# E-commerce Analysis

Divyansh Chawla

2025-09-24

## 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

---

## 2. Load scripts

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

```
##
## The downloaded binary packages are in
##  /var/folders/w5/mpsf811s6vv890vzk8xf1wxm0000gn/T//RtmpaMnaOi/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 | 541909 |

| | |
|---|---|
| Number of columns | 8 |
| Column type frequency: | |
| character | 5 |
| numeric | 3 |
| Group variables | None |

**Variable type: character**

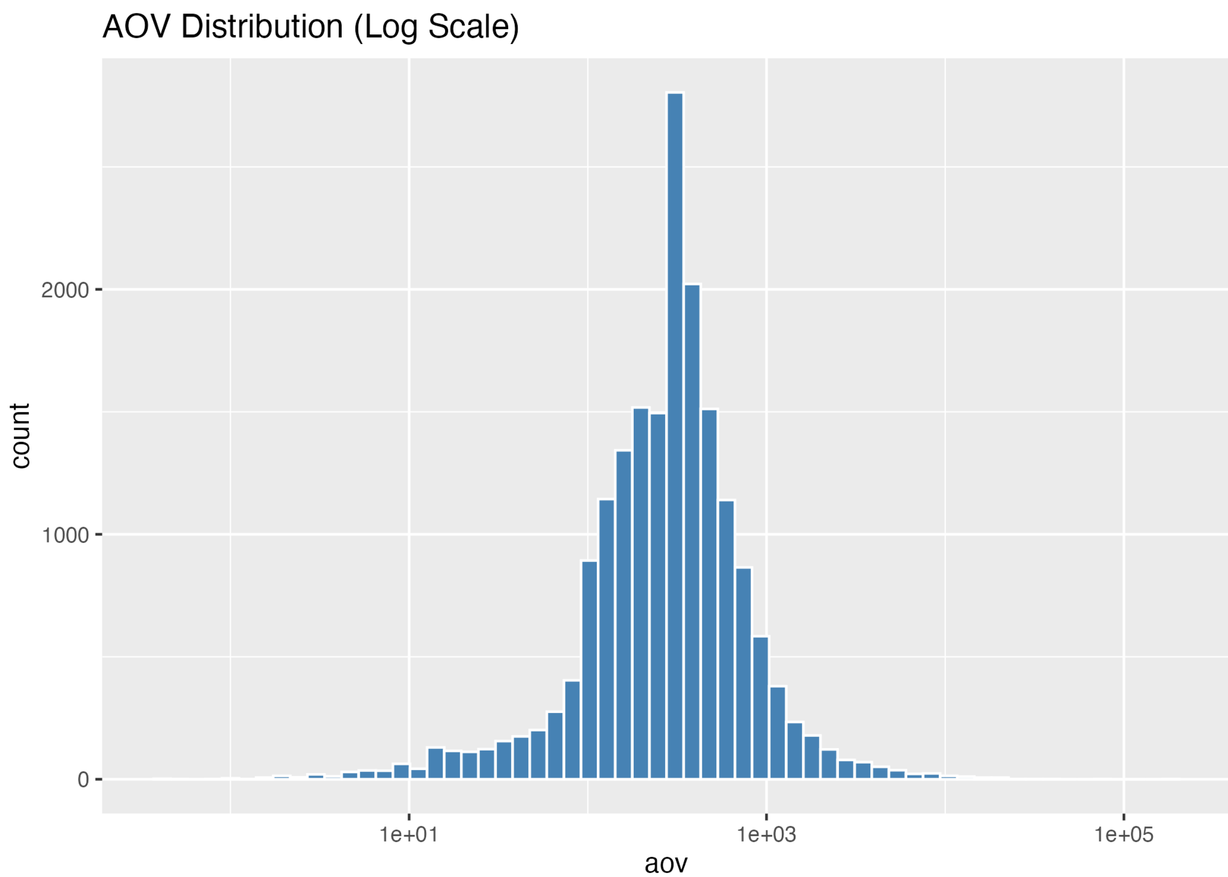
| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| invoice_no | 0 | 1 | 6 | 7 | 0 | 25900 | 0 |
