**Data understanding:**

**Inleiding:**

Deze fase richt zich op het begrijpen en analyseren van de beschikbare datasets binnen de 4FOOD-database. Data understanding is een cruciale stap, omdat de kwaliteit, structuur en relevantie van de data direct bepalen hoe betrouwbaar de uiteindelijke analyses en modellen zijn.

Met dit document stel ik vast:

* Welke variabelen bruikbaar zijn voor de analyse.
* Welke datakwaliteitsproblemen aanwezig zijn (zoals missende of foutieve waarden en outliers).
* Welke relaties en patronen in de data kunnen worden benut voor verdere modellering.

**Data Verzameling:**

De data is verzameld uit verschillende SQL-databases binnen de 4FOOD-Server. Via Python (pandas) en SQL-queries zijn de datasets ingelezen en verkend. Hiermee konden statistieken, datakwaliteit en patronen efficiënt in kaart worden gebracht.

Belangrijkste aandachtspunten bij de verzameling:

* Data ingelezen via SQL-verbinding met Python.
* Database-informatie beveiligd opgeslagen met een .env-bestand.
* Focus op datasets die relevant zijn voor analyse (irrelevante tabellen buiten scope gelaten).

**Data Beschrijving:**

**4FOOD.Matthijs.JJO**  
Omdat ik al bekend was met de 4FOOD-applicatie, heb ik deze database als eerste bekeken. Hierbij heb ik de tabel *orderinfo* geïdentificeerd als relevant. Op basis van deze tabel is een visualisatie gemaakt die de verdeling van orders per jaar en maand weergeeft.A graph with blue and orange lines

AI-generated content may be incorrect.

Periode weergegeven (exclusief 2025): 2017-05-01 t/m 2024-12-01

Beperkingen: er zijn geen prognosegegevens aanwezig, waardoor deze database minder geschikt is voor vergelijking met voorspelde waarden.

**4food.napoleon.productie**  
Deze database bleek omvangrijker en bevatte veel tabellen die niet relevant waren voor de analyse. Drie tabellen bleken wél bruikbaar:

* **Jaarschema:** bevat prognoses van de afgelopen 16 jaar, met onder meer de kolommen *jaar*, *maand* en *aantalhe*. Deze zijn gebruikt om trends per jaar en maand inzichtelijk te maken.
* **Orderinfo:** bevat de realisatiegegevens van de afgelopen 16 jaar, met onder meer de kolommen *orderDate*, *orderNumber, orderstatus, ID* en *clientId*. Elke orderinfo kan meerdere Opb-records bevatten.

Daarnaast bevat de tabel orderinfo een kolom ordertype, die essentieel is voor het bepalen van welke orders daadwerkelijk relevant zijn voor analyse. Alleen de volgende ordertypes zijn meegenomen:

**2** = definitief

**3** = backorder

**4** = afroep

**18** = retourzending

Andere waardes zoals **0** (concept), **1** (forecast), **5** (proforma), en **16–21** zijn uitgesloten, omdat deze geen representatieve realisatie-orders vormen.

* **Opb:** bevat de daadwerkelijke waarden van de realisatie, zoals gewicht en aantal, die niet in **orderinfo** zijn opgenomen. Belangrijke kolommen zijn *orderdatum*, *quantity*, *priceKG, atikelnummer, status,* en *tuWeight*. De koppeling met orderinfo verloopt via de kolom *orderInfoId*, die aansluit op de *id*-kolom in de orderinfo-tabel.

**Sterk punt:** bevat zowel prognoses als realisaties van 16 jaar - geschikt voor vergelijking.

**Beperkingen:** prognosegegevens zijn slechts beschikbaar voor één klant, waardoor analyses op klantniveau beperkt zijn.

**4food.pervasco.productie**

Uit deze database is de tabel *jaarschema* gebruikt. Deze bevat uitsluitend prognosegegevens, beginnend in 2016 en lopend tot en met 2026, goed voor ongeveer 10 jaar aan gegevens. De belangrijkste kolommen zijn *jaar*, *maand, klant* en *aantalhe*.

Hoewel deze database geen realisatiegegevens bevat, is het wel mogelijk om de prognoses te vergelijken met de gerealiseerde waarden uit andere databases, zoals 4food.napoleon.productie. Dit maakt de dataset bruikbaar voor het analyseren van verschillen tussen planning en realisatie over meerdere jaren.

**Sterke punten:** geschikt voor langetermijnanalyses en bruikbaar voor klantanalyses dankzij de aanwezigheid van klantgegevens.

**Beperkingen:** bevat uitsluitend prognoses en geen realisatiegegevens, waardoor op zichzelf geen validatie of controle mogelijk is.

**Data verkenning:**

Tijdens de verkenning zijn diverse analyses en visualisaties uitgevoerd:

A graph of blue bars

AI-generated content may be incorrect.

Jaarlijkse aantallen orders (orderinfo) uit database *4food.napoleon.productie.*

**A graph of blue bars

AI-generated content may be incorrect.Jaarlijkse aantallen bestelde producten** in aantal handelseenheden voor vergelijking prognose (orderinfo/opb) uit database *4food.napoleon.productie.*

A graph with blue lines and orange lines

AI-generated content may be incorrect.

A graph of blue bars

AI-generated content may be incorrect.

**Totaal in prognoses per jaar en maand** (*jaarschema,* kolom *aantalhe*) uit database *4food.napoleon.productie*.

**A graph with numbers and a bar

AI-generated content may be incorrect.**

**Totaal in prognoses per jaar en maand** (*jaarschema,* kolom *aantalhe*) uit database *4food.pervasco.productie*.

A graph with blue and orange bars

AI-generated content may be incorrect.

**Realsisaties vs Prognoses per jaar Napoleon**

(Realisaties = quantity/tuWeight uit *opb en orderinfo*)

(Prognoses = aantal\_he uit jaarschema) beide komen uit *4food.napoleon.productie.*

*A graph with orange bars

AI-generated content may be incorrect.*

**Prognoses per jaar Pervasco**

(Prognoses = aantal\_he uit jaarschema)En prognose komt uit *4food.pervasco.productie.*

A graph with lines and numbers

AI-generated content may be incorrect.

**Trends over de maanden** (jaarschema, kolom *aantalhe*)

**A graph with blue and white bars

AI-generated content may be incorrect.**

**Verdeling van quantity** uit *db.orderinfoNapoleon*

A graph with blue squares

AI-generated content may be incorrect.

**Overzicht van top 6 klanten uit prognose**

A graph with blue squares

AI-generated content may be incorrect.

**Overzicht top producten uit prognoses**

Daarna heb ik gekeken naar de statistical summery’s om verder inzicht te krijgen in hoe dat data is verdeeld:

Orderinfo/dbo:

A screen shot of a computer

AI-generated content may be incorrect.

In de dataset zijn nulwaarden aangetroffen op plaatsen waar deze niet plausibel zijn. En dat er outliers aanwezig zijn en waarschijnlijk incorrecte data. Tijdens de verkenning bleek dat de variabele ordertype een duidelijk onderscheid maakt tussen conceptorders, forecast-orders en daadwerkelijke uitgevoerde orders zoals definitief, backorder, afroep en retour.  
Voor de analyse zijn daarom uitsluitend ordertypes **2, 3, 4 en 18** meegenomen, omdat alleen deze daadwerkelijk gerealiseerde orders representeren.

Prognose/jaarschema:

A screenshot of a computer

AI-generated content may be incorrect.

Hieruit kon ik halen dat er negatieve waarden zijn aanwezig, terwijl deze in de context onmogelijk zijn. En dan de data inderdaad ver uit elkaar ligt dus dat er veel outliers/foute waardes aanwezig zijn.

**Conclusie dataverkenning**  
De verkenning laat zien dat de datasets bruikbaar zijn voor het analyseren van trends en het vergelijken van prognoses met realisaties. Wel zijn er datakwaliteitsproblemen vastgesteld, zoals nulwaarden, negatieve waarden, outliers en geannuleerde orders. Daarnaast zijn gegevens vóór 2007 en in 2025 onvolledig en daarom minder geschikt voor analyse. Opvallend is dat de meeste variabelen rechtsscheef verdeeld zijn, wat wijst op scheve spreidingen en extreme waarden. Ondanks deze beperkingen vormt de combinatie van prognoses en realisaties een sterke basis voor verdere analyses, mits de datavoorbewerking zorgvuldig wordt uitgevoerd.

**Datakwaliteit:**

Orderinfo en opb/Realisaties:

* Nulwaarden aanwezig op plaatsen waar dit niet mogelijk is.
* Outliers (aan de hoge en lage kant) en vermoedelijk incorrecte waarden in de velden *quantity, priceKG, tuWeight*.
* Status 4 = is gecancelled dus die moeten er allemaal uit in data preperation
* Alleen de ordertypes 2 (definitief), 3 (backorder), 4 (afroep) en 18 (retourzending) zijn meegenomen in de analyse de rest moet er uit in data preperation.
* Waarden vóór 2007 zijn schaars en niet representatief → uitgesloten van analyse.
* Waarden van 2025 zijn onvolledig (jaar nog niet afgerond) → uitgesloten van analyse.
* *orderDate* en *orderdatum* zijn bruikbaar als tijdsmaatstaf (zijn bijna precies hetzelfde).

Jaarschema/Prognoses

* Negatieve waarden aangetroffen → ongeldig.
* Grote spreiding en veel outliers → indicatie van foutieve waarden.
* Kolommen jaar en maand moeten meegenomen worden als tijdsmaatstaf (kan op deze manier worden vergeleken met realisatisch.
* Kolom aantalhe, klant(pervasco), jaar en maand zijn de belangrijkste indicatoren voor de prognoses.
* Kolommen als versioninfo zijn niet relevant en worden uitgesloten.
* Kolom klant is relevant is Pervasco database, in Napoleon niet aangezien het maar 1 klant is.
* Tabel artikel biedt inzicht per product

**Belangrijkste inzichten:**

* De realisatiegegevens bevatten outliers, nulwaarden en geannuleerde orders (*status = 4*) die in de datavoorbewerking verwijderd moeten worden.
* Naast filtering op statuscodes is filtering op ordertype noodzakelijk.  
  Hiermee wordt voorkomen dat concepten, forecasts of interne transacties de analyses verstoren of vertekenen.
* Gegevens van vóór 2007 zijn schaars en niet representatief → uitgesloten van de analyse.
* Gegevens van 2025 zijn onvolledig (jaar nog niet afgerond) → uitgesloten van de analyse.
* *orderDate*/*orderdatum* vormen de geschikte tijdsvariabelen voor realisaties; *jaar* en *maand* doen dit bij prognoses.
* In prognoses zijn de kolommen *aantalhe*, *jaar* en *maand* het belangrijkst; irrelevante kolommen zoals *versioninfo* worden uitgesloten.
* De kolom *klant* is alleen bruikbaar in de Pervasco-database (meerdere klanten); in Napoleon voegt deze geen waarde toe.