Analyseren van data met behulp van VUStat
Met afsluitende opdrachten

1. Analyseren van data met behulp van VUStat apps via https://www.vustat.eu/apps/

We gebruiken hier de apps van VUStat maar je kunt ook het programma VUStat downloaden via https://www.vusoft.be/vustat.html

Via https://youtu.be/g02wvJRptCE kun je ook een video met uitleg van VUStat apps.

Kies voor **Data Analyse**
We illustreren de data analyse met VUStat app aan de hand van het bestand pinguïn.
Haal dit bestand eerst van de Epsilon site en zet het bestand pinguin.json op je computer.
Ga naar Bestand Openen en upload het bestand pinguin.json.
Je ziet nu onderstaande tabel.

Bestand	Data	Grafiek	Tabel	Kentallen	Resampling	
##	species	island	bill_lengt...	bill_depth...	flipper_le...	body_ma...	sex
1	Adelie	Torgersen	39.10	18.70	181	3750	MALE
2	Adelie	Torgersen	39.50	17.40	186	3800	FEMALE
3	Adelie	Torgersen	40.30	18.00	195	3250	FEMALE
4	Adelie	Torgersen	**	**	*	*	*
5	Adelie	Torgersen	36.70	19.30	193	3450	FEMALE
6	Adelie	Torgersen	39.30	20.60	190	3650	MALE
7	Adelie	Torgersen	38.90	17.80	181	3625	FEMALE
8	Adelie	Torgersen	39.20	19.60	195	4675	MALE
9	Adelie	Torgersen	34.10	18.10	193	3475	*
10	Adelie	Torgersen	42.00	20.20	190	4250	*
11	Adelie	Torgersen	37.80	17.10	186	3300	*
12	Adelie	Torgersen	37.80	17.30	180	3700	*
13	Adelie	Torgersen	41.10	17.60	182	3200	FEMALE
14	Adelie	Torgersen	38.60	21.20	191	3800	MALE
15	Adelie	Torgersen	34.60	21.10	198	4400	MALE
16	Adelie	Torgersen	36.60	17.80	185	3700	FEMALE

Antarctica is het continent rond de zuidpool van de Aarde. In dit gebied leven naar schatting 5,9 miljoen pinguïns. Australische wetenschappers hebben onderzoek gedaan naar de kenmerken van pinguïns. Hierbij hebben ze van ruim 300 pinguïns de soortnaam genoteerd, het eiland waar ze voorkomen, de snavellengte (bill_length), snavelenbreedte (bill_depth), flipperlengte (flipper_length), lichaamsgewicht (body_mass) en geslacht (seks). Deze Palmer Archipelago (Antartica) pinguin data werden verzameld en beschikbaar gesteld door Dr. Kristen Gorman en het Palmer Station, Antarctica LTER (https://github.com/mwaskom/seaborn-data).

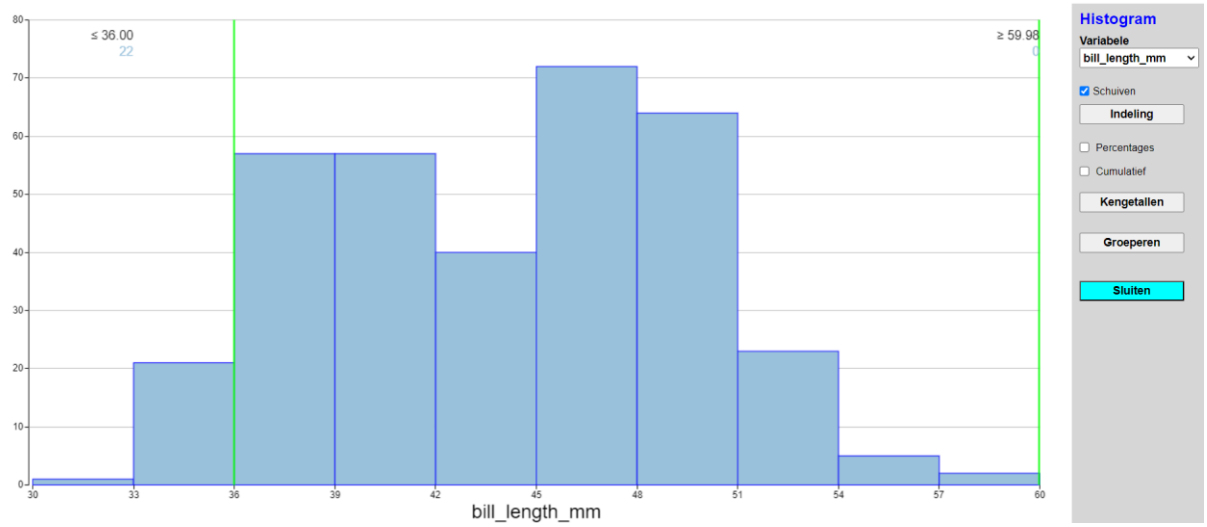
Via de menu's in de bovenste balk kun je kiezen voor grafieken, tabellen en kentallen. Zie als voorbeeld hieronder.

Bestand | Data | Grafiek | Tabel | Kentallen | Resampling | Opties

Dotplot
Histogram
Lijndiagram
Cirkeldiagram
Boxplot
Puntenwolk
Beslisbomen

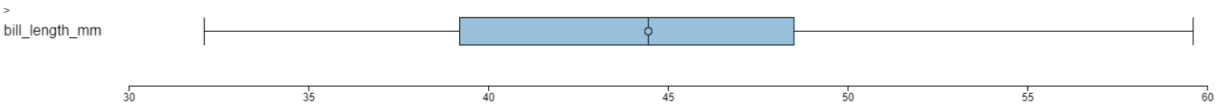
##	species	island	_depth...	flipper_le...	body_ma...	sex
1	Adelie	Torgers	18.70	181	3750	MALE
2	Adelie	Torgers	17.40	186	3800	FEMALE
3	Adelie	Torgers	18.00	195	3250	FEMALE
4	Adelie	Torgers	**	*	*	*
5	Adelie	Torgers	19.30	193	3450	FEMALE
6	Adelie	Torgers	20.60	190	3650	MALE
7	Adelie	Torgers	17.80	181	3625	FEMALE
8	Adelie	Torgers	19.60	195	4675	MALE
9	Adelie	Torgers	18.10	193	3475	*
10	Adelie	Torgers	20.20	190	4250	*
11	Adelie	Torgers	17.10	186	3300	*
12	Adelie	Torgersen	37.80	17.30	180	3700	*
13	Adelie	Torgersen	41.10	17.60	182	3200	FEMALE
14	Adelie	Torgersen	38.60	21.20	191	3800	MALE
15	Adelie	Torgersen	34.60	21.10	198	4400	MALE
16	Adelie	Torgersen	36.60	17.80	185	3700	FEMALE

We maken een histogram voor de snavellengte:

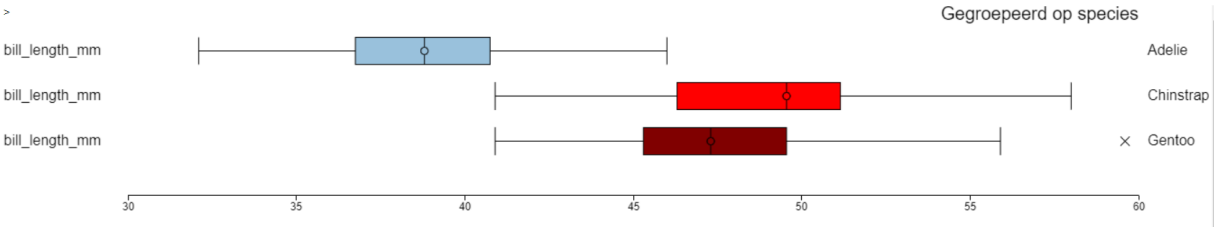


Je ziet verschillende mogelijkheden voor verdere analyse. Je kunt klassebreedte van de klassen aanpassen via indeling; je kunt procentuele verdeling geven (handig bij vergelijken van groepen); je kunt kentallen opvragen (bijvoorbeeld gemiddelde, mediaan, standaardafwijking, enz); tenslotte kun je groeperen, d.w.z. je kunt de groep splitsen in verschillende deelgroepen die je dan kunt vergelijken. Dit laatste demonstreren we in het volgende voorbeeld. Wil je weer in hoofdmenu komen dan moet je voor sluiten kiezen (rechts)

We maken nu een boxplot van de snavellengte:



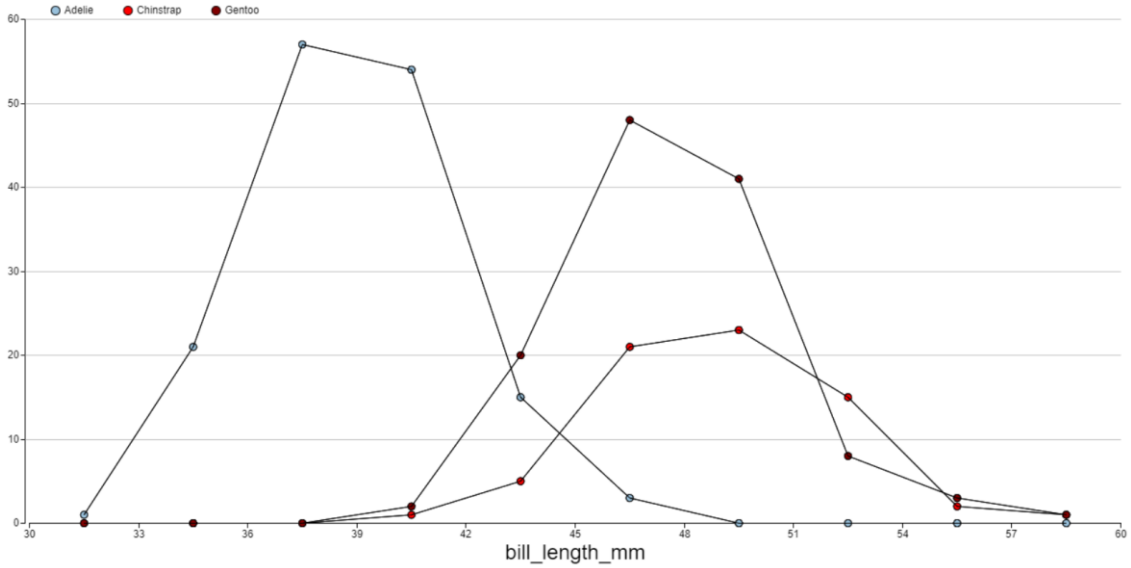
Via indeling kiezen we om de verschillende soorten (species) te vergelijken.



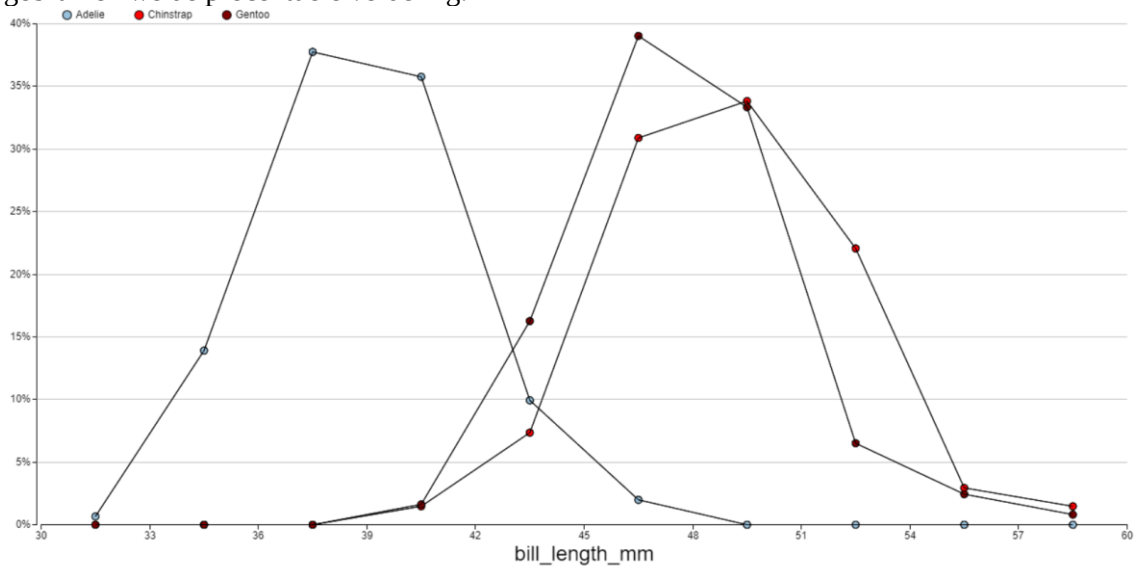
Ook kunnen de kentallen (via Kentallen in menu) voor de verschillende soorten berekend worden:

	bill_length_mm		
species	**Adelie**	**Chinstrap**	**Gentoo**
Waarnemingen	151	68	123
Gemiddelde	38.791	48.834	47.505
SD	2.663	3.339	3.082
Minimum	32.100	40.900	40.900
Eerste kwartiel	36.750	46.300	45.300
Mediaan	38.800	49.550	47.300
Derde kwartiel	40.750	51.150	49.550
Maximum	46.000	58.000	59.600

