**Learning Goals**

- Identify missing and suspicious values using `is.na()`, `complete.cases()`, `summary()`, `table()`, and simple plots to flag potential data quality issues.
- Distinguish between truly missing values and miscoded/implied missing values, and recode them properly as `NA`.
- Evaluate the impact of missing data on analysis by computing counts/proportions of missingness by row and by column, and deciding when dropping rows/columns is reasonable.
- Apply basic imputation strategies when appropriate and clearly document the assumptions being made.
- Export and re-import cleaned data with `write.csv()` and `read.csv()`, controlling for row.names, header issues, and column data types.

# Key Definitions / Functions

- `is.na()`:
- `complete.cases()`:
- `rowSums()` / `colSums()`:
- `summary()`:
- `table()` / `unique()`:
- `write.csv()`:
- `read.csv()`:

# Practice Problems

For each task below, write the R code you would use and briefly describe what you expect the output to look like.

1. Using the built-in `airquality` dataset, determine how many missing values occur in each column and which column has the most missing values.

2. Using `airquality`, display the rows where `Ozone` is missing. Then display the rows where `Solar.R` is missing. Briefly compare what you notice about which months these missing values occur in.

3. Using `airquality`, create a new dataset called `aq_clean` that removes any rows with missing values in `Ozone` or `Solar.R`. Report how many rows remain after cleaning.

4. Using `airquality`, create a new dataset called `aq_impute` where missing `Ozone` values are replaced with the overall mean `Ozone`. Then compute the mean `Ozone` in `airquality` and in `aq_impute` and state why they are the same or different.