Datavask og ryddige regneark

Aili Sarre og Katie Smart

Universitetsbiblioteket

13. feb. 2023

Image: Evita L. Bergstad

Agenda (60 min)

1. Beste praksis for å organisere data i regneark
   - Øvelse: finn feilene
   - Tidy Data Principles
2. Datavask og kvalitetskontroll i Excel
3. Andre verktøy

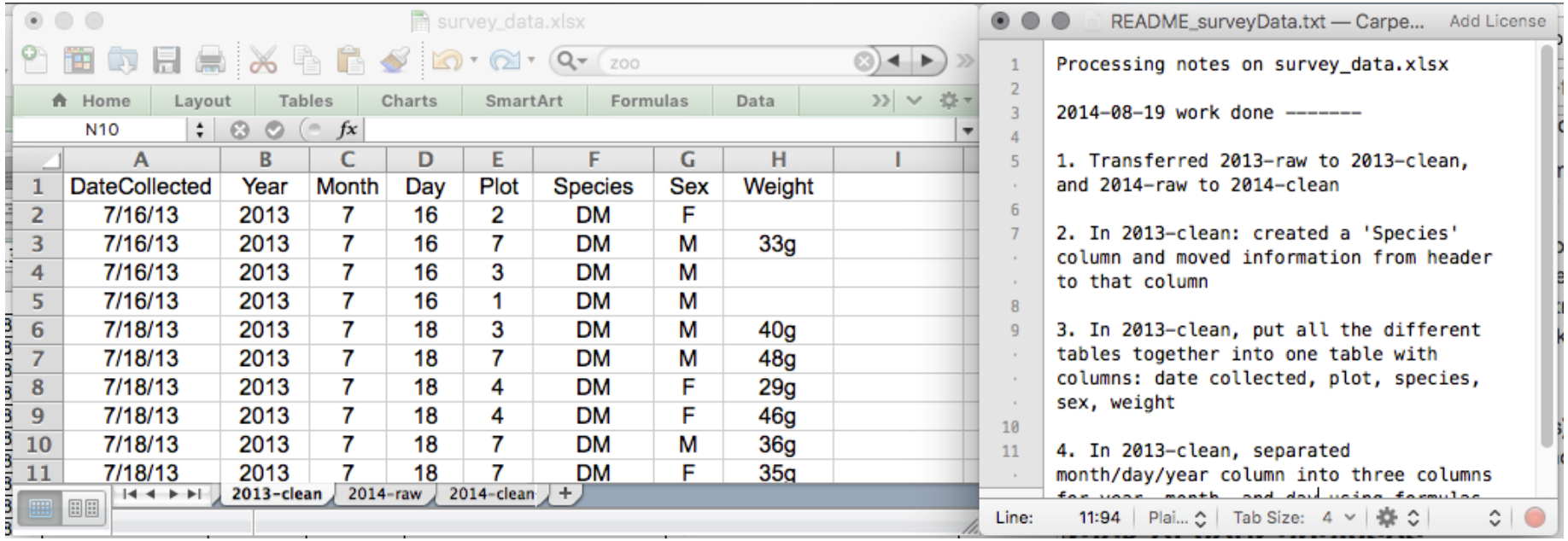
Organisering av data i regneark

Mål

- Dataene skal være forståelig for deg selv i fremtiden.
- Dataene skal være forståelig for andre.
  (R i FAIR prinsippene)
- Dataene skal være maskinlesbare
  (I i FAIR prinsippene).

Tommelfingerregel: Gjør aldri endringer i rådata-filene!

- Kopier filene før du gjør endringer.
- Ikke inkluder formler og kalkuleringer i rådata.
- Sikkerhetskopier
- Loggfør alle trinnene i databehandlingen i en ren tekstfil (README).



	A	B	C	D	E	F	G	H	I
1	DateCollected	Year	Month	Day	Plot	Species	Sex	Weight	
2	7/16/13	2013	7	16	2	DM	F		
3	7/16/13	2013	7	16	7	DM	M	33g	
4	7/16/13	2013	7	16	3	DM	M		
5	7/16/13	2013	7	16	1	DM	M		
6	7/18/13	2013	7	18	3	DM	M	40g	
7	7/18/13	2013	7	18	7	DM	M	48g	
8	7/18/13	2013	7	18	4	DM	F	29g	
9	7/18/13	2013	7	18	4	DM	F	46g	
10	7/18/13	2013	7	18	7	DM	M	36g	
11	7/18/13	2013	7	18	7	DM	F	35g	

Source: Data carpentry

Beskriv dataene i en README fil

- All nødvendig informasjon for at fremtidige deg og andre skal forstå og kunne gjenbruke innholdet i filene.
- Beskriv metodene for datainnsamling og prosessering.
- Definer kolonnevariabler, forkortelser og enheter.
- Loggfør alle endringer
- Lagre README fila nær datasettet det beskriver.

Explanation of column headings used in file Icecream_sales_2020
Column A contains temperature measuresments in degrees Celsius
Column B contains date in YYYY-MM-DD
Column C contains money earned per day in NOK (Norwegian kroner)
Column D contains …

Praktisk øvelse – break out rom

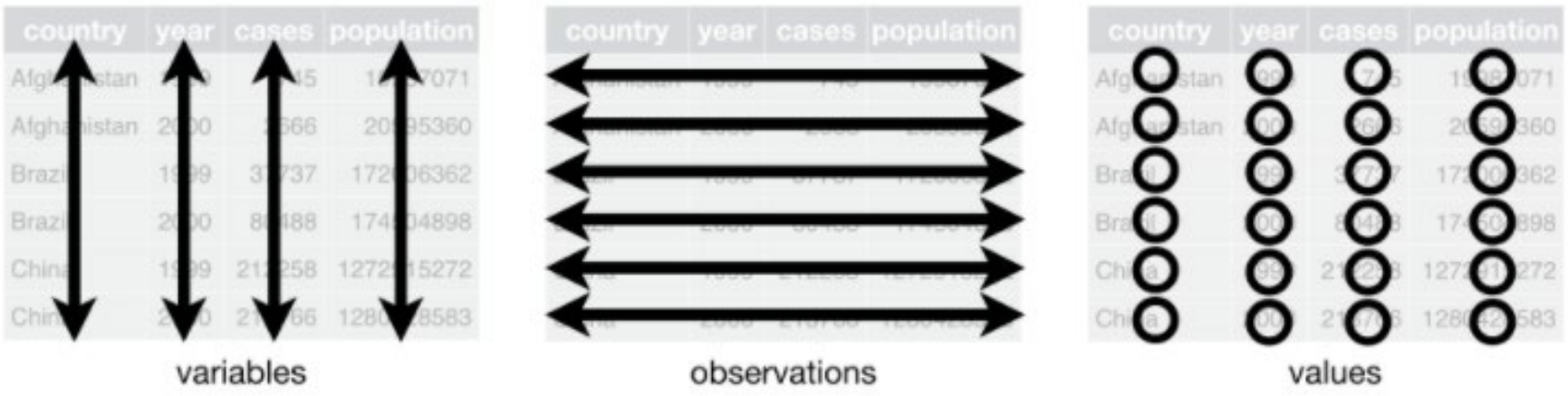
Lenke til regneark sendes i chat: survey_data_spreadsheet_messy.xls

- Hva er galt med regnearket?
- Hvor mange feil kan dere finne?
- Hvordan kan de utbedres?

5 minutter

Organiser regnearkene i h.h.t. «Tidy data»-prinsippene

1. Hver variabel må ha sin egen kolonne.
2. Hver observasjon må ha sin egen rad.
3. Hver verdi må ha sin egen celle.



The Touring Way

Kun en verdi per celle:

Date collected	Plot	Species-Sex	Weight
1/9/78	1	DM-M	40
1/9/78	1	DM-F	36
1/9/78	1	DS-F	135
1/20/78	1	DM-F	39
1/20/78	2	DM-M	43
1/20/78	2	DS-F	144
3/13/78	2	DM-F	51
3/13/78	2	DM-F	44
3/13/78	2	DS-F	146

Date collected	Plot	Species	Sex	Weight
1/9/78	1	DM	M	40
1/9/78	1	DM	F	36
1/9/78	1	DS	F	135
1/20/78	1	DM	F	39
1/20/78	2	DM	M	43
1/20/78	2	DS	F	144
3/13/78	2	DM	F	51
3/13/78	2	DM	F	44
3/13/78	2	DS	F	146

Løsning:
Legg til flere kolonner

Source: Data carpentry

Unngå enheter og kommentarer i cellene

18.07.2013	3	M	40g
18.07.2013	7	M	48g
18.07.2013	4	F	29g
18.07.2013	4	F	46g
18.07.2013	7	M	36g

Løsning:
Legg enheten i kolonneoverskriften eller i en egen kolonne.

13.11.2013	17	F	118
13.11.2013	11	F	126
13.11.2013	17	M	132 (scale not calibrated)
13.11.2013	14	F	113 (scale not callibtrated)
13.11.2013	11	F	122
13.11.2013	4	F	107

Løsning:
Legg informasjonen i en ny kolonne

Alternativ: Legg metadata i README fila

Ikke bruk formatering

Plot: 2			
Date collecte	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52
	measurement device not calibrated		

Date collecte	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y
1/8/14	DM	M	48	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	35	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	37	Y
1/8/14	PF	F	7	Y
1/8/14	DM	M	45	Y
1/8/14	OT			
1/8/14	DS	M	157	N
1/8/14	OX			
2/18/14	NA	M	218	N
2/18/14	PF	F	7	Y
2/18/14	DM	M	52	Y

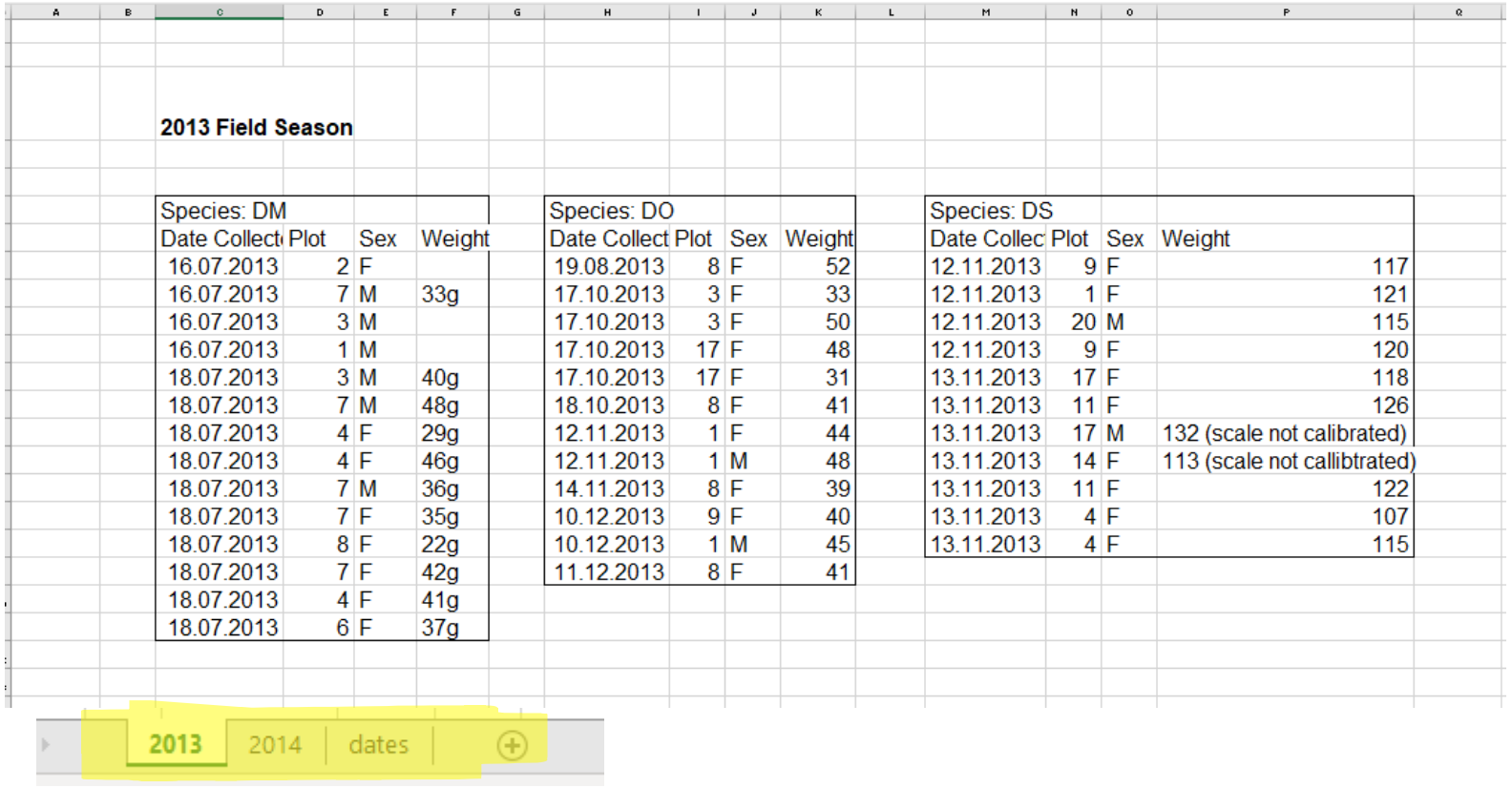
Species: DO			
Date Collect	Plot	Sex	Weight
19.08.2013	8	F	52
17.10.2013	3	F	33

Ikke slå sammen celler

Løsning:
Legg informasjonen i en ny kolonne

Source: Data carpentry

Unngå flere tabeller og ark i samme fil



2013 Field Season

Species: DM			
Date Collecte	Plot	Sex	Weight
16.07.2013	2	F	
16.07.2013	7	M	33g
16.07.2013	3	M	
16.07.2013	1	M	
18.07.2013	3	M	40g
18.07.2013	7	M	48g
18.07.2013	4	F	29g
18.07.2013	4	F	46g
18.07.2013	7	M	36g
18.07.2013	7	F	35g
18.07.2013	8	F	22g
18.07.2013	7	F	42g
18.07.2013	4	F	41g
18.07.2013	6	F	37g

Species: DO			
Date Collect	Plot	Sex	Weight
19.08.2013	8	F	52
17.10.2013	3	F	33
17.10.2013	3	F	50
17.10.2013	17	F	48
17.10.2013	17	F	31
18.10.2013	8	F	41
12.11.2013	1	F	44
12.11.2013	1	M	48
14.11.2013	8	F	39
10.12.2013	9	F	40
10.12.2013	1	M	45
11.12.2013	8	F	41

Species: DS			
Date Collec	Plot	Sex	Weight
12.11.2013	9	F	117
12.11.2013	1	F	121
12.11.2013	20	M	115
12.11.2013	9	F	120
13.11.2013	17	F	118
13.11.2013	11	F	126
13.11.2013	17	M	132 (scale not calibrated)
13.11.2013	14	F	113 (scale not callibtrated)
13.11.2013	11	F	122
13.11.2013	4	F	107
13.11.2013	4	F	115

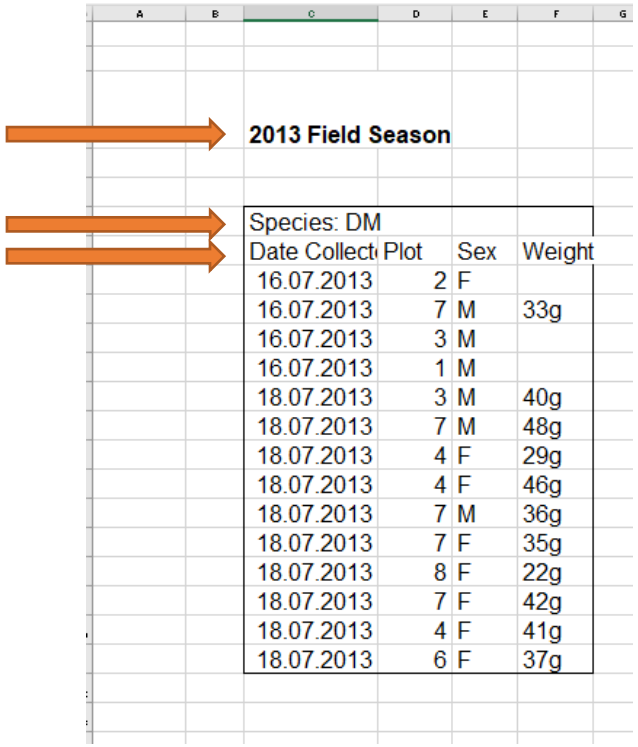
2013 | 2014 | dates

Regnearket skal kun inneholde en tabell med data.

Løsning:
Om mulig, kombiner alt i en tabell, eller lagre hver tabell som separate filer.

Source: Data carpentry

Kolonneoverskrifter



2013 Field Season				
Species: DM				
Date Collecti	Plot		Sex	Weight
16.07.2013		2	F	
16.07.2013		7	M	33g
16.07.2013		3	M	
16.07.2013		1	M	
18.07.2013		3	M	40g
18.07.2013		7	M	48g
18.07.2013		4	F	29g
18.07.2013		4	F	46g
18.07.2013		7	M	36g
18.07.2013		7	F	35g
18.07.2013		8	F	22g
18.07.2013		7	F	42g
18.07.2013		4	F	41g
18.07.2013		6	F	37g

- Bruk kun en tittel-linje!
- Unngå mellomrom og spesialtegn

Source: Data carpentry

Kolonneoverskrifter

	Habitat		
Species	X	Y	Z
A	0	3	0
B	1	0	2

Species	HabitatX	HabitatY	HabitatZ
A	0	3	0
B	1	0	2

Species	Habitat	Abundance
A	Y	3
B	X	1
B	Z	2

Bruk korte beskrivende kolonne-overskrifter

- Unngå spesialtegn i kolonneoverskrifter eller i dataene:

  / \ : * . ? ‘ < > [ ] ( ) & $ æÆ øØ åÅ äÄ

- Ingen mellomrom – bruk_understreking eller CamelCase
- Inkluder enheter

Good Name	Good Alternative	Avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell Type
Observation_01	first_observation	1st Obs

Source: Data carpentry

Bruk konsekvent dato-notasjon

Plot: 2			
Date collecte	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OI		

18.07.2013	3	M	40g
18.07.2013	7	M	48g
18.07.2013	4	F	29g
18.07.2013	4	F	46g
		M	36g

