

## Class 4 Data Wrangling with R Part I

Dr Wei Miao

UCL School of Management

October 11, 2023

## Section 1

### Overview

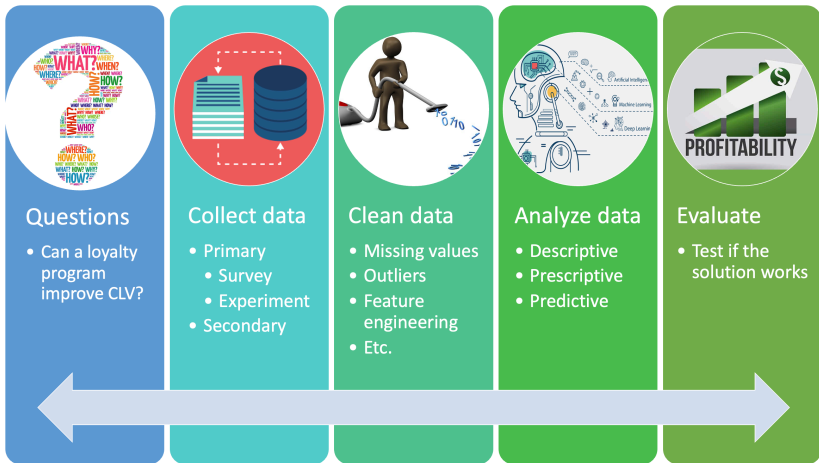
# Class Objectives

- Understand the major steps to conduct data analytics
- **Data collection:** Learn how to collect first-hand data
- **Data cleaning:** Learn how to use the `dplyr` package to collect, load, and clean data
- **Data analysis:** Learn how to conduct descriptive analytics

## Section 2

# Data Analytics Workflow

# Overview



# Collect Data

- **Primary Data:** Data that are generated by the researcher himself/herself, surveys, interviews, experiments, specially designed for understanding and solving the research problem at hand.
- **Secondary Data:** Existing data generated by the company's or consumer's past activities, as part of organizational record keeping.

| BASIS FOR COMPARISON     | PRIMARY DATA   | SECONDARY DATA  |
|--------------------------|--|---|
| Meaning                  | Primary data refers to the first hand data gathered by the researcher himself. | Secondary data means data collected by someone else earlier.                      |
| Data                     | Real time data   | Past data   |
| Process                  | Very involved  | Quick and easy  |
| Source                   | Surveys, observations, experiments, questionnaire, personal interview, etc.    | Government publications, websites, books, journal articles, internal records etc. |
| Cost effectiveness       | Expensive  | Economical  |
| Collection time          | Long   | Short   |
| Specific                 | Always specific to the researcher's needs.                                     | May or may not be specific to the researcher's need.                              |
| Available in             | Crude form   | Refined form  |
| Accuracy and Reliability | More   | Relatively less   |

# Collect Data: Marketing Surveys

- In a marketing survey, we typically would like to solicit the following
  - purchase intention
  - willingness to pay (WTP)
  - shopping basket
  - share of wallet (SoW)
  - demographics
- **Let's see an example of how to design a simple marketing survey!**
- Useful supplementary readings if you need to design marketing surveys
  - [The quick start guide on how to conduct market research](#)

## Section 3

# Data Wrangling with R



# Data Frames

- Data Frame is the R object that we will deal with most of the time in the MSc program. You can think of `data.frame` as a spreadsheet in excel
- Each row stands for an observation
- Each column stands for a variable; each column should have a **unique** name.
  - Each column must contain the same data type, but the different columns can store different data types.
    - compare with matrix?
- Each column must be of same length, because rows have the same length across variables.

# Install and Load the dplyr package

- In R, we will be using the dplyr package for data cleaning and manipulation.

```
1 install.packages("dplyr")
```

- Load the package

```
1 library(dplyr)
```

- Load a built-in dataset called mtcars using data()

```
1 data("mtcars")
```

- To browse the whole dataset, we can simply click the dataset in the environment
  - It may takes time to view a huge dataset

## Subset Rows Based on Conditions: `filter`

- We can use `filter()` to extract rows that meet logical criteria.
- We can also add multiple criteria separated by comma

```
1 # show all cars with more than 4 gears
2 filter(mtcars, gear == 4 )
```

