Quantitative Research: How Price, Reviews and Rating from TripAdvisor Affect the Restaurant Ranking in 31 European Cities

Student number: 20044050

1 Introduction

This paper aims to explore the information about restaurants for 31 European cities which is listed in TripAdvisor.com, a famous user-generated content (UGC) rating platform. The research question is how do rating, number of reviews and price range change can improve the restaurant ranking. After exploring key fields correlation analysis, the hypothesis will be demonstrated that rating and low price can positively affect the restaurant ranking. It will be revised after multiple regression modelling. The key finding is that ranking is highly positively related to the rating while different price can lead to different ranking change.

2 Literature Review

The potential commercial value of UGC has been widely recognised. Cao Duan's research team has conducted data mining on the UGC review data set in hotel review sites and established five-dimensional evaluation indicators to measure the quality of hotel services, which is conducive to the development of tourism industry (Duan et al., 2013). Simultaneously, research also evaluated a sentiment analysis-based decision model based on sentiment analysis in a restaurant selection case study using TripAdvisor reviews and helped build a restaurant review data set (Zuheros et al., 2021), which is of great significance for the behaviour research of UGC platform users.

In order to study the characteristics of TripAdvisor indicators such as reviews rating. In the research of review websites in the travel industry, we found that a research team analysed travellers' rating patterns between independent and chain hotels. They found that the rating patterns of hotels in different regions changed with the user's profile and region (Banerjee, 2016). This also inspired me to analyse by ranking and different variables. In addition, similar research has been done on Yelp website reviews. The difference is that, by regression analysis of reviews, researchers find restaurants revenues can grow nearly 10% when there is a one-star increase in their Yelp listing rating (Luca, 2011).

3 Data

This dataset can be accessed in Kaggle, a data science platform (the link will be given in Appendix section). The details about restaurants in a city have been obtained through scraping TripAdvisor.com (a famous tourism website). After cleaning and pre-processing this dataset, the final version TripAdvisor restaurant dataset in 31 European cities has been formed.

3.1 Key field

In this dataset, “Ranking” is the rank of the restaurant among the total number of restaurants in the city and “Rating” is on behalf of the rate of the restaurant on a scale from 1 to 5. The nominal field “Price Range” is the price range of the restaurant among 3 categories such as “$$ - $$$”. Finally, “Number of Reviews” is the number of reviews that customers have let to the restaurant listed in TripAdvisor. As shown in figure 1, there are four nominal and three ratio data in dataset.

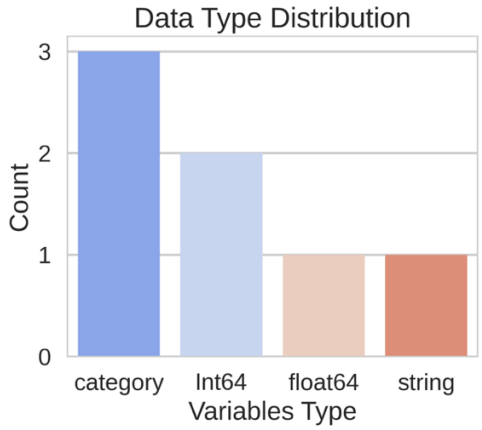

Figure 1. The distribution of variables types

3.2 Numeric Data Summary

It is appropriate to explore the descriptive statistics of numeric data (Ranking, Rating, and Number of Reviews). As presented in table 1, the max value of rating is close to the mean number, but the verse situation is seen in Number of Reviews. The mean value of it is much lower than the max value and standard deviation shows the great difference in Number of Reviews distribution.

Table 1. Summary of numeric data

	Ranking	Rating	Number of Reviews
count	74225	74225	74225
mean	2980.30	4.02	175.99
std	3372.20	0.55	362.27
min	1	1	2
50%	1674	4	68
max	16443	5	16478

Investigating the frequency of quantities of reviews, as shown in figure 2, it is obvious to see the number between 0-1000 appear frequently. In contrast, the restaurant who own a large number of reviews makes up for a small proportion. They can be seen as outliers, which means the majority of restaurants' comments number is lower than 1000.

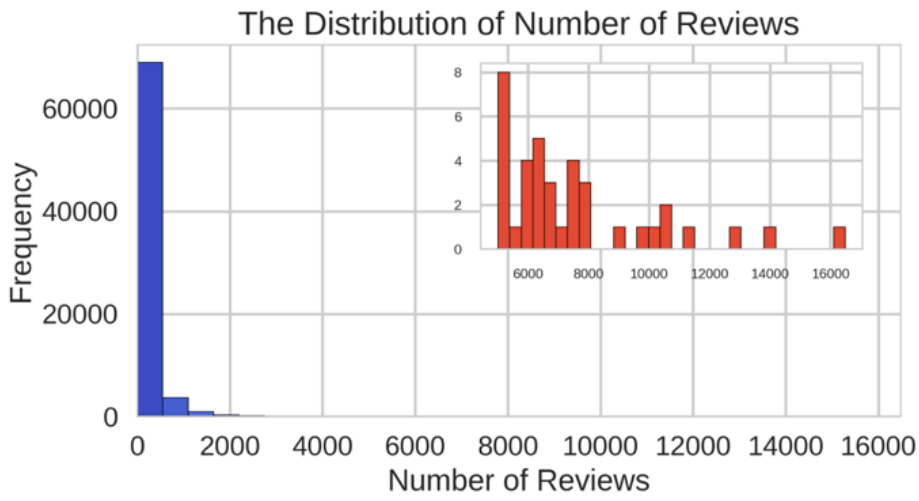

Figure 2. The distribution of number of reviews

As shown in figure 3, the distribution pattern of restaurant ranking is similar to that of rating. Restaurant ranking between 0-2500 or rating between 4.0-5.0 accounts for the largest part in this TripAdvisor dataset.

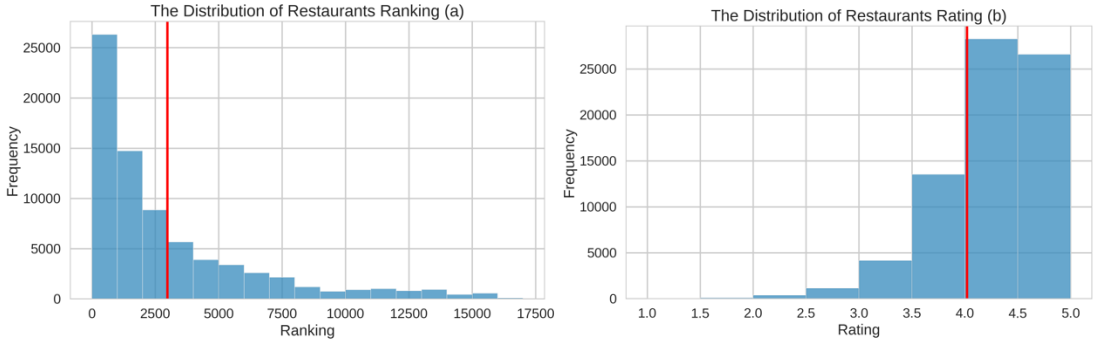

Figure 3. The distribution of restaurants ranking and rating

4 Methodology

The research clean and process the dataset in the beginning. For example, the categorical variable “Price Range” which has three categories is converted into three dummy variables (“low”, “medium”, “high”). This process is called “encoding of categorical variables”. Then the correlation analysis is applied in key fields exploration, which contributes to the liner regression.

Before building a liner regression model, the correlation analysis is applied in all numeric variables including processed variable “Price Range” to explore the relationship between key fields and find multicollinearity. Once the high correlation (above 0.8) between two variables has been found, one of them will be removed.

The ordinary least squares (OLS) method is utilised to estimate the unknown parameters in a linear regression model. The reason why the dependent variable

"Ranking" is processed in logarithm is that, as shown in the above third section, it follows a logarithmic distribution pattern. The formula that this research assumed to fit the Log-linear model is given below.

$$\log Ranking = \beta_1 \times Rating + \beta_2 \times Reviews + \beta_3 \times Price$$

In the log-linear model, the *Ranking* is the rank of the restaurant among the total number of restaurants in the city. *Rating* is the rate of the restaurant on a scale from 1 to 5 (At 0.5 intervals). *Reviews* is the number of reviews that customers have let to the restaurant listed in TripAdvisor. The dummy variables *Price* is the price range of the restaurant among 3 categories.

After the liner regressing, the F-test will be used to test the multiple regression model and T-test will be used to test every single variable regression results.

