Portfolio 101

In de eerste assignment hebben we gekeken naar wat data science inhoudt. Toen hebben we deze slide gemaakt:



Ondertussen zijn we een aantal weken bezig geweest met data science in Jupyter Notebooks met Python en onder andere Pandas. Maar ondanks dat we nog geen opdracht gemaakt hadden toen we de powerpoint maakten, klopt hij eigenlijk vrij goed. Alle opdrachten draaiden om een dataset en het verkrijgen van kennis daaruit, hoewel er in de slide grote hoeveelheden genoemd worden is dat niet persé nodig voor data science. Data science is een multidisciplinair vakgebied omdat er wetenschappelijk onderzoek wordt gedaan in een bepaald vakgebied zoals economie, biologie of psychologie. Dit wordt gedaan op basis van wiskundige statistische methodes door middel van software talen als python en R. Wiskunde wordt dus gecombineerd met een business case en computer science. Ook is toegepaste statistiek een groot onderdeel van data science, je wilt namelijk weten of de verbanden die je uit je dataset verkregen hebt ook daadwerkelijke verbanden zijn, dit kun je testen met statistiek. Julia en R hebben we niet gebruikt bij de opdrachten maar dat betekent niet dat dit geen valide talen zijn om data science mee te beoefenen.

Data science draait zoals eerder al vernoemd eigenlijk om het kennis verkrijgen uit data. We willen dus verbanden uit onze data te zien krijgen zodat we daarmee bijvoorbeeld voorspellingen kunnen doen of we geldige conclusies voor ons onderzoek uit kunnen trekken. In de eerste paar assements gingen we vooral even wennen aan de tools die we konden gebruiken, en gingen we aan de slag met univariate analyse. Univariate analysis is het kijken naar 1 kolom van de data. In portfolio assignment 3 hebben we een univariate analysis gedaan op de numerical data van de penguins dataset. We hebben barplots gemaakt van de categorical data en hiermee ontdekt wat de verhoudingen binnen onze dataset waren. Dit legt eigenlijk uit wat de eerste stap is van een Data scientist, je kijkt naar kolommen apart van elkaar om te kijken of er binnen die kolom dingen opvallen, zo kun je te maken hebben met veel ontbrekende gegevens of een hele scheve verhouding van 2 categorieën waarvan je zou verwachten dat ze dichter bij elkaar zouden liggen, de verhoudingen van mannetjes en vrouwtjes binnen een populatie is hier een voorbeeld van.

Nadat we in assignment 4 onze eigen dataset hadden gekozen gingen we hier in assignment 6, 7 en 8 mee aan de slag. Eerst hebben we in 6 een univariate analysis gedaan op numerical data. Hierbij hebben we gebruik gemaakt van wat statistische termen als gemiddelde, mediaan, mode en standaardafwijking. Deze assignment liet goed zien hoe onmisbaar statistiek is binnen datascience. Het is vrij lastig om op basis van een tabel te zien hoe de data bijvoorbeeld verdeeld is, liggen er veel gegevens aan de uiteinden van de range of ligt alles vrij in het midden. Dit laat de standaardafwijking goed zien. Je kunt je gegevens ook plotten waardoor je op een visuele manier naar je data kunt kijken. Hierdoor kun je vrij snel een goede overview krijgen van je numerical data. Dit laat goed zien hoe data science ook gaat om het inzichtelijk maken van data, je wilt conclusies trekken, en je kunt zaken heel goed onderbouwen met overzichtelijke grafieken die ook voor mensen die niks van je data weten toch leesbaar zijn. Hier zit natuurlijk ook weer een gevaar in maar daar zullen we nu niet verder op in gaan. In assignment 6 zetten we de eerste stappen van het conclusies trekken op basis van data waar data science zo om draait.

In assignment 7 en 8 kwamen we in aanraking met distribution en confidence intervals. We gingen kijken wat een normaalverdeling precies is en hoe we gebruik kunnen maken van een normaalverdeling binnen onze data. Bij assignment 8 hebben we gezien wat de betrouwbaarheid van data precies inhoudt en hoe je dit kunt testen. Als je bij data scientist conclusies wilt gaan trekken op basis van data wil je natuurlijk eerst zeker weten dat de data waar je je op baseert betrouwbaar is. Eén van de dingen die je dan kunt doen is het confidence interval uitrekenen. Dit getal geeft eigenlijk aan binnen welke range ons echte gemiddelde ligt en met hoeveel procent kans dat is. Op basis hiervan kun je kijken hoe zeker je eigenlijk bent van je gemiddelde in de data en dit geeft aan hoe goed je sample eigenlijk de populatie weerspiegeld.

Bivariate analyse was de logische vervolgstap in onze reis door data science land. We hebben al gekeken naar een kolom van de data opzichzelfstaand maar we willen hierna natuurlijk weten hoe de verbanden tussen de data liggen. Kunnen we verklaren wat we zien met de andere gegevens die we hebben? In assignment 9 hebben we alle dit gedaan met de penguins dataset van seaborn. Met behulp van een correlation table en vervolgens een scatterplot konden we duidelijk een verband aantonen tussen body mass en flipper length. Dit aantonen van verbanden is heel belangrijk voor data science omdat het ervoor zorgt dat je voorspellingen kunt gaan doen in de toekomst of nu al beslissingen kan nemen op basis van deze verbanden. Dit was een verband tussen 2 numerical columns, dat hebben we ook gedaan met onze eigen student performance dataset in 10.

In assignment 11 en 12 hebben we een bivariate analyse uitgevoerd met numerical vs categorical data. Hierbij kunnen we bijvoorbeeld kijken hoe verschillende soorten of groepen andere numerical waardes hebben. Zo hebben we gekeken naar de flipper length van penguins per soort en naar de cijfers van studenten per geslacht. Door deze analyse kunnen we de gegevens gaan groeperen op de categorical data in onze dataset dit is belangrijk omdat we hiermee groepen kunnen gaan vergelijken om verschillen of overeenkomsten te ontdekken.

In assignment 13 en 14 zijn we gestart met het vergelijken van categorical met categorical data door middel van contingency tables en plots hiervan. Hierdoor kunnen we van de eerder gevonden resultaten, bijvoorbeeld vrouwen scoren hoger voor toetsen dan mannen, gaan kijken of het misschien komt doordat vrouwen meer voorbereiding voor de toets gedaan hebben. We kunnen dus onder andere onze gevonden resultaten binnen een groep gaan verklaren of toelichten.

We hebben nu telkens 1 kolom bekeken of 2 kolommen met elkaar vergeleken, maar wat als we nu conclusies willen gaan trekken op meer dan 2 kolommen, voor mensen wordt dat vrij lastig vergelijken. We zijn in assignment 15 tot 18 dus gaan kijken naar decisiontrees die op basis van machine learning gaan voorspellen wat een kolom gaat zijn op basis van meerdere andere kolommen. We wilden natuurlijk wel weten of ons model klopt met de werkelijkheid, hiervoor hebben we onze dataset gesplitst in een test en een train set. Dit gedeelte van data science overlapt een beetje met AI maar zorgt ervoor dat we verbanden kunnen vinden die menselijk niet gevonden zouden kunnen worden. Dit leert ons over data science dat het erom gaat dat we verbanden vinden in data hierbij kunnen we machine learning gebruiken omdat dit heel snel en goed verbanden kan vinden die we zelf niet kunnen zien.

Als laatste assignment hebben we gekeken naar clustering op basis van de k-means. Hiermee kunnen we ontdekken of er misschien niet 1 groep in onze populatie zit, maar meerdere groepen, door deze op te splitsen kunnen onze conclusies nauwkeuriger worden.

Assignment 102

Een Data Scientist gaat als hij een dataset wilt gaan onderzoeken dus eerst kijken naar alle kolommen op zichzelf, een univariate analyse dus die we uitgevoerd hebben in assignment 3. Hierbij kan hij gebruik maken van gemiddelde, mediaan en standaardafwijking maar ook van frequency tables om de data visueel te maken. Maar het is ook belangrijk om bar- en boxplots te maken om ervoor te zorgen dat de Data Scientist geen bijzonderheden in de data over het hoofd ziet. Dit hebben wij gedaan in assignment 5 en 6. En om zeker te weten dat onze sample betrouwbaar is hebben we in assignment7 gekeken naar confidence intervals.

Hierna gaat hij kijken naar de verbanden tussen kolommen, een bivariate analyse dus, dit kan numerical vs numerical zijn maar ook met categorical data vs numerical data of zelfs categorical vs categorical zijn. Dit hebben we uitgevoerd door middel van simpele barplots waardoor de verbanden snel te zien waren en niet verloren raakten in de ingewikkelde grafieken.

Als hij nog verder in wilt gaan op de data kan hij ook nog decisiontrees gebruiken om een multivariate analyse uit te voeren en misschien zelfs voorspellingen gaat doen voor meer data. Dit hebben wij uitgevoerd in assignment 15-18 door middel van classification en regression.

Uiteindelijk gaat hij de data misschien nog clusteren om van specifiekere groepen nauwkeurigere data te krijgen. Dit hebben wij gedaan in assignment 19.