

Empowering the Future with Data Insights

Group Project
Programming for Data Science 2024/25



I. INTRODUCTION

In today's world, data has become an increasingly vital source of knowledge, driving innovation and shaping the future. We've explored various ways to harness the power of programming to extract valuable insights from data. This project offers a practical opportunity to apply the concepts and techniques learned throughout the course, deepening your understanding through a hands-on experience.

II. PROJECT GOALS

Working in data science goes beyond technical expertise; it demands strong critical thinking and problem-solving skills. You have the option to choose one of two projects for this course:

Project 1: Web Scraping & Data Analysis for the NOVA IMS Teachers

Project 2: Workforce Dynamics - Insights for Policy and Economic Planning

Each end-of-course group project is divided into three main components, each with its own set of objectives:

Data Wrangling and Analysis: Clean and preprocess your dataset, then perform the necessary data analysis to answer questions and extract meaningful insights.

Working with Advanced Topics: Apply one or more of the advanced topics covered in this course. This could involve working with web scrapping, data integration or any other advanced data type relevant to your project.

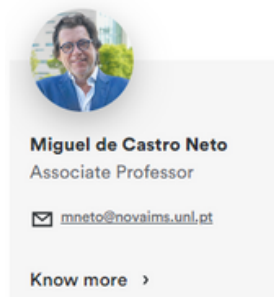
Data Science in Action: Integrate the pre-processed datasets and derive relevant insights. You have the flexibility to choose from a variety of provided techniques or use any other method you find suitable to explore the data and uncover patterns and relationships. This section is more open-ended, allowing you to choose the level of complexity you wish to pursue.

III. PROJECT 1: NOVA IMS Teachers

After being fired, a senior data engineer deleted his crucial web scraping pipeline in a fit of frustration, leaving NOV IMS with a critical problem: no functioning pipeline and no documentation. Only the dataset remains, and we need you to reconstruct the pipeline to retrieve the same data.

This project involves several objectives:

1. Web Scraping: Develop an algorithm to scrape data from the teaching staff page at NOVA IMS, accessible via the following link: [NOVA IMS Teaching Staff](#). The goal is to replicate the structure of the table on the next page, including information from each teacher's page accessed through the 'know more >' link.



2. Data Wrangling and Analysis: Use the initial dataset to begin data wrangling and perform analysis early in the semester.

3. Data Science in Action: Engage in more advanced data science activities.

Since web scraping will be covered later in the course, you will be provided with the dataset initially to start working on objectives 2 and 3. The scraping task can be completed towards the end of the semester. The dataset will be updated weekly to allow you to compare your results. The final evaluation will be based on the dataset obtained from the scraping algorithm on a specific date, which will be announced during the semester.

III. 1. PROJECT OBJECTIVES

Webscraping (Advanced Topic): You have to recreate the dataset 'nova_ims_teaching_staff.csv' given to you by leveraging webscraping techniques.

Data Wrangling and Analysis: You are given a set of questions about the data that require you to do some data wrangling and create visualizations (when possible and senseful) to answer them:

1. Treat missing values and duplicates, justify your approach
2. a) Which teachers have the highest wordcount* (top 1, name only)?
b) highest coursecount* (top 5) ?
c) most publications (top 10) ?
3. a) How are the wordcount, coursecount and publications distributed and related?
b) Are there differences in those variables for different types of teachers
4. a) How many different courses (unique course names) are taught at NOVA IMS?
b) How many courses (unique course names) are taught by only one teacher?
c) What's the probability of someone teaching 'Data Mining I' also teaching 'Data Mining II'?
5. Gain 3 additional insights of your choice


*wordcount (nr. of words in biography), we see a new word as a sequence of characters separated by a space

*coursecount (nr of courses taught by a teacher - if a similar course name appears twice, both of them are counted), we see a course as a sequence of characters separated by a comma + space it might be worth considering creating new columns for those variables

Data Science in Action: Apply Data Science knowledge to your data. You are free in your choice, but the level of complexity will be recognized for the grade. Here are some examples:

- “Build a Regression/Decision Tree to predict the number of publications or the type of teacher based on the other variables, visualize the feature importance and evaluate predictions”
- “Look for patterns in the biographies, are there words that are more common in the biographies of teachers with more publications or courses/different types of teachers? Maybe create a word cloud?”
- “Mine association rules to find out which courses are more likely to be taught together? Which rules have the highest confidence, lift and support?”

III. 2. PROJECT DATA



Name

Title

Contacts

Miguel de Castro Neto

Associate Professor
PhD in Agronomy Engineering (ISA/UTL)

✉ mneto@novaims.unl.pt

Course Units

- Business Intelligence
- Business Intelligence
- Business Intelligence I
- Business Intelligence II
- Business Intelligence in Tourism
- Data Governance
- Smart and Sustainable Cities

Courses

Biography

Miguel de Castro Neto is Dean of the NOVA Information Management School (NOVA IMS) at the Nova University of Lisbon. He is an Associate Professor at NOVA IMS, where he founded and runs NOVA Cidade - Urban Analytics Lab, dedicated to smart cities and urban analytics. He also founded the NOVA Business Intelligence & Analytics Lab in partnership with BI4All and Microsoft and was the co-founder of the NOVA Data-Driven Public Policy Lab. He is the scientific director of the Smart Campus Living Lab at the Nova University of Lisbon, an Advisory Member of the Order of Engineers, Chairman of the Board of Directors of Lisboa e-Nova, Lisbon Energy Agency and a member of the Platform for Sustainable Growth and member of the Executive Committee of the European Digital Innovation Hub AI4PA - Artificial Intelligence for Public Administration. He was Secretary of State for Spatial Planning and Nature Conservation in the XIX and XX Constitutional Governments of the Portuguese Republic. During this period, he was

Scientific Publications

Hassam, S., Alpalhão, N., & Neto, M. D. C. (2024)
A Spatiotemporal Comparative Analysis of Docked and Dockless Shared Micromobility Services. Smart Cities, 7(2), 880-912. <https://doi.org/10.3390/smartcities7020037>

Publications (The Screenshot shows just 2, it's 116 (might be more when you check if he publishes))

Moura, R., Pessanha Santos, N., Vala, A., Mendes, L., Simões, P., de Castro Neto, M., & Lobo, V. (2024)
Fisheries Inspection in Portuguese Waters from 2015 to 2023. Scientific Data, 11(1). <https://doi.org/10.1038/s41597-024-03088-4>

Name	Title	Courses	Publications	Biography
String	String	String (each course seperated by ', ')	Integer (number of Publications)	String
Miguel de Castro Neto	Associate Professor	Business Intelligence, Business Intelligence, Business Intelligence I, ...	116	Miguel the Castro Neto is...

Structure of 'nova_ims_teaching_staff.csv'

IV. PROJECT 2: Workforce Dynamics

A specialized branch of the European Union (EU) has tasked your team with conducting an in-depth analysis of the IT sector in Germany. Your objective is to extract valuable insights from the data provided by workers in the industry and derive meaningful tools and information to assess the current state of the workforce that can better shape government policy and economic planning.

The insights generated by this pilot project will be crucial in shaping strategies to support the IT workforce in Germany and potentially serve as a blueprint for similar analyses across other sectors and countries.

IV. 1. PROJECT OBJECTIVES

The goals for this project are three-fold:

Data Wrangling and Analysis: Your first task is to import, comprehend and preprocess the dataset. This will enable you to create compelling visualizations and identify key aspects of the IT sector in Germany. Moreover, you should be able to answer the following questions:

1. How did you deal with missing values and duplicates? Justify your approach.
2. a) For how many days were answers registered?
b) What was the day with the most answers? How many were registered?
3. a) What is the most used technology, from the main used technologies?
b) What are the top 3 most used techs, from the often used ones?
4. a) How are 'Age' and "Yearly brutto salary" distributed and related?
b) Are there differences in those variables for the 10 most common positions?
c) Where is the employee with the most years of experience in Germany from?
d) What is the probability that someone from Düsseldorf is a Software Engineer?
5. What are the other 3 insights you believe are meaningful for the company?

Data Integration (Advanced Topic): This step involves integrating data from other sources with the IT dataset you've been working on. You should formulate relevant questions and use data to answer them. You are expected to address the following question, as well as develop **two additional questions** of your own. You may utilize the datasets provided in class to support your analysis.

- **Is there a relation between the salaries of employees and the population of the cities they work in?** (datasets available online).

Data Science in Action: Integrate the pre-processed datasets and derive relevant insights. You should look into **creating something you believe would be important for the specialised branch of the EU**. You have the flexibility to choose from various provided techniques or use any other method you find suitable to explore the data and uncover patterns and relationships. Here are suggestions that could be (but don't have to be) a starting point:

- Build a Regression/Decision Tree to predict the yearly salary of an employee based on the other variables, visualize the feature importance and evaluate predictions"
- Look for patterns in your data, are there technologies that are more common for employees that get paid more or are in higher positions? Is there a pattern related to employees that were fired during covid?

Reminder: You are working on a study about only one country.

IV. 2. PROJECT DATA

The answers to the survey given to workers are stored in "**data_IT.xlsx**". This file contains two separate sheets:

"Section-General" contains answers to general questions about the employee and the IT field:

Variable	Description
<i>Timestamp</i>	Date and hour survey was answered
<i>Age</i>	Age of the employee
<i>Gender</i>	Gender of the employee
<i>City</i>	City where employee works from

<i>Position</i>	Position of the employee in the company
<i>Total years of experience</i>	Total years of experience in the IT field
<i>Years of experience in Germany</i>	Years of experience working in Germany
<i>Seniority level</i>	Seniority level in the company
<i>Your main technology / programming language</i>	Answers given by employee to question
<i>Other technologies/programming languages you use often</i>	Answers given by employee to question
<i>Yearly brutto salary (without bonus and stocks) in EUR</i>	Yearly brutto salary of employee (in Euros)
<i>Yearly bonus + stocks in EUR</i>	Yearly bonus + stocks received
<i>Annual brutto salary (without bonus and stocks) one year ago. Only answer if staying in the same country</i>	Yearly brutto salary of employee one year before answering the survey (in Euros)
<i>Annual bonus+stocks one year ago. Only answer if staying in same country</i>	Yearly bonus + stocks received one year ago.
<i>Number of vacation days</i>	Number of vacation day per year
<i>Employment status</i>	Current employment status
<i>Contract duration</i>	Duration of contract in current company
<i>Main language at work</i>	Language used at work
<i>Company size</i>	Number of employee company has (in range)
<i>Company type</i>	Type of company employee works at

“Section-COVID19” contains answers to questions about the COVID-19 pandemic:

- “Have you lost your job due to the coronavirus outbreak?”
- “Have you been forced to have a shorter working week (Kurzarbeit)? If yes, how many hours per week”
- “Have you received additional monetary support from your employer due to Work From Home? If yes, how much in 2020 in EUR”

V. DELIVERABLES

- A .ipynb notebook (or zip of multiple notebooks) featuring all the code you used throughout the project to:
 - a. Decide on your final solutions for the problem at hand.
 - b. Obtain your final results (code that helped you make decisions but does not directly contribute to reaching the final solution should be included but commented).
 - c. Any additional datasets required to run the notebook should be included in a zip file. (File naming format: **GroupXX_PDS_2425.ipynb**).
- A presentation (meant for 5 minutes) focusing on advanced topics and data science applications: Which insights did you gain? Why did you do what you did? What difficulties did you have? How did you solve them? ... (File naming format: **GroupXX_PDS_2425_Presentation.pdf**).

VI. EVALUATION

Criteria	Description	Max Grade (out of 20)
Data Wrangling and Analytics	Initial preprocessing, answers to given questions and extra insights found.	6v
Advanced topic	Mastery of one or more of the “advanced topics” covered in the course.	6v
Data Science in Action	How pertinent the approach and strategy were. The level of complexity will be recognized in your grade.	3v
Notebook Quality	Well structured, commented, efficient...	2v
Presentation and Defense	Appearance, timing, originality, argumentation and transmission of knowledge	3v

Students can submit all deliverables with a maximum delay of 3 days, incurring a penalty of 1v per day. Beyond these three days, submissions will not be accepted.

VIII. FINAL NOTES

1. Make sure to have a notebook that is understandable to someone reading it for the first time. A good structure, with appropriate comments showing that you understand what is being done at every block of code, goes a long way into that.
2. The trustworthiness of the information you provide is key. If you look for information outside the materials we provided, you should cite the source of the materials appropriately.
3. Before submitting, run your notebook from the start one last time, please comment unnecessarily lengthy cells that take too much time to run.
4. All the code you used that is unneeded to highlight the points you want to convey should be part of your submitted notebook(s), but it should be commented.
5. We will run your Jupyter Notebooks. So, please make sure we can run the notebook from start to finish in one go. Notebooks that do not fulfil this condition will be penalized.
6. **The notebook code will pass through a process of plagiarism and AI generation checking.**
7. **To avoid situations where we have conflicting versions, please make sure that you show, in the notebook, the version of the package you are using for each package you use.**
8. When determining the grade for your work, there will be a comparative component between your work and the works presented by your peers.
9. **Attendance at the presentation is mandatory for approval of the project. The presentation and discussion have a group component and an individual component. Considering this, a student's final grade can change during the defense depending on their performance, without any limitations.**