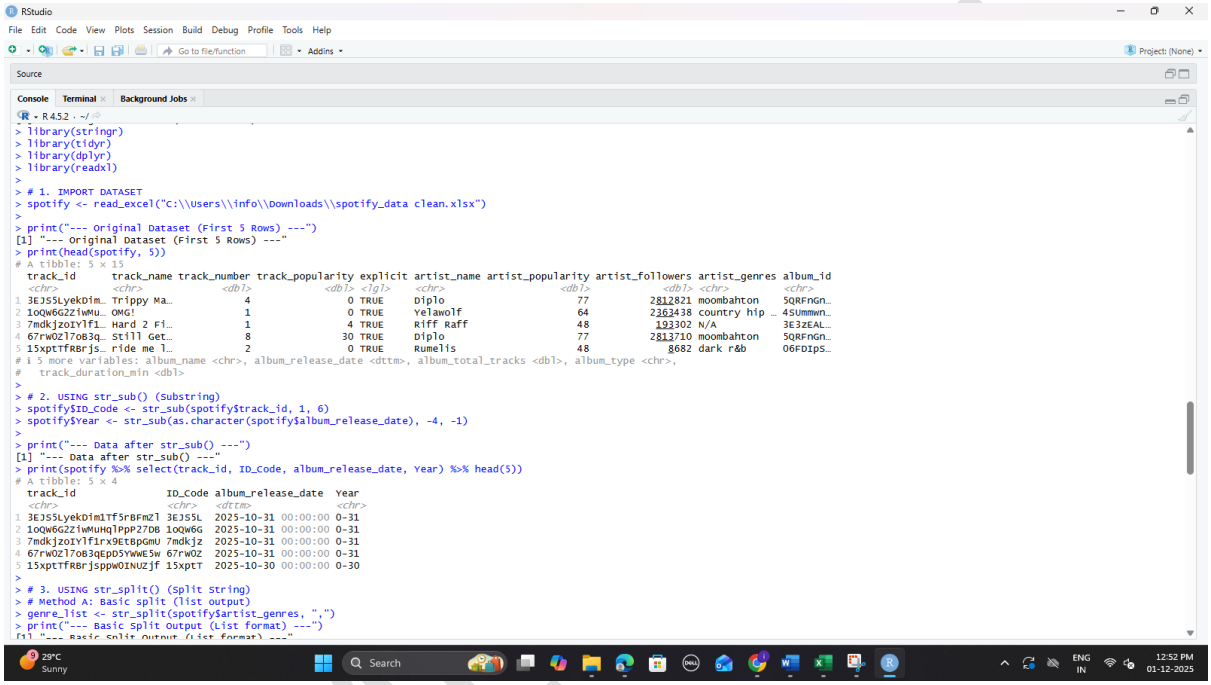
SHETH L.U.J AND SIR M.V. COLLEGE

SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R

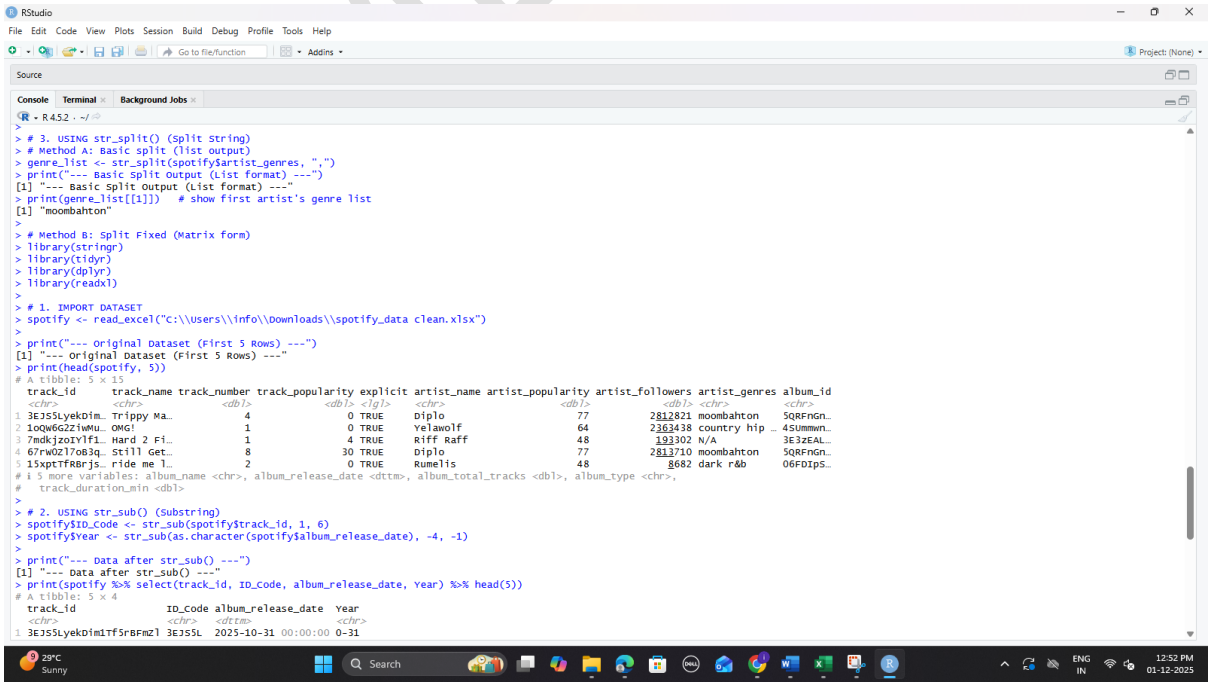
PRACTICAL – 9

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

OUTPUT:-



```
> library(stringr)
> library(tidyr)
> library(dplyr)
> library(readxl)
>
> # 1. IMPORT DATASET
> spotify <- read_excel("C:\\Users\\info\\Downloads\\spotify_data clean.xlsx")
>
> print("--- Original Dataset (First 5 Rows) ---")
[1] "--- Original Dataset (First 5 Rows) ---"
> print(head(spotify, 5))
# A tibble: 5 × 15
  track_id      track_name track_number track_popularity explicit artist_name artist_popularity artist_followers artist_genres album_id
  <chr>         <chr>            <dbl>            <dbl> <lgl>    <chr>                  <dbl>            <dbl> <chr>         <chr>
1 3EJS5LyekDim… Trippy Ma…           4                0 TRUE     Diplo                     77          2812821 moombahton    5QRFnGn…
2 1oQW6G2ZIwMu… OMG!                 1                0 TRUE     Yelawolf                  64          2363438 country hip … 4SUmmwn…
3 7mdkjzoIYlf1… Hard 2 Fi…           1                4 TRUE     Riff Raff                 48           193302 N/A           3E3zEAL…
4 67rWOZl7oB3q… Still Get…           8               30 TRUE     Diplo                     77          2813710 moombahton    5QRFnGn…
5 15xptTfR8rjs… ride me l…           2                0 TRUE     Rumelis                   48             8682 dark r&b      06FDIpS…
# ℹ 5 more variables: album_name <chr>, album_release_date <dttm>, album_total_tracks <dbl>, album_type <chr>,
#   track_duration_min <dbl>
>
> # 2. USING str_sub() (Substring)
> spotify$ID_Code <- str_sub(spotify$track_id, 1, 6)
> spotify$Year <- str_sub(as.character(spotify$album_release_date), -4, -1)
>
> print("--- Data after str_sub() ---")
[1] "--- Data after str_sub() ---"
> print(spotify %>% select(track_id, ID_Code, album_release_date, Year) %>% head(5))
# A tibble: 5 × 4
  track_id               ID_Code album_release_date  Year
  <chr>                  <chr>   <dttm>              <chr>
1 3EJS5LyekDim1Tf5rBFmZl 3EJS5L  2025-10-31 00:00:00 0-31
2 1oQW6G2ZIwMuHqlPpP27DB 1oQW6G  2025-10-31 00:00:00 0-31
3 7mdkjzoIYlf1rx9EtBpGmU 7mdkjz  2025-10-31 00:00:00 0-31
4 67rWOZl7oB3qEpD5YvWE5w 67rWOZ  2025-10-31 00:00:00 0-31
5 15xptTfR8rjsppwOINUZjf 15xptT  2025-10-30 00:00:00 0-30
>
> # 3. USING str_split() (Split String)
> # Method A: Basic split (list output)
> genre_list <- str_split(spotify$artist_genres, ",")
> print("--- Basic Split Output (List format) ---")
[1] "--- Basic Split Output (List format) ---"
```

