2MARKET

DATA ANALYSIS BY ALASDAIR BELL

2MARKET

CONTENTS

A. IDENTIFY AND DEFINE THE BUSINESS PROBLEM

1. BACKGROUND

Introduction

2Market is an international supermarket. They operate in numerous countries and offers several product lines. 2Market would like to understand more about their customers' demographics, the effectiveness of their advertising channels and the relationship between customer demographics and their product lines.

The Business Problem.

2Market is investing time and financial resources into their marketing and advertising efforts, yet they lack clarity on whether these efforts are targeting the most valuable customers, in the most effective way.

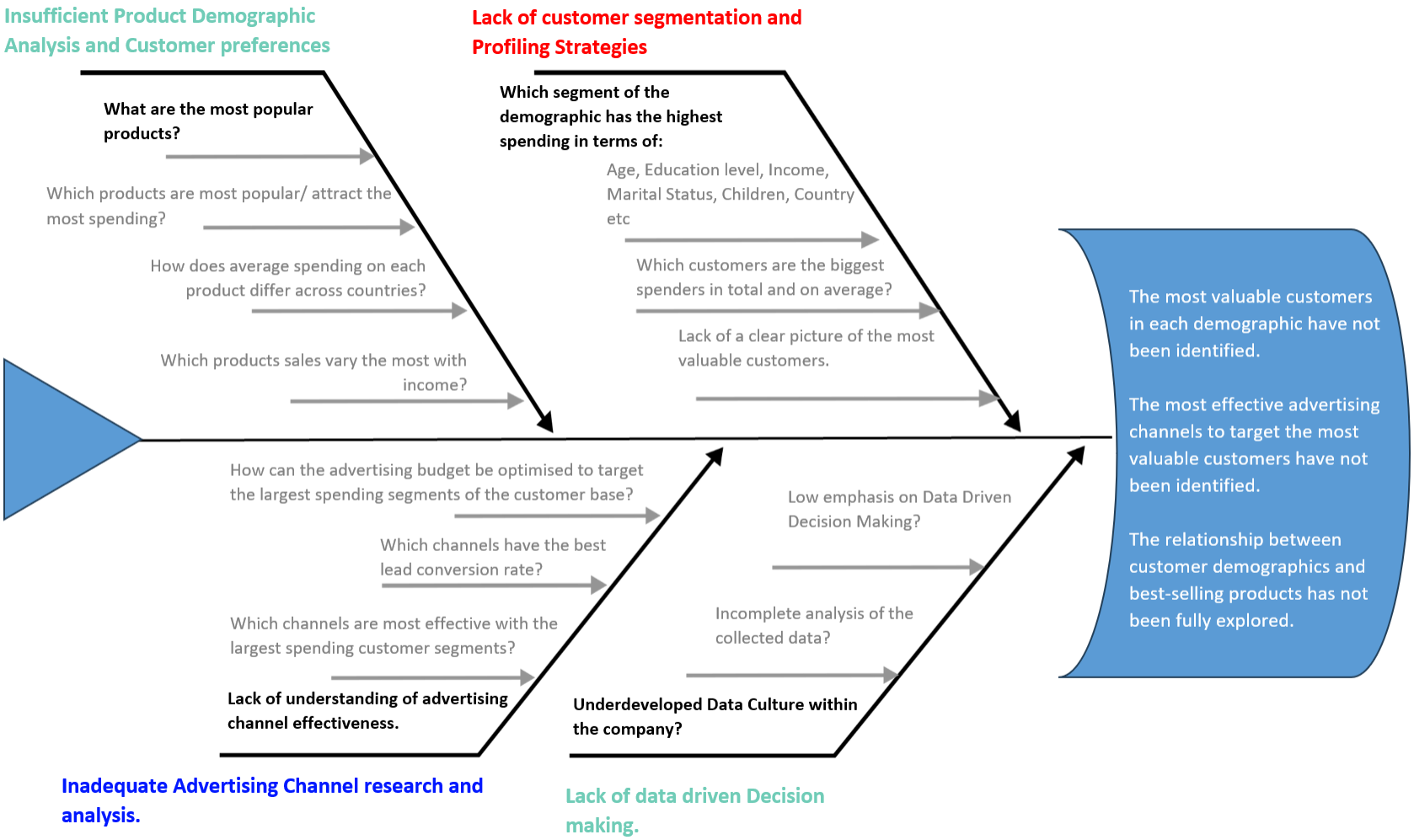
From a Business Problem to a Data Problem

The business problem can be further broken down:

- The most valuable customers in each demographic have not been identified.
- The most effective advertising channels to target the most valuable customers have not been identified.
- The relationship between customer demographics and best-selling products has not been fully explored.

B. EXPLORE AND DETERMINE THE ROOT CAUSE

The following fishbone diagram attempts to reframe the **Business Problem** as a series of **questions to ask of the data** and determine the root cause.



Limitations and Assumptions

To fully understand the business problem, I would require further information from 2Market.

Limitation	Assumption
What is the goal? Establish a need to solve the business problem. **Why** has 2Market asked for the analysis to be conducted?	2Market want to optimise their marketing and advertising budget to target the most valuable customers in the most efficient and cost effective way.
Justify the need to solve the business problem. **How** will 2Market use the results of the analysis and how does this align with their business objectives / strategy. What are the perceived benefits of the exercise?	2Market will use my analysis as the basis for a targeted marketing / advertising campaign. The analysis will enable 2Market to identify the most valuable customers, how to target them and what they tend to buy. My analysis will save them time and money.
Who benefits? I would like to know who the analysis is for so that I may **identify, assess and prioritise stakeholders.**	The analysis is for the marketing and advertising directors and will be used to justify a budget allocation from senior management. The analysis will improve their results. I will be preparing and presenting directly for them.
Business Context: I would like to understand the organisational and wider market context within which the problem has arisen.	2Market senior management have questioned whether the advertising expenditure is being well spent. They would like to maximise profits, increase efficiency and reduce costs.
How can I ensure Data Quality? How was the data collected, who collected it and what proportion of the whole customers base does the data represent.	The data has been collated by a reputable source and the volume of data is sufficiently large to infer insights about the wider population of 2Markets customers
Timeframes - when does the analysis need to be completed? When will the results be formed into an actionable plan?	The analysis needs to be completed expeditiously in time to be ingested before the advertising campaign commences.
What measures will be used to evaluate the analysis?	KPIs such as Accuracy, Completeness, Timeliness, Relevance and Actionability.
Cost of Sales in each country not available. Therefore country by country comparison is not accurately standardised. E.g. Germany has a higher average spend, but does it cost 2Market more to attain?	Assumed parity in the cost of doing business in each country.
Cost of Sales information is not available. Therefore profit margin cannot be calculated. This makes product by product comparison ineffective (vanity not sanity).	Assumed that profit margin is fairly consistent.

C. DEVELOP ALTERNATIVES

- More market research - assess customer preferences and preferred advertising channels.
- A/B testing of advertising channels.
- Sales team questionnaires to determine customer feedback.
- Thorough investigation of existing data.

D. SELECT A SOLUTION

- Thorough investigation of existing data.

E. IMPLEMENT

2. ANALYTICAL APPROACH

- Data Preparation: Validating, cleaning and organising the data in Excel.
- **Identify the most valuable customers in each demographic segment**.
- **Identify advertising channels to target the most valuable customers.**
- **Explore the relationship between customer demographics and best-selling products.**

2.1 Data Preparation

Data Validity

- Cross referenced with the metadata provided to ensure that the data was valid.
- Split into 24 fields.
- A data dictionary prepared:

	Field Name							
	ID	Year_Birth	Age	Education	Marital_Status	Income	Kidhome	Teenhome
Definition	Unique customer ID.	Customer's year of birth.	Age in years calculated with =YEAR(TODAY())-[@[Year_Birth]]	Educational qualification of the customer.	Customer's marital status .	Customer's annual income	Number of kids the customer has.	Number of teenagers the customer has.
Data Type	Number	Number	Number	Text	Text	Currency	Number	Number
Sample Value	5642	1980	29	Master	Together	$ 62,499.00	1	1
Mandatory	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Dt_Customer	Recency	AmtLiq	AmtVege	AmtNonVeg	AmtPes	AmtChocolates	AmtComm
Definition	Date of customer's registration with the company.	Number of days since customer's last purchase.	Amount spent on alcoholic beverages.	Amount spent on vegetables.	Amount spent on meat items.	Amount spent on fish products.	Amount spent on chocolates.	Amount spent on commodities.
Data Type	Date	Number	Number	Number	Number	Number	Number	Number
Sample Value	12/09/2013	99	140	4	61	25	30	197
Mandatory	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	NumDeals	NumWebBuy	NumWalkinPur	NumVisits	Response	Complain	Country	Count_success
Definition	Number of deals purchased made with a discount .	Number of purchases made from the website.	Number of in-store purchases.	Number of website visits per month.	If the customer had accepted the last campaign's offer (1) or not (0).	If the customer had complained in the last 2 years (1) or not (0).	Customer's location.	Total number of successful lead conversions.
Data Type	Number	Number	Number	Number	Boolean	Boolean	Text	Number
Sample Value	2	3	6	4	1 or 0	1 or 0	SP	1
Mandatory	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Data Clean in Excel

- Cleaned in accordance with the following itemised and repeatable procedure.
- Saved as separate worksheets after each stage to enable tracking and recourse to previous iterations.

Stage	Assumptions	Actions	Records
Data Accuracy			**2216**
Check for spelling errors.		Only Field headings in relation to the Data Dictionary were erroneous. Maintained for consistency.	
Values out of range *Check for initial and obvious outliers prior to full range check.*	Year_Birth : 1894, 1900 and 1901 assumed erroneous.	Removed.	-3
	Income: $666,666 assumed erroneous.	Removed.	-1
	Marital Status: YOLO and Absurd assumed erroneous.	Removed.	-4
	Outlier analysis conducted for Age and Income (see below)		
	Ages above 90.	Already removed above	
	Income above $117,360.	Removed	-3
Incorrect or invalid data types		Data types were checked using the 'Type' function in Excel. Income field changed from text to currency using find and replace to remove $ and formatting to currency. Dt_Customer field changed to date using 'textsplit' and 'date' function in Excel.	
Blank cells or spaces		Checked using Go to and trim functions. No blank spaces were found.	
Incorrect use of nulls		No observations found.	
Incorrect calculations		No observations found.	
Mistypes and other format errors.		Checked.	
Data Completeness			**2205**
Blanks in a required field		No observations found.	
Partial or incomplete data		No observations found.	
Incorrect or invalid calculations		No observations found.	
Missing values.	**One notable demographic was not included in the data namely, gender.**	No other observations found.	
Data Consistency			
Precision	The ID number (Unique Customer ID) is primarily composed of 4 digits (1756/2216 records). It seems inconsistent to have values that do not correspond to this format. However, nothing is specified in the metadata.		
Structure of data		Checked. Consistent.	
Case sensitivity		Checked. Consistent.	
Data type		Checked. Consistent.	
General Consistency	**2Market has two sales channels: In-store purchases and purchases made from the website. Filtering for zero on both channels results in 10 records that have used neither. This is inconsistent and those records were removed.**		**-10**

Data Uniqueness			**2195**
Entries with the same spelling but in a different case.		No observations found.	
Entries with different spelling.		No observations found.	
Different words but with the same meaning.	**Marital Status 'Alone' and 'Single' were assumed to be of the same meaning and therefore merged into 'Single'.**		
Words with alternate representations.		No observations found.	
Duplicate Values	**The customer IDs were all unique. However, excluding ID from the 'remove duplicates field', 47 duplicates were found and removed.**		-47
	Upon further inspection of the field data with conditional formatting, many other duplicates / triplicates were found. These records were identical other than in one field, for example the Response or Country field.	Use conditional formatting to highlight duplicate values. Remove Duplicates in Data tab.	
	Without knowing which record is correct, both records are unreliable for analysis and may skew the results. Therefore, both records of each duplicate were removed.		-304
Data timeliness			**1844**
Correct date format and type.		Dt_Customer field changed to date using 'textsplit' and 'date' function in Excel.	

Outlier Analysis

The Income and Age fields were tested for outliers.

Age	
QTR 1	45
QTR 3	63
Range	18
Lower	18
Upper	90

Income	
QTR 1	35782.25
QTR 3	68413.25
Range	32631
Lower	-13164
Upper	117360

Removed:

- Ages above 90 and below 18
- Incomes above $117,360

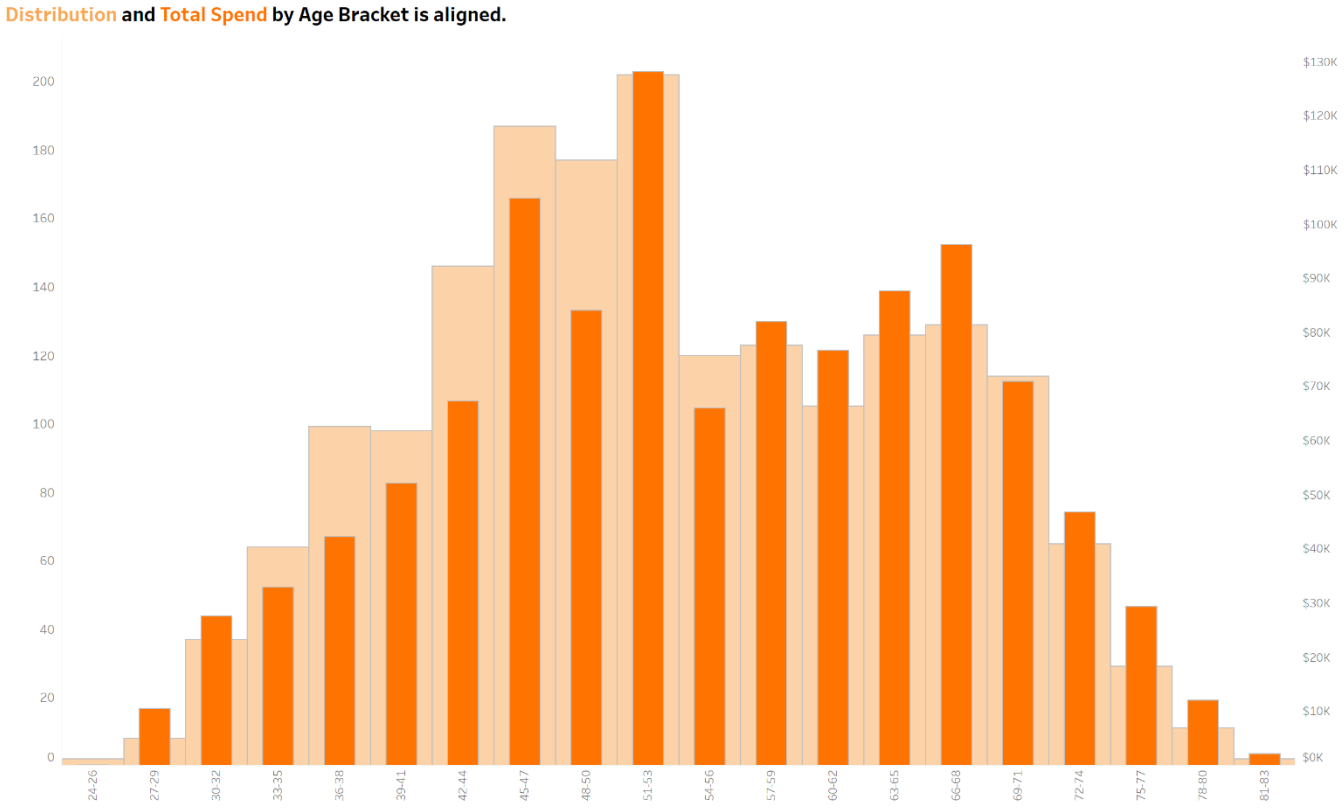
- The Ad data was found to be without duplicates and generally in good order.
- Exported to Tableau
- Tables created in SQL (See Appendix).

2.2 Identify the most valuable customers in each demographic segment.

The following analysis asks questions of the data and uses the dual axis function in Tableau. Sum, count, average aggregations were used, and data was aggregated into bins where necessary.

What age bracket has the most customers and is this aligned with spending?

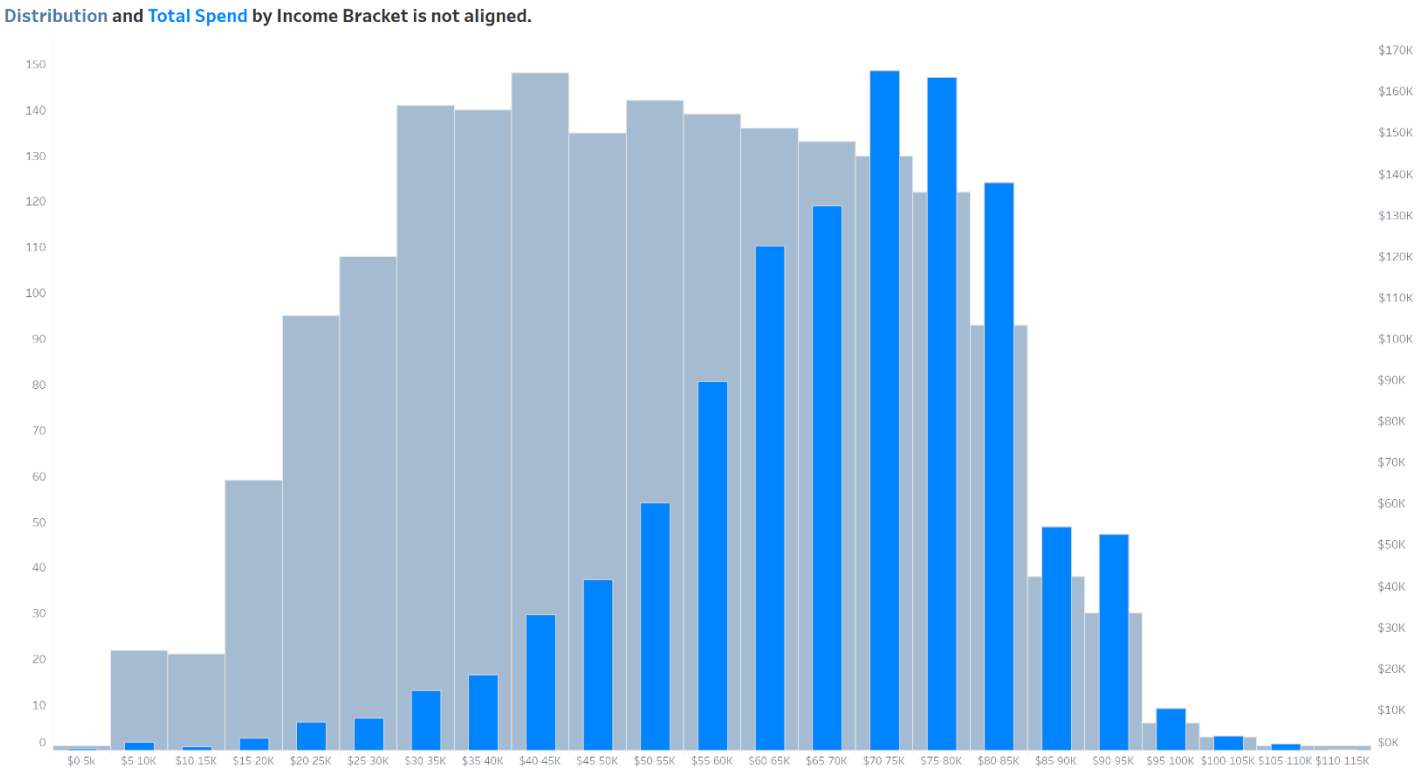
- 51-53 year olds.
- Yes



SUM(Age)	
Average:	53.17
Median:	52.00
Standard deviation:	11.62

What income bracket has the most customers and is this aligned with spending?

- $40-45K
- No. Highest spending bracket $70-75k

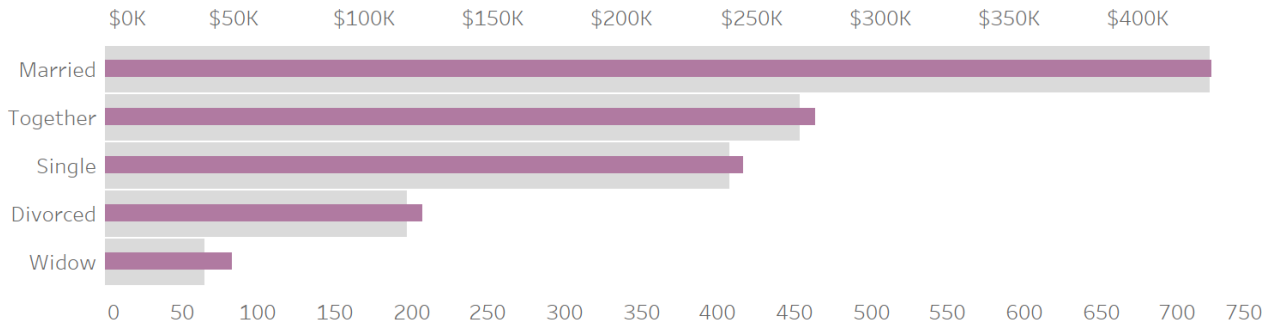


SUM(Income)	
Average:	$51,891
Median:	$51,684
Standard deviation:	20,607

Which Marital Status and Education segments have the most customers? Is this aligned with spending?

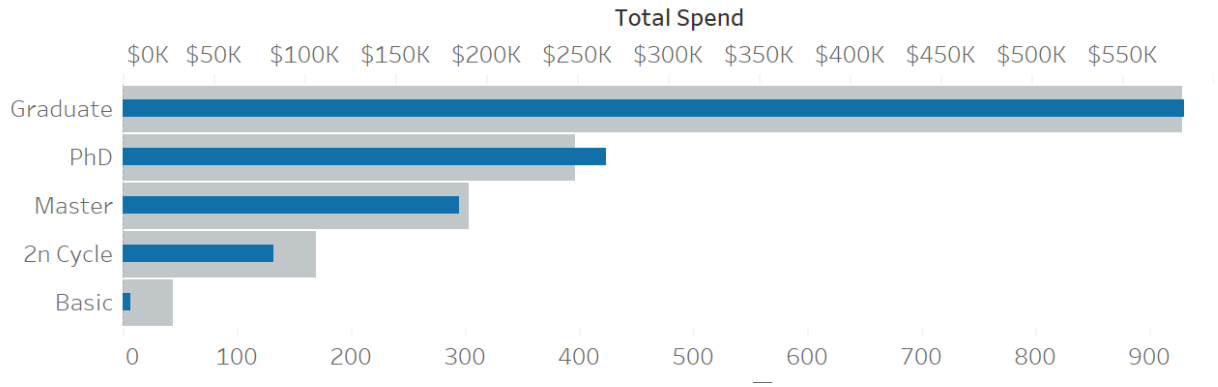
- Married, Graduate
- Yes

Marital Status Distribution vs Spend Distribution



	Customers	Total Spend
Married	721	$428,075
Together	453	$275,008
Single	408	$246,829
Divorced	197	$122,723
Widow	65	$48,855

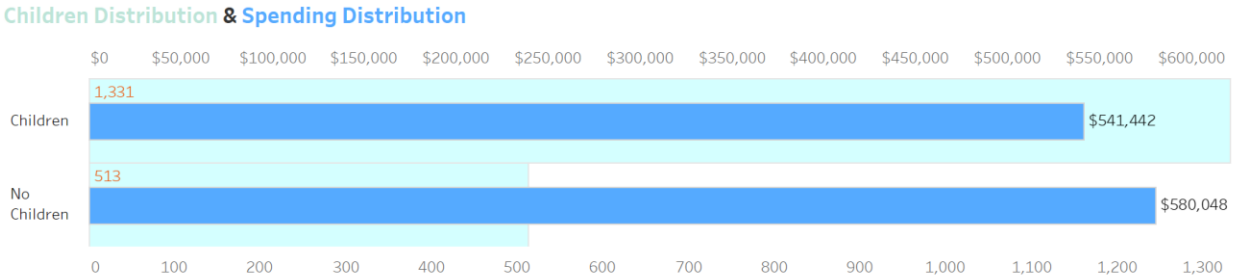
Education Distribution vs Spend Distribution

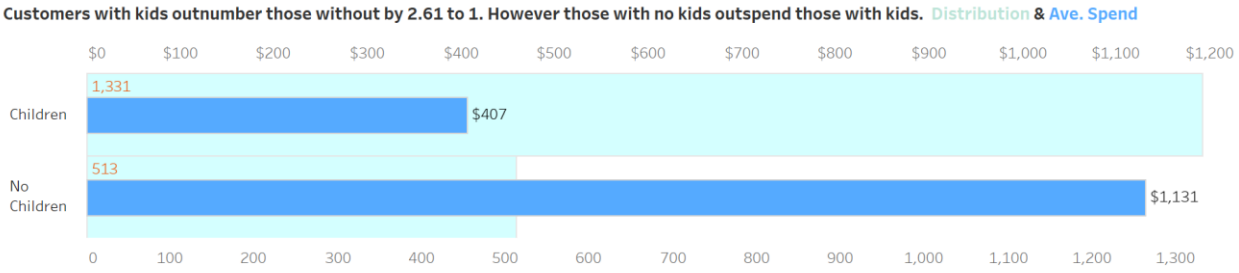


	Customers	Total Spend
Graduate	930	$583,556
PhD	397	$266,042
Master	303	$185,149
2n Cycle	170	$82,936
Basic	44	$3,807

How many customers have children? Is this aligned with spending?

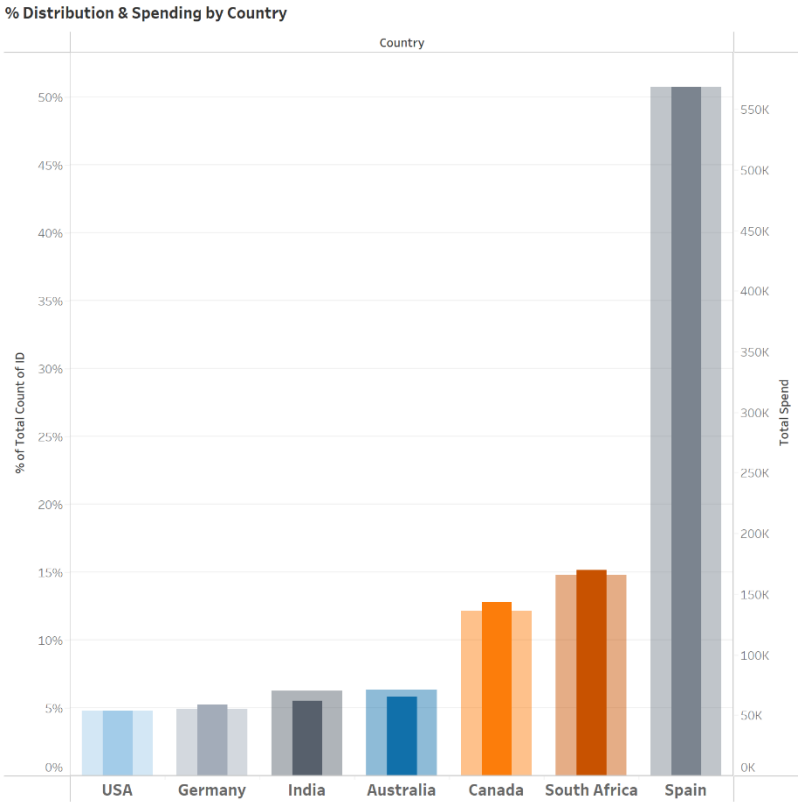
- 72%
- No. Customers without children spend more.





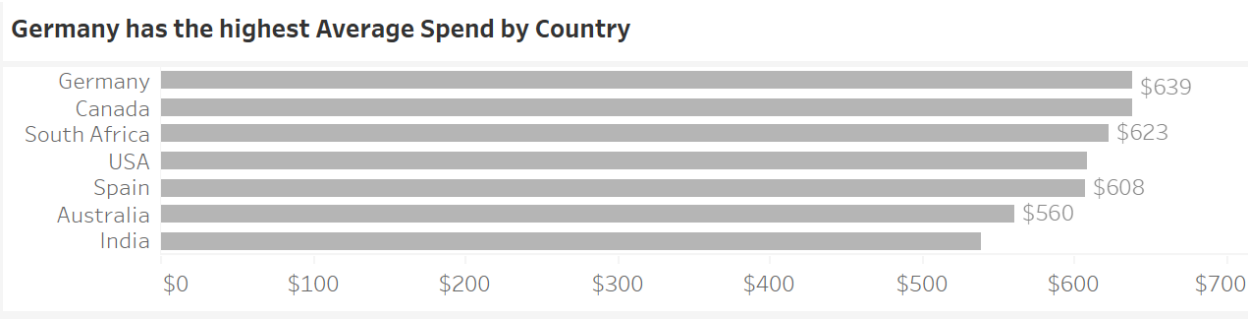
Which country has the most customers? Is this aligned with spending?

- Spain.
- Yes



NB. Montenegro was excluded from the analysis (only 1 record).

Which country's customers spend most on average?



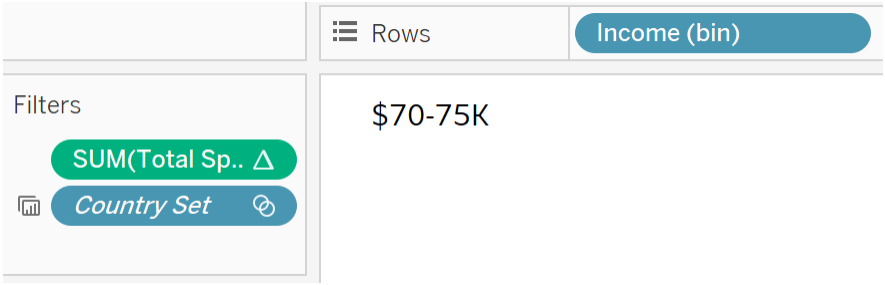
Summary

E.g.

Highest Spending Income Bracket	$70-75K
Highest Spending Age Group	51-53
Highest Spending Marital Status	Married
Highest Spending Education Level	Graduate

- Table calculation: Rank by Sum of Total Spend.
- Filter for Rank 1.

E.g.

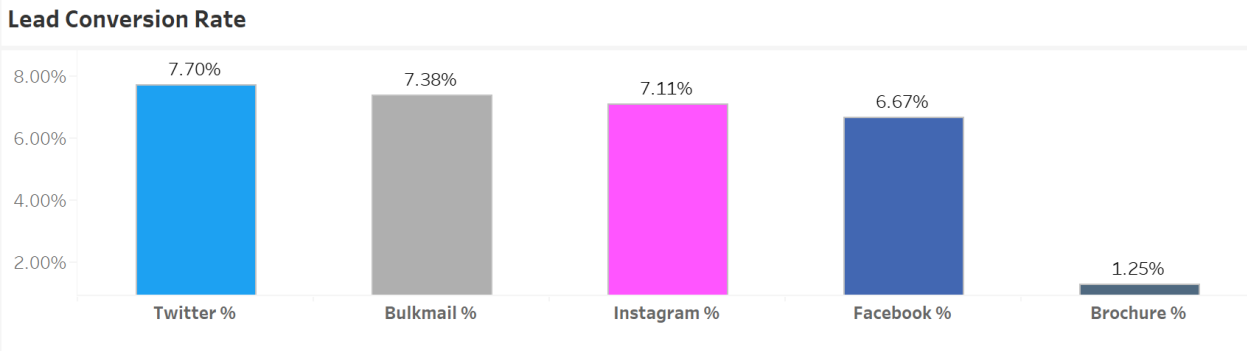


2.3 Identify which advertising channels to use to target the most valuable customers.

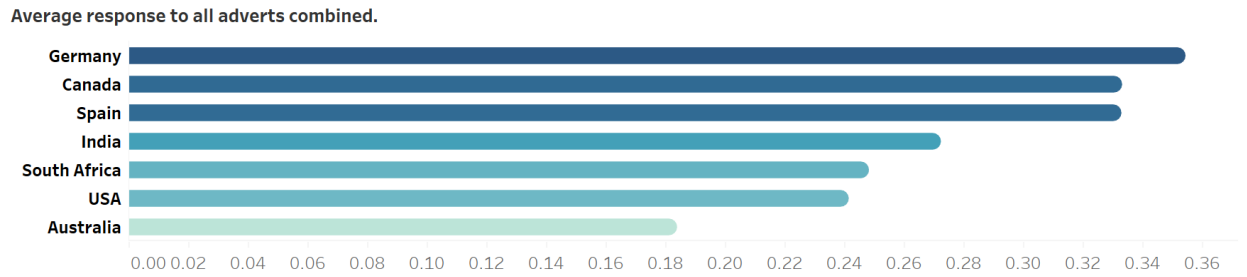
How can we determine the most effective ad channel?

- The percentage success rate of a channel - Lead Conversion Rate

```
SUM([Facebook ad])/COUNT([Facebook ad])
```



Which countries respond well to advertising channels in general?



Do Successful lead conversions translate into higher spending?

- It appears so.

Successful:

```
SUM(IF[Count success] != 0 THEN [Total Spend] END)/COUNT(IF[Count success] != 0 THEN [Total Spend] END)
```

Unsuccessful

```
SUM(IF[Count success] = 0 THEN [Total Spend] END)/COUNT(IF[Count success] = 0 THEN [Total Spend] END)
```

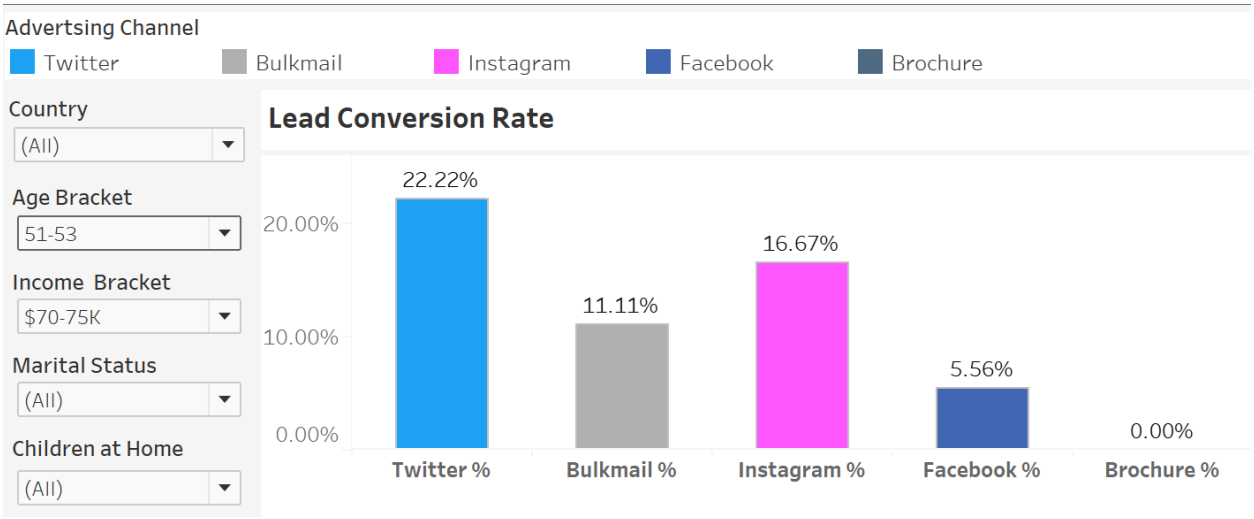
Can we apply the above to individual ad channels?

- No. Some spending is recorded with lead conversion success in multiple ad channels. The proportion between ad channels cannot be determined.

Which ad channels are best to target the highest spending demographics?

A filtering system was implemented to find the highest lead conversion rate.

E.g. 51-53 year olds in the $70-75k income bracket:



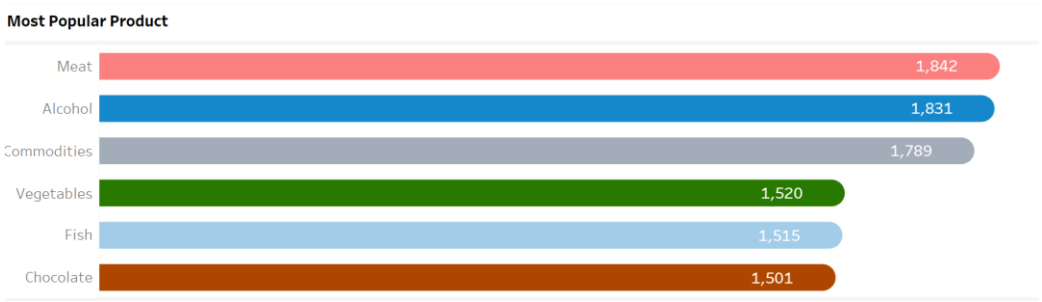
In this instance 2Market might consider concentrating on Twitter and Instagram.

2.4 Explore the relationship between customer demographics and the best-selling products.

- **Limitation:** No “cost of sales” info therefore no profit margin calculation.

What is the most popular product (most often bought)?

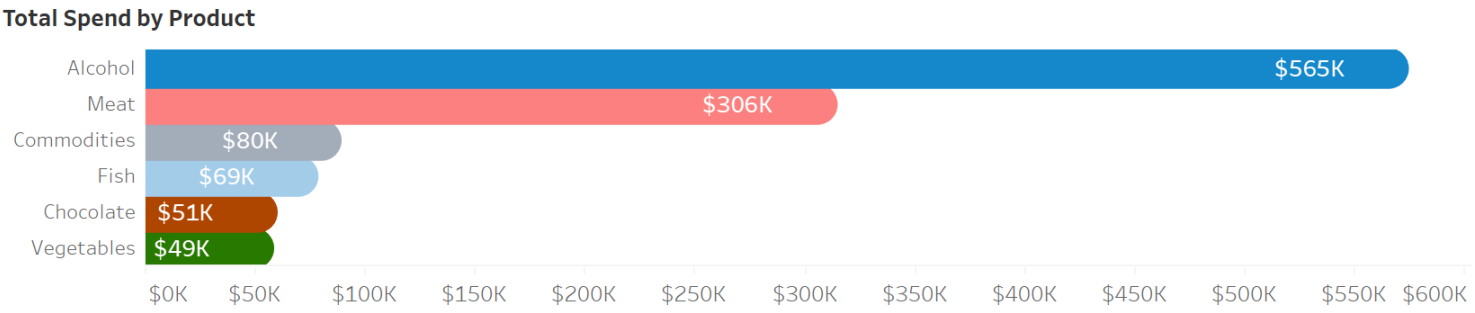
- **Meat**



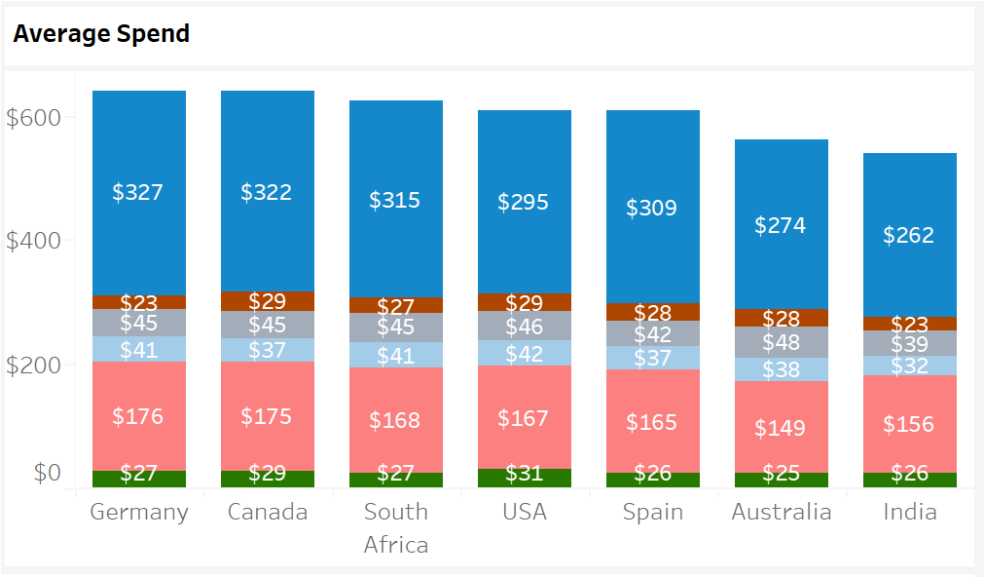
- See Appendix for SQL script.
- Zero values excluded using calculated field:

```
IF [Amount] != 0 THEN [Amount]
ELSE NULL
END
```

In total how much is spent on each product?

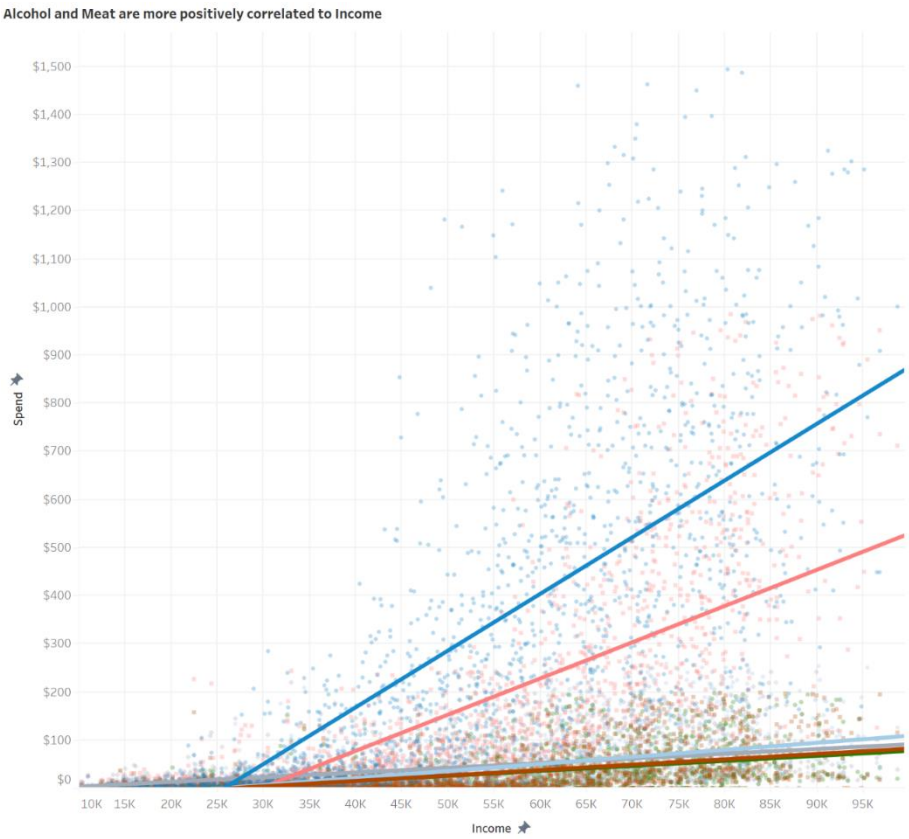


What does the average spend per product per country?



Do the products differ in their correlation with income?

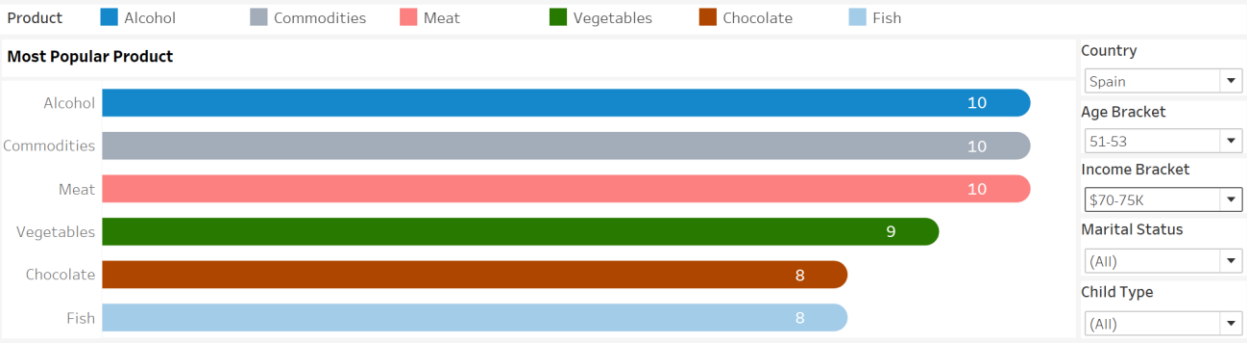
- **Yes. Alcohol and meat are more affected by income.**



Can we answer the questions above for specific demographics and countries?

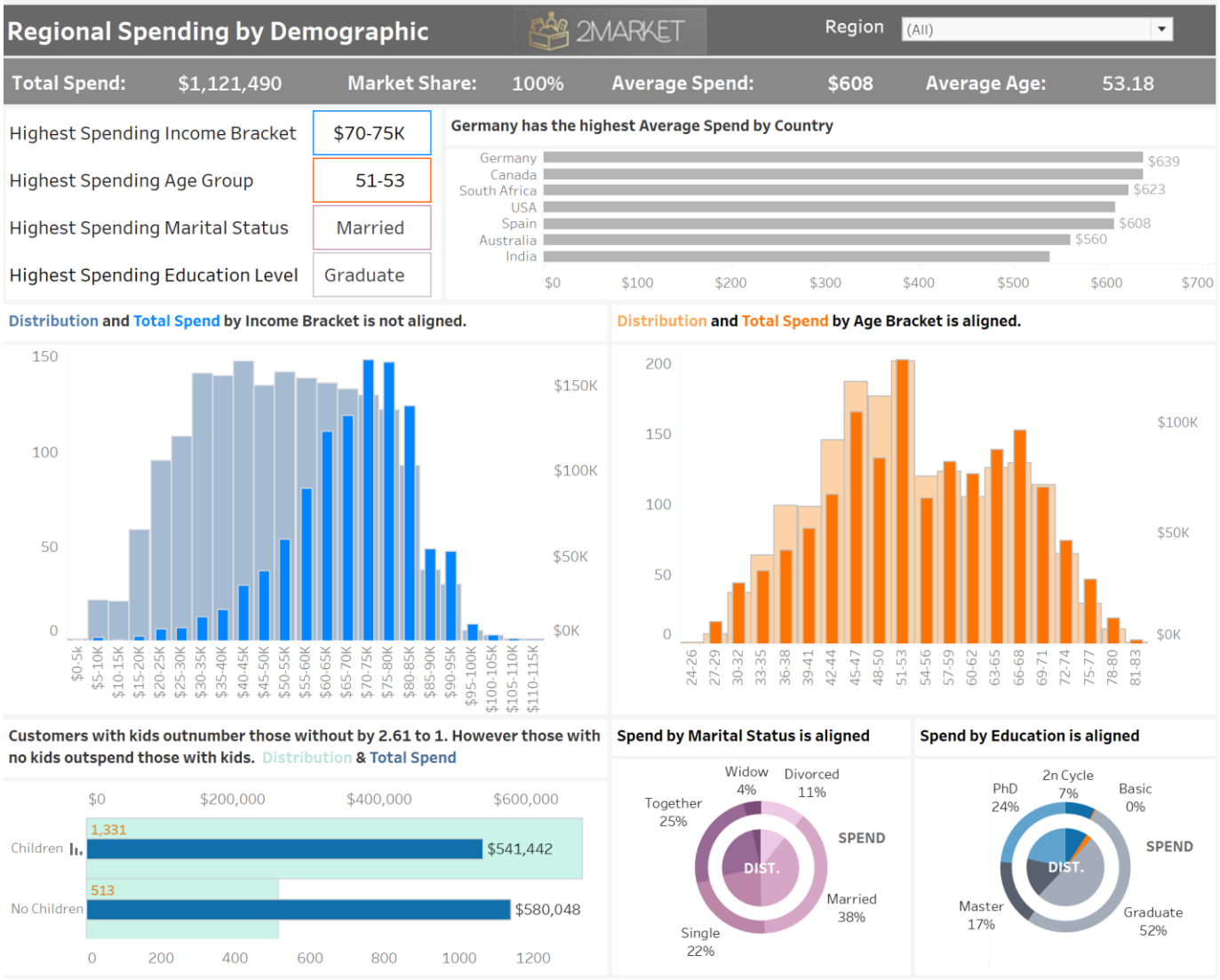
- Yes. By using filtering.

E.g.



3. Dashboard Design and Development

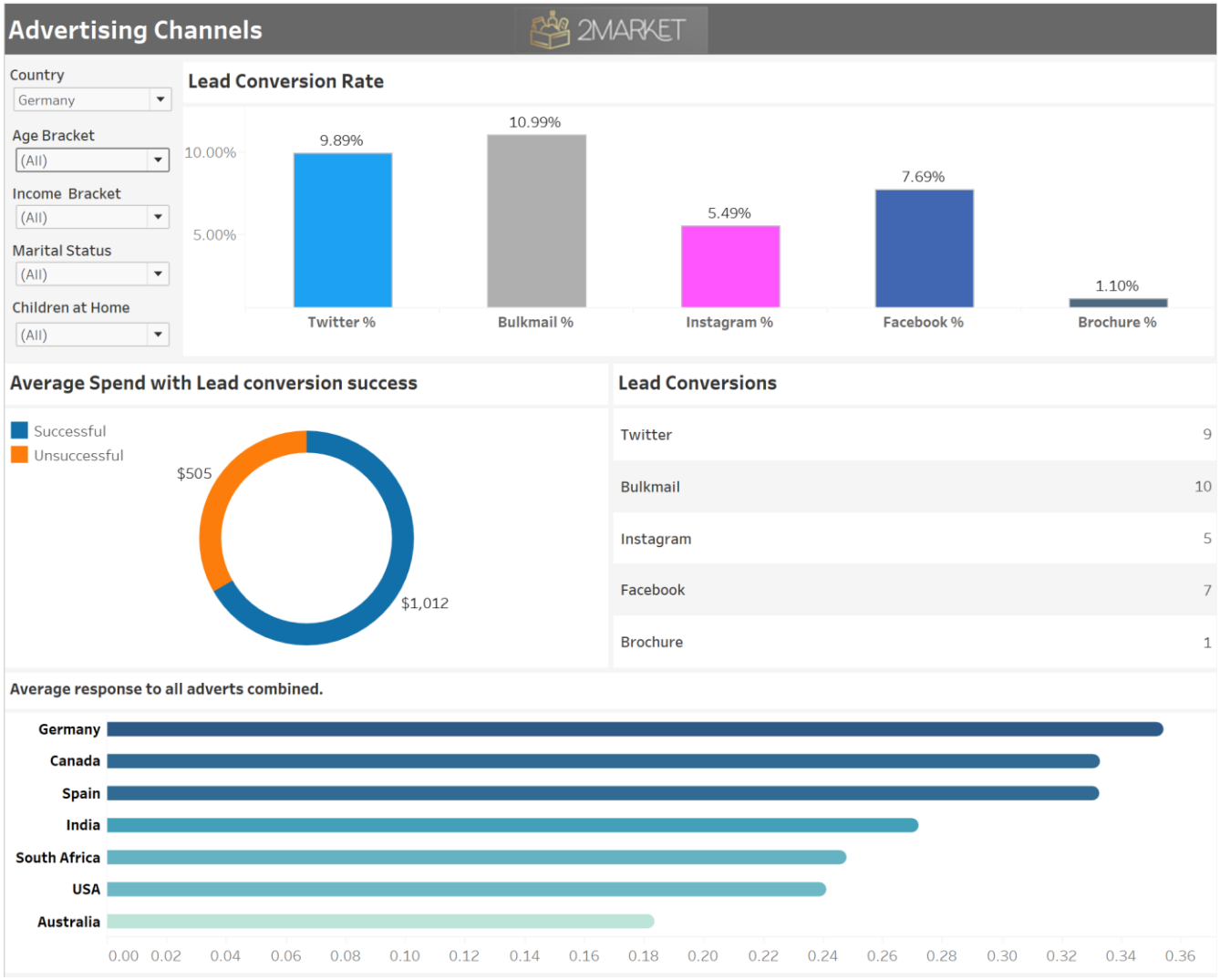
Due to the volume of information presented, 3 dashboards were produced. The above analysis justifies each chart's inclusion in the following dashboards:



Aim:

- **To Identify the highest spending customers in each demographic segment**.

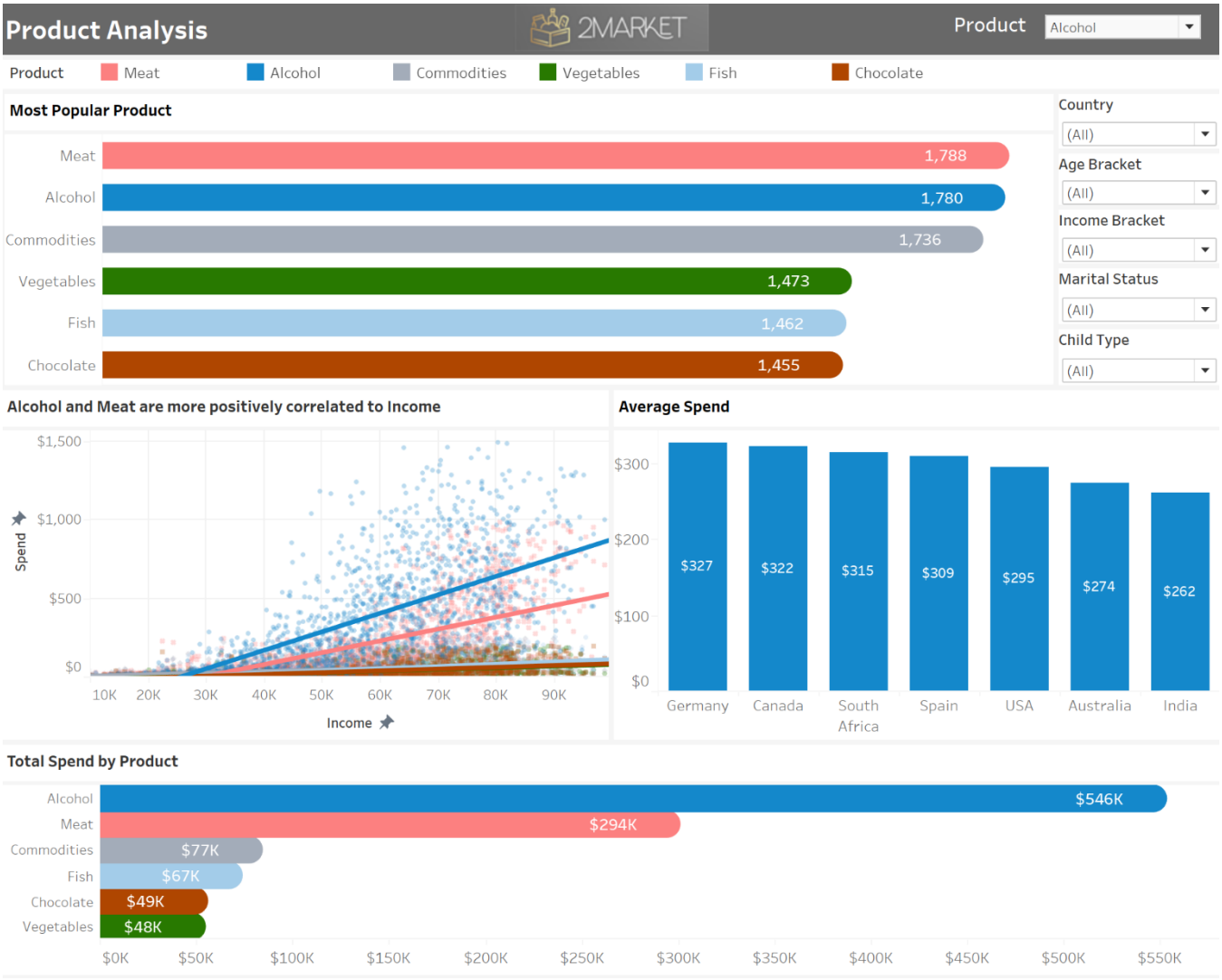
- To illustrate how demographic metrics relate to spending in each country.
- To use this information to inform the next 2 dashboards.



Aim:

- **To illustrate which advertising channels to use to target their highest spending customers.**

- To show whether lead conversion success translates to higher spending.
- To illustrate which countries respond well to advertising.



Aim:

- **To illustrate the relationship between customer demographics and products.**

- To highlight the most popular products, the highest spend per products, average spend per product per country and the correlation between product spend and income.
- To be fully customisable and enable filtering so that the best customers product preferences can be ascertained.

Notes on Dashboard Accessibility

Perceivable	Colours, fonts and text that allows visually impaired accessibility. I selected colours on the Ad effectiveness that coincide with the branding of the companies involved. On the Product Analysis dashboard I selected the colours that represent each product.
Operable	Ensured that the dashboard is navigable and operates as it should. Interactive elements have been labelled. Design is self explanatory. Most pertinent information at the top left corner of each dashboard.
Understandable	Dashboard shown to a 10 year old. He was able to understand and operate the dashboards.
Robust	Allowed for interpretation of the dashboard by a variety of user agents.

4. Patterns, Trends and Insights

Highest Spending Customers

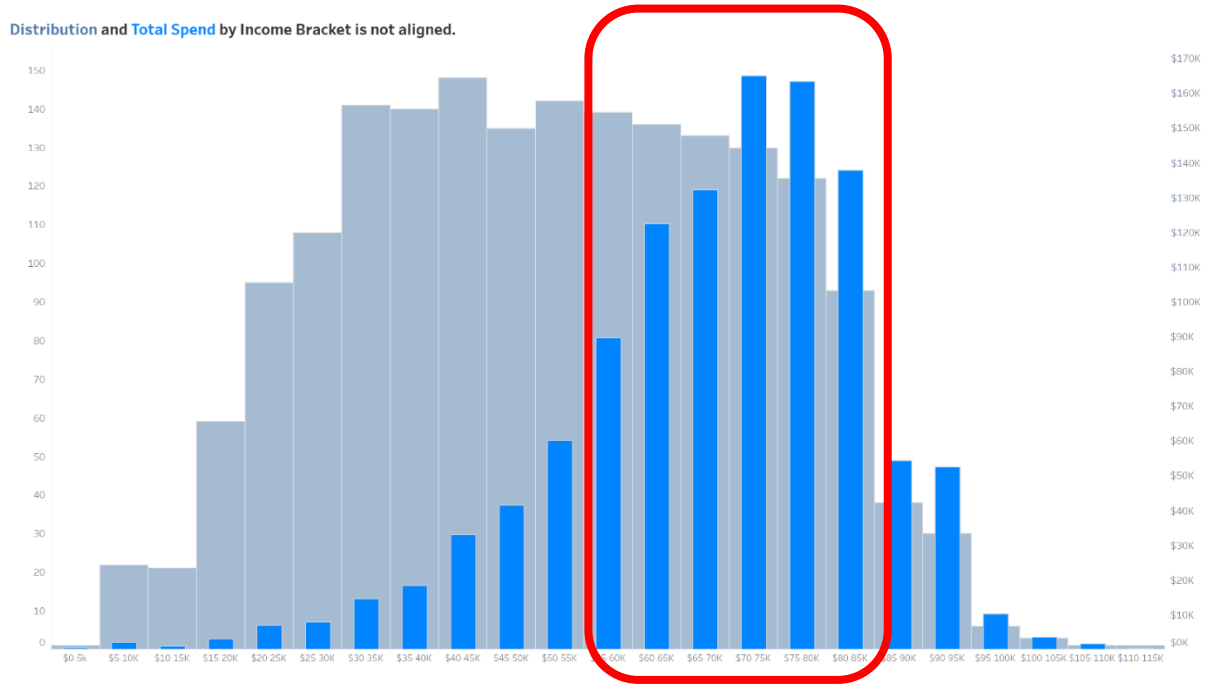
Insight
The demographic segment with the highest spending customers varies significantly in each country.

Recommendation
Target the advertising channels on a country by country basis.

Insight
Spending distribution is not always aligned with customer distribution.

Recommendation
Target the highest spending and highest customer distribution overlap.

E.g.



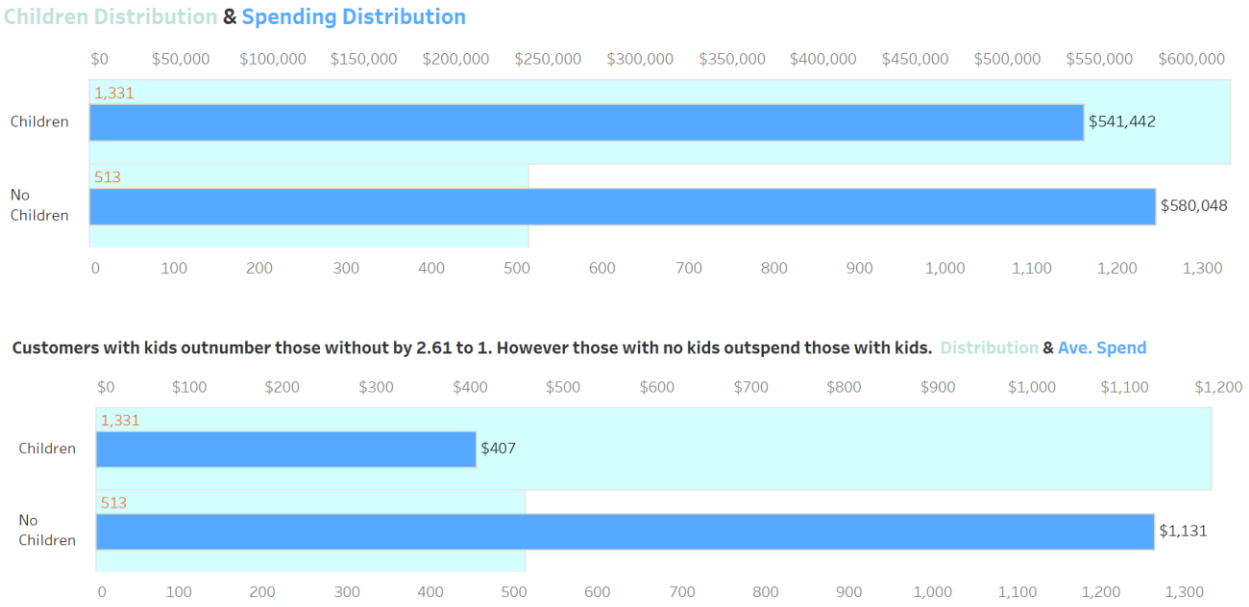
The income bracket with the most customers is $40 – 45K with 148 customers. The highest spending income bracket is $70-75K and has 130 customers. The $70-75K bracket outspend their more numerous counterparts in the $40 – 45K bracket by 5:1. There is an obvious asymmetry occurring in the income demographic.

Insight

Those with no children on average spend more than twice as much as those with children. Those with children outnumber those without. Expected distribution considering average customer age.

Recommendation

Diversify the customer targeting to account for this anomaly.

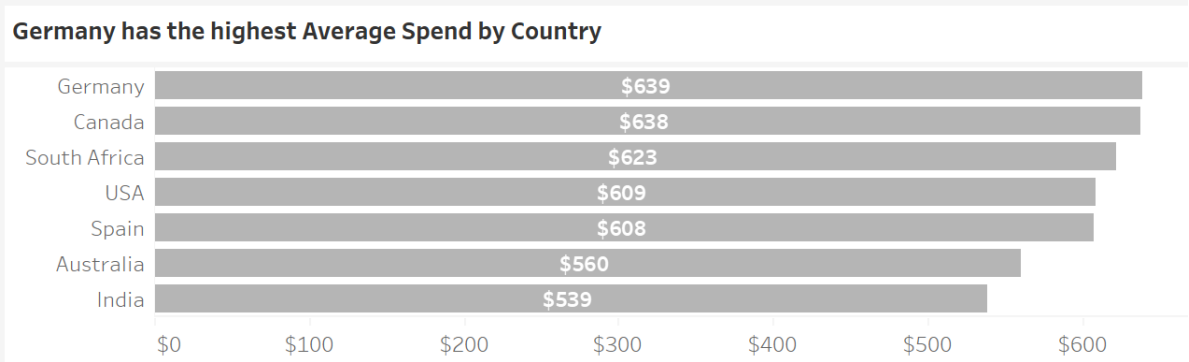


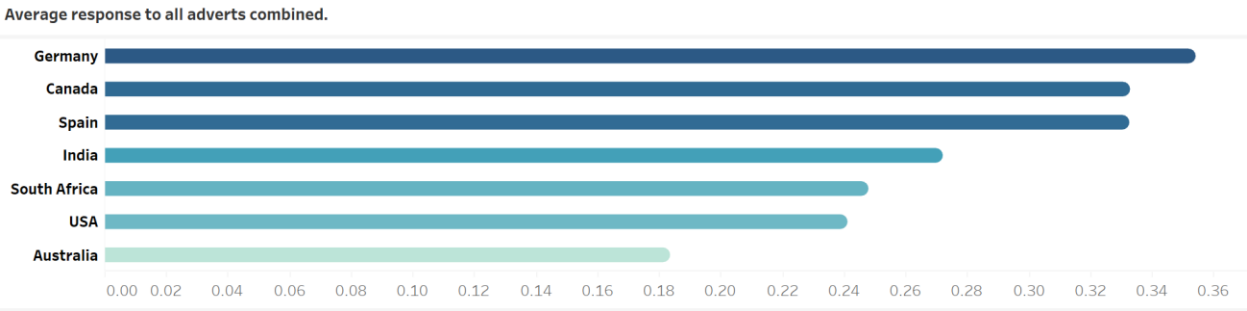
Insight

The average spend in Germany and Canada is above average. These countries also respond well to advertising.

Recommendation

Potential to optimise ad channel exposure in high spending countries.





Product Analysis

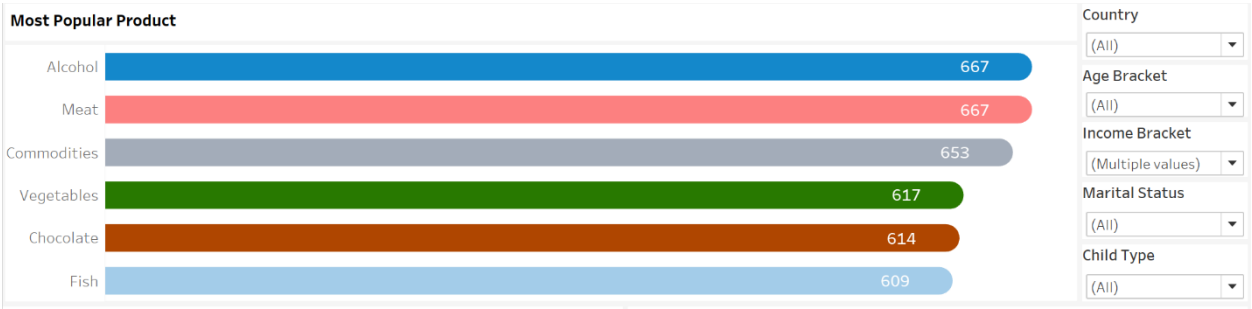
Insight

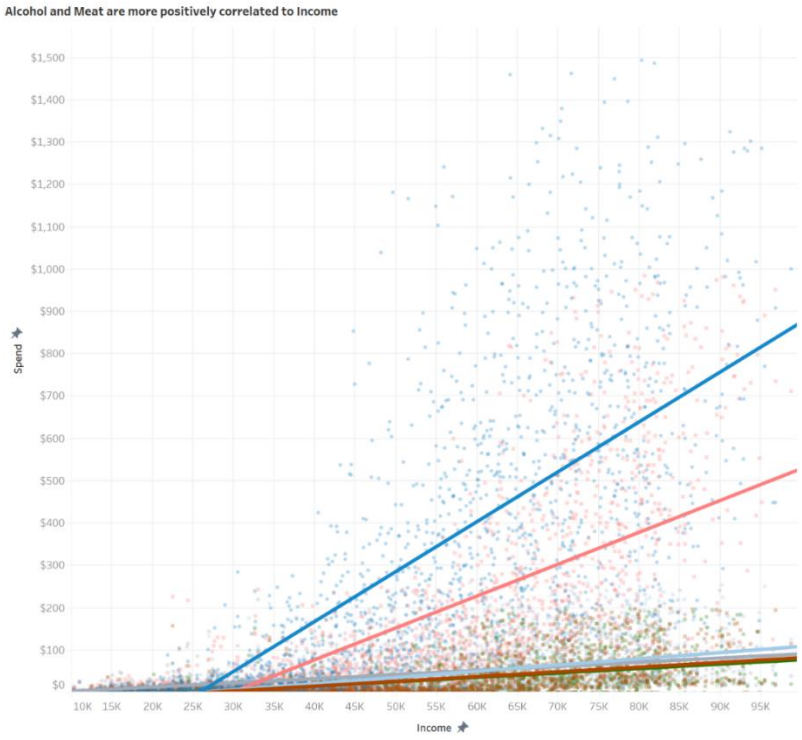
Alcohol and meat are favoured by the highest spending customers. They are also more positively correlated to income.

Recommendation

Use adverts referring specifically to meat and alcohol in the ad channels that favour high income segments.

Incomes over $60K



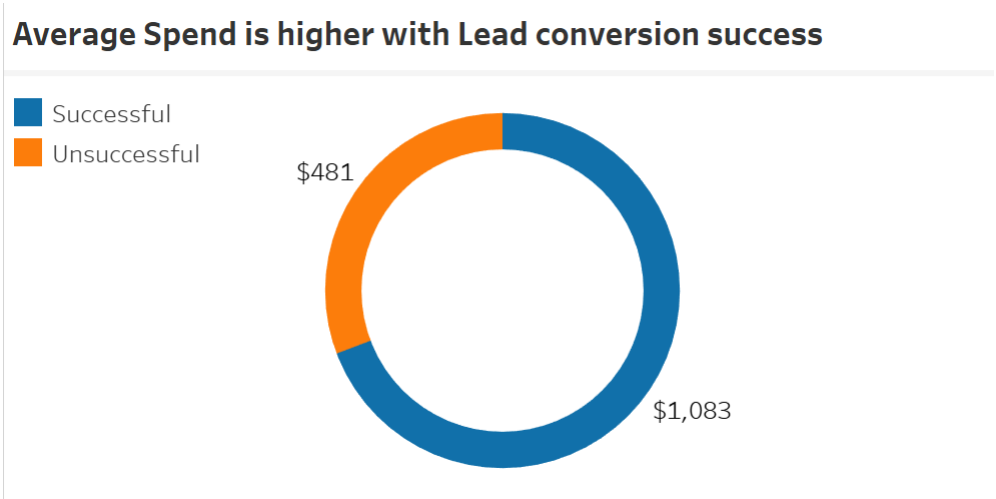


Insight

Average Spend is twice as high when accompanied with a successful lead conversion. Many records have multiple lead conversions so it is difficult to differentiate between ad channels.

Recommendation

Advertising seems to yield greater spending. Isolate the ad channels that result in higher average sales and concentrate on deploying them.



5. The Business Problem

Business Problem	Recommendation
The most valuable customers in each demographic have not been identified.	Use the Regional Spending by Demographic Dashboard to identify the most valuable customers by demographic in your chosen Country.
The most effective advertising channels to target the most valuable customers have not been identified.	Take the information above and use it to filter the Advertising Channels Dashboard and display the most effective channel to use.
The relationship between customer demographics and best-selling products has not been fully explored.	Select the desired demographic in the filter system in the Product Analysis dashboard. View the most popular products, the total amount spent and the average spent per country of your chosen demographic segment.

F. EVALUATE

To the best of my knowledge the analysis is:

- Accurate
- Complete
- Timely
- Relevant
- Actionable