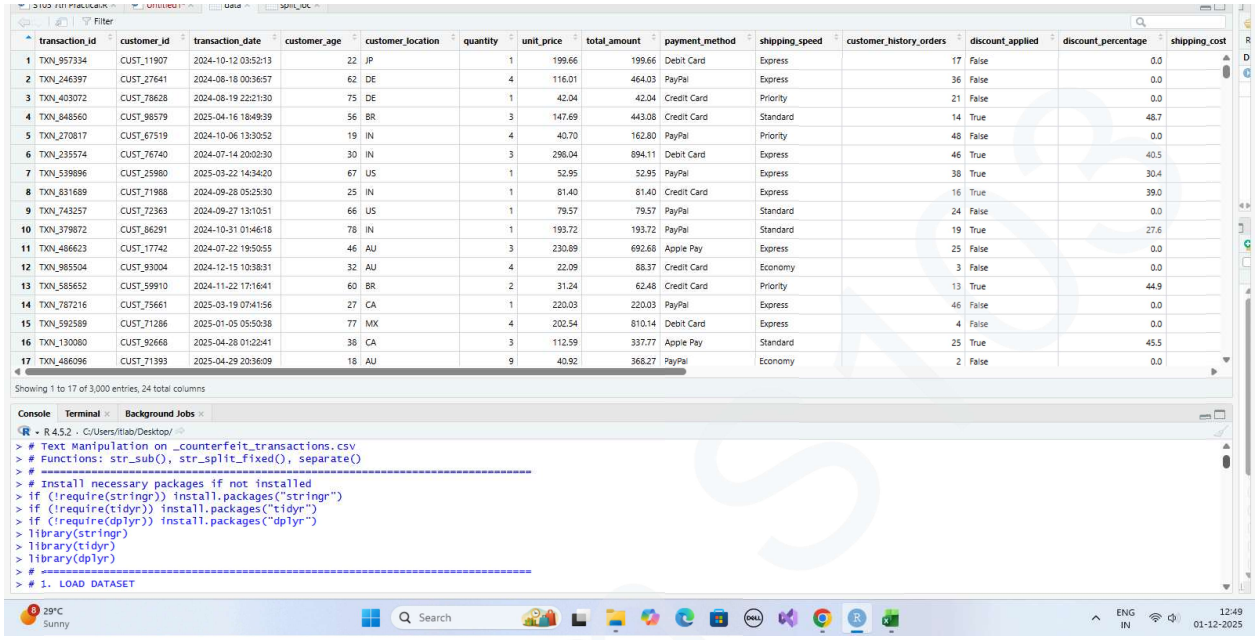
SHETH L.U.J. AND SIR M.V. COLLEGE
SUBJECT:Data Analysis with R

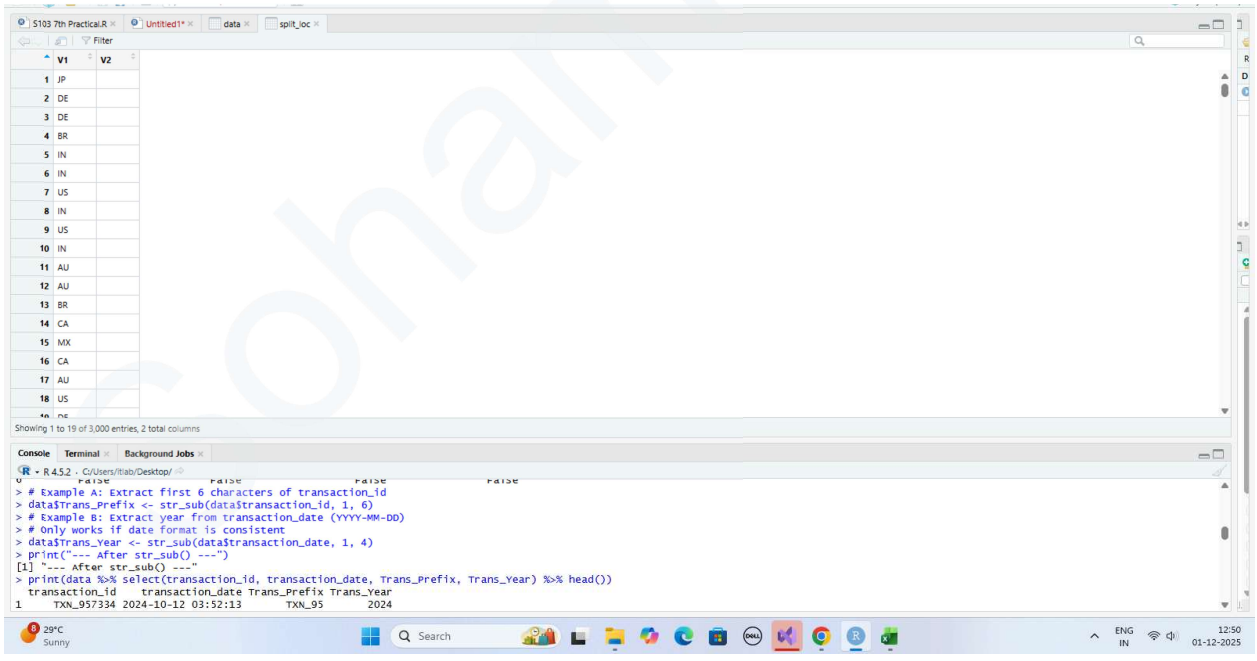
Aim: Performing text manipulation using str_sub(), str_split() (R). import dataset



	transaction_id	customer_id	transaction_date	customer_age	customer_location	quantity	unit_price	total_amount	payment_method	shipping_speed	customer_history_orders	discount_applied	discount_percentage	shipping_cost
1	TXN_957334	CUST_11907	2024-10-12 03:52:13	22	JP	1	199.66	199.66	Debit Card	Express	17	False	0.0	
2	TXN_246397	CUST_27641	2024-08-18 00:36:57	62	DE	4	116.01	464.03	PayPal	Express	36	False	0.0	
3	TXN_403072	CUST_78628	2024-08-19 22:21:30	75	DE	1	42.04	42.04	Credit Card	Priority	21	False	0.0	
4	TXN_848560	CUST_98579	2025-04-16 18:49:39	56	BR	3	147.69	443.08	Credit Card	Standard	14	True	48.7	
5	TXN_270817	CUST_67519	2024-10-06 13:30:52	19	IN	4	40.70	162.80	PayPal	Priority	48	False	0.0	
6	TXN_235574	CUST_76740	2024-07-14 20:02:30	30	IN	3	298.04	894.11	Debit Card	Express	46	True	40.5	
7	TXN_539896	CUST_25980	2025-03-22 14:34:20	67	US	1	52.95	52.95	PayPal	Express	38	True	30.4	
8	TXN_831689	CUST_71988	2024-09-28 05:25:30	25	IN	1	81.40	81.40	Credit Card	Express	16	True	39.0	
9	TXN_743257	CUST_72363	2024-09-27 13:10:51	66	US	1	79.57	79.57	PayPal	Standard	24	False	0.0	
10	TXN_379872	CUST_86291	2024-10-31 01:46:18	78	IN	1	193.72	193.72	PayPal	Standard	19	True	27.6	
11	TXN_486623	CUST_17742	2024-07-22 19:50:55	46	AU	3	230.89	692.68	Apple Pay	Express	25	False	0.0	
12	TXN_965504	CUST_93004	2024-12-15 10:38:31	32	AU	4	22.09	88.37	Credit Card	Economy	3	False	0.0	
13	TXN_585652	CUST_59910	2024-11-22 17:16:41	60	BR	2	31.24	62.48	Credit Card	Priority	13	True	44.9	
14	TXN_787216	CUST_75661	2025-03-19 07:41:56	27	CA	1	220.03	220.03	PayPal	Express	46	False	0.0	
15	TXN_592589	CUST_71286	2025-01-05 05:50:38	77	MX	4	202.54	810.14	Debit Card	Express	4	False	0.0	
16	TXN_130080	CUST_92668	2025-04-28 01:22:41	38	CA	3	112.59	337.77	Apple Pay	Standard	25	True	45.5	
17	TXN_486096	CUST_71393	2025-04-29 20:36:09	18	AU	9	40.92	368.27	PayPal	Economy	2	False	0.0	

Showing 1 to 17 of 3,000 entries, 24 total columns

```
> # Text Manipulation on _counterfeit_transactions.csv
> # Functions: str_sub(), str_split_fixed(), separate()
> # ================================================================================
> # Install necessary packages if not installed
> if (!require(stringr)) install.packages("stringr")
> if (!require(tidyr)) install.packages("tidyr")
> if (!require(dplyr)) install.packages("dplyr")
> library(stringr)
> library(tidyr)
> library(dplyr)
> # ================================================================================
> # 1. LOAD DATASET
```



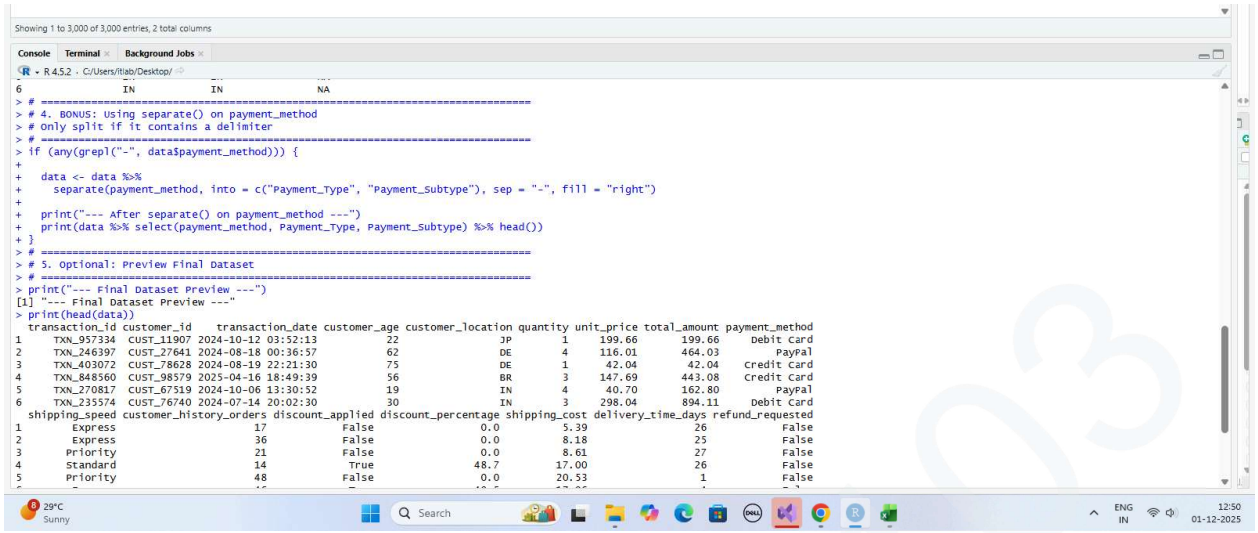
	V1	V2
1	JP	
2	DE	
3	DE	
4	BR	
5	IN	
6	IN	
7	US	
8	IN	
9	US	
10	IN	
11	AU	
12	AU	
13	BR	
14	CA	
15	MX	
16	CA	
17	AU	
18	US	

Showing 1 to 19 of 3,000 entries, 2 total columns

```
> # Example A: Extract first 6 characters of transaction_id
> data$Trans_Prefix <- str_sub(data$transaction_id, 1, 6)
> # Example B: Extract year from transaction_date (YYYY-MM-DD)
> # Only works if date format is consistent
> data$Trans_Year <- str_sub(data$transaction_date, 1, 4)
> print("--- After str_sub() ---")
[1] "--- After str_sub() ---"
> print(data %>% select(transaction_id, transaction_date, Trans_Prefix, Trans_Year) %>% head())
  transaction_id    transaction_date Trans_Prefix Trans_Year
1     TXN_957334 2024-10-12 03:52:13       TXN_95       2024
```

Soham Patil
Sycs S103

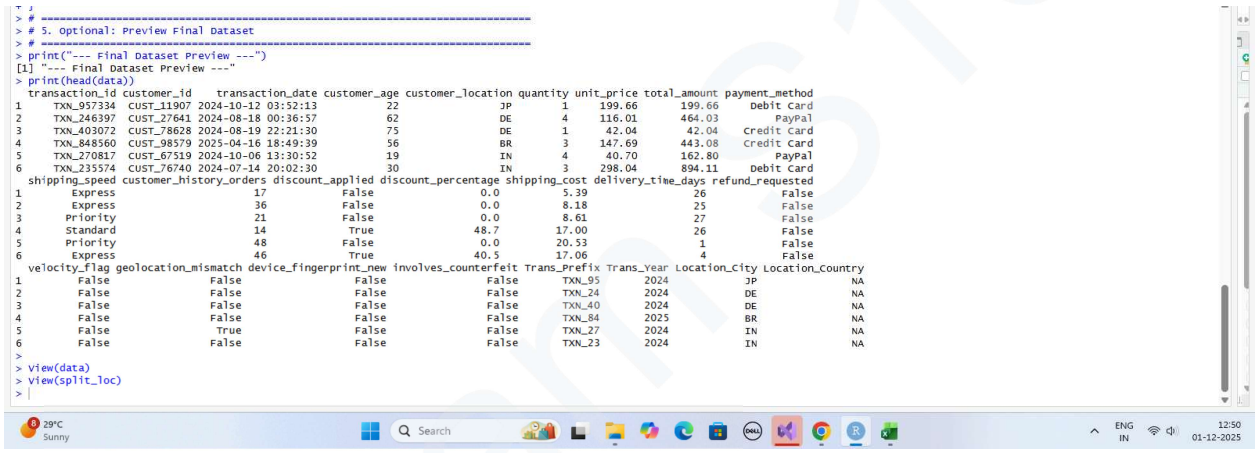
SHETH L.U.J. AND SIR M.V. COLLEGE
SUBJECT:Data Analysis with R



```
Showing 1 to 3,000 of 3,000 entries, 2 total columns

Console  Terminal  Background Jobs
R 4.5.2 · C:/Users/itlab/Desktop/

6                IN             IN                NA
> # ================================================================================
> # 4. BONUS: Using separate() on payment_method
> # Only split if it contains a delimiter
> # ================================================================================
> if (any(grepl("-", data$payment_method))) {
+
+   data <- data %>%
+     separate(payment_method, into = c("Payment_Type", "Payment_Subtype"), sep = "-", fill = "right")
+
+   print("--- After separate() on payment_method ---")
+   print(data %>% select(payment_method, Payment_Type, Payment_Subtype) %>% head())
+ }
> # ================================================================================
> # 5. Optional: Preview Final Dataset
> # ================================================================================
> print("--- Final Dataset Preview ---")
[1] "--- Final Dataset Preview ---"
> print(head(data))
  transaction_id customer_id    transaction_date customer_age customer_location quantity unit_price total_amount payment_method
1     TXN_957334  CUST_11907 2024-10-12 03:52:13           22                JP        1     199.66       199.66     Debit Card
2     TXN_246397  CUST_27641 2024-08-18 00:36:57           62                DE        4     116.01       464.03         PayPal
3     TXN_403072  CUST_78628 2024-08-19 22:21:30           75                DE        1      42.04        42.04    Credit Card
4     TXN_848560  CUST_98579 2025-04-16 18:49:39           56                BR        3     147.69       443.08    Credit Card
5     TXN_270817  CUST_67519 2024-10-06 13:30:52           19                IN        4      40.70       162.80         PayPal
6     TXN_235574  CUST_76740 2024-07-14 20:02:30           30                IN        3     298.04       894.11     Debit Card
  shipping_speed customer_history_orders discount_applied discount_percentage shipping_cost delivery_time_days refund_requested
1        Express                      17            False                 0.0          5.39                 26            False
2        Express                      36            False                 0.0          8.18                 25            False
3       Priority                      21            False                 0.0          8.61                 27            False
4       Standard                      14             True                48.7         17.00                 26            False
5       Priority                      48            False                 0.0         20.53                  1            False
```



```
> # ================================================================================
> # 5. Optional: Preview Final Dataset
> # ================================================================================
> print("--- Final Dataset Preview ---")
[1] "--- Final Dataset Preview ---"
> print(head(data))
  transaction_id customer_id    transaction_date customer_age customer_location quantity unit_price total_amount payment_method
1     TXN_957334  CUST_11907 2024-10-12 03:52:13           22                JP        1     199.66       199.66     Debit Card
2     TXN_246397  CUST_27641 2024-08-18 00:36:57           62                DE        4     116.01       464.03         PayPal
3     TXN_403072  CUST_78628 2024-08-19 22:21:30           75                DE        1      42.04        42.04    Credit Card
4     TXN_848560  CUST_98579 2025-04-16 18:49:39           56                BR        3     147.69       443.08    Credit Card
5     TXN_270817  CUST_67519 2024-10-06 13:30:52           19                IN        4      40.70       162.80         PayPal
6     TXN_235574  CUST_76740 2024-07-14 20:02:30           30                IN        3     298.04       894.11     Debit Card
  shipping_speed customer_history_orders discount_applied discount_percentage shipping_cost delivery_time_days refund_requested
1        Express                      17            False                 0.0          5.39                 26            False
2        Express                      36            False                 0.0          8.18                 25            False
3       Priority                      21            False                 0.0          8.61                 27            False
4       Standard                      14             True                48.7         17.00                 26            False
5       Priority                      48            False                 0.0         20.53                  1            False
6        Express                      46             True                40.5         17.06                  4            False
  velocity_flag geolocation_mismatch device_fingerprint_new involves_counterfeit Trans_Prefix Trans_Year Location_City Location_Country
1         False                False                  False                False        TXN_95       2024            JP               NA
2         False                False                  False                False        TXN_24       2024            DE               NA
3         False                False                  False                False        TXN_40       2024            DE               NA
4         False                False                  False                False        TXN_84       2025            BR               NA
5         False                 True                  False                False        TXN_27       2024            IN               NA
6         False                False                  False                False        TXN_23       2024            IN               NA
>
> view(data)
> view(split_loc)
>
```

Soham Patil
Sycs S103