Online Retail Analysis

Arihant Tripathi, Zihe(Peter) Zhang, Alan Wu

Description of Dataset

- Part of the UCI Repo of Datasets: Online Retail - UCI Machine Learning Repository
- Time Frame: December 2010 to December 2011.
- Key Data Fields: Invoice Number, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, Country.
- Data Volume: Approximately 540,000 records.

	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

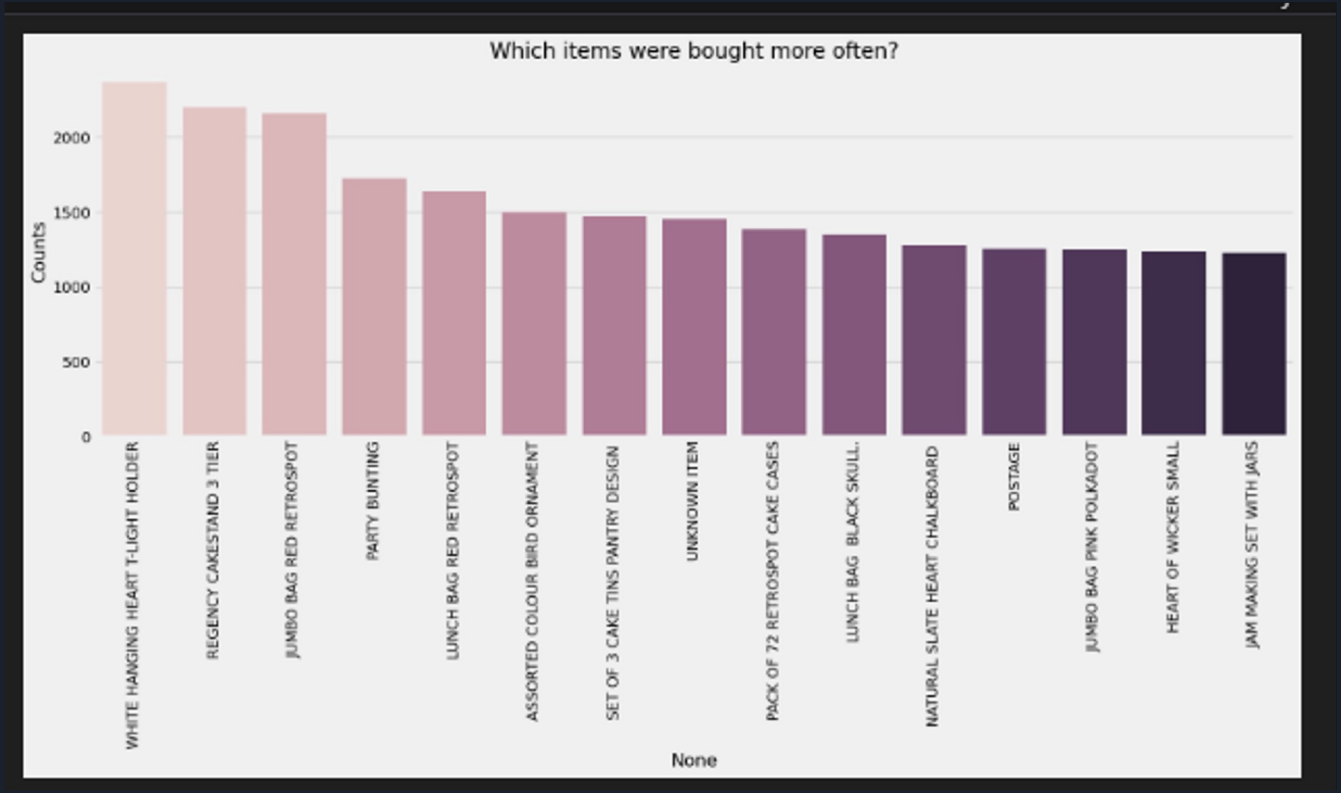
EDA Part 1

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

```
7]  df.isnull().sum()
```

```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
UnitPrice          0
CustomerID    135080
Country            0
dtype: int64
```

We decided to drop customerID.

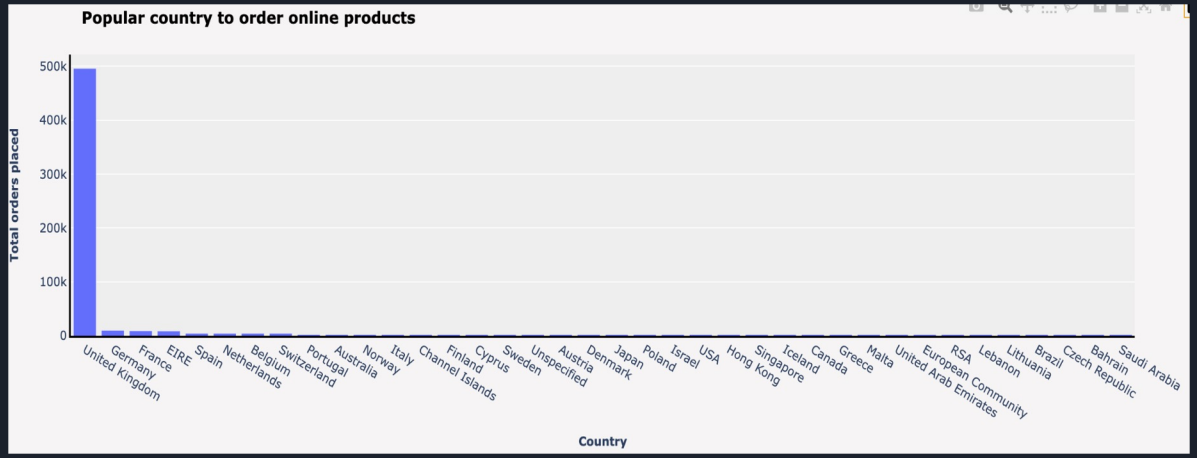


Most purchased item is “WHITE HANGING HEART T SHIRT HOLDER”

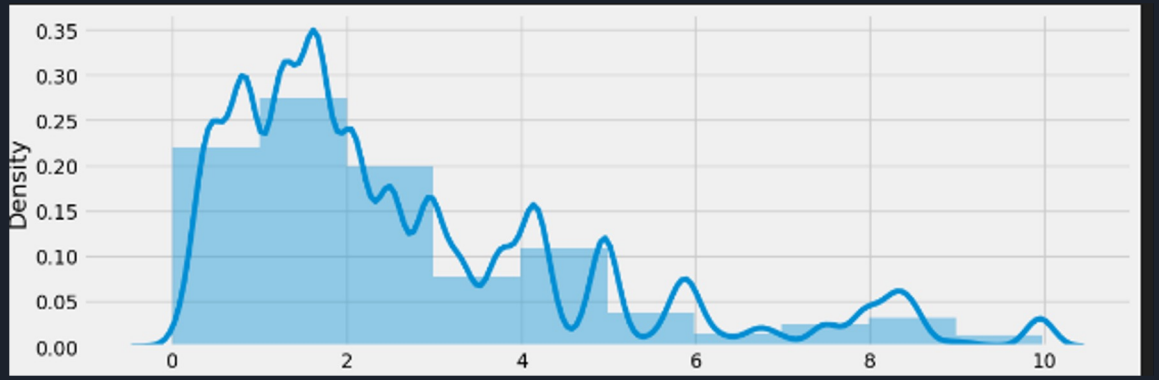
Least purchased item is “DUSTY CHRISTMAS TREE 30CM”


30cm
11.8in
22cm
8.6in

EDA Part 2

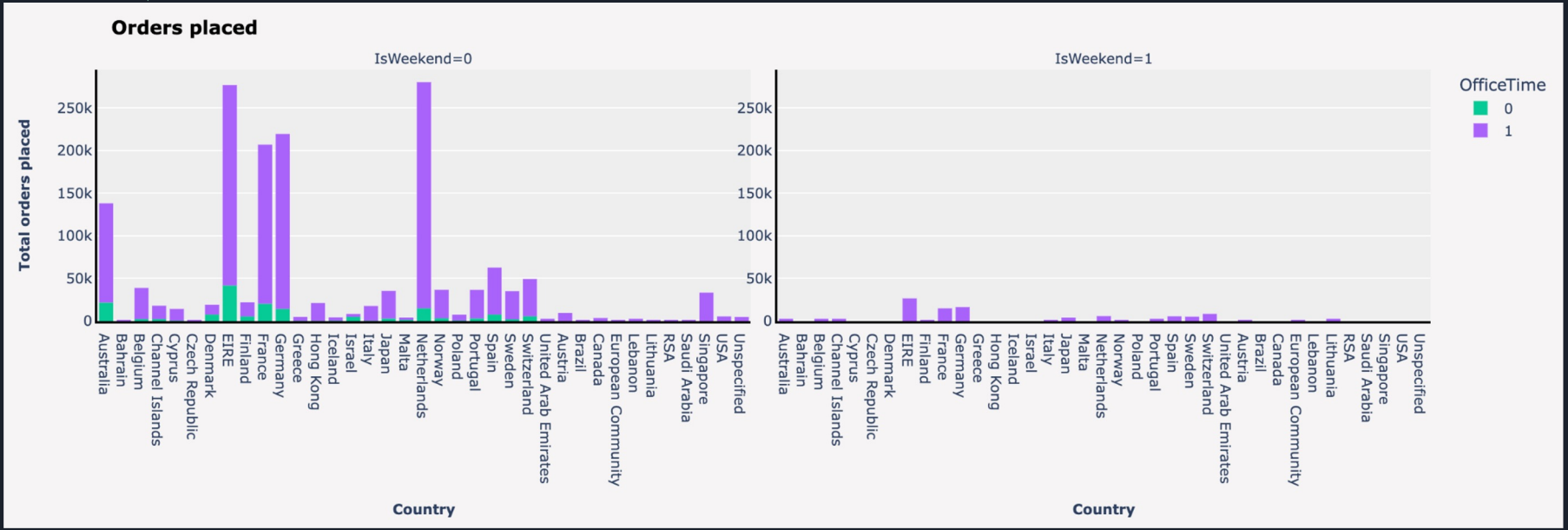


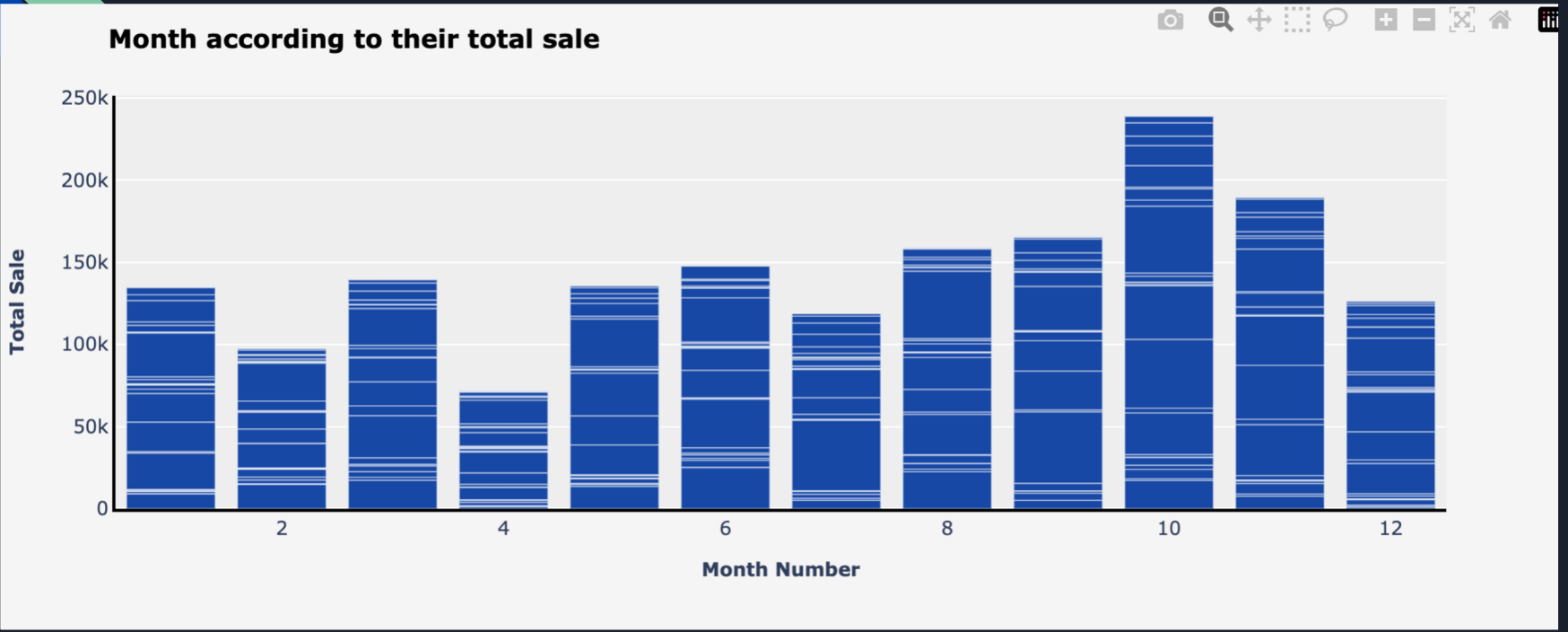
UK percentage of total data 91.43195628786383



Unit Price on the X-axis

Bonus EDA


Orders placed
IsWeekend=0
IsWeekend=1
Total orders placed
0
50k
100k
150k
200k
250k
Australia
Bahrain
Belgium
Channel Islands
Cyprus
Czech Republic
Denmark
EIRE
Finland
France
Germany
Greece
Hong Kong
Iceland
Israel
Italy
Japan
Malta
Netherlands
Norway
Poland
Portugal
Spain
Sweden
Switzerland
United Arab Emirates
Austria
Brazil
Canada
European Community
Lebanon
Lithuania
RSA
Saudi Arabia
Singapore
USA
Unspecified
Country
OfficeTime
0
1


Month according to their total sale
250k
200k
150k
100k
50k
0
Total Sale
2
4
6
8
10
12
Month Number

Unsupervised Approach - K Means VS RMF

Understanding RFM Metrics:

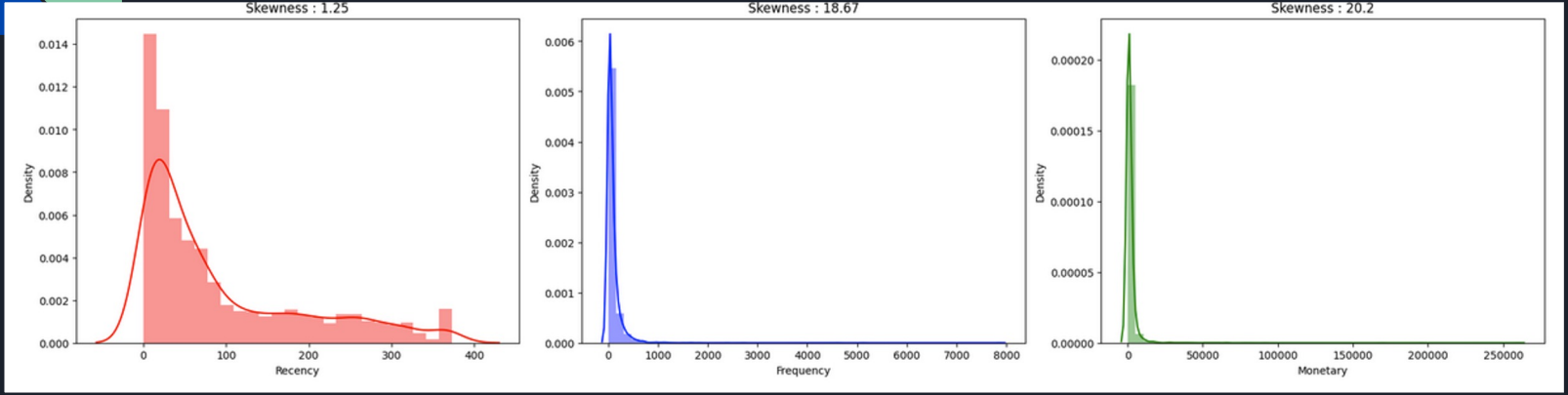
- Recency: Time since the last purchase.
- Frequency: Number of transactions in a given period.
- Monetary Value: Total amount spent in a given period.

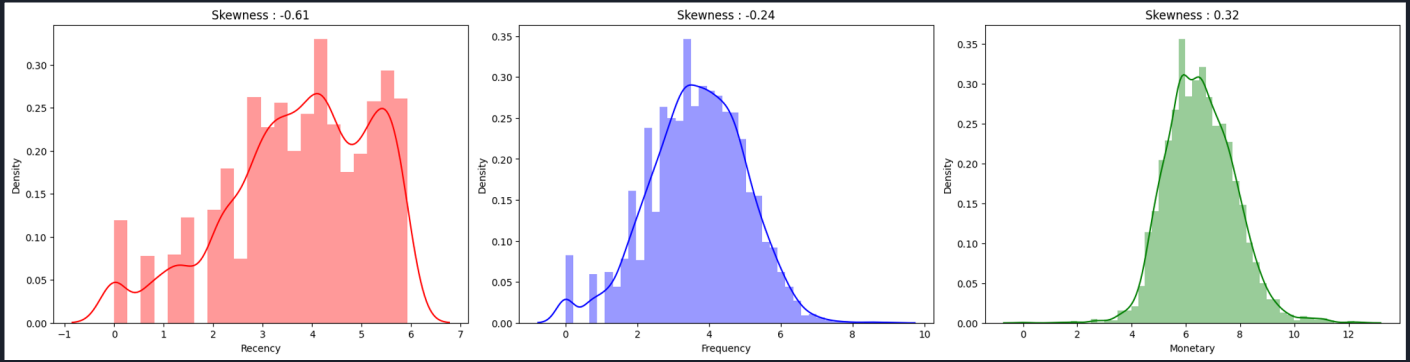
Use methods like the Elbow Method determine the optimal number of clusters.

Apply k-means clustering algorithm to the normalized RFM data.

Assign each customer to a cluster based on their RFM characteristics.

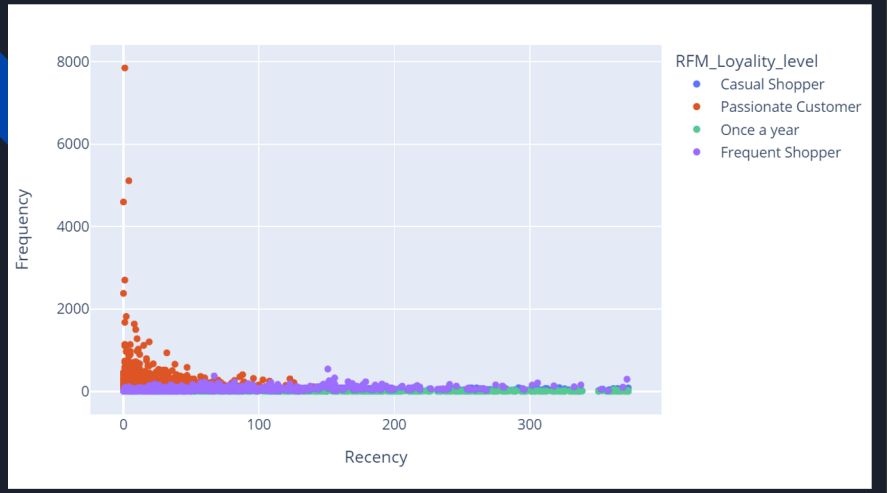
Log-transformation

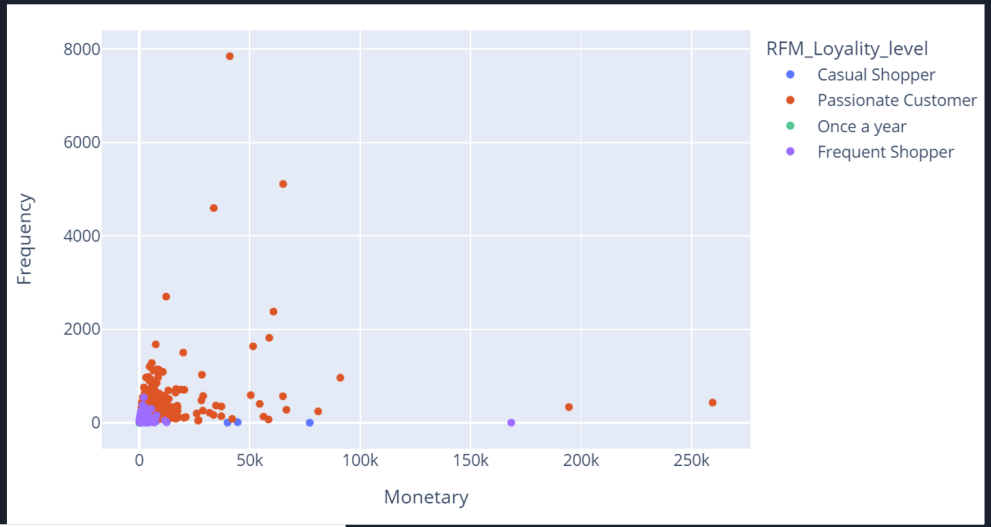


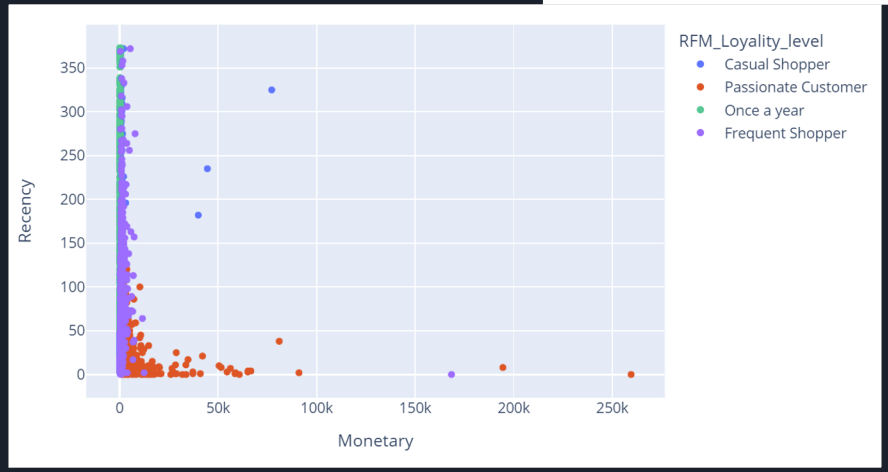


Unsupervised Approach - K Means VS RMF

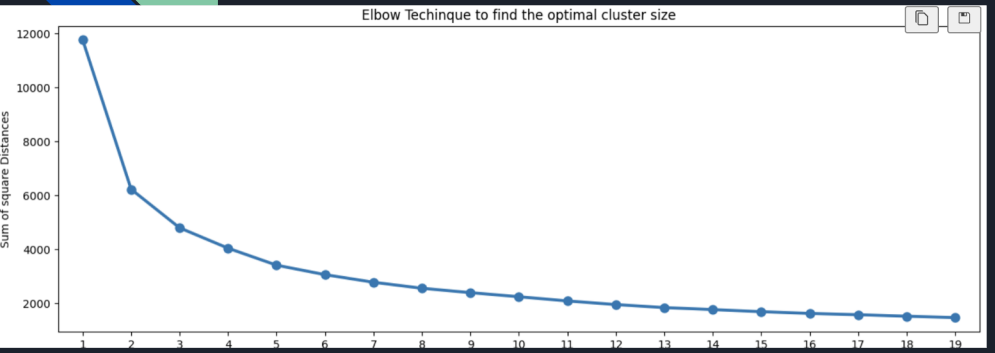
	CustomerID	Recency	Frequency	Monetary	R_score	F_score	M_score	RFM_Group	RFM_Score
0	12346.0	325	1	77183.60	4	4	1	441	9
1	12747.0	2	103	4196.01	1	1	1	111	3
2	12748.0	0	4596	33719.73	1	1	1	111	3
3	12749.0	3	199	4090.88	1	1	1	111	3

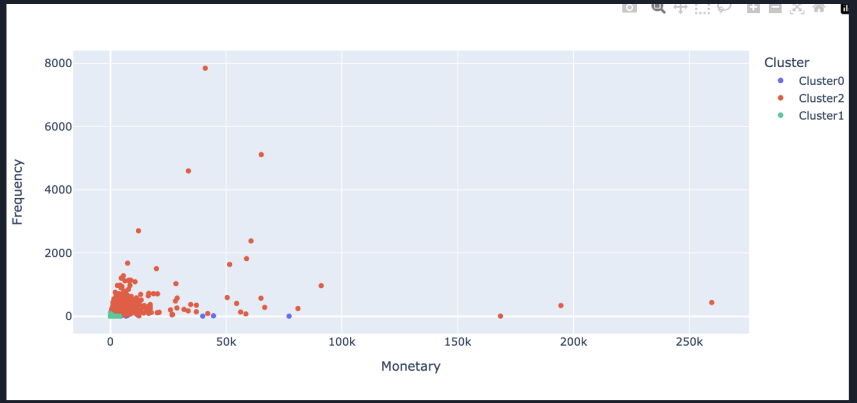

RFM_Loyality_level
Casual Shopper
Passionate Customer
Once a year
Frequent Shopper
Frequency
Recency
0
2000
4000
6000
8000
0
100
200
300

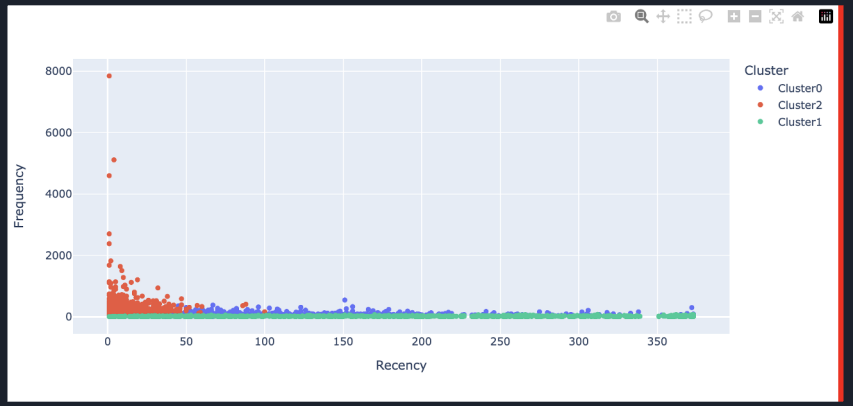

RFM_Loyality_level
Casual Shopper
Passionate Customer
Once a year
Frequent Shopper
Frequency
Monetary
0
2000
4000
6000
8000
0
50k
100k
150k
200k
250k

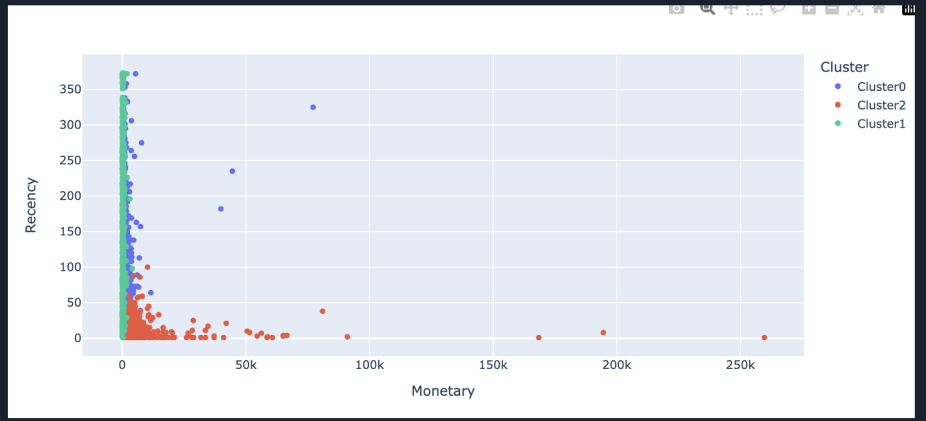

RFM_Loyality_level
Casual Shopper
Passionate Customer
Once a year
Frequent Shopper
Recency
Monetary
0
50
100
150
200
250
300
350
0
50k
100k
150k
200k
250k

Unsupervised Approach - K Means









Supervised Approach 1 - Linear Regression

- Parameter Selection: GridSearchCV can be used to optimize regularization parameters like alpha in Ridge and Lasso regression.
- Grid Setup: Define a parameter grid for alpha values (e.g., [0.001, 0.01, 0.1, 1, 10, 100]).
- Scoring Metric: RMSE (Root Mean Squared Error)
- Cross-Validation: 5-fold cross-validation
- Fitting Model: Fit the GridSearchCV with training data and retrieve the best parameters.

Supervised Approach 1 - Linear Regression Results

- Tuned Parameter: fit_intercept set to True.
- Best CV Score: 0.7539
- Mean Absolute Error (MAE): 0.7153
- Mean Squared Error (MSE): 1.7358
- R^2 Score: 0.7551, suggesting that about 75.51% of the variance in the dependent variable is explained by the model.

$$Y = -2.20 \times 10^{11} + (-2.00 \times 10^{-2}) \times \text{QuantityInv} + (3.40 \times 10^{-1}) \times \text{qr}_{(0,2]} + (1.00 \times 10^{-2}) \times \text{qr}_{(2,5]} + (1.40 \times 10^{-1}) \times \text{qr}_{(5,8]} + (1.30 \times 10^{-1}) \times \text{qr}_{(8,11]} + (1.10 \times 10^{-1}) \times \text{qr}_{(11,14]} + (5.00 \times 10^{-2}) \times \text{qr}_{(15,5000]} + (-2.58 \times 10^{9}) \times \text{pr}_{(0,1]} + (-2.58 \times 10^{9}) \times \text{pr}_{(1,2]} + (-2.58 \times 10^{9}) \times \text{pr}_{(2,3]} + (-2.58 \times 10^{9}) \times \text{pr}_{(3,4]} + (-2.58 \times 10^{9}) \times \text{pr}_{(4,20]} + (2.22 \times 10^{11}) \times \text{dr}_{(0,3]} + (2.22 \times 10^{11}) \times \text{dr}_{(3,6]} + (2.22 \times 10^{11}) \times \text{dr}_{(6,9]} + (2.22 \times 10^{11}) \times \text{dr}_{(9,12]}$$

Supervised Approach 2 - Decision Tree

- Parameter Selection: max_depth, min_samples_split, and min_samples_leaf.
- Grid Setup: Create a parameter grid with a range of values for the selected parameters (e.g., max_depth ranging from 1 to 10).
- Scoring Metric: Use R^2
- Cross-Validation: 5-fold CV
- Fitting Model: Run GridSearchCV with the Decision Tree model and the dataset, then identify the best parameter combination.

Supervised Approach 2 - Decision Tree Results

- Tuned Parameters: min_samples_leaf at 2, and min_samples_split at 2.
- Best CV Score: 0.7549
- Mean Absolute Error (MAE): 0.6499
- Mean Squared Error (MSE): 1.7032
- R^2 Score: 0.7597

Supervised Approach 3 - Random Forest

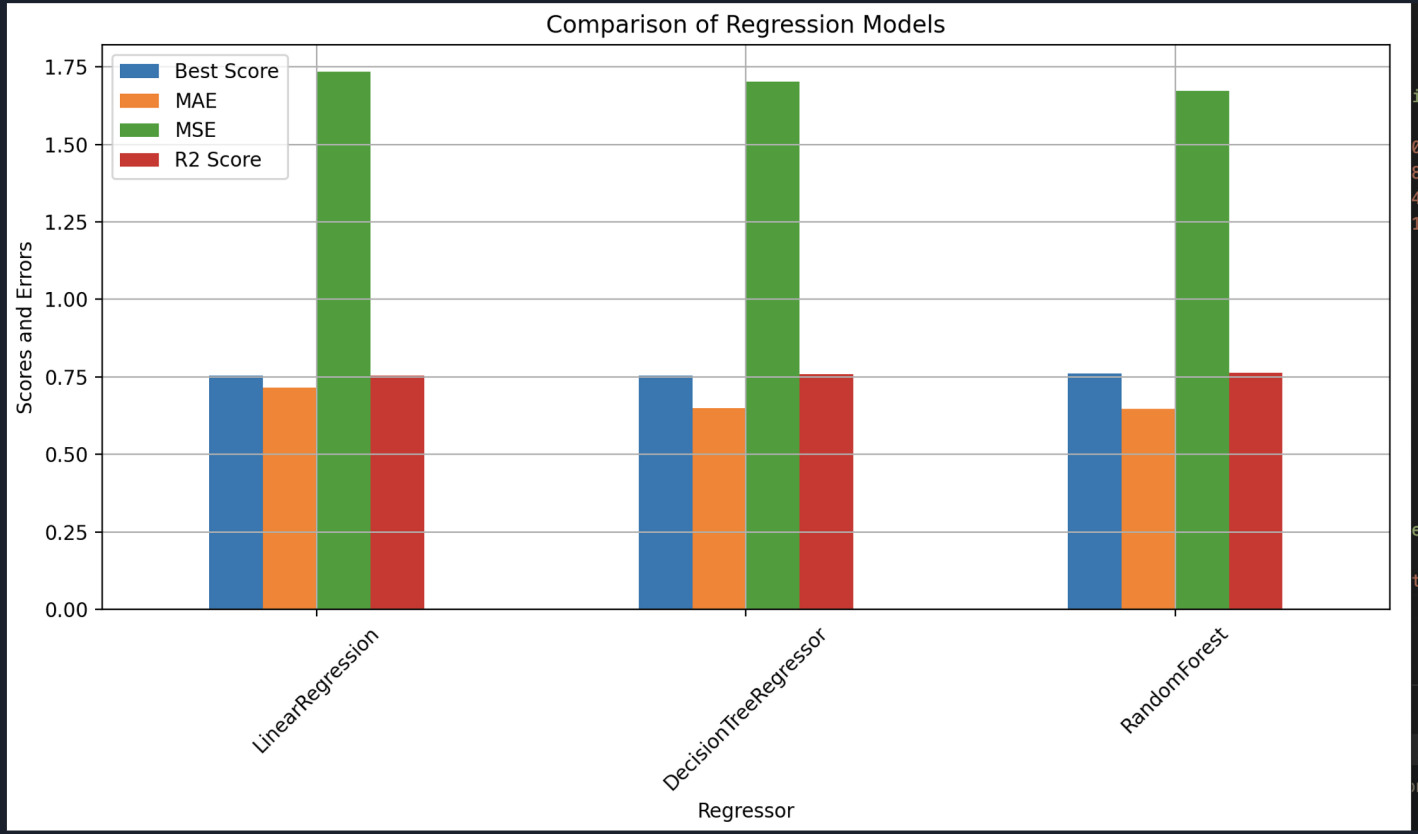
- Parameter Selection: Key parameters include n_estimators, max_features, max_depth, and min_samples_split.
- Grid Setup: Define a comprehensive grid for these parameters. For example, n_estimators could range from 10 to 100.
- Scoring Metric: F1-score
- Cross-Validation: Apply 5-fold CV
- Fitting Model: Execute GridSearchCV with the Random Forest model and determine the optimal parameter set.

Supervised Approach 3 - Random Forest Results

- Tuned Parameters: min_samples_leaf at 2, min_samples_split at 3, and n_estimators at 100.
- Best CV Score: 0.7605, the highest among the three
- Metrics:
- Mean Absolute Error (MAE): 0.6466
- Mean Squared Error (MSE): 1.6741
- R^2 Score: 0.7638

Grid Search Cross Validation Visualized

Regressor	Tuned Parameters	Best Score	MAE	MSE	R2 Score
LinearRegression	{'fit_intercept': True}	0.75	0.72	1.74	0.76
DecisionTreeRegressor	{'min_samples_leaf': 2, 'min_samples_split': 2}	0.75	0.65	1.7	0.76
RandomForest	{'min_samples_leaf': 2, 'min_samples_split': 3, 'n_estimators': 100}	0.76	0.65	1.67	0.76


Comparison of Regression Models
Best Score
MAE
MSE
R2 Score
Scores and Errors
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
LinearRegression
DecisionTreeRegressor
RandomForest
Regressor

Supervised Learning Conclusion

- Performance Ranking: Random Forest > Decision Tree > Linear Regression based on R^2 scores and error metrics.
- Error Metrics: Random Forest consistently showed lower error rates (MAE and MSE) across all models.
- Predictive Accuracy: Random Forest also had the highest R^2 score, indicating it's the most effective model for our dataset among the three.
- Tuning Impact: Hyperparameter tuning had a noticeable impact on the performance of Decision Tree and Random Forest models, while Linear Regression showed limited tuning scope.

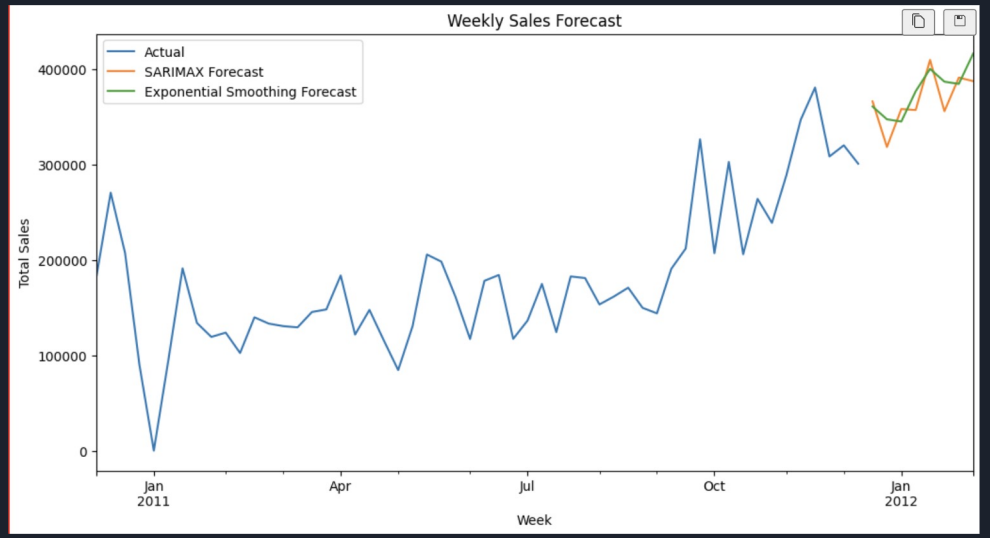
Experimental Methods - Time Series Prediction

- SARIMAX(Seasonal Autoregressive integrated moving average) and Exponential Smoothing Models
- Issues begin to come up as the minimum requirement to run time series prediction algorithm is 2 cycles of data
- Cycle is defined as 1 year of data
- Our case only had 1 cycle of data
- Outcomes were strange, no real trends predicted with the data

Trend	Seasonal		
	N	A	M
N	$\hat{y}_{t+h\|t} = \ell_t$ $\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = \ell_t s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t/\ell_{t-1}) + (1-\gamma)s_{t-m}$
A	$\hat{y}_{t+h\|t} = \ell_t + hb_t$ $\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = (\ell_t + hb_t)s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t/(\ell_{t-1} + b_{t-1})) + (1-\gamma)s_{t-m}$
A_d	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t$ $\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = (\ell_t + \phi_h b_t)s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t/(\ell_{t-1} + \phi b_{t-1})) + (1-\gamma)s_{t-m}$

$$y_t = c + \varphi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \Phi_1(y_{t-5} + \varphi_1 y_{t-6}) +$$
$$\Phi_2(y_{t-10} + \varphi_1 y_{t-11}) + \Theta_1(\varepsilon_{t-5} + \theta_1 \varepsilon_{t-6} + \theta_2 \varepsilon_{t-7}) + \varepsilon_t$$

Experimental Methods - Time Series Prediction



Data Science Project Stack

- Pandas
- Numpy
- Scikit-Learn
- Seaborn
- Matplotlib
- Statmodels (SARIMAX)