DaLiCo Data Literacy MOOC / Valencia

Author:
J. Alberto Conejero –
Universitat Politècnica
de València

Co-funded by the
Erasmus+ Programme
of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

DALICO
Data Literacy in Context

Valencia MOOC

PROJECT INFORMATION

PROJECT TITLE:
Data Literacy in Context

PROJECT ACRONYM:
DaLiCo

PROJECT NUMBER:
2019-1-DE01-KA203-005066

ERASMUS+ PROGRAMME:
KA2 Partnerships for Innovation and the Exchange of Good Practices

WEBSITE:
www.dalico.info

PARTNERS:

HAW HAMBURG

HOGESCHOOL UTRECHT UNIVERSITY OF APPLIED SCIENCES

UNIVERSITY of DEBRECEN

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Co-funded by the
Erasmus+ Programme
of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Valencia MOOC

AUTHORS

J. Alberto Conejero

CONTRIBUTORS

(all: Universitat Politècnica de València – UPV)

Andrea Conchado Peiró

José Miguel Carot Sierra

José Luis Poza Luján

ABSTRACT

Valencia MOOC is a contribution from UPV to IO 2 and IO 3 within the DaLiCo project.

DATE: 31.10.2022

This document was written by partners of the "DaLiCo" project and is published as open educational resource. The project was funded through the European Commission. The content of this document exclusively reflects the opinions of the authors. The EU Commission and the National Agency at the DAAD assume no liability arising from the use of the information presented here.

LICENCE

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Contents

DALICO
Data Literacy in Context

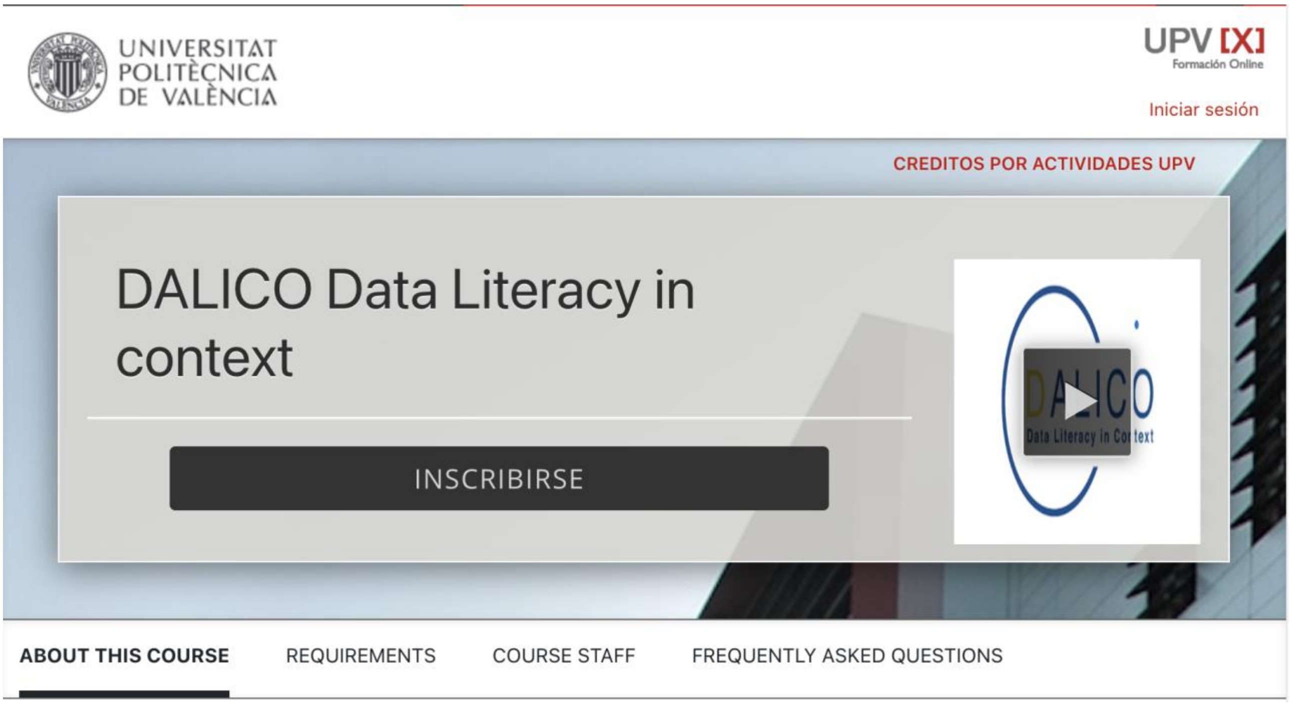
Co-funded by the
Erasmus+ Programme
of the European Union

Introduction

As a part of IO2 and in connection to IO3, we complement the Train-the-Trainer part with a MOOC oriented to the practical aspects of data literacy, showing a vision from the perspective of data science. The course has been uploaded to UPVx, the edX-based platform of MOOCS of Universitat Politècnica de València

https://upvx.es/courses/course-v1:poc+dalico201x+2022-01/about

Structure and Use of the MOOC



DALICO Data Literacy in context

UNIVERSITAT POLITÈCNICA DE VALÈNCIA
UPV [X] Formación Online
Iniciar sesión
CREDITOS POR ACTIVIDADES UPV
INSCRIBIRSE
DALICO Data Literacy in Context
ABOUT THIS COURSE
REQUIREMENTS
COURSE STAFF
FREQUENTLY ASKED QUESTIONS

Fig. 1: Screenshot of the MOOC landing page

Once approved by edX, it will be uploaded to its platform https://www.edx.org/

The MOOC contains videos, explanations, and exercises. The content is modular, and some videos have also been used within IO3. The list of videos is accessible from https://www.youtube.com/playlist?list=PL6kQim6ljTJv-2U_oYE1PwedJIQU7ZB5M and from https://www.youtube.com/playlist?list=PL6kQim6ljTJvcogG-0bWW7yM7aS8nyvKg

With this MOOC, we introduce the fundamental aspects of data literacy practically. It can be helpful for trainers and people interested in starting to work on data science projects. This MOOC has included feedback mainly from the summer schools within the DaLiCo project https://dalico.info/summer-schools/ As a matter of fact, one can also have an overview of the developed projects within the schools at https://projects.dalico.info/#/home

We present the contents using mainly Dataiku https://www.dataiku.com/ , a very accessible software with many statistical and basic machine learning tools already included and easy to use. It was successfully tested as a tool to start on Data Literacy in the Utrecht summer school. Besides, for small groups of up to 5 people working in the same project, the software is completely free.

Additionally, we also will use Datawrapper https://www.datawrapper.de/ a very easy to use software for data visualisation. This was already tested in the Hamburg summer school. For starting, it has a free license.

The MOOC is divided in the following parts:

- Firstly, we present an **introduction to data literacy from the perspective of data science**. We emphasize the importance of posing relevant questions either for our organization (in a broad sense). In order to answer them, we start by introducing some aspects of the **work of a data scientist** that a beginner can adopt in data literacy. Basically, her task consists of translating the organization's problems into questions linked with data and developing an analysis that can be understood by any other people making decisions in the organization. These analyses must be understood, interpreted, and contextualized within the organization.
- Secondly, we explain the process of **data collection**. We start indicating how to pose the right questions. We give some hints and characteristics. To illustrate some of them, we indicate that a research question has to be relevant, recent, open ended, answerable, and based on theory and prior practice.
- Then, we will have to look for different **data sources**. The **CRAAP test** will help us with the choice. It considers the following aspects: Currency, relevance, authority, accuracy, and purpose .If some data does not exist or it is not available, we will have to collect it. So as to, we introduce some fundamentals on **survey design** and **design of experiments** since both notions cover the spectrum of qualitative and quantitative data. For survey design, we introduce questionnaires and measurement scales, how to formulate the survey questions, and how to design of our sample, distinguishing between a probabilistic and a non-probabilistic approach. We show the basic principles of the design of experiments that are replication, blocking, and randomization. Also, we indicate some strategies in order to reduce the number of experiments, such as best-guess experiments, one factor at a time (OFAT), and full and fractional design.
- We explain how to prepare the data before starting our analysis. This is the most consuming part of the process. In **data preparation**, we start with **data cleansing**, checking the content's integrity, consistency, and density. Then, we perform the **data integration** while addressing the problem that heterogeneous data may present as different levels of details, abstraction, and synchronization times. After this, we should check the **data quality**. We will just focus on the presence of outliers that can be detected by looking at histograms and box plots and the existence of missing data. The decision to include these anomalous data is our choice, and it will depend mainly on what we know about the context. We also explain how errors can be introduced in the measurements, either by a bias or by randomness. After this, we introduce Dataiku, how to install it, and how to give the first steps in performing the data preparation and data quality over a

sample case that is based on information related to how countries are progressing to attain Sustainable Development Goals.

- Next, with fundamental notions of statistics, we show how to **analyse one-dimensional quantitative and qualitative variables**. We recall basic notions as mean, mode, maximum, minimum, quartiles, quantiles, and standard deviation, but we do not rephrase them. We provide some data sets Then. We show how to **compare two variables**. The tools to be used depending on the type of variables. We overview several visual tools such as mosaic plots, cross tables, contingency tables, stratified box plots, correlation matrices, and scatter plots.
- We present the options that Dataiku presents to perform hypothesis tests to evaluate if different samples correspond to the same population. We will rely on the t-student test for quantitative variables, and for qualitative variables, we will refer to the chi-square test. We know that these tests are pretty advance, so we will just mentioned this and we will mainly focused on the p-value and if this is significant enough, below 0.05.
- We will see how to use some of the predictive models that are incorporated in Dataiku. We will not only be interested in the predictions themselves, but also in the influence of different factors in the prediction. Dataiku will test the potential models that can be applied to all the existing variables, and then it will show us the accuracy of each model. It will also indicate what is the effect (positive or negative) of each variable in the predictions and how much each variable impacts the predictions. As basic models, we will explore the case of regression models or decision trees.
- Finally, we include a part of data visualisation using Datawrapper. It is very simple to use. One has just to upload the data, choose the desired chart, edit it, and that is all. We also can embed or download it.

Co-funded by the Erasmus+ Programme of the European Union