2Market: Exploratory analysis and presenting insights- Data Analysis Project

Lesson 1: Context and Problem Identification

1. Problem Statement

2Market is a global supermarket that sells products both online and in-store. They aim to optimize their marketing strategy to increase profitability. Specifically, they want to (a) identify which advertising channels work well in each country, (b) understand which products are most and least in demand across different locations and demographics, and (c) tailor their advertising spend to reach the right customers, in the right place, at the right time. By analyzing sales performance, advertising effectiveness, and customer demographics, this data analytics project will help 2Market minimize costs, focus on the best-performing marketing channels, and increase revenue by targeting the right audience.

2. Questions for the 2Market Team

- Could you provide more background on 2Market's expansion history? For instance, how long have you been operating in each country, and have you observed any region-specific marketing approaches or challenges?
- Is there any additional demographic or financial data (beyond what's in the CSV files) that we can leverage?

3. Additional Questions to Ask of the Data

- Has the marketing spend or distribution across channels changed over time, and has this influenced the number of successful conversions or overall revenue?
- Are there any seasonal or trend-based spikes in product categories that might help 2Market adjust marketing strategies or inventory? For example, do certain countries or demographic groups prefer specific product categories at certain times of the year?

If there is no additional data to address these questions, we can still investigate patterns in the existing data on how channel effectiveness and product popularity shift by customer demographics and location over different periods. This analysis will help 2Market decide where and when to allocate marketing resources for the highest return on investment.

Assumption: We are doing this analysis in 2025 and the key stakeholders are Senior Marketing Executives and Board members.

Lesson 2: Excel for Analysis

Data Cleaning and Transformation

The data in both CSV files was cleaned in Excel to ensure accuracy and consistency before analysis. Missing values were checked, and all date fields were reformatted using TEXTSPLIT to standardize the format as outlined in the metadata. Outliers were removed using the interquartile range method. Dollar signs were removed from numeric fields and added at the header level for clarity. Blanks, error values, and spelling mistakes were corrected. Inconsistent category labels were fixed, and low-frequency values (like “YOLO” and “Absurd”) were merged.

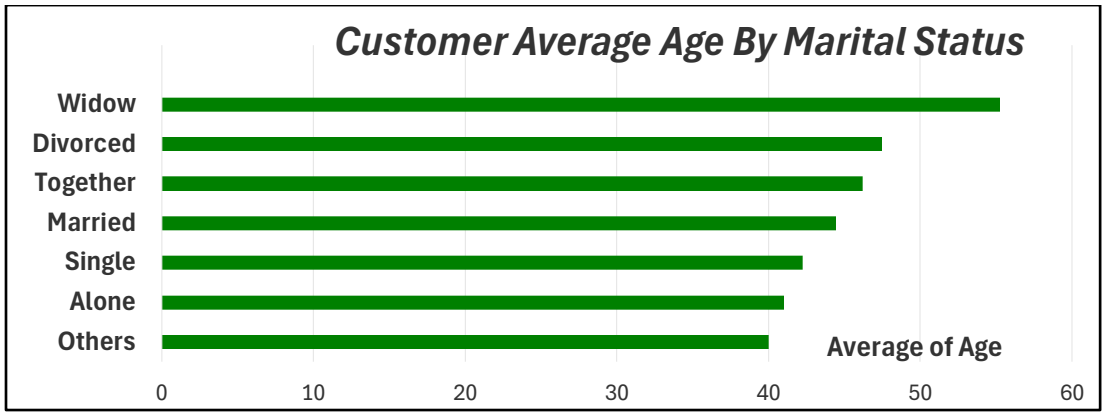
Two new columns were added for easier analysis: Customer Age and Total Spend by Customer. Duplicates were checked using customer ID—none were removed, as all entries were unique. Country codes were standardized (e.g., "ESP" to "Spain"), and all values were formatted appropriately for use in Tableau and SQL. This process ensured the data was clean, consistent, and ready for reliable analysis.

Below are key findings from analyzing customer age, income, and marital status.

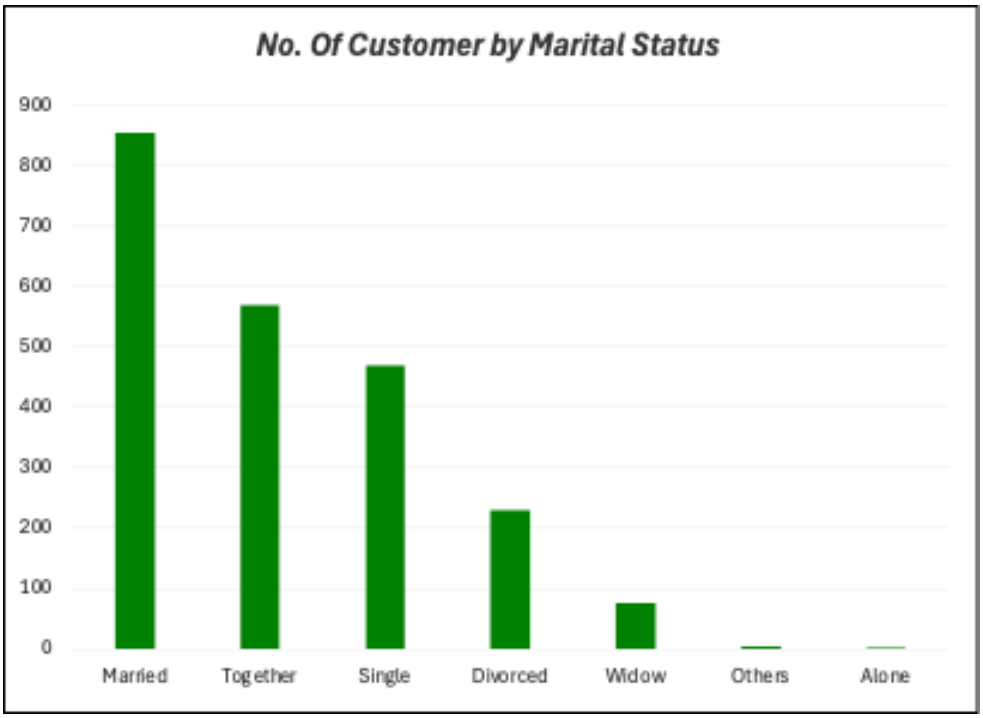
Marital_Status	Data Age	Frequency	Percentage
Alone	41	3	0.14%
Divorced	47.47391304	230	10.43%
Married	44.42622951	854	38.73%
Others	40	4	0.18%
Single	42.26170213	470	21.32%
Together	46.18485915	568	25.76%
Widow	55.23684211	76	3.45%
Grand Total	**45.09569161**	**2205**	**100.00%**

Income between 90k to 10k (usd)	
Average of Age	Total
Total	42.91

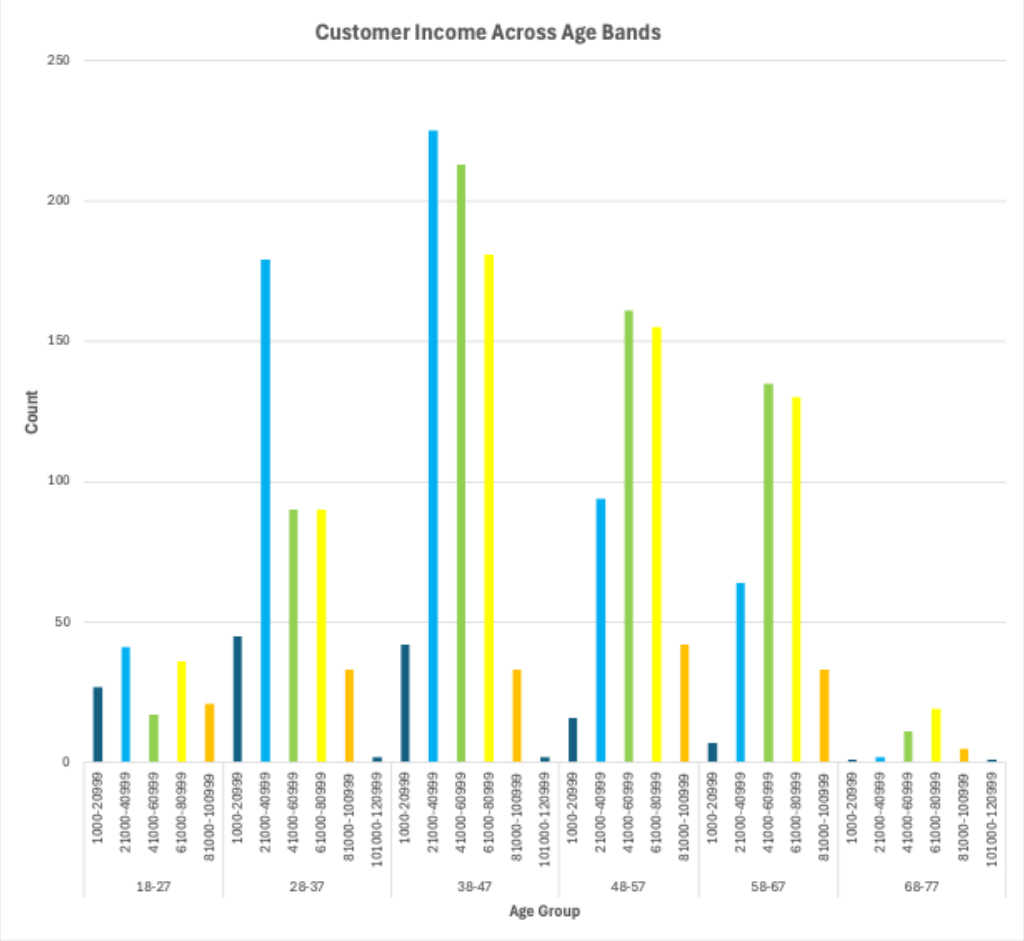
The column chart below illustrates the average age of customers by marital status to help visualize the variation across groups. Excluding the “Others” category due to its low sample size, the highest average age is observed among widowed customers, while single customers have the lowest average age.



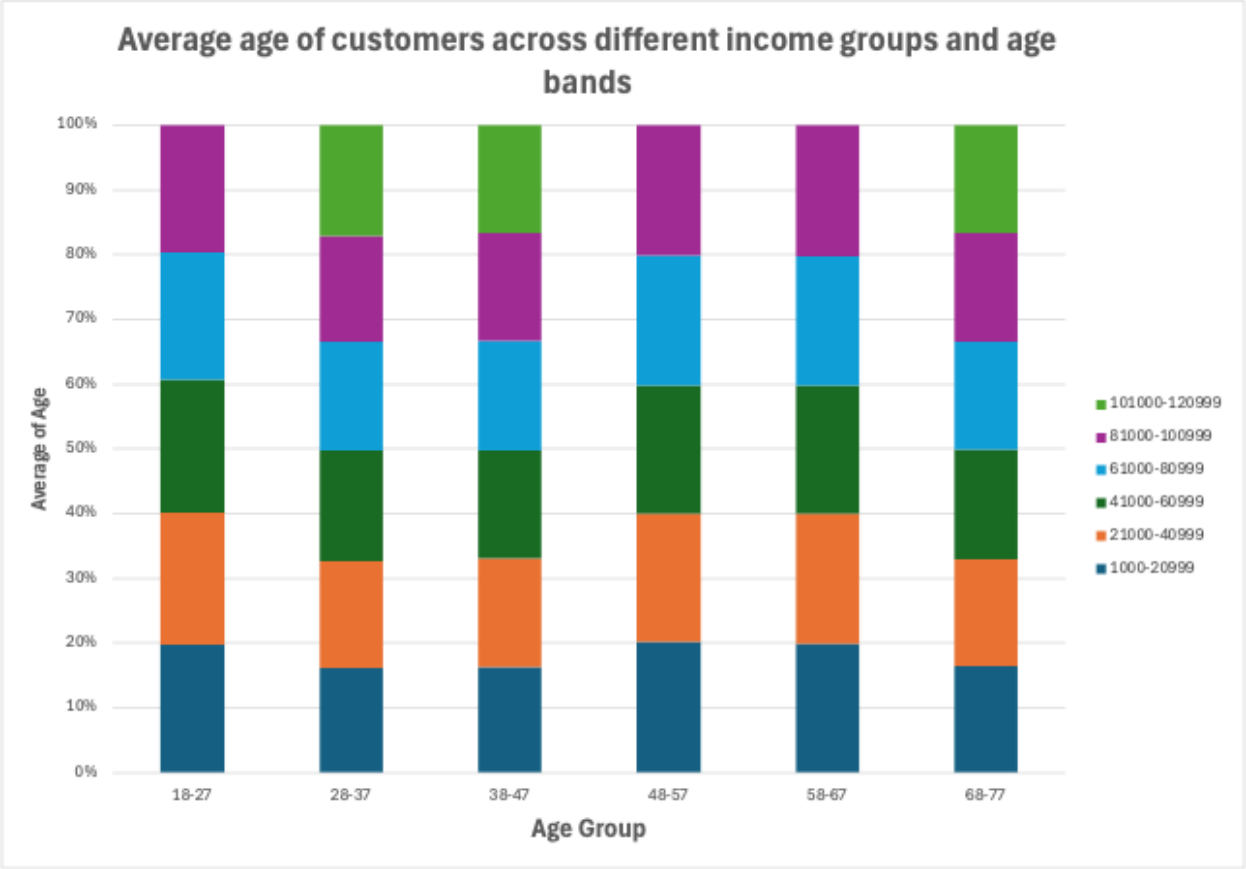
The data in the bar chart is arranged in descending order — from the highest number of customers **(Married)** on the left to the **lowest (Alone)** on the right.



The side-by-side bar chart illustrates the distribution of customer income across different age groups. It clearly shows that younger age groups (18–37) are primarily concentrated in the $21,000–$40,999 income range. However, as age increases, there is a noticeable shift, with the majority of customers aged 38 and above falling into the $41,000–$80,999 income bracket. This highlights a trend of increasing income with age.

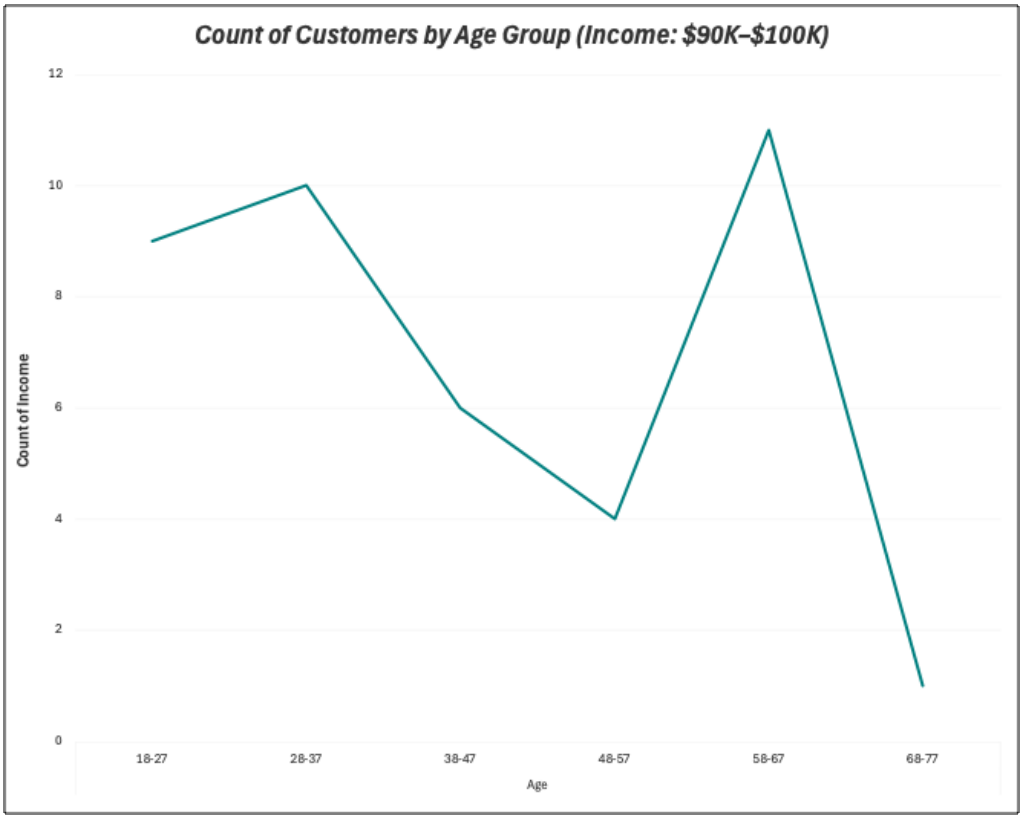


The stacked bar chart presents income distribution across various age bands. It reveals that the highest income brackets—specifically those earning above $90,000—are mostly concentrated within the 38–57 and 68–77 age groups. In contrast, younger age groups, particularly those aged 18–27, peak in the $75,000–$99,999 range but do not appear in the topmost income levels. This suggests that higher income levels are typically associated with older, more experienced customers.



The relationship between age and income for customers earning $90K–$100K shows no consistent pattern. While the highest number of earners falls in the 58–67 age group, other age groups also contribute significantly. The data distribution is irregular, indicating that high income is not strictly tied to age. Outliers include a noticeable dip in the 45–57 range and a sharp drop in the 65–77 group.

There is a significant dip in the 45–57 age group and a sharp drop in the 65–77 range, possibly due to retirement or reduced participation in the workforce. These patterns indicate variability and suggest that factors other than age—such as occupation, education, or location—may be more influential in determining income at this level.



Lesson 3 & 6

Dashboard Design and Development

The dashboard was built to analyse customer behaviour through demographics, spending patterns, and advertising performance. Bar charts were used to display age distribution, marital status, product sales, and spending by age group. These visuals revealed that the 32–48 age group forms the largest customer segment and accounts for the highest spend. Alcohol was identified as the top-selling product across all age ranges.

