

In [1]:

```
import pandas as pd
```

In [25]:

```
import pandas as pd
train_url = "http://s3.amazonaws.com/assets.datacamp.com/course/Kaggle/train.csv"
train = pd.read_csv(train_url)

test_url = "http://s3.amazonaws.com/assets.datacamp.com/course/Kaggle/test.csv"
test = pd.read_csv(test_url)
train.head()
```

Out[25]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	7
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	5
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8

In [3]:

```
test.head()
```

Out[3]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN

In [7]:

```
train.to_csv('/Users/liyue/Documents/Downloads/tatonic_train.csv')  
test.to_csv('/Users/liyue/Documents/Downloads/tatonic_test.csv')
```

In [8]:

```
train.shape
```

Out[8]:

(891, 12)

In [9]:

```
train.describe()
```

C:\Users\liyue\Anaconda2\lib\site-packages\numpy\lib\function_base.py:3834: RuntimeWarning: Invalid value encountered in percentile
RuntimeWarning)

Out[9]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Far
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.1
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.1
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.00
25%	223.500000	0.000000	2.000000	NaN	0.000000	0.000000	7.9
50%	446.000000	0.000000	3.000000	NaN	0.000000	0.000000	14.5
75%	668.500000	1.000000	3.000000	NaN	1.000000	0.000000	31.1
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512

In [11]:

```
import pandas as pd

train = pd.read_csv('/Users/liyue/Documents/Downloads/tatonic_train.csv',\
                    sep = ",", header=0)
```

In [12]:

```
train[:3]
```

Out[12]:

	Unnamed: 0	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	1
0	0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A
1	1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	F
2	2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	S

In [13]:

```
train["Survived"][:5]
```

Out[13]:

```
0    0
```

```
1    1
```

```
2    1
```

```
3    1
```

```
4    0
```

Name: Survived, dtype: int64

In [14]:

```
train["Survived"].value_counts()
```

Out[14]:

```
0    549
```

```
1    342
```

Name: Survived, dtype: int64

In [15]:

```
train["Survived"].value_counts(normalize = True)
```

Out[15]:

```
0    0.616162
```

```
1    0.383838
```

Name: Survived, dtype: float64

In [16]:

```
train['Sex'].value_counts()
```

Out[16]:

```
male    577
```

```
female  314
```

Name: Sex, dtype: int64

In [17]:

```
train["Survived"][train["Sex"] == 'male'].value_counts()
```

Out[17]:

```
0    468
```

```
1    109
```

Name: Survived, dtype: int64

In [18]:

```
train["Survived"][train["Sex"] == 'female'].value_counts()
```

Out[18]:

```
1    233
```

```
0     81
```

Name: Survived, dtype: int64

In [19]:

```
train["Survived"][train["Sex"] == 'male'].value_counts(normalize = True)
```

Out[19]:

```
0    0.811092
1    0.188908
Name: Survived, dtype: float64
```

In [20]:

```
train["Survived"][train["Sex"] == 'female'].value_counts(normalize = True)
```

Out[20]:

```
1    0.742038
0    0.257962
Name: Survived, dtype: float64
```

In [21]:

```
train["Child"] = float('NaN')
train.Child[train.Age < 5] = 1
train.Child[train.Age >= 5] = 0
print train.Child[:3]
```

```
0    0.0
1    0.0
2    0.0
Name: Child, dtype: float64
```

C:\Users\liyue\Anaconda2\lib\site-packages\ipykernel__main__.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy>

```
from ipykernel import kernelapp as app
```

C:\Users\liyue\Anaconda2\lib\site-packages\ipykernel__main__.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy>

```
app.launch_new_instance()
```

In [22]:

```
train.Survived[train.Child == 1].value_counts(normalize = True)
```

Out[22]:

```
1    0.675
0    0.325
Name: Survived, dtype: float64
```

In [23]:

```
train.Survived[train.Child == 0].value_counts(normalize = True)
```

Out[23]:

```
0    0.609792
1    0.390208
Name: Survived, dtype: float64
```

In [24]:

```
test_one = test
test_one['Survived'] = 0
test_one.Survived[test_one.Sex == 'female'] = 1
print test_one.Survived[:3]
```

```
0    0
1    1
2    0
Name: Survived, dtype: int64
```

C:\Users\liyue\Anaconda2\lib\site-packages\ipykernel__main__.py:3: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy>
app.launch_new_instance()

In [26]:

```
train["Sex"][train["Sex"] == "male"] = 0
train["Sex"][train["Sex"] == "female"] = 1
train["Embarked"] = train["Embarked"].fillna('S')
train["Embarked"][train["Embarked"] == "S"] = 0
train["Embarked"][train["Embarked"] == "C"] = 1
train["Embarked"][train["Embarked"] == "Q"] = 2
```

C:\Users\liyue\Anaconda2\lib\site-packages\ipykernel__main__.py:1: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy>

```
if __name__ == '__main__':
```

C:\Users\liyue\Anaconda2\lib\site-packages\ipykernel__main__.py:2: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy>

```
from ipykernel import kernelapp as app
```

C:\Users\liyue\Anaconda2\lib\site-packages\ipykernel__main__.py:4: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy>

C:\Users\liyue\Anaconda2\lib\site-packages\ipykernel__main__.py:5: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy>

C:\Users\liyue\Anaconda2\lib\site-packages\ipykernel__main__.py:6: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy>

In [27]:

```
train.sort_values(by = ['Age'])[:5]
```

Out[27]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
803	804	1	3	Thomas, Master. Assad Alexander	0	0.42	0	1	2625	8.5
755	756	1	2	Hamalainen, Master. Viljo	0	0.67	1	1	250649	14
644	645	1	3	Baclini, Miss. Eugenie	1	0.75	2	1	2666	19
469	470	1	3	Baclini, Miss. Helene Barbara	1	0.75	2	1	2666	19
78	79	1	2	Caldwell, Master. Alden Gates	0	0.83	0	2	248738	29

In [28]:

```
train.sort_values(by = ['Age'], ascending = False)[:5]
```

Out[28]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	0	80.0	0	0	27042	30
851	852	0	3	Svensson, Mr. Johan	0	74.0	0	0	347060	7.5
493	494	0	1	Artagaveytia, Mr. Ramon	0	71.0	0	0	PC 17609	49
96	97	0	1	Goldschmidt, Mr. George B	0	71.0	0	0	PC 17754	34
116	117	0	3	Connors, Mr. Patrick	0	70.5	0	0	370369	7.5

In [29]:

```
train.sort_values(by = ['Pclass', 'Age'], ascending = False)[:5]
```

Out[29]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
851	852	0	3	Svensson, Mr. Johan	0	74.0	0	0	347060	7.77
116	117	0	3	Connors, Mr. Patrick	0	70.5	0	0	370369	7.75
280	281	0	3	Duane, Mr. Frank	0	65.0	0	0	336439	7.75
483	484	1	3	Turkula, Mrs. (Hedwig)	1	63.0	0	0	4134	9.58
326	327	0	3	Nysveen, Mr. Johan Hansen	0	61.0	0	0	345364	6.23

In [30]:

```
train.groupby(['SibSp']).sum()
```

Out[30]:

	PassengerId	Survived	Pclass	Age	Parch	Fare
SibSp						
0	276865	210	1430	14788.25	113	15620.7530
1	91903	112	430	5506.42	137	9226.8004
2	11548	13	66	565.50	18	1449.1041
3	5145	4	41	167.00	21	1102.5418
4	6869	3	54	127.00	27	573.4000
5	1684	0	15	51.00	10	234.5000
8	3372	0	21	NaN	14	486.8500