# Project

| | | | |
|---|---|---|---|
| Site: | Eduvos LMS | Printed by: | Daniël De Waal |
| Course: | Data Mining and Data Administration Assessments | Date: | Monday, 10 March 2025, 6:15 PM |
| Book: | Project | | |

## Table of contents

## 1. Project

| | |
|---|---|
| **Faculty:** | Information Technology |
| **Module Code:** | ITDAA4-12 |
| **Module Name:** | Data Mining and Data Administration |
| **Content Writer:** | Taryn Michael |
| **Internal Moderation:** | Community of Practice |
| **Copy Editor:** | Mr. Kyle Keens |
| **Total Marks:** | 100 |
| **Submission Week:** | Week 6 |

This module is presented on NQF level 8

5% will be deducted from the student's project mark for each calendar day the project is submitted late, up to a maximum of three calendar days. The penalty will be based on the official campus submission date.

Projects submitted later than three calendar days after the deadline or not submitted will get 0%. [1]

This is an individual project.

**This project contributes 40% towards the final mark.**

[1] Under no circumstances will projects be accepted for marking after the projects of other students have been marked and returned to the students.

## 2. AI Checklist and Declaration

Before you submit an assignment, you should be able to confidently and honestly make all the below statements. For group work, you can also review the list, together, to hold one another accountable.

- I confirm that my submission reflects my personal learning, knowledge, skills, and understanding.
- If AI tools were employed for generating any part of this assignment (even in the drafting/research phase), I have referenced the use of AI in the text and/or declared the use of AI. I am willing to discuss the process and its contribution to my learning.
- I am aware that the lecturer may request a demonstration of my learning, such as explaining choices in approach, research, and the content I am submitting.
- I am aware that, if I did use AI in any phase of preparing this submitted work, it is recommended that I save a copy of the relevant chat history (prompts and answers), as this will help me demonstrate my writing/work process to my lecturer, if I am asked to do so.
- I have read the assignment instructions on whether AI tools are prohibited for this assignment, and if they are prohibited, I can confirm that I did not use AI tools.
- I understand that failure to agree to these terms may be deemed unethical, potentially leading to disciplinary action. I understand my responsibility for the integrity of my work, including seeking clarification from academic staff and adhering to instructions.

It is essential to acknowledge your use of ChatGPT or other generative AI in your learning. If you use ChatGPT or other generative AI to help you generate ideas or plan your process, you should still acknowledge how you used the tool, even if you don't include any AI-generated content in the assignment.

**Please note:** The following guiding questions that you will be asked in an AI declaration questionnaire below this assignment brief.

### AI Declaration

**It is compulsory to complete this AI declaration for each of your assignment submissions.**

| |
|---|
| I carefully read the assignment instructions, and the extent to which AI may be used for the assignment. |
| I used the following AI system(s)/tool(s): |
| I used it for the following: |
| If I quoted or paraphrased an AI output, I have referenced the relevant tool, version, and the date I used the tool. |

| |
|---|
| I still consider this work my own. (i.e., I have not outsourced the final product, or significant portions of it, to AI tools/systems). |
| If required, I can defend my argument/perspective, explain my choices and approach, and can show that I am knowledgeable about the details of my work. |

For further guidance on the use of AI at Eduvos, please refer to the AI FAQ glossary. You will locate the FAQs in the Artificial Intelligence tile on the myDocuments page of myLMS.

## 3. Instructions to Students

1. Please ensure that your answer file (where applicable) is named as follows before submission: **Module Code – Assessment Type – Campus Name – Student Number.**
2. **Use Python for this project**
3. Remember to keep a copy of all submitted projects.
4. All work must be typed.
5. Please note that you will be evaluated on your writing skills in all your projects.
6. All work must be submitted through Turnitin. The full originality report will be automatically generated and available for the lecturer to assess. Negative marking will be applied if you are found guilty of plagiarism, poor writing skills, or if you have applied incorrect or insufficient referencing. (See the "instructions to students" book activity before this activity where the application of negative marking is explained.)
7. You are not allowed to offer your work for sale or to purchase the work of other students. This includes the use of professional project writers and websites, such as Essay Box. You are also not allowed to make use of artificial intelligence tools, such as ChatGPT, to create content and submit it as your own work. If this should happen, Eduvos reserves the right not to accept future submissions from you.

## 4. Section A

### Section A

**Learning Objective**

For learners to apply Python programming skills to solve a problem and provide recommendations to a company

**Project Topic**

GitHub Survey Data Analysis

**Scope**

First Block Week 1 to Second Block Week 7

**Marking Criteria**

Marking is done at the discretion of the marker. Plagiarism is solely forbidden. Any project found to contain plagiarism will be awarded an immediate mark of 0.

## 4.1. Question 1

### Question 1 20 Marks

Study the scenario and complete the question(s) that follow:

**Scenario: Analyzing GitHub Survey (2022-2024)**

You have been employed by GitHub to analyze three years of data (2022, 2023, and 2024) they have collected to gain insights into their user base and developer preferences. Each year's data is stored as tables in a **database file** that will be provided to you.

The tables each include 80+ columns, but you will only work with the following columns:

- **YearsCode**: Number of years the respondent has been coding.
- **MainBranch**: The main role or area of focus of the respondent
- **Country**: The respondent's country of residence.
- **EdLevel**: The highest level of formal education attained.
- **LanguageHaveWorkedWith**: Programming languages the respondent has used professionally.
- **LanguageWantToWorkWith**: Programming languages the respondent aspires to use in the future.
- **DatabaseHaveWorkedWith**: Databases the respondent has used professionally.
- **DatabaseWantToWorkWith**: Databases the respondent aspires to use in the future.
- **Age**: The respondent's age.

**Task**

You are required to analyze the data using **Python** to clean, process, and generate a report with actionable insights.

You may download the dataset from the “Project dataset” folder. The database file name is “github.db”

Preprocessing the data

1.1 Use Python to perform the appropriate data cleaning and preprocessing steps needed to get your data into a more usable format. This will include the following steps:

· Combine the data from all three years into a single dataset by selecting only the relevant columns and adding an indicator column to specify the year of each survey.

(4 marks)

· Standardize the Country column, ensuring consistency in naming for countries with variations.

(4 marks)

· Handle missing data and subsetting the dataset to include a manageable selection of countries to facilitate clear visualizations and analysis. Justify your choices for missing value treatment and country selection.

(6 marks)

- Categorize the values in the MainBranch column into meaningful groups to provide a clear understanding of respondent roles.

(3 marks)

- Reclassify the EdLevel column into broader educational categories to create concise, interpretable groupings.

(3 marks)

End of Question 1

## 4.2. Question 2

## Question 2 50 Marks

2.1 Use visualizations and descriptive statistics to explore the dataset and answer the following questions. Provide clear insights to support your analysis.

**1. Demographics and Roles**

1. How has the number of users in each **MainBranch** category changed over three years?

(5 Marks)

2. What are the trends in users' **Age**, **Country**, **EdLevel**, and **YearsCode** across different MainBranch categories?

(12 Marks)

**2. Technology Trends and Preferences**

1. What are the top 5 most popular **databases** and **programming languages** that GitHub users currently use and want to use in the future?

(12 Marks)

2. How have the usage trends of the top 5 databases and programming languages changed over three years?

(6 Marks)

**3. Relationship Analysis**

1. How does **YearsCode** correlate with the use of the top 5 databases and programming languages currently in use?

(5 Marks)

Focus on clarity, relevance, and insightful analysis in your visualizations and interpretations.

2.2. Based on your above analysis, provide a short report detailing your findings / insights derived from the data that may be useful to GitHub to better understand their users. Additionally, provide recommendations on improvements or ideas to maintain their current users and attract new ones. Ensure that your findings are relevant based on your analysis and visualizations, and that they are articulated well. Your recommendations should be innovative, detailed and actionable.

(10 marks)

End of Question 2

# 4.3. Question 3

## Question 3

**30 Marks**

Study the scenario and complete the question(s) that follow:

**Sentiment Analysis: ChatGPT user reviews**

GitHub is in the process of developing an AI chatbot to assist developers with coding, troubleshooting, and project management tasks. To ensure the new chatbot meets user expectations, the development team is analyzing user reviews of the ChatGPT Android App. The goal is to understand what users like and dislike about ChatGPT and leverage these insights to build a more effective and user-friendly chatbot tailored to developers' needs.

You are tasked with performing a text analytics and sentiment analysis of the dataset containing daily-updated user reviews of the ChatGPT app. Your analysis will help GitHub identify key strengths and weaknesses in ChatGPT's design and functionality, along with trends in user preferences and feedback.

The dataset is named “chatgpt_reviews.csv” and can be found in the “Project Datasets” folder on myLMS

3.1. Preprocess your data as necessary and perform sentiment analysis on the content column to categorize user reviews as positive, neutral, or negative. Save the sentiments for each review in a new column called “Sentiment”

(10 marks)

3.2. Using visualizations, highlight features or issues that are frequently mentioned in both positive and negative reviews. Additionally, identify which aspects users seem to love about ChatGPT vs which features they dislike.

(5 marks)

3.3. Critically discuss how user satisfaction scores have changed over time. Motivate your answer using a line plot.

(5 marks)

3.4. Based on user feedback and sentiment analysis performed, summarize the strengths of ChatGPT that GitHub should emulate. Additionally, highlight weaknesses or pain points that GitHub should address in its chatbot.

(10 marks)

End of Question 3