Online Retail Data Analysis Report
Arihant Tripathi, Zihe(Peter) Zhang, Alan Wu

Introduction

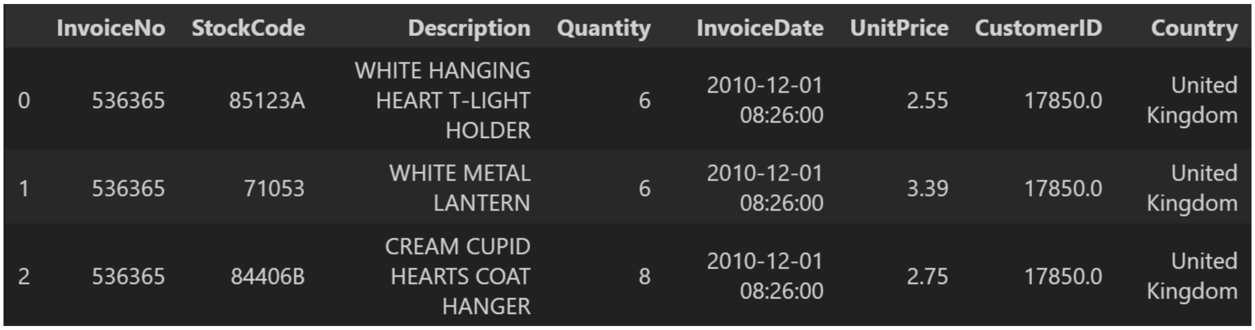
The dataset offers a detailed look into the sales transactions of an online retail business. The data includes various columns such as InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country, providing a comprehensive view of each transaction. The exploratory data analysis reveals some key insights: the average quantity per transaction is around 13 items, and the average unit price is approximately £3.28. Most transactions (80) are recorded in the United Kingdom, followed by 20 in France. The dataset reflects a diverse range of products and customer interactions, with invoice numbers and customer IDs indicating a variety of transactions.

In this project, our primary objective is to predict the unit price of items based on various features in the dataset, such as quantity, purchase date, and country groupings. To achieve this, we will employ four distinct approaches, with the first focusing on clustering and the subsequent three on predictive analysis. The first approach would be the comparison between RFM Customer Segmentation Analysis and K-means Clustering. The other predictive approaches are Regression Analysis, Decision Tree Analysis, and Random Forest Prediction.

At the conclusion of each approach, we will employ metrics such as R-squared (R^2) and Mean Squared Error (MSE) to evaluate the accuracy and predictive capability of each model. This will enable us to compare the effectiveness of different methodologies in predicting unit prices and to identify the best-performing models for potential practical applications.

One thing to note is that we will also try the time series prediction for total sales(total amount of monetary value spent on orders per month), although we might not have enough data to fully tune the requirements of the algorithms. This portion is not the task of the project.

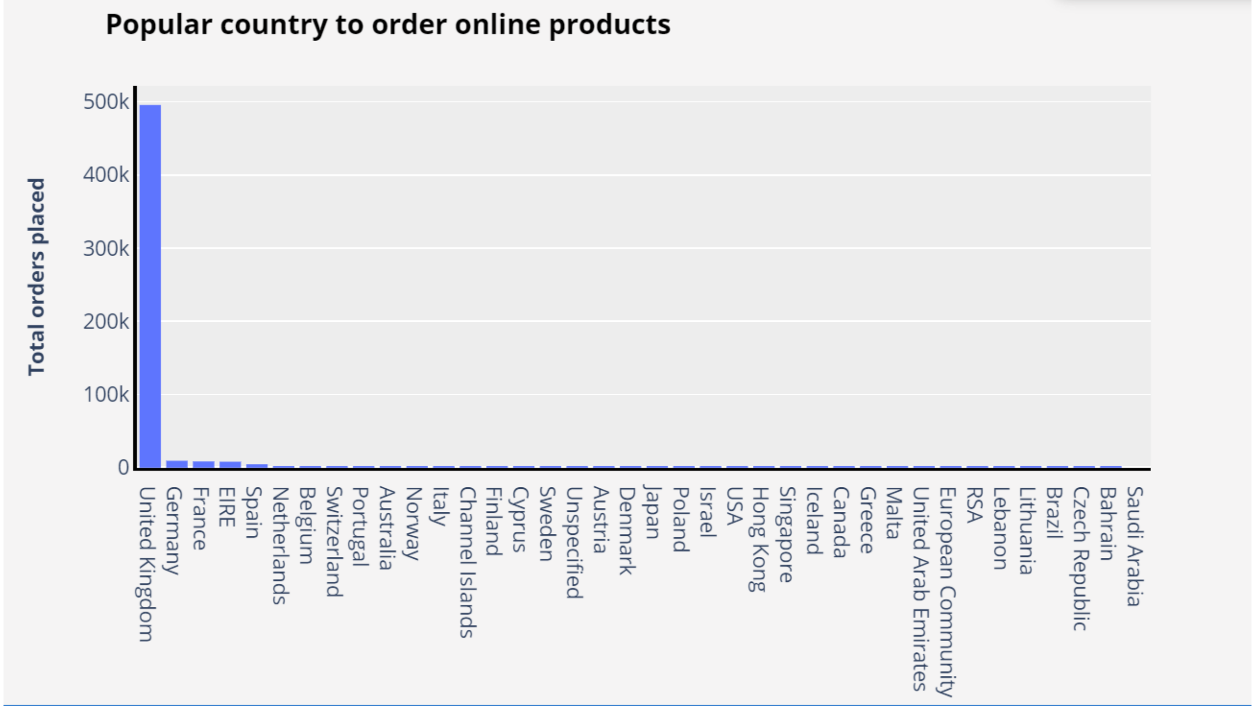
Sample data snippet:



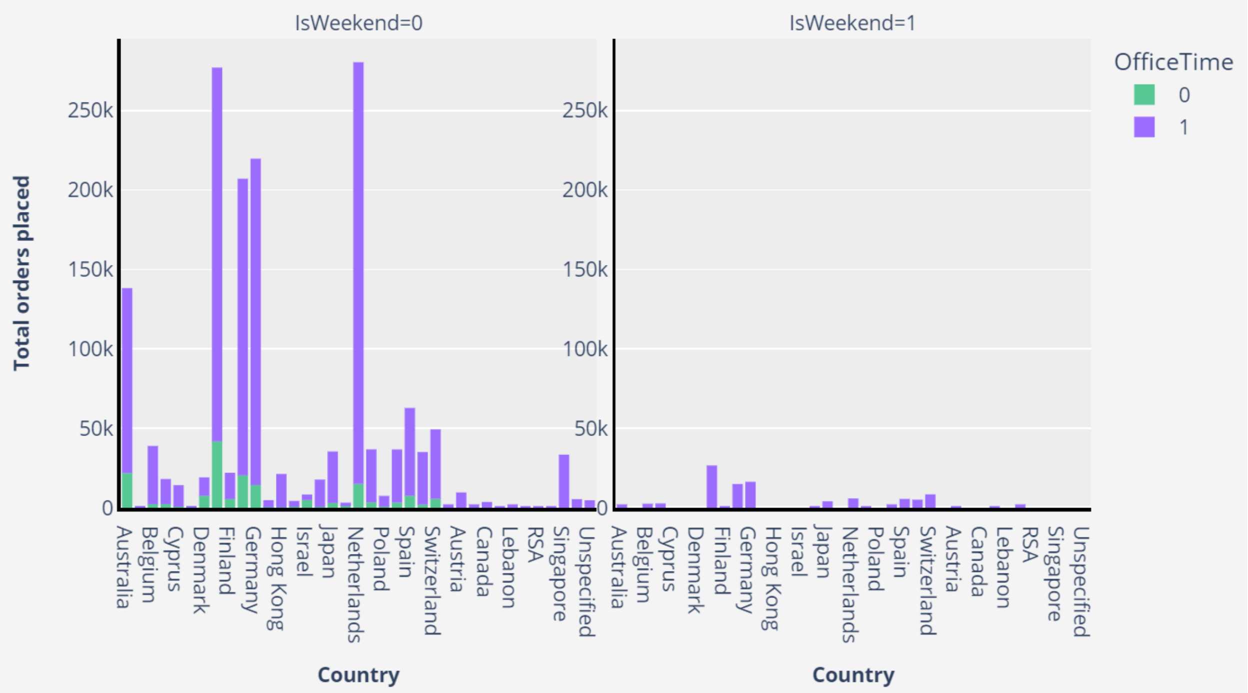
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom

Exploratory Data Analysis

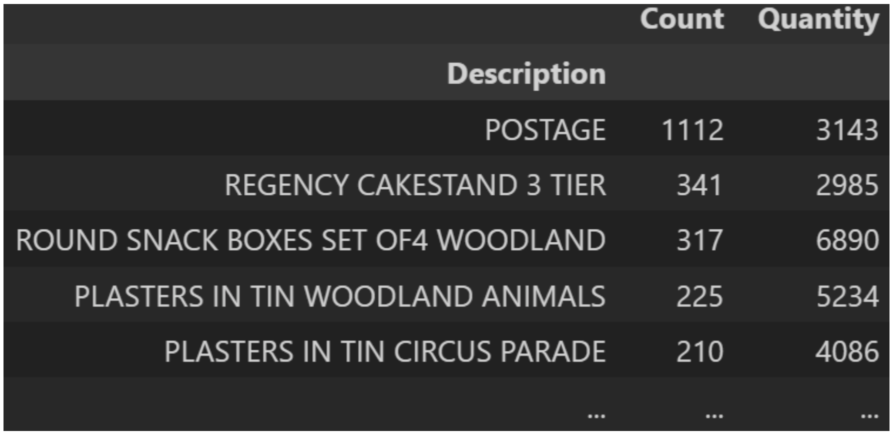
1. **Handling Missing Data**:
   - The columns “Description” and “CustomerID” contain missing data, with 1,454 and 135,080 missing entries respectively.
   - Due to the substantial amount of missing data and limited predictive value, we have decided to exclude the “CustomerID” column from our analysis.
   - For missing “Description” entries, we uniformly replaced them with “No Description.”
2. **Handling ‘InvoiceDate’ Time Variable**:
   - We extracted day, month, year, and time components, saving them as additional columns.
3. **Handling ‘Quantity’ and ‘UnitPrice’**:
   - Converted these columns from ‘object’ type to ‘int’ and ‘float’ respectively for ease of calculation.
4. **Country vs. Number of Orders**:
   - The majority of the data originates from UK clients, accounting for 91.43% of the total dataset.



