Projektname

Datenbericht

Letzte Änderung: 24.08.2021

# Rohdaten

## Übersichtstabelle der Rohdatensätze

|  |  |  |
| --- | --- | --- |
| **Datensatz Name** | **Quelle** | **Speicherort** |
| Dataset 1 History of Pandemics in Switzerland | Kaspar Staub and Wiebke Weber from  leaddata.ch | Uploaded Excel file: 1\_History\_Pandemics.xlsx in GitHub Repo: [Project\_PODSV/Data at main · plospen1/Project\_PODSV](https://github.com/plospen1/Project_PODSV/tree/main/Data) |
| Dataset 2 | … | … |
| Dataset 3  3\_Todesursachen Schweiz ohne Alter 1876-2002 | Kaspar Staub and Wiebke Weber from  leaddata.ch | Uplooaded Excel file: 3\_Todesursachen Schweiz ohne Alter 1876-2002 in GitHub Rep:  [Project\_PODSV/Data at main · plospen1/Project\_PODSV](https://github.com/plospen1/Project_PODSV/tree/main/Data) |
| full\_data | [covid-19-data/public/data/cases\_deaths at master · owid/covid-19-data](https://github.com/owid/covid-19-data/tree/master/public/data/cases_deaths) | Uplooaded Excel file: full\_data:  [Project\_PODSV/Data at main · plospen1/Project\_PODSV](https://github.com/plospen1/Project_PODSV/tree/main/Data) |

## Details Dataset 1

* **Description**:  
  Annual mortality statistics in Switzerland from 1880 to 2022, including pandemic-related deaths and excess mortality estimates.
* **Source Details:**

Historical data from Switzerland

COVID-19 data from Switzerland

* **Data Acquisition:¨**

File: 1\_History\_Pandemics.xlsx, contact Person Wiebke Weber

* **Legal Aspects:**

The data is publicly available and free to use for educational and analytical purposes.

* **Data Governance:**
* Category: Public
* No personal data is included.
* **Access Information:**
* Provided as an Excel file. On GitHub, linked at Speicherort.
* Can be accessed using common tools (Excel, Python pandas, etc.)

### Datenkatalog

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Spaltenindex** | **Spaltenname** | **Datentyp** | **Wertebereich / Validierung** | **Beschreibung** |
| 1 | Jahr | Integer | 1880–2022 | Year of observation |
| 2 | Todesfälle\_Grippe\_100000 | Float | ≥ 0 | Influenza deaths per 100,000 population |
| 3 | Todesfälle\_Covid\_100000 | Float | ≥ 0 | COVID-19 deaths per 100,000 population |
| 4 | TodesfälleGrippe | Float | ≥ 0 | Total number of influenza deaths |
| 5 | TodesfälleCOVID | Float | ≥ 0 | Total number of COVID-19 deaths |
| 6 | Population | Integer | ≥ 0 | Total population of Switzerland |
| 7 | Überasterblichkeit\_Alles | Float | Any value (%) | Total excess mortality percentage |
| 8 | ÜberasterblichkeitPlus | Float | ≥ 0 | Positive excess mortality percentage |
| 9 | ÜberasterblichkeitMinus | Float | ≤ 0 | Negative excess mortality percentage |

## Datenqualität Dataset 1

Methodology

* The analysis was performed using automated data profiling tools such as ydata.
* Descriptive statistics and a Pearson correlation matrix were used.
* The dataset includes 143 annual observations from 1880 to 2022.
* There are 9 variables total: 7 numeric and 2 categorical.

General Statistics

* Number of variables: 9
* Number of observations: 143
* Time span: 1880–2022
* Missing cells: 33.3% of all values
* No duplicate rows or invalid entries
* Mean year: 1951
* Mean population: approximately 5.2 million

Key Variables and Values

* TodesfälleGrippe (Influenza deaths): 3 to 21,491 per year, mean = 767, 2.1% missing.
* TodesfälleCOVID: 3 values only, 97.9% missing.
* Todesfälle\_Grippe\_100000: 0.0385 to 554 per 100k, mean = 18.2, minimal missing.
* Todesfälle\_Covid\_100000: 3 values, 97.9% missing.
* ÜbersterblichkeitMinus: -13.8% to 0%, mean = -2.87%, 52.4% missing.
* ÜbersterblichkeitPlus: 0% to 49.3%, mean = 3.74%, 46.2% missing.
* Übersterblichkeit\_Alles: nearly complete, -13.8% to 49.3%, mean = 0.66%.

Interpretation and Insights

* The dataset is ideal for long-term influenza mortality analysis.
* COVID-19 data is too incomplete for reliable modeling.
* Excess mortality is a valuable measure for both direct and indirect pandemic impact.
* Population normalization (per 100k) allows comparison across time.
* Negative excess mortality years may reflect successful health measures or mild seasons.

Data Quality Assessment

* Completeness: moderate (mainly missing values in COVID fields).
* Consistency: high (no duplicates or major outliers).
* Usability: excellent for flu and excess mortality; limited for COVID.

Conclusion

* Long-term flu mortality trends
* Excess mortality during pandemics
* Data-driven storytelling

## Details Dataset 2

* Beschreibung, welche Informationen die Daten enthalten
* Details zur Datenquelle
* Informationen zur Datenbeschaffung: Beschreibung und eventuell Verweise auf Ressourcen (Download-Scripts, Tools, Online-Services, …)
* Rechtliche Aspekte zur Nutzung der Daten, Lizenzen etc.
* Data Governance-Aspekte: Kategorisierung der Daten aufgrund Geschäftsinterner Vorgaben, z.B. Öffentlich, Geschäftsrelevant, Persönlich
* Wie kann auf die Daten zugegriffen werden: Beschreibung, Scripts, Tools, …
* Falls zutreffend: Unterteilung in abhängige (Zielgrösse, Regressand) und unabhängige (Regressor) Variablen
* …

### Datenkatalog Dataset 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Spaltenindex | Spaltenname | Datentyp | Werte (Wertebereich, ev. Validierungsregeln) | Kurze Beschreibung | Primary/ Foreign Key |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| … |  |  |  |  |  |

Beispiel eines Schemas von Daten in einer relationalen Datenbank:

Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

KI-generierte Inhalte können fehlerhaft sein.

### Datenqualität Dataset 2

Explorative Datenanalyse

* Beschreibung der Vorgehensweise und Verweise auf verwendete Ressourcen (Scripts, Tools, Software, …) und Output-Dateien -> Nachvollziehbarkeit und Sicherstellung der Reproduzierbarkeit
* Zusammenfassende Ergebnisse: Visualisationen und statistische Kenngrössen
* Erkenntnisse und relevante Zusammenhänge im Bezug auf die Modellierungs- und Analyseziele
* Einschätzung der Datenqualität

## Details Dataset 3

* **Description**:  
  Annual mortality statistics in Switzerland from 1880 to 2022, including pandemic-related deaths and excess mortality estimates.
* **Source Details:**

For more information please contact Wiebke Weber or Kaspar Staub

* **Data Acquisition:¨**

File: 3\_Todesursachen Schweiz ohne Alter 1876-2002.xlsx, contact Wiebke Weber

* **Legal Aspects:**

The data is publicly available and free to use for educational and analytical purposes.

* **Data Governance:**
* Category: Public
* No personal data is included.
* **Access Information:**
* Provided as an Excel file. On GitHub, linked at Speicherort.
* Can be accessed using common tools (Excel, Python pandas, etc.)

