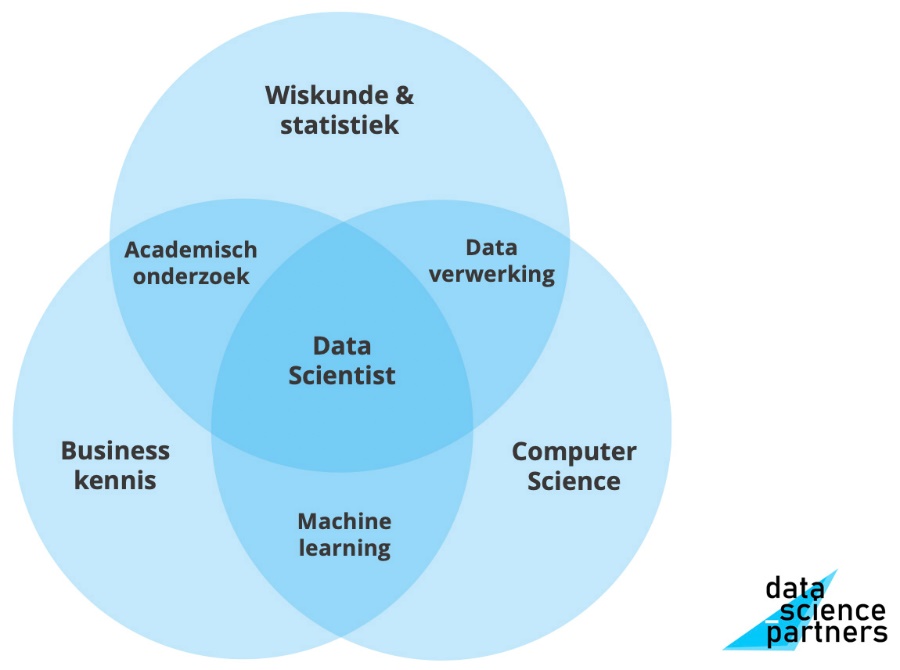
Intro data science

Assignment 101 & 102

Naam: Niels Kloos

Studentnummer: 2159340

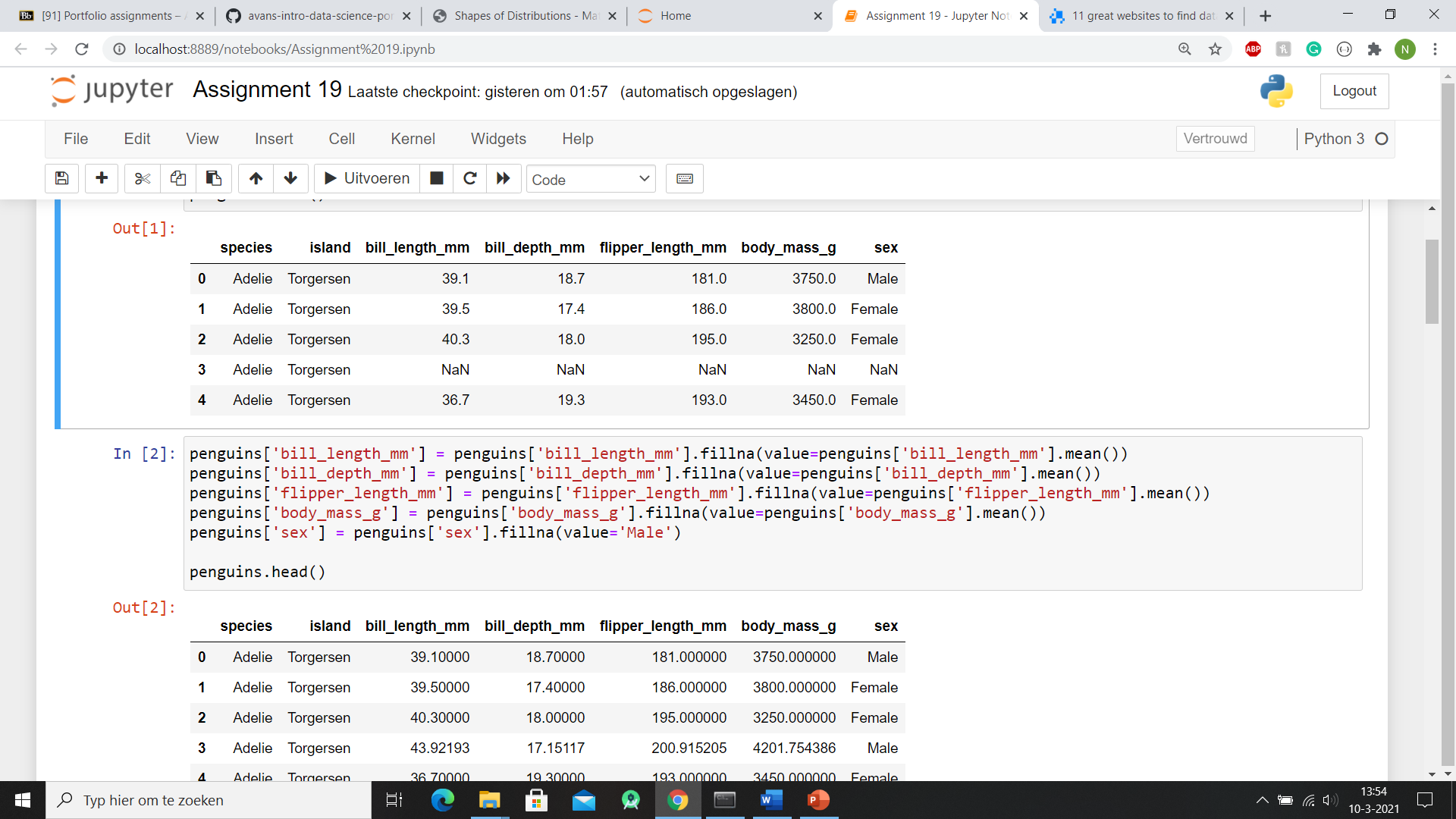
# Assignment 101

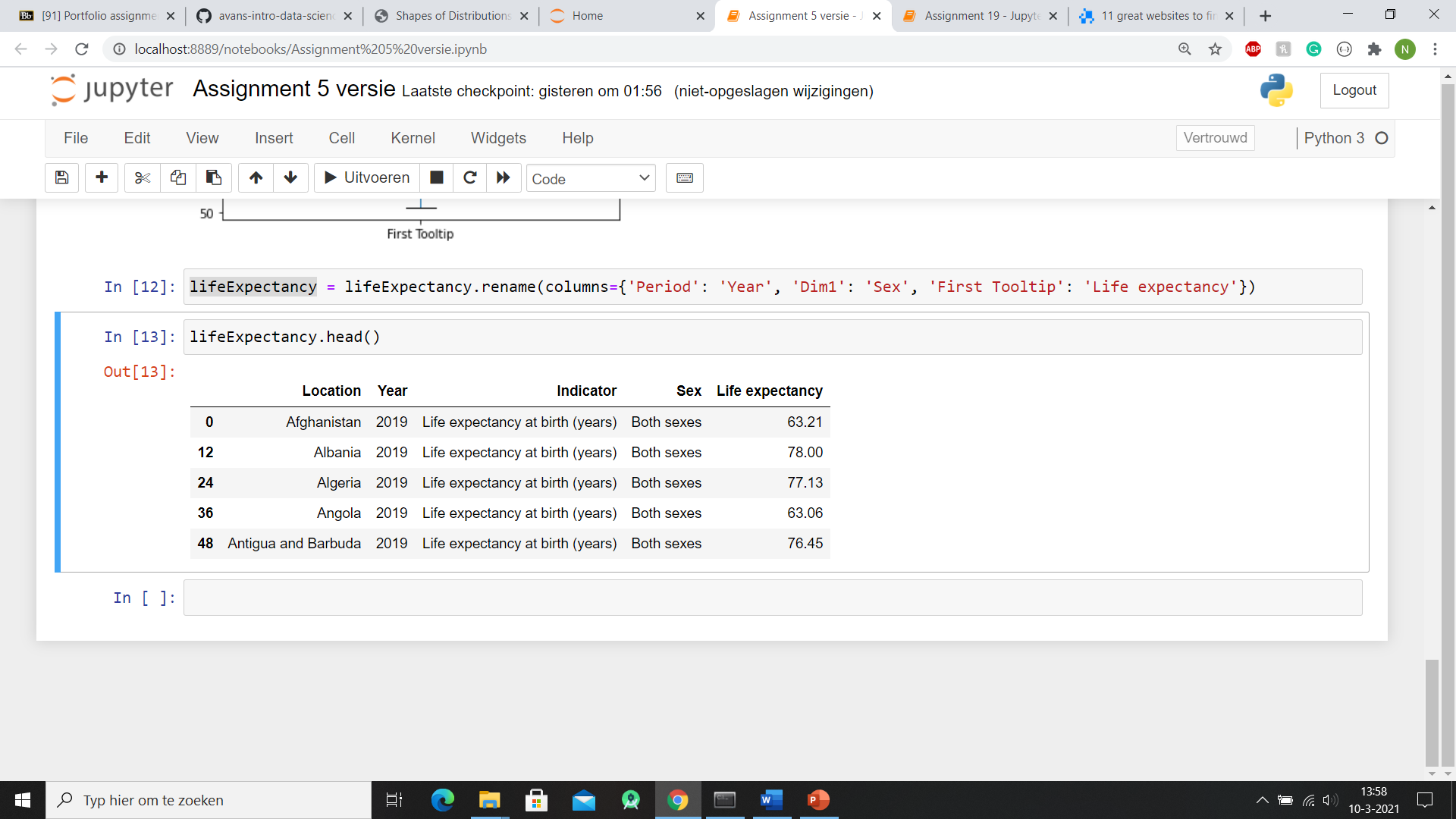
Data science is een vakgebied dat gericht is op het verkrijgen van inzichten uit data, dit wordt gedaan door middel van 3 verschillende vakgebieden: wiskunde & statistiek, computer science en business kennis.

De drie vakgebieden komen ook terug in de portfolio assignments die we maken. Zo gebruiken wij Pandas voor het analyseren van de data en Seaborn voor het visualiseren van de data en maken we gebruik van SciPy en scikit-learn voor het gebruik van machine learning en wiskundige berekeningen laten doen.

