EXPLORATORY DATA ANALYSIS AND VISUALIZATION ON ZOMATO RESTAURANTS DATA

PRAVEEN YARAGATTI¹, SIDDHARTH NANDA²
U.G. Student, SOE, ADYPU, Lohegaon, Pune, Maharashtra, India¹
Faculty-IT, iNurture, Bangalore, India²

Abstract

Present and future generation is the journey of e-commerce platform is booming in world. In India, several folks are drawn to technology. Indian shoppers are just about victimization the e-commerce platforms like on-line shopping, booking tickets, and health care, etc. at intervals the e-commerce platform food business is one of the amongst the foremost on-line food ordering business. This has brought just about every eating places are victimization e-food business, were in Asian nation, were differing kinds of business corporation area unit like zomato, swiggy, and foodpand, etc. Present vertical has showing tremendous growth over the last five years, signifying the untapped potential. The leader of the Republic of India has been Zomato. Zomato could also be related to each shoppers would associate to once he thinks of Associate in nursing e-food platform. The demo zomato restaurants graphics of the situation relies on k-means clusters algorithms, it’ll facilitate us to understand concerning the restaurants and rate of growth of the restaurants with the help of visualization. After this module is deploying in shiny web application, it provides clearly to know over browser of knowledge visual image of everywhere indian zomato data, were that city’s are unit most victimization the zomato on their restaurants, here we have a tendency to grasp concerning the top ten restaurants in specific town on the premise of various sort of cuisines providing restaurants and consumer review additionally it facilitate us in comparison between one with another city’s restaurants. This paper is facilitate us in digital promoting trade to create the promoting all told over restaurants additionally revenue of all restaurants weather is increasing or decreasing.

Keywords: E-commerce, E-food business, Zomato, K-means clustering algorithms, Visualization, Shiny, Restaurants, Food

I. INTRODUCTION

Zomato is one of the most important online food ordering businesses in today's generation, most of the customers are zomato abuse. It's kind of an online guide book for users of zomato and restaurants, bars, clubs, pubs, and dine-outs. The users of zomato are asked to give up the feedback and rate the restaurants on the basis of their expertise, visits, and selection. Zomato's website is useful for zomato customers to restaurants or pubs based on the presence or absence of feeder and non-vegetarian food, air conditioning,
smoking area, stag entrance, Wi-Fi internet, outside sitting area, alcohol is served correctly or not. Zomato web app displays alcohol-related data, served kitchens, restaurant opening time and shutdown time, estimated meal price for two. Whether or not a master card has been accepted. Restaurants change their business strategy to improve food and alcohol based on customer review or feedback. With the Foodiebay.com, the first zomato web application is launched, created by Deepinder Goyal. Foodiebay will be launched in July 2018 with 1,200 restaurants in the Delhi (urban center) NCR area, once this data is expanded to major cities such as Bangalore, Pune, Mumbai and Kolkata, etc. With half-consistent technology advancement being used in all business. Machine learning is one of the most trendy technology in digital technology, so in machine learning k-means clustering is a bunch rule that victimizes the idea of similarities to create the clusters. This room of an ordeal faculty involves a clump of k-means consumption of choice. K-means clustering is a little more technique that has generally been systematically separated into information that has been put aside under consideration by nationality. Basically, as long as a number of advantages  mercantilism clustering, it began relocate every purpose out calculator the intending to of the other limits. This isn’t in hot water iteratively till a specific amount defined criteria trends been achieved. The K-means agglomeration rule is employed to search out folks that haven’t been expressly tagged within the information and to search out patterns and create higher selections. Once the half rule tends been run and also the folks area unit outlined, any new information may be simply appointed to the foremost relevant cluster.

**Algorithms of K-means Clustering**

To start with the k-means algorithm, you first need to initialize points randomly called the centroids (K) cluster. K-means is a iterative algorithm that takes two steps:

1. **Cluster assignment**
   The algorithm passes through each of the data points and assigns data points to one of the three cluster centroids depending on which cluster is nearer.

2. **Move centroid**
   Here, K-means shifts the centroids to a cluster's average points. In other words, the algorithm calculates the average of all of a cluster's points and moves the centroid to that average location. This process is repeated until there is no cluster change.
possibly until some other condition of stoppage is fulfilled). K is selected randomly or by providing the user with specific initial starting points.

