2022

Visual Exploration of Survey Data

INTRODUCTION TO MULTIDIMENSIONAL PROJECTIONS

PABBI, KETHAN
121102356
CS6426

Introduction

In an era of advanced technology and data, there is ever growing of information to be stored or maintained constantly. The extraction and understandability of the information plays an important role for a company or individual. As the data volume is huge we would need an intermediator that simplifies the task of data interpretation. This is where data visualization plays a major role in simplifying data into visual ques which are easy on the human eye.

The dataset used in this report is Human Development Report (HDR) 2020. Making use of visualization tools such as Tableau and VisPipeline helps to represent multidimensional numerical in formats that a human can understand and interpret with ease.

Data Description

In task 1, the dataset used for this task is '2020_statistical_annex_all.xlsx'. It has 15 tables on various topics such as human development index, trends, work and employment, etc. I have extracted the following columns into 'Book1.xlsx': HDI rank, Country, Human Development Index (HDI) (value) 2019, Life expectancy at birth (years) 2019, Expected years of schooling (years) 2019, Mean years of schooling (years) 2019, Gross national income (GNI) per capita (2017 PPP$) 2019, Inequality-adjusted life expectancy index (Value) 2019, Inequality-adjusted income index (value) 2019, Gender Development Index (value) 2019, Human Development Index (female value) 2019, Human Development Index (male value) 2019, Life expectancy at birth (female years) 2019, Life expectancy at birth (male years) 2019, Expected years of schooling (female years) 2019, Expected years of schooling (male years) 2019, Mean years of schooling (female years) 2019, Mean years of schooling (male years) 2019, Estimated, Gross national income (GNI) per female capita (2017 PPP$) 2019, Estimated Gross national income (GNI) per male capita (2017 PPP$) 2019, Gender Inequality Index (value) 2019, Share of seats in parliament (% held by women) 2019, Labour force female participation rate (% ages 15 and older) 2019, Gross domestic product (GDP per capita 2017 PPP$) 2019, Total population (in millions) 2019, Label.

In task 2, made use of CBR, Coral, MedicalClasses, VSM and HDR datasets and used their data file to perform various visualization and analyze their projections.

Task 1

(a) Free Exploration:

Dataset used: HDI.csv. It contains the full series of Human Development Indices

(a.ii): Patterns Observed:

Human Development Index vs Gender Development Index (2019)(Fig 1.1):

We notice a higher value in gender development years with respect to human developed index in developed countries when compared to underdeveloped countries. There can be many factors affecting this case, being the developed countries have better education system and access to latest technology hence better chance of gender development. Also, the underdeveloped countries having poor services and lack of education might result in lower gender development index.

Human Development Index vs Gender Inequality Index (2019)(Fig 1.2):

We can see a decreasing somewhat linear trend in the relation between gender inequality index and human development index. The underdeveloped countries may have poor morale and social injustice due to unstable government hence gender inequality may rise. We can make a hypothesis that developed countries has better education and resources helped reduce the gender inequality index.

Human Development Index vs Inequality-Adjusted Income Index (2019)(Fig 1.3):

The "inequality-adjusted income index" is an income index calculated by considering inequality distribution factors to reflect justice and equality [1].

We can notice a huge difference between developed and underdeveloped countries based on Inequality Adjusted Income Index. I can hypothesize that countries like South Africa and Iran have similar values though they are of different labels because they have poor economic for growth.

Human Development Index vs Inequality-Adjusted Life Expectancy Index (2019)(Fig 1.4):

Inequality-Adjusted Life Expectancy Index combines a country's average achievements in health, education and income with how those achievements are distributed among country's population by "discounting" each dimension's average value according to its level of inequality [2]. From the packed bubbles we can observe that developed countries have higher IDHI life expectancy compared to underdeveloped countries. I can hypothesize that this is due to poor achievements in health education and income distributed among the population of the country.

(a.i): Visualizations:

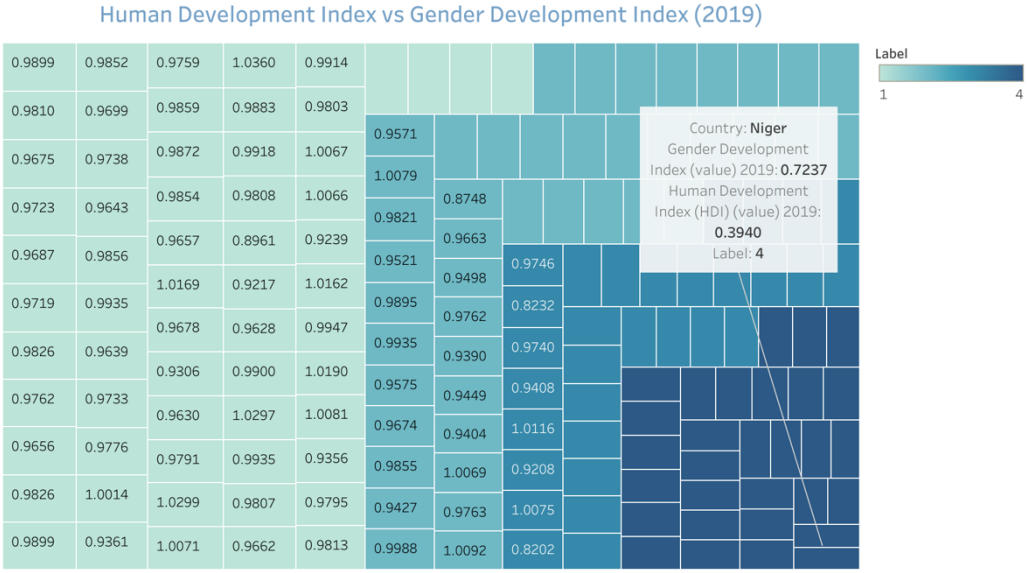

Fig 1.1. Human Development Index vs Gender Development Index (2019)

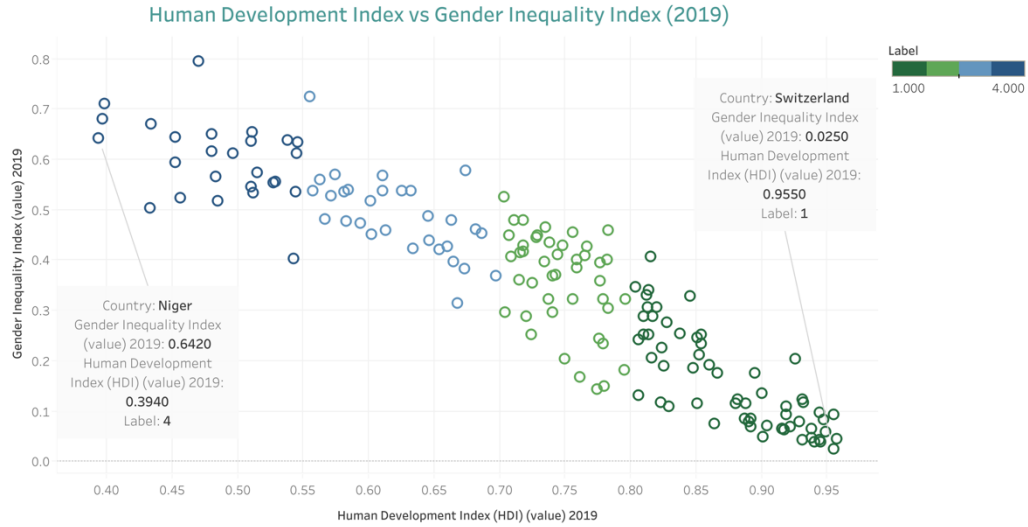

Fig 1.2. Human Development Index vs Gender Inequality Index (2019)

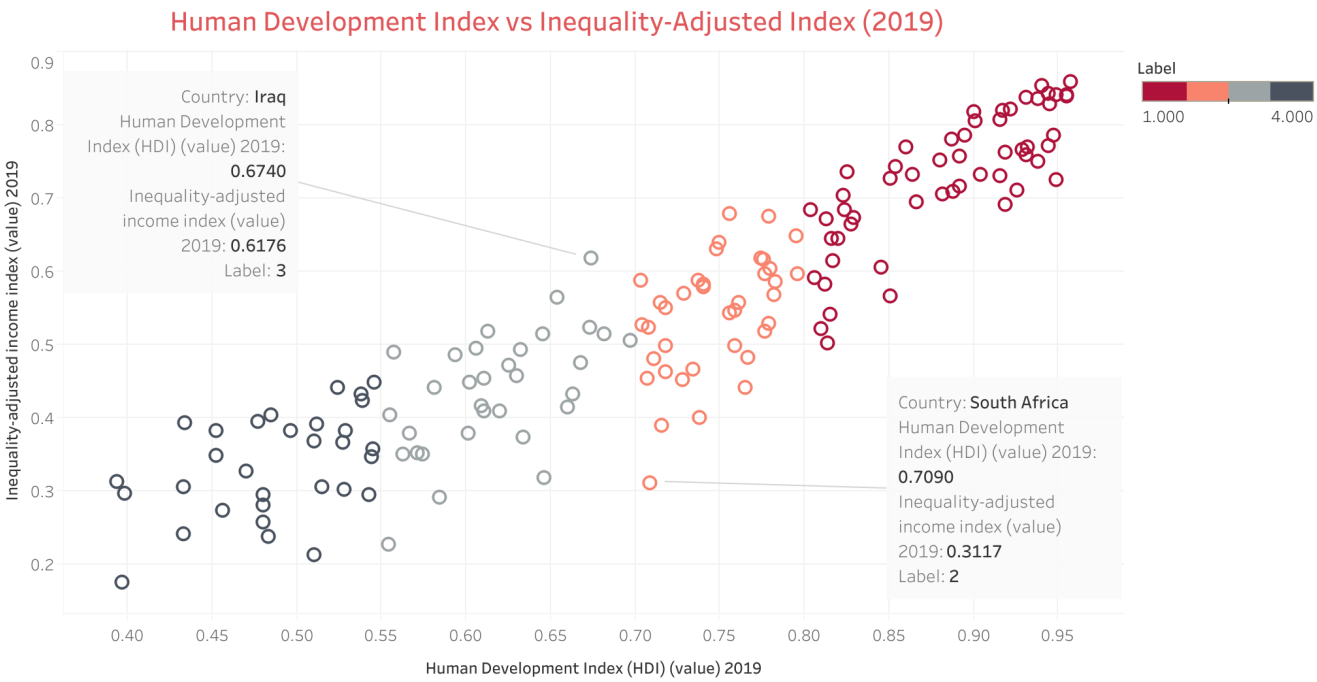

Fig 1.3. Human Development Index vs Inequality-Adjusted Income Index (2019)

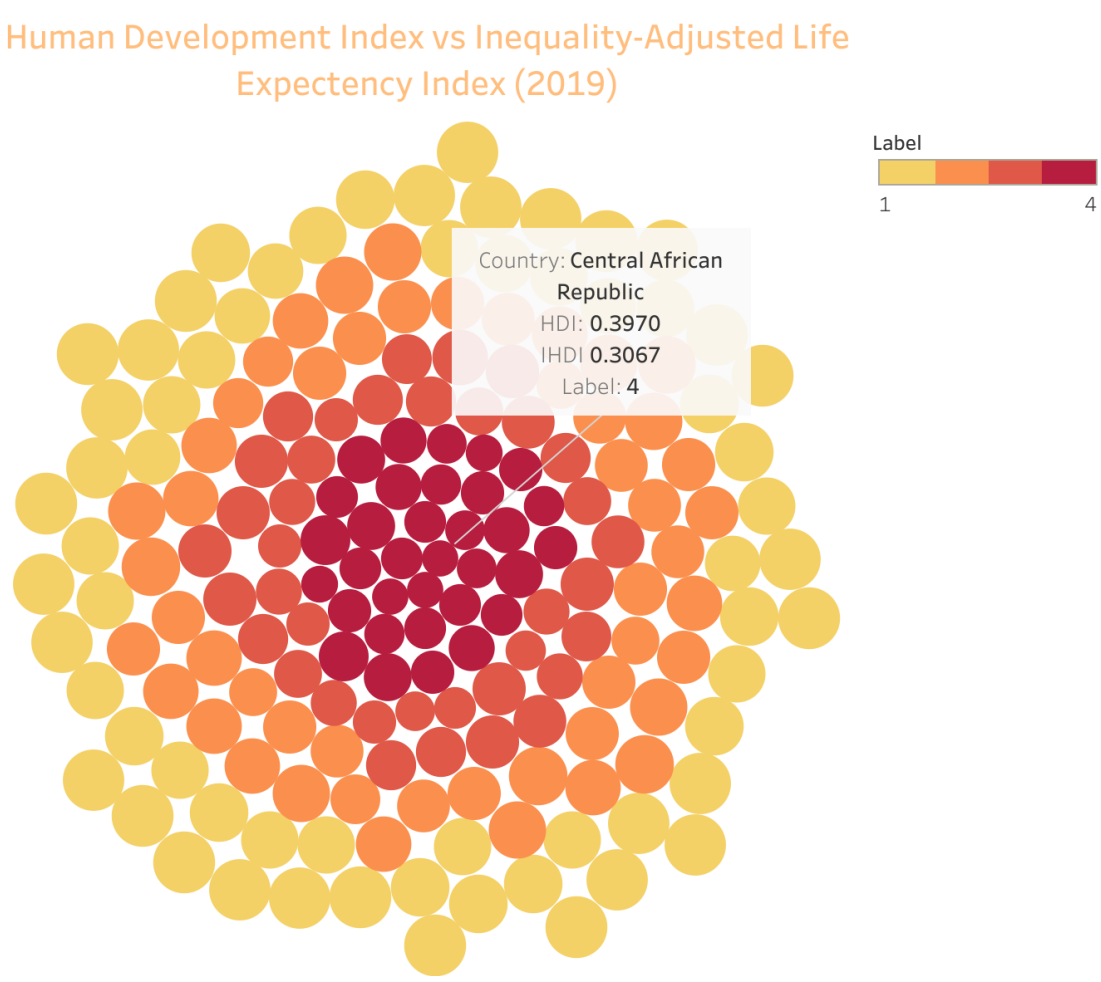

Fig 1.4. Human Development Index vs Inequality-Adjusted Life Expectancy Index (2019)

(b) Specific Exploration:

Variable of interest by country:

1. Share of seats in parliament (% held by women) 2019 (Fig 1.5):

From the text plot we can see how the parliament seats are divided in each country. We can observe a variation of trends depending on the countries.

Expected pattern: We expect countries that are label 1 or 2 i.e., developed countries to have an equal share of men to women seats in the parliament. The underdeveloped countries expect to usually have more men in the parliament compared to women as they are not well educated. This trend fairly follows for all the countries.

Unexpected pattern: Even though most countries followed the pattern we could see a few exceptions. Kuwait although being a developed country has a very poor percentage of seats for women. Similarly, Maldives too, being a label 2 country has very low percentage of seats for women. It is unexpected to see an underdeveloped country like Rwanda having one of the highest shares of parliament seats for women, being 55.66%.

Hypothesis: Kuwait and Maldives present a different scenario from their neighbors regarding share of parliament seats held by women. I hypothesize that Maldives being a small tourist country with limited resources and poor education, though being a label 2 country does not encourage women empowerment. Similarly, Kuwait being in an Islamic state has strict rules imposed on women which may not allow them to take part in social duties such as being a politician. Rwanda on the other side, being in Africa and quite remote might have helped the women to empower themselves and take ownership of the management of the country.

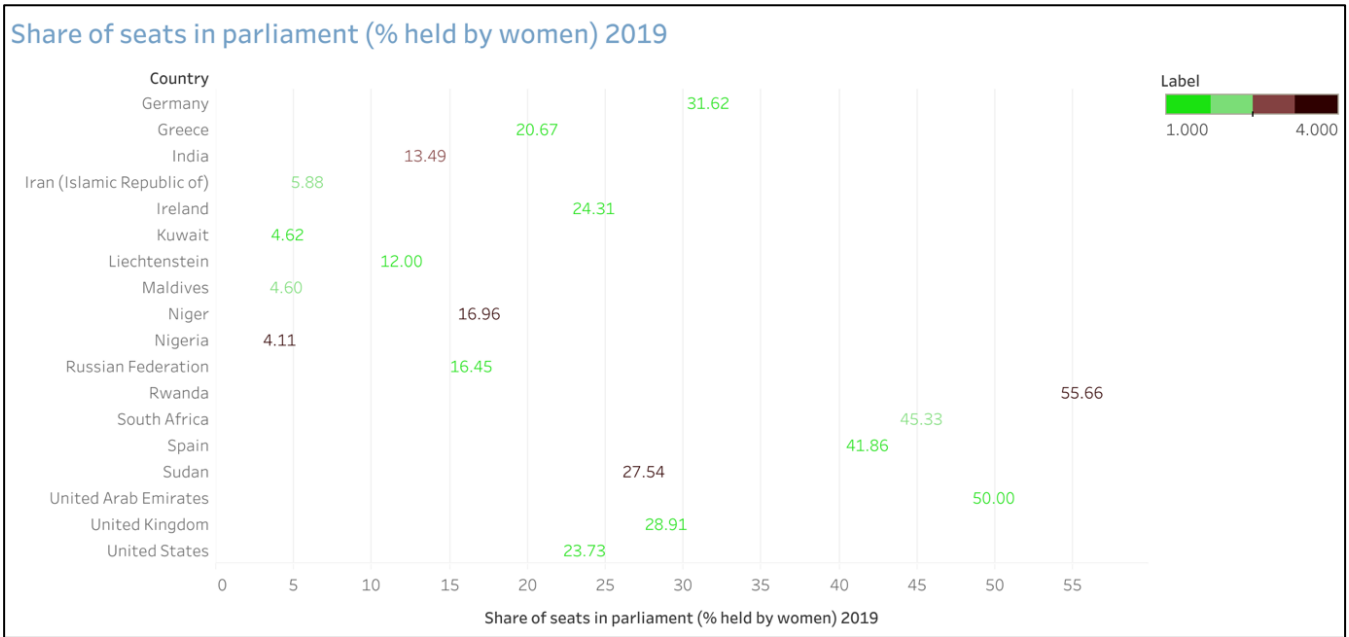

Fig 1.5. Share of seats in parliament (% held by women) 2019

2. Urban Population by Country (Fig 1.6):

From the plot we can see how urban population is in each country. We can observe a variation of trends depending on the countries.

Expected pattern: We expect countries that are label 1 or 2 i.e., developed countries to have a high urban population as they are developed and has good facilities. The underdeveloped countries expect to usually have less urban population as the country is behind in technology and does not have good urban planning implemented. This trend fairly follows for all the countries.

Unexpected pattern: Even though most countries followed the pattern we could see a few exceptions. Liechtenstein although being a developed country has a very low urban population in the country. Similarly, Papua New Guinea too, being a label 3 country has very low urban population. It is unexpected to few countries like Kuwait and Singapore having 100% urban population.

Hypothesis: Kuwait and Singapore present a unique pattern from their neighbors regarding urban population. I hypothesize that Singapore being a small highly developed country, has very limited space for expansion. Hence there are no lands for rural areas and only urban population is there. Similarly, Kuwait being a small Islamic state has no chance to have rural areas. Liechtenstein on the other hand, being in the remote part of Europe has good lands to develop crops and cattle and hence there is more rural population compared to the urban population. Also, Liechtenstein has no need to have major urban cities maybe cause they have more agricultural land compared to the service industry.

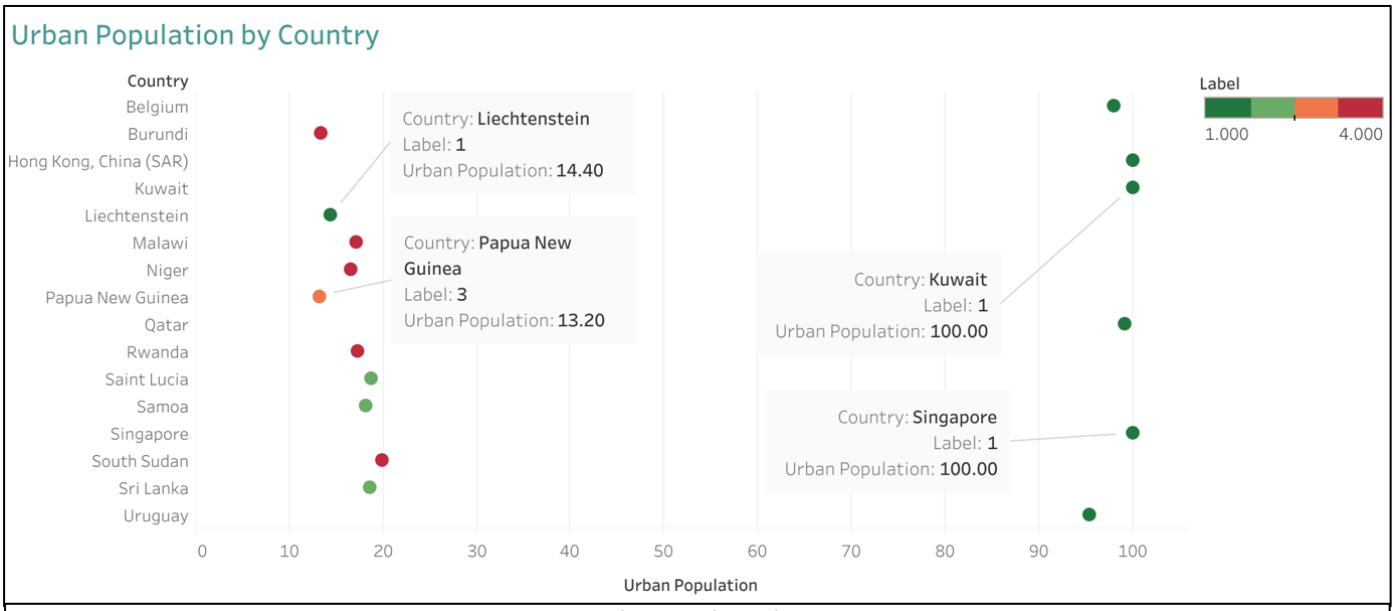

Fig 1.6. Urban Population by Country

Correlations between different indicators in HDR:

1. Mean Schooling vs Expected Schooling (Fig 1.7):

From the scatter plot we can notice a nonlinear relation between Mean years of schooling and Expected years of schooling. The variables are clearly correlated as they have almost constant variance and increasing mean.

Expected pattern: As they are correlated the countries must have same expected and mean years of schooling depending on their labels i.e., developed countries must be higher than the underdeveloped countries. Most of the countries follow this pattern which is not surprising.

Unexpected pattern: Even though most countries followed the pattern we could see a few exceptions. Maldives being a label 2 country (developed) has a lower expected to mean years schooling in 2019. Similarly, Kyrgyzstan which is a label 3 country is among the developed countries having high expected and mean years of schooling in 2019.

Hypothesis: In general, there is an equal distribution of the countries in relation to expected and mean years of schooling. Maldives and Kyrgyzstan present a different panorama from their neighbors regarding mean and expected years of schooling. I hypothesize that being a small tourist country with limited resources and job opportunities, Maldives though being a label 2 country does not have good education available. Similarly, Kyrgyzstan being in an isolated region and not much tourism might have increased jobs available and hence the better education.

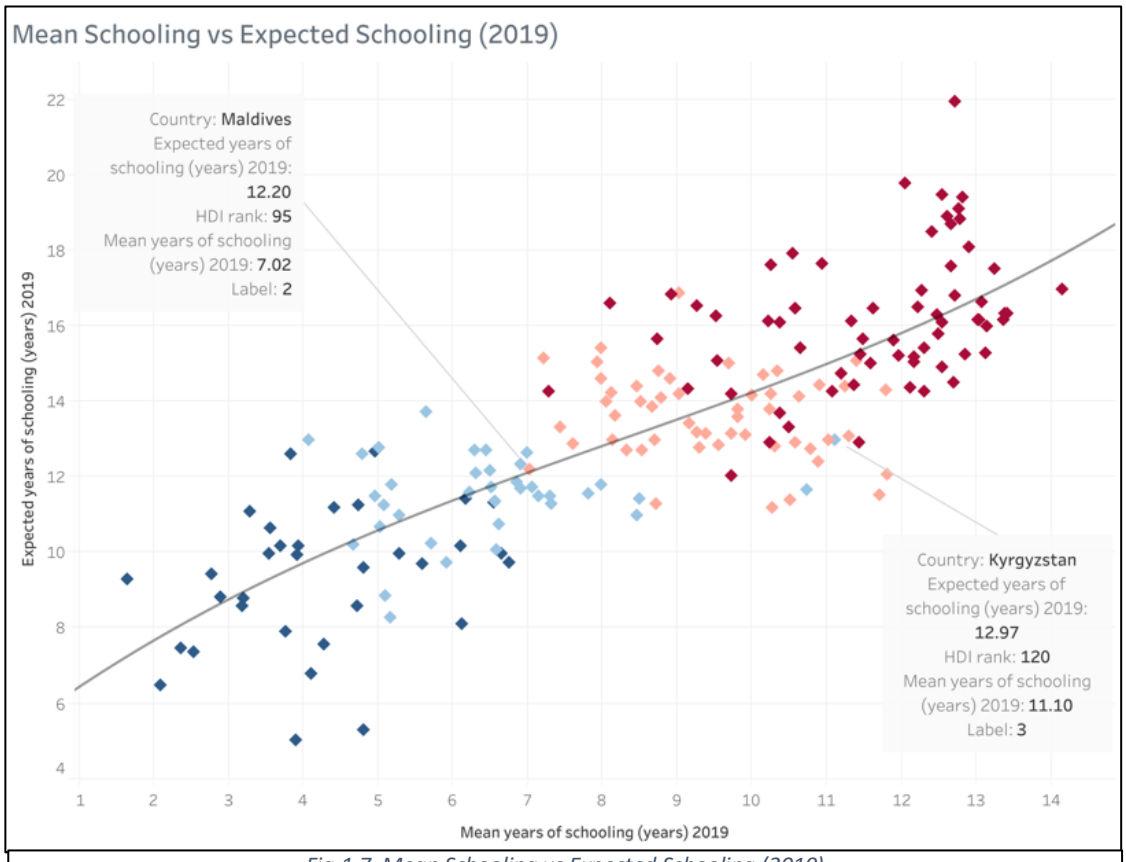

Fig 1.7. Mean Schooling vs Expected Schooling (2019)

2. GDP vs Exports and Imports (% of GDP) (Fig 1.8):

From the plot we can notice a nonlinear relation GDP and Exports and Imports (% of GDP). The variables are clearly correlated as they are uniformly grouped together.

Expected pattern: The countries with similar labels are all clustered together, black being label 4 and pink being developed countries. The pattern is uniformly distributed, underdeveloped countries all clumped together having similar GDP and % of Exports and Imports. We can understand that low GDP countries generally do not have a high Exports and Imports % of GDP, meaning there is not much export and import compared to GDP.

Unexpected pattern: Some of unexpected patterns found in this were the countries of Luxembourg, Qatar, and Djibouti. Luxembourg having the highest GDP in 2019, has also the highest Exports and Imports (% of GDP). Similarly, Qatar even though it has high GDP its Exports and Imports is very less, compared to other developed countries. Another expected pattern found was the country of Djibouti, located in Africa, having a very low GDP but very high Exports and Imports percentage.

Hypothesis: I hypothesize that Luxembourg being a high GDP country and a comparatively small area cold country, it has limited resources to make dailies necessaries. Hence, the high percentage of GDP spent on Exports and Imports to maintain staple elements. Qatar, being one of the highest GDP countries in 2019, has a very low percentage of Exports and Imports. This is probably the country can produce enough products within the country and Exports and Imports is not much encouraged or needed. When we look at Djibouti, we observe that though it has a low GDP it has a very high percentage of GDP for Exports and Imports. This is probably as it is a African country it has poor opportunities for high producing paychecks, in turn lower economic and GDP. Being a African country it is difficult to satisfy the general needs of the people due to poverty, low education and hunger. Hence the high percentage for Exports and Imports suggesting they may be exporting natural resources like timber and valuable metals in return for daily basic necessaries.

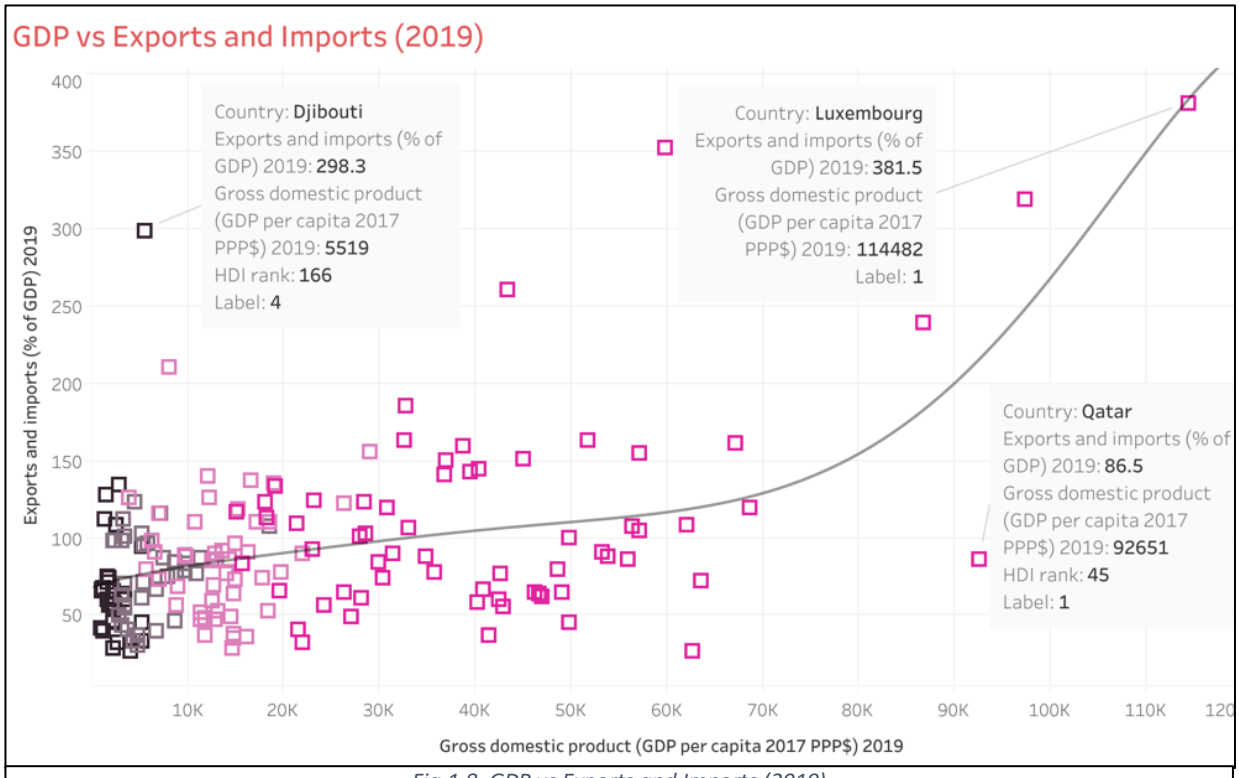

Fig 1.8. GDP vs Exports and Imports (2019)

Un-correlated indicators in HDR:

1. Expected Years of Schooling vs GDP (Fig 1.9):

From the plot we can notice an exponential relation between Expected Years of Schooling and GDP. The variables are clearly uncorrelated as they are pretty spread and random.

Expected pattern: The countries with similar labels are all clustered together. The pattern is uniformly distributed, underdeveloped countries all clumped together having similar GDP Expected years of schooling. We can expect that expected years of schooling must influence the GDP because in general more schooling yields better economy growth and more GDP in turn.

Unexpected pattern: Some of unexpected patterns found in this were the countries of Luxembourg, Australia, and Equatorial Guinea. Luxembourg having the highest GDP in 2019, has an average or median expected years of schooling. Similarly, Australia even though it has very high expected years of schooling, its GDP is comparatively less. Another unexpected pattern found was Equatorial Guinea, located in Africa, having a decent GDP even though it has a very low expected years of schooling.

Hypothesis: I hypothesize that Luxembourg being one of the developed country has focused on technology and IT sectors to have such a huge GDP. The expected years of schooling in Luxembourg is on median with other countries, this may be due to lack of education in theory and more chances of gaining an industrial experience. Likewise, Australia, has a high expected years of schooling has a low GDP compared to its neighbors. This may be because of less industrial job opportunities available in that location. When we look at Equatorial Guinea, we observe that though it has a decent GDP despite its low expected years of schooling. This is probably because it is an African country which has rich mineral resources. More efforts go into educating on mining and looking for these resources than educating the general like mathematics and science.

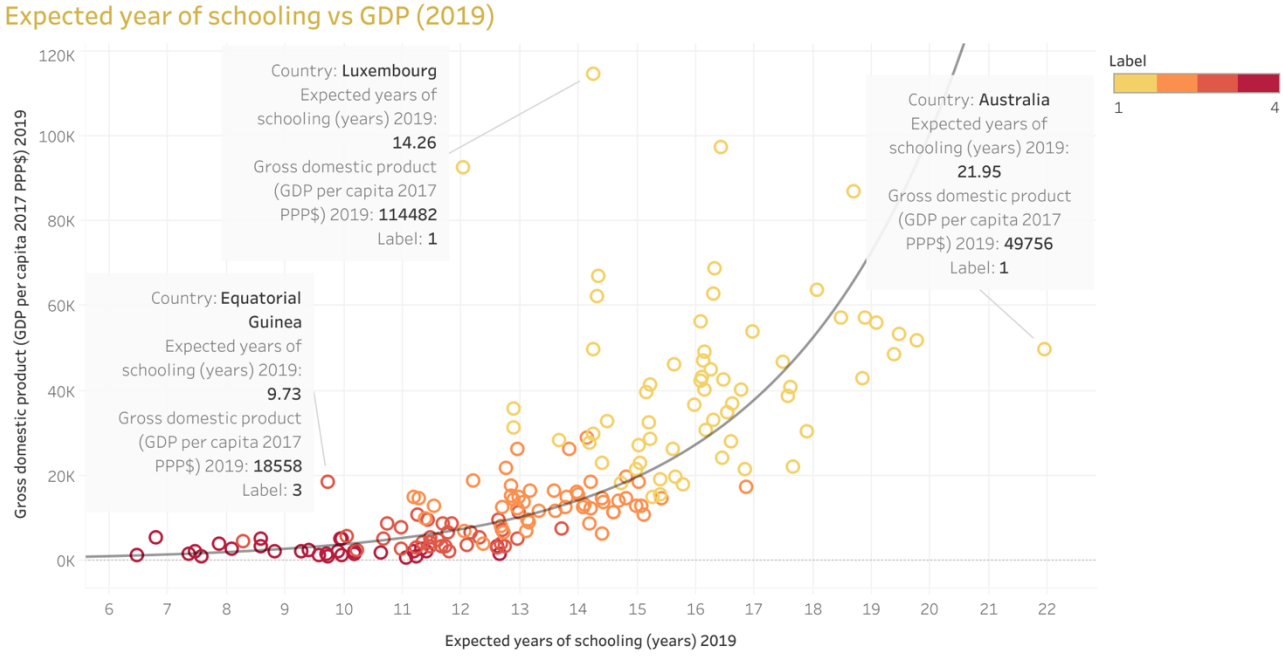

Fig 1.9. Expected Year of Schooling vs GDP (2019)

2. Urban Population vs Gender Development Index (Fig 1.10):

From the plot we can notice a flat relation between Urban Population and Gender Development Index. The variables are clearly uncorrelated as they are pretty spread and random.

Expected pattern: The countries with similar labels are all clustered together. The pattern is uniformly distributed, underdeveloped countries all clumped together having similar Urban Population and Gender Development Index. We can expect that Urban Population may influence the Gender Development Index because in general more Urban Population means better planning in turn better Human Development Index.

Unexpected pattern: Some of unexpected patterns found in this were the countries of Qatar, Yemen, Barbados and El Salvador. Qatar has a high Urban Population and Gender Development Index. Similarly, Barbados and El Salvador even though it has very high Gender Development Index, its Urban Population is comparatively less. Another unexpected pattern found was Yemen, has very low Gender Development Index and low Urban Population.

Hypothesis: I hypothesize that Qatar being one of the developed country has less space to expand and hence the high Urban Population. Also being one of the developed countries its Gender Development Index is very high. Countries like El Salvador and Barbados even though they have high Gender Development Index have low Urban Population. I can hypothesize that Barbados and El Salvador have good government hence the good gender development index. Yemen is low in both the Gender Development and Urban population; this maybe cause of poor government management and poor resources available in the country.

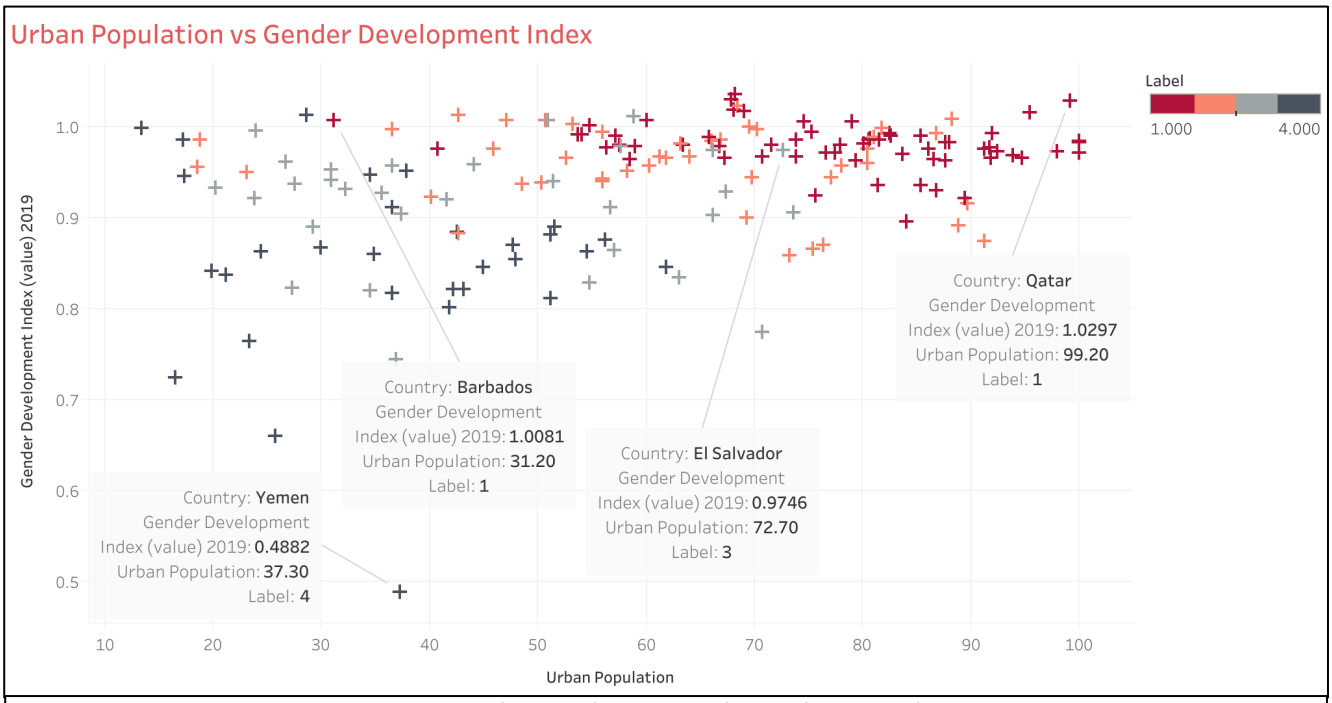

Fig1.10. Urban Population vs Gender Development Index

Usefulness or not Usefulness of Available Visualizations

1. Scatter Plot:

Helps show whether 2 variables are related or not and how much one variable affects another. Also is helpful to predict the behaviour of one variable (dependent) based on the measure of the other variable (independent). Downsides being, flat best-fit line gives inconclusive results. Interpretation can be subjective. Correlation does not mean and not show causation. Data on both axes have to be continuous data You cannot use Scatter diagrams to show the relation of more than two variables.

2. Tree Plot:

Compared to other algorithms decision trees requires less effort for data preparation during pre-processing. A decision tree does not require normalization, scaling of data. Missing values in the data also do not affect the process of building a decision tree. Downsides are, a small change in the data can cause a large change in the structure of the decision tree causing instability. The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.

3. Box and Whisker Plot:

It handles huge amounts of data easy. Exact values are not retained. Provides an clear summary and outliers present. Disadvantages being, it hides the multimodality and other features of distributions. Can be confusing for some audience. The Mean often is difficult to interpret.

4. Bar Plot:

Show each data category in a frequency distribution. Make trends easier to highlight than tables do it helps in studying patterns over long period of time it is used to compare data sets. Disadvantages are it often require additional explanation fail to expose key assumptions, causes, impacts and patterns can be easily manipulated to give false impressions do not show inter relationship between activities.

4. Histogram:

The frequency of score of datasets that have been categorized into classes or interval scale is plotted using a histogram. It's mostly used to figure out how data is distributed. Disadvantages are it does not allow you to read exact values because data is grouped into categories. It uses only with continuous data. In Histogram, it is not easy to compare two data sets.

5. Line Graph:

Area charts and line graphs are extremely similar. Line graphs are used to depict how data changes over time, whether it's a short or lengthy period. It comes in handy when comparing large amounts of data at once. Downsides are plotting too many lines over the graph makes it cluttered and confusing to read. A wide range of data is challenging to plot over a line graph. They are only ideal for representing data that have numerical values

5. Area Charts:

The time-series interaction is primarily displayed using an area chart. It depicts how data grows, shrinks, or changes over time. It's especially beneficial for keeping track of two or more closely connected groupings that make up a single category. Downsides being, its only useful for comparing trends, not exact values. Reading exact data values from an area chart is not something easy. Works efficiently only for smaller number of groups. Rendering multiple categories on an area chart will make it difficult for a user to understand the data.

Usefulness of Visualizations used in the previous task

1. GDP vs Exports and Imports (% of GDP) (Fig 1.11):

Insight: The plot (Fig 1.8) is useful in identifying the underlying trend between the variables easily. We can make use of the color and size to describe the data more effectively with plots.

New Insight: Fig 1.11 shows a box plot of GDP and Exports and Imports (% of GDP). We can observe that the outliers are easily identified using Box and Whisker plot. We can note the average and the quartiles very easily. We notice Singapore and Luxembourg both have high Exports and Imports and Gross Domestic Product in 2019.

