**Technical Report: Retail Data Analysis**

---

**Date**: October 4, 2024

---

**Introduction**

This project aimed to analyze and forecast retail sales data for an online retailer. The dataset used includes transactional data, product information, customer details, and sales metrics. The analysis focuses on understanding sales trends, performing customer segmentation using RFM (Recency, Frequency, Monetary) analysis, and predicting future sales using time-series forecasting techniques. Additionally, we implemented clustering models to identify customer groups for targeted marketing strategies.

**Data Cleaning/Preparation**

The dataset was sourced from UCI's online retail dataset. Key steps for cleaning and preparing the data included:

- **Loading the data** from a CSV file and inspecting it for missing values and inconsistencies.

- **Removing missing values**: Missing values in key columns such as *CustomerID* and *InvoiceDate* were dropped to maintain data integrity.

- **Date parsing**: The *InvoiceDate* column was converted to a proper datetime format to facilitate time-based analysis.

- **Creating new columns**: A TotalSales column was generated by multiplying *Quantity* and *UnitPrice* to compute the total sales for each transaction.

Only data from the United Kingdom was selected for analysis to narrow down the scope, as it had the most consistent data in the dataset.

---

**Exploratory Data Analysis (EDA)**

**Sales Analysis**

- **Total Sales**: The total sales from the United Kingdom were calculated as **£6,767,873.39**.

- **Average Sales per Transaction**: The average sales per transaction were **£34.16**.

- **Top Products by Sales**: The top 10 products by total sales were identified, with the highest-grossing product being the REGENCY CAKESTAND 3 TIER, contributing £103,122.85 in total sales.

**Visualizations**

- A bar chart was generated to display the top 10 products by sales, highlighting the most popular items purchased in the dataset.

---

**Model Selection**

**Time-Series Forecasting**

To predict future sales, we applied the Holt-Winters Exponential Smoothing model, which captures both trend and seasonality in the sales data. The data was grouped by month to facilitate monthly sales forecasting.

- **Model Details**: The model used the following parameters:

* Trend: Additive
* Seasonality: Additive with a seasonal period of 6 months.

- **Forecast**: The model forecasted sales for the next 12 months, with predictions showing an expected growth in sales during the second half of the following year.

**Clustering Analysis (Hierarchical Clustering)**

We used Hierarchical Clustering to group customers based on their RFM metrics. The “ward” method was employed to minimize the variance within clusters. A dendrogram was plotted to visualize the clustering hierarchy, and a distance threshold of 20 was used to cut the dendrogram and form clusters.

---

**Model Analysis**

**Sales Forecasting**

The forecasted sales showed a steady increase, with key predictions:

- January 2012: Forecasted sales of £660,994.93.

- May 2012: Forecasted peak sales of £1,028,587.85.

- November 2012: A significant increase in sales with a forecasted value of \*\*£1,202,434.27\*\*.

The model effectively captured the seasonal trend, highlighting peaks likely due to increased holiday sales.

**Customer Segmentation Using RFM**

- **RFM Scoring**: Customers were scored based on Recency, Frequency, and Monetary values, with scores ranging from 1 to 4 for each metric. An overall \*\*RFM Score\*\* was calculated to rank customers.

- **Segments**: Customers were grouped into the following segments:

* High Value: Customers with high RFM scores, indicating frequent purchases and high spending.
* Low Value: Customers with low RFM scores, indicating infrequent purchases and low spending.
* Potential Value and Valuable segments captured customers with moderate RFM metrics.

- **Cluster Summary**: After performing Hierarchical Clustering, customers were segmented into 9 clusters. Each cluster's characteristics, such as average Recency, Frequency, and Monetary value, were calculated.

---

**Conclusion and Recommendations**

The Holt-Winters model forecasted an increase in sales over the next year, with notable peaks around holiday seasons. This suggests a need for businesses to prepare for high demand during these times, optimizing inventory and marketing efforts accordingly.

RFM analysis and clustering of customer segmentation revealed that customers can be segmented into distinct groups, allowing for personalized marketing strategies:

**High-Value Customers** should be rewarded with loyalty programs and personalized offers.

**At-risk customers** (those with high recency and low frequency) should be targeted with re-engagement campaigns.

**Low-Value Customers** could be targeted with general promotions to increase frequency and spending.

This project demonstrated the value of time-series analysis and clustering techniques in retail data, providing actionable insights for improving customer engagement and sales strategies.

**Recommendations**

The following recommendations can be applied based on the data analysis:

- Use customer segmentation to deploy personalized marketing strategies based on the RFM segments, rewarding loyal customers while re-engaging less active ones.

- Prepare for high demand during forecasted peak periods (such as November) by ensuring sufficient inventory levels.

- Experiment with different seasonal periods and more advanced machine learning models for better sales predictions and customer behavior analysis.

Dua, D., & Graff, C. (2019). **UCI Machine Learning Repository: Online Retail Data Set**. University of California, Irvine, School of Information and Computer Sciences. <https://archive.ics.uci.edu/ml/datasets/online+retail>