In [2]:

```
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
import sklearn
from pandas import Series, DataFrame
from pylab import rcParams
from sklearn import preprocessing
from sklearn.linear_model import LogisticRegression
#from sklearn.cross_validation import train_test_split -- note deprecation war
ning below
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.metrics import classification_report
from sklearn import tree, metrics, model_selection, preprocessing
from IPython.display import Image, display
url="https://raw.githubusercontent.com/BigDataGal/Python-for-Data-Science/mast
er/titanic-train.csv"
titanic = pd.read_csv(url)
titanic.columns =['PassengerId','Survived','Pclass','Name','Sex','Age','SibSp'
,'Parch','Ticket','Fare','Cabin','Embarked']
```

In [3]:

```
titanic.shape
```

Out[3]:

```
(891, 12)
```

In [4]:

```
titanic.columns
```

Out[4]:

```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [5]:

```
titanic.index
```

Out[5]:

```
RangeIndex(start=0, stop=891, step=1)
```

In [6]:
```
titanic.describe()
```

Out[6]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.00
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.0000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.9104
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.32

In [7]:
```
titanic.head(5)
```

Out[7]:

	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	7.25
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05

In [8]:
```
titanic.isnull().sum(axis=0)
```

Out[8]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In [9]:
```
titanic = titanic.dropna()
```

In [10]:

```
titanic.head(20)
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.
21	22	1	2	Beesley, Mr. Lawrence	male	34.0	0	0	248698	13.
23	24	1	1	Sloper, Mr. William Thompson	male	28.0	0	0	113788	35.
27	28	0	1	Fortune, Mr. Charles Alexander	male	19.0	3	2	19950	26:
52	53	1	1	Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49.0	1	0	PC 17572	76.
54	55	0	1	Ostby, Mr. Engelhart Cornelius	male	65.0	0	1	113509	61.
62	63	0	1	Harris, Mr. Henry Birkhardt	male	45.0	1	0	36973	83.
66	67	1	2	Nye, Mrs. (Elizabeth Ramell)	female	29.0	0	0	C.A. 29395	10.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
75	76	0	3	Moen, Mr. Sigurd Hansen	male	25.0	0	0	348123	7.6
88	89	1	1	Fortune, Miss. Mabel Helen	female	23.0	3	2	19950	26
92	93	0	1	Chaffee, Mr. Herbert Fuller	male	46.0	1	0	W.E.P. 5734	61.
96	97	0	1	Goldschmidt, Mr. George B	male	71.0	0	0	PC 17754	34.
97	98	1	1	Greenfield, Mr. William Bertram	male	23.0	0	1	PC 17759	63.
102	103	0	1	White, Mr. Richard Frasar	male	21.0	0	1	35281	77.
110	111	0	1	Porter, Mr. Walter Chamberlain	male	47.0	0	0	110465	52.
118	119	0	1	Baxter, Mr. Quigg Edmond	male	24.0	0	1	PC 17558	24

In [11]:

```
titanic.isnull().sum(axis=0)
```

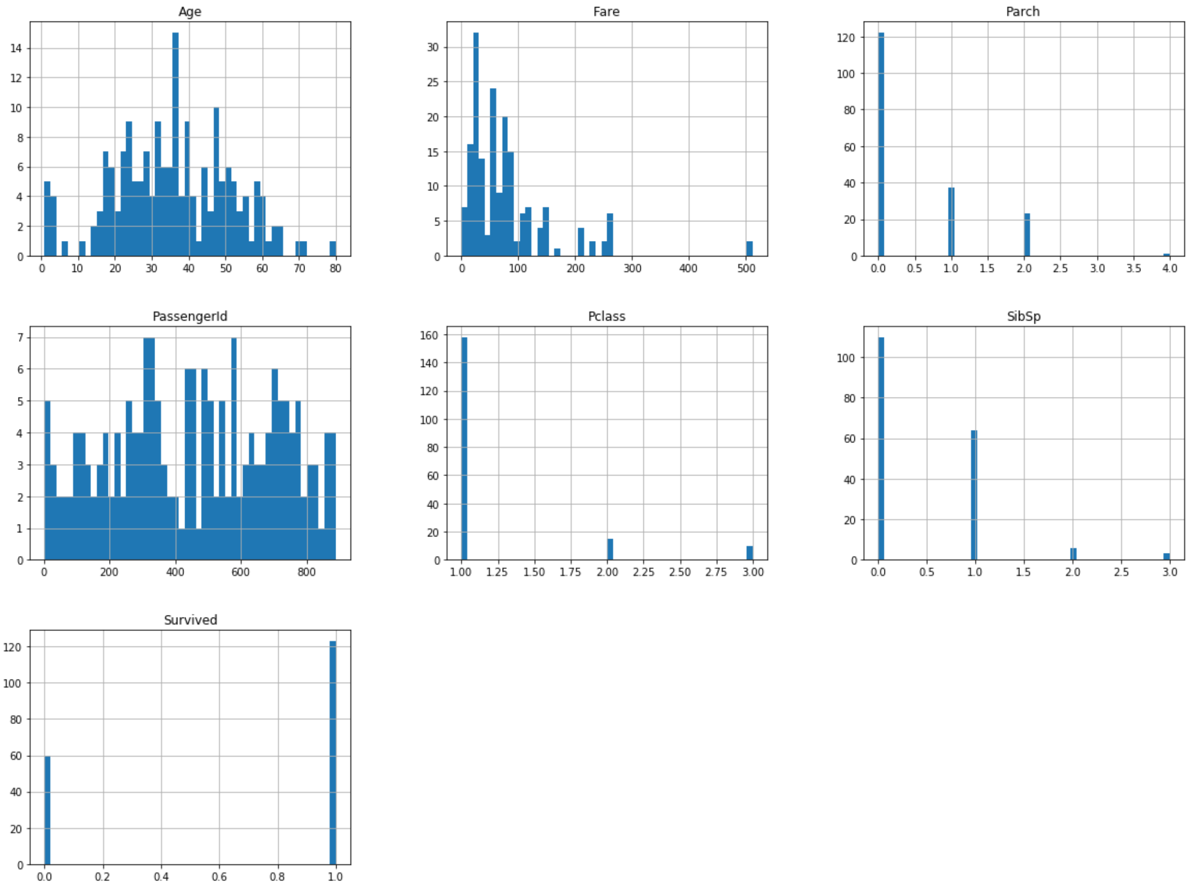
Out[11]:

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```

```
In [12]: %matplotlib inline
         import matplotlib.pyplot as plt
         titanic.hist(bins=50, figsize=(20,15))
         save_fig("attribute_histogram_plots")
         plt.show()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-12-d1a828756ee6> in <module>()
      2 import matplotlib.pyplot as plt
      3 titanic.hist(bins=50, figsize=(20,15))
----> 4 save_fig("attribute_histogram_plots")
      5 plt.show()

NameError: name 'save_fig' is not defined
```



In [13]:

```
titanic['Survived'] = titanic['Survived'].astype('int')
titanic.corr(method='pearson',min_periods=1).transpose().sort_values('Survive
d', ascending=False)
```

Out[13]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fa
Survived	0.148495	1.000000	-0.034542	-0.254085	0.106346	0.023582	0.13424
PassengerId	1.000000	0.148495	-0.089136	0.030933	-0.083488	-0.051454	0.02974
Fare	0.029740	0.134241	-0.315235	-0.092424	0.286433	0.389740	1.00000
SibSp	-0.083488	0.106346	-0.103592	-0.156162	1.000000	0.255346	0.28643
Parch	-0.051454	0.023582	0.047496	-0.271271	0.255346	1.000000	0.38974
Pclass	-0.089136	-0.034542	1.000000	-0.306514	-0.103592	0.047496	-0.3152
Age	0.030933	-0.254085	-0.306514	1.000000	-0.156162	-0.271271	-0.0924

In [14]:

```
titanic.head(15)
```

Out[14]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.10
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.86
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.70
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.55
21	22	1	2	Beesley, Mr. Lawrence	male	34.0	0	0	248698	13.00
23	24	1	1	Sloper, Mr. William Thompson	male	28.0	0	0	113788	35.50
27	28	0	1	Fortune, Mr. Charles Alexander	male	19.0	3	2	19950	263.0
52	53	1	1	Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49.0	1	0	PC 17572	76.72
54	55	0	1	Ostby, Mr. Engelhart Cornelius	male	65.0	0	1	113509	61.97
62	63	0	1	Harris, Mr. Henry Birkhardt	male	45.0	1	0	36973	83.47

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
66	67	1	2	Nye, Mrs. (Elizabeth Ramell)	female	29.0	0	0	C.A. 29395	10.50
75	76	0	3	Moen, Mr. Sigurd Hansen	male	25.0	0	0	348123	7.650
88	89	1	1	Fortune, Miss. Mabel Helen	female	23.0	3	2	19950	263.0
92	93	0	1	Chaffee, Mr. Herbert Fuller	male	46.0	1	0	W.E.P. 5734	61.17

In [15]:
```
titanic_gender = titanic.join(pd.get_dummies(titanic.Sex))
titanic_gender.head(5)
```

Out[15]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.10
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.86
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.70
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.55

In [16]:

```
titanic_gender['Survived'] = titanic_gender['Survived'].astype('int')
titanic_gender.corr(method='pearson',min_periods=1).transpose().sort_values('S
urvived', ascending=False)
```

Out[16]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fa
Survived	0.148495	1.000000	-0.034542	-0.254085	0.106346	0.023582	0.13424
female	0.025205	0.532418	0.046181	-0.184969	0.104291	0.089581	0.13043
PassengerId	1.000000	0.148495	-0.089136	0.030933	-0.083488	-0.051454	0.02974
Fare	0.029740	0.134241	-0.315235	-0.092424	0.286433	0.389740	1.00000
SibSp	-0.083488	0.106346	-0.103592	-0.156162	1.000000	0.255346	0.28643
Parch	-0.051454	0.023582	0.047496	-0.271271	0.255346	1.000000	0.38974
Pclass	-0.089136	-0.034542	1.000000	-0.306514	-0.103592	0.047496	-0.3152
Age	0.030933	-0.254085	-0.306514	1.000000	-0.156162	-0.271271	-0.0924
male	-0.025205	-0.532418	-0.046181	0.184969	-0.104291	-0.089581	-0.1304

In [17]:

```
y = titanic_gender['Survived']
X = titanic_gender[['Pclass', 'Age', 'SibSp', 'Parch', 'Fare', 'male', 'femal
e']]
```

In [18]:

```
X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y, test
_size=0.3, random_state=0)
```

In [19]:

```
dtree = tree.DecisionTreeClassifier(criterion='entropy', max_depth=3, random_s
tate=0)
dtree.fit(X_train, y_train)
```

Out[19]:

```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=3,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best')
```

In [20]:

```
y_pred = dtree.predict(X_test)
```

In [21]:

```
count_misclassified = (y_test != y_pred).sum()
print('Misclassified samples: {}'.format(count_misclassified))
accuracy = metrics.accuracy_score(y_test, y_pred)
print('Accuracy: {:.2f}'.format(accuracy))
```

```
Misclassified samples: 11
Accuracy: 0.80
```

In [22]:

```
from sklearn.model_selection import GridSearchCV
param_test1 = {
 'max_depth': range(2, 5),
 'min_samples_split': [2, 3, 5],
 'min_samples_leaf': [1, 2, 3]
}

grid_result = GridSearchCV(dtree, param_grid=param_test1, cv=10, n_jobs=-1, ve
rbose=1)
grid_result.fit(X_train, y_train)

print("Best: %f using %s" % (grid_result.best_score_, grid_result.best_params_
))
```

```
Fitting 10 folds for each of 27 candidates, totalling 270 fits

[Parallel(n_jobs=-1)]: Done  42 tasks      | elapsed:   46.6s
[Parallel(n_jobs=-1)]: Done 263 out of 270 | elapsed:   52.9s remaining:
1.3s
[Parallel(n_jobs=-1)]: Done 270 out of 270 | elapsed:   53.0s finished

Best: 0.757812 using {'max_depth': 4, 'min_samples_leaf': 1, 'min_samples_spl
it': 2}
```

In [23]:

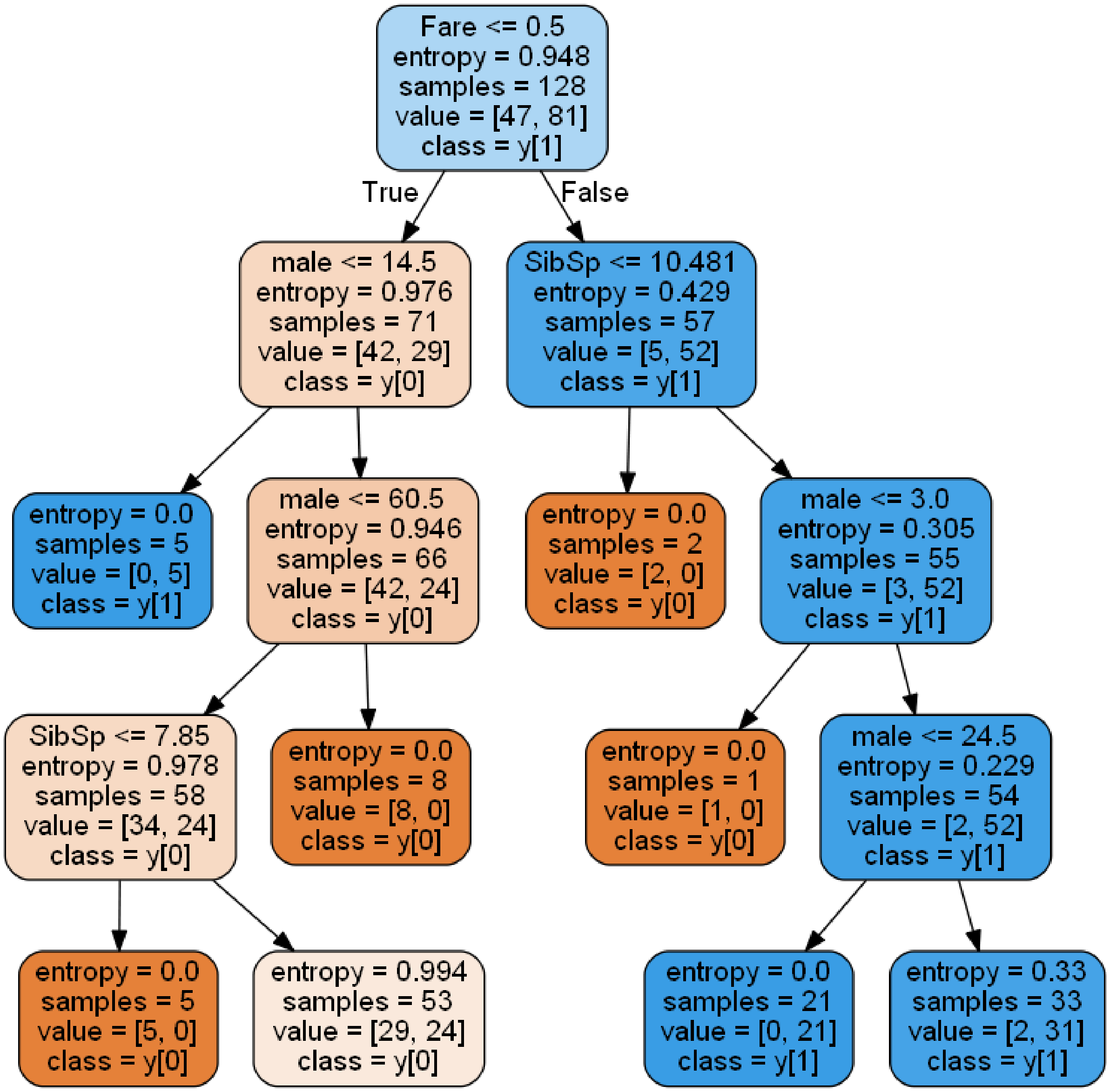
```
print("Accuracy for test data set:\n")
predicted = grid_result.predict(X_test)
print (format(metrics.accuracy_score(y_test, predicted) * 100,'.2f'), '%.')
```

```
Accuracy for test data set:

80.00 %.
```

```
In [24]: import pydotplus
         from sklearn import tree

         dot_data = tree.export_graphviz(grid_result.best_estimator_, out_file=None, fi
         lled=True, rounded=True,
                                         feature_names=['Pclass', 'male', 'female', 'Ag
         e', 'SibSp', 'Parch', 'Fare'],
                                         class_names=True)
         graph = pydotplus.graph_from_dot_data(dot_data)
         display(Image(graph.create_png()))
```


Fare <= 0.5
entropy = 0.948
samples = 128
value = [47, 81]
class = y[1]
True
False
male <= 14.5
entropy = 0.976
samples = 71
value = [42, 29]
class = y[0]
SibSp <= 10.481
entropy = 0.429
samples = 57
value = [5, 52]
class = y[1]
entropy = 0.0
samples = 5
value = [0, 5]
class = y[1]
male <= 60.5
entropy = 0.946
samples = 66
value = [42, 24]
class = y[0]
entropy = 0.0
samples = 2
value = [2, 0]
class = y[0]
male <= 3.0
entropy = 0.305
samples = 55
value = [3, 52]
class = y[1]
SibSp <= 7.85
entropy = 0.978
samples = 58
value = [34, 24]
class = y[0]
entropy = 0.0
samples = 8
value = [8, 0]
class = y[0]
entropy = 0.0
samples = 1
value = [1, 0]
class = y[0]
male <= 24.5
entropy = 0.229
samples = 54
value = [2, 52]
class = y[1]
entropy = 0.0
samples = 5
value = [5, 0]
class = y[0]
entropy = 0.994
samples = 53
value = [29, 24]
class = y[0]
entropy = 0.0
samples = 21
value = [0, 21]
class = y[1]
entropy = 0.33
samples = 33
value = [2, 31]
class = y[1]