In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [3]:
```
train = pd.read_csv('titanic_train.csv')
```

In [4]:
```
train.head()
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Nal
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Nal
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Nal

In [5]:
```
train.isnull()
```

Out[5]:

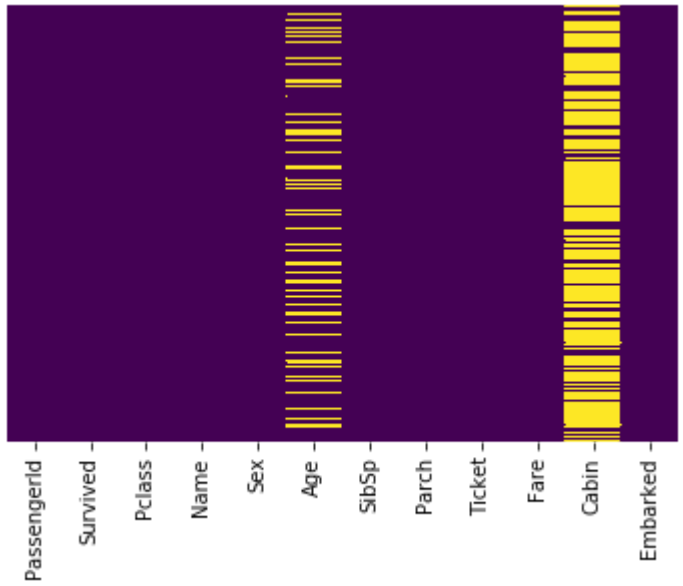
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emb
0	False	False	False	False	False	False	False	False	False	False	True	
1	False	False	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	False	True	
3	False	False	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	False	True	
5	False	False	False	False	False	True	False	False	False	False	True	
6	False	False	False	False	False	False	False	False	False	False	False	
7	False	False	False	False	False	False	False	False	False	False	True	
8	False	False	False	False	False	False	False	False	False	False	True	
9	False	False	False	False	False	False	False	False	False	False	True	
10	False	False	False	False	False	False	False	False	False	False	False	
11	False	False	False	False	False	False	False	False	False	False	False	
12	False	False	False	False	False	False	False	False	False	False	True	
13	False	False	False	False	False	False	False	False	False	False	True	
14	False	False	False	False	False	False	False	False	False	False	True	
15	False	False	False	False	False	False	False	False	False	False	True	
16	False	False	False	False	False	False	False	False	False	False	True	
17	False	False	False	False	False	True	False	False	False	False	True	
18	False	False	False	False	False	False	False	False	False	False	True	
19	False	False	False	False	False	True	False	False	False	False	True	
20	False	False	False	False	False	False	False	False	False	False	True	
21	False	False	False	False	False	False	False	False	False	False	False	
22	False	False	False	False	False	False	False	False	False	False	True	
23	False	False	False	False	False	False	False	False	False	False	False	
24	False	False	False	False	False	False	False	False	False	False	True	
25	False	False	False	False	False	False	False	False	False	False	True	
26	False	False	False	False	False	True	False	False	False	False	True	
27	False	False	False	False	False	False	False	False	False	False	False	
28	False	False	False	False	False	True	False	False	False	False	True	
29	False	False	False	False	False	True	False	False	False	False	True	
...	...	...	...	...	...	...	...	...	...	...	...	
861	False	False	False	False	False	False	False	False	False	False	True	
862	False	False	False	False	False	False	False	False	False	False	False	
863	False	False	False	False	False	True	False	False	False	False	True	

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emb
864	False	False	False	False	False	False	False	False	False	False	True	
865	False	False	False	False	False	False	False	False	False	False	True	
866	False	False	False	False	False	False	False	False	False	False	True	
867	False	False	False	False	False	False	False	False	False	False	False	
868	False	False	False	False	False	True	False	False	False	False	True	
869	False	False	False	False	False	False	False	False	False	False	True	
870	False	False	False	False	False	False	False	False	False	False	True	
871	False	False	False	False	False	False	False	False	False	False	False	
872	False	False	False	False	False	False	False	False	False	False	False	
873	False	False	False	False	False	False	False	False	False	False	True	
874	False	False	False	False	False	False	False	False	False	False	True	
875	False	False	False	False	False	False	False	False	False	False	True	
876	False	False	False	False	False	False	False	False	False	False	True	
877	False	False	False	False	False	False	False	False	False	False	True	
878	False	False	False	False	False	True	False	False	False	False	True	
879	False	False	False	False	False	False	False	False	False	False	False	
880	False	False	False	False	False	False	False	False	False	False	True	
881	False	False	False	False	False	False	False	False	False	False	True	
882	False	False	False	False	False	False	False	False	False	False	True	
883	False	False	False	False	False	False	False	False	False	False	True	
884	False	False	False	False	False	False	False	False	False	False	True	
885	False	False	False	False	False	False	False	False	False	False	True	
886	False	False	False	False	False	False	False	False	False	False	True	
887	False	False	False	False	False	False	False	False	False	False	False	
888	False	False	False	False	False	True	False	False	False	False	True	
889	False	False	False	False	False	False	False	False	False	False	False	
890	False	False	False	False	False	False	False	False	False	False	True	

891 rows × 12 columns

In [6]:
```
sns.heatmap(train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

Out[6]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xa8a2128>
```

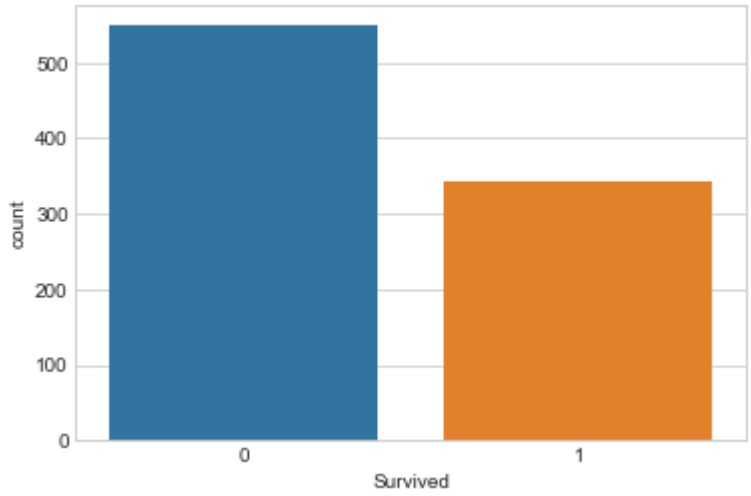


In [7]:
```
#this shows 20% data is missing and cabin also data is missiong
```

In [8]:
```
sns.set_style('whitegrid')
```

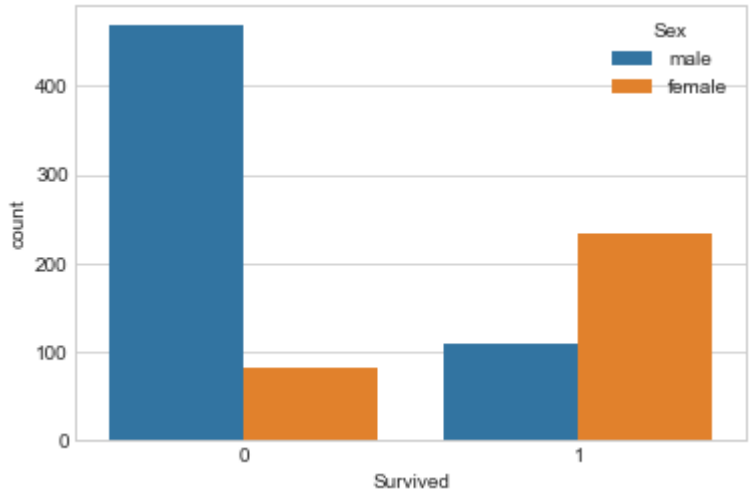
In [9]:
```
sns.countplot(x='Survived',data=train)
```

Out[9]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xac67e10>
```



In [10]:
```
sns.countplot(x='Survived',hue='Sex',data=train)
```

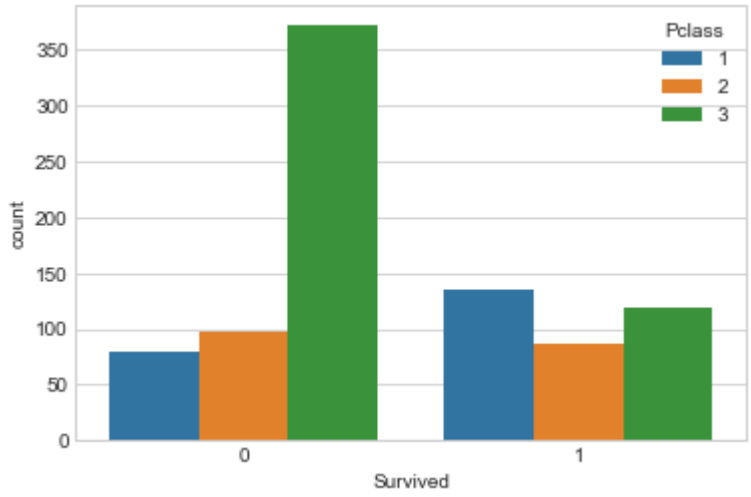
Out[10]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xacd8a58>
```



In [11]:
```
#no of more female survivded
```

In [12]:
```
sns.countplot(x='Survived',hue='Pclass',data=train)
```

Out[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xad40f98>
```

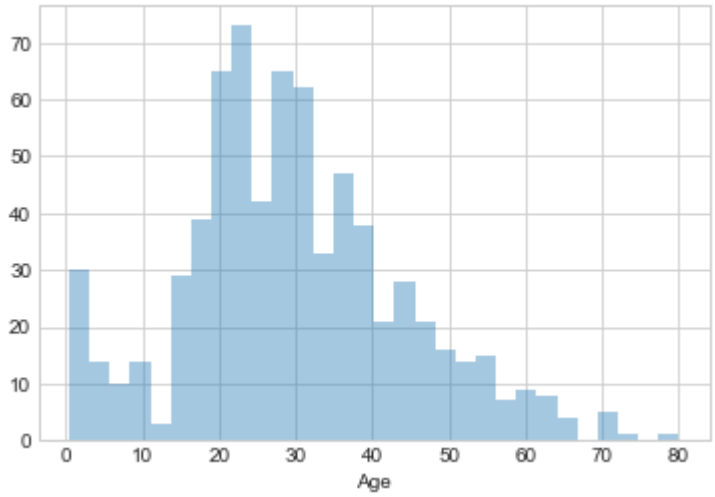


In [13]:
```
# Class 1 survived more
```

In [14]:
```
sns.distplot(train['Age'].dropna(),kde=False,bins=30)
```

```
C:\Users\q21\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWar
ning: The 'normed' kwarg is deprecated, and has been replaced by the 'density'
kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

Out[14]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xacd8c50>
```



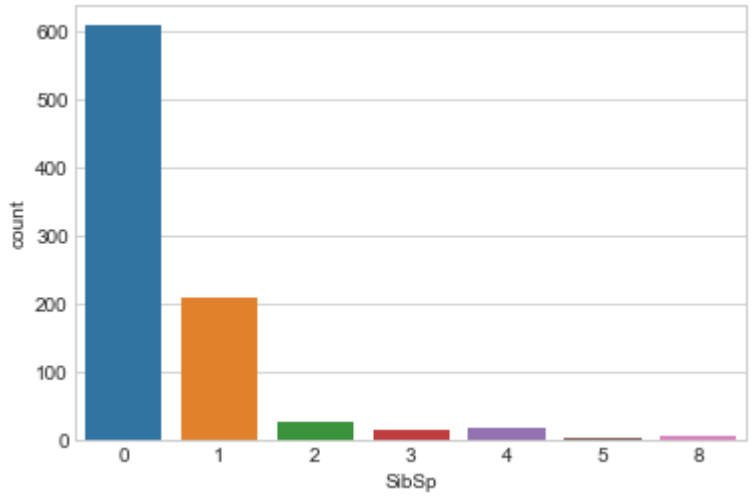
In [15]:
```
#Younger passenger on board more
```

In [16]:
```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

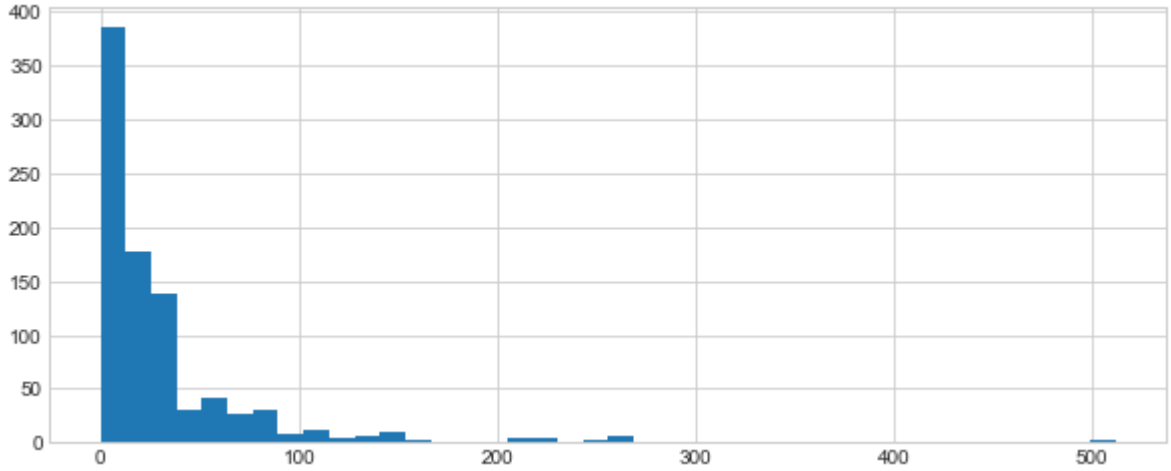
In [17]:
```
sns.countplot(x='SibSp',data=train)
```

Out[17]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xb537320>
```



In [18]:
```
train['Fare'].hist(bins=40,figsize=(10,4))
```

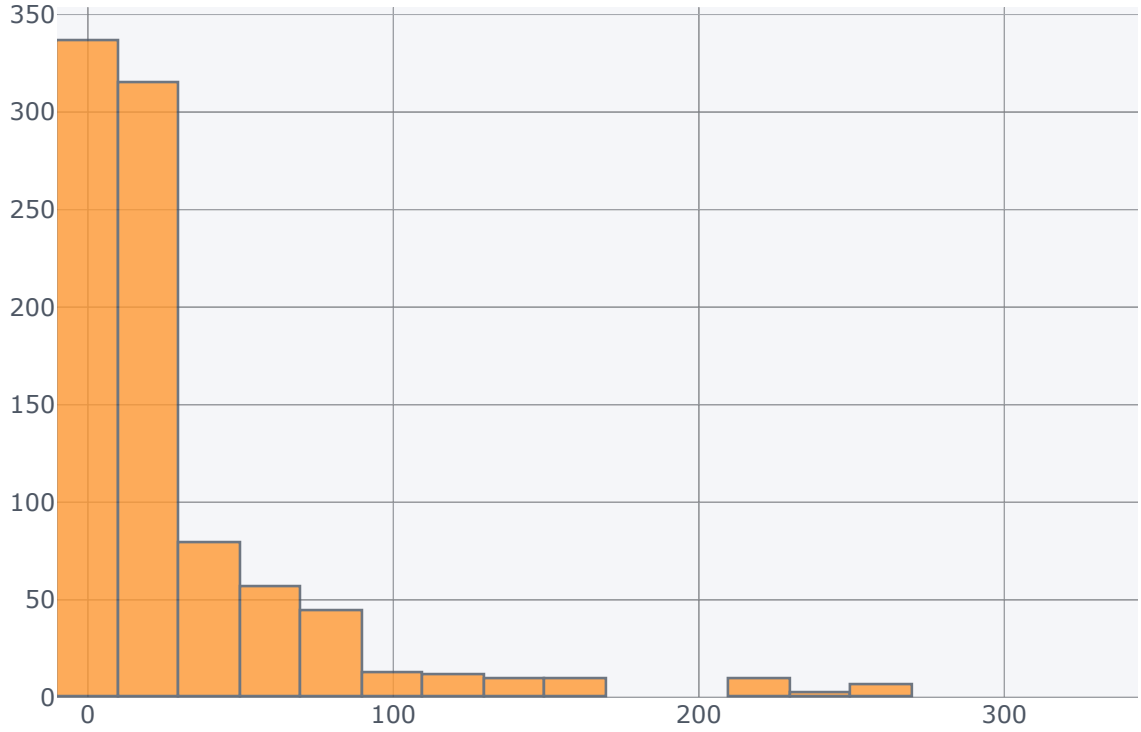
Out[18]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xb5a6048>
```



In [19]:
```
import cufflinks as cf
```

In [20]:
```
cf.go_offline()
```

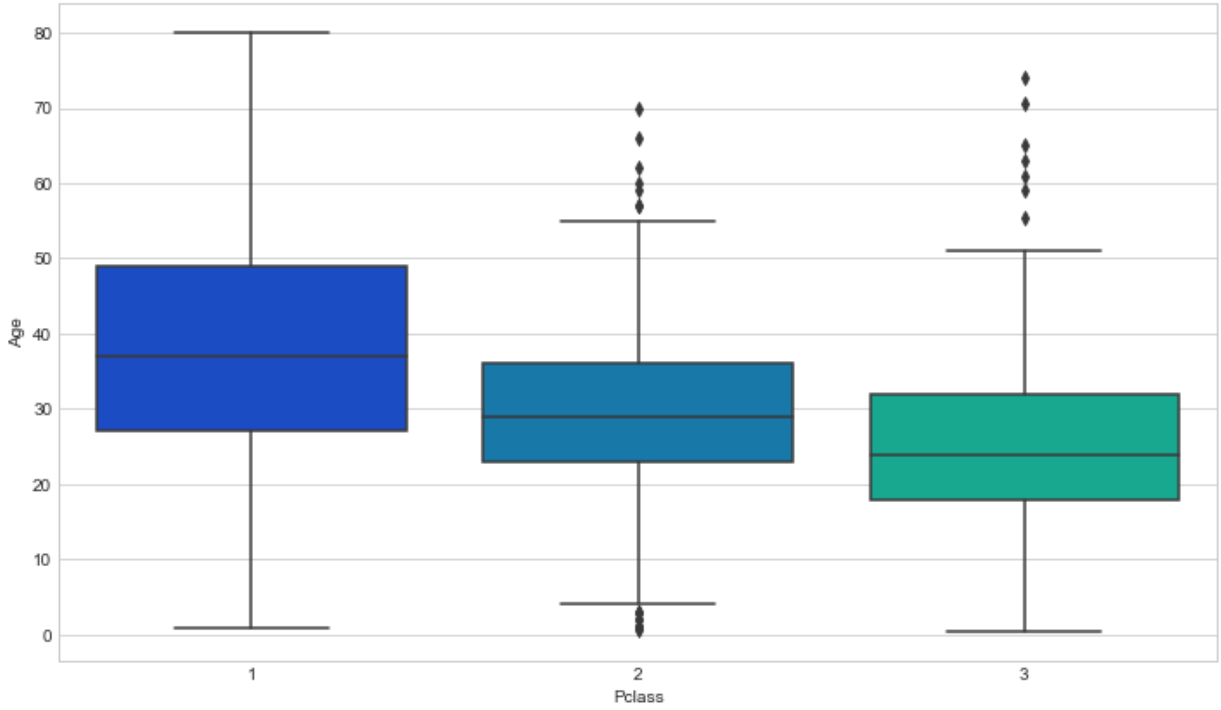
In [21]:
```
train['Fare'].iplot(kind='hist',bins=50)
```



In [22]:
```
#Clean Data
```

In [23]:
```
plt.figure(figsize=(12, 7))
sns.boxplot(x='Pclass',y='Age',data=train,palette='winter')
```

Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x79da278>



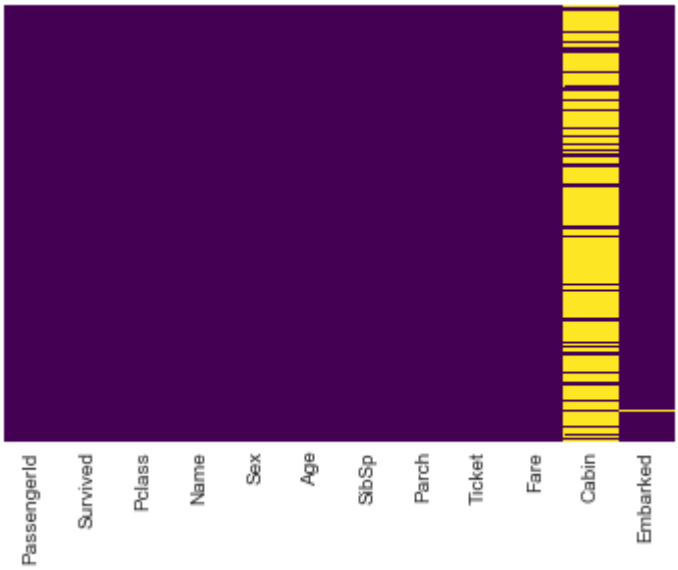
In [24]:
```
#older people travel in 1 & Second Class
```

In [25]:
```
def impute_age(cols):
    Age = cols[0]
    Pclass = cols[1]

    if pd.isnull(Age):

        if Pclass == 1:
            return 37

        elif Pclass == 2:
            return 29

        else:
            return 24

    else:
        return Age
```

In [26]:
```
train['Age'] = train[['Age','Pclass']].apply(impute_age,axis=1)
```

In [27]:
```
sns.heatmap(train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

Out[27]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xc3eb438>
```

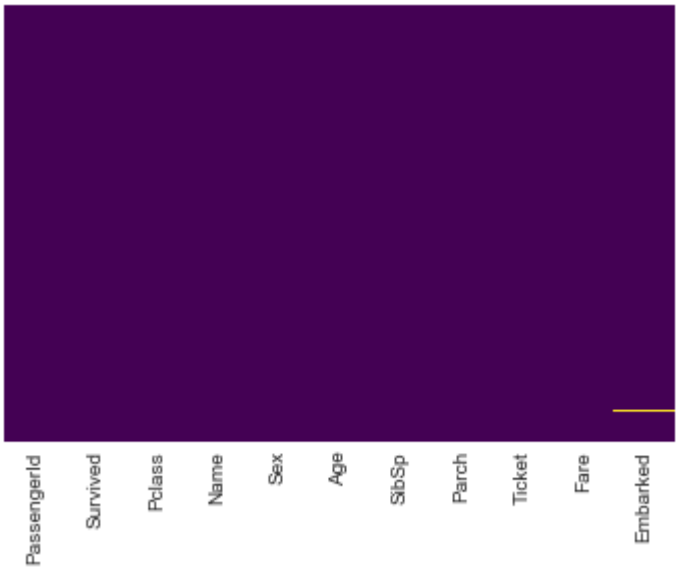


In [28]:
```
# Null values is imputed
```

In [29]:
```
train.drop('Cabin',axis=1,inplace=True)
```

In [30]:
```
sns.heatmap(train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

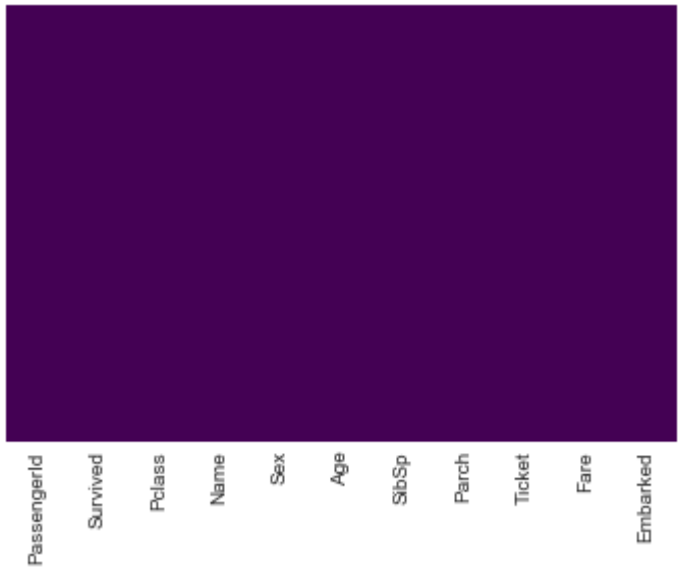
Out[30]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xb646470>
```



In [31]:
```
train.dropna(inplace=True)
```

In [32]:
```
sns.heatmap(train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

Out[32]:
```
<matplotlib.axes._subplots.AxesSubplot at 0xc465320>
```



In [33]:
```
# NO MISSING VALUES
```

In [34]:
```
# now we need to convert categorial value by dummy variable
```

In [35]:
```
pd.get_dummies(train['Sex'])
```

Out[35]:

	female	male
0	0	1
1	1	0
2	1	0
3	1	0
4	0	1
5	0	1
6	0	1
7	0	1
8	1	0
9	1	0
10	1	0
11	1	0
12	0	1
13	0	1
14	1	0
15	1	0
16	0	1
17	0	1
18	1	0
19	1	0
20	0	1
21	0	1
22	1	0
23	0	1
24	1	0
25	1	0
26	0	1
27	0	1
28	1	0
29	0	1
...	...	...
861	0	1
862	1	0
863	1	0

	female	male
864	0	1
865	1	0
866	1	0
867	0	1
868	0	1
869	0	1
870	0	1
871	1	0
872	0	1
873	0	1
874	1	0
875	1	0
876	0	1
877	0	1
878	0	1
879	1	0
880	1	0
881	0	1
882	1	0
883	0	1
884	0	1
885	1	0
886	0	1
887	1	0
888	1	0
889	0	1
890	0	1

889 rows × 2 columns

In [36]:
```
# Issue of Multi collinearity
```

In [37]:
```
pd.get_dummies(train['Sex'],drop_first=True)
```

Out[37]:

	male
0	1
1	0
2	0
3	0
4	1
5	1
6	1
7	1
8	0
9	0
10	0
11	0
12	1
13	1
14	0
15	0
16	1
17	1
18	0
19	0
20	1
21	1
22	0
23	1
24	0
25	0
26	1
27	1
28	0
29	1
...	...
861	1
862	0
863	0

	male
864	1
865	0
866	0
867	1
868	1
869	1
870	1
871	0
872	1
873	1
874	0
875	0
876	1
877	1
878	1
879	0
880	0
881	1
882	0
883	1
884	1
885	0
886	1
887	0
888	0
889	1
890	1

889 rows × 1 columns

In [38]:
```
sex = pd.get_dummies(train['Sex'],drop_first=True)
embark = pd.get_dummies(train['Embarked'],drop_first=True)
```

In [39]:
```
train.drop(['Sex','Embarked','Name','Ticket'],axis=1,inplace=True)
```

In [40]:
```
train = pd.concat([train,sex,embark],axis=1)
```

In [41]:
```
train.head()
```

Out[41]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	male	Q	S
0	1	0	3	22.0	1	0	7.2500	1	0	1
1	2	1	1	38.0	1	0	71.2833	0	0	0
2	3	1	3	26.0	0	0	7.9250	0	0	1
3	4	1	1	35.0	1	0	53.1000	0	0	1
4	5	0	3	35.0	0	0	8.0500	1	0	1

In [43]:
```
train.drop(['PassengerId'],axis=1,inplace=True)
```

In [44]:
```
train.head()
```

Out[44]:

	Survived	Pclass	Age	SibSp	Parch	Fare	male	Q	S
0	0	3	22.0	1	0	7.2500	1	0	1
1	1	1	38.0	1	0	71.2833	0	0	0
2	1	3	26.0	0	0	7.9250	0	0	1
3	1	1	35.0	1	0	53.1000	0	0	1
4	0	3	35.0	0	0	8.0500	1	0	1

In [45]:
```
#Train and Test Data
```

In [46]:
```
X =train.drop('Survived',axis=1)
```

In [47]:
```
y= train['Survived']
```

In [48]:
```
from sklearn.cross_validation import train_test_split
```

In [49]:
```
X_train, X_test, y_train, y_test = train_test_split(train.drop('Survived',axis=1)
                                                    train['Survived'], test_size=
                                                    random_state=101)
```

In [50]:
```
#Training and Predict by model
```

In [51]:
```
from sklearn.linear_model import LogisticRegression
```

In [52]:
```
logmodel =LogisticRegression()
```

In [53]:
```
logmodel.fit(X_train,y_train)
```

Out[53]:
```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

In [54]:
```
predictions = logmodel.predict(X_test)
```

In [55]:
```
# classification task
```

In [56]:
```
from sklearn.metrics import classification_report
```

In [57]:
```
print(classification_report(y_test,predictions))
```

```
             precision    recall  f1-score   support

          0       0.80      0.91      0.85       163
          1       0.82      0.65      0.73       104

avg / total       0.81      0.81      0.80       267
```

In [58]:
```
from sklearn.metrics import confusion_matrix
```

In [59]:
```
confusion_matrix(y_test,predictions)
```

Out[59]:
```
array([[148,  15],
       [ 36,  68]], dtype=int64)
```

In [ ]:
```

```