**titanic**: Titanic Passenger Survival Data Set.

Binary classifications problem.

# Load the dataset

Setting Survived column for test data to NA

# Note: titanic\_test$Survived <- NA

PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin Embarked

rbind(titanic\_train, titanic\_test)



## &#39;data.frame&

1309 obs. of 12 variables:

## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...

## $ Survived : int 0 1 1 1 0 0 0 0 1 1 ...

## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...

## $ Name :

## $ Sex : chr &quot;male&quot; &quot;female&quot; &quot;female&quot; &quot;female&quot;

## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ... Siblings/Spouses

## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...

## $ Ticket : chr &quot;A/5 21171&quot; &quot;PC 17599&quot; &quot;STON/O2.

## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

## $ Cabin : chr &quot;&quot; &quot;C85&quot; &quot;&quot; &quot;C123&quot; ...

## $ Embarked : chr &quot;S&quot; &quot;C&quot; &quot;S&quot; &quot;S&quot; ...

**Embarked** implies where the traveler mounted from.

Southampton, Cherbourg, and Queenstown

Up to question 7 we are doing basic EDA and data pre-processing for our model.

- str

-rbind

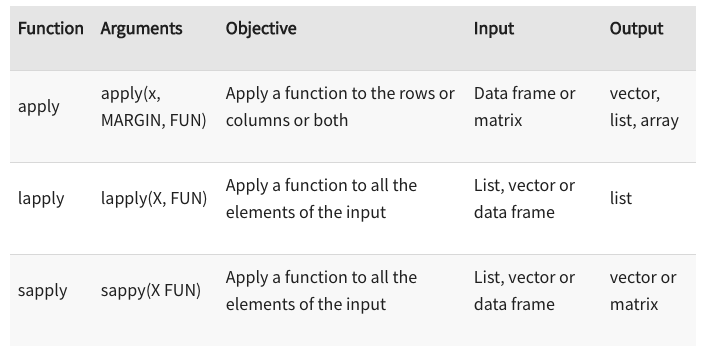
-NULL

-Missing values

Apply use to apply Functions Over any Array with given Margins

1 – Row

2 – column



8 – 10

Check number of unique values for each column

Removing unwanted columns

Convert to factors. (categorical variable. )

**11. Splitting training and test data.**

%>% piping function

**library("mice")**

# missing data pattern

md.pattern(df)

# predictive mean matching

# sample –

X – population data

Size –

There are two categories in survived 0 and 1.

By looking the summary of train dataset we see survived (0:549 ,1:342).

In order to train the model with equal level of survived 0 and 1, here we use upSample().

Upsampling is the process of inserting zero-valued samples between original samples to increase the sampling rate.

14 **ANOVA**

Basically, it is Used to find how a quantitative dependent variable changes according to the levels of one or more categorical independent variables.

**ANOVA** tests whether there is a difference in means of the groups at each level of the independent variable.

**Receiver Operator Characteristic (ROC)**

A **ROC curve** plots the false alarm rate against the hit rate for a probablistic forecast for a range of thresholds.

The area under the **curve** is viewed as a measure of a forecast's accuracy.

A measure of 1 would indicate a perfect model.

Area under curve