TUNKU ABDUL RAHMAN UNIVERSITY OF MANAGEMENT AND TECHNOLOGY

FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY

Western Specialty Restaurant Location Recommendation in Singapore

BMCS2114 MACHINE LEARNING

2023/2024

Student's name/ ID Number	:	RYAN KHO YUEN THIAN / 2204097
Student's name/ ID Number	:	THONG CHENG HOW / 2203154
Student's name/ ID Number	:	ONG WENG KAI / 2203309
Student's name/ ID Number	:	YONG ZEE LIN / 2203770
Programme	:	Bachelor of Computer Science in Data Science
Tutorial Group	:	2
Tutor's name	:	DR LIM SIEW MOOI

ABSTRACT

The competitive market demand and small size of Singapore make it challenging to start a restaurant there. Therefore, choosing a site is essential for our client hoping to launch a restaurant in Singapore. The purpose of this project is to help our client choose the best site for a new Western restaurant in Singapore. This project explored five clustering algorithms with varying performance measures to determine which cluster data points to group together.

With Kaggle acting as this project's main data source for our dataset, we created a Streamlit application, which is featured in the results section, to help visualise the clusters and pinpoint the area that gets the highest ratings. This helped find the best places to open the new western restaurant in Singapore.

Keywords: Western, Restaurant, Singapore, Food, Grab, Clustering, Location, Place

1) INTRODUCTION

i) Problem Statement

Choosing the right location is crucial when launching a business as it significantly influences the success of the business (Indarti, 2004). The wrong location can lead to repercussions such as lower sales, brand damage, a slow ramp-up time and the need for additional advertising (Lowder, 2017). Several factors should be considered when selecting the optimal location for one's startup. One factor is the market, where one must ensure that there is demand for one's product/service in the area being considered. It is also crucial to analyse the competition in the area. If numerous businesses are already offering similar products or services, achieving success might prove more challenging (FasterCapital, 2024).

We are tasked with assisting a client to select a location to set up a new **Western** restaurant in Singapore's competitive culinary landscape. The client would like to consider areas, where there is demand for **Western** food BUT the existing **Western** restaurants in those areas have Low Ratings. This gives some assurance to the client that the new restaurant will have customers and that the client has a good chance of competing against the existing **Western** restaurants.

Research question: Where should the new Western Restaurant be located in Singapore?

ii) Solution

To satisfy the client's demand, we will use a dataset from Kaggle, which contains data about Grab restaurants in Singapore. GrabFood is one of Singapore's leading food delivery service providers offering a variety of cuisines (Tan, 2024). Given Grab's status, the dataset should contain almost all the significant restaurants in Singapore.

Since the dataset is unlabelled, we will use an Unsupervised machine learning approach called Clustering. Clustering is the process of grouping data points into multiple groups so that the data points within each group are more similar to one other and less similar to the data points outside of each group (Jeldu & Tadele, 2018). We will be leveraging several clustering algorithms, namely: K Means, BIRCH, Agglomerative Hierarchical Clustering, DBScan, and Affinity Propagation.

After cleaning and pre-processing the dataset, we will filter out those restaurants that do not offer **Western** food. We then apply the selected clustering algorithms to discover clusters of restaurants. By analysing summary statistics and visualisations of the clusters, we can identify those clusters which fit our client's requirements. After the client has decided on the cluster(s), we can perform further analysis especially regarding competitors in the selected cluster(s).

iii) Objectives

- To clean and preprocess the Grab Restaurants dataset to ensure data integrity and consistency.
- To perform Exploratory Data Analysis (EDA) on the cleaned dataset to understand the data.
- To apply 5 clustering techniques (K-Means, BIRCH, DBScan, Agglomerative Hierarchical Clustering, and Affinity Propagation) on the data.
- To identify and visualise the distinct restaurant clusters formed.
- To assess the performance of 5 clustering techniques by using appropriate clustering performance metrics.
- To display the locations of the clusters on the Singapore map
- To interpret the clusters and recommend the best location(s) for the client.

2) LITERATURE REVIEW

1a) What is K-Means

Belonging to the category of partitioning-based clustering techniques, the aim is to find a set number of clusters or partitions that reduce the total sum of squared Euclidean distances between data points and the centroids of their respective clusters. But, the number of clusters (k) needs to

be established beforehand. Its performance can be influenced by the initial seed selection, making it vulnerable to getting stuck in local optima and potentially overlooking the global optimum. This method may converge towards less-than-ideal solutions. Moreover, since the mean value can be significantly impacted by a small amount of outliers, K-Means is sensitive to outliers. Furthermore, it cannot be used for clustering problems whose results cannot fit in the main memory (very high dimensionality dataset or an extremely huge desired number of clusters) (Sudhir Singh et al., 2013).

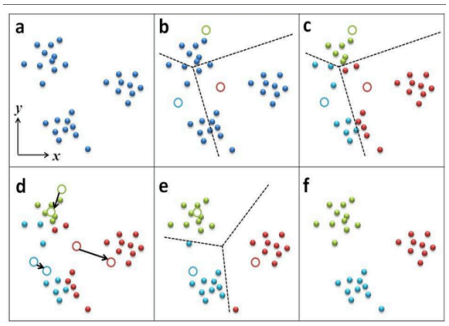


Figure 2.1 K-Means Algorithm

1b) Using K-Means for Recommending Restaurant Location

A useful application of K-Means clustering was showcased in a project by Kumar Shaswat (2020), which focused on identifying ideal localities for launching an Indian eatery in Delhi by using features, such as venue longitude, venue latitude and venue category. Kumar Shaswat (2020) stated that the ideal neighbourhoods would be those that have low density of Indian restaurants, have demand for Indian restaurants and show promise for growth, which can be identified by using heatmaps and clustering the neighbourhoods based on how similar their restaurant trends are. Kumar Shaswat (2020) used K-Means to cluster the neighbourhoods into 5 clusters but did not justify why he chose the number of clusters to be 5 and did not explore other clustering algorithms.

1c) Topic Modelling and Clustering Restaurant Areas in Vancouver

Lee (2021) aimed to investigate the geographical distribution of particular cuisines and whether discernible patterns exist in the clustering of restaurants. This would aid travellers in pinpointing attractive dining locales for specific cuisines and restaurateurs in identifying potential entry points while avoiding saturated areas. His Yelp restaurant dataset contained information like the restaurants' geographic location, cuisine types and a Bayesian estimate of the Mean Star Rating.

K-Means Algorithm (see Figure 2.1*)*:

1. Randomly select k items from a set of items to be clustered as the initial cluster centroids
2. Assign each item to the cluster it is most similar to by comparing it to the average value of items in the cluster.
3. Update the cluster centroids by computing the average value of the items within each cluster.
4. Repeat steps 2-3 until there is no change in the cluster

To generate labels identifying the culinary neighbourhoods, Lee (2021) used K-Means to cluster on the restaurant's location, using scikit-learn's silhouette score to optimise the labels' quality. He used K-Means the 2nd time when he had converted each neighbourhood into sets of cuisine groups (after applying NMF for topic modelling). This involved (based on the greatest reduction in inertia) grouping neighbourhoods according to their weights for each cuisine group to unveil patterns in the representation of cuisine groups across different neighbourhoods.

Performance metric used: **Silhouette Score**

2a) What is BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

BIRCH is a clustering algorithm that can deal with large datasets by first creating a condensed summary that preserves the majority of the distribution information and subsequently clustering the summary. Besides explicitly addressing time and memory limitations, BIRCH also capitalises on the insight that not all data points are equally significant for clustering, thus storing compact summaries for dense regions. This approach shifts the focus from clustering the original data points to clustering a much smaller set of summaries. These summaries produced by BIRCH capture the inherent proximity of data, enabling the calculation of distance-based metrics and can be maintained progressively and efficiently (Zhang et al., 1997).

