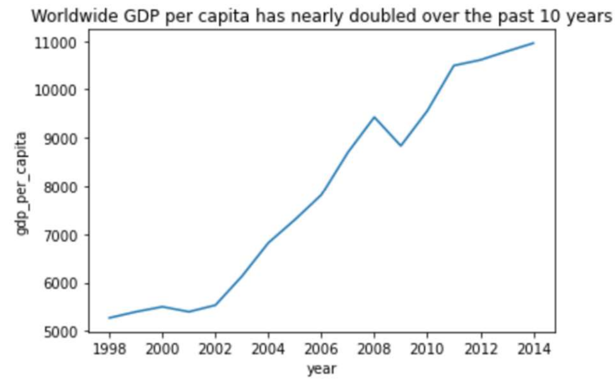
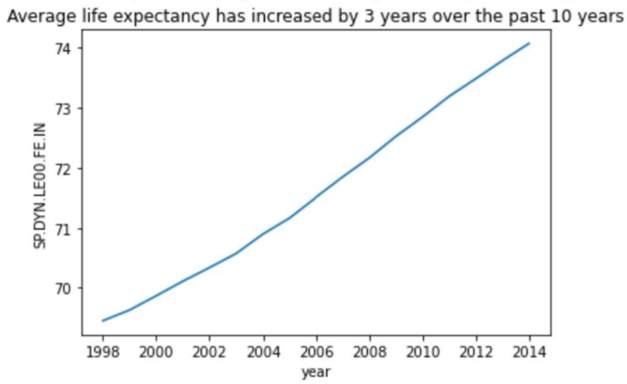
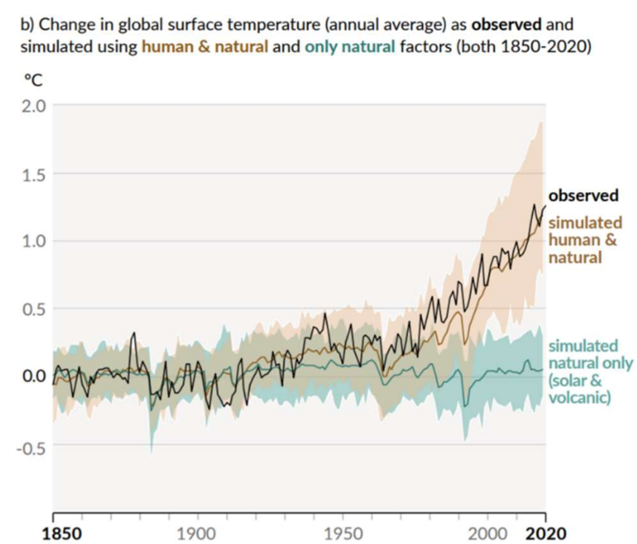
INTRODUCTION

We are living in an epoch of unprecedented human flourishing due to the major technological advances made during the first three industrial revolutions and the spread of democracy. This improvement is evidenced by increasing human life expectancy, and improving GDP per capita among other metrics (source: The World Bank). These are trends that every humanitarian hopes to continue going forward, but with every benefit there also comes a cost.







Over the past several decades we have observed meaningful impact on the global climate tied to human activities (source: IPCC 2021 report). Per the IPCC 2021 report, if these trends continue human flourishing may be at risk.

PROBLEM DESCRIPTION

The purpose of my analysis is to better understand the relationship between human flourishing and climate change through examination of high level data sourced from The World Bank. I will attempt to identify the key drivers that impact GDP per capita (human flourishing indicator). I will also perform similar analysis on greenhouse gas emissions per capita (climate change indicator). Finally, I will perform categorical supervised modeling to identify solutions that result in continued improvement in GDP per capita while reducing greenhouse gas emissions per capita. Another question I hope to study with this analysis is if we are seeing evidence of diminishing returns with regards to human flourishing due to increased greenhouse gas emissions. Evidence for this would be non-linear dependence within the data set for variables related to these targets.

DATA DESCRIPTION

The World Bank has made available an API that can be accessed freely by the public. In addition, a python library has been created to facilitate usage of this data within python using pandas (https://pypi.org/project/wbgapi/). The World Bank dataset is extensive and there are many issues with the data that need to be addressed prior to usage in predictive modeling. For example, data quality and completeness can vary meaningfully for each country. There are also issues that can occur due to changes in the political landscape like when the Soviet Union dissolved in 1991 to form 12 separate nation-states. I assume that many of these issues that can be seen in the data are due to government reporting protocol, government transparency (think North Korea), latency in reporting, and assumptions made by the analysts that gather and process the data from these various sources. Below is a table of data columns considered for inclusion in this analysis with a description of that data field. Also included are comments regarding the final inclusion of that data field for purposes of the final analysis.

Indicator	Code	Comment
Access to electricity (% of population)	EG.ELC.ACCS.ZS	
Electricity production from coal sources (% of total)	EG.ELC.COAL.ZS	
Electricity production from oil, gas and coal sources (% of total)	EG.ELC.FOSL.ZS	Redundant measure, not used in final modeling
Electricity production from hydroelectric sources (% of total)	EG.ELC.HYRO.ZS	
Electric power transmission and distribution losses (% of output)	EG.ELC.LOSS.ZS	
Electricity production from natural gas sources (% of total)	EG.ELC.NGAS.ZS	
Electricity production from nuclear sources (% of total)	EG.ELC.NUCL.ZS	
Electricity production from oil sources (% of total)	EG.ELC.PETR.ZS	
Electricity production (kWh)	EG.ELC.PROD.KH	Could not pull through API, did not use
Electricity production from renewable sources, excluding hydroelectric (% of total)	EG.ELC.RNWX.ZS	
Electric power consumption (kWh)	EG.USE.ELEC.KH	Could not pull through API, did not use
Exports of goods and services (% of GDP)	NE.EXP.GNFS.ZS	
Imports of goods and services (% of GDP)	NE.IMP.GNFS.ZS	
Agriculture, forestry, and fishing, value added (% of GDP)	NV.AGR.TOTL.ZS	

Manufacturing, value added (% of GDP)	NV.IND.MANF.ZS	Not used as it is a subset of NV.IND.TOTL.ZS
Industry (including construction), value added (% of GDP)	NV.IND.TOTL.ZS	
Services, value added (% of GDP)	NV.SRV.TOTL.ZS	
Labor force, female (% of total labor force)	SL.TLF.TOTL.FE.ZS	Sparse data set, did not use
Labor force, total	SL.TLF.TOTL.IN	Sparse data set, did not use
Population, total	SP.POP.TOTL	
GDP (current US$)	NY.GDP.MKTP.CD	
Total greenhouse gas emissions (kt of CO2 equivalent)	EN.ATM.GHGT.KT.CE	
Methane emissions (kt of CO2 equivalent)	EN.ATM.METH.KT.CE	Redundant measure, not used in final modeling
CO2 emissions from electricity and heat production, total (million metric tons)	EN.CO2.ETOT.MT	Could not pull through API, will not use
Life expectancy at birth, female (years)	SP.DYN.LE00.FE.IN	
Urban population	SP.URB.TOTL	
Fertility rate, total (births per woman)	SP.DYN.TFRT.IN	
Strength of legal rights index (0=weak to 12=strong)	IC.LGL.CRED.XQ	Sparse data set, did not use
Multidimensional poverty headcount ratio (% of total population)	SI.POV.MDIM	Sparse data set, did not use
Gini index (World Bank estimate)	SI.POV.GINI	Sparse data set, did not use

When these fields are pulled through The World Bank API, each row is indexed with the "economy" and the "time". The "economy" is typically the country and the "time" is a string that represents the calendar year represented by that data point. As I went through the process of preparing and cleaning the data set, I took several steps.

1) Exploratory data analysis
2) Removed columns with sparse data sets as shown above
3) Reviewed the complete data set for each "economy" individually to visually inspect and determine if that economy should be included based on the completeness of its data set. I kept data for 133 economies (one of which is a world level aggregation, but not used in modeling) and excluded data from 131 economies.
4) Imputed values into the EG.ELC.ACCS.ZS (access to electricity) column as it had issues with data sparseness, but was important for analysis. To impute values, I looked for null

