Semester : VI Subject : DAV Academic Year: 20 23 20 24

DIRTY DATA (or) DATA CLEANING IN R:

Data cleaning in R is the process to transform raw data into consistent data that can be easily analyzed. It is aimed at filtering the content of statistical statements based on the data as well as their reliability.

Purpose of Data cleaning:

The following are the various purpose of data cleaning.

- Eliminate Errors
- Eliminate Redundancy
- Increase Data Reliability
- Accuracy.
- Ensure Consistency
- Assure Completeness
- Standardize your approach.

let us consider starwars dataset and perform the following data cleaning using R:

(1) Select variables
(2) Filter variables
(3) Missing Data
(4) Duplicate values.

(1) Select variables:

```
#Load and view the dataset.
 library(tidyverse)
 view(starwars)
#To display the coloumn names in starwars dataset.
 names(starwars)
```

Semester : VI Subject : DAV Academic Year: 2023- 2024

Output:

'name', 'height', 'mass', 'hair_color', 'skin_color', 'eye_color', 'birth_year', 'sex', 'gender', 'homeworld', 'species', 'films', 'vehicles', 'starships'.

To display the coloumnames name, height and variables that ends with word color.

```
starwars %>%
   select (name, height, ends_with ("color"))
```

(2) Filter observations:

#Display the unique values present in the coloumn hair_color

```
unique(starwars $ hair_color).
```

#Display only the rows that contain haircolor types as blond, brown and having height less than 180.

```
starwars %>%
   select (name, height, ends_with ("color")) %>%
   filter ( hair_color %in% c("blond", "brown") & height <180)
```

This will filter the datas and display only the required once.

(3) Missing Data:

Calculate the mean of height coloumn.

```
mean (starwars $height)
```

This will not give proper output, since the coloumn consists of missing value 'NA'. To overcome this we use the below code.

```
mean( starwars$height, na.rm = TRUE)
```

Output:

```
174.60
```

The na.rm will eliminate the rows that has missing values and calculate the mean.

Semester : VI Subject : DAV Academic Year: 2023- 2024

To remove all NA from coloumn name, gender, hair_color, height.

```
starwars %>%
   select(name, gender, hair_color, height) %>%
   na.omit()
```

This will remove the rows that has NA in the mentioned coloumn name and displays the output.

If suppose we want to keep ~~some~~ NA of few coloums and delete the remaining.

To dislapy only the rows that has NA values in coloumn name, gender, hair_color, height.

```
starwars %>%
  select (name, gender, hair_color, height) %>%
  filter ( !complete. cases(.))
```

Delete the rows with NA values in height coloumn.

```
starwars %>%
   select (name, gender, hair_color, height) %>%
   filter (!complete.cases(.)) %>%
   drop_na(height)
```

Replace the rows with NA in coloumn hair_color with the value = "none".

```
starwars %>%
   select (name, gender, hair_color, height) %>%
   filter(!complete.cases(.)) %>%
   drop_na(height) %>%
    mutate ( hair_color = replace_na( hair_color, "none"))
```

The mutate() method will replace the null values in coloumn hair_color with value none.

(4) Duplicate values:

Lets create a dataframe with duplicate values:

```
Name ← c("Peter", "John", "Mark", "Peter")
Age ← c(22, 33, 44, 22)
emp ← data.frame(Names, Age)
emp
```

This code will display the dataframe with 2x4. values.
coloums rows.

In this Peter is a duplicated row.

```
#Remove the duplicated value and display only the unique values.
  emp[!duplicated(emp),]
```

These are the few examples of handling dirty data.

Characteristics of Clean data:

Cleandata is accurate, complete, and in a format that is ready to analyze. The characteristics are as follows.

- Free of duplicate rows.
- Error-free.
- Free of missing values.
- Free of outliers.
- Appropriate data type for analysis.