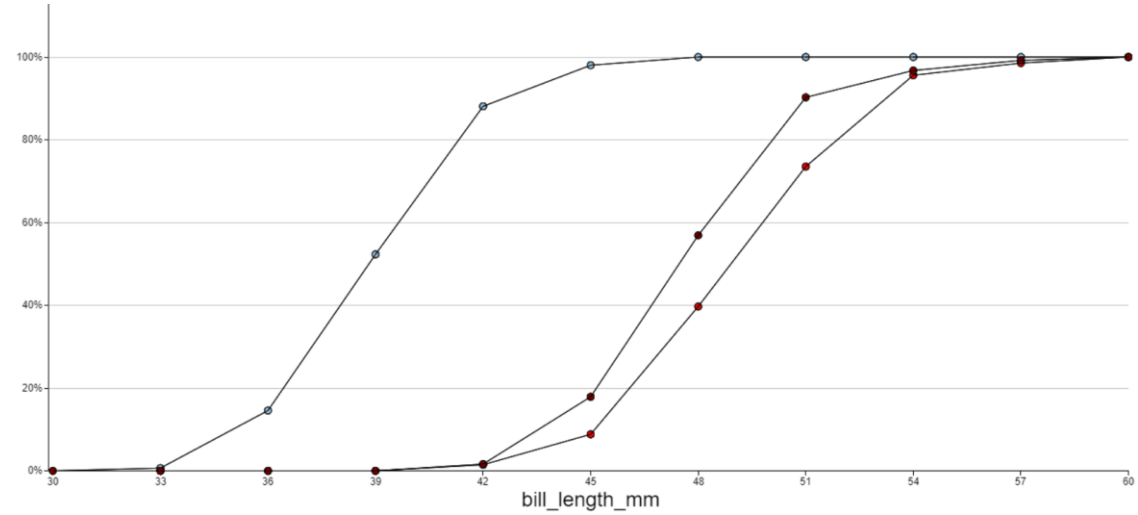
We kunnen ook frequentiepolygonen maken voor de verschillende soorten:



De aantallen van de verschillende soorten zijn niet gelijk. Om de soorten beter te kunen vergelijken gebruiken we de procentuele verdeling.



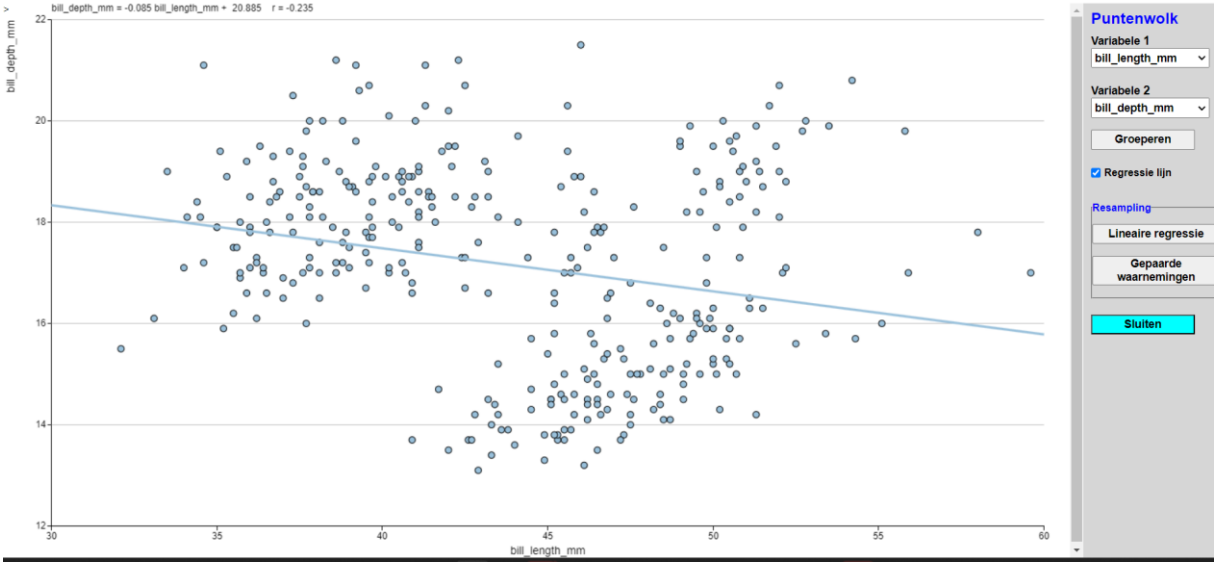
Of cumulatieve frequentiepolygonen, waarbij we voor procenten kiezen om ze beter te vergelijken.



Op basis van deze grafieken kun je allerlei conclusies trekken:

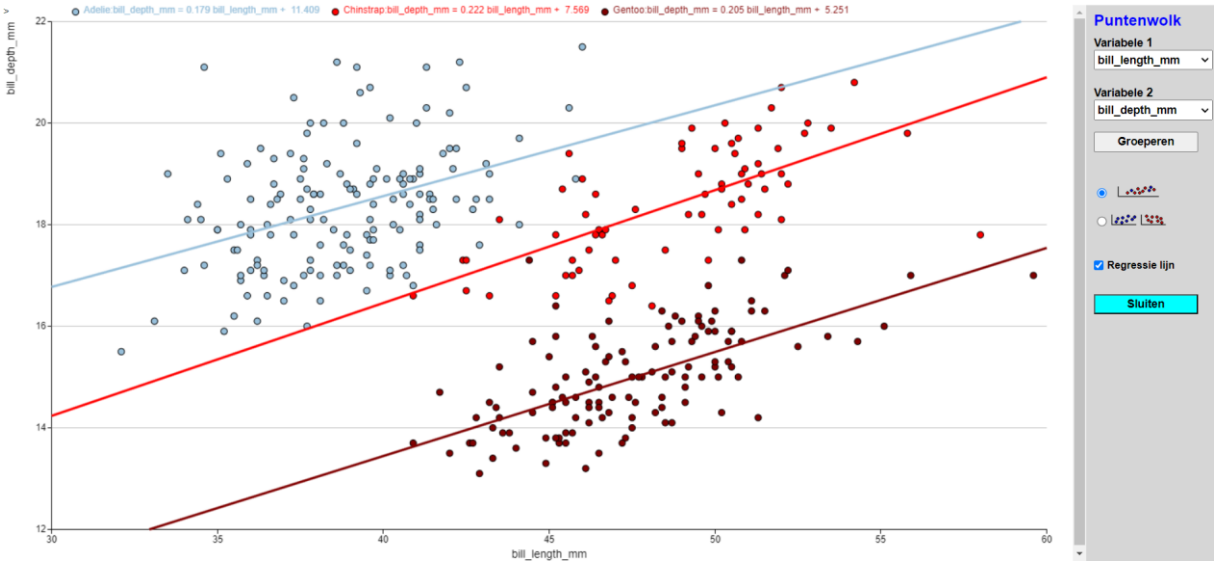
- Adelie heeft de kleinste snavellengte; die van Chinstrap en Gentoo zijn nagenoeg even lang
- Spreiding in snavellengte is bij alle soorten vrijwel even groot
- Dat zie je ook aan standaardafwijkingen en de interkwartielafstand (verschil tussen eerste en derde kwartiel
- Chinstrap komt minder voor

We kunnen ook naar verbanden tussen variabelen kijken
Via Grafiek en Puntenwolk kijken we naar het verband tussen snavellengte en snavelbreedte. Hieronder zie je de puntenwolk met een regressielijn. Klik daarvoor de regressielijn aan rechts in scherm; de vergelijking van regressielijn staat boven de puntenwolk; daarnaast staat de correlatiecoefficient.

