

Case Study 1: Online Grocery Retailing

Ling-Chieh Kung*

Note. For this case study, you will be given *real* data provided by a U.S. retailer. This data set is organized by the instructor specifically for this course. Therefore, they may be noisy and not “perfect” for a class practice. Nevertheless, they give us a chance to taste how data analysis may really help business decision makers in practice.

1 Introduction

Instacart is a U.S. company that operates as a same-day grocery delivery service. Customers select groceries through a web application from various retailers and delivered by a personal shopper. As of 2017, Instacart only has operations and services in the United States.

In this case study, we will play with a data set publicly posted by Instacart.¹ In this data set, Instacart publishes its online grocery shopping records with over three million grocery orders from more than 200,000 Instacart users. By investigating (a subset of) the data set, we may understand how people place orders, how popular a product is, their association, and many more. Please apply the visualization and summarization techniques we introduced in class to give a description to the data. With your findings, you may want to make some suggestions to Instacart. If so, convince the management that your suggestions make sense by showing them evidences obtained from the data.

*Department of Information Management, National Taiwan University; lckung@ntu.edu.tw.

¹<https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2>.

Some stories behind the data are here. You may want to read the article before you start your analysis. Do not forget that you will make some suggestions to the company. Before you do that, please study the company, its business, and its industry by yourselves.

2 Data

Part of these data are given in the MS Excel file “SDA106-1_case1_data.xlsx.”

Departments, aisles, and products. The spreadsheet “Department” contains two columns for department IDs and department names. The spreadsheet “Aisle” contains two columns for aisle IDs and aisle names. The spreadsheet “Product” contains four columns. The first two indicate the ID and name of a product, and the last two indicate the aisle and department that this product belongs to. One department contains several aisles, and one aisle contains several products. From the spreadsheet “Product,” we may construct a tree-type structure among departments, aisles, and products.

Orders. The spreadsheet “Order” contains the orders that were placed by 1,000 randomly selected users. For each order, we record the following information:

- *User ID*: The ID of the user who placed the order.
- *Order Number*: The order sequence number for this user (1 means the first).
- *Order ID*: A unique ID of this order.
- *Day of Week*: The day of the week the order was placed on (0 means Sunday, 1 means Monday, ..., and 6 means Saturday).
- *Hour of Day*: The hour of the day the order was placed on (0 means between 0:00 and 1:00, means between 1:00 and 2:00, etc.).
- *Days since Previous Order*: Days since the last order, capped at 30.

Order_Product. The spreadsheet “Order_Product” records the products that were purchased in the orders made by the 1,000 users. For each product, we record the following information:

- *Order ID*: The ID of the order that includes this product.
- *Product ID*: The ID of the product that is purchased.
- *Product Number*: The order in which this product was added into the cart.
- *Reordered*: 1 if the product has been ordered by this user in the past and 0 otherwise.

3 Tasks

The managers of Instacart are waiting for your report and/or presentation, which should address the following:

1. Do descriptive statistics to provide an overview for the online retailing business at that time. You may do it from any perspective with any method. You may include some basic summaries as well as some emphases on interesting findings.
2. Write down at most two decision problems (in marketing, operations, etc.) related to the online retailing business whose decision may be facilitated by data analysis. Then analyze the given data to address these problems. You may also leverage external data that you may obtain. Finally, if you believe some other internal data may be helpful, make proposals to the managers regarding how to use them.

4 Submission rules

Each team should have at most six students unless a special approval is obtained from the instructor. Each team must submit a written report to address the above tasks. You do not need to include the details of your solution process, but you may want to summarize your procedure. Focus more on the presenting your results.

The report, including everything (cover page, appendix, etc.), cannot be longer than **12 pages** (i.e., six double-sided pieces of paper). A report is considered good if it addresses the tasks correctly, precisely, and concisely. Including managerial implications found by analyzing the data is always a plus. Obviously, a well-formatted report is expected.

A hard copy of the written report must be submitted in class by **6:45 pm, November 1**. Electronic copies of the report (for all teams) and slides (for the presenting teams) should be uploaded to CEIBA by the same due time. For each team, only one member should upload the files. Submissions between 6:45 pm and 7:45 pm on the due date will get 20% off as a penalty. Submissions later than 7:45 pm are not accepted.

The report (for all teams) counts for 10% of the semester grades. The oral presentation (for presenting teams) counts for another 10%.