5 Result

5.1 Correlation Analysis

As shown in the figure 4, there is a high correlation can be found between Price_low and Price_medium, which means one of them might lead to multicollinearity in the following regressing analysis. In the next step, "Price_medium" is removed because its variance inflation factor is larger than "Price_low".

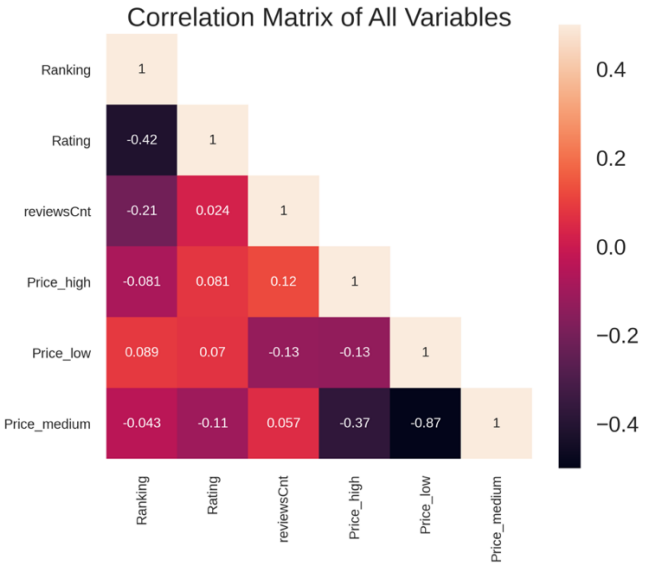

Figure 4. The correlation of ranking, rating, rumber of reviews and price range

Through the controlled variable method, a certain number of independent variables are selected for multiple regression analysis each time. Then, the goodness of the regression is tested and recorded in the table for comparison statistics. As shown in Table 2, it is found in the research that adding all independent variables to the regression model has the best fit.

Table 2. Multiple regression test statistics by controlling independences

Regression Test No.	Rating	reviewsCnt	Price Range (high, low)	R-squared
1	✓	✓	✓	**0.932**
2	✓	✓	-	0.931
3	✓	-	✓	0.929
4	-	✓	✓	0.367

***** The independents variables for each regression test are ticked by "✓"

5.2 Liner Regression Analysis

The hypothesis is true. All dependent values are included in the liner model, as shown in figure 5. R-squared value is 0.932, which demonstrated that the goodness of regression is high.

OLS Regression Results

Dep. Variable:	Ranking	**R-squared (uncentered):**	0.932
Model:	OLS	**Adj. R-squared (uncentered):**	0.932
Method:	Least Squares	**F-statistic:**	2.545e+05
Date:	Tue, 19 Jan 2021	**Prob (F-statistic):**	0.00
Time:	08:23:21	**Log-Likelihood:**	-1.5428e+05
No. Observations:	74225	**AIC:**	3.086e+05
Df Residuals:	74221	**BIC:**	3.086e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Rating	1.7977	0.002	785.813	0.000	1.793	1.802
reviewsCnt	-0.0011	1.99e-05	-53.140	0.000	-0.001	-0.001
Price_high	-0.7558	0.032	-23.828	0.000	-0.818	-0.694
Price_low	0.3292	0.017	19.370	0.000	0.296	0.363

Omnibus:	512.149	**Durbin-Watson:**	0.417
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	727.820
Skew:	0.077	**Prob(JB):**	9.03e-159
Kurtosis:	3.460	**Cond. No.**	1.81e+03

Figure 5. Ranking with all dependent values regression result

The result of log-liner model is given below:

$$\log Ranking = 1.800 \times Rating - 0.001 \times Reviews - 0.756 \times Price\_High + 0.329 \times Price\_Low$$

Each 1-unit increase in rating (one-star increase) multiplies the expected value of Ranking by $e^{1.8}$; For number of reviews, $e^{0.001}$is quite small (approximately $e^{0.001} = 1.001$). One-unit growth in number of reviews led to a decrease in ranking value of 1.001. The positive increase in *Price_High* value will cause a downward change in ranking value but *Price_Low* will result in an increase in dependent value.

