**Advanced Machine Learning Model for Customer Segmentation in Retail**

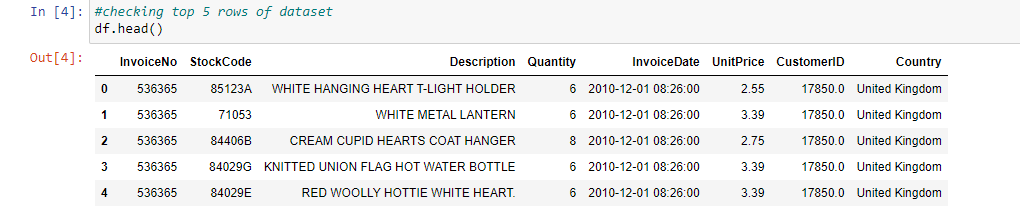
**Premchand M**

**Kasthur Reddy Paddarthi  
Saint Peter’s University  
Ds630  
Instructor  
February 14, 2024**

**Data Preprocessing and Exploratory Data Analysis**

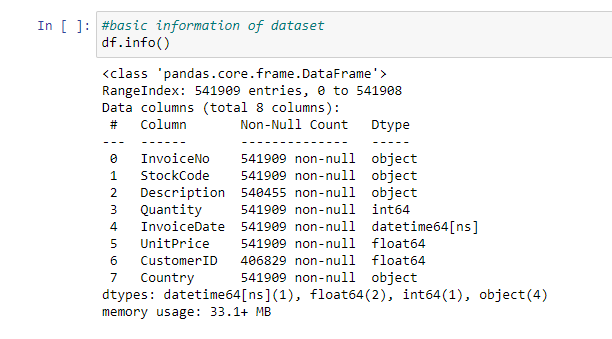
**Data Preview Report**

In my initial examination of the dataset, I observed columns for InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. This dataset is a detailed record of retail transactions, with each entry detailing an individual purchase, its description, and the customer's information. The preliminary review of this data is critical as it allows me to understand the structure and content I’m working with, setting the stage for the in-depth analysis.



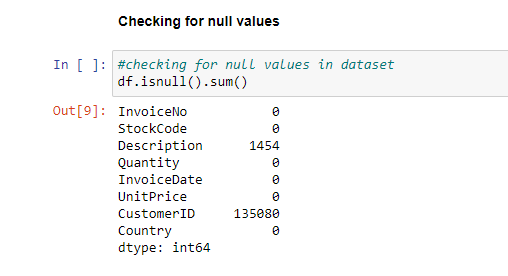
**Dataset Information Report**

The dataset consists of 541,909 entries, showing a high level of completeness, with most fields being non-null. However, I noted some gaps in the 'CustomerID' column, which is essential for the customer segmentation efforts. The data types are diverse, encompassing objects, integers, floats, and datetime formats. This variety presents us with rich opportunities for various analyses, from temporal trends to customer purchasing patterns.



**Data Cleaning Report**

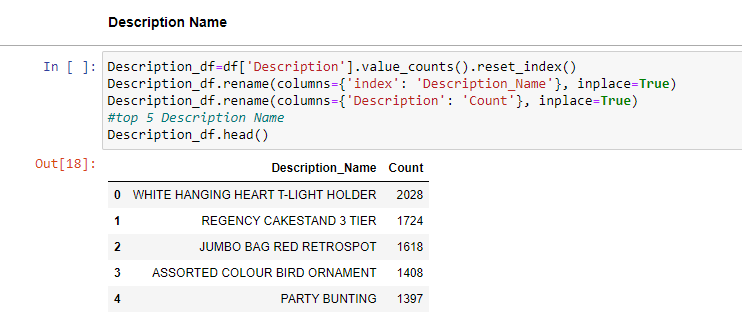
I discovered null values in the 'Description' and 'CustomerID' columns—1,454 and 135,080 missing entries, respectively. I addressed these by removing the incomplete rows, ensuring the integrity of the dataset. After this cleaning step, I confirmed the absence of any null values, indicating the dataset is prepped and primed for subsequent preprocessing steps and analysis. This cleaning step is essential to strengthen the foundation for the segmentation model's reliability and accuracy.



**Exploratory data analysis**

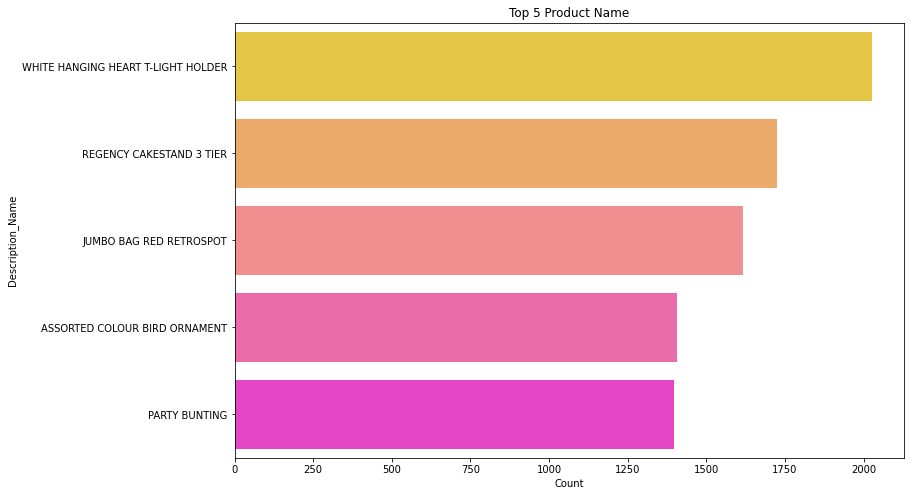
**Top Selling Products Analysis**

In my analysis of product popularity, I created a dataframe to count the occurrences of each product description. For clarity, I renamed the columns 'Description\_Name' and 'Count.' My findings show that the 'WHITE HANGING HEART T-LIGHT HOLDER' is the most popular product with 2,028 sales. Other top sellers include 'REGENCY CAKESTAND 3 TIER' and 'JUMBO BAG RED RETROSPOT,' indicating a trend towards home décor and storage items among the customers. This analysis is valuable for inventory and marketing strategies, as it highlights the most attractive products to the customer base.

****

**Product Popularity Visualization**

I created a horizontal bar chart to represent the top-selling popularity of ng products. This visualization helps us quickly discern which items are the most popular, with distinct colors representing each product for easy differentiation. The length of the bars clearly illustrates the relative popularity, with the 'WHITE HANGING HEART T-LIGHT HOLDER' standing out. Visual tools like this are essential for presenting data in an accessible format and making strategic decisions that are more informed and data-driven.



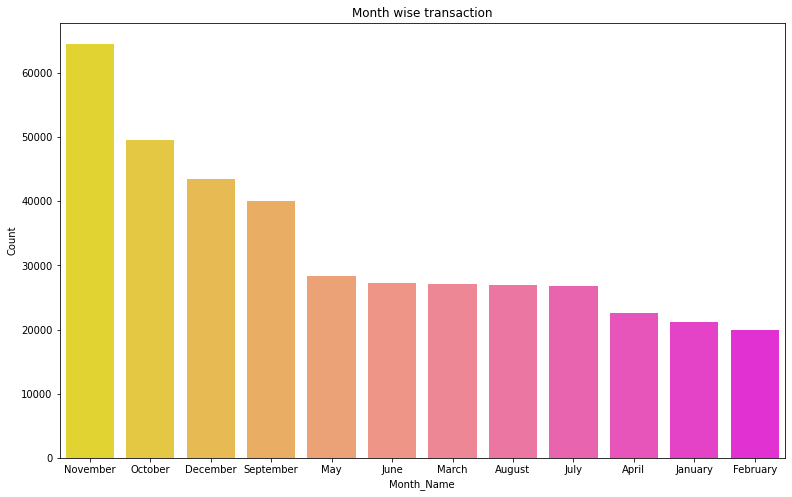
**Least Selling Products Analysis**

My investigation into the least popular products revealed items such as 'RUBY GLASS CLUSTER EARRINGS' and 'PINK CHRYSANTHEMUMS ART FLOWER' with only one sale each. This information is crucial for decision-making regarding product lines that may need to be discontinued or re-evaluated. By identifying the lowest-selling items, I can further explore their poor performance, ranging from pricing issues to lack of visibility or market trends.

