README.md

1. Predicting the Survival of Titanic Passengers

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

The task was to predict using Titanic on-board passengers' data that whether a person would be able to survive the accident based on various factors such as his/her age, sex, ticket fare, etc.

1.1. Overview of the dataset

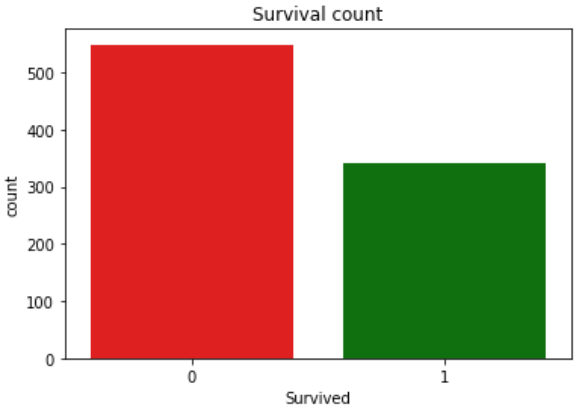
- 891 rows, 12 columns
- Column description:
  - "PassengerId" - Unique, ranges from 1 to 891
  - "Survived" - whether the passenger survived or not, 0 or 1
  - "Pclass" - class of passenger, 1 or 2 or 3
  - "Name" - Name of the passenger
  - "Sex" - sex of the passenger, male or female
  - "Age" - age of passenger
  - "SibSp" - number of sibling or spouse the passenger had
  - "Parch" - number of parents or children the passenger had
  - "Ticket" - ticket number
  - "Fare" - Fare of the ticket
  - "Cabin" - Cabin number of the passenger
  - "Embarked" - Place where the passenger embarked from (C = Cherbourg, Q = Queenstown, S = Southampton)

1.2. Data Visualization

Visualizing the dataset helps in better understanding of the data in hand and also helps in developing an intuition on whether a feature would be important in the final prediciton.

