🧠 Climate Project – Python Code Full Technical Summary

# 🔍 Objective & Context

This project is part of a climate awareness research initiative. The goal is to investigate attitudes, behaviors, and psychological variables related to climate change. The survey was designed by the group and distributed to a wide audience. All responses were cleaned, encoded, and analyzed in Python using data science techniques.  
The professor requested a complete pipeline from raw data to clustering, regression, and hypothesis validation. All code was executed in JupyterLab (Anaconda environment).

# 🧼 1. Dataset Cleaning & Encoding

• Input file: Raw Excel file containing responses (some in Italian, some in English).  
• Translation: Every string-based answer was mapped to English manually or via a dictionary.  
• Encoding:  
 - Likert scales (1–5) were created for frequency and agreement questions.  
 - Binary encoding (0/1) used for yes/no or support/do not support types.  
 - Complex options like 'rarely/sometimes/always' were mapped numerically (e.g., 1–3).  
• Output file: fully\_encoded\_dataset\_complete.xlsx → This is the master dataset for all analysis.

# 🌍 2. Cultural Behavior Analysis (water\_behavior\_analysis.py)

• Goal: Test if geographic or cultural origin influences sustainable water behavior.  
• Grouping:  
 - Respondents were grouped into Cultural\_Group based on country and region.  
 - For Italy: subgroups Italy\_North, Italy\_Center, Italy\_South created from the region.  
• Behavior:  
 - Water-use habits (e.g., reusing AC water, letting water run while brushing teeth) were extracted.  
 - A tidy format (long format) was created for frequency counts.  
• Output: Excel file with a summary of water-use responses by cultural group.

# 📊 3. Clustering Analysis (KMeans + PCA)

• Goal: Group people with similar environmental attitudes and behaviors.  
• Preprocessing:  
 - Selected all Likert-type questions with 3–7 unique values.  
 - Standardized with sklearn's StandardScaler.  
• KMeans:  
 - Elbow Method and Silhouette Score tested for K=2 to 10.  
 - Optimal K = 2 based on Silhouette Score.  
• PCA:  
 - Used to visualize clusters in 2D space.  
• Profiling:  
 - Cluster centers were interpreted and compared.  
 - Added cluster labels to dataset for further analysis.

# 🧪 4. Regression Hypothesis 1 — Belief → Willingness to Pay

• Hypothesis: People who believe climate change is real and harmful are more likely to support renewable energy financially.  
• X (Predictors):  
 - Belief that climate change is happening  
 - Worry level about global warming  
 - Harm to future generations  
 - Support for renewable energy, tax rebate, carbon tax policies  
• Y (Target): Willingness to pay more for energy (binary)  
• Method: Logistic Regression  
• Output: Accuracy score ~57%, confusion matrix, classification report  
• Interpretation: Weak predictive power, likely due to balanced classes and small sample.

# 📈 5. Regression Hypothesis 3 — Institutional Trust → Responsibility

• Hypothesis: Higher expectations of institutional action correlate with stronger personal responsibility.  
• X: Average of Q20–Q25 (attitudes toward action by institutions, government, citizens, companies).  
• Y: Q34 – Personal sense of responsibility for climate action.  
• Method: Linear Regression  
• Output: R² ≈ 0.17  
• Interpretation: Weak model but positive and meaningful trend. Citizens and companies matter more than state trust.

# 🏛️ 6. Regression Follow-Up — Public vs Civil Trust Split

• Idea: Instead of one trust score, divide between:  
 - Public Trust: Q21–Q24 (government-related)  
 - Civil Trust: Q20 + Q25 (companies, citizens)  
• Result:  
 - Civil trust coefficient: ~+0.48  
 - Public trust coefficient: ~+0.06  
 - R² improved slightly (~0.20)  
• Conclusion: Responsibility is more tied to trust in society than government.

# 🧩 7. Clustered Correlated Questions (CCQ)

• Goal: Find which questions most distinguish between the two KMeans clusters.  
• Method:  
 - Calculated point-biserial correlation for each Likert variable vs cluster label (0/1).  
 - Sorted results to highlight strongest signals.  
• Output:  
 - Table of top 15 most correlated items  
 - Heatmap of inter-item correlations among those  
• Interpretation: Supports that clustering has meaningful behavioral separation.

# 📦 8. Deliverable Files (to share)

- fully\_encoded\_dataset\_complete.xlsx — Final dataset

- water\_behavior\_analysis.py — Water behavior by culture

- institutional\_trust\_regression.py — Trust vs Responsibility

- trust\_subgroups\_regression.py — Public vs Civil trust

- clustered\_correlated\_questions.py — CCQ correlation explorer

- Optional: regression\_plots/, .ipynb notebooks, etc.

# 📅 Submission Deadline (AoE)

Final deadline: June 10, 14:00 Italy time (based on Midnight AoE)