Report

Thomas Nguyen

07/11/2021

Introduction

Background

From 8 September to 2 November, 2021, I worked on the **Monash University's Data Science: Data-Driven Decision Making** microcredential on FutureLearn, under the supervision of Rowan Peter from the host organisation (Monash Centre for Professional Development and Monash Online Education, informally referred to as the Centre).

Offered by the Monash School of Business and led by Professor Dianne Cook, the Data Science: Data-Driven Decision Making is an online postgraduate microcredential program comprising three courses at Level 8 of the Australian Qualifications Framework. The microcredential guide learners through practical programming exercises in R language to learn the process of tidying, harvesting and wrangling data and applying statistical models to simulate complex functions that solve a broad range of problems.

Motivation

By the end of November, 2021, the course will finish its fourth run. Overall, it went smoothly and was generally well-received by the learners. However, While the course is masterfully crafted, most of it was written 2 years ago. There is a need to update some of the materials that were outdated or no longer relevant. Furthermore, there are gaps in the existing resource that can be further enhanced to deliver smoother learning experience for the learners, especially those without statistics or coding background.

Based on those reasons, I review the materials and objects that have already been created, identify areas for enhancement and gaps, and then create the assets to fill those gaps, all within the scope of the microcredential.

The assets I created include written types (code chunks and worked examples) and media types (Tutorial with screen recording)

Objectives

The enhancements and assets that were

1. Creating well-documented and well-explained assets that support new learners with no background in statistics and R-coding.
2. Delivering cutting edges techniques that learners can apply in solving problems.
3. Providing detailed examples and code that learners can duplicate in their work.
4. Supplying functional practices that can improve learner's productivity.

Methodology

The assets created in the internship are based on the tidyverse approach. All of the new material utilised already-in-use datasets, creating smooth transition between old and new sessions. As each asset focus on different part of the course, further information is provided in corresponding section.

Asset 1: Functions (Tutorial)

In the first course, Wrangling and Workflow, of the micro-credential, the material teaches the learners to use functions to solve statistical problems. However, I believe this approach is not suitable. Many learners don't have a background in statistics or math, so they are not interested in using function in this fashion. Furthermore, various pre-written packages also supply a wide range of functions for this purpose.

Another way of using function, which is as a tool to avoid repetition, should be taught. The learners can apply the knowledge immediately to improve their productivity. Therefore, I created a new section to teach the learners about function as a tool to automate common tasks.

Dataset: Tb_long dataset from previous part of the course. Objective achieved: 1, 2, 3 and 4.

Key contributions: 1. Reasons to apply function as a tool to automate common tasks.
2. How to write function, and examples. 3. Introduction to curly curly to combine wrangling verbs and function.

Asset 2: Operators, Syntax & Shortcuts (Written)

Packages used: Base R and Knitr Objective 1, 3 and 4

Asset 3: Git and GitHub functionality (Tutorial)

Packages used: Base R Objective 3 ### Asset 4: Sample assignment (Written)

Tidyverse approach Objective 1, 3 and 4

Asset 5: KNN imputation and survivorship bias (Written)

In Course 2, Modelling and Visualisation, learners are introduced to missing values. I identified two gaps that need improvement in this course:

Firstly, the introduction is neither informative nor engaging, negatively affecting learner's experience.

Secondly, many learners struggled with neighbour imputation implementation on Step 1.16. "Dealing with missing data". The "bioconductor" package used in the section is outdated and hard to install. Furthermore, the material of this part is not documented, making it hard to understand.

Dataset: oceanbuoys from package naniar Objective 1, 2, 3 and 4

Key contributions:

1. Highlight the problems of missing values in real life, with survivorship bias and the Challenger disaster as examples.
2. Replace old knn_imputation method with more advanced tidymodels approach.
3. Provide better documented code.

Asset 6: Wordcloud (Written)

Dataset: Text data from "On the Origin of Species" (Charles Darwin) from package gutenbergr
Objective 1, 3 and 4

Asset 7: