Analysis for 2Market

Technical report

Prepared by Lilliana Golob

Contents

Background

2Market is a global supermarket retailer looking to increase sales. The marketing team want actionable insights by understanding their customers and identifying best-selling products and most effective channels. The challenge for them is knowing who to target and how.

Using the **Five Whys framework,** questions I'd like to answer:

1. What does a typical customer look like?
2. Who is the biggest spender?
3. What is the best seller?
4. Which channel converts the most?
5. Which demographic is most likely to convert?

Questions I'd like to ask:

1. Are the same products available online and in-store?
2. What and who have been targeted previously?
3. What were sales pre-advertising?

Analytical approach

I first reviewed the metadata file, then opened the CSV in Excel, created an 'original' copy, and checked that the data reflected the metadata information.

Exploratory analysis and cleaning in Excel

Field format

Changed fields with numbers (e.g. *ID*, *AmtLiq*) to numbers 0 decimals. Changed *Income* to currency but noticed leading $ still made it text so removed $. Changed *Dt_Customer* from General to Date.

Primary key

Used conditional formatting to check for duplicates and blanks in *ID*. None found.

Consistency and relevance

Education: Changed 2n Cycle[1] to Master and renamed Graduation to Bachelor because the terminology is more widely known.

Marital status: I added Alone (3 replacements), Absurd (2) and YOLO (2) to Single, consolidating the eight options to five so analysis is easier and outliers removed.

Country: Changed country codes to country names so visualisations are easier to read.

Completeness and validity

Checked for blanks and values out of range. There were no blanks and the data was within range (e.g. no negatives or too high).

Created *Age = Year - Year_Birth*. The last transaction was the end of 2014 so I assumed the year was 2015. Noticed three outliers where the age was 100. I did not delete them or impute new values; need to speak to marketing about further investigation.

Created *TotalSpend = AmtLiq + AmtVege + AmtNonVeg + AmtPes + AmtChocolates + AmtComm*

[1] 2n Cycle is a Master's level qualification following the Bologona Process.

Column headers

To minimise errors when importing and make analysis easier, I changed headers to a more descriptive name using lower_case format. Updated the metadata file.

metadata_2Market_updated — Edited

marketing_data.csv was updated using Excel during the initial data cleaning and exploratory analysis. The new file is now marketing_clean.csv. The ad_data.csv was updated using Excel during a quick data check and is now called ad_data_clean.csv.

This updated metadata document reflects the new columns plus the renaming of the column names. I simplified and made the names clearer, and also changed the naming convention formatting to lower_case. I did this to make analysis easier and to minimise potential errors when uploading and querying the data in SQL.

The two CSV files are: marketing_clean.csv (updated file)
ad_data_clean.csv (updated file)

****

marketing_clean.csv

There are 26 columns in this CSV file.

Column	Sample value	Interpretation of column
id	5642	Unique customer ID
birth_year	1980	Customer's year of birth
age	29	Customer's age (formula = 2015 - birth_year)
education	Master	Educational qualification of the customer
marital_status	Together	Customer's marital status: Married, Together, Single, Widow, Divorced
income	62499	Customer's annual income (in $ USD)
kid_home	1	Number of kids the customer has
teen_home	1	Number of teenagers the customer has
reg_date	25/09/2013	Date of customer's registration with the company (dd/mm/yyyy)
last_purch_days	99	Number of days since customer's last purchase
alcohol	140	Amount spent on alcoholic beverages
veg	4	Amount spent on vegetables
meat	61	Amount spent on meat items
fish	25	Amount spent on fish products
chocolate	30	Amount spent on chocolates
commodities	197	Amount spent on commodities
total_food	818	Total spent on food (veg + meat + fish)
total_spend	1198	Total spent across all categories
discount_deals	2	Number of deals purchased made with a discount
web_purch	3	Number of purchases made from the website
instore_purch	6	Number of in-store purchases
web_visits_pm	4	Number of website visits per month
campaign_resp	1	Boolean. If the customer had accepted the last campaign's offer (1) or not (0)
complaint	0	Boolean. If the customer had complained in the last 2 years (1) or not (0)
conv_success	1	Total number of successful lead conversions
country	Australia	Customer's location in country's full name.

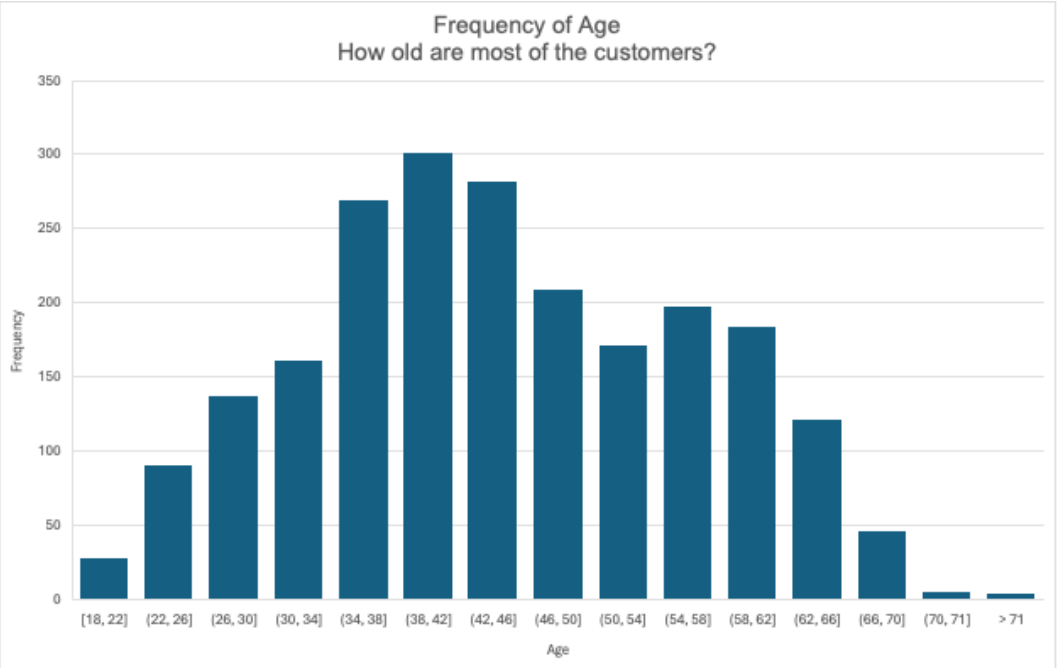
****

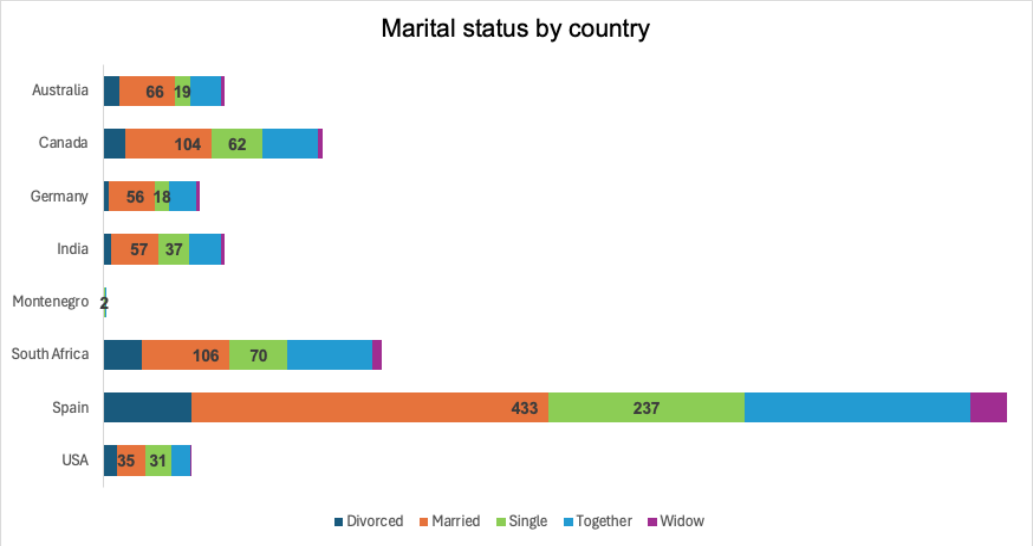
ad_data_clean.csv

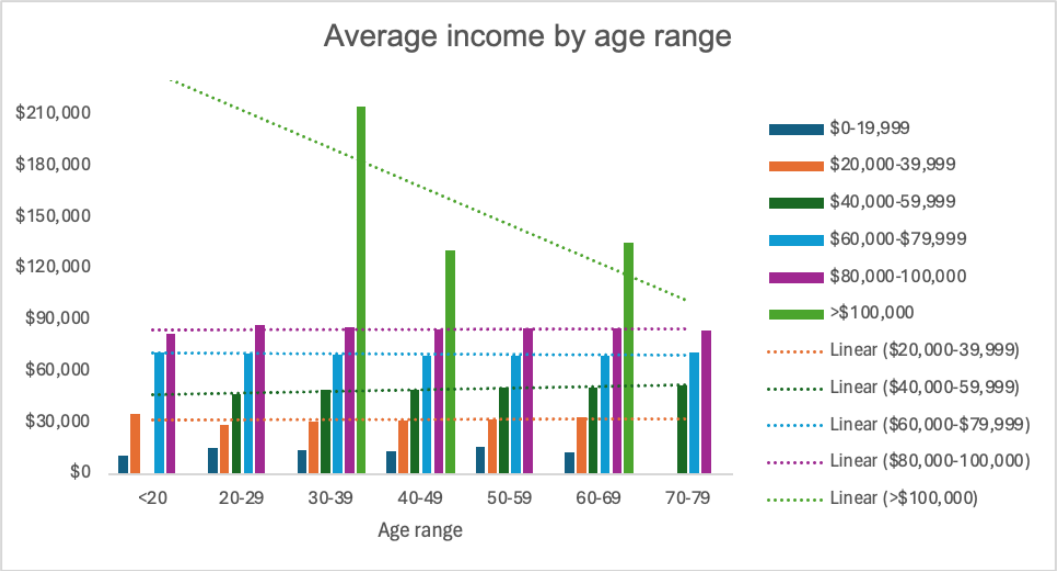
Column	Sample value	Interpretation of column
id	5642	Unique customer ID
bulkmail	1	Boolean. Successful lead conversions from bulk e-mails (1) or not (0) (potential customer received email and purchased a product)
twitter	1	Boolean. Successful lead conversions from Twitter (1) or not (0) (potential customer clicked on Twitter ad and purchased a product)
instagram	1	Boolean. Successful lead conversions from Instagram (1) or not (0) (potential customer clicked on Instagram ad and purchased a product)
facebook	1	Boolean. Successful conversions from Facebook (1) or not (0) (potential customer clicked on the Facebook ad and purchased a product)
brochure	1	Boolean. Successful conversions from brochures (1) or not (0) (potential customer received the company brochure and purchased a product)

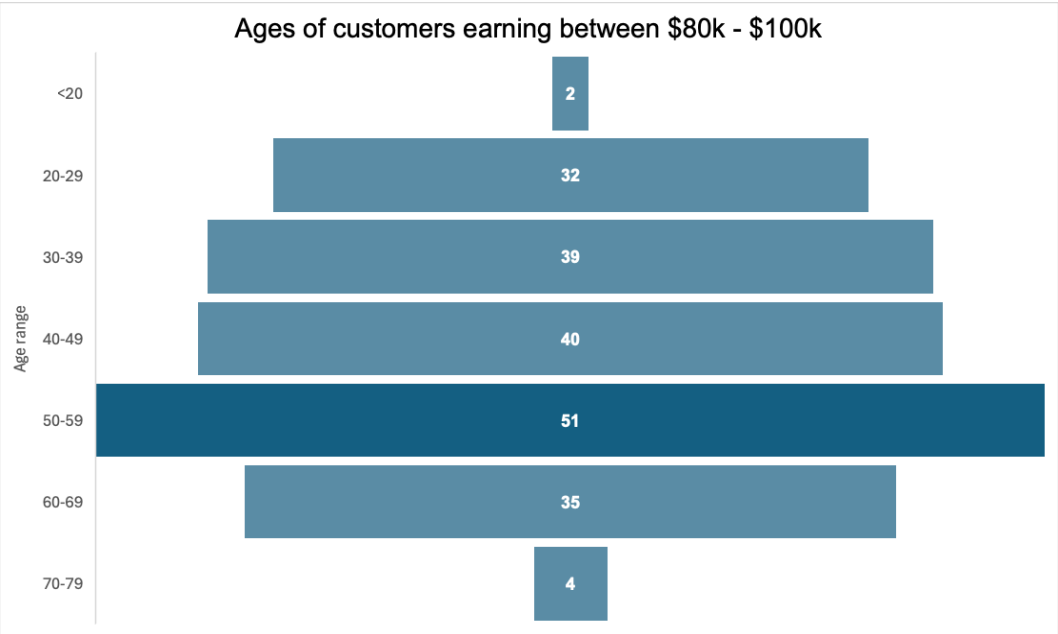
Pivot tables and charts

Explored the data using pivot tables and charts including:


Frequency of Age
How old are most of the customers?
Frequency
350
300
250
200
150
100
50
0
[18, 22]
(22, 26]
(26, 30]
(30, 34]
(34, 38]
(38, 42]
(42, 46]
(46, 50]
(50, 54]
(54, 58]
(58, 62]
(62, 66]
(66, 70]
(70, 71]
> 71
Age


Marital status by country
Australia
66
19
Canada
104
62
Germany
56
18
India
57
37
Montenegro
2
South Africa
106
70
Spain
433
237
USA
35
31
Divorced
Married
Single
Together
Widow





Exploratory analysis in SQL

Field format

Checked the date format used by SQL and updated the CSV file to yyyy-mm-dd.

Tables

Used CREATE TABLE and then imported the respective CSV files.

postgres/postgres@PostgreSQL 16

No limit

Query Query History

```sql
CREATE TABLE marketing
(
    id NUMERIC(6) PRIMARY KEY,
    birth_year NUMERIC(4),
    age NUMERIC(3),
    education VARCHAR(20),
    marital_status VARCHAR(20),
    income NUMERIC(10),
    kid_home NUMERIC(2),
    teen_home NUMERIC(2),
    reg_date DATE,
    last_purch_days NUMERIC(5),
    alcohol NUMERIC(10),
    veg NUMERIC(10),
    meat NUMERIC(10),
    fish NUMERIC(10),
    chocolate NUMERIC (10),
    commodities NUMERIC (10),
    total_food NUMERIC (10),
    total_spend NUMERIC (10),
    discount_deals NUMERIC (5),
    web_purch NUMERIC (5),
    instore_purch NUMERIC (5),
    web_visits_pm NUMERIC (3),
    campaign_resp NUMERIC (1),
    complaint NUMERIC (1),
    conv_success NUMERIC (1),
    country VARCHAR (50)
);
```

Later changed *complaint* and *campaign_resp* to BOOLEAN.

Query Query History

```sql
CREATE TABLE ad_data
(
    id NUMERIC(6) PRIMARY KEY,
    bulkmail BOOLEAN,
    twitter BOOLEAN,
    instagram BOOLEAN,
    facebook BOOLEAN,
    brochure BOOLEAN
);
```

Validation tests

Since the data was cleaned in Excel I worked from the two tables I created. I ran validation tests to check the data before any analysis.

```sql
/* Data validation: check for null values */

SELECT
    id,
    kid_home,
    teen_home,
    alcohol,
    veg,
    meat,
    fish,
    chocolate,
    commodities,
    total_food,
    total_spend,
    country
FROM public.marketing
WHERE
    id IS NULL OR
    kid_home IS NULL OR
    teen_home IS NULL OR
    alcohol IS NULL OR
    veg IS NULL OR
    meat IS NULL OR
    fish IS NULL OR
    chocolate IS NULL OR
    commodities IS NULL OR
    total_food IS NULL OR
    total_spend IS NULL OR
    country IS NULL;
```

