

PROGRAMMING FOR DATA MANAGEMENT & ANALYSIS I

**Bachelor in Data and Business Analytics BDBA SEP-2023
IPDS-DBA.1.M.A**

Area Others

Number of sessions: 25

Academic year: 23-24

Degree course: FIRST

Number of credits: 3.0

Semester: 1º

Category: COMPULSORY

Language: English

Professor: **PABLO MONFORT VINUESA**

E-mail: pmonfort@faculty.ie.edu

PABLO MONFORT VINUESA

Pablo is a Senior Data Scientist, who has worked for McKinsey for 3 years after obtaining his PhD in Mathematics. In 2017 he founded his own company, Canalytics, focused on Advance Analytics and Machine Learning.

He is a specialist in mathematical modeling, clustering, machine learning and data mining. He has been supporting several major client engagements by developing predictive models and data analysis. Pablo worked at the National Institute of Statistics (INE) as well.

Examples of his work include:

- Implementation of 4G telephony for a company leading Asian telecommunications.
- Clustering client base card of a major European bank.
- Prediction model for churned clients in Latin America telephone company.
- Sales analysis of a prominent Spanish textile company by segmenting its costume base and launch a product recommendation engine.

pmonfort@faculty.ie.edu

SUBJECT DESCRIPTION

Did you know the current most relevant companies in terms of market cap belong to tech sector? Apple, Microsoft, Alphabet, Amazon...

But not only these ones. All the companies in all the sectors are creating big analytics departments to take data-driven decisions. Programming skills to work with huge datasets, solve problems and look for insights based on data are required in most of current positions.

In this subject we will learn the basic on programming for data science. We will learn how a dataset is built, how to load, clean and transform data to obtain insights and to take decisions based on them.

There are some different programming languages to work with data such as R, Python, C++...

In this subject we will use R. According to the latest worldwide Kaggle survey, over 45% of data scientists use R on a regular basis. This language was designed as a statistical platform for data cleaning, analysis and representation. Some libraries such as dplyr for numerical analysis or ggplot2 for visualization have hundreds of pre-built useful commands to do this task a little bit easier. In fact thousands of questions are daily asked in websites for data scientists such as stackoverflow.

In summary, this course will cover the basics concepts and tools that will allow us to start to work using programming languages such as R. These concepts (organizing data, working with null values, loading datasets, defining functions, aggregating variables, exporting results) will be continuously used in other courses.

LEARNING OBJECTIVES

The overall objective of this course is to learn the basics of programming, namely acquiring data (csv, websites, excel, clipboard, etc.), defining functions and conditional commands, understanding the structure of a script, exporting results and insights, learning how a dataset can be built, etc.

Secondary goals include:

- Introduction to problem-solving using data analysis;
- Understanding programming languages and their importance in solving real situations based on data;
- Learning the basic of programming with R;
- Writing fully functional console or R programs.

TEACHING METHODOLOGY

Learning Activity	Weighting	Estimated time a student should dedicate to prepare for and participate in
Lectures	13.33 %	10.0 hours
Discussions	0.0 %	0.0 hours
Exercises in class, Asynchronous sessions, Field Work	20.0 %	15.0 hours
Group work	26.67 %	20.0 hours
Individual studying	40.0 %	30.0 hours
TOTAL	100.0 %	75.0 hours

PROGRAM

This subject is built around 4 main modules:

- M1: Introduction & coding basics
- M2: Loops & functions
- M3: Datasets

- M4: Data analysis

DESCRIPTION

SESSION 1 (LIVE IN-PERSON)

Course introduction and overview of the computer programming environment

Module: 1

Pre-work: read chapter 1.4.1, 1.4.2 and 1.5.

Topic: We will review the course logistics and organization together with the computer programming environment (IDEs, R, RStudio, etc...)

Post-work: install R and RStudio on your laptops

Book Chapters: Chapter 1.4.1, 1.4.2 and 1.5. (Book) (See Bibliography)

SESSION 2 (LIVE IN-PERSON)

What data science is?

Module: 1

Topic: We will talk about the importance and value of data, the phases in a data-based project and some of the most important current programming languages.

Work: read the slides and watch this link: <https://www.youtube.com/watch?v=X3paOmcTjQ>

Post-work: look for real examples where the value of data is shown and send, in groups, an email explaining the selected example.

Multimedia Material: Data Science In 5 Minutes | Data Science For Beginners | What Is Data Science? (YouTube)

SESSION 3 (LIVE IN-PERSON)

Data types & variables

Module: 1

Pre-work: read chapter 4.1 & 4.2

Topic: How to store information in R using variables. Different data-types for variables in R.

Post-work: Additional exercises in the slides will be proposed to be done by students after this session.

Book Chapters: Chapter 4.1 & 4.2

SESSION 4 (LIVE IN-PERSON)

Data types & variables

Module: 1

Pre-work: read chapter 4.1 & 4.2

Topic: How to store information in R using variables. Different data-types for variables in R.

Post-work: Additional exercises in the slides will be proposed to be done by students after this session.

Book Chapters: Chapter 4.1 & 4.2

SESSION 5 (LIVE IN-PERSON)

Data types and variables exercises

Module: 1

Work: Lot of exercises about variables string, numeric and boolean will be solved

Post-work: Read professor solutions documents in Campus

SESSION 6 (LIVE IN-PERSON)

Flow control statements, logic, scope and constants

Module: 2

Pre-work: read chapter 19.4 & 21.2.a

Topic: We will expand the functionality we can code with iterations, logic and flow control with if/else structures

Post-work: Additional exercises in the slides will be proposed to be done by students after this session.

Book Chapters: Chapter 19.4 & 21.2. (Book) (See Bibliography)

SESSION 7 (LIVE IN-PERSON)

Flow control statements, logic, scope and constants exercises

Module: 2

Topic: We will expand the functionality we can code with iterations, logic and flow control with if/else structures

Work: Read the slides and solve the exercises given.

Post-work: Read professors solution documents in Campus.

SESSION 8 (LIVE IN-PERSON)

Loops

Module: 2

Topic: We will explain the for and while loops, the logical of these 2 estructures and real examples

Book Chapters: Chapters 5.2 to 5.7 (Book) (See Bibliography)

SESSION 9 (LIVE IN-PERSON)

Loops

Module: 2

Topic: We will explain the for and while loops, the logical of these 2 estructures and real examples

SESSION 10 (LIVE IN-PERSON)

For and while loops exercises

Module: 2

Pre-work: read the theory in the last sessions

Topic: We will solve a lot of real exercises where loops are required.

Post-work: Check and read the solutions given by the professor.

Book Chapters: Chapter 7.1, 7.2, 7.3 & 7.5. (Book) (See Bibliography)

SESSION 11 (LIVE IN-PERSON)

For and while loops exercises

Module: 2

Pre-work: read the theory in the last sessions

Topic: We will solve a lot of real exercises where loops are required.

Post-work: Check and read the solutions given by the professor.

SESSION 12 (LIVE IN-PERSON)

Preparing the midterm

Module: 2

Topic: We will focus on reviewing with exercises all the aspects explained until now

Pre-work: Read the slides and review all the concepts explained in previous sessions.

Wok: A lot of exercises will be solved in groups. The best groups will receive additional points in the midterm

Post-work: Read professors solution documents in Campus.

SESSION 13 (LIVE IN-PERSON)

MidtermType:

We will solve some challenges in groups using R

SESSION 14 (LIVE IN-PERSON)

Functions, arguments and return types

Module: 2

Pre-work: read chapter 19.1, 19.2 & 19.3.

Topic: We will start encapsulating our code in methods that enable us to reuse the code preventing duplication and errors in large projects. The importance of returning an object. The difference between local and global variables

Work: Read the slides and solve the exercises given.

Post-work: Read professors solution documents in Campus.

Book Chapters: Chapter 19.1, 19.2 & 19.3. (Book) (See Bibliography)

SESSION 15 (LIVE IN-PERSON)

Functions, arguments and return types

Module: 2

Pre-work: read chapter 19.1, 19.2 & 19.3.

Topic: We will start encapsulating our code in methods that enable us to reuse the code preventing duplication and errors in large projects. The importance of returning an object. The difference between local and global variables

Work: Read the slides and solve the exercises given.

Post-work: Read professors solution documents in Campus.

SESSION 16 (LIVE IN-PERSON)

Functions exercises

Module: 2

Topic: We will solve multiple exercises using functions. We will define local and global variables.

Work: Read the slides and solve the exercises given.

Post-work: Read professors solution documents in Campus.

SESSION 17 (LIVE IN-PERSON)

Functions exercises

Module: 2

Topic: We will solve multiple exercises using functions. We will define local and global variables.

Work: Read the slides and solve the exercises given.

Post-work: Read professors solution documents in Campus.

SESSION 18 (LIVE IN-PERSON)

Arrays

Module: 3

Topic: We will learn how to store multiple items in vectors and data Frames in R. Also, we will learn what an array is and the most important functions and operators to be used with them.

Post-work: Solve some optional exercises

SESSION 19 (LIVE IN-PERSON)

Arrays

Module: 3

Topic: We will learn how to store multiple items in vectors and data Frames in R. Also, we will learn what an array is and the most important functions and operators to be used with them.

Post-work: Solve some optional exercises

SESSION 20 (LIVE IN-PERSON)

Arrays exercises

Module: 3

Topic: We will learn how to use arrays and the most important operations with them

Work: Review the commands about arrays explained in the last sessions

Post-work: Read professors solution documents in Campus.

SESSION 21 (LIVE IN-PERSON)

Reading and writing data

Module: 4

Pre-work: Review all the contents explained.

Topic: We will explain how to import and export datasets to be analyzed. The difference between different formats (csv, excel, json files....) will be explained

SESSION 22 (LIVE IN-PERSON)

Data Frames exercises

Module: 4

Pre-work: Review what a csv file is.

Topic: We will solve a lot of exercises about importing data. A few very basic commands about dataframes will be explained (head, tail, summary...)

SESSION 23 (LIVE IN-PERSON)

Recap exercises

Pre-work: Review all the contents explained.

Topic: We will solve some exercises in class. The best students solving the exercises will receive additional points in the final exam

SESSION 24 (LIVE IN-PERSON)

Preparing the final exam

Pre-work: Review all the contents explained.

Topic: We will solve different mock exams to be ready for the final exam

SESSION 25 (LIVE IN-PERSON)

Final exam

Type: Synchronous (F2F)

EVALUATION CRITERIA

A variety of teaching and learning strategies will be used in this course. You will be assigned a grade based on your demonstrated knowledge on in-class exercises, a midterm and final exam, and your participation in various class activities and discussions.

The scores for these different activities are as follows:

criteria	percentage	Learning Objectives	Comments
Final Exam	50 %		
Midterm in groups	25 %		
Workgroups	15 %		
Class Participation	10 %		

RE-SIT / RE-TAKE POLICY

Class Participation (10%)

Active participation in class activities, discussions, and labs is an especially important aspect in this course because our focus will be on understanding how the concepts discussed in class can be applied in real-world contexts.

Midterm Exam (25%)

The midterm exam will include material from the PowerPoint slides. This exam will be solved in groups and it will include several questions to be solved using R for a dataset that will be provided that day.

Class exercises (15%): Obligatory and Optional

There are two types of exercises that the student need to solve: Obligatory and they will form part of your evaluation criteria and optional (specified under each session). Students who do optional exercises will receive an extra 10 points as part of their continuous evaluation criteria.

Obligatory exercises will be given to be solved in groups using the concepts given in class. These exercises will be solved in groups at home after explaining these concepts and will always be sent to the professor before next session. No additional time will be allowed in any case. This exercises will be graded individually.

Final Exam (50%)

The final exam will include all the material explained in the course. The exam format will include multiple questions to be solved numerically and from a graphical point of view using R. The final exam will be held during the last session.

Marks:

- A minimum passing grade in the final exam (3.5) is required to pass the subject. If a student scores lower than this minimum, he will have to go to June retake, irrespective of their overall course grade. The overall passing course grade is 5.0.
- All the exercises/assignments/exams will be submitted via Turnitin on Campus Online. No other option will be accepted.

As per University Policy:

Each student has 4 chances to pass any given course distributed in two consecutive academic years (regular period and July period).

It is mandatory to attend 100% of the classes. Students who do not comply with at least 70% attendance will lose their 1st and 2nd chance, and go directly to the 3rd one (they will need to enroll again in this course the next academic year).

Grading for retakes will be subject to the following rules:

1. Those students who failed the subject in the first regular period will have to do a retake in July (except those not complying with attendance rules who are banned from this possibility).
2. Dates and location of the July retakes will be posted in advance and will not be changed. Please take this into consideration when planning your summer.
3. The maximum grade that a student may obtain in the 2nd exam session is 8 out of 10. Those students in the 3rd call will be required to attend 50% of the classes. If due to schedule overlap, a different option will be discussed with the professor in order to pass the subject.

BIBLIOGRAPHY

Compulsory

- Hadley Wickham & Garrett Grolemund. (2017). *R for Data Science*. O'Reilly. ISBN 9781491910399 (Digital)

BEHAVIOR RULES

Please, check the University's Code of Conduct [here](#). The Program Director may provide further indications.

ATTENDANCE POLICY

Please, check the University's Attendance Policy [here](#). The Program Director may provide further indications.

ETHICAL POLICY

Please, check the University's Ethics Code [here](#). The Program Director may provide further indications.