Plot: 3			
Date collecte	Species	Sex	Weight
1/8	PF	M	7
2/18	OT	M	24
2/19	OT	F	23
3/11	NA	M	232
3/11	OT	F	22
3/11	OT	M	26

Anbefalt: Det internasjonale standardformatet for dato YYYY-MM-DD (ISO-8601)

OBS!
Excels standard celleformat

Correspondence | Open Access | Published: 23 June 2004

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg, Joseph Riss, David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barrett & John N Weinstein

BMC Bioinformatics **5**, Article number: 80 (2004) | Cite this article

120k Accesses | **62** Citations | **620** Altmetric | Metrics

Comment | Open Access | Published: 23 August 2016

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Cite this article

142k Accesses | **81** Citations | **3085** Altmetric | Metrics

MICROSOFT \ REPORT \ SCIENCE

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Sometimes it's easier to rewrite genetics than update Excel

By James Vincent | Aug 6, 2020, 8:44am EDT

SHARE

MARCH1 — "Membrane Associated Ring-CH-Type Finger 1

SEPT2 - Septin 2

The Verge

OBS! Bruk av Excels dato-format kan gi problemer

	A	B	C
1	DATE	Number	How it was interpreted
2	Jul-10	40360	1-Jul-10
3	Jul-14	41821	1-Jul-14
4	Jul-15	42186	1-Jul-15
5	Jul-22	44743	1-Jul-22

Excel viser dato i mange ulike formater, men lagrer datoer som nummere.

Alternativer:

- Lagre datoer som en enkeltstreng YYYYMMDD
- Behandle datoer som flere datapunkter (separate kolonner for år, måned og dag)

Datavask

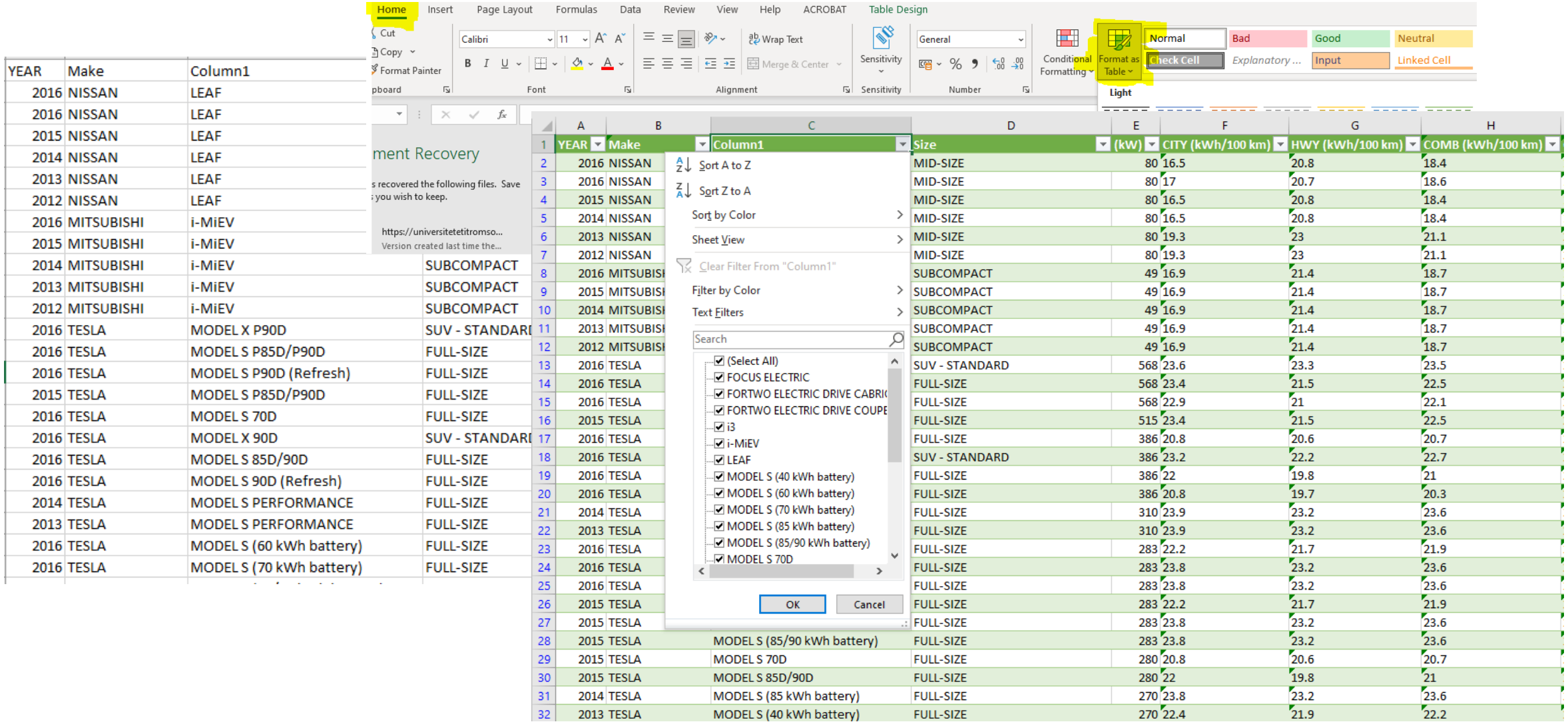
- Sletting av overflødige oppføringer
- Separere eller kombinere verdier i samme felt
- Konverteringer
- Grammatiske feil
- Inkonsistent bruk av navn
- Felt med «NULL» -verdier
- Datoformat

Fjern uønskede observasjoner

- Duplikater
- Irrelevante observasjoner
- Ufullstendige oppføringer
- Ugyldige data
- Motstridende data

Vurder nøye hvorvidt en observasjon skal fjernes!

For å bruke sortering og filtre i Excel, sørg for å bruke **tabulært format**.

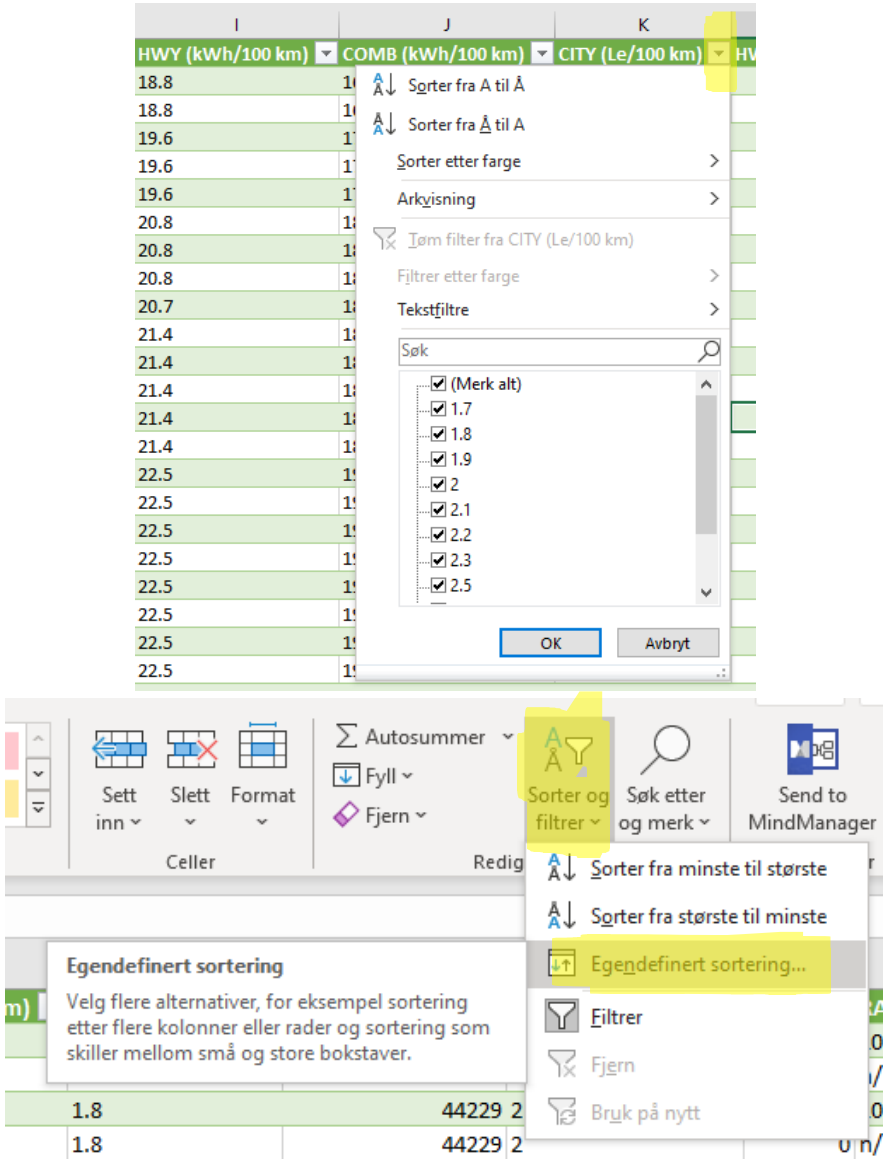


YEAR	Make	Column1	Size	(kW)	CITY (kWh/100 km)	HWY (kWh/100 km)	COMB (kWh/100 km)
2016	NISSAN	LEAF	MID-SIZE	80	16.5	20.8	18.4
2016	NISSAN	LEAF	MID-SIZE	80	17	20.7	18.6
2015	NISSAN	LEAF	MID-SIZE	80	16.5	20.8	18.4
2014	NISSAN	LEAF	MID-SIZE	80	16.5	20.8	18.4
2013	NISSAN	LEAF	MID-SIZE	80	19.3	23	21.1
2012	NISSAN	LEAF	MID-SIZE	80	19.3	23	21.1
2016	MITSUBISHI	i-MiEV	SUBCOMPACT	49	16.9	21.4	18.7
2015	MITSUBISHI	i-MiEV	SUBCOMPACT	49	16.9	21.4	18.7
2014	MITSUBISHI	i-MiEV	SUBCOMPACT	49	16.9	21.4	18.7
2013	MITSUBISHI	i-MiEV	SUBCOMPACT	49	16.9	21.4	18.7
2012	MITSUBISHI	i-MiEV	SUBCOMPACT	49	16.9	21.4	18.7
2016	TESLA	MODEL X P90D	SUV - STANDARD	568	23.6	23.3	23.5
2016	TESLA	MODEL S P85D/P90D	FULL-SIZE	568	23.4	21.5	22.5
2016	TESLA	MODEL S P90D (Refresh)	FULL-SIZE	568	22.9	21	22.1
2015	TESLA	MODEL S P85D/P90D	FULL-SIZE	515	23.4	21.5	22.5
2016	TESLA	MODEL S 70D	FULL-SIZE	386	20.8	20.6	20.7
2016	TESLA	MODEL X 90D	SUV - STANDARD	386	23.2	22.2	22.7
2016	TESLA	MODEL S 85D/90D	FULL-SIZE	386	22	19.8	21
2016	TESLA	MODEL S 90D (Refresh)	FULL-SIZE	386	20.8	19.7	20.3
2014	TESLA	MODEL S PERFORMANCE	FULL-SIZE	310	23.9	23.2	23.6
2013	TESLA	MODEL S PERFORMANCE	FULL-SIZE	310	23.9	23.2	23.6
2016	TESLA	MODEL S (60 kWh battery)	FULL-SIZE	283	22.2	21.7	21.9
2016	TESLA	MODEL S (70 kWh battery)	FULL-SIZE	283	23.8	23.2	23.6
2016	TESLA	MODEL S (40 kWh battery)	FULL-SIZE	283	23.8	23.2	23.6
2015	TESLA	MODEL S (60 kWh battery)	FULL-SIZE	283	22.2	21.7	21.9
2015	TESLA	MODEL S (70 kWh battery)	FULL-SIZE	283	23.8	23.2	23.6
2015	TESLA	MODEL S (85/90 kWh battery)	FULL-SIZE	283	23.8	23.2	23.6
2015	TESLA	MODEL S 70D	FULL-SIZE	280	20.8	20.6	20.7
2015	TESLA	MODEL S 85D/90D	FULL-SIZE	280	22	19.8	21
2014	TESLA	MODEL S (85 kWh battery)	FULL-SIZE	270	23.8	23.2	23.6
2013	TESLA	MODEL S (40 kWh battery)	FULL-SIZE	270	22.4	21.9	22.2

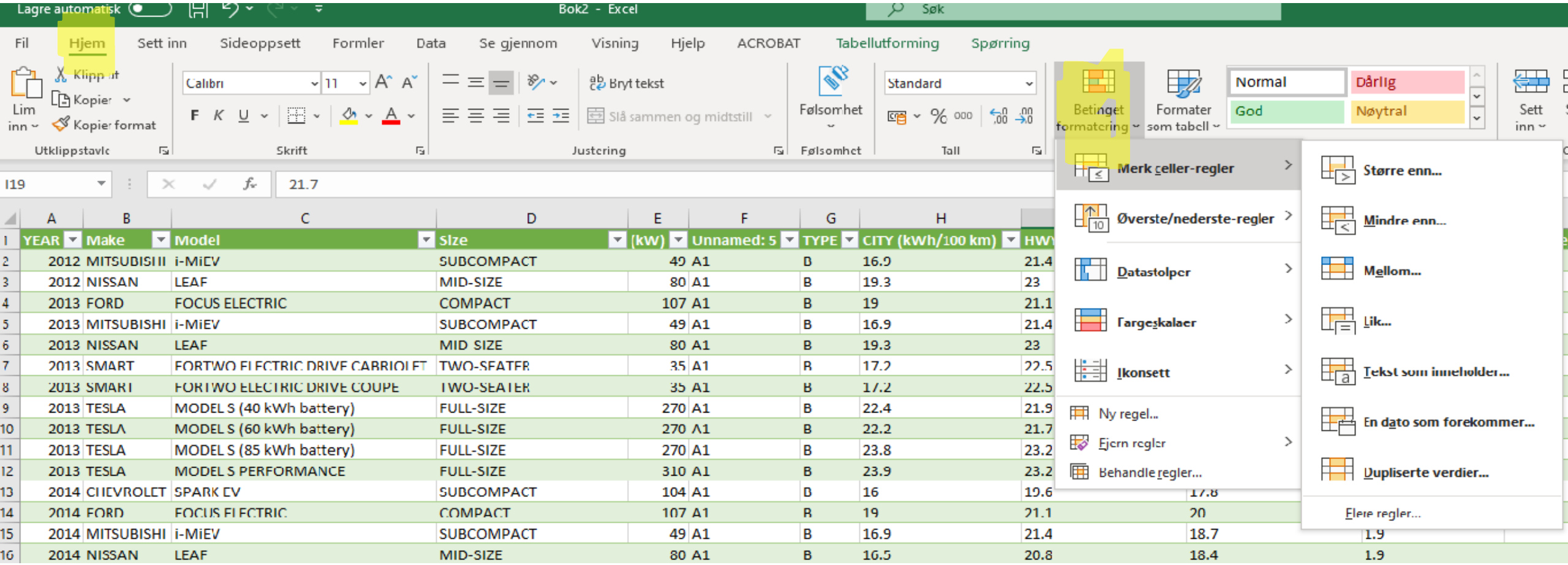
Bruk sortering for å identifisere avvikende verdier

- Avvikende verdier vil oftest havne øverst eller nederst.
- Gå gjennom hele regnearket ved å sortere for hver kolonne og sjekk for ugyldige verdier.
- Ikke bruk sortering dersom regnearket inneholder formler
- Vær obs på at opprinnelig rekkefølge vil ikke kunne gjenopprettes

  - Lagre original-fila før du starter og lag en kopi med passende navn og versjonsnummer.
  - Legg til en kolonne med den originale rekkefølgen



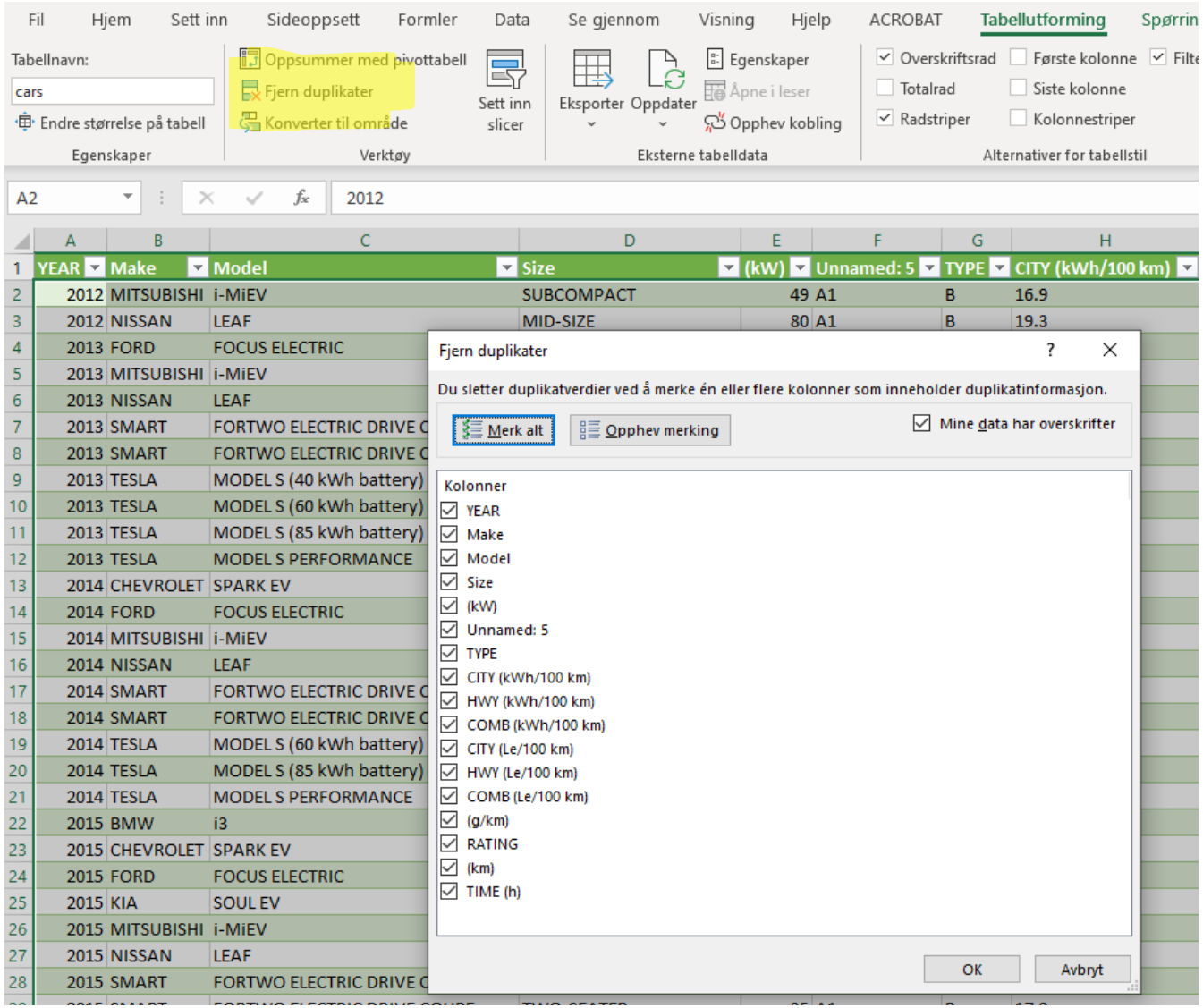
Bruk betinget formatering for å markere avvik



	A	B	C	D	E	F	G	H
1	YEAR	Make	Model	Size	(kW)	Unnamed: 5	TYPE	CITY (kWh/100 km)
2	2012	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9
3	2012	NISSAN	LEAF	MID-SIZE	80	A1	B	19.3
4	2013	FORD	FOCUS ELECTRIC	COMPACT	107	A1	B	19
5	2013	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9
6	2013	NISSAN	LEAF	MID-SIZE	80	A1	B	19.3
7	2013	SMART	FORTWO ELECTRIC DRIVE CABRIOLET	TWO-SEATER	35	A1	B	17.2
8	2013	SMART	FORTWO ELECTRIC DRIVE COUPE	TWO-SEATER	35	A1	B	17.2
9	2013	TESLA	MODEL S (40 kWh battery)	FULL-SIZE	270	A1	B	22.4
10	2013	TESLA	MODEL S (60 kWh battery)	FULL-SIZE	270	A1	B	22.2
11	2013	TESLA	MODEL S (85 kWh battery)	FULL-SIZE	270	A1	B	23.8
12	2013	TESLA	MODEL S PERFORMANCE	FULL-SIZE	310	A1	B	23.9
13	2014	CHEVROLET	SPARK EV	SUBCOMPACT	104	A1	B	16
14	2014	FORD	FOCUS ELECTRIC	COMPACT	107	A1	B	19
15	2014	MITSUBISHI	i-MiEV	SUBCOMPACT	49	A1	B	16.9
16	2014	NISSAN	LEAF	MID-SIZE	80	A1	B	16.5

	HWY		
1			
2	21.4		
3	23		
4	21.1		
5	21.4		
6	23		
7	22.5		
8	22.5		
9	21.9		
10	21.7		
11	23.2		
12	23.2		
13	19.6	17.8	
14	21.1	20	
15	21.4	18.7	1.9
16	20.8	18.4	1.9

Fjern duplikater



Før du fjerner duplikater:
Bekreft at du oppnår resultatene du forventer ved å først filtrer etter unike verdier

-> Sorter og filtrer –> Avansert –> Bare unike poster

Problematiske nullverdier

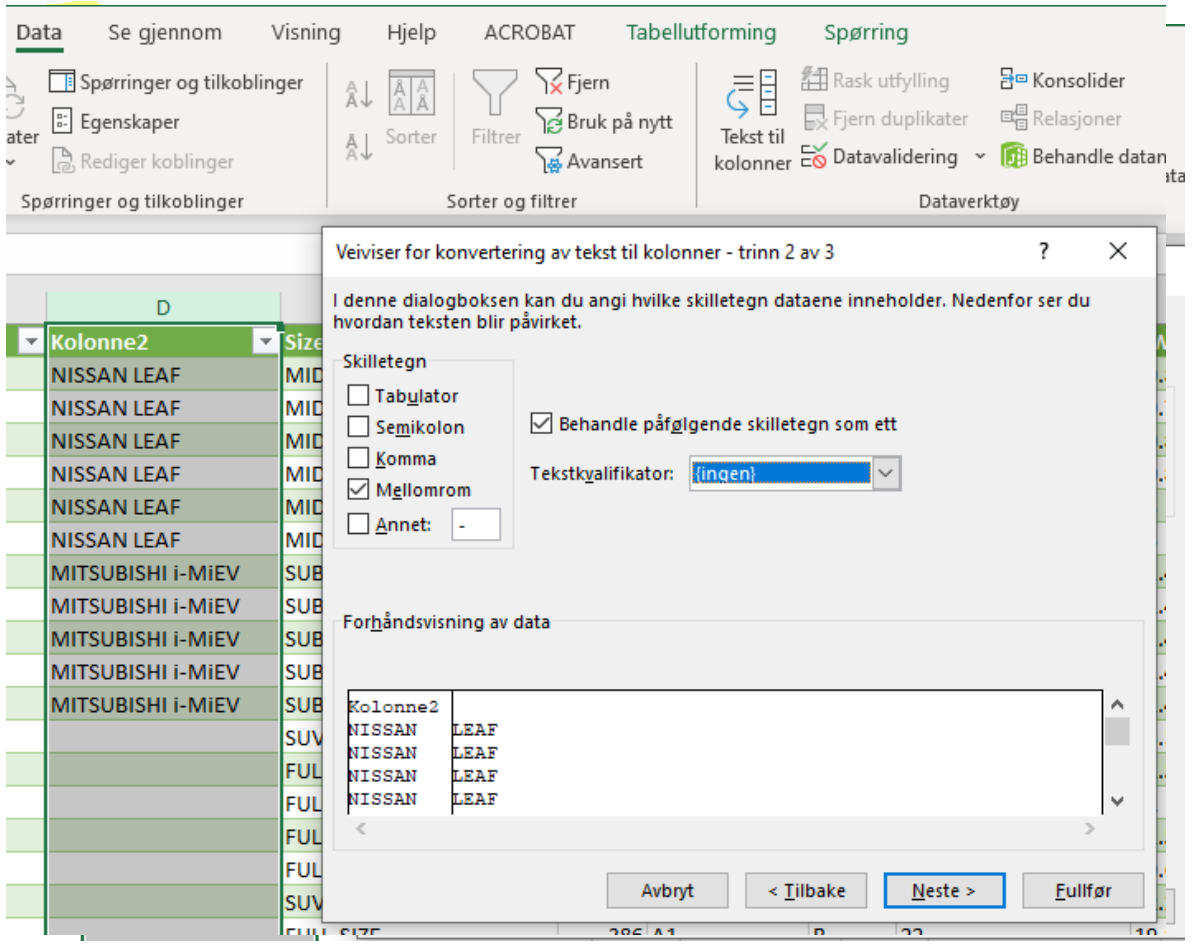
Date collecte	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y
1/8/14	DM	M	48	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	35	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	37	Y
1/8/14	PF	F	7	Y
1/8/14	DM	M	45	Y
1/8/14	OT			
1/8/14	DS	M	157	N
1/8/14	OX			
2/18/14	NA	M	218	N
2/18/14	PF	F	7	Y
2/18/14	DM	M	52	Y

Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

(Nine simple ways to make it easier to (re)use your data, White et al., 2013)

Dele en tekstkolonne til flere kolonner



2016	NISSAN	LEAF	MID-SIZE
2016	NISSAN	LEAF	MID-SIZE
2015	NISSAN	LEAF	MID-SIZE
2014	NISSAN	LEAF	MID-SIZE
2013	NISSAN	LEAF	MID-SIZE
2012	NISSAN	LEAF	MID-SIZE
2016	MITSUBISHI	i-MiEV	SUBCOMPACT
2015	MITSUBISHI	i-MiEV	SUBCOMPACT
2014	MITSUBISHI	i-MiEV	SUBCOMPACT
2013	MITSUBISHI	i-MiEV	SUBCOMPACT
2012	MITSUBISHI	i-MiEV	SUBCOMPACT

Rett opp feil

- Grammatiske feil
- Inkonsekvent bruk av store og små bokstaver
- Inkonsekvent bruk av kolonneoverskrifter. (Bruk standardiserte navn på tvers av datasett.)
- Sjekk for variabelkategorier eller verdier som representerer det samme, og som kan slåes sammen.

1. Gjør oppgaver som ikke krever kolonneredigering først (stavekontroll eller «Søk og Erstatt»).

2. Deretter oppgaver som krever kolonneredigering.

Kolonneredigering trinn for trinn:

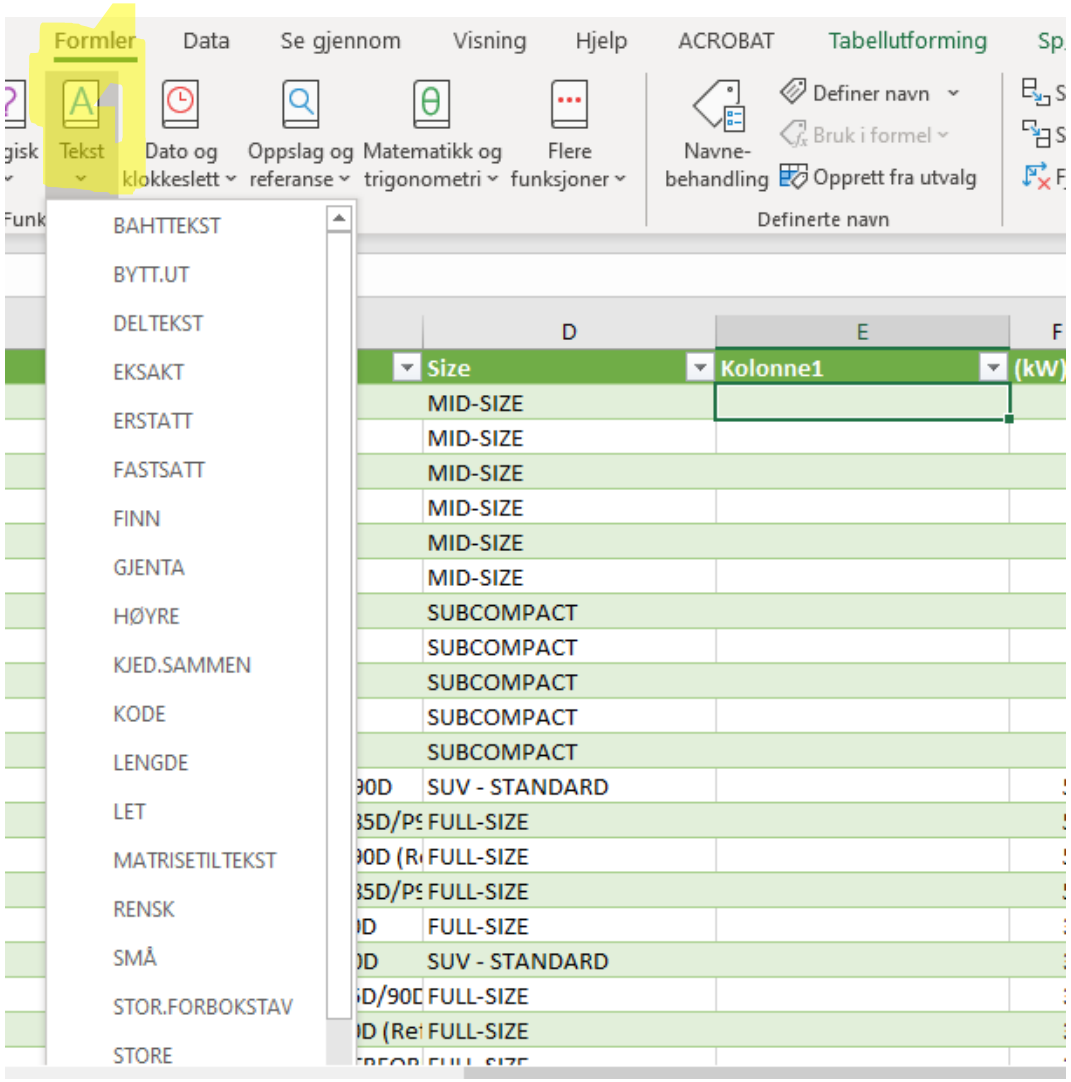
1. Sett inn en ny kolonne (B) ved siden av den opprinnelige kolonnen (A) som må ryddes opp i.

2. Legg til en formel som transformerer dataene fra (A) til den nye kolonnen (B).

3. Velg den nye kolonnen (B), kopier den og lim deretter inn som verdier i den nye kolonnen (B).

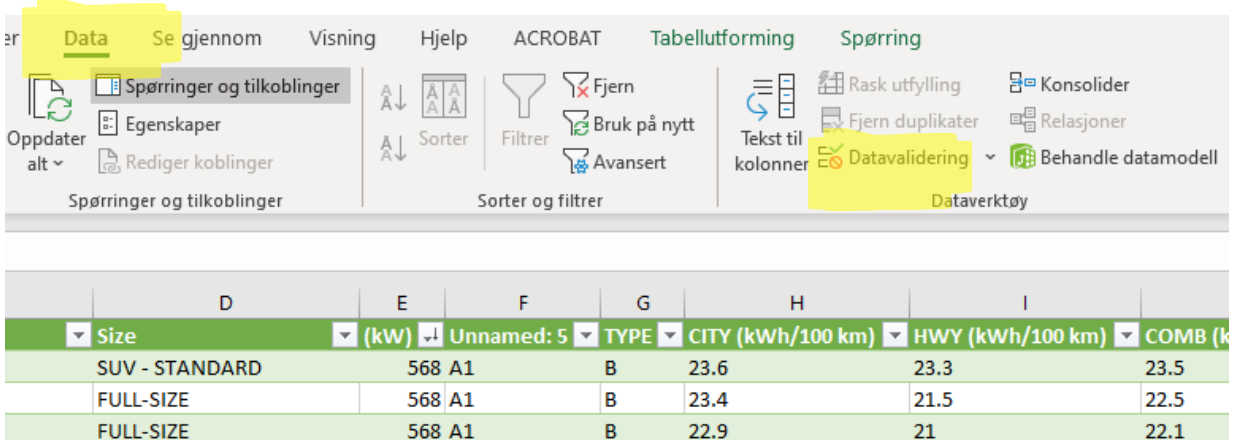
4. Fjern den opprinnelige kolonnen (A), som så gjør at den nye kolonnen konverteres fra B til A.

Bruk formler for redigering av tekst

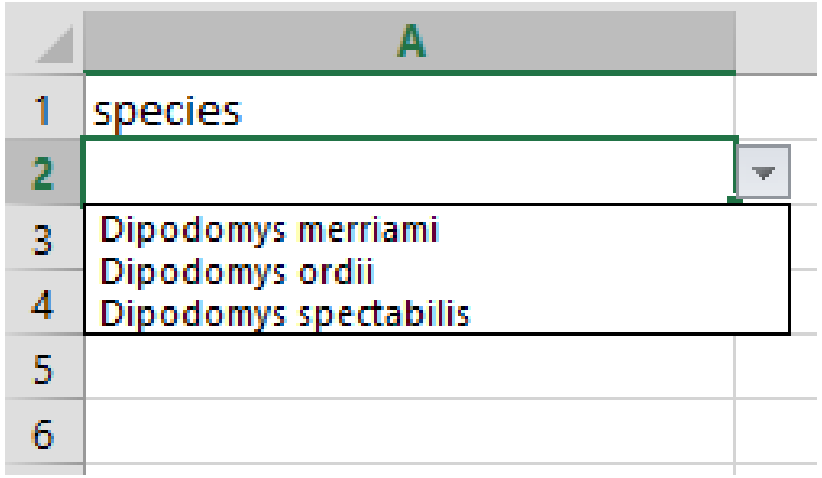


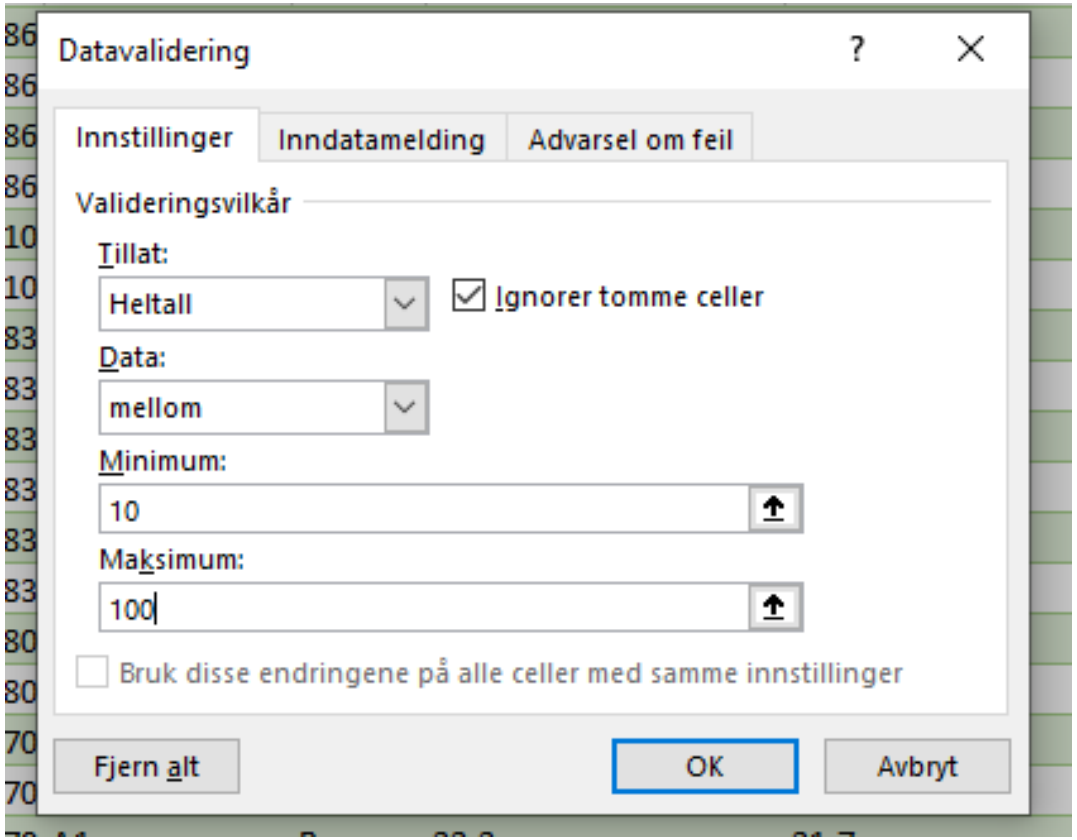
- SMÅ - Endre fra små til store bokstaver
- STORE – Endre fra store til små bokstaver
- TRIMME - Fjerner foranstilte og etterfølgende mellomrom fra teksten
- BYTT.UT - erstatte en bestemt tekst i en tekststreng
- ERSTATT - erstatte tekst som forekommer i en bestemt posisjon i en tekststreng

