The Engineering World #DataScience 3 & 4

May 31, 2018

AKKAL BAHADUR BIST
DATA SCIENTIST AT
KATHMANDU INSTITUTE OF APPLIED SCIENCES (KIAS)
Center for Conservation Biology (CCB)

1 HOW TO REMOVE DUPLICATE DATA

1.0.1 Remove duplicate data

```
In [1]: import numpy as np
        import pandas as pd
        from pandas import Series, DataFrame
```

```
In [2]: DF_obj = DataFrame({'Roll': [1, 2, 3, 4, 5, 6, 7, 8, 3, 3, 5, 5, 5],
                           'Name': ['Akkal', 'Janak', 'Laxman', 'Dinesh', 'Amin', 'Bikash', 'Sum
                           'Marks': [10, 20, 30, 40, 50, 60, 70, 80, 30, 30, 50, 50, 50]})
```

```
In [3]: DF_obj
```

```
Out[3]:     Marks    Name  Roll
        0      10   Akkal     1
        1      20   Janak     2
        2      30  Laxman     3
        3      40  Dinesh     4
        4      50    Amin     5
        5      60  Bikash     6
        6      70   Sumil     7
        7      80   Kiran     8
        8      30  Laxman     3
        9      30  Laxman     3
        10     50    Amin     5
        11     50    Amin     5
        12     50    Amin     5
```

```
In [4]: DF_obj.duplicated()
```

```
Out[4]: 0     False
        1     False
```

```
2     False
3     False
4     False
5     False
6     False
7     False
8      True
9      True
10     True
11     True
12     True
dtype: bool
```

```
In [5]: DF_obj.drop_duplicates()
```

```
Out[5]:    Marks    Name  Roll
        0     10   Akkal     1
        1     20   Janak     2
        2     30  Laxman     3
        3     40  Dinesh     4
        4     50    Amin     5
        5     60  Bikash     6
        6     70   Sumil     7
        7     80   Kiran     8
```

```
In [6]: DF_obj.drop_duplicates(['Marks'])
```

```
Out[6]:    Marks    Name  Roll
        0     10   Akkal     1
        1     20   Janak     2
        2     30  Laxman     3
        3     40  Dinesh     4
        4     50    Amin     5
        5     60  Bikash     6
        6     70   Sumil     7
        7     80   Kiran     8
```

```
In [7]: DF_obj
```

```
Out[7]:     Marks    Name  Roll
        0      10   Akkal     1
        1      20   Janak     2
        2      30  Laxman     3
        3      40  Dinesh     4
        4      50    Amin     5
        5      60  Bikash     6
        6      70   Sumil     7
        7      80   Kiran     8
        8      30  Laxman     3
```

```
9       30  Laxman     3
10      50    Amin     5
11      50    Amin     5
12      50    Amin     5
```

2 CONCATINATING AND TRANSFORMING DATA

2.0.1 Concatinating Data

```
In [8]: DF_obj = pd.DataFrame(np.arange(36).reshape(6,6))
```

```
In [9]: DF_obj
```

```
Out[9]:     0   1   2   3   4   5
        0   0   1   2   3   4   5
        1   6   7   8   9  10  11
        2  12  13  14  15  16  17
        3  18  19  20  21  22  23
        4  24  25  26  27  28  29
        5  30  31  32  33  34  35
```

```
In [10]: DF_obj_2 = pd.DataFrame(np.arange(15).reshape(5,3))
```

```
In [11]: DF_obj_2
```

```
Out[11]:     0   1   2
         0   0   1   2
         1   3   4   5
         2   6   7   8
         3   9  10  11
         4  12  13  14
```

```
In [12]: pd.concat([DF_obj,DF_obj_2], axis = 1)
```

```
Out[12]:     0   1   2   3   4   5     0     1     2
         0   0   1   2   3   4   5   0.0   1.0   2.0
         1   6   7   8   9  10  11   3.0   4.0   5.0
         2  12  13  14  15  16  17   6.0   7.0   8.0
         3  18  19  20  21  22  23   9.0  10.0  11.0
         4  24  25  26  27  28  29  12.0  13.0  14.0
         5  30  31  32  33  34  35   NaN   NaN   NaN
```

```
In [13]: pd.concat([DF_obj,DF_obj_2])
```

```
Out[13]:     0   1   2     3     4     5
         0   0   1   2   3.0   4.0   5.0
         1   6   7   8   9.0  10.0  11.0
         2  12  13  14  15.0  16.0  17.0
         3  18  19  20  21.0  22.0  23.0
```

```
4  24  25  26  27.0  28.0  29.0
5  30  31  32  33.0  34.0  35.0
0   0   1   2   NaN   NaN   NaN
1   3   4   5   NaN   NaN   NaN
2   6   7   8   NaN   NaN   NaN
3   9  10  11   NaN   NaN   NaN
4  12  13  14   NaN   NaN   NaN
```

