Sharon Kihara

136372

Submitted in partial fulfilment of the requirements of the Bachelor of Business in Information Technology at Strathmore University

Faculty of Information Technology

Strathmore University

Nairobi, Kenya

2nd November 2024

Declaration and Approval

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the research proposal contains no material previously published or written by another person except where due reference is made in the research proposal itself.

Student Name: Sharon Kihara

Admission Number: 136372

Student Signature: ______________________________ Date: _________________________

Abstract

This project explores the application of data analytics and machine learning techniques to address key business challenges at Classic Models. Using a relational database, the study focuses on customer behavior, product performance, demand forecasting, pricing optimization, and customer feedback analysis.

The data preparation phase involved handling missing values, removing duplicates, and standardizing data to create a clean and reliable dataset. Descriptive analytics revealed insights into revenue distribution, customer spending patterns, seasonal sales trends, and product performance, identifying key revenue drivers such as high-value customers and top-performing products like the 1992 Ferrari 360 Spider Red.

Predictive analytics emphasized demand forecasting and pricing optimization. Regression models underperformed in capturing non-linear trends and seasonality in demand. To address this, Facebook Prophet was implemented, providing accurate demand forecasts and actionable insights for inventory planning. Pricing optimization used polynomial regression to identify optimal price points, maximizing revenue while balancing customer affordability.

Lastly, topic modeling using Latent Dirichlet Allocation (LDA) analyzed customer comments to uncover key themes such as delayed deliveries, product quality issues, and customer support concerns. These findings align with earlier analyses, offering actionable recommendations to improve service, enhance customer satisfaction, and drive strategic growth.

Table of Contents

Contents

List of Figures

List of Tables

List of Abbreviations

Chapter 1: Introduction

1.1 Background

Classic Models is a globally recognized business specializing in the creation and sale of high-quality miniature replicas of classic vehicles, including cars, motorcycles, airplanes, and ships. These meticulously crafted scale models are designed to cater to the refined tastes of collectors and hobbyists who value the precision, detail, and artistry involved in such craftsmanship. The company has built its reputation on the uniqueness of its offerings, providing a diverse catalog of products that appeal to a broad spectrum of customers across the globe. With an expanding product range and a growing customer base, Classic Models is poised for significant growth but is also faced with several operational and strategic challenges.

Among the most pressing issues for Classic Models is the need to better understand customer purchasing patterns. With such a diverse catalog, identifying which products resonate most with customers is vital for maintaining inventory efficiency and ensuring customer satisfaction. Properly aligning inventory with demand not only minimizes the risk of overstocking slow-moving items but also helps avoid missed sales opportunities for popular products. Furthermore, as the business expands, managing operations efficiently across multiple regions and offices becomes increasingly complex. This includes evaluating employee contributions to revenue, understanding the performance of regional offices, and ensuring consistency in service delivery.

Another critical aspect of the business lies in understanding customer feedback. In today's competitive market, customer satisfaction is pivotal for success, and analyzing feedback in the form of comments, reviews, and suggestions can uncover recurring themes such as product quality issues, delayed deliveries, or gaps in service. Addressing these concerns proactively can help improve customer retention and loyalty. To tackle these challenges and maintain its competitive edge, Classic Models must leverage the vast amount of data it collects from its operations. This data, stored in a relational database, encompasses key aspects of the business, including customer details, order histories, product information, and financial transactions. Harnessing this data through advanced analytics can empower Classic Models to make informed, data-driven decisions that enhance both operational efficiency and customer satisfaction.

This project is designed to comprehensively analyze the Classic Models database using a combination of descriptive analytics, predictive modeling, and Natural Language Processing (NLP) techniques. The goal is to provide actionable insights that address key business questions, such as identifying top-performing products, forecasting demand, segmenting customers, and extracting themes from customer feedback. By aligning analytical methods with business needs, the project seeks to support Classic Models in achieving its strategic objectives and driving long-term growth.

1.2 Objectives

The analysis of the Classic Models database is driven by a clear set of objectives aimed at addressing the company's most pressing challenges and providing actionable insights. First and foremost, the project seeks to deepen the understanding of customer behavior. By analyzing purchasing patterns, the analysis aims to identify high-value customers—those who contribute the most to revenue—and understand their preferences and spending habits. This information can then be used to design targeted marketing campaigns and personalized promotions that drive customer engagement and loyalty. Additionally, the project seeks to quantify metrics such as average customer spend and purchase frequency, which are essential for profiling customer segments and tailoring business strategies to their needs.

Another critical objective of the analysis is to evaluate product performance. Classic Models offers a wide array of products, and understanding which items are top sellers is crucial for effective inventory management. By analyzing sales data at both the product and category levels, the project aims to identify trends in customer preferences and assess the revenue contributions of various product lines. This information is instrumental in making informed decisions about production, pricing, and promotional efforts.

Sales trends are another area of focus for this project. By examining historical sales data, the analysis aims to uncover seasonal patterns and identify peak periods of demand. Such insights are invaluable for resource planning, ensuring that the company is well-prepared to meet customer needs during high-demand periods while avoiding the inefficiencies of overproduction during low-demand months. Understanding sales trends also lays the foundation for accurate demand forecasting, which is essential for long-term planning and operational stability.

In addition to analyzing customer and product data, the project places significant emphasis on improving operational efficiency. Evaluating employee performance is a key component of this objective. By analyzing the contributions of sales representatives and regional offices to total revenue, the project seeks to identify areas of excellence and opportunities for improvement. Such insights can help the company allocate resources more effectively, provide targeted training for underperforming employees, and replicate best practices across the organization.

The project also extends beyond descriptive analytics to include predictive modeling. Using advanced machine learning techniques, the analysis aims to forecast product demand, optimize pricing strategies, and segment customers into meaningful groups based on their purchasing behavior. For example, clustering techniques are used to group customers based on Recency, Frequency, and Monetary (RFM) values, providing a clear picture of customer segments and their relative importance to the business. Regression models are applied to forecast future demand for specific products and optimize pricing to balance sales volume and revenue. Employee performance prediction models are also explored to help the company anticipate future outcomes based on historical data.

Finally, the project incorporates Natural Language Processing (NLP) techniques to analyze customer feedback. Customer comments often contain valuable insights that are difficult to extract through traditional analysis methods. By applying Latent Dirichlet Allocation (LDA), a popular topic modeling algorithm, the analysis aims to uncover key themes and sentiments expressed in customer feedback. Whether it is identifying recurring issues with product quality or pinpointing concerns about service delays, this analysis provides actionable insights that can guide improvements in customer experience and operational processes.

1.3 Scope

The scope of this project is defined by the comprehensive nature of the Classic Models database, a relational database that captures various dimensions of the company’s operations. The database schema consists of eight interconnected tables that collectively provide a holistic view of the business. The **customers** table stores detailed information about each customer, including their names, contact details, credit limits, and the sales representatives managing their accounts. This table forms the foundation for understanding customer demographics and purchasing behavior. The **products** table contains detailed information about the company’s offerings, including product names, categories, scales, vendors, descriptions, and pricing.

These attributes are critical for analyzing product performance and identifying trends in customer preferences.

The **productlines** table groups products into broader categories, such as "Classic Cars," "Motorcycles," and "Planes," providing a high-level view of product performance by category. The **orders** and **orderdetails** tables capture transactional data, including order dates, shipping dates, quantities ordered, and prices. Together, these tables enable a detailed analysis of sales trends and revenue contributions. The **payments** table tracks financial transactions, including payment dates and amounts, offering insights into customer payment behavior and financial health. The **employees** table contains information about the company's workforce, including job titles, contact details, and reporting structures, while the **offices** table provides details about the company's regional locations and their contributions to overall sales.

To analyze this data, Python was used in conjunction with libraries such as Pandas for data manipulation, SQLAlchemy for database connectivity, and machine learning libraries for predictive modeling. Visualizations were created using Matplotlib and Seaborn to present findings in an intuitive and accessible manner. By integrating data from these interconnected tables, the analysis provides a detailed and comprehensive view of the business, enabling Classic Models to make data-driven decisions that align with its strategic objectives.

Chapter 2: Data Preparation

The data preparation phase is a critical step in ensuring that the dataset is clean, consistent, and ready for analysis. For this project, the Classic Models relational database served as the foundation for deriving insights into customer behavior, product performance, and operational efficiency. This database contains eight interconnected tables: customers, products, productlines, orders, orderdetails, payments, employees, and offices. Each table contributes unique information about various aspects of the business, from customer details to product categories and financial transactions. To ensure the accuracy and reliability of the analysis, the data underwent a rigorous cleaning process, which included identifying and handling missing values, validating and standardizing data types, and removing duplicate records. These steps established a solid foundation for descriptive and predictive analytics.

2.1 Handling Missing Values

A comprehensive examination of all columns across the Classic Models database tables revealed notable gaps in the data.

For example, in the customers table, key fields such as state and postalCode exhibited missing values, alongside gaps in the salesRepEmployeeNumber column, which associates customers with specific sales representatives. Additionally, the addressLine2 column had 100 missing entries, although this field is often non-essential for analysis.

In the orders table, several records in the comment’s column were found to be incomplete. This field frequently contains useful information, such as customer feedback or special instructions related to orders. To maintain dataset integrity, these missing values were filled with a placeholder value, "No comments," ensuring no critical data was lost while preparing the dataset for analysis.

For numerical fields with gaps, such as the salesRepEmployeeNumber, careful evaluation is required to determine whether to impute values, remove records, or leave them as-is based on the analysis objective. Text fields, such as state, were left as missing when not critical to the analysis, minimizing unnecessary assumptions.

Addressing these missing values made the dataset more complete, consistent, and ready for further analytical processes.

```
Checking data for table: customers
Missing Values in customers:
customerNumber              0
customerName                0
contactLastName             0
contactFirstName            0
phone                       0
addressLine1                0
addressLine2              100
city                        0
state                      73
postalCode                  7
country                     0
salesRepEmployeeNumber     22
creditLimit                 0
dtype: int64
```

Figure 1 Output for missing values

2.2 Addressing Duplicate Records

Duplicate records can distort the results of analysis by artificially inflating metrics such as revenue or sales volume. To mitigate this risk, all tables were systematically checked for duplicates. For instance, duplicate entries in the customers table could result in double-counting customer activity, while duplicate records in the orderdetails table could misrepresent product performance.

Upon inspection, the majority of tables, including customers, products, and orders, were found to be free of duplicates. This indicates that the data collection process was well-managed. However, had duplicates been detected, they would have been removed to ensure the accuracy of the analysis. This verification step reinforced confidence in the integrity of the dataset.

```
Removed 0 duplicate rows from customers.
Checking data for table: products
Missing Values in products:
productCode          0
productName          0
productLine          0
productScale         0
productVendor        0
...
2      apt. 5A          NY     USA      10022         NA
3         None        None  France      75017       EMEA
4         None  Chiyoda-Ku   Japan   102-8578      Japan
```

Figure 2 Output for duplicate values

2.3 Final Dataset

The final step in the data preparation process involved verifying the integrity of the cleaned dataset. Sample records from each table were reviewed to ensure that the data was correctly processed and ready for analysis. For instance, a sample from the orders table was inspected to confirm that date formats were standardized, missing values in the comments column were appropriately filled, and no duplicates were present. These checks demonstrated that the data was clean, consistent, and aligned with the requirements of the analysis.

```
...    customerNumber                customerName contactLastName  \
    0             103           Atelier graphique         Schmitt
    1             112          Signal Gift Stores            King
    2             114  Australian Collectors, Co.        Ferguson
    3             119           La Rochelle Gifts         Labrune
    4             121          Baane Mini Imports       Bergulfsen

  contactFirstName         phone                 addressLine1 addressLine2  \
    0         Carine    40.32.2555               54, rue Royale         None
    1           Jean    7025551838              8489 Strong St.         None
    2          Peter  03 9520 4555             636 St Kilda Road      Level 3
    3         Janine    40.67.8555  67, rue des Cinquante Otages         None
    4          Jonas    07-98 9555         Erling Skakkes gate 78         None
```

Figure 3 Sample of final dataset

Chapter 3: Descriptive Analytics

Descriptive analytics is the important in understanding the current state of a business by providing a detailed view of past performance, customer behaviours, product trends, and operational efficiencies. For Classic Models, this analysis offers insights into its financial health, customer engagement, and operational strengths and weaknesses. By examining key metrics such as total revenue, average customer spending, product performance, sales trends, purchase behavior, and employee contributions, descriptive analysis establishes an g of how the company has performed over time. The insights generated at this stage set the stage for predictive and prescriptive analytics by uncovering patterns and trends that inform future decision-making.

Classic Models is assumed to operates in competitive market that thrives on understanding customer preferences, optimizing product offerings, and maintaining operational efficiency. To meet these demands, the descriptive analytics phase is structured around answering key business questions that help in understanding the total revenue generated by the company, assessing how customers engage with the business in terms of spending behavior and purchase frequency, evaluating the performance of individual products and product lines, and identifying seasonal or temporal sales trends. It also includes evaluating the contributions of employees and offices to revenue generation, which helps to identify top performers and areas for improvement.

To achieve these objectives, data from various interconnected tables within the Classic Models relational database is integrated, cleaned, and analyzed. The prepared dataset, consisting of tables such as orders, orderdetails, products, customers, and employees, forms the backbone of this analysis. Each table provides unique insights into different aspects of the business, which are synthesized to provide a comprehensive view of performance.

3.1 Total Revenue and Customer Spending Behaviour

Revenue is the cornerstone of any business's financial assessment, serving as a key indicator of operational scale, market presence, and profitability. For Classic Models, analyzing total revenue provides not only a snapshot of overall financial health but also serves as a foundation for deeper insights into customer behavior and product performance. The calculation of total revenue in this analysis leveraged the orderdetails table, which captures transactional details for every line item in customer orders. Each order line specifies the quantity of products purchased and their corresponding unit price. By multiplying these values for each transaction and aggregating them across the dataset, we derived the company's total revenue during the analysis period.

The results were impressive, with a total revenue of **$9,604,190.61**, underscoring Classic Models' strong market presence and operational scale. This significant revenue figure highlights the company's success in catering to a wide range of customers and delivering high-quality products. However, analyzing total revenue alone is insufficient to provide actionable business insights. It is crucial to dissect how this revenue is distributed across customers, as it helps identify patterns of customer engagement and spending behavior.

To achieve this, the analysis grouped revenue by individual customers and calculated the average spend per customer. This approach offered insights into customer engagement, revealing whether the company's revenue was predominantly driven by a small group of high-value customers or a larger, more evenly distributed customer base. The variability in customer spending was notable, with some customers contributing significantly more than others. For instance, high-value customers, such as corporate clients or loyal collectors, consistently made large purchases and accounted for a disproportionate share of the total revenue. This group represents a critical segment for Classic Models, requiring targeted retention and engagement strategies to maintain their loyalty.

Conversely, the analysis identified a segment of low-value customers whose spending patterns were sporadic and minimal. Understanding the barriers faced by this group—whether due to pricing, product availability, or customer awareness—presents opportunities for strategic growth. For example, offering tiered pricing models, discounts, or entry-level products could incentivize these customers to increase their spending. Additionally, personalized marketing campaigns tailored to their preferences and purchasing history could boost their engagement.

The analysis also examined the top-performing products contributing to revenue. As shown in the **Top 5 Best-Selling Products** table, the **1992 Ferrari 360 Spider Red** emerged as the leading product, generating revenue of **$276,839.98** with a total quantity of 1,808 units sold. Other high-performing products included the **1937 Lincoln Berline** and **American Airlines: MD-11S**, each generating substantial revenue and reflecting strong customer demand for classic car replicas and vintage-themed items. This data highlights the importance of product-level analysis in identifying customer preferences and aligning inventory with high-demand items.

Additionally, understanding the temporal patterns of revenue generation adds another layer of insight. The **Monthly Sales Trends** visualization illustrates seasonal spikes in revenue, particularly during the months leading up to the holiday season. Peaks observed in October 2003 and November 2004 align with typical consumer behavior, where demand for collectible items increases during gift-giving periods. These insights allow Classic Models to optimize inventory levels, align marketing campaigns with seasonal demand, and ensure that high-demand products are available during critical periods.

The findings emphasize the importance of focusing on both high-value customers and top-performing products. By tailoring loyalty programs, exclusive offers, and personalized communication to high-value customers, Classic Models can further solidify their loyalty and boost repeat purchases. Meanwhile, leveraging the popularity of best-selling products in promotional campaigns and ensuring their availability during peak demand periods can drive incremental revenue.

```
Total Revenue: $9,604,190.61
                                          total_quantity  total_revenue
productName
1992 Ferrari 360 Spider red                         1808      276839.98
1937 Lincoln Berline                                1111      102563.52
American Airlines: MD-11S                           1085       71753.93
1941 Chevrolet Special Deluxe Cabriolet             1076      102537.45
1930 Buick Marquette Phaeton                        1074       41599.24
```

Figure 4 Total revenue output

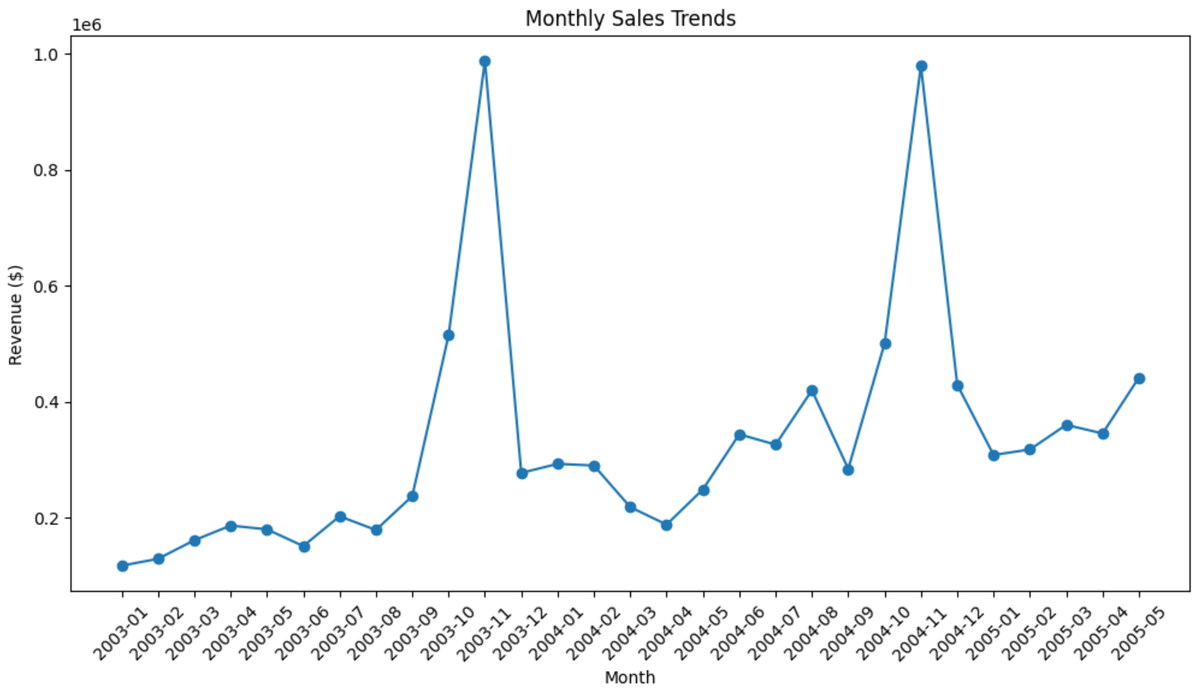


Figure 5 Customer spending habits monthly

3.2 Product Performance

Product performance is a critical metric for understanding revenue generation and customer engagement. By analyzing the performance of individual products, Classic Models can pinpoint which items resonate most with their customers and which require additional focus or improvement. This evaluation provides insights into customer preferences, informs inventory management, and drives targeted marketing strategies. For this analysis, two key metrics were considered: the total quantity of products sold, and the total revenue generated by each product. While total quantity reflects the popularity of a product in terms of units sold, total revenue provides a financial perspective, distinguishing between high-volume, low-revenue items and low-volume, high-revenue products.

The analysis used data from the orderdetails and products tables. By grouping the data by product names and aggregating the sales and revenue figures, the products were ranked to identify top performers. The results revealed significant patterns. For instance, the 1992 Ferrari 360 Spider Red stood out as the top-selling product with 1,808 units sold, generating an impressive $276,839.98 in revenue. Other top products, such as the 1937 Lincoln Berline and American Airlines MD-11S, also demonstrated strong sales and revenue, highlighting their popularity among collectors and enthusiasts. These findings indicate that moderately priced products often dominate in terms of volume, while premium products can achieve substantial revenue despite lower sales volumes.

The analysis also examined product performance at the category level, grouping products by their respective productLine. This broader view highlighted trends in customer preferences. Categories like "Classic Cars" and "Motorcycles" consistently performed well in both total sales volume and revenue, underscoring their strong demand among customers. Conversely, categories such as "Trucks and Buses" exhibited lower sales, pointing to potential opportunities for improvement. These insights are vital for inventory management, ensuring that high-demand categories are adequately stocked to meet customer needs while avoiding overstocking of slower-moving items.

Understanding product performance not only aids inventory optimization but also helps in designing effective marketing strategies. High-performing products, for example, can be leveraged in targeted promotional campaigns or bundled with less popular items to drive overall sales. On the other hand, slow-moving products may require additional marketing efforts, discounts, or repositioning to improve their visibility and appeal. For instance, offering promotional bundles that include a popular product like the 1992 Ferrari 360 Spider Red alongside a less popular item could increase the overall sales of both.

Visualizing the results further enhances the interpretation of product performance. The Top 10 Best-Selling Products by Quantity Sold chart clearly demonstrates the dominance of certain products in terms of units sold, with the 1992 Ferrari 360 Spider Red leading the pack. Similarly, the Top 10 Products by Revenue chart offers a financial perspective, emphasizing the revenue contributions of premium products like the 1956 Porsche 356A Coupe, which generated $134,240.71 in revenue. These visualizations provide actionable insights, enabling Classic Models to focus their efforts on sustaining the performance of top products and addressing gaps in underperforming ones.

Additionally, the insights derived from product performance analysis can inform pricing strategies. Products generating substantial revenue despite lower sales volumes, such as premium collectibles, could be evaluated for potential pricing optimization to maximize profit margins. Similarly, products with high sales volume but relatively low revenue may benefit from value-added features or slight price adjustments to increase profitability.

```
Top 10 Best-Selling Products:
                                          total_quantity  total_revenue
productName
1992 Ferrari 360 Spider red                         1808      276839.98
1937 Lincoln Berline                                1111      102563.52
American Airlines: MD-11S                           1085       71753.93
1941 Chevrolet Special Deluxe Cabriolet             1076      102537.45
1930 Buick Marquette Phaeton                        1074       41599.24
1940s Ford truck                                    1061      114232.79
1969 Harley Davidson Ultimate Chopper               1057       90157.77
1957 Chevy Pickup                                   1056      109946.21
1964 Mercedes Tour Bus                              1053      117669.66
1956 Porsche 356A Coupe                             1052      134240.71
```

Figure 6 Product performance output

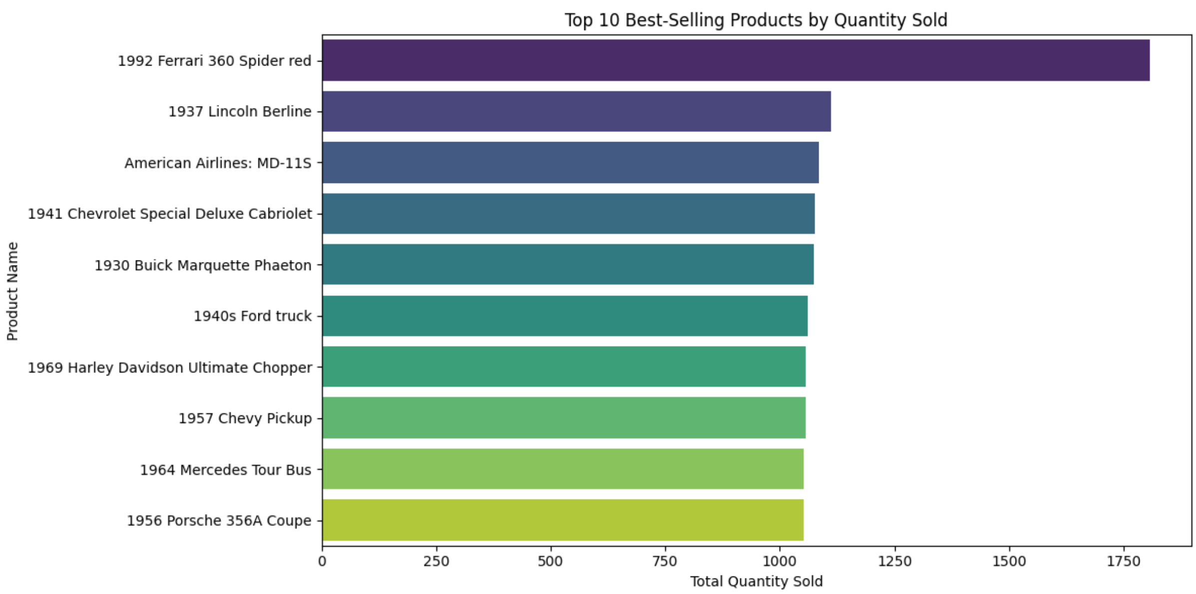


Figure 7 Bar chart of product performance

3.3 Sales Trends Over Time

By analyzing sales trends, Classic Models gains insights into seasonal patterns, peak periods, and potential areas of concern. This allows for strategic decisions regarding inventory management, promotional activities, and resource allocation. The analysis of sales trends used data from the orders and orderdetails tables, which record the dates of order placements and associated revenue details. By aggregating revenue by month, quarter, and across years, a comprehensive understanding of sales performance over time was achieved.

3.3.1 Monthly Sales Trends

The first phase of the analysis focused on monthly sales trends. By grouping revenue by month and plotting the results, distinct seasonal patterns emerged. For instance, sales exhibited significant spikes in **October 2003** and **November 2004**, as evident in the **Monthly Sales Trends** chart. These peaks align with typical consumer behavior, where demand increases during holiday seasons and special events, such as Christmas and end-of-year celebrations. The elevated revenue during these months reflects the heightened customer interest in collectible items as gifts, reinforcing the importance of aligning inventory and marketing efforts with these periods.

Conversely, certain months showed dips in revenue, indicative of periods with reduced customer activity. For example, the months following the holiday season, such as January and February, typically experienced slower sales. This information is invaluable for resource planning, as it highlights opportunities for promotional campaigns or discounts to boost sales during these low-demand periods.

