Rayguru Madan Mohan, *College of business*

Preparation and analysis of JD.com data, a detailed report and notebook files using python packages.

Modeling and analysis - making predictions and finding useful insights.

**By**

Group 18

Amith Joseph

Veena Ramachandani

Gowatami Rao

Sushima Panala

Sai Likhitha

Project work presented as requirement to complete the course **BCIS5110 Programming for Business Analytics**

Examining the analysis of JD.com data using python packages to find meaningful business insights and detailed analysis report.

In terms of

Using the JD sales data, which contains six different tables about the products and customers and inventory levels, we try to find the predictions for the given data in all terms.

**By**

Group 18

*Advised by*

Rayguru Madan Mohan

Rayguru Madan Mohan, College of Business

University of Louisville

Louisville, Kentucky

November 2024

**Acknowledgement**

To

Rayguru Madan Mohan

We want to thank you for your mentorship and guidance throughout this semester and the project. Our conversations and class discussions about data analysis in general made the experience wonderful for us.

# **Executive Summary**

JD.com, China's top online retailer, commands a commanding position in the market, with a staggering net revenue of US$67.2 billion in 2018 and a large customer base of over 320 million yearly active users. JD.com has set the standard for online shopping, positioning itself as a seller of premium, legitimate products and taking great pride in its unrivaled delivery speeds. JD.com, which is well known for its unwavering commitment to quality and authenticity across a wide range of product categories such as fresh foods, apparel, electronics, cosmetics, and more, distinguishes itself through a unique business model. JD.com protects product excellence by combining a first-party strategy that oversees the whole supply chain with a curated marketplace that prioritizes quality over quantity. Vendors, making use of a national fulfillment network that covers 99% of China's population.

The examination focuses on many key areas: To identify potential changes in order volumes, it first investigates consumer behavior by examining average item purchases across various order types (1P or 3P). Second, it investigates the spatial distribution of purchases across city tiers to assess whether users are equitably represented across city tiers. Third, to investigate order processing delays, it compares the average time for packages to leave the warehouse for 1P and 3P orders. It also studies the top-performing SKU\_IDs in sales to find the factors that lead to their long-term success. Furthermore, it searches for specific SKU-ID and Click channel combinations that improve sales to perhaps advise channel-specific activities. Finally, there is the study predicts future sales using user interaction timing and channels for specific SKUs, providing insights into potential sales trajectories and market trends.

# **Project motivation and background**

This data-driven approach is driven by the goal of improving customer happiness and sales operations in an e-commerce environment. By utilizing extensive datasets that include consumer information, sales transactions, and delivery logistics, the objective is to improve customer satisfaction and optimize the shopping process. The background is that to make wise judgments at every point of the sales funnel, enormous volumes of data must be gathered and analyzed.

1. How many items, on average, did customers buy across all order types (1P or 3P) (quantity field)? Do the typical order volumes for these two kinds of orders differ significantly? (orders table)
2. How were customer purchases distributed throughout cities at different levels (city\_level)? Are there more users from higher city levels, or are there people from all city levels represented fairly? (users)
3. What is the average duration (ship\_out\_time) for packages to leave the warehouse once an order is placed? Is there a big difference in processing times between 1P and 3P orders? (delivery table)
4. What SKU\_IDs have consistently outperformed the market regarding sales, and if so, what qualities or causes underlie their success? (SKUs)
5. What is the specific SKU-ID and Click channel combination that provides higher sales? If yes, then it can be used for channel-specific combination (clicks)
6. What are the future sales for specific SKUs based on the timing and channel of user interactions?

The data does have some difficulties with quality. The dataset has a significant number of missing values. The User\_Id has only a few missing entries, whereas the dc\_id has the most missing data. Because of some inconsistent values in Sku\_Id, that must be alphanumeric but includes some records with only numeric values. A limited number of records having numerical data can be erased to make the data consistent. Because it isn't utilized to anticipate the results, it's possible to disregard the variable dc\_id, which contains most of the missing information. The data collection process begins by gathering sales transaction records, capturing details like SKU, date, and quantity sold. Once obtained, the data is filtered to isolate transactions by specific days and SKUs. Aggregation follows, summing the total number of SKU units sold for each applicable day, consolidating the data by SKU. Findings are then presented, evaluating the overall sales for a particular SKU on a specific day, often through reports or visual representations. Concurrently, customer data, including PLUS membership status, might reside in a database, aiding in buyer identification. Delivery priority rules are established, potentially granting preference to PLUS members based on their membership level, order size, or location. Orders are evaluated for adherence to regulations, giving precedence to PLUS members in the fulfillment process. Ultimately, order prioritization occurs based on various factors, such as membership status and order status, ensuring efficient and rule-compliant order processing.

# **Data description**

Each SKU in the database can be classified as "first party owned" (1P) or "third-party owned" (3P), depending on who owns the inventory for that SKU.3

JD.com manages all 1P SKUs, including product assortments, inventory replenishments, product pricing, order deliveries, and after-sales customer service. Despite their varied operations, 1P and 3P SKUs compete for sales on the JD.com platform through different pricing tactics and marketing activities.

In general, 1P SKUs are top sellers in the category. JD.com can fully control the entire customer experience by owning these 1P products, ensuring guaranteed quality, fast delivery, and good customer service. Third-party merchants handle all 3P SKUs on the JD marketplace. In particular, the associated merchant is allowed to choose between using JD Logistics' logistics services and those of other logistics service providers to complete an order for a 3P SKU.4 The JD.com data sets give a thorough understanding of the operations related to every SKU in a single anonymized consumable category in March of 2018.

SKU Table and description

**Variables**

sku\_ID: A string value that acts as a special identification for every product in the dataset is stored in this field. An example of a sample value for a specific product is "b4822497a5".

type: An integer field designating the SKU's 3P (third-party) or 1P (first-party) category affiliation. '1', for example, might stand for 1P SKUs, and '3' for 3P SKUs.

brand\_ID: A unique identification code for the brand connected to every SKU is stored in this string field. For example, the unique brand identifier could be "c840ce7809".

attribute1 and attribute2: The SKU's category is indicated by these two integer elements, which represent important attributes. In a category, for example, "3" and "60" could stand for different attributes.

activate date: Using a string format ('YYYY-MM-DD'), this field provides the date the SKU was originally introduced or activated. For instance, the date of activation of a specific SKU is indicated by "2018-03-01".

deactivate date: This field contains the date that the SKU was discontinued or rendered inactive. The deactivation date of an SKU, for example, could be indicated by "2018-03-01".

With the help of this dataset, which offers detailed information on individual SKUs, such as identifiers, types (1P or 3P), related brand IDs, category attributes, activation, and deactivation dates, and more, it is possible to analyze and monitor the evolution of product lifecycles, brand affiliations, and category-specific attributes over time.

User table and description

**Variables**

user\_ID: String values in this field serve as each user's unique identifier. In the dataset, for example, "000000f736" designates a particular user.

user\_level is an integer field that represents the user's tier or level on the platform. '10', for instance, can indicate a particular degree of user interaction or access.

The first order month that the user placed on JD.com is represented by the string field first\_order\_month. 'yyyy-mm' is the format. For example, the month "2017-07" denotes the user's initial order month.

plus: A numeric field that indicates whether the user is a member of PLUS; values other than '0' may indicate that the user is not a member.

gender: A string field with the user's estimated gender indicated by letters, like 'F' for female.

age: A text field with the user's estimated age range entered. The age range, for example, could be represented by "26–35".

matri\_status: A text field that provides an estimate of the user's marital status. M, for instance, may stand for "married."

education: An integer field that may employ numerical codes to represent varying degrees of education to estimate the user's education level.

purchase\_power: An integer field that uses probabilities to approximate the user's spending power or purchasing power; various levels of purchasing power are probably represented by numerical values.

city\_level: An integer field that encodes the user's address's city level; it may employ numeric values to classify various city tiers.

Observations: A distinct user on JD.com is associated with each observation in this dataset. The observations encompass a range of attributes that describe the demographics, buying patterns, membership status, and other approximations about the characteristics of the users. These insights offer valuable information about the diversity and behavior of the platform's user base.

Table clicks and description.

**Variables**

sku\_ID: This field has string values that are used to uniquely identify SKUs (Stock Keeping Units), which are representations of certain products. An example of this would be "b4822497a5". SKU\_ID. user\_ID: String values in this field operate as distinct IDs for each user that interacts with an SKU. In the dataset, for example, "94ff800585" designates a particular user.

request\_time: A text field that stores the exact moment the customer clicked on the SKU item page. It follows the format 'yyyy-mm-dd HH:MM: SS'.

Channel: The user's click channel for interacting with the SKU is indicated by the string values in this field. "Wechat" could, for example, stand for a particular channel that the user used to access the SKU.

Observations: Every observation in this dataset is an instance of a communication that took place on JD.com between a particular user and an SKU. These observations provide details about the SKU that users engaged with, when they clicked on SKU item pages, who the user was interacting with, and the channel that the interaction took place through. This information offers perceptions.

Table orders and description

**Variables**

order\_ID: This field has string values that function as distinct order IDs. In the dataset, for example, "3b76bfcd3b" denotes a particular order.

user\_ID: This field has string values that function as special order-placing user identifiers. In the dataset, "3cde601074", for instance, designates a particular user.

sku\_ID: String values that serve as special IDs for SKUs (Stock Keeping Units)—specific products inside an order—are contained in this field. For example, a sample SKU\_ID linked to an order is "443fd601f0".

order\_date: A string field that holds the order placement date.

order\_time: A string field that contains the exact moment the order was placed.

quantity is an integer field that indicates how many units of the SKU were ordered in a specific transaction. '1', for instance, denotes the ordering of a single item.

type: An integer field designating a 1P (first-party) or 3P (third-party) order, respectively. One could imagine that '1' would stand for a 1P order.

assurance: An integer field that indicates how many days the order should take to arrive. A delivery within two days, for example, would be indicated by the number '2'.

original unit cost: a float field with the original list price of the SKU's single unit. '99.9', for instance, might

Observations: Every observation in this dataset corresponds to a distinct order placed by a JD.com user. The observations offer a wealth of information about the order, including the user, the SKUs purchased, the quantity, pricing data, any applied discounts, the order timing, and the anticipated delivery date. This information provides a thorough understanding of the purchasing behavior of users as well as transactional details.

Delivery table and description

**Variables**

package\_ID: This field has string values that function as distinct package identifiers. This ID matches the order\_ID if every SKU in the order is included in the package. For instance, the package "209a005c40" in the dataset denotes a particular package.

order\_ID: This field has string values that function as distinct order IDs. It symbolizes the package's associated order and is connected to it. If the package includes every SKU in the order, for example, "209a005c40" might also be used as the order ID.

type: An integer field informing us of the package's affiliation with a first party (1P) or third-party (3P) order. One could imagine that '1' would stand for a 1P order.

Observations: Every observation in this dataset relates to the current state of a particular package's delivery that is connected to an order. These observations include details regarding the package's journey, such as when it left the warehouse, when it reached a station, and when it was finally delivered to the client. By offering insights into the delivery schedule and logistics, this data makes it possible to analyze metrics related to order fulfillment and shipment effectiveness.

Table inventory and description

**Variables**

The Distribution Center ID is represented by the integer parameter dc\_ID. In the dataset, for example, '9' can represent a particular distribution center.

sku\_ID: String values that serve as distinctive IDs for SKUs (Stock Keeping Units), or individual products inside the inventory, are contained in this field. As an illustration, the sample SKU\_ID for the inventory is "fcc883f713".

date: A string field that holds the date the inventory status is updated. We use the format 'yyyy-mm-dd'.

Observations: Every observation in this dataset relates to a certain SKU's inventory state at a given distribution facility on a given date. These observations provide details regarding the stock levels of various SKUs at various distribution centers on various dates. By monitoring SKU availability and distribution center inventory over time, this data offers insights into inventory management, stock levels, and may even help with logistics planning and supply chain optimization.

Table Network and description

**Variables**

region\_ID: A field with an integer value that holds the Region ID. One possible use of '2' is to designate a particular area in the network dataset.

dc\_ID: An integer field denoting the District ID, which, in this case, is the same as the warehouse ID. In the network dataset, '6', for instance, can represent a particular neighborhood or warehouse.

Observations: All the observations in this dataset concern the connections between the network's districts and warehouses and its regions. The mapping or association between various regions and the matching districts or warehouses is disclosed by these observations. The geographic distribution of warehouses across different regions may be shown by this data, which could help with logistics planning and the comprehension of the network structure inside the business's operations.