Descriptive Analytics: Preliminary Customer Analysis*

MSIN0094 Case Study

Dr Wei Miao

2001-10-20

1 The Power of Descriptive Analytics

The amount of data created worldwide has been increasing exponentially over the past decade with some estimates placing the total at 59 zettabytes as of 2020 (Statista 2020). Data without analytics, however, is of little value to business decision-makers aiming to improve performance and increase growth. It is therefore no surprise that top-tier consulting companies, analytics firms and business schools have been promoting the positive returns to greater usage of analytics technology. It also explains why an increasing number of data analytics enthusiasts are willing to pay up to £40k tuition fee (that is 10,000 bubble teas!) to join the prestigious MSc Business Analytics program at the UCL School of Management (hmm, it's now week 3, too late to ask for a refund!).

By identifying patterns and trends in massive amounts of data, business analytics enables organizations to make better decisions and improve performance. Descriptive analytics is the simplest and most widely used type of analytics; it is used to generate key performance indicators (KPIs) and metrics for business reports and dashboards. The latest research shows that, even with the adoption of very simple descriptive analytics, businesses can improve their performance by a large extent — Berman and Israeli (2022)[1] use the synthetic difference-in-differences method to analyze the staggered adoption of a retail analytics dashboard by more than 1,500 e-commerce websites,[2] and find an increase of 4%–10% in average weekly revenues postadoption. The increase in revenue is not explained by price changes or advertising optimization. Instead, it is consistent with the addition of customer relationship management, personalization, and prospecting technologies to retailer websites. The adoption and usage of descriptive analytics also increases the diversity of products sold, the number of transactions, the numbers of website visitors and unique customers, and the revenue from repeat customers. These findings are consistent with a complementary effect of descriptive analytics that serve as a monitoring device that helps retailers control additional martech tools and amplify their value. Without using the descriptive dashboard, retailers are unable to reap the benefits associated with these technologies.

*This case was prepared by Dr. Wei Miao, UCL School of Management, University College London for MSIN0094 Marketing Analytics module. This case was developed to provide material for class discussion rather than to illustrate either effective or ineffective handling of a business situation. Names and data may have been disguised to assure confidentiality. Please do not circulate without permission.

[1] Berman, Ron, and Ayelet Israeli. 2022. "The Value of Descriptive Analytics: Evidence from Online Retailers." Marketing Science, March, mksc.2022.1352. https://doi.org/10.1287/mksc.2022.1352.

[2] We will cover the difference-in-differences technique to establish causal inference later in the module.

In practice, businesses use descriptive analytics to assess how well they are performing and if they are on pace to meet business objectives. Business leaders and financial specialists monitor common financial measures generated by descriptive analytics, such as revenue and spending growth on a regular basis. Marketing teams utilize descriptive analytics to analyze the efficacy of marketing campaigns by tracking data such as conversion rates and social media followers, and manage customer relationship by keeping track of customer lifetime values. Manufacturing organizations track indicators such as line throughput and downtime. Descriptive analytics enables everyone in the organization to make more informed decisions that move the business forward. It reveals trends that would otherwise remain buried in raw data, allowing marketing managers to quickly assess how well the firm is operating and identify areas for improvement. Additionally, descriptive analytics enables firms to convey information within departments and to external parties.

In the remaining of this case, we will explore (1) how to consolidate multiple databases from various sources using R and (2) how to conduct preliminary customer analysis using descriptive analytics.

2 Database Marketing at Tesco

We have learned in Week 2, how to compute the customer lifetime value for i-basket, an online grocery store. However, when computing the CLV, we used an "average" approach, which did not consider customer heterogeneity. That is, when considering each component in the CLV formula, such as customer spending in each period, their retention rate, etc., we took the average across all customers, and assumed customers are homogeneous. As a result, every customer would have the same CLV. Nevertheless, this is a strong assumption in practice — every customer is unique and should be treated differently.

The key to successful customer relationship management is to maintain a customer database that tracks detailed customer information, including their demograhpic information and past purchase history. This information would empower marketing analytics team to compute individual CLV for each customer, and conduct individualized targeted marketing.

2.1 Demographic Information

Knowing your consumer is a vital concept of running any business. Is the business selling fertilizer to farmers, apparel to teenage girls, or vacations to senior citizens? The distinctions are readily apparent in this comparison.

Demographics define the qualities of clients. To be successful, business owners must understand the demographics of their clients and the trends or changes that are occurring within those specific traits.

The following demographic information is usually of interest to business managers:

Age: Consumer behavior is strongly influenced by age. Younger consumers are more affluent and willing to spend more on entertainment, fashion, and movies. Seniors spend less on these items; they are less active, spend more time indoors, and require more medical treatment. Additionally, market segments can be defined by age groups. For instance, digital devices such as iPhones are targeted more towards millennials than at seniors.

While older adults are increasingly utilizing technology, they remain less digitally savvy than millennials and purchase fewer digital products.

Gender: Gender also matters. Males and females have vastly diverse demands and tastes, which influence their purchasing decisions. As a result, some products are created with a specific gender in mind. Macy's, Nordstrom, and The Gap all have departments dedicated to teenage girls' clothes, while Seiko has a specific line of diver watches for men only.