6 Discussion

Low correlation among key field can be found. In the correlation analysis, in addition to price range, it can be found that the correlation between key fields is generally low. For example, the number of reviews does not affect the rating and the price will not affect the quality of reviews. It can be seen that the various data of restaurants in TripAdvisor can be used as relatively independent variables that are better for Ranking.

The high rating can improve the ranking of restaurants. The evidence is that the values

of rating and the logarithm of ranking in the final regression model result are positively correlated, it can be obtained that the relationship between rating and ranking value is negative, which means the more rating is, the higher the ranking is. If a restaurant wants to improve its ranking, it should pay attention to the rate. As mentioned in the TripAdvisor website (2020), the rate is related to the quality and recency of the reviews, restaurants should offer up-to-date information to cater to consumers better, which helps to increase their exposure in the city.

On the contrary, the number of reviews may have a negative effect on restaurant rankings. As shown in result sections, an increase in the number of reviews will cause the restaurant's ranking to drop slightly. The reason may be that the more guests received, the more customers the restaurant cannot satisfy. Dissatisfied customers' own experience published on TripAdvisor may affect the overall ranking of the store. It can be seen that a restaurant should not try to obtain as many reviews as possible. Other dimensions of reviews are more worthy of consideration.

Regarding restaurant prices, higher prices may not lead to better rankings, while lower prices could improve rankings in the results. This is likely to be related to the psychological expectations of customers. According to some research, it can be found that service quality of budget hotel market was also related to customer expectation (Hua, 2009). Higher restaurant prices may bring high expectations to customers. The opposite situation is that low prices can reduce consumers' expectations of restaurant service quality. Then the positive sentiment generated will be reflected in the review scores, leading to higher rankings.

7 Conclusion

This report explores the key fields in the TripAdvisor restaurants in 31 European cities and answers the research question. For a restaurant on TripAdvisor, to improve ranking, it is suggested to concern their rating instead of a number of reviews and slightly reduce the expectation of consumers in the aspect of menu pricing. Regarding the limitation of this research, the residual analysis is not included and there are likely to be some variables which can affect the ranking such as cuisine style. According to some research that Naïve Bayes was applied in reviews on TripAdvisor (Laksono,2019). This research can be extended by natural language processing (NLP) with sentiment analysis in the further.

Word Count: 1,741

References

Banerjee, S. and Chua, A., 2016. In search of patterns among travellers' hotel ratings in TripAdvisor. Tourism Management, 53, pp.125-131.

Duan, W., Cao, Q., Yu, Y. and Levy, S., 2013. Mining Online User-Generated Content: Using Sentiment Analysis Technique to Study Hotel Service Quality. 2013 46th Hawaii International Conference on System Sciences.

Hua, Wen., Chan, Andrew., & Mao, Zhenxing., 2009. Critical Success Factors and Customer Expectation in Budget Hotel Segment — A Case Study of China. Journal of Quality Assurance in Hospitality & Tourism, 10(1), pp. 59-74.

Luca, M., 2015. User-Generated Content and Social Media. Handbook of Media Economics, pp.563-592.

Luca, M., 2011. Reviews, Reputation, and Revenue: The Case of Yelp.Com. SSRN Electronic Journal.

Parikh, A., Behnke, C., Vorvoreanu, M., Almanza, B. and Nelson, D., 2014. Motives for reading and articulating user-generated restaurant reviews on Yelp.com. Journal of Hospitality and Tourism Technology, 5(2), pp.160-176.

Laksono, R. A., Sungkono, K. R. Sarno, R. and Wahyuni, C. S., 2019. Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naïve Bayes. International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, pp. 49-54.

TripAdvisor. 2021. Tripadvisor Popularity Ranking: Key Factors And How To Improve. [online] Available at: <https://www.tripadvisor.com/TripAdvisorInsights/w722> [Accessed 12 January 2021].

Zuheros, C., Martínez-Cámara, E., Herrera-Viedma, E. and Herrera, F., 2021. Sentiment Analysis based Multi-Person Multi-criteria Decision Making methodology using natural language processing and deep learning for smarter decision aid. Case study of restaurant choice using TripAdvisor reviews. Information Fusion, 68, pp.22-36.

Appendix

Github: https://github.com/Hereislittlemushroom/CASA0007_Final_Assessment
Dataset: https://www.kaggle.com/damienbeneschi/krakow-ta-restaurans-data-raw