TITANIC

Marie Bai

INTRO

IMPORT-
ANT

Knn,
DA,
LR

Transfor-
mations

Missing
values

INTRO

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

'Pclass_1',
'Pclass_2',
'Pclass_3',

'female',
'male',

'cabin_a',
...
'cabin_t'

'embarked_s',
'embarked_c',
'embarked_q'

	Survived	SibSp	Parch	Fare	age_mean	cabin_a	cabin_b	cabin_c	cabin_d	cabin_e	...	cabin_g	cabin_t	Pclass_1	Pclass_2	Pclass_3	female
0	0	1	0	7.2500	22.000000	0	0	0	0	0	...	0	0	0	0	1	0
1	1	1	0	71.2833	38.000000	0	0	1	0	0	...	0	0	1	0	0	1
2	1	0	0	7.9250	26.000000	0	0	0	0	0	...	0	0	0	0	1	1
3	1	1	0	53.1000	35.000000	0	0	1	0	0	...	0	0	1	0	0	1
4	0	0	0	8.0500	35.000000	0	0	0	0	0	...	0	0	0	0	1	0

Missing values

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

AGE

EMBARKED

CABIN

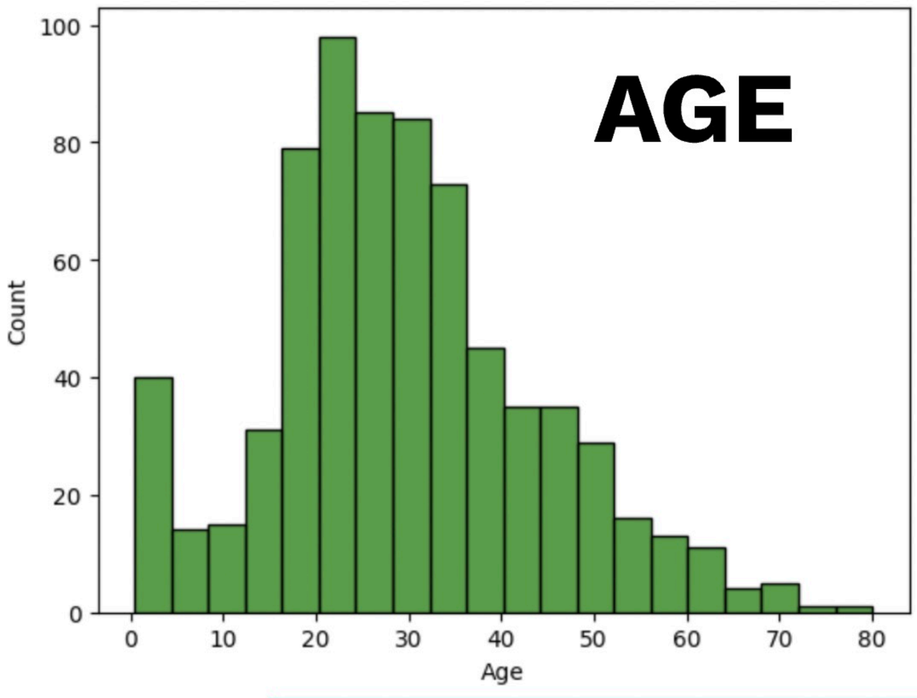
Missing values

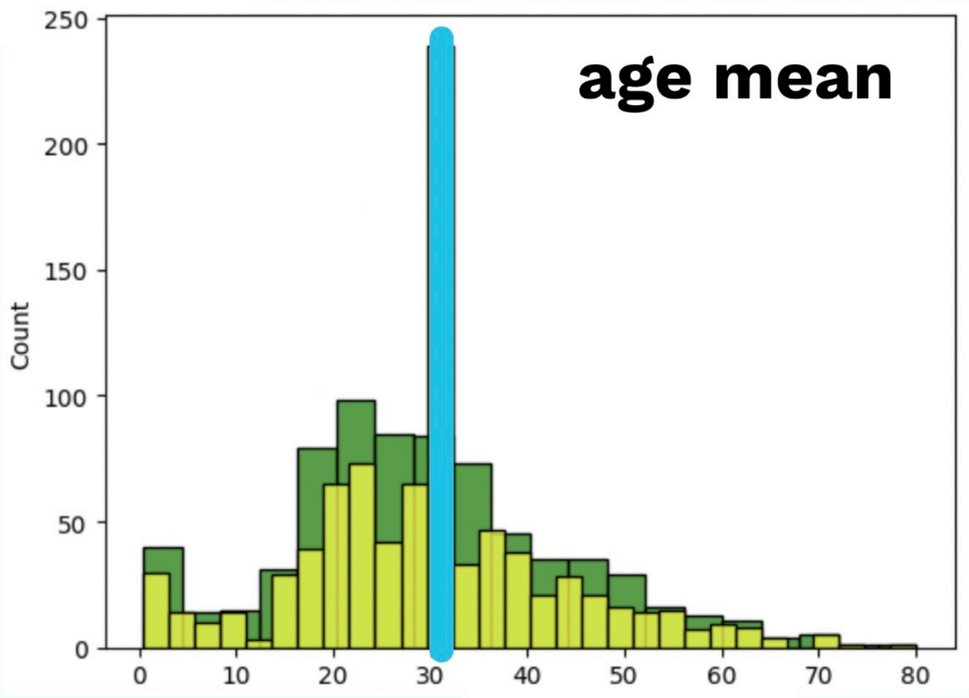
PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

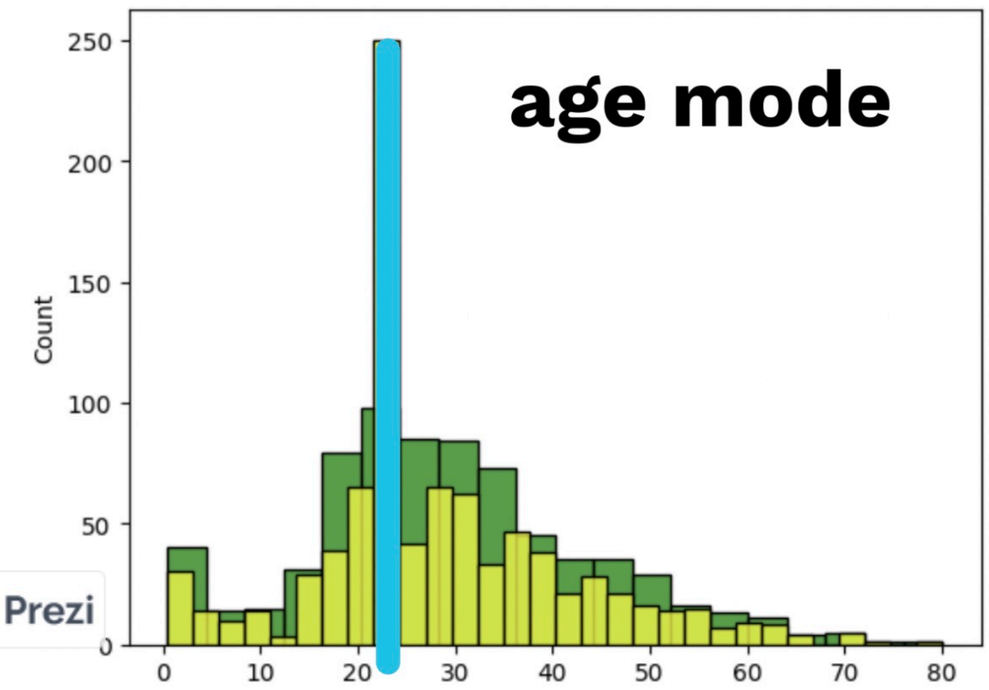
AGE

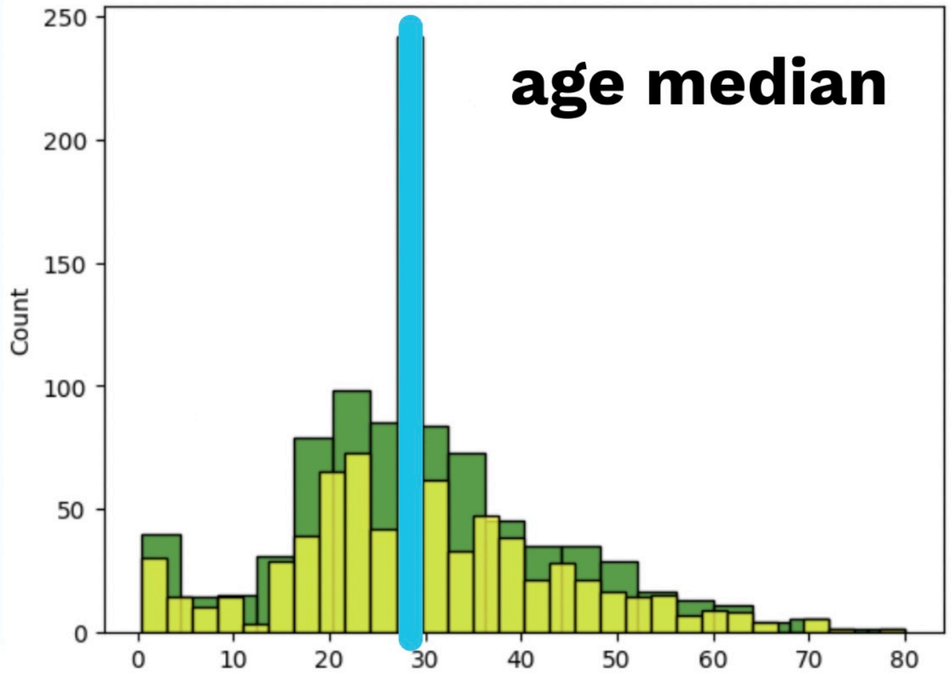
EMBARKED

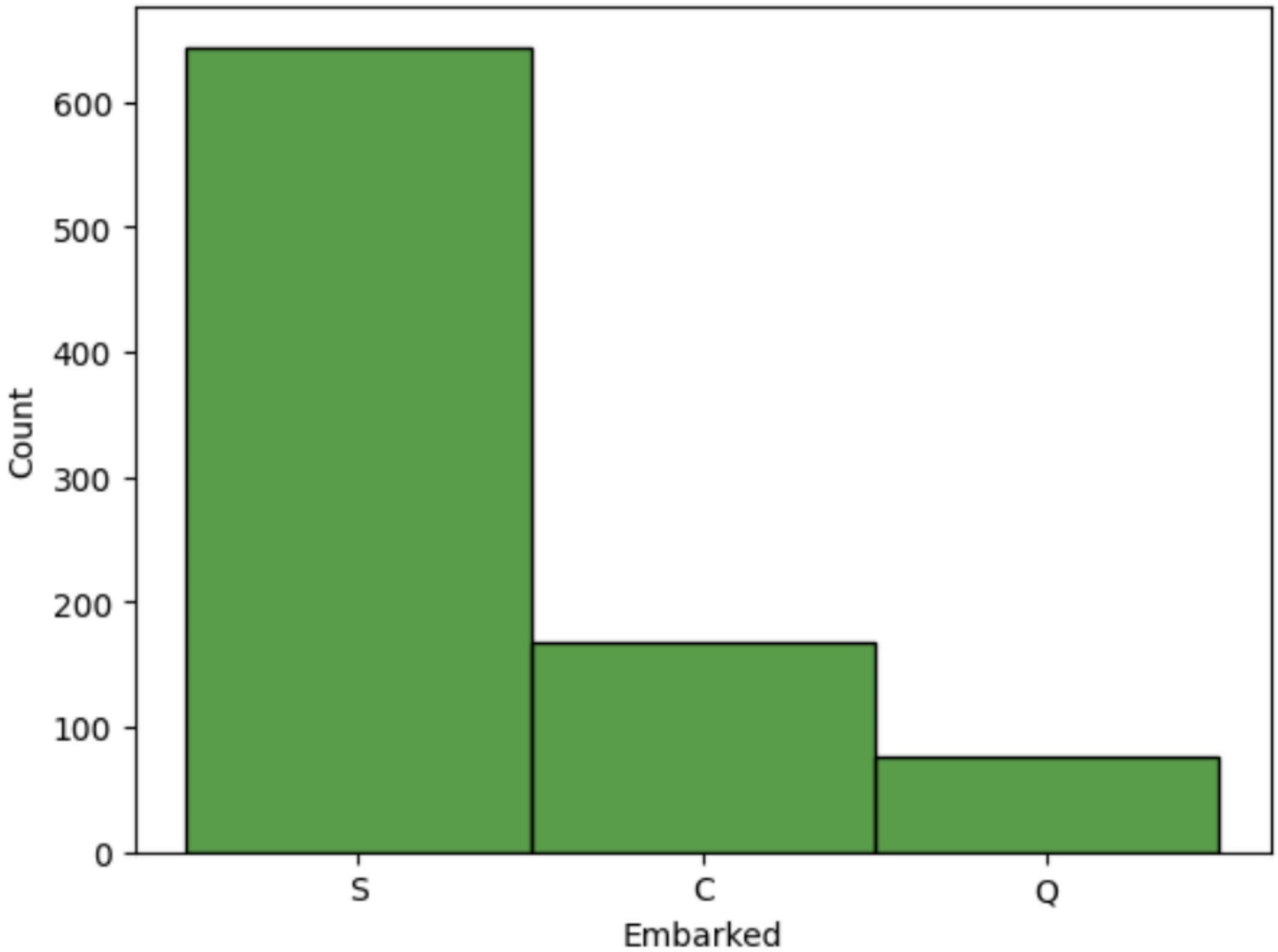
CABIN

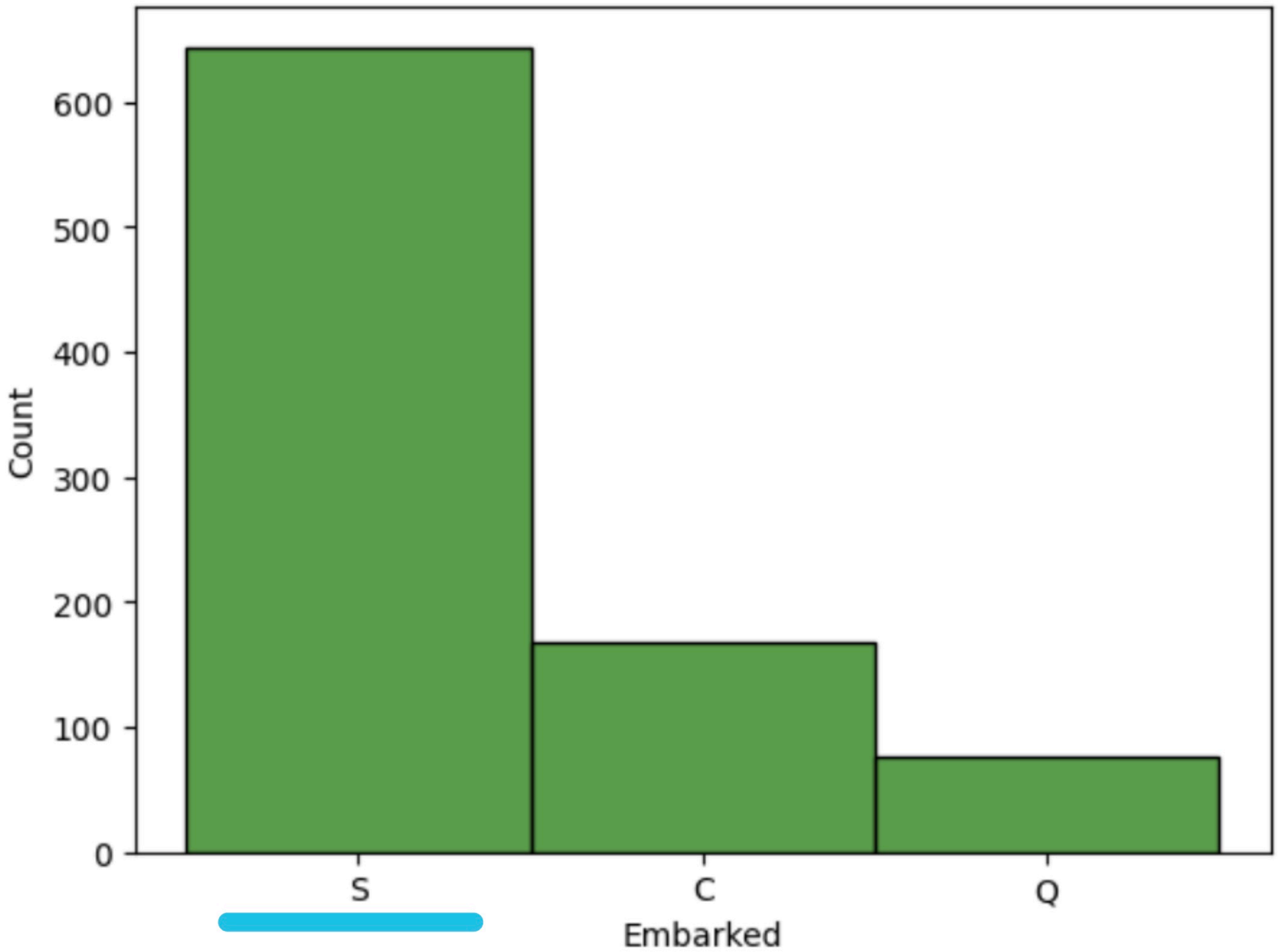

AGE
Count
Age
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80


age mean
Count
0
50
100
150
200
250
0
10
20
30
40
50
60
70
80


age mode
Count
0
50
100
150
200
250
0
10
20
30
40
50
60
70
80


age median
Count
0
50
100
150
200
250
0
10
20
30
40
50
60
70
80


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


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

Cabin
687 missing values

```
nan, nan, nan, nan, nan, 'D33', nan, 'B30', 'C52', nan, nan, nan,
nan, nan, 'B28', 'C83', nan, nan, nan, 'F33', nan, nan, nan, nan,
nan, nan, nan, nan, 'F G73', nan, nan, nan, nan, nan, nan, nan,
nan, nan, nan, nan, nan, 'C23 C25 C27', nan, nan, nan, 'E31', nan,
nan, nan, 'A5', 'D10 D12', nan, nan, nan, nan, 'D26', nan, nan,
nan, nan, nan, nan, nan, 'C110', nan, nan, nan, nan, nan, nan, nan,
'B58 B60', nan, nan, nan, nan, 'E101', 'D26', nan, nan, nan,
'F E69', nan, nan, nan, nan, nan, nan, nan, 'D47', 'C123', nan,
'B86', nan, nan, nan, nan, nan, nan, nan, nan, 'F2', nan, nan,
'C2', nan, nan, nan, nan, nan, nan, nan, nan, nan, nan, nan, nan,
nan, nan, 'E33', nan, nan, nan, 'B19', nan, nan, nan, 'A7', nan,
nan, 'C49', nan, nan, nan, nan, nan, 'F4', nan, 'A32', nan, nan,
nan, nan, nan, nan, nan, 'F2', 'B4', 'B80', nan, nan, nan, nan,
nan, nan, nan, nan, nan, 'G6', nan, nan, nan, 'A31', nan, nan, nan,
nan, nan, 'D36', nan, nan, 'D15', nan, nan, nan, nan, nan, 'C93',
nan, nan, nan, nan, nan, 'C83', nan, nan, nan, nan, nan, nan, nan,
nan, nan, nan, nan, nan, nan, nan, 'C78', nan, nan, 'D35', nan,
nan, 'G6', 'C87', nan, nan, nan, nan, 'B77', nan, nan, nan, nan,
'E67', 'B94', nan, nan, nan, nan, 'C125', 'C99', nan, nan, nan,
'C118', nan, 'D7', nan, nan, nan, nan, nan, nan, nan, nan, 'A19',
```

Embarked

embarked_s	embarked_c	embarked_q
1	0	0
0	1	0
1	0	0
1	0	0
1	0	0

PClass

Cabin

PClass, Sex

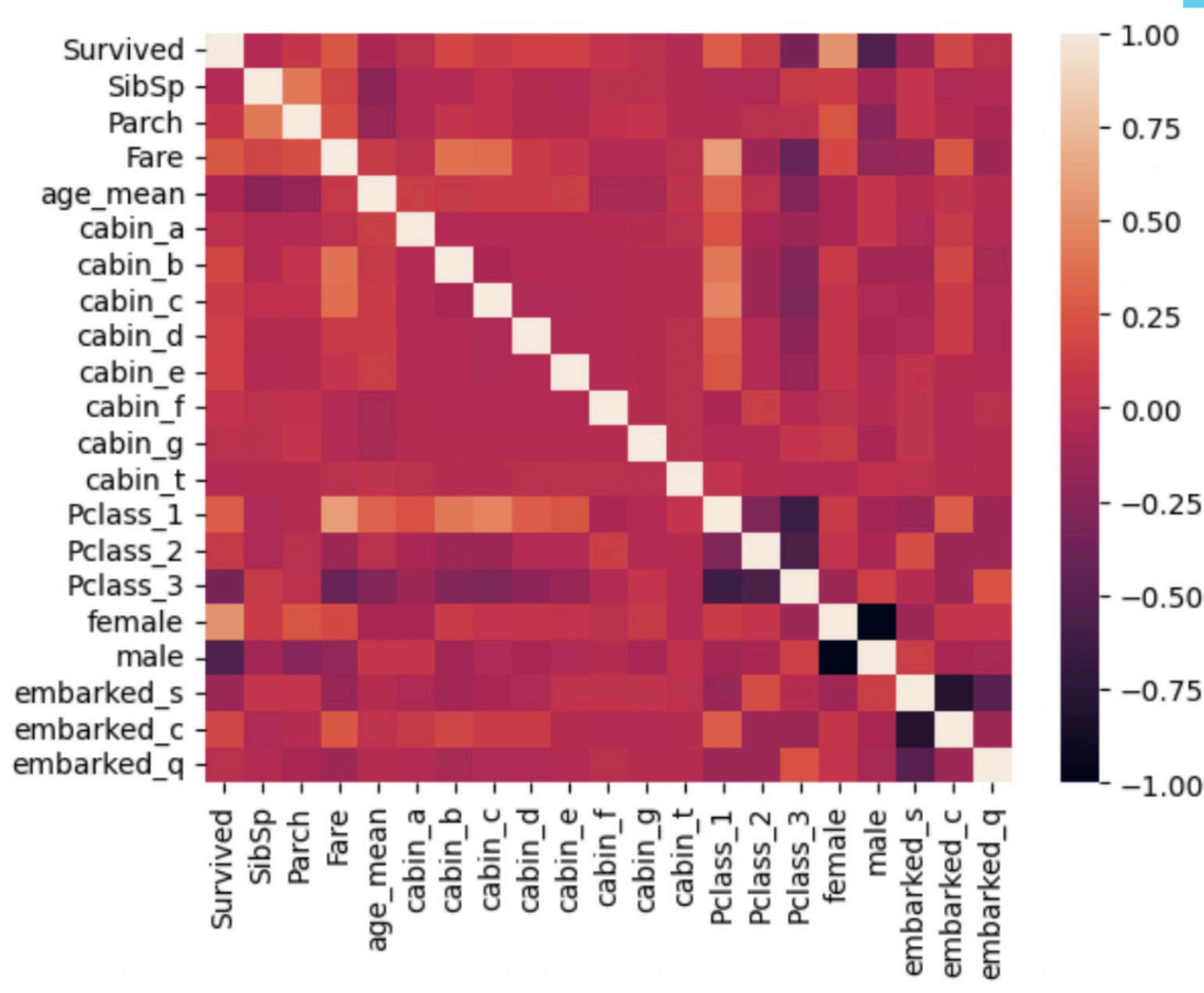
Pclass_1	Pclass_2	Pclass_3	female	male
0	0	1	0	1
1	0	0	1	0
0	0	1	1	0
1	0	0	1	0
0	0	1	0	1

Cabin

cabin_a	cabin_b	cabin_c	cabin_d	cabin_e	cabin_f	cabin_g	cabin_t
0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0

Correlation between all and survived

Survived	1.000000
female	0.543351
Pclass_1	0.285904
Fare	0.257307
cabin_b	0.175095
embarked_c	0.168240
cabin_d	0.150716
cabin_e	0.145321
cabin_c	0.114652
Pclass_2	0.093349
Parch	0.081629
cabin_f	0.057935
cabin_a	0.022287
cabin_g	0.016040
embarked_q	0.003650
cabin_t	-0.026456
SibSp	-0.035322
age_mean	-0.069809
embarked_s	-0.149683
Pclass_3	-0.322308
male	-0.543351



PREPARATION

	SibSp	Parch	Fare	age_mean	cabin_a	cabin_b	cabin_c	cabin_d	cabin_e	cabin_f	cabin_g	cabin_t	I
0	0.125	0.000000	0.014151	0.271174	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.125	0.000000	0.139136	0.472229	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
2	0.000	0.000000	0.015469	0.321438	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.125	0.000000	0.103644	0.434531	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
4	0.000	0.000000	0.015713	0.434531	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

F1 score, accuracy

Compare

F1 Score

The F1 score is a popular performance measure for classification and often preferred over

