Class 6 Descriptive Analytics for M&S

Dr. Wei Miao

UCL School of Management

October 16, 2024

Section 1

Data Cleaning

Missing Values

- In R, missing values are represented by the symbol `NA` (i.e., not available).
- Most statistical models cannot handle missing values, so we need to deal with them in R.
- If there are just a few missing values: remove them from analysis.
- If there are many missing values: need to replace them with appropriate values:
    - mean/median/imputation

Outliers

- **Outliers** are data points that are significantly different from other data points in the dataset, such as unusually large and small values.
- **Winsorization** is a common method to deal with outliers. It replaces the extreme values with the nearest non-extreme value, usually the 99th or 1th percentile (or other thresholds as appropriate).

Section 2

Descriptive Analytics

Two Major Tasks of Descriptive Analytics

- You can think of descriptive analytics as **creating a dashboard** to display the key information you would like to know for your business. For instance:

1. Describe data depending on your business purposes
   - “How much do our customers spend each month on average?”
   - “What percentage of our customers are unprofitable?”
   - “What is the difference between the retention rates across different demographic groups?”
2. Conduct statistical tests (such as t-tests) for hypothesis testing.
   - Is there any significant difference in the average spending between different age/gender groups?
   - Based on our test mailing, can we conclude that ad-copy A works better than ad-copy B?

Summary Statistics

- **Summary statistics** are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible.
- There are two main types of summary statistics used in evaluation:
  - **measures of central tendency**: number of observations, mean, min, 25 percentile, median, 75 percentile, max, etc.
  - **measures of dispersion:** range and standard deviation.
- It's important to include summary statistics table in your dissertation before any statistical analysis!

Summary Statistics with R

- In R, a power package to report summary statistics is called `modelsummary`.
- `datasummary_skim()` is a shortcut to conduct basic summary statistics
- For more features, refer to the package tutorial here

```r
pacman::p_load(modelsummary)
## Summary statistics for numeric variables
data_full %>%
  datasummary_skim(type = "numeric")

## Summary statistics for categorical variables
data_full %>%
  datasummary_skim(type = "categorical")
```

Section 3

M&S Descriptive Analytics

M&S Descriptive Analytics

Let’s move on to the Quarto document to see how we can apply the descriptive analytics to the M&S dataset.