Trends in Global Income Inequality from 1980-2010: A Visual Analytics Approach

Arshad Ahmed, MSc Data Science, City University

Abstract— We use open data from the World Bank to evaluate the global trends in Income Inequality. We find that from the period between 1980-2010 there has been an overall increase in the amount of wealth heald by the top 10% of the population across the globe. We also find that as a proportion GDP and Population some countries have experienced far worse increase in Income Inequality than others. We also show that the choise of data transformation has a very important bearing on thefinal output. In addition we utilise a Visual Analytics approach with an emphasis on an interplay between computation and visualisation with each step informing and complementing the next. This approach is found to be very useful as allows fast iteration and shortens the route to insight from the data. We find that the Min Max Scaling to [0, 1] interval is particularly effective at highlighting tax havens such as Panama and Seychelles in addition to lax tax jurisdictions such as Luxembourg. Also we find that K-means and BIRCH give identical results on this data.

Index Terms—Visual Analytics, Maps, Income Inequality, Clustering

INTRODUCTION

The purpose of this study is explore the global trends in Income Inequality using open data from the World Bank Group. To perform this analysis we consider the time period from 1980-2010. The reason for this choice is that the Income Indicator we use: Income held by the top 10% of the population does not exist prior to this time frame.

The motivation for this study can be summarised by the following analytical questions:

- ⊕ What are the global trends in Income Inequality? How has this changed over time? Which countries are most affected?
- ⊕ Can we derive any additional attributes that allows to segregate this data further?
- ⊕ What is the impact of data transformation on the final results?
- ⊕ How can clustering be used in this context?

The visual analysis tasks can then be summarised as follows:

- ⊕ Use of shaded maps to show the mean Income Inequality over the time period
- ⊕ Use of proportional symbol maps to present multiple Income Inequality attributes
- ⊕ Use of mosaic plots to show the countries that the attributes highlight with different data scaling.
- ⊕ Apply clustering and extract the cluster labels to compare to previous maps and explain variances

In addition we also aim to show that the Visual Analytics methodology is a useful approach for the analysis of this dataset as the data has spatio-temporal component because it is related to countries and we are considering a time series of the attributes of interest.

This paper is organised as follows in section we present a literature review of Visual Analytics methodology, clustering methods and map visualisations. In section 2, we explain in detail the data analysis and the results. In section 3, we present our discussion and answer the question we set out in the introduction. We finish with our conclusions and suggestions for further work.

1 LITERATURE REVIEW

1.1 Visual Analytics Methodology

According to [1], visual analytics can be defined as science of analytical reasoning supported by interactive visual interfaces. This approach aims to combine visualization, human factors, and data analysis. In the visualization aspect the field benefits from innovations in other fields such as information retrieval, data management & knowledge representation as well as data mining. This approach also allows for human factors, such as cognition and perception, to play an important role in the in the analysis and the resulting final output. [1]

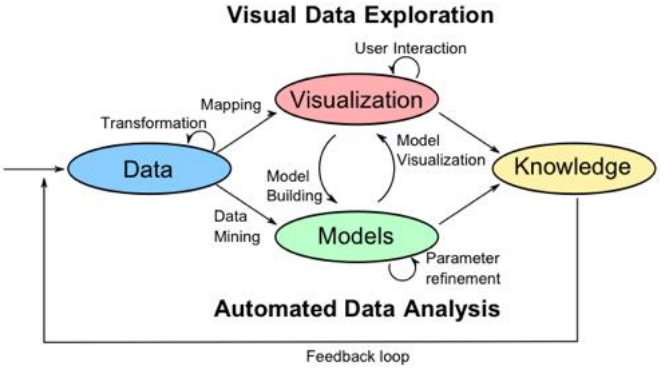

Figure 1: The VA methodology exemplified.[1]

1.2 Clustering Methods

Clustering is an unsupervised learning technique that aims to find structures in data utilizing either partitional, hierarchical, density, grid or model based methods [2]. Intuitively they can be thought of as grouping different objects within a dataset to groups of objects with similar properties. Therefore after clustering the properties of the items in a cluster should be more similar than those of items in a different cluster [3], [4].

Clustering is called an unsupervised learning technique because in contrast to supervised methods like classification

there do not exist any labels in the data. Thus the structure of the data must be learnt from the data. [3]

Clustering is difficult due to the many factors that must be considered and addressed before a successful clustering algorithm can be implemented. These factors are derivation of effective similarity measures, criterion functions and initial conditions. Also, the authors in [3] note that no clustering method is able to handle all aspects of a cluster structure such as different shapes, sizes and densities successfully. Inherently, some methods are better at handling some of these structural properties than others.

The clustering techniques presented here are not restricted to partitional and model based types but others such as graph based, density based, hierarchical and grid based methods also exist. Based on our reading of the literature, the partitional and model based ones are selected because of their ease of use. But the other types are mentioned here to illustrate the different methods available and how the literature influenced our choice of computation techniques.

1.2.1 Partitional Clustering

Partitional clustering algorithms finds divisions in the data rather than an agglomerative structure produced by hierarchical methods. These methods have the benefit of being scalable to large datasets and hence are widely used. Dendograms produced by agglomerative or hierarchical methods can become prohibitive computationally when dealing with large datasets. The partitions are produced my optimizing some criterion function defined either globally or locally. Since the potential combinations of labels within the datasets are large in practice these algorithms are run multiple times with different starting states to identify the best clustering output. Partitional methods are divided into centroid and medoid algorithms. The centroid methods represent a cluster by the center of gravity of the instance while the medoids methods represent each cluster by the means of instances that are closest to the center of gravity. [2], [3]

The K-means algorithm is a centroid based method which partitions data into k disjoint partitions by minimizing the objective function in Eq. 1.

$$E = \sum_{i=1}^{K} \sum_{O \in C_i} |O - \mu_i|^2.$$

Equation 1: K-means objective function

Here, the individual points are represented by O in a cluster, Ci where μ, represents the mean of the objects or the centroid of Ci. The squared distances between the objects within a cluster are minimized from their cluster centres. The K-means is a fast and simple algorithm which has complexity of O (I*k*n) where I is the number of iterations and k is the number of clusters. Some examples of k-medoids methods are CLARA and CLARANS. The CLARA method determines the medoids of the dataset through an iterative optimization and classifies samples of the dataset into partitions. Assuming a random sampling the medoids inferred from the samples are taken to represent the medoids of the dataset. The CLARANS methods uses a randomized graph search for the medoids in contrast to CLARA. [3]

The authors in [16] group the initialization of K into the following methods:

- Variance based – these aim to use intuitive or model based functions of a criterion to get extreme values at a correct K. Some examples of the statistics used under this approach is the gap statistic, jump statistic, Fisher wise criterion and Hartigans rule. All these measure incorporate the above intuition through different objective functions.
- •Structural approach – these compare within cluster cohesion with between cluster cohesion at different K values. The silhouette value is an example of this. We make use of the structural approach for K-means initialization as suggested here.
- •Consensus distribution – this approach suggests choosing K according to the distribution of the consensus matrix derived from clustering the data with different K values.
- •Hierarchical approach – choosing K by the results of the agglomerative or partitional clustering method
- •Resampling approach – choosing the K value based on the similarity of different k means clustering of the data with randomly perturbed or sampled data.

