0.02-PS-loading-and-cleaning

November 15, 2024

[1]:
```
import os
import sys

sys.path.append(os.path.dirname(os.getcwd()))

from src.load_covid19 import load_covid19
import pandas as pd
```

```
Warning: Your Kaggle API key is readable by other users on this system! To fix
this, you can run 'chmod 600 /home/philipp/.kaggle/kaggle.json'
```

[2]:
```
df = load_covid19()
```

```
Dataset already exists at /home/philipp/Dokumente/Master_Data_Science_Fernuni_Ha
gen/Projektpraktikum_Web_Science/covid-19-risiko-
erkennung/src/../data/raw/covid19-dataset. Skipping download.
```

[3]:
```
# Specify column types and map Boolean variables
bool_columns = ['PNEUMONIA', 'PREGNANT', 'DIABETES', 'COPD', 'ASTHMA',␣
 ↪'INMSUPR',
                'HIPERTENSION', 'CARDIOVASCULAR', 'RENAL_CHRONIC',␣
 ↪'OTHER_DISEASE', 'OBESITY', 'TOBACCO',
                'INTUBED', 'ICU']

missing_values=[97, 99]

# Convert Boolean columns to 'bool' and map values (Yes -> 1, No -> 0)
for col in bool_columns:
    df[col] = df[col].map({1: 1, 2: 0}).astype('boolean')
df.replace({col: missing_values for col in bool_columns if col in df.columns},␣
 ↪pd.NA, inplace=True)

df['SEX'] = df['SEX'].map({1: 'female', 2: 'male'})
df.replace('SEX', pd.NA, inplace=True)
df['SEX'] = df['SEX'].astype('category')

df['PATIENT_TYPE'] = df['PATIENT_TYPE'].map({1: 'returned home', 2:␣
 ↪'hospitalization'})
```

```
df.replace('PATIENT_TYPE', pd.NA, inplace=True)
df['PATIENT_TYPE'] = df['PATIENT_TYPE'].astype('category')

# DATE_DIED column missing value is '9999-99-99'
df['DATE_DIED'] = pd.to_datetime(df['DATE_DIED'].replace('9999-99-99', pd.NA),␣
 ↪errors='coerce')

# Replace DATE_DIED with DIED (True if actual date, False otherwise)
df['DIED'] = df['DATE_DIED'].notna().astype('boolean')

# Drop the original DATE_DIED column
df.drop('DATE_DIED', axis=1, inplace=True)


print("\nColumns and Data Types:\n", df.dtypes)
```

```
Columns and Data Types:
 USMER                    int64
MEDICAL_UNIT              int64
SEX                    category
PATIENT_TYPE           category
INTUBED                 boolean
PNEUMONIA               boolean
AGE                       int64
PREGNANT                boolean
DIABETES                boolean
COPD                    boolean
ASTHMA                  boolean
INMSUPR                 boolean
HIPERTENSION            boolean
OTHER_DISEASE           boolean
CARDIOVASCULAR          boolean
OBESITY                 boolean
RENAL_CHRONIC           boolean
TOBACCO                 boolean
CLASIFFICATION_FINAL      int64
ICU                     boolean
DIED                    boolean
dtype: object
```

[4]:
```
df
```

[4]:

	USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	INTUBED	PNEUMONIA	\
0	2	1	female	returned home	<NA>	True	
1	2	1	male	returned home	<NA>	True	
2	2	1	male	hospitalization	True	False	

```
3            2             1  female    returned home     <NA>      False
4            2             1    male    returned home     <NA>      False
…          …             …     …              …      …        …
1048570      2            13    male    returned home     <NA>      False
1048571      1            13    male  hospitalization    False      False
1048572      2            13    male    returned home     <NA>      False
1048573      2            13    male    returned home     <NA>      False
1048574      2            13    male    returned home     <NA>      False

          AGE  PREGNANT  DIABETES   COPD  …  INMSUPR  HIPERTENSION  \
0          65     False     False  False  …    False          True
1          72      <NA>     False  False  …    False          True
2          55      <NA>      True  False  …    False         False
3          53     False     False  False  …    False         False
4          68      <NA>      True  False  …    False          True
…       …        …        …     …  …      …           …
1048570   40      <NA>     False  False  …    False         False
1048571   51      <NA>     False  False  …    False          True
1048572   55      <NA>     False  False  …    False         False
1048573   28      <NA>     False  False  …    False         False
1048574   52      <NA>     False  False  …    False         False

          OTHER_DISEASE  CARDIOVASCULAR  OBESITY  RENAL_CHRONIC  TOBACCO  \
0                 False           False    False          False    False
1                 False           False     True           True    False
2                 False           False    False          False    False
3                 False           False    False          False    False
4                 False           False    False          False    False
…                 …              …      …           …      …
1048570           False           False    False          False    False
1048571           False           False    False          False    False
1048572           False           False    False          False    False
1048573           False           False    False          False    False
1048574           False           False    False          False    False

          CLASIFFICATION_FINAL    ICU   DIED
0                            3   <NA>   True
1                            5   <NA>   True
2                            3  False   True
3                            7   <NA>   True
4                            3   <NA>  False
…                         …    …    …
1048570                      7   <NA>  False
1048571                      7  False  False
1048572                      7   <NA>  False
1048573                      7   <NA>  False
1048574                      7   <NA>  False
```

```
[1048575 rows x 21 columns]
```

[5]:
```
df.head()
```

[5]:
```
   USMER  MEDICAL_UNIT     SEX     PATIENT_TYPE  INTUBED  PNEUMONIA  AGE  \
0      2             1  female    returned home     <NA>       True   65
1      2             1    male    returned home     <NA>       True   72
2      2             1    male  hospitalization     True      False   55
3      2             1  female    returned home     <NA>      False   53
4      2             1    male    returned home     <NA>      False   68

   PREGNANT  DIABETES   COPD  …  INMSUPR  HIPERTENSION  OTHER_DISEASE  \
0     False     False  False  …    False          True          False
1      <NA>     False  False  …    False          True          False
2      <NA>      True  False  …    False         False          False
3     False     False  False  …    False         False          False
4      <NA>      True  False  …    False          True          False

   CARDIOVASCULAR  OBESITY  RENAL_CHRONIC  TOBACCO  CLASIFFICATION_FINAL  \
0           False    False          False    False                     3
1           False     True           True    False                     5
2           False    False          False    False                     3
3           False    False          False    False                     7
4           False    False          False    False                     3

     ICU   DIED
0   <NA>   True
1   <NA>   True
2  False   True
3   <NA>   True
4   <NA>  False

[5 rows x 21 columns]
```

[ ]: