Automated EDA - ClfAutoEDA

November 14, 2022

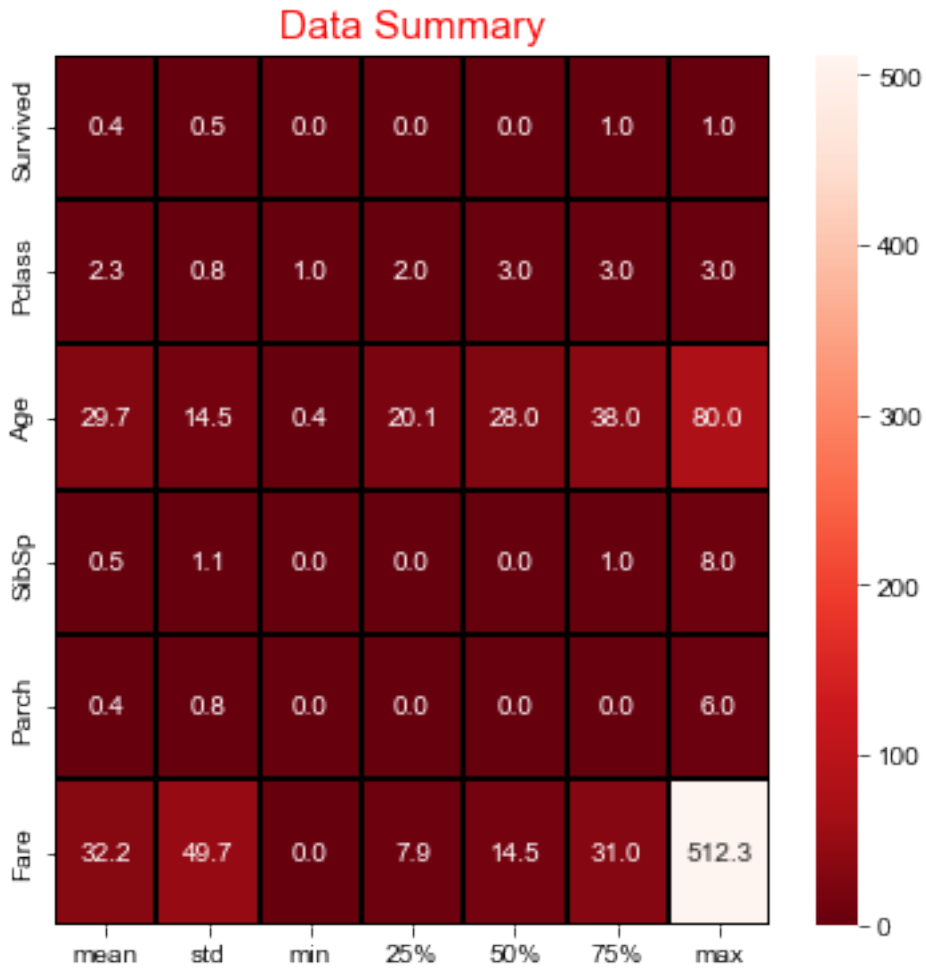
[1]:
```
from ClfAutoEDA import *
df=pd.read_csv('titanic_train.csv')
#Dropping Id related columns
df.drop(['PassengerId','Ticket'],axis=1,inplace=True)
#Setting parameter values
labels=["not survived","survived"]
target_variable_name='Survived'
df_processed,num_features,cat_features=EDA(df,labels,
                                           target_variable_name,
                                           data_summary_figsize=(6,6),
                                           corr_matrix_figsize=(6,6),
                                           corr_matrix_annot=True,
                                           pairplt=True)
```

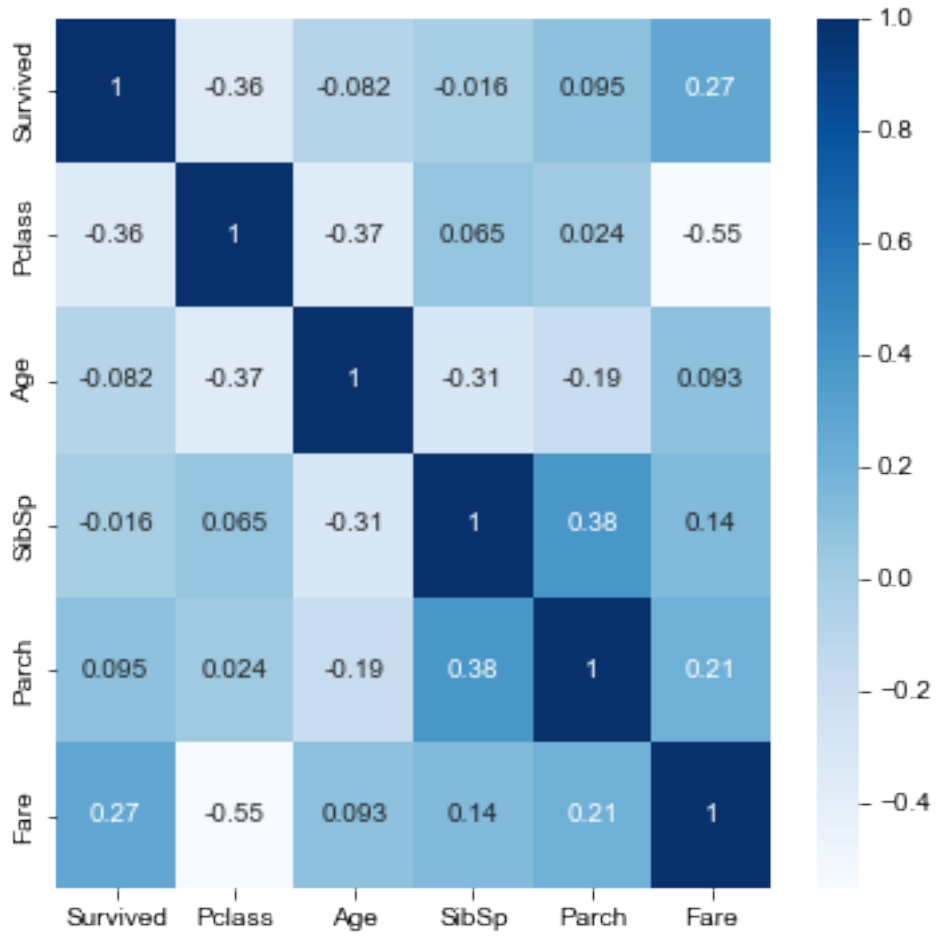
```
The data looks like this:
    Survived  Pclass                                                Name  \
0          0       3                             Braund, Mr. Owen Harris
1          1       1  Cumings, Mrs. John Bradley (Florence Briggs Th…
2          1       3                              Heikkinen, Miss. Laina
3          1       1       Futrelle, Mrs. Jacques Heath (Lily May Peel)
4          0       3                            Allen, Mr. William Henry

      Sex   Age  SibSp  Parch     Fare Cabin Embarked
0    male  22.0      1      0   7.2500   NaN        S
1  female  38.0      1      0  71.2833   C85        C
2  female  26.0      0      0   7.9250   NaN        S
3  female  35.0      1      0  53.1000  C123        S
4    male  35.0      0      0   8.0500   NaN        S

The shape of data is:  (891, 10)

The missing values in data are:
 Cabin       687
Age          177
Embarked       2
Survived       0
Pclass         0
Name           0
```

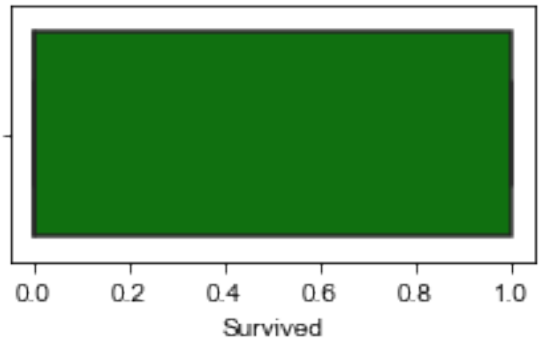
```
Sex           0
SibSp         0
Parch         0
Fare          0
dtype: int64

The summary of data is:
          Survived      Pclass         Age       SibSp       Parch        Fare
count  891.000000  891.000000  714.000000  891.000000  891.000000  891.000000
mean     0.383838    2.308642   29.699118    0.523008    0.381594   32.204208
std      0.486592    0.836071   14.526497    1.102743    0.806057   49.693429
min      0.000000    1.000000    0.420000    0.000000    0.000000    0.000000
25%      0.000000    2.000000   20.125000    0.000000    0.000000    7.910400
50%      0.000000    3.000000   28.000000    0.000000    0.000000   14.454200
75%      1.000000    3.000000   38.000000    1.000000    0.000000   31.000000
max      1.000000    3.000000   80.000000    8.000000    6.000000  512.329200

Some useful data information:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  891 non-null    int64
 1   Pclass    891 non-null    int64
 2   Name      891 non-null    object
 3   Sex       891 non-null    object
 4   Age       714 non-null    float64
 5   SibSp     891 non-null    int64
 6   Parch     891 non-null    int64
 7   Fare      891 non-null    float64
 8   Cabin     204 non-null    object
 9   Embarked  889 non-null    object
dtypes: float64(2), int64(4), object(4)
memory usage: 69.7+ KB
None

The columns in data are:
 ['Survived' 'Pclass' 'Name' 'Sex' 'Age' 'SibSp' 'Parch' 'Fare' 'Cabin'
 'Embarked']

The target variable is divided into:
 0    424
1    288
Name: Survived, dtype: int64

The numerical features are:
```

