**Titanic Dataset Analysis Report**

***~Disha Hegde***

**1️⃣ Introduction**

*This project focuses on analysing the famous Titanic passenger dataset to explore patterns related to survival during the maritime disaster of 1912. The dataset contains demographic and travel-related details of passengers, including survival outcome, age, gender, class, fare, family relationships, and embarkation port. The aim is to clean and enhance the data, perform exploratory analysis, and identify key factors influencing survival.*

**2️⃣ Project Goal**

*The goal of this project is to clean, enhance, and analyse the Titanic dataset to uncover patterns that influenced passenger survival during the disaster. By addressing missing data, engineering meaningful features, and visualizing key relationships (such as survival by gender, age, class, and family size), the project aims to gain insights that could help in understanding the social and economic factors affecting survival rates. This lays the groundwork for potential predictive modelling and further research.*

**3️⃣ Data Cleaning Steps**

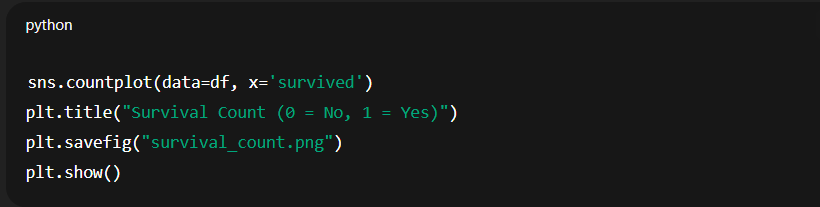
* ***Initial inspection:*** *Identified missing values in several columns, particularly age, deck, and embark\_town.*
* ***Column removal:*** *Dropped sparse or redundant columns — deck, embark\_town, and alive.*
* ***Missing value treatment:***
  + *Filled missing age values with the median age.*
  + *Dropped rows with remaining missing values.*
* ***Data type conversion:*** *Converted sex and embarked columns to categorical types for better efficiency.*
* ***Text standardization:*** *Capitalized values in the who column for consistency.*
* ***Feature engineering:*** *Created a family\_size column to represent total family members onboard (sibsp + parch + 1).*

**4️⃣ Visualizations and Findings**

* ***Survival count plot:*** *Showed that the majority of passengers did not survive.*
* ***Survival by gender:*** *Significantly higher survival rate among females compared to males.*
* ***Age distribution:*** *Most passengers were between 20–40 years old; the distribution was slightly right-skewed.*
* ***Correlation heatmap:***
  + *fare had a positive correlation with survival, indicating higher-paying passengers had better survival chances.*
  + *pclass was negatively correlated with survival (1st class passengers had better survival outcomes).*
  + *Strong correlation between sibsp and parch (indicating family groupings).*
* ***Violin plot (age vs survival by gender):***
  + *Females of all age groups, particularly 20–40, had higher survival rates.*
  + *Males had lower survival rates, mostly limited to younger passengers.*

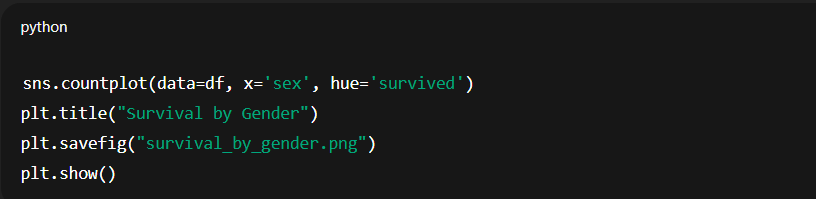
**1. Survival Count:**

* ***Description:*** *This bar plot shows the number of survivors vs. non-survivors*.



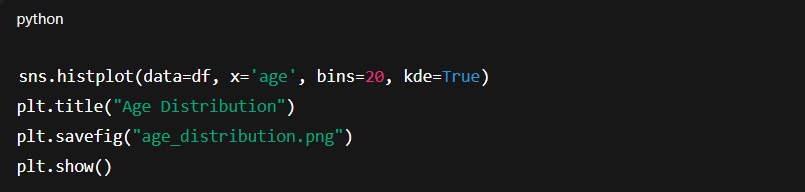
**2. Survival by Gender:**

* ***Description:*** *Bar plot grouped by gender and survival.*



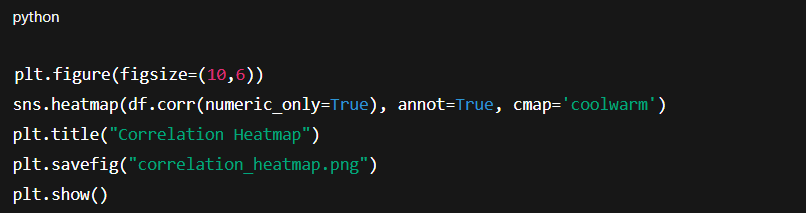
**3. Age Distribution:**

* ***Description:*** *Histogram of age distribution with KDE overlay.*



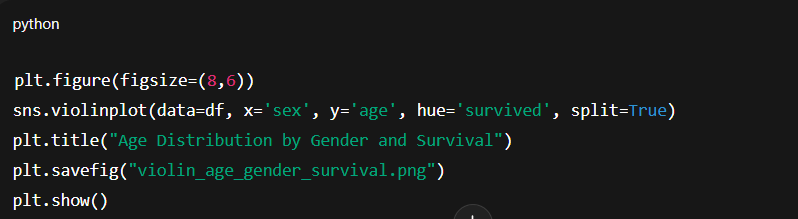
**4. Correlation Heatmap:**

* ***Description:*** *Heatmap showing correlation between numerical features.*



**5. Violin Plot: Age by Gender and Survival:**

* ***Description:*** *Age distribution split by gender and survival outcome.*



**5️⃣ Conclusion**

*The analysis highlighted the critical roles of* ***gender, class, fare****, and* ***family size*** *in influencing survival chances on the Titanic. These findings align with historical records, showing that women, children, and first-class passengers were prioritized during rescue efforts. The cleaned and enhanced dataset is now ready for further use, such as building predictive models or performing advanced statistical analysis.*