|               | mpg  | cyl | displacement | hp  | drat | wt    | qsec  | vs | am | gear | carb |
|---------------|------|-----|--------------|-----|------|-------|-------|----|----|------|------|
| Mazda RX4     | 21.0 | 6   | 160.0        | 110 | 3.90 | 2.620 | 16.46 | 0  | 1  | 4    | 4    |
| Mazda RX4 Wag | 21.0 | 6   | 160.0        | 110 | 3.90 | 2.875 | 17.02 | 0  | 1  | 4    | 4    |
| Datsun 710    | 22.8 | 4   | 108.0        | 93  | 3.85 | 2.320 | 18.61 | 1  | 1  | 4    | 1    |
| Merc 240D     | 24.4 | 4   | 146.7        | 62  | 3.69 | 3.190 | 20.00 | 1  | 0  | 4    | 2    |
| Merc 230      | 22.8 | 4   | 140.8        | 95  | 3.92 | 3.150 | 22.90 | 1  | 0  | 4    | 2    |
| Merc 280      | 19.2 | 6   | 167.6        | 123 | 3.92 | 3.440 | 18.30 | 1  | 0  | 4    | 4    |
| Merc 280C     | 17.8 | 6   | 167.6        | 123 | 3.92 | 3.440 | 18.90 | 1  | 0  | 4    | 4    |
| Fiat 128      | 32.4 | 4   | 78.7         | 66  | 4.08 | 2.200 | 19.47 | 1  | 1  | 4    | 1    |
| Honda Civic   | 30.4 | 4   | 75.7         | 52  | 4.93 | 1.615 | 18.52 | 1  | 1  | 4    | 2    |

# The Pipe Operator %>%

## Pipe Operator

%>%, or pipe operator, will forward a value, or the result of an expression, into the next function call/expression.

```
1 mtcars %>% filter(gear == 4) %>% head()
```

## Sort Rows: arrange

`arrange()` orders the rows of a data frame by the values of selected columns.

- The default is by ascending order; for descending order, put a minus sign before the variable.

```
1 # reorder mtcars based on hp
2 mtcars %>%
3   arrange(hp) %>%
4   head()
```

|                | mpg  | cyl | disp  | hp | drat | wt    | qsec  | vs | am | gear | carb |
|----------------|------|-----|-------|----|------|-------|-------|----|----|------|------|
| Honda Civic    | 30.4 | 4   | 75.7  | 52 | 4.93 | 1.615 | 18.52 | 1  | 1  | 4    | 2    |
| Merc 240D      | 24.4 | 4   | 146.7 | 62 | 3.69 | 3.190 | 20.00 | 1  | 0  | 4    | 2    |
| Toyota Corolla | 33.9 | 4   | 71.1  | 65 | 4.22 | 1.835 | 19.90 | 1  | 1  | 4    | 1    |
| Fiat 128       | 32.4 | 4   | 78.7  | 66 | 4.08 | 2.200 | 19.47 | 1  | 1  | 4    | 1    |
| Fiat X1-9      | 27.3 | 4   | 79.0  | 66 | 4.08 | 1.935 | 18.90 | 1  | 1  | 4    | 1    |
| Porsche 914-2  | 26.0 | 4   | 120.3 | 91 | 4.43 | 2.140 | 16.70 | 0  | 1  | 5    | 2    |

## Generate New Variables: mutate

**mutate()** adds new variables and preserves existing ones

```
1 mtcars %>%
2   mutate(sqrt_mpg = sqrt(mpg))%>%
3   head()
```

|                   | mpg  | cyl | disp | hp  | drat | wt    | qsec  | vs | am | gear | carb | sqrt_mpg |
|-------------------|------|-----|------|-----|------|-------|-------|----|----|------|------|----------|
| Mazda RX4         | 21.0 | 6   | 160  | 110 | 3.90 | 2.620 | 16.46 | 0  | 1  | 4    | 4    | 4.582576 |
| Mazda RX4 Wag     | 21.0 | 6   | 160  | 110 | 3.90 | 2.875 | 17.02 | 0  | 1  | 4    | 4    | 4.582576 |
| Datsun 710        | 22.8 | 4   | 108  | 93  | 3.85 | 2.320 | 18.61 | 1  | 1  | 4    | 1    | 4.774935 |
| Hornet 4 Drive    | 21.4 | 6   | 258  | 110 | 3.08 | 3.215 | 19.44 | 1  | 0  | 3    | 1    | 4.626013 |
| Hornet Sportabout | 18.7 | 8   | 360  | 175 | 3.15 | 3.440 | 17.02 | 0  | 0  | 3    | 2    | 4.324350 |
| Valiant           | 18.1 | 6   | 225  | 105 | 2.76 | 3.460 | 20.22 | 1  | 0  | 3    | 1    | 4.254410 |

# Important Tips

- Each dplyr operation does not overwrite the original data frame, therefore, we **must assign the object back** if we wish to overwrite the previous data frame.
- Exercises:
  - find car models with gear equal to 4 and mpg larger than 15
  - reorder the above dataset by wt from large to small.
  - generate a new column which computes the ratio of mpg to wt

## After-Class Exercise

- Data camp dplyr exercise
- Read “Preliminary Customer Analyses” dataset, and try to solve the case questions using the techniques learned today