Data pipeline:

- Cleaning & Processing
    - Removed StateCode X3, X5, US, DC, which conveniently removed all NA values
    - Changed MSN Codes and Amount columns into 29 new columns with each MSN and its corresponding amount value
        - Deleted MSN Code columns "BDFDB", "BFPRP", "CLPRP", "COPRK", "NGMPK", "NGMPP", "PAPRP", "TEPRB", "TETCB", "WDEXB", "WDPRB", "WDTCB", "WSTCB" because we saw that these columns were repeats (or similars) of other columns but with different units
    - Found external data sources and used merges, left joins, etc. to combine them to the original dataset
- Adding External Sources
    - After some initial analysis, we saw that our initial models were weak. We wanted to add more relevant parameters to make our model strong and more accurate.
    - We believe there are economic factors that play into the investments, so we searched for additional sources with economic data per state.
    - Additional sources used
        - Disbursement data: https://revenuedata.doi.gov/downloads/disbursements/
        - Energy prices data: https://www.eia.gov/electricity/state/archive/2016/
        - Population and GDP data: https://www.eia.gov/state/seds/seds-data-complete.php?sid=US#StatisticsIndicators
- Principal component analysis
    - Since we do have many features and want to avoid the curse of dimensionality, we decided to use PCA to reduce dimensionality.
    - We used feature extraction to shrink all features into one using the first principal component vector
    - Can see that PC1 accounts for lots of the variance
    - We continued to use PC1 in our regressions
- Clustering by K-Means to increase independent samples from per state to per cluster
    - Having only 5 observations for each state was not enough to perform many models
    - Decided to cluster states into “regions” to give us more observations to work with per cluster in order to have better regression results
    - Using silhouette score to determine optimal K value
        - Silhouette score denotes how much cluster regions overlap with another, the lower the score the more distinct and useful the clusters are for classification
- Models
    - Simple Linear Regression with one predictor PC1 predicting Assistance
        - P-values mostly small (< .05), statistically significant
    - Random Forest Clustering

        - Using multiple decision trees, we can use an unsupervised machine learning algorithm to create a model for predicting future values of investment
    - Partial Least Squares
        - Used for few observations but many variables
        - Limitations include assuming independence within each cluster; cannot be used for single-state clusters (two of them) due to lack of observations; included negative values
- Prediction
    - Cluster test data by the cluster model
    - Make predictions according to cluster
- Scoring and Assessment
    - RMSE: all relatively small, showing our models were accurate in predicting the renewable energy investment for 2020

| **Cluster** | **Model** | **RSME** |
|---|---|---|
| Cluster 1 | SLR; full model | 0.3972441 |
| Cluster 2 | SLR; full model | 3.297159 |
| Cluster 3 | SLR; full model | 3.699248 |
| Cluster 4 | SLR; full model | Inf |
| Cluster 5 | SLR; full model | 0.2403496 |
| Cluster 1 | SLR; no GDP, Population | 0.3960565 |
| Cluster 2 | SLR; no GDP, Population | 3.273118 |
| Cluster 3 | SLR; no GDP, Population | 3.665525 |
| Cluster 4 | SLR; no GDP, Population | Inf |
| Cluster 5 | SLR; no GDP, Population | 0.2400619 |
| Cluster 1 | SLR; no Net, GDP, Pop | 0.3511143 |
| Cluster 2 | SLR; no Net, GDP, Pop | 10.68734 |
| Cluster 3 | SLR; no Net, GDP, Pop | 2.191303 |
| Cluster 4 | SLR; no Net, GDP, Pop | Inf |
| Cluster 5 | SLR; no Net, GDP, Pop | 0.2296184 |
| Cluster 1 | SLR; no Retail, Net, GDP, | 0.3040538 |

| | | |
|---|---|---|
| | Pop | |
| Cluster 2 | SLR; no Retail, Net, GDP, Pop | 0.4187101 |
| Cluster 3 | SLR; no Retail, Net, GDP, Pop | 0.4938834 |
| Cluster 4 | SLR; no Retail, Net, GDP, Pop | Inf |
| Cluster 5 | SLR; no Retail, Net, GDP, Pop | 0.2868967 |
| Cluster 1 | SLR; no Disburse, Retail, Net, GDP, Pop | 0.3155448 |
| Cluster 2 | SLR; no Disburse, Retail, Net, GDP, Pop | 0.3879026 |
| Cluster 3 | SLR; no Disburse, Retail, Net, GDP, Pop | 1.29119 |
| Cluster 4 | SLR; no Disburse, Retail, Net, GDP, Pop | Inf |
| Cluster 5 | SLR; no Disburse, Retail, Net, GDP, Pop | 0.2312127 |
| Cluster 1 | SLR; no Net, Disburse, GDP, Pop | 19.07823 |
| Cluster 2 | SLR; no Net, Disburse, GDP, Pop | 379.8185 |
| Cluster 3 | SLR; no Net, Disburse, GDP, Pop | 3070.547 |
| Cluster 4 | SLR; no Net, Disburse, GDP, Pop | Inf |
| Cluster 5 | SLR; no Net, Disburse, GDP, Pop | 3.074248 |
| Cluster 1 | SLR; no Net, Retail | 0.3080473 |
| Cluster 2 | SLR; no Net, Retail | 0.4464448 |
| Cluster 3 | SLR; no Net, Retail | 0.560616 |
| Cluster 4 | SLR; no Net, Retail | Inf |

| | | |
|---|---|---|
| Cluster 5 | SLR; no Net, Retail | 0.2869295 |
| Cluster 1 | RandomForestClassifier | 35096318 |
| Cluster 2 | RandomForestClassifier | 14986823 |
| Cluster 3 | RandomForestClassifier | 697648 |
| Cluster 4 | RandomForestClassifier | 23145967 |
| Cluster 5 | RandomForestClassifier | 679829 |