Report

Statistical Analysis of Marketing Campaign in Omnichannel Retail

By: **Sina Tijani**

Table of Contents

List of Tables

List of Figures

Descriptive Statistics Measurements

Descriptive statistics simply summarizes and provides insights about a given data (Gandhi et al., 2021). When data tends to cluster around values like the mean, median, or mode, it's referred to as a measure of central tendency (Breslin, 2020). When data shows its spread, indicated by measures like standard deviation, variance, and quartiles, it's termed a measure of variability (Ruel, 2019).

The given dataset shows transactions of an omnichannel retail company. It will be used to further explain after cleaning.

Cleaning the Dataset

The dataset had only a few dirt to clean. Here is a summary of the cleaning process:

- Pandas, Numpy, and Matplotlib were libraries used to clean the dataset.
- Excluding the "ID" column, 201 perfect duplicates were found and dropped representing 8.97% of the dataset.
- 'Dt_customer' datatype was changed to datetime64[ns].
- 9 outliers in "Income" were found and dropped using the interquartile range representing 0.44% of the cleaned dataset.
- After dropping the outliers, 24 missing incomes were filled with the income mean.

Assumptions

1. The company is an omnichannel retail since the dataset shows both in-store and online activity.
2. The current operating year is 2014 as that is the last date a customer joined the company.

Table 1. Last Date a customer joined.

```
omni3[['dt_customer']].sort_values(by='dt_customer', ascending=False).head()
✓ 0.0s
```

	dt_customer
198	2014-06-29
954	2014-06-29
771	2014-06-28
45	2014-06-28
1666	2014-06-28

3. A business analyst working at the company will interpret the descriptive statistics to the managers.

Measuring Central Tendency

Mean:

The table below shows the business has 2,029 customers. Born in 1969 with the operating year of 2014, the customers are 45years on average. They have 1 teenager at home and they earn a yearly income of £51,735.

Table 2. Descriptive Statistics of the Dataset

	id	year_birth	income	kidhome	teenhome	dt_customer	recency	mntdrinks	mntfruits	mntmeatproducts
count	2029.0	2029.0	2029.0	2029.0	2029.0	2029	2029.0	2029.0	2029.0	2029.0
mean	5590.0	1969.0	51735.0	0.0	1.0	2013-07-12 02:42:31.404632832	49.0	305.0	26.0	165.0
min	0.0	1893.0	3502.0	0.0	0.0	2012-07-30 00:00:00	0.0	0.0	0.0	0.0
25%	2802.0	1959.0	35701.0	0.0	0.0	2013-01-17 00:00:00	24.0	24.0	2.0	16.0
50%	5510.0	1970.0	51537.0	0.0	0.0	2013-07-13 00:00:00	49.0	178.0	8.0	68.0
75%	8430.0	1977.0	68118.0	1.0	1.0	2014-01-01 00:00:00	74.0	505.0	33.0	228.0
max	11191.0	1996.0	113734.0	2.0	2.0	2014-06-29 00:00:00	99.0	1493.0	199.0	1607.0
std	3259.0	12.0	20551.0	1.0	1.0	NaN	29.0	336.0	40.0	218.0

Median:

When the customers are sorted by their income in ascending order, the table shows, half of the customers earn more than £51,537 shown as 50% percentile (Petrelli, 2021). They spend more than £178 on drinks, £8 on fruits and £68 on meat products.

Mode:

For each column, the frequent occurring values are the mode (Breslin, 2020). The tables below show most of the customers have graduated and have married. The company recorded the highest single-day signup on 2014-05-12. Most customers buy just 1 deal, purchase twice on the web and but do not buy anything in the company's mobile app.

Table 3. Mode 1

```
omni3[['year_birth', 'education', 'marital_status', 'income', 'kidhome',
        'teenhome', 'dt_customer', 'recency', 'mntdrinks', 'mntfruits', 'mntmeatproducts']].mode()
✓ 0.0s
```

	year_birth	education	marital_status	income	kidhome	teenhome	dt_customer	recency	mntdrinks	mntfruits	mntmeatproducts
0	1976	Graduation	Married	51537.0	0	0	2014-05-12	56	2	0	5

Table 4. Mode 2

```
omni3[['mntfishproducts', 'mntsweetproducts',
        'numdealspurchases', 'numwebpurchases', 'numapppurchases',
        'numstorepurchases', 'numwebvisitsmonth']].mode()
✓ 0.0s
```

	mntfishproducts	mntsweetproducts	numdealspurchases	numwebpurchases	numapppurchases	numstorepurchases	numwebvisitsmonth
0	0	0	1	2	0	3	7

Quartile:

This divides the data into four parts (Ruel, 2019). In Table 2, the 25% (first quartile or lower boundary) shows a quarter of the customers earn below £35,701. It has been 24 days since a quarter of them came interacted with the company shown under recency.

The 75% (third quartile or upper boundary) shows only a quarter of the customers spend above £50 on fish and £34 on sweets.

Table 5. Descriptive Statistics 2

```
omni3[['mntfishproducts', 'mntsweetproducts',
       'numdealspurchases', 'numwebpurchases', 'numapppurchases',
       'numstorepurchases', 'numwebvisitsmonth']].describe()
```

✓ 0.0s

	mntfishproducts	mntsweetproducts	numdealspurchases	numwebpurchases	numapppurchases	numstorepurchases	numwebvisitsmonth
count	2029.000000	2029.000000	2029.000000	2029.000000	2029.000000	2029.000000	2029.000000
mean	37.623460	27.344012	2.309019	4.116806	2.619024	5.797930	5.321833
std	54.786696	41.764201	1.851387	2.793988	2.740458	3.223676	2.403445
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3.000000	1.000000	1.000000	2.000000	0.000000	3.000000	3.000000
50%	12.000000	8.000000	2.000000	4.000000	2.000000	5.000000	6.000000
75%	50.000000	34.000000	3.000000	6.000000	4.000000	8.000000	7.000000
max	259.000000	263.000000	15.000000	27.000000	11.000000	13.000000	20.000000

Standard Deviation:

This measures the differences between the customers income from the average of £51,537 (Petrelli, 2021). The higher the standard deviation, the higher the difference. Table 1 shows income standard deviation of £20,551. That is huge. It means many customers earn way less and way higher than the mean income.

Hypothesis Testing

In this session,

H0 refers to the Null hypothesis, and

H1 is the Alternative hypothesis.

5% = significance level

In the following, business questions are asked with their hypothesis formulated, then the right statistical technique for it is defined and then used to conduct the analysis.

X_2 (Chi-square)

Question:

Does the educational level of customers influence their marital status?

H0: There is no link between Education and Marital Status ($\mu_1 = \mu_2 = \ldots = \mu_{k.}$)

H1: There is a significant link between Education and Marital Status (at least one μ is not equal.)

When association (link) between two categorical variables are of interest to be analyzed, X_2 (Chi-square) is used (Ji et al., 2020). The business question can be answered by using Education and Marital Status, two categorical variables in the dataset.

Table 6. Initial Chi-Square Test

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	23.480[a]	28	.709
Likelihood Ratio	23.967	28	.683
N of Valid Cases	2029		

a. 16 cells (40.0%) have expected count less than 5. The minimum expected count is .02.

40% expected count of less than 5 affects the test's reliability (Turhan, 2020). To address this, we grouped Alone, Absurd, YOLO into "Other" category to test again.

Before:

Table 7. Marital_Status value counts(vc)

```
omni3.marital_status.value_counts()
✓ 0.0s
marital_status
Married     784
Together    513
Single      445
Divorced    211
Widow        70
Alone         3
Absurd        2
YOLO          1
Name: count, dtype: int64
```

After:

Table 8. Marital_Status_grouped vc

```
# Grouping the 3 categories
group_three = ['Alone', 'Absurd', 'YOLO']


# Creating a new column
omni3['marital_status_grouped'] = omni3['marital_status'].replace(group_three, 'Other')


# Showing the value counts of the new groups
omni3.marital_status_grouped.value_counts()
✓ 0.0s
marital_status_grouped
Married     784
Together    513
Single      445
Divorced    211
Widow        70
Other         6
Name: count, dtype: int64
```

Results:

Table 9. Crosstabulation

education * marital_status_grouped Crosstabulation

			marital_status_grouped						
			Divorced	Married	Other	Single	Together	Widow	Total
education	2n Cycle	Count	22	69	0	37	53	5	186
		Expected Count	19.3	71.9	.6	40.8	47.0	6.4	186.0
	Basic	Count	1	18	0	18	11	1	49
		Expected Count	5.1	18.9	.1	10.7	12.4	1.7	49.0
	Graduation	Count	110	392	2	233	252	31	1020
		Expected Count	106.1	394.1	3.0	223.7	257.9	35.2	1020.0
	Master	Count	33	130	2	66	93	12	336
		Expected Count	34.9	129.8	1.0	73.7	85.0	11.6	336.0
	PhD	Count	45	175	2	91	104	21	438
		Expected Count	45.5	169.2	1.3	96.1	110.7	15.1	438.0
Total		Count	211	784	6	445	513	70	2029
		Expected Count	211.0	784.0	6.0	445.0	513.0	70.0	2029.0

Table 9. Cramer's V

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.097	.518
	Cramer's V	.048	.518
N of Valid Cases		2029	

Table 11. New Chi-Square test

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	19.054[a]	20	.518
Likelihood Ratio	20.148	20	.449
N of Valid Cases	2029		

a. 6 cells (20.0%) have expected count less than 5. The minimum expected count is .14.

Expected count less than 5 (20%) is within an acceptable range.

The analysis produced a p-value of 0.518 > 0.05 significance. Additionally, a Cramer's V value of 0.048 < 1 indicates a very weak association between Marital Status and Education.

- **Business Implication:** Educational levels do not significantly impact the distribution of marital statuses among customers. Tailoring marketing strategies based solely on education may not yield substantial variations in marital status distributions.

⇒ H0 Accepted
⇒ H1 Insufficient evidence to support it.

One-way ANOVA Test (including post hoc analysis)

Question:

Does the marital status of the customers have an impact on their engagement (recency) with the company?

H0: Customer engagement is the same regardless of their marital status ($\mu_1 = \mu_2 = \ldots = \mu_{k.}$)

H1: Customer engagement differs based on their marital status (at least one μ is not equal.)

One-way ANOVA which allows to compare the averages of three or more variables to know their significance will be used for this test (Gurvich & Naumova, 2021).

Table 10. ANOVA Effect_sizes

ANOVA Effect Sizes[a,b]

		Point Estimate	95% Confidence Interval Lower	Upper
recency	Eta-squared	.002	.000	.005
	Epsilon-squared	-.001	-.002	.002
	Omega-squared Fixed-effect	-.001	-.002	.002
	Omega-squared Random-effect	.000	.000	.000

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

b. Negative but less biased estimates are retained, not rounded to zero.

The effect sizes for "recency" in the ANOVA are small, implying minimal impact on the variability in income. The estimates, though negative, are not statistically significant.

Table 1311. One-way ANOVA test

ANOVA

recency

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3093.851	5	618.770	.737	.596
Within Groups	1699603.169	2023	840.140		
Total	1702697.020	2028			

The test yielded a non-significant result: [p-value = 0.596 > significance level 0.05].

This suggests there is no statistically significant difference in customer engagement means among various marital statuses.

- **Business Implication:** Marital status, as a standalone factor, may not significantly influence how frequently customers interact with the company. Therefore:

⇒ H0 accepted

⇒ H1 insufficient evidence to accept.

While there is no need for post-hoc analysis in this case, a look at the table below confirms its significance: p ranges from 0.705 to 1.0.

Table 1412. Post-hoc analysis

Multiple Comparisons

Dependent Variable: recency
Tukey HSD

(I) maritalstatus_nominal_score	(J) maritalstatus_nominal_score	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Single	Together	-1.075	1.878	.993	-6.43	4.28
	Married	1.165	1.720	.984	-3.74	6.07
	Divorced	.059	2.423	1.000	-6.85	6.97
	Widow	.648	3.727	1.000	-9.98	11.28
	Other	16.015	11.913	.760	-17.97	50.00
Together	Single	1.075	1.878	.993	-4.28	6.43
	Married	2.239	1.646	.751	-2.46	6.93
	Divorced	1.134	2.371	.997	-5.63	7.90
	Widow	1.723	3.693	.997	-8.81	12.26
	Other	17.090	11.902	.705	-16.86	51.04
Married	Single	-1.165	1.720	.984	-6.07	3.74
	Together	-2.239	1.646	.751	-6.93	2.46
	Divorced	-1.105	2.248	.996	-7.52	5.31
	Widow	-.516	3.616	1.000	-10.83	9.80
	Other	14.850	11.878	.812	-19.03	48.73
Divorced	Single	-.059	2.423	1.000	-6.97	6.85
	Together	-1.134	2.371	.997	-7.90	5.63
	Married	1.105	2.248	.996	-5.31	7.52
	Widow	.589	3.998	1.000	-10.81	11.99
	Other	15.956	12.000	.769	-18.27	50.19
Widow	Single	-.648	3.727	1.000	-11.28	9.98
	Together	-1.723	3.693	.997	-12.26	8.81
	Married	.516	3.616	1.000	-9.80	10.83
	Divorced	-.589	3.998	1.000	-11.99	10.81
	Other	15.367	12.330	.814	-19.80	50.54
Other	Single	-16.015	11.913	.760	-50.00	17.97
	Together	-17.090	11.902	.705	-51.04	16.86
	Married	-14.850	11.878	.812	-48.73	19.03
	Divorced	-15.956	12.000	.769	-50.19	18.27
	Widow	-15.367	12.330	.814	-50.54	19.80

Multiple Linear Regression Analysis (including multicollinearity and VIF analysis)

Question:

How do various customer behaviours and engagement metrics contribute to predicting their incomes?

H0: The income of customers is not influenced by various behaviours and engagement metrics.

H1: There is a significant influence of at least one behaviour or engagement metric on customers' income.

Multiple Linear Regression (MLR), as a statistical method, will be used here. It explores the association between a variable of interest (dependent) and two or more other independent variables/features (Ngige et al., 2023). From the business question, income becomes our dependent variable and behaviours such as in-app or web purchases, recency, and amount spent on drinks, fruits, and others become our independent variable. SPSS is used for the analysis. Potential issues like multicollinearity will be studied.

Table 13. Multiple Linear Regression(MLR)

```
Residuals:
   Min     1Q Median     3Q    Max
-75610  -5773    387   5657  37375

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        43428.7338  1021.4681  42.516  < 2e-16 ***
kidhome             2707.7081   513.0804   5.277 1.45e-07 ***
teenhome            6812.5013   438.1393  15.549  < 2e-16 ***
recency              -12.2807     7.0055  -1.753   0.0798 .
mntdrinks             17.3788     0.9936  17.490  < 2e-16 ***
mntfruits             17.8586     7.0019   2.551   0.0108 *
mntmeatproducts       18.7249     1.5322  12.221  < 2e-16 ***
mntfishproducts       10.6738     5.3190   2.007   0.0449 *
mntsweetproducts      31.4029     6.6224   4.742 2.27e-06 ***
numdealspurchases -1006.7921   141.0204  -7.139 1.30e-12 ***
numwebpurchases    1007.9574    96.9631  10.395  < 2e-16 ***
numapppurchases     674.2308   129.0910   5.223 1.94e-07 ***
numstorepurchases   759.7973    95.2943   7.973 2.57e-15 ***
numwebvisitsmonth -2608.8807   120.0055 -21.740  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9115 on 2015 degrees of freedom
Multiple R-squared:  0.8045,    Adjusted R-squared:  0.8033
F-statistic: 637.9 on 13 and 2015 DF,  p-value: < 2.2e-16
```

Table 14. Stepwise regression

```
Residuals:
   Min     1Q Median     3Q    Max
-75610  -5773    387   5657  37375

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        43428.7338  1021.4681  42.516  < 2e-16 ***
kidhome             2707.7081   513.0804   5.277 1.45e-07 ***
teenhome            6812.5013   438.1393  15.549  < 2e-16 ***
recency              -12.2807     7.0055  -1.753   0.0798 .
mntdrinks             17.3788     0.9936  17.490  < 2e-16 ***
mntfruits             17.8586     7.0019   2.551   0.0108 *
mntmeatproducts       18.7249     1.5322  12.221  < 2e-16 ***
mntfishproducts       10.6738     5.3190   2.007   0.0449 *
mntsweetproducts      31.4029     6.6224   4.742 2.27e-06 ***
numdealspurchases -1006.7921   141.0204  -7.139 1.30e-12 ***
numwebpurchases    1007.9574    96.9631  10.395  < 2e-16 ***
numapppurchases     674.2308   129.0910   5.223 1.94e-07 ***
numstorepurchases   759.7973    95.2943   7.973 2.57e-15 ***
numwebvisitsmonth -2608.8807   120.0055 -21.740  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9115 on 2015 degrees of freedom
Multiple R-squared:  0.8045,    Adjusted R-squared:  0.8033
F-statistic: 637.9 on 13 and 2015 DF,  p-value: < 2.2e-16
```

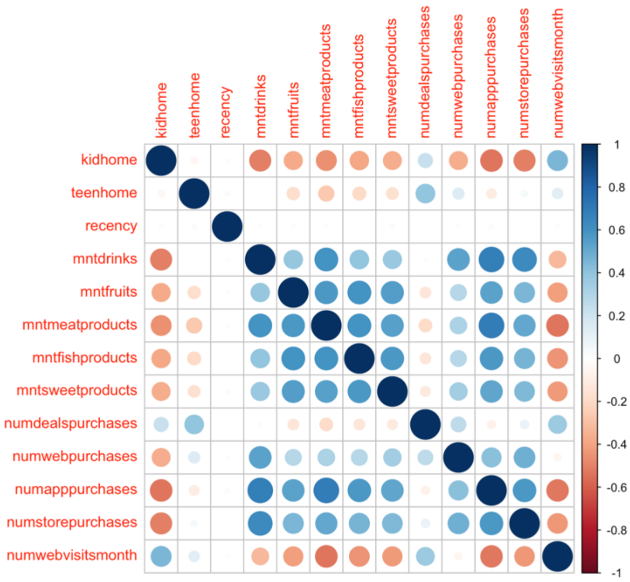
Only “recency” p-value=0.0798 > 0.05. Adjusted R-squared(0.8033) is used instead of Multiple R-squared and it gauges model fit representing the explained variance which is that 80.33% of variance in income can be explained from the independent variables (predictors) (Rasyidah et al., 2023).

Ultra-low p-value (< 2.2e-16) signifies high statistical significance of the model. Key predictors (kidhome, teenhome, mntdrinks, mntmeatproducts, numwebpurchases, etc.) are highly significant ($p < 0.05$).

Therefore

- ⇒ H0 rejected due to sufficient evidence.
- ⇒ H1 accepted due to ultra-high significance.

Figure 1. MLR Correlation Heatmap



A heatmap is a visual representation of data using colours to show the intensity of values in a matrix, making patterns and trends easily discernible (Mambang et al., 2022).

From the map, kidhome and mntdrinks are negatively correlated (-0.505). Mntdrinks and numwebpurchases are positively correlated (0.536)

Mntdrinks, mntmeatproducts, numwebpurchases, and numwebvisitsmonth exhibit significant correlations.

High correlations may signal multicollinearity impacting the regression reliability. Therefore, VIF is to be checked (Valerio-Hernández et al., 2023).

VIF (Variance Inflation Factor) | Multicollinearity

In a multiple linear regression model, some variables may be highly correlated (multicollinearity) which can affect the model's reliability. The measure to check for this is VIF (Cheng et al., 2022). To interpret,

- There is no significant correlation when the VIF gotten is 1.
- if it is bigger than 1 but smaller than 5, it is moderately correlated.
- And if it is bigger than 5, it indicates potential multicollinearity issues.

Table 1515. VIF

```
          kidhome          teenhome           recency          mntdrinks          mntfruits   mntmeatproducts
         1.860135          1.402542          1.005682           2.721450           1.903894          2.733605
  mntfishproducts mntsweetproducts numdealspurchases    numwebpurchases   numapppurchases numstorepurchases
         2.072639          1.867033          1.663686           1.791326           3.054575          2.303291
numwebvisitsmonth
         2.030408
```

The VIF values, all below 5, indicate no significant multicollinearity concerns among the predictor variables.

Conclusion

In omnichannel retail where in-store and online customer engagement is pivotal, the insights derived from rigorous statistical analyses play a crucial role in shaping effective marketing campaigns. The various analyses conducted, including Chi-square, One-way ANOVA, and Multiple Linear Regression, offer profound insights into customer behaviours, preferences, and spending patterns.

Understanding the nuanced relationship between variables such as education, marital status, and income allows marketers to tailor campaigns with precision. The absence of a significant association between marital status and customer engagement, for instance, prompts a strategic shift in focusing on other influential factors.

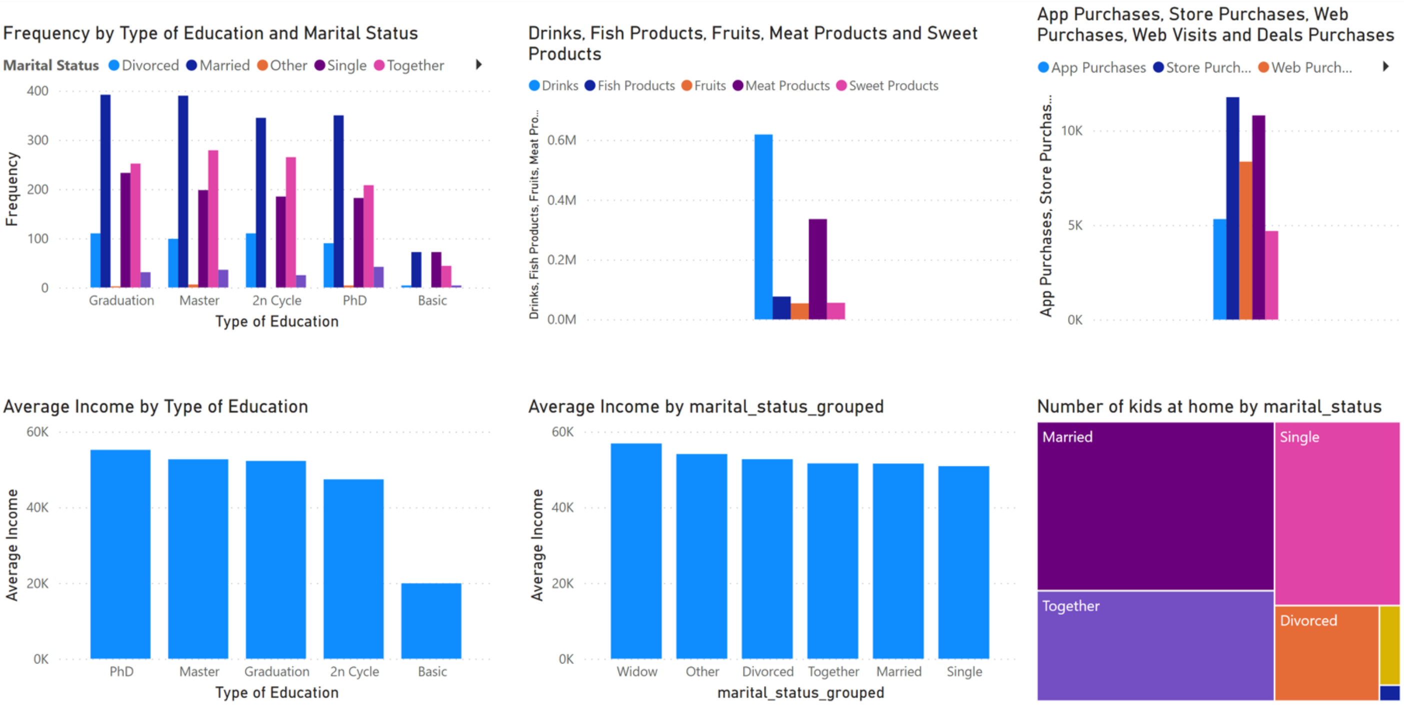
The Multiple Linear Regression analysis, with its comprehensive exploration of various customer behaviour and engagement metrics, becomes a compass for marketers. Identifying statistically significant predictors of income empowers campaigns to be finely tuned, ensuring resources are allocated where they are most impactful.

Moreover, the examination of correlations and the mitigation of multicollinearity concerns ensure that marketing decisions are founded on robust insights. For instance, recognizing the correlation between web and app purchases aids in orchestrating integrated campaigns across these channels.

As omnichannel retail businesses strive for synergy across platforms, the derived insights become the linchpin for creating cohesive and personalized marketing strategies. These statistical analyses, meticulously applied to real-world data, transcend mere numbers they pave the way for marketing campaigns that resonate with customers, fostering loyalty and driving business success. The fusion of data-driven decision-making and marketing prowess is the hallmark of a modern, effective omnichannel retail strategy.

Some visualizations from the dataset

Figure 2. Visualizations from the dataset



References

Al Adwan, A., Kokash, H., Al Adwan, R., & Khattak, A. (2023). Data analytics in digital marketing for tracking the effectiveness of campaigns and inform strategy. *International Journal of Data and Network Science*, *7*(2). https://doi.org/10.5267/j.ijdns.2023.3.015

Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and E-Business Management*. https://doi.org/10.1007/s10257-023-00640-4

Breslin, A. M. B. (2020). Descriptive Statistics. *SAGE Research Methods Foundations*. https://doi.org/10.4135/9781526421036917134

Cheng, J., Sun, J., Yao, K., Xu, M., & Cao, Y. (2022). A variable selection method based on mutual information and variance inflation factor. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *268*, 120652. https://doi.org/10.1016/J.SAA.2021.120652

Dogan, O., Hiziroglu, A., & Seymen, O. F. (2021). Segmentation of Retail Consumers with Soft Clustering Approach. *Advances in Intelligent Systems and Computing*, *1197 AISC*. https://doi.org/10.1007/978-3-030-51156-2_6

Frasquet, M., Ieva, M., & Ziliani, C. (2021). Online channel adoption in supermarket retailing. *Journal of Retailing and Consumer Services*, *59*. https://doi.org/10.1016/j.jretconser.2020.102374

Gandhi, P., Bhatia, S., & Dev, K. (2021). Data driven decision making using analytics. *Data Driven Decision Making Using Analytics*, 1–138. https://doi.org/10.1201/9781003199403

Gurvich, V., & Naumova, M. (2021). Logical contradictions in the one-way Anova and Tukey-Kramer multiple comparisons tests with more than two groups of observations. *Symmetry*, *13*(8). https://doi.org/10.3390/sym13081387

Ji, X., Gu, W., Qian, X., Wei, H., & Zhang, C. (2020). Combined Neyman–Pearson chi-square: An improved approximation to the Poisson-likelihood chi-square. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *961*, 163677. https://doi.org/10.1016/J.NIMA.2020.163677

Mambang, M., Hidayat, A., Wahyudi, J., & Marleny, F. D. (2022). Explanatory Data Analysis to Evaluate Keyword Searches for Educational Videos on YouTube with a Machine Learning Approach. *SinkrOn*, *7*(3). https://doi.org/10.33395/sinkron.v7i3.11502

Ngige, G. A., Ovuoraye, P. E., Igwegbe, C. A., Fetahi, E., Okeke, J. A., Yakubu, A. D., & Onyechi, P. C. (2023). RSM optimization and yield prediction for biodiesel produced from alkali-catalytic transesterification of pawpaw seed extract: Thermodynamics, kinetics, and Multiple Linear Regression analysis. *Digital Chemical Engineering*, *6*. https://doi.org/10.1016/j.dche.2022.100066

Paulo, M., Miguéis, V. L., & Pereira, I. (2022). Leveraging email marketing: Using the subject line to anticipate the open rate. *Expert Systems with Applications*, *207*. https://doi.org/10.1016/j.eswa.2022.117974

Petrelli, M. (2021). *Introduction to Python in Earth Science Data Analysis*. https://doi.org/10.1007/978-3-030-78055-5

Rasyidah, Efendi, R., Nawi, N. M., Deris, M. M., & Burney, S. M. A. (2023). Cleansing of inconsistent sample in linear regression model based on rough sets theory. *Systems and Soft Computing*, *5*. https://doi.org/10.1016/j.sasc.2022.200046

Ruel, E. (2019). 100 Questions (and Answers) About Survey Research. *100 Questions (and Answers) About Survey Research*. https://doi.org/10.4135/9781506348803

Tijani, S., & Microsoft Bing Image Creator. (2023). *A minimalist supermarket with drinks, meat, fish, and fruit*. https://www.bing.com/images/create/a-minimalist-supermarket-with-drinks2c-meat2c-fish2c-/1-6557e15dd39c40edb301549d839f56c5?id=R1sPaEaq%2fWsmoqhrME3Icg%3d%3d&view=detailv2&idpp=genimg&idpclose=1&FORM=SYDBIC

Turhan, N. S. (2020). Karl Pearson's Chi-Square Tests. *Educational Research and Reviews*, *16*(9), 575–580. https://doi.org/10.5897/ERR2019.3817

Valerio-Hernández, J. E., Pérez-Rodríguez, P., & Ruíz-Flores, A. (2023). Quantile regression for prediction of complex traits in Braunvieh cattle using SNP markers and pedigree. *Revista Mexicana De Ciencias Pecuarias*, *14*(1). https://doi.org/10.22319/rmcp.v14i1.6182

Appendix

Data cleaning process 1

Code | Markdown | Run All | Restart | Clear All Outputs | Variables | Outline | ··· | Python 3.12.0

```python
# by_sina_tijani
# explaining the cleaning process by displaying them all in a single cell
# the codes were originally written and run in different cells to check the progress.

# my libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# reading the data set directly as the working directory has already been set
omni = pd.read_csv("CW2(2324SepJan)_MarketingCampaignData.csv")

# formatting the columns To lowercase
omni.columns = omni.columns.str.lower()

# Check if the column heads were changed
omni.columns

# Checking for perfect duplicates
omni.duplicated().any()

# isolating the id column to check for duplicates across the other 18 columns
duplicate_check = ['year_birth', 'education', 'marital_status', 'income', 'kidhome',
       'teenhome', 'dt_customer', 'recency', 'mntdrinks', 'mntfruits',
       'mntmeatproducts', 'mntfishproducts', 'mntsweetproducts',
       'numdealspurchases', 'numwebpurchases', 'numapppurchases',
       'numstorepurchases', 'numwebvisitsmonth']

# Checking if there are any duplicates
omni.duplicated(subset=duplicate_check).any()

# printing the number and percent of duplicates
print(f'The dataset has {omni.duplicated(subset=duplicate_check).sum()} near perfect duplicates')
duplicates_percent = len(omni[omni.duplicated(subset=duplicate_check)]) / len(omni) * 100
print(f"And that represents {duplicates_percent:.2f}% of the dataset.")

# Check all the rows with their duplicates
omni[omni.duplicated(subset=duplicate_check, keep=False)].sort_values(by=['year_birth', 'income'], ascending=True)

# Dropping the duplicates while resetting the index
omni2 = omni.drop_duplicates(subset=duplicate_check).reset_index(drop=True)

# changing the data type for dt_customer
omni2.dt_customer = pd.to_datetime(omni2.dt_customer)

# checking if the data type was changed correctly
omni2.info()

# checking for total null values
omni2.isnull().sum()

# finding rows with missing income
missing_income = omni2[omni2.income.isnull()]
missing_income

# filling the missing income with the median of the incomes. and checking if they were filled
```

Ln 11, Col 78 LF

Data cleaning process 2

Code | Markdown | Run All | Restart | Clear All Outputs | Variables | Outline | Python 3.12.0

```python
# filling the missing income with the median of the incomes. and checking if they were filled
omni2.income.fillna(omni2.income.median(), inplace=True)
omni2.isnull().sum()

# finding descriptive statistics to check for outliers presence in the income column
omni2.income.describe().round()

# visualizing the outliers in income column with a boxplot
omni2[['income']].boxplot()
plt.show()

# using interquartile range to find and isolate the outliers
Q1 = omni2.income.quantile(0.25)
Q3 = omni2.income.quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1 * IQR
upper_bound = Q3 + 1.5 * IQR

iqr_outliers = omni2[(omni2.income < lower_bound) | (omni2.income > upper_bound)]
iqr_outliers.sort_values(by='income', ascending=True)

# printing the minimum and maximum income based on the IQR calculations
# and printing the count and percentage
print(f'The minimum income of IQR outliers {lower_bound}')
print(f'The maximum income of IQR outliers {upper_bound}\n')

print(f"The number of IQR outliers are {iqr_outliers.income.count()}.")
outlier_percent = len(iqr_outliers) / len(omni2) * 100
print(f"The outliers represents {outlier_percent:.2f}% of the dataset.")

# dropping the outliers from the dataset
omni3 = omni2.drop(iqr_outliers.index)
omni3

# finding descriptive statistics of the cleaned dataset
omni3.describe().round()

# finding mode across 1st part the dataset
omni3[['year_birth', 'education', 'marital_status', 'income', 'kidhome',
       'teenhome', 'dt_customer', 'recency', 'mntdrinks', 'mntfruits', 'mntmeatproducts']].mode()

# finding mode across 2nd part the dataset
omni3[['mntfishproducts', 'mntsweetproducts',
       'numdealspurchases', 'numwebpurchases', 'numapppurchases',
       'numstorepurchases', 'numwebvisitsmonth']].mode()

# checking the latest date the company is operating in
omni3[['dt_customer']].sort_values(by='dt_customer', ascending=False).head()

# exporting cleaned data to excel
omni3.to_excel('cw2_cleaned_dataset.xlsx')

# finding education value counts
omni3.education.value_counts()

# grouping Alone, Absurd and YOLO to a new Other group
omni3.marital_status.value_counts()
```

Ln 11, Col 78 LF

Data Cleaning Process 3 (end)

```
+ Code   + Markdown  |  ▷ Run All   ↺ Restart   Clear All Outputs  |  Variables   Outline   ···        Python 3.12.0

# grouping Alone, Absurd and YOLO to a new Other group
omni3.marital_status.value_counts()
group_three = ['Alone', 'Absurd', 'YOLO']
omni3['marital_status_grouped'] = omni3['marital_status'].replace(group_three, 'Other')

# exporting to a different excel file
omni3.to_excel('cw2_cleaned_dataset_gMS.xlsx')

# crosschecking the value count of the new marital status group
omni3.marital_status_grouped.value_counts()



# creating a new educational and marital status columns with their counts
# creating a dictionary to replace the educational categories with scores
edu_nominal_scores = {'Graduation': 1, 'PhD': 2, 'Master': 3, 'Basic': 4, '2n Cycle': 5}
# creating a new column for the education scores
omni3['education_nominal_score'] = omni3.education.replace(edu_nominal_scores)
# creating a dictionary to replace the marital categories with scores
marital_nom_scores = {'Single': 1, 'Together': 2, 'Married': 3, 'Divorced': 4,
                      'Widow': 5, 'Other': 6}
# creating a new column for the marital status scores
omni3['maritalstatus_nominal_score'] = omni3.marital_status_grouped.replace(marital_nom_scores)
# inspecting if the new columns are added
omni3

# exporting the latest changes to a new excel sheet
omni3.to_excel('cw2_cleaned_dataset_gMS_nsMS_nsE.xlsx')

[ ]                                                                                              Python
```

Performing multiple linear regression in R, checking for VIF and multicollinearity in R

```r
# by_sina_tijani
# installing package to read excel document, and reading it
install.packages("readxl")
library(readxl)

# installing and reading tibble
install.packages("tibble")
library(tibble)

# naming the dataset omni3
omni3 <- read_excel("cw2_cleaned_dataset_gMS_nsMS_nsE.xlsx")

# creating a multiple regression model
regmodel <- lm(income ~ kidhome + teenhome + recency + mntdrinks +
                 mntfruits + mntmeatproducts + mntfishproducts +
                 mntsweetproducts + numdealspurchases + numwebpurchases +
                 numapppurchases + numstorepurchases + numwebvisitsmonth, data = omni3)

# showing the summary of the regression model
summary(regmodel)

# selection some columns to check for correlation , multicollinearity
selected_omni3 <- omni3[, c('kidhome',
                            'teenhome', 'recency', 'mntdrinks', 'mntfruits',
                            'mntmeatproducts', 'mntfishproducts', 'mntsweetproducts',
                            'numdealspurchases', 'numwebpurchases', 'numapppurchases',
                            'numstorepurchases', 'numwebvisitsmonth')]

# creating a correlation matrix with the selected columns
cor_matrix <- cor(selected_omni3)

# checking the matrixes
cor_matrix

# installing a package to visualize the correlation
install.packages("corrplot")

library(corrplot)

# plotting the correlation into a heatmap
corrplot(cor_matrix)

# creating a stepwise regression to check for best columns to include in regression model
regmodel2 <- step(regmodel, direction = "both")

# showing summary of the model
summary(regmodel2)

#calculating for VIF
install.packages("car")  # Install the 'car' package
library(car)

# Assuming 'regmodel2' is your linear regression model
vif_values <- car::vif(regmodel2)

# View the VIF values
print(vif_values)

```