Assignment for Module 1: Data Analytics for Business

Problem statement

2Market is a supermarket with physical and online outlets. They need to understand:

- The demographics of their customers and how product sales vary by demographic
- Which advertising channels are most effective

The benefits will be that 2Market can maintain a stock of key products, customize offerings based on demographic insights, and focus on effective advertising channels.

Initial questions

- Are there any data security concerns about sharing this data?
- Does 2Market have any data about the number of units purchased and pricing?
- What is the audience for the analysis and how familiar are they with data analytics?

Additional questions it might be useful to answer

- Do the best-selling products vary by country?
- Does the effectiveness of social media advertising platforms vary by country?

Jonathan Shields

Analysis of the data using Excel and SQL

Excel was selected for data cleaning and exploratory analysis as it is suited to working with lower data volumes. Grouping and joining of data was performed using SQL as it more suitable for this purpose.

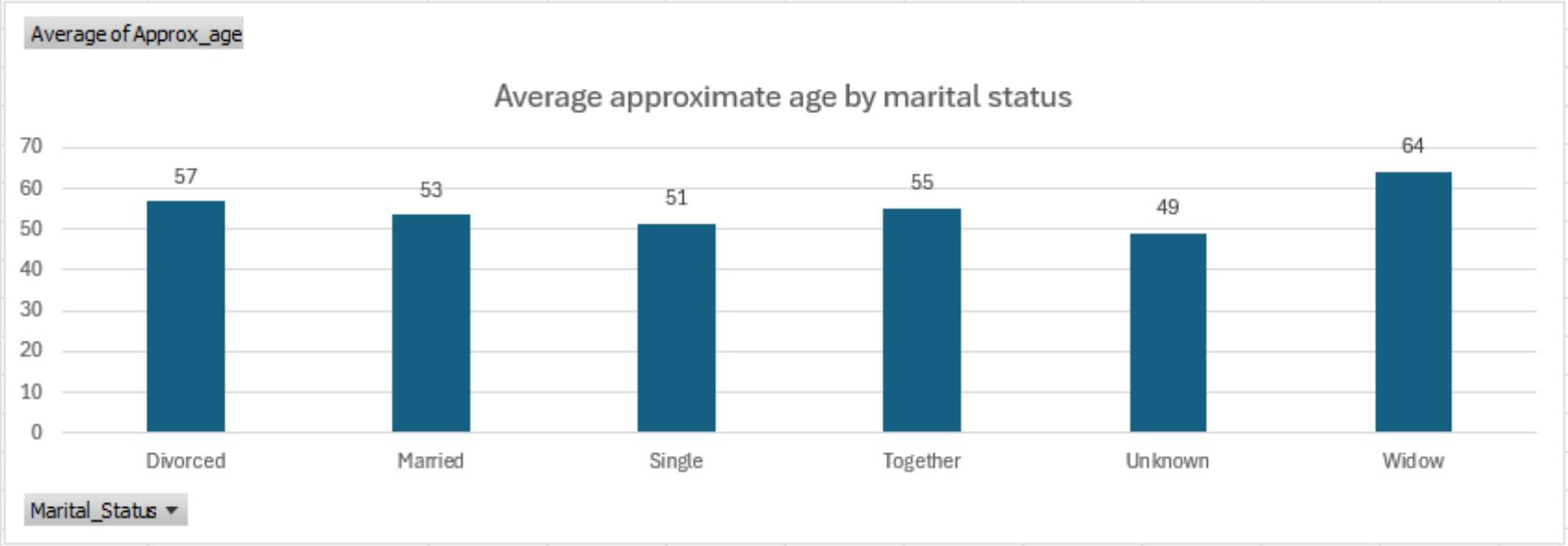
Data cleaning summary

No duplicates were found in the data and no records were discarded. Two outliers with high income or age were found but not removed as there was no clear reason to do so. Further details about data cleaning are in the Appendix.

Analysis of customer age

Birth year was used to derive an approximate age assuming a birth date of 1st July. The average age is 54 with a standard deviation of 11.98 and a 95% confidence interval of +- 0.5.

Age and marital status

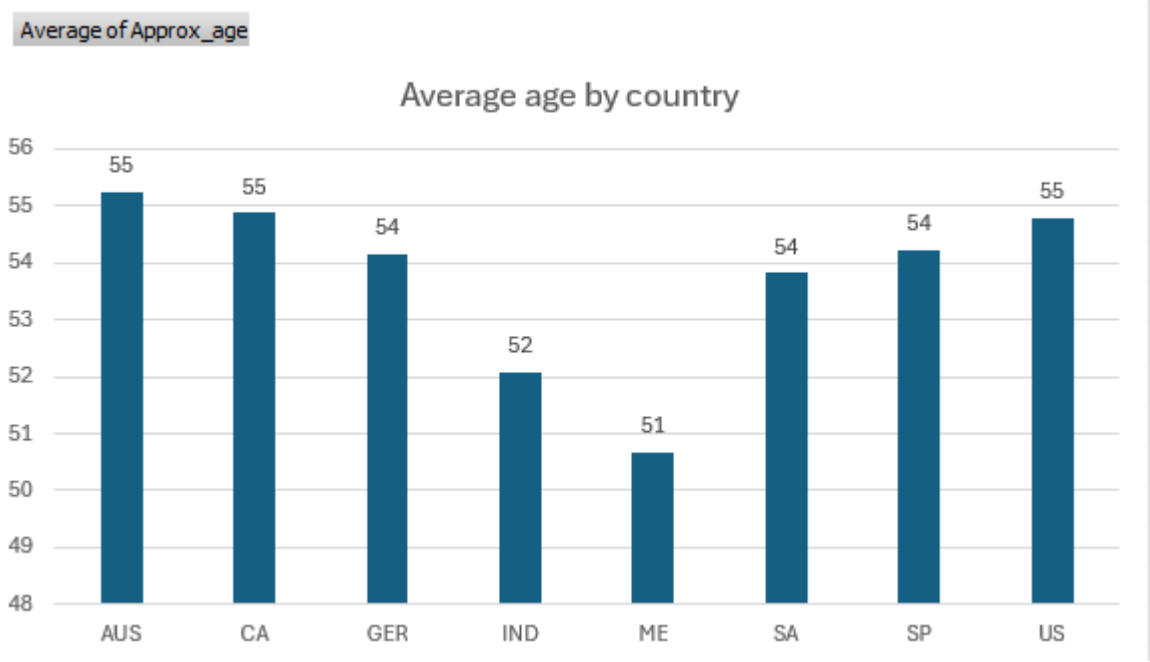


Single customers had the lowest average age of 51 and widowed customers had the highest at 64.

Jonathan Shields

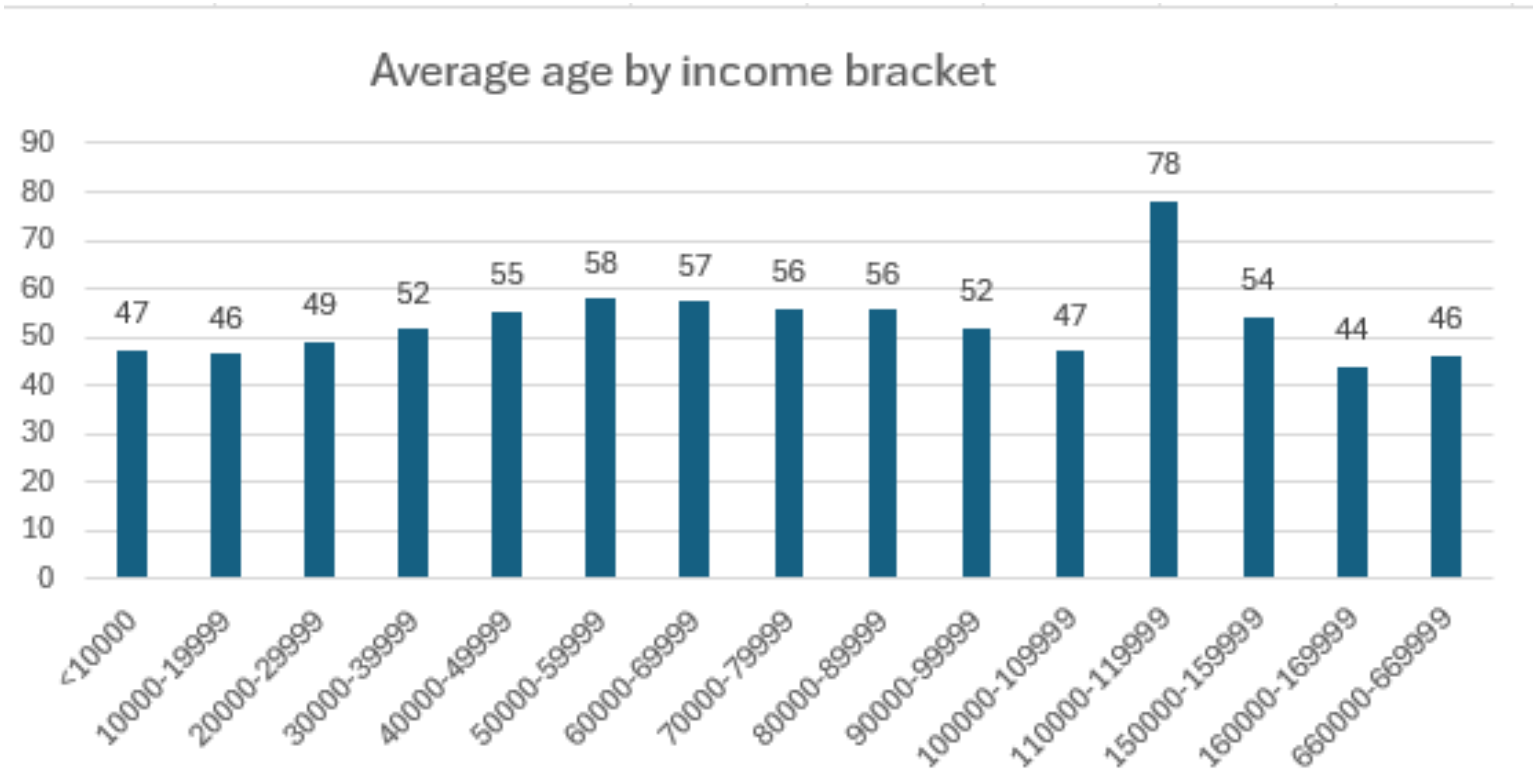
Age and country

The average age by country is similar although notably lower at 51 for Montenegro but this is based on only 3 customers.



Age and income

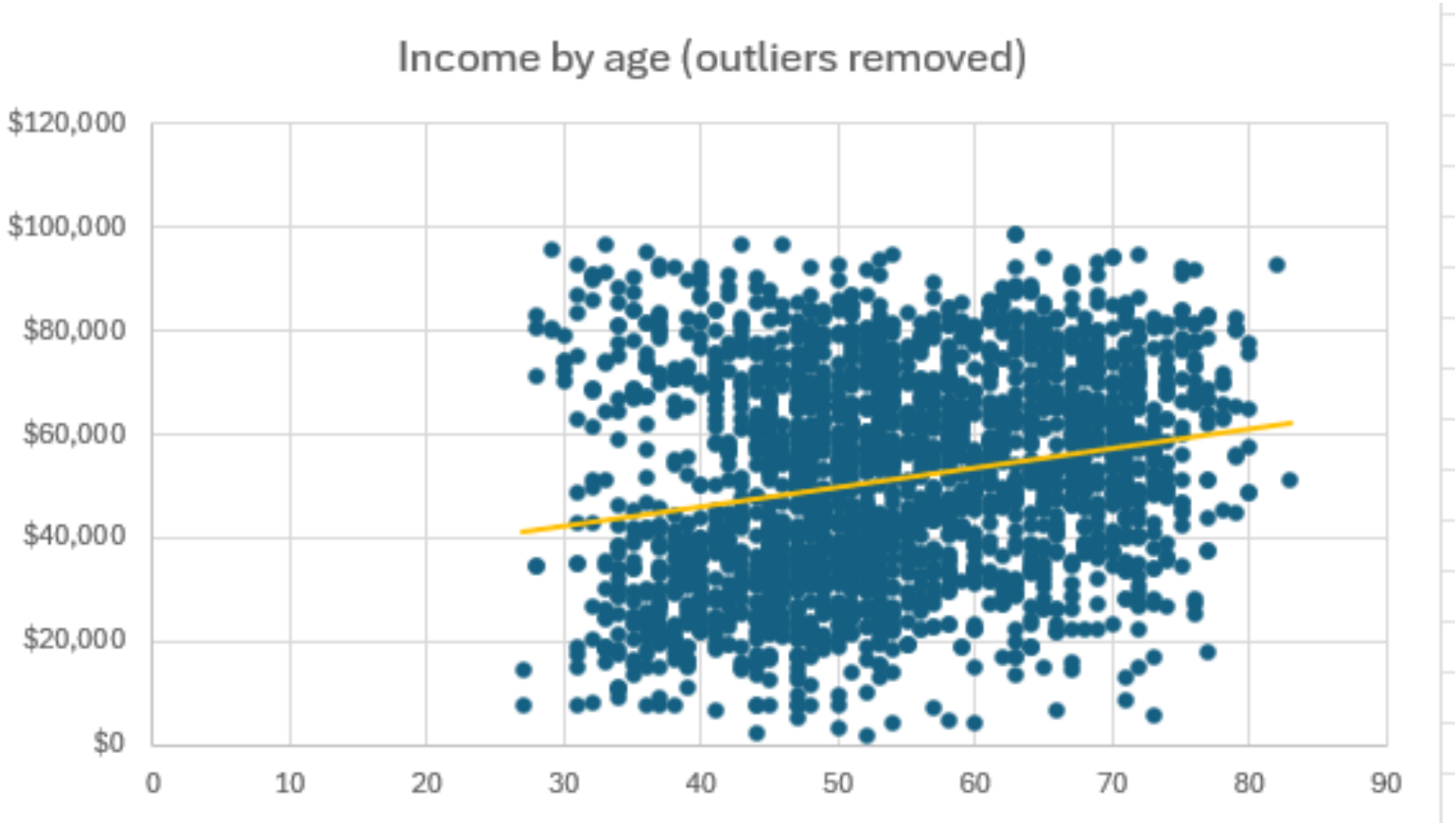
$10k income brackets were used to analyse the average age by income. Note that the 100k+ brackets have low customer numbers.



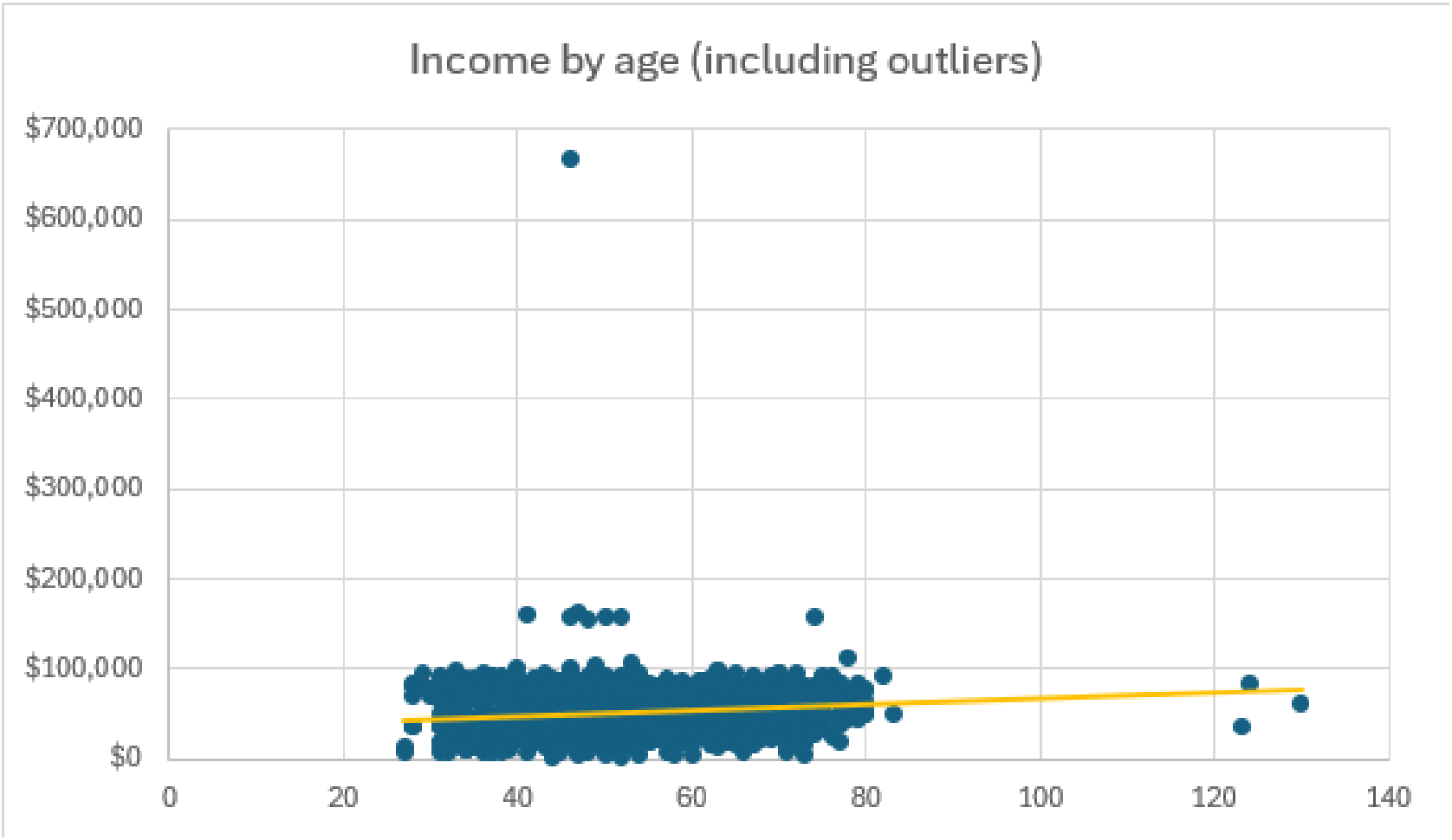
Ignoring brackets with a low population, $10k-20k has the lowest average age and $50-60k the highest. The average age increases with income until the $50 - 60k bracket then decreases.

Plotting income against age and removing outliers with income > $100k or age > 100 shows a gradual increase with a linear trendline. A similar trend is shown with the outliers included.

Jonathan Shields



Jonathan Shields



Jonathan Shields

$90-100k earners show a similar trend (note the y-axis here starts at 89k income).

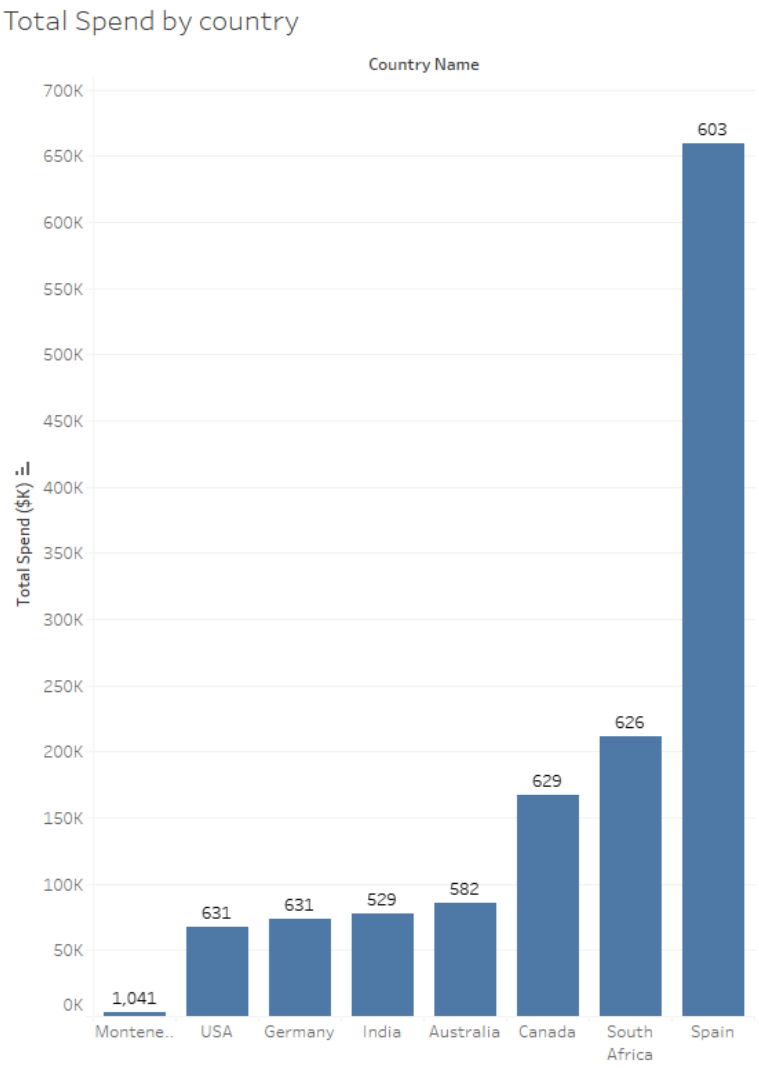
Further analysis using SQL

The data was imported into a Postgres database and queried to answer the following questions:

What is the total spend by country?

Spain (SP) has the highest spend.

Jonathan Shields



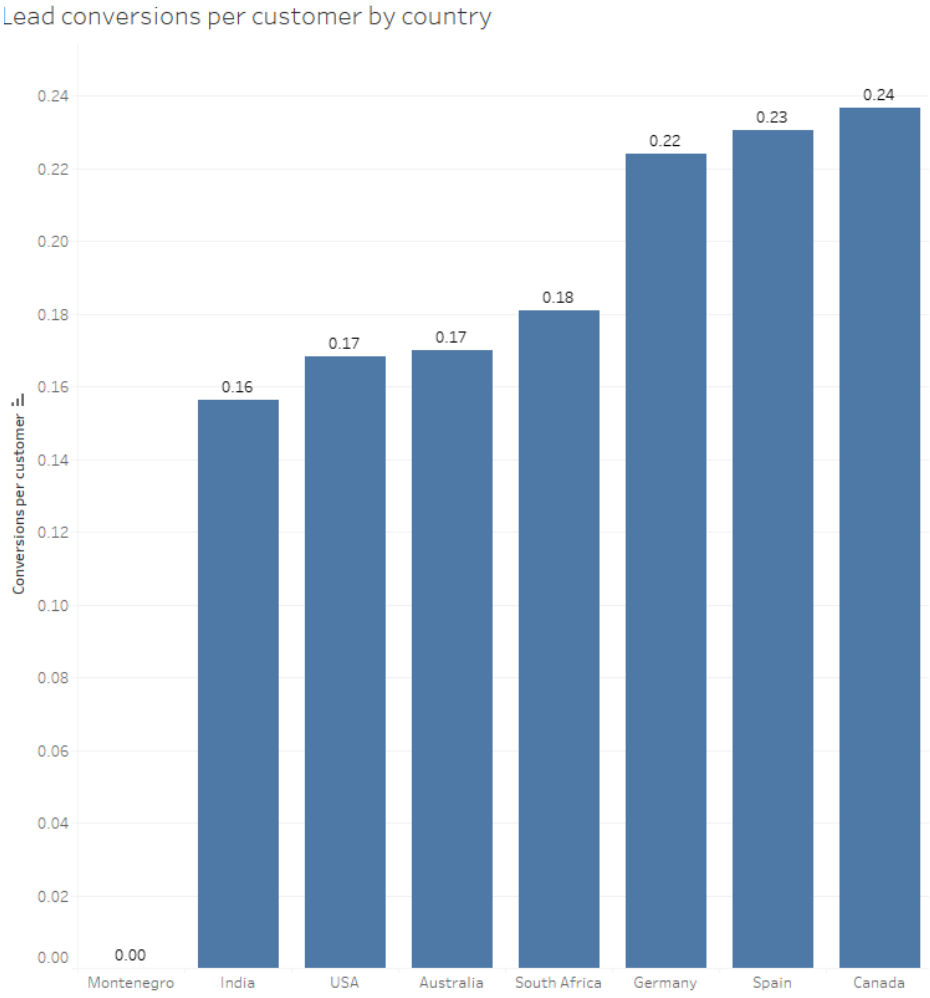
What are the top-selling products?

Liquor and non-vegetable products are the top-selling products whether sales are segmented by country, marital status, or number of children at home.

Jonathan Shields

Which country has the most social media conversions and conversions per customer?

Spain has the most conversions. However, it also has the most customers by a large margin so it is preferable to look at the conversions per customer. This shows Canada, Spain and Germany as the top performers.



Jonathan Shields

Which social media platform is most effective for each country?

The table shows the average spend by a conversion across the platforms. Instagram conversions spend more in all markets except the USA where Facebook conversions spend more. Facebook outperforms Twitter in most markets.

Lead conversions by spend

Country	Best platform	Avg spend for Facebook conversions($)	Avg spend for Instagram conversions($)	Avg spend for Twitter conversions($)	Social media conversions
Spain	Instagram	104.32	135.06	93.57	252
Canada	Instagram	103.95	116.37	94.97	63
South Africa	Instagram	92.61	106.05	79.37	61
Germany	Instagram	83.18	108.20	99.47	26
Australia	Instagram	71.29	125.88	53.84	25
India	Instagram	60.97	64.24	55.30	23
USA	Facebook	82.30	65.87	43.91	18
Montenegro		0.00	0.00	0.00	0

Is there a relationship between spending across categories and social media conversions?

The below table shows the percentage of total spending for each category and the percentage of the total conversions by platform. There is no clear correlation between conversions and spending in a particular category in markets with a significant number of conversions.

Country	Liquor % total	Veg % total	Nonveg % total	Fish % total	Choc % total	Comm % total	Twitter % total	Instagram % total	Facebook % total	Social media conversions
Spain	51.00%	4.29%	27.05%	6.09%	4.57%	7.00%	34.52%	35.32%	30.16%	252
Canada	50.22%	4.59%	27.43%	5.96%	4.54%	7.25%	38.10%	33.33%	28.57%	63
South Africa	50.18%	4.23%	27.67%	6.48%	4.27%	7.17%	32.79%	34.43%	32.79%	61
Germany	50.24%	4.07%	27.69%	6.29%	3.83%	7.88%	42.31%	30.77%	26.92%	26
Australia	49.96%	4.31%	26.09%	6.48%	4.82%	8.33%	24.00%	48.00%	28.00%	25
India	46.57%	4.87%	30.50%	6.19%	4.14%	7.73%	43.48%	26.09%	30.43%	23
USA	47.69%	4.49%	29.88%	6.53%	4.24%	7.16%	33.33%	27.78%	38.89%	18
Montenegro	55.38%	0.26%	26.17%	7.24%	3.91%	7.05%	0.00%	0.00%	0.00%	0

Jonathan Shields

Dashboard design and development

Tableau was chosen as it is a class-leading product that supports a wide range of visuals. Where suitable, simple and widely understood visualisations such as bar charts have been used. A palette suitable for colour blind users was used and alt text has been added to each visual to support screen readers.

The bubble chart on age distribution demonstrates at a glance the relative population of each age bracket and that almost all customers are between 30 and 79. The adjacent table uses colour to illustrate that almost 75% of spend happens within the 40-69 age brackets. The accompanying bar chart clearly shows how the average spend drops for 40-49 then slowly rises among the age groups with a significant population.

