**Data Mining Report**

**Aidan Seaborn**

## Executive Summary

The report is going to be split into two separate sections, one focusing on the Civic Motors dataset and the second will be the Marketing dataset. All sections may contain some insights into the code used but this will be kept to a minimum as this can be accessed within the Jupyter notebook itself.

Within this report there are some key insights into the dataset provided to me, including some statistics regarding the amount of customers within certain postcode ages, the total sales and the average amount of money spent by these customers.

Further analysis will go into the fiscal sales of each year, and the document will finish with some analysis into the marketing document and the return on investment. Throughout each section of this report, I will show my insight the analytics provided, and at the end of the I will be providing a brief conclusion of some insights and recommendations.

**s**

Contents

[Executive Summary 2](#_Toc140595342)

[Data Sources 3](#_Toc140595343)

[Introduction 3](#_Toc140595344)

[What are the age ranges of their customers? 4](#_Toc140595345)

[On average, how frequently are people changing their cars? 6](#_Toc140595346)

[How many customers do they have in the WR10 and WR11 Postcode Areas? & What is the total number of customers within this postcode area? 7](#_Toc140595347)

[The ages of these customers? 8](#_Toc140595348)

[The total value of the vehicles. 11](#_Toc140595349)

[The average value of the vehicles, accounting for outliers. 11](#_Toc140595350)

[What is the level of vehicle sales over a fiscal year? 12](#_Toc140595351)

[Analysis of the Marketing Channel 13](#_Toc140595352)

[Linear Regression 15](#_Toc140595353)

[Polynomial Regression 16](#_Toc140595354)

[Limitations 17](#_Toc140595355)

[What are the age ranges of their customers? 17](#_Toc140595356)

[How many customers do they have in the WR10 and WR11 Postcode Areas? *&* What is the total number of customers within this postcode area? 18](#_Toc140595357)

[The total value of the vehicles. 18](#_Toc140595358)

[Total value of sales in a Fiscal year 19](#_Toc140595359)

[Conclusion 20](#_Toc140595360)

## Introduction

In this report I shall be assuming the role of a Data Analyst, where I shall be manipulating, analysing, and presenting the data in a clear and concise manor. I shall be presenting my findings in a readable manor in which all the key information shall be presented accurately.

There will be some methods not discussed in much detail, such as Data Cleaning but this shall all be available in the Jupyter notebook for inspection as required.

Each section should have an appropriate visualisation to present the data in a way that anyone can understand what it represents, I have tried to use a variety, but I have mostly use bar graphs as personally this displayed the data in a more readable format.

## What are the age ranges of their customers?

Here I calculated the min, max and mean using a variable to store it in for future use, and then plotted them into a graph to present the data in a much more readable format. This will be an ongoing theme and shall only be mentioned briefly going forward.

Graphical user interface, text, application

Description automatically generated

Here we have the output which shows the age ranges across the whole dataset, including the duplicate customers because these are still relevant. The output does not represent this well enough, so I have also included a table.

A picture containing table

Description automatically generated

**Chart

Description automatically generated**

Although the graph clearly shows what age range from 18 to 90, and even though this answers the question more information can be gained by using the .describe method, such as the 25% 50% and 75th% values.

Text

Description automatically generated

Table

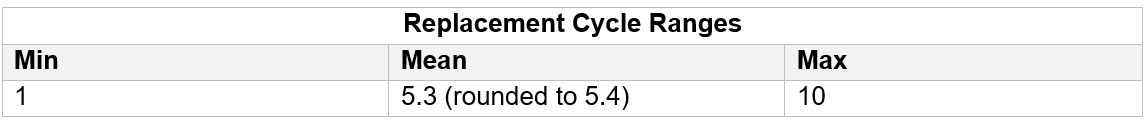
Description automatically generated

## On average, how frequently are people changing their cars?

**Graphical user interface, text, application

Description automatically generated**

As seen above we can see that the output for the code is 1, 5.3 and 10 for the min, mean and max respectively. This did undergo some cleaning / replacements which will be detailed in the reflections section.



**Chart

Description automatically generated**

The results give an insight into how often customers are replacing their cars, with the average being 5.4 years. However, the data itself is slightly inaccurate as some customers have a replacement cycle of 9,8,9 or 10, 4, 1, which does skew the results slightly, but this is still accurate to a degree.

## How many customers do they have in the WR10 and WR11 Postcode Areas? What is the total number of customers within this postcode area?

Here we have isolated two values in the Postcode column and then used the .count method to bring it all together. Then adding both these values together with a separate line to allow for further analysis as the next step is to put them both together.

**Graphical user interface, text

Description automatically generated**

Text

Description automatically generated with low confidence

**Chart

Description automatically generated**

My results show that more customers are based within the WR10 postcode area, with WR10 having 99 total customer and WR11 having only 84. Both combined make up a total of 180 which is only makes up 4.5% of the total business as the dataset contains a total of 4000 individual sales data.

## The ages of these customers?

**Text

Description automatically generated**

Our outputs here are displayed above as 18, 52.05 and 90 for the age range within the WR10 and WR11 postcodes when grouped together.



**Chart

Description automatically generated**

As seen in the graph the customer's ages within this area are between 18, which is our known minimum age and 90 which is our maximum. It has an average age of 52 which again is very similar to our average of the whole database. This shows that both these postcode areas contain a diverse customer’s age range. Although the information gathered here might seem limited it does pair well the graph below showing that overall the entire dataset had a very evenly distributed age range.

Chart, histogram

Description automatically generated

This graph is interactive within the Jupyter notebook (located at the end), using this graph you can more clearly identify the value (quantity) of the customers to their age (index), in this example shown above the age 50, which contains 54 values (customers) across the whole dataset. This could be useful if customers within a certain age range preferred certain cars as this could lead to more personalised advertisements.

## The total value of the vehicles.

**Text

Description automatically generated**

Here we have the total value or retail price of the customers within the WR10 and WR11 postcode areas. The WR10 contains a total of 11,326,533 and WR11 had a total of 9,431,415.

**Calendar

Description automatically generated with medium confidence**

**Chart, bar chart

Description automatically generated**

WR10 contains£11,326,533, whilst WR11 shows £9,431,415 and the total being £20,757,948. This could be directly linked to the fact that the WR10 postcode contains 12 more customers as seen below with the previous summary of total customers in these postcode areas.

## The average value of the vehicles, accounting for outliers.

**Text

Description automatically generated**

The total retail price of cars within WR10 is 63632.20, WR11 is 66416.41 as shown above which I have described again in the graph and bar chart below.

|  |  |
| --- | --- |
| Retail Price (Average) | |
| Wr10 | **WR11** |
| £63632.20 | £66416.41 |

**Chart, bar chart

Description automatically generated**

Although the total sales in WR10 is higher than WR11 here we can see that the average sales price within the WR11 postcode is slightly higher. This shows that on average customers within WR11 are spending £2,784.21 more on their cars than those who are within WR10.

## What is the level of vehicle sales over a fiscal year?

The fiscal year (April to March) is broken down into 4 sections and although the fiscal year of 2020-2021 is not finished I have included this in my results for comparison.

**Table

Description automatically generated**

****

**Chart, bar chart, histogram

Description automatically generated**

The fiscal year from 2017-2018 brought a total of£28**,**491**,**355 to the business which is not even half what 2018-2019 brought in doing more than double with a total of £84,974,133. The current last fiscal year is 2019-2020 as seen above £88,594,555 and 2020-2021 is currently at £58,375,532 although this fiscal year has not finished yet.

## Analysis of the Marketing Channel

Some data could be cleaned up within this dataset as seen below. due to the fact it pulls away slightly from the results as seen on the below boxplot.

**Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated**

However, in the analysis, I kept this in because, in the end, I wanted to represent the data as accurately as possible, and these values made minimal impact on the overall testing of this value.

**Square

Description automatically generated with medium confidence**

Here the heatmap shows the correction for the marketing database on a whole. Using this heatmap we can compare what is labelled on the X-axis to the Y-axis to visually identify what the correlations are by looking at the colours. Brighter colours show positive correlation, whilst darker values show a negative correlation, as seen above Newspaper and Sales have a very negative correlation, with Sales and post showing some positive correlation although not completely.

## Linear Regression

Chart, scatter chart

Description automatically generated

Here we have the linear regression for the Post advertising, the line fits well within the grouping showing a decent link between post advertisement and the total sales which represents a good return on investment for this advertising method.

Chart, scatter chart

Description automatically generated

Here we have the linear regression for the Online advertising, this regression does show some correlation along the line with regards to online advertising but not as strong as seen previously. You can expect a decent correlation between this advertising method and the amount of sales.

Chart, scatter chart

Description automatically generated

Here we have the linear regression for the Newspaper advertising, this has the worst correlation when compared to any of the others, with minimal impact on the sales. This ultimately shows a very poor return on investment.

## Polynomial Regression

Chart, scatter chart

Description automatically generated

Here we have the Polynomial regression for the Post Advertisement in relation to the total Sales, it obviously shows a very high correlation in both this graph and the linear graph and can also be represented in a correlation score as shown below.

Text

Description automatically generated

With the best possible score being 1.0 and the correlation showing as 0.845 this is considered very strong with a definitive link from the amount of money spent in this advertising method and the total sales.

Chart, scatter chart

Description automatically generated

Here we have the Polynomial regression for the Online Advertisement in relation to the total Sales, this again shows a correlation of .23 which is still a positive value although not as significant at Post.

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

Text

Description automatically generated

Here we have the Polynomial regression for the Newspaper Advertisement in relation to the total Sales. The correlation coefficient for this is -40 which is closer to -1. The closer to -1 the stronger the negative linear relationship indicating that an increase in one variable would decrease the other.

## Limitations

First major issue that I faces was although I achieved the tasks, I do believe some of my code could have been achieved with simpler methods. This also has a direct limitation on my analysis of my limitation as I am not sure of other ways in which I could achieve the same results. I also feel that the dataset itself was a limiting factor as the insights provided appear to be very limiting at times, proving slightly difficult to analyse.

### What are the age ranges of their customers?

“Never” is a string and cannot be computed with an integer, and therefore had to be converted to numeric with the following code to allow for computation.

**Graphical user interface, text, application, email

Description automatically generated**

### How many customers do they have in the WR10 and WR11 Postcode Areas? *&* What is the total number of customers within this postcode area?

During this step when trying to use a calculation on the customers within the postcode it would cause an error stating “no data to plot”. This was resolved by creating a dataframe using just these values and then computing.

Graphical user interface, application

Description automatically generated

### The total value of the vehicles.

For some reason whilst calculating the total value of the vehicles even though only the numeric data is here when plotting a graph an error with “no data to plot” would show. This was resolved by forcing it to become an int.

**Chart, bar chart

Description automatically generated**

### Total value of sales in a Fiscal year.

This displays 1e7 in the top left, and when attempting to clear the format intermittently caused an error preventing the graph from loading.

**Chart, bar chart, histogram

Description automatically generated**

## Conclusion

Overall, the customers age range shows that Civic Motors has a very varied customer range with no significant group favouring them. With this analysis it was interesting to note that although more customers are within the WR10 postcode area, and they generate more sales the customers in WR11 spend more on their car on average in comparison.

Based on my analysis of the Marketing document I can say that the best return on investment is Post, closely followed by Online as shown with the correlation coefficient. The worst is Newspaper which appears to have a negative correlation, and therefore I would drop as it brings no return to the business.