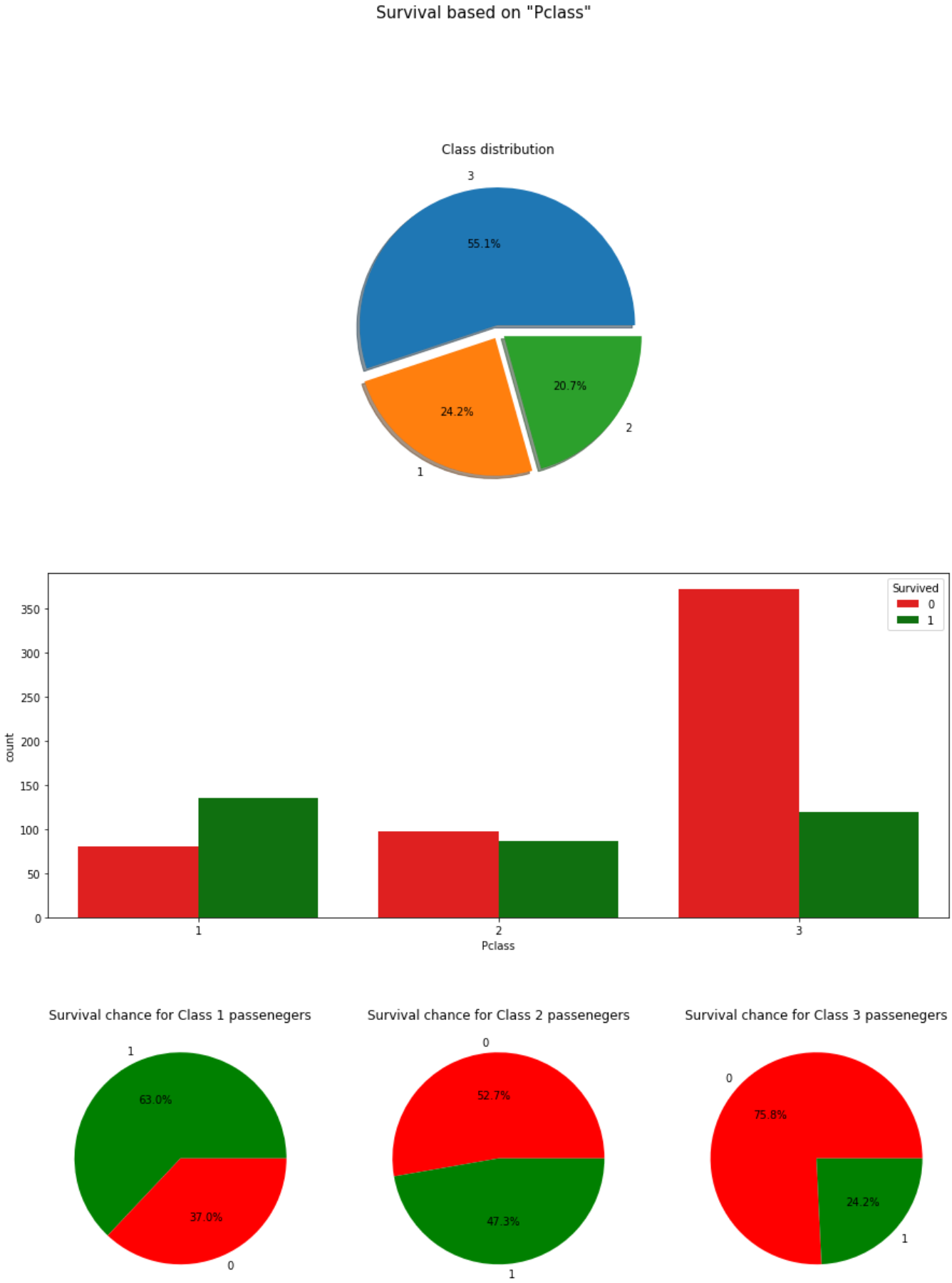
1.2.1. How many people survived

- Unfortunately, majority of the people didn't survive the Titanic accident.



