1 目的:

- 熟悉数据集
- 熟悉seaborn各种操作

In [3]:
```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
home = r'F:\database\pandas_dir\seaborn-data-master'
df = sns.load_dataset('titanic', data_home=home)
df.head()
```

Out[3]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

主要字段:

参数	说明
survived	是否获救
pclass	船舱等级
sex	性别
age	年龄
sibsp	家人数量
parch	船上父母或者孩子数量
deek	甲板
fare	票价
embark_town	登船港口
alone	是否单独一个人

2 数据整理

- 缺省值统计
- 缺省值处理：删除或补齐
- 数据二次处理

2.1 统计缺省值

In [ ]:
```

```

In [4]:
```
df.isnull().sum()
```

Out[4]:
```
survived       0
pclass         0
sex            0
age          177
sibsp          0
parch          0
fare           0
embarked       2
class          0
who            0
adult_male     0
deck         688
embark_town    2
alive          0
alone          0
dtype: int64
```

2.2 删除与填充

- 删除deck列

In [6]:
```
pdata = df.drop('deck', axis=1)
```

In [7]:
```
pdata.head()
```

Out[7]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	Southampton	no	True

- 年龄使用均值填充

In [8]:
```
#填充均值
pdata = pdata.fillna(pdata.mean())
#年龄分类
pdata['age_level'] = pd.cut(pdata.age,bins = [0,18,60,100], labels=['child','mid', 'old'])
```

In [ ]:
```

```

In [9]:
```
1 pdata.head()
```

Out[9]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	embark_town	alive	alone	age_level
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	Southampton	no	False	mid
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	Cherbourg	yes	False	mid
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	Southampton	yes	True	mid
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	Southampton	yes	False	mid
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	Southampton	no	True	mid

In [ ]:
```
1
```

3 数据统计

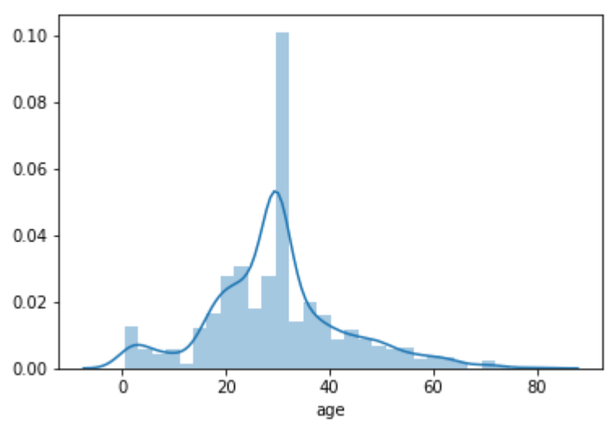
3.1 基础数据统计

- 年龄分布
- 船舱人数分布
- 男女分布
- 团队人数分布

- 年龄较分散，使用直方图进行展示

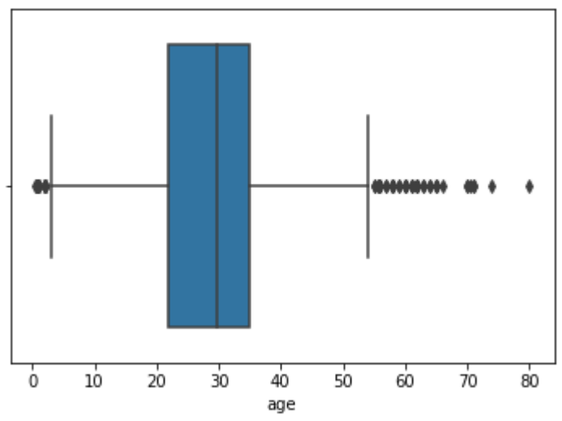
In [10]:
```
1 sns.distplot(pdata.age)
```

Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x1ba018dab70>



In [11]:
```
1 sns.boxplot(pdata.age)
```

Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x1ba01aa2b70>

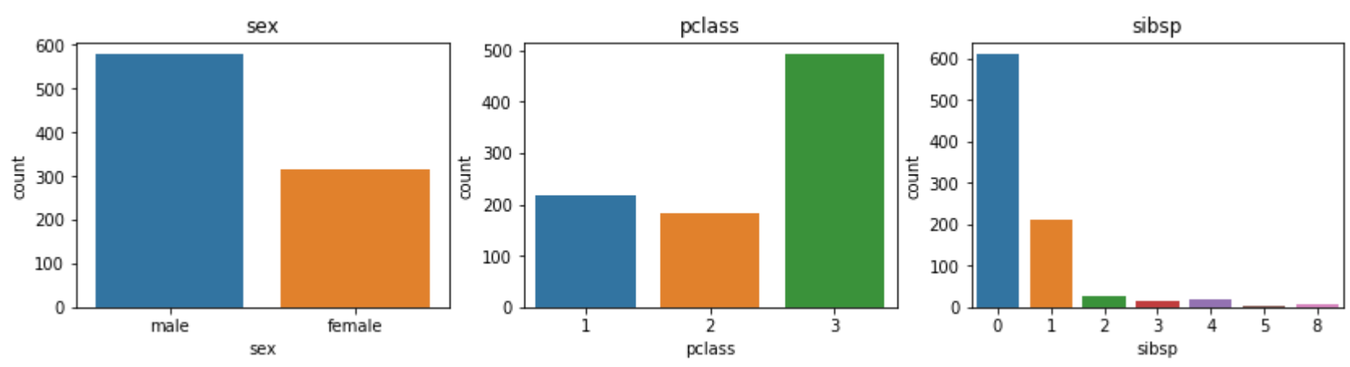


In [ ]:
```
1
```

In [ ]:
```
1
```

- 船舱人数，男女人数，团队人数(1个人，两个人，三个人对应的数量)使用柱状图进行展示

In [162]:
```
1 cols = ['sex', 'pclass', 'sibsp']
2 lens = len(cols)
3 plt.figure(figsize=(14,3))
4 for index, col in enumerate(cols):
5     plt.subplot(1, lens,index+1)
6     ax = sns.countplot(x=col, data=pdata)
7     ax.set_title(col)
```

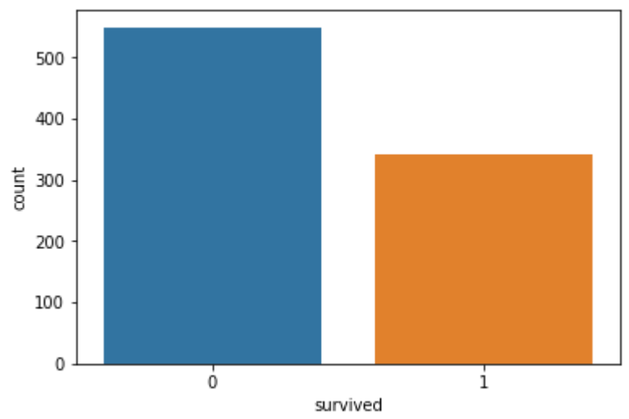


3.2 获救数据

- 获救人数与遇难人数
- 根据性别，统计获救与遇难人数
- 根据年龄段，统计获救与遇难人数
- 根据年龄段，性别，统计获救与遇难人数
- 根据年龄段，性别，船舱，统计获救与遇难人数

In [13]:
```
1 sns.countplot(x='survived', data=pdata)
```

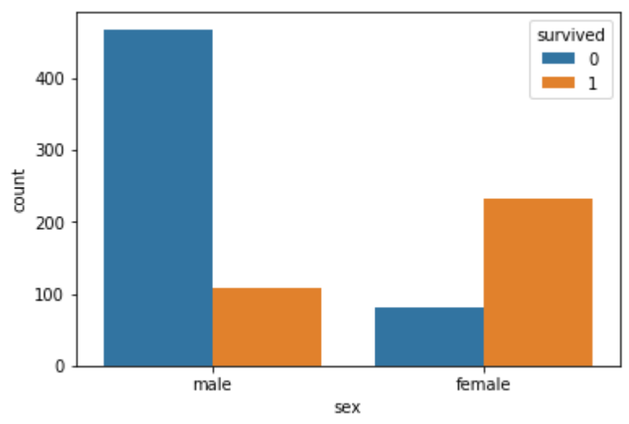
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x1ba02073a90>



- 根据性别进行分类

In [14]:
```
1 sns.countplot(x='sex', data=pdata, hue='survived')
```

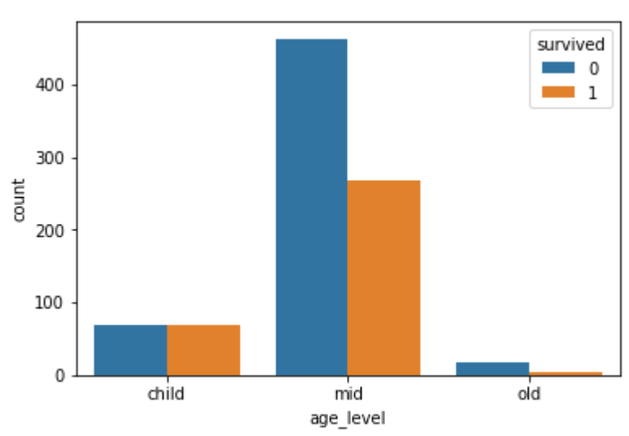
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x1ba021c36d8>



