



海豚大数据实验室

案例分析: 北京房价数据分析

本案例中我们将针对2012年欧洲杯数据进行分析, 训练Pandas库的使用

导入数据

导入所需的 pandas 库

```
In [4]: import pandas as pd  
import numpy as np
```

使用 `pd.read_csv` 导入数据

```
In [5]: data = pd.read_csv("/resources/jupyter/Train001/Train001-003_a/lianjia.csv")

/usr/local/lib/python3.5/dist-packages/IPython/core/interactiveshell.py:2705: DtypeWarning: Columns (1,11,12,13,14,27) have mixed types. Specify dtype option on import or set low_memory=False.
interactivity=interactivity, compiler=compiler, result=result)
```

房子的建造时间都有哪些？

```
In [6]: np.sort(data['Construction time'].unique())

Out[6]: array(['0', '1', '1906', '1914', '1933', '1934', '1944', '1950', '1952',
               '1953', '1954', '1955', '1956', '1957', '1958', '1959', '1960',
               '1961', '1962', '1963', '1964', '1965', '1966', '1967', '1968',
               '1969', '1970', '1971', '1972', '1973', '1974', '1975', '1976',
               '1977', '1978', '1979', '1980', '1981', '1982', '1983', '1984',
               '1985', '1986', '1987', '1988', '1989', '1990', '1991', '1992',
               '1993', '1994', '1995', '1996', '1997', '1998', '1999', '2000',
               '2001', '2002', '2003', '2004', '2005', '2006', '2007', '2008',
               '2009', '2010', '2011', '2012', '2013', '2014', '2015', '2016', '未知'], dtype=object)
```

建造时间最早的房子卖出去多少钱？

```
In [7]: data[data['Construction time']=='1906'][['Construction time', 'Total price']]
```

```
Out[7]:
```

	Construction time	Total price
298468	1906	950.0

哪个区卖出的房最多？

```
In [8]: district_mean=data["District"].value_counts().sort_values(ascending=False)
district_mean
```

```
Out[8]: 朝阳      107244
        昌平      38634
        海淀      38200
        西城      31293
        丰台      29338
        东城      17086
        大兴      15313
        通州      13974
        石景山    11371
        顺义      9202
        房山      2955
        亦庄开发区  2537
        门头沟    1704
Name: District, dtype: int64
```

距离地铁近的房子卖出去多少套？

```
In [9]: data['Subway'].value_counts()
```

```
Out[9]: 1.0    191646
        0.0    127173
Name: Subway, dtype: int64
```

海淀区的房价平均多少钱一平米？

```
In [10]: data[data['District']=='海淀']['Price'].mean()
```

```
Out[10]: 54855.08565445026
```

各种装修情况的房子的平均价格是多少？

```
In [11]: data['renovation condition'].value_counts()
```

```
Out[11]: 其他      118740
          精装      117438
          简装      77251
          毛坯       5390
          0         26
          1          6
          Name: renovation condition, dtype: int64
```

西城区的房子中有多少两居室

```
In [12]: data[(data['District'] == '西城') & (data['Bed Room']==2)].shape[0]
```

```
Out[12]: 13647
```

100平米以上的房子中，几居室最多

```
In [13]: data['Bed Room']=data['Bed Room'].apply(lambda x:str(x))
          data[data['Square']>100]['Bed Room'].value_counts()
```

```
Out[13]: 3      43615
          2      24290
          4      6452
          5      1082
          1       724
          6       217
          7        35
          8         7
          0         5
          #NAME?    5
          9         4
          Name: Bed Room, dtype: int64
```

三居室中塔楼的平均价格为多少

```
In [14]: data[(data['Bed Room']=='3') & (data['Building Type']=='塔楼')]['Price'].mean()
```

```
Out[14]: 44089.023590440316
```

Copyright © 2017 - 2018 海豚大数据实验室 (<http://www.dolphin-labs.com>)