Data cleaning and management

CHEATSHEET

Exam Mon, 7 Dec 2026

EXAM LIKELIHOOD Data cleaning and missing values (NA) **92%** · Outliers: investigate, do not blindly remove **92%** · Data inconsistencies (notation, types, units) **88%** · Transformations (log transform) against extreme values **85%** · Handling NA: filtering, imputation, deletion **55%** · Merging datasets (merge, primary key) **30%**

Data cleaning steps

1. Inspect the data
2. Missing values
3. Outliers and errors
4. Duplicates and types
5. Document the choices

Which test for which question?

Difference

Comparing means: t-test (2 groups) or ANOVA (3 or more).

Relationship

Two numeric variables: correlation. Two categorical: chi-square.

Prediction

Predict Y from X: regression (linear or logistic).

M 01 Principles

- Cleaning has no fixed recipe; keep raw data intact; document every step and new variable.
- Check after every operation: rows, NAs, Inf, duplicates.

M 02 Missing values (NA)

- Filtering (only for this analysis), imputing (mean, median: risk of bias), deleting (safe for a small part).
- Mean imputation: mean the same, standard deviation smaller.
- R: is.na(x), sum(is.na(x)), colSums(is.na(df)), mean(x, na.rm = TRUE), complete.cases(df), na.omit(df).
- Error codes (-999) are NAs, not outliers.

M 03 Inconsistencies

- Date formats, text variants, data types, units. Standardize, convert, find duplicates (duplicated, unique).
- str(df) shows types. as.factor(), as.numeric(), as.Date() or lubridate (ymd).
- Recoding: ifelse(), cut(), case_when().

M 04 Outliers

- Investigate first: error or rare but valid? Error: remove or NA. Skewed: log or other transformation.
- Detection: boxplot rule (1.5 x IQR) or z-score; document the criterion.

M 05 Merging

- rbind(): same variables, stack rows.
- merge(a, b, by = "key"): join columns. all.x = TRUE: keep all rows from the left (NA if no match). Default: inner join.
- Choose a primary key, remove duplicates first, match data types, check the row count.

M 06 New variables

- Transformation, binning, indicator, combination, date part, rank.
- Check NA and Inf after a transformation (log(0) = -Inf). Clear names.
- dplyr: mutate, filter, select, group_by, summarise, pipe. Long (row per observation) versus wide.

M 07 Scaling

- Normalization: (x – min) / (max – min), range [0, 1], sensitive to outliers.
- Standardization: z = (x – mu) / sigma, unbounded.
- Needed for K-Means, kNN, PCA, SVM.

M 08 Divvy Bikes workshop

- Load, types, remove docked_bike, NAs and outliers, duration and distance, clean weather data, merge on hourly timestamp, exploratory tests and correlations.

! Common mistakes

- Deleting rows with NA without looking at how many there are or why they are missing.
- Leaving error codes (-999, 9999) in as real values.
- Merging without checking the row count; duplicates in the key.
- Overwriting raw columns instead of creating a new variable.
- Seeing mean imputation as neutral: it reduces the spread.

★ Exam

- About 2 multiple-choice questions: read code and output, no code writing.