DS-670 Quantile a…

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

FINISHED

```
%pyspark
frame = DataFrame({'data1': np.random.randn(1000), 'data2': np.random.randn(1000)})
factor = pd.cut(frame.data1,4)
factor[:10]
```

FINISHED

```
0     (0.104, 1.701]
1     (0.104, 1.701]
2    (-1.493, 0.104]
3    (-1.493, 0.104]
4     (0.104, 1.701]
5    (-1.493, 0.104]
6    (-1.493, 0.104]
7    (-1.493, 0.104]
8    (-1.493, 0.104]
9     (0.104, 1.701]
Name: data1, dtype: category
Categories (4, object): [(-3.0969, -1.493] < (-1.493, 0.104] < (0.104, 1.701] < (1.701, 3.2
99]]
```

```
%pyspark
def get_stats(group):
  return {'min': group.min(), 'max': group.max(), 'count': group.count(), 'mean': group.mea

grouped = frame.data2.groupby(factor)
grouped.apply(get_stats).unstack()
```

FINISHED

```
                   count       max      mean       min
data1
(-3.0969, -1.493]   77.0  2.170535  0.047746 -2.673867
(-1.493, 0.104]    458.0  3.069384 -0.017941 -3.595508
(0.104, 1.701]     410.0  2.694214  0.112674 -3.054598
(1.701, 3.299]      55.0  2.598212  0.016491 -2.151662
```

```
%pyspark
# compute quantile numbers
grouping = pd.qcut(frame.data1, 10, labels=False)

grouped = frame.data2.groupby(grouping)
```

FINISHED

```
       count       max      mean       min
data1
0      100.0  2.170535 -0.000958 -2.673867
1      100.0  3.069384  0.117220 -2.416973
2      100.0  2.570141 -0.018772 -2.101784
3      100.0  2.701397  0.080765 -2.215289
4      100.0  1.689063 -0.155770 -3.595508
5      100.0  1.898770 -0.085276 -2.592917
6      100.0  2.694214  0.227859 -2.140839
7      100.0  2.438655  0.049405 -3.054598
8      100.0  2.521813  0.178774 -2.253160
9      100.0  2.598212  0.032378 -2.424888
```

```
%pyspark
```

FINISHED