INTRODUCTION

The datasets chosen for this exercise is from the World Bank specifically their World Development Indicator Data[1], Gender Statistics Data[2] and their Debt Statistics Data[3]. These datasets cover a huge range of socio economic indicators for countries at the macro level and are in the form of a time series from the 60's to the present time.

The aim was then to explore the relationship between Fertility Rates and wider macroeconomic indicators. Also I wanted to determine the most important feature set from the attributes calculated from that would enable the prediction of the median fertility rate for countries. The reason that median measure is chosen is because it is robust to outliers and wold therefore serve to reduce noise in the target value in our predictive modelling.

The motivation for exploring and predicting fertility rates is because this is a good indicator of demographic changes and potential problems related to it. So low fertility rates in countries such as Japan are well known and have caused a situation where there is a rapidly greying population with insufficient replacement leading to huge pressure on the economy, health system and by extension government expenditure. Such an inverse pyramid leads to questions such as how will care for such populations be funded? Where will the workers of the future come from? If there are insufficient workers what sort of impact would this have on the wider economy? Although the later questions are interesting but are not addressed here. The underlying cause for such questions such as the fertility rate are explored. As it would help to understand and thus enable for appropriate policy decisions to be made. The data processing flow is outlined in **Table 2.**

However, most of the time there are a lot of blanks in the data and the key challenge firstly was to pick interesting indicators that had sufficient amount of data. Initial exploration suggested that considering the time range from 2000 – 2013 would yield the most complete data for the set of chosen indicators shown in **Table 1** below.

The way this was done was by extracting each indicator separately from the source files into Python and then calculating the percentage of blank to non-blank values. The indicators chosen have typically less than 10% of the data missing. This was the most acceptable compromise as most other indicators had missing information in the range of 80% for the given time range. Thus considering such variable would introduce a high degree of bias towards variables that had more data and hence would not need to be filled with zeros. Also filling such variables with a large number of zeros would not lead to useful features.

Table 1: Selected indicators for analysis and source

Indicator Name	Source
Fertility Rate	Gender Statistics
Total Percentage of Labour Force that is female	Gender Statistics
GDP (current US$)	Word Development Indicators
GDP (current US$)	Word Development Indicators
GDP per capita, PPP (current international $)	Word Development Indicators
Health expenditure, total (% of GDP)	Word Development Indicators
Consumer price index (2010 = 100) (CPI)	Word Development Indicators
Population, total	Word Development Indicators
Gross National Income (GNI) (current US$)	Word Development Indicators
Total debt service (% of exports of goods, services and primary income) ->(Total Debt)	Word Development Indicators/Debt Statistics

The IMF databases mentioned collect over 480 indicators from all countries and not all of them are relevant for this purpose. There was a preference for broad macroeconomic and social indicators which are easy to understand without much domain knowledge.

I wanted to investigate whether using summary statistics can be used in place of the whole time series and only a few of these variables can then be used to predict our value of interest then we have successfully reduced our potentially big data problem to a smaller data problem.

ANALYSIS METHODOLOGY

The data analysis process started first seeing the data in Excel after downloading from the IMF site to see who it was structured. The data is structured by Countries with all their associated indicators and then the time series data. Based on the selection of the date range for this analysis from 2000-2013 I created a subset of this data with just this time range. This was read into Python using PANDAS[4] and all instances of the Indicators from this data for all the countries were extracted. This resulted in 10 data frame objects with all countries and a single indicator. After this I calculated completeness of the data as a percentage between complete and blank values. In addition to checking that all the indicators had Countries starting from Afghanistan to Zimbabwe giving 248 in total and the relevant time range. It was necessary to exclude aggregates as they would skew the data. This was done at the initial stage to prevent confusion later in the analysis. This then allowed easy calculation of the row statistics outlined in **Table 2** Step 1 and the column statistics shown in **Fig 6**.

The column statistics are considered separately. This is shown in the pair plot in **Fig 6**. Although **Fig 6**, is hard to use due to the large numbers of indicators considered. But it serves as a useful presentation of some key insights from the data. It also helps to contextualise the correlation and cluster maps. We see that for our attribute of interest which is the fertility rate the mean fertility rate is negatively correlated to almost all the indicators considered apart from Total Debt of a country.

Table 2: Shows Data Analysis flow and key decisions made and methods used.

Step 1: Extraction, Transform and Load	Step 2: Pre-processing	Step 3: Data Fusion	Step 4: Modelling & Visualization
1. Extract selected indicators from the relevant sources 2. Check amount of blank data if >10% pick a different indicator 3. Fill blank values with 0 4. Separate individual indicators for further processing 5. Delete first 34 rows as they consist of aggregations of the data such for OECD, Developing countries, only country wise data is considered 6. Calculate country wise statistics from the time series such a: • Mean • Median, • Standard deviation, • Interquartile Range • % Change 1 Year • % Change 5 Year, • % Change 10 Year • % Change 13 Year	7. Calculate column statistics for the indicators for the given time range 8. Data shows a variety in scale between indicators so test scaling methods maximum absolute scaling, min max scaling and standard scaling. 9. Produce IQR/Median plots with different scaling and choose best one 10. Produce columns statistics for time range and explore trends using [5]	11. Merge data based on Country Name and Country Code 12. Apply Maximum Absolute Scaling (MaxAbs) 13. From merged data calculate some additional features of interest such as means ratios of • Debt to GNI , • Population to GNI • Population to Labour Force %Female • CPI to Fertility Rate • GDP to GD per Capita • Health Expenditure to GDP • Health Expenditure to GDP growth • Health Expenditure to Fertility Rate • Health Expenditure to Population • Fertility Rate to GDP growth • Fertility Rate to GDP per Capita • Fertility Rate to Labour Force %Female • Fertility Rate to CPI 14. Calculate 'Ahmed Score' for countries based on the above features	15. The dataset is high dimensional with over 90 columns which makes visualization difficult therefore dimensionality reduction methods Tested the following: – PCA – ICA – FA – NMF – LLE – T-SNE 16. Assess Dimensionality Reduction methods by looking at explained variance, reconstruction error, and plots. 17. Use Gradient Boosting Regressor for feature selection 18. Use Gradient Boosting Regressor, Random Forest Regressor and Extra Trees Regressor with and without dimensionality reduction to predict median Fertility Rates and report Mean Square Errors (MSE) using [6]. 19. Use final data with all calculated statistics in previous steps and visualise in Tableau.

Since my aim is to predict median fertility rates I wanted to do so using summary statistics which reduce dimensionality of the data. Also it would be useful going forward as new data becomes available instead of having to deal with a large volume of data we have a means to compress the data and get our answer.

Next step was to merge the data using the Country Name and Country Code column and again checking the length, the names of the countries, and completeness of the data. It is important to do such Quality Control (QC) at each stage in processing to catch potential problems before progressing further in the analysis. After merging I saved the output as an Excel file to have a look the data, which showed that some columns had 'infinities' which are not caught by the PANDAS 'isnull' function so I used NUMPY[7] 'is.inf' and 'is.nan' functions to find the remaining problems in the data and fill them with zeros before progressing further. It is important to note that this would not have been as obvious from a dump of the data on the python console and this highlights the importance of having a variety of tools and leveraging their strengths in analysis. These extra blanks were caused by trying to calculate percentage values for example where one of entries was zero.

Then I investigated the need for scaling of data as shown in **Fig 3** and decided in this case the maximum absolute scaling worked best. The next step was to apply this scaling to the data and calculate the measure highlighted in **Table 2** Step 3. The reason I only calculated ratio of means is because there was not much variation when these measures were calculated with other statistics. Also, I designed the 'Ahmed Score' to combine these features into one score that could be used for ranking and visualization.

High dimensional datasets by their nature are difficult to visualize, so a useful method to visualize this dataset was by calculating the column IQR and Median and plotting them as scatterplot. This effectively compresses the dimensions into a point using robust statistics which is then easier to visualise. **Fig 3** shows the IQR-Median plot of data and highlights the need for data scaling. This is because we see a few points dominating and the rest of the dimensions getting drowned out. To make the most of the data available and to make them comparable it was necessary to scale them shown in **Fig 3.** Out of all the 3 methods tested the Max Abs Scaler is chosen as it exposes structures in the data not seen as clearly with the other two methods. It also highlights outliers such as Zimbabwe and Sao Tome et Principe which score poorly on most indicators considered.

Fig 5 shows a cluster map of the data which shows the detailed dependencies inferred from the data. This is to be expected as the correlation matrix heat maps in **Fig 4 and 9** show a high level of correlation in both the scaled and unscaled data. The correlation is invariant with the choice of scaling and here serves as a good quality control measure.

I wanted to be able to come with a single measure that would capture the relative importance of the means ratio measures calculated from the satistics (Table 2). The reason the Means Ratio of the indicators is used and not the others was because upon calculating the same measures with the other statsitics such as median there was no additional variance gained and they were broadly similar so I stuck with one set of ratio measures. Equation 1 shows how this is calculated, the ranking it produces in addition to its distribution.

$$\text{'Ahmed Score'} = \operatorname{arccosh}\left(\log\left(\sqrt[3]{\text{abs}\left(\sum \text{Features}\right)}\right)\right)$$

Equation 1: Definition of Ahmed Score. It's the logarithmic cube root weighting of summed features mapped as an angle to a hyperbolic plane. So bigger scores are better than lower ones. As Figure 1 shows it provides an easy way to rank countries and produces a bimodal distribution. This measure favours developing countries as shown in Figure 2.

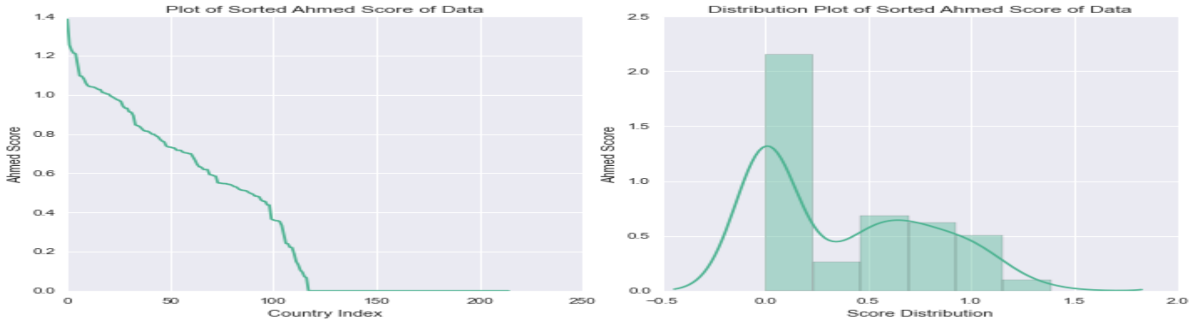


Figure 1: Left: Distribution of Ahmed Score greater than 0 from the data. Right: Sorted Ahmed Score plot of countries showing rankings produced.

The map in **Fig 2** shows how this can be visualised and the insights we can from it. We see that the median Fertility Rate is generally on the lower end compared to developing countries the usefulness of the score is captured in the size of the bubbles. To visualise the large number of ratios in this map would be difficult but one measure that captures them all makes it easier to see trends such as this score tends to favour developing countries. The reason for this could be that most of the developing countries have been undergoing recession for most of the time period considered so the indicators could have summed to zero. Also a lot of the growth over the same time period has been in emerging nations which would give them more positive scores. This is probably the underlying reason we observe the trends above.

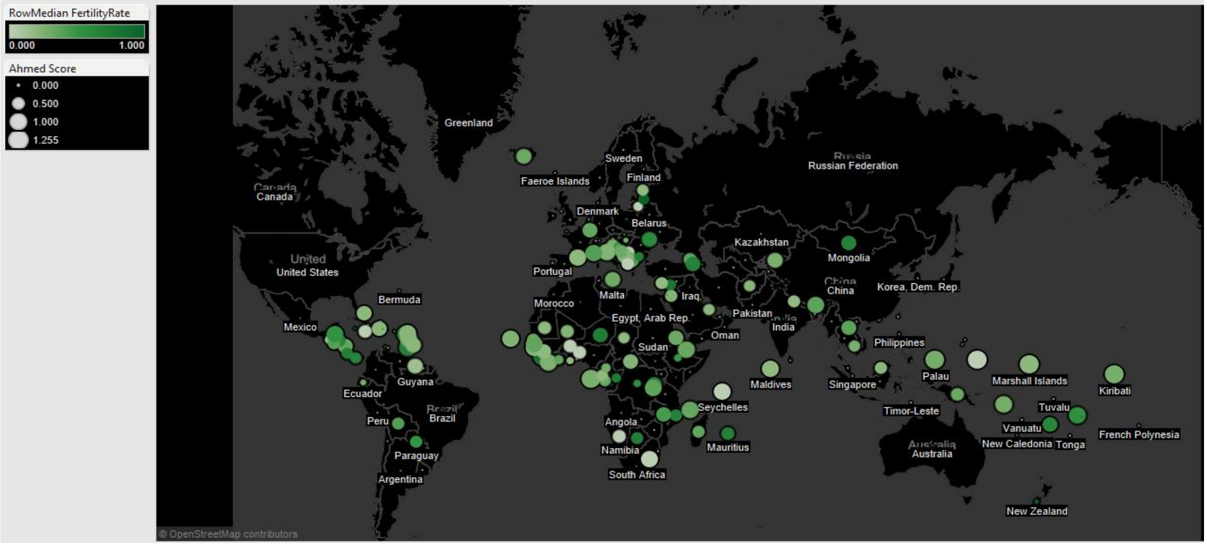


Figure 2: Map Countries Coloured by Median Fertility Rates and Sized by Ahmed Score showing that this measure favours developing countries which makes sense because most of the measures used to calculate it revolve around fertility rate ratios which are lower in developed countries than developing countries shown smaller circles here. Generated using Tableau.

Since the data is high dimensional, dimensionality reduction was a key component of the analysis and inspiration was taken from [8], [9] and the methods tested on this data were Principal Components Analysis (PCA), Independent Component Analysis (ICA), Factor Analysis (FA), Linear Local Embedding (LLE), t-distributed Stochastic Neighbour Embedding (t-SNE) and Non Negative Matrix Factorization (NMF).

The details of the dimensionality techniques are not discussed in detail here but references provided for those interested. But very briefly, PCA is an example of a linear dimension reduction technique that embeds data into a smaller subspace[9]. ICA is used for revealing hidden factors that underlie the dataset[10]. Factor Analysis removes redundancy from the data with a smaller set of derived variables and these factors are fairly independent of the initial variables[11]. LLE and t-SNE are examples of nonlinear methods for dimensionality reduction[9] in contrast to PCA. NMF is another approach to reducing dimensionality that aims to find non negative matrices whose product will approximate the non-negative data[12].

The first technique applied was PCA it was found that 3 components are sufficient to explain 99% of the variance in the data so all the other techniques are tested with 3 components and results presented in **Fig 7.** NMF on this data gave a very high reconstruction error so it was deemed unsuitable for application. The LLE produces a good projection of the data with a small reconstruction error rate but proved very difficult to integrate with regression methods so this is not considered further in the analysis but results of the testing are presented for completeness.

Regression methods are chosen because we are interested in a numerical value so classification methods are not appropriate for this type of problem. The methods chosen are ensemble methods such as Gradient Boosting Regression [13], Random Forest [14] and Extra Trees Regressor [15].

Ensemble methods are chosen because they have the benefit of being able to combine predictions of several underlying estimators which can contain combinations of strong and weak learners with the learning method, which improves the overall generalizability and robustness in contrast to using a single estimator. This is done by averaging like in Random Forests and Extra Trees or by boosting which minimizes the combined bias from the estimators like Gradient Boosting Regression. [16]

Prior to the application of the modelling techniques, the data had the target Median Fertility Rate removed and used as the target. All other statistics related to it such as IQR, Mean and Standard Deviation were dropped from the data. But the means ratios that included Fertility Rate were left in the data in addition to the others shown in **Table 2** Step 3**.** The training and test set was split into 90% for the training set with 10 fold cross validation and 10% for the test set. Mean Square Errors on the test set and average crass validation error are quoted here.

The results from the modelling with the different combinations are shown in **Table 3.**

Processing Applied	Machine Learning Method	MSE	Accuracy	CV Error
Data with scaling only	Gradient Boosting MSE	0.00769	99.2	0.7631
	Extra Trees MSE	0.00397	99.6	0.9200
	Random Forest MSE	0.00521	99.5	0.8738
Data scaling + PCA	Gradient Boosting w. PCA MSE	0.03538	96.5	0.0155
	Extra Trees w. PCA MSE	0.02593	97.4	0.0067
	Random Forest w. PCA MSE	0.02949	97.1	0.1384
Data Scaling + PCA + FA	Gradient Boosting w. PCA + FA MSE	0.03535	96.5	0.3467
	Extra Trees w. PCA + FA MSE	0.02629	97.4	0.4352
	Random Forest w. PCA + FA MSE	0.04405	95.6	0.3123
Data Scaling + PCA + ICA	Gradient Boosting w. PCA + ICA MSE	0.03542	96.5	0.4553
	Extra Trees w. PCA + ICA MSE	0.03114	96.9	0.4809
	Random Forest w. PCA + ICA MSE	0.04486	95.5	0.4884
Data Scaling + PCA + FA + ICA	Gradient Boosting w. PCA + FA +ICA MSE	0.03541	96.5	0.4553
	Extra Trees w. PCA + FA +ICA MSE	0.03282	96.7	0.4837
	Random Forest w. PCA + FA +ICA MSE	0.03405	96.6	0.4652
Data Scaling + ICA	Gradient Boosting ICA MSE	0.02712	97.3	0.2252
	Extra Trees ICA MSE	0.02406	97.6	0.5563
	Random Forest ICA MSE	0.03984	96	0.2752
Data Scaling + t-SNE	Gradient Boosting t-SNE MSE	0.02712	94	0.1652
	Extra Trees t-SNE MSE	0.02406	94.2	0.1482
	Random Forest t-SNE MSE	0.03984	95.1	0.2530
Data Scaling + PCA + FA + ICA + t-SNE	Gradient Boosting PCA + FA + ICA + t-SNE MSE	0.05989	96.5	0.0106
	Extra Trees PCA + FA + ICA + t-SNE MSE	0.05812	96.8	0.0067
	Random Forest PCA + FA + ICA + t-SNE MSE	0.04939	96.3	0.1619

Table 3: Table showing Mean Square Errors from the Machine Learning Models used to predict the Median Fertility Rate from our data with and without dimensionality reduction. With no dimensionality reduction the errors are lower but then there are over 90 dimensions in the data and the increase in error is less than 5%.

RESULTS

Fig 8 shows Feature Importance of the calculated features on the Median Fertility Rate, from here we can offer an explanation that the reason we might be observing a negative correlation between Fertility Rates and Total Debt in **Fig 6** is probably because the most important determinants seem to be GDP measures, Health Expenditure with respect to the Population, the Percentage of Labour Force that is Female and the amount by which price of things rise every year measured by CPI.

A possible explanations could be that if a Country has a high debt burden a lot of the Government expenditure would have to go towards servicing debt rather than on productive things of benefit to the economy such as healthcare, infrastructure and social policy which probably reduces opportunities for women and thus encourages them to probably marry and raise families. This is in contrast to more developed nations where better opportunities for women mean that career considerations might be affecting attitudes towards raising families as there is a significant cost involved personally and professionally for women. This is well captured my the custom measure designed for this dataset dubbed the 'Ahmed Score'.

The modelling suggests that we can successfully take summary statistics of the time series data and predict the median Fertility Rate of countries accurately. The models shown in Table 3 with different processing all show accuracy of over 95%. Since three different models are used with similar accuracy we can have us confidence in our results otherwise we would have expected much greater variation and lower accuracy from the models and also say that Regression Models especially the ensemble methods are highly successful on this problem.

It is important to note that dimensionality reduction worked very well on this data. We found that PCA with 3 components was sufficient to explain 99% of the variance in data and subsequent modelling showed that our accuracy does not drop below 95% and the difference is less than 5% than if the full dataset is used. This is an important finding that we can reduce our 90 dimension data into a small dataset and still make very accurate prediction of our value of interest. Since a wide range of methods are tested and combined we can see that our models are robust. With the exception of t-SNE no other processing yields final model results lower than 95%. Even then the result is only marginally lower than 95% accuracy but is still within the 5% window from the results of the full data set. The 5% can be considered as an acceptable limit on the variability of the results. However, PCA and ICA are the best methods for reducing dimensionality of this data. There also seems to be an upper limit as to how many dimensionality techniques that can be chained are actually useful. We see that with 4 methods combined we achieve around 96% accuracy which is lower than that gained by PCA and ICA. It was expected that combining PCA and FA would be quite effective since PCA is a special case FA. But we find that this combination produces the most variability in the accuracy across the 3 models. While applying ICA to the PCA and FA combination seems to stabilize the accuracy across methods. Thus we can hypothesise that if one wants the most accurate possible result with data dimensionality reduced on this problem PCA and ICA are good choices. If accurate and stable results across methods are required then the PCA, FA and ICA methods can be applied prior to modelling.

	ICA Comp1	PCA Comp1	FA Comp1	PCA + FA Comp1	PCA + ICA Comp1
0	HealthExp_toFertRate	RowMean_GNI	RowIQR_GNI	RowMean_GNI	%Chg5yrs_FertilityRate
1	%Chg13yrs_GDPgrowth	Debt_toGNI	%Chg13yrs_TotDebtService	%Chg13yrs_CPI	%Chg10yrs_GDPgrowth
2	%Chg13yrs_GDPperCap	RowIQR_Population	%Chg1yrs_TotDebtService	RowIQR_CPI	RowMedian_GDPperCap
3	%Chg1yrs_FertilityRate	%Chg5yrs_Population	RowStd_GNI	RowStd_CPI	%Chg13yrs_CPI
4	%Chg5yrs_FertilityRate	%Chg1yrs_LabForFemale	RowMean_GDP	RowMedian_CPI	RowStd_GDPperCap
5	%Chg1yrs_HealthExp	%Chg13yrs_Population	RowMedian_GNI	RowMean_CPI	FertRate_toLabForFem
6	%Chg13yrs_CPI	%Chg1yrs_Population	RowMean_GNI	%Chg10yrs_LabForFemale	RowMean_HealthExp
7	Pop_toGNI	%Chg10yrs_Population	%Chg5yrs_GNI	%Chg5yrs_LabForFemale	RowIQR_TotDebtService
8	RowMean_GDP	%Chg13yrs_LabForFemale	RowMedian_TotDebtService	%Chg1yrs_LabForFemale	%Chg1yrs_GDPperCap
9	Debt_toGNI	%Chg1yrs_CPI	%Chg1yrs_GNI	%Chg1yrs_CPI	%Chg1yrs_GNI
10	%Chg1yrs_GDP	RowMedian_Population	%Chg10yrs_TotDebtService	%Chg13yrs_LabForFemale	%Chg5yrs_LabForFemale

Table 4: Shows top 11 features from the first components of PCA , ICA, FA, PCA + FA and PCA + ICA.

Table 4 gives us additional insight into the features that the dimensionality reduction methods have identified as being important. This is to complement **Fig 8** and the full list is given in **Table 5**. We see PCA and FA find more global features in the data such as GNI while ICA finds more local features such Health Expenditure to Fertility Rate means ratio and %Changes in GDP. Since we see that PCA and ICA projected data have similar accuracy it is interesting to note the differences in the feature importance that these 2 methods return and gives us a different way to think about target attribute. This is useful because in **Fig 4** we see that there is high level of correlation and **Fig 5** shows there are some structures in the data, the dimensionality reduction methods allows us to determine which variables are significant within the structures we are observing.

SUGGESTIONS FOR FURTHER WORK

Although this is only observed in this data, **further work** could explore whether this dimensionality reduction processing flow has the same effect across even more models and on different datasets. A graphical modelling approach to this data could also another extension to this work.

COMMENTS ON SOFTWARE AND APPROACH

This analysis as noted has been conducted in Python with the pandas[4], numpy[7], seaborn[5], matplotlib[17], and scikit-learn[6] libraries. The map presented is generated using Tableau. Since this was a fairly large analysis task, it was helpful to split into stages. The first python script dealt with pre-processing such importing, cleaning, filling blanks, calculating statistics, merging and scaling data. The second script was the analysis stage which read in the excel file generated in the previous step and calculated ratio of means, custom scores and enabled testing of dimensionality reduction techniques before they were applied. An excel file of the final merged data with all the features are output to be visualised in Tableau at this stage. The third script was the application of the dimensionality reduction techniques deemed useful in the previous step with their associated parameters. The fourth script was the machine learning part where the data with and without dimensionality reduction along with the various combinations were split into training and test set to be used for predictive modelling.

CONCLUSION

In conclusion, it can be said we have successfully shown that median Fertility Rates are related to macroeconomics indicators such as GDP, CPI, GNI and Total Debt of a Country etc. Also, it is noted that median Fertility Rate of a country can be successfully predicted from features derived from time series data of these indicators exclusively. A custom measure to incorporate a range of features has been found to be surprisingly, an important predictor of our value of interest.

It is hoped that being able to predict trends in demographics changes can help countries prepare for the future with adequate social and economic policies. This can be with regards to increasing migration to combatting greying population, greater automation within the economy and also developing legislation which promotes a culture of mother-friendly work environments. This is because we note a trend that countries with a high proportion of women in the labour force have lower fertility rates.

APPENDIX OF FIGURES

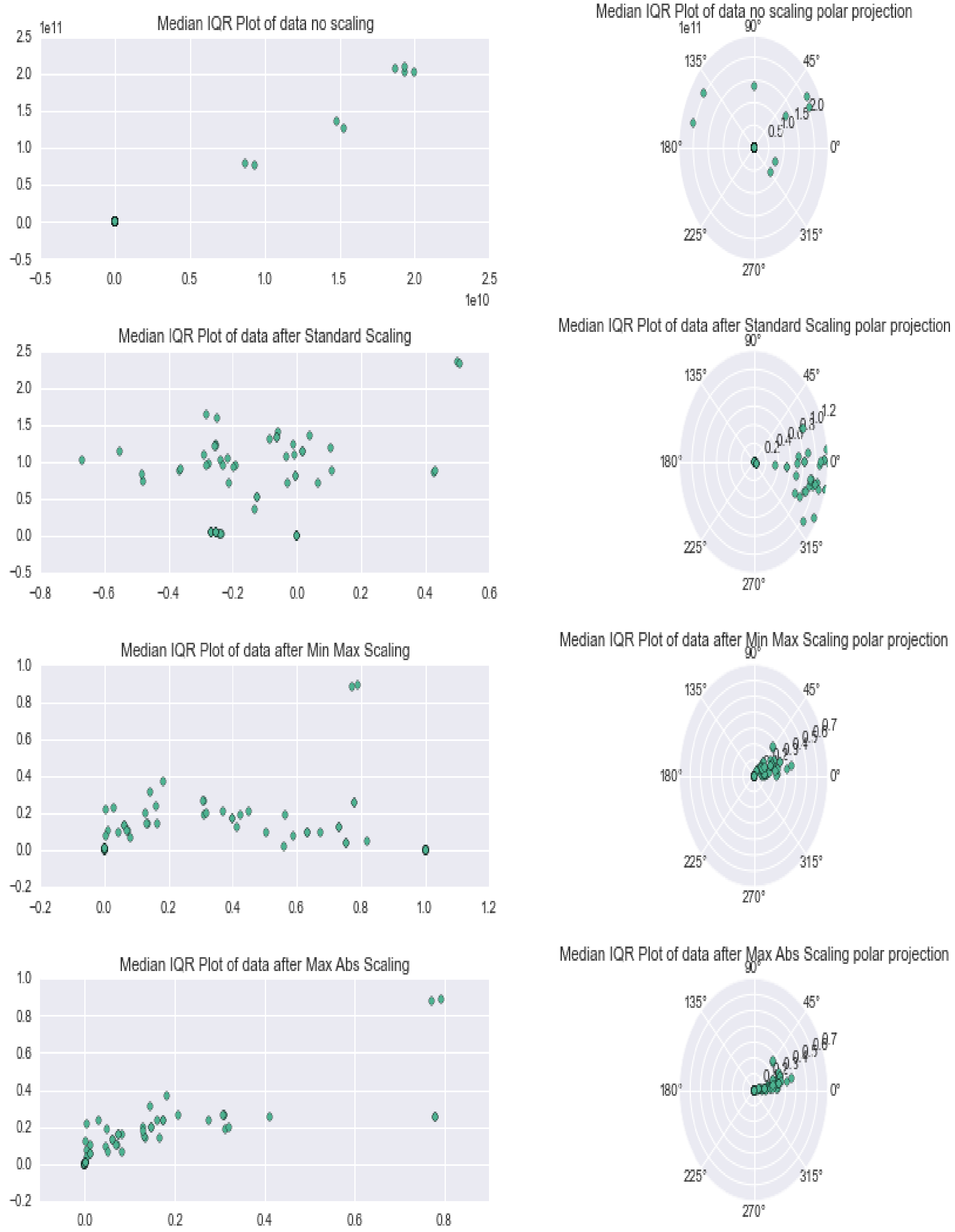
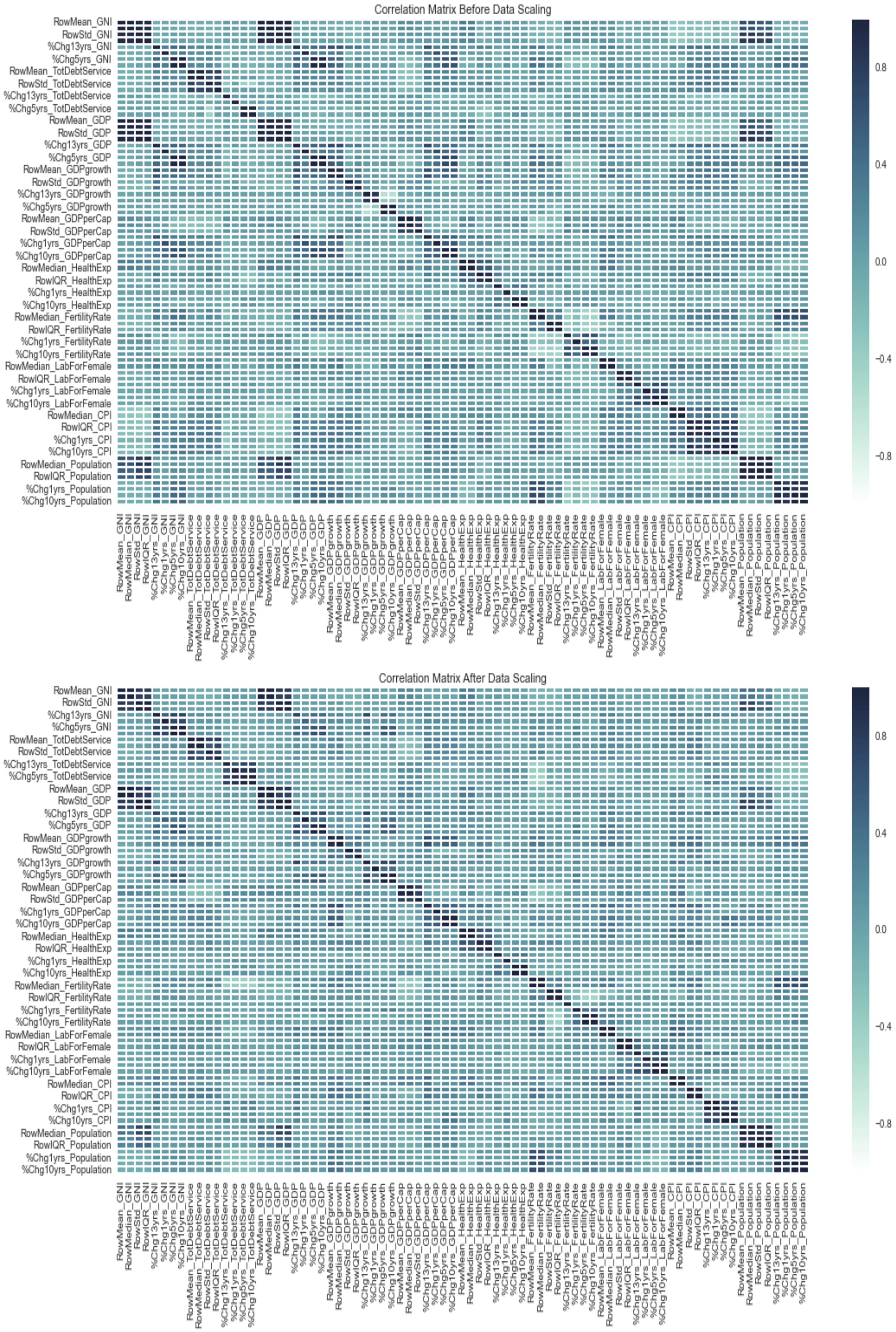


Figure 3: IQR Median Plots of data with and without various types of scaling. From the polar plots we see that without scaling the data is all over the place and hence there is a need for scaling. With the MaxAbs scaling we see that the data is scaled very well and that the outliers are exposed much more clearly than the other methods.

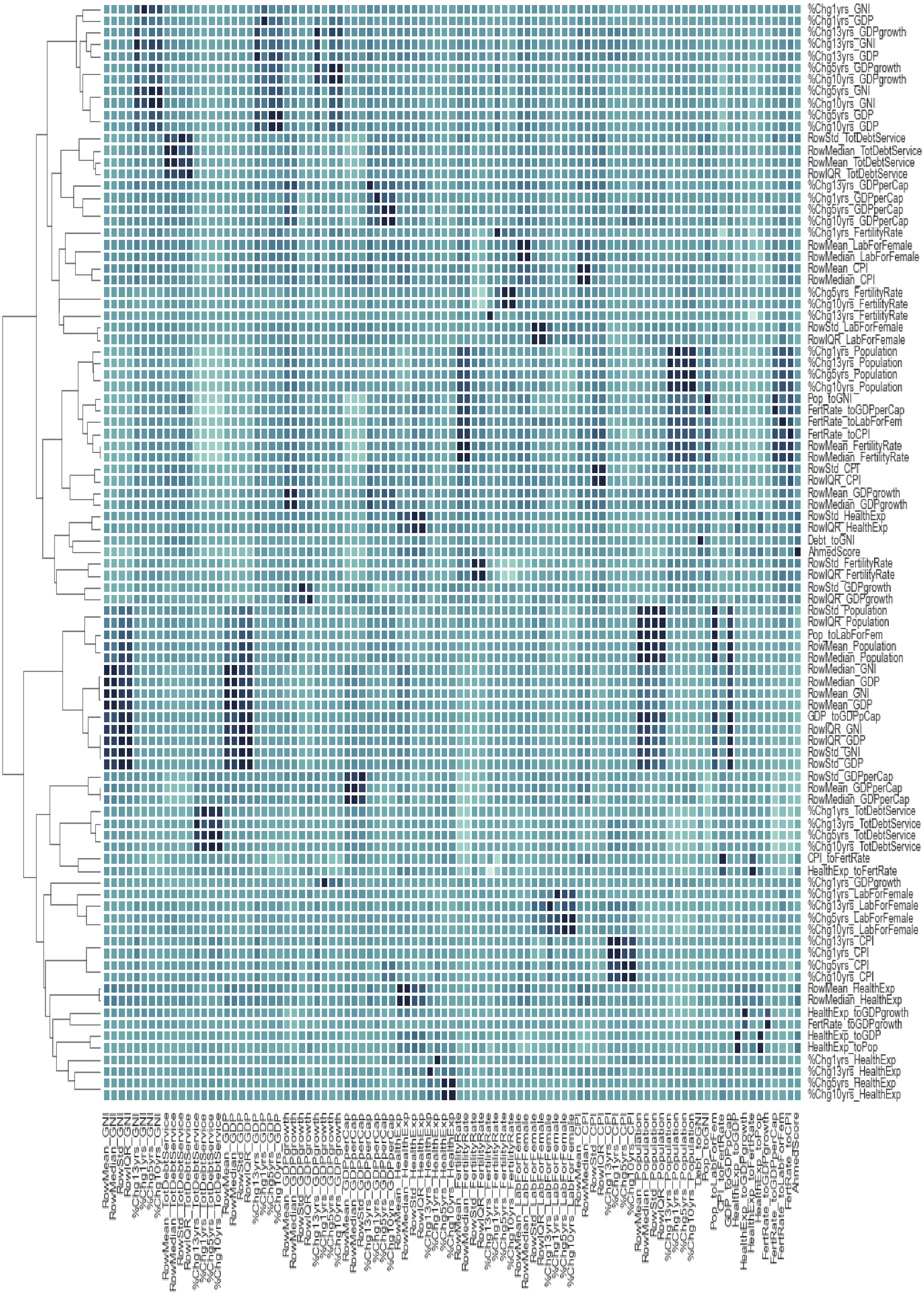
Figure 4: Correlation Matrix Heat map of data before and after Max Abs Scaling, showing there is a high degree of correlation among variables and that the correlations are scaling invariant.



Correlation Matrix Before Data Scaling

Correlation Matrix After Data Scaling

Figure 5: Clustergram of the pairwise correlation matrix of Max Abs scaled data. Agglomerative hierarchical clustering performed along rows using weighted distance as the linkage method. Only rows are used because square matrix rows and columns would be identical. The distance metric used is correlation distance as this was found to bring out structures better in this data. Darker colours indicate high correlation than lighter colours. We see that there is a high level of correlation on the data but some clusters such as % Labour force that is female related to Population indicators.



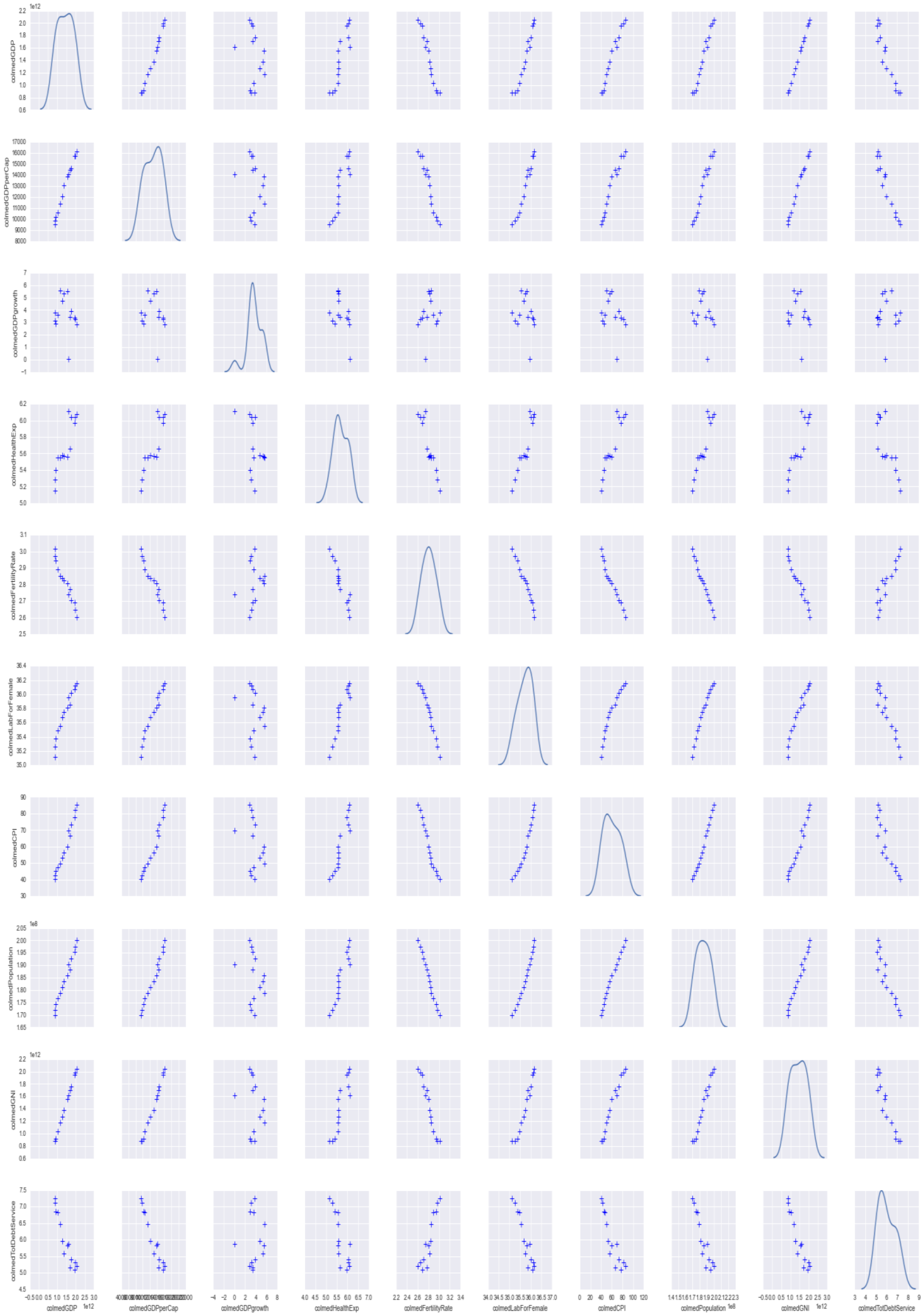


Figure 6: Pair plot of column medians of the data for the different indicators for the time range 2000-2013. We see some interesting correlations and shows high correlation among variables also we that Fertility Rate is negatively correlated with almost all indicators apart from Total Debt.

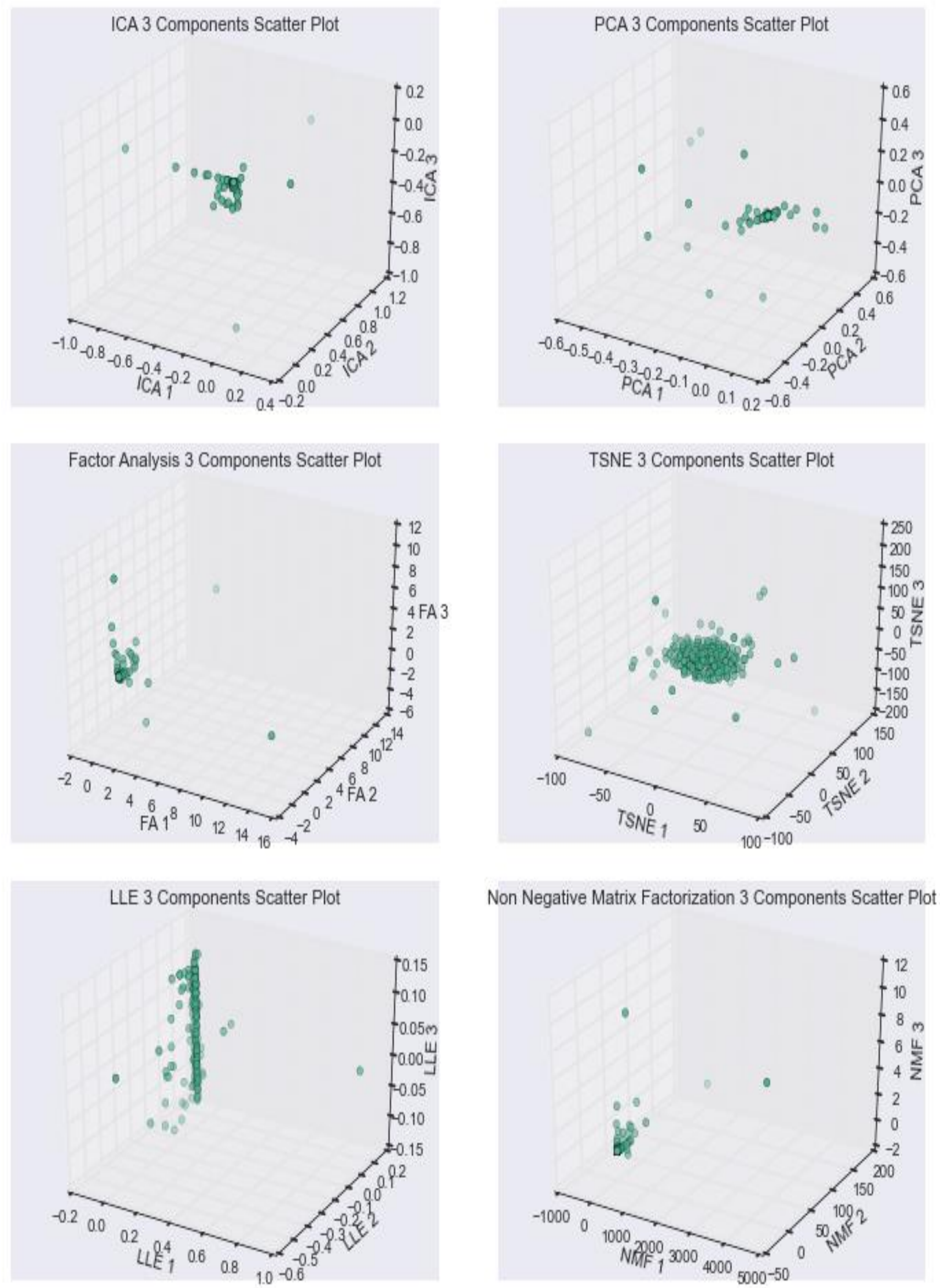


Figure 7: Plot of 3 components of dimensionality reduction methods tested on the data. From the top left we see scatter plots of ICA, PCA, FA, T-SNE, LLE and NMF. We see the first produce good stable results but TSNE produces a very spherical projection which is not ideal and NMF has a huge variation in scale between the components and its reconstruction error of the original data is very high, while that of LLE is very low.

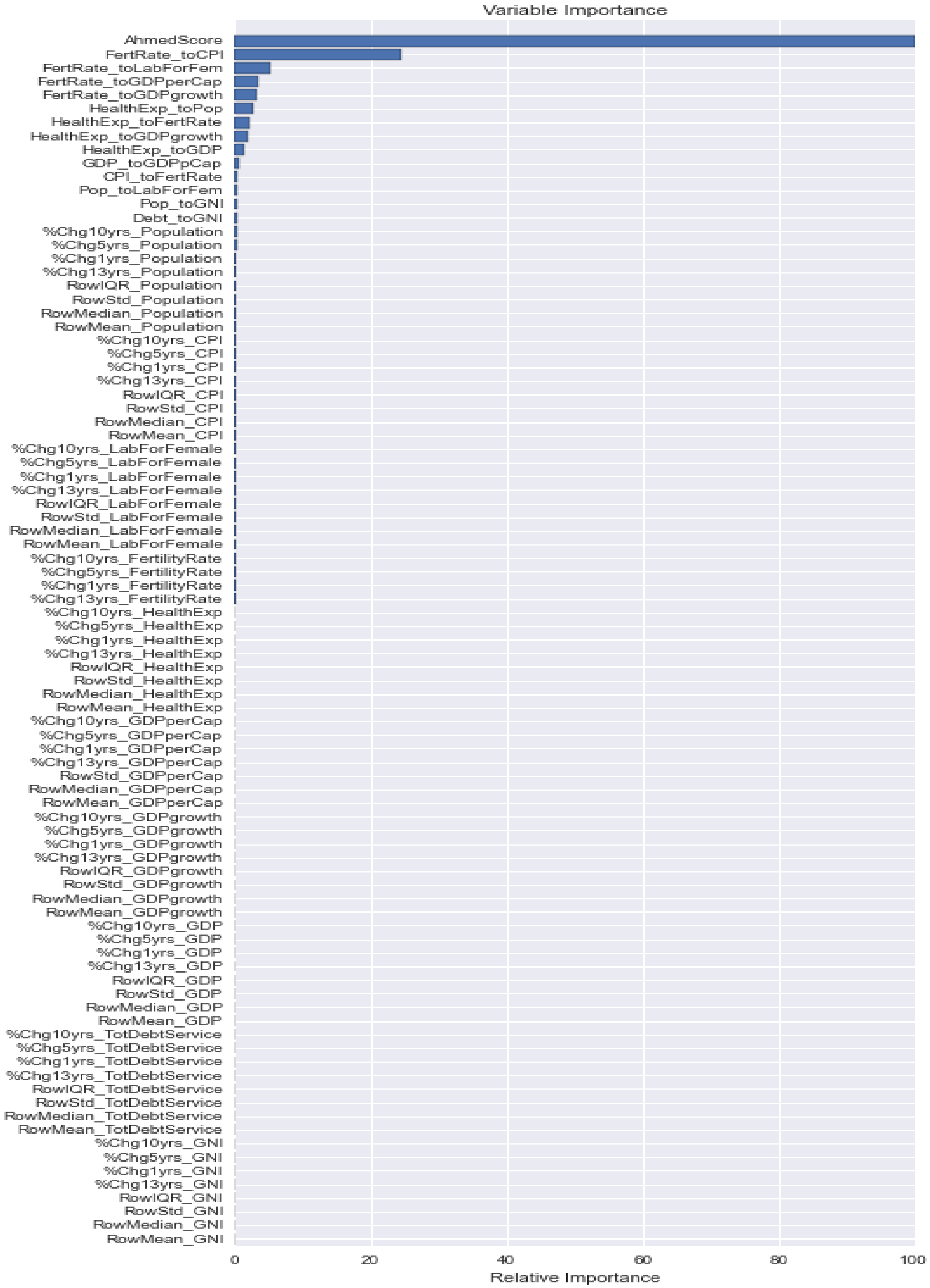


Figure 8: Feature Importance plot derived using Gradient Boosting Regressor with the target being Median Fertility Rate. We see that a low number of features are sufficient to explain our desired attribute from the data hence the effectiveness of dimensionality reduction techniques on this data set. Also not that features engineered as a proxy to the original time series are very effective especially the custom score derived from them.

