American International University-Bangladesh (AIUB)

Course Code: CSC4180

Course Title: *Introduction to Data Science*

Project Title : Student Performance Prediction Using KNN Classification Algorithm.

Submitted by

Name : RAHMAN MD, MAHFUZUR

ID : 20-43186-1

SECTION : [B]

Submitted to

TOHEDUL ISLAM

Assistant Professor

Faculty of Science and technology

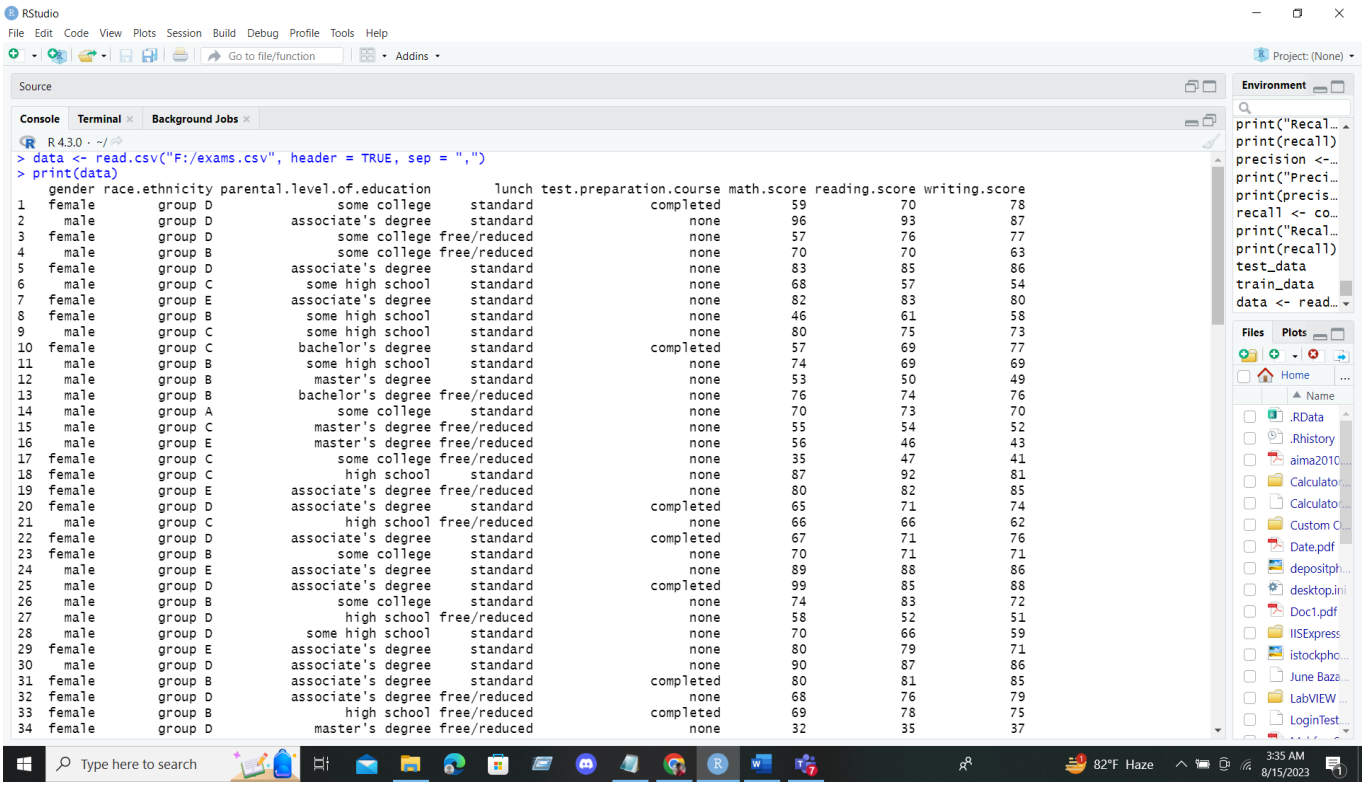
I. Project Description :

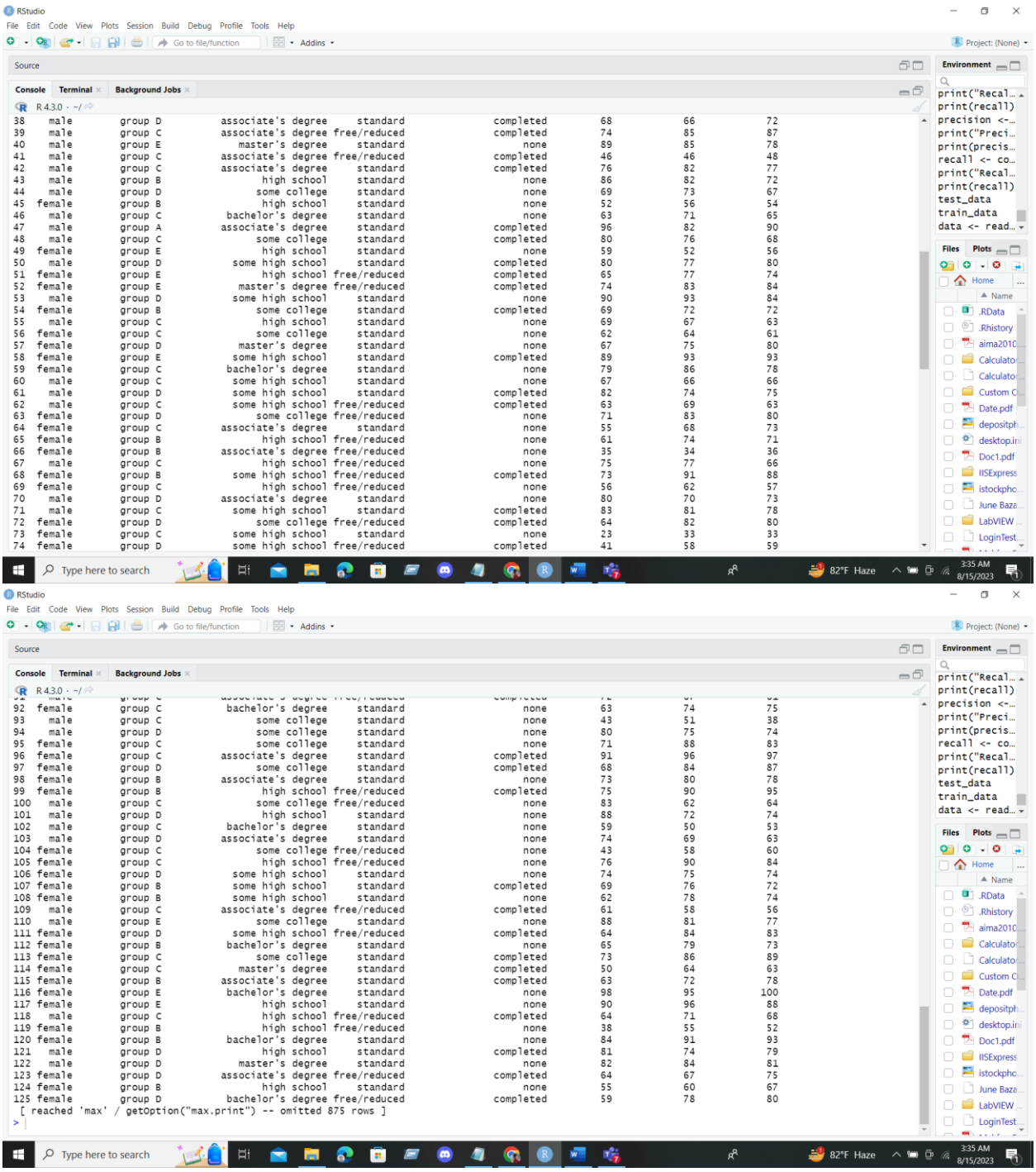
This project involves the development of a classification model to predict student math performance using the K-Nearest Neighbors (KNN) algorithm. The primary objective is to determine whether students pass or fail based on a set of relevant features. The project begins with data preprocessing, including encoding categorical variables and identifying feature correlations. Low correlation features are pruned to optimize model performance. The dataset is then split into training and testing sets for model evaluation. The KNN algorithm is implemented, and the optimal number of neighbors (K value) is determined using cross-validation. Precision and recall metrics assess the model's accuracy in classifying student performance. Insights gained from feature importance and model effectiveness guide potential enhancements.

Future work may involve refining hyperparameters, exploring alternative algorithms, and advanced feature engineering techniques. The project aims to deliver an efficient and robust classification model applicable in educational contexts, with findings that can contribute to further research and improvements in student performance prediction.

1. Load Libraries and Dataset:

In this initial phase, we start by loading the crucial R library "caret," which serves as a powerful toolkit for training and evaluating machine learning models. The library encompasses diverse functions and tools that are invaluable for our analysis. We then proceed to load the dataset "exams.csv" from the specified file location. This dataset embodies a comprehensive collection of information, encompassing student demographics, exam scores, and performance indicators. This preparatory step lays the foundation for the subsequent data manipulation and analysis.





