**1. Data Preprocessing**

**Motivations for choosing to fill up the empty slots with the mean of related available values instead of the median**:

* Using the mean meant preserving the variance
* This imputation strategy was also chosen because at a later stage we would need to detect any outliers, so using the median would reduce the visibility of the outliers

**2. K-means Clustering and Initialization**

* Please let it be of note that I fist did the PCA to reduce the data to smaller dimensions

This meant finding the optimal value for M= number of dimensions/indicators to keep that would preserve at least 95% of the variance

From a plot I made that number was 175

* I then used this data to do K-means and all that follows was conducted on both the original data and the PCA reduced data with 175 indicators
* And yes, using PCA to reduce the number of indicators did improve the performance, as observed the number of iterations was very much lower for the PCA reduced data while there was a small difference.

Based on the observations made a data-driven recommendation for the best initialization strategy for this dataset is random initialization: here are my motivations

After running the K-means multiple times, we saw that in both iterations of using either the M dimensional data or the original scaled data:

-random initialization is faster to converge and at the end it has a slightly lower loss, so it minimizes the losses

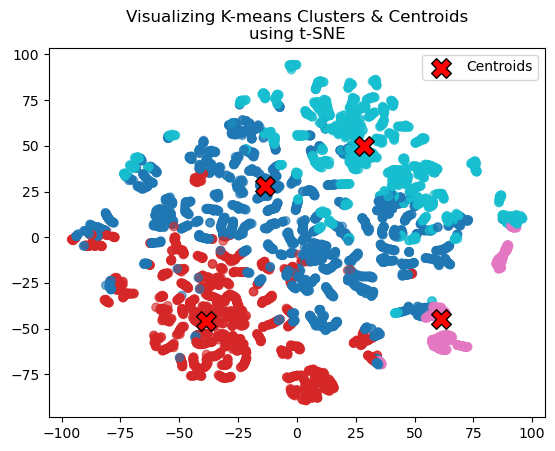
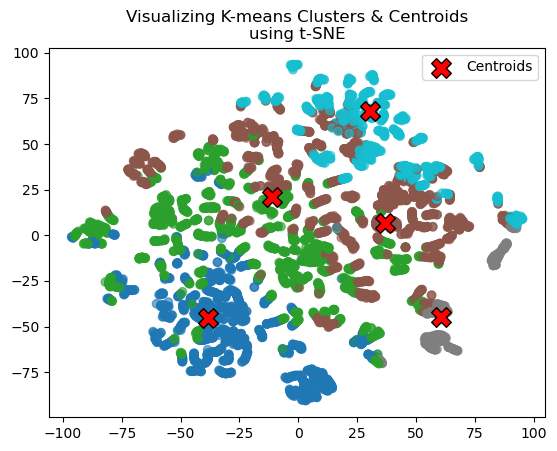
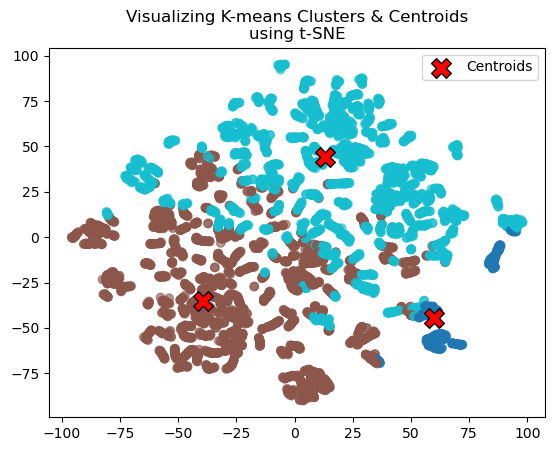
**4. Determining the Optimal K**

To experiment on consistency, I plotted the loss function for both the data with all its dimensions and for when it had only 175:

* **Using the loss function plot:** In both plots the loss seemingly starts to level off once we get to k=3/4 clusters, even though the graphs are not identical this proves to be a constant regardless of the numbers of indicators we are analyzing

**I also visualized the clusters:**

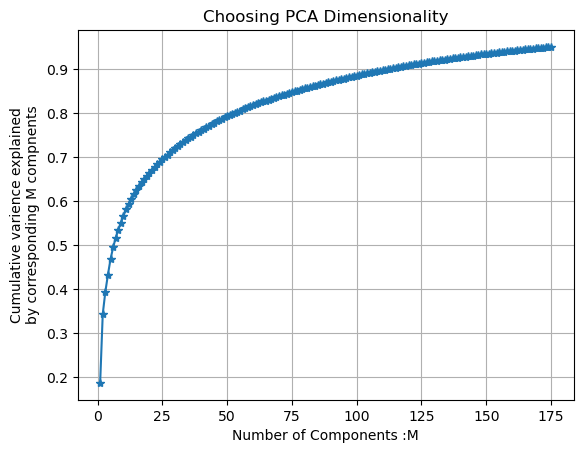
* Based on the visualized clusters below the seems to be a consistent pattern for values of k =3 ,4 and 5 even so 4 seems to be an optimal choice as it segments the data perfectly into 4 clusters (despite some data points clearly being close to the centroid of the pink cluster but being in the navy coloured cluster)



Both approaches are in strong agreement when it comes to the optimal value for the number of clusters.

**5. Dimensionality Reduction with PCA**

Using only 2 or 3D PCA greatly reduces the information



* From this plot i made earlier in my notebook it can be clearly seen that that any number of components less than 25 will only explain at most 70% of the variance in the dataset it can be argued visually that only considering 2 or 3 components will account for very little variance in the dataset
* and this is seen in the visualized plot in task 5 of the notebook it is nothing like the t-SNE visualization

**6: Cluster Interpretation**

**7. Creative Extensions**

* I used the silhouette method to help verify what i had learnt from the plot of the loss function and I observed:

The highest silhouette score in the graph produced is around 2 which even though being the highest is rather unreasonable considering this data is for the whole world

There was a second peak around the values of k=4 and k=5 which aligns with the number of clusters that was supported by our visualization and plot of the loss function.

Another interesting part of the graph was how it started to increase after the values of k=10, which could imply that the could be another better value for k that might improve cluster performance

Overall, there isn't much agreement between the silhouette graph and the loss function we plotted.