Ben Merrill

This is a ReadMe for a project to find a normalized projection for GDP in Africa over the next few years.

Data:

* This is a study on using the world development indicators dataset found here: <https://www.kaggle.com/worldbank/world-development-indicators>.

Goal:

* Better understand the types of time series' that represent Africas indicators
* With our dataset's huge number of indicators (>1000), we want to find k indicators that provide the best representation of the dataset in Africa. By pulling out the k most diverse indicators, we can understand many of the trends that are going on in the Africa, and which indicators have high correlation with each other.
* We will use the archetype that most closely represents GDP to project a normalized prediction of how GDP will change in the next 10 years.

Process:

* Clean and query the data for Africa for indicators where there is 50% or more data density
* Create a correlation matrix sorted by indicator
* Use the k-means algorithm to find k cluster centers that best represent the data, minimizing error
* Match GDP to it's cluster center, and use the center's projection to give a normalized prediction of GDP
* Display GDP projections in a time series

Clustering Algorithm

* For our K-means clustering algorithm, our goal is to create the k cluster centers that each represent an archetype contained in our indicators data set. In order to find the most effective number of cluster centers, k, we will to create an algorithm to measure the overall error of a given number of archetypes, k. We do this by building an objective fuction J, that for each cluster center j, sums the Euclidean distance from each point i in the cluster to its center j, normalized by the cluster standard deviation.
* We want to find the minimum of J, that would show that we've got the best fit for number of cluster centers k.We will run our K means algorithm 100 times for each k, with randomized cluster center location initializations to find the best fit for k.
* Note: We will look for a number of clusters between 4 and 16, so as not to have too many or too few archetypes.