```
# Programmieren mit Python
# Projektarbeit Klimadiagramm
# Teilbereich: Data Quality (Nutzung der Pandas Bibliothek)
# Entwickler: Gero Krikawa
# Datum: 26.01.2023
#
```

Kurzanleitung
Data Quality

unter Verwendung der Pandas Bibliothek

Rohdaten - CSV Datei: Mannheim_data.csv

Vorgehensweise:

1. **Data Quality Reports**
   Ziel: Überblick über die Datenqualität verschaffen

   Python File:
   Data_Quality_ProfileReport_csv_input.py
   (sehr umfangreich und detailliert)

   Data_Quality_report_csv_input.py
   (einfach)

2. **Analyse der Daten anhand der Data Quality Reports**
   Ziel: Qualität Probleme identifizieren

   Zum Beispiel:
   - Missing Values
   - Extremwerte
   - ….

   Beispiel:

   CSV Datei: sample: Mannheim_data.csv
   Variable: tmax

   empty cells: 20.04.1950 – 01.05.1950
   extreme values: 15.06.1950 – 18.06.1950

empty cells: 20.04.1950 – 01.05.1950

Time	Tavg	Tmin	Tmax	Prcp	Snow
1950-04-15	5.3	4.9	10.2	5.3	0
1950-04-16	6.5	3.1	9.8	2.5	0
1950-04-17	8	2.5	13	0	0
1950-04-18	10.3	2.2	15.4	0.1	0
1950-04-19	12.5	5.3	17.5	0	0
1950-04-20	12.4	10		0.4	0
1950-04-21	14.4	6.1		8.4	0
1950-04-22	12.7	10.1		0	0
1950-04-23	9.6	8.7		0	0
1950-04-24	6.6	5.7		5.4	0
1950-04-25	4.9	3.8		0	0
1950-04-26	5.9	2.8		1.3	0
1950-04-27	6.5	5.4		3.1	0
1950-04-28	8.3	2.2		0.4	0
1950-04-29	9.1	7.8		16.7	0
1950-04-30	14.4	9.2		0	0
1950-05-01	17	7.6		0	0
1950-05-02	17.8	9.1	24.4	4.8	0
1950-05-03	12	10	17.1	0	0
1950-05-04	11.9	6.6	16.2	6.8	0
1950-05-05	9.5	8	15.9	7	0

extreme values: 15.06.1950 – 18.06.1950

Time	Tavg	Tmin	Tmax	Prcp	Snow
1950-06-10	17.6	9.5	23.2	0	0
1950-06-11	19.8	9.6	26.2	0	0
1950-06-12	21.8	9.9	28.7	0	0
1950-06-13	19.4	13.7	25.2	7.9	0
1950-06-14	17.6	16.3	21.6	7.6	0
1950-06-15	19.1	14.7	70	10.2	0
1950-06-16	18	14.9	70	6.7	0
1950-06-17	16.8	14.8	70	38.6	0
1950-06-18	20	13	70	0	0
1950-06-19	19.8	13.9	25.7	0	0
1950-06-20	22.9	12.3	29.7	9	0
1950-06-21	17.4	15.6	22.5	2.1	0
1950-06-22	16	12.2	20.2	0.8	0
1950-06-23	15.4	12.5	20	3.6	0
1950-06-24	17.2	8.6	22	0	0
1950-06-25	16.7	12.8	23.2	15	0

3. Datenbereinigung

Ziel: bereinigte Datenbasis

Auf Basis der identifizierten Datenqualitätsprobleme Auswahl geeigneter Pandas Funktionen zur Bereinigung der Rohdaten.

Details dokumentiert in:
Python File: Data_Cleaning_csv_input.py

Beispiel:
Datenreduktion: unnötige Spalten gelöscht
Missing values ersetzt durch Interpolation

Time	Tavg	Tmin	Tmax	Prcp	Snow
1950-04-15	5.3	4.9	10.2	5.3	0
1950-04-16	6.5	3.1	9.8	2.5	0
1950-04-17	8	2.5	13	0	0
1950-04-18	10.3	2.2	15.4	0.1	0
1950-04-19	12.5	5.3	17.5	0	0
1950-04-20	12.4	10		0.4	0
1950-04-21	14.4	6.1		8.4	0
1950-04-22	12.7	10.1		0	0
1950-04-23	9.6	8.7		0	0
1950-04-24	6.6	5.7		5.4	0
1950-04-25	4.9	3.8		0	0
1950-04-26	5.9	2.8		1.3	0
1950-04-27	6.5	5.4		3.1	0
1950-04-28	8.3	2.2		0.4	0
1950-04-29	9.1	7.8		16.7	0
1950-04-30	14.4	9.2		0	0
1950-05-01	17	7.6		0	0
1950-05-02	17.8	9.1	24.4	4.8	0
1950-05-03	12	10	17.1	0	0
1950-05-04	11.9	6.6	16.2	6.8	0
1950-05-05	9.5	8	15.9	7	0

Time	Tavg	Tmin	Tmax
1950-04-15	5.3	4.9	10.2
1950-04-16	6.5	3.1	9.8
1950-04-17	8	2.5	13
1950-04-18	10.3	2.2	15.4
1950-04-19	12.5	5.3	17.5
1950-04-20	12.4	10	18.03
1950-04-21	14.4	6.1	18.56
1950-04-22	12.7	10.1	19.09
1950-04-23	9.6	8.7	19.62
1950-04-24	6.6	5.7	20.15
1950-04-25	4.9	3.8	20.68
1950-04-26	5.9	2.8	21.22
1950-04-27	6.5	5.4	21.75
1950-04-28	8.3	2.2	22.28
1950-04-29	9.1	7.8	22.81
1950-04-30	14.4	9.2	23.34
1950-05-01	17	7.6	23.87
1950-05-02	17.8	9.1	24.4
1950-05-03	12	10	17.1
1950-05-04	11.9	6.6	16.2

Extrem values ersetzt durch Interpolation

Time	Tavg	Tmin	Tmax	Prcp	Snow
1950-06-10	17.6	9.5	23.2	0	0
1950-06-11	19.8	9.6	26.2	0	0
1950-06-12	21.8	9.9	28.7	0	0
1950-06-13	19.4	13.7	25.2	7.9	0
1950-06-14	17.6	16.3	21.6	7.6	0
1950-06-15	19.1	14.7	70	10.2	0
1950-06-16	18	14.9	70	6.7	0
1950-06-17	16.8	14.8	70	38.6	0
1950-06-18	20	13	70	0	0
1950-06-19	19.8	13.9	25.7	0	0
1950-06-20	22.9	12.3	29.7	9	0
1950-06-21	17.4	15.6	22.5	2.1	0
1950-06-22	16	12.2	20.2	0.8	0
1950-06-23	15.4	12.5	20	3.6	0
1950-06-24	17.2	8.6	22	0	0
1950-06-25	16.7	12.8	23.2	15	0

Time	Tavg	Tmin	Tmax
1950-06-10	17.6	9.5	23.2
1950-06-11	19.8	9.6	26.2
1950-06-12	21.8	9.9	28.7
1950-06-13	19.4	13.7	25.2
1950-06-14	17.6	16.3	21.6
1950-06-15	19.1	14.7	22.42
1950-06-16	18	14.9	23.24
1950-06-17	16.8	14.8	24.06
1950-06-18	20	13	24.88
1950-06-19	19.8	13.9	25.7
1950-06-20	22.9	12.3	29.7
1950-06-21	17.4	15.6	22.5