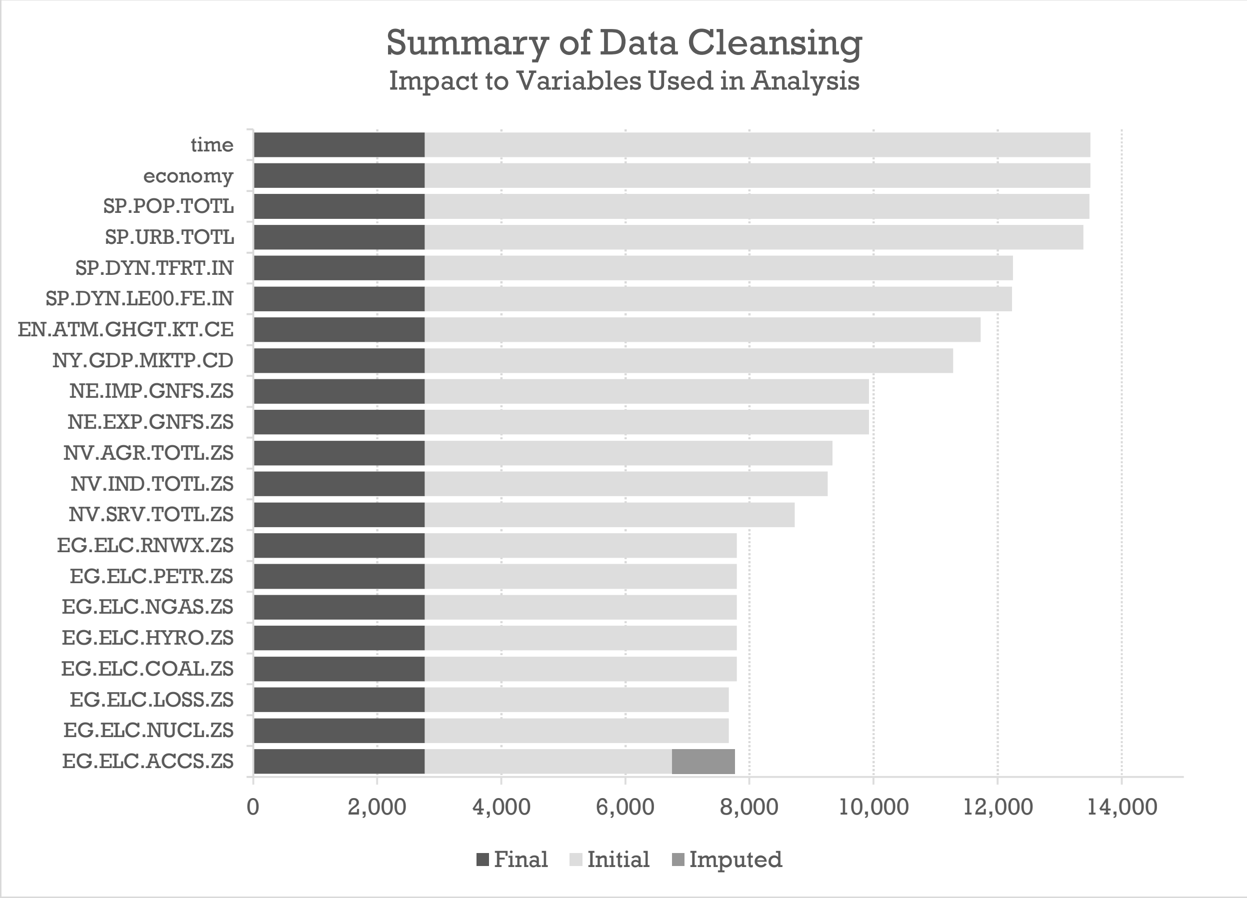
values that are preceded by a value of 100 (meaning 100% of the population has access to electricity) and assumed that the null value can be populated with a value of 100. This assumes that once an economy achieved 100% access to electricity, it stays at that level.

5) Removed rows with null values in the remaining data set
6) Created calculated columns
   a. GDP per capita: NY.GDP.MKTP.CD / SP.POP.TOTL
   b. Greenhouse gas emissions per capita: EN.ATM.GHGT.KT.CE / SP.POP.TOTL
   c. Urban population as a percent of total: SP.URB.TOTL / SP.POP.TOTL * 100
   d. Year: parsed the text string in the "time" column and converted the string representation of the 4 digit year to an integer

The cleaning process reduced the overall data set from a size of 13,495 rows to 2,768. The data sets have been published as csv files to one of my Github repositories (linked below).

- Raw data
- Cleaned data for the individual countries
- Cleaned data for the world in aggregate

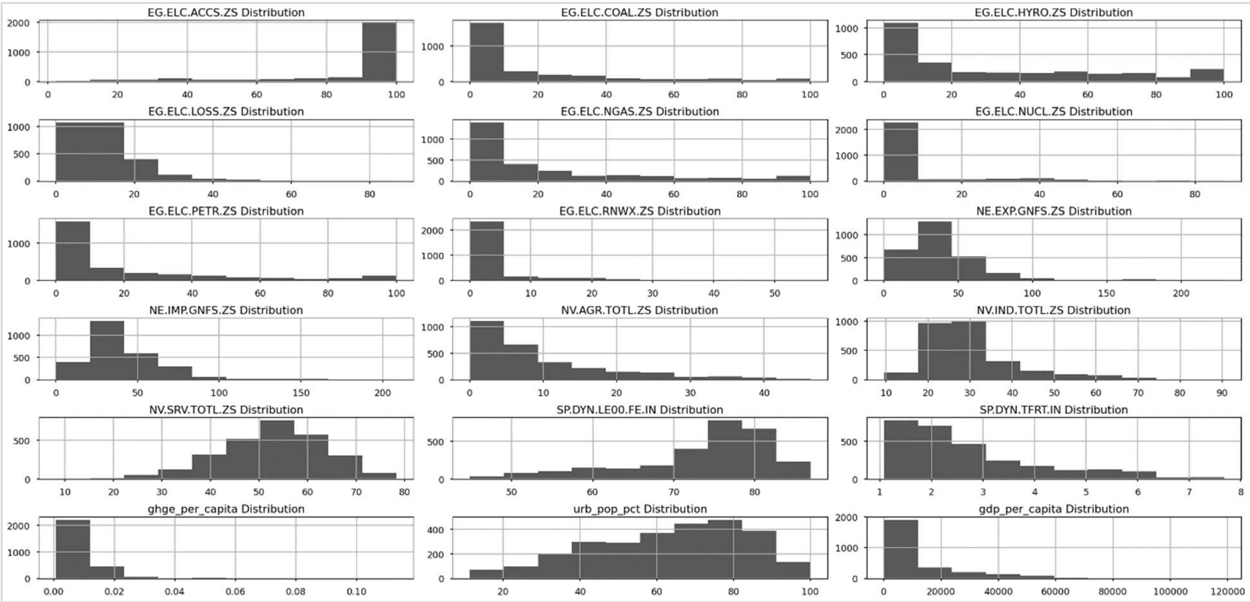
EXPLORATORY DATA ANALYSIS



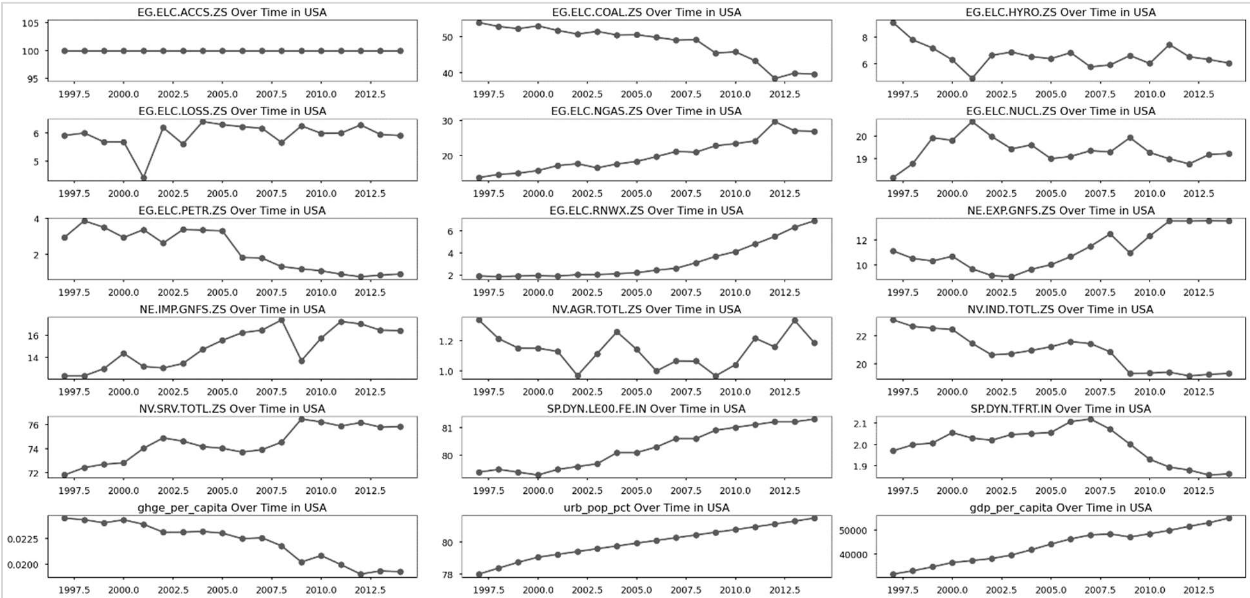
Country Level Descriptive Statistics (excludes 17 rows associated with World aggregate)

	count	mean	std	min	25%	50%	75%	max
EG.ELC.ACCS.ZS	2751	86.69	24.28	2.33	86.09	100.00	100.00	100.00
EG.ELC.COAL.ZS	2751	17.05	25.50	0.00	0.00	2.41	26.46	100.00
EG.ELC.HYRO.ZS	2751	31.40	32.32	0.00	2.76	17.45	56.51	100.00
EG.ELC.LOSS.ZS	2751	13.14	9.95	0.00	6.75	10.68	16.31	86.75
EG.ELC.NGAS.ZS	2751	21.12	27.74	0.00	0.00	9.61	31.79	100.00
EG.ELC.NUCL.ZS	2751	6.53	15.60	0.00	0.00	0.00	0.00	87.44
EG.ELC.PETR.ZS	2751	19.21	26.73	0.00	1.05	5.36	28.09	100.00
EG.ELC.RNWX.ZS	2751	2.96	6.01	0.00	0.00	0.40	2.52	55.85
NE.EXP.GNFS.ZS	2751	39.67	27.42	0.10	23.12	32.95	49.47	228.99
NE.IMP.GNFS.ZS	2751	41.13	24.88	0.06	25.85	34.73	51.11	208.33
NV.AGR.TOTL.ZS	2751	9.77	9.44	0.03	2.99	6.32	13.64	46.32
NV.IND.TOTL.ZS	2751	30.22	11.38	9.48	23.14	27.26	33.54	90.51
NV.SRV.TOTL.ZS	2751	52.66	10.72	8.15	45.86	53.51	60.20	78.31
SP.DYN.LE00.FE.IN	2751	73.48	8.71	44.85	70.23	75.93	79.60	86.83
SP.DYN.TFRT.IN	2751	2.72	1.40	1.08	1.67	2.24	3.34	7.68
ghge_per_capita	2751	8.3E-3	8.4E-3	7.3E-4	3.1E-3	6.3E-3	1.1E-2	1.1E-1
urb_pop_pct	2751	62.86	19.85	11.35	48.19	65.97	78.61	100.00
gdp_per_capita	2751	11,998	16,514	60.46	1,488	4,525	16,015	118,824