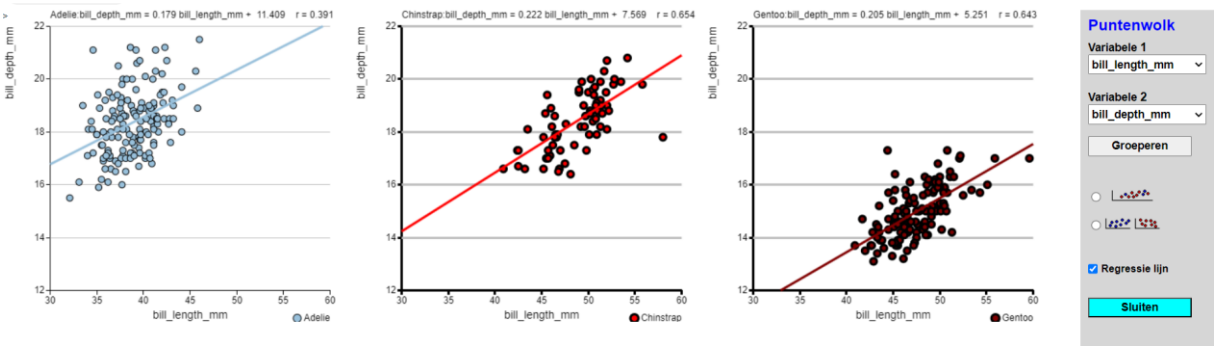


Er lijkt een negatief verband tussen snavellengte en snavelbreedte.

Als we echter weer kijken naar de verschillende soorten (via indeling) vinden we:



of



Je ziet dat de dalende regressielijn nu verdwenen is en dat voor ieder soort er een stijgend verband is tussen de snavellengte en snavelbreedte.

Als laatste nog twee zaken. Ten eerste is het noodzakelijk dat je bij het starten met een bestand dit opschoont, d.w.z. alle 'onzinnige' records moeten verwijderd worden. Dat kun je doen via DATA en dan records verwijderen, of via Filter.
Ten tweede zul je soms nieuwe variabelen, die aan de hand van de huidige data bepaald worden, willen introduceren. Je zou bijvoorbeeld de variabele snavellengte/snavelbreedte kunnen maken. Die doe je via DATA en dan Omrekenen.

2. Afsluitende Opdrachten

Zelf exploreren van data bij bestand pinguins.

1. Onderzoek de verbanden tussen de snavellengte en flipperlengte; ook tussen snavellengte en lichaamsgewicht. Houd weer rekening met de verschillende soorten.
2. Onderzoek of er verschillen zijn tussen de sexen met betrekking tot de variabelen snavellengte, snavelbreedte, flipperlengte, lichaamsgewicht.
3. Onderzoek of de verdeling van de verschillende soorten op de eilanden vergelijkbaar is. Geldt dat voor beide sexen?

Zelf data exploreren bij bestand Samplon.

Dit data-bestand gaat over het denkbeeldige land Samplonië. Dat is een is een eilandje met 1000 inwoners. Het heeft slechts twee provincies: Agrië, waar met name landbouw en veeteelt wordt bedreven, en Indusië, waar vooral industriële bedrijvigheid is en de grotere steden te vinden zijn.

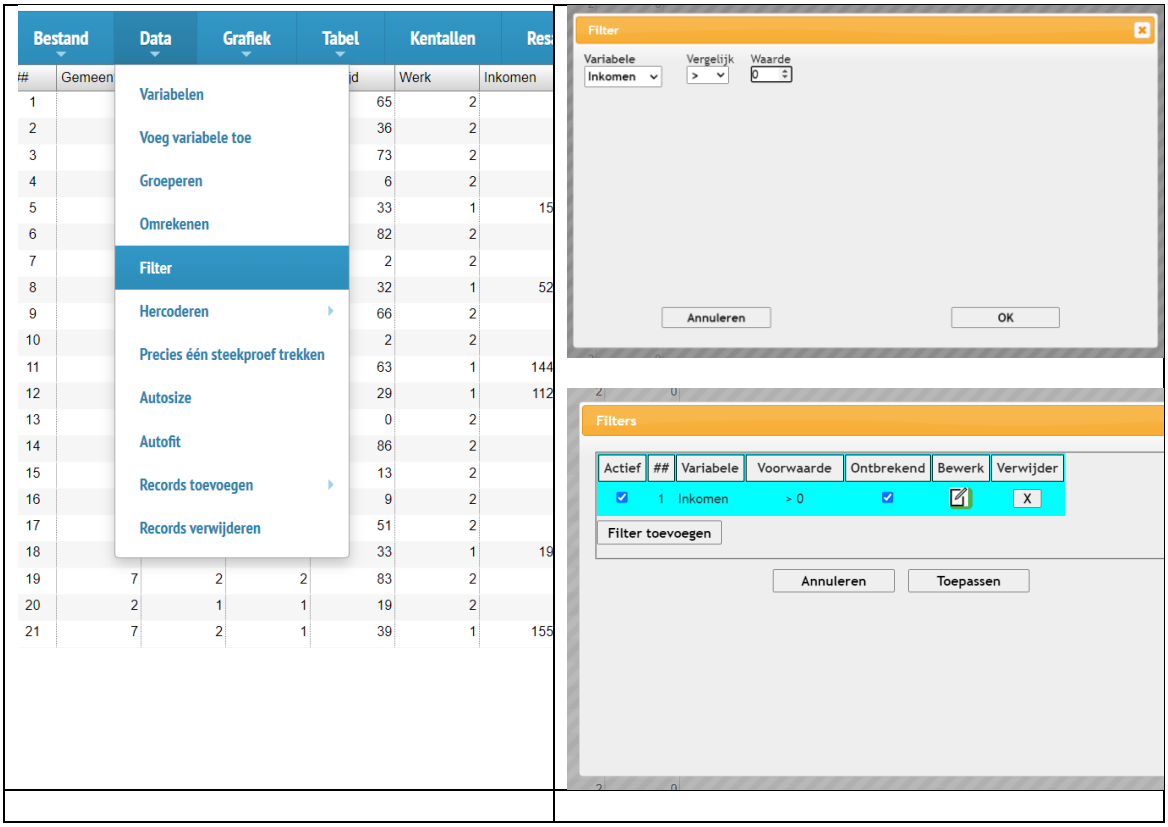
Variabele	Type	Codes & labels
Gemeente	Kwalitatief	1 = Akkerwinde 2 = Grasmalen 3 = Nieuwekans 4 = Lommerdal 5 = Smeulde 6 = Stapelrade 7 = Vuilpanne
Provincie	Kwalitatief	1 = Agrië 2 = Indusië
Geslacht	Kwalitatief	1 = Man 2 = Vrouw
Leeftijd	Kwantitatief	0 t/m 99 (2 cijfers, 0 decimalen)
Werk	Kwalitatief	1 = Werkzaam 2 = Werkloos
Inkomen	Kwantitatief	0 t/m 4500 (maximaal 4 cijfers, 0 decimalen)

1. Onderzoek de werkeloosheid

- Zijn vrouwen in Samplonie relatief meer werkeloos dan mannen?
- Is de werkeloosheid in beide provincies vergelijkbaar?
- Zijn 50-plussers vaker werkeloos dan mensen onder de 30?

2. Vergelijk de leeftijdsopbouw van werkenden en werkelozen. Ook die van mannen en vrouwen.
3. Zijn de werkeloosheidspercentages in de verschillende gemeenten vergelijkbaar?

4. Kies Data en gebruik Filter om de volgende selectie te maken (alle inkomens groter dan 0)



Onderzoek of er een verband is tussen leeftijd en inkomen.

5. Bij onderzoek naar verband tussen leeftijd en inkomen lijkt het erop dat dit inderdaad een fictief bestand is.
   Hoe zie je dat dit waarschijnlijk is?

Zelf exploreren van data bij bestand Ziekenhuizen.

Dit data-bestand is afkomstig van het Centraal Bureau voor de Statistiek (CBS) en bevat (onder andere) voor alle 431 Nederlandse gemeenten (in 2010) de gemiddelde afstand van de inwoners (in km) tot het dichtstbijzijnde ziekenhuis en tot de dichtstbijzijnde huisarts. Dit bestand bevat ook voor elke gemeente de naam, de omvang van de bevolking, de bevolkingsdichtheid (aantal inwoners per km^2) en het landsdeel.

1. Onderzoek in hoeverre er verschillen zijn tussen de landsdelen met betrekking tot de variabele huisarts, en de variabele ziekenhuis. Gebruik verschillende grafieken en de kentallen.
2. En met betrekking tot de gemiddelde bevolking en gemiddelde dichtheid?

3. Is er een verband tussen de variabelen huisarts en ziekenhuis? In hoeverre verschilt dit per landsdeel?