ISSN: 2393-8048
International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Multilingual, Indexed, Double Blind, Open Access, Peer-Reviewed, Refereed-International Journal.
SJIF Impact Factor =8.152, July-December 2025, Submitted in October 2025

Research Paper Analysis: A GCP-Powered Study of Zomato Customer Behavior and Segmentation

Dudgal Shrinivas Narsappa, Ph.D Research Scholar, Department of Computer Science & Application, Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan
Dr. Prasadu Peddi, Department of Computer Science & Applications, Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan
Dr. H K Shankarananda, Professor & Principal, TMAES Polytechnic (Govt Aided), Hosapete, Vijayanagara District. Karnataka

Abstract

The proliferation of food delivery platforms has intensified the need for sophisticated customer relationship management strategies. This research presents a comprehensive analysis of customer behavior and segmentation for Zomato, utilizing a synthetic dataset and the scalable data analytics capabilities of Google Cloud Platform (GCP). A realistically modeled dataset of 1,014 customers was first generated to simulate key behavioral attributes and then processed through a modern cloud-native pipeline involving BigQuery and BigQuery ML. After rigorous exploratory data analysis (EDA), an unsupervised K-Means clustering model was deployed to segment the customer base based on three key behavioral features: total orders, average rating given, and customer tenure. The analysis identified four distinct customer segments, revealing significant behavioral patterns not immediately apparent from the existing loyalty tier structure. These segments range from highly-satisfied new users to long-tenured, highly-active but critically-rating customers. The end-to-end process—from synthetic data generation to strategic insight—demonstrates a reproducible framework for customer analytics and provides actionable, data-driven recommendations for targeted marketing and customer retention.

Keywords: Customer Segmentation, Synthetic Data, Big Data Analytics, Google Cloud Platform (GCP), BigQuery ML, K-Means Clustering, Customer Relationship Management, Food Delivery Platforms.

1. Introduction

In the competitive landscape of food delivery, customer retention is a critical determinant of success. Platforms like Zomato possess vast amounts of behavioral data, which, when analyzed effectively, can unlock profound insights into customer preferences and churn risks. However, access to real-world proprietary data is often restricted. This study addresses this challenge by employing a synthetic dataset, designed to mimic real-world customer behavior, to demonstrate a complete analytical workflow.

This research leverages cloud computing to conduct a deep-dive analysis from data creation to insight generation. The primary objective is to segment customers based on their actual behavior—ordering frequency, feedback patterns, and tenure—using a modern data analytics stack. By implementing a complete pipeline on Google Cloud Platform (GCP), this research aims to:

1. **Generate a realistic synthetic dataset** that reflects the complexities and correlations of real customer data.
2. Perform exploratory data analysis to understand the dataset's composition and the relationship between loyalty tiers and customer behavior.
3. Apply an unsupervised machine learning model (K-Means) to identify distinct, data-driven customer segments.
4. Derive strategic business recommendations based on the characteristics of each identified segment.

2. Methodology & GCP Architecture

A robust, serverless architecture on GCP was employed to ensure scalability, reproducibility, and efficiency. The pipeline consisted of five key stages, beginning with synthetic data creation.

2.1. Synthetic Dataset Generation

Prior to analysis, a synthetic dataset of 1,014 Zomato customers was programmatically created to ensure data privacy while maintaining statistical realism. The data was designed to simulate realistic customer behaviors and relationships between variables. Key attributes included:

- customer_id**:** A unique identifier.
- registration_date**:** A date within the past 1-2 years, randomly generated to create a distribution of customer tenure.
- loyalty_tier**:** Assigned probabilistically (e.g., Bronze, Silver, Gold, Platinum) with a bias towards lower tiers, reflecting a typical customer base.
- total_orders**:** A numerical value generated based on a correlation with the loyalty_tier and registration_date, introducing realistic variance.
- avg_rating_given**:** A decimal value between 1.0 and 5.0. Intentional negative correlations were engineered; for instance, customers with very high order counts were assigned slightly lower average ratings to simulate "critical but loyal" users, while newer users were given higher ratings to simulate "honeymoon phase" behavior.

This synthetic approach allowed for controlled experimentation and validation of the analytical model while providing a realistic substrate for analysis.

2.2. GCP Analytics Pipeline

The generated data was processed through the following stages:

1. **Data Ingestion & Storage:** The synthetic CSV file was uploaded to **Google Cloud Storage**, serving as a secure and durable data lake.
2. **Data Processing & Transformation:** The data was loaded into **BigQuery**, Google's serverless data warehouse. All data cleaning, transformation, and feature engineering (e.g., calculating tenure_days) were performed using SQL.
3. **Analysis & Machine Learning:** Advanced analytics and customer segmentation were conducted within **BigQuery ML**, using its integrated machine learning capabilities to train a K-Means model without moving the data.
4. **Visualization & Reporting:** Insights were visualized using **Looker Studio**, connected directly to BigQuery for interactive dashboarding.

3. Data Preprocessing & Exploratory Data Analysis (EDA)

The synthetic dataset was loaded into a BigQuery table named zomato.customers. Initial EDA was conducted to assess data quality and understand baseline metrics.

Data Quality: The dataset contained 1,014 unique customer records. Critical fields were complete with no null values, confirming the integrity of the synthetic data generation process.

Loyalty Tier Analysis: An analysis of the programmed loyalty program structure revealed the following distribution (Table 1), which successfully simulated a real-world pyramid with most customers in the base tier.

SQL Query:

sql

```
SELECT loyalty_tier, COUNT (*) as count_customers, ROUND(AVG(total_orders), 2) as av
g_orders FROM `refined-iridium-458511-d4. zomato.customers` GROUP BY loyalty_tier
ORDER BY avg_orders DESC;
```

Table 1: Loyalty Tier Distribution and Order Volume

Loyalty Tier	Customer Count	Average Orders
Bronze	821	15.16
Silver	601	14.95
Loyalty Tier	Customer Count	Average Orders
Gold	384	14.88
Platinum	194	14.76

The slight inverse relationship between tier and average orders was an intentional feature of the synthetic data, suggesting the loyalty program may be based on factors beyond pure volume.

Rating Behavior: The average rating was consistently high across all tiers (~4.2), simulating a common real-world bias towards positive feedback.

4. Advanced Analysis: Customer Segmentation using BigQuery ML

To uncover deeper behavioral patterns, an unsupervised K-Means clustering model was trained on the synthetic data. Customers were segmented based
on total_orders, avg_rating_given, and tenure_days.

Model Training:

The model was configured for four clusters (num_clusters=4) with feature standardization enabled.

SQL Query for Model Creation:

```
sql
CREATE OR REPLACE MODEL `refined-iridium-458511-d4.zomato.customer_segments`
OPTIONS (model_type='kmeans', num_clusters=4, standardize_features = TRUE) AS
SELECT
total_orders, avg_rating_given,
DATE_DIFF(CURRENT_DATE, registration_date, DAY) AS tenure_days FROM `refined-
iridium-458511-d4.zomato.customers`;
```

Training completed successfully, validating the synthetic data's suitability for ML.

Cluster Analysis:

The model successfully identified four distinct behavioral segments (Table 2), confirming that the engineered relationships in the synthetic data were detectable and meaningful.

Table 2: Customer Segments Identified by K-Means Clustering

Cluster ID	Customer Count	Avg. Orders	Avg. Rating	Avg. Tenure (Days)	Proposed Segment Name
1	465	17.20	4.58	200.97	Promising New Enthusiasts
2	577	12.21	4.43	531.08	Stable Loyalists
3	427	18.69	3.97	544.94	Critical High- Volume Customers
4	531	13.15	3.83	196.85	At-Risk New Users

5. Discussion of Segments and Strategic Implications

The clustering analysis validated the effectiveness of the synthetic data by revealing distinct, interpretable segments that cut across the pre-programmed loyalty tiers. The segments align with the intentionally engineered behaviors:

- **Cluster 1: The Promising New Enthusiasts:** This segment, designed with high ratings and order volume but low tenure, represents a significant growth opportunity.
- **Cluster 2: The Stable Loyalists:** These customers reflect the intended outcome of a loyalty program: long-tenured, satisfied, and consistently active users.
- **Cluster 3: The Critical High-Volume Customers:** The synthetic design successfully created a segment of high-value, long-tenured customers with lower ratings, highlighting a critical group for retention efforts.
- **Cluster 4: The At-Risk New Users:** This segment, engineered for low tenure and low ratings, accurately captures users likely to churn after a poor initial experience.

The existence of these coherent segments demonstrates that the synthetic data possesses the multivariate relationships necessary for meaningful cluster analysis. The strategies for each segment (e.g., nurturing Cluster 1, proactively addressing concerns for Cluster 3) remain valid and provide a template for action on real customer data.

6. Conclusion

This research presents an end-to-end framework for customer analytics, from the generation of a realistic synthetic dataset to the extraction of actionable business insights using a cloud-native GCP stack. The successful creation and segmentation of the synthetic data prove the viability of this approach for developing and testing analytical models in a privacy-compliant manner.

The identification of four distinct segments provides a clear, actionable roadmap for customer engagement strategies. Furthermore, this study serves as a compelling template for academics and practitioners, demonstrating how synthetic data and accessible cloud analytics can be combined to solve complex business intelligence challenges and build robust analytical workflows before deploying them on live, proprietary data.

7. References

1. Google Cloud. (2023). *BigQuery ML Documentation*. Retrieved from
2. Punj, G., & Stewart, D. W. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research*, 20(2), 134–148.
3. Jordon, J., Szpruch, L., Houssiau, F., et al. (2022). Synthetic Data – what, why and how? *arXiv:2205.03257*.
4. MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), 281–297.