2Market

An exploratory analysis into a global supermarket's customer purchase behaviour

LSE Data Analytics Career Accelerator
Course 1: Data Analytics for Business

Jessica Krook
12 December 2022

Background of the business

2Market is a global omni-channel supermarket. The team wants to understand their customer purchase behaviour with a focus on customers' demographics, effective advertising channels, and best-selling products.

This project will support 2Market's objectives by giving insights to help the team make data-driven decisions based on customer purchase behaviour.

I used the Productive Thinking Model to unpack the business problem (**Figure 1**). This allowed me to shift my perspective to focus on what is success and what steps do we need to do to achieve it.

While I find the first three steps of the model applicable, the model is limiting at this point in the project. I have not analysed the data so I cannot move onto the step "Generate Answers", so it is impossible to finish the problem-solving framework.

Project questions:
Why is the project happening now?
Who are the stakeholders of the project?
Who is 2Market's target market?

Data questions:
Recency of the data?
Source of the data?

Analysis questions:
- Most attractive demographic?
- Best-selling products in each country?
- Most effective advertising channels?

Analytical approach

Cleaning the Data

Using Excel, I**:**

- Removed duplicates
- Added an age column (= 2022 - Year_Birth)
- Fixed Spelling mistakes in marital status column **(Figure 2)**
- Removed outliers of age and income **(Figure 3)**

Analysing the data

Using Excel, I**:**

1. Explored the demographics age and income
2. To explore patterns and trends, calculates:
    a. Average age

    b. Average age per marital status and income bracket
    a. Sales per marital status and income bracket
3. Created charts to visualise the patterns **(Figure 4, 5)**
4. Analysed the population of grouped data (**Figure 6, 7)**

Insights from my analysis:

- Income range of the data is wide
- Age range is narrow **(Figure 3)**
    - Indicating the supermarket does not attract customers based on income, but on age.
- The most populous age group falls around 51 (**Figure 6, 7)**

Using SQL, I:

1. Looked at best selling products and if they vary by demographics.
2. Aggregated each category, grouped them by country and sorted on total spend **(Figure 8).**
    a. Spain has the highest total spend **(Figure 9).**

Observations:

1. I was able to find the country with the highest spend (**Figure 8**) but it is difficult to make comparisons between countries with this table format.
   So, I created a new query to export the data into Tableau **(Figure 10).**

2. After creating the same query multiple times with small adjustments for a specific grouping, going forward I will run a query to create a table with all the data I need and create queries on that table to answer specific questions.

Next, using SQL I explored revenue generated from advertising channels by:

- Creating a table joining advertising and marketing datasets including the columns needed for my analysis **(Figure 11)**.
- Using formulas to evaluate the most successful advertising channel based on revenue by country and marital status **(Figure 11)**
- Exporting the table to Tableau for further analysis.

Observations:

Currently it is possible to only see total conversions per customers' total spend

This is limiting because:
- You cannot see which channel converted a sale
- You cannot link channel to a specific transaction.

- There is double counting of spend per channel when customers convert from multiple advertising sources.

Solution:

Add a transaction ID to both tables. Then it is possible to link conversion type to each sale.

Dashboard Design

I created my dashboard using best practices:

1. Applicability:

The dashboard answers the business question (**Figure 1**) by showing:

- Advertising channel conversions
- Country with highest advertising conversions
- Revenue per
    - Country
    - Advertising channel
    - Product type
- Sales per capita

2. Accessibility:

- Included a dashboard description
- Used colour and size on my visuals
- Captioned all visuals

3. Colour:

- Implemented a colour-blind palate on visuals.
- Each country has a constant colour

4. Interactivity:

- Visuals interact with each other based on country
- Highlights and Filters are used to drill down into country-specific data

5. Reasons for visualisations:

The dashboard shows 2Market's situation. I used simple visuals and layered details on them so users can see more datapoints while keeping the dashboard clean.

- Sales and Sales per capita packed bubbles: These two visuals work together to show that one cannot only use "sales per country" to select the most attractive market.
- Country conversions map focuses on conversions.

    - Size is total conversions
    - When hovering over the country there is a breakdown of channel conversions
- Product sales bar chart: The bar chart shows the sales of each product category with colours and details included for more insight, making the data more digestible
- Advertising channels sales bar chart: same reasons as above
- Advertising channel conversions table: a simple table highlighting total conversions so user can have context of advertising sales

NOTE: due to the limitations of the data (double counting of spend from multiple channel conversions) the two bar graphs are not comparable.

6. Layout:

Chosen so the eye naturally follows the dashboard:

1. Title
2. Description
3. Filter.
4. Packed bubbles - easy to understand and have a lot of information
5. The bar charts on the right half of the dashboard, so it is easy to read
6. The map is in the bottom – looks good with the rest of the visuals
7. The table is under it as it is used for a reference as more context

Trends and insights

Trends

1. Demographics:

- Customers who spend the most are:
    - Married
    - Salary of 61730-76729
- Spain has the highest sales value and conversions

2. Best-selling product per country:
    - Liquor

3. Most effective advertising channels:
    - Facebook has the most total conversions
    - Instagram has the highest value of converted sales

4. Spread of total sales between the countries is big but the average spend of each country’s spread is smaller

Insights

Spain has the highest conversions and the most sales. However, its sales per capita is lower than the total average sales per capita. If 2Market wants to grow their sales in other countries, they should look at Spain's marketing strategies and adopt it in other countries with higher sales per capita. This will compound the growth of profit, as each new customer in certain countries will spend more than Spanish customers.

Montenegro has 3 customers with a very high spend per customer and follows none of the general trends of the data. The data needs to be investigated for possible data capturing errors

Next Steps

Other countries should adopt Spain's marketing strategies. This will:

- Increase total population of customers
- Increase total conversions.
- Compound the increase in sales where countries have a higher per capita spend than Spain.

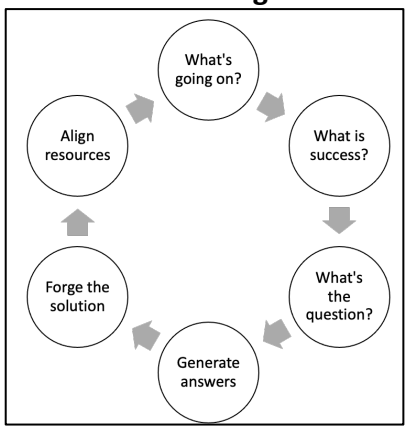
With the addition of a Transaction ID data point, it will be possible to analyse sales at basket level. We can gain further insight into which advertising channel converts which products. We can then answer questions like: if Instagram sales are the largest, what is being advertised on Instagram vs Facebook who has the highest conversions.

Word count: 1156

Appendix:

Figure 1 - First steps of the Productive Thinking Model:

Productive Thinking Model framework



Productive Thinking Model for 2Market

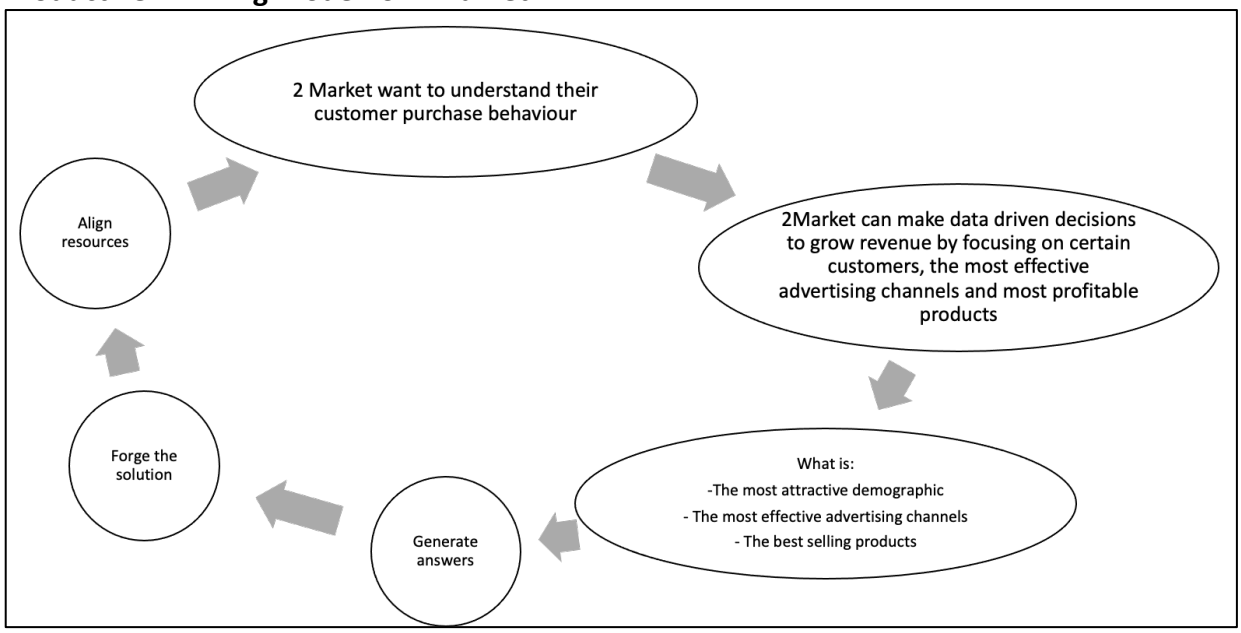


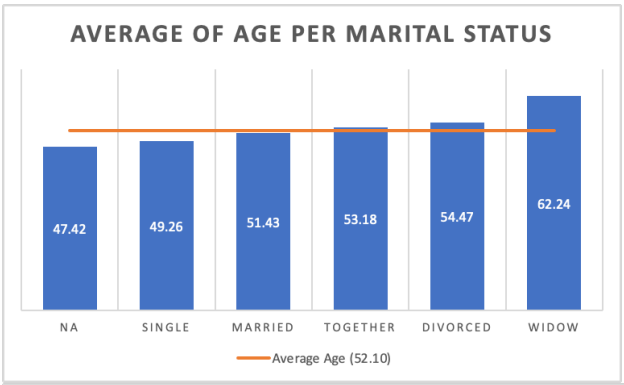
Figure 2 – Cleaning data by fixing spelling errors in marital status – any spelling errors were updated to NA as it is not possible to determine what the customer should have chosen:

Marital Status	Marital Status Cleaned
Absurd	NA
Alone	NA
Divorced	Divorced
Married	Married
Single	Single
Together	Together
Widow	Widow
Yes	NA
YOLO	NA

Figure 3 – Using measures of spread to identify outliers in age and income and calculating the range of age and income:

Measure	Age	Income
QTL 1	47.5	35,196.00
QTL 3	63	68,281.00
IQ range	15.5	33,085.00
Lower limit	24.25	(14,431.50)
Upper limit	86.25	117,908.50
Range	16.2	112,004.00

Figure 4 – visualisations of average age of each marital status group and bracket with a total average age line as a reference:



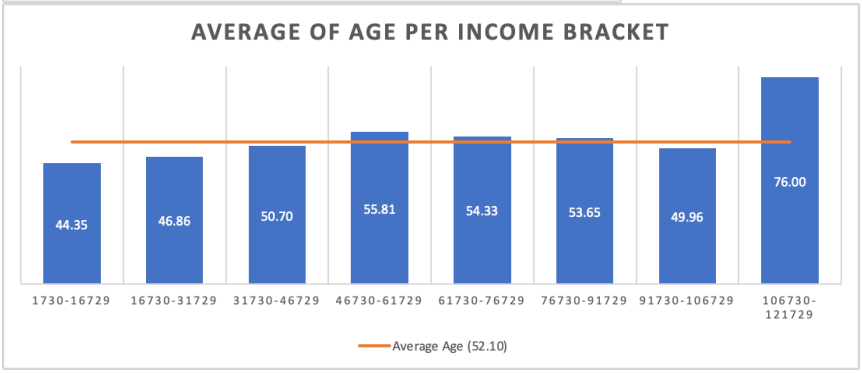


Figure 5 – Total Sales Value per marital status and income bracket

Income Bracket	Total Sales
1730-16729	6,613.00
16730-31729	24,866.00
31730-46729	88,297.00
46730-61729	247,554.00
61730-76729	534,341.00
76730-91729	387,454.00
91730-106729	48,640.00
106730-121729	277.00
Grand Total	**1,338,042.00**

Marital_Status	Total Sales
Divorced	141,825.00
Married	504,406.00
NA	4,010.00
Single	288,979.00
Together	345,626.00
Widow	55,401.00
Grand Total	**1,340,247.00**

Figure 6 – average age per marital status with a count of the population of each group:

Marital Status	Average of Age	Count
NA	47.42	7
Single	49.26	470
Married	51.43	857
Together	53.18	571
Divorced	54.47	231
Widow	62.24	76

Figure 7 – average age per income status with a count of the population of each group:

Income Bracket	Average of Age	Count
1730-16729	44.35	75
16730-31729	46.86	359
31730-46729	50.70	524
46730-61729	55.81	465
61730-76729	54.33	485
76730-91729	53.65	268
91730-106729	49.96	28
106730-121729	76.00	1

