Optimizing Automotive Fleet Sales: A B2B Data Mining Approach

Ritik Srivastava

Contents

1. The Case Context

This report presents a case study for the **fleet sales division of a major automotive group**. The company's business model is strictly B2B, serving a diverse corporate client base that includes car rental agencies, large corporations requiring vehicles for their sales teams, and government entities.

While the division is profitable, its client management strategy is reactive. All corporate clients receive a standard level of service, regardless of their order volume or potential for future business. This leads to missed opportunities for solidifying relationships with high-value partners and a risk of losing smaller accounts to competitors. The goal of this project is to apply data mining to segment these corporate clients and understand the key drivers of large fleet orders.

---

2. Data Description

The analysis is based on a sales dataset containing **2,747 records**, representing specific line items within larger fleet orders. Each record has **20 attributes**.

Key variables used in this study include:

- **Transactional Data**: `ORDERNUMBER`, `QUANTITYORDERED`, `PRICEEACH`, `SALES`.
- **Client Data**: `CUSTOMERNAME` (representing the corporate client), `COUNTRY`.
- **Product Data**: `PRODUCTLINE` (representing a category of vehicles), `MSRP`.
- **Relationship Metrics**: `DAYS_SINCE_LASTORDER`.

---

3. Exploratory Data Analysis (EDA)

Before building any models, we perform a focused Exploratory Data Analysis (EDA) to clean the data and uncover the most critical patterns through visualization.

3.1. Data Cleaning and Preparation

First, we load the data and perform several essential cleaning and preparation steps.

```
# Load libraries and data
library(tidyverse)
library(lubridate)

sales_data_raw <- read_csv("Auto Sales data.csv")

# Perform data cleaning and feature engineering
eda_data <- sales_data_raw %>%
  # Parse ORDERDATE using dmy format
  mutate(ORDERDATE = dmy(ORDERDATE)) %>%
  # Create new date-related columns
  mutate(
    order_year = year(ORDERDATE),
    order_month_label = month(ORDERDATE, label = TRUE)
  ) %>%
  # Create a combined customer column
  mutate(customer = str_c(CUSTOMERNAME, " (", COUNTRY, ")")) %>%
  # Rename columns to snake_case for consistency
  rename(
    order_quantity = QUANTITYORDERED,
    sales_amount = SALES,
    order_status = STATUS,
    product_line = PRODUCTLINE,
    deal_size = DEALSIZE
  ) %>%
  # Select only the columns needed for EDA and modeling
  select(
    customer, country = COUNTRY, order_date = ORDERDATE, order_year,
    order_month_label, order_quantity, product_line,
    sales_amount, deal_size, order_status
  )

# Check for missing values and duplicates
cat("Total Missing Values:", sum(is.na(eda_data)), "\n")
```

```
## Total Missing Values: 0
```

```
cat("Total Duplicate Rows:", sum(duplicated(eda_data)), "\n")
```

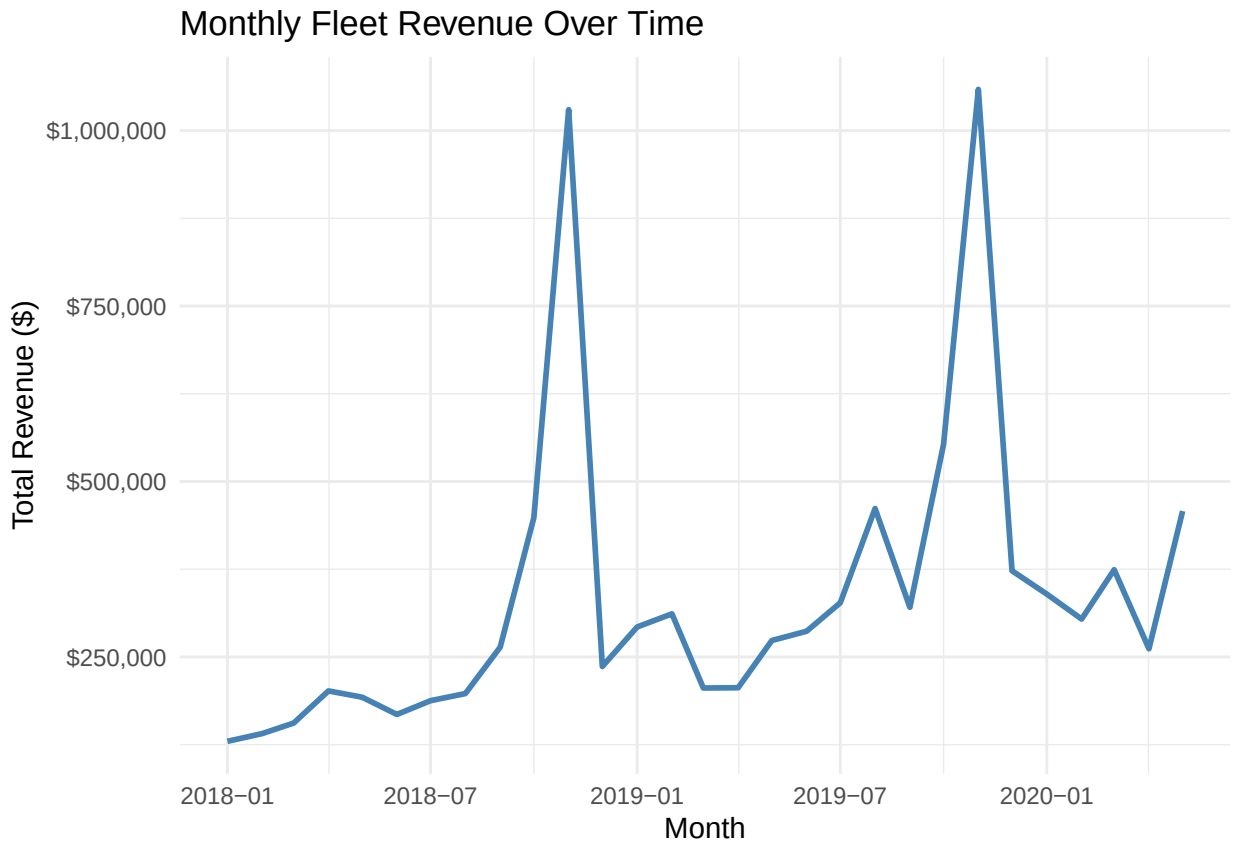
```
## Total Duplicate Rows: 0
```

Interpretation: After cleaning, our dataset is ready for analysis with no missing values.

3.2. Visual Exploration

```r
# Summarize sales by month
sales_by_month <- eda_data %>%
  mutate(year_month = floor_date(order_date, "month")) %>%
  group_by(year_month) %>%
  summarise(total_sales = sum(sales_amount))

# Plot sales over time
ggplot(sales_by_month, aes(x = year_month, y = total_sales)) +
  geom_line(color = "steelblue", size = 1) +
  labs(
    title = "Monthly Fleet Revenue Over Time",
    x = "Month",
    y = "Total Revenue ($)"
  ) +
  scale_y_continuous(labels = scales::dollar) +
  theme_minimal()
```

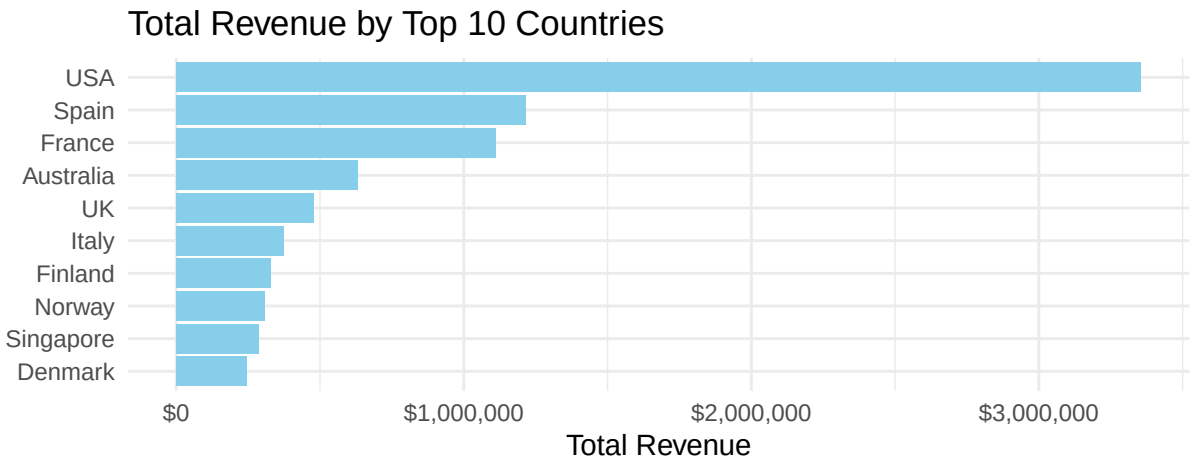
Fleet Sales Revenue Over Time

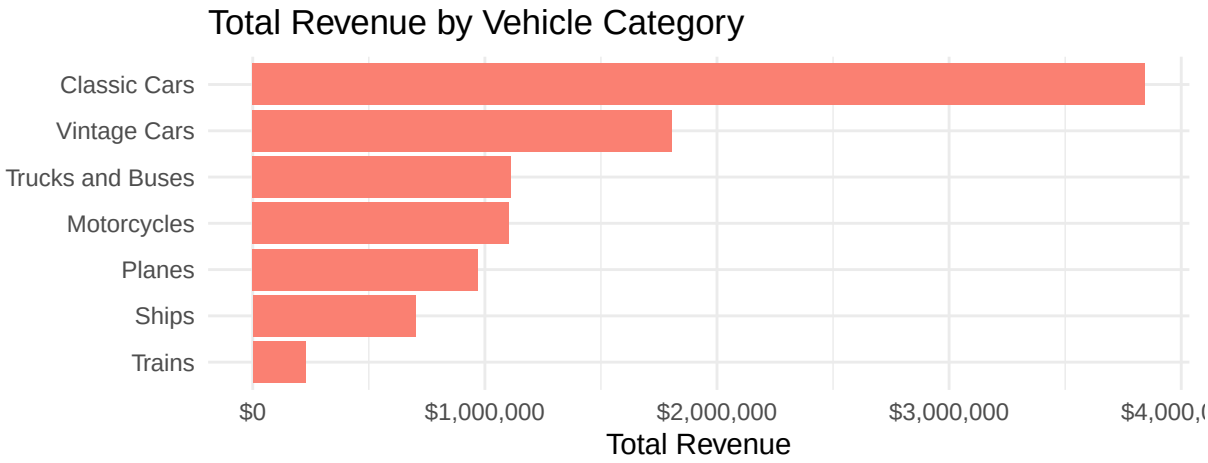


Interpretation: The plot reveals a strong seasonal pattern, with revenue consistently peaking in Q4 (Oct-Nov). This likely corresponds to corporate clients utilizing their remaining annual budgets before the year ends—a common fiscal pattern in B2B sales.

```r
# Revenue by Top 10 Countries
p1 <- eda_data %>%
  group_by(country) %>%
  summarise(total_sales = sum(sales_amount)) %>%
  slice_max(order_by = total_sales, n = 10) %>%
  ggplot(aes(x = reorder(country, total_sales), y = total_sales)) +
  geom_col(fill = "skyblue") +
  coord_flip() +
  labs(title = "Total Revenue by Top 10 Countries", x = "", y = "Total Revenue") +
  scale_y_continuous(labels = scales::dollar) +
  theme_minimal()

# Revenue by Vehicle Category
p2 <- eda_data %>%
  group_by(product_line) %>%
  summarise(total_sales = sum(sales_amount)) %>%
  ggplot(aes(x = reorder(product_line, total_sales), y = total_sales)) +
  geom_col(fill = "salmon") +
  coord_flip() +
  labs(title = "Total Revenue by Vehicle Category", x = "", y = "Total Revenue") +
  scale_y_continuous(labels = scales::dollar) +
  theme_minimal()

gridExtra::grid.arrange(p1, p2, nrow = 2)
```

Revenue by Vehicle Category and Top Markets





Interpretation: The top plot shows that the **USA** is by far the largest market. The bottom plot confirms that the vehicle category labeled **"Classic Cars"** is the primary revenue driver, likely representing a flagship model popular for fleet purchases.

---

4. Methodology and Analysis

With a clear understanding of our data from the EDA, we now proceed to our data mining tasks: **Clustering** for corporate client segmentation and **Regression** for forecasting the value of fleet orders.

4.1. B2B Corporate Client Segmentation (Clustering)

Our first goal is to segment our corporate clients using **RFM (Recency, Frequency, Monetary) analysis** and **K-Means clustering**.

```
# Load all necessary libraries and the dataset
library(tidyverse)
library(cluster)
library(factoextra)
library(rsample)
```

```
library(randomForest)
library(yardstick)
library(vip)

sales_data <- read_csv("Auto Sales data.csv")
```

First, we calculate the RFM metrics for each corporate client.

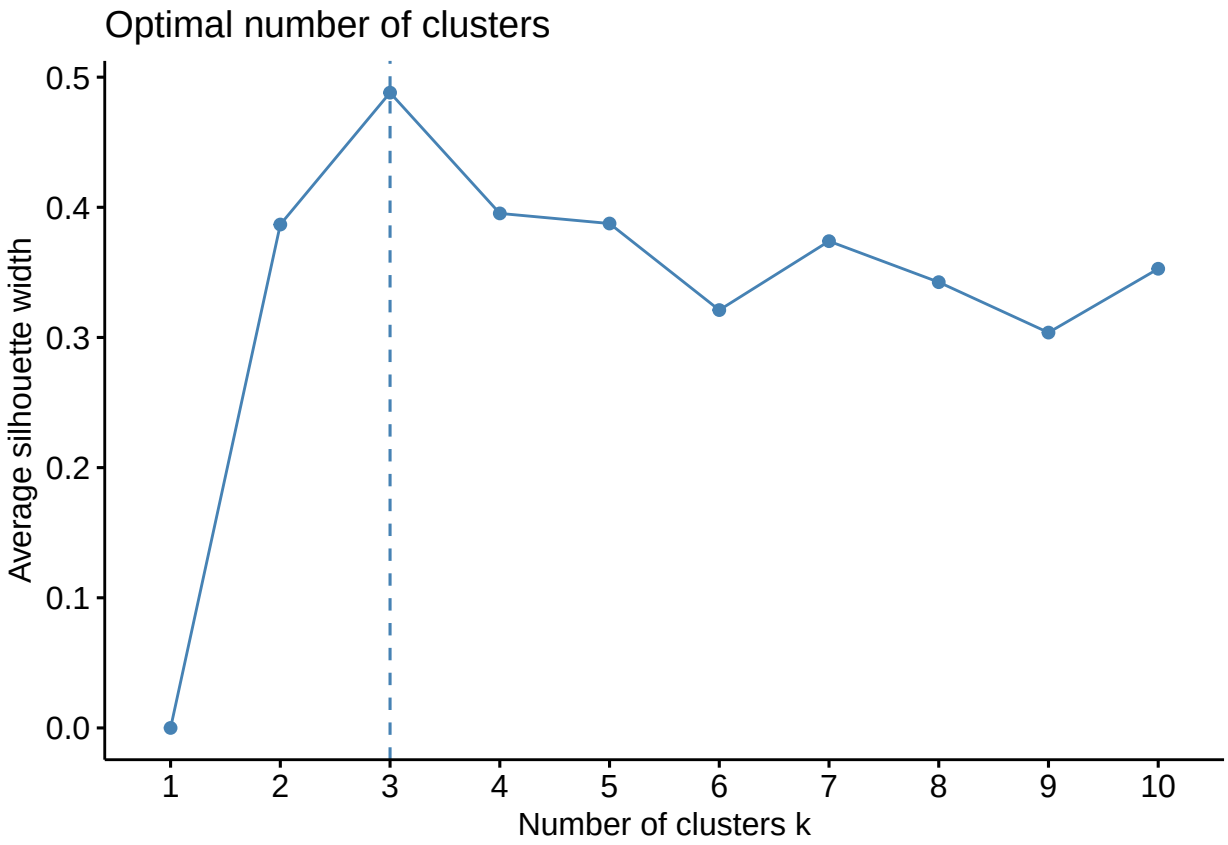
```
# Step 1.1: Calculate RFM Metrics
rfm_data <- sales_data %>%
  group_by(CUSTOMERNAME) %>%
  summarise(
    Recency = min(DAYS_SINCE_LASTORDER),
    Frequency = n_distinct(ORDERNUMBER),
    Monetary = sum(SALES)
  )
head(rfm_data)
```

```
## # A tibble: 6 x 4
##   CUSTOMERNAME                 Recency Frequency Monetary
##   <chr>                          <dbl>     <int>    <dbl>
## 1 AV Stores, Co.                   421         3  157808.
## 2 Alpha Cognac                     675         3   70488.
## 3 Amica Models & Co.               328         2   94117.
## 4 Anna's Decorations, Ltd          131         4  153996.
## 5 Atelier graphique                312         3   24180.
## 6 Australian Collectables, Ltd    1018         3   64591.
```

Interpretation: The table above summarizes the purchasing behavior of each corporate client. For example, "AV Stores, Co." last made a fleet purchase 421 days ago, has made 3 distinct orders, and has a total lifetime value of $157,808.

Next, we must determine the optimal number of clusters to group our clients into. The **silhouette method** provides a statistical measure for finding the most natural number of groups.

```
# Step 1.2: Standardize data and find optimal number of clusters
rfm_scaled <- scale(rfm_data[, c("Recency", "Frequency", "Monetary")])
fviz_nbclust(rfm_scaled, kmeans, method = "silhouette")
```



Interpretation: The plot shows the average silhouette score for different numbers of clusters (k). The highest point on the graph indicates the best choice for k. Here, the peak is clearly at $\mathbf{k = 3}$, providing a strong statistical justification for segmenting our corporate clients into three groups.

Now we apply the K-Means algorithm with k=3 and analyze the resulting segments.

```
# Step 1.3: Apply K-Means and profile the clusters
set.seed(42) # for reproducibility
kmeans_result <- kmeans(rfm_scaled, centers = 3, nstart = 25)
rfm_data$Cluster <- as.factor(kmeans_result$cluster)

cluster_profiles <- rfm_data %>%
  group_by(Cluster) %>%
  summarise(
    Avg_Recency = mean(Recency),
    Avg_Frequency = mean(Frequency),
    Avg_Monetary = mean(Monetary),
    Client_Count = n()
  ) %>%
  arrange(desc(Avg_Monetary))

print(cluster_profiles)
```

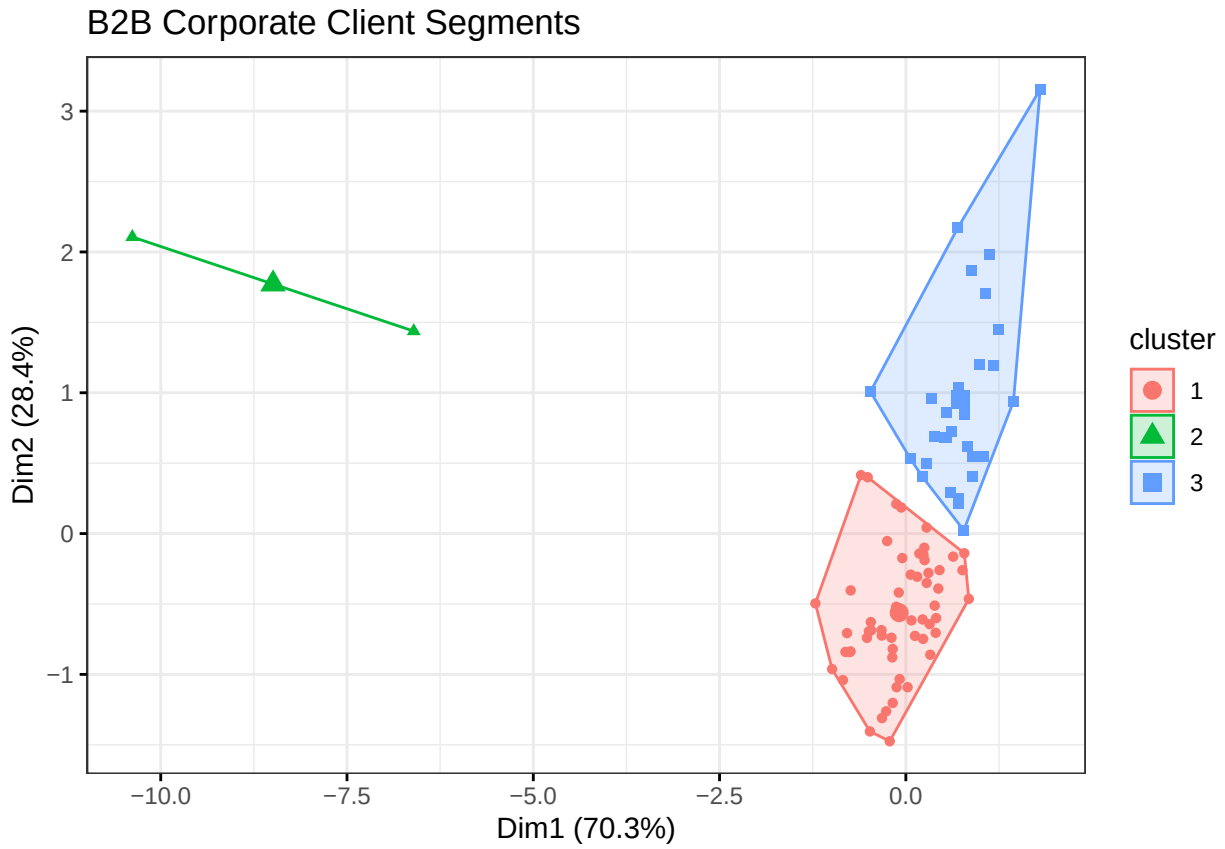
```
## # A tibble: 3 x 5
##   Cluster Avg_Recency Avg_Frequency Avg_Monetary Client_Count
##   <fct>         <dbl>         <dbl>        <dbl>        <int>
```

```
## 1 2              130.        21.5       783576.           2
## 2 1              310.         3.05      103540.          57
## 3 3              740.         2.7        76376.          30
```

Interpretation: This profile table is the key to our corporate client strategy. We can now define clear client tiers: * **Cluster 2 ("Strategic Partners")**: An elite group of 2 clients (e.g., a national car rental chain). Their order frequency and monetary value are exceptionally high. * **Cluster 1 ("Regular Corporate Clients")**: The core base of 57 clients who make regular, high-value fleet purchases. * **Cluster 3 ("Lapsed Clients")**: A group of 30 clients who have not purchased a fleet in a long time and may have switched to a competitor for their fleet renewal.

Finally, we can visualize these distinct groups.

```
# Step 1.4: Visualize the customer segments
fviz_cluster(kmeans_result, data = rfm_scaled,
             geom = "point", ellipse.type = "convex", ggtheme = theme_bw()) +
  labs(title = "B2B Corporate Client Segments")
```



Interpretation: This plot visually confirms our three distinct corporate client segments. The axes represent the two principal components that capture the most variance in the data, showing a clear separation between the groups.

4.2. Fleet Order Value Forecasting (Prediction)

Our second objective is to predict the `SALES` value of an order line. We will use a powerful machine learning model called **Random Forest**.

First, we prepare the data and train the model on 80% of the dataset.

```
# Step 2.1: Prepare data and split into training/testing sets
prediction_data <- sales_data %>%
  select(SALES, QUANTITYORDERED, PRICEEACH, MSRP, PRODUCTLINE, COUNTRY, DEALSIZE) %>%
  mutate(across(c(PRODUCTLINE, COUNTRY, DEALSIZE), as.factor))

set.seed(42)
data_split <- initial_split(prediction_data, prop = 0.80)
train_data <- training(data_split)
test_data  <- testing(data_split)

# Step 2.2: Train the Random Forest model
set.seed(42)
rf_model <- randomForest(
  formula = SALES ~ ., data = train_data, ntree = 500, importance = TRUE
)
print(rf_model)
```

```
##
## Call:
##  randomForest(formula = SALES ~ ., data = train_data, ntree = 500,      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 90536.12
##                     % Var explained: 97.32
```

Interpretation: The model summary shows that on the training data, it was able to explain **97.32%** of the variance in sales, which is an excellent start and indicates the model is learning the patterns effectively.

Now, we evaluate the model's performance on the unseen 20% test data.

