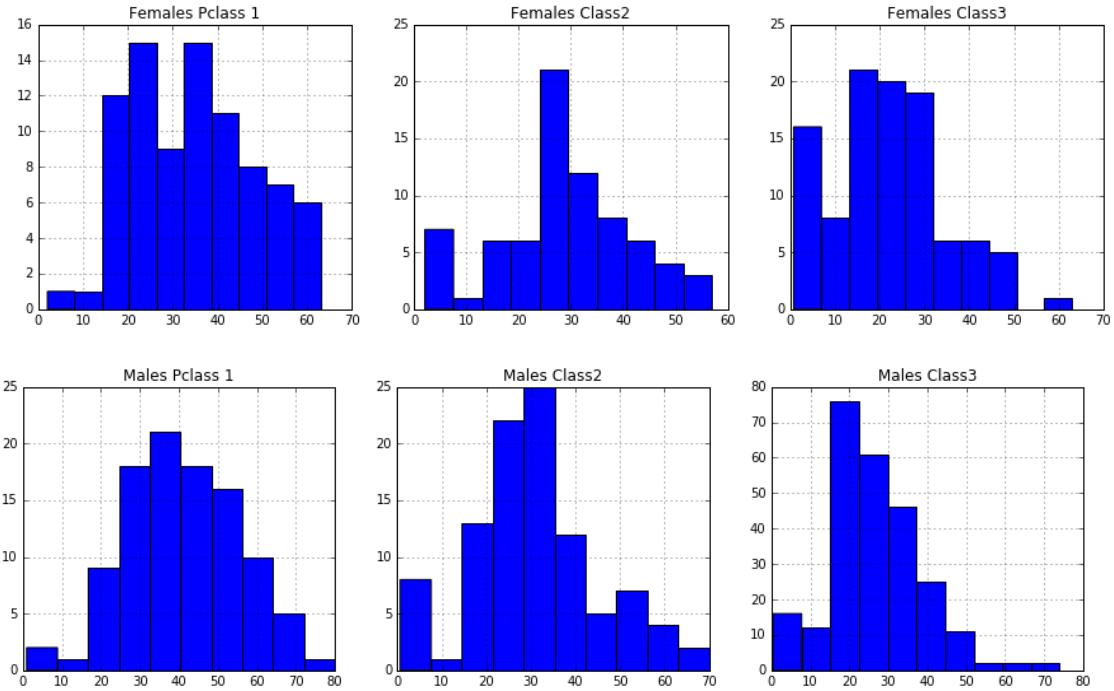
1. Imputing missing data in Age

1. The histograms below show that there is a relationship between age and Sex (older males than females) + PClass (socio economic status i.e. higher the class in which someone is travelling they are probably going to be older).



2. As a result each combination of sex (Male, Female) and Pclass (1,2,3) was treated as a separate group.
3. For each group, missing values of age was substituted by random numbers generated between (**mean+ standard deviation)** and (**mean – standard deviation)** for that group.

2. Create & Run Logistic Regression to predict survival of passengers

1. Converted some of the columns that were strings to integers.
    a. Sex : males > 0, female > 1
    b. Embarked: S > 0, C > 1, Q > 2
2. Used the following columns as features: "Pclass", "Sex", "Age", "SibSp", "Parch", "Fare", "Embarked". Used the “Survived” column as the target.
3. Model Output
    a. Here are the results of Logistic Regression:

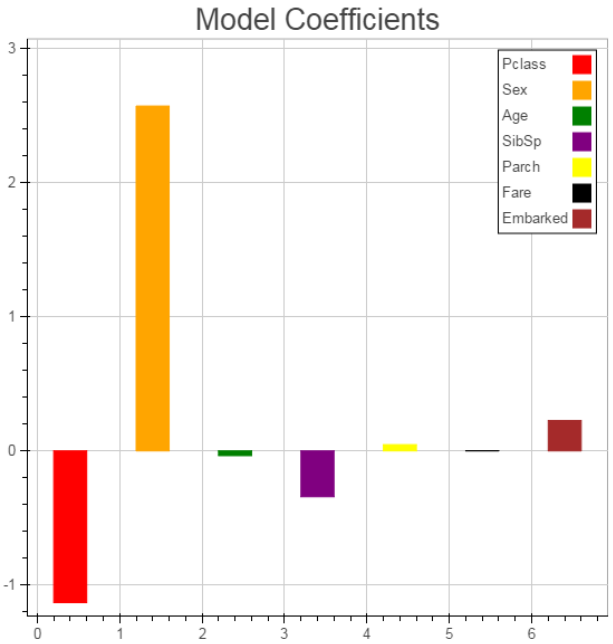
```
Results of Logistic Regression:

                Predicted Class 0  Predicted Class 1
Actual Class 0                 93                 13
Actual Class 1                 22                 51

Precision: 0.796875
Recall: 0.698630136986
```

b. Coefficient values for the different features:

	features	coeff	abs
1	Sex	2.568949	2.568949
0	Pclass	-1.130138	1.130138
3	SibSp	-0.341577	0.341577
6	Embarked	0.226091	0.226091
4	Parch	0.046608	0.046608
2	Age	-0.035362	0.035362
5	Fare	0.001794	0.001794



4. Features that are predictive for this logistic regression:
   a. Of all the features I initially picked the 7 features listed in the above graph. I did not include PassengerID, Name, Ticket and Cabin since intuitively they don't seem to be features that would help predict whether a passenger survived or not.
   b. From the graph above since the "Sex" feature has the highest positive coefficient value, followed by "Pclass". I would use these columns to predict the survival of a passenger.

i. Intuitively it does make sense to use “Sex” and “Pclass”. “Sex” feature does help predict whether someone survived the Titanic. Because they must have tried rescuing the females before males. “Pclass” represents the socio economic status (1 – Upper, 2 – Middle, 3 – Lower) which must have been a contributing factor to who survived. Probability of a upper class person surviving must have been higher than someone from a lower class.

3. Implement Cross-Validation

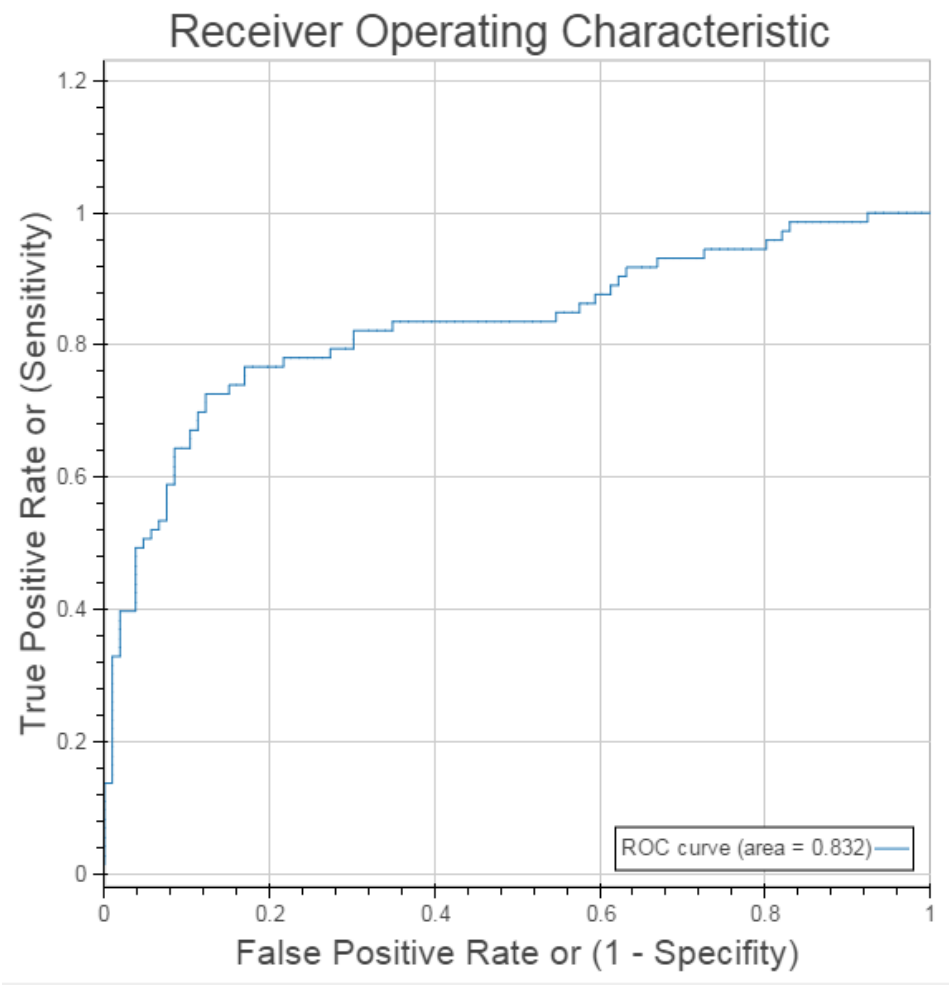
Ran the cross_val_score function for 3, 5, 10, 20 and 50 folds. The results are shown below:

```
[(3, 0.7957351290684623),
 (5, 0.7957734041613781),
 (10, 0.80027011689933025),
 (20, 0.80129227053140095),
 (50, 0.79611928104575158)]
```

So picked the 10 fold cross-validation as it has the most accuracy.

4. Create ROC Curve

ROC Curve plotted visually below:



AUC for the model is 0.832

Potential next steps to improve accuracy: Use the features with the top 3 highest coefficients only.

I would use a threshold of 0.298. Since this threshold has a high TPR of 0.79 and low FPR of 0.29.