**Data preparation documentatie**

**1. Doelstelling**

Het doel van deze datavoorbereiding is het creëren van een betrouwbare, opgeschoonde en gestandaardiseerde dataset waarmee gerealiseerde verkoopcijfers (realisaties) vergeleken kunnen worden met de geplande prognoses uit het jaarschema.

De initiële productiedata van de Napoleon-omgeving bevatte veel irrelevante of lege velden, onvolledige jaren en outliers. Door middel van deze voorbereiding wordt de data gezuiverd, samengevoegd en getransformeerd tot een analyse-klaar formaat.

De eindresultaten worden opgeslagen in de tabel dbo.Napoleon\_All binnen de database *4food.aidan*, en kunnen vervolgens worden gebruikt voor verdere analyses en modellen

**2. Bronnen en uitgangsdata**

De data is afkomstig uit verschillende databases binnen de 4food-omgeving:

| **Bron** | **Beschrijving** |
| --- | --- |
| 4food.napoleon.productie.dbo.orderinfo | Algemene informatie over orders (status, type, klant, datums enz.). |
| 4food.napoleon.productie.dbo.opb | Detailinformatie per orderregel (artikel, hoeveelheid, prijs, gewicht enz.). |
| 4food.napoleon.productie.dbo.jaarschema | Prognosedata per artikel per maand. |
| 4food.napoleon.productie.dbo.specificaties, specificationProductGroup, groepen | Metadata over productgroepen en artikelspecificaties. |
| 4food.pervasco.productie | Referentiedatabase voor de Pervasco-prognoses. |

**Verwijderde kolommen**

Tijdens het opschonen zijn vele kolommen verwijderd omdat ze volledig leeg waren, constant dezelfde waarde bevatten of buiten de scope van de analyse vielen. Voorbeelden hiervan zijn:

* createdOn, createdBy, username, externalReference, rowStatus, invoiceNumber — bevatten uitsluitend NULL of onrelevante waarden.
* vatgroupid, onderneming, paymentCondition, versioninfo — bevatten ofwel slechts één vaste waarde of zijn buiten de scope (administratief).
* planProdDate, loadingWarehouseID, deliveryConditionID — te veel ontbrekende waarden of niet relevant voor verkoopanalyse.

De motivatie was om enkel de noodzakelijke velden te behouden die relevant zijn voor tijdreeksanalyse, productvolumes en vergelijking met prognoses.

**3. Opschoning en filtering**

De ruwe data werd verder gefilterd om alleen zinvolle, consistente observaties over te houden. Belangrijkste stappen:

**3.1. Tijdsselectie**

* **Jaren met minder dan 6 maanden aan data** zijn verwijderd, omdat deze onvoldoende representatief zijn.
* Voor de definitieve versies is een strengere drempel van **minimaal 10 maanden** per jaar toegepast, om seizoensinvloeden juist te kunnen interpreteren.

**3.2. Orderstatus**

* **Orderstatus = 4** (geannuleerde orders) is uitgesloten, aangezien deze geen daadwerkelijke productie of levering vertegenwoordigen.

**3.3. Ordertype**

Alleen de volgende ordertypes zijn meegenomen:

* 2, 3, 4: normale en varianten van verkooporders.
* 18: retourzendingen — deze zijn omgezet naar **negatieve waarden** voor Quantity om een correcte balans te behouden in volumeberekeningen.

Deze selectie zorgt ervoor dat alleen echte commerciële bewegingen in de dataset terechtkomen.

**3.4. Nullwaarden en nulhoeveelheden**

Rijen waar Quantity, PriceKG of TuWeight gelijk zijn aan 0 of NULL zijn verwijderd. Deze records bevatten geen kwantitatieve betekenis en zouden gemiddelde of somberekeningen verstoren.

**3.5. Outlierverwijdering**

Om extreem hoge of foutieve waarden te elimineren is gebruikgemaakt van:

PERCENTILE\_CONT(0.01) en PERCENTILE\_CONT(0.99)

Hiermee zijn de 1% laagste en 1% hoogste waarden van Quantity verwijderd. Dit zorgt voor robuustere gemiddelden, terwijl retourzendingen (negatieve waarden) behouden blijven.

Het outlierfilter is toegepast in een CROSS JOIN met een CTE (Quantiles) die de globale percentielen berekent over de volledige dataset.

