Data Cleaning using R

Spoken Tutorial Project
https://spoken-tutorial.org
National Mission on Education through ICT
https://sakshat.ac.in

Tanmay Srinath
Madhuri Ganapathi
IIT Bombay

6 January 2022

Learning Objectives

We will learn about:

Learning Objectives

We will learn about:

- **Data Cleaning**

Learning Objectives

We will learn about:

- **Data Cleaning**
- **Reading Data from a text file**

Learning Objectives

We will learn about:

- **Data Cleaning**
- **Reading Data from a text file**
- **Type conversions**

Learning Objectives

We will learn about:

- **Data Cleaning**
- **Reading Data from a text file**
- **Type conversions**
- **Handling NA values**

Learning Objectives

We will learn about:

- Data Cleaning
- Reading Data from a text file
- Type conversions
- Handling NA values
- Encoding values to factors

System Specifications

System Specifications

- **Ubuntu Linux OS version 20.04**

System Specifications

- **Ubuntu Linux OS version 20.04**
- **R version 4.1.2**

System Specifications

- **Ubuntu Linux OS version 20.04**
- **R version 4.1.2**
- **RStudio version 1.4.1717**

System Specifications

- **Ubuntu Linux OS version 20.04**
- **R version 4.1.2**
- **RStudio version 1.4.1717**

System Specifications

- **Ubuntu Linux OS version 20.04**
- **R version 4.1.2**
- **RStudio version 1.4.1717**

Install R version 4.1.0 or higher

Pre-requisites

Pre-requisites

- **Basics of R Programming**

Pre-requisites

- **Basics of R Programming**
- **Dataframes, lists and vectors**

Pre-requisites

- **Basics of R Programming**
- **Dataframes, lists and vectors**

Pre-requisites

- **Basics of R Programming**
- **Dataframes, lists and vectors**

If not, please access the relevant tutorials on R on
https://spoken-tutorial.org/

What is Data Cleaning?

What is Data Cleaning?

Data Cleaning:

What is Data Cleaning?

Data Cleaning:

- **It involves detecting and correcting corrupt or inaccurate records in a dataset**

What is Data Cleaning?

Data Cleaning:

- **It involves detecting and correcting corrupt or inaccurate records in a dataset**
- **The correction may also lead to removal of specific inaccurate record**

Need for Data Cleaning

Need for Data Cleaning

Need for data cleaning:

Need for Data Cleaning

Need for data cleaning:

- **It improves data quality and data reliability**

Need for Data Cleaning

Need for data cleaning:

- **It improves data quality and data reliability**
- **Delivers accuracy and ensures consistency in data**

Need for Data Cleaning

Need for data cleaning:

- **It improves data quality and data reliability**
- **Delivers accuracy and ensures consistency in data**
- **Ensures that data is set for statistical analysis**

Reading Data from Text File

Reading Data from Text File

- **Data might not be available in convenient forms like CSV files**

Reading Data from Text File

- **Data might not be available in convenient forms like CSV files**
- **We need to learn how to extract data from text files**

Download Files

Download Files

We will use:

Download Files

We will use:

- **A script file** `DataCleaning.R`

Download Files

We will use:

- **A script file** `DataCleaning.R`
- **A dataset** `airquality.txt`

Download Files

We will use:

- **A script file** `DataCleaning.R`
- **A dataset** `airquality.txt`

Download Files

We will use:

- **A script file** `DataCleaning.R`
- **A dataset** `airquality.txt`

Download these files from the Code files link of this tutorial

Make a copy and then use them for practising

Purpose of Type Conversion

Purpose of Type Conversion

- **Most functions in R work solely using numeric data**

Purpose of Type Conversion

- **Most functions in R work solely using numeric data**
- **Hence, type conversion will make data suitable for analysis**

Handling NA values

Handling NA values

There are two ways of handling NA values:

Handling NA values

There are two ways of handling NA values:

- **Removing NA values**

Handling NA values

There are two ways of handling NA values:

- **Removing NA values**
- **Replacing them with appropriate values**

Replacing Missing Values

Replacing Missing Values

- **Removing NA values makes sense when we have few such entries**

Replacing Missing Values

- Removing NA values makes sense when we have few such entries
- However, removing a lot of missing values without replacement leads to data loss

Replacing Missing Values

- Removing NA values makes sense when we have few such entries
- However, removing a lot of missing values without replacement leads to data loss
- We should know how to replace the missing values

Summary

In this tutorial we have learnt about:

- **Data Cleaning**
- **Reading Data from a text file**
- **Type conversions**
- **Handling NA values**
- **Encoding values to factors**

Assignment

- **In the air quality dataset, replace the NA values with the mean of observations**

About the Spoken Tutorial Project

- Watch the video available at https://spoken-tutorial.org/What_is_a_Spoken_Tutorial
- It summarises the Spoken Tutorial project
- If you do not have good bandwidth, you can download and watch it

Spoken Tutorial Workshops

The Spoken Tutorial Project Team

- **Conducts workshops using spoken tutorials**
- **Gives certificates to those who pass an online test**
- **For more details, please write to contact@spoken-tutorial.org**

Answers for THIS Spoken Tutorial

- **Questions in THIS Spoken Tutorial?**
- **Visit https://forums.spoken-tutorial.org**
- **Choose the minute and second where you have the question**
- **Explain your question briefly**
- **The FOSSEE project will ensure an answer**

You will have to register to ask questions

Forum to answer questions

- Questions not related to the Spoken Tutorial?
- Do you have general / technical questions on the Software?
- Please visit the FOSSEE Forum

  https://forums.fossee.in/
- Choose the Software and post your question

Textbook Companion Project

- **The FOSSEE team coordinates the coding of solved examples of popular books and case study projects**
- **We give certificates to those who do this**

For more details, please visit these sites:

https://r.fossee.in/

https://fossee.in/

Acknowledgements

- **The** Spoken Tutorial **and** FOSSEE **projects are funded by the** Ministry of Education**, Govt. of India**

About the Contributors

- **This tutorial is contributed by Tanmay Srinath and Madhuri Ganapathi, IIT Bombay**