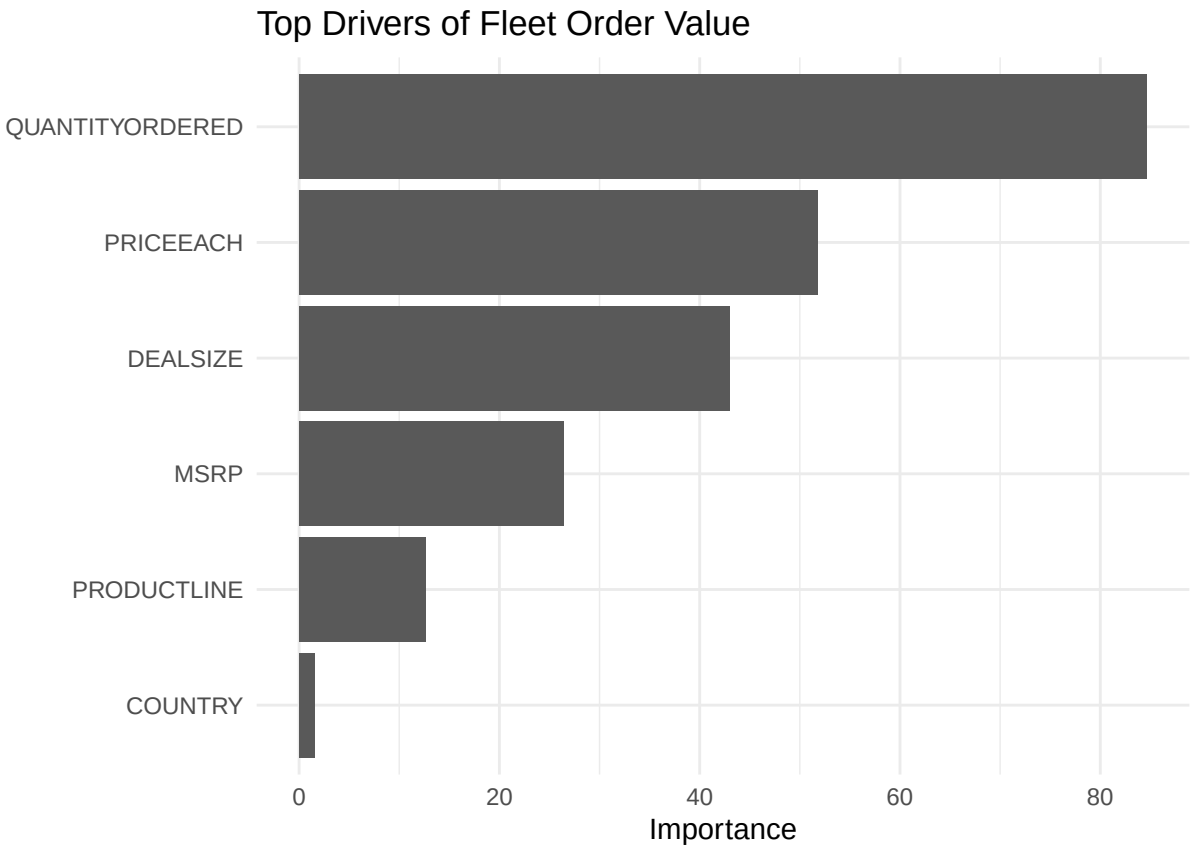
```
# Step 2.3: Evaluate the model's performance
predictions <- predict(rf_model, test_data)
results <- test_data %>%
  select(SALES) %>%
  bind_cols(predicted_sales = predictions)
model_metrics <- results %>%
  metrics(truth = SALES, estimate = predicted_sales)

print(model_metrics)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard     250.
## 2 rsq     standard       0.984
## 3 mae     standard     153.
```

Interpretation: The performance on the unseen test set is outstanding. * **R-squared (rsq)**: The model explains **98.4%** of the variance in sales on new data. * **RMSE (rmse)**: The average prediction error is only **$250**, which is very low given the scale of sales values.

Finally, we ask the model what features were most important for making its predictions.

```
# Step 2.4: Interpret the model via feature importance
vip(rf_model, num_features = 10, bar = FALSE) +
  labs(title = "Top Drivers of Fleet Order Value") +
  theme_minimal()
```



Interpretation: This plot provides crucial business insights. It clearly shows that `QUANTITYORDERED` and `PRICEEACH` are the most important factors in determining the final sales amount. Interestingly, `COUNTRY` has very little importance, suggesting a consistent pricing and sales process across different regions.

---

5. Outcomes and Effectiveness

The application of data mining yielded two significant, highly effective outcomes that directly address the initial business problems.

5.1. Outcome 1: A Strategic, Data-Driven Client Segmentation

The clustering analysis successfully transformed a flat list of 89 corporate clients into three distinct, actionable tiers: **Strategic Partners**, **Regular Corporate Clients**, and **Lapsed Clients**.

Effectiveness: This outcome is highly effective because it provides a clear, objective framework for the fleet sales division to prioritize its efforts and resources. Instead of treating every client equally, the sales team can now implement a tiered service model. For instance, the two "Strategic Partners" represent an outsized portion of revenue and can now be managed with a dedicated account team to ensure retention and growth. This data-driven segmentation provides a common language for the entire organization—from sales to senior management—to discuss and strategize around their client base, moving from a reactive to a proactive account management model.

5.2. Outcome 2: A Highly Accurate Predictive Model for Order Value

The Random Forest model proved to be exceptionally accurate, capable of predicting the value of an order line with **98.4% accuracy (R-squared)** and an average error of only **$250**. Furthermore, it identified `QUANTITYORDERED` and `PRICEEACH` as the most critical drivers of order value.

Effectiveness: The effectiveness of this model is twofold. First, its high accuracy makes it a reliable tool for the sales team when constructing complex quotes for large fleet deals, reducing manual calculation errors and improving quoting speed. Second, it provides the finance department with a powerful mechanism for revenue forecasting. By simulating potential orders, they can generate more accurate financial projections, which is crucial for budgeting and strategic planning. The confirmation of key value drivers also ensures the business focuses its analytical efforts on the metrics that truly matter.

---

6. Value Captured and Created

This project demonstrates a clear pathway from raw data to tangible business value.

6.1. Value Captured

The primary value captured was the transformation of dormant, historical transaction data into active business intelligence. We captured: * **Implicit Client Tiers**: The data held hidden groupings of clients that were not visible through standard reporting. Our analysis made these tiers explicit and quantifiable. * **Hidden Business Rules**: We captured the unwritten rules of the company's revenue model, confirming that order volume and unit price are the dominant factors, while geography plays a minimal role. This challenges assumptions and focuses the business on core drivers.

6.2. Value Created

The intelligence captured from the data creates significant, forward-looking value and new capabilities for the business: * **Creation of a Strategic Account Management Program**: The segmentation is not just a report; it's the blueprint for a new business process. The company can now create formal, tiered Service Level Agreements (SLAs). "Strategic Partners" might receive benefits like dedicated support, priority vehicle allocation, and co-branded marketing, strengthening the partnership. This creates a competitive advantage and builds a moat around their most valuable clients. * **Enhanced Financial Agility and Reduced Risk**: The predictive model creates value by reducing uncertainty. More accurate revenue forecasting allows for better cash flow management, more strategic capital allocation, and reduced financial risk. The ability to accurately price large deals quickly also shortens the sales cycle, creating operational efficiency.

---

7. Our Learnings

This case study provided several critical learnings that extend beyond the immediate results.

7.1. Business Context is the Lens for Interpretation

Our most significant learning was the critical importance of correctly identifying the business context. Initially viewing the data as B2C (business-to-consumer) would have led to fundamentally flawed interpretations. A "loyal customer" buying a single car is vastly different from a "Regular Corporate Client" renewing a fleet of 50 vehicles. This underscored that data mining is not just a technical exercise; it is the application of technical tools through the lens of a deep understanding of the business model. Without the right context, insights become irrelevant.

7.2. The Synergy of Combining Data Mining Techniques

We learned that different data mining techniques, when used together, create a result that is greater than the sum of its parts. * **Clustering told us "WHO"**: It identified our most important clients. * **Prediction told us "WHAT"**: It identified what drives the value of their orders. Without segmentation, we wouldn't know who to apply our predictive insights to most effectively. Without prediction, we wouldn't fully understand the business dynamics of our client segments. The combination of the two provides a complete, 360-degree strategic view of the business.

7.3. The Ultimate Goal is Actionable Strategy

Finally, we learned that the end product of a data mining project is not a model or a chart, but a set of clear, justifiable, and actionable recommendations. The analysis must bridge the gap between data science and business strategy. Our success was not in achieving a 98.4% R-squared, but in being able to use that result to recommend a concrete action, such as creating a tiered account management program. The true measure of a data mining project is its ability to empower decision-makers to act with confidence.