Shelves of food and drinks on shelves

Description automatically generated

By: **Sina Tijani**

Report

**Statistical Analysis of Marketing Campaign**

**in Omnichannel Retail**

Table of Contents

[Literature Review Error! Bookmark not defined.](#_Toc151395250)

[Search strategy Error! Bookmark not defined.](#_Toc151395251)

[Peer-reviewed Publications Error! Bookmark not defined.](#_Toc151395252)

[Summary and analysis of selected publications Error! Bookmark not defined.](#_Toc151395253)

[Themes and Trends Error! Bookmark not defined.](#_Toc151395254)

[Gaps in the Literature: Error! Bookmark not defined.](#_Toc151395255)

[Conclusion: Error! Bookmark not defined.](#_Toc151395256)

[Descriptive Statistics Measurements 4](#_Toc151395257)

[Cleaning the Dataset 4](#_Toc151395258)

[Assumptions 4](#_Toc151395259)

[Measuring Central Tendency 5](#_Toc151395260)

[Measuring Variability 6](#_Toc151395261)

[Hypothesis Testing 7](#_Toc151395262)

[X2 (Chi-square) 7](#_Toc151395263)

[Question: 7](#_Toc151395264)

[One-way ANOVA Test (including post hoc analysis) 9](#_Toc151395265)

[Question: 9](#_Toc151395266)

[Multiple Linear Regression Analysis (including multicollinearity and VIF analysis) 11](#_Toc151395267)

[Question: 11](#_Toc151395268)

[VIF (Variance Inflation Factor) | Multicollinearity 13](#_Toc151395269)

[Conclusion 14](#_Toc151395270)

[References 16](#_Toc151395271)

[Appendix 18](#_Toc151395272)

List of Tables

[Table 1. Last Date a customer joined. 4](#_Toc151395273)

[Table 2. Descriptive Statistics of the Dataset 5](#_Toc151395274)

[Table 3. Mode 1 5](#_Toc151395275)

[Table 4. Mode 2 5](#_Toc151395276)

[Table 5. Descriptive Statistics 2 6](#_Toc151395277)

[Table 6. Initial Chi-Square Test 7](#_Toc151395278)

[Table 7. Marital\_Status value counts(vc) 8](#_Toc151395279)

[Table 8. Marital\_Status\_grouped vc 8](#_Toc151395280)

[Table 10. Cramer's V 9](#_Toc151395281)

[Table 12. ANOVA Effect\_sizes 10](#_Toc151395282)

[Table 13. One-way ANOVA test 10](#_Toc151395283)

[Table 14. Post-hoc analysis 11](#_Toc151395284)

[Table . Multiple Linear Regression(MLR) 12](#_Toc151395285)

[Table . Stepwise regression 12](#_Toc151395286)

[Table 15. VIF 14](#_Toc151395287)

List of Figures

[Figure 1. MLR Correlation Heatmap 13](#_Toc151395288)

[Figure 2. Visualizations from the dataset 15](#_Toc151395289)

# 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 . Last Date a customer joined.

A screen shot of a computer

Description automatically generated

1. 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 . Descriptive Statistics of the Dataset

A screenshot of a computer

Description automatically generated

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 . Mode 1

A screenshot of a computer

Description automatically generated

Table . Mode 2

A screenshot of a computer

Description automatically generated

## Measuring Variability

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 . Descriptive Statistics 2

A screenshot of a computer screen

Description automatically generated

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.

## X2 (Chi-square)

### Question:

Does the educational level of customers influence their marital status?

H0: There is no link between Education and Marital Status (*μ*1 = *μ*2 = … = *μk.*)

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, X2 (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 . Initial Chi-Square Test

A screenshot of a graph

Description automatically generated

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 . Marital\_Status value counts(vc)

A black rectangular object with a black border

Description automatically generated

After:

Table . Marital\_Status\_grouped vc

A screen shot of a computer program

Description automatically generated

Results:

Table 9. CrosstabulationA screenshot of a table

Description automatically generated

Table . Cramer's V

A screenshot of a calculator

Description automatically generated

Table 11. New Chi-Square test

A screenshot of a test

Description automatically generated

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 (*μ*1 = *μ*2 = … = *μk.*)

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 . ANOVA Effect\_sizes

A screenshot of a graph

Description automatically generated

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 13. One-way ANOVA test

A table with numbers and letters

Description automatically generated

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 14. Post-hoc analysis

A screenshot of a graph

Description automatically generated

## 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 . Multiple Linear RegressionA screenshot of a computer

Description automatically generated(MLR)

Table . Stepwise regression

A screenshot of a computer

Description automatically generated

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 . MLR Correlation Heatmap

A diagram of a graph

Description automatically generated with medium confidence

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 15. VIF

A close-up of numbers

Description automatically generated

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

A screenshot of a graph

Description automatically generatedFigure . 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

A screenshot of a computer program

Description automatically generated

Data cleaning process 2

A screenshot of a computer program

Description automatically generated

Data Cleaning Process 3 (end)

A screen shot of a computer program

Description automatically generated

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

A screen shot of a computer program

Description automatically generated