| stock_code | 0 | 1 | 1 | 12 | 0 | 4070 | 0 |
| description | 1454 | 1 | 1 | 35 | 0 | 4211 | 0 |
| invoice_date | 0 | 1 | 13 | 16 | 0 | 23260 | 0 |
| country | 0 | 1 | 3 | 20 | 0 | 38 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| quantity | 0 | 1.00 | 9.55 | 218.08 | -80995.00 | 1.00 | 3.00 | 10.00 | 80995 | |
| unit_price | 0 | 1.00 | 4.61 | 96.76 | -11062.06 | 1.25 | 2.08 | 4.13 | 38970 | |
| customer_id | 135080 | 0.75 | 15287.69 | 1713.60 | 12346.00 | 13953.00 | 15152.00 | 16791.00 | 18287 | |

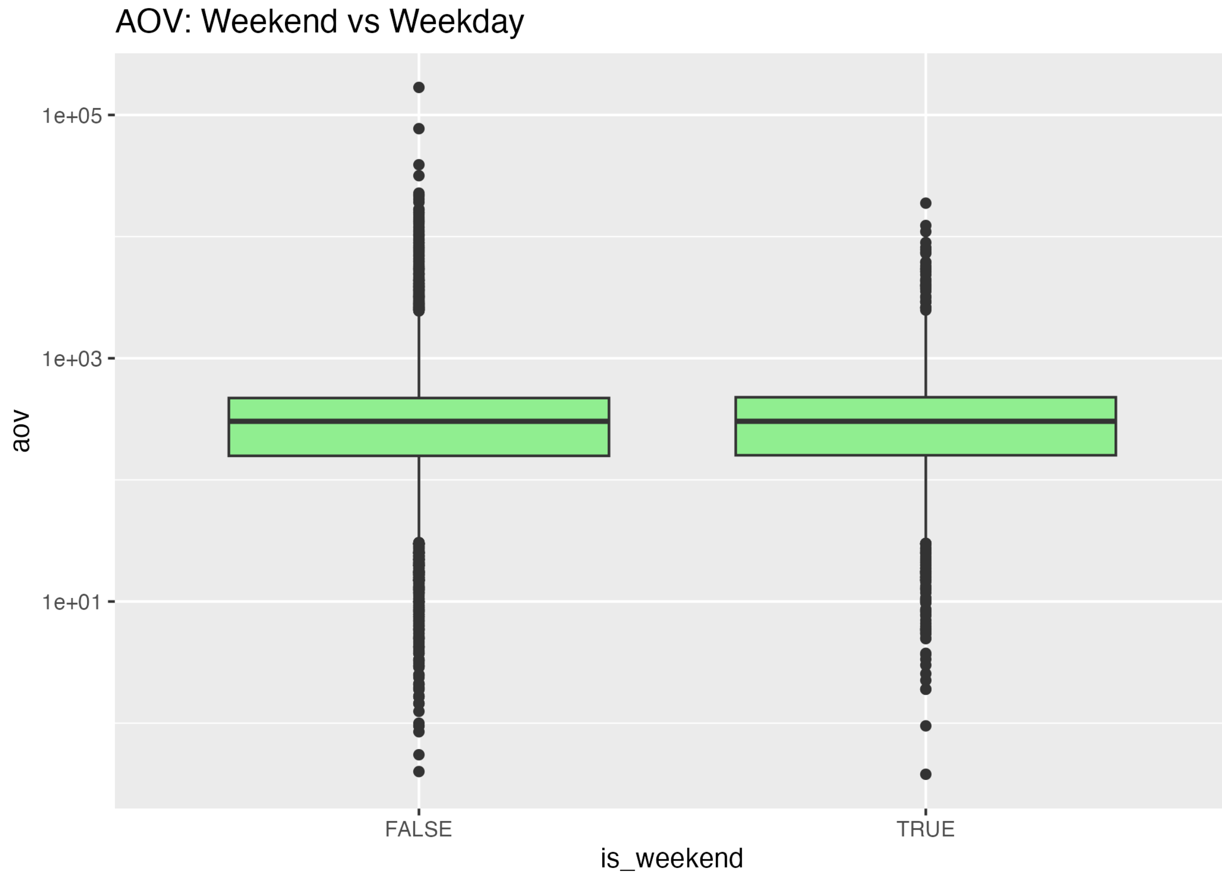
Average Order Value (AOV) Distribution:

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



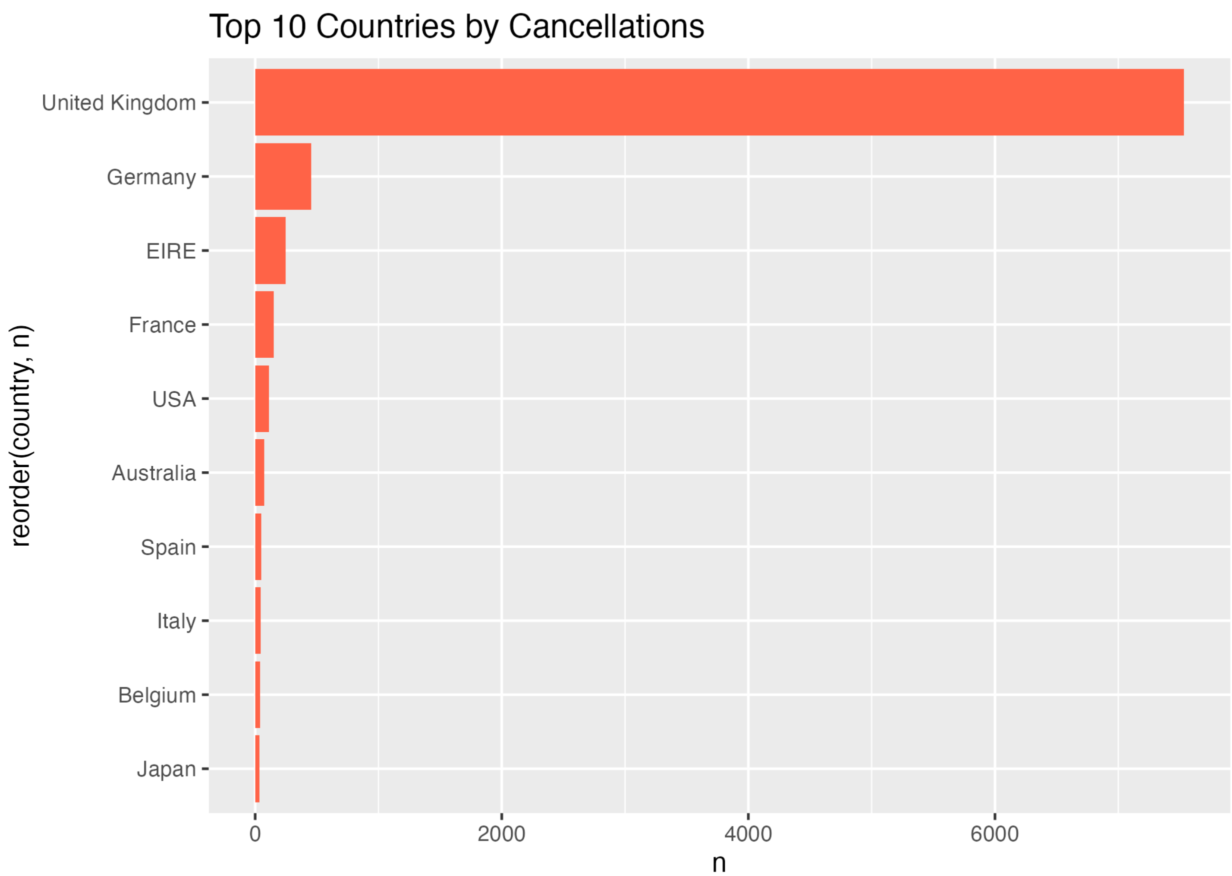

Weekend vs Weekday AOV:

```r
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
```

```
##
##  Welch Two Sample t-test
##
## data:  log_aov by is_weekend
## t = -0.41225, df = 3115.6, p-value = 0.6802
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.05613110  0.03662823
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            5.582376            5.592127
```

Top 20% Monetary Customers Frequency (t-test)

```
ttest_top20
```

```
##
##  Welch Two Sample t-test
##
## data:  frequency by top20
## t = -19.192, df = 881.82, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -12.59663 -10.25928
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            2.788043           14.216000
```

Chi-square Test: Cancellations by Top 10 Countries

```
chi_country
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  chi_table
## X-squared = 100.16, df = 1, p-value < 2.2e-16
```

Chi-square Test: Weekend vs Weekday Cancellations

```
chi_weekend_test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  chi_weekend
## X-squared = 19.331, df = 1, p-value = 1.099e-05
```

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
```

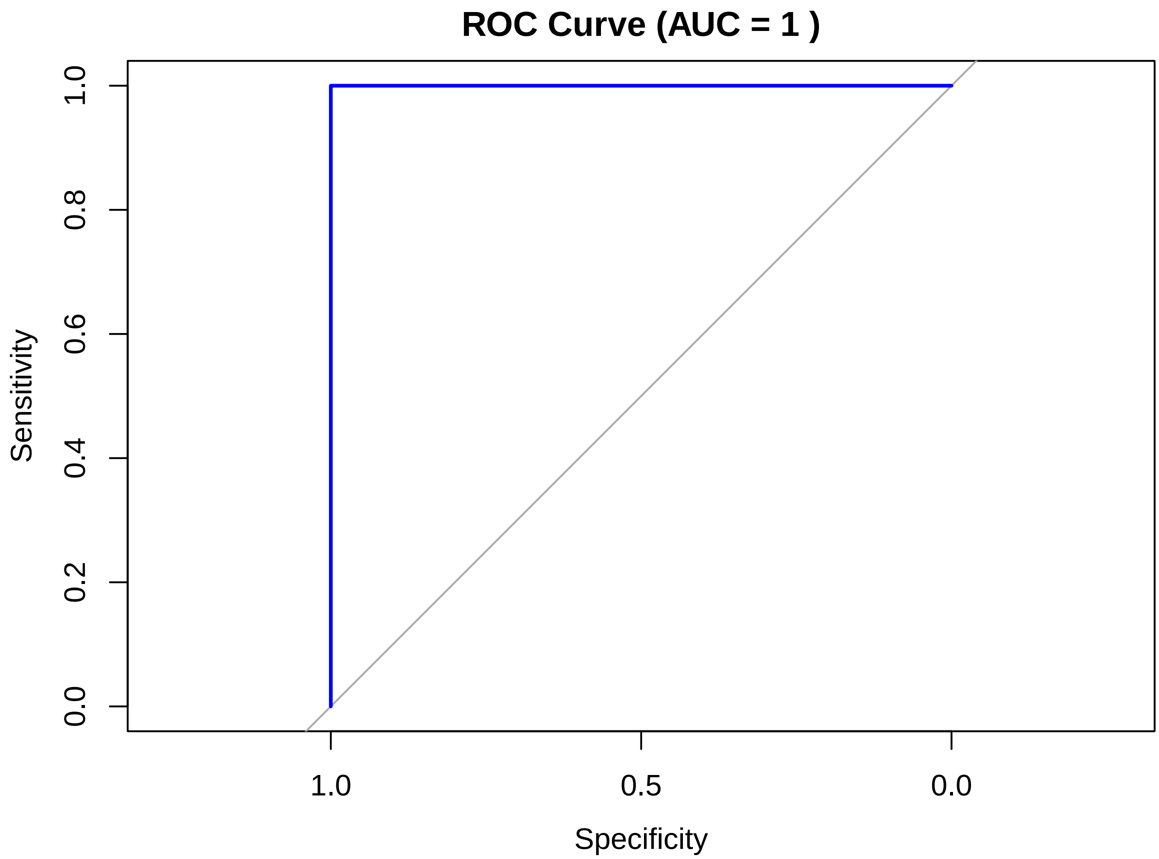
```
##          0 4239    0
##          1    0 1331
##
##                Accuracy : 1
##                  95% CI : (0.9993, 1)
##     No Information Rate : 0.761
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.000
##             Specificity : 1.000
##          Pos Pred Value : 1.000
##          Neg Pred Value : 1.000
##              Prevalence : 0.239
##          Detection Rate : 0.239
##    Detection Prevalence : 0.239
##       Balanced Accuracy : 1.000
##
##        '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

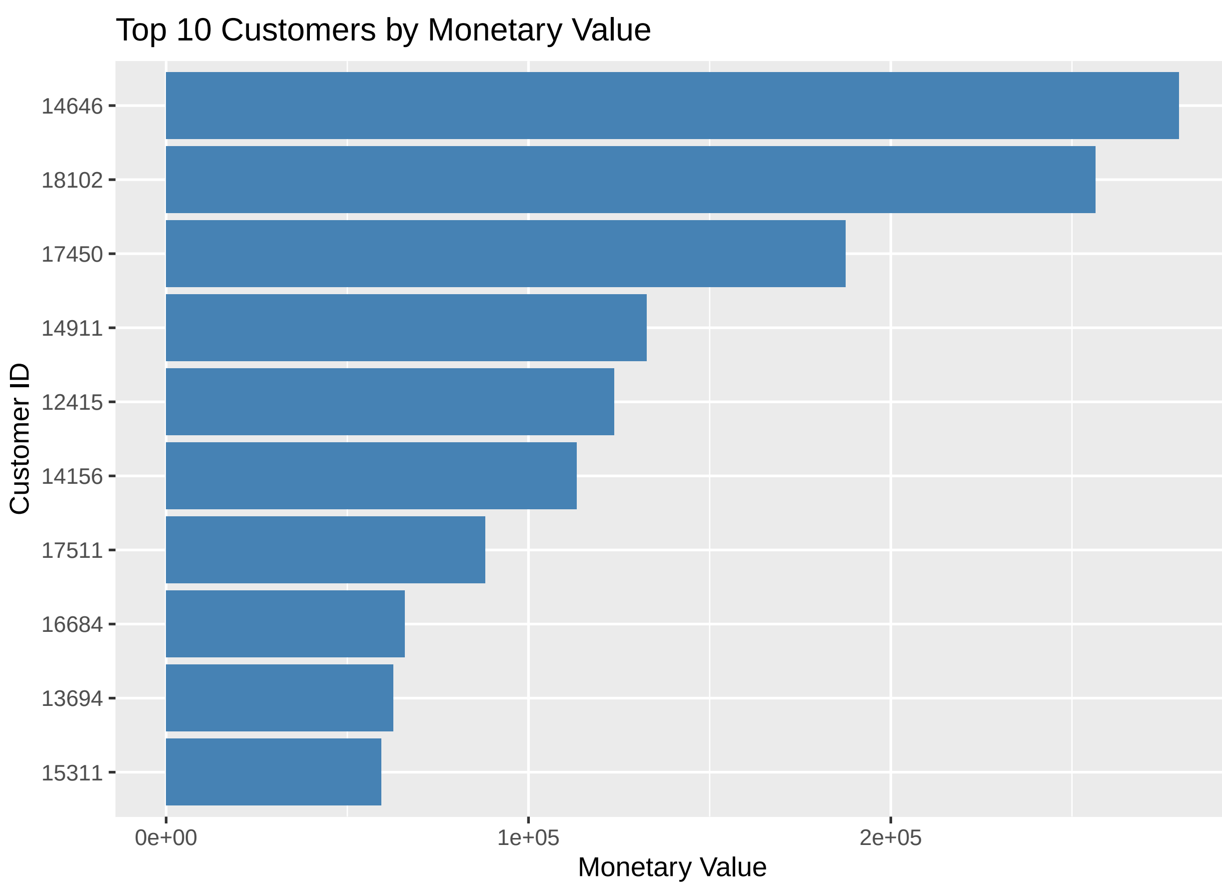
```
top_customers
```

```
## # A tibble: 20 x 9
##    customer_id recency_days frequency monetary top20 r_score f_score m_score
##          <dbl>        <dbl>     <int>    <dbl> <lgl>   <int>   <int>   <int>
##  1       14646           NA        76  279489. TRUE       NA       5       5
##  2       18102           NA        62  256438. TRUE       NA       5       5
##  3       17450           NA        55  187482. TRUE       NA       5       5
##  4       14911           NA       248  132573. TRUE       NA       5       5
##  5       12415           NA        26  123725. TRUE       NA       5       5
##  6       14156           NA        66  113384. TRUE       NA       5       5
##  7       17511           NA        46   88125. TRUE       NA       5       5
##  8       16684           NA        31   65892. TRUE       NA       5       5
##  9       13694           NA        60   62653. TRUE       NA       5       5
## 10       15311           NA       118   59419. TRUE       NA       5       5
```

```
## 11        13089           NA        118    57386. TRUE       NA        5        5
## 12        14096           NA         34    57121. TRUE       NA        5        5
## 13        15061           NA         55    54229. TRUE       NA        5        5
## 14        17949           NA         52    52751. TRUE       NA        5        5
## 15        15769           NA         29    51824. TRUE       NA        5        5
## 16        16029           NA         76    50993. TRUE       NA        5        5
## 17        14298           NA         45    50862. TRUE       NA        5        5
## 18        14088           NA         14    50415. TRUE       NA        5        5
## 19        17841           NA        169    40341. TRUE       NA        5        5
## 20        13798           NA         63    36351. TRUE       NA        5        5
## # i 1 more variable: rfm_score <chr>
```

Top 10 Customers Visualization

```r
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.

## 9. References


- Customer segmentation with RFM analysis
- Logistic regression and ROC/AUC for classification
- Tidyverse for reproducible data cleaning and visualization