check_data_xml_xlsx

February 14, 2022

[1]:
```
import pandas as pd
import os
import sys
```

[2]:
```
# aktuelle Arbeitsverzeichnis anpassen (soll das Verzeichnis sein, in dem das
# Skipt liegt, relative Pfade sind darauf ausgerichtet)
os.chdir('/home/cudok/Documents/GitHub/projektliste_bf/')
```

[3]:
```
path_modules = os.path.join(
    '../../GitHub/dvg_lib/ProjektListe/')  # zeigt auf den Ordner indem die
                                    # Datei auswertung.pytung.py liegt
sys.path.append(path_modules)
import auswertung as asw
```

[4]:
```
#pip freeze
```

1 Einlesen

[5]:
```
list_spalten = '02_Parameter_Dateien/Spalten_xml2csv_Vergl.csv'
xml2csv = '02_Parameter_Dateien/Spalten_dict_xml2csv_Vergl.csv'
path_xml = "../../../Nextcloud/Shared/Digitale_Vernetzung/Assis/03_Projekte/
 ↪DVG0001_BMWi_Wende/12_Daten/01_Enargus/Daten_von_Bosch_2022_02_01/enargus.
 ↪xml"
```

[6]:
```
df = asw.read_xml_enargus(path_xml, xml2csv, list_spalten)
```

[7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1647 entries, 0 to 1646
Data columns (total 34 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fkz                   1647 non-null   object
 1   db                    1647 non-null   object
 2   fi_von/iso8601        1647 non-null   object
 3   fi_ende/iso8601       1647 non-null   object
 4   v_thema               1647 non-null   object
```

```
 5   fi_sumbew/value         1647 non-null   object
 6   ver_bez                 1647 non-null   object
 7   lp_nr                   1647 non-null   object
 8   lp_text                 1647 non-null   object
 9   name_st                 1647 non-null   object
 10  plz_strasse_st          1647 non-null   object
 11  ort_st                  1647 non-null   object
 12  ad_str_st               1647 non-null   object
 13  land_st                 1647 non-null   object
 14  gem_gemkz_st            1647 non-null   object
 15  name_ze                 1647 non-null   object
 16  plz_strasse_ze          1647 non-null   object
 17  ort_ze                  1647 non-null   object
 18  ad_str_ze               1647 non-null   object
 19  land_ze                 1647 non-null   object
 20  gem_gemkz_ze            1647 non-null   object
 21  v_ressort               1647 non-null   object
 22  v_pt_detail             1647 non-null   object
 23  v_forschsp_text         1647 non-null   object
 24  v_prog_text             1647 non-null   object
 25  auf_bez_pub             1419 non-null   object
 26  auf_bez_pub_quelle      1419 non-null   object
 27  auf_bez_pub_en          1303 non-null   object
 28  auf_bez_pub_quelle_en   0 non-null      object
 29  pers_pl                 1647 non-null   object
 30  pers_titel_pl           1305 non-null   object
 31  pers_vname_pl           1647 non-null   object
 32  pers_name_pl            1647 non-null   object
 33  pers_email_pl           1647 non-null   object
dtypes: object(34)
memory usage: 437.6+ KB
```

[8]:
```
#cols_vec = asw.read_spalten_vor_csv('02_Parameter_Dateien/
 ↪Spalten_xml2csv_Vergl.csv')
#cols_vec
```

[9]:
```
#namenspaces_enargus = {'':"http://www.enargus.de/elements/0.1/begleitforschung/
 ↪", 'bscw':"http://bscw.de/bscw/elements/0.1/"}
```

[10]:
```
#cols = ['FKZ', 'Datenbank', 'Laufzeitbeginn']
#col_dic = {'FKZ':'fkz', 'Datenbank':'db', 'Laufzeitbeginn':'fi_von/iso8601'}
#df_test = asw.read_xml(path_xml,col_dic, cols, namespaces=namenspaces_enargus)
```

[11]:
```
#df_test
```

[12]:
```
path_excel = '../../../Nextcloud/Shared/WenDE/12_Daten/03_Gesamt_BF_Daten/
 ↪20220207_Verteiler_EWB_Projekte.xlsx'
```

[13]:
```
df_xlsx = pd.read_excel(path_excel, sheet_name='EnArgus Rohdaten')
```

[14]:
```
df_xlsx.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1647 entries, 0 to 1646
Data columns (total 46 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   http://bscw.de/bscw/elements/0.1/.oid   1647 non-null   int64
 1   http://bscw.de/bscw/elements/0.1/.name  1647 non-null   object
 2   fkz                                     1647 non-null   object
 3   db                                      1647 non-null   object
 4   v_thema                                 1647 non-null   object
 5   fi_sumbew.value                         1647 non-null   float64
 6   fi_sumbew.currency                      1647 non-null   object
 7   fi_von.type                             1647 non-null   object
 8   fi_von.iso8601                          1647 non-null   datetime64[ns]
 9   fi_ende.type                            1647 non-null   object
 10  fi_ende.iso8601                         1647 non-null   datetime64[ns]
 11  ver_bez                                 1647 non-null   object
 12  lp_nr                                   1647 non-null   object
 13  lp_text                                 1647 non-null   object
 14  name_st                                 1647 non-null   object
 15  plz_strasse_st                          1647 non-null   int64
 16  ort_st                                  1647 non-null   object
 17  ad_str_st                               1647 non-null   object
 18  land_st                                 1647 non-null   object
 19  gem_gemkz_st                            1647 non-null   int64
 20  name_ze                                 1647 non-null   object
 21  plz_strasse_ze                          1647 non-null   int64
 22  ort_ze                                  1647 non-null   object
 23  ad_str_ze                               1647 non-null   object
 24  land_ze                                 1647 non-null   object
 25  gem_gemkz_ze                            1647 non-null   int64
 26  v_ressort                               1647 non-null   object
 27  v_pt_detail                             1647 non-null   object
 28  v_forschsp_text                         1647 non-null   object
 29  v_prog_text                             1647 non-null   object
 30  v_kwort                                 1647 non-null   object
 31  auf_bez_pub                             1419 non-null   object
 32  auf_bez_pub_quelle                      1419 non-null   object
 33  auf_bez_pub_en                          1303 non-null   object
 34  auf_bez_pub_en_quelle                   1303 non-null   object
 35  pers_pl                                 1647 non-null   object
 36  pers_titel_pl                           1305 non-null   object
 37  pers_vname_pl                           1647 non-null   object
 38  pers_name_pl                            1647 non-null   object
```

```
 39  pers_email_pl                          1647 non-null   object
 40  laengengrad_st                         1629 non-null   float64
 41  breitengrad_st                         1629 non-null   float64
 42  Spalte1                                0 non-null      float64
 43  Nicht in Projektliste                  1647 non-null   object
 44  Bewilligung >= 2021                    1647 non-null   object
 45  Neues Projekt                          1647 non-null   bool
dtypes: bool(1), datetime64[ns](2), float64(4), int64(5), object(34)
memory usage: 580.8+ KB
```

xml teilweise nicht vollbesetzt 31 auf_bez_pub 1419 non-null object
32 auf_bez_pub_quelle 1419 non-null object
33 auf_bez_pub_en 1303 non-null object
34 auf_bez_pub_en_quelle 1303 non-null object
36 pers_titel_pl 1305 non-null object 40 laengengrad_st 1629 non-null float64
41 breitengrad_st 1629 non-null float64
42 Spalte1 0 non-null float64

2 Vergleich

funktioniert noch nicht. Es liegt wahrscheinlich an den Spaltennamen

[15]:
```
#df_xlsx[['fkz', 'db']].info()
```

[16]:
```
#df[['fkz', 'db']].info()
```

3 Für 11 Projekte stimmt die excel-Datei nicht mit der xml beim Enddatum überein

[17]:
```
col_xml = ['fkz', 'db', 'v_thema', 'fi_von/iso8601', 'fi_ende/iso8601']
col_xlsx = ['fkz', 'db','v_thema']
df_xml_part = df[col_xml]
df_xlsx_part = df_xlsx[col_xlsx]
# format und name anpassen
series_start_xlsx = df_xlsx['fi_von.iso8601'].dt.strftime('%Y-%m-%d').
 ↪rename('fi_von/iso8601')
df_xlsx_part = pd.concat([df_xlsx_part, series_start_xlsx], axis=1)
series_ende_xlsx = df_xlsx['fi_ende.iso8601'].dt.strftime('%Y-%m-%d').
 ↪rename('fi_ende/iso8601')
df_xlsx_part = pd.concat([df_xlsx_part, series_ende_xlsx], axis=1)

df_fkz_diff = pd.concat([df_xml_part, df_xlsx_part]).drop_duplicates(keep=False)
df_fkz_diff.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 22 entries, 79 to 878
Data columns (total 5 columns):
```

```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   fkz              22 non-null     object
 1   db               22 non-null     object
 2   v_thema          22 non-null     object
 3   fi_von/iso8601   22 non-null     object
 4   fi_ende/iso8601  22 non-null     object
dtypes: object(5)
memory usage: 1.0+ KB
```

[18]:
```
df_fkz_diff
```

[18]:
```
            fkz     db                                                v_thema  \
79    03ETW012D  PROFI  Verbundvorhaben: Verfahren zur softwaregestütz…
184    03ETW010  PROFI  TeBwA- Temperaturbasierte energetische Bilanzi…
192  03EGB0017B  PROFI  Verbundvorhaben: Coso - Entwicklung von Maßnah…
193  03EGB0017A  PROFI  Verbundvorhaben: CoSo - Entwicklung von Maßnah…
194  03EGB0017C  PROFI  Verbundvorhaben: Coso - Entwicklung von Maßnah…
477    03EN3001  PROFI  EnEff:Wärme - ZellFlex: Identifikation urbaner…
495   03ET1635C  PROFI  EnEff:Stadt: Drei Prozent Plus: Umsetzung des …
496   03ET1634E  PROFI  EnEff:Wärme: ErdEisII: Verbundprojekt: Erdeiss…
498   03ET1634C  PROFI  EnEff:Wärme: ErdEisII: Verbundprojekt: Erdeiss…
877  03EGB0012C  PROFI  EG2050: EffTecSomodIn: Energieeffiziente Moder…
878  03EGB0012B  PROFI  EG2050: EffTecSomodIn: Energieeffiziente Moder…
79    03ETW012D  PROFI  Verbundvorhaben: Verfahren zur softwaregestütz…
184    03ETW010  PROFI  TeBwA- Temperaturbasierte energetische Bilanzi…
192  03EGB0017B  PROFI  Verbundvorhaben: Coso - Entwicklung von Maßnah…
193  03EGB0017A  PROFI  Verbundvorhaben: CoSo - Entwicklung von Maßnah…
194  03EGB0017C  PROFI  Verbundvorhaben: Coso - Entwicklung von Maßnah…
477    03EN3001  PROFI  EnEff:Wärme - ZellFlex: Identifikation urbaner…
495   03ET1635C  PROFI  EnEff:Stadt: Drei Prozent Plus: Umsetzung des …
496   03ET1634E  PROFI  EnEff:Wärme: ErdEisII: Verbundprojekt: Erdeiss…
498   03ET1634C  PROFI  EnEff:Wärme: ErdEisII: Verbundprojekt: Erdeiss…
877  03EGB0012C  PROFI  EG2050: EffTecSomodIn: Energieeffiziente Moder…
878  03EGB0012B  PROFI  EG2050: EffTecSomodIn: Energieeffiziente Moder…

    fi_von/iso8601 fi_ende/iso8601
79      2019-01-01      2022-06-30
184     2019-03-01      2022-12-31
192     2019-03-01      2022-04-15
193     2019-03-01      2022-05-31
194     2019-03-01      2022-05-31
477     2019-05-01      2023-03-31
495     2019-01-01      2022-09-30
496     2019-03-01      2022-12-31
498     2019-03-01      2022-12-31
877     2018-07-01      2023-06-30
```

```
878     2018-07-01      2023-06-30
79      2019-01-01      2021-12-31
184     2019-03-01      2022-02-28
192     2019-03-01      2022-02-28
193     2019-03-01      2022-02-28
194     2019-03-01      2022-02-28
477     2019-05-01      2022-04-30
495     2019-01-01      2022-06-30
496     2019-03-01      2022-02-28
498     2019-03-01      2022-02-28
877     2018-07-01      2022-06-30
878     2018-07-01      2022-06-30
```

3.0.1 xlsx-Datei (auch in der xlsx-Datei in calc geprüft)

[19]:
```
df_xlsx_part[['fkz','fi_ende/iso8601']][df_xlsx_part['fkz']=='03ETW012D']
```

[19]:
```
          fkz fi_ende/iso8601
79  03ETW012D      2021-12-31
```

3.0.2 xml-Datei (auch in der xml-Datei in firefox geprüft)

[20]:
```
df_xml_part[['fkz','fi_ende/iso8601']][df_xml_part['fkz']=='03ETW012D']
```

[20]:
```
          fkz fi_ende/iso8601
79  03ETW012D      2022-06-30
```

1. Block: xml
2. Block: xlsx xml: Diffenzprojekte haben immer ein späteres End-Datum

- Vermutung: xlsx enthält die ursprünglichen/beantragten End-Daten, die Anpassung auf Grund von genehmigten Verlängerungen werden nicht in Verteiler-xlxs eingepflegt

3.1 Erste Null in der PLZ bei plz_strasse_st fehlt in der xlxs-Datei

- weiterer Vergleich ohne Ende-Datum
- Die eigentlichen Wert der PLZ sind gleich!!
- die ersten Nullen der PLZ sind schon in der xlxs nicht enthalten, dies ist noch ein Grund die nicht die xlxs als Datenquelle zu nutzen

[21]:
```
col_xml = ['fkz', 'db', 'v_thema', 'fi_von/iso8601', 'fi_sumbew/value',␣
 ↪'ver_bez','lp_nr','lp_text','name_st',
          'plz_strasse_st']
col_xlsx = ['fkz', 'db','v_thema','ver_bez','lp_nr','lp_text','name_st']
df_xml_part = df[col_xml]
df_xlsx_part = df_xlsx[col_xlsx]
# format und name anpassen
```

```
series_start_xlsx = df_xlsx['fi_von.iso8601'].dt.strftime('%Y-%m-%d').
 ↪rename('fi_von/iso8601')
df_xlsx_part = pd.concat([df_xlsx_part, series_start_xlsx], axis=1)
series_value_xlsx = df_xlsx['fi_sumbew.value'].astype('str').rename('fi_sumbew/
 ↪value')
df_xlsx_part = pd.concat([df_xlsx_part, series_value_xlsx], axis=1)

series_plz_xlsx = df_xlsx['plz_strasse_st'].astype('str')
df_xlsx_part = pd.concat([df_xlsx_part, series_plz_xlsx], axis=1)


df_fkz_diff = pd.concat([df_xml_part, df_xlsx_part]).drop_duplicates(keep=False)
df_fkz_diff.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 342 entries, 54 to 1623
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   fkz              342 non-null    object
 1   db               342 non-null    object
 2   v_thema          342 non-null    object
 3   fi_von/iso8601   342 non-null    object
 4   fi_sumbew/value  342 non-null    object
 5   ver_bez          342 non-null    object
 6   lp_nr            342 non-null    object
 7   lp_text          342 non-null    object
 8   name_st          342 non-null    object
 9   plz_strasse_st   342 non-null    object
dtypes: object(10)
memory usage: 29.4+ KB
```

[22]:
```
df_group = df_fkz_diff[['fkz','plz_strasse_st']].groupby(['fkz'])
df_group.groups
```

[22]:
```
{'0325550B': [249, 249], '0325871': [223, 223], '0327511A': [1192, 1192],
'0329663N': [1190, 1190], '0329663O': [1189, 1189], '03EGB0009A': [886, 886],
'03EGB0013A': [1187, 1187], '03EGB0013B': [1177, 1177], '03EGB0016A': [492,
492], '03EGB0016B': [489, 489], '03EGB0016C': [488, 488], '03EGB0020C': [165,
165], '03EGB0022': [1623, 1623], '03EN1001A': [1094, 1094], '03EN1001B': [1093,
1093], '03EN1001C': [1092, 1092], '03EN1006A': [1081, 1081], '03EN1009E': [1065,
1065], '03EN1020C': [1015, 1015], '03EN1022A': [1008, 1008], '03EN1028D': [983,
983], '03EN1028G': [979, 979], '03EN1029A': [978, 978], '03EN1029C': [976, 976],
'03EN1030A': [974, 974], '03EN1032D': [958, 958], '03EN1033B': [950, 950],
'03EN1033C': [949, 949], '03EN1033D': [948, 948], '03EN1034A': [946, 946],
'03EN1036A': [942, 942], '03EN1036D': [939, 939], '03EN1039C': [929, 929],
'03EN1044B': [915, 915], '03EN3001': [477, 477], '03EN3006C': [461, 461],
```

```
'03EN3006D': [460, 460], '03EN3006E': [459, 459], '03EN3006G': [457, 457],
'03EN3018C': [415, 415], '03EN3020D': [408, 408], '03EN3035A': [348, 348],
'03EN3035B': [347, 347], '03EN3035C': [346, 346], '03EN3035D': [345, 345],
'03EN3040B': [334, 334], '03EN3045B': [283, 283], '03EN6003D': [58, 58],
'03EN6004A': [57, 57], '03EN6004B': [56, 56], '03EN6004C': [55, 55],
'03EN6005A': [131, 131], '03EN6005B': [130, 130], '03EN6010A': [54, 54],
'03EN6011B': [120, 120], '03ESP225A': [867, 867], '03ESP225C': [865, 865],
'03ESP402B': [852, 852], '03ET1009C': [1572, 1572], '03ET1080A': [845, 845],
'03ET1080B': [844, 844], '03ET1080C': [843, 843], '03ET1119B': [1545, 1545],
'03ET1119D': [1543, 1543], '03ET1130B': [1538, 1538], '03ET1130C': [1537, 1537],
'03ET1155B': [832, 832], '03ET1166A': [1526, 1526], '03ET1171B': [827, 827],
'03ET1211A': [1488, 1488], '03ET1215A': [1485, 1485], '03ET1215B': [1484, 1484],
'03ET1215C': [1483, 1483], '03ET1215D': [1482, 1482], '03ET1230B': [818, 818],
'03ET1232D': [1468, 1468], '03ET1234A': [816, 816], '03ET1261B': [1452, 1452],
'03ET1267A': [1449, 1449], '03ET1268A': [1444, 1444], '03ET1268B': [1443, 1443],
'03ET1268C': [1442, 1442], '03ET1268D': [1441, 1441], '03ET1280A': [779, 779],
'03ET1284A': [1435, 1435], '03ET1287A': [1430, 1430], '03ET1299B': [1421, 1421],
'03ET1315B': [1404, 1404], '03ET1319A': [762, 762], '03ET1322A': [760, 760],
'03ET1338A': [745, 745], '03ET1358B': [726, 726], '03ET1359B': [1376, 1376],
'03ET1361A': [1372, 1372], '03ET1363A': [721, 721], '03ET1371E': [1033, 1033],
'03ET1372A': [1364, 1364], '03ET1372B': [1363, 1363], '03ET1374A': [114, 114],
'03ET1374B': [113, 113], …}
```

3.2 3 Abweichungen bei ad_str_st

- weiterer Vergleich ohne Ende-Datum und PLZ

[23]:
```
col_xml = ['fkz', 'db', 'v_thema', 'fi_von/iso8601', 'fi_sumbew/value',␣
 ↪'ver_bez','lp_nr','lp_text','name_st',
          'ort_st','ad_str_st']
col_xlsx = ['fkz',␣
 ↪'db','v_thema','ver_bez','lp_nr','lp_text','name_st','ort_st','ad_str_st']
df_xml_part = df[col_xml]
df_xlsx_part = df_xlsx[col_xlsx]
# format und name anpassen
series_start_xlsx = df_xlsx['fi_von.iso8601'].dt.strftime('%Y-%m-%d').
 ↪rename('fi_von/iso8601')
df_xlsx_part = pd.concat([df_xlsx_part, series_start_xlsx], axis=1)
series_value_xlsx = df_xlsx['fi_sumbew.value'].astype('str').rename('fi_sumbew/
 ↪value')
df_xlsx_part = pd.concat([df_xlsx_part, series_value_xlsx], axis=1)




df_fkz_diff = pd.concat([df_xml_part, df_xlsx_part]).drop_duplicates(keep=False)
df_fkz_diff.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6 entries, 457 to 1449
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   fkz              6 non-null      object
 1   db               6 non-null      object
 2   v_thema          6 non-null      object
 3   fi_von/iso8601   6 non-null      object
 4   fi_sumbew/value  6 non-null      object
 5   ver_bez          6 non-null      object
 6   lp_nr            6 non-null      object
 7   lp_text          6 non-null      object
 8   name_st          6 non-null      object
 9   ort_st           6 non-null      object
 10  ad_str_st        6 non-null      object
dtypes: object(11)
memory usage: 576.0+ bytes
```

[24]:
```
df_fkz_diff[['fkz','ad_str_st']]
```

[24]:

	fkz	ad_str_st
457	03EN3006G	Hainstr.1a
668	03ET1433G	Hainstr.1a
1449	03ET1267A	August-Bebel-Straße 30
457	03EN3006G	Leutragraben 1
668	03ET1433G	Leutragraben 1
1449	03ET1267A	George-Bähr-Str. 1

3.3 Erste Null in der PLZ bei plz_strasse_ze fehlt in der xlxs-Datei

- weiterer Vergleich ohne Ende-Datum, plz_strasse_st und ad_str_st
- Die eigentlichen Wert der PLZ sind gleich!!
- die ersten Nullen der PLZ sind schon in der xlxs nicht enthalten, dies ist noch ein Grund die nicht die xlxs als Datenquelle zu nutzen

[52]:
```
col_xml = ['fkz', 'db', 'v_thema', 'fi_von/iso8601', 'fi_sumbew/value',␣
 ↪'ver_bez','lp_nr','lp_text','name_st',
          'ort_st','land_st','gem_gemkz_st','name_ze','plz_strasse_ze']
col_xlsx = ['fkz',␣
 ↪'db','v_thema','ver_bez','lp_nr','lp_text','name_st','ort_st','land_st','name_ze']
df_xml_part = df[col_xml]
df_xlsx_part = df_xlsx[col_xlsx]
# format und name anpassen
series_start_xlsx = df_xlsx['fi_von.iso8601'].dt.strftime('%Y-%m-%d').
 ↪rename('fi_von/iso8601')
df_xlsx_part = pd.concat([df_xlsx_part, series_start_xlsx], axis=1)
```

```
series_value_xlsx = df_xlsx['fi_sumbew.value'].astype('str').rename('fi_sumbew/
↪value')
df_xlsx_part = pd.concat([df_xlsx_part, series_value_xlsx], axis=1)
series_gemkz_xlsx = df_xlsx['gem_gemkz_st'].astype('str')
df_xlsx_part = pd.concat([df_xlsx_part, series_gemkz_xlsx], axis=1)
series_plz_ze_xlsx = df_xlsx['plz_strasse_ze'].astype('str')
df_xlsx_part = pd.concat([df_xlsx_part, series_plz_ze_xlsx], axis=1)


df_fkz_diff = pd.concat([df_xml_part, df_xlsx_part]).drop_duplicates(keep=False)
df_fkz_diff.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 308 entries, 54 to 1623
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   fkz              308 non-null    object
 1   db               308 non-null    object
 2   v_thema          308 non-null    object
 3   fi_von/iso8601   308 non-null    object
 4   fi_sumbew/value  308 non-null    object
 5   ver_bez          308 non-null    object
 6   lp_nr            308 non-null    object
 7   lp_text          308 non-null    object
 8   name_st          308 non-null    object
 9   ort_st           308 non-null    object
 10  land_st          308 non-null    object
 11  gem_gemkz_st     308 non-null    object
 12  name_ze          308 non-null    object
 13  plz_strasse_ze   308 non-null    object
dtypes: object(14)
memory usage: 36.1+ KB
```

[53]:
```
df_group = df_fkz_diff[['fkz','plz_strasse_ze']].groupby(['fkz'])
df_group.groups
```

[53]:
```
{'0325550B': [249, 249], '0325871': [223, 223], '0327511A': [1192, 1192],
'0329663N': [1190, 1190], '0329663O': [1189, 1189], '03EGB0009A': [886, 886],
'03EGB0013A': [1187, 1187], '03EGB0013B': [1177, 1177], '03EGB0016A': [492,
492], '03EGB0016B': [489, 489], '03EGB0016C': [488, 488], '03EGB0018F': [186,
186], '03EGB0020C': [165, 165], '03EGB0022': [1623, 1623], '03EN1001A': [1094,
1094], '03EN1001C': [1092, 1092], '03EN1006A': [1081, 1081], '03EN1009E': [1065,
1065], '03EN1020C': [1015, 1015], '03EN1022A': [1008, 1008], '03EN1028D': [983,
983], '03EN1028G': [979, 979], '03EN1029A': [978, 978], '03EN1029C': [976, 976],
'03EN1030A': [974, 974], '03EN1032D': [958, 958], '03EN1033B': [950, 950],
'03EN1033C': [949, 949], '03EN1033D': [948, 948], '03EN1034A': [946, 946],
```

```
'03EN1036D': [939, 939], '03EN1039C': [929, 929], '03EN3001': [477, 477],
'03EN3006C': [461, 461], '03EN3006D': [460, 460], '03EN3006E': [459, 459],
'03EN3006G': [457, 457], '03EN3020D': [408, 408], '03EN3035A': [348, 348],
'03EN3035B': [347, 347], '03EN3035C': [346, 346], '03EN3035D': [345, 345],
'03EN3040B': [334, 334], '03EN3045B': [283, 283], '03EN6003D': [58, 58],
'03EN6004A': [57, 57], '03EN6004B': [56, 56], '03EN6004C': [55, 55],
'03EN6005A': [131, 131], '03EN6005B': [130, 130], '03EN6010A': [54, 54],
'03EN6011B': [120, 120], '03ESP225A': [867, 867], '03ESP225C': [865, 865],
'03ESP402B': [852, 852], '03ET1009C': [1572, 1572], '03ET1080B': [844, 844],
'03ET1080C': [843, 843], '03ET1119B': [1545, 1545], '03ET1119D': [1543, 1543],
'03ET1130B': [1538, 1538], '03ET1130C': [1537, 1537], '03ET1155B': [832, 832],
'03ET1166A': [1526, 1526], '03ET1171B': [827, 827], '03ET1211A': [1488, 1488],
'03ET1215A': [1485, 1485], '03ET1215B': [1484, 1484], '03ET1215D': [1482, 1482],
'03ET1230B': [818, 818], '03ET1232D': [1468, 1468], '03ET1234A': [816, 816],
'03ET1261B': [1452, 1452], '03ET1267A': [1449, 1449], '03ET1268A': [1444, 1444],
'03ET1268B': [1443, 1443], '03ET1268C': [1442, 1442], '03ET1268D': [1441, 1441],
'03ET1280A': [779, 779], '03ET1284A': [1435, 1435], '03ET1287A': [1430, 1430],
'03ET1299B': [1421, 1421], '03ET1315B': [1404, 1404], '03ET1319A': [762, 762],
'03ET1322A': [760, 760], '03ET1338A': [745, 745], '03ET1358B': [726, 726],
'03ET1359B': [1376, 1376], '03ET1361A': [1372, 1372], '03ET1363A': [721, 721],
'03ET1371E': [1033, 1033], '03ET1372B': [1363, 1363], '03ET1374A': [114, 114],
'03ET1374B': [113, 113], '03ET1374C': [112, 112], '03ET1382A': [713, 713],
'03ET1412B': [688, 688], '03ET1414B': [1339, 1339], '03ET1416B': [1333, 1333],
'03ET1423A': [1328, 1328], …}
```

3.4 2 Abweichungen in ad_str_ze

- weiterer Vergleich ohne Ende-Datum, plz_strasse_st, ad_str_st, plz_strasse_ze

[63]:
```
col_xml = ['fkz', 'db', 'v_thema', 'fi_von/iso8601', 'fi_sumbew/value',␣
↪'ver_bez','lp_nr','lp_text','name_st',
         'ort_st','land_st','gem_gemkz_st','name_ze', 'ort_ze', 'ad_str_ze']
col_xlsx = ['fkz',␣
↪'db','v_thema','ver_bez','lp_nr','lp_text','name_st','ort_st','land_st','name_ze',
           'ort_ze', 'ad_str_ze']
df_xml_part = df[col_xml]
df_xlsx_part = df_xlsx[col_xlsx]
# format und name anpassen
series_start_xlsx = df_xlsx['fi_von.iso8601'].dt.strftime('%Y-%m-%d').
↪rename('fi_von/iso8601')
df_xlsx_part = pd.concat([df_xlsx_part, series_start_xlsx], axis=1)
series_value_xlsx = df_xlsx['fi_sumbew.value'].astype('str').rename('fi_sumbew/
↪value')
df_xlsx_part = pd.concat([df_xlsx_part, series_value_xlsx], axis=1)
series_gemkz_xlsx = df_xlsx['gem_gemkz_st'].astype('str')
df_xlsx_part = pd.concat([df_xlsx_part, series_gemkz_xlsx], axis=1)
```

