# Summary:Key Drivers of Restaurant Success in Santa Barbara: 2019-2021

Group 11: Jinwen Xu, Yu Luan, Ruotong Zhang

December 6, 2023

## 1 Introduction and Tasks

### 1.1 Introduction

It has been important for business owners to extract actionable insights from complex data sets to drive informed decision-making. For our project, we will build models to analyze data from three sources, including Yelp, to help restaurant owners find key drivers of restaurant success. Focusing on the restaurants in Santa Barbara, California, we aim to extract the most effective attributes that contribute to restaurants' success from 2019 to 2021.

### 1.2 Define Tasks

We will consider both internal and external factors that is important for restaurants' success, including restaurants' characteristics, level of income, and population movement. We will follow the 3 steps which were necessary in coming up with a solution to our problem.

1. We will first research the data and compose relative attributes from different data resources.
2. We will construct the metrics. Considering the components the review data include, we will define the restaurant as successful according to star rating and review count.
3. After coming up with the attributes and metrics, we will develop models to evaluate how our candidate attributes influence the restaurants' success.

## 2 Data Collection and Preprocessing

### 2.1 Data collection

To address our defined problem, we collected data from three sources. The first source is from Yelp, comprising business.json and review.json files. The review.json file contains daily review data from 2005 to the beginning of 2022. Our second source is the distance.csv from the U.S. Department of Transportation's Bureau of Transportation Statistics, which records the number of people traveling various distances each day between 2019 and 2023. Lastly, the third source is from the U.S. Census Bureau, which includes the household annual income and population number for each postal code area from 2011 to 2021.

### 2.2 Data Preprocessing

#### 2.2.1 Data Integration

Seeking to identify key drivers of restaurant success in Santa Barbara from 2019 to 2021, we selected restaurant data from Santa Barbara in the Business.json file. Additionally, we used the Review.json file to calculate the annual average star rating and total review count for each restaurant in Santa

Barbara from 2019 to 2021. These data sets are merged based on the business ID, retaining only restaurants with reviews in all three years. Finally, the data set from the third source was filtered to select data from 2019 through 2021, and postal codes within the data set were merged in the previous step.

### 2.2.2 Handling Missing Values and Encoding Categorical Variables

For missing values in the restaurant features, we use 0 to fill them. In addition, for Categorical Variables in the features, we use integer values starting from 1 to encode them, where a higher original level corresponds to a larger replacement integer.

# 3 Number of Trips Trend Analysis and Visualization

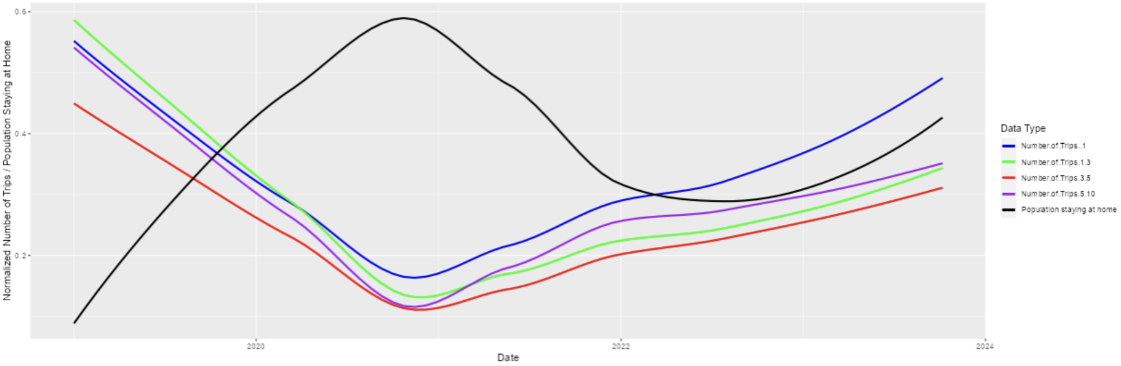


Figure 1: Number of people traveling different distances trend between 2019-2021

Figure 1 shows that in terms of travel distance, travel behavior varies considerably from 2019 to 2021, with an initial decrease followed by an increase in the number of people traveling different distances. This suggests that averaging the data for these years is misleading and that an analysis using annual average data is more appropriate.

# 4 Constructing Metric to Assess Business Success and Choosing Model Variables

## 4.1 Constructing Metric to Assess Business Success

From the collected data, we found that the variables star and review count reflect different aspects of the restaurant's success, the former reflecting the restaurant's reputation and the latter representing the number of people who go to eat. Specifically, we consider a restaurant successful in one year when the annual mean star and total review counts per year are larger than the median of the overall restaurants.

