In [46]:
```
# Nmae : Himanshu Agarwal , Net id: HXA180027 ,
#q2.1 Titanic dataset
#mporting libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

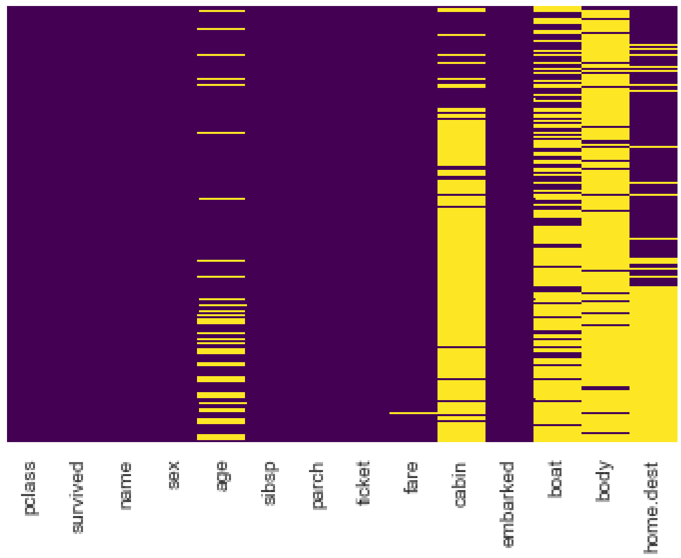
In [47]:
```
#import data using the pandas libraries.
titanic3_dataset = pd.read_csv('titanic3.csv')
titanic3_dataset.head()
```

Out[47]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
0	1	1	Allen, Miss. Elisabeth Walton	female	29.00	0	0	24160	211.3375	B5	S
1	1	1	Allison, Master. Hudson Trevor	male	0.92	1	2	113781	151.5500	C22 C26	S
2	1	0	Allison, Miss. Helen Loraine	female	2.00	1	2	113781	151.5500	C22 C26	S
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.00	1	2	113781	151.5500	C22 C26	S
4	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.00	1	2	113781	151.5500	C22 C26	S

In [48]:
```
#Heatmap to check the missig values
sns.heatmap(titanic3_dataset.isnull(),yticklabels=False,cbar=False,cmap='virid
is')
```

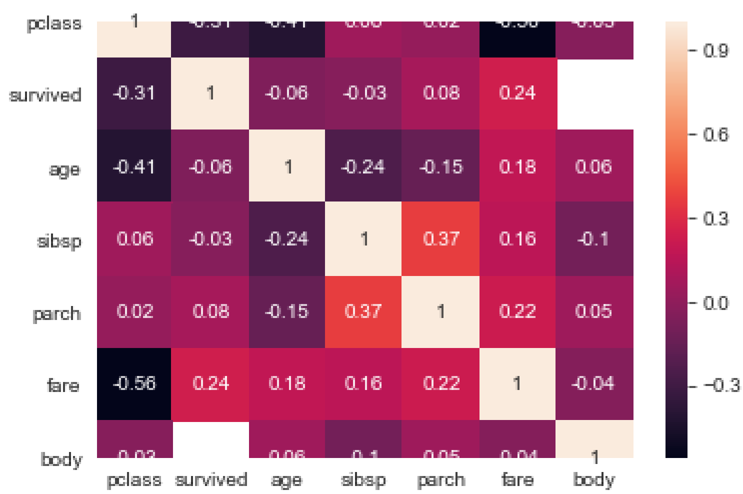
Out[48]: <matplotlib.axes._subplots.AxesSubplot at 0x1cb63676988>



In [ ]:

In [49]:
```
# create a correlation matrix that measures the linear relationships between t
he variables.
#The correlation matrix can be formed by using the corr function from the pand
as dataframe library.
#We will use the heatmap function from the seaborn library to plot the correla
tion matrix
correlation_matrix = titanic3_dataset.corr().round(2)
sns.heatmap(data=correlation_matrix, annot=True)
```

Out[49]: <matplotlib.axes._subplots.AxesSubplot at 0x1cb630479c8>

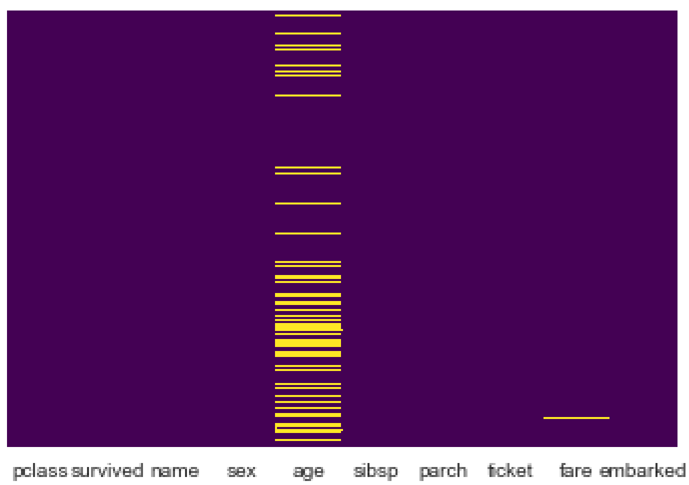


In [50]:
```
#Through Correlation matrix we can see that few columns are not present which
 shows they are independent and has no correlation.
#We can drop those features as they will not play any role in deciding the tar
get feature.
#So dropping features 'Cabin','body','boat', and 'home.dest'.Same way name and
ticket.
```

In [51]:
```
titanic3_dataset.drop('cabin',axis=1,inplace=True)
titanic3_dataset.drop('body',axis=1,inplace=True)
titanic3_dataset.drop('boat',axis=1,inplace=True)
titanic3_dataset.drop('home.dest',axis=1,inplace=True)
```

In [52]:
```
#Heatmap to check the missig values
sns.heatmap(titanic3_dataset.isnull(),yticklabels=False,cbar=False,cmap='virid
is')
```

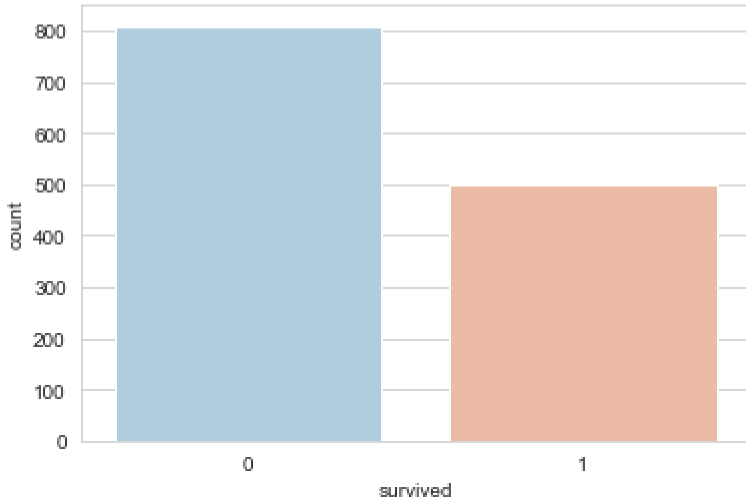
Out[52]: <matplotlib.axes._subplots.AxesSubplot at 0x1cb6315efc8>



In [53]:
```
#Roughly 20 percent of the Age data is missing. The proportion of Age missing
 is likely small enough for
#reasonable replacement with some form of imputation.
```

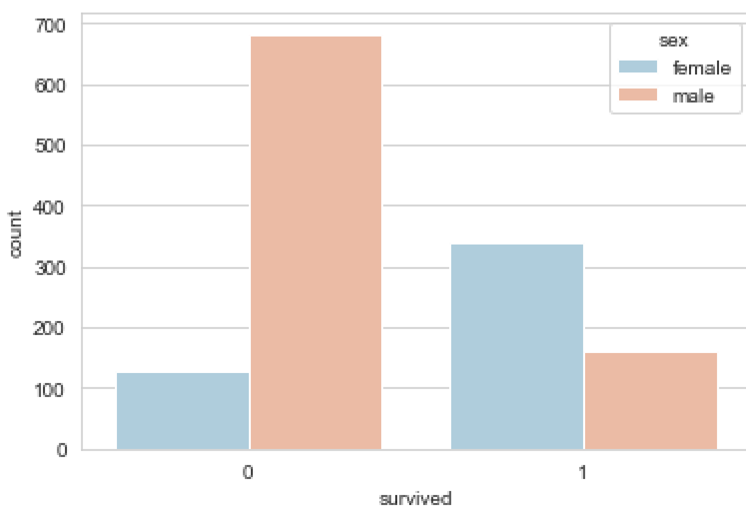
In [54]:
```
#Data Visualization
#using count plot checking how many passangers survived
sns.set_style('whitegrid')
sns.countplot(x='survived',data=titanic3_dataset,palette='RdBu_r')
```

Out[54]: <matplotlib.axes._subplots.AxesSubplot at 0x1cb631a7848>



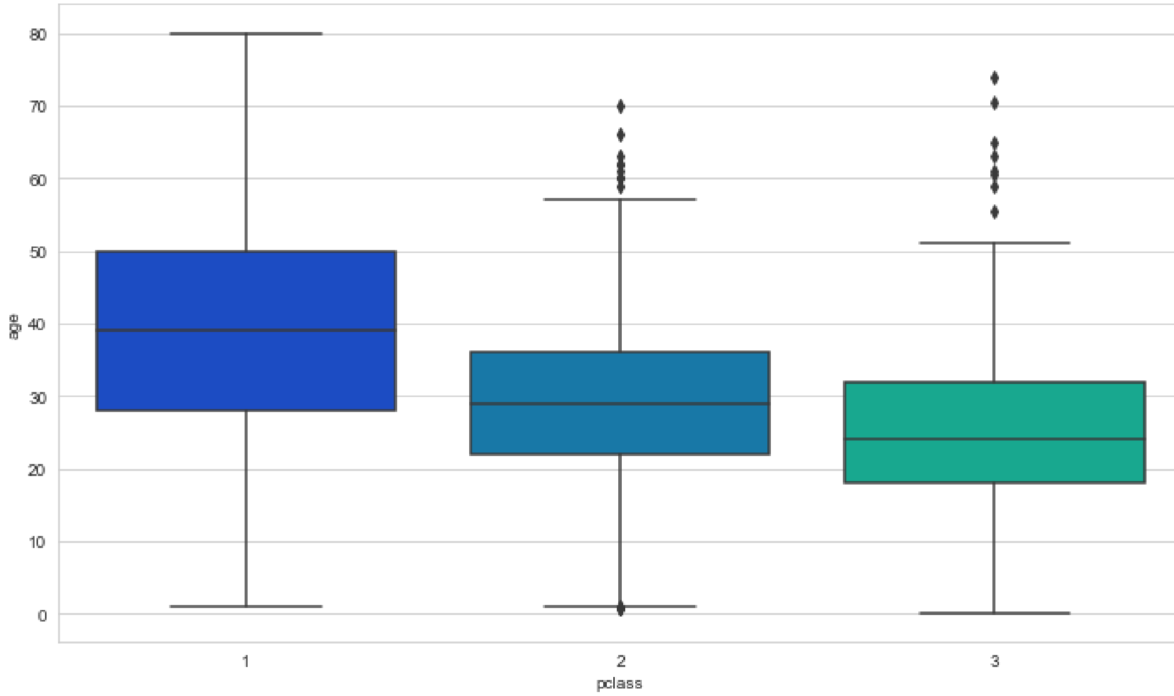
In [55]:
```
#using count plot checking how many passangers of different genders survived
sns.set_style('whitegrid')
sns.countplot(x='survived',hue='sex',data=titanic3_dataset,palette='RdBu_r')
```

Out[55]: <matplotlib.axes._subplots.AxesSubplot at 0x1cb631ff1c8>



In [56]:
```python
#Data cleaning
#using box plot checking class wise average age of passangers
plt.figure(figsize=(12, 7))
sns.boxplot(x='pclass',y='age',data=titanic3_dataset,palette='winter')
```

Out[56]: <matplotlib.axes._subplots.AxesSubplot at 0x1cb63243408>



In [57]:
```python
#fill in missing age data instead of just dropping the missing age data rows
#One way to do this is by filling in the mean age of all the passengers (imput
ation).
#However we can be smarter about this and check the average age by passenger c
lass.
```

In [58]:
```python
#Method to get the average age for a class
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 [59]:
```
#imputing the missing age values for each class.
titanic3_dataset['age'] = titanic3_dataset[['age','pclass']].apply(impute_age,
axis=1)

#to see the count of the missing values left in eachh column
titanic3_dataset.isnull().sum()
```

Out[59]:
```
pclass      0
survived    0
name        0
sex         0
age         0
sibsp       0
parch       0
ticket      0
fare        1
embarked    2
dtype: int64
```

In [60]:
```
#As the number of rows with null data for fare and embarked are too low as com
pared to total size of the dataset, so we can drop them.
titanic3_dataset.dropna(inplace=True)
```

In [61]:
```
#We’ll need to convert categorical features to dummy variables using pandas! L
ike 'sex' and 'embarked' features.
#Otherwise our machine learning algorithm won’t be able to directly take in th
ose features as inputs.
#the 'Name' and 'Ticket' column have no relationship with whether the person s
urvived or not,observed using correlation matrix.
#So we drop these 2 columns and we convert the other two columns into numerica
l values
```

In [62]:
```
sex = pd.get_dummies(titanic3_dataset['sex'],drop_first=True)
embark = pd.get_dummies(titanic3_dataset['embarked'],drop_first=True)
titanic3_dataset.drop(['sex','embarked','name','ticket'],axis=1,inplace=True)
```

In [63]:
```
#Data is ready to apply Logistic regression
titanic3_dataset.head()
```

Out[63]:

	pclass	survived	age	sibsp	parch	fare
0	1	1	29.00	0	0	211.3375
1	1	1	0.92	1	2	151.5500
2	1	0	2.00	1	2	151.5500
3	1	0	30.00	1	2	151.5500
4	1	0	25.00	1	2	151.5500

In [ ]:

In [64]:
```
#Preprocessing of Data for logistic regression
```

In [65]:
```
#spliting data in training and test set
#train the model with 80% of the samples and test with the remaining 20%.
n = int(len(titanic3_dataset)*0.80)
df_train, df_test = titanic3_dataset.iloc[:n, :], titanic3_dataset.iloc[n:, :]
titanic3_dataset.columns
```

Out[65]:
```
Index(['pclass', 'survived', 'age', 'sibsp', 'parch', 'fare'], dtype='objec
t')
```

In [66]:
```
# Initial Coefficients
B = np.array([0, 0, 0, 0, 0, 0]) #Weights array
alpha = 0.0001 # Learning rate
```

In [67]:
```
#Spliting the training and testing data in X,Y train and test sets.
dt = df_train
m = len(df_train.iloc[:,:-1])
x0 = np.ones(m)
Xtrain = np.array([x0, dt['pclass'],dt['age'],dt['sibsp'],dt['parch'],dt['far
e']]).T
ytrain = np.array(dt['survived'])

m = len(df_test.iloc[:,:-1])
x0 = np.ones(m)
dt= df_test
Xtest= np.array([x0, dt['pclass'],dt['age'],dt['sibsp'],dt['parch'],dt['fare'
]]).T
ytest= np.array(dt['survived'])
```

In [68]:
```
#Cost function
def cost_function(X, Y, B):
    m = len(Y)
    J = np.sum((X.dot(B) - Y) ** 2)/(2 * m)
    return J
```

In [69]:
```
#Sigmoid fuction to get the result in 0-1 range
def sigmoid(z):
        return 1 / (1 + np.exp(-z))
```

In [70]:
```
#mplementation of Logistic regression for the given numbers of iterations.
def gradient_descent(X, Y, B, alpha, iterations):
    cost_history = [0] * iterations
    m = len(Y)

    for iteration in range(iterations):
        z = X.dot(B)
        h = sigmoid(z)  #Calcultating the sigmoid of the  hypothesis
        # Difference b/w Hypothesis and Actual Y
        loss = h - Y
        # Gradient Calculation
        gradient = X.T.dot(loss) / m
        # Changing Values of B using Gradient
        B = B - alpha * gradient
        # New Cost Value
        cost = cost_function(X, Y, B)
        cost_history[iteration] = cost

    return B, cost_history
```

In [71]:
```
alpha= 0.001 #learning rate
iterations=10000
theta , cost = gradient_descent(Xtrain,ytrain, B, alpha, iterations)
```

In [72]:
```
#Method to predict the value of output feature based on the weights received a
fter training.
def predict(X, theta):
        return sigmoid(np.dot(X, theta)).round()
```

In [73]:
```
#using scikit-learn’s Libraries to get the reports.
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
```

In [74]:
```
predictions_train = predict(Xtrain,theta) # Making prediction on training set
print("Classification Report on Training data")
print(classification_report(ytrain,predictions_train))
print("Accuracy on Training data:",accuracy_score(ytrain, predictions_train))
```

```
Classification Report on Training data
              precision    recall  f1-score   support

           0       0.65      0.90      0.75       601
           1       0.71      0.33      0.45       443

    accuracy                           0.66      1044
   macro avg       0.68      0.62      0.60      1044
weighted avg       0.67      0.66      0.63      1044

Accuracy on Training data: 0.6590038314176245
```

In [75]:
```
predictions_test = predict(Xtest,theta) # Making prediction on test set
print("Classification Report on test data")
print(classification_report(ytest,predictions_test))
print("Accuracy on test data:",accuracy_score(ytest, predictions_test))
```

```
Classification Report on test data
              precision    recall  f1-score   support

           0       0.79      0.98      0.87       207
           1       0.00      0.00      0.00        55

    accuracy                           0.77       262
   macro avg       0.39      0.49      0.44       262
weighted avg       0.62      0.77      0.69       262

Accuracy on test data: 0.7709923664122137
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: