Project Overview:
I attempt to make a socio-economic index based on unsupervised learning algorithms. Instead of relying on weighted averages that currently used indices use, I aim to let the data explain the variation among districts. I plan to use two methods of ranking (i) ordinal and (ii) cardinal, where I rank clusters by using HDI as the ground truth for the former, and then subsequently attach numeric scores using the second approach. Lastly, I validate my current index through some of the checks mentioned in the project prompt.

Variable Selection:

I do not want to follow the usual weighted average methods prevalent with MDPI, but rather let the weights be figured out on their own through clustering mechanisms. I want to additionally manufacture more variables, from manipulation current data, to manage a more holistic factual representation.

1. *Gravity equations and distance from other areas: I want to make a variable that incorporates distances from each district, and essentially the algorithm figures out that some minimized distances from specific centers are an indicator of the socio-economic environment. I wish to primarily associate the same with distances from hubs of public goods, institutions, and per capita income measures.* **(?)**
2. Ageing population: If a district is aging significantly, it will have some downward trend on its socioeconomic index. Try to imagine this as a trend for future years, and can be incorporated as such. That's why birth and death rates are both added onto our large variable dataset.

Removed Variables in NFHS:

1. Similarity check for the list of removed variable names
2. *Unmet need for spacing* removed because it is already incorporated in *unmet family planning needs*, and same logic for a lot of these variables
3. *Children born at home who were taken to a health facility, Births in a private health facility that were delivered by caesarean section* removed due to missing values
4. *Institutional births in private facility (in the 5 years before the survey) (%)* removed because of linearity between public and total institutional births
5. Removed specific vaccination figures because of overfitting concerns
6. Segregation between diets of breastfeeding and non-breastfeeding children removed due to overfitting constraints and possible linearity among variables
7. Removed based on Spearman correlation –
    a. Women (age 1549) who are literate4 (%)
    b. Mothers who received postnatal care from a doctor/nurse/LHV/ANM/midwife/other health personnel within 2 days of delivery (for last birth in the 5 years before the survey) (%)
    c. Institutional births (in the 5 years before the survey) (%)'
    d. 'Female population age 6 years and above who ever attended school (%)

e. Births attended by skilled health personnel (in the 5 years before the survey)10 (%)
f. Children who received postnatal care from a doctor/nurse/LHV/ANM/midwife/ other health personnel within 2 days of delivery (for last birth in the 5 years before the survey)

Use HDI as ground truth for understanding training and test performance.

Ranking Approaches:

1. **Ordinal:** use k-means and then HDI as ground truth to cluster objects, experiment with different cluster counts, it is more beneficial for policy because rather than putting some fictional numbers onto it, we compare across similar levels of development
2. **Cardinal:** use gradients and vectors to identify the general direction of improvement, kind of like generating a heat map of improvement based on training data, and maybe test too, and then projecting units onto it to identify a score based on the gradient distance (NOT PROJECTION DISTANCE) and then arrive at a score