## Datenkatalog Dataset 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Spalten**  **index** | **Spaltenname** | **Datentyp** | **Wertebereich / Validierung** | **Beschreibung** |
| 1 | Jahr | Float | 1876 - 2004 | Durchschnittliches Beobachtungsjahr |
| 2 | Todesfälle\_Infektionen\_Total | Float | ≥ 0 | Gesamt: Infektions- und parasitäre Krankheiten |
| 3 | Pocken | Float | ≥ 0 | Pocken-Todesfälle |
| 4 | Scharlach | Float | ≥ 0 | Scharlach-Todesfälle |
| 5 | Masern | Float | ≥ 0 | Masern-Todesfälle |
| 6 | Typhus\_Paratyphus | Float | ≥ 0 | Typhus und Paratyphus |
| 7 | Diphtherie | Float | ≥ 0 | Diphtherie-Todesfälle |
| 8 | Keuchhusten | Float | ≥ 0 | Keuchhusten-Todesfälle |
| 9 | Tuberkulose\_Total | Float | ≥ 0 | Alle Tuberkulosefälle |
| 10 | Tuberkulose\_Lungen | Float | ≥ 0 | Tuberkulose der Lungen |
| 11 | Tuberkulose\_Andere\_Organe | Float | ≥ 0 | Tuberkulose anderer Organe |
| 12 | Tuberkulose\_Gehirnhaut | Float | ≥ 0 | Tuberkulose der Hirnhaut |
| 13 | Tuberkulose\_Knochen | Float | ≥ 0 | Tuberkulose der Knochen und Gelenke |
| 14 | Übrige\_Krankheiten\_Total | Float | ≥ 0 | Sonstige Infektionskrankheiten |
| 15 | Epidem\_Kinderlähmung | Float | ≥ 0 | Kinderlähmung |
| 16 | Rotlauf | Float | ≥ 0 | Rotlauf-Todesfälle |
| 17 | Syphilis | Float | ≥ 0 | Syphilis-Todesfälle |
| 18 | Aids | Float | ≥ 0 | Aids-Todesfälle |
| 19 | Atmungsorgane\_Total | Float | ≥ 0 | Todesfälle durch Erkrankungen der Atmungsorgane |
| 20 | Grippe | Float | ≥ 0 | Grippe-Todesfälle |
| 21 | Lungenentzündung | Float | ≥ 0 | Lungenentzündung |
| 22 | Bronchitis | Float | ≥ 0 | Bronchitis |
| 23 | Neubildungen\_Krebs\_Total | Float | ≥ 0 | Bösartige Neubildungen insgesamt |
| 24 | Krebs\_Gesamt | Float | ≥ 0 | Gesamtanzahl Krebsfälle |
| 25 | Krebs\_Magen\_Darm | Float | ≥ 0 | Magen-Darm-Krebs |
| 26 | Krebs\_Weibl\_Brustdrüse | Float | ≥ 0 | Brustkrebs (weiblich) |
| 27 | Krebs\_Bronchien\_Lunge | Float | ≥ 0 | Lungenkrebs/Bronchien |
| 28 | Nervensystem\_Kreislauf\_Total | Float | ≥ 0 | Gesamt: Nerven- und Kreislauferkrankungen |
| 29 | Nervensystem\_Total | Float | ≥ 0 | Nervenkrankheiten insgesamt |
| 30 | Hirnschlag | Float | ≥ 0 | Schlaganfall |
| 31 | Kreislauferkrankungen\_Total | Float | ≥ 0 | Herz-Kreislauf-Erkrankungen insgesamt |
| 32 | Herzkrankheiten | Float | ≥ 0 | Herzkrankheiten |
| 33 | Arterienverkalkung | Float | ≥ 0 | Arteriosklerose |
| 34 | Blut\_Stoffwechsel\_Total | Float | ≥ 0 | Erkrankungen des Blutes und Stoffwechsels |
| 35 | Anämie\_Leukämie | Float | ≥ 0 | Anämie & Leukämie |
| 36 | Diabetes | Float | ≥ 0 | Diabetes mellitus |
| 37 | Verdauungsorgane\_Total | Float | ≥ 0 | Erkrankungen des Verdauungssystems |
| 38 | MagenDarmKatarrh\_Kinder | Float | ≥ 0 | Magen-Darm-Katarrh bei Kindern |
| 39 | MagenDarmKatarrh\_Erwachsene | Float | ≥ 0 | ...bei Erwachsenen |
| 40 | MagenDarm\_Geschwür | Float | ≥ 0 | Magen-Darm-Geschwüre |
| 41 | Blinddarmentzündung | Float | ≥ 0 | Appendizitis |
| 42 | Leberzirrhose | Float | ≥ 0 | Leberzirrhose |
| 43 | Harnapparat\_Total | Float | ≥ 0 | Krankheiten des Harnapparats |
| 44 | Nierenentzündung | Float | ≥ 0 | Nierenentzündung |
| 45 | Gewaltsamer\_Tod\_Total | Float | ≥ 0 | Gewaltsamer Tod insgesamt |
| 46 | Selbsttötung | Float | ≥ 0 | Suizid |
| 47 | Unfälle\_Total | Float | ≥ 0 | Unfalltote insgesamt |
| 48 | Verkehrsunfälle\_Gesamt | Float | ≥ 0 | Verkehrsunfälle gesamt |
| 49 | Verkehrsunfälle\_Strasse | Float | ≥ 0 | Straßenverkehrsunfälle |
| 50 | Verkehrsunfälle\_Bahn | Float | ≥ 0 | Bahnverkehrsunfälle |
| 51 | Verkehrsunfälle\_Flug | Float | ≥ 0 | Flugunfälle |
| 52 | Sturz | Float | ≥ 0 | Sturzunfälle |
| 53 | Ertrinken | Float | ≥ 0 | Ertrinken |
| 54 | Mord\_Totschlag\_Total | Float | ≥ 0 | Mord/Totschlag + unklare Ursachen |
| 55 | Mord\_Totschlag | Float | ≥ 0 | Mord und Totschlag |
| 56 | Übrige\_Krankheiten\_Total | Float | ≥ 0 | Übrige bekannte Ursachen |
| 57 | Chronischer\_Alkoholismus | Float | ≥ 0 | Alkoholismus |
| 58 | Schwangerschaft\_Geburt | Float | ≥ 0 | Schwangerschaft & Geburt |
| 59 | Angeborene\_Krankheiten\_Säuglinge\_Total | Float | ≥ 0 | Angeborene Säuglingskrankheiten |
| 60 | Frühgeburt | Float | ≥ 0 | Frühgeburt |
| 61 | Alterschwäche | Float | ≥ 0 | Altersschwäche |
| 62 | Todesursache\_Arzt\_beigezogen | Float | ≥ 0 | Arzt beigezogen |
| 63 | Todesursache\_Arzt\_nicht\_beigezogen | Float | ≥ 0 | Arzt nicht beigezogen |
| 64 | Todesursache\_Unbekannt\_Total | Float | ≥ 0 | Todesursache unbekannt (gesamt) |
| 65 | Todesfälle\_Gesamt | Float | ≥ 0 | Gesamttodesfälle |

