EXECUTIVE SUMMARY

The Wirkkala 2020 Consumer Transaction Dataset, which comprises a sample of their customer demography from both Midlands and London retail locations, is the subject of this business research. Consumers were separated into four groups based on recency, frequency, and monetary value utilising customer relationship management methodologies. To construct qualifying criteria for each of these sections, dispersion and central tendency metrics were used. Wirkkala can make statistical decisions based on his knowledge of these unique customer segments' habits and behaviours. To prepare the data for analysis, proper data cleaning and converting techniques were used, the results of which are detailed in this report.

To glean observations from this set of data, descriptive and statistical analytic techniques have been often chosen, as well as a focus on providing advanced visualization illustrations to help Wirkkala management teams; include valuable information to power managerial business decisions on their operational processes, answer their main research questions, and assist Wirkkala in tracking and monitoring their customer base. Formal advice on how to approach the creation of a marketing campaign was given, with a significant emphasis on the age, gender, and regional factors that separate the client groups.

INTRODUCTION

Wirkkala commissioned this study, which included 2020 sample data from both retail locations (London and Midlands). The goal of this report is to use statistical and analytical methodologies to determine whether the data is representative of the respective region and general population so that data-driven management decisions can be made. Customers were divided into four groups based on recency, frequency, and monetary value using RFM analysis, a customer relationship management method. There are further analyses that look at the differences between these categories in terms of age, gender, and geography. The accompanying spreadsheet file supplements the findings and includes an interactive digital dashboard.

BACKGROUND

Fruit juice is the most widely consumed and most popular daily choice, consumed by 71% of individuals and 14% on a daily basis. Since 60% of people believe that fruit juices, juice drinks, and smoothies are a simple method to increase vitamin consumption. (Mintel,2021)

However, the state-wide lockdowns and local limitations imposed by the pandemic in 2020 had a significant impact on sales of these drinks in catering and on the go. The market recovery will be slowed by the January 2021 lockdowns, and on-premise sales are unlikely to resume for several months. (Mintel,2021)

The loss of on-the-go, lunchtime meal deal, and backpack occasions during the COVID-19 lockdowns in 2020 disproportionately impacted the juice drinks category in the retail channel. Furthermore, because the on-trade channel accounted for a larger percentage of juice products sales than fruit juice and smoothies, the juice drinks category was dealt the most severe damage from the channel's demise in 2020. With sales declining in the high-value on-trade channel and single-serve items, the pressure on market value was far more significant than the effect on volume. In 2020, value sales declined by 29% year on year. (mintel, 2022)

Additionally, adspend for fruit juice, beverages, and smoothies fell dramatically in 2020, with adspend between January and November 2020 totaling just £4.7 million, opposed to £13 million for the whole year of 2019. (Mintel,2021)

But nonetheless, one of the effects of covid-19 on customer behaviour is long-term working from home to benefit greater layouts. With greater working from home projected to be one of the pandemic's long-term implications, this will continue to enhance at-home occasions and, as a result, bigger forms of fruit juice, juice drinks, and shakes over added-value convenience alternatives. (Mintel, 2021)

But also, once boundaries from the new wave of lockdowns that began in January 2021 are lifted, which is expected to happen in the spring, the fruit juice, juice drinks, and smoothies category is expected to see some jumps, thanks to the gradual reopening of foodservice and hospitality establishments. (Mintel,2021)

Additional, Due to the successful vaccination and booster campaign, the on-trade channel should experience considerably fewer pandemic-related interruptions. (Mintel,2022)

RESEARCH QUESTIONS

* Which customer segments should be initially targeted in a new advertising campaign once restrictions due to the pandemic are over?
* How certain is Wirkkala that the sample dataset chosen for analysis is representative of the broader population?
* How do the two regions differ in terms of sales and customer interactions? To be consistent with Wirkkala's excellent service to customers and places a high value on customer relationships, particularly with repeat customers, and its correlation to total sales will be a vital indication for growth and success.

STRUCTURE OF THE DATA

All of the data in the selected sample is organised, allowing for a more efficient analysis procedure. The data sample with the submission of personal tracker data, including Post area, birthdate, purchase date, and items and quantity of each purchasing. It includes stock code and inventory details such as price. The data was obtained 12 months in 2020 in two locations. More sheets were included to conduct analysis, such as the interactive Digital Dashboard and RFM analysis. Some data collected has been added to enable more extensive research, such as the Age Distribution sheet, which details the distribution of customer age in the separate areas and has been sorted by region, and from the customer data sheet to allow for comparison.

TRANSFORMING/ CLEANING DATA

After the format checking all column format, the birthdate column was converted in consisted of UK format. In the duplication checking, one duplicated data was found and removed in inventory sheet.

Age data was generated from birthdate and added as a column in the customer order sheet. ‘Price’ column in Inventory sheet was transferred into Orders sheet to combine with the ‘Discount voucher’ to calculate the total value to create the ‘total paid’ column.

For the dataset, completeness analysis was performed to assess data quality. The presence of incomplete data was noticeable in the ’Title’ column (blank n=79,28%) and the ‘First Name’ column (blank n=34,12%), Furthermore, the 'First Name' data column comprised a range of distinct forms, containing partial names and letters, resulting in a low percentage of uniqueness. It was decided to remove these two columns from the dataset and eliminate them from the study. The reasoning behind this is that the data fields do not contribute business value and removing these fields would significantly enhance data quality and completeness.

Review Figure 1 for a description of the data type, in addition to an assessment of data quality

*Figure1: Data types and quality*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data** | **Data type** | **Number of Values** | **Number of Blank Values** | **Completeness** |
| Customer Account Number | Ordinal | 283 | 0 | 100% |
| Title | Nominal | 204 | 79 | 72% |
| Gender | Nominal | 283 | 0 | 100% |
| Postal Area | Ordinal | 283 | 0 | 100% |
| Region | Nominal | 283 | 0 | 100% |
| Birthdate | Interval | 283 | 0 | 100% |
| Age | Interval | 283 | 0 | 100% |
| First Name | Nominal | 249 | 34 | 88% |
| Last Name | Nominal | 283 | 0 | 100% |
| Purchase Date | Interval | 283 | 0 | 100% |
| Stock Code | Nominal | 283 | 0 | 100% |
| Quantity | Ratio | 283 | 0 | 100% |
| Discount voucher (%) | Ratio | 283 | 0 | 100% |
| Price | Ratio | 283 | 0 | 100% |
| Total Paid | Ratio | 283 | 0 | 100% |
| Description | Nominal | 23 | 0 | 100% |
| Price | Ratio | 23 | 0 | 100% |
| Number in Inventory | Ratio | 23 | 0 | 100% |

\*\* After the data processing step, data fields are suppressed and restricted from analysis.

DATA ANALYSIS AND RESULTS

Customer Demographics

Figure 2 displays an image of the live graph on the digital dashboard (check Excel file: Digital Dashboard), exhibiting general customer demographics using age, location, and gender as factors. We recognise that the majority of customers are female across all age groups and regions. Moreover, the number of customers in age group from 30 to 39 is significantly smaller than the other age groups, and it is the youngest age group in the statistics. Besides, the 50-59 age group is the most prevalent in both areas, with a slightly greater number (n=11) in the Midlands compared to London.

*Figure 2: Customer Demographics by age, region and gender*

Figures 3 and 4 further discusses the customer age distribution in the respective regions.

. In the Midlands, two outliers were identified: 33 (1.5\*IQR above Q1) and 78 (>1.5\*IQR above Q3). The IQR shows that 50% of customers in London are between 42.25-58.75 years, compared to 49-59 years in the Midlands. (appendix g)

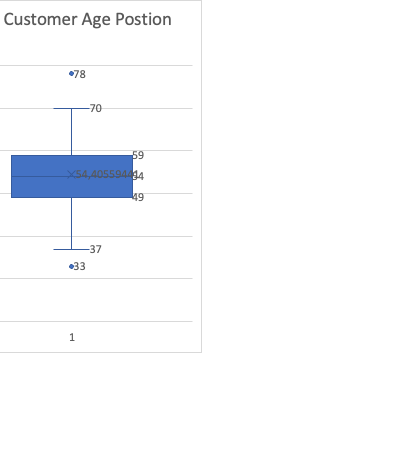
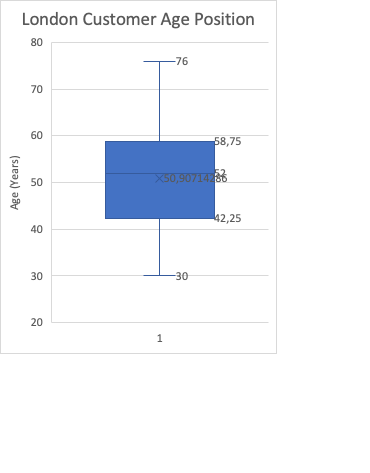
 

Figure 3: Customer age position in London Figure 4: Customer age position in Midlands

CUSTOMER SEGMENTATION

In this sample, RFM analysis were used to divide customers. RFM is an acronym that stands for Recency, Frequency, and Monetary Value. RFM analysis is a marketing approach that is used to evaluate client behaviour, how recently a customer bought (recency), how frequently the consumer bought (frequency), and how much the customer spent (monetary). (Birant, D., 2011)

Refer back to the descriptions while analysing the customer segments:

* Segment 4, High-value customer: high priority, lucrative, strong bargaining power, and frequent purchasers.
* Segment 3, Medium-value customer: medium priority, retaining priority, organisational interest
* Segment 2, Low-value customer: medium priority, occasional customer who contributes little to sales
* Segment 1, Non-value customer: low priority, primary emphasis on reattraction, single-shop buyers

Measures of dispersion and central tendency were utilised to generate criteria for segment categories for each R, F, and M. (appendix h)

Figure 5: The following segment criteria for each variable are shown below.

*Figure 5. Description and criteria for customer segmentation*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Customer Segment** | **R** | **London** | **Midlands** | **F** | **London** | **Midlands** | **M** | **London** | **Midlands** |
| 1 | Purchased in the past 21 days | 23 | 39 | Purchased 1-3 times in 2020 | 30 | 44 | Spent <£8.85 in 2020 | 28 | 38 |
| 2 | Purchased in the past 54 days | 35 | 30 | Purchased 4 times in 2020 | 43 | 35 | Spent from £8.86 - £12.5 in 2020 | 35 | 27 |
| 3 | Purchased in the past 116 days | 36 | 27 | Purchased 5 times in 2020 | 20 | 20 | Spent from £12.6 - £19.33 in 2020 | 30 | 34 |
| 4 | Purchased in the past 365 days | 30 | 35 | Purchased more than 6 times in 2020 | 31 | 32 | Spent > £19.34 in 2020 | 31 | 32 |

Figure 6 visualises the data in each segment. There is a substantial difference in 'frequency can be observed, with the higher number of customers in the segment 1 and 2 of the frequency count than that of segment 3 and 4

*Figure 6: Customer allocation in each section for individual R, F, and M values.*

Figure 7 illustrates the average total yearly expenditure, average number of purchases, and average date since last purchase for the typical customer in each sector to offer a holistic perspective of the average consumer in each segment.

*Figure 7: Average R, F and M value per Customer Segment*

|  |  |  |  |
| --- | --- | --- | --- |
| **Customer Segment** | **Average of Max Purchase Date** | **Average of Count of Purchase Date** | **Average of 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 |

To connect the customer demographics and customer segmentation, figure 8 shows the average age of customers within each of the four segments. The average age has also been subdivided by gender and area.

Figure 8 depicts the situation. By region, the male gender in London is younger (38.6) than the male gender in the Midlands (52.67). Females in London begin caring about a healthy lifestyle sooner (49,62) than women in the Midlands (52.34). In general, the Average customer age in the segment 3 and 4 show that the older spend more attention (time and money) on healthy items.

*Figure 8. Average Customer age distribution per segment by gender and area*

Figure 9 demonstrates a strong positive relationship between customer frequency and total sales (r = 0.97). Which can be understand as the more frequently a consumer purchases, the higher the company's income. Appendix a has a more comprehensive evaluation (table showing the number of purchases and average income of the juice bar by month) for more details.

*Figure 9. Customer Frequency and Total Sales Correlation*

REGIONAL ANALYSIS

Figure 10 depicts comparing regional sales statistics. The two areas show comparable monthly variations, but the months with the highest is different. Wirkkala's juice bar in the Midlands made the most money in December, while Juice bar in London made the most money in September, and both locations made the least money in July. By the end of 2020, the Midlands' juice bar was generated more income than London's bar. Refer Appendix b for the detail number.

*Figure 10: Sales by Region and Month in 2020*

CONCLUSION

*Figure 11: RRM Measure of Central Tendency. Dispersion and Confident Interval*

|  |  |  |  |
| --- | --- | --- | --- |
| **Column1** | **R** | **F** | **M** |
| Mean | 86 | 4.81 | 15.54 |
| Variance | 7896.98 | 8.25 | 114.01 |
| StDEV | 88.86 | 2.87 | 10.68 |
| (+1) StDEV | 174.86 | 7.68 | 26.21 |
| (-1) StDEV | 2.86 | -1.94 | -4.86 |
| Count above 1 STDEV | 42 | 27 | 28 |
| Count below 1 StDEV | 213 | 228 | 227 |
| Total Outliner | 11 | 9 | 9 |
| % within 1StDEV | 12% | 11% | 11% |
| Confident Interval 95% | 10.91 | 0.35 | 1.31 |
| C.I.Upper range | 96.91 | 5.16 | 16.85 |
| C.I.Upper range | 75.09 | 4.46 | 14.23 |

Figure 11 show the Confidence Interval at 95% of each criteria which can help Wirkkala answer their research question *’’How certain is Wirkkala that the sample dataset chosen for analysis is representative of the broader population?’’*

For forecast future sales, Wirkkala can base on the appendix c at the 95% confidence interval, Wirkkala can conclude, based on the sample data, that the real population mean of sales numbers per segment is between:

* Segment 1: £6.05 - £7.19
* Segment 2: £11.08 - £11.52
* Segment 3: £14.77 - £15.54
* Segment 4: £26.14 - £32.28

Wirkkala can try to anticipate sales margins based on how many clients they can transfer from a lower segment to a higher segment by utilizing these margins.

Wirkkala prides in placing a high priority on repeat consumers, which is evident when comparing RFM numbers across segments. First, refer to appendix d, High-value customer (Segment 4) stands for 47% of average yearly Wirkkala earnings, while segments 2 and 3 represent for 18% and 24%, respectively.

Furthermore, the high-value consumer (segment 4) purchases things from the business at a significantly higher frequency than any other section. Which can be refer a correlation between customer frequency and total sales in figure 9. According to figure 7, The high-value customer buys from Wirkkala 3.54 more times per year than consumer in segment3, comparable to a disparity of only 2.62 between the first and third segments.

We can be certain that the segmentation criteria were not influenced by outliers since the median, rather than the mean, was selected as the mid-point reference. The financial cultivation that these top consumers contribute may be utilised to drive organizational choices to promote these groups far more.

RECOMMENDATIONS

For the marketing campaign:

* As understanding, there is a significant positive linear link between client frequency and total sales, so customers can be sent automatic advertising campaigns through email if they haven't visited a shop in a particular amount of time. Based on the customer segment, Wirkkala can create a unique code for the customer in each segment to redeem in their next purchase. For example: For segment 3-4, Wirkkala should focus to maintain relationship with customers by giving the 10% discount on their birthday. For segment 1-2, Wirkkala can give the voucher when customers can register as frequency membership. In addition, Wirkkala can give some buy one to get one product form the less favourite products list after the 8th buying to introduce new flavour to consumers.
* According to customer demographics and average age in each segment, this sample does not contain any consumers under the age of 30. It's unclear if this is attributable to a shortage of under-30s buying at Wirkkala or a flaw in the sample method used. For the customer insight, Wirkkala should consider targeting the younger to expand brand name to the younger to grow the business.
* In customer demographics, most of the customer is the women, Wirkkala can expand the customer base by focusing on the man habit and buying behaviour to attract more male as a consumer.
* In Average Customer Age, the most consumer are about 50, to keep maintaining and attracting this group age. Wirkkala can send online survey to improve customer services and understand the older buys behaviour.
* According to Appendix e a & f, there are 4 most purchased items and 4 less favourite items, marketing team should launch a survey to compare and get client insight from these 2 list items, so they can improve products flavour and get the data to launch the next product. For example, Wirkkala can create a discount when consumer purchase the less favourite items to encourage people try new flavours.
* According to sales by region and month in the whole year 2020, we can observe that total revenue in Midlands is slightly higher than London (143.30). However, London had the second highest population, at 8.8 million people in the UK (Statista,2021). Wirkkala can consider implementing new commercial marketing to push brand awareness in London to attract more customer in the near future.
* Confidence intervals, as noted in the conclusion, can be utilised to fuel information decisions. If the marketing effort is successful, Wirkkala may forecast future sales revenue by analysing predicted average sales figures for the population to calculate how many consumers would fulfil the next qualification condition up from their previous segment. Wirkkala can launch membership program follow this segmentation and conduct analysis in the next event.
* Wirkkala should do more research that includes numerous years of archived archive data to present a more complete view of their consumers' yearly buying behaviour.

Reference

Birant, D., 2011. Data mining using RFM analysis. In *Knowledge-oriented applications in data mining*. IntechOpen.

Mintel (2021) Fruit Juice, Juice Drinks and Smoothies - UK - January 2021 Available at: https://reports.mintel.com/display/1042283/

Mintel (2022) Fruit Juice, Juice Drinks and Smoothies - UK - 2022 Available at: <https://reports.mintel.com/display/1101309/>

Statista (2021) Population of the United Kingdom in 2021, by region Available at: <https://www.statista.com/statistics/294729/uk-population-by-region/>

APPENDICES:

Appendix a: Amount Purchased, Sum of Total Paid and Average of Total Paid

|  |  |  |  |
| --- | --- | --- | --- |
| **Row Labels** | **Count of Purchase Date** | **Sum of Total Paid** | **Average of Total Paid** |
| Jan | 109 | £342.89 | £3.15 |
| Feb | 106 | £328.21 | £3.10 |
| Mar | 107 | £350.88 | £3.28 |
| Apr | 120 | £377.96 | £3.15 |
| May | 89 | £288.74 | £3.24 |
| Jun | 105 | £342.13 | £3.26 |
| Jul | 68 | £208.79 | £3.07 |
| Aug | 101 | £345.03 | £3.42 |
| Sep | 107 | £335.01 | £3.13 |
| Oct | 95 | £293.96 | £3.09 |
| Nov | 100 | £306.45 | £3.06 |
| Dec | 119 | £441.65 | £3.71 |
| **Grand Total** | **1226** | **£3,961.68** | **£3.23** |

Appendix b: 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** | **1909.19** | **2052.49** | **3961.68** |

Appendix c: Confidence Interval Calculations for Monetary Value

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Row Labels** | **Average of Sum of Total Price** | **Var of Sum of Total Price** | **StdDev of 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 d: Percentage of each segment

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Sum of Sum of Total Price** | **Percentage** |
| 1 | 436.9625 | 11% |
| 2 | 700.535 | 18% |
| 3 | 954.84 | 24% |
| 4 | 1869.345 | 47% |
| Grand Total | 3961.6825 |  |

Appendix e: Most purchased items

|  |  |  |
| --- | --- | --- |
| **Stock code** | **Count of Purchase** | **Price** |
| 84282 | 81 | 234.9 |
| 21669 | 81 | 234.9 |
| 21071 | 85 | 297.5 |
| 82580 | 84 | 243.6 |

Appendix f: Less favourite items

|  |  |  |
| --- | --- | --- |
| Stock code | Count | Value |
| 14557 | 5 | 13.75 |
| 15056 | 19 | 57 |
| 22167 | 14 | 44.8 |
| 64356 | 9 | 28.8 |

Appendix g: Measures of Dispersion and Central Tendency for Age by Region

|  |  |  |
| --- | --- | --- |
| **Age dispersion** | **London** | **Midlands** |
| Min | 30 | 33 |
| Q1 | 42.25 | 49 |
| Median | 52 | 54 |
| 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 |

Appendix h: Calculations to Determine RFM Segment Eligibility Criteria

|  |  |  |  |
| --- | --- | --- | --- |
| Segment criteria calculations | R | 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 i: 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 - £19.30 |
| 4 | 11/12/2020 - 31/12/2020 | >=6 | > £19.31 |

Appendix k: 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 |