2Market: Exploratory Analysis and Insights

By: Iulia-Diana Cristolovean

Last Updated: 14.12.2024

2Market, a global supermarket, faces challenges in understanding customer purchasing behaviour and its impact on sales performance. Its key objectives include making more data-driven marketing decisions, identifying shopping trends, assessing market opportunities, and optimizing resource allocation. To better understand the project, there are a few questions that need clarification:

- Are there specific customer segments that 2Market is interested in understanding better?
- What are 2Market's primary goals with these insights? (e.g. increase sales, optimize advertising spend)
- Does customer behaviour differ significantly between online and in-store purchases (e.g. types of products bought, amount spent)?

Analytical approach

The **objective** of this project is to analyse customer demographics, purchasing behaviour of customers, and how effective their advertising channels are.

The two data files ("marketing_data.csv", "ad_data.csv") were obtained from LSE company, providing raw data for the analysis. A description of the two .csv files was provided in the "metadata_2Market" text file.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	ID	Year_Birth	Education	Marital_Sta	Income	Kidhome	Teenhome	Dt_Customer	Recency	AmtLiq	AmtVege	AmtNonVe	AmtPes	AmtChoco	AmtComm	NumDeals	NumWebP	NumWalki	NumVisits	Response	Complain	Country	Count_success
2	1826	1971	Graduatio	Divorced	$84,835.0	0	0	6/16/14	0	189	104	379	111	189	218	1	4	6	1	1	0	SP	0
3	1	1962	Graduatio	Single	$57,091.0	0	0	6/15/14	0	464	5	64	7	0	37	1	7	7	5	1	0	CA	1
4	10476	1959	Graduatio	Married	$67,267.0	0	1	5/13/14	0	134	11	59	15	2	30	1	3	5	2	0	0	US	0
5	1386	1968	Graduatio	Together	$32,474.0	1	1	05/11/2014	0	10	0	1	0	0	0	1	1	2	7	0	0	AUS	0
6	5371	1990	Graduatio	Single	$21,474.0	1	0	04/08/2014	0	6	16	24	11	0	34	2	3	2	7	1	0	SP	1
7	7348	1959	PhD	Single	$71,691.0	0	0	3/17/14	0	336	130	411	240	32	43	1	4	5	2	1	0	SP	0
8	4073	1955	2n Cycle	Married	$63,564.0	0	0	1/29/14	0	769	80	252	15	34	65	1	10	7	6	1	0	GER	1
9	1991	1968	Graduatio	Together	$44,931.0	0	1	1/18/14	0	78	0	11	0	0	7	1	2	3	5	0	0	SP	0
10	4047	1955	PhD	Married	$65,324.0	0	1	01/11/2014	0	384	0	102	21	32	5	3	6	9	4	0	0	US	0
11	9477	1955	PhD	Married	$65,324.0	0	1	01/11/2014	0	384	0	102	21	32	5	3	6	9	4	0	0	IND	0
12	2079	1948	2n Cycle	Married	$81,044.0	0	0	12/27/13	0	450	26	535	73	98	26	1	5	10	1	0	0	US	0
13	5642	1980	Master	Together	$62,499.0	1	0	12/09/2013	0	140	4	61	0	13	4	2	3	6	4	0	0	SP	0
14	10530	1960	PhD	Widow	$67,786.0	0	0	12/07/2013	0	431	82	441	80	20	102	1	3	6	1	1	0	IND	0
15	2964	1982	Graduatio	Married	$26,872.0	0	0	10/16/13	0	3	10	8	3	16	32	1	1	2	6	0	0	CA	0
16	10311	1970	Graduatio	Married	$4,428.00	0	1	10/05/2013	0	16	4	12	2	4	321	0	25	0	1	0	0	SP	0
17	837	1978	Graduatio	Married	$54,809.0	1	1	09/11/2013	0	63	6	57	13	13	22	4	2	5	4	0	0	SP	0
18	10521	1978	Graduatio	Married	$54,809.0	1	1	09/11/2013	0	63	6	57	13	13	22	4	2	5	4	1	0	SP	0
19	10175	1959	PhD	Divorced	$32,173.0	0	1	08/01/2013	0	18	0	2	0	0	2	1	1	3	4	0	0	SP	0
20	1473	1961	2n Cycle	Single	$47,823.0	0	1	7/23/13	0	53	1	5	2	1	10	2	2	3	8	0	0	CA	0
21	2795	1959	Master	Single	$30,523.0	2	1	07/01/2013	0	5	0	3	0	0	5	1	1	2	7	0	0	CA	0
22	2285	1955	Master	Together	$36,634.0	0	1	5/28/13	0	213	9	76	4	3	30	3	5	5	7	0	0	SA	0
23	115	1967	Master	Single	$43,456.0	0	1	3/26/13	0	275	11	68	25	7	7	3	5	8	5	0	0	IND	0
24	10470	1980	Master	Married	$40,662.0	1	0	3/15/13	0	40	2	23	0	4	23	2	2	3	4	0	0	GER	0
25	4065	1977	PhD	Married	$49,544.0	1	0	02/12/2013	0	308	0	73	0	0	23	2	5	8	7	0	0	SP	0
26	10968	1970	Graduatio	Single	$57,731.0	0	1	11/23/12	0	266	21	300	65	8	44	4	8	6	6	0	0	IND	0
27	5985	1966	Master	Single	$33,168.0	0	1	10/13/12	0	80	1	37	0	1	3	3	2	4	7	0	0	SP	0
28	5430	1957	Graduatio	Together	$54,450.0	1	1	9/14/12	0	454	0	171	8	19	32	12	9	8	8	0	0	SP	0

	A	B	C	D	E	F	G	H	I
1	ID	Bulkmail_ad	Twitter_ad	Instagram_	Facebook_ad	Brochure_ad			
2	1826	0	0	0	0	0			
3	1	0	0	0	0	1			
4	10476	0	0	0	0	0			
5	1386	0	0	0	0	0			
6	5371	1	0	0	0	0			
7	7348	0	0	0	0	0			
8	4073	1	0	0	0	0			
9	1991	0	0	0	0	0			
10	4047	0	0	0	0	0			
11	9477	0	0	0	0	0			
12	2079	0	0	0	0	0			
13	5642	0	0	0	0	0			
14	10530	0	0	0	0	0			
15	2964	0	0	0	0	0			
16	10311	0	0	0	0	0			
17	837	0	0	0	0	0			
18	10521	0	0	0	0	0			
19	10175	0	0	0	0	0			
20	1473	0	0	0	0	0			
21	2795	0	0	0	0	0			
22	2285	0	0	0	0	0			
23	115	0	0	0	0	0			
24	10470	0	0	0	0	0			
25	4065	0	0	0	0	0			
26	10968	0	0	0	0	0			
27	5985	0	0	0	0	0			
28	5430	0	0	0	0	0			

Cleaning steps applied (see *APPENDIX pg. 10, Chapter 1: Cleaning steps*, for details):

1. The first steps of data cleaning were done in Excel, as the database provided is relatively small:
   - Check that the Primary Key is unique.
   - Age validation, meaning to check if there are customers older than 120, or younger than 18 (as the online department of the store sells alcohol).
   - “Marital_Status” column has 3 categories that stand apparat: “YOLO”, “Absurd”, “Alone”, so we replace “Alone” with “Single” to match one of the main categories, and “YOLO” and “Absurd” we will replace them with “Unknown”.
   - Remove the $ sign in the “Income” column.
   - Column “Dt_Customer” needs a consistent data formatting.
   - Check for missing values, or negative or invalid values in all the numeric columns.
   - Final check is that all the units are consistent with no typo errors or untidy text.

The result looks like this:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	Year_Birth	Age	Education	Marital_Status	Incom	Kidhome	Teenhome	Dt_Customer	Recency	AmtLiq	AmtVege	AmtNonVeg	AmtPes	AmtChocolates
2	1826	1971	53	Graduation	Divorced	84,835	0	0	2014-06-16	0	189	104	379	111	189
3	1	1962	62	Graduation	Single	57,091	0	0	2014-06-15	0	464	5	64	7	0
4	10476	1959	65	Graduation	Married	67,267	0	1	2014-05-13	0	134	11	59	15	2
5	1386	1968	56	Graduation	Together	32,474	1	1	2014-05-11	0	10	0	1	0	0
6	5371	1990	34	Graduation	Single	21,474	1	0	2014-04-08	0	6	16	24	11	0
7	7348	1959	65	PhD	Single	71,691	0	0	2014-03-17	0	336	130	411	240	32
8	4073	1955	69	2n Cycle	Married	63,564	0	0	2014-01-29	0	769	80	252	15	34
9	1991	1968	56	Graduation	Together	44,931	0	1	2014-01-18	0	78	0	11	0	0
10	4047	1955	69	PhD	Married	65,324	0	1	2014-01-11	0	384	0	102	21	32
11	9477	1955	69	PhD	Married	65,324	0	1	2014-01-11	0	384	0	102	21	32
12	2079	1948	76	2n Cycle	Married	81,044	0	0	2013-12-27	0	450	26	535	73	98
13	5642	1980	44	Master	Together	62,499	1	0	2013-12-09	0	140	4	61	0	13
14	10530	1960	64	PhD	Widow	67,786	0	0	2013-12-07	0	431	82	441	80	20
15	2964	1982	42	Graduation	Married	26,872	0	0	2013-10-16	0	3	10	8	3	16
16	10311	1970	54	Graduation	Married	4,428	0	1	2013-10-05	0	16	4	12	2	4
17	837	1978	46	Graduation	Married	54,809	1	1	2013-09-11	0	63	6	57	13	13
18	10521	1978	46	Graduation	Married	54,809	1	1	2013-09-11	0	63	6	57	13	13
19	10175	1959	65	PhD	Divorced	32,173	0	1	2013-08-01	0	18	0	2	0	0
20	1473	1961	63	2n Cycle	Single	47,823	0	1	2013-07-23	0	53	1	5	2	1
21	2795	1959	65	Master	Single	30,523	2	1	2013-07-01	0	5	0	3	0	0
22	2285	1955	69	Master	Together	36,634	0	1	2013-05-28	0	213	9	76	4	3
23	115	1967	57	Master	Single	43,456	0	1	2013-03-26	0	275	11	68	25	7
24	10470	1980	44	Master	Married	40,662	1	0	2013-03-15	0	40	2	23	0	4
25	4065	1977	47	PhD	Married	49,544	1	0	2013-02-12	0	308	0	73	0	0
26	10968	1970	54	Graduation	Single	57,731	0	1	2012-11-23	0	266	21	300	65	8
27	5985	1966	58	Master	Single	33,168	0	1	2012-10-13	0	80	1	37	0	1
28	5430	1957	67	Graduation	Together	54,450	1	1	2012-09-14	0	454	0	171	8	19

2. The second step is data validation in SQL (using PgAdmin 4) after creating the tables and importing the 2 data sets.

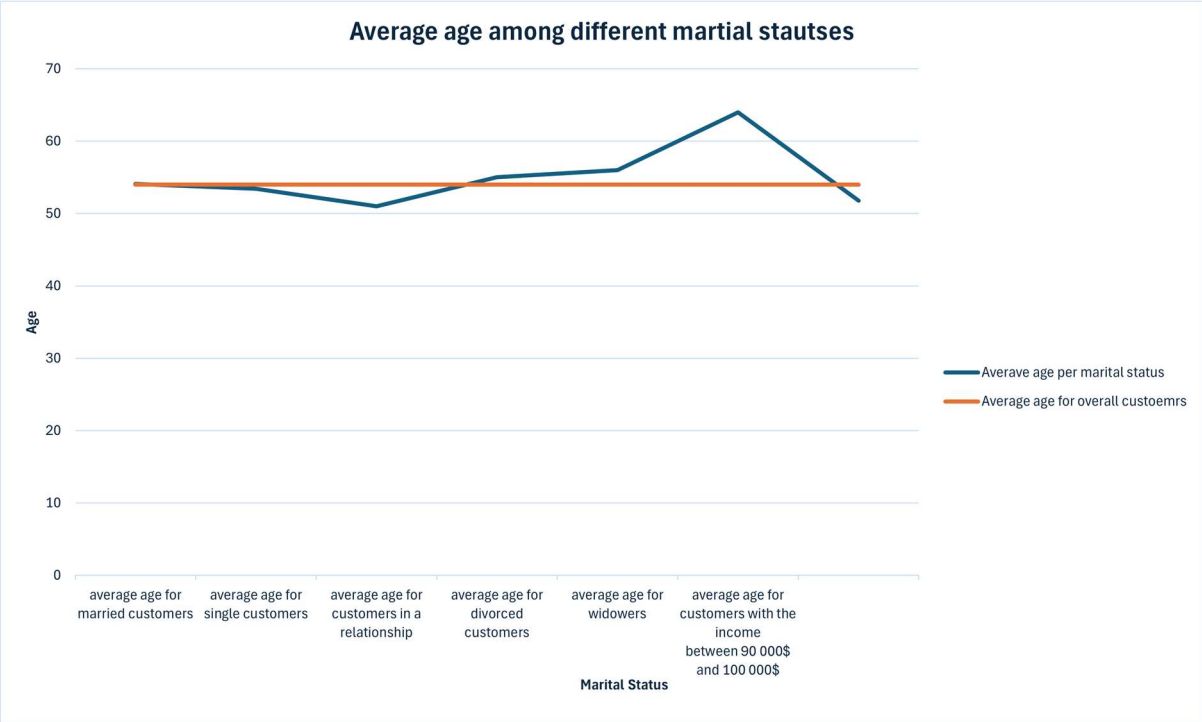
```sql
/* Create table for marketing_data.csv */

CREATE TABLE marketing_data (
    "ID" BIGSERIAL PRIMARY KEY,
    "Year_Birth" INT,
    "Age" INT,
    "Education" VARCHAR(20),
    "Marital_Status" VARCHAR(20),
    "Income" REAL,
    "Kidhome" INT,
    "Teenhome" INT,
    "Dt_Customer" DATE,
    "Recency" INT,
    "AmtLiq" INT,
    "AmtVege" INT,
    "AmtNonVeg" INT,
    "AmtPes" INT,
    "AmtChocolates" INT,
    "AmtComm" INT,
    "NumDeals" INT,
    "NumWebBuy" INT,
    "NumWalkinPur" INT,
    "NumVisits" INT,
    "Response" BOOLEAN,
    "Complain" BOOLEAN,
    "Country" VARCHAR(10),
    "Count_success" INT);
/*Create table for ad_data.csv */

CREATE TABLE ad_data(
    "ID" BIGSERIAL PRIMARY KEY,
    "Bulkmail_ad" BOOLEAN,
    "Twitter_ad" BOOLEAN,
    "Instagram_ad" BOOLEAN,
    "Facebook_ad" BOOLEAN,
    "Brochure_ad" BOOLEAN);
```

Because errors may appear when importing the data, it is good to redo some data validation: check for null values in the most important columns, Primary Key uniqueness in both data sets, check for duplicate rows, check the age range, and last, check that the date range is not invalid.

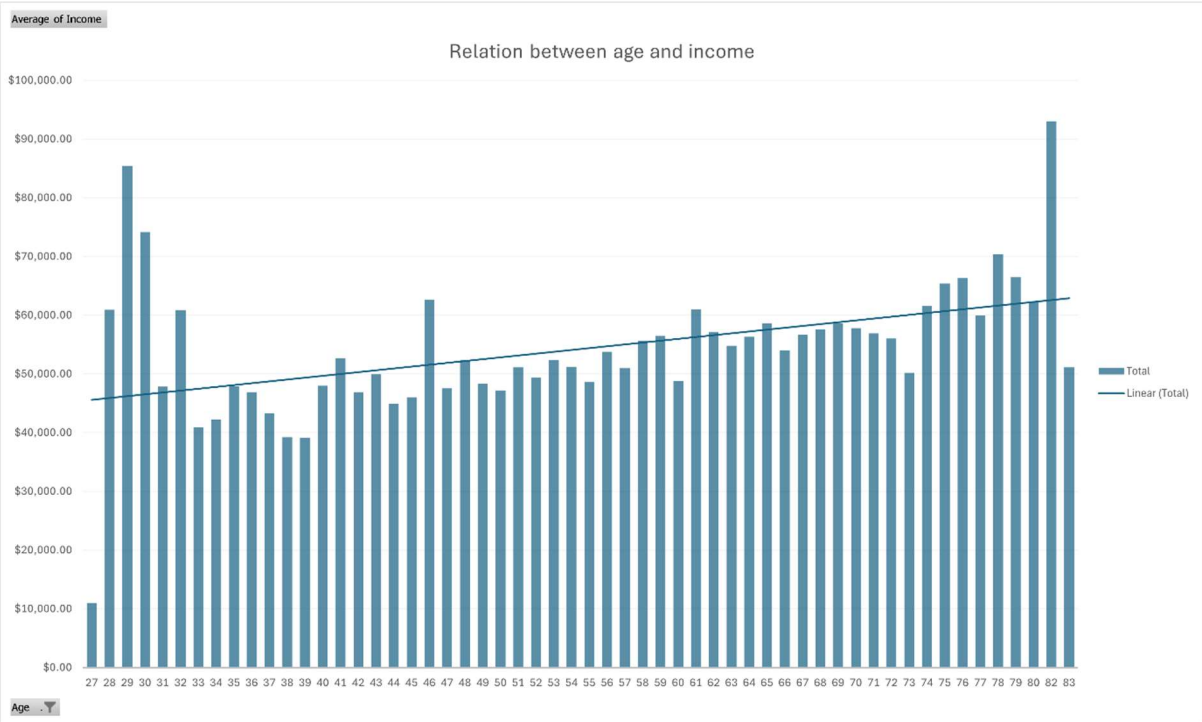
Dashboard design and development

After cleaning the data, I tried to gain some insights through basic visualisations in Excel. Using formulas and pivot tables I computed the average age and average age among different marital statuses.

average age for all customers	54
average age for married customers	53
average age for single customers	51
average age for customers in a relationship	55
average age for divorced customers	56
average age for widowers	64
average age for customers with the income between 90 000$ and 100 000$	52



The average age seems consistent among most of marital statuses, except for ‘widower’ where it is higher.

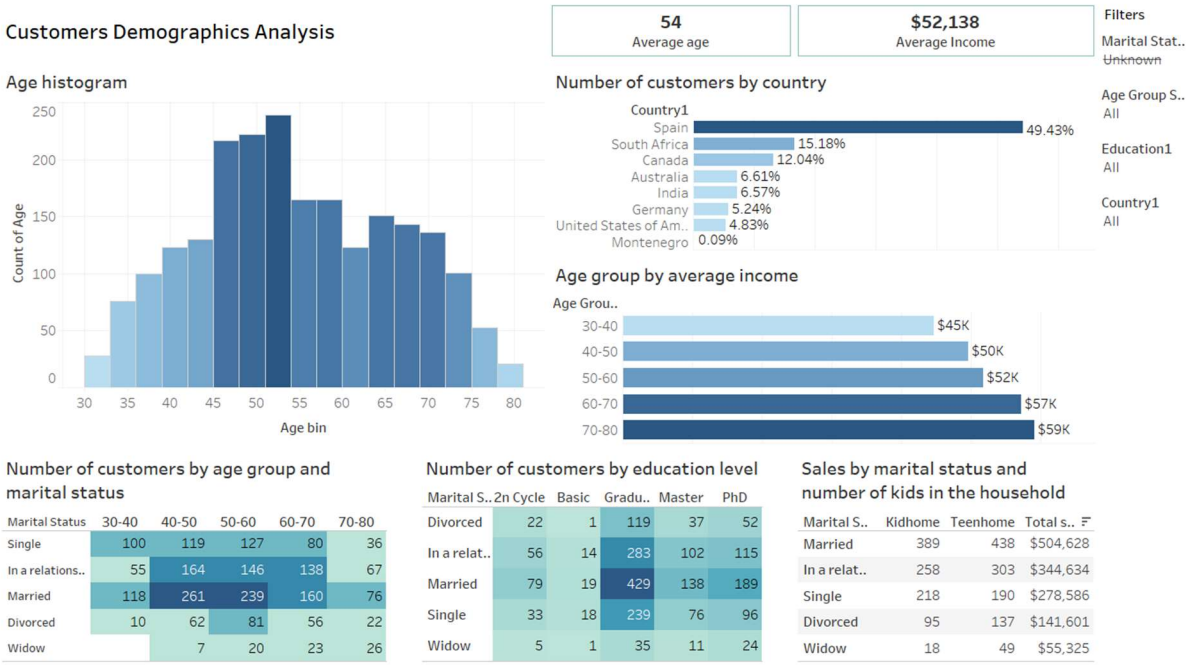


Also, the average income seems to be directly proportional with the age, as shown above by the trend-line.

I started the explorative analysis, using Tableau, to gain more in-depth information about the customers behaviour. I constructed the dashboards to be clear and accessible, using:

- bar charts to compare quantitative data.
- histograms for age distribution, as they effectively show frequency distribution over a continuous range.
- highlight tables for marital status and education levels, to capture dense categorical data and show patterns effectively.
- aggregated KPIs as stand-alone metrics, to give a quick summary.

The layout of the dashboards is divided into meaningful sections, each focusing on specific aspects of the data (demographics, sales, campaigns). My colour choices are mainly shades of blue to be easy on the eye, but also colourblind friendly, ensuring the data interpretation remains inclusive. The filters (for country, education, age group, marital status) allow users to focus on specific data subsets, making the dashboard adaptable to diverse needs.



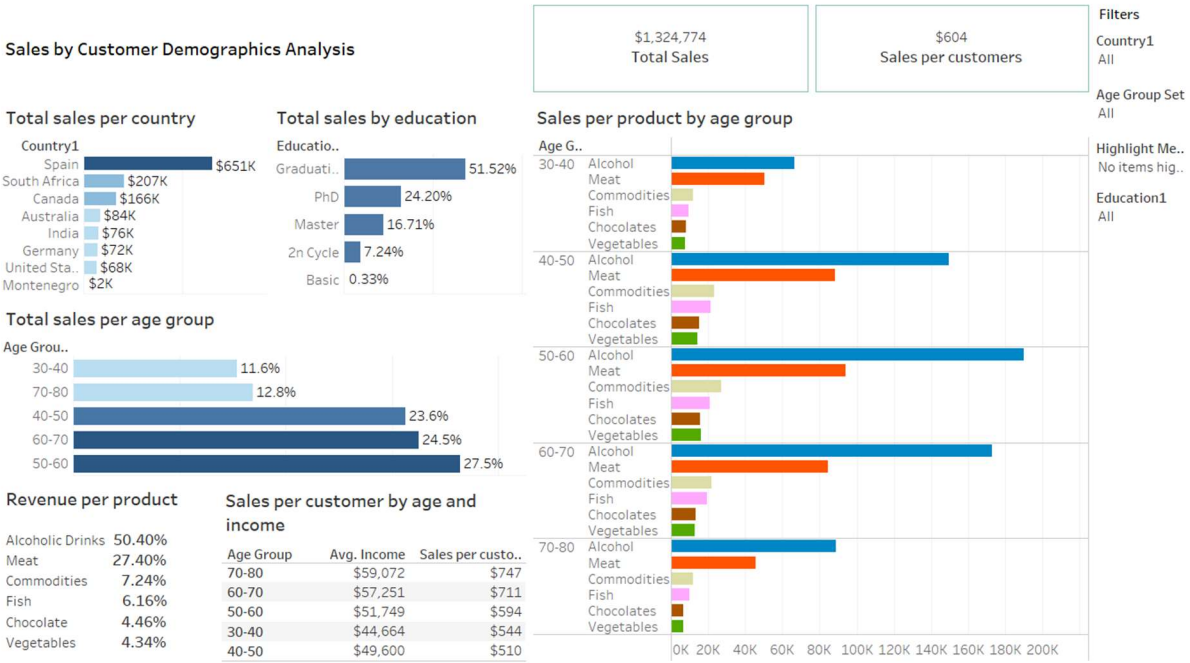
Marital Status	30-40	40-50	50-60	60-70	70-80
Single	100	119	127	80	36
In a relations..	55	164	146	138	67
Married	118	261	239	160	76
Divorced	10	62	81	56	22
Widow		7	20	23	26

