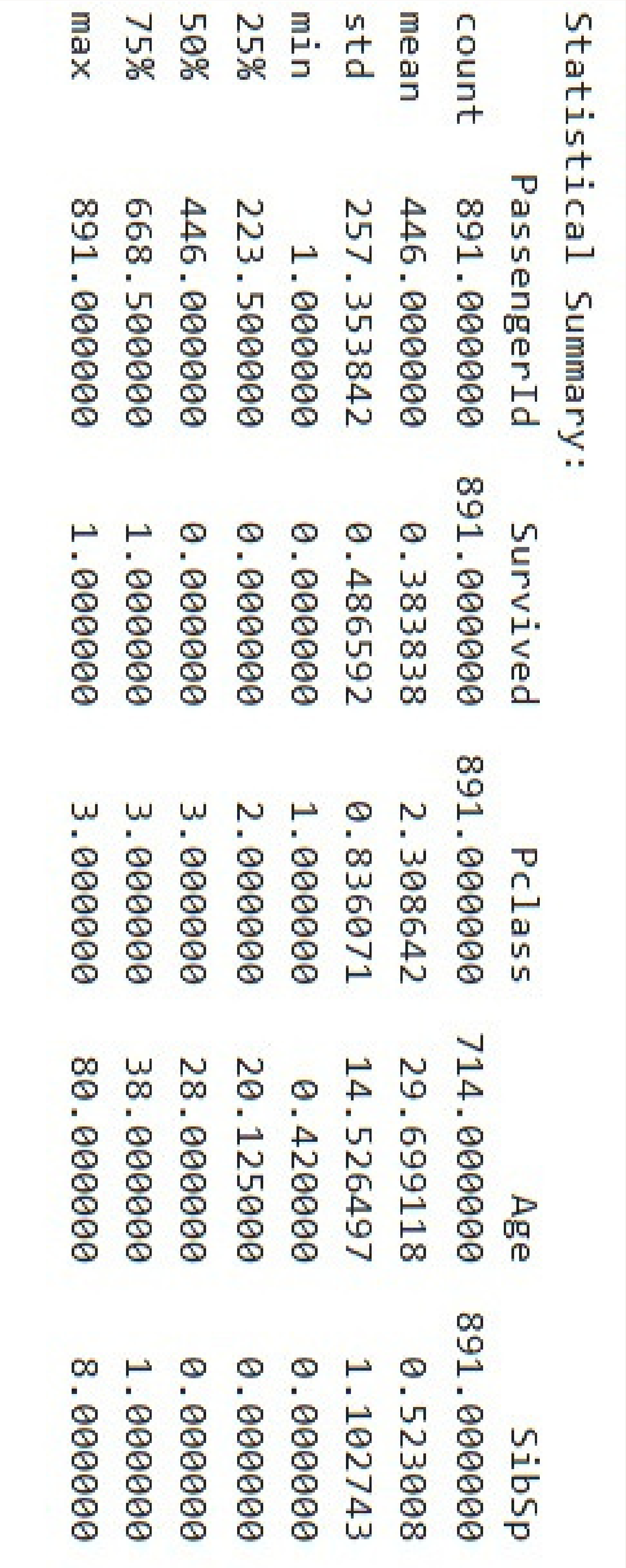
# Exploratory Data Analysis (EDA) on the Titanic Dataset

# 1. Dataset Overview

- **From the notebook's df.info() output:**
- **Class: pandas.core.frame.DataFrame**
- **Range: 0 to 890 (891 entries)**
- **Columns (12 total):**
  - **PassengerId: 891 non-null, int64**
  - **Survived: 891 non-null, int64**
  - **Pclass: 891 non-null, int64**
  - **Name: 891 non-null, object**
  - **Sex: 891 non-null, object**
  - **Age: 714 non-null, float64 (177 missing)**
  - **SibSp: 891 non-null, int64**
  - **Parch: 891 non-null, int64**
  - **Ticket: 891 non-null, object**
  - **Fare: 891 non-null, float64**
  - **Cabin: 204 non-null, object (687 missing)**
  - **Embarked: 889 non-null, object (2 missing)**
  - **Dtypes: float64(2), int64(5), object(5)**
  - **Memory usage: 83.7+ KB**
  - **The dataset has 891 passengers with some missing values in Age, Cabin, and Embarked**

# 2. Statistical Summary



```
Statistical Summary:
       PassengerId    Survived      Pclass         Age       SibSp
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000
```

# 3. Value Counts for Categorical Columns

From the notebook:

- Survived: 0: 549, 1: 342
- Pclass: 3: 491, 1: 216, 2: 184
- Sex: male: 577, female: 314
- Embarked: S: 644, C: 168, Q: 77

Majority are non-survivors, 3rd class, male, and embarked from Southampton.

# Visualizations and Observations

## a. Histogram for Age

- **Code**: sns.histplot(df['Age'], bins=20, kde=True)
- Description: Histogram of passenger ages with 20 bins and a kernel density estimate (KDE) line.
- Observation: The distribution is right-skewed, with a peak around 20-30 years and a tail extending to 80. There are fewer children and elderly passengers. Missing values in Age (177) may affect the plot, but the mean age is ~30, indicating a young adult-heavy demographic. This trend suggests potential survival advantages for younger groups.

## b. Histogram for Fare

Code: sns.histplot(df['Fare'], bins=30, kde=True)

- Description: Histogram of ticket fares with 30 bins and KDE.
- Observation: Highly right-skewed distribution, with most fares clustered below 50 (peak at low fares) and a long tail for high fares (outliers >200). This reflects economic inequality, with many low-class passengers paying minimal fares and a few luxury ones. The skewness (mean 32, max 512) highlights how fare correlates with class and survival.

# c. Boxplot for Age by Survived

Code: sns.boxplot(x='Survived', y='Age', data=df)

- Description: Boxplot comparing age distributions for survived (1) vs. not survived (0).
- Observation: Median age for survivors is slightly lower (~28) than non-survivors (~30), with wider interquartile range for survivors in lower ages. Outliers exist in elderly for both groups. This indicates that children and younger passengers had higher survival rates, likely due to prioritization during evacuation, though missing ages may slightly bias the view.

# d. Boxplot for Fare by Pclass

Code: sns.boxplot(x='Pclass', y='Fare', data=df)

- Description: Boxplot of fares grouped by passenger class (1, 2, 3).
- Observation: 1st class has the highest median fare (~60) with many outliers (up to 512), 2nd class ~14, and 3rd class ~8 with few outliers. The plot shows clear separation, with 1st class having greater variance due to luxury tickets. This trend underscores socioeconomic divisions, where higher class (lower Pclass number) paid more, linking to better survival odds.

# e. Scatterplot for Age vs Fare

Code: sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)

- Description: Scatter plot of age against fare, colored by survival status.
- Observation: Points are scattered with no strong linear trend (correlation ~0.1). Higher fares (>100) show more survivor clusters (hue=1), especially in 20-40 age range, while low fares have more non-survivors across all ages. This reveals that wealth (high fare) improved survival chances, independent of age, with some child survivors in low-fare groups.

# f. Pairplot for Selected Columns

Code: sns.pairplot(df[['Survived', 'Pclass', 'Age', 'Fare', 'SibSp', 'Parch']], hue='Survived')

- Description: Pairwise scatterplots and histograms for the selected columns, with points colored by survival.
- Observation: Diagonal histograms confirm skewed distributions (e.g., Age right-skewed, Fare highly skewed). Off-diagonals show relationships like negative Pclass-Survived trend (higher class = more survivors) and positive Fare-Survived. SibSp/Parch vs Age indicate larger families among younger passengers, with small families (1-3) surviving more. Overall, the pairplot highlights multivariate patterns, such as class and family size influencing survival.

# g. Heatmap for Correlation

Code: sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')

- Description: Annotated heatmap of correlations among numerical columns.
- Observation: Strong negative correlation between Pclass and Survived (-0.34) and Pclass and Fare (-0.55), positive between Fare and Survived (0.26). Age has weak negative correlation with Survived (-0.08) and Pclass (-0.37, older in higher classes). SibSp and Parch positively correlate (0.41), but weakly with survival. This visual emphasizes socioeconomic factors (class/fare) as key drivers of survival over age or family size.

# 5. Summary of Findings

The Titanic dataset EDA reveals a survival rate of approximately 38% (342 survivors out of 891), with significant disparities influenced by socioeconomic status, gender, and age. Higher classes (lower Pclass) and higher fares strongly correlate with better survival, reflecting access to lifeboats. Females (74% survival) and children outperformed males (19%) and adults, consistent with evacuation protocols. Distributions show skewness in Age (young adults dominant) and Fare (majority low-cost with luxury outliers), while family size is optimal at 1-3 members for survival. Anomalies include missing data (Age: 20%, Cabin: 77%) and zero fares (possible errors or crew). Identified trends include class-based inequalities and weak age effects. These insights provide a foundation for further analysis, such as machine learning models for survival prediction. The notebook demonstrates effective use of Pandas and Seaborn for exploration, though missing value handling could be added for robustness.