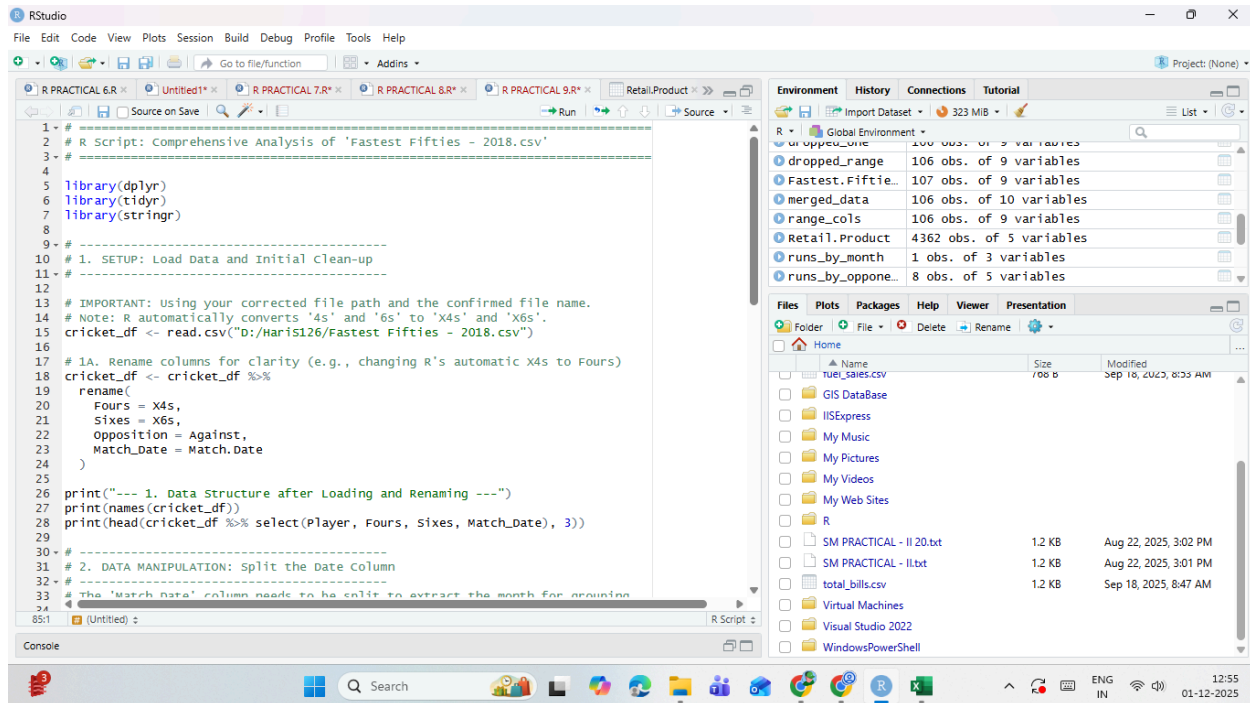


MVLU COLLEGE PRACTICAL NO.9

Aim: Performing text manipulation using `str_sub()`, `str_split()` (R). import dataset

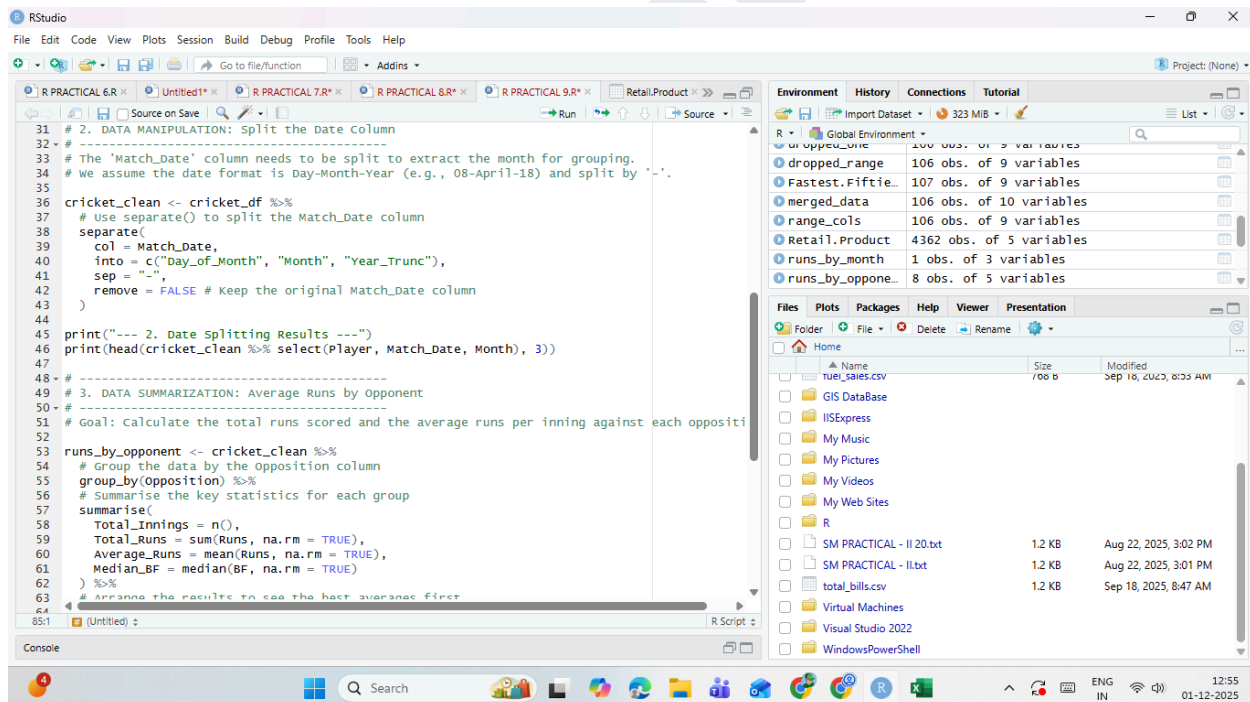
Code:



```
1 # R Script: Comprehensive Analysis of 'Fastest Fifties - 2018.csv'
2 #
3 #
4
5 library(dplyr)
6 library(tidyr)
7 library(stringr)
8
9 #-----
10 # 1. SETUP: Load Data and Initial Clean-up
11 #
12
13 # IMPORTANT: Using your corrected file path and the confirmed file name.
14 # Note: R automatically converts '4s' and '6s' to 'X4s' and 'X6s'.
15 cricket_df <- read.csv("D:/Haris126/Fastest Fifties - 2018.csv")
16
17 # 1A. Rename columns for clarity (e.g., changing R's automatic X4s to Fours)
18 cricket_df <- cricket_df %>%
19   rename(
20     Fours = X4s,
21     Sixes = X6s,
22     Opposition = Against,
23     Match_Date = Match.Date
24   )
25
26 print("--- 1. Data Structure after Loading and Renaming ---")
27 print(names(cricket_df))
28 print(head(cricket_df %>% select(Player, Fours, Sixes, Match_Date), 3))
29
30 #-----
31 # 2. DATA MANIPULATION: Split the Date Column
32 #
33 # The 'Match_Date' column needs to be split to extract the month for grouping
34 # We assume the date format is Day-Month-Year (e.g., 08-April-18) and split by '-'.
```

Environment pane:

Object	Class	Size	Modified
Global Environment			
cricket_df	data.frame	107 obs. of 9 variables	
dropped_range	data.frame	106 obs. of 9 variables	
Fastest.Fiftie...	data.frame	107 obs. of 9 variables	
merged_data	data.frame	106 obs. of 10 variables	
range_cols	data.frame	106 obs. of 9 variables	
Retail.Product	data.frame	4362 obs. of 5 variables	
runs_by_month	data.frame	1 obs. of 3 variables	
runs_by_oppone...	data.frame	8 obs. of 5 variables	



```
35
36 cricket_clean <- cricket_df %>%
37   # Use separate() to split the Match_Date column
38   separate(
39     col = Match_Date,
40     into = c("Day_of_Month", "Month", "Year_Trunc"),
41     sep = "-",
42     remove = FALSE # Keep the original Match_Date column
43   )
44
45 print("--- 2. Date Splitting Results ---")
46 print(head(cricket_clean %>% select(Player, Match_Date, Month), 3))
47
48 #-----
49 # 3. DATA SUMMARIZATION: Average Runs by Opponent
50 #
51 # Goal: calculate the total runs scored and the average runs per inning against each oppositi
52
53 runs_by_opponent <- cricket_clean %>%
54   # Group the data by the opposition column
55   group_by(opposition) %>%
56   # Summarise the key statistics for each group
57   summarise(
58     Total_Innings = n(),
59     Total_Runs = sum(Runs, na.rm = TRUE),
60     Average_Runs = mean(Runs, na.rm = TRUE),
61     Median_SF = median(SF, na.rm = TRUE)
62   ) %>%
63   # Arrange the results to see the best averages first
64   arrange(desc(Average_Runs))
65
66 print(head(runs_by_opponent, 5))
```

Environment pane:

Object	Class	Size	Modified
Global Environment			
cricket_clean	data.frame	107 obs. of 9 variables	
dropped_range	data.frame	106 obs. of 9 variables	
Fastest.Fiftie...	data.frame	107 obs. of 9 variables	
merged_data	data.frame	106 obs. of 10 variables	
range_cols	data.frame	106 obs. of 9 variables	
Retail.Product	data.frame	4362 obs. of 5 variables	
runs_by_month	data.frame	1 obs. of 3 variables	
runs_by_oppone...	data.frame	8 obs. of 5 variables	

MVLU COLLEGE PRACTICAL NO.9

```
54 # Group the data by the opposition column
55 group_by(opposition) %>%
56 # Summarise the key statistics for each group
57 summarise(
58   Total_Innings = n(),
59   Total_Runs = sum(Runs, na.rm = TRUE),
60   Average_Runs = mean(Runs, na.rm = TRUE),
61   Median_BF = median(BF, na.rm = TRUE)
62 ) %>%
63 # Arrange the results to see the best averages first
64 arrange(desc(Average_Runs))
65
66 print("--- 3. Summarized Runs by Opposition ---")
67 print(runs_by_opponent)
68
69 # -----
70 # 4. BONUS: Highest Scoring Month
71 # -----
72 # Goal: Find the total runs scored in each month.
73
74 runs_by_month <- cricket_clean %>%
75 # Group by the extracted Month column
76 group_by(Month) %>%
77 summarise(
78   Total_Runs_in_Month = sum(Runs, na.rm = TRUE),
79   Total_Fifties = n()
80 ) %>%
81 arrange(desc(Total_Runs_in_Month))
82
83 print("--- 4. Total Runs Scored by Month ---")
84 print(runs_by_month)
85
86
```

Output:

```
> library(strings)
> # IMPORTANT: using your corrected file path and the confirmed file name.
> # Note: R automatically converts '4s' and '6s' to 'X4s' and 'X6s'.
> cricket_df <- read.csv("D:/Harisi26/Fastest Fifties - 2018.csv")
> # 1A. Rename columns for clarity (e.g., changing R's automatic X4s to Fours)
> cricket_df <- cricket_df %>%
+   rename(
+     Fours = X4s,
+     Sixes = X6s,
+     Opposition = Against,
+     Match_Date = Match.Date
+   )
> print("--- 1. Data Structure after Loading and Renaming ---")
[1] "--- 1. Data Structure after Loading and Renaming ---"
> print(names(cricket_df))
[1] "Pos"      "Player"   "Runs"     "BF"       "Fours"    "Sixes"    "Opposition"
[8] "Venue"    "Match_Date"
> print(head(cricket_df %>% select(Player, Fours, Sixes, Match_Date), 3))
  Player Fours Sixes Match_Date
1  KL Rahul      6    4 08 April 2018
2  Ishan Kishan  5    6 09 May 2018
3  Sunil Narine  4    5 08 April 2018
> cricket_clean <- cricket_df %>%
+   # use separate() to split the Match_Date column
+   separate(
+     col = Match_Date,
+     into = c("Day_of_Month", "Month", "Year_Trunc"),
+     sep = "-",
+     remove = FALSE # Keep the original Match_Date column
+   )
warning message:
Expected 3 pieces. Missing pieces filled with `NA` in 106 rows [1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

Hariprasad Vishwakarma
S126
SYCS

MVLU COLLEGE PRACTICAL NO.9

```
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 - R4.1.2 ~ /
warning message:
Expected 3 pieces. Missing pieces filled with 'NA' in 106 rows [1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].

> print("--- 2. Date Splitting Results ---")
[1] "--- 2. Date Splitting Results ---"
> print(head(cricket_clean %>% select(Player, Match_Date, Month), 3))
  Player Match_Date Month
1  KL Rahul 08 April 2018 <NA>
2 Ishan Kishan 09 May 2018 <NA>
3 Sunil Narine 08 April 2018 <NA>
> runs_by_opponent <- cricket_clean %>%
+ # Group the data by the Opposition column
+ group_by(Opposition) %>%
+ # Summarise the key statistics for each group
+ summarise(
+   Total_Innings = n(),
+   Total_Runs = sum(Runs, na.rm = TRUE),
+   Average_Runs = mean(Runs, na.rm = TRUE),
+   Median_BF = median(BF, na.rm = TRUE)
+ ) %>%
+ # Arrange the results to see the best averages first
+ arrange(desc(Average_Runs))
> print("--- 3. Summarized Runs by Opposition ---")
[1] "--- 3. Summarized Runs by Opposition ---"
> print(runs_by_opponent)
# A tibble: 8 x 5
  Opposition Total_Innings Total_Runs Average_Runs Median_BF
  <chr>      <int>      <int>      <dbl>      <dbl>
1 SRH        13      1025      78.8       33
2 MI         10       734      73.4       35
3 RCB        15     1042      69.5       32
4 RR         14       949      67.8       32.5
5 CSK        14       934      66.7       31.5
```

```
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 - R4.1.2 ~ /
+ ) %>%
+ # Arrange the results to see the best averages first
+ arrange(desc(Average_Runs))
> print("--- 3. Summarized Runs by Opposition ---")
[1] "--- 3. Summarized Runs by Opposition ---"
> print(runs_by_opponent)
# A tibble: 8 x 5
  Opposition Total_Innings Total_Runs Average_Runs Median_BF
  <chr>      <int>      <int>      <dbl>      <dbl>
1 SRH        13      1025      78.8       33
2 MI         10       734      73.4       35
3 RCB        15     1042      69.5       32
4 RR         14       949      67.8       32.5
5 CSK        14       934      66.7       31.5
6 DC         13       866      66.6       28
7 KKR        12       738      61.5       30
8 PBKS       15       913      60.9       36
> runs_by_month <- cricket_clean %>%
+ # Group by the extracted Month column
+ group_by(Month) %>%
+ summarise(
+   Total_Runs_in_Month = sum(Runs, na.rm = TRUE),
+   Total_Fifties = n()
+ ) %>%
+ arrange(desc(Total_Runs_in_Month))
> print("--- 4. Total Runs Scored by Month ---")
[1] "--- 4. Total Runs Scored by Month ---"
> print(runs_by_month)
# A tibble: 1 x 3
  Month Total_Runs_in_Month Total_Fifties
  <chr>      <int>      <int>
1 NA         2201         106
```