TITANIC EXPLORATORY DATA ANALYSIS REPORT

By: Blessing Ilesanmi (WITS 24)

7th April, 2025

INTRODUCTION

The Titanic dataset is a widely recognized historical dataset that contains detailed information about the passengers who were aboard the RMS Titanic during its ill-fated maiden voyage in April 1912. The ship, which was considered unsinkable at the time, tragically struck an iceberg and sank in the North Atlantic Ocean, resulting in the deaths of over 1,500 passengers and crew members. The dataset includes various features such as passenger age, sex, ticket class, fare paid, port of embarkation, and survival status.

The goal of this analysis is to explore and understand the factors that influenced survival during the disaster. By performing exploratory data analysis (EDA), this report aims to uncover relationships and trends within the dataset, particularly focusing on how attributes like age, gender, and passenger class affected the likelihood of survival. Through statistical summaries and visual representations, this report provides insights into the underlying patterns of the Titanic tragedy, offering a data-driven perspective on one of the most well-known maritime disasters in history.

DATA LOADING AND CLEANING

The dataset used for this analysis was obtained from Kaggle and consists of three CSV files: *train.csv*, *test.csv*, and *gender_submission*.csv. For the purpose of this exploratory analysis, only the *train.csv* file was used, as it contains both the feature variables and the target variable (Survived). Upon loading the data into a Pandas DataFrame, the shape of the dataset was examined. The training set contains 891 row*s* and 12 columns, representing individual passengers and their respective features, such as age, sex, ticket class, fare, and survival status.

Data Loading

In [2]:
```
# Load the data
df = pd.read_csv('data/raw/train.csv')
test_df = pd.read_csv('data/raw/test.csv')
gender_submission_df = pd.read_csv('data/raw/gender_submission.csv')
```

In [3]:
```
df.head()
```

Out[3]:

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

In [4]:
```
df.shape
```

Out[4]: `(891, 12)`

Data cleaning inspection revealed that some columns contained missing values. The Age column had a moderate number of missing entries (177), while the Cabin column had a substantial proportion of missing values (687). The Embarked column was missing only two values.

Data Cleaning

```
In [5]: # Check missing values in each column
        df.isnull().sum()
```

```
Out[5]: PassengerId      0
        Survived         0
        Pclass           0
        Name             0
        Sex              0
        Age            177
        SibSp            0
        Parch            0
        Ticket           0
        Fare             0
        Cabin          687
        Embarked         2
        dtype: int64
```

To address this:

- Missing values in the Age column were filled using the median age to minimize the influence of outliers.
- The Cabin column was dropped due to the high volume of missing data and its limited utility for this analysis.
- Missing values in the Embarked column were imputed with the most frequent port of embarkation (mode).

```
In [6]: # Fill missing 'Age' with median
        df['Age'].fillna(df['Age'].median(), inplace = True)
```

```
In [7]: # Drop 'Cabin' due to too many missing values
        df.drop(columns = ['Cabin'], inplace = True)
```

```
In [8]: # Fill missing 'Embarked' with the mode (most frequent value)
        df['Embarked'].fillna(df['Embarked'].mode()[0], inplace = True)
```

The dataset was also checked for duplicate entries, and none were found.

```
In [9]: # Check and drop duplicates
        df.duplicated().sum()
```

```
Out[9]: 0
```

After these steps, the dataset was clean, with no missing or duplicate values, and was ready for further exploration and analysis.

```
In [10]: df.isnull().sum()
```

```
Out[10]: PassengerId    0
         Survived       0
         Pclass         0
         Name           0
         Sex            0
         Age            0
         SibSp          0
         Parch          0
         Ticket         0
         Fare           0
         Embarked       0
         dtype: int64
```

SUMMARY STATISTICS

The dataset was first analyzed to compute basic summary statistics, which helped in understanding the distribution of values across different columns. The key statistics for numerical columns are as follows:

- **Age**: The average age of passengers was approximately 29.36 years, with a minimum of 0.42 years (an infant) and a maximum of 80 years.
- **Fare**: The average fare paid was about 32.20, but the fare ranged widely, with some passengers paying as much as 512.33, indicating some high-class passengers.
- **Pclass**: The majority of passengers (55.11%) were in the third class, followed by first class (24.24%) and second class (20.65%).

Summary Statistics

In [11]:
```
# Summary statistics for numerical columns
df.describe()
```

Out[11]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.361582	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	13.019697	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	22.000000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	35.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [12]:
```
# Distribution of passengers across Pclass
pclass_distribution = df['Pclass'].value_counts(normalize=True) * 100
pclass_distribution
```

Out[12]:
```
Pclass
3    55.106622
1    24.242424
2    20.650954
Name: proportion, dtype: float64
```

Group Comparisons

Group comparisons revealed some notable insights:

- Survival Rate by Passenger Class: The survival rate for passengers in first class was significantly higher (approximately 62%) compared to those in second (47%) and third class (24%).
- Survival Rate by Sex: Female passengers had a significantly higher survival rate (about 74%) compared to male passengers (approximately 19%).
- Survival Rate by Age: Survival rates were higher among children (0-18 years), with a survival rate of 50%, compared to adults in other age groups.

Group Comparison

In [13]:
```
# Survival rate by Pclass
survival_by_pclass = df.groupby('Pclass')['Survived'].mean()
survival_by_pclass
```

Out[13]:
```
Pclass
1    0.629630
2    0.472826
3    0.242363
Name: Survived, dtype: float64
```

In [14]:
```
# Survival rate by Sex
survival_by_sex = df.groupby('Sex')['Survived'].mean()
survival_by_sex
```

Out[14]:
```
Sex
female    0.742038
male      0.188908
Name: Survived, dtype: float64
```

In [15]:
```
# Create age bins
bins = [0, 18, 35, 60, 100]
labels = ['0-18', '19-35', '36-60', '60+']
df['Age_group'] = pd.cut(df['Age'], bins=bins, labels=labels)

# Survival rate by Age group
survival_by_age = df.groupby('Age_group')['Survived'].mean()
survival_by_age
```

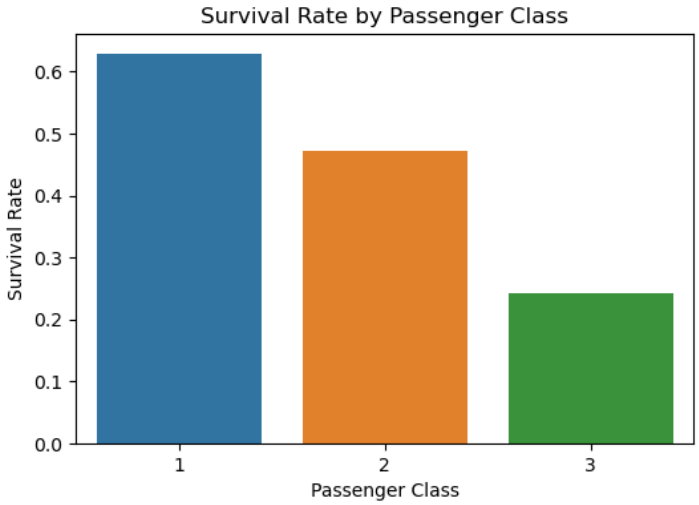
Out[15]:
```
Age_group
0-18     0.503597
19-35    0.353271
36-60    0.400000
60+      0.227273
Name: Survived, dtype: float64
```

DATA VISUALIZATION

To explore the patterns and trends in the dataset, several visualizations were created. These visualizations help in understanding how different factors, such as passenger class, sex, and age, influenced the survival rates of passengers aboard the Titanic.

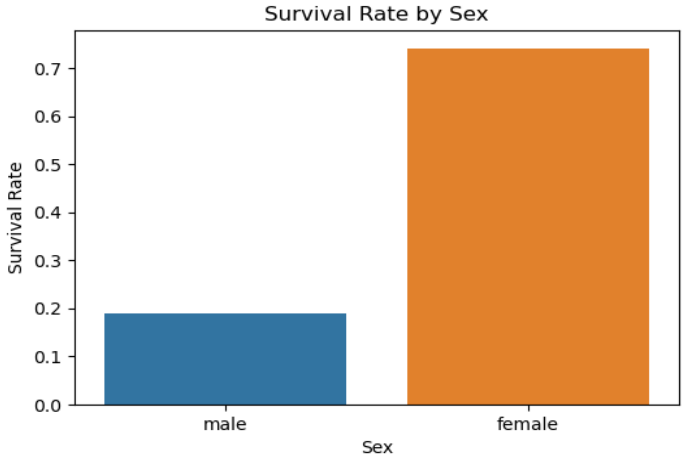
1. Survival Rate by Passenger Class

The bar plot below shows the survival rate across the different passenger classes. The survival rate was highest in first class, followed by second class, and lowest in third class. This indicates that passengers in higher classes had a better chance of survival.



