Proces

We zijn begonnen met het inlezen van de data in een panda data frame.

Hier hebben we eerst verschillende kolommen uit gehaald om te kijken hoe python deze zou clusteren.

We willen kmeans gebruiken omdat dit een snelle manier van clusteren is.

Grote outlier gevonden in de dataset. Warm lentil salad with Serrano chicken en rocket heeft 2012 carbs wat ons erg onwaarschijnlijk lijkt. Daarom hebben we besloten de andere outliers ook uit de dataset te gooien.

4/4/2018

Alle missende waardes nan eruit gegooid.

Proberen dubbele waardes eruit te filteren door de functie data.duplicate.

De nutritions uit de dataset verwijderen