Unsupervised Machine Learning (DS5230): PROJECT FINAL REPORT

Market Basket Analysis using RFM Modelling for Enhanced Decision Making

Market Basket Analysis decodes customer purchasing patterns through transactional data, revealing item associations. When integrated with RFM analysis, businesses gain a holistic understanding to alter strategies and enhance customer satisfaction. Our project merges Market Basket Analysis and RFM (Recency, Frequency, Monetary Value) Analysis to decode customer purchasing patterns. This combined approach uncovers correlations in product purchases and customer behavior, empowering us to tailor strategies for enhanced customer satisfaction. By leveraging these insights, we aim to optimize product placement, marketing strategies, and promotions to meet diverse customer preferences, fostering stronger relationships and loyalty

OBJECTIVES

Market basket Analysis

Utilize FP-growth algorithm

Discern Association rules among products using Apriori algorithm

RFM Analysis

Compute metrics for customer behavior.

Segment customers based on RFM values.

Understand customer engagement levels.

Sales Forecasting

Leverage ARIMA model.

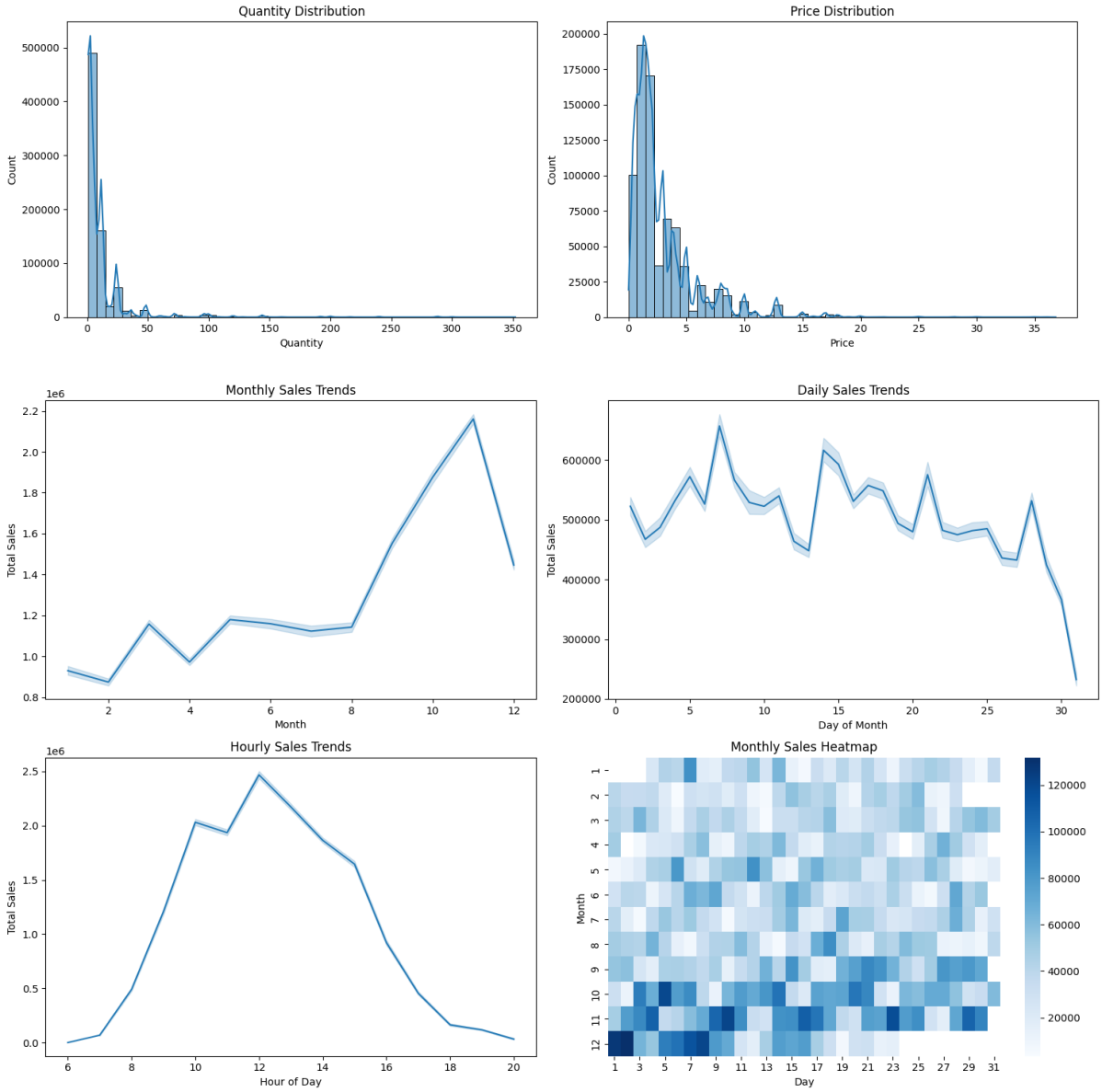
Analyze patterns in 'Invoice Date' and 'Quantity' columns.

Forecast future sales based on identified trends.

The comprehensive report explores three pivotal analytical methodologies—Market Basket Analysis (MBA), RFM Analysis, and Sales Forecasting utilizing the ARIMA model. Each method serves a distinct purpose in extracting insights crucial for strategic decision-making.

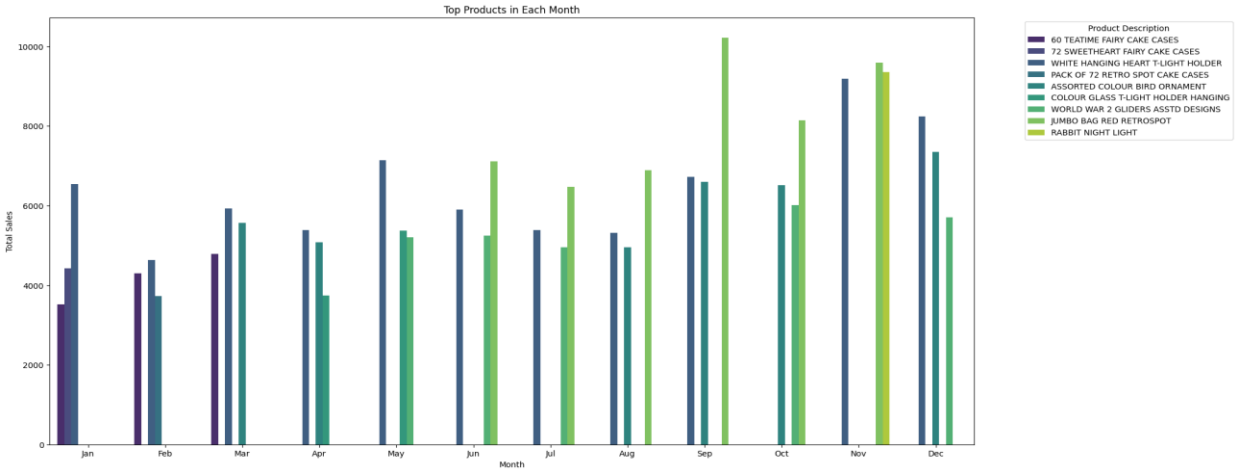
- Market Basket Analysis involves understanding customer purchasing behavior by discerning associations among products. The utilization of the FP-growth algorithm streamlines association rule mining, while insights derived from the Apriori algorithm highlight significant product associations and transaction patterns.
- RFM Analysis delves into customer behavior through Recency, Frequency, and Monetary metrics. By computing RFM metrics and segmenting customers based on these values, the analysis reveals nuanced customer engagement levels. Identification of customer segments, such as High-Value Customers and Churn Risk, provides actionable insights.
- The Sales Forecasting component, employing the ARIMA model, focuses on predicting future sales by identifying trends in 'Invoice Date' and 'Quantity' columns. Unveiling patterns within the temporal aspects of sales data allows for accurate predictions based on these insights.

EDA – Visualization

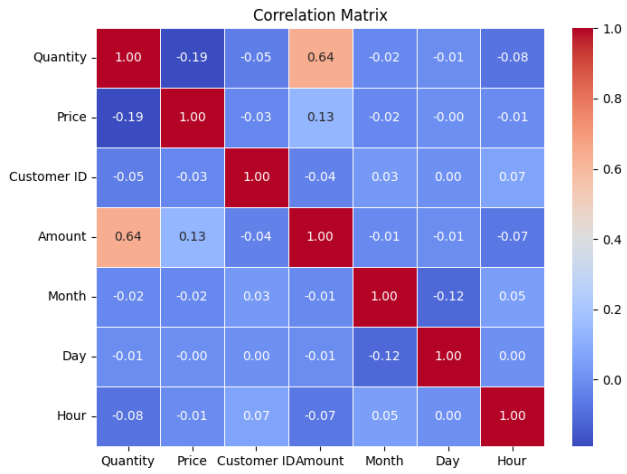


The visual analysis indicates an inverse relationship between price and quantity purchased. As the price increases, there is a noticeable decrease in the quantity of goods bought. This insight underscores the sensitivity of customer purchasing behavior to price fluctuations, emphasizing the need for strategic pricing considerations. Examining the time series data reveals intriguing patterns in sales. Over the last five months, there has been a noticeable uptick in goods sold, particularly in the latter half of the year. Daily sales trends remain relatively consistent, with a slight decline observed towards the end of each month. On an hourly basis, there is a distinct spike in sales during the midday hours, while sales decrease in the very early and late hours. The heatmap representation further supports these observations, highlighting increased sales in the last three months compared to the earlier months of the year. These insights provide valuable cues for optimizing inventory management and tailoring marketing strategies based on temporal sales patterns.

EDA – CONTINUED



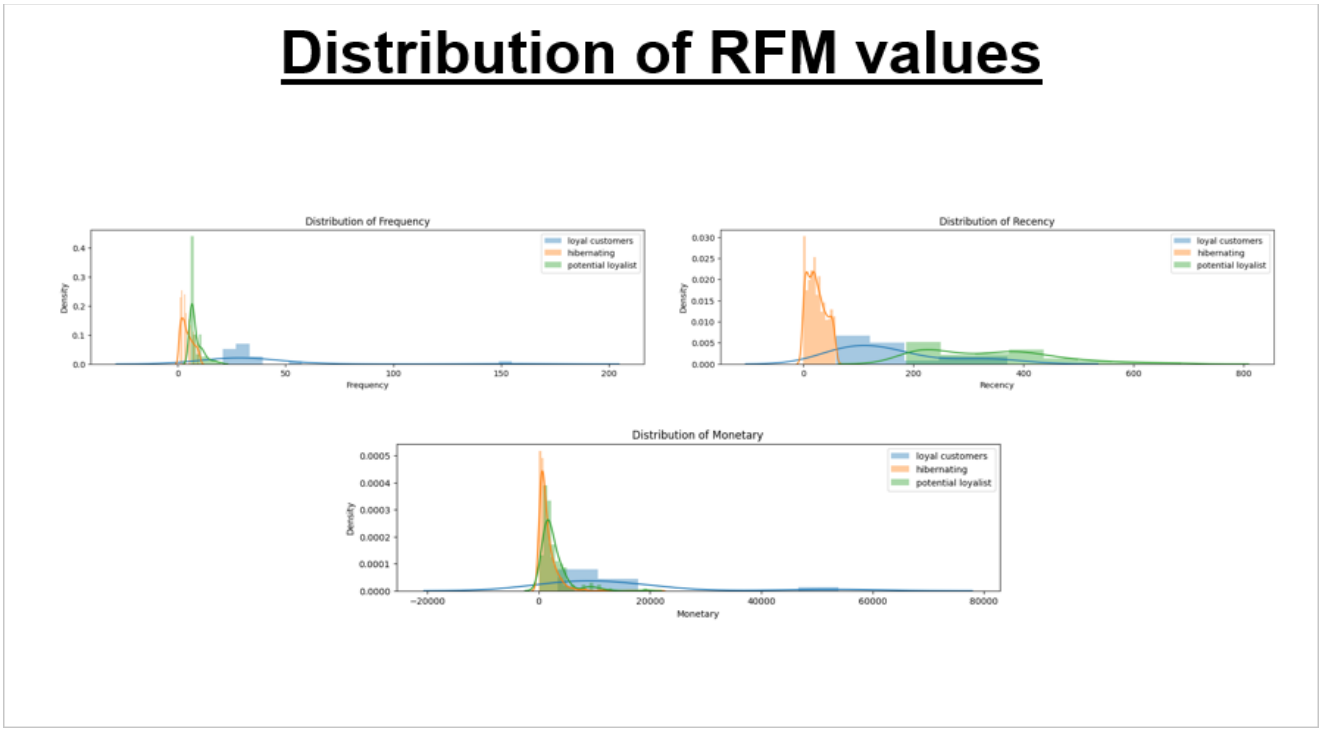
The analysis of specific product sales patterns reveals distinct insights that contribute to strategic decision-making. For "60 TEATIME FAIRY CAKE CASES," the observed surge in sales during the first three months of the year points to a seasonal trend, potentially associated with spring. This suggests a correlation between the product and seasonal preferences or events during this period. On the other hand, "WHITE HANGING HEART T-LIGHT HOLDER" emerges as a consistent bestseller, maintaining substantial sales volumes throughout the year. The steady demand underscores its popularity among customers, establishing it as a perennial favorite. "ASSORTED COLOUR BIRD ORNAMENT" exhibits a preference for the second half of the year, indicating a seasonal or thematic resonance during this period. Understanding such seasonal patterns can inform inventory management and targeted marketing strategies. Lastly, the sales of "JUMBO BAG RED RETROSPOT" showcase seasonal peaks from June to November, reaching it's zenith in September. Leveraging this seasonality insight can optimize targeted marketing efforts and inventory planning during these peak months. These nuanced insights provide valuable guidance for inventory stocking, marketing campaigns, and the implementation of effective seasonal promotions.



The correlation analysis conducted on the DataFrame using df.corr() indicates that, in general, there isn't a substantial correlation among the majority of variables. However, a notable exception is observed in the relationship between Quantity and Amount. This correlation is expected, given that the Amount variable is derived from the multiplication of Quantity and Price. The strong correlation between Quantity and Amount is explained by the fact that Amount is calculated by multiplying Quantity and Price. This relationship is anticipated and confirms the coherence in the data. The absence of significant correlations among other variables suggests a limited interdependence. Variables appear to operate relatively independently, emphasizing the diverse nature of the dataset.

MBA USING RFM-ANALYSIS:

Using our dataset, we perform a RFM (Recency-Frequency-Monetary) analysis to group customers based on their common characteristics. These customer segments are beneficial in marketing campaigns, in identifying potentially profitable customers, and in developing customer loyalty. A company might segment customers according to a wide range of factors, including demographics (age, gender, location, etc.), behavior (previous orders, responses to messaging), psychographics (values, interests, lifestyles) etc. We also use Apriori and FPGrowth to get the frequent itemsets and then go on to find association rules using lift or confidence.



The resulting distribution of Recency, Frequency and Monetary provide perspective, insights into the behaviors of the customer.

The RFM analysis has yielded three distinct customer segments based on their Recency (R), Frequency (F), and Monetary (M) behaviors:

- Loyal customers:
    - Recency (R): Customers in this cluster demonstrate consistent and frequent engagement over time, indicating loyalty to the brand or product.
    - Frequency (F) and Monetary (M): These customers exhibit a pattern of regular transactions and contribute significantly to the overall monetary value. They are the core, reliable customer base.
- Hibernating:
    - Recency (R): The hibernating segment displays a decline in recent interactions, suggesting a decrease in engagement over time.
    - Frequency (F) and Monetary (M): While they may have been active in the past, the current frequency and monetary contributions have diminished, indicating a potential decline in interest.

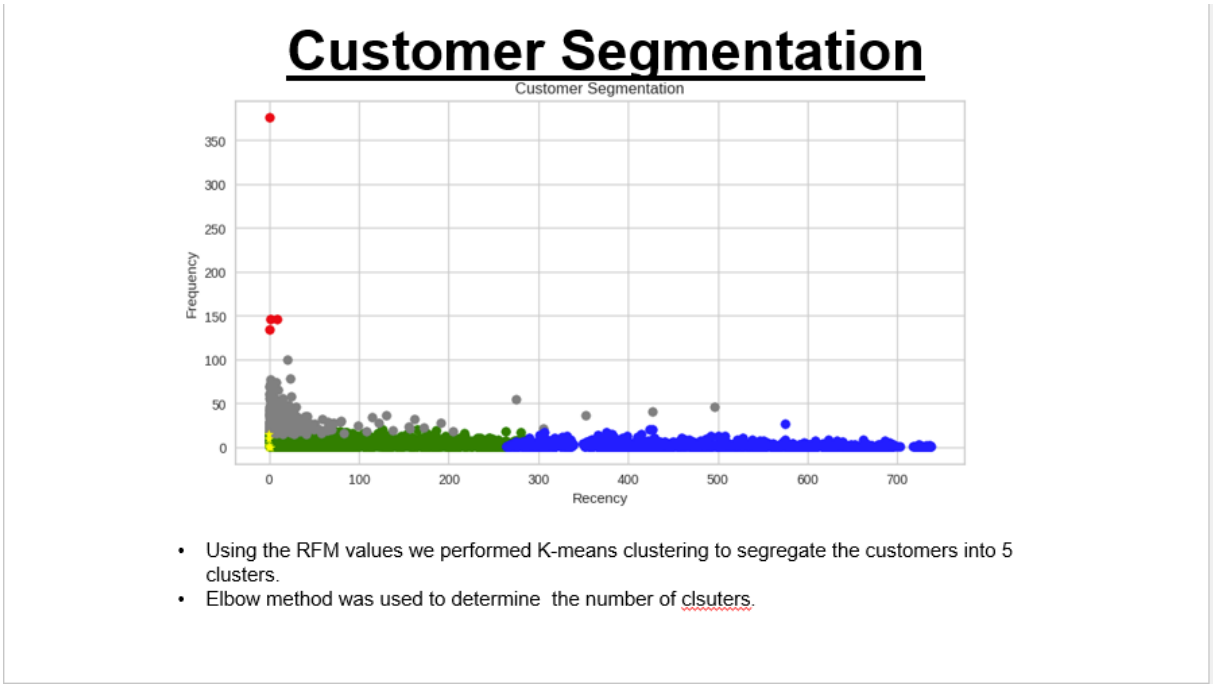
- Potential Loyalists
    - Recency (R): Like loyal customers, potential loyalists engage frequently in the recent past, showing promise for potential loyalty.
    - Frequency (F) and Monetary (M): Their current contribution is comparable to loyalists, but there's a decline as recency extends beyond a certain point. This group represents an opportunity for strategic interventions to convert potential into sustained loyalty.

K-MEANS CLUSTERING:

In our quest to enhance customer segmentation, we employed K-means clustering on RFM values, a powerful technique that partitions customers into distinct groups based on their Recency (R), Frequency (F), and Monetary (M) behaviors. The challenge, however, lies in determining the optimal number of clusters, K, to ensure meaningful segmentation.

K-means clustering is a widely used algorithm for partitioning data into K clusters, where each data point belongs to the cluster with the nearest mean. By applying this technique to our RFM values, we aim to uncover nuanced patterns within our customer base and tailor strategies to distinct segments.

The Elbow Method is a pragmatic approach for determining the optimal number of clusters in K-means clustering. It involves plotting the sum of squared distances between data points and their assigned cluster centers for different values of K. The "elbow" in the plot represents the point at which increasing the number of clusters ceases to significantly reduce the sum of squared distances, indicating the optimal K.



By leveraging the Elbow Method to pinpoint the optimal K value, we have refined our customer segmentation strategy and segregated them into 5 segments. This approach ensures that our

marketing efforts are precisely tailored to the diverse behaviors exhibited by our customers, setting the stage for enhanced engagement and satisfaction across the spectrum.

TIME SERIES FORECASTING

The time series forecasting analysis involved the utilization of the ARIMA (Autoregressive Integrated Moving Average) model to predict future sales based on historical data. The dataset, containing columns such as 'Invoice Date' and Amount,' was employed to uncover patterns and trends in the temporal aspects of sales. The initial step involved checking the stationarity of the time series data using the Augmented Dickey-Fuller (ADF) test. The Augmented Dickey-Fuller (ADF) test was employed to assess the stationarity of the time series data. The test yields a test statistic (ADF Statistic), a p-value, and critical values at different confidence levels. Here are the key insights from the ADF test results:

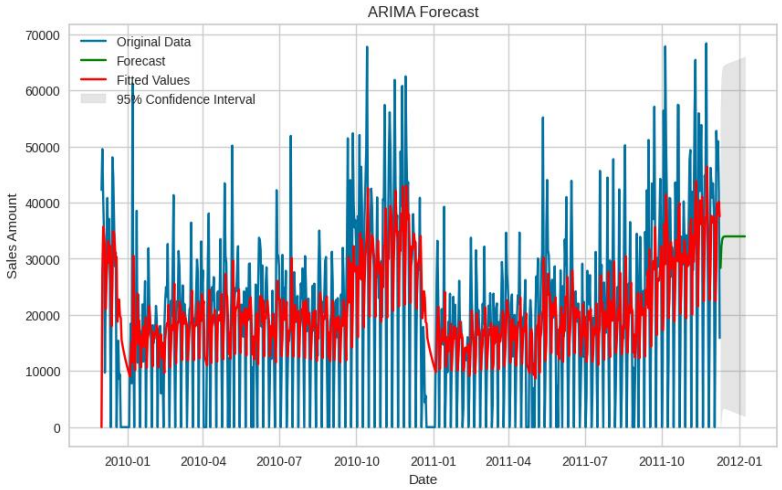
First Set of ADF Test Results: ADF Statistic: -2.621680358099256, p-value: 0.08859865237653525, Critical Values: {'1%': -3.439490435810785, '5%': -2.8655738086413374, '10%': -2.568918067209286}

The initial ADF test indicated a p-value of approximately 0.09, suggesting that the time series data may not be entirely stationary. However, the p-value is relatively close to the significance threshold of 0.05, indicating the need for further investigation.

Second Set of ADF Test Results: ADF Statistic: -9.811115628741115, p-value: 5.653726242069977e-17, Critical Values: {'1%': -3.439490435810785, '5%': -2.8655738086413374, '10%': -2.568918067209286}

These results suggest that the differenced series is now stationary. The test helps determine if difference is required to make the series stationary. The series was found to be non-stationary; therefore differencing was applied to make it stationary. This step is crucial for accurate modeling using ARIMA. The ARIMA model was constructed with the following parameters: p (AR parameter): Determined based on the autocorrelation function (ACF) plot, d (Integration parameter): Determined by the order of differencing required to achieve stationarity, q (MA parameter): Identified from the partial autocorrelation function (PACF) plot.

The ARIMA model was then trained on a subset of the time series data to learn the underlying patterns and relationships. The trained ARIMA model was then used to make future predictions for the time series, providing insights into potential future sales trends. This comprehensive approach to time series forecasting ensures that the model captures the inherent patterns and seasonality in the data, enabling accurate predictions for strategic decision-making. Adjustments to the ARIMA parameters can be made based on the specific characteristics of the dataset.



CONCLUSION:

This project, spanning from customer behavior analysis to market basket insights and time series forecasting, provided a thorough exploration of retail data. Leveraging RFM analysis, market basket algorithms, and the ARIMA model, we uncovered valuable patterns, identified product associations, and forecasted future sales trends. Visualizations and statistical analyses deepened our understanding of customer preferences and purchasing dynamics. The derived insights offer actionable strategies for inventory management and targeted marketing. This project not only enhances our grasp of retail analytics but also equips us with a versatile toolkit for informed decision-making in the dynamic retail landscape.