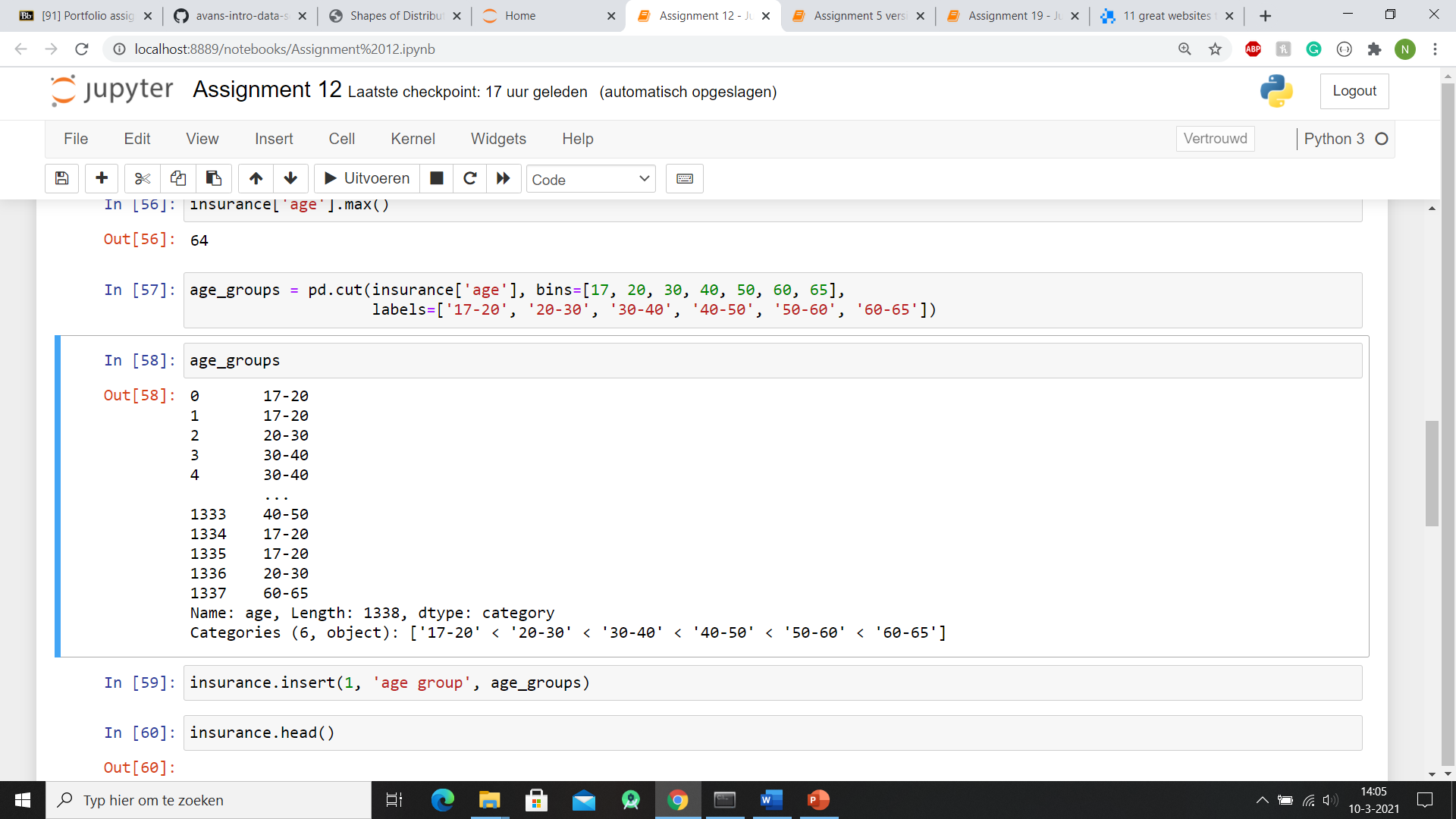
# Assignment 102

## Analyseren van data en het vergaren van data

Dit zie je in elke portfolio wel terug. Je haalt data op uit een bron zoals Seaborn of een andere dataset die je kan vinden via het sites zoals: Kaggle en Google Public Datasets. Als je dan een dataset hebt zal die niet altijd meteen klaar zijn voor gebruik. Het kan zien dat er waarde missen, die waarde moet je toevoegen, want niet alle machine learning technieken kunnen omgaan met legen waarde. Dan moet je de rijen verwijderen die leeg zijn of wat ik heb gedaan het gemiddelde toevoegen aan de legen waarde. 

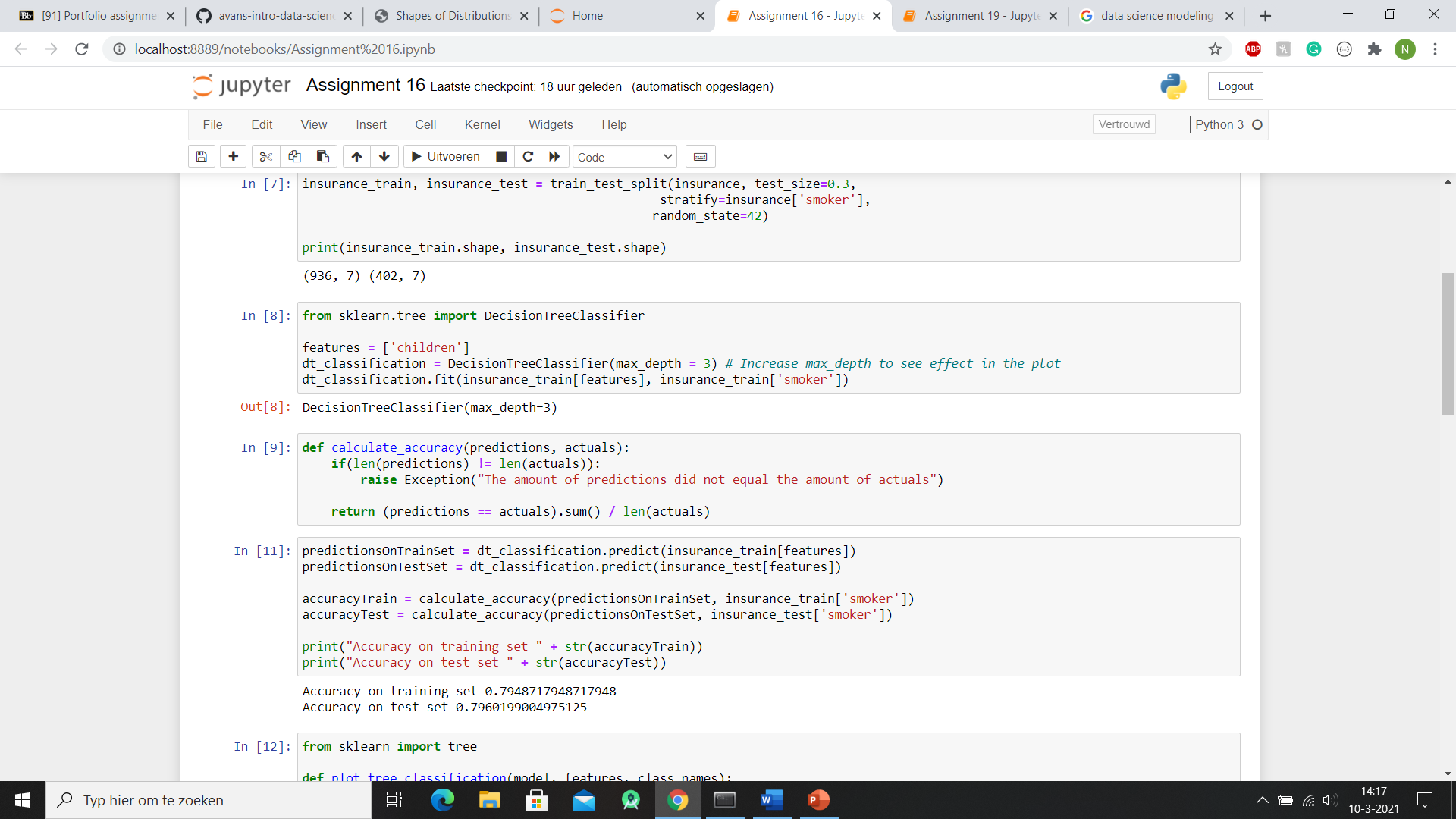
Het kan ook zijn dat de kolommen slechte naamgeving hebben, waardoor het voor een andere gebruiker onduidelijk is wat het betekend. 

Bij assignment 5 zijn de kolomnamen slecht gekozen. Het onduidelijk wat er met first tooltip bedoeld wordt daarom heb ik het verandert in life expectancy dat maakt het veel duidelijker wat er met die kolom bedoeld wordt.

Als laatste kan je data ook nog opdelen in groepen zoals leeftijdsgroepen dat heb ik gedaan in assignment 12. Ik heb de groepen ingedeeld in leeftijd groepen van 17-20, 20-30, 30-40, 40-50, 50-60 en 60-65. 

## Modeling

Dit is iets wat we vooral de laatste paar assignment deden. We moesten in assignment 16 met onze eigen dataset een decision tree trainen waar 70% train is een 30% test. Door een decision tree te trainen kan je data voorspellen door het gebruik van andere data.



In Assignment 17 doen we hetzelfde maar dan gebruiken we nummers met behulp van regression en RMSE. Zo wordt er berekend wat de gemiddelde afwijking is de daadwerkelijke waarden en de waarden die voorspeld waren.

