ICT233
Data Programming

Tutor-Marked Assignment

January 2026 Presentation

TUTOR-MARKED ASSIGNMENT (TMA)

This assignment is worth 24% of the final mark for ICT233, Data Programming.

The cut-off date for this assignment is **Sunday, 08 March 2026, 2355 hours.**

Note to Students:

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

1. Please provide answers into the given template ICT233_TMA_JAN25_STUDENT.ipynb between the following tags:

############### Your code starts here ###############

################ Your code ends here ################

2. Please do not modify the given code in the template, do not add/delete/split any code cell.

3. Please print out (use print() function) your analysis and insights.

3. If there is any discrepancy between the questions displayed in the template and those in the TMA paper (in pdf format), please refer to the TMA paper (in pdf format) as the source of truth.

Dataset Information

Important Note: Please use the dataset located in the folder provided with the notebook. There is no need to download it from the URL, as the dataset may change over time and affect your results. Additionally, the provided dataset includes minor modifications from the original source for the purposes of this notebook.

This dataset contains information about **world electricity generation** (https://github.com/owid/energy-data) from various sources and countries.

It is useful for analysing global energy trends, renewable vs non-renewable sources, and electricity generation patterns across different countries and regions.

Answer all questions

Question 1 (74 marks)

Objectives:

- Understand dataset with data scientist mindset.
- Understand and design computation logic and routines in Python.
- Assess use of Python only and Python data structures to perform extract, load, and transformation operations.
- Assess use of Pandas dataframe to perform extract, load, transformation and calculation operations.
- Structure code in appropriate methods (functions), looping and conditions.
- Design methods to perform extract, load, transformation and calculation operations on dataset.
- Conduct visualization in an appropriate way.

Data Cleaning and Exploratory Data Analysis (EDA) on World Electricity Generation Data

(a) Missing Values Analysis

1. **Overall Data Completeness Summary**
   Calculate the following:

   - **Total number of cells in the dataset**: `(n_columns)`: `n_rows × n_columns`
   - **Total number of missing values**: `Σ (missing values per cell)`
   - **Overall data completeness percentage**: `((Total Cells - Total Missing) / Total Cells) × 100`
   - **Overall missing rate percentage**: `100 - Overall Completeness (%)`

2. **Columns with Missing Values**
   Filter and display only the columns with missing values (i.e., missing count > 0). Include the following details:

   - Column name.
   - Count of missing values.
   - Percentage of missing values.

3. **Complete vs Incomplete Records**
   Write code to:

   - Count and display the number of complete records (rows with no missing values) and their percentage.
   - Count and display the number of incomplete records (rows with at least one missing value) and their percentage.

4. **Distribution of Missing Values per Record**

   Write code to analyse and display the distribution of missing values across incomplete records. Specifically:

   - Count the number of records with 1, 2, 3, etc., missing values.
   - Display the results in a readable format.

(5 marks)

(b) Analyse the metadata in `data/owid-energy-metadata.csv` to determine appropriate data types for all columns in the main dataset. Note that understanding the metadata is crucial for all subsequent questions.

(3 marks)

(c) Which top 5 countries generate the most electricity in 2024 (excluding regions like 'Europe' and country groups like 'High-income countries')?

(4 marks)

(d) Create a line plot showing the trend of total electricity generation of the top 5 countries from Question 1(c) over the years. Draw **TWO (2)** insights from the plot.

(4 marks)

(e) Design a visualization to visualize the electricity generation by source for China and the United States from 2010 to 2024. Provide **TWO (2)** insights based on the chart.

(8 marks)

(f) Design a visualization to visualize the electricity generation by source for China and the United States from 2010 to 2024, but this time normalize the data by population. Provide **TWO (2)** insights based on the chart.

(5 marks)

(g) Design a visualization to visualize the electricity generation by source for China and the United States from 2010 to 2024, but this time normalize the data by GDP. Provide **TWO (2)** insights based on the chart.

(5 marks)

(h) To analyse and address the followings:

1. Which country (excluding regions and country groups) has the highest decrease in fossil electricity generation per capita between 2010 and 2024?
2. For this country, find the replaceable energy source that has increased the most in electricity generation per capita during the same period. Provide the name of the energy source and the increase in generation per capita.

(5 marks)

(i) To analyse and perform the followings:

1. Visualize nuclear electricity generation of all countries (excluding regions and country groups) on a world map from 1965 to 2024 using Plotly visualization library (https://plotly.com/python/choropleth-maps). Note that the legend's minimum and maximum values must be global rather than year-specific.
2. Provide **TWO (2)** insights based on the visualization.

(7 marks)

(j) Design a visualization to show how electricity carbon intensity (gCO_2/kWh), carbon_intensity_elec, relates to the share of fossil electricity (%), fossil_share_elec, and share the insights on how this relationship has shifted over time.

(5 marks)

(k) Which countries reduced electricity carbon intensity the most without reducing total electricity demand from 2010 to 2024?

(5 marks)

(l) Which low-carbon electricity technologies (nuclear, hydro, solar, wind, bioenergy, and other renewables) are most strongly associated with reductions in electricity carbon intensity between 2010 and 2024?

Use a robust regression model (https://www.statsmodels.org/stable/rlm.html#technical-documentation) with M = sm.robust.norms.HuberT() to examine the relationship between **reductions** in electricity carbon intensity and **changes** in electricity generation from low-carbon technologies. Treat the reduction in electricity carbon intensity as the dependent variable (*y*) and the changes in generation from low-carbon technologies as the independent variables (*X*). Based on the model summary, identify the technologies with the **largest absolute coefficient values**, noting that positive coefficients indicate that increases in generation from those technologies are associated with reductions in electricity carbon intensity, and that **p-values below 0.05** indicate statistically significant associations.

(7 marks)

(m) How unequal is access to low-carbon electricity across countries with different income levels in 2024?

You may use records for countries whose names contain the word **“income”** and visualize the distribution of low-carbon electricity consumption per capita across different low-carbon energy sources and income levels using an appropriate plot and provide the insights.

(6 marks)

(n) Are electricity systems converging globally in carbon intensity?

1. Per year, compute the coefficient of variation (CV = standard deviation / mean) of electricity carbon intensity across all countries (excluding regions and country groups).
2. Create a line plot to visualize the trend of CV over the years.
3. Provide one insight based on the plot.

(5 marks)

Question 2 (26 marks)

Objectives:

- Understand dataset with data scientist mindset
- Design computation logic and routines in Python
- Conduct visualization in an appropriate way
- Assess use of Pandas dataframe to perform extract, load, transformation and calculation operations
- Design methods to perform extract, load, transformation and calculation operations on dataset
- Assess the design and use of database ORM / SQLite methods to perform extract, load, transformation and calculation operations

(a) Identify the top ten countries (excluding regions and country groups) with the highest net electricity imports as a share of demand, net_elec_imports_share_demand, in 2024, using:

1. Pandas
2. SQL (SQLite)
3. SQLAlchemy ORM

(9 marks)

(b) Do electricity-import-dependent countries decarbonize faster or slower? Compare 2024 against 2010 to reach a clear and logical conclusion. Conduct the analysis using:

1. Pandas
2. SQL (SQLite)
3. SQLAlchemy ORM
4. Provide a clear and logical conclusion based on your findings.

A country is considered electricity-import-dependent if its net electricity imports as a share of demand, net_elec_imports_share_demand, is greater than 0% in 2010. The decarbonization rate can be measured
by carbon_intensity_elec_2010 - carbon_intensity_elec_2024 / (2024 - 2010).

(17 marks)

----END OF ASSIGNMENT---