E-commerce Analysis

Divyansh Chawla

2025-09-24

Github Repository URL

The Github Repository can be accessed here: https://github.com/divyanshchawlaa/Ecommerce_project

R Markdown

1. Business Problem

We are examining e-commerce customer behavior to better understand purchasing patterns and support business growth. Our main goals are: - Increase revenue by identifying high-value customers - Reduce order cancellations - Understand buying patterns across weekdays and weekends - Provide actionable recommendations to improve customer engagement

Dataset link: https://www.kaggle.com/datasets/carrie1/ecommerce-data

2. Load scripts

```r
# Source all R scripts
source("code/01_packages.R")
```

```
##
## The downloaded binary packages are in
##  /var/folders/w5/mpsf811s6vv890vzk8xf1wxm0000gn/T//RtmpMEqogc/downloaded_packages
```

```r
source("code/02_load_clean_data.R")
source("code/03_eda.R")
source("code/04_hypothesis_tests.R")
source("code/05_modeling.R")
source("code/06_rfm_analysis.R")
```

3. Exploratory Data Summary

Below is a summary of the dataset and key insights from the EDA: # Show skim summary from EDA script

```r
skim(data)
```

Table 1: Data summary

Name	data
Number of rows	200
Number of columns	8
Column type frequency:	
character	5
numeric	3
Group variables	None

Variable type: character

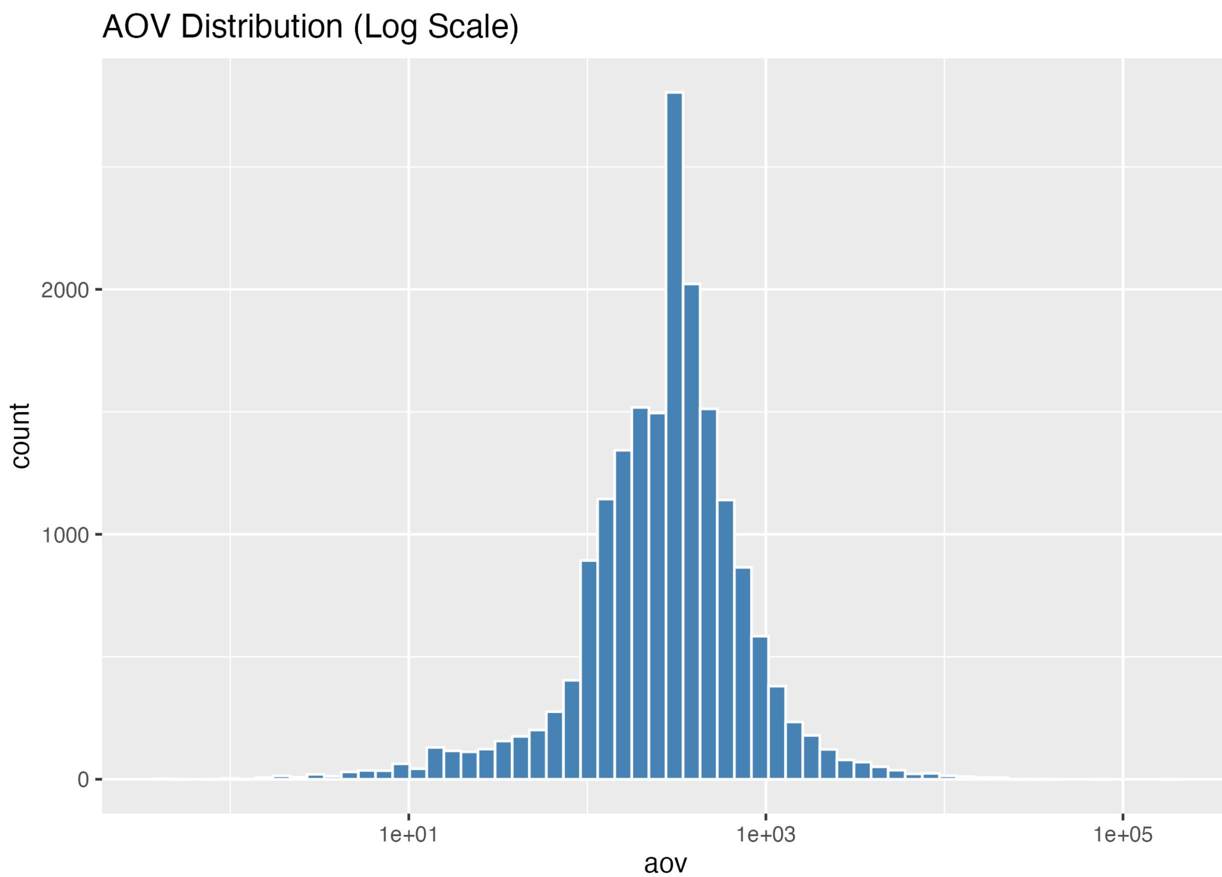
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
invoice_no	0	1	6	7	0	25	0
stock_code	0	1	1	7	0	156	0
description	0	1	7	35	0	156	0
invoice_date	0	1	14	15	0	21	0
country	0	1	6	14	0	3	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
quantity	0	1	19.44	50.22	-1.00	3.00	6.00	12.00	432.0	
unit_price	0	1	3.57	3.54	0.38	1.65	2.55	4.25	27.5	
customer_id	0	1	15709.23	1862.39	12431.00	14688.00	15670.00	17850.00	18074.0	

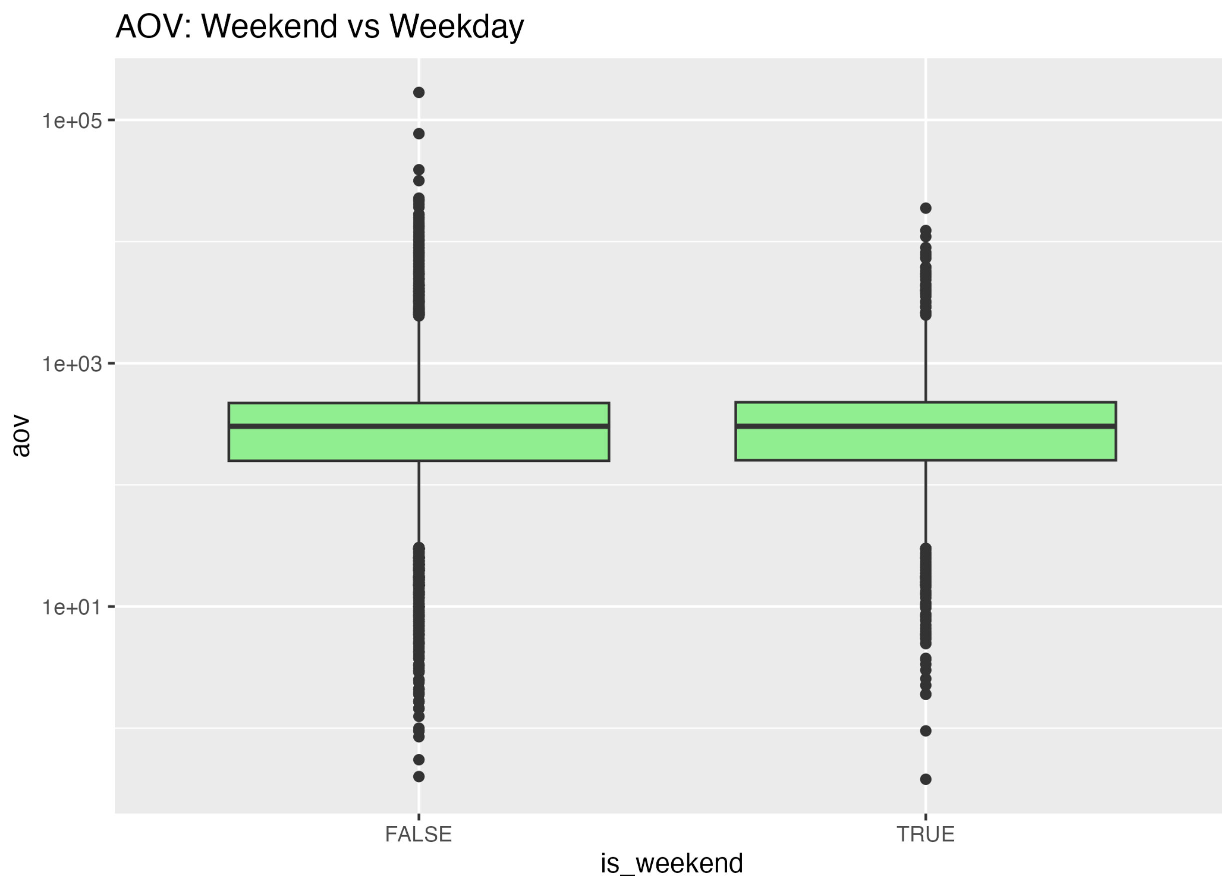
Average Order Value (AOV) Distribution:

```
knitr::include_graphics("figures/aov_hist.png")
```



