

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

SUBJECT: Data Analysis with R

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

Environment History Connections Tutorial

R • Global Environment • 210 MB •

Data

- clean_omit 100 obs. of 7 variables
- clean_rep1 157 obs. of 7 variables
- retail_df 157 obs. of 7 variables

values

avg_perce... 75.6847260273973

Files Plots Packages Help Viewer Presentation

Home

- My Music
- My Pictures
- My Videos
- My Web Sites
- NetBeansProjects
- Power BI Desktop
- S103 Cn 1stpdf 2.6 MB Nov 1
- scala for DS
- Sound Recordings
- T050_AnjaliT
- vgsales.csv 1.3 MB Nov 1
- Virtual Machines
- Visual Studio 18
- Visual Studio 2022
- WindowsPowerShell
- air_traffic_data.csv 1.9 MB Dec 1
- student_admission_record_dirty.csv 6.7 KB Dec 1

Console Terminal Background Jobs

```
R • R 4.5.2 • C:/Users/Itab/Desktop/
> retail_df <- read.csv("student_admission_record_dirty.csv", na.strings = c("", "NA"))
> print("1. Original Data (First 6 Rows) ---")
[1] "1. Original Data (First 6 Rows) ---"
> print(head(retail_df))
  Name Age Gender Admission.Test.Score High.School.Percentage City Admission.Status
1 Shehroz 24 Female 50 68.90 Quetta Rejected
2 Waqar 21 Female 99 60.73 Karachi <NA>
3 Bushra 17 Male 89 NA Islamabad Accepted
4 Aliya 17 Male 55 85.29 Karachi Rejected
5 Bilal 20 Male 65 61.13 Lahore <NA>
6 Murtaza 23 Female NA NA Islamabad Accepted
7 Asad 18 Male NA 97.31 Multan Accepted
8 Rabia 20 Female 82 55.67 Lahore Accepted
9 Rohail 17 Male 64 NA Karachi Accepted
10 Kamran 18 Male 53 98.98 Multan Rejected
11 Shafiq 17 Male 78 -10.00 Quetta Rejected
12 Nashit 18 Male 89 NA Lahore Accepted
13 NA 19 Male 66 88.17 NA NA
14 Ahmed 21 Male 62 79.03 Karachi Accepted
15 Fareeha 22 Male 86 50.77 Islamabad Rejected
16 Saim 19 Male 84 97.19 Peshawar Rejected
```

Environment History Connections Tutorial

R • Global Environment • 210 MB •

Data

- clean_omit 100 obs. of 7 variables
- clean_rep1 157 obs. of 7 variables
- retail_df 157 obs. of 7 variables

values

avg_perce... 75.6847260273973

Files Plots Packages Help Viewer Presentation

Home

- My Music
- My Pictures
- My Videos
- My Web Sites
- NetBeansProjects
- Power BI Desktop
- S103 Cn 1stpdf 2.6 MB Nov 1
- scala for DS
- Sound Recordings
- T050_AnjaliT
- vgsales.csv 1.3 MB Nov 1
- Virtual Machines
- Visual Studio 18
- Visual Studio 2022
- WindowsPowerShell
- air_traffic_data.csv 1.9 MB Dec 1
- student_admission_record_dirty.csv 6.7 KB Dec 1

Console Terminal Background Jobs

```
R • R 4.5.2 • C:/Users/Itab/Desktop/
> print("3. Data after replace_na() ---")
[1] "3. Data after replace_na() ---"
> print(clean_rep1[, 3]) # Check row 3 for replacements
  Name Age Gender Admission.Test.Score High.School.Percentage City Admission.Status
3 Bushra 17 Male 89 75.68473 Islamabad Accepted
> print(head(clean_rep1))
  Name Age Gender Admission.Test.Score High.School.Percentage City Admission.Status
1 Shehroz 24 Female 50 68.90000 Quetta Rejected
2 Waqar 21 Female 99 60.73000 Karachi Pending
3 Bushra 17 Male 89 75.68473 Islamabad Accepted
4 Aliya 17 Male 55 85.29000 Karachi Rejected
5 Bilal 20 Male 65 61.13000 Lahore Pending
6 Murtaza 23 Female 0 75.68473 Islamabad Accepted
> # Verify remaining NAs
> print("Remaining NAs after replacement ---")
[1] "Remaining NAs after replacement ---"
> nrow(na.omit(clean_rep1))
```

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

SUBJECT: Data Analysis with R

The screenshot displays the R Studio environment. The main window shows a data table with 15 rows and 7 columns. The console window at the bottom shows the execution of R code for data cleaning and summarization. The file explorer on the right shows the project files.

	Name	Age	Gender	Admission.Test.Score	High.School.Percentage	City	Admission.Status
1	Shehroz	24	Female	50	68.90000	Quetta	Rejected
2	Waqar	21	Female	99	60.73000	Karachi	Pending
3	Bushra	17	Male	89	75.68473	Islamabad	Accepted
4	Aliya	17	Male	55	85.29000	Karachi	Rejected
5	Blial	20	Male	65	61.13000	Lahore	Pending
6	Murtaza	23	Female	0	75.68473	Islamabad	Accepted
7	Asad	18	Male	0	97.31000	Multan	Accepted
8	Rabia	20	Female	82	55.67000	Lahore	Accepted
9	Rohail	17	Male	64	75.68473	Karachi	Accepted
10	Kamran	18	Male	53	98.96000	Multan	Rejected
11	Shafiq	17	Male	78	-10.00000	Quetta	Rejected
12	Nashit	18	Male	89	75.68473	Lahore	Accepted
13	NA	19	Male	66	88.17000	Unknown	Pending
14	Ahmed	21	Male	62	79.03000	Karachi	Accepted

```
R> # R452 C:\Users\Tab\Desktop\
R> # Name Age Gender Admission.Test.Score High.School.Percentage City Admission.Status
3 Bushra 17 Male 89 75.68473 Islamabad Accepted
> print(head(clean_replace))
  Name Age Gender Admission.Test.Score High.School.Percentage City Admission.Status
1 Shehroz 24 Female 50 68.90000 Quetta Rejected
2 Waqar 21 Female 99 60.73000 Karachi Pending
3 Bushra 17 Male 89 75.68473 Islamabad Accepted
4 Aliya 17 Male 55 85.29000 Karachi Rejected
5 Blial 20 Male 65 61.13000 Lahore Pending
6 Murtaza 23 Female 0 75.68473 Islamabad Accepted
> # Verify remaining NAs
> print("---- Remaining NAs after replacement ----")
[1] "---- Remaining NAs after replacement ----"
> print(colSums(is.na(clean_replace)))
      Name      Age      Gender Admission.Test.Score High.School.Percentage      City
      10         10         0              0              0              0
Admission.Status
```

Files Explorer:

- My Music
- My Pictures
- My Videos
- My Web Sites
- NetBeansProjects
- Power BI Desktop
- S103 Cn 1st.pdf (2.6 MB, Nov 1)
- scals for DS
- Sound Recordings
- T050_AnjaliT
- vgsales.csv (1.3 MB, Nov 1)
- Virtual Machines
- Visual Studio 18
- Visual Studio 2022
- WindowsPowerShell
- air_traffic_data.csv (1.9 MB, Dec 1)
- student_admission_record_dirty.csv (6.7 KB, Dec 1)