# Empowering Restaurant Owners with Machine Learning: Predicting Business Success and Identifying Opportunities for Growth

**Aranya Bhalla, Ayush Godbole, C S Vinayak, Rajat Srivastava, Dinesh Singh**
Student, Student, Student, Student, Professor/Guide
Computer Science Engineering,
PES University, Bangalore, India

**Abstract** - If we consider the scope of the restaurant industry in India was estimated to be worth |4,23,624 crore and growing at 16% annually. With 1.5 million eating establishments, less than 3,000 of which are organized, the industry is highly fragmented. Because of the pandemic, the numbers vary. The project caters to both existent and soon-to-be restaurant owners. We consider a hybrid market research approach (data collection) for the feasibility report for the restaurant industry and its analysis because it combines and mixes both qualitative and quantitative elements. This technique will evaluate the restaurant's potential location in order to forecast potential supply and demand. Prior to focusing and investing on other restaurants and areas, the idea proves to be feasible as it relies on on-shore analytics to inform better business decisions. what is being said is that data-driven decisions drive better feasibility of a restaurant. This will not only give investors and business owners peace of mind but is a great study to present to potential customers.

**Index Terms** - machine learning, restaurant trends, business intelligence, data-driven decision making, success probability, predictive analysis, review mining, review sentiment analysis.

## I. INTRODUCTION

In India, there is a huge number of restaurants being opened every year. But the success probability of the new restaurants is alarmingly low. Looking at the statistics, 60% of the new restaurants which are opened mostly don't make it past the principal year of their business and 80% go out of business within a year. In spite of these sorts of obstacles, numerous restaurant proprietors and administrators trust that for however long they are bringing in cash, they are doing all around okay. There is an apparent lack of resources for a new restaurant owners or even existing owners. The owners do not have any means to see what kind of sources are helping bigger and more popular restaurants stay afloat. Most of the guides which we come across on opening a restaurant aren't the best, due to the rapidly changing trends in the modern world due to social media and many more What people notice while coming to a restaurant is the value of money, hygiene, service, cuisine, and location. People prefer to choose a restaurant you can walk to. When you want to have a great time, driving afterward can pose a real problem. Sometimes particular cuisines are sold more than others. This paper is targeted at reducing the percentage of closed restaurants as well as helping new restaurant owners to successfully open profitable restaurants by giving the restaurant owners an analysis of restaurant trends over information collected from 50,000+ restaurants. The project will perform exploratory Design Analysis will be performed on the data-set to show the restaurant owner's trends such as popular cuisines, menu-items, favorable localities, etc. We would also be training a model from the data, and assigning restaurants having greater than 3.75 stars and over 500 reviews as popular (successful), given restaurant parameters such as locality, cuisine, price- for-two, etc; give an estimate of success probability. Additionally an in-depth sentiment analysis of the reviews using ML and DL models: an ML model for classifying restaurant reviews as Useful, Cool and Funny, a classification which is becoming increasingly popular for readers of reviews to pick the appropriate reviews, the model will be trained on a data set different from the one used for EDA; a DL model to recognize emotions in the review: positive and negative, the model will be trained on a data set different from the ones used predate send lastly a DL model to owners' sentiments and predict new review menu items positive or negative after Topic Modeling, the model will be trained on the same data set used for EDA.

## II. RELATED WORK

Data visualization, rating prediction, sentiment analysis and queried analysis are the four fundamental stages to analyze the restaurant.

*A)Data Visualization*

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even, animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand. The techniques can be used for visualizing word count vectors and tokenization.

*B)Rating Prediction*

Rating Prediction entails predicting the rating from a given set of inputs like cuisine, location, online ordering etc. The three models were Linear Regression, XGBoost and Deep Learning. Linear Regression is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. XGBoost provides parallel tree boosting and is the leading machine-learning library for regression, classifying stages, of and ranking problems. Deep Learning is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain.

*C)Sentiment Analysis*

Sentiment analysis is contextual mining of text which identifies and extracts subjective information from the source material, and helping a business understand the social sentiment of their brand. For analysis of reviews, we take reviews given by the user, either directly or through URL, and perform sentiment analysis. We perform sentiment analysis on each review and decided if the review was positive or negative, and found the percentage of reviews that were positive and negative. We also find the average sentiment based on sentiment scores and find the overall positive/negative score.

