Exploratory Analysis and Presenting Insights

Need for the Report

Business Problem – 2Market are a global supermarket business seeking to identify customer demographics to create a targeted advertising strategy. By analysing historical data, key demographics and highest selling products can be highlighted. This analysis will enable 2Market to select more effective advertising platforms to execute a focused advertising strategy.

The IDEAL (Bransford & Stein, 1993) and 5 Whys Frameworks were utilised to investigate root causes. Appendix 1 details the meeting transcripts. The conversation highlighted the following root causes for the report:

1. High Sunk Costs -> Historical advertising strategy covers multiple platforms. The Chief Marketing Office (CMO) aims to save costs by reducing platforms.
2. Increase Lead Conversions -> The CMO wants to increase advertising sales by narrowing the scope of products to a key demographic.

To develop solutions, these business questions aided insight research:

- What are the customer demographics?
- What is the best-selling product?
- Are there patterns between demographic and product?
- What are the most effective advertising platforms, and do these differ by demographic?

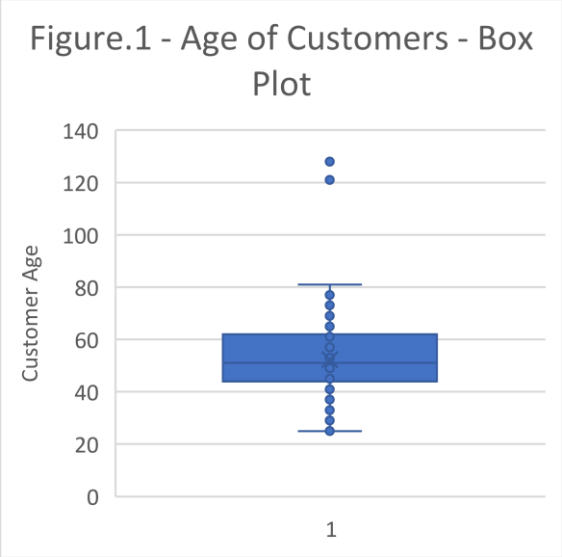
Cleaning the Data

The data was opened and cleaned in Excel power query to show a historical method for cleaning. Below outlines method and justification for the actions:

Initial Marketing dataset, n=2216

Data Quality - Characteristic	Error/Outlier Identified	Correction	Justification
Accuracy, Validity	Incorrect Data formats of Income	Replace function used to remove '$'. Converted to Income to currency format.	By converting to a continuous standardised measure, descriptive statistical analysis can be performed
Completeness	Check for Blank Cells	Table was highlighted and the find function searched for blanks. No blanks identified.	Missing data can create anomalous results during analysis
Consistency and Timeliness	Incorrect Data format of Dt_Customer	US date formats separated out and converted to UK format before changing column to date format.	Date now in a normalised format, which allows for analysis. Limitation -> Some US formats may still have been incorrectly converted
Uniqueness	Duplicate Rows Identified	Duplicates were grouped by column. 392 duplicate rows were identified and removed.	These rows were completely identical apart from Unique ID and Country. The rows were removed to avoid bias as this would have affected analysis relating to country demographics. 2Market can review these eliminated to decide whether these should remain

Secondly, data was visualised in excel to identify outliers within Age, Income, and Marital Status. Figures 1 and 2 highlight outliers of Age and Income respectively. Box-and-whisker plots identify the spread of data with interquartile range (IQR). Data outside whiskers of the box-and-whisker plot are outliers and were removed. Table.1 shows marital status of customers. In red are anomalous categories that due to their size were considered not relevant and were removed. The dataset is finalised for analysis where n = 1806.



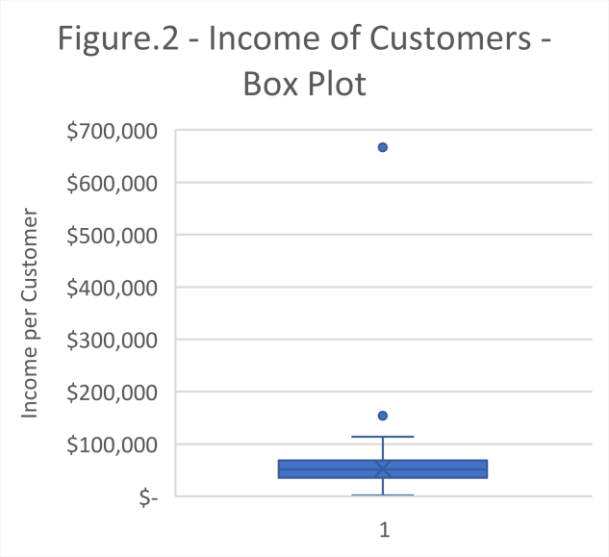


Table.1 - Summary of Customers by Marital Status

Marital Status	Absurd	Alone	Divorced	Married	Single	Together	Widow	Grand Total
Count	2	3	196	709	402	448	62	1822

Following excel, data was input to SQL. More complex analysis was performed by creating an SQL database and 2 tables for marketing and advertising. The data was imported via csv file, and a JOIN syntax assessed the tables together.

Dashboard Development

Tableau was selected for exploratory visualisation. A cross database inner join connected multiple sources. Inner join was chosen to exclude rows of outliers and duplicates. Although this inhibits the data volume, quality is maintained as there are no N/A values. Further investigation should assess anomalous data to debate its authenticity.

The dashboard was designed to present to the CMO and targets trends to show a key demographic, product, and platform. Colours were chosen to assist colour blind and is mobile friendly.

Age and Income:

- The visualisations show spreads of measure – centre and distribution. Tables identify demographic attributes whilst a box-and-whisker plot visualises data points to highlight the spread of age, indicating a normal distribution. Income 'bins' were established to group customers into Income brackets to show distribution.

Marital Status, Education, Kids at Home:

- Tables highlighted the highest and lowest categories, with colour indicating the size of grouping. Furthermore, data was organised in descending order to clearly highlight the most important demographic.

Country:

- Spatial data was identified and consequently a map created a visualisation to see customer base globally. To demonstrate size of population a choropleth map was chosen to highlight the customer countries. If more precise customer locations could be supplied, a symbol map would have been chosen to show customer distribution per country.

Comparisons between Products, Demographics and Advertising Platforms:

- Bar charts were chosen for these comparisons because there were few categories with many items. Product alias' were created to enhance understanding. Fonts for titles, measures and labels were boldened for clarity. Interactivity elements were included to support differences identified.

Liquor Vs Income and Number of Children:

- During SQL analysis, an insight was uncovered to test whether a proportion of the total variability in Liquor can be explained by Income. Consequently, a scatter chart was chosen, and a trend line was inputted to show the relationship.

Insights and Trends Identified

Demographics:

- The average age was 52 with a standard deviation of 12.
- The maximum age was 82 and minimum was 25. The IQR was 18 with no outliers. The most common age bracket was 46-50yrs (311 customers), the smallest was 76yrs+ (21 customers).
- 'Married' was the most common marital status and 'Graduation' for Education status.
- Most customers (930) had 1 child, followed by no children (501). The largest customer pool by country is Spain (50% of customers).

Product:

- Alcohol had the highest sales of $550k representing 50% of Total Sales.
- Per country, Alcohol is the highest selling product and is recommended to be the target product for advertising.
- By Income bracket, 70-80k earners spent the most alcohol ($156k) followed by 60-70k earners ($135k).
- Income was identified to test whether there is a variation in Alcohol Sales. The R^2 was 0.52 at $p<0.0001$, which means that 52% of Alcohol sales can be explained by Income. For future it is recommended to collect data that specifies the type of product per successful lead, enabling better regression analysis.

Advertising Platform:

- The most successful platform is Twitter, while Brochures are the least successful and it is recommended to drop.
- By Income bracket, the most successful leads were 80-90k Income earners, followed by 70-80k.
- By Income bracket, it is suggested to use Instagram and Facebook platform for mid-high-income earners. Twitter and Bulk-mail should target middle to low range income earners.

Further investigation:

- 70-90k earners have the most successful advertising leads, despite this, the highest sales for 2Market are within the 60-80k earners. Notably, 60-70k earners generate 23% of Total Sales, but only make up 13% of Successful Leads. Recommended future insight should investigate whether cheaper alcohol would be more effective to Mid-Low-Income earners to increase Successful Leads.

Appendices

Appendix 1 – 5 Whys Framework

Why are you searching to understand your customer demographics ?

- "Because we want to run a targeted advertising campaign"

Why ?

- "Because our previous attempts at running advertising campaigns have been costly and are not generating enough lead conversions. Furthermore, given the growth of social media we want to utilise the best advertising platform."

Why ?

- "Because we are currently advertising on multiple channels and are using all our products to advertise to our customers regardless of their demographic. We are also not able to be able to visualise our customer demographics clearly to help decision making regarding advertising strategy"

Why ?

- "Because we wanted to maximise our outreach to customers and so our initial advertising strategy was to expand onto as many platforms as we could. However, this is becoming costly and therefore we need to streamline who we target and which platforms we use. Furthermore, we constantly have challenges regarding using our data because it is messy and not in correct data formats. "

Why?

- " Because our Advertising KPIs are as follows:
  - Maximise the number of successful lead conversions per platform
  - Reduce costs by eliminating 'deadweight' platforms"