Organizing the Project

Before we get started working with a dataset, we should set up our project. The simplest project setup entails creating a folder structure to organize the files associated with the project. Here is an example of a basic folder structure.

Creating a project structure will help to keep the files organized. It is also good practice to break code into separate files with distinct functional units and give code files informative names. For example, the fifth step in the project where we build a random forest model could be called '05_build_randomForest_model.R'.

Analysis is an inherently iterative process, so we will end up overwriting code files over time. To avoid cluttering our file system with many files with '_v1', '_v2', '_v3', and so on, we recommend using some form of version or source control in developing your project. The basic idea is that project files are regularly copied to a repository where the differences between the new and old versions are automatically captured to allow the user to track changes and revert to previous versions. Version control software consists of a user interface or client that controls an underlying version control system. Popular version control systems include git and svn. Examples of Git clients include TortoiseGit and SmartGit.