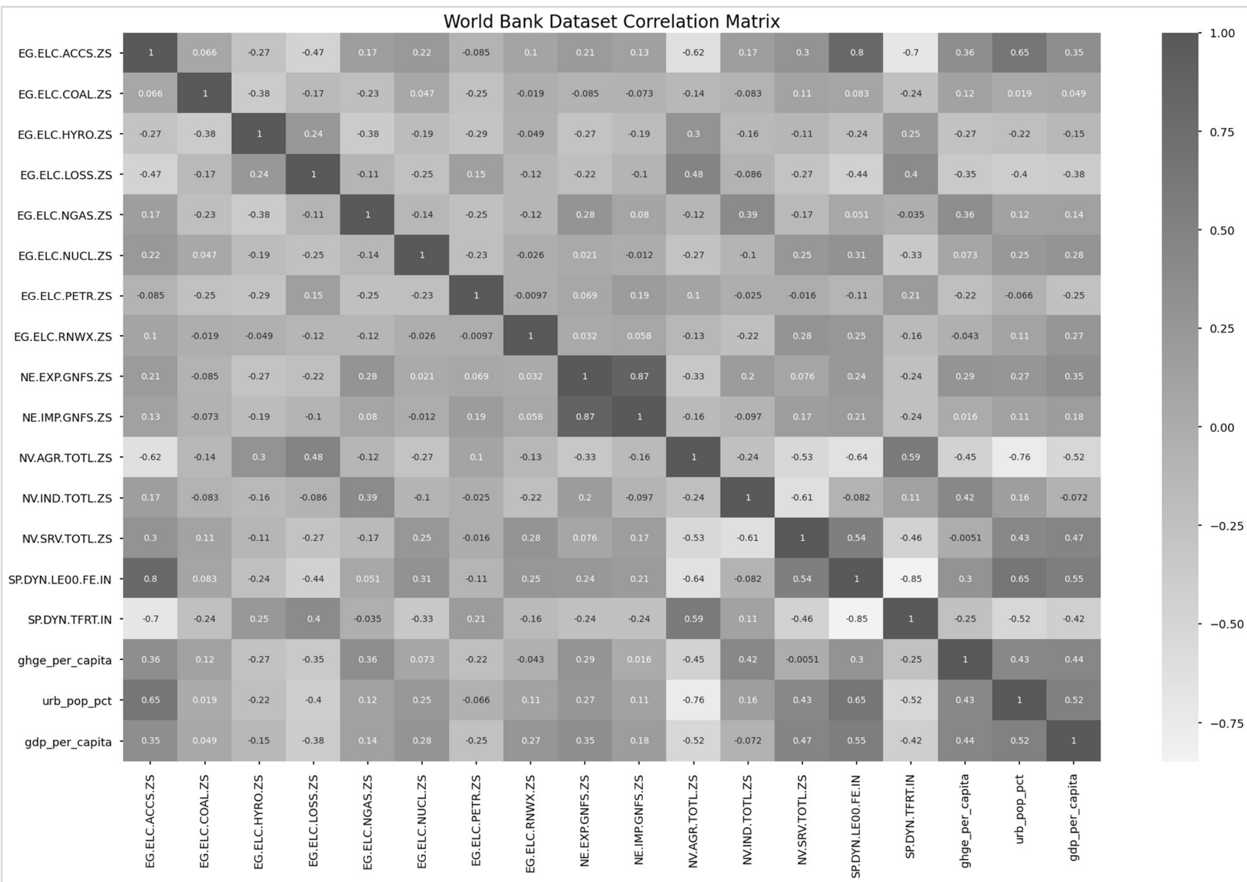
Distribution of Variables



Variables plotted over time (Filtered to USA to provide an example)



Correlation Matrix



SUMMARY OF ANALYSIS

QUESTION 1: WHAT FACTORS DRIVE QUALITY OF LIFE?

Target Variable: GDP per capita (gdp_per_capita)

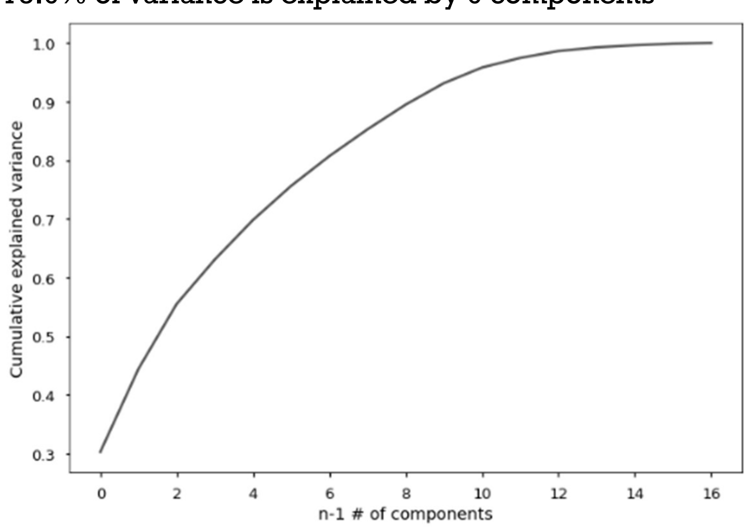
Analysis methods: PCA, k-means, linear regression, and random forest regression (decision trees using bagged ensemble method)

Process: First, I standardized the independent variables used for prediction and then began with PCA. Second, I used k-means clustering with the principal components as inputs. I then split the data set into testing and training groups using 75% for training. I applied linear regression to the training data set with the target variable transformed by the natural log which generated good results on both the training and test sets. The data used for linear regression consisted of the principal component values as continuous variables and the k-means clusters as a categorical variable. Finally, I used Random Forest regression where I ran 1000 decision tree permutations. For the Random Forest I did not use the principal components, but rather the raw independent variables and then the clusters as a categorical variable. I then created an initial model to identify important features which resulted in 13 features (that did not include the clusters) I used for the final model. The final Random Forest method generated excellent accuracy against the test data set however there is evidence of overfitting. The final Random Forest model is selected as the best model due to its accuracy and interpretability.

Conclusions: GDP per capita can be predicted with high accuracy and is increased by improved healthcare (as indicated by female life expectancy at birth), reduced economic contribution from agriculture, increased reliance on renewables for electricity, increased greenhouse gas emissions, and more urbanization.

Principal Component Analysis

75.6% of variance is explained by 6 components



Factor loadings table

	PC1	PC2	PC3	PC4	PC5	PC6
EG.ELC.ACCS.ZS	-0.358509	0.023526	-0.092499	-0.130759	-0.176933	-0.043863
EG.ELC.COAL.ZS	-0.082952	-0.110751	-0.210929	0.711069	0.166622	0.160493
EG.ELC.HYRO.ZS	0.189721	-0.179774	-0.113694	-0.606026	0.259375	-0.055726
EG.ELC.LOSS.ZS	0.275125	-0.020363	0.103546	-0.066080	-0.055392	-0.051698
EG.ELC.NGAS.ZS	-0.088053	0.441073	-0.028588	-0.086830	0.131903	0.269162
EG.ELC.NUCL.ZS	-0.166671	-0.161599	-0.150397	0.085632	0.206640	-0.570113
EG.ELC.PETR.ZS	0.070894	0.005494	0.412703	0.126620	-0.698385	-0.195021
EG.ELC.RNWX.ZS	-0.089288	-0.228786	0.121251	-0.114570	-0.120399	0.698577
NE.EXP.GNFS.ZS	-0.195706	0.266799	0.491553	0.002519	0.313671	-0.033401
NE.IMP.GNFS.ZS	-0.139858	0.092515	0.612277	0.039917	0.318214	-0.070839
NV.AGR.TOTL.ZS	0.373028	-0.027450	0.058174	0.054206	0.130250	0.042473
NV.IND.TOTL.ZS	-0.037800	0.534596	-0.188748	-0.061248	-0.179308	-0.079739
NV.SRV.TOTL.ZS	-0.241121	-0.409280	0.125980	-0.021286	-0.005320	0.059687
SP.DYN.LE00.FE.IN	-0.382823	-0.142543	-0.005065	-0.133192	-0.069714	0.010133
SP.DYN.TFRT.IN	0.357653	0.150016	0.029927	-0.006214	-0.104796	0.022918
ghge_per_capita	-0.227389	0.322939	-0.195763	0.012578	0.043727	0.128470
urb_pop_pct	-0.348737	0.022459	-0.064873	-0.186916	-0.204216	-0.089603

PC1

+ Access to electricity, service based economic activities, increased life expectancy, higher GDP per capita, more urban
– Electric transmission and distribution losses, agricultural economic activities, and increased fertility

PC2

+ Electricity from natural gas, exports, imports, GDP contribution from the industrial sector
– Service based economic activities

PC3

+ Electricity from oil, exports, imports
– GDP contribution from the industrial sector

PC4

+ Electricity from hydroelectric
– Electricity from coal and oil

PC5

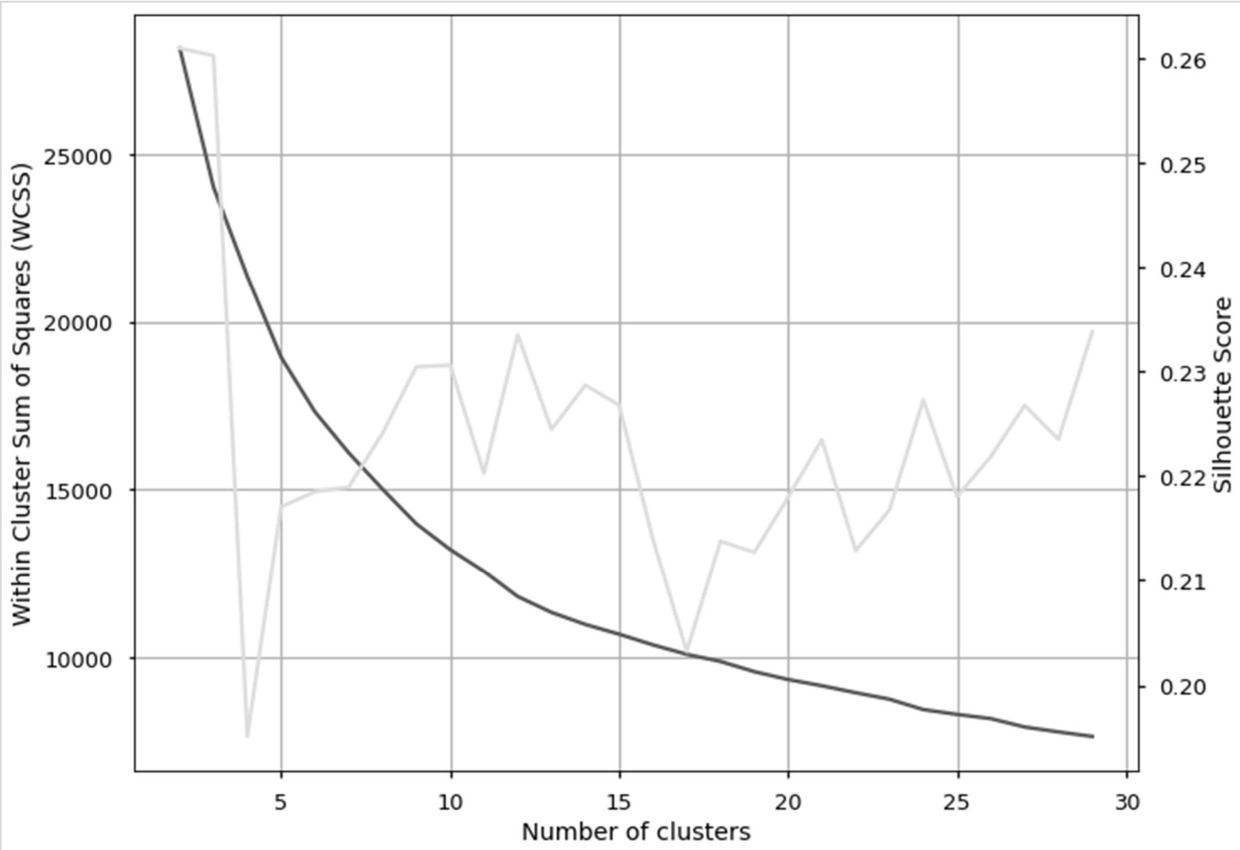
+ Electricity from coal, exports, imports
– Access to electricity, electricity from oil

PC6

+ Electricity from natural gas and renewables
– Electricity from nuclear

k-means Clustering

Selected k = 10 based on "elbow" plot shown below and silhouette score



Linear Regression

```
# Build linear regression model, transforming the target variable in order to meet normality requirement
qual_life_model = ols('np.log(gdp_per_capita) ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + C(Cluster_ID)', train_data).fit()
```

ANOVA

	df	sum_sq	mean_sq	F	PR(>F)
C(Cluster_ID)	9.0	2909.144515	323.238279	741.004484	0.000000e+00
PC1	1.0	674.394432	674.394432	1546.009028	2.221797e-252
PC2	1.0	2.496219	2.496219	5.722433	1.683922e-02
PC3	1.0	23.350074	23.350074	53.528653	3.640998e-13
PC4	1.0	26.636420	26.636420	61.062405	8.768681e-15
PC5	1.0	9.195186	9.195186	21.079415	4.673844e-06
PC6	1.0	3.902516	3.902516	8.946285	2.813530e-03
Residual	2047.0	892.934891	0.436216	NaN	NaN

Summary

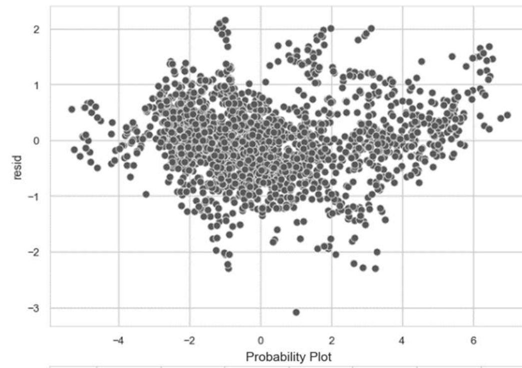
OLS Regression Results

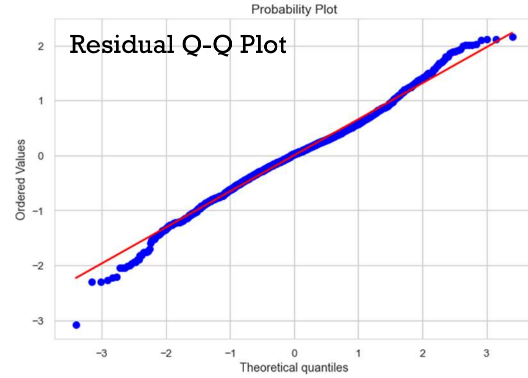
Dep. Variable:	np.log(gdp_per_capita)	R-squared:	0.803
Model:	OLS	Adj. R-squared:	0.802
Method:	Least Squares	F-statistic:	557.7
Date:	Wed, 17 Nov 2021	Prob (F-statistic):	0.00
Time:	12:48:43	Log-Likelihood:	-2063.5
No. Observations:	2063	AIC:	4159.
Df Residuals:	2047	BIC:	4249.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.9295	0.070	126.710	0.000	8.791	9.068
C(Cluster_ID)[T.1]	-0.4332	0.122	-3.562	0.000	-0.672	-0.195
C(Cluster_ID)[T.2]	-0.5976	0.094	-6.332	0.000	-0.783	-0.413
C(Cluster_ID)[T.3]	-0.3981	0.177	-2.249	0.025	-0.745	-0.051
C(Cluster_ID)[T.4]	-0.1789	0.112	-1.603	0.109	-0.398	0.040
C(Cluster_ID)[T.5]	-0.8961	0.092	-9.690	0.000	-1.077	-0.715
C(Cluster_ID)[T.6]	-0.6324	0.086	-7.372	0.000	-0.801	-0.464
C(Cluster_ID)[T.7]	-0.3721	0.125	-2.982	0.003	-0.617	-0.127
C(Cluster_ID)[T.8]	-0.2214	0.180	-1.232	0.218	-0.574	0.131
C(Cluster_ID)[T.9]	-0.8898	0.095	-9.411	0.000	-1.075	-0.704
PC1	-0.5945	0.015	-38.973	0.000	-0.624	-0.565
PC2	-0.0184	0.019	-0.987	0.324	-0.055	0.018
PC3	-0.0945	0.019	-5.092	0.000	-0.131	-0.058
PC4	-0.1756	0.020	-8.701	0.000	-0.215	-0.136
PC5	-0.0963	0.019	-5.179	0.000	-0.133	-0.060
PC6	0.0749	0.025	2.991	0.003	0.026	0.124

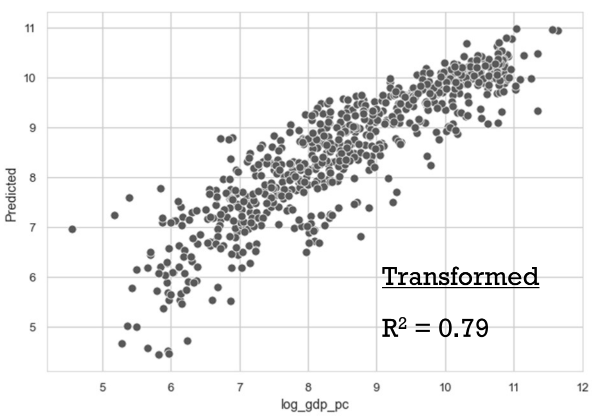
Omnibus:	38.189	Durbin-Watson:	2.043
Prob(Omnibus):	0.000	Jarque-Bera (JB):	73.736
Skew:	-0.062	Prob(JB):	9.74e-17
Kurtosis:	3.918	Cond. No.	45.1

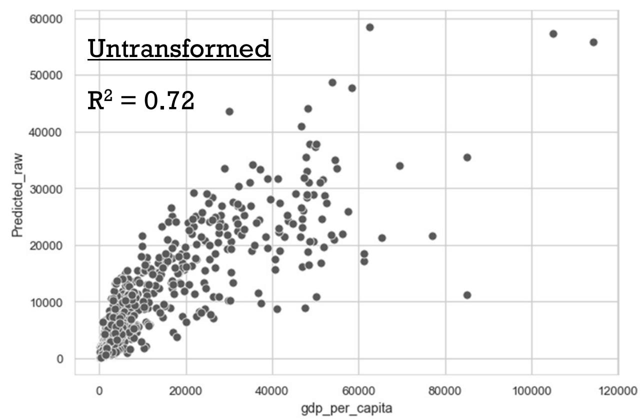
Residual Plot vs PC1





Predicted vs Actual on test data





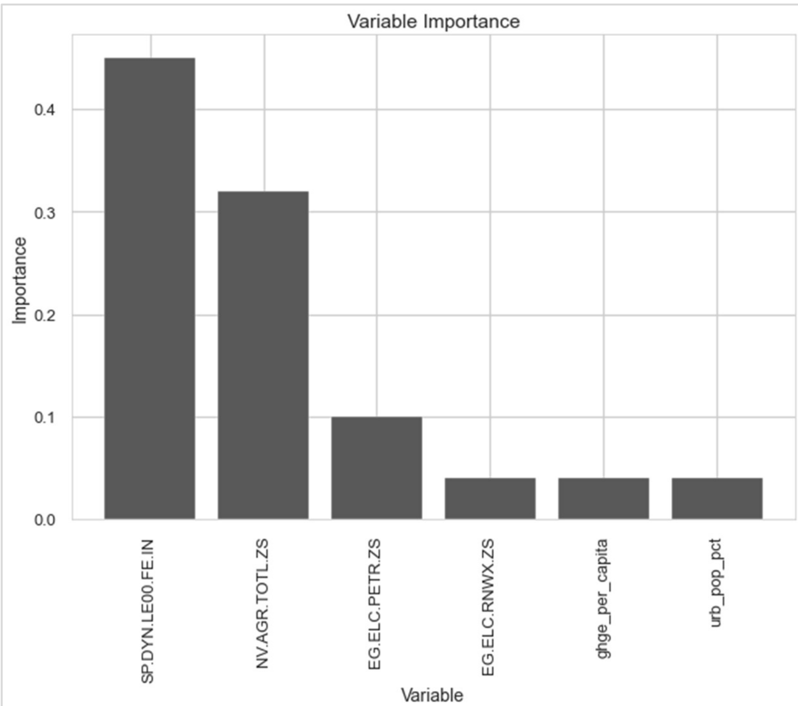
Comments on Regression

- A few of the predictors have high p-values and are not meaningful predictors. These include PC2, cluster 4, and cluster 8
- The residual plots all appear random and normally distributed
- Overfitting does not appear to be an issue comparing the R^2 values on training and testing
- Difficult to interpret given the individual predictors are embedded in principal components and clusters.

Random Forest

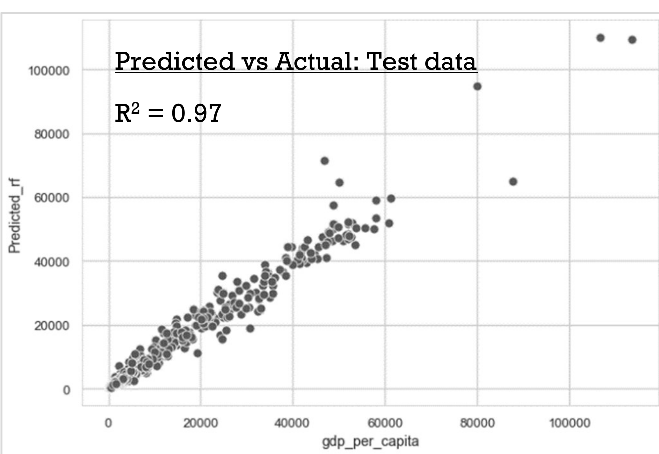
```
# Initiate model with 1000 decision trees
qual_life_model_rf2 = RandomForestRegressor(n_estimators = 1000)
qual_life_model_rf2.fit(X_train, y_train)
```

Feature Importance



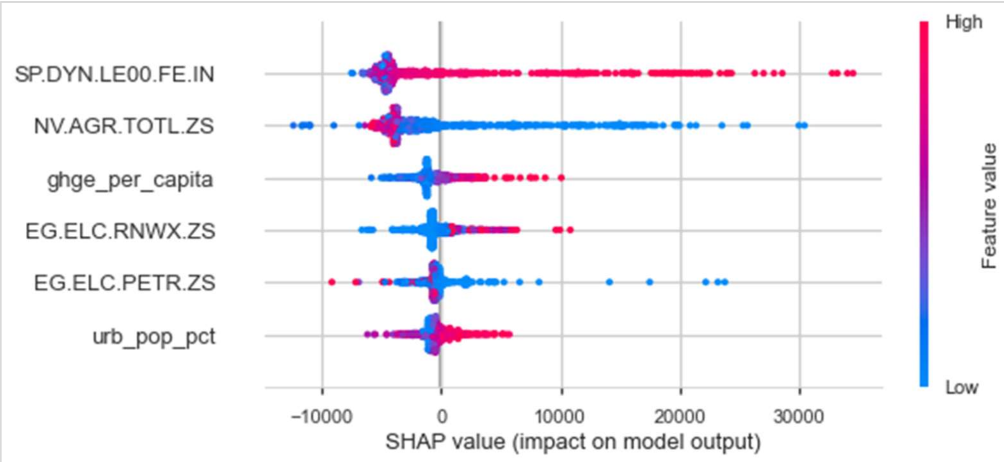
Model Evaluation

	Train	Test
Mean absolute error	559.6	1,278.2
Mean squared error	1,091,880	6,642,358
Root mean squared error	1,044.9	2,577.3



Model Interpretation with Shapley Additive Explanations

- Shapley values represent the average marginal contribution of an instance of a feature among all possible coalitions



Comments on Random Forest Model

- Evidence of overfitting comparing the error values between training and test data
- Model highly accurate against test data using only 6 independent variables
- Preferred model due to high predictive power and ease of interpretation

QUESTION 2: WHAT FACTORS DRIVE GREENHOUSE GAS EMISSIONS?

Target Variable: Greenhouse gas emissions per capita (ghge_per_capita)

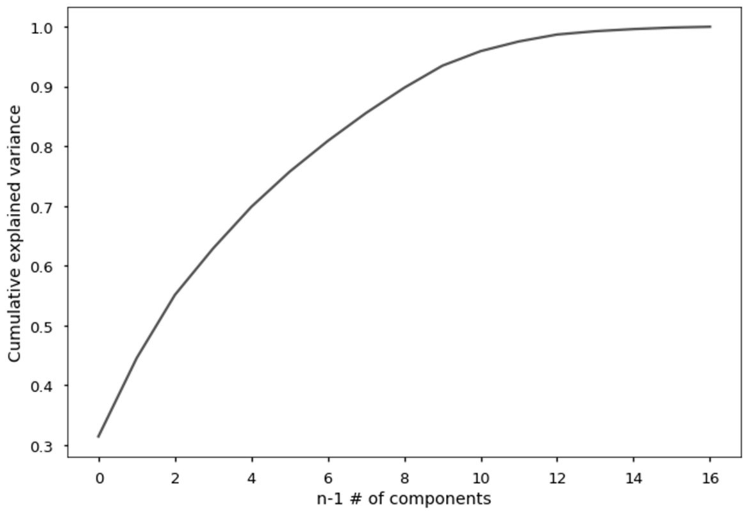
Analysis methods: PCA, k-means, linear regression, and random forest regression

Process: Same as described for question 1

Conclusions: GHGE per capita can be predicted with high accuracy and is increased by higher GDP per capita, higher access to electricity, more urbanization, and more use of fossil fuels for electricity generation.

Principal Component Analysis

75.7% of variance is explained by 6 components



Factor loadings table

	PC1	PC2	PC3	PC4	PC5	PC6
EG.ELC.ACCS.ZS	0.339965	0.065525	-0.161647	0.019417	-0.257646	-0.065084
EG.ELC.COAL.ZS	0.075724	-0.155820	-0.175683	-0.640594	0.346839	0.129881
EG.ELC.HYRO.ZS	-0.170438	-0.241456	-0.023756	0.616232	0.035796	-0.218105
EG.ELC.LOSS.ZS	-0.264792	-0.027999	0.132976	0.057745	-0.072876	-0.053616
EG.ELC.NGAS.ZS	0.066276	0.448780	-0.170237	0.154426	0.173468	0.337609
EG.ELC.NUCL.ZS	0.172988	-0.166984	-0.132988	-0.056189	0.198285	-0.493584
EG.ELC.PETR.ZS	-0.071492	0.108264	0.391475	-0.292817	-0.642392	-0.080036
EG.ELC.RNWX.ZS	0.107874	-0.204045	0.199162	0.127024	-0.084727	0.688909
NE.EXP.GNFS.ZS	0.189238	0.396140	0.403230	0.058378	0.279357	-0.101457
NE.IMP.GNFS.ZS	0.142806	0.255957	0.563113	-0.010895	0.251279	-0.175623
NV.AGR.TOTL.ZS	-0.360023	-0.058168	0.101244	-0.006304	0.144405	0.027497
NV.IND.TOTL.ZS	-0.000158	0.502099	-0.356545	0.032257	-0.155469	-0.021810
NV.SRV.TOTL.ZS	0.264194	-0.359284	0.232075	0.008539	-0.024647	0.047934
SP.DYN.LE00.FE.IN	0.381606	-0.089601	-0.012715	0.063592	-0.144818	-0.023748
SP.DYN.TFRT.IN	-0.352320	0.100199	0.039193	0.053756	-0.029372	0.099570
gdp_per_capita	0.297428	-0.048189	0.041781	0.227203	0.216614	0.174260
urb_pop_pct	0.338564	0.047903	-0.103959	0.118618	-0.248811	-0.065817

PC1

- \+ Access to electricity, service based economic activities, increased life expectancy, higher GDP per capita, more urban
- – Electric transmission and distribution losses, agricultural economic activities, and increased fertility

PC2

+ Electricity from natural gas, exports, imports, GDP contribution from the industrial sector
– Service based economic activities

PC3

+ Electricity from oil, exports, imports
– GDP contribution from the industrial sector

