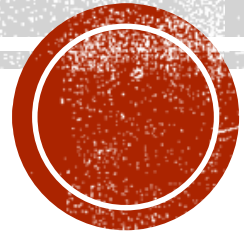
DATA ANALYSIS

Data cleaning and Summarizing with dplyr package



DATA CLEANING AND SUMMARIZING WITH DPLYR PACKAGE

Topics of this session

- Introduction of *dplyr* package in R

- Several examples and tips of how to use dplyr package for cleaning and transforming data.

- It's a complete tutorial on data manipulation and data wrangling with R

DPLYR PACKAGE

- One of the most powerful and popular package in R.
- written by the most popular R programmer **Hadley Wickham**
- He has written many useful R packages such as **ggplot2, tidyr** etc.

WHAT IS DPLYR?

- A powerful R-package to manipulate
- Useful to clean and summarize unstructured data.
- In short, it makes data exploration and data manipulation easy and fast in R.

WHAT'S SPECIAL ABOUT DPLYR?

- Comprises many functions that perform mostly used data manipulation operations such as:
  - applying filter
  - selecting specific columns
  - sorting data
  - adding or deleting columns
  - aggregating data

WHAT'S SPECIAL ABOUT DPLYR?

- it's very easy to learn and use dplyr functions.

- Also easy to recall these functions.

  For example, filter() is used to filter rows

  select() is used to select columns

DPLYR VS. BASE R FUNCTIONS

- dplyr functions process faster than base R functions as they were written in a computationally efficient manner.
- They are also more stable in the syntax
- Better supports data frames than vectors.

SQL – STRUCTURED QUERY LANGUAGE

- People have been utilizing SQL for analyzing data for decades.
- Every modern data analysis software such as Python, R, SAS etc supports SQL commands.
- But SQL was never designed to perform data analysis.
- It was rather designed for querying and managing data.

SQL QUERIES VS. DPLYR

- There are many data analysis operations where SQL fails or makes simple things difficult.
- Example, calculating median for multiple variables, converting wide format data to long format etc.
- Whereas, dplyr package was designed to do data analysis.

HOW TO INSTALL AND LOAD DPLYR PACKAGE

- To install the dplyr package, type the following command.

install.packages("dplyr")

- To load dplyr package, type the command below

library(dplyr)

IMPORTANT DPLYR FUNCTIONS TO REMEMBER

dplyr Function	Description	Equivalent SQL
select()	Selecting columns (variables)	SELECT
filter()	Filter (subset) rows.	WHERE
group_by()	Group the data	GROUP BY
summarise()	Summarise (or aggregate) data	-
arrange()	Sort the data	ORDER BY
join()	Joining data frames (tables)	JOIN
mutate()	Creating New Variables	COLUMN ALIAS

DATA : INCOME DATA BY STATES

- Example data contains income generated by states from year 2002 to 2015.

- This dataset contains 51 observations (rows) and 16 variables (columns).

	Index	State	Y2002	Y2003	Y2004	Y2005	Y2006	Y2007	Y2008	Y2009
1	A	Alabama	1296530	1317711	1118631	1492583	1107408	1440134	1945229	1944173
2	A	Alaska	1170302	1960378	1818085	1447852	1861639	1465841	1551826	1436541
3	A	Arizona	1742027	1968140	1377583	1782199	1102568	1109382	1752886	1554330
4	A	Arkansas	1485531	1994927	1119299	1947979	1669191	1801213	1188104	1628980
5	C	California	1685349	1675807	1889570	1480280	1735069	1812546	1487315	1663809
6	C	Colorado	1343824	1878473	1886149	1236697	1871471	1814218	1875146	1752387

	Y2010	Y2011	Y2012	Y2013	Y2014	Y2015
1	1237582	1440756	1186741	1852841	1558906	1916661
2	1629616	1230866	1512804	1985302	1580394	1979143
3	1300521	1130709	1907284	1363279	1525866	1647724
4	1669295	1928238	1216675	1591896	1360959	1329341
5	1624509	1639670	1921845	1156536	1388461	1644607
6	1913275	1665877	1491604	1178355	1383978	1330736