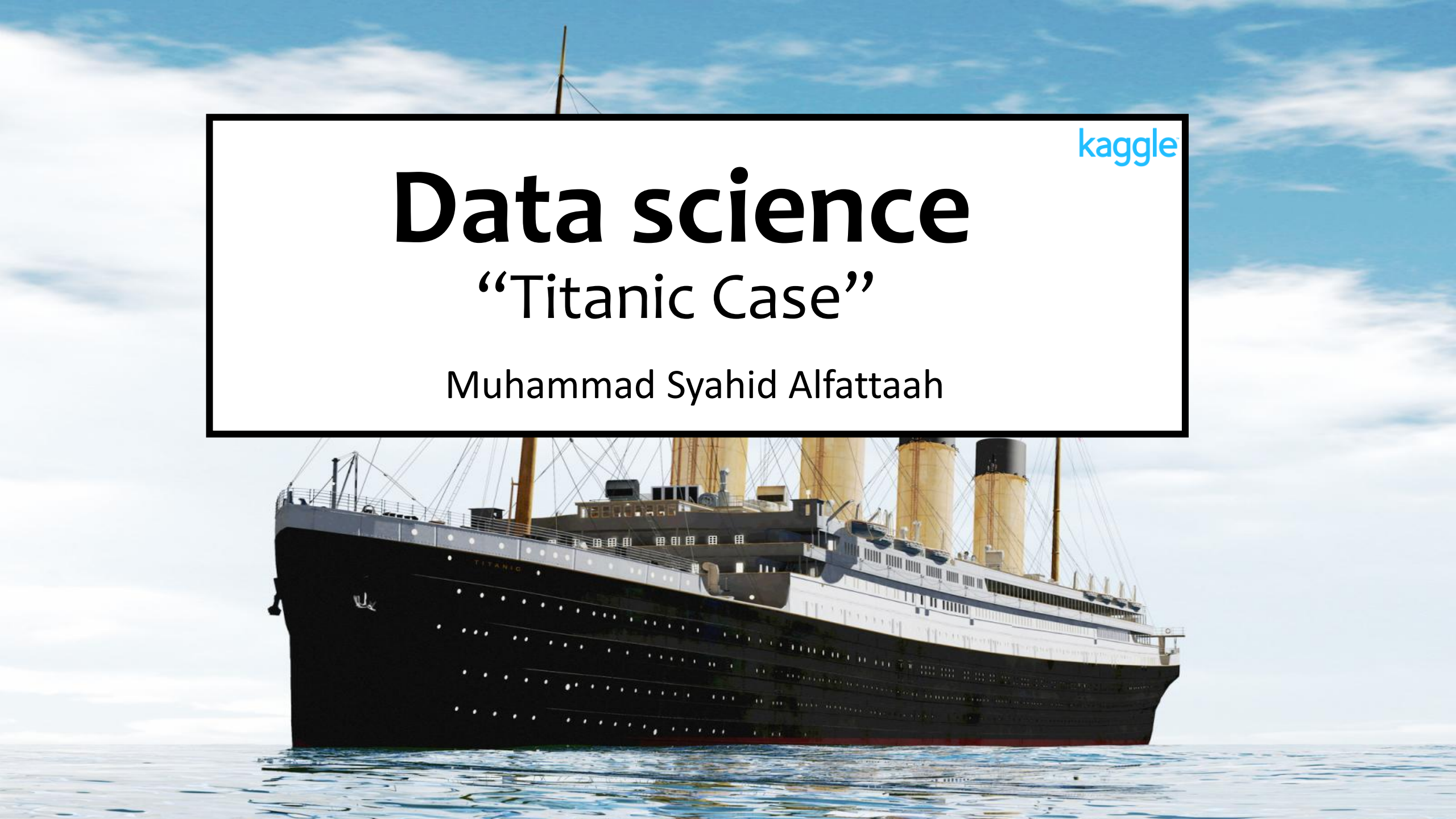
Data science

“Titanic Case”

Muhammad Syahid Alfattaah



Research Domain Knowledge

The Titanic, which was the largest and most expensive passenger ship of its time, sank on April 15, 1912 after hitting an iceberg in the North Atlantic Ocean. Of the 2,224 people on board, around 1,514 died in the disaster.

What sorts of people were more likely to survive?

Using passenger data (ie name, age, gender, socio-economic class, etc).

Understanding Every Feature In Data

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

The Data Have **891 rows X 12 Column**

Kode	Arti
Age	Umur Penumpang
SibSp	Jumlah Saudara Atau Pasangan
Parch	Jumlah Orang Tua atau Anak
Ticket	Nomor Ticket
Fare	Tarif
Cabin	Nomor Kamar
Embarked	Pelabuhan Pemberangkatan C = Cherbourg Q = Queenstown S = Southampton
Passengerld	ID Penumpang
Survived	0 = No (Meninggal) 1 = Yes (Hidup)
Pclass	1 = 1^{st} , 2 = 2^{st} , 3 = 3^{st}
Name	Nama Penumpang
Sex	Male = Laki-laki Female = Perempuan

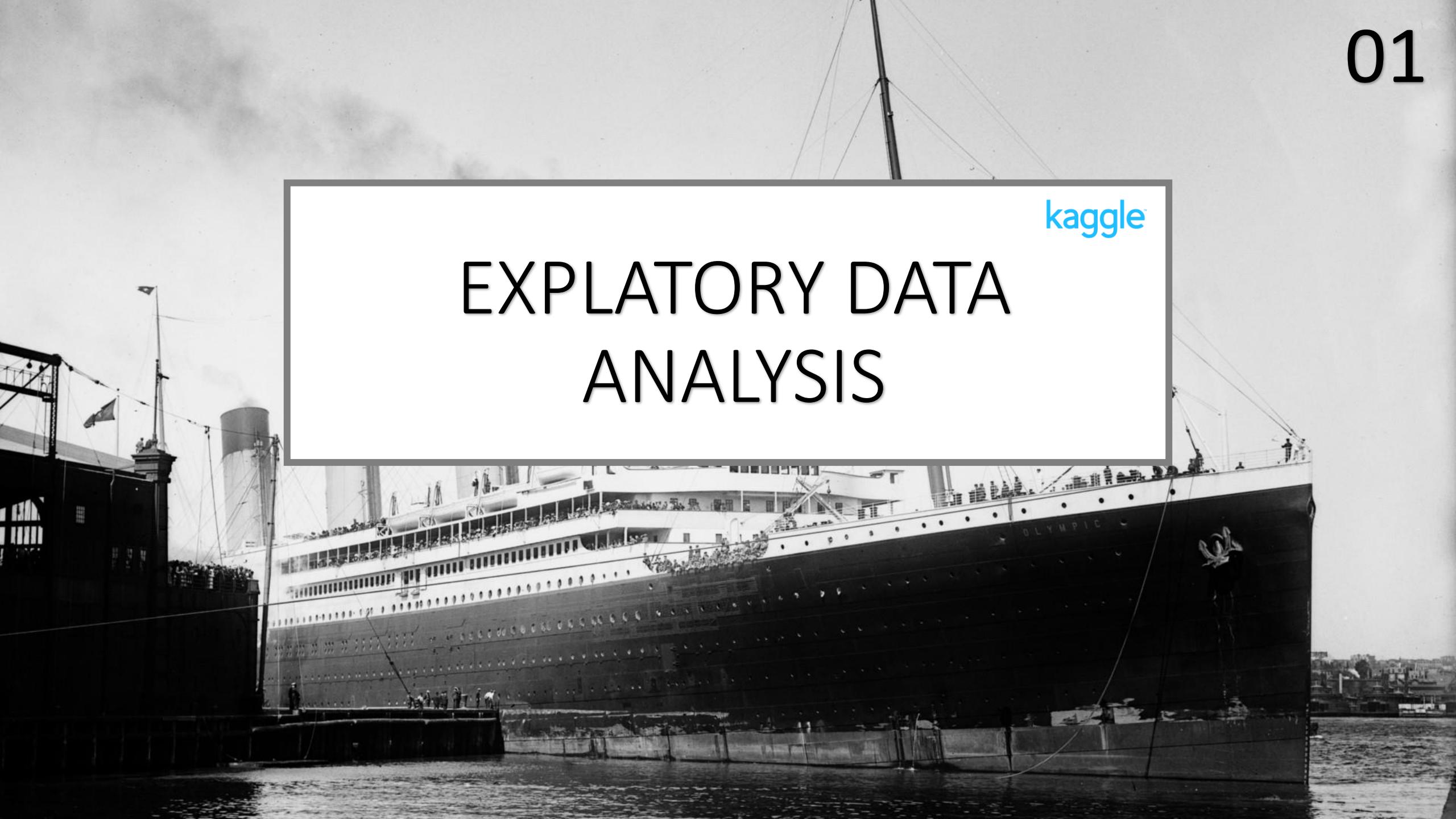
Data Cleaning

```
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
```

BEFORE

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

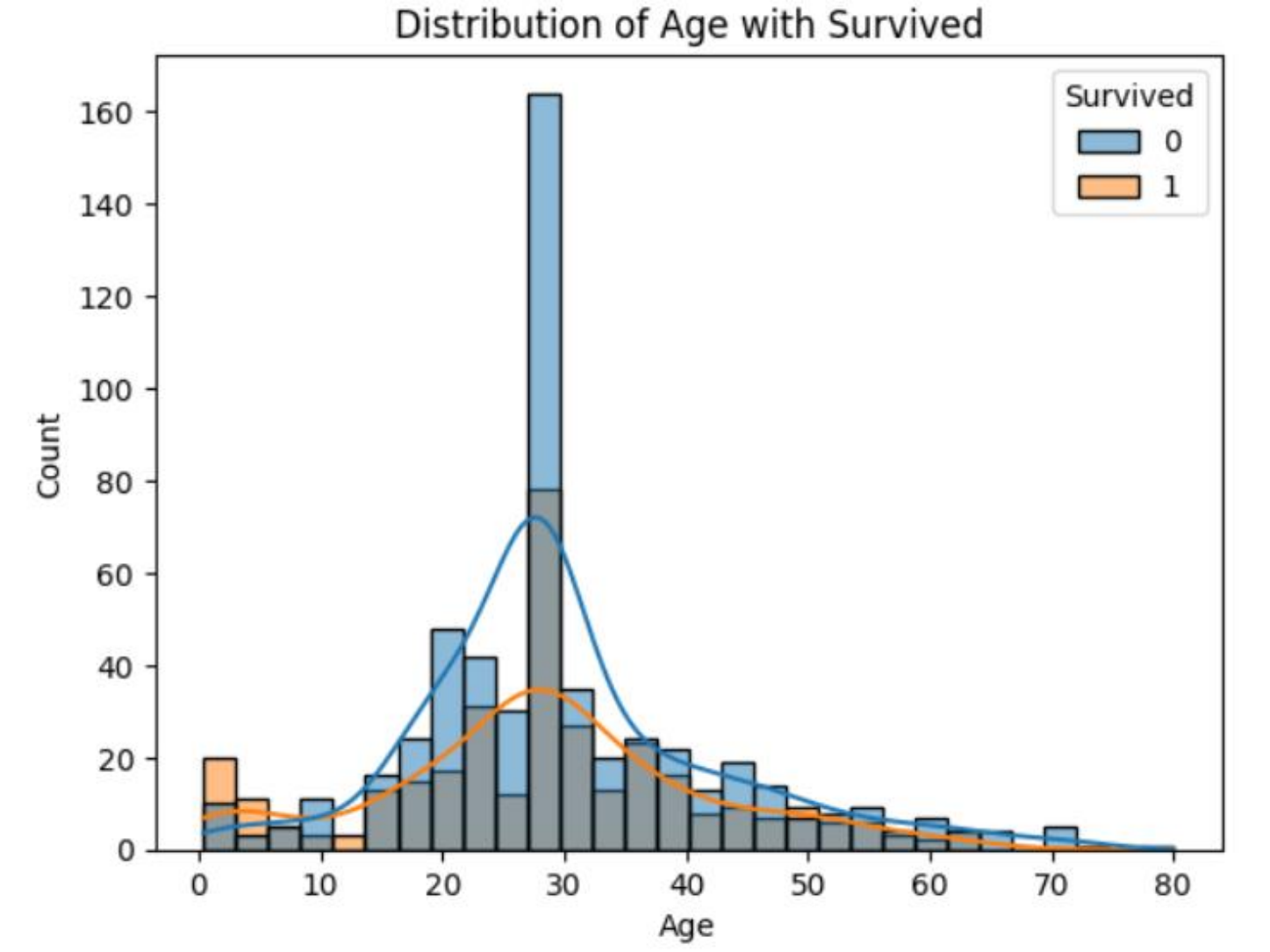
AFTER



kaggle

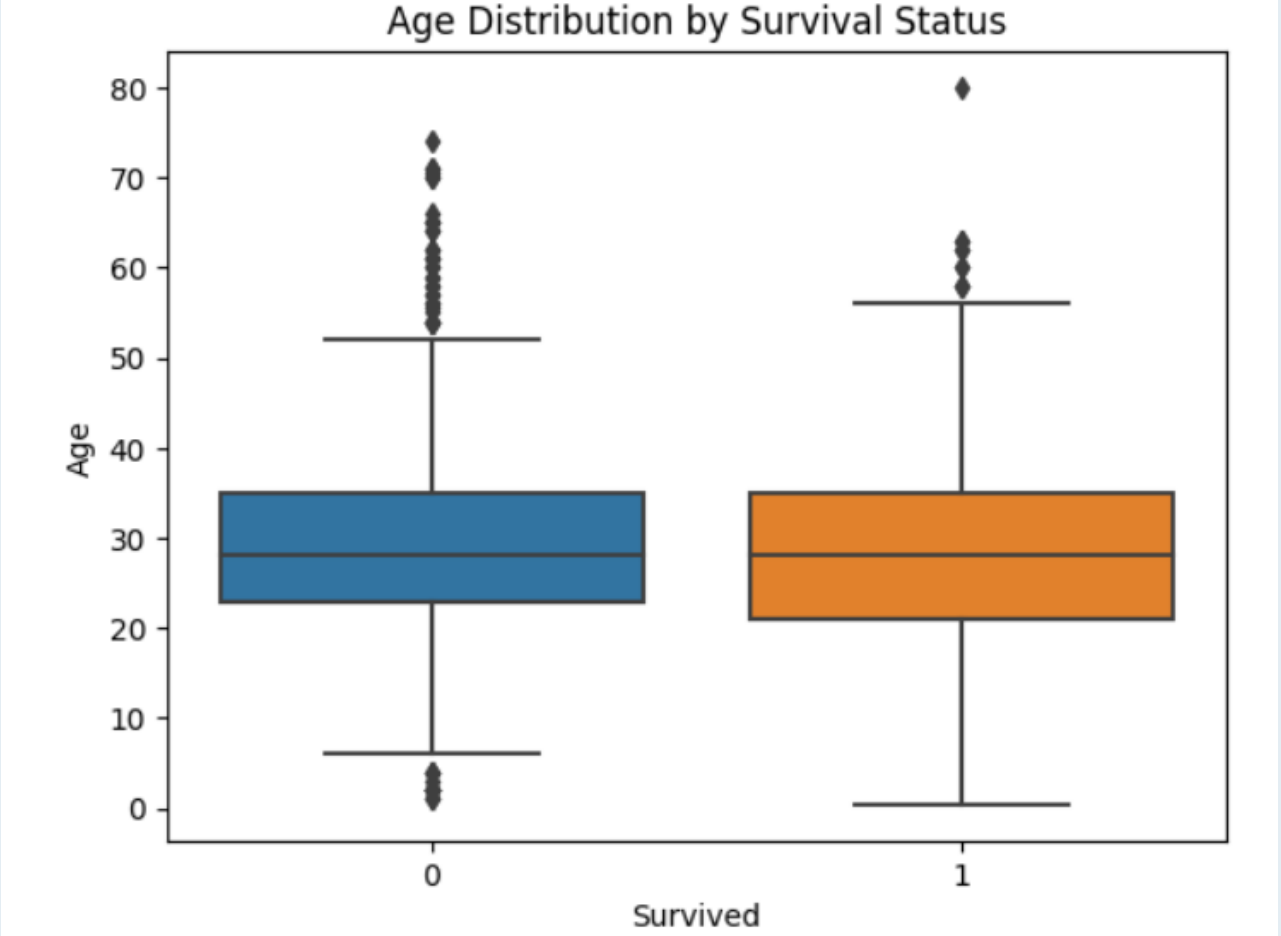
EXPLATORY DATA ANALYSIS

1 Analysis Column Age



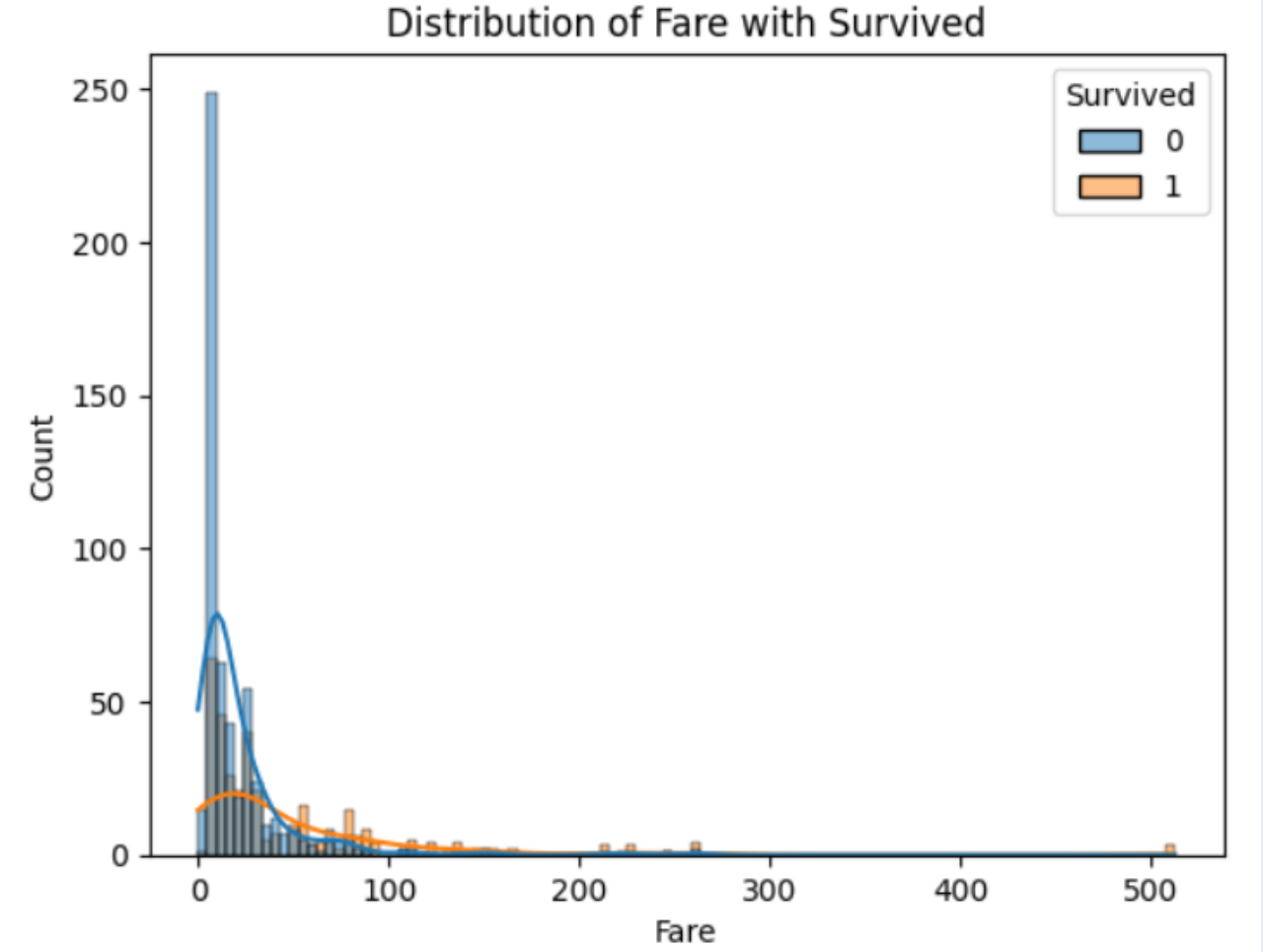
Children under the age of 10 have a greater chance of survival, possibly because they are prioritized for rescue using lifeboats.

