

HUIDGELEIDING – ANALYSE

Leerdoelen

- Ervaring opdoen met meten aan huidgeleiding
- Omgaan met grote datasets in Excel
- Het gebruik van gemiddelde als datafilter-techniek

Validatie

In het vorige werkcollege heb je gezien hoe huidgeleiding beïnvloedt wordt door het sympatisch zenuwstelsel, en hoe dit te gebruiken is om een leugen te detecteren. Vandaag ga je, naast het doorlopen van een huidgeleidingsmeting, ervaring opdoen met data-analyse van grote tijdseries in Excel.

Opdracht

0. Download de data

Op Brightspace vind je in week 4, werkcollege 2, een dataset die bestaat uit huidgeleidingsdata tijdens een leugendetectie-protocol. Download deze dataset *2122_B&F_wo2_GSRdata.xlsx*

Open de data in Excel. De data bestaat ditmaal uit 2309 rijen en 2 kolommen. De eerste kolom geeft de tijd in seconden aan sinds de start van de meting. De waarden in kolom B representeren het voltage in microvolt (μV) gemeten via de analoge input van de Arduino.

Tevens zie je een blok data in cellen F1:I13. In kolom F staat de tijd waarop de proefleider aangaf dat de meting gestart of gestopt was, of waarop de proefleider een getal uitvroeg. De gebeurtenis van dat moment staat beschreven in kolom G.

