Portfolio assignment 102

# Wat zijn de taken van een data scientist?

Er zijn een aantal vaste taken die een data scientist uitvoert. Deze taken zullen hieronder kort behandeld worden.

**Gegevensverzameling**Het verzamelen van gegevens kan voor een data scientist veel verschillende vormen aannemen. Zo moeten er soms handmatig enquêtes uit worden gevoerd om data binnen te halen, of kan de data simpelweg uit een database worden gehaald.   
Voor de assignments zijn er twee datasets gebruikt. Eén dataset, genaamd ‘penguins’, kan worden aangeroepen door de library van sns te importeren in het project. De andere dataset die is gebruikt voor de assignments heet ’Statbunker Football Statistics’. Deze dataset is gevonden middels google datasetsearch en bestond oorspronkelijk uit 42 verschillende kleine datasets. De datasets bestonden uit verschillende soort statistieken, zoals de spelersstatistieken, publiekopkomst en verdedigende en aanvallende statistieken van teams. Daarnaast waren de datasets ook verdeeld over verschillende jaren (voetbalseizoenen).   
Eerst moest er worden bepaald welk van de dataset(s) interessant genoeg waren. Al snel kwam er uit dat de spelersstatistieken de meest interessante dataset was. Alleen waren er spelersstatistieken van alle seizoenen van 2014 tot 2020, wat betekent dat er nog zes datasets over waren. Deze datasets moesten samengevoegd worden om er één grote dataset van te kunnen maken.

**Gegevensvoorbereiding**Het voorbereiden van de gegevens slaat op het transformeren, samenvoegen en schoonmaken van de gegevens, om er zo voor te zorgen dat er een nette en consistente dataset uitkomt. Bij transformeren kan er worden gedacht aan het aanpassen van datatypes en het veranderen van de bestandstype van de dataset. Bij samenvoegen gaat het over het combineren van gegevens om zo betere inzichten te kunnen geven, of om het combineren van verschillende datasets. En schoonmaken kan gaan over het verwijderen van dubbele data en het opvullen van NaN waardes. Niet alle onderdelen van de voorbereiding hoeven altijd uitgevoerd te worden, voor de ’Statbunker Football Statistics’ dataset, was dat wel het geval.   
Er is enige voorbereiding aan vooraf gegaan voordat deze dataset bruikbaar was. De eerste stap die is genomen was het samenvoegen van de zes overgebleven bestanden, dit is gedaan via de ‘cmd’ en heeft het gewenste resultaat opgeleverd van één grote dataset.   
Door het samenvoegen van deze bestanden kon Python Notebook niet meer identificeren wat de datatypes van de kolommen waren, dit zorgde voor problemen. Om deze problemen op te lossen moest er worden getransformeerd, door alle datatypes handmatig aan te geven.   
Voor een aantal assignments moesten er berekeningen worden toegepast op de datasets waarbij er geen NaN waardes mochten zitten in de data. Een voorbeeld van zo’n opdracht is bij machine learning. Door deze opdrachten moest de data worden schoongemaakt om zo alle NaN waardes er uit te halen. Soms word er voor gekozen om altijd alle NaN waardes uit de dataset te halen, deze keuze is niet gemaakt.

**Gegevensanalyse**Het goed klaar zetten van de data is erg belangrijk maar er uiteindelijk iets nuttigs mee kunnen doen, is nog veel belangrijker. De gegevensanalyse is het gebruiken van statistische technieken om inzichten uit gegevens te kunnen halen. Dit kan worden gedaan door trends en patronen te herkennen en door relaties tussen verschillende gegevens te spotten.   
Op de ’Statbunker Football Statistics’ dataset zijn er verschillende analyses toegepast op de data.

1. Univariate analysis

Analyse van één enkele variabele. In de assignments is dit aan bod gekomen in assignment 3, 5 en 6. Hierin is er voornamelijk gebruik gemaakt van histogrammen en box plotten om verschillende data inzichtelijk te maken.

1. Bivariate analysis

Analyse van twee variabelen. Deze analyse zoekt naar verbanden, hierdoor kan dus worden bepaald wat de relatie is tussen de twee variabelen. In de assignments 9 tot en met 14 is deze analyse aan bod gekomen. In de assignments is gebruik gemaakt van functies als correlatie en verschillende soorten vormgeving zoals ‘bar plotten’ en ’contingency tables’.

1. Multivariate analysis

Analyse van meer dan twee variabelen. Ook deze analyse zoekt, net als de bivariate, naar verbanden en zoekt relaties tussen de variabelen, alleen dan over meer variabelen. In de assignments 15 tot en met 18 is deze vorm van analyse te herkennen.