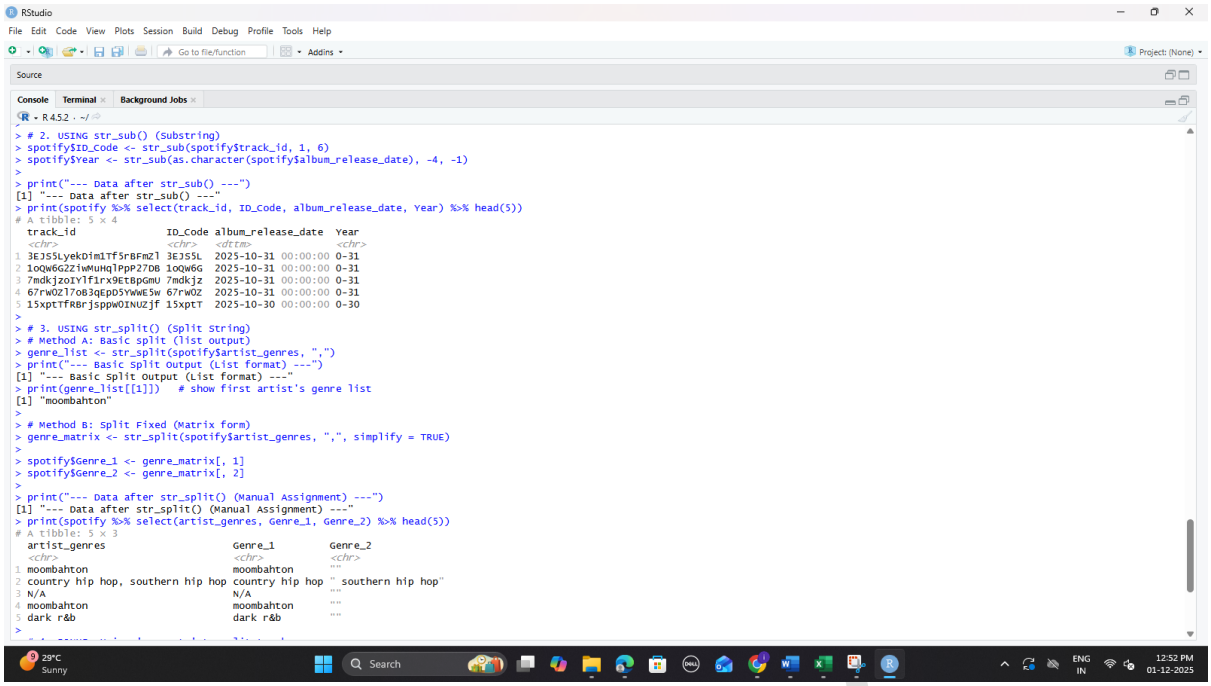


```
>
> # 3. USING str_split() (Split String)
> # Method A: Basic split (list output)
> genre_list <- str_split(spotify$artist_genres, ",")
> print("--- Basic Split Output (List format) ---")
[1] "--- Basic Split Output (List format) ---"
> print(genre_list[[1]])   # show first artist's genre list
[1] "moombahton"
>
> # Method B: Split Fixed (Matrix form)
> library(stringr)
> library(tidyr)
> library(dplyr)
> library(readxl)
>
> # 1. IMPORT DATASET
> spotify <- read_excel("C:\\Users\\info\\Downloads\\spotify_data clean.xlsx")
>
> print("--- Original Dataset (First 5 Rows) ---")
[1] "--- Original Dataset (First 5 Rows) ---"
> print(head(spotify, 5))
# A tibble: 5 × 15
  track_id      track_name track_number track_popularity explicit artist_name artist_popularity artist_followers artist_genres album_id
  <chr>         <chr>            <dbl>            <dbl> <lgl>    <chr>                  <dbl>            <dbl> <chr>         <chr>
1 3EJS5LyekDim… Trippy Ma…           4                0 TRUE     Diplo                     77          2812821 moombahton    5QRFnGn…
2 1oQW6G2ZIwMu… OMG!                 1                0 TRUE     Yelawolf                  64          2363438 country hip … 4SUmmwn…
3 7mdkjzoIYlf1… Hard 2 Fi…           1                4 TRUE     Riff Raff                 48           193302 N/A           3E3zEAL…
4 67rWOZl7oB3q… Still Get…           8               30 TRUE     Diplo                     77          2813710 moombahton    5QRFnGn…
5 15xptTfR8rjs… ride me l…           2                0 TRUE     Rumelis                   48             8682 dark r&b      06FDIpS…
# ℹ 5 more variables: album_name <chr>, album_release_date <dttm>, album_total_tracks <dbl>, album_type <chr>,
#   track_duration_min <dbl>
>
> # 2. USING str_sub() (Substring)
> spotify$ID_Code <- str_sub(spotify$track_id, 1, 6)
> spotify$Year <- str_sub(as.character(spotify$album_release_date), -4, -1)
>
> print("--- Data after str_sub() ---")
[1] "--- Data after str_sub() ---"
> print(spotify %>% select(track_id, ID_Code, album_release_date, Year) %>% head(5))
# A tibble: 5 × 4
  track_id               ID_Code album_release_date  Year
  <chr>                  <chr>   <dttm>              <chr>
1 3EJS5LyekDim1Tf5rBFmZl 3EJS5L  2025-10-31 00:00:00 0-31
```

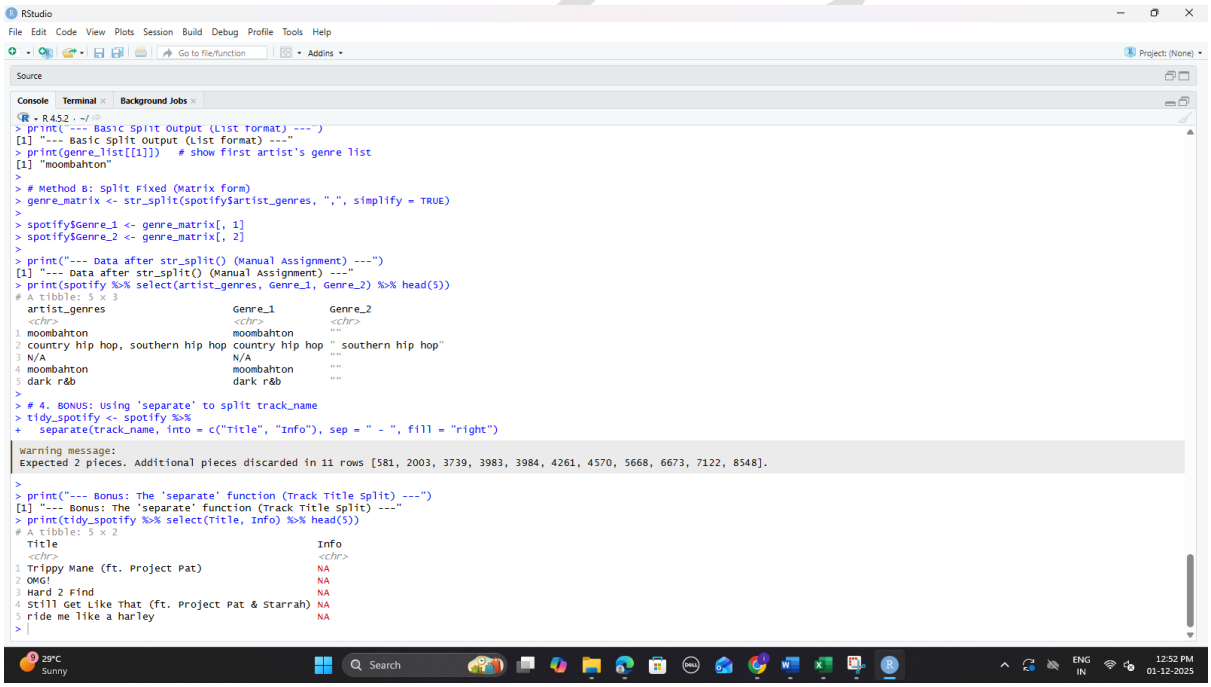
SUMEET JITENDRA YADAV
S124

SHETH L.U.J AND SIR M.V. COLLEGE

SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R



```
> # 2. USING str_sub() (Substring)
> spotify$ID_Code <- str_sub(spotify$track_id, 1, 6)
> spotify$Year <- str_sub(as.character(spotify$album_release_date), -4, -1)
>
> print("--- Data after str_sub() ---")
[1] "--- Data after str_sub() ---"
> print(spotify %>% select(track_id, ID_Code, album_release_date, Year) %>% head(5))
# A tibble: 5 × 4
  track_id               ID_Code album_release_date  Year
  <chr>                  <chr>   <dttm>              <chr>
1 3EJS5LyekDim1Tf5rBFmZl 3EJS5L  2025-10-31 00:00:00 0-31
2 1oQW6G2ZIwMuHqlPpP27DB 1oQW6G  2025-10-31 00:00:00 0-31
3 7mdkjzoIYlf1rX9EtBpGmU 7mdkjz  2025-10-31 00:00:00 0-31
4 67rWOZl7oB3qEpD5YvWESw 67rWOZ  2025-10-31 00:00:00 0-31
5 15xptTfRBrjsppwOINUZjf 15xptT  2025-10-30 00:00:00 0-30
>
> # 3. USING str_split() (Split String)
> # Method A: Basic split (list output)
> genre_list <- str_split(spotify$artist_genres, ",")
> print("--- Basic Split Output (List format) ---")
[1] "--- Basic Split Output (List format) ---"
> print(genre_list[[1]])   # show first artist's genre list
[1] "moombahton"
>
> # Method B: Split Fixed (Matrix form)
> genre_matrix <- str_split(spotify$artist_genres, ",", simplify = TRUE)
>
> spotify$Genre_1 <- genre_matrix[, 1]
> spotify$Genre_2 <- genre_matrix[, 2]
>
> print("--- Data after str_split() (Manual Assignment) ---")
[1] "--- Data after str_split() (Manual Assignment) ---"
> print(spotify %>% select(artist_genres, Genre_1, Genre_2) %>% head(5))
# A tibble: 5 × 3
  artist_genres                  Genre_1         Genre_2
  <chr>                          <chr>           <chr>
1 moombahton                     moombahton      ""
2 country hip hop, southern hip hop country hip hop " southern hip hop"
3 N/A                            N/A             ""
4 moombahton                     moombahton      ""
5 dark r&b                       dark r&b        ""
>
```



```
> print("--- Basic Split Output (List format) ---")
[1] "--- Basic Split Output (List format) ---"
> print(genre_list[[1]])   # show first artist's genre list
[1] "moombahton"
>
> # Method B: Split Fixed (Matrix form)
> genre_matrix <- str_split(spotify$artist_genres, ",", simplify = TRUE)
>
> spotify$Genre_1 <- genre_matrix[, 1]
> spotify$Genre_2 <- genre_matrix[, 2]
>
> print("--- Data after str_split() (Manual Assignment) ---")
[1] "--- Data after str_split() (Manual Assignment) ---"
> print(spotify %>% select(artist_genres, Genre_1, Genre_2) %>% head(5))
# A tibble: 5 × 3
  artist_genres                  Genre_1         Genre_2
  <chr>                          <chr>           <chr>
1 moombahton                     moombahton      ""
2 country hip hop, southern hip hop country hip hop " southern hip hop"
3 N/A                            N/A             ""
4 moombahton                     moombahton      ""
5 dark r&b                       dark r&b        ""
>
> # 4. BONUS: Using 'separate' to split track_name
> tidy_spotify <- spotify %>%
+   separate(track_name, into = c("Title", "Info"), sep = " - ", fill = "right")

Warning message:
Expected 2 pieces. Additional pieces discarded in 11 rows [581, 2003, 3739, 3983, 3984, 4261, 4570, 5668, 6673, 7122, 8548].

>
> print("--- Bonus: The 'separate' function (Track Title Split) ---")
[1] "--- Bonus: The 'separate' function (Track Title Split) ---"
> print(tidy_spotify %>% select(Title, Info) %>% head(5))
# A tibble: 5 × 2
  Title                                  Info
  <chr>                                  <chr>
1 Trippy Mane (ft. Project Pat)          NA
2 OMG!                                   NA
3 Hard 2 Find                            NA
4 Still Get Like That (ft. Project Pat & Starrah) NA
5 ride me like a harley                  NA
>
```


SUMEET JITENDRA YADAV

S124

SHETH L.U.J AND SIR M.V. COLLEGE

SUBJECT :- DATA ANALYSIS WITH SAS/SPSS/R

SUMEET JITENDRA YADAV
S124