Final Report

BSAN 6050 - Customer Relationship Management Analytics

Team Members: Megan Rice , Alena Sanchez, Muhammad Aoun, Oliver Lin

Instructor: Prof. Wang Sijun

Loyola Marymount University

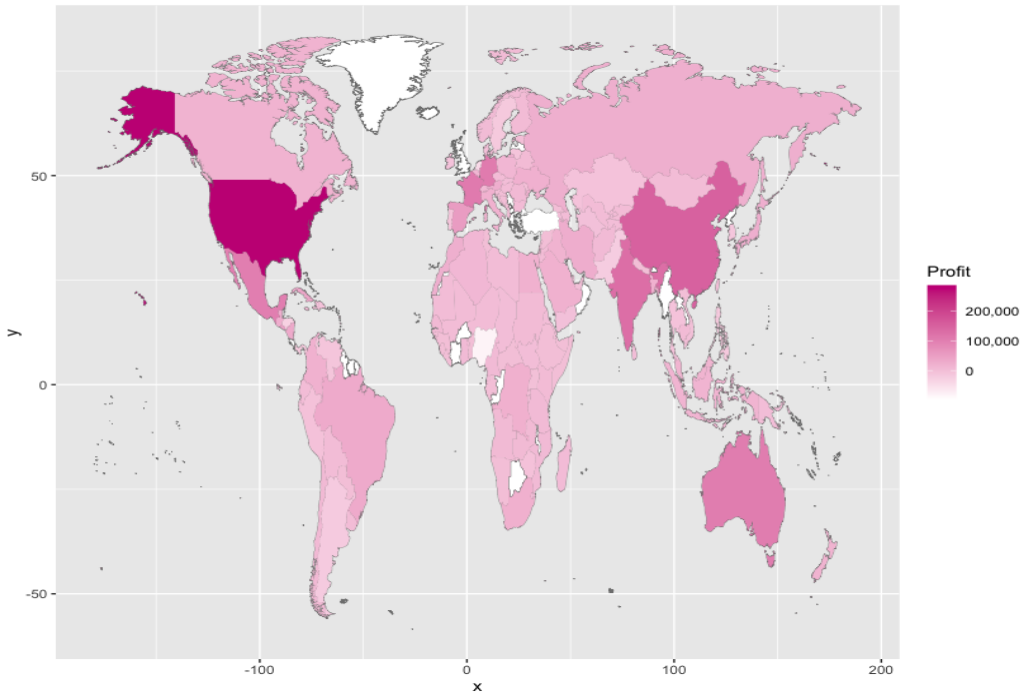
Summary

With our analysis of Office Supply Company, we intended our findings would better advise and equip the managers moving into the future. During our exploration of the company data, we exhausted many avenues of data analysis, including survival analysis, customer lifetime value calculations, logistic regression, and market basket analysis. While some of the analyses gave more insights than others, we believe our findings are impactful and give managers tangible recommendations. Throughout the inquiry, we discovered two main business problems. The company's most valuable customers have an overall lower survival rate compared to lesser valuable customers and managers are not maximizing profitable products and sub-categories effectively. To absolve these problems, we are prescribing a two-fold plan. The first part involves designing a more strategic discount model to address the randomness in the current model and prioritize the company's relationship with high-value customers. Customers are not equal, and therefore, should not be treated equally. Treatment, in this case, involves discounts given. Part two consists of implementing a product recommendation system for the website and using product relationships to develop target marketing campaigns. By recommending associated products frequently bought together, managers can organically increase the number of items purchased. Finally, an additional recommendation we discovered was that there are products purchased at a much higher frequency than all other products. This warrants special attention from managers. Managers should look into their profit margins and put high importance on the stock of these items, as they have a substantial impact on the bottom line.

Part 1 Descriptive Insights - Data Overview

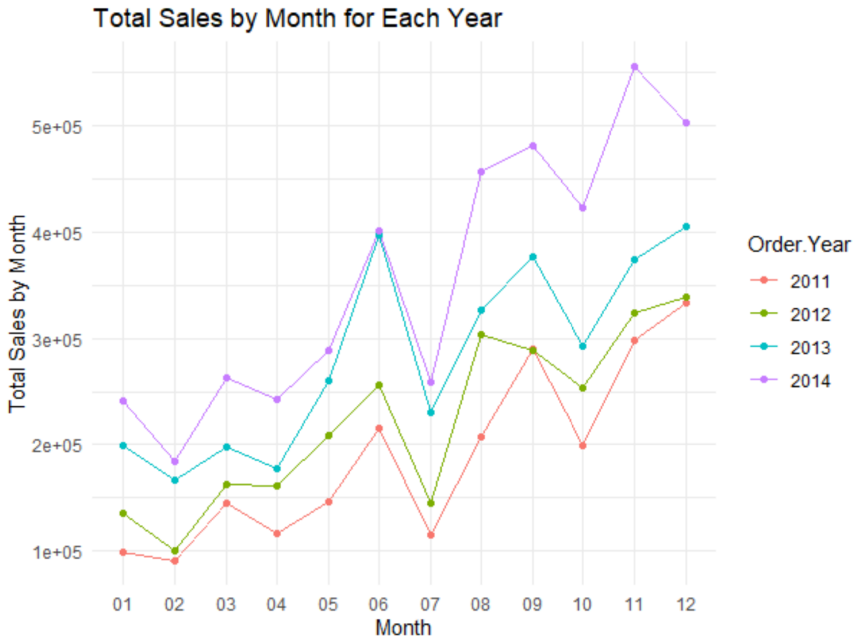
Office Supply Company sells various products ranging from paperclips to machines such as printers. This dataset contains about 50,000 records and 24 different variables that allowed us to dive into Customer Insights and Product Strategy.

- Global Presence



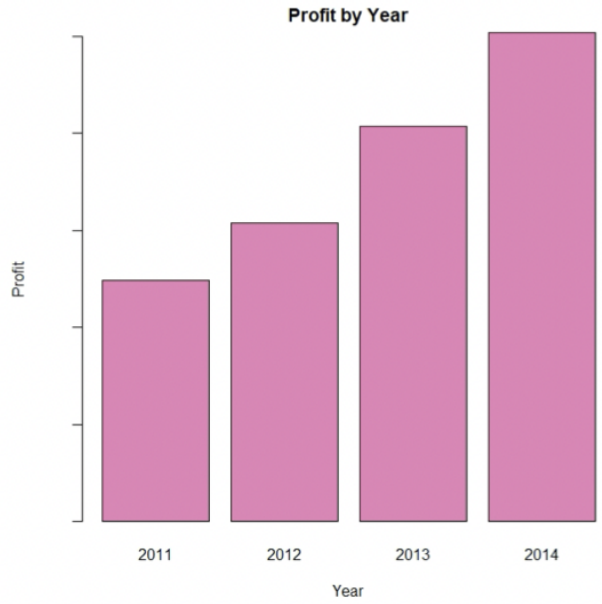
The company has a global presence, with the US being our largest market.

- Increasing Sales and Cyclicity

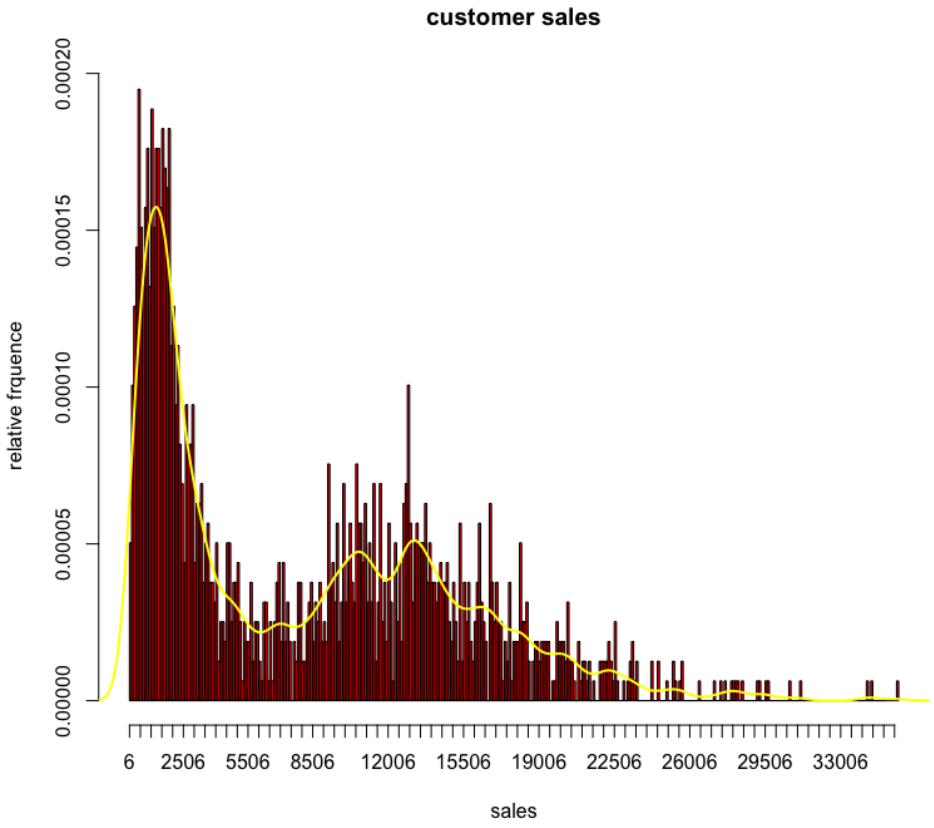


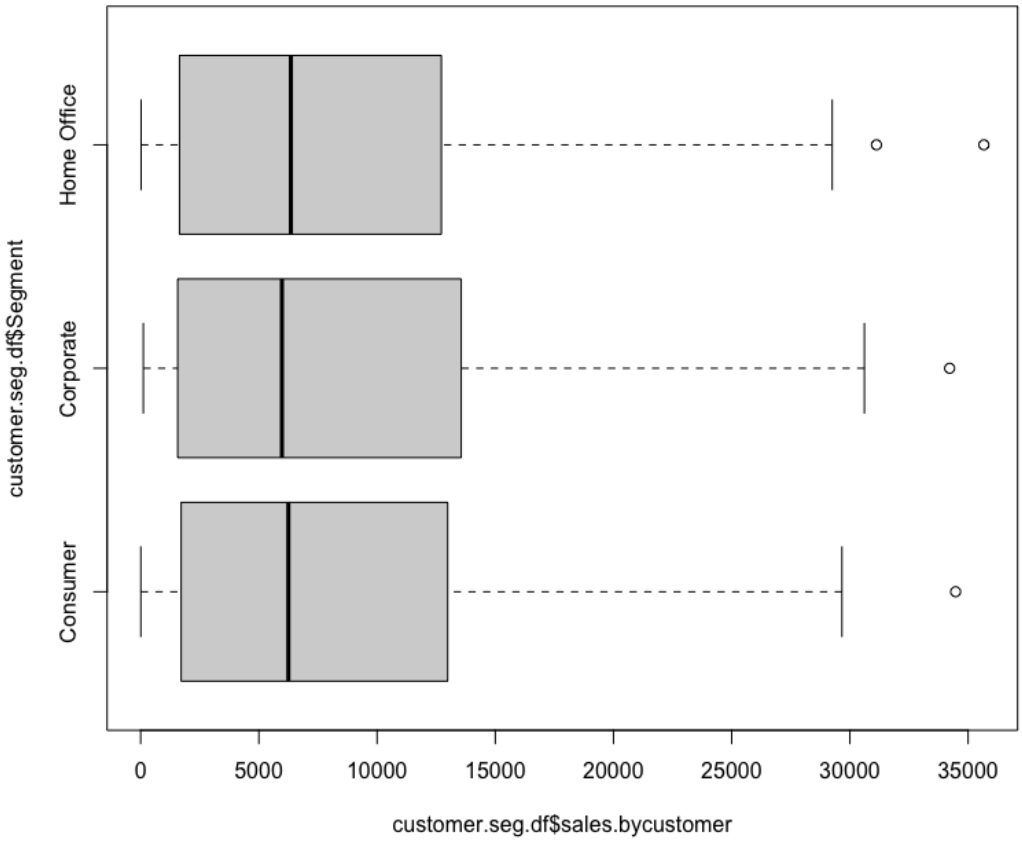
As shown in the above line chart, we can see that our company has seen consistent increase in sales from 2011 to 2014. And we also observe a clear cyclical pattern in our

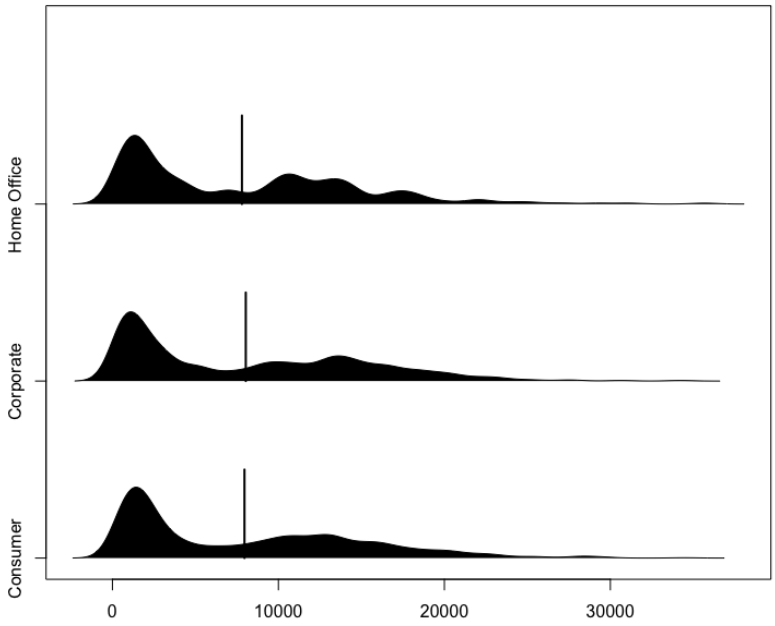
sales. We usually see a large increase in sales in June, which is almost always followed by an immediate decrease in July.

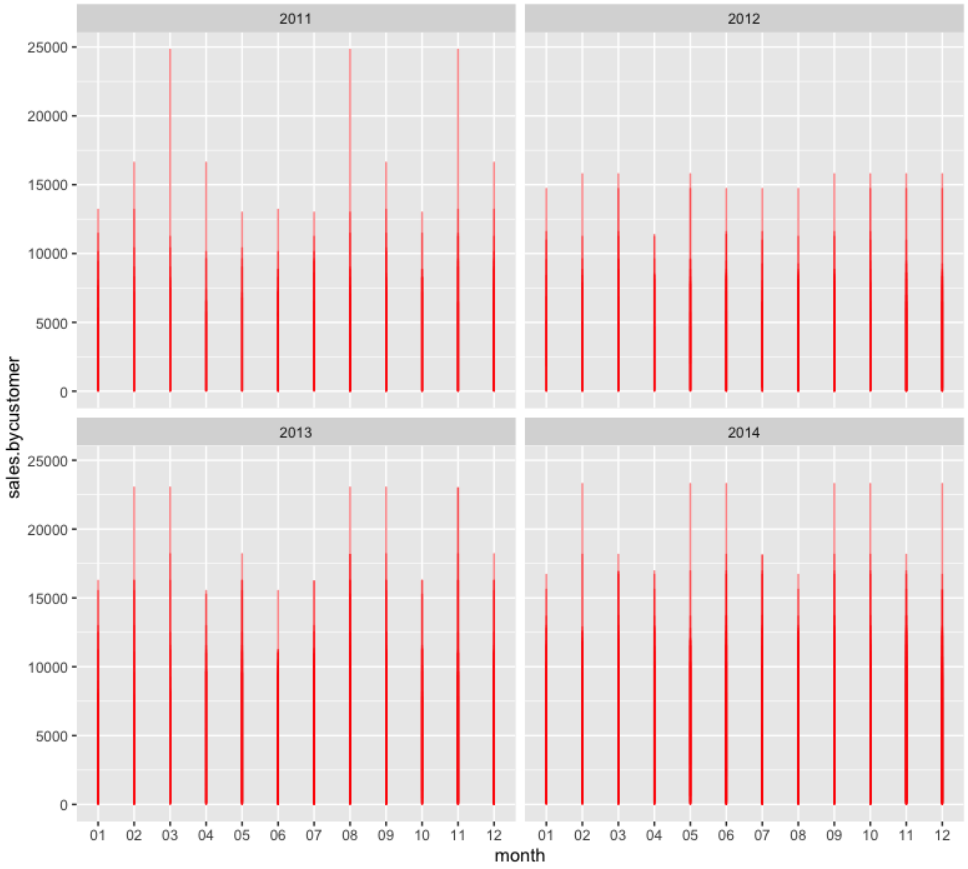


Furthermore, with the consistent increase in sales, it was found that the total profit also increases steadily each year, as seen in the figure above.


customer sales
relative frquence
0.00000
0.00005
0.00010
0.00015
0.00020
6
2506
5506
8506
12006
15506
19006
22506
26006
29506
33006
sales


customer.seg.df$Segment
Home Office
Corporate
Consumer
0
5000
10000
15000
20000
25000
30000
35000
customer.seg.df$sales.bycustomer


Home Office
Corporate
Consumer
0
10000
20000
30000


2011
2012
2013
2014
sales.bycustomer
25000
20000
15000
10000
5000
0
01
02
03
04
05
06
07
08
09
10
11
12
month

Part 2 Modelling-based Insights

Survival Analysis

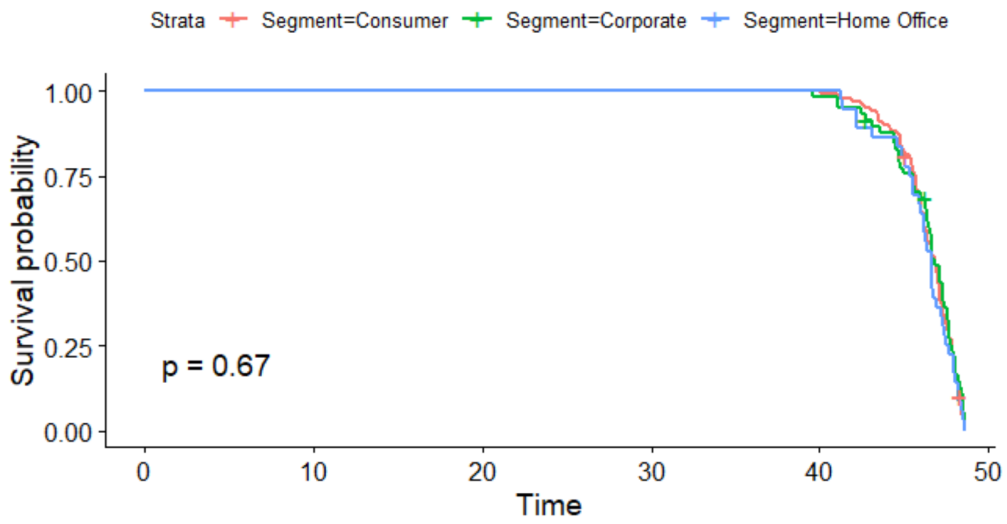
To conduct survival analysis, we first ran a series of codes to create a data frame containing all the customers with their unique customer names, and many aggregated data such as average sales, average profit etc. Survival analysis requires data on customer churn, thus we created a column containing the date of the last order for each customer, and then calculated days since the last order. By creating a frequency plot based on that information, our group decided that the cut off time for churning customers would be 100 days.

Using that cut-off, we created a churn column with values of 1 and 0 to indicate whether or not a customer has churned.

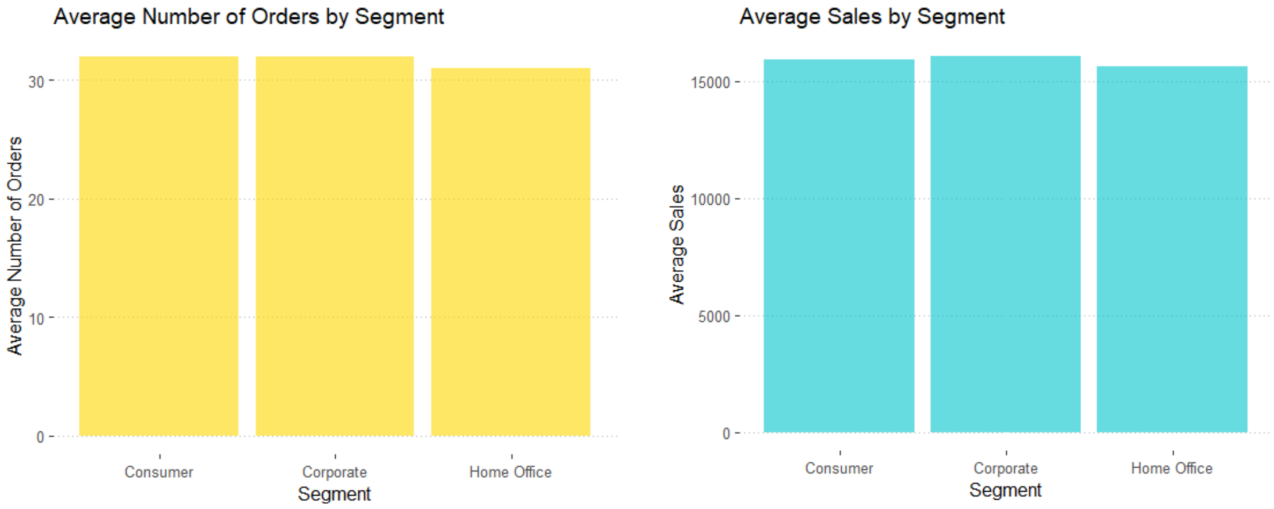
K-M Test:

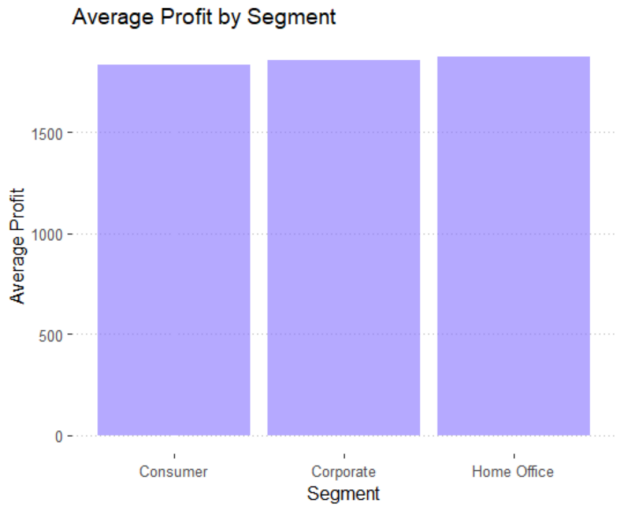
- Segment

For the first step of survival analysis, we want to conduct a K-M test. Since the K-M test only uses categorical variables, we considered using Segment as a criteria. However, the test does not result in any significant insights.



After seeing this result, we quickly took a look at the orders, sales and profit information for each customer segment. And the result explains why we cannot find any significance.





According to these visualizations, all three different customer segments seem to share an identical number of orders, average sales and average profit.
Unfortunately, there is not too much useful categorical data we can use for the purpose of the K-M test. There is geographical data in the data set, however, after a simple inspection we noticed that every customer seems to place orders from many different countries and regions, even markets. As a result, we decided to create our own categorical data.

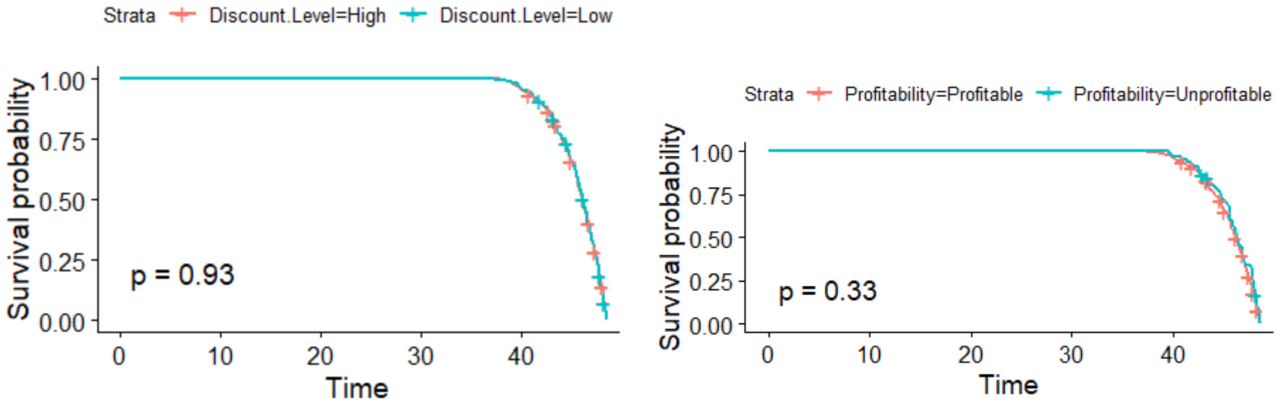
- Creating other categorical variables

We created a customer class variable based on how frequently each customer places orders. To do that, we first had to find out how many separate orders each customer creates and sum up the number of days between orders for every individual customer. And then we used the cut() function to divide all customers into 4 different classes from A to D based on the 25th, 50th and 75th quantiles, class A being those who place orders the most frequently and class D being those who order the least frequently. (This customer classification was used in customer lifetime value analysis).
In addition to the customer class, we also wanted to look into discount as a factor. And to convert discount, which is continuous, to a categorical variable, we simply divided all discounts by the median and called the top 50% “high” and bottom 50% “low”.
One other variable we created is profitability, which classifies customers into “profitable” and “unprofitable” based on the sum of each customer’s total profit.
Here is a sample of the customer data frame:

```
       Customer.Name Sum.Order.Gap.Adj Count.True Order.Freq Sum.Sale Average.Sale Sum.Profit
1      Aaron Bergman              1395         37         38 24644.63          666    4683.21
2      Aaron Hawkins              1337         34         39 20759.51          611    2450.93
3     Aaron Smayling              1358         31         44 14212.63          458     369.16
4    Adam Bellavance              1419         39         36 20186.78          518    4979.98
5          Adam Hart              1298         40         32 21718.20          543    1902.03
6 Adam Shillingsburg              1350         34         40 15444.68          454    1421.27
  Profit.Margin Profitability Customer.Class First.Order.Date Last.Order.Date
1    0.19002963    Profitable              A       2011-02-19      2014-12-15
2    0.11806300    Profitable              B       2011-04-22      2014-12-19
3    0.02597408    Profitable              C       2011-03-21      2014-12-08
4    0.24669511    Profitable              A       2011-01-07      2014-11-26
5    0.08757770    Profitable              A       2011-06-10      2014-12-29
6    0.09202327    Profitable              B       2011-04-06      2014-12-16
  Days.Since.Last.Order Average.Discount Discount.Level Sum.Discount Average.Discount.Amount
1                    16        0.1101124            Low         3799                       43
2                    12        0.1609286           High         4294                       77
3                    23        0.1676667           High         3679                       61
4                    35        0.1317647            Low         2783                       41
5                     2        0.0932381            Low         3746                       45
6                    15        0.1634085           High         4001                       56
      Segment Followtime Churn Survival  Pseudos id
1    Consumer       1411     1 47.03333 47.03681  1
2   Corporate       1349     1 44.96667 44.95139  2
3   Corporate       1381     1 46.03333 46.02743  3
4 Home Office       1454     1 48.46667 48.02172  4
5   Corporate       1300     1 43.33333 43.31685  5
6    Consumer       1365     1 45.50000 45.48740  6
```

- Discount Level & Profitability:

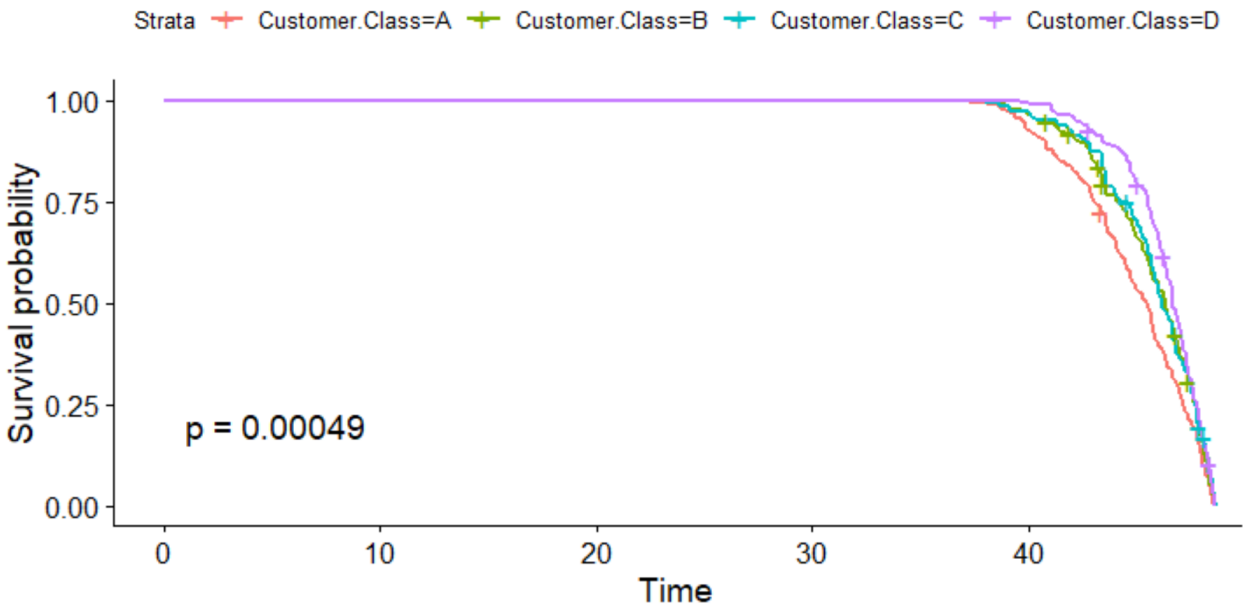
We ran the K-M test based on discount level and profitability. Surprisingly either showed any statistical significance.



We expected the discount level to have some influence over customer survival rate, but the result showed otherwise. So we set out to find out why this is the case. However, before that we wanted to see if segment and discount had a different effect based on different customer classes.

- Customer Class (aka Order Frequency)

The K-M test for customer class actually showed some significant results.



The graph shows a very distinct divergence in the survival rate of the four customer classes. Nevertheless, what it means was quite puzzling. The K-M test suggests that class A customer's survival rate actually drops off much faster than other classes, and Class D customer in fact has the highest survival rate. This indicates that the company is not capable of retaining its most valuable customer (confirmed in CLV analysis).

Pseudo Test:

The pseudo test showed some statistical significance in two variables, order counts and order frequency. On the other hand, discount, either in discount or dollar amount, does not show any significance.

- Order Counts

```
Call:
geese(formula = Pseudos ~ Count.True, id = id, data = customer.df,
    family = gaussian, scale.fix = FALSE, corstr = "independence",
    jack = TRUE)

Mean Model:
 Mean Link:                identity
 Variance to Mean Relation: gaussian

 Coefficients:
             estimate     san.se     ajs.se        wald           p
(Intercept) 44.28558106 0.51444062 0.51579621 7410.613928 0.000000000
Count.True   0.04151256 0.01577651 0.01581993    6.923681 0.008506173
```

The results suggest that with every additional order, a customer is likely to stay with our company for 0.04 month or 1.2 days longer. This is expected since a customer who orders more is likely to be more loyal to us.

- Order Frequency

```
Call:
geese(formula = Pseudos ~ Order.Freq, id = id, data = customer.df,
    family = gaussian, scale.fix = FALSE, corstr = "independence",
    jack = TRUE)

Mean Model:
 Mean Link:                identity
 Variance to Mean Relation: gaussian

 Coefficients:
              estimate     san.se     ajs.se       wald           p
(Intercept) 42.67457296 0.50177704 0.50457568 7232.97225 0.000000e+00
Order.Freq   0.06702975 0.01096564 0.01103649   37.36521 9.795393e-10
```

Order frequency is, in other words, average days between each order. This is the criteria that we based upon for the customer classes. The result from this pseudo test matches the K-M test result. The pseudo test result for order frequency suggests that, for every additional day between each order from a customer, they are likely to survive 0.06 days or 1.8 days.

- Discount

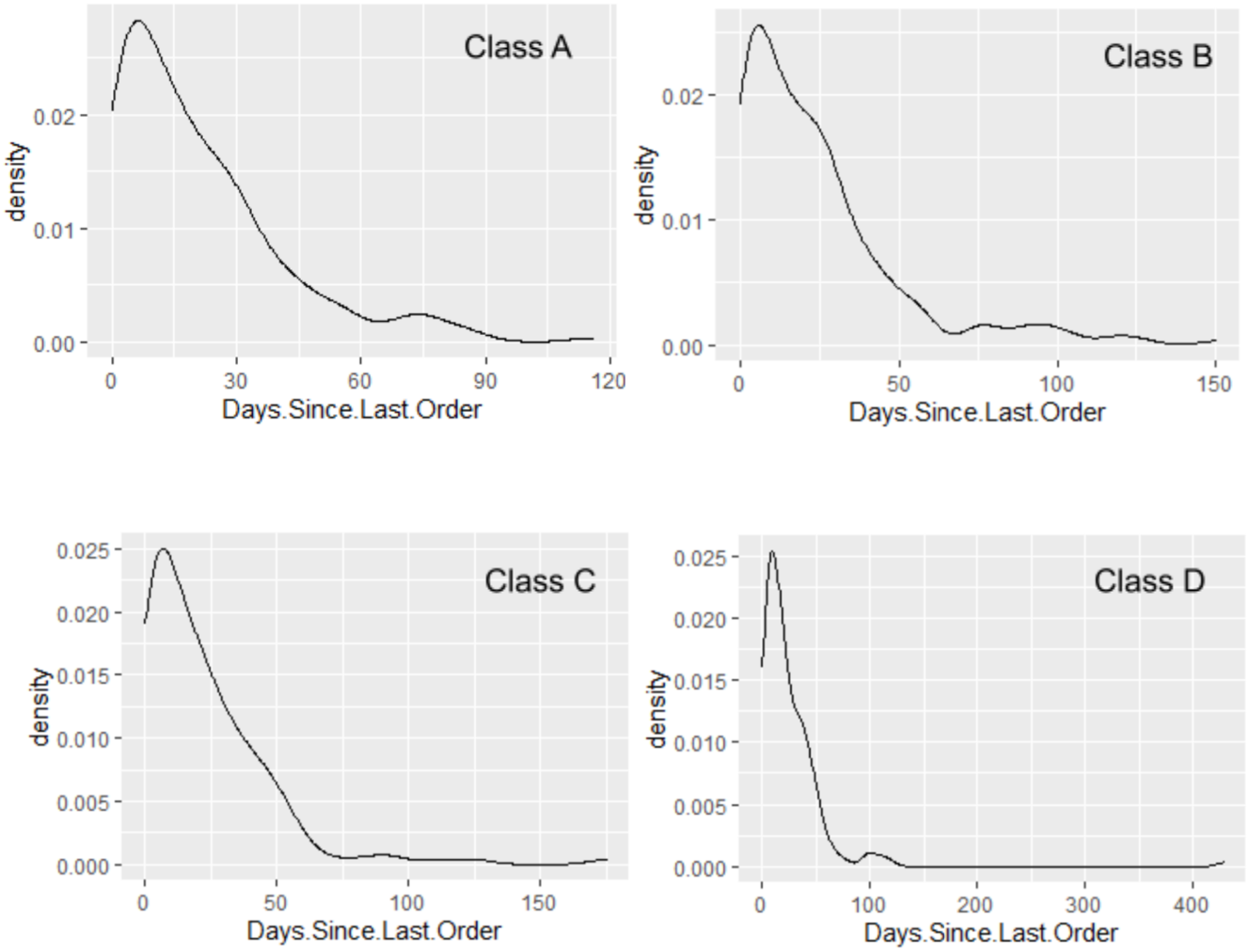
```
Call:
geese(formula = Pseudos ~ Average.Discount, id = id, data = customer.df,
    family = gaussian, scale.fix = FALSE, corstr = "independence",
    jack = TRUE)

Mean Model:
 Mean Link:                identity
 Variance to Mean Relation: gaussian

 Coefficients:
                estimate    san.se    ajs.se        wald        p
(Intercept)      45.833711 0.2783033 0.2791417 2.712273e+04 0.0000000
Average.Discount -1.634209 1.8930548 1.8998905 7.452278e-01 0.3879914
```

As expected from the K-M test, the pseudo test on average discount in dollar amount showed no statistical significance. This, in combination with the K-M test result, suggests that there is a serious problem with our discount model.

Survival Analysis for Each Customer Class:

Before looking into discounts, we decided to see if some of the segments and other variables make a difference within each customer class.

We used the filter() function to create 4 subsets of the original customer data frame. And then we ran similar analyses on the 4 subset data frames. The result of the “days since last order” density chart was very interesting.



The density chart showed that most class A customers placed an order about 60 days ago, and that number increased to around 100 days for class D. This was a promising distinction for different customer classes, but unfortunately when conducting K-M test and pseudo test, most results were not too interesting, but one thing that is statistically significant and somewhat unusual is the number of orders for each customer class.

```
Call:
geese(formula = Pseudos ~ Count.True, id = id, data = classD.df,
    family = gaussian, scale.fix = FALSE, corstr = "independence",
    jack = TRUE)

Mean Model:                                          Class A
 Mean Link:                 identity
 Variance to Mean Relation: gaussian

 Coefficients:
             estimate    san.se     ajs.se      wald           p
(Intercept) 39.2225162 1.6184741 1.67644961 587.29915 0.000000e+00
Count.True   0.2773215 0.0608789 0.06312152  20.75074 5.231123e-06
```

```
Call:
geese(formula = Pseudos ~ Count.True, id = id, data = classB.df,
    family = gaussian, scale.fix = FALSE, corstr = "independence",
    jack = TRUE)
                                                     Class B
Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian

 Coefficients:
             estimate     san.se     ajs.se      wald p
(Intercept) 20.3787022 2.35253218 2.43295036  75.03811 0
Count.True   0.7719307 0.07004575 0.07246353 121.44875 0
```

```
Call:
geese(formula = Pseudos ~ Count.True, id = id, data = classC.df,
    family = gaussian, scale.fix = FALSE, corstr = "independence",
    jack = TRUE)
                                                     Class C
Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian

 Coefficients:
            estimate     san.se     ajs.se      wald           p
(Intercept) 16.579959 2.16435845 2.22087496  58.68249 1.854072e-14
Count.True   0.987228 0.07161479 0.07348453 190.03323 0.000000e+00
```

```
Call:
geese(formula = Pseudos ~ Count.True, id = id, data = classD.df,
    family = gaussian, scale.fix = FALSE, corstr = "independence",
    jack = TRUE)
                                                     Class D
Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian

 Coefficients:
             estimate    san.se     ajs.se      wald           p
(Intercept) 39.2225162 1.6184741 1.67644961 587.29915 0.000000e+00
Count.True   0.2773215 0.0608789 0.06312152  20.75074 5.231123e-06
```

it seems that by dividing customer class and conducting pseudo tests, we see a large difference in customer survival length. Clearly, class D customers have much longer survival time than all other classes. An additional order seems to have more effect on survival time for classes that have a lower average survival time.

We initially hoped to see some statistical significance once we conducted the pseudo test on discount for separated customer classes. However, there was still nothing worth noting.

Analysis on Discount:

After realizing the strangeness in discount, we decided to look deeper into discount and attempt to find out why discount has no influence over customer's survival.

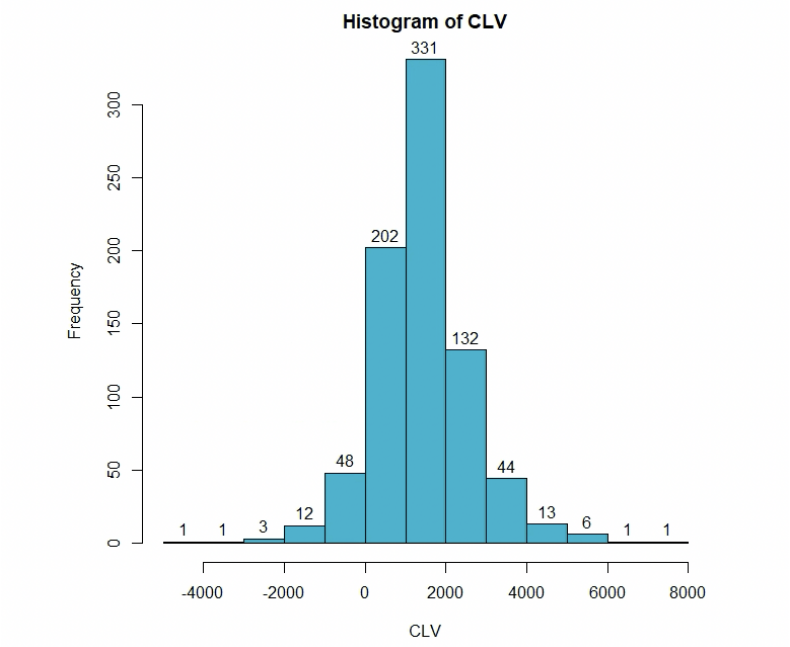
First, we created a bar chart of discount levels (all discounts divided in two two categories by the median) by customer class. The graph shows that all four customer classes seem to have the same ratio of high and low discount level. And class D customers seem to have a slightly higher amount of high discount than low discount.

Then, we looked at the average discount in dollar amount for each of the customer classes.

This bar chart indicates that each customer class is receiving almost the identical amount of discount. And class D customers are receiving slightly more discounts than other three classes of customers.
These analyses reflect that the current discount system does not have a clear strategy or plan, it seems to be offering discounts to customers randomly.

Customer Lifetime Value

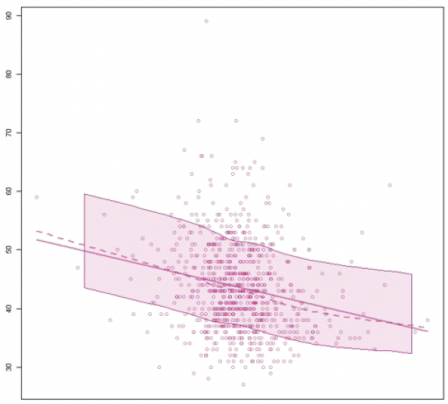
Next, we will now dive a bit deeper into our customer insights and discuss Customer Lifetime Value (CLV). By determining the lifetime value of the customers this allows the company to prioritize their spend against high value customers to optimize costs for a greater return. The CLV was calculated utilizing the past-value method by determining the Net Present Value, assuming a discount rate of 10% for each customer throughout the past four years.

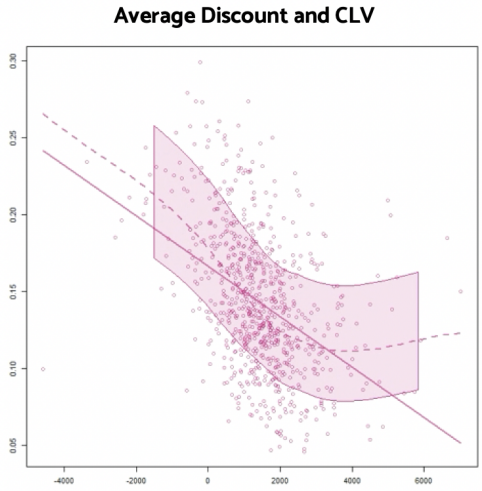


In the figure above, we can see that CLV is normally distributed for Office Supply Co., and with a simple calculation utilizing CLV it was found that the top 25% percent of customers contribute a greater amount of return compared to the bottom 75% of customers; approximately $16,800 more was returned from the top 25% compared to

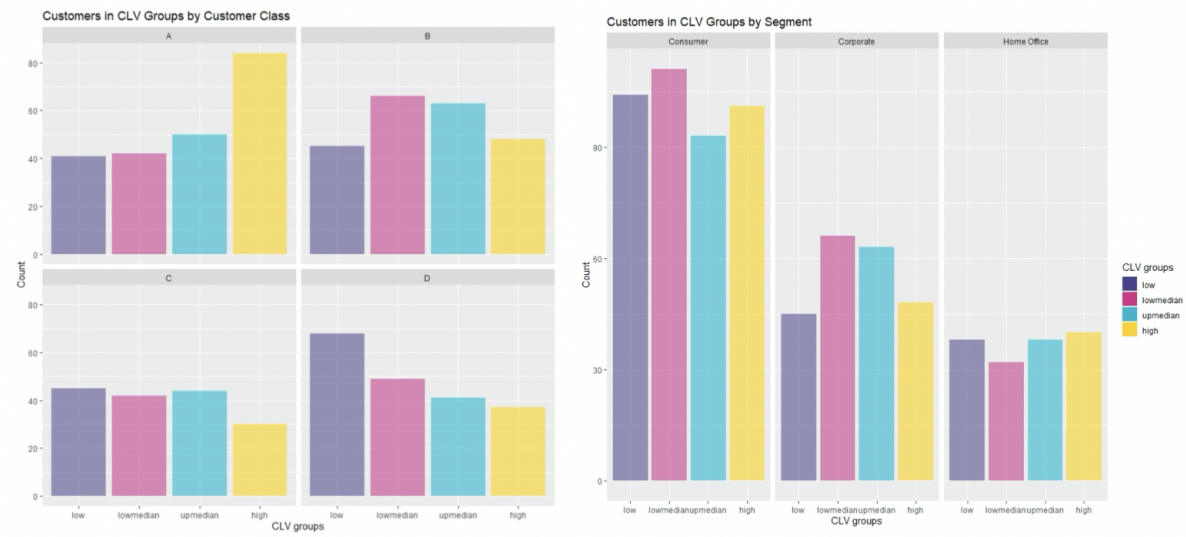
the bottom 75%. This indicates that the smaller number of customers, the company's most loyal and high value customers, contribute more than the entire bottom 75% of the company's customer base. With this, if the company invests more in the top 25%, the investment will have a greater proportionate impact on the company's expected return.

Additionally, with this calculation we wanted to find out which variables affected the customers' lifetime with the company. With the use of a scatter plot matrix we had the ability to spot any immediate correlations with lifetime. It was discovered that order frequency and average discount had the highest correlations to CLV. As seen in the figures below, we can see that average discount has a more significant correlation which we also found when running correlation and t-test. To our surprise, despite this strong correlation it is a negative correlation. This is surprising because as the average discount increases, customer lifetime value decreases. From this, we immediately knew that this was something that we had to take a deeper look into. Why are lifetime values decreasing when customers are being offered higher discounts?





After our findings with average discounts and customer lifetime value, we wanted to take the next step by gaining a better understanding of how CLV was distributed among the customer classes that we defined and discussed earlier in this report along with the predefined segments from the dataset. Our calculated CLV allowed for the customers to be grouped into 4 distinct categories based on their lifetime value with the company; low, lowmedian, upmedian, and high customer lifetime value groups. In the figure below, we can see how both groups differ in terms of lifetime with the company.



Seen on the left, it is clear that Class A customers have the most customers with a high lifetime value, while Class D has the highest number of customers in the low CLV category. Additionally, the graphic on the right compares CLV to the customer segments currently defined by the company. We see that not only are Consumers the largest segment, but we found that they actually have the lowest lifetime value average. This finding emphasizes that the number of customers within a specific customer segment does not mean that an investment in the larger customer segment will have a proportionate effect on the value to the company. Furthermore, if the current customer segmentation strategy is used we can see that the CLV has no differentiation. However, in our newly defined customer classes it is clearly seen which class holds the most high value customers. Hence, we suggest that our new segmentation method be adopted to run more accurate analyses in the future.

