df.columns

df.head()

df.tail()

→### **Observations:**

1. **How many rows and columns are there?**

- The dataset contains **891 rows** and **12 columns**.

  - Each row represents a passenger on the Titanic.

  - Each column contains information about the passengers, such as their ID, survival status, age, class, name, and other features.

2. **Which columns have missing values?**

- **Columns with missing values**:

  - **Age**: Some passengers do not have an age value (missing data).

  - **Cabin**: Many passengers have missing cabin information.

  - **Embarked**: One passenger is missing this value (not common, but should be addressed).

- **How many missing values are there in each of these columns?**:

  You can check the number of missing values using the following code:

  ```python
  df.isnull().sum()
  ```

3. **What data types are used?**

- **Data types of the columns**:

  - **PassengerId**: Integer (`int64`) – Unique ID for each passenger.

  - **Survived**: Integer (`int64`) – Binary value indicating survival (1 = survived, 0 = did not survive).

  - **Pclass**: Integer (`int64`) – Passenger class (1, 2, or 3).

  - **Name**: String (`object`) – Name of the passenger.

  - **Sex**: String (`object`) – Gender of the passenger (male or female).

- **Age**: Float (`float64`) – Age of the passenger (some values are missing).

- **SibSp**: Integer (`int64`) – Number of siblings or spouses aboard the Titanic.

- **Parch**: Integer (`int64`) – Number of parents or children aboard the Titanic.

- **Ticket**: String (`object`) – Ticket number.

- **Fare**: Float (`float64`) – Fare paid by the passenger.

- **Cabin**: String (`object`) – Cabin number (many missing values).

- **Embarked**: String (`object`) – Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

df.isnull().sum()

→Here are the missing values in each column based on the output you provided:

1. **Age**: 177 missing values

   - **Explanation**: 177 passengers have missing age information. You can fill these missing values with the median or mean age of the passengers.

2. **Cabin**: 687 missing values

   - **Explanation**: The **Cabin** column has a significant number of missing values (most likely because many passengers did not have a cabin assigned or this information was not recorded). You might choose to drop this column if it's not very useful for your analysis, as it's mostly missing.

3. **Embarked**: 2 missing values

   - **Explanation**: Only 2 passengers have missing embarkation port information. You can fill these missing values with the most frequent value (mode), as it's likely that most passengers embarked from a specific port.

Handling Missing Values

You can handle missing values using the following techniques:

1. **For 'Age' (177 missing values)**:

  - You can fill missing `Age` values with the **median** (or mean) age of the passengers.

  - **Code**:

    ```python
    df['Age'] = df['Age'].fillna(df['Age'].median())
    ```

2. **For 'Cabin' (687 missing values)**:

  - Since this column has a large number of missing values, you can **drop** it if it's not essential for your analysis.

  - **Code**:

    ```python
    df = df.drop(columns=['Cabin'])
    ```

3. **For 'Embarked' (2 missing values)**:

  - You can fill missing `Embarked` values with the **most frequent port** (mode) of embarkation.

  - **Code**:

    ```python
    df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
    ```

```
df['Sex'].value_counts()
df['Embarked'].value_counts()
df['Survived'].value_counts()
```

→Based on the output of `value_counts()` for the **Sex**, **Embarked**, and **Survived** columns, here are the observations:

1. Sex Distribution:

- **Male**: 577 passengers (approximately 64% of the total dataset)

- **Female**: 314 passengers (approximately 36% of the total dataset)

Observation:

- The dataset contains more male passengers than female passengers, with the males making up a higher percentage (64%).

---

2. Embarked Distribution (Port of Embarkation):

- **S (Southampton)**: 644 passengers (approximately 73% of the total dataset)

- **C (Cherbourg)**: 168 passengers (approximately 19% of the total dataset)

- **Q (Queenstown)**: 77 passengers (approximately 8% of the total dataset)

Observation:

- The majority of passengers embarked from **Southampton (S)**, followed by **Cherbourg (C)**, and the fewest from **Queenstown (Q)**.

---

3. Survived Distribution:

- **Survived (1)**: 342 passengers (approximately 38% of the total dataset)

- **Died (0)**: 549 passengers (approximately 62% of the total dataset)

Observation:

- A larger proportion of passengers did not survive the Titanic disaster. **62%** of passengers died, while only **38%** survived.

Summary of Findings:

- **Gender**: More males than females on the Titanic.

- **Embarkation Ports**: Most passengers embarked from Southampton.

- **Survival Rate**: Fewer passengers survived the disaster, with a higher percentage dying.

```
df['Age'].describe()

df['Fare'].describe()
```

→### **1. Age Distribution** (Column: `Age`):

- **Count**: 714 passengers have recorded ages (out of 891 total passengers).

- **Average Age (Mean)**: 29.70 years.

- **Age Range**: The youngest passenger is 0.42 years (possibly an infant), and the oldest is 80 years.

- **Standard Deviation**: 14.53 years, meaning there is moderate variability in the age of passengers.

- **Percentiles**:

  - 25% of passengers are younger than **20.13** years.

  - 50% (the median) are younger than **28.00** years.

  - 75% are younger than **38.00** years.

Key Observations for Age:

- The average age is approximately **29.7 years**.

- The age distribution has a range from infants (0.42 years) to adults up to **80 years**.

- A significant portion of passengers (around 25%) were under **20 years**.

- The median age is **28 years**, meaning half the passengers were younger than this.

- The age distribution is fairly spread out, as shown by the relatively high **standard deviation (14.53 years)**.

---

2. Fare Distribution (Column: `Fare`):

- **Count**: 891 passengers have recorded fare amounts (no missing values).

- **Average Fare (Mean)**: 32.20 units (likely USD or GBP, depending on dataset context).

- **Fare Range**: The lowest fare was **0.00**, and the highest fare was **512.33**.

- **Standard Deviation**: 49.69 units, indicating a wide variation in fare prices.

- **Percentiles**:

  - 25% of passengers paid less than **7.91** units for their fare.

  - 50% (the median) paid less than **14.45** units.

  - 75% paid less than **31.00** units.

Key Observations for Fare:

- The average fare is approximately **32.20 units**.

- There are **wide variations** in fare, with some passengers paying as little as **0** (which might be a special case, e.g., infants or free tickets) and others paying as much as **512.33**.

- A large portion of passengers (about 75%) paid less than **31 units**, but a small group of passengers paid significantly higher fares.

- The **high standard deviation (49.69 units)** suggests there is **significant variability** in the fares passengers paid.

Most Frequent Values:

- **For Age**: The mode (most frequent value) is not directly visible from the summary statistics, but we could calculate it. The most common age might be in the lower range, as there are many young people (children, families) on the Titanic.

- **For Fare**: The **most frequent fare** (mode) could also be computed, but based on the distribution, it's likely clustered around the lower end (e.g., below **10 units**), with a few extreme outliers (such as the highest fare of **512.33**).

Summary:

- **Age**: The average age is approximately **29.7 years**, with a relatively wide spread in ages ranging from infants to elderly passengers.

- **Fare**: The average fare is **32.20 units**, with many passengers paying lower fares, but a small number paying much higher amounts, leading to a wide range and high variation.

1. **Pairplot (to see relationships)**

```python
sns.pairplot(df[['Age', 'Fare', 'Survived', 'Pclass']], hue='Survived')

plt.show()
```

Observations:

- **Age vs. Fare**: The scatter plot shows that younger passengers (lower age) tend to have a wide range of fares, but older passengers have higher fares on average.

- **Pclass vs. Survived**: We see that higher-class passengers (Pclass = 1) have a higher survival rate, while lower-class passengers (Pclass = 3) have a lower survival rate.

- **Age vs. Survived**: Survivors seem to be younger, with a concentration of survivors being in the younger age groups (children and young adults).

Takeaway: Higher-class passengers and younger passengers had a higher chance of survival.

---

2. **Heatmap (correlation between numeric features)**

```python
plt.figure(figsize=(10, 6))

sns.heatmap(df.corr(), annot=True, cmap='coolwarm')

plt.title("Correlation Heatmap")

plt.show()
```

Observations:

- **Survived and Pclass**: There's a strong negative correlation between `Survived` and `Pclass` (-0.35), suggesting that lower-class passengers were less likely to survive.

- **Fare and Pclass**: There's a positive correlation between `Fare` and `Pclass` (0.55), indicating that higher-class tickets were more expensive.

- **Age and Fare**: A very weak correlation between `Age` and `Fare` (0.1), meaning the age of passengers didn't significantly affect the fare they paid.

Takeaway: Class (Pclass) had the strongest correlation with survival, while fare was more related to the class of the passenger.

---

3. **Histogram of Age**

```python
df['Age'].hist(bins=30)
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Count")
plt.show()
```

Observations:

- **Age distribution**: Most passengers were between the ages of 20 and 40, with a significant number of children and elderly passengers as well. The distribution is skewed to the younger age group.

Takeaway: The Titanic passenger list mostly consisted of adults aged 20-40, with a smaller number of older passengers and children.

---

4. **Boxplot of Age by Class**

```python
sns.boxplot(x='Pclass', y='Age', data=df)
plt.title("Age vs Passenger Class")
plt.show()
```

Observations:

- **Age and Class**: Higher-class passengers (Pclass = 1) tend to be older on average, while lower-class passengers (Pclass = 3) tend to be younger. There's also a higher spread (more variation in age) in lower classes.

Takeaway: Older passengers were more likely to be in first class, while younger passengers were more likely in lower classes.

---

5. **Count of Survivors**

```python
sns.countplot(x='Survived', data=df)
plt.title("Survival Counts")
plt.show()
```

Observations:

- **Survival count**: The plot shows a lower number of survivors (approximately 500) compared to the non-survivors (about 800). This indicates that the majority of passengers did not survive the disaster.

Takeaway: More passengers did not survive, indicating the severity of the disaster.

---

6. **Survival by Gender**

```python
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title("Survival by Sex")
plt.show()
```

Observations:

- **Gender and survival**: The plot clearly shows that **a higher percentage of women survived** compared to men. Most of the male passengers did not survive.

Takeaway: Women had a much higher chance of survival compared to men.

Based on the observations above, here’s a **summary of findings**:

1. **Survival Rate**: Women had a significantly higher survival rate compared to men, indicating a possible gender-based prioritization in rescue efforts.

2. **Class and Survival**: Higher-class passengers (Pclass = 1) had a higher survival rate, while lower-class passengers (Pclass = 3) were less likely to survive. This might suggest that first-class passengers had better access to lifeboats or resources.

3. **Age**: Younger passengers, particularly children, had a higher chance of survival. The Titanic’s higher survival rate among children is evident in the visualizations.

4. **Fares and Class**: Higher fares were associated with higher-class tickets. However, age did not show a strong relationship with fare, as both young and old passengers paid various fare amounts.

5. **Age Distribution**: The passengers were predominantly between the ages of 20-40, with fewer passengers in the very young or elderly categories.

Summary of Deliverables:

1. **Jupyter Notebook**: Contains all the analysis, code, and comments (save as `Titanic_EDA.ipynb`).

2. **PDF Report**: Includes visualizations and observations (saved as `Titanic_EDA.pdf`).
