Feature Selection and EDA Report

Null Values

Null Values by Columns/Features

Out of 12 total data columns, there are 3 columns with at least 1 null value.

Feature	Num of Nulls	Frac Null
Cabin	687	0.77
Age	177	0.2
Embarked	2	0.0

Null Values by Rows/Data Samples

Out of 891 total rows/data samples, 708 rows have at least one null value.

The row with the most NULL values has 2 NULLs.

Numeric vs Non-Numeric Features and Unique Values Count

Out of 11 total feature columns, there are 6 numeric columns and 5 non-numeric columns.

Numeric Feature	Num Unique Values
Pclass	3
SibSp	7
Parch	7

Non-Numeric Feature	Num Unique Values
Name	891
Ticket	681
Cabin	147

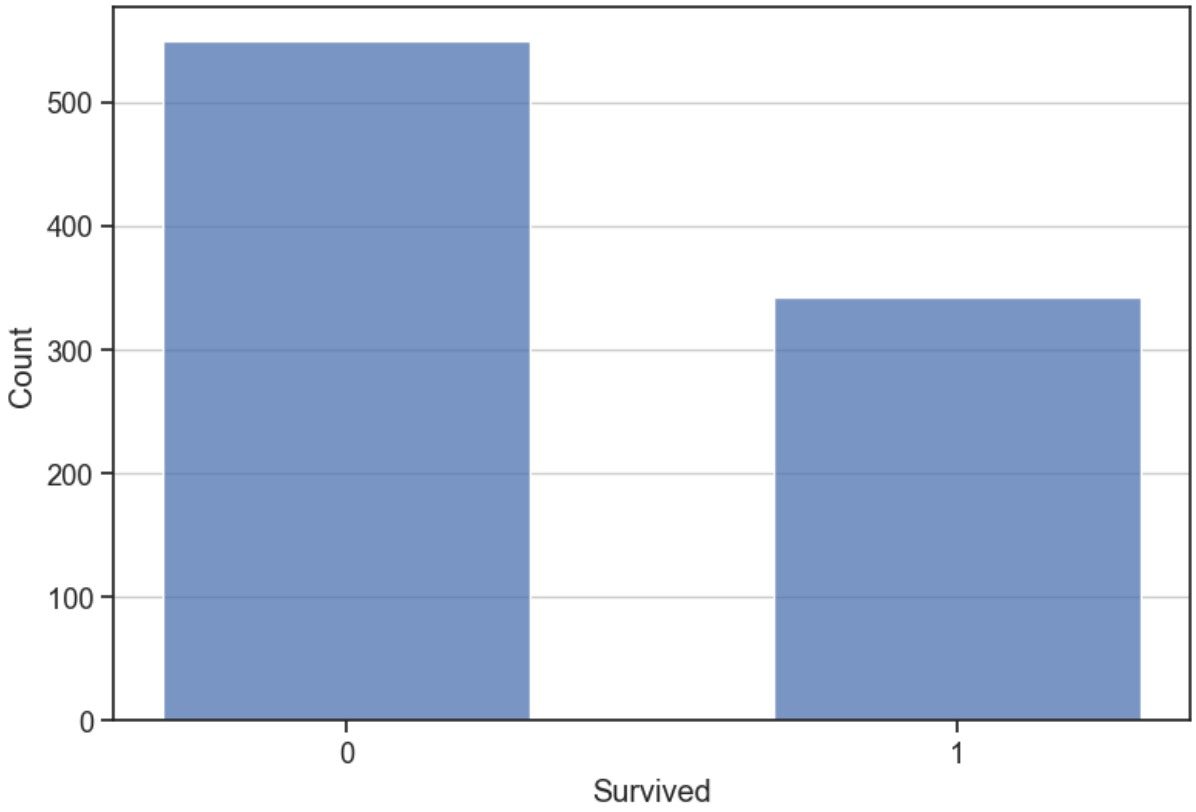
There are 0 non-numeric columns with just a single value and will be removed.

There are 3 non-numeric columns with a very large number of unique values and will be removed.

After the above adjustments, there are now 8 data columns, with 6 numeric columns and 2 non-numeric/categorical columns.

Target Column

For the chosen target column ('Survived'), this appears to be a classification problem.
The target column has 0 null values and 2 unique values.