```
df_fkz_diff = pd.concat([df_xml_part, df_xlsx_part]).drop_duplicates(keep=False)
df_fkz_diff.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4 entries, 457 to 668
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   fkz              4 non-null      object
 1   db               4 non-null      object
 2   v_thema          4 non-null      object
 3   fi_von/iso8601   4 non-null      object
 4   fi_sumbew/value  4 non-null      object
 5   ver_bez          4 non-null      object
 6   lp_nr            4 non-null      object
 7   lp_text          4 non-null      object
 8   name_st          4 non-null      object
 9   ort_st           4 non-null      object
 10  land_st          4 non-null      object
 11  gem_gemkz_st     4 non-null      object
 12  name_ze          4 non-null      object
 13  ort_ze           4 non-null      object
 14  ad_str_ze        4 non-null      object
dtypes: object(15)
memory usage: 512.0+ bytes
```

[64]:
```
df_fkz_diff[['fkz','ad_str_ze']]
```

[64]:

	fkz	ad_str_ze
457	03EN3006G	Hainstr.1a
668	03ET1433G	Hainstr.1a
457	03EN3006G	Leutragraben 1
668	03ET1433G	Leutragraben 1

3.5 6 Abweichungen in pers_pl

- weiterer Vergleich ohne Ende-Datum, plz_strasse_st, ad_str_st, plz_strasse_ze, ad_str_ze
- ohne auf_bez_pub_quelle_en, weil keine Einträge in xml und keine Spalte in xlsx

[80]:
```
col_xml = ['fkz', 'db', 'v_thema', 'fi_von/iso8601', 'fi_sumbew/value',␣
↪'ver_bez','lp_nr','lp_text','name_st',
         'ort_st','land_st','gem_gemkz_st','name_ze', 'ort_ze', 'land_ze',␣
↪'gem_gemkz_ze', 'v_ressort',
          'v_pt_detail', 'v_forschsp_text','v_prog_text', 'auf_bez_pub',␣
↪'auf_bez_pub_quelle', 'auf_bez_pub_en',
         'pers_pl']
```

```
col_xlsx = ['fkz',␣
 ↪'db','v_thema','ver_bez','lp_nr','lp_text','name_st','ort_st','land_st','name_ze',
            'ort_ze', 'land_ze','v_ressort', 'v_pt_detail','v_forschsp_text',␣
 ↪'v_prog_text', 'auf_bez_pub',
           'auf_bez_pub_quelle', 'auf_bez_pub_en', 'pers_pl']
df_xml_part = df[col_xml]
df_xlsx_part = df_xlsx[col_xlsx]
# format und name anpassen
series_start_xlsx = df_xlsx['fi_von.iso8601'].dt.strftime('%Y-%m-%d').
 ↪rename('fi_von/iso8601')
df_xlsx_part = pd.concat([df_xlsx_part, series_start_xlsx], axis=1)
series_value_xlsx = df_xlsx['fi_sumbew.value'].astype('str').rename('fi_sumbew/
 ↪value')
df_xlsx_part = pd.concat([df_xlsx_part, series_value_xlsx], axis=1)
series_gemkz_xlsx = df_xlsx['gem_gemkz_st'].astype('str')
df_xlsx_part = pd.concat([df_xlsx_part, series_gemkz_xlsx], axis=1)
series_gemkz_ze_xlsx = df_xlsx['gem_gemkz_ze'].astype('str')
df_xlsx_part = pd.concat([df_xlsx_part, series_gemkz_ze_xlsx], axis=1)


df_fkz_diff = pd.concat([df_xml_part, df_xlsx_part]).drop_duplicates(keep=False)
df_fkz_diff.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12 entries, 9 to 530
Data columns (total 24 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   fkz                 12 non-null     object
 1   db                  12 non-null     object
 2   v_thema             12 non-null     object
 3   fi_von/iso8601      12 non-null     object
 4   fi_sumbew/value     12 non-null     object
 5   ver_bez             12 non-null     object
 6   lp_nr               12 non-null     object
 7   lp_text             12 non-null     object
 8   name_st             12 non-null     object
 9   ort_st              12 non-null     object
 10  land_st             12 non-null     object
 11  gem_gemkz_st        12 non-null     object
 12  name_ze             12 non-null     object
 13  ort_ze              12 non-null     object
 14  land_ze             12 non-null     object
 15  gem_gemkz_ze        12 non-null     object
 16  v_ressort           12 non-null     object
 17  v_pt_detail         12 non-null     object
 18  v_forschsp_text     12 non-null     object
```

```
 19  v_prog_text          12 non-null     object
 20  auf_bez_pub          12 non-null     object
 21  auf_bez_pub_quelle   12 non-null     object
 22  auf_bez_pub_en       12 non-null     object
 23  pers_pl              12 non-null     object
dtypes: object(24)
memory usage: 2.3+ KB
```

[82]:
```
df_fkz_diff[['fkz','pers_pl']]
```

[82]:

	fkz	pers_pl
9	03EWR008E	Dr. Dietmar Schaal
10	03EWR008D	Dr. Dietmar Schaal
146	03ETW022A	Dr. Christoph Maurer
303	03EN3049B	Dr. Wiebke Hofacker
445	03EN3008G	Dr. Ing. Kevin Förderer
530	03ET1618A	Wirtschaftsingenieur Denise Graef
9	03EWR008E	Dr. Maximilian Seier
10	03EWR008D	Dr. Maximilian Seier
146	03ETW022A	Dr. Bruno Bueno
303	03EN3049B	Dr. Wiebke Harms
445	03EN3008G	Dr. Clemens Düpmeier
530	03ET1618A	Wirtschaftsingenieur David Pflegler

3.6 1 Abweichungen in pers_titel_pl

- weiterer Vergleich ohne Ende-Datum, plz_strasse_st, ad_str_st, plz_strasse_ze, ad_str_ze, pers_pl
- ohne auf_bez_pub_quelle_en, weil keine Einträge in xml und keine Spalte in xlsx

[85]:
```
col_xml = ['fkz', 'db', 'v_thema', 'fi_von/iso8601', 'fi_sumbew/value',␣
 ↪'ver_bez','lp_nr','lp_text','name_st',
         'ort_st','land_st','gem_gemkz_st','name_ze', 'ort_ze', 'land_ze',␣
 ↪'gem_gemkz_ze', 'v_ressort',
          'v_pt_detail', 'v_forschsp_text','v_prog_text', 'auf_bez_pub',␣
 ↪'auf_bez_pub_quelle', 'auf_bez_pub_en'
         , 'pers_titel_pl']
col_xlsx = ['fkz',␣
 ↪'db','v_thema','ver_bez','lp_nr','lp_text','name_st','ort_st','land_st','name_ze',
           'ort_ze', 'land_ze','v_ressort', 'v_pt_detail','v_forschsp_text',␣
 ↪'v_prog_text', 'auf_bez_pub',
          'auf_bez_pub_quelle', 'auf_bez_pub_en', 'pers_titel_pl']
df_xml_part = df[col_xml]
df_xlsx_part = df_xlsx[col_xlsx]
# format und name anpassen
series_start_xlsx = df_xlsx['fi_von.iso8601'].dt.strftime('%Y-%m-%d').
 ↪rename('fi_von/iso8601')
```

```
df_xlsx_part = pd.concat([df_xlsx_part, series_start_xlsx], axis=1)
series_value_xlsx = df_xlsx['fi_sumbew.value'].astype('str').rename('fi_sumbew/
↪value')
df_xlsx_part = pd.concat([df_xlsx_part, series_value_xlsx], axis=1)
series_gemkz_xlsx = df_xlsx['gem_gemkz_st'].astype('str')
df_xlsx_part = pd.concat([df_xlsx_part, series_gemkz_xlsx], axis=1)
series_gemkz_ze_xlsx = df_xlsx['gem_gemkz_ze'].astype('str')
df_xlsx_part = pd.concat([df_xlsx_part, series_gemkz_ze_xlsx], axis=1)


df_fkz_diff = pd.concat([df_xml_part, df_xlsx_part]).drop_duplicates(keep=False)
df_fkz_diff.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2 entries, 445 to 445
Data columns (total 24 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   fkz                  2 non-null      object
 1   db                   2 non-null      object
 2   v_thema              2 non-null      object
 3   fi_von/iso8601       2 non-null      object
 4   fi_sumbew/value      2 non-null      object
 5   ver_bez              2 non-null      object
 6   lp_nr                2 non-null      object
 7   lp_text              2 non-null      object
 8   name_st              2 non-null      object
 9   ort_st               2 non-null      object
 10  land_st              2 non-null      object
 11  gem_gemkz_st         2 non-null      object
 12  name_ze              2 non-null      object
 13  ort_ze               2 non-null      object
 14  land_ze              2 non-null      object
 15  gem_gemkz_ze         2 non-null      object
 16  v_ressort            2 non-null      object
 17  v_pt_detail          2 non-null      object
 18  v_forschsp_text      2 non-null      object
 19  v_prog_text          2 non-null      object
 20  auf_bez_pub          2 non-null      object
 21  auf_bez_pub_quelle   2 non-null      object
 22  auf_bez_pub_en       2 non-null      object
 23  pers_titel_pl        2 non-null      object
dtypes: object(24)
memory usage: 400.0+ bytes
```

[86]:
```
df_fkz_diff[['fkz','pers_titel_pl']]
```

[86]:
```
            fkz pers_titel_pl
445  03EN3008G      Dr. Ing.
445  03EN3008G           Dr.
```

3.7 7 Abweichungen in pers_vname_pl + pers_name_pl + pers_email_pl

- weiterer Vergleich ohne Ende-Datum, plz_strasse_st, ad_str_st, plz_strasse_ze, ad_str_ze, pers_pl, pers_titel_pl
- ohne auf_bez_pub_quelle_en, weil keine Einträge in xml und keine Spalte in xlsx

[100]:
```
col_xml = ['fkz', 'db', 'v_thema', 'fi_von/iso8601', 'fi_sumbew/value',␣
 ↪'ver_bez','lp_nr','lp_text','name_st',
          'ort_st','land_st','gem_gemkz_st','name_ze', 'ort_ze', 'land_ze',␣
 ↪'gem_gemkz_ze', 'v_ressort',
           'v_pt_detail', 'v_forschsp_text','v_prog_text', 'auf_bez_pub',␣
 ↪'auf_bez_pub_quelle', 'auf_bez_pub_en'
          , 'pers_vname_pl', 'pers_name_pl', 'pers_email_pl']
col_xlsx = ['fkz',␣
 ↪'db','v_thema','ver_bez','lp_nr','lp_text','name_st','ort_st','land_st','name_ze',
            'ort_ze', 'land_ze','v_ressort', 'v_pt_detail','v_forschsp_text',␣
 ↪'v_prog_text', 'auf_bez_pub',
           'auf_bez_pub_quelle', 'auf_bez_pub_en', 'pers_vname_pl',␣
 ↪'pers_name_pl', 'pers_email_pl']
df_xml_part = df[col_xml]
df_xlsx_part = df_xlsx[col_xlsx]
# format und name anpassen
series_start_xlsx = df_xlsx['fi_von.iso8601'].dt.strftime('%Y-%m-%d').
 ↪rename('fi_von/iso8601')
df_xlsx_part = pd.concat([df_xlsx_part, series_start_xlsx], axis=1)
series_value_xlsx = df_xlsx['fi_sumbew.value'].astype('str').rename('fi_sumbew/
 ↪value')
df_xlsx_part = pd.concat([df_xlsx_part, series_value_xlsx], axis=1)
series_gemkz_xlsx = df_xlsx['gem_gemkz_st'].astype('str')
df_xlsx_part = pd.concat([df_xlsx_part, series_gemkz_xlsx], axis=1)
series_gemkz_ze_xlsx = df_xlsx['gem_gemkz_ze'].astype('str')
df_xlsx_part = pd.concat([df_xlsx_part, series_gemkz_ze_xlsx], axis=1)