*D)Queried Analysis*

In this section, we use SQL to go through the data and generate insights that the user can use. Here, we take the location and cuisine given by the user. For a given location, we classify stages of the data to find the top-rated cuisines in that area, and for a given cuisine, we find the best and worst locations for that given cuisine, in terms of average rating. We will be looking at the following:

This section talks about the description of the data that is being taken under consideration. Understanding the way, the information works is imperative during the time spent making a decent answer for the main problem in consideration. The main input data present in the project is a data set of restaurant details in Bangalore. This data set is used to train the rating prediction machine learning and sentiment analysis deep learning model. The basic idea of analyzing the Zomato data set is to get a fair idea about the factors affecting the establishment of different types of restaurants at different places in Bengaluru, a city

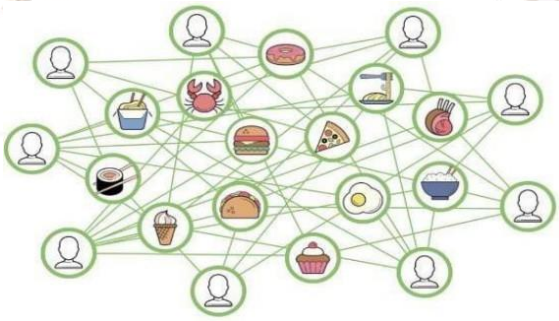

Fig.1 Food Network

with more than 50,000 restaurants serving dishes from all over the world. The demand has been consistently increasing and yet the industry isn't saturated. However, it has become difficult for new restaurants to compete with established restaurants. With such an overwhelming number of restaurants, it is important to study the demo- graphic.


Fig.2 Data Set Address, Name Count

Result: Data set with 50,000+ restaurants from all across Bangalore. There are 10 main relevant columns.

## III. METHODOLOGY

The pipeline consists of several components, each of which use different methodologies. The components and their corresponding methodologies are:

- Pre-Processing the data set for it to be viable,
- Performing Data Visualization to identify trends in the industry,
- Predicting the Rating of a restaurant on basis of given inputs,
- Performing sentiment analysis to identify the polarity of a review,
- Queried Analysis and Data Analytics.

To build the application we have used the Jupyter Note- book with libraries like XGBoost, FPDF, SQL, Pandas to name a few.

*(1)Pre-Processing*

As the data set contains some uninformative features that make it unsuitable for data analytics and rating prediction, it must be preprocessed. We declare a temporary data frame that points to the datasets so that we do not change the source data and any errors are not reflected in the original dataset.

1) Removing Blank locations from the data frame: Locations that are blank should be removed during pre- processing because they cannot be used for data analytics or model training, as location is an important factor in rating prediction and identifying the latest trends. Empty and blank values must be removed from the data frame to achieve this.

2) Removing Blank spaces and adding an underscore between names: As blank spaces between names complicate model training, they are replaced with underscore to make it easier to train. To accomplish this, we must use the replace function to swap out the blank space for underscore.

3) Assigning an integer index to every location, to replace in data frame: Since it would be easier to access the particular location using the index value, we assign a particular index to each location.

4) Dropping Unnecessary columns: Certain columns, such as URL, phone, address, dish liked, menu item, and reviews list, are removed from the data frame because they are not required for data analytics and do not aid in the creation of the model because they have no correlation with other columns or do not aid in rating prediction. URL, for example, has no rating correlation with it.

5) Replacing the 'YES' string with 1 and 0 respectively: We replace the" yes" and" no" values with "1" and "0" respectively because integers are intuitively easier to process from a computational standpoint.

6) Dropping rows with null values: After the following preprocessing is done rows with null values are dropped.

7) Converting into the same data type: All the values which are inputted into the model must be of the same data type so that the model can use that information without any problems. So, the data is converted into the float data type as it has a huge range in comparison to integer or short.

8) Removing special characters: Finally, we drop special characters like commas, exclamation marks etc. using the drop function.