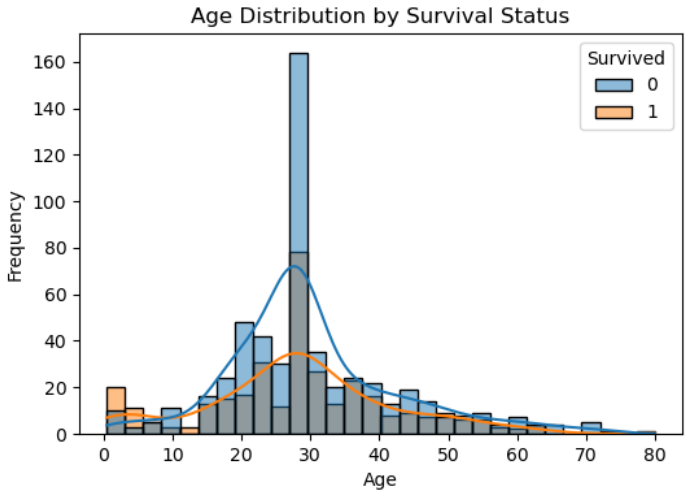
2. Survival Rate by Sex

The next plot visualizes the survival rate by sex. Female passengers had a significantly higher survival rate compared to male passengers, with nearly 75% of females surviving, while only about 20% of males survived. This stark difference highlights the gender-based survival disparity.



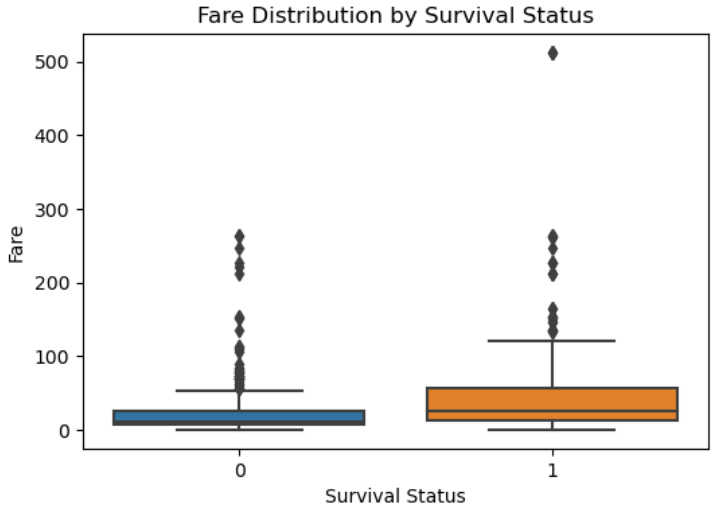
3. Age Distribution by Survival Status

This histogram shows the distribution of ages for survivors and non-survivors. We observe that children (those under the age of 18) had a higher survival rate compared to adults. The distribution for non-survivors appears to be concentrated among older passengers, indicating that age may have played a role in survival chances.



4. Fare Distribution by Survival Status

The box plot below shows the distribution of fares paid by survivors and non-survivors. It suggests that survivors tended to pay higher fares, which could indicate that wealthier passengers, likely in higher classes, had a higher chance of survival.



CONCLUSION

This analysis set out to explore the Titanic dataset and uncover which factors most significantly influenced passenger survival. Through a series of statistical summaries and visualizations, it became clear that passenger class, gender, and age were strongly associated with survival outcomes. Passengers in first class were more likely to survive compared to those in second or third class, with third-class passengers comprising the majority of the dataset (approximately 55%) and experiencing the highest fatality rate. Gender also played a major role: females had a substantially higher chance of survival than males, reflecting the historical "women and children first" policy that shaped evacuation priorities. Age distribution showed that children and younger passengers generally had a better chance of survival, although this varied depending on their class and gender.

Limitations

- The dataset does not include detailed information about why some individuals survived while others didn’t, factors like location on the ship, proximity to lifeboats, or crew assistance were not available.
- Cabin data was too sparse to include in the analysis and might have revealed more about a passenger’s location on the ship.

Overall, the analysis confirms that social and economic status, age, and gender were significant indicators of survival, reflecting the historical accounts of the Titanic tragedy.