

Sheth L.U.J. & Sir M.V. College

7. Selecting and dropping variables using select() in R. import dataset.

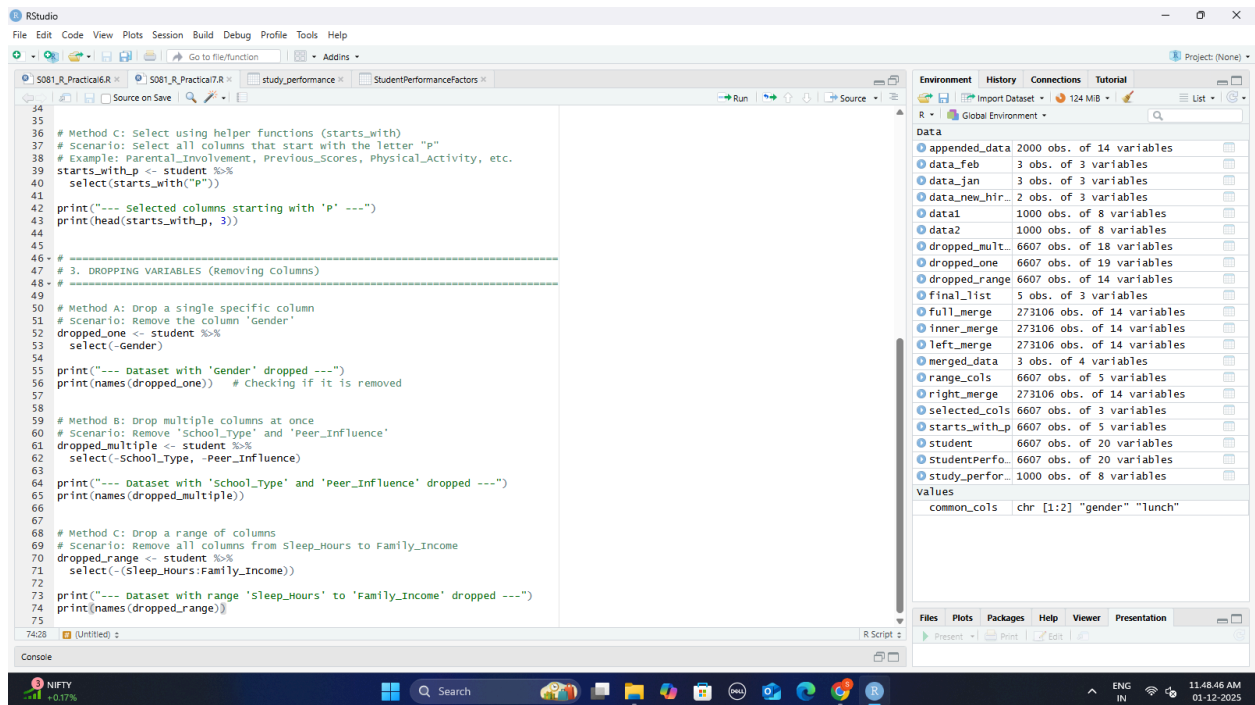
The screenshot displays the RStudio interface. The main editor shows a data frame with 26 columns and 6,607 rows. The columns are: Hours_Studied, Attendance, Parental_Involvement, Access_to_Resources, Extracurricular_Activities, Sleep_Hours, Previous_Scores, Motivation_Level, Internet_Access, Tutoring_Sessions, and Family_Income. The Environment pane on the right shows the loaded data.

The screenshot displays the RStudio interface with R code for importing and selecting variables. The code includes comments and function calls like `library(dplyr)`, `read.csv`, and `select`. The Environment pane on the right shows the resulting data frames.

Name :- Priya Gupta

Roll no. :- S081

Sheth L.U.J. & Sir M.V. College



The screenshot shows the RStudio interface with a script editor on the left and the Environment pane on the right. The script editor contains R code for data manipulation, including selecting columns with `starts_with()` and dropping columns with `select()`. The Environment pane lists various data objects created during the session.

```
# Method C: Select using helper functions (starts_with)
# Scenario: Select all columns that start with the letter "p"
# Examples: Parental_Involvement, Previous_Scores, Physical_Activity, etc.
starts_with_p <- student %>%
  select(starts_with("p"))

print("--- Selected columns starting with 'p' ---")
print(head(starts_with_p, 3))

# 3. DROPPING VARIABLES (Removing columns)

# Method A: Drop a single specific column
# Scenario: Remove the column 'Gender'
dropped_one <- student %>%
  select(-Gender)

print("--- Dataset with 'Gender' dropped ---")
print(names(dropped_one)) # checking if it is removed

# Method B: Drop multiple columns at once
# Scenario: Remove 'School_Type' and 'Peer_Influence'
dropped_multiple <- student %>%
  select(-School_Type, -Peer_Influence)

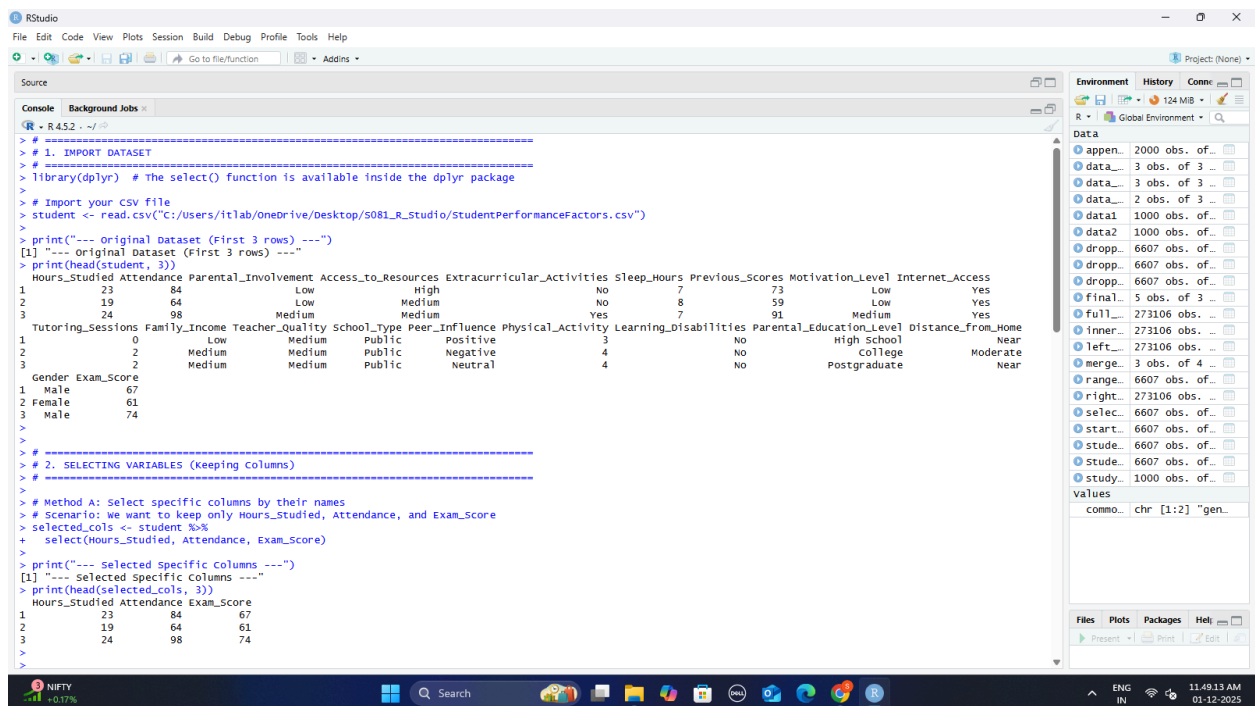
print("--- Dataset with 'School_Type' and 'Peer_Influence' dropped ---")
print(names(dropped_multiple))

# Method C: Drop a range of columns
# Scenario: Remove all columns from Sleep_Hours to Family_Income
dropped_range <- student %>%
  select(-(Sleep_Hours:Family_Income))

print("--- Dataset with range 'Sleep_Hours' to 'Family_Income' dropped ---")
print(names(dropped_range))
```

The Environment pane shows the following data objects:

Object	Size	Variables
appended_data	2000 obs.	14 variables
data_feb	3 obs.	3 variables
data_jan	3 obs.	3 variables
data_new_hir	2 obs.	3 variables
data1	1000 obs.	8 variables
data2	1000 obs.	8 variables
dropped_mult	6607 obs.	18 variables
dropped_one	6607 obs.	19 variables
dropped_range	6607 obs.	14 variables
final_list	5 obs.	3 variables
full_merge	273106 obs.	14 variables
inner_merge	273106 obs.	14 variables
left_merge	273106 obs.	14 variables
merged_data	3 obs.	4 variables
range_cols	6607 obs.	5 variables
right_merge	273106 obs.	14 variables
selected_cols	6607 obs.	3 variables
starts_with_p	6607 obs.	5 variables
student	6607 obs.	20 variables
studentPerfo	6607 obs.	20 variables
study_perfor	1000 obs.	8 variables



The screenshot shows the RStudio interface with a script editor on the left and the Console pane on the right. The script editor contains R code for importing a CSV file and selecting specific columns. The Console pane shows the output of the code, including the original dataset and the selected columns.

```
# 1. IMPORT DATASET
library(dplyr) # The select() function is available inside the dplyr package

# Import your csv file
student <- read.csv("C:/Users/itlab/OneDrive/Desktop/S081_R_Studio/StudentPerformanceFactors.csv")

print("--- Original Dataset (First 3 rows) ---")
print(head(student, 3))

# 2. SELECTING VARIABLES (Keeping columns)

# Method A: Select specific columns by their names
# Scenario: We want to keep only Hours_Studied, Attendance, and Exam_Score
selected_cols <- student %>%
  select(Hours_Studied, Attendance, Exam_Score)

print("--- Selected specific columns ---")
print(head(selected_cols, 3))
```

The Console output shows the following data:

```
Hours_Studied Attendance Parental_Involvement Access_to_Resources Extracurricular_Activities Sleep_Hours Previous_Scores Motivation_Level Internet_Access
1 23 84 Low High No 7 73 Low Yes
2 19 64 Low Medium Medium No 8 59 Low Yes
3 24 98 Medium Medium Medium Yes 7 91 Medium Yes

Tutoring_Sessions Family_Income Teacher_Quality School_Type Peer_Influence Physical_Activity Learning_Disabilities Parental_Education_Level Distance_From_Home
1 0 Low Medium Public Positive 3 No High School Near
2 2 Medium Medium Public Negative 4 No college Moderate
3 2 Medium Medium Public Neutral 4 No Postgraduate Near

Gender Exam_Score
1 Male 67
2 Female 61
3 Male 74
```

Name :- Priya Gupta

Roll no. :- S081

Sheth L.U.J. & Sir M.V. College

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Background Jobs
R - R 4.5.2 - ~/
1 23 84 67
2 19 64 61
3 24 98 74
>
> # Method B: Select a continuous range of adjacent columns
> # Scenario: Select from the first column to the 5th column
> # Example: Selecting columns from Hours_Studied to Extracurricular_Activities
> range_cols <- student %>%
+   select(Hours_Studied:Extracurricular_Activities)
>
> print("--- Selected Range of columns ---")
[1] "--- Selected Range of columns ---"
> print(head(range_cols, 3))
  Hours_Studied Attendance Parental_Involvement Access_to_Resources Extracurricular_Activities
1             23             84                 Low                High                No
2             19             64                 Low                Medium             No
3             24             98             Medium             Medium             Yes
>
> # Method C: select using helper functions (starts_with)
> # Scenario: Select all columns that start with the letter "p"
> # Example: Parental_Involvement, Previous_Scores, Physical_Activity, etc.
> starts_with_p <- student %>%
+   select(starts_with("p"))
>
> print("--- Selected columns starting with 'p' ---")
[1] "--- Selected columns starting with 'p' ---"
> print(head(starts_with_p, 3))
  Parental_Involvement Previous_Scores Peer_Influence Physical_Activity Parental_Education_Level
1             Low             73             Positive             3             High School
2             Low             59             Negative             4             college
3             Medium            91             Neutral             4             Postgraduate
>
> # =====
> # 3. DROPPING VARIABLES (Removing columns)
> # =====
> # Method A: Drop a single specific column
> # Scenario: Remove the column 'gender'
> dropped_one <- student %>%
+   select(-gender)
>
> print("--- Dataset with 'Gender' dropped ---")
[1] "--- Dataset with 'Gender' dropped ---"
> print(names(dropped_one)) # checking if it is removed
[1] "Hours_Studied"      "Attendance"         "Parental_Involvement" "Access_to_Resources" "Extracurricular_Activities"
[6] "Sleep_Hours"        "Previous_Scores"    "Motivation_Level"    "Internet_Access"    "Tutoring_Sessions"
[11] "Family_Income"      "Teacher_Quality"    "School_Type"         "Peer_Influence"     "Physical_Activity"
[16] "Learning_Disabilities" "Parental_Education_Level" "Distance_from_Home" "Exam_Score"         "Physical_Activity"
>
> # Method B: Drop multiple columns at once
> # Scenario: Remove 'School_Type' and 'Peer_Influence'
> dropped_multiple <- student %>%
+   select(-School_Type, -Peer_Influence)
>
> print("--- dataset with 'School_Type' and 'Peer_Influence' dropped ---")
[1] "--- dataset with 'School_Type' and 'Peer_Influence' dropped ---"
> print(names(dropped_multiple))
[1] "Hours_Studied"      "Attendance"         "Parental_Involvement" "Access_to_Resources" "Extracurricular_Activities"
[6] "Sleep_Hours"        "Previous_Scores"    "Motivation_Level"    "Internet_Access"    "Tutoring_Sessions"
[11] "Family_Income"      "Teacher_Quality"    "Physical_Activity"   "Learning_Disabilities" "Parental_Education_Level"
[16] "Distance_from_Home" "Gender"             "Exam_Score"         "Physical_Activity"   "Learning_Disabilities"
>
> # Method C: Drop a range of columns
> # Scenario: Remove all columns from Sleep_Hours to Family_Income
> dropped_range <- student %>%
+   select(-(Sleep_Hours:Family_Income))
>
> print("--- dataset with range 'Sleep_Hours' to 'Family_Income' dropped ---")
[1] "--- dataset with range 'Sleep_Hours' to 'Family_Income' dropped ---"
> print(names(dropped_range))
[1] "Hours_Studied"      "Attendance"         "Parental_Involvement" "Access_to_Resources" "Extracurricular_Activities"
[6] "Teacher_Quality"    "School_Type"        "Peer_Influence"     "Physical_Activity"   "Learning_Disabilities"
[11] "Parental_Education_Level" "Distance_from_Home" "Gender"             "Exam_Score"         "Physical_Activity"
>
> |
```

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console Background Jobs
R - R 4.5.2 - ~/
> # =====
> # Method A: Drop a single specific column
> # Scenario: Remove the column 'gender'
> dropped_one <- student %>%
+   select(-gender)
>
> print("--- Dataset with 'Gender' dropped ---")
[1] "--- Dataset with 'Gender' dropped ---"
> print(names(dropped_one)) # checking if it is removed
[1] "Hours_Studied"      "Attendance"         "Parental_Involvement" "Access_to_Resources" "Extracurricular_Activities"
[6] "Sleep_Hours"        "Previous_Scores"    "Motivation_Level"    "Internet_Access"    "Tutoring_Sessions"
[11] "Family_Income"      "Teacher_Quality"    "School_Type"         "Peer_Influence"     "Physical_Activity"
[16] "Learning_Disabilities" "Parental_Education_Level" "Distance_from_Home" "Exam_Score"         "Physical_Activity"
>
> # Method B: Drop multiple columns at once
> # Scenario: Remove 'School_Type' and 'Peer_Influence'
> dropped_multiple <- student %>%
+   select(-School_Type, -Peer_Influence)
>
> print("--- dataset with 'School_Type' and 'Peer_Influence' dropped ---")
[1] "--- dataset with 'School_Type' and 'Peer_Influence' dropped ---"
> print(names(dropped_multiple))
[1] "Hours_Studied"      "Attendance"         "Parental_Involvement" "Access_to_Resources" "Extracurricular_Activities"
[6] "Sleep_Hours"        "Previous_Scores"    "Motivation_Level"    "Internet_Access"    "Tutoring_Sessions"
[11] "Family_Income"      "Teacher_Quality"    "Physical_Activity"   "Learning_Disabilities" "Parental_Education_Level"
[16] "Distance_from_Home" "Gender"             "Exam_Score"         "Physical_Activity"   "Learning_Disabilities"
>
> # Method C: Drop a range of columns
> # Scenario: Remove all columns from Sleep_Hours to Family_Income
> dropped_range <- student %>%
+   select(-(Sleep_Hours:Family_Income))
>
> print("--- dataset with range 'Sleep_Hours' to 'Family_Income' dropped ---")
[1] "--- dataset with range 'Sleep_Hours' to 'Family_Income' dropped ---"
> print(names(dropped_range))
[1] "Hours_Studied"      "Attendance"         "Parental_Involvement" "Access_to_Resources" "Extracurricular_Activities"
[6] "Teacher_Quality"    "School_Type"        "Peer_Influence"     "Physical_Activity"   "Learning_Disabilities"
[11] "Parental_Education_Level" "Distance_from_Home" "Gender"             "Exam_Score"         "Physical_Activity"
>
> |
```

Name :- Priya Gupta

Roll no. :- S081