American International University-Bangladesh (AIUB)

Department of Computer Science
Faculty of Science & Technology (FST)

INTRODUCTION TO DATA SCIENCE

Report

Submitted By

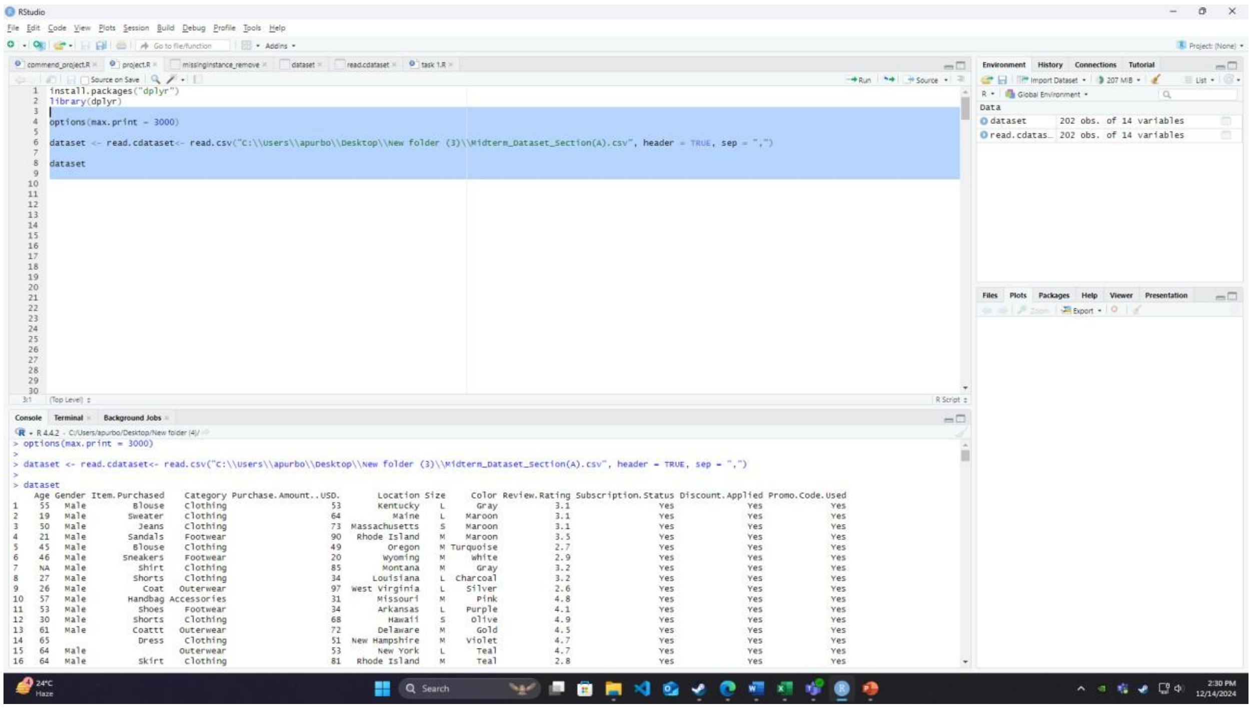
Semester: Fall_24_25	
Student Name	Student ID
Md.Shariar Hosain Sanny	**21-44677-1**

Short Description

The dataset contains detailed transactional information about individual purchases, including attributes such as Age, Gender, Item Purchased, Category, Purchase Amount (USD), Location, Size, Color, Review Rating, Subscription Status, Discount Applied, and Promo Code Used. It also includes behavioral metrics like Previous Purchases and Frequency of Purchases. The data features a mix of numerical values and boolean indicators, making it versatile for analysis. This dataset is labeled, as it includes target variables suitable for predictive analysis, and is designed for use in supervised learning tasks.

Loading and Displaying Dataset :

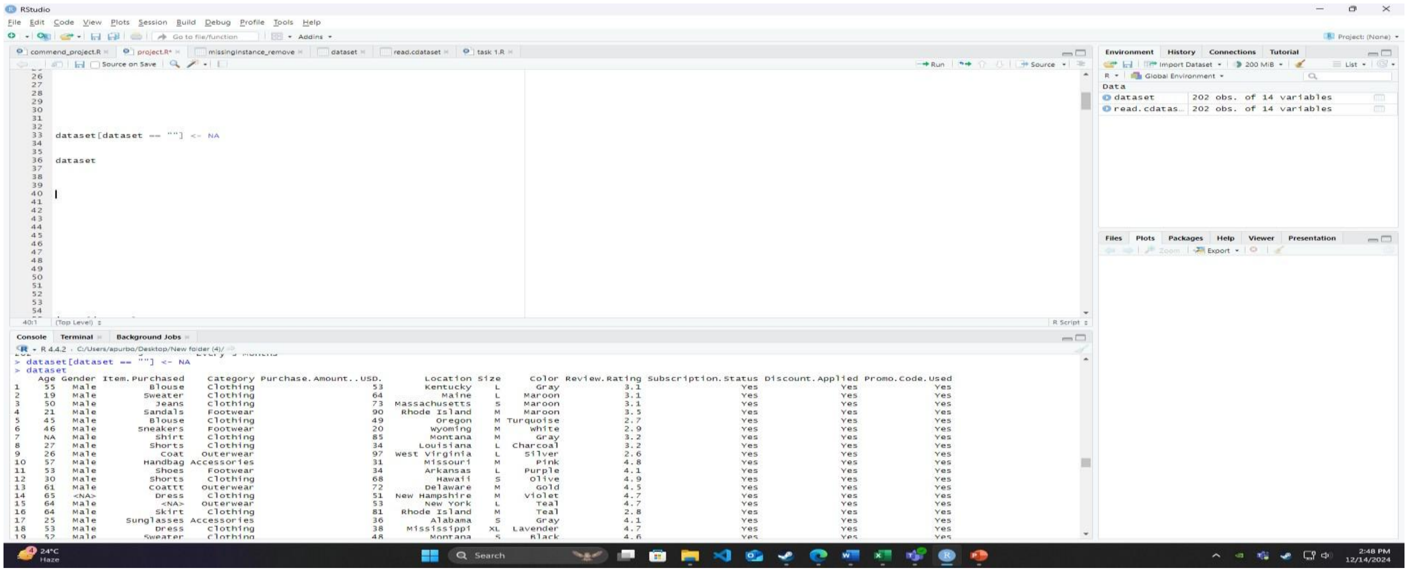
First we import the dataset using read.csv then we view the dataset by calling its variable.



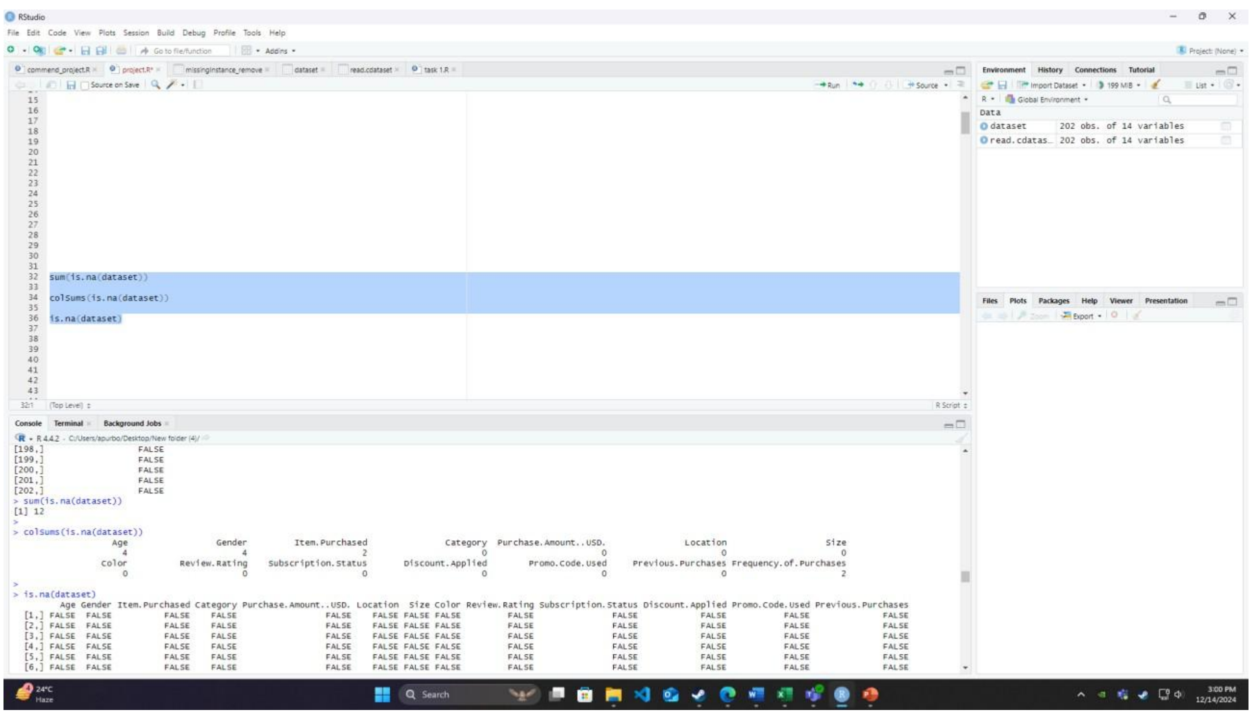
Find and handle the missing values:

```
> dataset
   Age Gender Item.Purchased    Category Purchase.Amount..USD.      Location Size     Color Review.Rating Subscription.Status Discount.Applied Promo.Code.Used
1   55   Male         Blouse    Clothing                    53      Kentucky    L      Gray           3.1                 Yes              Yes             Yes
2   19   Male        Sweater    Clothing                    64         Maine    L    Maroon           3.1                 Yes              Yes             Yes
3   50   Male          Jeans    Clothing                    73 Massachusetts    S    Maroon           3.1                 Yes              Yes             Yes
4   21   Male        Sandals    Footwear                    90  Rhode Island    M    Maroon           3.5                 Yes              Yes             Yes
5   45   Male         Blouse    Clothing                    49        Oregon    M Turquoise           2.7                 Yes              Yes             Yes
6   46   Male       Sneakers    Footwear                    20       Wyoming    M     White           2.9                 Yes              Yes             Yes
7   NA   Male          Shirt    Clothing                    85       Montana    M      Gray           3.2                 Yes              Yes             Yes
8   27   Male         Shorts    Clothing                    34     Louisiana    L  Charcoal           3.2                 Yes              Yes             Yes
9   26   Male           Coat   Outerwear                    97 West Virginia    L    Silver           2.6                 Yes              Yes             Yes
10  57   Male        Handbag Accessories                    31      Missouri    M      Pink           4.8                 Yes              Yes             Yes
11  53   Male          Shoes    Footwear                    34      Arkansas    L    Purple           4.1                 Yes              Yes             Yes
12  30   Male         Shorts    Clothing                    68        Hawaii    S     Olive           4.9                 Yes              Yes             Yes
13  61   Male         Coattt   Outerwear                    72      Delaware    M      Gold           4.5                 Yes              Yes             Yes
14  65                 Dress    Clothing                    51 New Hampshire    M    Violet           4.7                 Yes              Yes             Yes
15  64   Male                  Outerwear                    53      New York    L      Teal           4.7                 Yes              Yes             Yes
16  64   Male          Skirt    Clothing                    81  Rhode Island    M      Teal           2.8                 Yes              Yes             Yes
```

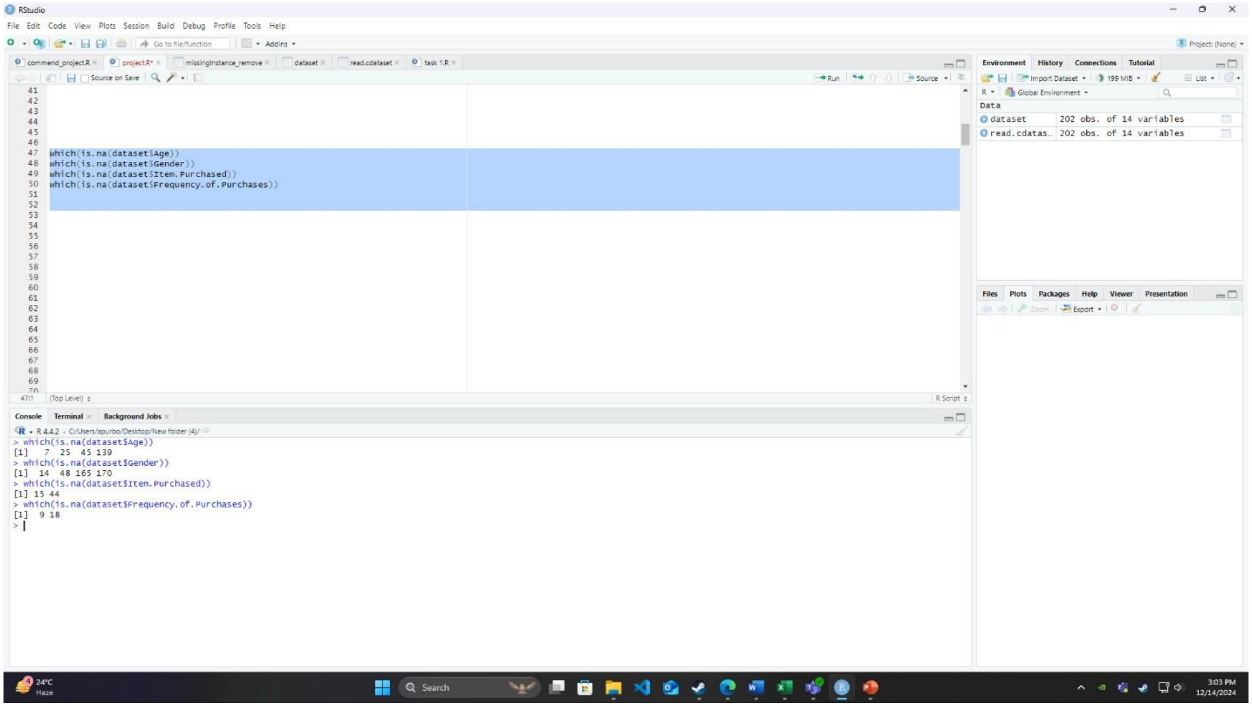
Some of the data are empty so we must first convert them to NA. We convert empty data using dataset[dataset == ""] <- NA



Then we use is.na(dataset) to find the missing values in the dataset sum(is.na(dataset)) to get the total count of missing values and colSums(is.na(dataset)) to count the missing values in each column.



We find the positions of missing values

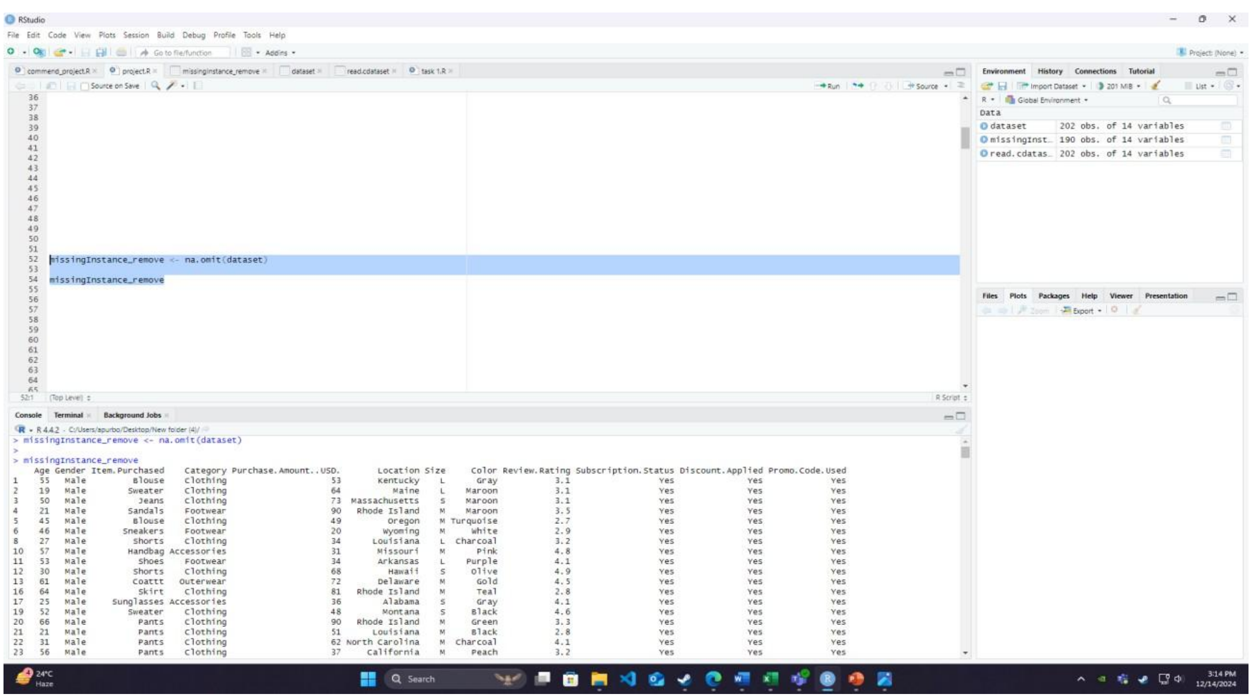


We use two methods handle the missing values

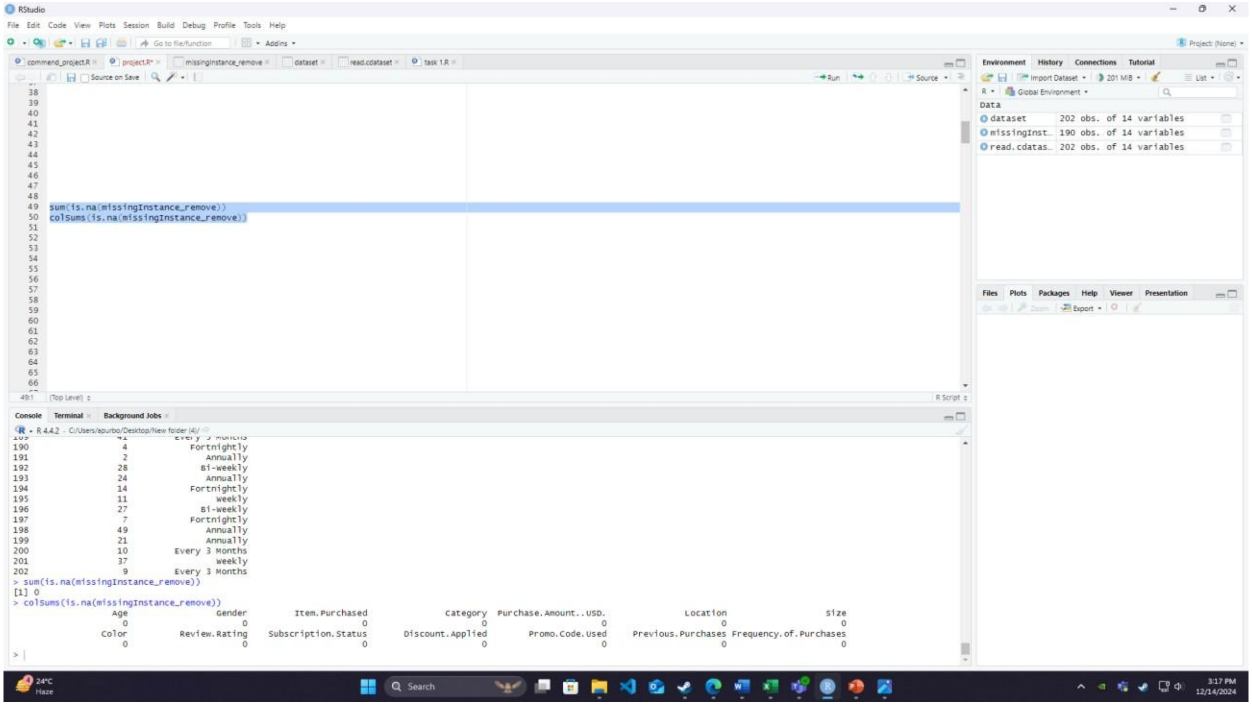
1. **Discard Instances**
2. **Replace by Most Frequent/Average Value**

Discard Instances:

Here we handle missing value by remove it

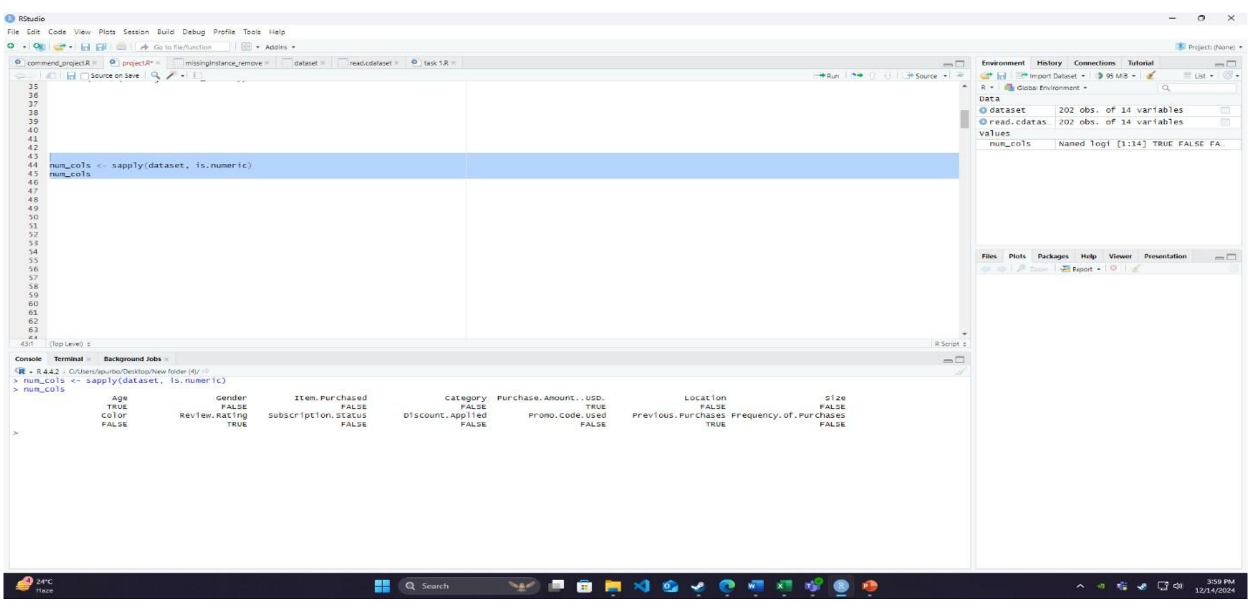


No missing values found.

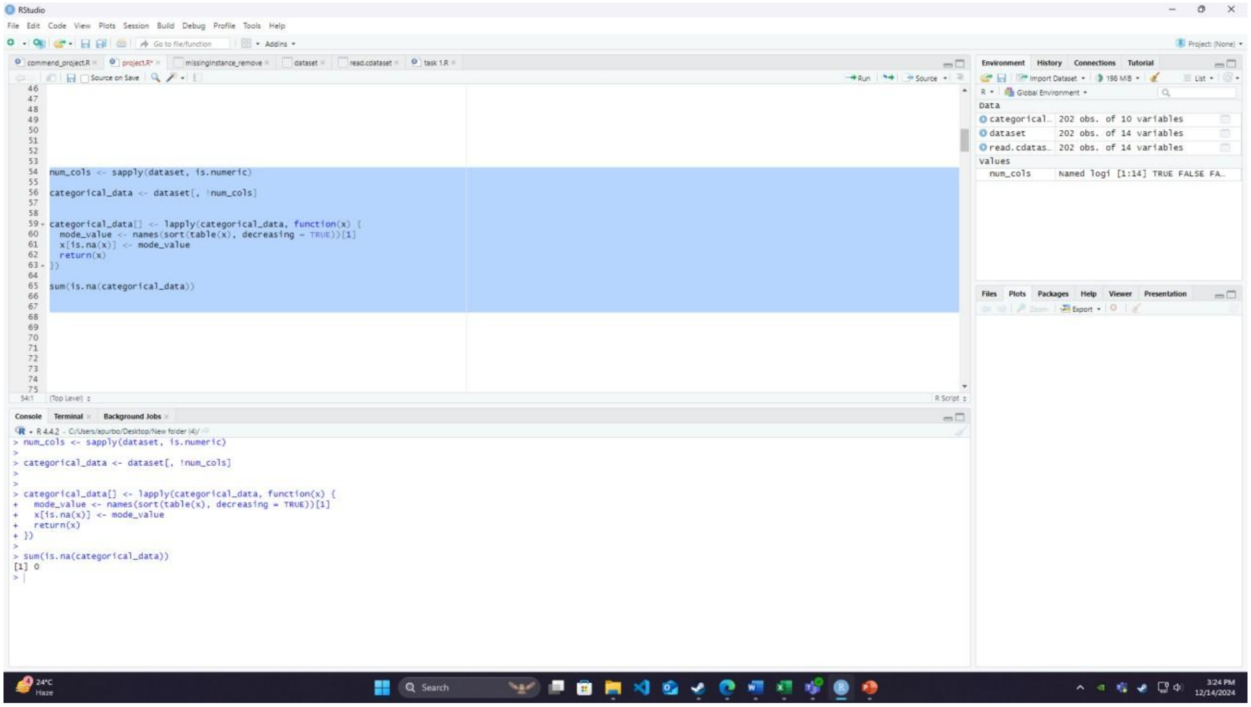


Replace by Most Frequent/Average Value

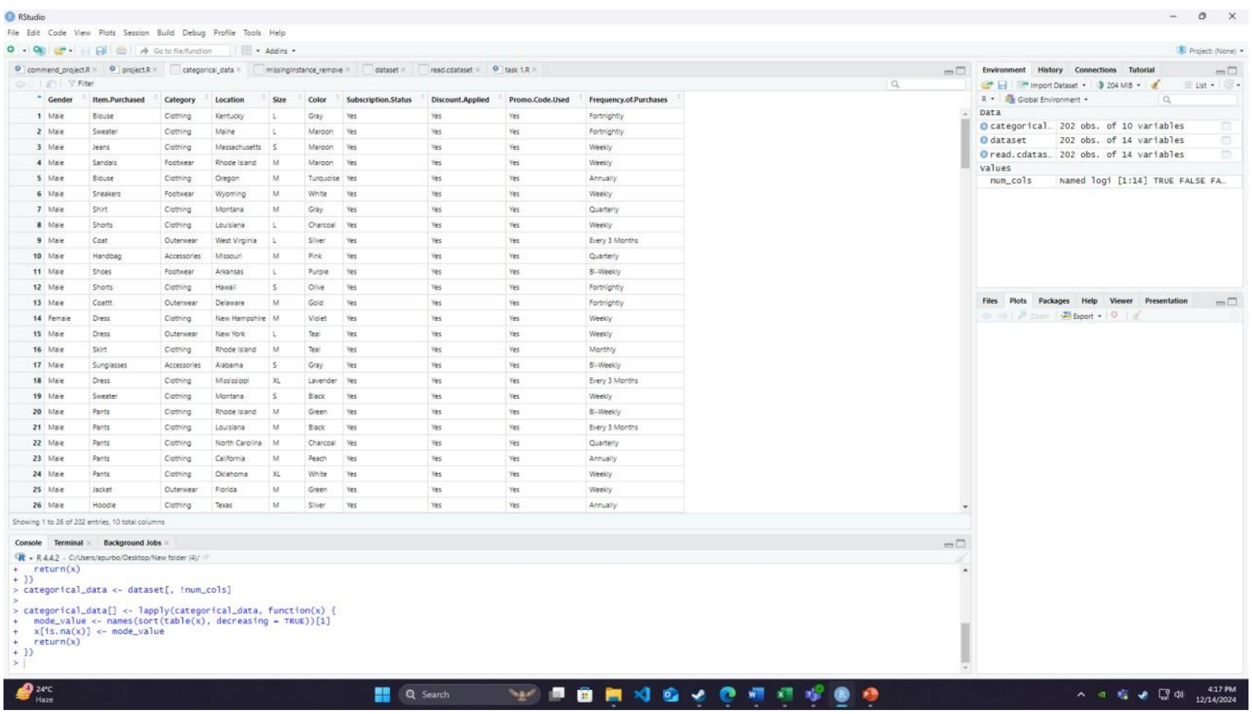
First, we detect categorical and numerical attributes. If the result is true the attribute is numerical, if False it is categorical



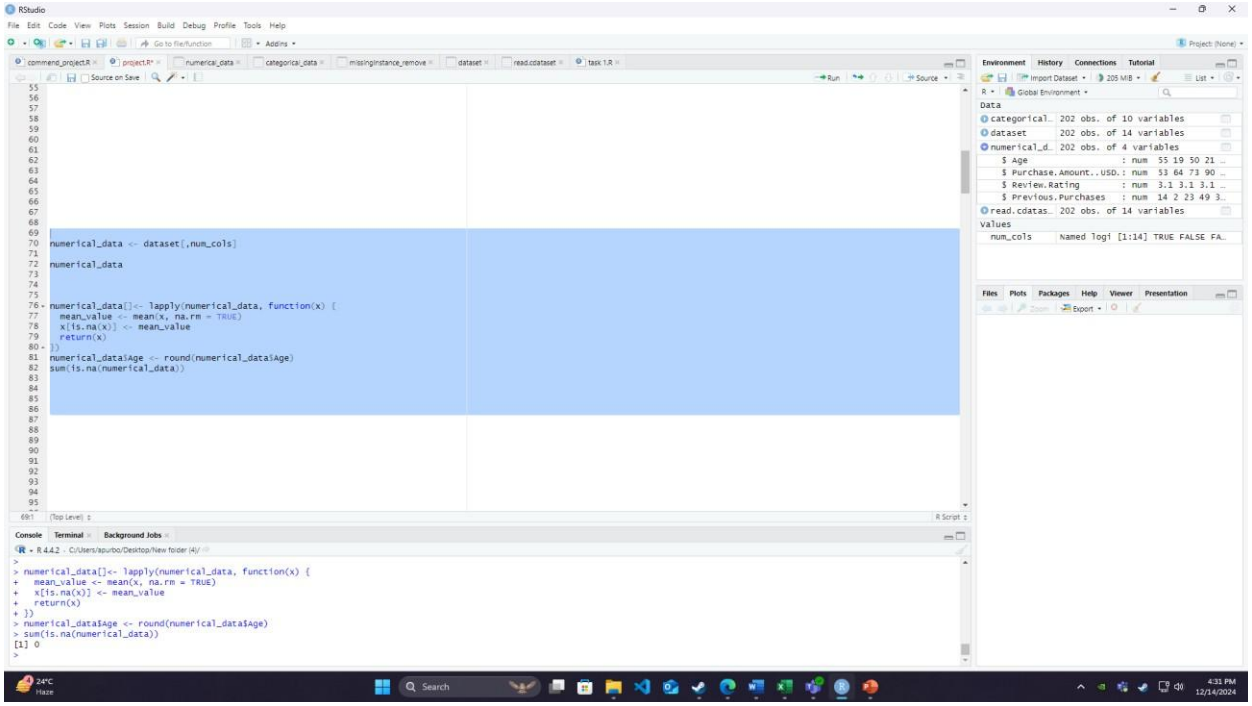
In the case of categorical attributes, we handle missing values by replacing them with the most frequent value (mode) in each instance.



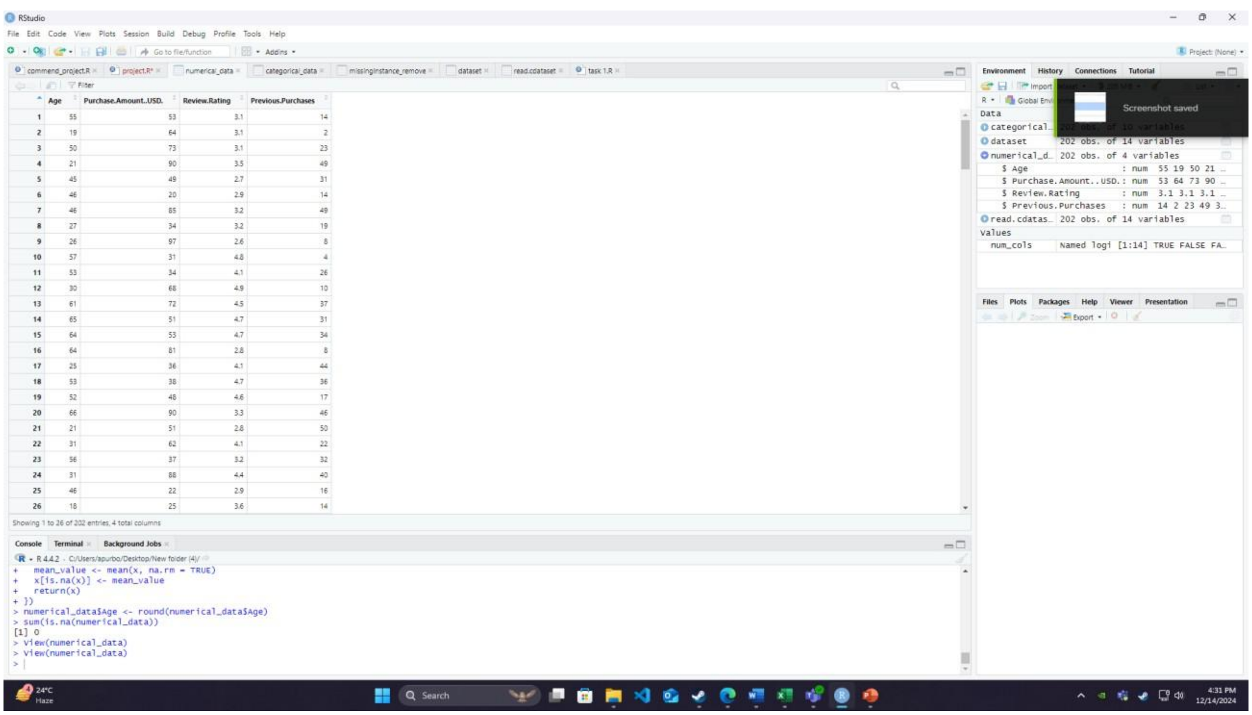
Replacing them with the most frequent value (mode) in each instance.



In the case of numerical attributes, we handle missing values by replacing them with the most average value (mean) in each instance.

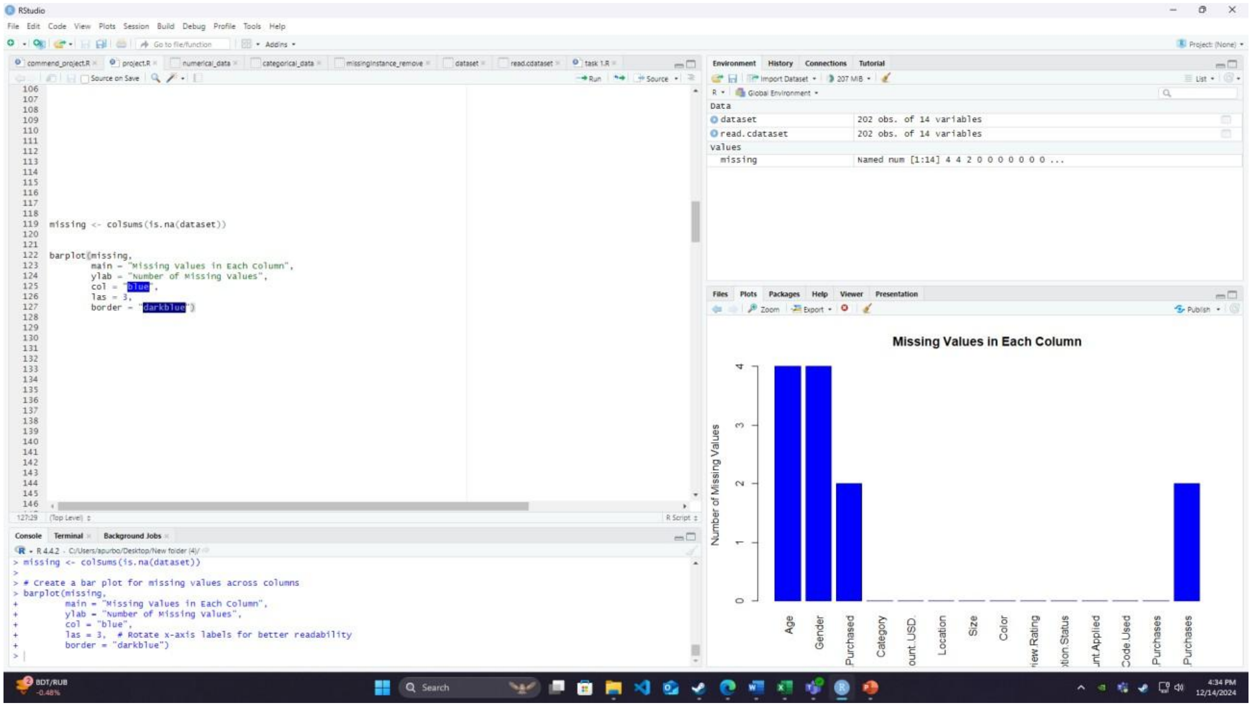


Replacing them with the most average value (mean) in each instance.



