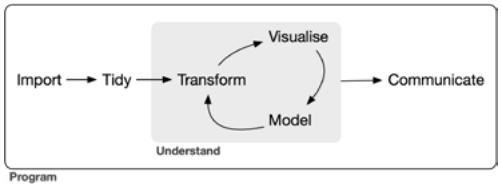
Spatial Data Science with R: An Introduction to the Tidyverse

R 4.2.2
Tidyverse 1.3.2

Welcome to the Data Science Task Sheet Series. This series supplements the Iowa State University Extension and Outreach Geospatial Technology Training Program's workshops and short courses by providing quick and easy instructions for performing a variety of mapping, data science, analysis, and visualization tasks.

The tidyverse is a collection of open source R packages that are useful for data scientists. With the packages loaded into projects, the common API design provides a coherent system, or toolbox, that makes it easy to perform common data science tasks such as importing, wrangling (tidying and manipulation), modeling data, presenting results, and programming. This task sheet will briefly introduce each of the core packages. For more information on the tidyverse, visit the official tidyverse website https://www.tidyverse.org. Additional task sheets utilizing features of the various tidyverse packages are also available in this task sheet series.

1. Tidyverse Overview



Developed initially by Hadley Wickham (PhD 2008, Iowa State University), the tidyverse utilizes a workflow, or model, of data science tools as illustrated in "R for Data Science (Wickham and Grolemund 2017)." This process starts with importing data, putting the data into a tidy format where every column is a variable, every row is an observation, every cell contains a single value and collectively this forms a table. Transforming, visualizing, and modeling are then used to understand the data before communicating the results to someone else. Additionally, the entire workflow can be programed to create new tools for data science projects.

While tidyverse provides many benefits, one drawback is that it can be slower than base R or data.table for some operations. However, one should take into consideration the ease in writing and reading tidyverse code to best determine which approach best meets the project's needs.

2. Installing Tidyverse Packages

The tidyverse consists of over 26 packages that follow regular patterns and syntax for variables, functions, and operations. This consistency ensures that a user can intuitively connect a sequence of commands to create a tidy pipeline. The eight most common packages, ggplot2, dplyr, tidyr, readr, purr, tibble, stringr and forcats, form the core of the tidyverse and can be installed and loaded with `install.packages("tidyverse")` and `library("tidyverse")`. Packages may also be loaded independently of each other. For example, use just ggplot2 with `install.packages("ggplot2")` and `library("ggplot2")`. One reason to load just the needed package(s) is to avoid naming conflicts with base R and other non-tidyverse packages. An example is `filter()` from `dplyr::filter()` which is loaded after base R and replaces the `filter()` from `stats::filter`. Note: using `stats::filter()` versus `filter()` solves the conflict. Run `tidyverse_conflicts()` to check for other conflicts.