Democratization of Data

Team HYJECS

Eleanor Jiang, Jingyi Lu, Chelsea Miao, Yue Wu, Yiren Xu, Hanyue Zhang

STATS 141SL
University of California, Los Angeles
March 2021

Contents

1 Abstract

As it becomes increasingly apparent that analysis of business data can generate great profit, we set out to investigate these benefits for small businesses in particular. Our client, Quantum Analytica, is a company with the goal of democratizing data for small businesses [1]. By examining the order, product, and customer data of Austin Custom Brass, a brass instrument store, we extracted useful information and recommended business strategies for increasing sales, curating product, and promoting customer engagement. With orders, we analyzed transaction date, shipment methods, coupons, etc., and found that higher order sales are associated with certain months of the year, free shipping, and not using coupons. Information from product data and customers' online activities implied that customers have loyalty to brands but not to the website. With the available customer information, we found that advertising through email is a relatively effective method, and during holiday seasons there tends to be a bigger wave of new customers. Also, we found that adding images to the website and making replenishment in time were important for total sales.

2 Intro

In modern-day society, companies have increasingly recognized the importance and usefulness of utilizing data. Large companies such as Facebook, Amazon, and Google have invested great amounts of human resources and serviceable tools to aggregate data and obtain valuable insights to promote development prospectively. However, the utilization of data has not been made effortlessly accessible to small businesses.

To demonstrate that data makes an enormous difference for small businesses, we worked with Quantum Analytica, a company devoted to the democratization of data for small businesses. In particular, we analyzed data provided by Austin Custom Brass, a music company that sold brass instruments and accessories, to help the company establish business strategies for sales promotion, product selection, and long-term engagement with customers.

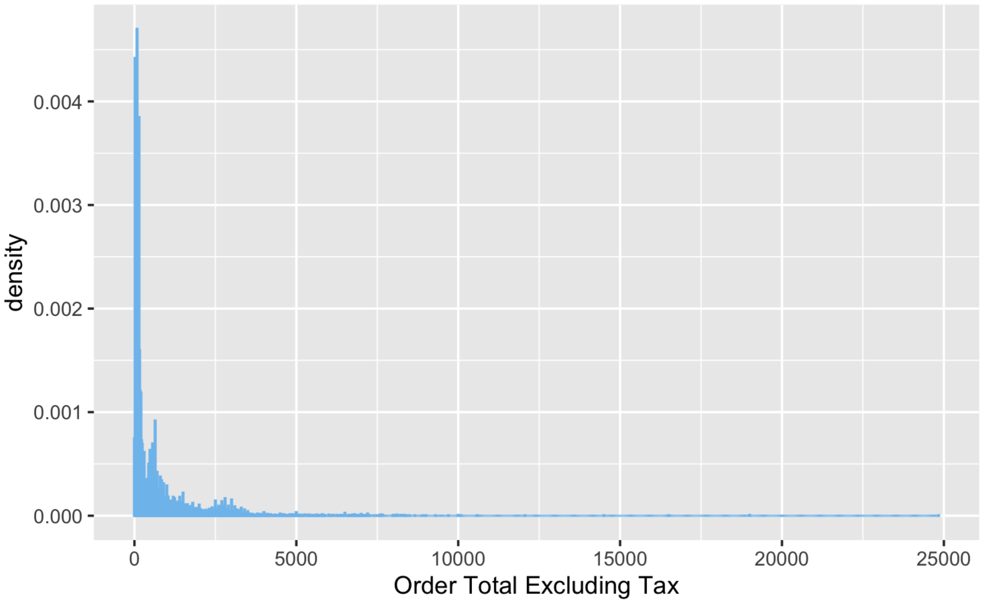
3 Data Discussion

3.1 Orders

With a dataset of each order placed at ACB (named orders.csv), we hoped to find out what promoted sales, so we examined only completed/shipped orders and deleted orders with 0 in the transaction total.

We re-grouped categorical variables (e.g. shipping and payment methods) for interpret-ability. We regrouped orders that were not pick-up with 0 shipping cost to free shipping. We also regrouped those that did not belong to any of four shipping methods as ”other methods”. Due to the original data having overlapping categories for payment methods and it being natural to pay in checks or card for large transactions, payment method could tell us limited information about what promoted sales. Therefore, we removed it from the predictor variables.

One possible problem in the data was that the distribution of orders total was right-skewed with a wide range from almost 0 to 24849 dollars per order. It could be due to that the differences among the different categories of products sold were too large (e.g., accessories like instruments stands cost less than a hundred dollars while the instruments themselves often cost thousands of dollars on average). We decided to log transform the orders total column.



Plot 3.1.1: Distribution of Order Total Excluding Tax

From Product Details, we also extracted quantities of products purchased to be later used in our Products model. From Coupon Details, we extracted the coupon codes used, also generating a binary variable of whether a coupon was used. Finally, the order date was changed into month to understand the seasonality of sales.

3.2 Products

The Products dataset contained vital information regarding individual products, including product ID, names, brands, descriptions, conditions, cost and sell prices, shipping fees, etc.

First, we removed columns that would not help much in our analysis, such as images, URLs, and warranty. Then, we used N/A to represent the missing values in product brands, and missing values in cost prices. Next, we converted all categorical columns into factors and converted dates into commonly used date expressions. After that, we extracted from the Category descriptions and retrieved the main categories to which the product belongs. Lastly, we removed the product with ID 3410, since this product is determined as invalid based on the product description.

3.3 Customers

The customer data we were given contained two files in total:

1. Customer_anonymous provided us information about the customers' company, notes, store credit, customer group, date joined, address, received review/abandoned cart emails, and tax exempt category.

2. Customer-2020-12-08-18-07-50 provided us information about the customers' names, company, email, phone, notes, store credit, customer group, date joined, address, received review/abandoned cart emails, and tax exempt category.

After reviewing the two customer files and observing completion and relevancy, we found that there are some potentially important analysis could be conducted toward the customers' geographical status, purchasing pattern and effective ways of advertising. There are abundant of data for street addresses, email advertisement responses and other potentially useful columns. We would pick out those columns and do further analysis. We also noticed that since customer information is hand-filled, there tends to be more missing or error information in the customer data. This bias will need to be considered in future processes.

3.4 Session JSON Data

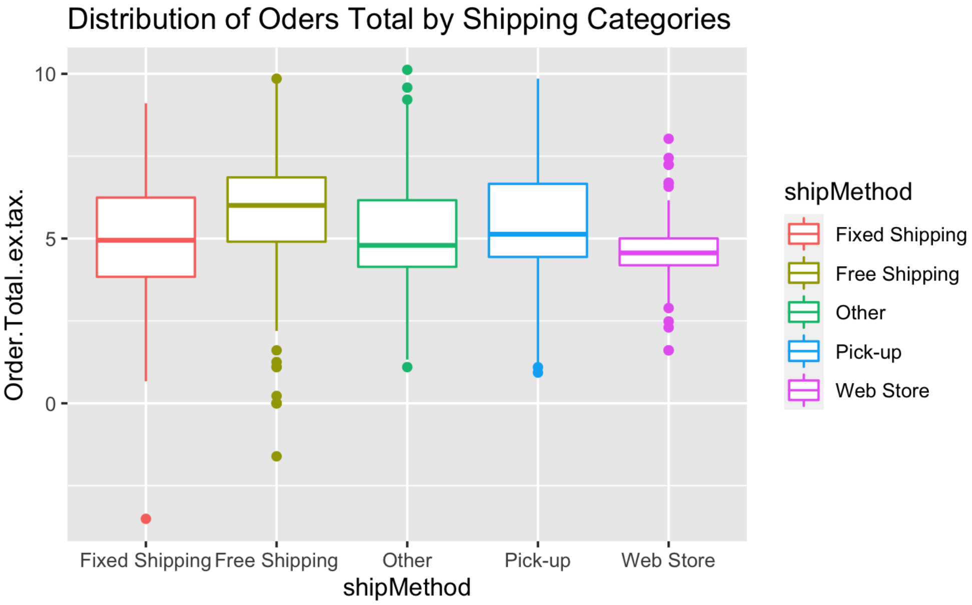
The Session JSON dataset provides customers' online activities, including whether they click, open, drop, and unsubscribe emails, what products they have viewed, and primary customers' demographics.

First, we summarize the total number of clicked, opened, dropped, and unsubscribed emails. Then, we transform the quintessential customer and product information from the Session dataset to become the supplementary information for the Product and Customer dataset. For the Product related data, we collect the number of times views by product. Also, for the Customer related data, we extract the various products viewed by each customer. We combine the products that has been viewed with the product that has been purchased by each customer, and we compute the rate of purchased in viewed.

4 Important Preliminary Findings

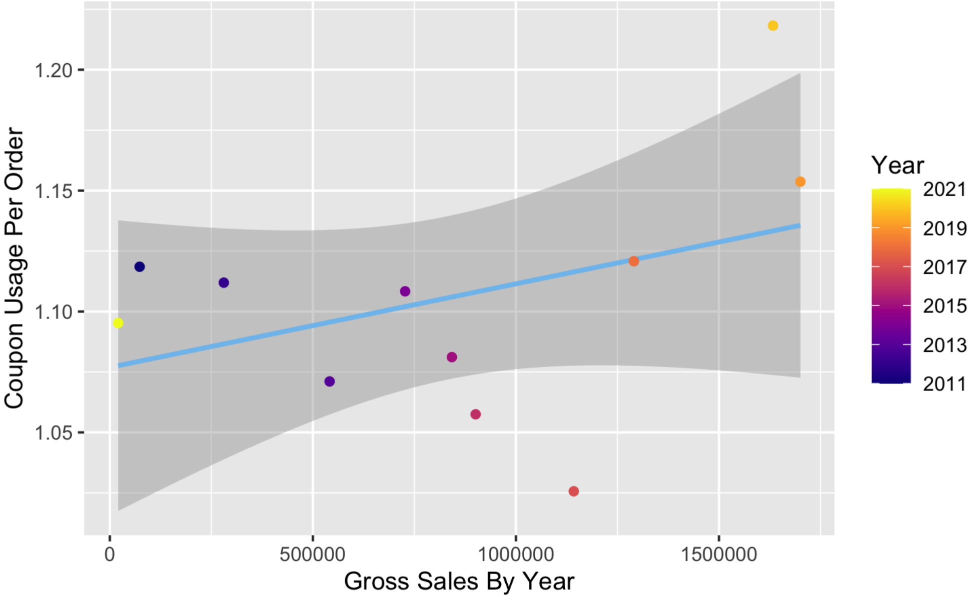
4.1 Orders

We plotted boxplots of gross sales by shipping categories. We could clearly observe the difference between shipping methods, with free-shipping achieving the highest mean gross sales.



Plot 4.1.1: Distribution of Orders Total By Shipping Categories

In other retail businesses, it is common to use coupon codes to promote sales. We hoped to investigate whether this pattern still applied to the brass store. Therefore, we plotted the gross sales against the coupon usage per order, grouping by year. We could observe a positive trend between the two variables.



Plot 4.1.2: Coupon Usage Per Order Grouped By Sales

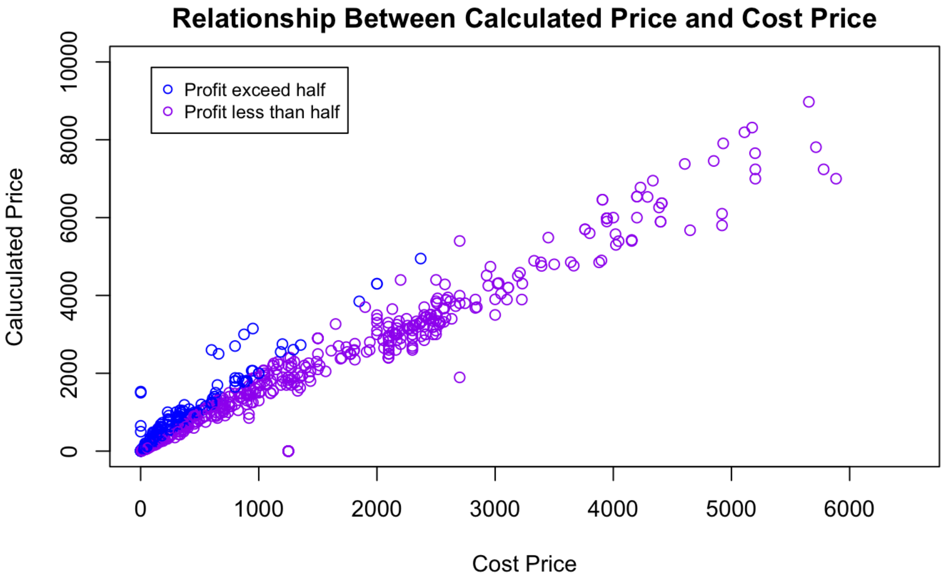
With a summary table of coupon usage, Table 4.1.3, we found that the most used coupons were FALLBACK10, followed by GOBBLE, FREESHIP, etc.

Coupon Code	Count
FALLBACK10	103
GOBBLE	59
FREESHIP	52
20ACC	38
JM15	33

Table 4.1.3: Coupon Usage

4.2 Product

We aimed to get insights about the demand in the specific time period of certain products, the relative profit of the product concerning all kinds of costs (cost of product, cost of stocking, shipping fee), customers' preferences of purchasing(e.g. whether sets with certain related products are more popular), in order to adjust the investment on different products.

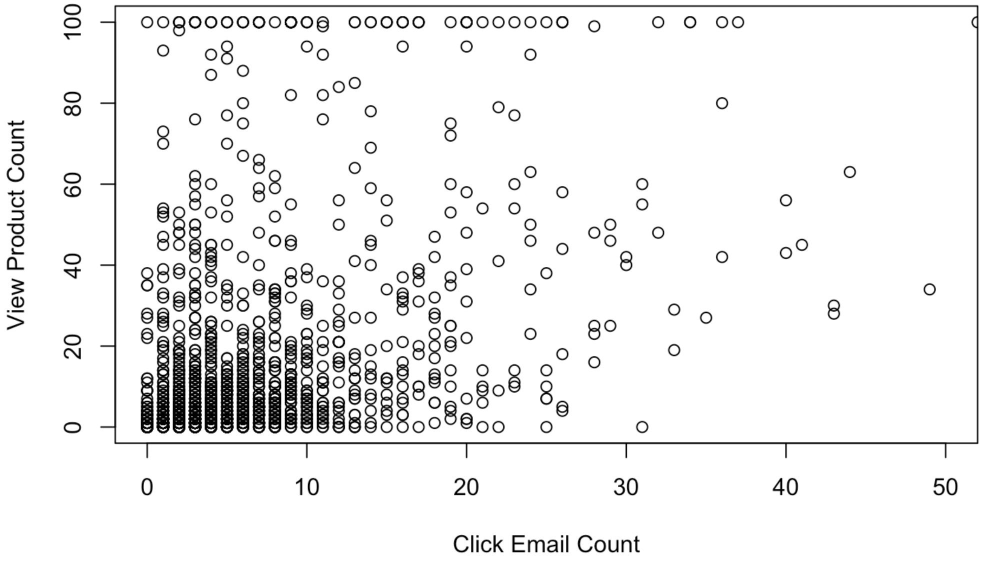


Plot 4.2.1: Relationship Between Calculated Price and Cost Price

Based on Plot 4.2.1, we observed that there existed a positive linear relationship between the cost price and calculated price, which indicated that the calculated price was set according to the cost price. Moreover, products which achieved more than half of the profit, had relatively low calculated prices and cost prices.

4.3 Customer Online Activities

We ignored the activities of dropping and opening email because of their limited occurrences and strong correlation with clicking emails.



Plot 4.3.1: Customer Online Activities Plot

We first checked the relationship between Click Email counts and View Product counts. According to Plot 4.3.1, we could see that, for customers who preferred to click emails (with clicking counts larger than 12), there was a positive relationship between email clicking and product viewing counts. However, for customers who did not click emails very often, most customers viewed products below 20 times and we could not see a clear relationship between clicking emails and viewing products. Nevertheless, taking customers' un-subscription rate 0.064 and customers' direct-product-view rate (no email activities) 0.0084 into consideration, we would still recommend sending emails about products more often to all customers, because it costs little but brings more potential sales.

4.4 Customer

We have gained valuable insights from preliminary observations and tests for customers. They reflected the company's potential sales force and future growth methods.We have conducted direct analysis for the targeted audience that the company has attracted or will attract in the future. Customer data is valuable and can help us further examine strategies to help the growth of the company.

For the customer data we are given, we gained information about the customers' names, company, email, phone, notes, store credit, customer group, date joined, address, received review/abandoned cart emails, and tax exempt category. We further evaluated potential combinations of tests.

Data Cleaning and findings: We first filled out blank spaces or unfilled parts of the data with N/A, then focused on the state allocation and received review/abandoned cart emails. we created a new csv with all the state names in it and used algorithms to extract them from the customer files for further analysis. For the received review/abandoned cart emails, we categorized them into "N" or "Y", and formulated a summary plot.

We discovered that with those two data sets that heavily relied on customers' self-filling system, there tends to be higher error rate with correct address information, telephone number etc. However, this can also be a good notice to be taken into consideration for future marketing strategies and customer organizations (for example, could have digital auto-fill systems from other accounts) we encountered some problems such as there are still many NAs in the state part, with either missing information or incorrect format. We also cleaned data based on Customer joined date and month, to check if there is a pattern that could lead us to further findings. The initial finding is that there tends to be more new customers toward the ending of the year, and there appears to be a stable growth each year. And the respond rate to email advertisement seem to be good.

5 Important Findings and Methodology

5.1 Orders

To understand the association of sales for each order as well as predictor variables including Order Date, Shipping Method, and Coupon, we first performed a one-way ANOVA F-test as an omnibus test.

We first transformed Order Totals with a log transform to satisfy the assumption for ANOVA tests.

Response: Order.Total.ex..tax.					
	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Order.Date	11	58	5.25	2.565	**0.00304**
shipMethod	4	1184	295.90	144.504	**<2e-16**
Coupon.YN	1	34	34.07	16.637	**4.57e-05**

Table 5.1.1: Orders Model Analysis of Variance Table

As shown in Table 5.1.1, p-values were significant for each variable, so we rejected the null hypothesis that the means of each category were the same.

Then, to further understand the difference between each group's mean sales, we performed a post-hoc Tukey's Honest Significant Differences test.

	diff
02-01	0.261735159
07-01	0.180889101
11-01	-0.083088896

Table 5.1.2a: Order.Date: Tukey's HSD Test

	diff
Free Shipping-Fixed Shipping	0.88315427
Web Store-Fixed Shipping	-0.44449415

Table 5.1.2b: shipMethod: Tukey's HSD Test

	diff
1-0	-0.2064949

Table 5.1.2c: Coupon.YN: Tukey's HSD Test

Shown in Table 5.1.2a, comparing mean sales for each month in variable Order Date, with January as the baseline, we found that February had the highest sales and November the lowest, on average. For shipping methods in Table 5.1.2b, with Fixed Shipping as the baseline, we found that Free Shipping had, on average, highest sales and Web Store the lowest. As for coupon in Table 5.1.2c, with using a coupon as the baseline, we found that orders not using coupons, on average, had higher sales.

5.2 Products

To comprehend the most popular products, we delved into the features of products which contributed the most towards the product's sales and popularity. The sold quantity of each product, which is extracted from the Order dataset, was determined as the standard of whether the product was popular. The features of products including brands, conditions, fixed shipping prices, cost prices, retail prices, times of viewing, whether there was free shipping, allowing purchases, visible, and inventoried, were chosen from the Product dataset and transformed from the Session JSON data.

In order to investigate the useful features and avoid multicollinearity and overfitting, we performed feature selection. Since the variables with lower variance could not improve the performance of the statistical model, we remove the Product.Condition and Allow.Purchases variables that had near zero variances. Additionally, to eliminate multicollinearity, we removed the Cost.Price variable, which was highly correlated to the Calculated.Price variable.

Since we wanted to explain variation in the sold quantity of each product, which can be ascribed to variation in the product variables, we applied the multi-linear regression method to quantify the relationship between sold quantity and various product features. We applied the log transformation on the Sold.Quantity and viewed variable since the range of the these two variables were more than one order of magnitude.

Response: log(Sold.Quantity)					
	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Brand	91	829.17	9.112	7.5293	**2.2e-16**
Calculated.Price	1	23.84	23.838	19.6976	**1.063e-05**
log(viewed)	1	131.41	131.411	108.5885	**<2.2e-16**
Product.Visible	1	77.24	77.238	63.8249	**6.045e-15**
Product.Inventoried	1	51.94	51.936	42.9161	**1.148e-10**

Table 5.2.1: Analysis of Variance Table

The Table 5.2.1 displays the performance of features. The p-value, which is a commonly used measure of statistical significance, indicated that the probability of obtaining a more extreme value compared to the observed value assuming the estimated coefficient of each specific feature was zero. Based on the last column in Table 5.2.1, the p-value of all these five features were small enough to be considered statistically significant. The Brand was a categorical variable that has 92 different brands and a small p-value, which indicated that at least one brand was significant. Since the estimated coefficient of Calculated.Price was close to zero, it did not have much explanatory power. Therefore, we performed a further investigation on Brand, Viewed, Product.Visible, and Product.Inventory.

5.2.1 Brand

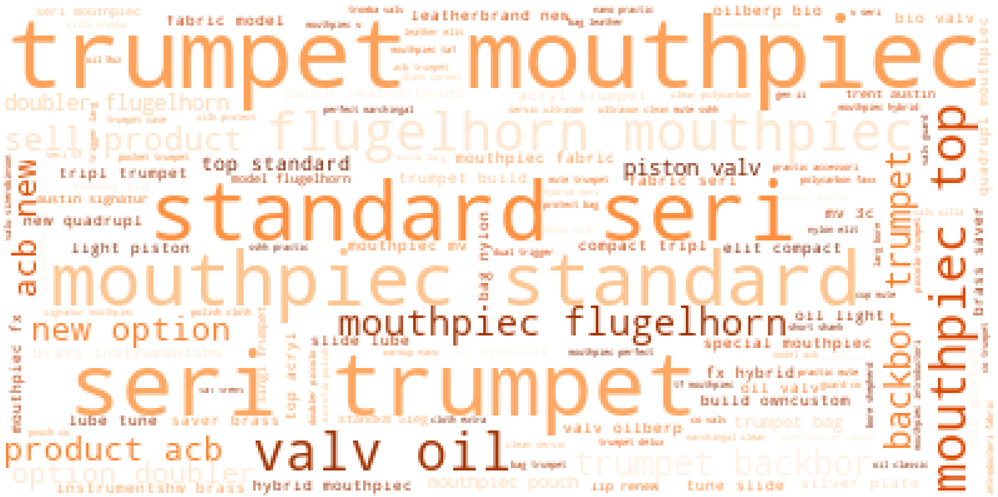
Since the Brand of the products had great influence on the product sold quantity, we extracted the sold quantity based on brand and further calculated the profit. The Table 5.2.3 shows the sold quantity and profit of the brands with top ten sold quantity.

Brand	Sold Quantity	Profit
Austin Custom Brass	8674	1177341.88
(blank)	1084	44193.14
Hetman	815	1555.18
Adams	768	225330.03
Gard	723	55500.55
Ultra-Pure	559	1319.35
Warburton	396	8951.20
Schagerl	257	139765.86
berp	254	1093.97
Leather Specialities	203	3496.97

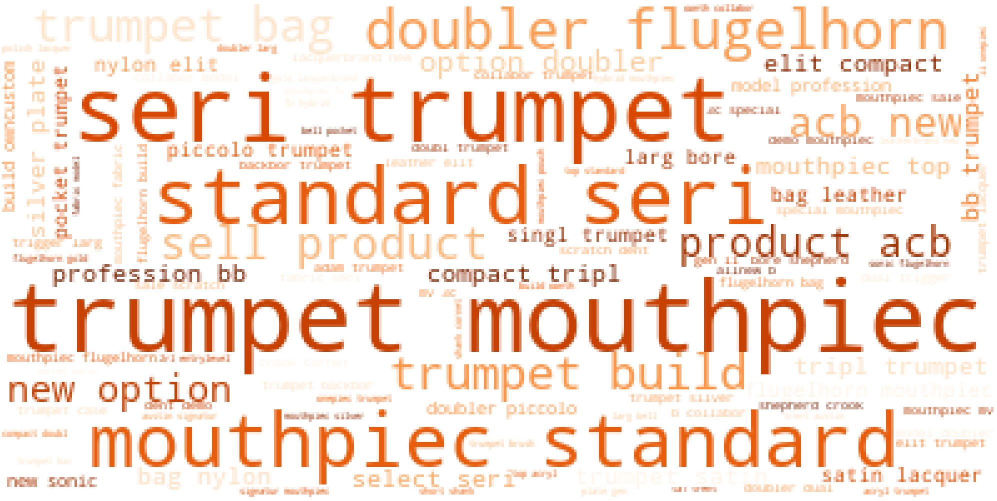
Table 5.2.2: Sales by Brand

According to Table 5.2.2, Austin Custom Brass was the most favorite brand that has the highest sales among all products. The Austin Custom Brass brand kept customers coming back and established a strong customer base. Also, the customers were willing to repeatably purchases items. In addition, although the Hetman ranked as the third highest sold quantity brand, its profit was still less than the Schageri due to the reason that the profit of individual Hetman's products were less than the profit of individual Schageri's products. Brand loyalty is a remarkable advantage that can increase user engagement. When promoting to registered customers, we could take advantage of the brand loyalty by recommending commodities with the same brand of their historical orders, because loyal customers always purchase items from the same brand repeatedly. For extending to new users or promoting to those customers who were not committed to certain brands, suggesting the brands with top sales might retain the customers and further bring more profit.

5.2.2 Viewed



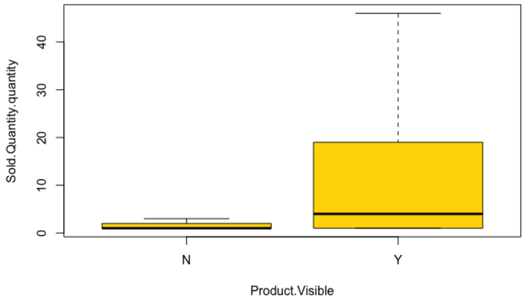
Plot 5.2.3.a: Most Popular Sold Product Type



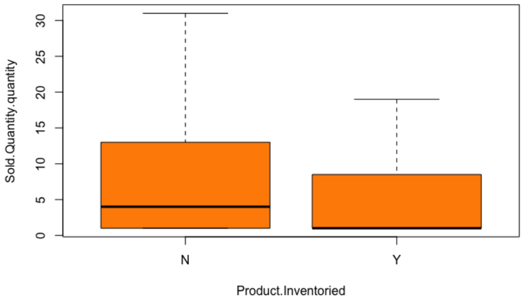
Plot 5.2.3.b: Most Popular Viewed Product Type

After gathering the products viewed and bought by customers, we ordered the products by their sales and views. Then we extracted keywords from product names and created a word cloud for each. We could easily see the most popular product types from the word clouds, which were mostly mouthpieces for different instruments such as trumpet and flugelhorn. This may help the store make decisions on product selection. In addition, from Plot 5.2.3.a and Plot 5.2.3.b, we could observe that the two word clouds were similar, which implied that the customers had a clear idea of what they will buy. These customers usually made their purchases very efficiently through direct searching. They might just check whether the products they want were sold and whether the price was expected. These two features implied that they have no loyalty to any website or store but they may have loyalty to brands. One suggestion we provide is that the website should optimize their search engines and make it easy and fast for customers to directly find the products they want. Also, since they maybe sensitive to the price, we suggest that the website (store) set prices reasonably.

5.2.3 Product Visible and Inventoried



Plot 5.2.4: Boxplot of Sold Quantity vs. Product Visible



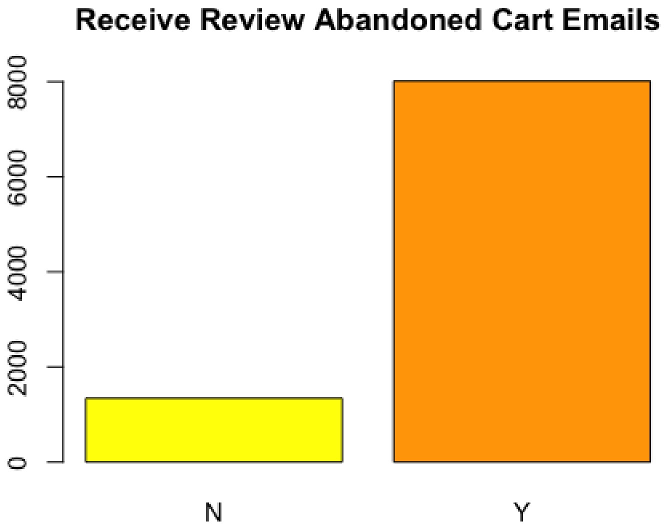
Plot 5.2.5: Boxplot of Sold Quantity vs. Product Inventoried

Besides the box-plots in Plot 5.2.4 and Plot 5.2.5, we performed t-tests on Product.Visible and Product.Inventoried, in order to statistically show the influence of visibility and inventory on total sell quantity of a certain product. The p-value for Product.Visible was 0.00072, which was less than 0.05, indicating that the mean difference between the paired observations (Products that are visible, and Products that are invisible) was significantly different from 0. The p-value for Product.Inventoried was 0.003, which was less than 0.05, indicating that the mean difference between the paired observations(Products that are inventoried, and Products that are not inventoried) was significantly different from 0. So we suggest the website to show at least one picture for each product, and replenish products in time. We also recommend sending reminder emails to workers in charge of replenishment if the inventory of certain products is under a certain amount.

5.3 Customers

After reviewing the two customer files and observing completion and relevancy (some useful categories have 0 completion, hence cannot be used) we concluded that the most useful information to extract are whether they receive review/ abandoned cart email,state allocations, and dates joined. Whether the customer receives or review/ abandoned cart email is important because we can determine if this way is an effective way of advertising, and hence decide future strategies for promoting or alter ways of doing so. State allocations can also showcase geographic information that indirectly show us many information such as whether we should emphasize on one state for more future advertisement, and patterns of customers' area. Dates joined can tie back to advertisement and suggest which period can potentially attract more customers.

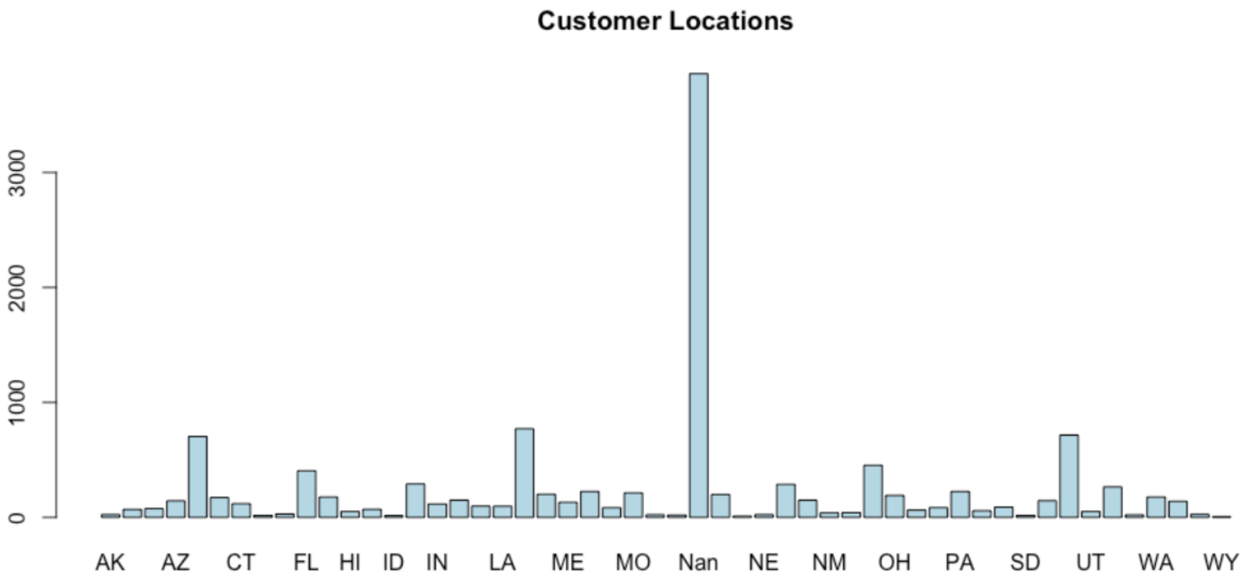
Received review/abandoned cart emails:



Plot 5.3.1: Received review/abandoned cart emails

From this graph, we can see that using cart emails to advertise for the company's brand, new products, and promotions is a relatively effective method. 8000 out of 9700 customers received or reviewed their email from the company. There is a 82 percent chance that the customer will review the information provided. At the same time, it is also a low cost way to advertise, so we suggest the company to keep this advertisement method and continue providing better content.

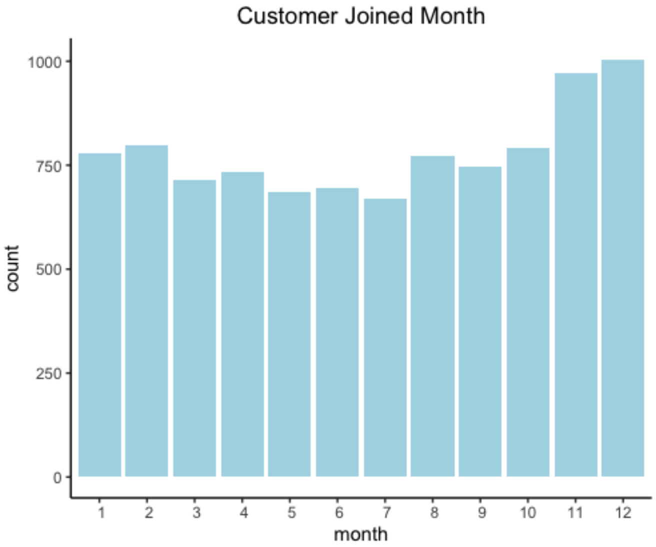
By State:



Plot 5.3.2: Customer State Allocations Graphs

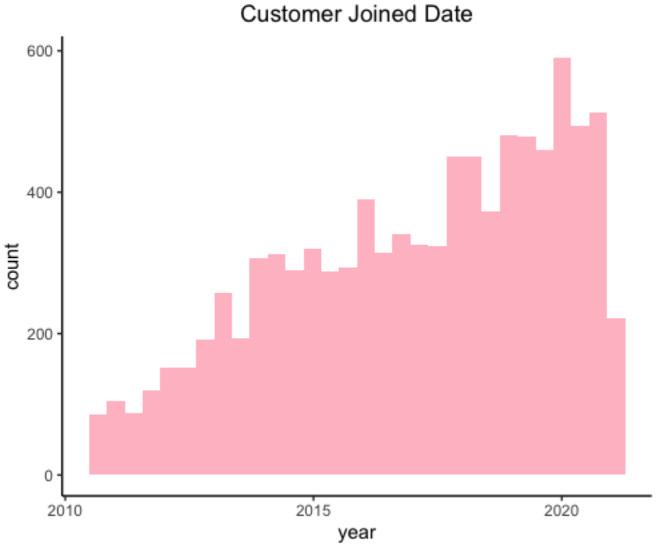
From the graph, we can see that CA, MA, and UT have the most customers. We can conclude that there might be a pattern with coastal regions having more customers. However, we have to take into consideration that there are a huge amount of NAs in this data due to wrongly filled or missing information.

Customer joined date/ month:



Plot 5.3.3: Customer joined month

The pattern for customer joined month is relatively even, however, we can still notice a higher count of almost 1000 toward the end of the year in November and December. A possible explanation for that is during Christmas and holiday seasons, people have more time and leisure to purchase the products or use the services of our company.



Plot 5.3.4: Customer joined year

We can see that the customer joined year has a positive growth, with some higher peaks in between (2016 and 2019). Those could be potentially affected by the economic situation of that year, and other environmental and political events. However, we can still see the pattern such as end of the year and steady growth that encourages us to continue improving based on the business operations right now. In the future when we make business plans, we can target those specific time periods for more profits.

6 Conclusions and Advice

In terms of order sales, we found that February and July welcomed the highest sales while November's sales were generally lowest. And we observed that customers tended to make bigger purchases when granted free shipping. However, the adoption of coupons did not bring an increase in average sales. Combined with the understanding that ACB might have a more specific customer base compared with other retail businesses, we considered a strategy on how to design and use coupons.

The brands, calculated prices, number of views by the customer, product visibility, and product inventory are the features that have a majestic contribution towards the product sales.

We also found that the cart emails are a relatively efficient way to advertise the company brand and the product, so the company could keep developing this aspect and make the content better. There also tend to have a bigger customer growth toward the end of the year with holiday seasons, so the marketing team can potentially target this period for more advertisements.

Advice

- The findings of negative difference in sales with coupons and positive difference with free shipping suggest that we can redesign coupons for free shipping; for example, use a "first-time-customer" coupon that new registered customers could have their first order free shipped.
- We suggest taking advantage of high sales in February and July to send more email promotions, and in the slower season of November, we suggest planning store maintenance or other activities accordingly.
- We can recommend products from the same preferred brands for advertising to subscribed customers based on their historical orders.
- We suggest recommending more products from the most popular brands to new customers or customers who do not have loyalty with certain brands.
- We suggest, on the website, to adjust the page layout and optimize search engines for customers to easily and quickly find the products they want, and to set prices for products reasonably.
- We suggest designing the website to show at least one picture for each product and replenish products in time.
- We suggest to continue advertising through email and target the end-of-the-year holiday seasons for more advertisements.

7 Bibliography

References

[1] *Quantum Analytica*. https://quantumanalytica.io/.