BM21 BUSINESS ANALYTICS · MODULE 4

Data cleaning en management

POSTER · DE MODULE IN ÉÉN BLIK

Tentamen ma 7 dec 2026 · 180 min

1 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

Eerst begrijpen, dan pas opschonen en vastleggen wat je deed.

2 Welke toets bij welke vraag?

Verschil

Gemiddelden vergelijken: t-test (2 groepen) of ANOVA (3 of meer).

Verband

Twee numerieke variabelen: correlation. Twee categorische: chi-square.

Voorspelling

Y voorspellen uit X: regressie (lineair of logistisch).

De vraag en het meetniveau bepalen de toets.

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%

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.

ZO SCHRIJF JE HET ANTWOORD

1. Term
model of begrip, Engelse vakterm

2. Definitie
definitie of kenmerken, precies

3. Casusfeit
concreet feit uit de casus

4. Conclusie
oordeel, advies of gevolg

KERNBEGRIPPEN

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