Final Project COGS 118B - Report

Anna Morozova - A16526155
Sebastian Balica - A15396318
Daniel Vega Lojo - A16569859
Agnes Wadell - A17243271
Pablo Moreno - A16112358

1. Introduction and motivation

For this project, we will implement the K-means algorithm and Principal Component Analysis on a data set containing information on 167 countries, regarding child mortality rate, exports, health, imports, income, inflation, life expectancy, fertility rates, and GDP. Our goal is to firstly, use the K-means algorithm to cluster the countries in order to find similarities among the data features. We then will use PCA to simplify the complexity of our high dimensional data while still retaining the essence of the data. To do this we reduce the number of dimensions, project them onto a smaller dimensional subspace, and maximize the variance of the projected points. What this does for us is it allows us to get the best summary of the data by reducing the features into principal components, rather than the data points, as with K-Means. We can then make better interpretations of the countries based on the reduction of features, without losing any important patterns.

2. Related work

In Chris Ding and Xiaofeng He's paper, "K-means Clustering via Principal Component Analysis" it is explained that the process of applying PCA to data and projecting it to a subspace where the K-means algorithm is then applied, is a very common process within unsupervised machine learning. As stated in the article, "K-means Clustering via Principal Component Analysis" K-means and PCA have a close relation even though initially their main purposes might seem different. The article suggests that it can be important to apply the two algorithms as a way to improve clustering results. Their evidence shows that after reducing their data from 1000 dimensions to 40, 20, 10, 6, 5, the smaller dimension's cluster accuracy increased.

3. Methods

K-means

The K-means algorithm (Bishop, 1995) is an algorithm for identifying K groups/clusters of data points in multidimensional spaces, the implementation for this section does not deviate too much from our algorithm first created from Hw3. We start by first by assigning randomly selected "k" cluster data points from the data set to be the initial cluster locations. The same helping functions from before calcSqDistance, determineRnk, and recalcMus work together to complete the algorithm. First, the calcSqDistance function takes the parameters of the data set and initial cluster locations, using the || x - μ || (euclidean norm) as a metric of distance we

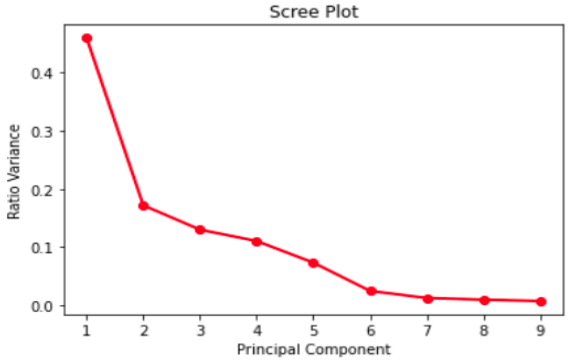
calculate the distances; We return the *sqMatrix* matrix that is NxK in dimension containing the distance to each cluster mean from all N data points. Next the determineRnk function uses the *sqMatrix* as the only parameter and will return the cluster locations in the *Rnk* Matrix NxK dimension, using the index_Argmins from all of the k clusters to determine which cluster it belongs to. As a form of latent variable, 1 is assigned to the cluster it is the minimum distance, while zero the remaining clusters. Lastly, the recalcMus function will use the result from the *Rnk* matrix and the data set as a way to recalculate the new mean locations of each cluster. The returned matrix is a KxD dimension, similar in shape to the initial cluster initialization. Setting the number of iterations to be around 1000 will ensure that there is a stopping criteria in the case that the means do not converge to the stopping point. Each cluster is then plotted using the world map package to visualize the clusters. Prior to running the algorithm we must first remove labels from the data set, converting them from a pandas data frame into a numpy array in the process. As a way to test our algorithm we decided to use 4 clusters to classify and group our data.

PCA

Principal component analysis (PCA) is a technique used for applications such as dimensionality reduction, lossy data compression, feature extraction, and data visualization. Given a set of data $\{x_n\}$ where n = 1, 2, …, N and x is a variable with dimensionality D. The goal of PCA is to reduce the dimensionality of each variable while maximizing the variance of the projected data.
For this project, we used the PCA section from the sklearn.decomposition module to implement this algorithm. First, we standardize the data to avoid features with large values having a large impact. Those features are selected based on the variance they cause in the output. The feature with the highest variance becomes the PCA. We plot the variance ratio against the principal components to find the optimal number of dimensions. We finally train our algorithm with the desired number of dimensions and the scaled data using fit_transform() and transform() functions.