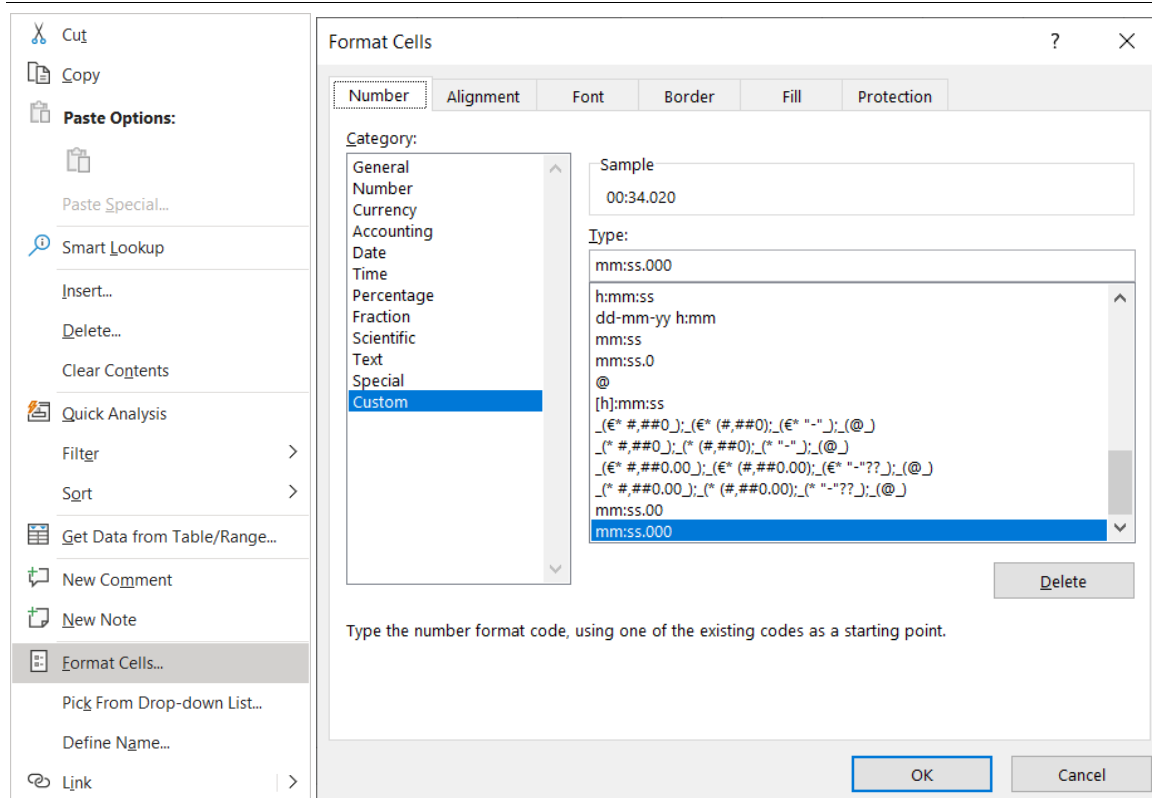
1. Formules doorgronden

Indien je datasets analyseert die je niet zelf gemeten hebt, kan het voorkomen dat je waarden of formules ziet die je misschien niet meteen begrijpt. Daarom is het belangrijk om bij iedere dataset die je tegenkomt, eerst eens door de data heen te bladeren en de verschillende variabelen te onderzoeken.

Bepaal voor jezelf eerst hoe kolom H relateert aan kolom F. Indien de data kolomtitels bevat, kan dit je al erg helpen. Ook het bekijken van de getallen kan misschien verraden wat de relatie ertussen is.

Antwoord:

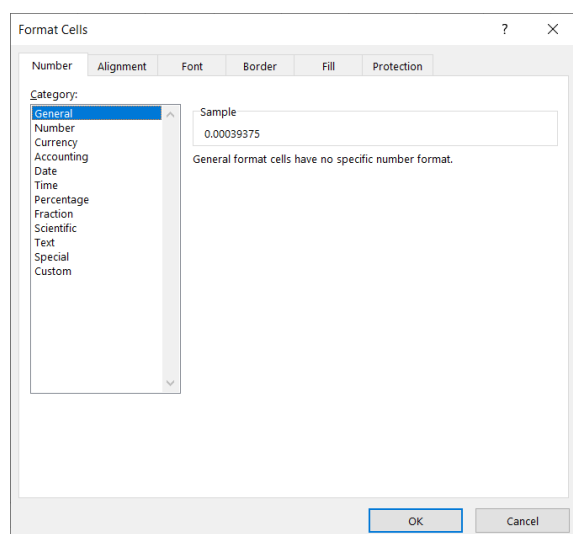
Het is handig om te weten dat Excel een getal op verschillende manieren kan tonen. Om dit te zien, klik met je rechtermuisknop op cel F2 en kies "Format Cells..."



Je kunt zien dat cel F2 een “custom” tijdsschaal toont, die onder “Type:” gespecificeerd staat: “mm:ss.000”. Minuten in 2 getallen, gevolgd door seconden in 2 getallen, gevolgd door 3 decimalen. Deze laatste waarden geven dus de milliseconden aan.

Intermezzo – Tijd in Excel

Zet voor extra inzicht even de categorie op “ruwe data” door links in het categorie-selectie menu op “General” te klikken. Excel zal je een voorbeeld tonen hoe de data eruit komt te zien in dit format. Zoals je kunt zien staat de tijd van 34 seconden en 20 milliseconden gelijk aan een getal van 0.00039375. Een waarde van 0 staat voor Excel gelijk aan 00:00:00, en een waarde 1 gelijk aan 24 uur later. Deze enorme kleine fractie is dus om te rekenen naar seconden door het te vermenigvuldigen met het aantal seconden dat er in 24 uur zit: $24 \cdot 60 \cdot 60 = 86400$.



Druk op “Cancel” om cel F2 weer het oorspronkelijk custom format (mm:ss.000) te laten behouden.

Nu je weet wat er in kolom H gebeurt, is het misschien inzichtelijker wat er gebeurt in kolom I. Dubbelklik op cel I2, en bekijk de formule. De functie MATCH of VERGELIJKEN heb je misschien nog niet eerder gezien. Klik daarom op de dikgedrukte functienaam die direct onder je cel staat, in de witte rechthoek. Er zal aan de rechterkant een help-window openen. Hierin kun je lezen wat deze functie doet. Deze help-functie is erg nuttig om formules te leren begrijpen en gebruiken.


Sheet View


Workbook Views


Show

Zoom

AVERAGE







=MATCH(H2,\$A\$1:\$A\$2309,1)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	1.58	556.64				tijd (mm:ss gebeurtenis)		tijd (s)	rij				
2	1.69	561.52				00:34.020 start		34.02	=MATCH(H2,\$A\$1:\$A\$2309,1)				
3	1.79	561.52				00:43.160	5	43.16	MATCH(lookup_value, lookup_array, [match_type])				
4	1.89	561.52				01:01.200	7	61.2	594				
5	1.98	561.52				01:13.530	4	73.53	717				
6	2.09	566.41				01:28.640	1	88.64	867				
7	2.19	551.76				01:36.770	3	96.77	948				
8	2.28	566.41				02:30.380	9	150.38	1482				
9	2.39	566.41				02:39.560	2	159.56	1573				
10	2.48	566.41				02:53.100	6	173.1	1708				
11	2.59	561.52				03:02.840	8	182.84	1805				
12	2.68	566.41				03:12.130	3	192.13	1897				
13	2.79	566.41				03:49.530 einde		229.53	2270				
14	2.89	566.41											
15	2.99	566.41											

Leg in eigen woorden uit wat MATCH/VERGELIJKEN doet:

Als je je afvraagt wat die laatste 1 in je formule precies doet, dan kun je deze 1 ook even verwijderen. Op het moment dat je in je formule aan bent gekomen bij de [match type], zal Excel met een extra uitleg window beschrijven welke opties er zijn.

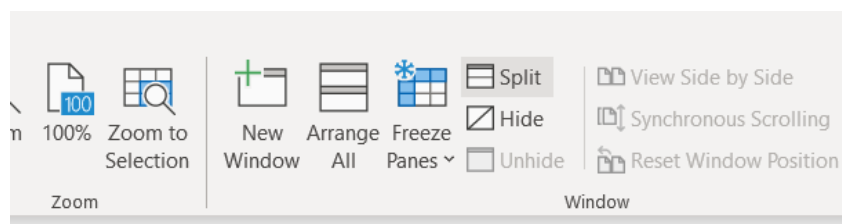
Je ziet dat de datareeks waarin we iets gaan opzoeken (A1:A2309) beschreven staat met dollartekens. Vraag jezelf af waarom dat handig zou zijn.

Antwoord:

2. Grafiek plotten van een lange datareeks

Je hebt in vorige opdrachten al grafieken gemaakt van kleine datasets door de data te selecteren en een grafiek in te voegen. Deze eerste stap wordt echter erg vervelend als je iedere keer door 2000+ rijen moet slepen. Gelukkig heeft Excel meerdere opties om het overzicht te behouden bij enorme datasets. Vandaag ga je gebruik maken van het genaamde "split view".

Klik op cel A16, en vervolgens in het "View" tabblad van Excel, in het subkopje "Window", voor de optie "Split".



Zodra je deze optie aanvinkt, zal je een dikke grijze lijn door je Excel sheet zien lopen. Je ziet aan de rechterkant van je scherm dat de verticale scroll-balk ook in tweeën is gedeeld. Je kunt nu onder en boven deze lijn onafhankelijk van elkaar door je data scrollen. Je data is volledig onveranderd, alleen het scherm toont nu meerdere delen van de datasheet.

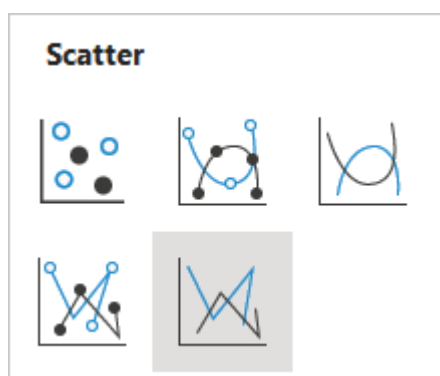
Lui als we zijn, gaan we natuurlijk ook niet helemaal met de hand naar beneden scrollen. Selecteer in plaats daarvan in het onderste deel van je scherm (onder je split) een van de cellen die data bevat in kolom A of B. Zodra je deze cel selecteert zie je de groene omranding van de geselecteerde cel, en het bekende "formule-doorsleep"-vierkantje in de rechteronderhoek.

17	3.20	56
18	3.29	56
19	3.39	56
20	3.50	55

De horizontale en verticale groene zijden hebben echter ook een eigen functie. Ga met je muis op de onderste horizontale groene streep van je cel staan (je muis-icoon verandert in een kruising van 4 pijltjes), en dubbelklik. Excel zal nu de alleronderste cel selecteren die ook data bevat. Dit werkt ook als je vanaf hier de bovenste, of meest rechter cel wilt bereiken die data bevat. Dubbelklik om deze cellen te bereiken respectievelijk op de bovenkant, of rechterzijkant van je geselecteerde cel. Dit werkt overigens ook bij cellen die geen data bevatten. Selecteer cel D9 maar eens in je view boven de split, en dubbelklik op de linker- en rechterkant van deze cel.

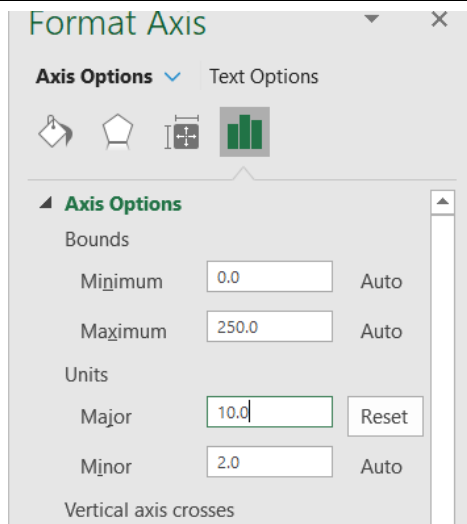
Zorg ervoor dat je bovenste split view de bovenkant van je dataset toont, en de onderste split view de onderkant. Selecteer vervolgens de data waar je een grafiek van wilt maken door in je bovenste view cel A1 aan te klikken. Houdt vervolgens SHIFT in, en klik in je onderste view op cel B2309. Je zult zien dat alle data geselecteerd is.

Ga in het menu vervolgens naar Insert/Toevoegen > Charts/Grafieken, en kies een "X-Y Scatter with straight lines"



Indien je grafiek in je onderste split view terecht komt, knip deze dan weg met CTRL+X (pc) CMD+X (mac). Selecteer in je bovenste view een cel als K1 en plak je grafiek.

Geef je grafiek een goede titel en astitels. Pas vervolgens het bereik van de verticale as aan door met de rechtermuisknop op de getallen van de y-as te klikken en selecteer "Format axis"/ "As aanpassen". Zet het minimum van de as op 500 en het maximum op 950. Pas ook de horizontale as aan door er met de rechtermuisknop op te klikken en "format axis/ as aanpassen" te kiezen.



Zet de major units op 10, zodat er iedere 10 seconden een tijdsmarkering op de as zichtbaar wordt.

Rek je grafiek nog wat uit, tot bijvoorbeeld kolom Z, en kijk of je al bijzonderheden in het signaal ziet.

3. Data relateren aan vragen

Voor je inzicht zou het mooi zijn als we de data van de gestelde vragen ook zichtbaar kunnen maken in de grafiek. Daarom gaan we een nieuwe datareeks maken die een extreme score bevat bij een belangrijke gebeurtenis.

Zorg dat je in je bovenste split view cellen F1:113 kunt zien. Scroll in je onderste split view naar de rij die het begin van een gebeurtenis aangeeft, en noteer in kolom C op die rij een groot getal, zoals bijvoorbeeld 1000. Ter voorbeeld: je kunt in cellen F2:I2 zien dat tijdstip 00:34.020 overeenkomt met rij 323 in de data. Scroll daarom in de onderste split view naar cel C323, en vul in deze cel het getal 1000 in.

Zodra je deze cellen in kolom C van een getal hebt voorzien, is het handig om de tussenliggende cellen te vullen met een andere extreme waarde, zoals 0. Tik in je bovenste split view de bovenste cel van je reeks (cel C1) het getal 0 in, en druk op Enter.

Scroll in je onderste split view eerst helemaal naar de bovenkant van je dataset. Klik vervolgens een lege cel in kolom C aan en dubbelklik op de onderste horizontale zijde om bij de "laatste" lege cel uit te komen, namelijk cel C322.

	A	B	C	D	E	F	G	H	I	J
1	1.58	556.64	0			tijd (mm:ss)	gebeurten	tijd (s)	rij	
2	1.69	561.52				00:34.020	start	34.02	323	
3	1.79	561.52				00:43.160		5	43.16	414
4	1.89	561.52				01:01.200		7	61.2	594
5	1.98	561.52				01:13.530		4	73.53	717
6	2.09	566.41				01:28.640		1	88.64	867
7	2.19	551.76				01:36.770		3	96.77	948
8	2.28	566.41				02:30.380		9	150.38	1482
9	2.39	566.41				02:39.560		2	159.56	1573
10	2.48	566.41				02:53.100		6	173.1	1708
11	2.59	561.52				03:02.840		8	182.84	1805
12	2.68	566.41				03:12.130		3	192.13	1897
13	2.79	566.41				03:49.530	einde		229.53	2270
14	2.89	566.41								
15	2.99	566.41								
16	3.10	566.41								
17	3.20	566.41								
316	33.22	581.05								
317	33.32	576.17								
318	33.42	576.17								
319	33.52	576.17								
320	33.62	576.17								
321	33.72	576.17								
322	33.82	576.17								
323	33.92	581.05								
324	34.02	576.17								
325	34.12	576.17								
326	34.23	576.17								

	A	B	C	D	E	F	G	H	I	J
1	1.58	556.64	0			tijd (mm:ss)	gebeurten	tijd (s)	rij	
2	1.69	561.52				00:34.020	start	34.02	323	
3	1.79	561.52				00:43.160		5	43.16	414
4	1.89	561.52				01:01.200		7	61.2	594
5	1.98	561.52				01:13.530		4	73.53	717
6	2.09	566.41				01:28.640		1	88.64	867
7	2.19	551.76				01:36.770		3	96.77	948
8	2.28	566.41				02:30.380		9	150.38	1482
9	2.39	566.41				02:39.560		2	159.56	1573
10	2.48	566.41				02:53.100		6	173.1	1708
11	2.59	561.52				03:02.840		8	182.84	1805
12	2.68	566.41				03:12.130		3	192.13	1897
13	2.79	566.41				03:49.530	einde		229.53	2270
14	2.89	566.41								
15	2.99	566.41								
16	3.10	566.41								
17	3.20	566.41								
318	33.42	576.17								
319	33.52	576.17								
320	33.62	576.17								
321	33.72	576.17								
322	33.82	576.17								
323	33.92	581.05								
324	34.02	576.17								
325	34.12	576.17								
326	34.23	576.17								
327	34.33	576.17								
328	34.43	581.05								

Klik nu in je bovenste split view op cel C1. Houdt SHIFT in en klik in je onderste split view op cel C322, zodat cellen C1:C322 geselecteerd zijn.

Houdt vervolgens CTRL in, en druk op D (CTRL+D, dus). Dit is de sneltoets waarmee Excel de formule of het getal dat in de eerste cel beschreven staat, automatisch doortrekt over de gehele geselecteerde range.

Selecteer in de bovenste split view een van de vers ingevulde cellen, en navigeer naar de onderkant van deze reeks door op de onderkant van deze cel te dubbelklikken.

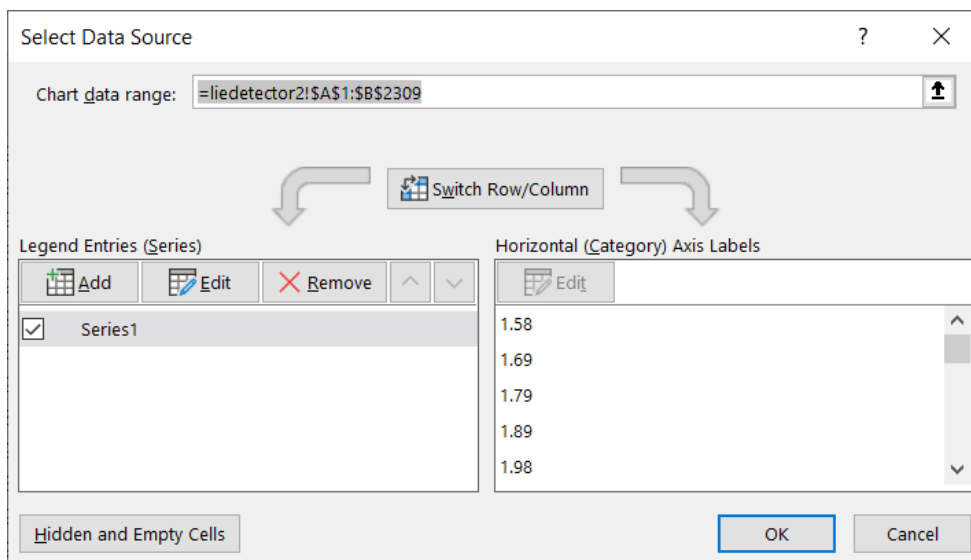
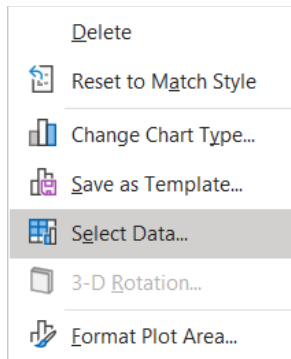
Kies in de onderste split view voor een lege cel in kolom C, en dubbelklik op de onderkant van deze cel om naar het begin van onze volgende gebeurtenis te navigeren.

	A	B	C	D
315	33.12	576.17	0	
316	33.22	581.05	0	
317	33.32	576.17	0	
318	33.42	576.17	0	
319	33.52	576.17	0	
320	33.62	576.17	0	
321	33.72	576.17	0	
322	33.82	576.17	0	
323	33.92	581.05	1000	
324	34.02	576.17		
325	34.12	576.17		
326	34.23	576.17		
327	34.33	576.17		
328	34.43	581.05		
329	34.53	576.17		
330	34.63	585.94		
331	34.73	581.05		
408	42.46	581.05		
409	42.56	585.94		
410	42.66	585.94		
411	42.76	590.82		
412	42.86	585.94		
413	42.96	585.94		
414	43.06	590.82	1000	
415	43.16	581.05		
416	43.26	585.94		
417	43.37	590.82		
418	43.47	585.94		

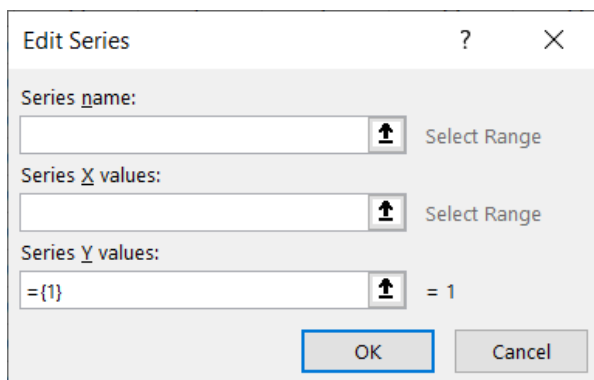
Vul in dit lege tussenstuk wederom allemaal 0 in door in de eerste cel een 0 te typen, de range te selecteren, en met CTRL+D de range te vullen met 0'en.

4. Datareeks toevoegen aan grafiek

Nu je een datareeks hebt die op dezelfde tijdschaal weergeeft dat er iets gebeurt, kun je deze reeks toevoegen aan je grafiek. Rechterklik in je grafiek en kies “select data” / “Selecteer gegevens”





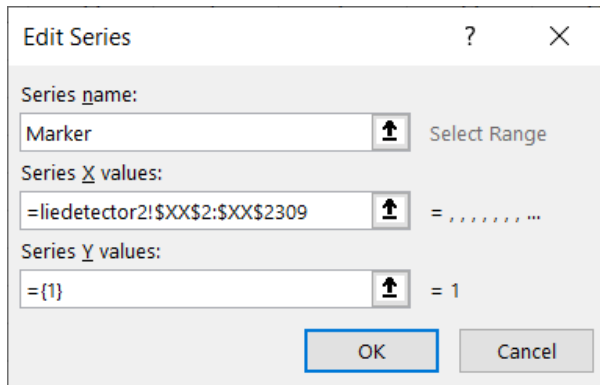
Je ziet dat er al een reeks data aanwezig is, met de naam “Series1”. Door deze reeks aan te klikken en te kiezen voor “Edit” is de naam nog aan te passen. Dit kun je later zelf doen. We gaan nu echter eerst een nieuwe reeks toevoegen. Klik op “Add”.



Kies voor de “Naam” een logische beschrijving, zoals “Marker”. Kies voor de X-waarden dezelfde tijdreeks als diegene die onze GSR-data gebruikt.

Dit is kolom:

Door op het  icoontje te klikken kun je de datareeks in je sheet selecteren. Doe dit ook weer handig met je split view, en selecteren met shift. Klik weer op het icoontje rechts van je invoerbalk  om terug te keren naar het voorgaande menu



Kies nu voor de Y-values de data die relateert aan onze marker.

Dit is kolom:

Druk op Ok, en laat een docent je mooie grafiek bewonderen.

5. Interpretatie

Gebaseerd op de huidgeleiding tijdens het protocol, is het waarschijnlijk dat de participant dacht aan het getal:

6. Ruis filteren met een gemiddelde

Misschien heb je bij het bestuderen van deze dataset, maar ook in het meten van huidgeleiding, gemerkt dat de output van deze meting snel kan fluctueren. Op een tijdschaal van seconden worden de waarden bepaald door de sympatische zweetrespons van de huid. Op kleinere tijdschaal lijkt het signaal echter ook ruis te bevatten, en snel heen en weer te springen. Hier kunnen we op een simpele manier al enigszins voor compenseren.

In vorige werkopdrachten heb je gemerkt dat je met het berekenen van een gemiddelde over meerdere waarden kunt compenseren voor een enkele uitschieter. De gemiddelde hartslag van een groep van 10 mensen zal slechts een beetje worden beïnvloed met het includeren van een elfde persoon die een 2x zo hoge hartslag heeft.

Dit concept kunnen we ook gebruiken bij de analyse van tijdseries. Door een gemiddelde te berekenen over meerdere samples, kunnen extreme uitschieters elkaar compenseren.

Ga in cel D3 staan, en bereken in de deze cel het gemiddelde van cellen B1:B5. Zorg ervoor dat deze berekening vervolgens gedaan wordt tot en met cel D2307. Voeg deze data toe aan je grafiek onder de naam "gemiddelde5".

Doe hetzelfde in kolom E, maar ditmaal voor een gemiddelde van 9 cellen. Begin dus in cel E5 met de eerste formule, en eindig met de laatste formule in E2305. Voeg ook deze data toe aan je grafiek met de naam "gemiddelde9".

Vergelijk de grafieken. Wat is het voordeel van het middelen?

Antwoord:

Bekijk nu een klein onderdeel van je data, bijvoorbeeld door een aparte grafiek te maken in de buurt van de "is het getal 3?"-vraag. Zie je verschillen tussen de grafieken als je let op de tijdschaal?

Wat zou je dus als nadeel kunnen noemen van het middelen van datapunten?

Antwoord:

Sla je excelsheet op in je B&F folder.

Al deze sheets bevatten formules die je later misschien weer kunt gebruiken.

Pro-tip: als je in een .CSV-bestand aan het werken bent, en je hebt grafieken en formules gebouwd, sla je werk dan ALTIJD OP ALS .XLS of .XLSX-bestand.

Csv-bestanden kunnen namelijk geen formules en grafieken bevatten. Opslaan als .csv zal er dus voor zorgen dat al je werk verloren gaat, en alleen getallen overblijven.