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.

**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?
* 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

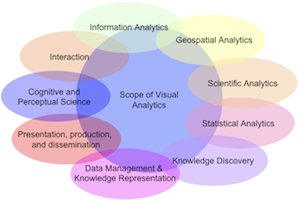
In addition we also aim to show that the Visual Analytics methodology is a useful approach for the analysis of this dataset as the data has spatio-temporal component because it is related to countries and we are considering a time series of the attributes of interest.

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.

# literature review

## 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]



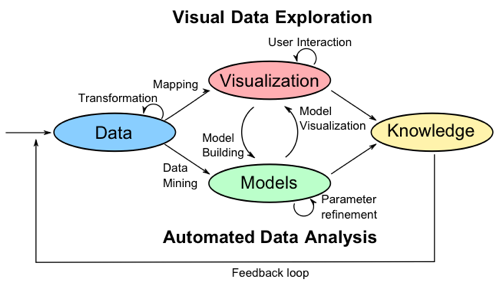


Figure : (Top) shows the scope of Visual Analytics to be a bridge between different fields of study. (Bottom) The VA methodology exemplified.[1]

## 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.

### 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.



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 (l\*k\*n) where l 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]

### 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]

## 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. Caartograms 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.

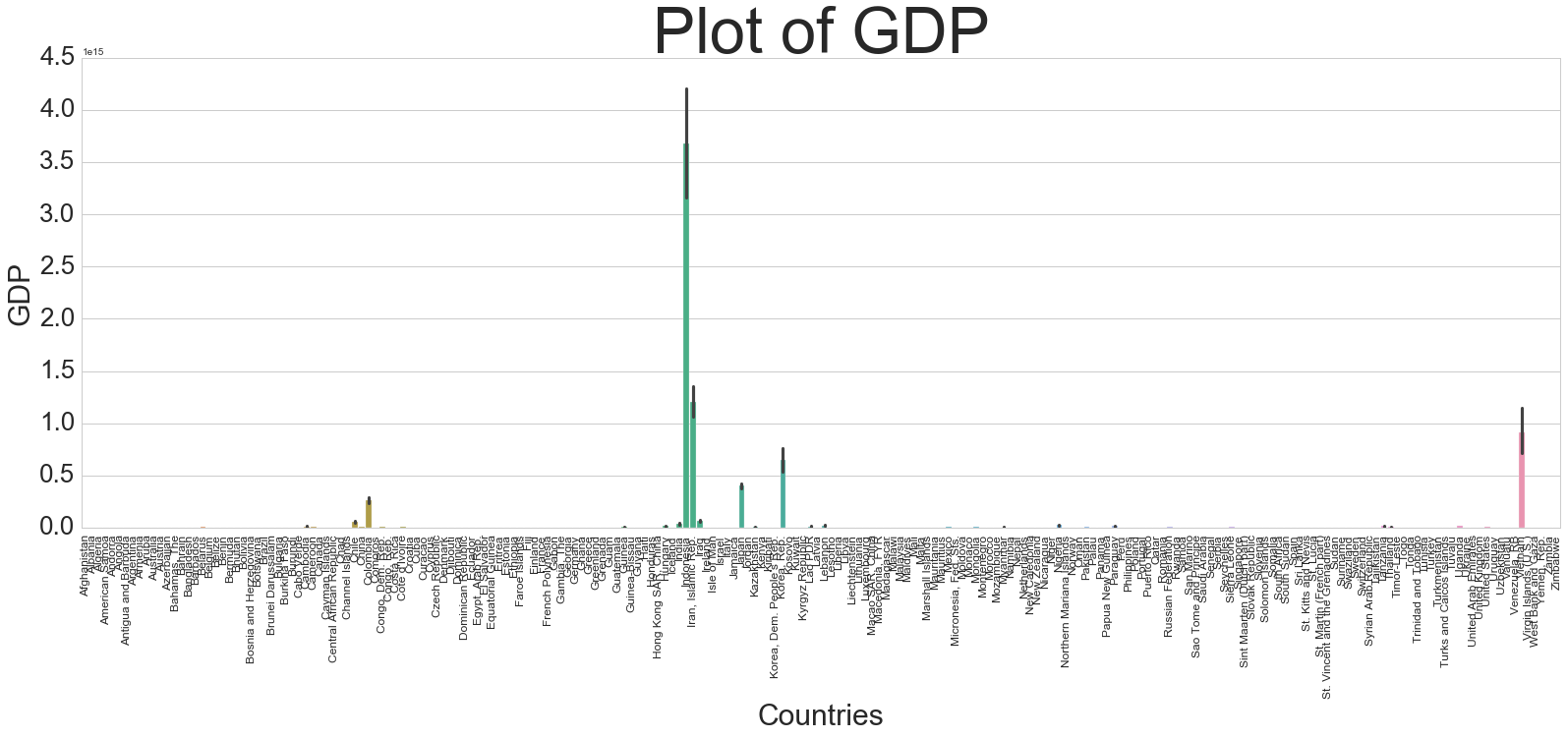
The colors can be inferred from the values being visualised and depending on the nature of the attribute the colors chosen can be harmonic, contrastive or clashing. [7]



Figure : (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.

# Data analysis and Results

## 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.

## 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.

## 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.

## 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.

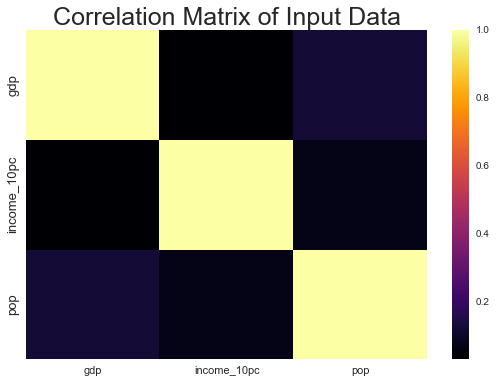


Figure : Pearson Correlation Coefficient among indicators in the input data before scaling

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.

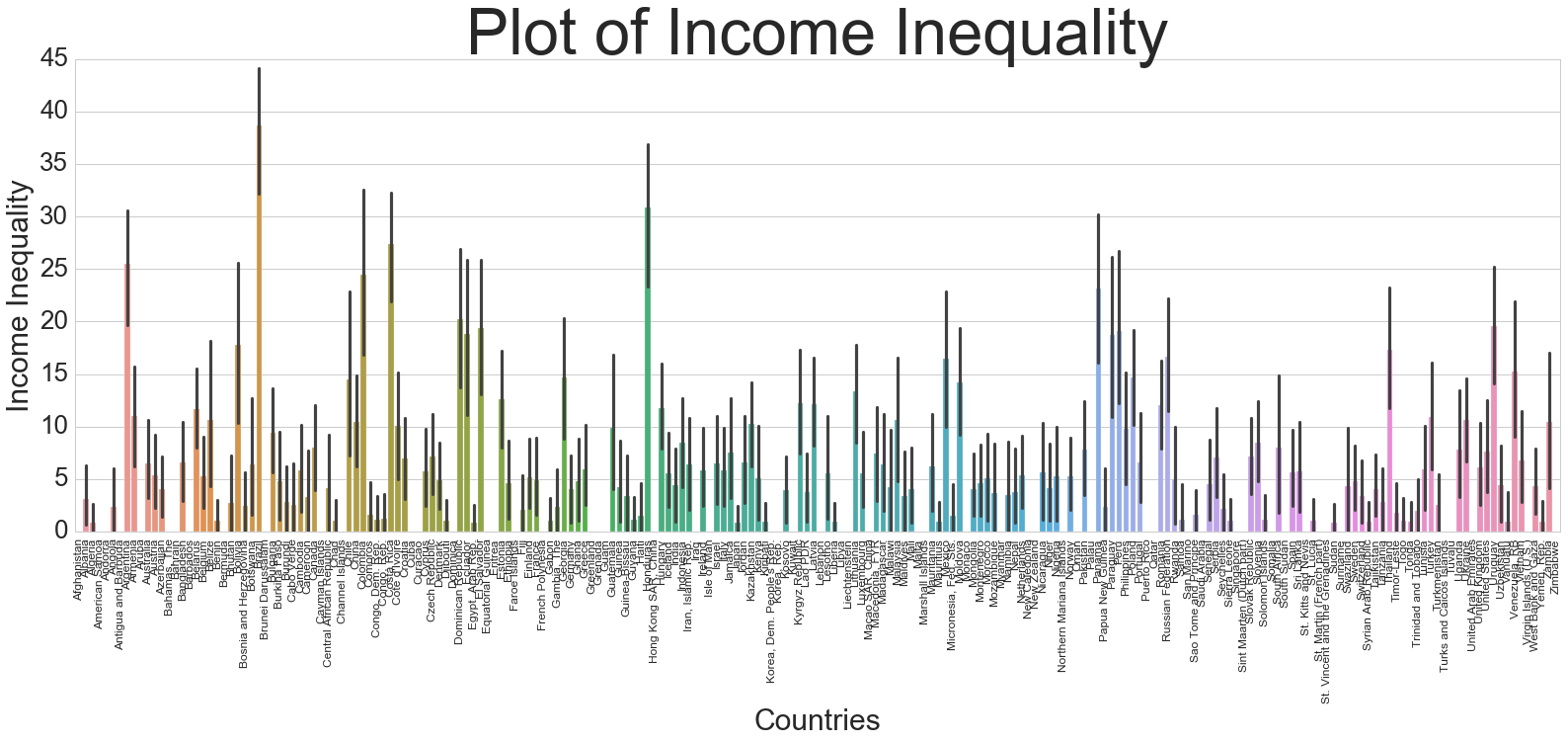
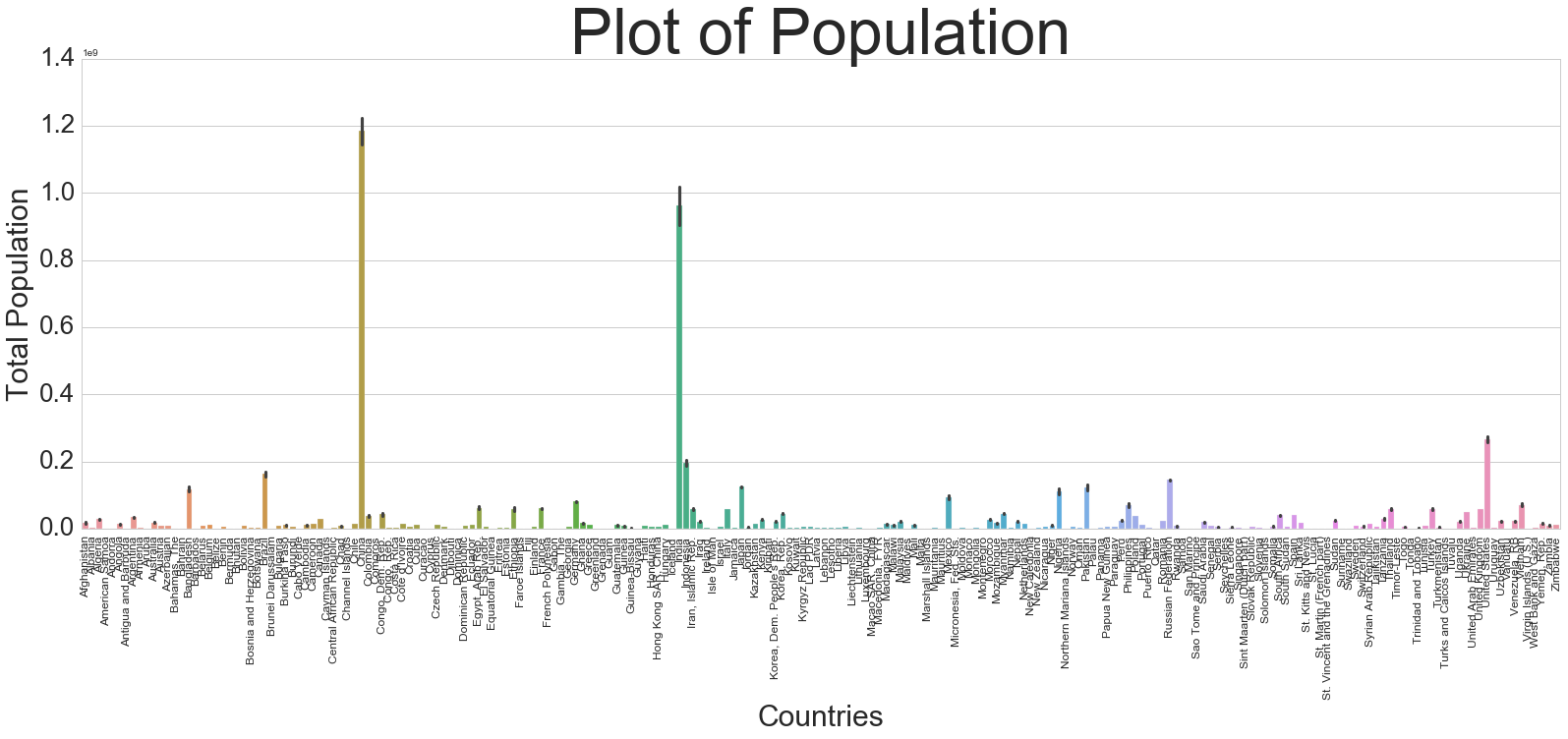


Figure : Bar plot of raw indicators after data loading and filling with zeros. We observe that there is a huge variation in scale among the indicators with GDP and Population being many magnitudes larger than the II variable. This shows the necessity of data transformation in this case.

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.