2 Analysis Column Age



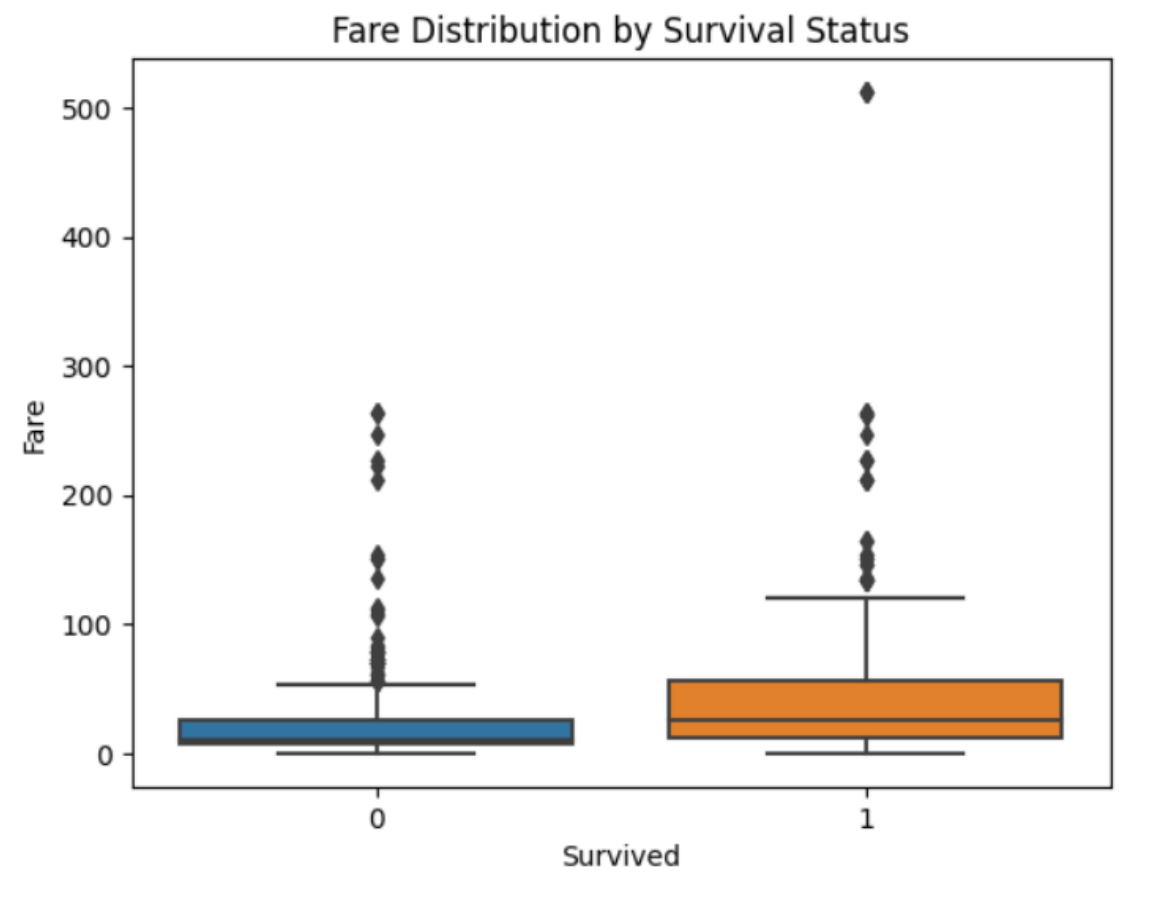
The area of the box indicating the 'Survived' category (Survived = 1) tends to be smaller than the area of the box indicating the 'Unsaved' category. This indicates that there are fewer individuals in the survivor category compared to the number of non-survivors in the analyzed data.

3 Analysis Column Fare



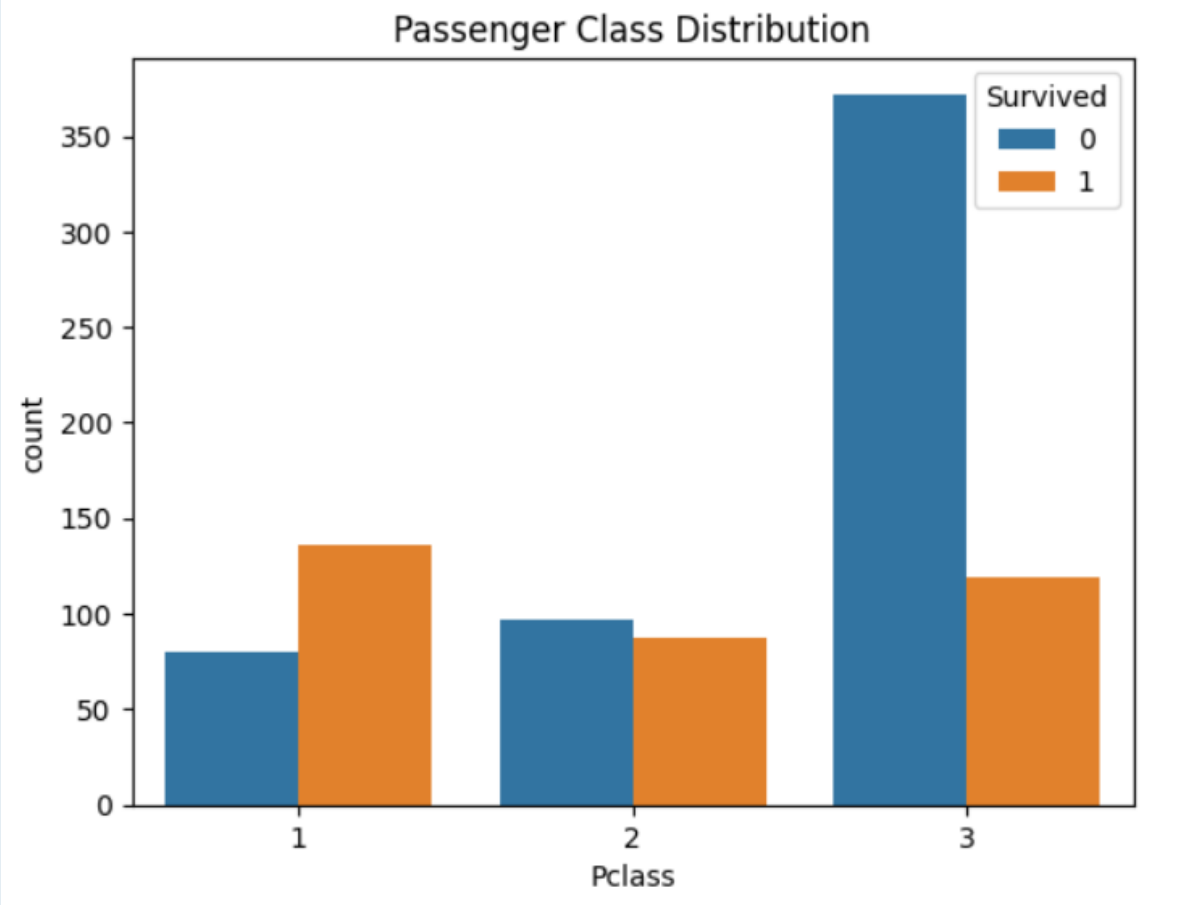
Passengers who pay higher travel costs are more likely to be safe than passengers who pay lower costs. This suggests a relationship between the amount of fare paid and the likelihood of safety, where passengers with higher fares may get better priority or facilities in emergency situations.

