Tutoriel 7

April 17, 2021

1 Detection and Removal of Outliers :

[5]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
```

[11]:

```
Data = pd.read_csv("/work/Data_VIIRS_NEW_REGIONS.csv")
Data.drop(Data.columns[0], axis=1, inplace=True)
Date = pd.date_range('2014-01-01', periods=84, freq='MS')
Data.insert(0, "Date", Date, True)
Data = Data.set_index('Date')
del Data['date']
Data
```

[11]:

```
                 Oriental  Tanger-Tétouan-Al Hoceima     Fès-Meknès  \
Date
2014-01-01  131843.684245               93092.002708   94494.472339
2014-02-01  129066.644328               85558.971136   85411.066837
2014-03-01  125839.063910               90001.294267   94277.234440
2014-04-01  114641.407431               95796.142398   92468.821872
2014-05-01  125517.813221              103743.237469   96385.390351
...                   ...                        ...            ...
2020-08-01  182982.674992              159545.892047  144114.066275
2020-09-01  192693.403151              141333.265296  144521.849355
2020-10-01  201118.837390              149759.585637  150276.394538
2020-11-01  234669.928200              157144.735590  167754.831608
2020-12-01  235208.519730              141878.598261  163949.861532

            Beni Mellal-Khénifra  Rabat-Salé-Kénitra  Casablanca-Settat  \
Date
2014-01-01          63062.878162        97389.183820      137652.065411
2014-02-01          62965.905268        88985.070791      139918.478415
2014-03-01          68628.444425        95817.741643      146854.777105
2014-04-01          63216.047767       100797.581497      159795.382707
2014-05-01          63336.806493       102448.139354      144307.671018
```

```
...                         ...            ...            ...
2020-08-01         96417.578373  144933.657921  209516.482522
2020-09-01         99028.404725  136486.119620  220802.403877
2020-10-01        104313.292250  159719.427016  230141.663064
2020-11-01        113609.527020  166976.017233  237822.540884
2020-12-01        118485.092392  149984.126055  212080.417283

            Marrakech-Safi  Draa-Tafilalet   Souss-Massa  Guelmim-Oued Noun  \
Date
2014-01-01   101159.418219    70574.035416  73239.149877       26154.604527
2014-02-01   106477.616256    73125.796180  85314.498283       35724.736972
2014-03-01   113422.364932    87102.006865  92265.322253       48753.137350
2014-04-01   108056.071635    67452.563176  81395.676098       39820.960684
2014-05-01    98142.952788    51233.214201  58526.327792       22925.139473
...                    ...             ...           ...                ...
2020-08-01   147459.564948   125832.830426 114311.901612       62033.772469
2020-09-01   154607.118698   144679.565911 138200.553976       81150.026141
2020-10-01   155396.329494   137628.628359 129131.499784       70231.886348
2020-11-01   173342.568565   180683.211590 158594.703654       96660.953949
2020-12-01   168090.190905   199489.895870 154871.727974       89366.522938

            Laayoune-Sakia-El-Hamra  Dakhla-Oued Ed-Dahab
Date
2014-01-01             71744.245857          32942.090562
2014-02-01             86537.865948          48540.585423
2014-03-01            159659.931900         119392.083914
2014-04-01            134870.148482         109252.289501
2014-05-01             74541.111365          44178.139672
...                             ...                   ...
2020-08-01            182427.896096         145599.580255
2020-09-01            261474.966889         207152.904109
2020-10-01            216067.795730         189088.816849
2020-11-01            295499.072240         242521.028459
2020-12-01            256522.026617         180940.780257

[84 rows x 12 columns]
```

[12]:

```
#Remove outliers columny by column
#Example
for x in ['Laayoune-Sakia-El-Hamra']:
    q75,q25 = np.percentile(Data.loc[:,x],[75,25])
    intr_qr = q75-q25

    max = q75+(1.5*intr_qr)
    min = q25-(1.5*intr_qr)

    Data.loc[Data[x] < min,x] = np.nan
    Data.loc[Data[x] > max,x] = np.nan
```

Having replaced the outliers with nan, let us now check the sum of null values or missing values using the below code:

```
[14]: Data.isnull().sum()
```

```
[14]: Oriental                    0
      Tanger-Tétouan-Al Hoceima   0
      Fès-Meknès                  0
      Beni Mellal-Khénifra        0
      Rabat-Salé-Kénitra          0
      Casablanca-Settat           0
      Marrakech-Safi              0
      Draa-Tafilalet              0
      Souss-Massa                 0
      Guelmim-Oued Noun           0
      Laayoune-Sakia-El-Hamra     0
      Dakhla-Oued Ed-Dahab        0
      dtype: int64
```

2 Imputation of the missing values with Mean, median or KNN :

https://www.askpython.com/python/examples/impute-missing-data-values