EXPLORATORY DATA ANALYSIS (EDA) – TITANIC DATASET

```
[1]  import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
```

```
df = pd.read_csv('/content/drive/My Drive/titanic/train.csv')
```

Import pandas, seaborn, and matplotlib.pyplot to handle data and create visualizations. We load the Titanic dataset from Google Drive using pd.read_csv() and store it in a DataFrame called df.

```
[9]  print("Jumlah data duplikat:", df.duplicated().sum())

     Jumlah data duplikat: 0
```

Use df.duplicated().sum() to check for duplicate rows, and the result shows there are none in the dataset.

```
[10] print("Missing value sebelum ditangani:")
     print(df.isna().sum())
```

```
Missing value sebelum ditangani:
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
```

By using df.isna().sum(), we identify the number of missing values in each column, revealing that the Age column has 177 missing entries, Cabin has 687, and Embarked has 2, which indicates the need for proper data cleaning before analysis.

```
[10] print("Missing value sebelum ditangani:")
     print(df.isna().sum())
```

```
Missing value sebelum ditangani:
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
```

By using df.isna().sum(), we identify the number of missing values in each column, revealing that the Age column has 177 missing entries, Cabin has 687, and Embarked has 2, which indicates the need for proper data cleaning before analysis.

```
df['Age'].fillna(df['Age'].median(), inplace=True)
```

```
<ipython-input-11-63d4fb902a4f>:1: FutureWarning: A value is trying to be set on a
The behavior will change in pandas 3.0. This inplace method will never work becaus

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.metho


  df['Age'].fillna(df['Age'].median(), inplace=True)
```

```
[12] df.drop('Cabin', axis=1, inplace=True)
```

```
[13] df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
```

To clean the dataset, we replaced missing values in the Age column with the median value to avoid distortion by outliers, dropped the Cabin column entirely due to a large number of missing entries, and filled the missing values in the Embarked column with its most frequent value (mode), ensuring the dataset is complete and ready for analysis.

```
[15] df.describe(include='all')
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
count	891.000000	891.000000	891.000000	891	891	891.000000	891.000000	891.000000	891	891.000000	891
unique	NaN	NaN	NaN	891	2	NaN	NaN	NaN	681	NaN	3
top	NaN	NaN	NaN	Dooley, Mr. Patrick	male	NaN	NaN	NaN	347082	NaN	S
freq	NaN	NaN	NaN	1	577	NaN	NaN	NaN	7	NaN	646
mean	446.000000	0.383838	2.308642	NaN	NaN	29.361582	0.523008	0.381594	NaN	32.204208	NaN
std	257.353842	0.486592	0.836071	NaN	NaN	13.019697	1.102743	0.806057	NaN	49.693429	NaN
min	1.000000	0.000000	1.000000	NaN	NaN	0.420000	0.000000	0.000000	NaN	0.000000	NaN
25%	223.500000	0.000000	2.000000	NaN	NaN	22.000000	0.000000	0.000000	NaN	7.910400	NaN
50%	446.000000	0.000000	3.000000	NaN	NaN	28.000000	0.000000	0.000000	NaN	14.454200	NaN
75%	668.500000	1.000000	3.000000	NaN	NaN	35.000000	1.000000	0.000000	NaN	31.000000	NaN
max	891.000000	1.000000	3.000000	NaN	NaN	80.000000	8.000000	6.000000	NaN	512.329200	NaN

Using df.describe(include='all'), we generate a summary of both numerical and categorical columns, showing statistics like mean, median, standard deviation for numeric data (Age, Fare, etc.) and frequency distribution for categorical data (Sex, Embarked), which helps us understand the overall distribution, spread, and common values in the dataset.

THANK YOU