An exploration on E-Commerce dataset

Team5: Ke Li, Eddie Tseng, Shuyi Ma, Yang Chen

March , 2019

1 Introduction

1.1 Project Background

Altogether, we have 9 tables that are related to E-commerce to analyze. We find out that this company is from Brazil and sells products in many different categories accross different states in Brazil. We have a total of 112,650 observations after cleaning up. The time duration covers from October 2016 to September 2018. However, data from November and December in 2016 is missing.

1.2 Analysis Objective

For this project, our objective is to explore the given e commerce data set. To be more specific, we want to find out whether there is any influence on the customer buying from the company.

2 Data Set

2.1 Data Format

Each table corresponds to different categories of information. A brief overview is as described below.

Table 1: Tables and Corresponding Description

Table Name	**Variable Name**
Customer Reviews	review_id, order_id, survey_score, survey_review_title, survey_review_content, survey_send_date, survey_completion_date
Customers	customer_id, customer_unique_id, customer_zip_code_prefix, customer_city, customer_state
Geolocation	geo_zip_code_prefix, geo_lat, geo_lng, geo_city, geo_state
Order Items	order_id, order_item_id, product_id, seller_id, shipping_limit_date, price, freight_value
Order Payments	order_id, payment_sequential, payment_type, payment_installments, payment_value
Orders	order_id, customer_id, order_status, order_purchase_timestamp, order_approved_at, order_carrier_delivery_date, order_customer_delivery_date, order_estimated_delivery_date
Product Category Name Translation	product_category_name, product_category_name_english
Products	product_id, product_category_name, product_name_lenght, product_description_lenght, product_photos_qty, product_weight_g, product_length_cm, product_height_cm, product_width_cm
Sellers	seller_id, seller_zip_code_prefix, seller_city, seller_state

All these attributes give us a deeper understanding to the data set, while also giving us the opportunity to explore relationships among the variables.

And here is a brief diagram showing the relationship between these tables.

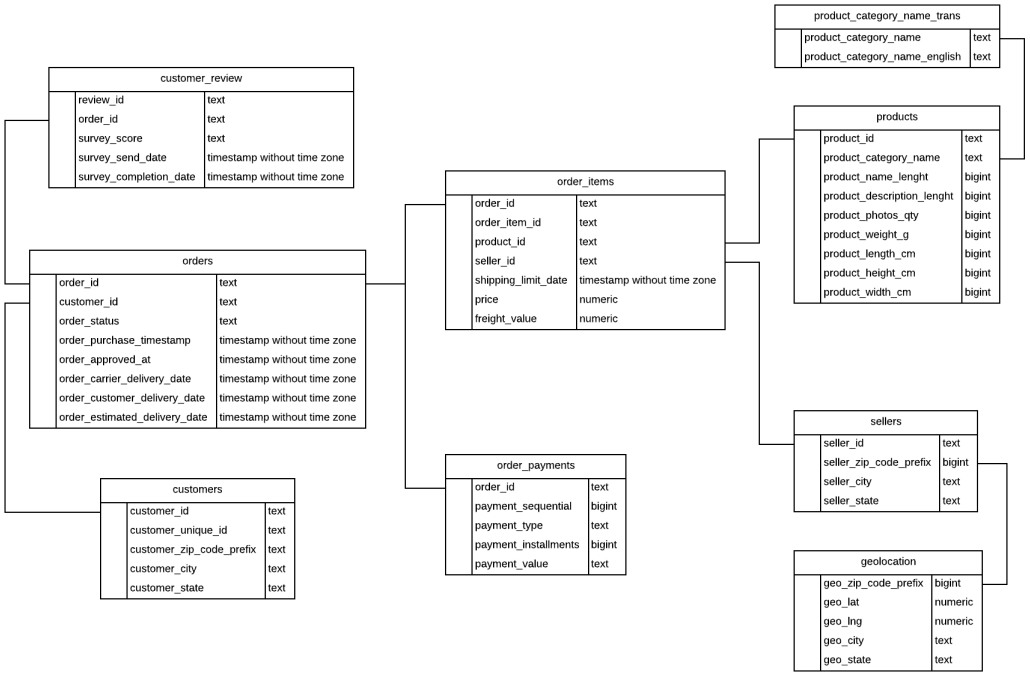


Figure 1: ETL Flow Diagram

2.2 ETL Data Preparation

We load the data into PostgresSQL using Pentaho Kettle. First off, in the input table, we change the columns with default data type of *Data* into *Timestamp* in order to be better prepared for future analyzing.
The other thing we do is to remove all the null values from our table for the purpose of cleaning up.
After changing types and cleaning up, the tables are outputted to Postgres database.

2.3 Postgres Database Analysis

We then answer the following analysis questions in SQL using Datagrip. Actual query commands are screenshots of results are in SQL.docx along with this report.

a. Which category of products receive most 5 ratings (survey_score)?

The category of products which receives most 5 ratings is *Health Beauty*, with a number of 5870.

b. What are the top 5 best selling products and what are their categories? What's the average rating of each of these products?

The top 5 best selling products are listed below.

Table 2: Top 5 Selling products

Product Id	Product Category	Average Score
389d119b48cf3043d311335e499d9c6b	garden_tools	4.11224
aca2eb7d00ea1a7b8ebd4e68314663af	furniture_decor	4.00759
422879e10f46682990de24d770e7f83d	garden_tools	3.94250
368c6c730842d78016ad823897a372db	garden_tools	3.91560
99a4788cb24856965c36a24e339b6058	bed_bath_table	3.86150

c. For which products are the buyers and sellers mostly from the same state?

For this question, we define the products that the buyers and sellers who are mostly from the same state as if two times the number of each products that of the same state is greater than the total number of sales of that product. On the other hand, we also want to eliminate those products with only one purchase.
Below is some of the examples that generated from our assumption.

Table 3: Products that the Buyers and Sellers Mostly from the Same State

Product Id	Proportion	Count Same	Total
001795ec6f1b187d37335e1c4704762e	0.667	36	54
00250175f79f584c14ab5cecd80553cd	0.909	100	110
003128f981470c3e5a2e7445e4a771cd	1	4	4
003c0b8f6580c850bd2e32044d2ac307	1	4	4
004636c889c7c3dad6631f136b7fa082	1	16	16

d. Give a regional break-up of total sales of heavier (weighing at least 1 kilogram) products.

We analyze this question by dividing the total sales of heavier products by the total sales of all the products according to each state. The region is defined by zip code.
Below is a brief overview of what the data set looks like:

Table 4: Regional Break-up of Total Sales of Heavier Products

Zip Code	Proportion
84925	1
14075	0.79245
5468	0.071428
14910	0.64102
95034	1

3 Exploratory Data Analysis

3.1 Trend Analysis

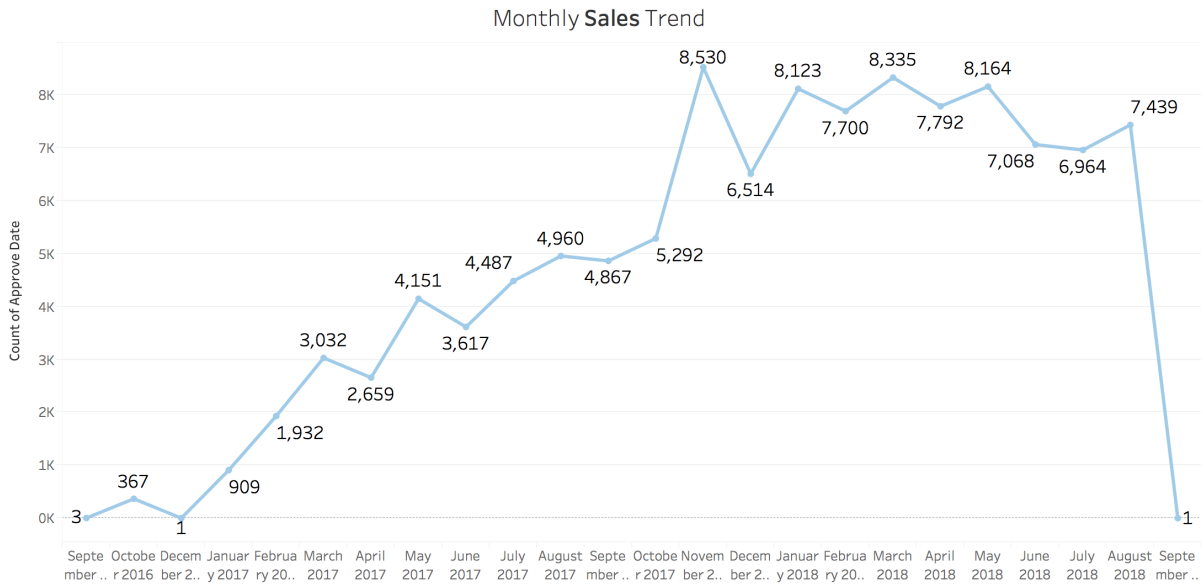


Figure 2: Monthly Sales growth

We first looked at the overall monthly sales across these two years. There's a very clear increasing trend in sales, which is obviously good news for the company. Also, given the dataset, we think there is no seasonality in terms of sales.

3.2 Geographics Analysis

Then, since we have geographic information, we mapped orders across state in order to identify if there's any observable difference .

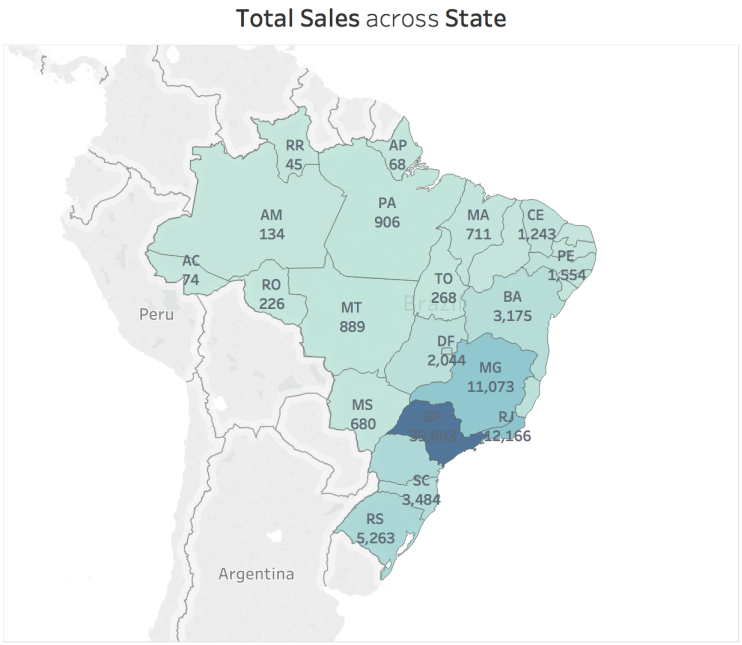


Figure 3: Distribution of Sales across states

We're not surprised to see this distribution in that this is indeed corresponding to the population distribution.
Thus further, we checked if the growth of sales is different across states:

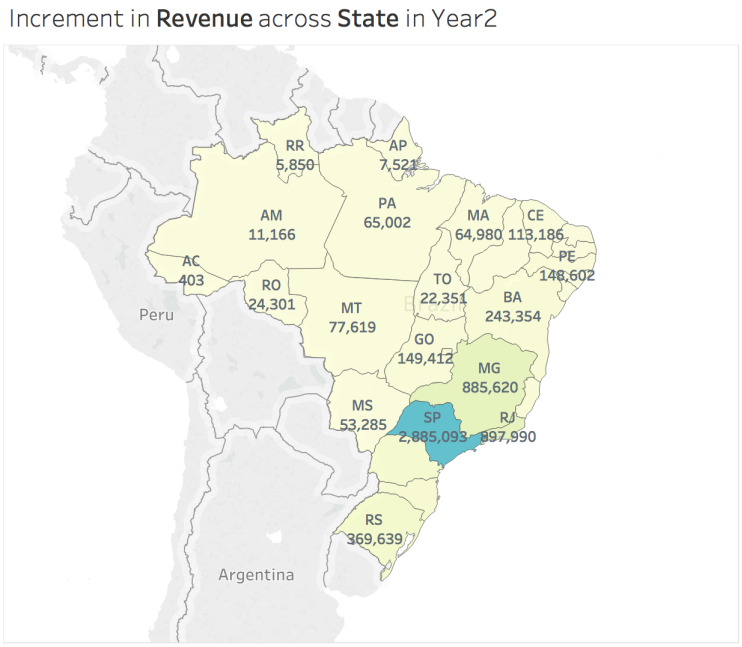


Figure 4: Increment in Sales across States

We found that the growth in sales is unbalanced across states. Growth in states where population density is high is much higher than states where population density is low. In addition on the states level, we also checked the average score given by customers from different states.

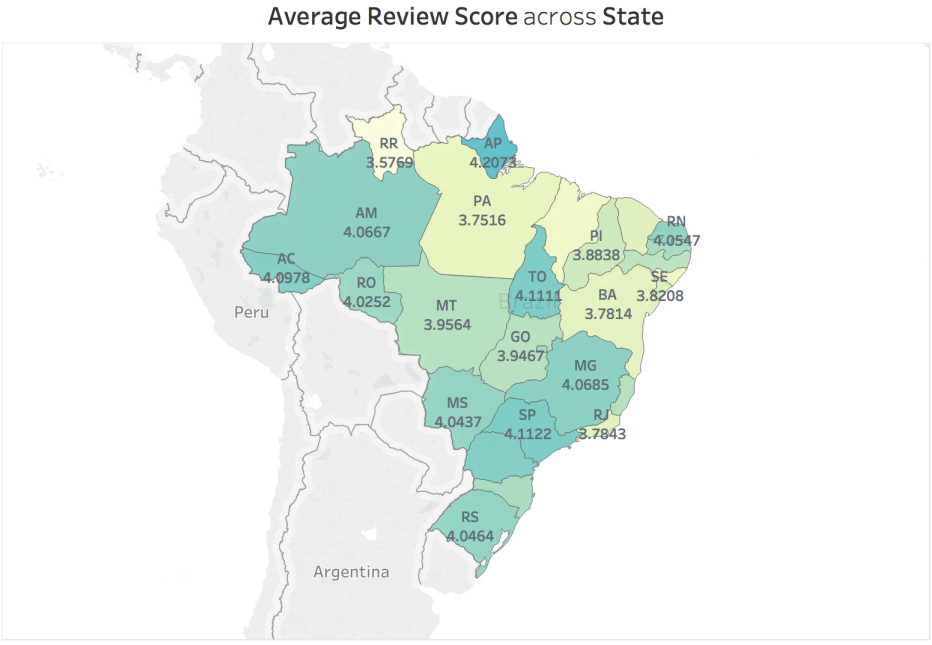
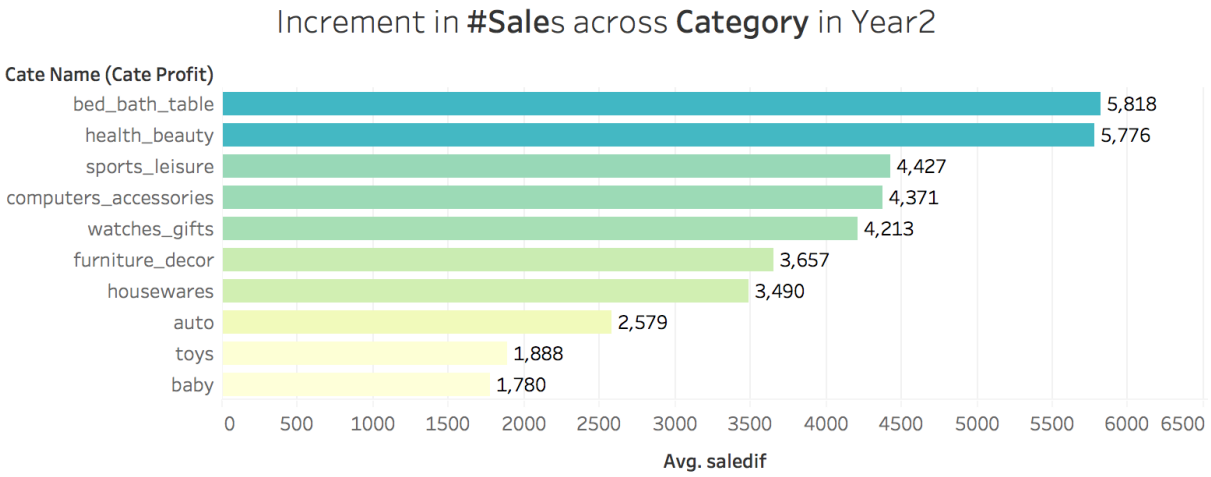


Figure 5: Screen shot of animation

It is interesting to notice that in states where total number of sales is high, the average scores given by customers from these are also higher. Unfortunately with no other information about customers we can't find out why this is the case.

3.3 Increments

Since we have two years of historical data, we're interested in finding out which type of products grow faster than other products in terms of sales and revenue.



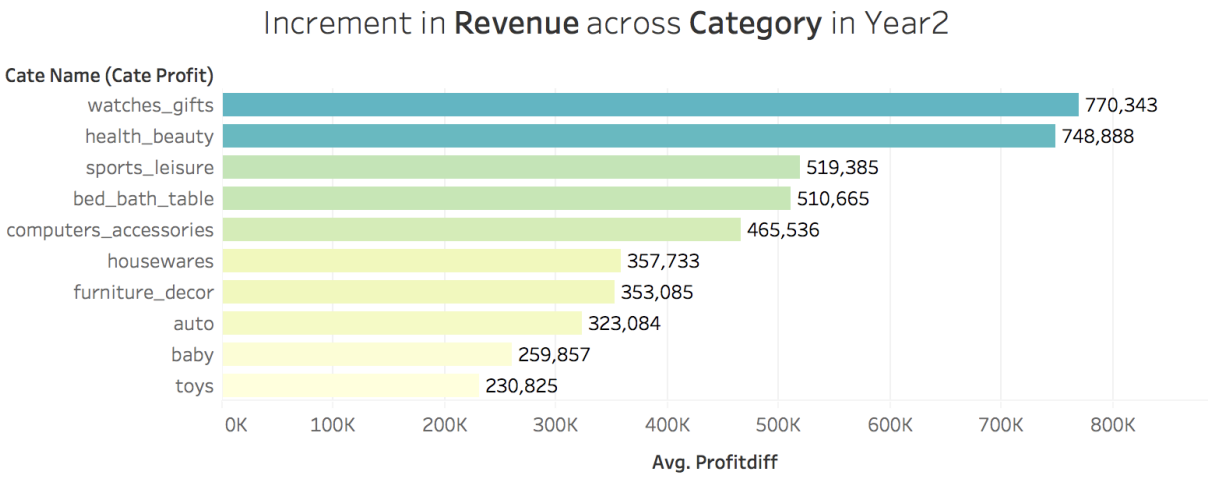


Figure 6: Comparison of Increments

It seems that sales of $\text{bed}_b ath_t able products increased tremendously in the second year, though it didn't perform that well i$

3.4 Total Sales Average Survey Score across Category

Then, continuing exploration with different category, we looked at total number of sales as well as average survey score. Here we observed something interesting.

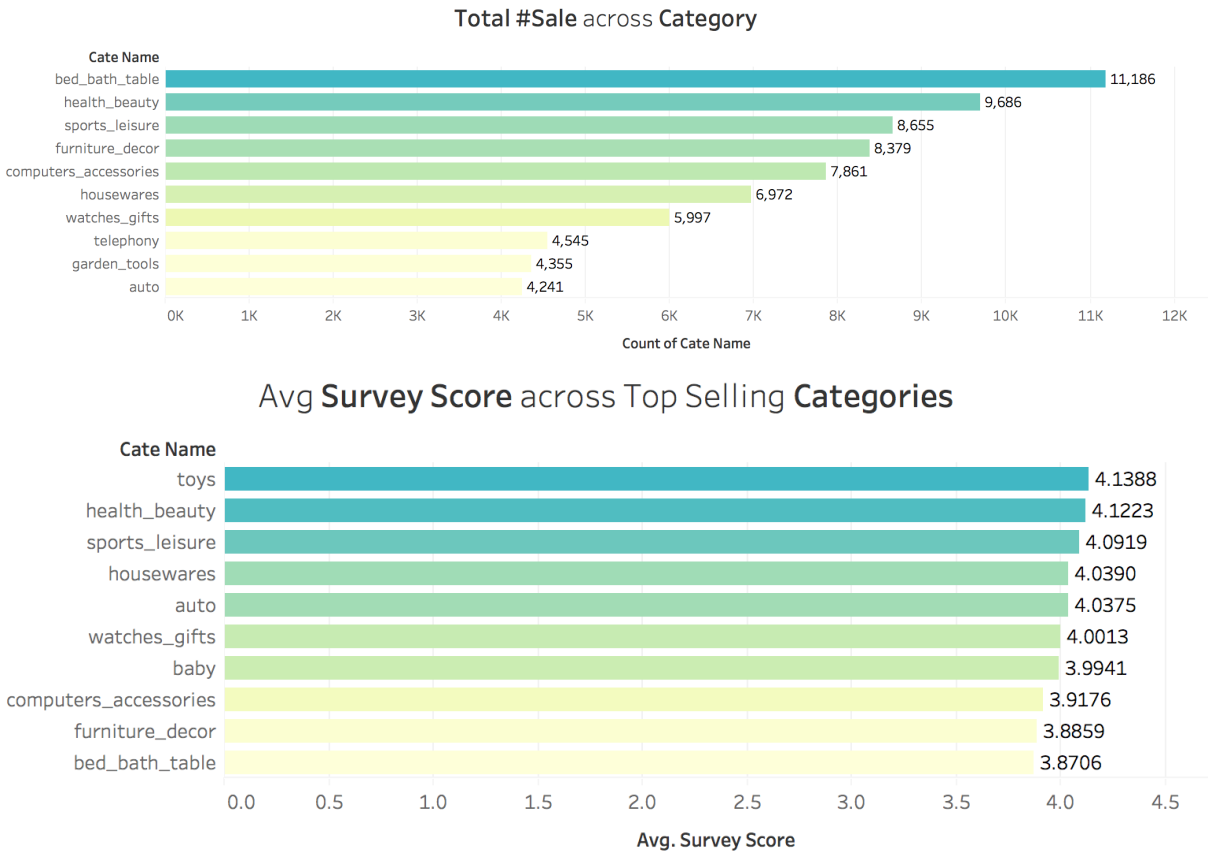


Figure 7: Comparison of Increments

While bed bath table products result in the highest total sales, they have relatively lower review score. Thus we begin wonder if low survey score will affect future sales?

In order to answer the question, for products that are purchased in the first year, we calculate their average survey score and when the average core is above 3 we label them as high otherwise low, in order

to compare with its performance in the second year.

Figure 8: Comparison of Increments

Looking at products in the 'low score group', their average increment in sales is significantly lower than products with high score, and the difference remains the same in revenue. That's why we decided to proceed our analysis and modeling on predicting review score. Because if we're able to know what drives survey score, we will be able to design precautionary actions in order to satisfy the customer.

3.5 Survey Score Distribution

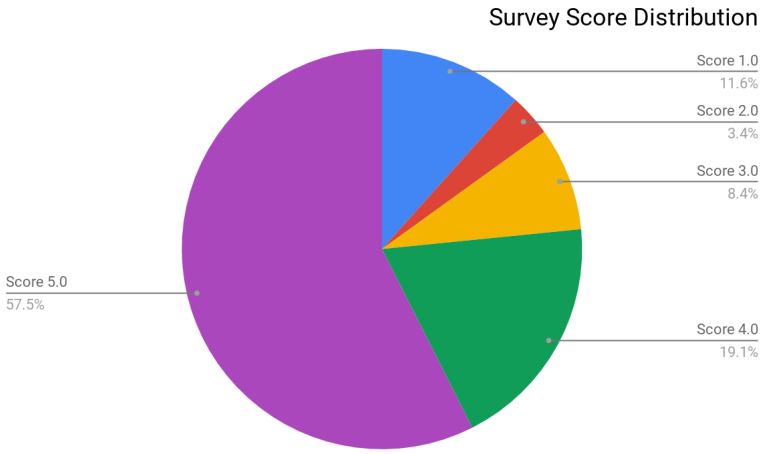


Figure 9: Comparison of Increments

Let's look at the distribution of survey score across products. Customers are generous on giving good survey scores, almost eighty percent products receive score of 4 or above. To find out the drivers of

survey score, intuitively, we assume that a customer will not be happy when an order is delivered later than estimated arrival, so we compared the average score based on whether the product is delivered late.

Figure 10: Impact of late delivery on survey score and sales

The above two charts tell us that most products are delivered on time, but when the products are delivered late, they're likely to experience a huge drop in average survey score.

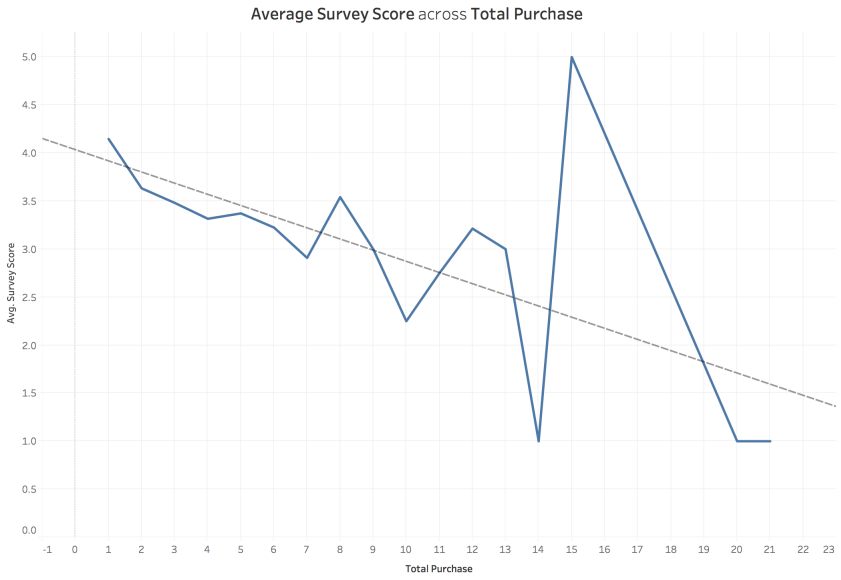


Figure 11: Impact of late delivery on survey score and sales

Besides, we also checked if our frequent buyers are likely to give high ratings or not. It seems that as total number of purchase increases, the standard deviation is increasing and average score given by the customer is decreasing.

4 Modeling

After loading our data, we need to split the entire data as train/test data. In our implementation, we used 75% of our whole data as training, 10% as validation, 15% as testing. As for the model we choose. Since we do not know the best model for this regression problem, we compare three different models.

4.1 Regression - Predict actual score value

The reason we choose regression as our model is that we like to predict the actual value of the survey score instead of classifying or clustering the data.

4.2 Data preparation

4.2.1 Features

- Price: the price of the product
- Freight Value: the freight value of each item
- Product ID: product name
- Product category: the category of each product
- Deliver - Estimate interval: duration between the actual delivery time and the estimate delivery time

4.2.2 Label

- Survey Score

4.2.3 Feature Engineering

- Onehot-encoding and down-sampling

The reason we choose these features is been discussed in the previous section. However, I liked to talk about re-sampling method, re-sampling is crucial for unbalanced data. There are lots of re-sampling methods, such as up-sampling, down-sampling or SMOTE etc.. We use down-sampling as our re-sampling method. First we calculate the smallest number of the score value (which is 2) and then randomly choose the data from the different score value to make the data balanced 16.

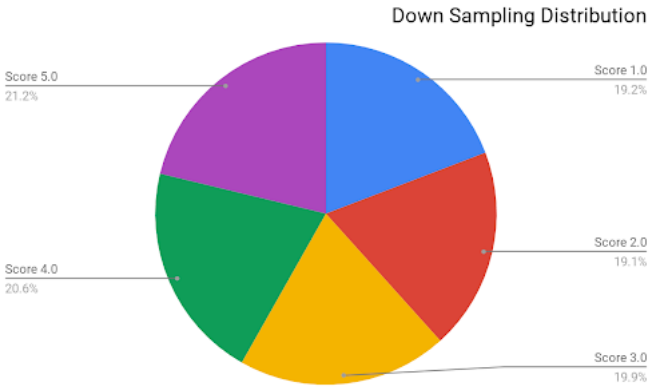


Figure 12: resampled data distirbution

4.3 Model

Since we do not know which is the best model for this regression problem, what we done is comparing three models, which are **Decision Tree**, **Multilayer Perceptron** and **Convolutional Neural Network**.

4.4 Result - Loss

Model	DT	MLP	CNN
Training Loss	NA	0.66	1.01
Validation Loss	1.35	0.967	0.997

From the table above, we can conclude that **Decision Tree** has the worst performance, and **Multilayer Perceptron** has the best. **Convolutional Neural Network** has the best training curve. Since MLP encountered the problem **overfitting**.

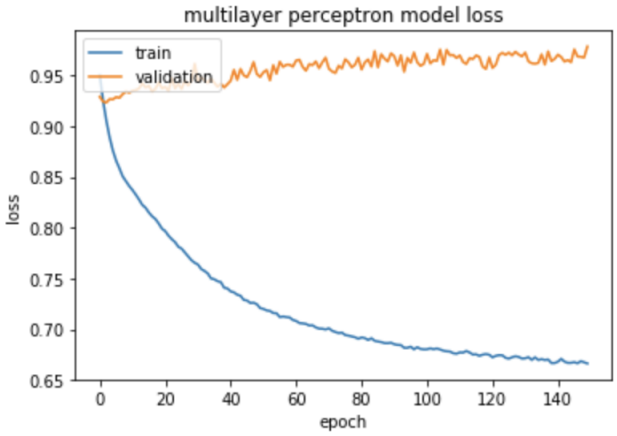


Figure 13: MLP MSE loss

From Fig.13, the training loss is converging but the validation loss is still high, so this might have some problem which is called **overfitting**. The model isn't training at all.

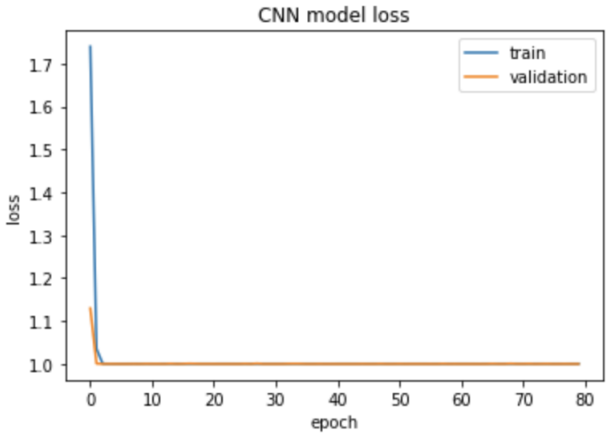


Figure 14: CNN MSE loss

From Fig.14, both the training and validation loss converged after five epochs. However, the loss is around 1. This means when we predict a score 3, the true value might be around 2 to 4. We can conclude that the dataset is too small to train well because of the method of re-sampling we use is down-sampling. Thus, we might want to use upsampling for another try.

4.5 Result - Prediction

When using our model for predicting the actual review score, first we need to see the distribution of the testing data. The distribution must look like the same as the original distribution.

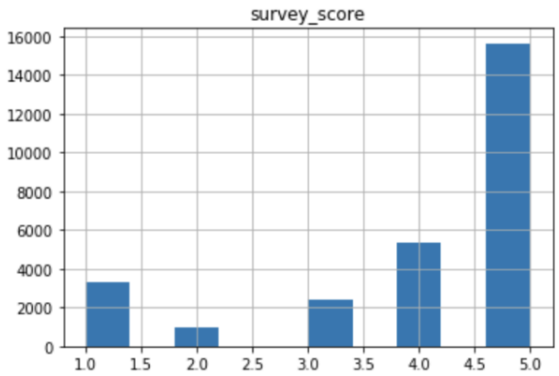


Figure 15: Testing data distribution

After predicting the score, we can find out that the result from our model predicts almost all the products with the score over five. Which means that our model (MLP) isn't working and we need to do something to improve it.

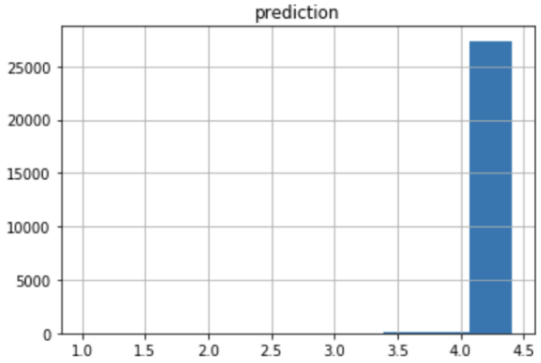


Figure 16: Testing data prediction

5 Extension

We also did extension on product category classification because we think recommendation is valuable for e-commerce industry.

5.1 Classification - Recommend Products to Customers based on Category

5.1.1 Features

- Frequency. Count of orders by customer and product category.
- Recency. Date part of (last order date by customer and product customer - last order date of the whole dataset), in days format
- Monetary. Sum of (order units * price) by customer and product category
- Average Customer Review Scores
- Customer State

5.1.2 Label

- Product category

5.1.3 Feature Engineering

- Onehot-encoding states and product category to index

5.1.4 Recommendation Accuracy

- Train - 0.34
- Test - 0.15

6 Summary

Based on our analysis, EDA and modeling, we can conclude in two points.

- 'Health Beauty' receives most 5 ratings and also is high in sales.
- The best model is multilayer perceptron model. After 150 epochs, the loss became 0.66. However, if we look at validation loss, the loss fluctuate. One explanation is the model isn't big enough to predict well. Therefore, we need to gather more information (features) or consider another resampling method in order to predict the survey score with higher accuracy.

7 Recommendation

We gained several business insights and we're now able to provide companies with the following two suggestions, based on our summary:

- More focus on 'Health Beauty'. Because 'Health Beauty' is a popular product category and also received higher review scores. We can invest more on this product category and maximize the sales, e.g. maintain relationships with good sellers and encourage these sellers to include more products of this type through our platform.
- More information of data. As data is limited, we cannot capture the useful information to help us understand the score. We suggest to include more feature(e.g. gender, age, return frequency) in the customer table.