Figure 8 – SQL Query aggregating total spend per product category of each country:

```
SELECT md.country
-- aggregate categories and rename to total_amt_"category"
                ,SUM(md.amtliq) total_amt_liqour
                ,SUM(md.amtvege) total_amt_vegetables
                ,SUM(md.amtnonveg) total_amt_non_vegegetables
                ,SUM(md.amtpes) total_amt_fish
                ,SUM(md.amtchocolates) total_amt_chocolates
                ,SUM(md.amtcomm) total_amt_comodities
                -- aggreagate each category to create a total amount of all
                ,SUM(md.amtliq) + SUM(md.amtvege) + SUM(md.amtnonveg) +
SUM(md.amtpes) total_amt
        FROM public.marketing_data md
-- group by country for a visual of spend per category per country and total spend
(total_amt)  per country
GROUP BY md.country
-- sort descending by country to see which country spends the most
ORDER BY total_amt DESC;
```

Figure 9 – Results of Query 8. Here you can see Spanish customers have the highest total spends:

country	total_amt_liqour	total_amt_vegetables	total_amt_non_vegegetables	total_amt_fish	total_amt_chocolates	total_amt_comodities	total_amt
SP	335,637	28,144	177,847	40,049	30,070	45,957	581,677
SA	105,901	8,923	58,375	13,655	9,018	15,115	186,854
CA	84,066	7,681	45,925	9,980	7,607	12,144	147,652
AUS	42,752	3,689	22,328	5,546	4,129	7,132	74,315
IND	36,221	3,782	23,721	4,811	3,217	5,989	68,535
GER	36,776	2,980	20,272	4,601	2,801	5,768	64,629
US	32,214	3,034	20,185	4,411	2,863	4,839	59,844
ME	1,729	8	817	226	122	220	2,780

Figure 10 – Query creating a new column that makes it possible to see the total spend of each product category per country and marital status. This query is used to create a table to export to Tableau, so the data is comparable in tables to get insights into the spend of different product categories. ID is included to join this data to other data in Tableau:

```
SELECT md.country
          ,md.marital_status
         ,md.customer_id
-- STEP 1: query that creates a new column called “Product” and create a line within that
column each category that reflects the total spend of the category
         ,'Liquor' product
         , SUM(md.amtliq) total_amt
         FROM public.marketing_data md
-- STEP 2: group by country to see spend per category per country
GROUP BY md.country
            ,md.marital_status
            ,md.customer_id
-- STEP 3: add all "Product" queries together to create one column with every category in it
```

```
UNION ALL
SELECT md.country
            ,md.marital_status
           , md.customer_id
           --Repeat STEP 1
           ,'Vegetable' product
           , SUM(md.amtvege) total_amt
FROM public.marketing_data md
-- Repeat STEP 2
GROUP BY md.country
            ,md.marital_status
           ,md.customer_id
-- Repeat STEP 3
UNION ALL
SELECT md.country
            ,md.marital_status
           ,md.customer_id
--Repeat STEP 1
           ,'Meat' product
           , SUM(md.amtnonveg) total_amt
FROM public.marketing_data md
-- Repeat STEP 2
GROUP BY md.country
                     ,md.marital_status
                     ,md.customer_id
-- Repeat STEP 3
UNION ALL
SELECT md.country
            ,md.marital_status
           ,md.customer_id
--Repeat STEP 1
           ,'Fish' product
           , SUM(md.amtpes) total_amt
           FROM public.marketing_data md
--- Repeat STEP 2
GROUP BY md.country
                     ,md.marital_status
                     ,md.customer_id
-- Repeat STEP 3
UNION ALL
SELECT md.country
            ,md.marital_status
           ,md.customer_id
--Repeat STEP 1
           ,'Chocolates' product
           , SUM(md.amtchocolates) total_amt
FROM public.marketing_data md
```

```
-- Repeat STEP 2
GROUP BY md.country
                ,md.marital_status
                ,md.customer_id
-- Repeat STEP 3
UNION ALL
SELECT md.country
          ,md.marital_status
        ,md.customer_id
--Repeat STEP 1
        ,'Commodities' product
        , SUM(md.amtcomm) total_amt
FROM public.marketing_data md
--- Repeat STEP 2
GROUP BY md.country
                ,md.marital_status
                ,md.customer_id
```

Figure 11 – Joining marketing and advertising datasets together:

```
-- Assumption: some customers have converted on more than one advertising channel
-- The revenue can be counted more than once
CREATE TABLE adchannelrevenue AS(
SELECT
-- include all relevant columns from both advertising_data and marketing_data needed for current or potential future analysis
        md.customer_id
        ,md.country
        ,md.marital_status
        ,ad.bulkmail_ad
        ,ad.instagram_ad
        ,ad.twitter_ad
        ,ad.facebook_ad
        ,ad.brochure_ad
        ,md.amtliq liquor
        ,md.amtvege vegetables
        ,md.amtnonveg meat
        ,md.amtpes fish
        ,md.amtchocolates chocolates
        ,md.amtcomm commodities
        ,md.numwebbuy
        ,md.numwalkinpur
-- create a column index of 1 called num_customers, this can be used for calculations
        ,1 num_customers
        ,ad.bulkmail_ad bulkmail
        ,ad.twitter_ad twitter
        ,ad.facebook_ad facebook
```

```
        ,ad.instagram_ad instagram
        ,ad.brochure_ad brochure
-- The code looks at if there is a value higher than 0 in each advertising channel type, then sums together the spend of all product categories
        ,CASE WHEN ad.bulkmail_ad > 0 THEN md.amtliq + md.amtvege + md.amtnonveg + md.amtpes + md.amtchocolates + md.amtcomm ELSE 0 END bulkmail_spend
        ,CASE WHEN ad.twitter_ad > 0 THEN md.amtliq + md.amtvege + md.amtnonveg + md.amtpes + md.amtchocolates + md.amtcomm ELSE 0 END twitter_spend
        ,CASE WHEN ad.facebook_ad > 0 THEN md.amtliq + md.amtvege + md.amtnonveg + md.amtpes + md.amtchocolates + md.amtcomm ELSE 0 END facebook_spend
        ,CASE WHEN ad.instagram_ad > 0 THEN md.amtliq + md.amtvege + md.amtnonveg + md.amtpes + md.amtchocolates + md.amtcomm ELSE 0 END instagram_spend
        ,CASE WHEN ad.brochure_ad > 0 THEN md.amtliq + md.amtvege + md.amtnonveg + md.amtpes + md.amtchocolates + md.amtcomm ELSE 0 END brochure_spend
-- join syntax joining two tables together on the same key: customer_id
FROM public.ad_data ad
    JOIN public.marketing_data md USING (customer_id)
);
```

Figure 12 – Query to calculate the total spend and the average spend per customer of each advertising channel:

```
SELECT country
-- calculate the total  of each advertising channel
        ,SUM(bulkmail_spend) bulkmail_spend
        ,SUM(twitter_spend) twitter_spend
        ,SUM(facebook_spend) facebook_spend
        ,SUM(instagram_spend) instagram_spend
        ,SUM(brochure_spend) brochure_spend
-- calculate the average spend per customer of each advertising channel
        ,SUM(bulkmail_spend)/SUM(num_customers) avg_bulkmail_spend
        ,SUM(twitter_spend)/SUM(num_customers) avg_twitter_spend
        ,SUM(facebook_spend)/SUM(num_customers) avg_facebook_spend
        ,SUM(instagram_spend) /SUM(num_customers) avg_instagram_spend
        ,SUM(brochure_spend) /SUM(num_customers) avg_brochure_spend
FROM adchannelrevenue
-- all calculations by country
GROUP BY country;
```

Figure 12:

```
SELECT marital_status
-- calculate the total  of each advertising channel
        ,SUM(bulkmail_spend) bulkmail_spend
        ,SUM(twitter_spend) twitter_spend
        ,SUM(facebook_spend) facebook_spend
        ,SUM(instagram_spend) instagram_spend
        ,SUM(brochure_spend) brochure_spend
```

```
-- calculate the average spend per customer of each advertising channel
        ,SUM(bulkmail_spend)/SUM(num_customers) avg_bulkmail_spend
        ,SUM(twitter_spend)/SUM(num_customers) avg_twitter_spend
        ,SUM(facebook_spend)/SUM(num_customers) avg_facebook_spend
        ,SUM(instagram_spend) /SUM(num_customers) avg_instagram_spend
         ,SUM(brochure_spend) /SUM(num_customers) avg_brochure_spend
FROM adchannelrevenue
-- all calculations by marital status
GROUP BY marital_status;
```

Figure 13 – Key metrics from analysis showcasing highest and lowest values for each category:

Metric	Highest	Value	Lowest	Value
Sales per Country	Spain	657,704	Montenegro	3,122
Spend per Capita	Montenegro	1,041	India	532
Customer Population	Spain	1,092	Montenegro	3
Sales per Channel	Instagram	260,009	Brochure	39,230
Conversions per Country	Spain	350	Montenegro	1
Product Sales	Liquor	675,296	Vegetables	58,241
Conversions per Channel	Facebook	164	Brochure	30