****

**Advanced Data Analytics**

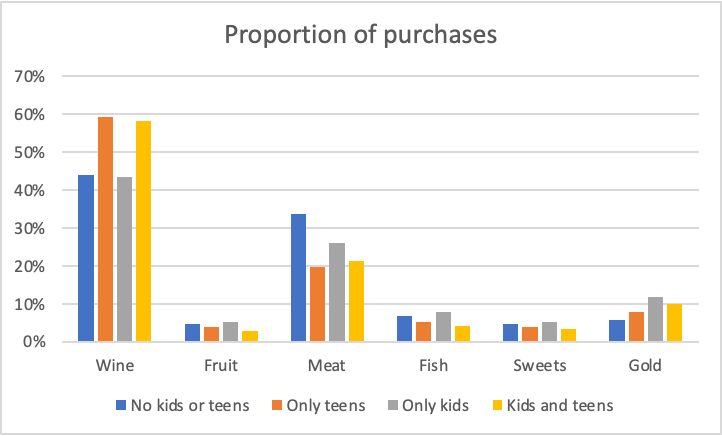
**A3: Team Project**

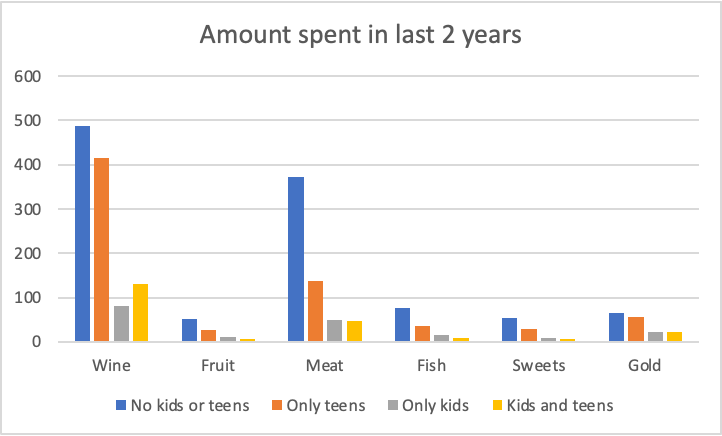
**Team 12 - Hasnain Ahmed, Pedro de Menezes Pereira Batistelli,**

**Hanna Läticia Gerharter, Martin Sandberg, Cha Boon Tan**

**11/26/2020**

|  |
| --- |
| You’ve been told by the Chief Marketing Officer that recent marketing campaigns have not been as effective as they were expected to be. You need to analyze the data in your data set to understand this problem. Use appropriate graphs and charts to present your analysis to the CMO. Some items to consider are below. Take some time to familiarize yourself with the data set before answering your questions. Please provide a two part analysis for the case as follows: |
| **Part I: Statistical tests and Regression** |
| **A.**What factors are significantly related to the number of web purchases? What would you suggest to CMO to improve web purchases? For full credit, provide a) Regression output, b) Analysis of output in no more than 350 words, c) Suggestions to CMO in no more than 500 words.  **Recommendation 1**  **Ho** :p value of X >= alpha 0.05 ( we understand there is a relationship between the variables of X in Y)  **Ha**:p value of X <= alpha 0.05 (we understand there is no significant impact of Y if X changes)  To find out what determines the variability of the number of website purchases denoted by Y dependant variable and using X as the independent containing the number or wine, meat and gold products which were most popularly bought items as a whole. With 680304 bottles of wine being sold, 373693 pieces of meat and 98527 pieces of gold being sold within the last 2 years. To confirm our hypothesis we ran a regression model which showed us that the f significance is less than alpha of 0.05 meaning the model is accurate, meanwhile the p value of all the x values shows they are greater than alpha 0.05 meaning we accept the null hypothesis that the values of X do impact the variability of Y meaning that the number of website purchases it's mostly catering the bulk of wine, meat and gold commodities rather than other offering such and candy, fruits and fish.  Therefore we recommend that the commodities of candy, fruits and fish be removed from the website and the bulk quality of wine, meat and gold be increased as it is most consumed regardless of marital status, income or education.  **Recommendation 2**  Furthermore we have also used 2 more variables to further strengthen our analysis which are the number of website visits denoted by Y and X denoted by the acceptance of campaign 3 and the purchase of catalogues. We did this as we noticed that there were significant visits reported from people who accepted campaign 3 compared to all other campaigns and the likelihood that people bought a catalogue also increased website visits. With the hypothesis being the same we tested using a regression model which showed us that the p values were all greater then alpha 0.05 meaning we accept null and reject the alternative hypothesis.  There is clear indication that we should provide more catalogues to potential customers and these should be free of cost, this would boost sales on the website through an increase of customer traffic also we should offer more campaigns similar to number 3 as it seems to be the most well received by customers compared to their campaigns and this should be advertised both on the website and the catalogue. |
| **B**. Does the US fare significantly better than RoW (Rest of the World) in terms of **total purchases.** For full credit, provide a) Regression output, b) Details of how you computed Rest of the World and Total purchases, in no more than 100 words, c) Your analysis of the results in no more than 100 words.  **Ho** :p value of X >= alpha 0.05 ( we understand there is a relationship between the variables of X in Y)  **Ha**:p value of X <= alpha 0.05 (we understand there is no significant impact of Y if X changes)  The output of the regression analysis shows that there is no correlation between total purchases and the buyer’s nationality. This can be inferred by analysing the output of the regression, which indicates a very low Adjusted R Square, a P-value smaller than alpha of 5% for the intercept but a high P-value for the variable dummy\_country. In order to run this analysis it was needed to calculate Total Purchases.  Total Purchase = NumWebPurchases + NumCatalogPurchases + NumStorePurchases  After that it was created a dummy variable called dummy\_country to transform the categorical (qualitative) data (Us and Rest of the world) into quantitative data in order to run the regression without errors. The next step was running the regression using Total Purchase as the Y range and dummy\_country as the X range. |
| **C**. Your supervisor insists that people who buy gold are more conservative and as such people who spent **an above average amount** on gold in the last 2 years would have more in store purchases. Justify/refute the statement using appropriate statistical tests. Hint: Use your knowledge of dummy variables. For full credit, provide a) Output of test(s), b) Your justification/rebuttal in no more than 100 words.  **Ho**: p value of X >= alpha 0.05 ( we understand there is a relationship between the variables of X and Y)  **Ha**: p value of X <= alpha 0.05 (we understand there is no relationship between the variables of X and Y)  The output of the regression analysis shows that people who spent an above average amount on gold in the last 2 years do not have more in store purchases. Furthermore, based on the Multiple R, which indicates 0,4071, we can conclude that there is a very weak relationship between X and Y. In addition, the p-value for people who spent an above average amount on gold in the last 2 years is 3,5573E-90, which is smaller than alpha = 0.05 and therefore we can reject the null hypothesis. |
| **D**. An **interaction effect** occurs when the effect of one variable on a dependent variable depends on the value of some other explanatory variable. This effect is captured by **creating** an interaction variable, by multiplying the two relevant explanatory variables. Use your knowledge of interaction variables/effects to answer this part. Fish has Omega 3 fatty acids, good for the brain, accordingly, do "Married PhD candidates" have a significant relation with the amount spent on fish? What other factors are significantly related to the amount spent on fish? For full credit, provide a) Regression output, b) Your answer to the questions stated above.  Based on the Adjusted R- square in the regression output, 1% of the variation of amount spent of fish around the mean are explained by the Married PhD candidate. There is a very weak relationship between the two variables. People with advanced degrees and married do not purchase more fish than others. Although fish has Omega 3 fatty acids, good for the brain, accordingly, people with advanced degrees do not purchase more fish than others. 25% of the variation of amount spent on fish around the mean are explained by the High Income Married candidate. There is a weak relationship between the two variables but higher than Married PhD candidates and Amount spent on fish. |
| **E**.Do any other analysis you deem relevant to show to your CMO. For this purpose, **propose a hypothesis and perform the appropriate tests**. **Be creative and insightful.** We can ask any questions about the last one. Just try to impress us! |

****

****

**E - 1**

**H(0): Customers with children spend the same amount as customers without**

**H(a): Customers without children spend more than customers with children**

The 95% two-sample confidence interval for the difference between sample means are $645 to $754. The range excludes 0 so we can reject the null hypothesis and be 95% confident that customers without children at home spend on average $645 to $754 more in the past two years.

**E - 2**

**H(0): Customers with teens buy wine in the same proportion as customers without**

**H(a): Customers with teens buy more wine in proportion to other products**

The 95% two-sample confidence interval for the difference between proportions are 0.11 to 0.19. The range excludes 0 so we can reject the null hypothesis and be 95% confident that for customers with teens at home wine makes up 11% to 19% more of total purchases than customers without teens.

**E - 3**

**H(0): Customers with teens buy fish in the same proportion as customers without**

**H(a): Customers with teens buy less fish in proportion to other products**

The 95% two-sample confidence interval for the difference between proportions are -0.052 to -0.015. The range excludes 0 so we can reject the null hypothesis and be 95% confident that for customers with teens at home fish makes up 5.2% to 1.5% less of total purchases than customers without teens.

**E - 4**

**H(0): Customers with children buy meat in the same proportion as customers without**

**H(a): Customers without children buy more meat proportionally to other products**

The 95% two-sample confidence interval for the difference between proportions are -0.169 to -0.086. The range excludes 0 so we can reject the null hypothesis and be 95% confident that for customers with children at home meat makes up 16.9% to 8.6% less of total purchases than customers without children.

**E - 5**

Exploring the data we found that 14 out of the 21 customers who had complained in the last two years were from Spain. This led us to propose the following hypothesis:

**H0: Customers from Spain do not complain more than customers from the rest of the world**

**Ha: Customers from Spain complain more than customers from the rest of the world**

The 95% two-sample confidence interval for the difference between proportions are -0.0014 to 0.0147. The range includes 0 so we cannot reject the null hypothesis. However, the lower end of the confidence interval is just below zero at the 95% confidence interval, it could be an association worth looking into.

