Titanic Data Analysis & Visualization

This project performs Exploratory Data Analysis (EDA) on the Titanic dataset using Python. The goal is to understand the structure of the data, clean it, and extract insights through statistics and visualizations.

Objective

Understand the Titanic dataset using descriptive statistics.

Visualize data distributions and relationships.

Identify missing values and clean the dataset.

Reveal patterns related to passenger survival.

Tools & Libraries

Pandas – Data manipulation & summary statistics

Matplotlib – Plotting histograms & boxplots

Seaborn – Pairplots, heatmaps, enhanced visuals

Plotly (optional) – For future interactive visualizations

Dataset

File: Titanic-Dataset.csv The dataset contains information about passengers on the Titanic, including age, class, sex, fare, and survival status.

Key Steps

1. Data Cleaning

Fill missing Age with median.

Fill missing Embarked with mode.

Drop Cabin column due to too many missing values.

2. Summary Statistics

Mean, median, mode

Standard deviation

Skewness and kurtosis

3. Visualizations

Histograms – Distributions of numeric features

Boxplots – Outlier detection

Heatmap – Correlation matrix

Pairplot – Relationship between key features (Survived, Age, Fare, etc.)

Sample Output

Mean Fare: $32.20

Age Skewness: 0.41

Strong correlation between Fare and P class

Survivors were more likely to be 1st class passengers

How to Run

Step 1: Install dependencies

pip install pandas matplotlib seaborn

Step 2: Run the Python file

python titanic\_analysis.py

### Future Work

\* Add interactive visualizations using Plotly or Streamlit

\* Build a \*predictive ML model\* (Logistic Regression / Random Forest)

\* Export stats summary to \*Excel or PDF\*

### Author

gopavarapu naga greeshma

B.Tech | Devops | Presidency University

Email: [gopavarapugreeshma@gmail.com](mailto:gopavarapugreeshma@gmail.com)