data_transformation

December 29, 2018

In [1]:
```
import pandas as pd
iris_file = 'iris.csv'
iris = pd.read_csv(iris_file, sep=',', decimal='.', header=None, names=['sepal_length'
                                                                        'petal_length',
                                                                        'target'])
```

In [2]:
```
iris.head(20)
```

Out[2]:

	sepal_length	sepal_width	petal_length	petal_width	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
11	4.8	3.4	1.6	0.2	Iris-setosa
12	4.8	3.0	1.4	0.1	Iris-setosa
13	4.3	3.0	1.1	0.1	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa
16	5.4	3.9	1.3	0.4	Iris-setosa
17	5.1	3.5	1.4	0.3	Iris-setosa
18	5.7	3.8	1.7	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa

In [3]:
```
iris.columns
```

Out[3]:
```
Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'target'], dtype=
```

In [4]:
```
target = iris['target']
print(target)
```

```
0         Iris-setosa
1         Iris-setosa
```

```
2         Iris-setosa
3         Iris-setosa
4         Iris-setosa
5         Iris-setosa
6         Iris-setosa
7         Iris-setosa
8         Iris-setosa
9         Iris-setosa
10        Iris-setosa
11        Iris-setosa
12        Iris-setosa
13        Iris-setosa
14        Iris-setosa
15        Iris-setosa
16        Iris-setosa
17        Iris-setosa
18        Iris-setosa
19        Iris-setosa
20        Iris-setosa
21        Iris-setosa
22        Iris-setosa
23        Iris-setosa
24        Iris-setosa
25        Iris-setosa
26        Iris-setosa
27        Iris-setosa
28        Iris-setosa
29        Iris-setosa
            ...
120    Iris-virginica
121    Iris-virginica
122    Iris-virginica
123    Iris-virginica
124    Iris-virginica
125    Iris-virginica
126    Iris-virginica
127    Iris-virginica
128    Iris-virginica
129    Iris-virginica
130    Iris-virginica
131    Iris-virginica
132    Iris-virginica
133    Iris-virginica
134    Iris-virginica
135    Iris-virginica
136    Iris-virginica
137    Iris-virginica
138    Iris-virginica
```

```
139    Iris-virginica
140    Iris-virginica
141    Iris-virginica
142    Iris-virginica
143    Iris-virginica
144    Iris-virginica
145    Iris-virginica
146    Iris-virginica
147    Iris-virginica
148    Iris-virginica
149    Iris-virginica
Name: target, Length: 150, dtype: object
```

```
In [5]: x = iris[['sepal_length', 'sepal_width']]
        print(x)
```

```
     sepal_length  sepal_width
0             5.1          3.5
1             4.9          3.0
2             4.7          3.2
3             4.6          3.1
4             5.0          3.6
5             5.4          3.9
6             4.6          3.4
7             5.0          3.4
8             4.4          2.9
9             4.9          3.1
10            5.4          3.7
11            4.8          3.4
12            4.8          3.0
13            4.3          3.0
14            5.8          4.0
15            5.7          4.4
16            5.4          3.9
17            5.1          3.5
18            5.7          3.8
19            5.1          3.8
20            5.4          3.4
21            5.1          3.7
22            4.6          3.6
23            5.1          3.3
24            4.8          3.4
25            5.0          3.0
26            5.0          3.4
27            5.2          3.5
28            5.2          3.4
29            4.7          3.2
```

```
..            ...          ...
120           6.9          3.2
121           5.6          2.8
122           7.7          2.8
123           6.3          2.7
124           6.7          3.3
125           7.2          3.2
126           6.2          2.8
127           6.1          3.0
128           6.4          2.8
129           7.2          3.0
130           7.4          2.8
131           7.9          3.8
132           6.4          2.8
133           6.3          2.8
134           6.1          2.6
135           7.7          3.0
136           6.3          3.4
137           6.4          3.1
138           6.0          3.0
139           6.9          3.1
140           6.7          3.1
141           6.9          3.1
142           5.8          2.7
143           6.8          3.2
144           6.7          3.3
145           6.7          3.0
146           6.3          2.5
147           6.5          3.0
148           6.2          3.4
149           5.9          3.0

[150 rows x 2 columns]
```

```
In [6]: #### lazy read
        iris_chunk = pd.read_csv(iris_file, header=None, names=['sepal_length', 'sepal_width',
                                                                'petal_length','petal_width',
                                                                'target'], chunksize=10)

        for chunk in iris_chunk:
            print("Dimensions", chunk.shape)
            print(chunk, "\n")
```

```
Dimensions (10, 5)
   sepal_length  sepal_width  petal_length  petal_width       target
0           5.1          3.5           1.4          0.2  Iris-setosa
1           4.9          3.0           1.4          0.2  Iris-setosa
```

2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa

Dimensions (10, 5)

	sepal_length	sepal_width	petal_length	petal_width	target
10	5.4	3.7	1.5	0.2	Iris-setosa
11	4.8	3.4	1.6	0.2	Iris-setosa
12	4.8	3.0	1.4	0.1	Iris-setosa
13	4.3	3.0	1.1	0.1	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa
16	5.4	3.9	1.3	0.4	Iris-setosa
17	5.1	3.5	1.4	0.3	Iris-setosa
18	5.7	3.8	1.7	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa

Dimensions (10, 5)

	sepal_length	sepal_width	petal_length	petal_width	target
20	5.4	3.4	1.7	0.2	Iris-setosa
21	5.1	3.7	1.5	0.4	Iris-setosa
22	4.6	3.6	1.0	0.2	Iris-setosa
23	5.1	3.3	1.7	0.5	Iris-setosa
24	4.8	3.4	1.9	0.2	Iris-setosa
25	5.0	3.0	1.6	0.2	Iris-setosa
26	5.0	3.4	1.6	0.4	Iris-setosa
27	5.2	3.5	1.5	0.2	Iris-setosa
28	5.2	3.4	1.4	0.2	Iris-setosa
29	4.7	3.2	1.6	0.2	Iris-setosa

Dimensions (10, 5)

	sepal_length	sepal_width	petal_length	petal_width	target
30	4.8	3.1	1.6	0.2	Iris-setosa
31	5.4	3.4	1.5	0.4	Iris-setosa
32	5.2	4.1	1.5	0.1	Iris-setosa
33	5.5	4.2	1.4	0.2	Iris-setosa
34	4.9	3.1	1.5	0.1	Iris-setosa
35	5.0	3.2	1.2	0.2	Iris-setosa
36	5.5	3.5	1.3	0.2	Iris-setosa
37	4.9	3.1	1.5	0.1	Iris-setosa
38	4.4	3.0	1.3	0.2	Iris-setosa
39	5.1	3.4	1.5	0.2	Iris-setosa

Dimensions (10, 5)

	sepal_length	sepal_width	petal_length	petal_width	target
40	5.0	3.5	1.3	0.3	Iris-setosa
41	4.5	2.3	1.3	0.3	Iris-setosa
42	4.4	3.2	1.3	0.2	Iris-setosa
43	5.0	3.5	1.6	0.6	Iris-setosa
44	5.1	3.8	1.9	0.4	Iris-setosa
45	4.8	3.0	1.4	0.3	Iris-setosa
46	5.1	3.8	1.6	0.2	Iris-setosa
47	4.6	3.2	1.4	0.2	Iris-setosa
48	5.3	3.7	1.5	0.2	Iris-setosa
49	5.0	3.3	1.4	0.2	Iris-setosa

Dimensions (10, 5)

	sepal_length	sepal_width	petal_length	petal_width	target
50	7.0	3.2	4.7	1.4	Iris-versicolor
51	6.4	3.2	4.5	1.5	Iris-versicolor
52	6.9	3.1	4.9	1.5	Iris-versicolor
53	5.5	2.3	4.0	1.3	Iris-versicolor
54	6.5	2.8	4.6	1.5	Iris-versicolor
55	5.7	2.8	4.5	1.3	Iris-versicolor
56	6.3	3.3	4.7	1.6	Iris-versicolor
57	4.9	2.4	3.3	1.0	Iris-versicolor
58	6.6	2.9	4.6	1.3	Iris-versicolor
59	5.2	2.7	3.9	1.4	Iris-versicolor

Dimensions (10, 5)

	sepal_length	sepal_width	petal_length	petal_width	target
60	5.0	2.0	3.5	1.0	Iris-versicolor
61	5.9	3.0	4.2	1.5	Iris-versicolor
62	6.0	2.2	4.0	1.0	Iris-versicolor
63	6.1	2.9	4.7	1.4	Iris-versicolor
64	5.6	2.9	3.6	1.3	Iris-versicolor
65	6.7	3.1	4.4	1.4	Iris-versicolor
66	5.6	3.0	4.5	1.5	Iris-versicolor
67	5.8	2.7	4.1	1.0	Iris-versicolor
68	6.2	2.2	4.5	1.5	Iris-versicolor
69	5.6	2.5	3.9	1.1	Iris-versicolor

Dimensions (10, 5)

	sepal_length	sepal_width	petal_length	petal_width	target
70	5.9	3.2	4.8	1.8	Iris-versicolor
71	6.1	2.8	4.0	1.3	Iris-versicolor
72	6.3	2.5	4.9	1.5	Iris-versicolor
73	6.1	2.8	4.7	1.2	Iris-versicolor
74	6.4	2.9	4.3	1.3	Iris-versicolor
75	6.6	3.0	4.4	1.4	Iris-versicolor
76	6.8	2.8	4.8	1.4	Iris-versicolor

```
77           6.7          3.0           5.0          1.7  Iris-versicolor
78           6.0          2.9           4.5          1.5  Iris-versicolor
79           5.7          2.6           3.5          1.0  Iris-versicolor

Dimensions (10, 5)
    sepal_length  sepal_width  petal_length  petal_width           target
80           5.5          2.4           3.8          1.1  Iris-versicolor
81           5.5          2.4           3.7          1.0  Iris-versicolor
82           5.8          2.7           3.9          1.2  Iris-versicolor
83           6.0          2.7           5.1          1.6  Iris-versicolor
84           5.4          3.0           4.5          1.5  Iris-versicolor
85           6.0          3.4           4.5          1.6  Iris-versicolor
86           6.7          3.1           4.7          1.5  Iris-versicolor
87           6.3          2.3           4.4          1.3  Iris-versicolor
88           5.6          3.0           4.1          1.3  Iris-versicolor
89           5.5          2.5           4.0          1.3  Iris-versicolor

Dimensions (10, 5)
    sepal_length  sepal_width  petal_length  petal_width           target
90           5.5          2.6           4.4          1.2  Iris-versicolor
91           6.1          3.0           4.6          1.4  Iris-versicolor
92           5.8          2.6           4.0          1.2  Iris-versicolor
93           5.0          2.3           3.3          1.0  Iris-versicolor
94           5.6          2.7           4.2          1.3  Iris-versicolor
95           5.7          3.0           4.2          1.2  Iris-versicolor
96           5.7          2.9           4.2          1.3  Iris-versicolor
97           6.2          2.9           4.3          1.3  Iris-versicolor
98           5.1          2.5           3.0          1.1  Iris-versicolor
99           5.7          2.8           4.1          1.3  Iris-versicolor

Dimensions (10, 5)
     sepal_length  sepal_width  petal_length  petal_width           target
100           6.3          3.3           6.0          2.5  Iris-virginica
101           5.8          2.7           5.1          1.9  Iris-virginica
102           7.1          3.0           5.9          2.1  Iris-virginica
103           6.3          2.9           5.6          1.8  Iris-virginica
104           6.5          3.0           5.8          2.2  Iris-virginica
105           7.6          3.0           6.6          2.1  Iris-virginica
106           4.9          2.5           4.5          1.7  Iris-virginica
107           7.3          2.9           6.3          1.8  Iris-virginica
108           6.7          2.5           5.8          1.8  Iris-virginica
109           7.2          3.6           6.1          2.5  Iris-virginica

Dimensions (10, 5)
     sepal_length  sepal_width  petal_length  petal_width           target
110           6.5          3.2           5.1          2.0  Iris-virginica
111           6.4          2.7           5.3          1.9  Iris-virginica
112           6.8          3.0           5.5          2.1  Iris-virginica
```

```
113           5.7          2.5           5.0          2.0  Iris-virginica
114           5.8          2.8           5.1          2.4  Iris-virginica
115           6.4          3.2           5.3          2.3  Iris-virginica
116           6.5          3.0           5.5          1.8  Iris-virginica
117           7.7          3.8           6.7          2.2  Iris-virginica
118           7.7          2.6           6.9          2.3  Iris-virginica
119           6.0          2.2           5.0          1.5  Iris-virginica

Dimensions (10, 5)
     sepal_length  sepal_width  petal_length  petal_width         target
120           6.9          3.2           5.7          2.3  Iris-virginica
121           5.6          2.8           4.9          2.0  Iris-virginica
122           7.7          2.8           6.7          2.0  Iris-virginica
123           6.3          2.7           4.9          1.8  Iris-virginica
124           6.7          3.3           5.7          2.1  Iris-virginica
125           7.2          3.2           6.0          1.8  Iris-virginica
126           6.2          2.8           4.8          1.8  Iris-virginica
127           6.1          3.0           4.9          1.8  Iris-virginica
128           6.4          2.8           5.6          2.1  Iris-virginica
129           7.2          3.0           5.8          1.6  Iris-virginica

Dimensions (10, 5)
     sepal_length  sepal_width  petal_length  petal_width         target
130           7.4          2.8           6.1          1.9  Iris-virginica
131           7.9          3.8           6.4          2.0  Iris-virginica
