

MSc Management & Business Analytics   
   
BUSI49150: Introduction to Business Analytics & Intelligence

Formulating a data-driven marketing campaign for Wirkkala juice bar informed by customer data and demographics.  
   
  
Word count: 2457

Submitted on 20/02/2022

By Robbie Rogers, Deeksha Deepak, Haris Anwar and Sharon Nnenna Achebe

Table of Contents

[1 Introduction 3](#_Toc96114858)

[2 Background 5](#_Toc96114859)

[2.1 Research questions 5](#_Toc96114860)

[2.1.1 KPI’s for Wirkkala 6](#_Toc96114861)

[2.1.2 Structure of the data 6](#_Toc96114862)

[3 Transforming and cleaning the data 7](#_Toc96114863)

[4 Data analysis and results 9](#_Toc96114864)

[4.1 Customer demographics 9](#_Toc96114865)

[4.2 Customer segmentation 10](#_Toc96114866)

[4.2.1 Segment 4 11](#_Toc96114867)

[4.2.2 Segment 3 11](#_Toc96114868)

[4.2.3 Segment 2 11](#_Toc96114869)

[4.2.4 Segment 1 11](#_Toc96114870)

[5 Conclusions 15](#_Toc96114871)

[6 Recommendation. 17](#_Toc96114872)

[7 APPENDIX 18](#_Toc96114873)

[APPENDIX A: RFM Measures of Central Tendency, Dispersion and Confidence Intervals 18](#_Toc96114874)

[APPENDIX B: Confidence Interval Calculations for Monetary Value. 18](#_Toc96114875)

[APPENDIX C: Amount Purchased, Sum of Total and Average of Total Paid per Month. 19](#_Toc96114876)

[APPENDIX D: 2019 Sales by Region and Month 19](#_Toc96114877)

[APPENDIX E: Most Purchased Items 20](#_Toc96114878)

[APPENDIX F: Calculations to Determine RFM Segment Eligibility Criteria 20](#_Toc96114879)

[APPENDIX G: RFM Segment Eligibility Criteria 20](#_Toc96114880)

[APPENDIX H: RFM Value Criteria to Determine Customer Segment 21](#_Toc96114881)

[APPENDIX I: Measures of Dispersion and Central Tendency of Age by Region. 21](#_Toc96114882)

[8 References 22](#_Toc96114883)

Executive summary

The following report aims to analyse the Wirkkala juice bar customer transactions dataset. The data itself contains a sample of the London and Midlands customer population comprised of orders, inventory and customer demographics. Throughout the report, a customer relationship management technique was undertaken to segment the sample based on recency, frequency and monetary value. Furthermore, measures of dispersion and central tendency were taken, which were utilised to segment the customers into four different segments based on value. The purpose of doing so was to inform data-driven decisions to maximise profit fundamentally.

Data was cleaned and transformed under the six data quality dimensions to add value to the business and ease of analysis. Various descriptive and statistical analytic techniques were employed to formulate insights within the dataset, which then help build visual models of the data to aid management to analyse the data. By doing so, the analysis is used to answer the research question formulated and allow for management to monitor the key performance indicators. Recommendations were then given to Wirkkala to help lay out steps that should be undertaken within their new marketing campaign that would best effectively target desired segments.

# Introduction

The following report summarises and analyses the Wirkkala customer transactions dataset for two regions, London and the Midlands. The report will draw upon a customer relationship management technique known as RFM analysis and other various statistical and analytical methods. Within the customer transactions dataset, data types such as nominal, ordinal, interval, and ratio represent customer demographics and orders. The segments established from the analysis will then enable the Wirkkala juice bar to make data-driven managerial decisions based upon cold hard data to inform a targeted marketing campaign.

Alongside this report, a spreadsheet that has had its data cleaned under the data quality dimensions suggested by Azeroual, Saake and Abuosba (2019) has also been submitted. It includes the cleaned and formatted data as well as an interactive digital dashboard to aid with graphical analysis and data visualisation.

# Background

The fruit juice industry within the UK is dynamic, containing mixed perceptions and health guidance. From 2015 to 2020, its total market value has grown around £200 million (Passport, 2020). That being said, due to increasing awareness and recommendations of reducing sugar intake, the juice industry is expected to see the volume of total sales dramatically drop (figure 1) from 1,626 million litres in 2019 to 1,516 million litres in 2024 (Mintel, 2020). As a result, companies are being forced to advertise their drinks as containing less sugar to combat industry trends (Kathie Canning, 2018).

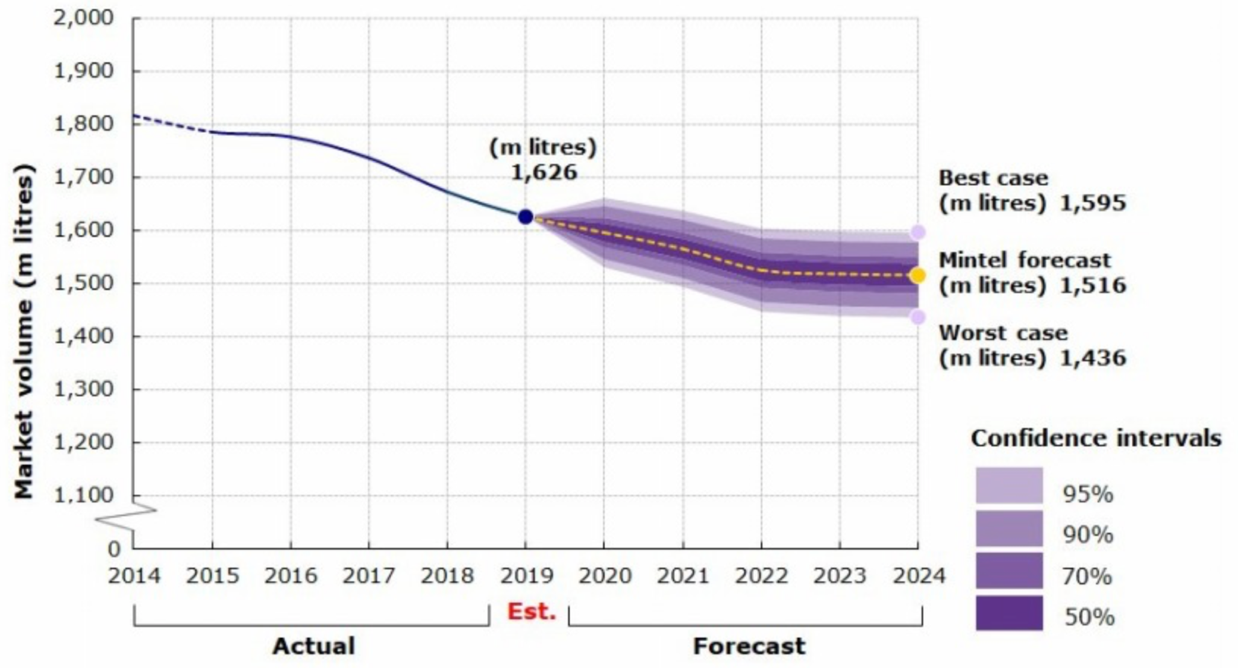
****

Figure 1- predicted volume of fruit juice sales in litres

Wirkkala wants to connect with customers uniquely. The firm prides itself on its quality service, and places great emphasises on holding a strong relationship with its customers, especially on customer retention and repeat purchases. The current task that has been raised is to create a new advertising campaign for its existing customers to boost sales once the pandemic restrictions end. As Chahal, Jyoti and Wirtz (2019, p3) suggest, the use of business analytics to aid data-driven decision making within marketing enables "organisations to improve profitability, increase market share and revenue".

## Research questions

* How confident can Wirkkala be that the sales for each sample reflect that of the population?
* Which of the customer segments identified would be most responsive to the advertising campaign?
* What are the differences in the value of sales and customer recency and frequency between the two regions?

### KPI’s for Wirkkala

Regarding Wirkkala's vision and customer values, the key performance indicator for growth and success will be ensuring customer retention and repeat purchases. The primary method of analysing this would be through the total sales and their correlation to the frequency of customer purchases.

### Structure of the data

The customer transaction dataset is presented as structured data. Given it is in this form, it allows analysis to be conducted with ease and efficiency. Within the excel document, sheets 1-3 contain the cleaned datasets and then sheets 4 and 5 include the RFM analysis and a digital dashboard, respectively.

The quality of data initially provided was low and did not meet the six measures for data quality. The measures are as follows: completeness, uniqueness, timeliness, validity, accuracy and consistency. Some of these dimensions were worse met than others. The following section looks at how these six dimensions were met and the steps undertaken in cleaning and improving the data.

# Transforming and cleaning the data

The first step undertaken in cleaning the data was addressing the consistency within the customer dataset, specifically, the inconsistency of customers' dates of birth. To rectify this, a text to columns function was employed. The purpose of this function is to format all dates into the DMY format as opposed to a mixture of DMY and MDY. From the corrected birthdates, an age collum was added calculated based on the date of 31/12/2021. Then a pivot table was created, grouping these ages in intervals of 9 years.

Within the customer's orders dataset, a VLOOKUP function was employed to look within the inventory dataset and transform the price for each transaction into the customer's orders dataset. Doing so allowed the total value paid column to be created. The column was calculated based on the price of an item multiplied by the quantity and the discount factor.

Also, cleaning was undertaken within the inventory dataset. It was noted that there was a duplicate stock code. To rectify this, the inventory values were combined between the two, and the two different prices were averaged under the same stock code.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data name** | **Data type** | **Number of values** | **Number of blank values** | **Completeness** |
| Title | Nominal | 6 | 79 | 72% |
| First name | Nominal | 100 | 34 | 88% |
| Last name | Nominal | 261 | 0 | 100% |
| Customer account no | Ratio | 284 | 0 | 100% |
| Date of birth | Interval | 280 | 0 | 100% |
| Age | Ratio | 43 | 0 | 100% |
| Gender | Nominal | 2 | 0 | 100% |
| Postal area | Interval | 80 | 0 | 100% |
| Region | Nominal | 2 | 0 | 100% |
| Stock code | Interval | 24 | 0 | 100% |
| Price | Ratio | 8 | 0 | 100% |
| Quantity | Ratio | 4 | 0 | 100% |
| Discount voucher (%) | Ratio | 5 | 0 | 100% |
| Purchase date | Interval | 352 | 0 | 100% |
| Total value paid | Ratio | 53 | 0 | 100% |
| Description | Nominal | 23 | 0 | 100% |
| Number in inventory  Figure 2 – completeness analysis conducted on all data fields | Ratio | 15 | 0 | 100% |

To assess the quality of data, a completeness analysis was undertaken (figure 2). From doing so, two data types were discovered to be incomplete, the titles and first name column, having 79 and 34 blank values, respectively. This provided a 72.08% completeness for titles, and for first names, it provided an 87.99% completeness. Secondly, first names were presented in various formats, which raised the issue of the data quality dimension of uniqueness. Given the incompleteness and low % uniqueness, the titles and first names column have been hidden and excluded from further analysis. The reasoning behind this boils down to these fields not adding any value to the business as customers can already be uniquely identified by account number. Furthermore, these fields do not allow further segmentation/marketing that the other fields already provide. The removal of these fields ultimately improves the data quality and completeness.

# Data analysis and results

## Customer demographics

Figure 3 – customer demographics

The demographics of general customers are depicted in figure 3 with age, region, and gender as variables. In both regions, the 50-59 age group appears to be the most dominant, with slightly higher numbers in the Midlands than in London. Moreover, there is a clear indication that female customers outnumber male customers.

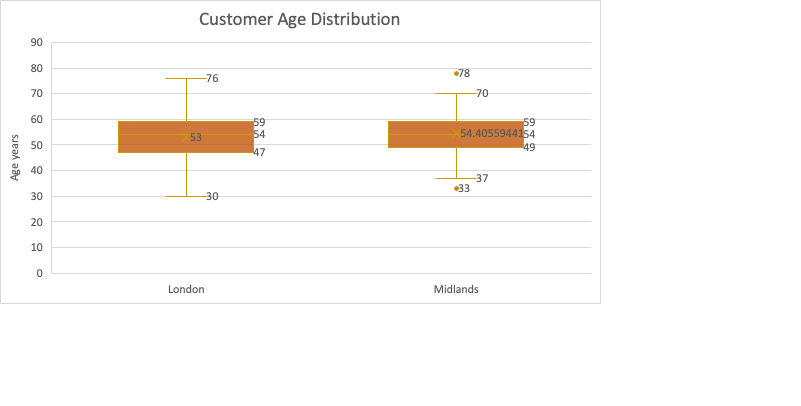


Figure 4 – age distribution

As show above in figure 4, the graphs depict the age distribution of customers in the respective regions. Two outliers have been detected in the Midlands region: 33 (1.5\*IQR below Q1) and 78 (>1.5\*IQR above Q3). According to the IQR, 50 percent of customers in London are between the ages of 48 and 60, while 50 percent of customers in the Midlands are between the ages of 50 and 60.

## Customer segmentation

Following data analysis gathered and cleaned up, the customer segmentation would be categorized using the Recency, Frequency and Monetary Value analysis.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Number of customers** | |  | **Number of customers** | |  | **Number of customers** | |
| **Segment** | **Recency** | **Lon** | **Mid** | **Frequency** | **Lon** | **Mid** | **Monetary value** | **Lon** | **Mid** |
| 1 | Purchased 3 months or more. | 50 | 45 | Visited between 0-2 times. | 14 | 16 | Spent 0-9 pounds | 31 | 14 |
| 2 | Purchased between 2 months but less than 3 months | 12 | 31 | visited between 3-5 times. | 87 | 80 | Spent 9-19 pounds | 70 | 14 |
| 3 | Purchased between a month and less than 2 months | 22 | 18 | visited between 6-8 times | 18 | 21 | Spent 20-29 pounds | 14 | 62 |
| 4 | Purchased less than a month | 43 | 34 | visited 9 or more times | 8 | 11 | Spent <=30 pounds | 8 | 38 |

Figure 5 – customer distribution within the segments

Shown above, figure 5 categorically explains where key customer strength lies, our customer segmentation would be dilapidated into:

### Segment 4

Premier Customers: These categories of individuals are placed as high priority by Wirkkala, they are categorized based on most recently visited, frequent purchases and high spenders, (Segment 4 category).

The Average number of customers in this region is:

* 22 customers in London
* 33 Customers in the Midlands

### Segment 3

Satisfied Customers: The satisfied Customers are classified under medium priority, key focus on retention, organizational interest, grouped under segment 3 of this data analysis, categorized by low quantified recent visits, minimal frequency visits to the store and an average/healthy spending capacity at Wirkkala store. The Average number of customers in this region is:

* 41 Customers in London
* 58 Customers the Midlands

### Segment 2

Margin Customers: Key analysis for this segment is medium priority, infrequent customer with lower sales contribution. Grouped under “Segment 2” categorized by minimal store visit, minimal recency grade and an average/minimal spending capacity at Wirkkala store. The Average number of customers in this region is:

* 33 Customers in London
* 18 Customers in the Midlands

### Segment 1

Disengaged Customers: This segment focuses on the low priority, single shop buyers. This set are the least beneficial customers to Wirkkala formed by analyzation of low store visits and very low budget for spending at the store. The Average number of customers in this region is:

* 29 Customers in London
* 20 Customers in the Midlands

Figure 6 – customer distribution within the 4 segments

The data from figure 5 is visualised in the spider diagram above (figure 6). As frequency leans toward segment 2, a distinct gap emerges. Segment 2 has a larger customer count, whilst segments 1 and 3 have relatively low customer counts.

Figure 7 – customer frequency and total sales

The scatter graph above (figure 7) illustrates that there is a positive correlation relationship between customer frequency and total sales (R = 0.92).

Figure 8 -sales by region and month

The graph above (figure 8) shows comparative regional sales statistics to help managers make better decisions. The two regions have similar monthly fluctuations, but their highest and lowest sales months are distinct. Most of the differentiation between the two regions begins September onwards as London branch hits it’s revenue at 200, which is also the greatest revenue for the London branch, with the lowest coming in July. In December, the Midlands region had the highest revenue, and in July, it had the lowest, which was identical to London's.

Figure 9 – average age of customers in each segment

The chart above (figure 9) shows the average age of customers in each of the four segments. The average age was also broken down by gender and area.

|  |  |  |  |
| --- | --- | --- | --- |
| Customer segment | Average of M (Sum of  Total Paid) | Average of F (Count of  Purchase Date) | Average of R(Max of purchase date) |
| 1 | 6.87 | 1.2 | 24/06/2019 |
| 2 | 14.27 | 3.95 | 14/10/2019 |
| 3 | 23.73 | 6.48 | 14/11/2019 |
| 4 | 43.12 | 12.1 | 17/12/2019 |
| Total | 87.99 | 23.73 | 09/10/2019 |

The table above provides a holistic view of the average customer in each segment. The figure represents the average total annual spend, average number of purchases, and average date since last purchase for the average customer in each segment.

# Conclusions

# Wirkkala has placed great importance on the “Premier Grade Customers”. From our analysis we can identify a clear case of Pareto categorization, clearing gauging that majority of the sales comes from this fragment of individuals (the 80:20 rule). Therefor this segment of customers generates 80% of Wirkkala total sales. Consequently, as an organization, it makes economic sense to position the “Premier Grade Customers” as high priority. Analysing data from the managerial point of view, we would also recommend that the marketing priority of the new Juice should be focused on the Premier Customers.

With the help of confidence intervals, Wirkkala can respond to the research question “How confident can Wirkkala be that the sales for each sample reflect to that of the population?”. According to the 95% confidence interval (appendix b) by using the sample data as a basis for analysis, Wirkkala might estimate that the true population mean of sales data per segment is between the estimated calculations.

Wirkkala can anticipate future sales margins based on how many customers they can transfer from a lower segment to the higher segment using those margins. When comparing RFM values between segments, Wirkkala prides itself on giving regular customers first priority. These individuals already have an implicit loyalty to Wirkkala, therefore dissemination of information and acceptance of a new Wirkkala product to them would be easier and less expensive as they are brand loyalist.

Top management of Wirkkala saddled with the responsibility of shareholders wealth maximization (achieved through sales and generation of profit via revenue) / new product growth and increased market share (achieved through brand loyalty) would be focused on faster and cheaper ways for Wirkkala to Break even and start making profit following the substantial and staple fixed expenses needed to be incurred in commencement of the new Juice project. It is therefore economically and strategically wise to prioritize initial marketing on the “Premier Customers”.

To validate this suggestion, we acknowledge that people within this segment in the society tend to be indirect influencers, as soon as they become brand loyalist, other segments would automatically want to be part of the new “Trend”, saving marketing expenses for the organization.

Secondary marketing strategy can henceforth be apportioned to the 20% classified customers within the 3-1 Customer Segmentation

# Recommendations

* From the analysis a clear link between total sales and customer frequency. There is a strong positive linear relationship of r=0.92. It is to be recommended to management that automated advertising campaigns should be shared via email to customers once its ascertained that its 6-month past since a customer went by and didn’t purchase, discount code could be provided to customers within this category to serve as rewards if they redeem their voucher in time.
* From the analysis The Midlands store enjoys more sales compared to the London based store. Geographical, London has a denser population than the Midlands thus providing a greater marketing within a smaller region. One proposed method would be to increase advertisement within the London region to boost sales.
* Ascertaining confidence individuals from the analysis is very beneficial in order to inform a successful campaign. Utilising the predicted mean sales data calculated would aid Wirkkala in understanding how many customers will meet criteria up from their current segment. Understanding this will allow Wirkkala to predict future sales revenue.
* The sample provided had a distinct faulty data as no customers under the age of 30 were included in this sample, only 4 samples under 35 years of age were provided. It is unclear whether this is due to lack of under 35s shopping at Wirkkala or a fault in sampling method. Regardless, it is important to note that the age bracket (18-35) configure a great mass of the labour population, it is recommended that Wirkkala looks to target this younger demographic by carrying our market research and including product ranges that appeal to the younger generation in order to expand their customer base.