	ICA Comp1	PCA Comp1	FA Comp1	PCA + FA Comp1	PCA + ICA Comp1
0	HealthExp_toFertRate	RowMean_GNI	RowIQR_GNI	RowMean_GNI	%Chg5yrs_FertilityRate
1	%Chg13yrs_GDPgrowth	Debt_toGNI	%Chg13yrs_TotDebtService	%Chg13yrs_CPI	%Chg10yrs_GDPgrowth
2	%Chg13yrs_GDPperCap	RowIQR_Population	%Chg1yrs_TotDebtService	RowIQR_CPI	RowMedian_GDPperCap
3	%Chg1yrs_FertilityRate	%Chg5yrs_Population	RowStd_GNI	RowStd_CPI	%Chg13yrs_CPI
4	%Chg5yrs_FertilityRate	%Chg1yrs_LabForFemale	RowMean_GDP	RowMedian_CPI	RowStd_GDPperCap
5	%Chg1yrs_HealthExp	%Chg13yrs_Population	RowMedian_GNI	RowMean_CPI	FertRate_toLabForFem
6	%Chg13yrs_CPI	%Chg1yrs_Population	RowMean_GNI	%Chg10yrs_LabForFemale	RowMean_HealthExp
7	Pop_toGNI	%Chg10yrs_Population	%Chg5yrs_GNI	%Chg5yrs_LabForFemale	RowIQR_TotDebtService
8	RowMean_GDP	%Chg13yrs_LabForFemale	RowMedian_TotDebtService	%Chg1yrs_LabForFemale	%Chg1yrs_GDPperCap
9	Debt_toGNI	%Chg1yrs_CPI	%Chg1yrs_GNI	%Chg1yrs_CPI	%Chg1yrs_GNI
10	%Chg1yrs_GDP	RowMedian_Population	%Chg10yrs_TotDebtService	%Chg13yrs_LabForFemale	%Chg5yrs_LabForFemale
11	RowIQR_GNI	Pop_toGNI	%Chg5yrs_TotDebtService	RowStd_LabForFemale	HealthExp_toPop
12	%Chg1yrs_GNI	%Chg5yrs_GDPperCap	RowMedian_GDP	RowMedian_LabForFemale	FertRate_toGDPgrowth
13	FertRate_toGDPperCap	%Chg10yrs_HealthExp	%Chg13yrs_LabForFemale	RowMean_LabForFemale	RowIQR_GNI
14	%Chg13yrs_HealthExp	%Chg10yrs_CPI	%Chg10yrs_CPI	%Chg10yrs_FertilityRate	HealthExp_toGDPgrowth
15	RowStd_GDP	RowIQR_LabForFemale	%Chg5yrs_CPI	%Chg5yrs_FertilityRate	HealthExp_toFertRate
16	%Chg5yrs_HealthExp	RowMedian_GDPperCap	%Chg1yrs_CPI	%Chg1yrs_FertilityRate	AhmedScore
17	%Chg13yrs_LabForFemale	RowStd_CPI	%Chg13yrs_CPI	%Chg13yrs_FertilityRate	%Chg10yrs_CPI
18	RowStd_HealthExp	%Chg10yrs_GDPgrowth	RowIQR_CPI	%Chg10yrs_HealthExp	%Chg10yrs_HealthExp
19	%Chg1yrs_TotDebtService	RowStd_Population	RowStd_CPI	RowIQR_LabForFemale	%Chg10yrs_GNI
20	RowIQR_CPI	RowIQR_CPI	RowMedian_CPI	%Chg5yrs_CPI	RowIQR_Population
21	%Chg13yrs_GDP	%Chg5yrs_HealthExp	RowMean_CPI	%Chg10yrs_CPI	%Chg13yrs_GNI
22	%Chg10yrs_FertilityRate	%Chg13yrs_CPI	%Chg10yrs_LabForFemale	RowMean_Population	RowMean_GDPgrowth
23	RowMedian_GDP	%Chg1yrs_TotDebtService	%Chg5yrs_LabForFemale	FertRate_toLabForFem	RowStd_LabForFemale
24	RowStd_TotDebtService	RowMean_Population	RowMedian_LabForFemale	FertRate_toGDPperCap	%Chg13yrs_GDPgrowth
25	HealthExp_toGDP	RowMean_GDPgrowth	RowStd_LabForFemale	FertRate_toGDPgrowth	RowMean_LabForFemale
26	CPI_toFertRate	%Chg13yrs_GDP	RowMean_Population	HealthExp_toPop	GDP_toGDPpCap
27	RowStd_CPI	RowStd_GDPgrowth	RowIQR_LabForFemale	HealthExp_toFertRate	%Chg10yrs_FertilityRate
28	RowMedian_CPI	%Chg1yrs_HealthExp	%Chg1yrs_LabForFemale	HealthExp_toGDPgrowth	RowStd_Population
29	RowMean_CPI	%Chg5yrs_GDP	RowMedian_Population	HealthExp_toGDP	%Chg1yrs_TotDebtService
30	%Chg13yrs_TotDebtService	%Chg13yrs_HealthExp	Debt_toGNI	GDP_toGDPpCap	Pop_toLabForFem
31	%Chg5yrs_CPI	%Chg5yrs_GDPgrowth	RowIQR_Population	CPI_toFertRate	RowIQR_GDPgrowth
32	AhmedScore	%Chg5yrs_CPI	FertRate_toLabForFem	Pop_toLabForFem	RowMedian_GDPgrowth
33	RowMean_GNI	RowIQR_TotDebtService	FertRate_toGDPperCap	Pop_toGNI	FertRate_toCPI
34	RowMedian_GDPperCap	RowMean_GDP	FertRate_toGDPgrowth	Debt_toGNI	RowStd_TotDebtService
35	RowMedian_Population	RowMean_TotDebtService	HealthExp_toPop	%Chg10yrs_Population	%Chg5yrs_GDPperCap
36	HealthExp_toGDPgrowth	RowMedian_TotDebtService	HealthExp_toFertRate	%Chg5yrs_Population	%Chg13yrs_HealthExp
37	%Chg10yrs_GNI	RowIQR_HealthExp	HealthExp_toGDPgrowth	%Chg1yrs_Population	%Chg13yrs_LabForFemale
38	%Chg5yrs_GDP	%Chg5yrs_GNI	HealthExp_toGDP	%Chg13yrs_Population	%Chg13yrs_GDP
39	RowMean_LabForFemale	%Chg1yrs_GNI	RowStd_Population	RowIQR_Population	%Chg13yrs_GDPperCap
40	RowIQR_HealthExp	RowStd_GDP	GDP_toGDPpCap	RowStd_Population	RowStd_GNI
41	%Chg13yrs_GNI	RowIQR_GNI	Pop_toLabForFem	RowMedian_Population	%Chg5yrs_CPI
42	RowMedian_GNI	RowStd_GNI	Pop_toGNI	%Chg5yrs_HealthExp	%Chg1yrs_GDP
43	RowMedian_TotDebtService	FertRate_toLabForFem	RowMean_LabForFemale	FertRate_toCPI	RowMedian_TotDebtService
44	RowMean_HealthExp	FertRate_toGDPperCap	%Chg10yrs_Population	%Chg1yrs_HealthExp	RowMedian_Population
45	RowMean_GDPgrowth	HealthExp_toGDPgrowth	%Chg5yrs_Population	RowIQR_HealthExp	%Chg13yrs_Population
46	FertRate_toCPI	HealthExp_toGDP	%Chg1yrs_Population	RowStd_GDP	RowMean_GNI
47	%Chg10yrs_LabForFemale	HealthExp_toPop	%Chg13yrs_Population	RowMedian_GDP	RowStd_GDPgrowth
48	%Chg10yrs_GDPgrowth	FertRate_toCPI	CPI_toFertRate	RowMean_GDP	RowIQR_GDP
49	RowMean_GDPperCap	HealthExp_toFertRate	%Chg10yrs_FertilityRate	%Chg10yrs_TotDebtService	%Chg10yrs_LabForFemale
50	%Chg10yrs_TotDebtService	Pop_toLabForFem	%Chg1yrs_HealthExp	%Chg5yrs_TotDebtService	%Chg5yrs_Population
51	GDP_toGDPpCap	CPI_toFertRate	%Chg1yrs_FertilityRate	%Chg1yrs_TotDebtService	RowMedian_CPI
52	%Chg1yrs_Population	GDP_toGDPpCap	RowStd_GDP	%Chg13yrs_TotDebtService	RowMean_GDP
53	%Chg5yrs_GDPgrowth	FertRate_toGDPgrowth	RowIQR_GDP	RowIQR_TotDebtService	FertRate_toGDPperCap
54	%Chg10yrs_Population	AhmedScore	%Chg13yrs_GDP	RowIQR_GDP	%Chg10yrs_TotDebtService
55	%Chg1yrs_GDPgrowth	RowMedian_GNI	%Chg1yrs_GDP	RowStd_TotDebtService	RowMedian_HealthExp
56	RowStd_LabForFemale	%Chg13yrs_GNI	%Chg5yrs_GDP	RowMean_TotDebtService	%Chg5yrs_GDPgrowth
57	RowIQR_GDPgrowth	%Chg10yrs_GNI	%Chg10yrs_GDP	%Chg10yrs_GNI	%Chg1yrs_LabForFemale
58	RowIQR_TotDebtService	%Chg13yrs_TotDebtService	RowMean_GDPgrowth	%Chg5yrs_GNI	RowMedian_GDP
59	HealthExp_toPop	%Chg13yrs_GDPgrowth	RowMedian_GDPgrowth	%Chg1yrs_GNI	RowIQR_CPI
60	RowMedian_GDPgrowth	%Chg10yrs_TotDebtService	RowStd_GDPgrowth	%Chg13yrs_GNI	%Chg13yrs_FertilityRate
61	%Chg10yrs_HealthExp	%Chg1yrs_GDP	RowIQR_GDPgrowth	RowIQR_GNI	%Chg5yrs_GNI
62	%Chg5yrs_Population	RowMedian_GDP	%Chg13yrs_GDPgrowth	RowStd_GNI	%Chg13yrs_TotDebtService
63	FertRate_toGDPgrowth	RowStd_TotDebtService	%Chg1yrs_GDPgrowth	RowMedian_GNI	RowStd_CPI
64	%Chg10yrs_GDPperCap	%Chg10yrs_FertilityRate	%Chg5yrs_FertilityRate	RowMedian_TotDebtService	%Chg10yrs_GDPperCap
65	RowIQR_LabForFemale	%Chg1yrs_GDPgrowth	%Chg10yrs_GDPgrowth	%Chg13yrs_GDP	%Chg5yrs_GDP
66	%Chg5yrs_LabForFemale	%Chg13yrs_FertilityRate	RowMean_GDPperCap	%Chg1yrs_GDP	CPI_toFertRate
67	Pop_toLabForFem	RowIQR_GDPgrowth	%Chg5yrs_GDPgrowth	%Chg5yrs_GDP	Pop_toGNI
68	RowIQR_Population	%Chg5yrs_TotDebtService	RowStd_GDPperCap	RowStd_HealthExp	RowMean_GDPperCap
69	RowMedian_HealthExp	%Chg10yrs_GDPperCap	%Chg13yrs_FertilityRate	RowMedian_HealthExp	%Chg1yrs_GDPgrowth
70	RowStd_GDPgrowth	RowMedian_HealthExp	%Chg10yrs_HealthExp	RowMean_HealthExp	RowStd_GDP
71	%Chg10yrs_CPI	%Chg13yrs_GDPperCap	%Chg5yrs_HealthExp	%Chg10yrs_GDPperCap	RowMean_TotDebtService
72	RowStd_GNI	RowIQR_GDP	RowMedian_GDPperCap	%Chg5yrs_GDPperCap	RowMedian_GNI
73	%Chg5yrs_GDPperCap	RowMedian_GDPgrowth	%Chg13yrs_HealthExp	%Chg1yrs_GDPperCap	RowIQR_LabForFemale
74	%Chg10yrs_GDP	%Chg5yrs_FertilityRate	RowIQR_HealthExp	%Chg13yrs_GDPperCap	RowMean_CPI
75	RowStd_GDPperCap	RowMean_HealthExp	FertRate_toCPI	RowStd_GDPperCap	%Chg1yrs_HealthExp
76	%Chg13yrs_Population	RowMean_LabForFemale	RowMedian_HealthExp	RowMedian_GDPperCap	%Chg1yrs_CPI
77	%Chg5yrs_GNI	RowStd_HealthExp	RowMean_HealthExp	RowMean_GDPperCap	%Chg1yrs_Population
78	%Chg1yrs_LabForFemale	%Chg1yrs_GDPperCap	%Chg10yrs_GDPperCap	%Chg10yrs_GDPgrowth	%Chg5yrs_HealthExp
79	RowMedian_LabForFemale	RowMedian_LabForFemale	%Chg5yrs_GDPperCap	%Chg5yrs_GDPgrowth	Debt_toGNI
80	%Chg1yrs_GDPperCap	%Chg10yrs_GDP	%Chg1yrs_GDPperCap	%Chg1yrs_GDPgrowth	%Chg10yrs_GDP
81	%Chg1yrs_CPI	%Chg1yrs_FertilityRate	%Chg13yrs_GDPperCap	%Chg13yrs_GDPgrowth	HealthExp_toGDP
82	RowStd_Population	RowMean_GDPperCap	RowStd_HealthExp	RowIQR_GDPgrowth	RowMean_Population
83	%Chg5yrs_TotDebtService	RowMean_CPI	AhmedScore	RowStd_GDPgrowth	%Chg1yrs_FertilityRate
84	RowIQR_GDP	%Chg5yrs_LabForFemale	%Chg10yrs_GNI	RowMedian_GDPgrowth	%Chg5yrs_TotDebtService
85	RowMean_TotDebtService	RowStd_LabForFemale	RowMean_TotDebtService	RowMean_GDPgrowth	RowStd_HealthExp
86	FertRate_toLabForFem	RowStd_GDPperCap	RowIQR_TotDebtService	%Chg10yrs_GDP	%Chg10yrs_Population
87	%Chg13yrs_FertilityRate	%Chg10yrs_LabForFemale	RowStd_TotDebtService	%Chg13yrs_HealthExp	RowIQR_HealthExp
88	RowMean_Population	RowMedian_CPI	%Chg13yrs_GNI	AhmedScore	RowMedian_LabForFemale

