# **📘 Assignment – Titanic Data Analysis with Probability**

This assignment will take you through a **real-world dataset** (Titanic 🚢) and help you practice:

* **Data Cleaning** (handling missing values)
* **Exploratory Data Analysis (EDA)** using **Pandas & Numpy**
* **Visualizations** with **Matplotlib & Seaborn**
* **Empirical vs Theoretical Probability** concepts

By the end, you will not only analyze the Titanic dataset but also **compare real probabilities with theoretical assumptions** to see how they differ.

## **Dataset (**[**https://www.kaggle.com/datasets/yasserh/titanic-dataset**](https://www.kaggle.com/datasets/yasserh/titanic-dataset)**)**

We’ll use the **Titanic dataset (train.csv)** which contains information about passengers such as:

* Survived → 0 = Did not survive, 1 = Survived
* Pclass → Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
* Sex → Male/Female
* Age → Passenger age
* SibSp → # of siblings/spouses aboard
* Parch → # of parents/children aboard
* Fare → Ticket price
* Embarked → Port of embarkation

## **Part A – Data Exploration (Pandas & Numpy)**

1. **Load the dataset**
   * Use Pandas to load the CSV.
   * Display the first 10 rows to get a quick look.
   * Print the column names and their data types.
   * Use .describe() to view basic statistics (mean, min, max, etc.).
2. **Handle Missing Values**
   * Check for missing values with .isnull().sum().
   * Replace missing **Age** values with the **mean age** (using Numpy).
   * Fill missing **Embarked** values with the **most common port**.
3. **Basic Questions**
   * How many passengers were onboard?
   * What is the average **Age**?
   * What is the average **Fare**?
   * How many **males** and **females** were there?

## **Part B – Data Analysis (Pandas)**

1. **Survival Analysis**
   * How many survived vs. did not survive?
   * What percentage of passengers survived?
2. **Survival by Class (Pclass)**
   * Find survival rates for 1st, 2nd, and 3rd class.
   * Which class had the **highest chance of survival**?
3. **Survival by Age Group**
   * Divide passengers into age groups:
     + 0–12 = Child
     + 13–19 = Teen
     + 20–40 = Adult
     + 41+ = Senior
   * Calculate survival rate for each group.

## **Part C – Visualization (Matplotlib & Seaborn)**

1. **Matplotlib**
   * Bar chart: Number of survivors (0 vs 1).
   * Histogram: Distribution of passenger ages (bins = 20).
   * Pie chart: Gender distribution.
2. **Seaborn**
   * Countplot: Survived by Sex (Who survived more?).
   * Boxplot: Fare by Pclass (Which class paid more?).
   * Violinplot: Age by Survived (Age distribution among survivors).
   * Heatmap: Correlation between numerical columns (Age, Fare, Pclass, Survived).

## **Part D – Probability Concepts**

Now let’s connect Titanic survival with **Probability Theory**.

### **1. Empirical Probability (From Data)**

Empirical probability = **Favorable Outcomes ÷ Total Outcomes**

Examples:

* Probability a passenger survived:  
   P(Survived) = Survivors / Total Passengers
* Probability a **female** survived:  
   P(\text{Female Survived}) = Surviving Females / Total Females
* Probability a **male** survived:  
   P(\text{Male Survived}) = Surviving Males / Total Males
* Compare probabilities across **Pclass**.

👉 These values come directly from the dataset.

### **2. Theoretical Probability (Assumptions)**

Imagine survival was **random** or equally likely for everyone.

* Then, P(Survived)=0.5 (like flipping a coin).
* Each class (Pclass) should have the **same chance** of survival → 1 / 3
* Gender should not matter.

👉 These values are **assumptions**, not based on data.

## **Part E – Insights & Reporting**

1. Answer the following in your notebook/report:
   * Did women survive at a higher rate than men?
   * Which class had the best survival chance?
   * Did age play a role in survival?
   * Was Fare (ticket price) related to survival?
   * How do **empirical probabilities** differ from **theoretical probabilities**?
2. Summarize your analysis in **5–7 bullet points**.
3. Save your work as **Titanic\_Analysis.ipynb** with:

* Cleaned data
* All visualizations
* Probability calculations
* Final insights