**REPRODUCIBILITY AND RISK AUDIT**

**UTS | FTDI | MDSI | DSP – SPRING’19 - 13313491**

*“An article about computational results is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result.”*

**Project repo -** <https://github.com/Lhogeshwaran/DSP_AT3>

**OVERVIEW**

A code or an analysis that can be walkthrough by someone step by step attracts more trust. It makes the work reliable. Reproducibility doesn’t essentially correlate with correctness of the research; however, it helps gain trust of the reviewers and also makes it easy to identify any overseen defects. Publishing a reproducible work also facilitates others to easily adopt or extend the findings of the research to new work, enabling repeatability with minimal effort.

In the modern world, reproducibility of a digital work is not affected only by the code itself, but also by a number of other factors such as choice of programming language, operating system, version of software, packages and other dependencies. With increasing number of dependencies and problems for reproducibility, also are increasing number of innovative solutions such as evolving code style such as built-in unit tests, containerisation methods like dockers, version control systems, APIs, etc.

In this document I will be reviewing briefly about some of these techniques alongside a practical demonstration of how these techniques have helped improve the reproducibility and repeatability of a work that I carried out in the first semester for the subject Data Science for Innovation. The work can be reproduced by following the literate programming technique adopted in this link.

**DATA**

The analysis that has been re-done as part of this document was originally performed in Spring last year for Data Science and Innovation assignment ‘Quantified Self’. The assessment required a group of people working as a team to collect one set of data in a common repository on the cloud for use by every individual and another set of data to be collected individually be the members only for their individual use. The aim of the assessment was to perform analysis on the collected datasets, acquire insights from the dataset and produce a report. The code was not required to be shared at that time. The assessment was aimed at using the Data Science skills required in that subject to quantify about self.

At the time, I had used R to analyse the dataset, along with minimal use of visualization tools like Tableau, and Excel with plug-in to perform sentiment analysis. The data was sourced in two excel sheets was a mixture of quantitative and qualitative data.

**IDENTIFIED REPRODUCIBILITY ISSUES**

***The project space***

The project directory was a mess! Due to lack of proper organizing and version control, the project folder has files randomly named and spread across the directory, thus making it harder for even the author to reproduce or repeat the results after a period of time. This makes it harder for the code to be readily shared and limits reproducibility.

A screenshot of a cell phone

Description automatically generated

***Version control***

Lack of a version control like git, which prevents the changes that were made to the script over time or identify any problems or issues that occurred at the time of creating the scripts. Due to this, even when shared, the code could be subject to encountering issues which were originally solved by workarounds at that time.

A black and blue text

Description automatically generated

***Lack of documentation and comments***

The code has no reliable source of documentation or proper commenting, which makes it hard to interpret the choices of packages, codes, etc. Thus, when questioned or when in doubt about some of the choices made in the code, there could be no answers.

A screenshot of a cell phone

Description automatically generated

***Improper coding style***

No appropriate coding style was followed making it hard to interpret the code and thus debug. Also, the spaghetti code is confusing, reducing the interpretability even for the author of the code!

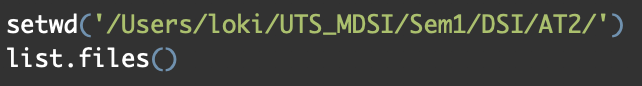
***Data confidentiality***

The data collected as a group had confidential details. Any information that could reveal the identity of an individual in the group has to be masked before sharing to audience.

**IDENTIFIED REPRODUCIBILITY RISKS**

***Use of absolute paths in the code***

Absolute paths were used in the code, which requires every user attempting to reproduce the results to alter these paths before they could verify the output of the analysis. Also, within the author’s computer environment, if the directory was to be renamed or moved, it would prevent the code from working in the future and require debugging.



***Lack of information of computational environment like package version, etc.***

There is no information available regarding the version of packages, OS environment the code was originally designed in, any path requirements etc. All these are having external dependencies that are subject to change or update overtime and could potentially cause the code to malfunction or stop functioning over a period of time.

***No mode of validating the data source for compatibility***

If the exercise was to be reproduced or repeated by another person, there is no method to validate the data source provided for the codebase. A mismatch in the column header, type of data, column order, etc. could stop the literate programming source from working as expected.

***Lack of installation instructions like shell/CLI commands***

There is no guide on how-to set up the system to run the code.

***Lack of code sharing framework like github/bitbucket/gitlab***

The code was never shared in the first place nor does it have a cloud repository, thereby putting the code at risk of being lost, if the author’s computer were to encounter any issues.

***Lack of APIs that facilitate repeatability***

Provided the analysis from the code was found to be insightful and someone wanted to repeat the exercise with a new dataset, there is no API they could call to which would ease this effort. Instead everyone who wished to repeat the exercise performed using this code has to undergo each step manually to reap the benefits of the exercise.

**IMPACT OF IDENTIFIED ISSUES AND RISKS**

***Issues***

* Project space does not facilitate easy sharing of code with others
* Lack of version control, unit tests, documentation and comments limits user from understanding the various issues encountered by the author at the time of development, and does not instruct on workarounds if the issues were to recur
* Coding style is confusing, thus making interpretation and debugging of the code difficult
* Data confidentiality restricts others from verifying the results of the analysis as performed by the author

***Risks***

* Absolute paths used in the code require alteration when executed on a new system
* Lack of computational environment information could lead to version/OS mismatch thus making the code unusable
* Lack of data source validation modules
* Lack of any set-up instructions makes it hard for others to have the project space setup for reproducing the results
* Lack of APIs to repeat the exercise makes it harder for the exercise to be replicated with new datasets

**REMEDIATION STRATEGIES**

Before diving into the remediation strategies that could help avoid the above identified risks and issues, let’s take some time to understand the need for Data Science projects to be reproducible and repeatable.

* Reproducibility is a key to gain trust of stakeholders. Often, people are satisfied only when they can verify an analysis by themselves.
* Often impressed by the one-off analysis work of a Data Scientist, there is a request for the analysis to be repeated on new datasets
* Data Science projects that begin as one-off could quickly turn into production grade requirements. Having a well-documented, structured and containerised code helps in getting the code from one-off to production-ready quickly.
* Data Science being hugely benefitting on knowledge-sharing platforms requires proper documentation to be able to share with fellow data scientists who may benefit from the work carried out and also help enhance the code overtime.
* Above all, a well modularized code allows re-usable modules/units to be quickly picked and plugged-in into new projects, thus making the life of Data Scientists so much easier!

**REMEDIENG IDENTIFIED REPRODUCIBILITY ISSUES**

The code for analysis was originally written in R. However, as R is not my primary programming language, for this assignment I have chosen python as the language of choice, primarily due to the fact that I am much more proficient in python and object-oriented programming style, in comparison to the functional programming style which R excels in. The choice of programming language in this case has no relation to the reproducibility.

***The project space, Version control***

Use of version control via git solves both of the issues identified here. Version control eliminates the need for having multiple versions of the codebase as separate files. All the changes can be made to a single file, referred to whenever necessary and also switched to if there is a requirement at any point of time. Version control also allows tracking of issues encountered, and workarounds added as features to the codebase over a timeline. Version control over cloud also has the advantage of facilitating multiple programmers to collaborate simultaneously over a single code.

Further, there are several GUIs that are available for download that helps visualize the changes to the code, compare them side-by-side easily.

Snapshot of project repo and version control in local -

A close up of text on a black background

Description automatically generated

***A close up of a screen

Description automatically generated***

***Lack of documentation and comments***

Literate programming in iPython notebook, an interactive notebook that let’s code, informative comments and interactive visualizations all live together simultaneously within a single file is used to capture the re-formatted code along with the comments on what the code does, how to use it and interpret the results.

<https://github.com/Lhogeshwaran/DSP_AT3/blob/master/13313491_AT3.ipynb>

A screenshot of a cell phone

Description automatically generated

***Improper coding style***

Object-oriented programming style is adopted to rewrite the code. OOP concepts come with its own set of benefits in programming and is the most commonly adopted programming style for python. Modularization of code will also be carried out to facilitate calling functions from both the iPython notebook and other python files that will work as APIs with command line interface.

A screenshot of a cell phone

Description automatically generated

***Data confidentiality***

The data with participant names masked will be made available. Also, instructions to create similar toy-dataset or self-recreate the dataset over a period of time will be provided.

**REMEDIENG IDENTIFIED REPRODUCIBILITY RISKS**

***Use of absolute paths in the code***

Replace all absolute paths with use relative paths and enhance with python packages os, sys. Also, CLI commands to navigate to the appropriate directories have been updated in the readme file of the repo.

***Lack of information of computational environment like package version, etc.***

Generate a requirement file by freezing the pip package requirements and making this available along with the code, and CLI installation instructions thereby facilitating any user to quickly set up their python environment with the same package versions as the author.



If possible, create a docker container to containerize the whole exercise.

***No mode of validating the data source for compatibility***

Add assertions to check if the data source used for reproducing the results of the analysis has the same structure as that of original data source.

A screenshot of a cell phone

Description automatically generated

***Lack of installation instructions like shell/CLI commands***

CLI instructions added in the github code repository as explained in the above step.

***Lack of code sharing framework like github/bitbucket/gitlab***

The codebase along with dataset, requirements file, and readme file for instructions made available in GitHub.

<https://github.com/Lhogeshwaran/DSP_AT3>

***Lack of APIs that facilitate repeatability***

Modularized python file that can be called from terminal will be added to the repo and can be used to quickly generate the outputs that are required to repeat the analysis with a new or same dataset.

By calling the ‘analysis\_rep\_api.py’ along with arguments of the name of the excel file and sheet, the plots from the analysis and the csv file with the sentiment scores can be generated for new data.

A close up of a logo

Description automatically generated

**CONTAINERS**

***What are containers and how do they operate?***

A medium to wrap up a software along with its dependencies into small units with all requirements that it needs to function is a container. The concept of containers was a development of virtualization. It is based on the principle of programming, where specific a specific software code with its dependencies is encapsulated and packaged so that it can run consistently and independently of any infrastructure.

Containers are not a new concept in the world of programming. They have been around for years. However, they gained popularity at all levels of programming when Docker made the containerisation easily available for everyone through their product. Soon everyone in the world of programming started harvesting the effectiveness of containers in a rapidly evolving world.

Some of the most common advantages of containers are –

* Independent host system in relation to application code
* Easy versioning and dependencies management
* Scalability, light weight, flexible
* Several containers can be executed in parallel easily and independent of each other

Containers contain only the libraries, frameworks, application and other dependencies they need. Several containers can be hosted in parallel in a single host system. The host operating system operates as the only operating system on the server while all running containers communicate directly with the host. This process structure helps the overhead to be extremely low.

***How containers can be used to create and share data science environments***

Majority of the Data Scientists are not software developers or programmers. The hassle of having to configure the computational environment to suit a specific data science project and later having to re-do the whole thing for another project is a common issue that most Data Scientists go through. Overcoming this hassle, once the data scientist has developed his data science product, when they send it over to another is when they notice there is a mismatch in the version of their python or R or one of the dependencies! Containers are potential solution to this problem!

A picture containing person

Description automatically generated

With a little practise and effort, a Data Science product can be containerized for good. All of the system environment can then be transferred to a software developer who is better fit to take the decision required for productionising the work.

Containerization ultimately reduces the hassle needed to re-create a specific virtual environment each time. The daemon image retains a snapshot of the environment as it was at the time the product was deemed to be working as expected and thus makes it easier for sharing work amongst the data science community.

***Containerization of this project***

While the complexity of the project I chose to rework was minimal and without much dependencies on versioning (as most of the packages are commonly used and updated) the product could still be containerized in a few ways. Using the docker images of python or jupyter-notebook would have been viable attempts as it could have reduced the effort required by anyone trying to replicate this exercise by having to follow the github readme file and follow every instruction line by line.

If this project was containerized by building on the image of the [jupyter/datascience-notebook](https://hub.docker.com/r/jupyter/datascience-notebook/), the effort required to reproduce the results of this exercise would have been as simple as pulling the docker image from the docker hub and directly port forwarding to the dockers iPython kernel. Also, this would have reduced the risk of the project every going outdated due to any dependency.

Another advantage would be that whenever, the project space in the local is tweaked causing the project to break, again the effort required to get the project working would be to re-download the docker image. This is also possible in git, but, if the computational environment of the user-system changes for some reason, docker container would be the better choice.

**CONCLUSION AND REFLECTIONS**

This document highlighted multiple reproducibility issues and risks that were originally present in my assignment from first semester. It also provided the opportunity to revisit all the skills I have picked up over the course and re-apply them to an old project. As a takeaway from this assignment, following are some of the areas I have to build up on in – object-oriented programming, statistical analysis, R programming, updating self with the latest improvements in the world of data science.