142	-1.95	0.051	-.424123	.0008529
parking2						
Yes	.0922891	.2206973	0.42	0.676	-.3402697	.5248478
bike_parking2						
Yes	-.069037	.1720047	-0.40	0.688	-.40616	.2680861
takeout2						
Yes	.0206099	.0496196	0.42	0.678	-.0766428	.1178625
takes reservations2						
Yes	-.8780424	.0317758	-27.63	0.000	-.9403218	-.8157629
wifi2						
Yes	.1746699	.0651707	2.68	0.007	.0469377	.3024021
noise_level2						
Quiet	.7489772	.4391724	1.71	0.088	-.1117849	1.609739
Loud	-.0027065	.4623406	-0.01	0.995	-.9088774	.9034644
Very Loud	-.1931908	.414276	-0.47	0.641	-1.005157	.6187752
/cut1	-4.127042	1.292528			-6.66035	-1.593734
/cut2	2.129842	.6003732			.9531326	3.306552
/cut3	6.510785	.7187632			5.102035	7.919535
/cut4	10.20456	1.196201			7.860051	12.54907
/cut5	14.08378	1.492349			11.15883	17.00874
/cut6	17.9769	1.857725			14.33582	21.61797
/cut7	20.34791	2.054265			16.32163	24.3742
/sigma2_u	.1792502	.0180682			.147116	.2184035

Figure 7: Ordered logit results

3 Interpretation of the results

Main result: we cannot reject the hypothesis of zero influence of violations on the Yelp rating in favor of the hypothesis of negative influence.

- One of the explanations is that violations occurring usually are not really noticed by the visitors, which is supported by the analysis of texts of reviews. This can be due to visitors not paying attention to such things or due to violations taking place somewhere

where visitors do not have access to.

- Or this can be also a consequence of more popular restaurants receiving more attention from the violation committee. As we saw from the exploratory data analysis, violation score increases with the price range just as the overall rating does

4 Conclusion

We cannot recommend elimination of violations as a means of increasing the Yelp rating and, consequently, attracting new customers. However, we can provide some possible managerial insights from the presence of significant factors in our regression:

- Visitors tend to give higher ratings to the quiet places, while there is no significant difference between ratings of places with the level of noise that is average or higher (ceteris paribus). This means that, if the manager of a place with an average level of noise knows of a way to decrease level of noise below average at some low cost, he should be advised to do so.

- Installing Wi-Fi or TV in the restaurant is an investment that is done to increase visitors' satisfaction with the place. And since places that have a TV all other things equal receive lower ratings that places that don't have a TV, we can suggest that this investment is probably a waste of resources. Installing Wi-Fi, in turn, might be a useful investment, since it is tied to higher ratings.

- One of the most robust results is the negative influence of the fact of taking reservations on rating, which suggests that this is not the best business decision, especially when we know that even in the highest price range only around 70% of restaurants take reservations. A possible explanation is that visitors might set their expectations too high when they are obliged to book in advance. Though here also can exist an omitted variables problem and it requires more thorough analysis to claim something more than correlation between these two variables.

4.1 Future improvements

It should be noted that all the found relations are to be studied in more detail, because these categorical features might, in fact, indicate a particular type of restaurant for the regression rather than carry some influence on score in itself. And as it was already mentioned, our constrain on the specific region and cuisine imposes constraints on the generalization of our results to the whole types of the restaurants. Therefore, there is still a lot of space for broadening of the analysis.