http://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Articles in Advance, pp. 1–9
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Data Driven Challenge

JD.com: Transaction-Level Data for the 2020 MSOM Data Driven Research Challenge

Max Shen,[a] Christopher S. Tang,[b] Di Wu,[c] Rong Yuan,[d] Wei Zhou[c]

[a] Department of Industrial Engineering and Operations Research, University of California, Berkeley, Berkeley, California 94720; [b] UCLA Anderson School of Management, University of California, Los Angeles, Los Angeles, California 90095; [c] JD.com American Technologies Corporation, Mountain View, California 94043; [d] Stitch Fix, San Francisco, California 94104

Contact: maxshen@berkeley.edu, https://orcid.org/0000-0003-4538-8312 (MS); chris.tang@anderson.ucla.edu (CST); di.wu@jd.com (DW); rongyuan.exe@gmail.com (RY); wei.zhou@jd.com (WZ)

Received: September 10, 2019
Revised: November 5, 2019
Accepted: December 20, 2019
Published Online in Articles in Advance: December 9, 2020

https://doi.org/10.1287/msom.2020.0900

Copyright: © 2020 INFORMS

Abstract. To support the 2020 MSOM Data Driven Research Challenge, JD.com, China's largest retailer, offers transaction-level data to MSOM members for conducting data-driven research. This article describes the transactional data associated with over 2.5 million customers (457,298 made purchases) and 31,868 stock keeping units (SKUs) over the month of March in 2018. We also present potential research questions suggested by JD.com. Researchers are welcome to develop econometric models or data-driven models using this database to address some of the suggested questions or examine their own research questions.

History: Gad Allon served as guest editor-in-chief for this article.

Keywords: e-commerce • transactional data • MSOM society • data-driven research

1. Introduction

The growth of e-commerce retailing (or E-tailing) has given rise to many new and challenging problems at both strategic and operational levels. To encourage operations management (OM) researchers to conduct data-driven research in E-tailing, we are collaborating with JD.com and the MSOM society to run a research competition based on their proprietary data. This competition is intended to enable researchers to examine research questions arising from customer purchasing decisions and supply chain operations in the context of E-tailing.

JD.com is China's largest retailer with a net revenue of US$67.2 billion in 2018 and over 320 million annual active customers. According to JD.com,

> [It] is committed to providing only high-quality, authentic products, and is known for its fast delivery speed. JD.com sets the standard for online shopping through its commitment to quality, authenticity, and its vast product offering covering everything from fresh food and apparel to electronics and cosmetics. JD.com combines its first-party business model, where it controls the entire supply chain, with a marketplace that intentionally limits the number of sellers, to ensure that it can maintain strict quality oversight. JD.com has a nationwide fulfillment network that covers 99% of China's population, and is able to provide standard same- and next-day delivery for approximately 90% of orders.

The data sets provided by JD.com capture a "full customer experience cycle" that begins as soon as a customer begins browsing on the platform and ends when the customer receives the delivered products. The data set describes 2.5 million customers (457,298 made purchases) and 30,000 stock keeping units (SKUs; from one product category) during the month of March in 2018.

Based on our discussion with the management of JD.com, we developed the following set of research questions. We encourage researchers to explore the provided data and develop innovative solutions to address the following problems (or other research problems of their own choosing):

1. Which product attributes and/or features have predictive power about the *customer's product choice*? Does *this product choice* differ by channel (e.g., purchasing via mobile phones vs. personal computers), region, and brand loyalty?
2. Would more products with similar attributes and features improve or hinder sales revenues for JD.com?
3. For a specific target customer segment (e.g., female customers in tier 1 cities), what should merchants and brands do to improve their sales performance?
4. What is the impact of various pricing and promotion strategies on product sales? How should JD.com improve its pricing and promotion strategy?

Table 1. Description of the *skus* Table

Field	Data type	Description	Sample value
sku_ID	string	Unique identifier of a product	b4822497a5
Type	int	1P or 3P SKU	1
brand_ID	string	Brand unique identification code	c840ce7809
attribute1	int	First key attribute of the category	3
attribute2	int	Second key attribute of the category	60
activate_date	string	The date at which the SKU is first introduced	2018-03-01
deactivate_date	string	The date at which the SKU is terminated	2018-03-01

In particular, among all the promotion methods (e.g., direct discounts, bundle discounts, and volume discounts), which one is more effective?

5. Do ordinary customers behave differently from JD.com's PLUS members?[1] How should JD.com improve its pricing and shipping strategy for its PLUS members?

6. How should JD.com improve its *demand forecast* accuracy for different geographic regions and different customer groups?

7. How should JD.com improve its fulfillment efficiency and customer experience with better *inventory allocation* strategies in a multilevel inventory network?

2. Data Description

We now describe the transaction-level data provided by JD.com. (We shall explain how to download the database in Section 4.) To ensure confidentiality, certain key identification information such as user ID and SKU ID are anonymized.[2]

To keep the size of the database more manageable, the database does not contain impression data, especially when JD.com may not have complete impression data from other channels (search, push notification, SMS messages, social media (e.g., WeChat), mobile ads, etc.). However, our data contain all product detail page click events for each customer, which can serve as a proxy. Instead, JD.com provides us with transaction-level data for the month of March 2018 during which there were no major holidays or promotions.[3] Hence, the March data can be viewed as a baseline, and researchers should be careful about extrapolating their results.

In the database, each SKU can be identified either as *first-party owned* (1P) or *third-party owned* (3P), depending on the ownership of the inventory of that SKU.[4] All 1P SKUs are managed by JD.com, including product assortments, inventory replenishments, product pricing, order deliveries, and after-sale customer services. Despite different operations, 1P and 3P SKUs compete on the JD.com platform for sales through different pricing strategies and marketing activities.

In general, 1P SKUs are usually top sellers within the category. By owning these 1P products, JD.com can fully control the entire customer experience to provide guaranteed quality, fast delivery, and good customer services. In contrast, all 3P SKUs are managed by third-party merchants on the JD marketplace. Specifically, to fulfill an order of a 3P SKU, the corresponding merchant can decide freely whether to use the logistics services provided by JD Logistics or other logistics service providers.[5]

The data sets provided by JD.com offer a detailed view of the activities associated with all SKUs within one anonymized consumable category during the month of March in 2018. This category could be beauty care (e.g., face moisturizers) or men's grooming (e.g., electric shavers) or something else. Owing to confidentiality, the specific category is not disclosed. The data set consists of seven tables that are labeled as (1) *skus*, (2) *users*, (3) *clicks*, (4) *orders*, (5) *delivery*,

Figure 1. (Color online) Distribution of Attribute 1 Across All SKUs

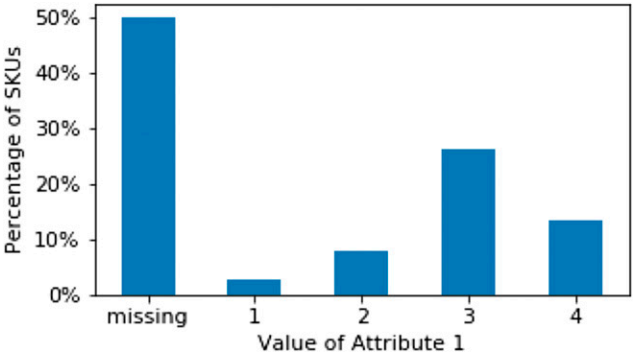


Figure 2. (Color online) Distribution of Attribute 2 Across All SKUs

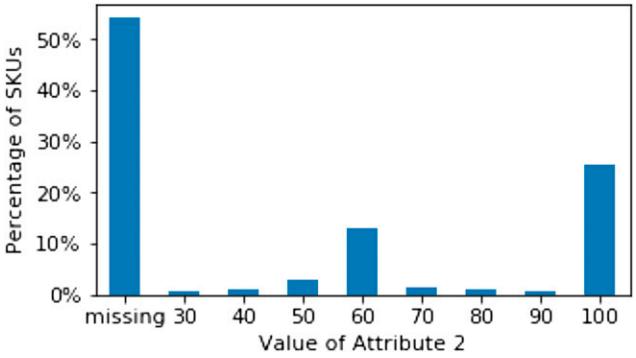


Table 2. Description of the *users* Table

Field	Data type	Description	Sample value
user_ID	string	User unique identification code	000000f736
user_level	int	User level	10
first_order_month	string	First month in which the customer placed an order on JD.com (format: yyyy-mm)	2017-07
plus	int	If user has a PLUS membership	0
gender	string	User gender (estimated)	F
age	string	User age range (estimated)	26–35
marital_status	string	User marital status (estimated)	M
education	int	User education level (estimated)	3
purchase_power	int	User purchase power (estimated)	2
city_level	int	City level of user address	1

(6) *inventory*, and (7) *network*. We now describe each of these seven tables.

2.1. Table 1: SKUs

The *skus* table (Table 1) describes the characteristics of all 31,868 SKUs that belong to a single product category receiving at least one click during March 2018.[6] As such, researchers should not generalize their results to other product categories. We now define each field and provide a brief description. Each entry in the *skus* table corresponds to a unique SKU (*sku_ID*). In addition, each SKU ID is "seller-specific." For example, an identical product that is sold by JD as a 1P product and by a third-party seller as a 3P product will be treated as *two separate SKUs with different SKU IDs*. Similarly, an identical product sold by multiple third-party sellers will be denoted by different SKU IDs.

Of these 31,868 SKUs, 1,167 of them are 1P SKUs (*type* value = 1) and the rest (30,701) are 3P SKUs (*type* value = 2). The brand information of each SKU is provided via the field (*brand_ID*). However, only 9,159 SKUs out of 32,343 were involved in purchase activities during March of 2018.

Each SKU also has two key attributes: the first attribute takes integer values between 1 and 4, and the second takes integer values between 30 and 100. For each attribute, a *higher value* indicates *better performance* of a certain functionality. For the face moisturizer category, these two attributes can be sun protection factor (SPF) and percentage of antiaging ingredients. Similarly, for the men's electric shaver category, these two attributes can be the number of shaves per charge and the number of personalized shaving modes. Hence, both attributes characterize the functionality of a product so that products with the same attribute values have the same functionality. The distributions of the value associated with these two attributes across all SKUs are depicted in Figures 1 and 2. Notice that many SKUs have missing values for various reasons, including that (a) the third-party merchants did not provide the attribute value, especially for certain slow-moving items, or (b) a certain attribute was not applicable to certain SKUs.[7]

For each SKU, the *skus* table provides two extra elements: *activate_date* and *deactivate_date*. The former specifies the date at which a SKU is first introduced on the JD.com platform and the latter specifies the date at which the SKU is terminated and removed from JD.com.[8] Note that the data set lists the *activate_date* and *deactivate_date* variables only when these dates fall in the month of March in 2018; thus, these variables are usually blank.

2.2. Table 2: Users

The *users* table (Table 2) describes the characteristics of all 457,298 users who purchased at least one of the SKUs in the given category during March of 2018.

Figure 3. (Color online) Distribution of Users: User Level

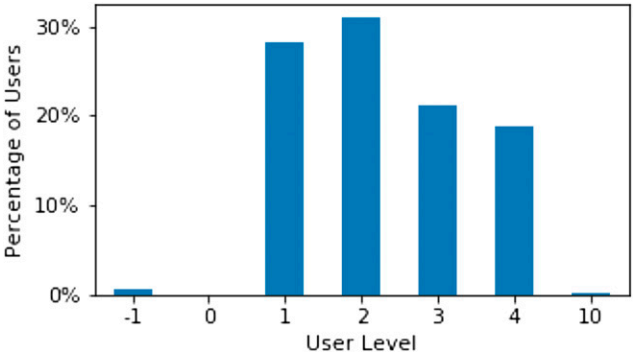


Figure 4. (Color online) Distribution of Users: Gender

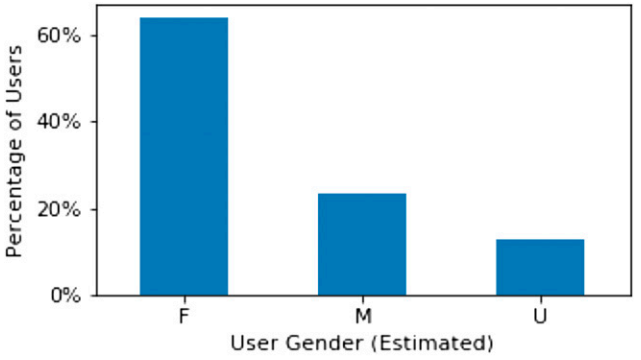
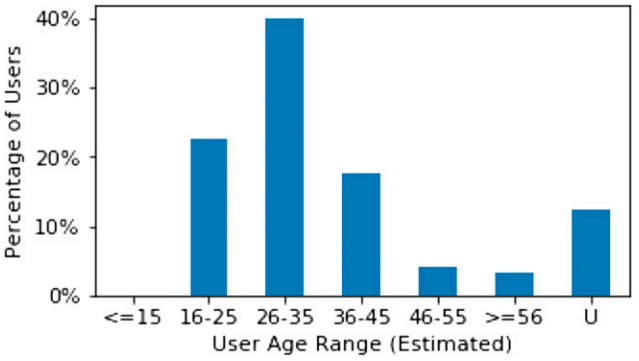


Figure 5. (Color online) Distribution of Users: Age



We now define each field and provide a brief description. Each entry in the *users* table corresponds to a unique customer (*user_ID*). The field *first_order_month* specifies the month when the user made his or her first purchase on JD.com.

For each repeat customer, the corresponding user is classified according to his or her past purchases so that the customer's *user_level* takes on 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. For users who are enterprise users (e.g., small shops in rural areas or small businesses), the corresponding *user_level* takes on a value of 10. However, for first-time purchasers, their *user_level* takes on the value −1.[9] Figure 3 depicts the distribution of user levels for all 457,298 customers.

The next variable is *plus*. This variable equals 1 when the corresponding user is an existing PLUS member on February 28, 2018.[10] (The variable *plus* is based on a snapshot on February 28, and we do not have information about the PLUS membership on a daily basis.) In addition to customer past purchase value and PLUS membership, the *users* table contains certain (*estimated*) user demographic information, because JD.com's customers are not required to provide any demographic information when making a purchase. However, JD.com has a sophisticated data-driven artificial intelligence system to estimate user demographics.

Figure 6. (Color online) Distribution of Users: Marital Status

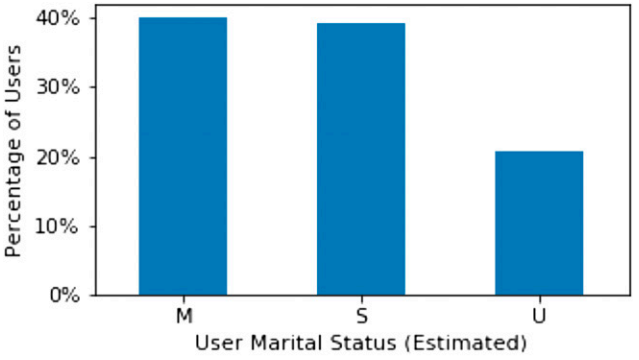
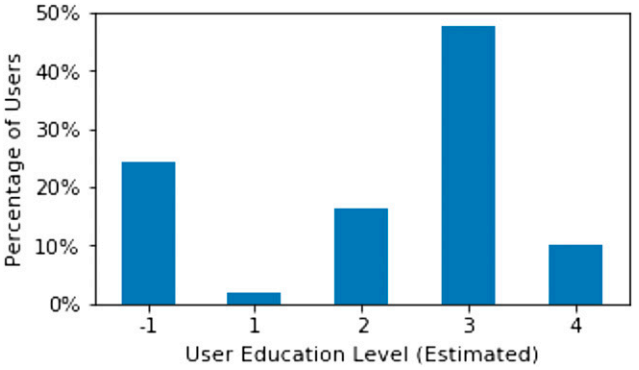


Figure 7. (Color online) Distribution of Users: Education Levels



The estimated user demographics for each user are (a) *gender* (F: female, M: male, U: unknown); (b) *age* (< =15: less than or equal to 15 years old, 16–25: 16 to 25 years old, 26–35: 26 to 35 years old, 36–45: 36 to 45 years old, 46–55: 46 to 55 years old, ≥ 56: greater than or equal to 56 years old, U: unknown); (c) *marriage*—user's marital status (M: married, S: single, U: unknown); (d) *education*—user's education level (1: less than high school, 2: high school diploma or equivalent, 3: bachelor's degree, 4: postgraduate degree, −1: unknown); and (e) *purchase_power*—user's estimated purchase power (ranging from 1 to 5 with 1 being the highest purchase power; −1 if there is no estimation).

In addition to those estimated demographics of each user, JD.com has provided actual information about the most commonly used shipping address for each user. This information is captured in the field *city_level*, which takes on values ranging between 1 and 5. JD.com developed its own classification scheme for different cities: level 1 corresponds to highly industrialized cities such as Beijing and Shanghai; level 2 cities correspond to provincial capitals; level 3–5 cities are smaller cities; if there are no data, then the value is −1. Notice that *city_level* is based on actual information.

Figure 8. (Color online) Distribution of Users: Purchase Power Levels

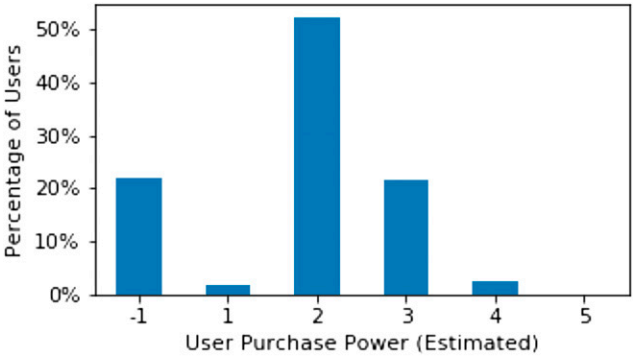


Figure 9. (Color online) Distribution of Users: City Levels

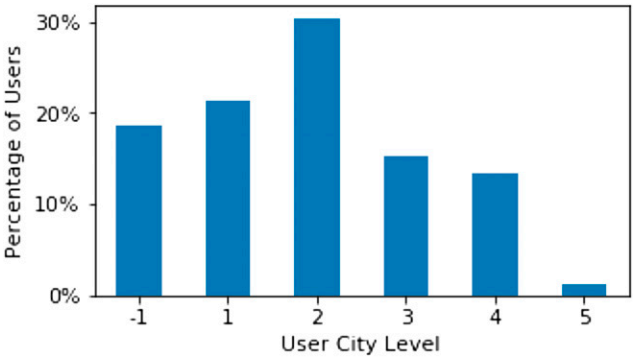


Figure 4 depicts the distribution of user gender across all 457,298 customers in the database, and Figure 5 summarizes the distribution of estimated user age. As shown in Figure 6, for this specific product category, more than 60% of all customers are estimated to be female and the estimated ages of these customers are in their 30s to 40s. From Figure 6, we observe a relatively even distribution between married and single customers. Figures 7 and 8 provide the customer's estimated education level and purchase power. Figure 9 summarizes the distribution of shipping address according to different city levels. It can be seen that most of the customers are from tier 1 and tier 2 cities.

2.3. Table 3: Clicks

The *clicks* table (Table 3) establishes the linkage between users and SKUs through their browsing history. Each entry in the *clicks* table represents a user's "click event" on a specific SKU page.[11] The date set contains over 20 million click records that are associated with the clicks of 2.5 million customers. Note that this table contains clicks contributed not only by the users identified in the *users* table (Table 2) who purchased at least one SKU but also by "other users" who did not end up completing a purchase order.

The records include the following: (a) the user who initiated a *click event* (*user_ID*), (b) the SKU associated with the click event (*sku_ID*), (c) the time at which the click event occurred (*request_time*), and (d) the channel in which the click event occurred (*channel*).[12] We classify the channel taken as five string values: *pc*, *mobile*, *app*, *wechat*, and *others*. Channels *pc* and *mobile* are associated with clicks through web browsers on personal computers and mobile devices, respectively. Channel *app* corresponds to JD.com's mobile app. Channel *wechat* corresponds to the miniprogram that runs on the social media app WeChat. Finally, channel *others* aggregates the clicks from all other channels.

The distribution of all click events across all channels is summarized in Figure 10. Because of the popularity of smartphones in China and the popularity of mobile payment options (e.g., WeChat payment), the majority of click events come from the *app* and *wechat* channels.

The field *request_time* provides extra granularity. It can be used to infer the customer browsing sequence and habits. In Figure 11, we plot the number of clicks during the day on March 1, 2018, within the *app* channel. We can clearly identify two peaks in the daily browsing activities: one from 8 a.m. to 4 p.m. in the day and the other in the late evening.

2.4. Table 4: Orders

The *orders* table (Table 4) contains 486,928 unique customer orders associated with our focused product category that were placed during the month of March in 2018. Each customer order (*order_ID*) in the *orders* table is based on a specific SKU (*sku_ID*) associated with a unique customer (*user_id*). (If a customer ordered multiple SKUs, then the same *order_ID* will appear in multiple rows of SKUs.)

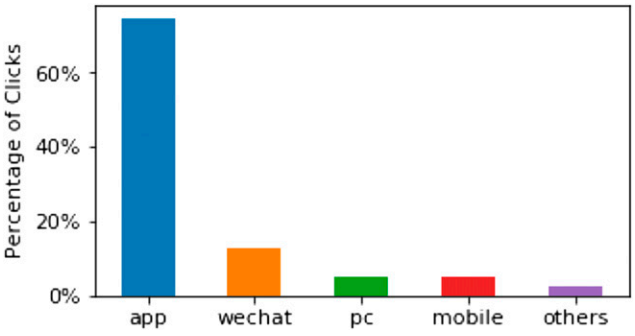
Other pieces of information associated with a customer order as shown in Table 4 include (a) order quantity for each SKU associated with the order (*quantity*), (b) the date and time when the ordering event took place (*order_date* and *order_time*), (c) the type of SKU being ordered (*type* = 1 if it is a 1P SKU and *type* = 2 if it is a 3P SKU), and (d) the promised delivery time of the order (*promise*).[13] Figure 12 demonstrates that most orders have promised delivery dates within two days. Figure 13 shows the total number of sales by date and by order type.

The *orders* table also offers information about product pricing and promotional activities for each SKU. For each entry, we denote the original list price of the SKU in the field *original_unit_price* and the actual paid price by the customer for the SKU as *final_unit_price*. The original list price of a SKU at any given time instant

Table 3. Description of the *clicks* Table

Field	Data type	Description	Sample value
sku_ID	string	SKU unique identification code	b4822497a5
user_ID	string	User unique identification code	94ff800585
request_time	string	The time at which the customer clicks the SKU item page (format: yyyy-mm-dd HH:MM:SS)	2018-03-01 23:57:53
channel	string	The click channel	wechat

Figure 10. (Color online) Distribution of All Click Events Across Different Channels



is the same for all customers, but the final price can vary among customers owing to various discounts or promotions.

The "gap" between the original price and the final price represents the coupons and discounts associated with different promotional activities for each SKU. There are four common types of promotional discounts on the JD.com platform:[14]

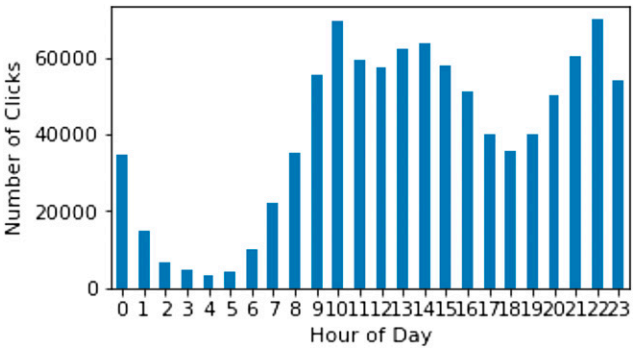
1. SKU direct discount: The seller of a SKU may offer a price cut in terms of a *direct discount*. This discount reflects the reduction in the list price as stated on the product detail page.

2. Group promotion: The seller of a SKU may offer a *quantity discount* to entice the customer to buy more. This quantity discount promotion can take different forms including "get an RMB 100 discount if buying over RMB 199" or "buy 3 and get 1 free." We note that the quantity discount promotion is usually on the order level and we apply a simple allocation rule to calculate the contribution provided by each SKU in the order.

3. Bundle promotion: The seller may offer a *bundle_discount* if a customer buys a "prespecified bundle" of SKUs within an order.

4. Gift items: The seller may offer a SKU as a "free gift" (*gift_item* value = 1) if the customer purchases a "prespecified set" of SKUs (e.g., get a free eraser if you buy x pencils and y pads of paper). The *final_unit_price* for each gift item is always equal to 0.

Figure 11. (Color online) Number of Click Events Occurring on March 1, 2018, Through JD.com's App Channel



Coupons can also be applied to the order after all other promotions are applied. In contrast to the four aforementioned promotion activities where discounts will be applied automatically once certain criteria are met, customers must "clip" (or claim) a coupon before making a purchase.[15] The field *coupon_discount* records the coupon promotional value associated with an order. Similar to *quantity discount* as explained earlier, the discount value of the coupon is allocated between items in the same order using an allocation rule when necessary.

Note that, for each entry in the *orders* table, the gap between *original_unit_price* and *final_unit_price* should always equal the sum of *direct_discount*, *group_discount*, *bundle_discount*, and *coupon_discount*.

Finally, for each order, we show from which district the order was shipped (*dc_ori*) and to which district the order was shipped (*dc_des*). The district here is defined by the warehouse ID that covers the demand of that district. In other words, one can think of *dc_ori* as the warehouse where the package is shipped from and *dc_des* as the warehouse that is nearest to the customer's designated shipping address. If *dc_ori* and *dc_des* are the same, this means that the package is shipped from the warehouse closest to the customer. Otherwise, it indicates that the package is fulfilled by some other warehouse in a different district. We note that in theory any warehouse in the nationwide network can fulfill the order for any customer in the country. However, in practice, there is a complicated order fulfillment logic that determines what inventory should be used to fulfill each customer order to optimize fulfillment resources while satisfying delivery promise.

One can trace the shipping path and time of each order by using Tables 4 and 5.[16] First, for each order denoted as *order_ID*, the order table (Table 4) provides information about the "origin" warehouse that the order is shipped from (via the variable *dc_ori*) and the "destination" warehouse that the order is shipped to (via the variable *dc_des*). By using the information provided in Tables 4 and 5, one can trace the shipping path of each order.

2.5. Table 5: Delivery

The *delivery* table (Table 5) establishes the linkage between each order (*order_ID*) and (possibly) multiple shipping packages (i.e., multiple *package_IDs*) in the event that an order is split into multiple delivery packages for logistical reasons (e.g., an order that involves in-stock and on-order items). The *delivery* table contains records for orders delivered with JD Logistics, which represents the majority of 1P orders and some 3P orders. The orders that cannot find a

Table 4. Description of the *orders* Table

Field	Data type	Description	Sample value
order_ID	string	Order unique identification code	3b76bfcd3b
user_ID	string	User unique identification code	3cde601074
sku_ID	string	SKU unique identification code	443fd601f0
order_date	string	Order date (format: yyyy-mm-dd)	2018-03-01
order_time	string	Specific time at which the order gets placed (format: yyyy-mm-dd HH:MM:SS)	2018-03-01 11:10:40.0
Quantity	int	Number of units ordered	1
Type	int	1P or 3P orders	1
Promise	int	Expected delivery time (in days)	2
original_unit_price	float	Original list price	99.9
final_unit_price	float	Final purchase price	53.9
direct_discount_per_unit	float	Discount due to SKU direct discount	5.0
quantity_discount_per_unit	float	Discount due to purchase quantity	41.0
bundle_discount_per_unit	float	Discount due to "bundle promotion"	0.0
coupon_discount_per_unit	float	Discount due to customer coupon	0.0
gift_item	int	If the SKU is with gift promotion	0
dc_ori	int	Distribution center ID where the order is shipped from	29
dc_des	int	Destination address where the order is shipped to (represented by the closest distribution center ID)	29

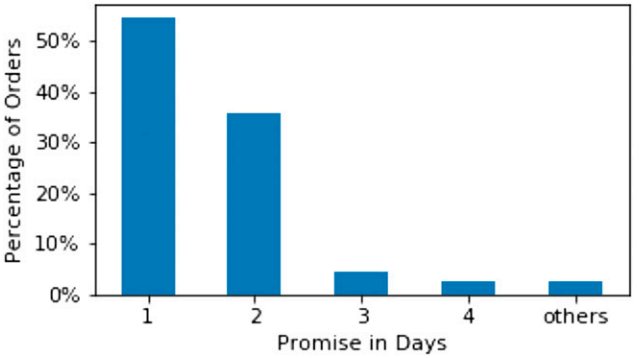
match record in the *delivery* table can be considered delivered by an alternative shipping method.

The *delivery* table contains 293,229 packages delivered by JD Logistics in the given time period, among which 244,333 orders involve 1P SKUs (*type* = 1) and 48,896 orders involve 3P SKUs (*type* = 0).[17] We further provide three key timestamps (up to hourly granularity) for each package delivery, namely, the time at which the package was shipped from the warehouse (*ship_out_time*), the time at which the package arrived at the delivery station[18] (*arr_station_time*), and the time at which the package was successfully delivered to the customer (*arr_time*).

2.6. Table 6: Inventory

The *inventory* table (Table 6) provides information about the availability of each SKU (*sku_id*) at each warehouse (*dc_ID*). We only disclose the availability of the inventory *at the end* of the day (*date*) instead of the amount of inventory. In addition, when a SKU is not available at a specific warehouse on a specific day, there will be no record of that SKU at that warehouse on that day.

Figure 12. (Color online) Distribution of Promise Delivery Time (1P Orders)



2.7. Table 7: Network

The *network* table (Table 7) provides information about the assignment of different warehouses located in different districts (*dc_ID*) to different geographical regions (*region_ID*). For each district, a designated warehouse (*dc_ID*) is responsible for fulfilling orders in the district. In addition, for different districts that are assigned to a geographical region, one of the (larger) warehouses will be designated as the "central warehouse" for that region. In JD.com's context, a central warehouse provides the "back-up fulfillment" option when other (typically smaller) warehouses in the region run out of inventory for their corresponding districts. Figure 14 shows the number of districts within each geographical region. We denote each central warehouse for each region by setting *dc_ID* = *region_ID*.

Figure 13. (Color online) Sales in Quantity by Date and Order Type

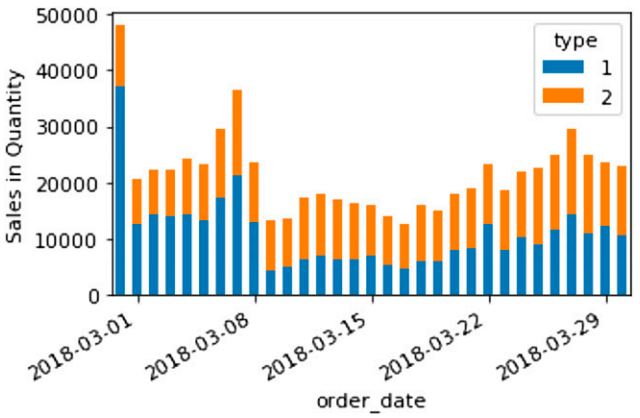


Table 5. Description of the *delivery* Table

Field	Data type	Description	Sample value
package_ID	string	Package unique identification code (same as order_ID if the package contains all SKUs in the order)	209a005c40
order_ID	string	Order unique identification code	209a005c40
Type	int	1P or 3P orders	1
ship_out_time	string	The timestamp when the package is shipped out from the warehouse (format: yyyy-mm-dd HH:MM:SS)	2018-03-01 08:37:33
arr_station_time	string	The timestamp when the package arrives at the delivery station (format: yyyy-mm-dd HH:MM:SS)	2018-03-01 15:37:31
arr_time	string	The timestamp when the package is delivered to the customer home (format: yyyy-mm-dd HH:MM:SS)	2018-03-01 18:49:03

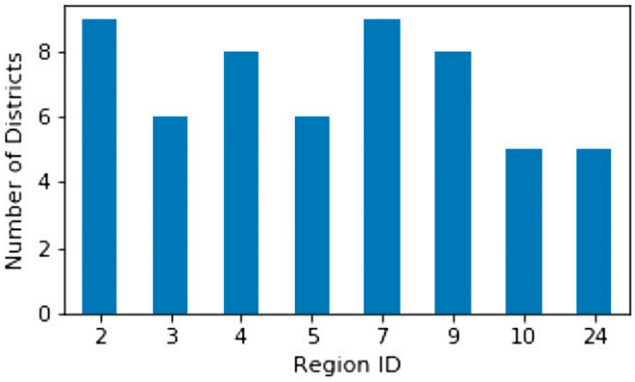
Table 6. Description of the *inventory* Table

Field	Data type	Description	Sample value
dc_ID	int	Distribution center ID	9
sku_ID	string	SKU unique identification code	fcc883f713
Date	string	Date (format: yyyy-mm-dd)	2018-03-01

Table 7. Description of the *network* Table

Field	Data type	Description	Sample value
region_ID	int	Region ID	2
dc_ID	int	District ID (same as warehouse ID)	6

Figure 14. (Color online) Number of Districts Within the Regions



3. Conclusion

The data sets provided by JD.com are presented in the seven tables described above. These data sets are based on the activities associated with 2.5 million users (457,298 made purchases) and 31,868 SKUs in March of 2018. Researchers are invited to analyze this database with data-driven models to address research questions posed by themselves or by JD.com as stated in Section 1.

The data sets include product information (attributes, pricing, etc.), customer information (demographics, total value of past purchases, PLUS membership, etc.), and logistics information (shipping networks, orders and inventories, delivery time, etc.). The data sets capture a "full customer experience cycle," which begins the moment a customer chooses the products on the platform and ends the moment the customer receives the products.

4. Downloading the Data and Python Code

MSOM members can access the data sets thorough the MSOM website (https://connect.informs.org/msom/events/datadriven2020). For easy access, we provide a Python notebook[19] with runnable sample code to facilitate reviewing and understanding of the data sets as well as to explain the relationships among the seven tables described in this paper. The code is provided in the online appendix, and a runnable version is available within the data set package.

Acknowledgments

The authors thank the guest editor, Gad Allon, the associate editor; and two anonymous reviewers for their valuable suggestions.

Endnotes

[1] JD.com's PLUS membership is a subscription-based program that provides its members certain benefits that range from free shipping to

member-specific price discounts. For details about JD.com's PLUS membership, see https://plus.jd.com/index.html (Chinese content).

[2] Note that the data provided by JD.com represent only a small sample of users and SKUs. Therefore, the database does not necessarily fully capture the business performance or business trends of JD.com.

[3] However, it is possible that some brands may launch "super brand day" promotions within March.

[4] All SKUs are displayed on JD.com's product page with the seller name and/or tags so that customers are fully aware of whether the corresponding SKU is a 1P SKU or a 3P SKU.

[5] The fulfillment process is usually described on the product page so that customers will know that the shipping process is managed by the merchant itself.

[6] There may be some SKUs that receive no click during March, but the information for these SKUs is not available.

[7] JD.com displays product ratings for each SKU. However, in the Chinese marketplace, most product ratings reported by customers are usually the highest rating. Because most ratings are rated 5, the information associated with product ratings has been shown to be uninformative. Consequently, product ratings are omitted in the database.

[8] Note that, even though a SKU is deactivated, it may still be able to be bought as a part of a bundled product or as the gift portion of a promotion.

[9] Regardless of different users' *user_level* values, they observe the same information and receive the same service from JD.com.

[10] JD PLUS membership costs up to US$45 per year and members enjoy a variety of perquisites including exclusive discounts, higher purchasing reward rate, free delivery, and return with no preconditions. About 18% of those 458,269 customers in the data set are JD PLUS members.

[11] It is worth noting that this table only contains click information on the SKU detail page. There are many other page types with which a customer can interact on JD.com, such as the website main page, category main page, various landing pages, search, recommendation page, and shopping cart page. Although those pages also contain information about SKUs and promotions, the customers still need to go to the SKU detail page to review the detailed description of the products and place the order.

[12] Note that these data capture the click event of a SKU initiated by a user, but each click event may not lead to the purchase of this SKU. In other words, a user may choose not to purchase this SKU even after the click event.

[13] When *promise* = 1, this refers to the standard same- and next-day delivery promise: Orders placed before 11 a.m. will be delivered on the same day, and orders placed before 11 p.m. will be delivered before 3 p.m. on the following day. When *promise* is x ($x > 1$), this indicates that the delivery will arrive at day $t + x$, where t is the day the order is placed. We note that *promise* information is not available for a small fraction of 1P orders and for most of the 3P orders.

[14] For third-party products (i.e., 3P SKUs), the discounts are controlled by the sellers. However, for JD-owned 1P SKUs, the discounts result from discussions between different vendors and JD.

[15] Coupons normally consist of a discount value, an eligibility criterion, and an expiration date. The discount value is the monetary amount that can be deducted from the order; the eligibility criterion specifies which SKU or SKU set is eligible for coupon use and whether there is a total purchase amount criterion. The expiration date shows when the coupon can be applied. There are many ways in which customers can receive a coupon. They can clip coupons from the product detail pages, promotional landing pages, or "coupon mall" (a specific section on the JD.com platform for coupon distribution). Customers can also receive personalized coupons based on their past activities.

[16] The data set, however, does not provide a detailed shipping path if the order is routed through multiple warehouses.

[17] The delivery table contains shipment information about orders involving 1P products owned by JD (and some 3P products owned by the third party). However, for other orders involving 3P products owned by the third party, the shipment information is not available to JD.com because the logistics providers are selected by the merchants.

[18] A delivery station corresponds to the "last mile" facility before a package is delivered to the customer. Hence, the *arr_station_time* specifies the time when the package arrives at the delivery station before it is delivered to the customers. The timestamps of internal transfers of a package routing through different warehouses are omitted.

[19] See https://jupyter.org/.