SAMENVATTING · OP TENTAMENNIVEAU

Module 4

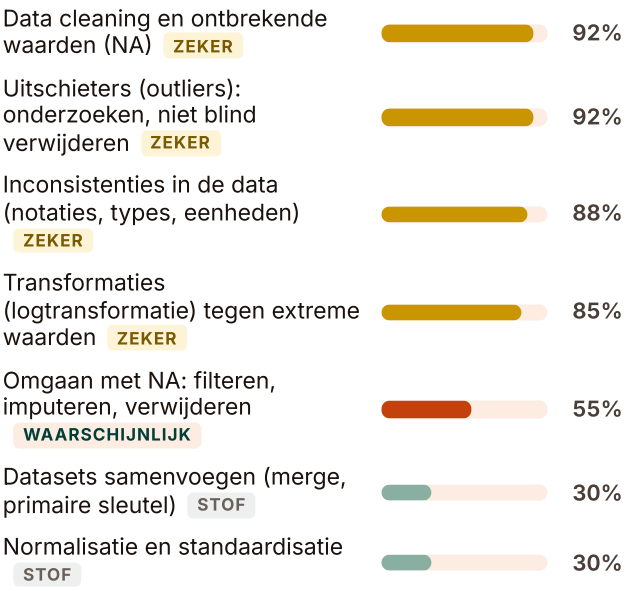
Data cleaning en management

Alles wat je voor de open vragen moet kunnen definiëren, uitleggen, toepassen en beoordelen. Met de modellen als figuur, de vergelijkingen als tabel en de valkuilen gemarkeerd.

WAT ERIN ZIT

01	Doel en plaats in het proces	2
02	Ontbrekende waarden	2
03	Inconsistenties	2
04	Outliers	3
05	Datasets samenvoegen	3
06	Nieuwe variabelen	3
07	Normaliseren en standaardiseren	4
08	De workshop: Divvy Bikes en weer	4
09	Veelgemaakte fouten	4
10	Op het tentamen	4
→	Zo schrijf je het antwoord	5
→	Checklist: kan ik dit uitleggen?	5
→	Overhoor jezelf	6

TENTAMENKANS



Inschatting per onderwerp, op basis van het voorbeeldtentamen en de nadruk in de stof.

KERNBEGRIPPEN

data cleaning · NA · filtering · imputation · deletion · data inconsistencies · factor · outliers

OP HET TENTAMEN

Blok 4

5 punten · 180 min totaal

MC	2 meerkeuzevragen	2.5 punten per vraag
7	Blok 7 (25 punten) bestaat uit drie open vragen over R-output.	200 / 250 / 150 woorden

Tentamen: ma 7 dec 2026 · 180 min · Vaktermen blijven Engels

01 Doel en plaats in het proces

Data cleaning is het opsporen en corrigeren van fouten en inconsistenties om de datakwaliteit (data quality) te verhogen (accuraat, consistent, volledig, betrouwbaar) voor modellering, rapportage en beslissingen. Het is geen standaardproces: de acties hangen van de data af. Het kost veel tijd (naar schatting 60% van het werk van data scientists) en slechte voorbereiding leidt tot slechte beslissingen. Module 4 is een workshopweek; het tentamen test het lezen van R-code en -uitvoer (circa 2 meerkeuzevragen), niet het schrijven ervan.

02 Ontbrekende waarden

Een **missing value** (meervoud: missing values) is een datapunt dat niet is vastgelegd of niet beschikbaar is (in R: **NA**). Oorzaken: verlies tijdens verzamelen of overdracht, invoerfouten of weglatingen, niet-toepasselijke antwoorden (skip-patronen). Gevolgen: bias, minder power, fouten in de analyse.

Aanpak	Wat	Voordeel of risico
Filteren	NA's alleen voor deze analyse weglaten	originele dataset blijft intact, flexibel bij herziening
Imputeren	vervangen door gemiddelde, mediaan of andere schatting	risico op bias; gemiddelde-imputatie verkleint de spreiding
Verwijderen	rijen met NA's schrappen	veilig bij een klein deel; bij veel ontbrekende data vertekening

Nuttige R-handelingen om te herkennen: **is.na(x)**, **sum(is.na(x))** (aantal NA's), **colSums(is.na(df))** (per kolom), **mean(x, na.rm = TRUE)**, **complete.cases(df)** en **na.omit(df)**. Let op het verschil tussen een percentage per kolom en het aantal **complete cases**: een dataset met 10% NA in drie kolommen kan veel meer dan 10% van de rijen verliezen bij **na.omit()**. Bij gemiddelde-imputatie verandert het gemiddelde niet, maar de standaardafwijking daalt (n groter, dezelfde som van kwadraten), en samenhangen worden zwakker.