Marital S..	2n Cycle	Basic	Gradu..	Master	PhD
Divorced	22	1	119	37	52
In a relat..	56	14	283	102	115
Married	79	19	429	138	189
Single	33	18	239	76	96
Widow	5	1	35	11	24

Marital S..	Kidhome	Teenhome	Total s..
Married	389	438	$504,628
In a relat..	258	303	$344,634
Single	218	190	$278,586
Divorced	95	137	$141,601
Widow	18	49	$55,325

Insights derived from the customer demographics:

- most customers are from Spain.
- highest concentration of customers is between 50 - 60 years old, married and with a higher education level.

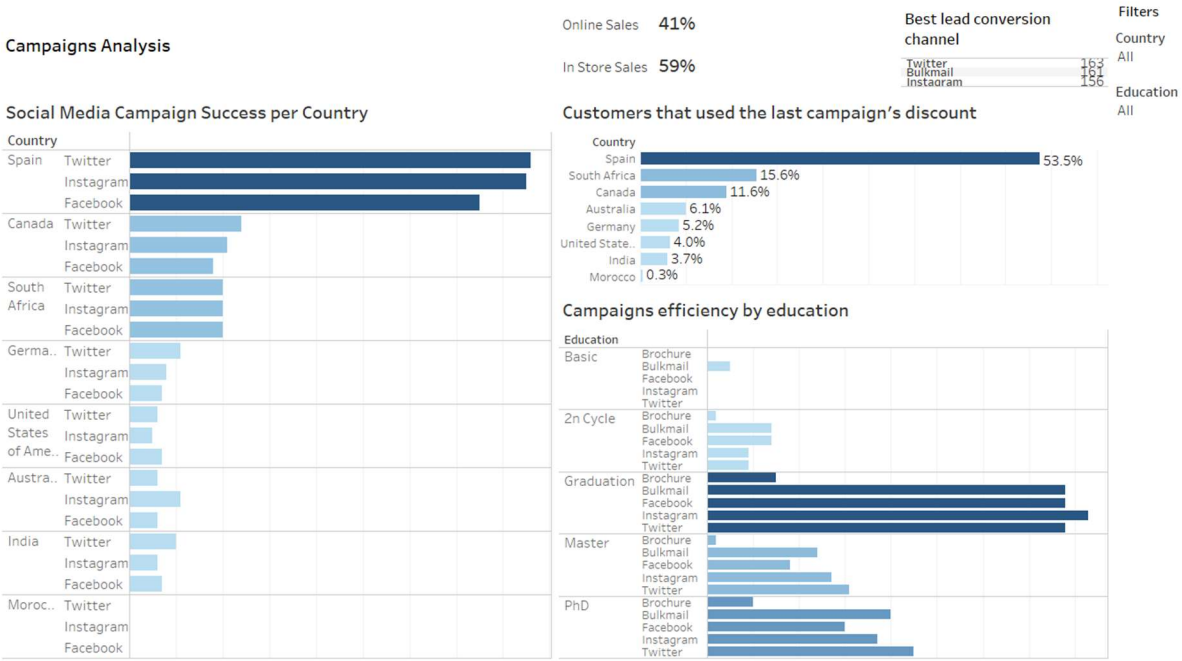


Age Group	Avg. Income	Sales per custo..
70-80	$59,072	$747
60-70	$57,251	$711
50-60	$51,749	$594
30-40	$44,664	$544
40-50	$49,600	$510

Insights derived from sales analysis:

- Spain dominates total sales, while Montenegro lags significantly.
- 50-60 age groups contribute most to total sales.
- alcohol and meat are top product across all customer demographics.

- customers with ‘Graduation’ education level contribute for over 50% of sales.



Campaign insights:

- Spain is the leading country in social media campaigns success, in alignment with customers using the most campaign discounts.
- higher education levels (Graduation, Master and PhD) show the most effective campaign responses, especially through Twitter and Instagram.

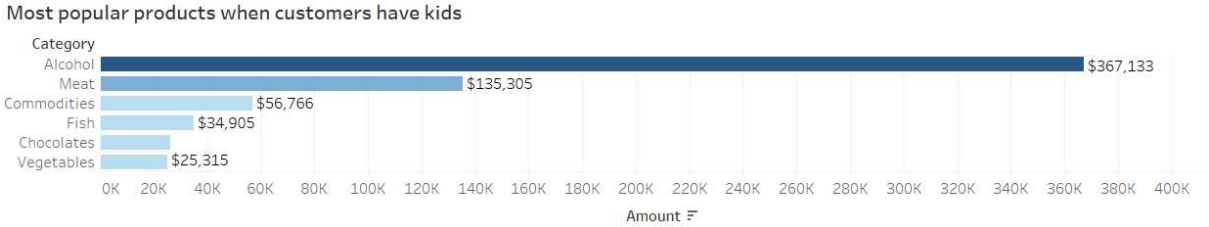
Patterns, trends, and insights

To check for more sales insights, I used SQL (for details see *APPENDIX pg. 15, Chapter 2: SQL analysis*). The results showed:

- The most sold product in each country: the table shows that alcoholic drinks are the best-selling products. Montenegro has the lowest sales, aligning with its minimal social media presence and possibly reflecting cultural restrictions on alcohol consumption.

	Country character varying (10)	best_selling_product text	total_sales_of_best_product bigint
1	SP	Alcoholic Drinks	336392
2	SA	Alcoholic Drinks	105918
3	CA	Alcoholic Drinks	84066
4	AUS	Alcoholic Drinks	42752
5	GER	Alcoholic Drinks	36776
6	IND	Alcoholic Drinks	36236
7	US	Alcoholic Drinks	32214
8	ME	Alcoholic Drinks	1729

- For product popularity amongst customers with kids, I exported the data from SQL and made a Tableau visualization.



- Ad conversion by marital status: Brochure conversions have the lowest overall conversion rates compared to the other digital channels, suggesting room for optimization in that marketing channel, while married customers seem to be the most responsive to marketing efforts.

	Marital_Status character varying (20)	facebook_conversions bigint	twitter_conversions bigint	instagram_conversions bigint	bulkmail_conversions bigint	brochure_conversions bigint
1	Together	32	42	44	37	12
2	Unknown	1	0	1	0	0
3	Married	62	62	66	63	7
4	Widow	5	10	7	4	1
5	Single	30	32	31	39	5
6	Divorced	12	18	13	20	5

- Success rate of the last ad campaign per country: given the relatively high success rate in Marocco at 66.67%, it could be worth expanding marketing efforts into this smaller but more responsive market.

	Country character varying (10)	total_number_of_customers bigint	total_sales bigint	accepted_campaign_offer bigint	success_rate text
1	SP	1093	659557	176	16.10%
2	SA	337	211071	52	15.43%
3	CA	266	167403	38	14.29%
4	AUS	147	85576	22	14.97%
5	GER	116	73198	17	14.66%
6	US	107	67546	13	12.15%
7	IND	147	77806	13	8.84%
8	ME	3	3122	2	66.67%

- Most popular product based on education level: higher education levels correlate with increased sales of alcoholic drinks, likely reflecting higher incomes.

	Education character varying (20)	best_selling_product text	total_sales_of_best_product bigint
1	Graduation	Alcoholic Drinks	318111
2	PhD	Alcoholic Drinks	195874
3	Master	Alcoholic Drinks	121538
4	2n Cycle	Alcoholic Drinks	40169
5	Basic	Commodities	1233

To better understand our customers, we need to know not just what they buy, but how often, how much, and why. Additional data that could help us target new opportunities would be:

- Purchase frequency, basket size, and average order value.
- Product preferences over time to track seasonality or trends.
- Competitor performance data in the same markets.
- Click-through rates, time spent on marketing content, and social media engagement metrics.
- Customer feedback and reviews for qualitative insights.
- Customer acquisition costs, churn rates, and retention metrics.

If we had more time, we could go further with advanced segmentation which may reveal hidden patterns and help us create micro-segments for hyper-personalized campaigns. We could explore how campaigns perform across different platforms (e.g. Twitter vs. Instagram) for distinct income groups, or to assess the Return on Investment (ROI) of offline vs. online channels.

Looking ahead, integrating all this data into a unified system could transform our analytics for the better, making campaigns smarter, markets more targeted, and decisions more actionable.

APPENDIX: Data Report for 2Market

By: Iulia-Diana Cristolovean

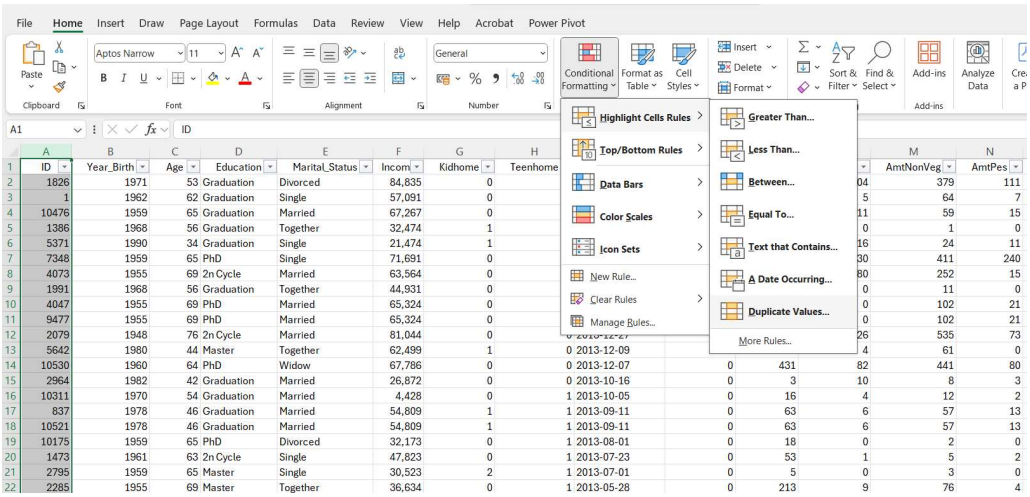
Last Updated: 14.12.2024

Contents:

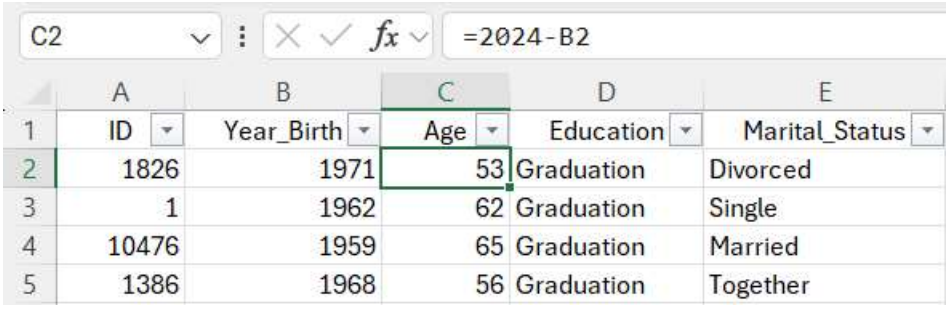
1. Cleaning steps applied:

The first steps of data cleaning will be done in Excel, as the database provided is not too large:

- Check that the Primary key, “ID” in each table, is unique, using *Conditional Formatting*. Not having unique customer ID can cause problems with duplicate data. No duplicate key found in any of the files, but I did notice that the customer IDs are not consecutive, so I need to ask the system administrator if this is an issue or there is another reason to have deleted data.



- Age validation, meaning check if there are customers older than 120, or younger than 18 (as the online department of the store sells alcohol). For this we are going to create another column to compute the age for each customer. We will apply the formula for every row.

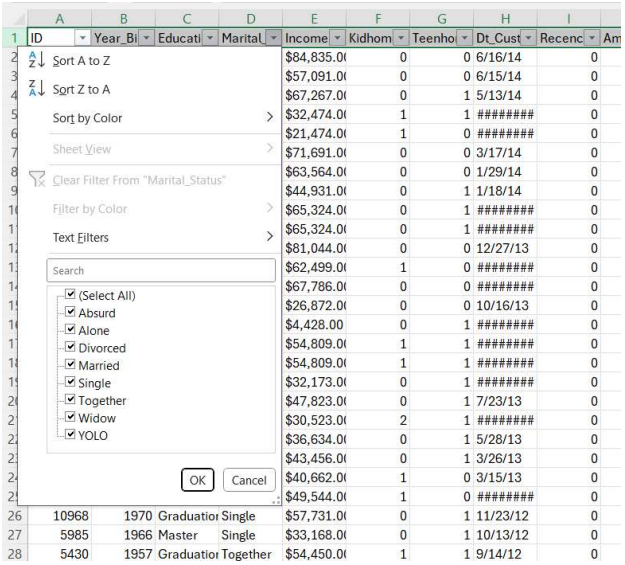


C2 =2024-B2

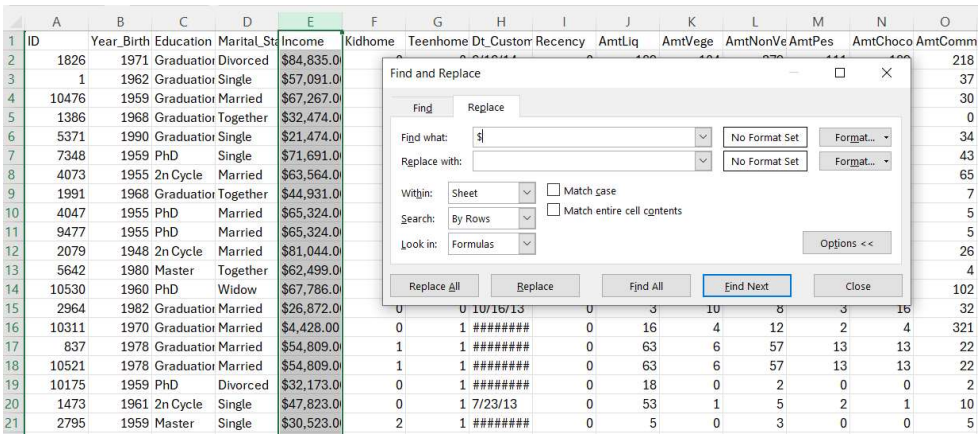
	A	B	C	D	E
1	ID	Year_Birth	Age	Education	Marital_Status
2	1826	1971	53	Graduation	Divorced
3	1	1962	62	Graduation	Single
4	10476	1959	65	Graduation	Married
5	1386	1968	56	Graduation	Together

There were 3 customers over the age of 120 and replace the “Age” and “Year_Birth” columns with ‘null’, although we can also delete them, as they are too few to affect our analysis.

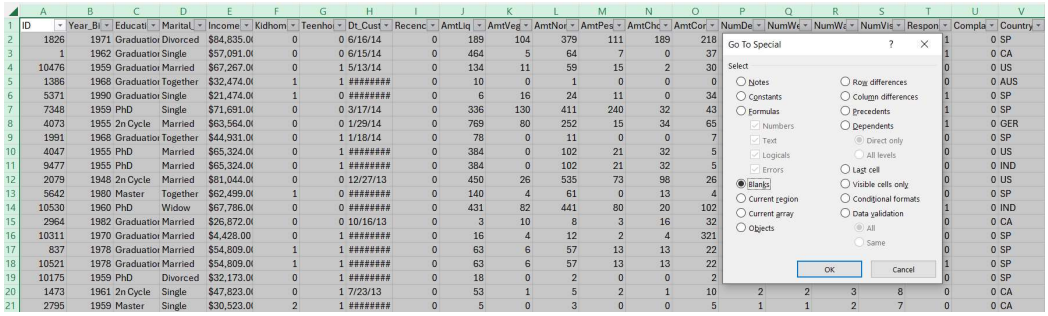
- From a quick view in the “Marital_Status” column filter we noticed that it has 3 categories that stand apparat: “YOLO”, “Absurd”, “Alone”, so we replace “Alone” with “Single” to match one of the main categories, and “YOLO” and “Absurd” we will rename them “Unknown”. There are two customers with “YOLO” marital status, but the ID and country are different so we will not consider them duplicates.



- Remove the $ sign in the “Income” column, as it is recognised as text, using *FIND AND REPLACE*.



- The last column that needs cleaning is “Dt_Customer”. The easiest way to is to use *TEXTSPLIT* and then combine the values from each row into the desired format (‘yyyy-mm-dd’). I checked that no date is in the future or to far into the past. At this point I noticed that the last year they registered a new customer was 2014, so I need to ask the system administrator why I am not receiving new data.
- Checked for missing values the *GO TO SPECIAL* feature, and no empty cells were found.



ID	Year_Bi	Educati	Marital_	Income	Kidhom	Teenhol	Dt_Cust	Recenc	AmtLiq	AmtVeg	AmtNor	AmtPes	AmtChc	AmtCor	NumDe	NumWe	NumWe	NumVis	Respon	Compla	Country
1826	1971	Graduatio	Divorced	$84,835.0	0	0	6/16/14	0	189	104	379	111	189	218					1	0	SP
1	1962	Graduatio	Single	$57,091.0	0	0	6/15/14	0	464	5	64	7	0	37					1	0	CA
10476	1959	Graduatio	Married	$67,267.0	0	1	5/13/14	0	134	11	59	15	2	30					0	0	US
1386	1968	Graduatio	Together	$32,474.0	1	1	########	0	10	0	1	0	0	0					0	0	AUS
5371	1990	Graduatio	Single	$21,474.0	1	0	########	0	6	16	24	11	0	34					1	0	SP
7348	1959	PhD	Single	$71,691.0	0	0	3/17/14	0	336	130	411	240	32	43					1	0	SP
4073	1955	2n Cycle	Married	$63,564.0	0	0	1/29/14	0	769	80	252	15	34	65					1	0	GER
1991	1968	Graduatio	Together	$44,931.0	0	1	1/18/14	0	78	0	11	0	0	7					0	0	SP
4047	1955	PhD	Married	$65,324.0	0	1	########	0	384	0	102	21	32	5					0	0	US
9477	1955	PhD	Married	$65,324.0	0	1	########	0	384	0	102	21	32	5					0	0	IND
2079	1948	2n Cycle	Married	$81,044.0	0	0	12/27/13	0	450	26	535	73	98	26					0	0	US
5642	1980	Master	Together	$62,499.0	1	0	########	0	140	4	61	0	13	4					0	0	SP
10530	1960	PhD	Widow	$67,786.0	0	0	########	0	431	82	441	80	20	102					1	0	IND
2964	1982	Graduatio	Married	$26,872.0	0	0	10/16/13	0	3	10	8	3	16	32					0	0	CA
10311	1970	Graduatio	Married	$4,428.00	0	1	########	0	16	4	12	2	4	321					0	0	SP
837	1978	Graduatio	Married	$54,809.0	1	1	########	0	63	6	57	13	13	22					0	0	SP
10521	1978	Graduatio	Married	$54,809.0	1	1	########	0	63	6	57	13	13	22					1	0	SP
10175	1959	PhD	Divorced	$32,173.0	0	1	########	0	18	0	2	0	0	2					0	0	SP
1473	1961	2n Cycle	Single	$47,823.0	0	1	7/23/13	0	53	1	5	2	1	10	2	2	3	8	0	0	CA
2795	1959	Master	Single	$30,523.0	2	1	########	0	5	0	3	0	0	5	1	1	2	7	0	0	CA

- Checked for negative or invalid values in all the numeric columns, ad also made sure that all the units are consistent.
- Remove outliers using Excel formulas for interquartile range (IQR), calculate the first and third quartiles, and the upper and lower limits.

Quartile 1	35303
Quartile 3	68522
Interquartile Range	33219
Standard Deviation	25173.07666
Variance	633683788.6
Outlier Lower Limit	-14525.5
Outlier Upper Limit	118350.5

From the values of our lower and upper limits, we can consider higher than 118350.5 as outliers and exclude them from the set, because the lower limit is negative. Because we found only 8, we decided not to delete them.

ID	Year_Birth	Age	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	AmtV	AmtVe	AmtNonV	AmtPe	AmtChocolat
4931	1978	46	Graduation	Together	157,146	0	0	29/04/2013	13	1	0	1725	2	1
1501	1983	41	PhD	Married	160,803	0	0	04/08/2012	21	55	16	1622	17	3
9432	1978	46	Graduation	Together	666,666	1	0	02/06/2013	23	9	14	18	8	1
1503	1977	47	PhD	Together	162,397	1	1	03/06/2013	31	85	1	16	2	1
5336	1972	52	Master	Together	157,733	1	0	04/06/2013	37	39	1	9	2	0
5555	1976	48	Graduation	Divorced	153,924	0	0	07/02/2014	81	1	1	1	1	1
11181	1950	74	PhD	Married	156,924	0	0	29/08/2013	85	2	1	2	1	1
8475	1974	50	PhD	Married	157,243	0	1	01/03/2014	98	20	2	1582	1	2

- Last, by using *PROPER* we can make sure that the text in each column is spelled correctly and is case consistent, while *TRIM* removes unwanted spaces and characters. After a check of all the columns we saw that the data is clean, with no typo errors or untidy text.

The result looks like this:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	Year_Birth	Age	Education	Marital_Status	Incom	Kidhome	Teenhome	Dt_Customer	Recency	AmtLiq	AmtVege	AmtNonVeg	AmtPes	AmtChocolates
2	1826	1971	53	Graduation	Divorced	84,835	0	0	2014-06-16	0	189	104	379	111	189
3	1	1962	62	Graduation	Single	57,091	0	0	2014-06-15	0	464	5	64	7	0
4	10476	1959	65	Graduation	Married	67,267	0	1	2014-05-13	0	134	11	59	15	2
5	1386	1968	56	Graduation	Together	32,474	1	1	2014-05-11	0	10	0	1	0	0
6	5371	1990	34	Graduation	Single	21,474	1	0	2014-04-08	0	6	16	24	11	0
7	7348	1959	65	PhD	Single	71,691	0	0	2014-03-17	0	336	130	411	240	32
8	4073	1955	69	2n Cycle	Married	63,564	0	0	2014-01-29	0	769	80	252	15	34
9	1991	1968	56	Graduation	Together	44,931	0	1	2014-01-18	0	78	0	11	0	0
10	4047	1955	69	PhD	Married	65,324	0	1	2014-01-11	0	384	0	102	21	32
11	9477	1955	69	PhD	Married	65,324	0	1	2014-01-11	0	384	0	102	21	32
12	2079	1948	76	2n Cycle	Married	81,044	0	0	2013-12-27	0	450	26	535	73	98
13	5642	1980	44	Master	Together	62,499	1	0	2013-12-09	0	140	4	61	0	13
14	10530	1960	64	PhD	Widow	67,786	0	0	2013-12-07	0	431	82	441	80	20
15	2964	1982	42	Graduation	Married	26,872	0	0	2013-10-16	0	3	10	8	3	16
16	10311	1970	54	Graduation	Married	4,428	0	1	2013-10-05	0	16	4	12	2	4
17	837	1978	46	Graduation	Married	54,809	1	1	2013-09-11	0	63	6	57	13	13
18	10521	1978	46	Graduation	Married	54,809	1	1	2013-09-11	0	63	6	57	13	13
19	10175	1959	65	PhD	Divorced	32,173	0	1	2013-08-01	0	18	0	2	0	0
20	1473	1961	63	2n Cycle	Single	47,823	0	1	2013-07-23	0	53	1	5	2	1
21	2795	1959	65	Master	Single	30,523	2	1	2013-07-01	0	5	0	3	0	0
22	2285	1955	69	Master	Together	36,634	0	1	2013-05-28	0	213	9	76	4	3
23	115	1967	57	Master	Single	43,456	0	1	2013-03-26	0	275	11	68	25	7
24	10470	1980	44	Master	Married	40,662	1	0	2013-03-15	0	40	2	23	0	4
25	4065	1977	47	PhD	Married	49,544	1	0	2013-02-12	0	308	0	73	0	0
26	10968	1970	54	Graduation	Single	57,731	0	1	2012-11-23	0	266	21	300	65	8
27	5985	1966	58	Master	Single	33,168	0	1	2012-10-13	0	80	1	37	0	1
28	5430	1957	67	Graduation	Together	54,450	1	1	2012-09-14	0	454	0	171	8	19

The second step is data validation in SQL (using PgAdmin 4), after creating the tables for the two data sets:

```sql
/* Create table for marketing_data.csv */

CREATE TABLE marketing_data (
    "ID" BIGSERIAL PRIMARY KEY,
    "Year_Birth" INT,
    "Age" INT,
    "Education" VARCHAR(20),
    "Marital_Status" VARCHAR(20),
    "Income" REAL,
    "Kidhome" INT,
    "Teenhome" INT,
    "Dt_Customer" DATE,
    "Recency" INT,
    "AmtLiq" INT,
    "AmtVege" INT,
    "AmtNonVeg" INT,
    "AmtPes" INT,
    "AmtChocolates" INT,
    "AmtComm" INT,
    "NumDeals" INT,
    "NumWebBuy" INT,
    "NumWalkinPur" INT,
    "NumVisits" INT,
    "Response" BOOLEAN,
    "Complain" BOOLEAN,
    "Country" VARCHAR(10),
    "Count_success" INT);
```

```sql
/*Create table for ad_data.csv */

CREATE TABLE ad_data(
    "ID" BIGSERIAL PRIMARY KEY,
    "Bulkmail_ad" BOOLEAN,
    "Twitter_ad" BOOLEAN,
    "Instagram_ad" BOOLEAN,
    "Facebook_ad" BOOLEAN,
    "Brochure_ad" BOOLEAN);
```

- Check for null values in the most important columns.

	ID [PK] bigint	Age integer	Education character varying (20)	Marital_Status character varying (20)	Income real	Country character varying (10)	AmtLiq integer	AmtVege integer	AmtNonVeg integer	AmtPes integer	AmtChocolates integer	AmtComm integer
1	11004	[null]	2n Cycle	Single	60182	SA	8	0	5	7	0	2
2	1150	[null]	PhD	Together	83532	SP	755	144	562	104	64	224
3	7829	[null]	2n Cycle	Divorced	36640	IND	15	6	8	7	4	25

The result showed the 3 customers who had the age greater than 120, which probably means that there was a typo in the birth year column, so we can make a note of them, as they will have no influence over the analysis.

- Check for duplicate Primary Key in both data sets.

```sql
SELECT
    "ID",
    COUNT (*) AS ordercount
FROM public.marketing_data
GROUP BY "ID"
HAVING COUNT(*)>1;
```

```sql
SELECT
    "ID",
    COUNT (*) AS ordercount
FROM public.ad_data
GROUP BY "ID"
HAVING COUNT(*)>1;
```

No duplicate IDs found in any of the tables.

- Check for duplicate rows in the ‘marketing_data’ table, based on a combination of columns.

```sql
/* Rows duplicate */

SELECT
    "ID",
    "Age",
    "Education",
    "Marital_Status",
    "Income",
    "Country",
    COUNT (*) AS duplicate_count
FROM public.marketing_data
GROUP BY
    "ID",
    "Age",
    "Education",
    "Marital_Status",
    "Income",
    "Country"
HAVING COUNT(*)>1;
```

No duplicate rows were found.

- Check the age range, as the company sells alcoholic products online.

```sql
/* Age checks*/

SELECT *
FROM public.marketing_data
WHERE "Age">120 OR
      "Age"<17;
```

No invalid data was found.

- Check that the date range are not invalid.

```sql
/* Date checks */

SELECT *
FROM public.marketing_data
WHERE "Dt_Customer"> CURRENT_DATE;
```

No invalid dates were found.

- Numeric range date checks, for columns with negative data.

```sql
/* NUmeric range chacks */

SELECT
    "ID",
    "Age",
    "Income",
    "AmtLiq",
    "AmtVege",
    "AmtNonVeg",
    "AmtPes",
    "AmtChocolates",
    "AmtComm"
FROM public.marketing_data
WHERE
    "ID" <0 OR
    "Age" <0 OR
    "Income" <0 OR
    "AmtLiq" <0 OR
    "AmtVege" <0 OR
    "AmtNonVeg" <0 OR
    "AmtPes" <0 OR
    "AmtChocolates" <0 OR
    "AmtComm" <0;
```

2. SQL Analysis

The syntax for the explorative analysis of the sales data that has been performed in SQL:

- The most sold product in each country:

```sql
WITH sales_data AS (
SELECT
    "Country",
    SUM ("AmtLiq")  AS total_alcoholic_drinks,
    SUM ("AmtVege") AS total_veggies,
    SUM ("AmtNonVeg") AS total_meat,
    SUM ("AmtPes") AS total_fish,
    SUM ("AmtChocolates") AS total_chocolate,
    SUM ("AmtComm") AS total_commodities
FROM marketing_data
GROUP BY "Country" )
SELECT
    "Country",
    CASE
        WHEN total_alcoholic_drinks >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities) T
        WHEN total_veggies >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities) THEN 'Vegg
        WHEN total_meat >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities) THEN 'Meat'
        WHEN total_fish >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities) THEN 'Fish'
        WHEN total_chocolate >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities) THEN 'Ch
        WHEN total_commodities >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities) THEN '
    END AS best_selling_product,
    GREATEST(
        total_alcoholic_drinks,
        total_veggies,
        total_meat,
        total_fish,
        total_chocolate,
        total_commodities
    ) AS total_sales_of_best_product
FROM sales_data
ORDER BY total_sales_of_best_product desc;
```

The result shows Alcoholic Drinks are the best-selling product across all countries, indicating a universal trend for this category. Spain tops the table with the highest total sales of alcoholic drinks ($336,392), far ahead of other countries.

	Country character varying (10)	best_selling_product text	total_sales_of_best_product bigint
1	SP	Alcoholic Drinks	336392
2	SA	Alcoholic Drinks	105918
3	CA	Alcoholic Drinks	84066
4	AUS	Alcoholic Drinks	42752
5	GER	Alcoholic Drinks	36776
6	IND	Alcoholic Drinks	36236
7	US	Alcoholic Drinks	32214
8	ME	Alcoholic Drinks	1729

- Product popularity amongst customers with kids:

```sql
/* which products are the most popular when there are children or teens in the home */

SELECT
    "Kidhome", "Teenhome",
    'AmtLiq' AS category,
    SUM("AmtLiq") AS amount
FROM "marketing_data"
WHERE "Kidhome" != '0' OR "Teenhome" !='0'
GROUP BY "Kidhome", "Teenhome"
UNION ALL
SELECT
    "Kidhome", "Teenhome",
    'AmtVege' AS category,
    SUM("AmtVege") AS amount
FROM "marketing_data"
WHERE "Kidhome" != '0' OR "Teenhome" !='0'
GROUP BY "Kidhome", "Teenhome"
UNION ALL
SELECT
    "Kidhome", "Teenhome",
    'AmtNonVeg' AS category,
    SUM("AmtNonVeg") AS amount
FROM "marketing_data"
WHERE "Kidhome" != '0' OR "Teenhome" !='0'
GROUP BY "Kidhome", "Teenhome"
UNION ALL
```

```sql
SELECT
    "Kidhome", "Teenhome",
    'AmtPes' AS category,
    SUM("AmtPes") AS amount
FROM "marketing_data"
WHERE "Kidhome" != '0' OR "Teenhome" !='0'
GROUP BY "Kidhome", "Teenhome"
UNION ALL
SELECT
    "Kidhome", "Teenhome",
    'AmtChocolates' AS category,
    SUM("AmtChocolates") AS amount
FROM "marketing_data"
WHERE "Kidhome" != '0' OR "Teenhome" !='0'
GROUP BY "Kidhome", "Teenhome"
UNION ALL
SELECT
    "Kidhome", "Teenhome",
    'AmtComm' AS category,
    SUM("AmtComm") AS amount
FROM "marketing_data"
WHERE "Kidhome" != '0' OR "Teenhome" !='0'
GROUP BY "Kidhome", "Teenhome"
ORDER BY amount DESC;
```

The result shows households with teenagers generate higher sales than those with younger kids, but the most popular product remain alcoholic drinks followed by meat products, showing a general trend regardless of the demographics.

	Kidhome integer	Teenhome integer	category text	amount bigint
1	0	1	AmtLiq	258984
2	0	1	AmtNonVeg	86357
3	1	1	AmtLiq	45805
4	1	0	AmtLiq	40949
5	0	1	AmtComm	34666
6	1	0	AmtNonVeg	24463
7	0	1	AmtPes	22684
8	0	1	AmtChocolates	17841
9	0	1	AmtVege	16840
10	1	1	AmtNonVeg	16785

- Ad conversion by marital status:

```sql
/* Which social media platform is the most effective method of advertising based on marital status? */

SELECT
    "Marital_Status",
    SUM(CASE WHEN "Facebook_ad" = TRUE THEN 1 ELSE 0 END) AS facebook_conversions,
    SUM(CASE WHEN "Twitter_ad" = TRUE THEN 1 ELSE 0 END) AS twitter_conversions,
    SUM(CASE WHEN "Instagram_ad" = TRUE THEN 1 ELSE 0 END) AS instagram_conversions,
    SUM(CASE WHEN "Bulkmail_ad" = TRUE THEN 1 ELSE 0 END) AS bulkmail_conversions,
    SUM(CASE WHEN "Brochure_ad" = TRUE THEN 1 ELSE 0 END) AS brochure_conversions
FROM joint_data
GROUP BY "Marital_Status";
```

The result shows Married individuals have the highest conversions across most platforms. Brochure conversions are the lowest among all channels, irrespective of marital status. Bulk mail shows moderate conversions for groups like 'Together', 'Married', and 'Single'.

	Marital_Status character varying (20)	facebook_conversions bigint	twitter_conversions bigint	instagram_conversions bigint	bulkmail_conversions bigint	brochure_conversions bigint
1	Together	32	42	44	37	12
2	Unknown	1	0	1	0	0
3	Married	62	62	66	63	7
4	Widow	5	10	7	4	1
5	Single	30	32	31	39	5
6	Divorced	12	18	13	20	5

- Success rate of the last ad campaign per country:

```sql
/* total sales by country and whether how many accepted the last campaing's offer */
/* what is the success rate of the last campaign's offer */

SELECT
    "Country",
    COUNT ("ID") AS total_number_of_customers,
    SUM("AmtLiq" + "AmtVege" + "AmtNonVeg" + "AmtPes" + "AmtChocolates" + "AmtComm") AS total_sales,
    SUM(CASE WHEN "Response" = TRUE THEN 1 ELSE 0 END) AS accepted_campaign_offer,
    ROUND(
    (CAST(SUM(CASE WHEN "Response" = TRUE THEN 1 ELSE 0 END) AS DECIMAL) /
     CAST(COUNT("ID") AS DECIMAL)) * 100,
     2) || '%' AS success_rate
FROM joint_data
GROUP BY "Country"
ORDER BY accepted_campaign_offer DESC;
```

The result shows Spain has the highest percentage of customers using discounts, which aligns with its social media campaign success, but given the relatively high success rate in Montenegro at 66.67%, it could be worth expanding marketing efforts into this smaller but more responsive market.

	Country character varying (10)	total_number_of_customers bigint	total_sales bigint	accepted_campaign_offer bigint	success_rate text
1	SP	1093	659557	176	16.10%
2	SA	337	211071	52	15.43%
3	CA	266	167403	38	14.29%
4	AUS	147	85576	22	14.97%
5	GER	116	73198	17	14.66%
6	US	107	67546	13	12.15%
7	IND	147	77806	13	8.84%
8	ME	3	3122	2	66.67%

- Most popular product based on education level:

```sql
/* which specific product is the most popular based on the education level */

WITH sales_data AS (
SELECT     "Education",
    SUM ("AmtLiq")  AS total_alcoholic_drinks,
    SUM ("AmtVege") AS total_veggies,
    SUM ("AmtNonVeg") AS total_meat,
    SUM ("AmtPes") AS total_fish,
    SUM ("AmtChocolates") AS total_chocolate,
    SUM ("AmtComm") AS total_commodities
FROM joint_data
GROUP BY "Education")
SELECT     "Education",
    CASE
        WHEN total_alcoholic_drinks >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities)
        WHEN total_veggies >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities) THEN 'Veg
        WHEN total_meat >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities) THEN 'Meat'
        WHEN total_fish >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities) THEN 'Fish'
        WHEN total_chocolate >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities) THEN 'C
        WHEN total_commodities >= GREATEST(total_alcoholic_drinks, total_veggies, total_meat, total_fish, total_chocolate, total_commodities) THEN
    END AS best_selling_product,
    GREATEST(
        total_alcoholic_drinks,
        total_veggies,
        total_meat,
        total_fish,
        total_chocolate,
        total_commodities
    ) AS total_sales_of_best_product
FROM sales_data
ORDER BY total_sales_of_best_product desc;
```

The result shows that the highest sales for alcoholic drinks occur among individuals with Graduation (318,111), followed by PhD holders (195,874) and Master graduates (121,538). This trend suggests higher sales among the more educated demographic. Individuals with a Basic education show minimal sales of alcoholic drinks (1,233) compared to other groups, reflecting either a lower disposable income or differing preferences in this group.

	Education character varying (20)	best_selling_product text	total_sales_of_best_product bigint
1	Graduation	Alcoholic Drinks	318111
2	PhD	Alcoholic Drinks	195874
3	Master	Alcoholic Drinks	121538
4	2n Cycle	Alcoholic Drinks	40169
5	Basic	Commodities	1233