9) Separating each restaurant and cuisine type in separate columns: We've noticed that some restaurants offer a variety of options. As a result, we create a separate column for each type, with a value of 0 or 1 depending on the restaurant. For example, if a restaurant is Quick Bites and Kiosk, it will have 1 in the Quick Bites and Kiosk column and 0 in every other column. Similarly, in the kitchen, we notice the same thing and perform the same task. After that, we will remove the old columns. After making all of these changes to the test data frame, we save it as a separate CSV file and copy the data into the new file so that it can be used for data analytics and ML-DL training models.

*(2)Data Visualization*

After the dataset has been pre-processed, we move on to data visualization. We must declare a new data frame on which to perform the data visualization on because changing the original dataset can make it riskier and less industry accurate. After making a copy of the dataset, we begin working on data visualization. The visualization graph contains a number of smaller tables and graphs. We use the axis function to place the subgraphs or sub-diagrams to prevent the graph from stacking on top of itself. Donut plots, count plots, bar plots, KDE plots, boxplots, line plots, pie charts, heat maps, geo plots, and word clouds are among the plots we employ. These variations can be formed using built-in functions extracted from various libraries, such as pyplot, word cloud, matplot, and geopy, to name a few. We add text to fill in the axis and give different titles to each plot after the plots are formed by using the datasets as input for the graphs. There are various correlations that can be observed, but one of the functionalities is comparing whether or not there is a correlation between the number of orders for restaurants and online orders.


Fig.3 Online Order and Number of Order Correlation

Hence this is how the data visualization is performed upon the dataset which is formed.

*(3)Rating Prediction*

After data analytics is performed, we will start rating prediction using ML and DL models. We have used three models to perform rating predictions namely, Linear Regression, XGBoost and Deep Learning. We import sklearn and TensorFlow to perform the ML and DL model. The sklearn library contains a lot of effective devices for ML and statistical modelling including classification, regression and clustering. We can use the TensorFlow platform to implement best practices for data automation, performance monitoring, model tracking, and model retraining. Before going ahead with training, the model, we must first take user inputs so that we can train the models. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient- boosted decision tree (GBDT) machine learning library. It provides a parallel tree boosting and is the leading machine- learning library for regression, classification, and ranking problems. Deep learning is a subset of machine learning, which essentially consists of three or more layers. These neural networks attempt to simulate the behaviour of the human brain. In linear regression, we take x and y and we have to apply scaling on it to do that we use linear regression. By applying linear regression, we get a mean square error of 0.22 the average in most cases is usually 0.2.

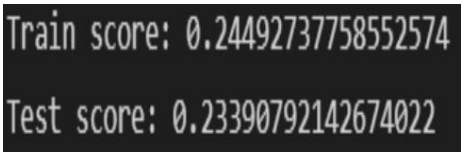


Fig.4 Train and Test Score

In XGBoost we get an average of 7.8% error and an MSE of 0.12 which is the desired value. We perform XGBoost for both rating and prediction. In Deep Learning we ran a model with three layers to 75 neurons, ReLU activation, RMS, and prop optimizer and we take mean scale error and mean average percentage error. With the deep learning model, we get an average error of around 9% between epochs, it varies between 9% and 11%.

*(4)Sentiment Analysis*

Before we begin using NLP on the review, we must pre-process the datasets in the same way that rating predictions are done.

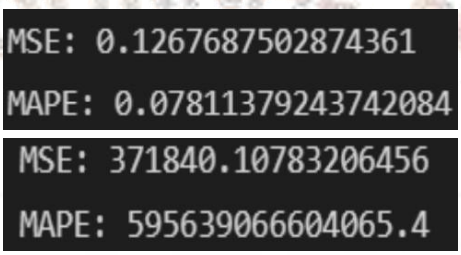


Fig.5 MSE and MAPE for various approaches

The reviews are either scraped from the official Zomato site or manually entered by the user. When the URL of the restaurant is provided, web scraping is performed. The scraping is done with Selenium Web Driver and Chrome Web Driver. After that, we calculate the average length of each review, which is around 12,000 characters. The longest review is approximately 1 million characters long. Because other columns aren't relevant in this model, we limit the dataset to just review and rating. The rating is divided into two parts: train and test. We even pad the review to around 1000 character. Following that, word tokenization is performed, in which we split the review and tokenize each review, assigning a value to each character. The tokenizer is then created. We divided the table into two parts: validate and train. Then, for the problem, we build a neural network with the first layer being the embedded layer, the second layer being the bidirectional layer, and the final layer being the dense layer, totalling 6 neurons. We classify the review using these six neurons. We used ReLU for the dense layer, SoftMax for the final activation, and spars categorical cross entropy for model compilation.

