# **README — Sentiment Survey Datasets**

# **Purpose**

This dataset contains **employee sentiment survey data** designed for integration as a **second data source** in the EUROSTAT\_load\_db relational database.  
It includes both **real** and **synthetic** responses, structured in **3NF** for efficient storage, querying, and analytics.

The processing script (data\_wrangeling\_sentiment\_data.py) performs:

* Cleaning and validation of the original survey CSV.
* Generation of **synthetic data** following real response patterns.
* Assignment of realistic demographics and regions.
* Normalization into three related tables (Respondent, Session, Response).
* Export of UTF-8 CSVs ready for database import.

## **Input**

**Source:** Sentiments\_data - Sheet1.csv  
 Contains survey responses on dimensions A, I, E, O, and H (Likert scale 1–5) plus respondent metadata.

### **Synthetic Data Logic**

* Synthetic records expand the dataset (~300×).
* Generated using **multivariate normal distributions** to preserve correlations across related questions (e.g., A1–A5).
* Demographic categories (Region, Afdeling, Dienstverband, Leeftijd) randomized within realistic ranges.
* Real vs. synthetic identified via is\_synthetic flag.

## **Output Files (3NF)**

| **File** | **Table** | **Description** |
| --- | --- | --- |
| Sentiments\_3NF\_Respondent.csv | Respondent\_2 | Demographic info per respondent |
| Sentiments\_3NF\_Session.csv | SurveySession\_2 | Metadata for each survey submission |
| Sentiments\_3NF\_Response.csv | SurveyResponse\_2 | Question-level responses (1–5) |

All files are UTF-8 encoded with headers and no index columns.

## **Database Schema (3NF)**

CREATE TABLE Respondent\_2 (

RespondentID TEXT PRIMARY KEY,

Region TEXT,

Afdeling TEXT,

Dienstverband TEXT,

Leeftijd TEXT,

UserLanguage TEXT,

is\_synthetic INT

);

CREATE TABLE SurveySession\_2 (

SessionID TEXT PRIMARY KEY,

RespondentID TEXT REFERENCES Respondent\_2(RespondentID),

ResponseId TEXT,

StartDate TIMESTAMP,

Finished TEXT,

DistributionChannel TEXT

);

CREATE TABLE SurveyResponse\_2 (

SessionID TEXT REFERENCES SurveySession\_2(SessionID),

QuestionCode TEXT,

Score INT

);

**Relationships**

Respondent\_2 (1) ───< SurveySession\_2 (1) ───< SurveyResponse\_2

* Each respondent can have multiple survey sessions.
* Each session contains multiple question responses.

## **Loading Examples**

**PostgreSQL**

\copy Respondent\_2 FROM 'Sentiments\_3NF\_Respondent.csv' CSV HEADER;

\copy SurveySession\_2 FROM 'Sentiments\_3NF\_Session.csv' CSV HEADER;

\copy SurveyResponse\_2 FROM 'Sentiments\_3NF\_Response.csv' CSV HEADER;

**MySQL**

LOAD DATA INFILE '/path/Sentiments\_3NF\_Respondent.csv'

INTO TABLE Respondent\_2

FIELDS TERMINATED BY ',' OPTIONALLY ENCLOSED BY '"'

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

Add referential constraints:

ALTER TABLE SurveySession\_2

ADD CONSTRAINT fk\_session\_respondent

FOREIGN KEY (RespondentID) REFERENCES Respondent\_2(RespondentID);

ALTER TABLE SurveyResponse\_2

ADD CONSTRAINT fk\_response\_session

FOREIGN KEY (SessionID) REFERENCES SurveySession\_2(SessionID);

## **Validation Queries**

-- Validate foreign keys

SELECT COUNT(\*) FROM SurveySession\_2 s

LEFT JOIN Respondent\_2 r ON s.RespondentID = r.RespondentID

WHERE r.RespondentID IS NULL;

SELECT COUNT(\*) FROM SurveyResponse\_2 res

LEFT JOIN SurveySession\_2 s ON res.SessionID = s.SessionID

WHERE s.SessionID IS NULL;

-- Synthetic vs. real data

SELECT is\_synthetic, COUNT(\*) FROM Respondent\_2 GROUP BY is\_synthetic;

| **Check** | **Result** |
| --- | --- |
| Scores strictly between 1–5 | Done |
| Referential integrity between tables | Done |
| UTF-8 encoding | Done |
| Synthetic data clearly marked | Done |

## **Typical Workflow**

Place Sentiments\_data - Sheet1.csv in the working directory.

Run:  
  
 python data\_wrangeling\_sentiment\_data.py

Verify console summary (row counts, export confirmation).

Import generated 3NF CSVs into the database.  
Apply foreign key constraints and validate joins.

## **Developer Notes**

* **Primary Keys:** RespondentID, SessionID.
* **Foreign Keys:** link sessions → respondents → responses.
* **Data Integrity:** maintained via synthetic indicators and controlled variation.
* **Temporal Dimension:** StartDate field.
* **Granularity:** question-level (one row per question per session).
* **Compatibility:** schema aligns with EUROSTAT\_load\_db for multi-source analytics.
* **No aggregation:** all values remain as collected/generated.