## Proces Opdracht 1

### Oriënteren

We zijn begonnen met het inlezen van de dataset in een pandas DataFrame. Daarna zijn we ons gaan oriënteren op de dataset. Zo hebben we even gauw enkele kolommen uit de data gehaald om hiermee te clusteren en te kijken wat hiermee gedaan wordt. Verder hebben we uit diverse kolommen gekeken wat de minimale en de maximale waarden zijn om zo eventuele uitschieters tegen te komen. Door het plotten van boxplotten wordt dit meteen visueel duidelijk. Zo kwamen we een grote uitschieter tegen bij de ‘Warm lentil salad with Serran chicken and rocket’ die 2012 carbs zou bevatten. Ons leek dit erg onwaarschijnlijk. Daarom hebben we besloten eventuele andere outliers ook uit de dataset te gooien m.b.v. van een simpele definitie.

De stappen die ondernomen moeten worden om de originele data te kunnen clusteren zijn als volgt:

* Data cleaning uitvoeren
* Data pre-processen
* Data clusteren
* Cluster valideren
* Feature selection

### Data Cleaning

Hierna zijn we verder gegaan met het cleanen van de data. Hiervoor hebben we de volgende stappen proberen uit te voeren.

* Uitschieters verwijderen

Het verwijderen van de uitschieters is niet correct gegaan, omdat het niet daadwerkelijk de corresponderende rij verwijdert, waar het de uitschieter vindt. Hierdoor krijg je wel een list terug waarin alle waardes zitten van de niet-uitschieters. Maar hier kun je dus niets mee, omdat je deze list niet zomaar terug kunt zetten in de dataset, omdat de indexen dan niet meer overeen komen.

* Alle rijen met missende NaN waardes eruit gooien
* Alle rijen met dubbele waardes eruit filteren
* Onnodige waarden zoals alle nutritional values eruit halen
* Alle (engelse) stopwoorden uit de tekstuele data filteren

Wat we verder nog geprobeerd hebben is om een stemming functie toe te passen op alle tekstuele data. Helaas is ons dit niet gelukt doordat we te maken kregen met diverse onbekende errors. Hierdoor hebben we besloten dit voor nu achterwege te laten.

### Pre-processing

Om geschikte features uit de opgeschoonde data te verkrijgen, moet deze data ge-pre-processed worden. Hiervoor hebben we de volgende stappen ondernomen

* Voor de kolommen met (variërende) tekstuele data hebben we een bag of words gecreëerd en hierna een tfidf (term frequency – inverse document frequency)
* Voor de kolom met skill\_level zijn de tekstuele categorieën omgezet naar categorieën met integers

### Clusteren

We hebben de dataset geclusterd met behulp van het K means model. Hierbij hebben we gekozen voor 3 clusters. De features die we gebruikt hebben om te clusteren zijn alle kolommen die in de vorige stap ge-pre-processed zijn. Zo hebben wij de tfidf van de kolommen title, description, ingredients en method gebruikt. Verder hebben we de omgezetten kolom skill\_level ook gebruikt als feature.

### Clusters interpreteren

Om de clusters te interpreteren hebben wij de labels vergeleken met de corresponderende titels van de gerechten. Helaas kwam hier niet een beduidend resultaat uit. Wat ons echter wel opviel, was dat in ons eerste cluster verassend veel gerechten voorkwamen die chocola of iets dergelijks bevatten. Echter wil dit niet zeggen dat in de andere twee clusters geen enkel gerecht voorkwam dat chocola bevatte.

## Proces Opdracht 2

Bij deze opdracht hebben we het pre-processen van de data hetzelfde aangepakt als bij opdracht 1. De feature om te voorspellen hebben we uit descriptions gehaald. Daarmee gaan we de rating voorspellen. Dit hebben we gedaan met lineare regressie omdat het een continue variabele is.

Om te voorkomen dat hier over-fitting bij plaatsvindt omdat we meer features hebben dan datapunten, hebben we een f-test toegepast. Zo worden de features die het beste het model fitten meegenomen in de regressie.

We hebben meerdere regressie geprobeerd, waaronder Logistische regressie en Multinomial Naive Bayes. Deze werkten niet omdat we te maken hebben met continue waarden in de labels. Hierbij is nog geprobeerd om de y in onze data discreet te maken en er vervolgens een test op uit te voeren. De scheiding is gemaakt voor de ratings van 0-2.5 en 2.5-5. Omdat veel waardes boven de 2.5 liggen is dit geen goede maat. De rating zijn onevenredig verdeeld waardoor het niet zo veel nut heeft om ze op deze manier te scheiden.