Data-Driven Research: Factors Affecting Different Customers' Online Shopping Spending and Product Choice

Fangze Wei*

School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China

*Corresponding author: sunnywfz@126.com

Keywords: Customer behavior, online shopping, e-commerce, data analysis.

Abstract: E-commerce is a rapidly expanding industry due to the acceleration of the digital economy and the COVID-19 pandemic. Based on the increased concern of customers' consumption behaviors in online shopping, empirical evidence on factors affecting the online shopping spending and product choice of different customers is essential. In this study, various factors were examined and the chosen data provided by JD.com were analyzed by applying the reliability test, multivariable regression, mediation effect, and so on. The study suggests that statistically, consumers with higher education and who are married have a preference for relatively cheap products in general, whereas consumers with higher user level and purchase power and those with JD. Com plus membership contribute a larger online shopping spending and are likely to buy relatively expensive products from this platform. Other factors including discounts, price level, and city level can also affect a customer's product choice. The result of this research advises online shopping platforms to adopt targeted advertising to different customer groups by utilizing quantitative and qualitative market research.

1. Introduction

The rapid development of the Internet has facilitated the growth in e-commerce, which makes shopping online become increasingly popular in recent decades. As one of the largest e-commerce companies and online retailers in China, JD.com earned net revenue of $82.9 billion in 2019 [1]. Millions of customers from all over the world choose to buy products from this platform because of its reliability, efficient delivery, product diversity, and so on. Compared with the traditional form of shopping, customers get a completely different shopping experience when shopping online. Therefore, since customers cannot have an intuitive impression of products online, how do they make product choices among a large variety of products? What factors affect a customer's online shopping spending? Many scholars have conducted experiments and researches to conclude that a customer's consumption behaviors and product choice are associated with psychological, personal, and social factors. Psychological factors include a customer's beliefs, perceptions, values, and so on. Personal factors include a customer's age, culture, and other demographic information. Social factors include a customer's income level, education, social class, and so on [2].

Based on the data provided by JD.com, the objective of this research is to determine the factors that influence the online shopping spending and product choice of different customers, especially the ordinary customers and JD. Com's plus members. Compared with ordinary customers, JD.com's Plus members are expected to have higher user levels and purchase power, which corresponds to their higher expenditure on online shopping and choice of relatively expensive products [3]. Other affecting factors for a customer's product choice and spending include gender, age, education, marital status, city level, and so on [4]. This research project plans to analyze the chosen data by applying the reliability test, multivariable regression, mediation effect, and so on.

2. Methods

2.1 Data Selection

To begin with, the first task for this research project is selecting the available data from the seven data sets provided by JD.com, which contains transactional data for more than 2.5 million customers. The skus, users, and orders tables share several same identification variables, including sku_ID in the skus table, user ID in the users table, and order ID, user ID, and sku_ID in the orders table. Therefore, by searching and matching with the vlookup function in Excel, all of the variables and data in these three tables are combined together with these key identification variables. After the matching and combining process, 24 variables and 8832 observations are achieved for the following analysis. Moving on, the adaptation of the multivariable regression analysis will further analyze the relationship between different variables and the correlation between the dependent and independent variables.

3. Results

Besides the four identification variables mentioned above, there are still three string variables in the dataset, including gender, age, and marital status, which cannot be analyzed directly by SPSS. Therefore, by using the substitution function in Excel, the alphanumeric data of these string variables can be substituted to numeric data. Take the variable gender as an example. The variable male is substituted with the value of 1, the variable female is substituted with value of 2, and the variable unknown is substituted with the value of 0. Moving on, in order to better understand and analyze the dataset, it is necessary to add two new variables into the dataset: total_money_spent and total_money_saved, which define the total money spent and saved respectively in every order. By using the multiplication and subtraction function in Excel, the variable total_money_spent represents the product of variables final_unit_price and quantity, and the variable total_money_saved represents the difference between variables original_unit_price and final_unit_price times the variable quantity.

Table 1. Descriptive Statistics.

Variable	N	Minimum	Maximum	Mean	Std. Deviation	Variance
type	8832	1	2	1.97	0.181	0.033
attribute1	6115	1	4	3.00	0.786	0.618
attribute2	5616	30	100	81.38	21.223	450.411
quantity	8832	1	101	1.15	1.965	3.860
promise	1410	1	8	3.39	1.617	2.613
original_unit_price	8832	0	12158	141.393	321.288	103226.240
final_unit_price	8832	-32	9223.372	120.390	318.915	101707.027
direct_discount_per_unit	8832	0	720	15.387	36.155	1307.161
quantity_discount_per_unit	8832	0	225	3.546	12.833	164.684
bundle_discount_per_unit	8832	0	39	0.030	0.840	0.705
coupon_discount_per_unit	8832	0	174	2.040	10.630	112.998
gift item	8832	0	1	0.06	0.233	0.054
user level	8693	-1	10	2.17	1.104	1.219
plus	8693	0	1	0.13	0.333	0.111
gender	8693	0	2	1.43	0.786	0.618
age	8693	0	6	2.67	1.591	2.530
marital status	8693	0	2	1.11	0.801	0.642
education	8693	-1	4	1.55	1.844	3.401
city level	8693	-1	5	1.50	1.760	3.098
purchase power	8693	-1	5	1.37	1.597	2.550
total_money_spent	8832	-34	12995	133.175	391.631	153375.144
total_money_saved	8832	0	748.50	21.004	39.969	1597.520
Valid N (listwise)	882					

Above table 1 presents the mean, standard deviation, variance, and other basic descriptive statistics for 22 numeric variables in this dataset. Most of the variables have more than eight thousand

observations, except attribute1, attribute2, and promise. The variables related to price and discount have relatively higher standard deviation and variance since price and discount vary for different orders.

3.1. Reliability and Validity Tests.

In this research project, reliability refers to the internal consistency of measurements. Meanwhile, validity refers to the extent to which the measurements represent the variables that they are intended to. After reviewing the whole dataset and the relationship between different variables, the 11 variables in table 4 are the most relevant variables for the reliability and validity tests and the regression analysis in the following parts [5].

Table 2. Reliability Test Summary.

		N	%
Cases	Valid	8693	98.4
	Excluded	139	1.6
	Total	8832	100

Table 3. Reliability Test Statistics.

Cronbach's Alpha	N of Items
0.763	11

Table 4. Reliability Test Item-Total Statistics.

Variable	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
type	13.11	55.386	0.080	0.770
quantity	12.01	52.135	0.030	0.807
gift item	13.09	55.505	0.021	0.771
user level	10.98	49.534	0.314	0.757
plus	13.02	54.375	0.233	0.765
gender	11.72	47.740	0.670	0.729
age	10.48	39.661	0.671	0.705
marital _status	12.04	46.745	0.753	0.722
education	11.60	36.191	0.723	0.693
city level	11.65	41.189	0.502	0.735
purchase power	11.78	39.250	0.691	0.701

The above three tables present the results for the reliability test. In this dataset, 98.4 percent of the cases are valid for the test, and the Cronbach's Alpha for the 11 variables is 0.763, which indicates a relatively good internal consistency and reliability. Then, the slight effect on the Cronbach's Alpha level, if any of the variables are deleted, proves that there is no need to delete any variables [6].

Table 5. KMO and Bartlett's Test.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.841
Bartlett's Test of Sphericity	Approx. Chi-Square	34770.034
	df	55
	Sig.	0.000

Table 5 shows the essential results for the validity test. The KMO value of the test is 0.841, which is higher than 0.8, indicating that the scale in this dataset is suitable for factor analysis. Meanwhile, in Bartlett's Test of Sphericity, the approximate Chi-Square value is 34770.034, which is relatively large.

This result proves that the corresponding p-value is 0.000 and is less than 0.05. Therefore, Bartlett's Test of Sphericity has significant significance [7].

Table 6. Total Variance Explained.

Components	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.074	37.038	37.038	4.074	37.038	37.038	3.799	34.539	34.539
2	1.442	13.108	50.146	1.442	13.108	50.146	1.688	15.344	49.883
3	1.026	9.328	59.475	1.026	9.328	59.475	1.055	9.592	59.475
4	0.998	9.074	68.549						
5	0.951	8.641	77.190						
6	0.694	6.311	83.501						
7	0.558	5.070	88.571						
8	0.456	4.148	92.719						
9	0.333	3.026	95.745						
10	0.255	2.316	98.061						
11	0.213	2.939	100.000						

The above table for total variance explained points out that the system extracts three principal factors. The explanatory degree of the three principal factors to the whole dataset can reach 59.475 percent.

Table 7. Rotated Component Matrix.

Variable	Component		
	1	2	3
gender	0.857		
purchase power	0.823		
age	0.821		
marital status	0.796		
city level	0.729		
education	0.724		
user level		0.816	
plus		0.747	
gift item			0.830
type			0.539
quantity			

Table 7, the rotated component matrix, indicates that the 11 variables from this dataset can be divided into three principal components. The first six variables, including gender, purchase power, age, marital status, city level, and education, can be seen as the personal and social factors that affect the customer's product choice. Then, variables user level and can explain different product choices between different types of customers. Besides, variables gift item and type are two factors decided by the merchants, which can also influence the customer's product choice.

3.2 Multivariable Regression Model

The next step after carefully reviewing the results of the reliability and validity test is running the multivariable regression analysis. In this model, the dependent variable is total_money_spent, and the independent variables include user level, plus, marital status, education, and purchase power. The

equation for this model is: $\text{total\_money\_spent} = \beta_1 \times \text{user\_level} + \beta_2 \times \text{plus} + \beta_3 \times \text{marital\_status} + \beta_4 \times \text{education} + \beta_5 \times \text{purchase\_power}$

Meanwhile, based on this model, there are five null hypotheses and the alternative hypotheses.

a) $H_0: \beta_1 \leq 0$

$H_1: \beta_1 > 0$

b) $H_0: \beta_2 \leq 0$

$H_1: \beta_2 > 0$

c) $H_0: \beta_3 \leq 0$

$H_1: \beta_3 > 0$

d) $H: \beta_4 \leq 0$

$H_1: \beta_4 > 0$

e) $H: \beta_5 \leq 0$

$H_1: \beta_5 > 0$

Table 8-a Multivariable Regression Model Summary.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.133[a]	0.018	0.017	364.865

Table 8-b Multivariable Regression Model Coefficients.

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	44.614	9.797		4.554	0.000	25.410	63.819
	user level	44.533	4.344	0.134	10.251	0.000	36.017	53.049
	plus	38.460	12.917	0.035	2.977	0.003	13.140	63.780
	marital status	-15.782	7.878	-0.034	-2.003	0.045	-31.224	-0.339
	education	-11.337	3.471	-0.057	-3.266	0.001	-18.141	-4.532
	purchase power	14.927	3.989	0.065	3.742	0.000	7.109	22.746

As shown in the above two tables, the estimated model equation for this regression model is:

$\text{total\_money\_spent} = 44.533 \times \text{user\_level} + 38.46 \times \text{plus} - 15.782 \times \text{marital\_status} - 11.337 \times \text{education} + 14.927 \times \text{purchase\_power}$.

The result of the hypothesis test can be obtained by analyzing the beta, t-value, and p-value for each of those variables.

a) Rejecting the null hypothesis. The variable user level has a positive effect on total_money_spent.

b) Rejecting the null hypothesis. The variable plus has a positive effect on total_money_spent.

c) Accepting the null hypothesis. The variable marital status has a negative effect on total _money_spent.

d) Accepting the null hypothesis. The variable education has a negative effect on total_mo ney_spent.

e) Rejecting the null hypothesis. The variable purchase power has a positive effect on tot al_money_spent.

The coefficient for the variable user level is 44.533, which indicates a positive correlation between the user level and total money spending in each order. It means that in each order, for every one level increase in user level, there is a corresponding 44.533 dollars increase in total money spending. This is statistically significant because the p-value is 0.000, which is less than 0.05.

Holding other variables fixed, in every order, being a JD. Com plus member will increase the total money spent by 38.460 dollars. This is statistically significant since the p-value is 0.003, which is less than 0.05.

Holding other variables fixed, compared with a single person, being a married person will decrease the total online shopping spending by 15.782 dollars. This is statistically significant since the p-value is 0.045, which is less than 0.05 at the 5 percent significance level.

The coefficients for the variables education and purchase power are respectively -11.337 and 14.927, which indicates a negative correlation between education and total money spending and a positive correlation between purchase power and total money spending. In each order, for every one level increase in education level, there is a corresponding 11.337 dollars decrease in total money spending. Meanwhile, a one-level increase in purchase power corresponds with a 14.927 dollar increase in total money spending. These are statistically significant because the p-values for these two variables are respectively 0.001 and 0.000, which are both less than 0.05 [8].

3.3. The Model with Mediation Effect

From the above multivariable regression model, it is expected that there is a mediation effect between variables total money spent, plus, and user level. The mediation effect can be examined by running three regression analyses between these three variables. The first step is examining the significance of the coefficient in the equation: $spending = c \times plus + e_1$

Table 9-a Regression Model Summary.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.73[a]	0.005	0.005	367.078

Table 9-b Regression Model Coefficients.

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	121.441	4.215		28.815	0.000	113.180	129.703
	plus	80.481	11.810	0.073	6.814	0.000	57.330	103.632

Table 9-c Regression Model Anova Test.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	152.988	978	0.156	1.484	0.000
Within Groups	813.043	7714	0.105		
Total	966.030	8692			

Table 9-a, 9-b, and 9-c present the result of the regression model and the ANOVA test. In this equation, the coefficient for the variable *plus* is 80.481. The p-value for this coefficient is 0.000, which is less than 0.05. Thus, the coefficient in this equation is statistically significant. There was a statistically significant difference between groups as demonstrated by one-way ANOVA (F (978, 7714) = 1.484, p=0.000). Then, the next step is examining the significance of the coefficient in the equation: $user\_level = a \times plus + e_2$

Table 10-a Regression Model Summary.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.400[a]	0.160	0.160	1.012

Table 10-b Regression Model Coefficients.

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	2.005	0.012		172.524	0.000	1.982	2.027
	plus	1.324	0.033	0.400	40.669	0.000	1.260	1.388

Table 10-c. Regression Model Anova Test.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	200.903	6	33.484	380.120	0.000
Within Groups	765.128	8686	0.088		
Total	966.030	8692			

The above three tables show the result for the ANOVA test and the regression model between variables user level and plus. In this equation, the coefficient for the variable plus is 1.324. The p-value for this coefficient is 0.000, which is less than 0.05 and is statistically significant. There was a statistically significant difference between groups as demonstrated by one-way ANOVA (F (6, 8686) = 380.12, p=0.000). The last step is examining the significance of the coefficients in the multivariable equation: $spending = c' \times plus + b \times user\_level + e_3$

Table 11-a Multivariable Regression Model Summary.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.124[a]	0.015	0.015	365.230

Table 11-b: Multivariable Regression Model Coefficients.

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	48.149	8.821		5.459	0.000	30.859	65.440
	plus	32.066	12.820	0.029	2.501	0.012	6.935	57.197
	user_level	36.562	3.871	0.110	9.445	0.000	28.973	44.150

Table 11-c: Multivariable Regression Model Anova Test.

