1. OVERVIEW

Subject area	Exploratory Data Analysis
Degree	Bachelor's Degree in Data Science
School/Faculty	Faculty of Science, Engineering and Design
Year	Second
ECTS	4.5 ECTS
Type	Compulsory
Language(s)	Spanish
Delivery Mode	Campus-based/Online
Semester	Second semester

2. INTRODUCTION

This is a compulsory subject within the syllabus for the Degree in Data Science at the Universidad Europea de Valencia. The aim of this subject is to teach the descriptive and inferential statistics methods applied to the first stages of a problem in the field of data science.

Through theory and practice, students will learn the techniques involved in the acquisition and preprocessing of data extracted from different sources and in different formats. They will also have an introduction to the data processing techniques particularly associated with the organisation and structure of data.

We will also learn how to display results in graphs and to explore the distributions and relationships between variables. Finally we will cover imputation methods for missing data, detection and erasure of possible outliers, feature selection and basic linear statistical models associated with regression and classification models.

3. SKILLS AND LEARNING OUTCOMES

Basic skills (CB, by the acronym in Spanish):

- CB1: Students have shown their knowledge and understanding of a study area originating from general secondary school education, and are usually at the level where, with the support of more advanced textbooks, they may also demonstrate awareness of the latest developments in their field of study.
- CB3: Students must have the ability to gather and interpret relevant data (usually within their study area) to form opinions which include reflecting on relevant social, scientific or ethical matters.
- CB4: Students can communicate information, ideas, problems and solutions to both specialist and non-specialist audiences.

Cross-curricular skills (CT, by the acronym in Spanish):

- CT01: Ethical values: ability to think and act in line with universal principles based on the value of a person, contributing to their development and involving commitment to certain social values.
- CT02: Independent learning: skills for choosing strategies to search, analyse, evaluate and manage information from different sources, as well as to independently learn and put into practice what has been learnt.
- CT05. Analysis and problem-solving: be able to critically assess information, break down complex situations, identify patterns and consider different alternatives, approaches and perspectives in order to find the best solutions and effective negotiations.

Specific skills (CE, by the acronym in Spanish):

- CE2. Ability to apply mathematical, statistical and optimisation techniques and models applied to data processing, decision support systems, the search for relationships between variables and making predictions.

Learning outcomes (RA, by the acronym in Spanish):

After passing the course the student will be able to:

- RA1: Use the language of mathematics and statistics to solve a problem.
- RA2: Search for, choose and process suitable data for the analysis process.
- RA3: Use programming language and IT software packages to apply statistical and optimisation techniques and models applied to data processing, decision support systems, the search for relationships between variables and making predictions.
- RA4: Generate reports to present the statistical analysis results including ethical criteria.

The following table shows how the skills developed in the subject area match up with the intended learning outcomes:

Skills	Learning outcomes
CB1, CB3 CT01, CT05 CE2	RA1
CB3 CT02, CT05 CE2	RA2
CB1, CB3, CB4 CT01, CT02, CT05 CE2	RA3 - RA4

4. CONTENTS

5. TEACHING/LEARNING METHODS

The types of teaching/learning methods are as follows:
UA 1. Introduction to Exploratory Data Analysis (EDA)

- What is EDA?
- Data acquisition and preprocessing techniques
- Sampling
- Types of statistics charts
- EDA for categorical and numerical variables

UA2. Descriptive statistics analysis

- Introduction to ggplot2
- Data distribution
- Relationships between variables
- Dealing with missing data
- Imputation techniques
- Detecting outliers.

UA 3. Data transformation

- Data capture
- Accessing data through APIs
- Data cleaning
- Extract, transform, load (ETL)

UA 4. Inferential statistics analysis

- Regression models
- Classification models

6. LEARNING ACTIVITIES

The types of learning activities, plus the amount of time spent on each activity, are as follows:

On campus:

Learning activity	Number of hours
Master lectures	30
Problem solving and case studies	14
Laboratory work	12
Formative assessment	6
Tutorials	6
Independent working	74
TOTAL	**142**

Online:

Learning activity	Number of hours
Problem solving and case studies	8
TOTAL	**8**

7. ASSESSMENT

The assessment methods, plus their weighting in the final grade for the subject area, are as follows:

On campus:

Assessment system	Weighting
On campus tests to assess theory/practical learning.	40%
On campus laboratory practice tests.	30%
Off-site tests to assess theory/practical learning.	20%

Online:

Assessment system	Weighting
Performance observation in digital work	10%

On the Virtual Campus, when you open the subject area, you can see all the details of your assessment activities and the deadlines and assessment procedures for each activity.

8. BIBLIOGRAPHY

The reference publication to accompany this subject area is:

- John W. Tukey (1970). Exploratory Data Analysis.

The recommended bibliography is indicated below:

- Pearson, Ronald K. (2018) Exploratory Data Analysis Using R.
- Pierson, L. (2015). Data Science for Dummies
- O'Neil C, Schutt R. (2013) Doing Data Science: Straight Talk from the Frontline, ed OReilly.
- Kannan R., Blum A., Hopcroft J. (2013). Foundations of Data Science.
- Gareth James (2013). An Introduction to Statistical Learning: With Applications in R.
- Wickham H., Grolemund G. (2016). R for Data Science.