A symbol map visualised total customer spend by country using bubble sizing. This highlighted that Spain leads in sales while Montenegro has the lowest, informing regional performance comparisons.

A scatter plot was used to show the relationship between income and customer spend. It revealed minimal variation in spending below $60K income, but a clear upward trend beyond that point, showing a positive correlation between higher income and spending.

A pie chart showed the effectiveness of advertisement channels. Instagram, Facebook, and Bulk Mail accounted for the highest reach, while Brochure ads were the least effective, suggesting they can be discontinued. A filter-enabled bar chart allowed ad performance to be explored by age group, marital status, and country.

Calculated fields such as Total Spend by Customer, as well as binned age and income ranges, were created for simpler analysis. Filters were added to explore trends by country and demographic groups, increasing dashboard flexibility.

Lesson 4

1. Total Spend by Country

```
SELECT country,
       SUM(amttotal) AS total_spend
FROM marketing_data
GROUP BY country
ORDER BY total_spend DESC;
```

2. Product Spend by Country

```
SELECT country,
       SUM(amtliq) AS liq,
       SUM(amtvege) AS veg,
       SUM(amtnonveg) AS nonveg,
       SUM(amtpes) AS fish,
       SUM(amtchocolates) AS choc,
       SUM(amtcomm) AS comm
FROM marketing_data
GROUP BY country;
```

3. Product Spend by Country (Sorted by Liquor)

```
SELECT country,
       SUM(amtliq) AS liq,
       SUM(amtvege) AS veg,
       SUM(amtnonveg) AS nonveg,
       SUM(amtpes) AS fish,
       SUM(amtchocolates) AS choc,
```

```
        SUM(amtcomm) AS comm
FROM marketing_data
GROUP BY country
ORDER BY liq DESC;
```

4. Product Spend by Marital Status

```
SELECT marital_status,
        SUM(amtliq) AS liq,
        SUM(amtvege) AS veg,
        SUM(amtnonveg) AS nonveg,
        SUM(amtpes) AS fish,
        SUM(amtchocolates) AS choc,
        SUM(amtcomm) AS comm
FROM marketing_data
GROUP BY marital_status
ORDER BY liq DESC;
```

5. Product Spend by Household Composition (Kids and Teens at Home)

```
SELECT kidhome,
        teenhome,
        SUM(amtliq) AS liq,
        SUM(amtvege) AS veg,
        SUM(amtnonveg) AS nonveg,
        SUM(amtcomm) AS comm,
        SUM(amtpes) AS fish,
        SUM(amtchocolates) AS choc
FROM marketing_data
```

```
GROUP BY kidhome, teenhome

ORDER BY liq DESC;
```

Lesson 5

1. Ad Channel Usage by Country

```
SELECT country,
       SUM(twitter_ad) AS twitter,
       SUM(facebook_ad) AS facebook,
       SUM(instagram_ad) AS instagram
FROM marketing_data
JOIN ad_data USING (id)
GROUP BY country;
```

- **Twitter** is most used in **Canada, India, and Germany**.
- **Instagram** dominates in **Australia, Spain, and Saudi Arabia**.
- **Facebook** is most used in the **United States**.
- **Montenegro** had **no recorded ad engagement**.

2. Ad Channel Usage by Marital Status

```
sql
CopyEdit
SELECT marital_status,
       SUM(twitter_ad),
       SUM(facebook_ad),
       SUM(instagram_ad)
FROM marketing_data
JOIN ad_data USING (id)
GROUP BY marital_status;
```

- **Twitter** is preferred by **Widowed, Divorced, and Single** customers.
- **Instagram** is most used by those who are **Married or in a relationship ("Together")**.
- **Facebook/Instagram** are equally used by customers labeled as **Others**.
- **Alone** categories show **no ad engagement**.

3. Conversion and Sales by Country

```
SELECT country,
       SUM(count_success) AS total_conv,
       SUM(twitter_ad),
       SUM(facebook_ad),
       SUM(instagram_ad),
       SUM(amtliq),
       SUM(amtvege),
       SUM(amtnonveg),
       SUM(amtpes),
       SUM(amtchocolates),
       SUM(amtcomm),
       SUM(amttotal)
FROM ad_data
JOIN marketing_data USING (id)
```

```
GROUP BY country
ORDER BY total_conv DESC;
```

This query compares total ad conversions with product sales by country. **Spain** ranks highest in both **conversion rate** and **total sales**.
There's a general trend that countries with higher ad conversions also generate higher sales. However, **Saudi Arabia** and **Australia** deviate from this pattern—despite having lower conversions, they still report higher overall sales than countries ranked above them for conversion success.