**Quantitative Research: How Price, Reviews and Rating from TripAdvisor Affect the Restaurant Ranking in London and 30 European Cities**

MSc Spatial Data Science and Visualisation

The Bartlett Centre for Advanced Spatial Analysis

Zeqiang Fang

## Introduction

With the development of rating APP,

Market

Yelp, TripAdvisor, Booking

Travel industry

User-generated content

Research question

This paper aim to … How Price, Reviews and Rating from TripAdvisor Affect the Restaurant Ranking in London and 30 European Cities? Data presentation, Methodology, result, discuss, key findings and evidence. Finally, result will be tested

## Literature Review

Restaurant Ranking evaluate – Yelp, TripAvisor

Ranking Rate Analysis

## Data

Where data comes from?

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). The datasets for the main European cities have been merged. After cleaning and pre-processing this dataset, the final version TripAdvisor restaurant dataset has been formed.

Limitation of Data

The data may not perfectly represent all the restaurants in these European cities because it merely includes the restaurants that are listed in the TripAdvisor.com database rather than all the restaurants in these areas.

To run a smooth analysis, some columns’ names and data types are changed(such as “Number of Reviews”becoming “reviewsCnt”). Besides, the null values also be cleaned and there are 74225 rows and 7 columns.

Key field

City: city location of the restaurant

Cuisine Style: cuisine style(s) of the restaurant

Ranking: rank of the restaurant among the total number of restaurants in the city

Rating: rate of the restaurant on a scale from 1 to 5

Price Range: price range of the restaurant among 3 categories, as a categorical type

Number of Reviews: number of reviews that customers have let to the restaurant, as a float object (108 020 non-null)

Data Type Statistics

As shown in figure 1, there are four nominal and three ratio data in dataset.

图表, 条形图

描述已自动生成

Figure 1. The distribution of variables types

Numeric Data Summary

As presented in the table 1, rating: max is close to mean

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 |
| 25% | 653 | 3.5 | 25 |
| 50% | 1674 | 4 | 68 |
| 75% | 4031 | 4.5 | 182 |
| 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, restaurant who own the large number of reviews make up for a small proportion. They can be seen as outliers, which means

图表

描述已自动生成

Figure 2. The distribution of Number of Reviews

Data Distribution

图表, 直方图

描述已自动生成图表, 直方图

描述已自动生成

Figure 3. The Distribution of Restaurants Ranking

## Methodology

### 4.1 Data Preprocessing

**Figure 1.** The flow diagram of

### 4.2 Correlation Analysis

徽标

中度可信度描述已自动生成

Where *Num* is the number of newly created EV charge points in a borough of London; *Area* is the area of the borough which is presented in square kilometre.

### Liner Regressing Analysis

Before undertaking regression, it is necessary to convert object(categorial) variables into dummy variables, which is conducive to liner regression. After encoding of categorical variables, …

Limitation: we ignore the multicollinearity, which may affect the results of regression.

**Hypothesis:**

Regressing Model:

Formula:

Test method:

### Logistic Regressing Analysis

**Hypothesis:**

Regressing Model:

Test method:

## Result

### 5.1 Correlation Analysis

As shown in the figure, Ranking and Rating, Ranking and Number of Reviews

图表

描述已自动生成

**Figure .** The distribution of all London boroughs in 2019 and 2020

### 5.2 Liner Regression Analysis

5.2.1 Ranking with all variables

表格

描述已自动生成表格

描述已自动生成

**Figure . Regression Result**

5.2.2 Ranking, Rating and reviewsCnt

手机屏幕截图

描述已自动生成

5.2.3 Ranking with Price

表格

描述已自动生成

5.2.4 Ranking with Price and reviewsCnt

表格

描述已自动生成

5.2.5

手机屏幕截图

中度可信度描述已自动生成

**Figure 3**.

As shown in Figure 4, .

**Figure 4.**

### Logistic Regression Analysis

图表, 树状图

描述已自动生成

**Figure . Logistic Regression Confusion Martix**

**As presented in the figure,**

**After testing its accuracy, 0.7 is not bad,**

## Discussion

### 6.1 Reflection on results

### 6.2 Limitations

## Conclusion

**Word Count:** 1,500

## References