Contents

[**1.** **Executive summary** 4](#_Toc127710800)

[**2.** **Introduction** 4](#_Toc127710801)

[**3.** **Background** 5](#_Toc127710802)

[3.1. Background to the case study 5](#_Toc127710803)

[3.2. Structure of the data 5](#_Toc127710804)

[3.3. Transforming/Cleaning the data 5](#_Toc127710805)

[**4.** **Data Analysis and Results** 6](#_Toc127710807)

[4.1. Customer demographics 6](#_Toc127710808)

[4.3. Regional analysis 10](#_Toc127710819)

[4.4. Discussion 11](#_Toc127710821)

[**5.** **Recommendations** 11](#_Toc127710822)

[**References** 11](#_Toc127710823)

[**Appendices** 11](#_Toc127710824)

[Appendix a: Confidence interval calculations for Monetary value 11](#_Toc127710825)

[Appendix b: RFM measures of Central Tendency, Dispersion and Confidence Intervals 12](#_Toc127710826)

[Appendix c: Amount purchased, Sum of Total and Average of Total Price per Month 12](#_Toc127710827)

[Appendix d: Sales by Region and Month in 2020 12](#_Toc127710828)

[Appendix e: Calculations to determine RFM segment eligibility criteria 13](#_Toc127710829)

[Appendix f: RFM Segment eligibility criteria 13](#_Toc127710830)

[Appendix g: RFM value criteria to determine customer segment 13](#_Toc127710831)

[Figure 1: Data fields, types and quality 6](#_Toc127710275)

[Figure 2: Customer demographics by age, region and gender. 6](#_Toc127710278)

[Figure 3: London customer age distribution. 7](#_Toc127710280)

[Figure 4: Midlands customer age distribution. 7](#_Toc127710282)

[Figure 5: Customer segmentation description and criteria. 8](#_Toc127710283)

[Figure 6: Spread of customers in each segment for R, F and M values. 8](#_Toc127710284)

[Figure 7: Average R, F, M values per Customer Segment. 9](#_Toc127710285)

[Figure 8: Average customer age per segment by gender and region. 9](#_Toc127710286)

[Figure 9: Relationship between customer frequency and total sales. 10](#_Toc127710287)

[Figure 10: Sales by Region and Month in 2020. 11](#_Toc127710289)

# **Executive summary**

This analysis has been conducted on behalf of Wirkkala using sample data in 2020 from both stores in London and Midlands. This report aims to use statistical and analytical techniques to determine whether the data is representative of the perspective region and general population in order to empower data-driven managerial decision-making. Customer relationship management technique, RFM analysis, was conducted in order to categorise customers into four segments based on recency, frequency, and monetary value. Customers were divided into four categories applying customer relationship management strategies based on recency, frequency, and monetary scores. Each segment eligibility criteria were developed using measures of dispersion and central tendency. Wirkkala can make more accurate data-driven judgements by comprehending the tendencies and fluctuations of these distinct client categories.

To deal with the data before processing, appropriate data cleaning and transforming procedures were applied, all of which are disclosed in this report. Descriptive and statistical analytic approaches were utilised to acquire insights from this dataset, as well as to create data visualisation models to assist organisation managers; incorporate vital data to influence managerial strategic choices on their business operations, answer their key research questions, and aid in building future expanding opportunities. Based on the study, appropriate and relevant recommendations for addressing the most effective development of a marketing campaign were developed, with a strong emphasis on the age, gender, and geographical factors between the customer segments.

# **Introduction**

This report was created on Wirkkala's behalf utilising 2020 sample data from both shop locations (London and Midlands). The purpose of this report is to employ statistical and analytical approaches to establish if the data is reflective of the respective region and population as a whole in order to empower data-driven management decision-making. RFM analysis, a customer relationship management approach, was applied to divide clients into four groups based on recency, frequency, and monetary score. Analysis of the differences between them, which were presented simultaneously this paper, augment the findings and include an interactive digital dashboard.