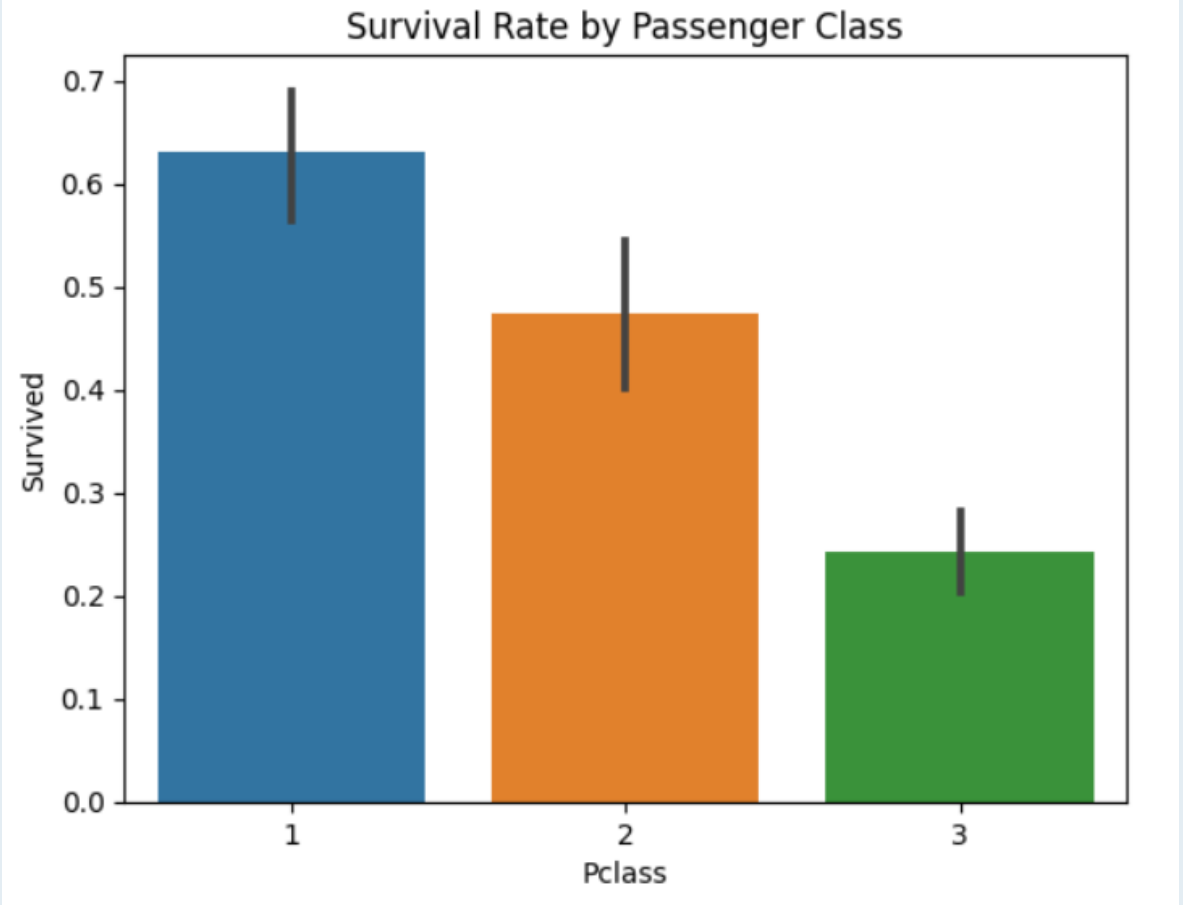
4 Analysis Column Fare



In the boxplot, the area indicating the 'Survived' category (Survived = 1) in the Fare column appears higher than the area for the 'Unsaved' category. This indicates that surviving passengers generally pay higher fares compared to non-surviving passengers.

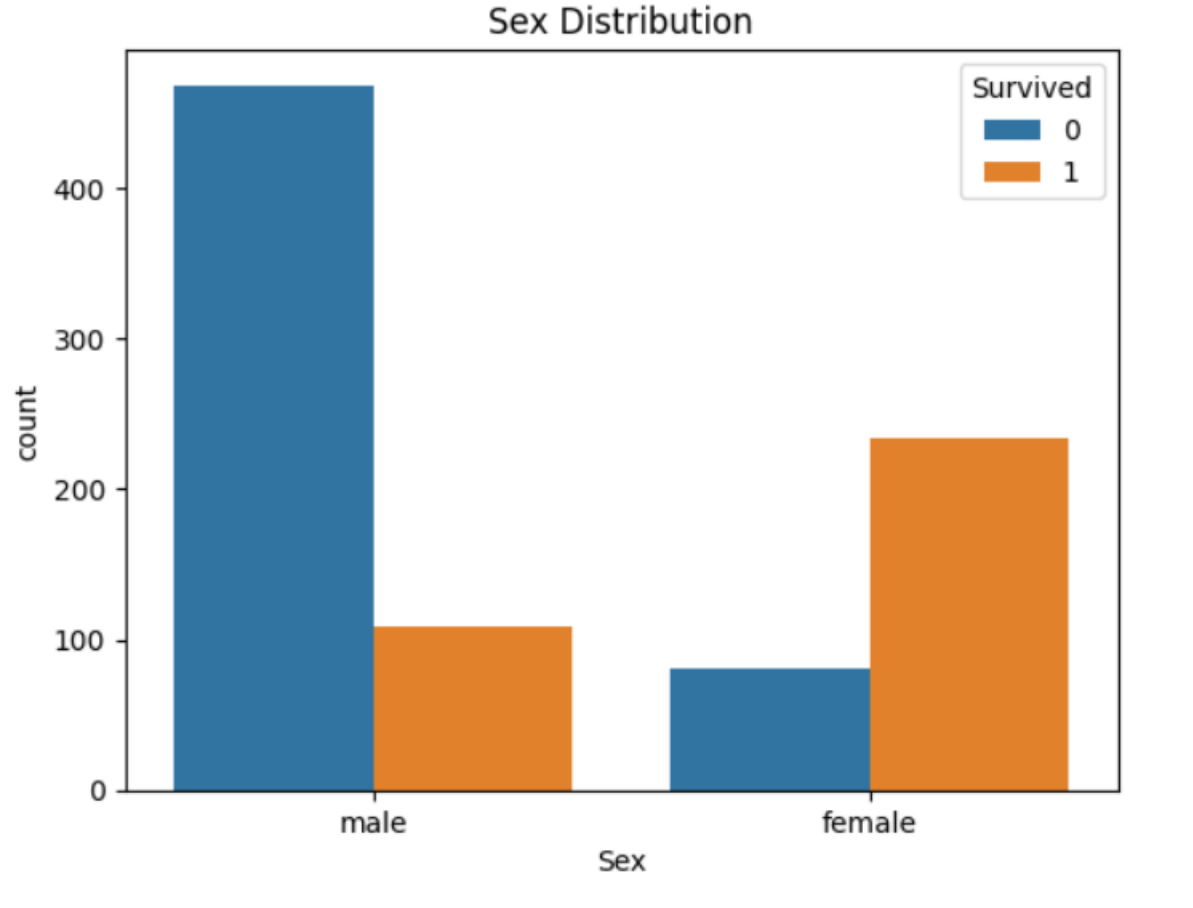
5 Analysis Column Pclass

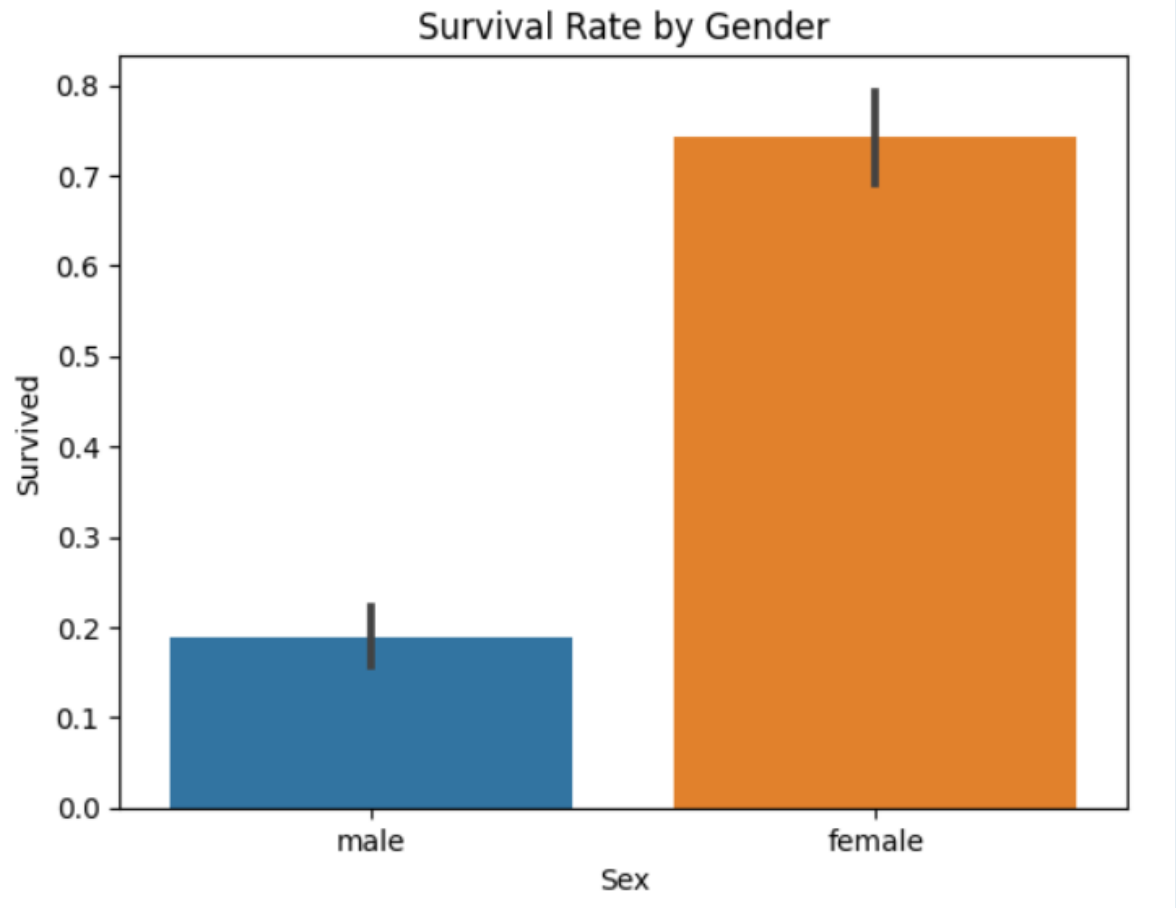




The percentage of survivors in Pclass = 1 is greater than Pclass 2 and 3.

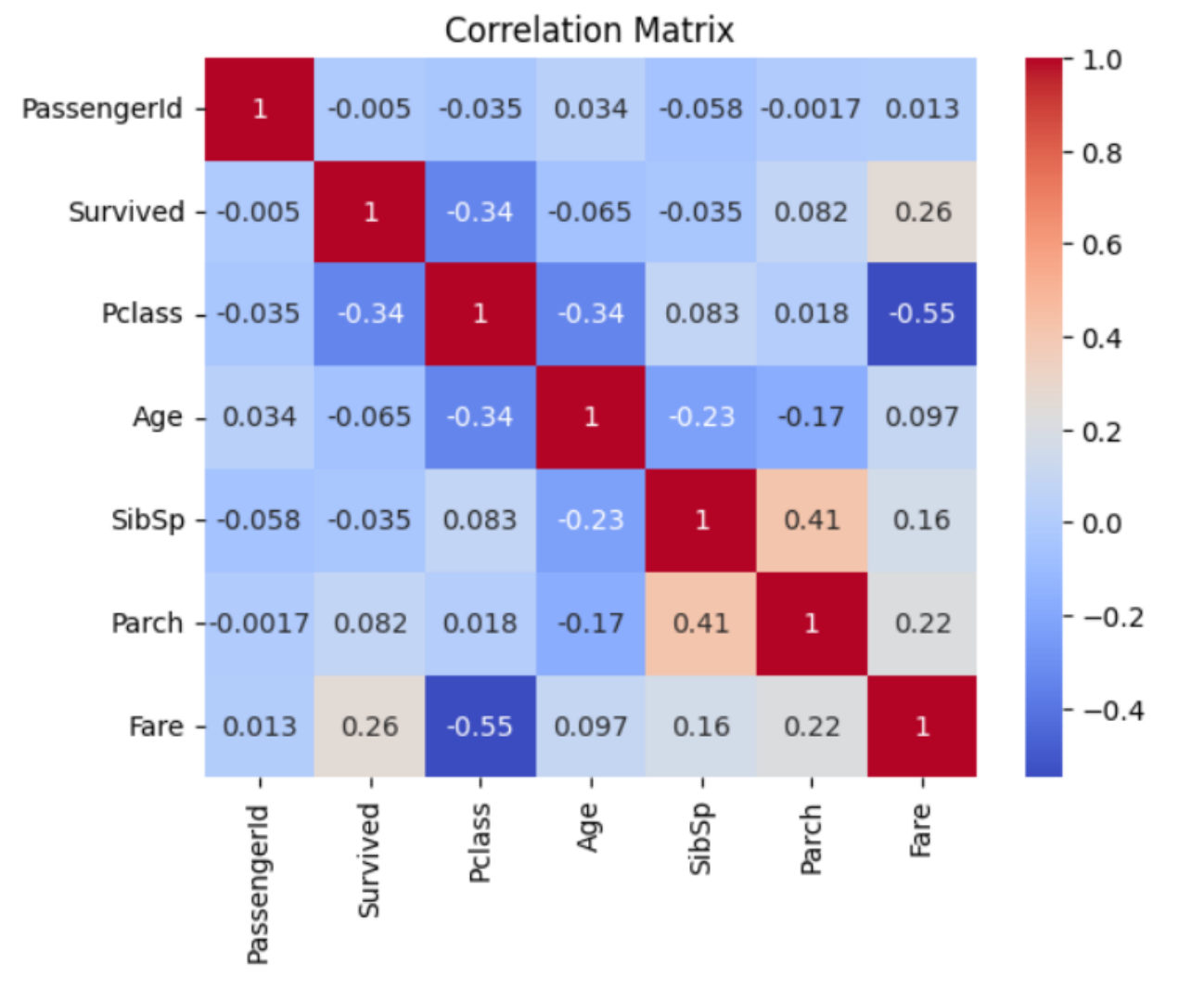
6 Analysis Column Sex





Percentage of survivors in Sex = female is higher than male

7 Correlation Matrix



Pclass dan Fare memiliki korelasi negativ, semakin rendah Pclass maka semakin tinggi Fare-nya.

