Journal of Global Research in Computer Science ISSN: 2229-371X

Predicting Ratings of Online Food Chain

Srishty Sri Nidhi*, Ravi Shankar Pandey

Department of Computer Science and Engineering. Birla Institute of Technology Mesra, Patna, India

Research Article

Received: 30-Nov-2022, Manuscript No. GRCS-22-81807; **Editor assigned:** 02-Dec-2022, PreQC No. GRCS-22-81807 (PQ); **Reviewed:** 16-Dec-2022, QC No. GRCS-22-81807; **Revised:** 30-Jan-2023, Manuscript No. GRCS-22-81807 (R); **Published:** 06-Feb-2023, DOI: 10.4172/2229-371X.14.1.001

***For Correspondence:** Srishty Sri Nidhi, Department of Computer Science and Engineering, Birla Institute of Technology Mesra, Patna, India;

Email: srinidhisrishty4@gmail.com

Citation: Nidhi SS, et al. Predicting Ratings of Online Food Chain. RRJ Glob Res Comput Sci.2023;14:001.

ABSTRACT

Online food ordering is going to be popular day by day and it requires customer satisfaction for more popularity in the society Several online food ordering System are available on internet like Zomato, Swiggy, Fresh menu, Dunzo, Guruhub, EatSure, UberEats, Deliveroo, dominos etc. All such kind of system requires customer satisfaction in the form of the feedback mechanism. This feedback mechanism helps to provide Appropriate food at location on the basis of customer rating.

In this paper we have analysed data of Zomato to incorporate location wise customer satisfaction to provide better restaurant for food ordering to customer.

We have used machine learning linear regression technique to separate better restaurant on the basis of customer satisfaction rating. We will also use this algorithm to predict aggregate ratings restaurants will receive based on different data points. We have tested our algorithm using Kaggle dataset.

Keywords: Online food ordering; Customer satisfaction; Algorithm; Kaggle dataset; Machine learning

INTRODUCTION

Online food ordering is going to be popular day by day and it requires customer satisfaction for more popularity in the society several online food ordering system are available on internet like Zomato, Swiggy, Fresh menu, Dunzo, GuruHub, EatSure, UberEats, Deliveroo, dominos etc. All such kind of system requires customer satisfaction in the form of the feedback mechanism. Data analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense, and recap, and evaluate data. Data analysis can help companies better understand their customers, evaluate their ad campaigns, personalize content, create content strategies, and develop products [1].

On this project we are applying our knowledge of data analysis on Zomato dataset where Zomato dataset is real time data set which gives information about restaurants, its cuisines, locality, ratings etc. City: Contains the neighbourhood in which the restaurant is located. Locality verbose: Exact place in that locality.

The technologies I am using in this analysis are Python, NumPy, Pandas, Seaborn, Matplotlib, Sklearn etc.

The basic idea of analysing the Zomato dataset is to get a fair idea about the factors affecting the establishment of different types of restaurants at different places of India. The national restaurant association of India, founded in 1982 represents over 5,00,000 restaurants, QSRs, Bars, cloud kitchens and Catering, pan India serving dishes from all over the world. With each day new restaurants opening the industry has not been saturated yet and the demand is increasing day by day. In spite of increasing demand, it however has become difficult for new restaurants to compete with established restaurants. Most of them serving the same food. Most of the people here are dependent mainly on the restaurant food as they do not have time to cook for themselves. With such an overwhelming demand of restaurants it has therefore become important to study the demography of a location. What kind of a food is more popular in a locality. Does the entire locality love vegetarian food [2]. If yes then is that locality populated by a particular sect of people for e.g., Jain, Marwaris, Gujaratis who are mostly vegetarian. This kind of analysis can be done using the data, by studying the factors such as

- Location of the restaurant.
- Approx price of food.
- Theme based restaurant or not.
- Which locality of that city serves those cuisines with maximum number of restaurants.
- The needs of people who are striving to get the best cuisine of the neighbourhood.
- Is a particular neighbourhood famous for its own kind of food.

Related works

The rapid growth of data collection has led to a new era of information. Data is being used to create more efficient systems and this is where recommendation systems come into play. Recommendation systems are a type of information filtering systems as they improve the quality of search results and provides items that are more relevant to the search item or are related to the search history of the user. They are active information filtering systems which personalize the information coming to a user based on his interests, relevance of the information etc. Recommender systems are used widely for recommending movies, articles, restaurants, places to visit, items to buy etc.

Different from those existing work, in this paper, we propose an algorithm based on machine learning, our contributions are:

- **We will be using content based filtering content based filtering:** This method uses only information about the description and attributes of the items users has previously consumed to model user's preferences.
- We have used machine learning linear regression technique to separate better restaurant on the basis of customer satisfaction rating. We will use linear regression algorithm to predict aggregate ratings restaurants will receive based on different data points. We will check co relation between different variables and select important features from the dataset for predicting the aggregate rating [3].

MATERIALS AND METHODS

Proposed method

We have used machine learning linear regression technique to separate better restaurant on the basis of customer satisfaction rating. We will use linear regression algorithm to predict aggregate ratings restaurants will receive based on different data points. We will check co relation between different variables and select important features from the dataset for predicting the aggregate rating. We will be using content based filtering content based filtering: This method uses only information about the description and attributes of the items users has previously consumed to model user's preferences [4-9].

Algorithm

Data collection: The data set is a matrix where the rows represent the details of the restaurants and the columns represent the factors or attributes (features) to be tested. The data in the dataset can be divided into two types: Categorical data and number of data. There are 17 different attributes and contains about 6000 related data. We collected our data set from Kaggle website. After collecting data we performed data cleaning process.

The data contains details of restaurants associated with an online e-commerce food aggregator. Various data points are available related to the restaurants. We explore the data points one by one and try to find best insights from it.

Feature selection: Feature selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data [10].

The choice of features is one of the major challenges to train a predictive learning algorithm. Feature selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you is trying to solve [11-14].

Data pre processing: Data pre processing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. In other words, it should be transformed in such a form so that it can be easily interpreted by different algorithms with producing higher accurate results. It is not necessary to have complete pure data in each and every dataset. There is always some missing data in each and every dataset in "NULL" form due to which the dataset becomes redundant and hence leads the models to predict results with poor accuracy. Hence, to overcome these poor accuracies and to attain higher and better accuracies, data pre processing came in genre. We usually clean the tuples having missing values by either dropping those tuples from the dataset or by imputing mean or median values of respective columns or some other hyper parameter optimization to attain the imputable values for replacing those missing values.

Data visualization: Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion [15].

In the world of big data, data visualization tools and technologies are essential to analyze massive amounts of information and make data driven decisions. We have analyzed our data on different grounds.

Prediction: Predictive modelling is a commonly used statistical technique to predict future behaviour. Predictive modelling solutions are a form of data mining technology that works by analysing historical and current data and generating a model to help predict future outcomes. In this research work I predicted using machine learning linear regression algorithm.

RESULTS

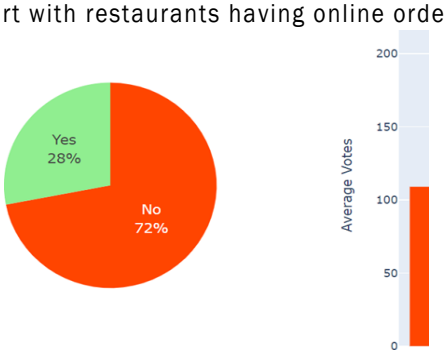
Data visualization results

Restaurants that offer online delivery to their customer: The majority of restaurants do not offer online delivery facility to their customers. Only 28% of restaurants offer this facility to their customers.

In conclusion, on an average, restaurants offering online delivery to their customers received more votes than the restaurants which don't offer one. This could be attributed to multiple reasons.

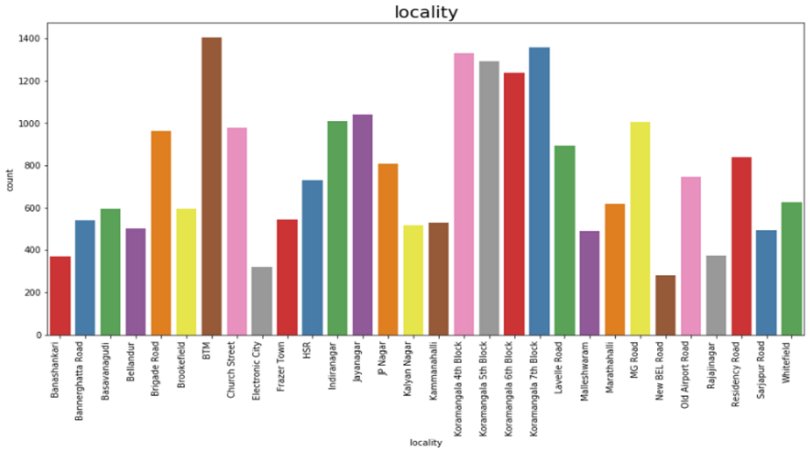
- One probable reason can be that people ordering food online are more tech-savvy and generally give their review for food they ordered.
- Another reason could be that online food delivery apps prompt their users to review the food they ordered after some time.
- Having online delivery facilities could also increase the reach of the restaurants which would result in more people consuming the services of restaurant and voting on the same (Figure 1).

Figure 1. Pie chart with restaurants having online order facility or not.



Number of restaurants received votes in different locations: The majority of votes received by the restaurants located in BTM and Koramangala from the figure we can see that they have received more than 13000 votes from this we concluded that if someone have to open restaurants then they should must chose areas where there are few restaurants because there is more chance to grow their business (Figure 2) [16].

Figure 2. Bar graph between location and number of votes.

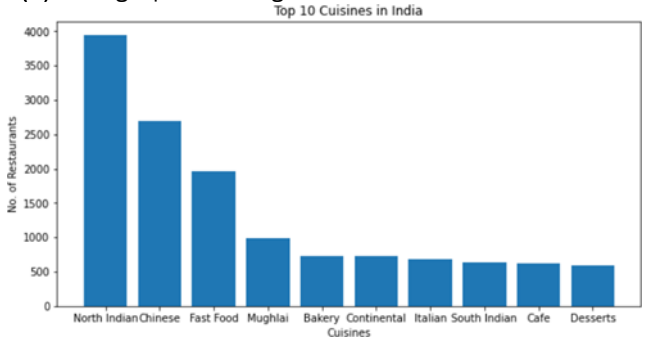


Word cloud or cuisine cloud:

- Here we have made a word cloud and bar graph showing us the common cuisines served in restaurants of cities which are predominantly located in the BTM Region. The bar graph shows a cuisine and number of restaurants serving that cuisine.
- Word cloud is a visual way to know the most common cuisine served in restaurants. The size of the cuisines shown in the word cloud will increase with the number of restaurants serving them. Hence we have words such as "North Indian", "Chinese", "fast food" in bigger size relative to other cuisines (Figures 3a and 3b).

Figure 3(a). World cloud of restaurant offering different cuisines.

Figure 3(b). Bar graph showing votes received for different cuisines.



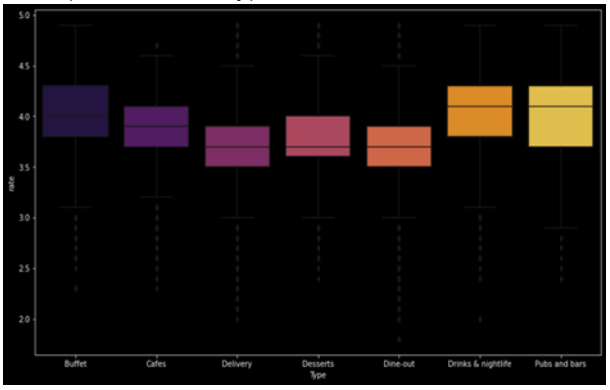
Types of restaurants versus rates: In Figure 4 I have taken boxplot it was plotted between different types of

restaurants and rate for that theme of restaurant. From this we analysed that the maximum average rating is given to the "drinks and nightlife" which means people of Bangalore more like restaurants of theme drinks and nightlife. So, if someone has to open a restaurant in Bangalore then they can choose "drinks and nightlife" Or "Pubs and Bars" kind of restaurant.

Restaurant ratings depend on many factors such as portion size, ambience, waiting time, and also on different facilities that are provided by restaurants such as valet parking, online delivery, table booking etc. having facilities like online delivery and table booking affect ratings. We will also see if there is a relation between number of cuisines and ratings [17].

Aggregate ratings are the averages of all individual ratings which are given to restaurants. These ratings are measured on the scale of 5. Ratings also have colour and text associated with them. We will define the range of different rating colours and rating text in terms of aggregate rating.

Figure 4. Box plot between types of restaurants and votes received.



Relation between services and ratings: We can infer from the below visuals that restaurants that offer facilities such as online delivery and table booking are more likely to have an above average rating *i.e*, rating in range of good, very good and excellent. Around 60% of restaurant that offer table booking have an above average rating.

The same is also true for online delivery facility but the difference is not as large in table booking scenario. Around 50% of restaurant that offer online delivery facility has an above average table booking and only 38% of restaurants that don't offer online delivery facility have above average rating [18].

One reason could be that offering services such as online delivery and table booking doesn't boost ratings significantly, but for many restaurants offering such services increases ratings slightly, to push them from the upper end of average rating segment to lower end of good rating segment.

The median and average ratings doesn't increase significantly. For example, the average rating increases from 3.31 for restaurants not having table booking to 3.55 for restaurants that have the facility.

From this we can make an calculated hypothesis that if an restaurants has facility such as online delivery and table booking it is more likely that it will have an above average rating (Figures 5a and 5b) [19].

Figure 5(a). Percentage of restaurants which has online delivery options.

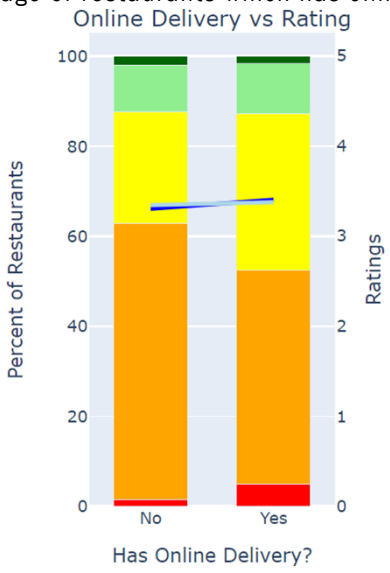
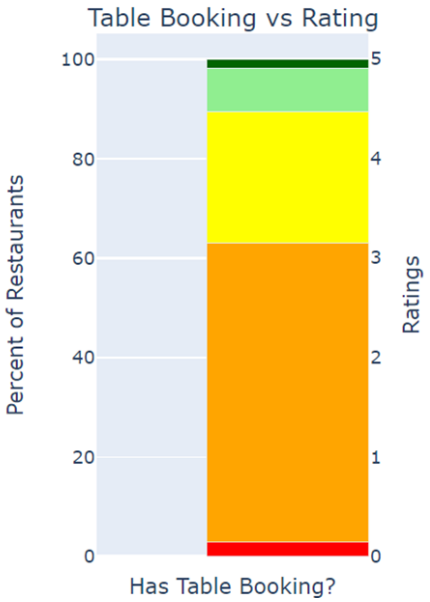


Figure 5(b). Percentage of restaurants having table booking facility.

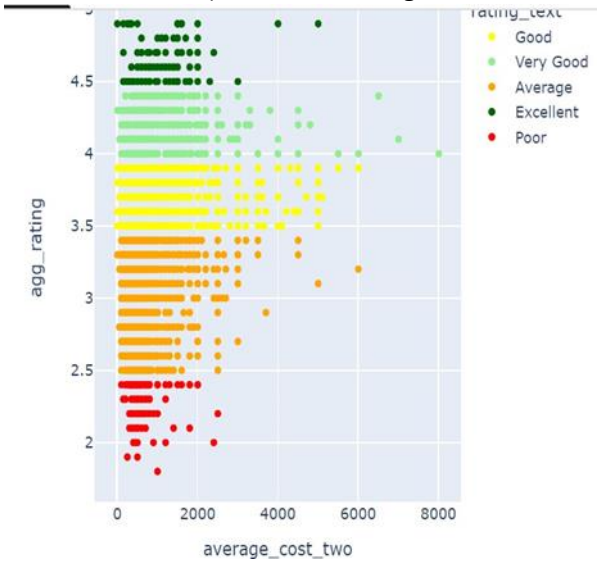


Here in the above figure:
Light blue indicates average ratings
Dark blue as median ratings
Dark green for excellent
Light green for very good
Orange for good
Red is for poor.

DISCUSSION

Relationship between average cost and ratings: There is no relationship between ratings and average cost for two as there are restaurants with high as well as low ratings for most price points bottom. Do not change the vertical spacing to align the bottoms of both columns (Figure 6) [20].

Figure 6. Relationship between average cost and ratings.



Factors affecting restaurants

We will use linear regression algorithm to predict aggregate ratings restaurants will receive based on different data

points.

Linear regression: What Is linear regression? Linear regression is a supervised learning algorithm that compares input (X) and output (Y) variables based on labeled data. It's used for finding the relationship between the two variables and predicting future results based on past relationships.

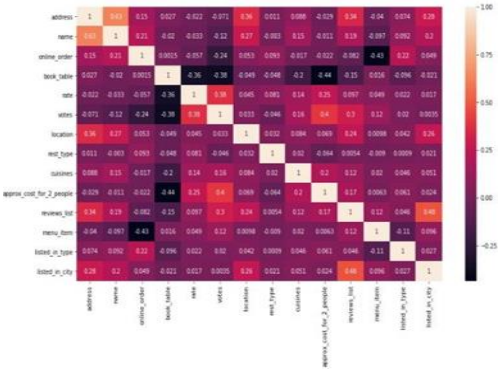
For Linear regression we performed some steps are:

- **Cuisines offered by zero restaurants are:** There are certain cuisines which have no restaurant serving them in the NCR region of India. We will also drop those cuisines as features as they don't have any impact on our model.
- **Finding the highly co-linear columns:** The highest correlation is between name and address which is 0.63 which is not of very much concern Splitting dataset into train and test.

Impact of highly colinear columns results in

- **Uncertainty in coefficient estimates or unstable variance:** Small changes (adding/removing rows/columns) in the data results in big change of coefficients.
- **Increased standard error:** Reduces the accuracy of the estimates and increases the chances of detection.
- **Decreased statistical significance:** Due to increased standard error, t-statistic declines which negatively impacts the capability of detecting statistical significance in coefficient leading to type-II error.
- **Reducing coefficient and p-value:** The importance of the correlated explanatory variable is masked due to collinearity.
- **Overfitting:** Leads to overfitting as is indicated by the high variance problem (Figure 7).

Figure 7. Co-relation between different data points.



- **Final feature selection:** We will select the most important data points that could affect rating. We have explored the dataset quite a bit and have an idea about various data points. Excluding price range as features because it is highly co-related to average cost two. We will also have votes columns as feature because there is some weak correlation between votes and aggregate rating.
- **Applying linear regression:** After applying regression analysis we got our regression model showing actual rating *vs.* predicted rating (Figures 8a and 8b).

Figure 8(a). Graph between predicted ratings versus actual rating.

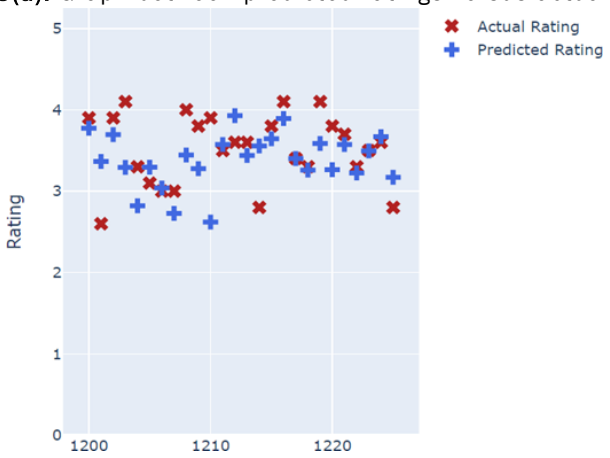
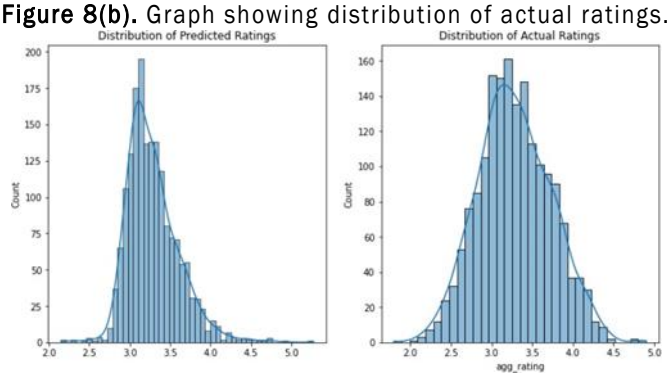


Figure 8(b). Graph showing distribution of actual ratings.



From this we got our linear regression value.
R square as: 0.3209984009348331.
Mean square error: 0.13915474370146483 mean absolute error: 0.28694827410331963.
Mean absolute percent error: 9.056286039506084.

These are metrics to analyze the accuracy of our linear regression model. We can infer from the 'predicted rating *vs.* actual rating' graph that for some restaurants we are missing the actual rating quite widely and for some we are predicting it quite accurately.
R-squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).
The R^2 value implies that there is 32.09% less variation around our linear regression line than the mean. If we had more relevant data points such as average waiting time, portion size, availability of various facilities such as AC, fan etc., this score could be improved.
The mean square error and mean absolute error are quite low and we are able to predict all our values close to the actual value. Mean absolute percent error tells us, on an average, we are 9 percent off from the actual rating.

CONCLUSION

After analysing we concluded that there are different points which affects the growth of the business of restaurants in world. It helped us to see, interact with, and better understand data. From predicted rating and actual rating graph from linear regression model I got to know that that for some restaurants we are missing the actual rating quite widely and for some we are predicting it quite accurately.

Acknowledgment

I would like to take this opportunity to thank the people who have made the implementation of this paper possible. Completion of this Project without any guidance would not be possible which was led by our professor and teaching staff. I express our sincere gratitude to our beloved guide Dr. Ravi Shankar Pandey Sir, assistant professor, who provided valuable guidance, suggestions and hand in hand cooperation throughout the completion of this project.
I also wish to extend our sincere gratitude towards the teaching and non-teaching staff of the department of computer

science and engineering for their technical support.

REFERENCES

1. Anscombe F, et al. Graphs in Statistical Analysis. Am Stat. 1973;27:17-21.
2. Box GEP, et al. Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. 2nd Edition. John Wiley and Sons. New York, United States of America. 1978.
3. Chambers, et al. Graphical Methods for Data Analysis. 1st Edition. Wadsworth and Brooks Publications. California, New York. 1983;395.
4. Chatfield C, et al. The Analysis of Time Series: An Introduction. 6th Edition. Chapman and Hall Publications. New York. 2003:352.
5. Cleveland, et al. Visualizing Data. Hobart Press. Summit, New Jersey. 1993.
6. Devaney JE, et al. Equation Discovery through Global Self-Referenced Geometric Intervals and Machine Learning. George Mason University ProQuest Dissertations Publishing. Fairfax, Virginia. 1998;1-24.
7. Draper, et al. Applied Regression Analysis. 3rd edition, John Wiley and Sons Publisher, Newyork. 1981;695
8. Du Toit, et al. Graphical Exploratory Data Analysis. 1st Edition. Springer-Verlag Publications, Newyork. 1986.
9. Evans, et al. Statistical Distributions. 4th Edition. John Wiley and Sons Publications, Hoboken. New Jersey. 2000.
10. McNeil, et al. Interactive Data Analysis. A Wiley-Interscience publication, New York and Toronto. 1977;186.
11. Natrella, et al. Experimental Statistics. National Bureau of Standards Handbook 91. Washington, DC. USA. 1963.
12. Neter J, et al. Applied Linear Statistical Models. 5th Edition. The McGraw-Hill Companies. Irwin. 1990.
13. Ryan, et al. Modern Regression Methods. 2nd Edition. John Wiley Publisher, Hoboken, New Jersey. 1997;127.
14. Scott, et al. Multivariate Density Estimation: Theory, Practice, and Visualization 2nd Edition. John Wiley and Sons Publications, New York. 1992;361.
15. Tufte, et al. The Visual Display of Quantitative. 2nd Edition. Graphics Press, China. 1983.
16. Tukey, et al. Exploratory Data Analysis. 1st Edition. Addison-Wesley, America. 1977;688.
17. Velleman, et al. The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis. J R Stat Soc Ser C Appl Stat. 1981;320-321.
18. Bortz J, et al. Statistic for Human-und Sozialwissenschaftler. 6th Edition. Heidelberg, Springer. New York, USA. 2004.
19. Selvin S, et al. Epidemiologic Analysis. Oxford University Press, United States of America. 2001.
20. Sir Bradford Hill A, et al. The environment and disease: Association or Causation?. Proc R Soc Med. 1965;58:295–300.