Chapter 2 - Data Preparation Basics

Segment 3 - Removing duplicates

```
import numpy as np
import pandas as pd

from pandas import Series, DataFrame


a=np.array([100,200,300,400,500])
b=np.array([1,2,3,4,5])
#print(np.subtract(a,b))
a.size
b=np.array([[22,33,44,55,77]])
print(b.T)
```

```
[[22]
 [33]
 [44]
 [55]
 [77]]
```

```
a=np.array([3,5+1j,3],dtype=complex)
print(a)
```

```
[3.+0.j 5.+1.j 3.+0.j]
```

```
np.std(np.array([]))
```

```
18.925115587493778
```

Double-click (or enter) to edit

```
from scipy.stats import norm
print(norm.var())
```

```
1.0
```

```
df=pd.DataFrame({
    'a':[1,4,5,2,6,9],
    'b':[0,3,2,1,8,7]
})
df.loc[[0,2,4][:]]
```

	a	b
0	1	0
2	5	2
4	6	8

```
np.std(np.array([23,26,21,16,33,35,46,42,51,18]))
```

```
11.614215427655887
```

```
import sys
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-5-b9ae7b76397a> in <module>()
      1 import sys
----> 2 info(sys)

NameError: name 'info' is not defined
```

SEARCH STACK OVERFLOW

Removing duplicates

```
import re
```

```
DF_obj= DataFrame({'column 1':[1,1,2,2,3,3,3],
```

```
                   'column 2':['a', 'a','b', 'b', 'c', 'c', 'c'],
                   'column 3':['A', 'A', 'B', 'B', 'C', 'C', 'C']})
#df=DataFrame({'co1'=[1,2,3], 'col2'=[3,1]})
DF_obj
```

	column 1	column 2	column 3
0	1	a	A
1	1	a	A
2	2	b	B
3	2	b	B
4	3	c	C
5	3	c	C
6	3	c	C

```
p=r'1(1+0)'


DF_obj.duplicated()
#df.duplicated()
```

```
0    False
1     True
2    False
3     True
4    False
5     True
6     True
dtype: bool
```

```
DF_obj.drop_duplicates()
#df.drop_duplicates()
```

	column 1	column 2	column 3
0	1	a	A

```
DF_obj= DataFrame({'column 1':[1,1,2,2,3,3,3],
                   'column 2':['a', 'a','b', 'b', 'c', 'c', 'c'],
                   'column 3':['A', 'A', 'B', 'B', 'C', 'D', 'C']})
DF_obj
```

	column 1	column 2	column 3
0	1	a	A
1	1	a	A
2	2	b	B
3	2	b	B
4	3	c	C
5	3	c	D
6	3	c	C

Double-click (or enter) to edit

```
DF_obj.drop_duplicates(['column 2'])
#df.drop_duplicates(['col 2'])
```

	column 1	column 2	column 3
0	1	a	A
2	2	b	B
4	3	c	C

Colab paid products - Cancel contracts here