Scientific Research Publishing

Intelligent Information Management, 2021, 13, 31-49
https://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

# A Data-Driven Research of Sales and Purchases on JD.com Platform

**Xiangyu Huang**

Shanghai Starriver Bilingual School, Shanghai, China
Email: b20140434@163.com

**How to cite this paper:** Huang, X.Y. (2021) A Data-Driven Research of Sales and Purchases on JD.com Platform. *Intelligent Information Management*, **13**, 31-49.
https://doi.org/10.4236/iim.2021.131002

**Received:** November 2, 2020
**Accepted:** January 17, 2021
**Published:** January 20, 2021

## Abstract

Unlike consumers in the mall or supermarkets, online consumers are "intangible" and their purchasing behaviors are affected by multiple factors, including product pricing, promotion and discounts, quality of products and brands, and the platforms where they search for the product. In this research, I study the relationship between product sales and consumer characteristics, the relationship between product sales and product qualities, demand curve analysis, and the search friction effect for different platforms. I utilized data from a randomized field experiment involving more than 400 thousand customers and 30 thousand products on JD.com, one of the world's largest online retailing platforms. There are two focuses of the research: 1) how different consumer characteristics affect sales; 2) how to set price and possible search friction for different channels. I find that JD plus membership, education level and age have no significant relationship with product sales, and higher user level leads to higher sales. Sales are highly skewed, with very high numbers of products sold making up only a small percentage of the total. Consumers living in more industrialized cities have more purchasing power. Women and singles lead to higher spending. Also, the better the product performs, the more it sells. Moderate pricing can increase product sales. Based on the research results of search volume in different channels, it is suggested that it is better to focus on app sales. By knowing the results, producers can adjust target consumers for different products and do target advertisements in order to maximize the sales. Also, an appropriate price for a product is also crucial to a seller. By the way, knowing the search friction of different channels can help producers to rearrange platform layout so that search friction can be reduced and more potential deals may be made.

## Keywords

E-Tailing, Data-Driven Research, Sales, Price-Discrimination, Search-Friction, Channels, Consumer Behavior

## 1. Introduction

In 2018, an estimated 1.8 billion people worldwide purchase goods online. This number is still increasing dramatically in recent years [1]. Total global online store sales are expected to reach the $2 trillion mark by the end of 2019. The increase will be 6 per cent compared with 2017 [2]. The Asia-Pacific region is home to the largest and fastest growing e-commerce market, with total online retail revenues set to nearly double from $733bn in 2015 to $1.4tn by 2020, with China being the largest of these, accounting for 80 per cent of the Asia-Pacific online retail market [3]. The growth of e-commerce retailing (or E-tailing) has given rise to many new and challenging problems at both strategic and operational levels.

In the context of E-tailing, this paper connects and contributes to three streams of literature. First, this research is related to a large literature on the determinants of consumers' online purchase behavior such as attitute, facilitating conditions, perceived usefulness, enjoyment, social pressure, transaction security, etc. [4]-[8]. For example, Tontini finds that the quality of e-services and online services are generally considered to be key determinants of competitive advantage and are related to measuring customers' purchase intention [4]. These studies have made important contributions to the understanding of the key factors in online shopping. However, most of researches adopted a questionaire survey and didn't take full advantage of the benefits of big data. Collecting sale data from JD.com, this research tries to comprehensively depict the digital image of users.

The second related stream of research is the operations management literature on dynamic pricing, which usually focuses on building pricing models and algorithms in different industries to maximize profits. This paper provides more evidence to design price strategies for sellers and platforms. For example, research on the pricing strategy of a firm to determine which price strategy, either dynamic pricing or preannounced pricing, is most beneficial to a firm [9]. Some other researches are focused on price promotions on retailing platforms and their effects on consumers' behaviors in both long-run and short-run period [10]. Moreover, some research is focused on matching rate and the factors that may affect it, including market thickness [11]. Dynamic pricing is based on the accurate understanding of consumers' consumption behavior. According to the acquired understanding of consumers, this paper also puts forward relevant suggestions. As you can see, many researches have focused on seller side in the e-tailing market, including pricing, discounts motivations and matching rate. However, in my research, the focus is more on the consumer side, revealing how different characteristics of products and platforms will affect consumers' purchasing behavior.

This research also provides insights for retailers about how they can increase their sales volume and revenue by allocating resources in different channels. Xu et al found results demonstrate that the tablet channel acts as a substitute for the

PC channel while it acts as a complement for the smartphone channel [12]. Chen et al concluded that the online retailer's profit critically depends on customer loyalty [13]. Ronghui *et al.* suggested that an increase in product return probability or retailer cost of handling a returned product can be beneficial to retailers [14]. Compared with risk aversion and service orientation, the convenience orientation of customers is higher, which urges consumers to choose online channels instead of offline channels. Knowing the demand and consumers' behavior when browsing products through various channels (PC, mobile devices, etc.) will help sellers better set price and decide on which channel to post their products in order to attract the most amount of potential buyers, getting most clicks from them, and reducing as much search friction as possible.

Using JD.com's proprietary data which captures a "full customer experience cycle" that begins as soon as a customer starts browsing on the platform and ends when the customer receives the delivered products, I hope to address several relationships related to product sales:

1) the relationship between sales and consumer's characteristics;

2) the relationship between each product's sales amount and their respective attributes;

3) and the demand curve of different products. Furthermore, there are several other interesting aspects related to consumer behaviors and pricing strategy, including search friction and price discrimination.

The contributions to the literature are two-fold. First, this study utilizes the big data to depict customers' characteristics in online retails, different from the questionnaire survey many previous studies adopted. Second, it provides a framework for sellers to analyze the users' behavior and design effective marketing strategies including allocating proper resources in different channels and designing pricing strategies.

The main methodologies in the paper include linear regression and various data analysis approaches. The remaining part of this paper begins with the introduction of the data and interesting statistical observations resulted from explorative analysis (Section 2). Methods and procedures adopted to solve the three main questions are discussed in Section 3, 4, 5, and 6, respectively. Section 7 concludes the paper.

## 2. Explorative Data Analysis

In this research, there are 457,298 potential JD consumers observed purchasing 31,867 products. All the data is stored in five datasets: "SKUs" table, which contains all the information related to products including the brand it belongs to, its attributes, and date when it enters the market and leaves the market; "users" table, which contains all the information related to a user including his or her user level, age, marital status, education, and city level; "clicks" table, which contains every clicking information of a user on a product via which channel; and "orders" table, which includes the information of an actual order including the price and discounts of this order. After first observing all the datasets and constructing a bar chart analyzing the sales distribution of all the products, I can find that the sales are highly skewed. The result is shown below (This graph only contains products that quantity of sales is greater than 500).

In Figure 1, the horizontal axis represents the product ID, and the vertical axis represents product sales. Those selling more than 1000 pieces accounted for only 1.0% of the total, while those selling more than 100 pieces accounted for 8.0% of the total. The most sold product "SKU_ID" was "068F4481B3", selling 25,769 units, while the average sales of samples were only 73.060 units. It can be seen that the sales distribution of JINGdong's products is highly skewed to the right, with only a small part of the sample selling extremely high.

What's more, we can find price-discrimination during the whole research. Taking the discount price of each consumer when purchasing products as the dependent variable and the characteristics of consumers as the independent variable, the linear regression model is established and the following results are obtained.

In Table 1, attributes of consumers with P values less than 0.05 are statistically significant and can be used for analysis. As we can see, gender, age, and marital status are all the aspects determining the final price of a product, which suggests the possible existence of certain degree of price discrimination. Take age for an example, the older you are, the products will be less expensive on average. Also, for the marital status, you will get possibly a lower price if you are married than those not.

Based on these results, sellers can adjust product features, target consumers, prices, and time of product entry and exit to achieve higher sales and maximum revenue.

## 3. Product Sales Analysis

Linear regression is used to find the relationship between a consumer's characteristics (education, city level, purchasing power, gender, marital status, etc.) and the sales volume of the products they bought. The data used needs to be preprocessed before these variables can be added. In the "Users" table, each consumer

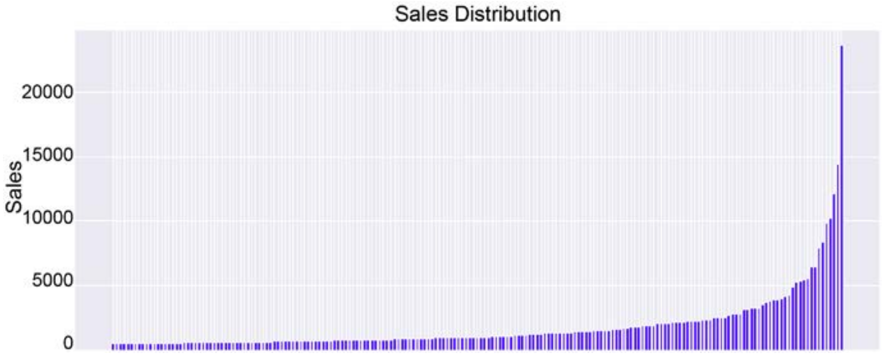


Figure 1. Sales distribution.

Table 1. Linear regression between discounted price and consumer's characteristics.

| | coef | std err | t | P > \|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 125.5511 | 2.015 | 62.790 | 0.000 | 122.601 | 130.502 |
| **user_level** | −0.1911 | 0.138 | −1.388 | 0.165 | −0.461 | 0.079 |
| **plus** | −0.0985 | 0.388 | −0.254 | 0.800 | −0.859 | 0.662 |
| **education** | 0.1348 | 0.204 | 0.662 | 0.508 | −0.264 | 0.534 |
| **city_level** | −0.1470 | 0.136 | −1.081 | 0.280 | −0.413 | 0.119 |
| **purchase_power** | −0.4686 | 0.305 | −1.537 | 0.124 | −1.066 | 0.129 |
| **length** | −0.0221 | 0.008 | −2.872 | 0.004 | −0.037 | −0.007 |
| **gender_F** | 43.2887 | 1.371 | 31.570 | 0.000 | 40.601 | 45.976 |
| **gender_M** | 42.1524 | 1.382 | 30.511 | 0.000 | 39.444 | 44.860 |
| **gender_U** | 41.1100 | 2.857 | 14.390 | 0.000 | 35.510 | 46.710 |
| **age_16-25** | 17.5876 | 3.325 | 5.289 | 0.000 | 11.070 | 24.106 |
| **age_26-35** | 18.3902 | 3.317 | 5.545 | 0.000 | 11.889 | 24.891 |
| **age_36-45** | 19.8418 | 3.330 | 5.959 | 0.000 | 13.315 | 26.368 |
| **age_46-55** | 17.8047 | 3.388 | 5.256 | 0.000 | 11.165 | 24.445 |
| **age_≤15** | 17.6498 | 21.393 | 0.825 | 0.409 | −24.282 | 59.582 |
| **age_≥56** | 16.7091 | 3.418 | 4.888 | 0.000 | 10.009 | 23.409 |
| **age_U** | 18.5677 | 4.706 | 3.945 | 0.000 | 9.343 | 27.792 |
| **marital_status_M** | 43.0995 | 0.881 | 48.908 | 0.000 | 41.372 | 44.827 |
| **marital_status_S** | 42.7529 | 0.876 | 48.820 | 0.000 | 41.036 | 44.469 |
| **marital_status_U** | 40.6986 | 1.105 | 36.821 | 0.000 | 38.532 | 42.865 |

has its own user ID and is given a number or letter in each condition to represent its user characteristics. The first five lines of the "users" Table 2 are shown below.

Some features are represented by specific Numbers, such as "user_level", which has a value of 0, 1, 2, 3, or 4, where a higher "user_level" is associated with a higher total purchase value in the past. If the consumer is a JD Plus member, the value of the "PLUS" column is 1, otherwise it is 0. "education" is valued according to consumers' education level, the greater the number, the higher the education level. "city_level" ranges from 1 to 5, with greater numbers representing less industrialized cities. "purchase power" is also ranged from 1 to 5, with greater numbers symbolizing lower purchasing ability. Figure 2 shows the percentage of users having different values under different categories.

A dummy variable is a numeric variable that represents categorical data. For example, in the "users" table, the gender of the user with the "user_ID" of "000089D6a6" is "F", indicating that the user is female. After being treated with dummy variables, the value of "gender_F" becomes "1" and the values of "gender_M" and "gender_U" become "0". By doing so, the computer will identify the user's gender. A portion of the "users" Table 3 after a dummy variable transformation is shown below.

