SUMMARY · AT EXAM LEVEL

Module 4

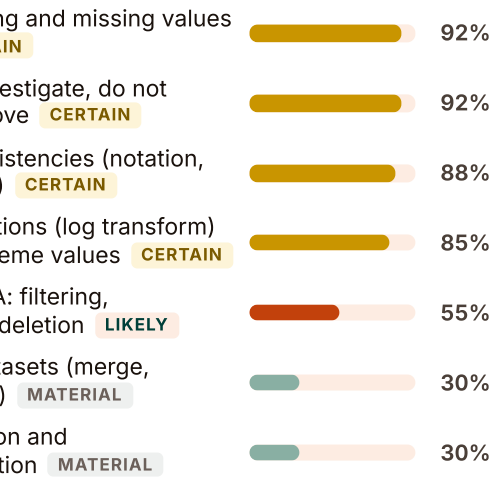
Data cleaning and management

Everything you must be able to define, explain, apply and assess for the open questions. With the models as figures, the comparisons as tables and the pitfalls marked.

WHAT IS INSIDE

01	Purpose and place in the process	2
02	Missing values	2
03	Inconsistencies	3
04	Outliers	3
05	Merging datasets	3
06	New variables	4
07	Normalizing and standardizing	4
08	The workshop: Divvy Bikes and weather	4
09	Common mistakes	5
10	At the exam	5
→	How to write the answer	5
→	Checklist: can I explain this?	6
→	Test yourself	6

EXAM LIKELIHOOD



Estimate per topic, based on the sample exam and the emphasis in the material.

KEY CONCEPTS

data cleaning · NA · filtering · imputation · deletion · data inconsistencies · factor · outliers · log transformation · merge · rbind · primary key · new variables · normalization

ON THE EXAM

Block 4

5 points · 180 min in total

MC	2 multiple-choice questions	2.5 points per question
7	Block 7 (25 points) consists of three open questions about R output.	200 / 250 / 150 words

Exam: Mon, 7 Dec 2026 · 180 min · Technical terms stay in English

01 Purpose and place in the process

Data cleaning is detecting and correcting errors and inconsistencies to raise data quality (accurate, consistent, complete, reliable) for modelling, reporting and decisions. It is not a standard process: the actions depend on the data. It takes much time (an estimated 60% of data scientists' work) and poor preparation leads to poor decisions.

02 Missing values



FIGURE 4.1 **Data cleaning steps.** Understand first, then clean, and record what you did.

A **missing value** is a data point that was not recorded or is not available (in R: **NA**). Causes: loss during collection or transfer, entry errors or omissions, not-applicable answers (skip patterns). Consequences: bias, less power, errors in the analysis.

Approach	What	Advantage or risk
Filtering	leave out NAs only for this analysis	original dataset stays intact, flexible if you revise
Imputing	replace by mean, median or another estimate	risk of bias; mean imputation reduces the spread
Deleting	drop rows with NAs	safe for a small part; distortion if much data is missing

Useful R operations to recognise: **is.na(x)**, **sum(is.na(x))** (number of NAs), **colSums(is.na(df))** (per column), **mean(x, na.rm = TRUE)**, **complete.cases(df)** and **na.omit(df)**. Note the difference between a percentage per column and the number of **complete cases**: a dataset with 10% NA in three columns can lose far more than 10% of the rows with **na.omit()**. With mean imputation the mean does not change, but the standard deviation falls (larger n, same sum of squares), and associations become weaker.

03 Inconsistencies

Examples: date formats (MM/DD/YYYY versus DD-MM-YYYY), categorical variants ("Male", "male", "M"), data types (character versus **factor**), units (euros and dollars, kilos and pounds). Approach: standardize formats (date/time, text case, types), convert units, **always check for duplicates** (**duplicated()**, **unique()**), fix typos and wrong entries with the same structure.

Types and factors: **str(df)** shows the types; **as.factor()**, **as.numeric()** and **as.Date()** (or lubridate functions such as **ymd()**) convert. A factor has **levels**; a variable with numbers stored as text must first become numeric before you can take a mean. Recoding can be done with **ifelse()**, **cut()** (binning) or dplyr **case_when()**.

04 Outliers

An outlier is a data point that differs strongly from the rest; it can point to an error, a rare event or real variation, and can distort the analysis.

Approach: **investigate** (error or rare but valid?), **remove** if it is clearly a mistake, **transform** (log or another function) for a skewed distribution. Do not remove blindly: understand why first. A common detection rule is the boxplot rule (beyond 1.5 times the IQR from the quartile) or a z-score; you must be able to justify and document the chosen criterion. An error code such as -999 for "unknown" is not an outlier but an NA that you must convert.

05 Merging datasets

Situation	Function
same variables, several files (months): stack rows	rbind()
same units, different variables: join columns on a key	merge(a, b, by = "customer_id")
many-to-one, keep all transactions	merge(a, b, all.x = TRUE)

Note: choose a **primary key**, remove duplicates beforehand, make data types match (no mix of integer and character), decide what to do with non-matches (a customer without transactions, a transaction from a customer not in the list) and check the output after merging (number of rows!). With **all.x = TRUE** unmatched rows from the left table stay with **NA** in the new columns; with default **merge()** they disappear (inner join). A common mistake is joining on a key with duplicates: the number of rows then increases.

