"Enhancing Customer Segmentation in Online Retail: A Comparative Analysis of Equal Width Binning, Equal Frequency Binning, and Anova Tests"

Nayana R[1], Navya T C[2]

[1,2]Department of Artificial Intelligence, Reva University, Bengaluru, India

ABSTRACT

This study employs RFM (Recency, Frequency, Monetary) analysis to segment customers based on transactional behavior in an online retail setting. The methodology encompasses data collection, preprocessing to ensure data integrity, and calculation of RFM metrics. Customer segmentation is achieved through both equal width and equal frequency binning methods, facilitating targeted marketing strategies. Statistical analyses including chi-squared and ANOVA tests evaluate relationships and disparities among RFM bins, offering insights into customer behavior and informing strategic decision-making in marketing and customer relationship management.

Keywords: RFM analysis, customer segmentation, online retail, equal width binning, equal frequency binning, chi-squared analysis, ANOVA analysis, customer behavior, marketing strategy, customer relationship management

INTRODUCTION

In the realm of online retail, understanding customer behavior through RFM analysis is pivotal for effective marketing and customer relationship strategies. This study delves into the methodology of RFM analysis, starting from data collection and preprocessing to the calculation of Recency, Frequency, and Monetary metrics. By segmenting customers using binning techniques such as equal width and equal frequency, the study aims to uncover distinct customer groups based on their purchasing patterns. Statistical analyses further explore associations and variances within these segments, offering actionable insights for enhancing customer engagement and optimizing business outcomes.

METHODOLOGY

The methodology adopted in this study involves several key steps: data collection, preprocessing, RFM metrics calculation, and customer segmentation through equal frequency binning.

Data Collection and Preprocessing

The dataset comprises detailed transactional data from an online retail store. Essential preprocessing steps include handling missing values, removing duplicates, and ensuring data consistency. These steps are crucial for accurate analysis and reliable results.

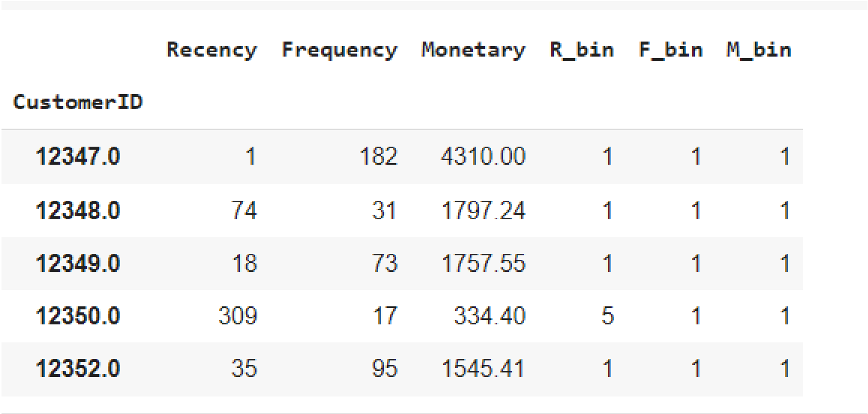
The initial dataset contained several inconsistencies, including missing values and duplicates. These were addressed through a systematic approach to ensure data integrity. Missing Customer IDs were the primary issue, resolved by dropping the affected rows. Missing descriptions were filled with a placeholder to maintain completeness. Duplicates were identified and removed to prevent skewed results. Additionally, outliers such as negative quantities and prices were corrected or removed to ensure a clean dataset for analysis.

RFM Metrics Calculation

RFM analysis, a cornerstone of customer segmentation in marketing and CRM, utilizes three key metrics—Recency, Frequency, and Monetary value—to profile customer behavior based on transactional data. By calculating how recently customers made purchases, their buying frequency, and total spending, businesses can categorize customers into distinct segments. These segments allow for tailored marketing strategies: engaging recently active customers to prevent churn, incentivizing frequent buyers to enhance loyalty, and strategically approaching high-spending customers for personalized service. Implemented through Python and Pandas, this study applied RFM analysis to an online retail dataset, computing metrics like recency in days, weeks, and months to reveal nuanced insights into customer engagement and purchasing patterns. Such insights not only optimize customer retention and acquisition strategies but also pave the way for predictive modeling and machine learning applications in retail analytics, ensuring businesses remain competitive by aligning marketing efforts precisely with customer preferences and behaviors. RFM analysis thus emerges as a pivotal tool in driving sustainable growth and fostering stronger customer relationships in the dynamic landscape of online retail.