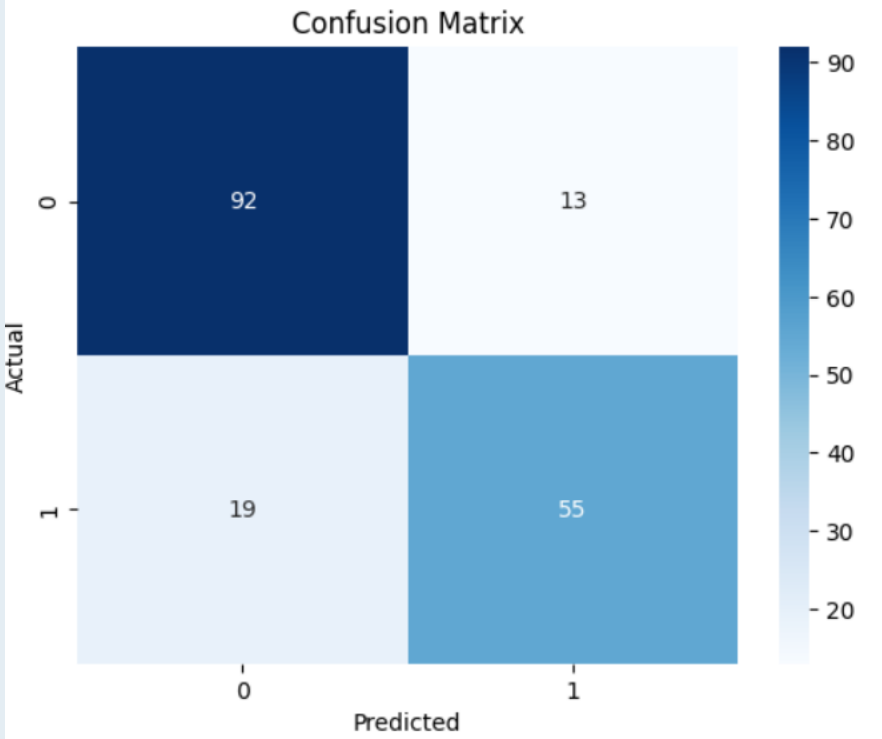
Fitur yang paling berpengaruh terhadap Survived adalah Pclass.

02

kaggle

MODELLING

Random Forest

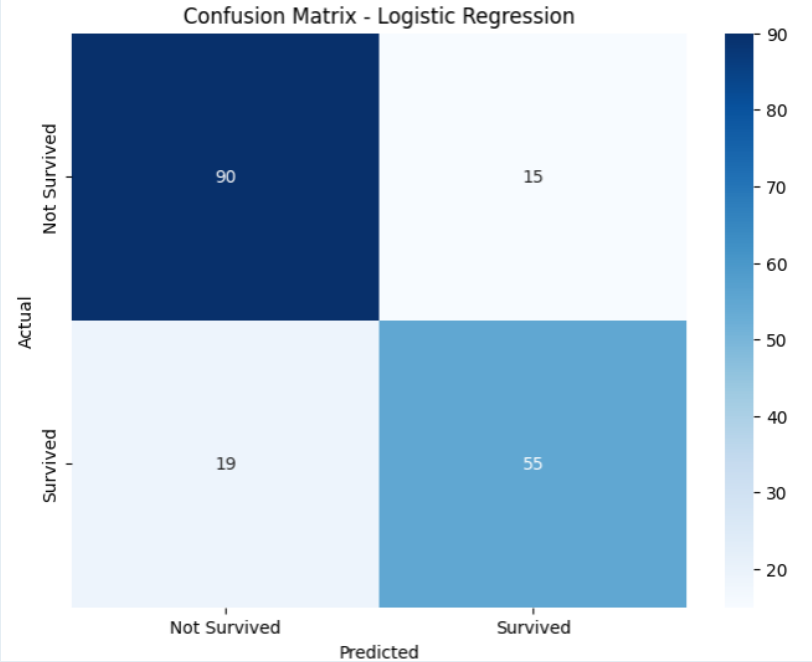


Classification Report:

	precision	recall	f1-score	support
0	0.83	0.88	0.85	105
1	0.81	0.74	0.77	74
accuracy			0.82	179
macro avg	0.82	0.81	0.81	179
weighted avg	0.82	0.82	0.82	179

- **Akurasi** yang tinggi (82%) menunjukkan bahwa model bekerja cukup baik dalam memprediksi kelas secara keseluruhan.
- **Precision** yang tinggi (80%) menunjukkan bahwa ketika model memprediksi "Selamat", ia sering benar.
- **Recall** yang lebih rendah (74%) menunjukkan bahwa model tidak berhasil mendeteksi semua kasus "Selamat", ada sejumlah kasus yang terlewatkan.
- **F1-Score** (77%) memberikan ukuran harmonis dari precision dan recall, memberikan gambaran seimbang tentang kinerja model

Logistic Regression



Logistic Regression

	precision	recall	f1-score	support
0	0.83	0.86	0.84	105
1	0.79	0.74	0.76	74
accuracy			0.81	179
macro avg	0.81	0.80	0.80	179
weighted avg	0.81	0.81	0.81	179

Accuracy: 0.8100558659217877

- **Akurasi**: 81% - Model cukup baik secara keseluruhan.
- **Precision**: 79% - Model relatif akurat ketika memprediksi kelas positif.
- **Recall**: 74% - Model mampu mendeteksi sebagian besar kasus positif tetapi ada beberapa yang terlewat.
- **F1-Score**: 76% - Keseimbangan antara precision dan recall yang baik.

CONCLUSION

- Random Forest showed a slight edge in all the evaluation metrics measured. So, if the main priorities are higher accuracy and precision and overall balanced model performance, then Random Forest is the better choice.
- Logistic Regression remains a reliable model and is almost comparable to Random Forest. Although slightly behind in terms of performance on Titanic data, it is still a useful model especially if you need ease of interpretation and faster processing.