Overall, this algorithm comprises 4 phases: Loading, Optional Condensing, Global Clustering and Optional Refining, which are illustrated in Figure 2.2.

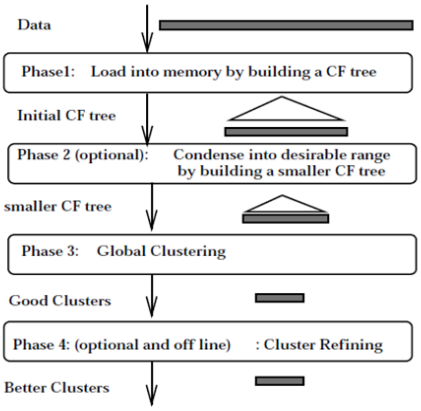


Figure 2.2 BIRCH algorithm

A drawback of BIRCH is that it can only handle metric attributes, whose values can be depicted using specific coordinates within a Euclidean space (Zhang et al., 1997).

2b) A segmentation analysis mapping tool for the energy sector

In this study, Liu et al. (2022) used the BIRCH algorithm to apply customer segmentation to analyse the daily energy consumption of residential buildings, at the individual and at the neighbourhood level. The resulting segmentation analysis platform, called SEGSys, offers an intuitive decision support tool for monitoring energy demand. Liu et al. (2022) selected BIRCH for several reasons: it is capable of identifying anomalies associated with irregular and scattered user behaviour, has a high efficiency, and a low memory footprint can be used to detect extreme values like extremely high or zero values. Moreover, BIRCH excels with big data, outperforming other methods like K-Means and EM clustering, and is ideal for high-performance applications and large datasets like IoT data. Unlike K-Means, it doesn't need the number of clusters as input and can detect extreme values as anomalies (Liu et al., 2022).

Performance metrics used: **entropy, standard deviation of cluster sizes, estimated threshold**

2c) Recommendation for Location of Digital Signage using Fusion of Multiple Information Sources

Xie et al. (2018) developed a sustainable model for suggesting suitable locations for digital signage by combining the spatial attributes of geographic areas with various data features from multiple sources. The study used several clustering algorithms (BIRCH, SOM, K-means and DBSCAN) that were used to split the study area into regions. Results showed that BIRCH had the second highest Calinski-Harabasz Index. Its recommendation quality is poor with a small recommendation score but its recommendation quality increases when the recommendation score threshold is huge (Xie et al., 2018). Among the 4 algorithms, the SOM algorithm had the highest Precision and Recall while the rest of the algorithms, including BIRCH, did not significantly affect the recommendation results.

Performance metrics used: Calinski-Harabasz Index, Maximum Information Coefficient, precision, recall, F-measure

3a) What is Affinity Propagation (AP)

AP is a type of unsupervised learning algorithm that does need to know the number of clusters beforehand because it figures it out automatically. It works by exchanging messages between clusters and data points, using different similarities between data points. It treats every data point as a possible centre for a cluster. Its main goal is to find the most representative cluster and group data points. To manipulate the data points, matrices are needed for identifying the relationship of each data point in a concise manner. (geeksforgeeks, 2023; Zhang & Gu, 2014).

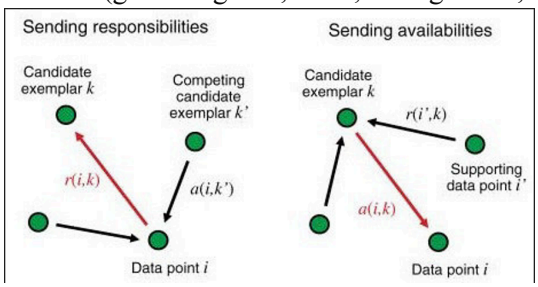


Figure 2.3: Concept of message passing (AP)

Concept of message passing

Message passing in AP is a key mechanism through data points sharing the same similarity to indicate the main node. It utilises "exemplars", which represent other data points within the same cluster. All nodes will eventually form groups. (R. Refianti et al., 2017)

1. **Similarity Matrix (S)**
   We figure out how much alike data points are by looking at their features, calculating 'similarity score' for each pair of points. (R. Refianti et al., 2017)
2. **Responsibility Matrix (R)**
   This matrix shows how good one data point is at being the main example with another data point. The value R (i, k) tells us how well data point 'i' fits as the main example for data point 'k'. (R. Refianti et al., 2017)
3. **Availability Matrix (A)**
   The availability matrix shows how ready each data point is to be the main example for others. It determines which data points are suitable to lead the clusters. (R. Refianti et al., 2017)

Key steps of Affinity Propagation (geeksforgeeks, 2023)

1. **Similarity Calculation**: The algorithm starts by figuring out how similar.
2. **Responsibility Calculation (R)**: Each value R(i, k) shows how much responsibility data point 'i' has to be the main example for data point 'k'.
3. **Availability Calculation**: Data points choose who will be the main exemplars by comparing the highest availability.
4. **Iterative Update**: The algorithm repeatedly updates the responsibility and availability matrices until they stop changing significantly.
5. **Net Responsibility Calculation**: calculates the net responsibility for each data point by adding up the responsibility and availability.
6. **Exemplar Selection**: Data points with high net responsibility. These examples become the centres of clusters.
7. **Cluster Assignment**: Each data point is assigned to the nearest exemplar based on similarity.

3b) Food recommendation system using machine learning for diabetic patients

Model	Cluster Size	Nutrient Score	Group Score	Speed
KMeans	17	46%	0%	1.22
Affinity Propagation	41	30%	0%	13.63
SOM	117	58%	90%	4.31

Table 01: The performance of clusting algorithms compared to baseline model

Figure 2.4: Performance of the algorithms

In this system, Phanich M et al., 2021 developed a food recommendation system using clustering analysis for diabetic patients. Foods are labelled into 3 groups which are normal food, limited food, avoidable food. AP was used as a substitution system to analyse other algorithms that may not cluster the food dataset properly. Although k-means algorithm took a shorter time to cluster, the number of clusters was not huge. AP took a longer time but the cluster size was almost triple of K means cluster size. Although the number of clusters for SOM was more than AP, AP also was equally performing and performance wise was similar. SOM was chosen by the researchers.

3c) Multi-stage hierarchical food classification

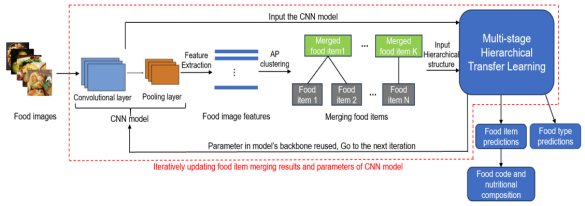


Figure 2.5: Illustration of food classification

Pan et al. (2023) developed a new dataset called VFN-nutrient by classifying each food image with specific food ingredients that contain the nutrient information. AP was used as one of the frameworks for food item classification. It was used for clustering, merging the food images during the training process. The researcher chose AP because the clustering did not need a predefined number of clusters. Therefore, the researchers can save time from identifying the number of clusters. From the above image, CNN was used for predicting the food images.

4a) What is Agglomerative hierarchical Clustering

Agglomerative Hierarchical Clustering (AHC) works by a bottom-up approach, starting with each data point as its cluster and merging them into larger clusters based on their similarity. This process involves several key steps:

1. *Initialization:* Each observation starts as its cluster.
2. *Proximity Matrix Computation:* Calculate the proximity matrix to determine the distances between all pairs of observations.
3. *Linkage*: Using a linkage criterion, determine which clusters to merge based on distance. Standard linkage methods include Complete Linkage, Single Linkage, Average Linkage, Centroid Linkage and Ward's Method.
4. *Merge*: The two closest clusters are merged into a single cluster.
5. *Update*: The proximity matrix is updated to reflect the merger.
6. *Repeat*: Steps 3 through 5 are repeated until all observations belong to a single cluster (Benhur, 2023).

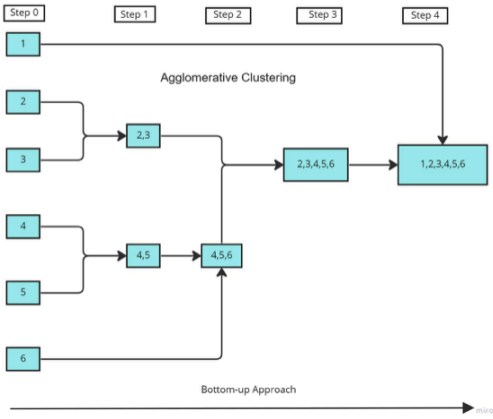


Figure 2.6: How Agglomerative Clustering Works

Hierarchical clustering typically produces results showcased in a dendrogram. Generally, the computational complexity of AHC is $O(n^3)$, rendering it inefficient for handling large datasets (K.Sasirekha & P.Baby, 2013).

4b) Implementing AHC using Food Ingredient Classification According to Nutritional Composition

The project is focused on enhancing health management through dietary insights by building a system that

categorises food based on its nutritional content. The system employs AHC with average linkage to analyse and categorise foods. This methodological approach is aimed at computing the average nutrient values across different food categories, offering an advanced tool for nutritional guidance. The AHC with the average linkage method operates by calculating the average distance between all object pairs in different clusters, providing a basis for merging clusters in a hierarchical fashion until a final structure is achieved. This approach allows for the creation of nuanced and informative classifications of food items based on their nutritional content, offering valuable insights for dietary planning and health optimisation. (Dalimunthe, 2021)

4c) Applying cluster analysis and preference mapping to assess consumer preference for different cooking temperatures of beef steaks.

Schmidt et al. (2010) evaluated consumer preferences for steaks from the Longissimus lumborum muscle, cooked to various degrees of doneness, and the relationship between these preferences and specific demographic characteristics. AHC, employing Euclidean distance and Ward's method, was used to group consumers based on their preference for the end-point temperature of steaks. This method enabled the segmentation of consumers into clusters with distinct preferences, enhancing the understanding of how different demographic groups and their desired degrees of steak doneness influenced their overall acceptability ratings. If notable distinctions arose among treatments within a cluster, the Least Significant Test (LSD) test was conducted to distinguish between treatment averages. External preference mapping was also applied to the descriptive data and consumer acceptability scores, bridging the gap between sensory attributes and consumer likes.

5a) What is DBSCAN (Density-Based Clustering of Applications with Noise)

Being a density clustering algorithm, DBSCAN is widely used for clustering in machine learning and data mining. It defines the clusters as dense regions and separates them by areas of low density. It can segment connected & high-density data into arbitrarily shaped clusters, handle noise effectively and compute efficiently (Chen et al., 2019).

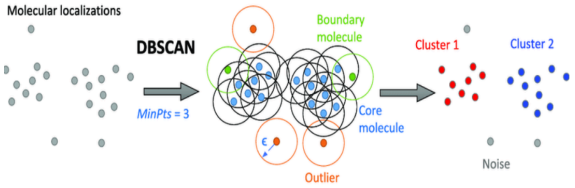


Figure 2.7 An example illustrating DBSCAN

It requires 2 parameters:

- **Epsilon (ε):** It defines the radius of the neighbourhood around a point.
- **MinPts:** It specifies the minimum number of the points within the ε-neighbourhood to consider a point to a core point.

How it works (Ester et al., 1996):

1. **Input Parameters:** Define parameters i.e. epsilon (eps) and minimum points (minPts).
2. **Identify Core Points:** Calculate core points using eps and minPts.
3. **Cluster Expansion:** Choose core points randomly and expands the cluster by adding neighbouring core points to it. This step will continue until there are no more core points within distance ε to add to the cluster.
4. **Handle Border Points:** Assign these points to the clusters that it belongs to. It falls within the ε-neighbourhood of a core point.but not core point.
5. **Handle Noise Points:** These points are not assigned to any cluster.
6. **Format Cluster:** Identifying core points, expanding cluster, assigning border points, and handling noise continues until all points in this dataset have been processed.
7. **Output:** The output of DBScan is a set of clusters. It will contain a group of points that are densely packed together. Furthermore, there will also be a set of noise points that do not belong to any cluster.

5b) Finding the Optimal Restaurant Location in Düsseldorf Germany

Xu (2021) aimed to find the optimal location for opening a restaurant in Düsseldorf by identifying an area with high population density and minimal nearby restaurants. Data, such as the restaurants' coordinates in each neighbourhood and the neighbourhoods' coordinates were used. To cluster the restaurants, Xu (2021) used DBSCAN, stating that it is particularly suited for arbitrarily shaped clusters, is highly resilient to outliers and does not necessitate specifying the number of clusters. The algorithm produced 8 clusters and

helped to identify restaurants that were outliers. However, Xu (2021) noted that since DBSCAN utilises the parameter min_samples, duplicates will inflate the count of data samples within the specified radius (epsilon), potentially impacting the results and making it necessary to remove duplicates.

5c) Opening a new pizzeria in Turin using DBSCAN

Huseynov (2021) stated that besides considering the premises' rental prices and overall availability of suitable spaces, the optimal location would have low density of restaurants and high density of residents. He used data about, for example, the city administrative division, zones geolocation and restaurants. Huseynov (2021) opted to cluster zones based on their proximity to nearby restaurants to ensure avoidance of duplication in the same category, prioritising this as a key criterion. This was achieved by using DBSCAN for density clustering, which differs from K-Means, primarily addressing the issue of centroid allocation. To visualise the best location on the map generated by the folium library, he used the alphashape library to create blue polygons representing areas occupied by pizzerias and marked outlier restaurants using icons.

3) RESEARCH METHODOLOGY

i) Overview of Methodology

Figure 3.1 shows an overview of the methodology.

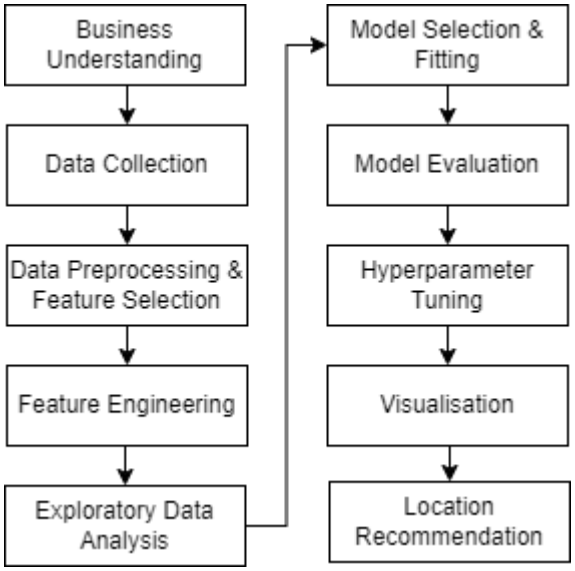


Figure 3.1 Methodology Overview

ii) Data Collection

The dataset was obtained in csv format from Kaggle, consisting of 16136 records and 19 features. Some columns were renamed for understandability.

Features	Meaning
id_source	Restaurant unique identifier
name	Restaurant Name.
address	Location of the restaurant
country	Country where the restaurant is located.
cuisine	Type or style of cuisine offered by each restaurant
currency	Currency used, such as SGD (Singapore dollars)
delivery_cost	Starting cost for delivery should be divided by 100.
latitude	Restaurant latitude coordinate
longitude	Restaurant longitude coordinate
opening_hours	Restaurant's regular opening hours
image_url	URL of the restaurant's image or logo.
radius	Distance between the restaurant and the searched location.
rating	Overall rating by users
no_of_reviews	Total number of reviews received.
delivery_options	Various delivery methods available.
promotion	Promotion used by customers.
loc_type	Type of store location.
delivery_by	Method used to deliver the food, which is either grab or merchant.
delivery_time	Estimated average delivery time.

Figure 3.2 Dataset Features Overview

iii) Data Preprocessing & Feature Selection

The features ***id_source, country, currency, image_url, radius,*** and ***delivery_by*** were dropped as they were not relevant to the problem domain. Records belonging to the 'FOOD' ***loc_type*** were kept, while those corresponding to the 'MART' ***loc_type*** were removed as the project's focus is on restaurants. The cuisine feature had missing values,

which were handled, and it was later feature-engineered (as explained in the next section) to facilitate the filtering of only Western restaurants. The ***name*** feature values were lowercased and trimmed, missing values were handled based on the address feature values, and inconsistent restaurant names were corrected. Given the impracticality of handling the extremely varied ***address*** values directly, we used the latitude and longitude values to extract address details, such as suburb, road and postcode using the GeoPy API (as explained in the next section). Missing values in both the ***delivery_cost*** and ***delivery_time*** features were replaced with the median, and outliers were handled using winsorization. In the ***rating*** column, restaurants with invalid ratings like -1 or missing ratings were removed as imputing missing or invalid ratings may create artificial data points that do not accurately reflect restaurant ratings, compromising dataset integrity. We imputed the value 0 for restaurants with missing values in the ***no_of_reviews*** feature. Finally, label encoding was applied to the ***delivery_options*** and ***promotion*** features. The ***delivery_options*** feature has a relatively small number of unique categories, and label encoding helps to indicate whether a restaurant is offering a promotion or not.

iv) Feature Engineering

For the ***cuisine*** feature, we lowercased the values, handled the missing values and applied *multilabel binarizer* as each value in the feature is a list of cuisine types. Lastly, using the newly created columns, we filtered out the restaurants that are specifically related to western cuisine.

In order to extract the various opening hours that are joined together from the ***opening_hours*** feature, we used a user-defined function to extract each day of the week's opening hours from each JSON and a lambda function, together with pandas dataframe, to extract the data and create new columns for the opening hours of the week. Finally, we applied label encoding due to the variations in opening hours.

v) Exploratory Data Analysis (EDA)

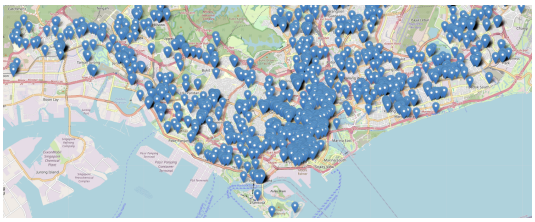


Figure 3.3 Visualising Competitor Locations on Map

Figure 3.3 allows us to visualise the locations of the Western Restaurant Competitors. We can clearly see that most of the restaurants are in the more populated areas of Singapore.

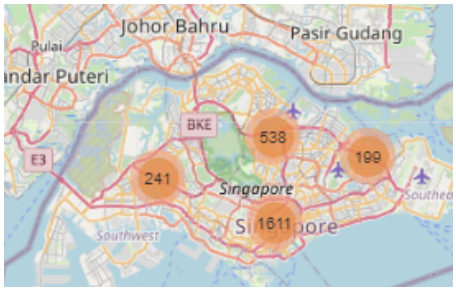


Figure 3.4 Grouping of Restaurants

Singapore has a total of 2589 western restaurants, more than half of which are located in the south area (Figure 3.4).

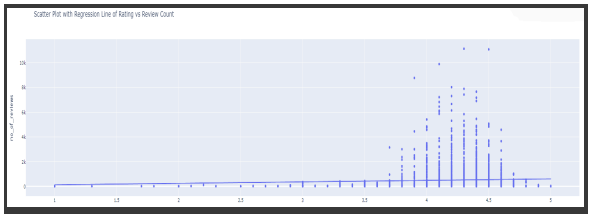


Figure 3.5 Scatter Plot with Regression Line of *Rating* VS *No of Reviews*

According to Figure 3.5, the higher the rating, the higher the number of reviews, indicating a positive correlation between the features.

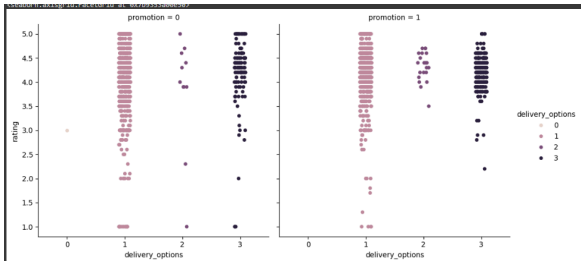


Figure 3.6 Categorical Plot of *Delivery Options* VS *Rating*, Investigating the Effect of Offering Promotions

Note:

0 = Delivery & Dine-in

1 = Delivery & Takeaway

2 = Delivery, Takeaway & Dine-in

3 = Only Delivery

Referring to Figure 3.6, the left side represents restaurants without promotions, while the right side shows those with promotions. Most offer both delivery and takeaway and have a rating of at least 3.0. Promo availability doesn't significantly impact delivery options or ratings.

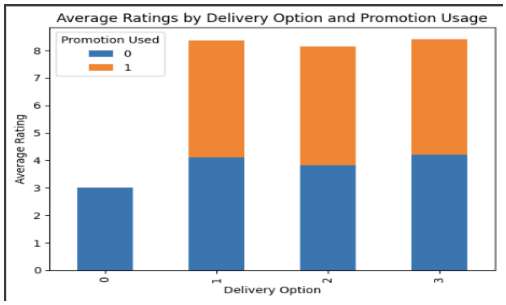


Figure 3.7 Stacked Bar Chart showing Average Ratings by Delivery Option and Promotion Usage

Based on Figure 3.7, delivery options 1, 2, and 3 enjoy higher customer satisfaction ratings, unaffected by the introduction of promotions.

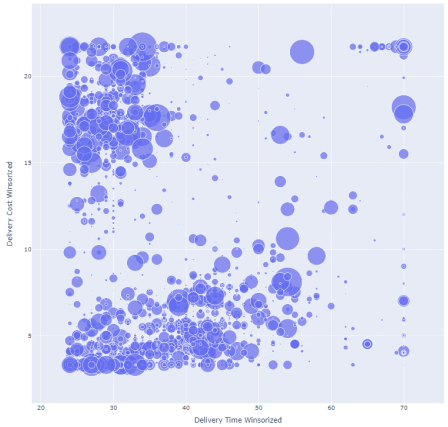


Figure 3.8 Bubble Chart of Delivery Cost vs Time, Sized by Number of Reviews

Referring to Figure 3.8, a dense cluster of data points around the lower to mid-range delivery time and cost indicate that most deliveries fall within this range. This could suggest that certain delivery times are typical, but the associated costs can vary significantly.

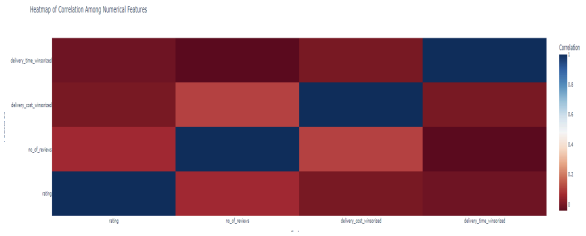


Figure 3.9 Heatmap Showing Correlation between selected numerical features

The features are most likely independent of each other as there is only a weak correlation (both positive and negative), indicating that there is little to no linear relationship between the pairs of features in the chosen features.

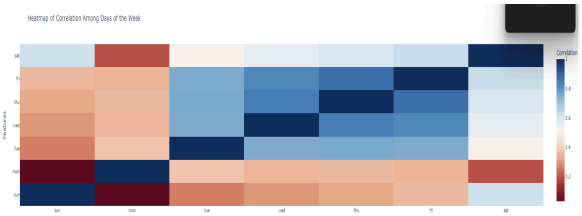


Figure 3.10 Heatmap Showing Correlation between Days of the Week

Wednesday and Thursday show a high positive correlation, it suggests that these days are likely to be busy for restaurants, indicating potential high revenue opportunities.

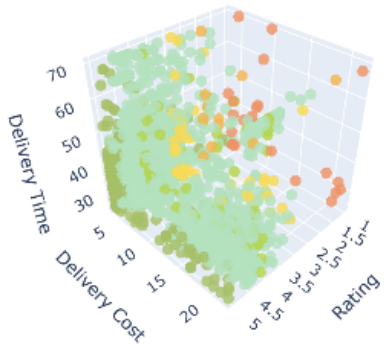


Figure 3.11 3D scatter plot of Rating VS Delivery Cost VS Delivery Time

Ratings don't really affect the delivery cost and time. Even if the delivery cost is high, it does not guarantee a faster delivery. The majority of the delivery times approximately range from 30 to 50 mins.

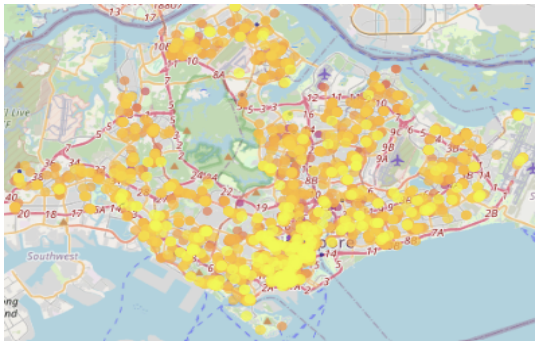


Figure 3.12 Scatter Mapbox based on rating on latitude and longitude

Referring to Figure 3.12, besides visualising the restaurant locations in Singapore, the Mapbox scatter plot shows that restaurant ratings are generally positive (most seem to be at least 3.8).

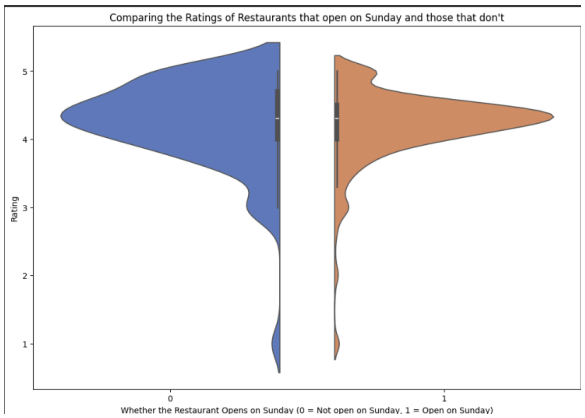


Figure 3.13 Violin Plot for Rating Distribution (Comparing Restaurants that open on Sunday and those that don't)

Figure 3.13 compares restaurant ratings based on Sunday opening status. Restaurants with higher ratings are more likely not to open on Sundays than those that do.

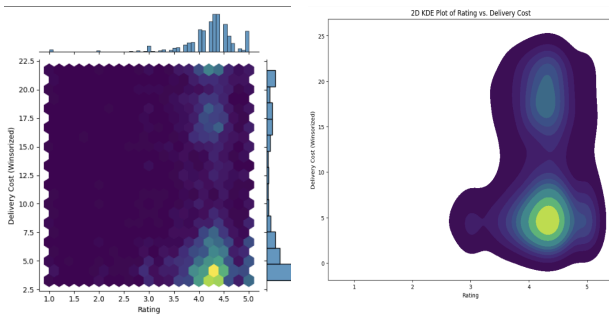


Figure 3.14 Joint Plot and KDE plot of Rating VS Delivery Cost

Figure 3.14 shows the relationship between restaurant ratings and delivery costs. Lighter regions indicate higher data point density. Based on this figure, we can see easily that the most dense area is between rating 4 and 4.5 and at a delivery cost between 2.5 to 5.0.

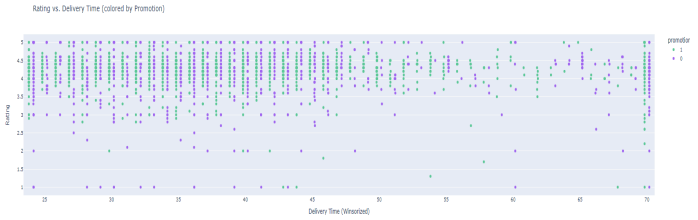


Figure 3.15 Strip Plot of Rating VS Delivery Time

Based on Figure 3.15, most of the data points are on the left side of the plot and that regardless of whether the restaurant offers promotion, the data point distribution is approximately the same.

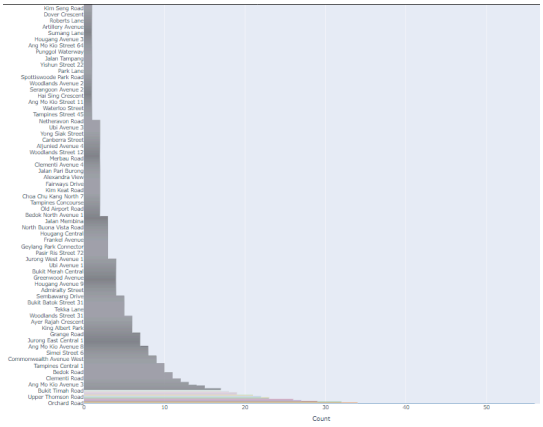


Figure 3.16 Horizontal Bar Chart for Road

Figure 3.16 compares the frequency of occurrence of different roads. Orchard Road has the highest frequency of 59 while Montreal Link has the lowest frequency of 1. This indicates that most of the western restaurants can be found on Orchard Road.

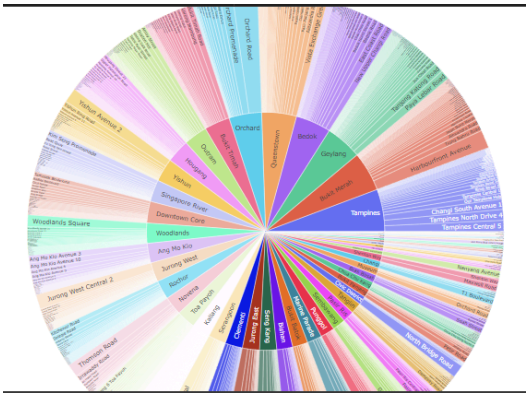


Figure 3.17 Sunburst Chart of Suburb-Road Combinations

Figure 3.17 visualises the distribution of counts for different combinations of suburbs and roads. Each hierarchical level is represented by one circle or ring. The root represents the top-level categories, which are the suburbs. The leaves represent the individual categories within each suburb, which are the roads. We can observe that most of the restaurants are situated at the Tampines suburb with a frequency of 148.

vi) Algorithms used

This project focuses on the K-Means, BIRCH, Agglomerative Hierarchical Clustering, Affinity Propagation, and DBSCAN algorithms. Initially, we fit the models with the dataset using default hyperparameter values. Subsequently, we fine-tuned them using techniques like GridSearchCV and RandomizedSearchCV. We evaluated their performance using five performance metrics explained in the next section. The clusters formed by each model are visualised on Folium maps.

vii) Performance Metrics

Since the dataset has no labels, we decided to use the following 5 performance metrics, which do not require ground truth or true labels.

- **Silhouette Score**: It measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates better defined clusters (Sauravkaushik8 Kaushik, 2023).
- **Davies-Bouldin Index**: It quantifies the average similarity between clusters. A lower Davies-Bouldin index indicates better clustering, with values closer to zero representing better separation (Artus, 2024).
- **Calinski-Harabasz Index**: It evaluates cluster validity based on the ratio of between-cluster dispersion and within-cluster dispersion. Higher values indicate better-defined clusters (Lindner, 2023).

- **Dunn Index:** It evaluates clustering by balancing cluster compactness and separation. It measures the ratio between the minimum inter-cluster distance and the maximum intra-cluster distance. Higher Dunn Index means a superior clustering outcome (Fatakdawala, 2023) .
- **Hubert & Levin C index**: Its aim is to assess how spread out clusters are compared to the total dispersion in a dataset. Ideally, the number of clusters that minimises the C index should correspond to the optimal number of clusters for partitioning the dataset (John, 2020).

viii) Feature Scaling & Dimensionality Reduction

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Figure 3.18 MinMaxScaler Formula

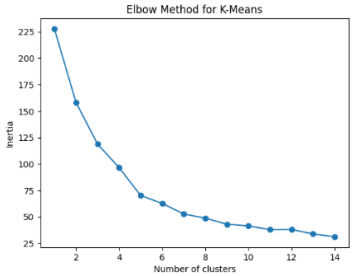
Before clustering, we'll focus on three columns: latitude, longitude, and rating. These values will be scaled using MinMaxScaler. Latitude and longitude fall within specific ranges, making MinMaxScaler ideal for preserving their spatial relationships. Similarly, rating values, ranging from 1 to 5, will be normalised to maintain their relative differences across the dataset. Dimensionality reduction isn't needed when clustering on just three columns because the dataset is already low-dimensional, making it easier to analyse without the need for additional techniques.

viv) Dashboard & Visualisation

We'll use Streamlit to visualise and interact with the clusters formed by the models. This dashboard will help identify the model with the clearest clusters and suggest optimal locations for the new western restaurant.

4) RESULTS & DISCUSSIONS

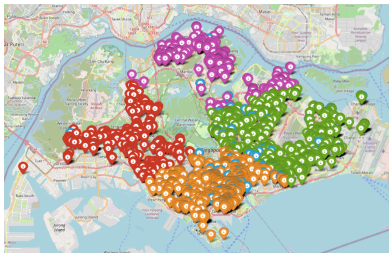
4.1 K MEANS



We used the elbow method to find the best K value for the k-means algorithm. From the graph, we see that the optimal K value is approximately 5, where the inertia begins decreasing linearly.

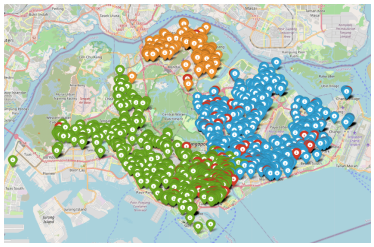
```
Optimal number of clusters (Silhouette Score): 5
```

This was further verified using a programmatic approach involving silhouette score.



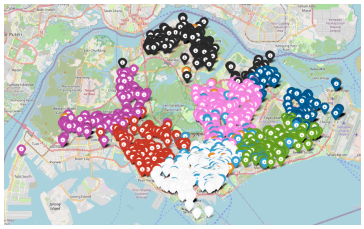
The K-Means model generated **5** distinct clusters, with the "blue" cluster overlapping with "orange", "green" and parts of "purple" clusters. Overall, the fine-tuned model replicates the clustering outcomes of the default model.

4.2 BIRCH



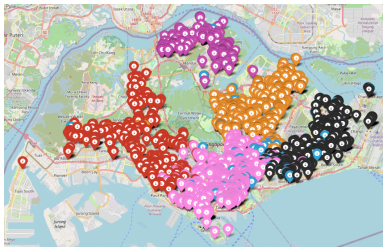
The fine-tuned Birch model identified **4** clusters. The "green", "blue" and "orange" clusters are clearly defined, but restaurants in the "red" cluster remain mixed with the others, particularly with the "green", "blue" and a few parts of the "orange" clusters.

4.3 Affinity Propagation



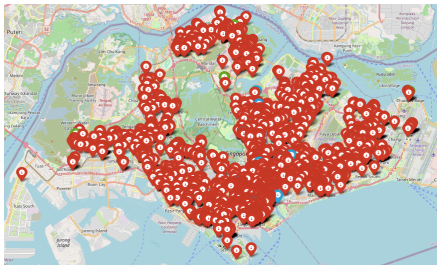
The fine-tuned Affinity Propagation model generated **9** clusters. Restaurants from the "light blue" and "orange" clusters intermingle with "red", “white”, "green" and "pink" clusters.

4.4 Agglomerative Hierarchical Clustering



The fine-tuned agglomerative hierarchical clustering yielded **6** clusters. The "red", "orange" and "purple" clusters are well-defined, but the "blue" cluster is scattered among the other three clusters.

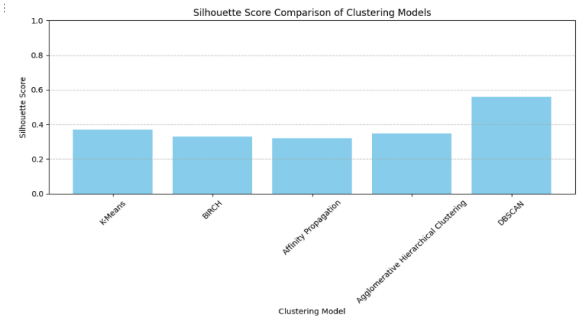
4.5 DBSCAN



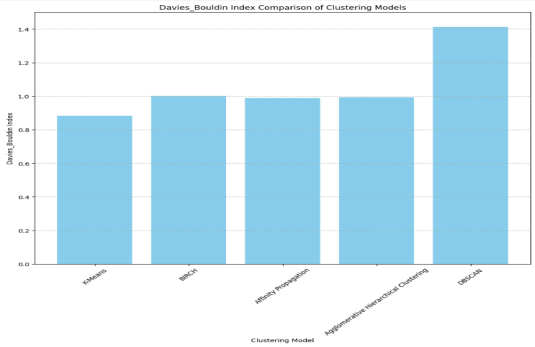
The fine-tuned DBSCAN generated **3** clusters. The "red" cluster is clearly defined, while the "blue" and "green" clusters appear in smaller groups scattered among the "red" clusters.

4.6 Performance Metrics Results

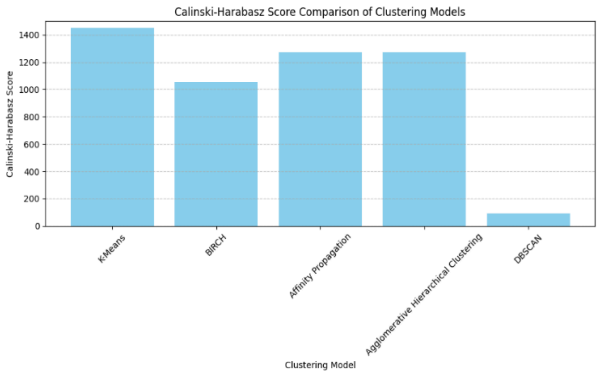
The clustering results from Folium maps suggest that most models effectively distinguished clusters, except for DBSCAN. Models, except DBSCAN, identified similar clusters. The AP model found 9 clusters, possibly complicating interpretation. DBSCAN found only 3 clusters, possibly oversimplifying patterns.



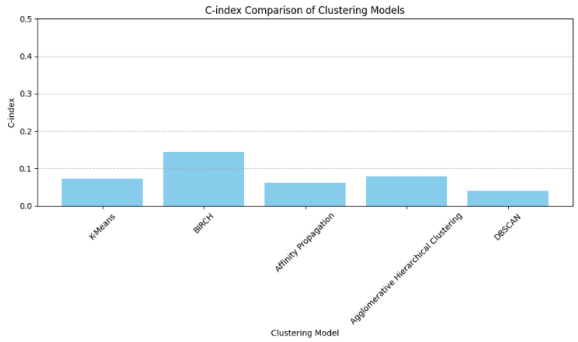
DBSCAN has the highest silhouette score, indicating that the points within clusters are densely packed and well-separated from points in other clusters but its visual representation of clusters on the map does not appear distinct. The other 4 models have roughly the same silhouette score, with Affinity Propagation being the lowest.



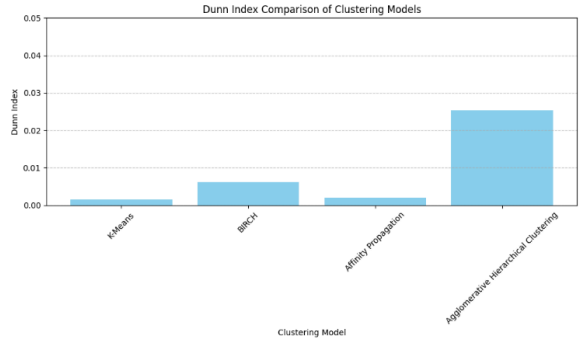
DBSCAN has the highest Davies-Bouldin index, which indicates that clusters are less well-separated or more dispersed. This means DBSCAN might not have performed well in distinguishing between different clusters. BIRCH, Affinity Propagation and AHC have approximately the same Davies-Bouldin index. K Means has the lowest Davies-Bouldin index, indicating that it has better defined clusters.



K Means achieves the highest Calinski-Harabasz score, indicating well-defined clusters with instances closely grouped within each and distant from instances in other clusters. Affinity Propagation and AHC score similarly. Birch outperforms DBSCAN, which fails to effectively separate clusters, leading to significant overlap.



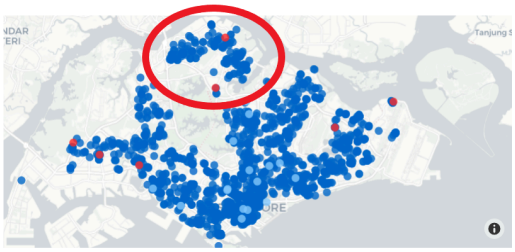
DBSCAN surprisingly has the lowest Hubert & Levin C-index, indicating that it produced the most optimal number of clusters (3), followed by Affinity Propagation, K Means and AHC. Birch has the highest index, indicating that its internal quality index is the worst.



AHC has the highest Dunn Index, indicating that its clusters are the most compact and well-separated from each other,

followed by Birch and Affinity Propagation. K Means, on the other hand, has the lowest Dunn Index, suggesting poorer clustering quality. The Dunn Index isn't applicable to DBSCAN due to its assumption of clear cluster boundaries, which doesn't align with the output of density-based clustering.

4.7 Restaurant Location Recommendation



We excluded DBSCAN's results from our analysis for their lack of visual appeal. After evaluating the remaining 4 clustering outcomes, **we suggest locating the new western restaurant in the upper northern area of Singapore (less competition)**, as depicted in the above figure. Reasons will be discussed thereafter.

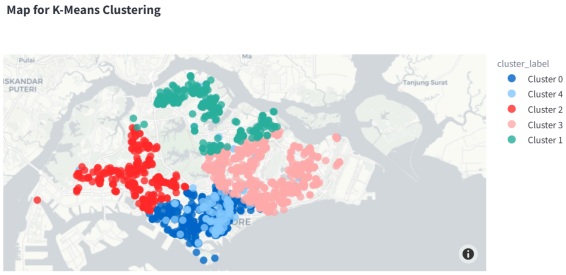


Figure 4.1 K-Means Clustering Results

Referring to Figure 4.1, the green cluster (2nd lowest average rating) is the most ideal location for setting up the restaurant. Although the light blue cluster has the lowest average rating, the restaurant would have to face intense competition from the dark blue cluster, which has the highest average rating.

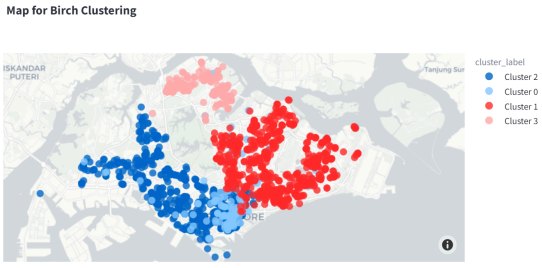


Figure 4.2 Birch Clustering Results

Referring to Figure 4.2, the beige cluster (2nd lowest average rating) is the most ideal location for setting up the restaurant. Although the light blue cluster has the lowest average rating, the restaurant would have to face intense competition from the dark blue cluster, which has the highest average rating.

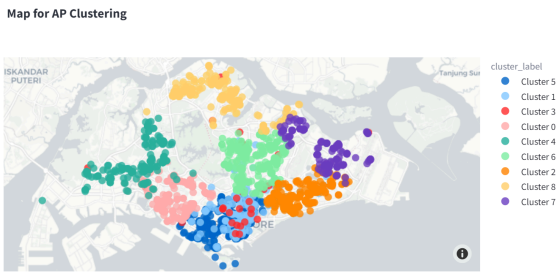


Figure 4.3 Affinity Propagation Clustering Result

Referring to Figure 4.3, the Yellow cluster (4th lowest average rating) is the most ideal location for the new restaurant. Although the average ratings of the light blue and red clusters are very low, setting up there would mean facing stiff competition from other clusters with higher average ratings.

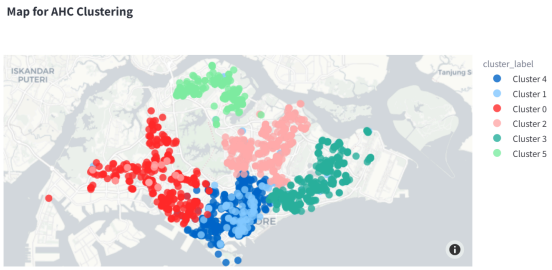


Figure 4.4 AHC Clustering Result

Referring to Figure 4.4, the Light Green cluster (3rd lowest average rating) is the most ideal location for the new restaurant. Although the average rating of the light blue cluster is the lowest, setting up there would mean facing stiff competition from other clusters with higher average ratings.

5) CONCLUSION & TECHNICAL FUTURE RECOMMENDATION

Based on the detailed analysis conducted using an unsupervised machine learning approach with a comprehensive dataset from Grab, key insights have been uncovered that directly influence the decision-making process for setting up a new Western restaurant in Singapore.

Applying multiple clustering algorithms—K Means, BIRCH, Agglomerative Hierarchical Clustering, Affinity Propagation and DBSCAN—has allowed for a thorough data exploration. After meticulous cleaning and preprocessing of the dataset and filtering to focus solely on Western restaurants, these algorithms identified clusters of restaurants across Singapore. The comprehensive analysis included an evaluation of summary statistics and visual representations of these clusters, enabling pinpointing the upper northern area of Singapore as an optimal location for the new establishment. This recommendation is based

on the lower average ratings of existing restaurants in that area and a higher density of Western restaurants, suggesting a substantial customer base and significant potential for outperforming the current offerings.

Despite these promising findings, it is crucial to acknowledge the limitations of this study. The dataset used was slightly outdated (June 2021), so it is possible that some of the restaurants may have already changed locations or closed down, and new restaurants may have entered the market. The focus was exclusively on Western restaurants within Singapore, potentially overlooking broader dining trends and preferences that could impact the success of new restaurants. Additionally, relying on unsupervised machine learning and specific datasets may carry inherent biases or exclude essential variables that could influence the results.

In conclusion, while the findings from this study provide a solid basis to choosing the upper northern area of Singapore as a promising location for a new Western restaurant, the decision should be tempered with consideration of the study's scope and the inherent uncertainties of predictive analytics in new business ventures.

Future Recommendations

To enhance the robustness of future strategic decisions and expand the applicability of this analysis, several recommendations are proposed:

1. **Experimentation with Diverse Models**: We could employ a broader array of machine learning models, like OPTICS and Gaussian Mixture Model, and explore different parameters within each model to uncover more nuanced insights or validate the current findings.
2. **Up-to-date Multi-Data Source Integration**: Utilising up-to-date data from multiple sources beyond GrabFood, such as social media sentiment analysis, competitor performance data, and customer demographics, could enrich the dataset and provide a more comprehensive view of the market landscape.
3. **Inclusion of Supervised Machine Learning**: Integrating supervised learning techniques could provide additional layers of analysis, such as predictive modelling based on historical data, which could refine understanding of customer preferences and success predictors.
4. **Practical Application for Restaurant Owners**: Sharing these findings with restaurant owners could provide valuable insights, helping them make informed decisions about where to locate new establishments based on a combination of customer demand, competition analysis, and market trends.

These steps will reinforce the decision-making process for new restaurant locations and enhance the strategic understanding of the dining industry's competitive dynamics, benefiting stakeholders across the market.

6) REFERENCES

BarkingData. (2022, April 24). *16000+ grab restaurants in Singapore*. Kaggle. https://www.kaggle.com/datasets/polartech/16000-grab-restaurants-in-singapore

Artus, V. (2024, February 29). *Clustering metrics: Evaluate the complex, make it simple*. Medium. https://medium.com/@vladimir-artus/%D1%81lustering-metrics-evaluate-the-complex-make-it-simple-6ae70c0f164b

Benhur, S. (2023, February 24). *Hierarchical clustering: Agglomerative + divisive clustering*. Built In. https://builtin.com/machine-learning/agglomerative-clustering

Chen, S., Liu, X., Ma, J., Zhao, S., & Hou, X. (2019). Parameter selection algorithm of DBSCAN based on k‐means two classification algorithm. *The Journal of Engineering*, *2019*(23), 8676–8679. https://doi.org/10.1049/joe.2018.9082

Chen, Y. Z., & Lai, Y. C. (2018). A schematic illustration of the K-means algorithm for two-dimensional data clustering [Online image]. Physical Review. https://www.researchgate.net/figure/A-schematic-illustration-of-the-K-means-algorithm-for-two-dimensional-data-clustering_fig2_324073652

Dalimunthe, S., & Hanafiah, A. (2021). Implementation of agglomerative hierarchical clustering based on the classification of food ingredients content of nutritional substances. IT Journal Research and Development, 6(1), 60–69. https://doi.org/10.25299/itjrd.2021.6872

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *A Density-Based Algorithm for Discovering Clusters*, 229–230. https://cdn.aaai.org/KDD/1996/KDD96-037.pdf

geeksforgeeks. (2023, November 23). *Affinity propagation*. GeeksforGeeks. https://www.geeksforgeeks.org/affinity-propagation/

Fatakdawala, M. (2023, June 3). *Dunn Index reveals the holy grail of optimal clustering*. Medium. https://medium.com/@mastmustu/dunn-index-reveals-the-holy-

grail-of-optimal-clustering-a48c5bc960e

Huseynov, A. (2021, July 19). *How to open new pizzeria with DBSCAN* [Post]. LinkedIn. https://www.linkedin.com/pulse/how-open-new-pizzeria-dbscan-arif-huseynov

Indarti, N. (2004). Business location and success: The case of internet café business in Indonesia. *Gadjah Mada International Journal of Business*, *6*(2), 171–172. https://doi.org/10.22146/gamaijb.5543

Jeldu, M. D., & Tadele, G. (2018). REVIEW ON DATA COLLECTION, PREPARATION AND CLUSTERING FOR QUALITY PURPOSE. *International Journal of Creative Research Thoughts*, *6*(1), 902–902. https://ijcrt.org/papers/IJCRT1872310.pdf

John. (2020, August 3). *C-index*. PyPI. https://pypi.org/project/c-index/

K.Sasirekha, & P.Baby. (2013). Agglomerative Hierarchical Clustering Algorithm- A Review. International Journal of Scientific and Research Publications, 3(3), 1–3. http://www.ijsrp.org/research-paper-0313.php?rp=P15831

Khater, I. M., Nabi, I. R., & Hamarneh, G. (2020). *An Example Illustrating the Density-Based DBSCAN Clustering Method Applied to SMLM Data* [Online image]. Patterns. https://www.researchgate.net/figure/An-Example-Illustrating-the-Density-Based-DBSCAN-Clustering-Method-Applied-to-SMLM-Data_fig4_342141592

Kumar Shaswat. (2020, January 22). *Restaurant location recommender (using K-means)*. Medium. https://medium.com/@shaswatd673/restaurant-location-recommender-using-k-means-6b3a54f27e64

Lee, W. (2021, March 26). *Clustering culinary neighborhoods in Vancouver, British Columbia*. Medium. https://towardsdatascience.com/clustering-culinary-neighborhoods-in-vancouver-british-columbia-d8c712399874

Lindner, J. (2023, December 19). *Must-know clustering metrics • gitnux*. GITNUX. https://gitnux.org/clustering-metrics/

Liu, X., Li, R., Wang, Y., & Nielsen, P. S. (2022). SEGSys: A mapping system for segmentation analysis in energy. *ARXIV*, 1–13. https://doi.org/10.48550/arXiv.2012.06446

Pan, et al. (2023). *(PDF) DeepFood: Automatic multi-class classification of food ingredients using Deep Learning*. https://www.researchgate.net/publication/319944339_DeepFood_Automatic_Multi-Class_Classification_of_Food_Ingredients_Using_Deep_Learning

Phanich M et al., 2021. (PDF) Food Recommendation System Using Machine Learning for Diabetic Patients.https://cepdnaclk.github.io/e16-4yp-Food-Recommendation-System-Using-Machine-Learning-for-Diabetic-Patients/

R. Refianti, A.B. Mutiara, & S. Gunawan. (2017). TIME COMPLEXITY COMPARISON BETWEEN AFFINITY PROPAGATION ALGORITHMS. *Journal of Theoretical and Applied Information Technology*, *95*(7), 1497–1505. https://www.jatit.org/volumes/Vol95No7/18Vol95No7.pdf

Sauravkaushik8 Kaushik. (2023, May 18). *40 questions & answers on Clustering Techniques for data science professionals (updated 2023)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/02/test-data-scientist-clustering/

Schmidt, T., Schilling, M. W., Behrends, J., Battula, V., Jackson, V., Sekhon, R., & Lawrence, T. (2010). Use of cluster analysis and preference mapping to evaluate consumer acceptability of choice and select bovine M. Longissimus Lumborum steaks cooked to various end-point temperatures. *Meat Science*, *84*(1), 46–53. https://doi.org/10.1016/j.meatsci.2009.08.016

Sudhir Singh, & Nasib Singh Gill. (2013). Analysis And Study Of K-Means Clustering Algorithm. *International Journal of Engineering Research & Technology (IJERT)*, *2*(7), 2546–2547. https://www.ijert.org/research/analysis-and-study-of-k-means-clustering-algorithm-IJERTV2IS70648.pdf

Tan, K. (2024, March 1). *Latest Grabfood Promo Codes Singapore (March 2024)*. SingSaver. https://www.singsaver.com.sg/blog/grab-food-promo-codes

Xie, X., Zhang, X., Fu, J., Jiang, D., Yu, C., & Jin, M. (2018). Location recommendation of digital signage based on Multi-Source Information Fusion. *Sustainability*, *10*(7), 1–17. https://doi.org/10.3390/su10072357

Xu, X. (2021, April 4). *The battle of neighborhoods - find the best location for a new restaurant in Dusseldorf* [Post]. LinkedIn. https://www.linkedin.com/pulse/battle-neighborhoods-find-best-location-new-restaurant-xinya-xu

Zhang, K., & Gu, X. (2014). An affinity propagation clustering algorithm for mixed numeric and categorical datasets. *Mathematical Problems in Engineering*, *2014*, 1–8. https://doi.org/10.1155/2014/486075

Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH Overview [Online image]. Data Mining and Knowledge Discovery. https://doi.org/10.1023/a:1009783824328

Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH: A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery*, *1*(2), 141–182. https://doi.org/10.1023/a:100978382