Table 5: Full list of feature ranking derived from dimensionality reduction methods. Note only the first component is quoted as it's the most significant.

REFERENCES

[1] "World Development Indicators | Data." [Online]. Available: http://data.worldbank.org/data-catalog/world-development-indicators. [Accessed: 29-Nov-2015].

[2] "Gender Statistics | Data." [Online]. Available: http://data.worldbank.org/data-catalog/gender-statistics. [Accessed: 29-Nov-2015].

[3] "International Debt Statistics | Data." [Online]. Available: http://data.worldbank.org/data-catalog/international-debt-statistics. [Accessed: 29-Nov-2015].

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

[5] "An introduction to seaborn ." [Online]. Available: http://stanford.edu/~mwaskom/software/seaborn/introduction.html. [Accessed: 29-Nov-2015].

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

[7] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, Mar. 2011.

[8] L. J. P. van der Maaten and H. J. van den H. , E.O. Postma, "Dimensionality Reduction: A Comparative Review." [Online]. Available: http://pages.iai.uni-bonn.de/zimmermann_joerg//dimensionality_reduction_a_comparative_review.pdf. [Accessed: 01-Dec-2015].

[9] M. Carreira-Perpinán, "A review of dimension reduction techniques," *Dep. Comput. Sci. Univ. Sheffield. Tech. Rep. CS-96-09*, pp. 1–69, 1997.

[10] A. Hyvärinen, "What is Independent Component Analysis ?" [Online]. Available: http://www.cs.helsinki.fi/u/ahyvarin/whatisica.shtml. [Accessed: 01-Dec-2015].

[11] E. Garrett-Mayer, "Introduction to Factor Analysis - Lecture8.pdf." [Online]. Available: http://ocw.jhsph.edu/courses/statisticspsychosocialresearch/PDFs/Lecture8.pdf. [Accessed: 01-Dec-2015].

[12] "sklearn.decomposition.NMF ." [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html#sklearn.decomposition.NMF. [Accessed: 01-Dec-2015].

[13] "sklearn.ensemble.GradientBoostingRegressor — scikit-learn 0.17 documentation." [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html. [Accessed: 01-Dec-2015].

[14] "sklearn.ensemble.RandomForestRegressor ocumentation." [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#sklearn.ensemble.RandomForestRegressor. [Accessed: 01-Dec-2015].

[15] "sklearn.ensemble.ExtraTreesRegressor." [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html#sklearn.ensemble.ExtraTreesRegressor. [Accessed: 01-Dec-2015].

[16] "1.11. Ensemble methods — scikit-learn 0.17 documentation." [Online]. Available: http://scikit-learn.org/stable/modules/ensemble.html#ensemble. [Accessed: 01-Dec-2015].

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