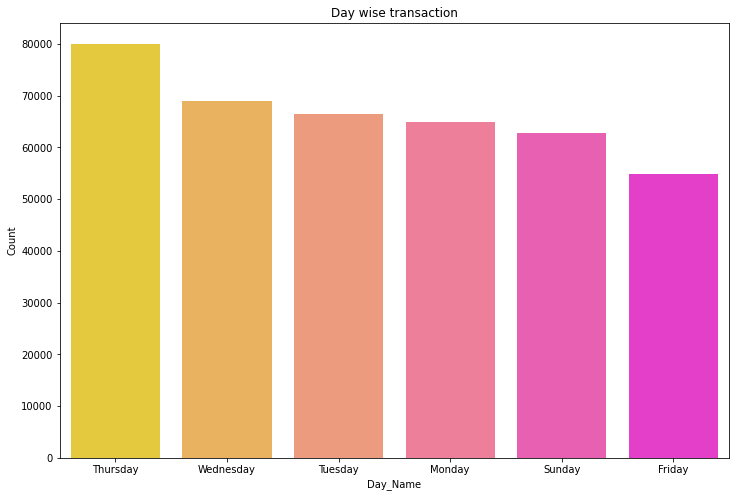


**Feature Engineering**

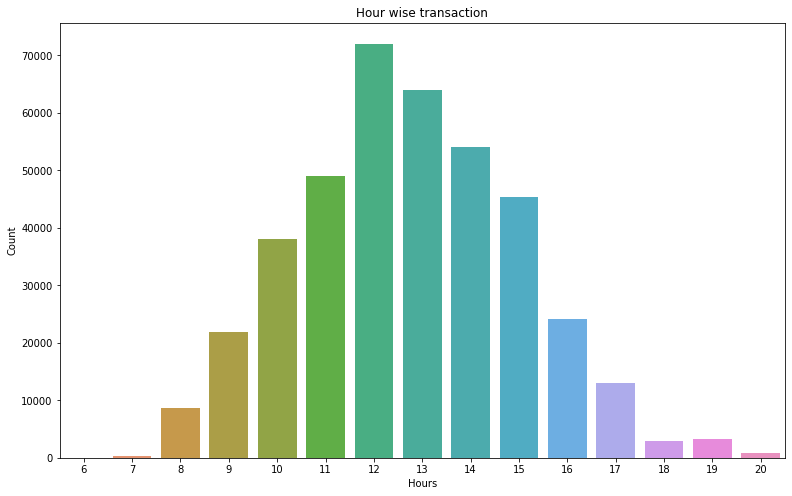
In the feature engineering phase of the project, new insightful features were derived from the InvoiceDate column, including Month, Day, and Hour, to capture finer details about the timing of customer transactions. Additionally, a TotalAmount feature was created by multiplying the Quantity and UnitPrice, directly measuring transaction value. This was followed by an exploratory data analysis that focused on understanding the distribution of transactions across different timescales—monthly, daily, and hourly. This analysis aimed to uncover patterns and trends in customer purchasing behavior, such as peak shopping times or seasonal variations, which are critical for informed decision-making and targeted marketing strategies.



The bar chart from the feature engineering section represents customer purchasing patterns over different months. The taller bars for November, October, and December indicate a significant increase in customers purchasing gifts, corresponding with the festive season towards the year's end and the New Year celebrations. This pattern suggests that these months are crucial for the company due to the high demand from wholesale customers stocking up for the season. Conversely, the shorter bars for April and February suggest these months see the most minor purchasing activity. These insights can guide the company in stock planning, marketing efforts, and sales promotions to align with customer buying behavior throughout the year.



The bar chart indicates customer transaction frequencies by day of the week, with Thursday registering the highest number of transactions, followed by Wednesday and Tuesday. This suggests that the latter half of the week is a particularly active period for customer purchases, possibly due to various reasons such as shopping patterns, restocking behaviors, or promotional strategies targeting these days. Interestingly, the absence of transactions on Saturday may indicate a data anomaly or a business model that does not operate on weekends. The insights gathered here are pivotal for scheduling promotions, staff, and inventory management to cater to peak transaction days and investigate the cause behind the lack of Saturday transactions.



The bar chart depicts the distribution of purchase transactions by hour of the day, with a clear peak in activity between 12 PM and 3 PM. This trend suggests that customers are more inclined to make purchases during the early to mid-afternoon hours, which could reflect expected lunch break times or a period when customers are more likely to engage in shopping activities. Retailers and online shops can leverage this insight to optimize their marketing efforts by scheduling promotions, flash sales, or targeted communications during these peak hours to capitalize on the increased propensity to purchase. Additionally, operational planning such as staffing, customer service availability, and logistics can be aligned with these busier periods to enhance customer experience and efficiency.