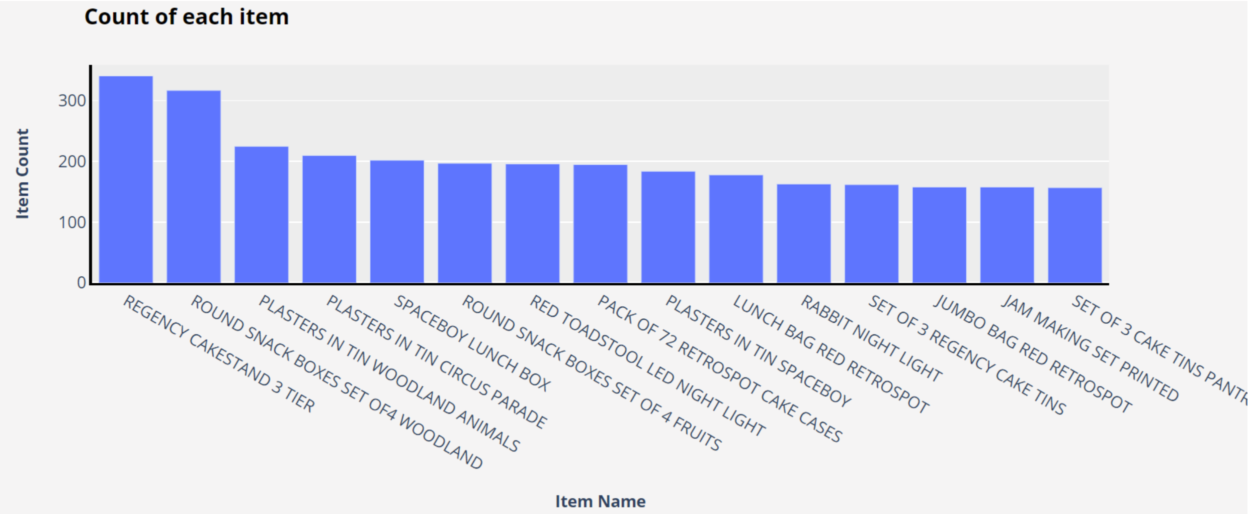
5. **New Columns: ‘isWeekend’ and ‘OfficeTime’**:
    - ‘isWeekend’: Indicates whether the purchase was made on a weekend (value = 1 for Friday to Sunday; 0 otherwise).
    - ‘OfficeTime’: Reflects if the order was placed during office hours (9:00 to 17:00), with 1 indicating within office hours.

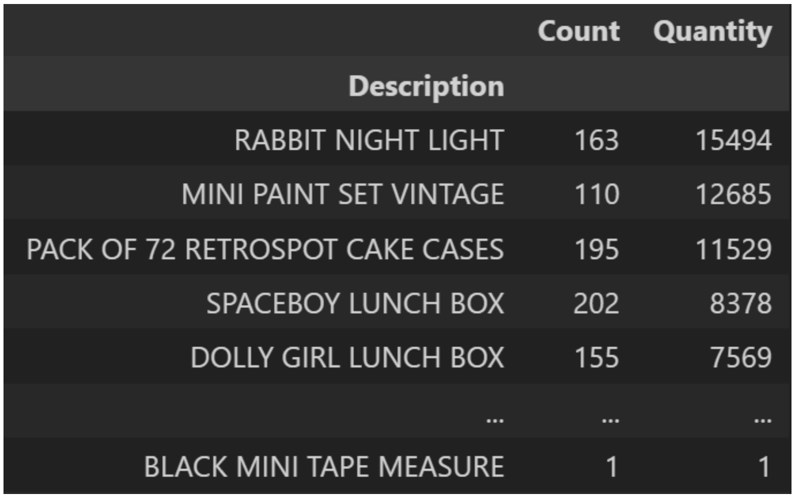


6. **Order Patterns by Time**:
    - Most orders, regardless of weekday or weekend, are placed during office hours.
    - There are significantly more purchases on weekdays compared to weekends.
7. **Insights from ‘Description’ for non-UK dataset**:
    - The most frequently purchased item is “Postage,” with 1,112 orders and a total quantity of 3,143. The least popular item is “ZINC WIRE SWEETHEART LETTER TRAY,” purchased only once with 8 units.
    - The most ordered item by quantity is “RABBIT NIGHT LIGHT” with 16,394 units across 163 orders. The least ordered item is “KINGS CHOICE GIANT TUBE MATCHES,” with just one unit sold.



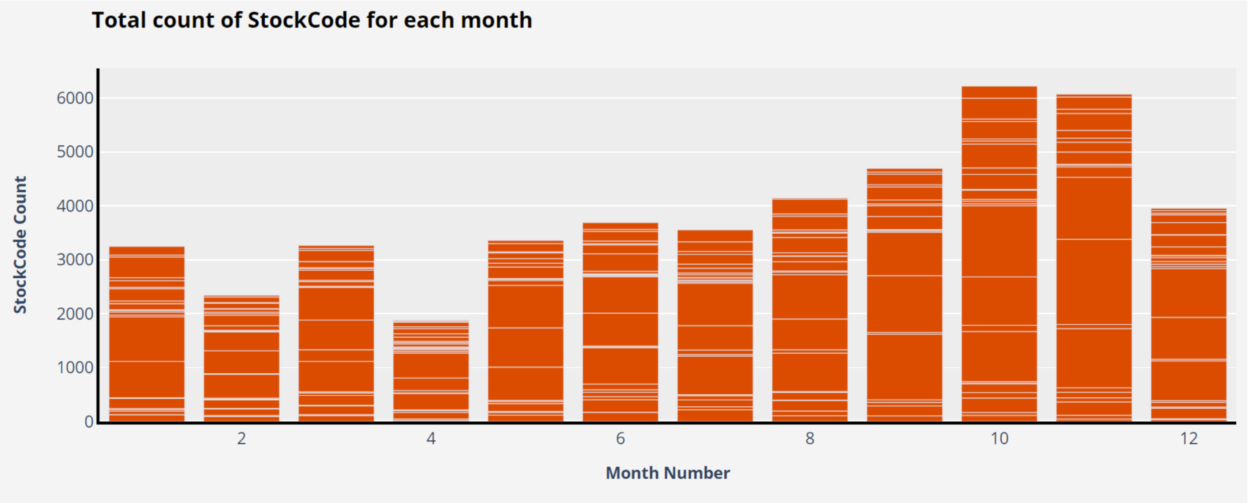
	Count	Quantity
Description		
POSTAGE	1112	3143
REGENCY CAKESTAND 3 TIER	341	2985
ROUND SNACK BOXES SET OF4 WOODLAND	317	6890
PLASTERS IN TIN WOODLAND ANIMALS	225	5234
PLASTERS IN TIN CIRCUS PARADE	210	4086
...	...	...