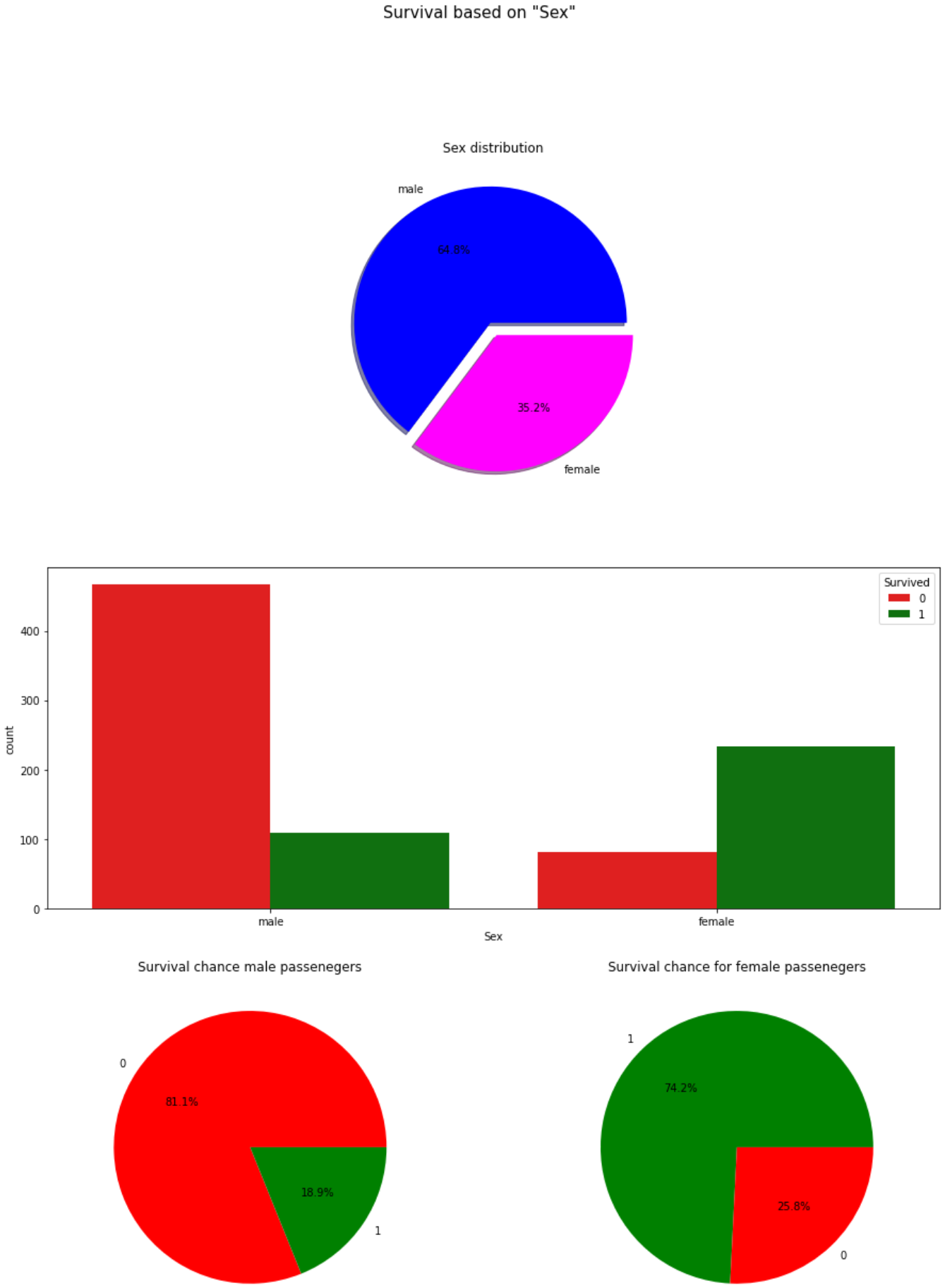
1.2.2. Survival based on "Pclass"

- Majority of the people from class1 survived.
- Almost half of people from class2 survived.
- Majority of the people from class3 did not survive.



1.2.3. Survival based on "Sex"

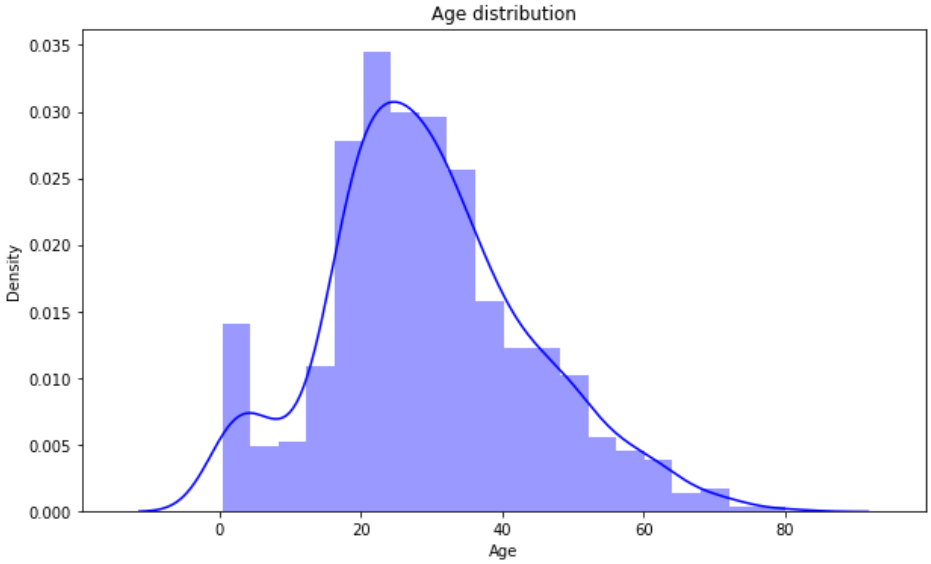
- Lesser proportion of males were able to survive the accident.
- Greater proportion of females were able to survive the accident.

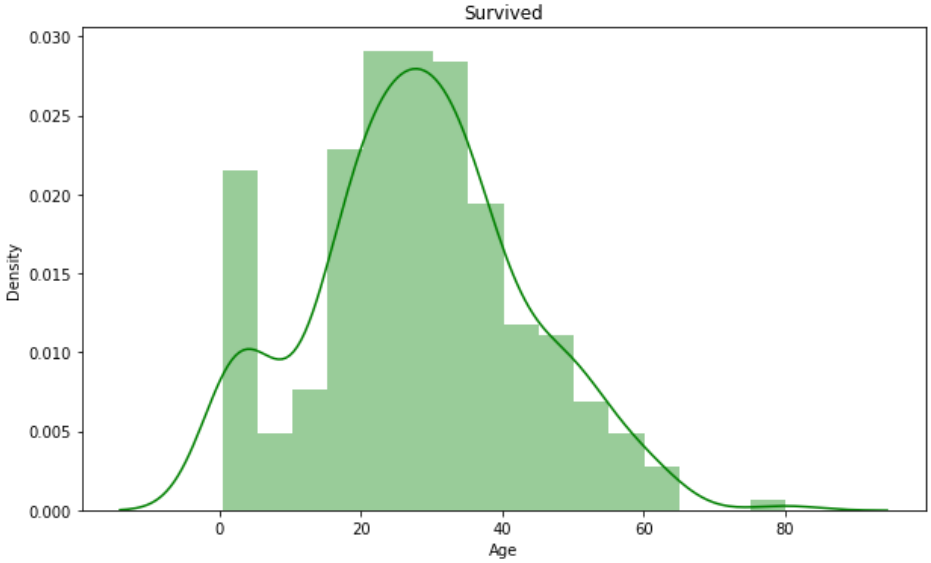


1.2.4. Survival based on "Age"

- Majority of passengers were between 20-40 years of age.
- Most of the infants (0-5 year olds) were able to survive.
- Many young adults (20-30 year olds) were not able to survive.

Survival based on "Age"

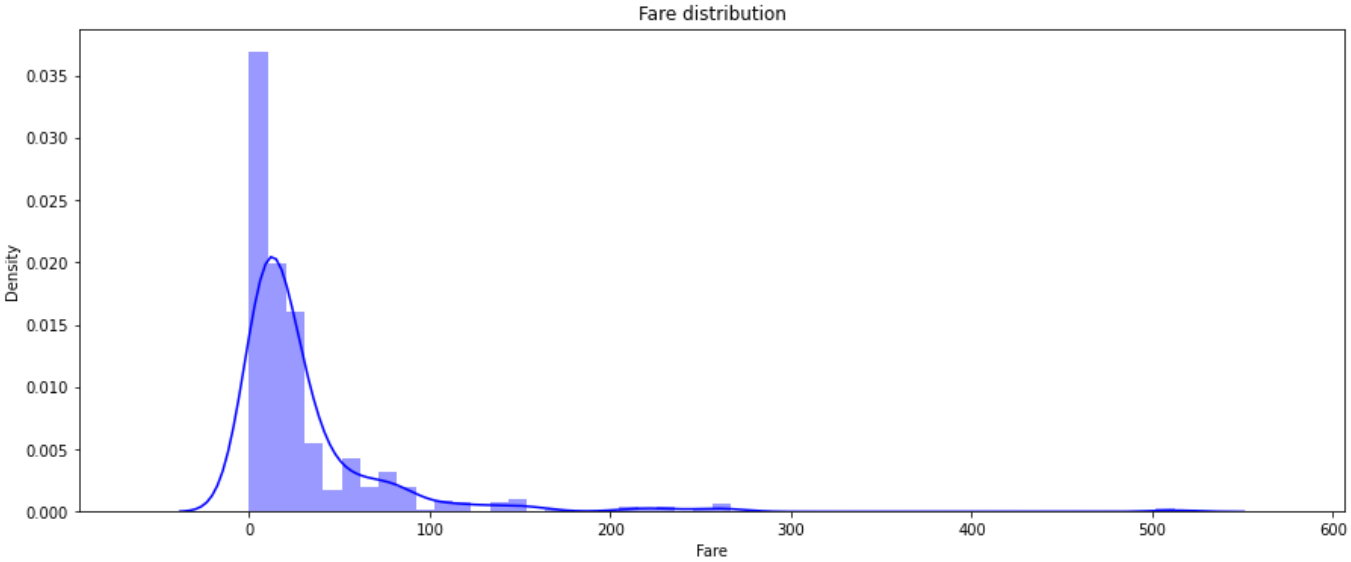


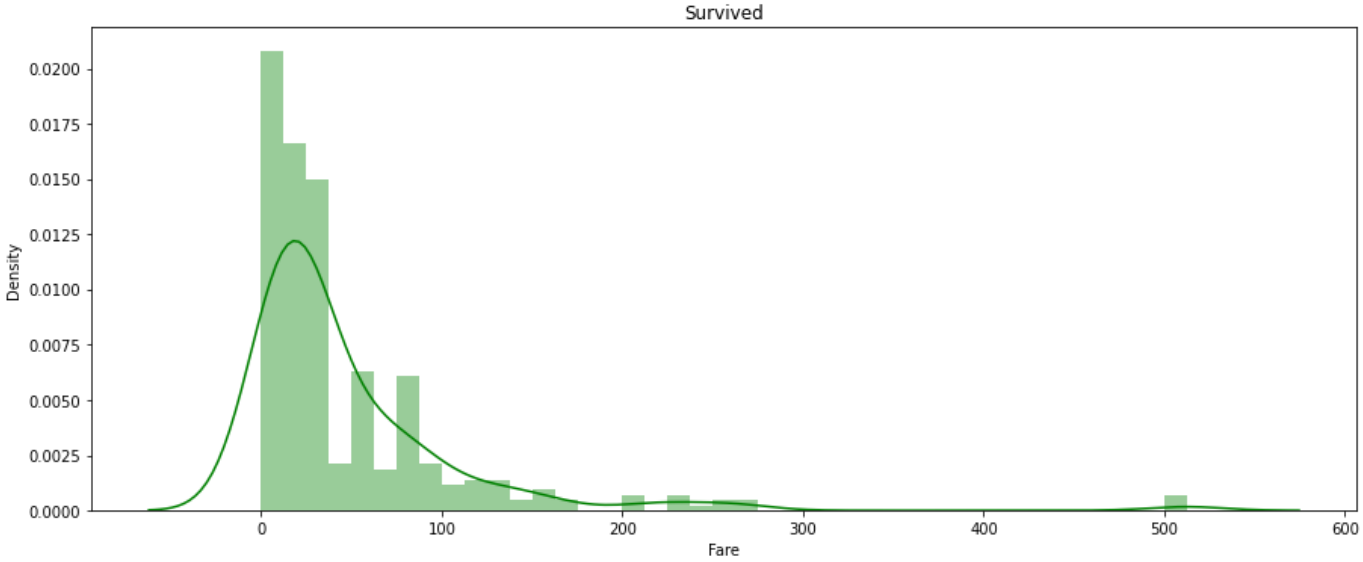


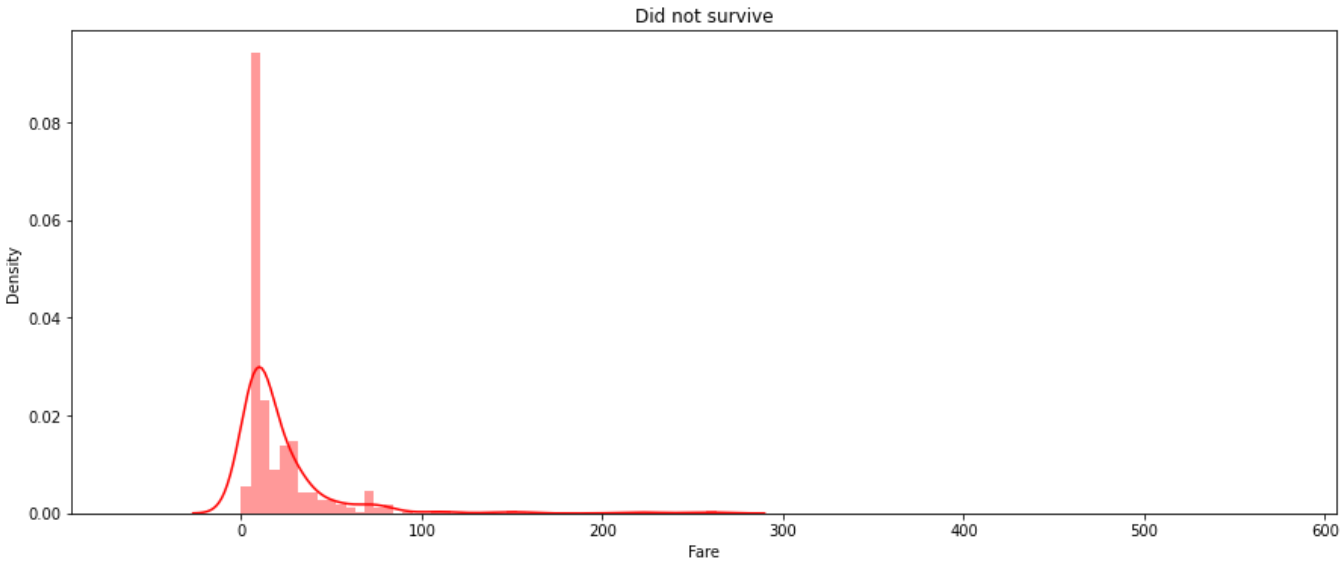
1.2.5. Survival based on "Fare"

- Most of the passengers were took low-fare tickets.
- Greater proportion of mid and high-fare ticket passengers were able to survive.
- Lesser proportion of low-fare ticket passengers were able to survive.

Survival based on "Fare"







1.3. Data Pre-processing

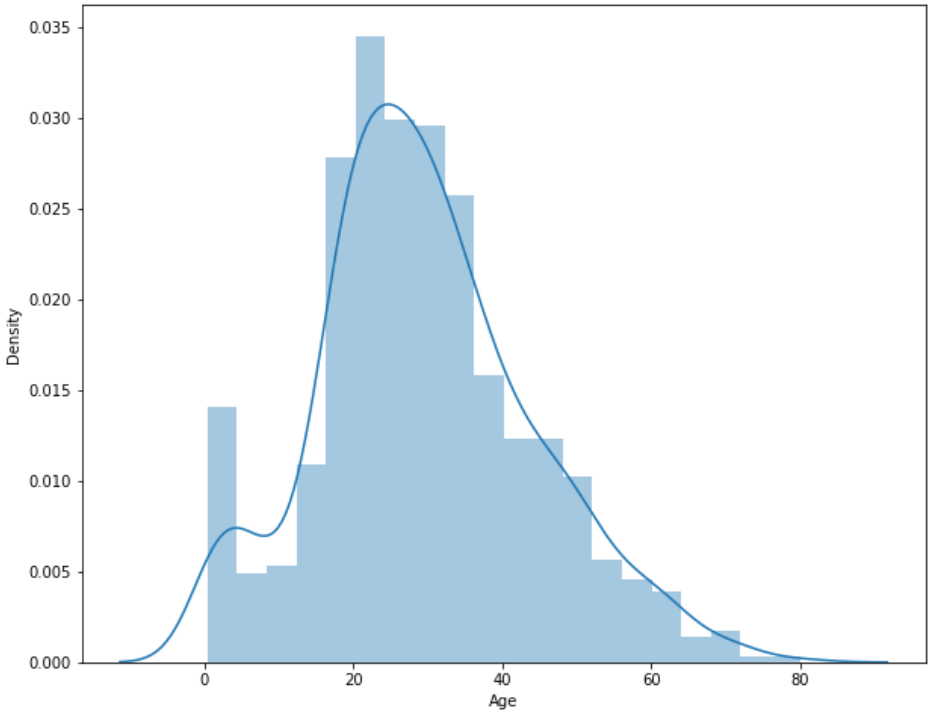
Pre-processing data means removing unwanted attributes from the dataset, filling missing values or removing the tuples with missing values, etc. It overall makes the data useful for the problem or usecase in hand.

1.3.1. Drop columns

- "Name" - general intuition that survival of a person should not depend on his/her name
- "Ticket" - general intuition that survival of a person should not depend on his/her ticket number
- "Cabin" - too many missing values

1.3.2. Filling NaN values

- Used a distribution plot to have an intuition about skewness of the data to make the decision whether to fill NaN value with mean/median/mode.



1.3.3. Replacing characters and strings with numerical data

- Machine learning models operate only in numerical data. Hence it is important to map any string/character type data to some numerical value.

1.4. Building the machine learning model

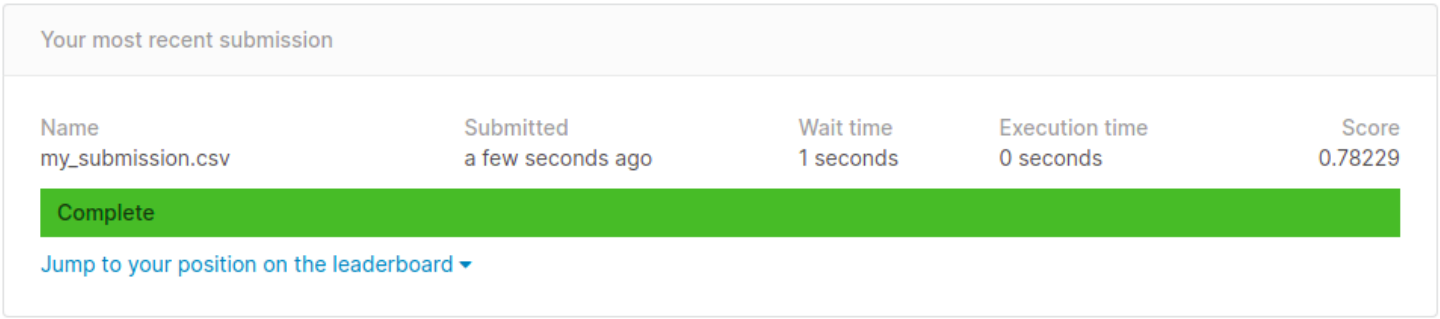
After preparation of the dataset, the final step is to build our machine learning model which is basically an algorithm which takes in our processed data and gives a relevant output based on our use-case.

1.4.1. Reasons to use Random Forest Classifier

- The problem in hand is a classification problem.
- RFC makes used of a number of decision trees and combines their result for better and stable predictions.
- Easier to implement using *sklearn*.

1.5. Submission and assessment score

On submitting the predicted values for the challenge, the following score was achieved.



1.6. Key learnings

- Learnt about basics of Random Forest Classifier and decision trees.
- The concept of bagging in machine learning which says that a combination of ,machine learning models increases the overall result favourably.