Laporan Praktikum
Pembelajaran Mesin

Disusun oleh :
Kelompok 3

1. Gede Rangga Wira Aditya (082011633048)
2. Muhammad Rahmadhani Ferdiansyah (082011633068)
3. Arya Danu Triatmodjo (082011633069)
4. Mukhamad Ikhsanudin (082011633086)

Kelas I3

PROGRAM STUDI S1 SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS AIRLANGGA

SURABAYA

2021/2022

Chapter 2

Data Preprocessing

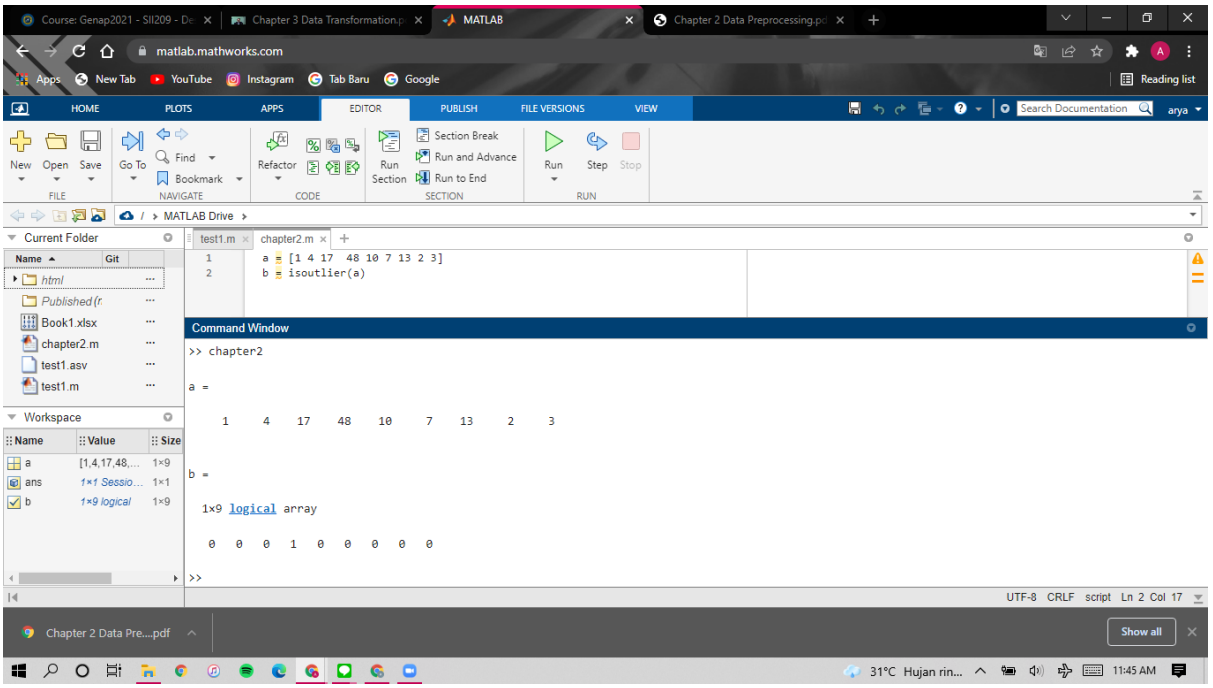
MATLAB

1. Outlier

1.1 Outlier Detection

Outlier detection used to handle outlier. Outlier detection used is isoutlier *function*

Result :



Value 0 = not *outliers*, and value 1 = *outliers*.

1.2 Outlier Handling

1. filloutlier

Source code :

```
3
4   a2 = [57 59 65 70 59 58 57 58 350 61 62 60 62 58 57]
5   c2 = std(a2)
6   Outlier = 3*c2
7   b2 = filloutliers(a2,'nearest','mean')
8
9
```

Result :

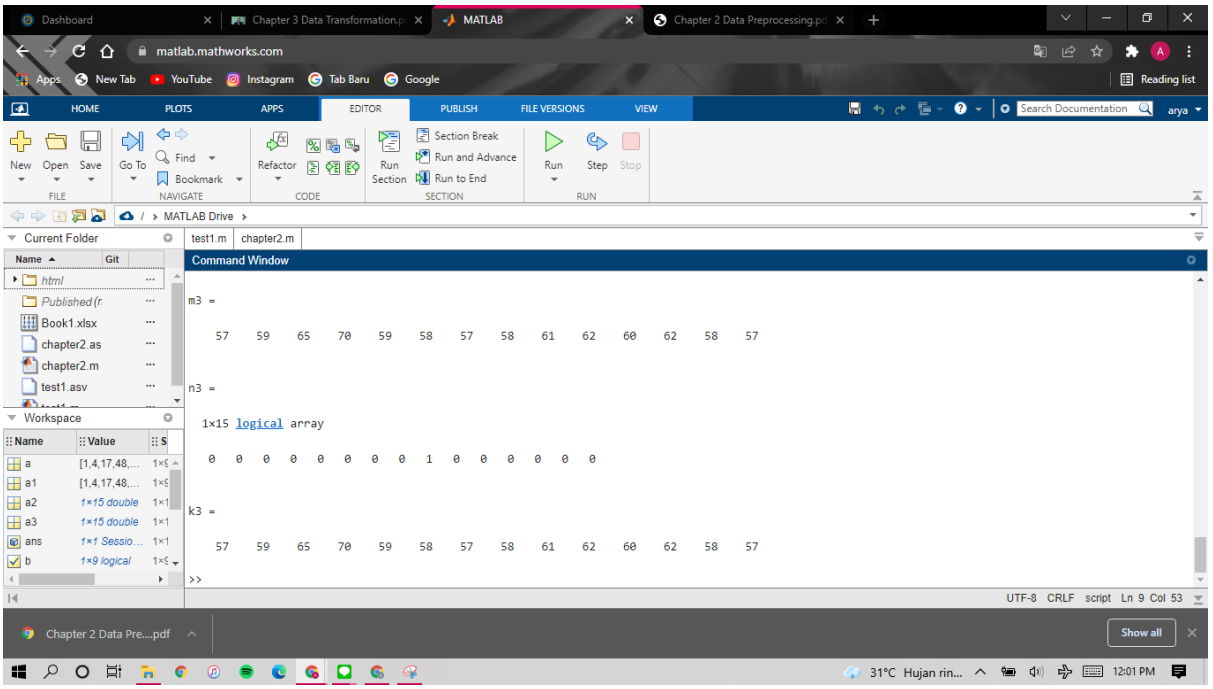
```
c2 =

   74.9055


Outlier =

  224.7166


b2 =

    57    59    65    70    59    58    57    58    61    61    62    60    62    58    57
```

2. rmoutlier

Source code :

```
 8
 9      a3 = [57 59 65 70 59 58 57 58 350 61 62 60 62 58 57]
10      [m3, n3] = rmoutliers(a3,'mean')
11      k3 = rmoutliers(a3,'mean')
```

Result :



Variables M and K contain data whose outliers have been removed. Variable N is used to detect outliers. Value 0 = not outlier, value 1 = outlier.

2. Missing Value

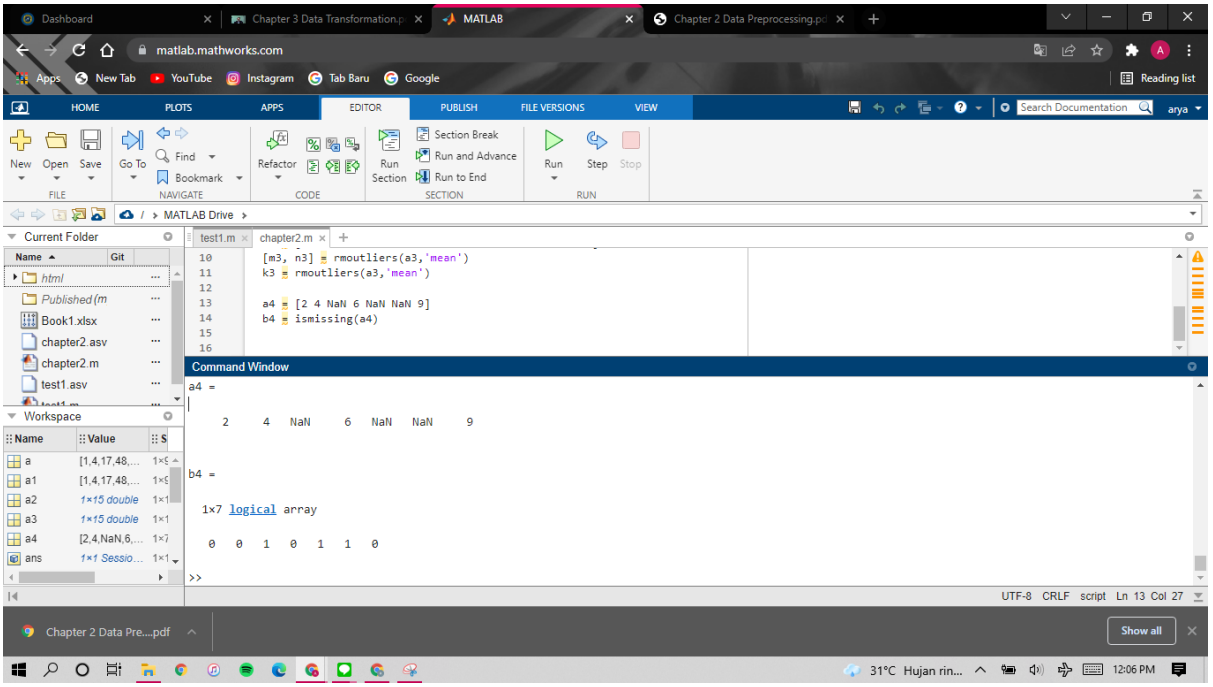
2.1 Missing value detection

ismissing function.

Source code:

```
a4 = [2 4 NaN 6 NaN NaN 9]
b4 = ismissing(a4)
```

Result :



Value 0 = not missing value, value 1 = missin value.

2.2 Handling missing value

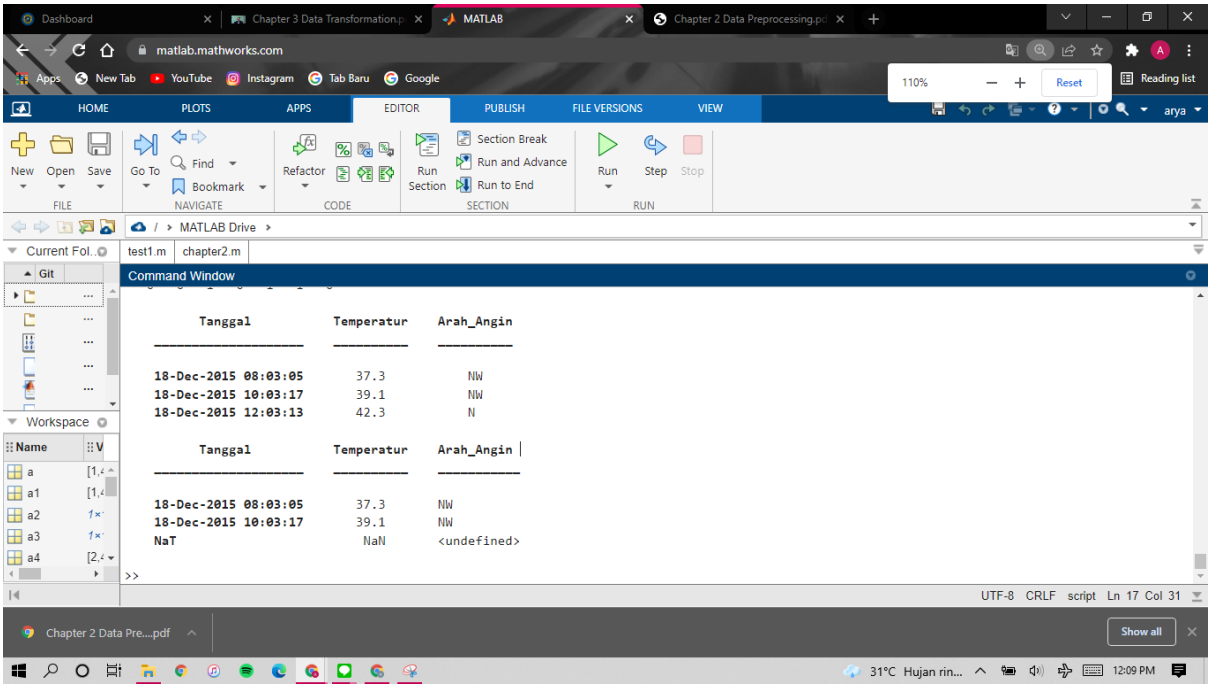
Functions that can be used in handling missing data are as follow :

1. missing

```
Tanggal = datetime({'2015-12-18 08:03:05';'2015-12-18 10:03:17';'2015-12-18 12:03:13'});
Temperatur = [37.3;39.1;42.3];
Arah_Angin = categorical({'NW';'NW';'N'});
TT = timetable(Tanggal,Temperatur,Arah_Angin);
disp(TT)
TT.Tanggal(3) = missing;
TT.Temperatur(3) = missing;
TT.Arah_Angin(3) = missing;
disp(TT)
```

Command Window

result :

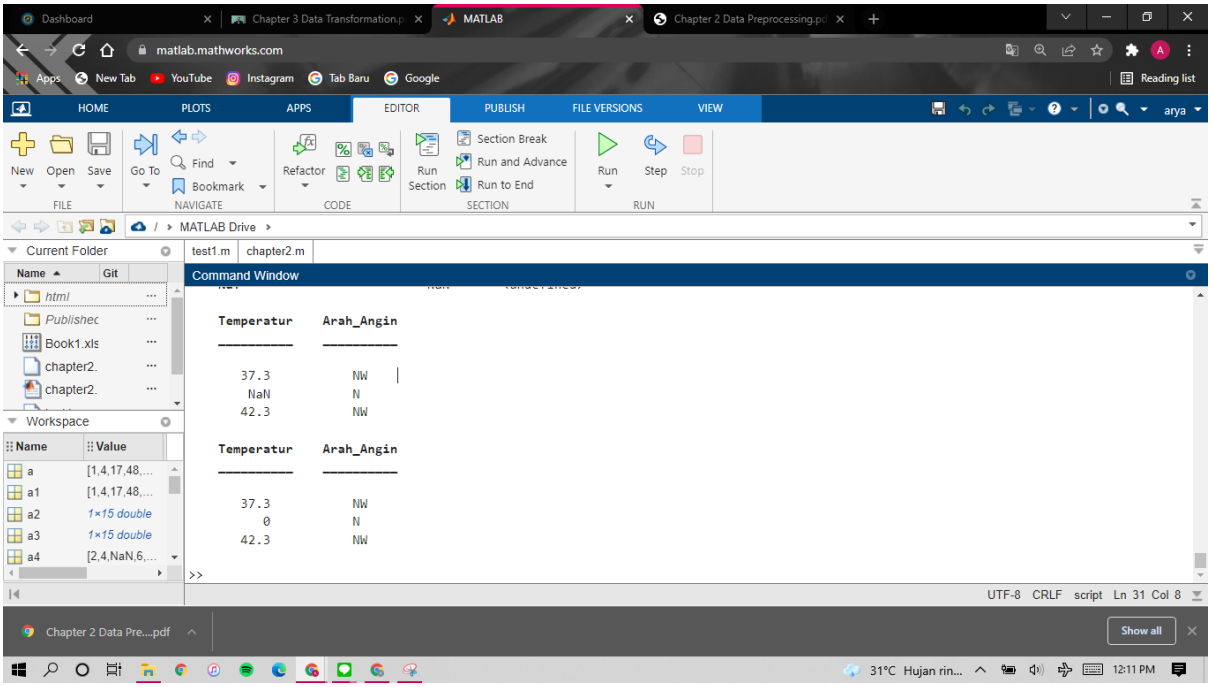


2. fillmissing

Source code :

```
Temperatur = [37.3;NaN;42.3];
Arah_Angin = categorical({'NW';'N';'NW'});
TT = table(Temperatur,Arah_Angin);
disp(TT)
F = fillmissing(TT,'constant',0,'DataVariables',@isnumeric);
disp(F)
```

Result :

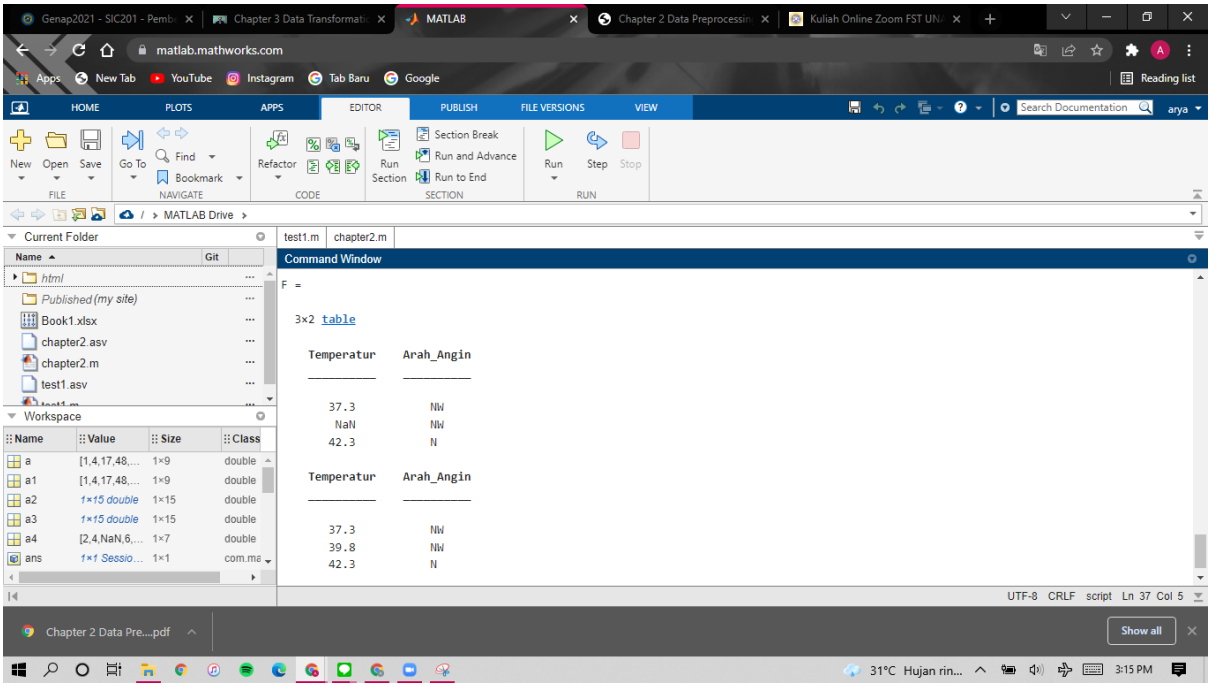


F = fillmissing(TT,'constant',0,'DataVariables',@isnumeric);

Source code :

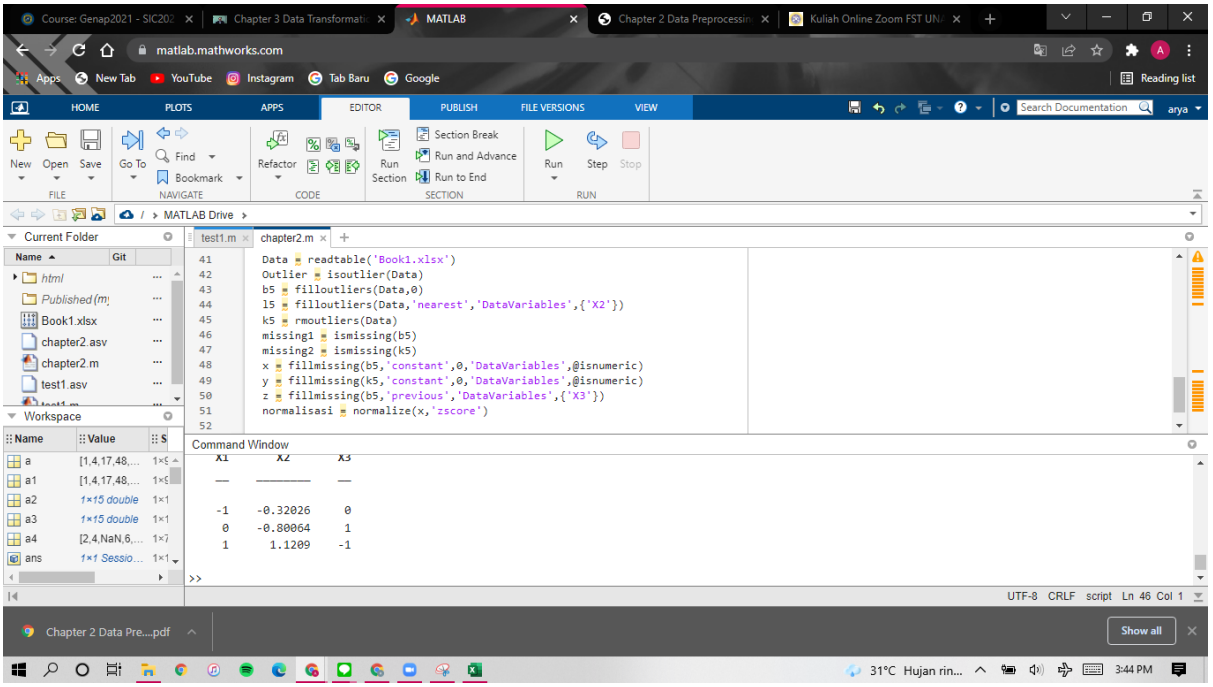
```
33    Temperatur = [37.3;NaN;42.3];
34    Arah_Angin = categorical({'NW';'';'N'});
35    TT = table(Temperatur,Arah_Angin);
36    disp(TT)
37    F = fillmissing(TT,'previous','DataVariables',{'Arah_Angin'})
38    G = fillmissing(F,'pchip','DataVariables',{'Temperatur'});
39    disp(G)
```

Result :



'DataVariable' used to fill missing data on certain variables.

Source Code



Result:

Data

```
Data =

  3×3 table

    X1    X2     X3
    __    ___    ___

    1       2      3
    4     150      6
    7       8    NaN
```

1. Outlier Detection

```
Outlier =

  3×3 logical array

   0   0   0
   0   1   0
   0   0   0
```

2. Handling Outlier

Replace with value “ 0 ”

```
b5 =

  3×3 table

    X1    X2     X3
    __    __     ___

    1     2        3
    4     0        6
    7     8      NaN
```

Replace with previous value

```
15 =

  3×3 table

    X1    X2    X3
    __    __    ___

    1     2       3
    4     8       6
    7     8     NaN
```

Remove Outlier

```
k5 =

  2×3 table

    X1    X2    X3
    __    __    ___

    1     2       3
    7     8     NaN
```

Missing Value detection

```
missing1 =

  3×3 logical array

   0   0   0
   0   0   0
   0   0   1


missing2 =

  2×3 logical array

   0   0   0
   0   0   1
```

Handling Missing Value

Fill with value 0

```
x =

  3×3 table

    X1    X2    X3
    __    __    __

    1     2     3
    4     0     6
    7     8     0


y =

  2×3 table

    X1    X2    X3
    __    __    __

    1     2     3
    7     8     0
```

Fill with closest value

```
z =

  3×3 table

    X1    X2    X3
    __    __    __

    1     2     3
    4     0     6
    7     8     6
```

4. Normalization

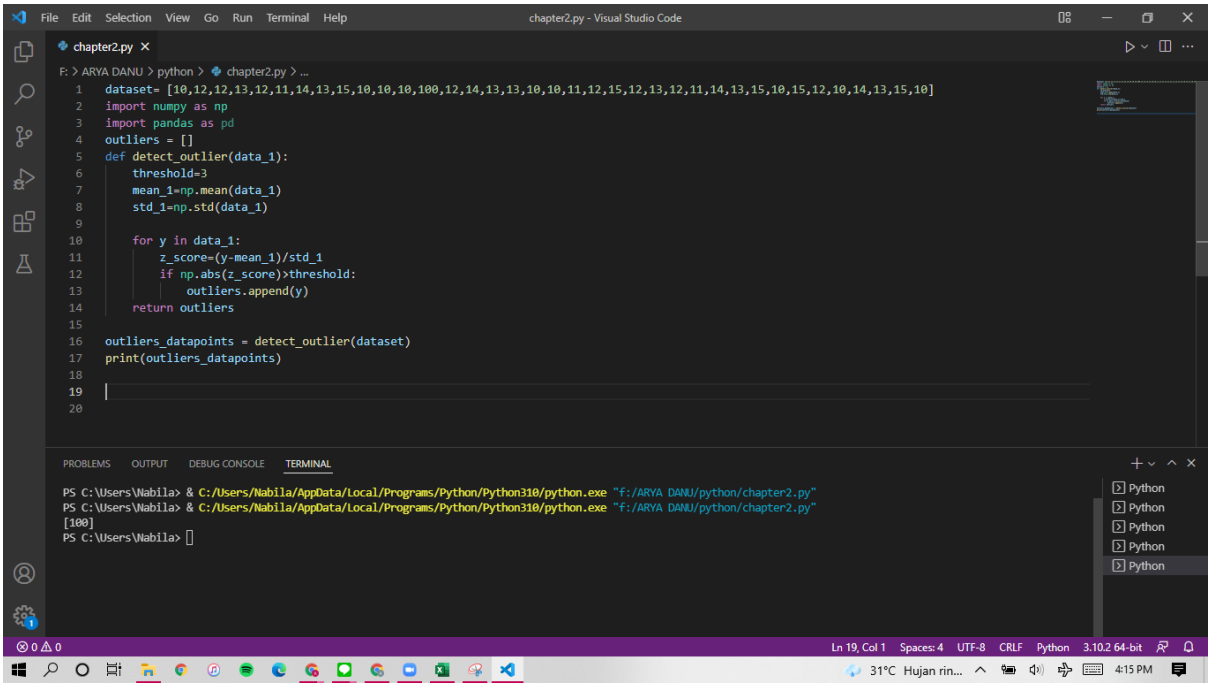
Z-score

```
normalisasi =

  3×3 table

    X1        X2        X3
    __    ________    __

    -1    -0.32026     0
     0    -0.80064     1
     1      1.1209    -1
```

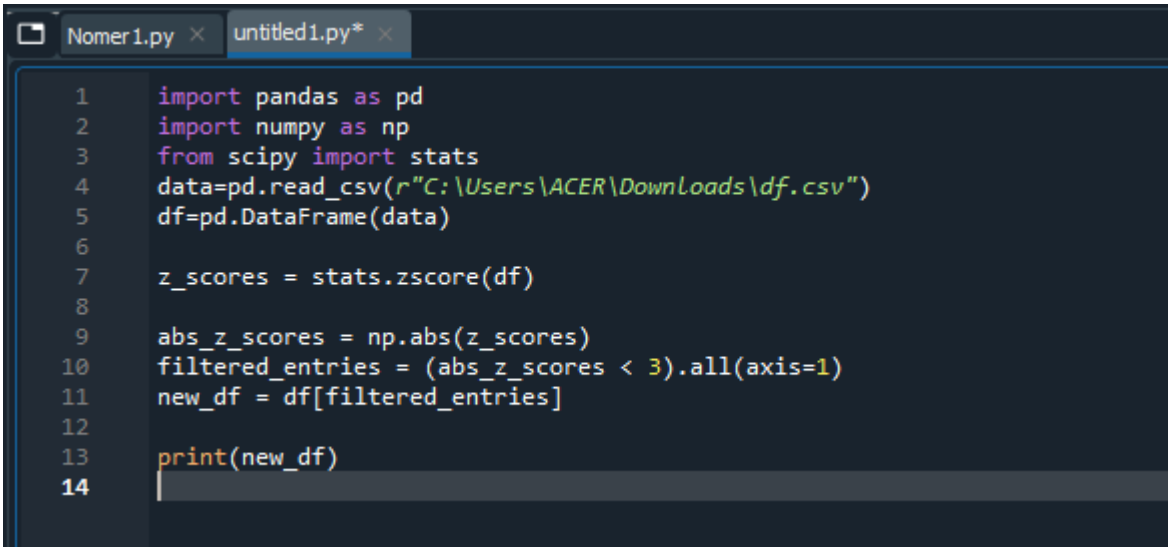
Phyton

1. **Outlier**

Result :



Data (df)

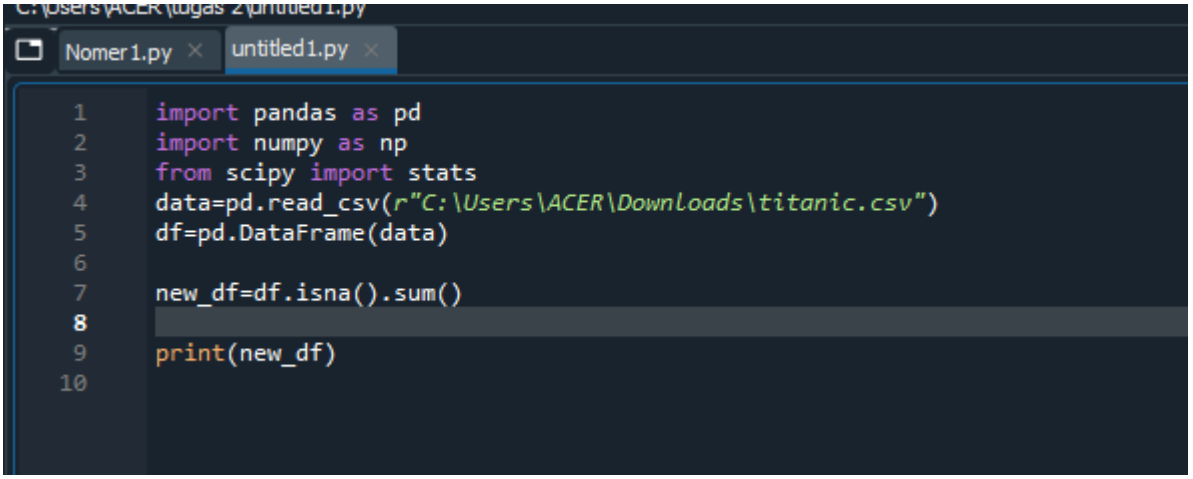


Output :

```
In [2]: runfile('C:/Users/ACER/tugas 2/untitled1.py',
wdir='C:/Users/ACER/tugas 2')
   Column_1  Column_2
0         1         1
1         1         1
2         1         1
3         1         1
5         1         1
6         1         1
7         1         1
8         1         1
9         1         1
```

2. Missing Value

2.1 Missing Value Detection



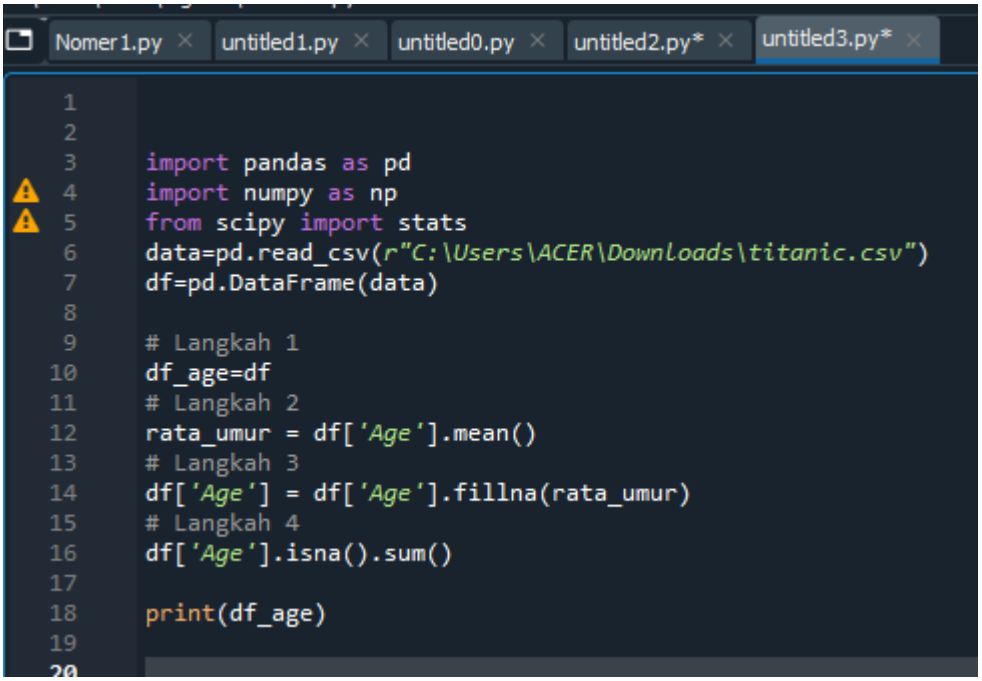
Output

```
In [4]: runfile('C:/Users/ACER/tugas 2/untitled1.py',
wdir='C:/Users/ACER/tugas 2')
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Dengan bantuan fungsi *isna()* dan *sum()* kita tahu bahwa dalam dataset semua kolom tidak ada nilai yang kosong kecuali kolom Age dengan 177 missing value, Kolom Cabin 687 dan kolom Embarked 2

2.2 Handling Missing Value

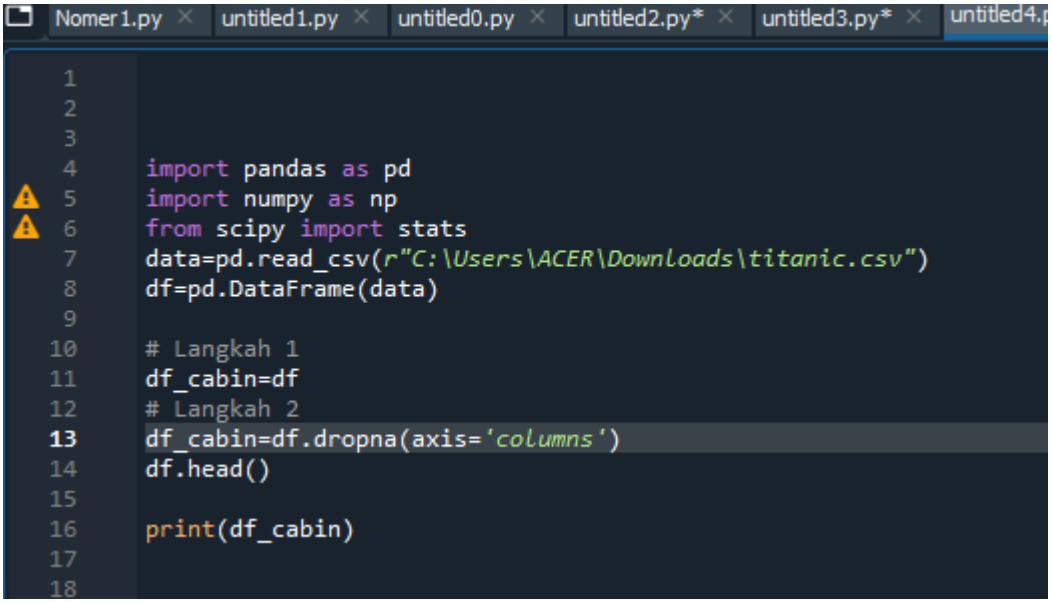
Replace missing value with mean



Output

```
In [2]: runfile('C:/Users/ACER/tugas 2/untitled3.py', wdir='C:/Users/ACER/tugas 2')
     PassengerId  Survived  Pclass  ...     Fare Cabin  Embarked
0              1         0       3  ...   7.2500   NaN         S
1              2         1       1  ...  71.2833   C85         C
2              3         1       3  ...   7.9250   NaN         S
3              4         1       1  ...  53.1000  C123         S
4              5         0       3  ...   8.0500   NaN         S
..           ...       ...     ...  ...      ...   ...       ...
886          887         0       2  ...  13.0000   NaN         S
887          888         1       1  ...  30.0000   B42         S
888          889         0       3  ...  23.4500   NaN         S
889          890         1       1  ...  30.0000  C148         C
890          891         0       3  ...   7.7500   NaN         Q

[891 rows x 12 columns]

In [3]:
```

Remove missing value



Output

```
In [6]: runfile('C:/Users/ACER/tugas 2/untitled4.py', wdir='C:/Users/ACER/
tugas 2')
     PassengerId  Survived  Pclass  ... Parch            Ticket     Fare
0              1         0       3  ...     0         A/5 21171   7.2500
1              2         1       1  ...     0          PC 17599  71.2833
2              3         1       3  ...     0  STON/O2. 3101282   7.9250
3              4         1       1  ...     0            113803  53.1000
4              5         0       3  ...     0            373450   8.0500
..           ...       ...     ...  ...   ...               ...      ...
886          887         0       2  ...     0            211536  13.0000
887          888         1       1  ...     0            112053  30.0000
888          889         0       3  ...     2        W./C. 6607  23.4500
889          890         1       1  ...     0            111369  30.0000
890          891         0       3  ...     0            370376   7.7500

[891 rows x 9 columns]
```

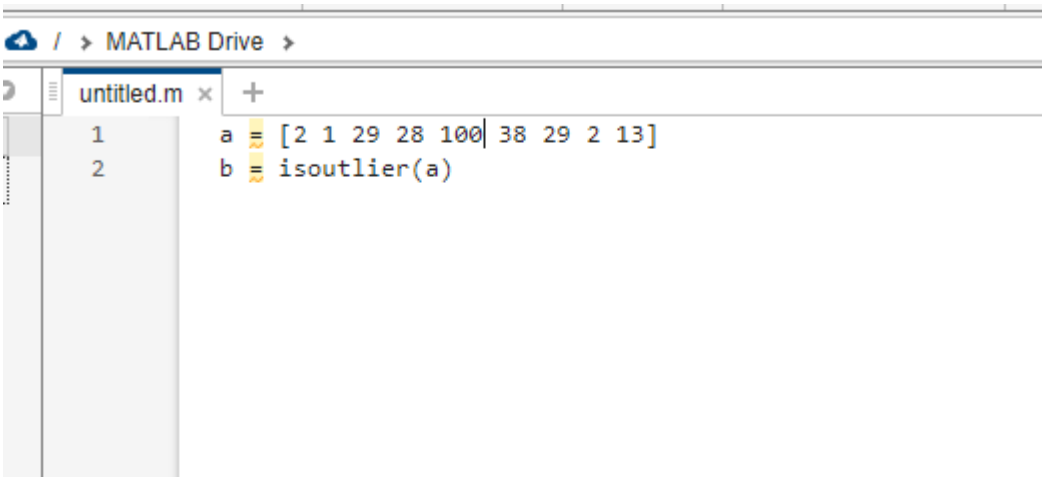
ASSIGNMENT:

1. Try all the source code that has been written above
2. Look for any data that can be downloaded
3. Do preprocessing on data you have found using Matlab and Python
4. Make a report containing a print screen of the results of the code that has been written and provide an explanation
5. Name the file with " laporan preprocessing_KelompokXXX.pdf

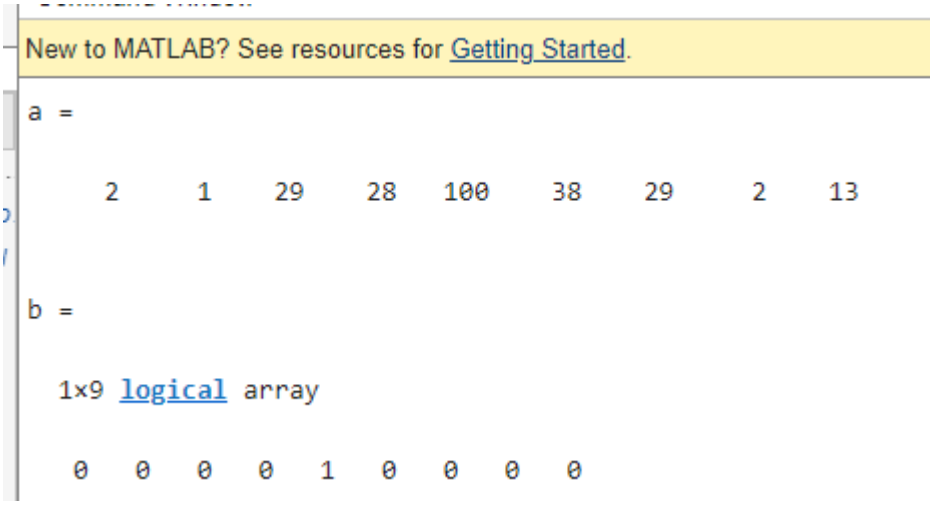
MATLAB:

1. **Outlier**

1.2 Outlier Detection



Result :



1.2 Oulier Handling

1. fill outlier

```
a2 = [94 73 47 34 73 840 63 44 75 62 72 81 73]
c2 = std(a2)
Outlier = 3*c2
b2 = filloutliers(a2,'nearest','mean')
```

Result :

```
New to MATLAB? See resources for Getting Started.

c2 =

  215.3058


Outlier =

  645.9175


b2 =

  Columns 1 through 10

    94    73    47    34    73    63    63    44    75    62
```

2. **rmoutlier**

```
8
9       a3 = [23 47 29 384 62 36 42 93 84 76 23 84]
10      [m3,n3] = rmoutliers(a3,'mean')
11      k3 = rmoutliers(a3,'mean')
12
```

Result :

```
m3 =

    23    47    29    62    36    42    93    84    76    23    84


n3 =

  1×12 logical array

   0   0   0   1   0   0   0   0   0   0   0   0


k3 =

    23    47    29    62    36    42    93    84    76    23    84
```

2. Missing Value

2.1 Missing value detection

```
a4 = [1 2 NaN 3 NaN NaN 8]
b4 = ismissing(a4)
```

Result:

```
a4 =

     1     2   NaN     3   NaN   NaN     8


b4 =

  1×7 logical array

   0   0   1   0   1   1   0
```

2.2 Handling missing value

1. **missing**

```
Tanggal = datetime({'2022-03-16 08:03:05';'2022-03-16 10:03:17';'2022-03-16 12:03:13'});
Temperatur = [35.3;37.1;45.3];
Arah_Angin = categorical({'NW';'NW';'N'});
TT = timetable(Tanggal,Temperatur,Arah_Angin);
disp(TT)
TT.Tanggal(3) = missing;
TT.Temperatur(3) = missing;
TT.Arah_Angin(3) = missing;
disp(TT)
```

Result:

```
      Tanggal          Temperatur    Arah_Angin
____________________    __________    __________

16-Mar-2022 08:03:05       35.3          NW
16-Mar-2022 10:03:17       37.1          NW
16-Mar-2022 12:03:13       45.3          N

      Tanggal          Temperatur    Arah_Angin
____________________    __________    __________

16-Mar-2022 08:03:05       35.3       NW
16-Mar-2022 10:03:17       37.1       NW
NaT                         NaN       <undefined>
```

2. **fillmissing**

```
untitled.m ×  +
25
26  Temperatur = [35.3;NaN;41.3];
27  Arah_Angin = categorical({'NW';'N';'NW'});
28  TT = table(Temperatur,Arah_Angin);
29  disp(TT)
30  F = fillmissing(TT,'constant',0,'DataVariables',@isnumeric);
31  disp(F)
32
33  Temperatur = [35.3;NaN;41.3];
```

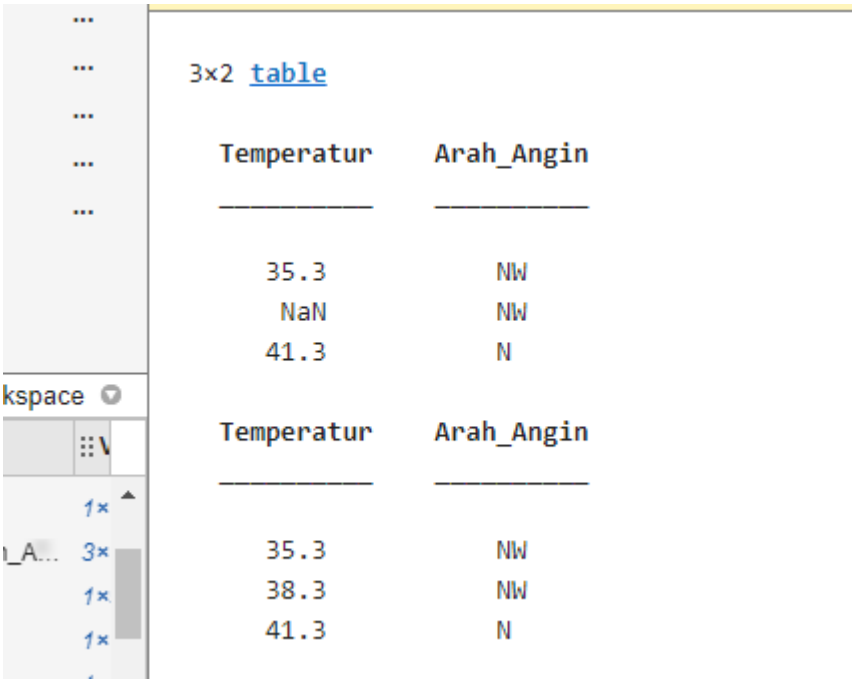
Result:

```
Temperatur    Arah_Angin
__________    __________

   35.3           NW
    NaN           N
   41.3           NW

Temperatur    Arah_Angin
__________    __________

   35.3           NW
      0           N
   41.3           NW
```

source code :

```
32
33  Temperatur = [35.3;NaN;41.3];
34  Arah_Angin = categorical({'NW';'';'N'});
35  TT = table(Temperatur,Arah_Angin);
36  disp(TT)
37  F = fillmissing(TT,'previous','DataVariables',{'Arah_Angin'})
38  G = fillmissing(F,'pchip','DataVariables',{'Temperatur'});
39  disp(G)
40
```

Result :



3×2 table

Temperatur	Arah_Angin
35.3	NW
NaN	NW
41.3	N

Temperatur	Arah_Angin
35.3	NW
38.3	NW
41.3	N

Source code :

```
40
41  Data = readtable('BenderlyZwick.csv')
42  Outlier = isoutlier(Data)
43  b5 = filloutliers(Data,0)
44  l5 = filloutliers(Data,'nearest','DataVariables',{'growth'})
45  k5 = rmoutliers(Data)
46  missing1 = ismissing(b5)
47  missing2 = ismissing(k5)
48  x = fillmissing(b5,'constant',0,'DataVariables',@isnumeric)
49  y = fillmissing(k5,'constant',0,'DataVariables',@isnumeric)
50  z = fillmissing(b5,'previous','DataVariables',{'inflation'})
51  normalisasi = normalize(x,'zscore')
```

Result data :

```
Data =

  31x6 table

    Var1     returns    growth    inflation    growth2    inflation:
    ____     _______    ______    _________    _______    _________

      1          NaN       NaN          NaN        3.9          2.2
      2          NaN       NaN          NaN          4          2.1
      3           53       6.7         -0.4       -1.3          0.6
      4         31.2       2.1          0.4        5.6          1.3
      5          3.7       1.8          2.9        2.1          1.9

     :            :         :            :          :            :

     27         -1.3       2.8          7.9        5.3          7.3
     28          8.6      -0.3          9.8        2.5          9.2
     29         22.2       2.6         10.2       -0.2         10.7
     30        -12.2      -1.9          7.3        1.9          9.2
```

1. **outlier detection**

```
Outlier =

  31x6 logical array

   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
```

2. **Handling outlier**

b5 =

31×6 table

Var1	returns	growth	inflation	growth2	inflation2
1	NaN	NaN	NaN	3.9	2.2
2	NaN	NaN	NaN	4	2.1
3	53	6.7	-0.4	-1.3	0.6
4	31.2	2.1	0.4	5.6	1.3
5	3.7	1.8	2.9	2.1	1.9
:	:	:	:	:	:
27	-1.3	2.8	7.9	5.3	7.3
28	8.6	-0.3	9.8	2.5	9.2
29	22.2	2.6	10.2	-0.2	0
30	-12.2	-1.9	7.3	1.9	9.2
31	NaN	NaN	NaN	-2.5	5.7

Replace with previous value

l5 =

31×6 table

Var1	returns	growth	inflation	growth2	inflation2
1	NaN	NaN	NaN	3.9	2.2
2	NaN	NaN	NaN	4	2.1
3	53	6.7	-0.4	-1.3	0.6
4	31.2	2.1	0.4	5.6	1.3
5	3.7	1.8	2.9	2.1	1.9
:	:	:	:	:	:
27	-1.3	2.8	7.9	5.3	7.3
28	8.6	-0.3	9.8	2.5	9.2
29	22.2	2.6	10.2	-0.2	10.7
30	-12.2	-1.9	7.3	1.9	9.2
31	NaN	NaN	NaN	-2.5	5.7

Remove Outlier

k5 =

29×6 table

Var1	returns	growth	inflation	growth2	inflation2
1	NaN	NaN	NaN	3.9	2.2
2	NaN	NaN	NaN	4	2.1
3	53	6.7	-0.4	-1.3	0.6
4	31.2	2.1	0.4	5.6	1.3
5	3.7	1.8	2.9	2.1	1.9
:	:	:	:	:	:
26	-13.1	5	5.9	4.7	6.5
27	-1.3	2.8	7.9	5.3	7.3
28	8.6	-0.3	9.8	2.5	9.2
30	-12.2	-1.9	7.3	1.9	9.2
31	NaN	NaN	NaN	-2.5	5.7

Missing value detection

missing1 =

31×6 logical array

0	1	1	1	0	0
0	1	1	1	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

```
missing2 =

  29×6 logical array

   0   1   1   1   0   0
   0   1   1   1   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
   0   0   0   0   0   0
```

Handling missing value

fit with value 0

```
X =

  31×6 table

    Var1    returns    growth    inflation    growth2    inflation2
    ____    _______    ______    _________    _______    __________

      1           0         0          0         3.9         2.2
      2           0         0          0           4         2.1
      3          53       6.7       -0.4        -1.3         0.6
      4        31.2       2.1        0.4         5.6         1.3
      5         3.7       1.8        2.9         2.1         1.9

      :         :         :          :           :           :

     27        -1.3       2.8        7.9         5.3         7.3
     28         8.6      -0.3        9.8         2.5         9.2
     29        22.2       2.6       10.2        -0.2           0
```

y =

29×6 table

Var1	returns	growth	inflation	growth2	inflation2
1	0	0	0	3.9	2.2
2	0	0	0	4	2.1
3	53	6.7	-0.4	-1.3	0.6
4	31.2	2.1	0.4	5.6	1.3
5	3.7	1.8	2.9	2.1	1.9
:	:	:	:	:	:
26	-13.1	5	5.9	4.7	6.5
27	-1.3	2.8	7.9	5.3	7.3
28	8.6	-0.3	9.8	2.5	9.2

Fill with closent value

z =

31×6 table

Var1	returns	growth	inflation	growth2	inflation2
1	NaN	NaN	NaN	3.9	2.2
2	NaN	NaN	NaN	4	2.1
3	53	6.7	-0.4	-1.3	0.6
4	31.2	2.1	0.4	5.6	1.3
5	3.7	1.8	2.9	2.1	1.9
:	:	:	:	:	:
27	-1.3	2.8	7.9	5.3	7.3
28	8.6	-0.3	9.8	2.5	9.2
29	22.2	2.6	10.2	-0.2	0

Normalization

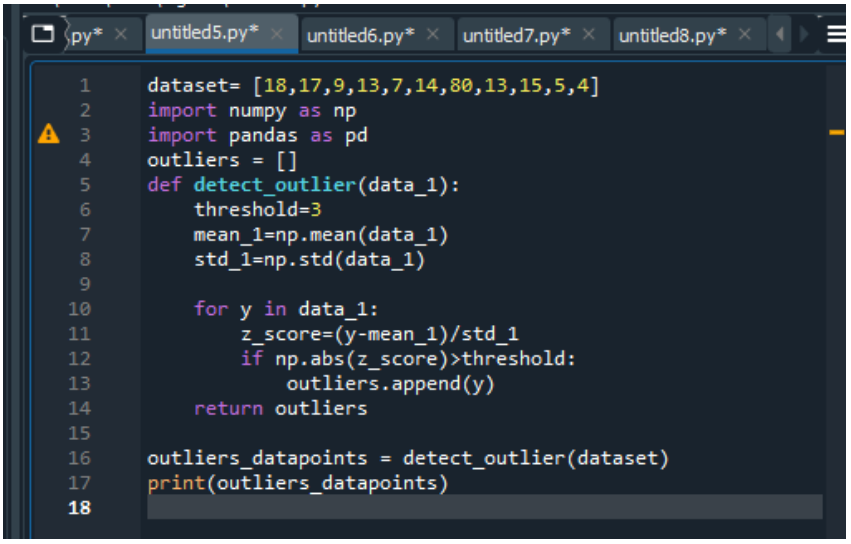
normalisasi =

31×6 table

Var1	returns	growth	inflation	growth2	inflation2
-1.6498	-0.34644	-1.1085	-1.1803	0.41391	-0.52834
-1.5398	-0.34644	-1.1085	-1.1803	0.45413	-0.56697
-1.4298	2.3783	1.4468	-1.3065	-1.6777	-1.1464
-1.3198	1.2576	-0.30756	-1.054	1.0977	-0.876
-1.2098	-0.15622	-0.42198	-0.26477	-0.31011	-0.64423
:	:	:	:	:	:
1.2098	-0.41327	-0.040598	1.3137	0.97704	1.4417
1.3198	0.095689	-1.2229	1.9135	-0.14922	2.1757
1.4298	0.79487	-0.11687	2.0397	-1.2352	-1.3782
1.5398	-0.97364	-1.8331	1.1242	-0.39056	2.1757
1.6498	-0.34644	-1.1085	-1.1803	-2.1604	0.82367

PYTHON:

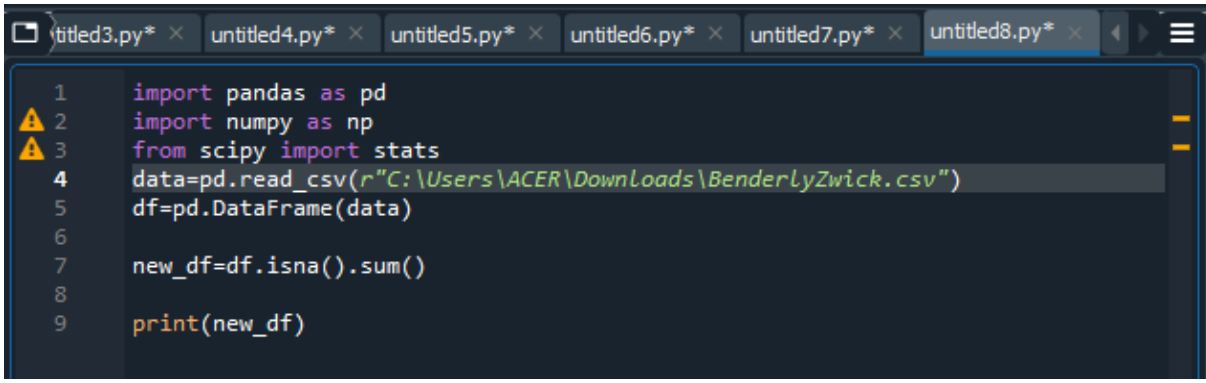
1. Outlier



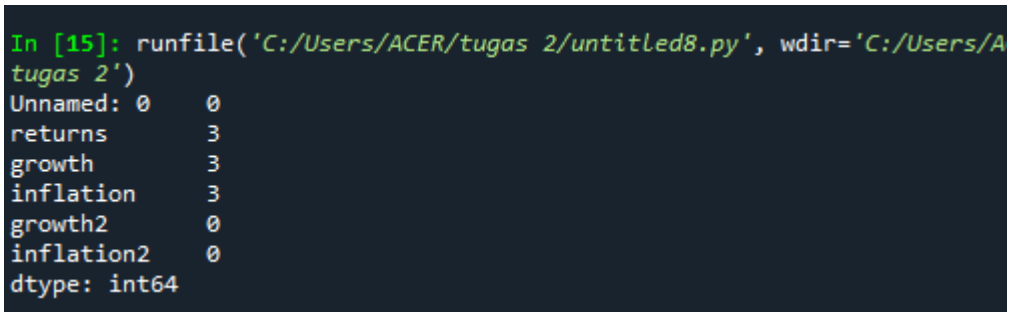
Output

```
In [30]: runfile('C:/Users/ACER/tugas 2/untitled5.py', wdir='C:/Users/ACER/tugas 2')
[80]
```

2. Missing Value
    a. Missing Value Detection



Output



    b. Handling Missing Value

Replace missing value with mean

untitled0.py | untitled2.py* | untitled3.py* | untitled4.py* | untitled5.py* | untitled6.py*

```python
import pandas as pd
import numpy as np
from scipy import stats
data=pd.read_csv(r"C:\Users\ACER\Downloads\BenderlyZwick.csv")
df=pd.DataFrame(data)

# Langkah 1
df_growth=df
# Langkah 2
rata_umur = df['growth'].mean()
# Langkah 3
df['growth'] = df['growth'].fillna(rata_umur)
# Langkah 4
df['growth'].isna().sum()

print(df_growth)
```

output

```
In [10]: runfile('C:/Users/ACER/tugas 2/untitled6.py', wdir='C:/Users/ACER/
tugas 2')
    Unnamed: 0  returns    growth  inflation  growth2  inflation2
0            1      NaN  3.217857        NaN      3.9         2.2
1            2      NaN  3.217857        NaN      4.0         2.1
2            3     53.0  6.700000       -0.4     -1.3         0.6
3            4     31.2  2.100000        0.4      5.6         1.3
4            5      3.7  1.800000        2.9      2.1         1.9
5            6    -13.8 -0.400000        3.0      1.7         3.2
6            7     41.7  6.000000        1.7     -0.8         1.8
7            8     10.5  2.100000        1.5      5.8         2.2
8            9     -1.3  2.600000        1.8      2.2         1.9
9           10     26.1  5.800000        0.8      2.6         1.2
10          11    -10.5  4.000000        1.8      5.3         1.8
11          12     21.2  5.300000        1.6      4.1         1.5
12          13     15.5  6.000000        1.0      5.3         1.7
13          14     10.2  6.000000        2.3      5.8         1.7
14          15    -13.3  2.700000        3.2      5.8         3.1
15          16     21.3  4.600000        2.7      2.9         2.5
16          17      6.8  2.800000        4.3      4.1         4.5
17          18    -13.5 -0.200000        5.0      2.4         4.3
18          19     -0.4  3.400000        4.4     -0.3         4.6
```

Remove missing value

titled2.py* | untitled3.py* | untitled4.py* | untitled5.py* | untitled6.py* | untitled7.py*

```
1   import pandas as pd
2   import numpy as np
3   from scipy import stats
4   data=pd.read_csv(r"C:\Users\ACER\Downloads\BenderlyZwick.csv")
5   df=pd.DataFrame(data)
6
7   # Langkah 1
8   df_returns=df
9   # Langkah 2
10  df_returns=df.dropna(axis='columns')
11  df.head()
12
13  print(df_returns)
14
```

Output

```
In [14]: runfile('C:/Users/ACER/tugas 2/untitled7.py', wdir='C:/Users/ACER/
tugas 2')
    Unnamed: 0  growth2  inflation2
0            1      3.9         2.2
1            2      4.0         2.1
2            3     -1.3         0.6
3            4      5.6         1.3
4            5      2.1         1.9
5            6      1.7         3.2
6            7     -0.8         1.8
7            8      5.8         2.2
8            9      2.2         1.9
9           10      2.6         1.2
10          11      5.3         1.8
11          12      4.1         1.5
12          13      5.3         1.7
13          14      5.8         1.7
14          15      5.8         3.1
15          16      2.9         2.5
16          17      4.1         4.5
17          18      2.4         4.3
18          19     -0.3         4.6
19          20      2.8         4.7
20          21      5.0         4.0
21          22      5.2         6.2
22          23     -0.5        10.5
23          24     -1.3         8.0
24          25      4.9         5.7
25          26      4.7         6.5
26          27      5.3         7.3
27          28      2.5         9.2
28          29     -0.2        10.7
29          30      1.9         9.2
30          31     -2.5         5.7
```