JOURNAL OF DATA SCIENCE | Vol.2024:35
eISSN:2805-5160

KopiCulture: Unveiling Customer Loyalty in Malaysia's Coffee Market through Clustering Algorithms for Local Cafe Insights

*[1]Tuan Norhafizah Tuan Zakaria, [2]Che Ku Nuraini Che Ku Mohd

[1]College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Negeri Sembilan, Kampus Kuala Pilah, 72000 Kuala Pilah, Negeri Sembilan, Malaysia
[2]Department of Interactive Media, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, 76100, Durian Tunggal, Melaka, Malaysia

***Email:** tuannorhafizah@uitm.edu.my[1*], cknuraini@utem.edu.my[2]

Abstract

In recent years, the coffee market in Malaysia has expanded significantly, propelled by an expanding cafe culture and consumer demand for unique coffee experiences. It is crucial for coffee retailers, such as global chains and local cafes, to comprehend customer loyalty in this dynamic environment. This study aims to identify customer loyalty patterns in the Malaysian coffee market, focusing on the Malaysia Starbucks customer survey dataset. Using clustering algorithms such as KMeans, KMeans with Principal Component Analysis (PCA), single linkage, complete linkage, DBScan, and DBScan in conjunction with PCA, we identify distinct customer segments based on loyalty patterns. Our findings give Starbucks and local coffee shops valuable insights, allowing them to tailor their marketing strategies and improve customer retention efforts. Through this analysis, we contribute to the expanding body of knowledge on customer loyalty in the context of the Malaysian coffee market and offer implications for coffee retailers seeking to thrive in this competitive environment.

Keywords

Customer loyalty, clustering algorithms, coffee market, machine learning, marketing strategies

Introduction

Malaysia's coffee sector has undergone a tremendous expansion and revolution. 164 487 bags of coffee were reportedly drunk worldwide in 2019–2020 (Ramanathan & Nashath 2021). The Hainanese Chinese opened kopitiams during the Colonial era (kopi is the Malay word for coffee, and tiam is the Hokkien/Hakka word for shop), which marked the beginning of Malaysia's coffee culture (Khoo, 2009), as cited in Khazanah Research Institute (2019); The Sun (2017); Ramanathan et al., (2021). Due to the establishment of a thriving cafe culture and growing consumer interest in specialty coffee, the industry has become highly competitive. Coffee stores, international chains, and neighborhood cafes always look for new ways to attract customers in this competitive market.

Submission: 27 July 2024; **Acceptance:** 28 October 2024

In this paper, we aim to explore customer loyalty patterns within the context of the Malaysian coffee market. Specifically, we focus on the Malaysia Starbucks customer survey dataset, which provides valuable insights into customer preferences, behaviors, and satisfaction levels. By leveraging clustering algorithms and incorporating Principal Component Analysis (PCA), we aim to identify distinct customer segments based on their loyalty patterns.

Understanding customer loyalty is crucial for the success of coffee retailers. Loyal customers contribute to recurring revenue and act as brand advocates, influencing others to choose specific coffee establishments. Retailers can develop targeted marketing strategies by uncovering loyalty patterns, identifying customer segments, personalizing customer experiences, and improving customer retention efforts.

We employ various clustering algorithms to achieve our research objective, including KMeans, single linkage, complete linkage, and DBScan. These algorithms allow us to group customers with similar loyalty-related attributes, providing valuable insights into their behaviors and preferences. Additionally, the integration of PCA enhances the clustering process by reducing data dimensionality and improving the visualization of loyalty patterns.

This study holds practical implications for Starbucks and local coffee cafes in Malaysia. By understanding loyalty patterns and identifying distinct customer segments, coffee retailers can tailor their marketing strategies to engage and retain customers effectively. The findings of this research contribute to the existing knowledge on customer loyalty in the Malaysian coffee market and provide valuable insights for industry players seeking to thrive in this competitive environment.

The rest of the paper is organized as follows: Section 2 comprehensively reviews the literature on customer loyalty, clustering algorithms, and their application in the retail and coffee industries. Section 3 describes the methodology, including data collection, preprocessing, and clustering algorithms. Section 4 presents the results and analysis, highlighting the customer segments and loyalty patterns. Section 5 discusses the practical implications and applications of the findings in the context of the Malaysian coffee market. Finally, Section 6 concludes the paper, summarizing the essential findings and suggesting directions for future research.

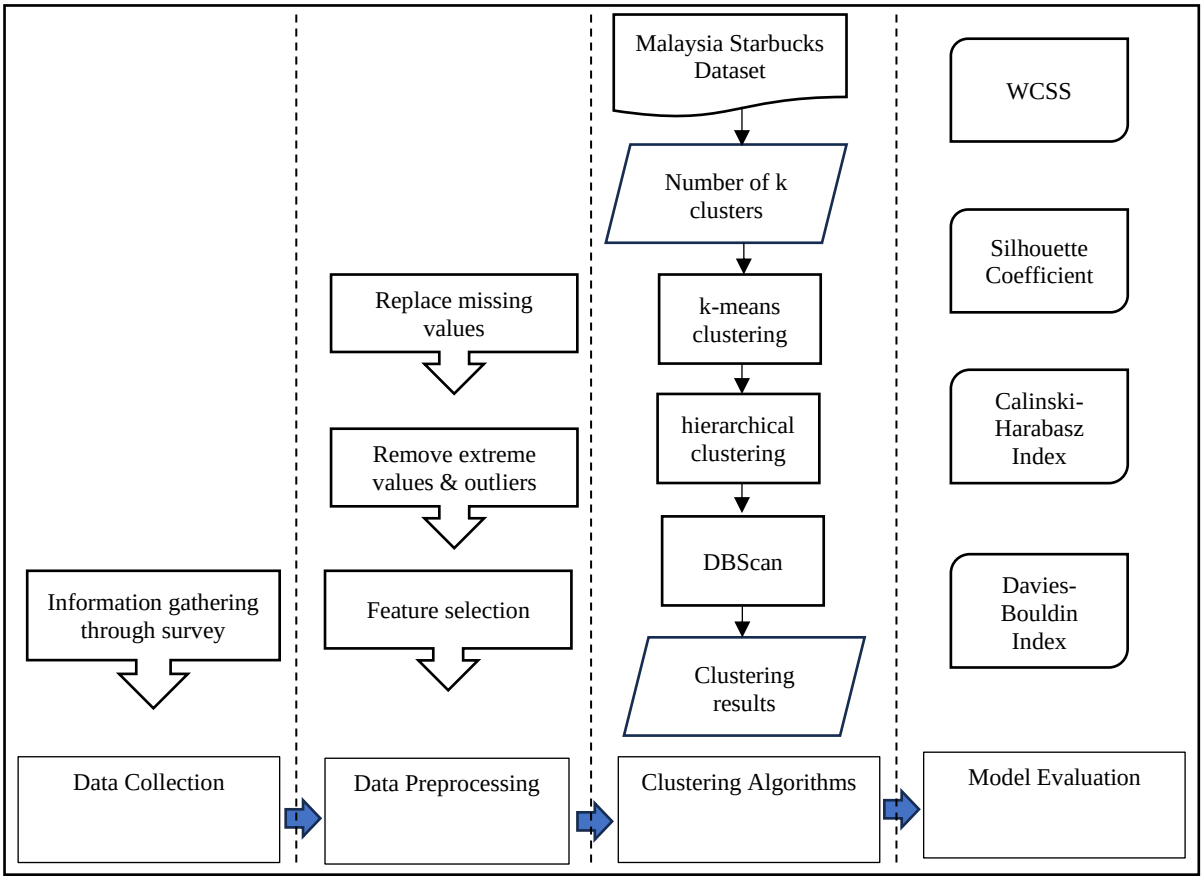
By examining customer loyalty patterns in the Malaysian coffee market, this study aims to contribute to understanding customer behavior and assist coffee retailers in developing effective strategies to enhance customer loyalty and business success.

Methodology

In this research, customer loyalty was thoroughly analyzed using six different clustering techniques: k-means clustering, k-means clustering with PCA, single linkage hierarchical clustering, complete linkage hierarchical clustering, DBScan, and DBScan with PCA. The investigation followed a comprehensive four-step process, which included Data Collection, Data

Preprocessing, Clustering Algorithms, and Model Evaluation. These steps were presented in Figure 2.

Figure 2
Approach Clustering Steps



For the Data Collection phase, information was gathered through a survey conducted among Malaysia Starbucks customers across diverse locations in the country. The dataset comprised 20 variables, such as gender, age, employment status, annual income, and other relevant attributes. Data Preprocessing emerged as a crucial step before the actual clustering analysis. This phase involved several essential preprocessing steps to ensure the data's quality and compatibility with the clustering algorithms. Firstly, Data Cleaning procedures were implemented to handle any missing or erroneous data points, ensuring data integrity.

Next, Feature Selection techniques were applied to identify the most important attributes that influence customer loyalty, thus optimizing the clustering process. Additionally, Data Transformation was carried out to normalize the dataset and minimize the influence of attribute scales on the clustering outcomes.

The Clustering Algorithms phase involved the application of six distinct clustering techniques to segment customers based on their loyalty patterns. These algorithms comprised traditional k-means clustering, k-means clustering enhanced with Principal Component Analysis

(PCA), hierarchical clustering using single linkage, hierarchical clustering using complete linkage, DBScan, and DBScan with PCA.

Finally, Model Evaluation was conducted to assess the performance of each clustering algorithm. Several evaluation metrics, including Within Cluster Sum of Squares (WCSS), Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index, were employed to gauge the effectiveness of the clustering models in capturing meaningful customer loyalty patterns. Through this comprehensive research process, insights into customer loyalty and distinct customer segments were derived, facilitating the development of targeted marketing strategies and enhancing customer experiences for Starbucks in Malaysia.

Results and Discussion

In this study, we employed correlation analysis using a heatmap in Jupyter Notebook to analyze the factors influencing customer loyalty. The heatmap allowed us to visualize the strength and direction of the linear relationships between the various attributes and the target variable, "loyal." By representing the correlation coefficients as color gradients, we gained valuable insights into which features are positively or negatively associated with customer loyalty. The correlation analysis using the heatmap served as an essential step in our data exploration and attribute selection process, providing a visual and quantitative means to identify meaningful relationships between variables. This analysis lays the groundwork for subsequent modeling and analysis, guiding us toward building predictive models and making informed decisions to enhance customer loyalty and satisfaction in our business or research setting. Table 1 shows the correlation values between various attributes toward customer loyalty.

Table 1

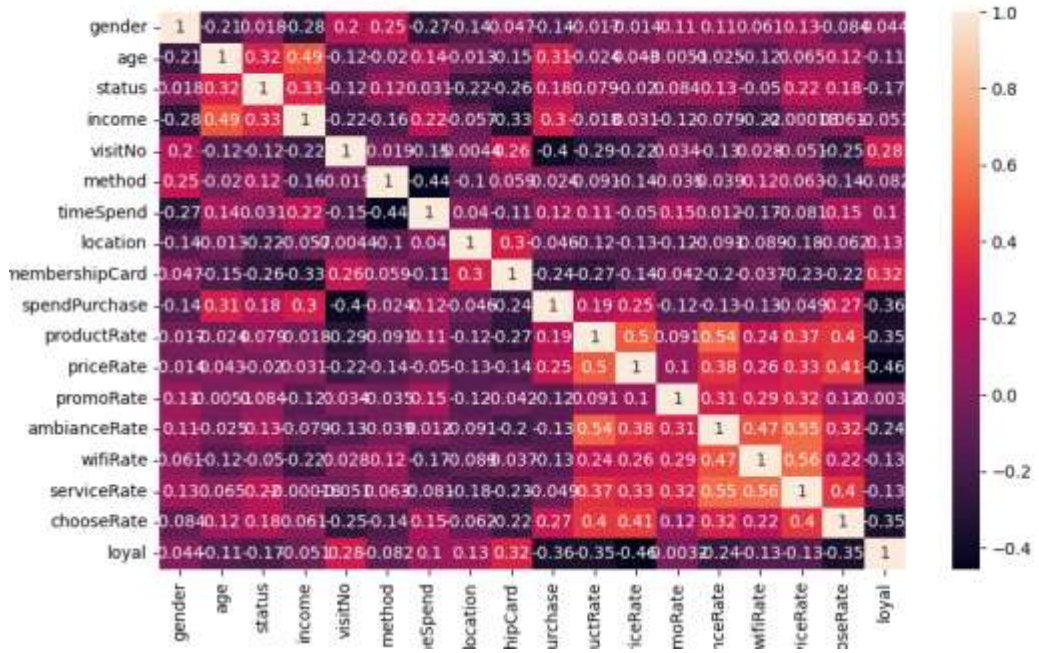
Correlation values between the various attributes and customer loyalty

Attributes	Correlation
priceRate	-0.46
spendPurchase	-0.36
productRate	-0.35
chooseRate	-0.35
membershipCard	0.32
visitNo	0.28
ambianceRate	-0.24
status	-0.17
location	0.13
wifiRate	-0.13
serviceRate	-0.13
age	-0.11
timeSpend	0.10
method	-0.08
income	-0.05

gender	0.04
promoRate	0.00

Table 1 shows the correlation values between the various attributes and customer loyalty. The correlation coefficients, represented as numerical values, indicate the strength and direction of the linear relationships between each attribute and the target variable, "loyal." Positive correlation coefficients suggest a direct relationship, meaning an increase in one attribute is associated with increased customer loyalty. In contrast, negative correlation coefficients indicate an inverse relationship, implying that higher values in one attribute are associated with lower customer loyalty. Next, we represent the correlation values as a heatmap, as in Figure 3.

Figure 3
Heatmap displaying the correlation of attributes with loyalty toward Starbucks.



The heatmap illustrates the relationships between different attributes and their impact on customer loyalty, with warmer colors indicating stronger positive correlations and cooler colors indicating weaker or negative correlations. The heatmap serves as a visual aid for feature selection, highlighting the essential attributes that exhibit significant correlations with loyalty and helping identify key factors influencing customer loyalty in the analyzed data. The correlation values between attributes and customer loyalty provide insights into how each attribute is related to customer loyalty. A positive correlation indicates that as the value of one attribute increases, the value of customer loyalty also tends to increase. On the other hand, a negative correlation suggests that as the value of one attribute increases, customer loyalty tends to decrease. Based on the correlation values, the analysis reveals that membership card (0.32), visit frequency (0.28), and location (0.13) have a significant positive impact on customer loyalty. Conversely, attributes such as price rate (-0.46), spending on purchases (-0.36), product rate (-0.35), and choosing Starbucks for meetings and hangouts (-0.35) are associated with a notable negative impact on customer

loyalty. These findings highlight the crucial factors influencing customer loyalty and provide valuable insights for the coffee shop to focus on strengthening positive aspects and addressing areas of concern to enhance overall customer loyalty.

Next, we created multiple clustering models to group customers based on their loyalty levels. The performance of each clustering algorithm was assessed using various evaluation metrics, including the Within Cluster Sum of Squares (WCSS), Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index. Table 2 displays the evaluation results for each algorithm applied to the Malaysia Starbucks customer survey dataset. These metrics provide valuable insights into the effectiveness and quality of the clustering solutions, aiding in the selection algorithm for customer segmentation and analysis. Table 2 presents the WCSS, Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index values obtained from various clustering models. These metrics indicate the clustering algorithms' performance in segmenting the Malaysia Starbucks customer survey dataset. By analyzing these evaluation measures, we gain valuable insights into the efficacy and appropriateness of each clustering approach for customer segmentation and analysis.

Table 2

The evaluation metrics for each clustering algorithm

Algorithm	WCSS	Silhouette Coefficient	Davies-Bouldin Index	Calinski-Harabasz Index
K-means	186.21	0.41	0.95	73.38
Kmeans + PCA	354.48	0.31	1.16	61.92
Hierarchical (Single)	36.24	0.17	0.79	2.75
Hierarchical (Complete)	61.58	0.37	1.14	53.14
DBScan	Not applicable	0.45	1.07	16.19
DBScan + PCA	522.00	-0.07	5.39	3.18

Table 2 depicts the evaluation metrics for each clustering algorithm using four evaluations: WCSS, Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index. Based on this analysis, the K-Means algorithm achieved a WCSS value of 186.21, indicating a relatively low value for this metric. Lower WCSS values imply compact and well-separated clusters, suggesting that the data points within each cluster are closely related. The Silhouette Coefficient of 0.41 means a moderate level of separation between the clusters. A positive Silhouette Coefficient indicates that the clusters are well-defined and distinct. The Davies-Bouldin Index of 0.95 indicates relatively well-separated clusters, with lower values suggesting better-defined clusters. With a Calinski-Harabasz Index of 73.38, this algorithm exhibits high inter-cluster variance and compactness of the clusters, which is favorable.

The K-Means algorithm with PCA shows different results compared to the regular K-Means. The WCSS increased to 354.48, indicating a slightly less compact clustering solution compared to K-Means alone. The Silhouette Coefficient remains at 0.31, suggesting a similar level of separation between the clusters as the regular K-Means. The Davies-Bouldin Index increased to 1.16, indicating slightly less well-separated clusters. The Calinski-Harabasz Index decreased to 61.92, indicating slightly lower inter-cluster variance and compactness than K-Means alone.

The Hierarchical Clustering using Single Linkage achieved a WCSS value of 36.24, indicating relatively compact and well-separated clusters. Lower WCSS values suggest well-defined clusters. The Silhouette Coefficient of 0.17 suggests a moderate level of separation between the clusters, although it is lower than K-Means. The Davies-Bouldin Index of 0.79 indicates relatively well-separated clusters. The Calinski-Harabasz Index of 2.75 indicates high inter-cluster variance and compactness of the clusters.

The Hierarchical Clustering using Complete Linkage achieved a WCSS value of 61.58, indicating compact and relatively well-separated clusters, although slightly higher than Single Linkage. The Silhouette Coefficient of 0.37 suggests a moderate level of separation between the clusters, which is higher than a Single Linkage. The Davies-Bouldin Index of 1.14 indicates relatively well-separated clusters. The Calinski-Harabasz Index of 53.14 indicates high inter-cluster variance and compactness of the clusters.

DBScan does not have a WCSS value since it does not use the concept of clusters based on centroids. The Silhouette Coefficient of 0.45 indicates a relatively well-defined clustering solution, showing moderate separation between the clusters. The Davies-Bouldin Index of 1.07 suggests relatively well-separated clusters. The Calinski-Harabasz Index 16.19 indicates a relatively good balance between inter-cluster variance and compactness. The DBScan with PCA shows different results compared to DBScan alone. The WCSS value increased significantly to 522.00, indicating a less compact clustering solution. The Silhouette Coefficient decreased to -0.07, indicating that the clustering solution might not be well-defined or meaningful. The Davies-Bouldin Index increased to 5.39, suggesting less well-separated clusters. The Calinski-Harabasz Index decreased to 3.18, indicating lower inter-cluster variance and compactness.

Based on these evaluation metrics, the K-Means algorithm is the most suitable method for identifying customer loyalty patterns in the Malaysian coffee market. Its ability to generate compact, well-separated, and distinct clusters allows for a comprehensive understanding of customer preferences and behaviors, empowering coffee retailers like Starbucks and local cafes to tailor marketing strategies and effectively enhance customer retention. Consequently, the utilization of K-Means in this study contributes significantly to the expanding body of knowledge on customer loyalty in the context of the Malaysian coffee market. We calculate the mean for each attribute from the clustered segment using the K-means algorithm. Table 3 depicts the means for gender, number of visits, time spent, location of nearest Starbucks store, and membership card.

Table 3

Mean Values from Clustered Segment

Attributes	Mean	Detail
gender	0.5	Male
age	1.4	20 - 29 years old
status	1.8	self-employed
income	1.2	25 - 50 k
visitNo	2.5	monthly - rarely
method	1.1	drive-thru
timeSpend	0.7	below 30 - 1 hour
location	1.1	1km - 3km
membershipCard	0.3	yes
itemPurchaseCoffee	1.0	yes
itempurchaseCold	1.0	yes
itemPurchasePastries	1.0	yes
itemPurchaseJuices	1.0	yes
itemPurchaseSandwiches	1.0	yes
itemPurchaseOthers	1.0	yes
spendPurchase	1.6	less than 20 - 40 rm
productRate	3.8	neutral - good
priceRate	3.0	neutral
promoRate	3.9	good
ambianceRate	3.9	good
wifiRate	3.2	neutral
serviceRate	3.9	good
chooseRate	3.7	neutral - good
promoMethodApp	1.0	yes
promoMethodSoc	1.0	yes
promoMethodEmail	1.0	yes
promoMethodDeal	1.0	yes
promoMethodFriend	1.0	yes
promoMethodDisplay	1.0	yes
promoMethodBillboard	1.0	yes
promoMethodOthers	1.0	yes

Discussion

The dataset offers valuable insights into customer preferences and behaviors within a coffee shop setting. The gender distribution appears relatively balanced, with males and females equally represented among the customers. Most customers fall within the age group of 20 to 29 years old, indicating that the coffee shop attracts a younger audience. Interestingly, many customers are self-employed, suggesting that the coffee shop may be popular among freelancers or individuals with flexible work arrangements. Additionally, most customers seem to have a moderate income range, falling between 25,000 to 50,000 RM, which aligns with the coffee shop's affordability and appeal to a diverse customer base.

Customer visits to the coffee shop are primarily monthly to rare, indicating that it may not be the daily go-to spot for everyone but is likely a preferred destination for occasional visits or special occasions. The drive-thru method is preferred by many customers, probably due to its convenience and time-saving benefits. During each visit, customers spend a relatively short amount of time, typically less than 30 minutes to 1 hour, suggesting that the coffee shop may cater to customers seeking quick and efficient service. Most customers reside within a distance of 1 to 3 kilometers from the coffee shop, indicating that it may attract a local and loyal customer base.

While only a small portion of customers possess a membership card, most customers make regular purchases of various items offered by the coffee shop, including coffee, cold beverages, pastries, juices, sandwiches, and other offerings. The spending range of customers falls within the 20 to 40 RM range, implying that the coffee shop offers affordable options for its products.

Furthermore, customer ratings for product quality, pricing, promotions, ambiance, wifi availability, service, and choice are generally positive, with promotions and ambiance receiving higher ratings. Interestingly, most customers utilize various promotional methods, such as apps, social media, emails, deals, word-of-mouth through friends, and advertisements displayed on billboards or within the coffee shop, indicating that it effectively engages with its customer base through multiple channels.

In conclusion, the dataset offers valuable insights that can significantly benefit local coffee shops by understanding their customers' diverse preferences and behaviors. Armed with this information, local coffee shops could optimize their marketing strategies, elevate customer experiences, and build strong customer loyalty.

Implications & Recommendations

The valuable insights obtained from the clustering analysis of Starbucks' customer loyalty patterns can be leveraged by other local coffee shops to enhance their competitiveness and achieve similar success. By understanding and implementing the following implications and recommendations, local coffee shops can effectively improve customer engagement and retention:

i. Marketing approach:

Local coffee shops can adopt a targeted marketing approach similar to Starbucks by tailoring their advertising campaigns and promotions to appeal to specific customer segments. Identifying the dominant demographic, such as females, and designing marketing strategies that resonate with their preferences and interests will enable local coffee shops to build stronger connections with their target audience.

ii. Customer service:

Time optimization is crucial for enhancing the customer experience. Local coffee shops can take inspiration from Starbucks and ensure efficient service, comfortable seating areas, and convenient amenities to cater to customers' time preferences. A well-designed space encouraging customers to stay longer and enjoy their coffee will increase customer satisfaction and loyalty.

iii. Store location:

Strategic store location planning is another crucial aspect to consider. Understanding the proximity preferences of the target customer segment, as Starbucks did with their analysis, will help local coffee shops strategically position their outlets for maximum accessibility and convenience. Opening new stores or optimizing existing ones in areas where the target customers reside will attract more footfall and drive customer satisfaction.

iv. Loyalty program:

Customer retention strategies are paramount for local coffee shops to increase customer loyalty and encourage repeat visits. Implementing loyalty programs and personalized offers, similar to Starbucks, will incentivize customers to choose their establishment more frequently. By making customers feel valued and appreciated, local coffee shops can foster a sense of loyalty and affinity for their brand.

Besides, local coffee shops can enhance their membership or loyalty programs to rival the success of Starbucks. Local coffee shops can strengthen customer engagement and build a loyal customer base by offering exclusive perks, personalized recommendations, and rewards that cater to the preferences and behaviors of their identified customer segment.

In conclusion, the insights gained from Starbucks' customer loyalty analysis serve as a valuable blueprint for other local coffee shops aiming to reach the same level of success. By adopting targeted marketing strategies, implementing effective customer retention initiatives, optimizing the customer experience, strategically planning store locations, and enhancing membership programs, local coffee shops can elevate their brand to rival Starbucks and thrive in the competitive coffee market.

References

International Coffee Organization. (2020). Coffee development report 2020: The value of coffee – sustainability, inclusiveness, and resilience of the coffee global value chain. Retrieved July 10, 2021, from https://www.ico.org/documents/cy2020-21/ed-2336e-coffee-development-report.pdf

Jaafar, N., & Aziz, R. (2018). The Rise of Specialty Coffee in Malaysia: Trends and Sustainability Issues. *Malaysian Journal of Social Sciences and Humanities*, 3(1), 57-65. https://coffexcoffee.com/2018/10/09/the-rise-of-specialty-coffee-in-malaysia/?srsltid=AfmBOoo0DTT14v4FHugaHacISlGZynTB01NBbbBHXD8SUNI0-EqeJWLG

Khazanah Research Institute (2019). Coffee: the world's favourite cuppa' caffeine, Research article by Tumin, S.A., & Shaharudin, A. A. A. Retrieved June 30, 2020, from http://www.krinstitute.org/assets/contentMS/img/template/editor/Views_Coffee_v314012020.pdf

Khoo, G.C. (2009). Kopitiam: discursive cosmopolitan spaces and national identity in Malaysian culture and media. In A. Wise & S. Velayutham (Eds), Everyday Multiculturalism (pp.87-104). London, UK: Palgrave Macmillan Publication. https://doi.org/10.1057/9780230244474_5

Ministry of Agriculture and Food Industries Malaysia. (2021). *The Coffee Industry in Malaysia: An Overview*. Retrieved August 15, 2021, from https://www.mafi.gov.my/documents/coffee-industry-malaysia.pdf

Osman, S., & Ariffin, M. Z. (2020). Exploring Malaysian Coffee Culture: The Emergence of Hipster Cafés. In M. A. Gomez & N. Fernandez (Eds.), *The Cultural Politics of Food and Eating* (pp. 120-134). Kuala Lumpur, Malaysia: University of Malaya Press.

Ramanathan, R., & Ali, N. (2021). Coffee Consumption and the Sustainability of the Coffee Industry in Malaysia. Trends in Undergraduate Research, 4(2), g1-10. https://doi.org/10.33736/tur.3465.2021

The Sun (2017). Time for a break. Pressreader, Retrieved June 29, 2020, from https://www.pressreader.com/malaysia/the-sun-malaysia/20170511/281870118356282.

Teoh, L. M. (2019). The Economic Impact of Coffee Production on Local Farmers in Southeast Asia. *Journal of Agricultural Economics and Development*, 8(3), 132-142.