## 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.

## 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.

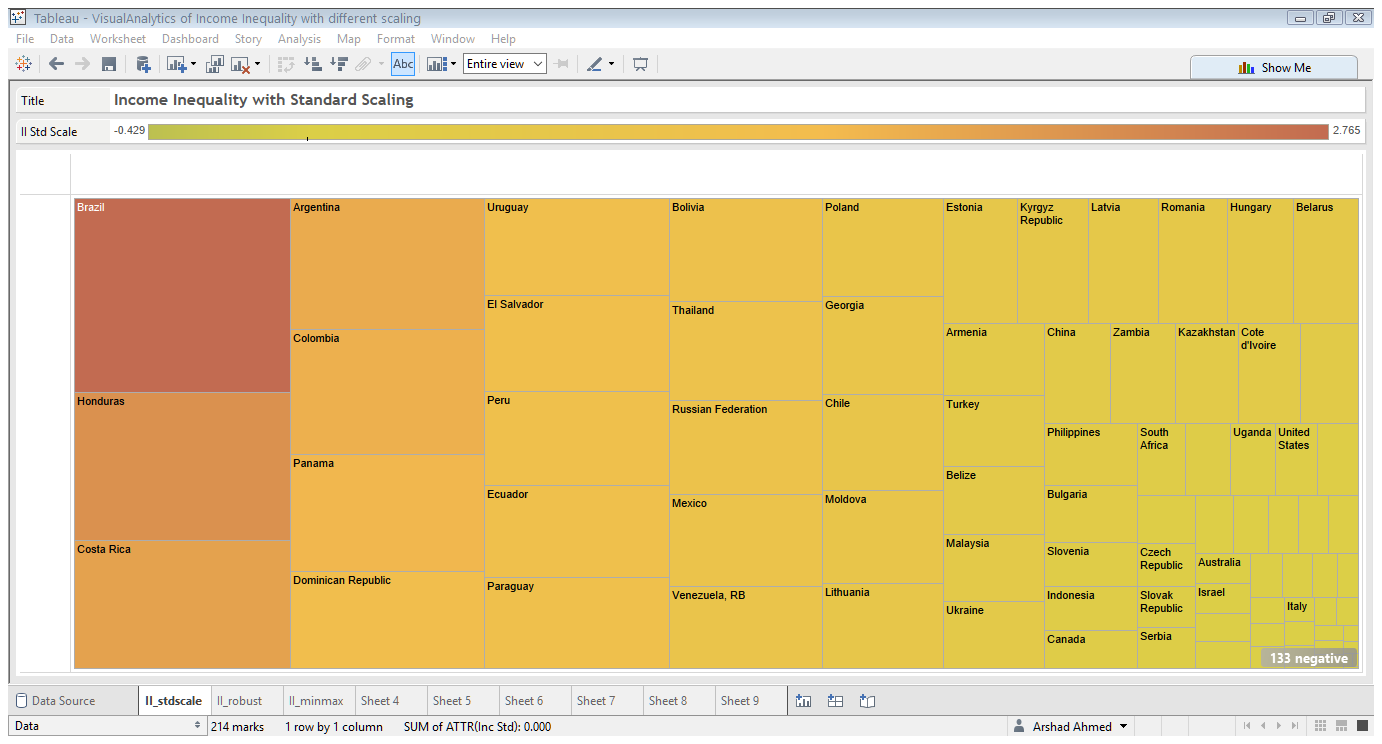
## Data Merging and Aggregation

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].

