Managing Data Frames with dplyr - Introduction

Contents

1 Managing Data Frames with dplyr - Introduction

- arrange
- filter
- select
- mutate
- rename

1.1 dplyr

The data frame is a key data structure in statistics and in R. There is one observation per row Each column represents a variable or measure or characteristic Primary implementation that you will use is the default R implementation Other implementations, particularly relational databases systems

Developed by Hadley Wickham of RStudio An optimized and distilled version of plyr package (also by Hadley) Does not provide any "new" functionality per se, but greatly simplifies existing functionality in R Provides a "grammar" (in particular, verbs) for data manipulation Is very fast, as many key operations are coded in C++

1.2 dplyr Verbs

`select`: return a subset of the columns of a data frame filter: extract a subset of rows from a data frame based on
logical conditions `arrange`: reorder rows of a data frame
`rename`: rename variables in a data frame `mutate`: add new variables/columns or transform existing variables `summarise` / `summarize`: generate summary statistics of summarise / summarize: generate summary statistics of dierent variables in the data frame, possibly within strat

There is also a handy print method tha

1.3 dplyr Properties

The first argument is a data frame. I The subsequent arguments describe what to do with it, and you can refer to columns in the data frame directly without using the $ operator (just use the names). I The result is a new data frame I Data frames must be properly formatted and annotated for this to all be useful