README Template for Reproducibility Packages

* This README template is designed to be included in reproducibility packages to provide directions for replicating the results in the research paper. The sections mentioned below should be included to guide users through the replication process. This template is meant to offer guidance, and authors should feel free to adapt and expand upon it.
* This is a simplified version of the social science data editors template, see [here](https://github.com/social-science-data-editors/template_README/blob/release-candidate/templates/README.md.)

## Contents

1. [Overview](#overview)
2. [Data Availability](#data-availability)
3. [Instructions for Replicators](#instructions-for-replicators)
4. [List of Exhibits](#list-of-exhibits)
5. [Requirements](#requirements)

The rest of the contents of this README file is highly desirable, but not strictly needed for reproducibility. The points above are needed.

1. [Code Description](#code-description)
2. [Folder Structure](#folder-structure)

## Overview

Begin by offering a concise overview for the replicator regarding the materials included in the package and provide a brief guide on how to proceed from start to finish. Ensure to include any crucial information that replicators should be aware of.

## Data Availability

This section will outline where and how the data supporting the findings of the study can be accessed and used. This is crucial, especially for replicating the results, as only the same data will be able to produce consistent results. Make sure to list all the datasets used and categorize them as follows:

* ☐ All data are publicly available.
* ☐ Some data cannot be made publicly available.
* ☐ No data can be made publicly available.

### Data Sources

Provide detailed information about the data sources, whether obtained from public repositories, institutional databases, or other sources. Include instructions on how others can access the data, including where it can be downloaded and the names under which it is cataloged. This is particularly important for replicators to ensure consistent results by using the same datasets. For example, if the package uses the World Bank’s World Development Indicators, ensure to add the URL, variable names, and file names exactly as they appear on the portal, and the year the data was accessed.

You can use the following as a template. Make sure to fill out this information for each of the data files used:

* **Filename 1:** Exact file name as shown on the source website
* **Source:** Name of the source website
* **URL:** Exact downloadable URL of the data used
* **Access year:** Date when the data was accessed. This is especially important as data can be updated, and replicators should know the exact time when the data was downloaded.
* **Variable names (optional):** This is useful when only data for certain variables is downloaded, such as when using the World Bank’s World Development Indicators data.
* **License (optional):** While this is not mandatory, it is great to know under which license the data is available to understand if it is public or private, or publication limitations.

### Statement about Rights

* ☐ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
* ☐ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permission are documented in the LICENSE.txt file.

## Instructions for Replicators

New users should follow these steps to run the package successfully: - Users must first have access to all data files if they are not included in the reproducibility package. They should go to the mentioned links, download the listed files, and place them in the data folder. - Update the following files with your directory paths

* main\_dofile.do
* Ensure all required software and dependencies are installed as listed in the [Requirements](#requirements) section.
* Run the main\_dofile.do file.

## List of Exhibits

Clearly identify and document the tables and figures as they appear in the manuscript by their corresponding numbers. If file names do not correspond to exhibit numbers, provide detailed explanations.

If not all data is provided in the reproducibility package, as described in the data section, then the list of tables should clearly indicate which tables, figures, and in-text numbers can be reproduced with the public material provided.

Example template for exhibit identification:

The provided code reproduces:

* ☐ All numbers provided in text in the paper
* ☐ All tables and figures in the paper
* ☐ Selected tables and figures in the paper, as explained and justified below

| Exhibit name | Output filename | Script | Note |
| --- | --- | --- | --- |
| Table 1 | Balancetable.xls | 02\_analysis.do (line 23) | Found in Outputs/tables/main |
| Figure 1 | Regresults.png | 02\_analysis.do (line 40) | Found in Outputs/figures/annex, Image Format: Portable Network Graphic (PNG), Bits Per Pixel: 32, Color: Truecolour with alpha, Dimensions: 970 x 544, Interlaced: Yes, XResolution: 144, YResolution: 144 |

## Requirements

### Computational Requirements

In this section, specify operating system requirements, software dependencies, environment setup instructions, and any other relevant information essential for replicating the results. Each of these factors plays an important role in ensuring successful replication.

### Software Requirements

List all software requirements, including versions, dependencies, libraries, environment setup, and packages installed. Using different versions of the same software could lead to variations in results. If multiple software are used, include details for all.

Example:

* **Stata version 15**
  + estout
  + rdrobust
* **Python 3.6.4**
  + pandas 0.24.2
  + numpy 1.16.4

### Memory and Runtime and Storage Requirements

Provide consistent information about memory resources for reliable computation. Include runtime information for replicators to assess processing times and detect potential issues with the code. It would be best to describe how much storage is required in addition to the space visible in the typical repository, for instance, because data will be unzipped, data downloaded, or temporary files written.

## Code Description

Give an overview of the program files and their purposes. Remove redundant or obsolete files from the replication archive. For example, main.do sets file paths, installs necessary ADO packages, and executes all other dofiles. Meanwhile, cleaning.do loads data, handles missing values, and analysis.do performs basic statistical analysis and generate visualizations.

Make sure to also include any crucial information that replicators should be aware of to facilitate a one-click run of the code.

## Folder Structure

Details about folder structure are crucial because a well-organized layout enables replicators to navigate quickly to the desired files or directories without searching through cluttered or disorganized folders. Include only the files necessary for replication and delete any unnecessary files.

An ideal folder structure for a reproducibility package should look something like this:

Data  
 ├── Raw  
 └── Cleaned  
Code  
 ├── Main\_dofile.do  
 ├── 01\_cleaning.do  
 └── 02\_analysis.do  
Outputs  
 ├── Main  
 │ ├── Tables  
 │ └── Figures  
 └── Annex  
 ├── Tables  
 └── Manuscript