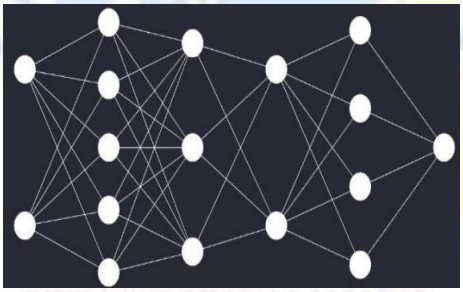

Fig.6 Neural Network

*(4)Queried Analysis and Data Analytics*

We proceed with queried analysis after the dataset has been pre-processed. We must declare a new data frame on which to perform the requested analysis because changing the original dataset can make it riskier and less industry accurate. In data analytics, we will identify the relevant correlation between two columns with an arbitrary value of 0.33, as this indicates significant correlation between the columns. To accomplish this, we use a nested loop, declare separate columns, and check if the correlation value is greater than 0.33 or less than -0.33. The correlation staple is 0.33 because 0 - 0.3 indicate weak positive correlation and anything above 0.33 indicates strong correlation. From this, we can infer that:

Votes which indicate the popularity of a restaurant is positively correlated to" booking table"," rating", and" cost". This means that:

- Restaurant is more popular if online booking is available.
- Higher rated restaurants are more popular
- Higher costing restaurants are more popular

Furthermore, the cost and rating of a restaurant have a significant correlation with the booking table, indicating that if online booking is available, the rating and cost are both higher, with the cost being especially strong, indicating that expensive restaurants have online bookings. Finally, the data shows that the higher the rating, the higher the restaurant's price. Now once the correlation has been determined we created a database using SQ, so that we can show the data for restaurants with ratings of more than 3.75 and votes of more than 177. After the database has been formed, we perform various queries like:

1) For each type of cuisine what's the highest rated restaurant.
2) Finding the top 5 highest and lowest rated restaurants location.

3) The lowest average rated restaurants in a particular area.
4) Finding average rating of each cuisine type in a particular area.

```
['votes', 'book_table_int', 0.3931857146696946]

['votes', 'rating', 0.4352563931175224]

['votes', 'cost', 0.3665560490444662]

['restaurants_in_area', 'locations_int', -0.35057819196065143]

['book_table_int', 'rating', 0.42606964633625094]

['book_table_int', 'cost', 0.6142939270061158]

['rating', 'cost', 0.3853836383941031]

['type', 'cost', 0.37608617371844394]
```

Fig.7 Correlation Table

## IV. RESULTS

The target dataset consists of several different types of outputs. We have separated them into four different classes:
a. Data Visualization
b. Review Analysis
c. Rating Prediction
d. Queried Analysis

A)*Data Visualization*

There will be useful charts presented that can be used to gain intuition about the data. To assist users in drawing their own conclusions, the Exploratory Data Analysis will be divided into topics. Questions that can be answered with some ratings or table analysis will be posed onto the exploratory analysis in each topic. These are some of the generated data visualization graphs and tables.

```
Average rating of Sushi restaurants in JP_Nagar
3.7474999999999996


Average rating of Juices restaurants in JP_Nagar
3.587878787878787


Average rating of Coffee restaurants in JP_Nagar
3.6968750000000004


Average rating of Greek restaurants in JP_Nagar
3.7543478260869563
```

