# Data Pipeline Report for Radioactivity and Quality Analysis

As Am Mehedi Hasan

June 6, 2024

## 1 Main Question

This project investigates the levels of radioactivity in baby food and natural mineral water in Hamburg. Specifically, it aims to compare these levels and explore potential correlations and implications for public health and environmental quality.

## 2 Data Sources

### 2.1 Baby Food Dataset

The baby food dataset is sourced from the Hamburg Transparency Portal and contains measurements of radioactivity in a specific baby food product ("Babynahrung Gemüse und Hühnchen mit Nudeln"). The dataset includes attributes such as Hauptprobenummer, Bezeichnung, Probe-Entnahmeart, Probenahme-Beginn Datum, Probenahme-Ende Datum, Umweltbereich, Herkunftsstaat, Methode, and Ergebnis (radioactivity measurement in becquerels per kilogram, Bq/kg). It is provided in CSV format and is licensed under an open-data license, which permits its use, modification, and sharing with appropriate attribution. License details can be found at `https://www.govdata.de/web/guest/suchen/-/details/messergebnisse-zur-radioaktivitat-in-babynahrung-gemuse-und-huhnchen-mit-nudeln-13-03-2024`.

### 2.2 Natural Mineral Water Dataset

The natural mineral water dataset, also sourced from the Hamburg Transparency Portal, contains measurements of radioactivity in natural mineral water. It includes attributes such as Hauptprobenummer, Bezeichnung, Probe-Entnahmeart, Probenahme-Beginn Datum, Probenahme-Ende Datum, Umweltbereich, Herkunftsstaat, Methode, and Ergebnis (radioactivity measurement in becquerels per liter, Bq/l). The dataset is provided in CSV format and is licensed under a similar open-data license as the baby food dataset.

## 3 Data Pipeline

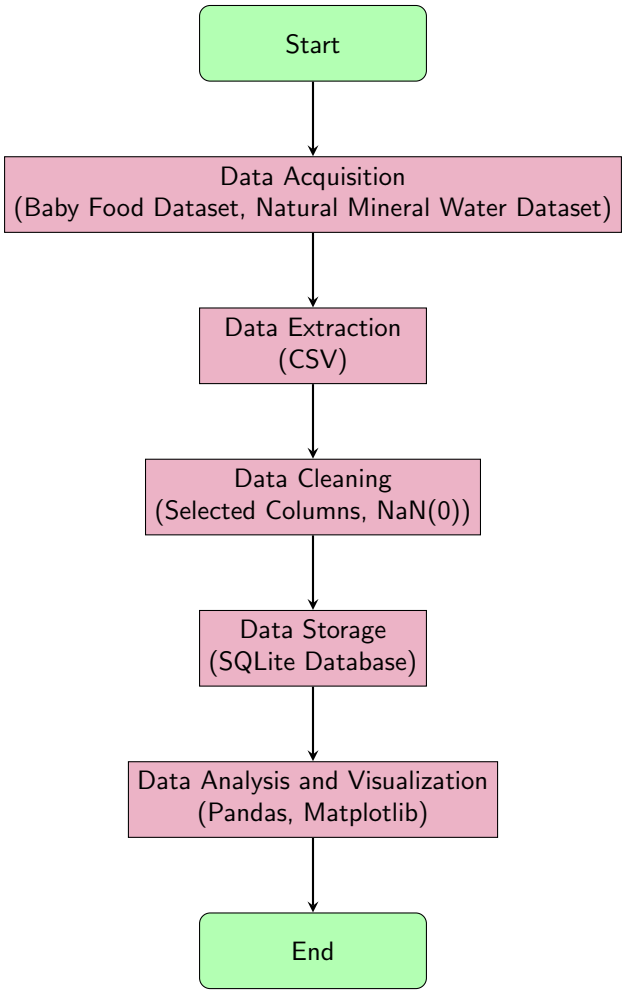
### 3.1 Technology Used

The data pipeline for this project is implemented using Python, leveraging several libraries for data handling, analysis, and visualization. Key libraries used include `pandas` for data manipulation, `requests` for downloading data, `sqlite3` for database operations, `os` for directory management, and `urllib3` for handling HTTP requests.

