Titanic EDA - Partial Analysis Summary

Summary of Findings Based on Initial Analysis

1. Dataset Overview:

- Loaded Titanic dataset with .read_csv().

- Displayed top 5 rows using df.head() to inspect structure.

2. Data Structure & Summary:

- Used df.info() to check data types and missing values.

- df.describe() gave statistics like mean, median, and std deviation.

3. Missing Data:

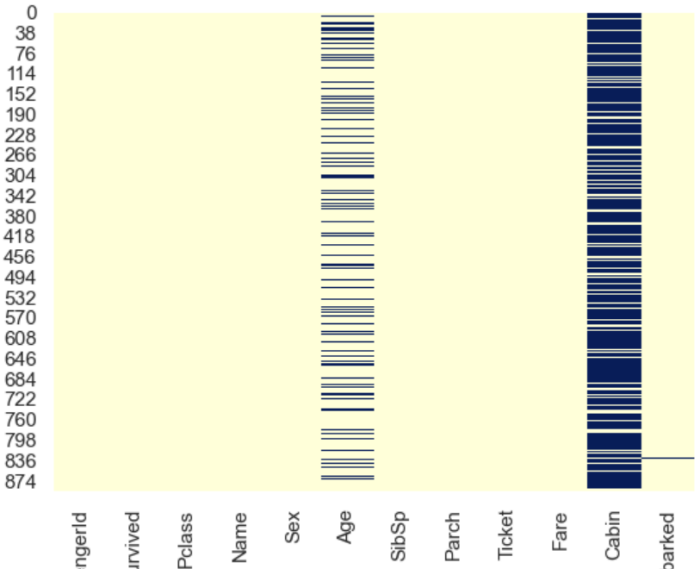
- Used df.isnull().sum() and sns.heatmap() to find missing values.

- Observed that columns like Age and Cabin have missing values.

4. Categorical Analysis:

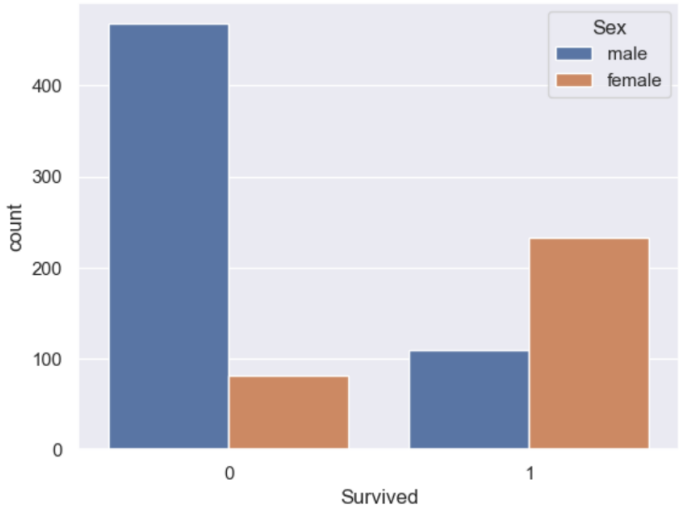
- df['Sex'].value_counts() showed more male passengers than females.

- df['Pclass'].value_counts() revealed most passengers were from class 3.



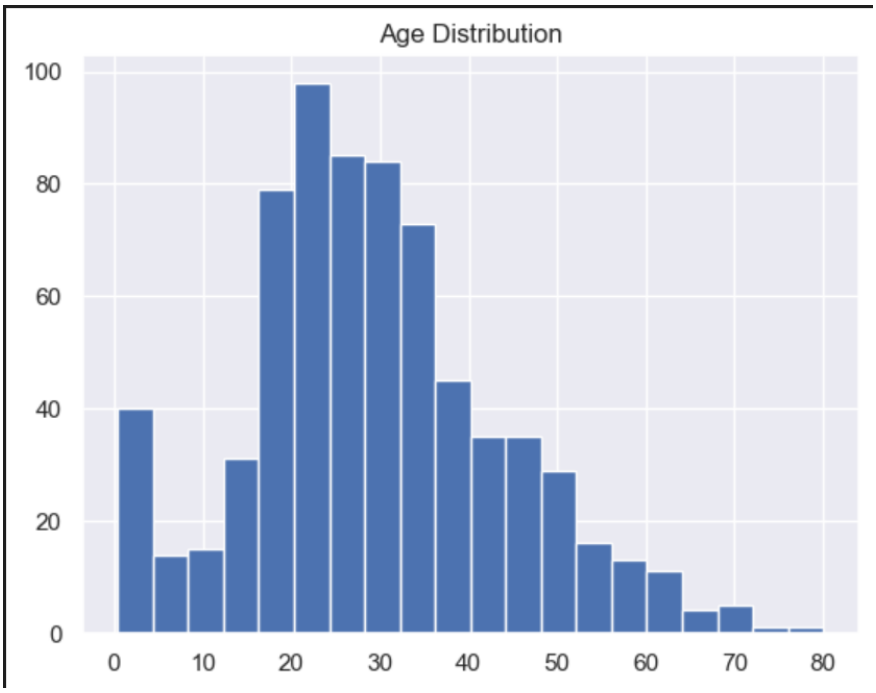
5. Visualization of Survival by Gender:

- sns.countplot() showed females had higher survival compared to males.



6. Age Distribution:

- df['Age'].hist() indicated most passengers were between 20-40 years.



These initial insights provide a base to continue deeper EDA including relationships and trends across other features.

✅ **Interview Questions and Answers on EDA**

1. What is EDA and why is it important?
Exploratory Data Analysis (EDA) is the process of analyzing datasets using statistical summaries and visualizations. It helps in understanding the structure, patterns, relationships, and anomalies in the data before applying any model. It's important for cleaning data, discovering trends, and making informed decisions.

2. Which plots do you use to check correlation?

- sns.heatmap() for visualizing the correlation matrix.
- sns.pairplot() to visually explore relationships between numerical features.

3. How do you handle skewed data?

- Apply transformations like **log**, **square root**, or **Box-Cox** to reduce skewness.
- Use appropriate scaling methods.
- Consider binning or grouping values.

4. How to detect multicollinearity?

- Use the **correlation matrix** to check for highly correlated variables.
- Calculate **VIF (Variance Inflation Factor)**; VIF > 5 or 10 indicates multicollinearity.

5. What are univariate, bivariate, and multivariate analyses?

- **Univariate**: Analysis of one variable (e.g., histograms for Age).
- **Bivariate**: Analysis between two variables (e.g., scatter plot of Age vs Fare).
- **Multivariate**: Analysis of more than two variables (e.g., pairplots, heatmaps).

6. What is the difference between heatmap and pairplot?

- **Heatmap**: Shows correlation between features using color shades (e.g., sns.heatmap).
- **Pairplot**: Shows scatter plots and histograms for multiple feature combinations (e.g., sns.pairplot).

7. How do you summarize your insights?

- Note observations below each chart or visualization.

- Write a clear summary of patterns, outliers, and trends.
- Mention key takeaways such as which features impact the target variable most.