

EDA Project: Titanic Dataset Analysis

Objective

To explore the Titanic dataset and uncover insights about passenger demographics, travel class, and survival patterns.

Step 1: Import Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Step 2: Load Dataset

```
df = pd.read_csv("Titanic.csv")
df.head()
```

Step 3: Basic Information

```
df.info()
df.describe()
df.isnull().sum()
```

Step 4: Data Cleaning

```
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
df.drop(['Cabin'], axis=1, inplace=True)
```

Step 5: Univariate Analysis

```
sns.countplot(x='Survived', data=df)
plt.title('Survival Count')
```

```
sns.histplot(df['Age'], kde=True)
plt.title('Age Distribution')
```

Step 6: Bivariate Analysis

```
sns.barplot(x='Sex', y='Survived', data=df)
plt.title('Survival Rate by Gender')
```

```
sns.barplot(x='Pclass', y='Survived', data=df)
plt.title('Survival Rate by Class')
```

```
sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Age vs Survival')
```

Step 7: Correlation Analysis

```
corr = df.corr(numeric_only=True)
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
```

Step 8: Insights

- Women had higher survival rates than men.
- Passengers in 1st class had higher chances of survival.
- Younger passengers tended to survive more.

Step 9: Optional – Feature Engineering

```
df['FamilySize'] = df['SibSp'] + df['Parch'] + 1
sns.barplot(x='FamilySize', y='Survived', data=df)
```

Step 10: Summary

Conclusion:

1. Survival chances depended heavily on gender and class.
2. Age also influenced outcomes slightly.
3. Data cleaning was crucial due to missing values in Age, Cabin, and Embarked.