# **Background**

* 1. Background to the case study

A growing number of consumers around the world are switching to a plant-based diet or introducing plant-based products (PBPs) into their meals (Bord Bia, 2018; Mintel, 2020a). As a result, the plant-based products market sector is gaining significance and is a potentially profitable industry for both new entrants and existing plant-based food and beverage makers (Peschel et al., 2019). In 2018, it was predicted that the European market, with the United Kingdom (UK) holding the highest sub-market share, would expand from €1.5 billion to €2.4 billion by 2025 (Deloitte, 2019). As the juice market in the United Kingdom is more and more developing, Wirkkala has full potential to create and maintain strong relationships with customers with an in-dept and informative analysis so as to launch a new marketing strategy. With the sample dataset, the company is able to acquire a portfolio of their customers in both areas, London and Midlands, which supports them in classifying buying behaviours and tendencies as they are the most effective components that contribute to a higher profit margin.

* Research questions:
* How responsive each customer segment is to the store products?
* Can the sample dataset be the representative for a whole population?
* What are the differences between the sales and customer commitment between London and Midlands?
  1. Structure of the data

This sample's data is entirely organised, enabling for a more efficient analysis approach. New sheets, such as the interactive Digital Dashboard and RFM analysis, were included to undertake analysis. Some existing data has been merged to enable for more in-depth analysis, such as combining customer and order data to allow for a regional comparison of yearly sales (find: Figure 10 sheet on Excel document). Data from the client datasheet was sorted and extracted by region to enable for regional comparison (find: Figure 3 sheet and Figure 4 sheet).

* 1. Transforming/Cleaning the data

Since there is an amalgam of American and United Kingdom birthdates format types in the Customer dataset, the column was converted into UK birthdates format. Moreover, age data was calculated from birthdate and added as a column in the Customer dataset. In the inventory dataset, the price dataset and number in inventory dataset were combined in order to calculate the total value of Wirkkala’s inventory. On the other hand, total price data was created by utilising the generated total value column and discount one in orders dataset.

For the dataset, completeness analysis was executed to assess data quality. In the Title (blank values: 79, 27.9%) and First Name (blank values: 34, 10.7%) columns, there was insufficient data. Moreover, because of the inconsistence of the First Name data including both full names and initials, the uniqueness factor was low. Both Title and First Name column were elected to be hidden from the dataset and eliminated from the study.  The reasoning behind this is that the data fields themselves do not contribute business value because other data fields to identify customers, such as surname and gender, have been utilised, and removing these fields would significantly enhance data quality and completeness.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data** | **Type** | **Number of values** | **Number of blank values** | **Completeness (%)** |
| Customer Account Number | Ratio data | 283 | 0 | 100% |
| Title | Nominal data | 204 | 79 | 72.1% |
| First Name | Nominal data | 283 | 34 | 89.3% |
| Last Name | Nominal data | 283 | 0 | 100% |
| Birthdate | Interval data | 283 | 0 | 100% |
| Gender | Nominal data | 283 | 0 | 100% |
| Postal Area | Interval data | 283 | 0 | 100% |
| Region | Nominal data | 283 | 0 | 100% |
| Age | Ratio data | 283 | 0 | 100% |
| Purchase date | Interval data | 1226 | 0 | 100% |
| Stock code | Nominal data | 1226 | 0 | 100% |
| Quantity | Ratio data | 1226 | 0 | 100% |
| Discount voucher | Ratio data | 1226 | 0 | 100% |
| Description | Nominal data | 23 | 0 | 100% |
| Price | Ratio data | 23 | 0 | 100% |
| Number in inventory | Ratio data | 23 | 0 | 100% |

Figure 1: Data fields, types and quality

# **Data Analysis and Results**

* 1. Customer demographics

Figure 2 demonstrates a snapshot of the interactive graph on the digital dashboard (see Excel document: Digital Dashboard), displaying general consumer demographics with age, location, and gender as variables. The 50-59 age group is the most distinct in both areas, with a relatively greater number in the Midlands than in London.

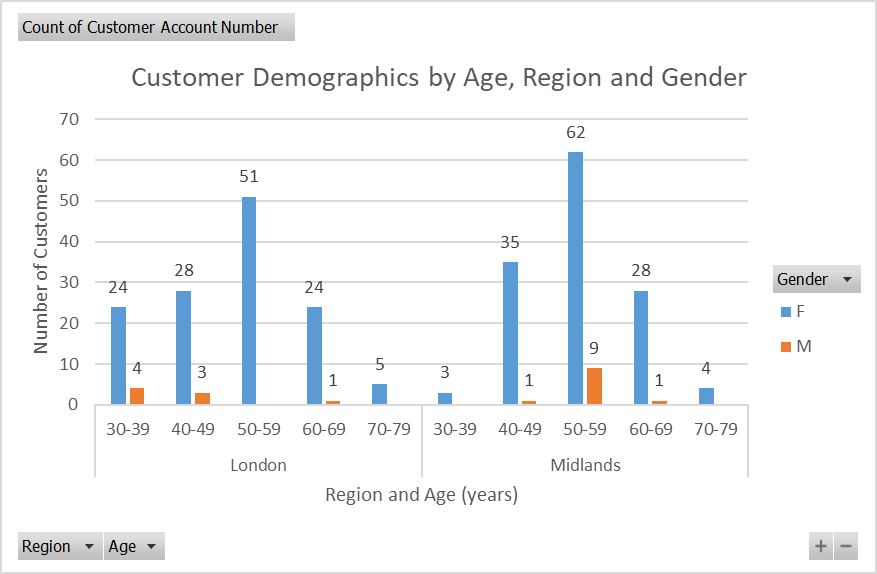


Figure 2: Customer demographics by age, region and gender.

Figure 3 and 4 below further illustrate the age distribution of customers in the separate areas. In the Midlands, two outliers were identified: 33 (<1.5\*IQR below Q1) and 78 (>1/5\*IQR above Q3). The IQR shows that 50% of customers in London are between 42 – 59 years, compared to 49 - 59 years in the Midlands.

|  |  |
| --- | --- |
| Figure 3: London customer age distribution. | Figure 4: Midlands customer age distribution. |

* 1. **Customer Segmentation**

In this section, there are three criteria which were considered in order to classify customers into 4 value segments: recency (max purchased date), frequency (count of purchase date) and monetary value (sum of total price).

The four customer segments were described as below:

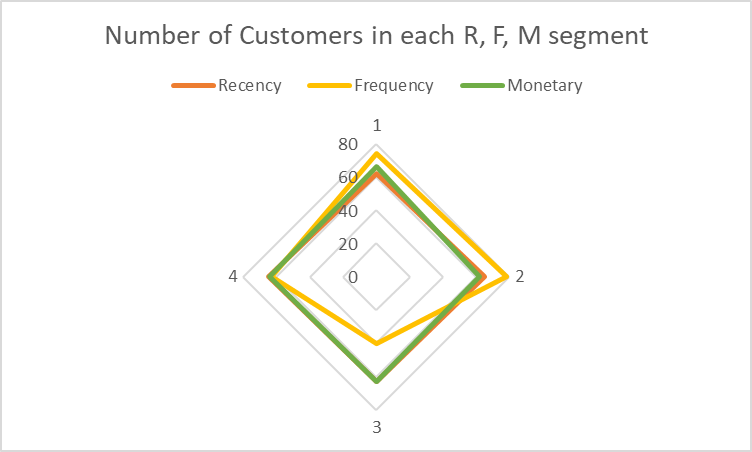
* Segment 4, Loyal customer: high priority, profitable, strong bargaining power, regular purchasers.
* Segment 3, Satisfied customer: medium priority, organisational interest
* Segment 2, Irregular customer: medium priority, irregular customer with lower sales contribution
* Segment 1, Disengaged customer: low priority, emphasis on re-attracting more, single shop purchasers

Measures of dispersion and central tendency for each criterion were applied to derive segment classifications for each R, F, and M. (Excel functions: Min, Q1, Median, Q3, Max). Figure 5 depicts the following segment criteria for each variable.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Customer Segment** | **R** | **No. of customer** | |  | **No. of customer** | |  | | **No. of customer** | |
| **Lon** | **Mid** | **F** | **Lon** | **Mid** | | **M** | **Lon** | **Mid** |
| ***1*** | Purchased more than 116 days | 34 | 28 | Purchased less than 4 times within the month | 36 | 14 | | Spent less than £8.86 within the month | 33 | 33 |
| ***2*** | Purchased between 55 and 116 days | 34 | 31 | Purchased 4 times within the month | 36 | 66 | | Spent between £8.86 and £12.5 within the month | 28 | 34 |
| ***3*** | Purchased between 22 and 54 days | 25 | 38 | Purchased 5 times within the month | 27 | 25 | | Spent between £12.6 - £19.29 within the month | 38 | 25 |
| ***4*** | Purchased within the last 21 days | 34 | 31 | Purchased more than 5 times within the month | 28 | 23 | | Spent more than £19.29 within the month | 28 | 36 |

