# Wat is het process van een data scientist

Volgens Berkely houdt datascience het volgende in:

“Effective data scientists are able to

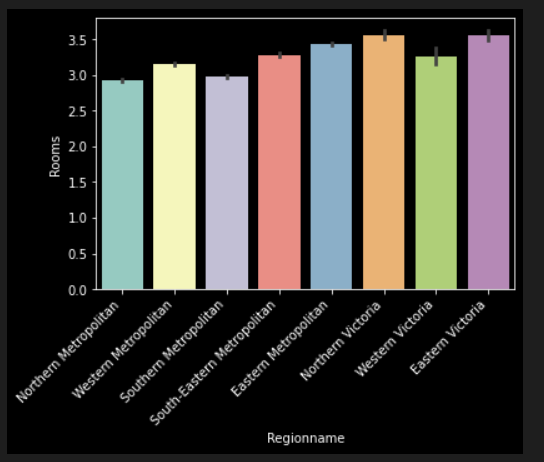
* identify relevant questions,
* collect data from a multitude of different data sources,
* organize the information,
* translate results into solutions,
* and communicate their findings in a way that positively affects business decisions.”   
  - Berkeley

De definitie van Berkely geeft inzicht in het process waar een data scientist doorheen gaat en deze stappen zijn te vinden in de opdrachten in dit portfolio.

Ik zal de stappen laten zien door middel van opdracht 12.

De eerste stap in de definitie van Berkely is “identify relevant questions”. Omdat we in ons geval vooral oefenen met data science en de motivatie dus leren is i.p.v het beantwoorden van specifieke vragen van organisaties, is deze stap minder duidelijk terug te vinden in de opdrachten. In het geval van opdracht 12 zou je kunnen zeggen dat de vraag als volgt is: “Welke variabelen hebben de meeste impact op de prijs van een huis dat te koop staat in Melbourne”.

Na het opzetten van een vraag, volgen stap 2: “collect data from a multitude of different data sources” en stap 3: “organize the information”. Omdat deze vraag beantwoord kan worden met 1 bron, gebruiken we hier maar 1 bron. In een realistischer scenario zouden de vragen complexer zijn en zouden er dus waarschijnlijk meer bronnen nodig zijn. In ons geval is de informatie ook al goed georganiseerd. Bij de dataset in opdracht 6, 7 en 8 was de dataset minder georganiseerd en heb ik best een aantal transformaties en filters gebruikt om bruikbare data uit de dataset te halen.



Als de data klaarstaan volgt mogelijk de belangrijkste stap: “translate results into solutions”. In ons geval analyseren we de verschillen tussen verschillen categorieën en numerieke variabelen, bijvoorbeeld het gemiddeld aantal kamers per te koop staand huis per regio. In de grafiek kunnen we zien dat deze verschillen niet heel groot zijn. Omdat we niet echt een “probleem” hadden is het niet mogelijk om een oplossing te geven.

Dan de laatste stap: “… and communicate their findings in a way that positively affects business decisions.” Omdat we in deze casus geen bedrijf of organisatie hebben, is het niet mogelijk om deze stap te ondernemen. Het is natuurlijk wel een belangrijk onderdeel van het werk van een datascientist.