Titanic Dataset Analysis Report

Date of Analysis: June 2, 2025

1. Introduction

This report details the exploratory data analysis (EDA) performed on the Titanic dataset. The primary objective was to extract meaningful insights from the passenger data using the pandas library and to visualize these findings for analytical purposes using Matplotlib and Seaborn. The analysis covers the basic structure of the data, a preview of its contents, identification of missing values, summary statistics, value counts for categorical variables, and a correlation analysis. Furthermore, various plots were generated to visually represent key aspects of the data, such as survival rates, gender distribution, age demographics, and relationships between different passenger attributes.

2. Data Overview and Initial Insights (Pandas)

The initial phase of the analysis involved loading the dataset and gaining a fundamental understanding of its structure and content using pandas.

2.1. Basic Structure

The dataset comprises **891 rows (passengers)** and **12 columns (features)**. The columns in the dataset are: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked.

The data types for each column are as follows:

- PassengerId: int64
- Survived: int64
- Pclass: int64
- Name: object
- Sex: object
- Age: float64
- SibSp: int64
- Parch: int64
- Ticket: object
- Fare: float64
- Cabin: object
- Embarked: object

2.2. Data Preview

- A preview of the first five rows was examined to get a glimpse of the data entries:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

2.3. Missing Values

The analysis identified the following number of missing values in these columns:

- Age: 177 missing values
- Cabin: 687 missing values
- Embarked: 2 missing values

Other columns (PassengerId, Survived, Pclass, Name, Sex, SibSp, Parch, Ticket, Fare) had no missing values. The high number of missing values in 'Cabin' suggests it might be challenging to use this feature directly without imputation or dropping it. 'Age' also has a significant number of missing values that would need addressing.

2.4 Exploratory Data Analysis (EDA) - Pandas Insights

Further exploration using pandas provided deeper statistical insights.

- **Summary Statistics**:
    - Descriptive statistics (count, mean, std, min, 25th/50th/75th percentiles, max) were generated for numerical columns (PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare).
    - For categorical columns (Name, Sex, Ticket, Cabin, Embarked), the summary included count, unique values, top (most frequent), and freq (frequency of top). For example, 'male' was the most frequent gender (577 out of 891).
- **Value Counts for Categorical Variables**:
    - Detailed value counts were examined for each object-type column:
        - Name: 891 unique names, indicating one entry per passenger.
        - Sex: male (577), female (314).
        - Ticket: 681 unique ticket numbers, with some tickets shared (e.g., '347082' appeared 7 times).
        - Cabin: 147 unique cabin numbers, with 'G6', 'C23 C25 C27', and 'B96 B98' appearing most frequently (4 times each). The high number of missing values is reiterated here.
        - Embarked: S (Southampton - 644), C (Cherbourg - 168), Q (Queenstown - 77).
- **Correlation Matrix (Numerical Only)**:
    - A correlation matrix was computed for all numerical columns. Key observations include:
        - Survived has a notable negative correlation with Pclass (-0.34), suggesting passengers in higher classes (lower Pclass number) had a better chance of survival.
        - Survived has a positive correlation with Fare (0.26).
        - Age has a negative correlation with Pclass (-0.37) and SibSp (-0.31).
        - Fare is negatively correlated with Pclass (-0.55), meaning higher fares are associated with higher classes.

3. Data Visualization Analysis (Matplotlib and Seaborn)

Visualizations were created using matplotlib and seaborn to better understand distributions and relationships.

3.1 Survival Count (Matplotlib Bar Chart):

- **Plot Type**: Bar Chart
- **Description**: Showed the count of passengers who survived (1) versus those who did not (0).
- **Insight**: Visually confirmed that more passengers did not survive than survived.

3.2 Gender Distribution (Matplotlib Pie Chart):

- **Plot Type**: Pie Chart

- **Description**: Illustrated the proportion of male and female passengers.
- **Insight**: Clearly showed that males constituted a larger portion of the passengers (64.8%) compared to females (35.2%).

3.3 Age Distribution (Matplotlib Histogram):

- **Plot Type**: Histogram
- **Description**: Displayed the frequency distribution of passenger ages.
- **Insight**: The age distribution is skewed towards younger adults, with a peak in the 20-30 age range. There are also a number of infants and young children.

3.4 Survival by Passenger Class (Matplotlib Stacked Bar Chart):

- **Plot Type**: Stacked Bar Chart
- **Description**: Showed the count of survived and not survived passengers for each passenger class.
- **Insight**: Indicated that survival rates were higher for passengers in 1st class and significantly lower for those in 3rd class.

3.5 Age vs Fare Colored by Survival (Matplotlib Scatter Plot):

- **Plot Type**: Scatter Plot
- **Description**: Plotted passenger age against fare, with points colored by survival status.
- **Insight**: This plot helps visualize if combinations of age and fare had any bearing on survival. Generally, it showed a wide spread, but higher fares (often associated with 1st class) had more survivors. Very high fares were predominantly survivors.

3.6 Age Distribution by Survival (Matplotlib Box Plot):

- **Plot Type**: Box Plot
- **Description**: Compared the age distributions of passengers who survived versus those who did not.
- **Insight**: Survivors tended to be slightly younger on average, and children had a higher survival rate. The median age for non-survivors was higher than for survivors.

3.7 Passengers by Embarkation Port (Matplotlib Horizontal Bar Chart):

- **Plot Type**: Horizontal Bar Chart
- **Description**: Showed the number of passengers who embarked from each port (S, C, Q).
- **Insight**: The vast majority of passengers embarked from port 'S' (Southampton).

3.8 Survival Count (Seaborn Count Plot / Bar Chart):

- **Plot Type**: Count Plot (similar to a bar chart)
- **Description**: Visualized the number of survived vs. not-survived passengers using seaborn's countplot.
- **Insight**: Reconfirmed the findings from the matplotlib bar chart regarding survival counts.

3.9 Age Distribution with KDE (Seaborn Histogram):

- **Plot Type**: Histogram with Kernel Density Estimate (KDE)
- **Description**: Showed the age distribution with a smooth KDE curve overlaid.
- **Insight**: Provided a smoother representation of the age distribution, confirming the peak in young adulthood.

3.10 Survival by Passenger Class (Seaborn Count Plot / Bar Chart with Hue):

- **Plot Type**: Count Plot with Hue

- **Description**: Displayed survival counts (0 or 1) grouped by passenger class.
- **Insight**: Reemphasized that 1st class passengers had a higher survival count (and proportion) compared to 2nd and especially 3rd class.

3.11 Age Distribution by Survival (Seaborn Box Plot):

- **Plot Type**: Box Plot
- **Description**: Compared age distributions for survived and not-survived passengers using seaborn.
- **Insight**: Similar to the matplotlib box plot, it showed that children had a better survival chance and the median age of survivors was lower.

3.12 Fare Distribution by Class and Survival (Seaborn Violin Plot):

- **Plot Type**: Violin Plot
- **Description**: Showed the distribution of fares for each passenger class, split by survival status.
- **Insight**: This plot combines aspects of a box plot and a KDE. It highlighted that:
  - 1st class fares were generally much higher and more spread out. Survivors in 1st class often paid higher fares than non-survivors.
  - 2nd and 3rd class fares were lower and less variable. In 3rd class, the fare distribution for survivors and non-survivors was quite similar, though slightly higher for some survivors.

3.13 Correlation Heatmap (Seaborn Heatmap):

- **Plot Type**: Heatmap
- **Description**: Visualized the correlation matrix of numerical features with annotations.
- **Insight**: Provided a clear, color-coded visual of the correlations identified earlier, making it easier to spot strong positive or negative relationships (e.g., Pclass and Fare, Pclass and Survived).

3. Conclusion:

The exploratory data analysis performed in Task5.ipynb has yielded valuable initial insights into the passenger dataset. Key takeaways include the identification of significant missing data in 'Age' and 'Cabin', the demographic makeup of passengers (more young adult males, majority embarked from Southampton), and strong correlations between survival and features like passenger class and fare. Visualizations effectively highlighted these patterns, such as the better survival odds for first-class passengers and females, and the higher survival rate among children.