## 4.2 Selection of variables for fitting the model

Due to the large number of restaurant features, we should select appropriate variables. The criteria for selection are as follows: (1). Variables should not have a high number of missing values, which reduces the amount of effective information and leads to significant analysis errors. (2). The values of the variables should be as balanced as possible; if most of a variable's values belong to one category, it can lead to sample imbalance and large analysis errors.

Therefore, based on the above considerations, we choose the average annual household income after normalization, RastaurantsPriceRange, RastaurantsTableService and QuietLevel. Specifically, all missing values in these variables were filled with 0. The restaurant price range ranges from 0 to 4, with larger values representing a wider price range. Restaurant table service is binary data, with 1 representing the presence of restaurant service. The Quiet Level ranges from 0 to 3, with higher values indicating a quieter environment.

# 5 Model Construction and Analysis of Results

## 5.1 Model Construction

We use the annual success of restaurants as the dependent variable. Since it is binary data, we employ multivariate logistic regression, using the variables selected in the previous step as predictors.

## 5.2 Analysis of Results

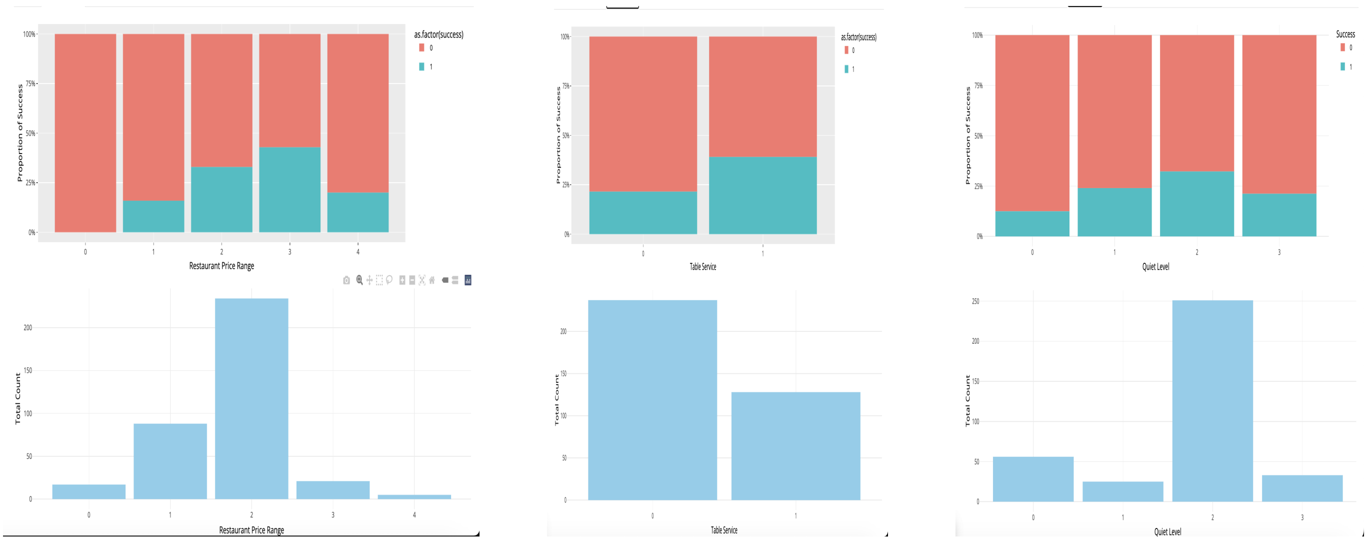
### 5.2.1 Model coefficient analysis

| Coefficients | Estimate | Std. Error | z value | $Pr(>\|z\|)$ |
|---|---|---|---|---|
| (Intercept) | -2.27863 | 0.24697 | -9.226 | $<$ 2e-16 |
| income_c | 0.13606 | 0.07223 | 1.884 | 0.059587 |
| RastaurantsPriceRange | 0.469877 | 0.11922 | 3.932 | 8.43e-05 |
| RastaurantsTableService | 0.56300 | 0.15019 | 3.749 | 0.000178 |
| QuietLevel | 0.18278 | 0.09448 | 1.935 | 0.053038 |

Table 1: coefficient analysis

As shown in Table 1, the p-values for all variable coefficients are less than 0.1, indicating that these variables significantly influence the success of a restaurant annually. Moreover, the coefficients of these variables are all positive, which implies that higher household incomes in the area where the restaurant is located, a wide price range of food in the restaurant offering extensive choices to customers, the availability of table service, and a quiet environment all contribute to the success of the restaurant.

### 5.2.2 Visual Analysis of the Impact of These Predictors



(a) price range

(b) table service

(c) quiet level

Figure 2: Visual Analysis of the Impact of These Predictors in 2020

As shown in Picture 2, taking 2020 as an example, we visualize and analyze the distribution of the success or failure of the restaurant in that year under the independent variables, and find that the results are basically consistent with those obtained from the model. However, we got additional information from the visualization, including the fact that when the price range of the restaurant is too large, it can instead act as a side effect. Due to the scarcity of data when fitting the model with a restaurant price range of 4, this part is not actually represented in the model.

### 5.3 Model Diagnostic

We checked the following assumptions for our multiple logistic regression model. First, we examine the response variable for the logistic regression model. Since we use the success or failure of the restaurant each year as the response variable value, and this is a binary variable, this assumption holds. Second, we assessed multicollinearity among the independent variables using the Variance Inflation Factor (VIF). The VIF values for our variables were as follows: ‘income‘ at 1.044108, ‘Restaurants Price Range‘ at 1.210411, ‘Restaurants Table Service‘ at 1.174550, and ‘Quiet Level‘ at 1.081635. Given that these predictor VIFs are all slightly above than 1, these slight covariances do not reach the level of multicollinearity that is common in logistic regression models. Therefore, we can be confident that our model estimates are robust and not unduly influenced by multicollinearity among the predictors.

### 5.4 Strengths and limitations of the analysis

#### 5.4.1 Strengths

1. We used data from multiple sources to explore the factors contributing to restaurant success from various perspectives, including the restaurant's own attributes and economic impacts. 2. Based on previous analyses, we observed significant temporal variations in the data. Therefore, we analyzed annual average and total data, rather than averaging over the three years. 3. By effectively filtering and encoding variables, we reduced the number of variables before building the model, greatly facilitating subsequent modeling. 4. We established an appropriate model based on the nature of the data, and the p-values indicate that the chosen independent variables have a significant impact on the dependent variable.

#### 5.4.2 Limitations

In analyzing the key drivers of Santa's restaurant success, we utilized data from multiple sources. Consequently, we filtered out data common to these sources, thereby reducing the overall amount of data. As shown in Figure 2, the amount of data is especially small when the price range is at its maximum. This makes it difficult for the model to capture information in this region.

## 6 Conclusion

In addressing the problem of finding the attributes that are important for restaurants' operations in Santa Barbara, California from 2019 to 2021, we explored and analyzed the data from different resources including Yelp, U.S. Census Bureau, and U.S. Department of Transportation's Bureau of Transportation Statistics. After evaluated candidate attributes with our logistic regression model, we identified four important factors of restaurants' success. We believe that restaurants operating in higher-income areas will contribute to the success of the restaurants. In addition, restaurant owners should consider a wider price range in their pricing strategy; providing customers with more choices will lead to higher ratings and attract more people to dine. Increasing high-quality table services is also necessary for increasing customer satisfaction, which brings a higher rating. A quiet environment for dinning is usually demanded by customers, and our model shows that a restaurant with higher quiet level tends to be more successful. Thus, we will suggest restaurants' owners consider more about income, restaurants' price range, restaurants' table service, and restaurants' quiet level.

# Contributions and References

## Contributions

| Contributions | Jinwen Xu | Yu Luan | Ruotong Zhang |
|---|---|---|---|
| Presentation 1 | Responsible for Slides 2, 6, and 7 (Introduction, Variable:Number of Trips with different distance, Models) and review slides. | Responsible for slides 5-10. Reviewed/edited slides 1-4. | Responsible for slides 3 and 6, and provided feedback on all slides. |
| Presentation 2 | Responsible for Slides 3, 4, 6, 7 and review slides. | Responsible for Slides 2,5. | Provided feedback on slides. |
| Summary | Responsible for sections 2, 3, 4, and 5 of the summary and revising the entire summary. | Provided feedback on summary. | Responsible for sections 1,6 of the summary. |
| Code | Responsible for data preprocessing, time series visualization, model building and diagnostic code and reviewing other parts of the code. | Responsible for Figure 2 Visualization code and shiny code. | Reviewed and provided feedback on model code. |
| Shiny App | Review and revise shiny code | The main person responsible for shiny code | Provided feedback on Shiny app |

Table 2: Contributions Table

## References

1 U.S. Census Bureau. Mean Income in the Past 12 Months (in 2022 Inflation-Adjusted Dollars), S1902. 2022
`https://data.census.gov/table/ACSST1Y2022.S1902?t=Income%20(Households,%20Families,%20Individuals)&g=040XX00US06,06$8600000_060XX00US0608390378,0608390380,0608390630,0608391200,0608391708,0608391710,0608392890,0608392908,0608392910,0608392950`

2 U.S. Census Bureau. Total Population, B01003. 2022
`https://data.census.gov/table/ACSDT5Y2021.B01003?t=Population+Total&g=040XX00US06$8600000`

3 U.S. Department of Transportation's Bureau of Transportation Statistics. Trips by Distance. 2023

4 Yelp review.json and business.json