**BCIS 5110 Project Guidelines – Spring 2023**

**Due on March 5th 11:59pm, 2023**

**Objective**

You will apply your knowledge and skills of Python programming and business analytics to organize and analyze real-life data for meaningful insights.

Note: The following descriptions are excerpts from the document

Shen et al. (2019) “JD.com: Transactional data for the 2020 MSOM Data Driven Research Challenge”. You may find it here: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3511861>

**Context**

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 the official description provided by JD.com,

“*JD.com 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* ***business model of first party****, 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 covers 99% of China’s population and is able to provide standard same- and next-day delivery as standard for approximately 90% of orders*.”

**Data**

The data sets provided by JD.com capture a “full customer experience cycle” that begins at the

moment when a customer browses through products available on the platform before placing their order and ends at the moment when the customer receives the products at their designated location. The data sets provide information on **2.5 million customers** over **30,000 SKUs** in **one specific product category** during **the month of March in 2018**.

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.

We focus on three tables of this large dataset.

1. **Table JD\_sku\_data.csv**

The sku table describes the characteristics of each of the 31,868 SKUs that belong to a single product category. Each entry in the skus table corresponds to a unique SKU (sku\_ID). In addition, each SKU ID is “sellerspecific.” 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.

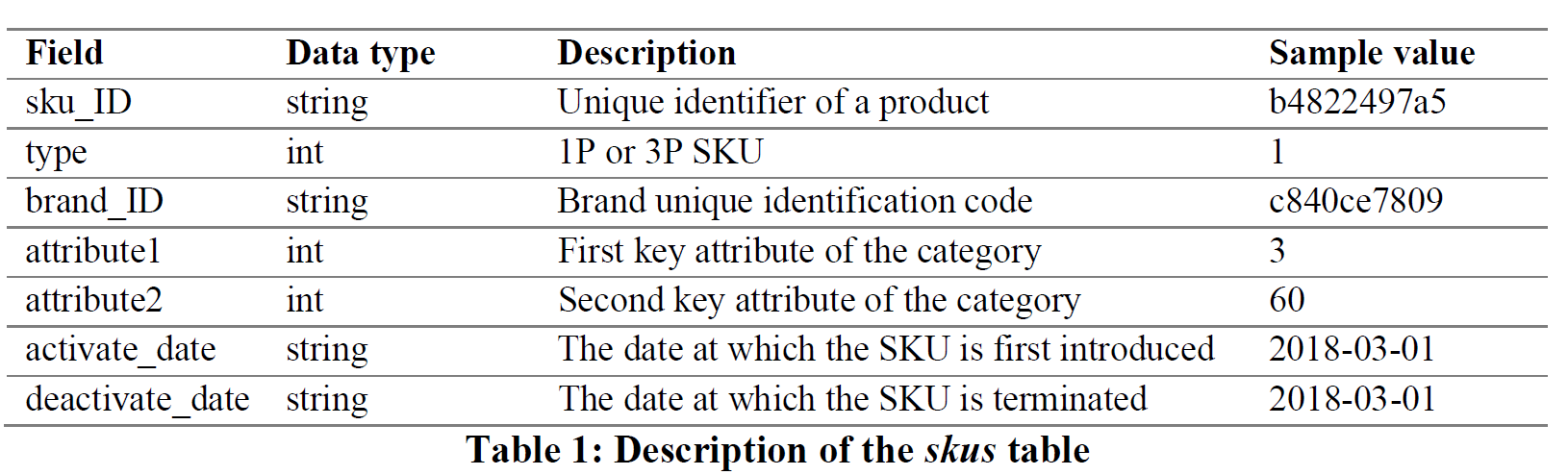
Each SKU also has two key attributes: the first attribute takes on a value that ranges from 1 to 4,

and the second attribute takes on a value that ranges from 30 to 100. For each attribute, a higher

value indicates better performance of a certain functionality (e.g., longer battery life and higher

screen resolution).

For each SKU, the skus table provides two extra elements: activate\_date and deactivate\_date. The former specifies the date at which an 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.6 Notice that the data set only lists a valid activate\_date and deactivate\_date when one of these dates occurred within the month of March in 2018. If one of these fields is empty, this means that the SKU was activated before March 2018 and/or deactivated after March 2018.



1. **Table JD\_user\_data.csv**

The users table describes the characteristics of each of the 457,298 users who purchased at least one of the SKUs in the given category during March of 2018. 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 their “first purchase” on JD.com.

For each repeat customer, the corresponding user is classified according to their 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). Regardless of different users’ user\_level values, they observe the same information and receive the same service from JD.com.

If the field value of plus is 1, this denotes that the corresponding user is an existing PLUS member before March of 2018.8 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.

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);

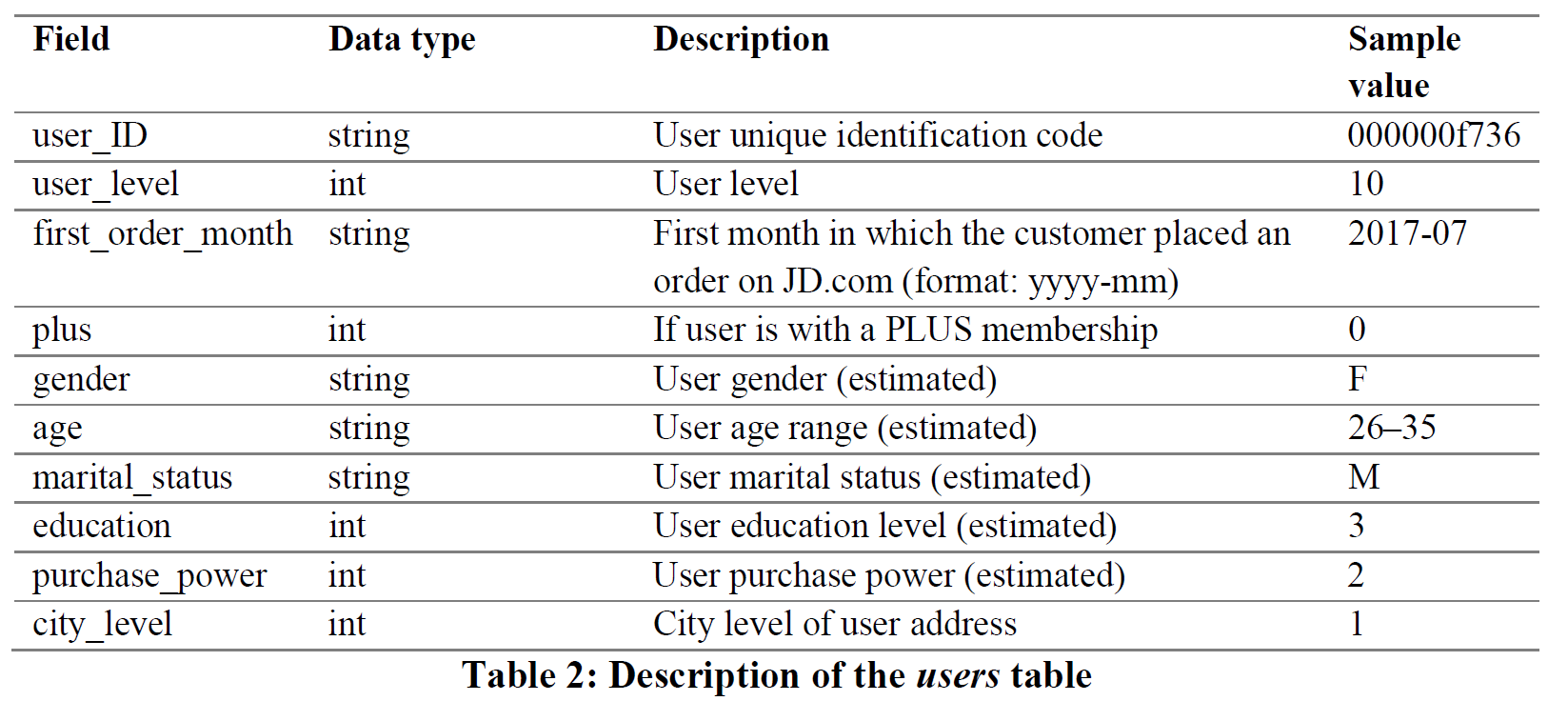
(c) marriage – user’s marital status (M: Married, S: Single, U: Unknown);

(d) education – user’s education level (1: less than high school, 2: high school diploma or equivalent, 3: Bachelor’s degree, 4: post-graduate degree, −1: unknown);

(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. Here, level 1 corresponds to highly industrialized cities such as Beijing and Shanghai; level 2 cities correspond to provincial capitals; level 3 to 5 cities are smaller cities; if there are no data then the value is −1.



1. **Table JD\_order\_data.csv**

The orders table 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 information associated with a customer order 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)

(d) the promised delivery time of the order (promise)

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 an SKU at any given time instant 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:

**(1) SKU direct discount:** The seller of an 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 an 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.

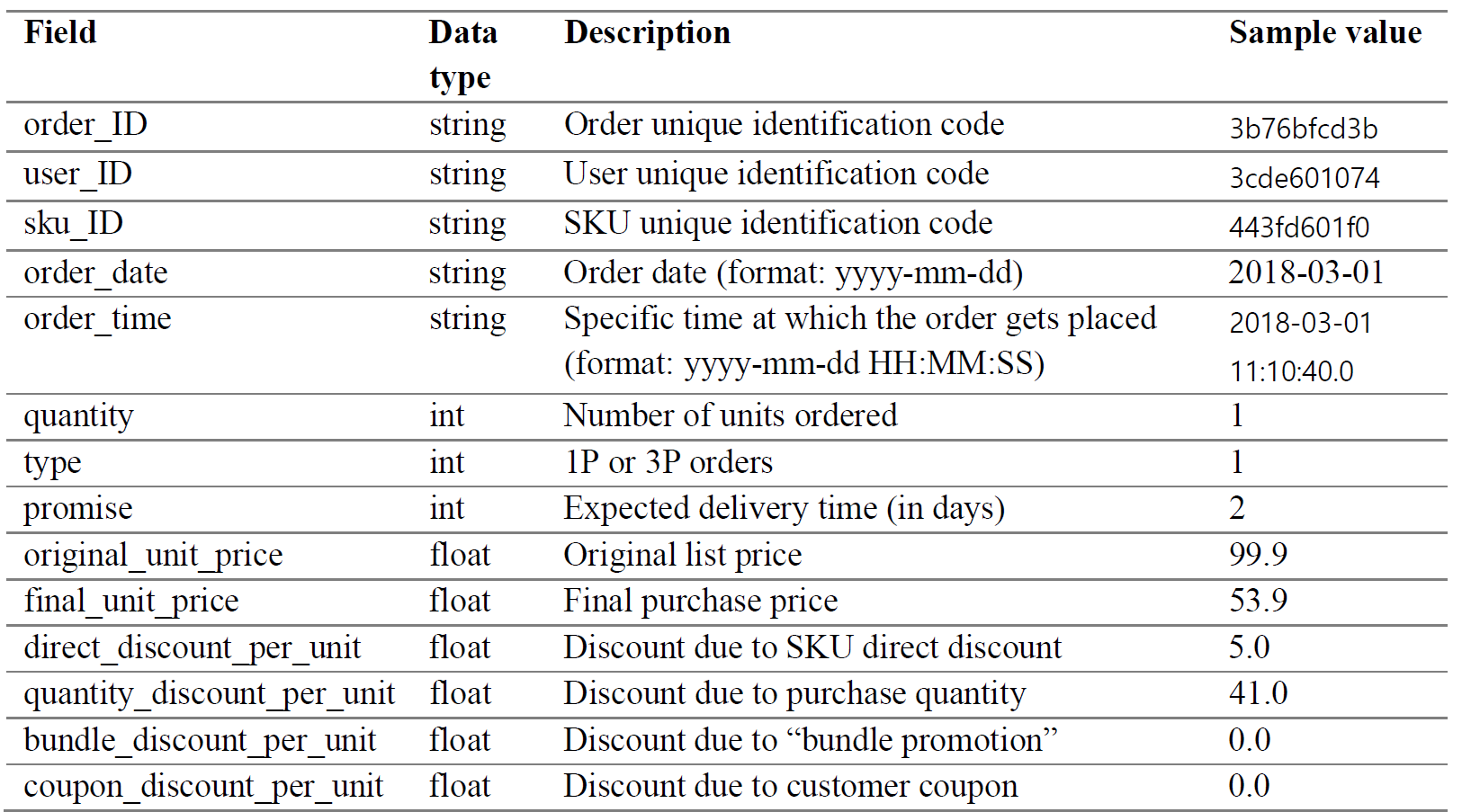
**(3) bundle promotion:** The seller may offer a bundle\_discount if a customer buys a “pre-specified bundle” of SKUs within an order.