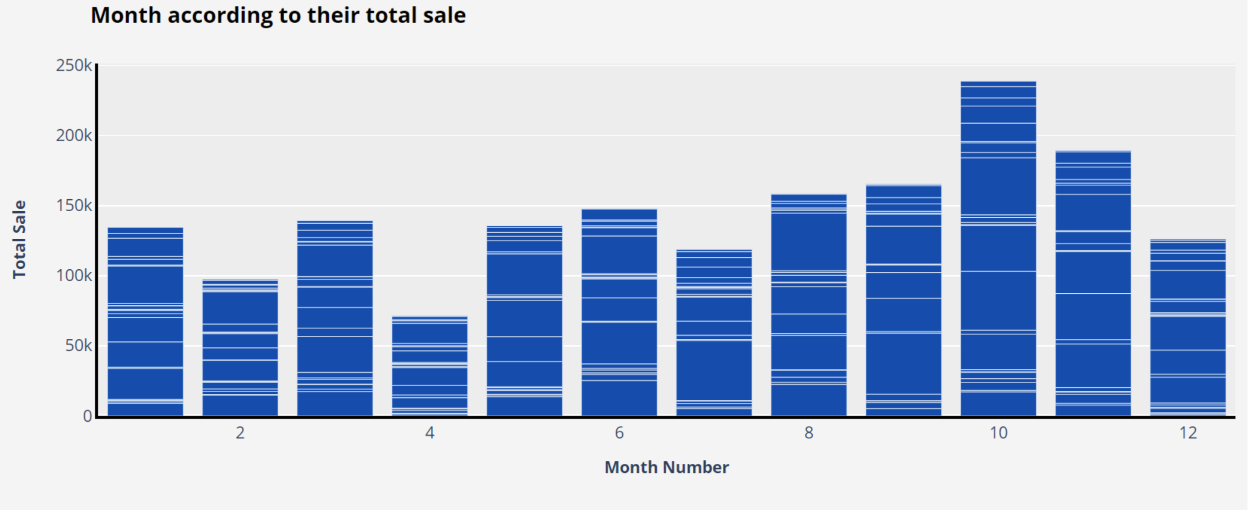




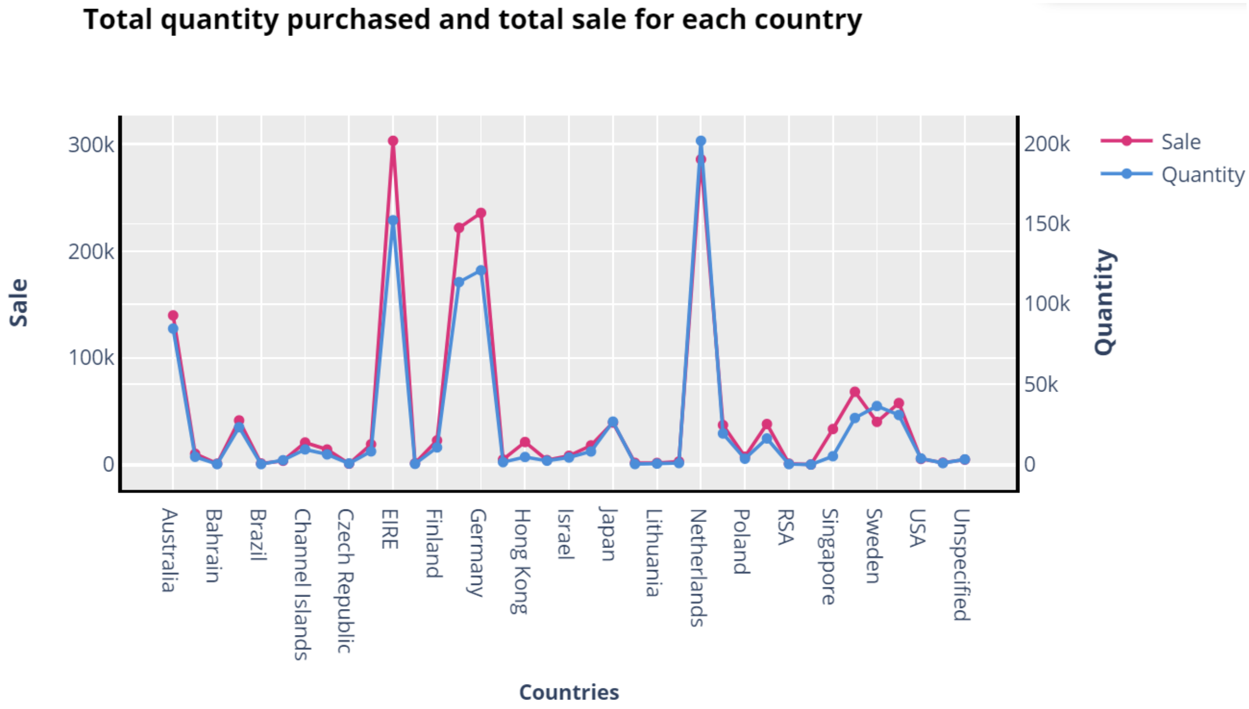
	Count	Quantity
Description		
RABBIT NIGHT LIGHT	163	15494
MINI PAINT SET VINTAGE	110	12685
PACK OF 72 RETROSPOT CAKE CASES	195	11529
SPACEBOY LUNCH BOX	202	8378
DOLLY GIRL LUNCH BOX	155	7569
...	...	...
BLACK MINI TAPE MEASURE	1	1

8. **Stock Code Count and Sales Value by Month**:
   - October records the highest stock code reads and sales values.
   - There is a notable correlation between stock code frequency and sales values across months.

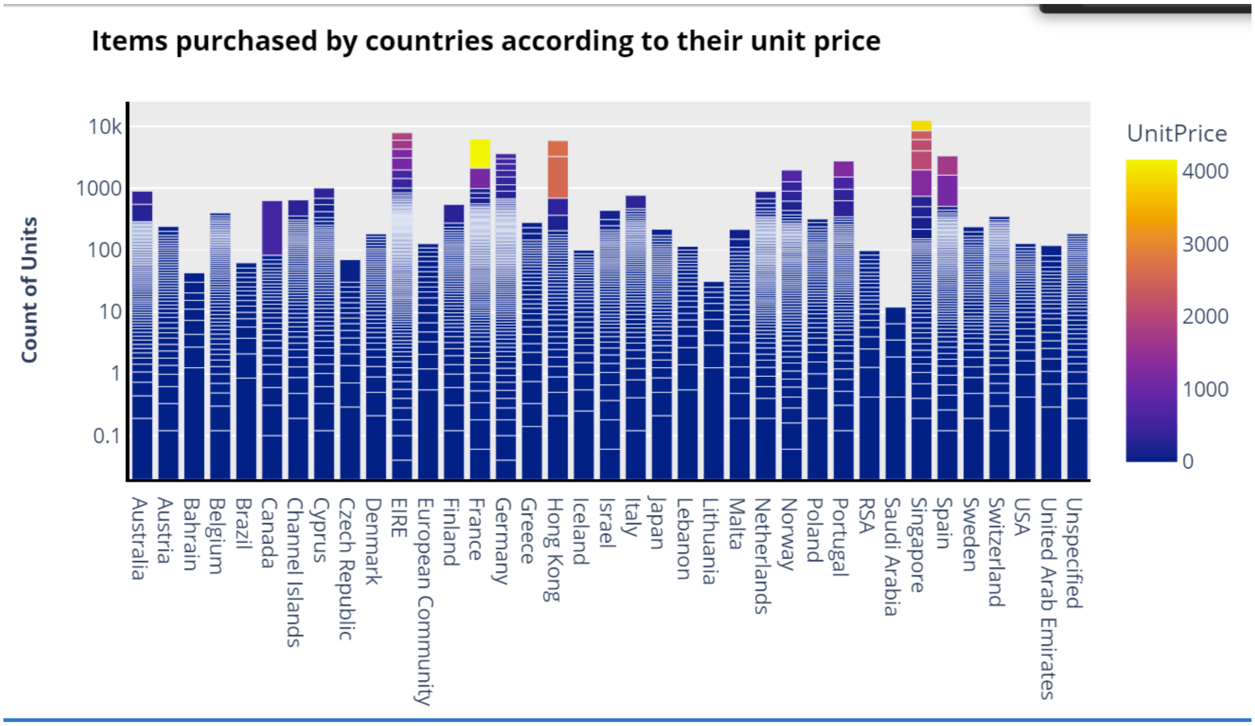




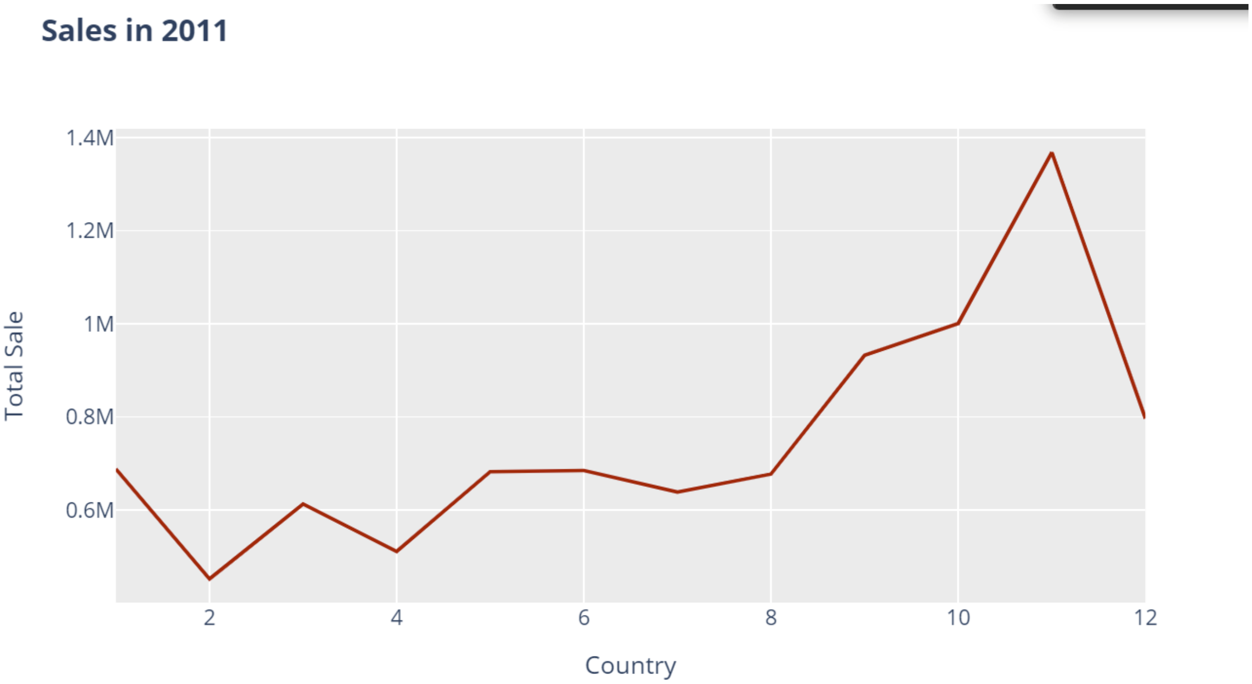
9. **Sales and Quantity by Country (Excluding UK)**:
    - EIRE and the Netherlands show notable trends: EIRE with higher sales value but lower quantities, and the Netherlands with higher quantities but lower sales value.



11. **Item Purchases by Country by Unit Price**:
    - France stands out for ordering items priced over £4,000. Singapore and Hong Kong follow with the second and third highest unit price purchases.

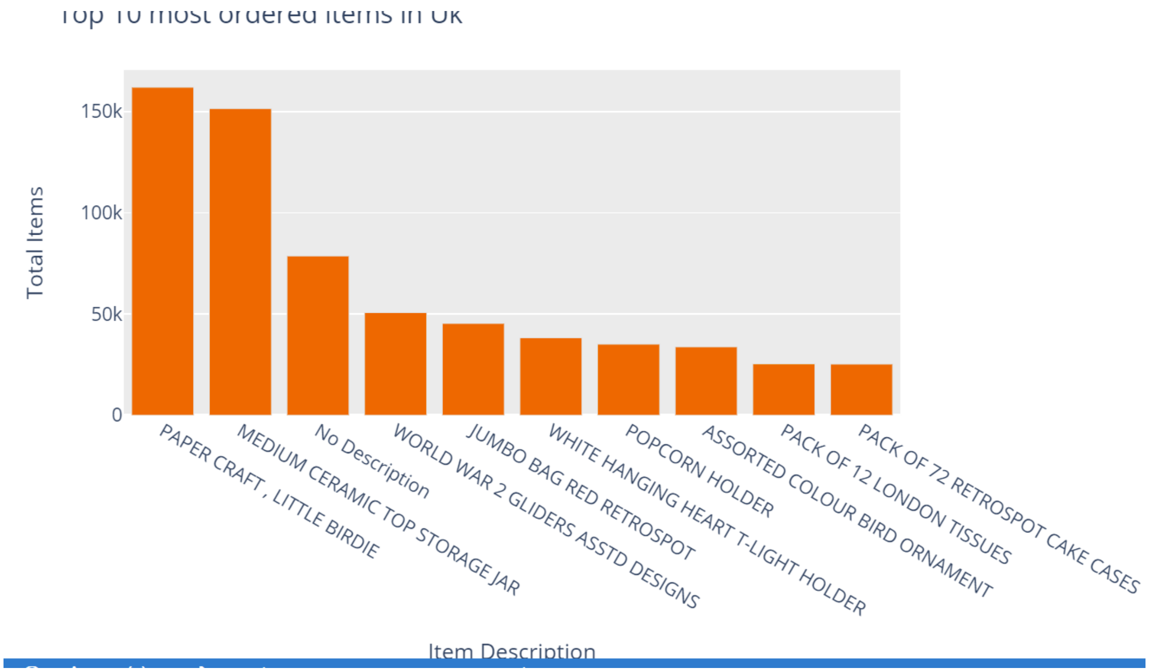

Items purchased by countries according to their unit price

12. **Sales Value Per Month in the UK**:
    - November peaks in sales with a total value of £1.37 million.

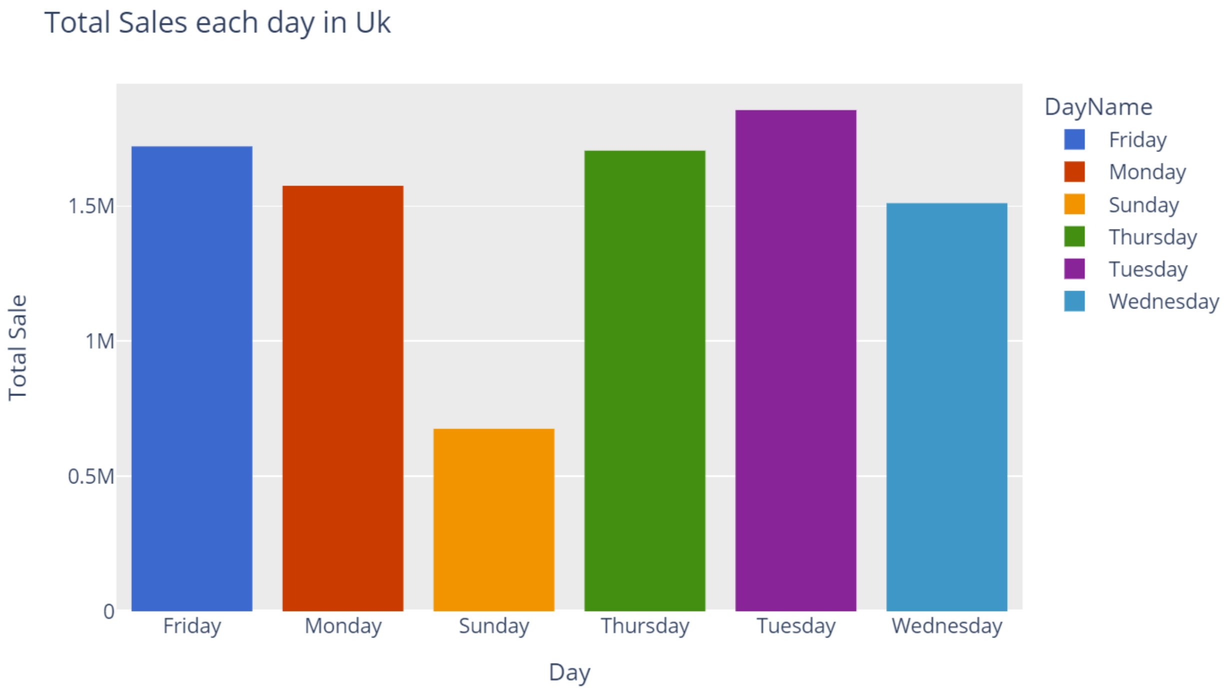

Sales in 2011

13. **Top 10 Purchased Items in the UK by Quantity**:

- The most purchased item is “Paper craft, little birdie.” The second is “Medium ceramic top storage jar.” Surprisingly, “No Description” ranks third, highlighting a significant number of unlabeled items.

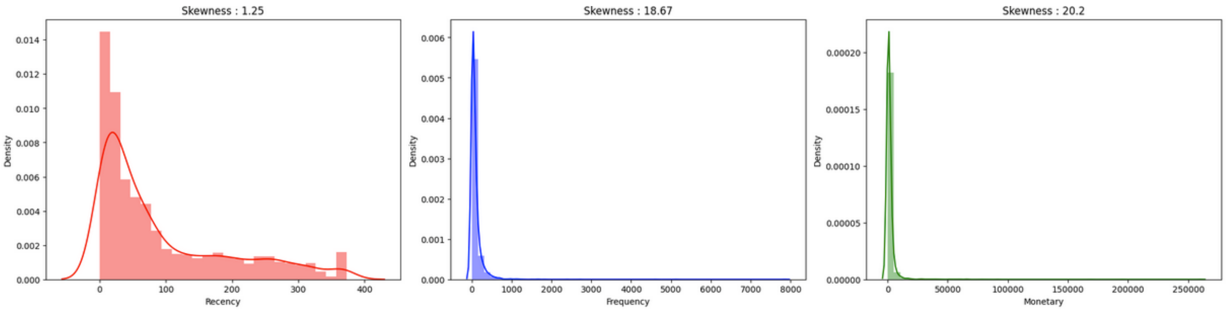


14. **Insights from ‘Total Sales in UK’**:
    - Peak purchasing days are Friday and Tuesday, with over 1.7M+ total sales each. Sunday has the least activity, with 645k total sales.



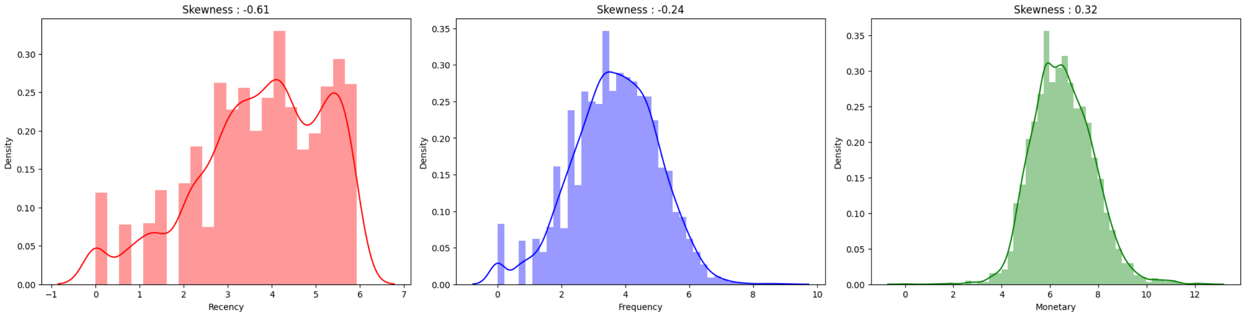
Model Method Algorithms

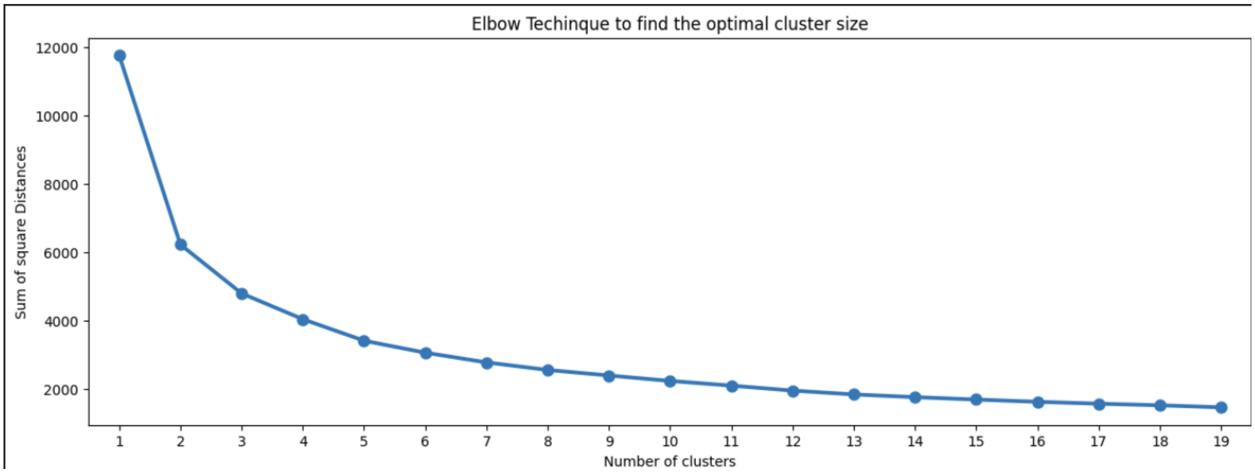
1. **RFM Customer Segmentation Analysis**
    - **Goal**: Categorize customers into loyalty groups based on purchasing behavior. Our initial approach involves segmenting customers using RFM (Recency, Frequency, Monetary) analysis to identify distinct shopping patterns. In parallel, we will implement k-means clustering. The aim here is to compare and contrast the clusters formed by the RFM model and the k-means algorithm, gaining insights into different customer behaviors.
    - **Methodology**:
        - Conducted RFM (Recency, Frequency, Monetary) analysis on the UK dataset, which has no missing data.
        - Recency: Time since last purchase; Frequency: Number of total orders; Monetary: Unit price multiplied by quantity.
        - Density plots for RFM variables are right-skewed, indicating recent purchases, frequent small orders, and a majority of orders valued between 0 - 25,000.
        - Boxplot quantiles used to assign RFM scores. Recency scores inversely related to value (higher score, less value), while Frequency and Monetary scores are directly related (higher score, better value).
        - Created "RFM_Group" and "RFM_Score" columns for segmentation.
        - Developed groups of "loyalty_level" classification: 'Passionate Customer', 'Frequent Shopper', 'Casual Shopper', 'Once a Year' based on quantiles of RFM scores.



2. **K-means Clustering**
    - **Goal**: Segment client data using k-means algorithm.
    - **Methodology**:
        - Transformed negative or zero values to positive for consistency.

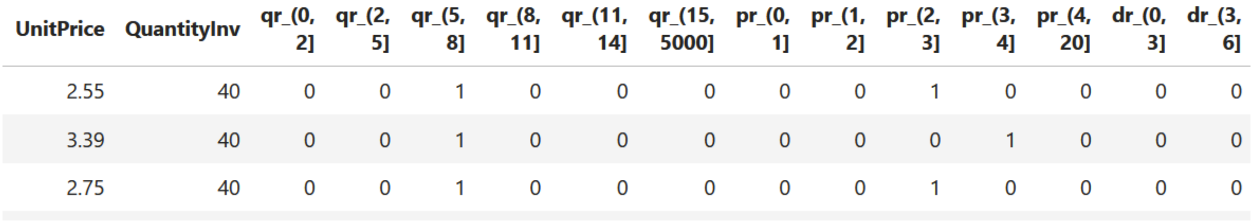
        - Applied log-transformation to address skewness in RFM variables.
        - Normalized and scaled data post-transformation.
        - Used elbow technique to determine optimal cluster number, selecting 3 as ideal.





3. **Time Series Prediction**
    - **Goal**: Predict future monthly total sales values.
    - **Methodology**:
        - Analyzed monthly sales data, noting stable trends with a peak in late 2011.
        - Employed SARIMAX and Exponential Smoothing Models for prediction, despite having only one year of data.
4. **Regression Analysis**
    - **Goal**: Predict unit price using various predictors. The second approach involves classical regression analysis, utilizing both numerical and categorical variables to predict the unit price of purchased items. This method will help us understand the direct relationships between various factors and the unit price.
    - **Methodology**:

        - Response variable (y): unit price. Predictors (X): scaled quantity per invoice, dummy variables for quantity range, price range, and date range.
        - Quantity range will be separated into 6 groups: (0,2] , (2,5], (5,8], (8,11], (11,14], and (14,5000].
        - Price range will be separated into 5 groups: (0,1], (1,2], (2,3], (3,4], and (4,20].
        - Date range will be separated into 4 groups: (0,3], (3,6], (6,9], and (9,12].
        - Applied linear regression model to UK-only data set.



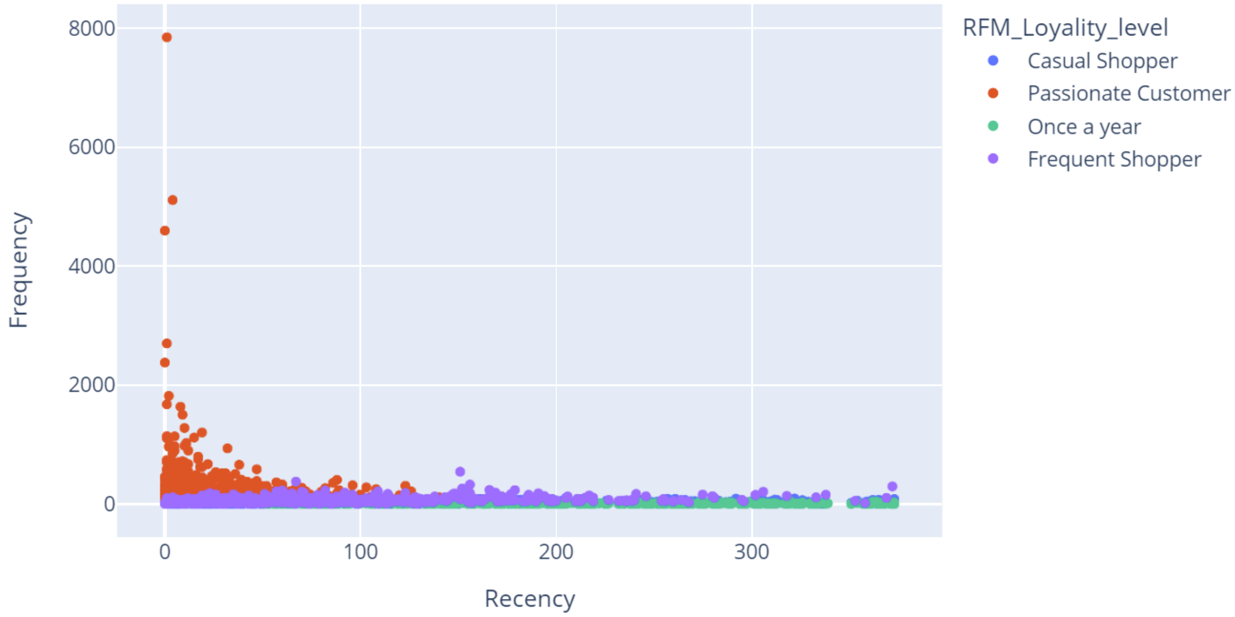
UnitPrice	QuantityInv	qr_(0, 2]	qr_(2, 5]	qr_(5, 8]	qr_(8, 11]	qr_(11, 14]	qr_(15, 5000]	pr_(0, 1]	pr_(1, 2]	pr_(2, 3]	pr_(3, 4]	pr_(4, 20]	dr_(0, 3]	dr_(3, 6]
2.55	40	0	0	1	0	0	0	0	0	1	0	0	0	0
3.39	40	0	0	1	0	0	0	0	0	0	1	0	0	0
2.75	40	0	0	1	0	0	0	0	0	1	0	0	0	0

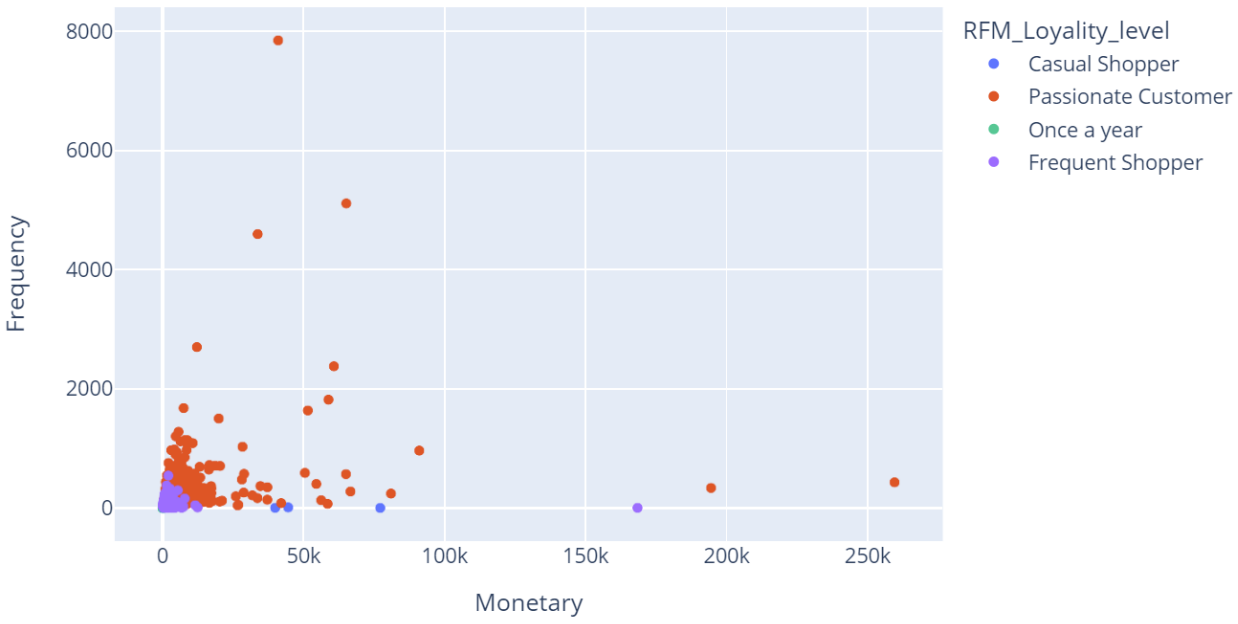
Note: the picture didn’t fit the columns of dr_(6,9] and dr(9,12]. They should be after dr_(3,6]

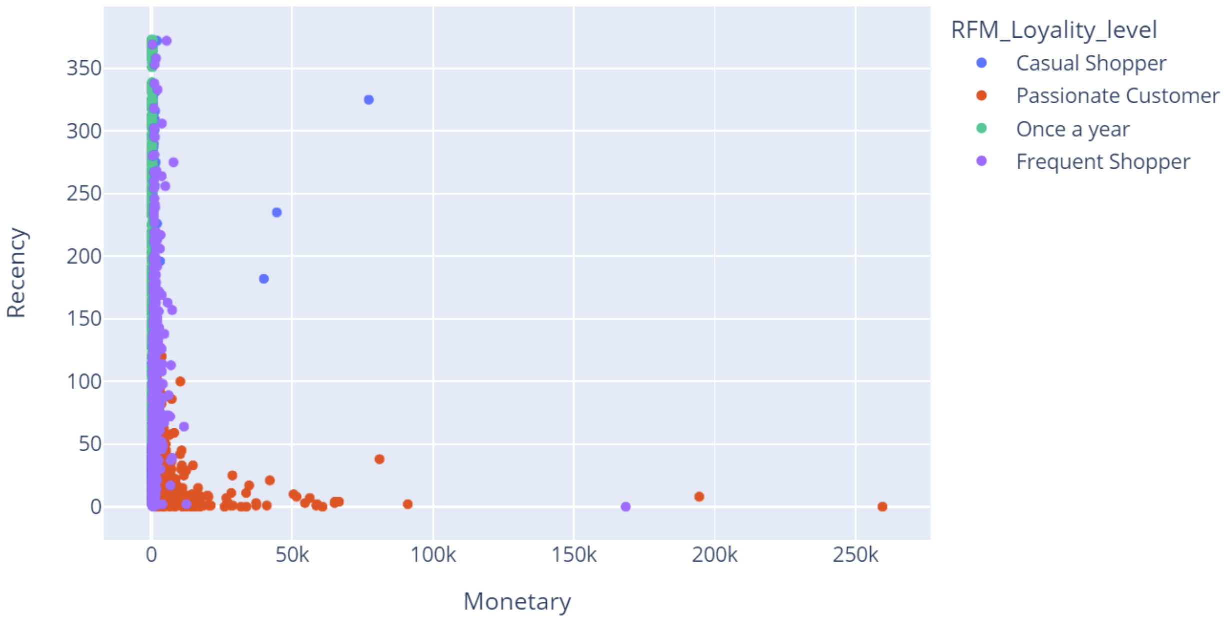
5. **Decision Tree**
    - **Goal**: Cluster similar purchase habits and predict unit price. In our third approach, we plan to use a decision tree model to group similar purchase habits and predict the unit price of items. This method will enable us to visualize and understand the decision paths that lead to different unit price outcomes.
    - **Methodology**:
        - Utilized a single decision tree.
        - Uncertainty about performance compared to regression, depending on non-linear relationships between predictors and unit prices.
6. **Random Forest**
    - **Goal**: Achieve accurate predictions of unit price.
    - **Methodology**:
        - An advanced approach using multiple trees to balance variance and bias.
        - Anticipated as the most accurate method due to its computational complexity.

