Task 5: Exploratory Data Analysis (EDA) – Titanic Dataset

1. Objective

The main goal of this project was to explore the famous **Titanic dataset** from Kaggle and find patterns that explain who survived the disaster and why.
This process is called **Exploratory Data Analysis (EDA)** — where we clean the data, understand each column, visualize relationships, and summarize our findings.

By the end of this analysis, the aim was to:

- Handle missing or messy data correctly
- Understand trends and distributions
- Visualize which factors most affected survival
- Build a clear picture of the data before modeling

---

2. Tools Used

- **Python** – Main programming language
- **Pandas, NumPy** – For data handling and statistics
- **Matplotlib, Seaborn** – For data visualization
- **VS Code + Jupyter Notebook** – For running and documenting the analysis

---

3. About the Dataset

The Titanic dataset comes from Kaggle's "Machine Learning from Disaster" competition.
It includes information about passengers such as:

- **Pclass:** Passenger class (1st, 2nd, 3rd)
- **Name, Sex, Age:** Personal details
- **SibSp, Parch:** Number of siblings/spouses and parents/children aboard
- **Ticket, Fare:** Ticket number and price
- **Cabin:** Cabin number (mostly missing)
- **Embarked:** Port of boarding (C = Cherbourg, Q = Queenstown, S = Southampton)
- **Survived:** Target column (1 = survived, 0 = not survived)

In total, there are **891 rows and 12 columns** in the training data.

---

4. Data Cleaning

Before any visualizations, I checked for missing values and inconsistent data.
Here's what I found and fixed:

- **Age** had several missing entries — I replaced them with the **median age** so the distribution stayed realistic.
- **Embarked** had a few missing values — filled them with the **most common port** (mode).
- **Cabin** was almost completely missing, so I **removed it** from the dataset.

After cleaning, no missing data remained, and all columns were ready for analysis.

---

5. Understanding the Data

I first explored the general structure using .info() and .describe().
Some quick findings:

- Around **65% of passengers were male**, and **most traveled in 3rd class**.
- The **average age** was about **29 years**.
- Ticket **fares** varied widely — from very cheap to extremely high (first-class luxury).

This already hinted that **gender, class, and fare** might play big roles in survival chances.

---

6. Univariate Analysis

This step focuses on **one column at a time** — to see how values are distributed.

- **Age:** Most passengers were between 20–40 years old, with a few very young and older travelers.
- **Fare:** The fare distribution was highly skewed; most passengers paid low fares, while a few paid very high prices.
- **Sex:** More males than females were onboard.
- **Class:** 3rd class had the most passengers.
- **Embarked:** Most passengers boarded from **Southampton (S)**.

These insights helped set the stage for seeing how these features interact with survival later.

---

7. Bivariate Analysis

Now I compared two variables — mainly to see how they relate to **Survival**.

Survival by Gender

This was the most obvious pattern:

- **Females had a much higher survival rate** than males.
- The “women and children first” rule clearly shows in the data.

Survival by Passenger Class

- **1st class passengers** had the highest survival rate.
- **3rd class** passengers had the lowest.
  This shows that social class (and possibly access to lifeboats) played a huge role.

Survival by Age

- **Younger passengers**, especially children, had better survival chances.
- Older passengers were less likely to survive.

Survival by Fare

- People who paid higher fares (first class) were more likely to survive.
- This again connects survival with wealth and access.

Survival by Port of Embarkation

- Passengers who boarded from **Cherbourg (C)** survived more often than those from Southampton (S) or Queenstown (Q).

---

8. Correlation and Multivariate Analysis

To understand how numeric features relate, I created a **correlation heatmap**.

Key relationships:

- Fare was positively correlated with Survived (higher fare → more survival).
- Pclass was negatively correlated (lower class number → higher chance of survival).
- Other variables like Age, SibSp, and Parch had weak correlations individually but could still be useful together.

A **pairplot** (scatter matrix) also showed clear separation between survivors and non-survivors based on fare and class.

---

9. Key Insights and Patterns

After all visualizations and analysis, here are the main takeaways:

1. **Gender:** Women had a much higher chance of survival than men.
2. **Class:** First-class passengers had far better outcomes than second or third class.
3. **Fare:** Higher ticket prices (indicating wealth) were strongly linked to survival.
4. **Age:** Children and young adults were more likely to survive.
5. **Embarked Port:** Passengers from Cherbourg seemed to have better survival odds.

Together, these patterns reflect a mix of **social status, location, and gender priority** affecting survival.

---

10. Exporting Results

After cleaning and analysis, I saved the cleaned version of the dataset as **train_cleaned.csv**.
This file can be used later for machine learning models like logistic regression or decision trees.

---

11. Conclusion

This EDA helped transform a raw dataset into clear, meaningful insights.
Through visual and statistical exploration, we learned that:

- Survival wasn't random — it depended strongly on **gender, class, and age**.
- The dataset required minimal but crucial cleaning to make it usable.
- Visual storytelling made these patterns easy to understand at a glance.

Overall, this project strengthened my understanding of:

- How to clean and prepare real-world data
- How to use Python visualization libraries effectively
- How to interpret and communicate data-driven findings

---

12. Deliverables

- Titanic_EDA.ipynb — Jupyter Notebook with full analysis
- train_cleaned.csv — Cleaned dataset
- EDA_Report.pdf — This report summary

Final Note

This EDA not only explains **who survived** the Titanic tragedy but also demonstrates a structured approach to real-world data analysis — from cleaning to storytelling through visuals.