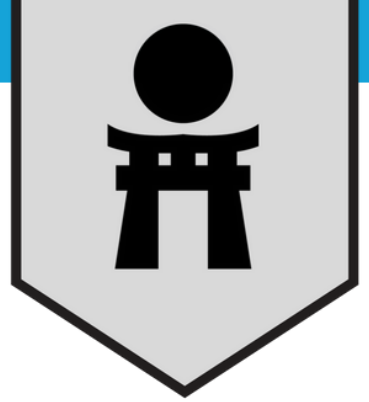

Community Talk

Data Manipulation with dplyr

datasciencedojo
data science for everyone

About Arham Akheel

- Business Analyst and Instructor at Data Science Dojo
- Worked as an Application Engineer
  - Schematics
  - Assembly
- Worked in a startup, managed
  - Operations
  - Information Systems
- Masters in Technology Management.

Introduction to dplyr

- dplyr provides efficient manipulation of datasets

- Perform group-by aggregation on a dataset, select columns, filter rows, and find overlapping and distinct values from two data sources.

- Functions: *arrange, group_by, summarize, select, filter, intersect, setdiff,* etc.

Introduction to dplyr

- Goal: To perform basic data manipulation and be able to holistically approach and dissect data using group-by aggregation.
- Demonstration with common functions of dplyr while working with an example dataset on wine ratings from Kaggle.
- Quick way to get beginners up to speed and help them select segments they find useful.

Getting Started

- Obtain R, install RStudio, and load the wine ratings dataset from Kaggle.
- Install and load *dplyr* and *ggplot* packages.
- Walk-through of how to properly load the dataset into RStudio, including how to correctly import foreign characters.

Reshape, Subset, and Summarize

- Perform group-by aggregation using *group_by* and *summarize*.
- Manipulate order of data using *arrange*.
- Subset columns and rows using *select* and *filter*.

Feature Engineering

- Locate overlapping values from two data sources using *intersect*.
- Locate distinct values of two data sources using *setdiff*
- *Missingness imputation using ifelse from base R.*
- Feature engineering with *mutate* and *transmute* from dplyr.
- Different joins to combine datasets the way you want it with *full_join, inner_join, left_join, or right_join*.