Milk Quality Prediction

Student Name: Md. Minhazul Bari Fahim
ID: 20-42176-1
Dept Name: CSE
Institute Name: American International University-Bangladesh
City, Country: Dhaka, Bangladesh
email address: minhazulbarifahim@gmail.com

Data set description: This dataset, which was generated through manual observations, is a vital resource for developing machine learning models to predict milk quality. The data set consists of 7 independent variables: pH, temperature, taste, odor, fat, turbidity, and color. The overall Grade or Quality of the milk is significantly influenced by these factors taken together. They perform an important role in facilitating predictive analysis for determining milk quality. The milk Grade, which can be binary in nature, is the desired variable in this situation. When the ideal conditions for Taste, Odor, Fat, and Turbidity are present, a value of 1 is assigned; otherwise, a value of 0 is given. The dataset also offers accurate pH and temperature readings.

i. **pH:** Its define pH, ranges from 3 to 9.5 max: 6.25 to 6.90
ii. **Temprature:** This column define temperature which ranges from 34'C to 90'C max: 34'C to 45.20’C
iii. **Taste, Odor, Fat, Turbidity:** Those columns define Categorical data 0 (Bad) or 1 (Good). Max: 1 (Good)
iv. **Colour:** This Column defines Colour of the milk which ranges from 240 to 255. Max: 255
v. **Grade**: This Column defines Grade (Target) of the milk which is categorical data Where Low (Bad) or Medium (Moderate) High

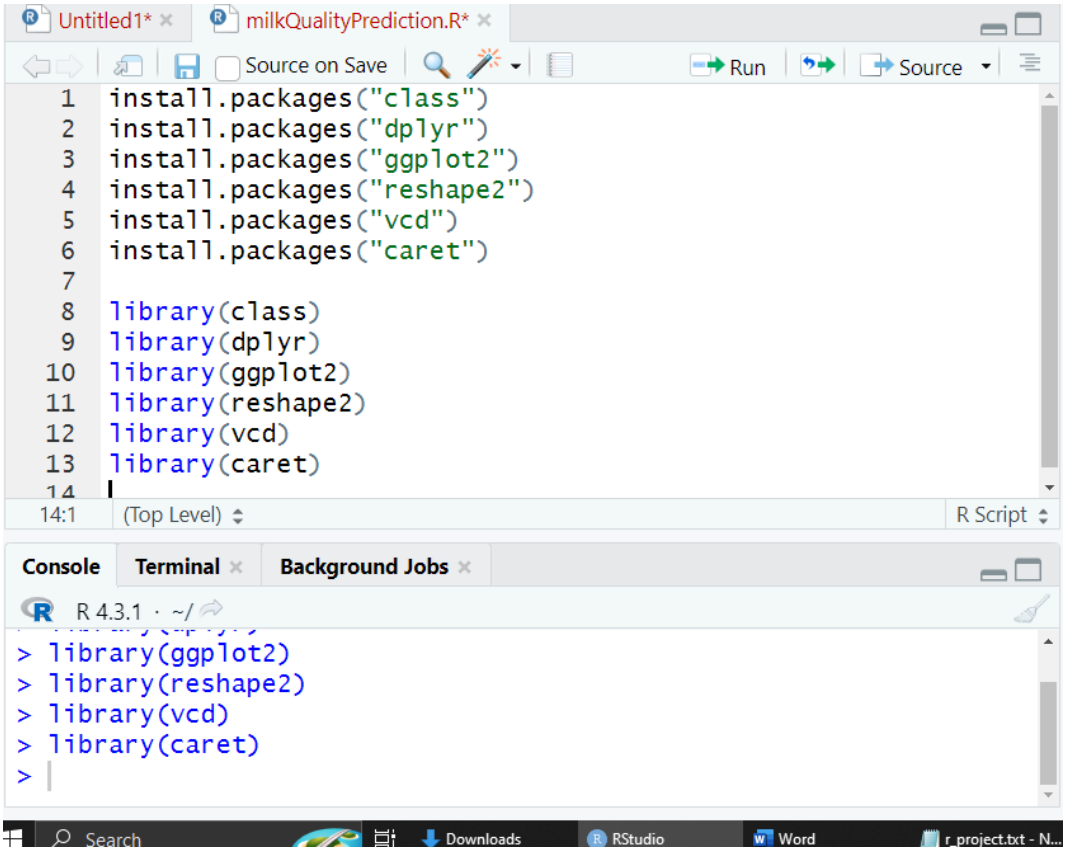
Methodology: The provided project performs the following steps:

i. Data loading and preprocessing (handling missing values, converting categorical data, normalization, and feature selection).
ii. Correlation analysis and visualization.
iii. Data splitting into training and test sets.
iv. Building a KNN classification model and evaluating its performance using accuracy.
v. Performing 10-fold cross-validation to select the best KNN model.
vi. Generating a confusion matrix and calculating precision and recall.