03 Inconsistenties

Voorbeelden: datumformaten (MM/DD/YYYY tegen DD-MM-YYYY), categorische varianten ("Male", "male", "M"), datatypes (data types: character tegen **factor**), eenheden (euro's en dollars, kilo's en ponden). Aanpak: formaten standaardiseren (datum/tijd, hoofdletters, types), eenheden omrekenen, **altijd op duplicaten controleren** (**duplicated()**, **unique()**), typefouten en foute invoer herstellen met dezelfde structuur.

Typen en factors: **str(df)** toont de types; **as.factor()**, **as.numeric()** en **as.Date()** (of lubridate-functies als **ymd()**) converteren. Een factor heeft **levels**; een variabele met cijfers als tekst moet eerst numeriek worden voor je er een gemiddelde van kunt

nemen. Recoderen kan met **ifelse()**, **cut()** (binning) of dplyr **case_when()**.

04 Outliers

Een outlier is een datapunt dat sterk afwijkt van de rest; het kan wijzen op een fout, een zeldzame gebeurtenis of echte variatie, en kan de analyse vertekenen.

Aanpak: **onderzoeken** (fout of zeldzaam maar geldig?), **verwijderen** als het duidelijk een fout is, **transformeren** (log of een andere functie) bij een scheve verdeling. Niet blind verwijderen: begrijp eerst waarom. Een gangbare detectieregel is de boxplot-regel (buiten 1,5 keer de IQR vanaf het kwartiel) of een z-score; het gekozen criterium moet je kunnen rechtvaardigen en documenteren. Een foutcode zoals -999 voor "onbekend" is geen outlier maar een NA die je moet omzetten.

05 Datasets samenvoegen

Situatie	Functie
zelfde variabelen, meerdere bestanden (maanden): rijen stapelen	rbind()
zelfde eenheden, andere variabelen: kolommen koppelen op een sleutel	merge(a, b, by = "customer_id")
many-to-one, alle transacties behouden	merge(a, b, all.x = TRUE)

Let op: kies een **primary key**, verwijder duplicaten vooraf, laat datatypes overeenkomen (geen mix van integer en character), beslis wat te doen met non-matches (een klant zonder transacties, een transactie van een klant die niet in de lijst staat) en controleer de uitvoer na het samenvoegen (aantal rijen!). Bij **all.x = TRUE** blijven niet-gematchte rijen uit de linkertabel met **NA** in de nieuwe kolommen; bij standaard **merge()** verdwijnen ze (inner join). Een veelgemaakte fout is joinen op een sleutel met duplicaten: het aantal rijen neemt dan toe.

06 Nieuwe variabelen

Soorten: transformatie (**log(income)**, **sqrt(price)**), binning (leeftijd naar kind, volwassene, senior), indicator of flag (**is_high_income = income > 100000**), combinatie (**revenue = price * quantity**), datumdelen (**lubridate::year(order_date)**) en rangorde of score. Regels: **documenteer** nieuwe variabelen, controleer op **NA** en **Inf** na een transformatie (zoals **log(0)**), pas de ruwe kolommen niet direct aan, gebruik duidelijke namen.

Met dplyr gebruik je **mutate()** voor nieuwe kolommen, **filter()** voor rijen, **select()** voor kolommen, **group_by()** met **summarise()** voor samenvattingen en de pipe **%>%** of **|>** om stappen te koppelen. De data kan **long** (een rij per observatie, bijvoorbeeld per uur of per meting) of **wide** (een kolom per meting) zijn; modellen en grafieken verwachten meestal long, samenvattingstabellen zijn vaak wide.

07 Normaliseren en standaardiseren

Reden: modellen die gevoelig zijn voor schaal (K-Means, kNN, PCA, SVM) laten variabelen met een groot bereik domineren; schalen maakt vergelijking eerlijk.

Methode	Formule	Uitkomstbereik
Normalisatie	x' = (x – min) / (max – min)	[0, 1]
Standaardisatie	z = (x – mu) / sigma	onbegrensd

Normalisatie is gevoelig voor outliers (min en max bepalen alles); standaardisatie verandert de vorm van de verdeling niet.

08 De workshop: Divvy Bikes en weer

Taken: dataset laden en verkennen (**str**, **summary**), variabeletypes (factor, tijdobject) controleren, de rijen met **docked_bike** verwijderen, NA's behandelen, outliers in reisduur en afstand aanpakken, nieuwe variabelen maken (duur in minuten of uren, afstand met de Haversine-formule), de weerdata opschonen en uitbreiden (dagelijkse gemiddelden voor temperatuur en windsnelheid, dagelijks totaal voor neerslag), beide datasets samenvoegen op een **uurlijkse tijdstempel** die je in beide maakt, en verkennende (modelvrije) analyses met toetsen en correlaties.

09 LET OP Veelgemaakte fouten

- ✗ Rijen met NA verwijderen zonder te kijken hoeveel het er zijn of waarom ze ontbreken.
- ✗ Foutcodes (-999, 9999) als echte waarden laten staan.
- ✗ Merge zonder controle van rijaantal; duplicaten in de sleutel.
- ✗ Ruwe kolommen overschrijven in plaats van een nieuwe variabele maken.
- ✗ Gemiddelde-imputatie zien als neutraal: het verkleint de spreiding.
- ✗ Outliers automatisch verwijderen.

10 Op het tentamen

Circa 2 meerkeuzevragen over module 4. Geen code schrijven; wel lezen: wat telt **sum(is.na(x))**, welk effect heeft **all.x = TRUE**, welke aanpak past bij deze outlier, wat is het risico van deze imputatie? Het formuleblad is hier nauwelijks relevant; de redenering telt.

VOORBEELDANTWOORD

→ Zo schrijf je het antwoord

Vier bouwstenen, in deze volgorde. Elke steen levert punten op; laat er geen weg.

1. Term
Noem het model of begrip met de Engelse vakterm.

2. Definitie
Geef de definitie of de kenmerken, precies.

3. Casusfeit
Koppel aan een concreet feit uit de casus.

4. Conclusie
Sluit af met een oordeel, advies of gevolg.

OPDRACHT 1 · WOORDLIMIET 75 WOORDEN (WOORDLIMIET AANGENOMEN)

In een dataset ontbreekt bij 4 van de 500 rijen een waarde. Wat doe je, en wat leg je vast?

TERM Eerst inspecteren, dan een bewuste keuze maken voor de missing values. DEFINITIE Je kunt de rijen verwijderen of de waarden imputeren, en je weegt af wat dat doet met de steekproef. CASUSFEIT Het gaat om weinig rijen (4 van 500), dus verwijderen verandert de steekproef nauwelijks, mits de ontbrekende waarden niet systematisch zijn. CONCLUSIE Leg de keuze en het aantal verwijderde rijen vast, zodat de analyse herhaalbaar blijft.

65 / 75 woorden

KERNBEGRIPPEN

✓ Checklist: kan ik dit uitleggen?

Vink af wat je uit je hoofd in één zin kunt uitleggen, met de Engelse term.

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

FLASHCARDS

Overhoor jezelf

Dek de rechterkolom af en zeg het antwoord hardop.

Vraag	Antwoord
1 Wat is data cleaning?	Het opsporen en herstellen van fouten en inconsistenties zodat data accuraat, consistent, compleet en betrouwbaar zijn voor analyse.
2 Wat zijn typische oorzaken van missing values?	Dataverlies bij verzamelen of overdracht, invoerfouten of weglatingen, en niet-van-toepassing-antwoorden (skip patterns).
3 Welke drie manieren zijn er om met missing values om te gaan?	Filteren, vervangen (imputation) of verwijderen (deletion) van de rijen met ontbrekende waarden.
4 Hoe verwijder je rijen met een NA in R?	Met na.omit(data) of drop_na(data); met drop_na(kolom) alleen de rijen met een NA in die kolom.
5 Wat doet as.numeric() op de tekst "n/a"?	Het geeft NA met een waarschuwing, want "n/a" is niet als getal te lezen.
6 Wat doet recode (bijvoorbeeld ifelse) met een categorische variabele?	Het vervangt waarden of groepen door nieuwe waarden, bijvoorbeeld M en F door male en female.
7 Welke vraag stel je eerst bij een outlier?	Is het een fout of een zeldzaam maar geldig geval? Het antwoord bepaalt of je verwijdert of behoudt.
8 Hoe omzeil je nullen bij een log-transformatie?	Met log(x + 1), mits je dat bewust doet en documenteert.
9 Wat doet all.x = TRUE in merge()?	Alle rijen van de eerste dataset blijven behouden, ook zonder match; ontbrekende waarden uit de tweede worden NA.
10 Wat doet mutate()?	Het maakt nieuwe kolommen of past bestaande aan, bijvoorbeeld mutate(prijs_k = prijs / 1000).
11 Wat is binning?	Een numerieke variabele indelen in klassen, bijvoorbeeld met cut(), zoals temperatuur in koud, mild en warm.
12 Wat is het verschil tussen long en wide data?	Wide heeft een kolom per variabele of tijdstip; long heeft een rij per waarneming met een kolom voor de variabele en een voor de waarde.