Data Manipulation using dplyr Package

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

http://sakshat.ac.in/

Script: Varshit Dubey
Narration: Sudhakar Kumar

IIT Bombay
7 July 2019

Learning Objectives

We will learn about:

Learning Objectives

We will learn about:

- **Data manipulation**

Learning Objectives

We will learn about:

- **Data manipulation**
- **dplyr package**

Learning Objectives

We will learn about:

- Data manipulation
- dplyr package
- How to use filter and arrange functions

Pre-requisites

Pre-requisites

- **Basics of Statistics**

Pre-requisites

- Basics of Statistics
- Basics of ggplot2 package

Pre-requisites

- Basics of Statistics
- Basics of ggplot2 package
- Data frames

Pre-requisites

- Basics of Statistics
- Basics of ggplot2 package
- Data frames

Please locate the relevant tutorials on
https://spoken-tutorial.org/

System Specifications

System Specifications

- **Ubuntu Linux OS v 16.04**

System Specifications

- **Ubuntu Linux OS v 16.04**
- **R v 3.4.4**

System Specifications

- **Ubuntu Linux OS v 16.04**
- **R v 3.4.4**
- **RStudio v 1.1.463**

System Specifications

- **Ubuntu Linux OS v 16.04**
- **R v 3.4.4**
- **RStudio v 1.1.463**

R version 3.2.0 or higher

Download Files

We will use:

Download Files

We will use:

- **A data frame moviesData.csv**

Download Files

We will use:

- **A data frame moviesData.csv**
- **A script file myVis.R**

Download Files

We will use:

- A data frame moviesData.csv
- A script file myVis.R

Download these files from the Code files link of this tutorial

Need for Data Manipulation

Need for Data Manipulation

In real life, it is rare that we get the data in exactly the right form we need

Need for Data Manipulation

Often we'll need to

Need for Data Manipulation

Often we'll need to

- **create some new variables or summaries**

Need for Data Manipulation

Often we'll need to

- **create some new variables or summaries**
- **rename the variables**

Need for Data Manipulation

Often we'll need to

- **create some new variables or summaries**
- **rename the variables**
- **reorder the observations in order to make the data a little easier to work with**

About dplyr Package

About dplyr Package

- **dplyr is a package for data manipulation, written and maintained by Hadley Wickham**

About dplyr Package

- **dplyr** is a package for data manipulation, written and maintained by Hadley Wickham
- It comprises many functions that perform mostly used data manipulation operations

Functions in dplyr Package

Functions in dplyr Package

- **filter - to select cases based on their values**

Functions in dplyr Package

- **filter** - to select cases based on their values
- **arrange** - to reorder the cases

Functions in dplyr Package

- **filter** - to select cases based on their values
- **arrange** - to reorder the cases
- **select** - to select variables based on their names

Functions in dplyr Package

- **filter** - to select cases based on their values
- **arrange** - to reorder the cases
- **select** - to select variables based on their names
- **mutate** - to add new variables that are functions of existing variables

Functions in dplyr Package

- **summarise** - **to condense multiple values to a single value**

Functions in dplyr Package

- **summarise** - to condense multiple values to a single value

All these functions can be combined with group_by function. It allows us to perform any operation by a group

Summary

We have learnt about:

- Data manipulation
- dplyr package
- How to use filter and arrange functions

Assignment

1. **Consider the built-in data set *mtcars*. Find the cars with *hp* greater than 100 and *cyl* equal to 3**
2. **Arrange the *mtcars* data set based on *mpg* variable**

About the Spoken Tutorial Project

- Watch the video available at http://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**

Forum to answer questions

- Do you have questions in THIS Spoken Tutorial?
- Choose the minute and second where you have the question
- Explain your question briefly
- Someone from the FOSSEE team will answer them. Please visit

http://forums.spoken-tutorial.org/

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

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

Textbook Companion Project

- The FOSSEE team coordinates coding of solved examples of popular books
- We give honorarium and certificates to those who do this

For more details, please visit these sites:

https://r.fossee.in/

https://fossee.in/

Acknowledgements

- Spoken Tutorial Project is a part of the Talk to a Teacher project
- It is supported by the National Mission on Education through ICT, MHRD, Government of India
- More information on this Mission is available at: http://spoken-tutorial.org/NMEICT-Intro

Thank You

- The script for this tutorial was contributed by Varshit Dubey (CoE Pune)
- The video has been created by Sudhakar Kumar, IIT Bombay