21BDS0340

Abhinav Dinesh Srivatsa

Programming for Data Science Lab

Digital Assignment – II

Code

```
# Load necessary libraries
library(dplyr)

setwd("/Users/abhi/College Work/Year 3 Semester 2 (Sem 6)/Programming for Data
Science Lab/Assignment 2/")
print(getwd())
file_path <- "Contacts.csv"

df <- read.csv(file_path)

# Displaying a summary of the dataset
summary(df)

# Check for missing values
missing_values <- sum(is.na(df))
cat("Number of missing values:", missing_values, "\n")

# Handle missing values (replace with mean, median, or remove)
# Example: Replace missing values with the mean of the column
df <- df %>% mutate_if(is.character, ~ifelse(is.na(.), "empty", .))

# Remove duplicate rows
df <- distinct(df)

# Check for outliers and handle them if necessary
# Example: Remove outliers from a numeric column
# outliers <- boxplot.stats(df$numeric_column)$out
# df <- df %>% filter(!numeric_column %in% outliers)

# Save the cleaned dataset to a new file
# Replace 'cleaned_dataset.csv' with your desired file name
write.csv(df, 'cleaned_dataset.txt', row.names = FALSE)
head(df)
```

Output

```
> # Load necessary libraries
> library(dplyr)
>
> setwd("/Users/abhi/College Work/Year 3 Semester 2 (Sem 6)/Programming for Data
Science Lab/Assignment 2/")
> print(getwd())
```

```
[1] "/Users/abhi/College Work/Year 3 Semester 2 (Sem 6)/Programming for Data
Science Lab/Assignment 2"
> file_path <- "Contacts.csv"
>
> df <- read.csv(file_path)
>
> # Displaying a summary of the dataset
> summary(df)
    Source            FirstName           LastName          Companies       Title
Emails
 Length:68          Length:68          Length:68          Mode:logical
Mode:logical   Length:68
 Class :character   Class :character   Class :character   NA's:68        NA's:68
Class :character
 Mode  :character   Mode  :character   Mode  :character
Mode  :character
 PhoneNumbers        CreatedAt          Addresses       Sites
InstantMessageHandles FullName
 Length:68          Length:68          Mode:logical   Mode:logical   Mode:logical
Mode:logical
 Class :character   Class :character   NA's:68        NA's:68        NA's:68
NA's:68
 Mode  :character   Mode  :character
 Birthday       Location       BookmarkedAt   Profiles
 Mode:logical   Mode:logical   Mode:logical   Mode:logical
 NA's:68        NA's:68        NA's:68        NA's:68

>
> # Check for missing values
> missing_values <- sum(is.na(df))
> cat("Number of missing values:", missing_values, "\n")
Number of missing values: 680
>
> # Handle missing values (replace with mean, median, or remove)
> # Example: Replace missing values with the mean of the column
> df <- df %>% mutate_if(is.character, ~ifelse(is.na(.), "empty", .))
>
> # Remove duplicate rows
> df <- distinct(df)
>
> # Check for outliers and handle them if necessary
> # Example: Remove outliers from a numeric column
> # outliers <- boxplot.stats(df$numeric_column)$out
> # df <- df %>% filter(!numeric_column %in% outliers)
>
> # Save the cleaned dataset to a new file
> # Replace 'cleaned_dataset.csv' with your desired file name
> write.csv(df, 'cleaned_dataset.txt', row.names = FALSE)
> head(df)
           Source             FirstName LastName Companies Title
Emails   PhoneNumbers
```

```
1 MOBILE_CONTACTS                    Me                NA    NA
tanushsrivatsa@gmail.com    74836 85981
2 MOBILE_CONTACTS Shyam Sir - Chemistry                NA    NA
shyamsundermatta@gmail.com    99581 37588
3 MOBILE_CONTACTS                Archit   Murali       NA    NA
+917517066578
4 MOBILE_CONTACTS                 Swami    Vibhu       NA    NA
+919442504602
5 MOBILE_CONTACTS               Nishank      Das       NA    NA
+916299926210
6 MOBILE_CONTACTS                 Ayush   Mishra       NA    NA
+917581903399
         CreatedAt Addresses Sites InstantMessageHandles FullName Birthday Location
BookmarkedAt Profiles
1 6/24/23, 5:22 AM        NA    NA                   NA       NA       NA       NA
NA       NA
2 6/24/23, 5:22 AM        NA    NA                   NA       NA       NA       NA
NA       NA
3 6/24/23, 5:22 AM        NA    NA                   NA       NA       NA       NA
NA       NA
4 6/24/23, 5:22 AM        NA    NA                   NA       NA       NA       NA
NA       NA
5 6/24/23, 5:22 AM        NA    NA                   NA       NA       NA       NA
NA       NA
6 6/24/23, 5:22 AM        NA    NA                   NA       NA       NA       NA
NA       NA
```