Warmart Lab

# Change of Customer Engagement over time

Jiawen Huang
2-4-2020

## Data set

Our data set consists of three segments, which include the basic information of 100 household customers, the information of our products sold in stores, and the information of transactions within 104 weeks (Year 2017 and Year 2018).

The detailed information included in each segment are as below:

1. Household customers data: household number, loyalty or not, range of age, maritage, income range, homeowner, composition, size, and number of kids.
2. Products data: product number, department, commodity type, brand type, and organic or not.
3. Transaction data: household number (buyer), purchase date, product number, money spent, units bought, region of store, week number, and year.

## Report Goal

In the 104 weeks of year 2017 and year 2018, I wanted to know if the customers spent more or less money as time went by. In order to see if there is a tendency on sales, I first did exploratory data analysis.

If an obvious pattern existed, I needed to find out the reason, such as the origin of the loss or benefit, the category that contributed most to the pattern, and the specific type of household customers contributed most to the pattern.

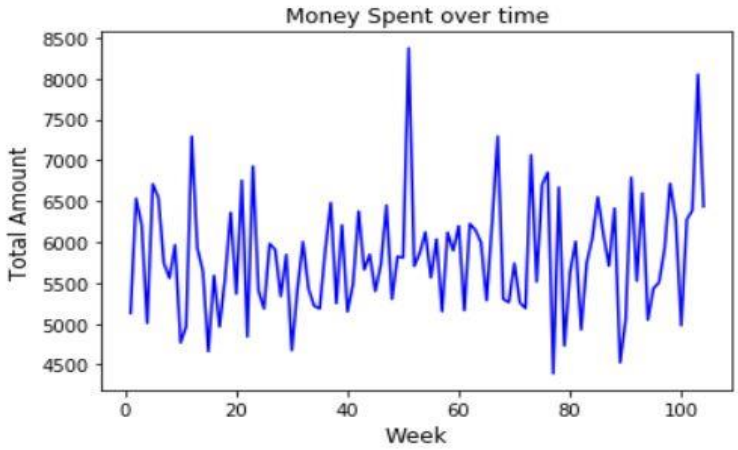
If such a pattern was not existent, I wanted to further analyzed the sales data and found if some factors of customers played important roles and the four main factors were age range, maritage condition, income range, and size. By checking the influence of these factors with statistical tools, I could draw the conclusion and made recommendation for expanding specific group of household customers if the influence was significant. If it was significant, we could give valuable marketing suggestions based on the result, which will help improve sales in the future.

# Analysis

## I. EDA Findings

For checking if there is a tendency over time shown by transaction data, a plot of money spent for each week was presented as below.

Figure 1 Money spent overview




It is obvious that no pattern exists which indicates if the customers spend less or more. Basically, the money spent by customers are quite even over time, although we can see that for several week, the sales were extremely high. By investigating into those abnormal weeks, I found that the weeks are close to festivals and the high sales phenomenon during the weeks happened both in 2016 and 2017. To conclude, money spent do not show a time pattern based on data.

After excluding time factor, I wanted to deepen my investigation and see if the difference of age range, maritage condition, income range, and size among household customers would have effect on the amount of money they spend. First, I wanted to take a glance at the distribution of money spent for each group divided by the previous factors. Figure 2 below was an example of the descriptive analysis of money spent for households of 5 different sizes. The mean seemed to be huge different which motivated me to do further analysis, although the large std (standard deviation) cast a concern about the normality of the samples, which might result in failure of statistical test.

Figure 2 Money spent for households of different size overview

| SIZE | mean | median | count_nonzero | std |
|---|---|---|---|---|
| 1 | 7124.877273 | 5034.535 | 22.0 | 8681.331549 |
| 2 | 4430.452857 | 2029.310 | 21.0 | 4694.871863 |
| 3 | 4966.613913 | 934.430 | 23.0 | 7111.578864 |
| 4 | 6042.765000 | 4620.485 | 4.0 | 6374.671240 |
| 5+ | 10672.836875 | 6153.235 | 16.0 | 12354.639633 |

I identified the values of the four factors as categorical values, since they were all divided into several baskets, so instead of seeking correlation between two numeric values, I wanted to figure out the association of these factors and money spent. In order to achieve this goal, I wanted to do ANOVA to see if the money spent are different for households divided into multiple groups by their features. Before doing so, I cleaned data and drop customers who have no listed information for the four factors. The actual data size for analysis was 86.

## II. Assumptions and Statistical Analysis

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. Here, we used One-Way ANOVA to check the impact of age range, maritage condition, income range, and size separately.

There are three basic assumptions used in ANOVA:

- the samples are independent
- the variances of all errors are equal to each other
- they are normally distributed

The first assumption is easy to satisfy since each household is supposed to be independent. As for the other two assumptions on the experimental errors, I would use Levene's test to test homogeneity of variance and Shapiro method to test normality if we could reject the null hypothesis of ANOVA.

The null hypothesis for ANOVA is that the means in all the groups were the same, which means the factor did not result in significant difference. The alternative hypothesis is that there was significant different among the means of all the groups.

We could define the following terms in order to the test:

- Sum of Squares A (sum of explanatory variable A's sum of squares)
- Sum of Squares Total (the sum of squared deviation of all data points from the distribution mean)
- Sum of Squares Error (Error (Residual) Sums of Squares)
- MSA (the variance of the group means around the grand mean)
- MSE (the variation of the errors around the group means)
- F ratio (=MSA/MSE)

After doing One-Way ANOVA analysis for each of the four factors, I found that all the realized F-statistics were small and the P-values were quite larger than 0.05, which indicated that we could not reject that all the means are not different. The F-statistics and P-values were listed in table 1. Since the F-statistics are not significant, I did not need to do the following ANOVA analysis.

Table 1 ANOVA table overview

| | F-statistics | P-values |
|---|---|---|
| Age range | 1.071 | 0.387 |
| Maritage | 0.1314 | 0.718 |
| Income range | 0.8251 | 0.535 |
| Size | 1.577 | 0.188 |

## III. Conclusion and Next Step

For the given data, I drew the conclusion that customers did not spend less or more as time went by and for the four main factors age range, maritage condition, income range and size, they did not make a big difference on the time spent by households.

Nevertheless, we could still see the small size of sample did affect the result of ANOVA and the normality of experimental errors. What we need to do next is to increase the sample size and collect data for more households. If sample size is large enough, due to the central limit theorem, the normality of errors can be assured.

Especially, the F-statistics of age range and size were relative larger than the other two factors which may indicate potential association, so I could do more analysis for these two factors and analyzed the model effect and influence power if given more data. Even more, we could investigate on whether the time pattern existed if we divided the customers based on the critical factors, which would give valuable point for the future advertisements and marketing.