Fig. 2 K-Means at work.

Shiny is a RStudio package that can be used with R to create interactive web pages. While this may sound scary due to the words "web pages," it's geared towards R users who have 0 web development experience and you don't need to know any HTML / CSS / JavaScript. With Shiny you can do a lot: think of it as an easy way to create an interactive web page, and that web page can interact seamlessly with R and display R objects (plots, tables, anything else you do in R). You can visit my Shiny site, which hosts some of my own Shiny applications, to get a sense of the wide range of things you can do with Shiny. Each Shiny app consists of two parts: a web page showing the user the app, and a computer powering the app. The computer running the app could be either your own laptop (such as running an app from RStudio) or another server. You need to write these two parts as the Shiny app developer (you're not going to write a computer, you're going to write the code that power the app). They're named UI (user interface) and database in Shiny jargon.

UI is just a web file that can be used by the user, it is HTML that you write using the functions of Shiny. It is the UI that creates the app's layout and tells Shiny exactly where things are going. The server is responsible for the application's logic; it is the set of instructions that tell the web page what to display when the user interacts with the page.

Fig. 3 Image of Shiny web application.
Once you want to support more than one user at a time, scoping is very important to understand in Shiny. Since your app can be hosted online, your app can be used by multiple users at the same time. If there are any variables that should be shared by all users (such as data sets or global parameters), then you can define them globally safely. But there should be no universal description of any parameter that should be unique to the session of each user.

II. LITREATURE SURVEY

[1] Mitali Gupta stated that Online ordering took a storm from the food business. Innovation has buried the business sector, technology has changed the whole restaurant industry system, and it will continue to do a great job. A technologically built digital food ordering system has radically changed the culture of the restaurant and is bringing people around the globe a new amazing comfort zone. The key goal. In India, there are a couple of food delivery applications that can be downloaded from advanced cells to arrange food in a hurry and from home comfort. The typical Indian's changing urban lifestyle is emotional enough to be ideal to evolve at higher rates for the food-on - the-go and fast home delivery models.

[2] Author says that data and method are the most important factors in order to make precise predictions and provide expert recommendations. We came across the fact that Facebook and Yelp are most successfully used datasets by thoroughly analyzing the literature. We also present a survey of various techniques and advantages that have been used. The recommendation system for restaurants can provide users with accurate and effective information about restaurants based on user profile information and preferences. This paper analyse existing techniques ' different approaches and challenges of existing techniques.

[3] Anupriya Saxena stated that we would understand the drivers of online food sites from this research paper. Different services provided by application that make consumer happy and satisfied. Comfort and convenience that make consumers more inclined to order food online. Doorstep Delivery at any place at any time is the feature that attracts the most customers. Customers are mostly inspired when they earn some Rewards & Cashbacks followed by loyalty points or benefits.

[4] Author explained aim of this paper is to provide insights into the marketing habitude of a web-based business which influenced people to change their understanding of food ordering and to explore the dimension by which the organization managed the processes of this growing field. This paper includes a case study on Zomato, India's largest online platform for food ordering. Zomato became a strong brand of all kinds. Despite the presence of Zomato's difficulties, however, it had to resolve those in order to operate effectively in the international arena, Zomato had performed quite well on the market, which is why it is considered one of the best online food delivery companies in the world. The business model of Zomato is perfect for meeting the needs of both current and prospective customers.
[5] Author stated that food tourism has recently been on the rise, along with the growth of social media support for food and restaurants. Nevertheless, there is often a shortage of funding hampering the process of making food tourism open to more audiences. The author will discuss the implementation of the branch and bound algorithm to optimize food tourism planning with the help of Zomato A as a 0/1 knapsack problem. Jakarta has an average price overall more expensive culinary attraction, while Bandung is more affordable. Branch and bound algorithm proved to be efficient in solving 0/1 knapsack problem from the practical perspective. Integrating Zomato API with 0/1 knapsack resolution allows the user to plan food tourism with their desired budget in advance.

[6] Author described about new method is proposed in this paper to find the better initial centroids and provide an efficient way to assign data points to appropriate clusters with reduced time complexity. The proposed algorithm has the more accuracy with less computational time compared to the original k-means clustering algorithm, according to our experimental results. One of the most popular clustering algorithms is the k-means clustering algorithm, but this method relies heavily on the initial centroids, which are selected randomly, for the quality of the final clusters. In addition, the k-means algorithm is also very expensive in computational terms. Compared with the original k-means algorithm, the proposed algorithm is found to be more accurate and effective. This method has been proposed.

[7] In this paper author performed an experimental study to apply the iterative initial-point refinement algorithm of Bradley and Fayyad to the clustering of k-modes to improve the accuracy and repeatability of the clustering results[ cf. P. Bradley, U. Fayyad, Refining K-mean Clustering Initial Points, in: Proceedings of the 15th International Machine Learning Conference, Morgan Kaufmann, Los Altos, CA. Experiments show that using refined initial points, the k-modes clustering algorithm results much more efficiently than the random selection approach without refinement, thereby making the refinement procedure relevant to many categorical data-mining applications.

[8] Author suggests a method for making the algorithm more efficient and effective to achieve better clustering with reduced complexity. This paper provides an improved k-means algorithm that combines a comprehensive approach to defining initial centroids and an efficient way to assign data points to clusters. This method guarantees the complete cycle of O(n2) time clustering without losing the precision of the cluster. The previous improvements of the k-means algorithm compromise either on accuracy or performance. A drawback of the proposed algorithm is that, irrespective of the distribution of data points, the value of k, the number of desired clusters, must still be supplied as output. For future research, it is suggested that certain statistical methods should be developed to determine the value of k, based on the distribution of data.

[9] Author proposed in this paper an algorithm for calculating initial cluster centers for the algorithm k-means. For illustrative purposes, the algorithm is applied to several
different datasets of different dimensions. The newly proposed algorithm is observed to have good performance to obtain the initial cluster centers for the algorithm k-means. The proposed algorithm is very effective, converges to better results of clustering, and almost all clusters contain some data. In contrast with the random initial cluster centers, experimental results indicate improved and stable cluster structures. However, according to the previously proposed algorithm in the literature, the new algorithm is much simpler and easier to implement.

[10] Author explained algorithms, a real click-stream data set is used to derive three typical user interest patterns. More specifically, more calculations based on different click-stream data sets check accuracy and reliability of these three interest patterns. This study shows that the proposed algorithm and the derived interest patterns can provide significant website optimization assistance and personalized recommendations. The method predicts the ratings of users on each item and recommends to the user top-rated items. The CF recommendation could be refined by basing it on the browsing behaviour of users and preferences of users, given the novel approach proposed in this paper. Firstly, the preference of user u on item I can be formalized according to the user's browsing behaviour, including browsing sequence, frequency and duration. Then, using the proposed rough leader clustering algorithm, the top k similar users will be selected. Finally, the preferential information of similar top k users can be used to predict user u preferences, and some common interest items are recommended to user u. In the meantime, the recommendation's accuracy should be assessed and compared to that of the traditional rating-based CF method. To this end, we can use the two recall and accuracy indices that are commonly used to measure the quality of recommendations.

[11] Author explained about to create a web application that clusters data on the hotspots. Using Shiny web framework for R programming language, this application implements DBSCAN algorithm. In 2002-2003, clustering takes place on a hotspot data set on Kalimantan Island and South Sumatra Province. The spread pattern of hotspot resulting from this clustering can be used as a predictive model of occurrence of forest fires and can be accessed via the internet browser. This research developed a web-based application clustering using the R programming language with a Shiny framework using the DBSCAN algorithm.

[12] Author described work is to develop a web-based application using the R programming language Shiny framework. This application provides several functions including summary and visualization of selected data, clustering of hotspot data using k-means algorithm, visualization of clustering results and sum square error (SSE), and display of global and collective outlier and visualization of outlier spread on Riau Province Map. Web-based hotspot detection applications using k-means clustering algorithm and Shiny framework provides several major features. First, the selected data summary and plot.
Second, clustering data from hotspot with the algorithm k-means. Third, the clustering results are visualized. All of the main function that works well on the application.

III. Proposed system approach

The viable system approach proposes an in-depth analysis of the dichotomy of structure/systems, suggesting that each system represents a recognizable entity emerging from a specific changing structure (set of individual elements with assigned roles, activities and tasks in accordance with rules and constraints).