PC4

+ Electricity from hydroelectric
– Electricity from coal and oil

PC5

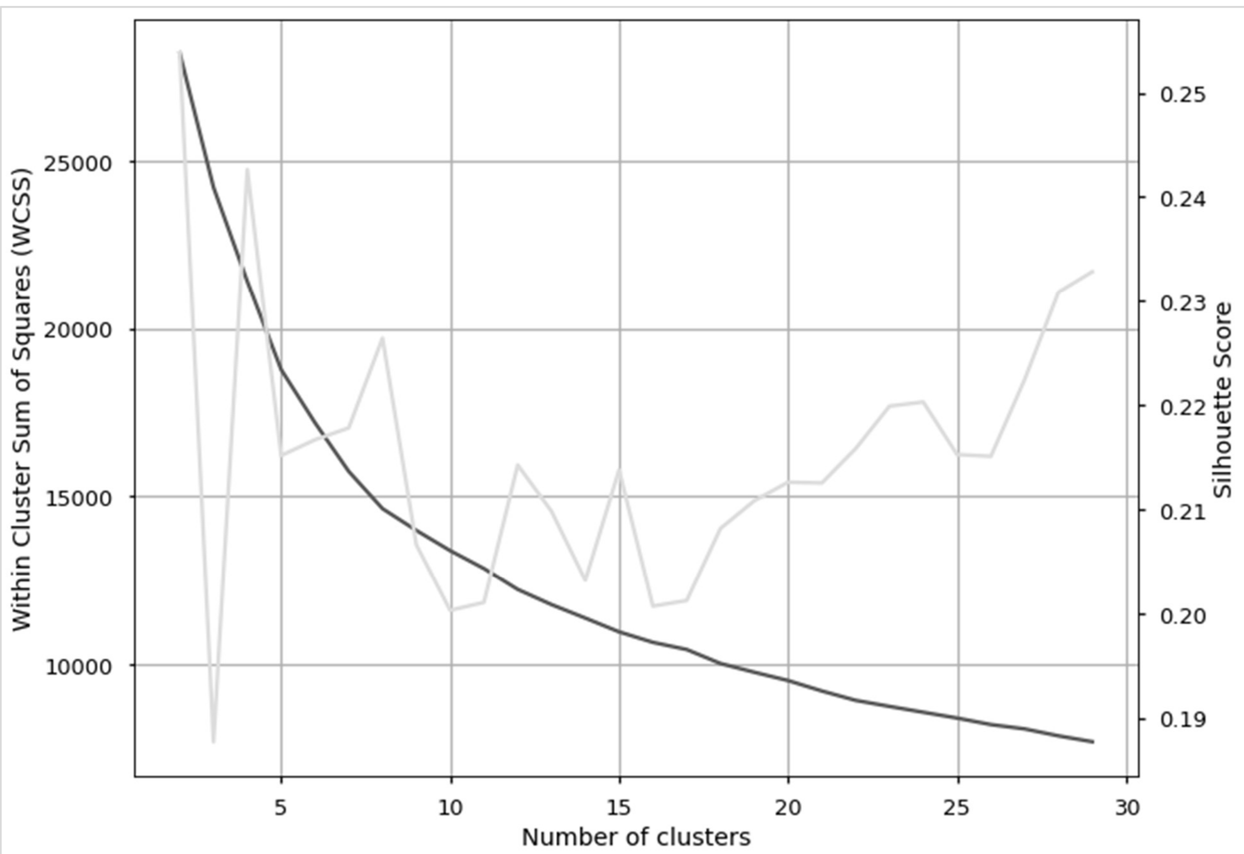
+ Electricity from coal, exports, imports
– Access to electricity, electricity from oil

PC6

+ Electricity from natural gas and renewables
– Electricity from nuclear

k-means Clustering

Selected k = 10 based on "elbow" plot shown below



Linear Regression

```
# Build linear regression model, transforming the target variable in order to meet normality requirement
ghge_model = smf.ols(formula = 'np.log(ghge_per_capita) ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + C(Cluster_ID)', data = train_data).fit()
```

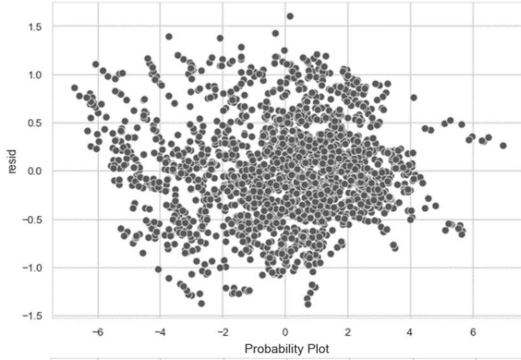
ANOVA

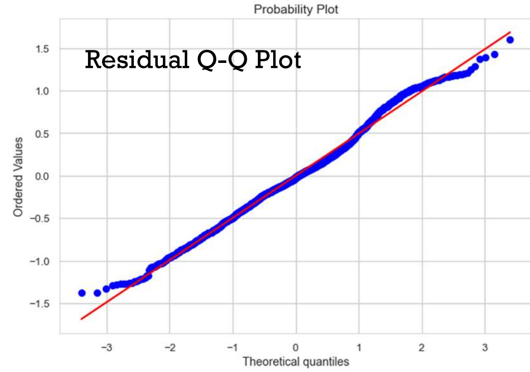
	df	sum_sq	mean_sq	F	PR(>F)
C(Cluster_ID)	9.0	853.419372	94.824375	381.116071	0.000000e+00
PC1	1.0	116.910203	116.910203	469.882953	5.697155e-94
PC2	1.0	43.709442	43.709442	175.676041	1.571550e-38
PC3	1.0	50.853500	50.853500	204.389286	2.891711e-44
PC4	1.0	0.063966	0.063966	0.257090	6.121810e-01
PC5	1.0	1.876505	1.876505	7.542008	6.080431e-03
PC6	1.0	1.332894	1.332894	5.357137	2.073615e-02
Residual	2047.0	509.308082	0.248807	NaN	NaN

Summary

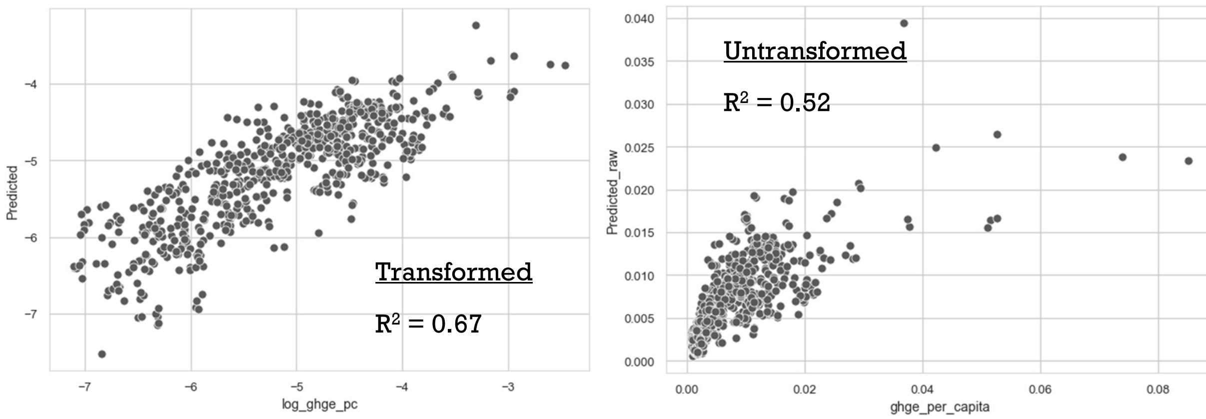
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     np.log(ghge_per_capita)   R-squared:                  0.677
Model:                                 OLS   Adj. R-squared:             0.675
Method:                      Least Squares   F-statistic:                286.2
Date:                     Fri, 12 Nov 2021   Prob (F-statistic):          0.00
Time:                             14:02:13   Log-Likelihood:           -1484.3
No. Observations:                     2063   AIC:                        3001.
Df Residuals:                         2047   BIC:                        3091.
Df Model:                               15
Covariance Type:                 nonrobust
==============================================================================
                      coef    std err          t      P>|t|     [0.025    0.975]
------------------------------------------------------------------------------
Intercept          -5.0632      0.055    -92.001      0.000     -5.171    -4.955
C(Cluster_ID)[T.1] -0.1119      0.072     -1.549      0.122     -0.254     0.030
C(Cluster_ID)[T.2]  0.0432      0.077      0.560      0.576     -0.108     0.194
C(Cluster_ID)[T.3]  0.0836      0.084      0.998      0.318     -0.081     0.248
C(Cluster_ID)[T.4] -0.3439      0.073     -4.689      0.000     -0.488    -0.200
C(Cluster_ID)[T.5]  0.0051      0.142      0.036      0.971     -0.274     0.284
C(Cluster_ID)[T.6] -0.3194      0.078     -4.088      0.000     -0.473    -0.166
C(Cluster_ID)[T.7] -0.2293      0.083     -2.777      0.006     -0.391    -0.067
C(Cluster_ID)[T.8] -0.2546      0.086     -2.952      0.003     -0.424    -0.085
C(Cluster_ID)[T.9]  0.1552      0.061      2.564      0.010      0.036     0.274
PC1                 0.2651      0.012     22.102      0.000      0.242     0.289
PC2                 0.1749      0.014     12.519      0.000      0.148     0.202
PC3                -0.1954      0.014    -13.883      0.000     -0.223    -0.168
PC4                 0.0152      0.015      1.041      0.298     -0.013     0.044
PC5                -0.0352      0.015     -2.421      0.016     -0.064    -0.007
PC6                -0.0361      0.016     -2.315      0.021     -0.067    -0.006
==============================================================================
Omnibus:                    11.785   Durbin-Watson:                 2.023
Prob(Omnibus):               0.003   Jarque-Bera (JB):             11.949
Skew:                        0.185   Prob(JB):                    0.00254
Kurtosis:                    2.953   Cond. No.                       43.5
==============================================================================
```

Residual Plot vs PC1





Predicted vs Actual on test data



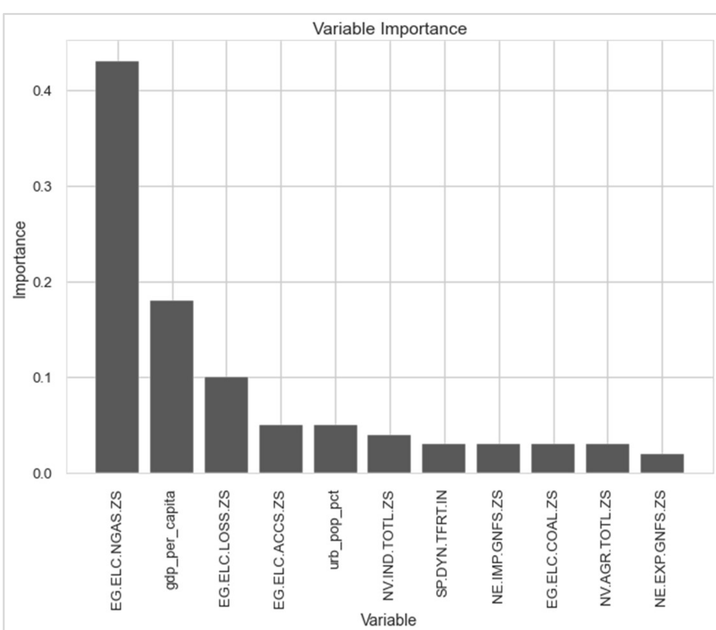
Comments on Regression

- A few of the predictors have high p-values and are not meaningful predictors. These include PC4, cluster 1, cluster 2, cluster 3, and cluster 5
- The residual plots all appear random and normally distributed
- Overfitting does not appear to be an issue comparing the R^2 values on training and testing
- Difficult to interpret given the individual predictors are embedded in principal components and clusters.

Random Forest

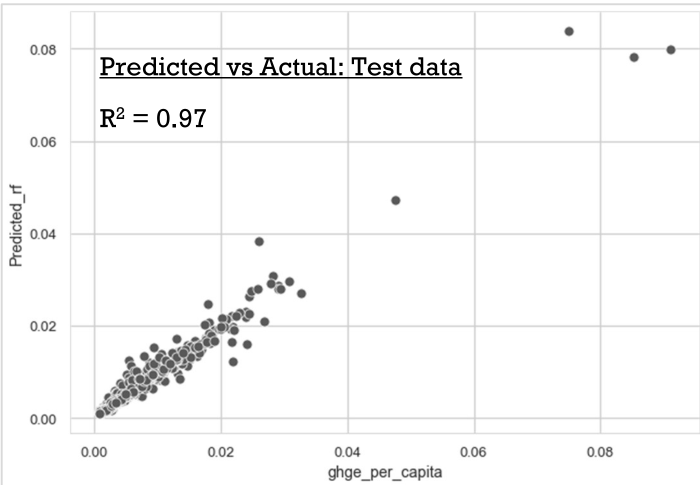
```
# Initiate model with 1000 decision trees
ghge_model_rf2 = RandomForestRegressor(n_estimators = 1000)
ghge_model_rf2.fit(X_train, y_train)
```

Feature Importance

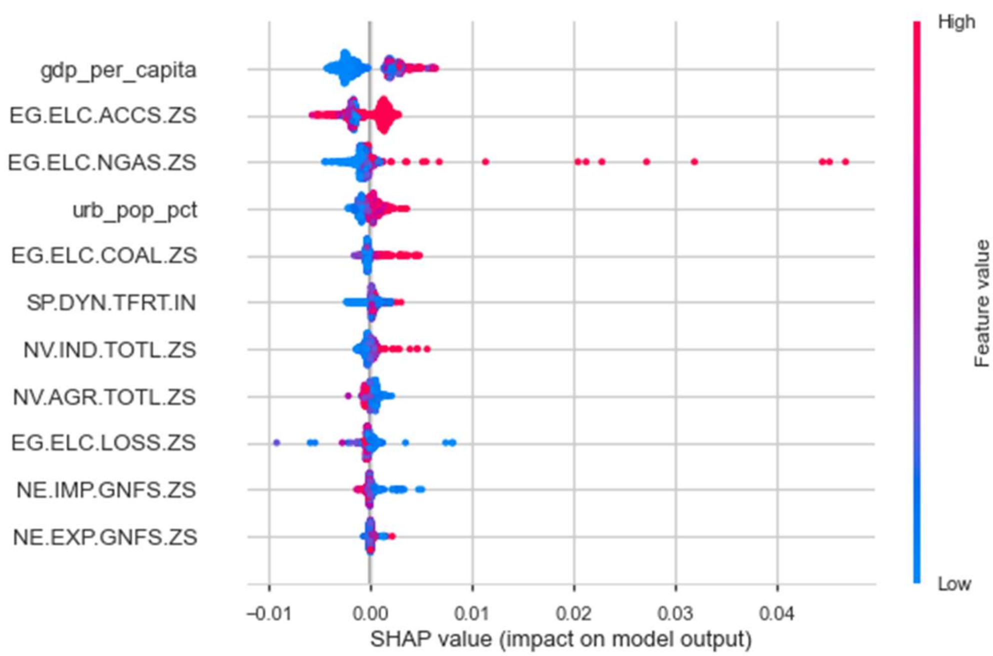


Model Evaluation

	Train	Test
Mean absolute error	0.0003	0.0008
Mean squared error	0.0	0.0
Root mean squared error	0.0007	0.0015



Model Interpretation with Shapley Additive Explanations



Comments on Random Forest Model

- Evidence of overfitting comparing the error values between training and test data
- Model highly accurate against test data using 11 independent variables
- Preferred model due to high predictive power and ease of interpretation

QUESTION 3: WHAT CAN WE DO TO CONTINUE TO IMPROVE HUMAN FLOURISHING WHILE REDUCING GREENHOUSE GAS EMISSIONS?

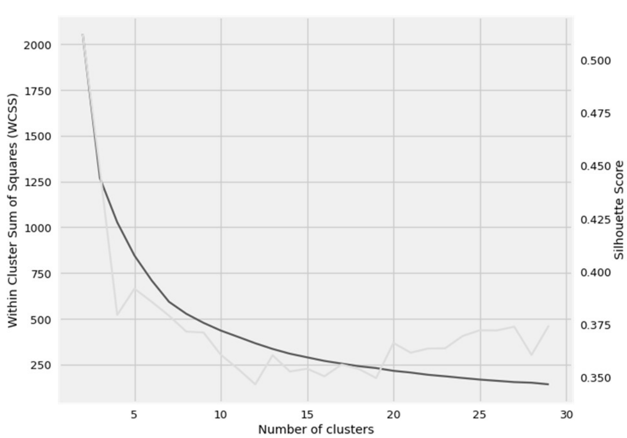
Target Variable: Categorical variable that represents groups binned by GDP per capita and greenhouse gas emissions per capita. For example high GDP per capita, high greenhouse gas emissions and high GDP per capita and low greenhouse gas emissions.

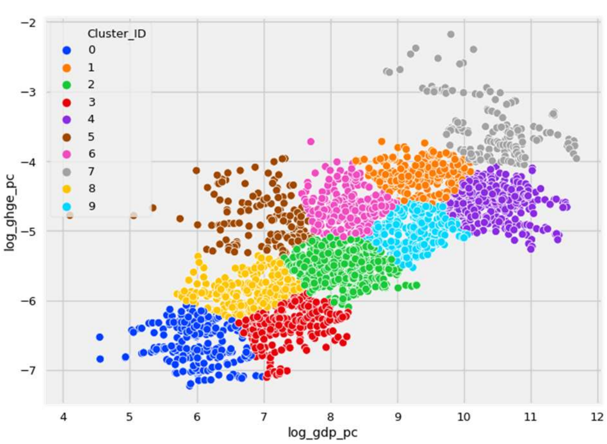
Analysis methods: k-means to create target variable, Random Forest classification, kNN

Process: First I created 10 clusters using k-means analysis with standardized ln(GDP per capita) and standardized ln(GHGE per capita) as independent variables. The resulting 10 clusters became the target variable for the analysis associated with the 3rd question. Using this target variable, I ran a Random Forest classification model, and a kNN classification model. The Random Forest model was run twice. First, to determine the most important features, and then second only including the most important features. The kNN model used the standardized values of the important features as derived from the first Random Forest model. The Random Forest and kNN models both resulted in perfect predictions using training data and roughly 90% accuracy on test data indicating issues with overfitting, however highly accurate nonetheless.

Conclusions: Marginal improvements can achieved in both GDP per capita and GHGE per capita by increasing the amount of energy we produce from low or no carbon sources such as renewable or nuclear sources.

Producing the target variable using k-means clustering

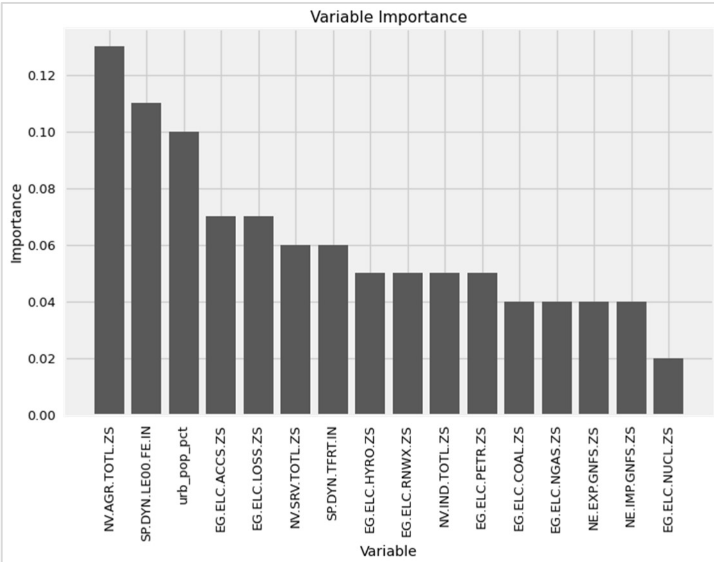




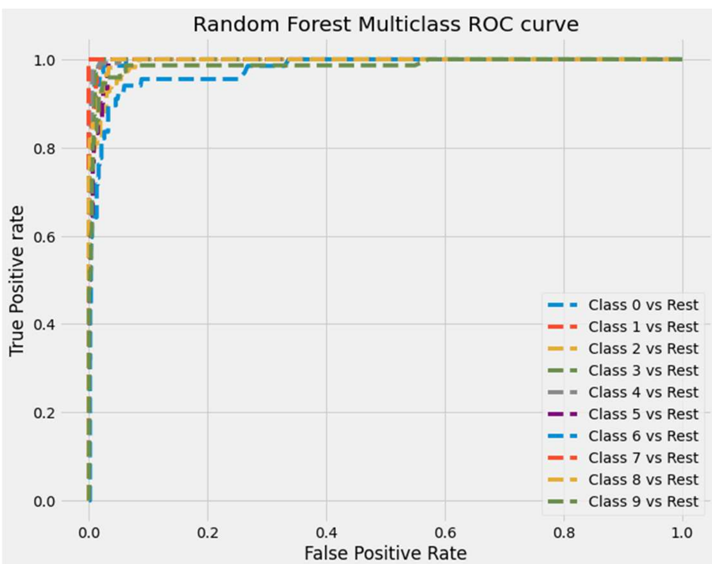
Random Forest Classification

```
# Initiate model with 1000 decision trees
ideal_model_rf2 = RandomForestClassifier(n_estimators = 1000)
ideal_model_rf2.fit(X2_train, y2_train)
```

Feature Importance



ROC Curve (AUC score = 0.994)



Confusion Matrix on Test Data (Accuracy = 0.89)

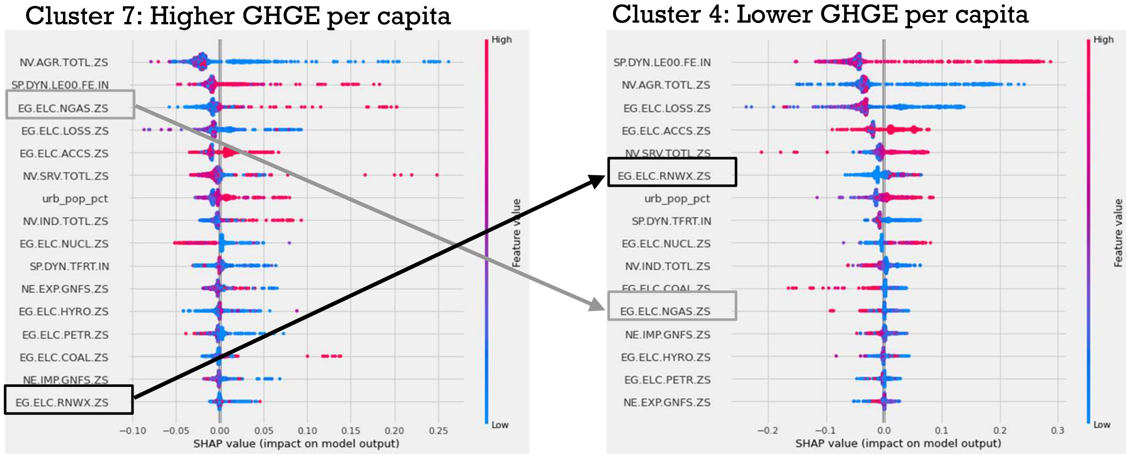
	0	1	2	3	4	5	6	7	8	9
0	67	0	0	1	0	0	0	0	1	0
1	0	50	0	0	2	0	3	0	0	0
2	0	0	96	2	0	2	2	0	3	5
3	6	0	3	61	0	0	0	0	2	0
4	0	2	0	0	116	0	0	0	0	0
5	0	0	3	0	0	25	2	0	0	0
6	0	2	4	0	0	5	51	0	0	5
7	0	1	0	0	0	0	0	38	0	0
8	5	0	3	1	0	0	0	0	46	0
9	0	1	2	0	3	0	2	0	0	65

Model Interpretation with Shapley Additive Explanations

- Comparison of clusters 7 and 4 (highest GDP per capita clusters)
- Charts sorted by order of feature importance

Cluster 7: Higher GHGE per capita

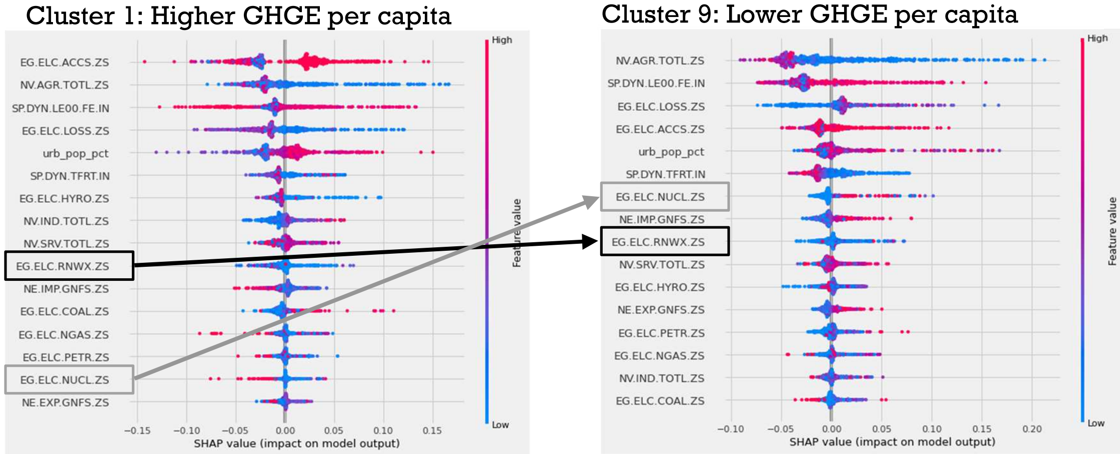
Cluster 4: Lower GHGE per capita



- Comparison of clusters 1 and 9 (2nd highest GDP per capita)

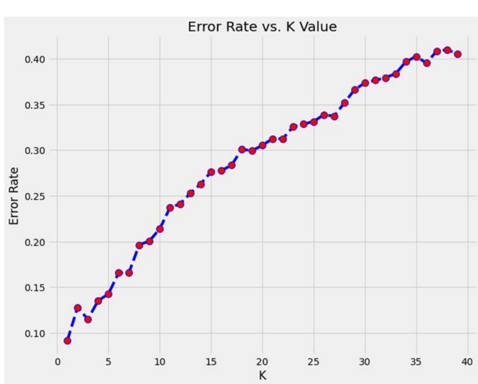
Cluster 1: Higher GHGE per capita

Cluster 9: Lower GHGE per capita



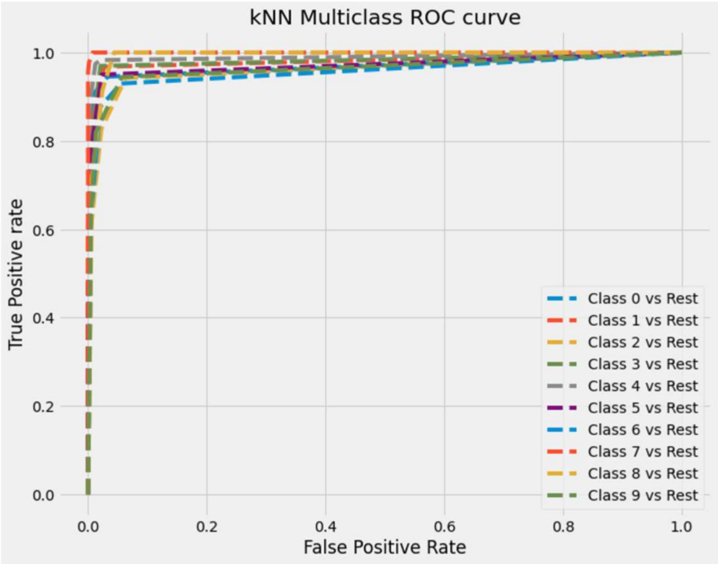
kNN Classification

Ran scenarios against different k-values to determine optimal k for modeling. Used k = 3.



```
# Initiate model with k = 3
ideal_model_knn = KNeighborsClassifier(n_neighbors = 3)
ideal_model_knn.fit(X_train, y_train)
```

ROC Curve (AUC = 0.976)



Confusion Matrix on Test Data (Accuracy = 0.89)

	0	1	2	3	4	5	6	7	8	9
0	50	0	0	4	0	0	0	0	1	0
1	0	53	0	0	5	0	3	0	0	1
2	0	0	88	0	0	0	0	0	12	4
3	6	0	2	56	0	0	0	0	2	0
4	0	0	0	0	113	0	0	1	0	2
5	0	0	6	0	0	31	3	0	0	0
6	0	1	3	0	0	2	50	0	0	1
7	0	1	0	0	0	0	0	44	0	0
8	1	0	2	2	0	1	0	0	65	0
9	0	1	3	0	2	0	7	0	0	59

CONCLUSIONS

This modeling exercise was valuable for learning the various modeling approaches while using real data that required significant cleaning and preparation. The problem that I am attempting to address with this analysis is much too complex for the lack of granularity associated with this data set and thus would not be deployed to solve a real problem. However, the analysis supports conclusions that all of us now find "common sense". Addressing climate change must be done carefully such that we do not cause harm. It is logical that draconian measures to mitigate climate change would impact the most vulnerable in our society. We must find ways to maintain or advance human flourishing while mitigating greenhouse gas emissions. The most logical course as supported by this analysis is to switch to low or no carbon fuels as our primary energy sources.