## 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.



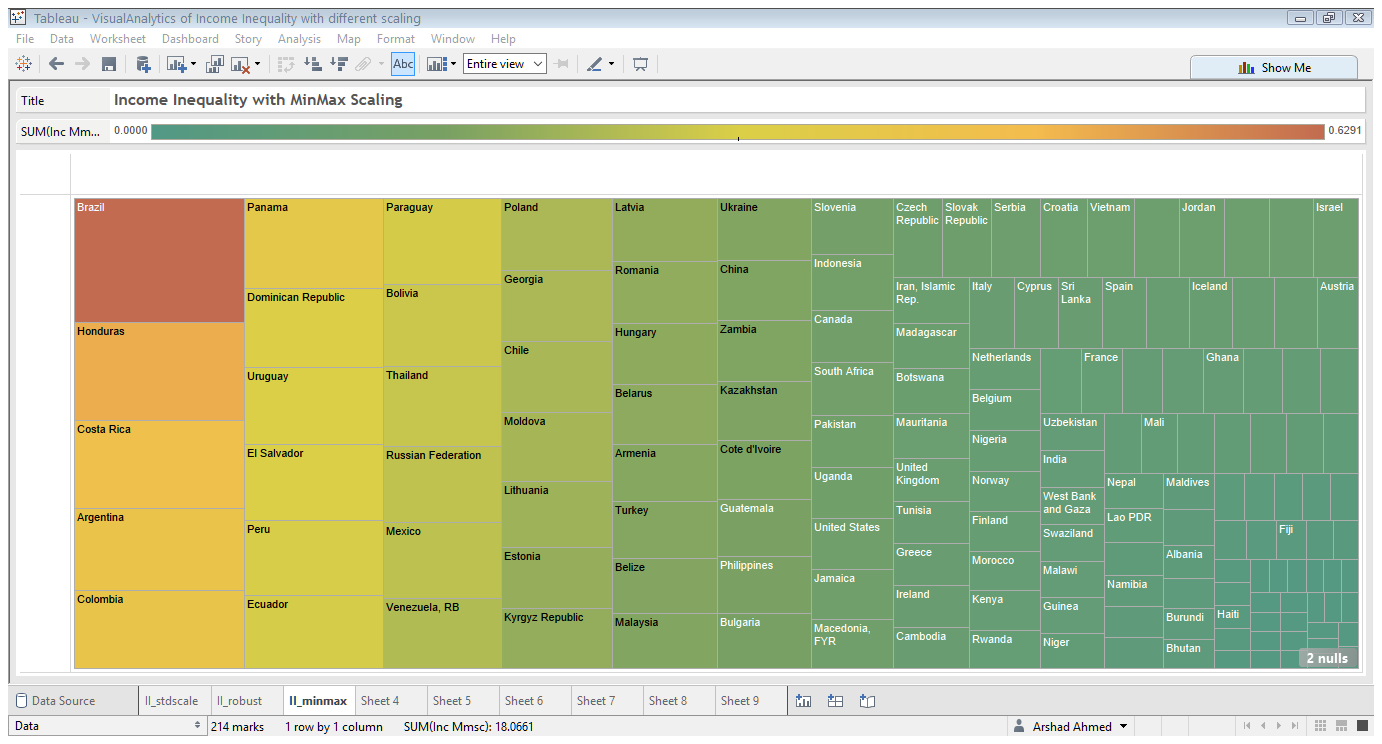
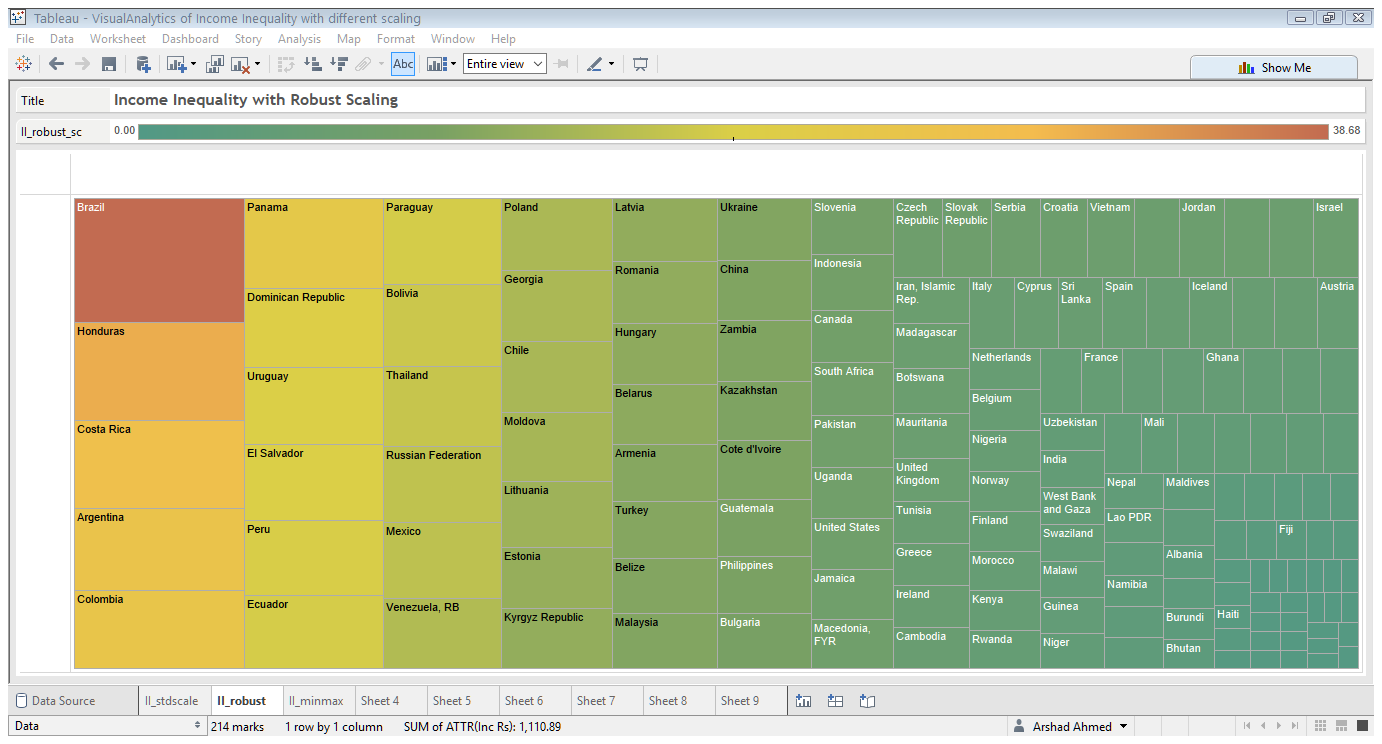
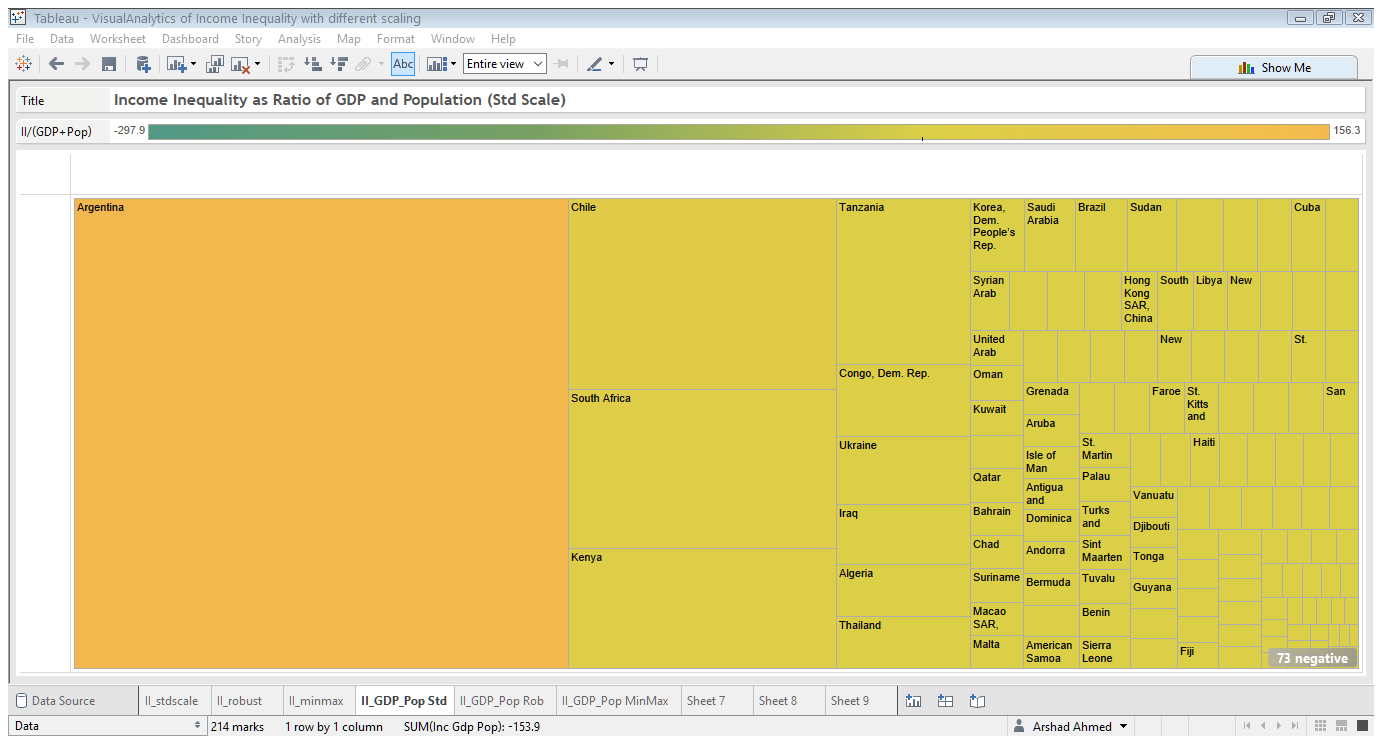
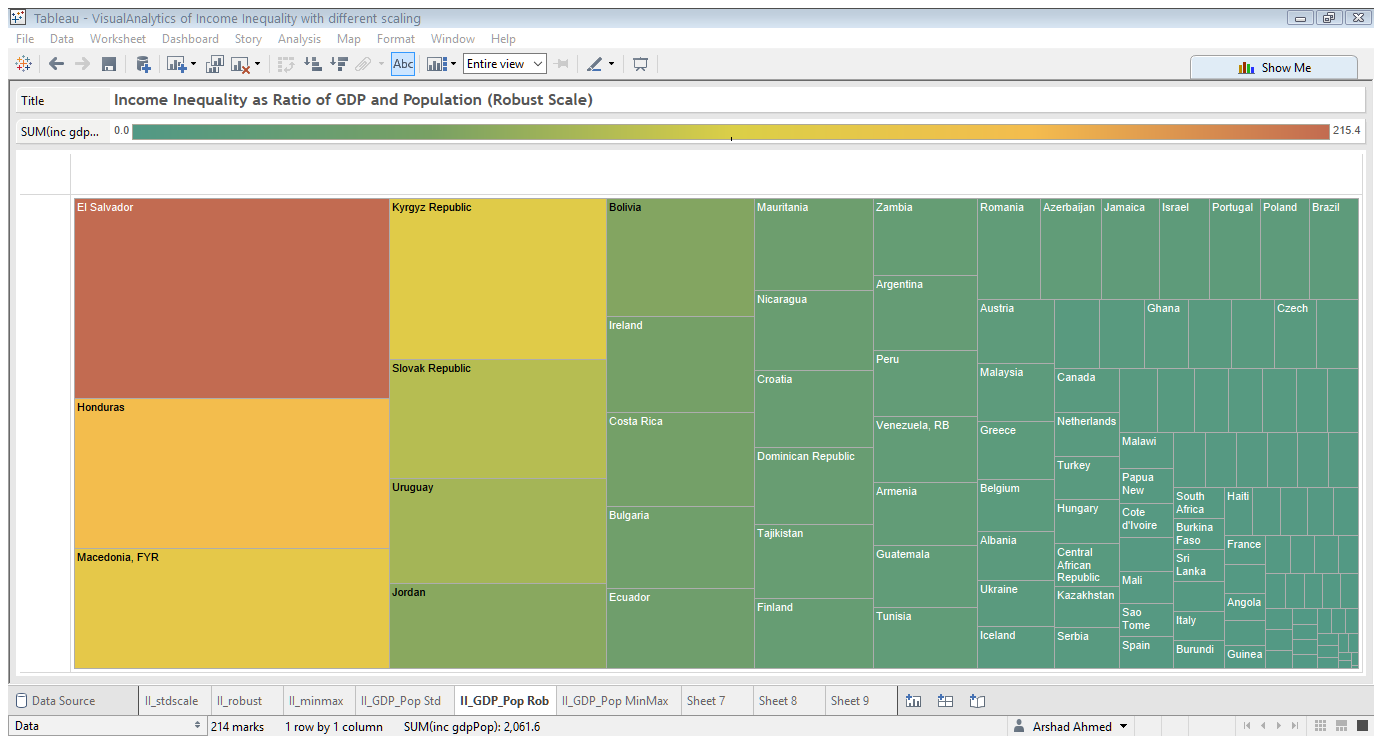


Figure : Mosaic Plot of II variable with (top) Standard Scaling (middle) robust scaling and (bottom) Min Max Scaling.

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.





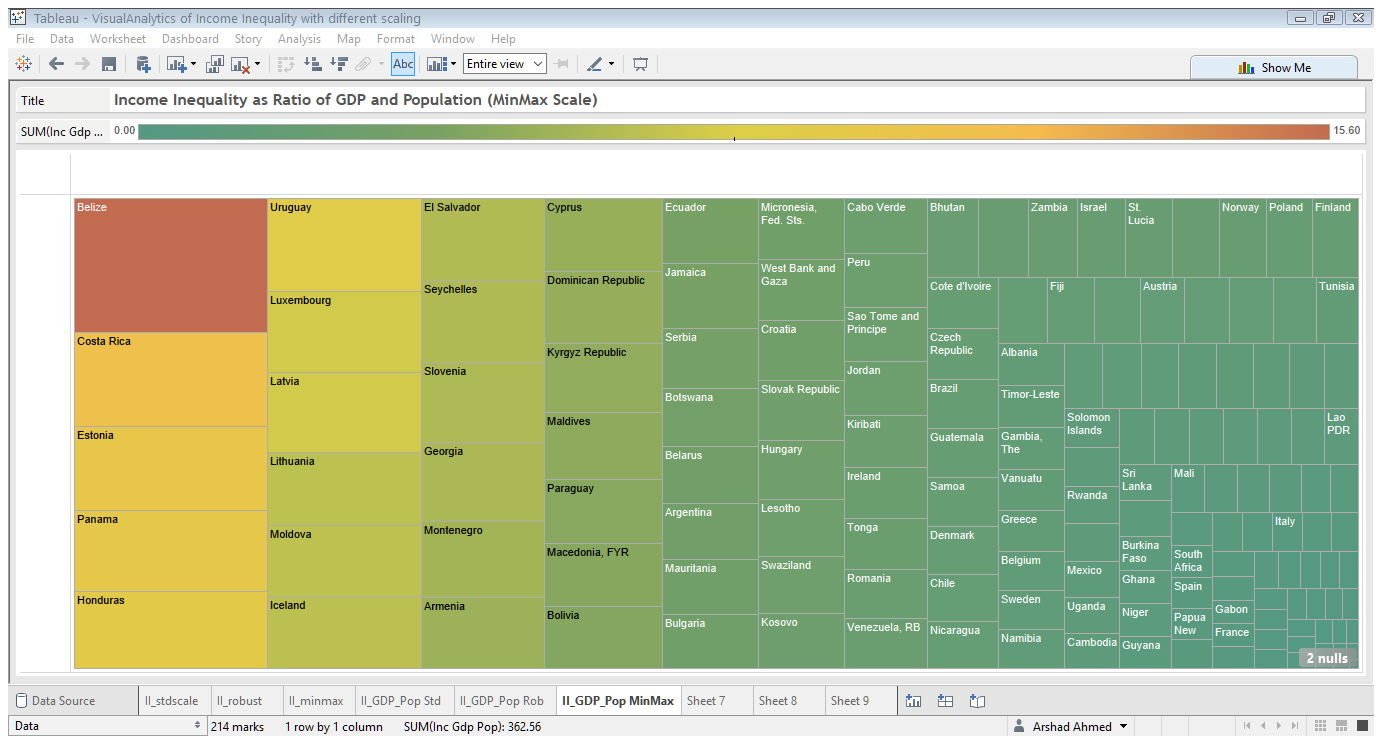
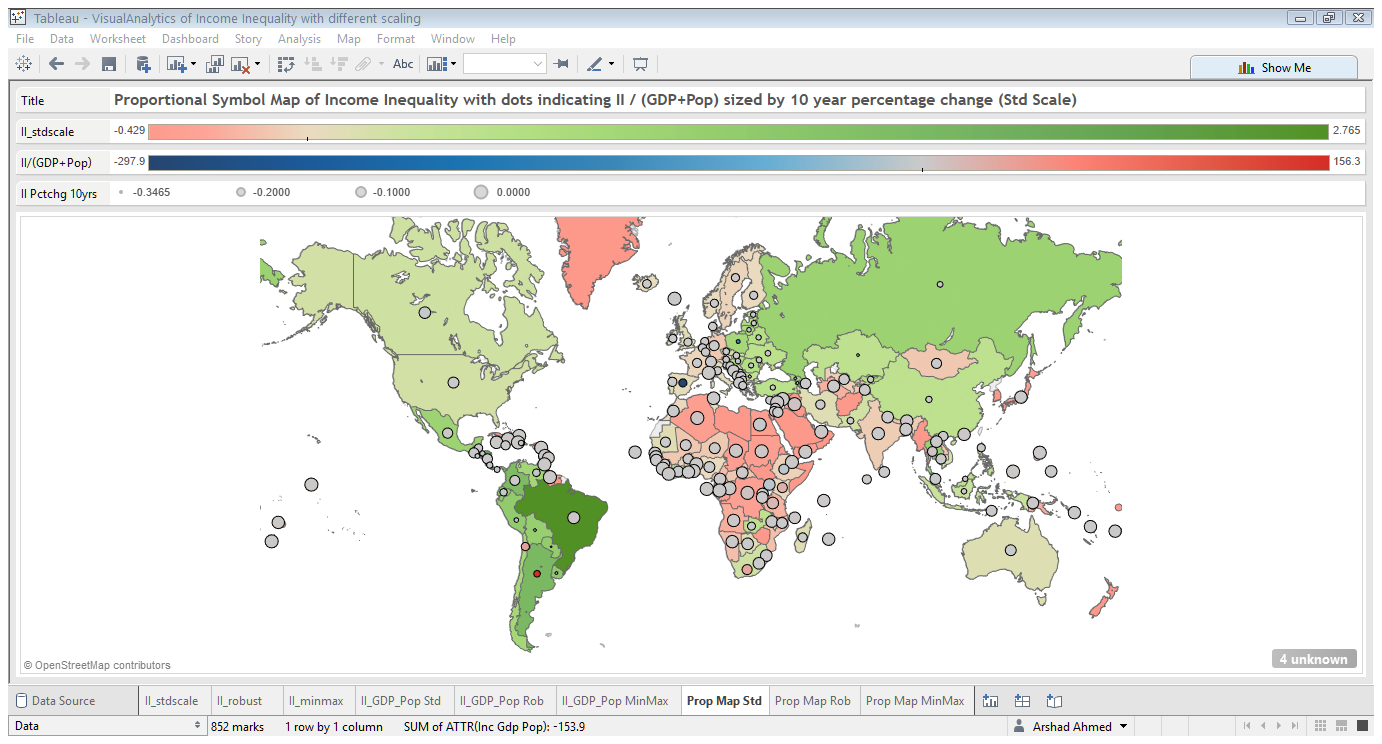
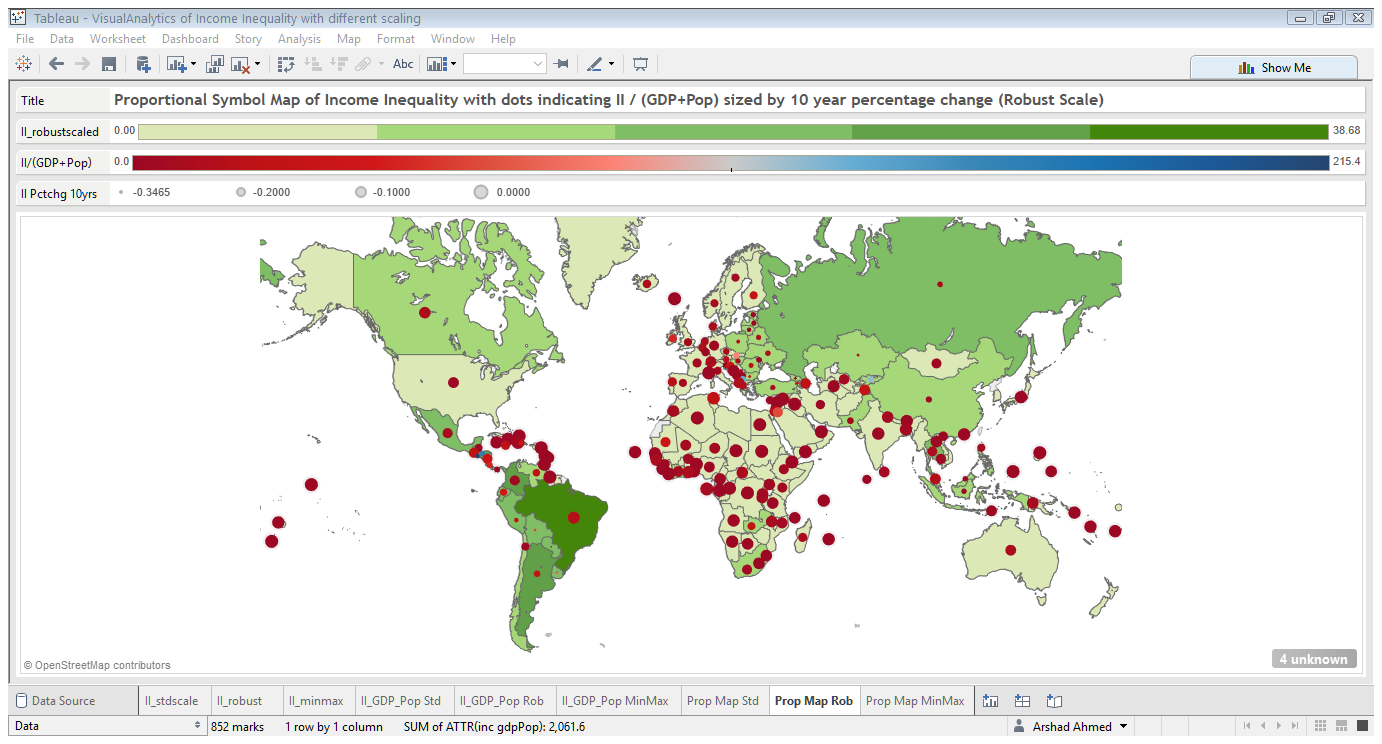


