PARSHWANATH CHARITABLE TRUST'S



Subject Incharge: Prof. Sarala Mary Page No._

A.P. SHAH INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering
Data Science



Department of CSE-Data Science | APSIT

Semester :VI	Subject : DAV	Acad	demic Year: 20 23 20 20
DIRTY TATA MATA	DATA METALING	N D.	
Dala cleaning	in R is the pro	uss 16 trans	form Hawaii
into consistent o	dala thal can be	easily analy	1300
aimed at filterin	g the content of	statistical sto	tements
based on the d	alā as well as	their reliabil	राष्ट्र :
Purpose of Dala	dealing.	5 - 5 1	- Jamina
The following	are the various	purpose of dal	a cleaning.
* Eliminali	rors		
* Eliminale	Redundancy		
* Increase	Dala Reliability		
* Accuracy			
Inquire Co	onsistency		
Aceure Co	ompleteres.		
		n .	un de llowing
let us consider	starwars datasel	and perform	True gers (
data deaning u	urng R:		
11) Select variables	l .		
(2) Filter variable	\$		
(3) Missing Dala			
(4) Duplicate Van	ies.		
1-1-1-1-1-1-1-1-1-1	•	200	
Hond and Well	office occurre		
Ochrany Chagvers	e)		
new (darwars)	coloumn names in	glarwan datere	f .
# to display the	colournis	1	
names (slarwar	4)		





and calculate the mean.

Subject Incharge: Prof. Sarala Mary Page No. 2

A.P. SHAH INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering

Data Science



Department of CSE-Data Science | APSIT

	Semester : VI Subject : DAY Academic Year: 2023- 20 2
	output:
	'name', 'huight', 'mase', 'hair-color', 'skin-color', 'eye-color',
	'name', 'huight', 'mase', 'hair-color', 'skin-color', 'eye-color', 'birth-year', 'sex', 'gender', 'homeworld', 'species', 'films',
	'vehicles', 'slärships'.
	the to display the colour names name, height and variables that
	ends with word color.
	1 - 0/ 0/
	select (name, height, ends-will
	(2) Filter observations: #Display the unique values present in the coloumn hair color
	# Display the unique values present in the
	unique (slarwars & hair_color).
	# Display only the nows that contain haircolor types as
	#Display only the rows that contain haircolor types as blond brown and having height less than 180.
	slarwars %>% 8 elect (name, height, ends-with ("color")) %>% 8 elect (name, height, ends-with ("color")) %>% 8 height < 180)
	select (name, height, ends-with (color) & height <180) felter (hair-color %in% c("blond", "brown") & height <180)
	felter (hair color rolling and display only the required
	This will felter the datas and display only the required
П	once.
	(3) Missing Data: # Calculate the mean of height coloumn. # Calculate the mean of height coloumn.
١.	te Calculate via "]
	mean (slarware sherght) mean (slarware sherght) since the coloumn consists
	This will not give proper out this we use the below code.
	mean (slarwars sheight) This will not give proper output, since the coloumn consists of missing value 'NA!. To overcome this we use the below code.
	mean (8 la round)
	output:
	174.60
•	The name will eliminate the sours that has missing values



A.P. SHAH INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering
Data Science



Semester : VI Academic Year: 20 23- 2024 Subject : DAV # To remove all NA from coloumn name, gender, hair-color, height. 8 arwars % >% Select (name, gender, hair-color, height) %>% na.omitc) This will remove the rows that how NA in the mentioned coloumn name and displays the output. It suppose we want to keep some NA of few colours and # To dislapy, only the slows that has NA values in coloumn name, gender, heir-color, height. Slarwars %>% select (name, gender, hair-color, height) %>% felter (1 complete. cases (.)) # Delete the 91000s with NA values in height coloumn. Starwars %>% Select (name, gender, hair-color, height) %>% filter (!complete.cases(.)) %>% #Replace the slower with NA in coloumn hair_color with the value = "none". setect (name, gender, hair-color, height) %>% slarwars %>% filter (!complete.cases(.)) %>7% drop-na (height) %>% mutate (halr-color = replace na (halr-color, "none")) The mutater, method will replace the null values is coloumn hair-color Subject Incharge: Prof. Sarala Mary Page No. 3 Department of CSE-Data Science | APSIT



A.P. SHAH INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering **Data Science**



Semester: Subject: Academic Year: 20.03-20
with value none. Academic Year: 20 23 20
(4) Duplicate values:
Lets create a data frame with duplicate values:
Name < c("Peter", "John", "Mark", "Peter")
Age← c(&2,33,44,22)
emp < data. frame (Names, Age)
This code will display the dataframe with 2x4. values.
In this Peter is a duplicated 91000. #Remove the duplicated value and display only the unique value
Remove the duplicated value and display only the wings
emp[!dup rcated (emp) 1]
These are the few enamples of handling dirty data.
a manager of Cleandala:
Characteristics of characters complete, and in a format that is

ready to analyze. The characteristics are as follows.

- * Free of duplicate rows.
- * Froot-free.
- * Free of missing values.
- * Free of outliers.
- * Appropriate data lype for analysis.