Equal Width Binning

This study understands the importance of customer segmentation and employs a binning process on three variables: Recency (R), Frequency (F), and Monetary (M), which are key metrics in customer segmentation analysis. The binning process involves dividing the continuous values of these metrics into discrete intervals or bins. The method used here is equal-width binning, meaning that the entire range of values for each metric is divided into a specified number of bins that each cover an equal portion of the range. In this case, the number of bins is set to five, meaning that the values for Recency, Frequency, and Monetary are each divided into five equal-width bins. Each value of Recency, Frequency, and Monetary is assigned a bin number from one to five, where one indicates the lowest values and five the highest. To store these bin numbers, three new columns are created, named R_bin, F_bin, and M_bin respectively. Finally, the result of this binning process is displayed, typically showing the first few rows of the modified data to illustrate the bin assignments. This categorization of continuous RFM values into discrete bins facilitates easier segmentation and analysis of customer behavior based on their recency, frequency, and monetary value metrics.



	Recency	Frequency	Monetary	R_bin	F_bin	M_bin
CustomerID						
12347.0	1	182	4310.00	1	1	1
12348.0	74	31	1797.24	1	1	1
12349.0	18	73	1757.55	1	1	1
12350.0	309	17	334.40	5	1	1
12352.0	35	95	1545.41	1	1	1

Fig 1 Equal Width Binning

It sets up a figure with three subplots side by side. For each metric (Recency, Frequency, and Monetary), it plots a histogram showing how many data points fall into each bin.

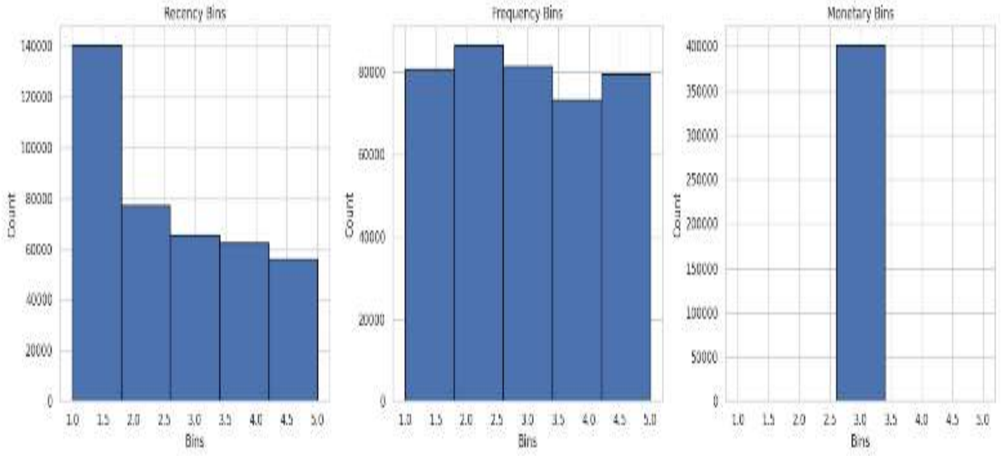
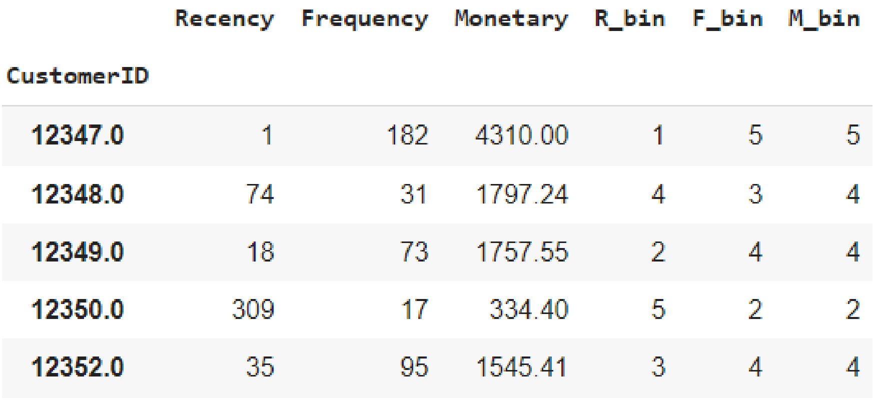


