EDA_Begin

January 31, 2021

[3]:
```
import pandas as pd
import numpy as np
import seaborn as sb

# URL: https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/
 ↪dd4580c810204019a7b8eb3e0b329dd6_0
df = pd.read_csv("RKI_COVID19.csv")
df_a = df["Altersgruppe"]

x = df.describe()

print(x)
```

```
           ObjectId  IdBundesland    AnzahlFall  AnzahlTodesfall  \
count  1.134996e+06  1.134996e+06  1.134996e+06     1.134996e+06
mean   5.674985e+05  7.925296e+00  1.931618e+00     4.908828e-02
std    3.276453e+05  3.577952e+00  3.087737e+00     2.809964e-01
min    1.000000e+00  1.000000e+00 -5.000000e+00    -1.000000e+00
25%    2.837498e+05  5.000000e+00  1.000000e+00     0.000000e+00
50%    5.674985e+05  8.000000e+00  1.000000e+00     0.000000e+00
75%    8.512472e+05  9.000000e+00  2.000000e+00     0.000000e+00
max    1.134996e+06  1.600000e+01  1.640000e+02     1.500000e+01

         IdLandkreis     NeuerFall  NeuerTodesfall    NeuGenesen  \
count  1.134996e+06  1.134996e+06    1.134996e+06  1.134996e+06
mean   8.253901e+03  6.401785e-03   -8.636367e+00 -1.136446e+00
std    3.541520e+03  8.455908e-02    1.774146e+00  3.001681e+00
min    1.001000e+03 -1.000000e+00   -9.000000e+00 -9.000000e+00
25%    5.566000e+03  0.000000e+00   -9.000000e+00  0.000000e+00
50%    8.216000e+03  0.000000e+00   -9.000000e+00  0.000000e+00
75%    9.762000e+03  0.000000e+00   -9.000000e+00  0.000000e+00
max    1.607700e+04  1.000000e+00    1.000000e+00  1.000000e+00

       AnzahlGenesen  IstErkrankungsbeginn
count   1.134996e+06          1.134996e+06
mean    1.672409e+00          6.793460e-01
std     2.955976e+00          4.667282e-01
min    -3.000000e+00          0.000000e+00
```

```
25%     1.000000e+00          0.000000e+00
50%     1.000000e+00          1.000000e+00
75%     2.000000e+00          1.000000e+00
max     1.640000e+02          1.000000e+00
```

[4]:
```
df_a.describe()
```

[4]:
```
count     1134996
unique          7
top       A35-A59
freq       386139
Name: Altersgruppe, dtype: object
```

[5]:
```
df.describe(include=object)
```

[5]:
```
                Bundesland   Landkreis Altersgruppe Geschlecht  \
count              1134996     1134996      1134996    1134996
unique                  16         412            7          3
top     Nordrhein-Westfalen  SK Hamburg      A35-A59          W
freq                231920       14034       386139     579892

                Meldedatum             Datenstand             Refdatum  \
count              1134996                1134996              1134996
unique                 380                      1                  394
top     2021/01/05 00:00:00  29.01.2021, 00:00 Uhr  2020/12/16 00:00:00
freq                 13109                1134996                10996

           Altersgruppe2
count            1134996
unique                 1
top     Nicht übermittelt
freq             1134996
```

[6]:
```
df.describe(include=np.number)
```

[6]:
```
           ObjectId  IdBundesland    AnzahlFall  AnzahlTodesfall  \
count  1.134996e+06  1.134996e+06  1.134996e+06     1.134996e+06
mean   5.674985e+05  7.925296e+00  1.931618e+00     4.908828e-02
std    3.276453e+05  3.577952e+00  3.087737e+00     2.809964e-01
min    1.000000e+00  1.000000e+00 -5.000000e+00    -1.000000e+00
25%    2.837498e+05  5.000000e+00  1.000000e+00     0.000000e+00
50%    5.674985e+05  8.000000e+00  1.000000e+00     0.000000e+00
75%    8.512472e+05  9.000000e+00  2.000000e+00     0.000000e+00
max    1.134996e+06  1.600000e+01  1.640000e+02     1.500000e+01

        IdLandkreis     NeuerFall  NeuerTodesfall    NeuGenesen  \
count  1.134996e+06  1.134996e+06    1.134996e+06  1.134996e+06
mean   8.253901e+03  6.401785e-03   -8.636367e+00 -1.136446e+00
```

```
std    3.541520e+03  8.455908e-02    1.774146e+00  3.001681e+00
min    1.001000e+03 -1.000000e+00   -9.000000e+00 -9.000000e+00
25%    5.566000e+03  0.000000e+00   -9.000000e+00  0.000000e+00
50%    8.216000e+03  0.000000e+00   -9.000000e+00  0.000000e+00
75%    9.762000e+03  0.000000e+00   -9.000000e+00  0.000000e+00
max    1.607700e+04  1.000000e+00    1.000000e+00  1.000000e+00

       AnzahlGenesen  IstErkrankungsbeginn
count   1.134996e+06          1.134996e+06
mean    1.672409e+00          6.793460e-01
std     2.955976e+00          4.667282e-01
min    -3.000000e+00          0.000000e+00
25%     1.000000e+00          0.000000e+00
50%     1.000000e+00          1.000000e+00
75%     2.000000e+00          1.000000e+00
max     1.640000e+02          1.000000e+00
```

[7]:
```
df["NeuerFall"].describe()
```

[7]:
```
count    1.134996e+06
mean     6.401785e-03
std      8.455908e-02
min     -1.000000e+00
25%      0.000000e+00
50%      0.000000e+00
75%      0.000000e+00
max      1.000000e+00
Name: NeuerFall, dtype: float64
```

[11]:
```
df_single = df["NeuerFall"]
df_single.describe().map('{:,.2f}'.format)
```

[11]:
```
count    1,134,996.00
mean             0.01
std              0.08
min             -1.00
25%              0.00
50%              0.00
75%              0.00
max              1.00
Name: NeuerFall, dtype: object
```

[12]:
```
df_NeuerFall = df["NeuerFall"]
df_NeuerFall.head()
```

[12]:
```
0    0
1    0
```

```
2    0
3    0
4    0
Name: NeuerFall, dtype: int64
```

[16]:
```
for col in df.columns:
    print(col)
```

```
ObjectId
IdBundesland
Bundesland
Landkreis
Altersgruppe
Geschlecht
AnzahlFall
AnzahlTodesfall
Meldedatum
IdLandkreis
Datenstand
NeuerFall
NeuerTodesfall
Refdatum
NeuGenesen
AnzahlGenesen
IstErkrankungsbeginn
Altersgruppe2
```

[27]:
```
df["NeuerFall"].describe().map('{:,.2f}'.format)
```

[27]:
```
count    1,134,996.00
mean             0.01
std              0.08
min             -1.00
25%              0.00
50%              0.00
75%              0.00
max              1.00
Name: NeuerFall, dtype: object
```

[31]:
```
df[["NeuerFall","AnzahlFall"]].describe()
```

[31]:

	NeuerFall	AnzahlFall
count	1.134996e+06	1.134996e+06
mean	6.401785e-03	1.931618e+00
std	8.455908e-02	3.087737e+00
min	-1.000000e+00	-5.000000e+00
25%	0.000000e+00	1.000000e+00
50%	0.000000e+00	1.000000e+00

```
75%    0.000000e+00  2.000000e+00
max    1.000000e+00  1.640000e+02
```

[33]:
```
df["AnzahlFall"].describe().map('{:,.2f}'.format)
```

[33]:
```
count    1,134,996.00
mean             1.93
std              3.09
min             -5.00
25%              1.00
50%              1.00
75%              2.00
max            164.00
Name: AnzahlFall, dtype: object
```

[39]:
```
df_sub = df.loc[df["NeuerFall"]==1]
df_sub.head()
```

[39]:

	ObjectId	IdBundesland	Bundesland	Landkreis	Altersgruppe	\
507	508	1	Schleswig-Holstein	SK Kiel	A15-A34	
508	509	1	Schleswig-Holstein	SK Kiel	A15-A34	
570	571	1	Schleswig-Holstein	SK Flensburg	A35-A59	
731	732	1	Schleswig-Holstein	SK Flensburg	A35-A59	
767	768	1	Schleswig-Holstein	SK Flensburg	A60-A79	

	Geschlecht	AnzahlFall	AnzahlTodesfall	Meldedatum	IdLandkreis	\
507	M	1	0	2021/01/28 00:00:00	1002	
508	M	1	0	2021/01/28 00:00:00	1002	
570	unbekannt	6	0	2021/01/28 00:00:00	1001	
731	W	2	0	2021/01/28 00:00:00	1001	
767	M	2	0	2021/01/28 00:00:00	1001	

	Datenstand	NeuerFall	NeuerTodesfall	Refdatum	\
507	29.01.2021, 00:00 Uhr	1	-9	2021/01/25 00:00:00	
508	29.01.2021, 00:00 Uhr	1	-9	2021/01/28 00:00:00	
570	29.01.2021, 00:00 Uhr	1	-9	2021/01/28 00:00:00	
731	29.01.2021, 00:00 Uhr	1	-9	2021/01/28 00:00:00	
767	29.01.2021, 00:00 Uhr	1	-9	2021/01/28 00:00:00	

	NeuGenesen	AnzahlGenesen	IstErkrankungsbeginn	Altersgruppe2
507	-9	0	1	Nicht übermittelt
508	-9	0	0	Nicht übermittelt
570	-9	0	0	Nicht übermittelt
731	-9	0	0	Nicht übermittelt
767	-9	0	0	Nicht übermittelt

[45]:
```
df_sub = df.loc[df["NeuerFall"]==1]

df_sub.head()
```

[45]:

	ObjectId	IdBundesland	Bundesland	Landkreis	Altersgruppe	\
507	508	1	Schleswig-Holstein	SK Kiel	A15-A34	
508	509	1	Schleswig-Holstein	SK Kiel	A15-A34	
570	571	1	Schleswig-Holstein	SK Flensburg	A35-A59	
731	732	1	Schleswig-Holstein	SK Flensburg	A35-A59	
767	768	1	Schleswig-Holstein	SK Flensburg	A60-A79	

	Geschlecht	AnzahlFall	AnzahlTodesfall	Meldedatum	IdLandkreis	\
507	M	1	0	2021/01/28 00:00:00	1002	
508	M	1	0	2021/01/28 00:00:00	1002	
570	unbekannt	6	0	2021/01/28 00:00:00	1001	
731	W	2	0	2021/01/28 00:00:00	1001	
767	M	2	0	2021/01/28 00:00:00	1001	

	Datenstand	NeuerFall	NeuerTodesfall	Refdatum	\
507	29.01.2021, 00:00 Uhr	1	-9	2021/01/25 00:00:00	
508	29.01.2021, 00:00 Uhr	1	-9	2021/01/28 00:00:00	
570	29.01.2021, 00:00 Uhr	1	-9	2021/01/28 00:00:00	
731	29.01.2021, 00:00 Uhr	1	-9	2021/01/28 00:00:00	
767	29.01.2021, 00:00 Uhr	1	-9	2021/01/28 00:00:00	

	NeuGenesen	AnzahlGenesen	IstErkrankungsbeginn	Altersgruppe2
507	-9	0	1	Nicht übermittelt
508	-9	0	0	Nicht übermittelt
570	-9	0	0	Nicht übermittelt
731	-9	0	0	Nicht übermittelt
767	-9	0	0	Nicht übermittelt

[42]:
```
df["Meldedatum"].min()
```

[42]:
```
'2020/01/02 00:00:00'
```

[ ]:
```

```