

# Gaurav Kakde | Data Science | Week 2 Task

## 1. Task Overview

This task involved performing data cleaning and exploratory data analysis (EDA) on the Titanic dataset from Kaggle.

The objective was to uncover survival patterns and prepare the dataset for further modeling. I used Python and libraries like Pandas, NumPy, Matplotlib, and Seaborn.

## 2. Tools and Platforms Used

- Python 3.x
- Jupyter Notebook
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Kaggle (for dataset)

## 3. What I Did

- Loaded and explored the Titanic dataset
- Handled missing values (Age, Embarked), and dropped columns like Cabin
- Removed outliers in Fare column using 99th percentile capping
- Encoded categorical variables like Sex and Embarked
- Created visualizations for survival analysis based on features like Sex, Pclass, Age, and Fare
- Generated insights from the charts to understand survival trends

## 4. Screenshots or Links

GitHub link :<https://github.com/phoenix845/titanic-eda-data-cleaning.git>

Key file: Titanic\_EDA\_Report.pdf

## 5. Challenges Faced & Solutions

- Some columns like 'Cabin' had excessive missing values. I dropped it after assessing its sparsity.

## **Gaurav Kakde | Data Science | Week 2 Task**

- Fare had extreme outliers; I applied 99th percentile capping to handle them.
- For visualizations, interpreting legends and labels clearly was necessary, so I customized legends and annotations for better clarity.
- Unicode characters in PDF generation caused errors - resolved by replacing bullets with dashes.