**Titanic Dataset EDA Report**

**Name:** Harsh Goyal

**Enrollment no.** 23114034

**Course code:** DAI-101

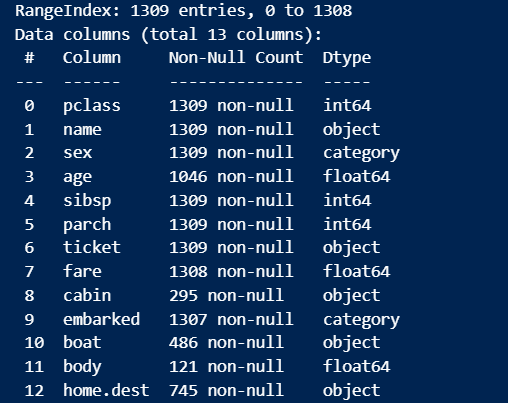
**Professor:** Mrs. Shalini Priya

**INTRODUCTION:**

This report presents an in-depth exploratory data analysis (EDA) of the scikit-learn Titanic dataset, which contains information about the passengers aboard the Titanic, including their fare, age, sex, class, and other attributes. This dataset doesn’t include their survival status, though. The objective of this analysis is to analyze the features both individually and their relationships with each other.

**DATASET OVERVIEW:**

* Dataset contains 1309 entries and 13 columns. Basic information regarding columns is given below:



* The columns: age, fare, cabin, embarked, boat, body, home.dest are the ones which have missing values.
* Even among them, the columns: cabin, boat and body have less than even 1/3rd of the data, so we will not be considering them in our analysis.
* For the columns: age and fare, we fill the missing values with median of the respective columns. As for categorical columns, the missing values are being filled by the mode of the respective columns.
* Duplicates are not present in the dataset.
* The outliers in the numerical columns are found using IQR method and thus removed from the dataset.
* The columns Name and parch are removed from the dataset as these were not contributing any valuable information in the dataset.

**UNIVARIATE ANALYSIS:**

Univariate analysis involves examining each variable individually to understand its distribution and characteristics.

**Summary Statistics:**

pclass: The mean passenger class is approximately 2.47, indicating that most passengers were in the second or third class.

age: The mean age is around 28.55 years, with a standard deviation of 9.31 years. The age ranges from 3 to 54 years.

sibsp: The mean number of siblings/spouses aboard is 0.27, with most passengers having no siblings or spouses.

fare: The mean fare is approximately 17.10, with a standard deviation of 13.72. The fare ranges from 0 to 65.

**Frequency Distributions:**

sex: There are 698 males and 327 females in the dataset.

embarked: The majority of passengers embarked from Southampton (S: 742), followed by Cherbourg (C: 173) and Queenstown (Q: 110).

**Visualizations:**

Histograms: Histograms are plotted for numerical columns (pclass, age, sibsp, fare) to visualize their distributions.

Count Plots: Count plots are used to visualize the frequency distributions of categorical variables (sex, embarked).

**Key Observations:**

* The distribution of age is slightly right-skewed, with most passengers being between 20 and 40 years old.
* The fare distribution is highly right-skewed, indicating that most passengers paid a lower fare, with a few paying significantly higher fares.
* The majority of passengers were male and embarked from Southampton.

**BIVARIATE ANALYSIS:**

**1. Correlation Matrix**:

* Confirms strong negative correlation between `pclass` and `fare`, indicating higher-class tickets were more expensive.
* Weak correlations among other numerical variables, suggesting limited direct linear relationships.

**2. Scatterplots**:

* `age` vs. `fare`: No clear linear trend, but some older passengers paid high fares, likely indicating first-class travelers.
* `sibsp` vs. `fare`: Outliers exist where families with many siblings/spouses paid much more, possibly large families in first class.

**3. Boxplots**:

* Shows fare distributions across different numerical variables, with several outliers in high fare values.
* Passengers with lower `pclass` (1st class) have higher median fares, with greater variability.

**4. Violin Plots**:

* Captures the distribution density of fares across classes and age groups, reinforcing fare-skewed distributions in higher classes.

**MULTIVARIATE ANALYSIS:**

**Pairplot Observations**:

* The scatterplots reveal potential clusters based on pclass, indicating that passenger class influences age and fare distributions.
* fare and pclass likely show an inverse relationship, where higher-class passengers (lower pclass values) tend to have higher fares.
* sibsp (number of siblings/spouses aboard) shows some concentration around lower values, likely indicating most passengers traveled alone or with only one family member.

**Correlation Heatmap (sns.heatmap) Observations**:

* Strong negative correlation between pclass and fare, reinforcing that higher-class passengers paid significantly more.
* Weak correlation between age and fare, meaning older passengers didn't necessarily pay more for their tickets.
* Low correlation between sibsp and fare, suggesting that traveling with family members didn't strongly affect ticket price.

**Boxplots (sns.catplot) Observations**:

* Fare Differences by Sex, Embarked, and Pclass:
* Male and female passengers show differences in fare distributions, with first-class females potentially paying more.
* Embarked location impacts fare, as passengers from some ports (e.g., C vs. S) might have paid more.
* Age vs. Pclass, Grouped by Sex and Sibsp:
* First-class passengers tend to be older on average.
* Males and females have different age distributions across classes.
* Passengers with many siblings/spouses tend to be younger, possibly indicating families traveling together.

**CONCLUSION:**

Overall, the analysis highlights clear socioeconomic stratifications among passengers, reinforcing historical records that higher-class passengers had better access to resources. These insights could be useful for predictive modeling, such as survival analysis, where passenger class and fare may serve as key features.