Table 2. "Users" table after getting dummies example.

| | User_ID | User_level | First_order_month | plus | gender | age | Marital_status | education | City_level | Purchase_power |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 000089d6a6 | 1 | 2017-08 | 0 | F | 26 - 35 | S | 3 | 4 | 3 |
| 1 | 0000babd1f | 1 | 2018-03 | 0 | U | U | U | −1 | −1 | −1 |
| 2 | 0000bc018b | 3 | 2016-06 | 0 | F | ≥56 | M | 3 | 2 | 3 |
| 3 | 0000d0e5ab | 3 | 2014-06 | 0 | M | 26 - 35 | M | 3 | 2 | 2 |
| 4 | 0000dce472 | 3 | 2012-08 | 1 | U | U | U | −1 | −1 | −1 |

Table 3. "Users" table after getting dummies example.

| gender_F | gender_M | gender_U | age_16-25 | age_26-35 | age_36-45 | age_46-55 | age_≤15 | age_≥56 | age_U | maeital_status_M | maeital_status_S | maeital_status_U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |

After combining all the information in the "Users" and "Orders" Table 3 and Table 4, linear regression can be used to determine the relationship between the sale of the product purchased by the consumer and the characteristics of that consumer. The "Y" value is the total sales of the products purchased by each consumer, while the "X" value is the different characteristics of each consumer, including user level, marital status, gender, etc. The code for linear regression can be seen in Appendix (Code A).

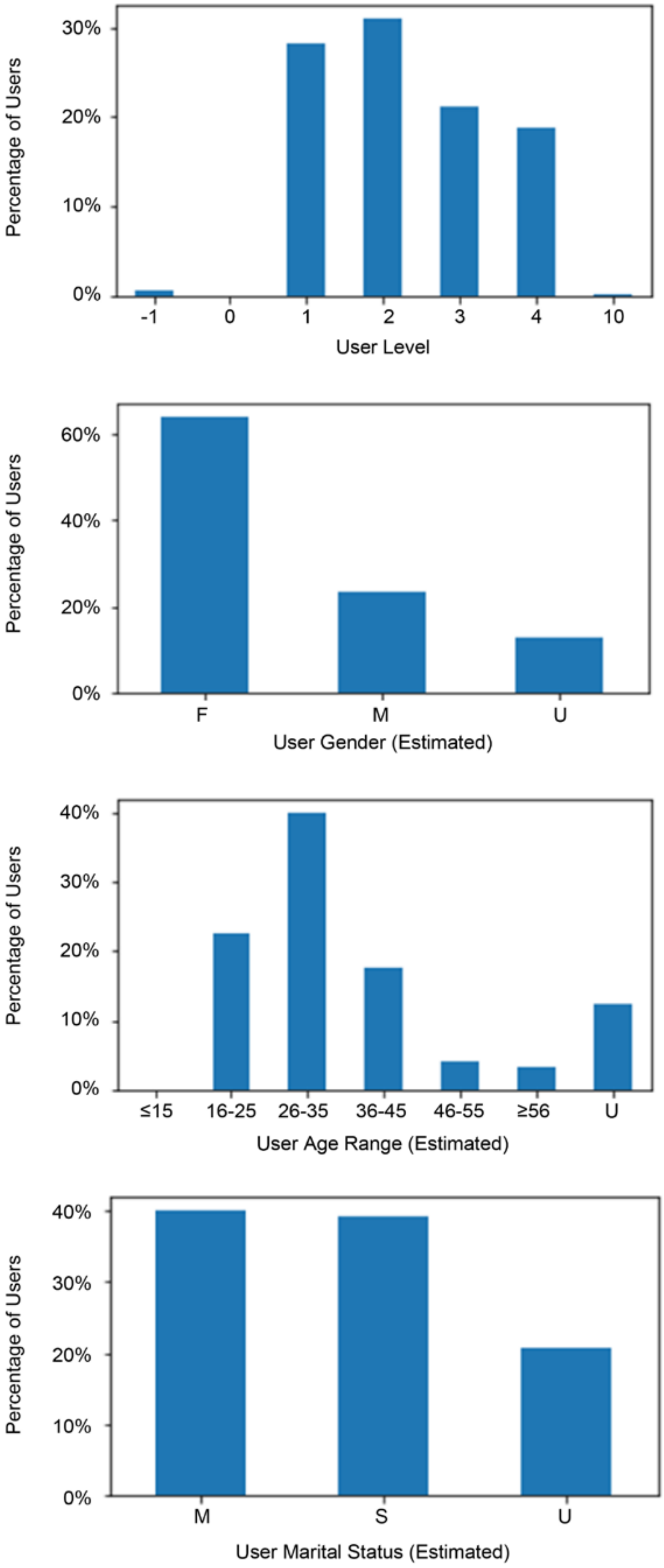
The result of the linear regression is as follows:

As you can see from Table 5, "user_level", "city_level", "purchasing_power", "gender", and "marital_status" are not statistically significant. This indicates that JD + membership, education level and age have no obvious relationship with product sales, and these three factors have little impact on sales. The higher the user level, the higher the consumption. Consumers living in industrialized cities tend to bring in higher sales. In terms of marital status, single people tend to buy more than married people.

## 4. Quantity of Sales Analysis

In this section, the main goal is to find the relationship between product sales and their respective attributes. When working on this, the "SKUs" Table 6, which contains all the information of a typical product, becomes important. The first five lines of the "SKUs" Table 6 are shown below.

In Figure 3, the first attribute takes an integer value between 1 and 4 (unknown data is denoted by −1), and the second ranges between 30 and 100


Percentage of Users
30%
20%
10%
0%
-1
0
1
2
3
4
10
User Level
Percentage of Users
60%
40%
20%
0%
F
M
U
User Gender (Estimated)
Percentage of Users
40%
30%
20%
10%
0%
≤15
16-25
26-35
36-45
46-55
≥56
U
User Age Range (Estimated)
Percentage of Users
40%
30%
20%
10%
0%
M
S
U
User Marital Status (Estimated)

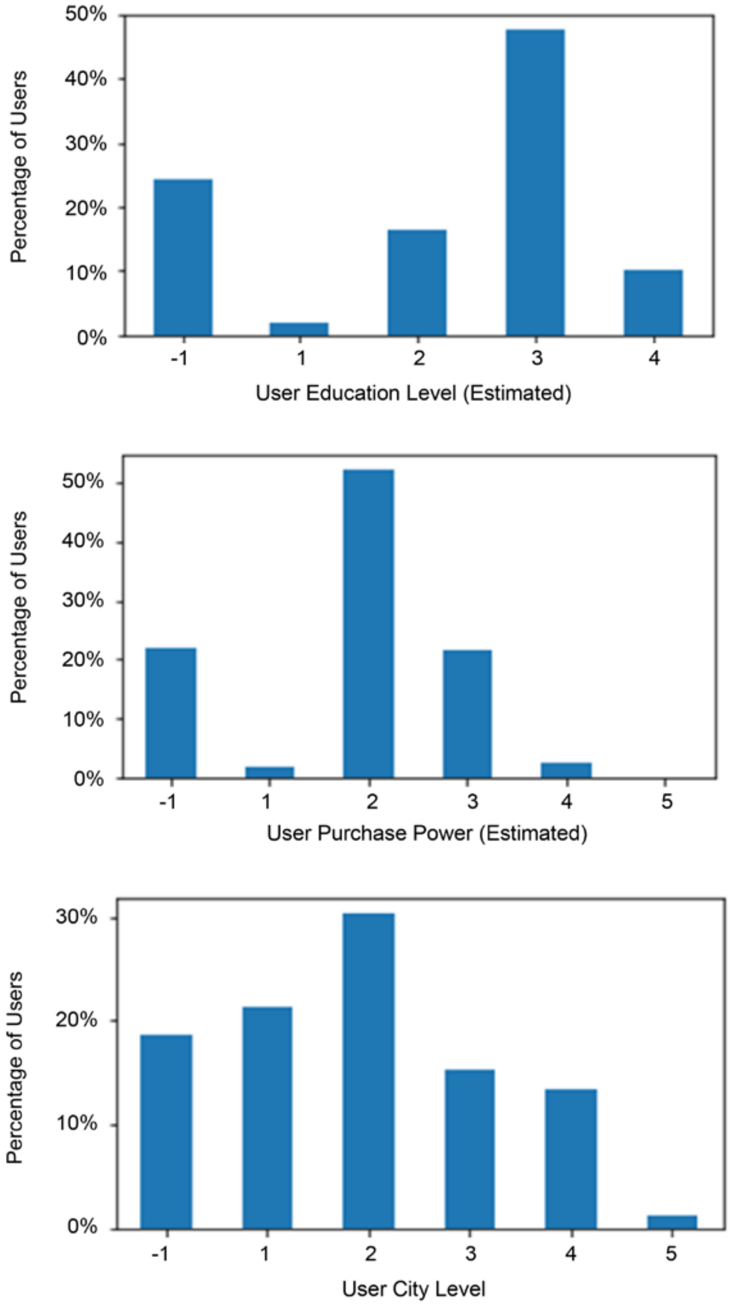


**Figure 2.** Distribution of consumers with different values under different categories.

(unknown data is also −1). For each attribute, a higher value indicates better performance of certain functionality. The distribution of all the products having different values under different attributes is shown below.

If linear regression is directly used by providing the value of attribute as independent variables and sales amount as dependent variables, the relationship cannot be fairly analyzed because for some products with the same attributes, due to its huge number of products carrying the same value for this attribute, the

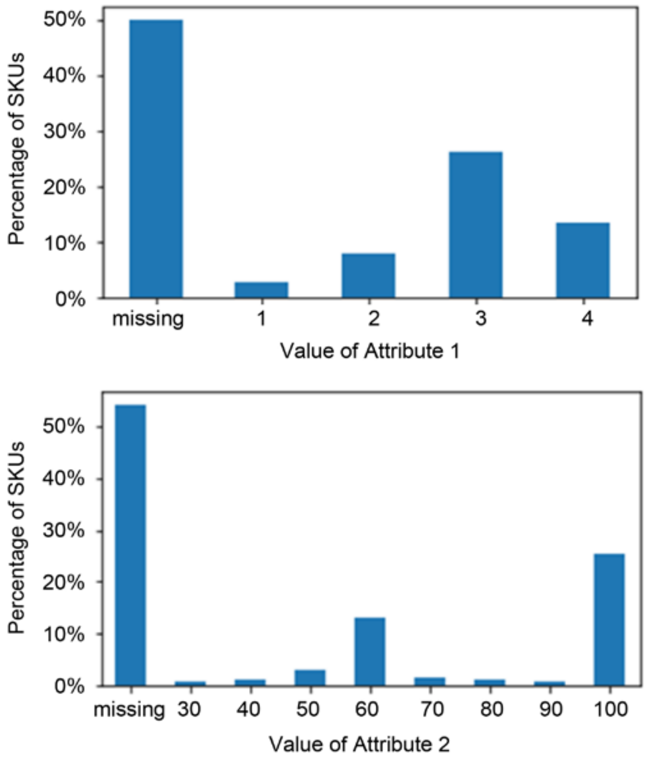


**Figure 3.** Distribution of products having different value under different attributes.

**Table 4.** "Orders" table example.

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **order_ID** | d0cf5cc6db | 7444318d01 | f973b01694 | 8c1cec8d4b | d43a33c38a |
| **user_ID** | 0abe3ef2ce | 33a9e56257 | 4ea3cf408f | b87cb736cb | 4829223b6f |
| **sku_ID** | 581d5b54c1 | 067b673f2b | 623d0a582a | fc5289b139 | 623d0a582a |
| **order_date** | 2018-03-01 | 2018-03-01 | 2018-03-01 | 2018-03-01 | 2018-03-01 |
| **order_time** | 2018-03-01 17:14:25.0 | 2018-03-01 11:10:40.0 | 2018-03-01 09:13:26.0 | 2018-03-01 21:29:50.0 | 2018-03-01 19:13:37.0 |
| **quantity** | 1 | 1 | 1 | 1 | 1 |
| **type** | 2 | 1 | 1 | 1 | 1 |
| **promise** | - | 2 | 2 | 2 | 1 |
| **original_unit_price** | 89 | 99.9 | 78 | 61 | 78 |
| **final_unit_price** | 79 | 53.9 | 58.5 | 35 | 53 |
| **direct_discount_per_unit** | 0 | 5 | 19.5 | 0 | 19 |
| **quantity_discount_per_unit** | 10 | 41 | 10 | 26 | 0 |
| **bundle_discount_per_unit** | 0 | 0 | 0 | 0 | 0 |
| **coupon_discount_per_unit** | 0 | 0 | 0 | 0 | 6 |
| **gift_item** | 0 | 0 | 0 | 0 | 0 |
| **dc_orl** | 4 | 28 | 28 | 4 | 3 |
| **dc_des** | 28 | 28 | 28 | 28 | 16 |

Table 5. Result of linear regression.

| | coef | std err | t | P > \|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 25.9750 | 3.461 | 7.504 | 0.000 | 19.191 | 32.759 |
| **user_level** | 20.8349 | 0.377 | 55.205 | 0.000 | 20.095 | 21.575 |
| **plus** | 0.8692 | 0.911 | 0.955 | 0.340 | −0.916 | 2.654 |
| **education** | 0.2536 | 0.333 | 0.762 | 0.446 | −0.399 | 0.906 |
| **city_level** | −0.8834 | 0.247 | −3.580 | 0.000 | −1.367 | −0.400 |
| **purchase_power** | −2.5657 | 0.493 | −5.199 | 0.000 | −3.533 | −1.598 |
| **length** | −0.2916 | 0.017 | −17.394 | 0.000 | −0.324 | −0.259 |
| **gender_F** | 12.9252 | 2.419 | 5.342 | 0.000 | 8.183 | 17.667 |
| **gender_M** | −1.4812 | 2.456 | −0.603 | 0.547 | −6.295 | 3.333 |
| **gender_U** | 14.5310 | 5.095 | 2.852 | 0.004 | 4.544 | 24.518 |
| **age_16-25** | 5.5536 | 5.702 | 0.974 | 0.330 | −5.623 | 16.730 |
| **age_26-35** | 7.6701 | 5.691 | 1.348 | 0.178 | −3.484 | 18.824 |
| **age_36-45** | 7.4198 | 5.711 | 1.299 | 0.194 | −3.774 | 18.614 |
| **age_46-55** | 9.9128 | 5.813 | 1.705 | 0.088 | −1.480 | 21.305 |
| **age_≤15** | −12.6273 | 36.623 | −0.345 | 0.730 | −84.407 | 59.152 |
| **age_≥56** | 3.0317 | 5.856 | 0.518 | 0.605 | −8.447 | 14.510 |
| **age_U** | 5.0144 | 8.198 | 0.612 | 0.541 | −11.053 | 21.081 |
| **marital_status_M** | 7.2924 | 1.481 | 4.925 | 0.000 | 4.390 | 10.195 |
| **marital_status_S** | 15.7439 | 1.490 | 10.570 | 0.000 | 12.824 | 18.663 |
| **marital_status_U** | 2.9388 | 1.787 | 1.645 | 0.000 | −0.563 | 6.441 |
| **Omnibus:** | 2,357,776.594 | | **Durbin-Watson** | | 1.999 | |
| **Prob (Omnibus):** | 0.000 | | **Jarque-Bera (JB):** | | 131,442,555,492,929.266 | |
| **Skew:** | 245.365 | | **Prob (JB):** | | 0.00 | |

Table 6. “SKUs” table example.

| | sku_ID | type | brand_ID | attribute1 | attribute2 | activate_date | deactivate_date |
|---|---|---|---|---|---|---|---|
| **0** | a234e08c57 | 1 | c3ab4bf4d9 | 3.0 | 60.0 | NaN | NaN |
| **1** | 6449e1fd87 | 1 | 1d8b4b4c63 | 2.0 | 50.0 | NaN | NaN |
| **2** | 09b70fcd83 | 2 | eb7d2a675a | 3.0 | 70.0 | NaN | NaN |
| **3** | acad9fed04 | 2 | 9b0d3a5fc6 | 3.0 | 70.0 | NaN | NaN |
| **4** | 2fa77e3b4d | 2 | b681299668 | - | - | NaN | NaN |

quantity of sales will be undoubtedly greater than those only have a relatively small numbers of products carrying the other value for the same attribute, that is, the sales amount is distorted.

To understand this, take “attribute 1” as an example as shown in Table 7. There are 813 products with the value of 1 and 2491 products with the value of 2 for “attribute 1” There are 7952 quantities of sales for all the products with the value of 1 for “attribute 1” and 91,708 quantities of sales for all the products with

Table 7. The number of products and their sales under certain value for "attribute 1".

| Values | The quantity of products | The quantity of sales |
|---|---|---|
| 1 | 813 | 7952 |
| 2 | 2491 | 91,708 |
| 3 | 8351 | 260,601 |
| 4 | 4252 | 82,150 |
| −1 | 15,961 | 85,866 |

the value of 2 for "attribute 1". So, one may conclude that the greater the value for "attribute 1", the more the quantity of sales would be. However, this result is highly biased because there exists a possibility that it is the greater number of products having the value of 2 makes the quantity of sales to be greater than those with value of 1. So, in order to exclude this biased condition in this linear regression, I need to apply volume factor twice, one based on the quantity of products in "attribute 1" and one based on the quantity of products in "attribute 2". By dividing the quantity of sales for each value of "attribute 1" or "attribute 2" by the quantity of products for each value of "attribute 1" or "attribute 2" respectively, the above biased condition will be largely excluded. The codes I apply based on both attributes, "attribute 1" (Code B) and "attribute 2" (Code C), will be shown in APPENDIX.

The linear regression's independent variables include the value of the attributes and the dependent value is the sales amount. The result of linear regression done based on "attribute 1" which resolves the distortion is shown in Table 8.

The results of linear regression based on "attribute 2" for resolving distortion are shown in Table 9.

According to the results of both trials, the relationship between sales quantity and the value of an attribute will be more pronounced if the trial is unbiased based on that respective attribute. In other words, if the result is processed unbiased based on "attribute 1", then the relationship revealed between quantity of sales and attribute 1 will be more apparent than that of attribute 2. Since a smaller P value suggests higher significance. In the above table, for example, the P value for "attribute 1" is 0.000 when using "attribute 1" to avoid biased results, whereas the P value for "attribute 2" is 0.480 when using "attribute 2" to avoid biased results, suggesting that the result will be more significant if the unbiased adjustment is done based on respective attribute. In addition, from these two linear regression results, we can see that the lower the value of the two attributes, the higher the sales. This means that the better the performance and function of the product, the higher the sales.

## 5. Demand Analysis

In this section, demand curve will be the main focus. As shown in Figure 4, Demand Curve is a curve showing the seller the willingness to buy for each

Table 8. Results of the linear regression.

| | | | |
|---|---|---|---|
| **Dep. Variable:** | attribute 1 | **R-squared:** | 0.0 |
| **Model:** | OLS | **Adj. R-squared:** | 0.0 |
| **Method:** | Least Squares | **F-statistic:** | 18.0 |
| **Date:** | Tue, 25 Aug 2020 | **Prob (F-statistic):** | 3.66e− |
| **Time:** | 22:50:49 | **Log-Likelihood:** | −461 |
| **No. Observations:** | 8832 | **AIC:** | 9.221e+ |
| **Df Residuals:** | 8827 | **BIC:** | 9.225e+ |
| **Df Model:** | 4 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P > \|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| const | 6.7210 | 5.239 | 1.283 | 0.200 | −3.548 | 16.990 |
| type_x | 2.1030 | 2.627 | 0.801 | 0.423 | −3.046 | 7.252 |
| attribute1 | −2.0150 | 0.456 | −4.423 | 0.000 | −2.908 | −1.122 |
| attribute2 | −0.0289 | 0.016 | −1.766 | 0.077 | −0.061 | 0.003 |
| avtivate_date | 0.4424 | 1.679 | 0.263 | 0.792 | −2.849 | 3.733 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 20,512.942 | **Durbin-Watson** | 2.017 |
| **Prob (Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 171,025,144.444 |
| **Skew:** | 22.875 | **Prob (JB):** | 0.00 |
| **Kuetosis:** | 683.183 | **Cond.No** | 825. |

Table 9. Results of the linear regression.

| | | | |
|---|---|---|---|
| **Dep. Variable:** | attribute 2 | **R-squared:** | 0.0 |
| **Model:** | OLS | **Adj. R-squared:** | 0.0 |
| **Method:** | Least Squares | **F-statistic:** | 16.0 |
| **Date:** | Sat, 21 Mar 2020 | **Prob (F-statistic):** | 3.83e− |
| **Time:** | 21:47:11 | **Log-Likelihood:** | −4655 |
| **No. Observations:** | 8832 | **AIC:** | 9.312e+ |
| **Df Residuals:** | 8827 | **BIC:** | 9.315e+ |
| **Df Model:** | 4 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P > \|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **const** | 5.1338 | 5.515 | 0.931 | 0.352 | −5.677 | 15.945 |
| **type_x** | 2.4199 | 2.766 | 0.875 | 0.382 | −3.001 | 7.841 |
| **attribute1** | 0.2211 | 0.480 | 0.461 | 0.645 | −0.719 | 1.161 |
| **attribute2** | −0.0992 | 0.017 | −5.763 | 0.000 | −0.133 | −0.065 |
| **avtivate_date** | 0.5419 | 1.768 | 0.307 | 0.759 | −2.923 | 4.007 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 20,521.351 | **Durbin-Watson** | 2.019 |
| **Prob (Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 168,478,701.701 |
| **Skew:** | 22.908 | **Prob (JB):** | 0.00 |
| **Kuetosis:** | 678.073 | **Cond.No** | 825. |

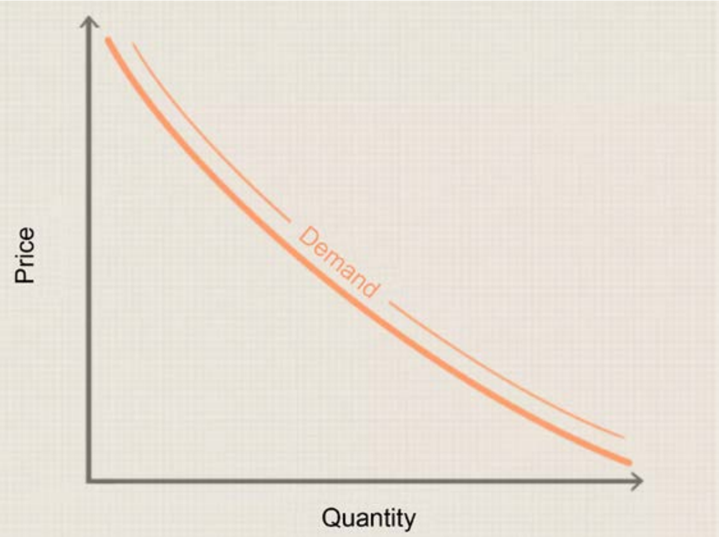


**Figure 4.** Demand curve example[1].

consumer under each price for a product. So, I choose the product with the most quantity of sales to do the research, which is the product with "sku_ID": "068f4481b3". In order to construct a demand curve, I need to first know the price when a consumer is viewing the product. Since the "SKUs" table (shown in section IV) contains several discounts that do not fully reflect the consumer's willingness to pay when seeing the product, including bundle discount, coupon discount, and quantity discount, I use final unit price as each consumer's willingness to pay for a product.

In order to discover for a specific product, the price associated with most sales, which is captured by the demand curve, I need to process the data in a way that avoids biases similar to that discussed in section IV. There is a certain possibility that the sales quantity of a product is affected by the time period this price is shown to all consumers. For example, if a tube of toothpaste costs $25 for a year, meaning the price for it will be $50 if you see this product during the year, and the quantity of sales is 100. Then, its price changes to $25 for only a month and the quantity of sales is 10. Can you say that consumers' willingness to pay for this toothpaste is at $50 because this price level has more quantity of sales? Of course, the answer will be negative because this conclusion is based on consumers having a higher possibility of buying it at a price of $50 due to this price level's time period is longer. So, in order to exclude this bias, I apply the techniques as in part IV, which uses the frequency of a certain price for a product divide by the duration of time it is sold for that price as y-axis. The final unit price will be shown on the x-axis. By doing so, I can find out at which price a product has most sales in a less biased way.

The result is shown in **Figure 5**.

According to the results shown above, most people prefer the price to be

[1] https://www.investopedia.com/terms/d/demand-curve.asp

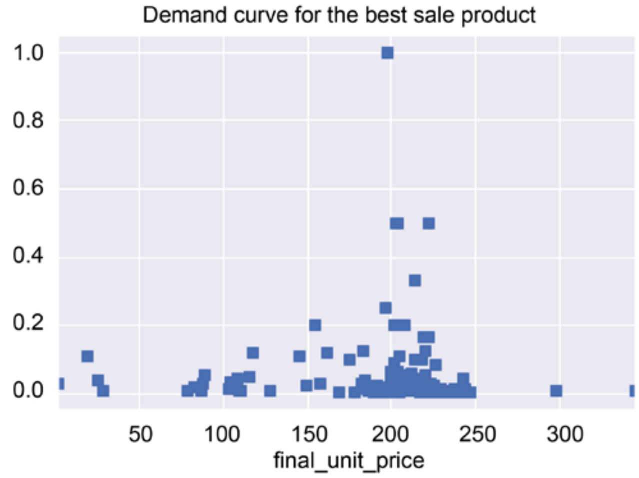


**Figure 5.** Demand curve for the product with "sku_ID" - "068f4481b3".

around 200 RMB for this product. Concerning prices from 0 RMB to 100 RMB and 250 RMB to 250 RMB, there are more people who prefer cheap prices than really expensive price. This result reveals consumer behavior. Consumers will prefer an "appropriate" price more than prices that are extremely high or extremely low. The reason may be: if a price is too cheap, although it becomes more affordable, the product may seem bad and low-quality in consumers' minds, leading to relatively lower quantity of sales. If the price is too high, then it is too unaffordable and exceeds the value in consumers' minds. As a result, only an appropriate price will lead to better quantity of sales. After seeing the best sold products, the demand curve for the top 5 best sold can also be taken into concern, as shown in Figure 6.

As you can see in Figure 6, there are prices which have relatively high quantity of sales (we call it the "best" price in the following paragraphs), proving that setting an appropriate price is crucial to a producer. An appropriate price, instead of a cheap price, will lead to a higher quantity of sales. Furthermore, there is something special and different if we look at the five graphs together. The "best" price is distributed at different points in the pricing range of the product. For example, for the best sold product, the "best" price is $200, which is located almost in the middle of the price range: $0 - $350. For the third one, the "best" price is located at price $40, to the right of the price range. For the rest, the "best" price occurs frequently, usually appears when price increases by a certain amount. This finding may imply the product type of those products. For "best" price only appears rarely and located in the middle or slightly to the right of the price range, the products may be food, electronic devices or those may have potential safety hazard, making human disbelieve extra-cheap price and refuse extra-expensive price. For "best" price appears frequently among the price range, the products may be toys or recreational equipment, that are valued usually different depending on the wealth of the consumers. In conclusion, knowing the

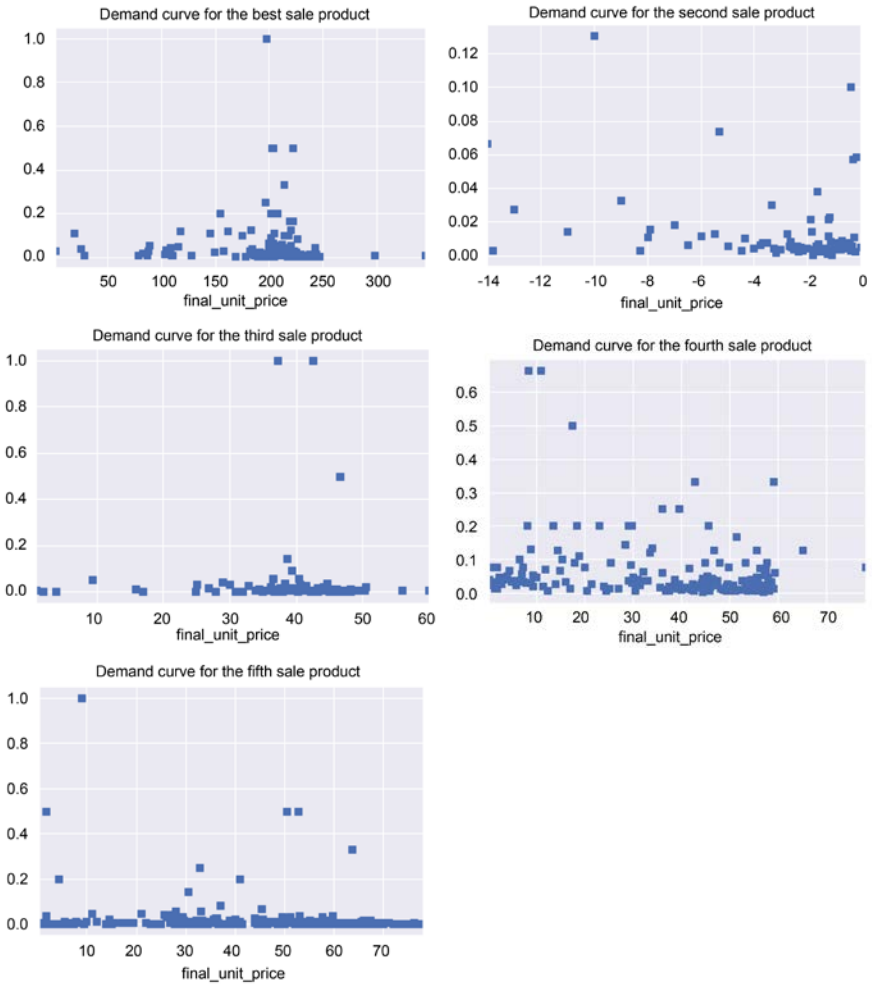


Figure 6. Demand curve for top 5 best sold products.

demand curve of several products may help producers characterize the products and specialize the pricing of each product.

## 6. Search Friction in the Online Marketplace

In this section, search-friction of different platforms will be the main topic. Search friction is determined by the amount of clicks a consumer is willing to spend on the product via this channel. Smaller search friction means a higher possibility of consumer's purchase due to more voluntary clicking. As a result, all manufacturers want their products to be more searchable. To determine the number of searches, the "Clicks" table is used to store the number of channels or platforms on which consumers are clicking on products. An example of the "Clicks" Table 10 is shown below.

In Figure 7, the distribution of clicks between different channels is shown. It turns out that app has the most hits, followed by WeChat. Another key indicator of search resistance across different channels is the time spent browsing products through that channel.

Table 10. “Clicks” table example.

| | sku_ID | user_ID | request_time | channel |
|---|---|---|---|---|
| **0** | a234e08c57 | 4c3d6d10c2 | 2018-03-01 23:57:53 | wechat |
| **1** | 6449e1fd87 | - | 2018-03-01 16:13:48 | wechat |
| **2** | 09b70fcd83 | 2791ec4485 | 2018-03-01 22:10:51 | wechat |
| **3** | 09b70fcd83 | eb0718c1c9 | 2018-03-01 16:34:08 | wechat |
| **4** | 09b70fcd83 | 59f84cf342 | 2018-03-01 22:20:35 | wechat |

Table 11. An example of the average time spent by consumers in different channels.

| channel | request_time |
|---|---|
| app | 2.353836 |
| mobile | 4.969490 |
| others | 3.970588 |
| pc | 3.097241 |
| wechat | 2.166059 |

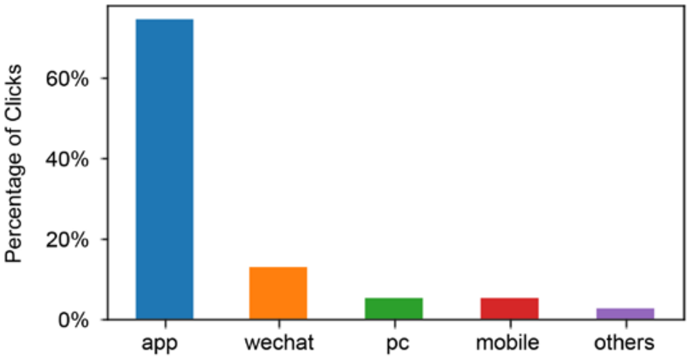
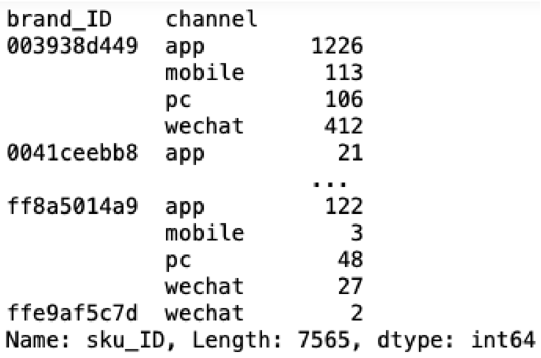


Figure 7. Click distribution among different channels.



```
brand_ID    channel
003938d449  app        1226
            mobile      113
            pc          106
            wechat      412
0041ceebb8  app          21
                        ...
ff8a5014a9  app         122
            mobile        3
            pc           48
            wechat       27
ffe9af5c7d  wechat        2
Name: sku_ID, Length: 7565, dtype: int64
```

Figure 8. Total number of clicks for a brand under different channels.

Table 11 shows that mobile has the longest time a consumer spent on clicking, which suggests that it has the lowest search friction; whereas WeChat has the highest search friction. By knowing the search friction in different channels, sellers can efficiently decide which channel could be the best for a specific product and which channel is needed to be adjusted in order to reduce search friction.

Figure 8 shows the total number of clicks for a brand under different channels. Utilizing this information, sellers can clearly know the search friction of different channels for a brand and adjust its selling strategy or even improve channels with relatively higher search friction. Take the brand with "brand_ID" "003938d449" as an example, the clicking frequencies of app and WeChat are significantly larger than that of pc and mobile, indicating higher search friction in the pc and mobile channels. In this way, the seller who sells the product with this "brand_ID" can adjust his or her selling strategy accordingly. Maybe he or she can increase advertisement spent on app and decrease ads spent on pc or redesign and improve the website for pc users in order to make it more attractive and convenient to browse, thus reducing search friction.

## 7. Conclusions

In this research, I mainly focused on two aspects: consumers' purchasing behavior towards different products and strategy a producer can take to increase sales.

First focus is on how different consumer characteristics affect sales. Quantity of sales for products is highly skewed in e-tailing area, with extremely high quantity of sales for a tiny portion of all the products. Furthermore, I found that people live in a more industrialized city, with higher purchasing power and user level, being singles will lead to higher consumption on buying things. Moreover, a higher performance or better function of a product will always lead to higher quantity of sales.

Second focus is on pricing. I found that an appropriate price, not too expensive and too cheap, leads to higher quantity of sales of a product. Also, demand curve for several products helps producers characterize the products and specialize the pricing of each product. What's more, according to the results get after researching on search frictions of different channels, it is better to concentrate on app selling due to its lower search friction for consumers among most of the products. At the same time, some platforms with high search friction, for example, personal computer, can consider reforming the platform and rearranging layout so that search friction can be reduced.

After analyzing major results, I found throughout the research, there are still some improvements we can make for the further research. In this paper, we focus on linear regression methods to study the various relationships in the e-commerce data from JD.com. Other regression analysis, such as nonlinear regression and generalized linear models, could be utilized to gain more insights. Machine learning methods such as decision trees, nearest neighbors and random forests, can also be applied to make predictions on sales and consumer clicking and purchasing behavior.

In the demand curve analysis, an issue exists as the price data is censored—only prices of sold products are recorded. We hope to mitigate this issue by collecting more posted data in addition so as to get a full picture of the demand curve.

## Acknowledgements

I would like to extend my special thanks to my supervisor, Danye Ma, who provides me with class time to learn about the knowledge of economics and modeling, innumerable times of assistance, sufficient resources, and helpful advice for revising the thesis. I would also like to express my thanks to my teachers, who taught me useful methods for completing the program. Moreover, I am also very grateful for my parents, who emotionally support me when facing challenges and difficulties.

## Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

## References

[1] Kamberaj, B. (2019) The Importance of E-Commerce. https://doi.org/10.2139/ssrn.3434315

[2] Chaffey, D., Hemphill, T. and Edmundson-Bird, D. (2019) Digital Business and e-Commerce Management. Pearson UK, New York, 5-15.

[3] Chen, H.-M., Wu, C.-H., Tsai, S.-B., Yu, J., Wang, J.T. and Zheng, Y.X. (2016) Exploring Key Factors in Online Shopping with a Hybrid Model. *SpringerPlus*, **5**, Article No. 2046. https://doi.org/10.1186/s40064-016-3746-4

[4] Tontini, G. (2016) Identifying Opportunities for Improvement in Online Shopping Sites. *Journal of Retailing and Consumer Services*, **31**, 228-238. https://doi.org/10.1016/j.jretconser.2016.02.012

[5] Matatu, L.T. (2019) Determinants of Consumer's Online Purchase Behavior in Zimbabwe. *Journal of Management and Humanity Research*, **1**, 1-10

[6] Al-Maghrabi, Talal, et al. (2011) Determinants of Customer Continuance Intention of Online Shopping. *International Journal of Business Science & Applied Management* (*IJBSAM*), **6**, 41-66.

[7] Bulut, Z.A. (2015) Determinants of Repurchase Intention in Online Shopping: A Turkish Consumer's Perspective. *International Journal of Business and Social Science*, **6**, 55-63.

[8] Madahi, A. and Sukati, I. (2016) Determinants of the Channel Selection and Choice Intention: A Marketing Perspective. *Journal for Global Business Advancement*, **9**, 357-389. https://doi.org/10.1504/JGBA.2016.079882

[9] Li, J. and Netessine, S. (2019) Higher Market Thickness Reduces Matching Rate in Online Platforms: Evidence from a Quasiexperiment. *Management Science*, 66, 271-289.

[10] Zhang, D.J., Dai, H.C., Dong, L.X., Qi, F.F., Zhang, N.N., Liu, X.F., Liu, Z.Y. and Yang, J. (2019) The Long-Term and Spillover Effects of Price Promotions on Retailing Platforms: Evidence from a Large Randomized Experiment on Alibaba.

*Management Science*, **66**, 2589-2609.

[11] Papanastasiou, Y. and Savva, N. (2016) Dynamic Pricing in the Presence of Social Learning and Strategic Consumers. *Management Science*, **63**, 919-939.

[12] Xu, K.Q., Chan, J., Ghose, A. and Han, S.P. (2017) Battle of the Channels: The Impact of Tablets on Digital Commerce. *Management Science*, **63**, 1469-1492. https://doi.org/10.1287/mnsc.2015.2406

[13] Chen, L., Nan, G.F. and Li, M.Q. (2018) Wholesale Pricing or Agency Pricing on Online Retail Platforms: The Effects of Customer Loyalty. *International Journal of Electronic Commerce*, **22**, 576-608. https://doi.org/10.1080/10864415.2018.1485086

[14] Wang, R.H., Chen, L. and Li, M.Q. (2020) Channel Integration Choices and Pricing Strategies for Competing Dual-Channel Retailers. *IEEE Transactions on Engineering Management*. https://doi.org/10.1109/TEM.2020.3007347