# Analyse calorieën van verschillende voedingsproducten

Naam student: Jannie Janssen

Studentnummer: 123456

Datum: 2025-04-10

## Inleiding

In de huidige maatschappij is er een wijdverspreide bewustwording ten aanzien van voedselconsumptie. De kwantificering van calorische inname en de reductie van lipide consumptie vormen gangbare aanbevelingen van diëtisten en voedingswetenschappers. Daarom is kennis omtrent de nutriëntensamenstelling van diverse voedingsmiddelen van evident belang. De in dit rapport geanalyseerde dataset omvat een comma-separated values (CSV) bestand met gegevens van meer dan 300 voedingsmiddelen, waarbij voor elk item de kwantiteit van calorieën, lipiden, proteïnen, gesatureerde vetzuren, koolhydraten en voedingsvezels is geëtiketteerd. Bovendien zijn de voedingsmiddelen geclassificeerd in verschillende categorieën, zoals desserts, groenten, fruit, enzovoort.

## Dataset

De dataset komt van Kaggle ([Nutritional Facts for most common foods](https://www.kaggle.com/datasets/niharika41298/nutrition-details-for-most-common-foods)). De dataset gebruikt Wikipedia als referentie ([Table of food nutrients - Wikipedia](https://en.wikipedia.org/wiki/Table_of_food_nutrients)).

## Doel

Het primaire doel van de data-analyse van deze dataset is het kwantificeren en vergelijken van het caloriegehalte tussen verschillende voedselgroepen en binnen individuele voedselgroepen.

Specifiek beoogt de analyse de volgende inzichten te verschaffen:

* Identificatie van de voedselgroep(en) met het gemiddeld hoogste caloriegehalte. Dit stelt ons in staat om te bepalen welke categorieën van voedingsmiddelen over het algemeen de grootste bijdrage leveren aan de totale energie-inname.
* Identificatie van het specifieke voedingsmiddel binnen de voedselgroep met de hoogste gemiddelde hoeveelheid calorieën. Dit pinpoint individuele voedselitems die potentieel significant bijdragen aan een hoge calorieconsumptie.
* Identificatie van het item binnen de hele dataset met het hoogste aantal calorieën.
* Markeren van items met een hogere calorische hoeveelheid dan de mediaan.

## Inladen van de data

Allereerst is de data bekeken in Visual studio code. Hieronder is een screenshot te zien.

!A screen shot of a computer program

AI-generated content may be incorrect.

Figuur 1: analyse van csv file in Visual Studio.

Aan de file te zien betreft het een csv file met de komma als kolom separator. Decimalen gesepareerd met een punt.

Ik heb daarna de file ingeladen in Excel via Power Query:

A screenshot of a computer

AI-generated content may be incorrect.

Figuur 2. Inladen data via Power Query.

Vervolgens heb ik de tabel selectie naam veranderd (Formulas > Name manager) naar “data” om latere selecties makkelijker te maken.

A screenshot of a computer

AI-generated content may be incorrect.

Figuur 3. Name manager. Selectie naam veranderd.

## Cleaning

Allereerst heb ik gekeken of er lege waarden zijn in de kolom Calories:

Ik heb hiervoor de functie COUNTBLANC gebruikt.

Er bleek 1 missende waarde te zijn.

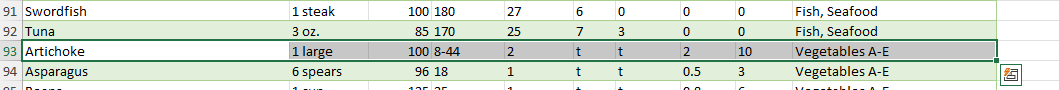
Met XLOOPUP heb ik deze geïdentificeerd:

A white grid with black text

AI-generated content may be incorrect.

Figuur 4. Missende waarde.

Een inspectie van de data laat zien dat er bij sommige rijen data-ranges gebruikt zijn:



Figuur 5. Data range in plaats van waarde.

Ik heb daarom eerst gefilterd en deze waarden verwijderd.



Figuur 6. Lege waarde voor Frozen peas.

Frozen peas bleek een lege waarde voor deze kolom te hebben. Deze rij is ook verwijderd.

Ik heb deze data uit het bronbestand (csv) verwijderd en opnieuw ingeladen zodat de calorieën niet meer als range geïnterpreteerd worden:

A screenshot of a computer

AI-generated content may be incorrect.

Figuur 7. Opgeschoonde data opnieuw ingeladen.

Met deze opgeschoonde data is nu een pivot chart te maken.

## Identificatie van de voedselgroep(en) met het gemiddeld hoogste en laagste caloriegehalte

Om voedselgroepen te identificeren met het hoogste vetgehalte heb ik een draaitabel grafiek gemaakt:

A graph of food categories

AI-generated content may be incorrect.

Figuur 8. Draaitabel grafiek van calorieën uitgezet tegen categorieën.

Te zien in deze grafiek is dat zaden en noten gemiddeld het hoogste en groenten gemiddeld het laagste aantal calorieën bevatten.

Om wat meer inzicht te krijgen in de spreiding van de data heb ik ook een Boxplot gemaakt:

A graph with blue squares and black text

AI-generated content may be incorrect.

Figuur 9. Boxplot van calorieën in verschillende voedingscategorieën.

Ook hierbij is te zien dat zaden en noten het hoogst uitkomen (hoogste mediaan). Brood en ontbijt laat een grote spreiding zien.

## Identificatie van het specifieke voedingsmiddel binnen de voedselgroep met de hoogste gemiddelde hoeveelheid calorieën

Om voedingsmiddelen te identificeren met het hoogste aantal calorieën binnen de hoogste gemiddelde categorie heb ik eerst een barplot gemaakt van alle items:

A graph with blue lines and white text

AI-generated content may be incorrect.

Figuur 10. Barplot van calorieën versus alle items.

En daarna gefilterd op seeds and nuts:

A graph of calories per item

AI-generated content may be incorrect.

Figuur 11. Seeds and nuts.

Hieruit blijkt dat Brazil nuts en Roasted and salted als hoogste calorieen geklassificeerd kunnen worden.

Ik heb vervolgens ook een taartdiagram deze analyse gemaakt:

A pie chart with numbers and a number of calories per item

AI-generated content may be incorrect.

Figuur 12. PIE chart van de data.

## Identificatie van het item binnen de hele dataset met het hoogste aantal calorieën.

Omdat er eerst per categorie gekeken is, zou het kunnen dat er een individueel item is dat een zeer hoge categorische waarde heeft maar niet in dezelfde categorie valt als de items die gemiddeld de hoogste categorische waarde hebben.

Daarom is er ook een analyse gedaan op de hoogste waarde en de laagste waarde.

Ik heb hiervoor de functie MIN, MAX en XLOOPUP gebruikt. De volgende resultaten werden hierbij verkregen:

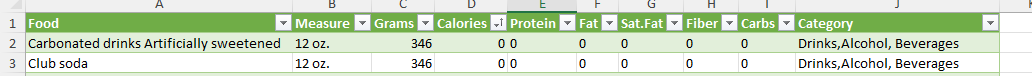
A screenshot of a graph

AI-generated content may be incorrect.

Figuur 13. Identificatie van voedingsmiddel met max calorieën met XLOOKUP

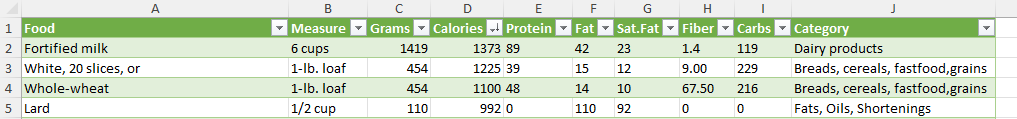
Deze werden gecontroleerd door te sorteren:

Er werden 2 items gevonden met 0 calorieën:



Figuur 14. Gesorteerd op hoeveelheid calorieën. Minimum.

En inderdaad werd hetzelfde item gevonden voor het maximale aantal calorieën:



Figuur 15. Gesorteerd op hoeveelheid calorieën. Max aantal.

## Markeren van items met een hogere calorische hoeveelheid dan de mediaan.

Om te markeren welke voedingsmiddelen een hoge calorische waarden hebben, heb ik de mediaan gebruikt. De mediaan is robuuster dan het gemiddelde en minder gevoelig voor een individuele uitbijter. Ik heb eerst de mediaan berekend van het aantal calorieën. Daarnaast geteld hoeveel van de rijen groter zijn:

A white grid with black text

AI-generated content may be incorrect.

Figuur 16. Mediaan van aantal calorieën en aantal rijen dat hoger is dan de mediaan.

Vervolgens heb ik deze waarde gebruikt om met Conditional Formatting aan te geven welke items een hoge calorische waarde hebben. De dataset is alfabetisch gesorteerd op voedselitems.

A screenshot of a computer

AI-generated content may be incorrect.

Figuur 17. Conditional formatting voor markeren hoog calorische items.

## Conclusie en discussie:

De analyse van de caloriegehalten in meer dan 300 voedingsmiddelen, gecategoriseerd in diverse groepen, onthult significante variaties. Zaden en noten werden geïdentificeerd als de voedselgroep met het gemiddeld hoogste aantal calorieën, terwijl groenten gemiddeld het laagste caloriegehalte vertoonden. Binnen de categorie zaden en noten bleken Brazil nuts en geroosterde, gezouten noten de hoogste individuele calorieën te bevatten. Opvallend was de identificatie van pure plantaardige olie als het voedingsmiddel met het absolute maximum aan calorieën in de gehele dataset. Daarentegen werden er twee voedingsmiddelen met een caloriegehalte van nul geconstateerd. Door de mediaan van het caloriegehalte te berekenen en toe te passen voor conditionele formattering, werden de voedingsmiddelen met een relatief hoog caloriegehalte visueel gemarkeerd in de dataset. Deze bevindingen benadrukken het brede scala aan calorieën binnen verschillende voedselgroepen en individuele producten. De resultaten kunnen waardevol zijn voor consumenten die hun calorie-inname willen monitoren, evenals voor professionals in de voedingsindustrie en gezondheidszorg. De grote spreiding binnen sommige categorieën, zoals brood en ontbijtgranen, suggereert dat verdere, meer gedetailleerde analyses binnen deze groepen nuttig kunnen zijn.

## Verantwoording Checklist:

A screenshot of a computer

AI-generated content may be incorrect.

Te zien is dat er nog twee elementen in het portfolio ontbreken:

* Clustered barplot
* XY-scatterplot