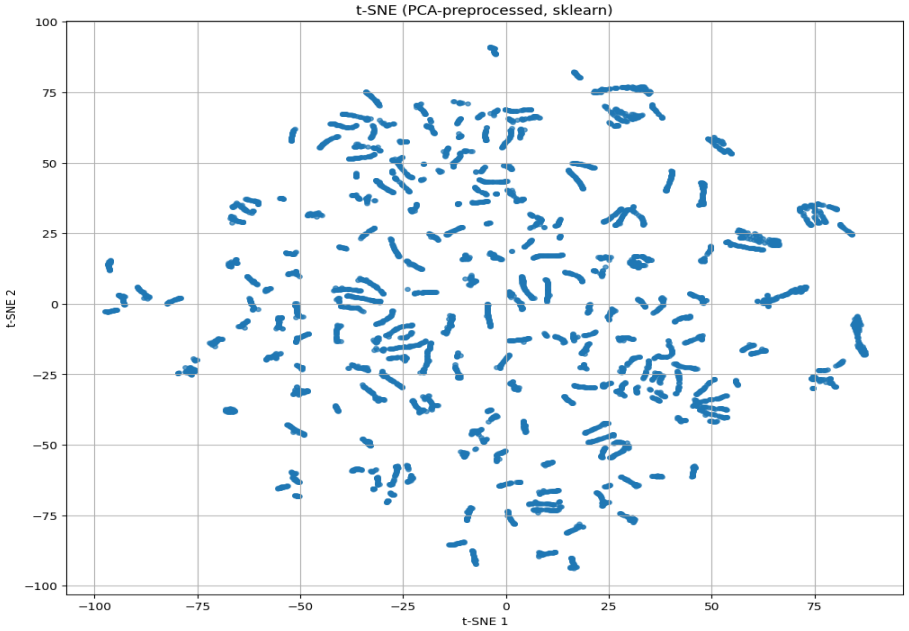
1. Data Preprocessing:

We processed our data and cleaned, using linear imputation and removing rows and columns with too many NaN values. Data was also normalized so that our distances would be smaller and easier to compute. Removing rows and columns at the thresholds set in the notebook lead to our data shrinking from around 17000 vectors/rows/country-year pairs to 6364, and our columns/indicators/dimensions from 1500 dimensions to 571. We achived a good balance of coverage and completeness. Below is a visual representation of our cleaned data using PCA and TSNE.



2. K-Means Clustering and Initialization:

We made our K-Means clustering class from first principles and included methods like K++ initialization. This was the standard functions and K-Means class taught to us in lectures.

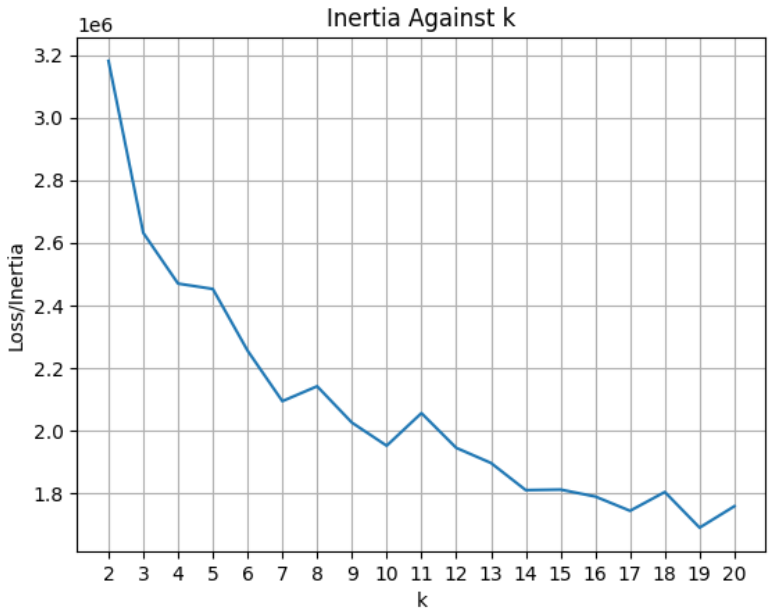
3. Convergence Criteria:

We used four different convergence criteria. We converged either by max iterations, centroid change, inertia change between the current and previous value and average inertia change over a window or n values. All yielded the same result as the centroids would always converge first to a change in distance compared to the previous iteration of 0. Hence throughout the rest of the notebook and assignment I use convergence by max iterations as it has a stopping criterion, namely if centroid change is 0, to prevent unnecessary computation.

4. Determining the Optimal K:

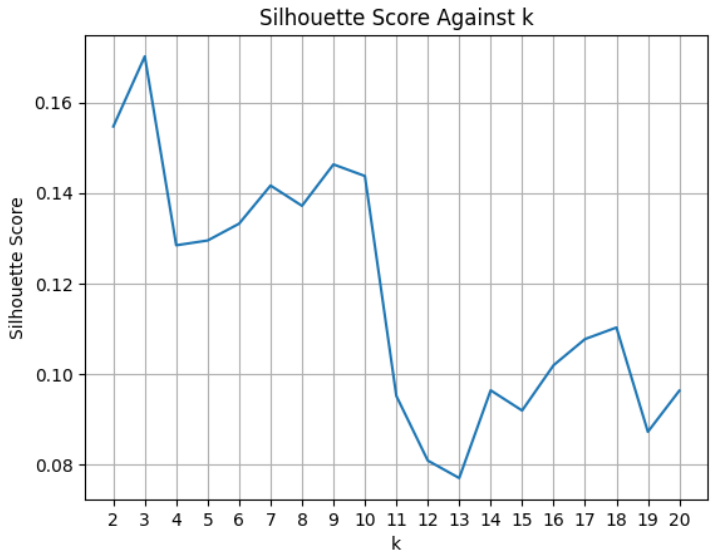
The optimal K will be determined by using a mix of the best or a relatively good silhouette score and by using the elbow method.

The elbow method gave us the following graph.



In order to compute the silhoutte score using sci-kit's built in silhoutte score, we had to make a labels function. Which takes the clusters array and assigns each value in a new array of size X (our data) to the corresponding cluster out of our k clusters, it tells us which cluster the i'th point is in. For example, [1, 2, ..., 3] means point 1 in X is in cluster 1, 2 is in cluster 2 and N is in cluster 3.

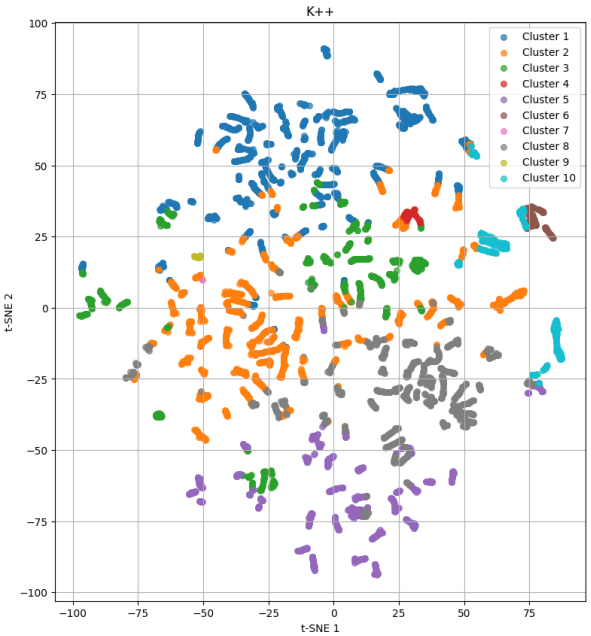
The silhoutte scores against K are below

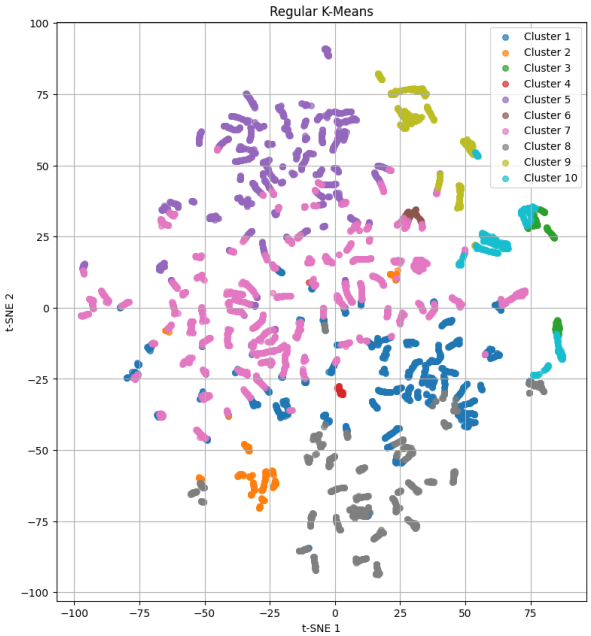


We do not want k to be too small as then we don't discover good patterns and unrelated things get clustered. We also don't want K to be too big as then we get overfitting.

According to the above data, K=10 is a reasonably good candidate for an optimal k as the loss starts to flatten after K=10 and the silhoutte score is relatively high compare to that of the other K values. We will be using K=10 from now.

Using this optimal K, we also compared K++ and regular K-Means, below are our results.





K++ Loss: 2052010.533849431

K++ Silhouette score: 0.08668216381120003

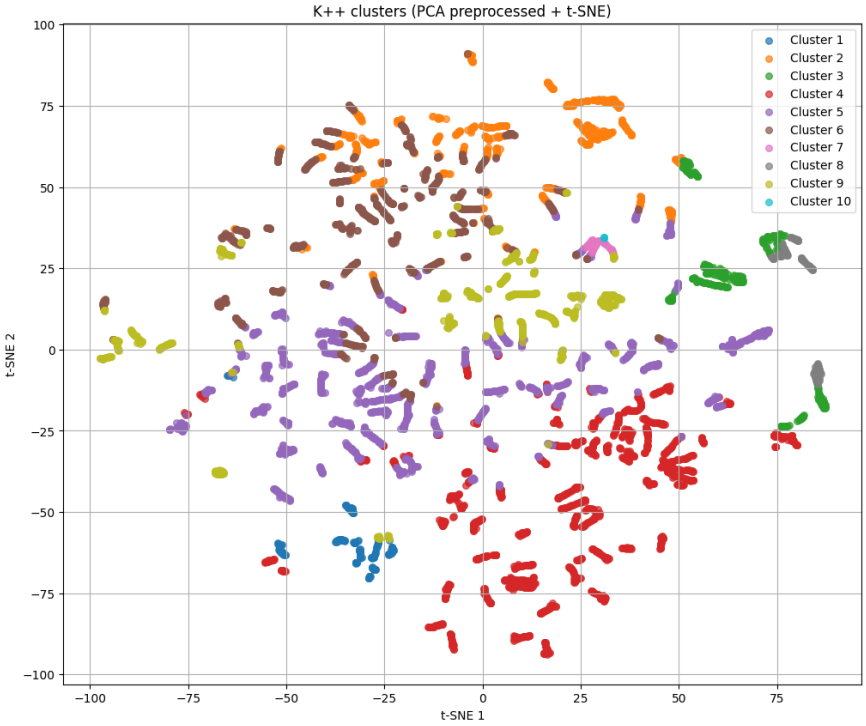
Loss: 2029175.8493517677

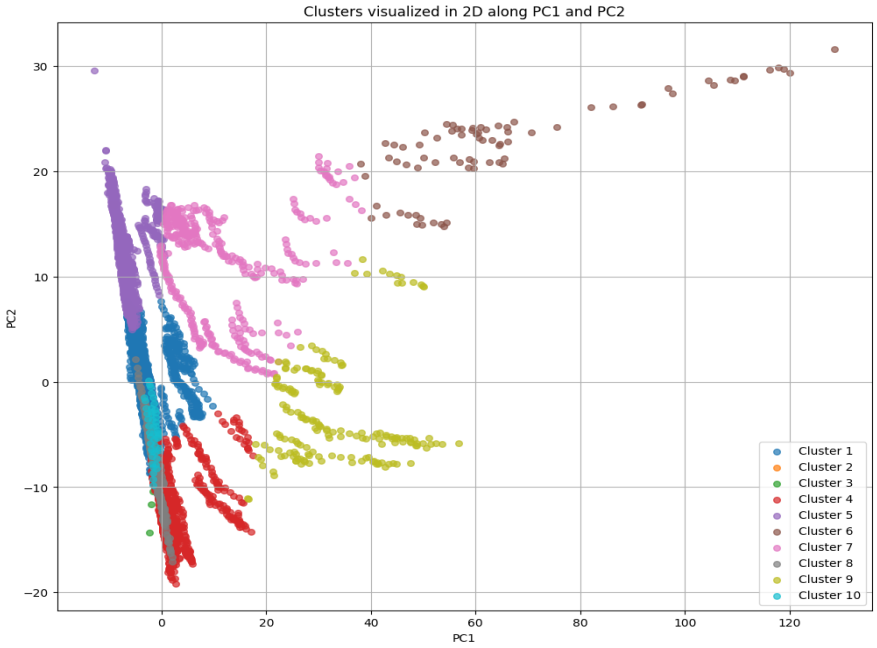
Regular K-Means Silhouette score: 0.09473910691625936

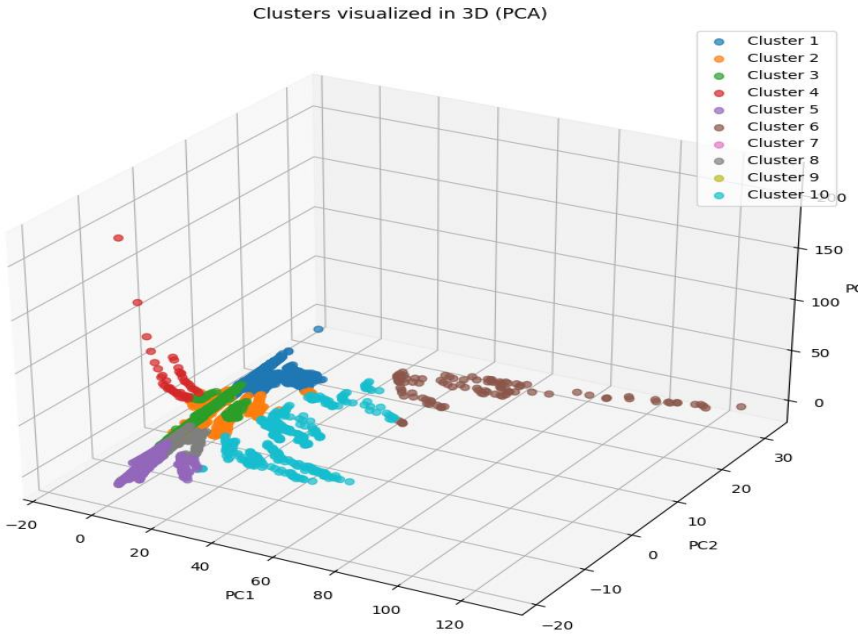
Our end results are virtually the same whether using K-Means or K++, however, K++ was a lot faster as it completed in 50 iterations compared to K-Means completing in 90 iterations.

5. Dimensionality Reduction With PCA:

We performed clustering after reducing our data no just 50 dimensions using PCA below are visualisations of our data.







Loss: 1477802.5744257437

Silhoutte score: 0.13637578248702378

Obviously, our loss goes down because we have way less components, but our silhouette score has improved significantly, which means PCA helped our clustering and lead to better clusters. The top 50 principal components account for 85% of the total variance in our data, a good representation.

Below are the features that contribute to the top 3 PC's:

Top features for PC1:

GDP, PPP (constant 2021 international $) 0.085189

Population ages 65 and above, female 0.084320

Total greenhouse gas emissions excluding LULUCF (Mt CO2e) 0.084221

Top features for PC2:

Population ages 0-14, male (% of male population) 0.089727

Population ages 0-14 (% of total population) 0.089403

Population ages 00-04, male (% of male population) 0.088984

Top features for PC3:

Households and NPISHs Final consumption expenditure (current LCU) 0.158152

Final consumption expenditure (current LCU) 0.157762

GDP: linked series (current LCU) 0.156923

We can see that PC2 corresponds to the general life expectancy and hence overall health of the country or region at that time. PC3 represents the consumption expenditure, indicating trade and purchasing power and things like that. PC1 is economic strength or size, income and industrialization.

6. Cluster Interpretation:

According to the results in our notebook, each cluster was basically just a country, and years all close to each other. This sadly didn't yield much insight into the clusters themselves. Comparing it to the HRD just doesn't make sense or give us any information.

7. Creative Extensions:

1. We implemented the optional early stopping mechanism if below an improvement threshold is seen in clustering for a specified number of iterations.
2. We used silhouette score extensively to evaluate our clusterings and help us find the optimal k. We also used this in PCA
3. We used a Gaussian Mixture model to visually compare to our K-Means or K++ clusters. The GMM gave us very similar look clusters and a similar silhouette score, verifying our results.