

2Market Assignment

Exploratory Analysis and Presenting Insights

Damian Ferguson

16th December 2023

Background

Problem statement

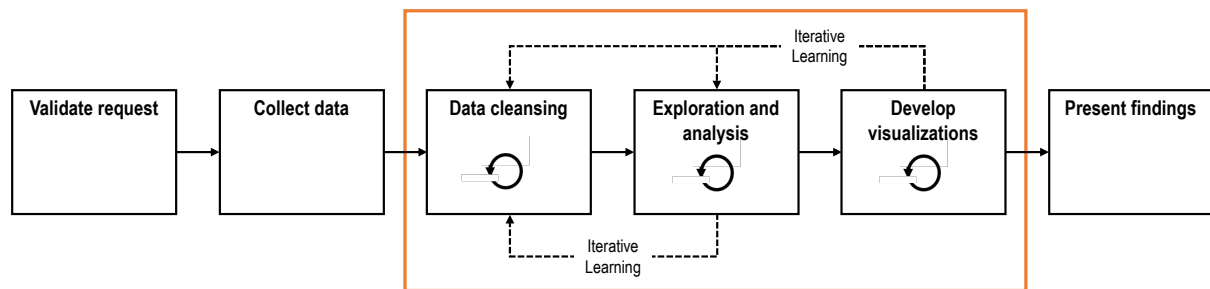
2Market is a global supermarket that sells products in store and online in eight countries. They want to grow sales revenues through targeted marketing campaigns. Customer demographics and previous campaign data will be used to identify the best opportunities for future campaigns. The data will be visualised to aid decision making.

The problem statement was developed using 5 Why's methodology, based on current knowledge and understanding.

Further questions for 2Market

- 1) Who are the stakeholders and what are their levels of engagement?
- 2) Is the problem statement correct? Any changes required?
- 3) What is the growth target?
- 4) What is the source of the data? Is it complete? Is it recent?

Overall Approach



The overall approach used is shown above, with the analytical and visualisation steps highlighted in the orange rectangle. Each step was iterative. Learning from later steps was incorporated in new iterations of earlier steps.

Analytical Approach

Data cleansing in Excel

The marketing and ad data were imported into Excel (2,216 records). Care was taken with Customer_Dt column as the CSV file had a different date format to Excel. Import parameters were set to ensure the data imported as expected.

After the data was imported it was reviewed for accuracy, completeness, consistency, uniqueness, and timeliness of the data. Cleansing was non-destructive. Each iteration was retained in case of error. The following actions were taken:

- Check for duplicate records. None found.
- Removed leading "\$" sign values in the Income column and converted to numeric to allow arithmetic operations (see Appendix A).
- Fifteen records were deleted due to outliers or unusual values being present for Age, Income and Marital Status (see Appendix B).
- Data values in the Marital Status, Education and Country columns were amended to improve understanding (see Appendix C).

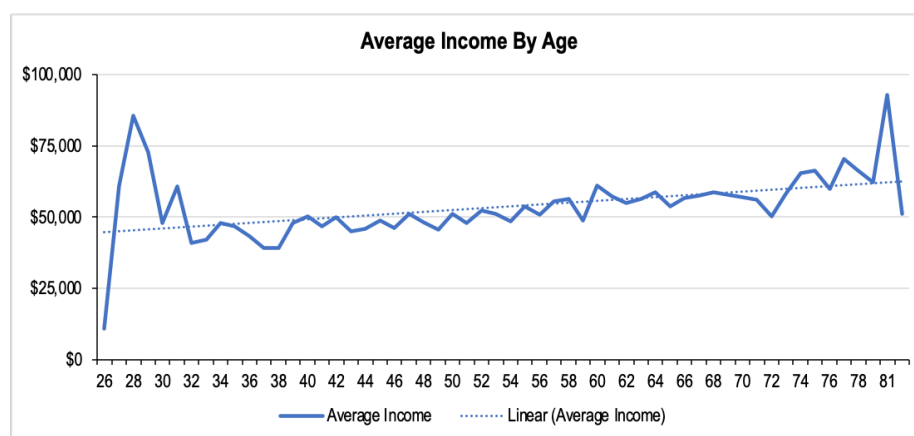
- Three new columns were added to aid analysis: Age, Age Band and Income Band.
- Column headings were amended to improve understanding and make analysis more efficient.