## Datenqualität Dataset 3

Methodology

* The analysis was performed using automated data profiling tools such as ydata.
* Descriptive statistics and a Pearson correlation matrix were used.
* The dataset includes 250 annual observations.
* There are 82 variables total: all numerical

General Statistics

* Number of variables: 82
* Number of observations: 240
* Time span: 1876–2002
* Missing cells: 25.1% of all values
* 1 duplicate rows
* 81 unsupported variable types
* No statistical measures can be calculated

Key Variables and Values

Jahr: Average: 1939.8

Infektions-und parasitäre Krankheiten: Title

* Pocken, Scharlach, Masern, Typhus, Diphtherie, Keuchhusten: Title
  + Total: Average 825
  + Pocken: Average: 31.6
  + Scharlach 1): Average: 98
  + Masern 2): Average: 134
  + Typhus Paratyphus: Average: 142
  + Diphtherie: Average: 386
  + Keuchhusten: Average: 232
* Tuberkulose
  + Total: Average: 3’163
  + Atmungsorgane
    - Total: Average: 2’417.1
    - Lungen-Tb: Average: 2’331
  + Übrige Organe
    - Total: Average: 776.5
    - Gehirnhaut, Hirnhaut: Average: 388
    - Knochen, Gelenke: Average: 273.4
* Übrige Krankheiten 4)
  + Total: Average: 607
  + Epidem. Kinderlähmung: Average: 29.5
  + Rotlauf: Average: 65
  + Syphilis: Average: 88.5
  + Aids: Average: 228
* Total: Average: 4’581

Atmungsorgane:

* Total: Average: 4’917
* Grippe: Average: 630.3
* Lungenentzünding: Average: 2’273
* Bronchitis: Average: 689.8

Jahr: Average: 1939.8

Neubildungen:

* Total Krebs (bösartige Geschwülste); Average: 7’058.1
  + Total: Average: 5’492
  + Magen, Darm (Total): Average: 2’430.6
  + Weibl. Brustdrüse: Average: 596
  + Bronchien, Lunge 7): Average: 831.6

Nervensystem und Kreislauforgane

* Total : Average: 15’009
* Nervensystem:
  + Total: Average: 2’309.6
  + Hirnschlag: Average: 1’845
* Kreislauforgane
  + Total: Average: 13’385.6
  + Herzkrankeiten: Average: 6’295
  + Arterienverkalkung: Average: 4’778.4

Blut, Stoffwechsel 8)

* Total: Average: 1’117
* Anämie, Leukämie: Average: 278.7
* Zuckerkrankheit (Diabetes mellitus): Average: 569

Ernährungsapparat, Verdauuungsorgane

* Total: Average: 3’116.9
* Magen-Darm-Kattarrh
  + Säuglinge und Kleindkinder 9): Average: 1’155
  + Erwachsene: Average: 312.1
* Magen-Darm-geschwür: Average: 250
* Blinddarm entzündung: Average: 176.2
* Leberzirrhose: Average: 443

Harnapparat

* Total: Average: 951.7
* Nierenentzündungen: Average: 686

Jahr: Average: 1939.8

Gewaltsamer Tod:

* Total: Average: 2’989
* Selbsttötung: Average: 844.5
* Unfälle
  + Total 12): Average: 2’010
  + Verkehrsunfälle:
    - Total: Average: 577.2
    - Strassenverkehr: Average: 705
    - Bahnverkehr: Average: 70.7
    - Flugverkehr; Average: 17
  + Sturz: Average: 672.1
  + Ertrinken: Average: 183
* Mord, Totschlag, unklare Ursache
  + Total: Average: 135.2
  + Mord, Totschlag: Average: 55

Übrige bekannte Todesursachen

* Total: Average: 4’296.4
* Chronischer Alkoholismus: Average: 167
* Schwangerschaft u. Geburt: Average: 229.7
* Angeborene Säuglingskrankheiten Total: Average: 1’775
  + Frühgeburt 14): Average: 1’301.8
* Alterschwäche; Average: 965

Unbekannte Todesursachen

* Arzt beigezogen: Average: 437.7
* Arzt nicht beigezogen: Average: 1’653
* Total: Average: 1’973.5

Gesamttotal: Average: 45’535

Interpretation and Insights

* Ideal for looking at different causes of death in more or less detail

Data Quality Assessment

* Completeness: poor- moderate (lot of missing values).
* Consistency: moderate (no duplicates or major outliers).
* Usability: poor, because of how the title are distributed (unclear for when reading the table)

Conclusion

* Insight to cause of death
* Data-driven storytelling

## Details full\_data

* **Description:**

Daily COVID-19 case and death statistics from various global locations, including cumulative and rolling aggregates (weekly, biweekly), spanning the pandemic timeline.

* **Source Details:**

Data sourced from a global COVID-19 tracking initiative (e.g., Our World in Data, Johns Hopkins University).

* **Data Acquisition:**

File: full\_data.csv, in the GitHub Repo

* **Legal Aspects:**

The dataset is public

* **Data Governance:**

Category: Public

No personal data is included.

* **Access Information:**

Provided as a CSV file.

Can be accessed using common tools (Excel, Python pandas, R, etc.)

### Datenkatalog

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Spaltenindex** | **Spaltenname** | **Datentyp** | **Wertebereich / Validierung** | **Beschreibung** |
| 1 | date | Date | ISO format (YYYY-MM-DD) | Date of record |
| 2 | location | String | Valid location names | Country or region name |
| 3 | new\_cases | Integer | ≥ 0 | New reported COVID-19 cases (daily) |
| 4 | new\_deaths | Integer | ≥ 0 | New reported COVID-19 deaths (daily) |
| 5 | total\_cases | Integer | ≥ 0 | Total confirmed cases to date |
| 6 | total\_deaths | Integer | ≥ 0 | Total confirmed deaths to date |
| 7 | weekly\_cases | Integer | ≥ 0 | Aggregated cases over the past 7 days |
| 8 | weekly\_deaths | Integer | ≥ 0 | Aggregated deaths over the past 7 days |
| 9 | biweekly\_cases | Integer | ≥ 0 | Aggregated cases over the past 14 days |
| 10 | biweekly\_deaths | Integer | ≥ 0 | Aggregated deaths over the past 14 days |

## Datenqualität full\_data

**Methodology**

* Data collected and processed by a centralized public health data provider.
* Descriptive analysis done via automated profiling tools.
* Rolling aggregates (weekly, biweekly) computed from raw daily data.
* Dataset includes over 411,000 rows across 10 variables.

**General Statistics**

* Number of variables: 10
* Number of observations: 411,804
* Duplicate rows: 0
* Primary key: (date, location)

**Key Variables and Values**

* new\_cases: varies widely, occasional spikes during outbreak peaks, 0.4% missing.
* new\_deaths: similar pattern, 0.29% missing.
* weekly\_cases, weekly\_deaths: ~0.6–0.7% missing.
* biweekly\_cases, biweekly\_deaths: ~1% missing.
* Cumulative fields (total\_cases, total\_deaths) are complete.

**Interpretation and Insights**

* Highly suitable for time-series analysis of pandemic spread and impact.
* Useful for comparing countries/regions and identifying outbreak waves.
* Rolling aggregates provide smoothing for trend analysis.

**Data Quality Assessment**

* **Completeness**: High (less than 1.2% missing overall)
* **Consistency**: High (no duplicates, logical numeric values)
* **Usability**: Excellent for trend analysis, comparison, and modeling

# Prozessierte Daten

## Übersichtstabelle der Prozessierten Daten

|  |  |  |
| --- | --- | --- |
| Name | Input-Datensätze | Speicherort |
| data\_set3\_cleaned | Dataset 3 | Can be found in the Github Repo under the name data\_set3\_cleaned:  [Project\_PODSV/Data at main · plospen1/Project\_PODSV](https://github.com/plospen1/Project_PODSV/tree/main/Data) |
| data\_set3\_infectious\_diseases | Dataset 3 | Can be found in the Github Repo under the name data\_set3\_infectious\_diseases  [Project\_PODSV/Data at main · plospen1/Project\_PODSV](https://github.com/plospen1/Project_PODSV/tree/main/Data) |

## Details Processed data\_set3\_cleaned

* **Description:**
* This dataset contains yearly death statistics in Switzerland (1876–2004), categorized by causes such as infectious diseases (e.g., measles, tuberculosis), respiratory diseases, cancers, cardiovascular issues, and accidents
* **Processing Steps:**
* **Headers:** The original table had multi-row headers (rows 3–5). These were merged using forward fill and joined with " | " to create clear column names.
* **Data Extraction:** Actual data starts at row 10 (index 9). Earlier rows were removed.
* **Cleanup:** Index was reset and column names verified.
* **Access and Tools:**
* **Available at:**  
  GitHub: [Project\_PODSV/Data at main · plospen1/Project\_PODSV](https://github.com/plospen1/Project_PODSV/tree/main/Data)
* **Tools:** Python and pandas
* **Script:** Processing done via a reproducible script using standard pandas functions.

## Details Processed data\_set3\_infectious\_diseases

* **Description:**
* This dataset contains annual death counts in Switzerland (1876–1891) specifically related to infectious diseases. It includes categories such as smallpox, scarlet fever, measles, typhoid/paratyphoid, diphtheria, and whooping cough, alongside total infectious disease-related deaths per year.
* **Processing Steps:**
* **Header Extraction:** Column names were taken from row 6 (index 5) of the Excel sheet. These were cleaned of empty and duplicate names, with placeholders added where needed.
* **Data Selection:** The actual data starts from row 8 (index 7). Only these rows were kept.
* **Filtering:** Only rows with valid numeric years were retained. The “Year” column was converted to int type.
* **Column Selection:** Only the first 8 columns (A–H) were kept, which correspond to the relevant disease categories.
* **Renaming Columns:** All columns were renamed to English names
* **Access and Tools:**
* **Available at:** GitHub: [Project\_PODSV/Data](https://github.com/plospen1/Project_PODSV/tree/main/Data) ,File name: data\_set3\_infectious\_diseases
* **Tools:** Python and pandas
* **Script:** Processing done via a reproducible script using pandas, including header cleaning, filtering valid rows, renaming, and exporting to .csv.