Fig.8 Average rating of a cuisine in a particular area

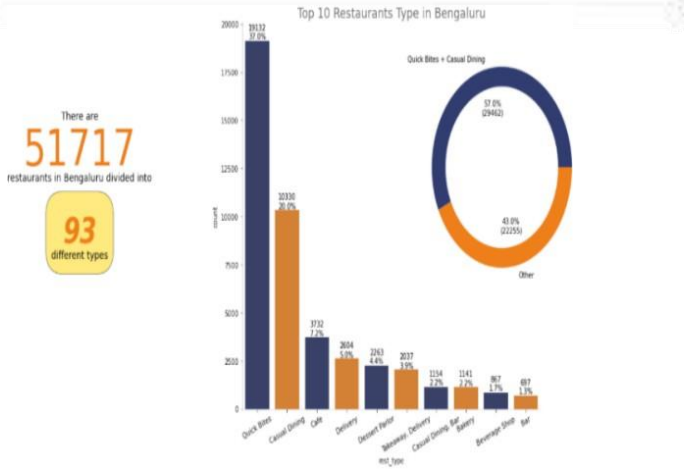


Fig.9 a) Count of restaurants b) Top 10 restaurants in Bangalore c) Number of Quick bites and Casual Dining restaurants

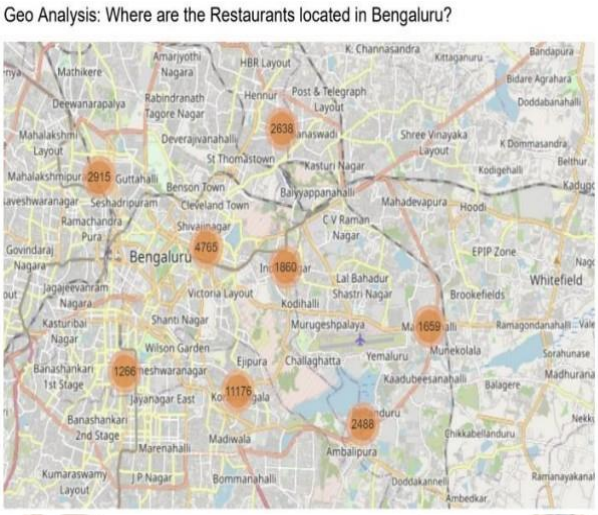


Fig.10 Geo Analysis for restaurants located in Bangalore

There are additional graphs that have been created, such as how many restaurants offer ordering services, how Ratings and Cost-for-Service are calculated, and so on. Two distributed among Restaurant Types, what are the most popular Dishes in Bengaluru, is there a link between Online Order option and Restaurant Rating, and various other questions

B)*Review Analysis*


Fig.11 Percentage of positive and negative reviews andoverall sentiment.

C)*Rating Prediction*

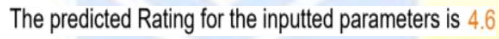


Fig.12 Predicted rate after input details are provided

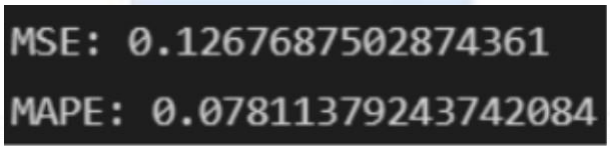


Fig.13 MSE and MAPE for XGBoost.

D)*Queried Analysis*

The Lowest and Highest Average Rated Cuisines in Church Street

The Lowest Rated Cuisines for given Location are:

| | Number of Restaurants in Area | Rating | Cuisine |
|---|---|---|---|
| 0 | 4 | 3.575000 | Oriya |
| 1 | 7 | 3.614286 | Seafood |
| 2 | 9 | 3.711111 | Bakery |
| 3 | 8 | 3.725000 | Mithai |
| 4 | 9 | 3.755556 | Andhra |

The Highest Rated Cuisines for given Location are:

| | Number of Restaurants in Area | Rating | Cuisine |
|---|---|---|---|
| 99 | 10 | 4.210000 | Sandwich |
| 100 | 6 | 4.233333 | Coffee |
| 101 | 9 | 4.244444 | Japanese |
| 102 | 4 | 4.300000 | Burger |
| 103 | 2 | 4.400000 | Spanish |

Fig.14 Highest and Lowest Rated Cuisines for a given location query

## V. CONCLUSION

To demonstrate the most recent trends in the restaurant industry, various data visualization methods were used. Before using the dataset, it must be cleaned, and various methods were used to do so. Unneeded columns such as URL are removed. To simplify the dataset, various values were reduced to larger districts, and various reviews that were in different languages were changed to English so that no problems arose. Redundant information was removed, and restaurants with no reviews or ratings were removed from the dataset. Visualization graphs such as scatter plots, bar graphs, line graphs, and pie charts were used to demonstrate various industry relationships and trends. Stating a few examples of line graphs for various restaurants, with prices and ratings as coordinates, as well as the number of restaurants in each region and their prices. This research also focuses on developing a machine learning classification model to predict the success probability of a restaurant given certain parameters. We determined the restaurant rating by providing a certain minimum number of reviews and the rating of that particular restaurant by also providing a certain minimum which must be satisfied. To solve this type of problem, many state-of-the-art Machine learning algorithms were used. One of the ML algorithms used to solve this problem, decision tree, outperformed the other machine learning models. The dataset used for this problem contains more than 15,000 restaurants. Finally, the proposed machine learning-based data may be a promising method for predicting restaurant success. The NLP model is used to determine the review sentiment

## VI. ACKNOWLEDGMENT

We would like to thank the project advisor, Dr.Dinesh Singh, and Dr.Shylaja Sharath, the Chairperson, Department of Computer Science and Engineering at PES University, for their support and guidance in completing the project.

## VII. REFERENCES

[1] Alamoudi, Eman Seed and Sana Al Azwari. “Exploratory Data Analysis and Data Mining on Yelp Restaurant Review.” 2021 National Computing Colleges Conference(NCCC). IEEE 2021
[2] Asghar, Nabiha. “Yelp dataset challenge: Review rating prediction.” arXiv preprint arXiv: 1605.05362. 2016.
[3] Shibata, Akito Kamei, Sayaka Nakano, Koji. Category-oriented Sentiment Polarity Dictionary for Rating Prediction of Japanese Hotels. (2020). 440-444. 10.1109/CANDARW51189.2020.00090
[4] Rehman, Anwar Ur, et al. ”A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis.” Multimedia Tools and Applications 78.18 (2019): 26597- 26613.
[5] Batbaatar, Erdenebileg, Meijing Li, and Keun Ho Ryu. ”Semanticemotion neural network for emotion recognition from text.” IEEE Access 7 (2019): 111866-111878.
[6] Hicks, Stephanie C., and Rafael A. Irizarry. ”A guide to teaching data science.” The American Statistician 72.4 (2018): 382-391.
[7] Grus, Joel. Data science from scratch: first principles with python. O’Reilly Media, 2019.
[8] Gibert, Karina, et al. ”Which method to use? An assessment of data mining methods in Environmental Data Science.” Environmental modelling software 110 (2018): 3-27.
[9] Zhang, Amy X., Michael Muller, and Dakuo Wang. ”How do data science workers collaborate? roles, workflows, and tools.” Proceedings of the ACM on Human-Computer Interaction 4.CSCW1 (2020): 1-23.
[10] Kabir, Ahmed Imran, Koushik Ahmed, and Ridoan Karim. ”Word Cloud and Sentiment Analysis of Amazon Earphones Reviews with R Programming Language.” Informatica Economica 24.4 (2020): 55-71.
[11] Janiesch, Christian, Patrick Zschech, and Kai Heinrich. ”Machine learning and deep learning.” Electronic Markets 31.3 (2021): 685- 695.
[12] Glielmo, Aldo, et al. ”Unsupervised learning methods for molecular simulation data.” Chemical Reviews 121.16 (2021): 9722-9758.
[13] Ghazal, Taher M., et al. ”Performances of K-means clustering algorithm with different distance metrics.” (2021).
[14] Sarker, Iqbal H. ”Machine learning: Algorithms, real-world ap- plications and research directions.” SN Computer Science 2.3 (2021): 1-21.
[15] Moustafa, Sayed SR, et al. ”Development of an optimized regression model to predict blastdrivenground vibrations.” IEEE Access 9 (2021): 31826-31841.
[16] Sahoo, G., and Yugal Kumar. ”Analysis of parametric non parametric classifiers for classification technique using WEKA.” International Journal of Information Technology and Computer Science (IJITCS) 4.7 (2012): 43.
[17] Potok, Thomas. ”Adiabatic quantum linear regression.” Scientific reports 11.1 (2021): 1-10.
[18] Schober, Patrick, and Thomas R. Vetter. ”Logistic regression in medical research.” Anesthesia and analgesia 132.2 (2021): 365.