

American International University - Bangladesh

Introduction to Data Science [A]

Midterm Project Report

**Submitted to -**

Tohedul Islam Sir

Dataset – Titanic Survivors

**Submitted by –**

FARUQUE, SAKIF AL

20-42308-1

BSc. in CSE

Dataset Description:

This dataset is about the accident of Titanic sinking. Some data have been exhibited. Here, 6 of the attributes are categorical that are, gender, embarked, class, who, alone, survived and rest of the 4 attributes are numerical that are, age, sibsp, parch, fare.

Retrieving the dataset from .csv file and general analysis,

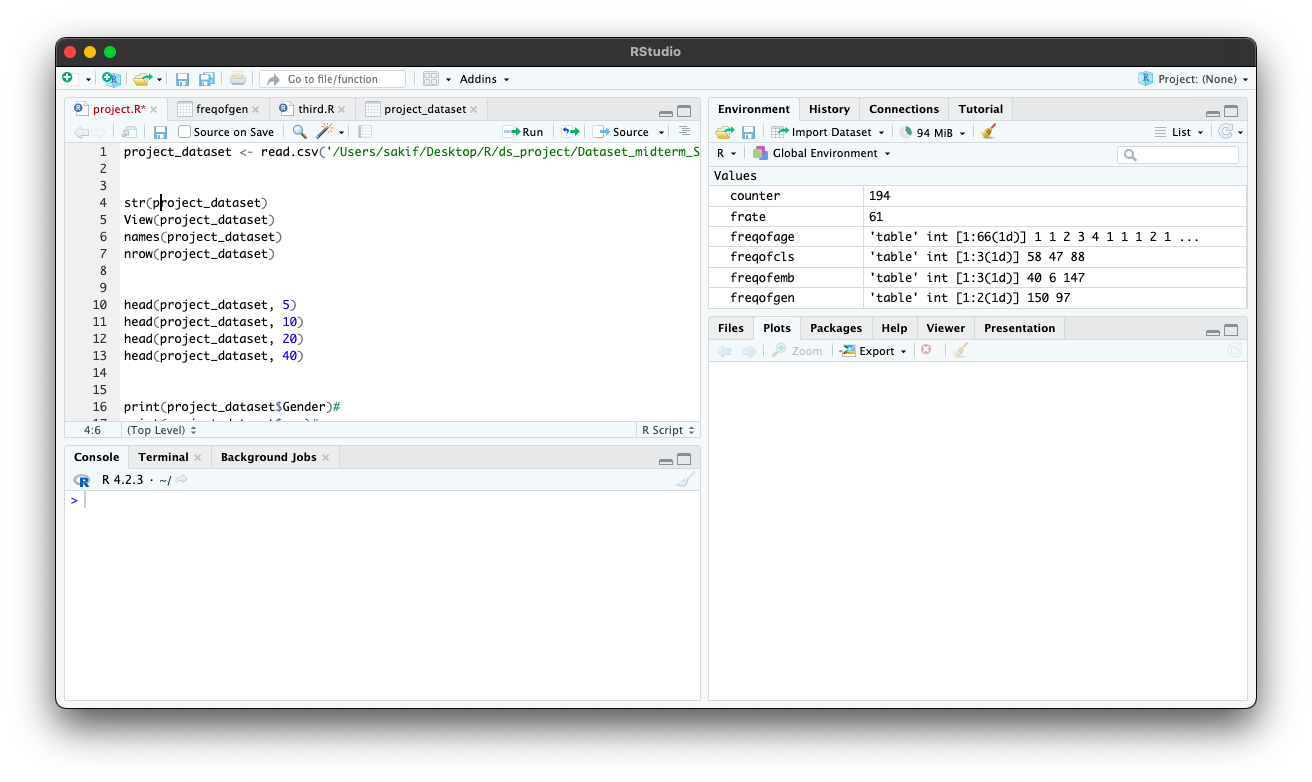


Figure : Overview of instructions

Line-1 is used for reading the data from csv.

Line-4 provides a short summary of the dataset,

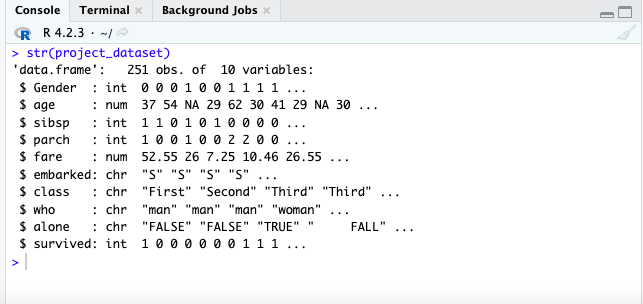


Figure : Type of the attributes in the Dataset

Line-5 provide dataset view,

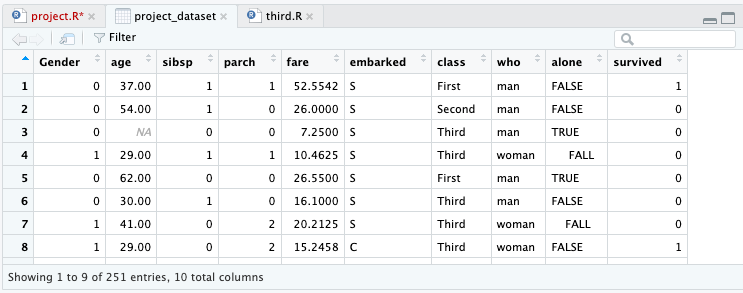


Figure : View of the uncleaned dataset

Line-6 & 7 provides the names of the columns and no. of row of dataset.

The definitions of the attributes have been shown below,

|  |  |  |
| --- | --- | --- |
| sex | Sex/Gender | male/female |
| age | Age |  |
| sibsp | siblings of the passenger | 0/1 /2 … |
| parch | parents / children aboard the Titanic | 0/1/2 … |
| fare | Passenger fare |  |
| embarked | Port of Embarkation | C : Cherbourg, Q : Queenstown, S : Southampton |
| class | Ticket class | First / Second / Third |
| who | categories to passengers | male, female, child |
| alone | he was alone in ship or no | 0/1 |
| survived |  | 0/1 |

Defining male and female values in the Gender attribute-

Form the website, the number of male is greater than the female of the Gender attribute.

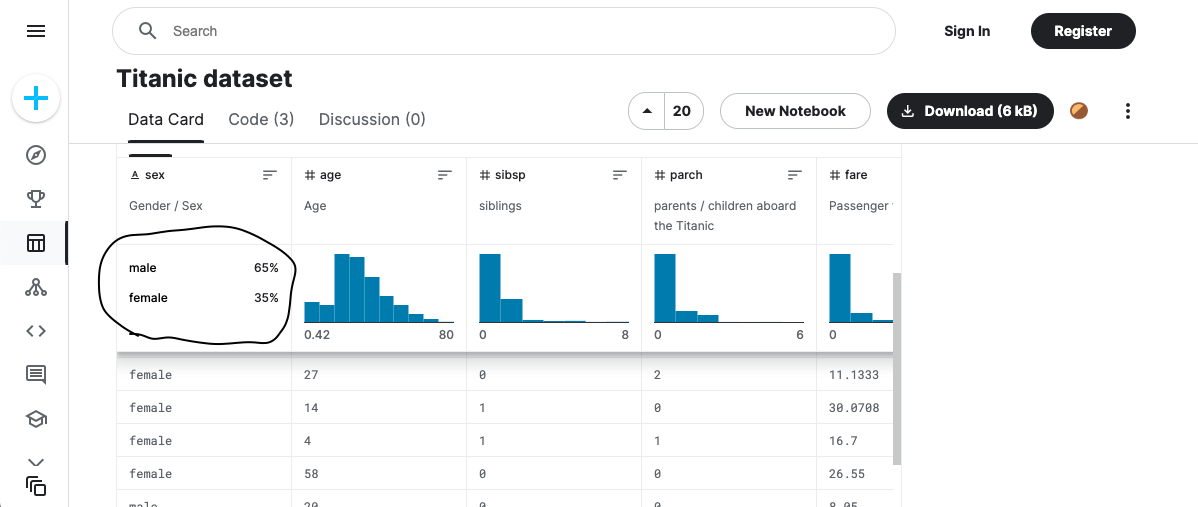


Figure : Percentages of male & female in the dataset

And in the dataset, the number of 0’s is greater than the 1’s.

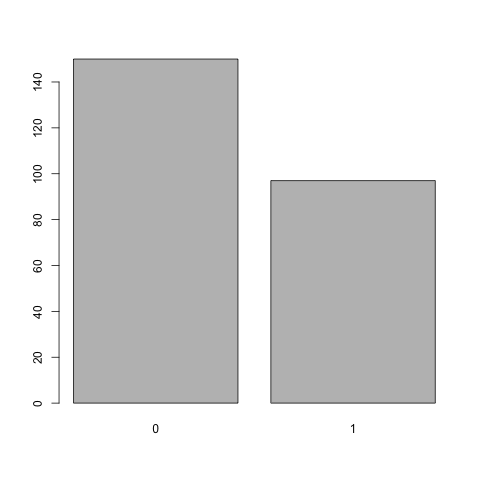


Figure : Graph of gender frequency from the dataset

Code:



Figure : This code is producing the above gender bar plot

So, in the dataset, 0 represents male and 1 represents female. [0 -> male | 1 -> female]

In the survival attribute, according to the general hypothesis,

0 represents FALSE -> not survived.

1 represents TRUE -> survived.

Missing values:

Below command shows a general overview about those attributes which has missing values.

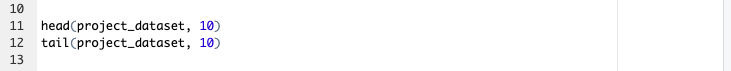


Figure : To show top and bottom 10 rows

Overview:

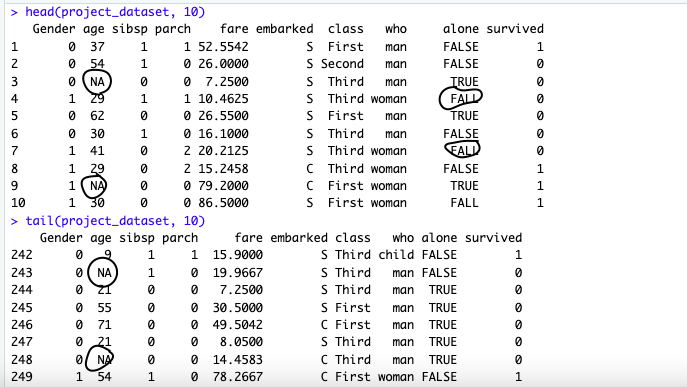


Figure : Overview of first and last ten data

Now, every attribute is checked individually for missing or irrelevant values,



Figure : Individual attribute checking

Overview for gender: It has some missing values,

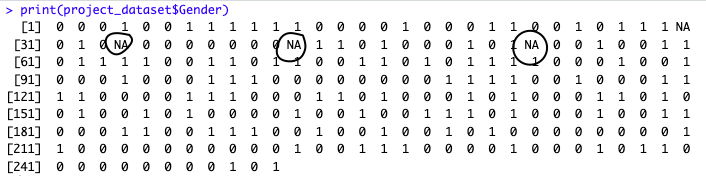


Figure : Missing values in gender

Overview for age: It has some missing values,

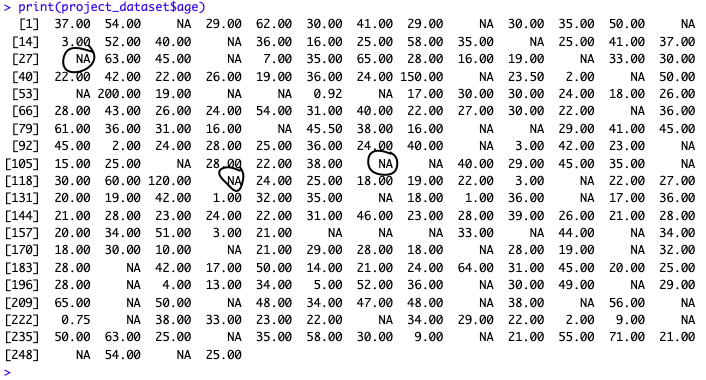


Figure : Missing values in age

Overview for class: It has some missing values,

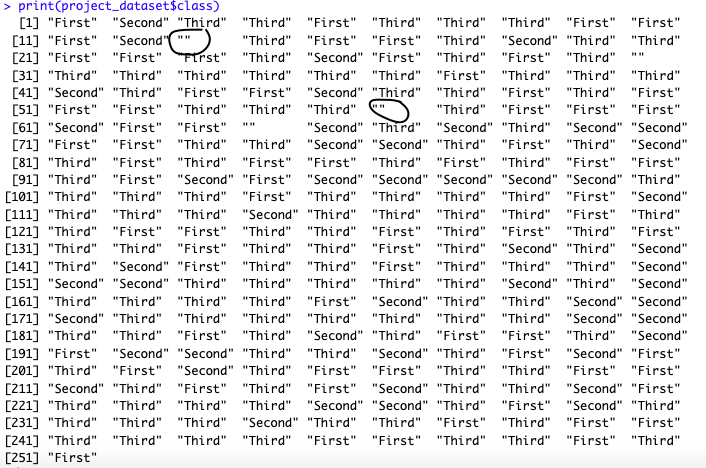


Figure : Missing values in class

Overview for alone: It has some irrelevant values,

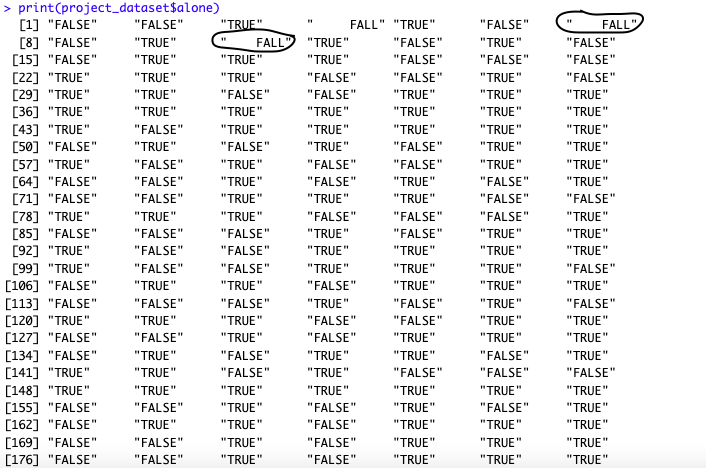
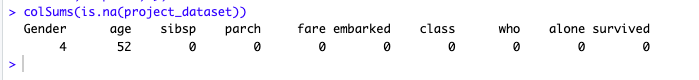


Figure : Irrelevant values in alone

The summary of the missing values is (before transformation),



After transformation stage, there will be more NA values. These NA values will be discarded after applying the transformation.

Cleaning all the NA values from the dataset,



Figure : Commands for showing the count of NA values and discard all the NA values

Data Quality:

* Data freshness

This dataset is updated in the last year. Considering the context of data, this dataset is about historical incident. So, it doesn’t matter how old the data are.

* Data correctness

According to the subject of that incident, all the attributes are relevant.

* Data completeness

For sampling data is quite complete.

Data types:

Before applying the type conversion of data, the overview of alone attribute,

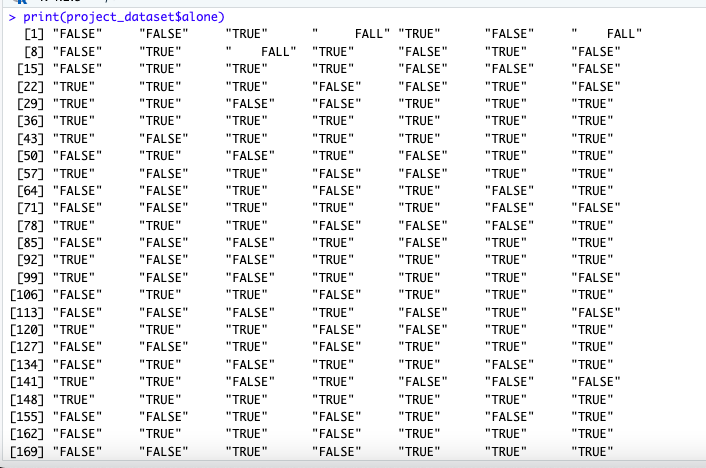


Figure : Values of alone before type conversion

Data type conversion is need for alone attribute. It is in string type and needs to convert in Boolean type.

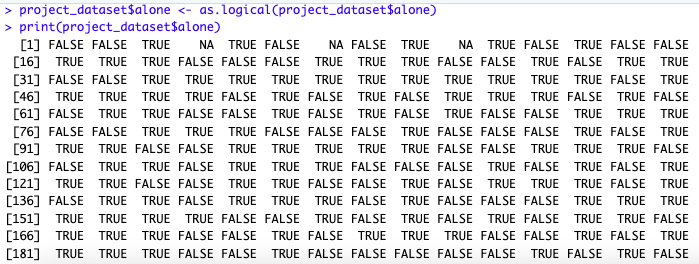


Figure : Values of alone after type conversion

Data transformation:

In the age attribute, outlier and missing values should be replaced with NA value,



Figure : Data transformation of age

In the fare attribute, outlier and missing values should be replaced with NA value,



Figure : Data transformation of fare

In the class attribute, there are some irrelevant data which needs to transform into NA.



Figure : Data transformation of class

In the alone attribute, garbage values should be discarded (or replaced with NA),

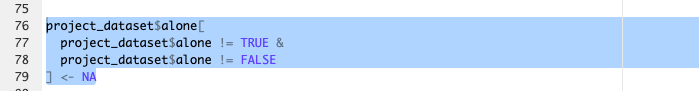


Figure : Data transformation of alone

After transformation, the summary of the missing values is,

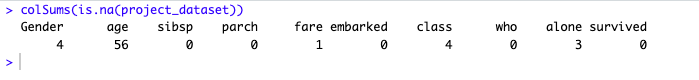


Figure : Missing values count

And the view of the dataset,

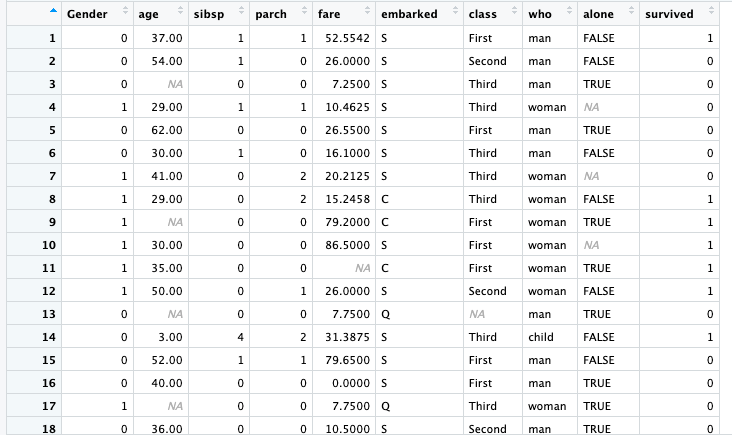


Figure : Dataset view with missing values

Outliers:

For categorical data, barplot has been used to identify the outliers.

Outlier data found in class attribute,

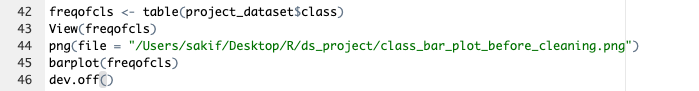


Figure : Bar plot command for the frequency of class

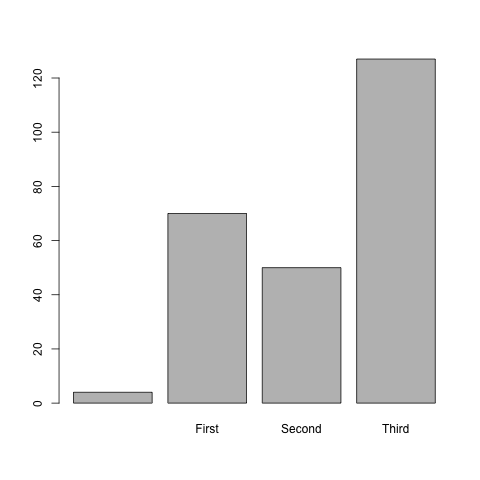


Figure : Bar plot of frequency of class

Outlier data found in alone attribute,

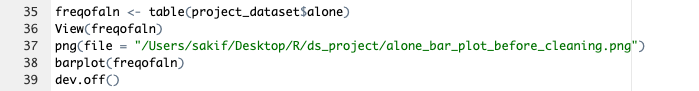


Figure : Bar plot command for the frequency of alone



Figure : Bar plot of frequency of alone

For numerical data, boxplot has been used to identify the outliers.

Outlier data are found in age attribute,

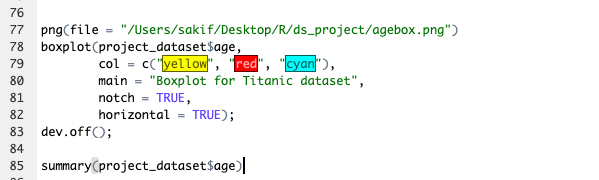


Figure : Box plot command for the frequency of age

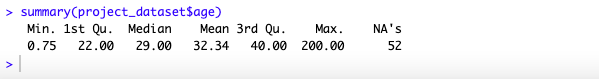


Figure : Summary of age column

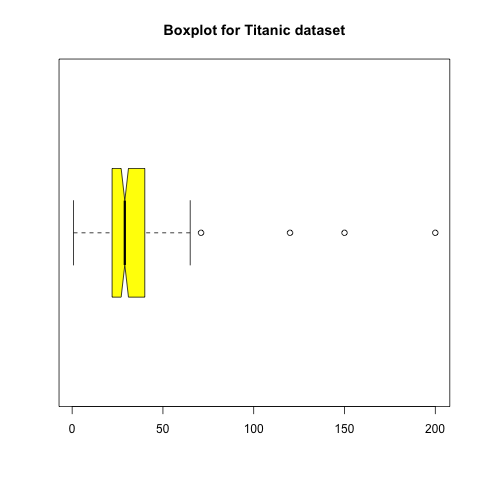


Figure : Box plot of age

Outlier data are found in fare attribute,

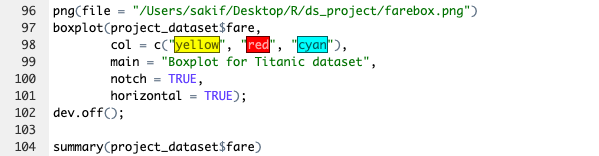


Figure : Box plot command for the frequency of fare



Figure : Summary of fare column

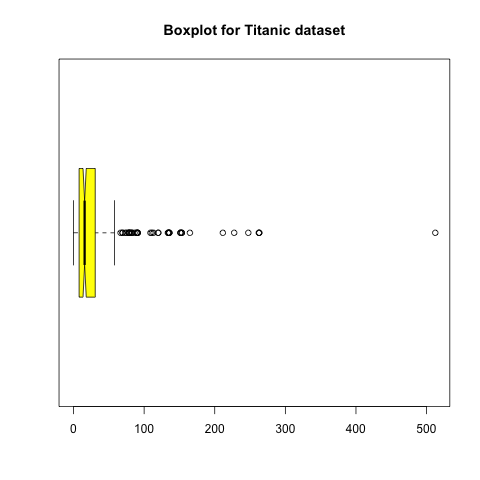


Figure : Box plot of fare

The outlier data are transformed into NA values in transformation stage.

Data exploration:

After cleaning the dataset, the overview,

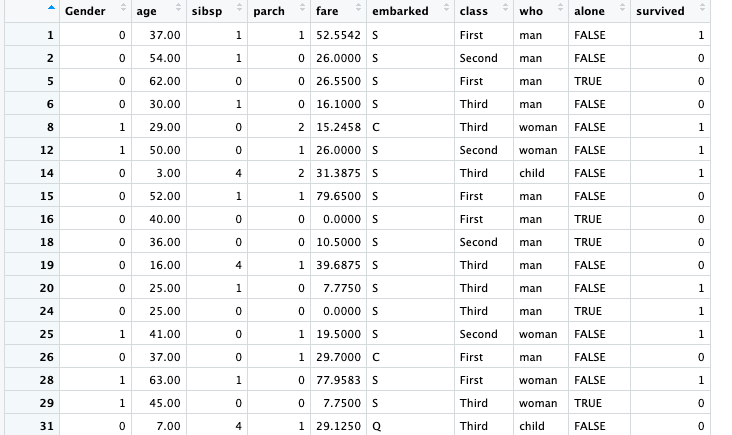


Figure : Clean Dataset

No. of observations: 161

No. of columns: 10

Summary of every attribute presents in the dataset,

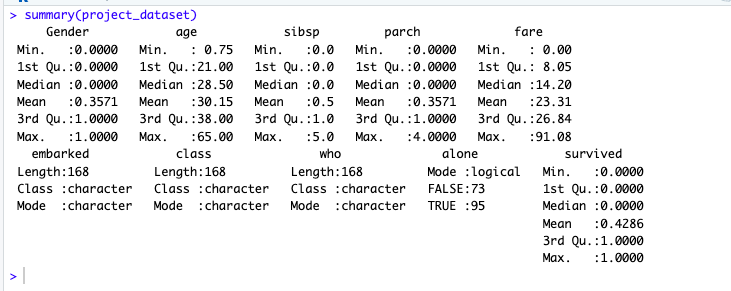


Figure : Summary of every attribute

There were two attributes which have outlier data and high standard deviation. Now, the outlier values of those attributes are handled and the values of standard deviation got decreased.

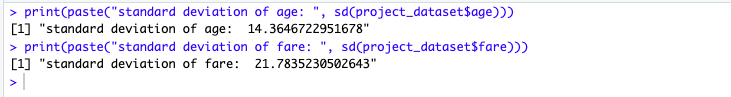


Figure : Standard deviation of age & fare

Below the hist graphs are represented for better understanding of the deviation of age and fare attributes.

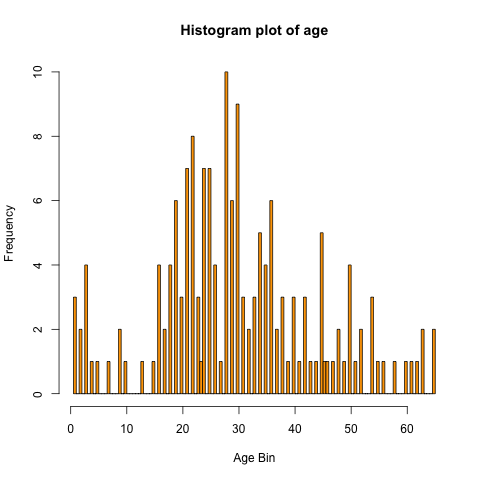


Figure : Histogram plot of age

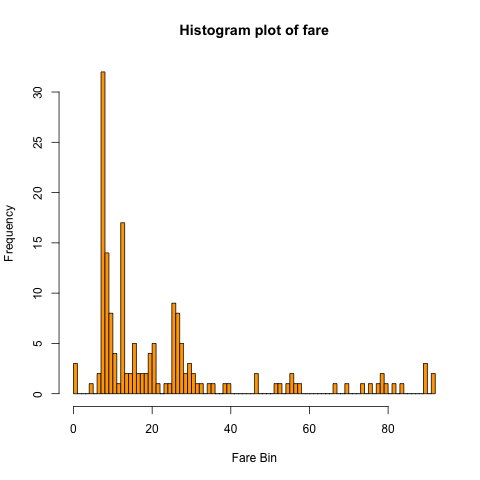


Figure : Histogram plot of fare

Feature selection

For this dataset, the sibsp, parch and who attributes have less priority. So, these attributes can be discarded.

Final attributes:

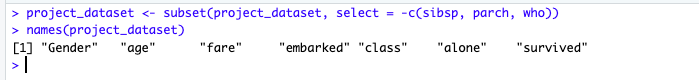


Figure : Selected features

Sampling

After cleaning, the dataset looks like this, which has 161 observations.