Results of Analysis

1. **RFM Clustering Results**
   - **RFM Analysis**:
     - Recency vs Frequency: Right-skewed; passionate customers are concentrated within 0-100 recency days. 'Once a year' customers are found mostly on the second half of the x-axis.
     - Monetary vs Frequency: All loyalty levels are clustered in the bottom left corner, with only passionate shoppers showing high frequency and monetary values.
     - Monetary vs Recency: Vertical clustering around the y-axis. Passionate shoppers often have high monetary values, with some casual shoppers also showing relatively high monetary and recency values.

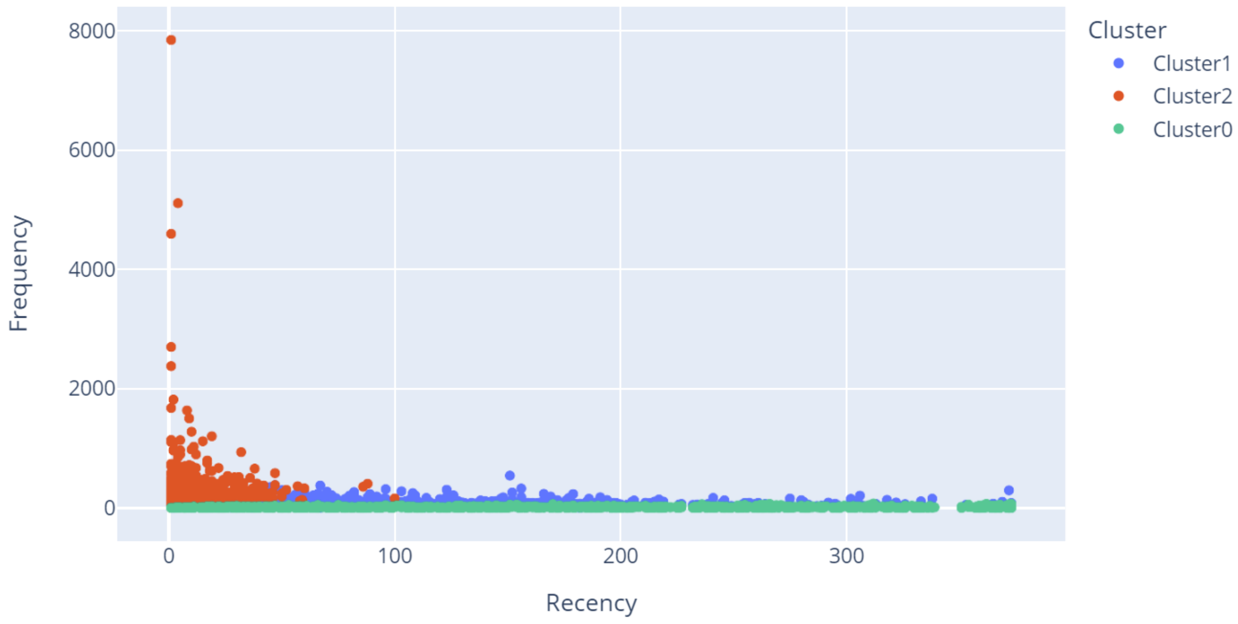


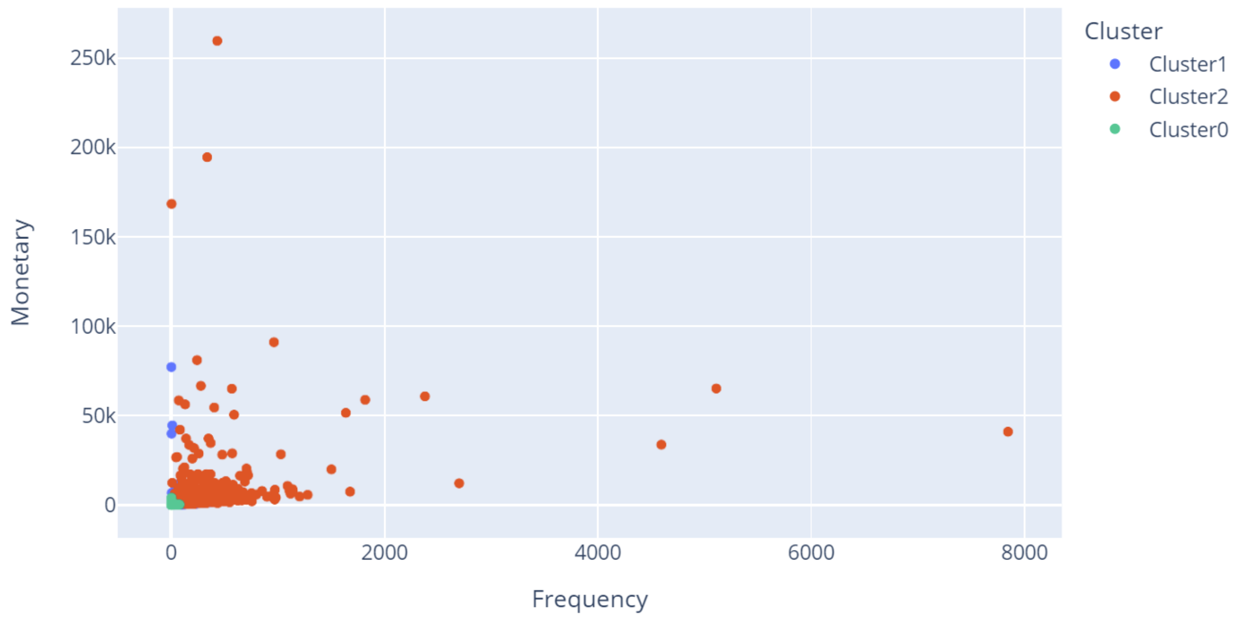


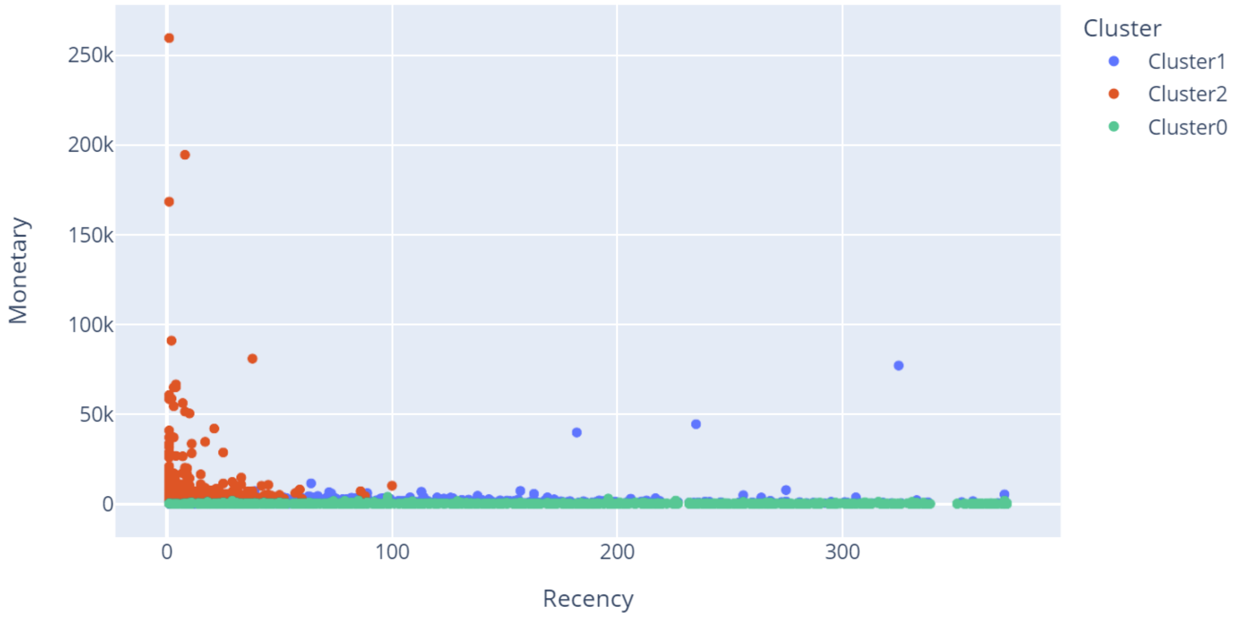


2. **K-means Clustering Results**
    - **Clustering Patterns**:
        - Recency vs Frequency: Clusters 0 and 1 are close to the x-axis, resembling 'frequency' and 'once a year' shoppers. Cluster 2 aligns with passionate shoppers.
        - Monetary vs Frequency: Majority are in cluster 2 (passionate shoppers), with some in cluster 1 (frequent shoppers).

- Monetary vs Recency: Clusters 0 and 1 are near the y-axis, with cluster 2 extending to high monetary values, indicating passionate shoppers.