Fig: 2 Plot of Equal Width Binning

Equal Frequency Binning

This study understands the importance of effectively segmenting customers and applies equal frequency binning to the RFM metrics—Recency (R), Frequency (F), and Monetary (M). This method divides these metrics into quintiles, resulting in five equal-sized groups for each metric, facilitating the comparison of customers within each segment. In our analysis of customer transaction data from an online retail dataset, we employed quantile binning to derive meaningful segments based on RFM metrics. This statistical method divides the dataset into equal-sized bins based on the distribution of the data, ensuring each bin contains approximately the same number of observations. By segmenting customers into quintiles using quantile binning, we enhance the comparison of customers within each segment. Specifically, the code utilized in our RFM analysis divided Recency, Frequency, and Monetary values into five equally sized groups, categorizing 'Recency' values from 'Very Recent' to 'Not Recent,' 'Frequency' values from 'Very Frequent' to 'Infrequent,' and 'Monetary' values from 'Very High' to 'Very Low.' This stratification enables a nuanced understanding of customer purchasing behavior and supports the development of targeted marketing strategies. Additionally, the initial rows of the resulting DataFrame were displayed to illustrate the segmented data.



	Recency	Frequency	Monetary	R_bin	F_bin	M_bin
CustomerID						
12347.0	1	182	4310.00	1	5	5
12348.0	74	31	1797.24	4	3	4
12349.0	18	73	1757.55	2	4	4
12350.0	309	17	334.40	5	2	2
12352.0	35	95	1545.41	3	4	4

Fig 3 Equal Frequency Binning

Fig visualize the distribution of customers across different RFM bins. It creates a figure with three subplots, each representing the distribution of customers based on Recency, Frequency, and Monetary values. For each subplot, a count plot is generated, showing the count of customers in each bin. This visualization allows for a clear understanding of customer distribution and helps identify patterns in purchasing behavior. Adjustments are made to the layout for better spacing between subplots, ensuring clarity in presentation.

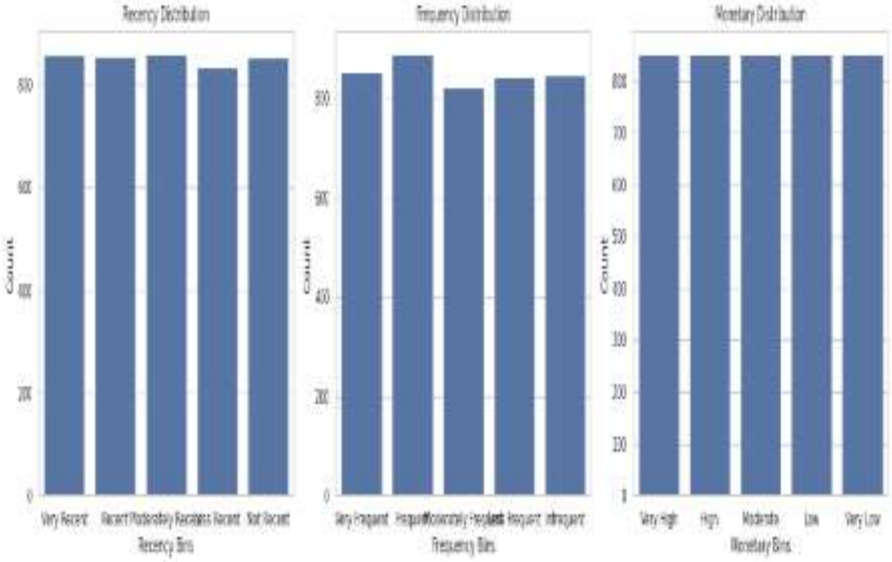


Fig: 4 Plot of Equal frequency Binning

The Fig: analysis involves segmenting customers based on Recency, Frequency, and Monetary value (RFM analysis). The average monetary value for each Recency and Frequency bin is calculated to understand how monetary value varies with these segments. Similarly, the average recency for each Monetary bin is determined.

The visualization consists of three bar graphs: one for Recency bins, one for Frequency bins, and one for Monetary bins. Each bin's bar color represents the average value of the corresponding metric, highlighting high-value customers. This approach identifies key customer segments, aiding in targeted marketing strategies.

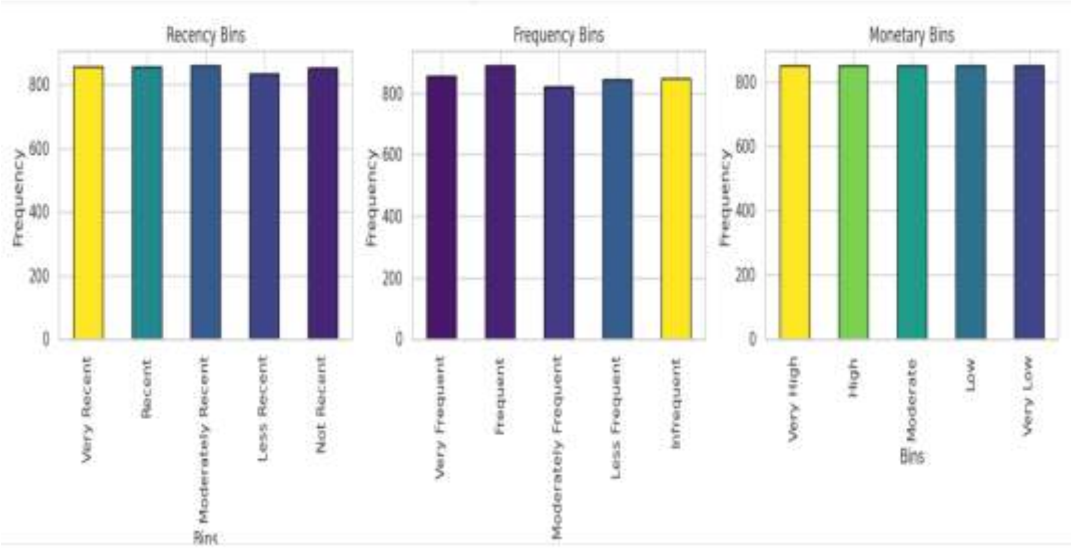


Fig: 5

A Chi-Squared Analysis

In this study, a chi-squared test was conducted to investigate the association between Recency bins and both Monetary and Frequency bins. Initially, Recency, Frequency, and Monetary values were grouped into bins using quantiles. Subsequently, Recency bins were categorized as 'Low' or 'High' based on their recency. Contingency tables were then generated to display the distribution of Monetary and Frequency bins across the mapped Recency bins. Chi-squared tests were employed to evaluate whether a significant association existed between Recency and Monetary, as well as Recency and Frequency. The significance level (alpha) was set at 0.05, and p-values were compared against this threshold. If the p-value was less than alpha, the null hypothesis (no association) was rejected, indicating a significant disparity between the bins. Conversely, if the p-value was greater than or equal to alpha, the null hypothesis could not be rejected, suggesting no substantial difference.

An Anova Analysis

The conducted analysis involves ANOVA (Analysis of Variance) tests to assess disparities in means across distinct Recency, Frequency, and Monetary bin levels. Initially, Recency, Frequency, and Monetary values undergo binning through quantile categorization. Subsequently, individual ANOVA tests are executed for each variable to ascertain the presence of statistically noteworthy mean discrepancies among the bins. For each ANOVA test, both the F-value and p-value are computed. A p-value below the standard significance level of 0.05 signifies substantial differences between at least two groups, while a p-value equal to or exceeding 0.05 suggests an absence of significant differences. These ANOVA test outcomes furnish insights into the potential influence of Recency, Frequency, or Monetary bins on customer behavior.

CONCLUSION

This study demonstrates the application of RFM analysis in understanding customer behavior within an online retail context. By utilizing transactional data and applying rigorous preprocessing techniques, including handling missing values and removing duplicates, the study ensured data integrity for accurate RFM metric calculation. Equal width and equal frequency binning methods effectively segmented customers based on Recency, Frequency, and Monetary metrics, revealing distinct customer groups with varying purchasing behaviors. Statistical analyses such as chi-squared and ANOVA tests provided further insights into relationships and disparities among RFM segments, aiding in the formulation of targeted marketing strategies. The findings underscore the importance of RFM analysis in optimizing customer engagement and retention strategies, thereby enhancing business performance in competitive online markets.

REFERENCES

[1]. Pacific Business Review International. (2021). Customer Segmentation Using RFM Analysis: Realizing Through Python Implementation. Volume 13 Issue 11. Retrieved from Pacific Business Review.