Now to go back to our finding on Discount Amount and Customer Lifetime Value, we wanted to determine which class was receiving the highest discounts. In the figure above, also shown earlier in this report, we can see that our least valuable customers in Class D are receiving the highest average discount amount. This was immediately seen as a potential issue, and explains why Class A customers are expected to be the quickest to leave the company. If the top 25% of customers are bringing in the greatest return, why are they not being rewarded? With this finding, we believe that there is a huge opportunity for the company to drive growth by increasing the value of existing and loyal customers rather than investing in gaining new customers. If the company adjusts their discount strategy and offers more discounts to those customers in the top 25%, the company can expect to have a greater return on their investment. Overall through our CLV analysis, we found that the most effective changes that should be implemented immediately include adopting our new customer segmentation method and shifting the discount strategy to treat the top 25% and more loyal customers better.

Logistic Regression

The next analysis that we worked on was logistic regression to create a model that would help us gain more insights from the data and so it would help the company understand which of our customers to focus their efforts on to keep them profitable.

What we want to understand is how the current discounting methodology impacts our customers. For the model we had to decide on a certain hypothesis and the one we chose was which of our top and bottom 25% customers would likely remain profitable for us and which ones would keep giving us losses based on discounts.

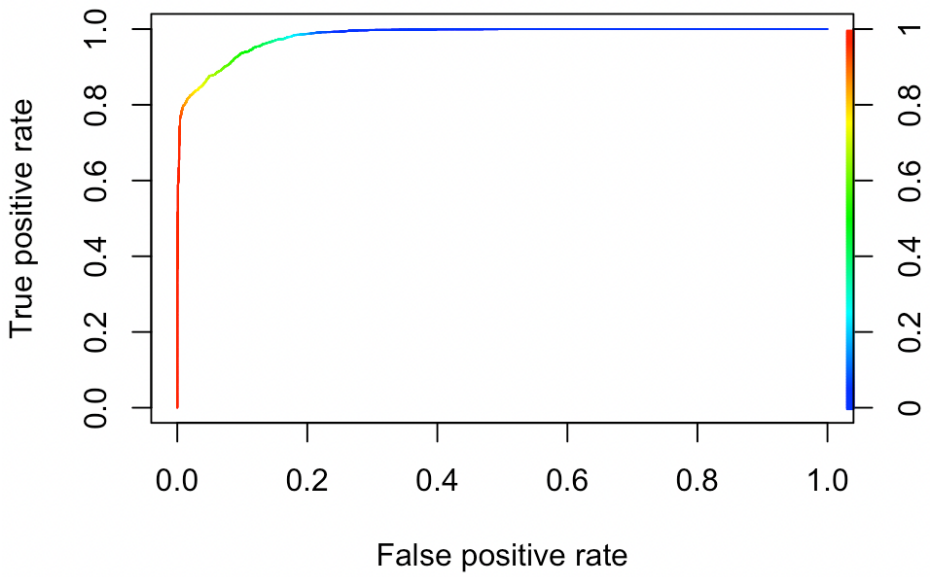
After conducting correlation and t tests, we found that three variables were significantly impacting the profitability of the customers individually as shown below. We divided our model data into 70% for training and 30% for validation. Our model has a confidence interval of 93% which shows the model has a good ability to predict the results.

```
glm(
  prob ~ Discount + Quantity + Sales
  data = train,
  family = binomial
)
```

As a result of the logistic regression we have the confusion matrix that would help us explain whether the model we have prepared will provide us with accurate results. For example, we don't want to be providing discounts to customers who are losing us money or have stopped purchasing our products. The Confusion Matrix shows that 95.2% of our Profitable observations and 87.4% of our non-profitable observations were predicted correctly. There is a small percentage of observations that were falsely predicted.

	FALSE	TRUE
0	0.874	0.048
1	0.126	0.952

The curve below gives us a good True Positive to True Negative ratio which indicates the model's ability to identify the right customers that we want to target with our Sales and Marketing team.



The results of the model here reiterate what we found from our previous analysis. Even though Discount is a significant variable for the model, there seems to be no impact on the profitability of our customers. It even reduces the quality of the model as a whole. This tells us that we should be looking further into which customers are being given discounts and what criteria are we using.

	Df	Deviance	AIC
<none>		6175.8	6207.8
- Quantity	1	6252.8	6282.8
- Sales	1	6343.8	6373.8
- Region	12	6521.4	6529.4
- Discount	1	20393.8	20423.8

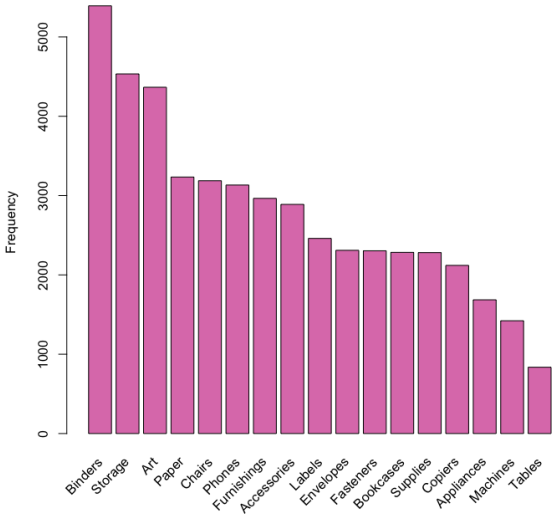
The second thing to note is that our model tells us something that we expected, increasing the quantity in the order will increase our profits. It tells us that the odds of a customer being in the top 25% will increase by 14.1% if we increase the quantity of the order they buy. We can use this information along with a basket analysis as to how we can target these customers.

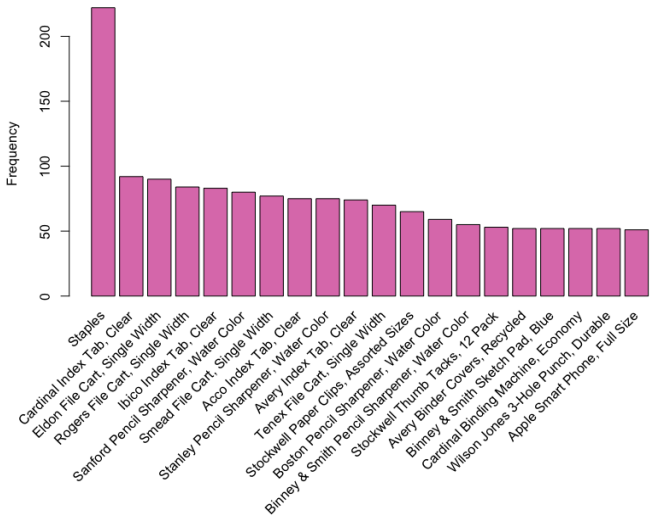
Discount	Quantity	Sales
0.000	1.141	1.001

Our model could be improved if we have more data regarding the customers to help us understand better how to retain them for longer and also keep them profitable.

Basket Analysis

After analyzing the general behavior of our customers, we wanted to look more deeply into their purchasing behavior when it came to products. Specifically, we wanted to see if the collected transactional data had predictive power. Therefore, we performed a basket analysis. Market Basket Analysis is used to uncover associations between items, or in other words, it allows us to identify relationships between the items that people buy. We decided to perform the analysis on product sub-category and product name. The graphs below display all seventeen sub-categories organized by purchase frequency, along with the same graph product name but only showing the top twenty. Binders is the most popular purchase sub-category, while staples is the most for an individual product.

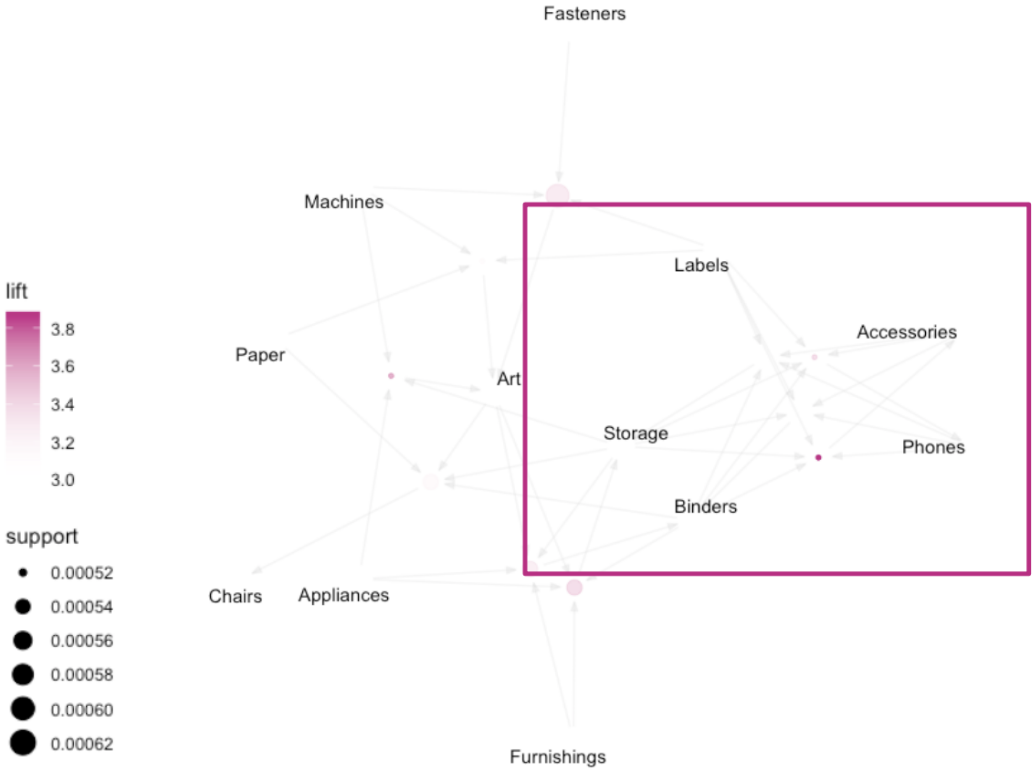




The output from a basket analysis is a series of rules, which are essentially If-Then scenarios or statements. Starting with our first analysis, below are the top 10 rules for sub-categories.

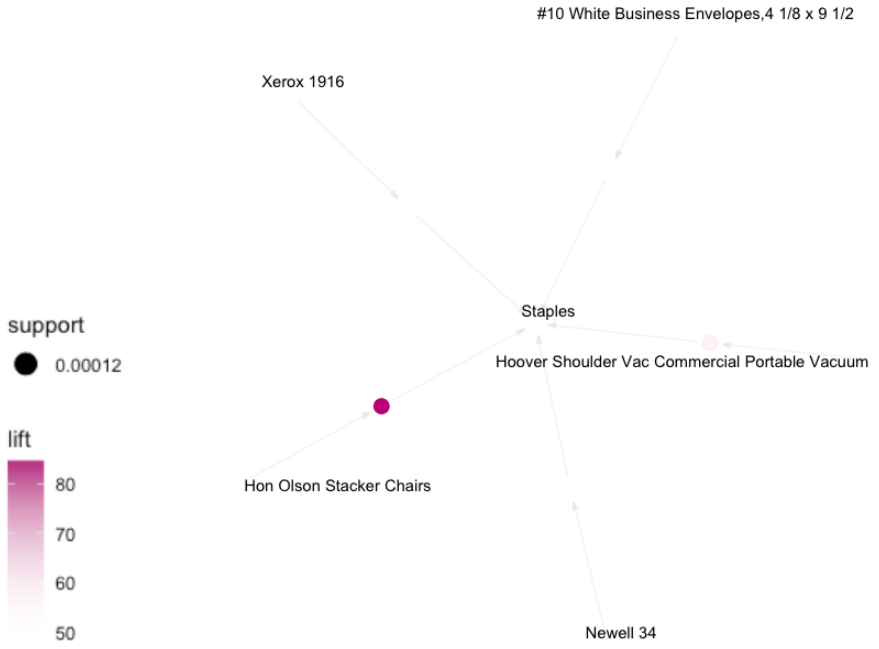
	lhs		rhs	support	confidence	coverage	lift	count
[1]	{Binders, Labels, Phones, Storage}	=>	{Accessories}	0.00052	0.45	0.00116	3.9	13
[2]	{Appliances, Machines, Storage}	=>	{Art}	0.00052	0.62	0.00084	3.5	13
[3]	{Appliances, Art, Binders, Furnishings}	=>	{Storage}	0.00056	0.61	0.00092	3.4	14
[4]	{Accessories, Binders, Labels, Storage}	=>	{Phones}	0.00052	0.42	0.00124	3.4	13
[5]	{Fasteners, Labels, Machines}	=>	{Art}	0.00064	0.57	0.00112	3.3	16
[6]	{Appliances, Art, Furnishings, Storage}	=>	{Binders}	0.00056	0.70	0.00080	3.3	14
[7]	{Art, Binders, Paper, Storage}	=>	{Chairs}	0.00056	0.40	0.00140	3.1	14
[8]	{Labels, Machines, Paper}	=>	{Art}	0.00052	0.54	0.00096	3.1	13
[9]	{Accessories, Labels, Phones, Storage}	=>	{Binders}	0.00052	0.65	0.00080	3.0	13
[10]	{Accessories, Binders, Labels, Phones}	=>	{Storage}	0.00052	0.54	0.00096	3.0	13

To show the implications of this analysis the first rule can be read as follows. If a customer has already purchased from the binders, labels, phones, and storage sub-categories 45% of the time they will also purchase from the Accessories sub-category. These rules can be plotted as shown below, where they form clusters.



We ran the same analysis for the product name and found five rules that all included the item staples. It appears staples has a relationship with many products including stacker chairs, portable shoulder vacuums, a Xerox product, white envelopes, and a Newell product.

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{Hon Olson Stacker Chairs}	=>	{Staples}	0.00012	0.75	0.00016	85	3
[2]	{Hoover Shoulder Vac Commercial Portable Vacuum}	=>	{Staples}	0.00012	0.50	0.00024	56	3
[3]	{Xerox 1916}	=>	{Staples}	0.00012	0.43	0.00028	48	3
[4]	{#10 White Business Envelopes,4 1/8 x 9 1/2}	=>	{Staples}	0.00012	0.43	0.00028	48	3
[5]	{Newell 34}	=>	{Staples}	0.00012	0.43	0.00028	48	3



Part 3 Managerial Implications (3 pages)

Based on findings

Throughout all the analyses, we discovered two main business problems. First, the more valuable customers seem to have an overall lower survival rate than our less valuable customers. Second, product sales seem to be somewhat clueless. In other words, managers are not focusing on the most profitable products and sub-categories. The marketing system is less than effective in its current state.

To address these problems, our recommendations consist of a two-fold plan. First, there is a need to design a more strategic discount model. Currently, there is a disproportional discount offered to customers, as they are identical to all of our customers regardless of their status. The managerial focus should be on relationship

development with high-value customers instead of incentivizing lower-value customers to spend more. Customer Class A, the top twenty-five percent of customers in order frequency, have a much higher customer lifetime value. There is an opportunity to increase revenue through investing and marketing towards Customer Class A, by replacing the current model to discount customers universally. Additionally, the company can run more thorough analyses in the future if our new customer segmentation method is adopted. With this new method the company will have the ability to see if a new discount strategy in fact affects the Customer Lifetime Value within various customer classes.

Secondly, Managers can use the insights gained from market basket analysis in a couple of ways. First, they can develop a product recommendation feature for the website. These are common on major sites like Amazon, where near checkout it says something like “customers who purchased this product also viewed this product.” By recommending associated products frequently bought together, managers can organically increase the number of items purchased. In addition, managers can develop target marketing campaigns and promotions to customers and entice them to purchase related products for items they recently purchased. It was also revealed during our basket analysis that binders and staples have a high purchase frequency compared to other sub-categories and items. This warrants special attention from managers. Managers should look into their profit margins and put high importance on the stock of these items, as they have a substantial impact on the bottom line.

Another interesting trend that we saw was that almost 25% of our orders were giving us losses which gave us a 3:1 Profit/Loss ratio of our orders. This is quite alarming and the company despite a positive trend in profits is losing a lot of money. There is a decision to be made on whether we should cut these customers or not.