Data cleaning en management

CHEATSHEET

Tentamen ma 7 dec 2026

TENTAMENKANS Data cleaning en ontbrekende waarden (NA) **92%** · Uitschieters (outliers): onderzoeken, niet blind verwijderen **92%** · Inconsistenties in de data (notaties, types, eenheden) **88%** · Transformaties (logtransformatie) tegen extreme waarden **85%** · Omgaan met NA: filteren, imputeren, verwijderen **55%** · Datasets samenvoegen (merge, primaire sleutel) **30%**

Data cleaning in stappen

1. Inspecteer de data
2. Ontbrekende waarden
3. Uitbijters en fouten
4. Dubbele rijen en typen
5. Leg de keuzes vast

Welke toets bij welke vraag?

Verschil	Verband	Voorspelling
Gemiddelden vergelijken: t-test (2 groepen) of ANOVA (3 of meer).	Twee numerieke variabelen: correlation. Twee categorische: chi-square.	Y voorspellen uit X: regressie (lineair of logistisch).

M 01 Principes

- Cleaning is geen vast recept; ruwe data blijft intact; documenteer elke stap en nieuwe variabele.
- Na elke bewerking controleren: rijen, NA's, Inf, duplicaten.

M 02 Ontbrekende waarden (NA)

- Filteren (alleen voor deze analyse), imputeren (gemiddelde, mediaan: risico op bias), verwijderen (veilig bij klein deel).
- Gemiddelde-imputatie: gemiddelde gelijk, standaardafwijking kleiner.
- R: is.na(x), sum(is.na(x)), colSums(is.na(df)), mean(x, na.rm = TRUE), complete.cases(df), na.omit(df).
- Foutcodes (-999) zijn NA's, geen outliers.

M 03 Inconsistenties

- Datumformaten, tekstvarianten, datatypes, eenheden. Standaardiseren, omrekenen, duplicaten zoeken (duplicated, unique).
- str(df) toont types. as.factor(), as.numeric(), as.Date() of lubridate (ymd).
- Recoderen: ifelse(), cut(), case_when().

M 04 Outliers

- Eerst onderzoeken: fout of zeldzaam maar geldig? Fout: verwijderen of NA. Scheef: log of andere transformatie.
- Detectie: boxplot-regel (1,5 x IQR) of z-score; criterium documenteren.

M 05 Samenvoegen

- rbind(): zelfde variabelen, rijen stapelen.
- merge(a, b, by = "key"): kolommen koppelen. all.x = TRUE: alle rijen van links behouden (NA bij geen match). Standaard: inner join.
- Primary key kiezen, duplicaten eerst weg, datatypes gelijk, rijaantal controleren.

M 06 Nieuwe variabelen

- Transformatie, binning, indicator, combinatie, datumdeel, rang.
- Controleer NA en Inf na transformatie (log(0) = -Inf). Duidelijke namen.
- dplyr: mutate, filter, select, group_by, summarise, pipe. Long (rij per observatie) tegen wide.

M 07 Schalen

- Normalisatie: (x – min) / (max – min), bereik [0, 1], gevoelig voor outliers.
- Standaardisatie: z = (x – mu) / sigma, onbegrensd.
- Nodig voor K-Means, kNN, PCA, SVM.

M 08 Workshop Divvy Bikes

- Laden, types, docked_bike eruit, NA's en outliers, duur en afstand, weerdata opschonen, merge op uurlijkse tijdstempel, verkennende toetsen en correlaties.

M 09 Tentamen

- Circa 2 meerkeuzevragen: code en uitvoer lezen, geen code schrijven.

! 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.