EDA Report on Titanic Dataset

Introduction

This report presents the results of an Exploratory Data Analysis (EDA) conducted on the Titanic dataset, aiming to identify key patterns, relationships, and potential outliers within the data.

Data Overview

Dataset Source: Kaggle

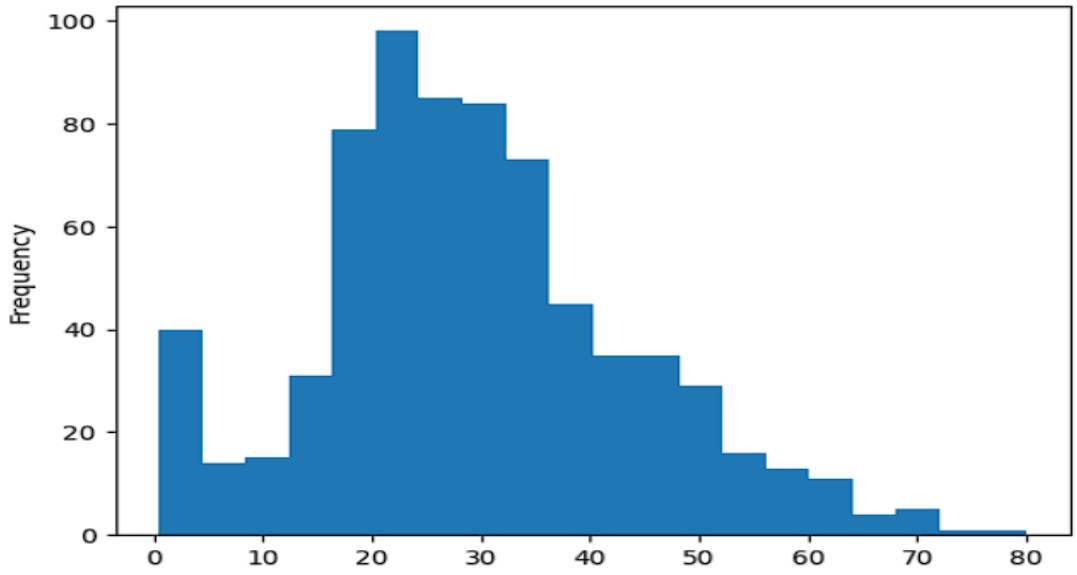
Types of Columns:

- Numerical – Age, Fare, PassengerId
- Categorical – Survived, Pclass, Sex, SibSp, Parch, Embarked
- Mixed – Name, Ticket, Cabin

Univariate Analysis of the numerical columns

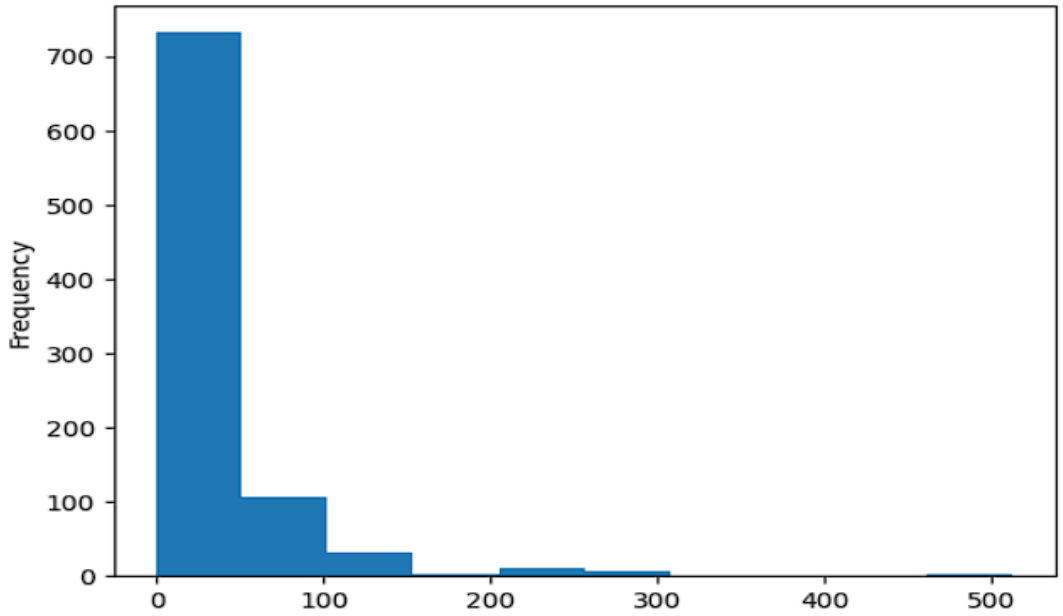
Age Column

- Age is normally(almost) distributed.
- 20% of the values are missing.
- Mean age is 30 years, Max age is 80 years.



Fare Column

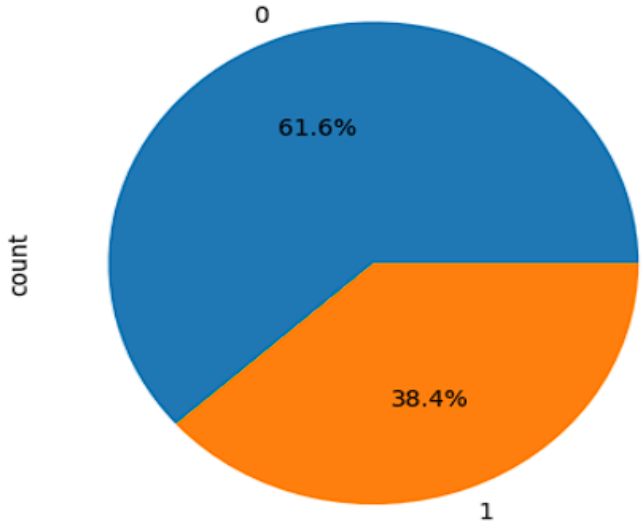
- The data is highly skewed.
- Fare column contains the group fare and not the individual fare.
- We need to create a new column called individual fare.



Univariate Analysis of the categorical columns

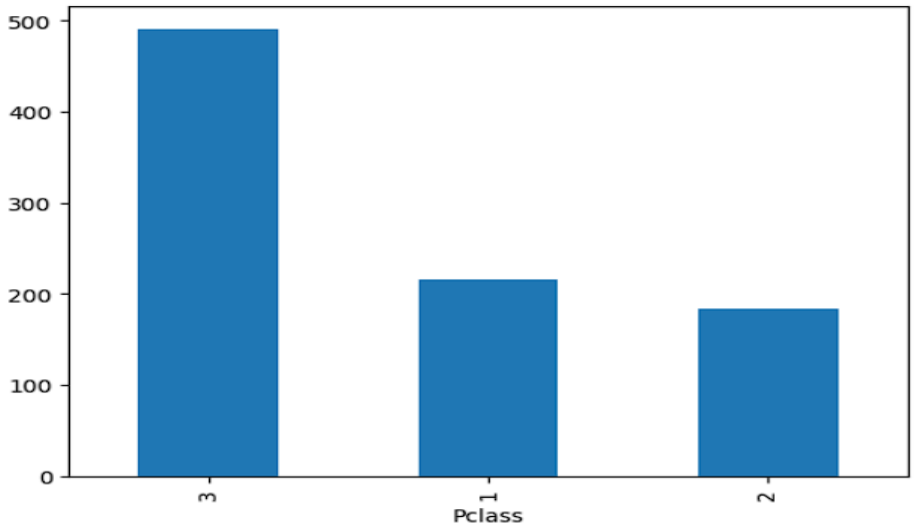
Survived Column

- The accident was very deadly, more than 60% people could not survive.
- No null values



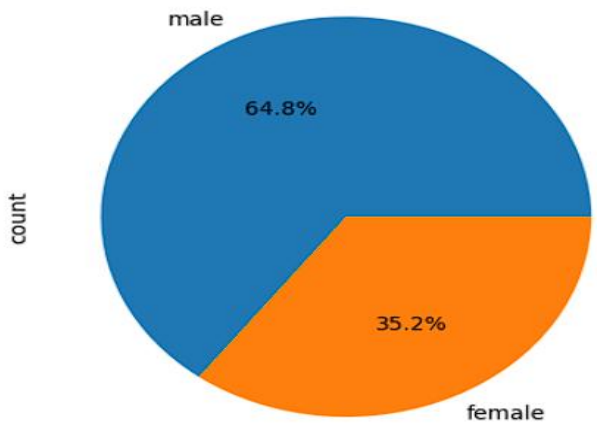
Pclass Column

- Around 55% of the passengers travel in class 3, 24% in class 1 and 20% in class 2.



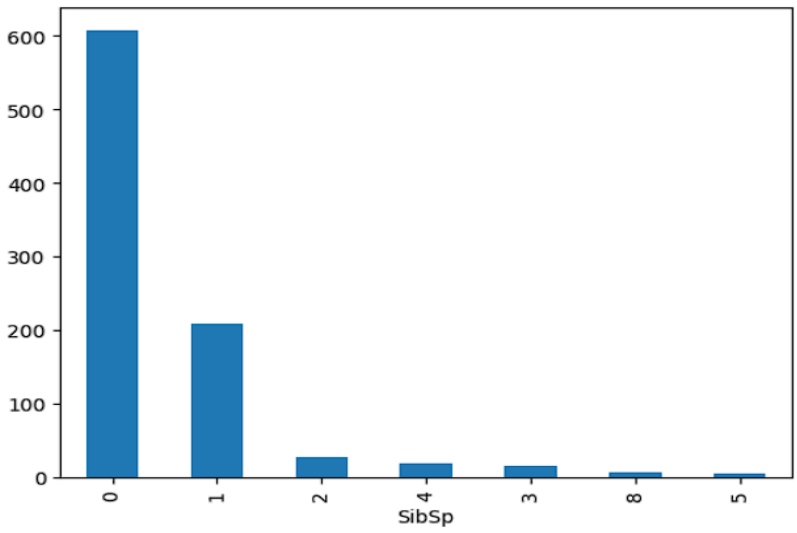
Sex Column

- Around 65% of the passengers are males.



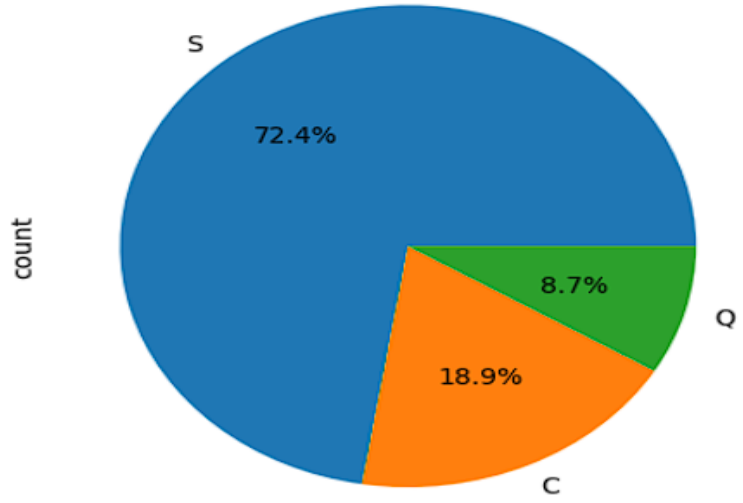
SibSp (Siblings and Spouse) Column

- Most of the people are either travelling alone or with one member.



Embarked Column

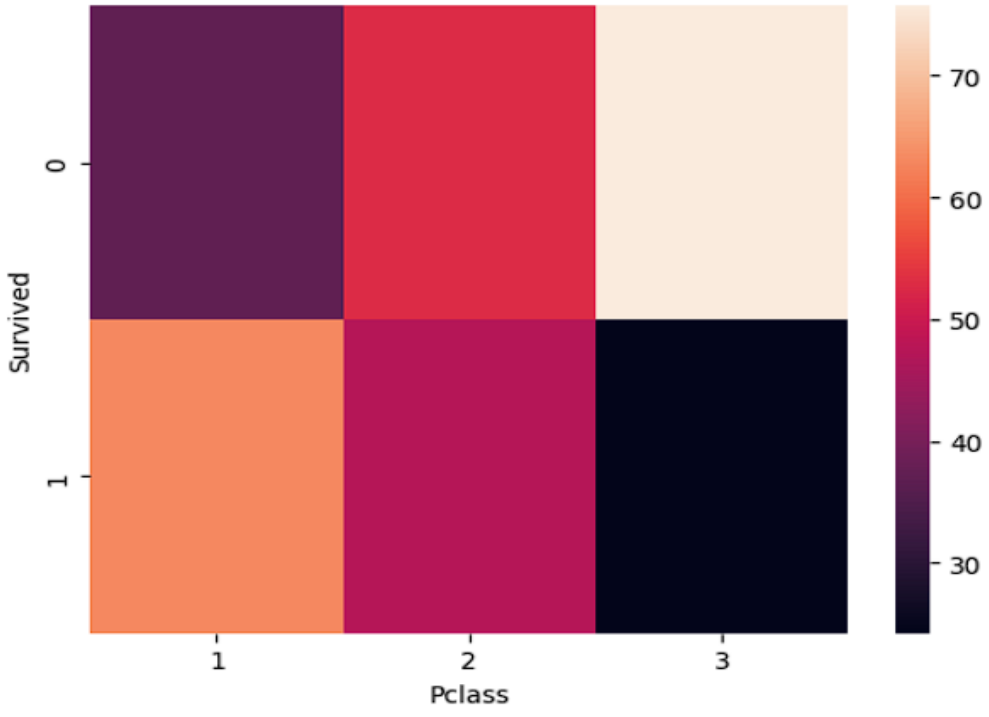
- Around 72% (approx.) of the people boarded from Southampton, 19%(approx.) from Cherbourg and 9%(approx.) from Queenstown.



Bivariate Analysis of columns

Survived and Pclass Columns

- 63% people survived in class 1, 47% people survived in class 2 and 24% people survived in class 3.
- Class 1 was safer than class 2 and class 3.

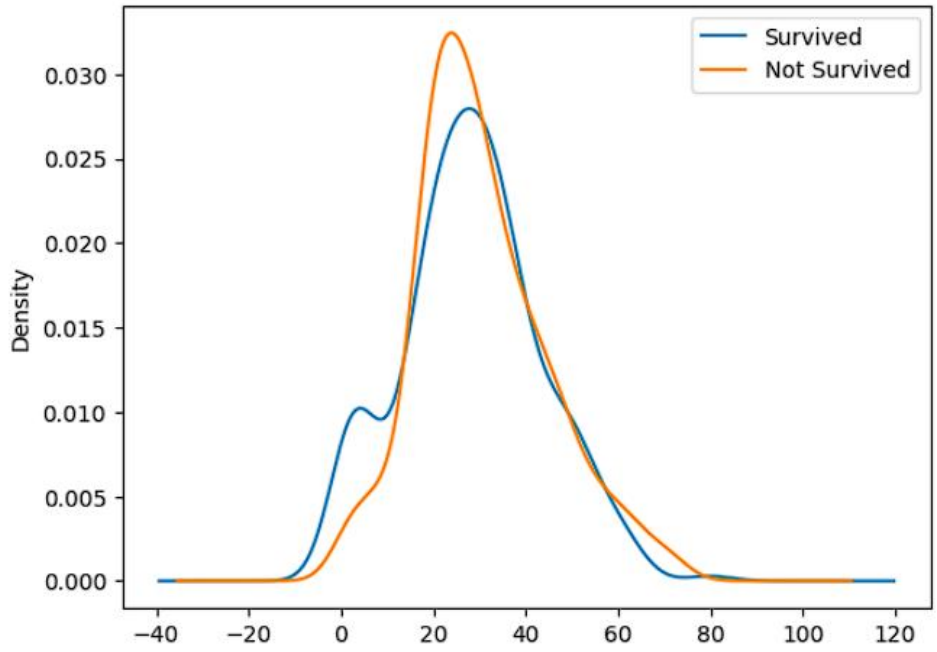


Survived and Sex Column

- 74% females and 24% males survived the accident.

Survived and Age Column

- Age group 0-10 - percentage survived is more than not survived.
- Age group 10-30 - percentage not survived is more than survived.
- Age group 30-60 - percentage survived and not survived is almost equal.
- Age group 60-80 - percentage not survived is more than survived.



Creating New Columns

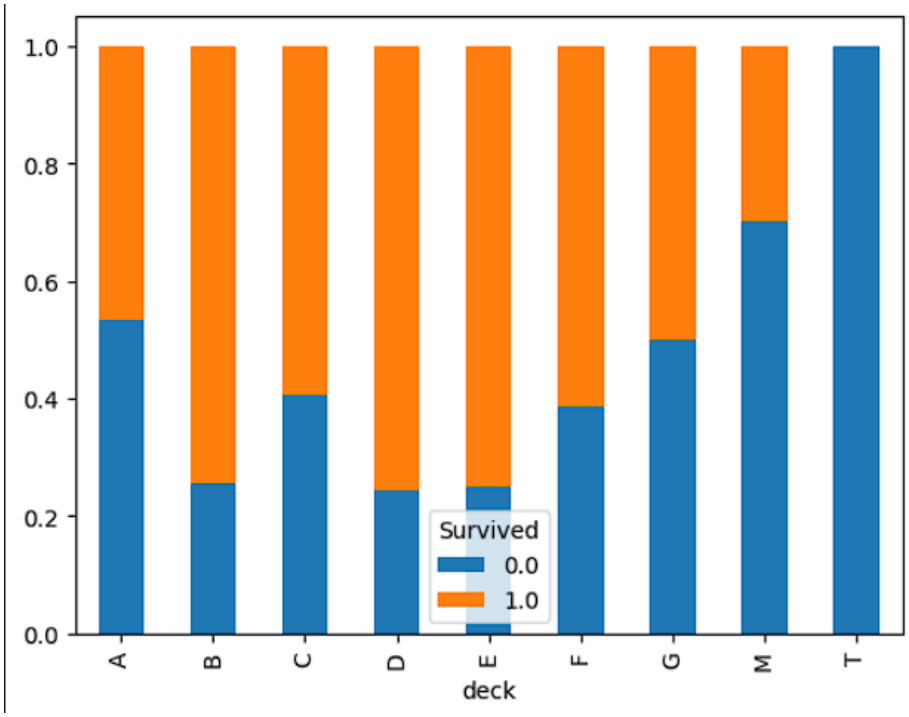
- Individual_fare: For separating individual fare of each passenger.
- family_size: This column will show family size of each passenger.

family_size is further transformed into family_type column.

- New column deck is created from the Cabin column.

Survived and deck column

- Probability of survival is varying deck to deck.



Key Insights

- The accident was very deadly, more than 60% people could not survive.
- 74% females and 24% males survived the accident.
- Class 1 was safer than class 2 and class 3.
- People with small family size (up to 4 members) had more chance of survival.

By Saptrishi Tiwari