df_fkz_diff = pd.concat([df_xml_part, df_xlsx_part]).drop_duplicates(keep=False)
df_fkz_diff.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 14 entries, 9 to 530
Data columns (total 26 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   fkz                 14 non-null     object
 1   db                  14 non-null     object
```

```
 2   v_thema             14 non-null     object
 3   fi_von/iso8601      14 non-null     object
 4   fi_sumbew/value     14 non-null     object
 5   ver_bez             14 non-null     object
 6   lp_nr               14 non-null     object
 7   lp_text             14 non-null     object
 8   name_st             14 non-null     object
 9   ort_st              14 non-null     object
 10  land_st             14 non-null     object
 11  gem_gemkz_st        14 non-null     object
 12  name_ze             14 non-null     object
 13  ort_ze              14 non-null     object
 14  land_ze             14 non-null     object
 15  gem_gemkz_ze        14 non-null     object
 16  v_ressort           14 non-null     object
 17  v_pt_detail         14 non-null     object
 18  v_forschsp_text     14 non-null     object
 19  v_prog_text         14 non-null     object
 20  auf_bez_pub         14 non-null     object
 21  auf_bez_pub_quelle  14 non-null     object
 22  auf_bez_pub_en      14 non-null     object
 23  pers_vname_pl       14 non-null     object
 24  pers_name_pl        14 non-null     object
 25  pers_email_pl       14 non-null     object
dtypes: object(26)
memory usage: 3.0+ KB
```

[97]:
```
df_fkz_diff[['fkz','pers_vname_pl', 'pers_name_pl', 'pers_email_pl']]
```

[97]:
```
           fkz pers_vname_pl pers_name_pl  \
9    03EWR008E       Dietmar       Schaal
10   03EWR008D       Dietmar       Schaal
146  03ETW022A     Christoph       Maurer
303  03EN3049B        Wiebke     Hofacker
390  03EN3024C     Christian        Hüttl
445  03EN3008G         Kevin     Förderer
530  03ET1618A        Denise        Graef
9    03EWR008E    Maximilian        Seier
10   03EWR008D    Maximilian        Seier
146  03ETW022A         Bruno        Bueno
303  03EN3049B        Wiebke        Harms
390  03EN3024C     Christian        Hüttl
445  03EN3008G       Clemens     Düpmeier
530  03ET1618A         David     Pflegler

                            pers_email_pl
9                       d.schaal@enbw.com
```

```
10                         d.schaal@enbw.com
146       christoph.maurer@ise.fraunhofer.de
303  wiebke.hofacker@stadtwerke-karlsruhe.de
390      christian.huettl@siemens-energy.com
445                  kevin.foerderer@kit.edu
530                   denise.graef@kea-bw.de
9                           m.seier@enbw.com
10                          m.seier@enbw.com
146            bruno.bueno@ise.fraunhofer.de
303     wiebke.harms@stadtwerke-karlsruhe.de
390              christian.huettl@siemens.com
445                 clemens.duepmeier@kit.edu
530                  david.pflegler@kea-bw.de
```

4 Zusammenfassung

4.1 Abweichung in:

- fi_ende.iso8601 (Ende-Datum),
- plz_strasse_st,
- ad_str_st,
- plz_strasse_ze,
- ad_str_ze, pers_pl,
- pers_titel_pl
- pers_vname_pl
- pers_name_pl
- pers_email_pl

4.2 Keine Daten in

- auf_bez_pub_quelle_en
- ohne auf_bez_pub_quelle_en, weil keine Einträge in xml und keine Spalte in xlsx

4.3 Unvollständig in

- auf_bez_pub 1419 non-null object
- auf_bez_pub_quelle 1419 non-null object
- auf_bez_pub_en 1303 non-null object
- auf_bez_pub_en_quelle 1303 non-null object
- pers_titel_pl 1305 non-null object
- laengengrad_st 1629 non-null float64
- breitengrad_st 1629 non-null float64

- Spalte1 0 non-null float64

4.4 Empfehlung: xml-Datei nutzen für die EnArgus-Infos, weil die xlxs (Verteiler) nicht den aktuellsten Stand der EnArgus-Infos enthält