Income: Income has a substantial influence on consumer behavior and product purchases. Middle-income customers make purchases with due regard for the utility of money. They do not have unlimited money to spend, and hence the money spent on one item may be used on something else. On the other hand, consumers with higher incomes tend to be less price sensitive and have a higher willingness to pay.

Education: Consumers' level of education has an effect on their impressions of the world around them and on the amount of research they conduct prior to making a purchase. Individuals with a higher level of education will spend more time educating themselves before investing their money. Education has an impact on fashion, film, and television programming. Consumers with a higher level of education can be more distrustful of commercials and the facts offered.

Tesco has collected rich customer demographic information through its loyalty program, Tesco Clubcard membership. In the demograhpics.csv dataset, the data scientist team has the following demographic variables:

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company

2.2 Purchase History

"History doesn't repeat itself, but it often rhymes."

This popular aphorism, frequently (and perhaps incorrectly) attributed to Mark Twain, is frequently invoked to demonstrate that, while past events do not always provide a clear indication of future events, they do provide valuable context. This sentiment is especially true for marketing managers, where a consumer's purchase history provides invaluable insight into their future purchasing habits.

Tesco's data engineering team has assembled a cross-sectional customer purchase history data, with variables including

- ID: Customer's unique identifier
- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years

- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- NumDealsPurchases: Number of purchases made with a discount
- NumWebPurchases: Number of purchases made through the company's web site
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's web site in the last month
- Complain: 1 if customer complained in the last 2 years, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise
- Recency: Number of days since customer's last purchase

3 Data Wrangling

3.1 Data Loading

To work on the datasets, we first need to load the raw data into R. The demographic information data are stored as csv files. In R, we can use read.csv(filepath) to load the data into R environment.

For your convenience, I have stored the `demographic.csv` and purchase.csv files on my Dropbox. We can directly feed the url links to `read.csv()` to download and create the dataset.

After running the above code blocks, you should see two datasets in your RStudio environment.

Now, click into each dataset, take a look, and get a sense of how these two datasets look like.

3.2 Data Consolidation

In reality, to accomplish a data analytics task, data scientists often need to collect data from various sources, and assemble them into a larger dataset as needed.

Now we have two Tesco datasets at hand, and we should assemble them into a larger data frame.

- Merge the demographic information into purchase history data. Name the joined data as "data_full"
  - try left_join(), right_join(), inner_join(), and full_join().
  - Do they give you the same results? Why? When would you get different results?

3.3 Data Types

- *Task*: Check all data types in `data_full` are correct and as expected
- *Discussion*: If the variables types are incorrect, think about how would you make it right using `dplyr`?

3.4 Missing Values

- *Tasks*: Are there any missing values in the data?
  - tip: use `datasummary_skim()`, which reports the number of missing values
- *Tasks*: Clean missing values in the dataset.
  - tip: use `mean(var, na.rm = T)` to get the average income; and then replace missing values in `data_full` with the average income using `replace()` function. See below an example or check `replace()` help file to find out its syntax.

4 Preliminary Customer Analysis (Descriptive Analytics)

Next, once the final dataset is ready, we can proceed to use descriptive analytics to conduct preliminary customer analysis using `dplyr`.

Descriptive analytics is concerned with summarizing and highlighting patterns in current and historical data in order to assist businesses in comprehending what has occurred thus far. However, it makes no attempt to explain why something occurred or to forecast what may occur in the future.[3] To answer those questions, businesses must combine descriptive analytics with other types of analysis.

Your task is to 'get to know' the data by conducting some statistical analysis using Tesco's customer database.

1. Provide the summary statistics of the cleaned data. From summary statistics, do you already see any insights? (open question; from central tendency and dispersion perspectives)
2. Report the percent of customers in each Marital Status group and compare the average spendings in each group. What insights can you draw? Tips: at least two methods to do this task
   - Method 1: use `group_by()` + `summarise()`
   - Method 2: use `datasummary_balance()`
3. Which education group accounts for the largest percentage of customers? Answering this question using `dplyr` only.
   - tip: in dplyr, there is a function called `n()`, which counts the number of rows in the group after `group_by()`
4. What is the average Total £ spent on wine and fruit products by customers with and without kids?
   - tip: first mutate a variable called `has_kid`, which equals 1 if the number of kids in the household is larger than 0, and otherwise 0; then group by this `has_kid` variable
5. Which product categories have the most sales? Which have sold the least?

[3] "why something occurred" belongs to the scope of causal inference; "forecast what may occur in the future" falls in the scope of predictive analytics.

6. For both complainers and non-complainers, find the total number and also the percent of customers who responded to the offer.

7. Compute the individual CLV and identify the top 50% customers, assuming the following:

- the annual retention rate is 60% for complainers and 80% for non-complainers (generate a new column called retention rate, based on complainer or non-complainer)
- consider 5 years for customer life
- average COGS 60%
- Variable marketing cost: 15% of total spending
- 10% annual discount rate