**Mid Term Project**

**Introduction to Data Science**

**Submitted by: Shahriar Soudeep**

**ID: 20-43823-2**

**Section: C**

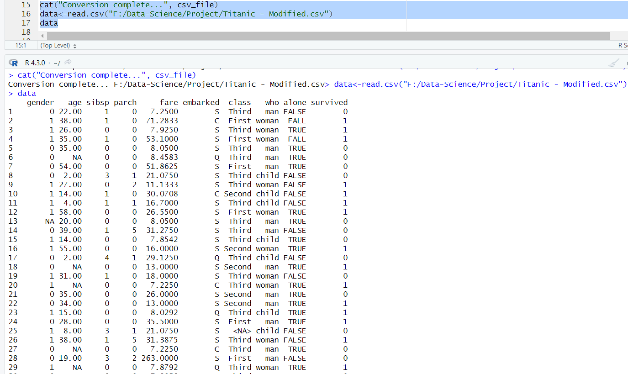
**Submitted to: DR. ABDUS SALAM**

**Dataset Description:**

The Titanic dataset is a well-known historical dataset that contains information about the passengers.

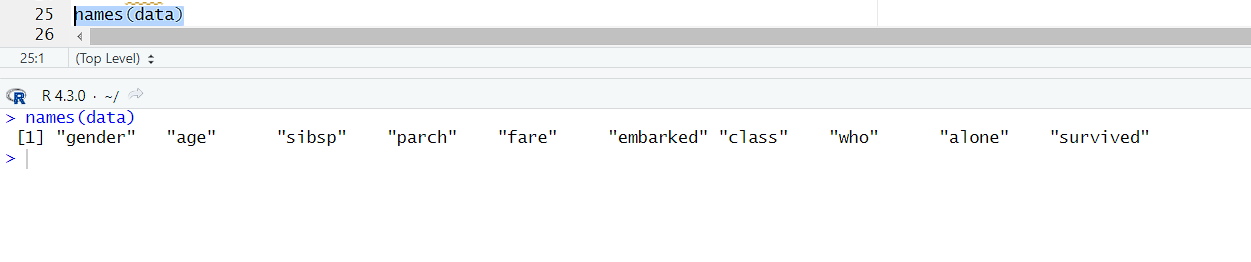
The dataset typically includes various attributes for each passenger, such as age, sex, ticket class (Class), ticket fare (Fare), number of siblings/spouses aboard (Sibs), number of parents/children aboard (Parch), port of embarkation (Embarked), and survival status (Survived). The 'Survived' variable indicates whether a passenger survived the disaster (0) or did not survive (0). It provides an opportunity to study the characteristics and demographics of the passengers aboard the Titanic and to analyze factors that may have influenced their chances of survival.

**Convert XLSX to CSV and Import CSV and print the dataset:**



**Description:**

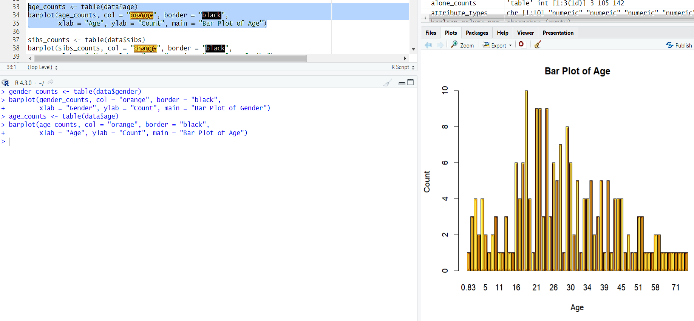
The code is for read the dataset and also to prove the successful conversion of a xlsx file to csv file. Also, the output of the dataset by using the name of the variable name ‘data’.