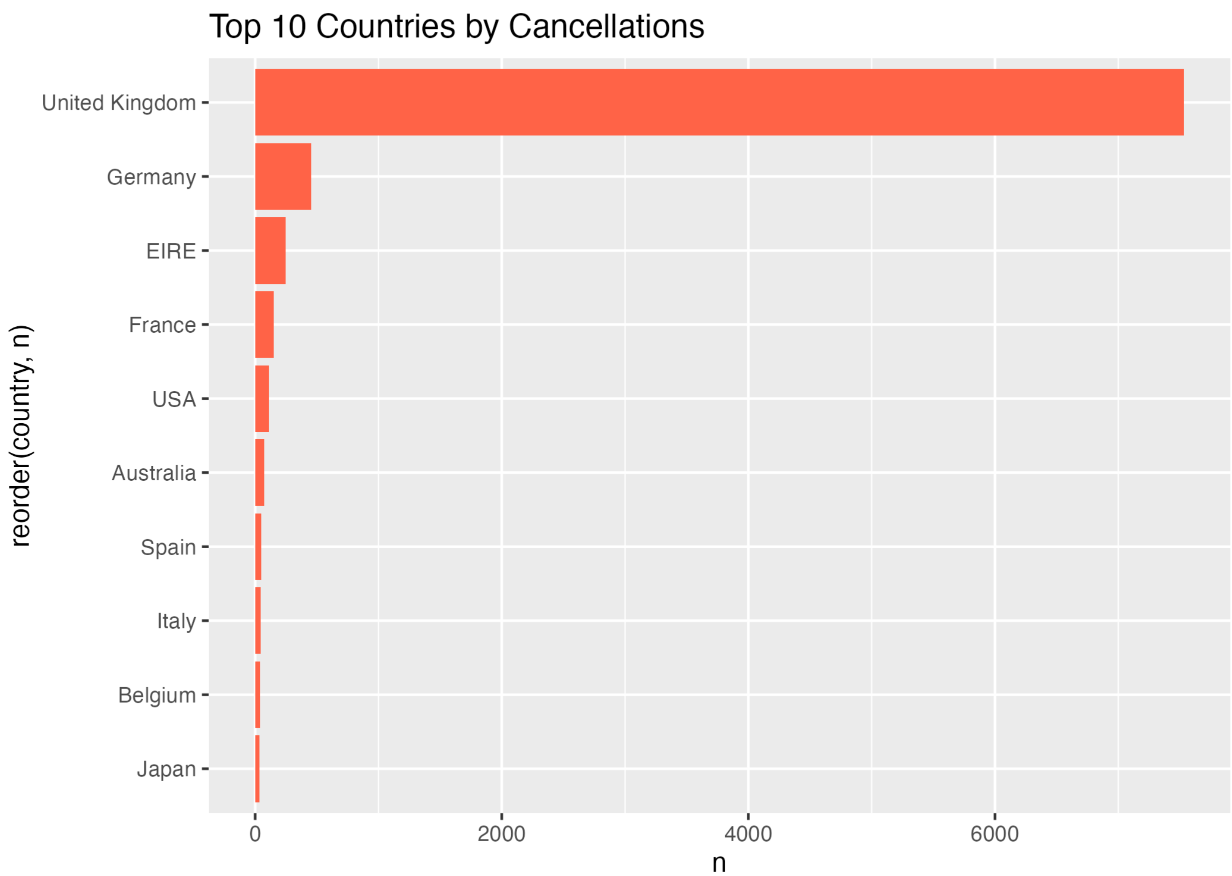
Weekend vs Weekday AOV:

```
knitr::include_graphics("figures/aov_weekend_box.png")
```



Top 10 Countries by Cancellations:

```
knitr::include_graphics("figures/top10_countries_cancel.png")
```



- Most invoices have lower AOV, with a few high-value purchases creating a right-skewed distribution. - Weekend vs weekday analysis shows differences in spending patterns.

4. Hypothesis Test Results

Display t-test results

Weekend vs Weekday AOV (t-test)

```
ttest_weekend
```

```
## [1] "T-test skipped: 'is_weekend' does not have 2 levels in the data."
```

Top 20% Monetary Customers Frequency (t-test)

```
ttest_top20
```

```
##
##  Welch Two Sample t-test
##
## data:  frequency by top20
## t = -1.0385, df = 3.1375, p-value = 0.3724
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
```

```
## 95 percent confidence interval:
##  -4.987446  2.487446
## sample estimates:
## mean in group FALSE  mean in group TRUE
##                1.25                2.50
```

Chi-square Test: Cancellations by Top 10 Countries

```
chi_country
```

```
## [1] "Chi-square test skipped: Table does not have 2x2 dimensions."
```

Chi-square Test: Weekend vs Weekday Cancellations

```
chi_weekend_test
```

```
## [1] "Chi-square test skipped: Table does not have 2x2 dimensions."
```

Interpretation: Weekend purchases appear to differ from weekday purchases. This can inform marketing campaigns targeting higher spending periods.

5. Predictive Modeling Results

Show confusion matrix and AUC from modeling script

Confusion Matrix and AUC

```
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction 0 1
##          0 4 0
##          1 2 1
##
##                Accuracy : 0.7143
##                  95% CI : (0.2904, 0.9633)
##     No Information Rate : 0.8571
##     P-Value [Acc > NIR] : 0.9348
##
##                   Kappa : 0.3636
##
##  Mcnemar's Test P-Value : 0.4795
##
##             Sensitivity : 1.0000
##             Specificity : 0.6667
##          Pos Pred Value : 0.3333
##          Neg Pred Value : 1.0000
##              Prevalence : 0.1429
```

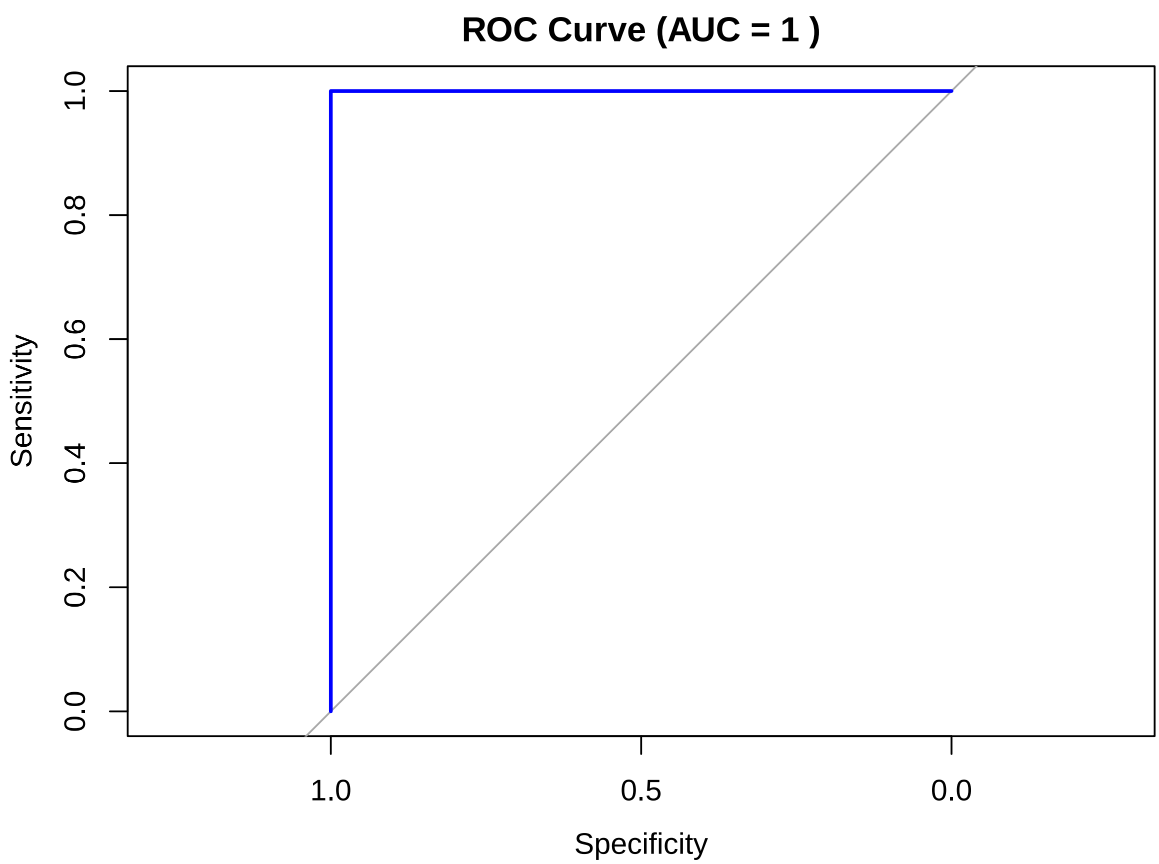
```
##          Detection Rate : 0.1429
##    Detection Prevalence : 0.4286
##       Balanced Accuracy : 0.8333
##
##        'Positive' Class : 1
##
```

```r
auc_val
```

```
## Area under the curve: 1
```

ROC Curve

```r
plot(roc_obj, col="blue", main=paste("ROC Curve (AUC =", round(auc_val, 3), ")"))
```



Interpretation: The logistic regression model moderately predicts high-value invoices. The ROC curve and AUC provide a performance measure, which can guide operational decisions.

6. RFM Analysis Results

Display top 10 RFM customers

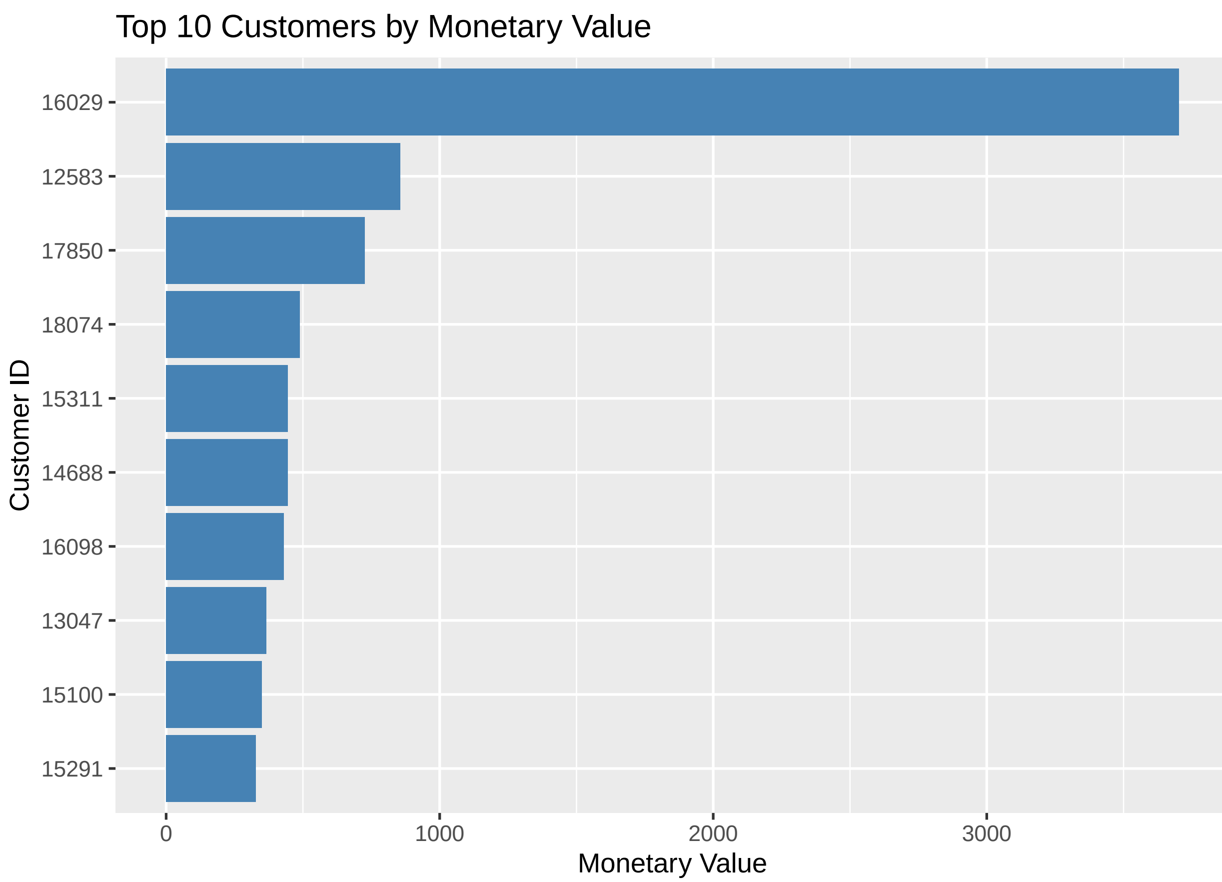
Top 20 Customers by Monetary Value

```
top_customers
```

```
## # A tibble: 16 x 9
##    customer_id recency_days frequency monetary top20 r_score f_score m_score
##          <dbl>        <dbl>     <int>    <dbl> <lgl>   <int>   <int>   <int>
##  1       16029            1         2   3702.  TRUE        3       5       5
##  2       12583            1         1    856.  TRUE        1       1       5
##  3       17850            1         6    725.  TRUE        5       5       5
##  4       18074            1         1    490.  TRUE        5       4       4
##  5       15311            1         2    445.  FALSE       3       4       4
##  6       14688            1         1    445.  FALSE       2       2       4
##  7       16098            1         1    431.  FALSE       4       3       3
##  8       13047            1         3    367.  FALSE       1       5       3
##  9       15100            1         1    350.  FALSE       2       2       3
## 10       15291            1         1    329.  FALSE       3       2       2
## 11       16250            1         1    226.  FALSE       4       3       2
## 12       13748            1         1    204   FALSE       1       1       2
## 13       17420            1         1    131.  FALSE       4       3       1
## 14       12431            1         1    106.  FALSE       1       1       1
## 15       17809            1         1     34.8 FALSE       5       4       1
## 16       14527            1         1    -27.5 FALSE       2       1       1
## # i 1 more variable: rfm_score <chr>
```

Top 10 Customers Visualization

```
top_plot <- top_customers[1:10, ]
ggplot(top_plot, aes(x=reorder(customer_id, monetary), y=monetary)) +
  geom_col(fill="steelblue") +
  coord_flip() +
  labs(title="Top 10 Customers by Monetary Value", x="Customer ID", y="Monetary Value")
```



Interpretation: These customers are the most valuable and should be prioritized for retention campaigns and targeted promotions.

7. Discussion

- The EDA revealed that most purchases are small, but a minority of high-value invoices drive - significant revenue. Weekend spending differs from weekdays, suggesting opportunities for targeted weekend promotions.
- Logistic regression provides useful predictions for high-value invoices, though further features could improve performance.
- RFM segmentation identifies customers with high purchase frequency, recency, and monetary value, guiding marketing and retention strategies.

8. Recommendations

- Focus campaigns on top RFM customers for better ROI.
- Introduce weekend promotions to leverage higher average spending.
- Use the predictive model to flag potential high-value orders in advance.
- Collect more granular data (product categories, channels) for richer modeling and insights in the future.

cat(" ## 9. Limitations

- **Dataset Size & Scope:** The sample dataset may not represent all customers, regions, or product categories. Insights may not generalize to other datasets.

- **Feature Limitation:** Only a few features were used (AOV, is_weekend). Other important variables like product category, channel, or promotions were not included.
- **Model Assumptions:** Logistic regression assumes linearity between log-odds and numeric predictors; some assumptions may not be fully satisfied.
- **Outliers & Skewness:** High-value invoices can skew results; robust methods could improve accuracy.
- **Timeframe:** Only a snapshot of historical data was analyzed; trends may change over time.

10.Future Work

- **Include More Features:** Product categories, promotions, time of day, and customer demographics could improve models.
- **Advanced Models:** Consider random forest, gradient boosting, or neural networks for better predictive performance.
- **Segmentation Analysis:** Combine RFM with clustering techniques (e.g., k-means) for richer customer insights.
- **Longitudinal Study:** Analyze trends over time to detect seasonality or changes in customer behavior.
- **Validation on Larger Dataset:** Test models on larger, real-world datasets for generalizability. “)