# R Developer Case Study

Streamline builds custom data pipelines to organize data for our clients.  This code runs on a schedule whenever the data is updated.  For many clients, the data is in Excel spreadsheets with varied structures.  Many of these structures are not rectangular and one of our first tasks is to reformat the data.

In this coding challenge, you are tasked with creating an R package for data cleaning and manipulation of the Enron email dataset, specifically an XLS attachment of data, and a small Shiny app to display results. The Enron email dataset was made public following the company's collapse and contains emails from Enron employees. Your task is to write a package that includes functions for cleaning the dataset. The package should also include documentation and tests for each function.  The app should use the cleaned data to display things

**Data**

The data is from Enron, which was a gas pipeline company.  After a scandal, a number of their spreadsheets were released to the public, so this example represents a real format that analysts have in practice.  The original spreadsheet is located at https://github.com/StreamlineDataScience/enron-example/blob/main/andrea_ring_000_1_1.pst.0.xls. You can also see a version on Google Sheets.

The data contains deliveries and receipts for a number of locations/facilities (e.g. ACADIAN, BRIDGELINE).  We would like the data to be reformatted into the following data columns: location, date, deliveries, receipts, where date is a date object, which should go from 2001-09-01 to 2002-02-05, with all dates included. Prior month averages and the month averages can be removed.  The CSV located at https://github.com/StreamlineDataScience/enron-example/blob/main/long_enron_data.csv can be used as a test for the correct final output.

**Instructions**

1. Fork the https://github.com/StreamlineDataScience/str.rdev repository
2. Create an R package named "str.rdev" with the following functions:

   a. enron_download_data() - A function that downloads the Enron XLS file.
   b. enron_process_data() - A function that cleans the dataset by transforming the columns into location, date, deliveries, receipts.

3. Create a unit test in the R package that compares the output to the benchmark CSV.
4. Create a small Shiny app (one app.R file) in it with 1 visualization of the data, specifically a line plot of receipts over time for each location.
   a. The app should live in the inst/app folder (a blank app.R is in StreamlineDataScience/str.rdev). We will install your version of str.rdev, and run shiny::runApp(system.file(“app”, package = "str.rdev")) to see the output.  Please test accordingly.

The package should pass R CMD check –as-cran with no warnings, errors, or notes. This will be checked via GitHub actions.

**Deliverables**

1. A pull request (PR) of the package that includes the above functions and documentation for each function.
   a. (If you would like to not send a PR to this public repository, we can provide you with a private repo option)
2. A paragraph or 2, as a README of the repository, that describes where there’d be issues with generalizability of the code.  For example, “My code relies on columns 1-5 containing this data.  If that shifted, the above code would break”.  Overall, this explanation is a discussion of the limitations and robustness of the code that you have identified.
3. Information about any insights from the Shiny app described in another section like 2. above, either about improvements to the app or what would be good to add in..

We hope the project will take you < 2.5 hours but you may spend as much time as you prefer.

Note: You are free to use any R packages that you find useful for the analysis and package development.