# Preprocessing & Feature Engineering Report

## Team Lead: Kevin Puduseril

## Recorder: Maxwell Tuttle

## Spokesperson: Melissa Best

# Recap: Background & Question

Our research question is:
What combination of state characteristics (political leaning, grid capacity, median income, existing energy mix, etc.) and renewable energy policy designs (e.g., Renewable Portfolio Standard targets, subsidy types) best predict a successful change in fossil fuel shares per state in the U.S.?

Hypothesis: States with liberal political climates, robust grid capacity, higher median incomes, and ambitious, well-designed renewable energy policies will demonstrate greater success in reducing fossil fuel energy shares.

Prediction: A machine learning model will show these features as significantly and positively correlated with fossil fuel reduction over the past decade.

# Methods

## Overview of Plan and Revisions

Previously, our plan involved merging multiple datasets and cleaning them for analysis. Based on our EDA, we revised the plan by:

- Defining a clear outcome variable: percent change in fossil fuel energy share (2000-2019).

- Creating a binary success variable (1 if a state's reduction > national median).
- Identifying missing data in RPS variables and income values that required imputation or exclusion
- Noting that simple binary indicators (e.g., RPS presence) may not be sufficient—requiring engineered composite or interaction features.

# Preprocessing Methods

## Merging and Key Matching

- Merged datasets using consistent state identifiers (e.g., state name or codes).
- Ensured consistent year ranges across datasets where possible (primarily 2000-2019)

## Handling Missing Values

- For median_income and political_score: Removed or imputed based on neighboring state medians or national trends.
- For RPS variables: Missing data treated as "no policy" only when appropriate (e.g., if no data and no evidence of legislation).
- For numerical data: Used mean/median imputation where patterns were stable.
- For the policies that were observed, after converting them to binary sources, some had 0s in all values in a column. We chose to remove these columns, as they added no significant value. Filtering out territories from the dataset removed the zero values.
- For the senate data, there were some entries with NaN values. We chose to handle these by inputting zeros for those values, because zeros are at least something we can do calculations with.

## Collapsing Variables and Encoding

- Categorical variables like policies were binary encoded and combined into a single variable (Total_Policies) capturing the policies taking effect in a given year.
- Income and RPS target values were binned into quartiles to reduce noise and facilitate stratified analysis.
- Created interaction terms such as:
    - rps_present x political_bucket
    - median_income x political_bucket

# Feature Engineering Methods

## 1. Unsupervised Methods

### Principal Component Analysis (PCA)

- Applied PCA to identify dominant dimensions in the energy mix and policy design features.

### Cluster Analysis (K-Means)

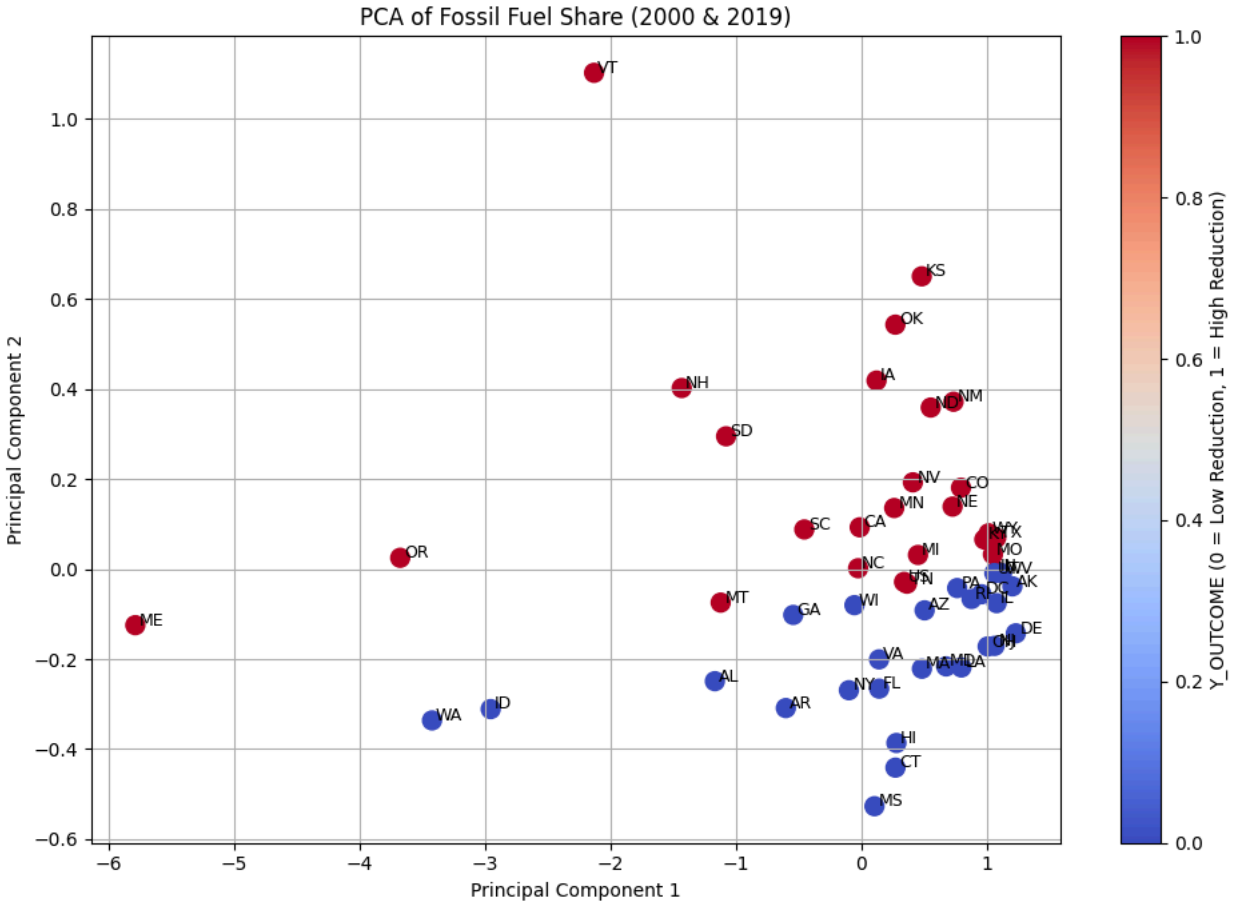
- Applied K-Means clustering on states using:
    - Energy source proportions
    - Median Income
    - Political Score
    - RPS target percentages

## 2. Supervised Feature Engineering Plan

We are preparing for supervised modeling by:

- Selecting key predictors based on EDA (e.g., delta_fossil_share, rps_target_pct, median_income, political_score, policy_count, and interactions).
- Considering polynomial features and logarithmic transformation for non-linear variables (e.g., total energy generated).
- Planning to apply Lasso/Ridge regression and tree-based models (e.g., Random Forest, XGBoost) to assess variable importance and reduce multicollinearity.

# Results & Interpretation




To better understand patterns in state-level fossil fuel dependency and its relationship to energy transition success, we performed Principal Component Analysis (PCA) using fossil fuel share data from the years 2000 and 2019. The resulting PCA plot visualizes the structural variance in fossil reliance across U.S. states and reveals important insights into energy transition trajectories.

The two input features — fossil fuel share in 2000 and 2019 — were standardized and transformed into two principal components (PC1 and PC2), which together capture the primary axes of variation in the data.

**Principal Component 1 (PC1): Overall Fossil Fuel Dependency**
The horizontal axis (PC1) appears to reflect the overall level of fossil fuel reliance across both time points. States on the left of the graph, such as Vermont (VT), Maine (ME), and Oregon (OR), likely either:

- began the period with low fossil fuel dependency,
- achieved significant reductions by 2019, or

- exhibited atypical energy generation patterns.

In contrast, states on the right of the PC1 axis (e.g., Mississippi, Arkansas, Alabama) tended to maintain consistently high reliance on fossil fuels. These states showed limited progress in reducing fossil fuel share, as confirmed by their classification in the outcome variable (Y_OUTCOME = 0).

**Principal Component 2 (PC2): Directional or Source-Specific Change**

The vertical axis (PC2) may capture source-specific or temporal aspects of fossil fuel change — such as transitions from coal to gas or the timing of reductions. States with higher PC2 values (e.g., Kansas, Vermont) may have undergone more distinctive shifts in energy composition compared to those lower on the axis.

**Outcome-Based Color Coding**

Each state is color-coded by Y_OUTCOME, a binary variable indicating whether a state reduced its fossil fuel share more than the national median:

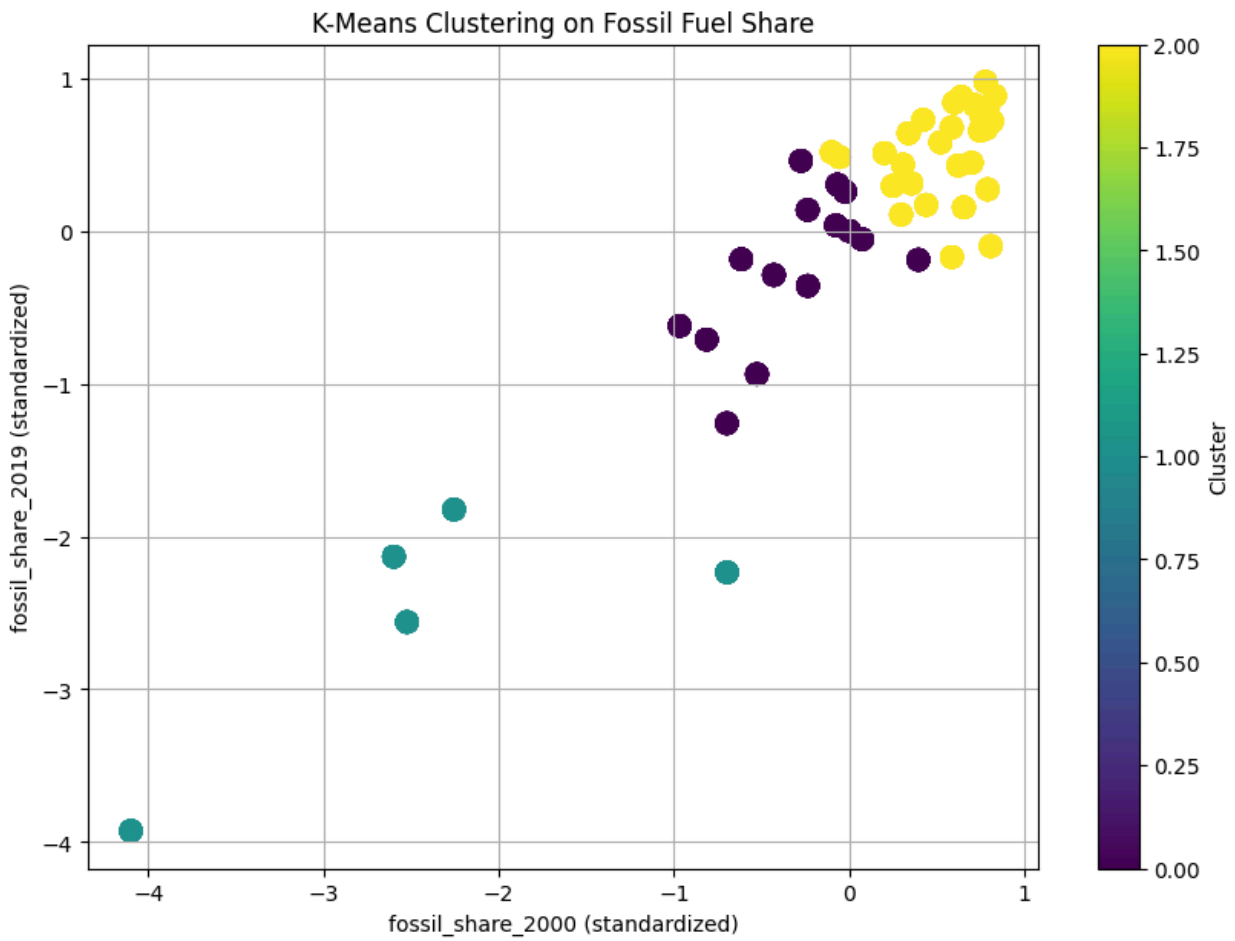
- **Red points** represent states with high fossil reduction (successful transitions).
- **Blue points** represent states with low or negative reduction.

A clear pattern emerges: **successful states cluster to the left**, indicating that PC1 is strongly associated with positive fossil fuel reduction outcomes. Meanwhile, **unsuccessful states are concentrated on the right**, suggesting they retained higher fossil dependence over time.

**Interpretation and Implications**

This PCA analysis reveals that the combination of fossil fuel share in 2000 and 2019 effectively distinguishes states with successful energy transitions. The visual separation of outcome groups along PC1 suggests that fossil fuel reliance alone—even without policy variables—contains meaningful predictive information about a state's transition progress.

This supports the hypothesis that structural energy characteristics (such as baseline fossil dependence and trajectory of change) are key indicators of energy transition outcomes. These insights can guide future modelling efforts, including supervised learning approaches, by emphasizing the predictive value of fossil share trends.




This scatter plot shows the results of applying K-means clustering (k=3) to standardized fossil fuel share values for U.S. states in 2000 and 2019.

1. Axes
    a. X-axis: Standardized fossil fuel share in 2000.
    b. Y-axis: Standardized fossil fuel share in 2019.
    c. Since both variables are standardized, 0 represents the mean value for that year, and values above/below show how far a state deviates from the average.
2. Clusters
    a. Cluster 0 (Purple) -

        States with moderate fossil fuel share in 2000 but lower than average in 2019.

        - Likely transitional states that made moderate progress in reducing fossil fuel reliance.

- Positioned around the middle on X-axis but below average on Y-axis.

b. Cluster 1 (Teal) -

States with well below average fossil fuel share in both 2000 and 2019.

- These are likely early adopters of renewable energy or states with inherently low fossil reliance (e.g., hydro/nuclear-heavy states).
- Positioned far left and low on the plot

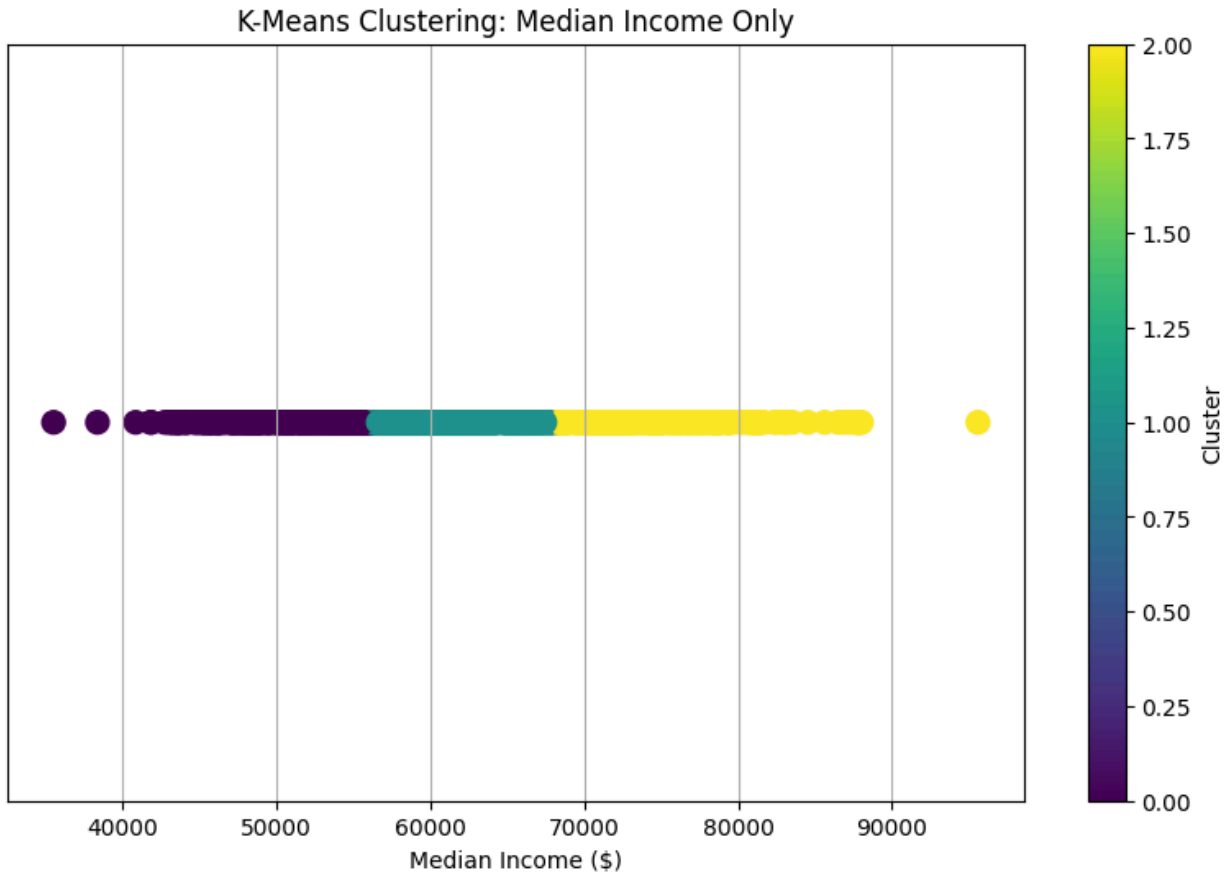
c. Cluster 2 (Yellow) -

States with above average fossil fuel share in both 2000 and 2019.

- These are the high-dependency states that made little relevant change over time.
- Concentrated in the upper-right quadrant.

3. Insights

- Cluster separation is largely along PC1-like behavior: overall fossil dependence is the primary factor dividing groups.
- The transition states (Cluster 0) form a bridge between high-dependency and low-dependency states.
- Cluster 1 is small — meaning only a few states had consistently low fossil fuel reliance.
- Cluster 2 is the largest group, indicating that many states maintained high fossil dependency over two decades.




This plot shows the results of K-Means clustering (k=3) applied only to Median Income ($) values.

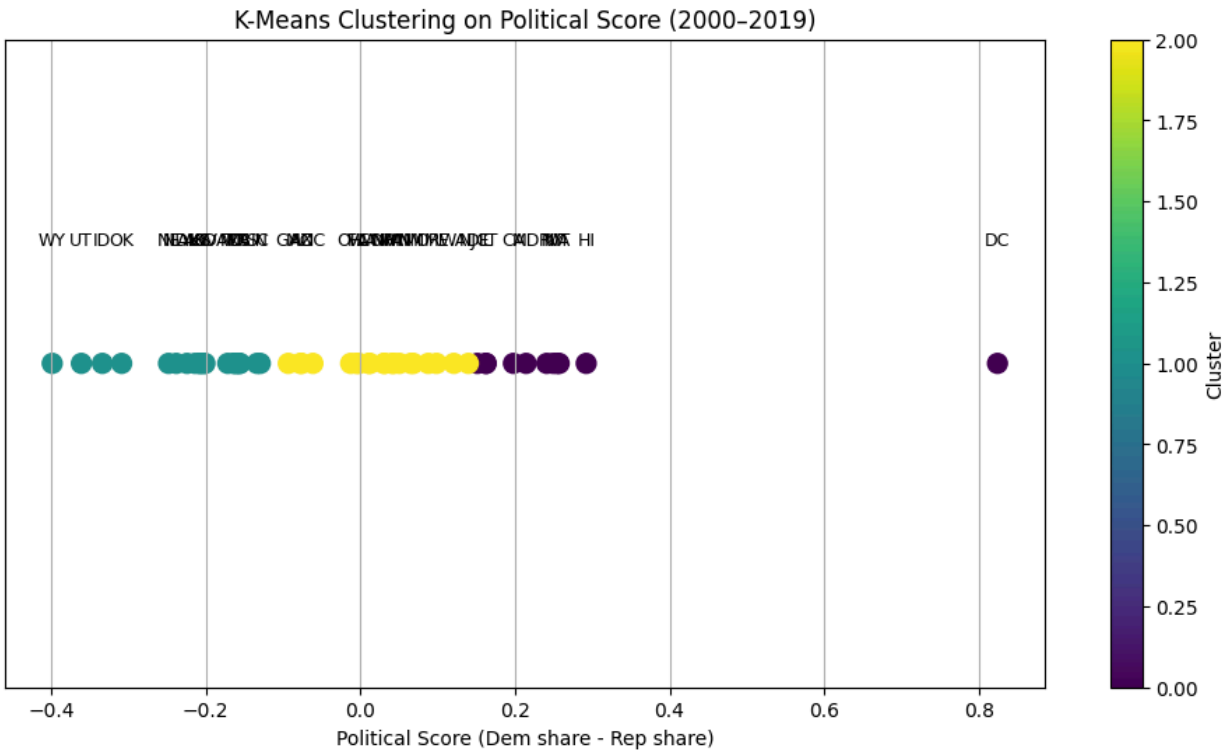
1. Axes
    - X-axis: Median household income in U.S. dollars.
    - Y-axis: Flat (set to 0) since there's only one feature — the vertical axis is just a placeholder to separate points visually.
    - Color: Indicates the cluster assigned by K-Means.

2. Clusters
    - Cluster 0 (Purple) -
        - Lowest income group
        - Median incomes roughly below $58-60k.

        - These states may face greater financial constraints in implementing large-scale renewable energy transitions

- Cluster 1 (Teal) -
    - Middle-income group
    - Median incomes around $60k-$70k.
    - These states sit in the middle economically—they may have moderate resources for transition policies.
- Cluster 3 (Yellow) -
    - Highest income group
    - Median incomes roughly above $70k, extending to nearly $95k.
    - These states have stronger economic capacity, potentially enabling larger investments in renewable energy infrastructure.

3. Insights
    a. The boundaries between clusters are purely based on income—no energy-related variables were used here.
    b. Most data points fall into the middle and high-income clusters, meaning relatively few states are in the lowest-income range.




This plot shows the clustering of U.S. states based on their average political score between 2000 and 2019, where:

Political Score = Average Democratic Vote Share - Average Republican Vote Share

1. Axes and Color Coding
    a. X-axis: Political score (range: ~-0.4 to +0.8)
        i. Negative scores -> More Republican-leaning
        ii. Positive scores -> More Democratic-leaning
        iii. Near zero -> Balanced or swing states
    b. Y-axis: Flat, since there's only one dimension — the vertical spacing is purely for visualization
    c. Colors: Cluster assignments from K-Means (k=3)
        i. Cluster 0 (Purple): Strong Democratic lean
        ii. Cluster 1 (Teal): Strong Republican lean
        iii. Cluster 2 (Yellow): Moderate or swing states

2. Cluster Insights
    a. Cluster 1 (Teal, left side):
        i. Consistently Republican, with negative political scores below -0.15
    b. Cluster 2 (Yellow, middle):
        i. Political scores near 0, indicating mixed voting patterns over the two decades.
    c. Cluster 0 (Purple, right side):
        i. Scores well above 0.15
3. Patterns and Observations
    a. There's a clear separation between strongly Republican, swing, and strongly Democratic states.
    b. Swing states form the middle band, bridging the extremes.
    c. Washington D.C. stands out as an extreme Democratic outlier with a much higher score than any state.

# Discussion & Next Steps

## Key Takeaways

Our research question was: *What combination of state characteristics (political leaning, grid capacity, median income, existing energy mix, etc.) and renewable energy policy designs (e.g., Renewable Portfolio Standard targets, subsidy types) best predict a successful change in fossil fuel shares per state in the U.S.?* After some preprocessing and feature engineering, this question still remains unanswered.

One important task that was completed in this week of the project was the completion of the merging of the datasets and initial cleaning. This initial cleaning included removing columns that were deemed not necessary, whether they were repetitive or not important for our purpose. Subsetting the various datasets to only include values from the years 2000-2019 to ensure that all entries had the proper data. We ensured that the data only included the 50 US states, not including territories or DC. We also performed calculations, such as generating percentages, for columns where it was appropriate, such as republican versus democrat votes for presidential and senate elections.

## Next Steps

Our next steps to prepare for the Modeling Plan in the next Module are as follows: We will finalize our feature set for modeling. This will include ensuring that the features we plan on using are normally distributed or transformed to behave as such. We plan to run multiple regression and tree-based classification models to assess feature importance and prediction accuracy. We plan to apply k-fold cross-validation to ensure model competency. We will refine clustering and dimensional reduction methods based on model results.