

SHETH L.U.J & SIR M.V COLLEGE OF SCIENCE

SUBJECT : Data Analysis with SAS / SPSS / R

AIM : Applying basic data cleaning functions: handling missing values using `na.omit()/replace_na()` in R. import dataset.

OUTPUT

The first screenshot shows the initial data import and cleaning process. The console output is as follows:

```
> video_df <- read_csv("video_data.csv", na.strings = c("", "NA"))
> print("--- 1. Original Data ---")
[1] "--- 1. Original Data ---"
> print(head(video_df))
  video_id title duration_sec hashtags_count views likes comments shares upload_hour category
1 vid_1000 Short Video #0 43 9 198775 21933 3228 400 8 Tech
2 vid_1001 Short Video #1 56 2 290336 20063 3719 1942 16 Comedy
3 vid_1002 Short Video #2 33 6 264206 37032 3228 1817 7 Food
4 vid_1003 Short Video #3 19 9 85076 27269 2371 980 1 Lifestyle
5 vid_1004 Short Video #4 47 8 90780 8041 2891 1109 23 Tech
6 vid_1005 Short Video #5 12 3 153617 14488 3756 249 12 Travel
> print("--- Count of Missing Values per Column ---")
[1] "--- Count of Missing Values per Column ---"
> print(colSums(is.na(video_df)))
  video_id title duration_sec hashtags_count views likes comments shares upload_hour category
0 0 0 0 0 0 0 0 0 0
> clean_omit <- na.omit(video_df)
> print("--- 2. Data after na.omit() ---")
[1] "--- 2. Data after na.omit() ---"
> print(paste("Original rows:", nrow(video_df)))
[1] "Original rows: 300"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 300"
> print(head(clean_omit))
  video_id title duration_sec hashtags_count views likes comments shares upload_hour category
1 vid_1000 Short Video #0 43 9 198775 21933 3228 400 8 Tech
2 vid_1001 Short Video #1 56 2 290336 20063 3719 1942 16 Comedy
3 vid_1002 Short Video #2 33 6 264206 37032 3228 1817 7 Food
4 vid_1003 Short Video #3 19 9 85076 27269 2371 980 1 Lifestyle
```

The second screenshot shows the replacement of missing values. The console output is as follows:

```
> avg_duration <- mean(video_df$duration_sec, na.rm = TRUE)
> avg_views <- mean(video_df$views, na.rm = TRUE)
> clean_replace <- video_df %>%
+   replace_na(list(
+     category = "Unknown",
+     likes = 0,
+     shares = 0,
+     duration_sec = avg_duration,
+     views = avg_views
+   ))
> print("--- 3. Data after replace_na() ---")
[1] "--- 3. Data after replace_na() ---"
> print(clean_replace[3, ])
  video_id title duration_sec hashtags_count views likes comments shares upload_hour category
3 vid_1002 Short Video #2 33 6 264206 37032 3228 1817 7 Food
> print(head(clean_replace))
  video_id title duration_sec hashtags_count views likes comments shares upload_hour category
1 vid_1000 Short Video #0 43 9 198775 21933 3228 400 8 Tech
2 vid_1001 Short Video #1 56 2 290336 20063 3719 1942 16 Comedy
3 vid_1002 Short Video #2 33 6 264206 37032 3228 1817 7 Food
4 vid_1003 Short Video #3 19 9 85076 27269 2371 980 1 Lifestyle
5 vid_1004 Short Video #4 47 8 90780 8041 2891 1109 23 Tech
6 vid_1005 Short Video #5 12 3 153617 14488 3756 249 12 Travel
> print("--- Remaining NAs after replacement ---")
[1] "--- Remaining NAs after replacement ---"
> print(colSums(is.na(clean_replace)))
  video_id title duration_sec hashtags_count views likes comments shares upload_hour category
0 0 0 0 0 0 0 0 0 0
  shares upload_hour category
0 0 0
```

SHETH L.U.J & SIR M.V COLLEGE OF SCIENCE

SUBJECT : Data Analysis with SAS / SPSS / R

RStudio interface showing a data frame with 18 rows and 10 columns. The data includes video IDs, titles, durations, and engagement metrics. The console shows the command 'View(clean_omit)'.

	video_id	title	duration_sec	hashtags_count	views	likes	comments	shares	upload_hour	category
1	vid_1000	Short Video #0	43	9	198775	21933	3228	400	8	Tech
2	vid_1001	Short Video #1	56	2	290336	20063	3719	1942	16	Comedy
3	vid_1002	Short Video #2	33	6	264206	37032	3228	1817	7	Food
4	vid_1003	Short Video #3	19	9	85076	27269	2371	980	1	Lifestyle
5	vid_1004	Short Video #4	47	8	90780	8041	2891	1109	23	Tech
6	vid_1005	Short Video #5	12	3	153617	14488	3756	249	12	Travel
7	vid_1006	Short Video #6	25	0	22689	5669	4225	390	2	Tech
8	vid_1007	Short Video #7	43	1	274318	4400	94	1805	15	Tech
9	vid_1008	Short Video #8	23	0	351605	39137	2674	334	5	Comedy
10	vid_1009	Short Video #9	27	4	389318	36686	1955	628	18	Food
11	vid_1010	Short Video #10	15	4	366177	12115	793	1542	6	Travel
12	vid_1011	Short Video #11	15	6	10348	45102	1834	136	9	Education
13	vid_1012	Short Video #12	28	8	417930	6354	2116	882	8	Education
14	vid_1013	Short Video #13	57	8	39102	17244	4106	687	1	Tech
15	vid_1014	Short Video #14	40	2	75460	24314	3325	1863	14	Comedy
16	vid_1015	Short Video #15	44	2	365778	40218	2919	1521	4	Comedy
17	vid_1016	Short Video #16	28	2	90930	15283	2597	826	21	Education
18	vid_1017	Short Video #17	7	3	7801	6338	3205	726	3	Travel

RStudio interface showing the same data frame as the first image, but with the command 'View(clean_replace)' in the console. A Snipping Tool window is also visible in the bottom right corner.

	video_id	title	duration_sec	hashtags_count	views	likes	comments	shares	upload_hour	category
1	vid_1000	Short Video #0	43	9	198775	21933	3228	400	8	Tech
2	vid_1001	Short Video #1	56	2	290336	20063	3719	1942	16	Comedy
3	vid_1002	Short Video #2	33	6	264206	37032	3228	1817	7	Food
4	vid_1003	Short Video #3	19	9	85076	27269	2371	980	1	Lifestyle
5	vid_1004	Short Video #4	47	8	90780	8041	2891	1109	23	Tech
6	vid_1005	Short Video #5	12	3	153617	14488	3756	249	12	Travel
7	vid_1006	Short Video #6	25	0	22689	5669	4225	390	2	Tech
8	vid_1007	Short Video #7	43	1	274318	4400	94	1805	15	Tech
9	vid_1008	Short Video #8	23	0	351605	39137	2674	334	5	Comedy
10	vid_1009	Short Video #9	27	4	389318	36686	1955	628	18	Food
11	vid_1010	Short Video #10	15	4	366177	12115	793	1542	6	Travel
12	vid_1011	Short Video #11	15	6	10348	45102	1834	136	9	Education
13	vid_1012	Short Video #12	28	8	417930	6354	2116	882	8	Education
14	vid_1013	Short Video #13	57	8	39102	17244	4106	687	1	Tech
15	vid_1014	Short Video #14	40	2	75460	24314	3325	1863	14	Comedy
16	vid_1015	Short Video #15	44	2	365778	40218	2919	1521	4	Comedy
17	vid_1016	Short Video #16	28	2	90930	15283	2597	826	21	Education
18	vid_1017	Short Video #17	7	3	7801	6338	3205	726	3	Travel

SHETH L.U.J & SIR M.V COLLEGE OF SCIENCE

SUBJECT : Data Analysis with SAS / SPSS / R

The screenshot displays the RStudio environment. The main editor window shows a data frame with 18 rows and 10 columns. The columns are: video_id, title, duration_sec, hashtags_count, views, likes, comments, shares, upload_hour, and category. The data represents various short videos and their engagement metrics. The Environment pane on the right shows three data objects: clean_omit, clean_repl, and video_df, each containing 300 observations. The Files pane shows the project directory structure, including folders like .RData, .Rhistory, and .Rproj, and files like final_data.csv and video_data.csv. The Console pane at the bottom shows the R prompt and the command 'View(video_df)'.

video_id	title	duration_sec	hashtags_count	views	likes	comments	shares	upload_hour	category	
1	vid_1000	Short Video #0	43	9	198775	21933	3228	400	8	Tech
2	vid_1001	Short Video #1	56	2	290336	20063	3719	1942	16	Comedy
3	vid_1002	Short Video #2	33	6	264206	37032	3228	1817	7	Food
4	vid_1003	Short Video #3	19	9	85076	27269	2371	980	1	Lifestyle
5	vid_1004	Short Video #4	47	8	90780	8041	2891	1109	23	Tech
6	vid_1005	Short Video #5	12	3	153617	14488	3756	249	12	Travel
7	vid_1006	Short Video #6	25	0	22689	5669	4225	390	2	Tech
8	vid_1007	Short Video #7	43	1	274318	4400	94	1805	15	Tech
9	vid_1008	Short Video #8	23	0	351605	39137	2674	334	5	Comedy
10	vid_1009	Short Video #9	27	4	389318	36686	1955	628	18	Food
11	vid_1010	Short Video #10	15	4	366177	12115	793	1542	6	Travel
12	vid_1011	Short Video #11	15	6	10348	45102	1834	136	9	Education
13	vid_1012	Short Video #12	28	8	417930	6354	2116	882	8	Education
14	vid_1013	Short Video #13	57	8	39102	17244	4106	687	1	Tech
15	vid_1014	Short Video #14	40	2	75460	24314	3325	1863	14	Comedy
16	vid_1015	Short Video #15	44	2	365778	40218	2919	1521	4	Comedy
17	vid_1016	Short Video #16	28	2	90930	15283	2597	826	21	Education
18	vid_1017	Short Video #17	7	3	7801	6338	3205	726	3	Travel