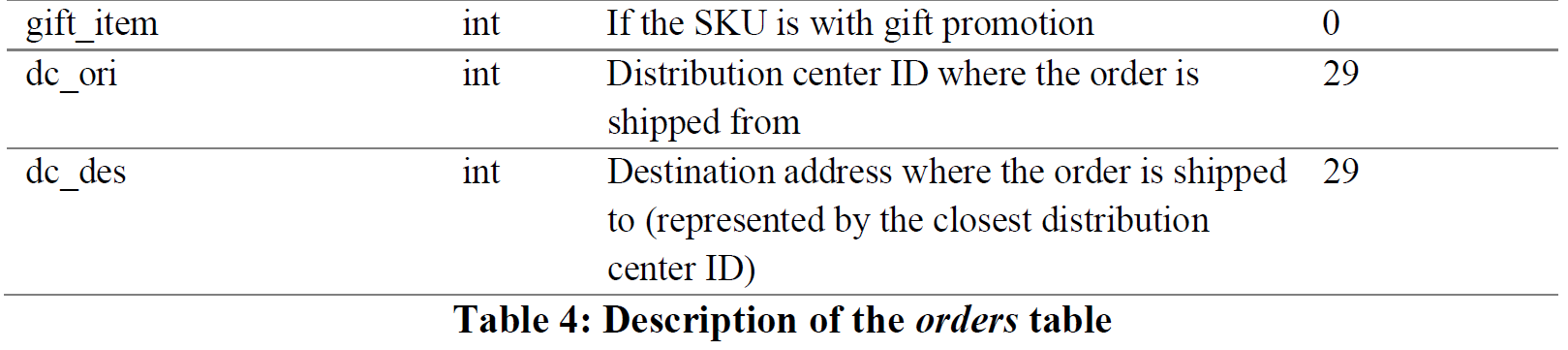
**(4) gift items**: The seller may offer an SKU as a “free gift” (gift\_item value = 1) if the customer purchases a “pre-specified set” of SKUs (e.g., get a free eraser if you buy x pencils and y pads of paper). The final\_unit\_price for each gift item is always equal to 0.

Coupons can also be applied to the order after all other promotions are applied. In contrast to the previous four promotion activities where discounts will be applied automatically once certain criteria are met, customers need to “clip” (or claim) a coupon before making a purchase. The field coupon\_discount records the coupon promotional value associated with an order.

We 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 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.



****

**Initial Report** (Due at Week 7)

Visit JD.com to have a brief understanding of the business. Examine the data files and understand the data structure and the meaning of every variable. You will make an initial report to:

1. **Identify research questions (20 points)**

Based on your understanding of the data, you may already find interesting questions to explore. There are two types of questions in general: descriptive vs. predictive. Descriptive questions describe what the data is. Examples include: what is the average number of downloads for an App in Google Play Store? Who is the best salesman in the Northeast region? How does the price change over the years? Is housing price correlated with zip code? You can usually answer them with summary statistics or graphs. They are part of data exploration. You can have plenty of descriptive questions to understand your data. You may present the most interesting ones in your report or presentation.

 Predictive questions intend to predict variable outcomes based on data. Examples include: what is the prediction for next month’s sale? Is the customer going to default on their loan? What might be the price for this house? What is the probability of the patient getting readmitted? You will need to build predictive models to answer these questions.

**You will generate at least one predictive question (Think about your target variable) and five descriptive questions based on the JD.com data.**You do not need to consider how these questions can be answered at this stage.

1. **Explore the data (20 points)**

Examine the data. Answer the following questions in words (not code).

* Are there quality issues in the dataset (missing data, extreme values, inconsistency, etc.)? Give one example variable with potential quality issues.
* What would you propose to deal with the quality issue for the example variable you give in the question above?
* How to calculate the sales for a specific SKU for a specific day? Describe the steps. You do not need to write the code.
* Can we merge the three tables together? How?
* After merging, we may ask questions based on the merged table, such as “Which type of customer (based on gender, education, etc.) purchase the most of product X”. Please come up with two such questions.