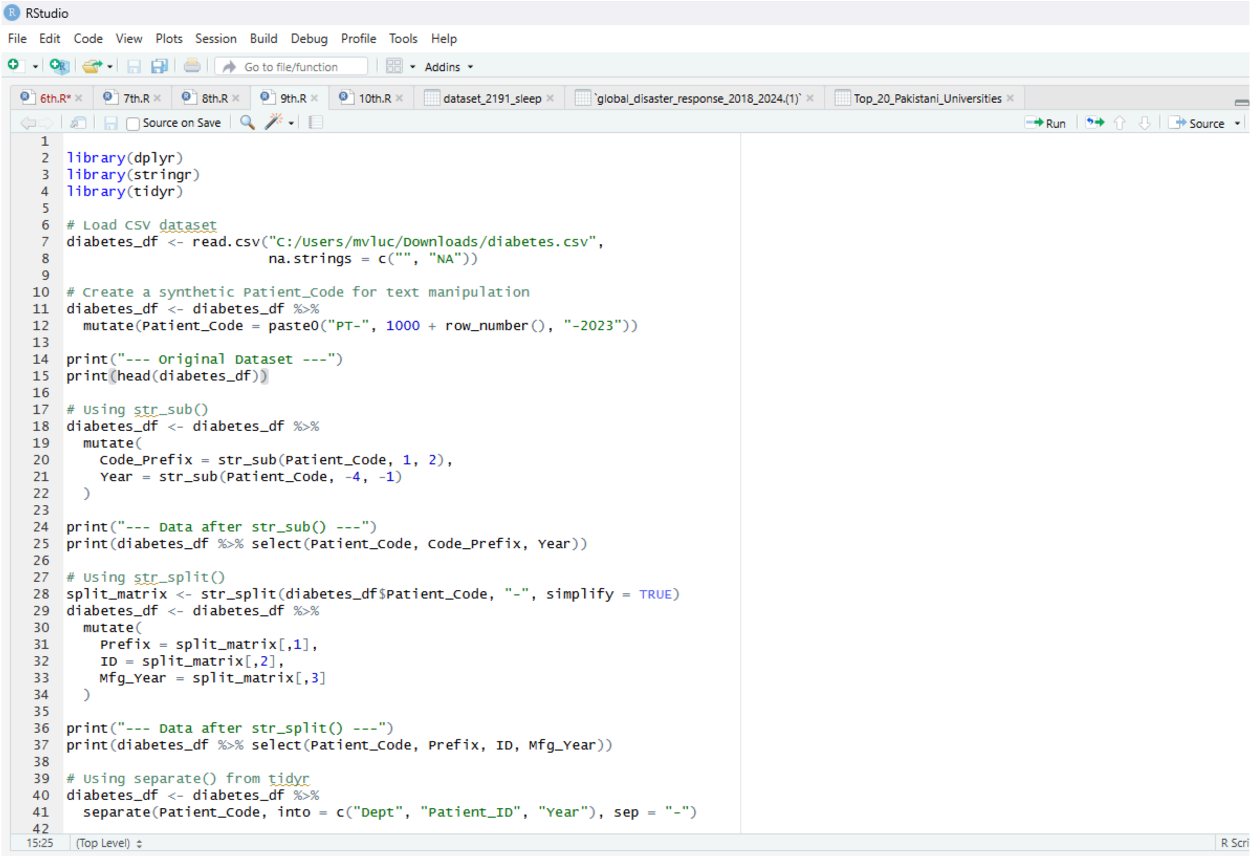
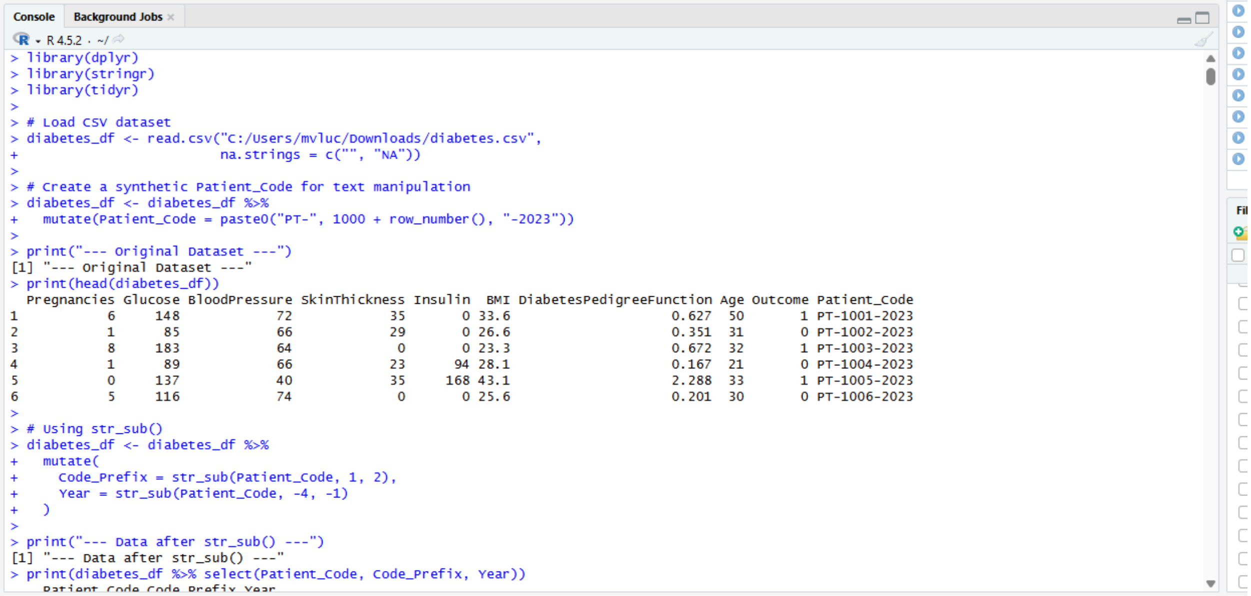
Sheth l.u.j. And sir m.v. college of arts science and commerce

Practical 9 R

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

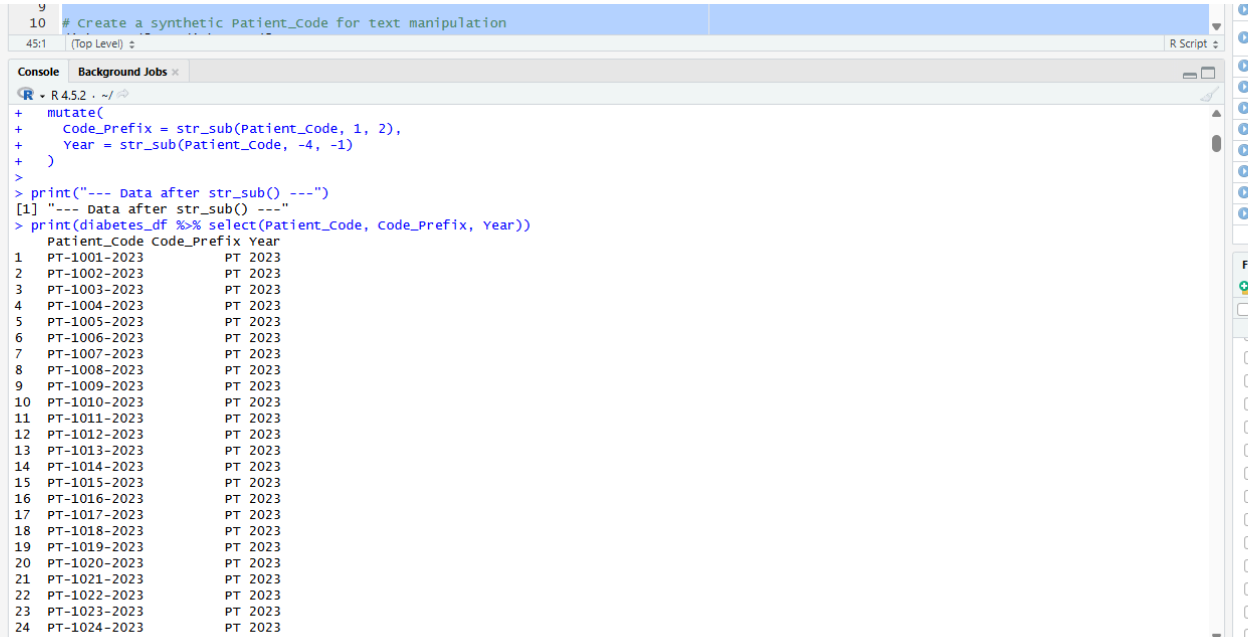

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
6th.R*
7th.R
8th.R
9th.R
10th.R
dataset_2191_sleep
`global_disaster_response_2018_2024.(1)`
Top_20_Pakistani_Universities
Source on Save
Run
Source
15:25
(Top Level)
R Scri

```r

library(dplyr)
library(stringr)
library(tidyr)

# Load CSV dataset
diabetes_df <- read.csv("C:/Users/mvluc/Downloads/diabetes.csv",
                        na.strings = c("", "NA"))

# Create a synthetic Patient_Code for text manipulation
diabetes_df <- diabetes_df %>%
  mutate(Patient_Code = paste0("PT-", 1000 + row_number(), "-2023"))

print("--- Original Dataset ---")
print(head(diabetes_df))

# Using str_sub()
diabetes_df <- diabetes_df %>%
  mutate(
    Code_Prefix = str_sub(Patient_Code, 1, 2),
    Year = str_sub(Patient_Code, -4, -1)
  )

print("--- Data after str_sub() ---")
print(diabetes_df %>% select(Patient_Code, Code_Prefix, Year))

# Using str_split()
split_matrix <- str_split(diabetes_df$Patient_Code, "-", simplify = TRUE)
diabetes_df <- diabetes_df %>%
  mutate(
    Prefix = split_matrix[,1],
    ID = split_matrix[,2],
    Mfg_Year = split_matrix[,3]
  )

print("--- Data after str_split() ---")
print(diabetes_df %>% select(Patient_Code, Prefix, ID, Mfg_Year))

# Using separate() from tidyr
diabetes_df <- diabetes_df %>%
  separate(Patient_Code, into = c("Dept", "Patient_ID", "Year"), sep = "-")

```



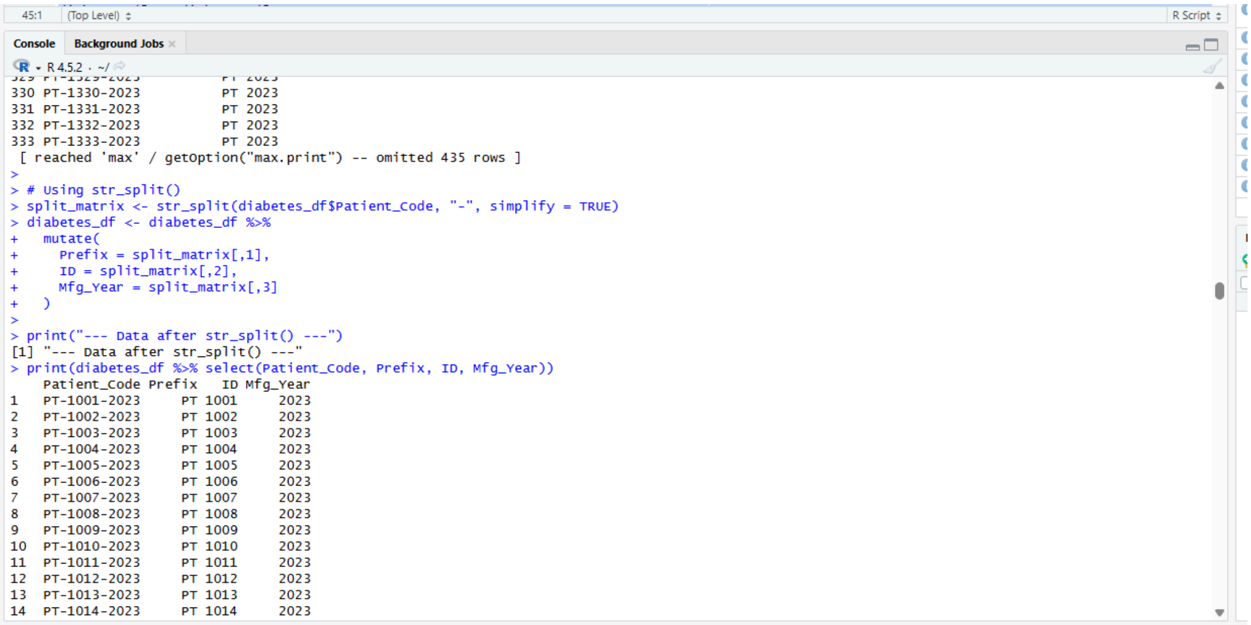
```
> library(dplyr)
> library(stringr)
> library(tidyr)
>
> # Load CSV dataset
> diabetes_df <- read.csv("C:/Users/mvluc/Downloads/diabetes.csv",
+                         na.strings = c("", "NA"))
>
> # Create a synthetic Patient_Code for text manipulation
> diabetes_df <- diabetes_df %>%
+   mutate(Patient_Code = paste0("PT-", 1000 + row_number(), "-2023"))
>
> print("--- Original Dataset ---")
[1] "--- Original Dataset ---"
> print(head(diabetes_df))
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age Outcome Patient_Code
1           6     148            72            35       0 33.6                    0.627  50       1 PT-1001-2023
2           1      85            66            29       0 26.6                    0.351  31       0 PT-1002-2023
3           8     183            64             0       0 23.3                    0.672  32       1 PT-1003-2023
4           1      89            66            23      94 28.1                    0.167  21       0 PT-1004-2023
5           0     137            40            35     168 43.1                    2.288  33       1 PT-1005-2023
6           5     116            74             0       0 25.6                    0.201  30       0 PT-1006-2023
>
> # Using str_sub()
> diabetes_df <- diabetes_df %>%
+   mutate(
+     Code_Prefix = str_sub(Patient_Code, 1, 2),
+     Year = str_sub(Patient_Code, -4, -1)
+   )
>
> print("--- Data after str_sub() ---")
[1] "--- Data after str_sub() ---"
> print(diabetes_df %>% select(Patient_Code, Code_Prefix, Year))
   Patient_Code Code_Prefix Year
```

Name: Simran S113

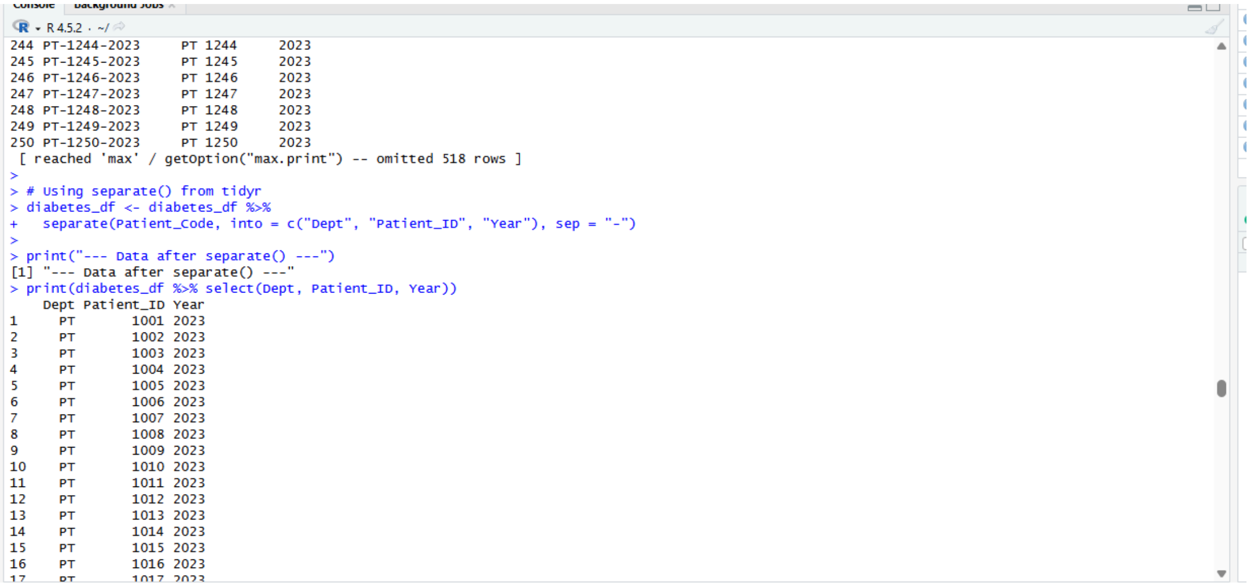


```
# Create a synthetic Patient_Code for text manipulation
```

```
+   mutate(
+     Code_Prefix = str_sub(Patient_Code, 1, 2),
+     Year = str_sub(Patient_Code, -4, -1)
+   )
> 
> print("--- Data after str_sub() ---")
[1] "--- Data after str_sub() ---"
> print(diabetes_df %>% select(Patient_Code, Code_Prefix, Year))
   Patient_Code Code_Prefix Year
1  PT-1001-2023          PT 2023
2  PT-1002-2023          PT 2023
3  PT-1003-2023          PT 2023
4  PT-1004-2023          PT 2023
5  PT-1005-2023          PT 2023
6  PT-1006-2023          PT 2023
7  PT-1007-2023          PT 2023
8  PT-1008-2023          PT 2023
9  PT-1009-2023          PT 2023
10 PT-1010-2023          PT 2023
11 PT-1011-2023          PT 2023
12 PT-1012-2023          PT 2023
13 PT-1013-2023          PT 2023
14 PT-1014-2023          PT 2023
15 PT-1015-2023          PT 2023
16 PT-1016-2023          PT 2023
17 PT-1017-2023          PT 2023
18 PT-1018-2023          PT 2023
19 PT-1019-2023          PT 2023
20 PT-1020-2023          PT 2023
21 PT-1021-2023          PT 2023
22 PT-1022-2023          PT 2023
23 PT-1023-2023          PT 2023
24 PT-1024-2023          PT 2023
```



```
330 PT-1330-2023          PT 2023
331 PT-1331-2023          PT 2023
332 PT-1332-2023          PT 2023
333 PT-1333-2023          PT 2023
 [ reached 'max' / getOption("max.print") -- omitted 435 rows ]
> 
> # Using str_split()
> split_matrix <- str_split(diabetes_df$Patient_Code, "-", simplify = TRUE)
> diabetes_df <- diabetes_df %>%
+   mutate(
+     Prefix = split_matrix[,1],
+     ID = split_matrix[,2],
+     Mfg_Year = split_matrix[,3]
+   )
> 
> print("--- Data after str_split() ---")
[1] "--- Data after str_split() ---"
> print(diabetes_df %>% select(Patient_Code, Prefix, ID, Mfg_Year))
   Patient_Code Prefix   ID Mfg_Year
1  PT-1001-2023     PT 1001     2023
2  PT-1002-2023     PT 1002     2023
3  PT-1003-2023     PT 1003     2023
4  PT-1004-2023     PT 1004     2023
5  PT-1005-2023     PT 1005     2023
6  PT-1006-2023     PT 1006     2023
7  PT-1007-2023     PT 1007     2023
8  PT-1008-2023     PT 1008     2023
9  PT-1009-2023     PT 1009     2023
10 PT-1010-2023     PT 1010     2023
11 PT-1011-2023     PT 1011     2023
12 PT-1012-2023     PT 1012     2023
13 PT-1013-2023     PT 1013     2023
14 PT-1014-2023     PT 1014     2023
```



```
244 PT-1244-2023     PT 1244     2023
245 PT-1245-2023     PT 1245     2023
246 PT-1246-2023     PT 1246     2023
247 PT-1247-2023     PT 1247     2023
248 PT-1248-2023     PT 1248     2023
249 PT-1249-2023     PT 1249     2023
250 PT-1250-2023     PT 1250     2023
 [ reached 'max' / getOption("max.print") -- omitted 518 rows ]
>
> # Using separate() from tidyr
> diabetes_df <- diabetes_df %>%
+   separate(Patient_Code, into = c("Dept", "Patient_ID", "Year"), sep = "-")
>
> print("--- Data after separate() ---")
[1] "--- Data after separate() ---"
> print(diabetes_df %>% select(Dept, Patient_ID, Year))
    Dept Patient_ID Year
1     PT       1001 2023
2     PT       1002 2023
3     PT       1003 2023
4     PT       1004 2023
5     PT       1005 2023
6     PT       1006 2023
7     PT       1007 2023
8     PT       1008 2023
9     PT       1009 2023
10    PT       1010 2023
11    PT       1011 2023
12    PT       1012 2023
13    PT       1013 2023
14    PT       1014 2023
15    PT       1015 2023
16    PT       1016 2023
17    PT       1017 2023
```