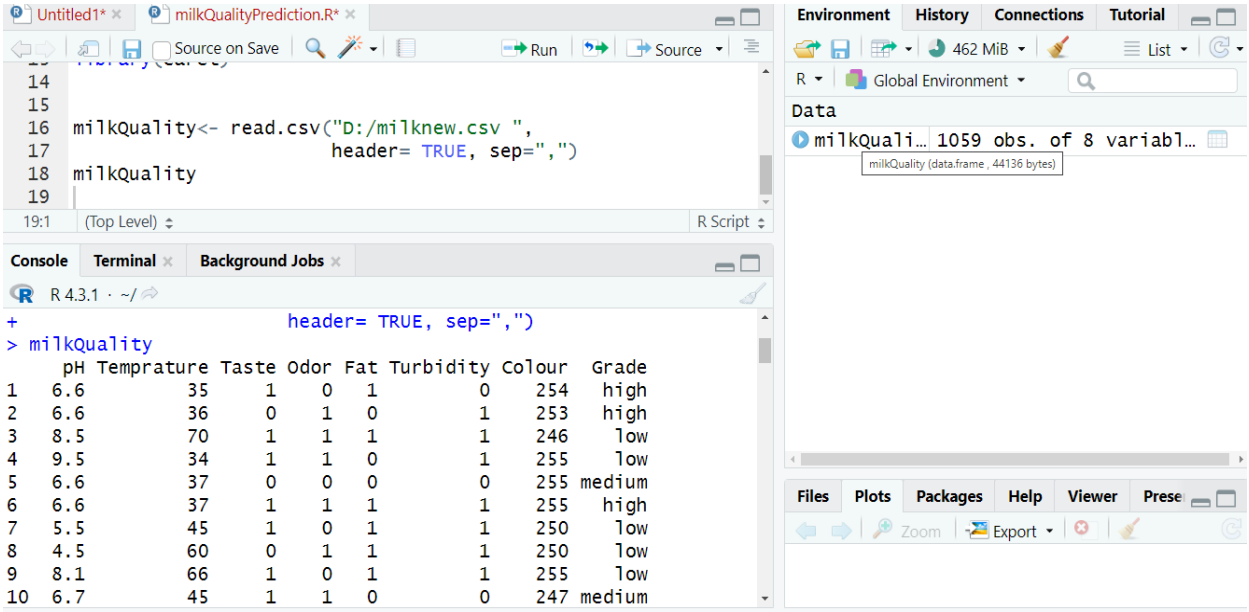
The project is a comprehensive implementation of a KNN classification workflow with data preprocessing, model training, evaluation, and performance analysis.

Data set link: https://www.kaggle.com/datasets/cpluzshrijayan/milkquality

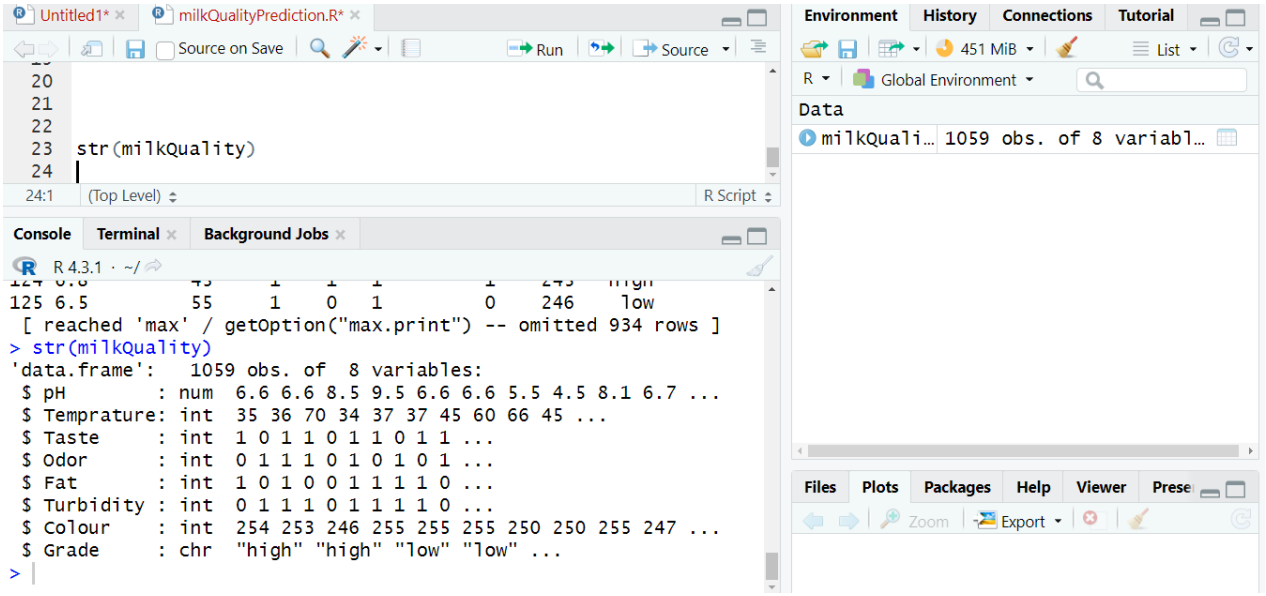
Load all the necessary libraries for the dataset.



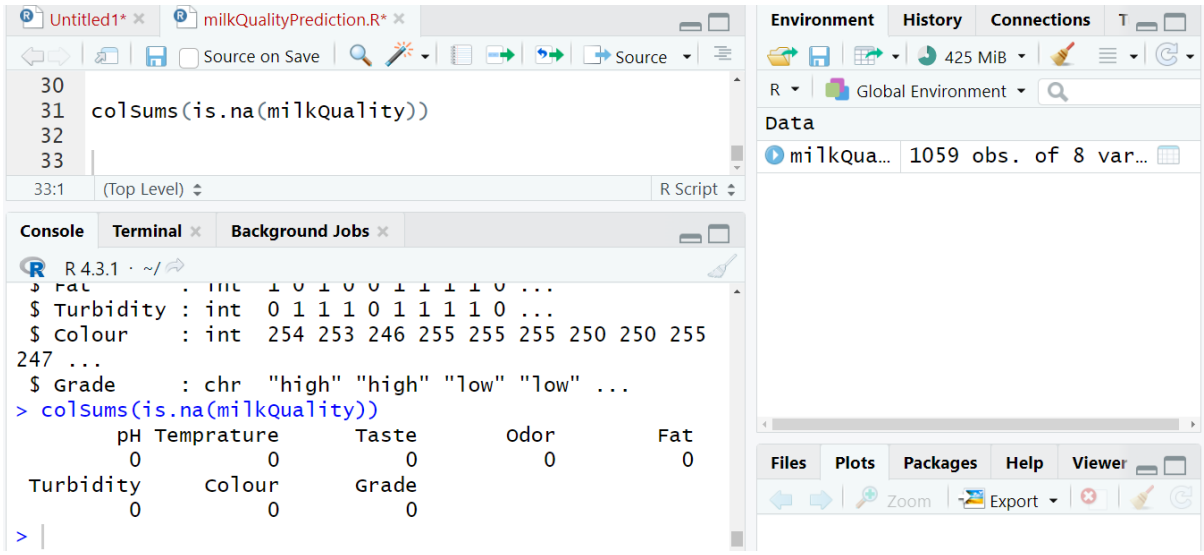
Importing dataset (csv file) name as milkQuality and printing the dataset.



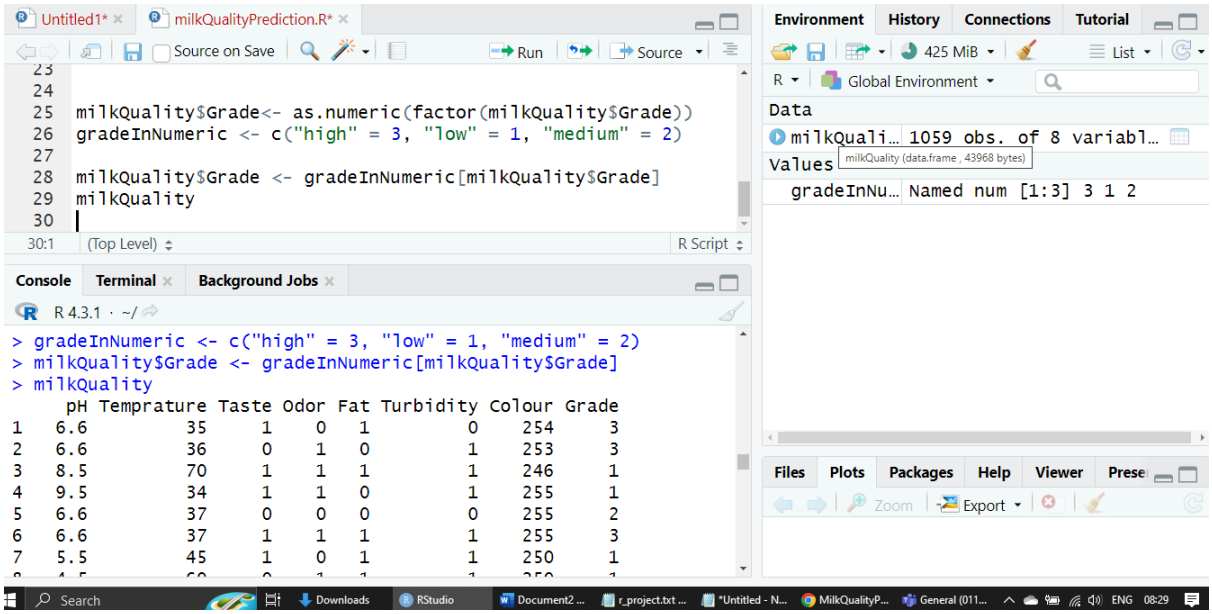
Display detailed information about dataset.



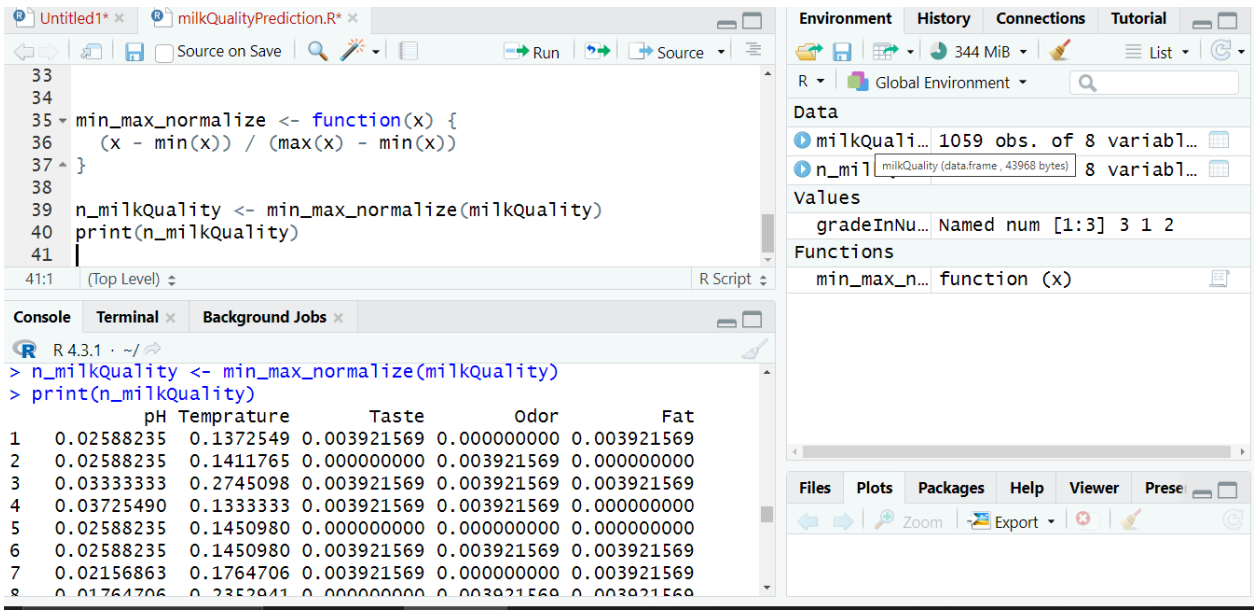
Show the number of missing values in each column.



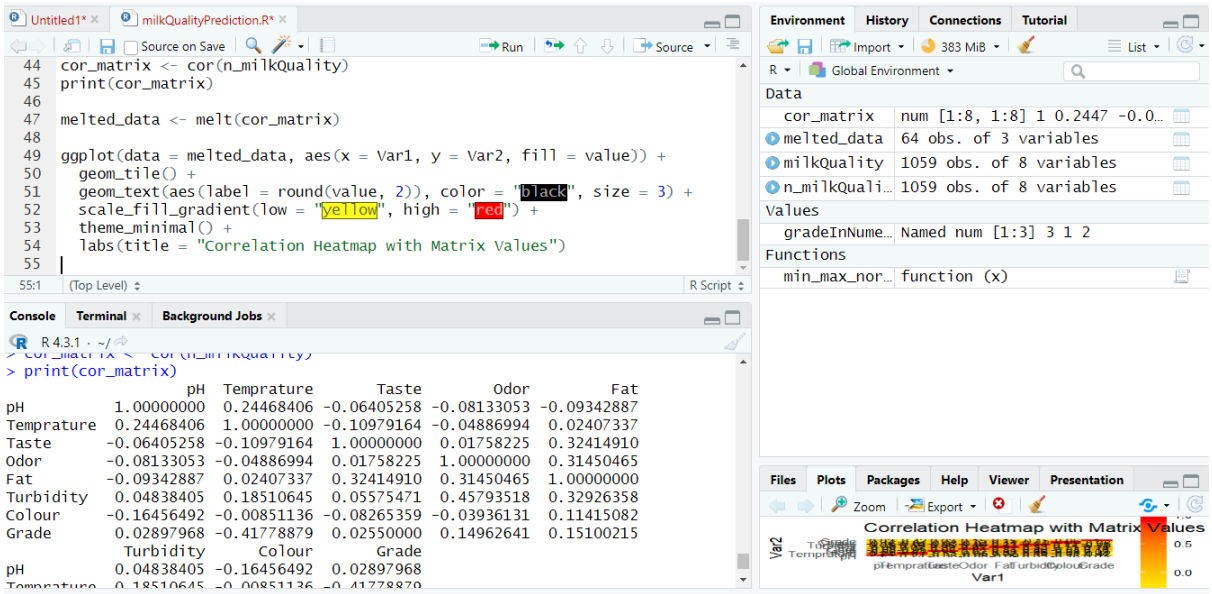
Convert categorical to numeric data.

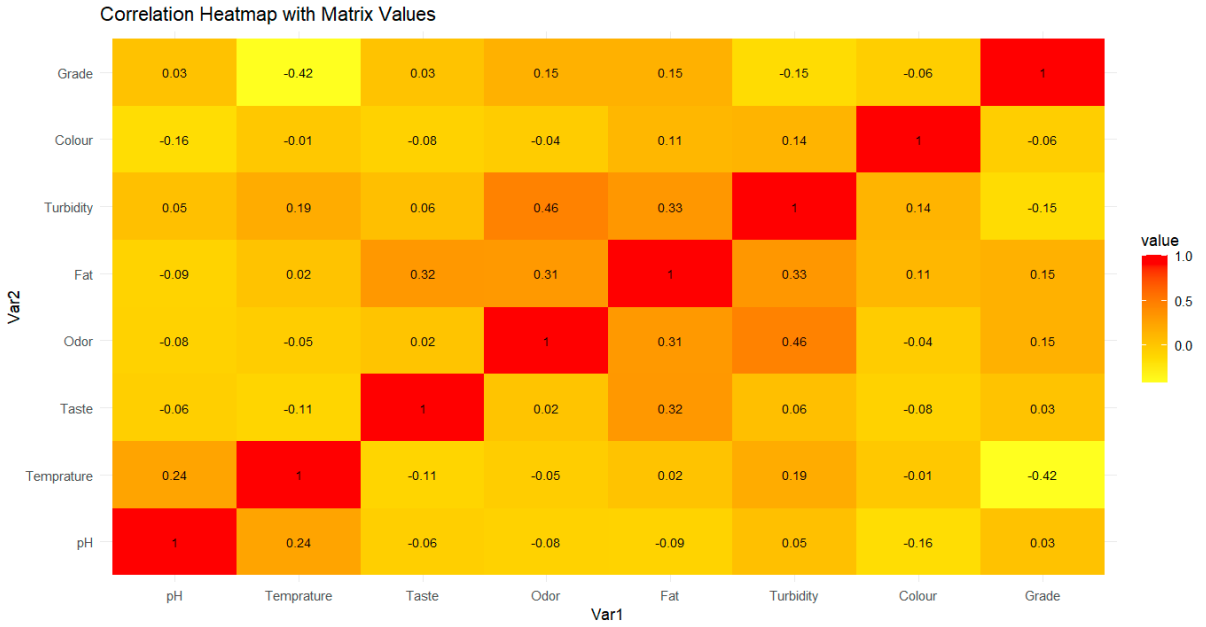


Use the min-man normalization after that save the normalize dataset name as n_milkQuality.

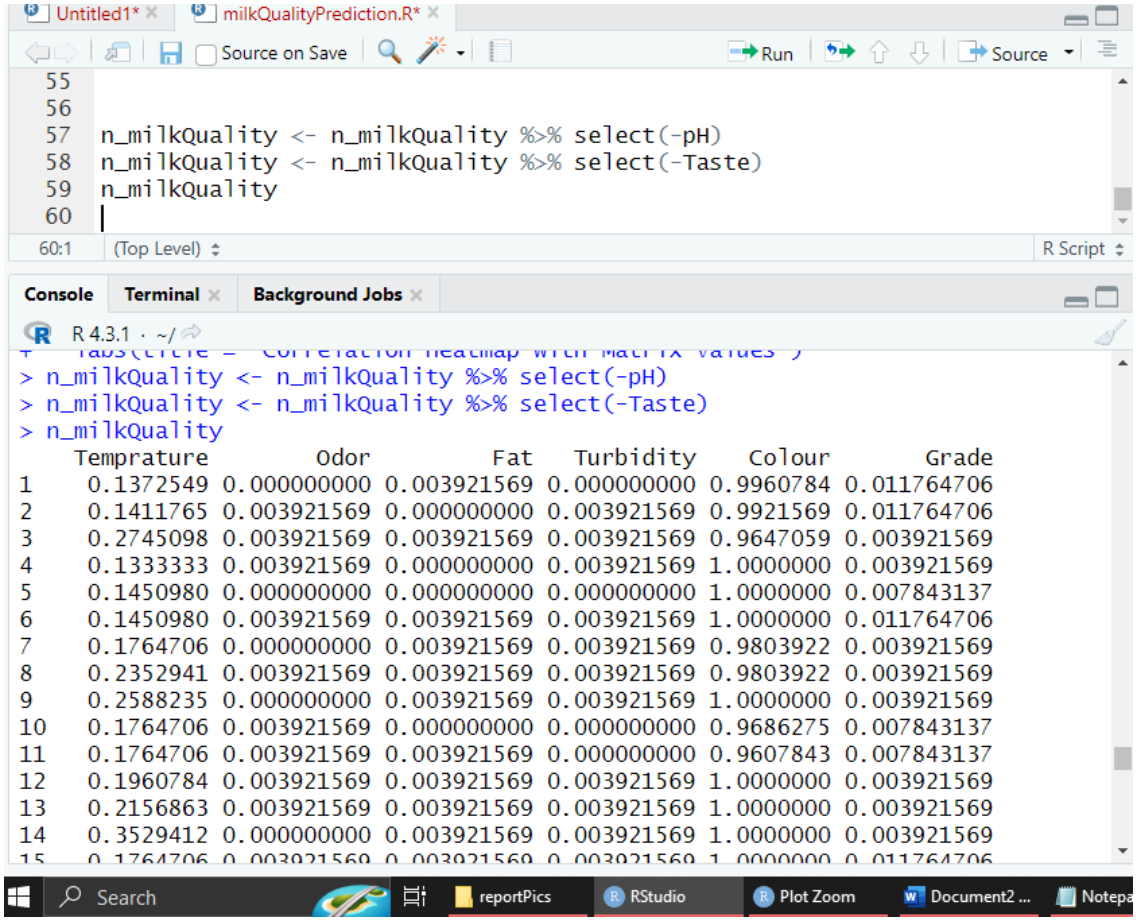


Showing correlation heatmap with matrix values.

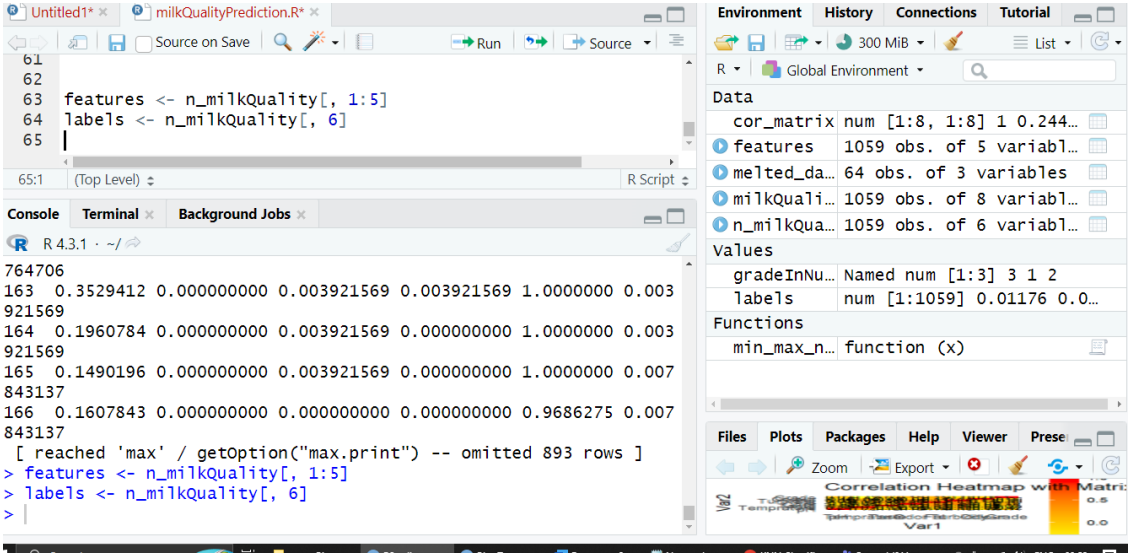




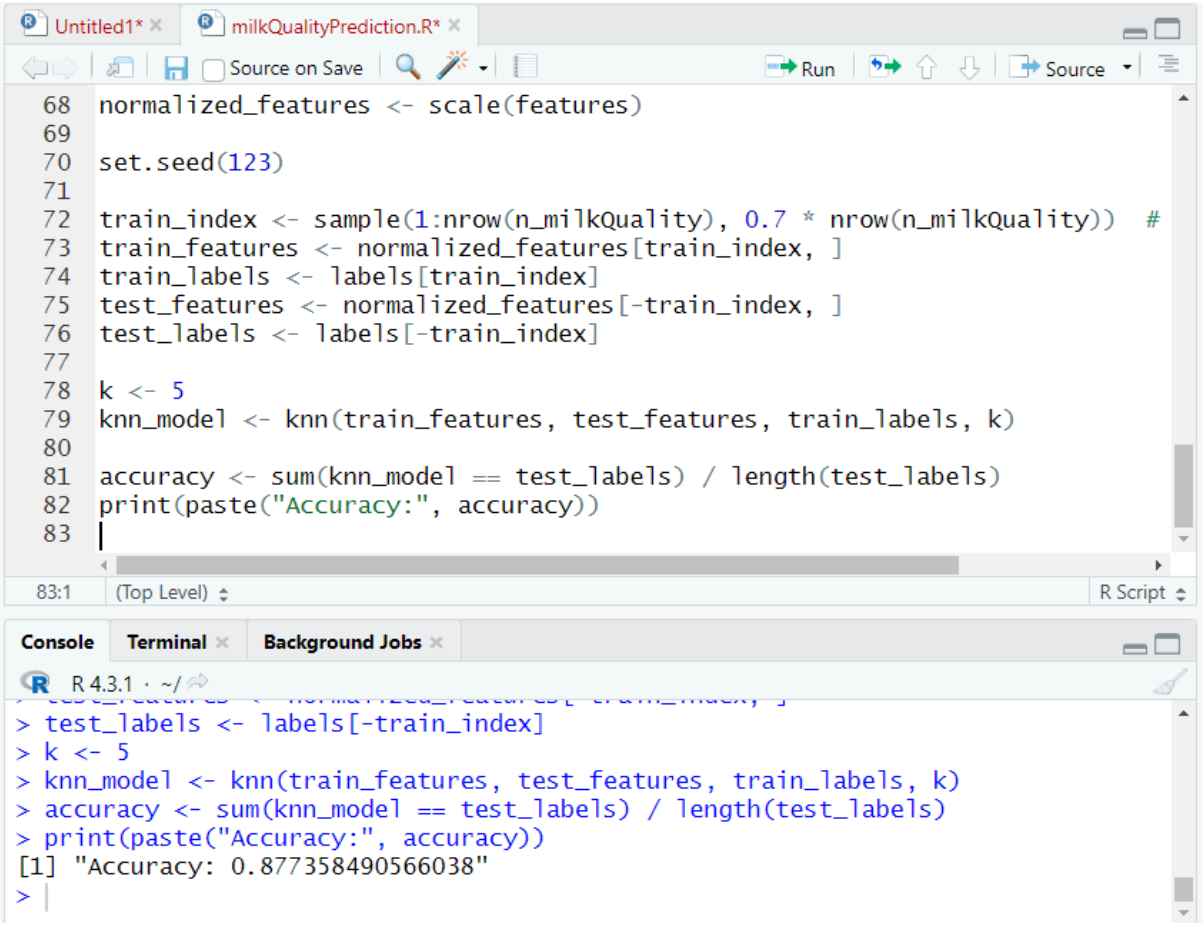
- **After checking correlation, drop two column.**



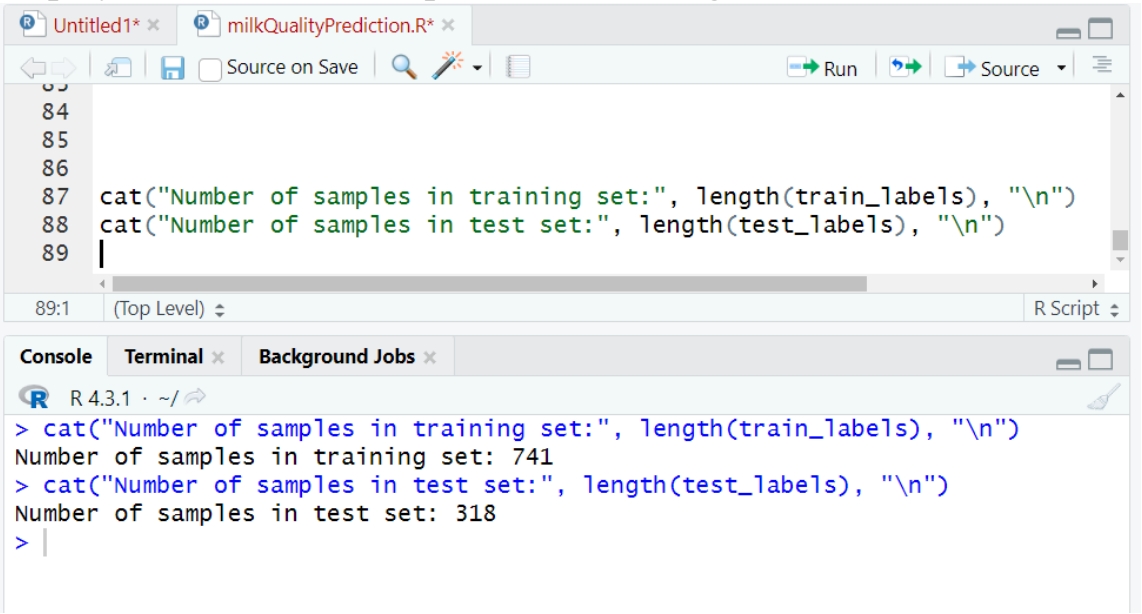
- **Assuming the last column is the target variable and the rest are features**



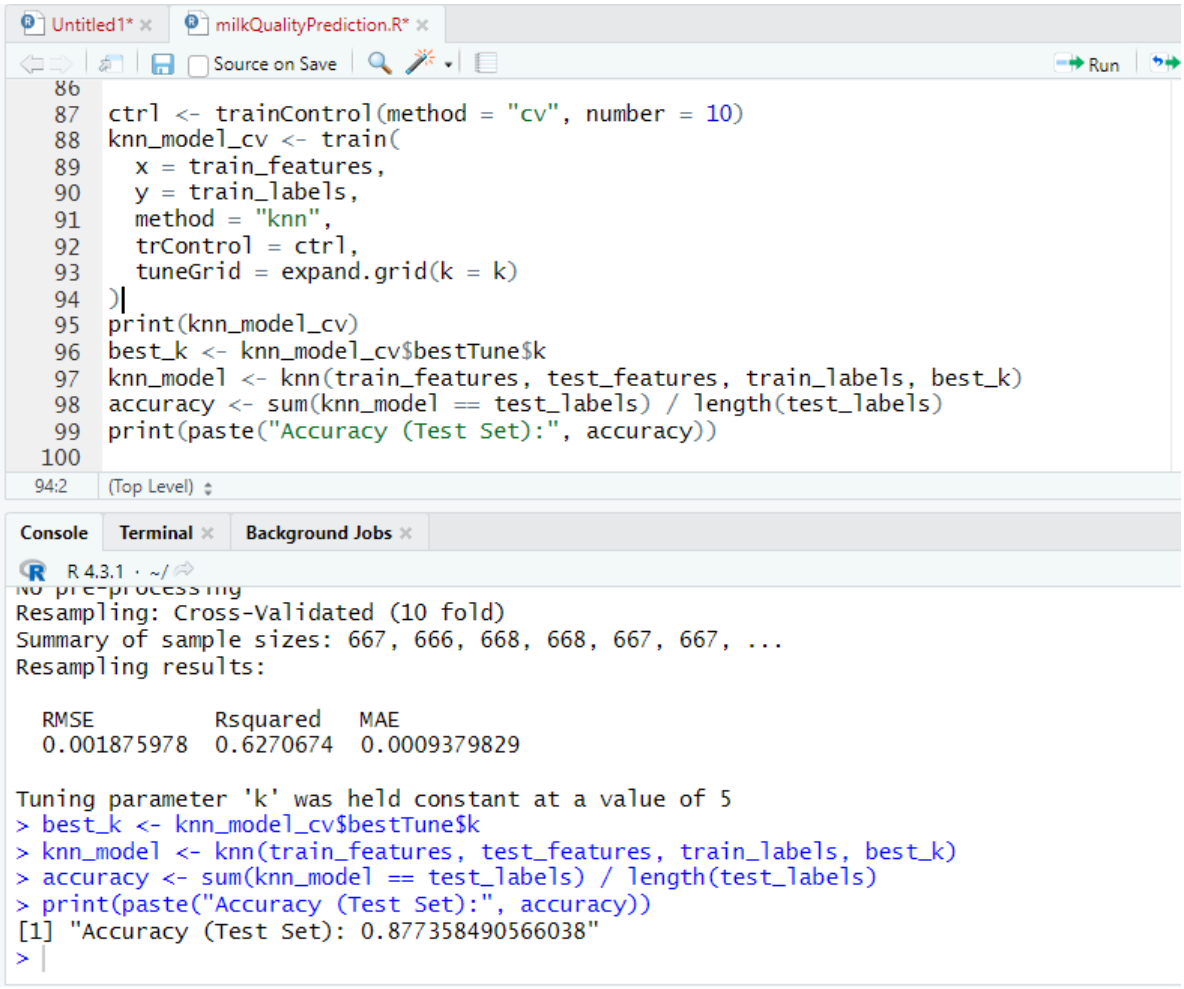
Built KNN model and find the accuracy.



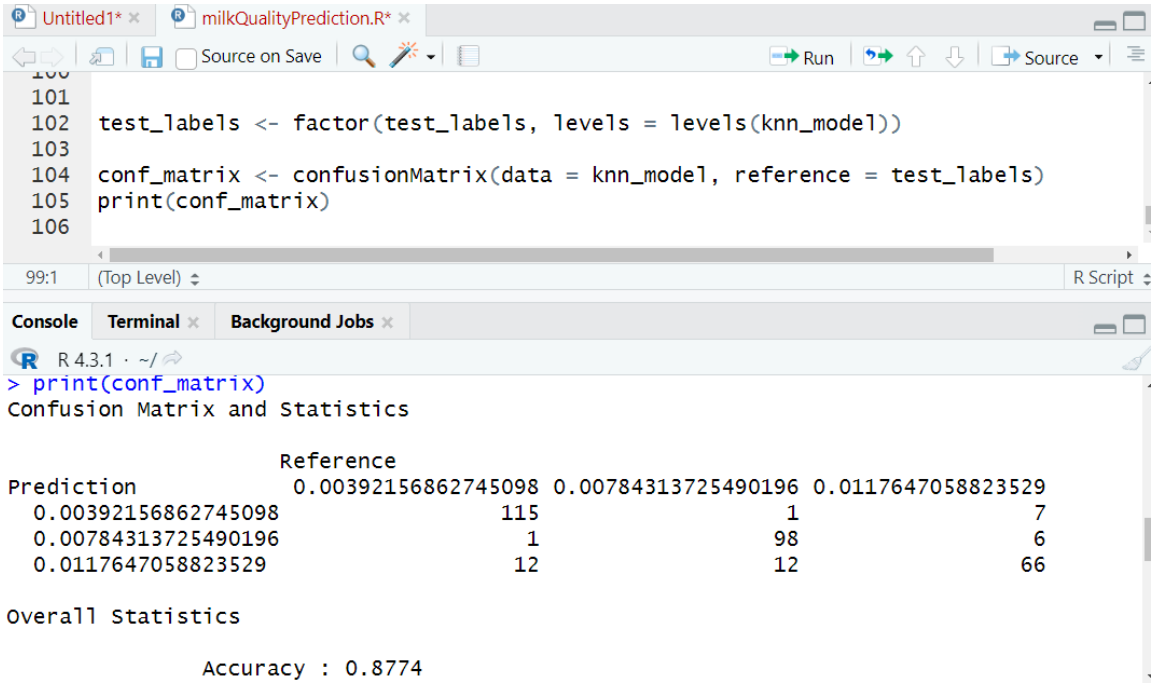
Display the number of samples in the training and test sets



Using 10-fold cross validation and predicting accuracy.



Confusion matrix and statistics



```
R 4.3.1 · ~/
Confusion Matrix and Statistics

                     Reference
Prediction            0.00392156862745098 0.00784313725490196 0.0117647058823529
  0.00392156862745098                 115                   1                   7
  0.00784313725490196                   1                  98                   6
  0.0117647058823529                   12                  12                  66

Overall Statistics

               Accuracy : 0.8774
                 95% CI : (0.8362, 0.9113)
    No Information Rate : 0.4025
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8138

 Mcnemar's Test P-Value : 0.3455

Statistics by Class:

                     Class: 0.00392156862745098 Class: 0.00784313725490196
Sensitivity                              0.8984                      0.8829
Specificity                              0.9579                      0.9662
Pos Pred Value                           0.9350                      0.9333
Neg Pred Value                           0.9333                      0.9390
Prevalence                               0.4025                      0.3491
Detection Rate                           0.3616                      0.3082
Detection Prevalence                     0.3868                      0.3302
Balanced Accuracy                        0.9282                      0.9245
                     Class: 0.0117647058823529
Sensitivity                             0.8354
Specificity                             0.8996
Pos Pred Value                          0.7333
Neg Pred Value                          0.9430
Prevalence                              0.2484
Detection Rate                          0.2075
Detection Prevalence                    0.2830
Balanced Accuracy                       0.8675
>
```

Finding the precision and recall.

```
Untitled1* × milkQualityPrediction.R* ×
Source on Save | Run | Source
107
108
109
110 precision <- conf_matrix$table[2, 2] / sum(conf_matrix$table[, 2])
111 recall <- conf_matrix$table[2, 2] / sum(conf_matrix$table[2, ])
112
113 cat("Precision:", precision, "\n")
114 cat("Recall:", recall, "\n")
115
116
117
115:1 (Top Level) R Script

Console Terminal × Background Jobs ×
R 4.3.1 · ~/
> precision <- conf_matrix$table[2, 2] / sum(conf_matrix$table[, 2])
> recall <- conf_matrix$table[2, 2] / sum(conf_matrix$table[2, ])
> cat("Precision:", precision, "\n")
Precision: 0.8828829
> cat("Recall:", recall, "\n")
Recall: 0.9333333
>
```