# APPENDIX

## APPENDIX A: RFM Measures of Central Tendency, Dispersion and Confidence Intervals

|  |  |  |  |
| --- | --- | --- | --- |
|  | **R Values** | **F Values** | **M values** |
| Mean | 85.95 | 4.62 | £ 15.68 |
| Variance | 7874.35 | 7.20 | £ 116.16 |
| Standard Deviation | 88.74 | 2.68 | £ 10.78 |
| +1 standard deviation | 174.69 | 7.31 | £ 26.46 |
| -1 standard Deviation | -2.78 | 1.94 | £ 4.90 |
| Count above 1 stdev of mean | 42 | 25 | £ 28.00 |
| Count Below 1 Stdev of mean | 0 | 24 | £ 17.00 |
| Total outliers | 20 | 19 | 14 |
| %1 within Stdev | 92.15% | 92.54% | 94.5% |
| Confidence Interval | 10.891 | 0.33 | 1.32 |
| C.I Upper Range | 96.84 | 4.95 | £ 17.00 |
| C.I Lower Range | -75.06 | 4.29 | £ 14.36 |

## APPENDIX B: Confidence Interval Calculations for Monetary Value.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Average of M** | **Variance of M** | **Standard Deviation of M** | **Confidence interval** | **Upper CI** | **Lower CI** |
| 1 | £6.87 | £5.74 | £2.40 | 0.558995 | 7.432999 | 6.307001 |
| 2 | £ 14.27 | £8.17 | £ 2.86 | 0.492013 | 14.76201 | 13.77799 |
| 3 | £ 23.73 | £9.42 | £3.07 | 1.189962 | 24.91996 | 22.54004 |
| 4 | £ 43.12 | £164.64 | £12.83 | 5.689069 | 48.80907 | 37.43093 |
| **Grand total** | **£ 87.99** | **£187.96** | **£21.15** |  |  |  |

## APPENDIX C: Amount Purchased, Sum of Total and Average of Total Paid per Month.

|  |  |  |  |
| --- | --- | --- | --- |
| **Months** | **Count of Customer Account Number** | **Sum of Total value Paid** | **Average of Total value Paid** |
| Jan | 104 | 324.18 | 3.12 |
| Feb | 101 | 316.68 | 3.13 |
| Mar | 104 | 340.08 | 3.27 |
| Apr | 118 | 375.15 | 3.18 |
| May | 82 | 264.94 | 3.23 |
| Jun | 102 | 334.22 | 3.28 |
| Jul | 66 | 204.19 | 3.09 |
| Aug | 98 | 334.31 | 3.41 |
| Sep | 106 | 336.26 | 3.17 |
| Oct | 92 | 287.35 | 3.12 |
| Nov | 96 | 295.65 | 3.07 |
| Dec | 110 | 402.17 | 3.65 |
| **Grand Total** | **1179** | **3815.20** | **3.23** |

## APPENDIX D: 2019 Sales by Region and Month

|  |  |  |  |
| --- | --- | --- | --- |
| **Sum of Total value Paid** | **Regions** | |  |
| **Months** | **London** | **Midlands** | **Grand Total** |
| Jan | 140.16 | 184.03 | 324.19 |
| Feb | 164.99 | 151.70 | 316.68 |
| Mar | 166.92 | 173.17 | 340.08 |
| Apr | 185.75 | 189.41 | 375.16 |
| May | 129.51 | 135.44 | 264.94 |
| Jun | 166.23 | 168.00 | 334.23 |
| Jul | 105.19 | 99.01 | 204.19 |
| Aug | 164.53 | 169.79 | 334.31 |
| Sep | 200.65 | 135.61 | 336.26 |
| Oct | 111.15 | 176.21 | 287.36 |
| Nov | 134.03 | 161.62 | 295.65 |
| Dec | 196.35 | 205.83 | 402.17 |
| **Grand Total** | **1865.43** | **1949.78** | **3815.21** |

## APPENDIX E: Most Purchased Items

|  |  |  |
| --- | --- | --- |
| **MOST PURCHASED PRODUCT** |  |  |
| **Stock code** | **Price** | **Count of purchase** |
| 21071 | £317.625 | 91 |
| 22077 | £310.27 | 89 |
| 22593 | £309.925 | 89 |

## APPENDIX F: Calculations to Determine RFM Segment Eligibility Criteria

|  |  |  |  |
| --- | --- | --- | --- |
| **Segment criteria calculations** | **R** | **F** | **M** |
| Min | 15/01/2019 | 1 | £2.50 |
| Q1 | 06/09/2019 | 3 | £8.85 |
| Median | 07/11/2019 | 4 | £12.88 |
| Q3 | 10/12/2019 | 5 | £19.465 |
| Max | 31/12/2019 | 17 | £86.47 |

## APPENDIX G: RFM Segment Eligibility Criteria

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria** | **R(Date)** | **F(Times)** | **M ( total spent)** |
| 1 | 15/7/2019-31/12/2019 | 0-2 | <£9.00 |
| 2 | 15/4/2019-14/7/2019 | 3-5.0 | £10.00-£19.00 |
| 3 | 15/2/2019-14/4/2019 | 6-8.0 | £20.00-£29.00 |
| 4 | 15/01/2019-14/2/2019 | 9-12.00 | >£30 |

## APPENDIX H: RFM Value Criteria to Determine Customer Segment

|  |  |
| --- | --- |
| **SEGMENT** | **RFM AVG CRITERIA** |
| 1 | 1-1.67 |
| 2 | 1.68-2.67 |
| 3 | 2.68-3.33 |
| 4 | 3.34-4.00 |

## APPENDIX I: Measures of Dispersion and Central Tendency of Age by Region.

|  |  |  |
| --- | --- | --- |
| **Age Dispersion** | **Midlands** | **London** |
| Min | 33 | 30 |
| Q1 | 49 | 47 |
| Median | 54 | 54 |
| Mean | 54.40 | 53 |
| Q3 | 59 | 59 |
| Max | 78 | 76 |
| IQR | 10 | 12 |
| IQR\*1.5 | 15 | 18 |
| Upper Limit | 74 | 77 |
| Lower Limit | 34 | 29 |

# References

Azeroual, O., Saake, G. and Abuosba, M., 2019. *Data quality measures and data cleansing for research information systems*. [online] Cornell University. Available at: <https://arxiv.org/pdf/1901.06208.pdf> [accessed on 13/02/22]

Canning, K. 2018. *Market Trends – orange juice sales continue to tumble.* [online] ebscohost.com. Available at: <https://web.p.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=4&sid=93649007-0a4b-49a9-a554-21eac67b9cfd%40redis> [accessed on 14/02/22]

Chahal, H., Jyoti, J. and Wirtz, J., 2019. *Business analytics: Concept and applications.* [online] Understanding the role of Business Analytics (pp. 1-8). Springer, Singapore. Available at: <https://link.springer.com/content/pdf/10.1007%2F978-981-13-1334-9.pdf> [accessed on 14/02/22]

Mintel. 2020*. Fruit Juice, Juice Drinks and Smoothies - UK - January 2020.* [online] reports.mintel.com. Available at:

<https://reports.mintel.com/display/1001152/?fromSearch=%3F_ga%3D2.196609452.148406303.1645015953-773689319.1645015953%26freetext%3DUK%2520Fruit%2520Juice%252C%2520Juice%2520Drinks%2520and%2520Smoothies%2520Market%2520Report%25202020> [accessed on 14/02/22]

Passport. 2020. *Market and Prices – Historical.* [online] Euromonitor.com. Available at: <https://www.portal.euromonitor.com/portal/statisticsevolution/index> [accessed on 14/02/22]