University College Cork

CS3205 Lab Report 1

HDI Trends and Multidimensional Projections

Jack O'Connor

February 26, 2022

Contents

1 Introduction

1.1 List of Acronyms

UCC (University College Cork), HDR (Human Development Report), HDI (Human Development Index), GNI (Gross Naional Income), GDP (Gross Domestic Product), PCA (Principal Component Analysis), LSP (Least Squares Projection), ProjClus (Projection Clustering), t-SNE (t-distributed Stochastic Neighbour Embedding), CBR (Case-Based Reasoning).

1.2 The Report

This report was created to document my experience developing my data visualisation skills as part of UCC's CS3205 Data Visualisation module. It is divided into two separate parts, each tackling its own separate area of data visualisation. As well as documenting my experience exploring the specified datasets and the tools I chose to use, this report also served as an instructive exercise in typesetting and formatting an academic paper, which is sure to prove useful next year when I tackle my final year project.

Part 1 of this report deals with visualising survey data, namely in the form of the annual Human Development Report (HDR) dataset, to find attribute patterns using a wide variety of methods. Part 2 of this report uses the HDR Dataset as well as several other sample labeled datasets (which will be discussed in depth in the data description section) to create comprehensible projections into the 2-dimensional plane of the point-like datasets using several multi-dimensional scaling techniques.

As one of the primary goals of this report is to compare the different visualisation methods against each other, I have formatted the document into two or three columns where appropriate, such that multiple visualisations are visible on each page. However, this does come at the cost of reducing the size of each visualisation. To offset this issue each image in this report is hyperlinked to a full size version hosted on either GitHub or Tableau. I would suggest viewing this pdf report in your browser such that the need to switch applications when viewing full size images is eliminated.

The full project repository containing all scripts, datasets and other miscellaneous files needed to reproduce this report can be found here.

1.3 The Role of Visualisation

Humans being a strongly visually oriented species means visualisations are a key part of any data analysis. A well formulated visualisation can turn an incomprehensible raw dataset into a graphic full of valuable information for our highly optimised pattern seeking brains.

That does not mean that all visualisation techniques are suitable for all data analysis tasks. Care must be taken with the transformation of the raw data into visualisations that the resulting visual is actually meaningful, and not just misleading noise. Tasks 1 and 2 of this report are a good example of distinguishing when and when not to use different visualisation techniques.

Task 1 uses the HDR dataset which has a (compared to Task 2's datasets) relatively small number of attributes, each of which is meaningful in its own right i.e. corresponds to an attribute of a country which has a meaningful, physical interpretation such as a country's total population or gross domestic produce (GDP). Techniques which compare individual attributes directly against each other can expose correlations between attributes which might not be obvious at first.

Task 2 on the other hand uses highly multidimensional data such as images where each pixel of an image can be attribute of that data, or text documents where the words of each document are embedded into a vector space with hundreds or thousands of dimensions. Each individual attribute of these datasets on its own does not carry much weight in the context of an image or docuement and directly comparing them to each other is unlikely to yield any salient information. In such an instance it is much more useful to project each data point into a 2 or 3-dimensional space and seek more broad patterns between documents such as clustering and dissimilarity.

2 Data Description

All **datasets** used are linked here.
All **scripts** used are linked here. (Scripts use relative filepaths based on project structure found in the GitHub repo.)

2.1 Task 1 Datasets

Two source datasets were used to complete task 1, on which data transformations were performed:

1. **CS3205_2020_statistical_annex_all.xlsx** is an Excel spreadsheet containing survey data collected by the Human Development Centre from their HDR 2020 report.

2. **hdi.csv**, a csv file containing the Human Development Index (HDI) score assigned to each country in the world from the years 1990 to 2019 provided to all students taking CS3205 on Canvas.

CS3205_2020_statistical_annex_all.xlsx contains all of the original sheets supplied in the HDR, unmodified. When selecting which attributes to include in the task 1 visualisations I found it most expedient to simply copy and paste the columns from the existing sheets to a new sheet called ReformattedData. This was possible due to the relatively small scall of the number of attributes and records in the dataset.

I then was able to manually add the HDI Label column by selecting groups of cells between each of the highlighted HDI Level rows which were in the original sheets (possible since countries are ordered by HDI rank). Finally I removed the highlighted HDI Level rows since they were only for the benefit of human observers and do not actually contain any attribute information.

hdi.csv was used to create two additional datasets. **hdi_with_levels.csv** is identical to hdi.csv except for each YEAR_HDI attribute of the dataset a corresponding YEAR_HDI_LEVEL attribute following the same scheme as in the HDR is added using the script **add_hdi_levels.R**.

hdi_pivoted.csv was created using **pivot_hdi_levels.R** and it performs a pivot transformation such that instead of having a YEAR column for each year in the period, a YEAR column and HDI column pair is used. This drastically reduces the number of repetitive attributes in the dataset at the expense of introducing repeat values for COUNTRY_NAME AND COUNTRY_CODE to make up for the increased number of rows in the dataset. This transformation was especially useful for aggregating data by year and country, since it cut down on the number of attributes which had to be included in any visualisation. A HDI_Label column was also added to this dataset.

2.2 Task 2 Datasets

A total of five primary multidimensional data sets and two auxiliary stopwords datasets for CBR and Headlines textual datasets were used to create projections in VisPipeline. **hdr.data** was the only dataset which was not readily provided on Canvas.

1. **Corel** is a collection of images with 10 distinct classes: African tribes, beaches, buildings, buses, dinosaurs, elephants, flowers, food, horses, mountains.

2. **CBR** is a collection of paper abstracts and references spanning four main topics: case-based reasoning (CBR), inductive logic programming (ILP), information retrieval (IR), sonification (SON) and six intruders.

3. **Medical** is a collection of X-Rays of the human body mainly of the skull and spine from different angles.

4. **Headlines** is a collection of articles from AP, BBC, Reuters and CNN which were all published during the same time period in the mid 2000's.

5. **HDR** contains all of the same attributes as were included in ReformattedData but excludes HDI and HDI_Rank.

To create the hdr.data file it was enough to create a new spreadsheet in Excel, remove the requisite columns, move Country column to the beginning and Label column to the end, export the new sheet to csv using a semi-colon as the delimiter and finally remove the Country and Label headers in the exported cvs while adding the necessary .data read-in file parameters to the top of the file.

3 Task 1

3.1 Free Exploration

In the free exploration section of the assignment I have mainly (but not exlusively) created visualisations which give a broad overview of the general distribution of HDI across countries.

Each of the following visualisations will include:

- A graphic
- An explanation of how to interpret the graphic
- A description of a pattern (or lack of)
- A hypothesis as to the cause of this pattern
- The dataset from which the data came

3.1.1 HDI Histogram

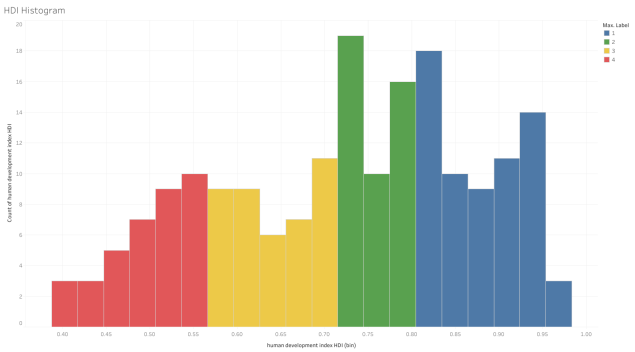

Figure 3.1: HDI Distribution with HDI Level as Colour

The histogram in Figure 3.1 gives a general idea as to the number of countries in each HDI score 'bins' ranging from about 0.38 to 0.98 in with each bin covering the next approximately 0.04 increase in HDI. The counts in each bin range from 3 to 19.

There is no mingling between colours since colour corresponds to HDI Level, which is single-handedly determined by a country's HDI score.

This histogram is appears to be left skewed, which implies that there are more countries with very high and high HDI than there are with medium and low HDI. This is a very positive thing to see as it is undoubtedly preferable to have more countries with a good standard of living than less.

Dataset: CS3205_2020_statistical_annex_all.xlsx

3.1.2 HDI Global Heatmap

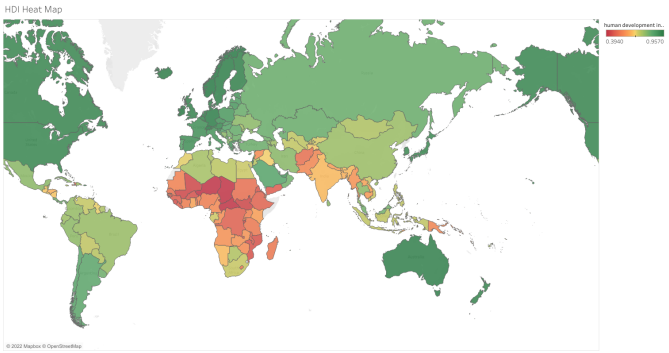

Figure 3.2: Darker greens imply higher HDI, darker reds lower HDI

The global heat map in Figure 3.2 is coloured by a dark-red to dark-green spectrum with the darkest red corresponding to the lowest HDI score and the darkest green corresponding to the highest HDI score. Using this visualisation it is very easy to see the influence a country's geographical location has on its HDI.

From looking at this visualisation it becomes readily apparent that there are very few low HDI countries that are not found in Africa (except for its northern coast) or the south of Asia.

Dataset: CS3205_2020_statistical_annex_all.xlsx

3.1.3 Average HDI Trend

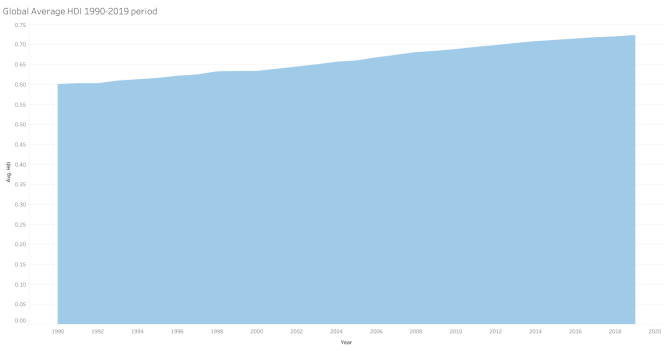

Figure 3.3: Global average HDI strictly rises over time

The area chart in Figure 3.3 represents the average HDI score across all countries in the world for each of the years between 1990 to 2019 inclusive.

By looking at this visualisation it is possible to interpret the fact that HDI across the world has on average been steadily increasing. In fact, it is possible to see from the chart that there hasn't been a single year in the given period in which average global HDI has not increased.

Dataset: hdi_pivoted.csv

3.1.4 Min/Max HDI

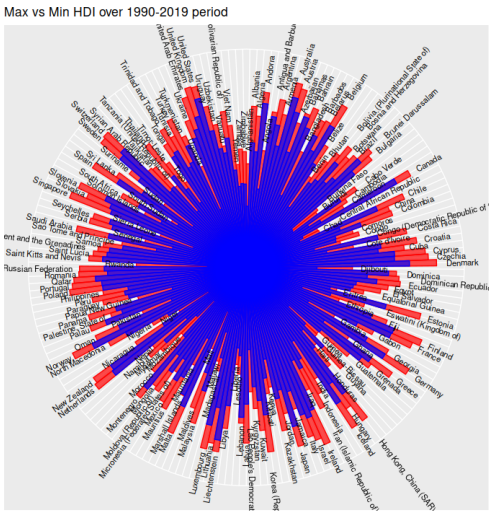

Figure 3.4: Max HDI over period in red, Min HDI over period in blue

The radial plot in Figure 3.4 was created using ggplot2 in R. The name of the script used is **plots.R**. The length of a bar represents the HDI score for the country whose name appears at the end of the bar. The blue bars correspond to the minimum value a country's HDI score over the 1990-2019 period and the red bars correspond to the maximum HDI score over the same period.

This plot gives an indication of which countries are developing at the fastest rates over the last 30 years, but it does not indicate whether that development is for the better or worse. Example: Iceland has quite a lot of red bar showing and it would be safe to say that's due to better living conditions with how popular it is to travel to Iceland for its hot springs and unique climeate, but Libya also has a lot or red bar showing and this is most likely due to the country taking a steep turn for the worst when the Libyan Civil war started in 2011.
Dataset: hdi_pivoted.R

3.1.5 Male/Female Mean Years Education

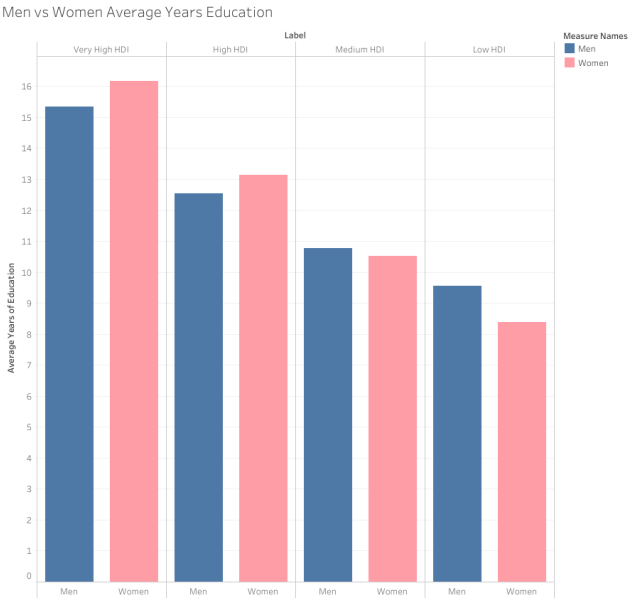

Figure 3.5: Male mean years education represented by blue, female pink

The side-by-side bar chart in Figure 3.5 shows the mean years of education for men (blue) and women (pink) by HDI label, very high to the left and low to the right.

This chart shows an an interesting trend in that, typically, a country going from medium to high HDI happens in tandem with women overtaking men in terms of number of years of education. More interestingly this effect is excacerbated at the ends of HDI Level, with women spending even more or less years in education respectively than at the middle two levels.
Dataset: CS3205_2020_statistical_annex_all.xlsx

3.1.6 Hypotheses for Patterns Found

Hypothesis 1
My first hypothesis is that *there are more countries that have a good standard of living than do not*. I believe this is to case from having observed the patterns that: the distribution of HDI has more mass in the positive direction of the HDI scale; the HDI global heat map has much more green on it than it does red; and the fact that global average HDI is trending upwards.

I believe that this hypothesis is justified on the basis that a country having high or very high HDI Levels should be a good proxy for saying if that country has a good standard of living.

Hypothesis 2
My second hypothesis is that *the primary factor which determines a country's HDI is its climate*. I believe this

to be the case from having observed the pattern that the reddest regions in the HDI global heatmap consistently overlap with regions of desert or dry savannah, which are notoriously inhospitable locations to live.

The only exception to this on the map is Australia, but it is worth noting that Australia was colonised relatively recently by Europeans on a historic scale and also that the majority of Australians live on its west coast, far from its most inhospitable deserts.

Hypothesis 3

My third hypothesis is that *in countries where the standard of living is not high, women are encouraged to quit school earlier than men*. I base this hypothesis on observing the pattern in the male/female mean years education bar chart of women staying in school longer than men in more developed countries and women staying in school shorter than men in less developed countries.

From my own experience as a citizen of a first world country, both boys and girls are equally encouraged when it comes to schooling. The fact that women stay in school longer then suggests that women value education more then men do on average. Knowing this, it seems like that external societal pressures play a role in them dropping out sooner in less developed countries, as their natural inclination seems to be to enjoy education.

Hypothesis 4

My fourth and final hypothesis, which perhaps belongs in the realm of speculation, is that *in the future women globally will be significantly more educated than men on average*. I come to this hypothesis by observing that currently men and women seem to be about equally educated on average when looked at from a global perspective (see Section 3.1.5) and the fact that the global average HDI is steadily increasing over time (see Section 3.1.3).

Using these two observed patterns together it seems natural to conclude that in the future HDI levels have increases even further and more and more countries start to have high and very high HDI Levels, the number of women across the globe with significantly more years of education than their male counterparts will only get higher and higher.

3.2 Specific Observations

In the specific observations section of the assignment I have created visualisaions in line with the task list specifications.

3.2.1 Spurious Values

Expected Population Growth

Figure 3.6: Distribution of future to current population ratio by country, colour represents HDI Level

For the boxplots in figure 3.6, I have annotated the highest and lowest population growth factor countries for each HDI level. Two of the annotated countries which I was very surprised by were Lithuania (lowest growth factor of Very High HDI countries) and Palestine (highest growth factor of High HDI countries).

My reasoning for being surprised at Lithuania's placing is that when I was attending secondary school I knew as many Lithuanians as I knew Polish students, and Ireland is a known historically popular destination for Polish emigrants. I believed that with their being so many Lithuaning immigrants in Cork, Lithuania's population must be booming. However it appears I was wrong and there are actually more Lithuanians leaving the country than there are to replace them.

The reason I was also surprised by Palestine's placing is because with the prosecution of Palestinian by Israel I thought that Palestine's future looked far too bleak to support a high level of population growth. It may be that Israel's treatment of Palestine has deteriorated so fast that the fact the HDR was published in 2020 which is 2 years ago may not account for current Palistinian affairs.

Health Expenditure ordered by Healthy Life Expectancy

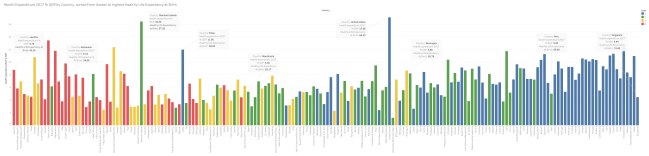

Figure 3.7: Health expenditure as a fraction of GDP, from left to right in order of increasing average life expectancy

The two countries which most stood out to me as being spurious when plotting health expenditure as a percentage of GDP, and then ordering the countries by healthy life expectancy were Marshall Islands and Palau.

The thing that I found most interesting about the Marshall Islands and what set it apart from most other countries was that its health expenditure as a fraction of GDP was only topped by the United States of America and yet despite spending all that money it still only has an average healthy life expectancy of 57 years, which does not even place it out of the bottom quarter countries.

The reason I consider Palau to be an outlier then is that despite being a Very High HDI country and its health expenditure as a percentage of GDP being in line with other Very High HDI countries, its average life healthy life is by far the lowest of all Very High HDI countries at only 59 years. If I had to suggest a reason for this, despite being completely unfamiliar with the country before now, I would say it could be down to a large societal class diving existing in the country, similar to the United States which also has a lower than average healthy life expectancy for Very High HDI countries.

3.2.2 Correlated Attributes

Fertility Rate vs Adolescent Birth Rate

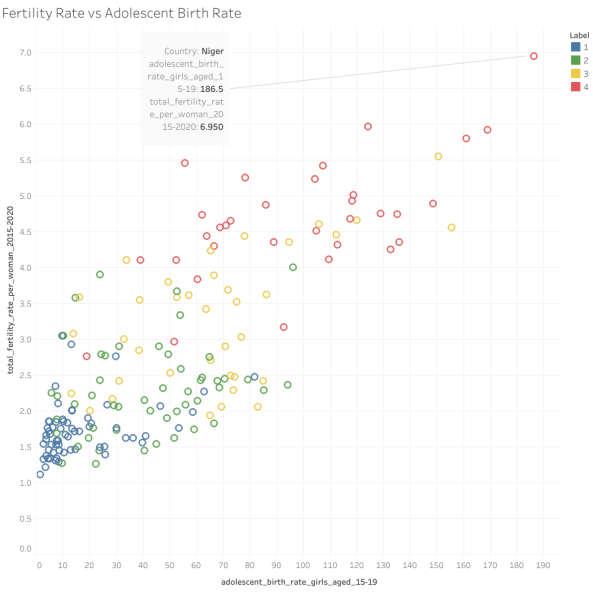

Figure 3.8: Fertility rate highly correlates with adolescent birth rate, colour represents HDI Level

The scatterplot in Figure 3.8 shows us that there is a clear linear relationship between the number of children a woman will have and whether she has her first child while still an adolescent. By looking at the colours of the points in the plot as well you can see a correlation between the fertility rate and a country's HDI Level as well as adolescent birth rate and a country's HDI Level.

From the scatterplot in Figure 3.9, it can be seen that HDI is extremely correlated with a countries GNI. The relationship is almost so perfectly linear that I would be surprised if GNI was not the heaviest weighted variable when it came to calculating a country's final HDI score.

HDI vs Gross National Income (GNI)

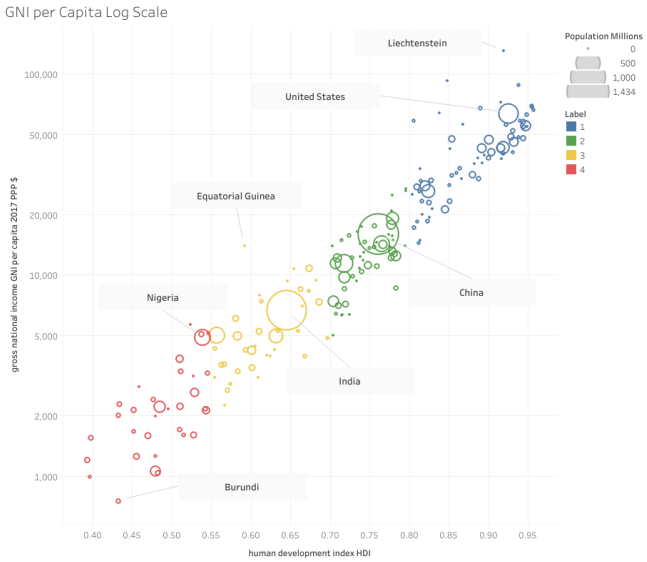

Figure 3.9: GNI accounts almost totally for HDI, colour represents HDI Level, size represents total population

3.2.3 Uncorrelated Attributes

Women Share of Seats in Government vs Average Years Education

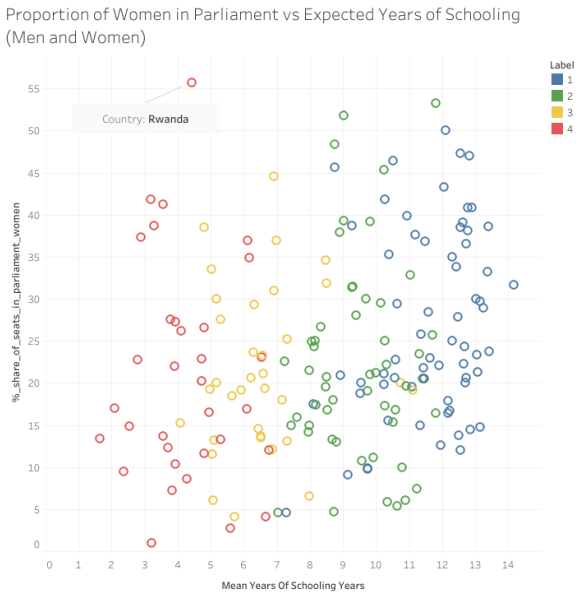

Figure 3.10: Average years of schooling plotted against percentage of women in government, colour represents HDI Level

The scatterplot in Figure 3.10 shows that the percentage of seats in government occupied by women in a country is completely uncorrelated with the average years of education of its people and that percentage of seats in government occupied by women is also not correlated with HDI level.

The country of Rwanda which is marked in the plot is a rather interesting outlier in having such a high proportion of women in government especially being a low HDI Level country. I am not sure if this is just down to lasting cultural reasons or has anything to do with the Rwandan genocide.

Labour Force Participation vs GDP

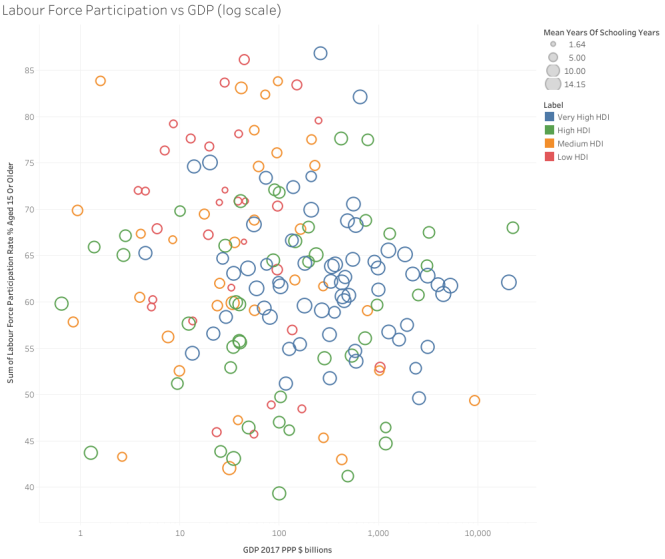

Figure 3.11: Labour force participation plotted against GDP, colour represents HDI Level, size represents mean years schooling

Another set of uncorrelated variables shown in Figure 3.11 are labour force participation rate and GDP. Even adding in mean years of schooling as a size attribute for each point still does not appear to materialise any pattern among country's labour force participation rates.

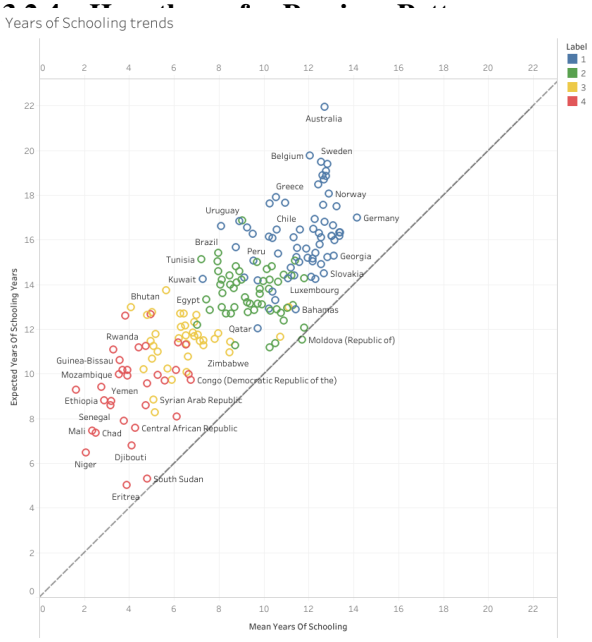

Figure 3.12: Amount of education is expected to rise, colour represents HDI Level

Expected Correlation 1

Figure 3.12 shows us that the expected mean years of education in the future for a country is highly correlated with its current mean years of education. We also see that for every country except Moldova for some reason, expected mean years of education is higher than current mean years of education.

This correlation makes perfect sense as we would not expect a country which currently have five years less education than another country to all of a sudden catch up. The general pattern is that expected years education is set to rise by about four years across the board. It is worth noting that expected years education can not be greater than current forever since ceiling effects will start to come into play eventually, even if there are no signs of it yet.

Expected Correlation 2

The second expected correlation I found was the correlation between fertility rate and adolescent birth rate found in Figure 3.8. From a purely mathematical view point this correlation makes perfect sense as the sooner your first child, the longer the window to have your second child and so on. There is also the biological factor that fertility is highest when women are young and lessens with time, so women have the easiest time conceiving when they are younger.

Unexpected Lack of Correlation 1

It was entirely unexpected in Figure 3.10 to see that there is a complete lack of correlation between percentage of women in government and mean years of education. This lack of correlation stands in start contrast to women having greater participation in education in more developed countries. Based on this I would hypothesise that female participation in government is entirely a cultural issue and is not necessarily indicative of a country's development in standard of living.

Unexpected Lack of Correlation 2 I was also shocked that a country's labour force participation rate had no correlation with that country's GDP. It seemed natural to me that country's that produced more goods would have more jobs. Labour force participation rate does not even seem to correlate with HDI Level.

I can't think of any singular reason that could cause this. I think it is more likely that there are multiple factors that affect labour force participation rate independent of a country's GDP and each country is experiencing a high degree of one or two of those factors, shifting them up and down the scale seemingly at random. Possible reasons could be a society becoming highly automated with a great standard of living versus a war which cripples a country's economy.

3.2.5 Usefulness of All Visualisations

Overall I am happy with usefulness of the visualisations I chose. In my opinion scatterplots are the most useful since between colour, size and shape of each point, you can fit as many as five dimensions onto a single plot and still reasonably compare them. And even when just comparing two variables the density of points you can fit into a small space makes scatterplots extremely readable.

If I had to pick a least useful visualisation it would be the radial plot, mainly because I lacked the ggplot2 skills to truly make it look nice. Even the attributes I chose to look at min and max hdi over the period are more novelty than truly informative since you do not have a sense of direction of time.

I very much lke the very long bar charts when viewed in tableau since it properly supports dynamic scrolling while maintaining visibility of the axes. I must admit though in the two column pdf setup of the report, very long bar charts are completely unreadable. Luckily for the bar charts it is possible to have live links for every visualisation so I feel there is no information lost, only a minor inconvenience caused.

3.2.6 Alternative Visualisations

Min/Max HDI Alternative

Figure 3.13: Click off white rectangle above to view in Tableau or navigate to visualisations folder

Alternative to: Figure 3.4, radial plot
For comparing Min vs Max HDI Figure 3.13 side by side circle plots are an interesting alternative to radial plots since despite looking so dissimilar on the surface they are just one functional transformation away from becoming one another, by unwinding the circumference of the circle. The side by side circle plots are definitely easier to interpret than the radial plots, but that is at the expense of compactness which is also useful in a visualisation. Insight-wise I would say these visualisations are the same.

Mean vs Expected Years Schooling

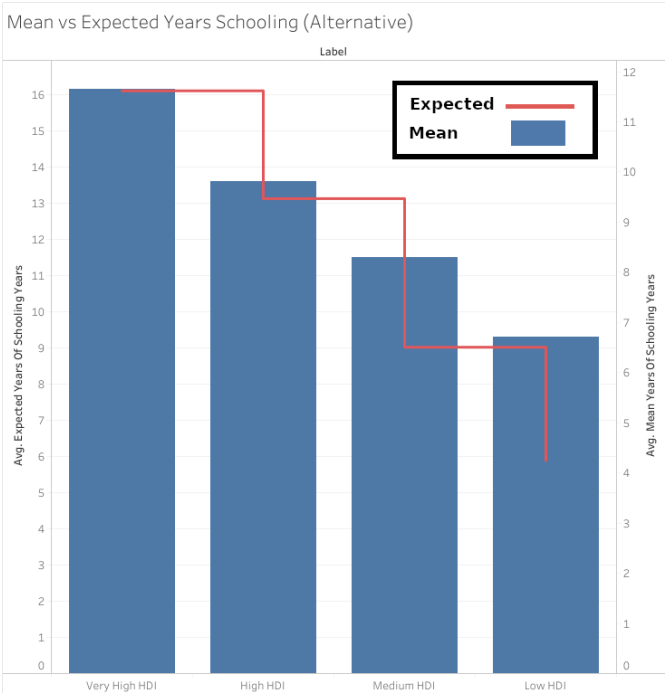

Figure 3.14: Lose individual countries

Alternative to: Figure 3.12, scatterplot
The bar and line graph Figure 3.14 is a very good alternative to the scatterplot as it can provide new and meaningful insight due to very clearly aggregating the data by label, all the while not sacrificing readability or compactness. An example of an insight which would be immediately obvious in the bar and line graph and not so in the scatterplot is the largest gap by Level between expected and mean years education. Unfortunately for me I forgot to make sure both axes are on the same scale and comparing the height of the bars to the height of the line is completely meaningless.

4 Task 2

My results for task 2 are mainly in the form of images rather than text since images are most illustrative of what I am trying to show and I lack the necessary academic background to talk about the exact technical differences between projection techniques to any meaningful degree.

The main thing I would say I accomplished while doing out task 2 besides very savvy with my folder structure and file naming scheme is just the skill of learning a niche piece of academic software to a usable degree, even without fully understanding how it works. I believe I demonstrated this by being able to create noticeably different projections for the same datasets only by adjusting a techniques parameters.

It was also fun to see the names of advanced distance techniques beyond euclidean distance such as city-block, cosine-based and in particular the cool sounding infinity-norm.

4.1 Corel Projections Comparison

4.1.1 LSP

Figures 4.1-4.6 provide an overview of the different projections I did while changing the parameters of LSP. Images which were hard to separate are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

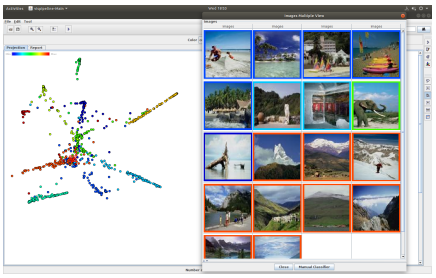

Figure 4.1: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 10, No: Neighbours: 10, Dissimilarity: Euclidean

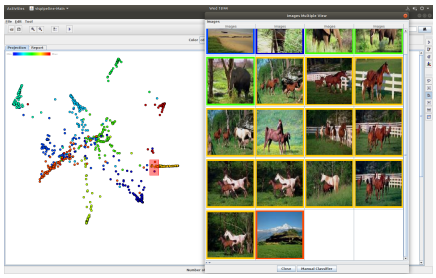

Figure 4.3: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 12, No: Neighbours: 10, Dissimilarity: Cosine

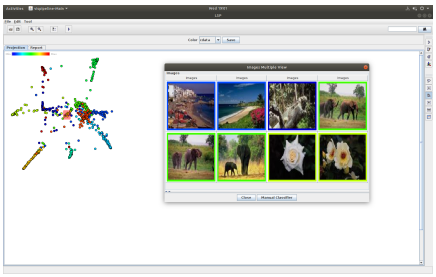

Figure 4.5: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 10, No: Neighbours: 15, Dissimilarity: Euclidean

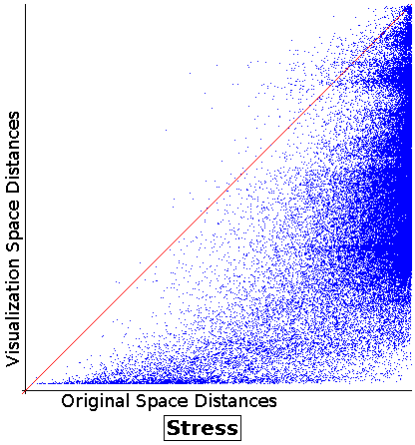

Figure 4.2: Cosine Stress Curve

Silhouette Coefficient: 0.5155

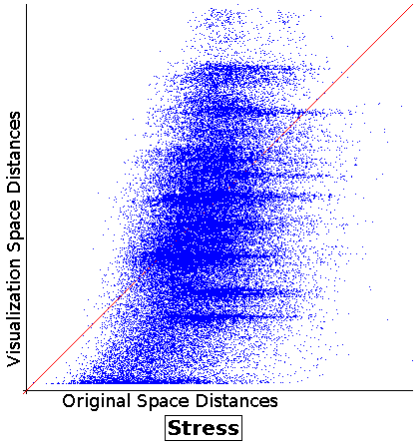

Figure 4.4: Euclidean Stress Curve

Silhouette Coefficient: 0.4635

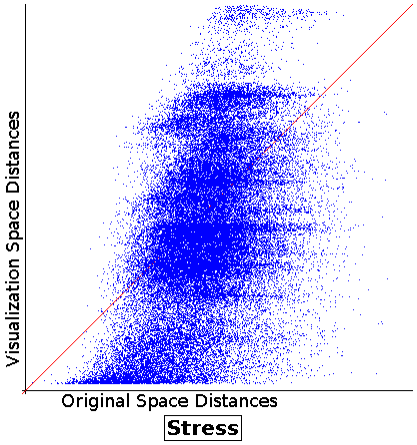

Figure 4.6: Euclidean Stress Curve

Silhouette Coefficient: 0.4762

4.1.2 t-SNE

Figures 4.7-4.12 provide an overview of the different projections I did while changing the parameters of t-SNE. Images which were hard to separate are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

One thing I have noticed about t-SNE from using it is that while it's very good at creating clusters, it is noticeably slower than LSP for larger, higher dimension datasets.

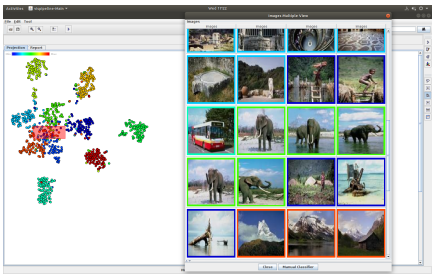

Figure 4.7: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Euclidean

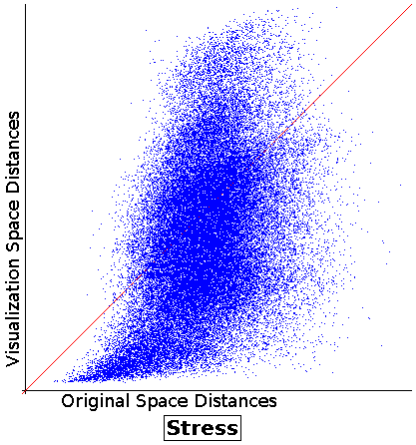

Figure 4.8: Euclidean Stress Curve

Silhouette Coefficient: 0.4693

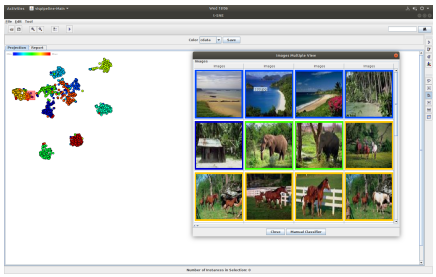

Figure 4.9: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Cosine

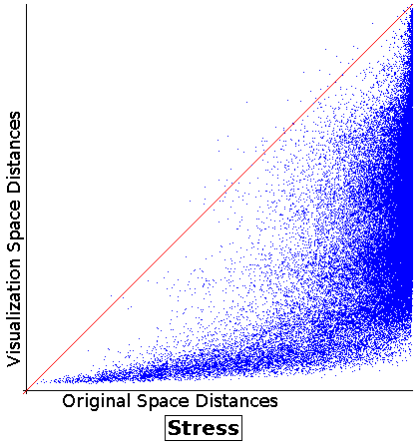

Figure 4.10: Cosine Stress Curve

Silhouette Coefficient: 0.4961

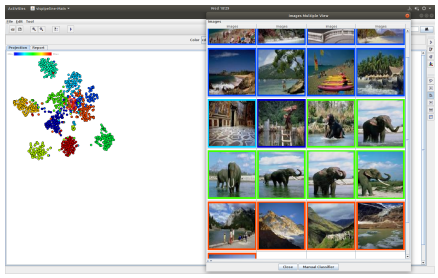

Figure 4.11: Initial Dimensions: 15, Target Dimension: 2, Perplexity: 60, Max No. Iterations: 1000, Dissimilarity: Euclidean

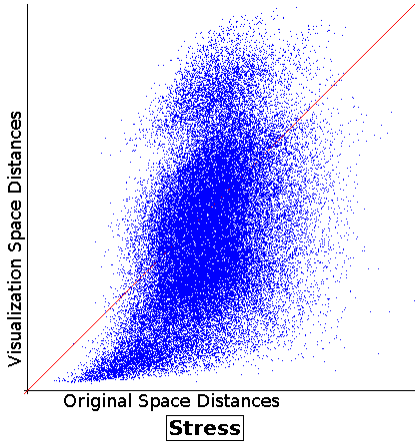

Figure 4.12: Euclidean Stress Curve

Silhouette Coefficient: 0.4484

4.1.3 PCA

Figures 4.13-4.14 provide an overview of the different projections I did using Principal Component Analysis (PCA). Images which were hard to separate are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

PCA is noticeably worse at separating out clusters of images than LSP and t-SNE but one thing I do find interesting that PCA shows is that the (faded) orange points which I unfortunately can not confirm at home but I believe were buildings are completely and utterly like all the other images in the dataset.

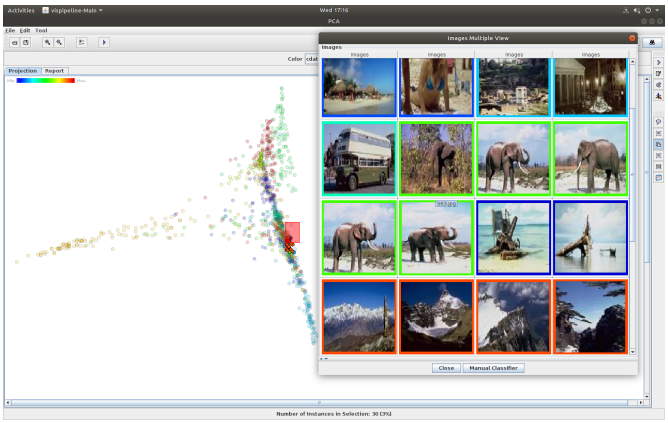

Figure 4.13: No parameters

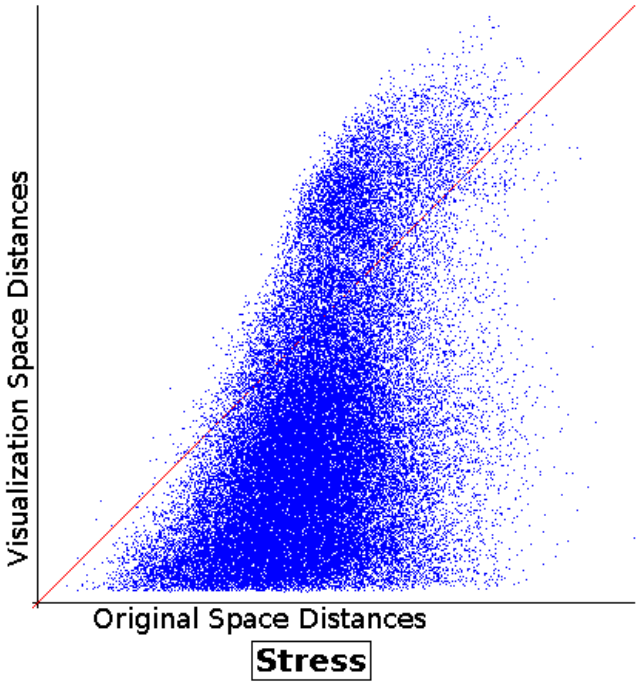

Figure 4.14: Stress Curve

Silhouette Coefficient: 0.5201

4.2 CBR Projections Comparison

4.2.1 LSP

Figures 4.15-4.20 provide an overview of the different projections I did while changing the parameters of LSP. Papers which were hard to separate are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

Figure 4.15: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 10, No: Neighbours: 10, Dissimilarity: Euclidean

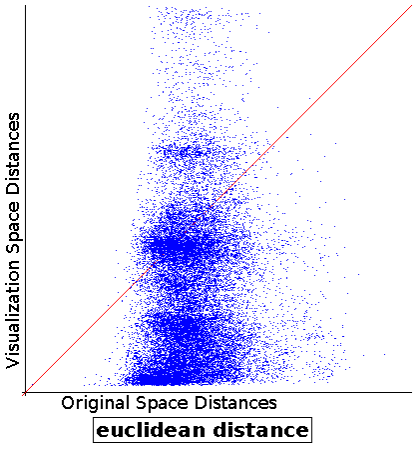

Figure 4.16: Stress Curve

Silhouette Coefficient: 0.3550

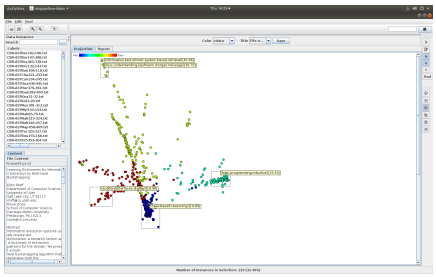

Figure 4.17: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 10, No: Neighbours: 10, Dissimilarity: Cosine

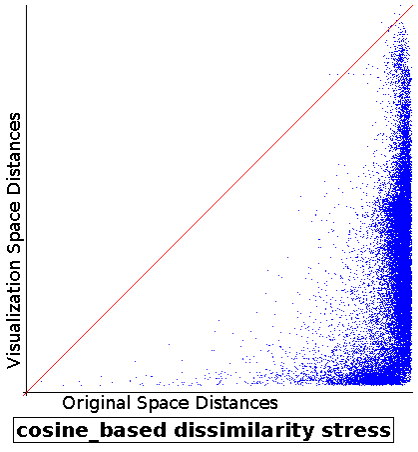

Figure 4.18: Stress Curve

Silhouette Coefficient: 0.5100

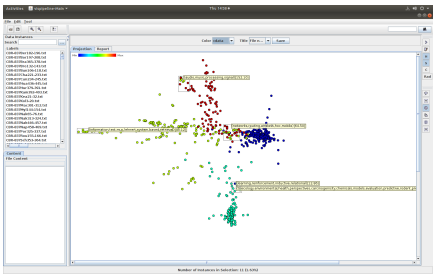

Figure 4.19: No. Iterations: 80, Fraction of Delta: 8.0, No. Control Points: 20, No: Neighbours: 12, Dissimilarity: Cosine

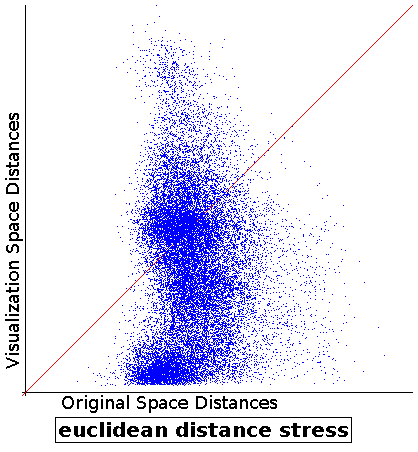

Figure 4.20: Stress Curve

Silhouette Coefficient: 0.5116

4.2.2 t-SNE

Figures 4.21-4.26 provide an overview of the different projections I did while changing the parameters of t-SNE. Papers which were hard to separate are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

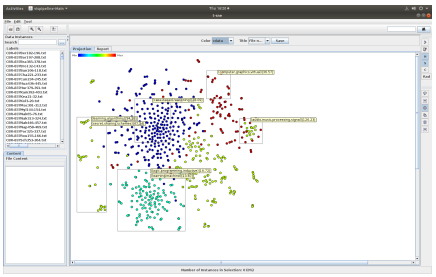

Figure 4.21: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Euclidean

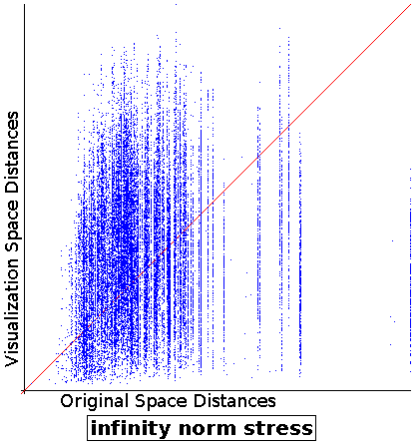

Figure 4.22: Stress Curve

Silhouette Coefficient: 0.2731

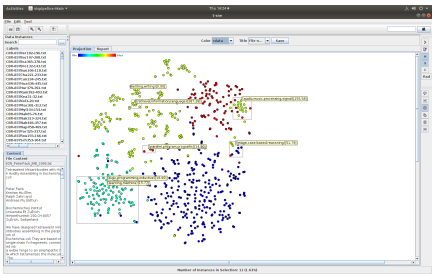

Figure 4.23: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 60, Max No. Iterations: 1500, Dissimilarity: Cosine

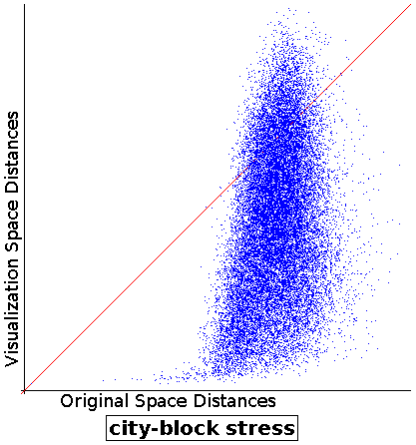

Figure 4.24: Stress Curve

Silhouette Coefficient: 0.3481

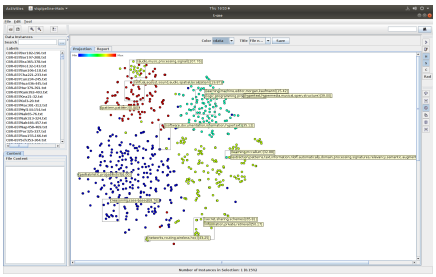

Figure 4.25: Initial Dimensions: 500, Target Dimension: 2, Perplexity: 100, Max No. Iterations: 2000, Dissimilarity: Cosine

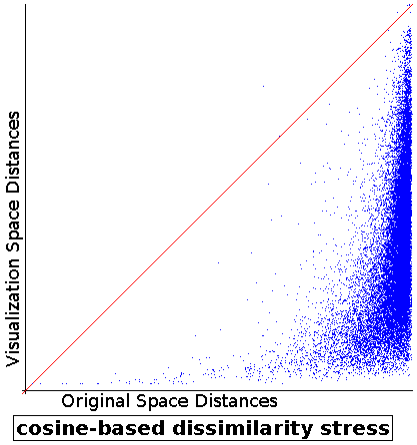

Figure 4.26: Stress Curve

Silhouette Coefficient: 0.2612

4.2.3 ProjClus

Figures 4.27-4.32 provide an overview of the different projections I did while changing the parameters of ProjClus. Papers which were hard to separate are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

I used Projection Clustering instead of PCA since PCA was incredibly slow computing the eigenvectors for the highly multidimensional text datasets. I really liked projection clustering since even though it isn't the best at separating labels, the clusters it makes do typically have a lot in common and the projections can vary wildly with the parameters.

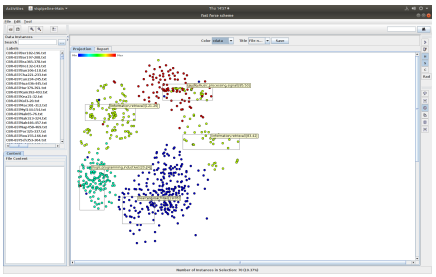

Figure 4.27: No. Iterations: 50, Fraction of Delta: 8.0, Cluster Factor: 4.5, Type of Projection: Nearest Neighbor Projection, Dissimilarity: Cosine

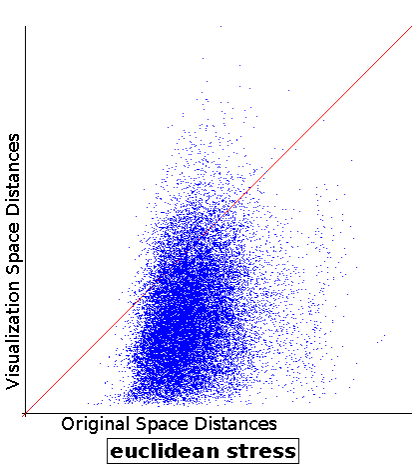

Figure 4.28: Stress Curve

Silhouette Coefficient: 0.3777

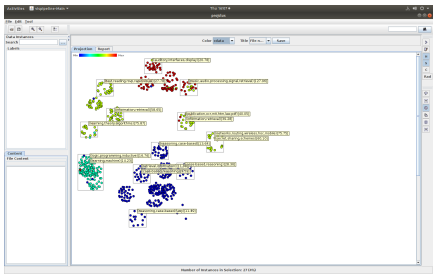

Figure 4.29: No. Iterations: 50, Fraction of Delta: 8.0, Cluster Factor: 9.0, Type of Projection: Nearest Neighbor Projection, Dissimilarity: Cosine

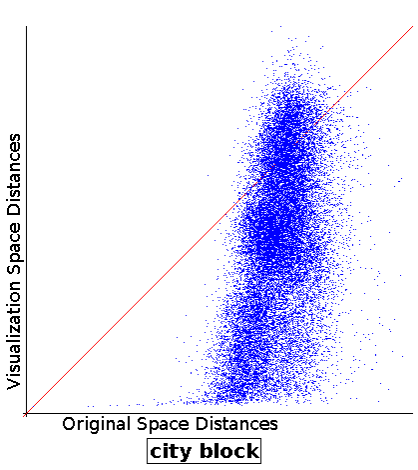

Figure 4.30: Stress Curve

Silhouette Coefficient: 0.4225

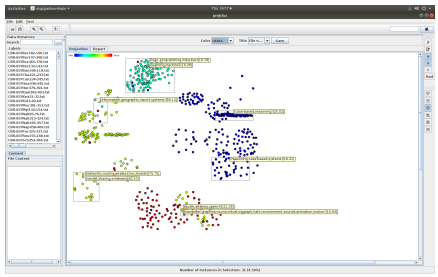

Figure 4.31: No. Iterations: 200, Fraction of Delta: 15.0, Cluster Factor: 7.0, Type of Projection: Fastmap, Dissimilarity: Cosine

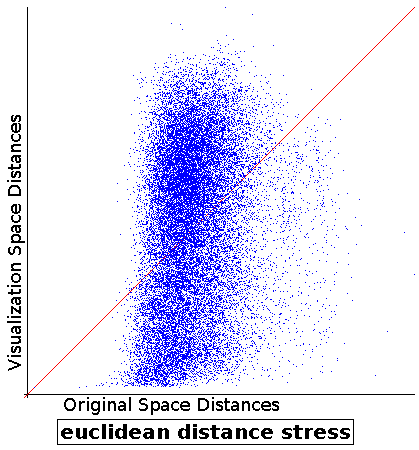

Figure 4.32: Stress Curve

Silhouette Coefficient: 0.4637

4.3 Medical Images Projections Comparison

It's worth noting that the medical images collection is the first set of images to score very poorly on both silhouette coefficient and stress curve metrics.

This does make sense as the images lack any colour and are comprised of very similar classes such as a skull viewef from two different angles, or a topdown x-ray of a skull but to two different depths.

4.3.1 LSP

Figures 4.33-4.38 provide an overview of the different projections I did while changing the parameters of LSP. Images which were hard to separate are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

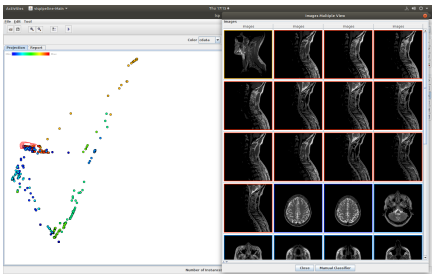

Figure 4.33: No. Iterations: 100, Fraction of Delta: 4.0, No. Control Points: 12, No: Neighbours: 80, Dissimilarity: Cosine

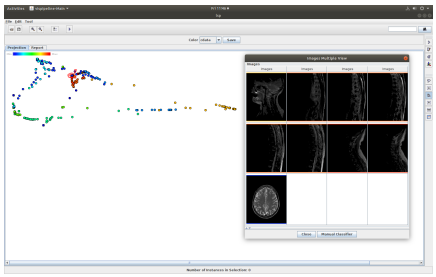

Figure 4.35: No. Iterations: 100, Fraction of Delta: 8.0, No. Control Points: 12, No: Neighbours: 6, Dissimilarity: Cosine

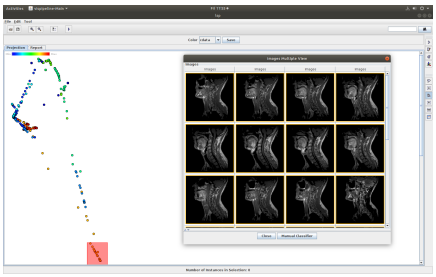

Figure 4.37: No. Iterations: 100, Fraction of Delta: 2.0, No. Control Points: 14, No: Neighbours: 6, Dissimilarity: Cosine

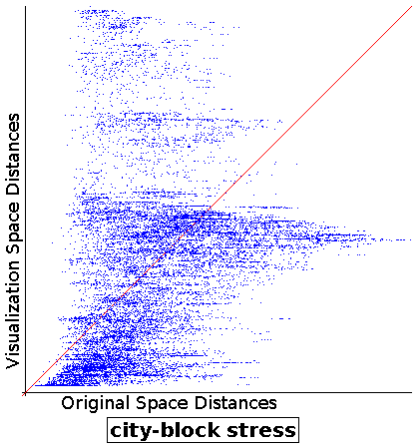

Figure 4.34: Stress Curve

Silhouette Coefficient: 0.0375

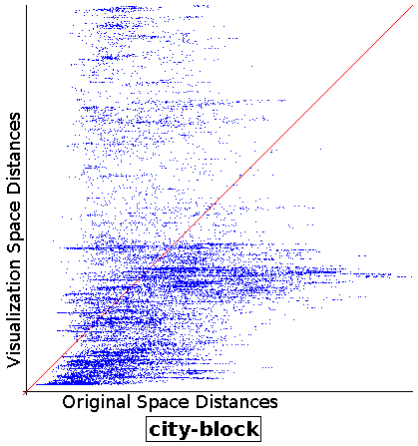

Figure 4.36: Stress Curve

Silhouette Coefficient: -0.01736

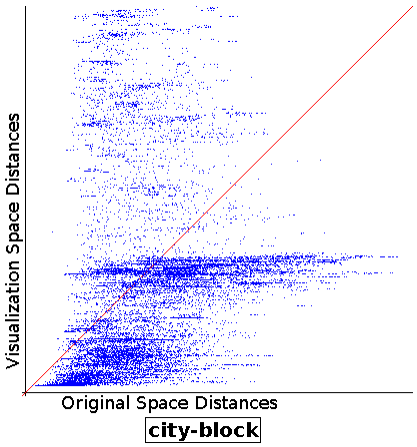

Figure 4.38: Stress Curve

Silhouette Coefficient: -0.0921

4.3.2 t-SNE

Figures 4.39-4.44 provide an overview of the different projections I did while changing the parameters of t-SNE. Images which were hard to separate are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

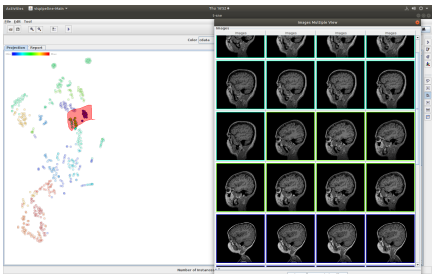

Figure 4.39: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Euclidean

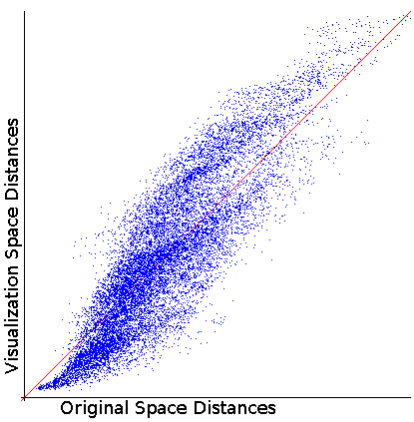

Figure 4.40: Stress Curve

Silhouette Coefficient: 0.1451

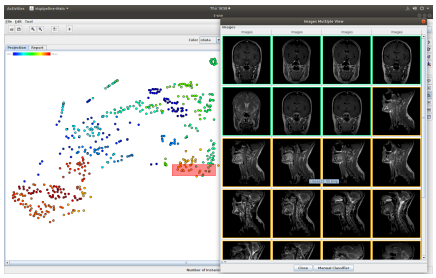

Figure 4.41: Initial Dimensions: 4, Target Dimension: 2, Perplexity: 60, Max No. Iterations: 1500, Dissimilarity: Euclidean

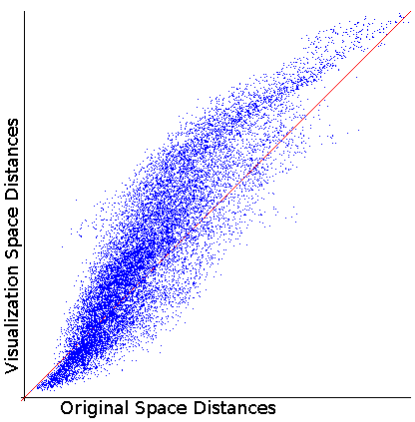

Figure 4.42: Stress Curve

Silhouette Coefficient: 0.1089

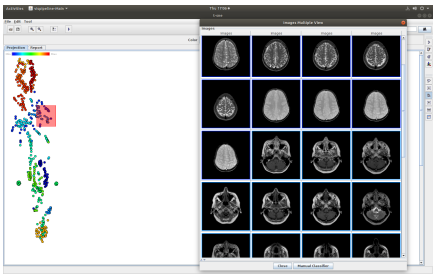

Figure 4.43: Initial Dimensions: 256, Target Dimension: 2, Perplexity: 50, Max No. Iterations: 1500, Dissimilarity: Cosine

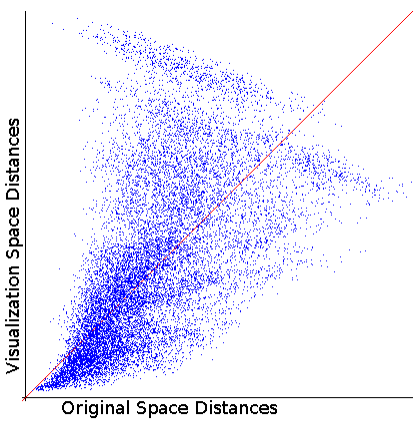

Figure 4.44: Stress Curve

Silhouette Coefficient: 0.0301

4.3.3 PCA

Figures 4.45-4.46 provide an overview of the different projections I did using Principal Component Analysis (PCA). Images which were hard to separate are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

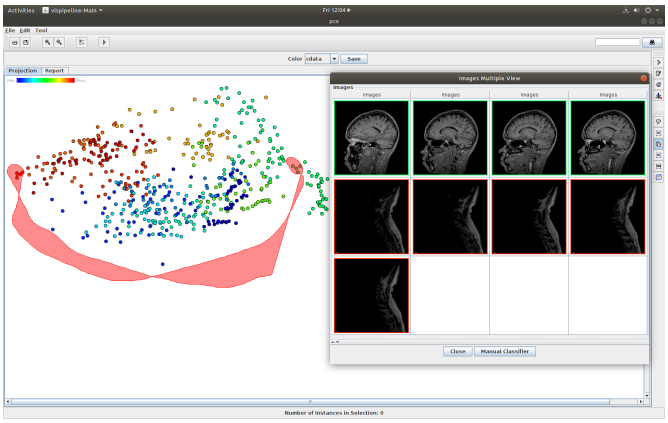

Figure 4.45: No parameters

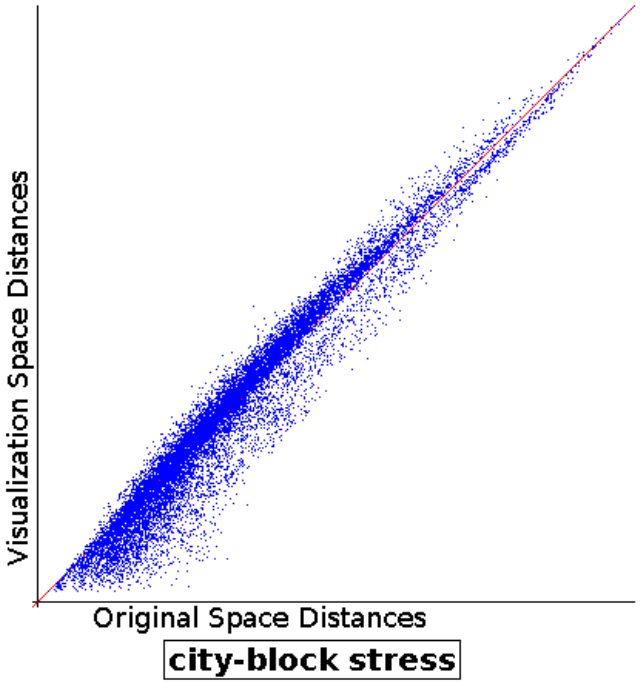

Figure 4.46: Stress Curve

Silhouette Coefficient: 0.3501

4.4 Headlines Projections Comparison

4.4.1 LSP

Figures 4.47-4.52 provide an overview of the different projections I did while changing the parameters of LSP. Articles which were hard to separate are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

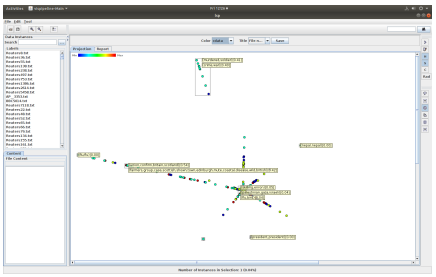

Figure 4.47: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 10, No: Neighbours: 10, Dissimilarity: Euclidean

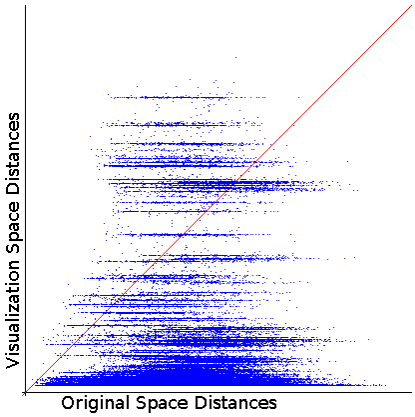

Figure 4.48: Stress Curve

Silhouette Coefficient: -0.1885

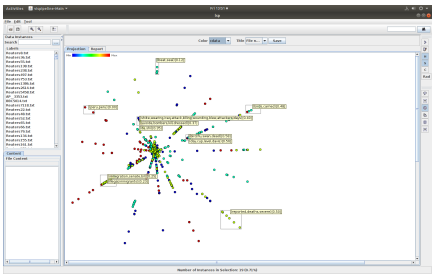

Figure 4.49: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 20, No: Neighbours: 10, Dissimilarity: Cosine

