# *Advanced Topics in Computer Science II (420-G50-HR)*

# *Lab 10 – Data Analysis and Visualization*

Date due: **March 27, 2025**

**Learning Objectives**

Upon successful completion of this lab exercise, the student will be able to:

* Perform data analysis using Pandas
* Create visual plots from data using Matplotlib

Lab Set-Up

1. Create a folder called *initials*G50L10 where *initials* are your initials.
2. Put the **pldb.csv** file inside your folder

Problem Statement

We want to analyze the programming language dataset comprising of over 4000 records to answer the following questions:

* What is the oldest programming language?
* How many programming languages appeared in the year 2022? What are they?
* What are the top 10 most popular programming languages by number of users and jobs?
* What are the top 10 least popular programming languages by number of users and jobs?
* How many programming languages existed before the year 2000?
* How many programming languages were created in the year 2000 and afterwards?
* What are the various programming language types based on the dataset?
* Which programming languages have the highest number of books written about them?
* Is there a relationship between any of the columns in the dataset? If so, can we visualize it?

To Do

**Part A – Analysing and Visualizing**

Use the steps below to complete the task. Use [Markdown](https://www.markdownguide.org/getting-started/) to document each step of the process.

1. In *initials*G50L10 folder, create an interactive python notebook called **plang.ipynb**.
2. Copy the pldb.csv file into the same directory as plang.ipynb
3. After importing pandas, read in the CSV file using the appropriate function into a variable (give it any name that makes sense)
4. Determine how many rows and columns are in the dataset
5. Given the huge number of columns in this dataset, you can use [set\_option()](https://pandas.pydata.org/docs/reference/api/pandas.set_option.html) to see all columns without it being truncated. To see this in action, add this line right after the import pandas statement - **pandas.set\_option('display.max\_columns', None)**
6. Display the descriptive statistics for the dataframe
7. Determine the different programming language types. (**Hint**: call unique() on the column of interest)
8. From the previous step, determine how many programming languages belong to the different ‘**type’** categories. (**Hint**: consider using value\_counts())
9. Determine the oldest programming language (**Note**: Although the dataset is about programming languages, not all the records are programming languages (i.e., ‘pl’). Refer to steps 7 and 8 above)
10. Determine how many programming languages appeared in 2022 alone.
11. Display in a table all the languages returned in the previous step. Limit the columns to title, appeared and type
12. Determine the top 10 programming languages by:
    1. **numberOfUsers** – (sort in descending order) – extract title and numberOfUsers
    2. **numberOfJobs** – (sort in descending order) – extract title and numberOfJobs
13. Plot a bar graph of the top 10 programming languages by number of users.
14. Plot a bar graph of the top 10 programming languages by number of jobs.
15. Determine the top 10 least popular programming languages by **numberOfUsers** and **numberOfJobs**
16. Determine how many programming languages existed before the year 2000
17. Determine how many programming languages appeared from the year 2000 and onwards
18. Determine the top 10 programming languages with a high number of published books.
19. Create a pie chart from the previous step. Remember to give it a suitable title
20. To determine if there is any strong relationship between the columns of the dataset, you can use corr(). The syntax is: **dataset.corr()**. You can also directly compare two columns in the dataset like so: dataset[‘column1’].corr(dataset[‘column2’])
    1. Start by checking if there is a strong relationship between the bookCount and tiobe.currentRank columns
    2. Next, check if there is a strong relationship between the bookCount and numberOfUsers columns
    3. Now, use [.corr()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html) on the entire dataset and see what columns have the strongest relationship to each other. (**Hint**: the closer the values are to one the better. Also, remember to set the **numeric\_only=True**)
    4. Once you determine the columns with the strongest relationships, plot a scatter diagram to see the result

**To submit**

When you have completed the lab exercise, call the Teacher’s attention and we’ll go over it together. Then, create a single zip file called *initials*G50L10.zip and copy the file to the Moodle page for the course.