In [ ]:

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.linear_model import Lasso

import matplotlib.pyplot as plt
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import cross_val_score
```

In [ ]:

```
df = pd.read_csv('inequality_class_final.csv')
df.info
```

Out[ ]:

```
<bound method DataFrame.info of                          country  federalism_GT
id  \
0                         Mexico           2.0      NaN
1                       Suriname           0.0      NaN
2                         Sweden           0.0  17568.0
3                    Switzerland           2.0  17777.0
4                          Ghana           0.0      NaN
..                           ...           ...      ...
200                      Vanuatu           1.0      NaN
201        United Arab Emirates           2.0      NaN
202                Vatican City           0.0      NaN
203  Palestine (British Mandate)           NaN      NaN
204                      Hungary           0.0   8600.0

                  region_wb          gdp  statehiste1500_02n  origtime2  \
0                       NaN          NaN            0.311557    15000.0
1                       NaN          NaN                 NaN        NaN
2    Europe and Central Asia  49296.81030            0.124440     8000.0
3    Europe and Central Asia  64943.53591            0.358282    45000.0
4                       NaN          NaN            0.082024   135000.0
..                      ...          ...                 ...        ...
200                     NaN          NaN                 NaN        NaN
201                     NaN          NaN                 NaN        NaN
202                     NaN          NaN                 NaN        NaN
203                     NaN          NaN                 NaN        NaN
204  Europe and Central Asia  25795.26227            0.287663    45000.0

          eleva  avg_temp  Maddison_gdppc_1990_estimate_ln  ...  \
0    1076.740352      21.6                         9.079312  ...
1     184.250987      26.0                         8.904918  ...
2     359.346328       3.6                        10.248469  ...
3    1317.581088       7.0                        10.436217  ...
4     208.329245      27.6                         7.401412  ...
..           ...       ...                              ...  ...
200   324.979599      24.6                         7.850108  ...
201   138.717195      27.7                        10.400395  ...
202          NaN       NaN                              NaN  ...
203          NaN       NaN                              NaN  ...
204   174.981066      11.6                         9.341803  ...
```

	sd_emeanclip	Urbanpopulationoftotalpop	dist2suitable_km_new	\
0	0.776437	76.920	191207.980	
1	0.133851	66.547	175987.630	
2	0.248912	84.588	108188.940	
3	0.625552	73.530	358067.160	
4	0.074167	48.669	351368.590	
..	...	...	...	
200	NaN	23.687	33445.219	
201	0.135800	83.023	64342.836	
202	NaN	NaN	NaN	
203	NaN	NaN	NaN	
204	0.106408	67.421	456766.720	

	Ygini_disp	Fixedtelephonesubscriptionsp	Employmentinagricultureof	\
0	46.400002	18.318046	13.790000	
1	45.799999	16.041285	8.460000	
2	25.000000	60.024409	2.250000	
3	29.900000	65.373435	4.000000	
4	43.000000	1.639566	53.939999	
..	...	...	...	
200	40.400002	4.018736	66.070000	
201	NaN	22.460032	4.390000	
202	NaN	NaN	NaN	
203	NaN	NaN	NaN	
204	27.000000	32.428040	4.590000	

	Accesstoelectricityofpopu	pln_sxHr_mean	agyears_ext	popd_1500AD
0	98.110626	2626.828031	4100.0	12.078997
1	94.783394	6696.253076	1500.0	0.057579
2	100.000000	0.000000	5500.0	0.842480
3	100.000000	877.940256	5500.0	19.928775
4	56.975372	6889.873359	3500.0	7.670432
..	...	...	...	...
200	33.624620	NaN	NaN	NaN
201	100.000000	0.000000	7600.0	0.315265
202	NaN	NaN	8000.0	NaN
203	NaN	NaN	10500.0	NaN
204	100.000000	5454.954688	7400.0	10.708198

go ahead and drop variables that are not needed - name and region are not quantitative and we dont necessarily want to use each country to map to its respective income inequality

Country name, regional dummy,