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) 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,[1] 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.

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

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

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.

```
## use read.csv() to download and load the data, assign it to an R data object
## header = T argument is to tell read.csv() to keep the dataset header (first row)

# Load demograhpic data, and call it data_demo
data_demo <- read.csv(file = "https://www.dropbox.com/s/a0v38lpydls2emy/demographics.csv?dl=1",
                      header = T)

# Load purchase history data, and call it data_purchase
data_purchase <- read.csv(file = "https://www.dropbox.com/s/de435r8zdxydnhg/purchase.csv?dl=1" , header = T)
```

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?

```
pacman::p_load(dplyr)
# left join
```

```r
data_full <- data_purchase %>%
  left_join(data_demo, by = "ID")
```

```r
# right join
data_full_right_join <- data_purchase %>%
  right_join(data_demo, by = "ID")
```

```r
# inner_join
data_full_inner_join <- data_purchase %>%
  inner_join(data_demo, by = "ID")
```

```r
# full_join
data_full_full_join <- data_purchase %>%
  full_join(data_demo, by = "ID")
```

3.3 Data Types

- *Task*: Check all data types in `data_full` are correct and as expected

```r
# can use str() to get the structure of data
# Lina covered this in the induction week
str(data_full)
```

```
'data.frame':   2000 obs. of  22 variables:
 $ ID                 : int  5524 2174 4141 6182 5324 7446 965 6177 4855 5899 ...
 $ MntWines           : int  635 11 426 11 173 520 235 76 14 28 ...
 $ MntFruits          : int  88 1 49 4 43 42 65 10 0 0 ...
 $ MntMeatProducts    : int  546 6 127 20 118 98 164 56 24 6 ...
 $ MntFishProducts    : int  172 2 111 10 46 0 50 3 3 1 ...
 $ MntSweetProducts   : int  88 1 21 3 27 42 49 1 3 1 ...
 $ MntGoldProds       : int  88 6 42 5 15 14 27 23 2 13 ...
 $ NumDealsPurchases  : int  3 2 1 2 5 2 4 2 1 1 ...
 $ NumWebPurchases    : int  8 1 8 2 5 6 7 4 3 1 ...
 $ NumCatalogPurchases: int  10 1 2 0 3 4 3 0 0 0 ...
 $ NumStorePurchases  : int  4 2 10 4 6 10 7 4 2 0 ...
 $ NumWebVisitsMonth  : int  7 5 4 6 5 6 6 8 9 20 ...
 $ Complain           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Response           : int  1 0 0 0 0 0 0 0 1 0 ...
 $ Year_Birth         : int  1957 1954 1965 1984 1981 1967 1971 1985 1974 1950 ...
 $ Education          : chr  "Graduation" "Graduation" "Graduation" "Graduation" ...
```

```
$ Marital_Status     : chr  "Single" "Single" "Together" "Together" ...
$ Income             : int  58138 46344 71613 26646 58293 62513 55635 33454 30351 5648 ...
$ Kidhome            : int  0 1 0 1 1 0 0 1 1 1 ...
$ Teenhome           : int  0 1 0 0 0 1 1 0 0 1 ...
$ Dt_Customer        : chr  "04/09/2012" "08/03/2014" "21/08/2013" "10/02/2014" ...
$ Recency            : int  58 38 26 26 94 16 34 32 19 68 ...
```

- *Discussion*: If the variables types are incorrect, think about how would you make it right using `dplyr`?
  - For instance, we can observe that `Dt_Customer` is of `character` type, we need to convert it to `date` type in R.

```r
# use mutate to overwrite the previous Dt_Customer
# use as.Date() to make the conversion from character to date type
data_full <- data_full %>%
  mutate(Dt_Customer = as.Date(Dt_Customer, format = "%d/%m/%Y" ))
```

Now, if we check the data type of `Dt_Customer` using `class()`, it is `date` type now!

```r
class(data_full$Dt_Customer)
```

```
[1] "Date"
```

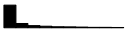
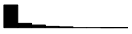
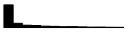
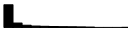
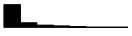
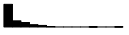
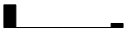
3.4 Missing Values

- *Tasks*: Are there any missing values in the data?
  - tip: use `datasummary_skim()`, which reports the number of missing values

```r
pacman::p_load(modelsummary)
datasummary_skim(data_full)
```

- *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.

```r
# the below code generates a vector a with 2 NAs (not relevant to case study)
# the 2nd and 4th elements are missing values
# the last line of code replaces any missing value with 2
a <- c(1,NA,3,NA)
replace(a, is.na(a), 2)
```

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
ID	2000	0	5599.2	3242.0	0.0	5492.0	11 191.0	
MntWines	738	0	306.1	338.3	0.0	176.5	1493.0	
MntFruits	157	0	26.4	39.9	0.0	8.0	199.0	
MntMeatProducts	532	0	167.9	225.3	0.0	68.0	1725.0	
MntFishProducts	179	0	37.6	54.6	0.0	12.0	259.0	
MntSweetProducts	175	0	27.5	41.8	0.0	8.0	263.0	
MntGoldProds	207	0	43.8	51.7	0.0	24.0	362.0	
NumDealsPurchases	15	0	2.3	2.0	0.0	2.0	15.0	
NumWebPurchases	15	0	4.1	2.8	0.0	4.0	27.0	
NumCatalogPurchases	14	0	2.7	3.0	0.0	2.0	28.0	
NumStorePurchases	14	0	5.8	3.3	0.0	5.0	13.0	
NumWebVisitsMonth	15	0	5.3	2.5	0.0	6.0	20.0	
Complain	2	0	0.0	0.1	0.0	0.0	1.0	
Response	2	0	0.2	0.4	0.0	0.0	1.0	
Year_Birth	59	0	1968.8	12.0	1893.0	1970.0	1996.0	
Income	1783	1	52 139.7	21 492.4	1730.0	51 518.0	162 397.0	
Kidhome	3	0	0.4	0.5	0.0	0.0	2.0	
Teenhome	3	0	0.5	0.5	0.0	0.0	2.0	
Recency	100	0	49.2	29.0	0.0	50.0	99.0	

```
[1] 1 2 3 2
```

```r
data_full <- data_full %>%
  mutate(Income = replace(Income, is.na(Income), mean(Income,na.rm = T)))
```

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.[2] 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)

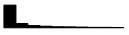
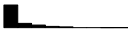
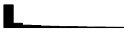
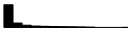
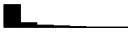
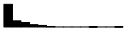
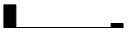
```r
# remember to load package modelsummary before calling this function
datasummary_skim(data_full)
```

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
   - Tips: use `group_by()` + `summarise()`

To compute the percentage of customers in each marital status group

```r
# The below code computes the percentage of customers in each marital status group
data_full %>%
  group_by(Marital_Status)%>%
  summarise(n_customer_each_group = n())%>% #n() counts the # of customer in each group
  ungroup() %>%
  mutate(total_customers = sum(n_customer_each_group))%>% # total number of customers is the sum of number of
  mutate(percentage_customers = n_customer_each_group/total_customers)
```

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

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
ID	2000	0	5599.2	3242.0	0.0	5492.0	11 191.0	
MntWines	738	0	306.1	338.3	0.0	176.5	1493.0	
MntFruits	157	0	26.4	39.9	0.0	8.0	199.0	
MntMeatProducts	532	0	167.9	225.3	0.0	68.0	1725.0	
MntFishProducts	179	0	37.6	54.6	0.0	12.0	259.0	
MntSweetProducts	175	0	27.5	41.8	0.0	8.0	263.0	
MntGoldProds	207	0	43.8	51.7	0.0	24.0	362.0	
NumDealsPurchases	15	0	2.3	2.0	0.0	2.0	15.0	
NumWebPurchases	15	0	4.1	2.8	0.0	4.0	27.0	
NumCatalogPurchases	14	0	2.7	3.0	0.0	2.0	28.0	
NumStorePurchases	14	0	5.8	3.3	0.0	5.0	13.0	
NumWebVisitsMonth	15	0	5.3	2.5	0.0	6.0	20.0	
Complain	2	0	0.0	0.1	0.0	0.0	1.0	
Response	2	0	0.2	0.4	0.0	0.0	1.0	
Year_Birth	59	0	1968.8	12.0	1893.0	1970.0	1996.0	
Income	1783	0	52 139.7	21 400.8	1730.0	51 844.5	162 397.0	
Kidhome	3	0	0.4	0.5	0.0	0.0	2.0	
Teenhome	3	0	0.5	0.5	0.0	0.0	2.0	
Recency	100	0	49.2	29.0	0.0	50.0	99.0	

Marital_Status	n_customer_each_group	total_customers	percentage_customers
Alone	3	2000	0.0015
Divorced	206	2000	0.1030
Married	767	2000	0.3835
Single	436	2000	0.2180
Together	521	2000	0.2605
Widow	67	2000	0.0335

To compare the average spendings of customers in each marital status group

```
data_full %>%
  mutate(total_spending = MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGol
  group_by(Marital_Status)%>% # group by
  summarise(avg_spending = mean(total_spending))%>% # compute the mean of spending for each group
  ungroup()
```

Marital_Status	avg_spending
Alone	256.6667
Divorced	620.2379
Married	591.8214
Single	606.8578
Together	619.7620
Widow	722.9552

In fact, the two questions can be answers in one code block, due to the utmost elegance of R dplyr!

```
data_full %>%
  mutate(total_spending = MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGol
  group_by(Marital_Status)%>%
  summarise(n_customer_each_group = n(),#n() counts the # of customer in each group
            avg_spending = mean(total_spending))%>% # compute the mean of spending for each group
  ungroup() %>%
  mutate(total_customers = sum(n_customer_each_group))%>% # total number of customers is the sum of number of
  mutate(percentage_customers = n_customer_each_group/total_customers)
```

Marital_Status	n_customer_each_group	avg_spending	total_customers	percentage_customers
Alone	3	256.6667	2000	0.0015
Divorced	206	620.2379	2000	0.1030
Married	767	591.8214	2000	0.3835
Single	436	606.8578	2000	0.2180
Together	521	619.7620	2000	0.2605
Widow	67	722.9552	2000	0.0335

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()`

```
data_demo%>% # Think about why data_demo? what if we use data_full?
  group_by(Education)%>% # group by Education
  summarise(n_customers_in_group = n()) %>% # compute the number of customers in each education group
  ungroup() %>% # important to do the ungroup after each group_by
  mutate(percent_customer = n_customers_in_group/ sum(n_customers_in_group)) %>% # compute the percentage of
  arrange(-percent_customer) # descending order
```

Education	n_customers_in_group	percent_customer
Graduation	992	0.4960
PhD	453	0.2265
Master	327	0.1635
2n Cycle	185	0.0925
Basic	43	0.0215

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

```
data_full %>%
  mutate(has_kid = ifelse(Kidhome>0,1,0)) %>% # generate a new variable indicating if there is kid
  group_by(has_kid) %>%
  summarise(avg_spending_wine = mean(MntWines, na.rm = T),
            avg_spending_fruit = mean(MntFruits, na.rm = T)) %>%
  ungroup()
```

has_kid	avg_spending_wine	avg_spending_fruit
0	453.7390	39.61798
1	103.3677	8.15777

5. Which product categories have the most sales? Which have sold the least? Use `dplyr` only.

```
data_full %>%
  select(starts_with("Mnt")) %>% # this is a shortcut `select` function by dplyr; check its help manual
  sapply( sum) %>%
  sort()
```

```
MntFruits MntSweetProducts  MntFishProducts     MntGoldProds
```

```
          52715            54990            75228            87572
MntMeatProducts         MntWines
          335770           612115
```

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

```
data_full %>%
  group_by(Complain) %>%
  summarise(n_responder = sum(Response, na.rm = T),
            percent_responder = n_responder / n(),
            n_non_responder = n() - n_responder,
            percent_non_responder = n_non_responder / n()
            ) %>%
  ungroup()
```

Complain	n_responder	percent_responder	n_non_responder	percent_non_responder
0	298	0.1505051	1682	0.8494949
1	3	0.1500000	17	0.8500000

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

> **i** User Defined Function
>
> UDF is covered in the induction week Friday session (by Lina). You can also refer to the following tutorials to learn more about UDF.
>
> - simple
> - advanced

```
# define a function that can compute CLV for each customer.
compute_CLV <- function(retention,N,spending,discount,COGS, marketing_cost){
  # revenue sequence
  revenue_seq <- rep(spending,N)

  # profit sequence g = M-c
  profit_seq <- revenue_seq * (1 - COGS - marketing_cost)

```

```r
  # apply retention rate

  profit_after_churn <- profit_seq * (retention ^ (seq(1,N) - 1) )

  # discount the future profits

  profit_after_churn_discount <- profit_after_churn * ( 1/ (1+discount)) ^ seq(1:N)

  # CLV; note that CAC is not available in this case, so the CLV here does not include CAC. If CAC information

  CLV <- sum(profit_after_churn_discount)

  return(CLV)

}

data_full <- data_full %>%
  mutate(retention_rate = ifelse(Complain == 1, 0.6, 0.8),# generate individual specific churn rate
         total_spending = MntFishProducts + MntFruits + MntGoldProds + MntMeatProducts + MntSweetProducts + Mn
  rowwise()%>% # row-wise operation for each row to compute the CLV
  mutate(CLV = compute_CLV(retention_rate,5,total_spending/2,0.1,0.6,0.15))%>%
  ungroup() %>%
  arrange(-CLV)
```

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.