**COMP2445 Data Mining**

**Student number: 22011877**

**Submission date: 24/11/2023**

**Assignment 1**

# Executive Summary

This detailed report presents a comprehensive analysis of Civitas Motors' marketing strategies and customer insights, leveraging advanced statistical techniques to offer actionable intelligence. The analysis is segmented into two primary sections:

1. **Exploratory Data Analysis (EDA):**

This section covers exploratory findings, encompassing the popularity of car makes among repeat customers, city-wise sales distribution, and monthly sales trends. The EDA provides a holistic view of customer preferences and sales patterns, crucial for aligning marketing and sales strategies with customer behaviour and market opportunities.

1. **Customer Demographics:**

In-depth analysis of Civitas Motors’ customer database reveals significant insights into age demographics and vehicle replacement cycles. This aspect of the report delves into understanding the profile of the customer base, informing targeted marketing strategies and product offerings.

1. **Marketing Channel ROI Analysis:**

A critical evaluation of different marketing channels, including online advertising, mailshot marketing, and newspaper ads, is conducted to assess their effectiveness. Using linear and polynomial regression models, the report predicts the sales impact of each channel, identifying those that offer the highest return on investment.

The report concludes with strategic recommendations for Civitas Motors, focusing on optimising marketing spends, tailoring approaches based on customer demographics, and enhancing overall customer engagement. Acknowledging the limitations of the analysis, such as data scope and statistical assumptions, the report aims to guide Civitas Motors in strengthening its market presence, improving customer satisfaction, and boosting sales growth in the automotive industry.

# Table of Contents

**Contents**

[Executive Summary 2](#_Toc151022841)

[Table of Contents 3](#_Toc151022842)

[Data Sources 3](#_Toc151022843)

[Introduction 3](#_Toc151022844)

[EDA 4](#_Toc151022845)

[Analysis of Customer Base 8](#_Toc151022846)

[Analysis of Sales Across a Fiscal Year 13](#_Toc151022847)

[Analysis of marketing channels 15](#_Toc151022848)

[Limitations 24](#_Toc151022849)

[Conclusion and Recommendations 24](#_Toc151022850)

# Data Sources

The primary dataset under scrutiny originates from Civitas Motors' internal systems.

# Introduction

This report offers a comprehensive exploration of Civitas Motors' marketing effectiveness and customer trends. Utilising statistical analysis, it provides insights into customer demographics, sales distribution, and marketing channel efficacy, crucial for informed strategic decision-making. The report aims to equip Civitas Motors with the knowledge to optimise marketing strategies and customer engagement in a competitive automotive market.

# EDA

**Key Insights**

This EDA report offers Civitas Motors valuable insights into customer preferences and sales patterns, essential for informed decision-making.

1. **Favourite Car Makes Among Repeat Customers:**

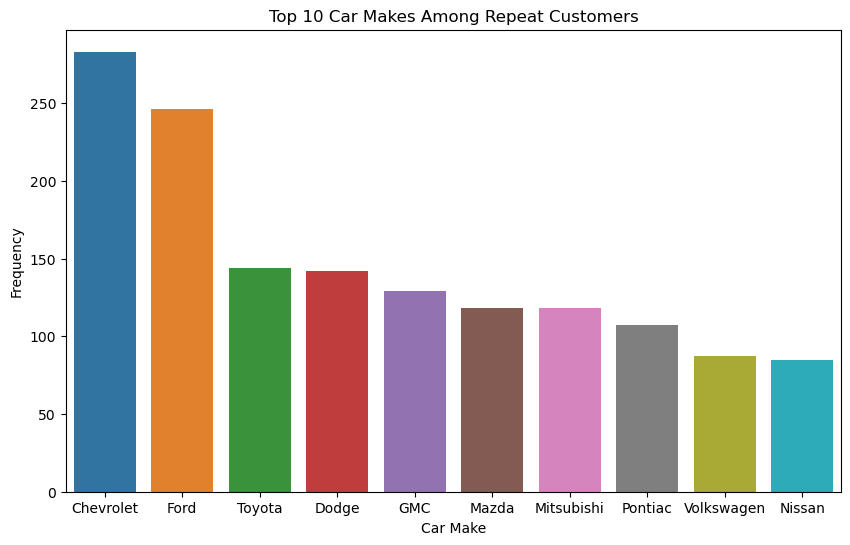
****

Figure 0a

* + A bar chart (Figure 0a) highlights the top 10 car makes favoured by repeat customers, underscoring brand loyalty and customer preferences.

1. **Sales Distribution by City:**

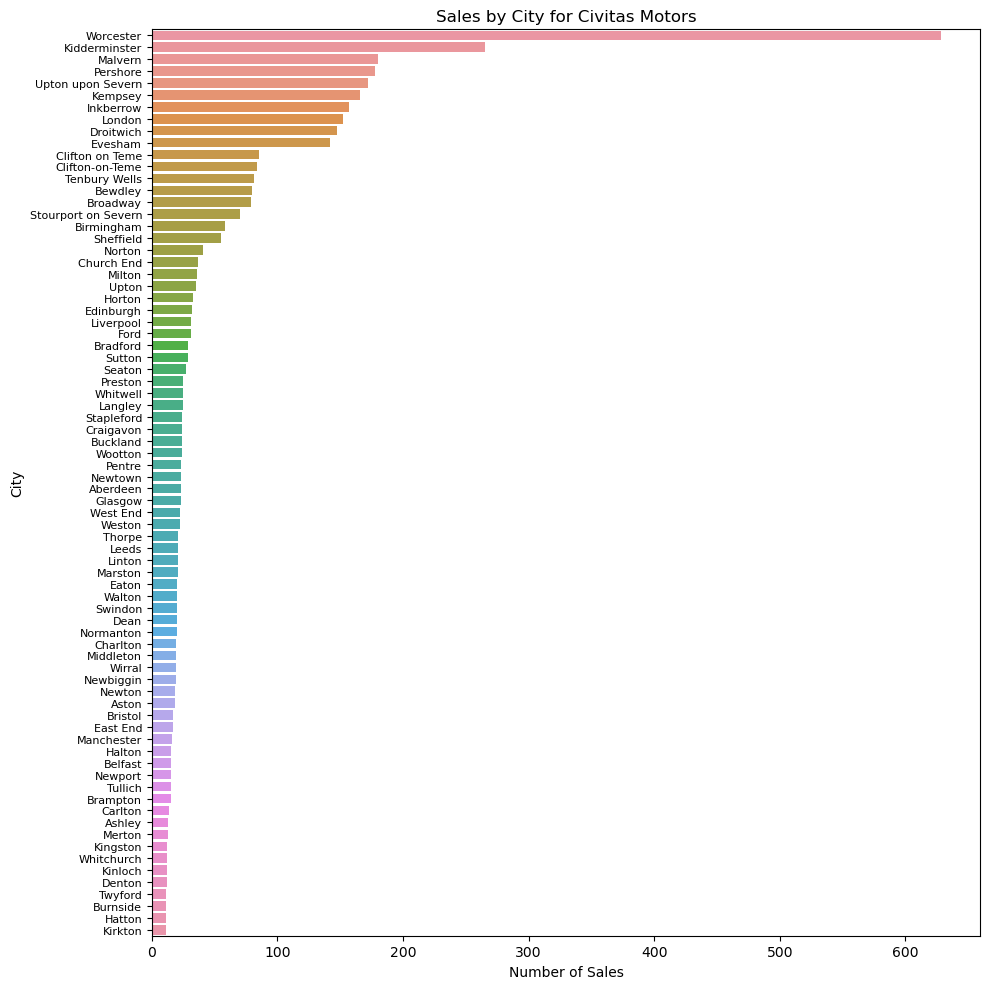
****

Figure 0b

* + Sales data grouped by city (Figure 0b) reveals key regions with the highest sales, indicating market dominance and potential growth areas.

1. **Monthly Sales Trends:**

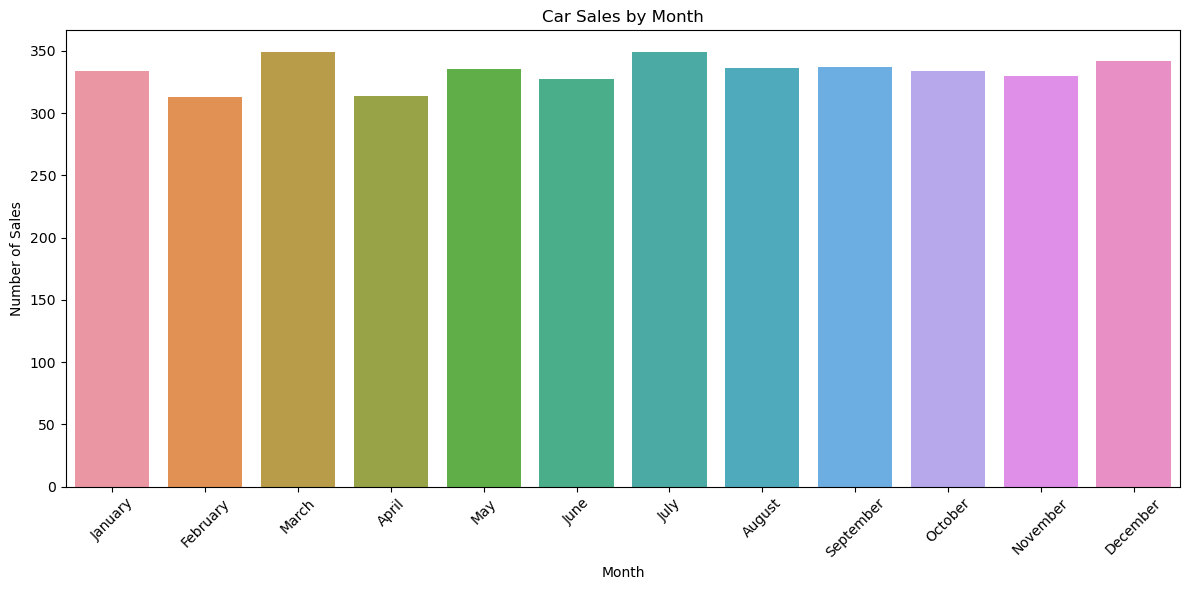
****

Figure 0c

* + Analysis of monthly sales trends (Figure 0c) identifies peak sales periods, vital for targeted marketing and inventory management.

1. **Vehicle Replacement Cycles:**

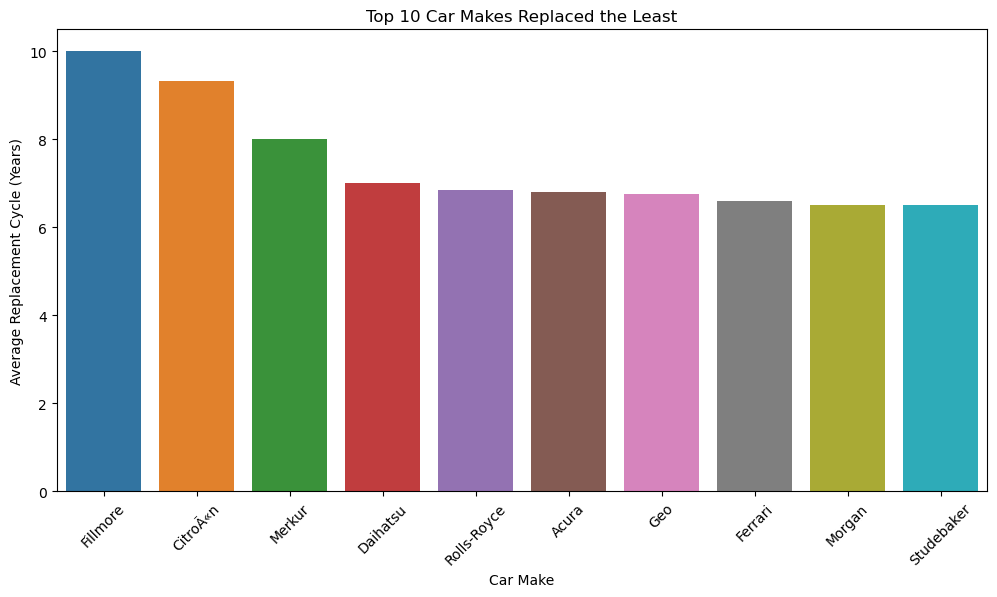
****

Figure 0d

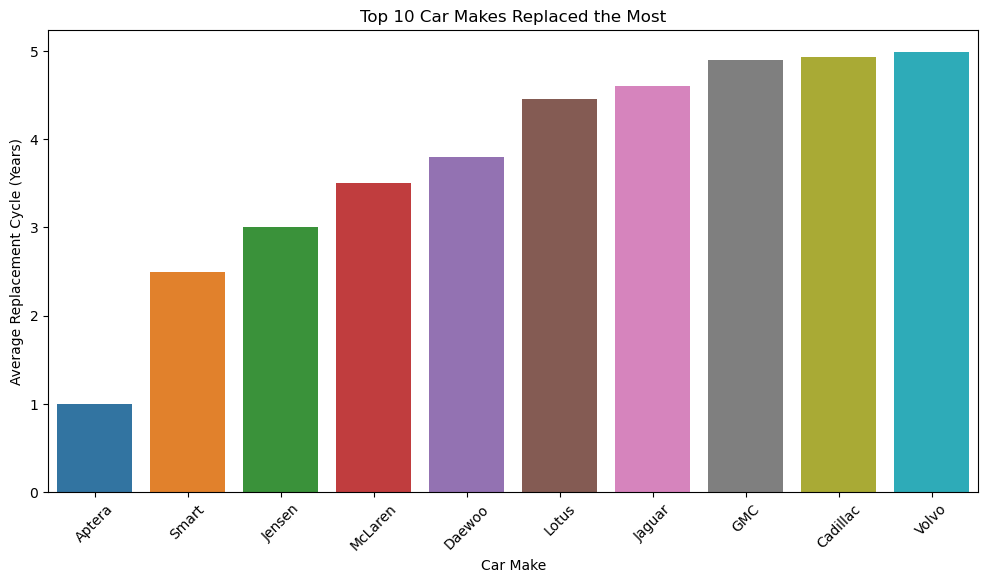


Figure 0e

* + The study of replacement cycles (Figures 0d and 0e) contrasts the most and least frequently replaced car makes, providing insights into customer satisfaction and vehicle longevity.

**Conclusion and Recommendations:**

* **Targeted Marketing Initiatives:** Capitalise on favourite car makes to enhance customer loyalty.
* **Strategic Regional Focus:** Increase marketing and inventory in lower-sales regions to capture new markets.
* **Seasonal Sales Strategy:** Align stock and marketing efforts with peak sales months for optimal sales performance.
* **Customer Retention Analysis:** Investigate short replacement cycles to improve customer satisfaction and retention.

**Utilisation of Insights**

These findings enable Civitas Motors to tailor marketing strategies, manage inventory efficiently, and enhance customer engagement, driving growth and profitability in the competitive automotive market.

# Analysis of Customer Base

This section of the report provides an overview of Civitas Motors' customer database integrity and a profile of the customer base. The analysis includes data verification, customer age distribution, vehicle replacement cycles, and customer concentration in specific postal code areas.

**Data Verification**

The dataset consists of 4,000 records. Utilising .shape[0] and .count() methods confirmed the absence of missing values across all fields, indicating a complete and robust dataset ready for further analysis.

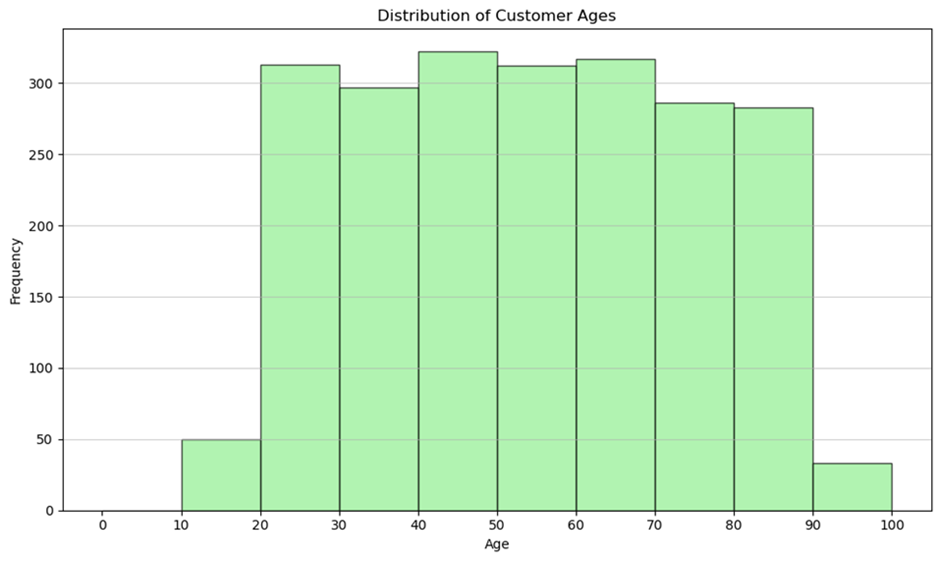
Screens screenshot of a computer screen

Description automatically generated

**Customer Age Distribution**

To avoid skewed interpretations due to multiple purchases by the same customer, duplicate records were removed based on 'Customer ID'. The age distribution of unique customers was visualised using a histogram (see Figure 1a), which revealed a comprehensive view of the demographic spread.

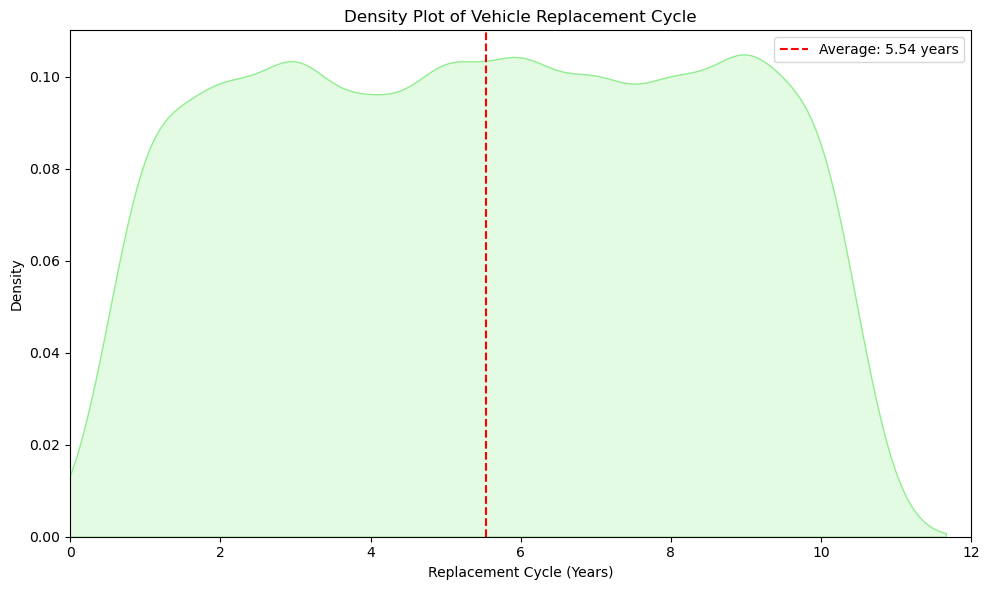
Figure 1a:



**Vehicle Replacement Cycle**

The average vehicle replacement cycle was calculated to be approximately 5.5 years. A density plot (see Figure 1b) was chosen over a histogram for its ability to illustrate the distribution of replacement cycles more smoothly, providing a clear view of the frequency at which customers change their vehicles.

Figure 1b:



**Customer Concentration in WR10 and WR11**

Analysis of customers in the WR10 and WR11 postcode areas showed a total of 180 unique customers. A histogram (see Figure 1c) displayed the age distribution, indicating the highest concentration in the 40-50 age range.

Figure 1c:

A graph with green bars

Description automatically generated

**Financial Analysis of Customer Vehicles**

The total value of vehicles owned by customers in these areas was over £20 million. To evaluate individual vehicle values and exclude statistical outliers, z-scores were computed. No outliers were detected using a threshold of three standard deviations from the mean, suggesting a homogenous vehicle value distribution within this customer segment.

A boxplot visualisation (see Figure 1d) was selected to depict the retail price distribution of vehicles. This plot efficiently highlights the average vehicle value, interquartile range, and potential outliers. However, as no outliers were present, the boxplot serves to confirm the consistency of vehicle values.

Figure 1d:

A graph with a green rectangular box and red line

Description automatically generated with medium confidence

**Conclusion**

Civitas Motors maintains a complete and comprehensive customer database, with a diverse age demographic and a stable vehicle replacement cycle. The concentration of customers in the WR10 and WR11 areas presents a significant portion of the customer base with a consistent vehicle value distribution.

# Analysis of Sales Across a Fiscal Year

This section presents an in-depth analysis of Civitas Motors' vehicle sales performance across different fiscal years, utilising time-series data to identify sales trends and inform future business strategies.

**Data Preparation and Analysis Methodology**

The purchase dates were first converted into a monthly time-series format, providing a foundation for a detailed examination of sales trends. Adjustments were made to reflect the company's fiscal year, ensuring that sales data align with financial reporting periods.

Retail prices were normalised to ensure consistency in the financial data. The analysis was focused on comparing sales during the fiscal years 2018 and 2019, offering a direct and relevant evaluation of the company's year-on-year performance.

**Monthly Sales Trends**

A line plot of monthly sales (see Figure 2a) provided initial insights into the company's sales trends over time, suggesting patterns that may correlate with marketing initiatives or seasonal fluctuations.

Figure 2a:

A graph with green lines

Description automatically generated

**Fiscal Year Sales Performance**

By aligning the sales data with the fiscal year calendar, a bar chart (see Figure 2b) was created to visualise the total sales for fiscal years 2018-2019 and 2019-2020. This visualisation aids in understanding annual sales performance in relation to the company's fiscal reporting.

Figure 2b:

**A graph of a vehicle sales report

Description automatically generated**

**Conclusions**

The graphical representations of sales data over monthly and annual periods provide Civitas Motors with actionable insights. Understanding these patterns is crucial for strategic planning, resource allocation, and targeted marketing campaigns to enhance sales performance in subsequent fiscal years.

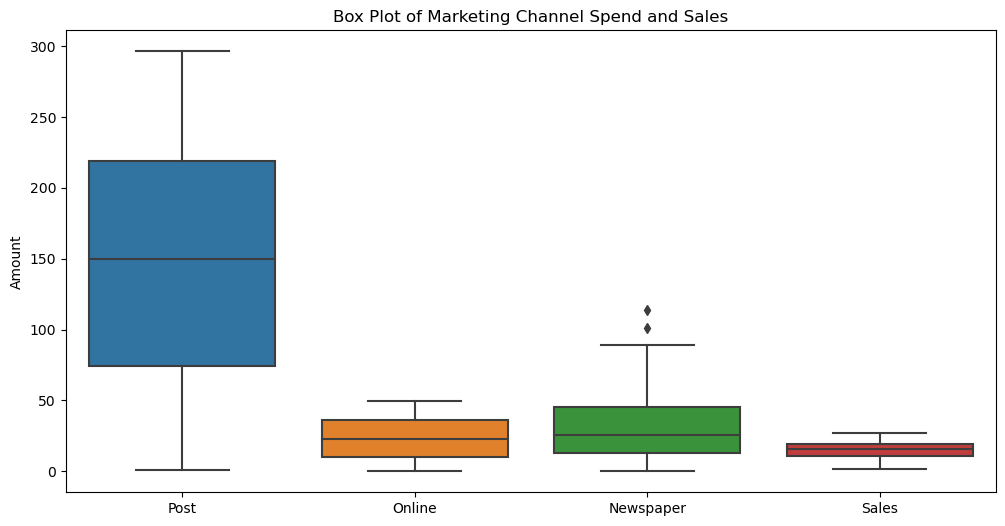
# Analysis of marketing channels

Civitas Motors aims to determine the most effective marketing methods in terms of ROI. The marketing methods in question are online advertising (Google and Facebook), mailshot marketing, and advertising in the Worcester News.

**Data Preparation and Preliminary Analysis**

The marketing dataset, encompassing expenditures across three channels and corresponding sales figures, underwent rigorous preprocessing. An initial box plot (see Figure 3a) visualisation identified the presence of outliers, subsequently addressed through Z-score calculations with a threshold of 3 for outlier removal. This process refined the dataset, ensuring integrity and reliability for the subsequent analysis.

Figure 3a:

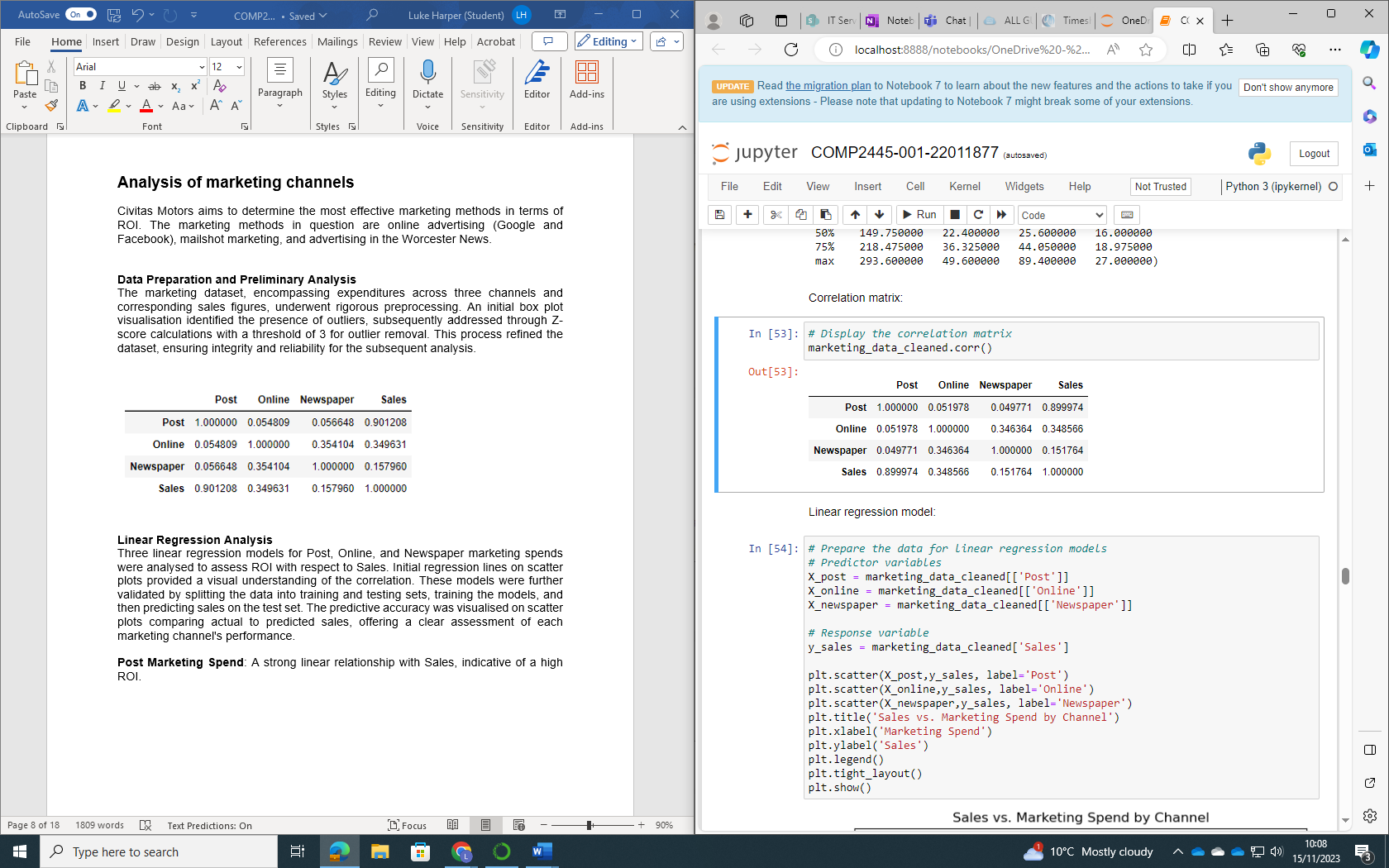


**Statistical Summary and Correlation Analysis**

Post-cleansing, the dataset included 198 entries. Key statistical measures, such as mean and standard deviation, were described for each marketing channel and sales. These figures provided a foundational understanding of the distribution and central tendencies of the marketing expenditures and their relationship with sales.

A correlation matrix (see Figure 3b) was developed to quantify the relationships between each marketing channel and sales. This analysis revealed a substantial positive correlation between Post spend and sales, indicating a significant impact of this channel on sales performance. The Online and Newspaper channels showed relatively lower, yet notable, correlations.

Figure 3b:



**Linear Regression Analysis**

Three linear regression models for Post (Figures 3c and 3d), Online (Figures 3e and 3f), and Newspaper (Figure 3g and 3h) marketing spends were analysed to assess ROI with respect to Sales. Initial regression lines on scatter plots provided a visual understanding of the correlation. These models were further validated by splitting the data into training and testing sets, training the models, and then predicting sales on the test set. The predictive accuracy was visualised on scatter plots comparing actual to predicted sales, offering a clear assessment of each marketing channel's performance.

**Post Marketing Spend**: A strong linear relationship with Sales, indicative of a high ROI.

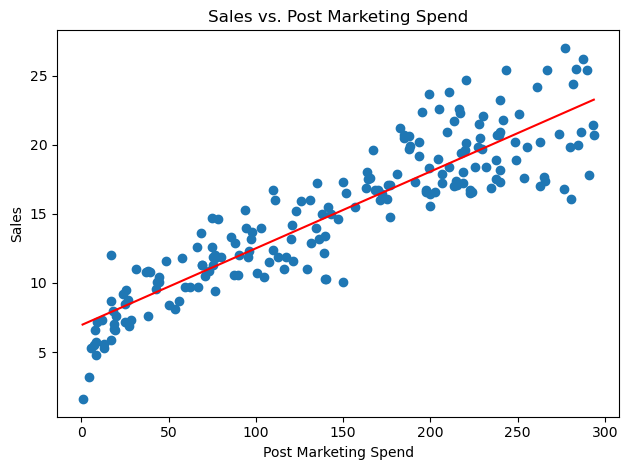


Figure 3c

****

Figure 3d

**Online Marketing Spend**: A moderate linear relationship with Sales, suggesting a fair ROI.

**A graph with blue dots and a red line

Description automatically generated**

Figure 3e

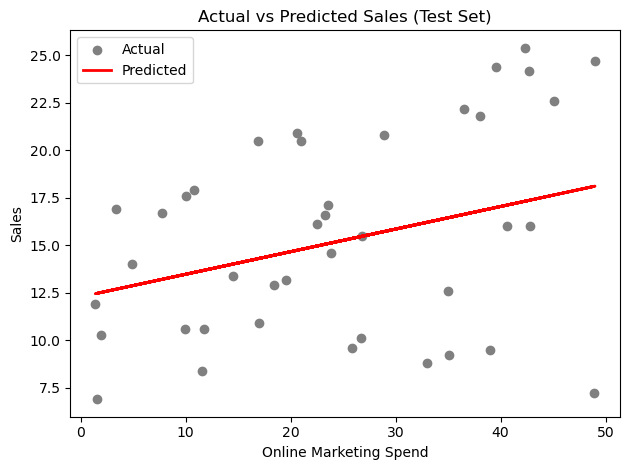
****

Figure 3f

**Newspaper Marketing Spend**: The weakest linear relationship with Sales, implying the lowest ROI.

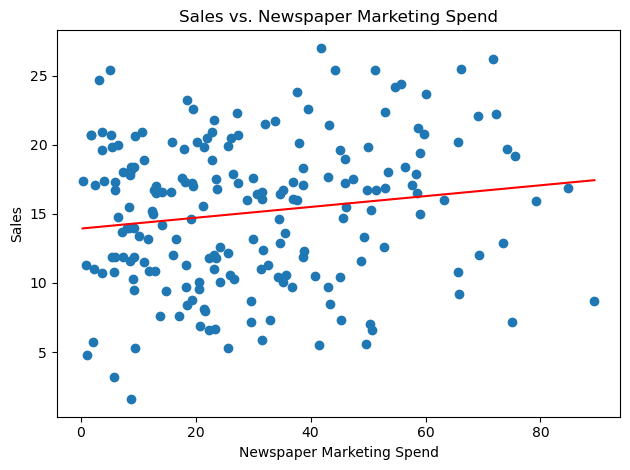


Figure 3g

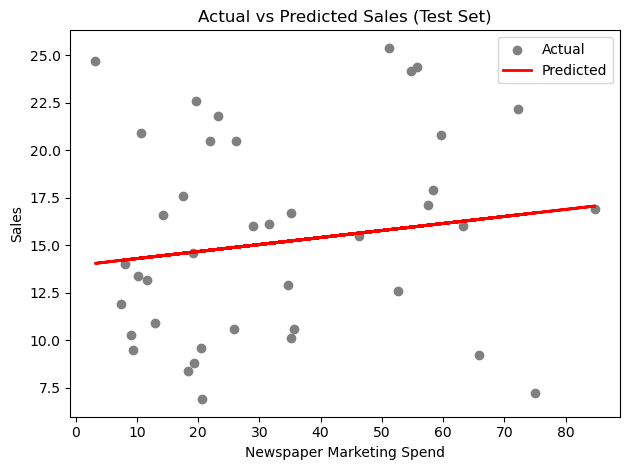


Figure 3h

The evaluation metrics computed for each marketing channel were:

* **Post Spend Model Evaluation Metrics:**
  + Mean Absolute Error (MAE): 1.6079
  + Mean Squared Error (MSE): 4.1400
  + Root Mean Squared Error (RMSE): 2.0347
* **Online Spend Model Evaluation Metrics:**
  + Mean Absolute Error (MAE): 4.1786
  + Mean Squared Error (MSE): 24.0075
  + Root Mean Squared Error (RMSE): 4.8997
* **Newspaper Spend Model Evaluation Metrics:**
  + Mean Absolute Error (MAE): 4.2563
  + Mean Squared Error (MSE): 27.0234
  + Root Mean Squared Error (RMSE): 5.1984

The Post Spend Model exhibited a robust relationship with sales, indicating the highest ROI. While the Online Spend Model was less efficient, it showed potential for ROI and could serve as a supplementary channel. Conversely, the Newspaper Spend Model, marked by the highest prediction errors and weakest correlation, emerged as the least effective channel.

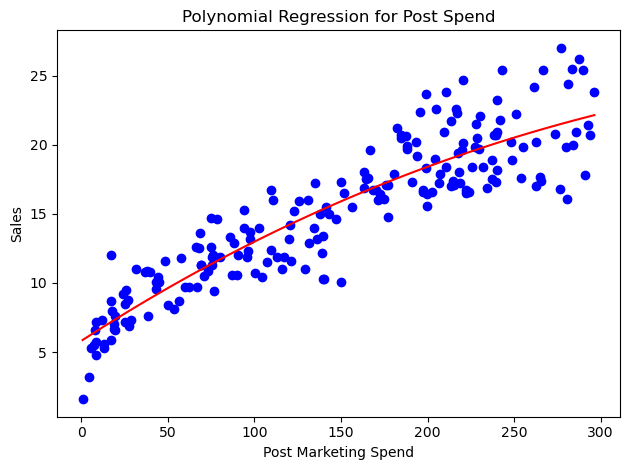
**Multivariate Regression Analysis**

A scenario with equal expenditure across Post, Online, and Newspaper channels projected sales around 329.25 units. The coefficients revealed Online marketing as the most influential (0.1069 coefficient), followed by Post (0.0545) and Newspaper (0.0009). This suggests prioritising Online and Post marketing for better investment returns, with Newspaper marketing being less impactful.

**Polynomial Regression Analysis**

Polynomial regression models (see Figures 3i, 3j and 3k) were applied to capture non-linear trends. The models were assessed with an R-squared statistic, which measures the proportion of variance in the dependent variable predictable from the independent variables:

**Post Spend Model:** R-squared = 0.8209

Figure 3i

**Online Spend Model:** R-squared = 0.1344

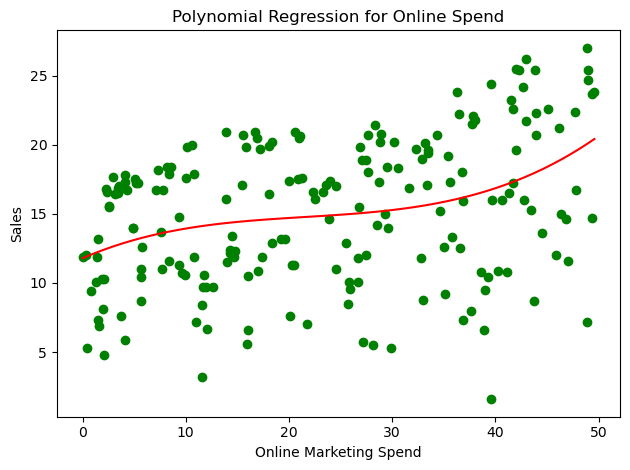
****

Figure 3j

**Newspaper Spend Model:** R-squared = 0.0534

A graph showing a red line and orange dots

Description automatically generated

Figure 3k

The Post Spend Model achieved a significantly higher R-squared value, implying a strong non-linear relationship with Sales and suggesting that mailshot marketing efficiency may increase at a higher rate with increased spending.

**Conclusion**

The analysis revealed that Post (mailshot marketing) provides the highest ROI, with the strongest linear and non-linear relationships with Sales. Online marketing, while less predictive, still presents a fair ROI and may benefit from its broader reach and targeting capabilities. Newspaper advertising shows the least effectiveness in driving Sales.

# Limitations

This report's analysis of Civitas Motors' marketing channels and customer demographics, while comprehensive, has certain limitations:

1. **Data Scope and Timeliness:**

The analysis relies solely on Civitas Motors' internal data, not accounting for post-collection market changes or evolving consumer behaviour, which could affect the findings' relevance.

1. **Statistical Methodology:**

The use of linear and polynomial regression models implies a specific relationship between marketing spend and sales, possibly oversimplifying complex market dynamics.

1. **Outlier Management:**

Employing Z-scores for outlier removal might have excluded significant data points, potentially omitting valuable insights, especially in unique marketing scenarios.

1. **Dataset Limitation:**

The conclusions are drawn exclusively from Civitas Motors' data, without considering external influences like economic variations, competitor strategies, or industry trends.

1. **Predictive Modelling Assumptions:**

The multivariate regression model's forecasts are based on equal spending scenarios across channels, which may not reflect actual market conditions.

Recognising these constraints is vital for Civitas Motors in applying these insights strategically. Future analyses would benefit from broader data sources, external market considerations, and possibly more advanced analytical methods.

# Conclusion and Recommendations

**Analysis Summary**

The analysis for Civitas Motors has provided pivotal insights into the impact of various marketing channels and customer demographics. Using advanced statistical methods, the study has effectively demonstrated the relationship between marketing spends in Post, Online, and Newspaper channels and their respective influence on sales.

1. **Post Marketing:** Evidently the most effective channel, Post marketing showed the strongest correlation with sales, indicating significant ROI.
2. **Online Marketing:** Demonstrated moderate impact but possesses considerable potential as indicated by its coefficient in the multivariate analysis, suggesting opportunities for enhanced engagement and conversion.
3. **Newspaper Marketing:** Proved to be the least effective, aligning with the shift towards digital platforms and evolving media consumption habits.
4. **Customer Demographics:** Insights into the customer base, particularly age demographics and vehicle replacement cycles, are critical for crafting targeted marketing and sales strategies.

**Strategic Recommendations**

* **Prioritise Post Marketing:** Increase budget allocation to Post marketing and innovate with mailshot campaigns to maximise ROI.
* **Enhance Online Marketing:** Boost investment in online channels and employ data-driven digital strategies to strengthen online engagement.
* **Reassess Newspaper Marketing:** Consider reducing investment in newspaper advertising, reallocating resources to more effective channels.
* **Focus on Customer Demographics:** Leverage customer demographic insights for more focused marketing strategies.
* **Continued Data Analysis:** Maintain and expand data analysis efforts to keep strategies relevant and effective.
* **Integrated Marketing Strategies:** Explore combining different channels, especially the synergistic potential between Post and Online marketing.
* **Adapt to Market Changes:** Remain responsive to market shifts and evolving consumer trends for agile marketing strategy adjustments.

**Concluding Thoughts**

Civitas Motors can significantly benefit from these data-driven insights. Aligning marketing strategies with these findings positions the company for enhanced market presence, improved customer satisfaction, and increased sales growth in the competitive automotive sector.