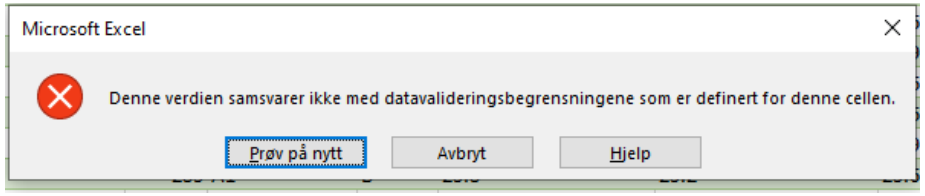
Bruk datavalidering for å unngå å legge inn feilaktige verdier



	D	E	F	G	H	I	
	Size	(kW)	Unnamed: 5	TYPE	CITY (kWh/100 km)	HWY (kWh/100 km)	COMB (k
	SUV - STANDARD	568	A1	B	23.6	23.3	23.5
	FULL-SIZE	568	A1	B	23.4	21.5	22.5
	FULL-SIZE	568	A1	B	22.9	21	22.1







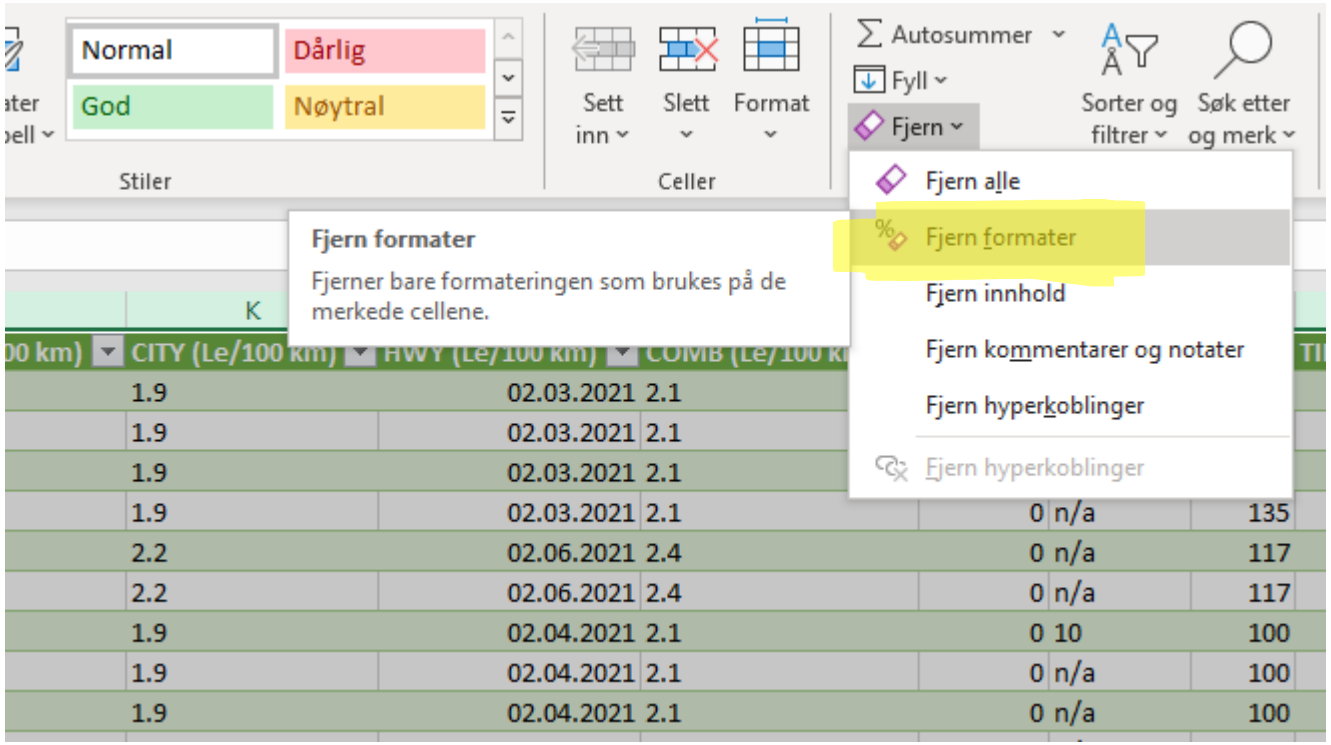
Source: [DataCarpentry]

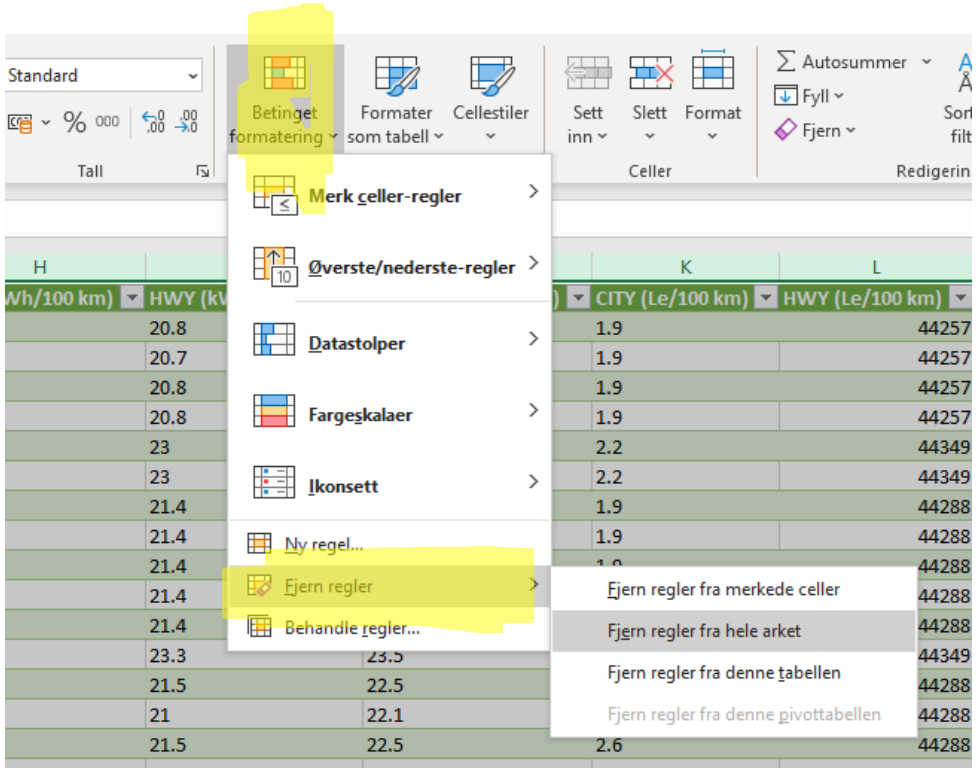
Klargjør dataene for deling og arkivering