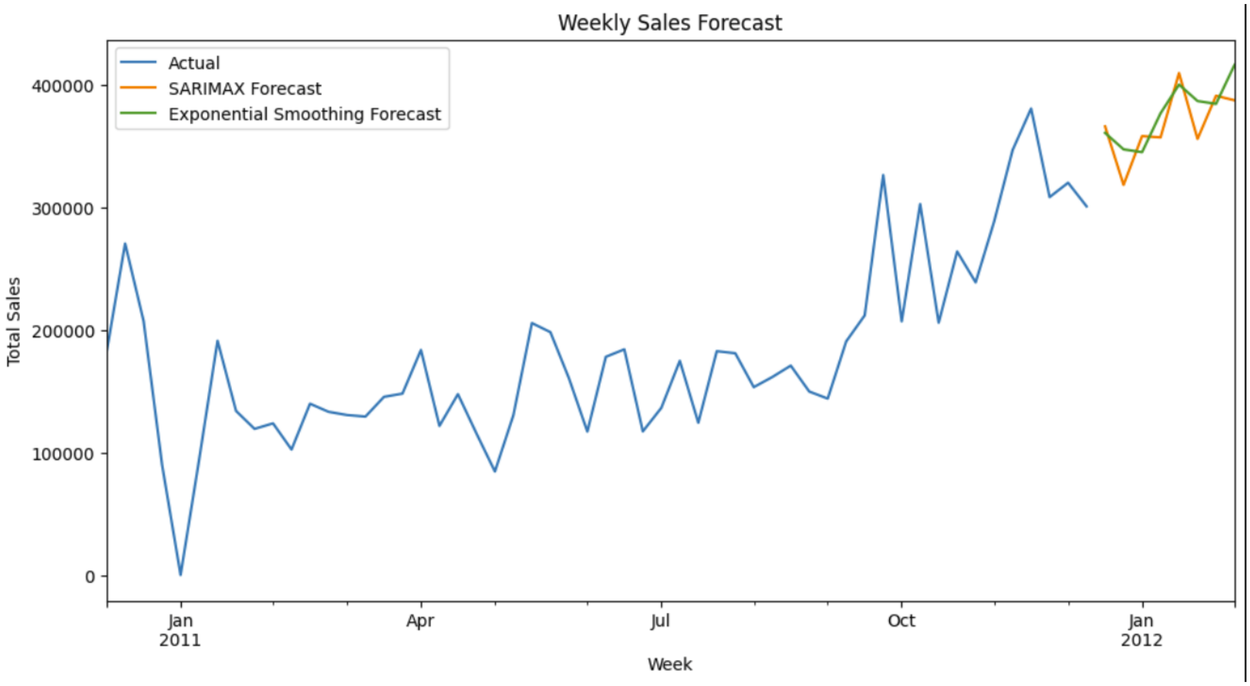


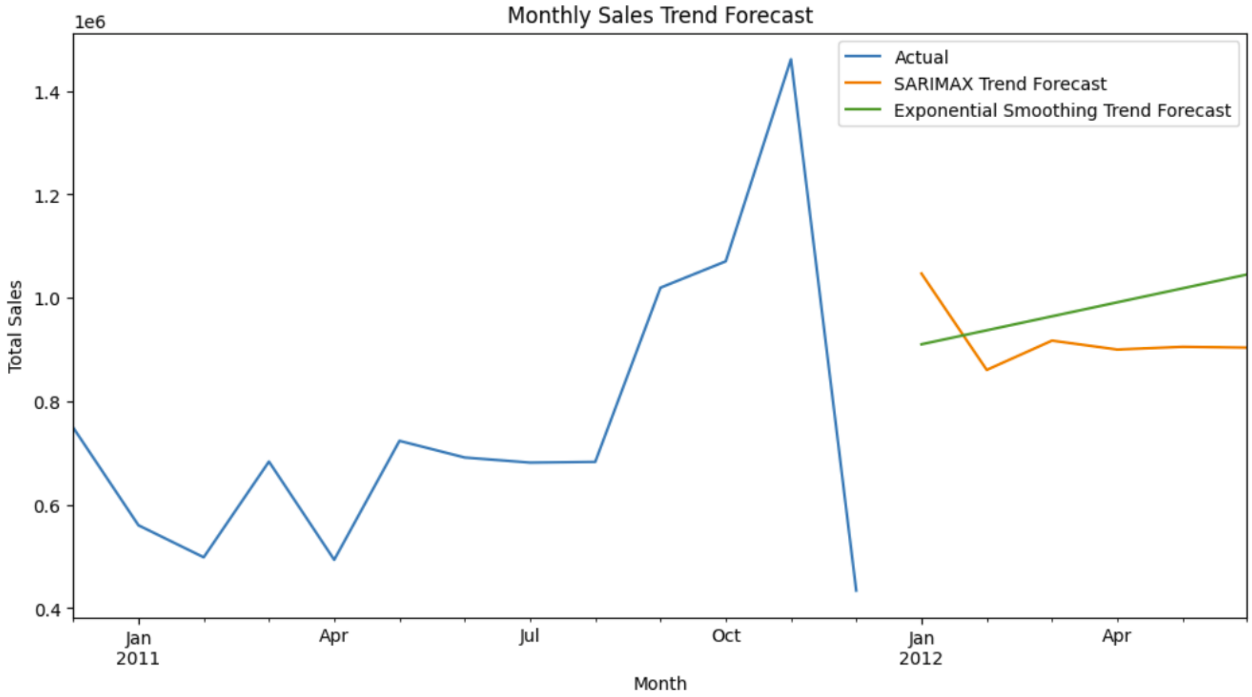




- **Comparison with RFM Clustering**: K-means tends to combine casual and frequent shoppers, highlighting their similarities. It distinctively separates passionate shoppers and 'once a year' customers.

3. **Time Series Analysis Results**
    - **Weekly Decomposition**:
        - Initial 8-week prediction for 2012 indicates higher sales than most of the original data.
    - **Monthly Decomposition**:
        - First 6-month prediction for 2012 shows stable, high sales compared to original data.
    - **Model Comparison**: SARIMAX model exhibits more significant changes in slopes compared to Exponential Smoothing Model. However, none of the models seems to provide a valid prediction.

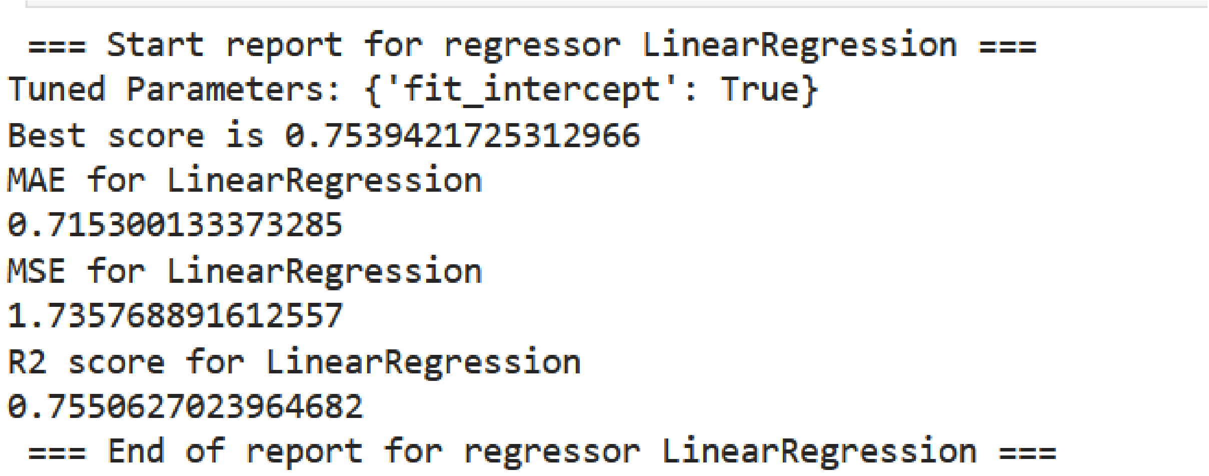




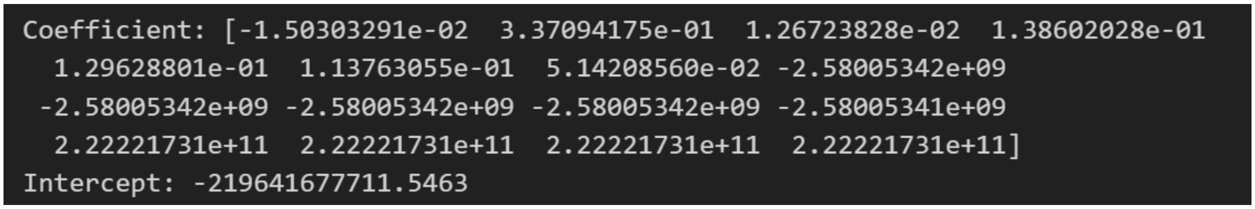
4. **Regression Analysis Results**
   - **Regression Equation**: Y = -2.20e+11 + [complex regression formula with various predictors].
   - Y = -2.20e+11 + (-2.00e-02)*QuantityInv + (3.40e-01)*qr_(0, 2] + (1.00e-02)*qr_(2, 5] + (1.40e-01)*qr_(5, 8] + (1.30e-01)*qr_(8, 11] + (1.10e-01)*qr_(11, 14] + (5.00e-02)*qr_(15, 5000] + (-2.58e+09)*pr_(0, 1]

+ (-2.58e+09)*pr_(1, 2] + (-2.58e+09)*pr_(2, 3] + (-2.58e+09)*pr_(3, 4] + (-2.58e+09)*pr_(4, 20] + (2.22e+11)*dr_(0, 3] + (2.22e+11)*dr_(3, 6] + (2.22e+11)*dr_(6, 9] + (2.22e+11)*dr_(9, 12]

- **Key Predictors**: Range of prices and dates are the most impactful, with their beta values being extremely high.
- **Performance Metrics**:
  - MSE: 1.736
  - MAE: 0.715
  - R²: 0.755
  - Best Score: 0.754
- **Interpretation**: High R² indicates good model performance, but there is room for improvement.

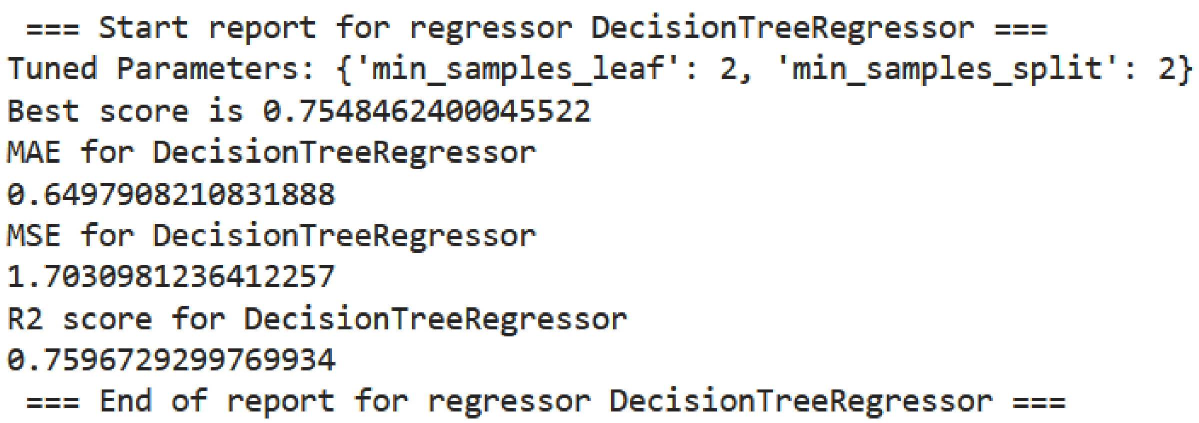


```
 === Start report for regressor LinearRegression ===
Tuned Parameters: {'fit_intercept': True}
Best score is 0.7539421725312966
MAE for LinearRegression
0.715300133373285
MSE for LinearRegression
1.735768891612557
R2 score for LinearRegression
0.7550627023964682
 === End of report for regressor LinearRegression ===
```



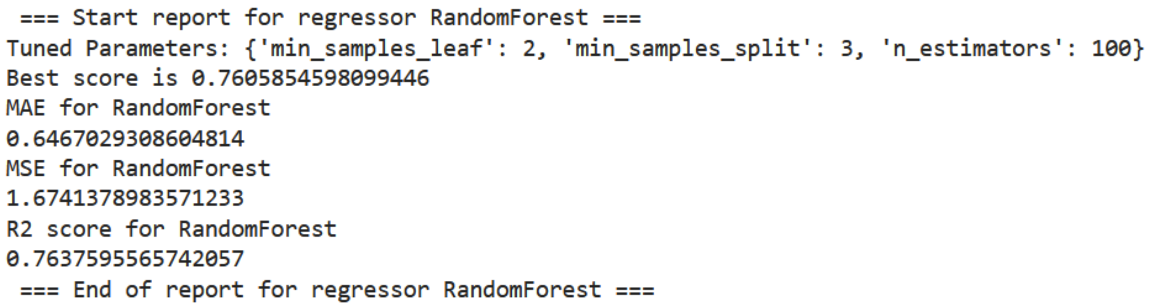
```
Coefficient: [-1.50303291e-02  3.37094175e-01  1.26723828e-02  1.38602028e-01
  1.29628801e-01  1.13763055e-01  5.14208560e-02 -2.58005342e+09
 -2.58005342e+09 -2.58005342e+09 -2.58005342e+09 -2.58005341e+09
  2.22221731e+11  2.22221731e+11  2.22221731e+11  2.22221731e+11]
Intercept: -219641677711.5463
```

5. **Decision Tree Analysis Results**
   - **Best Parameters**: {'min_samples_leaf': 2, 'min_samples_split': 2}
   - **Performance Metrics**:
     - MSE: 1.703
     - MAE: 0.650
     - R²: 0.760
     - Best Score: 0.755
   - **Interpretation**: Slight improvement over the linear regression model in terms of error reduction and R².

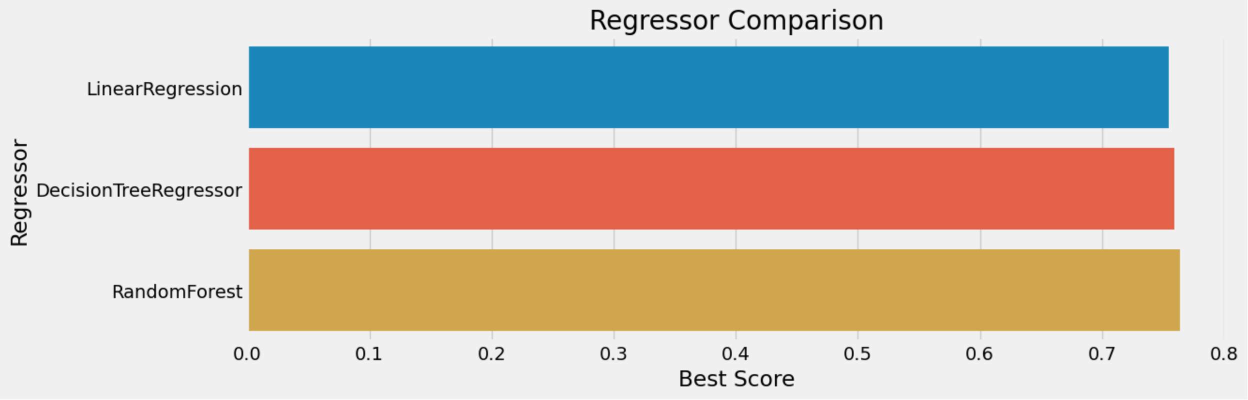


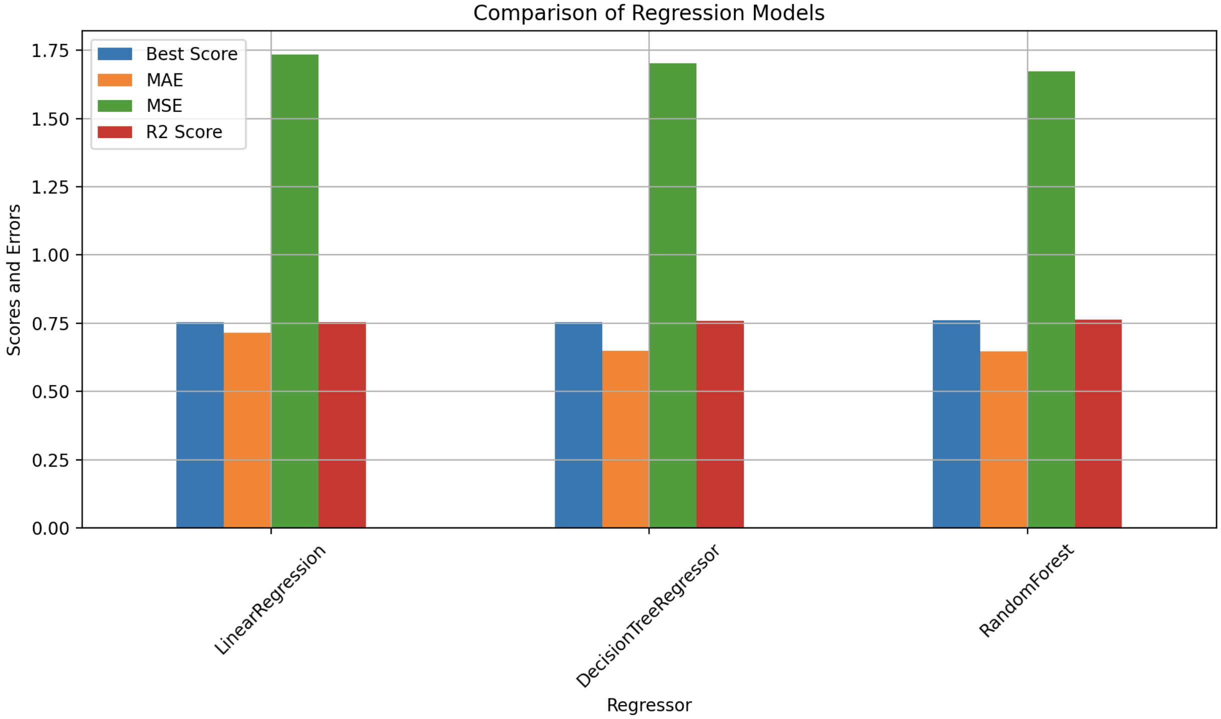
```
 === Start report for regressor DecisionTreeRegressor ===
Tuned Parameters: {'min_samples_leaf': 2, 'min_samples_split': 2}
Best score is 0.7548462400045522
MAE for DecisionTreeRegressor
0.6497908210831888
MSE for DecisionTreeRegressor
1.7030981236412257
R2 score for DecisionTreeRegressor
0.7596729299769934
 === End of report for regressor DecisionTreeRegressor ===
```

6. **Random Forest Analysis Results**
    - **Best Parameters**: {'min_samples_leaf': 2, 'min_samples_split': 3, 'n_estimators': 100}
    - **Performance Metrics**:
        - MSE: 1.674
        - MAE: 0.647
        - R²: 0.764
        - Best Score: 0.761
    - **Interpretation**: The random forest model shows the best performance with the lowest error terms and highest R², indicating improved predictive accuracy.



```
 === Start report for regressor RandomForest ===
Tuned Parameters: {'min_samples_leaf': 2, 'min_samples_split': 3, 'n_estimators': 100}
Best score is 0.7605854598099446
MAE for RandomForest
0.6467029308604814
MSE for RandomForest
1.6741378983571233
R2 score for RandomForest
0.7637595565742057
 === End of report for regressor RandomForest ===
```





Further discussion

In our supervised learning methods, the random forest model demonstrated the best prediction accuracy and the lowest error rate compared to the regression and decision tree methods. However, the differences among these three methods are quite small, particularly in terms of their R² values. When we take into account factors such as the complexity of the model and the ease of interpreting results, random forest might not be the optimal choice.

Comparing k-means and RFM clustering, both methods provided similar results. The classic RFM clustering identified four distinct groups of shopping habits, while k-means effectively merged two middle groups (frequent and casual shoppers) into one.

This indicates that both models are consistent and validate each other, highlighting that the data exhibits clear signals from shoppers who buy frequently and those who rarely make purchases.

A major challenge we encountered in time series prediction was the dataset's duration, which spanned only one year. Typically, our algorithms require multiple years of data to train the model effectively across several cycles. With our limited data, we were unable to establish and validate our predictions robustly. Obtaining more comprehensive data over a longer period is crucial for enhancing the accuracy of our sales amount predictions.

If more data were available, particularly more numerical variables, we could explore more complex models such as polynomial regression, variable selection through stepwise regression, and regularization techniques. These advanced methods could potentially improve accuracy, but there's also a risk of overfitting that must be carefully managed.

With additional time, we would like to delve deeper into both unsupervised and supervised machine learning algorithms to further explore the relationships within each column of our dataset. We are particularly interested in conducting more dimension reduction analysis, such as PCA, to reinforce our findings regarding the most distinct groups in our RFM analysis: the passionate shoppers and those who shop just once a year.