main

July 27, 2021

1 Akhil Shukla

1.1 Roll No,.- 180057

Variable Definition Key * survival | Survival | 0 = No, 1 = Yes <> * pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd * sex | Sex
* Age | Age in years
* sibsp | Number of siblings / spouses aboard the Titanic
* parch | Number of parents / children aboard the Titanic
* ticket | Ticket number
* fare | Passenger fare * cabin | Cabin number
* embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton

Top 5 tuples of the given data

1.2 Basic Data Understanding

[93]:

	PassengerId	Survived	Pclass	\
0	1	0	3	
1	2	1	1	
2	3	1	3	
3	4	1	1	
4	5	0	3	

	Name	Sex	Age	SibSp	\
0	Braund, Mr. Owen Harris	male	22.0	1	
1	Cumings, Mrs. John Bradley (Florence Briggs Th…	female	38.0	1	
2	Heikkinen, Miss. Laina	female	26.0	0	
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	
4	Allen, Mr. William Henry	male	35.0	0	

	Parch	Ticket	Fare	Cabin	Embarked
0	0	A/5 21171	7.2500	NaN	S
1	0	PC 17599	71.2833	C85	C
2	0	STON/O2. 3101282	7.9250	NaN	S
3	0	113803	53.1000	C123	S
4	0	373450	8.0500	NaN	S

Count of missing values in the dataset.

[94]:
```
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
```

Data-type of each attribute

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Statistical description of numerical attribute.

[96]:
```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000
```

	Parch	Fare
count	891.000000	891.000000
mean	0.381594	32.204208
std	0.806057	49.693429
min	0.000000	0.000000
25%	0.000000	7.910400
50%	0.000000	14.454200
75%	0.000000	31.000000
max	6.000000	512.329200

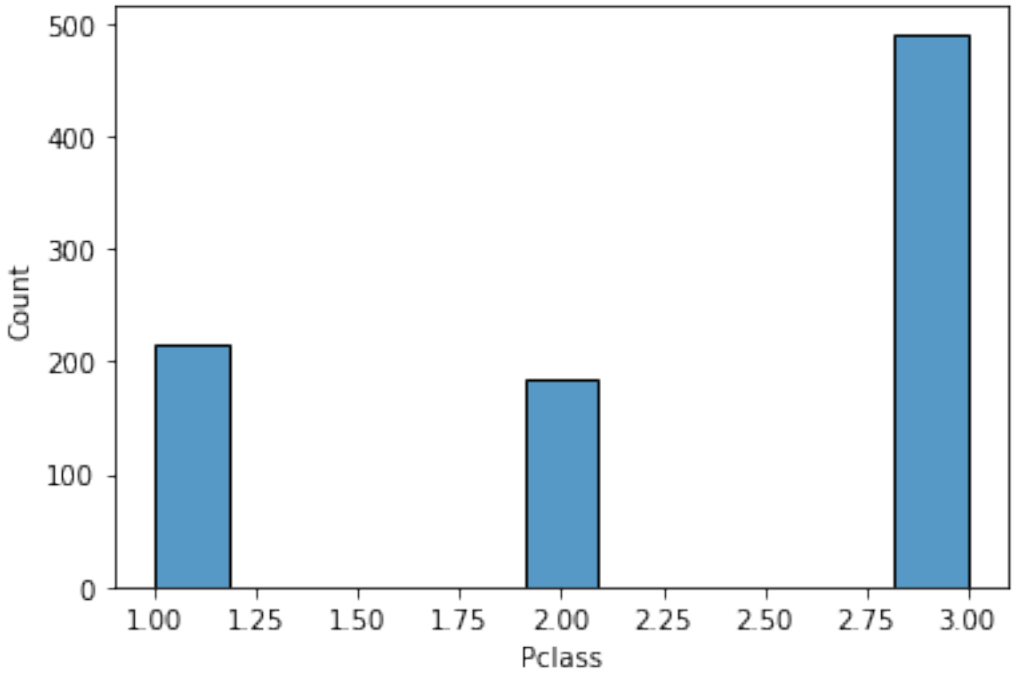
1.3 Data Analysis,Cleaning and Transformation

