NOVA SCHOOL OF
BUSINESS & ECONOMICS

Group Assignment 1

Marketing Analytics

Team:

Catarina Ramos, 29219

Diogo Miguel Pereira Sá, 46105

João Gil Neto Ribeiro, 32399

Melania Padolecchia, 45733

Instructors:

Rodrigo Belo

Carolina Lourenço

Introduction

It is key for a company to understand the effectiveness of previously implemented actions. This process, first of all, requires the acquisition of meaningful data which - after careful analysis – could give insights into the performance of a specified operation.

In this case, this procedure has been activated for an upcoming online retail specialized in selling books and technology, named Books&Tech. The task that the team was instructed on, was to evaluate the performance of an e-mail marketing campaign targeting customers who have interacted at least once with the store and completed a purchase in the last 12 months.

The magnitude of the data acquired - which has been recorded throughout 4 weeks - counts up to 45,572 customers which have been separated into three categories:

- Customers who received an e-mail campaign featuring books.
- Customers who received an e-mail campaign featuring electronics.
- Customers who did not receive an e-mail campaign.

The analysis conducted by the team has been structured differently according to the different desired results to be achieved.

In the first part, an exploratory analysis has been carried out to understand properly the nature of the data, as well as the one of the randomization. This step is essential because it works as a preparatory phase that helps the reader to be comfortable in managing different types of information and understanding the various distinctions.

In the second part, a more elaborated investigation has been implemented. This section aims to understand in detail if the campaign's design and results were successful in terms of sales.

Another aspect that needs to be highlighted is that some modules of this report will be dedicated to answering specific questions to clear any doubts and make the reading as fluent as possible. Moreover, in some cases, recommendations will be elaborated on the output obtained so that the reader will be able to follow the basic theme of thought which guides the development of this entire report.

Exploratory Data Analysis and Randomization

Numerical and Binary variables

What emerged from an initial glance and basic descriptive analysis of the Numerical and Binary variables will be explained as follows.

The Last Purchase mean is around June - indicated as "month 6" - but, at the same time, it should not be forgotten that there is a large standard deviation, and the distribution may not be normal.

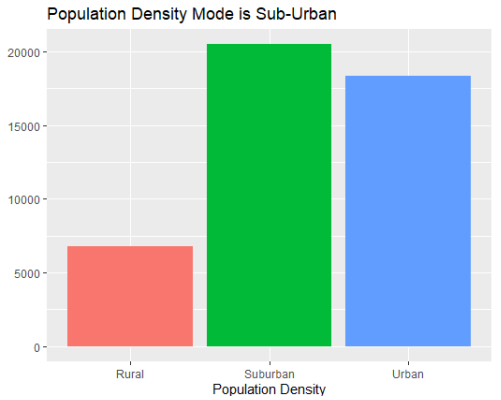
The historical spending of last year has a very large standard deviation. With this, it is possible to infer that its distribution must be skewed with a long right tail, given the large maximum value – which is 3,345 – and, also, considering the fact of having 75% of all variables up to only 325.

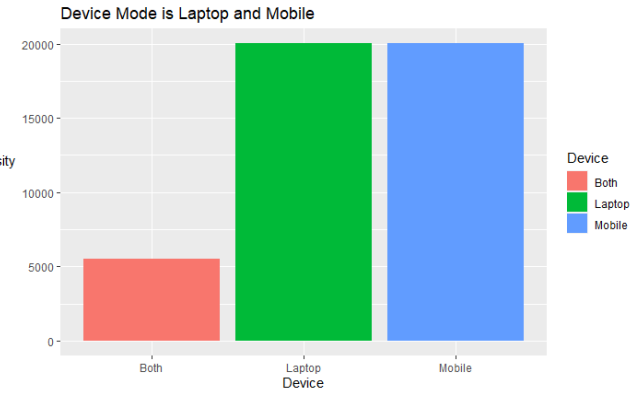
Electronics and books share approximately the same mean, both having more than 50% of customers buying the respective product. This has been inferred by the fact that it is a binomial variable, and the mean will be the sum of all entries divided by all entries.

The "new_customer" variable appears to be perfectly balanced with an almost equal number of regular/old customers and new customers.

Categorical variables

Regarding categorical variables, the mode for the variables *Population density* and *Device* is, respectively, Sub-Urban and Laptop/Mobile.





Predictor features

This section's structure consists of a sequence of exploratory questions aiming to better understand the problem and the dataset.

When are customers (old and new) making their last *purchases?*

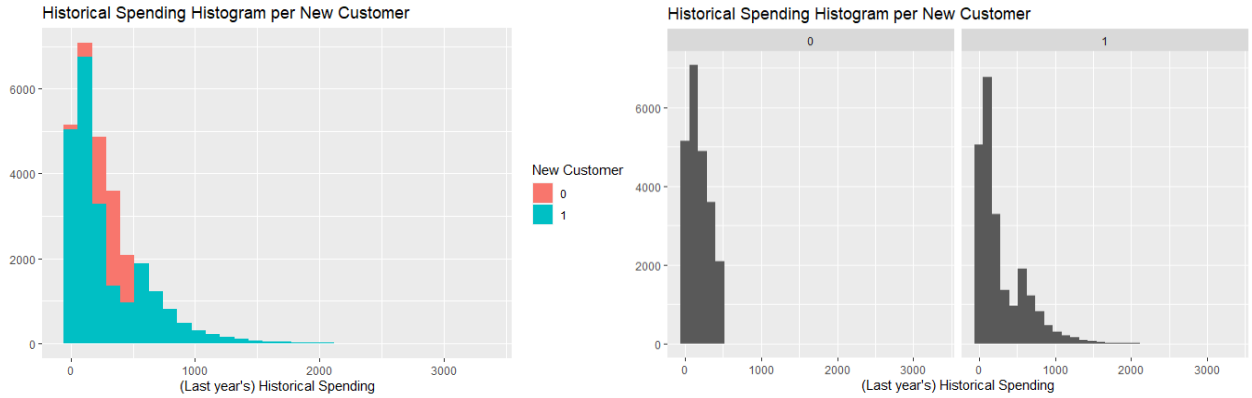
It appears that last purchases seem to have an increase in the last month of the year, which coincides with the Christmas season. Regular/old and new customers make the majority of their last purchases, with new customers being the largest share of buyers for this period.

This lead is sustained over regular customers until September. Then regular customers become the largest share of last purchases as fewer people are making their last purchases in the summer months of August through to May. In April and March, it is possible to observe again a large number of last purchases from both regular/old customers, now with regular customers being the largest share of the respective months. Then, in February and January, the lowest number of last purchases in the year is recorded.

Further investigation might be required to verify if there is any external reason to prompt the increase in last purchases that early in the year (March and April). Except that, it is possible to assume that the Easter holidays can partially explain that increase.

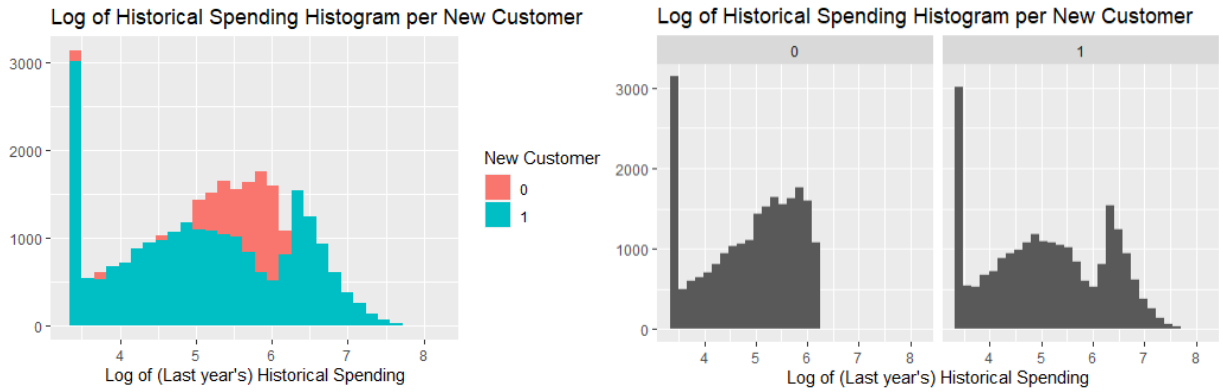
Are new customers spending more than regular customers?

The response to this question is affirmative, historical spending (last year's spending) per customer yields distribution a long right tail clearly shows that new customers spend much more than regular/old customers.

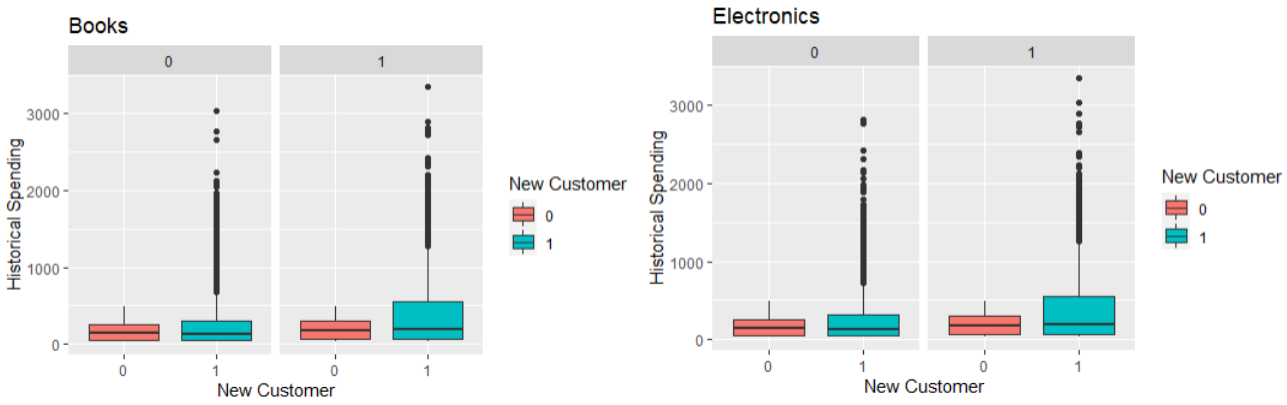


This is what is called a "Power Law" distribution, by applying a log it is possible to see something closer to a normal distribution. However, in this case, we observe that a very significant number of customers spends a very small amount, with older customers then following a more normal distribution and new customers either buying less than average or more than average (if we consider these to be normal distributions without the clear and significant outlier on the left tail).

It also appears to exist a barrier of spending that is shared among a significant share of regular customers, around the 500 to 600$. This could be a significant insight when considering whether to retain regular/old customers or shifting the attention to acquiring new customers.

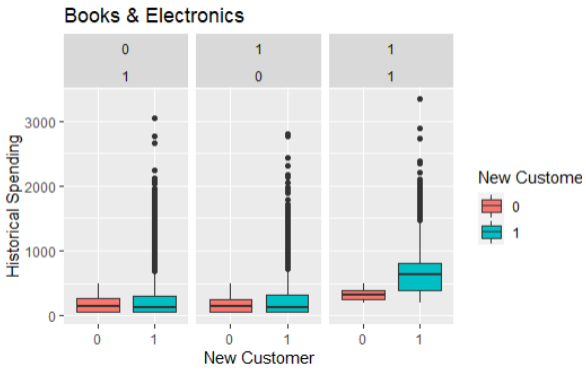


What is the relation between new customers and books/electronics?



There is a statistically significant difference in means between new and old customers for both Books and Electronics historical spending [Check R file for t-test results].

New customers have a significant number of outliers in spending, with the largest one being a person that bought both books and electronics last year. This tells us that for customers that buy either only books or only electronics last year, their distribution will have a longer tail than that of a regular customer. There is a clear difference in means between new and old customers buying both books and electronics, which is verified by performing the F-test where the output shows that books and electronics are jointly statistically significant.




```
Call:
lm(formula = new_customer ~ books + electronics, data = dt.em)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6053 -0.4918  0.3947  0.5082  0.5146

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.371876   0.008822   42.16   <2e-16 ***
books       0.119918   0.008102   14.80   <2e-16 ***
electronics 0.113497   0.008105   14.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4988 on 45569 degrees of freedom
Multiple R-squared:  0.005029,  Adjusted R-squared:  0.004985
F-statistic: 115.2 on 2 and 45569 DF,  p-value: < 2.2e-16
```

The order of graph columns is (book – first row (0,1,1), electronics – second row (1,0,1)).

How much do the predictor variables explain new and old customers?

By doing a multilinear regression on *new_customer*, *hist_spend* (last year expenditure), and *deviceLaptop* and *deviceMobile* are statistically significant at a 99.9% Confidence Interval. Interestingly, the time since the last purchase is not statistically significant when it comes to explaining new customers [Check R file for Multilinear regression output].

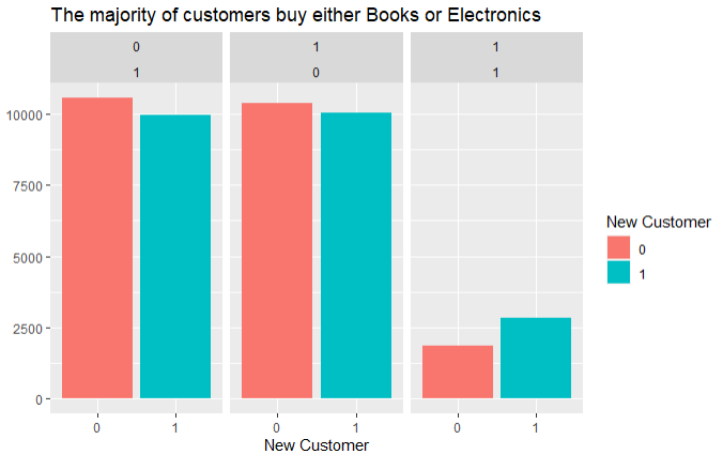
Inspection of relations between all combinations of variable pairs

The correlation between historical spending (last year's spending) and last purchase is statistically significant at -0.246. Therefore, there is a weak negative correlation implying that we expected to have higher spending with customers that made their last purchase at the end of the year, which seems intuitive given that the end of the year has Christmas season a well know and planned expense for many customers. The relatively weaker correlation comes from the spike in March and April in what would otherwise be a downward trend throughout the year. It is important to reinforce the need to investigate what is causing such a significant increase in last purchases this early in the year, as the ideal would be to have a strong negative correlation between historical spending and time since last purchase as it would imply larger revenue for the company.

Books and electronics slightly negative correlations with last_purchase suggest that there may be little influence in the timing of the last purchase and this being books or electronics, actually, with this data it is necessarily one or the other.

Books and electronics slightly positive correlation with historical spending (last year's spending) which likely comes from having 2 out of 3 possible states of purchase. These are, (Books, N° Electronics), (N° Books, Electronics), or (Books, Electronics). As seen in the Box plot for Books and Electronics, there is a statistically significant difference in means, between old and new customers with the highest historical spending mean of all combinations coming from the purchase of both books and electronics.

Books and electronics have a strong negative correlation, which is to be expected given that only a small share of customers bought both products during the year.



The order of graph columns is (book – first row (0,1,1), electronics – second row (1,0,1)).

Randomization

Are both groups comparable?

Firstly, it was carried out a comparison to see if the values of the variables from the group that received the e-mail campaign featuring electronics are different from the one that did not receive any e-mail (control group). If the randomization is properly done, the p_value of each test should be greater than 0.1 and it would be possible to reject the null hypothesis.

For the variables *(visit_after, purchased_after, spend_after*), the p-value is expected to be close to 0, since these are the output variables and it is expected that the variables behave differently across the two groups, otherwise the campaign would not have had any result (e.g. a group of customers that receive an e-mail campaign featuring electronics will probably spend more dollars than one that did not receive any e-mail campaign).

All tests yielded a p-value greater than 0.1 (except, as expected, the last 3) [check R file for the results], Therefore this treatment group (the ones who received an e-mail campaign featuring electronics) is properly randomly assigned and both, treatment, and control groups, are comparable.

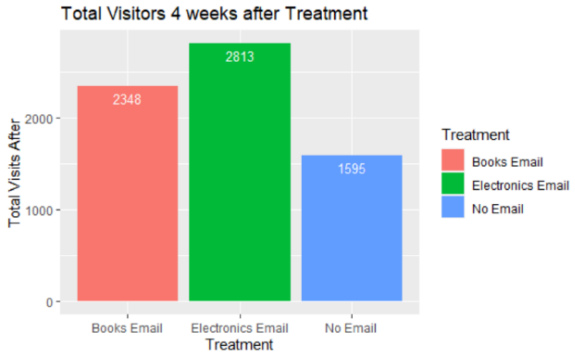
Furthermore, t-tests were made to compare values from the group that received the e-mail campaign featuring books with the ones that did not receive an e-mail and the p-values were also greater than 0.1, implying that the randomization was properly done [check R file for results].

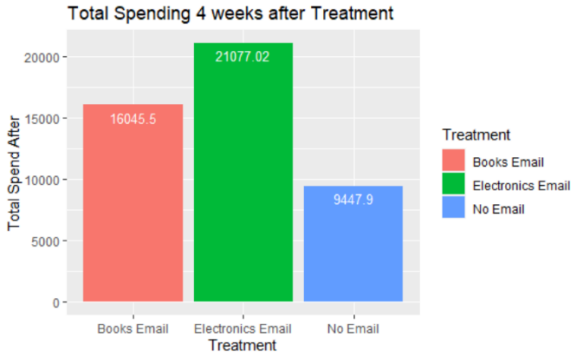
Finally, to make sure that both treatment groups are comparable between them, so that the results of the experimentation are valid (if both groups are not comparable there might be a case where the results are due to the characteristics of the sample and not due to the experimentation), t-tests were realized to check if the values of the variables from the group that received the e-mail campaign featuring electronics are different from the one that did receive featuring books. These tests yielded a p-value greater than 0.1, implying that results are valid, and the randomization was properly done [check R file for results].

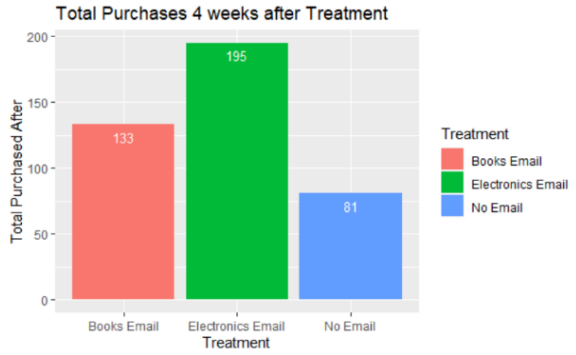
Analysis of the Experiment Design and Results

Which campaign performed the best overall, the Books version, or the Electronics version?

Both campaigns seem to have had a positive impact on the visits and sales (both number of purchases and amount spent). However, the Electronics version of the campaign performed better than the Books version, resulting in an extra 465 visits, 62 purchases, and 5031.52€ spent.



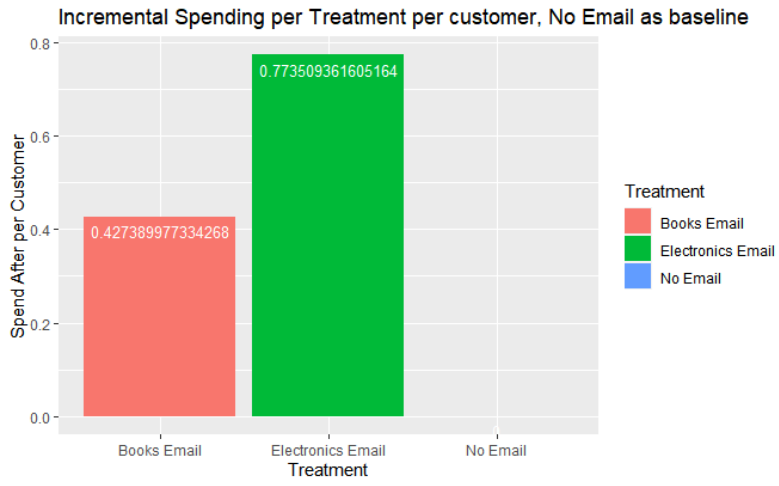




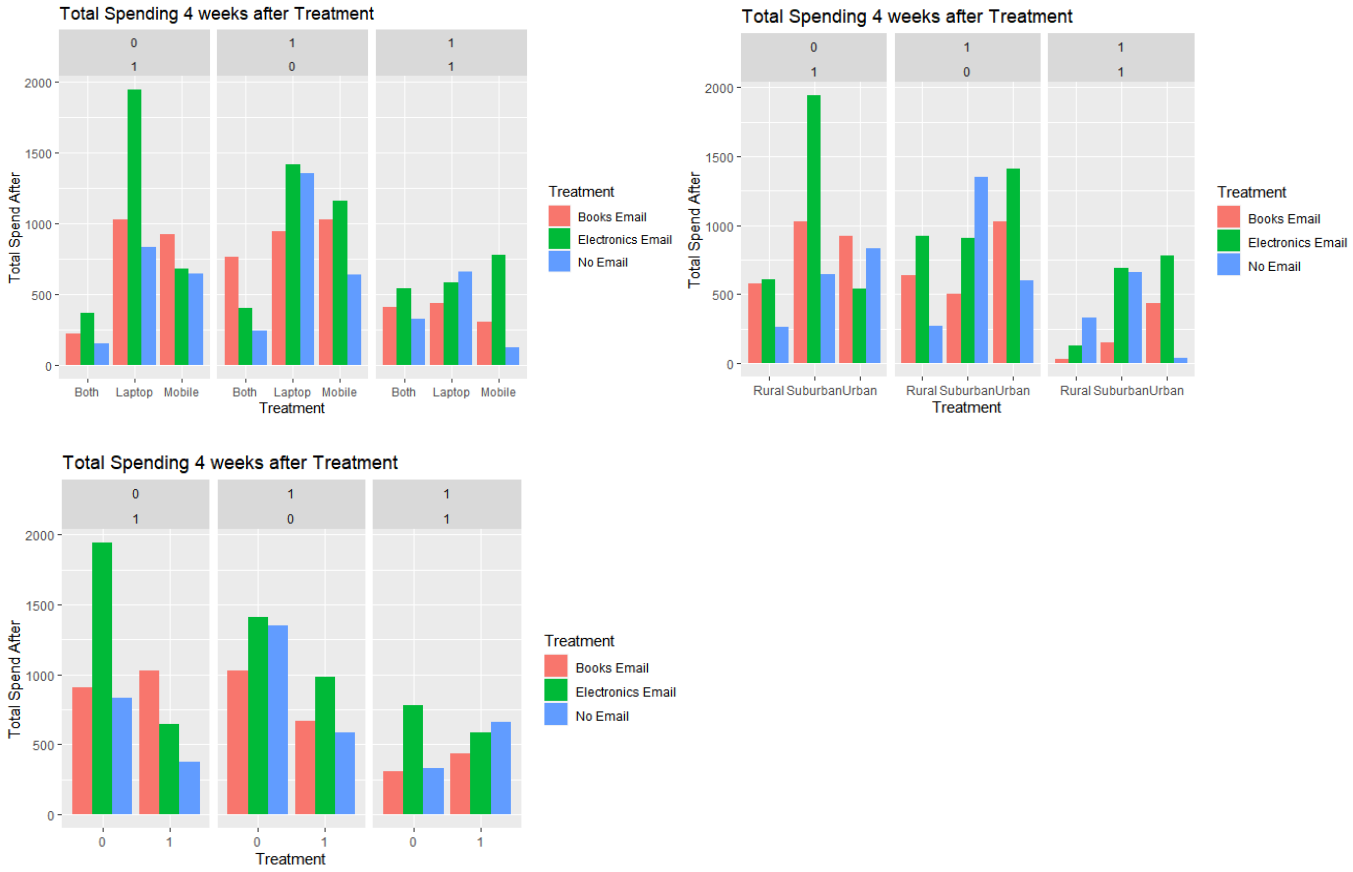
How much incremental sales per customer did the Books version of the campaign drive? And how much incremental sales per customer did the Electronics version of the campaign drive?

The books version of the campaign yielded an incremental 0.42739 € per customer. With 15287 customers spending 16045.5 €. Which equates to 34.45427% of the total sales 4 weeks after the experiment

The electronics version of the campaign yielded an incremental 0.7735094 € per customer. With 15101 customers spending 21077.02 €. Which equates to 45.25839 % of the total sales 4 weeks after the experiment.



Is cross-selling a good strategy for Books&Tech? In other words, which audience would you target the Books version to, and the Electronics version to, given the obtained results?



The order of graph columns is (book – first row (0,1,1), electronics – second row (1,0,1)).

Considering cross-selling as customers that bought a product last year and received an email for a different product, we have that, under columns (book=0, electronic=1) new customers, from Rural or Sub-Urban areas using either a Mobile or Laptop were the most responsive to cross-selling books having previously bought only electronics in the past year.

Regarding cross-selling electronics to previous book buyers, new customers, from Rural or Urban areas using a Mobile show the most response to buying electronics after receiving the Electronics Email and having only purchased books in the past year.

This is a good strategy further supported by the previous finding that new customers spend more money than old ones, especially when buying both books and electronics (at least during their first year as customers) and that there is a large amount of untapped sales given most customers buy either one or the other product.

Next steps and Experience Re-design

Evaluating the experiment, it has been concluded that its size seems to be sufficient, although to have a better grasp of trends a longer timeline of the data would be desirable as some relevant questions cannot be answered without knowing if the analyzed year is a representative year for the company.

Regarding the metrics used, it would be suitable to introduce a new variable to assess if customers that receive the coupon use it. Without it is not possible to properly evaluate the effect of the treatment since old customers, for example, can keep buying without checking the email. Another good contribution would be the historical average for monthly purchases per customer, this would allow seeing differences in months and access when are customers buying more or less. At the moment, it can only be seen a difference between new and old customers' spending but are unable to see when it occurs during the year.

Regarding the duration of the experiment, it is not possible to access if 4 weeks are enough to see a significant difference in treatments, but it is evident that the patterns of purchase are very different throughout the year, with results tending to vary depending on the month of the experiment. A test in December would have very different effects from one in January as seen from the analysis above.

Finally, another key element to be re-assessed should be the number of treatments. It would be relevant to include a new treatment group in which the customers receive an e-mail campaign combining both electronics and books given the sales for both products yield the highest historical spending for new customers and that cross-selling is a good strategy. However, attention would have to be paid to erosion of sales or possible overwhelming of customers with too many choices (choice paralysis).