06 New variables

Kinds: transformation (**log(income)**, **sqrt(price)**), binning (age into child, adult, senior), indicator or flag (**is_high_income = income > 100000**), combination (**revenue = price * quantity**), date parts (**lubridate::year(order_date)**) and ranking or score. Rules: **document** new variables, check for **NA** and **Inf** after a transformation (such as **log(0)**), do not change the raw columns directly, use clear names.

With dplyr you use **mutate()** for new columns, **filter()** for rows, **select()** for columns, **group_by()** with **summarise()** for summaries and the pipe **%>%** or **|>** to chain steps. Data can be **long** (one row per observation, for example per hour or per measurement) or **wide** (one column per measurement); models and plots usually expect long, summary tables are often wide.

07 Normalizing and standardizing

Reason: models that are sensitive to scale (K-Means, kNN, PCA, SVM) let variables with a large range dominate; scaling makes the comparison fair.

Method	Formula	Output range
Normalization	x' = (x – min) / (max – min)	[0, 1]
Standardization	z = (x – mu) / sigma	unbounded

Normalization is sensitive to outliers (min and max determine everything); standardization does not change the shape of the distribution.

08 The workshop: Divvy Bikes and weather

Tasks: load and explore the dataset (**str**, **summary**), check variable types (factor, time object), remove the rows with **docked_bike**, handle NAs, deal with outliers in trip duration and distance, create new variables (duration in minutes or hours, distance with the Haversine formula), clean and extend the weather data (daily averages for temperature and wind speed, daily total for precipitation), merge both datasets on an **hourly timestamp** that you create in both, and exploratory (model-free) analyses with tests and correlations.

09

WATCH OUT

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.
- ✖ Removing outliers automatically.

10

At the exam

About 2 multiple-choice questions on module 4. The formula sheet is hardly relevant here; the reasoning counts.

→

MODEL ANSWER

How to write the answer

Four building blocks, in this order. Each one earns points; leave none out.

1. Term
Name the model or concept with the English term.

2. Definition
Give the definition or features, precisely.

3. Case fact
Link it to a concrete fact from the case.

4. Conclusion
End with a judgement, advice or consequence.

QUESTION 1 · WORD LIMIT 75 WORDS (WORD LIMIT ASSUMED)

In a dataset, 4 of 500 rows have a missing value. What do you do, and what do you record?

TERM Inspect first, then make a deliberate choice for the missing values. DEFINITION You can delete the rows or impute the values, and you weigh what this does to the sample. CASE FACT Few rows are affected (4 of 500), so deleting barely changes the sample, provided the missing values are not systematic. CONCLUSION Record the choice and the number of deleted rows, so the analysis stays reproducible.

63 / 75 words

KEY CONCEPTS

✓ Checklist: can I explain this?

Tick what you can explain from memory in one sentence, using the English term.

- ☐ data cleaning
- ☐ NA
- ☐ filtering
- ☐ imputation
- ☐ deletion
- ☐ data inconsistencies
- ☐ factor
- ☐ outliers
- ☐ log transformation
- ☐ merge
- ☐ rbind
- ☐ primary key
- ☐ new variables
- ☐ normalization
- ☐ standardization
- ☐ long and wide

FLASHCARDS

? Test yourself

Cover the right-hand column and say the answer out loud.

Question	Answer
1 What is data cleaning?	Detecting and correcting errors and inconsistencies so that data are accurate, consistent, complete and reliable for analysis.
2 What are typical causes of missing values?	Data loss during collection or transfer, entry errors or omissions, and not-applicable answers (skip patterns).
3 Which three ways are there to deal with missing values?	Filtering, replacing (imputation) or removing (deletion) the rows with missing values.
4 How do you remove rows with an NA in R?	With na.omit(data) or drop_na(data); with drop_na(column) only the rows with an NA in that column.
5 What does as.numeric() do on the text "n/a"?	It gives NA with a warning, because "n/a" cannot be read as a number.
6 What does recoding (for example ifelse) do to a categorical variable?	It replaces values or groups with new values, for example M and F with male and female.
7 Which question do you ask first about an outlier?	Is it an error or a rare but valid case? The answer decides whether you remove or keep it.
8 How do you get around zeros in a log transformation?	With log(x + 1), provided you do it deliberately and document it.
9 What does all.x = TRUE do in merge()?	All rows of the first dataset are kept, even without a match; missing values from the second become NA.
10 What does mutate() do?	It creates new columns or changes existing ones, for example mutate(price_k = price / 1000).

Question	Answer
11 What is binning?	Dividing a numeric variable into classes, for example with cut(), such as temperature into cold, mild and warm.
12 What is the difference between long and wide data?	Wide has one column per variable or time point; long has one row per observation with a column for the variable and one for the value.