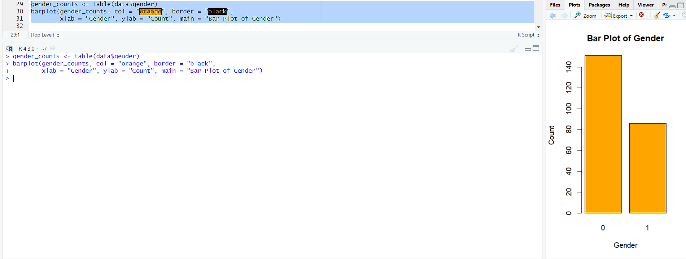
**See Column Name:**

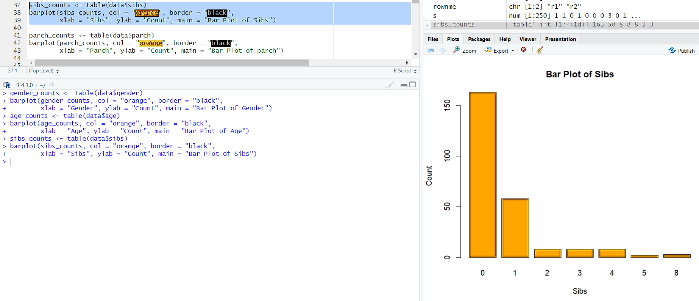
**Description:**

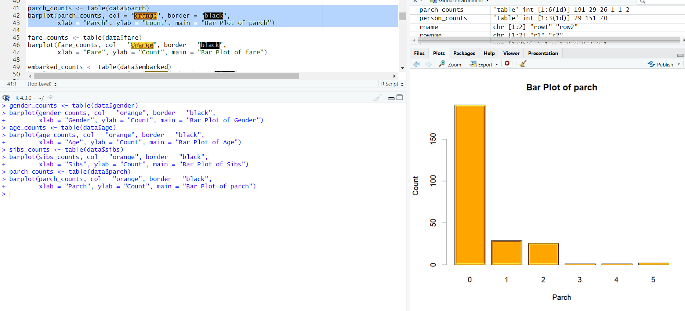
By using the ‘names()’ function we can see the column name of the dataset. To get the actual idea we need to see that how many column are there. So, that we can further apply the data preparation steps.

1. **Exploration steps:**



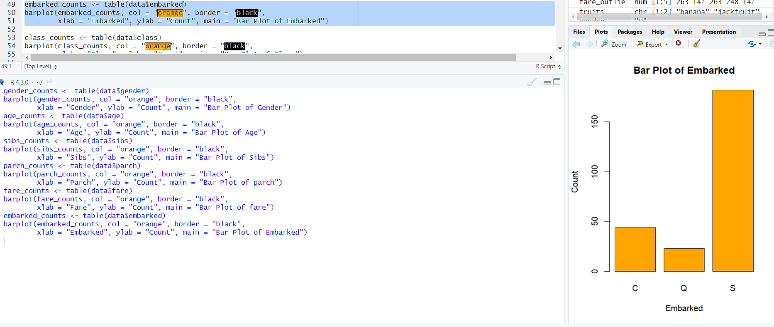


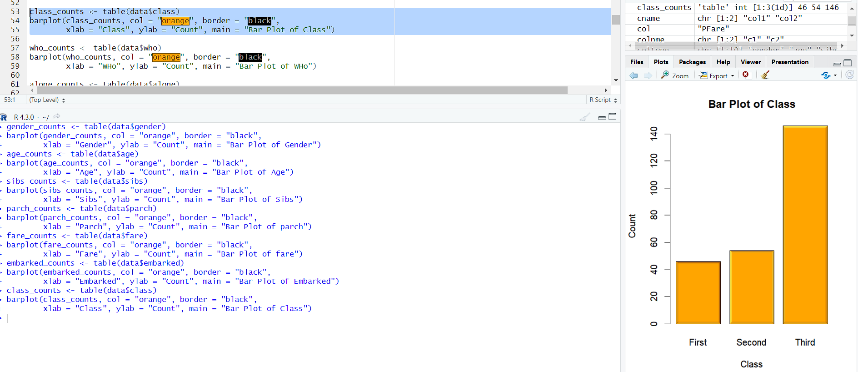


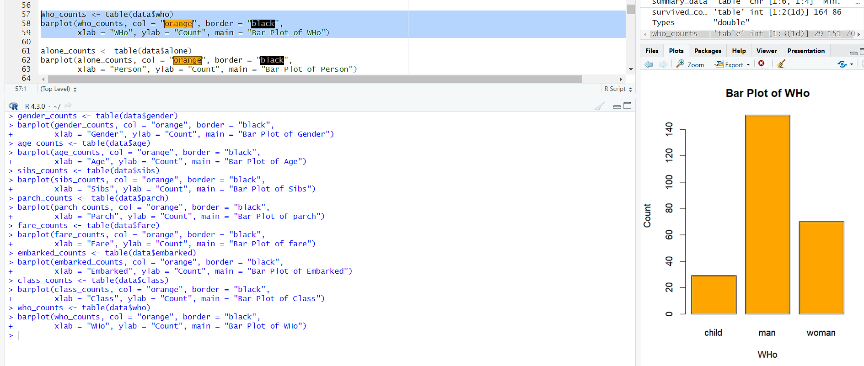


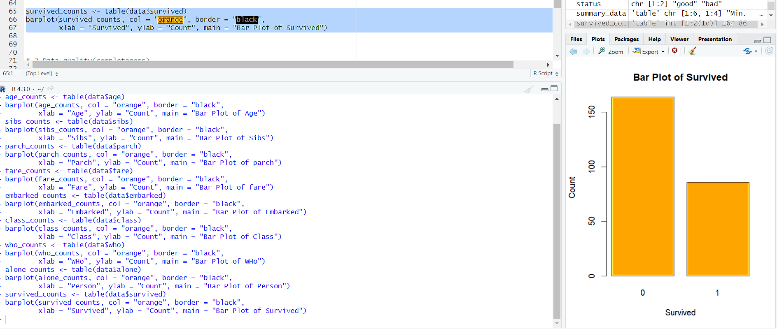
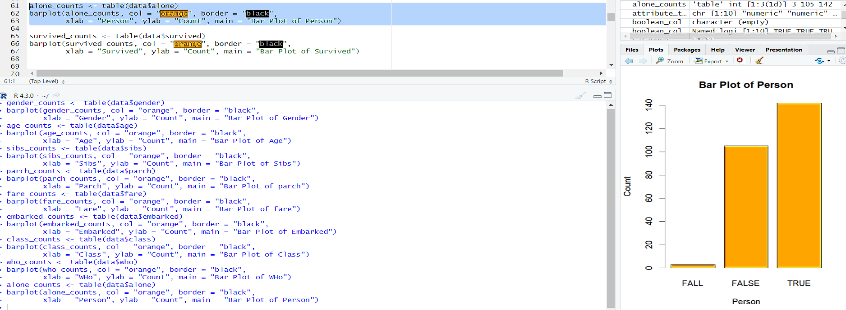
A screenshot of a computer

Description automatically generated





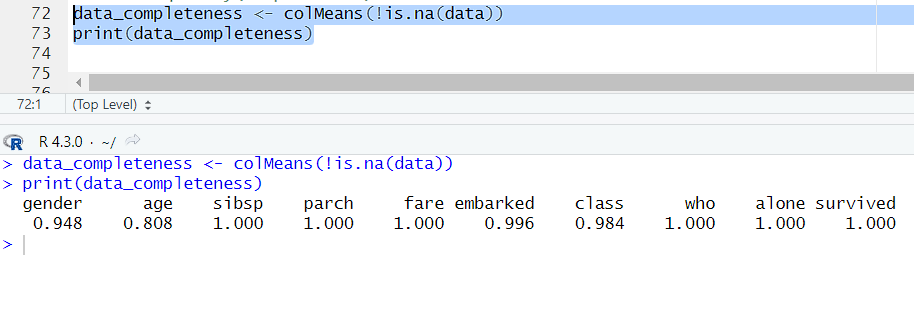




**Description:**

At first count the total value of every column and than using ‘barplot()’ function draw the Bar-chart to visualize the dataset of every column.

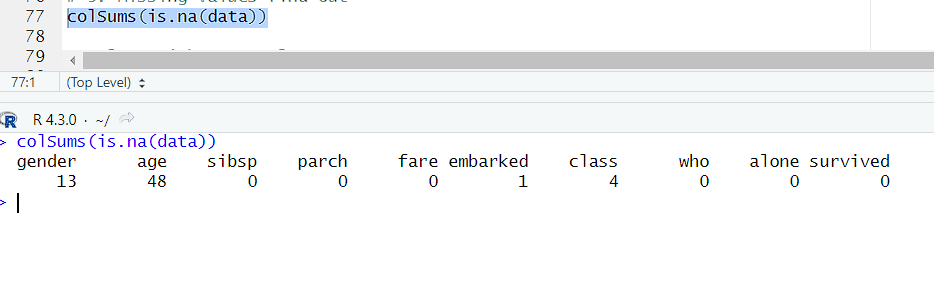
1. **Data-Quality(Check Data-Completeness):**



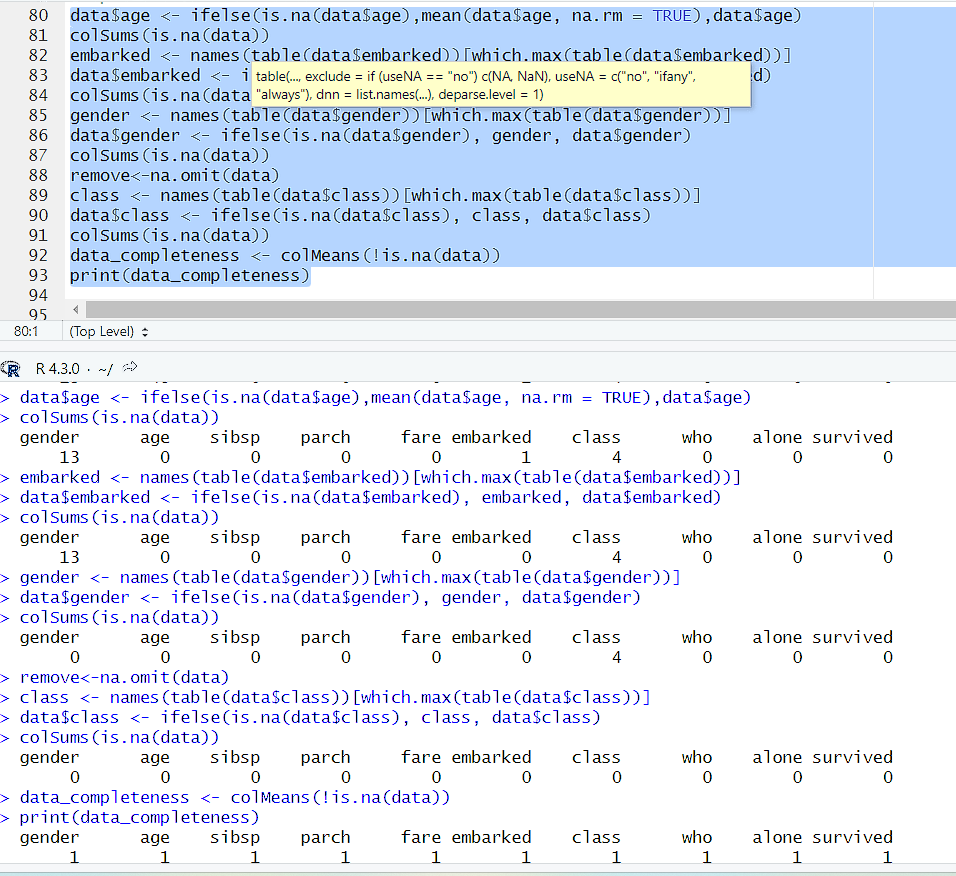
**Description:**

To check the dataset is complete or not we use ‘colMeans()’ function. If,not complete than it shows less than 1 value but if it is complete than it shows the value 1.

1. **Find Missing Values:**



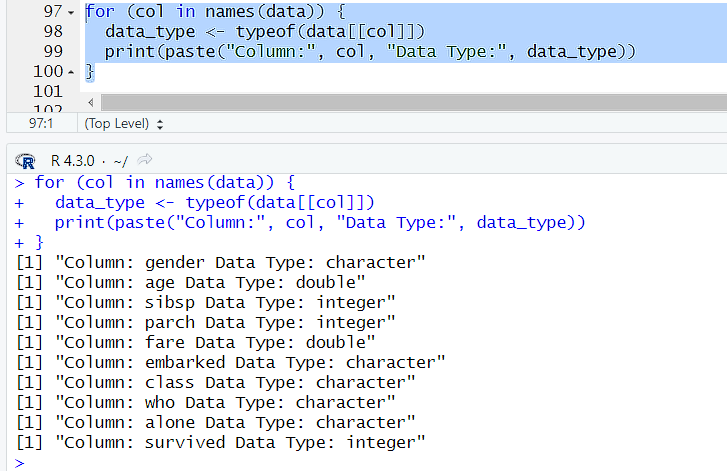
Handling Missing values:



**Description:**

First find the missing value from every column by using colSums() function. Than, check the datatype if numeric than replace with mean values but if it is categorical than replace with most frequent values.

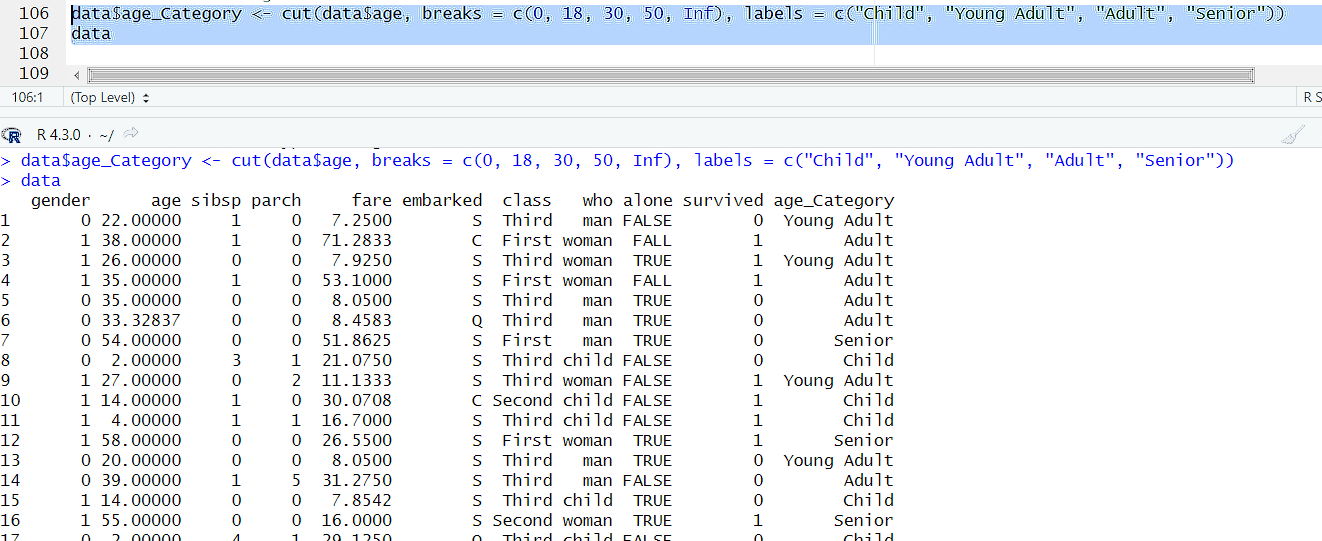
**4.Check Data-Types and Conversion:**



**Description:**

Check the datatype by using names() function. It help to find the class and also using typeof() function we can find the exact data type of every column.

**Conversion:**



**Description:**

We can convert data types. Here we do binning. Numeric data types to categorical data types. With cut() function.

1. **Descriptive Statistics(SD,MEAN,MEDIAN for numeric values):**

A screenshot of a computer

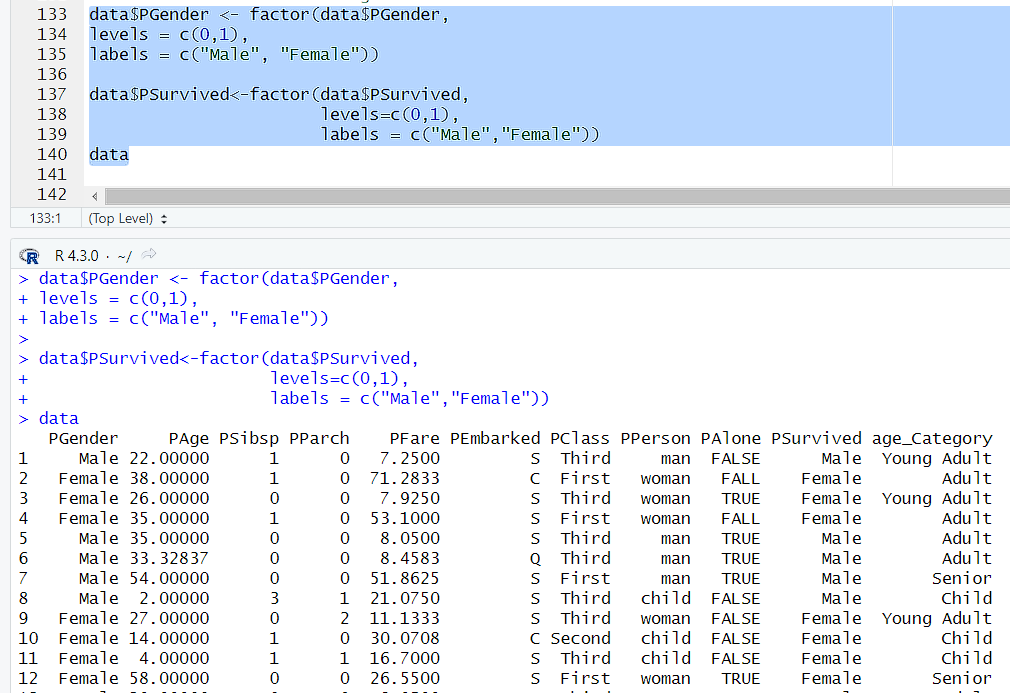
Description automatically generated

**Description:**

With summarise\_if() function we can get the standard deviation, Mean, Median values. %>% summarise\_if() we can get the multiple columns sd, mean, median values if the colums is numeric.

1. A screenshot of a computer

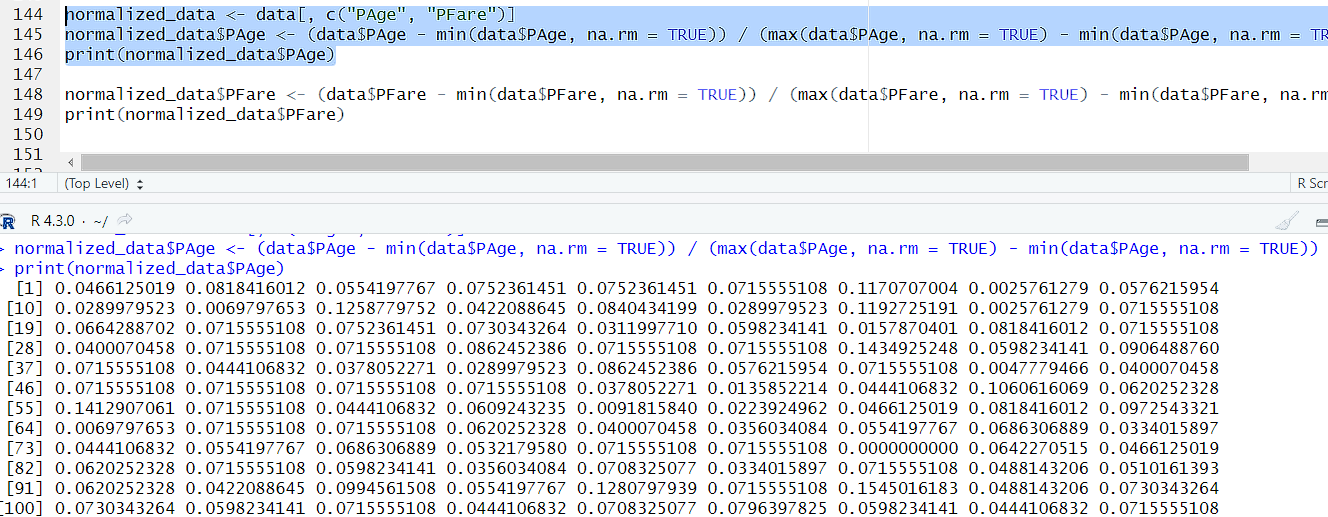
   Description automatically generated**Annotation for categorical value(Column and value):**

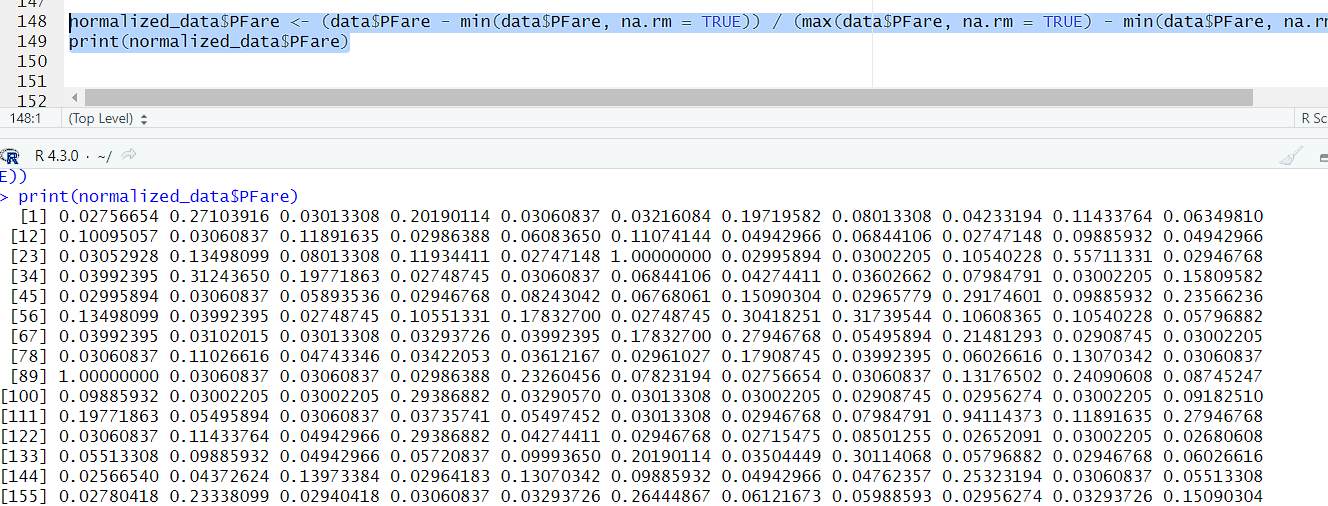


**Description:**

After finding the data-types, for categorical valus we do the annotation. 1st picture the column annotation with names[1]<- new. Than, 2nd picture we annotate the values of column values of categorical data.

1. **Transformation(Normalization):**

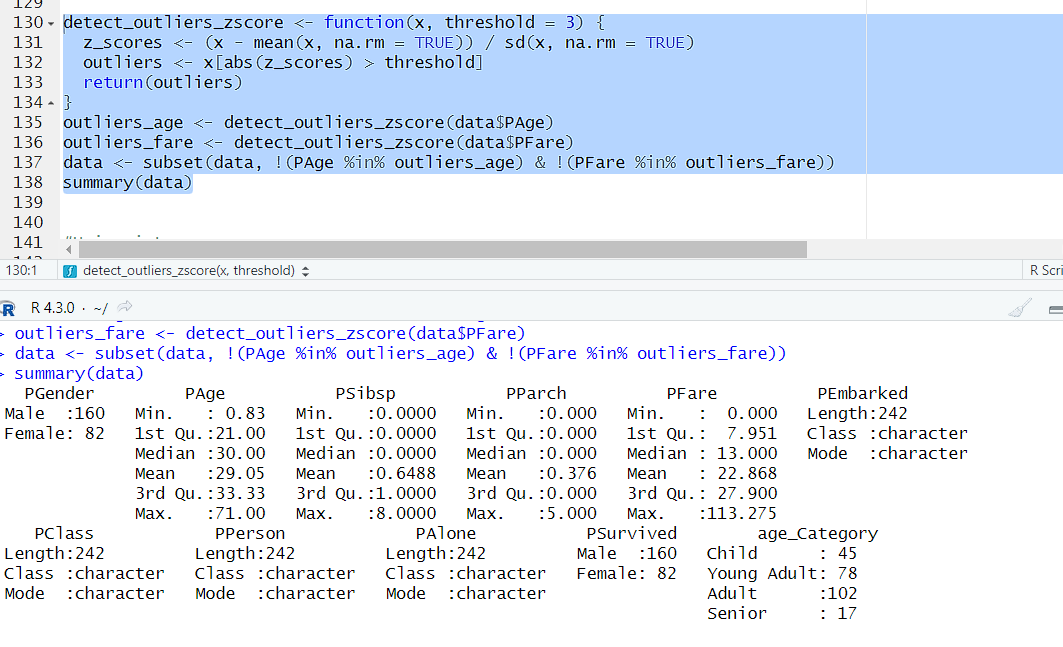




**Description:**

For transformation, here did the normalization. Simple formula (value – min value) / (max value – min value)

1. **Outliers (Detect and Handle)**

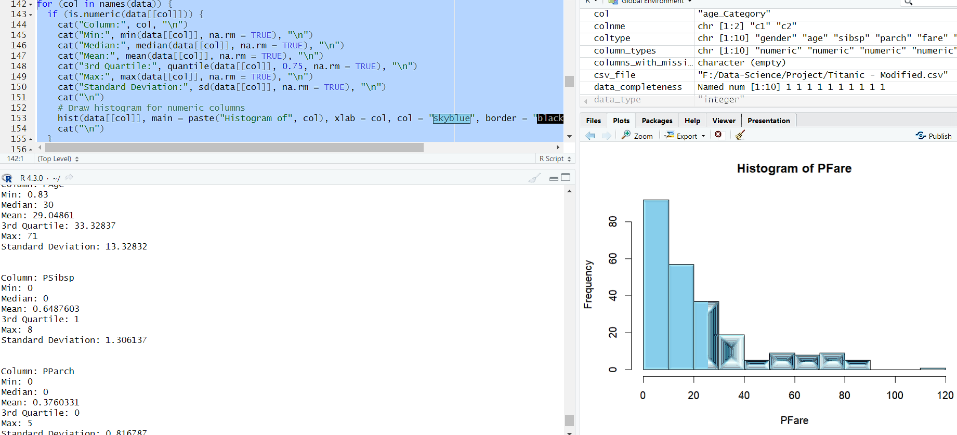
****

**Description:**

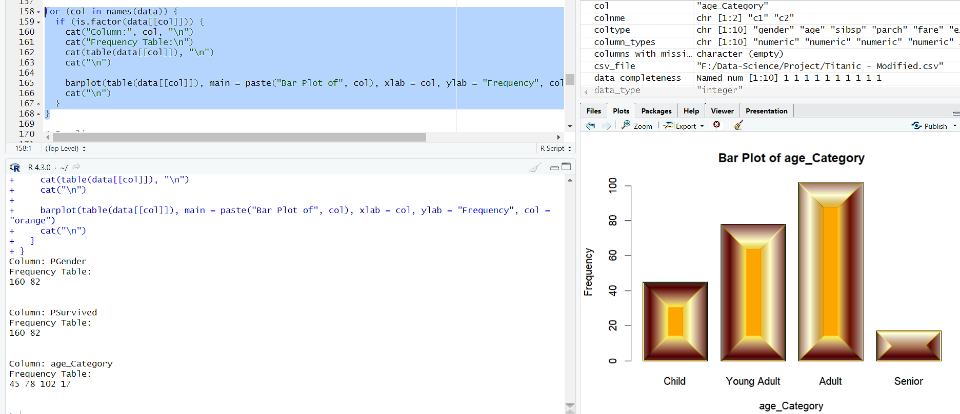
Here first detect the outliers with z\_score() function. Than, we remove this value from the dataset by using subset() function. First in, age column there was outlier 356 and it was the highest value. After removing this value the highest value is 71.00. So,successfully handled the outliers.

1. **Univariate:**

Numeric(Histogram)

****

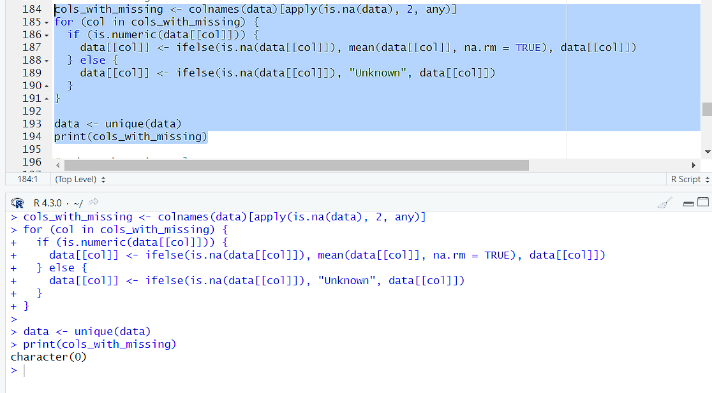
Categorical (Frequency-table)



**Description:**

For numeric data we find the min value, max value, mean, sd, median values and also draw the histogram in picture-1. In second picture the value was categorical for this we use frequency-table to draw a bra-chart by using barplot() function.

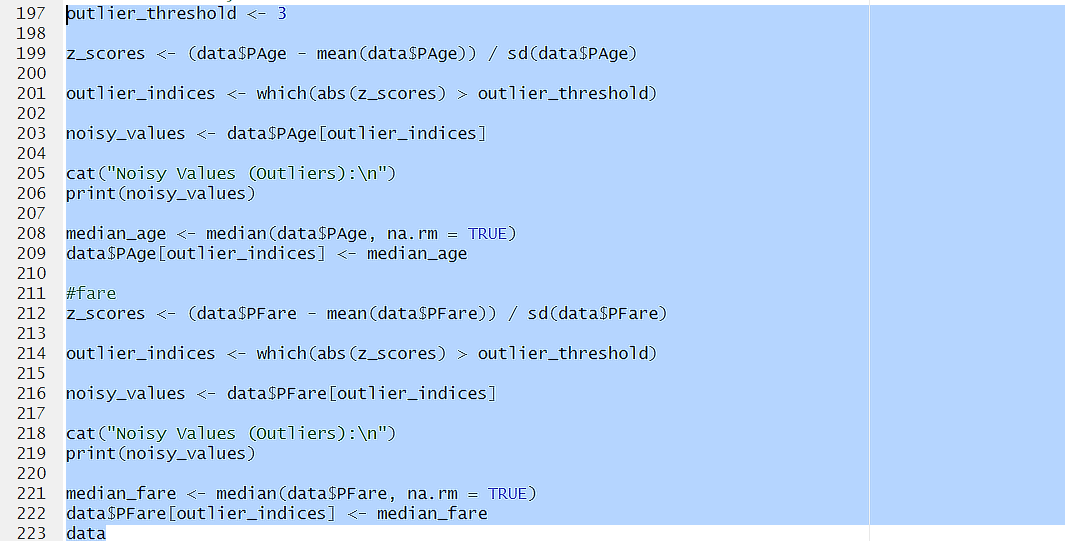
1. **Data-Cleaning:**

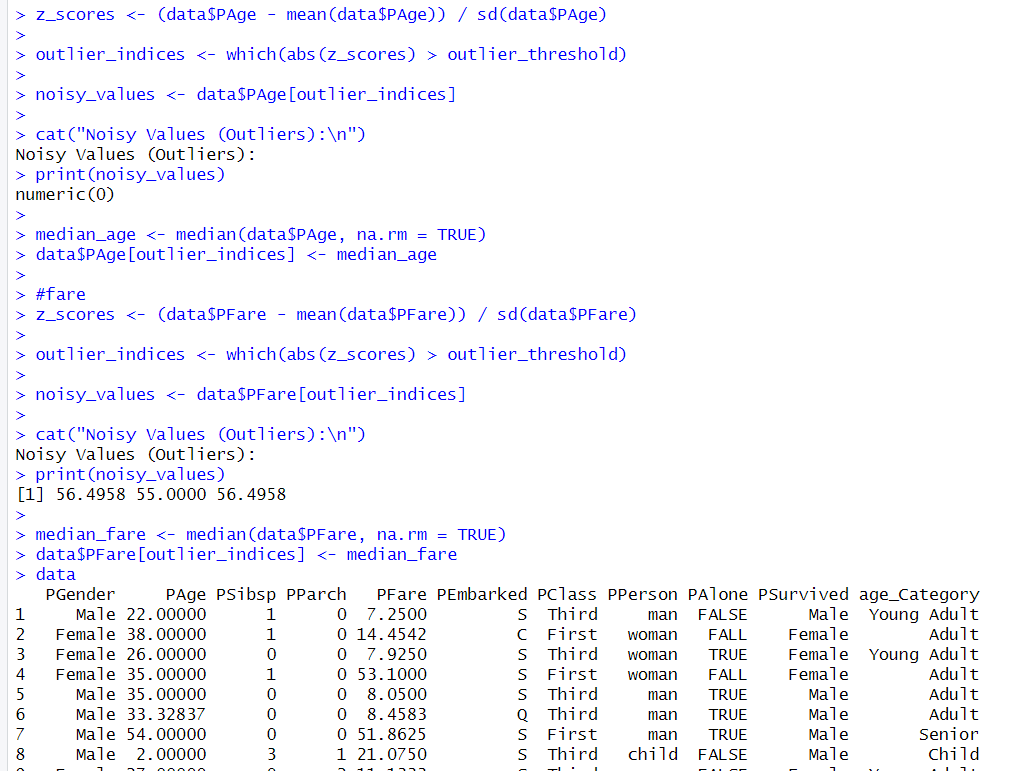
****

**description:**

First check is there any any missing values or not. Than, check is there any duplicate value or not by using unique() function. In our dataset there is no missing values for data cleaning.

**11.Noisy Value:**

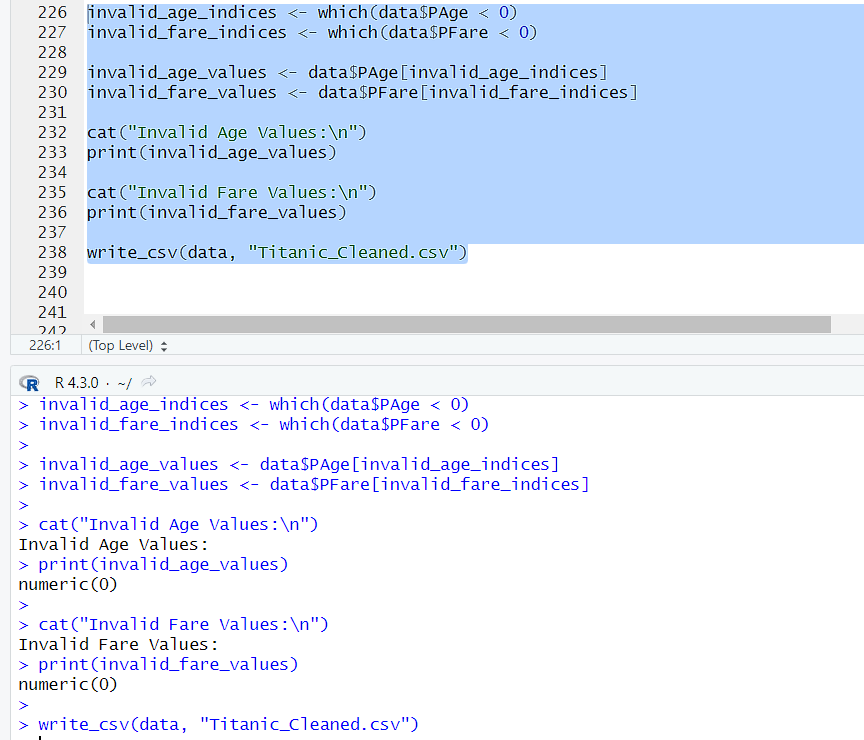
****

****

**Description:**

First check in age column is there any noisy value or not by using z\_score() function. Than,it return no noisy value found. But, in fare column there found the noisy value. Than, handle it to replace there median values of that column.

**13. Invalid value:**

****

**Description:**

Here we can see there is no invalid values so, don’t need to handle it just write the new file for new clean dataset.