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

# *Lab 9 – Using Pandas*

Date due: **March 21, 2025**

**Learning Objectives**

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

* Performing data analysis using Pandas

Lab Set-Up

1. Create a folder called *initials*G50L09 where *initials* are your initials.
2. Data source - https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS

Problem Statement

We want to analyze the youth global unemployment dataset for the year range of 2017 to 2021 to answer the following questions:

* What countries in 2021 had youth unemployment greater than or equal to 25%?
* Which countries had youth unemployment of less than or equal to 1% in 2021?
* Top 10 countries that had an average high rate of youth unemployment from 2017 to 2021
* Top 10 countries that had an average low rate of youth unemployment from 2017 to 2021
* Top 10 countries that had a high COVID impact (for years 2019 and 2020)
* Top 10 countries that had less COVID impact (for years 2019 and 2020)

To Do

**Data Analysis using Pandas**

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

1. In the G50L09\_files folder you downloaded from Moodle, create an interactive python notebook called **main.ipynb**.
2. Ensure the **global\_unemployment.csv** file is in the same directory as main.ipynb
3. After importing pandas, read in the CSV file using the appropriate function into a variable called df
4. Preview the df variable (**Note**: Do not use print()). You will observe that the data is a bit untidy
5. Use [drop()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.drop.html) to remove rows (0, 1 and 2) for a start. Preview df to ensure the change took place.
6. Now make row number 3 the header. There are [many ways](https://sparkbyexamples.com/pandas/pandas-convert-row-to-column-header-in-dataframe/#:~:text=columns()%20to%20Convert%20Row,by%20extracting%20the%20first%20row.) to do this but remember that since you deleted the first three rows, row number 3 is now the first row in the data frame
7. Delete the first row now since you have made it the header. Use .drop(labels=3, axis=0, inplace=True). **Note**: Setting the **inplace** parameter to True simply means you do not want to make a copy of the dataframe you are modifying
8. Next, check to see how many rows and columns your dataframe has.
9. Check to see what the various **columns** in the dataframe are.
10. Create a new dataframe called **five\_year\_df** that consists of only columns: “Country Name”, “Country Code”, 2017.0, 2018.0, 2019.0, 2020.0, 2021.0 and preview it.
11. Check for the first 7 records and the last 7 records respectively.
12. Check for null values in the dataframe by chaining isnull() and sum() to the dataframe. Observe that column 2021 has 33 null values unlike the other columns that have either 0 or 31.
13. Fill in the null records of the dataframe using [ffill()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.ffill.html#pandas.DataFrame.ffill). Copy the results to a new data frame called **new\_five\_year\_df**
14. Do a quick check of the first seven records in the new dataframe (i.e., **new\_five\_year\_df**). You should observe that the row for Aruba is completely null so we can drop that record using dropna() like so: df = df.dropna()where df is the name of your dataframe.
15. Confirm that there are no more records with null values by repeating step 12.
16. Check if there are duplicates in the new dataframe
17. Check the dtypes of the columns in the dataframe
18. Sort the dataframe by country name.
19. Set the index of the dataframe to be the “Country Code” column (**Hint**: use [set\_index()](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.set_index.html))
20. Add a new column to the dataframe called ‘**5yrs\_avg**’. This column should display the average of data from columns 2017 through to 2021. (Hint: df[‘5yrs\_avg’] = df[[selected\_columns]].mean(axis=1))
21. Add another column called ‘**covid\_yr\_diff**’. This column should subtract data in the 2019 column from data in the 2020 column of the dataframe (i.e. **2020\_data** – **2019\_data**)
22. List all countries with unemployment greater than or equal to 25% in 2021
23. List all the countries where unemployment is less than or equal to 1% in 2021
24. List the top 10 countries that had an average high rate of unemployment from 2017 to 2021
25. List the top 10 countries that had an average low rate of unemployment from 2017 to 2021
26. List the top 10 countries with a high COVID impact on employment rate
27. List the top 10 countries that had a low COVID impact on employment rate

**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*G50L09.zip and copy the file to the Moodle page for the course.