In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [ ]:
```
train_data = pd.read_csv('data/train.csv')
data_test = pd.read_csv('data/test.csv')
train_data.head()
```

In [ ]:
```
train_data.info()
print('-'*40)
data_test.info()
```

In [ ]:
```
# First look at the overall survival ratio
fig = plt.figure(figsize=(6,6))
train_data['Survived'].value_counts().plot.pie(autopct = '%1.2f%%')
```

In [ ]:
```
train_data['Embarked'][train_data['Embarked'].isnull()]
train_data['Embarked'][train_data['Embarked'].isnull()] = train_data['Embarked'].dropr
```

In [ ]:
```
train_data['Cabin'] = train_data['Cabin'].fillna('U0')
```

In [ ]:
```
from sklearn.ensemble import RandomForestRegressor

age_df = train_data[['Age','Survived','Fare', 'Parch', 'SibSp', 'Pclass']]
age_df_notnull = age_df.loc[(train_data['Age'].notnull())]
age_df_isnull = age_df.loc[(train_data['Age'].isnull())]
X = age_df_notnull.values[:,1:]
Y = age_df_notnull.values[:,0]
# use RandomForestRegression to train data
RFR = RandomForestRegressor(n_estimators=1000,n_jobs=-1)
RFR.fit(X,Y)
predictAges = RFR.predict(age_df_isnull.values[:,1:])
train_data.loc[train_data['Age'].isnull(), ['Age']]= predictAges
```

In [ ]:
```
# Next, look at the complementary data
train_data.info()
```

In [ ]:
```
#3-Preliminary data analysis
#3-1-The relationship between gender and survival or not (Sex)
```

In [ ]:
```
train_data.groupby(['Sex','Survived'])['Survived'].count()
```

In [ ]:
```
survived_by_sex = train_data[['Sex','Survived']].groupby('Sex').mean()
type(survived_by_sex)
survived_by_sex.plot.bar()
```

In [ ]:
```
#3-2-The relationship between cabin class and survival or not Pclass
```

In [ ]:
```
train_data.groupby(['Pclass','Survived'])['Pclass'].count()
```

In [ ]:
```
train_data[['Pclass','Survived']].groupby(['Pclass']).mean().plot.bar()
```

In [ ]:
```
train_data.groupby(['Sex', 'Pclass', 'Survived'])['Survived'].count()
```