School: Computer Science

Academic year: 2023-24

Student Name	Student ID
Aishwarya Shahu	20607987

Category	Details
Title	Fundamentals of Information Visualisation
Module Code	COMP3021

Exploratory Data Analysis of Auto Sales Data

Abstract

This report presents an exploratory data analysis (EDA) of auto sales data, which details information on 18,619 auto sales orders over a three-year period (January 2018 to May 2020). The EDA aims to unravel patterns, trends, and relationships among sales, customer behaviour, and product characteristics. The findings provide valuable insights into the dynamics of auto sales, aiding in informed decision-making and strategic planning for the automotive industry.

1. Introduction

1.1. Background

Automobile sales data holds immense value in shaping business strategies and optimizing marketing efforts. By analysing sales patterns, customer preferences, and product performance, companies can gain a deeper understanding of market trends and identify opportunities for growth. This EDA delves into a comprehensive analysis of auto sales data, shedding light on key aspects that influence sales performance.

1.2. Objectives

The primary objectives of this EDA are to:

1. Uncover the distribution of sales across different product lines, countries, deal sizes, and time periods.
2. Identify the product lines with the highest average sales and the most volatile sales trends.
3. Assess how sales vary by country and the proportion of different deal sizes.
4. Analyse the evolution of sales over the three-year period and identify any significant trends or patterns.
5. Evaluate the correlation between sales and order date, customer country, deal size, and month.
6. Gain insights into customer behaviour and preferences based on sales data.
7. Identify potential areas for optimization and improvement in sales performance.

These objectives are essential for a comprehensive understanding of the auto sales landscape and inform decision-making processes in the industry.

2. Data Description

2.1.Dataset Overview

The data set for this exploratory data analysis (EDA) consists of seven variables:

- SALES: The total sales value of an order (numerical)
- ORDERDATE: The date the order was placed (date)
- STATUS: The status of the order (shipped, cancelled, disputed, on hold) (categorical)
- PRODUCTLINE: The type of product sold (classic cars, vintage cars, trucks-n-buses, motorcycles, trains, aircraft, ships) (categorical)
- COUNTRY: The country of the customer (USA, Spain, France, Australia, UK, Germany, Canada, Italy, Netherlands, Belgium, Switzerland, Portugal, Austria, Finland, Sweden, Norway, Denmark, Poland, Russia, China) (categorical)
- DEALSIZE: The size of the order (small, medium, large) (categorical)

2.2.Data Cleaning

- Data Selection: The data cleaning process begins with selecting the relevant variables for analysis.
- Data Conversion: The ORDERDATE variable is converted from a character string to a date format using the as. Date () function. This is necessary for performing date-based analysis, such as calculating the average sales per month.
- Data Standardization: The SALES variable is standardized using the scale() function to ensure that all variables are on the same scale. This can be helpful when comparing the relative importance of different variables.
- Outlier Detection and Removal: Outliers are extreme values that can skew the results of statistical analysis. In this case, outliers are identified using the IQR (interquartile range) method. Values that fall outside the interquartile range are considered outliers and are removed from the dataset.
- Data Imputation: Missing values are handled using the fillna() function. In this case, missing values in the SALES variable are imputed with the median of the SALES variable.
- Data Aggregation: The aggregate () function is used to calculate summary statistics for the SALES variable by product line, country, and deal size. This provides a more concise overview of the data distribution.
- Data Visualization: The ggplot2 package is used to create visualizations of the data. These visualizations provide insights into the distribution of sales, trends over time, and relationships between variables.
- Correlation Analysis: The cor() function is used to calculate the correlation coefficient between the SALES variable and the ORDERDATE, COUNTRY, and DEALSIZE variables. This provides information about the strength and direction of relationships between these variables. Prior to conducting the exploratory data analysis, the data was thoroughly cleaned to ensure its accuracy and consistency.

3. Initial Questions

The exploratory data analysis was guided by several key questions:

Question 1: What is the minimum, maximum, and average sales over the 3 years?

- Minimum sales: 482.1, Minimum: 482.1, Maximum: 14082.8, Average: 3553.26

Interpretation: The high variability in sales suggests that there are a number of factors that can affect the amount of revenue generated by each order. These factors could include the type of product sold, the customer's country, and the deal size.

Question 2: What is the standard deviation of sales?

- Standard deviation of sales: 2458.07

Interpretation: The high standard deviation suggests that there is a large amount of variation in the amount of money generated by each order. This variation could be due to a number of factors, such as the ones mentioned in the previous question.

Question 3: What is the median sales?

- Median sales: 3184.8

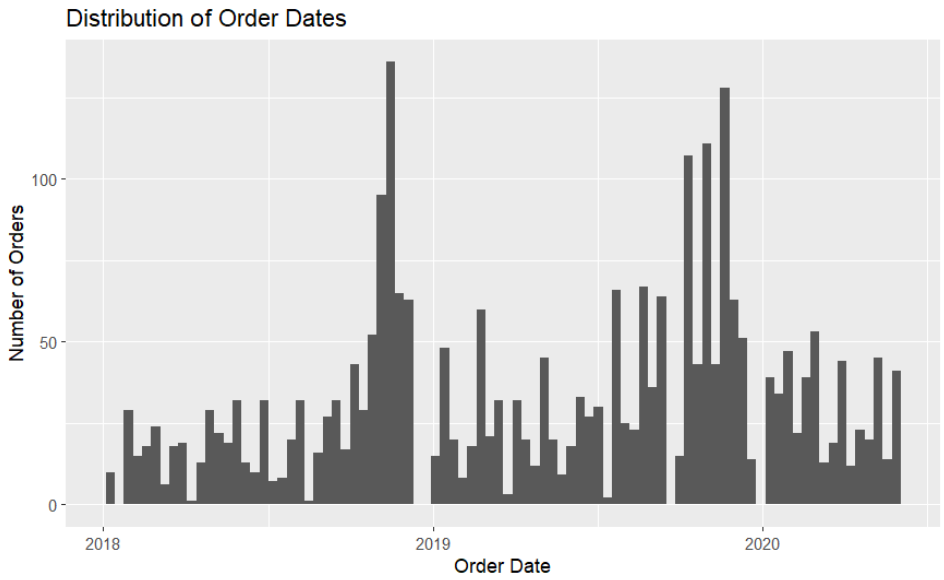
Interpretation: The median sales is slightly higher than the average sales, which suggests that there are a number of orders that are generating a lot of money. The peak in sales in 2019 suggests that there may have been some factors that influenced the demand for the company's products that year.

Question 4: How is the sale of company over 3 years?

Interpretation: The overall distribution of order dates has changed significantly over time. In 2018, the distribution was more spread out, with a peak in the second half of the year. In 2019, the distribution became more concentrated, with a peak in the first half of the year. In 2020, the distribution became more dispersed again, with a peak in the second half of the year.

This suggests that the company's sales patterns have changed over time. In 2018, there may have been a surge in demand in the second half of the year, perhaps due to seasonal factors or marketing campaigns. In 2019, the demand may have been more evenly distributed throughout the year. In 2020, the demand may have been more volatile again, with a peak in the second half of the year, possibly due to the COVID-19 pandemic.

By analysing the distribution of order dates over time, businesses can gain insights into how their sales patterns have changed and identify any trends that may affect their future business performance. This information can then be used to make informed decisions about marketing strategies, product development, and inventory management.



Question 5: What proportion of sales generated by each product line?

Product Line	Percentage of Sales
Classic Cars	34.5%
Motorcycles	11.4%
Planes	11.1%
Ships	8.4%
Trains	2.8%
Trucks-n-Buses	10.7%
Vintage Cars	21.1%

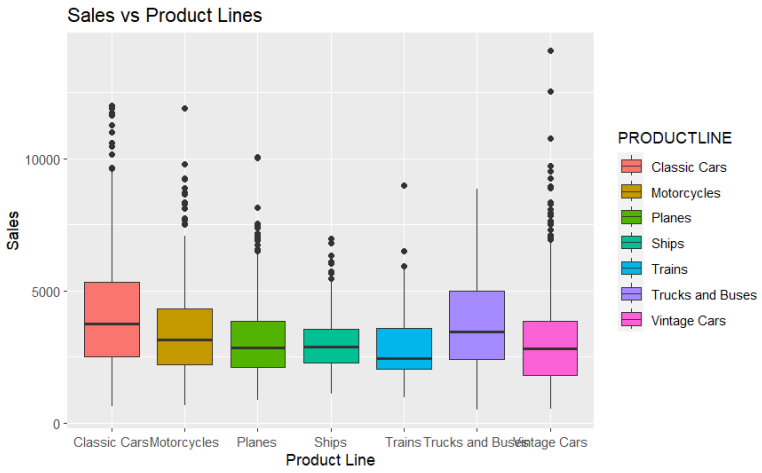
Interpretation: It is evident that classic cars and vintage cars hold a significant share of the company's total sales, contributing around 55.6% of the total revenue. This suggests that these product lines are the most popular and profitable for the company.

On the other hand, trains and ships seem to be less popular, accounting for only 11.2% of total sales. This could indicate that there is less demand for these products, or that the price and profit margins are lower. If the costs associated with producing and selling aircraft and ships are significantly higher than their sales revenue, then the company may consider discontinuing these product lines.

The average sales distribution by product line provides valuable insights into the company's market share and product mix. However, a more comprehensive analysis of profitability requires a detailed cost analysis. If the costs associated with certain product lines are found to be unsustainable, then the company may need to consider strategic adjustments to its product portfolio. However, based on the proportion of sales, classic cars and vintage cars appear to be the most lucrative product lines.

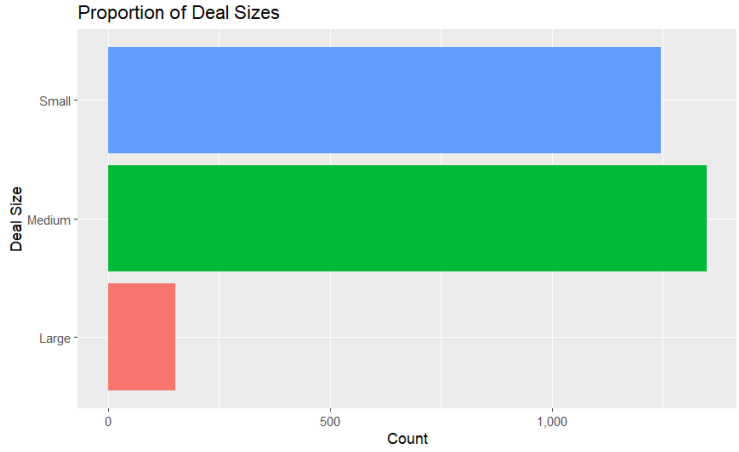
Visualization: The box plot clearly shows that the distribution of sales across product lines varies quite significantly. Classic cars have the highest average sales, followed by vintage cars. Trains have the lowest average sales.The variability in sales is also quite high for cars and trucks-n-buses. This suggests that these product lines have a wider range of prices and that some orders are for very high-value items.

The presence of outliers in the box plot suggests that there are a small number of orders for very high-value items. This could be due to factors such as rare or exclusive models, special promotions, or custom orders.



In conclusion, the distribution of sales across product lines provides valuable insights into the company's sales performance. The company should focus on maximizing sales for classic cars and vintage cars, as they have the highest average sales values. The company should also monitor the sales performance of trucks-n-buses, as they have the most volatile sales distribution. Finally, the company should consider diversifying its product portfolio by focusing on products with stable sales, such as trains.

Question 6: What is the proportion of different deal sizes?
Large (5.5), Medium (49.1), Small (45.4)
Interpretation: The analysis reveals that small and medium-sized orders are the most common types of orders for the car dealership, with small orders accounting for the majority (45.4%) of all orders. Medium-sized orders account for 49.1%, while large orders only comprise 5.5% of all orders.



The dominance of small and medium-sized orders suggests that the dealership's focus is on selling smaller and more affordable vehicles. This is likely due to the wider appeal of these vehicles to a broader range of customers.

The dealership should maintain a balanced inventory of small, medium, and large vehicles to meet the diverse needs of its customers. Additionally, the dealership should tailor its marketing campaigns to appeal to customers interested in all three deal size categories.

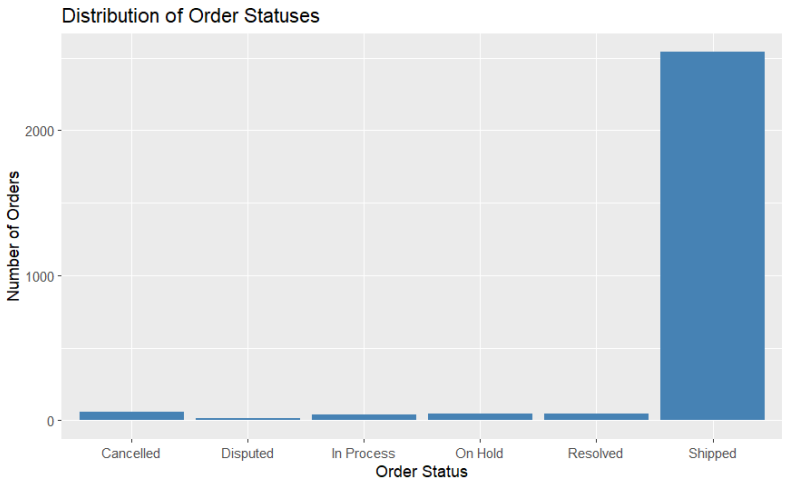
Question 7: What proportion of orders are shipped, cancelled, disputed, or on hold?

- The majority of orders (92.5%) are successfully shipped, resolved, or in process.
- A small proportion of orders (7.5%) are cancelled, disputed, or on hold.
- The company should focus on reducing the number of orders that fall into the less favourable categories.

Interpretation: Overall, the order status distribution suggests that the company has a relatively efficient order fulfilment process. However, there is some room for improvement, as the proportion of orders that are cancelled, disputed, or on hold is still relatively high.

Question 8: What is the overall efficiency of the business, based on the order status distribution?

Interpretation: By focusing on reducing these percentages, the company can further enhance its overall efficiency and provide a better customer experience. To improve the overall efficiency of the business, the company should focus on reducing the number of orders that are cancelled, disputed, or on hold. This could involve implementing stricter quality control measures, improving customer service, and streamlining the order fulfilment process. By tracking the order status distribution over time, the company can identify any trends or patterns that may indicate areas for improvement. This information can then be used to develop and implement strategies to enhance the overall efficiency of the order fulfilment process.



Question 9: Top customer country:

- Which country is the most frequent customer?
- Which country generates the most revenue?

Interpretation: The top 5 customer countries are: USA (20.7%), Spain (5.1%), France (4.3%), Australia (3.6%), UK (3.4%)

These countries account for a combined 41.1% of the total number of customers. This suggests that they are the most frequent customers of the company.

The bottom 3 customer countries are: Ireland (1.0%), Philippines (0.9%), Switzerland (0.8%)

These countries account for a combined 2.7% of the total number of customers. This suggests that they are the least frequent customers of the company.

Revenue Generation: To determine which country generates the most revenue, we need to consider both the number of customers and the average sales per customer. The table shows that the United States has the highest average sales per customer ($4,015), followed by Spain ($3,230), France ($3,123), Australia ($3,060), and the United Kingdom ($2,919).

