Exploratory Data Analysis (EDA) and Logistic Regression Model on the Titanic dataset

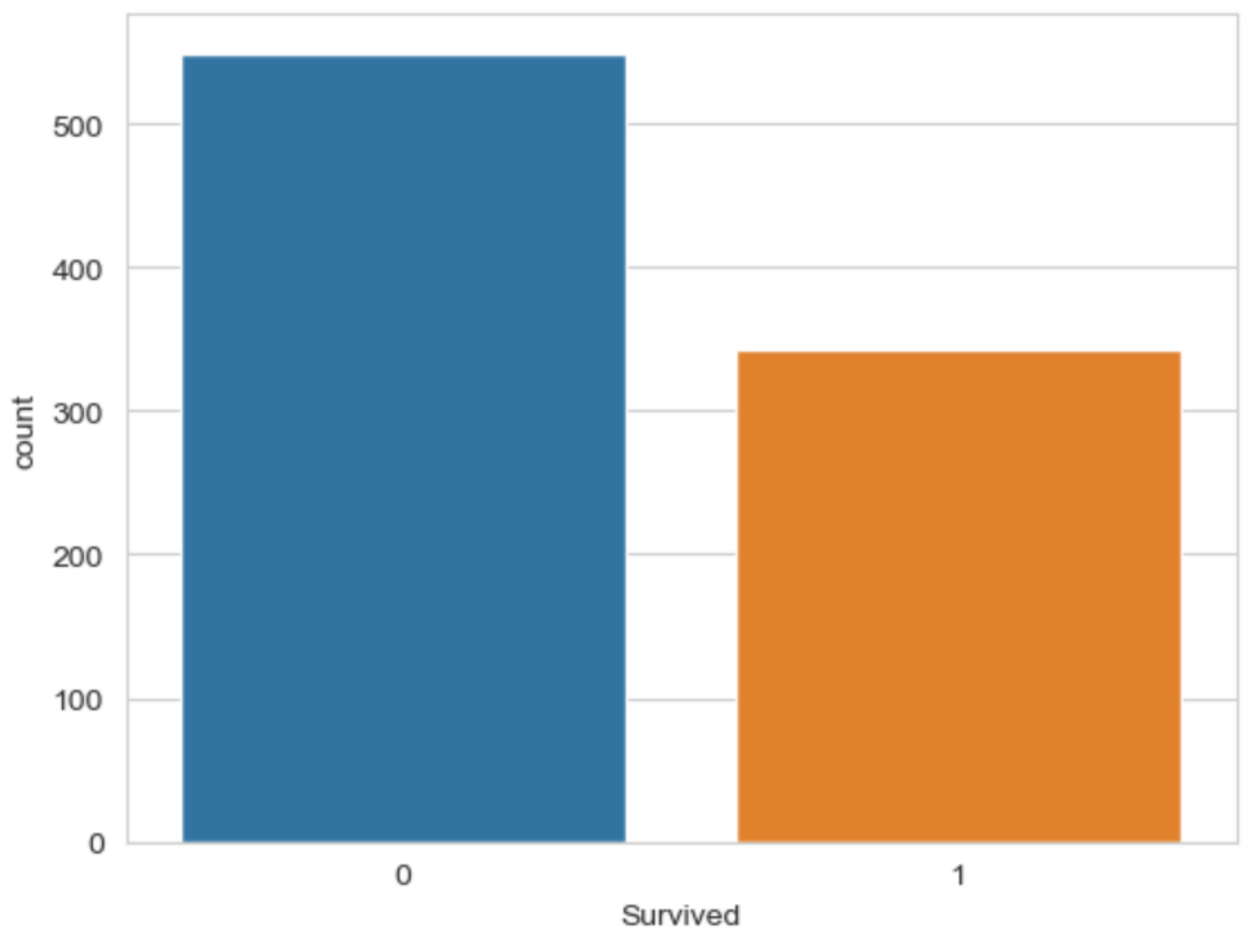
Lara Onipede and Lat Leger

Overview of the Titanic Dataset

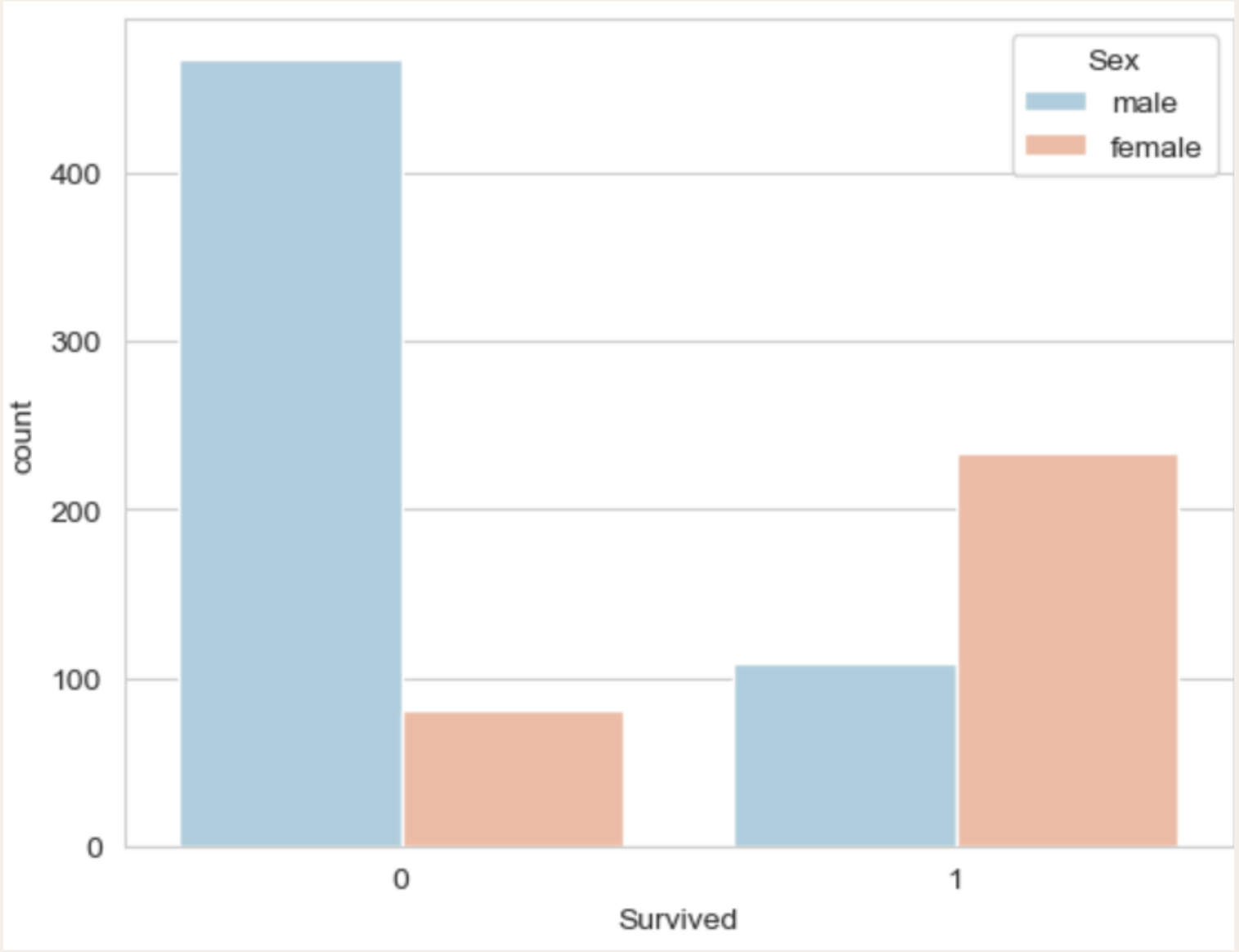
The titanic dataset in brief

Features of the Dataset

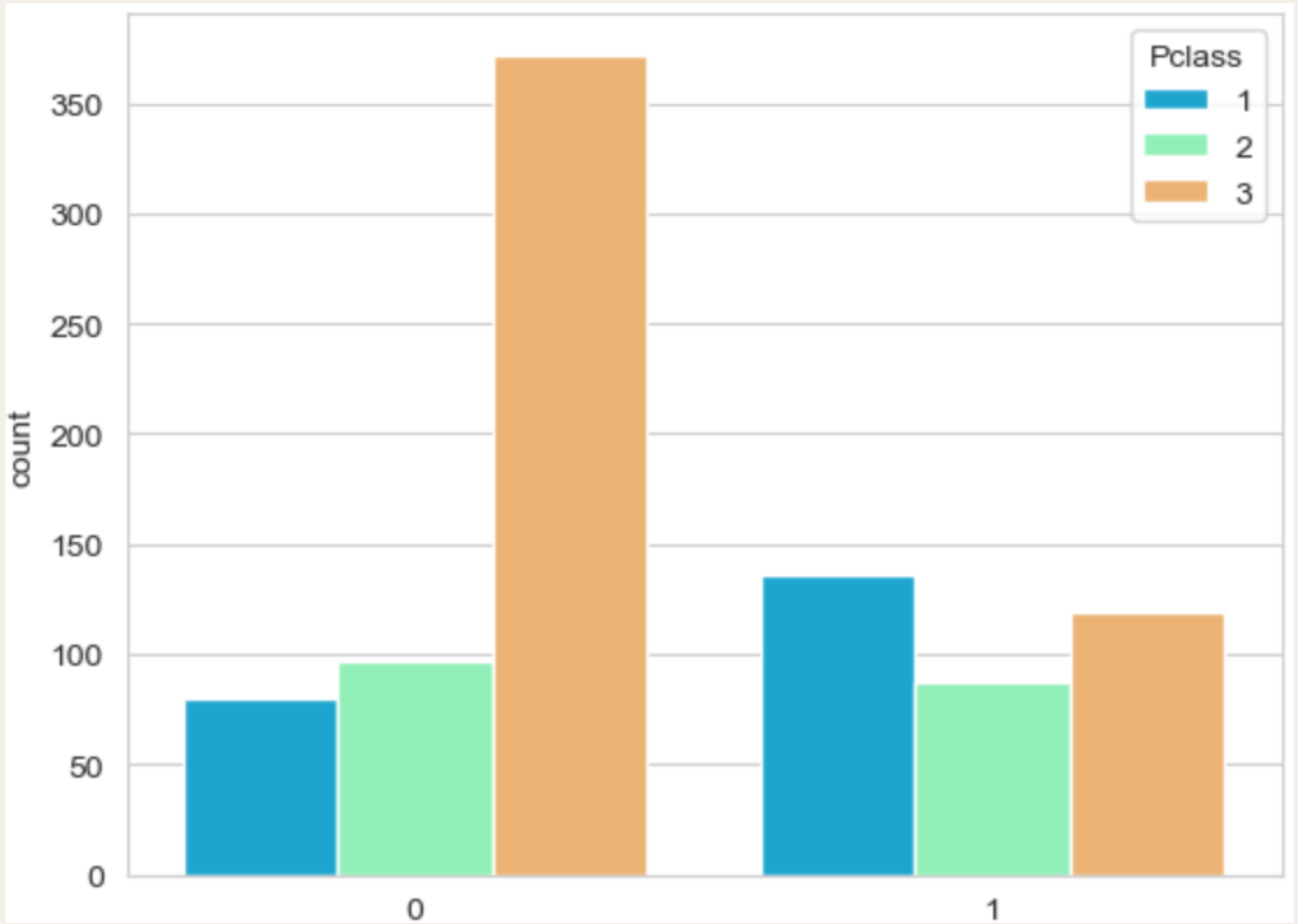
Visual EDA



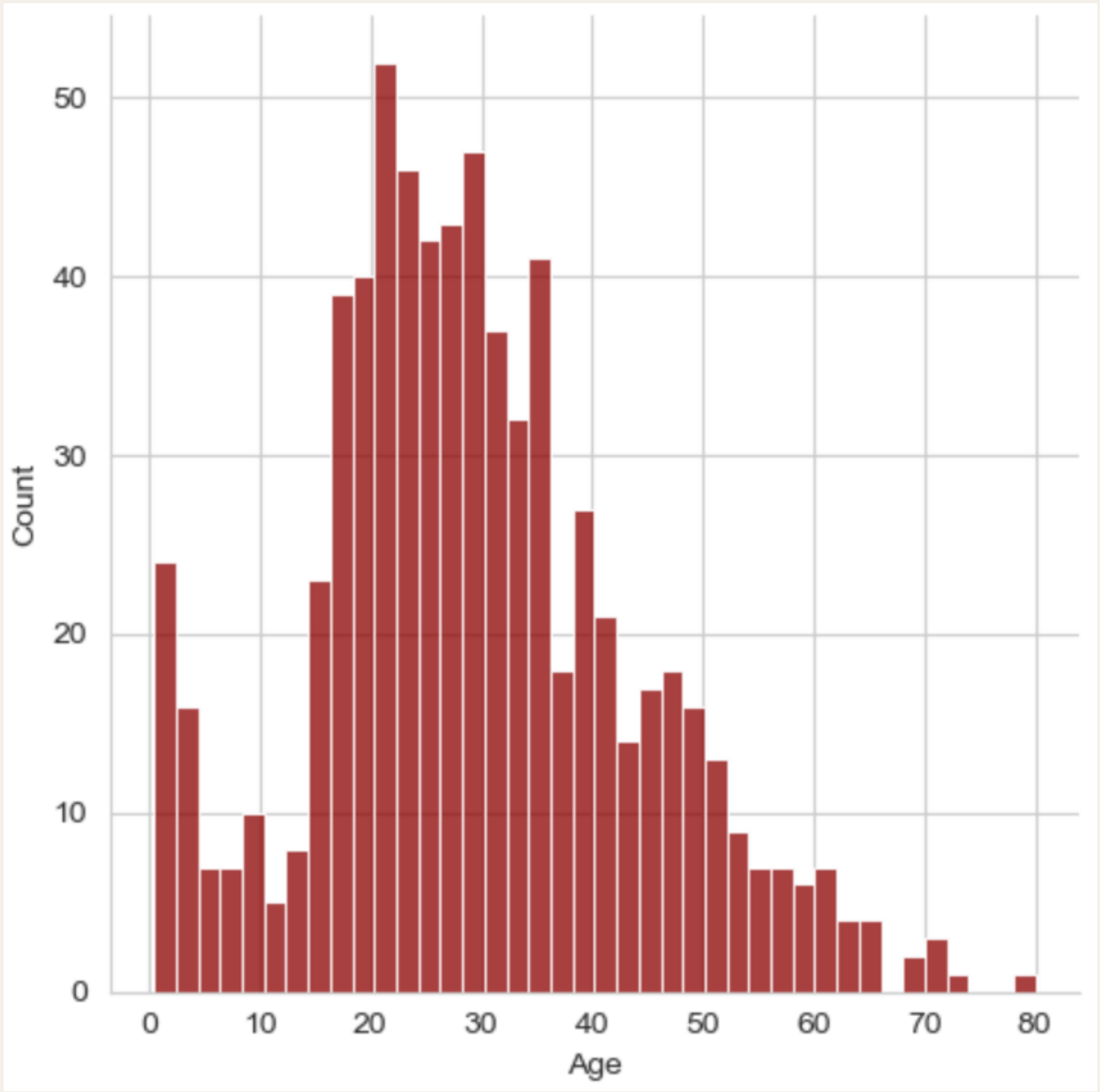
Count of survived passengers



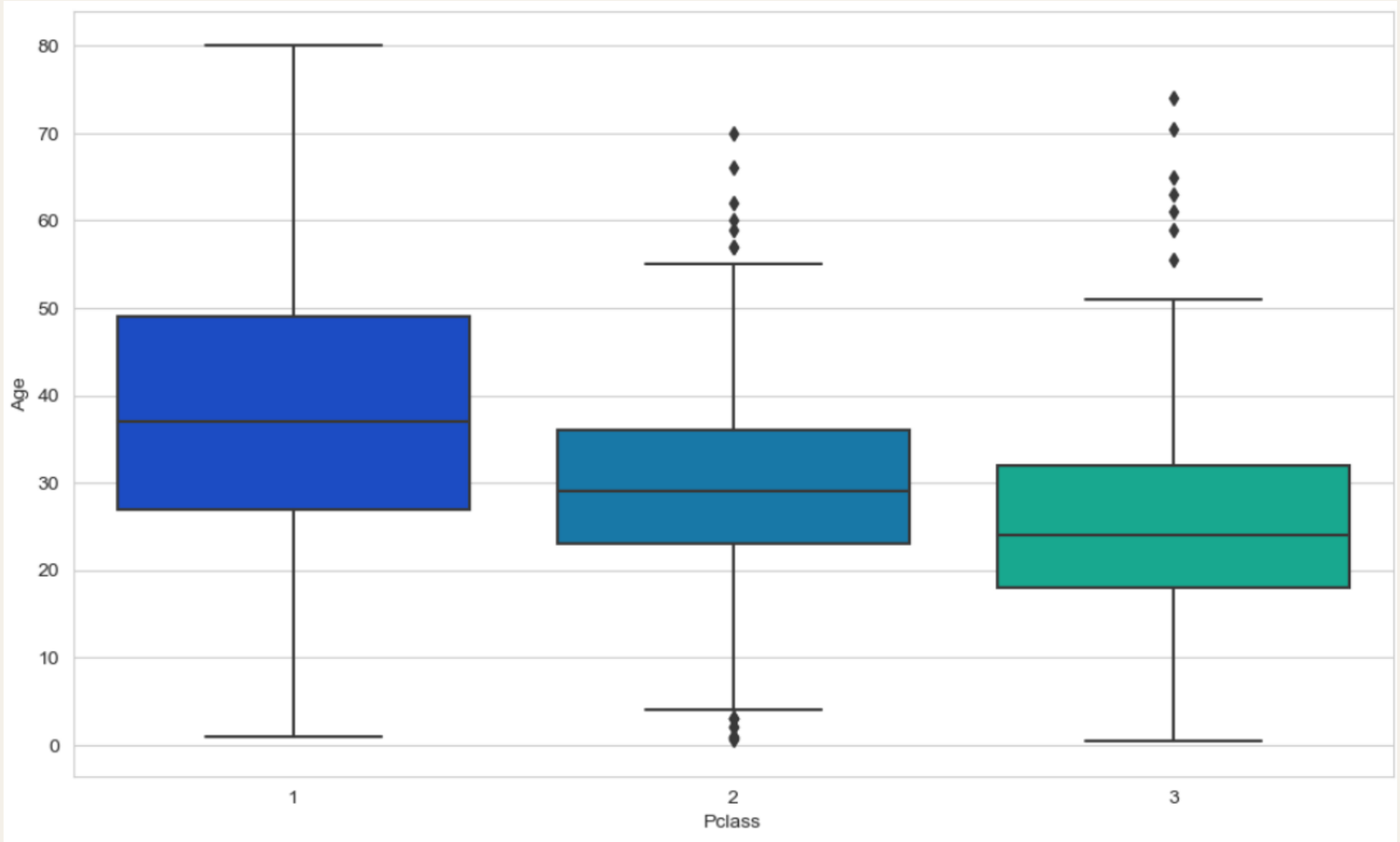
Gender Count of survived



Count of survived per passenger Class



Age distribution of passengers



Average age of passengers by Ticket Class

Logistic Regression Model

Logit Regression Results

Dep. Variable:	Survived	No. Observations:	889
Model:	Logit	Df Residuals:	884
Method:	MLE	Df Model:	4
Date:	Fri, 14 Jul 2023	Pseudo R-squ.:	0.3220
Time:	11:41:52	Log-Likelihood:	-401.00
converged:	True	LL-Null:	-591.41
Covariance Type:	nonrobust	LLR p-value:	3.885e-81

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	4.9200	0.538	9.153	0.000	3.866	5.974
male	-2.5966	0.188	-13.822	0.000	-2.965	-2.228
Age	-0.0367	0.008	-4.762	0.000	-0.052	-0.022
Fare	0.0003	0.002	0.164	0.870	-0.004	0.004
Pclass	-1.2272	0.142	-8.636	0.000	-1.506	-0.949

Logistic Regression Model

	precision	recall	f1-score	support
0	0.82	0.91	0.86	163
1	0.84	0.68	0.75	104
accuracy			0.82	267
macro avg	0.83	0.80	0.81	267
weighted avg	0.83	0.82	0.82	267

Conclusions

The logistic regression model is perfect for datasets with categorical features.

While we look forward to into ML, the titanic dataset is good and handy for beginners to practice.

Thanks

Lara Onipede and Lat Leger