2Market: Exploratory Data Analysis and Predictive Modelling of Customer Spending

Student: Monica Urquiza Ribeiro Baracho

Table of Contents

1. Overview

This report presents insights from an exploratory data analysis conducted for **2Market**, a global supermarket with both online and in-store channels. The goal is to inform 2Market's 2025 marketing strategy by analysing customer demographics, purchasing behaviour, and advertising effectiveness.

Two datasets were analysed—**customer profiles** and **advertising conversions**—using **Excel**, **SQL**, **Tableau**, and **R**. The analysis addressed three core questions:

- Who are our customers? (age, income, education, marital status)
- Which advertising channels drive the most conversions?
- Which products are most popular across customer segments?

Findings are supported by dashboards and a regression model designed to inform data-driven campaign planning.

2. Tools and Methodology

Data Sources

- marketing_data.csv: Contains customer demographics, spending behaviours, income, education, and marital status.
- ad_data.csv: Records customer responses to various marketing channels (e.g., Instagram, Facebook, brochures).

Tools

- Excel: For initial data cleaning, handling missing values, calculating age, and generating basic visualisations.
- SQL (PostgreSQL): Used to join datasets, calculate spending by segment, and summarize category spend by education level.
- Tableau: Built two interactive dashboards to present key insights to stakeholders.
- RStudio: regression analysis

Data Preparation

- Cleaning: Handled missing values, removed duplicates, and standardized data types (e.g., formatted income fields).
- Category Grouping: Combined "YOLO", "Alone", and "Single" into a single "Single" category to reflect similar behavioural patterns in spending and campaign response (see Appendix A1 – Average Age by Marital Status).
- Invalid Values: Replaced "Absurd" entries with #N/A to prevent skewed analysis.
- Feature Engineering: Calculated age from Year_Birth using 2025 as the reference year, and created total spend variables by category.
- Data Join: Merged marketing_data and ad_data using the ID field.

While the dataset is historical (last recorded in 2014), this analysis projects demographic attributes (e.g., age) to 2025 to align insights with current business strategy needs. Education

is assumed stable over time, while marital status and channel effectiveness may have changed. These assumptions are acknowledged, and results should be interpreted as indicative rather than definitive. Future models would benefit from updated datasets to improve predictive accuracy and reflect changing consumer behaviour, especially in digital engagement.

SQL queries[1] were used extensively to support data cleaning, aggregation, and exploration. For instance, queries were designed to calculate total product spend by marital status, identify top-spending customer segments, and group data by family type.

3. Key Findings

Three interactive Tableau dashboards were developed to present key customer insights clearly and engagingly.

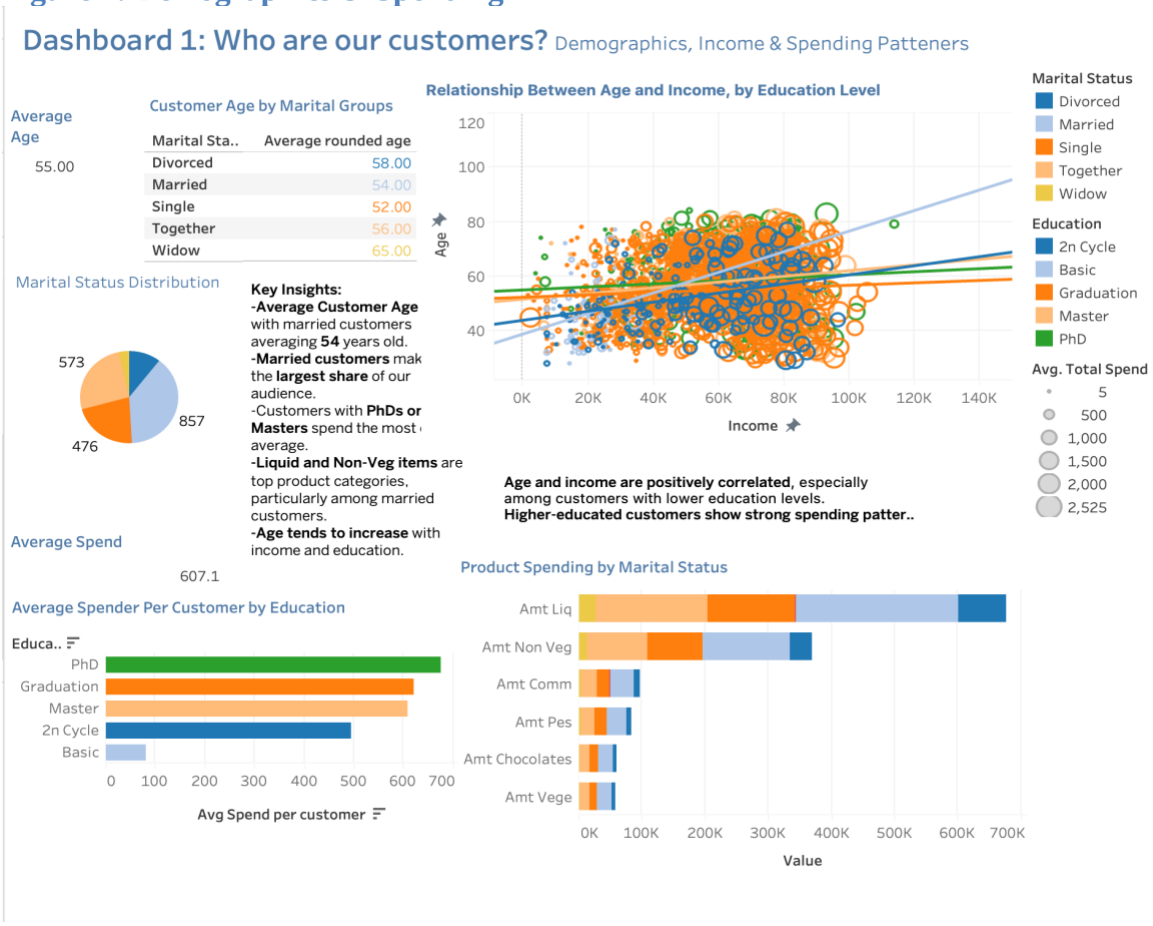
The Demographics & Spending Patterns dashboard highlights the following:

- The average customer age is approximately 55 years. (44 Years if we had considered 2014 as the base)
- Married individuals represent the most significant customer segment, with around 857 customers 38.7% followed by together 573 with 26% of the total customers
- There is a positive correlation between age, income, and education level.
- Customers with master's or PhD degrees demonstrate the highest average spending across all segments.
- Liquor and non-vegetables are the most purchased product categories, particularly among higher-educated and married customers.

This scatterplot illustrates the relationship between income and age, segmented by education level. Each point represents an individual customer, with a circle size indicating their average total spend.

[1] The full SQL queries can be found in Appendix **B – SQL Queries**

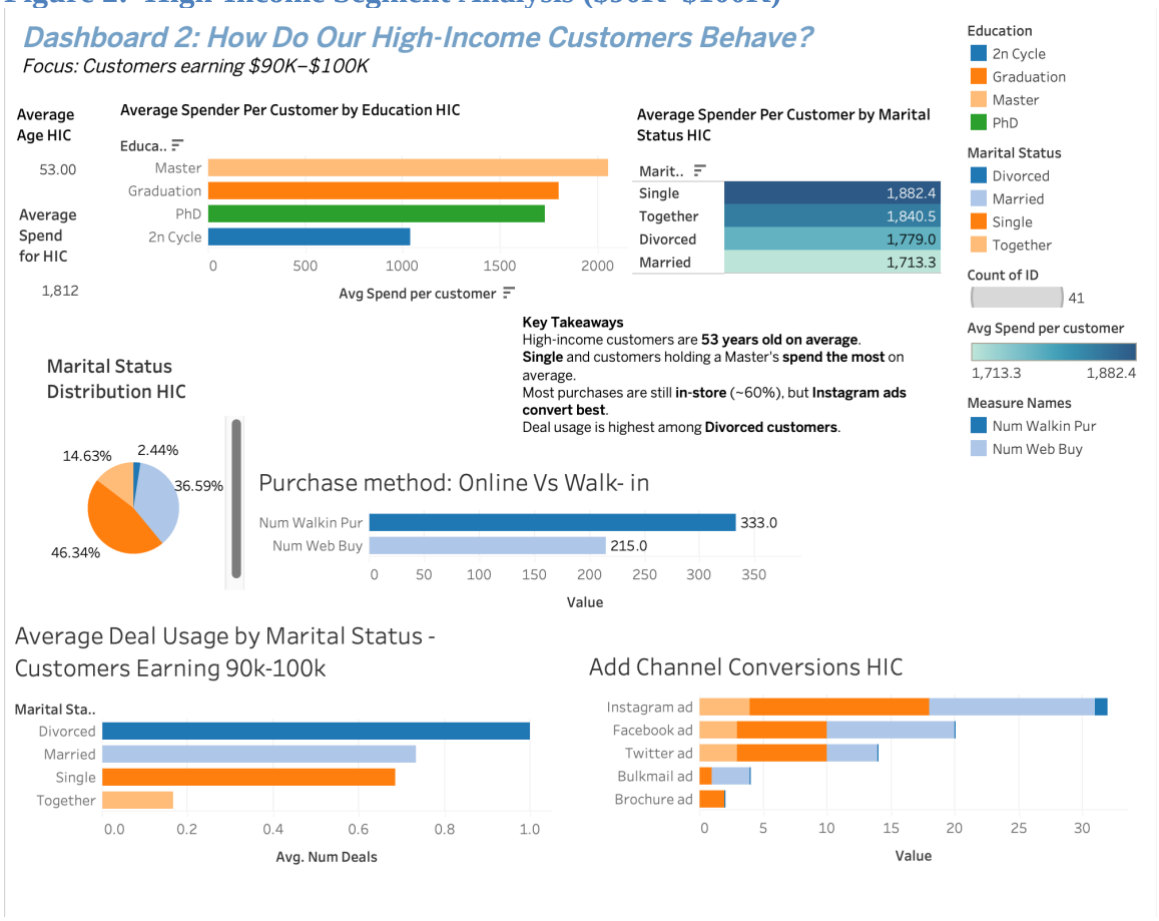
Figure 1: Demographics & Spending



Our analysis shows that high-income customers (earning $90K–$100K) are typically around 53 years old, with single and divorced individuals spending the most. Spending is also strongly linked to education, with master's degree holders leading in expenditure. Surprisingly, 60% of purchases still occur in-store, even among this digitally engaged group—highlighting the value of in-store experiences.

Instagram stands out as the top-performing ad channel, appealing to this segment's preference for visually rich, aspirational content. SQL analysis further reveals that households without young children or teens spend more on premium products like liquor, chocolates, and meat. This suggests a strong opportunity for Instagram-driven premium campaigns targeting older, high-income, educated customers without young kids.

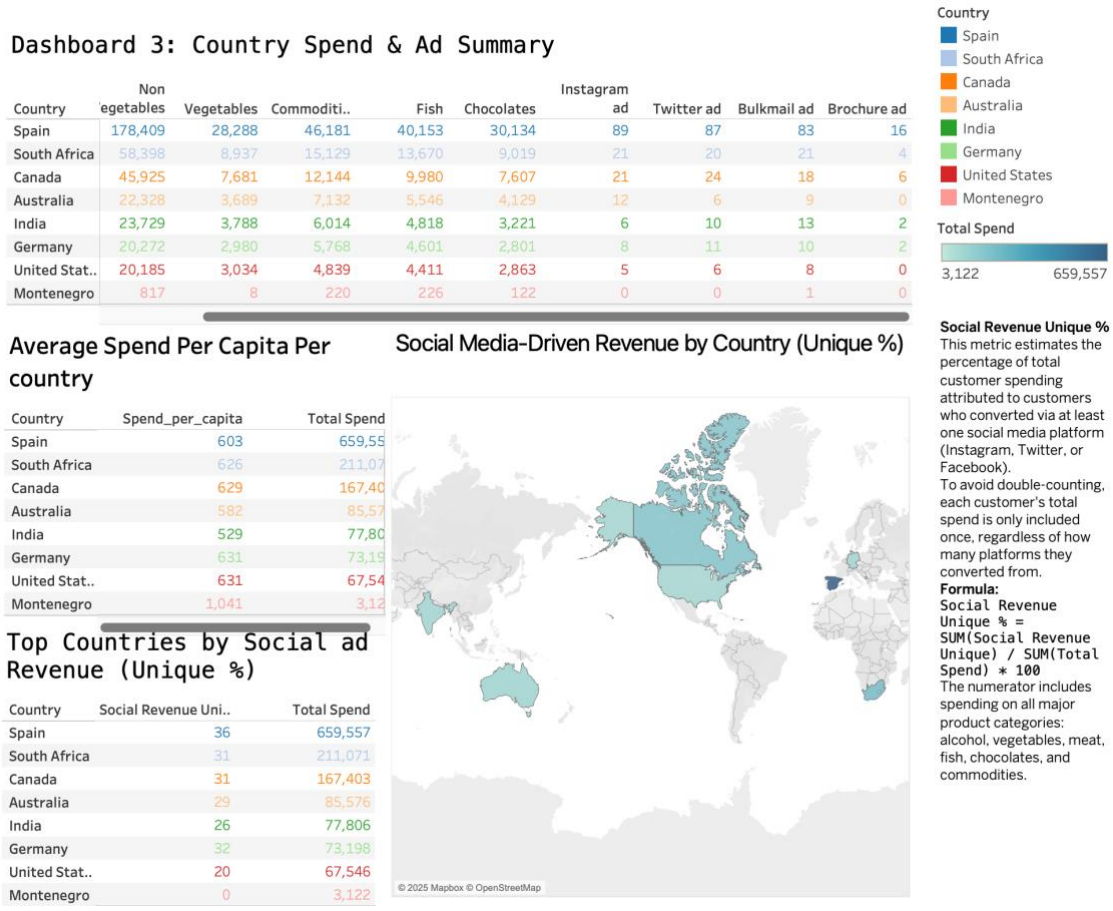
Figure 2: High-Income Segment Analysis ($90K–$100K)



Dashboard 3 below shows country-level customer spending, response to social media ads, and Unique % of Social Revenue, which is the share of spending from customers who converted via at least one social platform.

- Spain had the highest total spend ($659,557) and Social Revenue Unique % (36%).
- Canada and South Africa followed, each with 31%.
- Montenegro had no revenue from social media conversions. This insight helps identify where social ads are most effective and where to focus future campaigns.

Figure 3: Country Spend & Ad Summary



Regression Analysis Summary

We used a multiple linear regression model to understand and forecast customer spending behaviour for the 2025 campaign. This approach allows us to quantify how different factors — such as income, purchase behaviour, and ad exposure — influence total spending.

We applied a **log transformation to income** to correct for skewness and better capture its relationship with spending. The model was trained on a cleaned dataset, excluding extreme outliers and statistically insignificant variables like Twitter ads. We validated the model using **robust standard errors** to account for potential heteroskedasticity and confirmed all included predictors were statistically significant.

The final model explains **approximately 73% of the variance in customer spending**, making it a reliable and interpretable tool for guiding campaign strategy and resource allocation. (see **Appendix C**

Appendix C.1-)

4. Recommendations

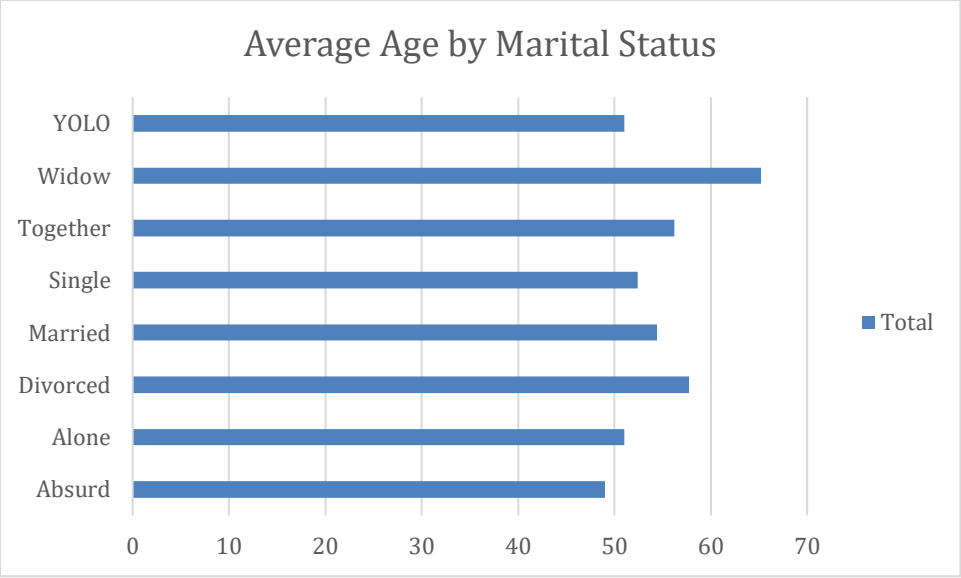
Descriptive and regression analysis reveal that 2Market's highest spenders are older (45–65), high-income, highly educated customers without children**.** They prefer in-store and online purchases and are concentrated in Spain, South Africa, and Canada**.**

Key spending drivers include income, purchase behaviour, and Instagram ad exposure, while households with children spend less. We recommend focusing campaigns on this core segment, prioritizing Instagram**,** enhancing in-store experiences, and exploring TikTok **t**o stay ahead of digital trends.

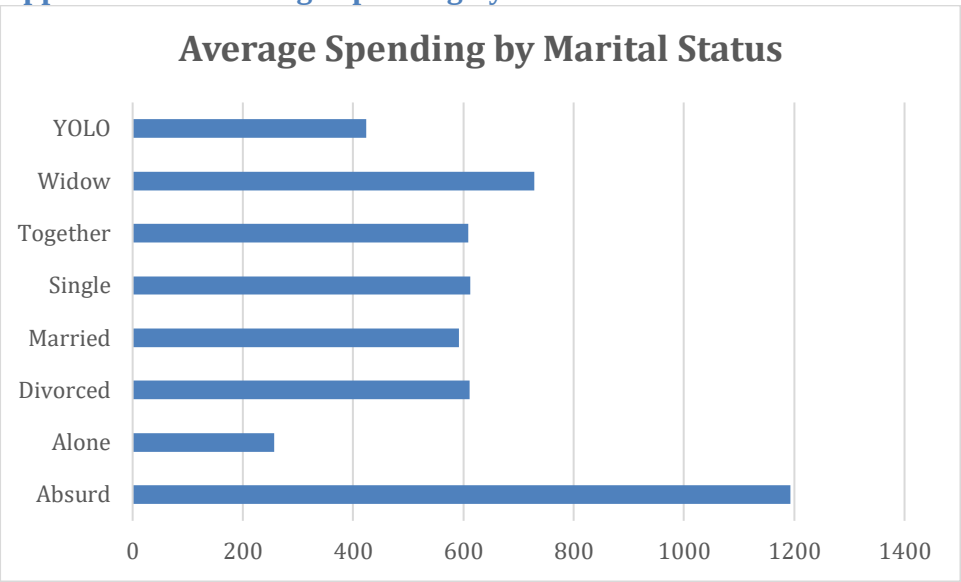
Appendices

Appendix A: Excel Visualizations

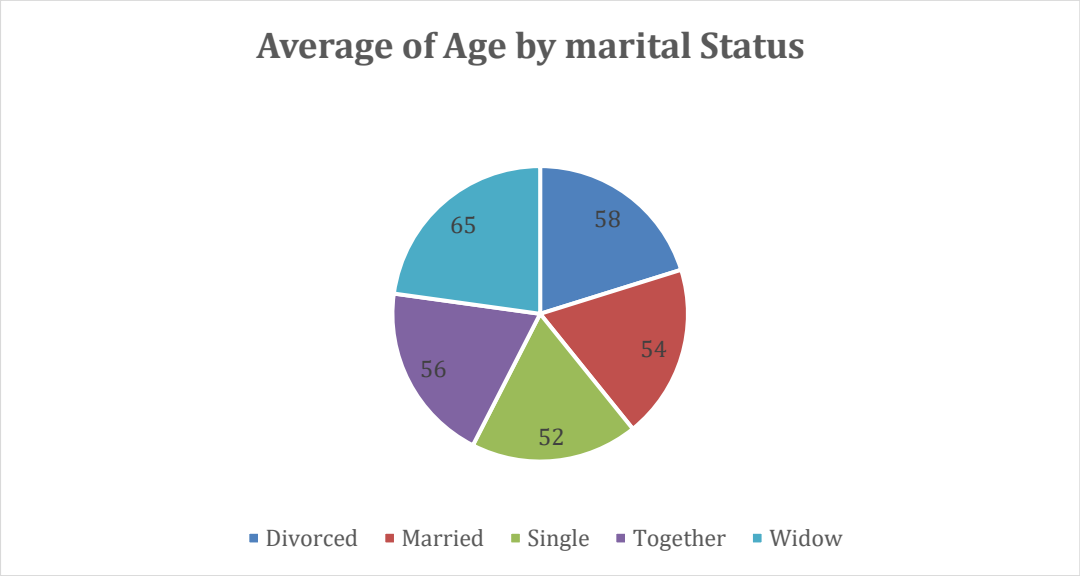
Appendix A1 – Average Age by Marital Status



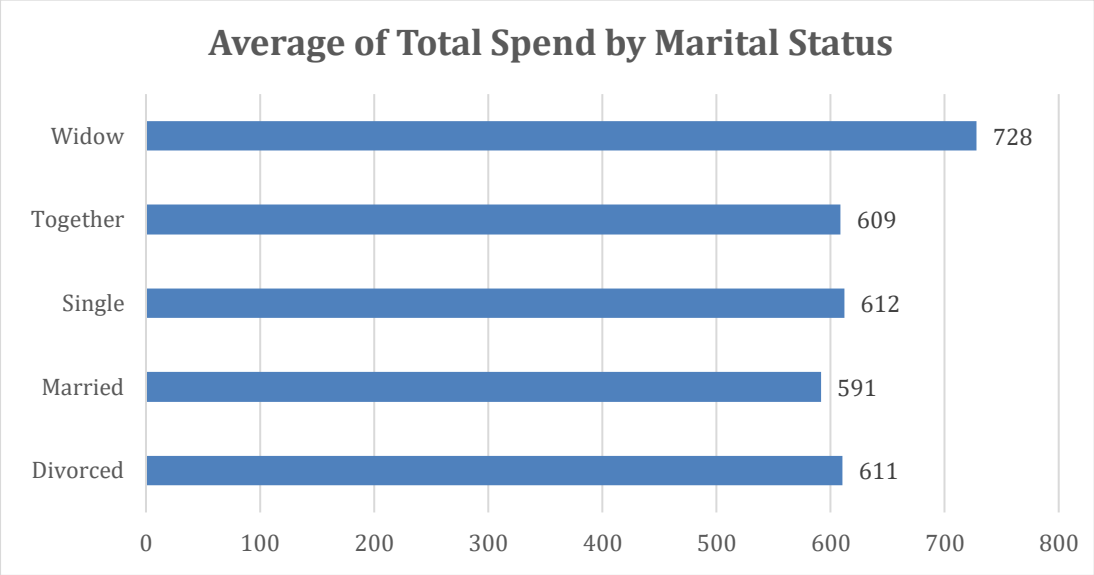
Appendix A2 – Average Spending by Marital Status



Appendix A.3 – Average of Age by Marital Status after joining "YOLO, Single and Alone" into a single category



Appendix A.4 – Average Spend by Marital Status after joining "YOLO, Single, and Alone" into a single category



Appendix B – SQL Queries

Appendix B.1– Country-Level Total Spend and Product Category Breakdown with Ranking

Query | Query History

```
58    WITH country_spend AS (
59     SELECT
60         Country,
61         SUM(AmtLiq) AS Liquor,
62         SUM(AmtVege) AS Vegetables,
63         SUM(AmtNonVeg) AS Meat,
64         SUM(AmtPes) AS Fish,
65         SUM(AmtChocolates) AS Chocolates,
66         SUM(AmtComm) AS Commodities,
67         SUM(AmtLiq + AmtVege + AmtNonVeg + AmtPes + AmtChocolates + AmtComm) AS Total_Spend
68     FROM marketing_data
69     GROUP BY Country
70 )
71 SELECT *,
72        RANK() OVER (ORDER BY Total_Spend DESC) AS Spend_Rank
73 FROM country_spend;
74
75
```

Data Output | Messages | Graph Visualiser | Notifications

Showing rows: 1 to 8 | Page No: 1 of 1

	country text	liquor bigint	vegetables bigint	meat bigint	fish bigint	chocolates bigint	commodities bigint	total_spend bigint	spend_rank bigint
1	SP	336392	28288	178409	40153	30134	46181	659557	1
2	SA	105918	8937	58398	13670	9019	15129	211071	2
3	CA	84066	7681	45925	9980	7607	12144	167403	3
4	AUS	42752	3689	22328	5546	4129	7132	85576	4
5	IND	36236	3788	23729	4818	3221	6014	77806	5
6	GER	36776	2980	20272	4601	2801	5768	73198	6
7	US	32214	3034	20185	4411	2863	4839	67546	7
8	ME	1729	8	817	226	122	220	3122	8

Total rows: 8 | Query complete 00:00:00.093 | LF | Ln 73, Col 20

Appendix B.2 - Marital Status-Level Spend and Product Category Analysis with Rank

```sql
WITH marital_spend AS (
    SELECT
        Marital_Status,
        SUM(AmtLiq) AS Liquor,
        SUM(AmtVege) AS Vegetables,
        SUM(AmtNonVeg) AS Meat,
        SUM(AmtPes) AS Fish,
        SUM(AmtChocolates) AS Chocolates,
        SUM(AmtComm) AS Commodities,
        SUM(AmtLiq + AmtVege + AmtNonVeg + AmtPes + AmtChocolates + AmtComm) AS Total_Spend
    FROM marketing_data
    GROUP BY Marital_Status
)
SELECT *,
       RANK() OVER (ORDER BY Total_Spend DESC) AS Spend_Rank
FROM marital_spend;
```

Data Output | Messages | Graph Visualiser | Notifications

Showing rows: 1 to 5 | Page No: 1 of 1

	marital_status text	liquor bigint	vegetables bigint	meat bigint	fish bigint	chocolates bigint	commodities bigint	total_spend bigint	spend_rank bigint
1	Married	256976	21981	137888	30395	22926	36719	506885	1
2	Together	176715	14612	95374	22383	15031	24754	348869	2
3	Single	139126	13027	87868	18704	12839	20970	292534	3
4	Divorced	75364	6363	34848	8130	6222	10739	141666	4
5	Widow	27902	2422	14085	3793	2878	4245	55325	5

Total rows: 5 | Query complete 00:00:00.094 | LF | Ln 130, Col 1

Appendix B.3 - Family Type and Most Popular Product Category Analysis

Query | Query History

```
147 WITH family_product_spend AS (
148     SELECT
149         CASE WHEN Kidhome + Teenhome = 0 THEN 'No Children/Teens' ELSE 'Has Children/Teens' END AS Family_T
150         'Liquor' AS Product,
151         SUM(AmtLiq) AS Total_Spend
152     FROM marketing_data
153     GROUP BY Family_Type
154
155     UNION ALL
156     SELECT
157         CASE WHEN Kidhome + Teenhome = 0 THEN 'No Children/Teens' ELSE 'Has Children/Teens' END AS Family_T
158         'Vegetables',
159         SUM(AmtVege)
160     FROM marketing_data
161     GROUP BY Family_Type
162
163     UNION ALL
```

Data Output | Messages | Graph Visualiser | Notifications

Showing rows: 1 to 2 | Page No: 1 of 1

	family_type text	most_popular_product text	total_spend bigint
1	Has Children/Teens	Liquor	367133
2	No Children/Teens	Liquor	308950

Total rows: 2 | Query complete 00:00:00.074 | LF | Ln 206, Col 1

Appendix B.4 - Country-Level Advertising Channel Conversion Analysis

```
209
210  -- Step 2: Total Lead Conversions by Social Media Platform per Country
211  SELECT
212      m.Country,
213      SUM(CAST(a.Twitter_ad AS INT)) AS Twitter_Leads,
214      SUM(CAST(a.Instagram_ad AS INT)) AS Instagram_Leads,
215      SUM(CAST(a.Facebook_ad AS INT)) AS Facebook_Leads,
216      SUM(CAST(a.Bulkmail_ad AS INT)) AS Bulkmail_Leads,
217      SUM(CAST(a.Brochure_ad AS INT)) AS Brochure_Leads
218  FROM marketing_data m
219  JOIN ad_data a ON m.ID = a.ID
220  GROUP BY m.Country
221  ORDER BY m.Country;
222  -- Step 3 (updated): Total Lead Conversions by Ad Channel per Marital Status
223  SELECT
224      m.Marital_Status,
225      SUM(CAST(a.Twitter_ad AS INT)) AS Twitter_Leads,
226      SUM(CAST(a.Instagram_ad AS INT)) AS Instagram_Leads
```

Data Output | Messages | Graph Visualiser | Notifications

Showing rows: 1 to 8 | Page No: 1 of 1

	country text	twitter_leads bigint	instagram_leads bigint	facebook_leads bigint	bulkmail_leads bigint	brochure_leads bigint
1	AUS	6	12	7	9	0
2	CA	24	21	18	18	6
3	GER	11	8	7	10	2
4	IND	10	6	7	13	2
5	ME	0	0	0	1	0
6	SA	20	21	20	21	4
7	SP	87	89	76	83	16
8	US	6	5	7	8	0

Total rows: 8 | Query complete 00:00:00.087 | LF | Ln 221, Col 20

Appendix B.5 - Ad Channel Conversion Analysis by Marital Status

```
221  ORDER BY m.Country;
222  -- Step 3 (updated): Total Lead Conversions by Ad Channel per Marital Status
223  SELECT
224      m.Marital_Status,
225      SUM(CAST(a.Twitter_ad AS INT)) AS Twitter_Leads,
226      SUM(CAST(a.Instagram_ad AS INT)) AS Instagram_Leads,
227      SUM(CAST(a.Facebook_ad AS INT)) AS Facebook_Leads,
228      SUM(CAST(a.Bulkmail_ad AS INT)) AS Bulkmail_Leads,
229      SUM(CAST(a.Brochure_ad AS INT)) AS Brochure_Leads
230  FROM marketing_data m
231  JOIN ad_data a ON m.ID = a.ID
232  GROUP BY m.Marital_Status
233  ORDER BY m.Marital_Status;
234
235  -- Step 4 (updated): Product Spend + All Ad Channels per Country
236  SELECT
237      m.Country,
```

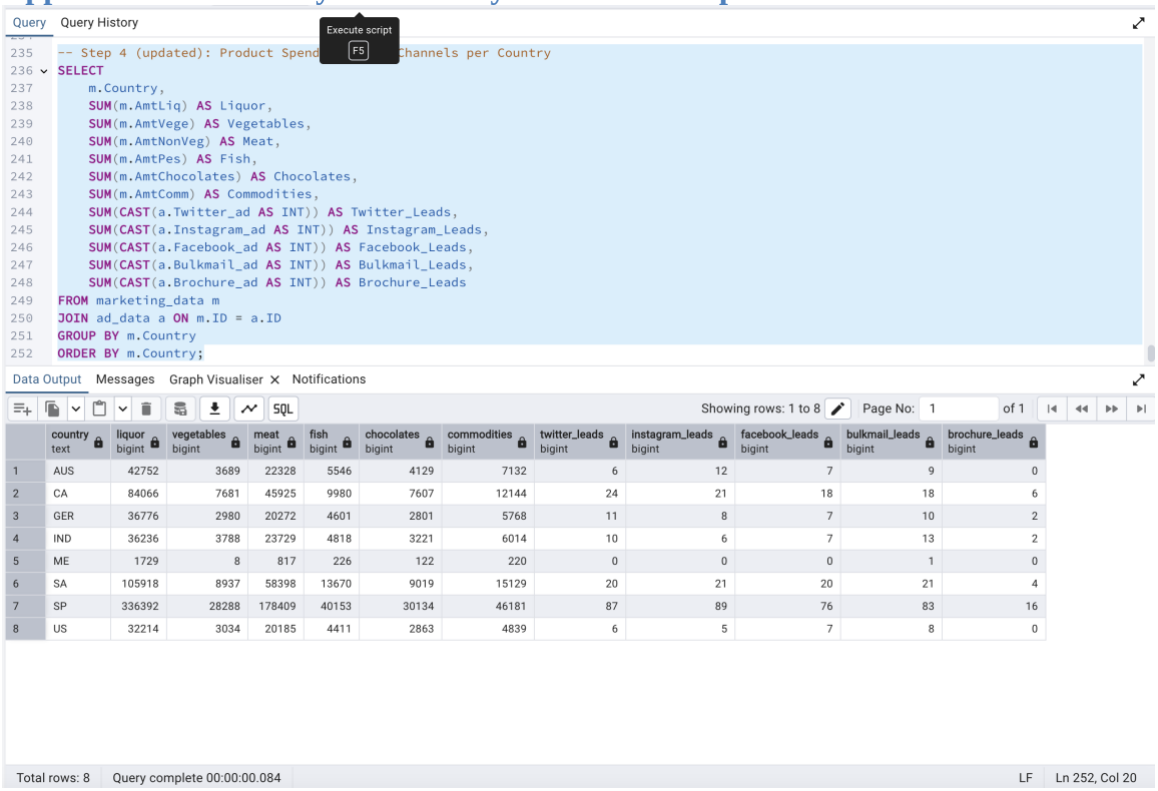
Data Output | Messages | Graph Visualiser | Notifications

Showing rows: 1 to 5 | Page No: 1 of 1

	marital_status text	twitter_leads bigint	instagram_leads bigint	facebook_leads bigint	bulkmail_leads bigint	brochure_leads bigint
1	Divorced	18	13	12	20	5
2	Married	62	66	62	63	7
3	Single	32	32	31	39	5
4	Together	42	44	32	37	12
5	Widow	10	7	5	4	1

Total rows: 5 | Query complete 00:00:00.090 | LF | Ln 233, Col 27

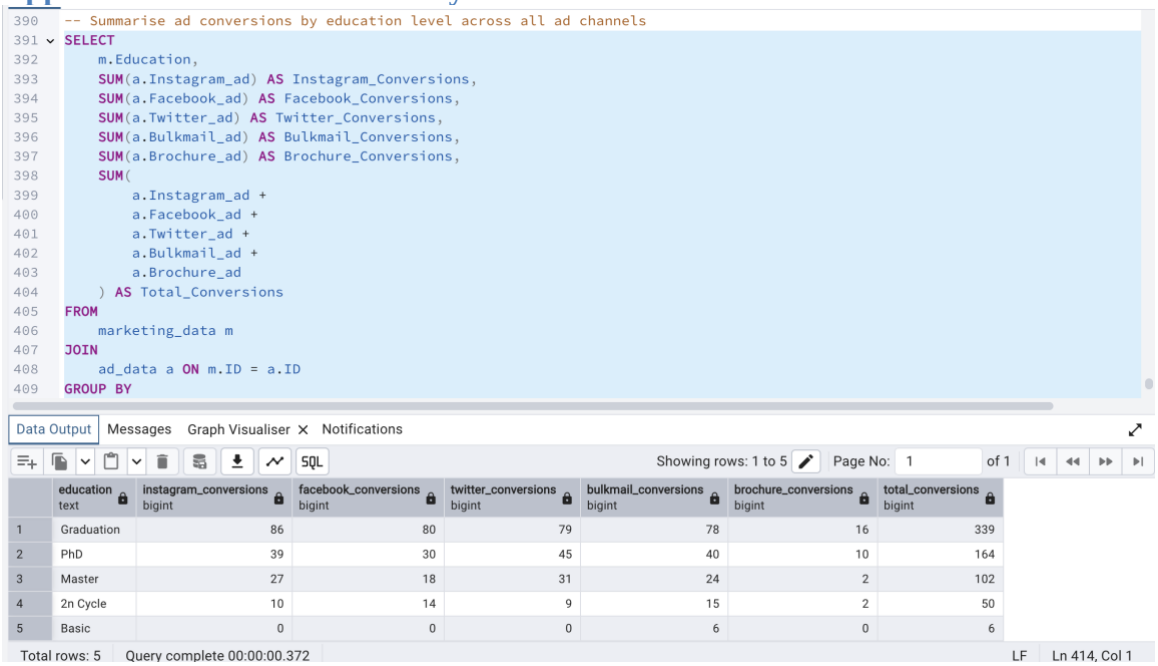
Appendix B.6 - Country-Level Analysis of Product Spend and Ad Channel Conversions



```
-- Step 4 (updated): Product Spend & Ad Channels per Country
SELECT
    m.Country,
    SUM(m.AmtLiq) AS Liquor,
    SUM(m.AmtVege) AS Vegetables,
    SUM(m.AmtNonVeg) AS Meat,
    SUM(m.AmtPes) AS Fish,
    SUM(m.AmtChocolates) AS Chocolates,
    SUM(m.AmtComm) AS Commodities,
    SUM(CAST(a.Twitter_ad AS INT)) AS Twitter_Leads,
    SUM(CAST(a.Instagram_ad AS INT)) AS Instagram_Leads,
    SUM(CAST(a.Facebook_ad AS INT)) AS Facebook_Leads,
    SUM(CAST(a.Bulkmail_ad AS INT)) AS Bulkmail_Leads,
    SUM(CAST(a.Brochure_ad AS INT)) AS Brochure_Leads
FROM marketing_data m
JOIN ad_data a ON m.ID = a.ID
GROUP BY m.Country
ORDER BY m.Country;
```

	country text	liquor bigint	vegetables bigint	meat bigint	fish bigint	chocolates bigint	commodities bigint	twitter_leads bigint	instagram_leads bigint	facebook_leads bigint	bulkmail_leads bigint	brochure_leads bigint
1	AUS	42752	3689	22328	5546	4129	7132	6	12	7	9	0
2	CA	84066	7681	45925	9980	7607	12144	24	21	18	18	6
3	GER	36776	2980	20272	4601	2801	5768	11	8	7	10	2
4	IND	36236	3788	23729	4818	3221	6014	10	6	7	13	2
5	ME	1729	8	817	226	122	220	0	0	0	1	0
6	SA	105918	8937	58398	13670	9019	15129	20	21	20	21	4
7	SP	336392	28288	178409	40153	30134	46181	87	89	76	83	16
8	US	32214	3034	20185	4411	2863	4839	6	5	7	8	0

Appendix B.8 – ad conversions by education level across all ad channels



```
-- Summarise ad conversions by education level across all ad channels
SELECT
    m.Education,
    SUM(a.Instagram_ad) AS Instagram_Conversions,
    SUM(a.Facebook_ad) AS Facebook_Conversions,
    SUM(a.Twitter_ad) AS Twitter_Conversions,
    SUM(a.Bulkmail_ad) AS Bulkmail_Conversions,
    SUM(a.Brochure_ad) AS Brochure_Conversions,
    SUM(
        a.Instagram_ad +
        a.Facebook_ad +
        a.Twitter_ad +
        a.Bulkmail_ad +
        a.Brochure_ad
    ) AS Total_Conversions
FROM
    marketing_data m
JOIN
    ad_data a ON m.ID = a.ID
GROUP BY
```

	education text	instagram_conversions bigint	facebook_conversions bigint	twitter_conversions bigint	bulkmail_conversions bigint	brochure_conversions bigint	total_conversions bigint
1	Graduation	86	80	79	78	16	339
2	PhD	39	30	45	40	10	164
3	Master	27	18	31	24	2	102
4	2n Cycle	10	14	9	15	2	50
5	Basic	0	0	0	6	0	6

Appendix B.10 - Top Advertising Channel by Country (Based on Leads)

Query | Query History

```
    CASE
        WHEN Country = 'AUS' THEN 'Australia'
        WHEN Country = 'CA' THEN 'Canada'
        WHEN Country = 'GER' THEN 'Germany'
        WHEN Country = 'IND' THEN 'India'
        WHEN Country = 'ME' THEN 'Montenegro'
        WHEN Country = 'SA' THEN 'South Africa'
        WHEN Country = 'SP' THEN 'Spain'
        WHEN Country = 'US' THEN 'United States'
        ELSE Country
    END AS Country_Name,
    Channel AS Top_Channel,
    Leads
FROM ranked_channels
WHERE channel_rank = 1
ORDER BY Country_Name;
```

Data Output | Messages | Graph Visualiser | Notifications

Showing rows: 1 to 8 | Page No: 1 of 1

	country_name text	top_channel text	leads bigint
1	Australia	Instagram	12
2	Canada	Twitter	24
3	Germany	Twitter	11
4	India	Bulkmail	13
5	Montenegro	Bulkmail	1
6	South Africa	Bulkmail	21
7	Spain	Instagram	89
8	United States	Bulkmail	8

Total rows: 8 | Query complete 00:00:00.081 | LF | Ln 390, Col 1

Appendix B.11 - average amount spent on each product category by households, segmented by: Number of kids or Teens

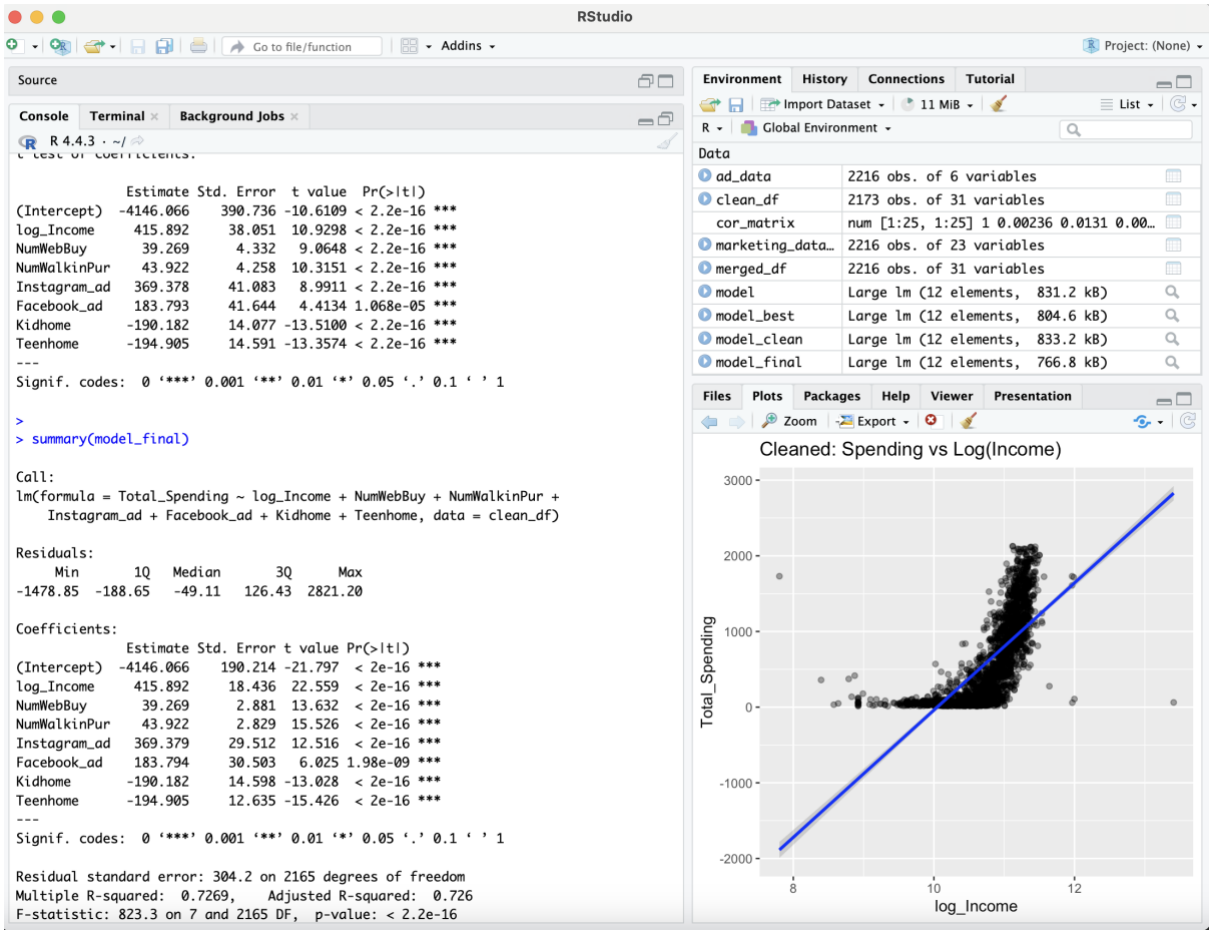
```
743
744 SELECT
745     Kidhome,
746     Teenhome,
747     CASE
748         WHEN Kidhome = 0 AND Teenhome = 0 THEN 'No Kids or Teens'
749         WHEN Kidhome > 0 AND Teenhome = 0 THEN 'Kids Only'
750         WHEN Kidhome = 0 AND Teenhome > 0 THEN 'Teens Only'
751         WHEN Kidhome > 0 AND Teenhome > 0 THEN 'Kids and Teens'
752     END AS family_type,
753     ROUND(AVG(AmtLiq), 1) AS avg_liquor,
754     ROUND(AVG(AmtVege), 1) AS avg_vegetables,
755     ROUND(AVG(AmtChocolates), 1) AS avg_chocolates,
756     ROUND(AVG(AmtComm), 1) AS avg_commercial,
757     ROUND(AVG(AmtPes), 1) AS avg_fish,
758     ROUND(AVG(AmtNonVeg), 1) AS avg_meat
759 FROM marketing_data
760 GROUP BY Kidhome, Teenhome
761 ORDER BY Kidhome, Teenhome;
762
```

Data Output | Messages | Graph Visualiser | Notifications

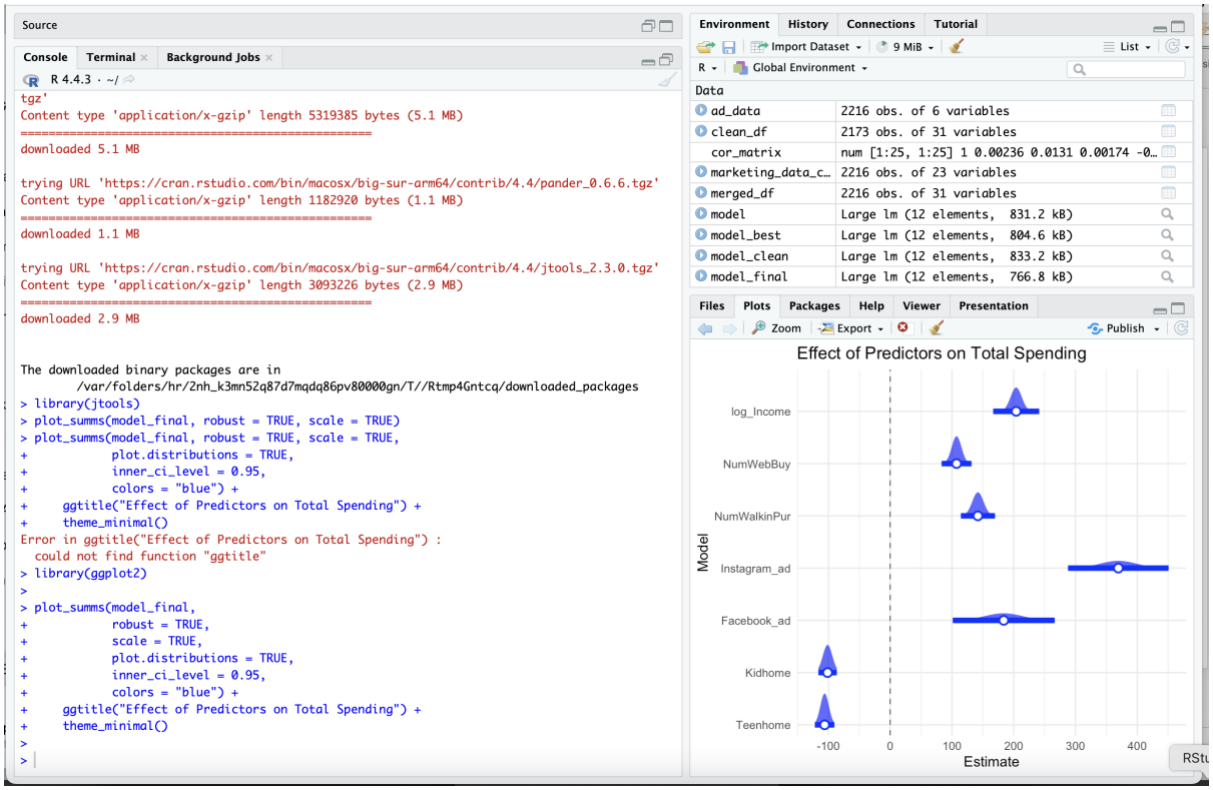
	kidhome integer	teenhome integer	family_type text	avg_liquor numeric	avg_vegetables numeric	avg_chocolates numeric	avg_commercial numeric	avg_fish numeric	avg_meat numeric
1	0	0	No Kids or Teens	488.1	52.3	53.2	64.2	76.6	370.9
2	0	1	Teens Only	417.7	27.2	28.8	55.9	36.6	139.3
3	0	2	Teens Only	409.6	20.7	19.1	57.0	33.7	133.5
4	1	0	Kids Only	82.4	9.9	9.4	21.3	14.6	49.2
5	1	1	Kids and Teens	124.1	6.5	7.4	22.8	9.4	45.5
6	1	2	Kids and Teens	276.0	12.9	10.1	28.9	8.6	110.1
7	2	0	Kids Only	61.2	14.6	8.1	28.0	13.4	42.1
8	2	1	Kids and Teens	78.3	1.0	1.4	10.3	3.0	23.1

Appendix C

Appendix C.1- Final Linear Regression Model with Robust Standard Errors



Appendix C.2 – Effect predictors on Total Spending


Source
Console
Terminal
Background Jobs
R 4.4.3 · ~/
tgz'
Content type 'application/x-gzip' length 5319385 bytes (5.1 MB)
==
downloaded 5.1 MB
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.4/pander_0.6.6.tgz'
Content type 'application/x-gzip' length 1182920 bytes (1.1 MB)
==
downloaded 1.1 MB
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.4/jtools_2.3.0.tgz'
Content type 'application/x-gzip' length 3093226 bytes (2.9 MB)
==
downloaded 2.9 MB
The downloaded binary packages are in
/var/folders/hr/2nh_k3mn52q87d7mqdq86pv80000gn/T//Rtmp4Gntcq/downloaded_packages
> library(jtools)
> plot_summs(model_final, robust = TRUE, scale = TRUE)
> plot_summs(model_final, robust = TRUE, scale = TRUE,
+ plot.distributions = TRUE,
+ inner_ci_level = 0.95,
+ colors = "blue") +
+ ggtitle("Effect of Predictors on Total Spending") +
+ theme_minimal()
Error in ggtitle("Effect of Predictors on Total Spending") :
could not find function "ggtitle"
> library(ggplot2)
>
> plot_summs(model_final,
+ robust = TRUE,
+ scale = TRUE,
+ plot.distributions = TRUE,
+ inner_ci_level = 0.95,
+ colors = "blue") +
+ ggtitle("Effect of Predictors on Total Spending") +
+ theme_minimal()
>
>
Environment
History
Connections
Tutorial
Import Dataset
9 MiB
List
R
Global Environment
Data
ad_data 2216 obs. of 6 variables
clean_df 2173 obs. of 31 variables
cor_matrix num [1:25, 1:25] 1 0.00236 0.0131 0.00174 -0…
marketing_data_c… 2216 obs. of 23 variables
merged_df 2216 obs. of 31 variables
model Large lm (12 elements, 831.2 kB)
model_best Large lm (12 elements, 804.6 kB)
model_clean Large lm (12 elements, 833.2 kB)
model_final Large lm (12 elements, 766.8 kB)
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Effect of Predictors on Total Spending
log_Income
NumWebBuy
NumWalkinPur
Instagram_ad
Facebook_ad
Kidhome
Teenhome
Model
-100
0
100
200
300
400
Estimate