**Mid Term Project**

**Introduction to Data Science**

**Faculty name: DR. ABDUS SALAM**

**Name: Most. Lilun Nahar Aurthy**

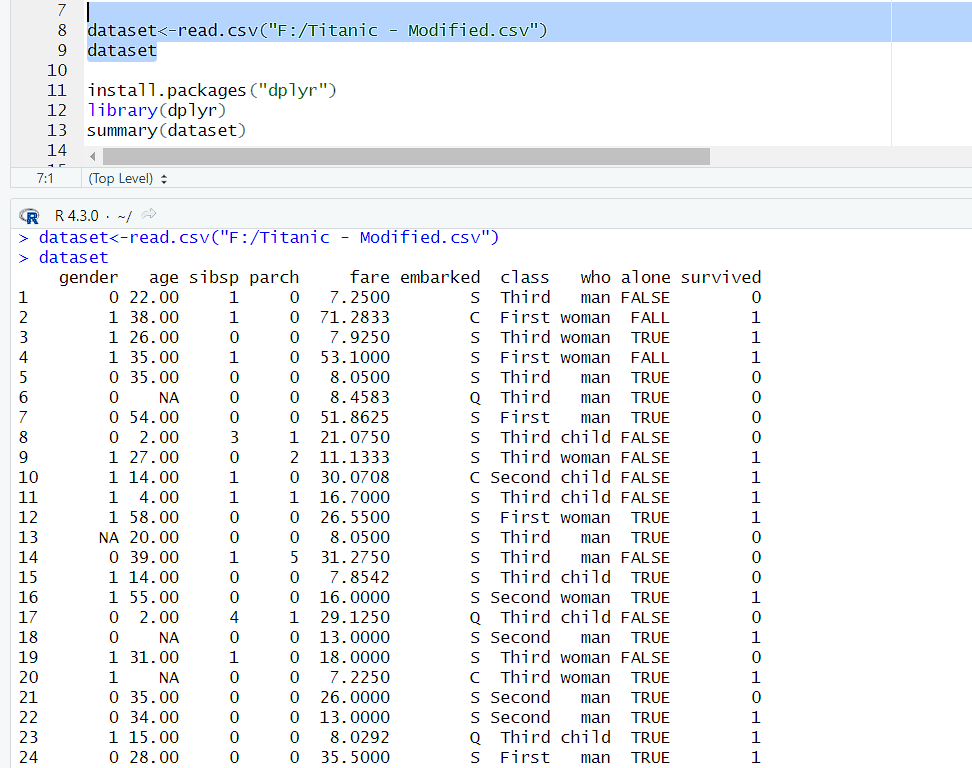
**ID: 20-43997-2**

**Section: C**

**Dataset Description:**

The Titanic dataset is a well-known dataset that provides information about the passengers aboard the RMS Titanic. The dataset captures various attributes of the passengers, including their demographic information, cabin class, ticket fare, survival status, and more. It has been widely used in data science and machine learning to explore patterns and factors that influenced the survival of passengers during the Titanic disaster.  
 The dataset includes the following columns:  
• gender: Represents the gender or sex of the individual.  
• age: Represents the age of the individual in years.  
• sibsp: Represents the number of siblings/spouses the individual had aboard the Titanic.  
• parch: Represents the number of parents/children the individual had aboard the Titanic.  
• fare: Represents the fare or ticket price paid by the individual for the Titanic voyage.  
• embarked: Represents the port of embarkation for the individual.  
• class: Represents the passenger class or ticket class of the individual.  
• who: Represents the demographic category of the individual.  
• alone: Indicates whether the individual was traveling alone or with family members.  
• survived: Represents the survival status of the individual, indicating whether they survived or did not survive the Titanic disaster.

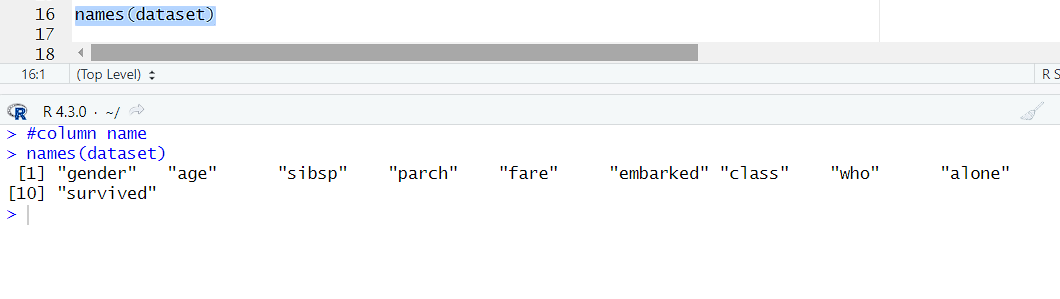
**Import CSV and print the dataset:**



**Description:**

Here , read the csv file by using read() function.

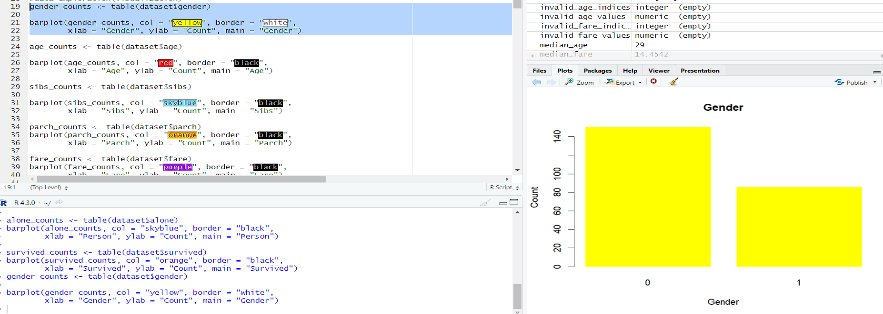
**See Column Name:**

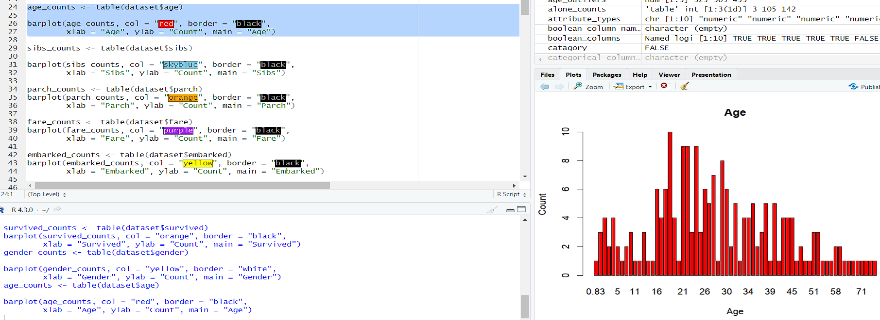
****

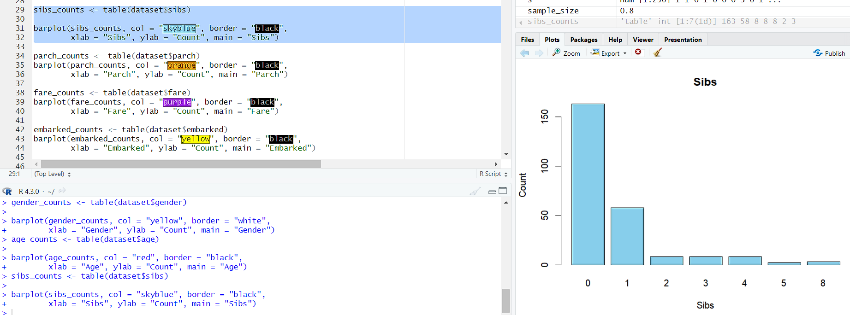
**Description:**

Here, display all the column names by using names() function.

**Exploration:**

****

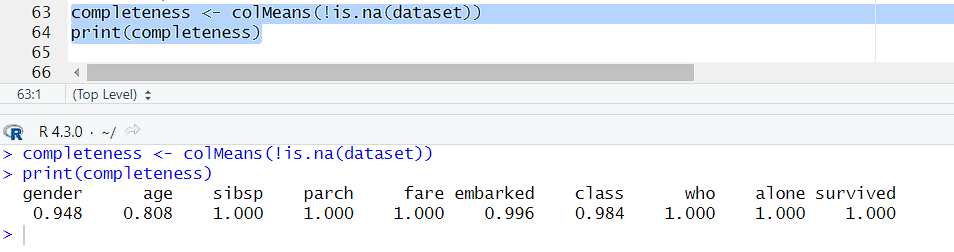




**Description:**

Data exploration approaches involve computing descriptive statistics and visualization of data. By using barplots() function I did the data visualization.

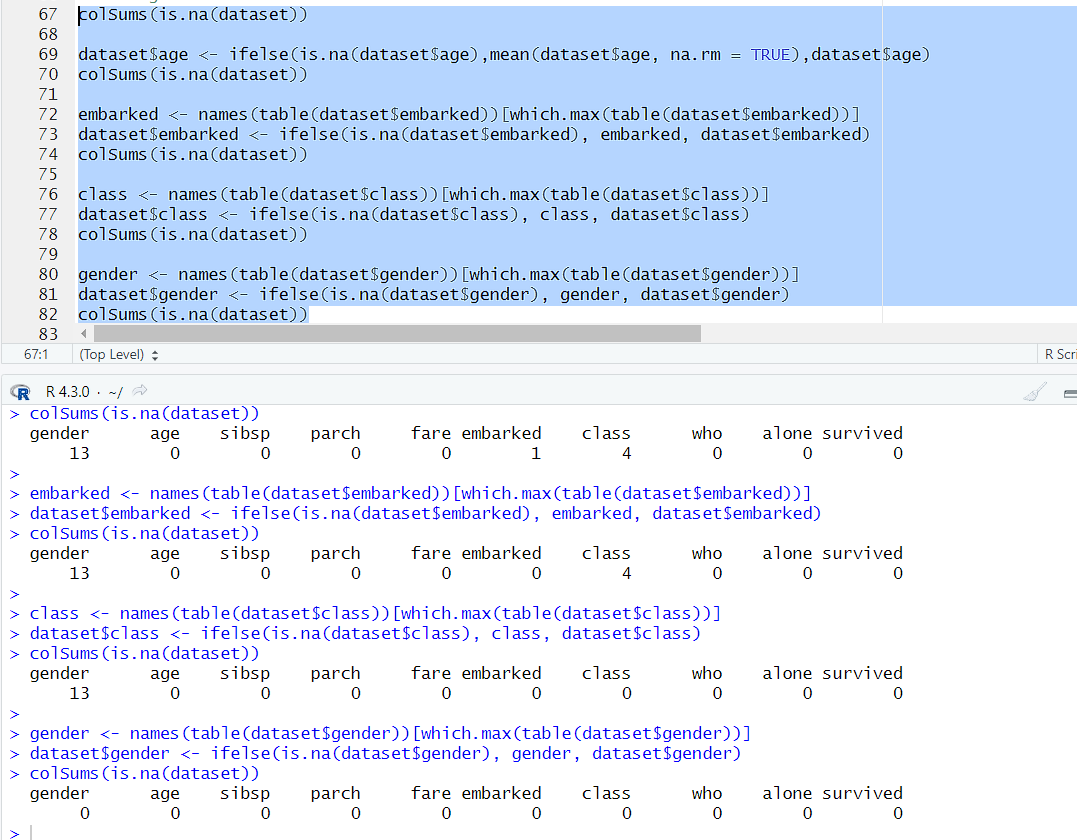
**Data-Quality(Data-Completeness):**



**Description:**

A measure which describes how whole and complete a data asset is. By using colMeans() function here checked that data is correct or not.

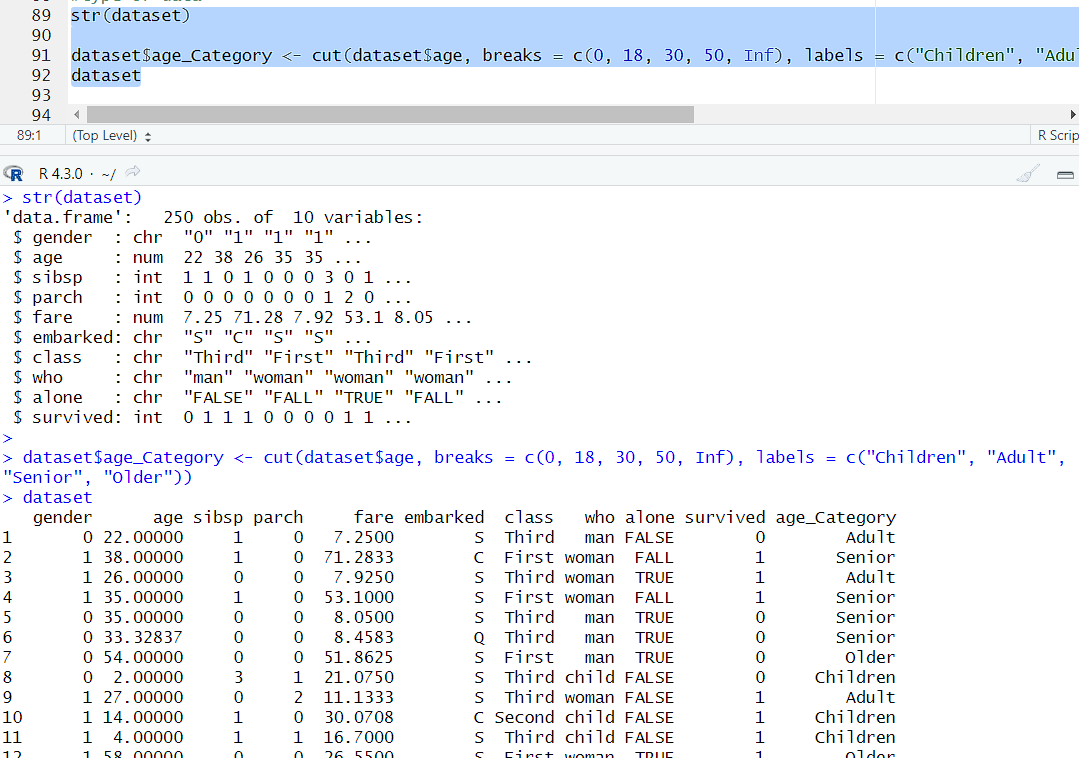
**Find Missing Values:**



Description:

One of the most common data quality issues is that some records have missing attribute values. By using colSums() function here checked the missing values.

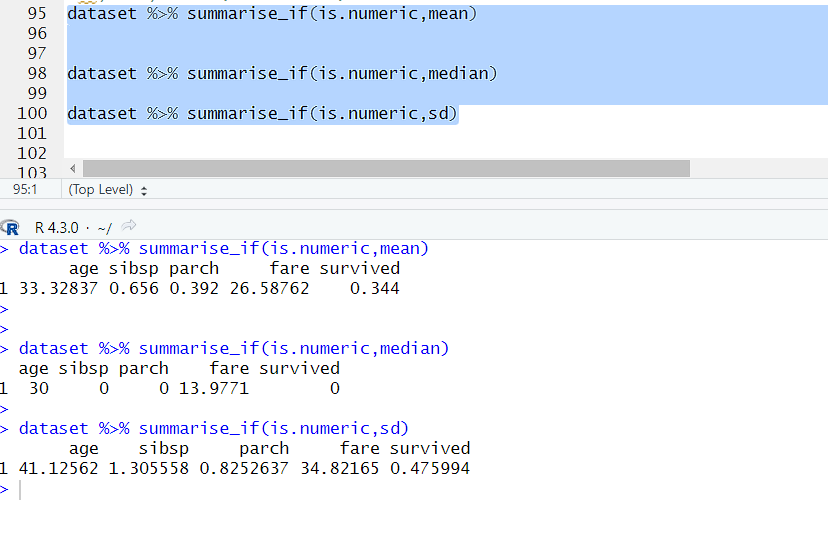
**Check Data-Types and Conversion:**

****

**Description:**

To finding the datatype I can do the further operation. By using str() function here checked the datatypes. Also, here convert data type numeric value to categorical values.

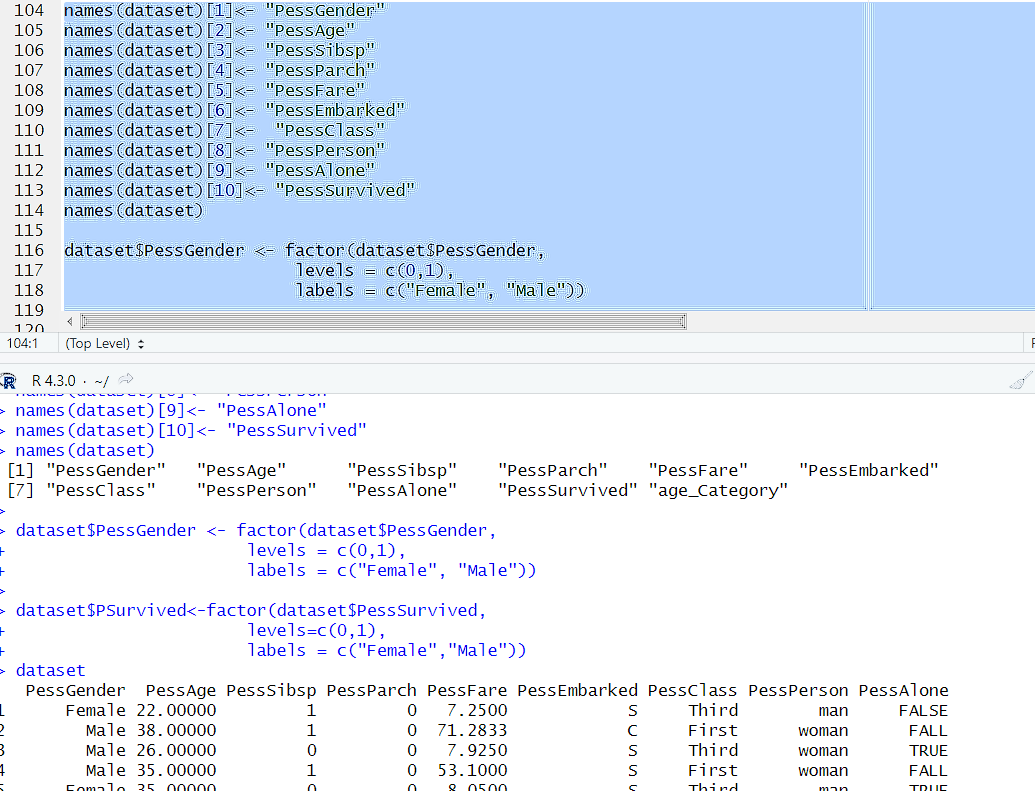
**Descriptive Statistics:**



Description:

In descriptive stastics, works with only numerical values and find the sd,mean,median. By using summarise() function here perform all the operation of all column sd,mean and median.

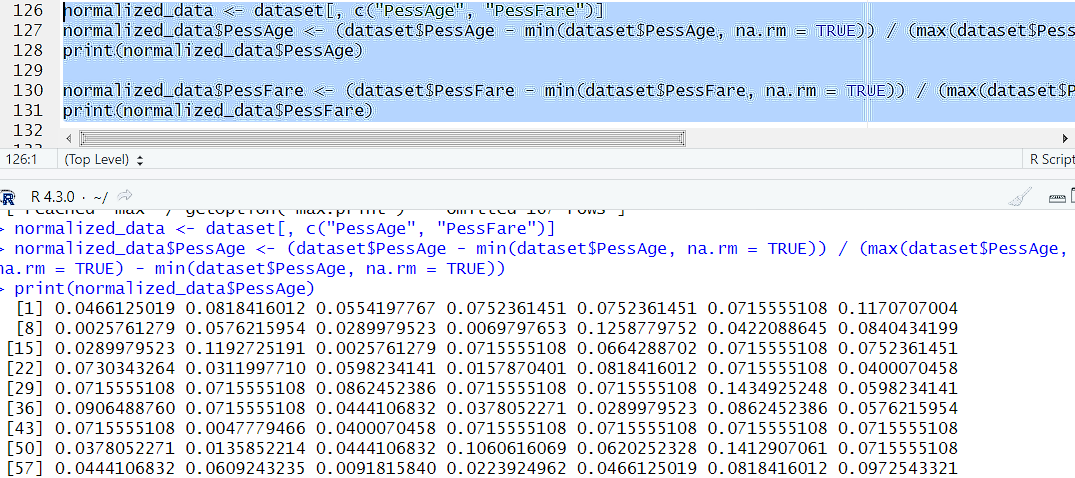
**Annotation :**



Description:

Change the columns name or change the categorical values from the dataset for this here used the annotation.

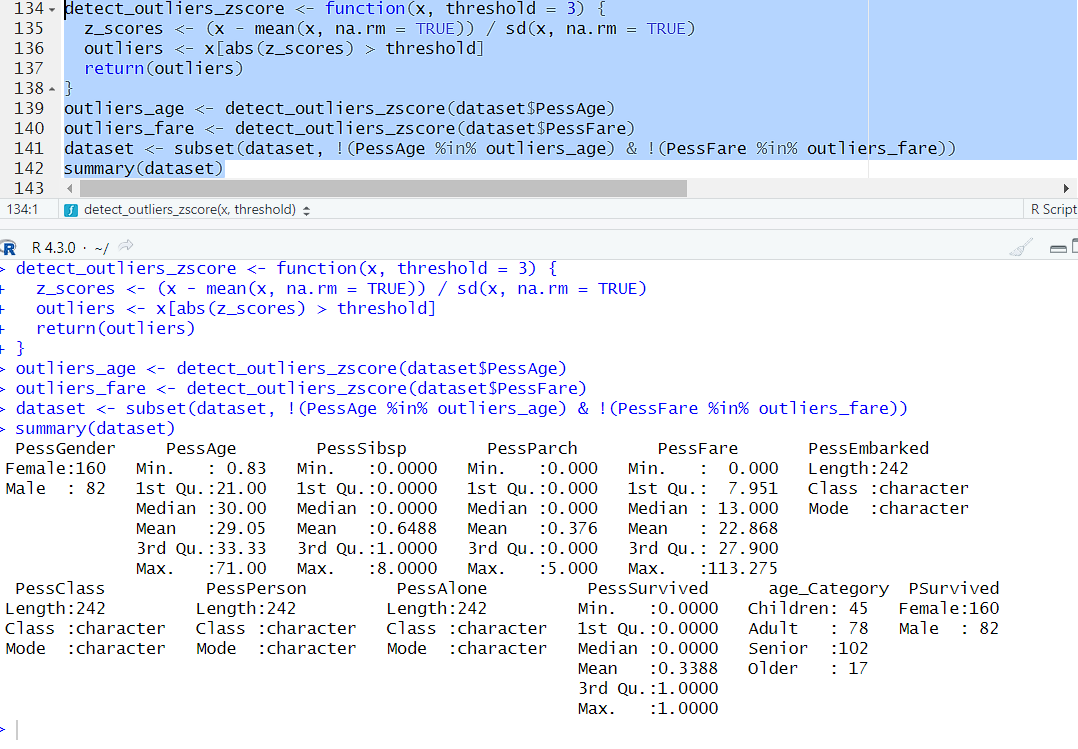
**Transformation(Normalization):**



Description:

Here did the normalization process of transformation. Normalization prevents one attribute dominating the distance results because of large values.

**Outliers:**



Description:

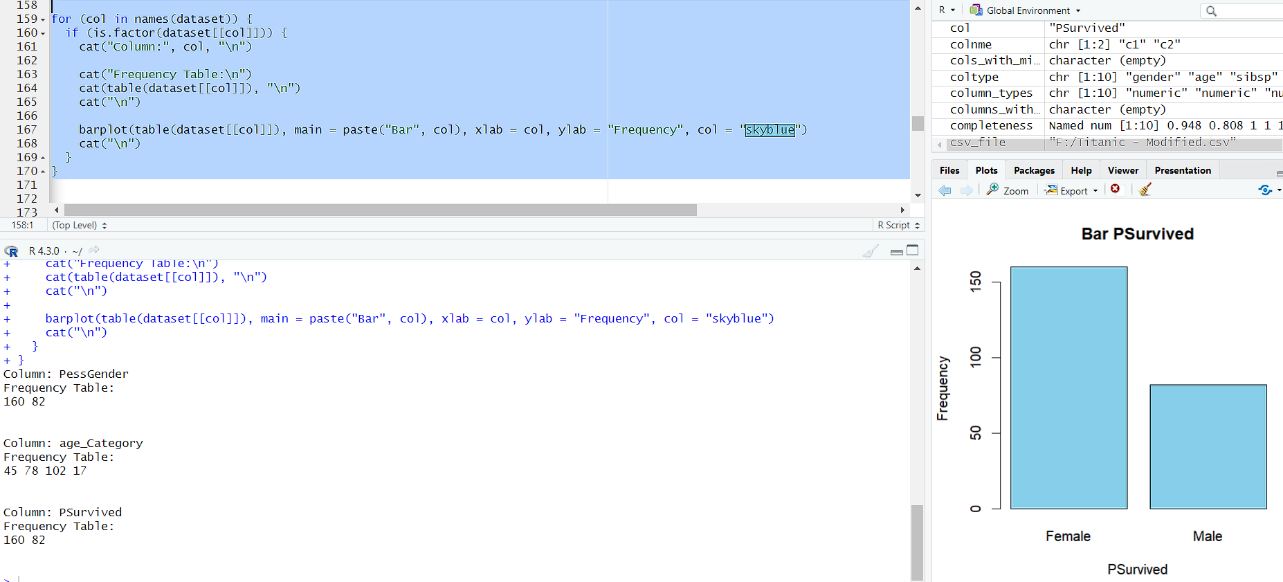
The unexpected value or the value which is different from the normal data values. By using z-scores() function here find out the outliers value. Than, with the subset() function replace the outliers value.

**Univariate:**

Numeric:



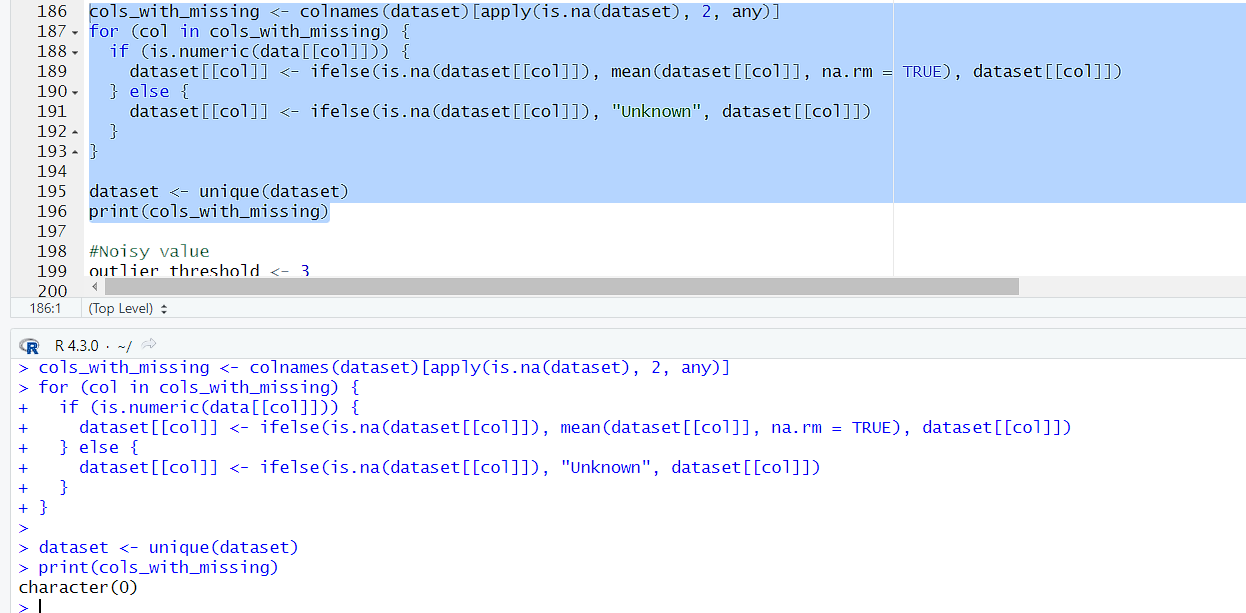
Categorical:



Description:

Univariate is a type of data which consists of observations on only a single characteristic or attribute. Here, for numerical and categorical values I did 2 different operations. For numerical values find the sd, median, mean also find the histogram by using hist() function. For, categorical values using the function table than draw the bar-chart by using barPlots() function.

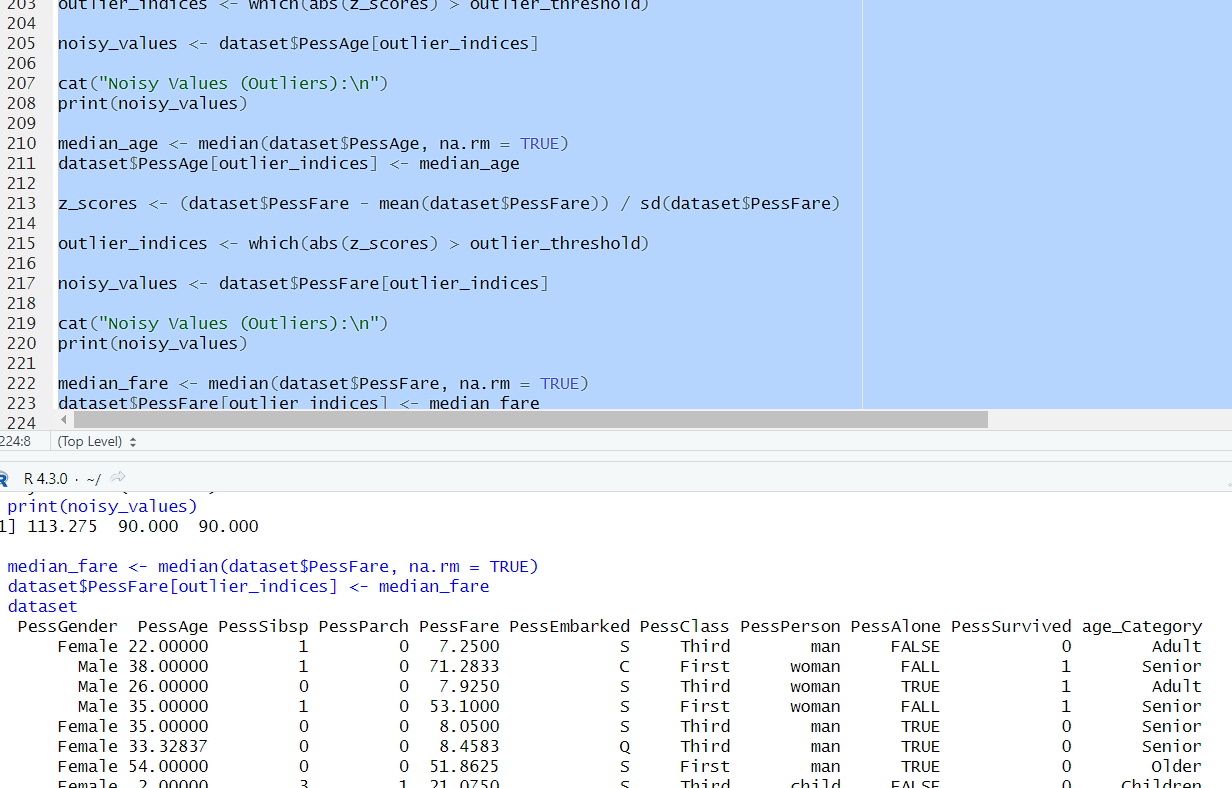
**Data-Cleaning:**



Description:

Here will check is there have any missing value or not. If missing value is available than clean the value. Also have to check is there any value which is repeated or not by applying unique() function.

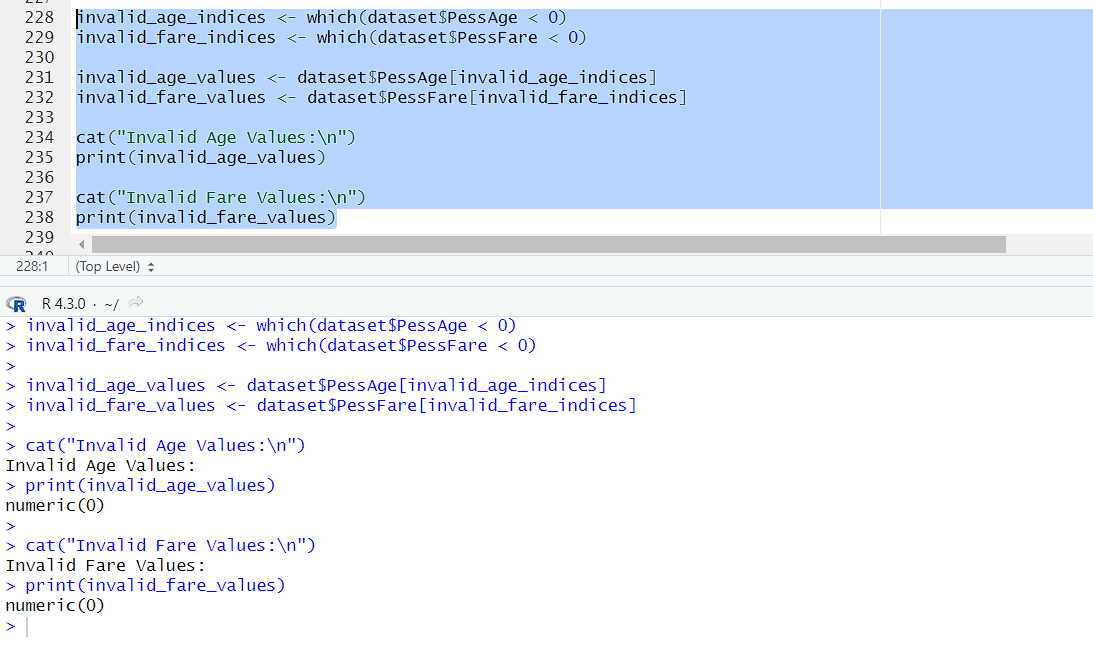
**Noisy Value:**

****

**Description:**

a noisy value to mean one that is valid for the dataset, but is incorrectly recorded. Here, we use z\_scores() function to do check is there any missing values or not.

**Invalid Attribute:**

****

**Description:**

In data science, an invalid value refers to a data point that is missing, incorrect, or does not conform to the expected format or range of values for a particular variable. Invalid values can arise due to data entry errors, data corruption, or inconsistencies in data sources. These values can negatively impact data analysis and modeling tasks, requiring cleaning and preprocessing steps to handle or remove them appropriately.

**Thank you**