132           6.4          2.8           5.6          2.2  Iris-virginica
133           6.3          2.8           5.1          1.5  Iris-virginica
134           6.1          2.6           5.6          1.4  Iris-virginica
135           7.7          3.0           6.1          2.3  Iris-virginica
136           6.3          3.4           5.6          2.4  Iris-virginica
137           6.4          3.1           5.5          1.8  Iris-virginica
138           6.0          3.0           4.8          1.8  Iris-virginica
139           6.9          3.1           5.4          2.1  Iris-virginica

Dimensions (10, 5)
     sepal_length  sepal_width  petal_length  petal_width         target
140           6.7          3.1           5.6          2.4  Iris-virginica
141           6.9          3.1           5.1          2.3  Iris-virginica
142           5.8          2.7           5.1          1.9  Iris-virginica
143           6.8          3.2           5.9          2.3  Iris-virginica
144           6.7          3.3           5.7          2.5  Iris-virginica
145           6.7          3.0           5.2          2.3  Iris-virginica
146           6.3          2.5           5.0          1.9  Iris-virginica
147           6.5          3.0           5.2          2.0  Iris-virginica
148           6.2          3.4           5.4          2.3  Iris-virginica
149           5.9          3.0           5.1          1.8  Iris-virginica
```

```
In [7]: #### lazy read iterator
        iris_iterator = pd.read_csv(iris_file, header=None, names=['sepal_length', 'sepal_width
                                                                   'petal_length','petal_width',
                                                                   'target'], iterator=True)
        print(iris_iterator.get_chunk(10).shape)
        print(iris_iterator.get_chunk(2))
```

```
(10, 5)
    sepal_length  sepal_width  petal_length  petal_width       target
10           5.4          3.7           1.5          0.2  Iris-setosa
11           4.8          3.4           1.6          0.2  Iris-setosa
```

```
In [8]: #### batch read CSV
        import csv
        import numpy as np
        def batch_read(file_name, batch = 5):
            with open(file_name, 'rt') as data_stream:
                batch_output = list()
                for n, row in enumerate(csv.reader(data_stream, dialect='excel')):
                    if n > 0 and n % batch == 0:
                        yield(np.array(batch_output))
                        batch_output = list()
                    batch_output.append(row)
                leftyield(np.array(batch_output))

        batch =  batch_read(iris_file, batch=3)
        print(next(batch))
        print(next(batch))
```

```
[['5.1' '3.5' '1.4' '0.2' 'Iris-setosa']
 ['4.9' '3.0' '1.4' '0.2' 'Iris-setosa']
 ['4.7' '3.2' '1.3' '0.2' 'Iris-setosa']]
[['4.6' '3.1' '1.5' '0.2' 'Iris-setosa']
 ['5.0' '3.6' '1.4' '0.2' 'Iris-setosa']
 ['5.4' '3.9' '1.7' '0.4' 'Iris-setosa']]
```

```
In [9]: #### Custom data frame
        custom_data_frame = pd.DataFrame({'Col1': range(5), 'Col2':[1.0] * 5, "Col3": 1, 'Col4
        print(custom_data_frame)
```

```
   Col1  Col2  Col3       Col4
0     0   1.0     1  Test Test
1     1   1.0     1  Test Test
2     2   1.0     1  Test Test
3     3   1.0     1  Test Test
4     4   1.0     1  Test Test
```

```
In [10]: #### Maks data
         mask_feature = iris['sepal_length'] > 6.0
         print(mask_feature)
```

```
0      False
1      False
2      False
3      False
4      False
5      False
6      False
7      False
8      False
9      False
10     False
11     False
12     False
13     False
14     False
15     False
16     False
17     False
18     False
19     False
20     False
21     False
22     False
23     False
24     False
25     False
26     False
27     False
28     False
29     False
       ...
120     True
121    False
122     True
123     True
124     True
125     True
126     True
127     True
128     True
129     True
130     True
131     True
132     True
```

```
133     True
134     True
135     True
136     True
137     True
138    False
139     True
140     True
141     True
142    False
143     True
144     True
145     True
146     True
147     True
148     True
149    False
Name: sepal_length, Length: 150, dtype: bool
```

```
In [11]: ##### Mask data #2 set new value
         mask_target = iris['target'] == 'Iris-virginica'
         iris.loc[mask_target, 'target'] = 'New value'
         print(iris['target'].unique())
```

```
['Iris-setosa' 'Iris-versicolor' 'New value']
```

```
In [12]: ##### Grouping
         grouped_mean = iris.groupby(['target']).mean() # mean
         print(grouped_mean)
         grouped_variance = iris.groupby(['target']).var() # variance
         print(grouped_variance)
```

```
                sepal_length  sepal_width  petal_length  petal_width
target
Iris-setosa            5.006        3.418         1.464        0.244
Iris-versicolor        5.936        2.770         4.260        1.326
New value              6.588        2.974         5.552        2.026
                sepal_length  sepal_width  petal_length  petal_width
target
Iris-setosa         0.124249     0.145180      0.030106     0.011494
Iris-versicolor     0.266433     0.098469      0.220816     0.039106
New value           0.404343     0.104004      0.304588     0.075433
```

```
In [13]: #### sorting
         iris.sort_values(by='sepal_length').head()
```

