

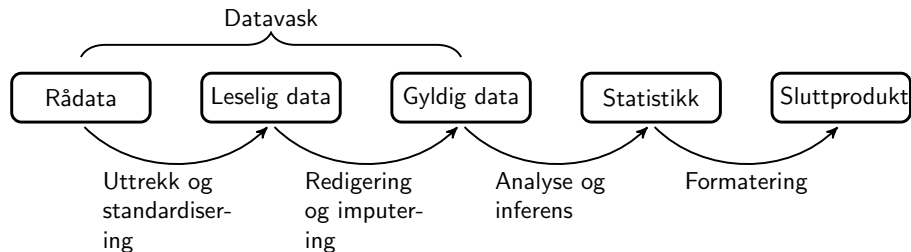
Datavask og datastrukturer - klargjøring av data for analyse

Kevin Thon

SKDE | 

Hva er datavask

- Begrepet datavask har ingen entydig definisjon
 - Typisk: Sett av valideringsregler i en datainnhentingskjede



Hva er datavask

- Denne forelesningen tar (som oftest) utgangspunkt i at man har mottatt et *ferdig* datasett:
 - Hvilke type problemer kan det være med datasett
 - Hvordan identifiserer man mangler
 - Hvordan utbedrer man manglene (hvis mulig)
 - Hvordan skape et analyseklart datasett
- Vil gå gjennom en rekke forhold man må være bevisst
- Vil foreslå noen metoder for sjekk og omforming/vasking av data

- På hvilke måter kan et datasett være være skittent?
 - Mangel!
 - Kan noen ganger bedres ved datavaskmetoder
 - Noen ganger vil konklusjonen være at man ikke kan gjøre ønskede analyser basert på tilgjengelig datasett.
 - Garbage in, garbage out (GIGO)

- Mulige problemer med datasett:
 - Viktig data i form av fritekst (f.eks. fra scannerløsning):
 - Datoer: Ulike formater.
 - Duplikater
 - Ugyldige verdier
 - Tomme felter
 - Misvisende defaultverdier
 - Uhåndterlige variabelnavn
 - Tall som tekst
 - Upraktisk struktur
 - Data levert i flere tabeller, behov for kobling
 - Encoding

Skitne data: Fritekst

- Bør unngås
- Fortsatt svært vanlig å motta data i fritekst
- Utelukker mulighet for valideringsregler ved innregistrering
- Vanlige problemer:
 - Mellomrom før og etter ord
 - Blanding av små og store bokstaver
 - Ved scanning: 1 tolket som l eller l, O som 0, og motsatt
 - Encoding: Lesing og visning av æ,ø,å

Skitne data: Fritekst

Eksempel: Ønsker å lage tabell over diagnoser i et register

- I registeret registreres diagnoser strukturert som ICD-10 koder.
- Ved én avdeling har ikke personvernombud godkjent elektronisk innregistreringsløsning: Data registreres på papir og scannes.
- Vanligste diagnoser:

M54.5	M54.2	M545	M51.1	M54.9	M542	M54.6	M79.1
742	363	85	69	42	41	29	20
m79.1	M48.0	M50.1	M5496	M 79.1	M43.1	M47.2	M549
19	17	16	15	9	9	9	7

Skitne data: Fritekst

- Gjør alle bokstaver store
- Fjern alle mellomrom
- Sett inn ledende M der koden begynner med et tall (Vi vet at det er ryggdiagnoser)
- Erstatt 'l' og 'L' med 1, og O med 0.
- Fjern komma og punktum
- Vanligste diagnoser:

M545	M542	M511	M791	M549	M546	M5496	M480		
827	411	77	57	55	30	22	19		
M501	M431	M472	M419	USPES1F1KKENAKKESMERTER				S32	
19	9	9	6					6	5

Skitne data: Fritekst

Opprinnelig	Antall	Vasket	Antall
M54.5	742	M545	827
M54.2	363	M542	411
M545	85	M511	77
M51.1	69	M791	57
M54.9	42	M549	55
M542	41	M546	30
M54.6	29	M5496	22
M79.1	20	M480	19
m79.1	19	M501	19
M48.0	17	M431	9
M50.1	16	M472	9
M5496	15	M419	6

Skitne data: Datatyper

- Hvilke datatyper/formater som har vært spesifisert ved eksport fra database kan ha betydning ved innlesing:
- Numeriske kolonner kan inneholde enhet ("42%", "42 \$", "42 km", osv.)
- Numeriske kolonner kan ha lokalitetsspesifikke formateringer: "4.2"vs "4,2", eller "4.200,42"vs "4,200.42"
- Løsning: Fiks kolonnene, fjern irrelevant info og sett datatype (numerisk).
- Personnummer vil ofte leses inn som tall og miste ledende nuller: Må spesifiseres som tekstvariabel ved innlesning.

Skitne data: Datoer

- Vær oppmerksom på at datoformatet kan variere, DD.MM.YYYY, YYYY-DD-MM
 - På tvers av datovariabler
 - Innad i variabel - kan kreve betinget manipulering av tekststrenger
- De fleste (alle?) statistikkprogrammer vil la deg spesifisere format ved innlesning
 - Datoer i ikke-standard format vil som regel kunne identifiseres ved at de ikke lar seg lese inn som datovariabler.
 - Noen ganger viktig å ta hensyn til tidssoner, spesielt hvis tidsdifferanser inngår i analysen.
- Vær forsiktig med å åpne og lagre datafil i Excel: Tall kan ha blitt tolket som dato og motsatt
 - Artikkel i Genome Biology fra august 2016 slår fast at: ...
approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Skitne data: Duplikater

- Mange årsaker til duplikater
 - Innleggelse over år - opptrer i flere årssammenstillinger
 - Pasient kan være flyttet i løpet av forløpet - opptrer for flere sykehus/enheter
- Kan i noen tilfeller være enkelt å identifisere:
 - Samme unike pasient- eller forløps-ID
 - samme innleggelses/utskrivningsdato
- Potensielt dilemma: Har identifisert dobbeltregistrering, men med avvikende verdier i enkelte variabler: Hva beholder man?

Skitne data: Ugyldige verdier

- Det anbefales å innføre logiske sjekker av datamaterialet
 - Utskrivelsesdato må være etter innleggsdato
 - All behandling etter dødsdato bør betviles
 - Registreringer basert på kodeverk (ICD-10, NCSP, osv.): Er det gyldige koder?
 - Konsulter med fagfolk hvis mulig
- Dersom man har dokumentasjon på variabelsettet:
 - Maksimum og minimumsverdier
 - Svaralternativer i samsvar med kodebok

Skitne data: Uhåndterlige variabelnavn

- I en del tilfeller vil variabelnavn i utlevert datasett være vanskelig/umulig å jobbe med
- Noen programmer har begrensning på antall tegn
- Eksempel: *32. I løpet av de siste fire ukene, hvor mye av tiden har den fysiske helsen din eller følelsesmessige problemer påvirket dine sosiale aktiviteter (som å besøke venner, slektninger osv)?*
- Innlest i R:

```
X32..I.løpet.av.de.siste.fire.ukene..hvor.mye.av.tiden.har.den.fysiske.helsen.din.eller.følelsesmessige.problemer.påvirket.dine.sosiale.aktiviteter
```

- Nødvendig å omnavne variablene: Enkelt i de fleste statistikkprogram.
- Potensielt problematisk hvis encoding er ukjent
- Innlest i R uten å spesifisere UTF-8 (med BOM):

```
X32..I.IÃ.pet.av.de.siste.fire.ukene..hvor.mye.av.tiden.har.den.fysiske.helsen.din.eller.fÃ.lelsesmessige.problemer.pÃ.virket.dine
```

Skitne data: Upraktiske variabelverdier/kategorier

- Kategoriske variabler kan leveres med kodeverdi eller kategorinavn
 - Eks.: kodebok sier at variabel har verdiene 1 = Dårlig, ..., 5 = Svært bra
 - Faktiske verdier: Daarlig til SvaertBra
 - Kodebok og datadump må stemme overens!
- Mer alvorlig tilfelle: Samme kode kan ha forskjellig betydning for registreringer ved forskjellige tider.
- Kode 1 går fra å bety *mann* til å bety *kvinne*
- Håndterbart hvis dokumentasjonen er god.

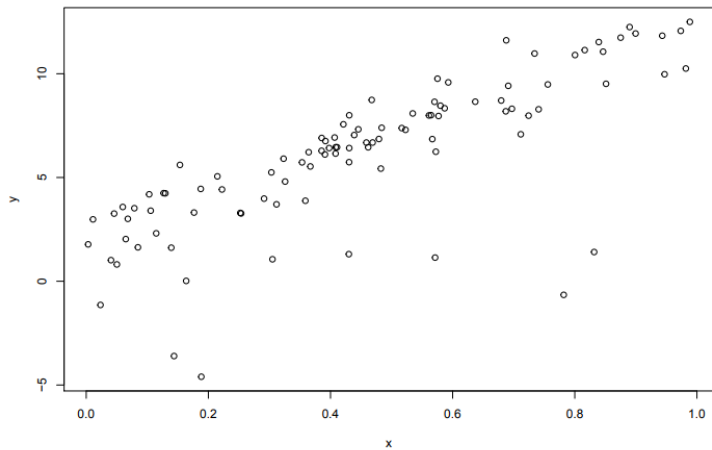
Skitne data: Feil gjennom defaultverdier

- Registerdata er ofte registrert i en elektronisk innregistreringsløsning
- Viktig å vite hvordan ikke-besvarte variabler er representert
- Eksempel fra norsk register: Beregning av gjennomsnittsverdi
 - Opprinnelig var ikke-besvart representert ved tomme felter
 - mean-funksjonen i R lar deg angi at tomme felter ignoreres ved beregning
 - Ved oppdatering av kjerneversjon i innregistreringsløsning ble ikke-besvart erstattet med verdien -1.
 - Alle beregninger basert på variabelen feilet
- Er det mulig å skille *Ikke besvart* fra *Skal ikke besvares*?

Generelle råd

- Se på data!
- Bruk ditt foretrukne verktøy
 - Sorter
 - Gjør utvalg
 - Sammenlign
 - Plot data
 - Lag krysstabeller

Generelle råd: Plot data



Generelle råd: bruk tilgjengelige verktøy

- Enkle rutiner for datasjekk

```
##      isl_1          vekt          blod1
##  Min.    :298.0   Min.    :  -4.00   Min.    :107
##  1st Qu.:300.0   1st Qu.:  59.75   1st Qu.:109
##  Median :300.0   Median :   74.00   Median :110
##  Mean   :300.1   Mean   : 100.94   Mean   :110
##  3rd Qu.:301.0   3rd Qu.:   96.00   3rd Qu.:111
##  Max.   :303.0   Max.   :2444.00   Max.   :112
##  NA's   :13
```

Omstrukturering: Kobling av data

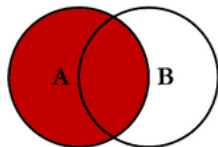
- En database vil typisk bestå av flere tabeller, f.eks.:
 - Pasientinfo
 - Behandlerskjema/intervensjon
 - Oppfølging
- Som regel ønsker registerfolk/forskere data i én fil - kobling overlates til IT leverandør
- Ikke alltid praktisk mulig f.eks. for kronikerregister med vilkårlig antall oppfølginger
- De fleste statistikkprogrammer har funksjoner for kobling av data

Omstrukturering: Kobling av data

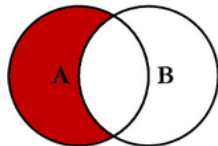
- Hver tabell vil ha én eller flere koblingsnøkler
 - Personnummer
 - Pasient-ID
 - Foløps-ID
- Ved en kobling matcher man rader med samme koblingsnøkkel
- Forskjellige måter å koble data:
 - Inner join: Behold rader med koblingsnøkkel i begge tabeller
 - Left outer join: Behold alle rader i venstre tabell, de som ikke har matchende koblingsnøkkel i høyre tabell blir tomme
 - Full outer join: Behold alle rader, ikke-matchede (venstre eller høyre) blir tomme
 - Right outer join: Behold alle rader i høyre tabell, de som ikke har matchende koblingsnøkkel i venstre tabell blir tomme

Omstrukturering: Kobling av data

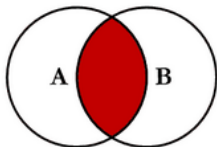
SQL JOINS



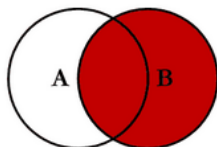
```
SELECT <select_list>  
FROM TableA A  
LEFT JOIN TableB B  
ON A.Key = B.Key
```



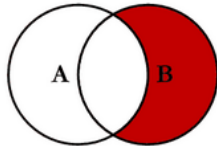
```
SELECT <select_list>  
FROM TableA A  
LEFT JOIN TableB B  
ON A.Key = B.Key  
WHERE B.Key IS NULL
```



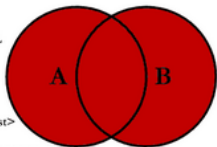
```
SELECT <select_list>  
FROM TableA A  
INNER JOIN TableB B  
ON A.Key = B.Key
```



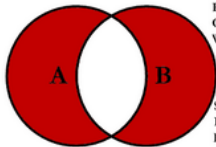
```
SELECT <select_list>  
FROM TableA A  
RIGHT JOIN TableB B  
ON A.Key = B.Key
```



```
SELECT <select_list>  
FROM TableA A  
RIGHT JOIN TableB B  
ON A.Key = B.Key  
WHERE A.Key IS NULL
```



```
SELECT <select_list>  
FROM TableA A  
FULL OUTER JOIN TableB B  
ON A.Key = B.Key
```



```
SELECT <select_list>  
FROM TableA A  
FULL OUTER JOIN TableB B  
ON A.Key = B.Key
```

Omstrukturering

- Ofte vil strukturen i en database være designet/optimalisert for lagring
- Rasjonell struktur for lagring trenger ikke være rasjonell for analyse
- Hadley Wickham presenterer noen prinsipper for ryddig (tidy) data
 - Hver variabel utgjør en kolonne
 - Hver observasjon utgjør en rad
- En observasjon inneholder alle verdier målt på samme enhet (f.eks person, operasjon, dag) på tvers av attributter.
- En variabel inneholder alle verdier som måler samme underliggende attributt på tvers av enheter. Eksempler: Høyde, blodtrykk, O2-opptak.

Omstrukturering

- Hva som er observasjonsenhet påvirker hva som er variabler
- I repeterte målinger (repeated measures):
 - Wide format: Én rad per subjekt (f.eks. individ, land, osv.), én kolonne per respons
 - Long format: Hver rad er ett tidspunkt per subjekt

Omstrukturering

Wide:

id	trt	T1	T2	T3	T4
1	treatment	0.0851360	0.6158293	0.1135090	0.0519033
2	control	0.2254366	0.4296715	0.5959253	0.2641777
3	control	0.2745305	0.6516557	0.3580500	0.3987907
4	treatment	0.2723051	0.5677378	0.4288094	0.8361341

Omstrukturering

Long:

id	trt	key	value
1	treatment	T1	0.0851360
2	control	T1	0.2254366
3	control	T1	0.2745305
4	treatment	T1	0.2723051
1	treatment	T2	0.6158293
2	control	T2	0.4296715
3	control	T2	0.6516557
4	treatment	T2	0.5677378
1	treatment	T3	0.1135090
2	control	T3	0.5959253
3	control	T3	0.3580500
4	treatment	T3	0.4288094
1	treatment	T4	0.0519033
2	control	T4	0.2641777

- Mange måter data kan være rotete
 - Kolonnenavn kan være verdier (ikke variabelnavn)
 - Flere variabler kan være lagret i én kolonne
 - Variabler kan være lagret i både rader og kolonner

Omstrukturering

- Finnes mange verktøy for rydding av rotete data
- To R-pakker er velegnet, *tidyr* og *reshape2*, SPSS og andre statistikkprogrammer vil ha tilsvarende.
- Med tre funksjoner kan man rydde i de fleste datasett
 - `tidyr::pivot_longer` el. `tidyr::gather` (`reshape2::melt`), konverterer fra wide til long
 - `tidyr::separate` (`reshape2::colsplit`), deler tekstvariabel i flere kolonner
 - `tidyr::pivot_wider` el. `tidyr::spread` (`reshape2::dcast`), konverterer fra long til wide

Omstrukturering

Kolonnenavn er verdier:

religion	<\$10k	\$10-20k	\$20-30k	\$30-40k
Agnostic	27	34	60	81
Atheist	12	27	37	52
Buddhist	27	21	30	34
Catholic	418	617	732	670
Don't know/refused	15	14	15	11
Evangelical Prot	575	869	1064	982
Hindu	1	9	7	9
Historically Black Prot	228	244	236	238
Jehovah's Witness	20	27	24	24
Jewish	19	19	25	25

Omstrukturering

Ryddig form:

religion	income	freq
Agnostic	<\$10k	27
Agnostic	\$10-20k	34
Agnostic	\$20-30k	60
Agnostic	\$30-40k	81
Agnostic	\$40-50k	76
Agnostic	\$50-75k	137
Agnostic	\$75-100k	122
Agnostic	\$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96
Atheist	<\$10k	12
Atheist	\$10-20k	27
Atheist	\$20-30k	37
Atheist	\$30-40k	50

Datavask

Omstrukturering

Flere variabler i én kolonne:

	country	year	m014	m1524	m2534	m3544	m4554
11	AD	2000	0	0	1	0	0
37	AE	2000	2	4	4	6	5
61	AF	2000	52	228	183	149	129
88	AG	2000	0	0	0	0	0
137	AL	2000	2	19	21	14	24
166	AM	2000	2	152	130	131	63
179	AN	2000	0	0	1	2	0
208	AO	2000	186	999	1003	912	482
237	AR	2000	97	278	594	402	419
266	AS	2000	-	-	-	-	1

Omstrukturering

Steg 1: gather():

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AF	2000	m65	10

Omstrukturering

Steg 2: separate():

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AF	2000	m	65+	10

Omstrukturering

Norsk kvalitetsregister:

KL	KM	KN	KO	KP	KQ	KR	KS	KT
KSNITT	KSNITT_F	KOMPL_IF	KOMPLIKASJONER	VANNAVC	ABRUPTI	PLACENT	RUPTUR	BLODNIN
			F_81_kp_27;					3
2	1		F_F34_fv_31;F_P39_pl_29;F_P75_pl_29;F_13_kp_27;					
			F_F34_fv_31;F_P39_pl_29;F_P75_pl_29;F_13_kp_27;F_85_kp_27;					3
			F_26_kp_27;					
			F_F34_fv_31;F_P39_pl_29;F_P75_pl_29;F_26_kp_27;					
			F_81_kp_27;					3
2	1		F_F34_fv_31;F_13_kp_27;					
2	1		F_4_fl_23;F_58_kp_27;F_67_kp_27;					
2	1		F_P12_pl_29;F_P46_pl_29;F_5_fl_23;F_58_kp_27;					
			F_18_kp_27;F_26_kp_27;F_81_kp_27;					3
		1	F_F34_fv_31;F_P39_pl_29;F_P75_pl_29;					
2	1		F_24_kp_27;					
			F_12_kp_27;					
			F_22_kp_27;F_26_kp_27;					
2	1		F_P46_pl_29;F_24_kp_27;					
			F_F11_fv_31;F_P42_pl_29;	2				
2			F_XC35_kp_34;F_XC42_kp_34;F_XC59_kp_34;		1			
2	1		F_F34_fv_31;F_13_kp_27;					
			F_F34_fv_31;					
2	1		F_P57_pl_29;F_41_kp_27;F_5_fl_23;F_58_kp_27;		1			
			F_5_fl_23;F_58_kp_27;					
			F_F34_fv_31;F_P46_pl_29;F_13_kp_27;					
2	1		F_95C_kp_27;					
			F_P46_pl_29;F_P73_pl_29;F_13_kp_27;F_45_kp_27;F_64_kp_27;					
2	1		F_13_kp_27;F_81_kp_27;					3
2	1		F_13_kp_27;					

Omstrukturering

Variabler lagret i både rader og kolonner:

id	year	month	element	d1	d2	d3	d4	d5
MX17004	2010	1	tmax	-	-	-	-	-
MX17004	2010	1	tmin	-	-	-	-	-
MX17004	2010	2	tmax	-	27.3	24.1	-	-
MX17004	2010	2	tmin	-	14.4	14.4	-	-
MX17004	2010	3	tmax	-	-	-	-	32.1
MX17004	2010	3	tmin	-	-	-	-	14.2
MX17004	2010	4	tmax	-	-	-	-	-
MX17004	2010	4	tmin	-	-	-	-	-
MX17004	2010	5	tmax	-	-	-	-	-
MX17004	2010	5	tmin	-	-	-	-	-

Omstrukturering

Steg 1: gather():

id	date	element	value
MX17004	2010-01-30	tmax	27.8
MX17004	2010-01-30	tmin	14.5
MX17004	2010-02-02	tmax	27.3
MX17004	2010-02-02	tmin	14.4
MX17004	2010-02-03	tmax	24.1
MX17004	2010-02-03	tmin	14.4
MX17004	2010-02-11	tmax	29.7
MX17004	2010-02-11	tmin	13.4
MX17004	2010-02-23	tmax	29.9
MX17004	2010-02-23	tmin	10.7

Omstrukturering

Steg 2: spread():

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

Oppsummering

- Datakvalitet er viktig!
- Mye kan gjøres med datasett i etterkant, men GIGO
- Må stille krav til dokumentasjon
 - betydning av variabler
 - tilatte verdier
 - formater
 - endringer
- Ulike analyser kan kreve ulik strukturering av data. Bør kunne omstrukturere datasett.
- Vær bevisst mulige problemer!