**VARIABLE SELECTION**

WHY NOT CONSIDERED?

ID,XCOO,YCOO are used for plotting and identifying in map but not for distribution analysis and clustering .

Silver,and gold do not contribute much to pollution even though one of them is similar normal distribution.

Beryllium ,Molybdenum ,Sodium and cobalt are considered a pollutant when it is present in the environment in high concentrations and can cause harmful effects on human health and the environment.So beryllium is not in high frequencies and not well spread,and not high ,this indicates that it is not a contributor to pollution here in the data aspect.

Potassium is not typically considered a pollutant in the context of environmental pollution. It is a naturally occurring element and is a key nutrient for plant and animal life.

Boron can be pollutants in certain concentrations. High levels of Boron can be toxic to plants and animals,.So it is not considered due to its thin spread and frequency.

WHY CONSIDERED?

Barium can be pollutants in certain concentrations .Barium can accumulate in soil and water and have harmful effects on humans and animals if ingested in large quantities.So Barium has wide distribution spread and is considered.

The presence of LOI and pH levels in water can have an impact on the solubility and mobility of specific pollutants. Furthermore, electrical conductivity can be an essential indicator of water salinity, which can affect the survival and growth of aquatic organisms.And they do have wide distribution and are considered .

"Pb" "Rb" "S" "Sb" "Sc" "Si" "Sr" "Th" "Tl" "U" "V" "Y" "As" "Ba" "Ca" "Cd" "Cr" "Hg" "La" "Mg" "Mn" "Ni" "P" have either high frequency in distribution or normal distribution or with a combination of both.And they all belong to pollutant family if in high concentrations

"Zn" "C" "H" "N" contribute both to environment optimistically and harmfully if there is a variation change .So they are added too .

**OUTLIER HANDLING AND TRANSFORMATION**

It is common actually to find outliers in real life elemental distribution dataset.So log transformation needs to be applied to converted the distributions to a near normal one

Generally it is not compulsory for data to be normally distributed for clustering but skewed data to symmetric data is beneficial.

If the original data follows a log-normal distribution or approximately so, then the log-transformed data follows a normal or near normal distribution.

So lets apply log transformation to all variables of work\_data because some features have skewed distributions

**CLUSTERING:**

**HEIRARCHIAL CLUSTERING**