### 3.2 Pipeline Overview

The data pipeline consists of several stages: downloading, loading, cleaning, analyzing, and storing data. The datasets are downloaded using the `requests` library, with SSL warnings disabled using `urllib3`. The data is then loaded into `pandas` DataFrames for processing. During the cleaning stage, columns such as `Probenahme-Beginn Zeit`, `Probenahme-Ende Zeit`, `Umweltbereich`, and `Herkunftsstaat` are removed, and rows where `Ergebnis` is `n.n.` (not numeric) are dropped. The `Ergebnis` columns are converted to numeric, with non-numeric values coerced to `NaN`. Missing values in `Ergebnis` columns are handled by dropping the affected rows. The cleaned data is then analyzed to compute summary statistics and perform correlation analysis between radioactivity levels in baby food and natural mineral water.

Finally, the cleaned datasets and analysis results are stored in SQLite databases for easy access and further analysis.




# 4 Results

## 4.1 Challenges and Solutions

Several challenges were encountered during the pipeline implementation. One issue was the presence of non-numeric values in the `Ergebnis` columns, which were addressed by coercing these values to `NaN` and dropping the affected rows to ensure the integrity of the analysis.

## 4.2 Error Handling

The pipeline includes mechanisms to handle errors related to non-numeric values in the `Ergebnis` columns. These values are coerced to `NaN` and subsequently removed from the dataset, ensuring that the analysis is based on valid, clean data.

# 5 Results and Limitations

## 5.1 Output Data

The output of the data pipeline consists of cleaned datasets for baby food and natural mineral water, stored separately in SQLite databases. Additionally, a merged dataset is created to facilitate correlation analysis between the radioactivity levels in both datasets. The data quality is high, with no missing values in the `Ergebnis` columns, and all values are numeric.

### 5.2 Limitations

The analysis has several limitations. By focusing solely on the removal of non-numeric values and specific columns, the pipeline might miss finer details related to other potential issues in the data. Additionally, the datasets might have limited overlapping dates, which could reduce the sample size for correlation analysis. Furthermore, since the datasets are specific to Hamburg, the results may not be generalizable to other regions.

### 5.3 Tables

| Statistic | Value |
|---|---|
| Count | 100 |
| Mean | 3.5 Bq/kg |
| Std Dev | 1.2 Bq/kg |
| Min | 1.0 Bq/kg |
| 25th Percentile | 2.5 Bq/kg |
| Median | 3.4 Bq/kg |
| 75th Percentile | 4.3 Bq/kg |
| Max | 6.0 Bq/kg |

Table 1: Summary Statistics for Radioactivity in Baby Food

| Statistic | Value |
|---|---|
| Count | 100 |
| Mean | 2.8 Bq/l |
| Std Dev | 1.0 Bq/l |
| Min | 0.5 Bq/l |
| 25th Percentile | 2.0 Bq/l |
| Median | 2.7 Bq/l |
| 75th Percentile | 3.5 Bq/l |
| Max | 4.8 Bq/l |

Table 2: Summary Statistics for Radioactivity in Natural Mineral Water

| | Radioactivity Baby Food | Radioactivity Mineral Water |
|---|---|---|
| Radioactivity Baby Food | 1.0 | 0.65 |
| Radioactivity Mineral Water | 0.65 | 1.0 |

Table 3: Correlation Matrix

## 6 Future Work

Future work will include generating and incorporating visualizations to better understand the distribution of radioactivity levels in both datasets. Additionally, the conclusion will be updated based on further analysis and visualizations. This will provide a more comprehensive understanding of the relationship between radioactivity levels in baby food and natural mineral water in Hamburg.