Out[13]:

	sepal_length	sepal_width	petal_length	petal_width	target
13	4.3	3.0	1.1	0.1	Iris-setosa
42	4.4	3.2	1.3	0.2	Iris-setosa
38	4.4	3.0	1.3	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
41	4.5	2.3	1.3	0.3	Iris-setosa

In [14]:
```
#### Use one apply, execute function for rows or columns
iris.apply(np.count_nonzero, axis=1).head()
```

Out[14]:
```
0    5
1    5
2    5
3    5
4    5
dtype: int64
```

In [15]:
```
#### Use one apply, execute function for rows or columns
iris.apply(np.count_nonzero, axis=0).head()
```

Out[15]:
```
sepal_length    150
sepal_width     150
petal_length    150
petal_width     150
target          150
dtype: int64
```

In [16]:
```
#### Use one apply, execute function for rows or columns
iris.apply(lambda x: x is not 2, axis=1).head()
```

Out[16]:
```
0    True
1    True
2    True
3    True
4    True
dtype: bool
```

In [17]:
```
#### Use one applymap
iris.applymap(lambda el: len(str(el))).head()
```

Out[17]:

	sepal_length	sepal_width	petal_length	petal_width	target
0	3	3	3	3	11
1	3	3	3	3	11
2	3	3	3	3	11
3	3	3	3	3	11
4	3	3	3	3	11

In [29]:
```
#### Get data
print(iris['sepal_width'][20])
```

```
print(iris.loc[20, 'sepal_width'])
print(iris.iloc[20, 1])

print(iris[['sepal_width', 'petal_length']][18:20])

print(iris.loc[[18,19,20], ['sepal_width', 'petal_length']])
```

```
3.4
3.4
3.4
    sepal_width  petal_length
18          3.8           1.7
19          3.8           1.5
    sepal_width  petal_length
18          3.8           1.7
19          3.8           1.5
20          3.4           1.7
```

In [ ]:

In [ ]: