

Titanic Dataset - Exploratory Data Analysis (EDA)

Objective

To understand the dataset using descriptive statistics and visualizations.

Tools Used

- Pandas: For data manipulation and statistics
- Matplotlib & Seaborn: For data visualization

1. Summary Statistics

Descriptive statistics of numeric features:

Feature	Count	Mean	Std Dev	Min	25%	Median	75%	Max
PassengerId	891	446.00	257.35	1	223.5	446.0	668.5	891.0
Survived	891	0.384	0.487	0	0	0	1	1
Pclass	891	2.309	0.836	1	2	3	3	3
Age	714	29.70	14.53	0.42	20.13	28.0	38.0	80.0
SibSp	891	0.523	1.103	0	0	0	1	8
Parch	891	0.382	0.806	0	0	0	0	6
Fare	891	32.20	49.69	0	7.91	14.45	31.0	512.3

2. Visualizations

Histograms and Boxplots for:

Titanic Dataset - Exploratory Data Analysis (EDA)

- Age
- Fare
- SibSp
- Parch

These plots help in understanding the distribution and identifying outliers.

3. Feature Relationships

Correlation Matrix & Pairplot (Python code):

```
import seaborn as sns
```

```
import matplotlib.pyplot as plt
```

```
corr = df[['Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']].corr()
```

```
sns.heatmap(corr, annot=True, cmap='coolwarm')
```

```
plt.title("Correlation Matrix")
```

```
plt.show()
```

```
sns.pairplot(df[['Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']], hue='Survived')
```

```
plt.show()
```

Folder Descriptions

- data/: Contains the original dataset file.
- notebooks/: Jupyter notebooks for EDA work.

Titanic Dataset - Exploratory Data Analysis (EDA)

- scripts/: (Optional) Python scripts for reusable code.
- outputs/: Saved visualizations and statistical summaries.
- requirements.txt: Python package dependencies.