

TITANIC PROJECT



DIGITAL SKILL FAIR 310



Table of Contents

1. About Dataset

2. Data Dictionary

3. Data Preprocessing

4. EDA

5. Feature Engineering

**6. Feature Selection
& Train-Test Split**

7. Modeling

8. Kesimpulan

About Dataset

	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... Heikkinen, Miss. Laina	female	38.0	1	0	PC 17599 STON/O2. 3101282	71.2833	C85	C
2	3	1	3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	26.0	0	0	113803	53.1000	C123	S
3	4	1	1	Allen, Mr. William Henry	male	35.0	1	0	373450	8.0500	NaN	S
4	5	0	3	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
...
886	887	0	2	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
887	888	1	1	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W/C. 6607	23.4500	NaN	S
888	889	0	3	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
889	890	1	1	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q
890	891	0	3									

Dataset ini berasal dari kompetisi Titanic di Kaggle, yang berisi informasi tentang penumpang Titanic dan apakah mereka selamat atau tidak. Dataset Titanic ini terdiri dari 891 baris, yang mewakili setiap penumpang yang ada di kapal Titanic, dan 12 kolom. Tujuan dari analisis ini adalah untuk memprediksi kelangsungan hidup penumpang berdasarkan karakteristik tertentu.

Data Dictionary

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	# of siblings / spouses aboard the Titanic	
parch	# 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

Data Preprocessing

```
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
```

Sebelum

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

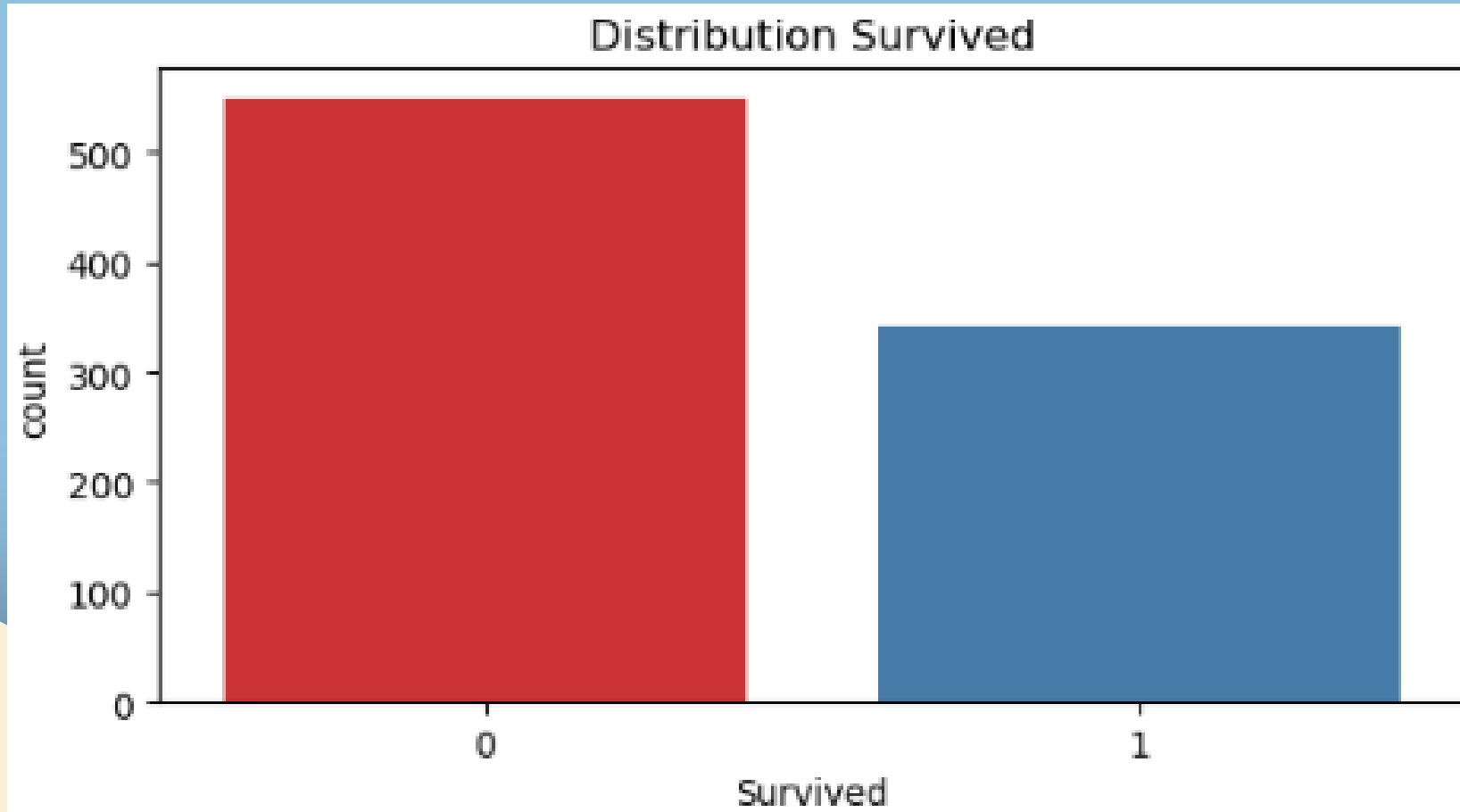
Sesudah

Pada tahap data preprocessing, kita membersihkan data dari nilai-nilai yang hilang atau kosong (missing values).

Pada gambar di sebelah kiri (**Sebelum**), terlihat bahwa kolom Age memiliki 177 nilai yang hilang, kolom Cabin memiliki 687 nilai yang hilang, dan kolom Embarked memiliki 2 nilai yang hilang. Setelah proses pembersihan, seluruh nilai hilang telah diisi atau dihapus, sehingga data siap untuk dianalisis lebih lanjut.

EDA

Distribution Survived

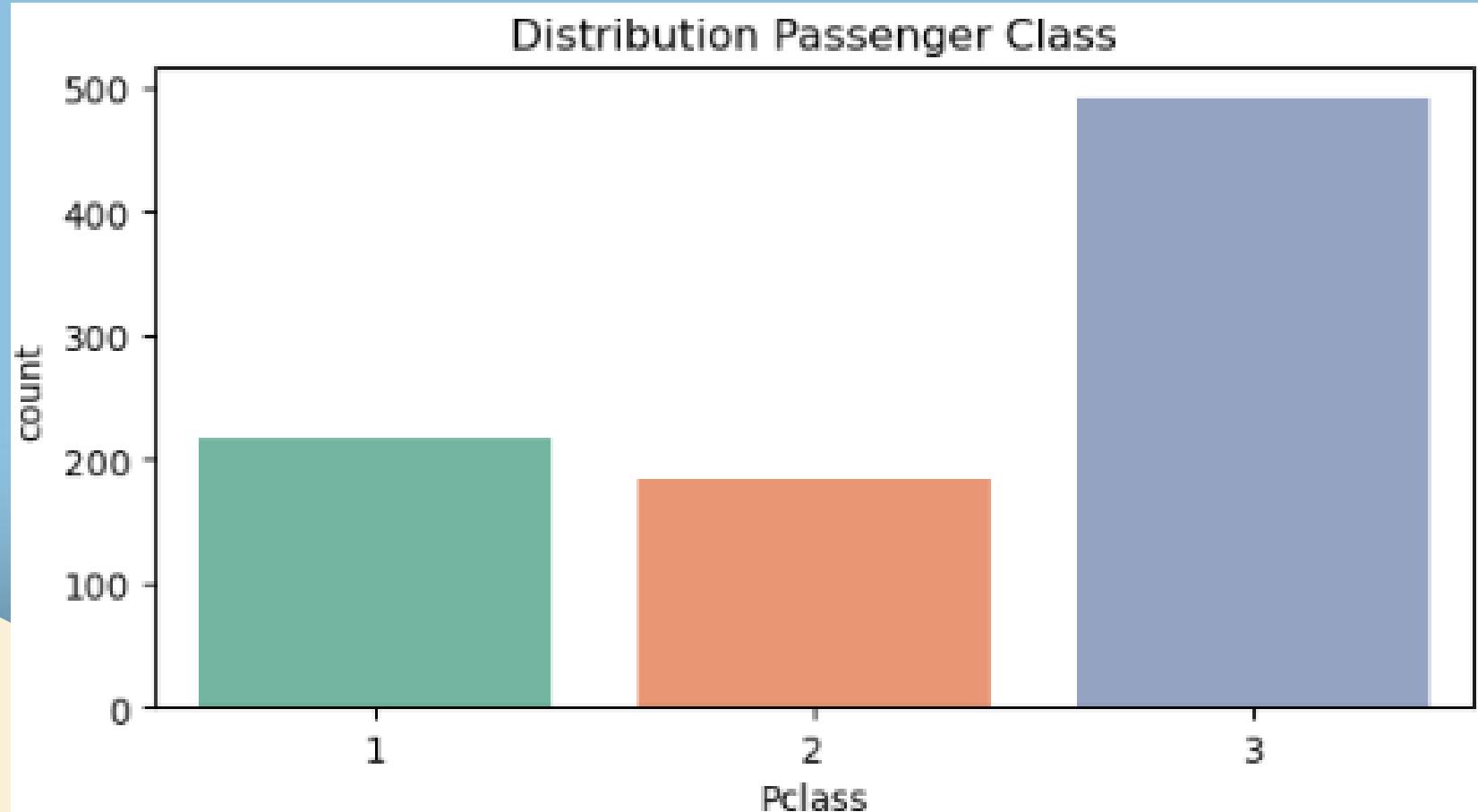


Distribution Survived

Grafik ini menunjukkan jumlah penumpang yang tidak selamat (0) jauh lebih tinggi (~500) dibandingkan yang selamat (1) (~300).

EDA

Distribution Passenger Class

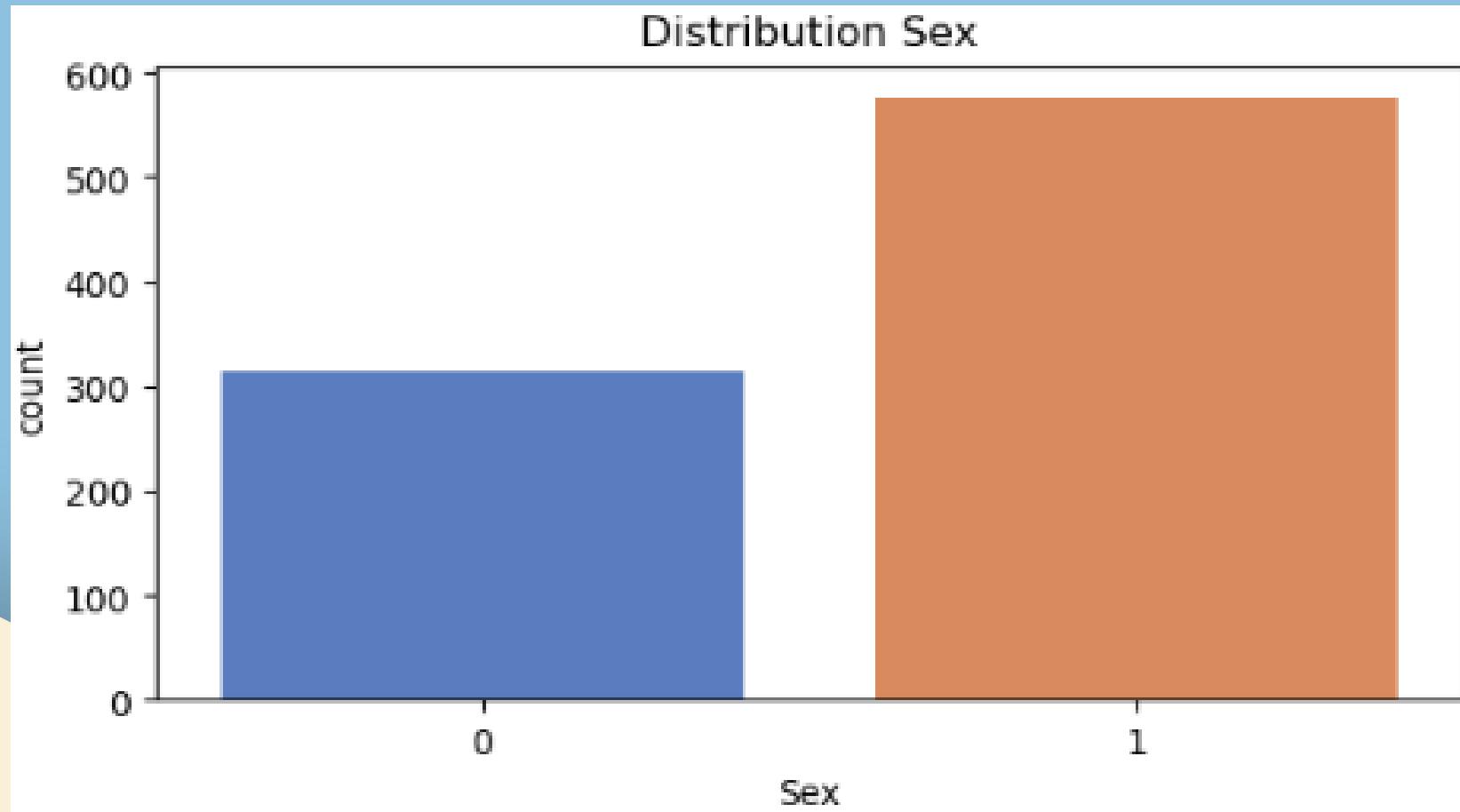


Distribution Passenger Class

Grafik ini menggambarkan mayoritas penumpang berada di kelas ketiga (3), dengan jumlah yang jauh lebih besar dibanding kelas pertama (1) dan kedua (2). Ini menunjukkan adanya distribusi yang tidak merata antara kelas penumpang.

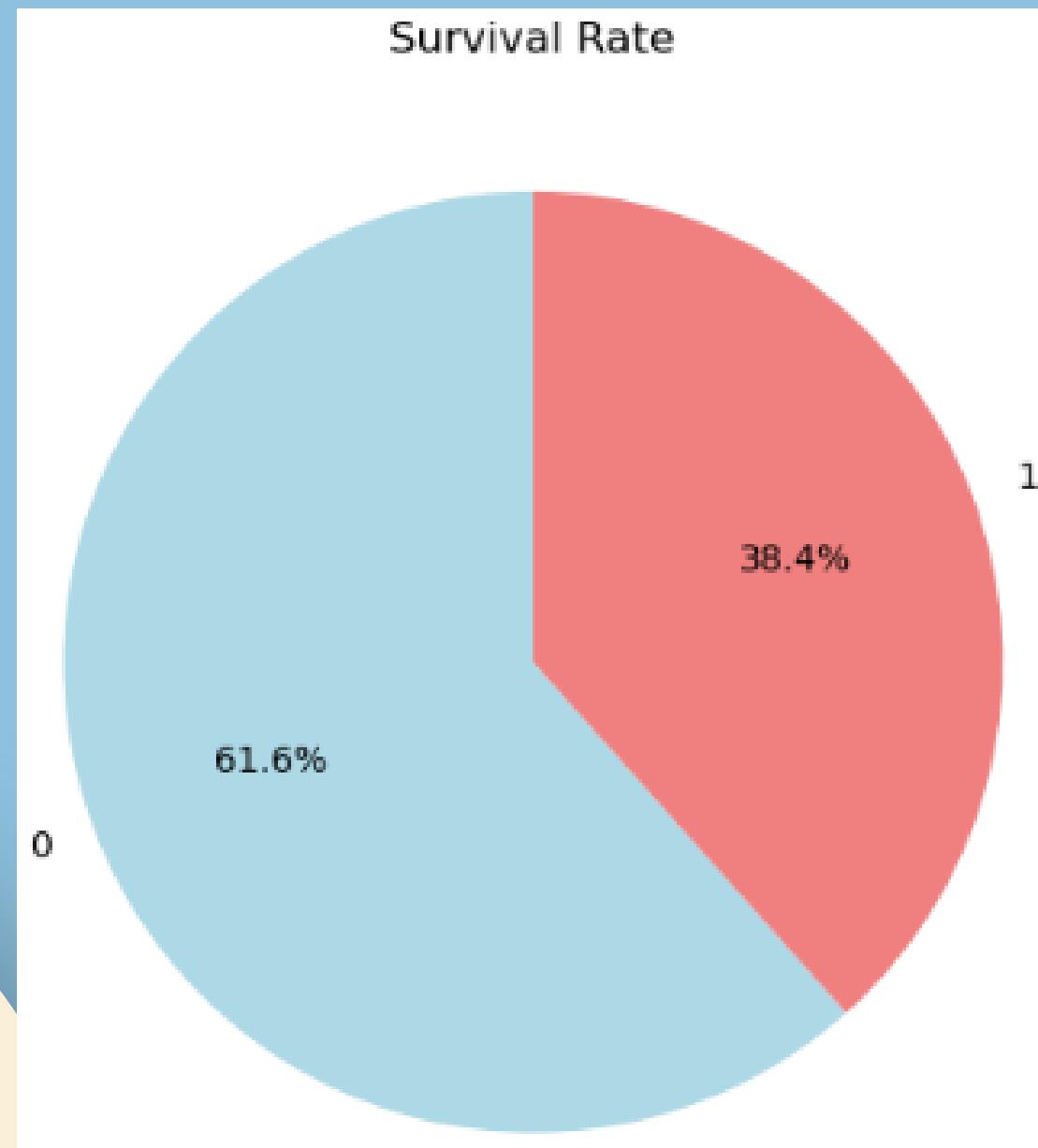
EDA

Distribution Gender



Grafik ini menunjukkan mayoritas penumpang berjenis kelamin laki-laki (1), dengan jumlah hampir dua kali lipat dibanding perempuan (0). Ini penting karena jenis kelamin berkorelasi signifikan dengan peluang bertahan hidup.

EDA

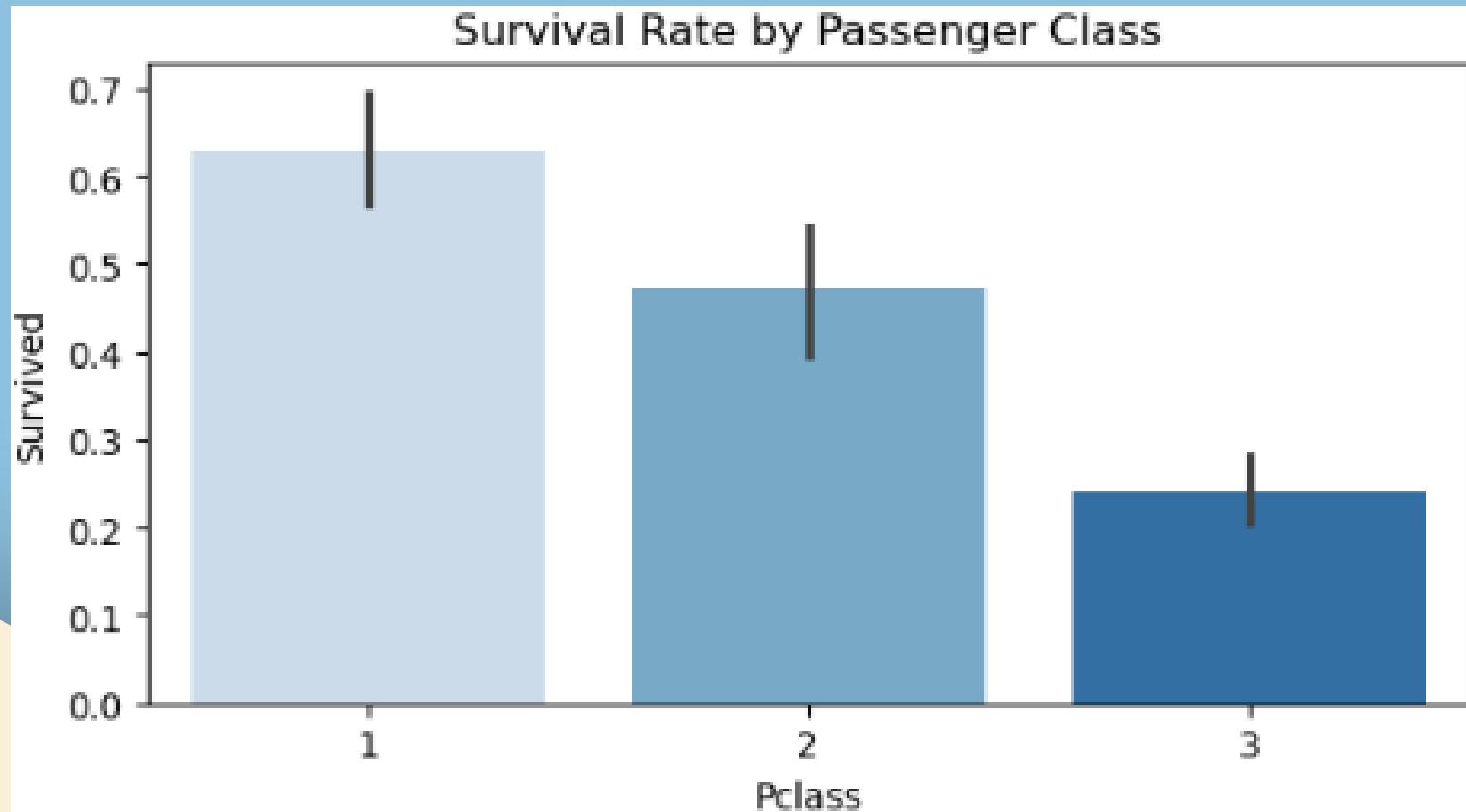


Survival Rate

- Dari Grafik tersebut:
 - 61.6% penumpang tidak selamat (label 0, warna biru).
 - 38.4% penumpang selamat (label 1, warna merah).
- Ini mengindikasikan bahwa sebagian besar penumpang tidak selamat.

EDA

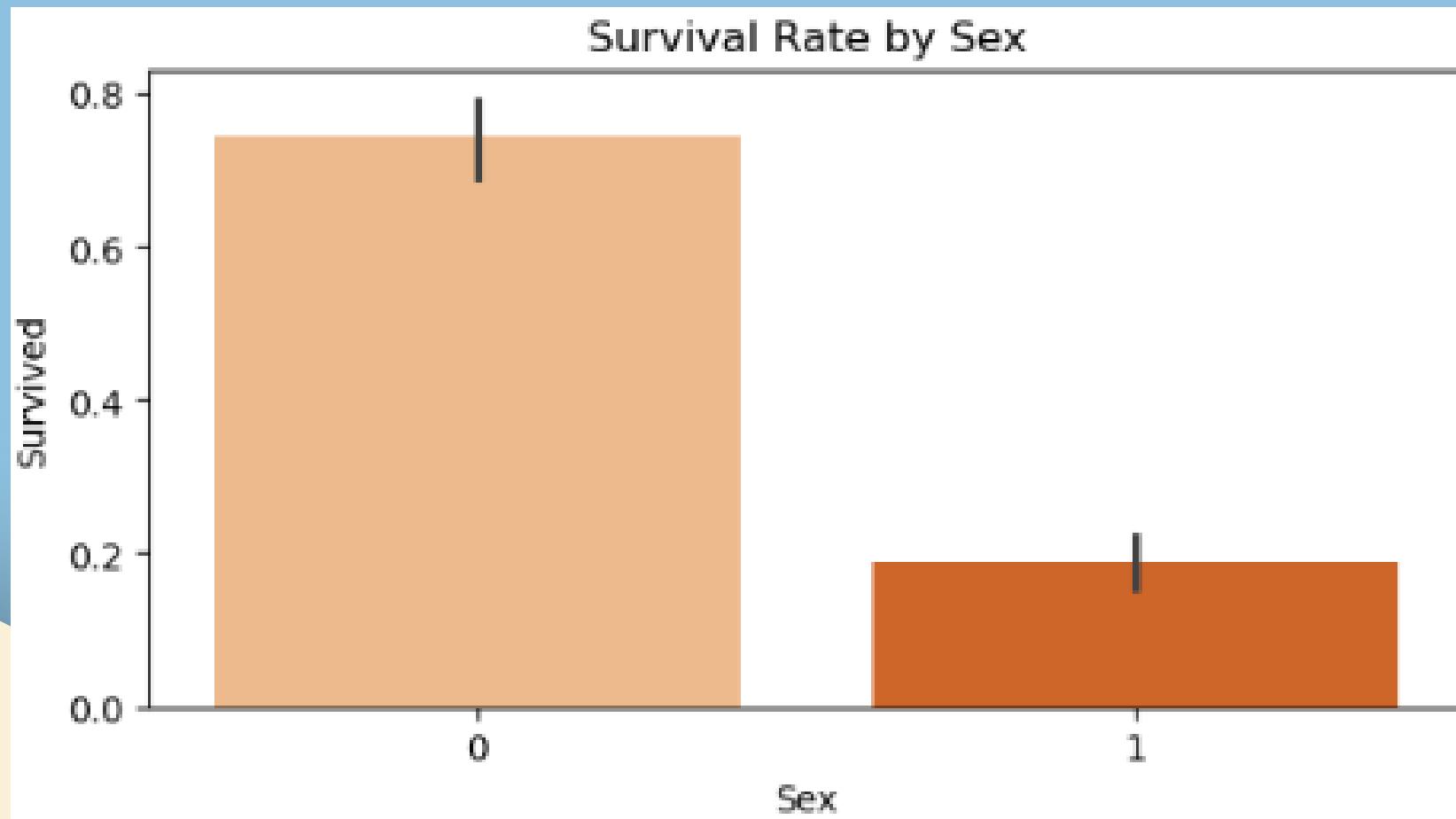
Survival Rate by Passenger Class



Survival Rate by Passenger Class

Grafik ini menunjukkan bahwa penumpang kelas pertama memiliki peluang terbesar untuk selamat (~0.6), diikuti oleh kelas kedua (~0.4), dan yang paling rendah adalah kelas ketiga (~0.2). Ini menunjukkan adanya hubungan yang kuat antara kelas penumpang dan tingkat keselamatan.

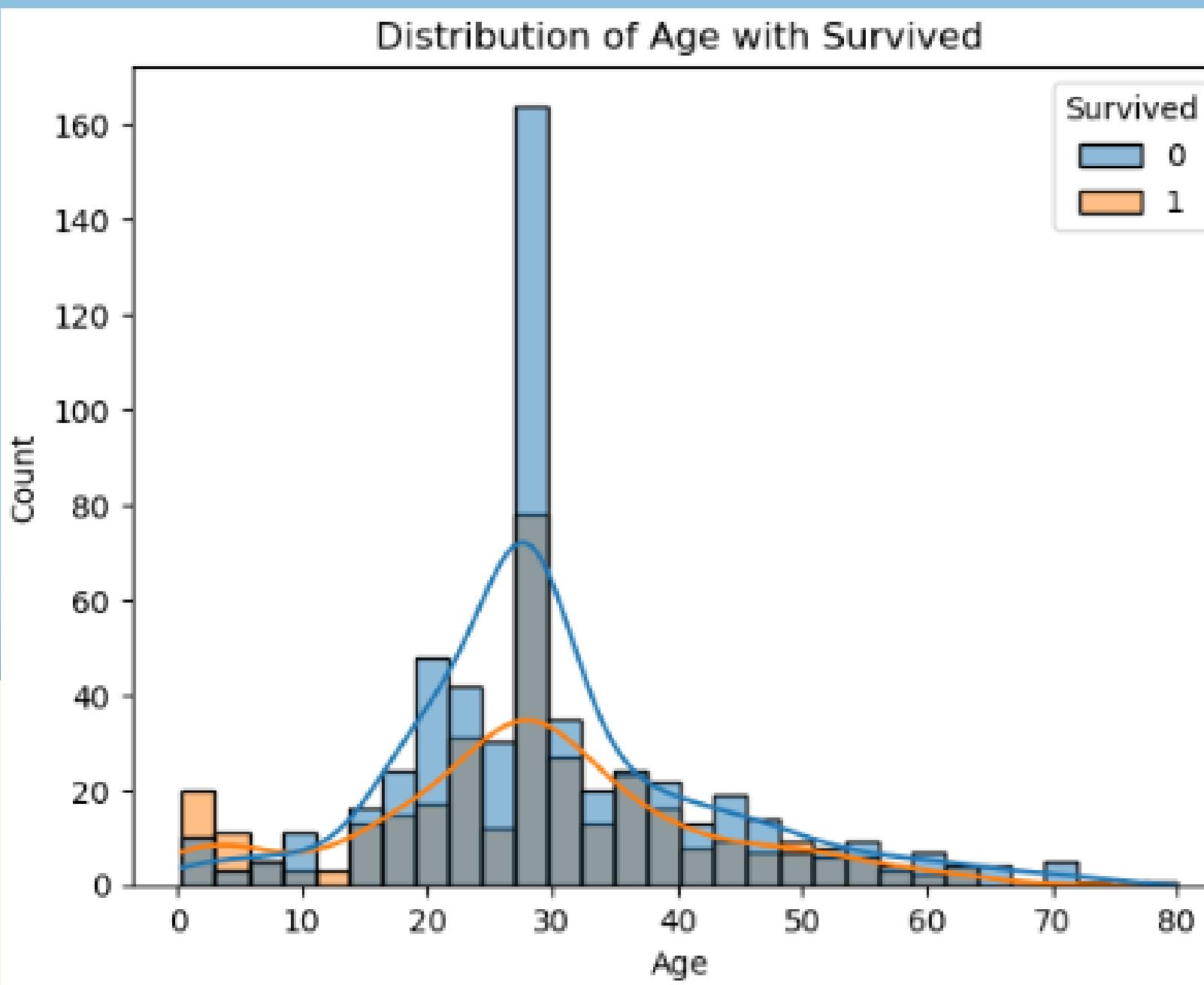
EDA



Survival Rate by Gender

Grafik ini menunjukkan bahwa perempuan (0) memiliki tingkat keselamatan yang jauh lebih tinggi (~0.75) dibandingkan laki-laki (1), yang tingkat keselamatannya jauh lebih rendah (~0.2). Ini menguatkan kesimpulan dari korelasi sebelumnya bahwa jenis kelamin merupakan faktor penting dalam bertahan hidup.

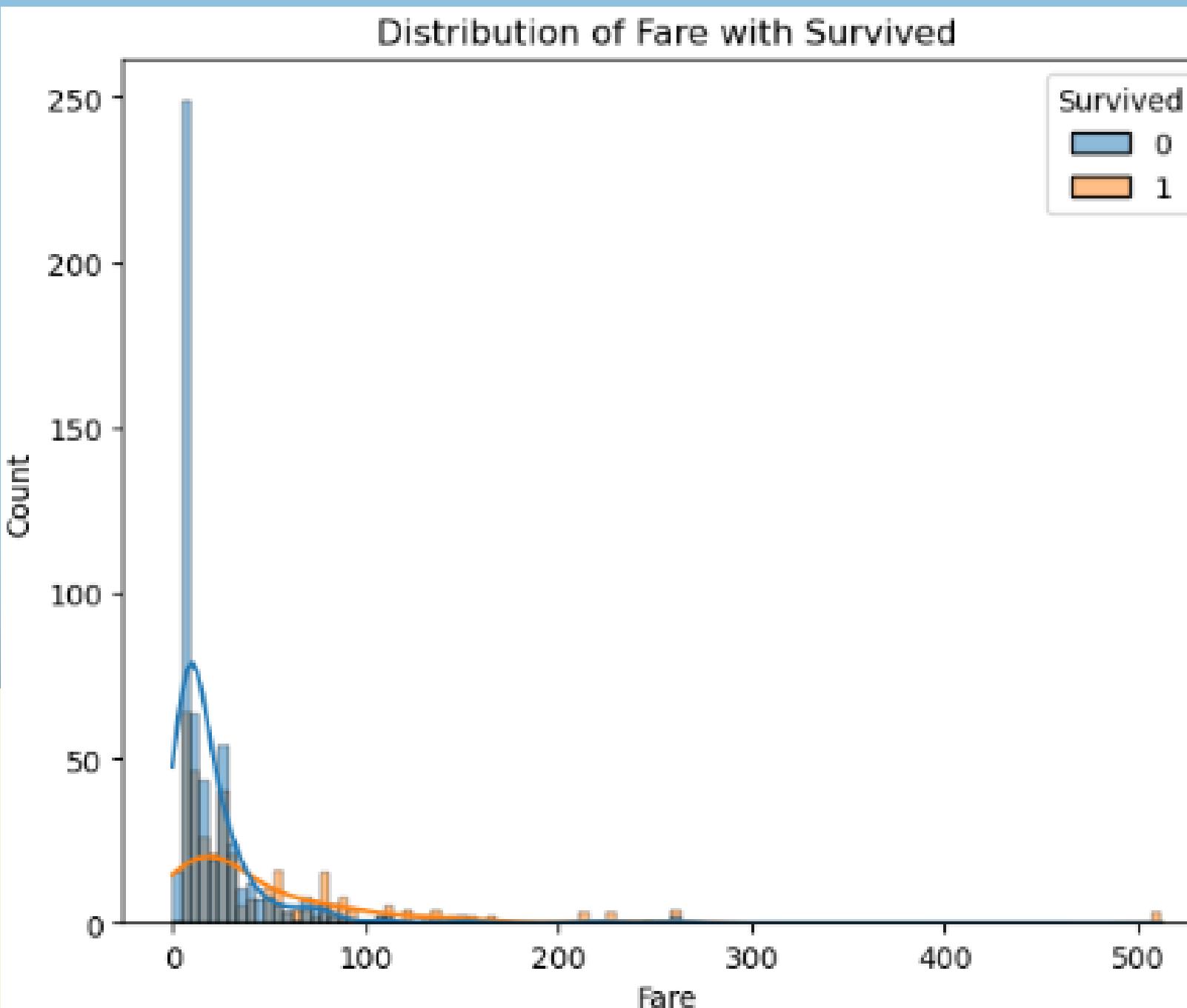
EDA



Distribution of Age with Survived

- Sebagian besar penumpang yang tidak selamat berada di rentang usia sekitar 20-40 tahun, dengan puncak di usia 30-an.
- Penumpang yang selamat cenderung tersebar di berbagai kelompok usia, namun jumlah mereka lebih sedikit dibandingkan yang tidak selamat, terutama pada rentang usia yang lebih muda (0-10 tahun) dan usia di atas 50 tahun.
- Kernel Density Estimate (KDE) juga menampilkan pola distribusi dengan garis yang lebih halus.

EDA

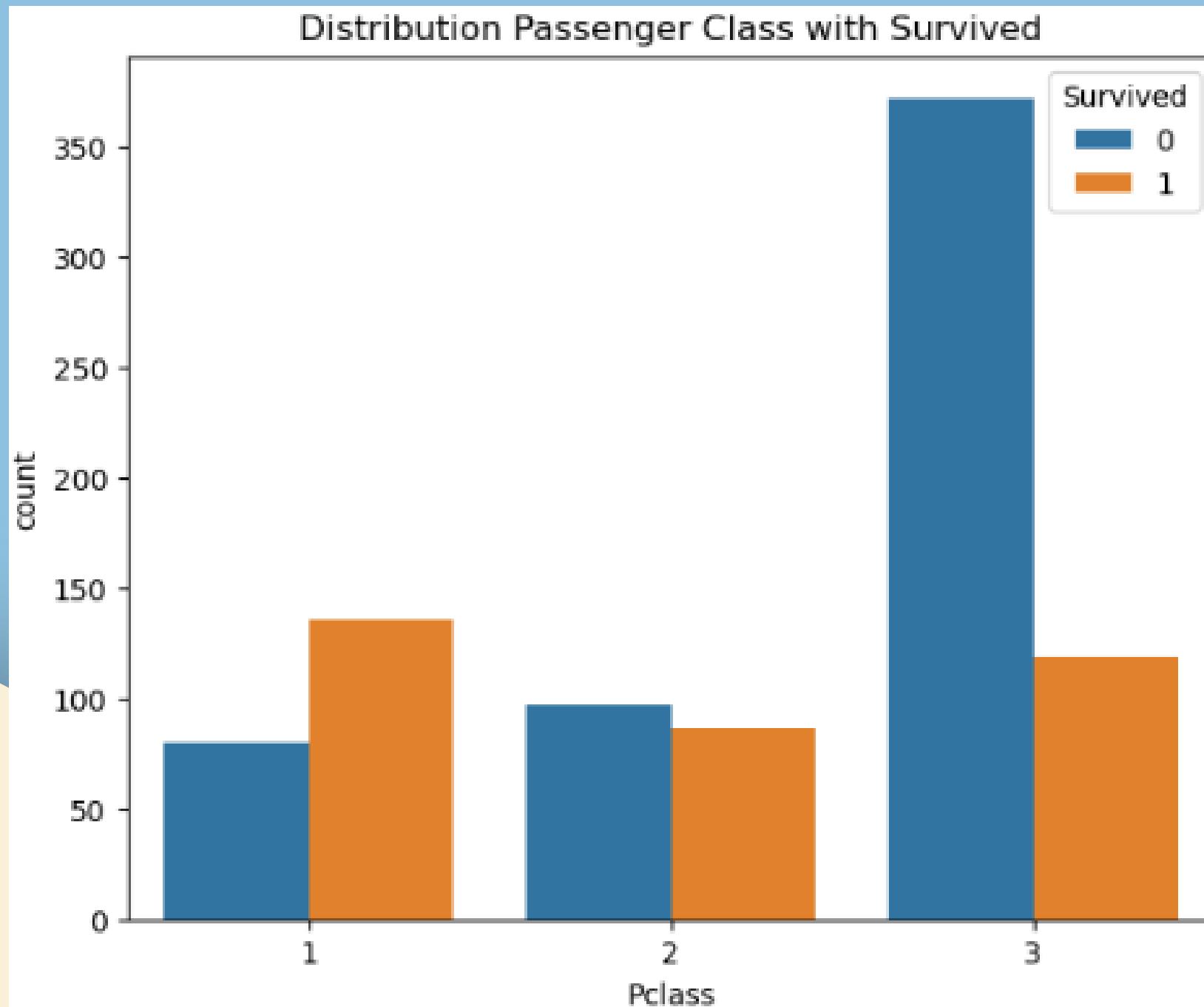


Distribution of Fare with Survived

- Sebagian besar penumpang membayar tarif yang lebih rendah (sekitar 0-100), dan banyak dari mereka tidak selamat. Penumpang dengan tarif lebih tinggi cenderung memiliki peluang lebih besar untuk selamat, meskipun jumlah mereka lebih sedikit.
- Grafik ini menunjukkan bahwa ada korelasi antara harga tiket dan kemungkinan selamat, dengan harga tiket yang lebih tinggi cenderung terkait dengan tingkat keselamatan yang lebih besar.

EDA

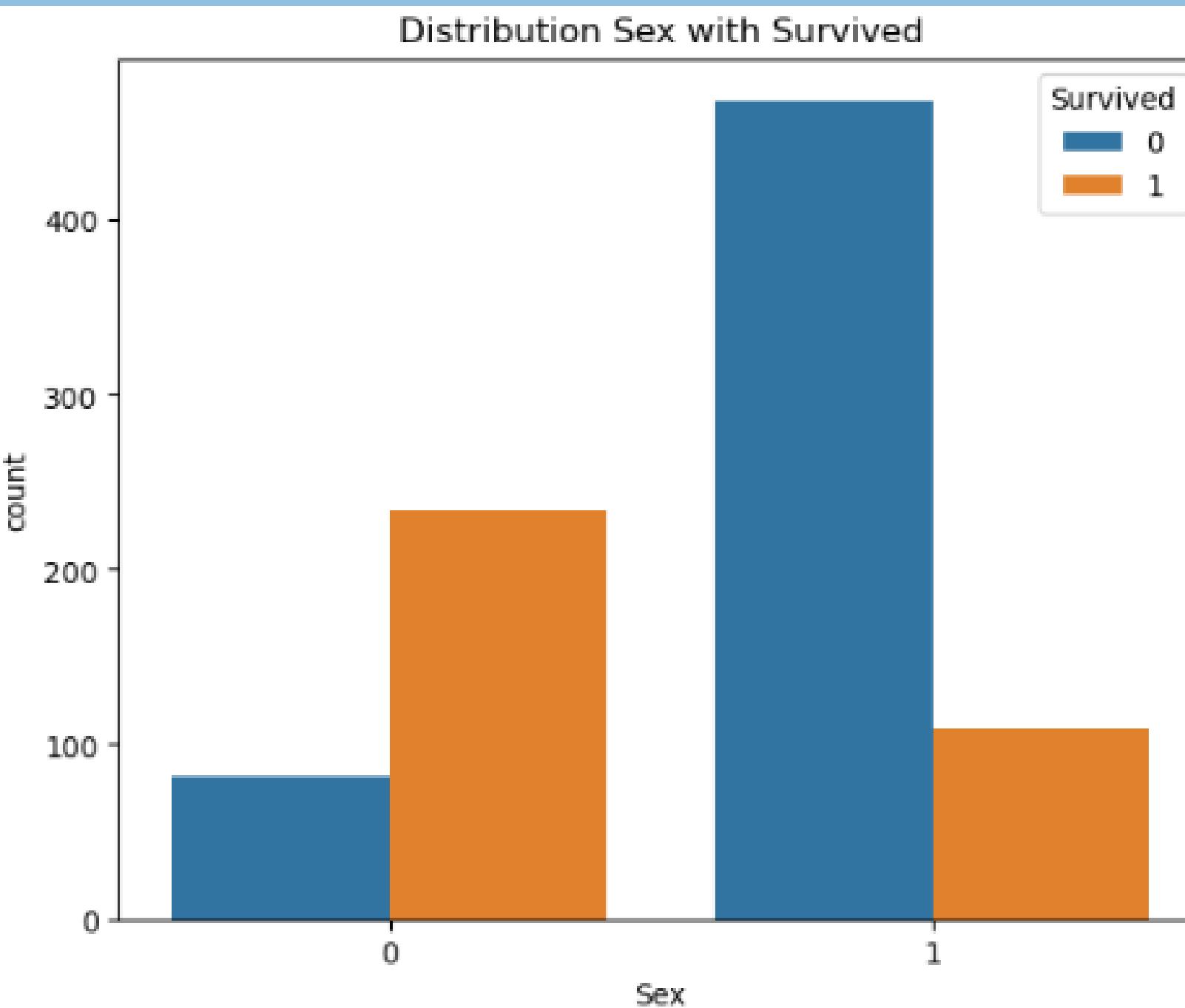
Distribution Passenger Class with Survived



Distribution of Passenger Class with Survived

- Penumpang kelas 3 mendominasi penumpang yang tidak selamat, sementara penumpang kelas 1 memiliki tingkat kelangsungan hidup yang lebih tinggi.
- Hal ini menunjukkan adanya hubungan antara kelas penumpang dan kemungkinan keselamatan, di mana penumpang kelas yang lebih rendah (kelas 3) memiliki peluang lebih kecil untuk selamat dibandingkan dengan penumpang kelas 1.

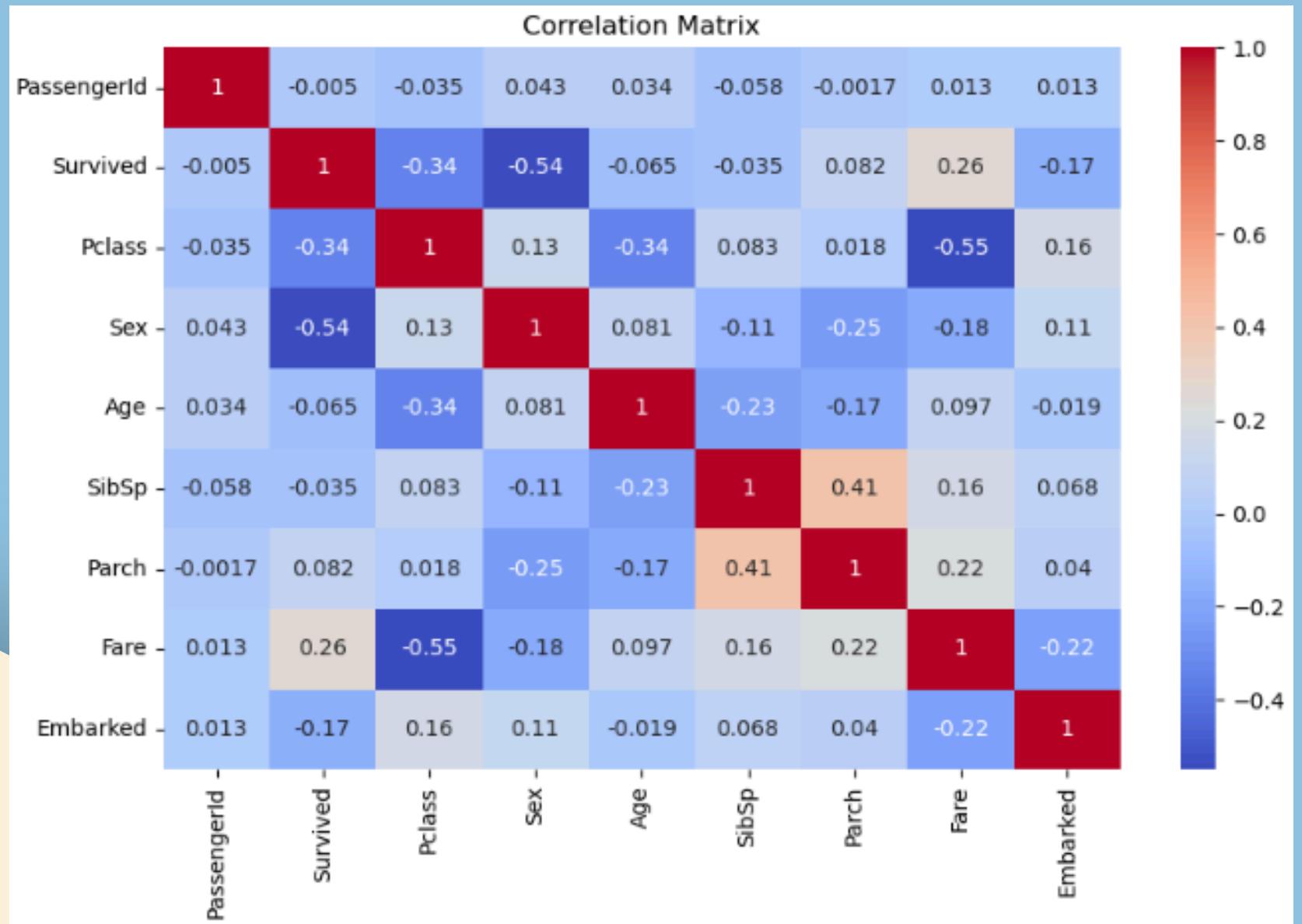
EDA



Distribution of Gender with Survived

- Penumpang laki-laki ($\text{Sex} = 1$) jauh lebih banyak yang tidak selamat ($\text{Survived} = 0$, warna biru) dibandingkan penumpang perempuan ($\text{Sex} = 0$).
- Sebaliknya, perempuan memiliki peluang lebih besar untuk selamat, menunjukkan bahwa jenis kelamin merupakan faktor penting dalam menentukan peluang kelangsungan hidup.

EDA



Correlation Matrix

- Korelasi yang negatif kuat antara Sex dan Survived (-0.54), menunjukkan bahwa jenis kelamin berhubungan signifikan dengan kemungkinan bertahan hidup (kemungkinan besar, wanita lebih mungkin bertahan hidup dibanding pria).
- Korelasi positif sedang antara Fare (tarif) dan Survived (0.26), yang mungkin menunjukkan bahwa penumpang dengan tarif lebih tinggi memiliki peluang lebih besar untuk selamat.
- Korelasi negatif antara Pclass dan Survived (-0.34), mengindikasikan bahwa penumpang dari kelas yang lebih rendah (Pclass) memiliki peluang lebih kecil untuk selamat.

Feature Engineering

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

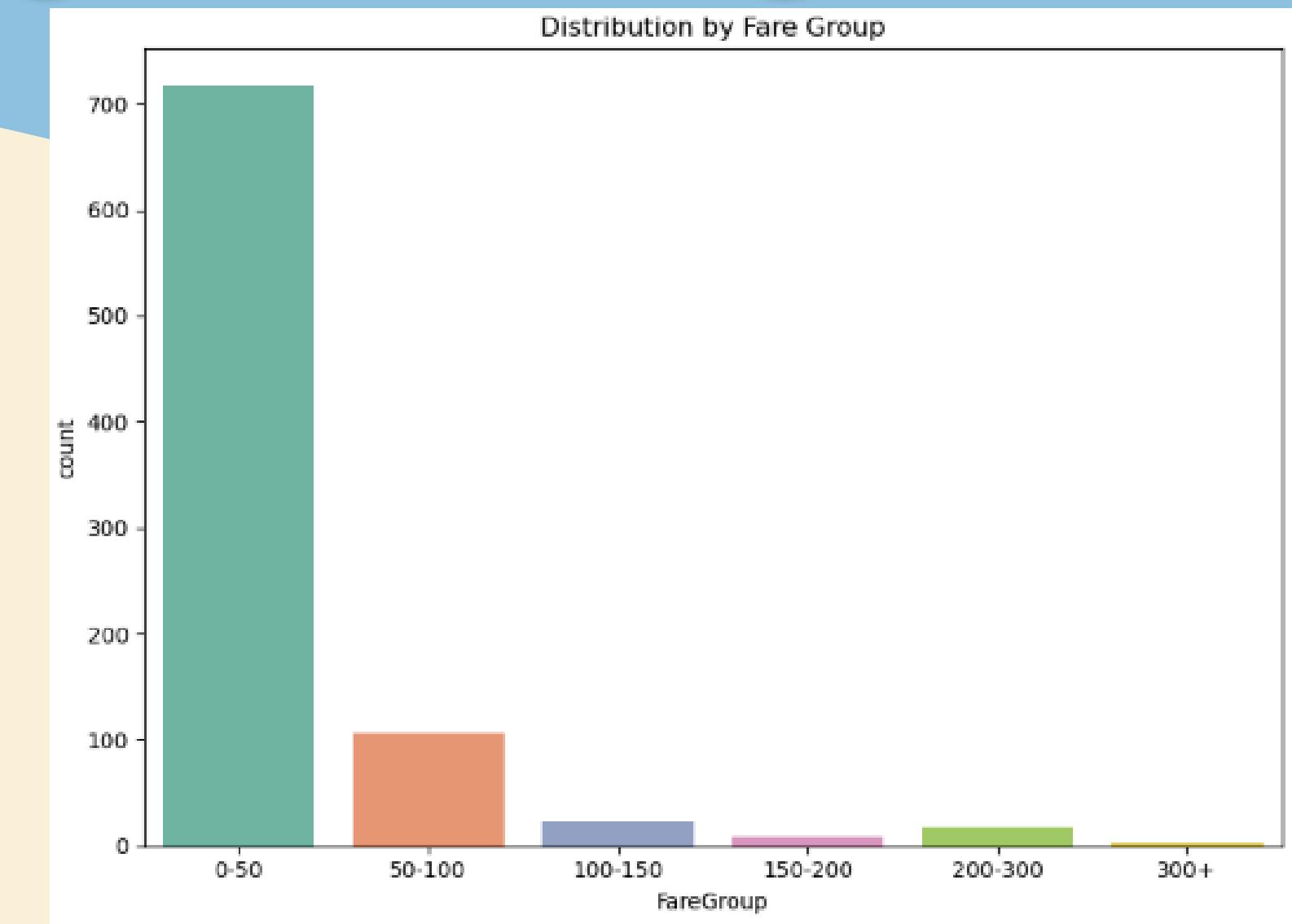
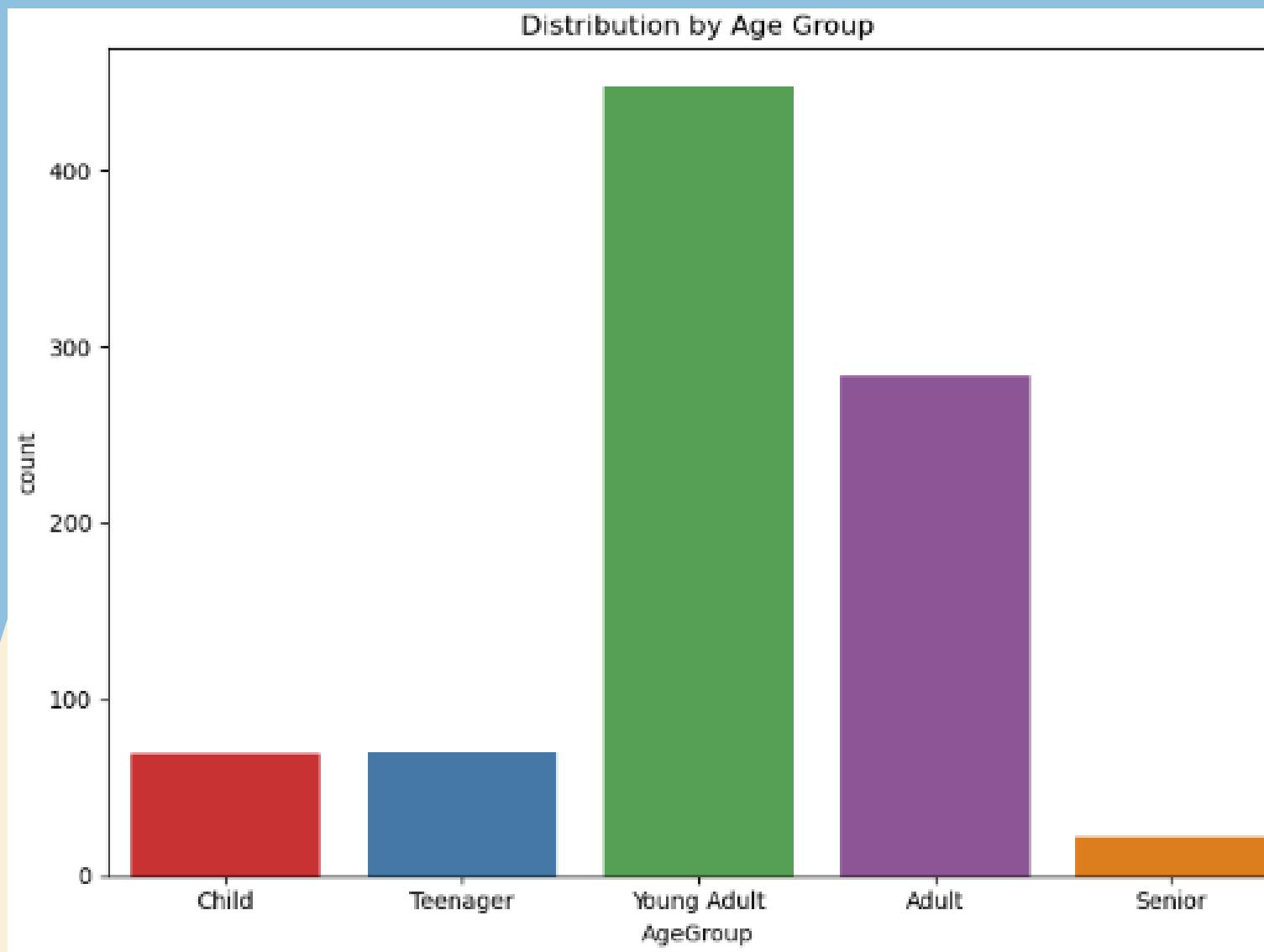
Sebelum

Sesudah

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

Pada tahap feature engineering, digunakan LabelEncoder untuk mengubah data kategorik menjadi numerik. Fitur-fitur yang diubah adalah 'Sex' dan 'Embarked'. Kolom 'Sex', yang awalnya berisi nilai laki-laki dan perempuan, diubah menjadi angka 1 dan 0. Sementara itu, kolom 'Embarked' yang berisi titik keberangkatan seperti C (Cherbourg), Q (Queenstown), dan S (Southampton) diubah menjadi angka 0, 1, dan 2. Transformasi ini diperlukan agar data lebih siap digunakan dalam model pembelajaran mesin, karena sebagian besar algoritma hanya dapat bekerja dengan data numerik.

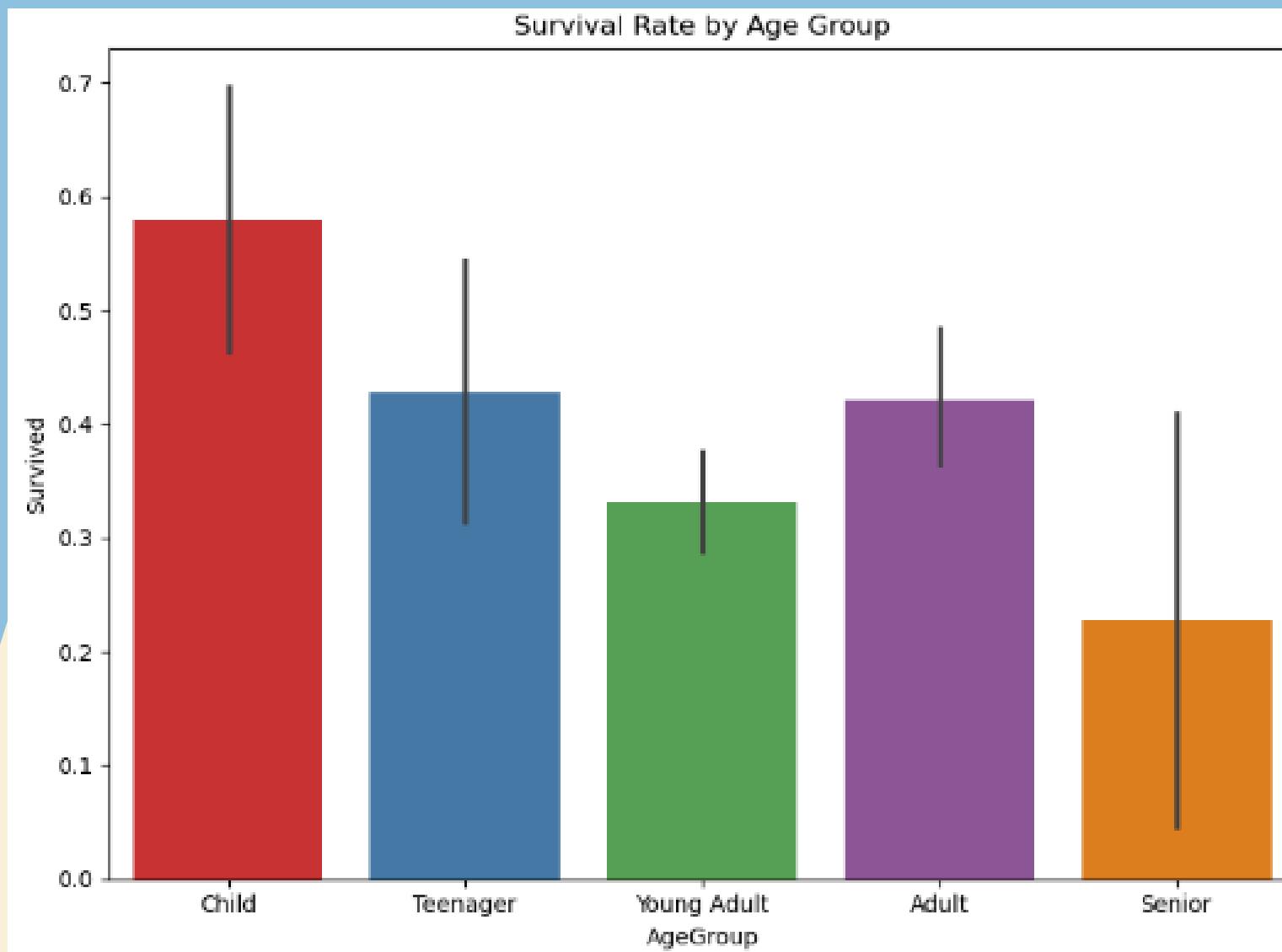
Feature Engineering



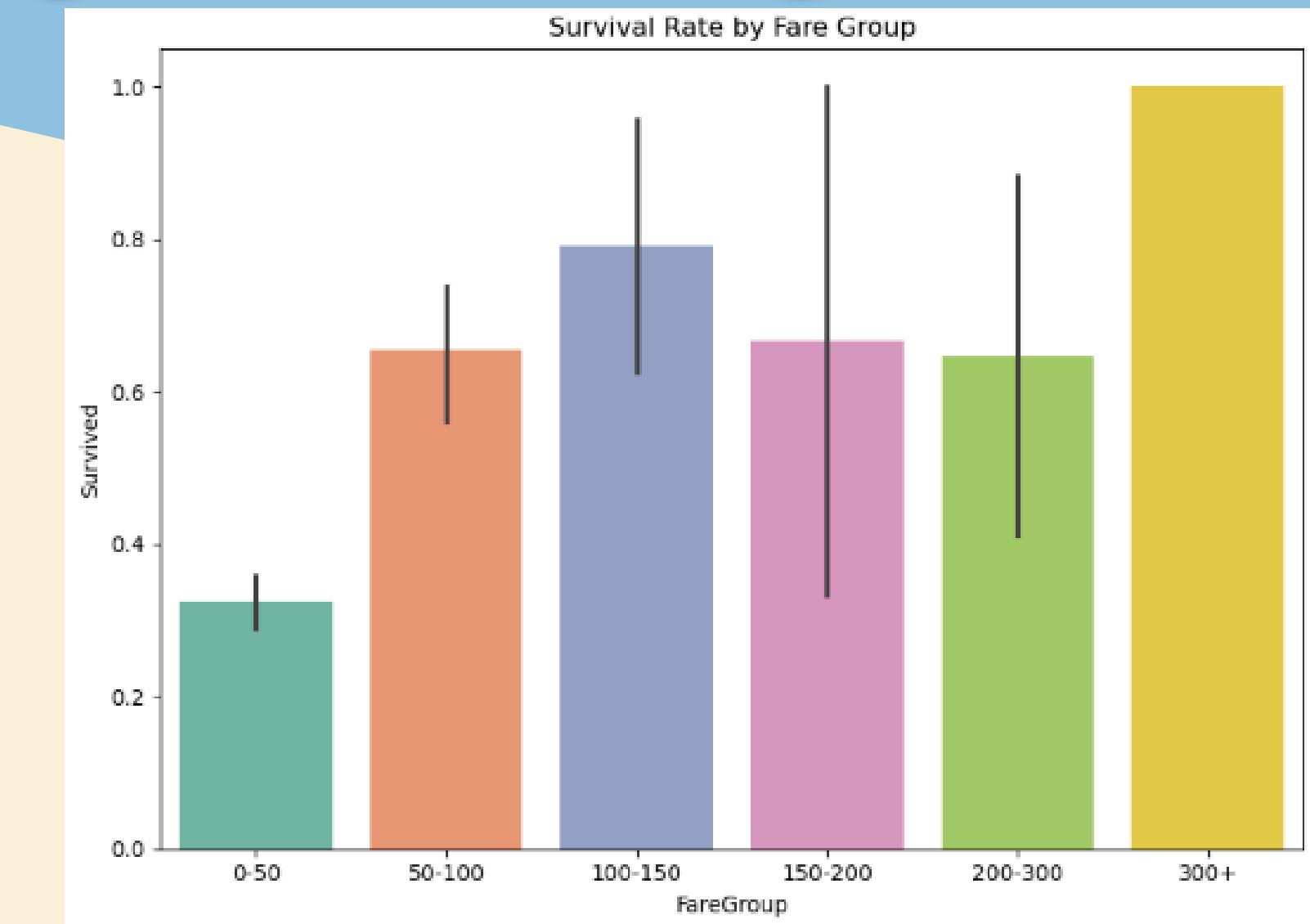
Selain untuk mengubah, di tahap ini kita juga bisa membuat fitur baru dari data mentah yang ada. Contohnya, usia penumpang dapat dibagi ke dalam beberapa kategori, seperti Child (Anak-anak), Teenager (Remaja), Young Adult (Dewasa Muda), Adult (Dewasa), dan Senior (Lansia). Dari hasilnya, terlihat bahwa mayoritas penumpang berada di kelompok Young Adult, sementara kelompok dengan jumlah penumpang paling sedikit adalah Senior.

Grafik ini menunjukkan distribusi penumpang berdasarkan kelompok tarif (Fare Group) yang dibayar. Penumpang yang membayar tarif paling rendah, dalam rentang 0-50, mendominasi jumlah penumpang. Sedangkan, hanya sedikit penumpang yang membayar lebih dari 300 untuk tarif mereka.

Feature Engineering



Grafik ini menunjukkan tingkat kelangsungan hidup (survival rate) berdasarkan kelompok usia. Child (Anak-anak) memiliki tingkat kelangsungan hidup tertinggi, sedangkan kelompok Senior memiliki tingkat kelangsungan hidup yang paling rendah. Ini mungkin menunjukkan adanya prioritas penyelamatan untuk anak-anak, sedangkan lansia memiliki tingkat kelangsungan hidup lebih rendah.



Grafik ini menunjukkan tingkat kelangsungan hidup berdasarkan kelompok tarif. Penumpang yang membayar tarif lebih tinggi, khususnya dalam rentang 300+, memiliki tingkat kelangsungan hidup tertinggi. Sebaliknya, mereka yang membayar tarif rendah (khususnya rentang 0-50) memiliki tingkat kelangsungan hidup yang lebih rendah. Ini menunjukkan adanya hubungan antara kemampuan ekonomi (fare) dengan peluang kelangsungan hidup pada tragedi Titanic.

Feature Selection & Train-Test Split

```
# Feature Selection  
features = ['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Embarked']  
X = df[features]  
y = df['Survived']
```

Pada tahap ini dipilih beberapa fitur yang relevan untuk model prediksi yaitu Pclass, Sex, Age, SibSp, Parch, Fare, dan Embarked. Variabel targetnya adalah Survived yang menyatakan apakah penumpang selamat atau tidak.

```
# Train Test Split, Memisahkan data untuk dilatih dan diuji  
from sklearn.model_selection import train_test_split  
  
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
```

X_train							
	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
301	3	1	28.0	2	0	23.2500	1
309	1	0	30.0	0	0	56.9292	0
516	2	0	34.0	0	0	10.5000	2
120	2	1	21.0	2	0	73.5000	2
570	2	1	62.0	0	0	10.5000	2
...
715	3	1	19.0	0	0	7.6500	2
767	3	0	30.5	0	0	7.7500	1
72	2	1	21.0	0	0	73.5000	2
235	3	0	28.0	0	0	7.5500	2
37	3	1	21.0	0	0	8.0500	2

X_test							
	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
862	1	0	48.0	0	0	25.9292	2
223	3	1	28.0	0	0	7.8958	2
84	2	0	17.0	0	0	10.5000	2
680	3	0	28.0	0	0	8.1375	1
535	2	0	7.0	0	2	26.2500	2
...
796	1	0	49.0	0	0	25.9292	2
815	1	1	28.0	0	0	0.0000	2
629	3	1	28.0	0	0	7.7333	1
421	3	1	21.0	0	0	7.7333	1
448	3	0	5.0	2	1	19.2583	0

y_train	y_test
381	1
309	1
516	1
120	0
570	1
...	...
715	0
767	0
72	0
235	0
37	0

```
print(X_train.shape)  
print(X_test.shape)  
print(y_train.shape)  
print(y_test.shape)
```

(712, 7)
(179, 7)
(712,)
(179,)

Data kemudian dibagi menjadi set pelatihan dan pengujian menggunakan fungsi train_test_split. Set pelatihan digunakan untuk melatih model, sedangkan set pengujian digunakan untuk mengevaluasi kinerjanya. Proporsi pembagian adalah 80% untuk pelatihan dan 20% untuk pengujian dengan parameter random_state=1 untuk memastikan hasil yang konsisten.

Modeling

Random Forest Classifier

Logistic Regression

XGBoost

K-Nearest Neighbors

Modeling

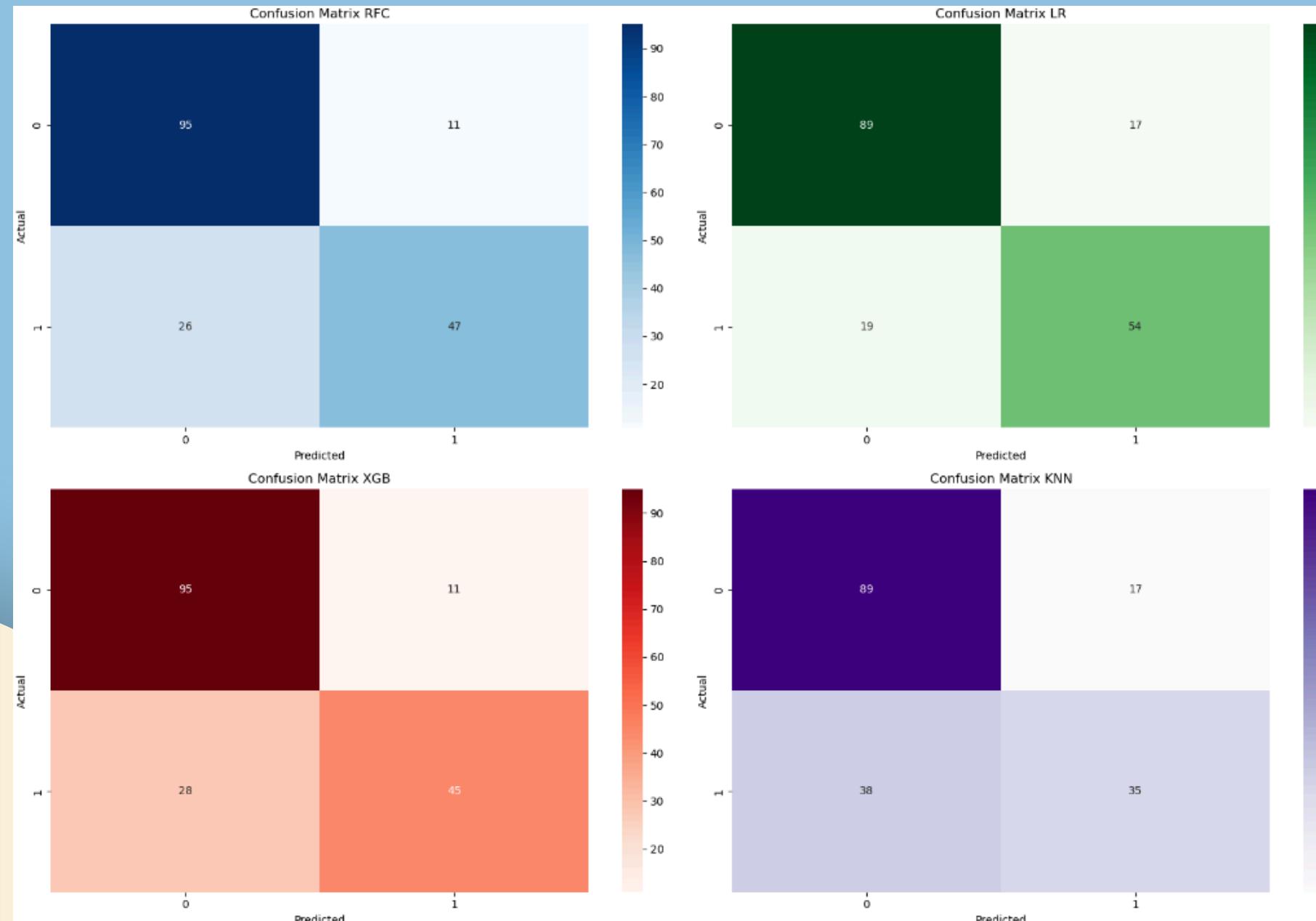
Actual	RFC_Pred	LR_Pred	XGB_Pred	KNN_Pred
862	1	1	1	1
223	0	0	0	0
84	1	1	1	0
680	0	1	1	1
535	1	1	1	1
...
796	1	1	1	1
815	0	0	1	0
629	0	0	0	0
421	0	0	0	0
448	1	1	1	1

179 rows × 5 columns

Data Test Prediction Comparison

- Tabel ini menampilkan hasil prediksi dari empat model machine learning: Random Forest Classifier (RFC), Logistic Regression (LR), XGBoost (XGB), dan K-Nearest Neighbors (KNN), dibandingkan dengan nilai aktual.
- Tabel ini memungkinkan kita untuk melihat secara langsung bagaimana perbedaan algoritma menghasilkan variasi dalam prediksi, baik dalam hal prediksi yang akurat maupun kesalahan yang terjadi.

Modeling



Confusion Matrix Comparison

- Confusion matrix untuk masing-masing model memberikan informasi lebih rinci tentang performa setiap model dalam memprediksi kelas positif dan negatif.
- Dalam hal prediksi kelas negatif, RFC dan XGB memberikan hasil yang serupa, namun RFC sedikit lebih baik dalam mendeteksi kelas positif dibanding XGB.
- Logistic Regression menampilkan keseimbangan yang baik dalam mendeteksi kedua kelas, namun memiliki tingkat kesalahan yang lebih tinggi pada kelas negatif.
- KNN, di sisi lain, memiliki kesalahan yang lebih besar pada prediksi kelas positif, yang menunjukkan bahwa model ini kurang baik dalam mendeteksi kelas positif dibanding yang lain.

Modeling

	Feature	Importance
1	Sex	0.263622
5	Fare	0.261463
2	Age	0.256590
0	Pclass	0.090348
3	SibSp	0.052552
4	Parch	0.040127
6	Embarked	0.035298

Feature Importance Random Forest Classifier

Tabel ini menunjukkan peringkat feature importance yang dihasilkan oleh model Random Forest Classifier. Fitur yang paling berpengaruh dalam model ini adalah "Sex" (jenis kelamin) dengan bobot sekitar 26.36%, diikuti oleh "Fare" (harga tiket) sebesar 26.15%, dan "Age" (usia) sebesar 25.66%. Fitur lainnya seperti "Pclass", "SibSp", "Parch", dan "Embarked" memiliki pengaruh yang lebih rendah dalam memprediksi hasil.

Modeling

Models Performances Comparison

Tabel ini menunjukkan evaluasi performa empat model machine learning berdasarkan akurasi, precision, recall, F1-score, dan ROC AUC. Hasil utama:

- Logistic Regression memiliki akurasi tertinggi (0.798) dan juga unggul dalam recall, F1-Score, serta ROC AUC, yang menunjukkan kinerja yang seimbang.
- Random Forest Classifier memiliki precision tertinggi (0.810), yang berarti lebih baik dalam meminimalkan prediksi positif palsu.
- XGBoost menunjukkan kinerja yang cukup baik dengan precision dan recall yang seimbang.
- K-Nearest Neighbors menunjukkan performa paling rendah di semua metrik.

Models Performances Comparison					
	Accuracy	Precision	Recall	F1-Score	ROC_AUC
Random Forest Classifier	0.793296	0.810345	0.643836	0.717557	0.770031
Logistic Regression	0.798883	0.760563	0.739726	0.75	0.789674
XGBoost	0.782123	0.803571	0.616438	0.697674	0.756332
K-Nearest Neighbors	0.692737	0.673077	0.479452	0.56	0.659537

Kesimpulan

Berdasarkan keseluruhan analisis yang dilakukan dalam proyek ini kita dapat menyimpulkan bahwa:

- Prediksi kelangsungan hidup penumpang Titanic sangat dipengaruhi oleh beberapa faktor utama seperti kelas penumpang, jenis kelamin, usia dan tarif tiket yang dibayarkan.
- Model yang digunakan seperti Random Forest, Logistic Regression, XGBoost dan K-Nearest Neighbors memberikan hasil yang bervariasi dengan Logistic Regression memiliki performa terbaik dalam hal akurasi dan keseimbangan deteksi kelas.
- Selain itu faktor-faktor seperti jenis kelamin dan harga tiket terbukti memiliki pengaruh signifikan terhadap tingkat keselamatan penumpang.



THANK
YOU