		Sum of Squares	df	Mean Square	F	Sig.
plus	Between Groups	152.988	978	0.156	1.484	0.000
	Within Groups	813.043	7714	0.105		
	Total	966.030	8692			
user_level	Between Groups	1983.455	978	2.028	1.817	0.000
	Within Groups	8611.641	7714	1.116		
	Total	10595.096	8692			

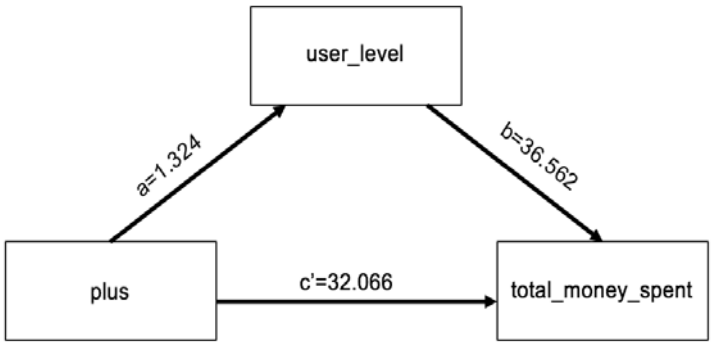

Fig 1. Multivariable Regression Model.

The multivariable regression model above indicates that the coefficients for the variables plus and user level are respectively 32.066 and 36.562. Meanwhile, the p-values for these two coefficients are 0.012 and 0.000, which are both less than 0.05 at the 5 percent significance level. Therefore, these two coefficients are statistically significant, and the mediation effect is tenable in this regression model [9]. For the variable plus, there was a statistically significant difference between groups as demonstrated

by one-way ANOVA (F (978, 7714) = 1.484, p=0.000). For the variable user level, there was a statistically significant difference between groups as demonstrated by one-way ANOVA (F (978, 7714) = 1.817, p=0.000). This regression model points out that compared with ordinary customers, JD. Com's plus members spend more on online shopping and are more likely to buy relatively expensive products. Meanwhile, holding other variables fixed, for every level increased in user level, there is a corresponding 36.562 dollars increase in spending for every order.

4. Conclusion

To sum up, the above analysis proves that a customer's online shopping spending and product choice are affected by user level, JD. Com plus membership, marital status, education, and purchase power. Education and marital status have negative relationships with a customer's online shopping spending, which also indicates a preference for relatively cheap products. By contrast, the user level, JD. Com's plus membership, and purchase power have positive relationships with a customer's online shopping spending. These positive relationships point out that a JD. Com's plus member with a higher user level and higher purchase power is more likely to buy relatively expensive products from this platform. Besides the factors discussed above, many other factors can also affect a customer's product choice, including discounts, price level, city level, and so on. The result of this research project advises JD.com to adopt targeted advertising to recommend the most suitable product to different customers.

References

[1] JD.com. (2020, March 02). JD.com Announces Fourth Quarter and Full Year 2019 Results. Retrieved January 10, 2021, from https://ir.jd.com/news-releases/news-release-details/jdcom-announces-fourth-quarter-and-full-year-2019-results#:~:text=For%20the%20full%20year%20of%202019%2C%20JD.com%20reported%20net,the%20full%20year%20of%202018.

[2] Cetină, Iuliana, et al. (2012). "Psychological and Social Factors That Influence Online Consumer Behavior." Procedia, Social and Behavioral Sciences, vol. 62, pp. 184–88, doi:10.1016/j.sbspro.2012.09.029.

[3] Javadi, M. H. M., Dolatabadi, H. R., Nourbakhsh, M., Poursaeedi, A., & Asadollahi, A. R. (2012). An analysis of factors affecting on online shopping behavior of consumers. International journal of marketing studies, 4(5), 81.

[4] Gong, W., Stump, R. L., & Maddox, L. M. (2013). Factors influencing consumers' online shopping in China. Journal of Asia Business Studies.

[5] Taherdoost, Hamed. (2016). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. International Journal of Academic Research in Management. 5. 28-36. 10.2139/ssrn.3205040.

[6] Leech, N. L., Barrett, K. C., & Morgan, G. A. (2014). IBM SPSS for intermediate statistics: Use and interpretation. Routledge.

[7] Dziuban, C. D., & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. Psychological bulletin, 81(6), 358.

[8] Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2013). Applied regression analysis and other multivariable methods. Cengage Learning.

[9] Cuevas, A., Febrero, M., & Fraiman, R. (2004). An anova test for functional data. Computational statistics & data analysis, 47(1), 111-122.