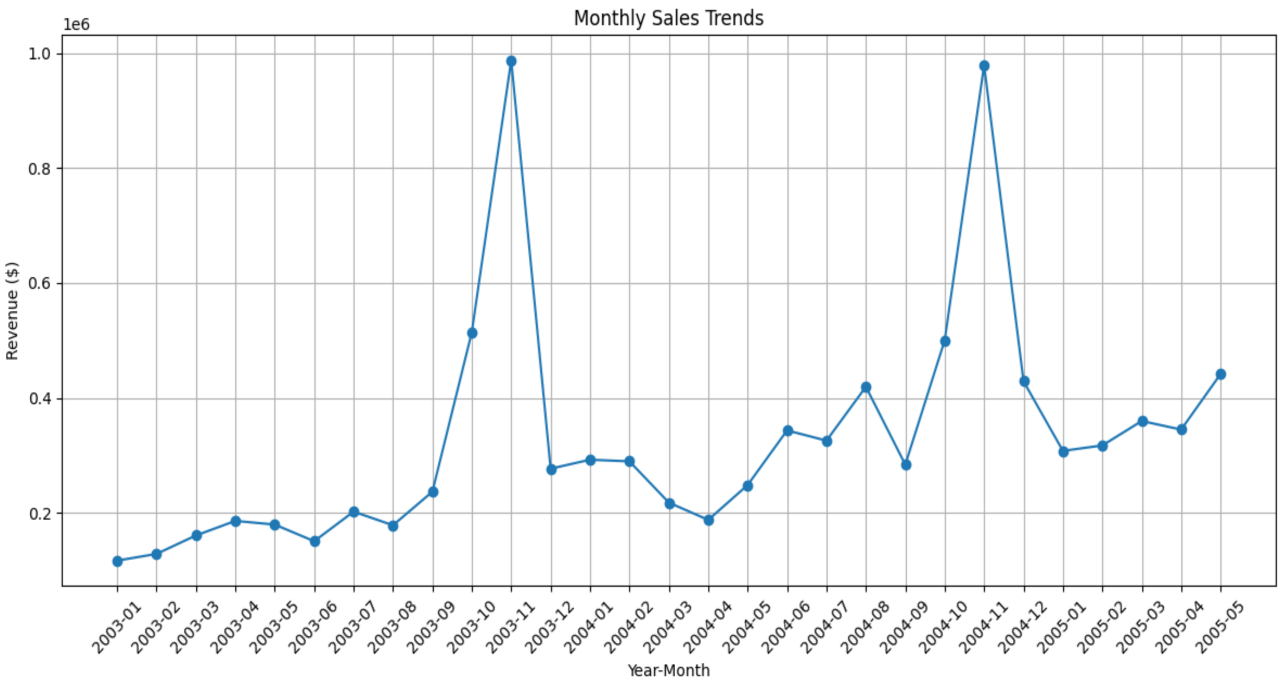


Figure 8 Monthly Sales Trends: Identifying seasonal peaks and off-peak periods for strategic planning.

3.3.2 Quarterly Sales Trends

To gain a broader perspective, revenue was also analyzed on a quarterly basis. The **Quarterly Sales Trends** revealed that the fourth quarter consistently outperformed other quarters, with **Q4 2003** and **Q4 2004** standing out as the highest-grossing periods. This underscores the importance of the holiday season in driving revenue and customer engagement. Such trends provide actionable insights for strategic planning, enabling the company to allocate resources effectively and ensure high-demand products are well-stocked during these critical periods.

Quarterly analysis also highlighted periods of steady revenue growth, such as the transitions from Q2 to Q3 in 2003 and 2004, reflecting sustained customer interest and engagement. These insights allow Classic Models to align their operational strategy with long-term trends and maintain their competitive edge.

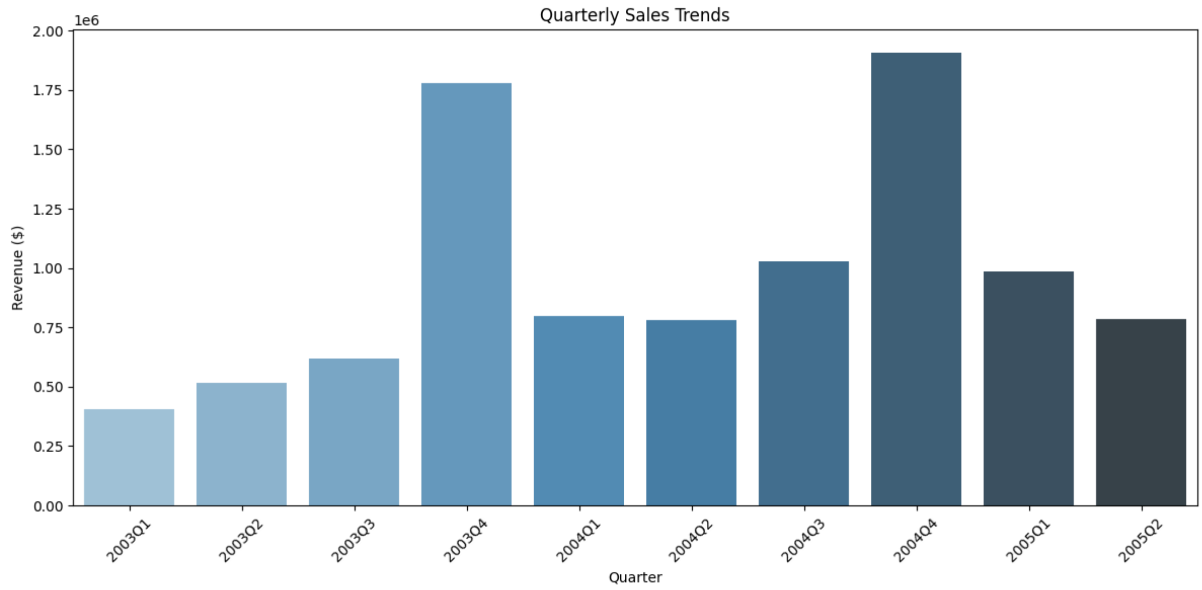


Figure 9 Quarterly Sales Trends: Demonstrating the importance of Q4 in revenue generation

3.3.3 Seasonal Trends

Seasonal trends were further explored by analyzing the average revenue for each calendar month across the dataset. The **Seasonal Trends: Average Monthly Revenue** chart revealed that months like **April** and **May** consistently outperformed others, while months like **January** and **August** showed relatively lower performance. These findings highlight the cyclical nature of customer demand, enabling Classic Models to anticipate sales patterns and prepare accordingly.

For example, increased sales in spring months such as April may reflect customer preferences for new purchases during this period, while slower months such as August may indicate opportunities for targeted promotional campaigns. By understanding these patterns, the company can refine its marketing strategies to maximize customer engagement and sales during slower months.

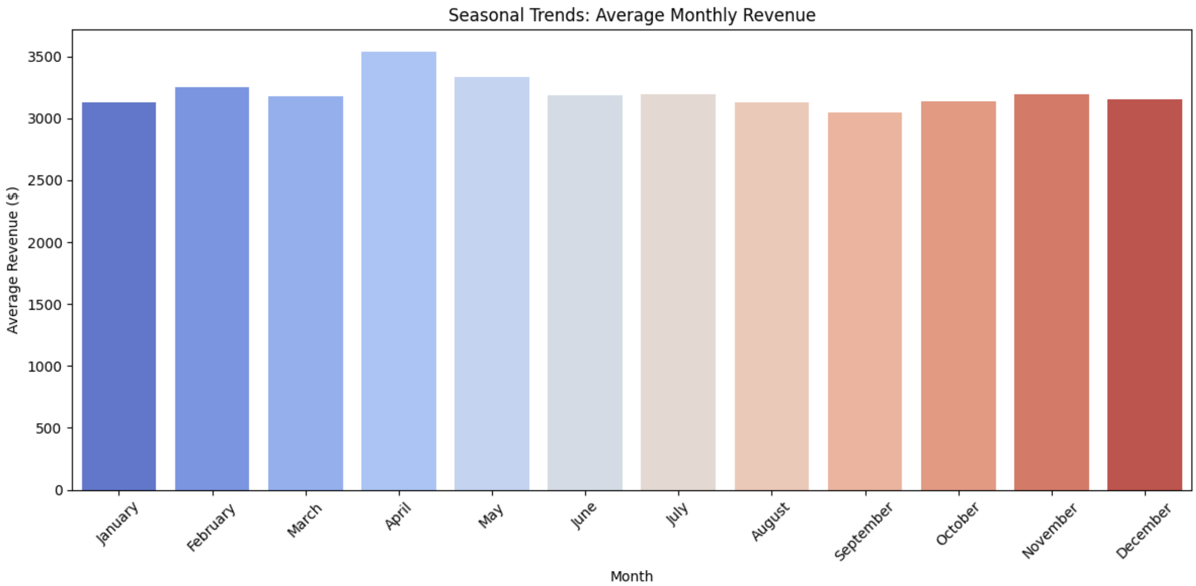


Figure 10 Seasonal Trends: Highlighting average monthly revenue to uncover patterns in customer demand

The findings from these time-based analyses are critical for both short-term and long-term planning. During high-demand periods, such as Q4 or holiday seasons, Classic Models must ensure adequate inventory levels to prevent stockouts and capitalize on heightened customer interest. Effective staffing and logistical planning are also essential during these periods to ensure seamless operations. Conversely, during low-demand periods, strategies such as promotional offers, targeted discounts, or bundled deals can be employed to stimulate customer activity and drive revenue.

In addition to seasonal trends, the analysis uncovered broader long-term growth patterns. The gradual increase in monthly and quarterly revenue over the analysis period indicates sustained growth and expanding customer engagement. This reinforces confidence in the company's strategic direction and highlights opportunities for further investment in high-performing areas.

By integrating these insights into their operational strategy, Classic Models can better align their resources with customer demand, enhance customer satisfaction, and sustain profitability. The clear identification of sales patterns enables the company to stay ahead of market trends and remain competitive in the ever-evolving collectibles market.

.

3.4 Customer Purchase Behavior

Customer purchase behavior plays a crucial role in understanding engagement, loyalty, and the overall health of Classic Models' customer base. Through analyzing purchase frequency and spending patterns, the company can segment its customers more effectively and devise targeted strategies to maximize retention and growth. This analysis utilized data from the orders and order details tables, focusing on metrics such as the number of unique orders placed by each customer, their total spending, and the recency of their purchases.

The analysis revealed significant variability in customer engagement. While most customers placed one to three orders, a smaller yet highly influential segment exhibited frequent purchases, with some customers making over 20 transactions during the analysis period. These high-frequency buyers not only contribute significantly to total revenue but also represent a loyal customer segment with high retention potential. These insights highlight the importance of prioritizing frequent buyers for personalized marketing strategies, loyalty programs, and other engagement initiatives.

The spending behavior analysis revealed a similar range of variability. The average spending per customer was approximately $98,000, with the top 10% of customers contributing disproportionately to total revenue. These high-value customers had an average spend of $268,149, emphasizing their critical role in driving business profitability. This subset of customers should be the focus of strategic retention efforts, such as exclusive offers, personalized communication, and premium services.

Understanding recency, which measures the time elapsed since a customer's last purchase, is another critical aspect of analyzing customer behavior. The analysis indicated that customers typically returned within an average of 183 days. However, there were substantial deviations, with some customers making purchases as recently as the analysis cutoff date, while others had not purchased for over 500 days. The histogram of recency distribution underscored distinct peaks, likely representing seasonal or periodic buying patterns.

These findings provide actionable insights into the lifecycle stages of Classic Models' customers. One-time buyers are likely in the acquisition stage, requiring further incentives to transition to repeat buyers. Conversely, high-frequency, high-spending customers represent the retention stage and necessitate strategies that reinforce their loyalty, such as early access to new products, targeted promotions, or personalized communications. Additionally, identifying customers with declining purchase recency allows the company to intervene with re-engagement strategies before they churn.

Visualizations further elucidated these patterns. The histogram of purchase frequencies revealed a right-skewed distribution, emphasizing the prevalence of low-frequency buyers. Similarly, the spending distribution highlighted a clustering around moderate spending levels, with a few high outliers corresponding to the company's most valuable customers. Finally, the boxplot of spending demonstrated the presence of significant outliers, reaffirming the impact of top-tier customers on overall revenue.

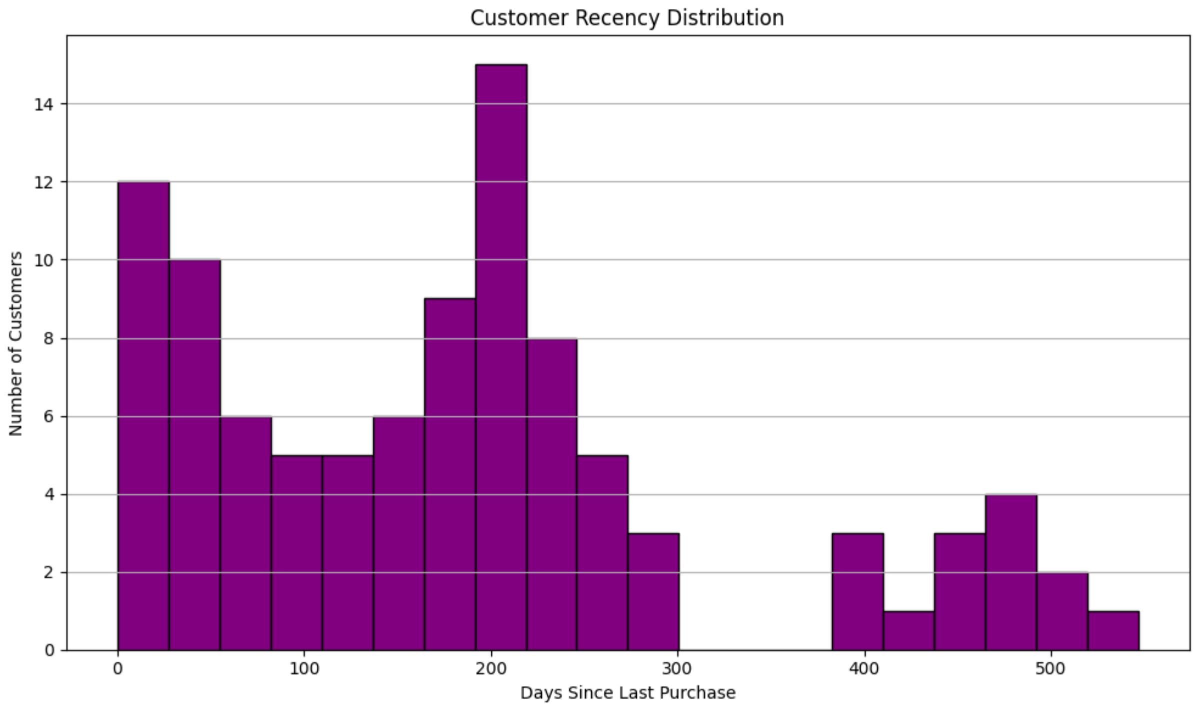


Figure 11 Customer recency

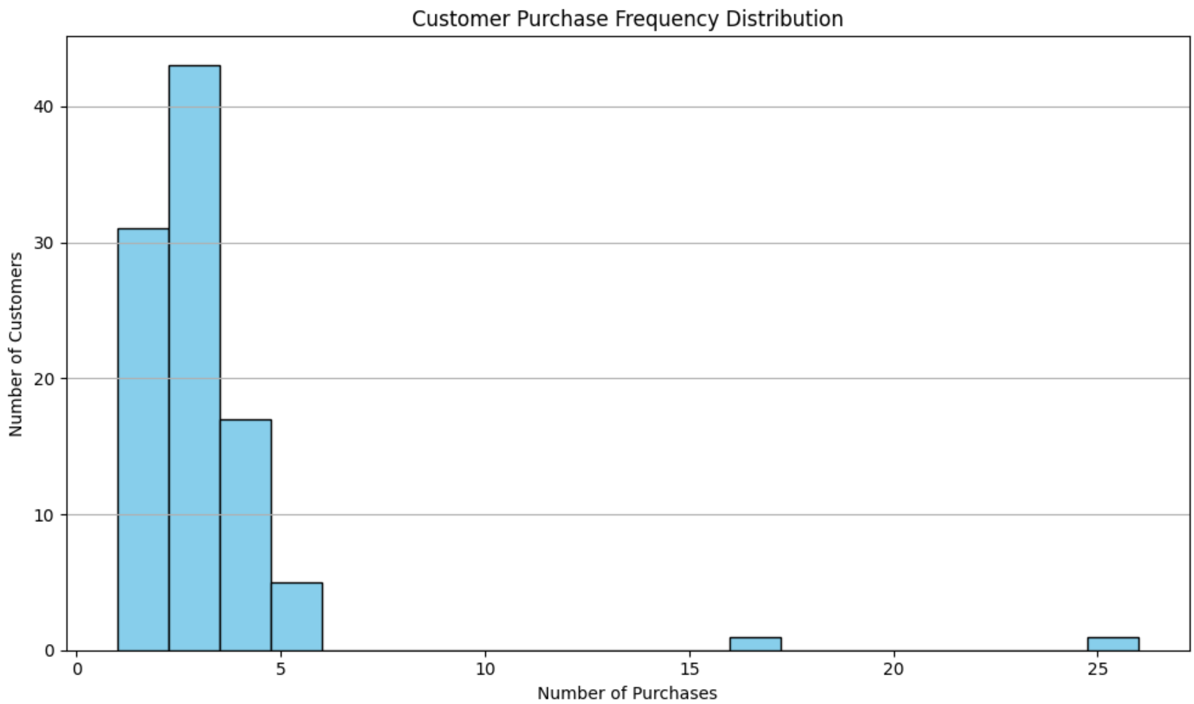


Figure 12 Customer Purchase Frequency table

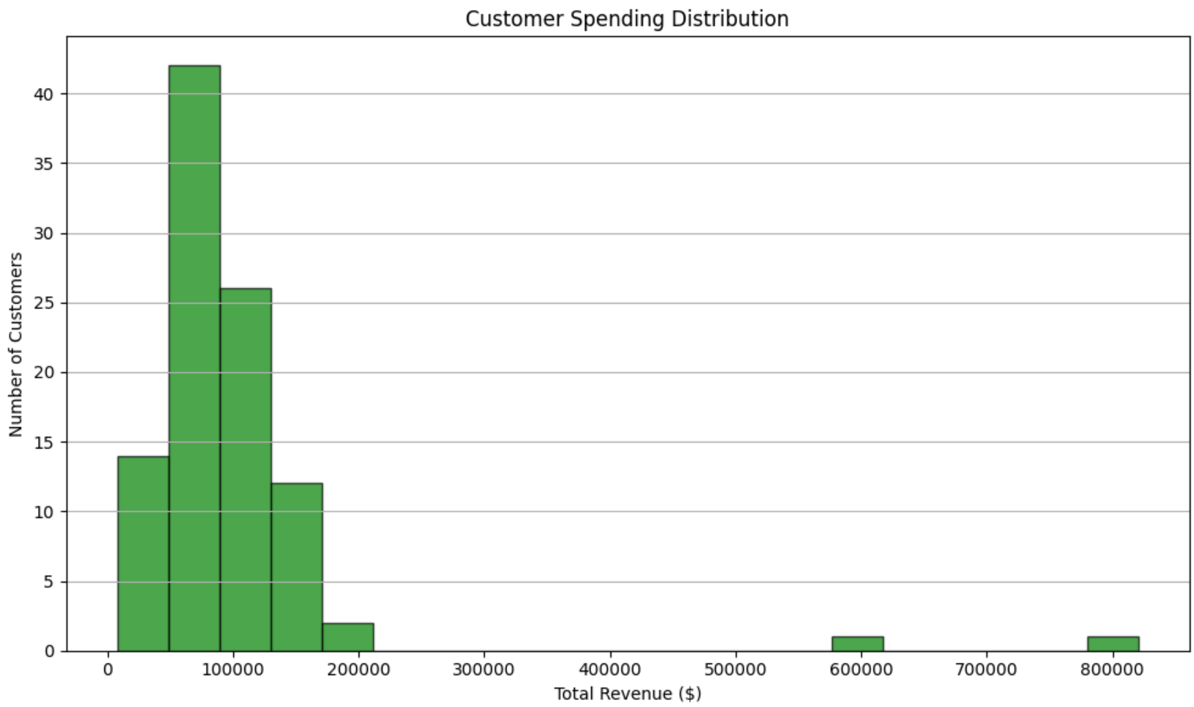


Figure 13 Cusytomer spending distribution

Chapter 4: Predictive Analytics

4.1 Demand Prediction for Products

Demand forecasting is a vital component of inventory management and strategic planning at Classic Models. Accurate demand predictions enable the company to meet customer needs effectively, avoiding the costs associated with overstocking or shortages. Initially, traditional regression methods, including Linear Regression and Random Forest, were implemented to predict demand for the 1992 Ferrari 360 Spider Red, a top-performing product. While these approaches captured some aspects of the demand trend, their overall performance was unsatisfactory. The evaluation metrics, such as Mean Squared Error (MSE) and R-squared (R^2), highlighted significant limitations. For instance, the regression models struggled to account for seasonal variations and non-linear demand patterns, which are characteristic of sales data for collectible items. These limitations necessitated a shift to a more robust forecasting method. After careful consideration, Facebook Prophet was chosen for its ability to handle time-series data with built-in support for seasonality, trends, and uncertainty.

The decision to transition from regression models to Prophet was driven by several factors. First, the poor R^2 values from the regression analysis (-0.45) indicated that the models failed to explain the variability in demand effectively. Second, the inability of linear regression and Random Forest to capture complex seasonality limited their utility for accurate forecasting. These models operate best when the relationship between features and the target variable is either linear or straightforwardly hierarchical, making them unsuitable for the nuanced patterns in sales data. Prophet, on the other hand, is specifically designed to address these challenges. It provides a modular framework that decomposes time-series data into trend, seasonal, and residual components, making it highly effective for capturing recurring patterns and long-term trends.

To implement Prophet, the historical monthly sales data for the Ferrari 360 Spider Red was aggregated, with each month's total quantity ordered forming the dependent variable. This data was formatted into Prophet's required structure, with a ds column for dates and a y column for observed values. Prophet's ability to automatically detect and model seasonality made it particularly suitable for this dataset, where demand tends to spike during holiday seasons and special occasions. Once the data was prepared, the model was trained on historical trends, and a future data frame was created to generate predictions for the next six months.

Prophet's forecast yielded valuable insights into future demand. The model identified a consistent upward trend in sales, indicative of growing customer interest in the product. Seasonal components revealed spikes in demand during specific months, such as October, reflecting holiday-driven purchasing behavior. The forecast included confidence intervals, providing a range of expected values for each prediction. For instance, the demand forecast for October was 271 units, with lower and upper bounds of 253 and 291 units, respectively. These intervals allowed Classic Models to plan for variability, ensuring adequate inventory to meet demand while avoiding overstocking.

Visualization of the results further underscored the advantages of Prophet. The decomposition plot showed how the model separated the long-term trend from seasonal variations, offering a clear understanding of demand drivers. The forecast plot compared historical data with predicted values, demonstrating the model's accuracy in capturing sales patterns. These visualizations not only validated the model's performance but also provided actionable insights for inventory planning and marketing strategies.

The transition to Prophet also highlighted important lessons for demand forecasting. One key advantage was its ability to incorporate uncertainty, represented by confidence intervals, which regression models lacked. Moreover, Prophet's modular approach enabled the company to focus on specific components, such as trend or seasonality, for more granular analysis. However, the model's reliance on historical data means that its accuracy depends on the availability and quality of past sales records. External factors, such as economic changes or competitive actions, remain unaccounted for, suggesting potential areas for future.

Prophet's ability to model trends, seasonality, and uncertainty provided deeper insights into future demand, enabling the company to align production schedules, inventory levels, and marketing efforts with anticipated customer needs. This transition not only addressed the shortcomings of regression models but also empowered Classic Models with a predictive framework that is both scalable and adaptable to other products. By leveraging Prophet across its portfolio, the company can achieve more precise forecasts, optimize inventory management, and maintain a competitive edge in the market.

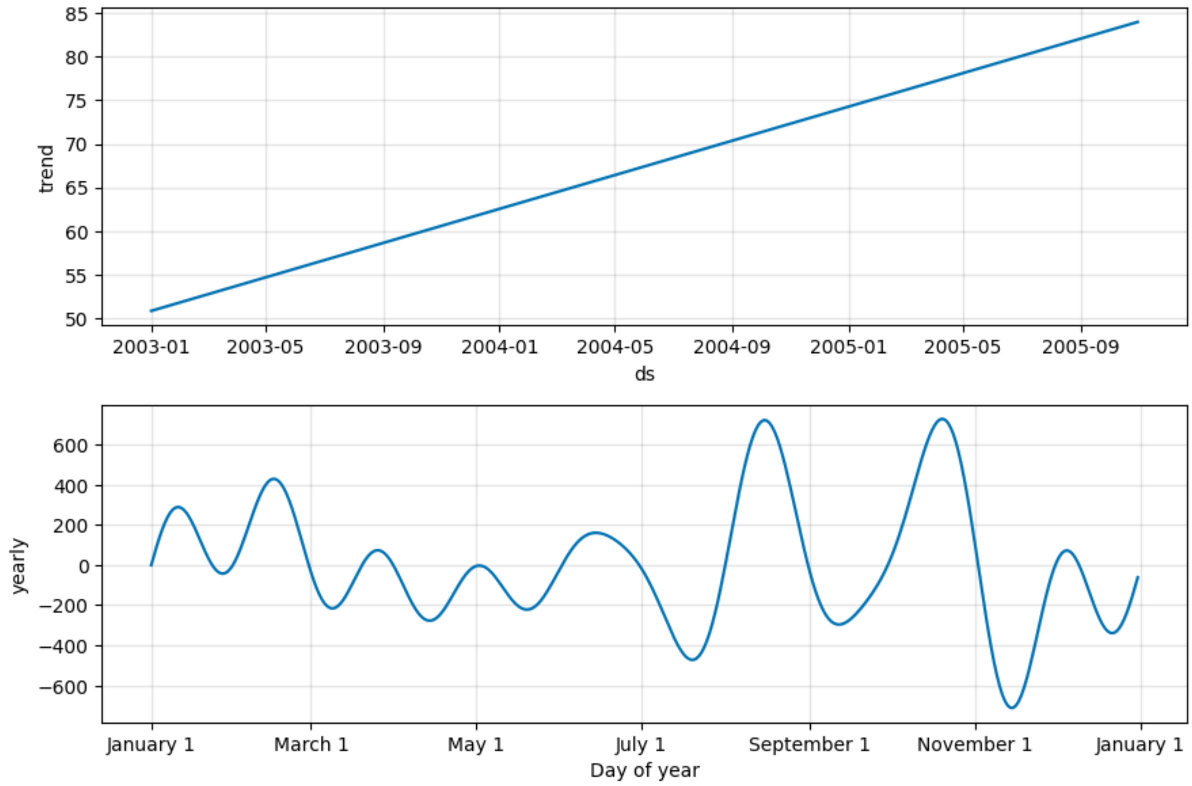


Figure 14 Future demand forecast graph

```
Future Demand Forecast:
          Date  Forecasted Quantity  Lower Bound  Upper Bound
28 2005-05-31            40.505249    22.935494    59.219453
29 2005-06-30            80.785983    61.898796   100.085816
30 2005-07-31            25.562055     7.811368    44.962347
31 2005-08-31           100.239300    81.394741   118.247764
32 2005-09-30            89.761060    72.335569   108.399710
33 2005-10-31           271.731366   252.968497   290.658341
```

Figure 15 Prophet output on future demand forecast

4.2 Pricing Optimization

Pricing is a critical factor in maximizing revenue, balancing profitability, and maintaining customer satisfaction. For Classic Models, pricing optimization focuses on understanding the intricate relationship between product prices, quantities sold, and total revenue. By leveraging historical sales data and advanced modeling techniques, this analysis provides actionable insights into setting optimal price points for the company's products, ensuring maximum revenue and profitability.

The analysis began by extracting data from the orderdetails and products tables to calculate key metrics, including the average price, average quantity sold, total revenue, and profit margins for each product. This aggregated data provided a foundation for examining the relationship between price and demand. By applying statistical methods and polynomial regression models, the analysis captured the non-linear dynamics of pricing, enabling the identification of optimal price points where revenue and demand align most effectively.

One of the key findings of this analysis was the determination of the optimal price point for a high-performing product. For instance, the analysis revealed that setting the price at **$136.44** would maximize sales quantity for a top-selling product. This price point represents a balance between affordability for customers and profitability for the company, enabling higher sales volumes while maintaining sustainable revenue per unit.

The relationship between price and demand was further explored through detailed visualizations. A scatter plot of average price vs. average quantity ordered illustrated the non-linear trend, showing that demand peaks within a specific price range before declining as prices increase. The polynomial regression model applied to this data identified the precise turning point where higher prices begin to negatively impact demand. Similarly, an analysis of price vs. total revenue showed that revenue increases steadily with price until reaching a plateau, emphasizing that excessive price increases can lead to diminished returns.

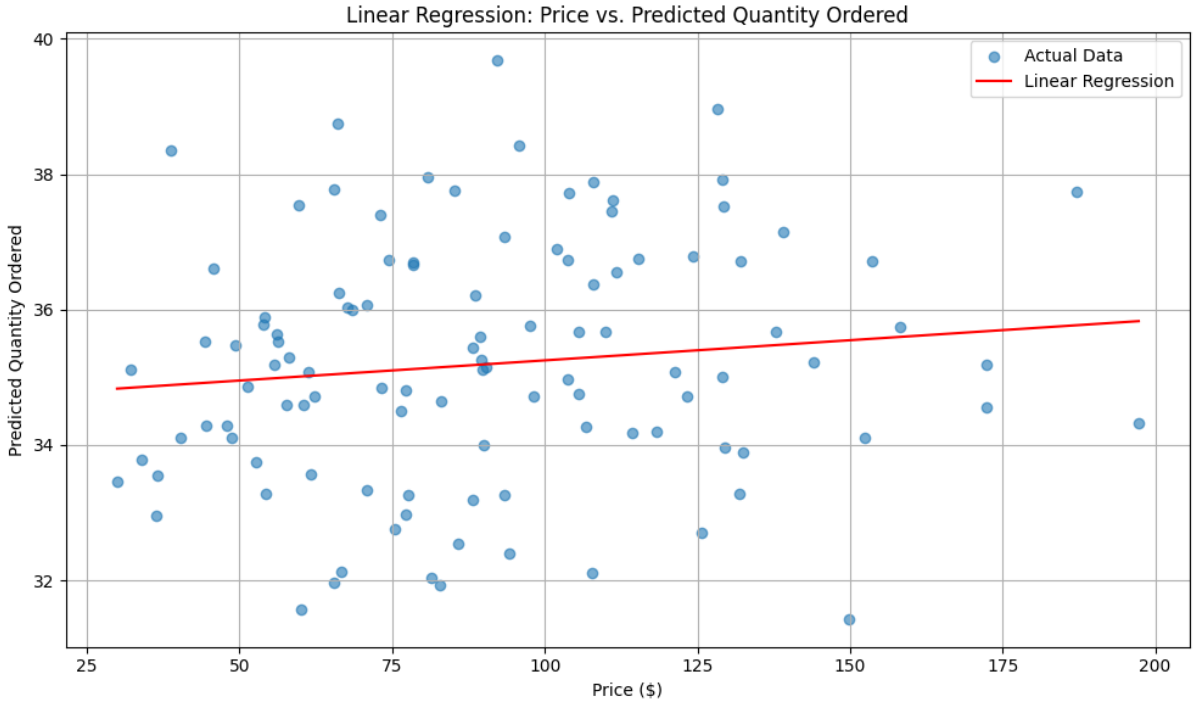
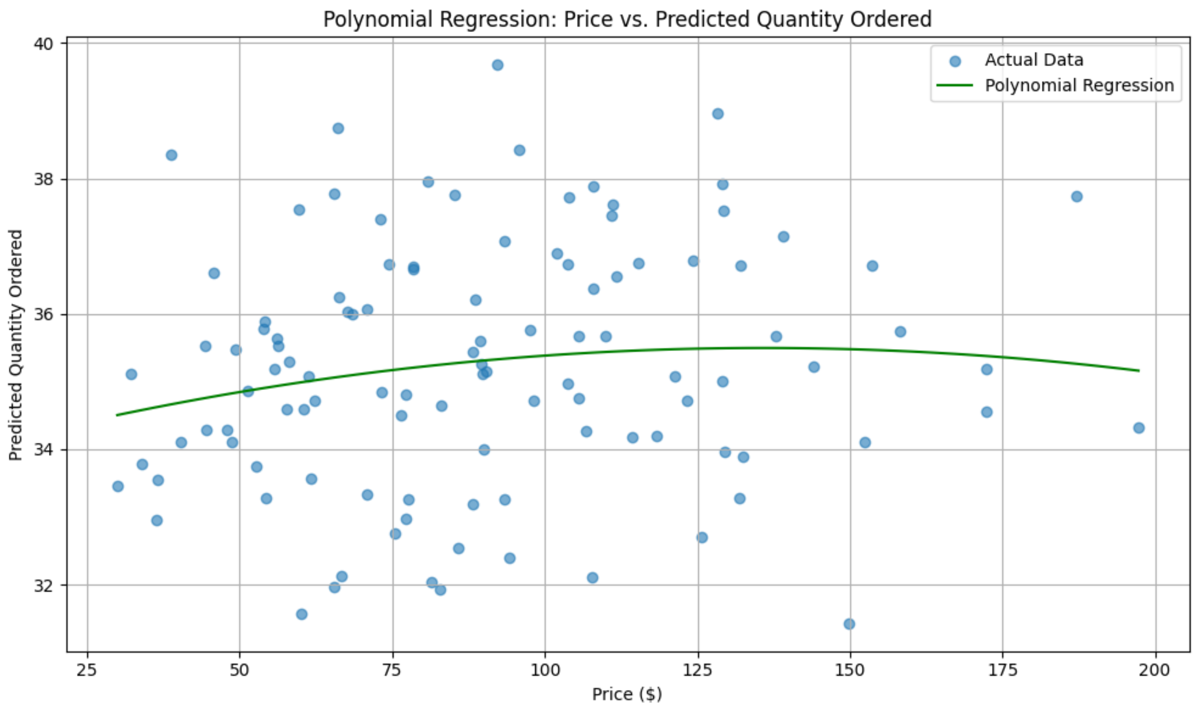


Figure 16 Price vs quantity ordered



In addition to revenue and demand, profit margins were analyzed to ensure pricing strategies are financially sustainable. Products with high profit margins and moderate price points emerged as key drivers of profitability. By identifying such products, the company can focus on these items in its marketing and inventory strategies, ensuring a robust contribution to overall revenue while maintaining competitive pricing.

Figure 17 Price vs revenue visualization

Visualizations played a pivotal role in communicating the findings. The scatter plot of price vs. revenue highlighted the trend of increasing revenue with price, showcasing the plateau at the optimal price point. Similarly, the scatter plot of price vs. quantity ordered, overlaid with the polynomial regression curve, provided a clear representation of the demand dynamics. Additionally, a detailed table summarized the pricing optimization results, listing key metrics for multiple products, such as average price, quantity sold, total revenue, and profit margins.

This pricing optimization analysis has several business implications for Classic Models. Identifying the optimal price points for products enables the company to better align its pricing strategy with customer behavior and market demand. Products with upward trends in revenue and profitability should be prioritized for production and marketing, while items with declining demand can benefit from targeted promotions or bundling strategies. Moreover, understanding the profit margins associated with different products allows for more informed decision-making when allocating resources and setting promotional budgets.

```
...   Pricing Optimization Summary:
        productCode   avg_price  avg_quantity  total_revenue  profit_margin  \
    0      S10_1678   85.174286     37.750000       90157.77      36.364286
    1      S10_1949  197.309286     34.321429      190017.96      98.729286
    2      S10_2016  110.018929     35.678571      109998.82      41.028929
    3      S10_4698  172.288214     35.178571      170686.00      81.268214
    4      S10_4757  124.245714     36.785714      127924.32      38.565714
    ..          ...         ...           ...            ...            ...
    104   S700_3505   89.522222     35.259259       84992.25      38.432222
    105   S700_3962   88.202222     33.185185       78919.06      34.572222
    106   S700_4002   65.967857     38.750000       71753.93      29.697857
    107    S72_1253   44.427143     34.285714       42692.53      11.657143
    108    S72_3212   49.402593     35.481481       47550.40      16.102593
```

Figure 18 Pricing optimization summary

Chapter 5: Topic Modelling using LDA

The objective of this analysis was to identify and interpret the key themes within customer comments using Latent Dirichlet Allocation (LDA). By uncovering hidden topics, Classic Models can gain actionable insights into customer sentiment, preferences, and recurring issues. This understanding enables targeted strategies to improve service quality, product offerings, and overall customer satisfaction.

To achieve this, a systematic process was followed. The comments data underwent preprocessing, which included stop word removal, lemmatization to reduce words to their base forms, and tokenization for text segmentation. Rare words appearing in fewer than five comments were also filtered out to enhance the relevance of the topics identified. The LDA algorithm was applied to this preprocessed data, with the number of topics varying between three and six to determine the most coherent and interpretable set. Ultimately, five topics were selected based on their alignment with observed patterns in the comments.

The analysis uncovered five key themes. The first theme focused on **delayed deliveries**, with customers expressing concerns about late shipments and emphasizing the need for timely order fulfillment. Keywords such as "caution," "cancel," and "order" highlighted their frustrations. This points to a need for improved logistics and proactive communication regarding delays. The second theme, **product quality issues**, captured feedback about defective or mismatched items. Keywords like "mismatch," "risk," and "concern" suggested dissatisfaction with product standards, indicating an opportunity to enhance quality control processes.

The third theme revolved around **positive experiences**, where customers praised product quality and service. Keywords such as "marketing," "contact," and "sale" reflected appreciation for the company’s efforts, highlighting the importance of maintaining high standards. The fourth theme, **pricing and discounts**, featured discussions about pricing policies and promotional offers. Keywords like "renegotiate," "warehouse," and "material" suggested a need for transparent pricing strategies and targeted discount campaigns. Finally, the fifth theme focused on **customer support**, emphasizing feedback related to customer service interactions. Keywords like "agreement," "payment," and "received" revealed the importance of responsive and effective communication.

The themes were further visualized using word clouds, which provided a clear representation of the dominant keywords for each topic. These insights align with findings from earlier descriptive analyses, reinforcing the need for operational improvements in logistics, quality assurance, and customer support. By addressing these recurring themes, Classic Models can enhance customer satisfaction and foster loyalty.

```
Interpretation of Topics:
Topic 1: Delayed Deliveries - Customers discussing delays in receiving orders.
Topic 2: Product Quality Issues - Complaints or comments about defective products.
Topic 3: Positive Experiences - Praise for product quality or customer service.
Topic 4: Pricing and Discounts - Discussions related to pricing or discounts.
Topic 5: Customer Support - Issues or feedback regarding customer service interactions.
```

Figure 19 Interpretation of topics

Figure 20 Word cloud for topic 1

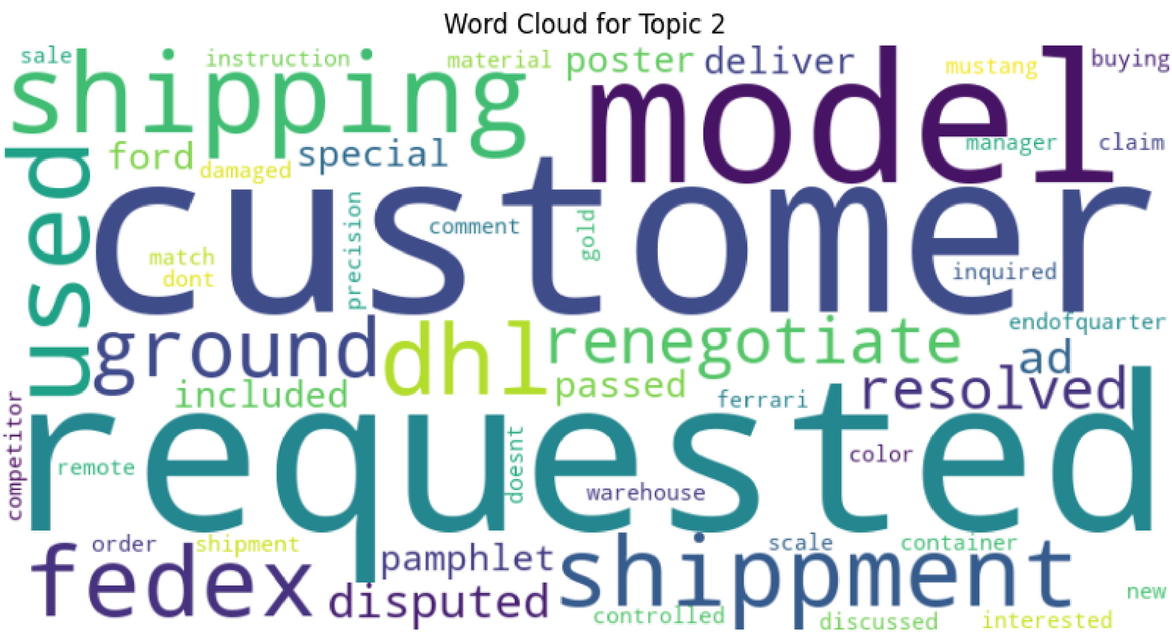


Figure 21 Word cloud for topic 2

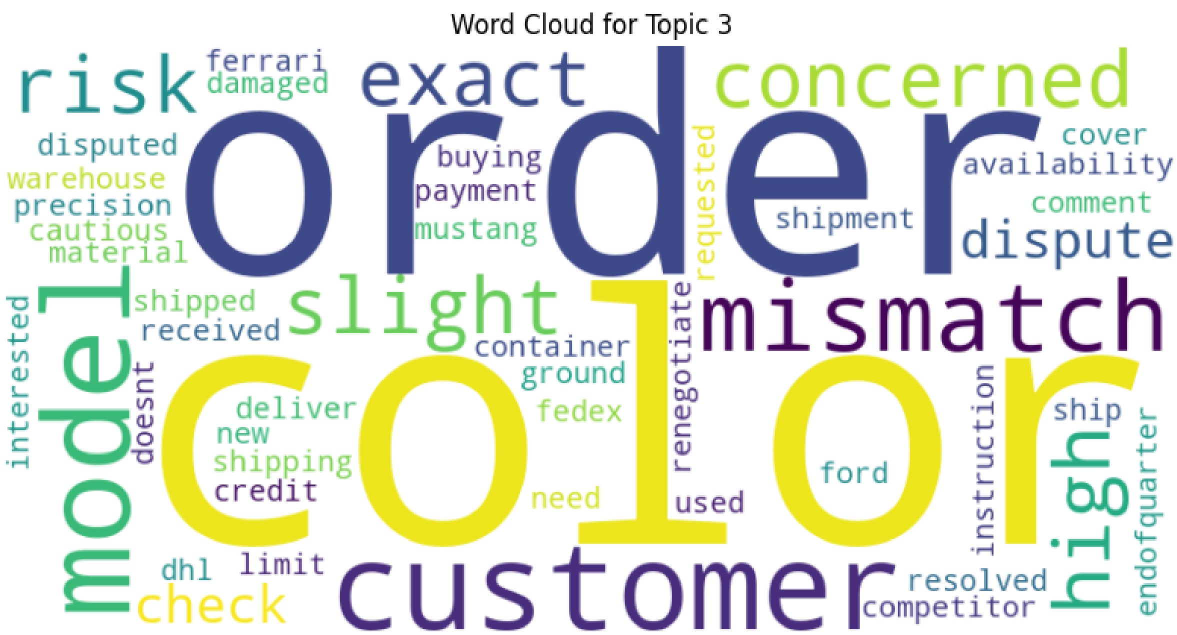


Figure 22 Word cloud for topic 3

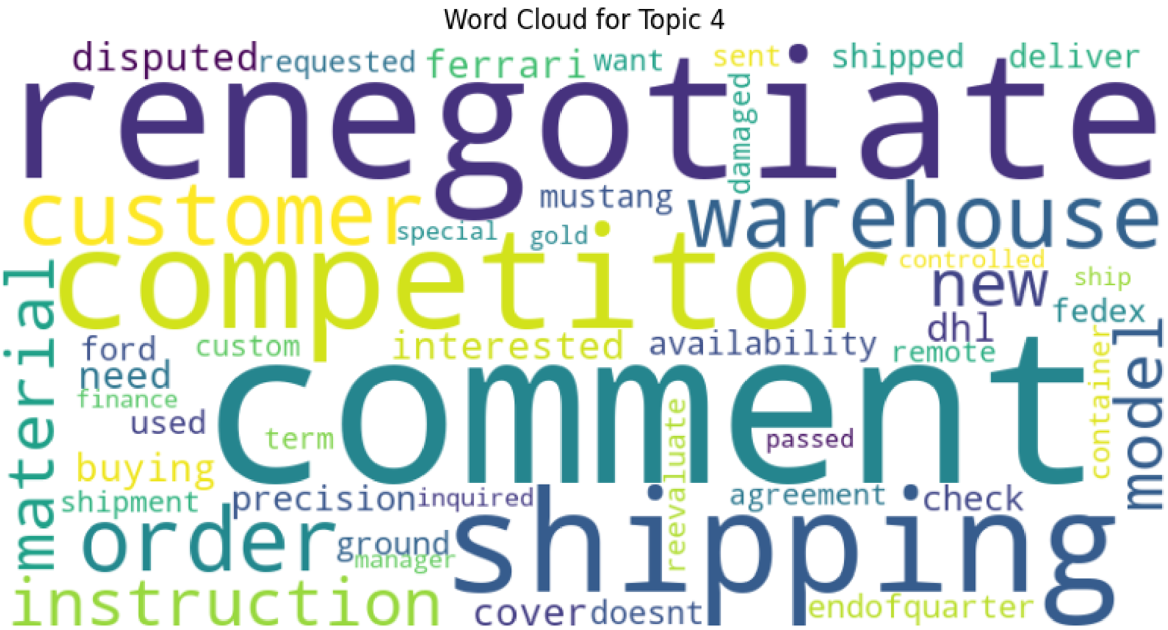


Figure 23 Word cloud for topic 4

Figure 24 Word cloud for topic 5