Titanic Dataset Analysis – Step-by-Step Summary

Step 1: Load and Preview the Data

Start by reading the dataset into a Data Frame using pandas. Examine the first few rows to understand its structure.

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

Step 2: Understand the Columns

Identify the meaning of each column:

- PassengerId: Unique ID for each passenger
- Survived: Survival status (0 = No, 1 = Yes)
- Pclass: Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd)
- Name, Sex, Age: Self-explanatory
- SibSp: # of siblings/spouses aboard
- Parch: # of parents/children aboard
- Ticket: Ticket number
- Fare: Ticket fare
- Cabin: Cabin number (often missing)
- Embarked: Port of embarkation (C, Q, S)

Step 3: Check for Missing Data

Identify missing values in the dataset:

python

Copy code

```
df.isnull (). sum()
```

Common missing values:

- Age
- Cabin (often heavily missing)
- Embarked (sometimes missing)

Step 4: Descriptive Statistics

Generate basic statistics to understand numerical columns.

python

Copy code

```
df.describe ()
```

You'll see count, mean, std, min, max, etc. for Age, Fare, etc.

Step 5: Clean the Data

Handle missing or inconsistent data:

- Fill missing Age values (e.g., with median)
- Drop or impute Cabin
- Fill Embarked with the most common value

Step 6: Explore Survival Rates

Basic survival insights:

- Overall survival rate: df['Survived']. mean()
- By sex: df.groupby('Sex') ['Survived']. mean()
- By class: df.groupby('Pclass') ['Survived']. mean()
- By age group: create bins like children, adults, elderly

Step 7: Visualizations (optional)

Use matplotlib or seaborn to visualize data:

- Bar plots for survival by class/sex
- Histograms for age and fare
- Heatmap for correlation matrix

Step 8: Feature Engineering (optional)

Create new columns:

- Family Size = SibSp + Parch + 1
- Categorized age groups
- Extract title from Name

Step 9: Modeling (optional)

Train a simple predictive model:

- Logistic Regression or Decision Tree
- Use features like Pclass, Sex, Age, etc.
- Evaluate using accuracy or cross-validation