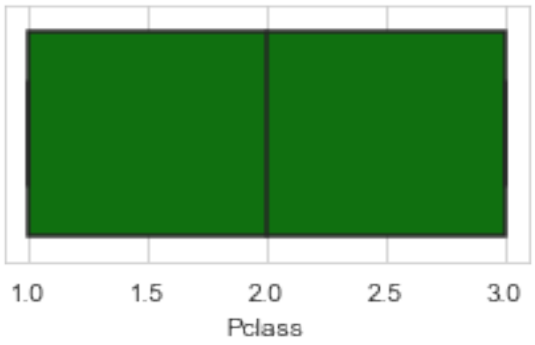
```
 ['Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']

The categorical features are:
 ['Name', 'Sex', 'Embarked']

The categorical variable is divided into:
 Braund, Mr. Owen Harris                 1
West, Miss. Constance Mirium             1
Palsson, Miss. Torborg Danira            1
Bonnell, Miss. Elizabeth                 1
Heikkinen, Miss. Laina                   1
                                        ..
Sutehall, Mr. Henry Jr                   1
Rice, Mrs. William (Margaret Norton)     1
Montvila, Rev. Juozas                    1
Graham, Miss. Margaret Edith             1
Dooley, Mr. Patrick                      1
Name: Name, Length: 712, dtype: int64

The categorical variable Name has too many divisions to plot


The categorical variable is divided into:
 male      453
female    259
Name: Sex, dtype: int64

The categorical variable is divided into:
 S    554
C    130
Q     28
Name: Embarked, dtype: int64

Execution Time for EDA: 0.10 minutes
```

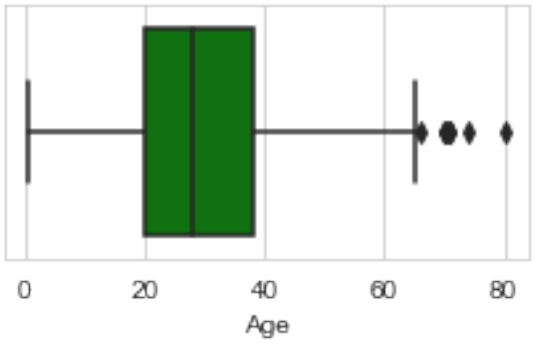

Missing Values Summary
0
43
86
129
172
215
258
301
344
387
430
473
516
559
602
645
688
731
774
817
860
1.0
0.8
0.6
0.4
0.2
0.0
Survived
Pclass
Name
Sex
Age
SibSp
Parch
Fare
Cabin
Embarked

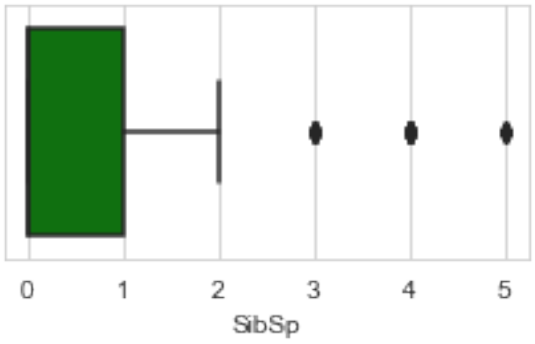

Data Summary
Survived
Pclass
Age
SibSp
Parch
Fare
0.4
0.5
0.0
0.0
0.0
1.0
1.0
2.3
0.8
1.0
2.0
3.0
3.0
3.0
29.7
14.5
0.4
20.1
28.0
38.0
80.0
0.5
1.1
0.0
0.0
0.0
1.0
8.0
0.4
0.8
0.0
0.0
0.0
0.0
6.0
32.2
49.7
0.0
7.9
14.5
31.0
512.3
mean
std
min
25%
50%
75%
max
500
400
300
200
100
0

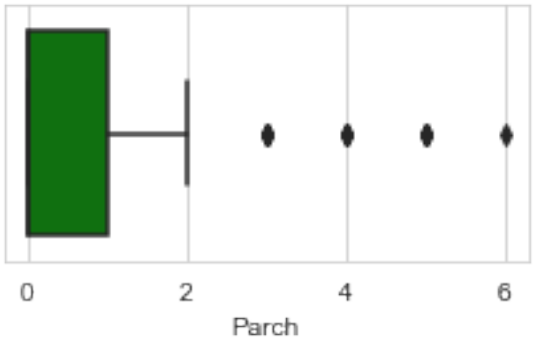

Survived
Pclass
Age
SibSp
Parch
Fare
1
-0.36
-0.082
-0.016
0.095
0.27
-0.36
1
-0.37
0.065
0.024
-0.55
-0.082
-0.37
1
-0.31
-0.19
0.093
-0.016
0.065
-0.31
1
0.38
0.14
0.095
0.024
-0.19
0.38
1
0.21
0.27
-0.55
0.093
0.14
0.21
1
Survived
Pclass
Age
SibSp
Parch
Fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

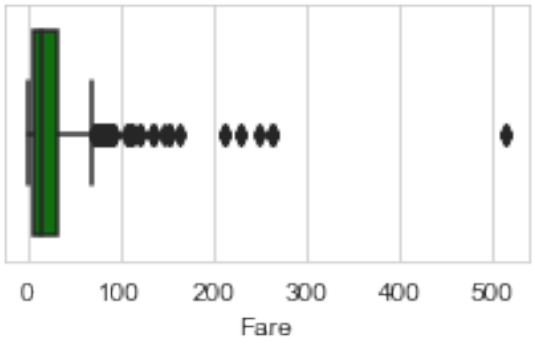

0.0
0.2
0.4
0.6
0.8
1.0
Survived

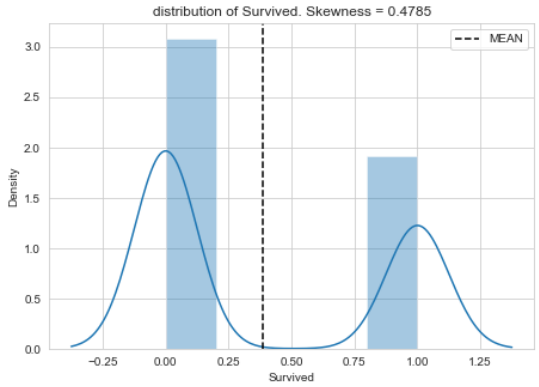

1.0
1.5
2.0
2.5
3.0
Pclass

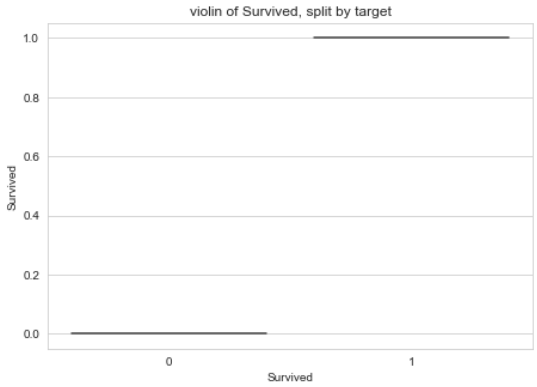

0
20
40
60
80
Age

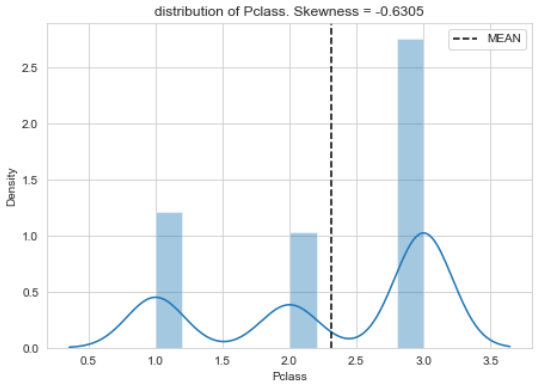

0
1
2
3
4
5
SibSp

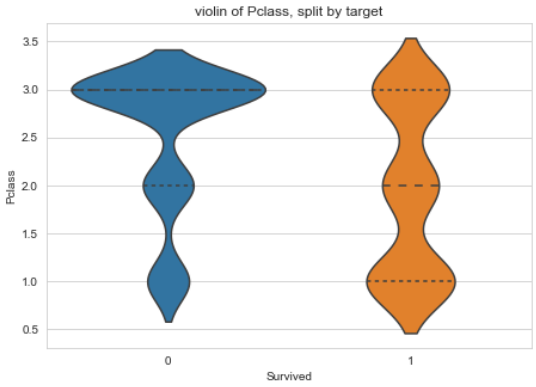

0
2
4
6
Parch

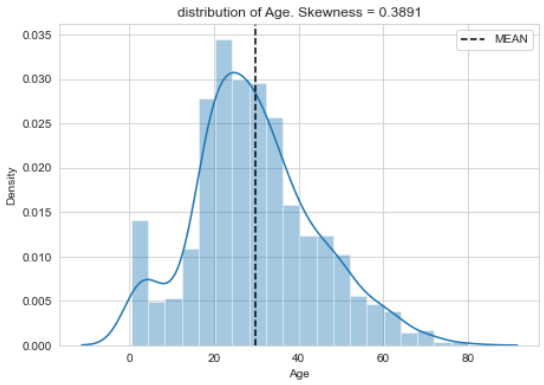

0
100
200
300
400
500
Fare

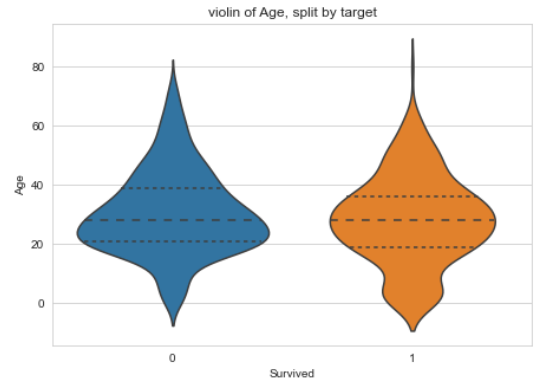

distribution of Survived. Skewness = 0.4785
MEAN
Density
Survived

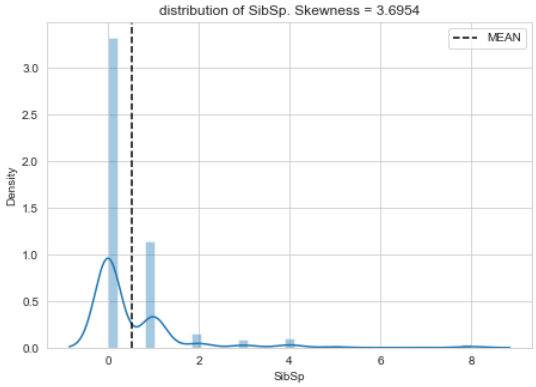

violin of Survived, split by target
Survived
Survived

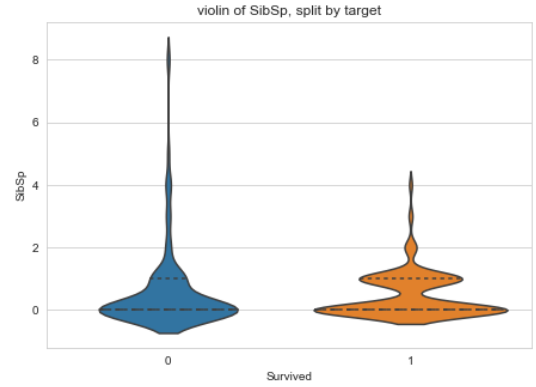

distribution of Pclass. Skewness = -0.6305
MEAN
Density
Pclass

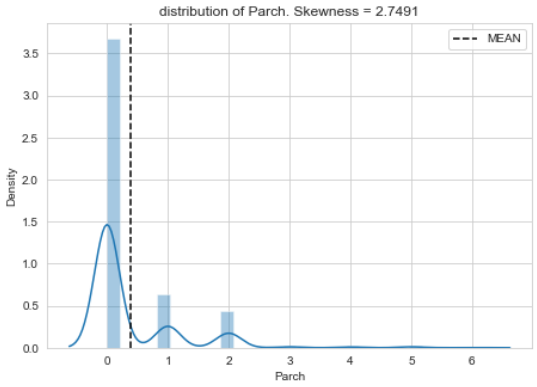

violin of Pclass, split by target
Pclass
Survived

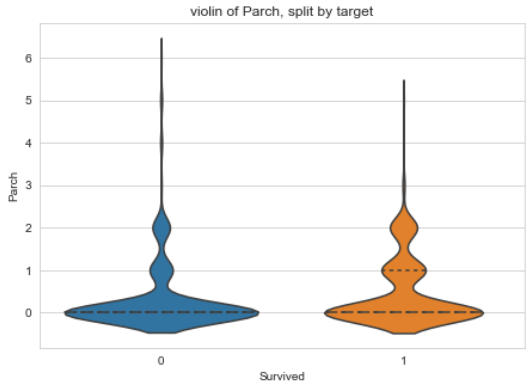

distribution of Age. Skewness = 0.3891
MEAN
Density
Age

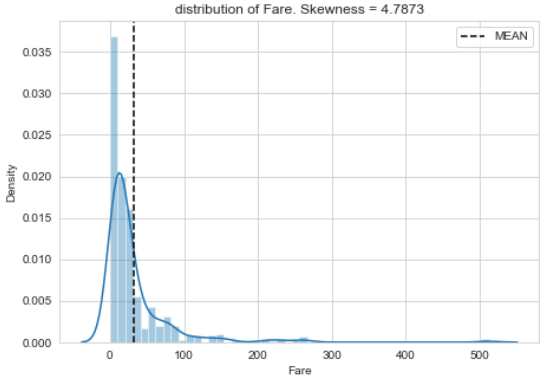

violin of Age, split by target
Age
Survived

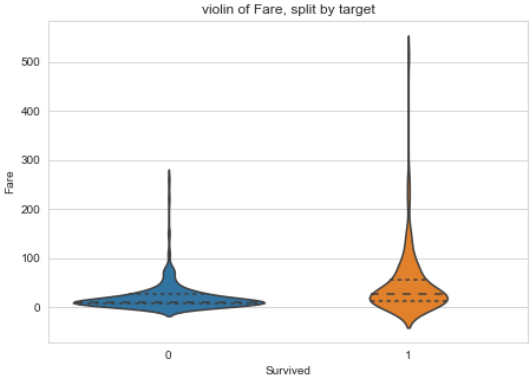

distribution of SibSp. Skewness = 3.6954
MEAN
Density
SibSp

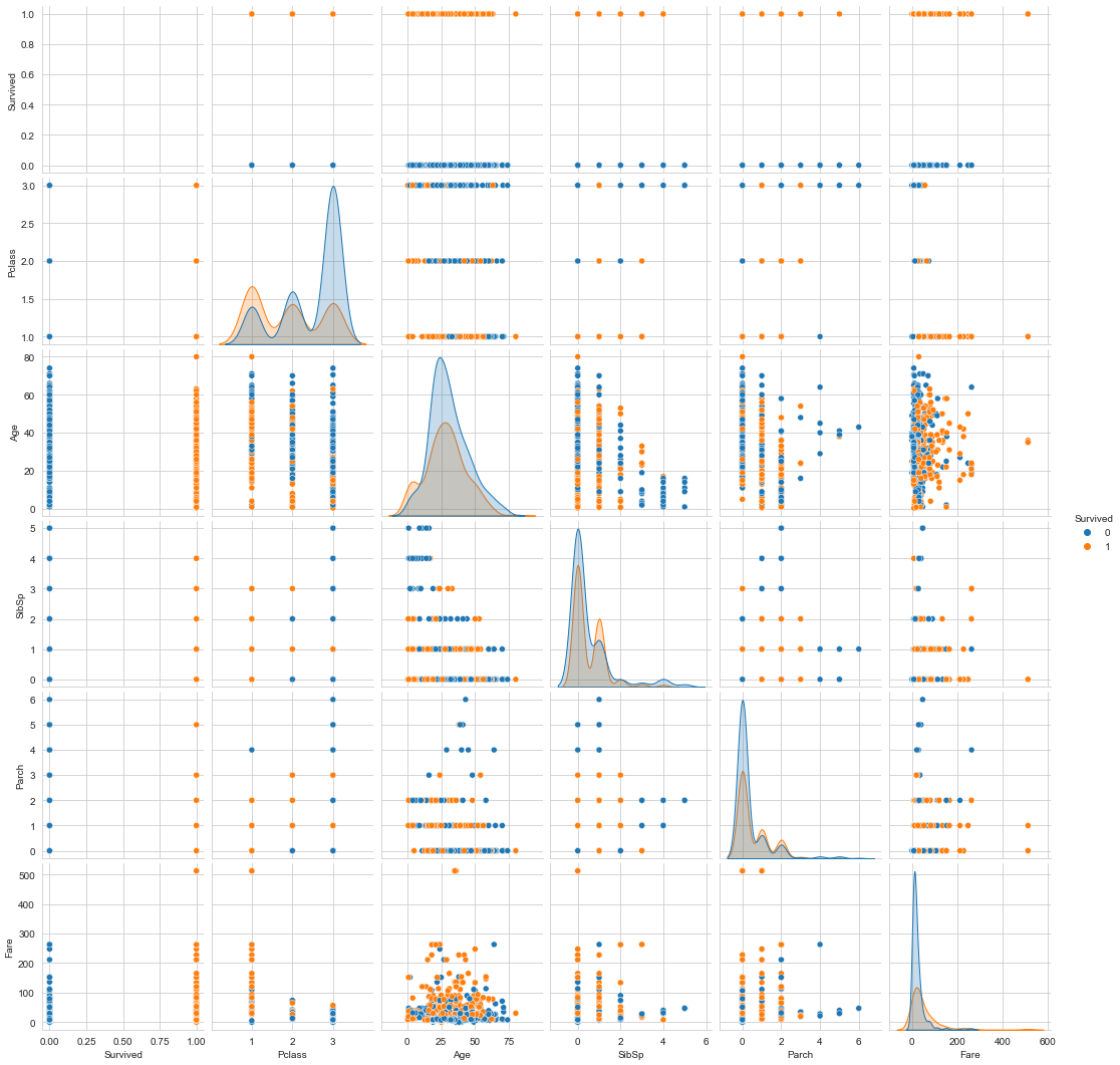

violin of SibSp, split by target
SibSp
Survived

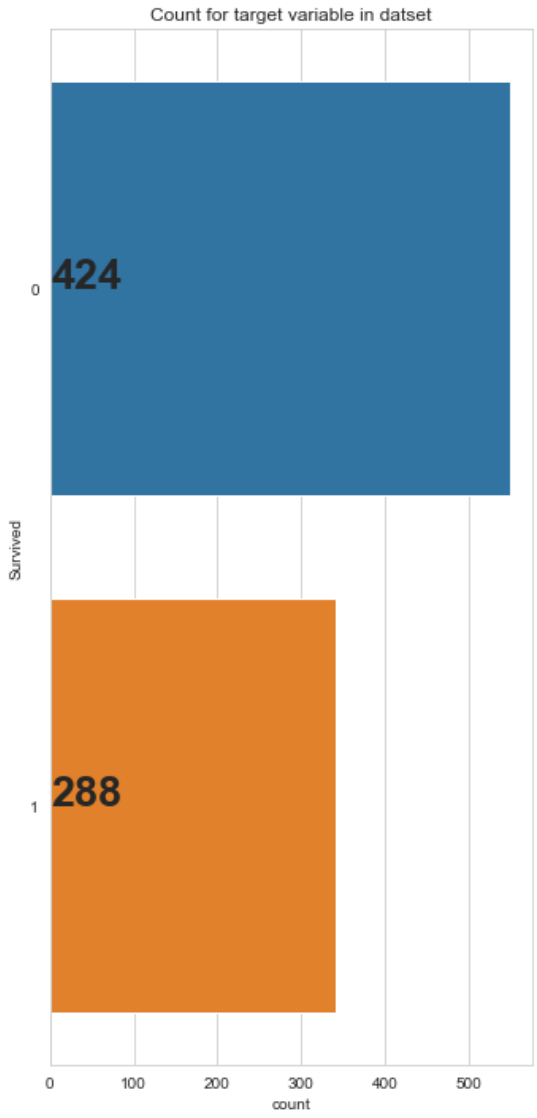

distribution of Parch. Skewness = 2.7491
MEAN
Density
Parch

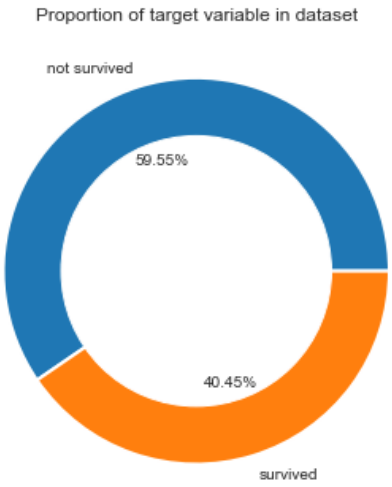

violin of Parch, split by target
Parch
Survived

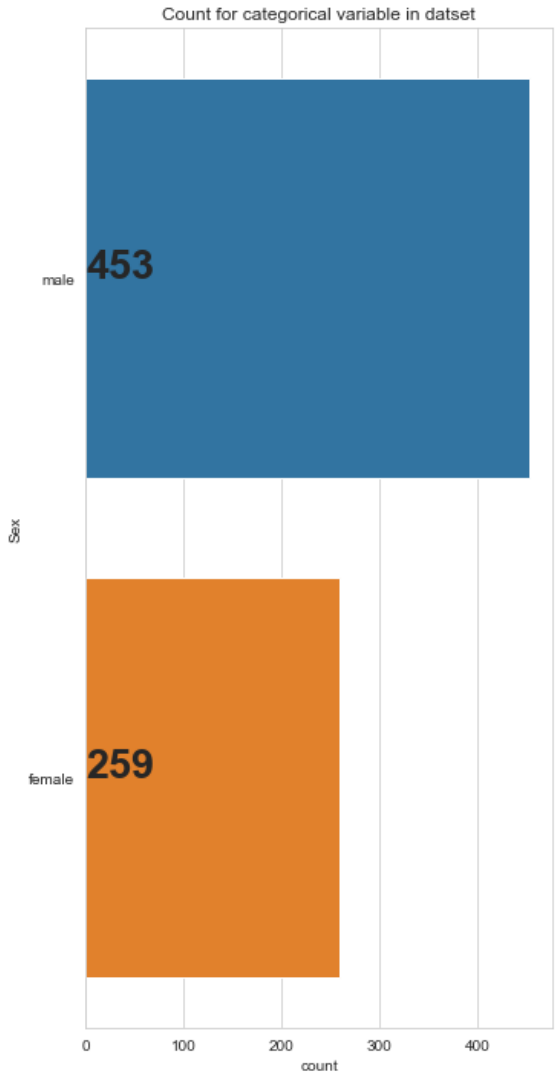

distribution of Fare. Skewness = 4.7873
MEAN
Density
Fare

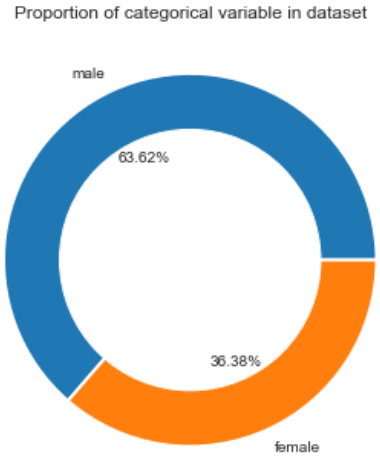

violin of Fare, split by target
Fare
Survived

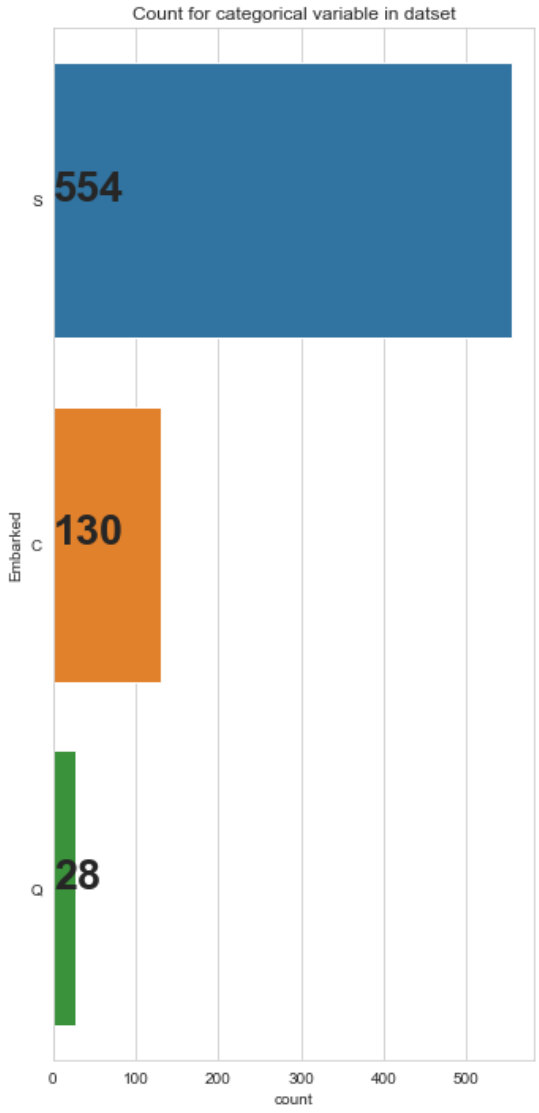

Survived
Pclass
Age
SibSp
Parch
Fare
Survived
0
1

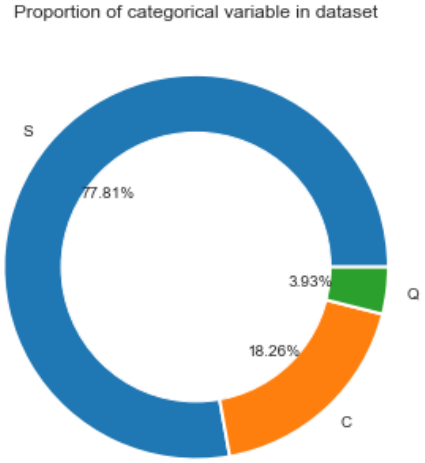

Count for target variable in datset
Survived
0
1
424
288
0
100
200
300
400
500
count


Proportion of target variable in dataset
not survived
59.55%
40.45%
survived


Count for categorical variable in datset
male
453
female
259
Sex
0
100
200
300
400
count


Proportion of categorical variable in dataset
male
63.62%
36.38%
female


Count for categorical variable in datset
S
554
C
130
Q
28
Embarked
0
100
200
300
400
500
count


Proportion of categorical variable in dataset
S
77.81%
3.93%
Q
18.26%
C

[ ]: