EE3801 Data Engineering

Laboratory Exercise (LAB-III)

Assignment release date: **Sept 13, 2023**

Date submission due: **Sept 21, 2023**

Grading: **Your ASSIGNMENT will be graded out of 100 marks and the final weight of this assignment is 15%. The guidelines explained in the template will also carry marks. So please adhere to the guidelines.**

Note: From Lab 3 and for all labs in PART II of this module, you will be using Python 3.9, as it may use some special features.

Concepts used: Data frames, data wrangling - extraction, handling missing data, transforming a DF to a target DF, command line utilities, AWS EC2

Consider the food production dataset (***FAOSTAT_Lab3_Original.csv***) given to you.

Nomenclature: Dataframe – DF

Attempt all the following.

Important: Set up according to briefing video/slides before attempting this lab. All screenshots should be 'full-screen', with date/time shown, unless otherwise stated.

1. Login to your AWS account. Select "ap-southeast-1" as region. Submit a **full-screen** screenshot of the AWS EC2 Management Console page from the web browser. [5 Marks]

2. Write a python script (lab3.py) (clearly, logically named) that defines and calls a function that does the following:

   a) Takes in 3 arguments as an input. [6 Marks]

      1. File name (eg. 'FAOSTAT_Lab3_Original.csv')
      2. Item (eg. 'Apples')
      3. Element (eg. 'Production')

   b) Read the csv file into a dataframe using Pandas [4 Marks]

      Hint: if you encounter issues reading the file, consider: pd.read_csv(filename, encoding= 'unicode_escape')

   c) Generates a new dataframe that filter for a specific 'Item' and 'Element' combination. [5 Marks]

   d) List the countries that has nan as the values. [5 Marks]

   e) For the areas that have the actual values for the 'Element', find the Year that has the maximum values of the 'Item', the respective Values, and the median values throughout the years. The output dataframe should have the structure as follows: [25 Marks]

Area	Year	Maximum Values	Median
Afghanistan	2019	250324	89403
Albania	2018	…	
…	…	…	

   f) Save this dataframe as csv, with file name '**output_<Item>_<Element>.csv**'. [5 Marks]

   g) You should be able to run it from command line to produce the output csv file, see below for example:

```
$ python lab3.py "FAOSTAT_Lab3_Original.csv" "Apples"
"Production"
```

'output_Apples_Production.csv' should be created and should contain a list of countries starting with Afghanistan and ending with Zimbabwe. You can check the output by using this command:

```
$ cat 'output_Apples_Production.csv'
```

Hint: using pandas and sys libraries will be enough for this task.

3. Start a **Linux-based t2.micro EC2** instance using the AWS CLI, submit a screenshot of the command and the corresponding output. (Screenshot of the terminal window) [5 Marks]

4. Using AWS CLI, filter for your newly created instance's following properties:
   a) Public DNS name (ends with '.amazonaws.com').
   b) Launch Time

   Submit a screenshot of the command and the corresponding output. [12 Marks]

   Hint: modify command(s) used in briefing video.

5. Copy your script and FAOSTAT_Lab3_Original.csv to the EC2 instance. Submit a screenshot of the command and the corresponding output. [5 Marks]

6. On the EC2 instance, run the following lines to set up a conda environment with python 3 and pandas library installed (accept and enter yes to any prompts):

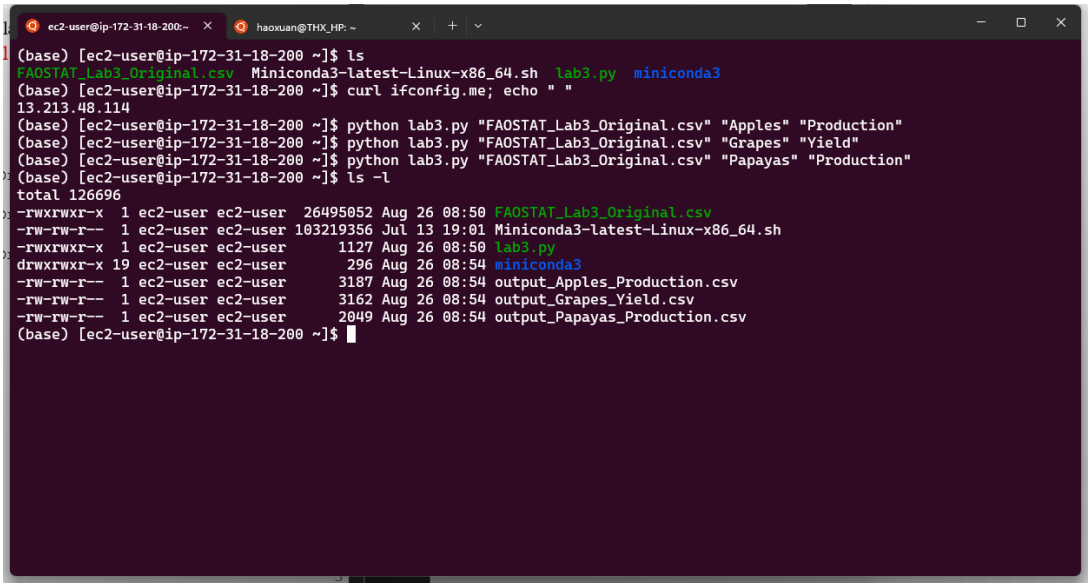
```
$ wget https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.sh

$ bash Miniconda3-latest-Linux-x86_64.sh

$ source ~/.bashrc

$ conda install pandas
```

7. Run these commands back-to-back, replacing only the script name, and take a screenshot of all the outputs together: [15 Marks]

```
$ ls
```

```
$ curl ifconfig.me; echo " "

$ python lab3.py "FAOSTAT_Lab3_Original.csv" "Apples" "Production"

$ python lab3.py "FAOSTAT_Lab3_Original.csv" "Grapes" "Yield"

$ python lab3.py "FAOSTAT_Lab3_Original.csv" "Papayas" "Production"

$ ls -l
```

Your screenshot should look like this:



8. Transfer the output csv files back to your local machine. Submit a screenshot of the command and the corresponding output. [5 Marks]

9. Submit a zip file containing:

   a) your python script: EE3801_Lab3_<your_name>.py

   b) all the output csv files

   c) Lab3 Submission Template with the following file name convention: EE3801_Lab3_<your_name>.pdf

   The zip file should be named : EE3801_Lab3_ <your_name>.zip

10. **Remember to terminate your instances!** Submit a screenshot of the command and the corresponding output (showing no instance is running). [3 Marks]