1.2.2 Model Based Clustering

Model based clustering assumes that the data are generated by a mixture of probability distributions. The statistical approach to model based clustering often assumes a Gaussian mixture model which then allows for maximum likelihood type approaches such as Expectation Maximization to be applied iteratively to arrive at parameter vectors of the component densities. However, non-parametric density estimation methods such as those based on the Parzen window have been used to search for bins with large counts in a multi-dimensional histogram of the input data. The other approach is to use neural networks. Both approaches attempt to improve the fit of the data to an underlying model. The Self Organizing Map (SOM) is the best known method in this category. [2], [3]

An SOM can be thought of as a two layer neural network, where each neuron represents an n-dimensional weight vector. The dimensions of the weight vector corresponds to the dimensions of the input data. The SOM is trained iteratively and the neurons act like the centers of the cluster. At each training step a vector is chosen at random from the input and then the distance between it and all the weight vectors are calculated using some distance measure. After this step the neuron with the weight vector which most closely matches the weight of the input vector are moved closer to the input vector. The topological neighbors of these matching units are treated in a similar manner. The SOM is a very robust technique that can be used for outlier detection and can deal with missing values. [4], [5]

1.2.3 Hierarchical Clustering

Hierarchical clustering algorithms typically fall into the single-link, complete link and minimum variance approaches. The difference between these methods are their similarity measures for computing clusters. In the single link case the distance between two clusters is defined as the minimum distance between all pairs of points drawn from the two cluster i.e. one point from cluster 1 and one point from cluster 2. In the complete link case, the distance between two clusters is defined as the maximum of all pairwise distances between the points in two clusters. [4]

These algorithms perform a sequence of partitioning operations either in a bottom up way which is like performing repeated consolidations of data points until a user defined threshold is reached. This can also be done top down where recursive

partitions of the data are computed until a threshold is reached. [5]

Some examples of these methods are CURE, ROCK, and BIRCH. The CURE algorithm is bottom up method which use well formed group of points to define intra cluster distance rather than a centroid based approach. CURE begins by choosing a constant number of well scattered points from a cluster and then shrinks the selected points towards the centroid of the cluster using some predetermined fraction. ROCK uses concept of links to measure similarity of a pair of points. The number of links is the number of common neighbors of these points. The merging of clusters also utilizes links instead of distances which allow this method to be extended to non-metric similarity measures. BIRCH is an integrated hierarchical clustering algorithm that uses clustering feature and clustering feature tree to summarize clusters. This allows the BIRCH method to be scalable to large datasets, while being fast and is also suitable for clustering of incremental and/or dynamic inputs. This method applies multiple phases of clustering, where the first phase produces a basic clustering and further iterations can be applied to refine its output. [5]

1.3 Cartographic Techniques

We conduct a brief search of the literature to get a sense of what type of map visualizations have been proposed and what their relative strengths and weaknesses are. Since the data relates to countries a Visual representation through maps would be very effective.
With regards to map visualization's we identify a few types below from [6]:

- ⊕ •Choropleth Maps
- ⊕ •Cartogram
- ⊕ •Dot distribution Maps
- ⊕ •Proportional Symbol Maps
- ⊕ •Dasymetric Maps

Chloropleth maps are thematic maps in which areas are shaded by color using some attribute. Cartograms are thematic maps that deform the area of the map using the variable that is to be represented. Dot distribution maps represent density of an attibute by using scatter to shown spatial patterns. The value of the attribute is represented as dots. The proportional symbol techniques uses symbols of different sizes to represent data associated with different locations.the dasymetric map is an extension or alternative to the chloropleth map where the attributes are represented by enumeration units on the map so the regions appear uniform but additional information is used to model the internal distribution of the variable. So it has an additional distribution layer on top of the chloropleth map.

For our analysis we will use mainly chlorpleth maps with a divergent color scheme. The reason for ths being that even though we have shown some of the other visualisations that are available they are not appropriate given the data. The data from the World Bank is it a country level and at the year level. So it is a highly aggregated time series and a lot of the data at the year for our time range is missing for a lot of countries. Therefore in the analysis we resort to using further aggregations such as percentage change in the attributes over 10 years and yearly averages for all countries to get a sense of global trends. Therefore chloropleth maps represent the simplest and easiest representaion of our data. Using a area deforming approach would not be ideal because smaller countries next to big countries with similar trends would not be as visible. Also as the data is at the country level it would make no sense to use proportional symbols or daysmetric methods as there is insufficient granularity in the data.

The colors can be inferred from the values being visualised and depending on the nature of the attribute the colors chosen can be harmonic, contrastive or clashing. [7]

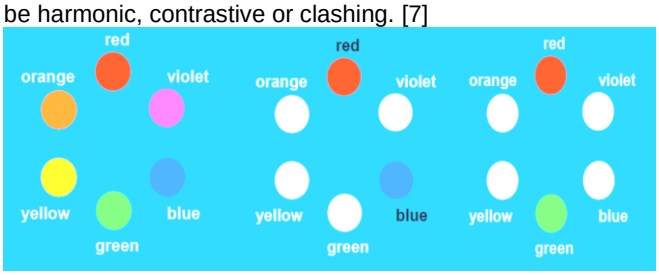

Figure 2: (Left) a color wheel showing harmonic colors. Harmonic colors are those that are next to each other on the wheel such as yellow and green. (Middle) Shows contrastive colors which are colors with one neighbor skipped such as red and blue. (Right) Shows clashing colors which are two neighbors removed.

2 Data Analysis and Results

2.1 Tools

This analysis is performed using a wide range of tools which are both open and commercial. The bulk of the analysis is conducted in IPython[8] using the PANDAS[9], SCIPY[10], SEABORN[11], MATPLOTLIB[12] and SCIKIT-LEARN[13] libraries. In addition we use Tableau[14] for the Visual Analysis tasks and use MATLAB[15] for the initial cluster analysis.

Note that Python zero based indexing and especially when talking about the clustering results the cluster 0 is referred to as cluster 1 sometimes and so on.

2.2 Data Selection and Loading

The indicators that we use along with their WB code as follows:

- ⊕ Gross Domestic Product (GDP) : NY.GDP.MKTP.KN
- ⊕ Total Population (Pop): SP.POP.TOTL
- ⊕ Income share held by highest 10% (II): SI.DST.10TH.10

The Income share held by highest 10% is our variable of interest which we will take to represent a measure of income inequality. The first step in our analysis is to retrieve these indicators from the World Bank using the built in API in the PANDAS[9] IPython[8] library for the selected time range.

2.3 Data Preprocessing

We then checked for missing values and filled all missing values with zeros. Also it was noticed that the data had some aggregates such as those for High Income, Low Income, High Debt, Low Debt countries in addition to the country level data. As a preprocessing step we delete these aggregations and extract the country data from this to derive our raw dataset. This forms the basis for all subsequent processing. This leads to an input dataset size of 7688 by 4.

2.4 Exploratory Analysis

The next step was the exploratory analysis step to understand our data a bit better which would inform the subsequent

analysis. We do this by using traditional line, bar plots and correlation matrices.

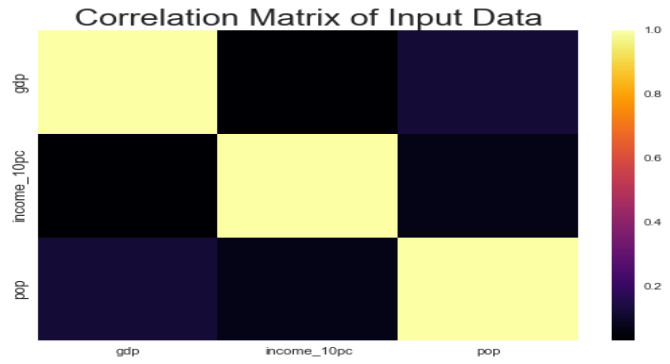

Figure 3: Pearson Correlation Coefficient among indicators in the input data before scaling

From this plot we already start to see the correlations among our variables and get a sense of what to expect. We observe that all the variables are negatively correlated none of the variables show a positive correlation among them. This means that based on this data we can expect to see that II increases even if GDP and population decrease or it can mean that wealth concentration at the top is much faster than it is for the other 90% of the population so we don't get a linear mapping of the variables. This could be an illustration of the rich get richer principle.

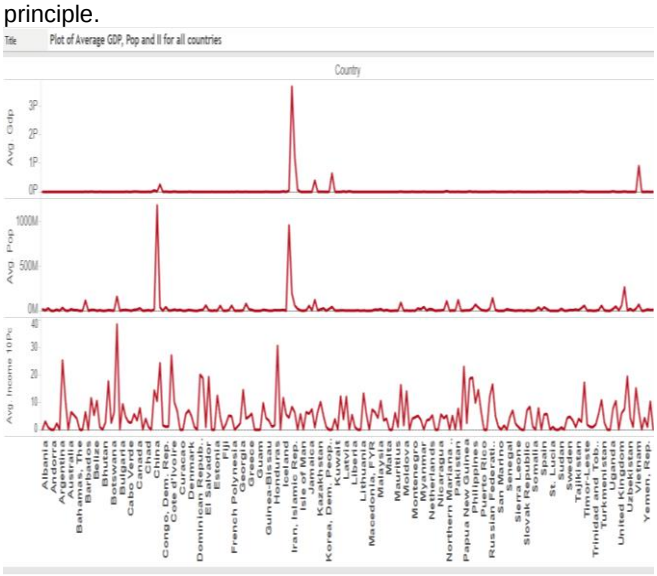

Figure 4: Plot of raw indicators after data loading and filling with zeros. We observe that there is a huge variation in scale among the indicators with GDP and Population being many magnitudes larger than the II variable. This shows the necessity of data transformation in this case.

We see from Fig 4, that due to the large number of countries 214 to be exact, the plot becomes cluttered and it becomes impossible to gain any insight from it. The proportions of the plot that would allow any meaningful analysis becomes prohibitive when there is a space constraint such this report format. But also it becomes difficult for the analyst to glean any other insight apart from the distribution of points which in case of the GDP is very sparse because of presence of large values. This is the motivation for the use of visual analysis that allows for more meaningful interaction and analysis of this data. But for any approach to be successful we need to scale the data to make them comparable.

2.5 Data Transformation

As Fig 4, clearly shows there is a need for scaling. So initially we tested one type of scaling which Min Max scaling to interval of [0, 1]. But one of the issues of this is that as we have already observed that there are some very large values in this dataset which are valid. So using these to scale the data would force a lot of points to close or at zero. Hence we have created an artifact from the processing of the data. This was the motivation to explore two other type of scaling the Standard Scaling, which is mean centering and scaling to unit variance and Robust Scaling which uses median and Interquartile Range to scale data in a manner that is more resistant to outliers because these measures have high breakdown points when faced with outliers compared to the mean and standard deviation which are very sensitive and not robust to outliers.

2.6 Attribute Derivation

As our initial analysis suggested that the time series has a lot of missing data so doing any representative comparison on a yearly basis would not be entirely accurate approach. The analysis would be dominated by countries for which there are data and we would miss any insights for places where there is little data. As a way to counter the sample imbalance and mitigate issues arising from different imputations of input data a simple but effective approach was found to be to fill the missing data with zeros, scale the data and derive attributes. After this step we grouped the data by country and aggregated by the mean.

Prior to data aggregation, we derived additional attributes to gain additional insight into the data. Here we explain these additional attributes.

The attributes that we calculated were as follows:

- ⊕ II / GDP Ratio
- ⊕ II / Population Ratio
- ⊕ II / (GDP + Population) Ratio
- ⊕ 10 year percentage change in II, GDP, Population
- ⊕ 30 year percentage change in II, GDP, Population

The attributes are calculated from the scaled indicators. As our analysis will show that splitting the data along these attributes allow us to gain insight into the data based on the different scaling that we used that were otherwise not obvious from the raw data. The visual analysis is a key component of this and it was to confirm the findings of the min max scaled data through visual analysis that we performed the additional tests with the scaling.

2.7 Data Merging and Aggregation

The data was grouped by country and then merged on the country and the year keys. Additionally, we also derived a dataset aggregated by the year to get an idea of the global trends over the time range.

After aggregation we write out the dataset from IPython[8], to Excel spreadsheet format for visual analysis in Tableau[14].

2.8 Results

The use of Tableau facilitated the Visual Analysis tasks that were identified at the beginning of this paper. Firstly, we look at

the mosaic plot of the data with different scaling to see its effect on the data. We use a diverging scale to show different values.

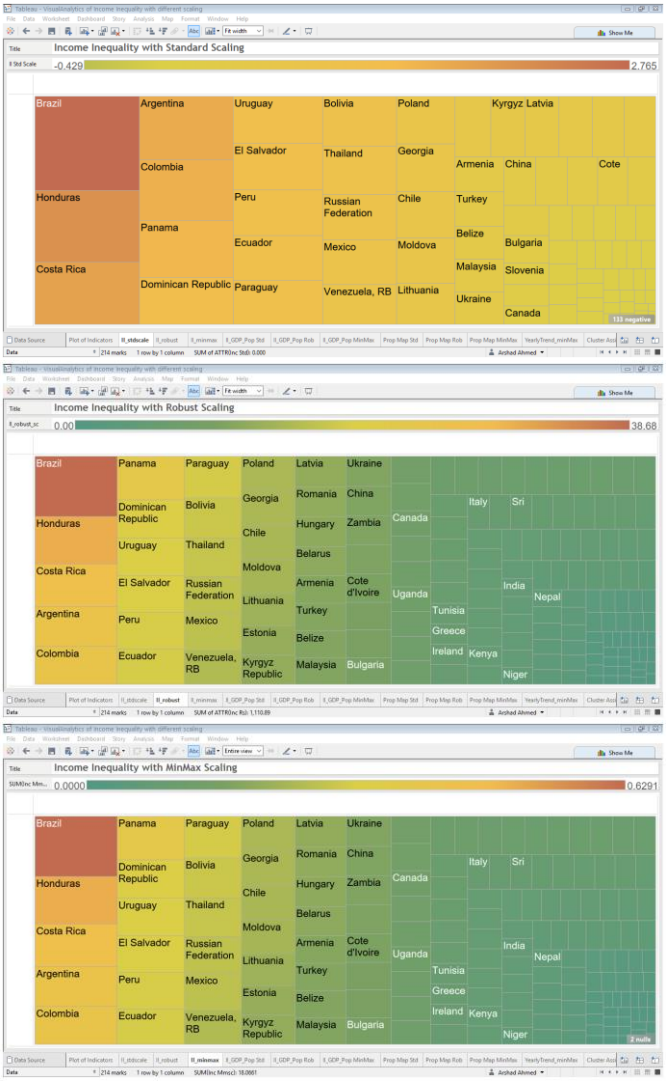

Figure 5: Mosaic Plot of II variable with (top) Standard Scaling (middle) robust scaling and (bottom) Min Max Scaling.

From Fig 5, the effect of data transformation is clear. This is easily shown in the Visual Analysis, because with all three scaling's we retrieve the countries in the same order starting with Brazil at the top followed by Honduras, Costa Rica, Argentina, Colombia and the tax haven of Panama. We have empirically shown that the data under the different transformations yields the same structure but the proportion is over estimated by the Standard Scaling of the data but is almost identical with Robust and Min Max Scaling of the data.

To assess the impact of this further we consider the mosaic plots of the II/(GDP + Pop) attribute that we calculated earlier but with different scaling.

Figure 6: Mosaic Plot of II/(GDP +Pop) with (top) Standard Scaling (middle) robust scaling and (bottom) Min Max Scaling.

From Fig 6, we start to see divergence in the attribute for the dataset but as a result of the scaling. The Standard Scaling highlights Argentina, Chile, South Africa and Kenya while the same attribute with the Robust Scaling yields, El Salvador, Honduras, Macedonia and Kyrgyz Republic as the top countries with the highest value for this attribute. The Min Max scaling produces the more interesting results by highlighting Panama, Luxembourg, Estonia and Iceland in addition Honduras. This scaling allows us to remove what almost appears to be a very strong South American bias in the scaled data. The picture with the Min Max Scaling is more nuanced as it highlights some of the nations highlighted by the Panama Papers scandal such as Seychelles and Panama.

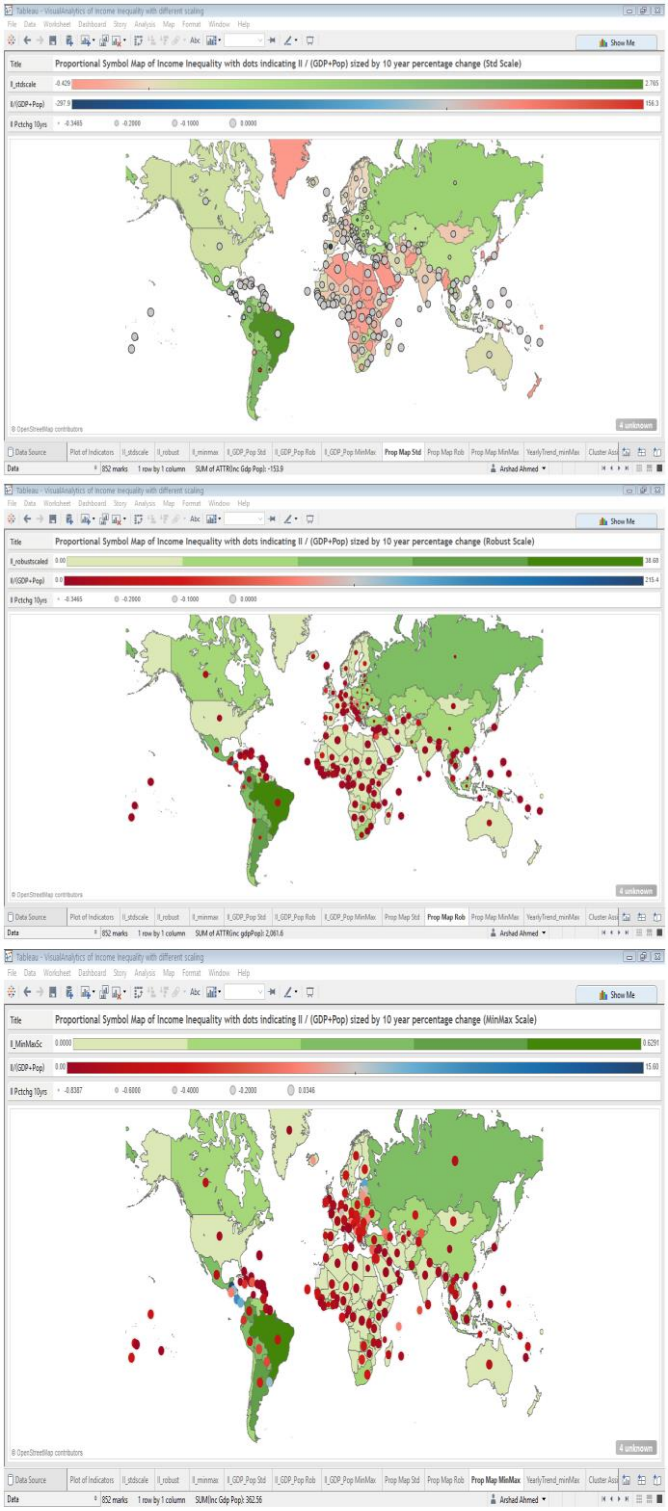

Figure 7: A combination map which shows the countries shaded using a diverging color scheme using the scaled II attribute. The circles are proportional symbols which are colored by II/ (GDP + Pop) attribute and the size is encoded using the 10 year percentage change in the II variable. A diverging color scale is used to represent the values of the symbols and the scale also shows the values that the different circle sizes correspond to.

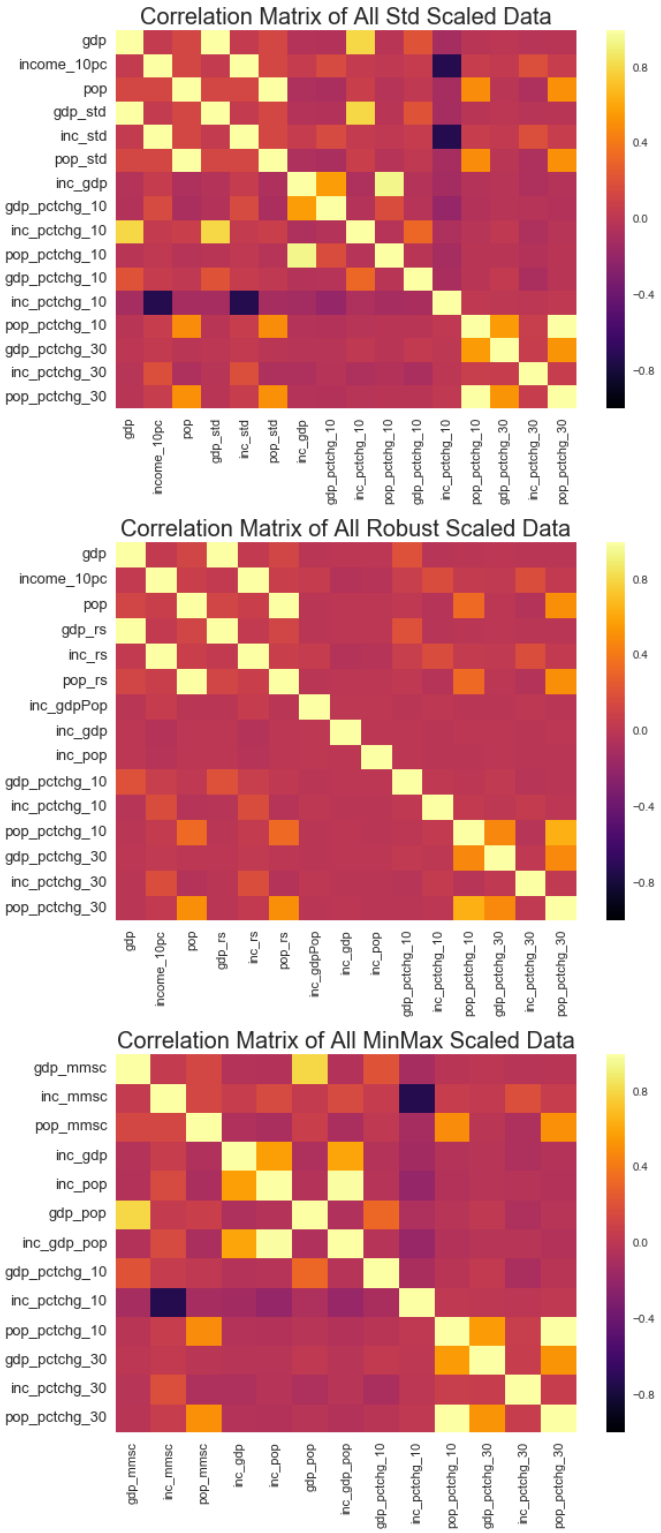

Figure 8: Pearson Correlation Matrices for aggregated datasets containing all attributes with (top) Standard Scaling (middle) robust scaling and (bottom) Min Max Scaling. The Standard and Min Max scaling produces similar correlations among variables while the Robust Scaling produces a slightly different correlation matrix where the negative correlations in the other two matrices are less strong. But apart from this the relationships are broadly similar among the three matrices.

2.9 Cluster Analysis

Out of the three transformations considered the Min Max is perhaps the more interesting one because it does not seem to be biased towards South American countries as the other two transformation types. Therefore, we use this dataset for cluster analysis.

We utilize three clustering techniques in our cluster analysis. We use a K-means, BIRCH and SOM. Our novel suggestion is to use SOM to initialize K which was not encountered as a suggestion in the literature. As a validation we also compare our K-means evaluated structurally and calibrated by SOM to BIRCH which derives the optimal cluster numbers from the data.

The aim of the cluster analysis firstly is to determine the optimal number of clustering of the data. We take a structural approach to evaluating K as suggested by the literature. To do this we use a K-means with different values of K. The evaluation of the K-means is done using the Silhouette Value. Also the distance function used is City block distance as opposed to Euclidean distance as it is a more sophisticated measure. As an additional verification of the clustering solution derived with K-means we try another clustering method to see if we get the same results. We try the neural network analogue of K-means the SOM.

The Silhouette value measures the similarity of inter cluster objects to that of intra cluster objects. It can be thought of as a measure of cohesion between objects in a cluster compared to the separation of clusters. This value ranges from -1 to 1. A high value indicates a good match to its current cluster and a low value indicates a poor match to neighboring clusters. If the objects derived from the clustering solution have mostly high values then the configuration can be thought of as being optimal while if there are many negative points than this indicates that we have defined an insufficient number of clusters for the data. We evaluate the k values of [2, 3, 5, 7, 9, 11] on this data evaluate the silhouette value for each solution and then plot the mean silhouette value against increasing number of clusters.

We cluster the data with an SOM using a [5, 5], [10, 10] and [15, 15] map. Since the SOM network takes a square form the dimensions translate to 25, 100 and 225 neurons in the network. The input data has a weight vector of 214. So the first two set ups cluster and reduce dimensionality of the data. For the SOM we present the U-Matrix which is a two dimensional representation of the higher dimensional luster centers. The blue hexagons in these plots represent the neurons and the red lines connect the neighboring neurons. The colors in the regions that contain the red lines indicate the distances between the neurons in the SOM network. The darker colors represent greater distances, while lighter colors represent smaller distances.

2.9.1 K-Means Clustering

For the K-means we apply clustering to the input data with different values for K. We use K values of 2, 3, 5, 7, 9 and 11. The silhouette plot is evaluated for each value of K and then we look at the mean silhouette value for each K value and pick the cluster number with the highest score. This is shown in Fig 7.

The mean silhouette value plot shows a U shaped behavior which means that we have two ideal clustering's for the data. The 2 and 11 cluster option yield the highest scores and the rest experience a dramatic fall in the score. This could be that with 2 cluster option it finds the countries with high II as we have shown with our tile maps but with the option of more clusters it is possible that the algorithm is able to achieve more granularity in partitioning the countries using all the attributes that we have. But it is also possible that with such high number of clusters the last clusters are very arbitrary. The cluster labels will allow us to evaluate this.

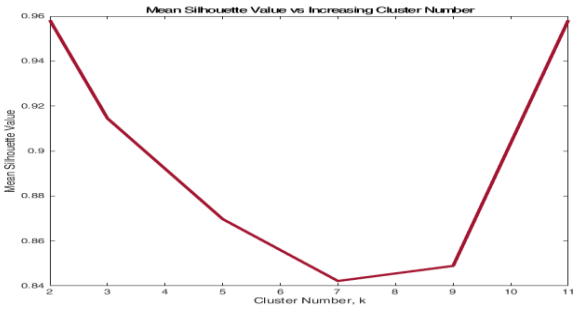

Figure 9: Mean Silhouette Value for K-means with increasing K-values.

2.9.2 SOM Clustering

For the SOM network we evaluate the U-matrix to observe the clustering of the data.

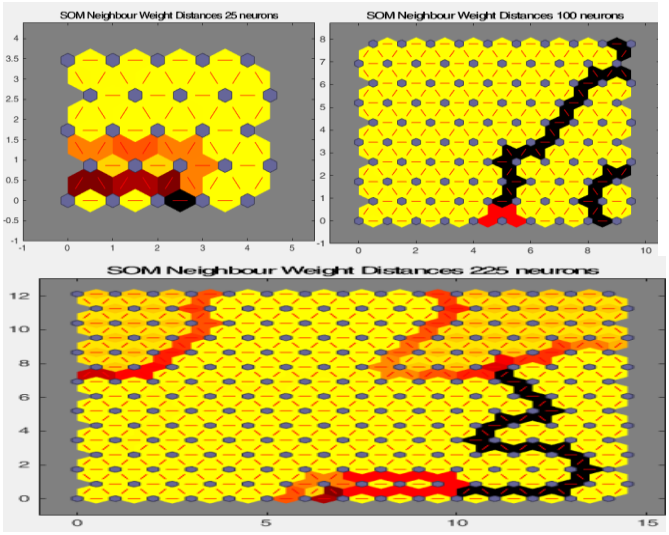

Figure 10: SOM U-Matrix with (top-left) 25, (top-right) 100 and (bottom) 225 neurons.

The SOM suggests that the ideal clustering should be 3 clusters this is hinted at with the U matrix of the 25 neuron network but is very clear in the 225 neuron network. Also, we observe that the silhouette value between k = 2 and k =3 only falls by 4.7% and is nowhere as dramatic as the 10% fall between k =2 and k= 5. To take this new information into account we will conduct the K-means with k = 3 for starters.

2.9.3 K-Means Clustering, k= 3

In our clustering analysis, we start with clustering the data into 3 clusters and then visualizing the output. Since the input data is high dimensional as it consists of 13 columns clearly it is not possible to visualize this a scatter plot. Therefore, to enable the visualization we will apply a Principal Component

decomposition of the clustering to reduce dimensionality to 2 dimensions to enable plotting.

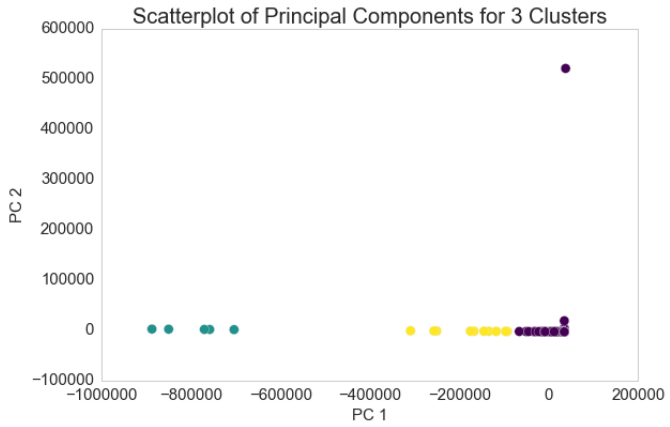

Figure 11: PCA Decomposition Scatter plot of 3 cluster solution

From the above figure we clearly see that the 3 cluster solution is a good solution. However, upon first inspection it would appear that the last two clusters represented in yellow and purple are not well separated hence could be merged as suggested by the silhouette plot. But once we look at the cluster member labels it will become obvious that the 3 cluster solution is a better and valid solution.

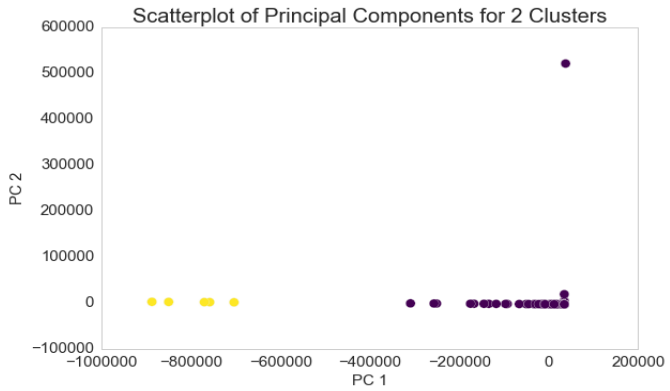

Figure 12: PCA Decomposition Scatter plot of 2 cluster solution

Cluster 0	Cluster 1	Cluster 2
Azerbaijan	Argentina	Belize
El Salvador	Armenia	Kiribati
Georgia	Aruba	Micronesia, Fed. Sts.
Jordan	Australia	Montenegro
Kosovo	Austria	Timor-Leste
Liberia	Bahamas, The	
Moldova	Bahrain	
Panama	Bangladesh	
Tajikistan	Barbados	
Tonga	Belarus	
	Belgium	
	Benin	
	Bermuda	
	Bhutan	
	Bolivia	
	Bosnia and Herzegovina	
	Botswana	
	Brazil	

Table 1: Cluster Memberships for K= 3 showing only selected members for cluster 2 and all members for cluster 1 and 3.

As expected using two clusters merges clusters 2 and 3 as shown in Fig 12. This was done in response to Fig 11.

If we use the 2 cluster solution we miss this detail hence the choice of K informed by the SOM is validated and is far more efficient than running multiple clustering's with different K values. But we have demonstrated both approaches and shown a better way of initializing K-means. It should be recalled that Argentina appeared among the top countries for scaled II indicator. Brazil also makes an appearance in cluster 1 for both of the clustering solutions so this additional validation of this solution as we can take these initial findings from the raw data to be ground truth for the our cluster evaluation.

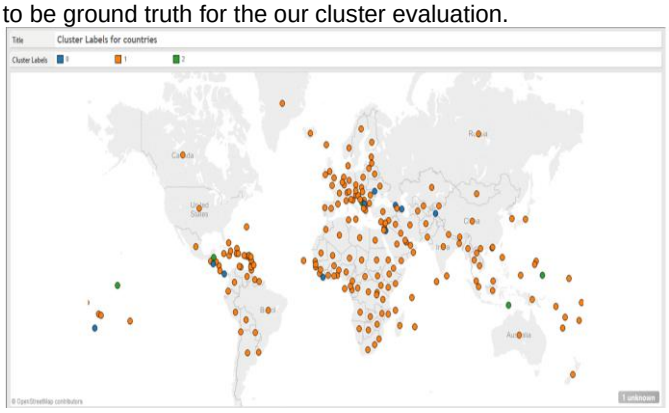

Figure 13: Map showing cluster assignment of countries

2.9.4 Cluster Group Means, K-means = 3

Now that we have explained and validated our choice of clustering we merged the cluster labels to the data to look cluster group means. The means of the attributes of each cluster are shown in the heat maps. This is a more convenient visualization of the attributes because we see that the clustering is dominated by the II/GDP attribute due to its scale.

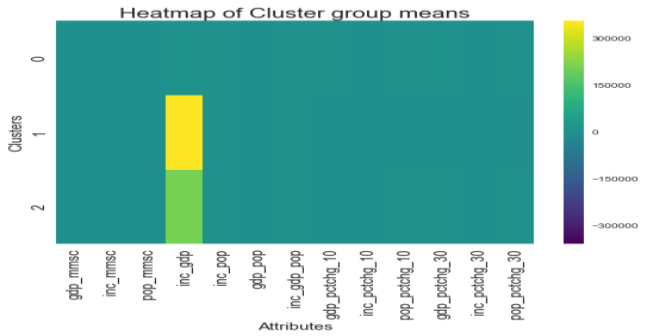

Figure 14: Heat map of cluster group means

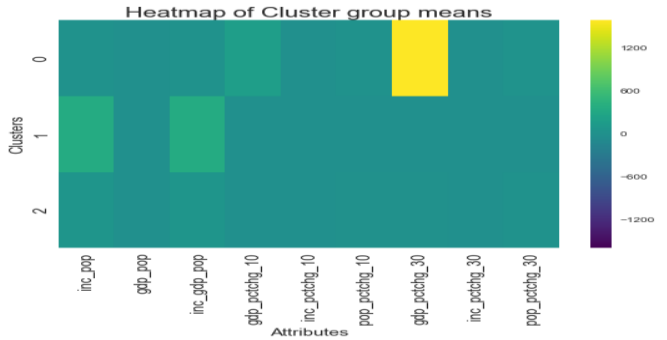

Figure 15: Heat map of cluster group means with II/GDP excluded

2.9.5 BIRCH Hierarchical Clustering

We also compare the K-means clustering with the BIRCH hierarchical clustering method. This method is chosen because it is a recursive partitional method that also be used to initialize K-means. But with this method there is no need to specify the cluster numbers as it infers them from the data. We find that the BIRCH method produces an almost identical result to the K-means and also finds 3 clusters.

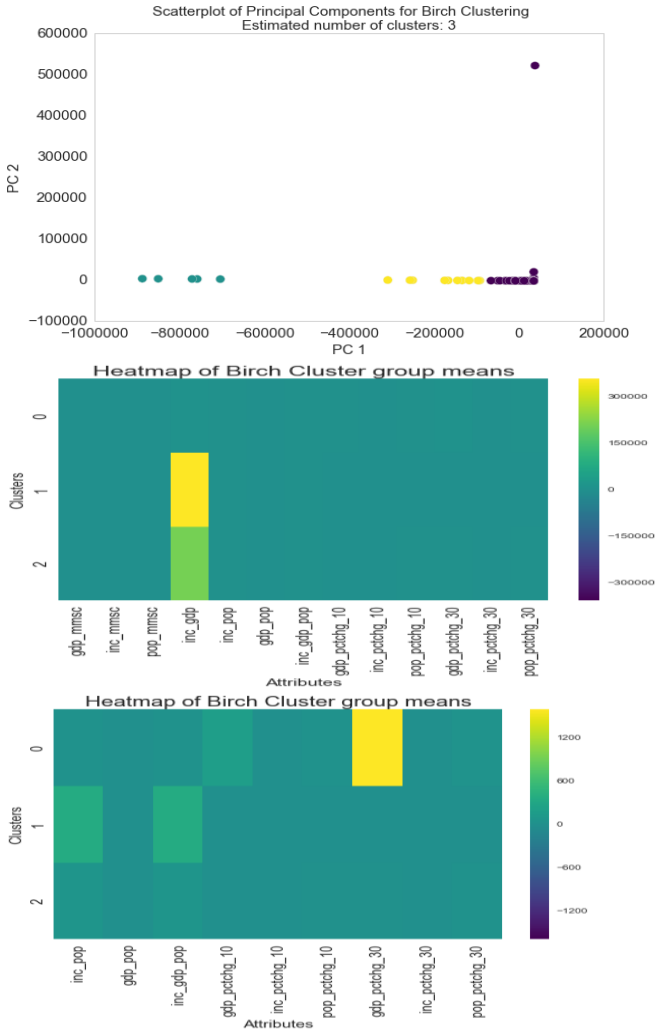

Figure 17: (top) PCA Scatter Plot of BIRCH clustering (mid) heat map of BIRCH cluster means and (bottom) with II/GDP excluded.

3 Discussion

In our discussion we will seek to answer the questions that we identified at the beginning.

3.1 What are the global trends in Income Inequality? How has this changed over time? Which countries are most affected?

To answer this we refer to dataset we generated by grouping the years and aggregating by the mean. This allows us to observe global trends across the indicators.

In Fig 17, we plot scaled II, II/(GDP+Pop), II/GDP, II/Pop, GDP and Population. The median is shown in black with the area between the upper and lower quartiles highlighted in grey.

It is clear that II has risen over the time period while GDP has steadily increased. The II trend shows that it was rising slowly for most part of the 30 year period but increased from 2000-2010. This could be due to greater implementation of neo-liberal economic policies. The II/(GDP+Pop), II/GDP and II/Pop attributes show that in this time period from 2006-2010 experienced greatest positive change above the median compared to the whole time series. This could be as a result of bailout of banks which could be seen as a transfer of wealth from state to private insttitions and people running them. The period between 2000-2010 is notable also because out of the whole time series the changes in the indicators are higher than the median and upper quartile values. The spike in 2006 is notable but its appearance 2 years prior to the Lehmans Brother crash in 2008 could be seen as a potential precursor.

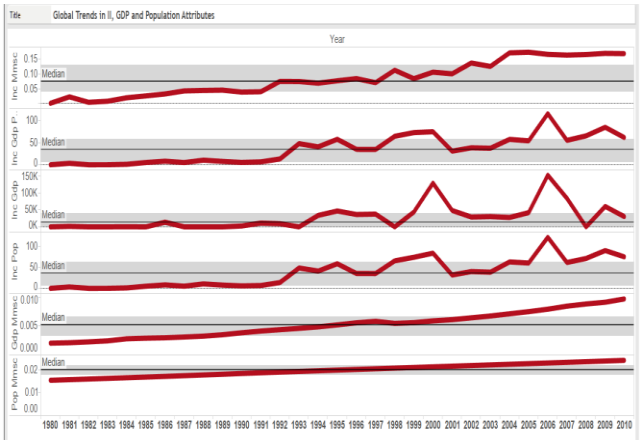

Figure 16: Plot of Scaled Attributes Mean Yearly Trends

3.2 Can we derive any additional attributes that allows to segregate this data further?

We have shown that additional attributes derived from the data is very effective in segmenting data and deriving additional insight. In this analysis we derived additional attributes such of II based on other indicators. These were derived from scaled attributes.

From the visual analysis using the combination maps and mosaic plots of the II/ (GDP+Pop) attributes we see that with different scaling we are able to gain slightly different insights. For example this attribute for the standard scaling shows that Argentina, Chile and South Africa of having the highest values while the robust scaling shows El Salvador, Macedonia and Honduras. The min max scaling for this attribute is most interesting because this attribute highlights wealthier countries such as Luxembourg and Iceland. This is first time in the analysis that these countries are encountered alongside lower income countries and tax havens such as Panama.

From the combination map with the standard scaling we observe that Spain, South Africa, Chile and Argentina stand out when we encode the II / (GDP + Pop) attribute with percentage change over 10 years for II averaged over the time period that Spain stands out as the nation with the greatest negative value while South Africa, Chile and Argentina have the most positive values out of the rest of the nations. This probably indicates that Spain underwent a great change in either its underlying population or economic dynamics or both which had an overall negative impact on Income Inequality in society. The combination map for this attribute with the robust scaling is less interesting because most values scale at or close to zero and

we do not observe much divergence. But Honduras and Kyrgyz Republic show the largest positive change in II indicating these countries also experienced rising income share asymmetry in society. The combination map for this attribute with the min max scaling is the most interesting as we start to see more divergence in the symbols. Here we see a lot of countries having percentage change at or close to zero. Most African nations appear to have negative changes in this attribute while there are some countries such as Estonia, Panama, Costa Rica and Uruguay among others that show a positive change in this attribute. This is surprising given that we are considering the same attribute but with different scaling. But the visual analysis in the form of a combination map which combines choropleth and proportional symbols are key to uncover this because this is not as obvious from the mosaic plots we presented but they are very good for providing a high level overview.

Perhaps more interesting is the fact where these plots agree and we start to see common countries among them even though their relative proportions for the attributes are different. We observe countries such as Iceland, Luxembourg alongside tax havens such as Panama and Seychelles which appear along BRIC countries such as Brazil. The unifying factor appears to be that even though they have experienced economic growth the inequality of the distribution is stronger in some sense here than the rest of the world.

3.3 What is the impact of data transformation on the final results?

We have clearly shown that the choice of data transformation has big impact on the final results. This is clearly seen from the cluster membership shown in Table 1. For standard and robust scaled data Belize does not show up as one of the top countries when using our II attributes. But the min max scaling extracts this country to be the top in terms of II as a proportion of GDP and Population. This shows the merit of derived attributes because we hardly observe any differences in II raw indicator among the different transformations. Since the min max scaled data is used for the K-means clustering Belize appears in a separate cluster with small island nations such as Kiribati and Micronesia etc. The Min Max scaling is ultimately chosen because it highlights tax havens such as Panama and Seychelles as having high II/ (GDP + Pop) values.

But future work could rightly explore other transformation types such as maximum absolute scaling and perform the clustering with different data transformations. The initial motivation for exploration of the effect of data transformation was to verify the findings of the min max scaling. This is another example of the visual analysis informing subsequent computations with the data being analyzed through different computational parameters.

3.4 How can clustering be used in this context?

The choice of K-means motivated by its ease of use and interpretability. This is because centroid based methods such as K-means optimize a well-known distance function. The SOM is a very powerful and robust technique but it is difficult to integrate the outputs into further analysis. The SOM network outputs weights after training and the U matrix shows us the clustering of the data. We use the SOM as a means of initializing the choice of K for the K-means.

In this study we showed the effective of a novel approach to K-means initialization through the use of the SOM. Using the structural approach it was suggested that the ideal clustering should be either 2 or 11. But the SOM shows very clearly the 3 distinct clusters of the data. Since this is a robust method that deals with outliers well we see that using this to initialize K-means leads to good results and considerably shortens the testing time of k-means.

Also we compared the K-means clustering to the BIRCH hierarchical clustering method. As the scatter plot suggests that both methods discover identical clustering of the data and the ideal number of clusters that BIRCH discovers is also 3. This further validates our choice of K. The cluster group statistics are also therefore not different from the K-means and we see that cluster 1,2 have high II/GDP values while cluster 0 have high percentage change in GDP over 30 years.

Partitional type methods are preferred as we want to group similar items and separate dissimilar ones. A Density based approach is not appropriate because our data has spatial meaning. Graph approaches such as Spectral Clustering [16] are good for detection of non-convex structures in data. But application of Spectral clustering to this dataset yielded 8 clusters a reason for this being that the matrix does not produce a fully connected graph which leads to some arbitrarily small clusters.

4 Further work

Further work could explore the clustering of the data with different transformations some of which have been tried here but not used for clustering such as robust and standard scaling. Other transformations such Maximum Absolute scaling could also be tried.

The World Bank Group has a wealth of other indicators available only the imagination is the limit as to what other indicators this analysis could be complemented by. But some pertinent ones to consider for future work might be health and education expenditure. The relevance of these variables to Income Inequality question can be explored and they can be used to derive attributes as we have done here and their effectiveness could be evaluated.

5 Conclusion

In conclusion, it can be said that we have successfully utilized the iterative data analysis process that combines computation and visualization as suggested by the Visual Analytics methodology. The visual feedback was used to inform computation and generate new lines of inquiry. For example our initial choice of scaling was visualized and then others were tried the visualization of which lead to interesting findings such as extraction of tax havens and lax tax jurisdiction areas such as Luxembourg. We proposed and validated a new method of K means initialization based on the visualization of the SOM U-Matrix. We did not encounter this method in our reading of the literature. We also tried a number of different types of clustering solutions on the data such as partition based K-means, artificial neural network based SOM and hierarchical based BIRCH. We compared the outputs of the K-means and BIRCH and found that they produce nearly identical results. We also derived cluster statistics for these two methods and found that cluster 0 had relatively low income countries, cluster 1 had most of the countries in the world and cluster 2 had the extreme cases such as very poor island nations.

6 References

[1] J. J. Thomas and K. a Cook, "Illuminating the path: The research and development agenda for visual analytics," *IEEE Computer Society*, 2005. [Online]. Available: http://vis.pnnl.gov/pdf/RD_Agenda_VisualAnalytics.pdf . [Accessed: 13-Apr-2016].

[2] P. M. Latha, "A Review on Clustering Techniques 1," vol. 11, no. 5, pp. 14–17, 2014.

[3] A. K. Mann and N. Kaur, "Review Paper on Clustering Techniques," vol. 13, no. 5, pp. 803–806, 2013.

[4] a. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[5] S. Roy and D. K. Bhattacharyya, "Data Clustering Techniques – A Review," *Rev. Lit. Arts Am.*, no. April, 2016.

[6] A. Zoss, "LibGuides: Introduction to Data Visualization: Visualization Types."

[7] "Cartographic Techniques : Color Hues."

[8] F. Pérez and B. E. Granger, "IPython: a System for Interactive Scientific Computing," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 21–29, May 2007.

[9] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, 2010, pp. 51–56.

[10] E. Jones, T. Oliphant, and P. Peterson, "SciPy: Open Source Scientific Tools for Python. Version: 0.16.0." 2015.

[11] M. Waskom, "Seaborn: statistical data visualization," 2012. [Online]. Available: http://stanford.edu/~mwaskom/software/seaborn/. [Accessed: 06-Apr-2016].

[12] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.

[13] F. Pedregosa and G. Varoquaux, "Scikit-learn: Machine Learning in Python," *J. Mach. …*, vol. 12, pp. 2825–2830, 2011.

[14] Tableau, "Business Intelligence and Analytics | Tableau Software." [Online]. Available: http://www.tableau.com/. [Accessed: 06-Apr-2016].

[15] "MATLAB - MathWorks - MathWorks United Kingdom." [Online]. Available: http://uk.mathworks.com/products/matlab/. [Accessed: 13-Apr-2016].

[16] M. Planck and U. Von Luxburg, "A Tutorial on Spectral Clustering A Tutorial on Spectral Clustering," *Stat. Comput.*, vol. 17, no. March, pp. 395–416, 2006.

7 Table of Figures