**Task 5: Exploratory Data Analysis (EDA) Tools: Python (Pandas, Matplotlib, Seaborn)**

**Objective: Extract insights using visual and statistical exploration.**

**Tools Needed: (install)**

* **Python (Libraries: pandas, numpy, matplotlib, seaborn)**
* **Jupyter Notebook**

**Step 1. Setup & Load Data**

* **Note the shape of the dataset**
* **Mention what each column appears to represent**

**Step 2. Understand Data Structure**

* **Count of null values**
* **Which columns are numerical/categorical**
* **Any anomalies in data types (e.g., numbers stored as strings)**

**Step 3. Visualization (Pairplot, Heatmap)**

* **Pairplot Insights:**

**- There is a visible separation between survivors and non-survivors in age and fare.**

**- Fare is positively skewed with some high-value outliers.**

* **Heatmap Insights:**

**- Fare and Pclass show a negative correlation (lower class = lower fare).**

**- Survival has weak positive correlation with Fare and Age, suggesting higher fare (wealthier) had better survival chances.**

**Step 4. Relationships and Trends Identified:**

**- Survival is strongly related to gender: 75% of women survived vs. 20% of men.**

**- Higher fare is associated with higher survival: passengers in 1st class had better chances.**

**- Age and survival show mild trends: younger passengers had slightly higher survival.**

**- Pclass and Fare are negatively correlated.**

**- Fare has outliers; some passengers paid extremely high prices.**

**Step 5. Observations for visual**

**1. Histograms – Distribution of a Single Variable**

**Histogram of Age**

* **The age distribution is right-skewed.**
* **Most passengers were between 20 and 40 years old.**
* **There is a noticeable drop in frequency after age 50.**
* **A small number of children (under 10) were also onboard.**

**Histogram of Fare**

* **Fare values are highly right-skewed.**
* **Most passengers paid under $100, but there are a few outliers above $200.**
* **The peak (mode) is around $10–$30, indicating this was the most common fare range.**

**2. Boxplots – Distribution & Outliers by Category**

**Boxplot of Age**

* **The median age is around 28–30 years.**
* **There are a few outliers below age 5 (children).**
* **The age spread is relatively even, with no extreme upper outliers.**

**Boxplot of Fare**

* **The majority of fares were below $50.**
* **There are many outliers, especially above $100, indicating a few passengers paid premium prices.**
* **The median fare is around $15.**

**Boxplot: Age by Passenger Class**

* **1st class passengers tend to be older, with a median age over 35.**
* **3rd class passengers are generally younger, with more variability.**
* **This suggests a relationship between socioeconomic status and age.**

**3. Scatterplots – Relationships Between Two Numeric Variables**

**Scatterplot: Age vs Fare**

* **Most data points are clustered between ages 20–40 and fares $0–$50.**
* **A few older passengers paid very high fares (outliers).**
* **No strong linear relationship is visible between age and fare.**

**Scatterplot: Age vs Fare Colored by Survival**

* **Survivors are often in the higher fare category.**
* **Many non-survivors cluster in the lower-fare range, possibly 3rd class.**
* **There is no obvious trend based on age alone, but fare has a visible impact on survival.**

**Step 6. Summary**

**General Overview**

* **The dataset contains information about passengers on the Titanic, including demographic features (age, gender), ticket and fare details, and whether or not they survived.**
* **There are some missing values in important columns like Age and Cabin.**

**Key Insights from Univariate Analysis**

* **Age Distribution: Most passengers were between 20 and 40 years old. A few children and elderly were present.**
* **Fare Distribution: The majority of passengers paid under $50, but there are some extremely high fares, indicating wealthy passengers.**
* **Gender: More males than females were on board.**
* **Passenger Class: Most passengers traveled in 3rd class, followed by 1st and 2nd.**

**Bivariate & Multivariate Relationships**

* **Survival by Gender:**
  + **A much higher percentage of women survived compared to men.**
* **Survival by Passenger Class:**
  + **1st class passengers had the highest survival rate, followed by 2nd class.**
  + **3rd class passengers had the lowest survival rate.**
* **Fare vs Survival:**
  + **Passengers who paid higher fares were more likely to survive.**
* **Age vs Survival:**
  + **Younger children had relatively high survival rates, but overall, age was not a strong predictor.**

**Correlation Analysis**

* **Fare and Pclass are negatively correlated – higher class = higher fare.**
* **Survived has weak positive correlation with Fare, and a slight negative correlation with Pclass.**
* **No strong correlation between Age and Survived.**

**Missing Values**

* **Age and Cabin columns contain missing values.**
* **Cabin has a significant number of missing entries, suggesting it may not be useful without cleaning.**

**Outliers**

* **Detected in Fare (some passengers paid significantly more than average).**
* **Outliers were visualized using boxplots.**

**Step 7. Save & share.**