TASK 1

We are using Medical Appointment No Shows.

Our objective is to solve the given Problem, following code we are using –

```
is.na(`KaggleV2.May.2016[1]`)          # Returns TRUE/FALSE matrix
sum(is.na(`KaggleV2.May.2016[1]`))     # Count of all NA values
colSums(is.na(`KaggleV2.May.2016[1]`)) # NA count per column

library(dplyr)
`KaggleV2.May.2016[1]` <- `KaggleV2.May.2016[1]` %>% distinct()  # Remove full duplicate rows

`KaggleV2.May.2016[1]`$Gender <- tolower(`KaggleV2.May.2016[1]`$Gender)

`KaggleV2.May.2016[1]`$Gender[`KaggleV2.May.2016[1]`$Gender %in% c("m", "male")] <- "Male"
`KaggleV2.May.2016[1]`$Gender[`KaggleV2.May.2016[1]`$Gender %in% c("f", "female")] <- "Female"

str(`KaggleV2.May.2016[1]`)          # Structure of the dataset
sapply(`KaggleV2.May.2016[1]`, class)  # Class of each column
```

Some code are not applied because dataset fulfill our creteria.

If missing values are few and not critical, use na.omit() in R.

Use duplicated() in R

Use dropna() when missing data is not usable, and fillna() when you want to keep the record but replace the missing parts.

Outliers are values that differ significantly from most other data points. They can be due to errors or real extreme values.

Standardizing means converting data to a common scale without distorting differences.

Handling date/time inconsistencies:

- Detect different formats (e.g., "12/03/2024" vs "2024-03-12")
- Convert them to a common format using:
    - **R**: as.Date() or lubridate functions

common data cleaning challenges

- Missing or incomplete data
- Duplicate entries
- Inconsistent formatting (case sensitivity, date formats)
- Outliers and anomalies
- Mixed data types in one column
- Typos or spelling errors
- Misleading or incorrect data labels

We can assess data quality through:

- **Missing Value Analysis**
- **Duplicate Checks**
- **Data Type Verification**
- **Range and Validity Checks**
- **Outlier Detection**
- **Consistency Audits** (e.g., column naming, units)
- **Summary statistics** (mean, min, max, frequency tables)