SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

ICT233

End-of-Course Assessment - January Semester 2024

Data Programming

INSTRUCTIONS TO STUDENTS:

1. This End-of-Course Assessment paper comprises **6** pages (including the cover page).

2. You are to include the following particulars in your submission: Course Code, Title of the ECA, SUSS PI No., Your Name, and Submission Date.

3. Late submission will be subjected to the marks deduction scheme. Please refer to the Student Handbook for details.

IMPORTANT NOTE

ECA Submission Deadline: Sunday, 07 April 2024 12:00 pm

ICT233 Copyright © 2024 Singapore University of Social Sciences (SUSS)

ECA Submission Guidelines

Please follow the submission instructions stated below:

This ECA carries 70% of the course marks and is a compulsory component. It is to be done individually and not collaboratively with other students.

Submission

You are to submit the ECA assignment in exactly the same manner as your tutor-marked assignments (TMA), i.e. using Canvas. Submission in any other manner like hardcopy or any other means will not be accepted.

Electronic transmission is not immediate. It is possible that the network traffic may be particularly heavy on the cut-off date and connections to the system cannot be guaranteed. Hence, you are advised to submit your assignment the day before the cut-off date in order to make sure that the submission is accepted and in good time.
Once you have submitted your ECA assignment, the status is displayed on the computer screen. You will only receive a successful assignment submission message if you had applied for the e-mail notification option.

ECA Marks Deduction Scheme

Please note the following:

(a) Submission Cut-off Time – Unless otherwise advised, the cut-off time for ECA submission will be at 12:00 noon on the day of the deadline. All submission timings will be based on the time recorded by Canvas.

(b) Start Time for Deduction – Students are given a grace period of 12hours. Hence calculation of late submissions of ECAs will begin at 00:00 hrs the following day (this applies even if it is a holiday or weekend) after the deadline.

(c) How the Scheme Works – From 00:00 hrs the following day after the deadline, 10 marks will be deducted for each 24-hour block. Submissions that are subject to more than 50 marks deduction will be assigned zero mark. For examples on how the scheme works, please refer to Section 5.2 Para 1.7.3 of the Student Handbook.
Any extra files, missing appendices or corrections received after the cut-off date will also not be considered in the grading of your ECA assignment.

Plagiarism and Collusion

Plagiarism and collusion are forms of cheating and are not acceptable in any form of a student's work, including this ECA assignment. You can avoid plagiarism by giving appropriate references when you use some other people's ideas, words or pictures (including diagrams). Refer to the American Psychological Association (APA) Manual if you need reminding about quoting and referencing. You can avoid collusion by ensuring that your submission is based on your own individual effort. The electronic submission of your ECA assignment will be screened through a plagiarism detecting software. For more information about plagiarism and cheating, you should refer to the Student Handbook. SUSS takes a tough stance against plagiarism and collusion. Serious cases will normally result in the student being referred to SUSS's Student Disciplinary Group.

(Full marks: 100)

Question 1

Objectives:

- Understand dataset with Data Scientist mindset.
- Exposure to real-world dataset analysis.
- Understand and design computation logic and routines in Python.
- Assess use of Pandas and Dataframes to perform extract, load, transformation and calculation operations.
- Assess the Design and use of Database method to perform create and load operations.
- Conduct visualization in an appropriate way.
- Structure code in appropriate methods (functions), looping and conditions.

You will scrape, analyze, and visualize data on water distribution on Earth from Wikipedia using the URL 'https://en.wikipedia.org/w/index.php?title=Water_distribution_on_Earth&oldid=1193884082'.

Question 1a

Using BeautifulSoup, design a script to scrape and parse data from the table locating at the 'Distribution of saline and fresh water' section into the Pandas data frame with the following columns: 'source', 'volume', 'total_water_percent', 'salt_water_percent', 'fresh_water_percent', 'liquid_surface_fresh_water_percent' and 'parent'. The 'parent' field in each record denotes the hierarchical parent, for instance, 'Pacific Ocean' has 'Oceans' as its parent, and 'Caspian Sea' has 'Saline lakes' as its parent, which in turn is a child of 'Lakes'. Select the correct types for all fields.

(12 marks)

Question 1b

Process the 'volume' column to ensure all values are stored in cubic kilometers.

(3 marks)

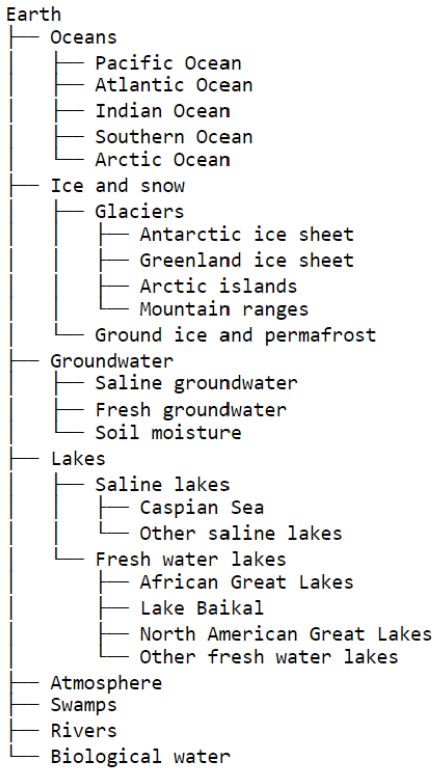
Question 1c

Insert a row into the data frame with 'source' set as 'Earth', and 'volume' calculated as the 'Oceans' volume divided by its total water percentage. Set 'total_water_percent', 'salt_water_percent', 'fresh_water_percent', and 'liquid_surface_fresh_water_percent' all to 100%." Set the 'parent' field of all rows with a None value to 'Earth', except for the row where the 'source' is 'Earth'.

(5 marks)

Question 1d

Design and write a Python script to visualize the data frame as shown in the following image using anytree (https://anytree.readthedocs.io/en/latest/api/anytree.render.html):



(5 marks)

Question 1e

Design and create a Plotly's icicle chart (https://plotly.com/python/icicle-charts/) to visualize water volume within a hierarchical tree structure defined by the 'parent' attribute. Draw **ONE (1)** insight.

(5 marks)

Question 1f

The data frame contains missing data. For instance, the parent record 'Groundwater' has NA values in the 'salt_water_percent' and 'fresh_water_percent' fields, while its child records like 'Saline groundwater' have values for 'salt_water_percent', and other children have values for 'fresh_water_percent'. In such cases, replace the NA values in the parent record with the sum of its children's values for each specific column: 'salt_water_percent', 'fresh_water_percent', and 'liquid_surface_fresh_water_percent'. However, do not alter the values of parent records that already have data sourced from the Wiki page.

(8 marks)

ICT233 Copyright © 2024 Singapore University of Social Sciences (SUSS)

Question 1g

Compute the total volume, in cubic kilometers, of freshwater, saltwater, and liquid surface freshwater, considering that liquid surface freshwater constitutes 0.3% of the total freshwater volume.

(3 marks)

Question 1h

Use matplotlib to design visualization and perform the following tasks:

- Create a pie chart to visualize the volumes of salt and fresh water.
- Create a pie chart to compare the volumes of non-liquid surface freshwater with liquid surface freshwater. The sum of these two categories is the total volume of fresh water.
- Draw one insight.

(6 marks)

Question 1i

Design the program using Pandas, SQLite and SQLAlchemy respectively to identify the smallest ocean.

(6 marks)

Question 2

Objectives:

- Manipulate dataset with data scientist mindset.
- Exposure to real-world dataset analysis.
- Design computation logic and routines in Python.
- Structure code in appropriate methods (functions), looping and conditions.
- Design methods to extract and parse information from the internet.
- Assess use of Pandas and Dataframes to perform extract, load, transformation and calculation operations.
- Conduct visualization in an appropriate way.

The dataset 'fresh_water_withdrawal.csv' details annual fresh water withdrawals by country, measured in billion cubic meters, spanning from 1962 to 2020. Additionally, 'population.csv' provides country-wise population data from 1960 to 2022. Finally, the file 'irrigated_agricultural_land.csv' provides data on the percentage of irrigated agricultural land relative to the total agricultural land in each country, spanning the years 2001 to 2021. All three datasets, originating from 'data.worldbank.org', have undergone preprocessing.

Question 2a

Load in the dataset and perform data explorations. Which country has the earliest recorded year for tracking annual fresh water withdrawals?

(4 marks)

ICT233 Copyright © 2024 Singapore University of Social Sciences (SUSS)

Question 2b

The 'country' column in the dataset comprises both countries and regions. Divide the data frame into two separate ones: one specifically for countries and another for regions. Print out the list of regions.

(5 marks)

Question 2c

Design the logic using dataframe to obtain which country and region (excluding 'World') recorded the highest annual freshwater withdrawals in 2020?

(4 marks)

Question 2d

Design and visualize the annual fresh water withdrawals over the years for the region and country identified in Q2(c) using matplotlib.

(4 marks)

Question 2e

Visualize the annual freshwater withdrawals over the years on a map using Plotly (https://plotly.com/python/scatter-plots-on-maps/). Draw and share **ONE (1)** insight.

(7 marks)

Question 2f

Utilize the data from 'population.csv' to calculate the per capita annual freshwater withdrawal for each country over the years. Create a world map visualization of the per capita values. Plotly is allowed for the visualization.

(8 marks)

Question 2g

To investigate if there's a correlation between the percentage of irrigated agricultural land out of the total agricultural land, as recorded in 'irrigated_agricultural_land.csv', and the annual fresh water withdrawal. Keep only the rows of countries that have at least 10 non-NA values for 'irrigated_percent' for the correlation analysis. Design the visualization, extract and substantiate **ONE (1)** insight based on the visualization.

(10 marks)

Question 2h

Is there any trend in global annual fresh water withdrawal per capita over the years? Share **ONE (1)** observed insight based on the visulization.

(5 marks)

----- END OF ECA PAPER -----