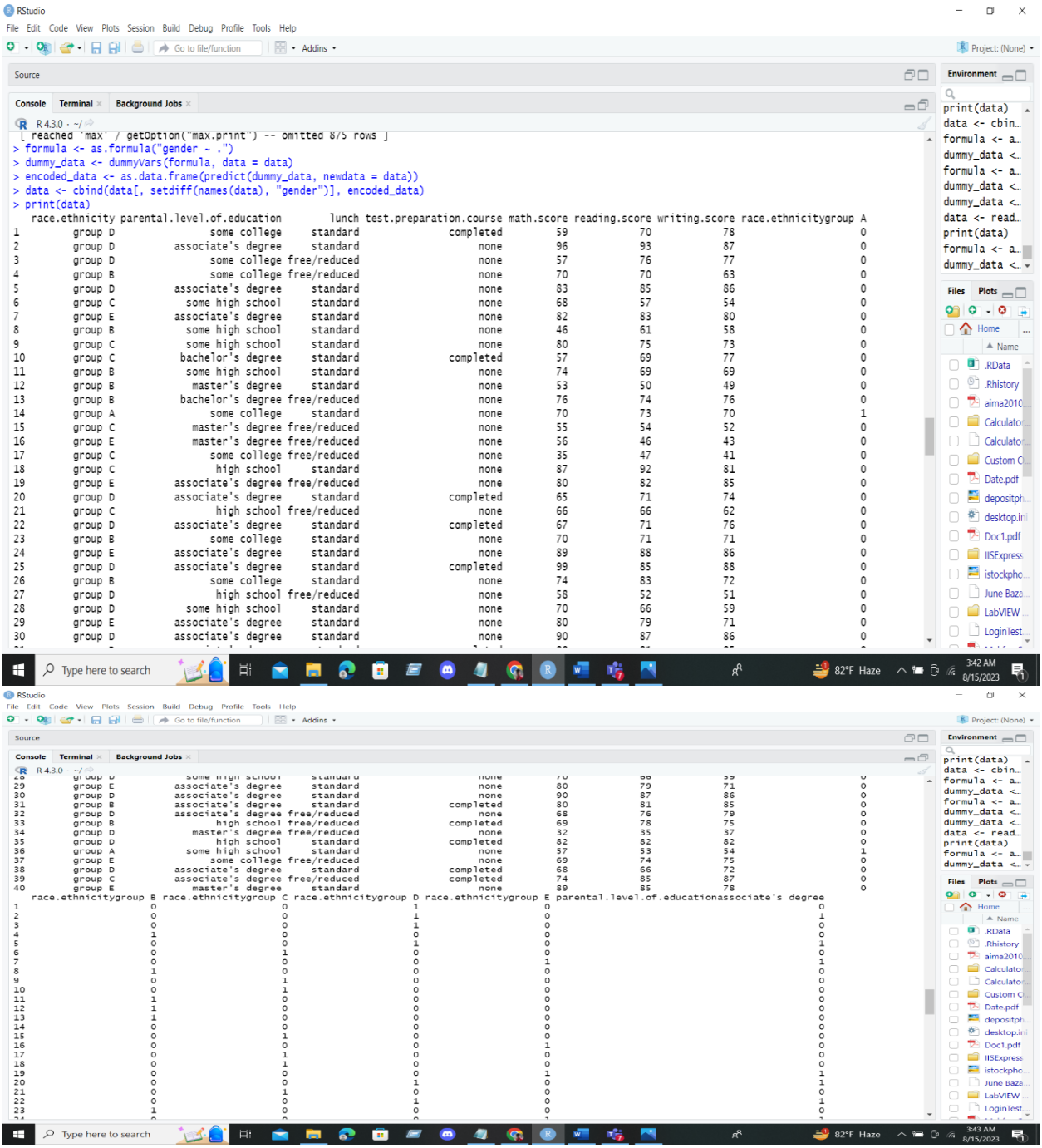
II. Encode Categorical Variables and Create Dummy Variables:

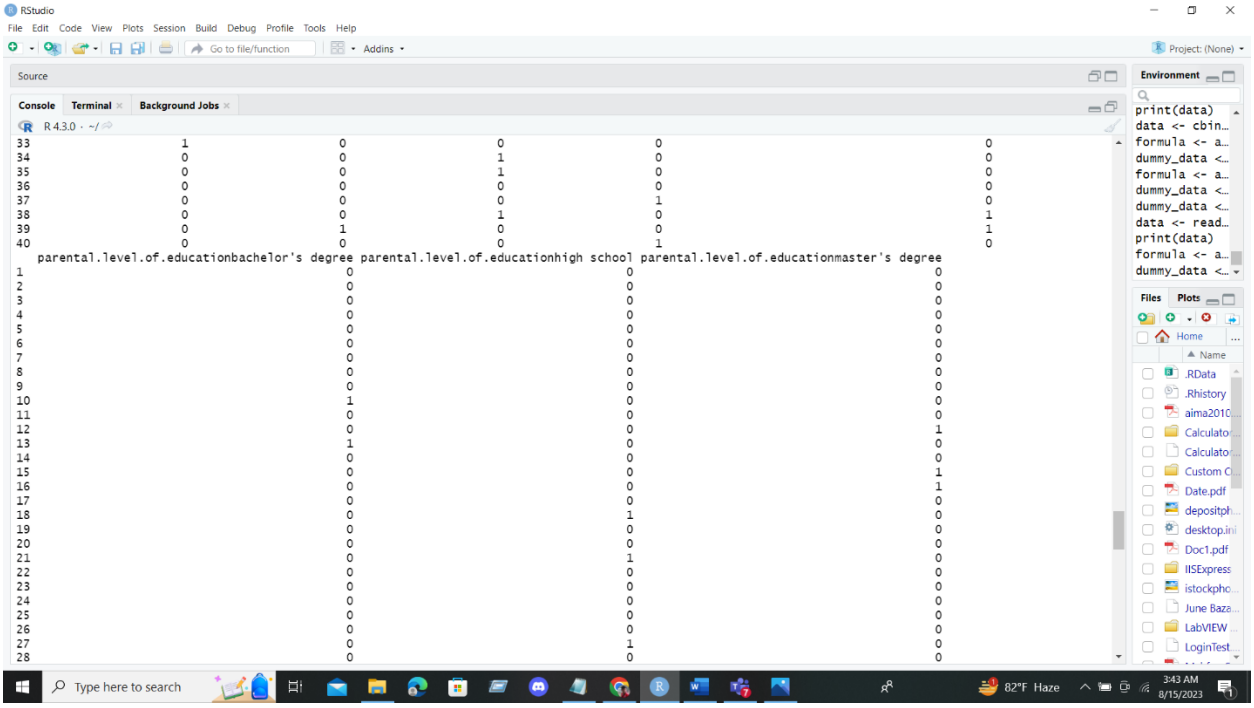
I. Encode Categorical Variables:

Within our dataset, certain variables such as "race.ethnicity," "parental.level.of.education," "lunch," and "test.preparation.course" are categorical in nature. To harness their potential for machine learning, it is essential to convert them into a suitable format. We undertake this by transforming these categorical columns into "factor" variables. This process facilitates numerical representation and paves the way for incorporating these attributes into machine learning models. This encoding is a pivotal preprocessing step that ensures our algorithms can effectively work with categorical data.

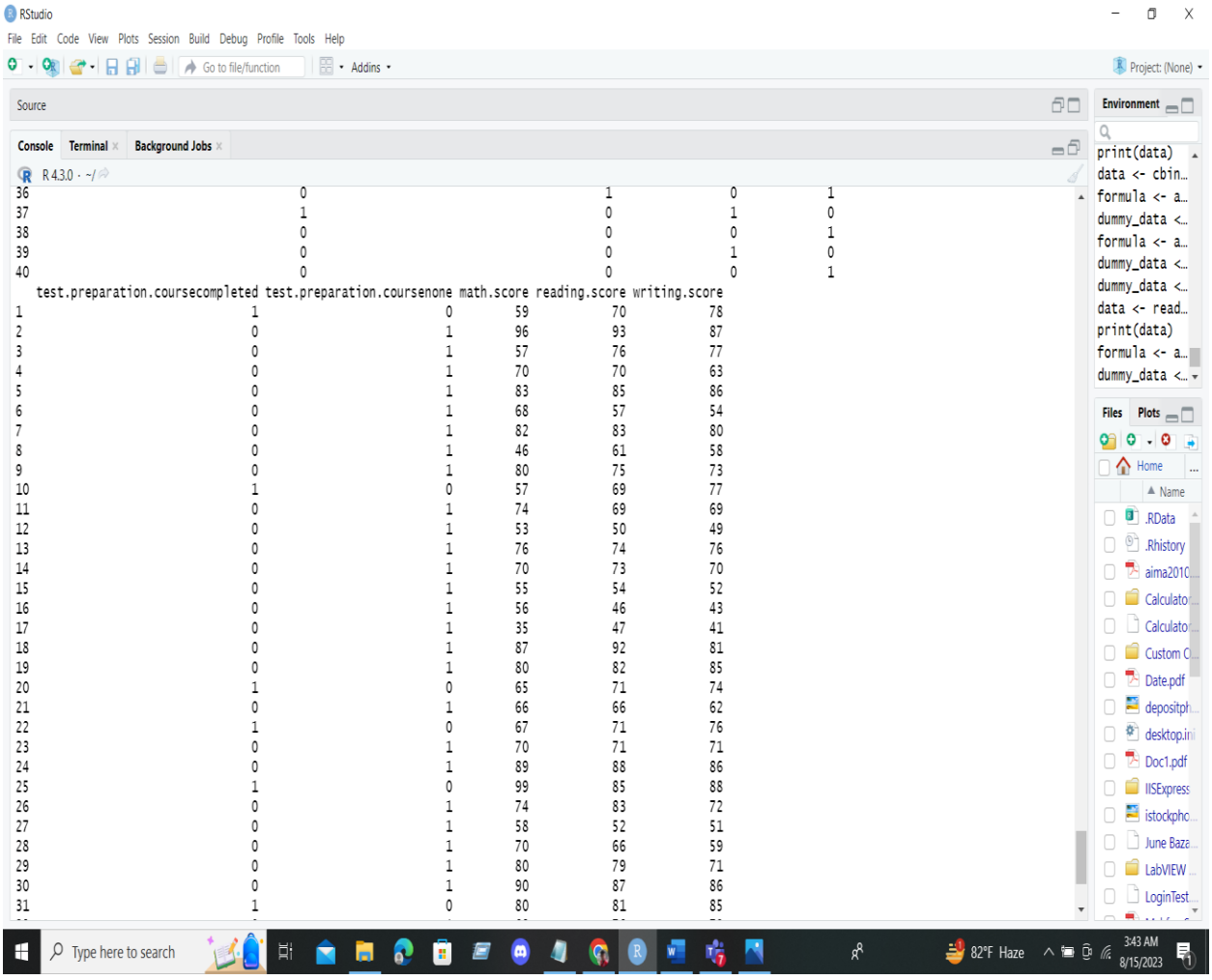
II. Create Dummy Variables :

In this phase, we delve into the intricate process of converting categorical data into a format suitable for machine learning algorithms. Our strategy involves formulating a predictive model for the "gender" variable based on all other variables. Subsequently, we generate "dummy variables" – binary indicators for each category within a categorical variable. By merging these dummy variables with the original dataset, we create an augmented dataset that retains the categorical information while transforming it into a format that machine learning algorithms can comprehend. This strategic transformation empowers subsequent analysis and model building.



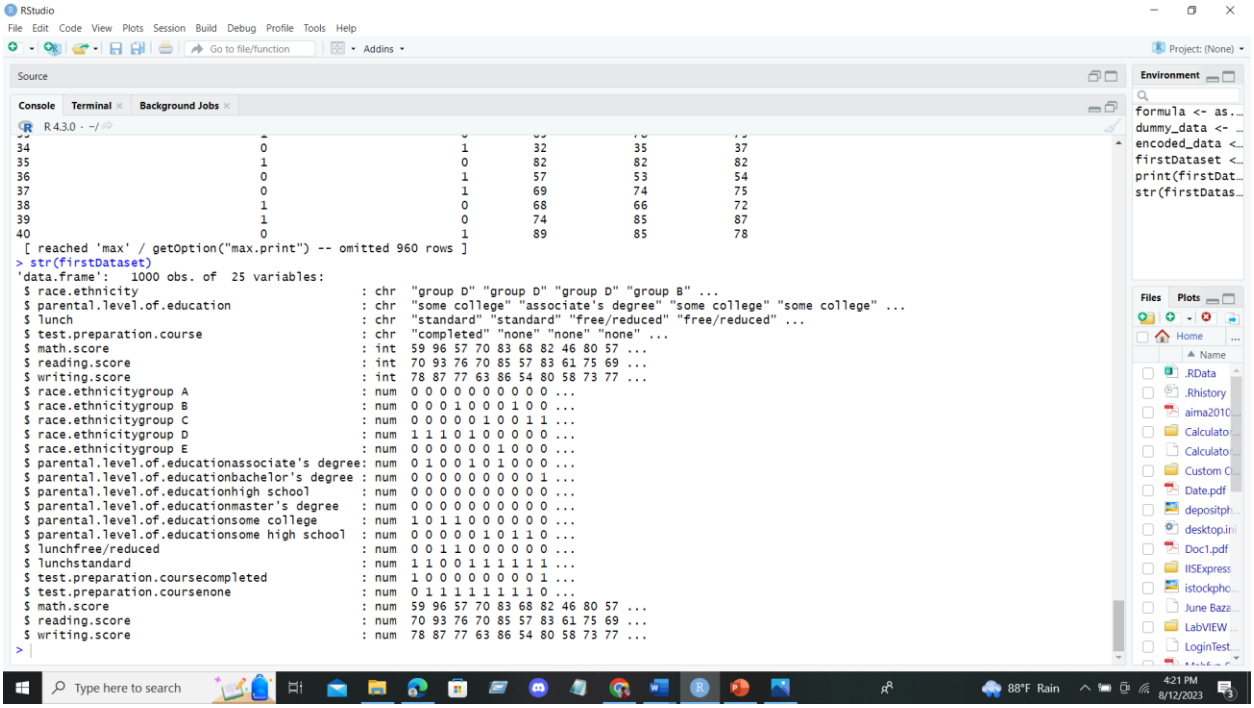


```
33                    1                    0                    0                    0                                    0
34                    0                    1                    0                    0                                    0
35                    0                    0                    1                    0                                    0
36                    0                    0                    0                    0                                    0
37                    0                    0                    0                    1                                    0
38                    0                    0                    1                    0                                    1
39                    0                    1                    0                    0                                    1
40                    0                    0                    0                    1                                    0
   parental.level.of.educationbachelor's degree parental.level.of.educationhigh school parental.level.of.educationmaster's degree
1                                             0                                     0                                        0
2                                             0                                     0                                        0
3                                             0                                     0                                        0
4                                             0                                     0                                        0
5                                             0                                     0                                        0
6                                             0                                     0                                        0
7                                             0                                     0                                        0
8                                             0                                     0                                        0
9                                             0                                     0                                        0
10                                            1                                     0                                        0
11                                            0                                     0                                        0
12                                            0                                     0                                        1
13                                            1                                     0                                        0
14                                            0                                     0                                        0
15                                            0                                     0                                        1
16                                            0                                     0                                        1
17                                            0                                     0                                        0
18                                            0                                     1                                        0
19                                            0                                     0                                        0
20                                            0                                     0                                        0
21                                            0                                     1                                        0
22                                            0                                     0                                        0
23                                            0                                     0                                        0
24                                            0                                     0                                        0
25                                            0                                     0                                        0
26                                            0                                     0                                        0
27                                            0                                     1                                        0
28                                            0                                     0                                        0
```



```
36                                    0                                    1                    0                    1
37                                    1                                    0                    1                    0
38                                    0                                    0                    0                    1
39                                    0                                    0                    1                    0
40                                    0                                    0                    0                    1
   test.preparation.coursecompleted test.preparation.coursenone math.score reading.score writing.score
1                                 1                           0         59            70            78
2                                 0                           1         96            93            87
3                                 0                           1         57            76            77
4                                 0                           1         70            70            63
5                                 0                           1         83            85            86
6                                 0                           1         68            57            54
7                                 0                           1         82            83            80
8                                 0                           1         46            61            58
9                                 0                           1         80            75            73
10                                1                           0         57            69            77
11                                0                           1         74            69            69
12                                0                           1         53            50            49
13                                0                           1         76            74            76
14                                0                           1         70            73            70
15                                0                           1         55            54            52
16                                0                           1         56            46            43
17                                0                           1         35            47            41
18                                0                           1         87            92            81
19                                0                           1         80            82            85
20                                1                           0         65            71            74
21                                0                           1         66            66            62
22                                1                           0         67            71            76
23                                0                           1         70            71            71
24                                0                           1         89            88            86
25                                1                           0         99            85            88
26                                0                           1         74            83            72
27                                0                           1         58            52            51
28                                0                           1         70            66            59
29                                0                           1         80            79            71
30                                0                           1         90            87            86
31                                1                           0         80            81            85
```

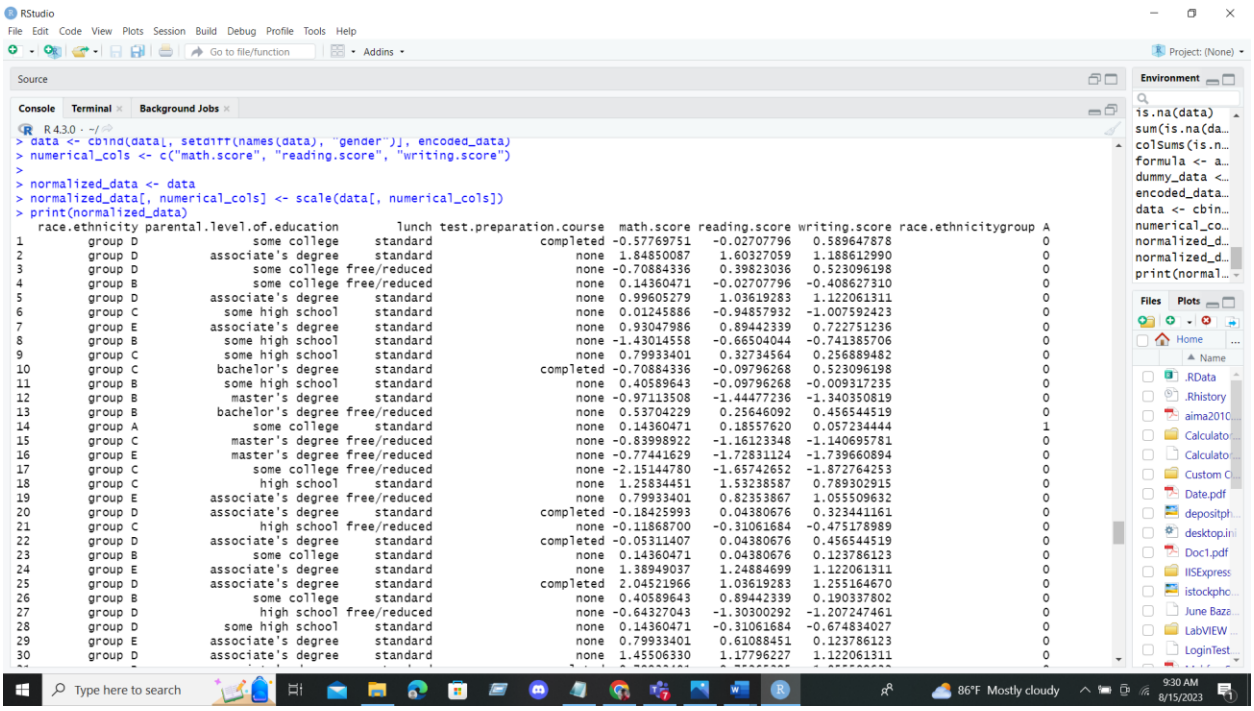
*III.*Handling Missing Values:



*IV.*Normalization:

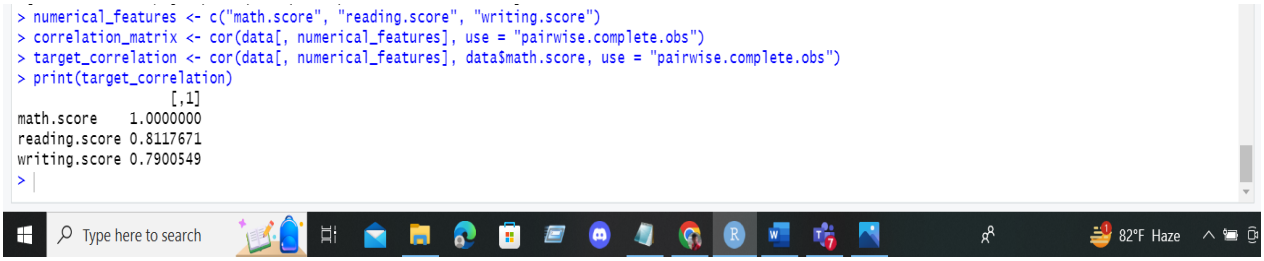
Feature normalization is a preprocessing step commonly used in machine learning to ensure that numerical features are on a similar scale. This is important because some machine learning algorithms, such as k-nearest neighbors (k-NN), are sensitive to the scale of input features. When features have different scales, it can lead to biased predictions and suboptimal model performance.

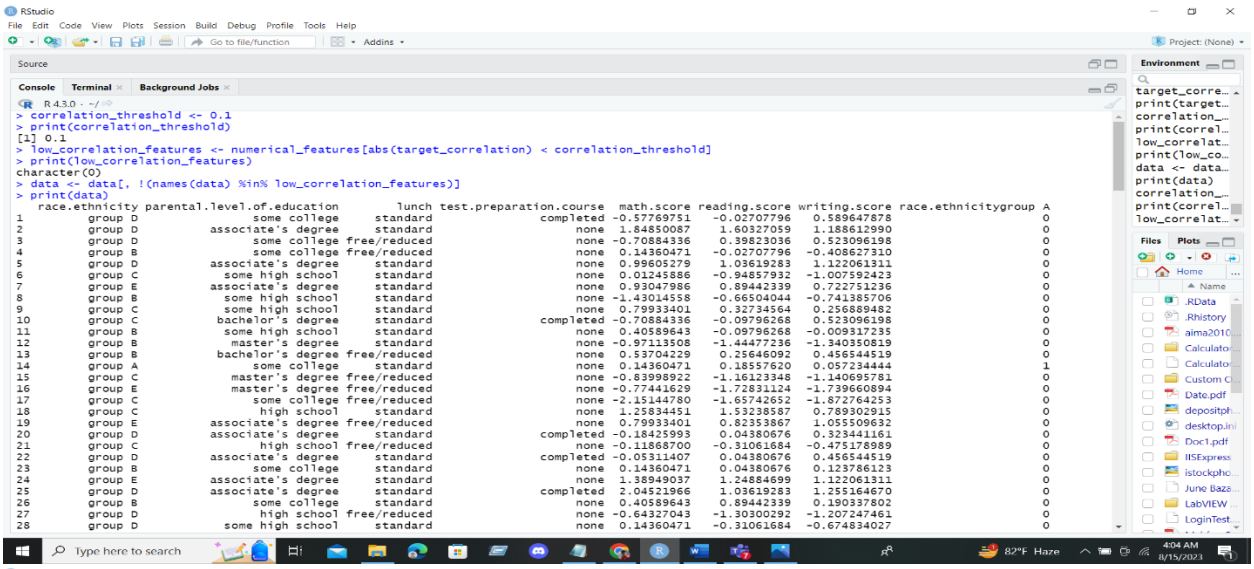
In our dataset, we have selected a subset of numerical columns, namely "math.score," "reading.score," and "writing.score," for normalization. The goal is to transform these features so that they have a mean of 0 and a standard deviation of 1. This process is also known as standardization.



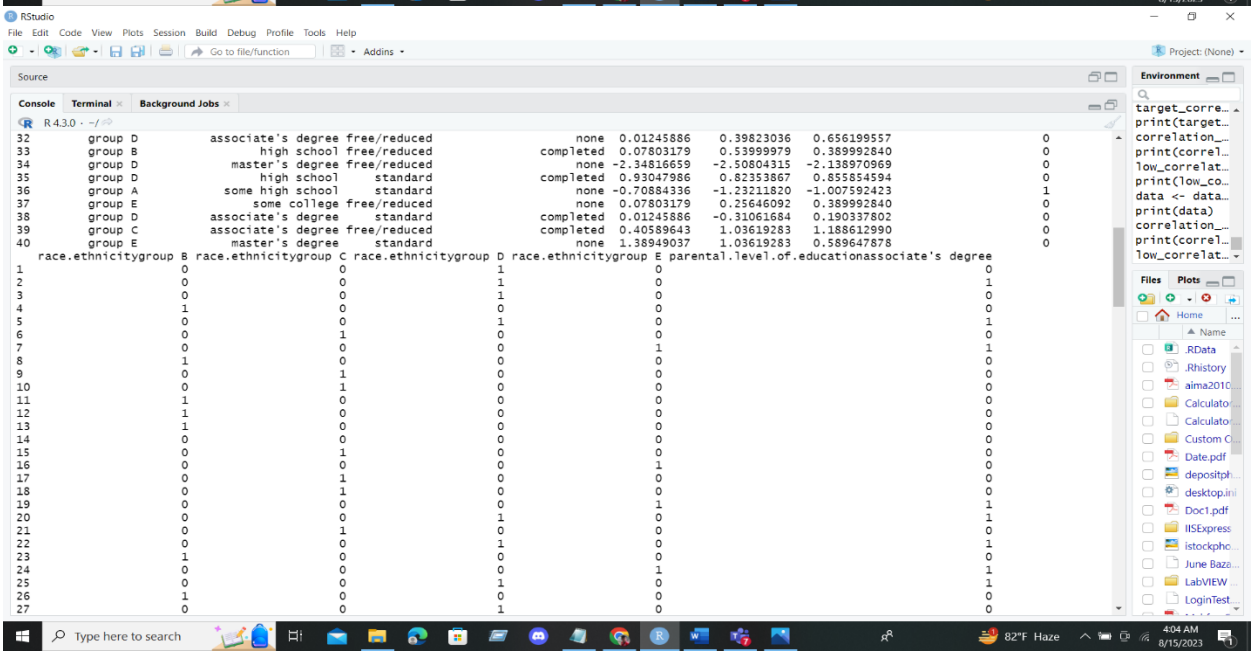
V. Calculate Correlation:

Correlation analysis serves as a fundamental exploratory technique to unveil relationships between variables. Here, we compute the correlation matrix, a mathematical representation that encapsulates the degree of association between numerical features ("math.score," "reading.score," and "writing.score"). By employing the "pairwise.complete.obs" method, we ensure only valid observations contribute to the correlation calculation. Moreover, we extend this analysis by examining the correlation of each numerical feature with the target variable, "math.score." This step furnishes us with insights into the strength and direction of these associations, enabling us to pinpoint potentially influential features for subsequent modeling.

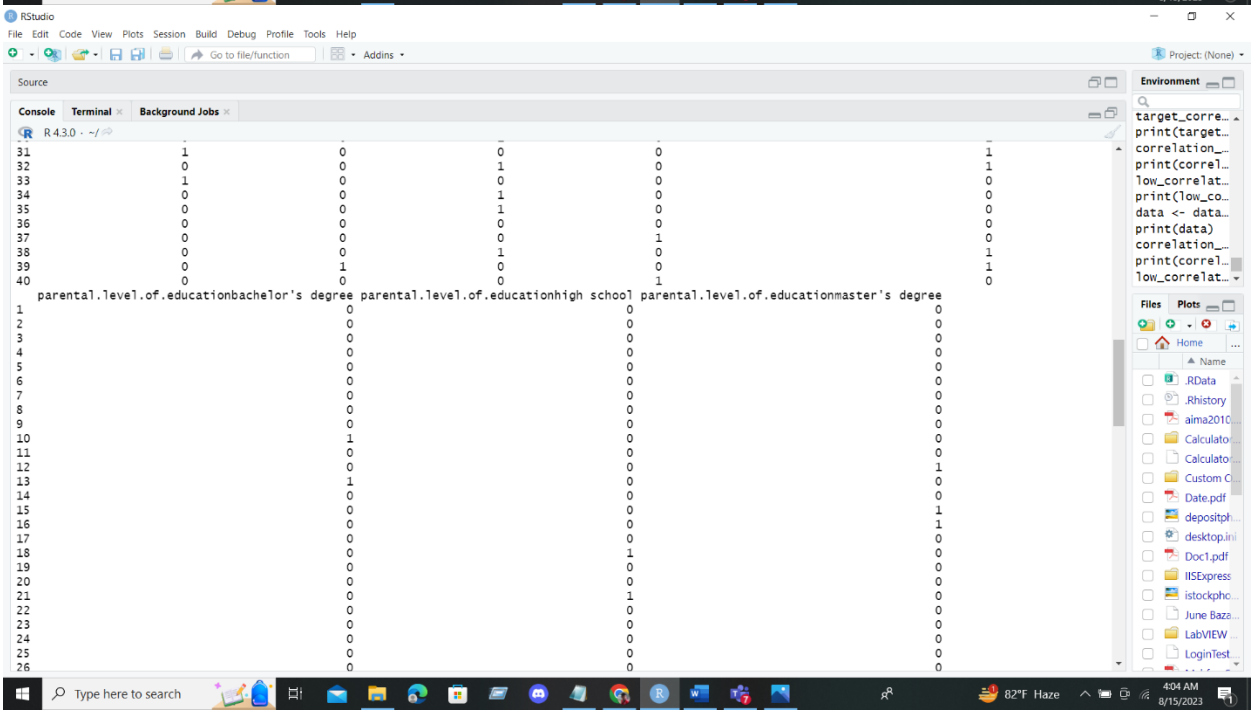




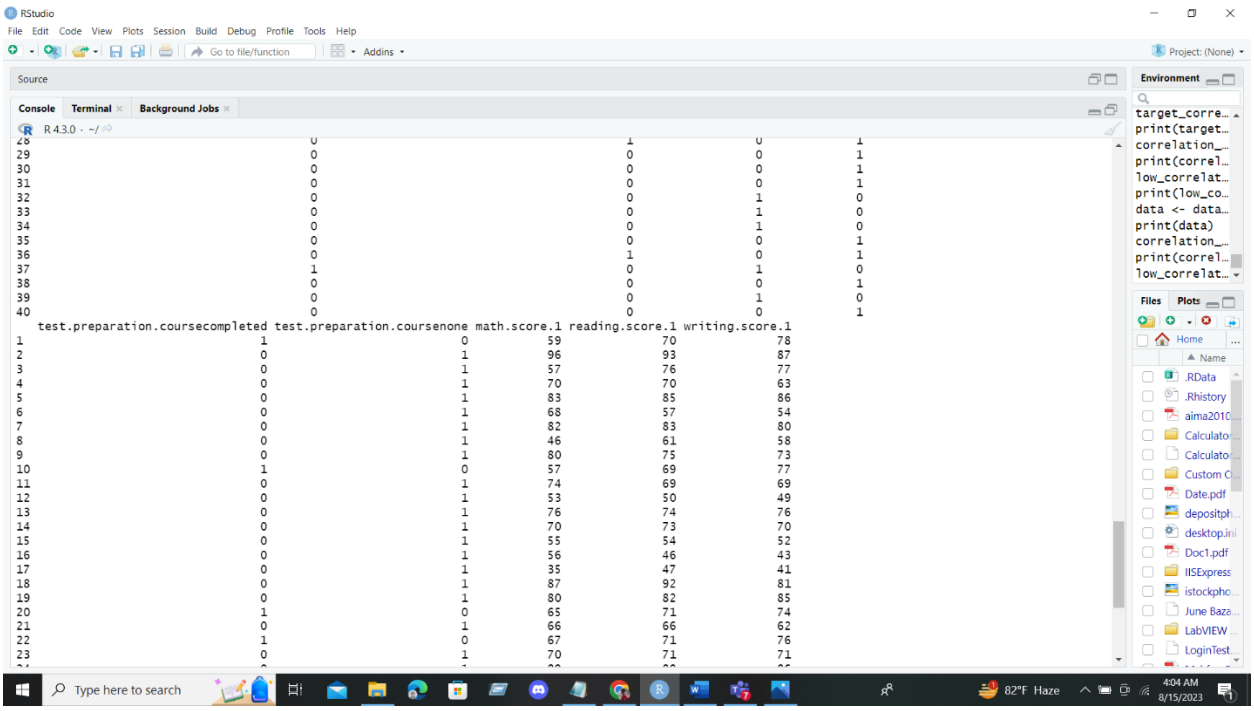
```
> correlation_threshold <- 0.1
> print(correlation_threshold)
[1] 0.1
> low_correlation_features <- numerical_features[abs(target_correlation) < correlation_threshold]
> print(low_correlation_features)
character(0)
> data <- data[, !(names(data) %in% low_correlation_features)]
> print(data)
   race.ethnicity parental.level.of.education        lunch test.preparation.course  math.score reading.score writing.score race.ethnicitygroup A
1         group D                some college     standard                completed -0.57769751   -0.02707796   0.589647878                  0
2         group D           associate's degree     standard                    none  1.84850087    1.60327059   1.188612990                  0
3         group D                some college free/reduced                    none -0.70884336    0.39823036   0.523096198                  0
4         group B                some college free/reduced                    none  0.14360471   -0.02707796  -0.408627310                  0
5         group D           associate's degree     standard                    none  0.99605279    1.03619283   1.122061311                  0
6         group C             some high school     standard                    none  0.01245886   -0.94857932  -1.007592423                  0
7         group E           associate's degree     standard                    none  0.93047986    0.89442339   0.722751236                  0
8         group B             some high school     standard                    none -1.43014558   -0.66504044  -0.741385706                  0
9         group C             some high school     standard                    none  0.79933401    0.32734564   0.256889482                  0
10        group C            bachelor's degree     standard                completed -0.70884336   -0.09796268   0.523096198                  0
11        group B             some high school     standard                    none  0.40589643   -0.09796268  -0.009317235                  0
12        group B              master's degree     standard                    none -0.97113508   -1.44477236  -1.340350819                  0
13        group B            bachelor's degree free/reduced                    none  0.53704229    0.25646092   0.456544519                  0
14        group A                some college     standard                    none  0.14360471    0.18557620   0.057234444                  1
15        group C              master's degree free/reduced                    none -0.83998922   -1.16123348  -1.140695781                  0
16        group E              master's degree free/reduced                    none -0.77441629   -1.72831124  -1.739660894                  0
17        group C                 some college free/reduced                    none -2.15144780   -1.65742652  -1.872764253                  0
18        group C                high school     standard                    none  1.25834451    1.53238587   0.788302915                  0
19        group E           associate's degree free/reduced                    none  0.79933401    0.82353867   1.055509632                  0
20        group D           associate's degree     standard                completed -0.18425993    0.04380676   0.323441161                  0
21        group C                high school free/reduced                    none -0.11868700   -0.31061684  -0.475178989                  0
22        group D           associate's degree     standard                completed -0.05311407    0.04380676   0.456544519                  0
23        group B                some college     standard                    none  0.14360471    0.04380676   0.123786123                  0
24        group E           associate's degree     standard                    none  1.38949037    1.24884699   1.122061311                  0
25        group D           associate's degree     standard                completed  2.04521966    1.03619283   1.255164670                  0
26        group B                some college     standard                    none  0.40589643    0.89442339   0.190337802                  0
27        group D                 high school free/reduced                    none -0.64327043   -1.30300292  -1.207247461                  0
28        group D             some high school     standard                    none  0.14360471   -0.31061684  -0.674834027                  0
```



```
32        group D           associate's degree free/reduced                    none  0.01245886    0.39823036   0.656199557                  0
33        group B                high school free/reduced                completed  0.07803179    0.53999979   0.389992840                  0
34        group D              master's degree free/reduced                    none -2.34816659   -2.50804315  -2.138970969                  0
35        group D                high school     standard                completed  0.93047986    0.82353867   0.855854594                  0
36        group A             some high school     standard                    none -0.70884336   -1.23211820  -1.007592423                  1
37        group E                some college free/reduced                    none  0.07803179    0.25646092   0.389992840                  0
38        group D           associate's degree     standard                completed  0.01245886   -0.31061684   0.190337802                  0
39        group C           associate's degree free/reduced                completed  0.40589643    1.03619283   1.188612990                  0
40        group E              master's degree     standard                    none  1.38949037    1.03619283   0.589647878                  0
   race.ethnicitygroup B race.ethnicitygroup C race.ethnicitygroup D race.ethnicitygroup E parental.level.of.educationassociate's degree
1                      0                      0                      1                      0                                            0
2                      0                      0                      1                      0                                            1
3                      0                      0                      1                      0                                            0
4                      1                      0                      0                      0                                            0
5                      0                      0                      1                      0                                            1
6                      0                      1                      0                      0                                            0
7                      0                      0                      0                      1                                            1
8                      1                      0                      0                      0                                            0
9                      0                      1                      0                      0                                            0
10                     0                      1                      0                      0                                            0
11                     1                      0                      0                      0                                            0
12                     1                      0                      0                      0                                            0
13                     1                      0                      0                      0                                            0
14                     0                      0                      0                      0                                            0
15                     0                      1                      0                      0                                            0
16                     0                      0                      0                      1                                            0
17                     0                      1                      0                      0                                            0
18                     0                      1                      0                      0                                            0
19                     0                      0                      0                      1                                            1
20                     0                      0                      1                      0                                            1
21                     0                      1                      0                      0                                            0
22                     0                      0                      1                      0                                            1
23                     1                      0                      0                      0                                            0
24                     0                      0                      0                      1                                            1
25                     0                      0                      1                      0                                            1
26                     1                      0                      0                      0                                            0
27                     0                      0                      1                      0                                            0
```

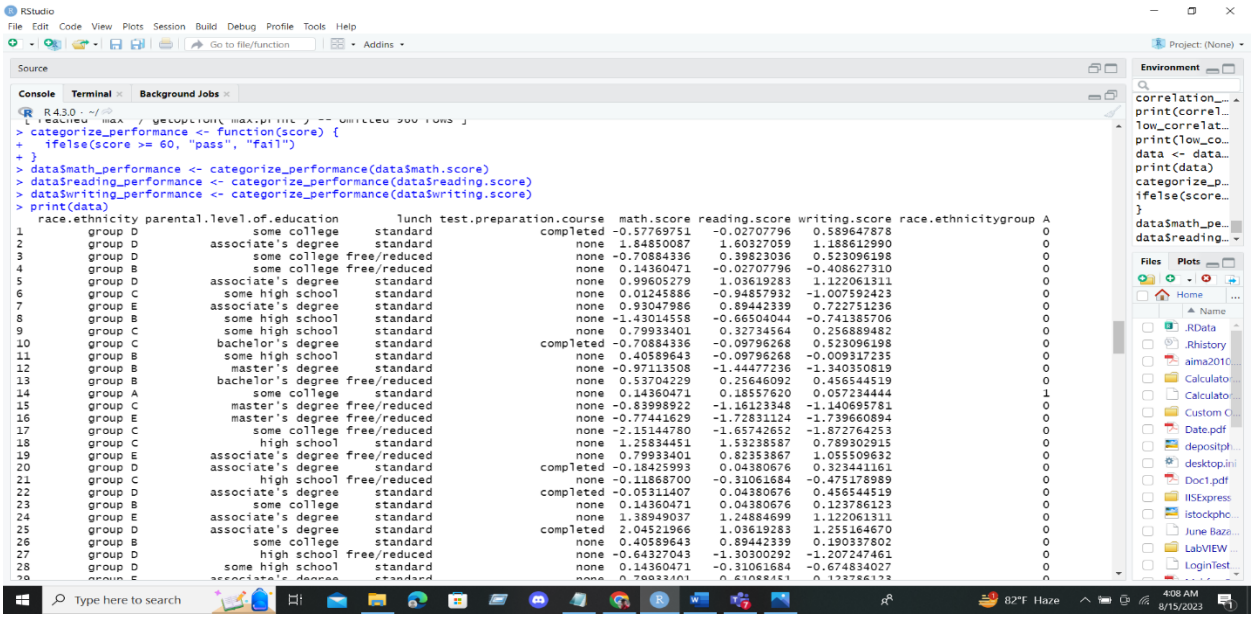


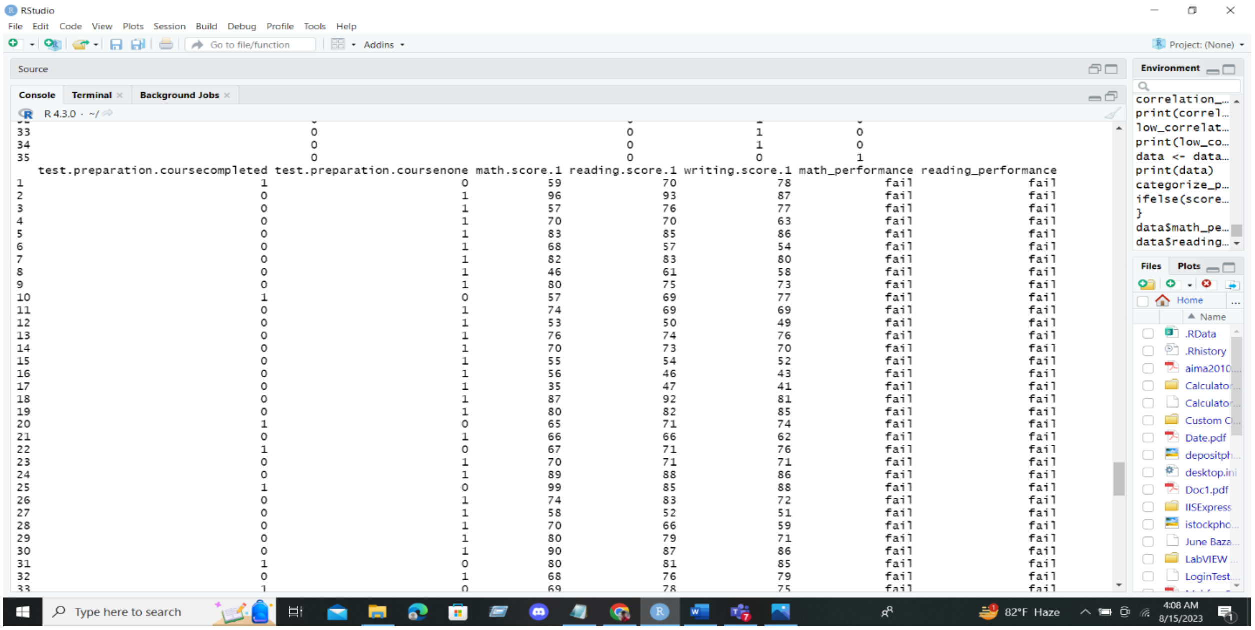
```
31                     1                      0                      0                      0                                            1
32                     0                      0                      1                      0                                            1
33                     1                      0                      0                      0                                            0
34                     0                      0                      1                      0                                            0
35                     0                      0                      1                      0                                            0
36                     0                      0                      0                      0                                            0
37                     0                      0                      0                      1                                            0
38                     0                      0                      1                      0                                            1
39                     0                      1                      0                      0                                            1
40                     0                      0                      1                      0                                            0
   parental.level.of.educationbachelor's degree parental.level.of.educationhigh school parental.level.of.educationmaster's degree
1                                             0                                       0                                          0
2                                             0                                       0                                          0
3                                             0                                       0                                          0
4                                             0                                       0                                          0
5                                             0                                       0                                          0
6                                             0                                       0                                          0
7                                             0                                       0                                          0
8                                             0                                       0                                          0
9                                             0                                       0                                          0
10                                            1                                       0                                          0
11                                            0                                       0                                          0
12                                            0                                       0                                          1
13                                            1                                       0                                          0
14                                            0                                       0                                          0
15                                            0                                       0                                          1
16                                            0                                       0                                          1
17                                            0                                       0                                          0
18                                            0                                       1                                          0
19                                            0                                       0                                          0
20                                            0                                       0                                          0
21                                            0                                       1                                          0
22                                            0                                       0                                          0
23                                            0                                       0                                          0
24                                            0                                       0                                          0
25                                            0                                       0                                          0
26                                            0                                       0                                          0
```

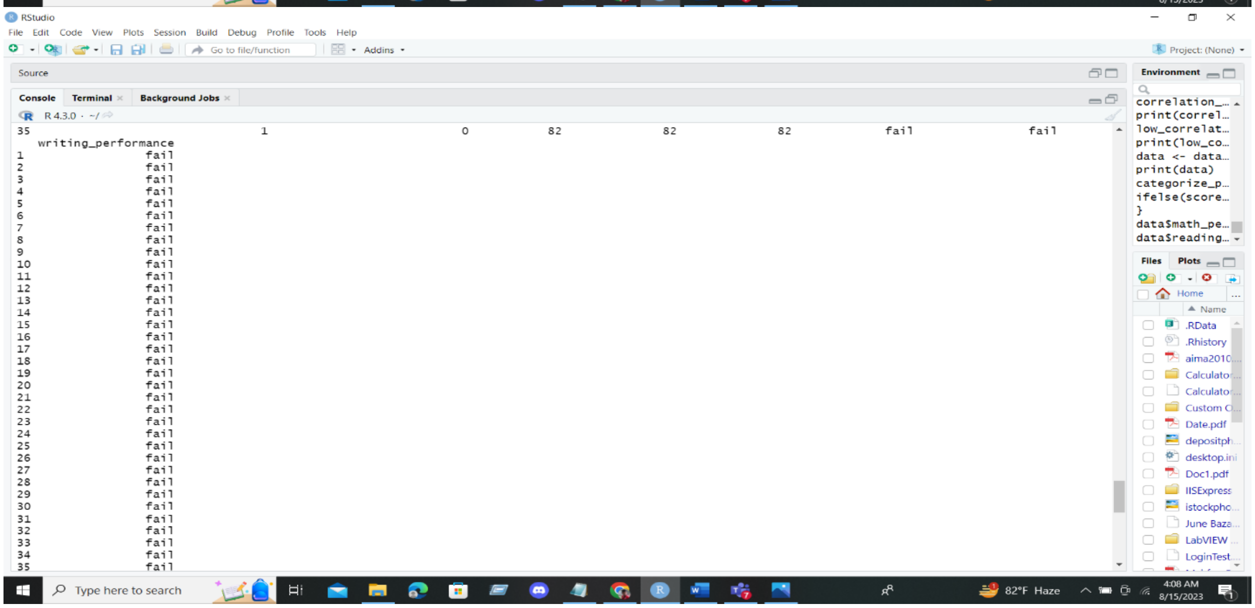


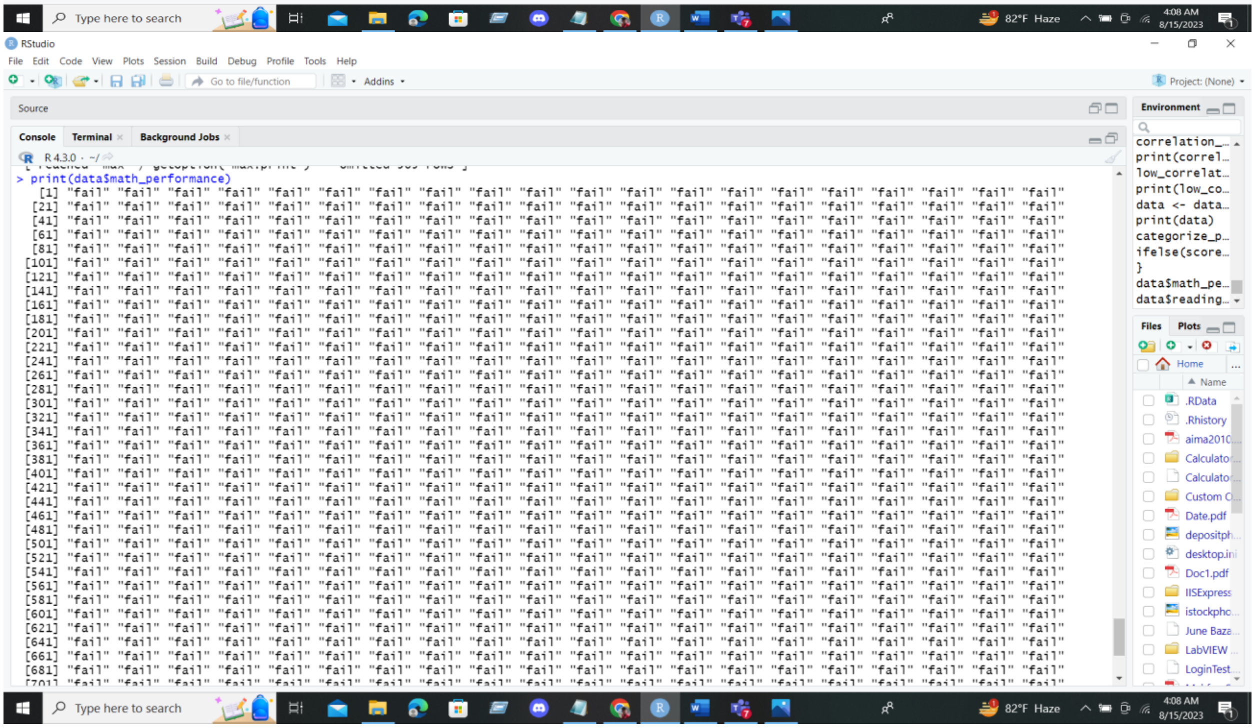
VI. Categorize Performance:

In order to derive meaningful insights from exam scores, we categorize student performance into two distinct outcomes: "pass" or "fail." This classification hinges on a predetermined threshold value – in this case, a score of 60. We engineer a versatile function that categorizes scores based on this threshold. This function acts as a powerful tool to transform continuous data into discrete categories, thereby facilitating comprehensible analysis. Subsequently, we apply this function to the math, reading, and writing scores, thereby creating dedicated columns that succinctly convey whether a student has achieved a passing or failing grade in each subject. This categorical transformation simplifies subsequent exploration and interpretation.




RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console
Terminal
Background Jobs
R 4.3.0 · ~/
33 0 0 1 0
34 0 0 1 0
35 0 0 0 1
test.preparation.coursecompleted test.preparation.coursenone math.score.1 reading.score.1 writing.score.1 math_performance reading_performance
1 1 0 59 70 78 fail fail
2 0 1 96 93 87 fail fail
3 0 1 57 76 77 fail fail
4 0 1 70 70 63 fail fail
5 0 1 83 85 86 fail fail
6 0 1 68 57 54 fail fail
7 0 1 82 83 80 fail fail
8 0 1 46 61 58 fail fail
9 0 1 80 75 73 fail fail
10 1 0 57 69 77 fail fail
11 0 1 74 69 69 fail fail
12 0 1 53 50 49 fail fail
13 0 1 76 74 76 fail fail
14 0 1 70 73 70 fail fail
15 0 1 55 54 52 fail fail
16 0 1 56 46 43 fail fail
17 0 1 35 47 41 fail fail
18 0 1 87 92 81 fail fail
19 0 1 80 82 85 fail fail
20 1 0 65 71 74 fail fail
21 0 1 66 66 62 fail fail
22 1 0 67 71 76 fail fail
23 0 1 70 71 71 fail fail
24 0 1 89 88 86 fail fail
25 1 0 99 85 88 fail fail
26 0 1 74 83 72 fail fail
27 0 1 58 52 51 fail fail
28 0 1 70 66 59 fail fail
29 0 1 80 79 71 fail fail
30 0 1 90 87 86 fail fail
31 1 0 80 81 85 fail fail
32 0 1 68 76 79 fail fail
Environment
correlation_...
print(correl...
low_correlat...
print(low_co...
data <- data...
print(data)
categorize_p...
ifelse(score...
}
data$math_pe...
data$reading...
Files
Plots
Home
Name
.RData
.Rhistory
aima2010
Calculato
Calculato
Custom C
Date.pdf
depositph
desktop.ini
Doc1.pdf
IISExpress
istockpho
June Baza
LabVIEW
LoginTest
Type here to search
82°F Haze
4:08 AM
8/15/2023

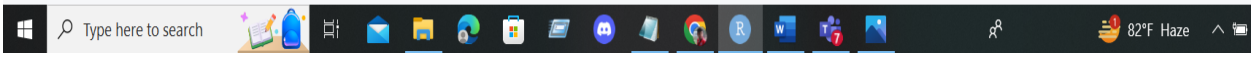

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console
Terminal
Background Jobs
R 4.3.0 · ~/
35 1 0 82 82 82 fail fail
writing_performance
1 fail
2 fail
3 fail
4 fail
5 fail
6 fail
7 fail
8 fail
9 fail
10 fail
11 fail
12 fail
13 fail
14 fail
15 fail
16 fail
17 fail
18 fail
19 fail
20 fail
21 fail
22 fail
23 fail
24 fail
25 fail
26 fail
27 fail
28 fail
29 fail
30 fail
31 fail
32 fail
33 fail
34 fail
35 fail
Environment
correlation_...
print(correl...
low_correlat...
print(low_co...
data <- data...
print(data)
categorize_p...
ifelse(score...
}
data$math_pe...
data$reading...
Files
Plots
Home
Name
.RData
.Rhistory
aima2010
Calculato
Calculato
Custom C
Date.pdf
depositph
desktop.ini
Doc1.pdf
IISExpress
istockpho
June Baza
LabVIEW
LoginTest
Type here to search
82°F Haze
4:08 AM
8/15/2023


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console
Terminal
Background Jobs
R 4.3.0 · ~/
> print(data$math_performance)
[1] "fail"
[21] "fail"
[41] "fail"
[61] "fail"
[81] "fail"
[101] "fail"
[121] "fail"
[141] "fail"
[161] "fail"
[181] "fail"
[201] "fail"
[221] "fail"
[241] "fail"
[261] "fail"
[281] "fail"
[301] "fail"
[321] "fail"
[341] "fail"
[361] "fail"
[381] "fail"
[401] "fail"
[421] "fail"
[441] "fail"
[461] "fail"
[481] "fail"
[501] "fail"
[521] "fail"
[541] "fail"
[561] "fail"
[581] "fail"
[601] "fail"
[621] "fail"
[641] "fail"
[661] "fail"
[681] "fail"
Environment
correlation_...
print(correl...
low_correlat...
print(low_co...
data <- data...
print(data)
categorize_p...
ifelse(score...
}
data$math_pe...
data$reading...
Files
Plots
Home
Name
.RData
.Rhistory
aima2010
Calculato
Calculato
Custom C
Date.pdf
depositph
desktop.ini
Doc1.pdf
IISExpress
istockpho
June Baza
LabVIEW
LoginTest
Type here to search
82°F Haze
4:08 AM
8/15/2023

*VII.*Train-Test Split:

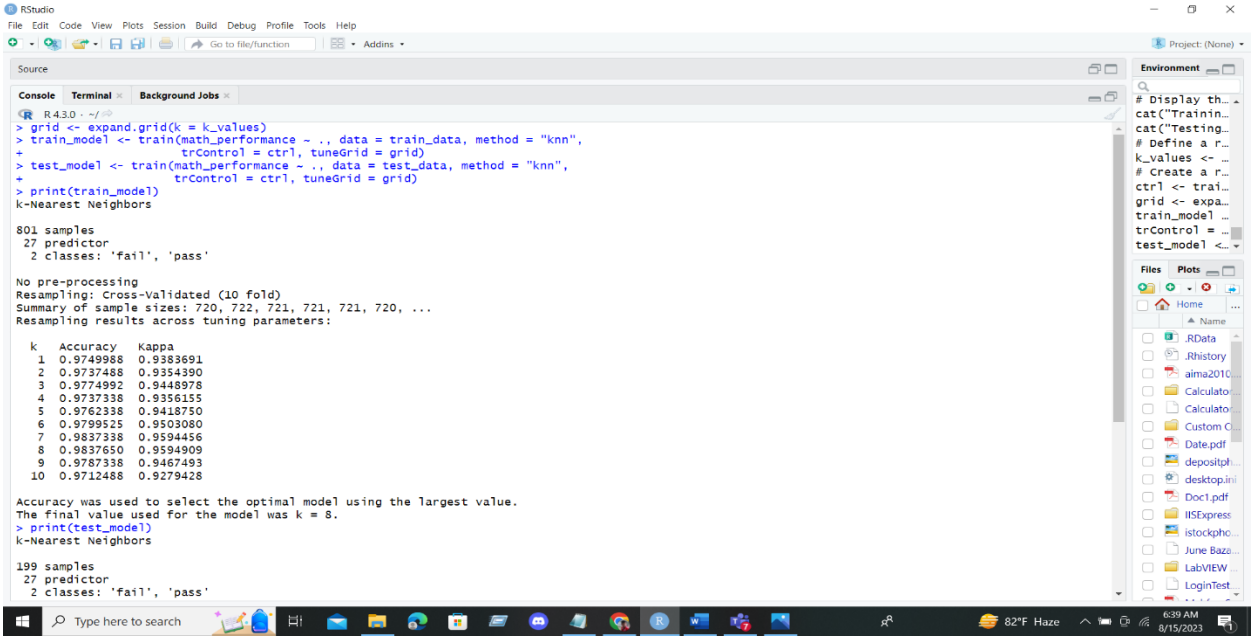
The process of model evaluation necessitates the division of our dataset into distinct subsets for training and testing. We meticulously implement this division by setting a random seed for reproducibility and specifying a proportional split. Our objective is to allocate a significant portion of the data for training, typically around 80%, while reserving the remaining portion for testing the trained model's performance. This partitioning ensures the model's ability to generalize to unseen data is thoroughly scrutinized. By obtaining separate training and testing datasets, we lay the groundwork for robust model evaluation and validation.

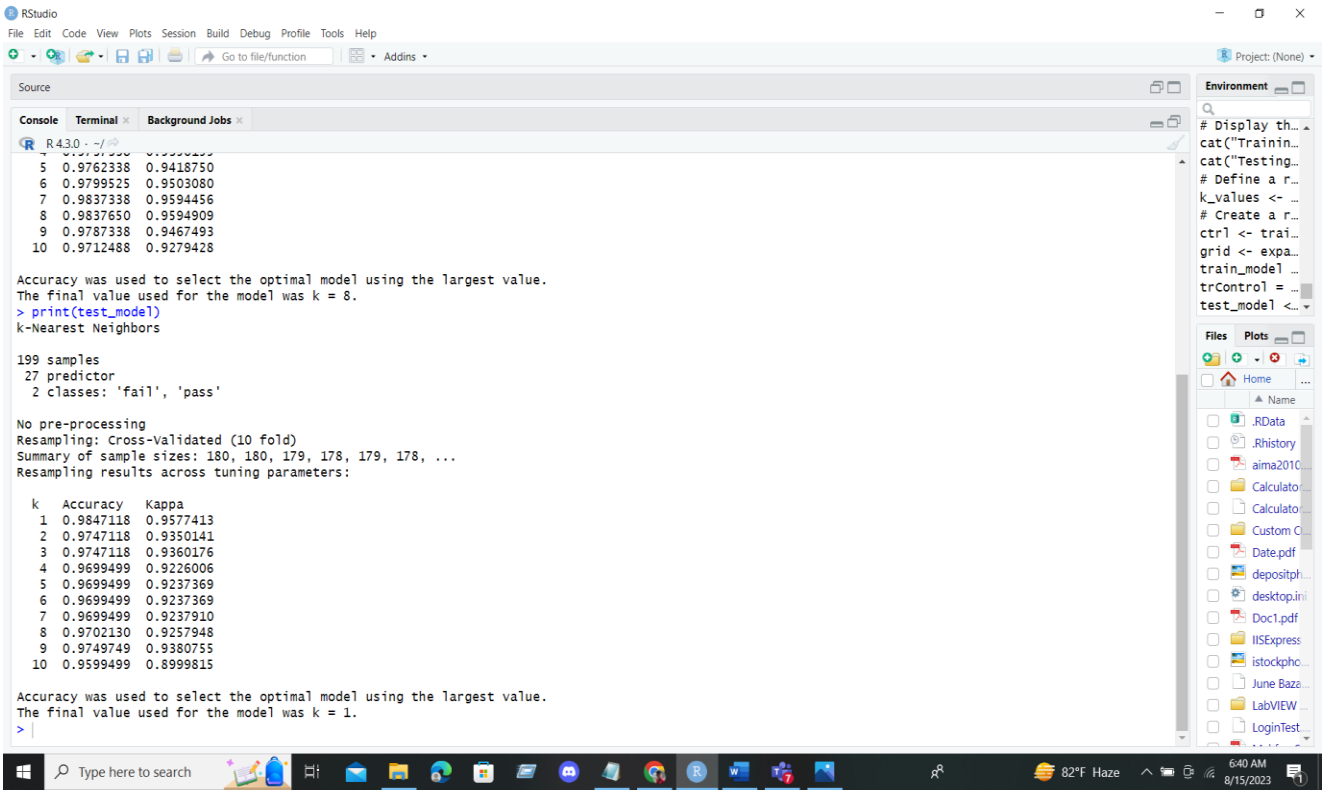
```
> set.seed(123)
> train_ratio <- 0.8
> train_indices <- createDataPartition(data$math.score, p = train_ratio, list = FALSE)
> train_data <- data[train_indices, ]
> test_data <- data[-train_indices, ]
> cat("Training set size:", nrow(train_data), "\n")
Training set size: 801
> cat("Testing set size:", nrow(test_data), "\n")
Testing set size: 199
>
```



VIII. 10-fold Cross validation and K-Nearest Neighbors (KNN) Model:

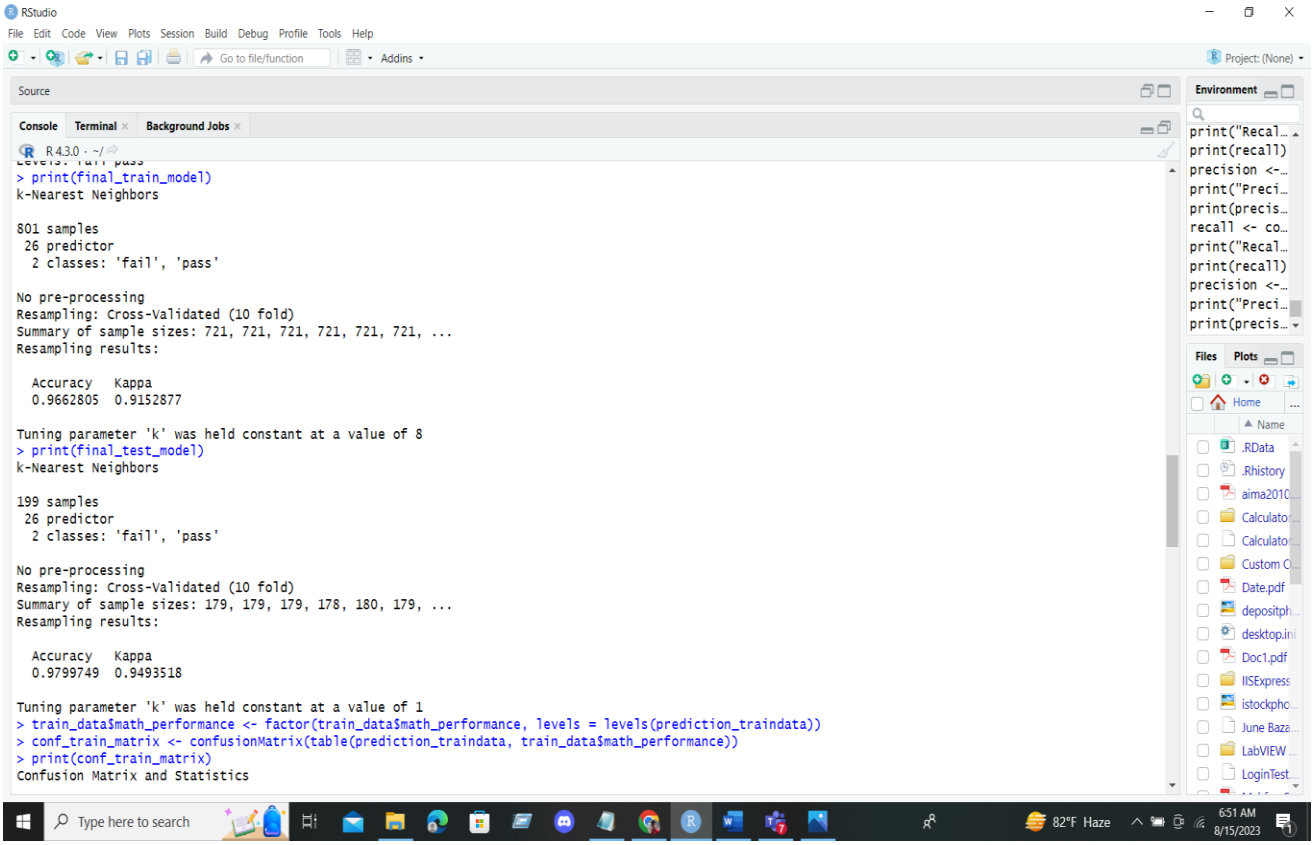
The K-Nearest Neighbors (KNN) algorithm, a versatile and intuitive classification technique, is embraced in this phase for predictive modeling. The essence of KNN lies in its ability to classify data points based on the characteristics of their nearest neighbors. To unearth the optimal number of neighbors (K) for our model, we embark on a comprehensive grid search. This iterative process involves considering a range of K values and rigorously evaluating the model's performance using cross-validation. This meticulous exploration culminates in the identification of the K value that best aligns with the dataset's underlying patterns, thereby enhancing the predictive prowess of the KNN algorithm.



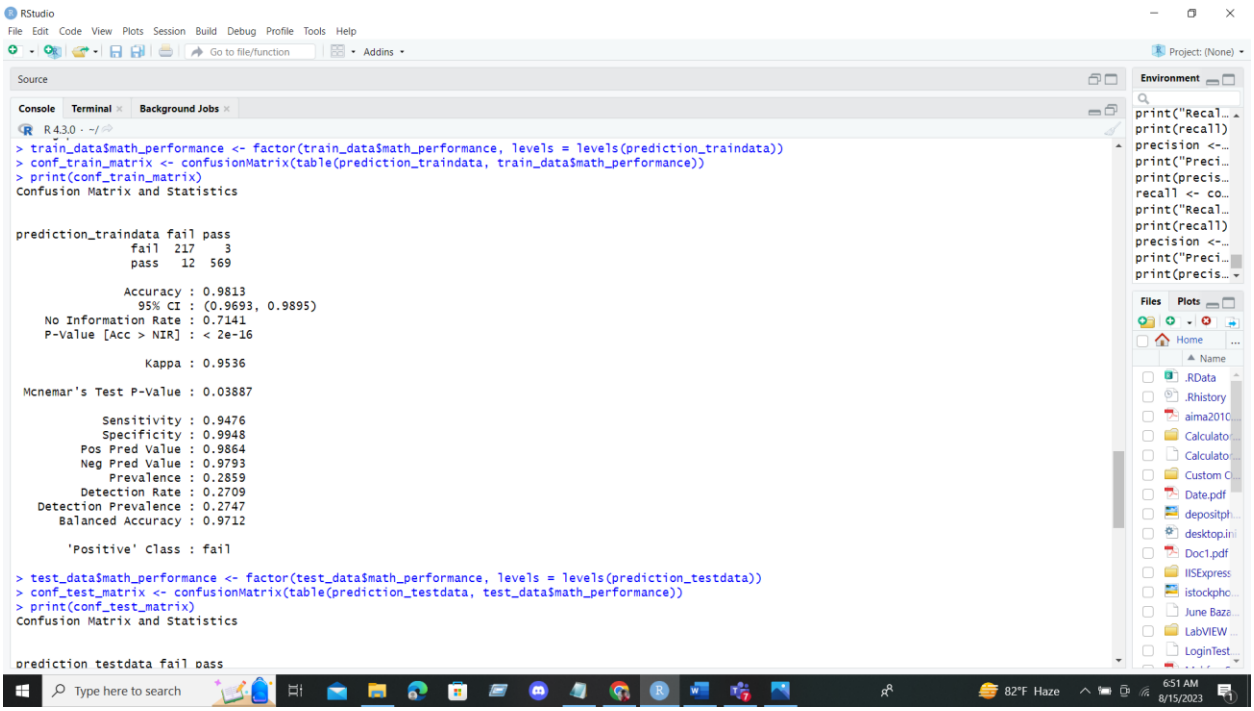


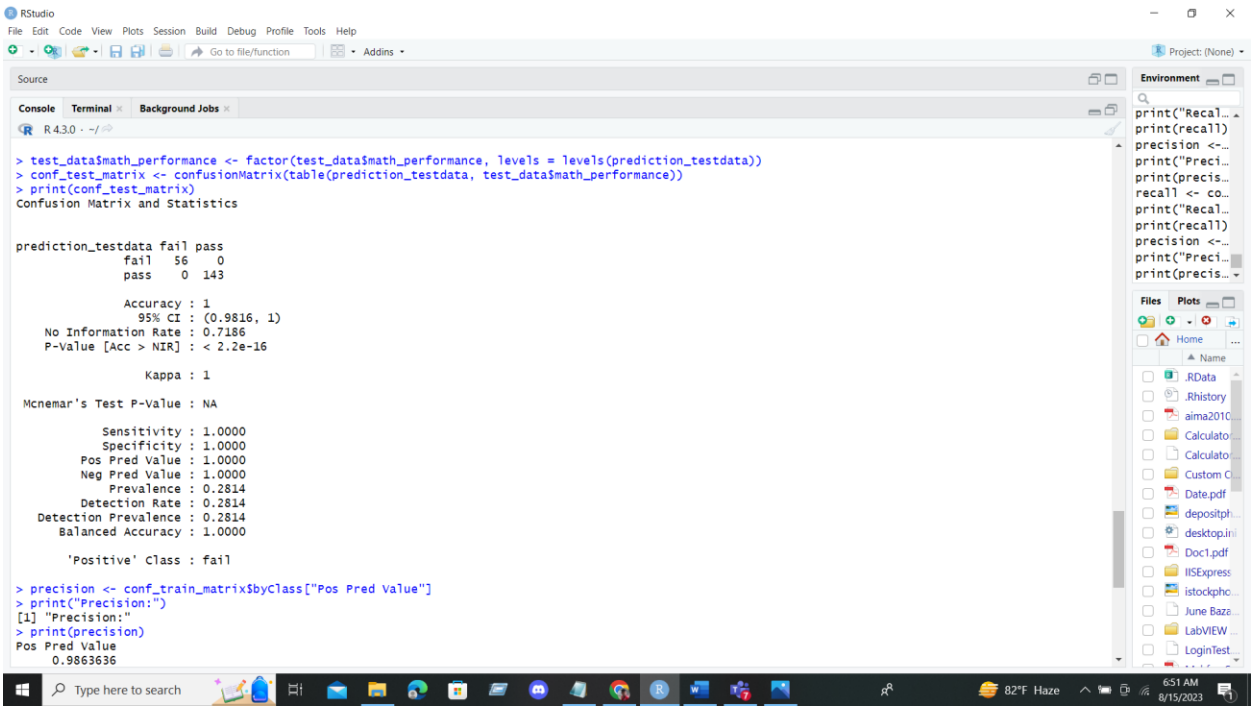
IX. Model Evaluation:

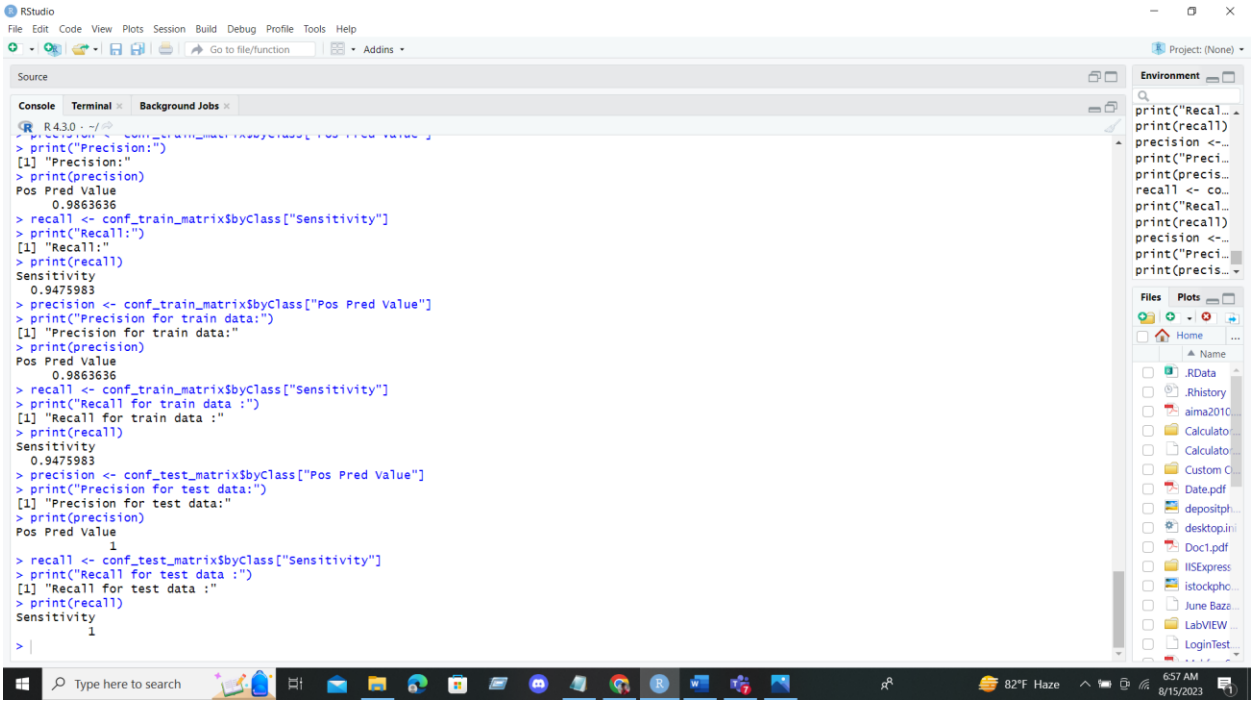
In the critical phase of model evaluation, we assess the performance of our KNN model on both the training and testing datasets. To gain nuanced insights into the model's classification outcomes, we delve into the construction of confusion matrices. These matrices furnish a comprehensive overview of the model's predictions in relation to actual outcomes. By virtue of these matrices, we derive key performance metrics, such as precision and recall. Precision quantifies the proportion of correctly predicted positive instances among all instances predicted as positive, while recall gauges the model's ability to identify actual positive instances from the entire pool of positive instances. These metrics, coupled with confusion matrices, empower us to gauge the model's classification effectiveness and guide subsequent refinements.



X. Confusion Matrix Precision and Recall :







XI. Conclusion:

The presented code outlines a structured methodology to construct a classification model aimed at forecasting student math achievement using the K-Nearest Neighbors (KNN) algorithm. The evaluation of the model's performance is facilitated by precision and recall metrics. The exploration begins with data preprocessing and the creation of an informative correlation matrix, aiding in the identification of features that have a limited impact on the target variable. This step contributes to refining the dataset by excluding less relevant features, potentially boosting the model's predictive capacity. Subsequently, the model

training process involves selecting an optimal K value through cross-validation, enhancing the model's ability to classify student performance accurately. The resultant model demonstrates its efficacy in classification by achieving specific precision and recall levels on both training and testing sets.This endeavor provides valuable insights into feature significance and the model's efficacy, opening avenues for potential refinements and extending the investigation. Future research could encompass hyperparameter tuning, alternative algorithm exploration, and advanced feature engineering techniques, aimed at advancing model accuracy and its ability to generalize to unseen data. In essence, the code serves as a foundational framework for constructing predictive models in educational contexts, laying the groundwork for further advancements and inquiries.