classwork/data_grp Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled U

Zeppelin

classwork/data_grp

default

Untitled

FINISHED

```
%pyspark
col = ['Identification',  'Salesprice' , 'Finishedsquarefeet', 'Numberofbedrooms' ,'Numberofb
,'Pool', 'Yearbuilt', 'Quality' , 'style', 'Lotsize' , 'Adjacenttohighway']
```

Took 0 sec. Last updated by anonymous at March 30 2017, 7:00:21 PM.

FINISHED

```
%pyspark
start = timeit.timeit()
data1 = pd.read_csv('/Users/jyothi/Desktop/real.csv')
data1.columns  = col
end = timeit.timeit()
print(end - start)
```

```
0.006752923989552073
```

Took 0 sec. Last updated by anonymous at March 30 2017, 7:05:52 PM.

FINISHED

```
%pyspark
import timeit
start = timeit.timeit()
print("hello")
end = timeit.timeit()
print(end - start)
```

```
hello
-0.00791193799523171
```

Took 0 sec. Last updated by anonymous at March 30 2017, 8:01:57 PM.

FINISHED

```
%pyspark
data1.head()
```

```
   Identification  Salesprice  Finishedsquarefeet  Numberofbedrooms  \
0               2      340000                2058                 4
1               3      250000                1780                 4
2               4      205500                1638                 4
3               5      275500                2196                 4
4               6      248000                1966                 4
   Numberofbathrooms  Airconditioning  Garagesize  Pool  Yearbuilt  Quality  \
0                  2                1           2     0       1976        2
1                  3                1           2     0       1980        2
2                  2                1           2     0       1963        2
3                  3                1           2     0       1968        2
4                  3                1           5     1       1972        2
   style  Lotsize  Adjacenttohighway
0      1    22912                  0
```

classwork/data_grp

classwork/data_grp

Zeppelin

Took 0 sec. Last updated by anonymous at March 30 2017, 7:02:14 PM.

classwork/data_grp

default

FINISHED

```
%pyspark
from pandas import Series, DataFrame
import pandas as pd
import numpy as np
```

Took 0 sec. Last updated by anonymous at March 30 2017, 7:02:17 PM.

FINISHED

```
%pyspark
start = timeit.timeit()
grouped = data1.groupby('Yearbuilt')
end = timeit.timeit()
print(end - start)
```

```
-0.00442620599642396
```

Took 0 sec. Last updated by anonymous at March 30 2017, 7:40:33 PM.

FINISHED

```
%pyspark
start = timeit.timeit()
data1.info()
end = timeit.timeit()
print(end - start)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 521 entries, 0 to 520
Data columns (total 13 columns):
Identification        521 non-null int64
Salesprice            521 non-null int64
Finishedsquarefeet    521 non-null int64
Numberofbedrooms      521 non-null int64
Numberofbathrooms     521 non-null int64
Airconditioning       521 non-null int64
Garagesize            521 non-null int64
Pool                  521 non-null int64
Yearbuilt             521 non-null int64
Quality               521 non-null int64
style                 521 non-null int64
Lotsize               521 non-null int64
Adjacenttohighway     521 non-null int64
dtypes: int64(13)
memory usage: 53.0 KB
```

Took 0 sec. Last updated by anonymous at March 30 2017, 8:02:02 PM.

FINISHED

```
%pyspark
start = timeit.timeit()
data1[-4:]
end = timeit.timeit()
print(end - start)
```

classwork/data_grp

classwork/data_grp Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled U

Zeppelin

Took 0 sec. Last updated by anonymous at March 30 2017, 8:02:05 PM.

classwork/data_grp

default

FINISHED

```
%pyspark
start = timeit.timeit()
lotsize_corr = lambda x: x.corrwith(x['Lotsize'])
end = timeit.timeit()
print(end - start)
```

```
0.008321376008098014
```

Took 0 sec. Last updated by anonymous at March 30 2017, 8:02:10 PM.

FINISHED

```
%pyspark
start = timeit.timeit()
import statsmodels.api as sm
def regression(data, yvar, xvars):
  Y = data[yvar]
  X = data[xvars]
  X['intercept'] = 1.
  result = sm.OLS(Y,X).fit()
  return result.params
  end = timeit.timeit()
print(end - start)
```

```
0.0040694080089451745
```

Took 0 sec. Last updated by anonymous at March 30 2017, 8:02:13 PM.

FINISHED

```
%pyspark
start = timeit.timeit()
xvar = [ 'Finishedsquarefeet', 'Numberofbedrooms' ,'Numberofbathrooms', 'Airconditioning',
'style', 'Lotsize' , 'Adjacenttohighway']
end = timeit.timeit()
print(end - start)
```

```
-0.0036503760056803003
```

Took 0 sec. Last updated by anonymous at March 30 2017, 8:02:16 PM.

FINISHED

```
%pyspark
start = timeit.timeit()
by_Numberofbathrooms = data1.groupby('Numberofbathrooms')
end = timeit.timeit()
print(end - start)
```

```
0.004003459005616605
```

Took 0 sec. Last updated by anonymous at March 30 2017, 7:39:50 PM.

FINISHED

```
%pyspark
start = timeit.timeit()

by_Numberofbathrooms.apply(regression,'Salesprice',xvar)
```

```
end = timeit.timeit()
print(end - start)
```

Zeppelin

classwork/data_grp

default

Took 0 sec. Last updated by anonymous at March 30 2017, 7:42:20 PM. (outdated)

FINISHED

```
%pyspark


by_Numberofbathrooms.apply(regression,'Salesprice',xvar)
```

	Finishedsquarefeet	Numberofbedrooms	Numberofbathrooms	Numberofbathrooms
0	0.799334	0.000000	0.000000e+00	
1	41.704401	7789.667268	-3.624188e+05	
2	89.617662	7156.517226	-5.236697e+05	
3	117.966949	-8478.624895	-5.865658e+05	
4	112.261047	469.580325	-1.217455e+06	
5	217.747483	-24555.066427	-4.799156e+06	
6	0.852219	0.001028	1.028216e-03	
7	79.841580	0.093354	2.869320e-01	

	Airconditioning	Garagesize	Pool	Yearbuilt	Numberofbathrooms
0	0.000375	0.001126	0.000000	0.747897	
1	3884.646014	5832.481388	9457.310636	422.877244	
2	11422.173069	-5425.120839	14027.973629	666.404923	
3	-25415.981403	2418.531835	28820.314519	1137.685472	
4	-81398.314280	1412.029447	39438.656304	2758.055840	
5	-57923.740050	1084.397776	41449.484623	12554.121394	

Took 1 sec. Last updated by anonymous at March 30 2017, 7:42:47 PM.

FINISHED

Took 0 sec. Last updated by anonymous at March 30 2017, 8:00:41 PM.

READY