- 年龄与获救关系

In [15]:
```
1 sns.countplot(x='age_level', data=pdata, hue='survived')
```

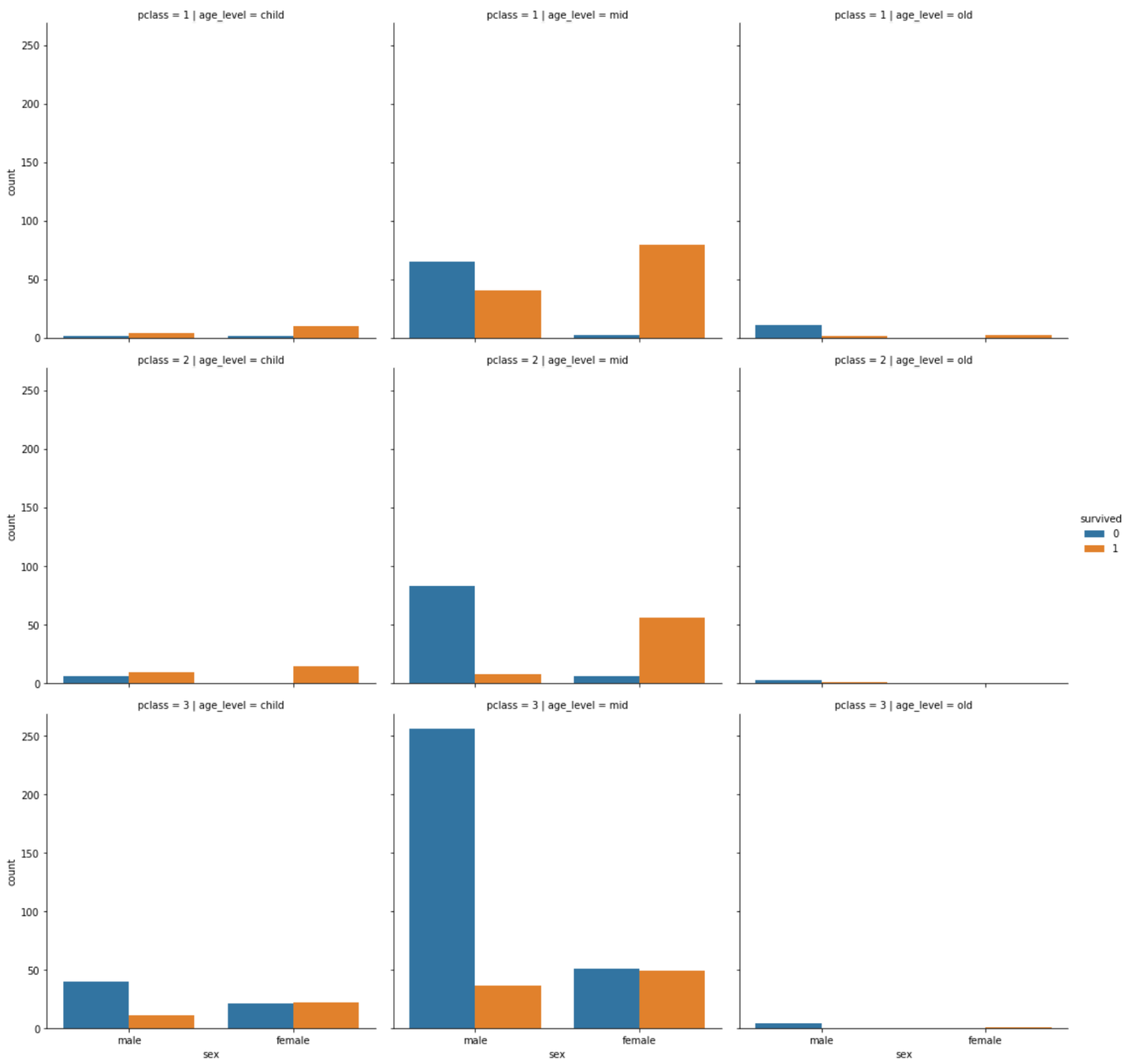
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x1ba021e3080>



- 性别，获救，年龄段，船舱获救统计

In [16]:
```
1 sns.catplot(x='sex',  hue='survived', data=pdata, kind='count', col='age_level', row='pclass')
```

Out[16]: <seaborn.axisgrid.FacetGrid at 0x1ba02062cf8>



In [ ]:
```
1
```