Figure : Mosaic Plot of II/(GDP +Pop) with (top) Standard Scaling (middle) robust scaling and (bottom) Min Max Scaling.

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.





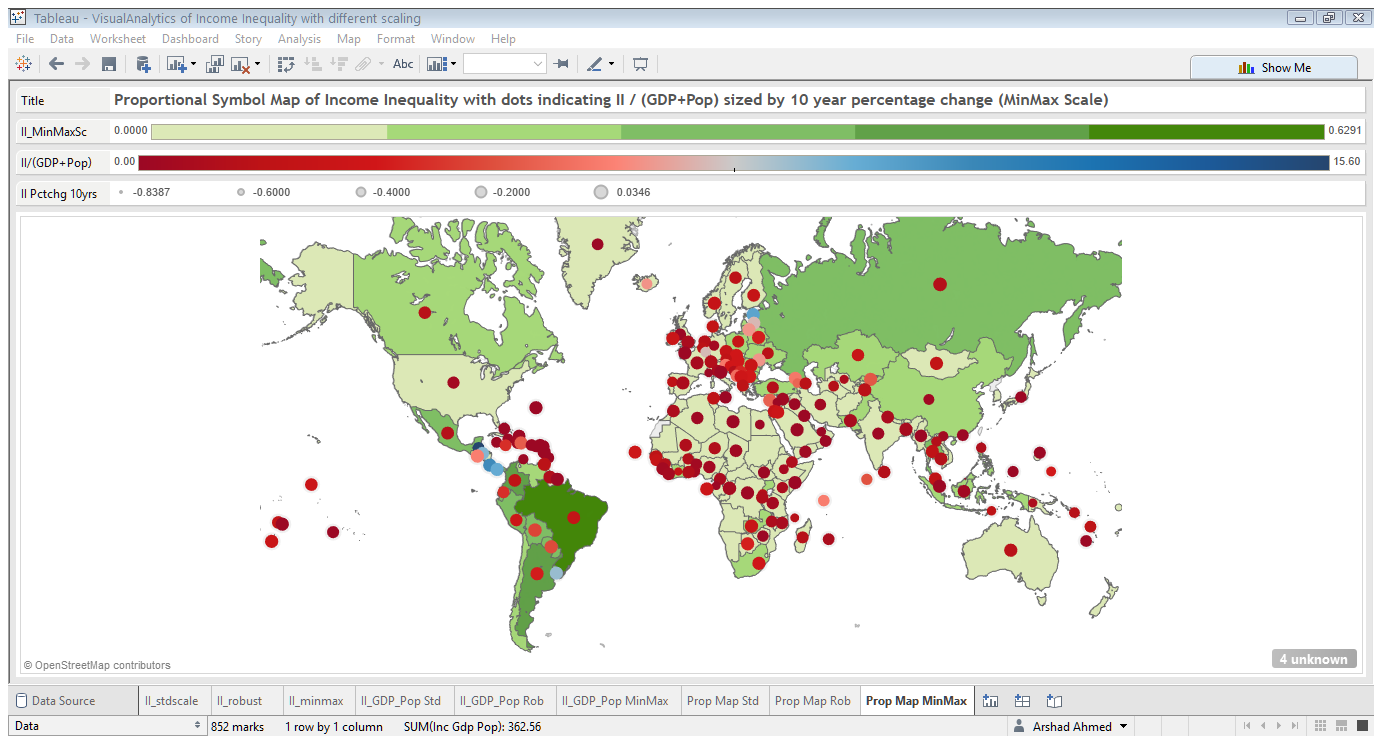
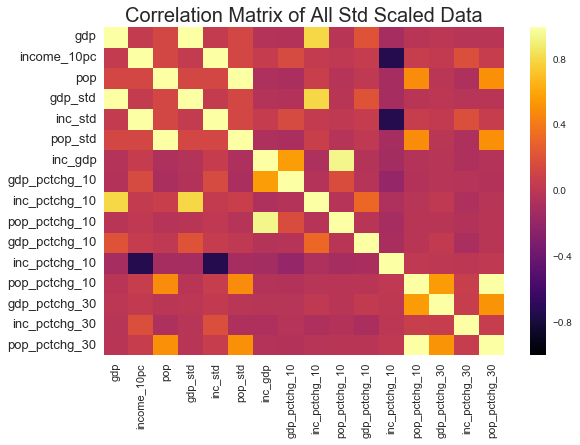