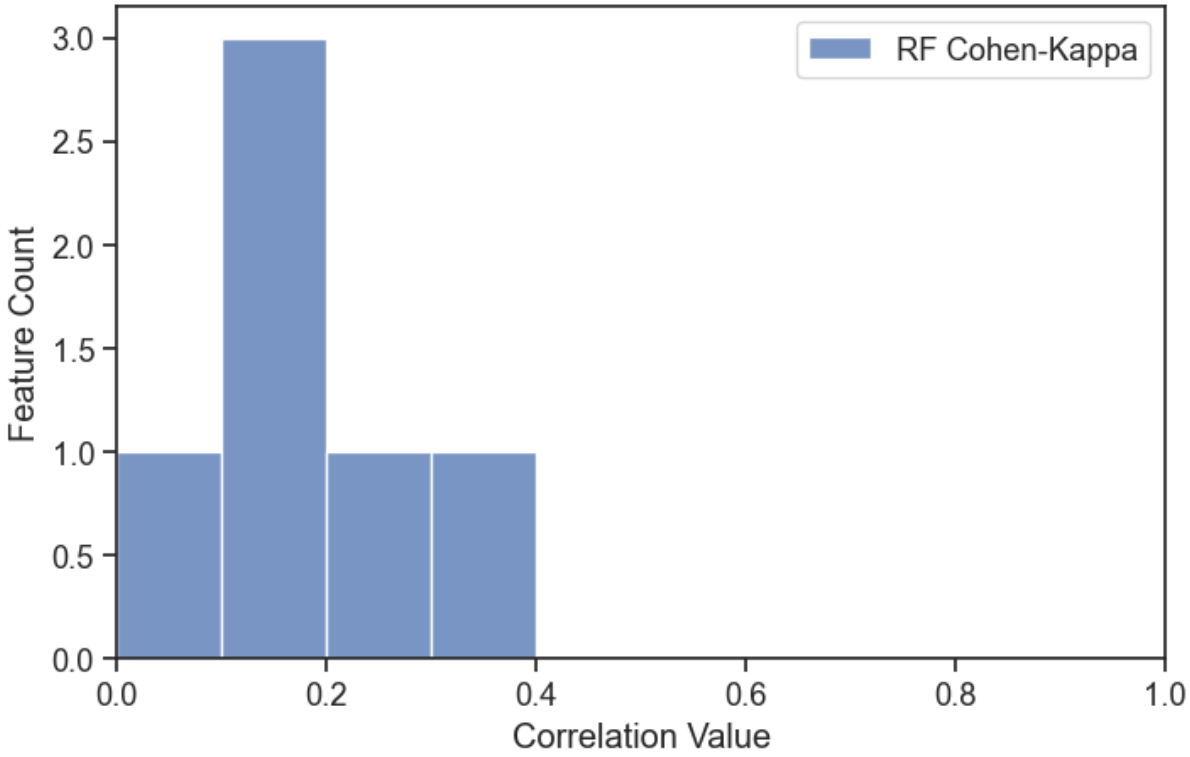


The above plot shows the distribution of values in the target column.

Feature Correlations

Correlations of Numeric Features with Target Variable

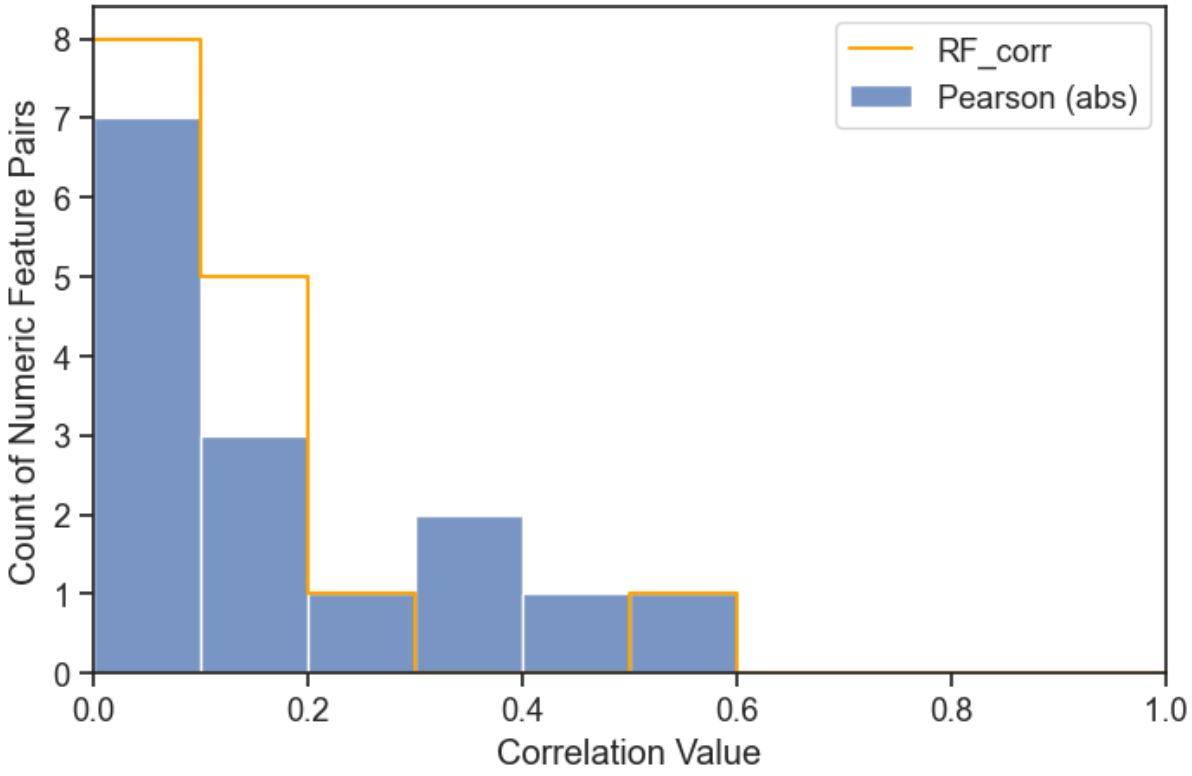
Numeric Feature	Count non-Null	Mutual Info	RF Corr
Fare	891	0.12	0.35
Pclass	891	0.06	0.27
SibSp	891	0.01	0.16
Age	714	0.03	0.13
Parch	891	0.04	0.11
PassengerId	891	0.01	0.08



The above plot shows a histogram of all numeric features and their correlation value with the target variable.

Correlations between Numeric Features

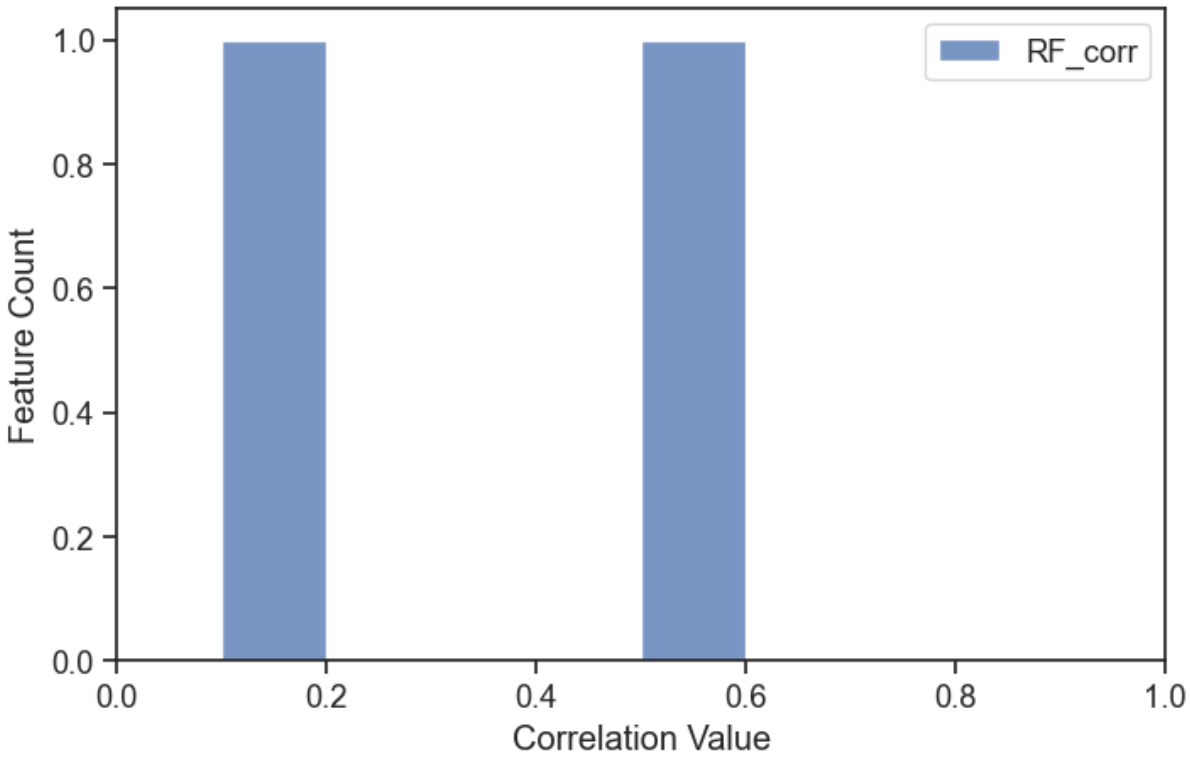
Numeric Feature	Avg Pearson Correlation	Avg RF Correlation	Feat with Max Pear Corr	Max Pear Corr	Feat with Max RF Corr	Max RF Corr
Pclass	0.21	0.15	Fare	-0.55	Fare	0.53
Fare	0.21	0.19	Pclass	-0.55	Pclass	0.53
SibSp	0.2	0.12	Parch	0.41	Parch	0.21
Parch	0.17	0.11	SibSp	0.41	SibSp	0.21
Age	0.2	0.12	Pclass	-0.37	SibSp	0.18
PassengerId	0.03	0.03	SibSp	-0.06	Age	0.05



The above plot shows a histogram of all unique pairs of numeric features and the correlation between the two features of each the pair.

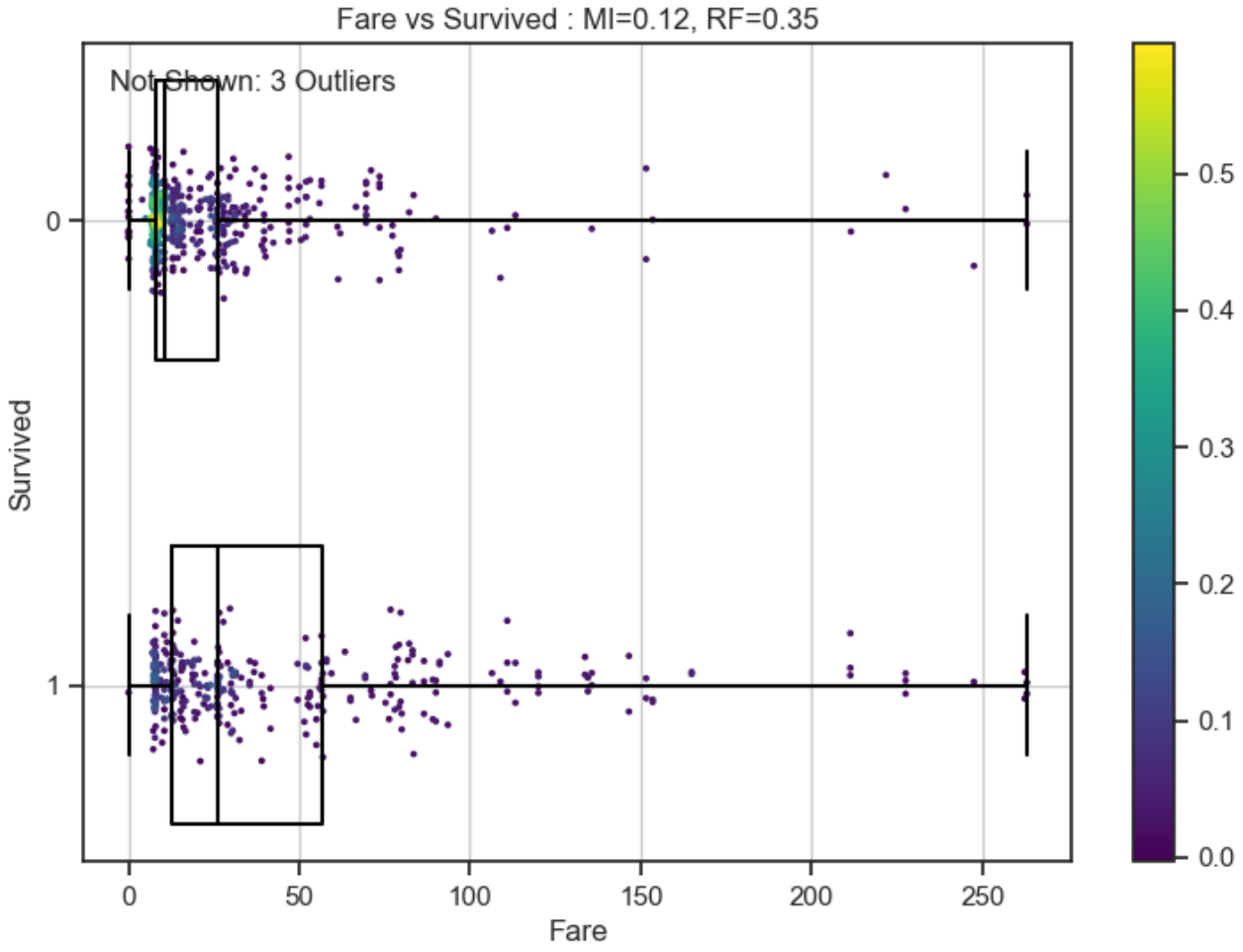
Correlations of Non-Numeric Features with Target Variable

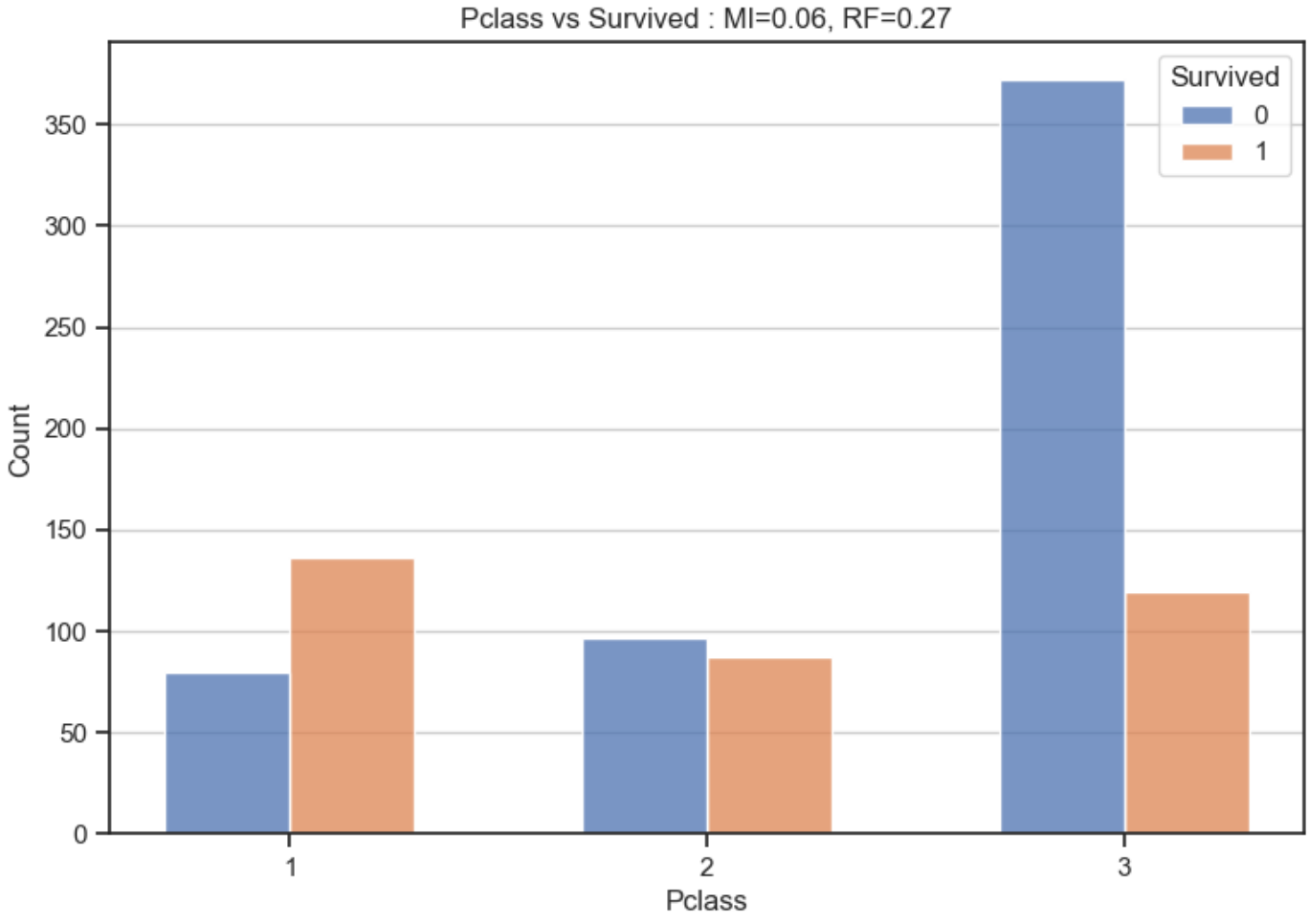
Non-Numeric Feature	Count non-Null	Num Unique	Mutual Info	RF Corr
Sex	891	2	0.15	0.54
Embarked	889	3	0.01	0.15

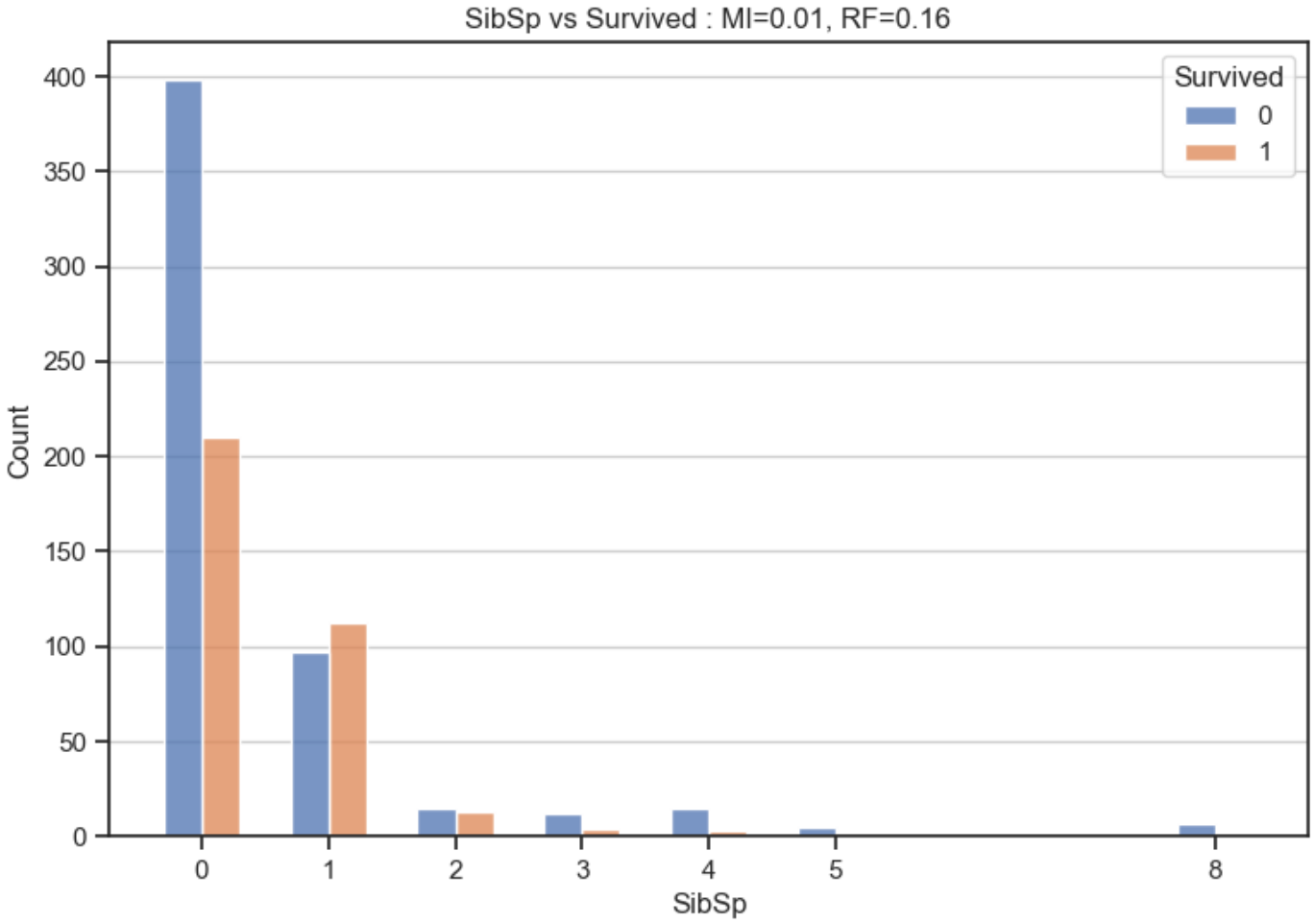


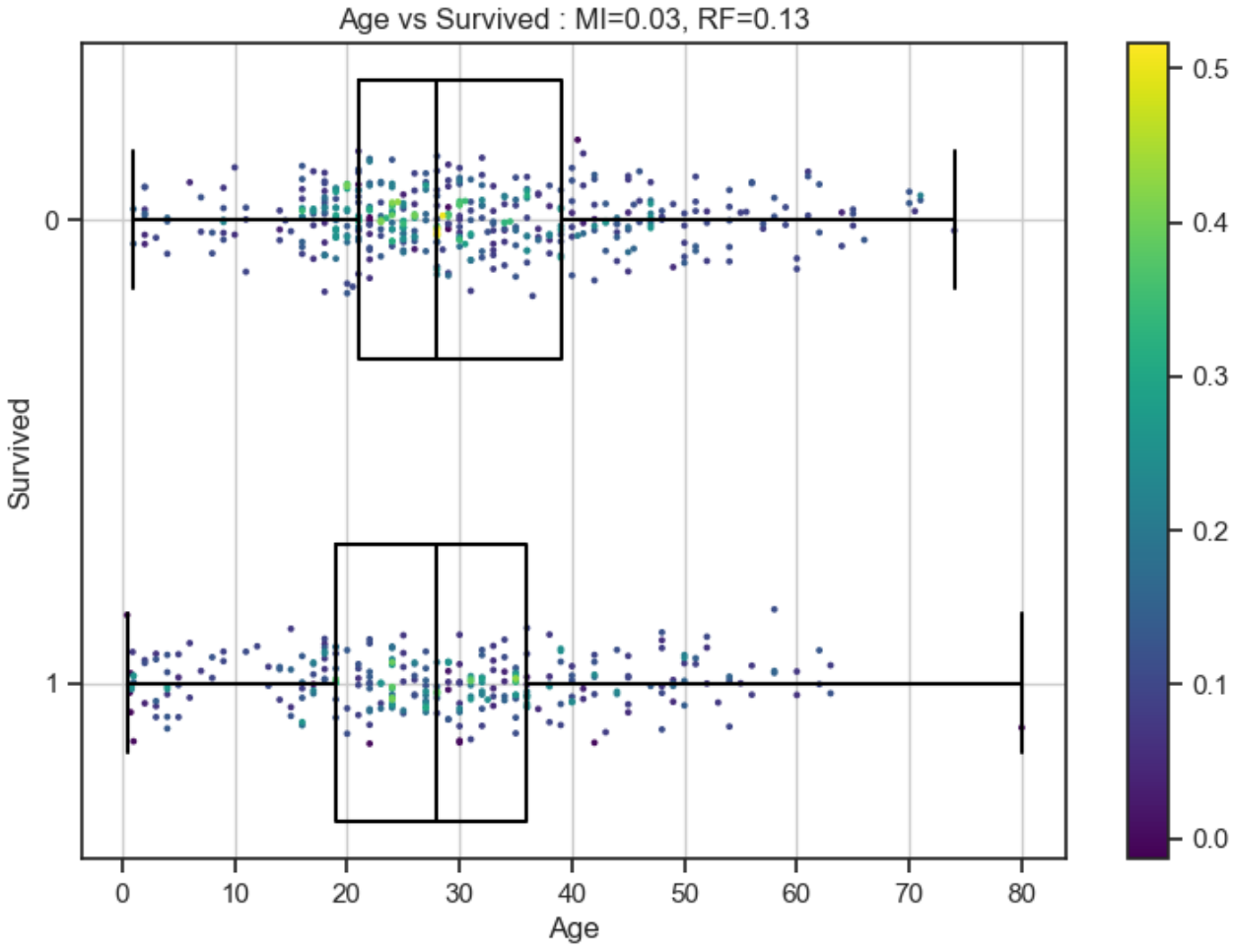
The above plot shows a histogram of all non-numeric features and their correlation value with the target variable.

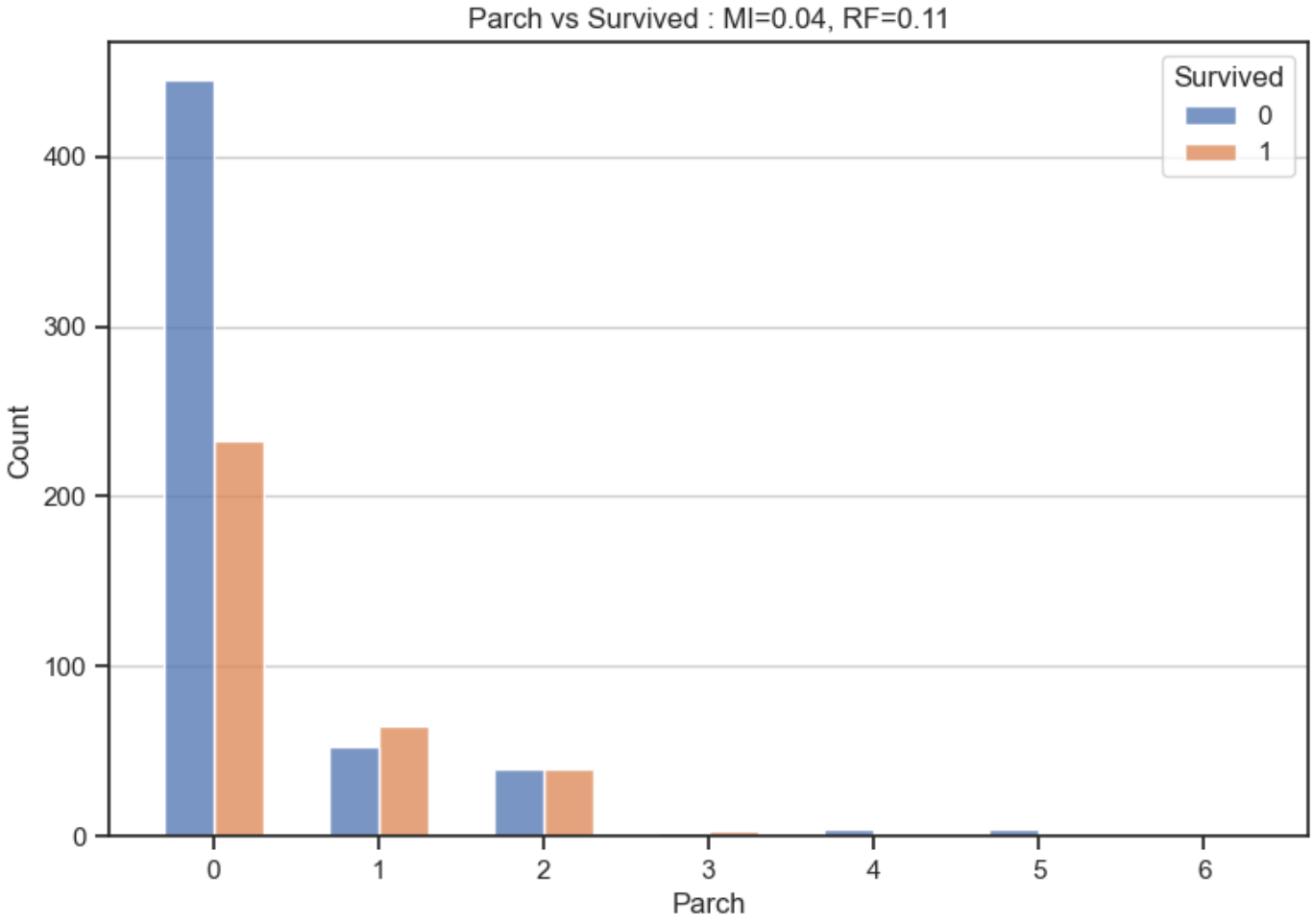
Plots of Numeric Columns versus the Target Variable

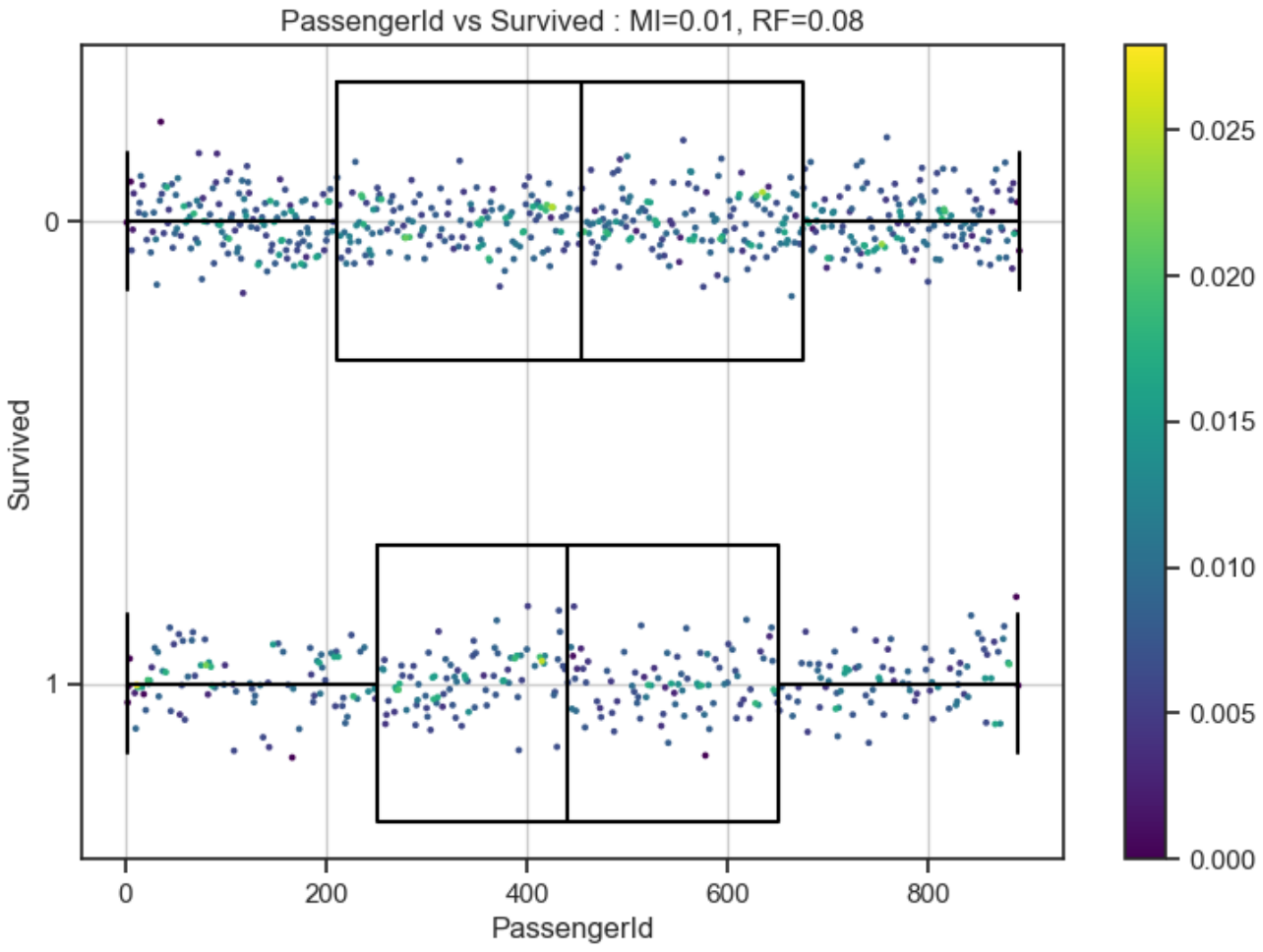

Fare vs Survived : MI=0.12, RF=0.35
Not Shown: 3 Outliers
Survived
0
1
0
50
100
150
200
250
Fare
0.5
0.4
0.3
0.2
0.1
0.0


Pclass vs Survived : MI=0.06, RF=0.27
Survived
0
1
Count
350
300
250
200
150
100
50
0
1
2
3
Pclass

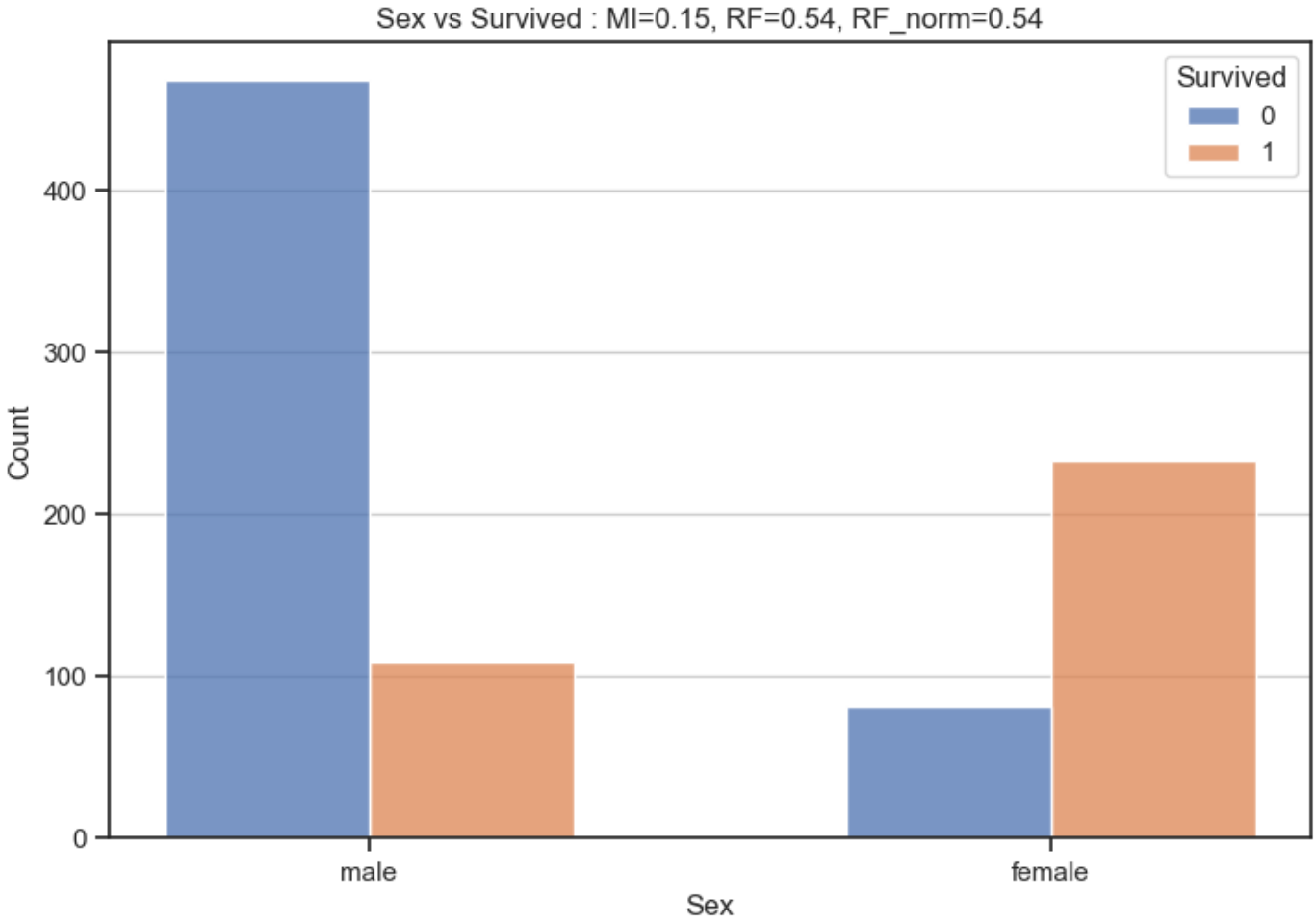

SibSp vs Survived : MI=0.01, RF=0.16
Survived
0
1
Count
SibSp
0
1
2
3
4
5
8
0
50
100
150
200
250
300
350
400

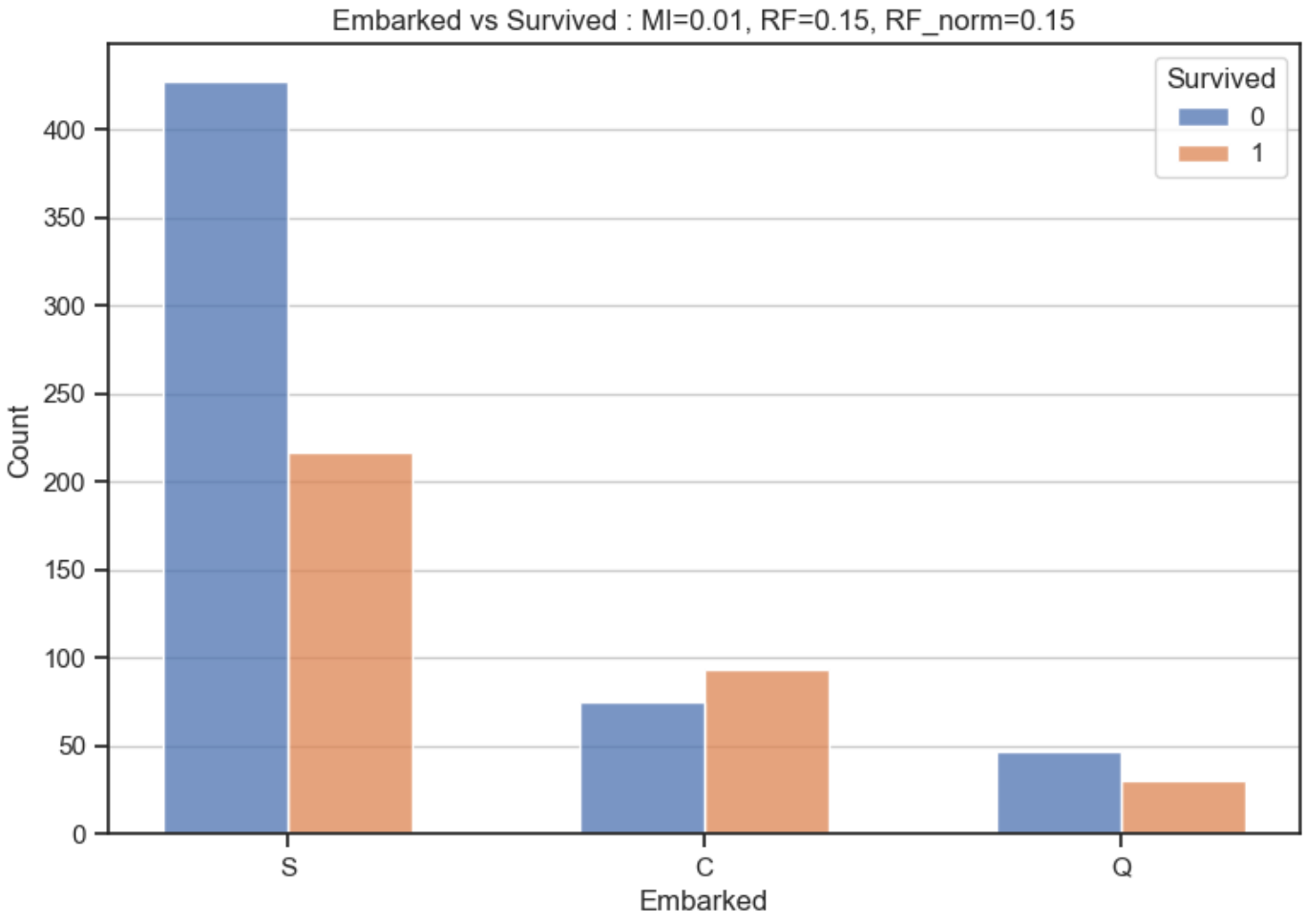

Age vs Survived : MI=0.03, RF=0.13
Survived
Age
0
1
0
10
20
30
40
50
60
70
80
0.0
0.1
0.2
0.3
0.4
0.5


Parch vs Survived : MI=0.04, RF=0.11
Survived
0
1
Count
400
300
200
100
0
0
1
2
3
4
5
6
Parch


PassengerId vs Survived : MI=0.01, RF=0.08
Survived
0
1
0
200
400
600
800
PassengerId
0.025
0.020
0.015
0.010
0.005
0.000

Plots of Non-Numeric Columns versus the Target Variable


Sex vs Survived : MI=0.15, RF=0.54, RF_norm=0.54
Survived
0
1
Count
400
300
200
100
0
male
female
Sex


Embarked vs Survived : MI=0.01, RF=0.15, RF_norm=0.15
Survived
0
1
Count
400
350
300
250
200
150
100
50
0
S
C
Q
Embarked