# Tayko

*Data Mining Spring, 2022*

## Abstract

One of the largest concerns amongst all businesses is securing future clients with the correct advertising. Millions of dollars can be lost when spent on advertising to individuals who will never buy a product. Yet, much more can be lost when the correct target audiences are not met. This is why Tayko joined a consortium of firms to share customer pools to expand their reach while still ensuring some relevance to the individual. When pooling mailing groups and securing 200,000 individuals' information, Tayko wanted to make maximum profit. In order to calculate this Tayko first did a sample group of 20,000. In this they found the return rate of purchasers, which was quite low at 0.053, or 5.3%. In turn the decision was made to create a sample with an equal number of purchasers and non-purchasers to find those who are most likely to buy a product after receiving a catalog. Assumptions that have been made are that all addresses are valid, there are no repeat individuals and that this is the first time the individual is receiving the catalog.

## Introduction

Creating maximum profit with our data can be done by manipulating and investigating the data to find patterns and trends. What we plan to produce from this data are the clients that will most likely buy a product after receiving a catalog. This will be done by first creating an environment in which the data can be investigated for information. This will include studying the data's attributes, cleaning the data, feature selections, and normalizing features.

## Attribute Inspection

Source a - w -

Source catalog for the record (15 possible sources). 1: yes, 0: no.

Freq -

Number of transactions in last year at source catalog

mean 1.417

std 1.405738

min 0.00

25% 1.00

50% 1.00

75% 2.00

max 15.00

Last_update_days_ago -

How many days ago was last update to customer record

mean 2155.101

std 1141.302846

min 1.00

25% 1133.00

50% 2280.00

75% 3139.25

max 4188.00

1st_update_days_ago -

How many days ago was last update to customer record

mean 2435.6015

std 1077.872233

min 1.00

25% 1671.25

50% 2721.00

75% 3353.00

max 4188

Web_order -

Customer placed at least 1 order via web

1: yes, 0: no

Mean 0.426

Gender=male -

The value of this attribute is either 1 or 0. 1 represents that the user is a male, 0 represents the user is not a male (is a female).

Mean 0.5245

Address_is_res -

The value of this attribute is either 1 or 0. 1 represents the address being a resident address, 0 represents it being something other than a resident address. As mentioned above, we will be assuming that all addresses are valid, thus there are no further steps to take to clean this data.

Mean 0.221

Purchase -

The value of this attribute is either 1 or 0. 1 represents the user has purchased something, 0 represents the user has not.

Mean 0.5

Spending -

The value of this attribute is a float that represents the amount spent on a purchase. If they did not make a purchase, the value is 0.

mean 102.62500

std 186.78261

min 0.00000

25% 0.00000

50% 2.00000

75% 153.00000

max 1500.00000

%0 0

Type float

Partition -

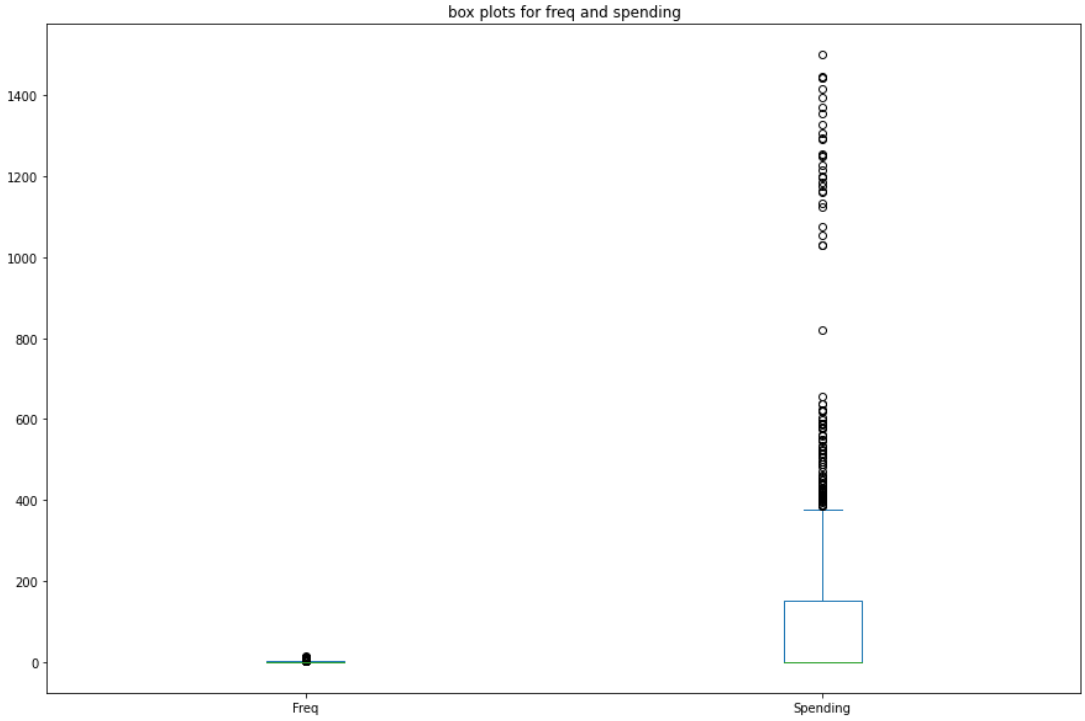
Variable indicating which partition the record will be assigned to. p1: training p2: validation

freq 1500

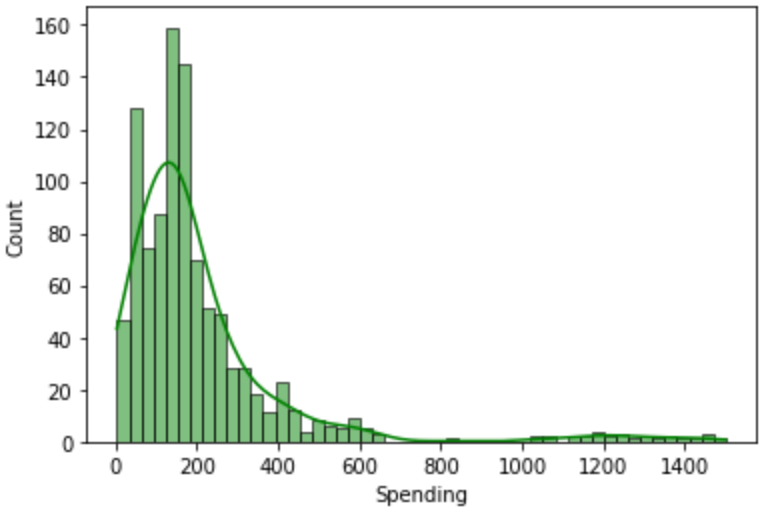
Top P1

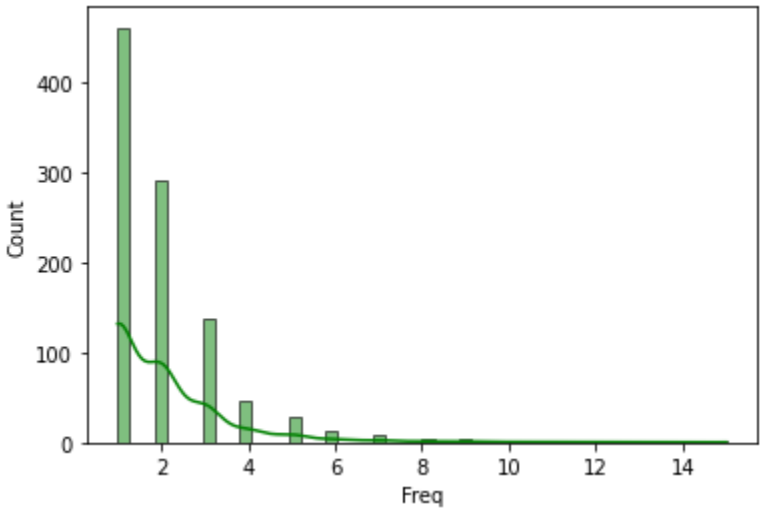
Type object

## Data Cleaning




We had two attributes in which outliers were present, Freq and Spending. However after initially studying those columns we came to realize that those outliers hold a lot of information and importance, and as it will be shown later in this report, both these attributes are highly correlated with each other. One way to explain the presence of outliers in these columns is to consider the high spenders as 'whales'. Whales mean that they are very big spenders, and this phenomenon is common when it comes to consumers. A company will usually have just a few very big spenders who stand out from the rest of the pack and it would be wrong to simply change their values as it would eliminate an entire group of people from our dataset. Thus, we elected to bin these two columns into 4 categories each. The choice for the size of our bins was made based on studying the attributes and graphs of the two features. Essentially, they are split as follows: the non-spenders, low spenders, big spenders, and the whales. The reason for and values of the bins will be explained when we talk about our model below. However, when we were binning our spending column we realized that there were only 999 people who had spending equal to zero. Since we know that our dataset has 1000 non purchasers, we had an odd one out. To fix this issue we found the odd one out and set their spending to zero.




We decided to drop ‘1st_update_days_ago’ and ‘Last_update_days_ago’ as common sense dictates these are irrelevant for our results. These attributes are meant to indicate how long ago the system updated the data for a specific user, thus it does not have any bearing on or relation to the user’s behavior. Additionally, we elected to remove the ‘Purchase’ column as this information will also be indicated in the ‘Spending’ column on which we intend to focus (i.e. if the customer did not make a purchase, then ‘Spending’ will be 0). This leaves our attributes to be ‘Freq’, ‘ Web_order’, ‘Gender=male’, ‘Address_is_res’, and ‘Spending’. Lastly, we removed the partition column as it was something that someone else created and that we were not going to use.

We also wanted to group all of our source_* columns into one column. The reason for this is because every row only belonged to one source_* column and never two. This will ultimately come in handy when we do our Categorical Naive Bayes classification later on and in doing so we realized that

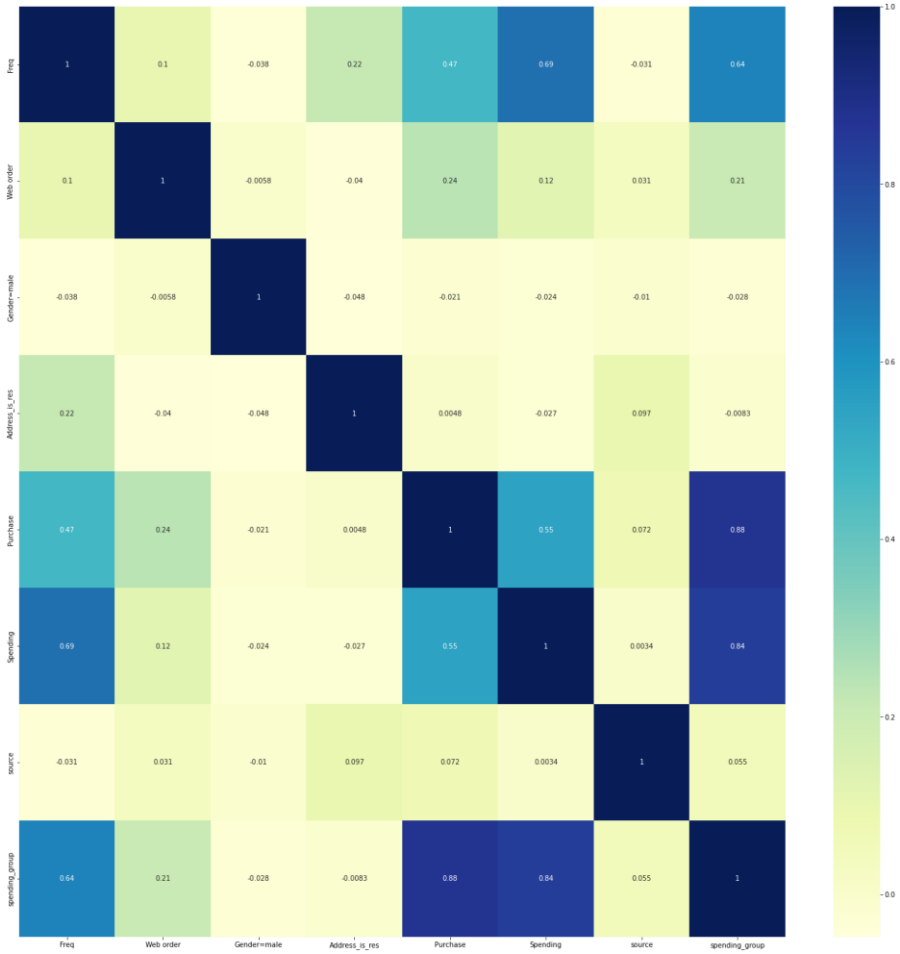
some rows didn't belong to any of the source_* columns provided. In this case, we condired them to be our 'default' source and named them as such in our new aggregated source column.

## Feature Selection

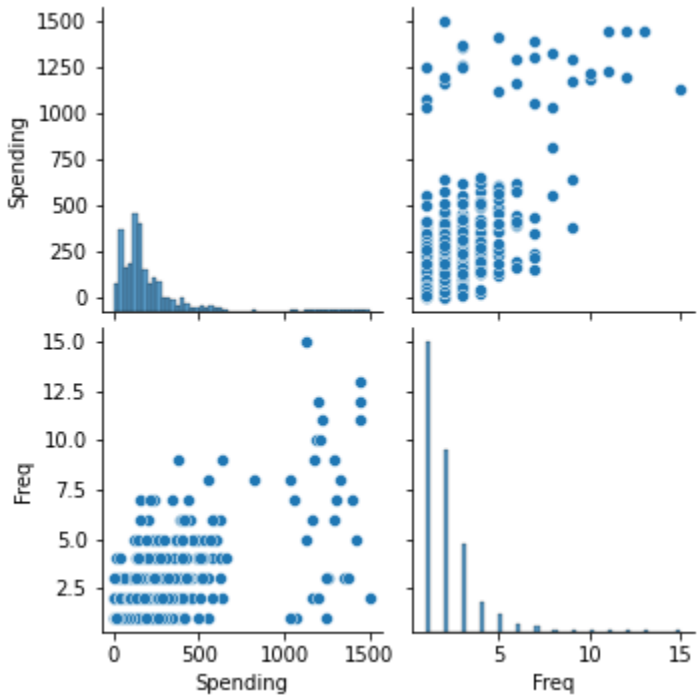
We will use different methods of feature selections to find the best one. We will start with information gain, which calculates the reduction in entropy from the transformation of a dataset. It is used by evaluating the Information gain of each variable in the context of the target variable. The information gain is given by the following equation:

Gain(A) = Info(D) - InfoA(D)

To further aid our decision, we created a heatmap to visualize the relationship between different variables.



The image above is the resulting heatmap. Each square represents the relation between two variables, resulting in a triangular matrix. We are interested in the relation specifically between any given factor and spending, so we only need to look at the 'Spending' column. As indicated in the image, frequency and spending seem to have the highest correlation, which is further supported by various other plots.




## Methods

We decided to use two models, both Bayesian. The first which we tested was the classic Naive Bayes Model. The second was Bernoulli Naive Bayes.

### Class Labels

In order to use the Naive Bayes theorem on our data, it is necessary to create class labels using bins. This resulted in us creating four bins for both the 'Spending' category and the 'Freq' category. Within 'Spending', the first group is those that do not spend anything, the second being the group that depends 1 to 300 dollars, the third being the group that spends 301 to a thousand dollars and the fourth being the spenders that spend more than a thousand dollars. Our 'Freq' data was binned into 0, 1 to 2, 3 to 7, and greater than 7.

| binned | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **0** | 1000.0 | 0.76 | 0.73 | 0.0 | 0.0 | 1.0 | 1.0 | 4.0 |
| **1-300** | 837.0 | 1.71 | 0.95 | 1.0 | 1.0 | 1.0 | 2.0 | 7.0 |
| **301-1000** | 134.0 | 3.38 | 1.71 | 1.0 | 2.0 | 3.0 | 4.0 | 9.0 |
| **>1000** | 29.0 | 6.62 | 4.07 | 1.0 | 3.0 | 7.0 | 10.0 | 15.0 |

### Categorical Naive Bayes

We did categorical Naive Bayes calculations with our values we labeled as categorical. This resulted in a prediction accuracy of 0.67 (+/- 0.05).

### Bernoulli Naive Bayes

In addition to categorical Naive Bayes we also did Bernoulli Naive Bayes calculations which resulted in an accuracy of 0.57 (+/-0.05). Bernoulli assumes that all values are binary, thus we changed the data again, shown by the table below.

| source_r | source_s | ... | source_u | source_p | source_x | source_w | Web order | Gender=male | Address_is_res | Spen_binned | Freq_bin | predict |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | ... | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1-300 | 1-2 | 1-300 |
| 0 | 0 | ... | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1-300 |
| 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1-300 | 1-2 | 0 |
| 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1-2 | 0 |
| 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1-2 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 0 | 0 | ... | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1-300 | 1-2 | 1-300 |
| 0 | 1 | ... | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1-300 | 1-2 | 1-300 |
| 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | ... | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1-2 | 0 |
| 1 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1-2 | 0 |

## Results

In our calculations we found that the most correlated features were spending and frequency. This was found initially using heat mapping. From there we confirmed the findings using Naive and

Bernoulli Bayes Classification , as well as using benchmark, chi-square, mutual information modeling. From this we confidently say that those who have interacted the most with other sources will be most likely to buy a product and spend more. Those with higher frequencies are customers whose information was provided from another source and who had already interacted with a purchase from the other source. This would result in the highest return on investment.

The following are tables showing the results we obtained from using the Categorical Naive Bayes algorithm. Each table compares our model's prediction on spending and compares it to a given feature.

## Predictive Tables

Table 1 and 2: Prediction vs. Source_*

| | count | unique | top | freq |
|---|---|---|---|---|
| predict | | | | |
| **0** | 1123 | 16 | E | 187 |
| **1-300** | 861 | 16 | A | 148 |
| **301-1000** | 3 | 3 | R | 1 |
| **>1000** | 13 | 6 | A | 7 |

| sourceALL | count | unique | top | freq |
|---|---|---|---|---|
| A | 253 | 3 | 1-300 | 148 |
| B | 120 | 2 | 0 | 80 |
| C | 112 | 4 | 0 | 56 |
| D | 83 | 4 | 0 | 42 |
| E | 302 | 3 | 0 | 187 |
| H | 105 | 2 | 0 | 65 |
| M | 33 | 2 | 0 | 21 |
| O | 67 | 2 | 0 | 55 |
| P | 12 | 2 | 1-300 | 10 |
| R | 137 | 4 | 0 | 74 |
| S | 94 | 2 | 0 | 64 |
| T | 43 | 2 | 1-300 | 28 |
| U | 238 | 3 | 1-300 | 121 |
| W | 275 | 2 | 0 | 150 |
| X | 36 | 2 | 0 | 22 |
| default | 90 | 2 | 0 | 77 |

Tables 1 and 2 are our model's predictions on how much spending each source would generate. At a first glance we could clearly see that source A would be the most stable source in terms of generating revenue. We say this because of the 13 who spent >1000 dollars 7 of them belong to source A as well as having the highest amount of spenders between the 1 to 300$ range. However, looking at the second table clearly shows us that there are other valuable sources to consider. For example, source P has 12 spenders and of those 12, 10 of them spent something. That is the highest percentage of spenders within a source, but it is important to keep in mind that we only have 12 data points for source P and would probably want more to accurately give this conclusion.

Table 3: Prediction vs. Web order

| Web order | count | unique | top | freq |
|---|---|---|---|---|
| **0** | 1148 | 3 | 0 | 1016 |
| **1** | 852 | 4 | 1-300 | 736 |

Table 3 emphasizes the importance of web orders given that the distribution between the two is fairly even. Those who purchase through online means are more likely to buy our product compared to those who come in stores.

Table 4: Prediction vs. Freq_bin

| Freq_bin | count | unique | top | freq |
|---|---|---|---|---|
| **0** | 398 | 1 | 0 | 398 |
| **1-2** | 1340 | 2 | 0 | 725 |
| **3-7** | 246 | 1 | 1-300 | 246 |
| **>7** | 16 | 2 | >1000 | 13 |

Table 4 reaffirms the correlation we found early on with our heat map. The more products an individual or business buys, the higher the spending will be on their account. This shows that customers who have made purchases in the past are more likely to spend in the future. Furthermore, our model tells us that those who have a frequency of 3 to 7 will spend between 1 and 300 dollars and those who have a frequency higher than 7 will spend over 1000 dollars. For the ones at or below a frequency of 2 are predicted to not spend a dime.

Table 5: Prediction vs. Address_is_res

| Address_is_res | count | unique | top | freq |
|---|---|---|---|---|
| 0 | 1558 | 3 | 0 | 899 |
| 1 | 442 | 3 | 0 | 224 |

There is no significant difference between an address being residential versus not being residential.

Table 6: Prediction vs. Gender=male

| Gender=male | count | unique | top | freq |
|---|---|---|---|---|
| 0 | 951 | 4 | 0 | 533 |
| 1 | 1049 | 3 | 0 | 590 |

There is no significant difference between a customer being a male versus not being a female.

## Conclusion and Discussion

In conclusion, our study found that the best candidates to target for catalog distribution would be a person of any gender who has already made at least three purchases. To further increase the possibility of higher profits, this person should come from Source A. Additionally, we found that since online ordering generates more revenue, Tayko should focus on their online audience. In other words, they should aim to develop their online presence whilst reducing their retailer presence in order to make more and save more overall. In terms of advertising, Tayko may want to look at ways of engaging new customers since most of their revenue comes from previous customers. This would allow for them to grow their customer base and thus generate more revenue over the long run. Overall, they have good customer retention and should aim to bring more onboard by targeting the sources that are predicted to generate revenue for them.

Numerous limitations are important to recognize. Firstly, our study is completely limited to Naive Bayes techniques and has a fairly low accuracy of 67%. This may be fixed by using different methods to have a model that is better able to predict which spending categories a customer will fall in.

However, we do believe that our current model does a good job of showing the initial picture and the results we have derived from our model will help Tayko succeed in the future.