PyCon 2015 大会 上海 - Python大数据分析与可视化

丁来强

- Email: wjo1212 at 163.com
- Wechat: LaiQiangDing

In [1]:

```
%matplotlib inline
%run ./env.py
```

Spark分析读取AFF数据

In [2]:

```
from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)

df = sqlContext.read.json("./aff_big.json")

df.head(n=3)
```

Out[2]:

```
[Row(age=46, language=u'english'),
 Row(age=54, language=u'english'),
 Row(age=65, language=u'english')]
```

average age

In [3]:

```
df.groupBy().avg().collect()
```

Out[3]:

```
[Row(AVG(age)=33.408620073141)]
```

选择所有20~40 英语人士

In [4]:

```
df=df.filter((df.age >= 20) & (df.age <= 40) & (df.language=="english"))\
         .drop("language")
```

转为Pandas DataFrame

In [5]:

```
df = df.toPandas()
```

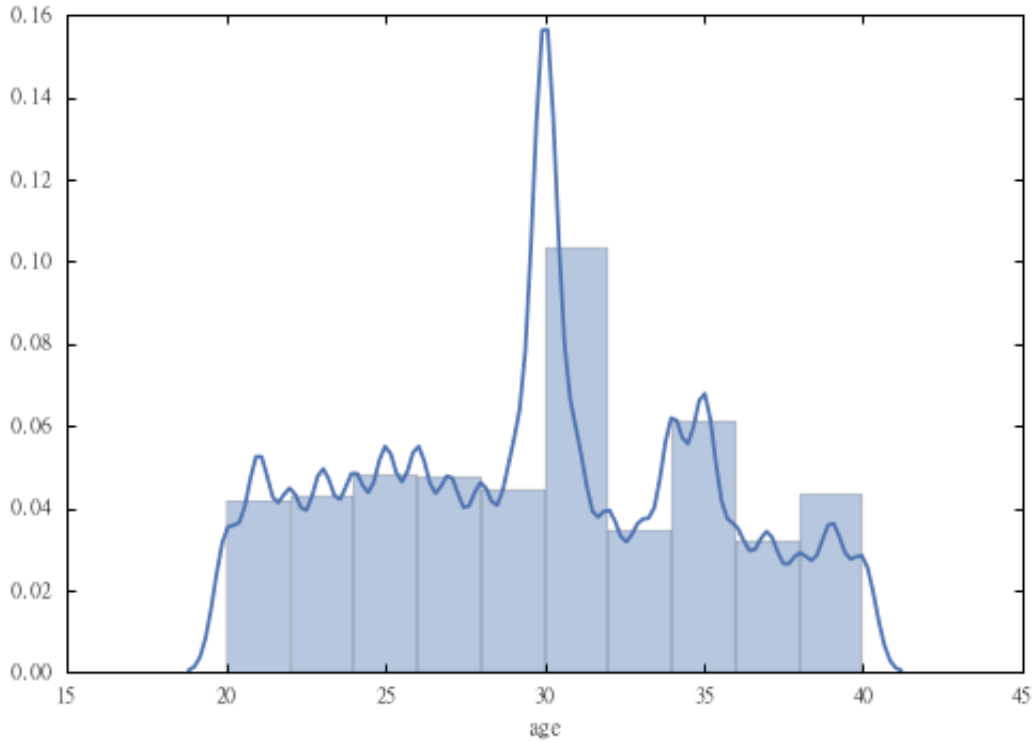
查看年龄分布

In [6]:

```
sns.distplot(df.age, bins=10)
```

Out[6]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x109236810>
```

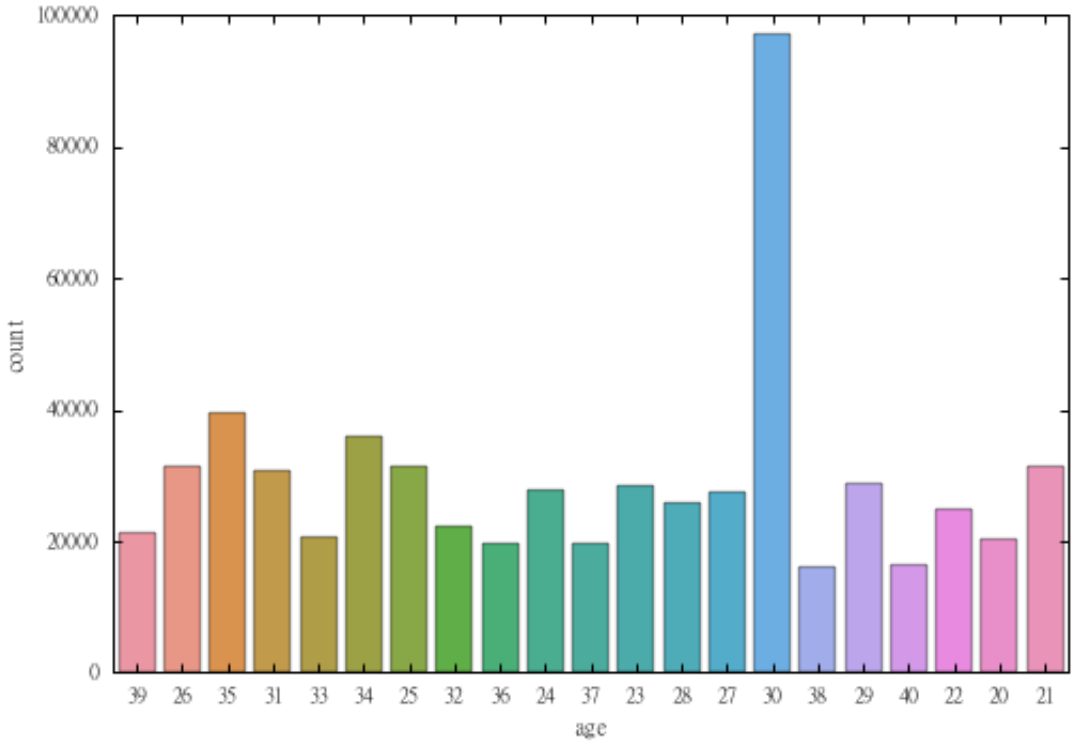


In [7]:

```
sns.countplot(df.age)
```

Out[7]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1092362d0>
```



In [ ]:

```

```