**4. Aggregatie en transformatie**

Na het opschonen is de data geaggregeerd en samengevoegd om een maandelijkse structuur te creëren die direct vergelijkbaar is met prognoses.

**4.1. Tijdsdimensie**

Met:

YEAR(deliveryDate) AS Jaar,

MONTH(deliveryDate) AS Maand

worden alle gegevens samengevat per maand per product.

**4.2. Aggregatie**

Binnen elke combinatie van jaar, maand en artikel worden de volgende aggregaties berekend:

* SUM(Quantity) → **TotaalQuantity**
* AVG(PriceKG) → **GemiddeldePrijs** (indien aanwezig)
* AVG(TuWeight) → **GemiddeldTuWeight**

Deze metrics representeren de gerealiseerde verkoop per product per maand.

**4.3. Toevoeging van productgroepen**

Via joins op specificaties, specificationProductGroup en groepen is per artikel de bijbehorende **GroepNaam** toegevoegd.  
Deze stap maakt het mogelijk om resultaten op hoger niveau (bijv. per productgroep of categorie) te analyseren.

**5. Samenvoeging met prognoses**

Om realisaties te kunnen vergelijken met geplande prognoses, is een tweede dataset (Jaarschema\_CTE) aangemaakt waarin de prognosehoeveelheden uit de tabel jaarschema per artikel en maand worden samengevat.

Daarna zijn beide datasets gecombineerd via een FULL OUTER JOIN, zodat:

* maanden/artikelen zonder realisaties maar mét prognose zichtbaar blijven,
* én omgekeerd: realisaties zonder prognose ook worden weergegeven.

Voorbeeld:

FULL OUTER JOIN Jaarschema\_CTE j

ON d.Jaar = j.jaar

AND d.Maand = j.maand

AND d.Artikelnummer = j.artikel

Hierbij worden ontbrekende waarden vervangen door 0 of '-' via COALESCE() om de volledigheid te garanderen.

**6. Eindresultaat**

De uiteindelijke tabel dbo.Napoleon\_All bevat de volgende velden:

| **Kolom** | **Beschrijving** |
| --- | --- |
| Jaar | Jaar van levering. |
| Maand | Maand van levering. |
| Artikelnummer | Unieke code van het product. |
| Artnaam | Naam van het product. |
| GroepNaam | Productgroepnaam (indien beschikbaar). |
| TotaalQuantity | Totale verkochte hoeveelheid in die maand. |
| PrognoseWaarde | Geplande hoeveelheid volgens het jaarschema. |

Deze tabel vormt de schakel tussen operationele productiegegevens en prognosedata. Het resultaat is direct bruikbaar voor trendanalyses, afwijkingsmetingen (bijv. realisatie vs. forecast) en rapportering per product of productgroep.

**7. Belangrijkste ontwerpkeuzes en motivatie**

| **Keuze** | **Redenering** |
| --- | --- |
| **Verwijderen van kolommen met alleen NULL of vaste waarden** | Minimaliseert datasetgrootte en verhoogt leesbaarheid. |
| **Orderstatus ≠ 4** | Geannuleerde orders vervalsen realisatiecijfers. |
| **Ordertype 18 negatief gemaakt** | Retouren moeten worden afgetrokken van de totale verkoop. |
| **Outlierfilter (1%–99%)** | Verwijdert extreme meetfouten zonder structurele patronen te verliezen. |
| **Minstens 10 maanden per jaar** | Waarborgt representatieve jaardata en voorkomt seizoensbias. |
| **Gebruik van deliveryDate in plaats van orderDate** | Biedt een realistischer weergave van daadwerkelijke levering en productieactiviteit. |
| **FULL OUTER JOIN tussen realisatie en prognose** | Garandeert dat zowel ontbrekende prognoses als ontbrekende realisaties zichtbaar zijn voor analyse. |
| **Toevoegen van productgroepen** | Maakt rapportages mogelijk op aggregatieniveau (bijv. per productgroep of categorie). |

**8. Conclusie**

Door deze gestructureerde datavoorbereiding is een betrouwbare, analytisch bruikbare dataset gecreëerd die de brug vormt tussen operationele productiegegevens en planningsprognoses.  
De data is opgeschoond, genormaliseerd, samengevoegd en voorzien van relevante metadata. Dit vormt de basis voor verdere analyse, machine learning toepassingen of dashboarding binnen de 4food-omgeving.