Achieving Full Reproducibility in R: A Docker and renv Strategy

Research Computing Team

2025-05-13

Abstract

This white paper presents a comprehensive approach to achieving reproducibility in R workflows by combining two powerful tools: renv for R package management and Docker for containerizing the computing environment. Together, these tools ensure that an R workflow runs identically across different systems with the same packages, R version, and system libraries as the original setup. The paper includes a practical case study demonstrating collaborative development using this approach.

Table of contents

1 Executive Summary

Reproducibility stands as a cornerstone of professional data analysis, yet in practice, achieving it consistently with R workflows remains challenging. R projects frequently break when transferred between computers due to mismatched R versions or package dependencies, leaving developers in what is colloquially known as "dependency hell." This white paper presents a comprehensive approach to solving this problem by combining two powerful tools: **renv** for R package management and **Docker** for containerizing the computing environment. Together, these tools ensure that an R workflow runs identically across different systems with the same packages, R version, and system libraries as the original setup.

2 Introduction

2.1 The Challenge of Reproducibility in R

R has become a standard tool for data science and statistical analysis across numerous disciplines. However, as R projects grow in complexity, they often develop intricate webs of dependencies that can make sharing and reproducing analyses difficult. Some common challenges include:

- Different R versions across machines
- Incompatible package versions
- Missing system-level dependencies
- Operating system differences
- Conflicts with other installed packages

These challenges often manifest as the frustrating "it works on my machine" problem, where analysis code runs perfectly for the original author but fails when others attempt to use it. This undermines the scientific and collaborative potential of R-based analyses.

2.2 A Two-Level Solution

To address these challenges comprehensively, we need to tackle reproducibility at two distinct levels:

1. **Package-level reproducibility**: Ensuring exact package versions and dependencies are maintained
2. **System-level reproducibility**: Guaranteeing consistent R versions, operating system, and system libraries

The strategy presented in this white paper leverages **renv** for package-level consistency and **Docker** for system-level consistency. When combined, they provide a robust framework for end-to-end reproducible R workflows.

3 renv: Package-Level Reproducibility

3.1 What is renv?

renv (Reproducible Environment) is an R package designed to create isolated, project-specific library environments. Instead of relying on a shared system-wide R library that might change over time, renv gives each project its own separate collection of packages with specific versions.

3.2 Key Features of renv

- **Isolated project library**: renv creates a project-specific library (typically in `renv/library`) containing only the packages used by that project. This isolation ensures that updates or changes to packages in one project won't affect others.
- **Lockfile for dependencies**: When you finish installing or updating packages, `renv::snapshot()` produces a `renv.lock` file - a JSON document listing each package and its exact version and source. This lockfile is designed to be committed to version control and shared with collaborators.
- **Environment restoration**: On a new machine (or when reproducing past results), `renv::restore()` installs the exact versions of packages specified in the lockfile. This creates an R package environment identical to the one that created the lockfile, provided the same R version is available.

3.3 Basic renv Workflow

The typical workflow with renv involves:

```
# One-time installation of renv
install.packages("renv")

# Initialize renv for the project
renv::init()  # Creates renv infrastructure

# Install project-specific packages
# ...

# Save the package state to renv.lock
renv::snapshot()

# Later or on another system...
renv::restore()  # Restore packages from renv.lock
```

While renv effectively handles package dependencies, it does not address differences in R versions or system libraries. This limitation is where Docker becomes essential.

4 Docker: System-Level Reproducibility

4.1 What is Docker?

Docker is a platform that allows you to package software into standardized units called containers. A Docker container is like a lightweight virtual machine that includes everything needed to run an application: the code, runtime, system tools, libraries, and settings.

4.2 Docker's Role in Reproducibility

While renv handles R packages, Docker ensures consistency for:

- **Operating system**: The specific Linux distribution or OS version
- **R interpreter**: The exact R version
- **System libraries**: Required C/C++ libraries and other dependencies
- **Computational environment**: Memory limits, CPU configuration, etc.

By running an R Markdown project in Docker, you eliminate differences in OS or R installation as potential sources of irreproducibility. Any machine running Docker will execute the container in an identical environment.

4.3 Docker Components for R Workflows

For R-based projects, a typical Docker approach involves:

1. **Base image**: Starting from a pre-configured R image (e.g., from the Rocker project)
2. **Dependencies**: Adding system and R package dependencies
3. **Configuration**: Setting working directories and environment variables
4. **Content**: Adding project files
5. **Execution**: Defining how the project should run

A simple Dockerfile for an R Markdown project might look like:

```
# Use R 4.1.0 on Linux as base image
FROM rocker/r-ver:4.1.0

# Set the working directory inside the container
WORKDIR /workspace

# Install renv and restore dependencies
RUN R -e "install.packages('renv', repos='https://cloud.r-project.org')"

# Copy renv lockfile and infrastructure
COPY renv.lock renv/activate.R /workspace/

# Restore the R package environment
RUN R -e "renv::restore()"

# Default command when container runs
CMD ["/bin/bash"]
```

This Dockerfile creates a consistent environment with a specific R version and packages, regardless of the host system.

5 Combining renv and Docker: A Comprehensive Approach

5.1 Why Use Both?

Using renv or Docker alone improves reproducibility, but combining them provides the most comprehensive solution:

- **Docker** guarantees the OS and R version
- **renv** guarantees the R packages and their versions
- **Together** they achieve end-to-end reproducibility from operating system to package dependencies

This combined approach creates a fully portable analytical environment that can be shared and will produce identical results across different computers.

5.2 Integration Strategy

The recommended workflow integrates renv and Docker in the following manner:

1. **Develop locally with renv**: Create your R project with renv to manage package dependencies.
2. **Snapshot dependencies**: Use `renv::snapshot()` to create a lockfile.
3. **Containerize with Docker**: Create a Dockerfile that uses a specific R version and incorporates the renv lockfile.
4. **Share both**: Distribute both the code (with lockfile) and the Docker configuration.
5. **Execute consistently**: Run analyses in the Docker container for guaranteed reproducibility.

This strategy ensures that your R Markdown documents and analyses will run identically for anyone who has access to your Docker container, regardless of their local setup.

6 Practical Example: Collaborative R Markdown Development

The following case study demonstrates how two developers can collaborate on an R Markdown project using renv and Docker to ensure reproducibility.

6.1 Project Scenario

Two data scientists are collaborating on an analysis of the Palmer Penguins dataset. Developer 1 will set up the initial project structure and create a basic analysis. Developer 2 will extend the analysis with additional visualizations. They'll use GitHub for version control and DockerHub to share the containerized environment.

6.2 Step-by-Step Implementation

6.2.1 Developer 1: Project Setup and Initial Analysis

Step 1: Create and Initialize the GitHub Repository

Developer 1 creates a new GitHub repository called "penguins-analysis" and clones it locally:

```
git clone https://github.com/username/penguins-analysis.git
cd penguins-analysis
```

Step 2: Initialize renv for Dependency Management

```
install.packages("renv")  # If not already installed
renv::init()  # Initialize renv for the project
```

This creates the necessary renv infrastructure, including an initial `renv.lock` file.

Step 3: Install Required R Packages

```
install.packages("ggplot2")
install.packages("palmerpenguins")
renv::snapshot()  # Save package versions to renv.lock
```

Step 4: Create Initial R Markdown Analysis

Developer 1 creates a file named `peng1.Rmd` with the following content:

Listing 1 `peng1.Rmd`

````
---
title: "Palmer Penguins Analysis"
author: "Developer 1"
date: "`r Sys.Date()`"
output: html_document
---

```r
#| label: setup
#| include: false
library(ggplot2)
library(palmerpenguins)
```
````

7 Flipper Length vs. Bill Length

```
#| label: flipper-bill-plot
ggplot(penguins, aes(x = flipper_length_mm, y = bill_length_mm)) +
  geom_point() +
  theme_minimal() +
  ggtitle("Flipper Length vs. Bill Length")
```

````
**Step 5: Create a Dockerfile**

Developer 1 creates a Dockerfile that deliberately excludes the R Markdown file to ensure th

```{.dockerfile filename="Dockerfile"}
# Use R 4.1.0 as base image
FROM rocker/r-ver:4.1.0

# Set the working directory inside the container
WORKDIR /workspace
````
````

```
# Install renv and restore dependencies
RUN R -e "install.packages('renv', repos='https://cloud.r-project.org')"

# Copy only the renv.lock and renv infrastructure
COPY renv.lock renv/activate.R /workspace/

# Restore the R package environment
RUN R -e "renv::restore()"

CMD ["/bin/bash"]
```

Step 6: Build and Push the Docker Image

```
docker build -t username/penguins-analysis:v1 .
docker login
docker push username/penguins-analysis:v1
```

Step 7: Commit and Push to GitHub

Developer 1 commits the project files (excluding the R Markdown document from the Docker image):

```
git add .
git commit -m "Initial renv setup and Docker environment (without Rmd)"
git push origin main
```

Step 8: Communicate with Developer 2

Developer 1 provides these instructions to Developer 2:

1. Clone the GitHub repository
2. Pull the prebuilt Docker image from DockerHub
3. Run the container interactively, mounting the local repository
4. Extend the analysis in the peng1.Rmd file
5. Push changes back to GitHub

7.0.1 Developer 2: Extending the Analysis

Step 1: Clone the Repository and Pull the Docker Image

```
git clone https://github.com/username/penguins-analysis.git
cd penguins-analysis
docker pull username/penguins-analysis:v1
```

Step 2: Run Docker Interactively

Developer 2 runs the container with the local repository mounted:

```
docker run --rm -it -v "$(pwd):/workspace" -w /workspace username/penguins-analysis:v1 /bin/
```

This approach: - Uses the renv-restored environment from the container - Allows Developer 2 to access and modify files directly from their local machine

Step 3: Extend the Analysis

Developer 2 modifies `peng1.Rmd` to add a second plot for body mass vs. bill length:

Listing 2 `peng1.Rmd (modified)`

````
---
title: "Palmer Penguins Analysis"
author: "Developer 2"
date: "`r Sys.Date()`"
output: html_document
---

```r
#| label: setup
#| include: false
library(ggplot2)
library(palmerpenguins)
```
````

8 Flipper Length vs. Bill Length

```r
#| label: flipper-bill-plot
ggplot(penguins, aes(x = flipper_length_mm, y = bill_length_mm)) +
  geom_point() +
  theme_minimal() +
  ggtitle("Flipper Length vs. Bill Length")
```

9 Body Mass vs. Bill Length

```r
#| label: mass-bill-plot
ggplot(penguins, aes(x = body_mass_g, y = bill_length_mm)) +
  geom_point() +
  theme_minimal() +
  ggtitle("Body Mass vs. Bill Length")
```

Step 4: Commit and Push Changes Back to GitHub

```bash
git add peng1.Rmd
git commit -m "Added second plot: Body Mass vs. Bill Length"
```

```
git push origin main
```

9.1 Key Benefits Demonstrated in This Example

This collaborative workflow demonstrates several advantages of the renv + Docker approach:

1. **Dependency consistency**: Both developers work with identical R package versions thanks to renv.
2. **Environment consistency**: The Docker container ensures the same R version and system libraries.
3. **Separation of concerns**: The R Markdown document remains outside the Docker image, allowing for easier collaboration.
4. **Workflow flexibility**: Developer 2 can work in the container while editing files locally.
5. **Full reproducibility**: The entire analysis environment is captured and shareable.

10 Best Practices and Considerations

10.1 When to Use This Approach

The renv + Docker approach is particularly valuable for:

- **Long-term research projects** where reproducibility over time is crucial
- **Collaborative analyses** with multiple contributors on different systems
- **Production analytical pipelines** that need to run consistently
- **Academic publications** where methods must be reproducible
- **Teaching and education** to ensure consistent student experiences

10.2 Tips for Efficient Implementation

1. **Keep Docker images minimal**: Include only what's necessary for reproducibility.
2. **Use specific version tags**: For both R packages and Docker base images, specify exact versions.
3. **Document system requirements**: Include notes on RAM and storage requirements.
4. **Leverage bind mounts**: Mount local directories to containers for easier development.
5. **Consider computational requirements**: Particularly for resource-intensive analyses.

10.3 Potential Challenges

Some challenges to be aware of:

- **Docker image size**: Images with many packages can become large
- **Learning curve**: Both Docker and renv require some initial learning
- **System-specific features**: Some analyses may rely on hardware features
- **Performance considerations**: Containers may have different performance characteristics

11 Conclusion

Achieving full reproducibility in R requires addressing both package dependencies and system-level consistency. By combining renv for R package management and Docker for environment containerization, data scientists and researchers can create truly portable and reproducible workflows.

This approach ensures that the common frustration of "it works on my machine" becomes a thing of the past. Instead, R Markdown projects become easy to share and fully reproducible. A collaborator or reviewer can launch the Docker container and get identical results, without worrying about package versions or system setup.

The case study presented demonstrates how two developers can effectively collaborate on an analysis while maintaining reproducibility throughout the project lifecycle. This strategy represents a best practice for long-term reproducibility in R, meeting the high standards required for professional data science and research documentation.

By adopting this two-tool approach, the R community can make significant strides toward the goal of fully reproducible research and analysis.

12 References

1. Thomas, R.G. "Docker and renv strategy."
2. "Palmer Penguins Analysis." Developer 2.
3. The Rocker Project. https://www.rocker-project.org/
4. renv documentation. https://rstudio.github.io/renv/