Figure : A combination map which shows the countries shaded using a diverging color scheme using the scaled II attribute. The circles are proportional symbols which are colored by II/ (GDP + Pop) attribute and the size is encoded using the 10 year percentage change in the II variable. A diverging color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



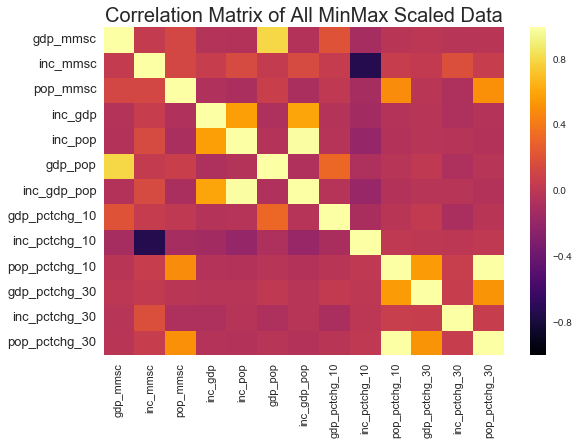
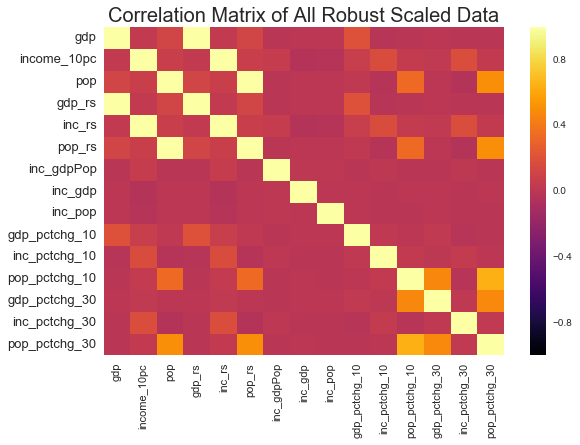


Figure : Pearson Correlation Matrices for aggregated datasets containing all attributes with (top) Standard Scaling (middle) robust scaling and (bottom) Min Max Scaling. The Standard and Min Max scaling produces similar correlations among variables while the Robust Scaling produces a slightly different 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.

## Cluster Analysis

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

The aim of the cluster analysis firstly is to determine the optimal number of clustering of the data. 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.

### 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.

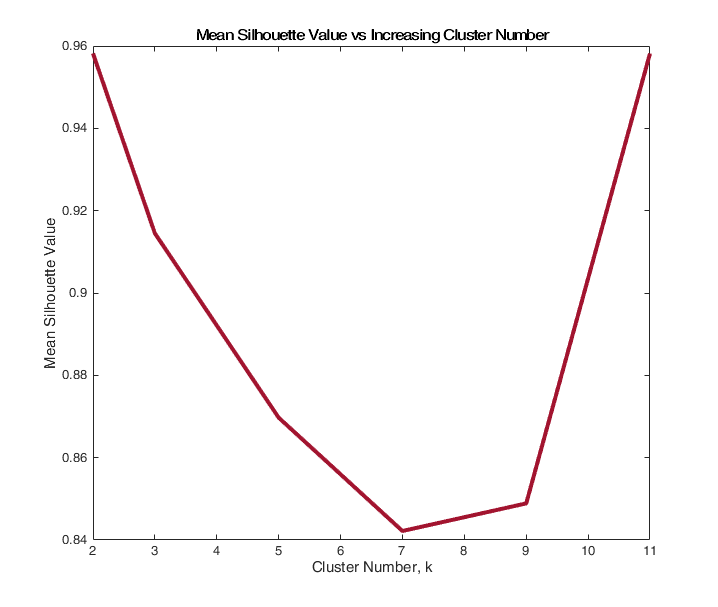


Figure : Mean Silhouette Value for K-means with increasing K-values.

### SOM Clustering

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

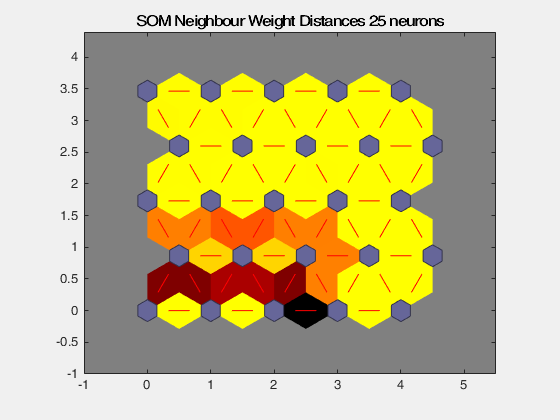
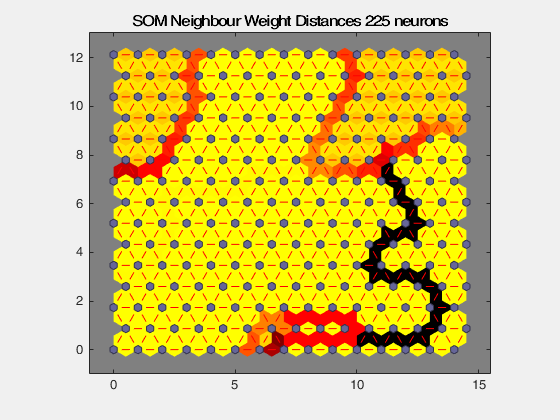


Figure : SOM U-Matrix with (top-left) 25, (top-right) 100 and (bottom) 225 neurons.

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.

### 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.

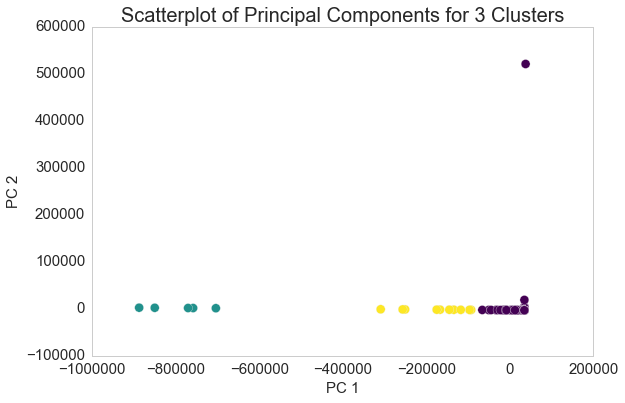


Figure : PCA Decomposition Scatter plot of 3 cluster solution

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.

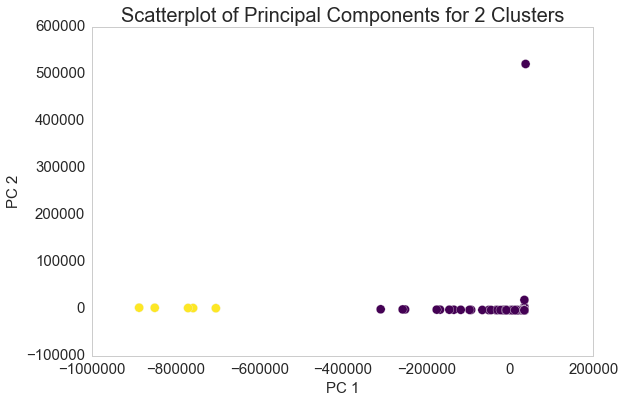


Figure : PCA Decomposition Scatter plot of 2 cluster solution

|  |  |  |
| --- | --- | --- |
| **Cluster 1** | **Cluster 2** | **Cluster 3** |
| 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 : Cluster Memberships for K=2 and 3 showing only first 10 items.

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

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.

### Cluster Group Means

Now that we have explained and validated our choice of clustering we merged the cluster labels to the data to look cluster group means.

| **K** | **GDP** | **II** | **Pop** | **II/GDP** | **II/Pop** | **II/GDP+Pop** |
| --- | --- | --- | --- | --- | --- | --- |
| **1** | 5.553083e-03 | 0.081683 | 0.021133 | 6127.132101 | 27.724753 | 25.313437 |
| **2** | 7.155142e-08 | 0.063421 | 0.000272 | 828932.147330 | 361.464635 | 361.155704 |
| **3** | 8.246970e-07 | 0.149418 | 0.002855 | 209206.323204 | 70.769338 | 70.715097 |

Table : Cluster group means of selected attributes

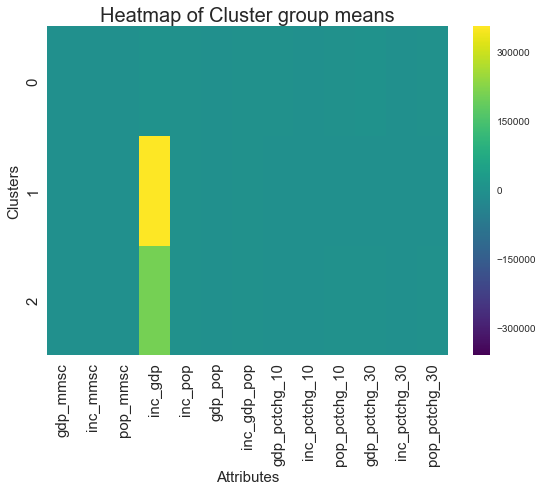


Figure : Heat map of cluster group means

From the table above we can say that the clustering was strongly influenced by the II/GDP attribute due to its scale hence the clusters we observe are those where II/GDP is high and medium. This is because cluster 2 and 3 have mean of II/GDP 2 orders of magnitude higher than cluster 1.

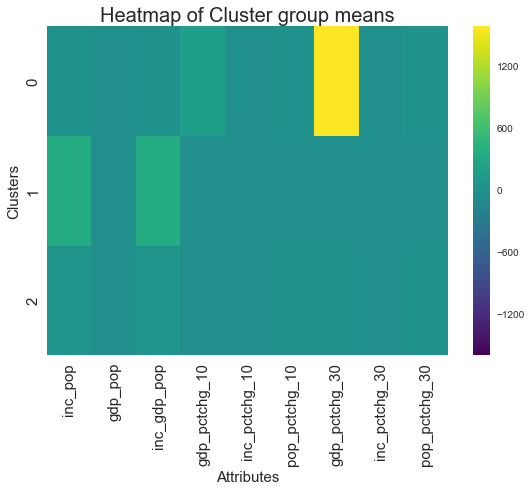


Figure : Heat map of cluster group means with II/GDP excluded

Incorporating the feedback from Fig 13, we see that Fig 14 shows that cluster 1 also has the greatest percentage change in income over the 30 year period in II as opposed to the other clusters. Also for cluster 2, II/Pop and II/(GDP + Pop) is higher than the other clusters. Therefore these countries experienced greater share of the wealth going to top 10% as a proportion of their GDP and Population compared to others.

# Discussion

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

## What are the global trends in Income Inequality? How has this changed over time?

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.

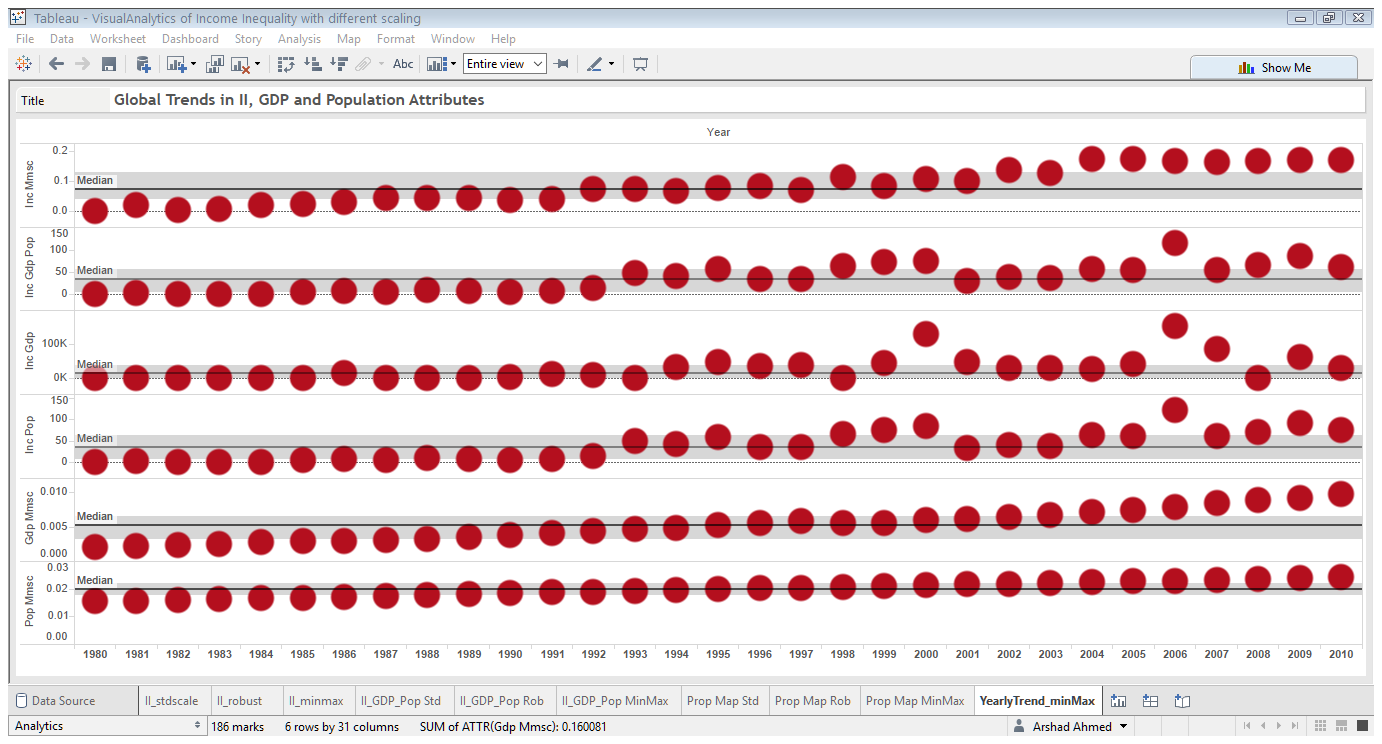


Figure : Yearly trends in II attributes, GDP and Population.

In Fig 15, we plot scaled II, II/(GDP+Pop), II/GDP, II/Pop, GDP and Population. The median is shown in black with 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 instittions and people running them.

## 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. As the mosaic plots in Fig 6, 7 and the clustering shows these are very effective in additional segmentation of the data. Also the conclusions derived from them are refinements of what we observed with the initial II indicator.

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

## How can clustering be used in this context?

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].