Exploration and analysis in Excel

The data was explored using charts and pivot tables:

	Count of ID	Average of Age
Cohabiting	568	54.18
Divorced	230	55.47
Married	854	52.43
Single	473	50.25
Widow	76	63.24
(blank)		
Grand Total	2,201	53.10

Income Band	Single	Cohabiting	Married	Divorced	Widow
More than \$100,000	2	2		1	
\$80,001 to \$100,000	57	45	75	20	6
\$60,001 to \$80,000	115	164	245	71	26
\$40,001 to \$60,000	124	168	245	72	32
\$20,001 to \$40,000	144	162	232	54	12
Up to \$20,000	31	27	57	12	



Initial insights:

- Average age 53
- Over half of customers earn \$40K to \$80K
- Over half of customers are married or cohabiting
- Income increases with age.

Five additional records were identified deleted to simplify analysis, Country = Montenegro and Age Group = 80's.

The cleansed data was then exported for database implementation.

Database Implementation (PostgreSQL)

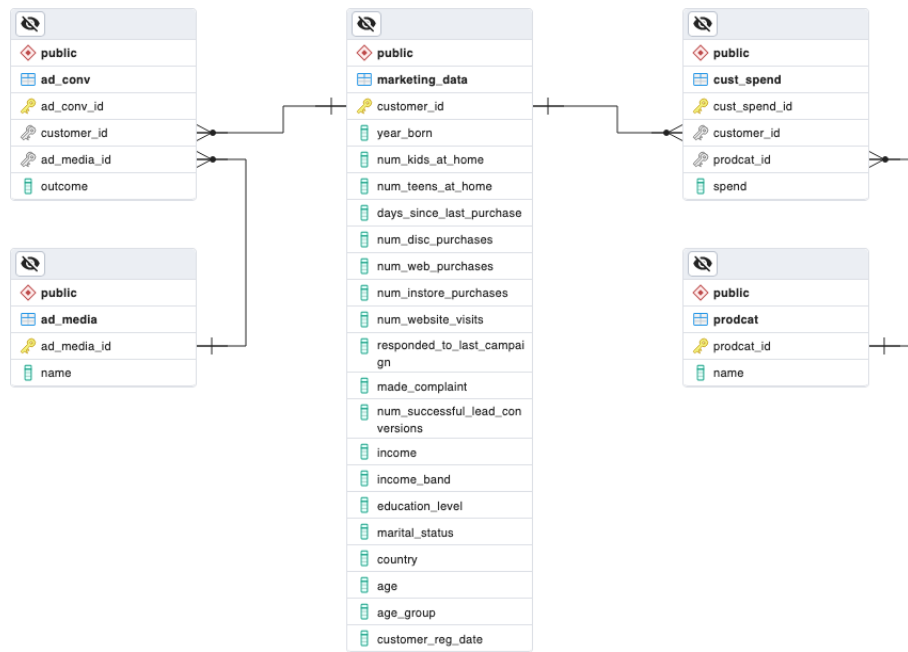
A 2Market database was created in PostgreSQL and the cleansed data was imported.

BOOLEAN data was imported as INTEGER to enable arithmetic operations (values constrained using CHECK statement). Customer Registration Date was imported as TEXT and converted DATE. Appendix C and D show evidence of SQL syntax and validation.

The database was validated against the original Excel files.

Data Normalisation (PostgreSQL)

The data required normalisation (see *Prototype Dashboard* below). Four new tables were created enabling arithmetic operations and improved efficiency. The revised structure is illustrated in the ERD below.

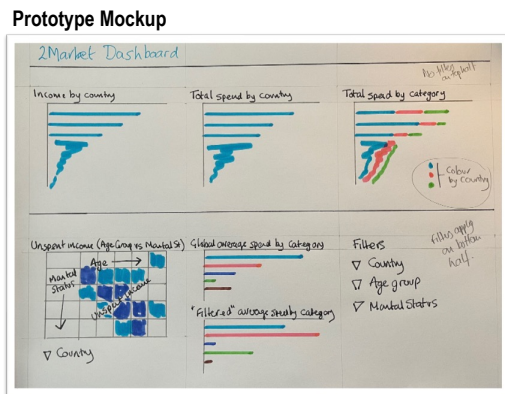


The new tables were validated against the original columns. Redundant columns and the Ad Data table were removed from the database. Appendices E and F show evidence of SQL syntax and validation.

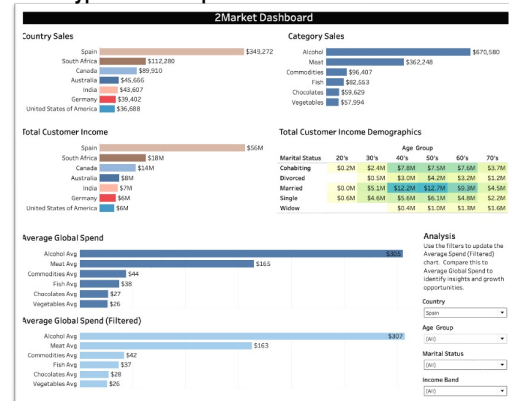
The normalised data was exported for use in Tableau.

Dashboard Design & Development

Prototype Dashboard



Prototype Tableau Implementation



The design approach was based on the principle of “overview; zoom and filter; details on demand”. A prototype sketch was created to guide implementation in Tableau, focusing on the key metrics of sales, customer demographics and customer incomes (not carried forward to the final dashboard).

Horizontal bar charts were the preferred chart type due to their ease of understanding. A heatmap table was also included to show the relationship between Age, Marital Status and Income. Filters were implemented for one of the bar charts.

The prototype highlighted a problem with the data.

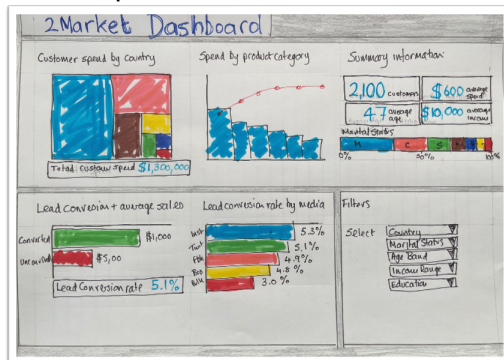
The sales by product category were held in separate columns (Amt*) and Tableau treated these as “different things”. This made arithmetic operations and analysis more complex than necessary. As a result of this, the data was normalised in ProgreSQL (see *Data Normalisation* above).

Additional insights:

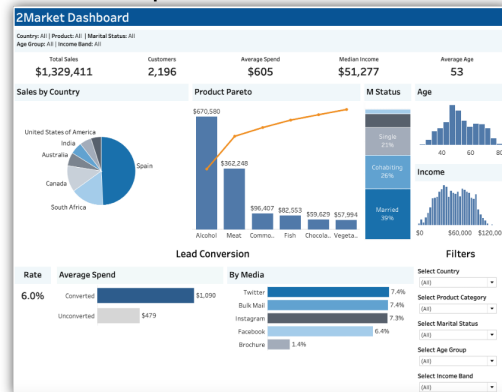
- Spain accounts for ~50% of sales
- Meat and alcohol account for ~75% of sales

Final Dashboard

Final Mockup



Final Tableau Implementation



A prototype sketch was created to guide implementation in Tableau, focusing on the key metrics of sales, customer demographics and lead conversion. This was not fully adhered to, and additional visualizations were added for Age and Income. A mixture of chart types and tables were implemented to attract and engage the user.

The final dashboard has four main sections:

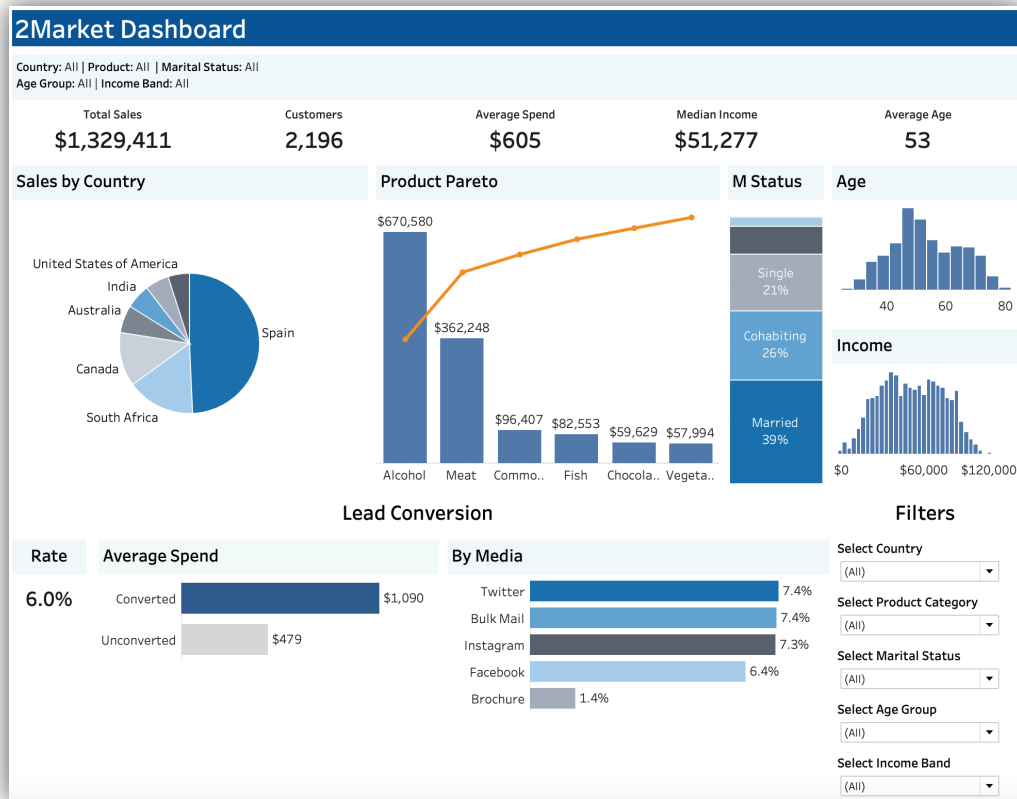
- Overview table across the top, showing summarised data.
- Sales demographics (Country, Products, Marital Status, Age and Income)
- Lead conversion
- Filters that allow the user to interact with the dashboard.

Key elements incorporated into the dashboard include:

- Colour scheme that provides accessibility for colour blind users
- Black text on white background for improved contrast
- Titles in bold with coloured background to differentiate visualisations.
- Dynamic title that changes when filter values change
- Simple meaningful headings
- Filter titles that encourage the user to investigate further.

Final Dashboard & Insights

Final Dashboard



Insights & Recommendations

Insights:

- 2Market has total sales of \$1.3M, 2,196 customers with average spend of \$605.
- The average age of customers is 53, median income is \$51K.
- Spain has the highest sales, 49% of the total.
- Three countries generate 77% of the total.
- Alcohol and meat account for 78% of sales.
- Married and cohabiting couples account for 65% of sales.
- Advertising lead conversion rate is 6.0%.
- Customers who respond to advertising spend \$611 more than those who don't.
- Brochure advertising is the least successful media with 1.4% lead conversion.

Analysis of marital status identified the widow demographic behaving differently. Widows:

- Have highest average spend of \$728, \$123 above overall.
- Respond better to advertising with 7.1% lead conversion, 1.1% above overall.
- Are responsive to advertising on Twitter with a 13.2% conversion rate, 7% above overall.


Based on the analysis so far, the following actions are recommended:

- Reduce or stop brochure advertising; redirect budget to better performing media.
- Focus on increasing lead conversion rate from 6.0% as customers who respond to advertising spends more than those who don't.

- Target new customers in the widow demographic (highest average spend). However, the ethics of this should be considered.
- Conduct further investigation to uncover additional insights in the data.
- Identify opportunities for dashboard enhancements.

APPENDIX

A. Data Cleansing

Income Column			
FROM		TO	Comments
\$7,500.00		7,500	Removal of leading "\$" sign and conversion to numeric enabling arithmetic operations.
\$14,421.00		14,421	
\$34,824.00		34,824	
\$34,824.00		34,824	
\$71,163.00		71,163	

B. Deleted Records

ID	Age	Marital_Status	Income	Country	Comments
11004	129	Single	\$60,182.00	SA	Unusual and / or unlikely value for Age
1150	123	Together	\$83,532.00	SP	Unusual and / or unlikely value for Age
7829	122	Divorced	\$36,640.00	IND	Unusual and / or unlikely value for Age
6663	82	Single	\$51,141.00	SP	Simplify analysis one of two records in Age Group "80's"
6932	81	Married	\$93,027.00	SP	Simplify analysis one of two records in Age Group 80's
4369	65	Absurd	\$65,487.00	CA	Unusual and / or unlikely value for Marital Status
7734	29	Absurd	\$79,244.00	AUS	Unusual and / or unlikely value for Marital Status
11133	49	YOLO	\$48,432.00	IND	Unusual and / or unlikely value for Marital Status
492	49	YOLO	\$48,432.00	CA	Unusual and / or unlikely value for Marital Status
9432	45	Together	\$666,666.00	SA	Unusual and / or unlikely value for Income
11181	73	Married	\$156,924.00	CA	Outside expected statistical range for Income
5336	51	Together	\$157,733.00	SP	Outside expected statistical range for Income
8475	49	Married	\$157,243.00	IND	Outside expected statistical range for Income
5555	47	Divorced	\$153,924.00	SP	Outside expected statistical range for Income
1503	46	Together	\$162,397.00	SP	Outside expected statistical range for Income
4931	45	Together	\$157,146.00	SA	Outside expected statistical range for Income
1501	40	Married	\$160,803.00	US	Outside expected statistical range for Income
5080	29	Single	\$70,515.00	ME	Simplify analysis one of three records in Country "ME"
9323	73	Together	\$49,912.00	ME	Simplify analysis one of three records in Country "ME"
2920	47	Single	\$52,614.00	ME	Simplify analysis one of three records in Country "ME"

C. Amended Field Values

Education Column

FROM	TO
Graduation	Graduate
Master	Masters
2n Cycle	Masters

Comments

Changed values for consistency. "2n Cycle" found to be equivalent to "Masters" following internet research.

Marital Status Column

FROM	TO
Alone	Single
Together	Cohabiting

Comments

Changed values to improve understanding.

Country Column

FROM	TO
AUS	Australia
CA	Canada
GER	Germany
IND	India
ME	Montenegro
SA	South Africa
SP	Spain
US	United States of America

Comments

Changed values to improve understanding, using long description from meta data

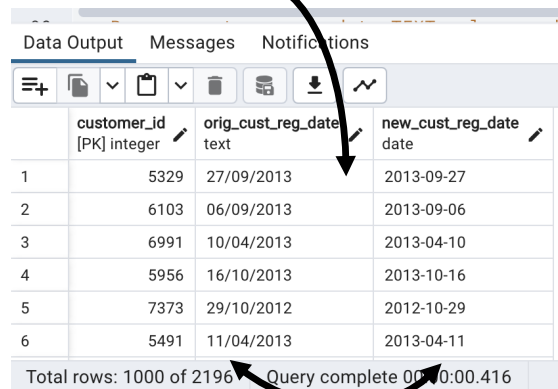
D. SQL Syntax

```
-- CREATE marketing data table
CREATE TABLE marketing_data(
  customer_id INTEGER PRIMARY KEY,
  year_born INTEGER,
  num_kids_at_home INTEGER,
  num_teens_at_home INTEGER,
  customer_reg_date TEXT,
  days_since_last_purchase INTEGER,
  spend_alc NUMERIC(7),
  spend_veg NUMERIC(7),
  spend_meat NUMERIC(7),
  spend_fish NUMERIC(7),
  spend_choc NUMERIC(7),
  spend_comm NUMERIC(7),
  num_disc_purchases INTEGER,
  num_web_purchases INTEGER,
  num_instore_purchases INTEGER,
  num_website_visits INTEGER,
  responded_to_last_campaign INTEGER,
    CHECK (responded_to_last_campaign = 0 OR responded_to_last_campaign = 1),
  made_complaint INTEGER,
    CHECK (made_complaint = 0 OR made_complaint = 1),
  num_successful_lead_conversions INTEGER,
  income NUMERIC(7),
  income_band VARCHAR(15),
  education_level VARCHAR(15),
  marital_status VARCHAR(15),
  country VARCHAR(30),
  age INTEGER,
  age_group CHAR(4)
);
```

E. Validation of Date Conversion

```
-- CREATE a temp table to convert date
CREATE TEMP TABLE tmp_date_conv(
    customer_id INTEGER PRIMARY KEY,
    orig_cust_reg_date TEXT,
    new_cust_reg_date DATE
);

-- POPULATE temp table
INSERT INTO tmp_date_conv(customer_id, orig_cust_reg_date, new_cust_reg_date)
SELECT customer_id, customer_reg_date, TO_DATE(customer_reg_date, 'DD/MM/YYYY')
FROM public.marketing_data;
```



	customer_id [PK] integer	orig_cust_reg_date text	new_cust_reg_date date
1	5329	27/09/2013	2013-09-27
2	6103	06/09/2013	2013-09-06
3	6991	10/04/2013	2013-04-10
4	5956	16/10/2013	2013-10-16
5	7373	29/10/2012	2012-10-29
6	5491	11/04/2013	2013-04-11

Total rows: 1000 of 2196 Query complete 00:00:00.416

Validate date
conversion

```
-- REMOVE customer_reg_date TEXT column, add customer_reg_date DATE column
ALTER TABLE public.marketing_data
    DROP COLUMN customer_reg_date;

ALTER TABLE public.marketing_data
    ADD COLUMN customer_reg_date DATE;

-- CHECK actions on table after each step
SELECT *
FROM public.marketing_data;

-- UPDATE new customer_reg_date DATE column from temp table
UPDATE public.marketing_data
SET customer_reg_date =(
    SELECT new_cust_reg_date
    FROM tmp_date_conv
    WHERE tmp_date_conv.customer_id = marketing_data.customer_id
);

-- CHECK dates loaded correctly against customer_id
SELECT customer_id, customer_reg_date
FROM public.marketing_data
ORDER BY customer_id
LIMIT 10;

SELECT *
FROM tmp_date_conv
ORDER BY customer_id
LIMIT 10;      -- with a visual comparison of both these queries

-- REMOVE temp table
DROP TABLE tmp_date_conv;
```

F. SQL Syntax

```
-- CREATE customer spend table
CREATE TABLE cust_spend(
  cust_spend_id SERIAL PRIMARY KEY,
  customer_id INTEGER,
  prodcats_id INTEGER,
  spend NUMERIC(7)
);

-- INSERT data into cust spend table for prodcats Alcohol
INSERT INTO public.cust_spend(customer_id, prodcats_id, spend)
SELECT customer_id, (VALUES(1)), spend_alc
FROM public.marketing_data;
```

G. Validation of New Customer Spend Table

Summarized data from new table

```
SELECT pc.prodcats_id, pc.name, SUM(cs.spend)
FROM public.prodcats pc
JOIN public.cust_spend cs USING(prodcats_id)
GROUP BY pc.prodcats_id, pc.name
ORDER BY pc.prodcats_id;
```

Data Output			
prodcats_id [PK] integer	name character varying (20)	sum numeric	
1	Alcohol	670580	
2	Vegetables	57994	
3	Meat	362248	
4	Fish	82553	
5	Chocolate	59629	
6	Commodities	96407	

Summarized data from original table

```
SELECT SUM(spend_alc) AS alcohol, SUM(spend_veg) AS vegetables,
SUM(spend_meat) AS meat, SUM(spend_fish) AS fish,
SUM(spend_choc) AS chocolate, SUM(spend_comm) AS commodities
FROM public.marketing_data;
```

Data Output						
alcohol numeric	vegetables numeric	meat numeric	fish numeric	chocolate numeric	commodities numeric	
1	670580	57994	362248	82553	59629	96407

Validate the two data sets match