2.0.2 Trandforming Data

Dropping data

```
In [14]: DF_obj.drop([0,2])
```

```
Out[14]:     0   1   2   3   4   5
         1   6   7   8   9  10  11
         3  18  19  20  21  22  23
         4  24  25  26  27  28  29
         5  30  31  32  33  34  35
```

```
In [15]: DF_obj.drop([0,2], axis = 1)
```

```
Out[15]:     1   3   4   5
         0   1   3   4   5
         1   7   9  10  11
         2  13  15  16  17
         3  19  21  22  23
         4  25  27  28  29
         5  31  33  34  35
```

Adding Data

```
In [16]: series_obj = Series(np.arange(6))
         series_obj.name = 'aded_variables'
         series_obj
```

```
Out[16]: 0    0
         1    1
         2    2
         3    3
         4    4
         5    5
         Name: aded_variables, dtype: int64
```

```
In [17]: variable_added = DataFrame.join(DF_obj,series_obj)
```

```
In [18]: variable_added
```

```
Out[18]:     0   1   2   3   4   5  aded_variables
         0   0   1   2   3   4   5               0
         1   6   7   8   9  10  11               1
         2  12  13  14  15  16  17               2
         3  18  19  20  21  22  23               3
         4  24  25  26  27  28  29               4
         5  30  31  32  33  34  35               5
```

```
In [19]: added_datatable = variable_added.append(variable_added, ignore_index = False)
```

```
In [20]: added_datatable
```

```
Out[20]:     0   1   2   3   4   5  aded_variables
         0   0   1   2   3   4   5               0
         1   6   7   8   9  10  11               1
         2  12  13  14  15  16  17               2
         3  18  19  20  21  22  23               3
         4  24  25  26  27  28  29               4
         5  30  31  32  33  34  35               5
         0   0   1   2   3   4   5               0
         1   6   7   8   9  10  11               1
         2  12  13  14  15  16  17               2
         3  18  19  20  21  22  23               3
         4  24  25  26  27  28  29               4
         5  30  31  32  33  34  35               5
```

```
In [21]: added_datatable = variable_added.append(variable_added, ignore_index = True)
         added_datatable
```

```
Out[21]:      0   1   2   3   4   5  aded_variables
         0    0   1   2   3   4   5               0
         1    6   7   8   9  10  11               1
         2   12  13  14  15  16  17               2
         3   18  19  20  21  22  23               3
         4   24  25  26  27  28  29               4
         5   30  31  32  33  34  35               5
         6    0   1   2   3   4   5               0
         7    6   7   8   9  10  11               1
         8   12  13  14  15  16  17               2
         9   18  19  20  21  22  23               3
         10  24  25  26  27  28  29               4
         11  30  31  32  33  34  35               5
```

Sorting data

```
In [22]: DF_sorted = DF_obj.sort_values(by = [5], ascending = [False])
```

```
In [23]: DF_sorted
```

```
Out[23]:     0   1   2   3   4   5
         5  30  31  32  33  34  35
         4  24  25  26  27  28  29
         3  18  19  20  21  22  23
         2  12  13  14  15  16  17
         1   6   7   8   9  10  11
         0   0   1   2   3   4   5
```

```
In [24]: DF_sorted = DF_obj.sort_values(by = [5], ascending = [True])
```

```
In [25]: DF_sorted
```

```
Out[25]:     0   1   2   3   4   5
         0   0   1   2   3   4   5
         1   6   7   8   9  10  11
         2  12  13  14  15  16  17
         3  18  19  20  21  22  23
         4  24  25  26  27  28  29
         5  30  31  32  33  34  35
```

```
In [26]: DF_sorted = DF_obj.sort_values(by = [5])
         DF_sorted
```

```
Out[26]:     0   1   2   3   4   5
         0   0   1   2   3   4   5
         1   6   7   8   9  10  11
         2  12  13  14  15  16  17
         3  18  19  20  21  22  23
         4  24  25  26  27  28  29
         5  30  31  32  33  34  35
```