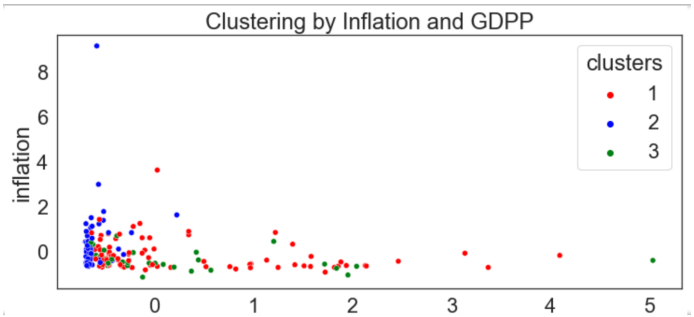
4. Results

Before training the PCA algorithm, we needed to find the optimal dimension. For this purpose, we plotted the variance ratio against the principal components.

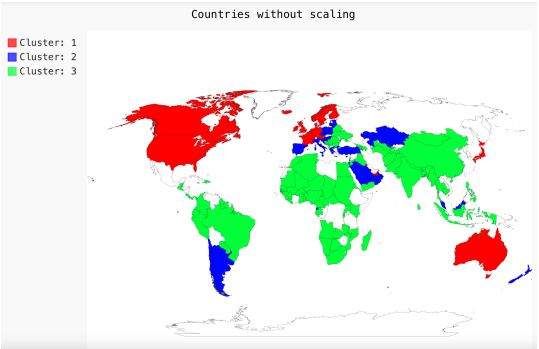


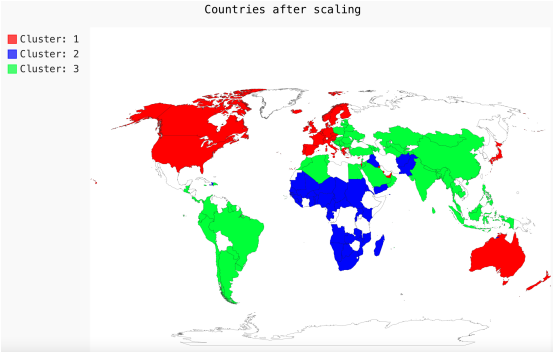
As we can see, using dimensions from 5 up to 9 would have little impact on the predictions since they produce very little variance. The optimal dimension here would be three since those dimensions produce the most variance.

We performed clustering by 2 features first, and we saw that clustering didn't work very well - there was no clear division between clusters, they overlapped with each other.

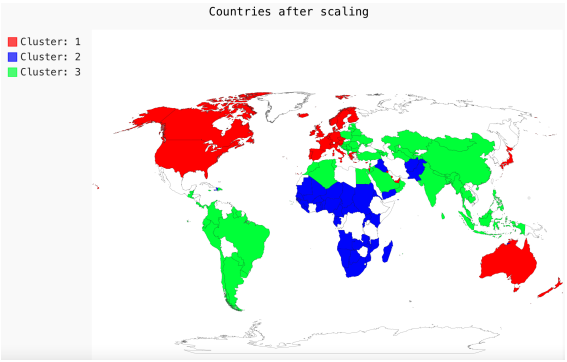


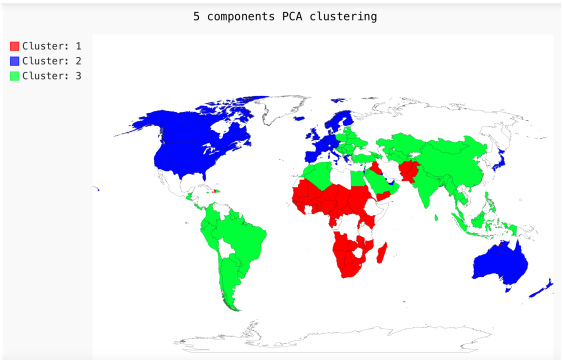
Then we did clustering on all the features, and we illustrated our results on a world map. We scaled our data and saw that it helped to reduce variance for our data and get rid of any outliers, the clusters turned out to look more generalized. However, our algorithm didn't include some countries in the clusters at all (Mexico, Russia, Greenland).



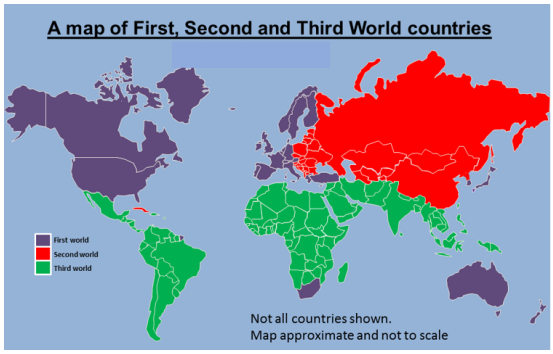


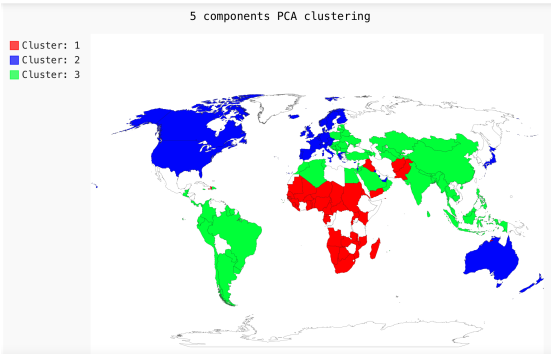
We also performed PCA to reduce dimensions of our data to 5. However, it didn't make any difference in our clustering. The clusters' borders stayed exactly the same.





We also compared our results to the map of the first, second and third world countries. Cluster 2 corresponds to the First world countries. Cluster 3 corresponds to Second world countries including South America in it. Cluster 1 vaguely corresponds to Third world countries excluding South America though.





5. Discussion

In this section, the key learnings of the project will be discussed together with suggestions for improvements as well as possible extensions for further studies.

5.1 Key learnings

K-means by itself can be used to make sense of real-world datasets
The country clustering is easy to understand based on previous knowledge. Accordingly, our project clearly shows that K-means can be used to split data into clusters that make sense. This learning may sound simple, however, given that the results from the K-means clustering are easy to interpret, it definitely shows the strength of the algorithm.

Standardization of data important to avoid dominant features
We could see in our project that given that K-means is based on spherical assumptions, the variables with larger variances such as gdp and income will impact the final clustering to a greater extent. We learned that standardization of the data reduced this impact.

PCA limits the interpretability of the analysis
When clustering the data before the PCA analysis, the produced clusters made intuitive sense. However, a disadvantage with PCA is that after applying it, we could not tell anymore how the different features impacted the final result. This makes it harder to use the result for action oriented insights, e.g. to determine key features of improvement to help a country move from one cluster to another.

5.2 Improvements and extensions

Improve initialization of cluster means
We saw in our project that the clustering result depended on the cluster initializations. The reason for this is that K-means only finds a local minima. Accordingly, it would be of interest to use more sophisticated methods for initial cluster allocation to ensure that the clusters represent the data well.

Soft clustering with GMM is a potential extension of the project

Given that K-means clustering assumes that the variance of the distribution of each variable is spherical and identical, some data points might get a counterintuitive cluster assignment.. It would accordingly be of interest to run GMM on the data set to allow for clusters of different shapes and sizes. GMM also allows for soft clustering which could be useful to identify differences between countries within the same cluster.

6. Additional sections

Contributions

Pablo Moreno:

- K-means algorithm code fixing/variable naming, implementing the displayCluster world map helping function, and K-means algorithm for the Method section in report. PCA algorithm in video. Data cleaning and plots was done as a collective

Agnes Sofia Wadell:

- Code: Modification of data set for it to be compatible with to K-means algorithm, implementation of K-means algorithm, scaling of dataset to prior to clustering, screeplotting to determine dimensions for PCA
- Report and presentation: Discussion section

Sebastian Balica

- Discussed and collaborated on implementation of K-means algorithm and PCA visualization
- Intro and motivation and related work section for report and presentation

Anna Morozova:

- Code: algorithm for world map illustration, discussed and collaborated on K-means clustering, finalyzing and final edit of the repository
- Report and presentation: Results section
- Video editing and compilation

Daniel Vega Lojo

- Code: Overall fixing/cleaning code redundancy, variable namings, implementation of functions such as getClusters(), createListofList(), and getClusterIndex. Code documentation.
- Report: PCA section and optimal dimension result in Result section.
- Video: K-means algorithm section.

Code

https://github.com/anna-morozova-ucsd/COGS118B---Project---Coutries-analysis/blob/main/Final_project_COGS_118B_masterV6.ipynb

References

Pasi Fränti, Sami Sieranoja. *"How much can k-means be improved by using better initialization and repeats?".* Link: https://www.sciencedirect.com/science/article/pii/S0031320319301608#:~:text=K%2Dmeans%20clustering%20algorithm%20can,results%20of%20the%20initialization%20technique.

Kaggle. *"Unsupervised learning on country data".* Link: https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data/code

https://ranger.uta.edu/~chqding/papers/KmeansPCA1.pdf

https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

https://www.pygal.org/en/3.0.0/documentation/types/maps/pygal_maps_world.html

https://www.acid.uk.com/acid-member-call-to-action-tell-us-how-you-deal-with-ip-in-third-world-countries-deadline-imminent/

Video

https://youtu.be/z04KlL3Ds50