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

TP = number of true positives

FP = number of false positives

FN = number of false negatives

1st N

K VALUE

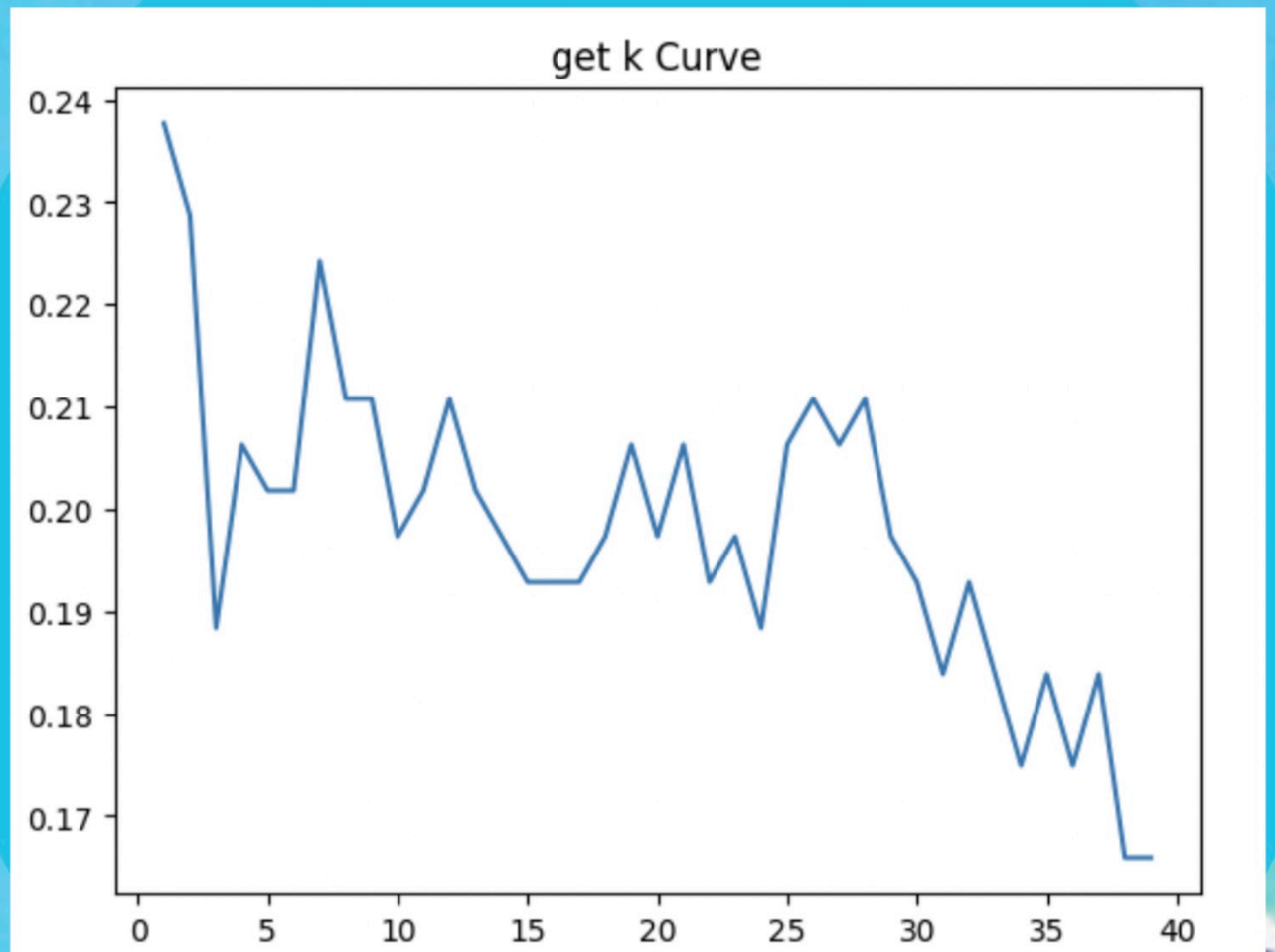
ACC, MSE

N = 10,
F1 Score = 0.713

F1 score Interpretation

> 0.9	Very good
0.8-0.9	Good
0.5-0.8	OK
< 0.5	Not good

K VALUE



N = 36,
F1 Score = 0.719

Accuracy Score

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Mean squared error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values

KNN Accuracy = 0.825

KNN MSE = 0.175

COMPARE

Accuracy KNN = 0.834
Accuracy LDA = 0.807
Accuracy LR = 0.780

KNN mean_squared_error = 0.175
LDA mean_squared_error = 0.193
LR mean_squared_error = 0.2197