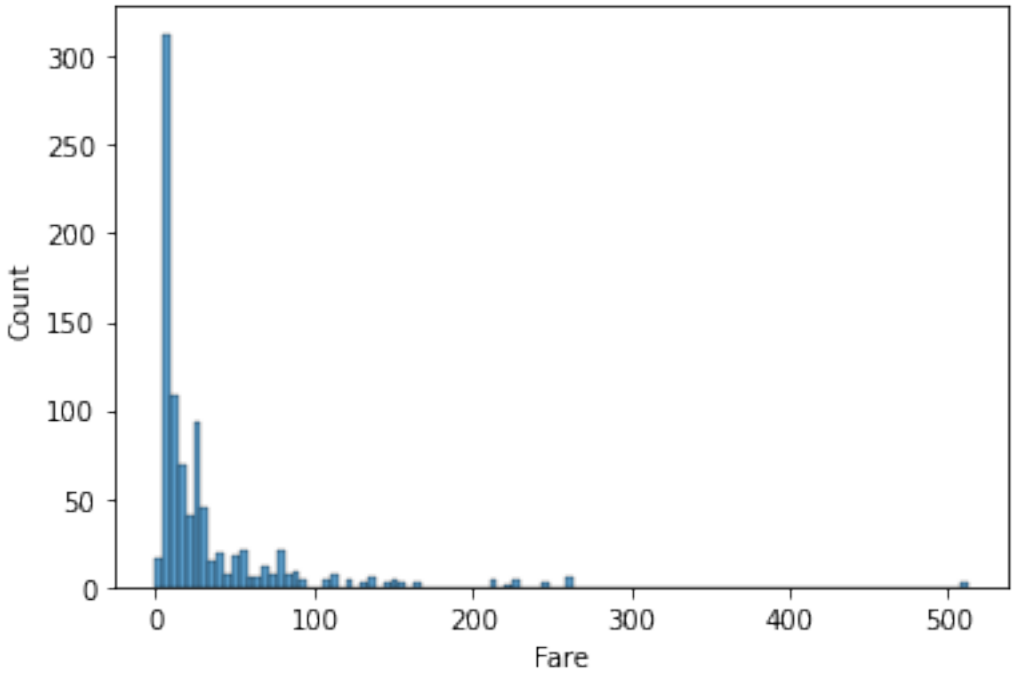
Dividing the columns, and extracting Titles (Mr. Mrs, Miss. , ..etc) from the names of the passenger and dropping the rest. Names are useful but the title depict the rank of the passenger in the society, and it might affect the survival. Fillinf the missing values using interpolation.

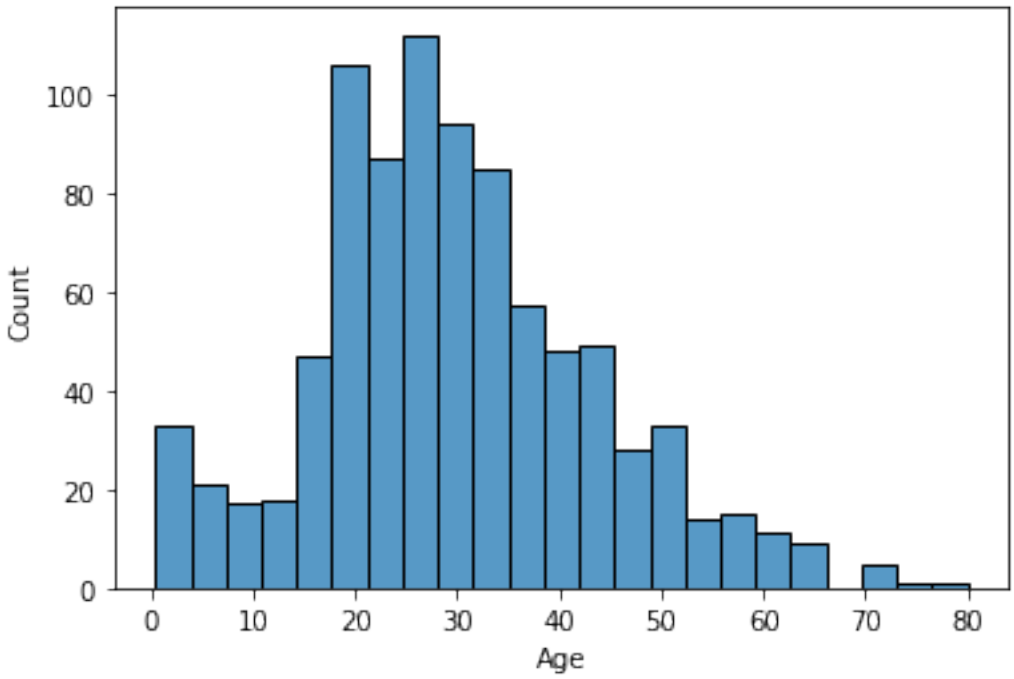
1.3.1 Data Cleaning

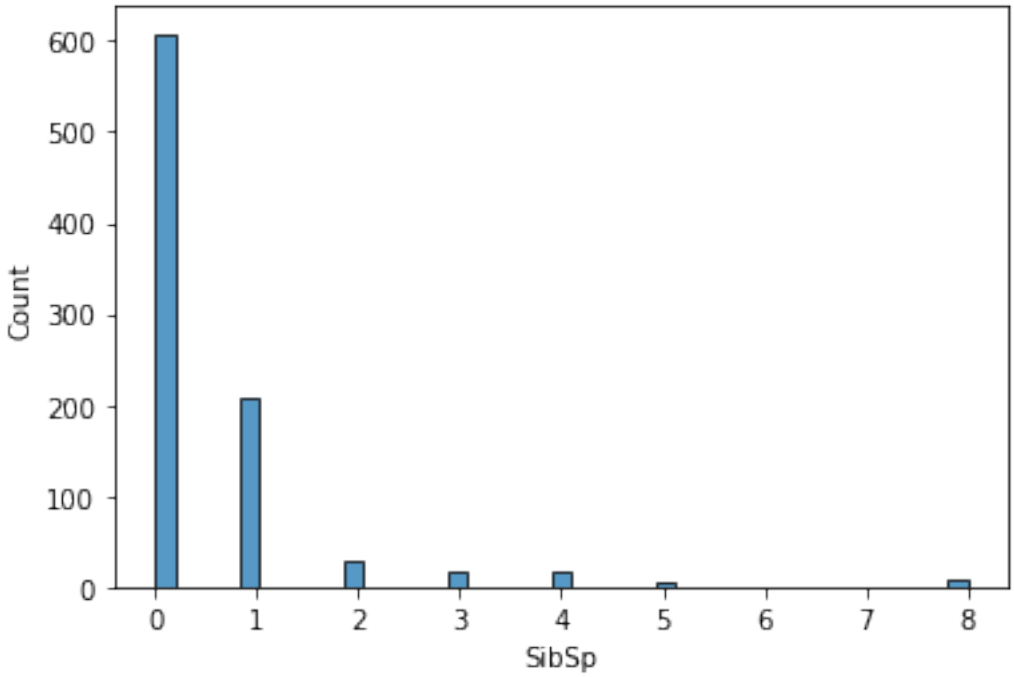
1.3.2 Plotting Data

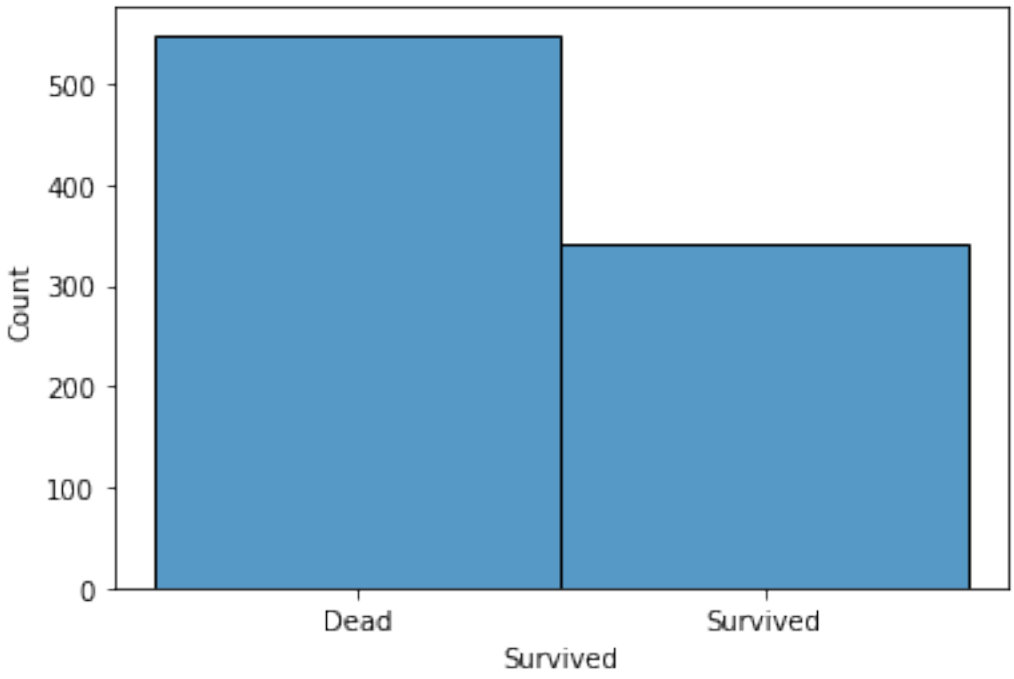
Understanding the distribution of some attributes.

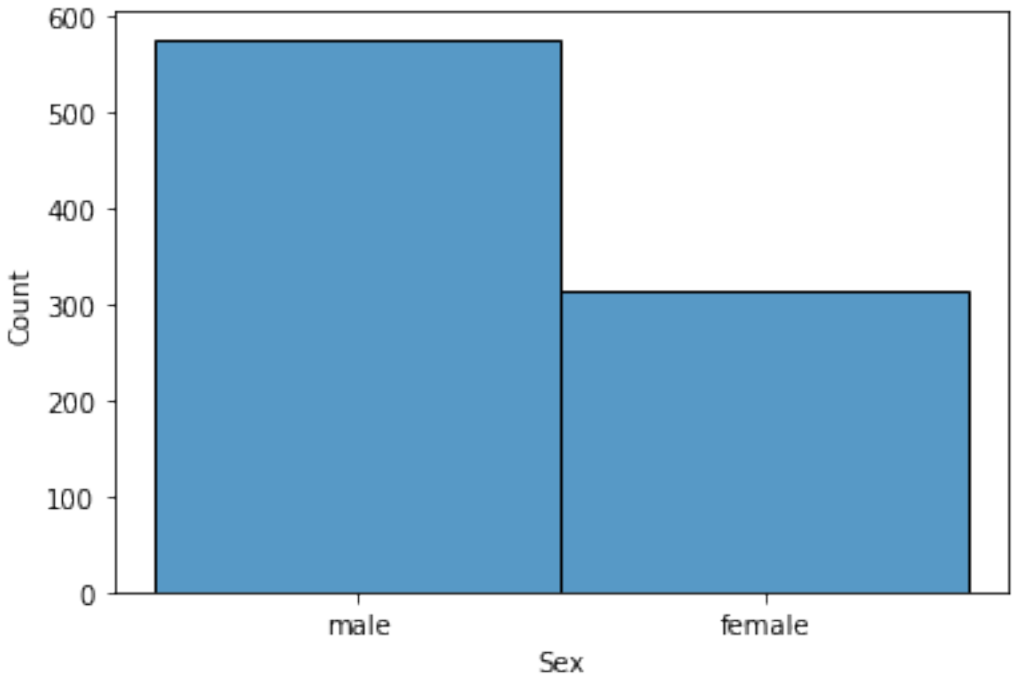


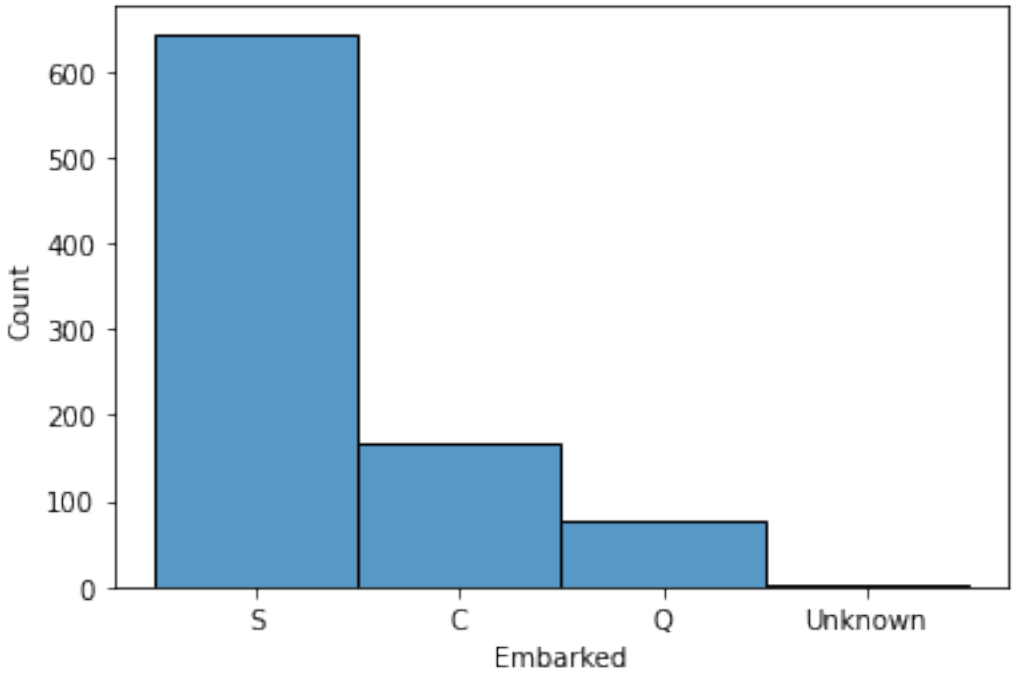

300
250
200
150
100
50
0
Count
0
100
200
300
400
500
Fare

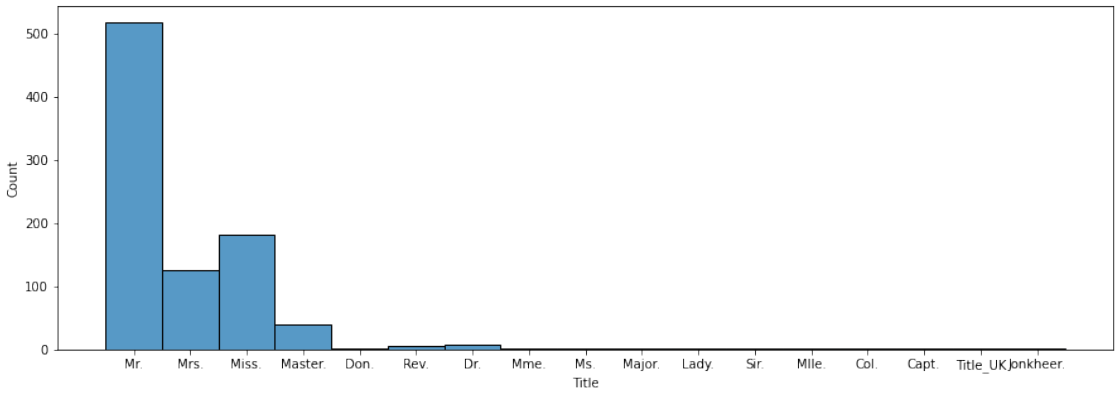

100
80
60
40
20
0
Count
0
10
20
30
40
50
60
70
80
Age


600
500
400
300
200
100
0
Count
0
1
2
3
4
5
6
7
8
SibSp


500
400
300
200
100
0
Count
Dead
Survived
Survived

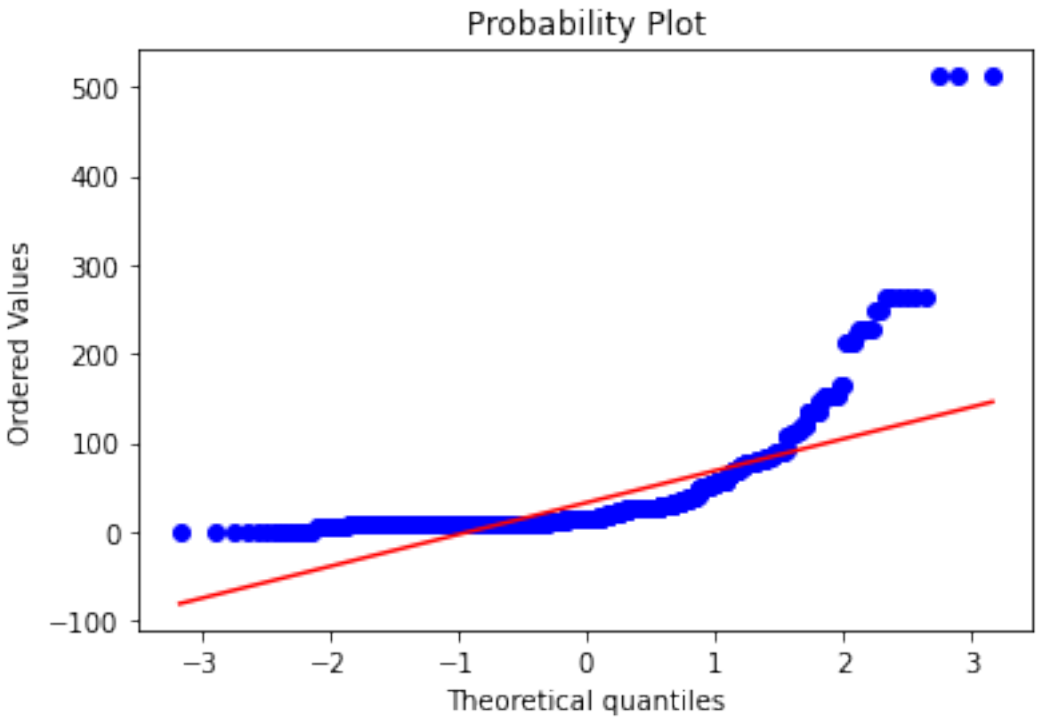

600
500
400
300
200
100
0
Count
male
female
Sex

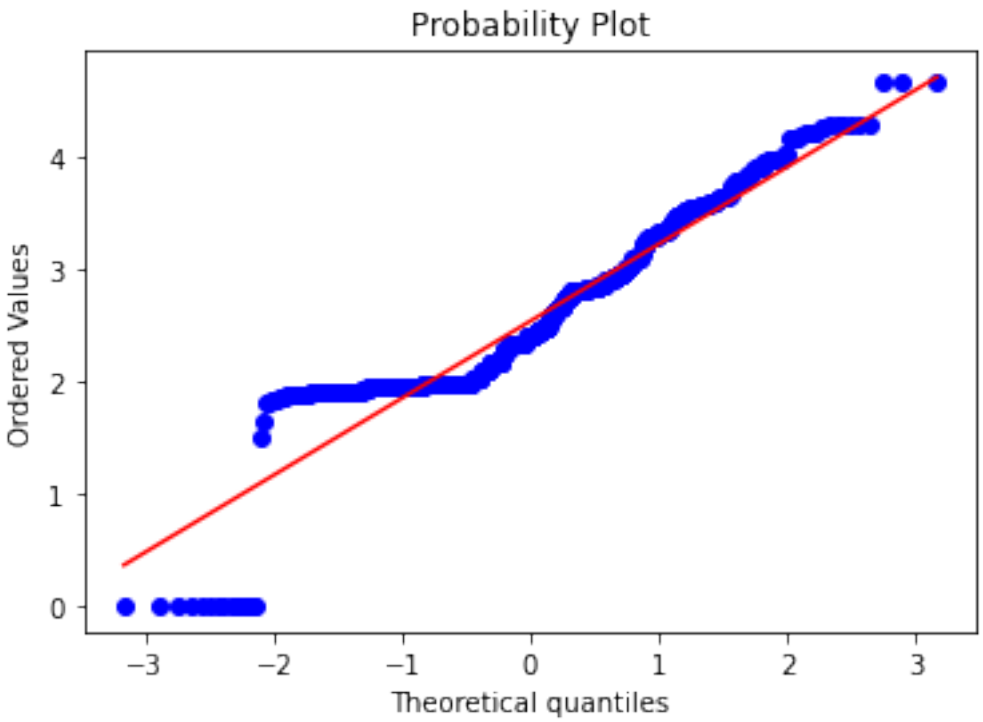

600
500
400
300
200
100
0
Count
S
C
Q
Unknown
Embarked



1.3.3 Data Transformation

From the above plots, we see that *Fare* is skewed, so plotted the Quantile Plot, and applied yeo-johnson transformation, so the transformation gets closer to normal distribution. Yeo-johnson was chosen cause log and simple power transformer were insufficient, and boxcox was also not yielding good result.





Binning the Age into 4 groups (Child, Teen, Adult, and Old) for ease of interpretation and building simpler model

1.4 Model building for generating Association Rules

Preparing the dataset for Apriori algorithm, it requires data to be in boolean form. The resulting dataset has 2003 columns

[102]:

	-0.0	0	1	1.491313199205757	1.643571232325522	\
0	False	True	True	False	False	
1	False	True	True	False	False	
2	False	True	False	False	False	
3	False	True	True	False	False	
4	False	True	False	False	False	

	1.799515862481875	1.8219481450652837	1.8233281111866173	\
0	False	False	False	
1	False	False	False	
2	False	False	False	
3	False	False	False	
4	False	False	False	

	1.8283626763046636	1.855705398817095	…	Zimmerman, Mr. Leo	\
0	False	False	…	False	

```
1              False             False  …              False
2              False             False  …              False
3              False             False  …              False
4              False             False  …              False

   de Messemaeker, Mrs. Guillaume Joseph (Emma)  de Mulder, Mr. Theodore  \
0                                         False                    False
1                                         False                    False
2                                         False                    False
3                                         False                    False
4                                         False                    False

   de Pelsmaeker, Mr. Alfons  del Carlo, Mr. Sebastiano  female   male    nan  \
0                      False                      False   False   True   True
1                      False                      False    True  False  False
2                      False                      False    True  False   True
3                      False                      False    True  False  False
4                      False                      False   False   True   True

   van Billiard, Mr. Austin Blyler  van Melkebeke, Mr. Philemon
0                            False                        False
1                            False                        False
2                            False                        False
3                            False                        False
4                            False                        False

[5 rows x 2003 columns]
```

Using Apriori algorithm on the dataset, and the result has 263 itemset that have support greater than our *min_support=0.2*

[103]:
```
    support itemsets
0  0.840449      (0)
1  0.444944      (1)
2  0.296629      (2)
3  0.557303      (3)
4  0.683146  (Adult)
```

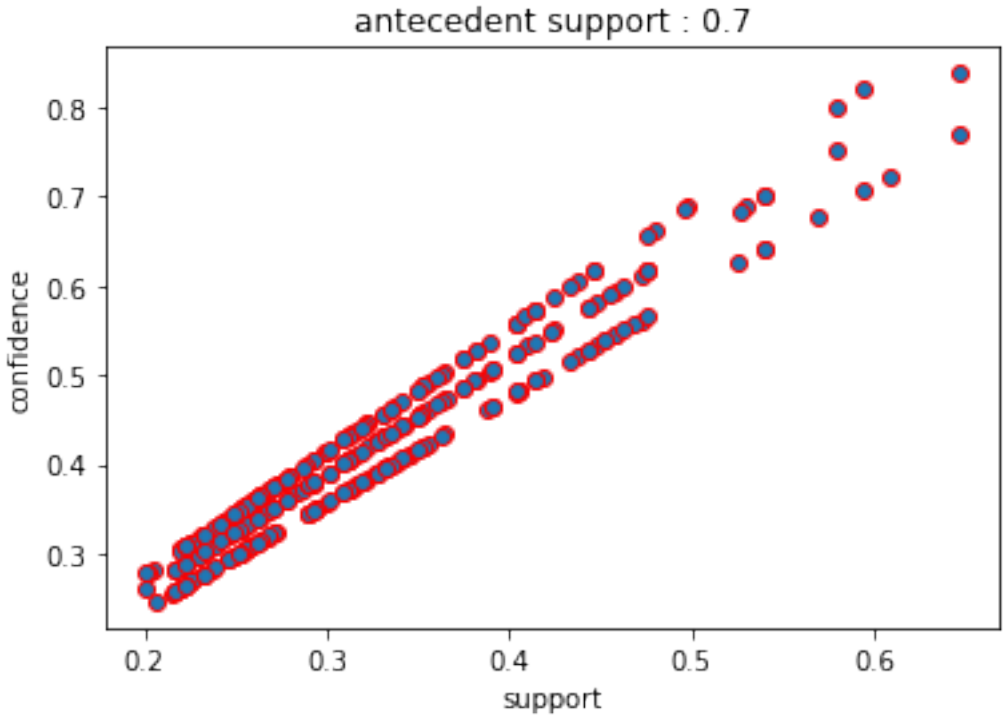
Association Rules:

```
['antecedents' 'consequents' 'antecedent support' 'consequent support'
 'support' 'confidence' 'lift' 'leverage' 'conviction']
------------------------------------
Configuration :  antecedent support  :  0.7
------------------------------------
   antecedents                      consequents  antecedent support  \
0          (0)                              (1)            0.840449
1          (0)                              (2)            0.840449
```

```
2           (0)                                (3)           0.840449
3           (0)                           (Adult)            0.840449
4           (0)                            (Dead)            0.840449
..          …                                 …                …
345         (0)      (Mr., nan, 3, male, Dead, S)            0.840449
346         (S)      (Mr., nan, 3, 0, male, Dead)            0.723596
347       (nan)    (Mr., Adult, 0, male, Dead, S)            0.770787
348         (0)  (Mr., nan, Adult, male, Dead, S)            0.840449
349         (S)  (Mr., nan, Adult, 0, male, Dead)            0.723596
```

	consequent support	support	confidence	lift	leverage	conviction
0	0.444944	0.316854	0.377005	0.847310	-0.057099	0.890949
1	0.296629	0.206742	0.245989	0.829282	-0.042560	0.932839
2	0.557303	0.458427	0.545455	0.978739	-0.009958	0.973933
3	0.683146	0.607865	0.723262	1.058722	0.033715	1.144960
4	0.615730	0.525843	0.625668	1.016140	0.008352	1.026549
..	…	…	…	…	…	…
345	0.241573	0.222472	0.264706	1.095759	0.019442	1.031461
346	0.289888	0.222472	0.307453	1.060595	0.012711	1.025364
347	0.264045	0.232584	0.301749	1.142795	0.029062	1.053998
348	0.248315	0.232584	0.276738	1.114465	0.023888	1.039299
349	0.293258	0.232584	0.321429	1.096059	0.020384	1.041514

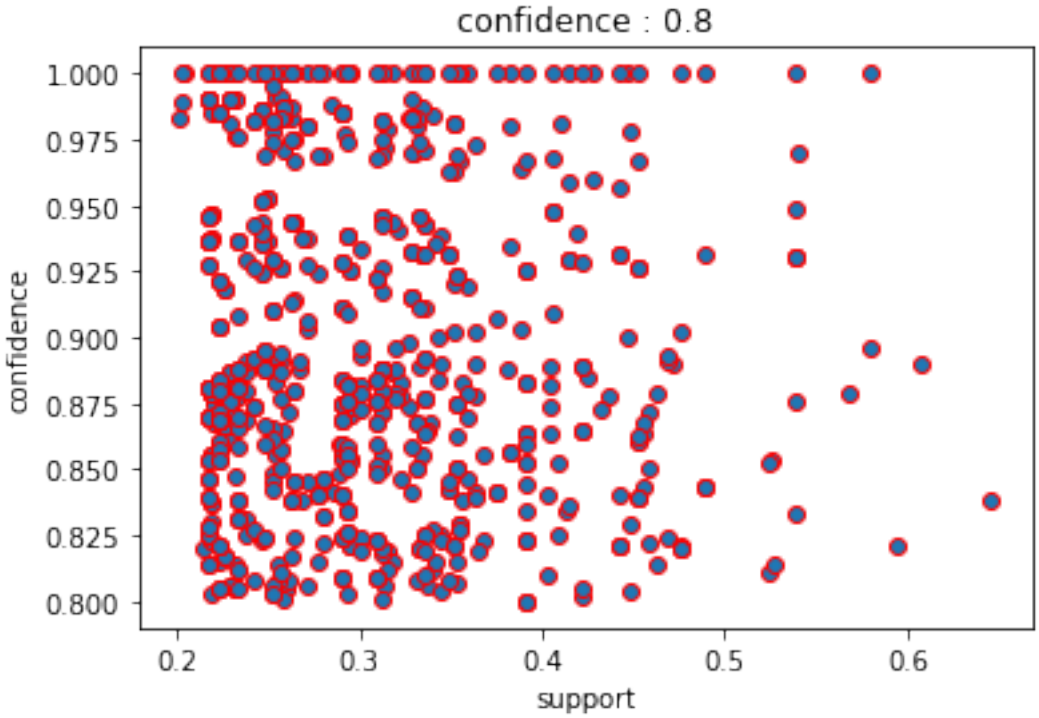
[350 rows x 9 columns]



```
['antecedents' 'consequents' 'antecedent support' 'consequent support'
 'support' 'confidence' 'lift' 'leverage' 'conviction']
-------------------------------------
Configuration :  confidence  :  0.8
-------------------------------------
                   antecedents      consequents  antecedent support  \
0                          (3)              (0)            0.557303
1                      (Adult)              (0)            0.683146
2                       (Dead)              (0)            0.615730
3                        (Mr.)              (0)            0.579775
4                          (S)              (0)            0.723596
..                         …                …                …
800  (Adult, 0, male, Dead, S)       (Mr., nan)            0.270787
801       (Mr., S, nan, Adult)  (0, male, Dead)            0.277528
802      (Mr., S, Adult, Dead)   (nan, male, 0)            0.279775
803      (nan, male, S, Adult)   (Mr., 0, Dead)            0.286517
804     (male, S, Adult, Dead)    (Mr., 0, nan)            0.288764

     consequent support   support  confidence      lift  leverage  conviction
0              0.840449  0.458427    0.822581  0.978739 -0.009958    0.899285
1              0.840449  0.607865    0.889803  1.058722  0.033715    1.447862
2              0.840449  0.525843    0.854015  1.016140  0.008352    1.092921
3              0.840449  0.539326    0.930233  1.106828  0.052054    2.286891
4              0.840449  0.594382    0.821429  0.977368 -0.013763    0.893483
..                  …       …           …       …       …          …
800            0.475281  0.232584    0.858921  1.807186  0.103885    3.719332
801            0.468539  0.232584    0.838057  1.788658  0.102551    3.281770
802            0.462921  0.232584    0.831325  1.795824  0.103070    3.184109
803            0.452809  0.232584    0.811765  1.792731  0.102847    2.906952
804            0.442697  0.232584    0.805447  1.819412  0.104749    2.864539

[805 rows x 9 columns]
```

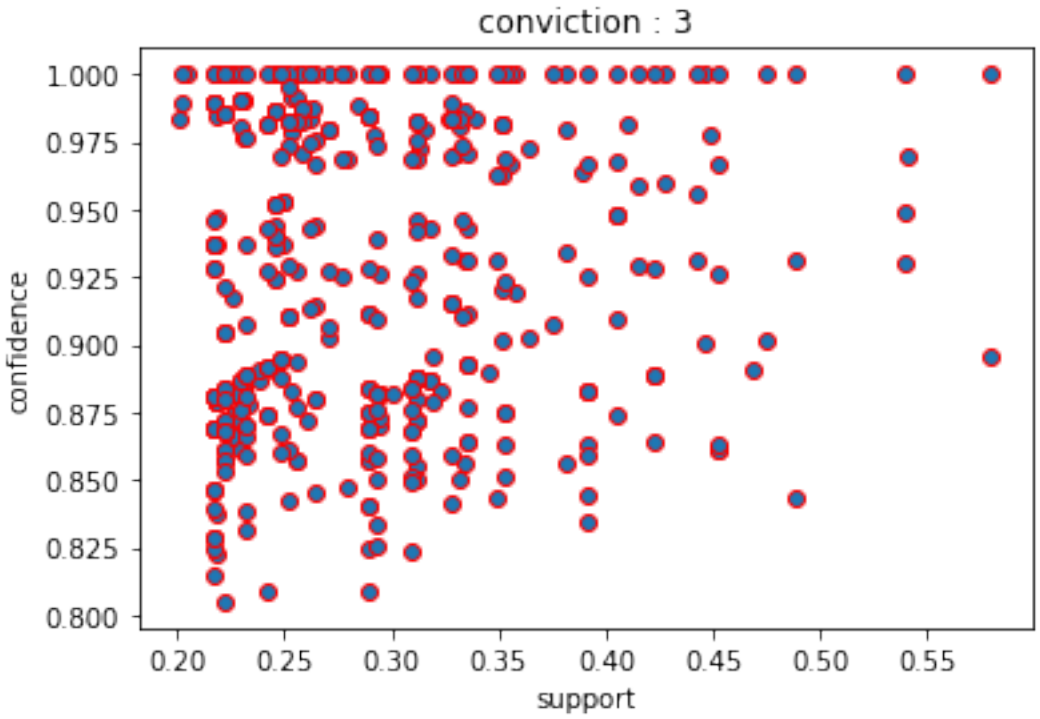



```
['antecedents' 'consequents' 'antecedent support' 'consequent support'
 'support' 'confidence' 'lift' 'leverage' 'conviction']
------------------------------------
Configuration :  conviction  :  3
------------------------------------
                     antecedents      consequents  antecedent support  \
0                            (3)            (nan)            0.557303
1                       (Miss.)         (female)            0.204494
2                         (Mr.)           (male)            0.579775
3                         (male)           (Mr.)            0.647191
4                         (3, 0)            (nan)            0.458427
..                           …              …                 …
455     (nan, Adult, 0, Dead, S)      (Mr., male)            0.267416
456  (nan, Adult, male, Dead, S)         (Mr., 0)            0.256180
457    (Adult, 0, male, Dead, S)       (Mr., nan)            0.270787
458         (Mr., S, nan, Adult)  (0, male, Dead)            0.277528
459        (Mr., S, Adult, Dead)   (nan, male, 0)            0.279775

     consequent support   support  confidence      lift  leverage  conviction
0              0.770787  0.540449    0.969758  1.258141  0.110888    7.579326
1              0.352809  0.204494    1.000000  2.834395  0.132347         inf
2              0.647191  0.579775    1.000000  1.545139  0.204550         inf
3              0.579775  0.579775    0.895833  1.545139  0.204550    4.034157
4              0.770787  0.448315    0.977941  1.268758  0.094965   10.391011
```

```
..          ...       ...       ...       ...       ...       ...
455    0.579775  0.232584  0.869748  1.500147  0.077543  3.226241
456    0.539326  0.232584  0.907895  1.683388  0.094420  5.001605
457    0.475281  0.232584  0.858921  1.807186  0.103885  3.719332
458    0.468539  0.232584  0.838057  1.788658  0.102551  3.281770
459    0.462921  0.232584  0.831325  1.795824  0.103070  3.184109

[460 rows x 9 columns]
```



[107]: (263, 2)