Data Output | Messages | Notifications

id [PK] numeric (6)	kid_home numeric (2)	teen_home numeric (2)	alcohol numeric (10)	veg numeric (10)	meat numeric (10)	fish numeric (10)	chocolate numeric (10)	commodities numeric (10)	total_food numeric (10)	total_spend numeric (10)	country character v

```sql
/* Data validation: duplicate primary id */

SELECT id, COUNT(id)
FROM public.marketing
GROUP BY id
HAVING COUNT(id) > 1;
```

Data Output | Messages | Notifications

id [PK] numeric (6)	count bigint

```sql
/* Data validation: duplicate primary id in 2nd table */

SELECT id, COUNT(id)
FROM public.ad_data
GROUP BY id
HAVING COUNT(id) > 1;
```

Data Output | Messages | Notifications

id [PK] numeric (6)	count bigint

```sql
/* Data validation: invalid quantities */

SELECT
    MIN(alcohol) AS alcohol_min,
    MIN(veg) AS veg_min,
    MIN(meat) AS meat_min,
    MIN(fish) AS fish_min,
    MIN(chocolate) AS chocolate_min,
    MIN(commodities) AS commodities_min,
    MIN(total_food) AS min_food,
    MIN(total_spend) AS min_spend
FROM public.marketing;
```

Data Output Messages Notifications

	alcohol_min numeric	veg_min numeric	meat_min numeric	fish_min numeric	chocolate_min numeric	commodities_min numeric	min_food numeric	min_spend numeric
1	0	0	0	0	0	0	1	5

Query Query History

```sql
/* Data validation: duplicate records based on columns */

SELECT
    birth_year,
    education,
    marital_status,
    income,
    kid_home,
    teen_home,
    reg_date,
    last_purch_days,
    total_spend,
    country,
    COUNT(*) AS duplicate_count
FROM public.marketing
GROUP BY
    birth_year,
    education,
    marital_status,
    income,
    kid_home,
    teen_home,
    reg_date,
    last_purch_days,
    total_spend,
    country
HAVING COUNT(*) > 1;
```

Data Output Messages Notifications

	birth_year numeric (4)	education character varying (20)	marital_status character varying (20)	income numeric (10)	kid_home numeric (2)	teen_home numeric (2)	reg_date date	last_purch_days numeric (5)	total_spend numeric (10)	country character varying (50)	duplicate_count bigint
1	1956	Master	Married	62972	0	1	2012-08-03	39	587	Spain	2
2	1961	Bachelor	Together	63381	0	1	2012-10-05	78	1005	Spain	2
3	1985	Bachelor	Together	35196	1	0	2012-11-13	68	497	Spain	2
4	1986	Bachelor	Single	70596	0	0	2012-10-05	68	968	Spain	2
5	1957	Master	Together	50943	0	1	2013-06-21	49	46	Spain	2
6	1972	Bachelor	Together	34600	1	1	2013-01-01	8	318	Spain	2
7	1990	Bachelor	Single	30279	1	0	2012-12-30	13	37	Spain	2

Total rows: 54 of 54 Query complete 00:00:00.055

There are potentially 54 duplicate records. I've left these in because they all have unique IDs. I would like more info (e.g. name and/or email) to determine if they are duplicates.

I saved this query as a view.

Query Query History

```sql
/* Data validation: duplicate records based on columns */

CREATE VIEW marketing_duplicates AS

SELECT
    birth_year,
    education,
    marital_status,
    income,
    kid_home,
    teen_home,
    reg_date,
    last_purch_days,
    total_spend,
    country,
    COUNT(*) AS duplicate_count
FROM public.marketing
GROUP BY
    birth_year,
    education,
    marital_status,
    income,
    kid_home,
    teen_home,
    reg_date,
    last_purch_days,
    total_spend,
    country
HAVING COUNT(*) > 1;
```

Data Output Messages Notifications

```
CREATE VIEW

Query returned successfully in 79 msec.
```

Exploration

Query Query History

```sql
/* Total spend per product per country */

SELECT
    country,
    SUM(alcohol) AS total_alcohol,
    SUM(veg) AS total_veg,
    SUM(meat) AS total_meat,
    SUM(fish) AS total_fish,
    SUM(chocolate) AS total_choc,
    SUM(commodities) AS total_comm,
    SUM(total_spend) AS total_spend
FROM public.marketing
GROUP BY country
ORDER BY country;
```

Data Output Messages Notifications

	country character varying (50)	total_alcohol numeric	total_veg numeric	total_meat numeric	total_fish numeric	total_choc numeric	total_comm numeric	total_spend numeric
1	Australia	42752	3689	22328	5546	4129	7132	85576
2	Canada	84066	7681	45925	9980	7607	12144	167403
3	Germany	36776	2980	20272	4601	2801	5768	73198
4	India	36236	3788	23729	4818	3221	6014	77806
5	Montenegro	1729	8	817	226	122	220	3122
6	South Africa	105918	8937	58398	13670	9019	15129	211071
7	Spain	336392	28288	178409	40153	30134	46181	659557
8	USA	32214	3034	20185	4411	2863	4839	67546

```sql
/* Total spend per product per country sorted by total spend */

SELECT
    country,
    SUM(alcohol) AS total_alcohol,
    SUM(veg) AS total_veg,
    SUM(meat) AS total_meat,
    SUM(fish) AS total_fish,
    SUM(chocolate) AS total_choc,
    SUM(commodities) AS total_comm,
    SUM(total_spend) AS total_spend
FROM public.marketing
GROUP BY country
ORDER BY total_spend DESC;
```

Data Output Messages Notifications

	country character varying (50)	total_alcohol numeric	total_veg numeric	total_meat numeric	total_fish numeric	total_choc numeric	total_comm numeric	total_spend numeric
1	Spain	336392	28288	178409	40153	30134	46181	659557
2	South Africa	105918	8937	58398	13670	9019	15129	211071
3	Canada	84066	7681	45925	9980	7607	12144	167403
4	Australia	42752	3689	22328	5546	4129	7132	85576
5	India	36236	3788	23729	4818	3221	6014	77806
6	Germany	36776	2980	20272	4601	2801	5768	73198
7	USA	32214	3034	20185	4411	2863	4839	67546
8	Montenegro	1729	8	817	226	122	220	3122

Sorting by *total_spend* it's easy to see the best seller is alcohol.

```sql
/* Most popular products based on marital status */

SELECT
    marital_status,
    SUM(alcohol) AS total_alcohol,
    SUM(veg) AS total_veg,
    SUM(meat) AS total_meat,
    SUM(fish) AS total_fish,
    SUM(chocolate) AS total_choc,
    SUM(commodities) AS total_comm,
    SUM(total_spend) AS total_spend
FROM public.marketing
GROUP BY marital_status
ORDER BY total_spend DESC;
```

Data Output Messages Notifications

	marital_status character varying (20)	total_alcohol numeric	total_veg numeric	total_meat numeric	total_fish numeric	total_choc numeric	total_comm numeric	total_spend numeric
1	Married	256976	21981	137888	30395	22926	36719	506885
2	Together	176715	14612	95374	22383	15031	24754	348869
3	Single	139126	13027	87868	18704	12839	20970	292534
4	Divorced	75364	6363	34848	8130	6222	10739	141666
5	Widow	27902	2422	14085	3793	2878	4245	55325

```sql
/* Most popular products based on kids/teens in the household */

SELECT
    kid_home AS number_kids,
    teen_home AS number_teens,
    SUM(alcohol) AS total_alcohol,
    SUM(veg) AS total_veg,
    SUM(meat) AS total_meat,
    SUM(fish) AS total_fish,
    SUM(chocolate) AS total_choc,
    SUM(commodities) AS total_comm,
    SUM(total_spend) AS total_spend
FROM public.marketing
GROUP BY kid_home, teen_home
ORDER BY total_spend DESC;
```

Data Output Messages Notifications

	number_kids numeric (2)	number_teens numeric (2)	total_alcohol numeric	total_veg numeric	total_meat numeric	total_fish numeric	total_choc numeric	total_comm numeric	total_spend numeric
1	0	0	308950	33090	234758	48500	33663	40661	699622
2	0	1	258984	16840	86357	22684	17841	34666	437372
3	1	0	40949	4907	24463	7259	4682	10608	92868
4	1	1	45805	2399	16785	3455	2745	8399	79588
5	0	2	12287	621	4004	1011	574	1711	20208
6	1	2	5796	270	2312	180	212	606	9376
7	2	1	2271	29	669	88	42	300	3399
8	2	0	1041	249	715	228	137	476	2846

Alcohol is still best-selling regardless of the number of kids/teens.

Query Query History

```sql
/* Average spend based on kids/teens at home */

SELECT
    kid_home + teen_home AS minors_at_home,
    ROUND(AVG(alcohol),2) AS avg_alcohol,
    ROUND(AVG(veg),2) AS avg_veg,
    ROUND(AVG(meat),2) AS avg_meat,
    ROUND(AVG(fish),2) AS avg_fish,
    ROUND(AVG(chocolate),2) AS avg_choc,
    ROUND(AVG(commodities),2) AS avg_comm,
    ROUND(AVG(total_spend),2) AS avg_spend
FROM public.marketing
GROUP BY minors_at_home
ORDER BY AVG(total_spend) DESC;
```

Data Output Messages Notifications

	minors_at_home numeric	avg_alcohol numeric	avg_veg numeric	avg_meat numeric	avg_fish numeric	avg_choc numeric	avg_comm numeric	avg_spend numeric
1	0	488.07	52.27	370.87	76.62	53.18	64.24	1105.25
2	1	268.52	19.47	99.21	26.81	20.16	40.53	474.70
3	3	161.34	5.98	59.62	5.36	5.08	18.12	255.50
4	2	142.15	7.86	51.69	11.28	8.31	25.45	246.74

Dashboard

The marketing team is responsible for delivering targeted campaigns that maximise conversions and revenue generation. Using the 4Ps as inspiration, I created **four dashboards** that answer the following questions to help inform future marketing activities.

1. Customer
    a. Where do customers live?
    b. What is their marital status?
    c. How much do they earn?
2. Product
    a. What products are they buying?
    b. How much are they spending on average?
3. Place
    a. Instore purchase frequency
    b. Web purchase frequency
4. Promotion
    a. How many conversions did we get?
    b. Which channels are most effective?

Visualisations

I built the dashboards for **Desktop (1000 x 800)** since most marketers I know tend to use a laptop and/or multiple screens.

For customers by *country*, I chose a **treemap** to show the relative size of each group. I used **packed bubbles** for *marital status* because I wanted to show the relative quantity using a different visualisation.

For financial values (e.g. *income*, *spend*), I used **vertical and horizontal bar charts** because they're easy to read and allow quick comparison. I included **reference lines** to show the average value (e.g. *average age*, *income*, *conversion*).

I chose **histograms** for *in-store* and *web* purchases to show the frequency of purchases.

Interactivity

During the revisit, I included four multi-select **filters** (*country, marital status, income and age range*) on Product, Place and Promotion to add consistency to the analysis.

You can filter by a combination of factors, the filters apply to all the charts on the dashboard, and there are **dynamic titles** – making it easier for the stakeholder.

Tooltips show specific values when you hover over a bubble or bar.

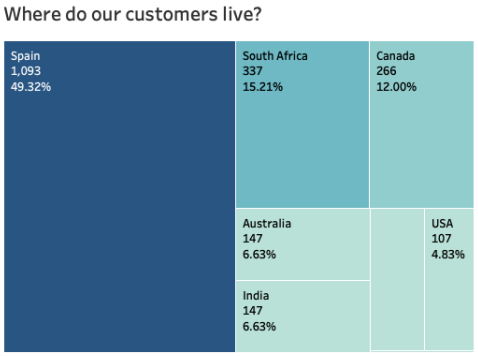
Colours

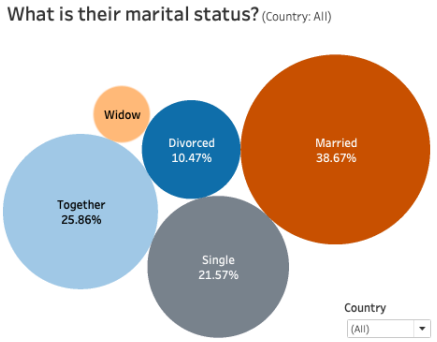
I used the **colour blindness palette** because I wanted to make the dashboards accessible as well as pretty.

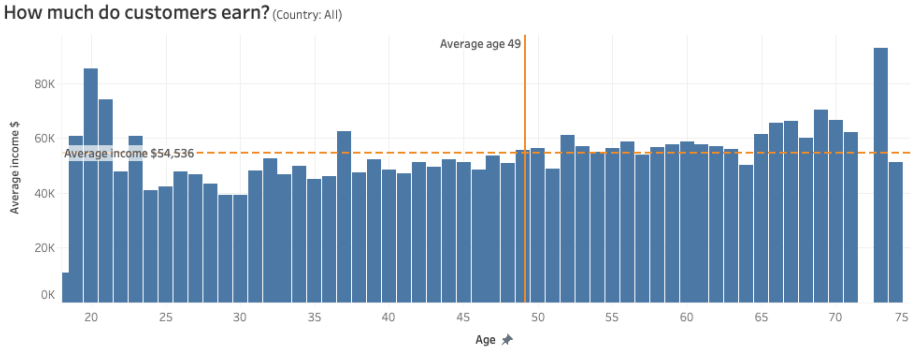
Dashboard screenshots

See next page

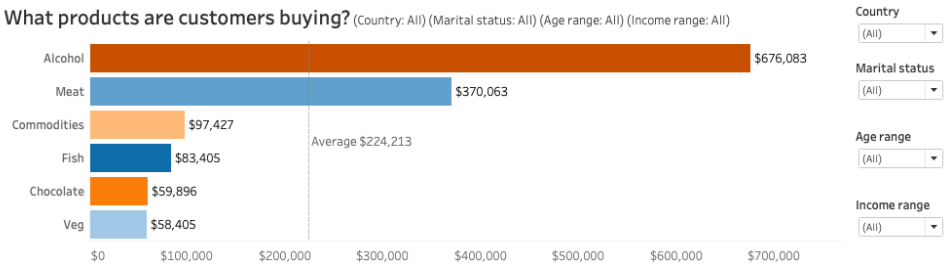
Customer dashboard

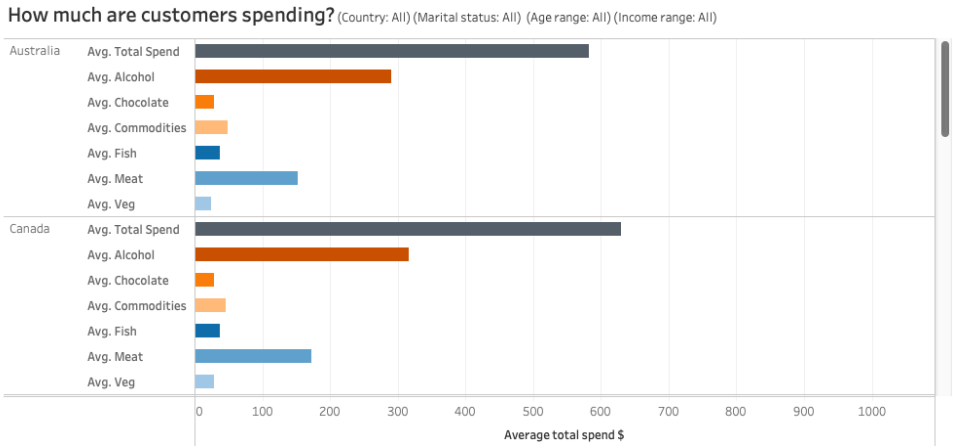




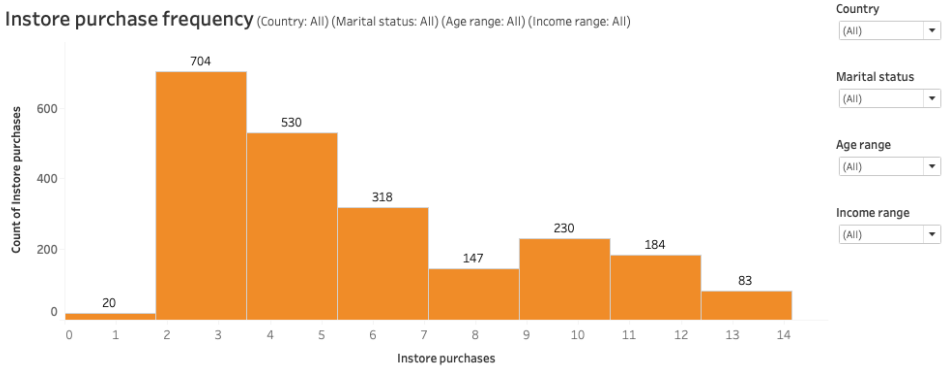


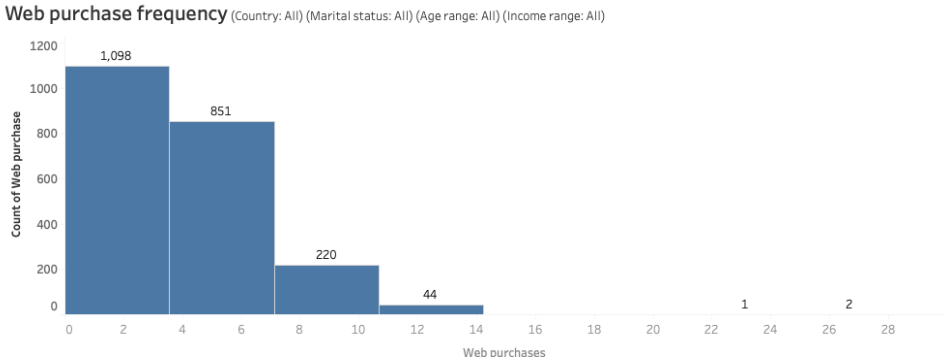
Product dashboard



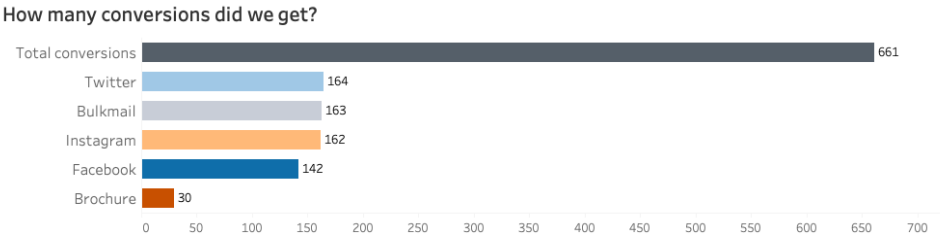


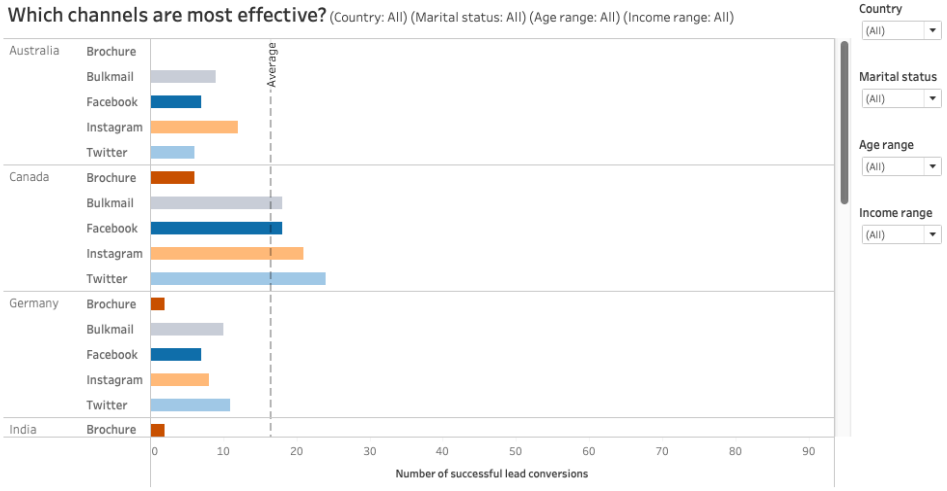
Place dashboard





Promotion dashboard





Insights

Country

Nearly half of the customers are Spanish, the other 50% is split across seven countries. Customers from Montenegro account for 0.14% but they spend the most on average.

Product

Alcohol is the best seller across countries and marital status. Spending on meat is above average, however, on the remaining products it's below average.

Channel

Twitter and Instagram are most effective overall, though email outperforms all channels in India and USA.

Household size

Customers with no kids/teens spend the most on average, while those with two kids/teens spend the least.

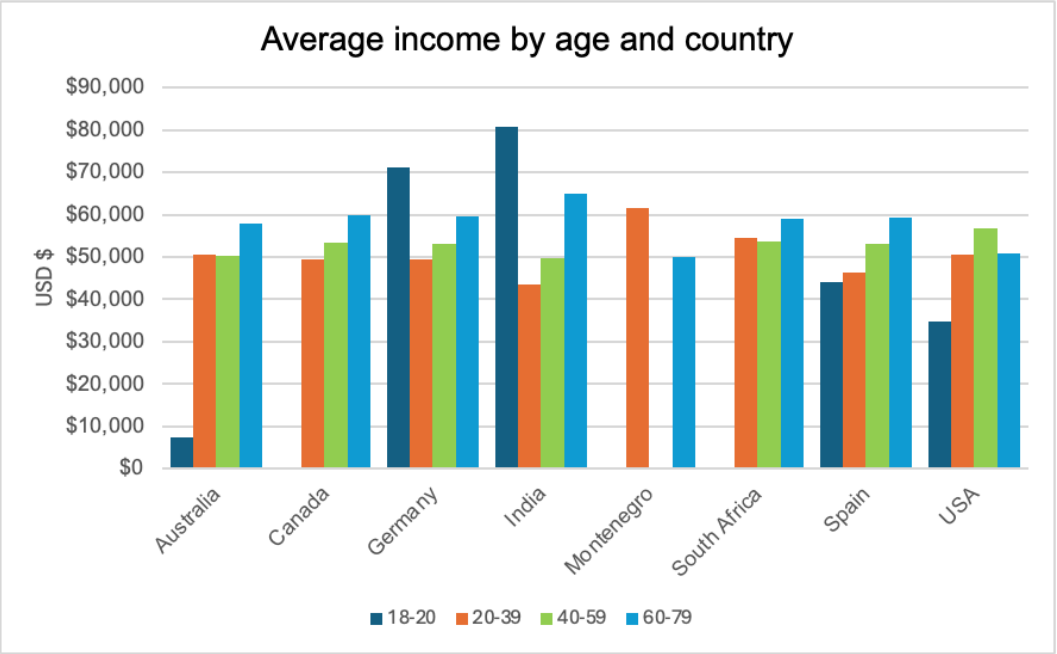
Marital status

64.5% of customers are married or together, and aged 30-50 years old.

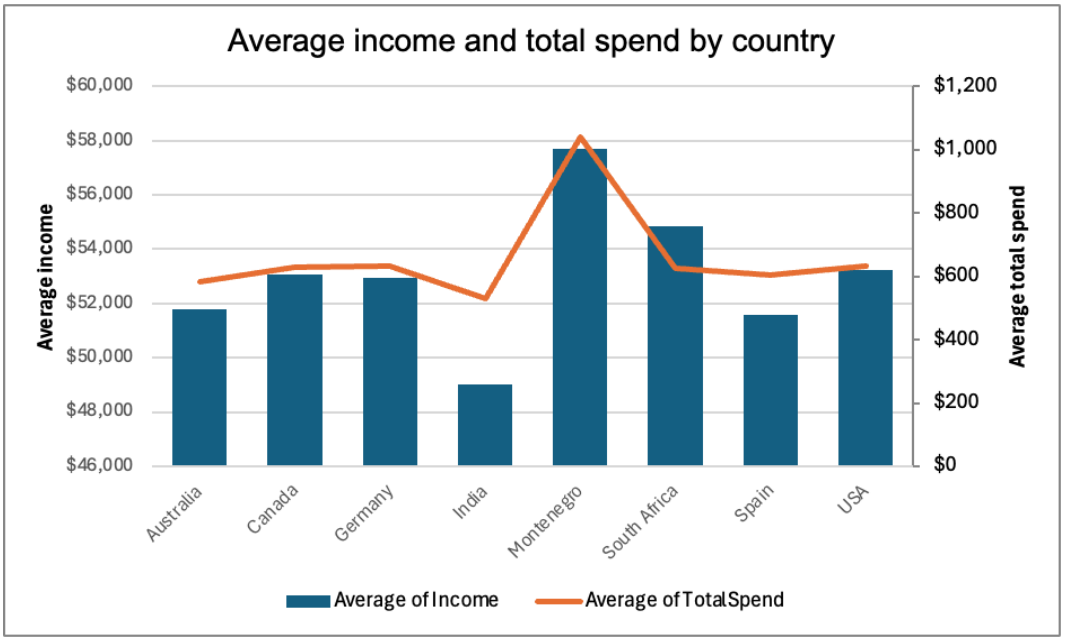
Age	18-20	20-29	30-39	40-49	50-59	60-69	70-79	% of total
Married	0.1%	3.6%	10.3%	12.5%	7.2%	4.7%	0.3%	38.7%
Together	0.0%	1.7%	7.0%	7.3%	5.9%	4.0%	0.0%	25.8%
Single	0.2%	3.9%	5.4%	5.8%	4.0%	2.1%	0.1%	21.6%
Divorced	0.0%	0.4%	2.1%	3.8%	2.6%	1.5%	0.1%	10.4%
Widow	0.0%	0.0%	0.2%	0.9%	1.0%	1.3%	0.1%	3.4%
% of total	0.3%	9.6%	25.0%	30.2%	20.7%	13.5%	0.7%	100.0%

Income

German and Indian customers are the youngest (18-20) above-average earners.



Generally, average total spend increases as average income does (except South Africa).



Recommendations

My **five recommendations** for the marketing team:

1. Discontinue brochure.
2. Continue email activities.
3. Focus on Twitter and Instagram.
4. Promote below-average selling products.
5. Diversify customer base geographically.

Further analysis

This report summarises my approach, key insights and recommendations for the marketing team. I'd like to do more analysis to further help the marketing team.

Further analysis:

- **Age outliers –** investigate where age is >100.
- **Potential duplicates** – investigate using other info (e.g. name or email).
- **Margins –** evaluate profitability by product.
- **Channel costs –** analyse cost per acquisition.

Get in touch

lilliana.golob@gmail.com | 07501 450668