0.03-PS-testing-load-and-clean-utils

November 15, 2024

[1]:

```
import os
import sys

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

from src.load_covid19 import load_clean_covid19
import pandas as pd
```

[2]:

```
df = load_clean_covid19()
```

```
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'
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.
Saving clean dataset to: /home/philipp/Dokumente/Master_Data_Science_Fernuni_Hag
en/Projektpraktikum_Web_Science/covid-19-risiko-erkennung/data/interim/covid-
data-clean.csv
Saved
Loading clean dataset from: /home/philipp/Dokumente/Master_Data_Science_Fernuni_
Hagen/Projektpraktikum_Web_Science/covid-19-risiko-erkennung/data/interim/covid-
data-clean.csv
```

[3]:

```
df
```

[3]:

```
         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]
```

[4]:
```
df.dtypes
```

[4]:
```
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
```

[5]:

```
missing_values_count = df.isna().sum()
print(missing_values_count)
```

```
USMER                        0
MEDICAL_UNIT                 0
SEX                          0
PATIENT_TYPE                 0
INTUBED                 855869
PNEUMONIA                16003
AGE                          0
PREGNANT                527265
DIABETES                  3338
COPD                      3003
ASTHMA                    2979
INMSUPR                   3404
HIPERTENSION              3104
OTHER_DISEASE             5045
CARDIOVASCULAR            3076
OBESITY                   3032
RENAL_CHRONIC             3006
TOBACCO                   3220
CLASIFFICATION_FINAL         0
ICU                     856032
DIED                         0
dtype: int64
```

[ ]: