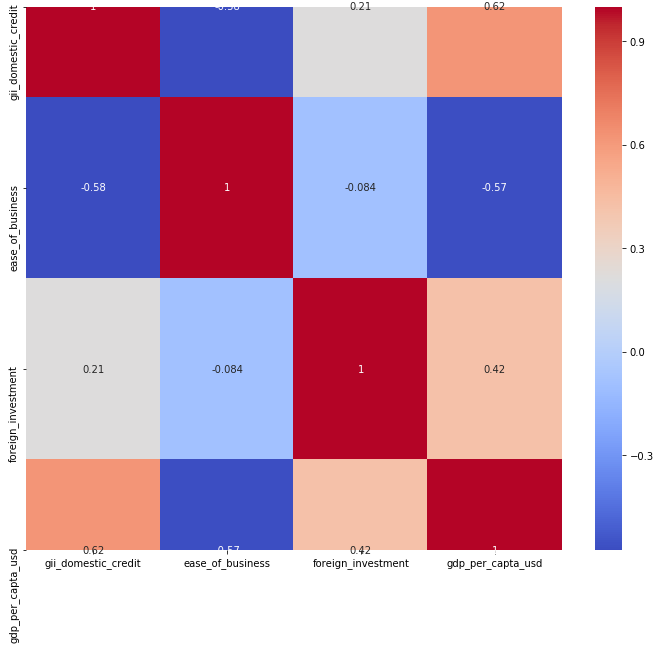
Unsupervised workflow and results

* Team collected a dataset with 46 features from 5 different sources (Global Innovation Index, UNESCO, World Bank, WHO, World Population Review)
* Pair-wise testing of features using correlation matrices revealed a high level of correlation between 26 features. Example: correlation matrix of economic factors:



* As a result, unsupervised workflow performed on two datasets, one featuring all 46 features (*full*) and one featuring 20 features (*curated*)
* K-means clustering algorithm was used for the unsupervised classification, testing cluster sizes from 3-8
* As a QC step, the clustering results from the two datasets were plotted against the *global\_innovation\_index*target variable. Results are plotted below.
* The x-axis contains the *global\_innovation\_index* values with the k-means clustering results on the y-axis, with a range of 3-8 clusters. The first set of plots uses the full 46-feature dataset, while the second set of plots uses the curated 20-feature dataset.
* It was determined that the most robust result came from the 3 cluster model using all features (top left). It is relatively straightforward to identify 3 clusters (0,1,2). The curated dataset was too “noisy”, ie. too difficult to identify the appropriate number of clusters. Even in the 3 cluster example, only two obvious trends (0, 2) can be identified. This is likely due to the sparse nature of the collected data. The entire dataset is required to get a stable result.