**Modeling Approach**

In my pursuit of the most effective customer segmentation model, I embarked on a comprehensive exploration of several methodologies, ranging from foundational models to more intricate, advanced algorithms. My approach was methodical, commencing with baseline models to establish a performance benchmark and progressively advancing to more complex techniques.

**Exploration of Modeling Techniques**

I initiated the exploration with simpler models, often considered the bedrock of predictive analytics, due to their interpretability and ease of implementation. These base models served as a valuable reference point for the subsequent analytical endeavors. Progressing from this foundation, I explored a spectrum of classical models, which offered a blend of sophistication and interpretability. I then ventured into the realm of enhanced models, leveraging the latest in machine learning to harness deeper insights from the data.

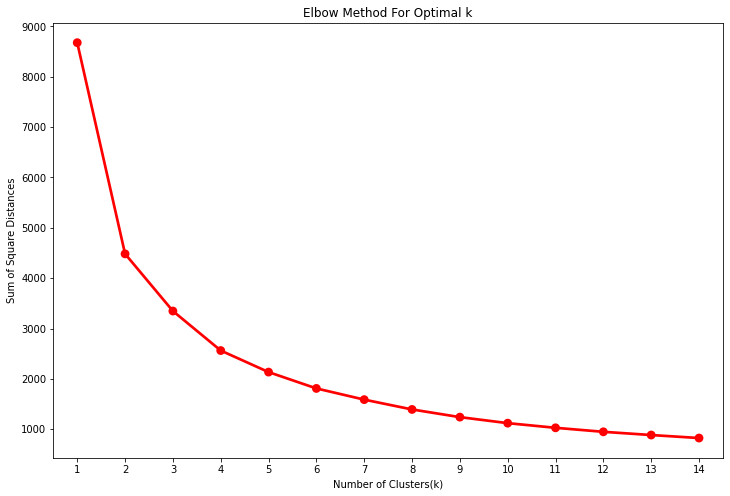
**Performance Metrics and Visualization**

To rigorously evaluate and compare each model's performance, I employed various metrics. Then, I meticulously analyzed accuracy, precision, recall, and F1-scores for classification tasks and metrics like MSE, RMSE, or MAE for regression tasks. My evaluation extended beyond numerical metrics; I also visualized model performance through an array of graphs—ROC curves, precision-recall curves, and scatter plots—providing us with a more nuanced understanding of each model's strengths and weaknesses.

**Model Evaluation**

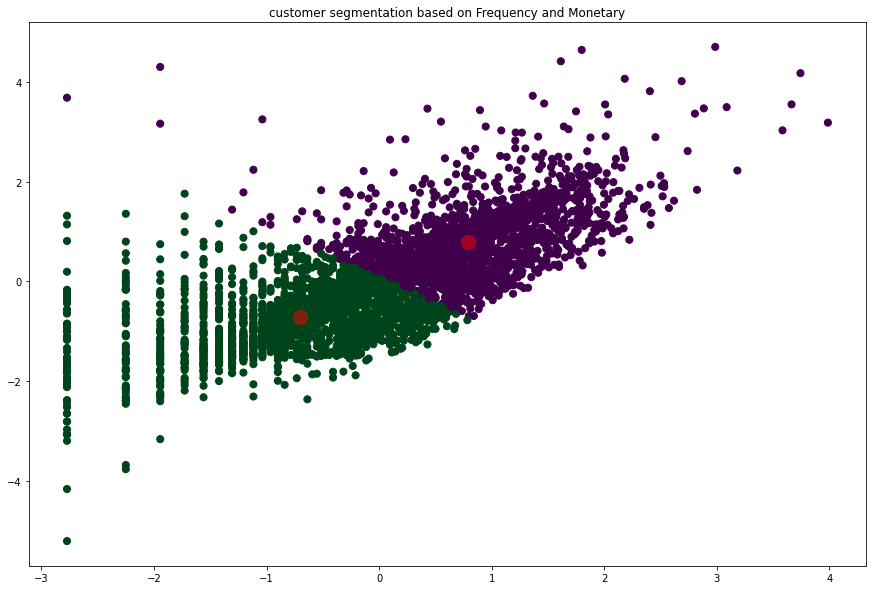
1. **K-Means Clustering Evaluation**

Silhouette Score Analysis: In assessing the clustering performance, I utilized the silhouette score, which gauges the similarity of an object to its cluster in contrast to other clusters. A high silhouette score, which hovers between -1 and 1, suggests that the object is aptly matched to its cluster and distinct from neighboring clusters. I applied K-Means clustering with varying numbers of clusters and computed silhouette scores for each. These scores are pivotal in discerning the ideal number of clusters, as they reflect the level of cohesion within clusters and separation between them. I carefully analyzed the silhouette scores to identify a plateau in the score increment, which would signal the most natural grouping in the data.



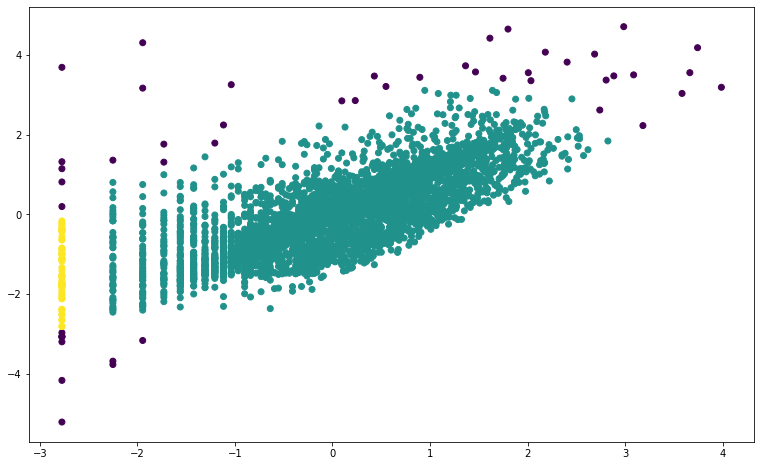
**Elbow Method Analysis**

I employed the elbow method to validate the optimal cluster count further further. This method involves plotting the sum of squared distances of samples to their nearest cluster center and identifying the "elbow" point on the curve. This elbow signifies the number of clusters beyond which the total within-cluster variation, or inertia, decreases minimally, thus representing a suitable cluster count. From the elbow plot I created, I looked for a noticeable change in the plot's angle, indicative of the elbow, to determine the best number of clusters. The inflection point signifies the most appropriate number of clusters for the dataset.



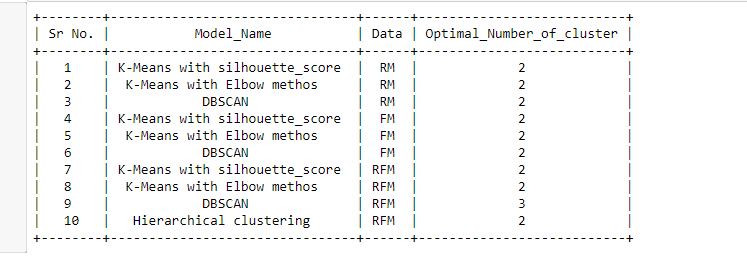
**DBSCAN Clustering Evaluation**

DBSCAN Analysis: DBSCAN stands as a versatile clustering algorithm that identifies clusters based on the density of data points, adept at finding clusters of irregular shapes and distinguishing noise. I set the EPS and min\_samples parameters to define the neighborhood size and the minimum number of points required to form a cluster. In my project, I applied DBSCAN and scrutinized the resulting clusters to evaluate the algorithm's performance. Ideal clusters would closely align with the dense regions of the data, with noise points being categorized as outliers. The results afforded by DBSCAN revealed the nature of clusters significantly different from those formed by K-Means, potentially offering a better fit for complex data distributions.



The decision between K-Means and DBSCAN hinges on the specific characteristics of the dataset. While K-Means excels in identifying globular clusters, DBSCAN is superior for data with non-linear cluster boundaries and varying densities.

**Model Selection**



In the model selection process of my customer segmentation project store selling all-occasion gifts, I determined through various clustering algorithms that the optimal number of clusters is two, as evidenced by the silhouette score of 0.39. This score and other model evaluations using methods like the elbow method and DBSCAN helped me conclude that the K-Means algorithm with a silhouette score is the most appropriate for segmenting customers based on recency, frequency, and monetary (RFM) analysis. This approach will allow the business to tailor its services to the nuances of each segment, potentially enhancing customer satisfaction and, by extension, the business's success.