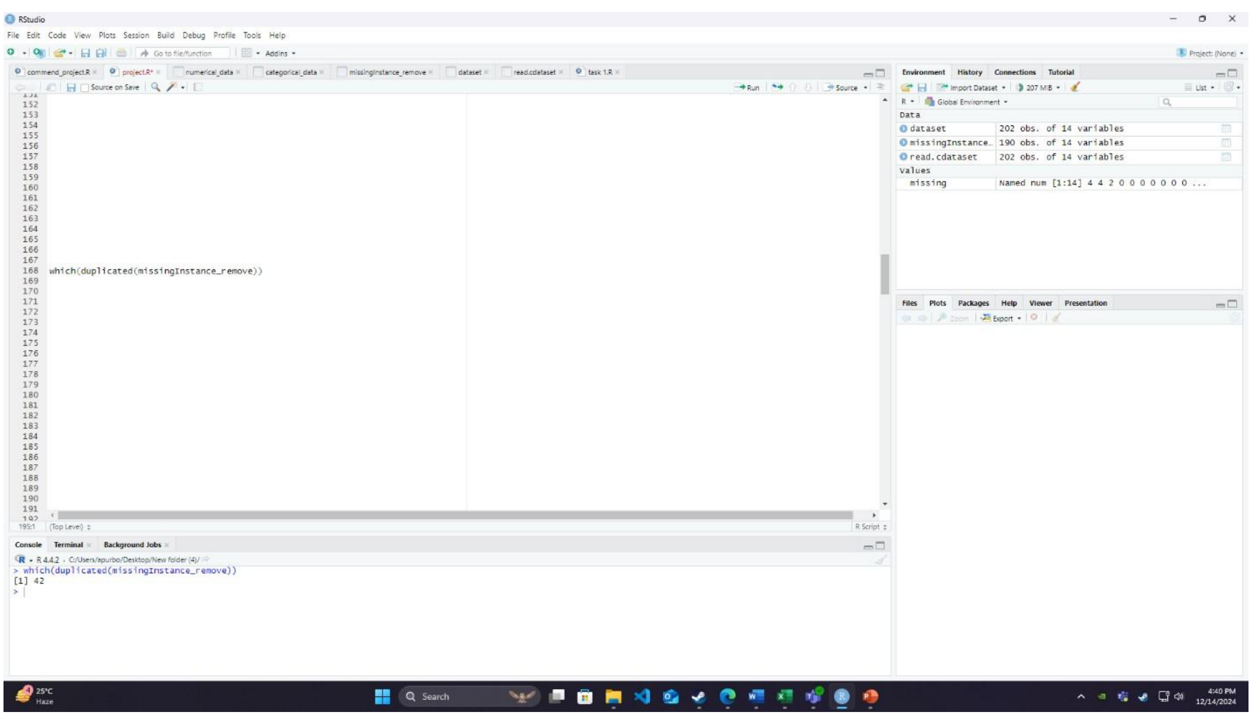
Missing value in graph

We use barplot for the graph

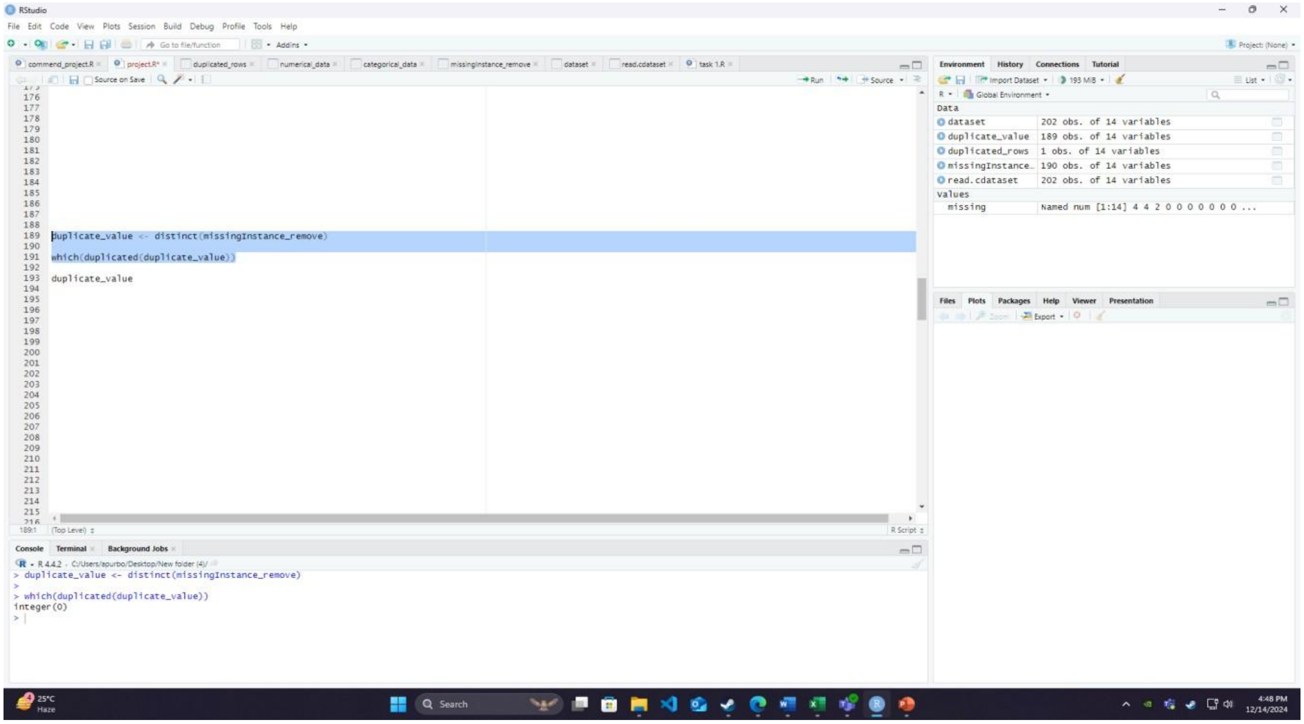


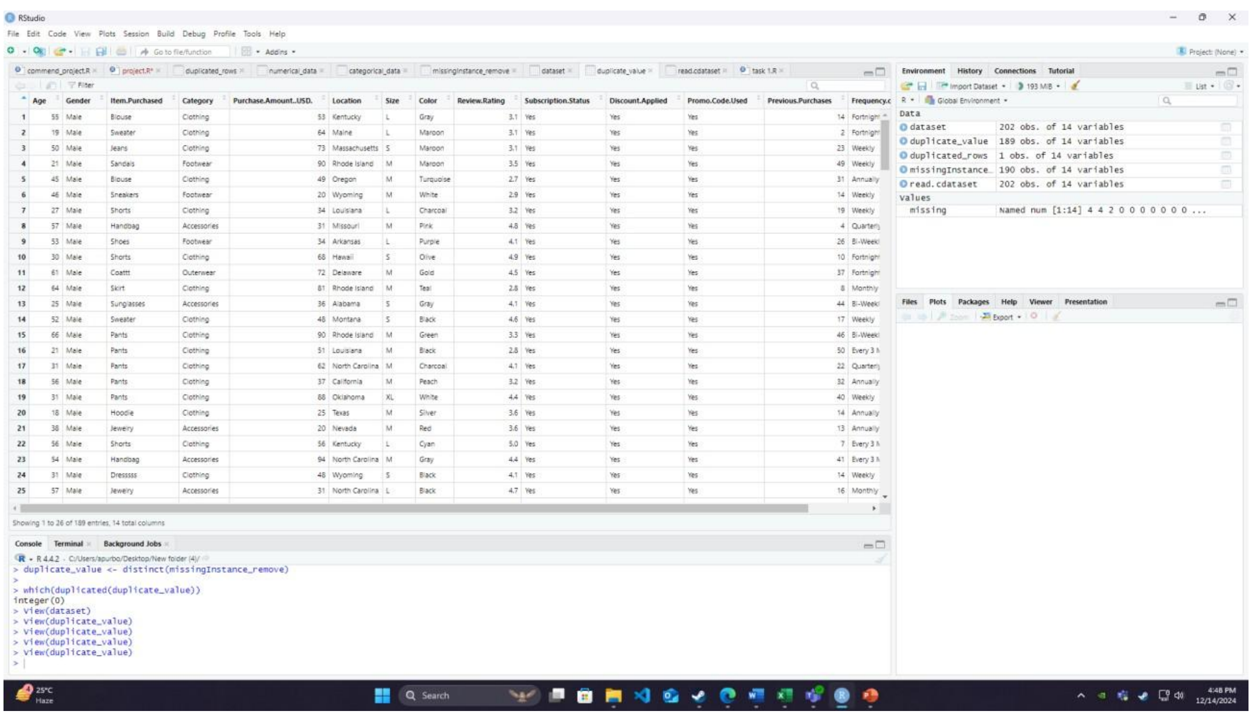
Find and remove duplicate values.

Find and identify the positions duplicate value.



Remove duplicate and show the output



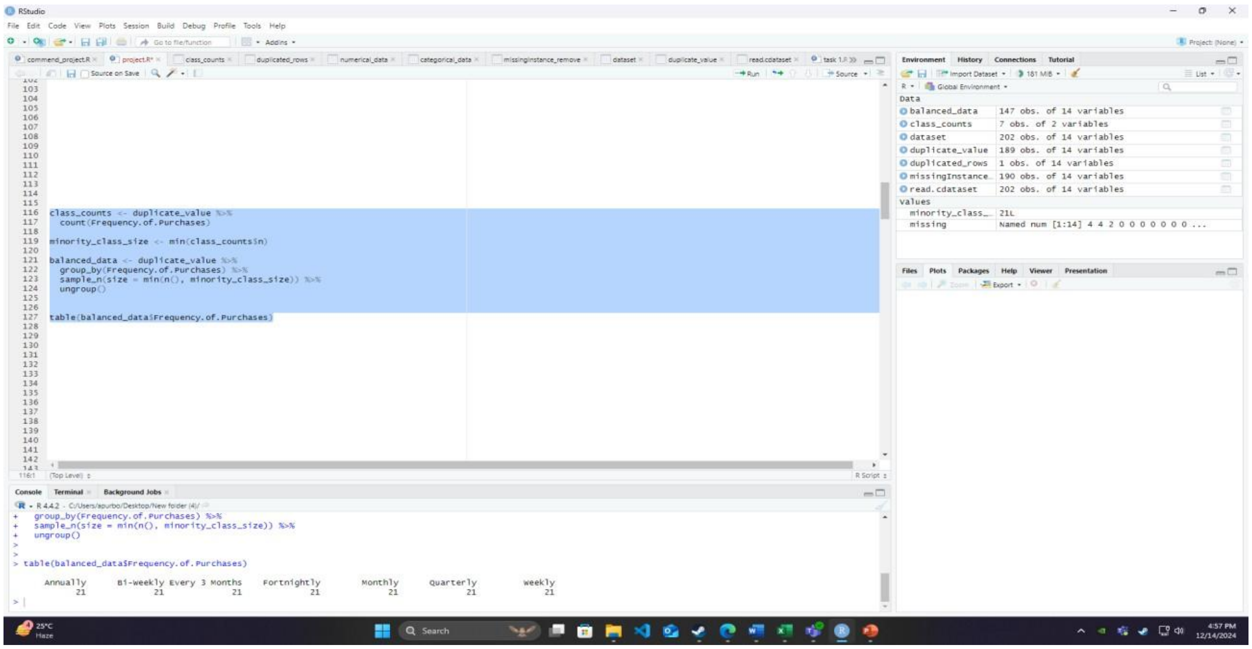


Convert the imbalanced data set into the balanced data set

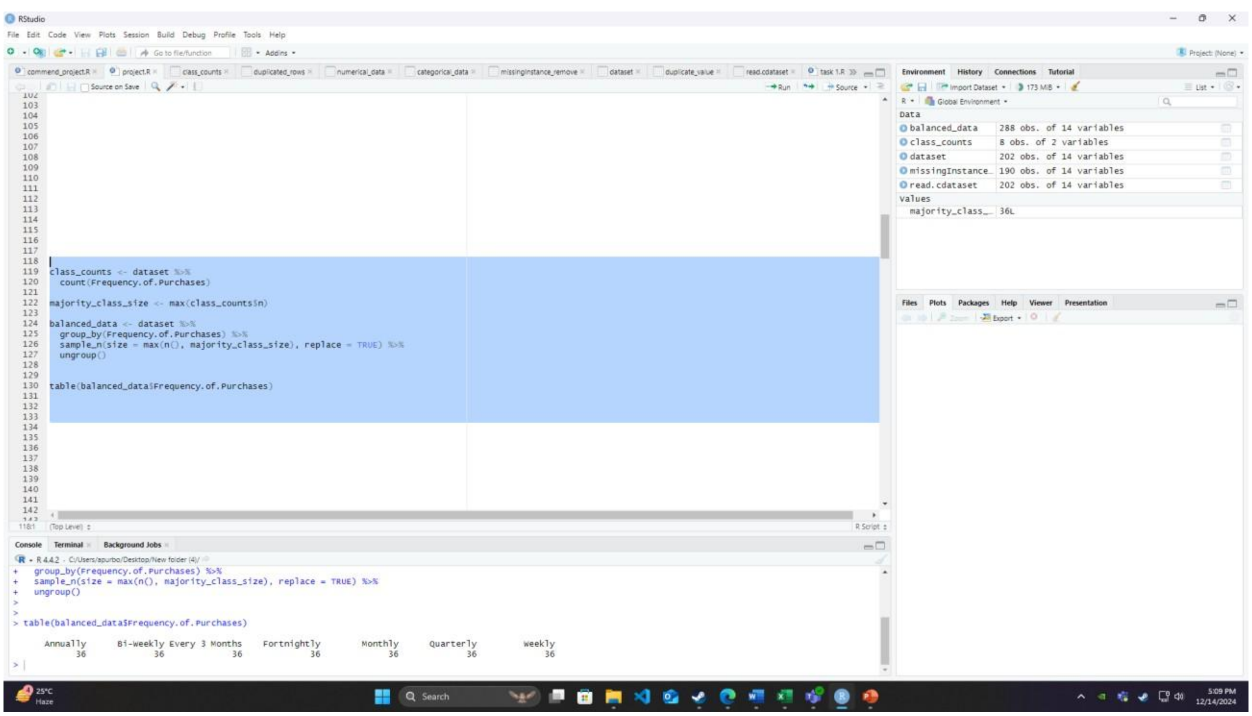
We have applied two method for Converting imbalanced to balanced dataset

1. Undersampling
2. Oversampling

Undersampling

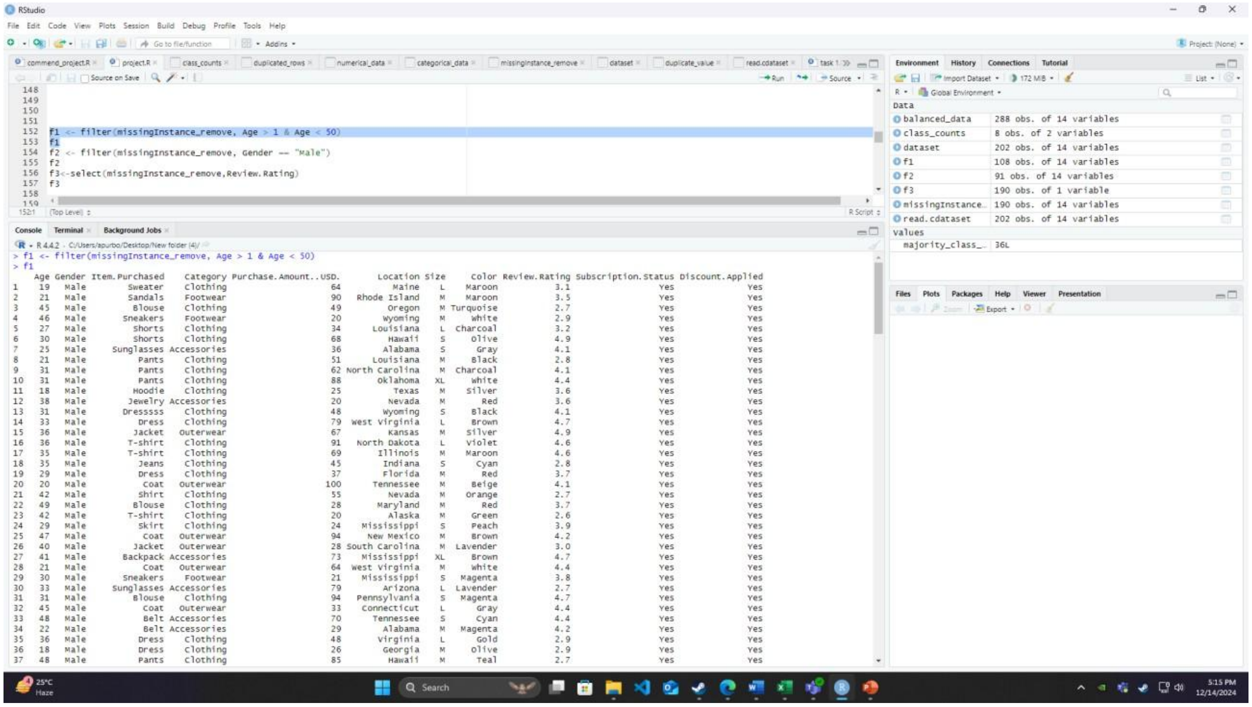


Oversampling

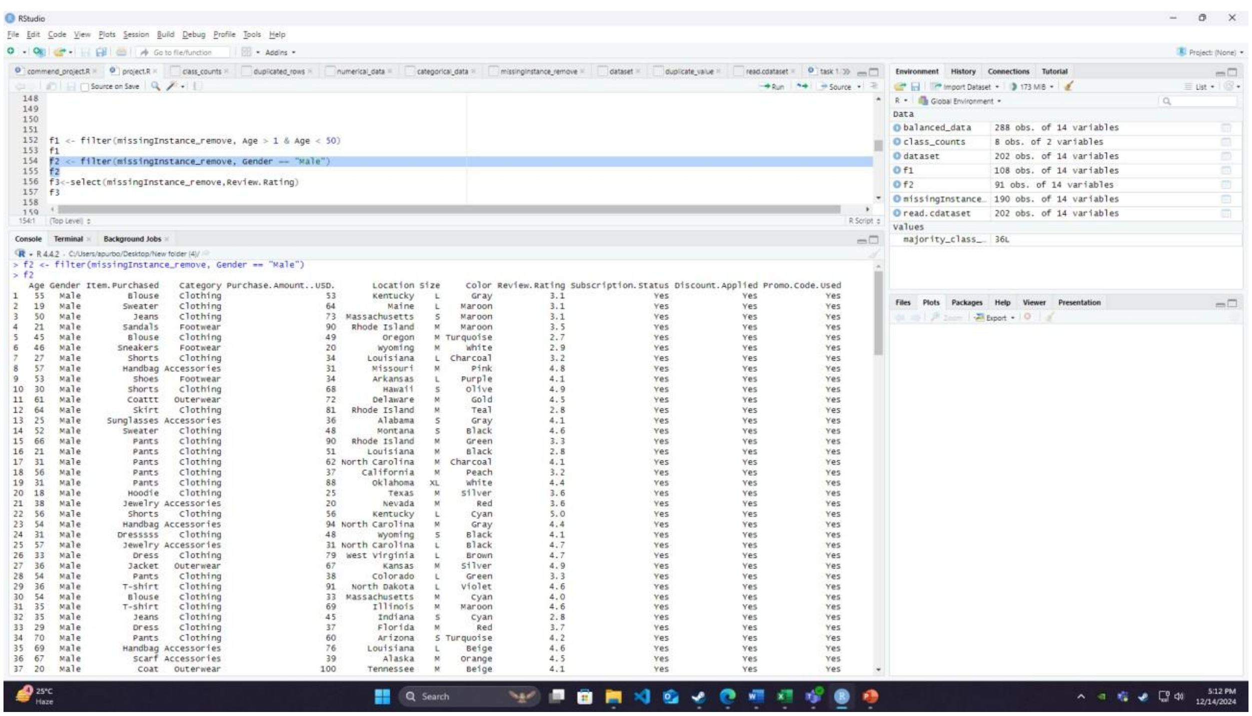


filtering methods

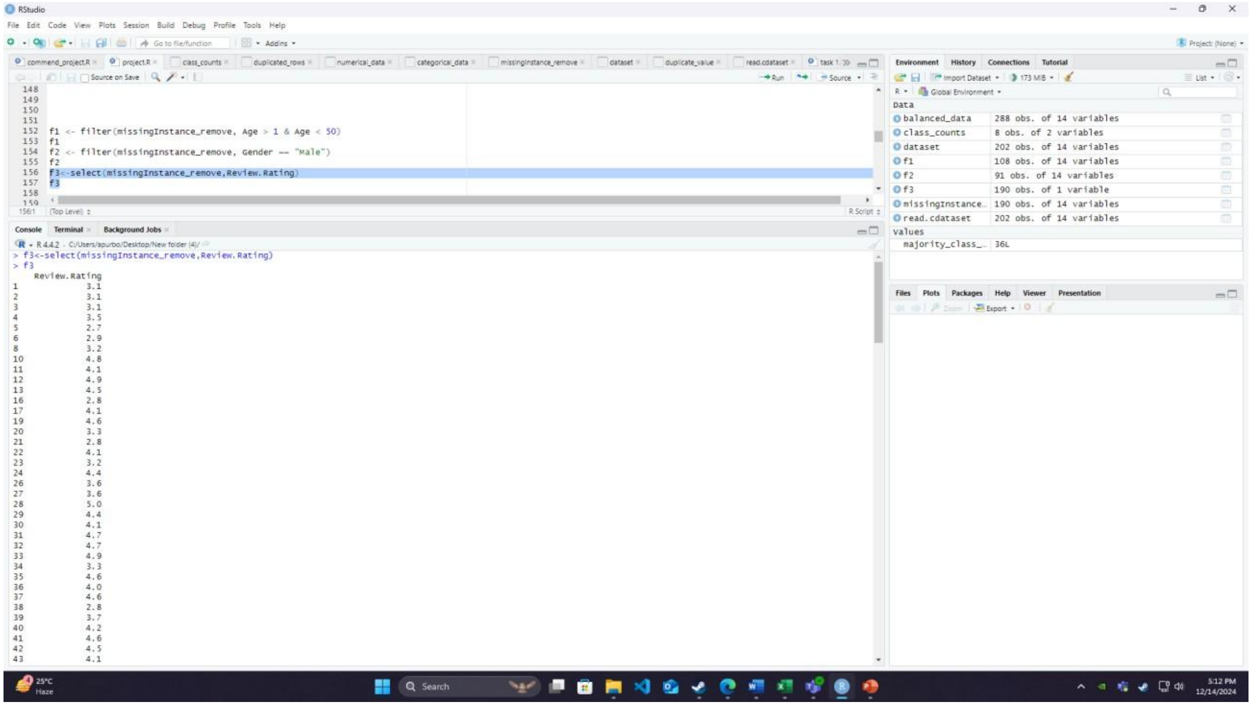
Filter age between 1 to 50



Filter Gender male only

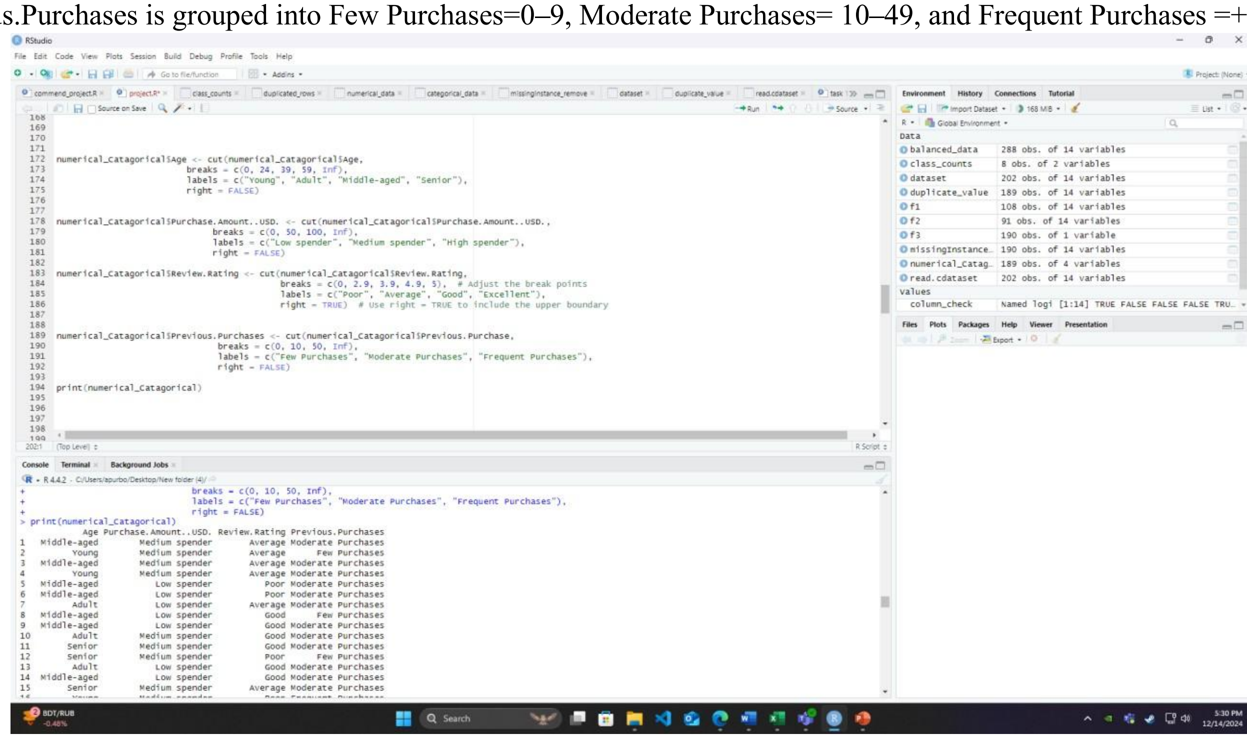


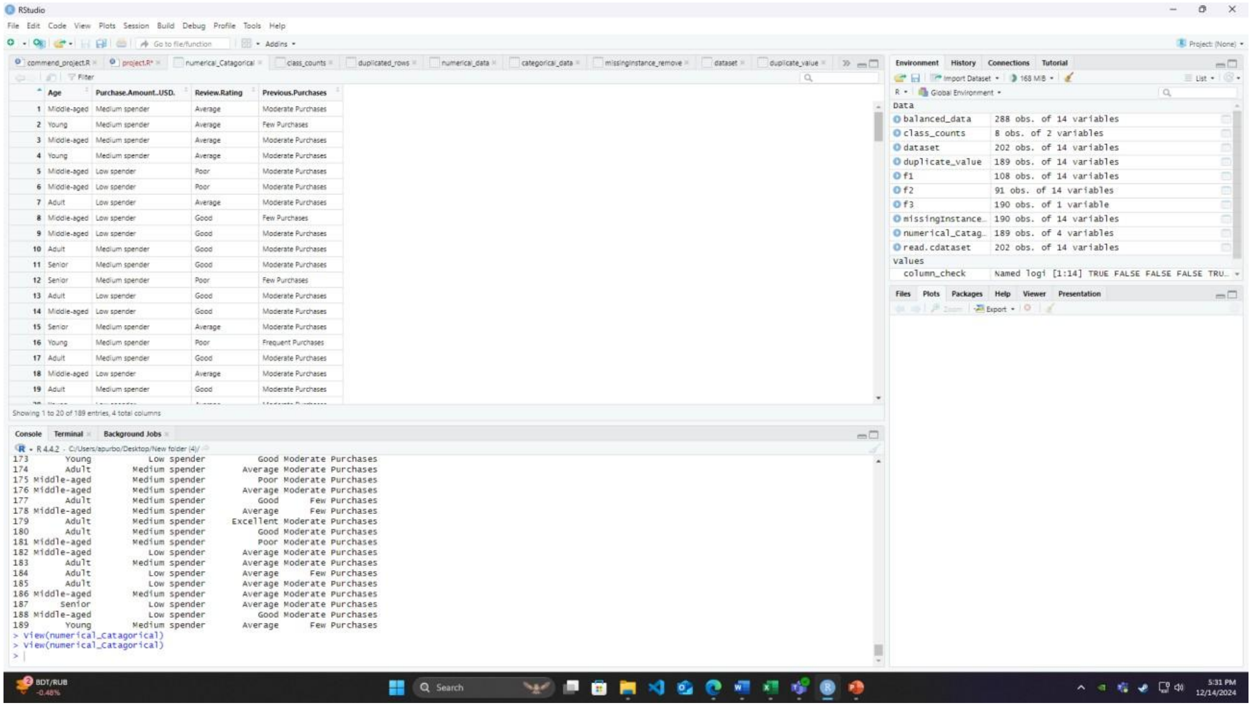
we select only the review.rating column from the dataset



Convert attributes from numeric to categorical

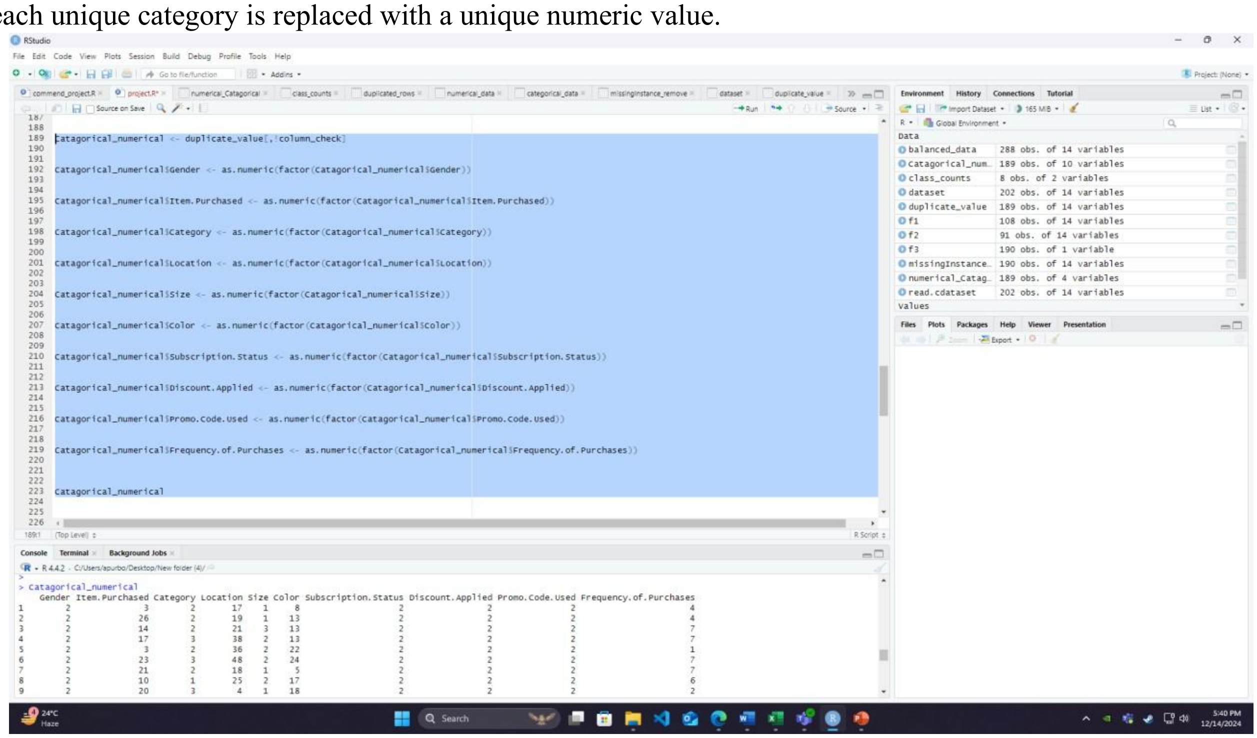
We categorizes numerical columns into groups. Age is categorized as Young =0–23, Adult =24–38, Middle-aged =39–58, and Senior =+59. Purchase.Amount..USD. is grouped into Low spender =0–49, Medium spender=50–99, and High spender=+100. Review.Rating is labeled as Poor= 0–2.9, Average =3–3.9, Good =4–4.9, and Excellent= 5.
Previous.Purchases is grouped into Few Purchases=0–9, Moderate Purchases= 10–49, and Frequent Purchases =+50.

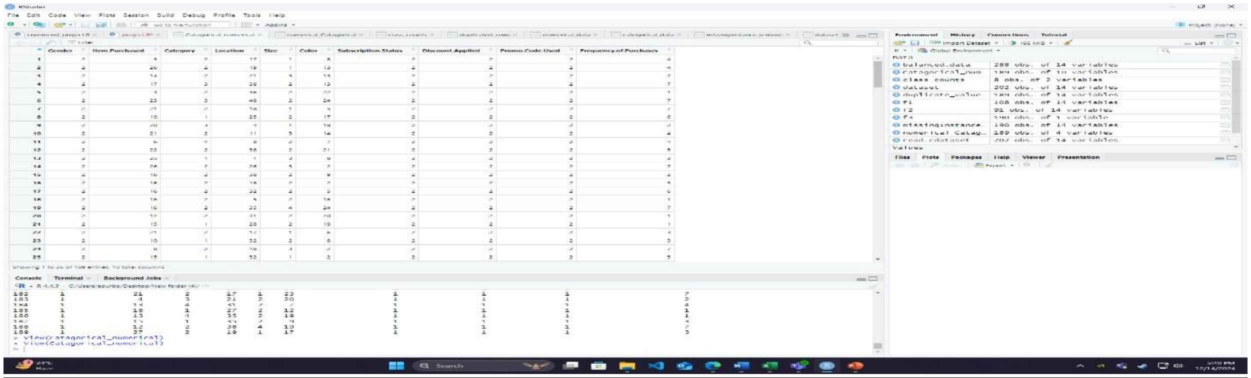




convert attributes from categorical to numeric

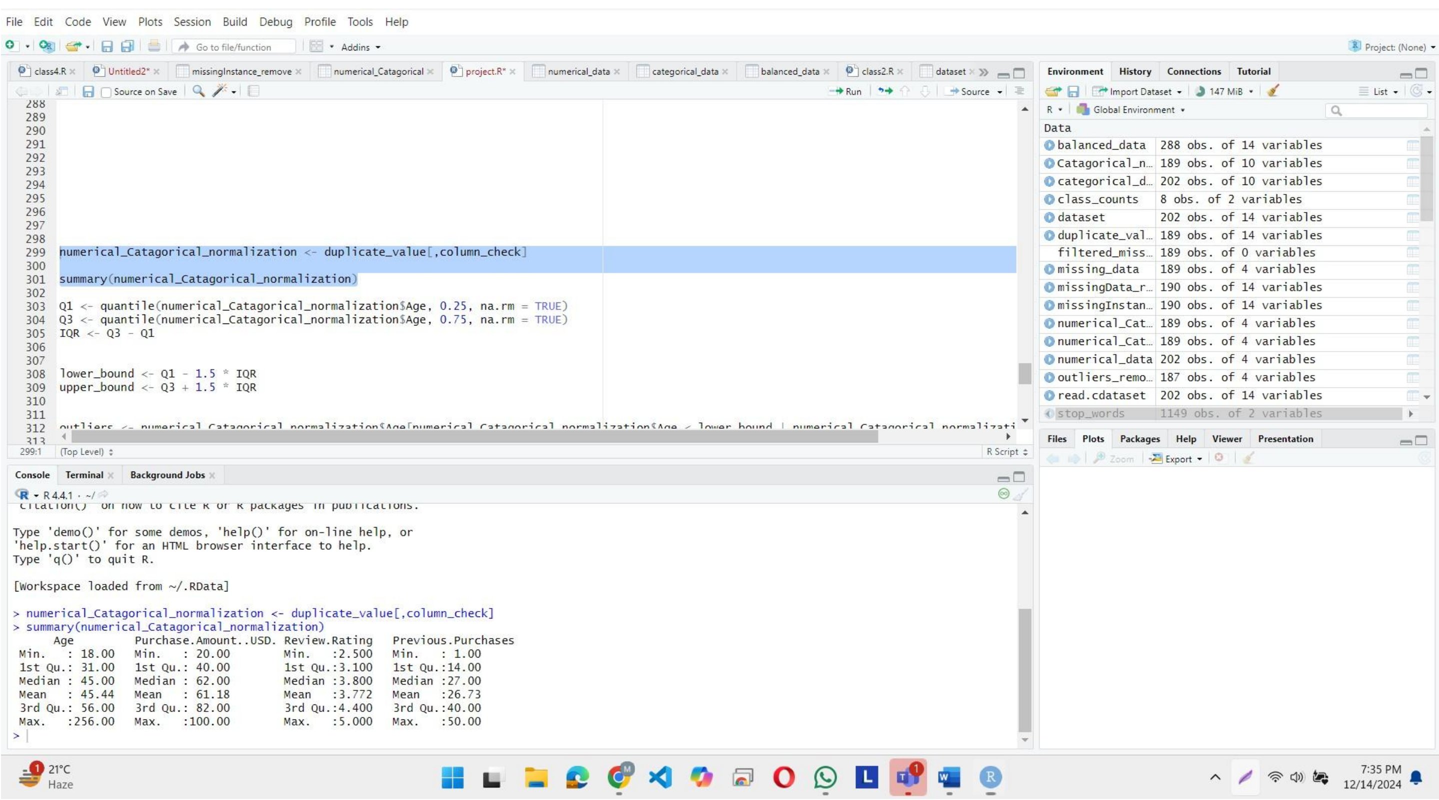
We convert categorical columns into numeric values. Gender, Item.Purchased, Category, Location, Size, Color, Subscription.Status, Discount.Applied, Promo.Code.Used, and Frequency.of.Purchases are all converted to numeric codes where each unique category is replaced with a unique numeric value.



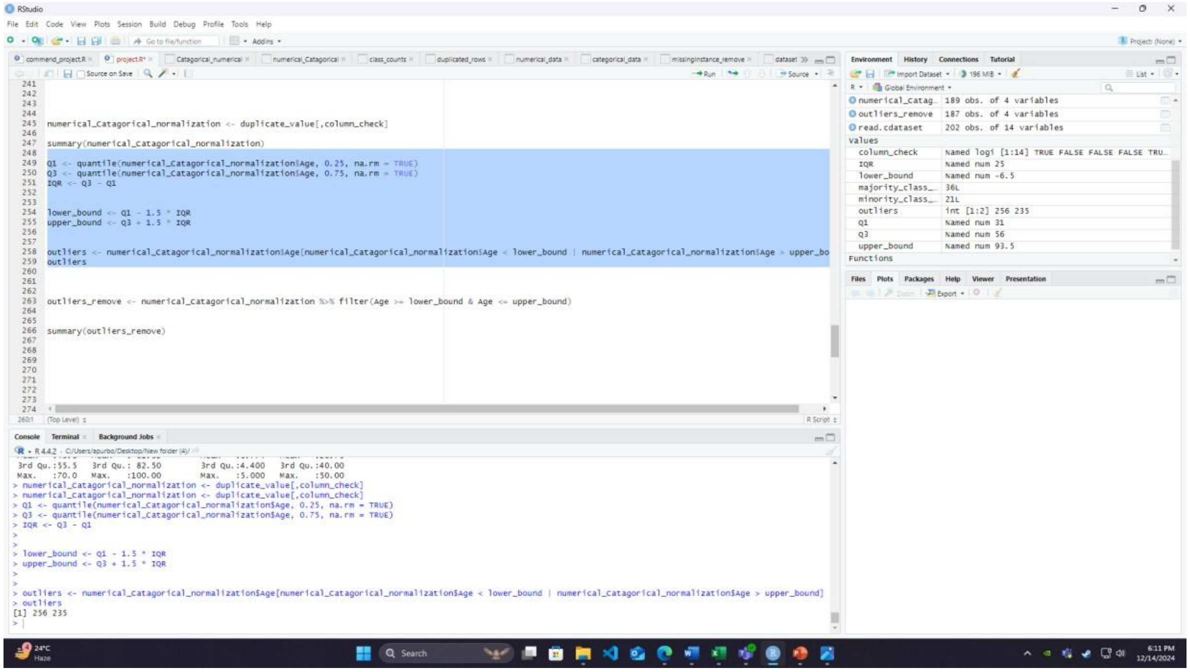


Outliers

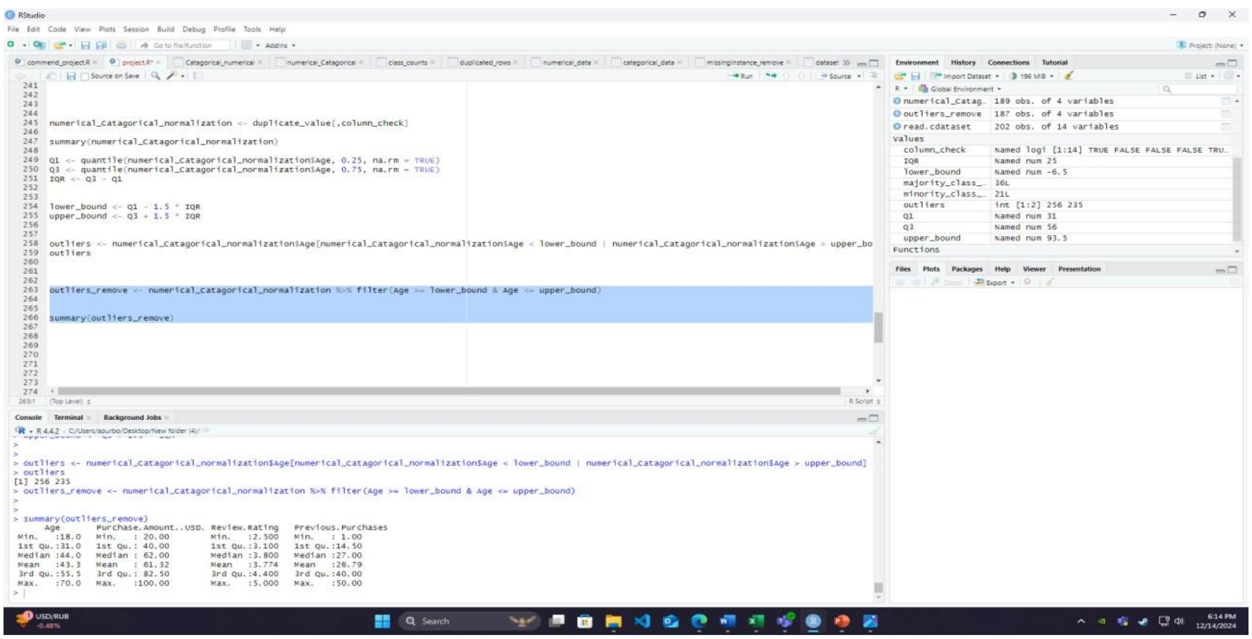
We select specific attributes from the dataset, identified by column_check, and store them in numerical_Catagorical_normalization. Then, we generate a summary of the dataset, which provides statistical details about the data.



We calculate the first (Q1) and third quartiles (Q3) of the age column in categorical normalization. Then we compute the Interquartile Range (IQR) by subtracting Q1 from Q3. The lower and upper bounds are determined using 1.5 times the IQR. At last we identify the outliers in the Age attribute that fall outside these bounds and store them in the outliers variable.

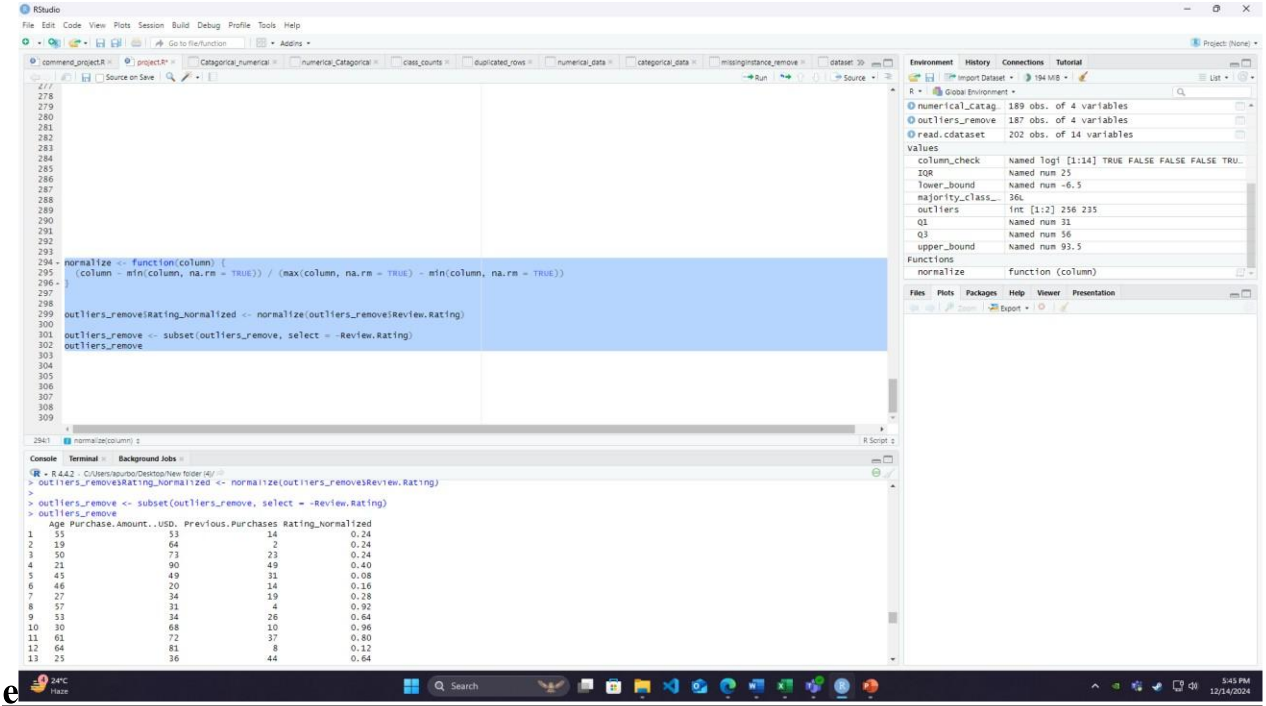


We filter the Catagorical_normalization_normalization dataset to remove the outliers from the Age column by keeping only the intence where Age is between the lower bound and upper bound. The resulting dataset outliers remove is then summarized to provide statistical details.



normalization method

Normalizes the Review.Rating column to a 0–1 scale adds it as a new column (Rating_Normalized) and then removes the original Review.Rating column from the dataset.



e