

SHETH L.U.J AND SIR M.V COLLEGE

Aim: Applying basic data cleaning functions: handling missing values using na.omit()/replace_na() in R. import dataset.

```
R - R 4.5.2 - ~/R
> market_df <- read.csv("annex1.csv", na.strings = c("", "NA"))
> print("--- 1. Original Data (First 6 Rows) ---")
[1] "--- 1. Original Data (First 6 Rows) ---"
> print(head(market_df))
  Item.Code Item.Name Category.Code Category.Name
1 1.029e+14 Niushou Shengcai 1011010101 Flower/Leaf Vegetables
2 1.029e+14 Sichuan Red Cedar 1011010101 Flower/Leaf Vegetables
3 1.029e+14 Local Xiaomao Cabbage 1011010101 Flower/Leaf Vegetables
4 1.029e+14 White Caitai 1011010101 Flower/Leaf Vegetables
5 1.029e+14 Amaranth 1011010101 Flower/Leaf Vegetables
6 1.029e+14 Yunnan Shengcai 1011010101 Flower/Leaf Vegetables
>
> print("--- Count of Missing Values per Column ---")
[1] "--- Count of Missing Values per Column ---"
> print(colSums(is.na(market_df)))
  Item.Code Item.Name Category.Code Category.Name
         0         0         0         0
>
> clean_omit <- na.omit(market_df)
>
> print("--- 2. Data after na.omit() ---")
[1] "--- 2. Data after na.omit() ---"
> print(paste("Original rows:", nrow(market_df)))
[1] "Original rows: 251"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 251"
> print(head(clean_omit))
  Item.Code Item.Name Category.Code Category.Name
1 1.029e+14 Niushou Shengcai 1011010101 Flower/Leaf Vegetables
2 1.029e+14 Sichuan Red Cedar 1011010101 Flower/Leaf Vegetables
3 1.029e+14 Local Xiaomao Cabbage 1011010101 Flower/Leaf Vegetables
4 1.029e+14 White Caitai 1011010101 Flower/Leaf Vegetables
5 1.029e+14 Amaranth 1011010101 Flower/Leaf Vegetables
6 1.029e+14 Yunnan Shengcai 1011010101 Flower/Leaf Vegetables
>
> clean_replace <- market_df %>%
+   replace_na(list(
+     Item.Code = "Unknown",
+     Item.Name = "Unknown Item",
+     Category.Code = "0000000000"
+   ))
```

```
R - R 4.5.2 - ~/R
[1] "--- 1. Original Data (First 6 Rows) ---"
> print(head(market_df))
  Item.Code Item.Name Category.Code Category.Name
1 1.029e+14 Niushou Shengcai 1011010101 Flower/Leaf Vegetables
2 1.029e+14 Sichuan Red Cedar 1011010101 Flower/Leaf Vegetables
3 1.029e+14 Local Xiaomao Cabbage 1011010101 Flower/Leaf Vegetables
4 1.029e+14 White Caitai 1011010101 Flower/Leaf Vegetables
5 1.029e+14 Amaranth 1011010101 Flower/Leaf Vegetables
6 1.029e+14 Yunnan Shengcai 1011010101 Flower/Leaf Vegetables
>
> print("--- Count of Missing Values per Column ---")
[1] "--- Count of Missing Values per Column ---"
> print(colSums(is.na(market_df)))
  Item.Code Item.Name Category.Code Category.Name
         0         0         0         0
>
> clean_omit <- na.omit(market_df)
>
> print("--- 2. Data after na.omit() ---")
[1] "--- 2. Data after na.omit() ---"
> print(paste("Original rows:", nrow(market_df)))
[1] "Original rows: 251"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 251"
> print(head(clean_omit))
  Item.Code Item.Name Category.Code Category.Name
1 1.029e+14 Niushou Shengcai 1011010101 Flower/Leaf Vegetables
2 1.029e+14 Sichuan Red Cedar 1011010101 Flower/Leaf Vegetables
3 1.029e+14 Local Xiaomao Cabbage 1011010101 Flower/Leaf Vegetables
4 1.029e+14 White Caitai 1011010101 Flower/Leaf Vegetables
5 1.029e+14 Amaranth 1011010101 Flower/Leaf Vegetables
6 1.029e+14 Yunnan Shengcai 1011010101 Flower/Leaf Vegetables
>
> clean_replace <- market_df %>%
+   replace_na(list(
+     Item.Code = "Unknown",
+     Item.Name = "Unknown Item",
+     Category.Code = "0000000000",
+     Category.Name = "Uncategorized"
+   ))
```

YUKTA SONAWANE S120
R PROGRAMMING
PRACTICAL NO : 8