- Fjern formateringer
- Eksporter regnearkene til varige formater

Fjern all formatering

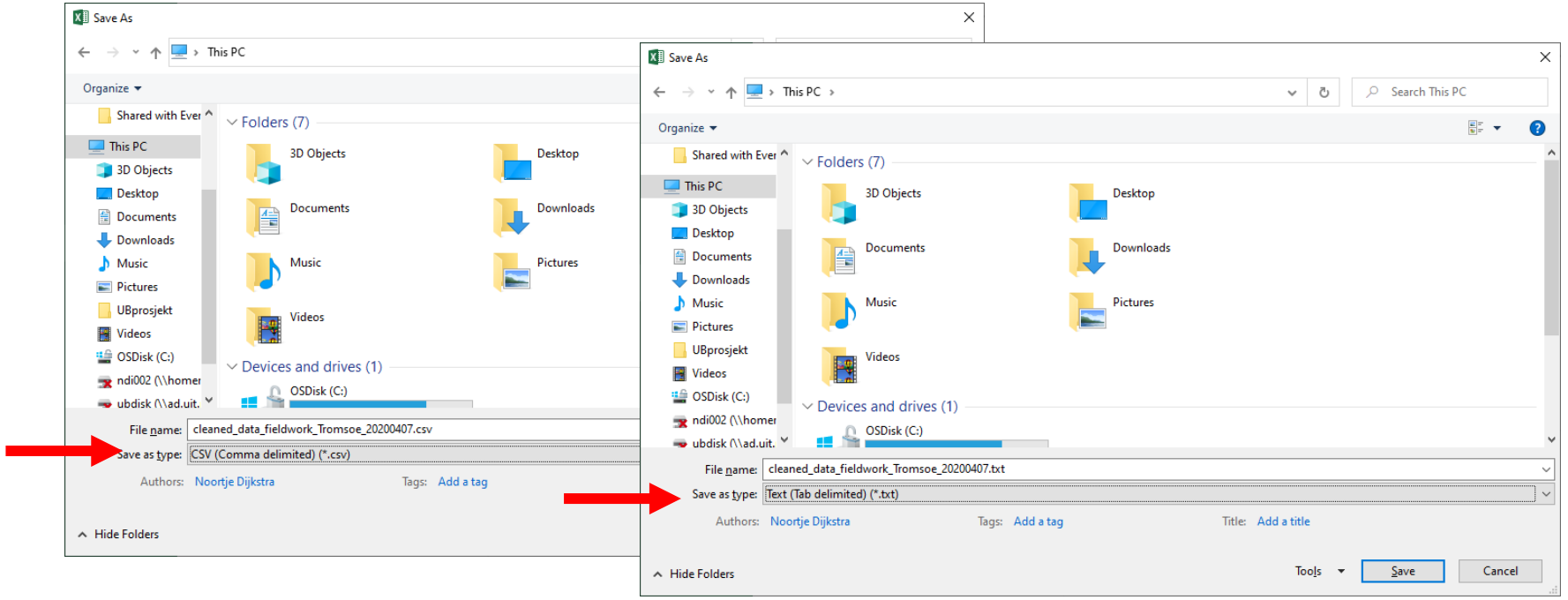
- Også betinget formatering





Eksporter rengjort data til et tekstbasert format

- Lesbart i de fleste analyse programmene (I i FAIR data prinsippene)
  - CSV file (.csv) – OBS! vurder om separatorsymbolet er brukt i celleverdiene (komma, semikolon)
  - TAB delimited (.txt)
  - Husk å eksportere hvert ark separat.



Source: Data carpentry

Etabler rutiner for dine data

Etabler en arbeidsflyt for datavask som passer dine data - og bruk den konsekvent!

- Bruk gjennomtenkte filnavn og versjonskontroll for å holde oversikt over prosessene.
- Dokumenter alle endringer i en README-fil.

Select the line representing the data
Right click the line
Select – Format Data Series
Select – Line Style
Check – the box for Smoothed line

Select the line representing the data
Click again on data point to be edited
Right click data point to be edited
Select – Format Data Point
Select – Marker Options
Select – Built-in
For Type, Select – the circle
For Size, Select – 8

Select – Marker Fill
Select – Solid Fill
Select – Color (the paint bucket)
Select – White
Select – Marker Line Color
Select – Solid Line
Select – Color (the paint bucket)
Select – the same color you used for the main line
Select – Marker Line Style
For Width, increase to 2 pt

Regneark

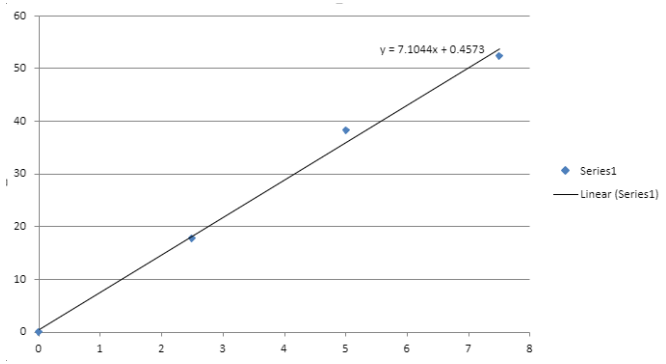
Excel, Google Sheets, LibreOffice

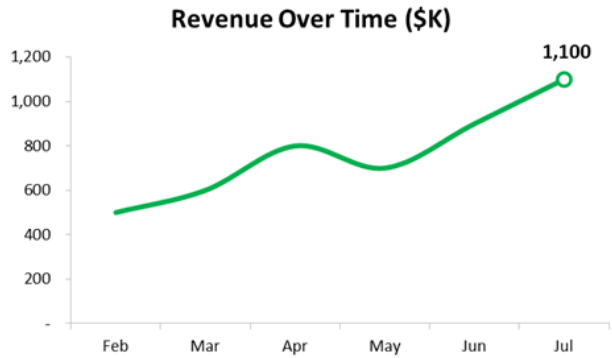
Fordeler:

- Tilgjengelig, lett å bruke, raskt overblikk over data

Ulemper:

- Black box; lite kontroll over prosessene
- Vanskelig å spore endringer eller reprodusere arbeidsflyten





Løsning: Bruk Excel Macros for å gjøre et opptak av arbeidsflyten.

Verktøy for datavask

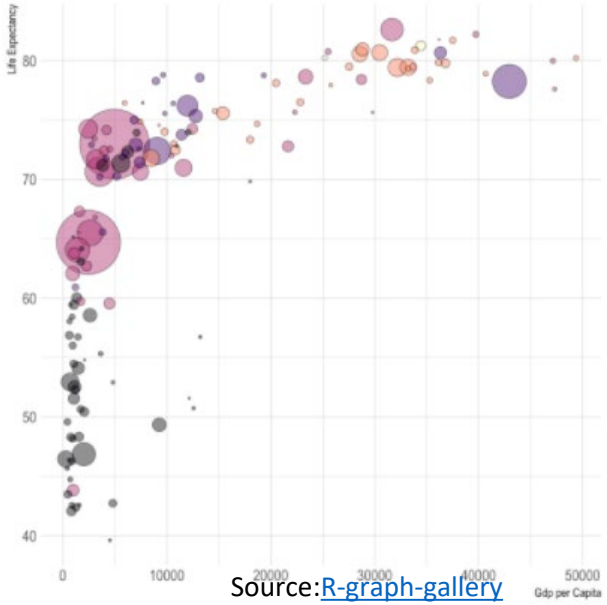
Programmeringsspråkene Python og R

Fordeler:

- Gratis, open-source, fungerer på tvers av platformer (FAIR-principles: Accessible and Interoperable)
- Full kontroll over prosessene
- Lett å spore endringer
- Reproduserbar!
- Lett å finne hjelp på nett (e.g. Stack Overflow, RStudio community, Python community)
- Visualisering av høy kvalitet

Ulempe:

- Må lære programmeringsspråk



Source:R-graph-gallery

Datavask med OpenRefine (grafisk)

- Gir god oversikt over datasettet.
- Lett å identifisere og fikse feil og uregelmessigheter.
- Lett å kombinere data fra flere kilder.
- OpenRefine endrer ikke på originalfila
- Alle handlinger kan lett reverseres
- Alle handlinger loggføres og denne dokumentasjonen kan publiseres sammen med datasettet.
- Arbeidsflyten kan lagres og brukes på flere datasett.

Egen workshop 25. april kl 12-15

Videre selvstudier (Data Carpentries leksjoner)

Kvalitetskontroll og datavask (Data Carpentry leksjon)

- Data Organization in Spreadsheets for Ecologists
- Data Cleaning with OpenRefine for Ecologists
- Data Organization in Spreadsheets for Social Scientists
- OpenRefine for Social Science Data

Data visualisering i R eller Python (Data Carpentry leksjon)

- R for Social Scientists
- Data Analysis and Visualization with Python for Social Scientists
- Data Analysis and Visualization in R for Ecologists
- Data Analysis and Visualization in Python for Ecologists

Webinarer om forskningsdata V2023

Kurs ved UiT er åpne for alle. For mer informasjon og registrering, klikk her.

Datavask og ryddige regneark
13. FEBRUAR 2023

Lagring av forskningsdata
13. FEBRUAR 2023

Hvordan arkivere forskningsdata
14. FEBRUAR 2023

Hvordan arkivere data i UiT Open Research Data
14. FEBRUAR 2023

Hvordan søke og sitere forskningsdata
15. FEBRUAR 2023

Forskningsdata: Rettigheter og lisenser
15. FEBRUAR 2023

Hvordan bruke en elektronisk labnotatbok?
16. FEBRUAR 2023

Hvordan skrive en datahåndteringsplan
16. FEBRUAR 2023

Introduction to research data management @ UiT
22. MARS 2023

How to manage sensitive data
13. APRIL 2023

How to structure and document research data
17. APRIL 2023

How to store research data
17. APRIL 2023

Data cleaning
17. APRIL 2023

How to archive research data
18. APRIL 2023

How to archive data in UiT Open Research Data
18. APRIL 2023

How to search and cite research data
19. APRIL 2023

Research data: Rights and licenses
19. APRIL 2023

How to use an electronic lab notebook
20. APRIL 2023

How to write a Data Management Plan
20. APRIL 2023

Egen OpenRefine workshop 25. april kl 12-15

Mer informasjon og hjelp

[Forskningsdataportalen ved UiT]

Email: researchdata@hjelp.uit.no

Referanser

Teal et al., 2019, datacarpentry/spreadsheet-ecology-lesson: Data Carpentry: Data Organization in Spreadsheets for Ecologists, June 2019: Zenodo, doi:10.5281/zenodo.3269869.

White et al., 2013. Nine simple ways to make it easier to (re)use your data, Ideas in Ecology and Evolution 6(2): 1-10 Special Issue-Data Sharing in Ecology and Evolution https://ojs.library.queensu.ca/index.php/IEE/article/view/4608

Hadley Wickham, *Tidy Data*, Vol. 59, Issue 10, Sep 2014, Journal of Statistical Software. http://www.jstatsoft.org/v59/i10.