### Figure 5: Customer segmentation description and criteria.

Figure 6 depicts the data from the preceding table. There is a noticeable table discrepancy in "frequency," with a larger client count in the bottom segments (1 and 2) and a low customer count in segment 3.



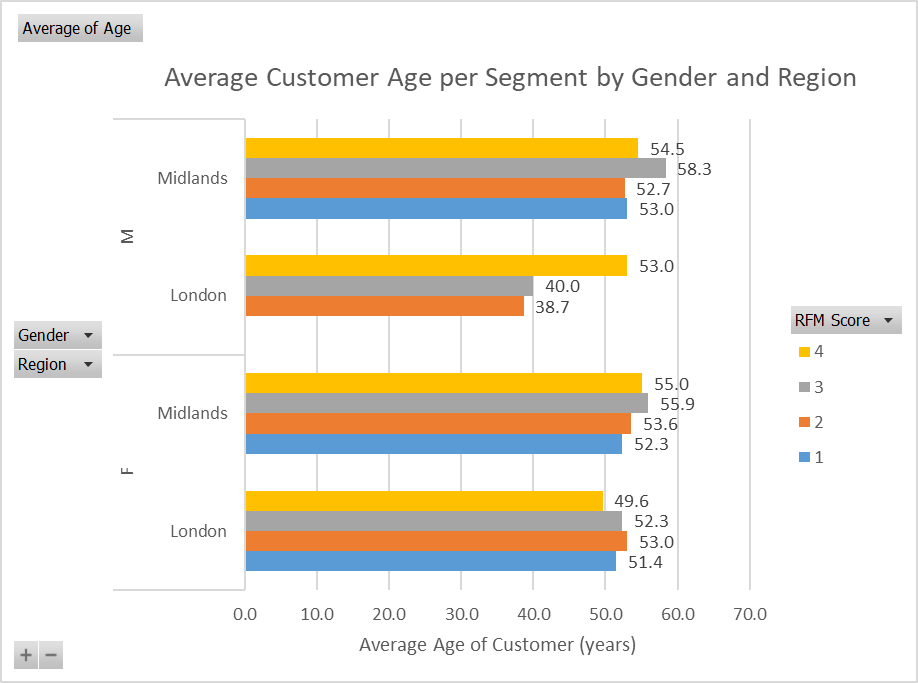
### Figure 6: Spread of customers in each segment for R, F and M values.

Figure 7 depicts the typical customer in each sector's average date since the previous purchase, average number of purchases, and average total yearly expenditure to offer a holistic perspective of the average customer in each category.

|  |  |  |  |
| --- | --- | --- | --- |
| **Customer Segment** | **Average of R (Max Purchase Date)** | **Average of F (Count of Purchase Date)** | **Average of M (Sum of Total Price)** |
| *1* | 25/05/2020 | 2.38 | 6.62 |
| *2* | 09/10/2020 | 4.00 | 11.30 |
| *3* | 24/11/2020 | 5.00 | 15.16 |
| *4* | 21/12/2020 | 8.54 | 29.21 |
| **Grand Total** | **06/10/2020** | **4.81** | **15.54** |

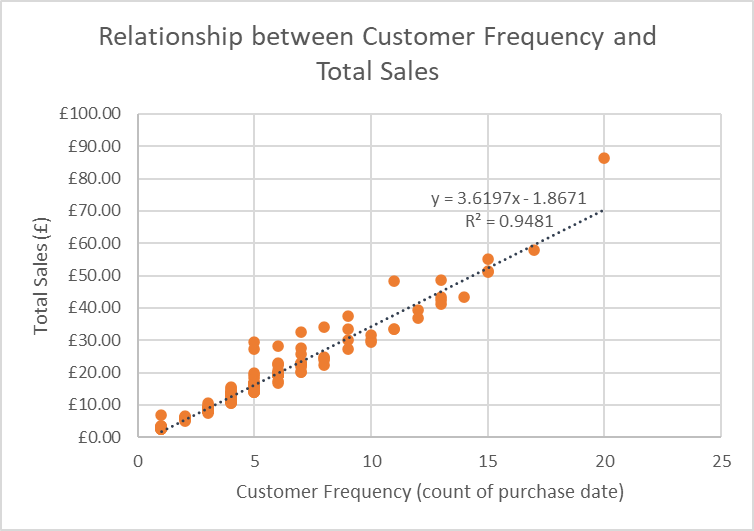
### Figure 7: Average R, F, M values per Customer Segment.

The figure below, Figure 8, demonstrates the average age of buyers in each of the four divisions. The average age has also been divided by gender and area.



### Figure 8: Average customer age per segment by gender and region.

Figure 9 demonstrates a strong positive correlation between customer frequency and total sales (r = 0.97).



### Figure 9: Relationship between customer frequency and total sales.

* 1. Regional analysis

Figure 10 displays relative regional sales statistics to enable managers to make decisions. The two areas show comparable monthly variations, although the highest and lowest sales months are dissimilar months. Wirkkala store experienced its highest revenue in both September and December and lowest in July in London, likewise, the store in Midlands witnessed the highest sales in December and lowest in July.

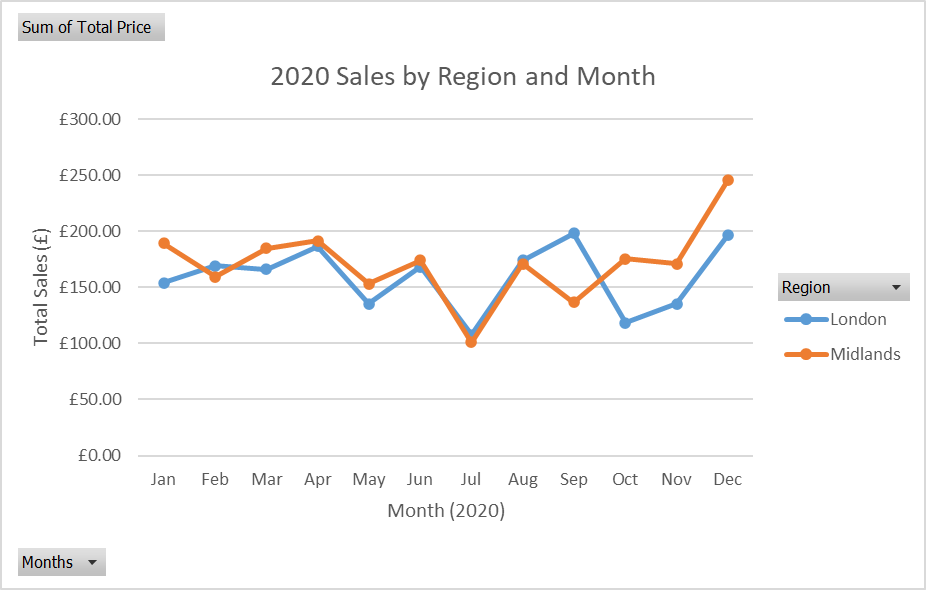


Figure 10: Sales by Region and Month in 2020.

* 1. Discussion

Wirkkala can use confidence intervals to help answer their research question “Can the sample dataset be the representative for a whole population?”. Wirkklala can infer at the 95% confidence interval (refer to appendix b) that by utilising the sample data as a core principle for analysis, the actual population mean of sales data per segment is between:

* £6.05 – £7.19 (customer segment 1)
* £11.08 – £11.52 (customer segment 2)
* £14.77 – £15.54 (customer segment 3)
* £26.14 - £32.38 (customer segment 4)

Wirkkala can predict future sales margins based on how many purchasers they can transfer from a lower segment to a higher segment utilising these margins.

Wirkkala prides themselves on placing a high priority on frequent customers, this is extremely apparent when comparing RFM scores between segments. The segmentation criteria were not influenced by outliers because the median was adopted as the mid-point reference rather than the mean. The business value that these top consumers contribute can be utilised to drive managerial decisions to target these groups even more.

# **Recommendations**

Because customer frequency and total sales have been justified to have a significant positive linear association (r = 0.97), automated advertising campaigns can be delivered through email to customers who meet a specific time without visiting a site. For instance, if a customer does not make a purchase for 6 months, a unique discount code will be supplied for the customer to use with benefits if the voucher is redeemed on time.

The Midlands store generates £143,3 greater sales than the London store. With London's larger population density as comparison to the Midlands, Wirkkala might consider increasing its marketing and analytics expenditure for the London retail site.

Confidence intervals, as suggested in the conclusion, have the potential to drive data-driven decisions. If the marketing effort is effective, Wirkkala can apply expected mean sales data for the population to determine how many consumers will satisfy the next eligibility condition up from their present segment and hence forecast future sales income.

It can be seen that there is no customer under the age of 30 were included in this dataset sample. It is ambiguous if this is attributable to a shortage of under-30s buying at Wirkkala or a flaw in the sample method used. Notwithstanding, the under-30-adult population (18-29) accounts for a sizable proportion of the overall UK population, so Wirkkala should look to target this younger demographic by conducting market research and including market segments that appeal to the younger generation in order to expand their customer source.

Because some consumers did not satisfy the lowest RFM requirement and so were not included in any customer segment. Given that the average Wirkkala customer spends £3.23 per visit, it is advised that the business target those customers via email and phone advertisements while using social media that personalises the interest and tendency to buy in order to attract them to visit and purchase that item at the stores.

# **References**

Bord Bia. (2018). Dietary lifestyles report. Bord Bia. <https://www.bordbia.ie/globalassets/bordbia2020/industry/insights/consumer-insights/dietary-lifestyles-report-november-2018.pdf>.

Deloitte. (2019). Plant-based alternatives. https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consumer-business/deloitte-uk-plant-based-alternatives.pdf.

Mintel. (2020a). Veganism in 2020 – Mintel analysts’ outlook.

Peschel, A. O., Kazemi, S., Liebichova, M., Sarraf, S. C. M., & Aschemann-Witzel, J. (2019). Consumers’ associative networks of plant-based food product communications. Food Quality and Preference, 75, 145–156.

# **Appendices**

## Appendix a: Confidence interval calculations for Monetary value

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Segment** | **Average of M (Sum of Total Price)** | **Var of M (Sum of Total Price)** | **StdDev of M (Sum of Total Price)** | **Confidence Interval** | **Upper C.I.** | **Lower C.I.** |
| 1 | £6.62 | £5.68 | £2.37 | 0.571 | £7.19 | £6.05 |
| 2 | £11.30 | £0.82 | £0.90 | 0.223 | £11.52 | £11.08 |
| 3 | £15.16 | £2.48 | £1.56 | 0.385 | £15.54 | £14.77 |
| 4 | £29.21 | £159.59 | £12.53 | 3.071 | £32.28 | £26.14 |
| **Grand Total** | **£15.54** | **£114.01** | **£10.66** |  |  |  |

## Appendix b: RFM measures of Central Tendency, Dispersion and Confidence Intervals

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Recency Score** | **Frequency Score** | **Monetary Score** |
| *Mean* | 2.51 | 2.36 | 2.49 |
| *Standard Error* | 0.07 | 0.07 | 0.07 |
| *Median* | 3 | 2 | 2 |
| *Mode* | 4 | 2 | 1 |
| *Standard Deviation* | 1.12 | 1.14 | 1.13 |
| *Sample Variance* | 1.25 | 1.31 | 1.27 |
| *Kurtosis* | -1.36 | -1.35 | -1.38 |
| *Skewness* | -0.01 | 0.26 | 0.01 |
| *Range* | 3 | 3 | 3 |
| *Minimum* | 1 | 1 | 1 |
| *Maximum* | 4 | 4 | 4 |
| *Sum* | 641 | 602 | 635 |
| *Count* | 255 | 255 | 255 |
| *Largest(1)* | 4 | 4 | 4 |
| *Smallest(1)* | 1 | 1 | 1 |
| *Confidence Level (95.0%)* | 0.14 | 0.14 | 0.14 |

## Appendix c: Amount purchased, Sum of Total and Average of Total Price per Month

|  |  |  |  |
| --- | --- | --- | --- |
| **Month** | **Count of Purchase Date** | **Sum of Total Paid** | **Average of Total Paid** |
| Jan | 109 | £358.98 | £3.29 |
| Feb | 106 | £326.39 | £3.08 |
| Mar | 107 | £335.03 | £3.13 |
| Apr | 120 | £393.57 | £3.28 |
| May | 89 | £278.79 | £3.13 |
| Jun | 105 | £335.60 | £3.20 |
| Jul | 68 | £239.42 | £3.52 |
| Aug | 101 | £321.01 | £3.18 |
| Sep | 107 | £352.15 | £3.29 |
| Oct | 95 | £304.88 | £3.21 |
| Nov | 100 | £338.40 | £3.38 |
| Dec | 119 | £374.79 | £3.15 |
| **Grand Total** | **1226** | **£3,959.00** | **£3.23** |

## Appendix d: Sales by Region and Month in 2020

|  |  |  |  |
| --- | --- | --- | --- |
| **Sum of Total Price** | **Column Labels** |  |  |
| **Row Labels** | **London** | **Midlands** | **Grand Total** |
| *Jan* | £153.86 | £189.03 | £342.89 |
| *Feb* | £169.07 | £159.15 | £328.21 |
| *Mar* | £166.12 | £184.77 | £350.88 |
| *Apr* | £186.35 | £191.61 | £377.96 |
| *May* | £135.41 | £153.34 | £288.74 |
| *Jun* | £168.13 | £174.00 | £342.13 |
| *Jul* | £107.89 | £100.91 | £208.79 |
| *Aug* | £173.95 | £171.09 | £345.03 |
| *Sep* | £198.25 | £136.76 | £335.01 |
| *Oct* | £118.55 | £175.41 | £293.96 |
| *Nov* | £135.33 | £171.12 | £306.45 |
| *Dec* | £196.31 | £245.34 | £441.65 |
| **Grand Total** | **£1,909.19** | **£2,052.49** | **£3,961.68** |

## Appendix e: Calculations to determine RFM segment eligibility criteria

|  |  |  |  |
| --- | --- | --- | --- |
| **Segment criteria calculations** | **R (date)** | **F** | **M** |
| *Min* | 15/01/2020 | 1 | £2.50 |
| *Q1* | 06/09/2020 | 3 | £8.85 |
| *Median* | 07/11/2020 | 4 | £12.50 |
| *Q3* | 10/12/2020 | 5 | £19.29 |
| *Max* | 31/12/2020 | 20 | £86.47 |

## Appendix f: RFM Segment eligibility criteria

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria** | **R (date)** | **F (times)** | **M (total spent)** |
| *1* | 15/01/2020 - 06/09/2020 | 1 - 3 | < £8.86 |
| *2* | 07/09/2020 - 07/11/2020 | 4 | £8.86 - £12.5 |
| *3* | 08/11/2020 - 10/12/2020 | 5 | £12.51 - £19.29 |
| *4* | 11/12/2020 - 31/12/2020 | >=6 | > £19.29 |

## Appendix g: RFM value criteria to determine customer segment

|  |  |
| --- | --- |
| **Segment** | **RFM average criteria** |
| *1* | 1 - 1.67 |
| *2* | 1.68 - 2.33 |
| *3* | 2.34 - 3.33 |
| *4* | 3.34 - 4 |

Appendix h: Measures of dispersion and central tendency for age by region

|  |  |  |
| --- | --- | --- |
|  | **London** | **Midlands** |
| *Min* | 30 | 33 |
| *Q1* | 42.35 | 49 |
| *Median* | 52 | 54 |
| *Mean* | 50.9 | 54.4 |
| *Q3* | 58.75 | 59 |
| *Max* | 76 | 78 |
| *IQR* | 16.5 | 10 |
| *IQR\*1.5* | 24.75 | 15 |
| *Upper limit* | 83.5 | 74 |
| *Lower limit* | 17.5 | 34 |