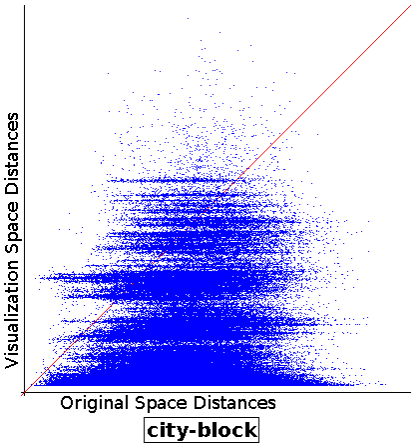

Figure 4.50: Stress Curve

Silhouette Coefficient: -0.0756

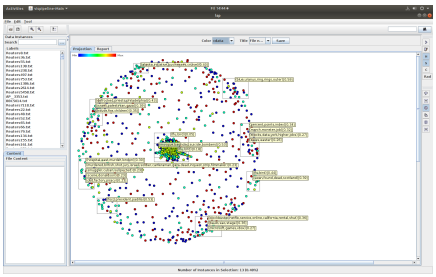

Figure 4.51: No. Iterations: 100, Fraction of Delta: 35.0, No. Control Points: 600, No: Neighbours: 200, Dissimilarity: Cosine

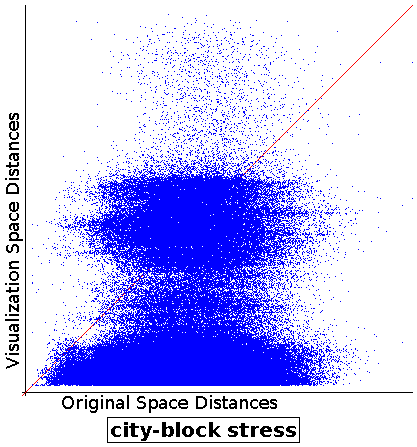

Figure 4.52: Stress Curve

Silhouette Coefficient: -0.1269

4.4.2 t-SNE

Figures 4.53-4.58 provide an overview of the different projections I did while changing the parameters of t-SNE. Articles which were hard to separate are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

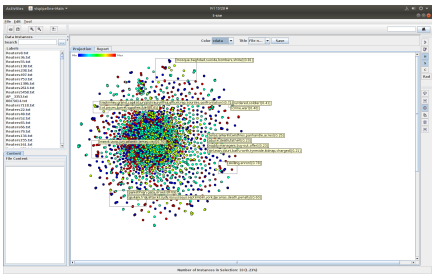

Figure 4.53: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Euclidean

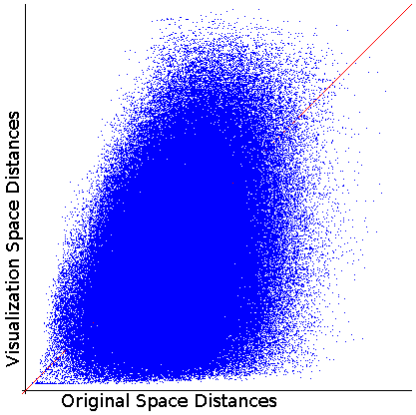

Figure 4.54: Stress Curve

Silhouette Coefficient: -0.0714

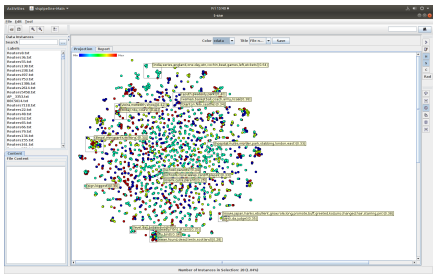

Figure 4.55: Initial Dimensions: 60, Target Dimension: 2, Perplexity: 40, Max No. Iterations: 5000, Dissimilarity: Cosine

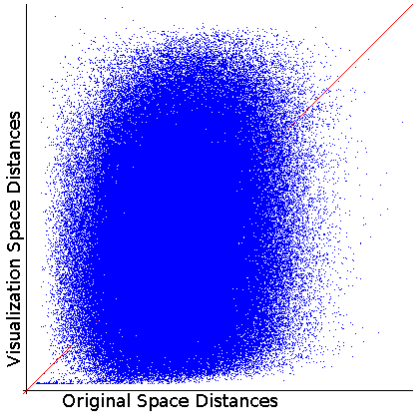

Figure 4.56: Stress Curve

Silhouette Coefficient: -0.0275

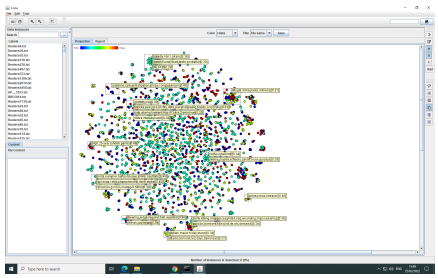

Figure 4.57: Initial Dimensions: 240, Target Dimension: 2, Perplexity: 80, Max No. Iterations: 750, Dissimilarity: Cosine

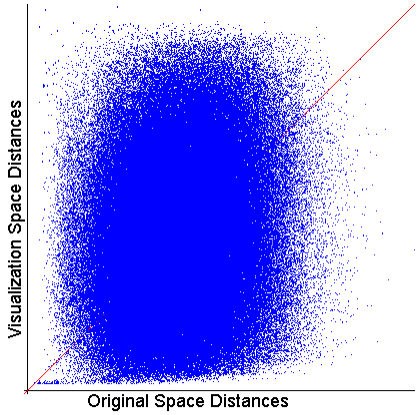

Figure 4.58: Stress Curve

Silhouette Coefficient: -0.0197

4.4.3 ProjClus

Figures 4.59-4.64 provide an overview of the different projections I did while changing the parameters of LSP. Articles which were hard to separete are included next to each projection in the figure graphic and can be viewed full size by clicking the picture as they are all live links.

For the highly clustered headlines in Figure 4.63 it's very interesting how high the similarity score for each article is. It does make me think that labeling news sources in this instance by company isn't particularly useful at clustering points together and that clustering by content is actually very effective.

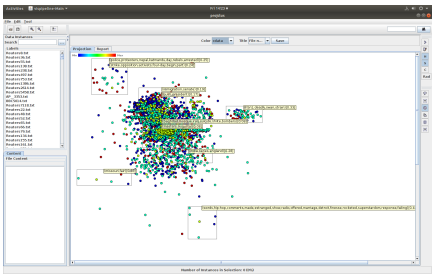

Figure 4.59: No. Iterations: 50, Fraction of Delta: 8.0, Cluster Factor: 4.5, Type of Projection: Nearest Neighbor Projection, Dissimilarity: Cosine

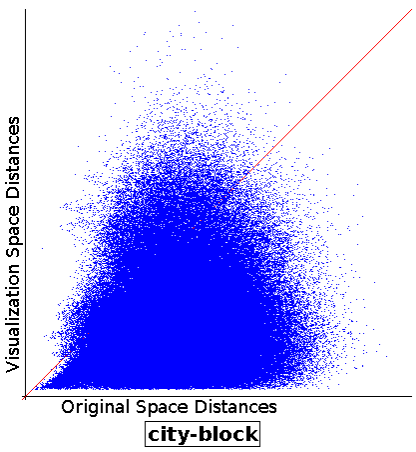

Figure 4.60: Stress Curve

Silhouette Coefficient: -0.1216

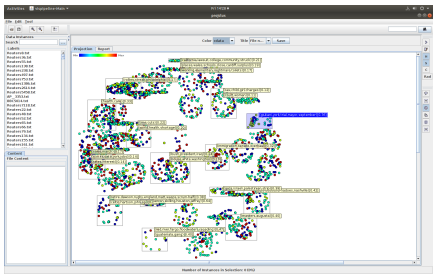

Figure 4.61: No. Iterations: 100, Fraction of Delta: 8.0, Cluster Factor: 8.0, Type of Projection: Nearest Neighbor Projection, Dissimilarity: Cosine

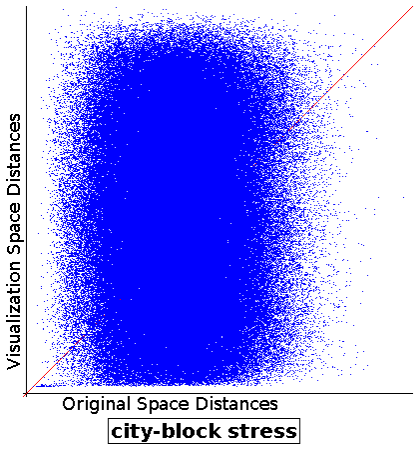

Figure 4.62: Stress Curve

Silhouette Coefficient: -0.04515

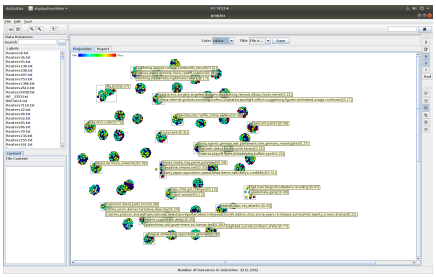

Figure 4.63: No. Iterations: 75, Fraction of Delta: 40.0, Cluster Factor: 20.0, Type of Projection: Nearest Neighbor Projection, Dissimilarity: Cosine

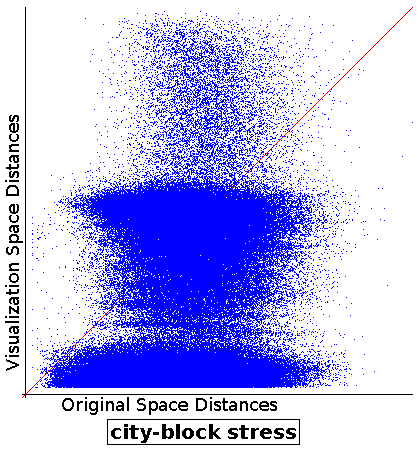

Figure 4.64: Stress Curve

Silhouette Coefficient: -0.0195

4.5 Projections of HDR

I did remove HDI Rank and HDI from my HDR dataset before making my hdr.date file. The reason being that these variables are so highly correlated with HDI Level that the projections and clusters would perfectly separate out the countries by HDI Level using HDI and HDI Rank.

I did however still include GDP. This seems completely innocent at first until you cast your mind back to Figure 3.9 which shows that GNI almost perfectly correlates with HDI, which means it almost perfectly correlates with HDI Level. In a plot not included in this report I but that is included in the visualisations folder of the GitHub repo I show that GDP also essentially perfectly correlates with HDI.

Because of this, the following graphs are all obviously incredibly good at separating out countries by HDI Level. If I could rub a genie's lamp and go back in time, I would remove GDP and GNI from by HDR dataset as well.

4.5.1 LSP

Figures 4.65-4.70 provide an overview of the different projections I did while changing the parameters of LSP. Countries are once again coloured by HDI Level.

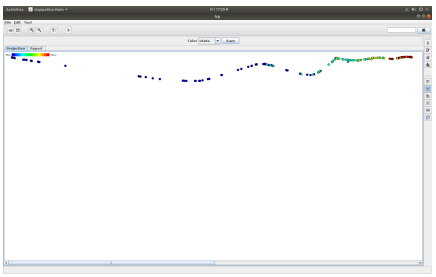

Figure 4.65: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 18, No: Neighbours: 10, Dissimilarity: Euclidean

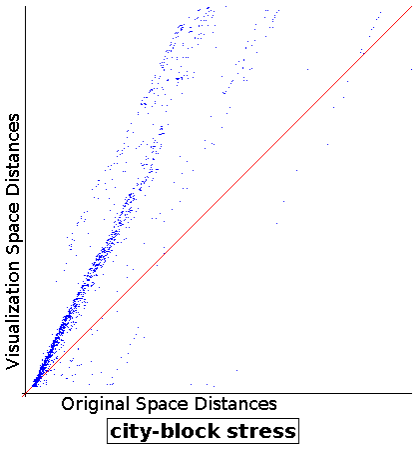

Figure 4.66: Stress Curve

Silhouette Coefficient: 0.2117

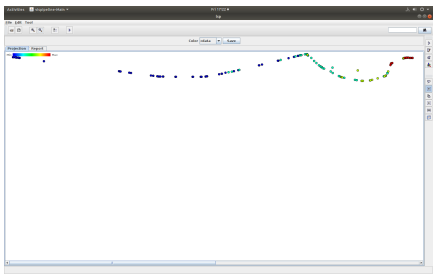

Figure 4.67: No. Iterations: 50, Fraction of Delta: 4.0, No. Control Points: 6, No: Neighbours: 8, Dissimilarity: Euclidean

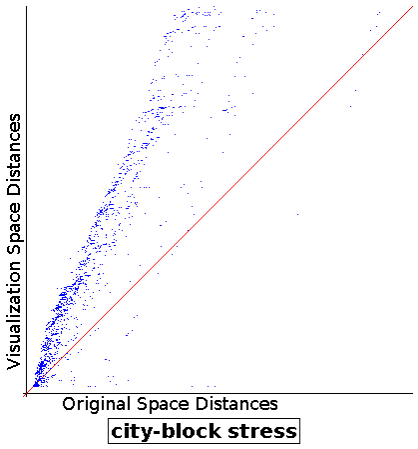

Figure 4.68: Stress Curve

Silhouette Coefficient: 0.2338

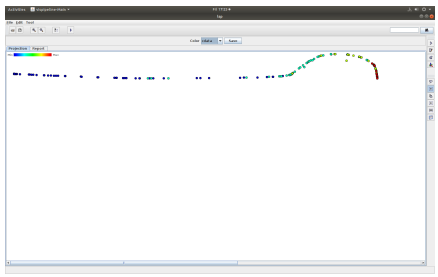

Figure 4.69: No. Iterations: 75, Fraction of Delta: 6.0, No. Control Points: 4, No: Neighbours: 15, Dissimilarity: Euclidean

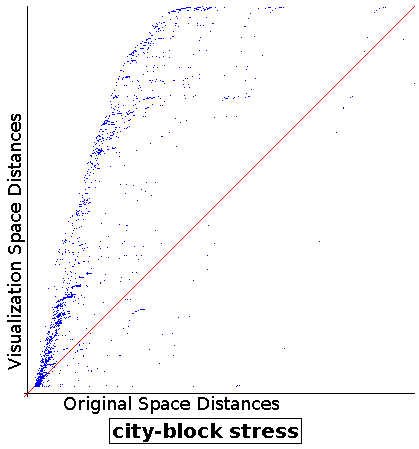

Figure 4.70: Stress Curve

Silhouette Coefficient: 0.2779

4.5.2 t-SNE

Figures 4.71-4.76 provide an overview of the different projections I did while changing the parameters of t-SNE. Countries are once again coloured by HDI Level.

t-SNE managed to make a loop out of red countries which is impressive that it managed to still find pick out some variation among countries despite GDP beind an all-powerful discriminating variable.

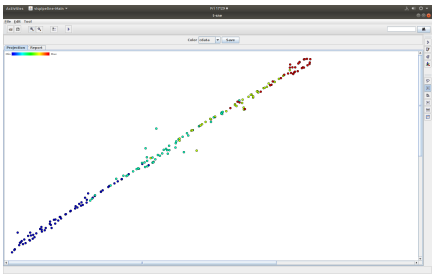

Figure 4.71: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Euclidean

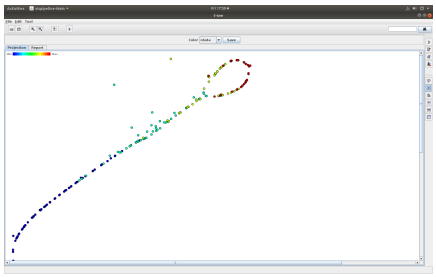

Figure 4.73: Initial Dimensions: 24, Target Dimension: 2, Perplexity: 100, Max No. Iterations: 1000, Dissimilarity: Euclidean

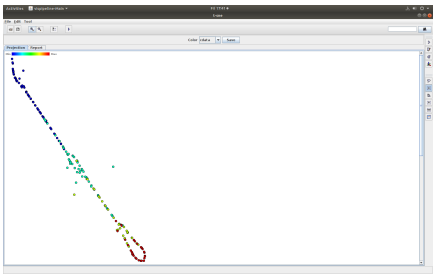

Figure 4.75: Initial Dimensions: 6, Target Dimension: 2, Perplexity: 60, Max No. Iterations: 1000, Dissimilarity: Infinity Norm

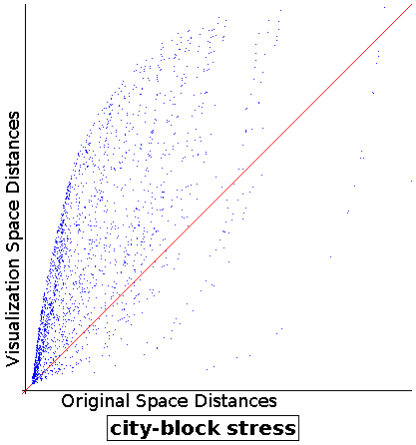

Figure 4.72: Stress Curve

Silhouette Coefficient: 0.3648

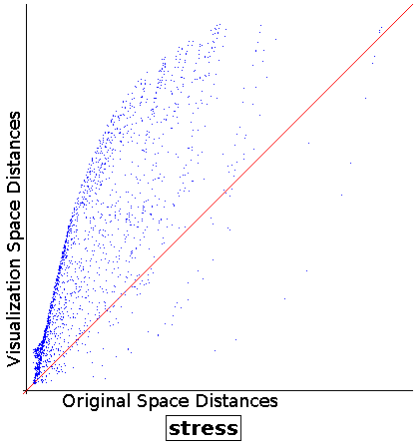

Figure 4.74: Stress Curve

Silhouette Coefficient: 0.3267

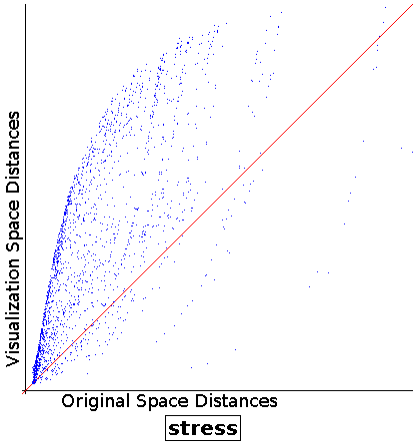

Figure 4.76: Stress Curve

Silhouette Coefficient: 0.3427

4.5.3 PCA

Figures 4.77-4.78 provide an overview of the different projections I did using Principal Component Analysis (PCA). Countries are once again coloured by HDI Level.

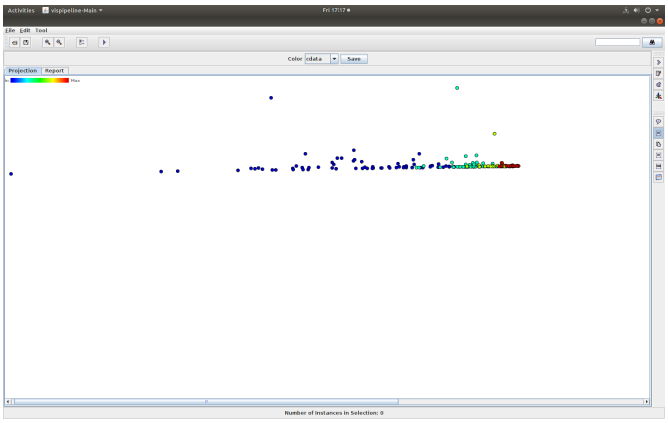

Figure 4.77: No parameters

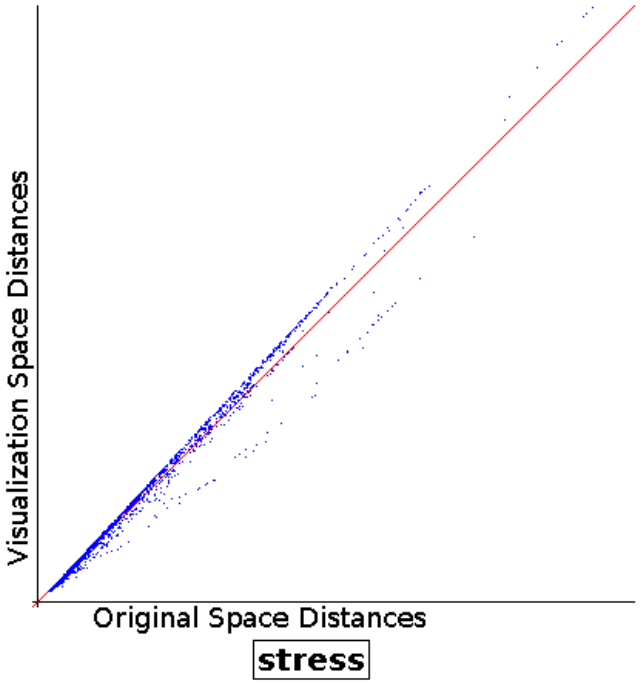

Figure 4.78: City-Block Stress Curve

Silhouette Coefficient: 0.1985

5 Conclusions

The datasets supplied to us for this assignment were very appropriate and had lots of depth information-wise without a large amount of heavy-duty cleaning which would have worked against the purposes of the assignment, namely visualisation.

The HDR dataset was particularly fun to use as it had so many variables associated with it anyone could pick out country attributes they were intereseted in and still have plenty of scope for creating numberous comparisons and visualisations. The fact that each variable also had a real-world representation meant that each visualisation really did mean something as well which is a great feeling to see that what you are doing is actually creating value.

Also, for every visualisation I did that had results I could have anticipated, there was another visualisation that would truly did take me by surprise and cause me to reevaluate how I think about certain country dynamics.

The multidimesional scaling techniques and projections of Task 2 were great to get experience with for me personally since I hope at some point in the future to either be a statistician or data scientist working with big data. Learning about the techniques which make sense of the large highly multidimensional datasets I hope to encounter in the future is a great stepping stone for whatever I decide to go on and study next.

For both tasks I would say the visualisations were extremely appropriate. For HDR each dimension corresponeded to an attribute so things like the measure of a point in a certain dimension were easily interpreted and for the VisPipeline datasets since they were both images and text documents, both of which are usually consumed visually, seeing the final projections and being able to see which words or images were most common in a group was very insightful and easy to digest.

6 References

HDR 2020 dataset, Human Development Data Center, https://hdr.undp.org/en/2020-report