The average spend by category bar chart illustrates that alcohol and meat are the product categories with most sales and that alcohol sales rise with age across the age brackets with a significant population. The interactive filter allows age groups to be easily compared with one another as needed. A similar chart grouping alcohol and meat products separately from other products shows no obvious trend with age across other products.

The total sales by country map allows users to get a geographical sense of sales across markets. It is accompanied by average spend by country and customers by country bar charts so that average sales, total sales, and customer numbers can be compared across markets on one dashboard.

The treemap showing average spending by the number of children/teens at home is a visually intuitive way of showing the considerable increase in spending for customers with no children vs 1, 2, or 3 children. Pairing this with the bar chart of customers by number of children shows that customers with no children represent only around 28% of the total despite having the highest average spend.

The final dashboard uses a bar chart to show social media conversions per customer by country in a straightforward way. The table beneath clearly shows each platform's percentage of the total conversions with Instagram performing better in all markets except the USA, which had a low number of social media conversions.

Jonathan Shields

Conclusions

The following conclusions can be drawn from the data:

- Older customers tend to spend more
- Customers spend the most on liquor and more so as they get older
- Customers with no children spend a lot more money
- Spain has the highest sales by a large margin
- Customers who respond to Instagram ads spend more than other social media platforms in most markets

How can 2Market use this information to increase revenue?

2Market will have their own ideas on how to act on these insights but some suggestions could be:

- Introduce premium products that may appeal to customers aged 50+, particularly alcohol
- Make premium products for the Spanish market a priority
- Ensure that inventory of liquor and meat products is always sufficient
- Introduce more products likely to appeal to customers without children
- Target Instagram as the preferred social media advertising channel

Jonathan Shields

Appendix

Data cleaning during the initial analysis stage

No errors or blank rows were evident in marketing_data.csv or ad_data.csv

The dt_customer field had multiple date formats and used a 2 digit year. It was assumed that all 2 digit years were post-2000 and the field was converted into an mm/dd/yyyy format date field.

Numeric/currency fields were explicitly converted into the appropriate Excel type and checking for errors was performed.

No duplicated rows were found.

The marital_status field contained a number of ambiguous or equivalent values. To facilitate analysis: "Absurd" was replaced with Unknown, "Alone" was replaced with Single, and "YOLO" was replaced with Unknown

Further information on outliers

Customer id 11004 had a birth year of 1894. At the time of writing, the maximum documented age for a human is 122. However, this data point was obtained in 2014, making them approximately 120 at that time. For this reason, the data point was not removed.

Customer 9432 had an income of $666666.66. While this is around 5 times higher than the next highest income, it is not outside of credible bounds for a high salary so the data point was not removed.

Differences between age demographics for social media platforms

See https://www.investopedia.com/articles/markets/100215/twitter-vs-facebook-vs-instagram-who-target-audience.asp

SQL queries used during the analysis

--Total spend per country

```
SELECT country,SUM(amtLiq+amtvege+amtnonveg+amtpes+amtchocolates+amtcomm) As TotalSpend
FROM marketing_data
```

Jonathan Shields

```
GROUP BY country
ORDER BY TotalSpend DESC
--Total spend per product per country/Which products are most popular in each country

SELECT country,
SUM(amtliq) As "Liquor Spend ($)",
SUM(amtvege) As "Vegetable Spend ($)",
SUM(amtnonveg) As "Non-Vegetable Spend ($)",
SUM(amtpes) As "Fish Spend ($)",
SUM(amtchocolates) As "Chocolates Spend ($)",
SUM(amtcomm) As "Commodities Spend ($)",
SUM(amtLiq+amtvege+amtnonveg+amtpes+amtchocolates+amtcomm) As TotalSpend
FROM marketing_data
GROUP BY country
ORDER BY TotalSpend DESC

--Which products are most popular by marital status
SELECT marital_status,
SUM(amtliq) As LiquorSpend,
SUM(amtvege) As VegetablesSpend,
SUM(amtnonveg) As NonVegetablesSpend,
SUM(amtpes) As FishProductsSpend,
SUM(amtchocolates) As ChocolatesSpend,
SUM(amtcomm) As CommoditiesSpend,
```

```
SUM(amtLiq+amtvege+amtnonveg+amtpes+amtchocolates+amtcomm) As TotalSpend
FROM marketing_data
GROUP BY marital_status
ORDER BY TotalSpend DESC

--Which products are most popular based on the number of kids and teens in the home
SELECT kidhome + teenhome As NoOfKidsAndTeens,
SUM(amtliq) As LiquorSpend,
SUM(amtvege) As VegetablesSpend,
SUM(amtnonveg) As NonVegetablesSpend,
SUM(amtpes) As FishProductsSpend,
SUM(amtchocolates) As ChocolatesSpend,
SUM(amtcomm) As CommoditiesSpend,
SUM(amtLiq+amtvege+amtnonveg+amtpes+amtchocolates+amtcomm) As TotalSpend
FROM marketing_data
GROUP BY kidhome+ teenhome
ORDER BY NoOfKidsAndTeens

/* Which social media platform is most effective in each country in terms of the number of conversions */
SELECT m.country,
sum(CAST(a.twitter_ad AS INTEGER)) As twitter_conversions,
sum(CAST(a.instagram_ad AS INTEGER)) As instagram_conversions,
sum(CAST(a.facebook_ad AS INTEGER)) As facebook_conversions,
sum(CAST(a.twitter_ad AS INTEGER)) + sum(CAST(a.instagram_ad AS INTEGER)) + sum(CAST(a.facebook_ad AS INTEGER)) As all_social_media_conversions,
```

```
count(*) As Customers
FROM marketing_data m
INNER JOIN ad_data a
ON m.id=a.marketing_data_id
GROUP BY m.country
ORDER BY all_social_media_conversions DESC
/* Which social media platform is most effective for each marital status in terms of the number of conversions */
SELECT m.marital_status,
sum(CAST(a.twitter_ad AS INTEGER)) As twitter_conversions,
sum(CAST(a.instagram_ad AS INTEGER)) As instagram_conversions,
sum(CAST(a.facebook_ad AS INTEGER)) As facebook_conversions,
sum(CAST(a.twitter_ad AS INTEGER)) + sum(CAST(a.instagram_ad AS INTEGER)) + sum(CAST(a.facebook_ad AS INTEGER)) As all_social_media_conversions,
count(*) As Customers
FROM marketing_data m
INNER JOIN ad_data a
ON m.id=a.marketing_data_id
GROUP BY m.marital_status
ORDER BY all_social_media_conversions DESC
/*Spend by country and conversions by country */
WITH spend_by_country(country,liquor_spend,vegetable_spend,nonvegetable_spend,fish_spend,chocolate_spend,commodity_spend,total_spend,customers) AS
(SELECT m.country,
SUM(amtliq) As liquor_spend,
SUM(amtvege) As vegetable_spend,
SUM(amtnonveg) As nonvegetable_spend,
```

```
SUM(amtpes) As fish_spend,
SUM(amtchocolates) As chocolate_spend,
sum(amtcomm) As commodity_spend,
SUM(amtliq) + SUM(amtvege) + SUM(amtnonveg) + SUM(amtpes) + SUM(amtchocolates) + sum(amtcomm) As total_spend,
COUNT(*) As Customers
FROM marketing_data m
GROUP BY country),
conversions_by_country(country,twitter_conversions,instagram_conversions,facebook_conversions,all_social_media_conversions)
AS
(SELECT m.country,
sum(CAST(a.twitter_ad AS INTEGER)) As twitter_conversions,
sum(CAST(a.instagram_ad AS INTEGER)) As instagram_conversions,
sum(CAST(a.facebook_ad AS INTEGER)) As facebook_conversions,
sum(CAST(a.twitter_ad AS INTEGER)) + sum(CAST(a.instagram_ad AS INTEGER)) + sum(CAST(a.facebook_ad AS INTEGER)) As all_social_media_conversions
FROM marketing_data m
INNER JOIN ad_data a
ON m.id=a.marketing_data_id
GROUP BY m.country)
SELECT sc.country,
liquor_spend,
vegetable_spend,nonvegetable_spend,fish_spend,chocolate_spend,commodity_spend,
total_spend,
twitter_conversions,instagram_conversions,facebook_conversions,
all_social_media_conversions,
```

```
customers
FROM spend_by_country sc
LEFT JOIN conversions_by_country cc
USING (country)
ORDER BY customers DESC
/* Average spend for conversions */
SELECT country,
AVG(CASE WHEN a.twitter_ad=true THEN (amtliq+amtvege+amtnonveg+amtpes+amtchocolates+amtcomm)
ELSE 0
END) As Twitter_avg_spend,
AVG(CASE WHEN a.instagram_ad=true THEN (amtliq+amtvege+amtnonveg+amtpes+amtchocolates+amtcomm)
ELSE 0
END) As insta_avg_spend,
AVG(CASE WHEN a.facebook_ad=true THEN (amtliq+amtvege+amtnonveg+amtpes+amtchocolates+amtcomm)
ELSE 0
END) As facebook_avg_spend,
sum(CAST(a.twitter_ad AS INTEGER)) + sum(CAST(a.instagram_ad AS INTEGER)) + sum(CAST(a.facebook_ad AS INTEGER)) As all_social_media_conversions
FROM marketing_data m
INNER JOIN ad_data a
ON m.id=a.marketing_data_id
GROUP BY country
ORDER BY all_social_media_conversions DESC
```