Data Discovery

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

February 2021

1 Understanding of Data

In the modern world of society, more and more companies have realized the importance and usefulness of data. Big companies such as Facebook, Amazon, and Google have invested a lot of human resources and useful tools to aggregate data, and then get the insights behind to make wiser decisions and promote their development in the future. However, one problem is that the utilization of data is not accessible to small businesses. In order to show that data makes a huge difference for small businesses, we started analyzing the data provided by "Austin Custom Brass", a music company that sells trumpets in brass instruments, and then make inferences to help them make business strategies about advertising, long-term engagement with customers, and product selections.

1.1 Orders

At the first glance of the data, we found out the entries for categorical variables (e.g. ship and payment methods) were too specific to cause hundreds of unique levels. We extracted the most common levels and re-categorized the variables.
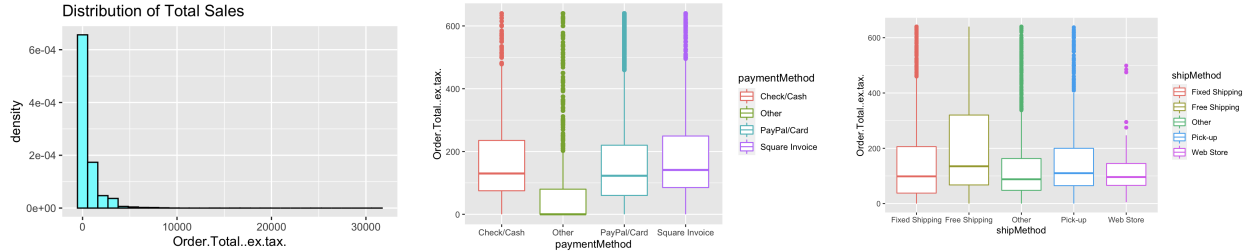
The distribution of the variable, orders total, is seriously right skewed. Its range is also wide from 0 to 31176 per order. We computed the ANOVA table between the categorical variables with the orders total and the resulted p-values were all significant. The boxplots we drew also conveyed the same idea. However, the huge amount of outliers of the orders total made us hard to make analysis. We come up with two possible solutions. Firstly, we can calculate the profit by combining with the cost price from the product dataset to see if the right skewed can be reduced. Secondly, due to the volatile categories from different price range that the store sold, we can further separate the orders total to different product categories.

Apart from calculating how many times an individual customer made purchases to investigate into the brand loyalty, we also made a time plot by summing up the daily sales. We found out that high gross sales took place at the start and end of a year.

With the rest of the variables, we noticed that "Product Details" and "Coupon Details" include useful information pertaining to the amount of each product purchased and the coupon events used. We then extracted coupon codes and products and quantities of each product of every order. This information is best combined with Products and Customers data.

Sales Per Day
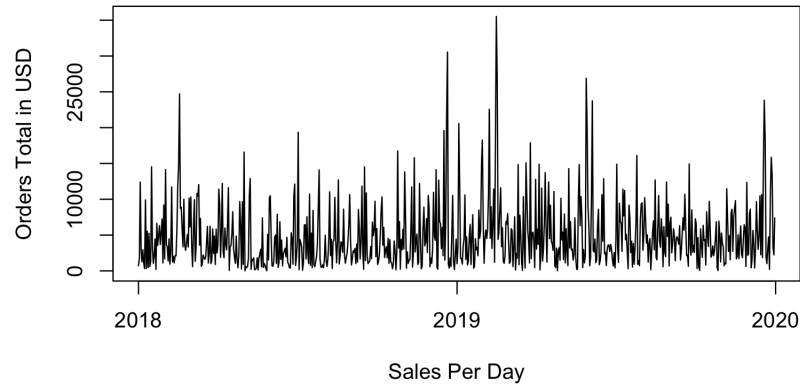
1.2 Products

The dataset of Products contains essential information about products themselves, including names, brands, descriptions, conditions, cost and sell prices, shipping fee, and images. It also includes information especially for the company to manage and track their selling, including ProductID, stock, Categories, measures of size, ect,. We would get insights about the demand in the specific time period of certain products, the relative profit of the product concerning the all kinds of costs(cost of product, cost of stocking, shipping fee), customers' preferences of purchasing(eg. sets with certain related products are more popular), in order to adjust the investments on different products.

Histogram of Date Added

Date Added
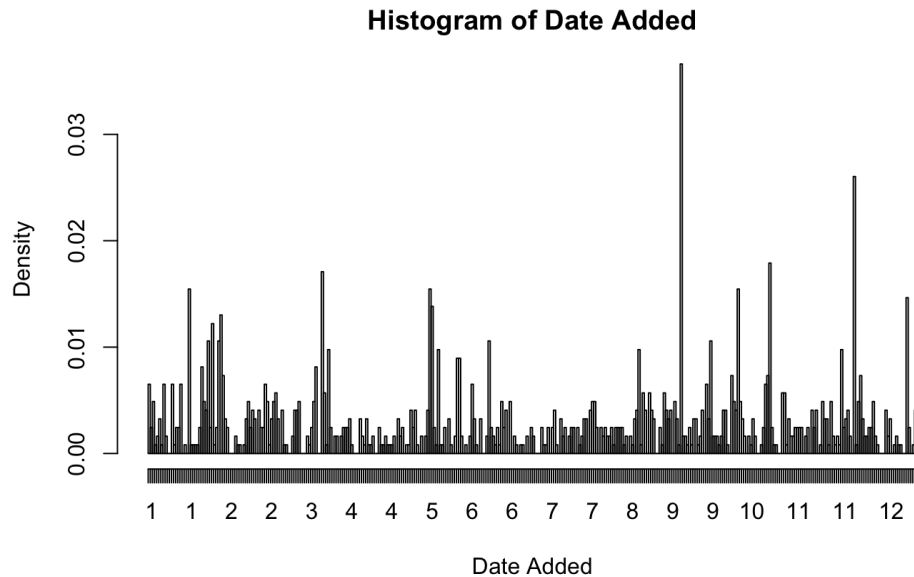
We got some insights regarding the demand of certain product over a specific time period. There's a pattern that nearly every two months, the company needs to replenish stocks. However, the demand seems particularly low in summer, but particularly high in September and November. We assume that this is because most students would like to purchase at the beginning of their school year, or at the discount season such as Black Friday or Christmas.

Relationship Between Calculated Price and Cost Price
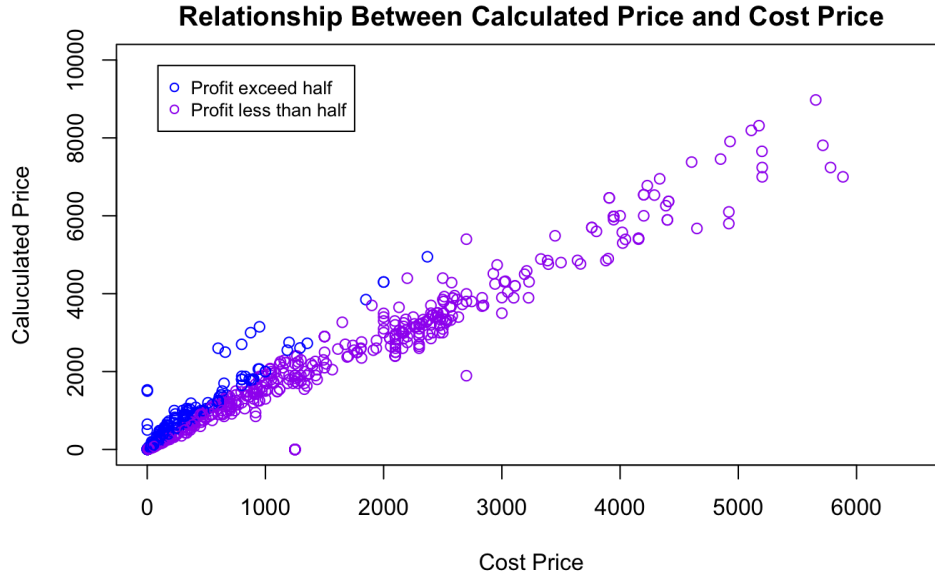
Additionally, we observed that there exists a positive linear relationship between the cost price and calculated price, which indicates that the calculated price is set according to the cost price. Moreover, the product, which achieved more than half of the profit, has a relatively low calculated price and cost price.

1.3   Customers

Customer information can provide the most direct analysis for the targeted audience that the company has attracted or will attract in the future. Accurate contact information also prepares for future businesses and potential re-use of data for other agencies. To conclude, 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, there are 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 (some useful categories have 0 completion, hence cannot be used) I concluded that the most useful information to extract from the two customer files are state allocations and whether they receive review/ abandoned cart email. State allocations can 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. Whether the customer receives a review/ abandoned cart email is also 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.

2   Ideas for Investigation

1. We can combine the orders and products data to investigate brand loyalty and calculate profits from the orders total and cost price.

2. We can use the categories of products in the inventory csv to analyze which category of products generate the highest profit. The total annual profit for the products can help the company make decisions on product selections.

3. Which coupon codes are used most frequently? Are sales improved by events/holidays? Can the shop benefit from promoting coupon codes on their website or through brand deals on social media?

4. Which category of products is most popular? Are there possible package deals to promote sales?