Text

Description automatically generated

**MS5105 - Statistical Techniques for Business Analytics: Group Assignment**

**Group 23:**

|  |  |  |
| --- | --- | --- |
| **Members** | **ID** | **e-mail** |
| Raviraj Bhandari | 24230491 | r.bhandari1@universityofgalway.ie |
| Navaneetha Mulgund | 24237749 | n.mulgund1@universityofgalway.ie |
| Aditya Kshirsagar | 24240142 | a.kshirsagar2@universityofgalway.ie |
| Devraj Raghuwanshi | 24232193 | d.raghuwanshi1@universityofgalway.ie |
| Mariana Parra Aldrete | 24259483 | m.parraaldrete1@universityofgalway.ie |

**Words**: 1,832

*“In submitting this work, we confirm that it is entirely our own work. We acknowledge that we may be invited to online interview if there is any concern in relation to the integrity of my submission, and we are aware that any breach will be subject to the University's Procedures for dealing with breaches of Exam Regulations. We are aware of what the University of Galway plagiarism policy entails.”*

## **Objective**

The objective of this project is to demonstrate our acquired knowledge of statistical techniques such as Logistic Regression, the Kruskal-Wallis Test, Mixed Model ANOVA, and One-Way ANCOVA. Additionally, the project aims to showcase our understanding of the data to be analysed, as well as our ability to interpret and present the results obtained from these techniques.

## **Data description**

The data set selected is categorised as a “Sales data”, 1,000 inputs with no missing values generated through random logic in VBA; it could be assumed that it is a business that conducts sales both online and offline, offering various types of products worldwide.

## **Logistic Regression**

Question:   
Can customer demographics and purchase behaviour predict whether a customer will purchase high-priority items?

Statistical Test:

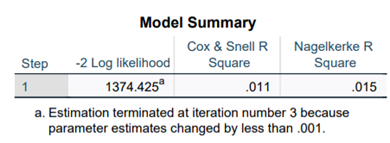
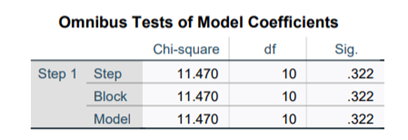
We performed Logistic Regression Analysis as our objective was to predict a binary outcome (purchase of high-priority items: Yes/No) based on a combination of continuous and categorical predictors.

Data Preprocessing:

Before running the logistic regression, the following steps were performed:

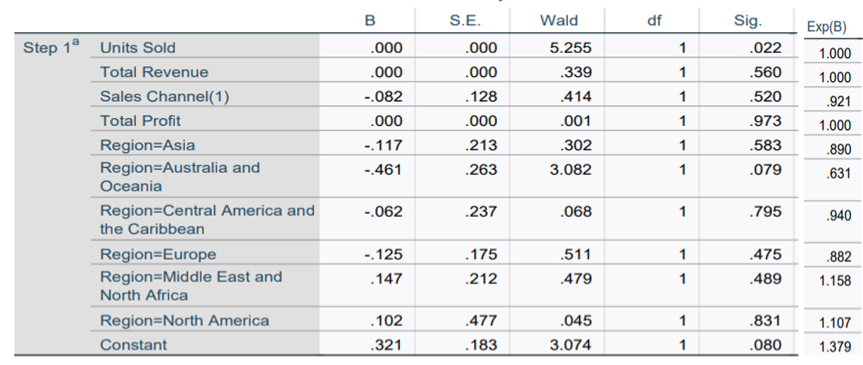
1. Creation of the Dependent Variable:
   1. The column Order Priority was recoded into a binary variable:
      1. 1 for High (H) and Critical (C) priorities.
      2. 0 for Medium (M) and Low (L) priorities.
2. Handling Categorical Variables:
   1. Region: Dummy variables were created for all region categories, such as:
      1. Region\_Asia, Region\_Europe, Region\_NorthAmerica, etc.
      2. One category was excluded as the reference group to avoid multicollinearity.
   2. Sales Channel: Included as a categorical covariate (Online/Offline), with SPSS automatically treating Offline as the reference group.

Outputs:



The logistic regression model was executed, and the results are summarized below:

1. Model Fit:
   1. Omnibus Tests of Model Coefficients: Chi-Square =11.470, p = 0.322.
      1. This indicates that the model, as a whole, is not statistically significant.
   2. Nagelkerke R²: 0.015.
      1. The predictors explain only 1.5% of the variance in high-priority purchases.
2. Variables in the equation



Interpretation of results:

1. Units Sold:
   1. This is the only statistically significant predictor (p = 0.022), indicating that an increase in units sold has a slight positive effect on the likelihood of purchasing high-priority items.
2. Sales Channel (Online):
   1. While Online purchases slightly reduce the odds of high-priority purchases compared to Offline (Exp(B) = 0.921), this effect is not statistically significant (p = 0.520).
3. Region:
   1. The category "Australia and Oceania" had a marginally significant negative effect (p = 0.079), reducing the odds of a high-priority purchase (Exp(B) = 0.631).
   2. Other regions showed no significant impact.
4. Other Predictors (Total Revenue, Total Profit):
   1. These variables did not significantly contribute to predicting high-priority purchases.

* Units Sold is a significant predictor of high-priority purchases, albeit with a minor effect.
* Region (Australia and Oceania) has a marginally significant negative impact on the outcome.
* Other predictors, including Total Revenue, Total Profit, and Sales Channel, do not have significant effects.

## **KRUSKAL-WALLIS Test**

Question:

Does the number of units sold differ significantly across different regions?

Statistical Test:

We have run the normality test by adding region (categorical variable) in Label by factor by options to make it up for the multiple levels and comes out that the data is not normally distributed, we have used Non-Parametric Test called as Kruskal Wallis Test.

Data Processing:

Normality Test:

* Each region has 100% valid cases, meaning there are no missing data points for "Units Sold."
* The total number of cases matches the valid cases for every region (e.g., Asia has 136 total cases, all of which are valid).

The Kruskal-Walli’s test is appropriate for comparing regions due to non-normal distributions and significant differences in standard deviations.

Non-Parametric Test: P value 0.590, which is greater than the significance level of 0.05, We retain the null hypothesis, indicating no statistically significant difference in total esteem scores across different levels of educational attainment. This suggest that total number of units sold does not have a significant impact on different regions.

Outputs:

A screenshot of a test

Description automatically generatedA close-up of a test

Description automatically generated

* Test Statistic: 4.65
* p-value: 0.59

Since the p-value (0.59) is much greater than the typical significance level of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant difference in the number of units sold across the different regions.

The "Test Statistic" column indicates the difference in rank sums between two regions, and the "Sig." column shows the significance (p-value) of these comparisons. Since the adjusted significance values (Adj. Sig.) are greater than 0.05 for all comparisons (after Bonferroni correction for multiple tests), there is no evidence to reject the null hypothesis that the distributions of the groups being compared are the same.

A graph showing the different states of the same type

Description automatically generated with medium confidence

This box plot represents the results of an independent-samples Kruskal-Walli’s test, comparing the distributions of "Units Sold" across different regions. The x-axis lists the regions (Asia, Australia and Oceania, Central America and the Caribbean, Europe, Middle East and North Africa, North America, Sub-Saharan Africa). The y-axis represents the number of units sold.

Interpretation of results:

A Kruskal-Walli’s test was conducted to evaluate whether there was statistically significant difference in number of units sold across difference regions. The results of the test indicated there is no statistically significant difference in number of units sold across difference regions.

H (4) =4.65, p=0.59. This suggests that the distribution of total number of units sold is the same across different regions.

## **Mixed Model ANOVA**

Question:

The primary objective of this study was to evaluate whether Sales Channel (Online vs. Offline) and Region influenced Total Profit over time (2010–2017). The following specific questions were addressed:

1. Does Total Profit significantly change across years?
2. Does Sales Channel affect Total Profit?
3. Does Region influence Total Profit?
4. Are there interaction effects between Year, Sales Channel, and Region?

Statistical Test:

We utilized a Mixed Model ANOVA approach, with the following factors:

* Within-Subjects Factor: Year (8 levels: 2010–2017).
* Between-Subjects Factors: Sales Channel (Online, Offline) and Region.
* Dependent Variable: Total Profit.

Data processing:

* Extracted Year from the Order Date column for repeated measures.
* Restructured the dataset to include Total Profit as the repeated measure across years, while retaining Sales Channel and Region as between-subject factors.

Outputs:

Main Effects

1. Year: No significant effect of Year on Total Profit (F (6.79, 986) = 0.489, p = 0.838).  
   Conclusion: Total Profit remained relatively stable over time.
2. Sales Channel: No significant difference in Total Profit between Online and Offline channels (F (1, 986) = 0.029, p = 0.864).  
   Conclusion: Sales Channel did not significantly influence Total Profit.
3. Region: No significant effect of Region on Total Profit (F (6, 986) = 0.204, p = 0.975).  
   Conclusion: Regional variations in Total Profit were not significant.

Interaction Effects

1. Year × Sales Channel: No significant interaction (F (6.79, 986) = 0.785, p = 0.596).  
   Conclusion: Changes in Total Profit over time were consistent across Sales Channels.
2. Year × Region: No significant interaction (F (41.59, 986) = 1.256, p = 0.125).  
   Conclusion: The effect of Year on Total Profit did not differ by region.
3. Year × Sales Channel × Region: A weak trend was observed (F (6.00, 986) = 1.817, p = 0.093), suggesting a possible combined effect of Year, Sales Channel, and Region, though not statistically significant.

Visual Representation

1. Year vs. Sales Channel: Estimated Marginal Means by Year and Sales Channel  
   The profile plot shows consistent trends in Total Profit across years for Online and Offline channels, with no significant divergence.

**A graph with lines and numbers

Description automatically generated**

1. Year vs. Region: (Refer to Figure 2): Estimated Marginal Means by Year and Region.  
   Variability in Total Profit is observed across regions, but the differences were not statistically significant.

**A graph of a number of lines

Description automatically generated with medium confidence**

Interpretation of results:

The analysis revealed that Total Profit was not significantly influenced by Year, Sales Channel, or Region, and no significant interaction effects were observed. The results suggest that Total Profit remained stable over the years, irrespective of sales channels or geographic regions.

Addressing the Research Questions

1. Does Total Profit significantly change across Years?  
   No, Total Profit remained stable across years.
2. Does Sales Channel (Online vs. Offline) affect Total Profit?  
   No, there was no significant difference between Online and Offline channels.
3. Does Region influence Total Profit?  
   No, Total Profit did not significantly differ across regions.
4. Are there interaction effects between Year, Sales Channel, and Region?  
   No significant interactions were observed, although a weak trend was identified for the combined effect of Year, Sales Channel, and Region.

## **One-Way ANCOVA**

Question:

Does Region significantly influence the unit price, after adjusting for the sales channel and item types?

Statistical Test:

ANCOVA is used to control external factors (item type and sales channel) which might have influence on the outcome (unit price).

* Independent Variable: Region
* Dependent Variable: Unit Price
* Covariables: Sales Channel and Item types

H0: The mean of the Unit price is the same across all seven regions.

H1: The mean of the Unit price is distinct for at least one region compared to the rest.

Data processing:

The values in Region,

Sales Channel and Item type were recoded, and new tables are created with newly assigned numeric values for each unique values for given columns.

Outputs:

A table with numbers and a few black text

Description automatically generated

One-way Anova is insignificant as pvalue, but we do not know which region has insignificant unit price.

A screenshot of a graph

Description automatically generated

The Eta-Squared indicates small effect of 0.4% variance in the unit price across seven regions.

The mean difference is not significant between any regions as there are no asterisks on any of the mean differences.

Possible Bias:

* The effect of Region on Unit price may depend on the predominant Sales Channel and item types in that region.
* Does the relationship between Region and Unit Price vary depending on the predominant Sales Channel (Online vs. Offline) and item types?
* Underlined Assumption – The values of the covariate cannot vary across the different levels of the independent variable which had been assessed by ANOVA test. The assumption is met here.

Interpretation of results:

Though ANCOVA refines the impact of sales channel and item types, the effect size for the region only increases for item type from 0.04 to 0.025 while it remains constant at 0.05 for sales channel, this adds more light on assessment of the region’s contribution to unit price. Overall, there is no change in the unit price for the region.

## **Overall conclusion**

Based on the results obtained through various techniques, we can assume that:

1. The only variable that helps predict behaviours is "Units Sold," which indicates that, based on the observed behaviour, the purchase of "high-priority items" is increasing.
2. There is no significant difference in the purchase of products across different regions.
3. Total Profit was not significantly influenced by Year, Sales Channel, or Region, and no significant interaction effects were observed. The results suggest that Total Profit remained stable over the years, regardless of sales channels or geographic regions.
4. Region does not influence the "unit price."

## **Suggestions to business:**

Run an analysis based on seasons to understand further the behaviour of the customers, in order to prevent sales and also could run marketing campaigns to push sales or consider the idea of adapting the price to the local market for the purpose of improve the revenue.

## **Reference:**

Verma, V.A. (2021) Downloads 18 - sample CSV files / data sets for testing (till 5 million records) - sales, Excel BI Analytics. Available at: https://excelbianalytics.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/ (Accessed: November 2024).