Exploratory Data Analysis of the Titanic Dataset

The report represents an exploratory Data Analysis of the Titanic Dataset. The intention of this report is to understand the structure of the dataset, identify the missing data, study the correlation between major variables and examine how different passenger relate to survival. This titanic dataset is widely used in machine learning lessons by utilizing various tools such as python, R, spreadsheets, SQL and many other tools.

In this report, Excel spreadsheet is used for studying the generic overview of the dataset and several visualizations including bar charts, donut charts, heatmaps as well as data processing and data cleaning were mainly conducted by using Python.

1. Data Overview

Just like many other datasets, Titanic Dataset also contains a mixture of numerical and categorical variables as well as many missing data under especially in *"Age"* and "*Cabin*" where only 2 data was missing in "*Embarked*". The dataset contains a mix of **numerical columns** (e.g. *Age, Fare, SibSp, Parch)* and **categorical columns** (e.g. *Sex, Pclass, Embarked, Name, Ticket, Cabin)*.

- ***Identify the total number of rows and columns.***

  The dataset got **891 rows and 14 columns**

```
RangeIndex: 891 entries, 0 to 890
Data columns (total 14 columns):
```

- ***List all the features and their data types (numerical, categorical, boolean).***

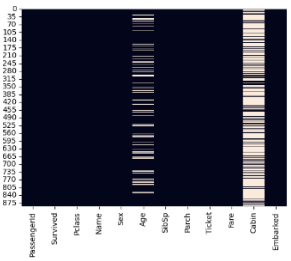
```
data_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     891 non-null    object
 12  AgeGroup     891 non-null    object
 13  FamilySize   891 non-null    int64
dtypes: float64(2), int64(6), object(6)
memory usage: 97.6+ KB
```

- ***Highlight any columns with missing values.***

```
data_df.isnull().sum()

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
```




```
data_missing = (data_df.isnull().mean()*100).round(2)
data_missing

PassengerId     0.00
Survived        0.00
Pclass          0.00
Name            0.00
Sex             0.00
Age            19.87
SibSp           0.00
Parch           0.00
Ticket          0.00
Fare            0.00
Cabin          77.10
Embarked        0.22
dtype: float64
```

2. Summary Statistic

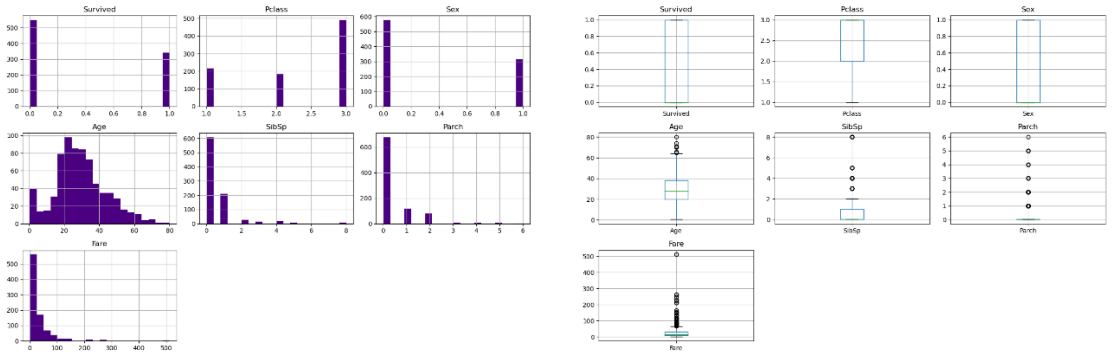
- ***Numerical Features***

  Except for PassengerId, the rest of numerical data columns are organized into "features" and dataframe with new "features" is used with *.describe()* to pull out statistics. Boxplot and histograms are used to detect outliers and skewness.

```
In [532]: features = ['Survived', 'Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare']
In [533]: data_df[features].describe()
```

Out[533]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	0.352413	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	0.477990	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	0.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	0.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	1.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	1.000000	80.000000	8.000000	6.000000	512.329200



Age Outliers:

Based on the IQR method and the Age boxplot, several passengers aged 65–80 appear as outliers. These values are realistic and represent older passengers rather than data errors. The histogram also shows a right tail with fewer elderly passengers, confirming that older ages are rare but valid.

Fare Outliers:

Fare shows strong right skewness, which is clear in the histogram. Most passengers paid low fares, while a small group paid extremely high fares (mostly first-class luxury tickets). These high values are genuine and expected due to differences in ticket class.

- ***Categorical Features***

1. Sex :

Sex	Count	Imbalance:
Male	577	There are significantly more male passengers than female passengers.
Female	314	

2. Passenger Class (Pclass)

Class	Count	Remarks:
1st	216	Third-class passengers make up more than half of the dataset.
2nd	184	
3rd	491	

3. Embarked

Port	Count	Risk	Remarks:
S(Southampton)	646	Majority	Maximum passengers Embarkation port is Southampton.
C(Cherbourg)	168	Moderate	
Q(Queenstown)	77	Lowest	

3. Missing Values

To understand the data quality, Firstly, checked for missing values using data_df.isnull().sum() and plotted a missing-value heatmap. The dataset contains:

- A large number of missing values in Age
- A small number of missing values in Embarked
- Many missing values in Cabin (expected in Titanic dataset)

- Handling Strategy

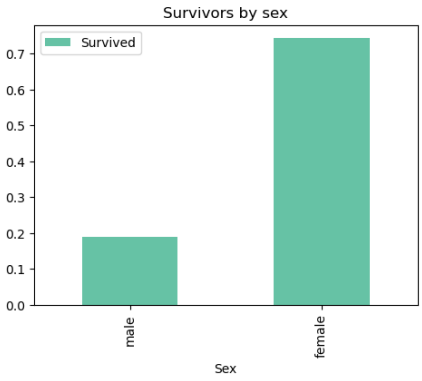
  Since the assignment focuses on exploratory analysis rather than modeling, I did not remove any rows. Instead, I handled missing values in the following way:

  - Embarked:
    Filled missing Embarked values using the most common category ("S" for Southampton), because it is the standard and simplest imputation method for categorical data:
    data_df['Embarked'] = data_df['Embarked'].fillna('S')
  - Age:
    Did not impute or drop missing Age values.
    Instead, allowed plotting functions to automatically ignore NaN values.
    For a full ML pipeline, the recommended approach would be median imputation by Pclass and Sex, but this was not required for this analysis.
  - Cabin:
    Cabin has too many missing values to impute meaningfully.
    Since Cabin was not used during analysis hence simply left it unchanged.

4. Feature Relationships Vs Survival

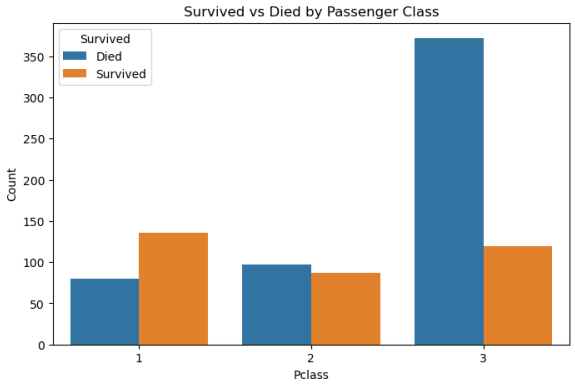
- **Sex and Survival**

The plots show that female passengers had a much higher survival rate than males. This confirms the "women and children first" evacuation practice and matches the strong correlation found earlier.



- **Passenger Class and Survival**

  First-class passengers have the highest survival rate, followed by second class, while third class has the lowest. This reflects differences in cabin location and access to lifeboats.

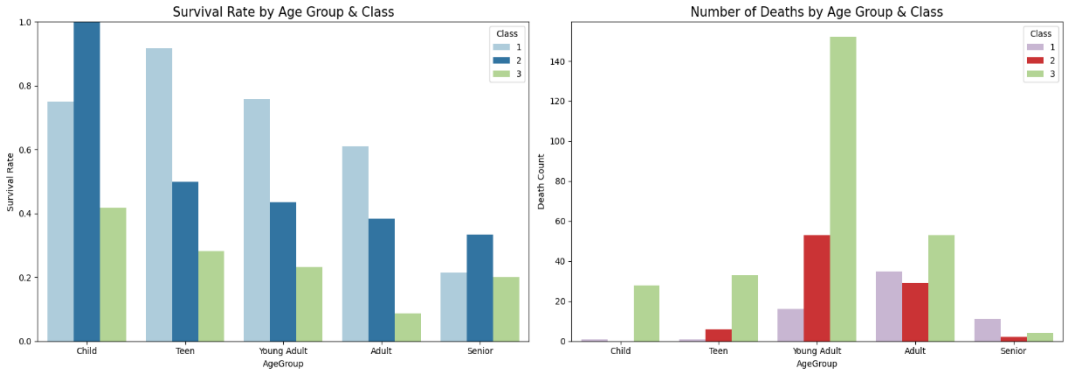


- **Age and Survival**

  In order to have clear vision, Age group are created to make the pattern clearer and followed by plot to visualise the survival rate. Since age pattern alone is could provide the surface level of analysis, further analysis was performed by using Pclass's age group survival and death rate. Subplot is used here to have better visionary.

  First-class passengers have the highest survival rate, followed by second class, while third class has the lowest. This reflects differences in cabin location and access to lifeboats.

AgeGroup	Age
Child	57.971014
Teen	47.727273
Adult	41.776316
Young Adult	35.055351
Unknown	29.378531
Senior	26.923077



From these charts:

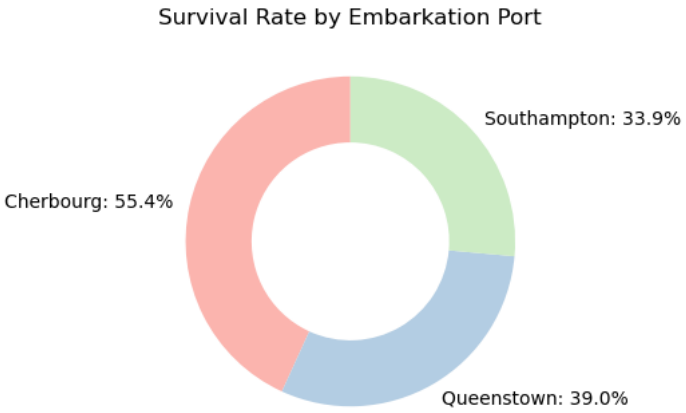
- **Children**, especially in 1st and 2nd class, have clearly higher survival rates.
- **Adults and Young Adults** in 3rd class represent many of the deaths.
- **Seniors** have low survival in all classes.

This shows that age interacts with class: it is not only about being young, but also about where that specific person was on the ship.

- **Embarked and Survival**

  I visualised this with a ring (donut) pie chart. Passengers embarking at Cherbourg (C) have the highest survival rate, while Southampton (S) has the lowest. This again connects to class distribution, since more wealthy passengers boarded at Cherbourg.

Embarkation Port	Survival Rate
Cherbourg	55.357143
Queenstown	38.961039
Southampton	33.900929

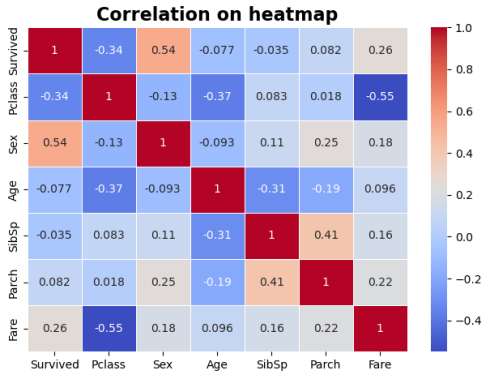


- **Suggest which features appear most influential in survival outcomes.**

  Across all classes, females and children survive more than adult males, but the gap is especially clear in 1st and 2nd class. Third-class males have very low survival.

5. Correlation Analysis

This heatmap is produced using by only numerical features however, passengerId is not necessary hence, it was dropped.



- Survived has the strongest positive correlation with Sex, meaning females are more likely to survive.
- Pclass has a negative correlation with Survived (higher class → higher survival).
- Fare is also weakly related to Survived, but this is partly because Fare is linked to Pclass.
- Other numerical features (SibSp, Parch, Age) have much weaker correlations.

6. Observations and Insights

Survival Differences by Gender

Female passengers had much higher survival rates than males. This was the strongest pattern in the dataset and reflects the "women and children first" evacuation rule. The correlation between Sex and Survived also supports this finding.

Impact of Passenger Class (Pclass)

Passenger class strongly influenced survival. First-class passengers survived at the highest rates, followed by second class, while third class had the lowest survival. Differences in cabin locations and access to lifeboats likely explain this pattern.

Effects of Age on Survival

Age alone had only a weak correlation with survival. However, when grouped, children—especially in higher classes—survived more often, while seniors had lower survival rates. Age outliers above 65 were kept because they represent real older passengers.

Influence of Fare

Higher fares are associated with slightly higher survival chances. Because fare is closely tied to class, this suggests that wealthier passengers had better evacuation conditions.

Trends in Embarkation Ports

Passengers boarding at Cherbourg (C) showed a higher survival rate than those from Southampton (S) or Queenstown (Q). This pattern is mainly driven by more first-class passengers embarking at Cherbourg.

Missing Data and Outliers

The dataset contains missing values in Age, Embarked, and many in Cabin. Embarked was filled with the most common value. Age and Cabin were left unchanged for EDA. Upper-end Age outliers (65–80) were detected, but these represent genuine elderly passengers.

Insights from Family Size (Additional Observation)

Exploring family size provided an extra perspective. Smaller family groups tended to have higher survival rates, while individuals traveling alone and very large families had lower survival. This suggests that being accompanied by a small group may have supported evacuation, while large family groups faced more difficulty staying together during the emergency.

FamilySize	PassengerCount	SurvivalRate(%)
4	29	72.413793
3	102	57.843137
2	161	55.279503
7	12	33.333333
1	537	30.353818
5	15	20.000000
6	22	13.636364
8	6	0.000000
11	7	0.000000