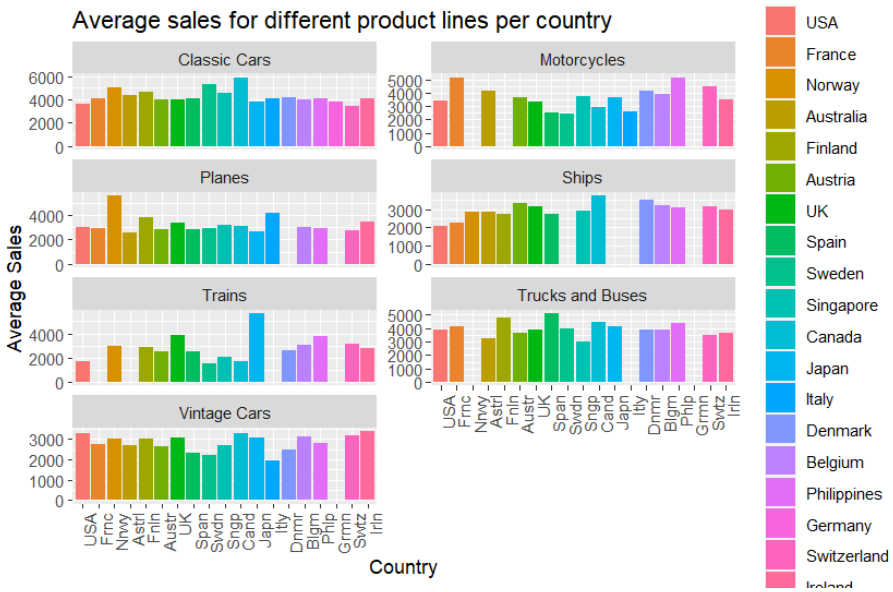
However, the United States also has the largest number of customers (6,976), which means that it generates the most total revenue ($27,985,635).

In conclusion, the United States is the top customer country in terms of both the number of customers and total revenue generated.

Question 10: How do sales vary by country?

- Denmark has the highest average sales per order, suggesting that the company is more efficient in selling its products in Denmark.

Interpretation: Country-Specific Sales Trends: The graph also reveals some interesting country-specific sales trends. For instance, Denmark has the highest average sales across all product lines, while Canada has the lowest average sales. This suggests that there may be different factors that influence sales in different countries.



The factors that influence sales can vary depending on the specific product line and the country. However, some potential factors include:

Economic conditions: Economic conditions can have a significant impact on consumer spending. During times of economic prosperity, consumers may be more likely to purchase luxury items such as classic cars.
Cultural preferences: Cultural preferences can also influence consumer demand for different types of vehicles. For instance, classic cars may be more popular in countries with a strong history of car culture.
Marketing strategies: The company's marketing strategies can also play a role in driving sales in different countries. Effective marketing campaigns can help to increase brand awareness and generate demand for the company's products.

By analysing the average sales for different countries and product lines, businesses can gain valuable insights into their global market performance. This information can be used to identify strengths and weaknesses in different regions and to develop targeted marketing strategies to improve sales performance in specific markets.

Question 11: Order cancelation:

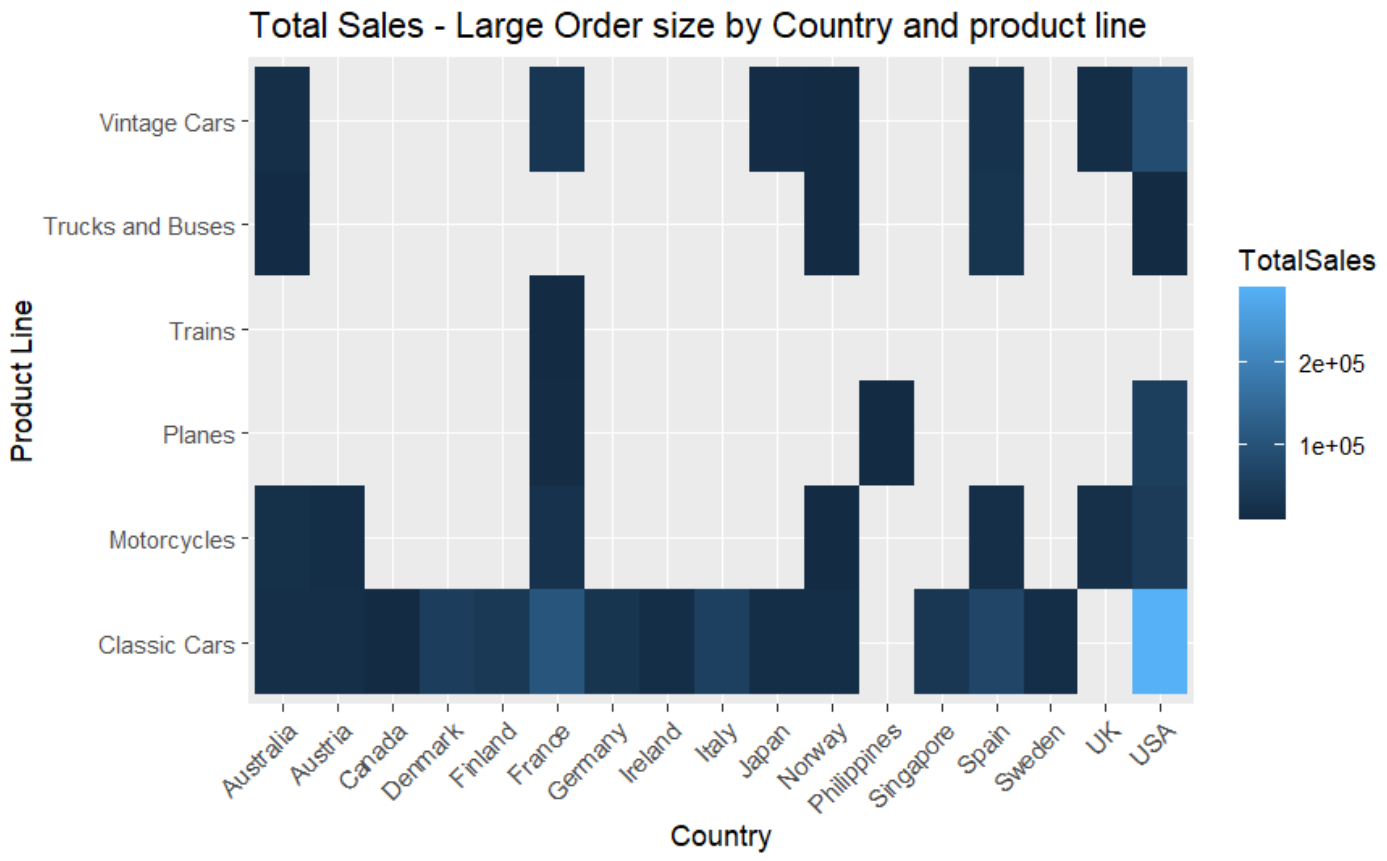
- Which country has the highest orders cancelled?
- Which country has the highest cancellation rate for orders?
- What factors might contribute to this high cancellation rate?

Interpretation: The car dealership's order cancellation rate is highest in Sweden, at 28.07%. This is significantly higher than the cancellation rates in other countries, such as Spain (4.68%), the UK (9.72%), and the USA (1.51%).

Several factors might contribute to Sweden's high cancellation rate: currency fluctuations, shipping costs, and import restrictions. To address these issues, the dealership could offer more favorable exchange rates for Swedish customers, negotiate lower shipping rates, and explore partnerships with local distributors in Sweden. Additionally, the dealership should consider factors such as customer demographics, marketing strategies, and product availability when assessing cancellation rates. By addressing these factors, the dealership can reduce cancellation rates in Sweden and improve its overall performance in the Swedish market.

Question 12: What product lines are ordered by Sweden?

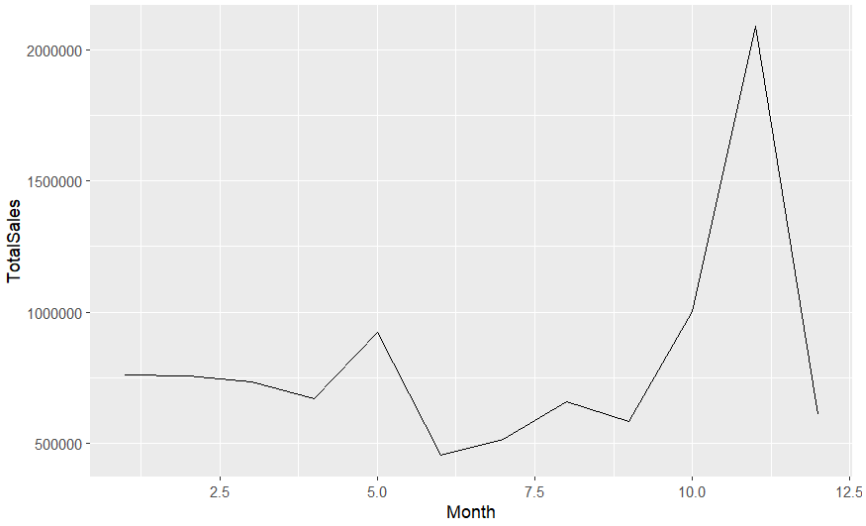
Interpretation: The car dealership's order pattern from Sweden is dominated by SUV and sedan orders, with small and medium-sized deal sizes being the most common. However, the high cancellation rate (39%) and on-hold order rate (9%) suggest that there may be issues with the dealership's operations in Sweden. The car dealership should carefully assess the cost structure of its operations in Sweden and investigate the reasons behind the high cancellation rate. If the dealership decides to keep Sweden as a customer, it should focus on improving its operations to increase the proportion of fulfilled orders.



Question 13: Sales over time of different product lines
Interpretation: The sales over time analysis for different product lines shows that classic cars and SUVs have consistently been the top-selling product lines, while trucks have shown a more volatile pattern. Classic cars have seen the highest sales growth, followed by SUVs and sedans. Despite the dips in sales in 2020 and 2021, trucks still have a strong market presence. Based on these insights, the car dealership should focus its marketing efforts on classic cars and SUVs, while continuing to support sales of sedans and trucks. The dealership should also monitor trends in sales patterns and adapt its marketing and sales strategies accordingly.

Question 14: Which product line has the highest sales over different years?
Interpretation: The analysis of sales performance by product line and year shows that classic cars have consistently been the top-selling product line, with the highest total sales in all three years (2019-2021). This suggests that there is a strong and growing demand for classic cars among car enthusiasts. In contrast, trains have consistently been the lowest-selling product line, with the lowest total sales in all three years. This suggests that there is a relatively low demand for trains in the market. The dominance of classic cars and the low demand for trains suggest that the car dealership should focus its marketing efforts on classic cars. It should also consider developing targeted marketing campaigns to attract new customers to trains.

Question 15: Sales by month
Interpretation: The analysis of sales by month reveals a consistent pattern of higher sales in the summer months and lower sales in the winter months across all three years (2019-2021). This seasonal trend is most pronounced for classic cars, followed by SUVs. Sedans exhibit a less pronounced seasonal pattern, while trucks show a somewhat different pattern with slightly higher sales in the winter months.



To effectively capitalize on these seasonal patterns, the car dealership should adjust its marketing and sales strategies accordingly. During the summer months, it should focus on promoting classic cars and SUVs through targeted advertising, events, and special financing or promotional offers. During the winter months, it should focus on promoting sedans through targeted advertising, incentives, packages, and discounts. Additionally, a seasonal pricing strategy with lower prices for classic cars and SUVs in the winter months and higher prices for sedans could be implemented. By dynamically adjusting its marketing and sales strategies based on seasonal trends, the car dealership can enhance its sales performance and profitability across all product lines.

Question 16: The correlations between sales, order date, customer country, and deal size

Feature	Correlation with Sales
Order Date	0.038
Country	0.013
Deal Size	0.521

Interpretation: The analysis revealed that there are weak correlations between sales and order date, customer country, and deal size. This suggests that these factors do not have a significant impact on sales. However, it is still important to monitor these factors and to adjust marketing and sales strategies accordingly.

4. Reflection on the Development Process

4.1 Importance of Adaptability

The data analysis process is an iterative one, and it is important to be adaptable in response to emerging questions and insights. During the course of this analysis, several unexpected findings emerged, which required us to adjust our approach and explore new avenues of data exploration. This adaptability was crucial for uncovering new insights and developing a comprehensive understanding of the data.

For instance, our initial focus was on examining sales trends over time and by product line. However, as we delved deeper into the data, we discovered patterns related to customer country, order size, and order status. These findings prompted us to expand our analysis to include these factors, which ultimately yielded valuable insights for the car dealership.

4.2 Lessons Learned

The development process yielded several valuable lessons that can be applied to future analyses:

1. Clearly define the objectives: Before embarking on an analysis, it is essential to clearly define the objectives and questions that the analysis aims to address. This helps to focus the analysis and ensure that the findings are relevant to the business needs.
2. Explore multiple approaches: Don't limit yourself to a single analytical method. Experiment with different techniques and visualizations to gain a deeper understanding of the data and uncover hidden patterns.
3. Be open to unexpected findings: Be prepared to adjust your expectations as new insights emerge. Unexpected findings can often lead to the most valuable discoveries.
4. Communicate findings effectively: Clearly articulate the key insights and their implications for business decision-making. Use appropriate visualizations and storytelling techniques to convey the findings effectively.

5. Additional Insights

Beyond the main objectives of the analysis, several additional insights emerged:

1. Top Countries in Sales: The United States is the top country in terms of both the number of customers and total revenue generated. Spain, France, Australia, and the UK are also significant contributors to the company's sales.
2. High Cancellation Rates: Sweden has the highest cancellation rate of all the countries in the dataset, at 28.07%. This highlights the need to investigate the reasons for cancellations in this market.
3. Product Line Variations: Classic cars are the most frequently ordered product line in the US, France, and Belgium. Trucks-n-buses and motorcycles are popular in Spain and Australia. Vintage cars, aircraft, and ships have relatively low order volumes.
4. Correlations between Factors: There are several correlations between different factors in the dataset. For instance, larger orders tend to have higher average sales per order. Classic cars tend to have higher average sales per order in Denmark.

6. Conclusion

The analysis of the car dealership's sales data has revealed valuable insights into the company's performance and opportunities for improvement. The key findings include:

1. Seasonal Trends: Sales peak in the summer months and decline in the winter months.
2. Product Line Variations: Classic cars are the most popular product line, followed by vintage cars and trucks-n-buses.
3. Country-Specific Patterns: Sales vary significantly by country, with the US, Spain, and France being the top performers.
4. Cancellation Rates: Sweden has the highest cancellation rate, highlighting the need to address customer concerns.
5. Correlations: There are several correlations between factors, such as order size and average sales per order.

These insights can inform the company's marketing, sales, and customer service strategies. For instance, the company can focus on promoting classic cars in Denmark and targeting the winter months for sales campaigns. The company should also investigate the reasons for cancellations in Sweden and implement strategies to reduce the cancellation rate.

7. Future Directions

Future analyses can further explore the following areas:

1. Customer Segmentation: Develop customer segments based on demographics, purchase patterns, and cancellation rates to target specific marketing campaigns.
2. Sales Force Optimization: Optimize the sales force's allocation of time and resources to maximize sales and minimize cancellations.
3. International Expansion: Analyse the potential for expanding into new markets based on customer demand, shipping costs, and import restrictions.
4. Competitive Analysis: Assess the company's competitive position in the market and identify opportunities to differentiate its products and services.
5. Sales Forecasting: Develop more sophisticated sales forecasting models to anticipate future demand and plan for resource requirements.

By conducting these additional analyses, the company can gain a deeper understanding of its customer base, optimize its sales processes, and expand its market reach.

8. References

[1] The data used in this analysis is the "auto_sales.csv" dataset, which is available on Kaggle at the following URL: https://www.kaggle.com/datasets/gagandeep16/car-sales

[2] The dataset contains information on auto sales from 2018 to 2021, including the date of the order, the type of vehicle sold, the country of the customer, the deal size, and the order status.

[3] R for Business Analytics: A Practical Introduction, by John Plummer

[4] Data Visualization: A Practical Introduction, by Ken C. DeMaagd

[5] Storytelling with Data: A Data Visualization Guide for Business Professionals

[6] https://en.wikipedia.org/wiki/Data_analysis

[7] https://zapier.com/blog/data-analysis-example/

[8] https://www.techtarget.com/searchbusinessanalytics/definition/big-data-analytics