The flow chart of proposed system approach is given in figure. In diagram each block show a process and describe the working.

![Flow chart of Project](image1)

**Preparation of data set.**

The data is gathered from the kaggle.com, you can also collect data in zomato web with the help of API zomato key.

![Zomato API web site](image2)
Here you can download the all over Indian restaurants zomato data.

1. **Parameter of the data set.**

   Statistical parameters are an important part of any statistical analysis. Simply put, a parameter is any number quantity that characterizes or some aspect of a given population. This means we are told something about the entire population by the parameter.

![Fig.6 Structure of data set.](image)

2. **Making the model of k-means clustering algorithms.**

   To perform this algorithm we required some packages are like tidyverse, dplyr, ggplot, tidytext.The K_means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the starting points for each cluster, and then performs iterative (repetitive) calculations to optimize the centroid positions.

   - If one, it avoids building and optimizing clusters.
   - The centroids have stabilized — there is no change in their values because the clustering has been successful.
   - The defined number of iterations has been achieved.

   **K-Means Clustering Pseudocode**
   - Choose the cluster(K) number and get the data points
   - Place the c 1, c 2,..... C k 3 at random.
   - Repeat steps 4 and 5 until convergence or until a fixed number of iterations 4 has been completed.
   - For each data point x_i:
     - Find the centroid closest(c 1, c 2 ..... c k)
     - Assign the item to the cluster
   - For each cluster j = 1..K
     - New centroid = average of all given points to that level.
• End
There would be some cases where the number of clusters would not be identified. So how can we pick the K value?? There's a method called the process of the elbow. You pick a different number of clusters in this process and start plotting the distance within the cluster to the centroid. The graph looks like the one below.

![Code for K-means Model](image1)

Fig. 7 Code for K-means Model

![Elbow method graph of K-means cluster](image2)

Fig. 8 Elbow method graph of K-means cluster

An average score for all clusters and the total square distance from each point to their allocated center is measured above the map.

3. Visualization of data.
Visualization of data is an art of turning numbers into practical knowledge. R Programming helps you to learn this art by offering a collection of built-in functions and libraries for creating visualizations and presenting information.
We understand the common cuisines in a specific Indian city data set from fig.10.

4. Geo map of India.

A map is a visual representation of a place's selected features, usually drawn on a flat surface. Maps present world information in a simple, visual manner. By showing country sizes and shapes, locations of features, and distances between places, they teach about the world. Here we get information about the restaurant information.
Fig.11 Code for GEO Map.

Fig.12 Geo Map of pune data set.

Fig.13 shows us what kind of restaurants are located in a particular region and what kind of food is served there as well as details of restaurants.

5. **Shiny User Interface.**

The R language is widely used in the design of statistical technology and data analysis by statisticians and data miners. While R has an interface for command line, there are a variety of graphical user interfaces, such as an integrated development environment, R Studio.

Fig. 13 Shiny User Interface.
From fig.14 we are able to understand the result of our data with the help of a brilliant web application, so the brilliant user interface helps us easily to visualize the data set, the data structure, the summary of data sets and the diagram or graph displayed.

IV. Future Scope And Discussion.

This project helps zomato customers who can easily order their food, flexible payments, real-time tracking, and zomato web application loyalty points. It can be used to expand the business through analysis and observation, and they can achieve their goals such as financial strategy, i.e. fund and sales, marketing strategy to increase their clients, globalization strategy, i.e. spreading around the globe.

V. Conclusion

Following my completion of this work, we can conclude that clustering k-means is a commonly used machine learning algorithm. It can be used in various data sets and grouping types. As a user interface, we used shiny app and grouped the zomato data set. We used Geo map to define the restaurants’ location. We found in my final result that we are getting a good representation of clusters and a better model that delivers better results.

VI. Reference

2. Josh i, S., Dubey, J., graduate Scholar, P. and Vishwavidyalaya, S.V.V., A Survey on Restaurant Recommendation System Based on Feature Selection and Classification Method.
5. Wirasuta, I.P.G., Branch and Bound Algorithm Implementation on Zomato Restaurant Data for Food Tourism Planning.