Takeaway: We can say box and whisker plot is better in identifying the measure values (mean, median, etc) easier than the scatterplot. The scatter plot is useful to identify the data based on a filter (by country, by rank, by population etc)

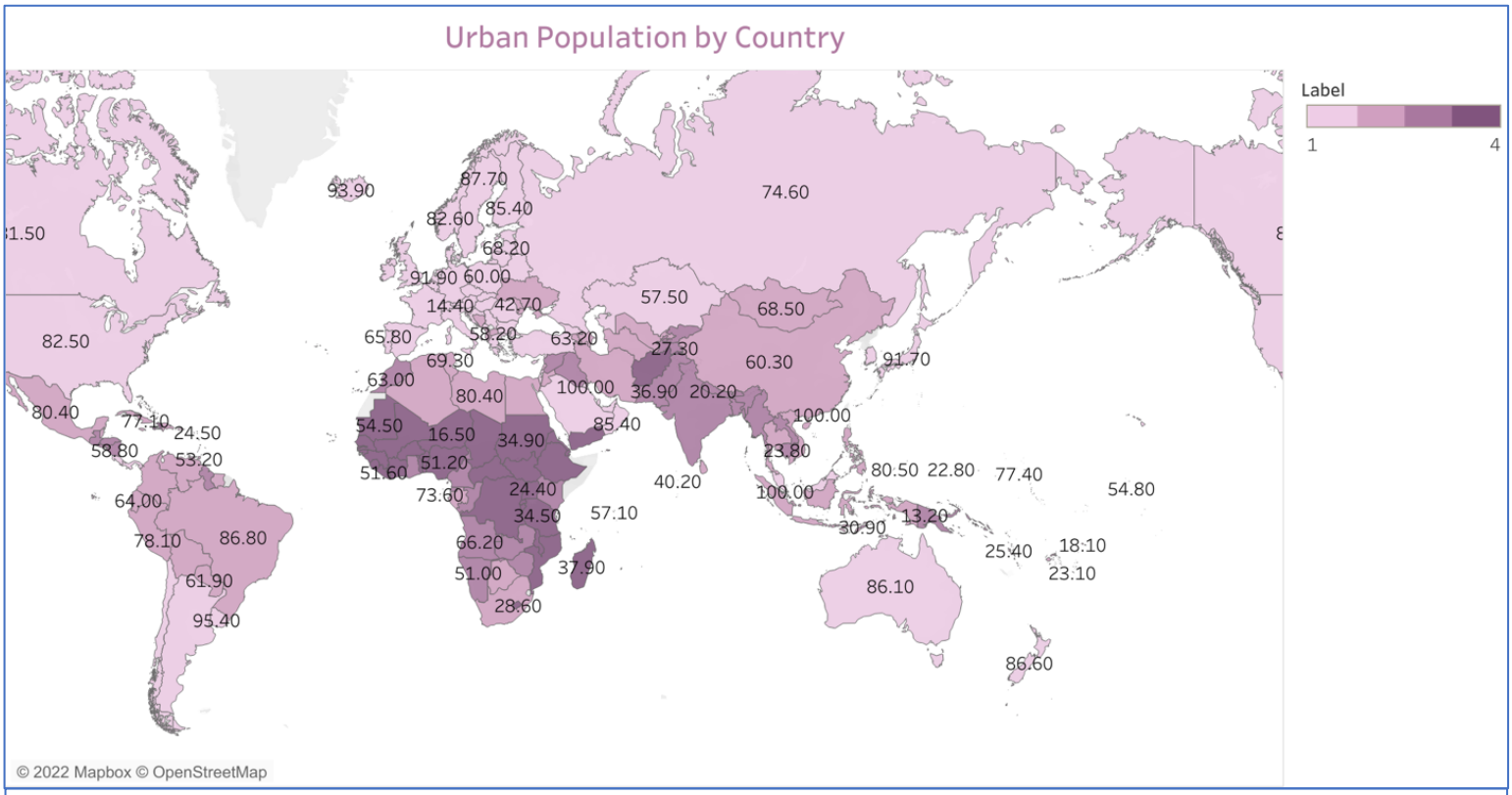

Fig 1.12. Urban Population by Country

2.Urban Population by Country (Fig 1.12):

Insight: The plot (Fig 1.6) is useful in identifying the outliers easily based on the countries. The values are displayed on the side of each point hence can be interpreted easy.

New Insight: Fig 1.12 shows a map plot of Urban based on countries. We can observe the whole global map in one view and the Urban Population values associated with them. We can see the values by hovering over a country which makes it simple and easy. We notice that countries such as Kuwait, Singapore and USA have high Urban Population.

Takeaway: We can compare and conclude saying map plot is useful when there is single attribute or single variable to be evaluated based on the countries (like urban population, total population etc). When there are multiple variables of interest the map plot can become cluttered and it becomes harder to interpret them.

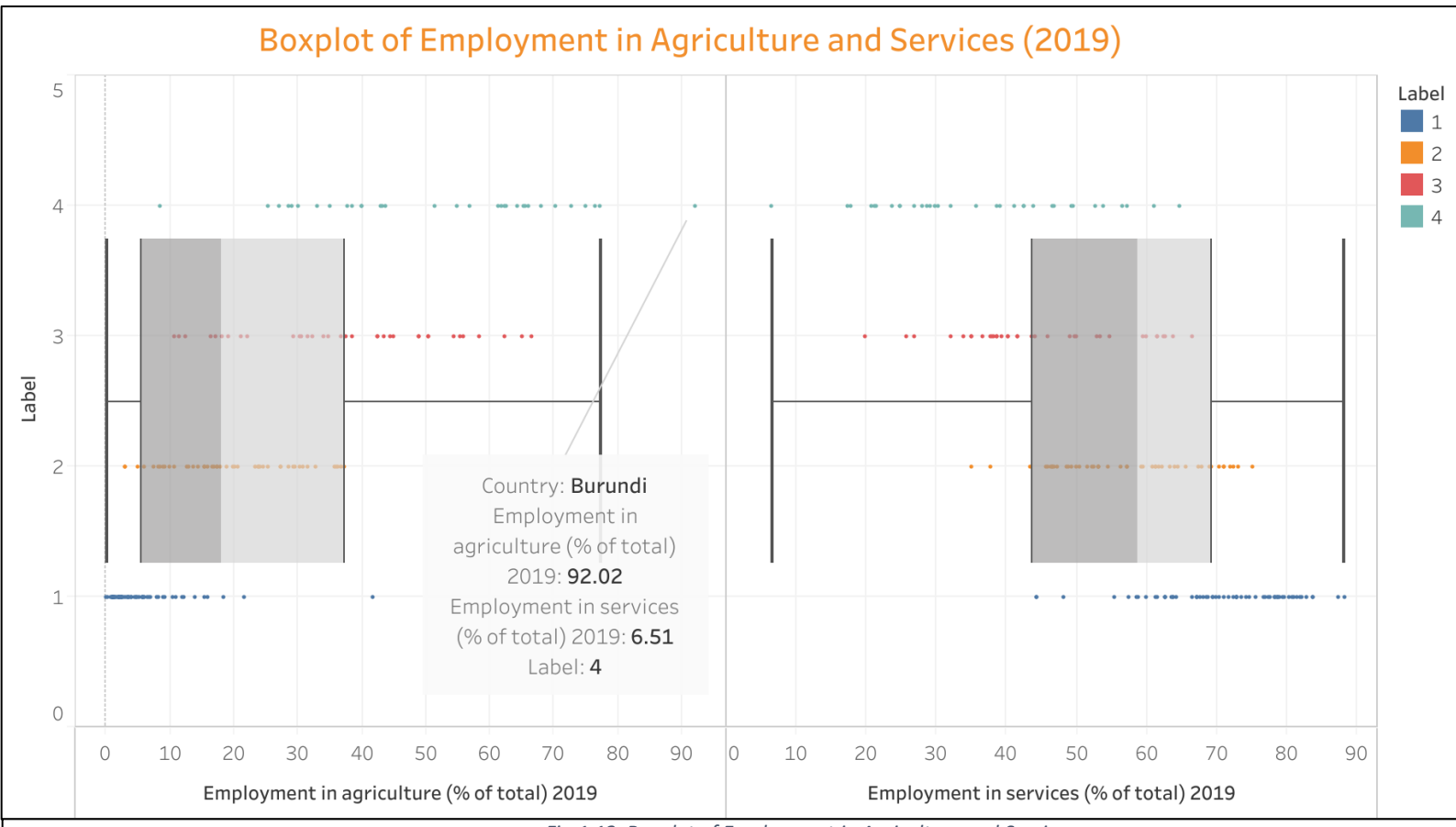

Fig 1.12. Boxplot of Employment in Agriculture and Services

Task 2

For this task, we must make use of VisPipeline and perform visual analysis on the following datasets: Coral, CBR and HDR. We can make use of multiple projection techniques such as LSP, T-SNE, NJ, Isomaps, Sammon, Rapid Sammon, MDS to name a few. I also make use of the silhouette coefficients to compare the efficiency of the projections.

The silhouette score falls within the range [-1, 1]. The silhouette score of 1 means that the clusters are very dense and nicely separated. The score of 0 means that clusters are overlapping. The score of less than 0 means that data belonging to clusters may be wrong/incorrect [3].

(a) Exploration of an image collection:

The data set Corel in the data archive is a vector space composed of image features, for a collection of photographs and drawings. There are 10 labels and 1000 items in the data set. I apply different projections with different parameters to that data set, making observation in regards to segregation of labels and the differences between projections. Also have given examples of images that seem to be difficult to discriminate from other labels and the reason behind it.

Fig 2.1 depicts the pipeline I have used. It has a Detailed Points Matrix Reader which takes the 'coral.data' file as the input which is in a matrix format. It is connected to multiple projections such as Neighbour-Joining algorithm(NJ), Least Square Projection(LSP), Classical Multidimensional Scaling(MDS) and t-Distributed Stochastic Neighbour Embedding(TSNE). NJ is connected to a Tree model and then a Radical Layout and then a Tree View Frame to view the output of projection. Radical Layout also has a Tree Model Silhouette Coefficient to capture the projection efficiency. TSNE is connected to a topic Projection Model and a Topic Projection View Frame to view the projection. Similarly, MDS and LSP is connected to a Labelled Projection Model and a Labelled Projection View Frame and also has a Labelled Projection Model Silhouette Coefficient.

Figures 2.2, 2.3, 2.4, 2.5 show the projections I obtained running the pipeline.

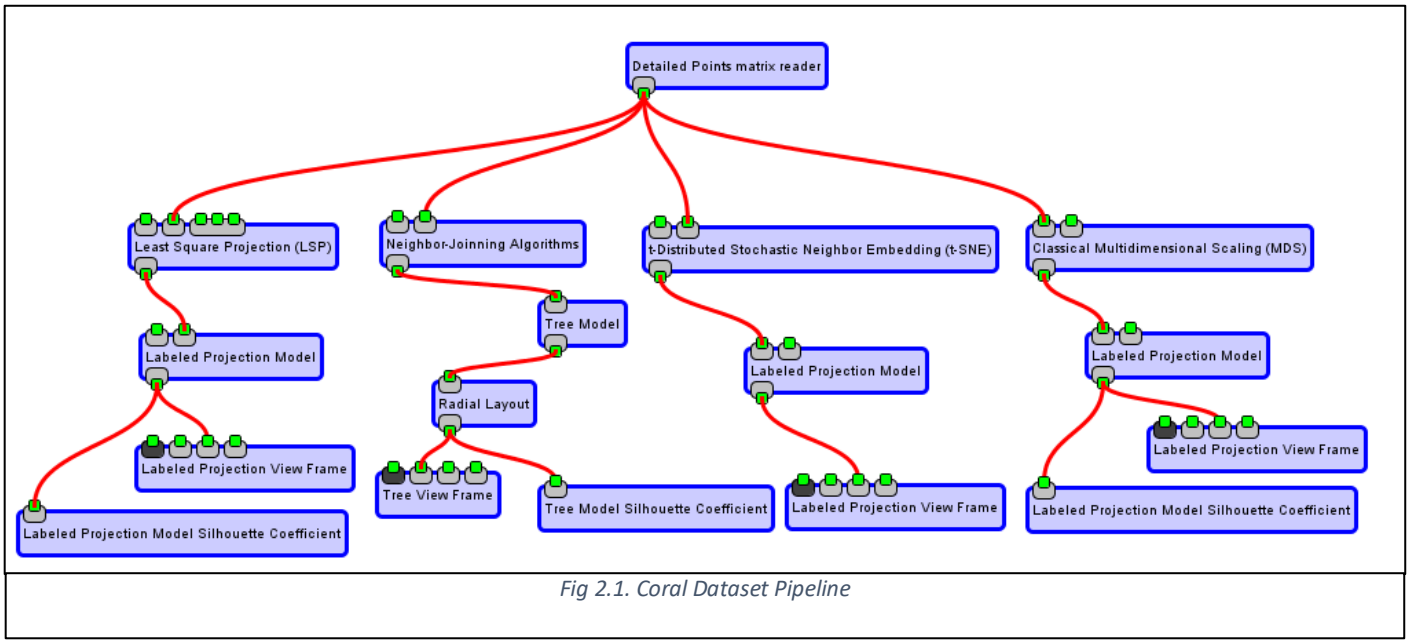

Fig 2.1. Coral Dataset Pipeline

LSP

In LSP projection a linear system is formulated and its solution refers to the projection of the remaining points in the convex hull of its k nearest neighbours.[4]
Fig 2.2 is the LSP Projection, we can see that it has classified decently when the coral data is input.

Parameters:

Number of iterations: 50
Fraction of delta: 8.0
Number of control points: 50
Number of Neighbors: 10
Dissimilarity: Euclidean
LSP silhouette coefficient obtained (Euclidean): 0.287
Hence we can conclude the projection though not perfect is good enough to classify the coral dataset, though we find a few misclassified objects such as a man being grouped together with food items.

t-SNE

t-SNE is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map [5]. It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability.

Parameters:

Initial dimension: 30
Target dimensions: 2
Perplexity: 30
Maximum iterations: 1000
Dissimilarity: Euclidean
T-SNE (fig 2.3) does a very good job in classifying the data into different clusters so we can observe it is the best among all projections.

MDS

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset [6]. MDS is used to translate "information about the pairwise 'distances' among a set of objects or individuals" into a configuration of points mapped into an abstract Cartesian space. MDS (fig 2.4) also does a decent job in classifying though there are overlaps and the data interpretability is difficult.

Parameters:

Dissimilarity: Euclidean
MDS silhouette coefficient obtained (Euclidean): 0.053
Hence it does a poor job compared to LSP and TSNE.

NJ

Neighbour Joining is a bottom-up (agglomerative) clustering method for the creation of phylogenetic trees [7]. NJ (fig 2.4) has many branches to classify the data into multiple clusters of similar data.

Parameters:

Dissimilarity: Euclidean.
NJ Algorithm: Rapid Neighbor-Joining
NJ silhouette coefficient obtained: 0.271
We can conclude that it does a better job than MDS but not as good as LSP and TSNE.

Conclusion

We can observe that t-SNE has the best projection and classification among all the projections used. LSP, NJ and MDS have positive silhouette coefficient which states that the clusters are nicely separated and dense though not perfect.

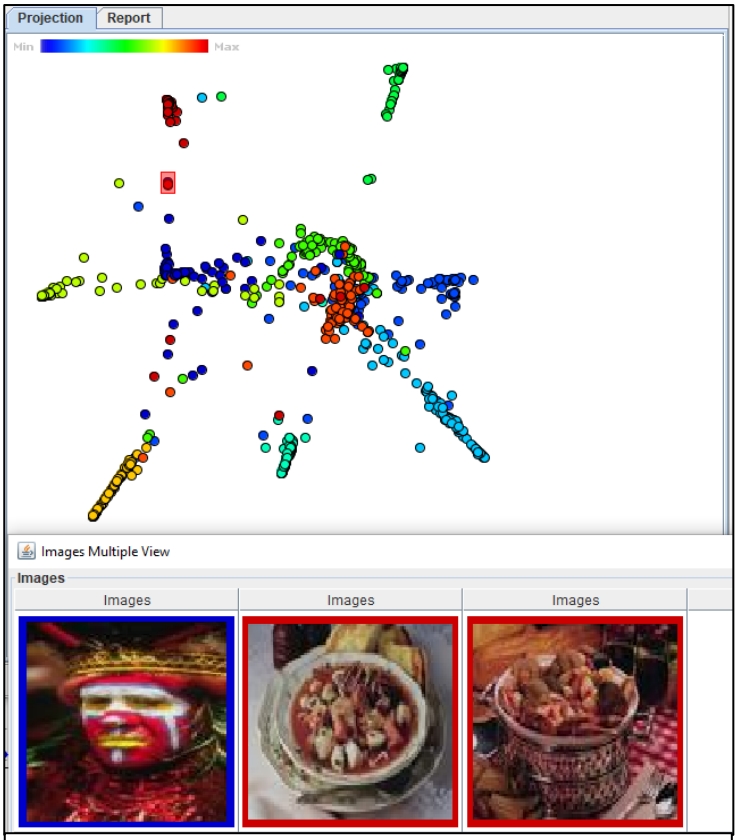

Fig 2.2. LSP Projection

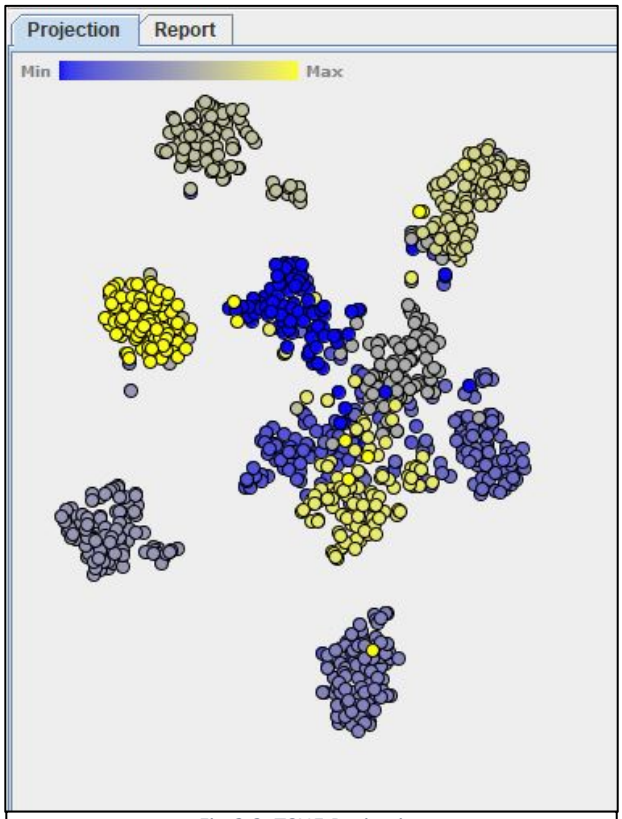

Fig 2.3. TSNE Projection

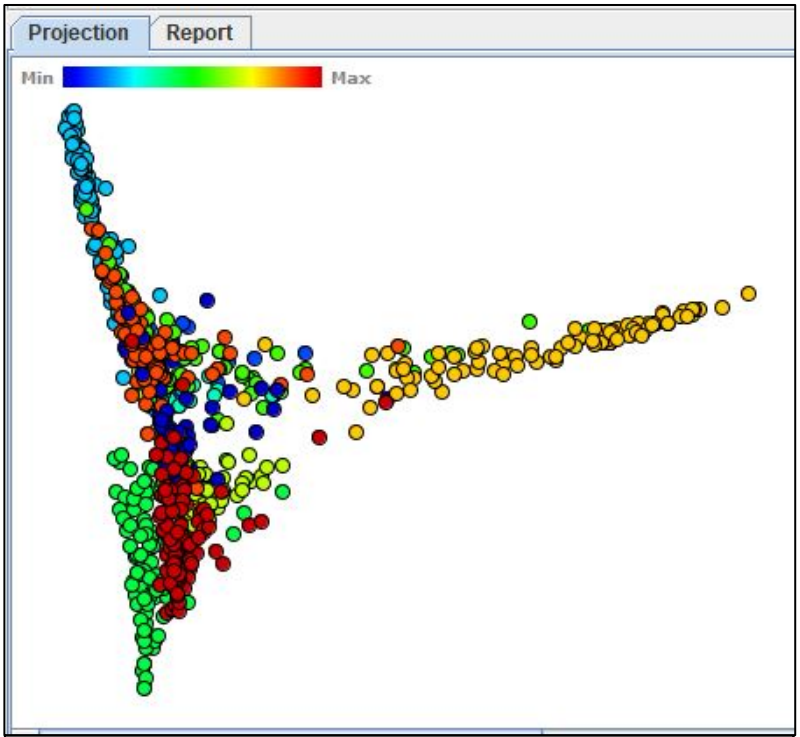

Fig 2.4. MDS Projection

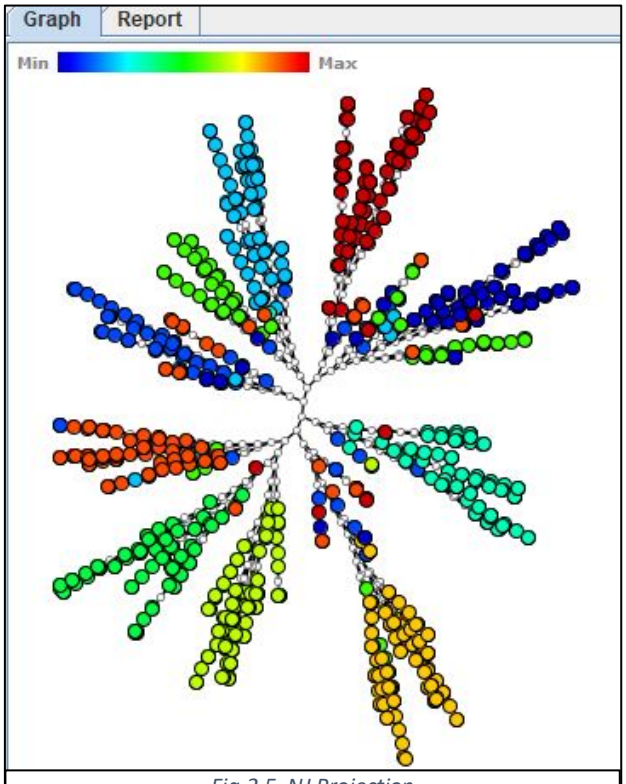

Fig 2.5. NJ Projection

(b) Exploration of a document collection:

The data named CBR is a collection of article abstracts from certain areas of knowledge. It comprises a vector space model extracted from 600+ documents. It consists of multiple articles. Fig 2.6 shows the pipeline connections I used to set up for the document processing and analysis. First we have to select the file 'cbr-ilp-ir-son-int-int.zip' in the Zip pre-processor and connect to Points Matrix Writer and execute to get the 'CBR_int.data' file. We input this file in Detailed Points Matrix Reader. The detailed points matrix reader is connected to NJ, ISOMAP, LSP, TSNE. NJ is connected to tree model, then radial layout then Tree View frame to view the projection, also has a tree model silhouette coefficient. TSNE is connected to a topic Projection Model and a Topic Projection View Frame to view the projection. Similarly, ISOMAP and LSP is connected to a Labelled Projection Model and a Labelled Projection View Frame and also has a Labelled Projection Model Silhouette Coefficient.

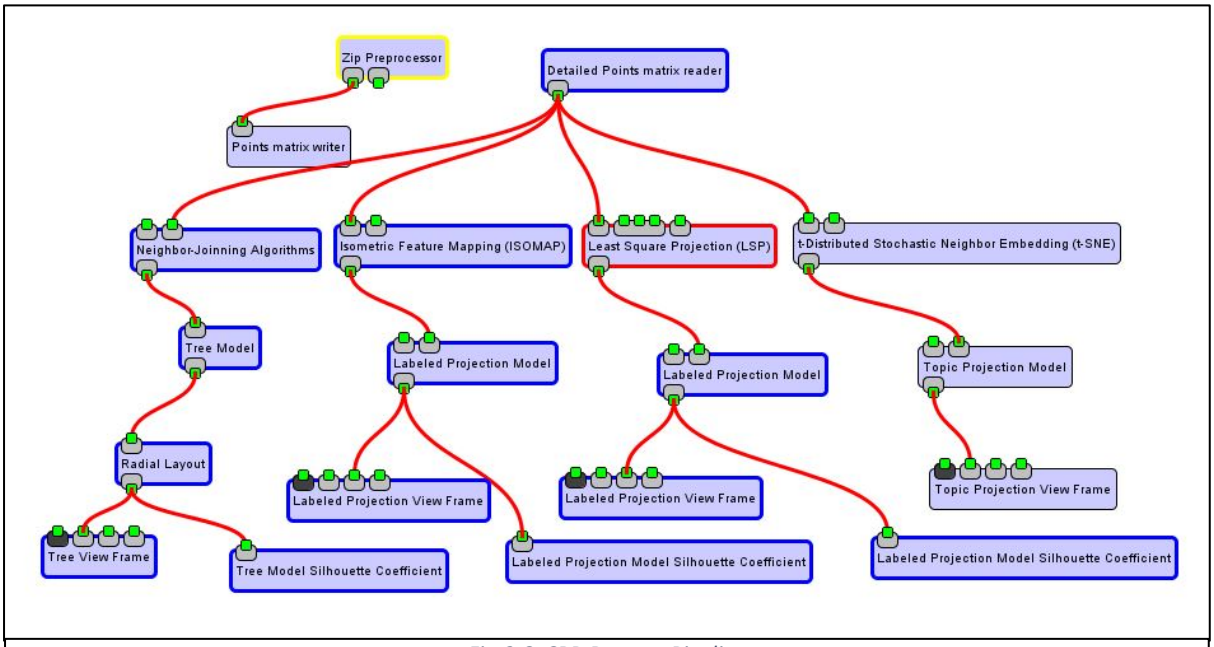

Fig 2.6. CBR Dataset Pipeline

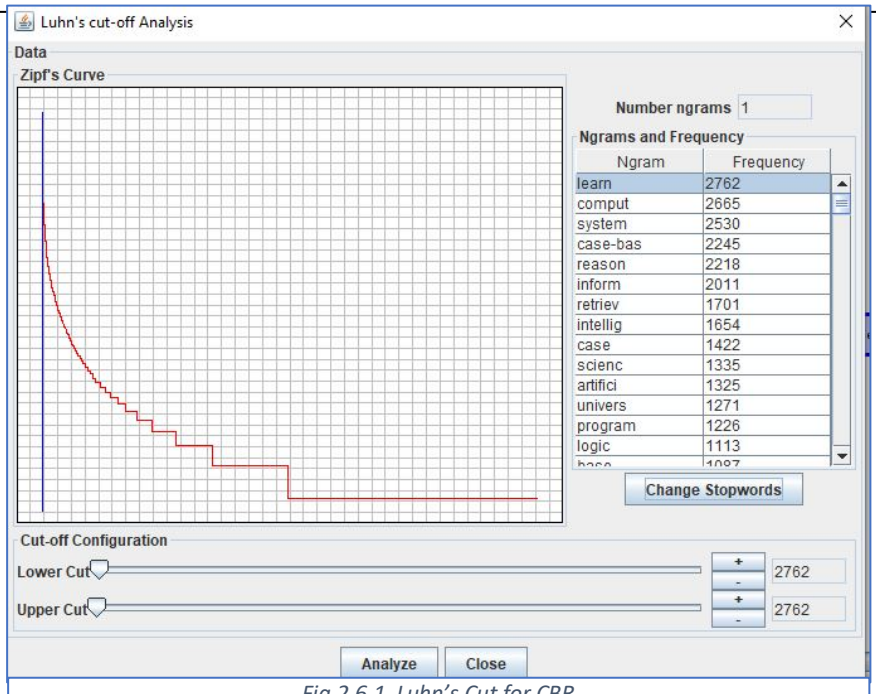

Fig 2.6.1. Luhn's Cut for CBR

Fig 2.6.1 shows the frequency of the words repeated in the document (Luhn's Cut). I have removed the stop words using 'stopwords_article.stw' file. Figures 2.7, 2.8, 2.9, 2.10 show the projections I obtained running the pipeline.

LSP

Fig 2.7.1 shows the LSP Projection, we can see that the classification is not that good as there are overlaps and spread-out points.

Parameters:

Number of iterations: 60
Fraction of delta: 8.0
Number of control points: 10
Number of Neighbors: 10
Dissimilarity: Cosine-based dissimilarity
LSP silhouette coefficient obtained (Cosine-based dissimilarity): -0.155
LSP silhouette coefficient obtained after changing(Cosine-based dissimilarity): 0.297
The projection is not that great. We can get a better projection by tuning the parameters a bit, we can get a better view by increasing the control points from 10 to 20 (fig 2.7.2).

t-SNE

t-SNE models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability.

Parameters:

Initial dimension: 30
Target dimensions: 2
Perplexity: 30
Maximum iterations: 1000
Dissimilarity: Cosine-based dissimilarity
t-SNE silhouette coefficient obtained (Cosine-based dissimilarity): 0.199
t-SNE (fig 2.8) does a very good job in classifying the data into different clusters so we can conclude it is the best among all projections. There are some files misclassified together.

NJ

Neighbour Joining is a bottom-up (agglomerative) clustering method for the creation of phylogenetic trees [7]. NJ (fig 2.9) has many branches to classify the data into multiple clusters of similar data.

Parameters:

Dissimilarity: Cosine-based dissimilarity.
NJ Algorithm: Rapid Neighbor-Joining
NJ silhouette coefficient obtained: 0.1295
We can conclude that it does a worse job in segregation than the others.

ISOMAP

Isomap is a nonlinear dimensionality reduction method (fig 2.10). It is one of several widely used low-dimensional embedding methods. Isomap is used for computing a quasi-isometric, low-dimensional embedding of a set of high-dimensional data points. The algorithm provides a simple method for estimating the intrinsic geometry of a data manifold based on a rough estimate of each data point's neighbours on the manifold [8]. Isomap is highly efficient and generally applicable to a broad range of data sources and dimensionalities.

Parameters:

Number of neighbors: 15

Dissimilarity: Cosine-based dissimilarity
Isomap silhouette coefficient obtained (Cosine-based dissimilarity): 0.2253 It has a better projection when compared to NJ and LSP as Isomaps specialize in classification and nonlinear dimensionality reduction.

Conclusion

We can observe that LSP has the best projection and segregation among all the projections used. The silhouette of LSP is 0.297 which states that the clusters are very well segregated and useful. NJ and ISOMAP both have positive silhouette coefficient which states that the clusters are nicely separated and dense though not perfect.

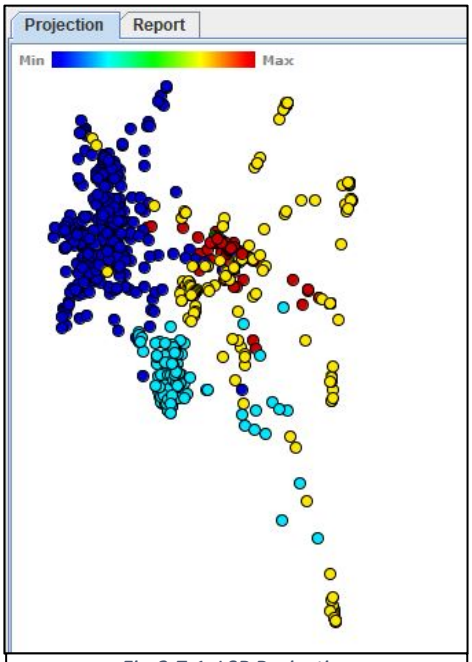

Fig 2.7.1. LSP Projection

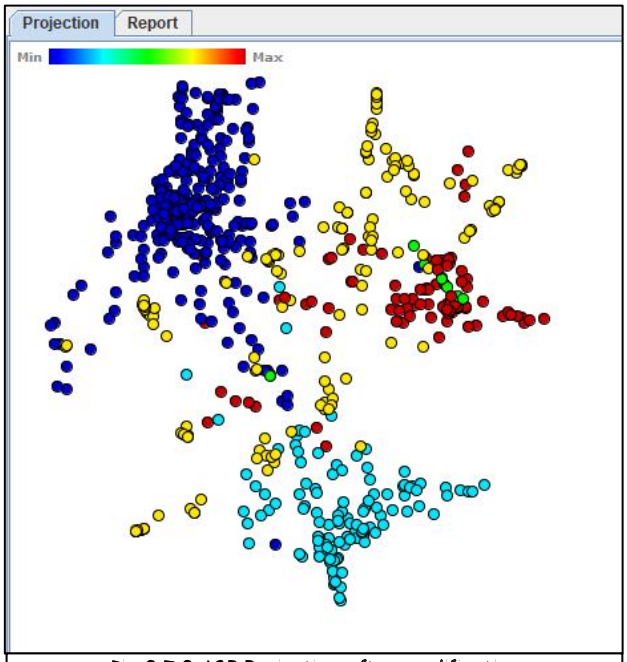

Fig 2.7.2. LSP Projection after modification

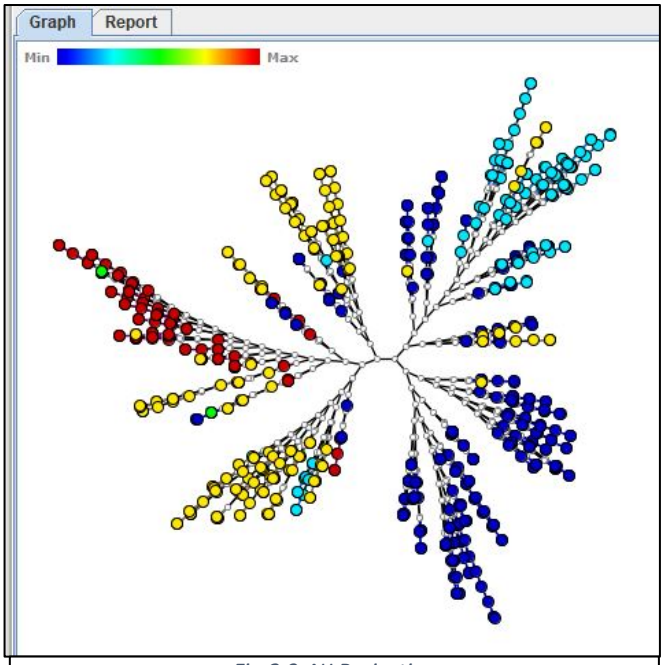

Fig 2.9. NJ Projection

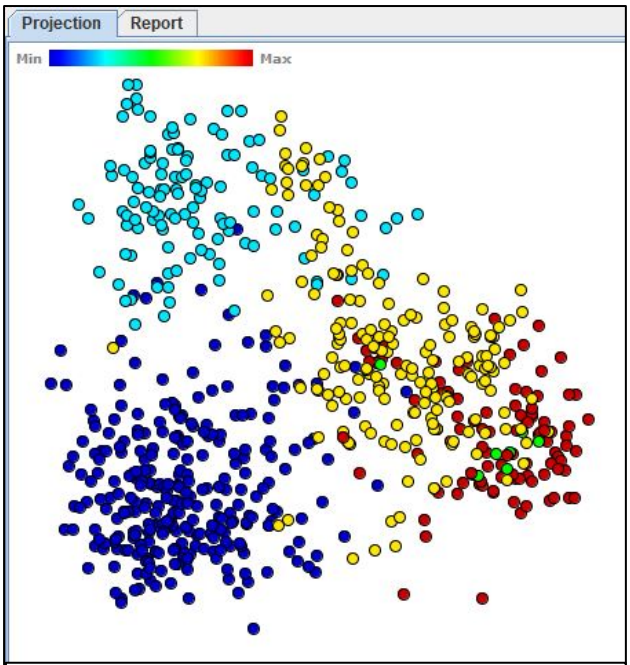

Fig 2.10. ISOMAP Projection

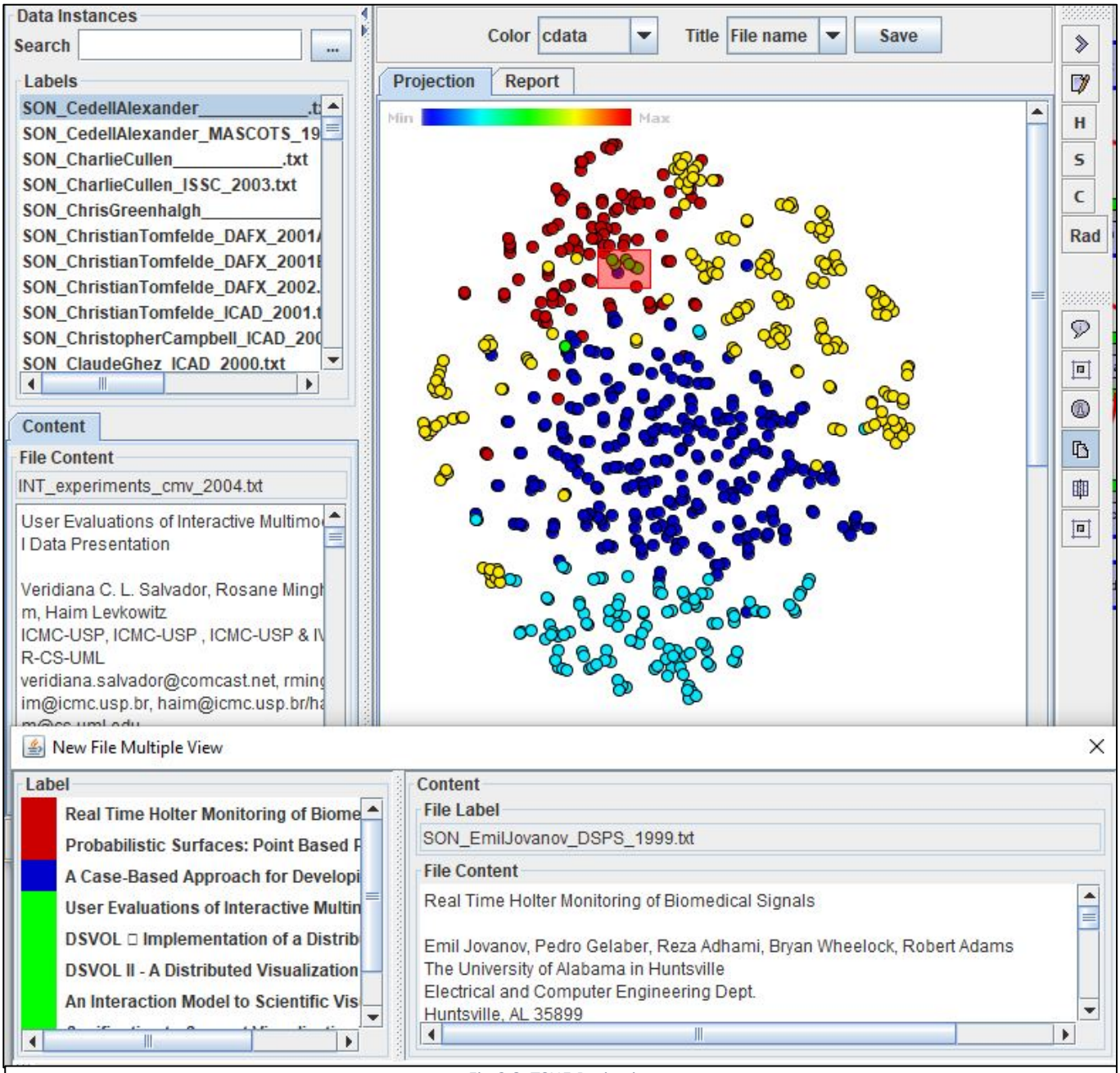

Fig 2.8. TSNE Projection

(c) Exploration of another datasets:

1. VSM data

Fig 2.11 shows the pipeline for VSM data file. We input the file in Detailed Points Matrix Reader. The detailed points matrix reader is connected to MDS, LSP and TSNE. TSNE is connected to a topic Projection Model and a Topic Projection View Frame to view the projection. Similarly, MDS and LSP is connected to a Labelled Projection Model and a Labelled Projection View Frame and also has a Labelled Projection Model Silhouette Coefficient.

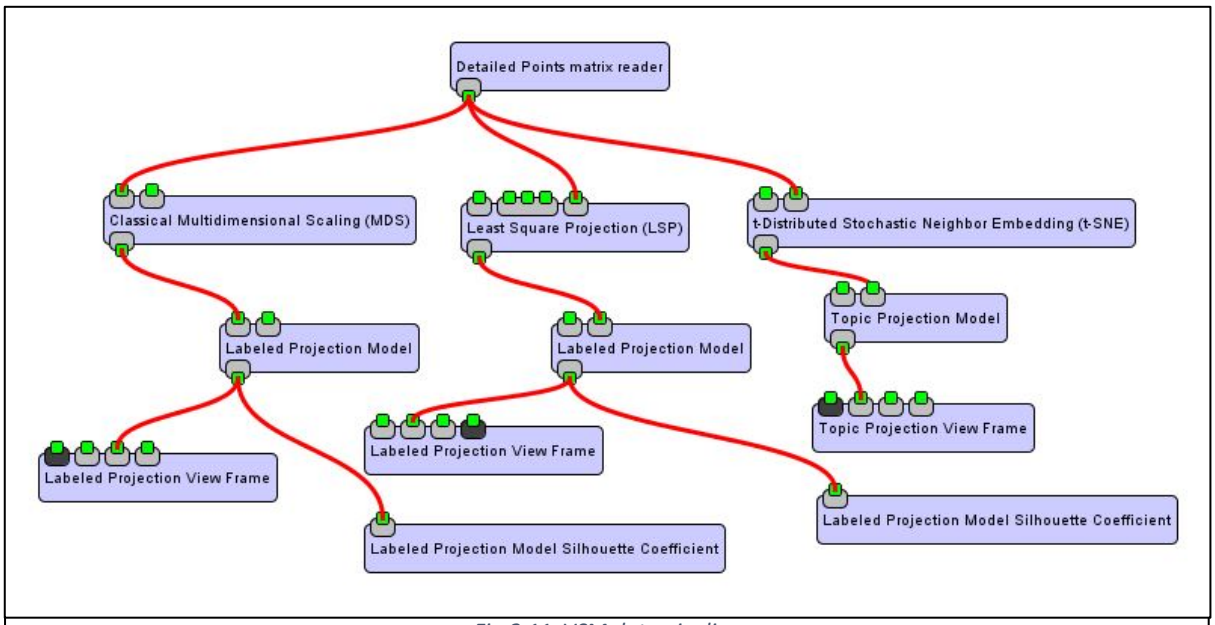

Fig 2.11. VSM data pipeline

Figures 2.12, 2.13, 2.14 show the projections I obtained running the pipeline.

LSP

Fig 2.12 shows the LSP Projection, we can see that the classification is not that good as there are overlaps and spread-out points.

Parameters:

Number of iterations: 50
Fraction of delta: 5.0
Number of control points: 15
Number of Neighbors: 10
Dissimilarity: Euclidean
LSP silhouette coefficient obtained: 0.0720
LSP has a positive silhouette coefficient, hence can conclude it is performing decent clustering,

t-SNE

t-SNE models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability.

Parameters:

Initial dimension: 30
Target dimensions: 2
Perplexity: 30
Maximum iterations: 1000
Dissimilarity: Cosine-based dissimilarity

t-SNE (fig 2.13) does a very good job in classifying the data into different clusters so we can conclude it is the best among all projections. There are some files misclassified together, but not that major of an issue.

MDS

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset. MDS is used to translate "information about the pairwise 'distances' among a set of objects or individuals" into a configuration of points mapped into an abstract Cartesian space. MDS (fig 2.14) also does a decent job in classifying though there are overlaps and the data interpretability is difficult.

Parameters:

Dissimilarity: Euclidean

MDS silhouette coefficient obtained (Euclidean): -0.0456

Hence it does a poor job compared to LSP and TSNE.

Conclusion:

PCA did the worst among the three. Dimensionality reduction sometimes causes a loss in the data which might lead to cases of being misclassified. LSP even though its decent compared to other projections is not better than t-SNE. t-SNE is also a method to reduce the dimension. One of the most major differences between PCA and t-SNE is it preserves only local similarities whereas PA preserves large pairwise distance maximize variance. It takes a set of points in high dimensional data and converts it into low dimensional data [9].

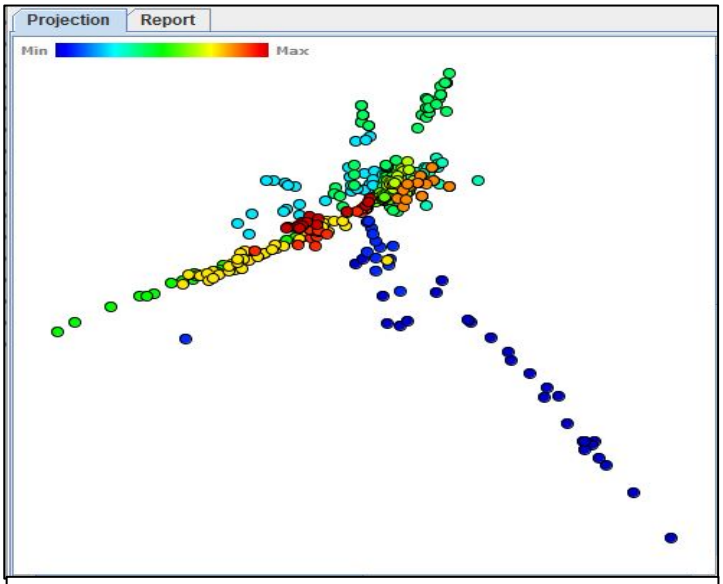

Fig 2.14. MDS Projection

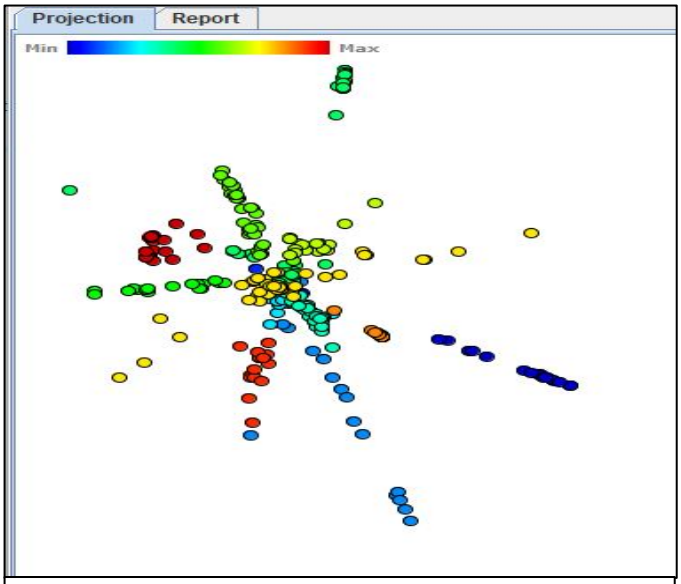

Fig 2.12. LSP Projection

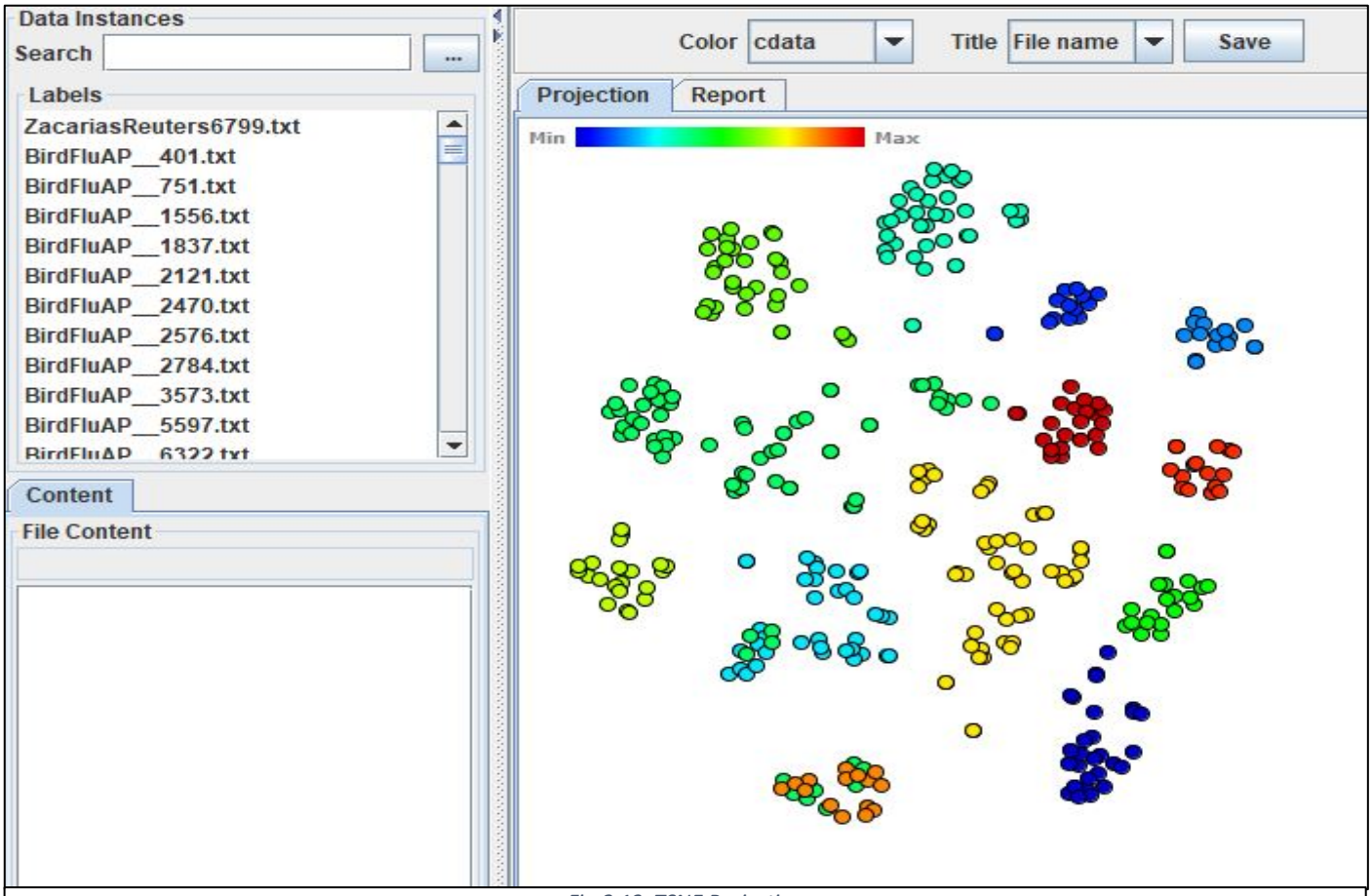

Fig 2.13. TSNE Projection

2. Medical12Classes

The data named Medical12Classes is a collection of MRI scans. It consists of multiple MRI scans images. Fig 2.15 shows the pipeline connections I used to set up for the document processing and analysis. First we have to select the file 'Medical12Classes.zip' in the Zip pre-processor and connect to Points Matrix Writer and execute to get the 'Medical12Classes.data' file. We input this file in Detailed Points Matrix Reader. The detailed points matrix reader is connected to NJ, LSP, TSNE. NJ is connected to tree model, then radial layout then Tree View frame to view the projection, also has a tree model silhouette coefficient. TSNE is connected to a topic Projection Model and a Topic Projection View Frame to view the projection. LSP is connected to a Labelled Projection Model and a Labelled Projection View Frame and also has a Labelled Projection Model Silhouette Coefficient.

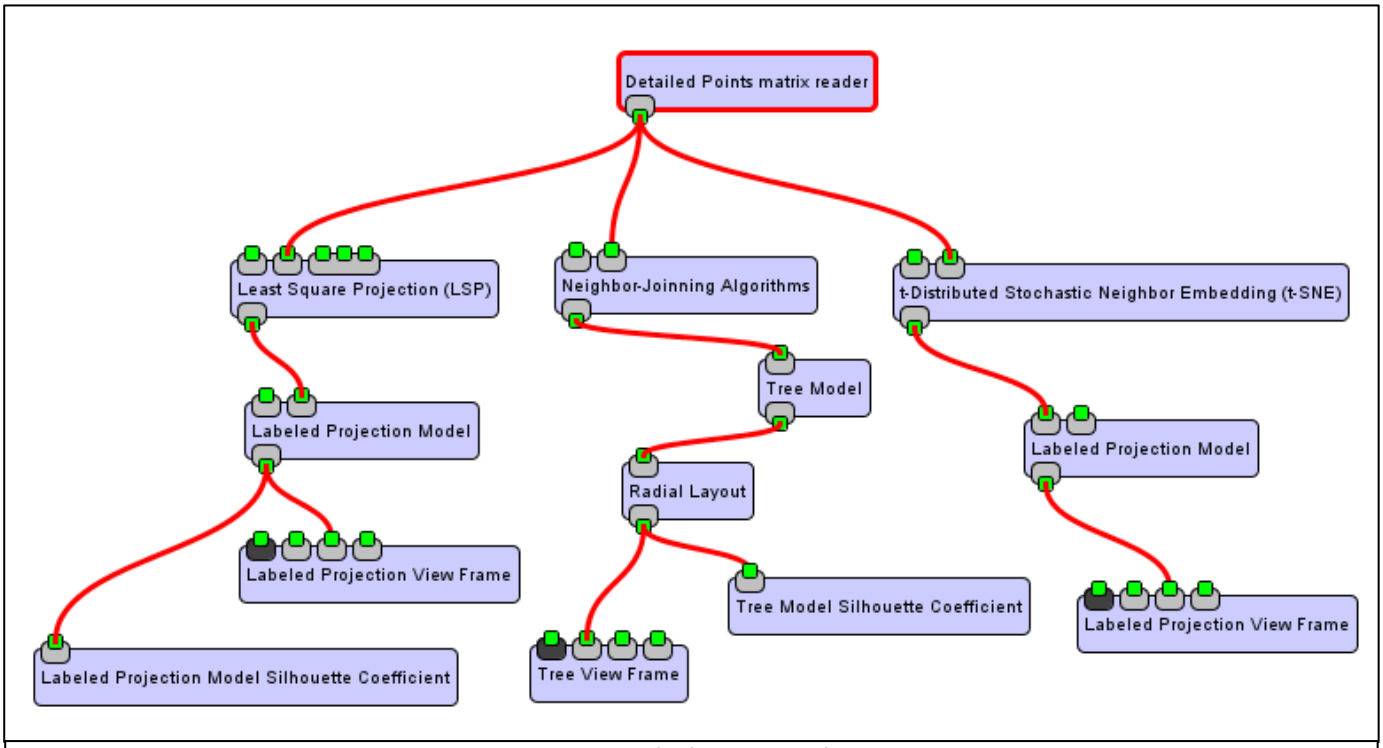

Fig 2.15. Medical Dataset Pipeline

Figures 2.16, 2.17, 2.18 show the projections I obtained running the pipeline.

LSP

Fig 2.16 shows the LSP Projection, we can see that the classification is not that good as there are overlaps and spread-out points.

Parameters:

Number of iterations: 200
Fraction of delta: 8.0
Number of control points: 20
Number of Neighbors: 20
Dissimilarity: Euclidean
LSP silhouette coefficient obtained (Euclidean): -0.1891
The projection is not that great. We can see from the figure that some classifications are not accurate. Figure 2.16.1 shows MRI scans of brain, mouth and face all being classified together as same.

t-SNE

TSNE models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability.

Parameters:

Initial dimension: 30
Target dimensions: 2
Perplexity: 30
Maximum iterations: 1000
Dissimilarity: Euclidean

TSNE (fig 2.17) does a very good job in clustering the data into different clusters so we can conclude it is the best among all projections. There are some files misclassified together.

NJ

Neighbour Joining is a bottom-up (agglomerative) clustering method for the creation of phylogenetic trees. NJ (fig 2.18) has many branches to classify the data into multiple clusters of similar data.

Parameters:

Dissimilarity: Euclidean.
NJ Algorithm: Rapid Neighbor-Joining
NJ silhouette coefficient obtained: 0.0596

We can conclude that it does a better job in segregation than LSP.

Conclusion

LSP did the worst among the three. Dimensionality reduction sometimes causes a loss in the data which might lead to cases of being misclassified. T-NSE even though its decent compared to LSP. t-SNE is also a method to reduce the dimension. One of the most major differences between PCA and t-SNE is it preserves only local similarities whereas PA preserves large pairwise distance maximize variance. It takes a set of points in high dimensional data and converts it into low dimensional data [9]. Here NJ algorithm did a good job in segregation of the images.

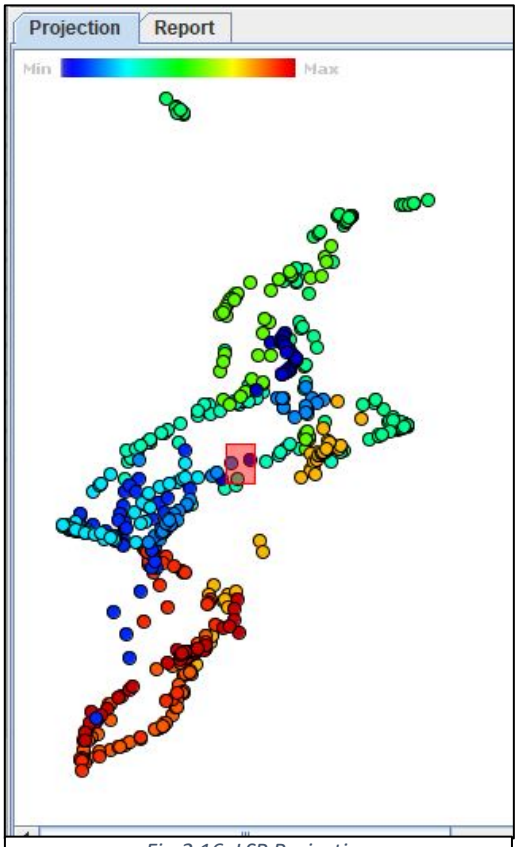

Fig 2.16. LSP Projection

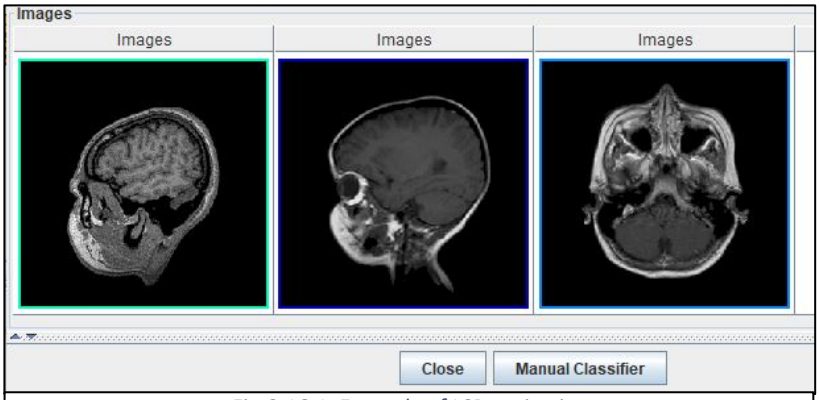

Fig 2.16.1. Example of LSP projection

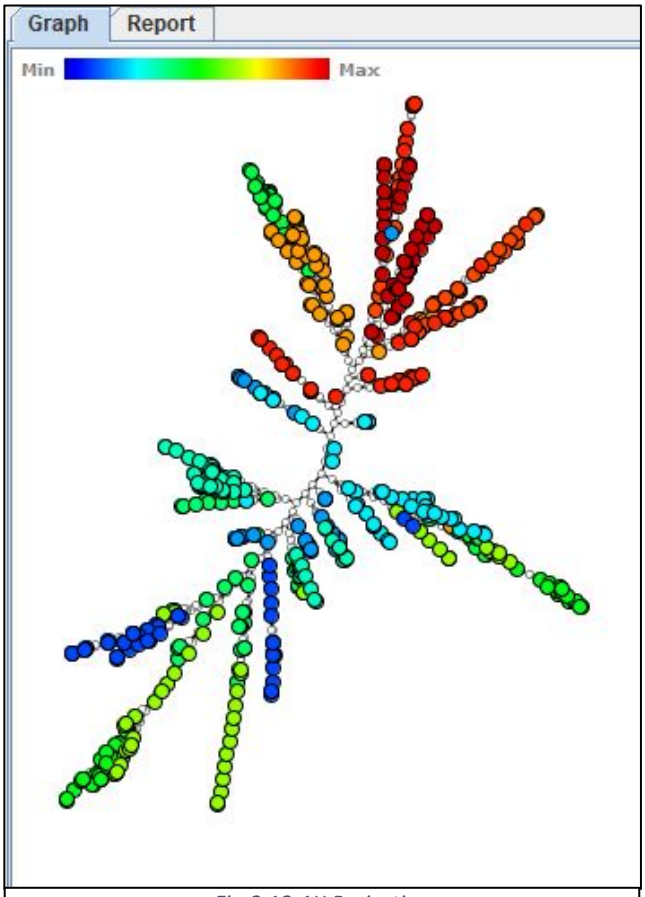

Fig 2.18. NJ Projection

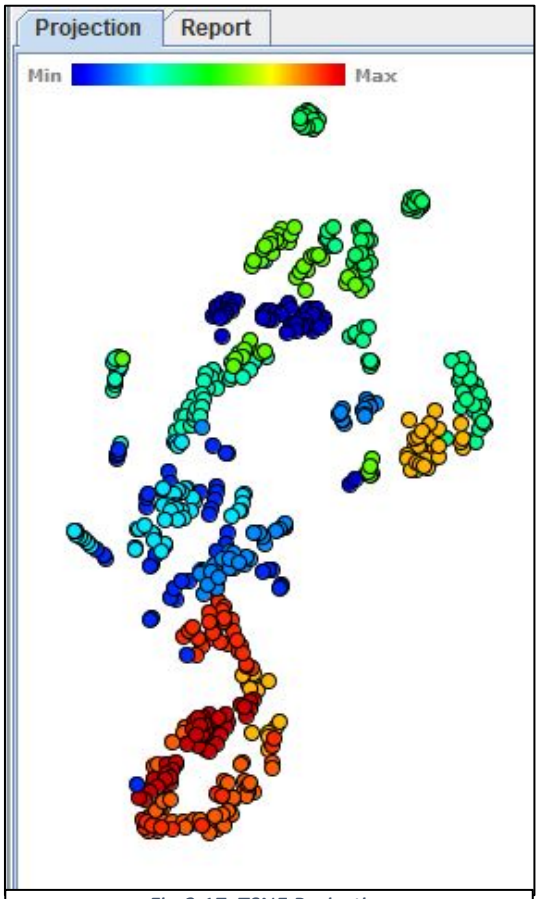

Fig 2.17. TSNE Projection

(c) Exploration of HDR dataset:

We're looking at the HDR dataset with the help of a document gathering workflow. There are 189 nations as in HDR dataset, each with 21 variables that describe various aspects of the country, such as life expectancy, GDP, and Gross National Income (GNI). We are explicitly converting the.csv file to a.data file, which is then imported as input for the predictions. The HDR data - set is subjected to the following unique projection techniques: Least-Square Projection (LSP), t-SNE, and Neighbour-Joining Tree. Fig 2.18 shows the pipeline for the HDR dataset.

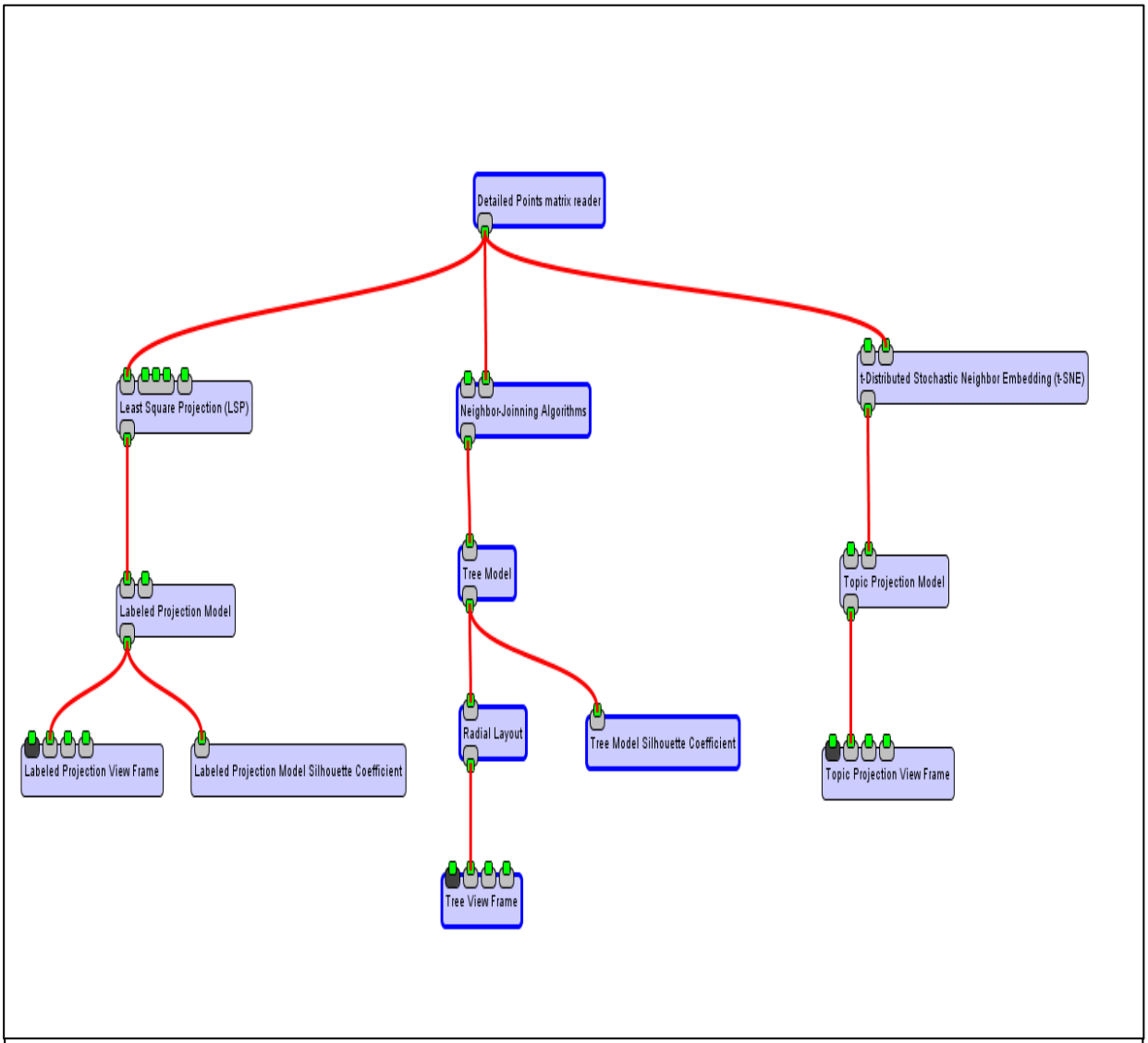

Fig 2.18. HDR Dataset Pipeline

LSP

Least Square Projection (LSP), tries to preserve neighbourhood. The core idea behind LSP is to project a subset of points and interpolate the rest which leads to preserving of neighbourhood. The HDR data has 189 different countries with 21 attributes.

Parameters:

Number of iterations: 50
Fraction of delta: 8.0
Number of control points: 100

Number of Neighbors: 60
Dissimilarity: Euclidean
LSP silhouette coefficient obtained (Euclidean): 0.24247903
The projection is not that great. We can see from the figure that some classifications are not accurate. Increasing the control points and neighbours seems to yield better result

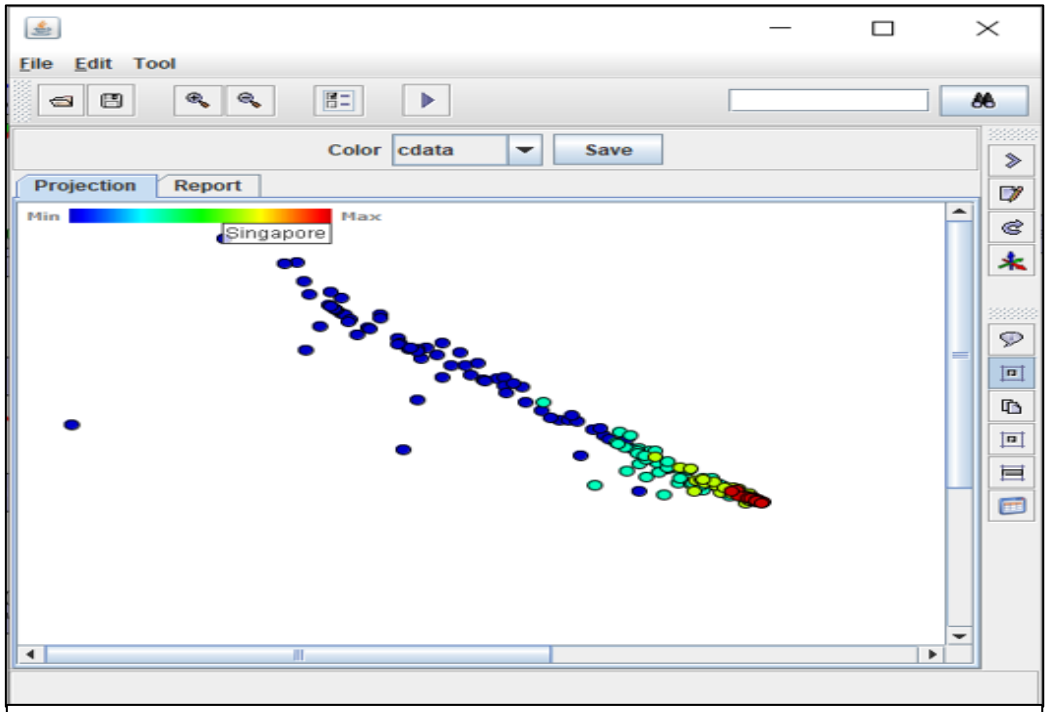

Fig 2.19. LSP Projection

t-SNE

TSNE models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modelled by nearby points and dissimilar objects are modelled by distant points with high probability.

Parameters:

Initial dimension: 30
Target dimensions: 2
Perplexity: 30
Maximum iterations: 1000
Dissimilarity: Euclidean
Silhouette coefficient (Euclidean): 0.3659109
TSNE (fig 2.20) does a very good job in clustering the countries into different clusters based on the labels. Hence we can conclude it is the best among all projections.

NJ

Neighbour Joining is a bottom-up (agglomerative) clustering method for the creation of phylogenetic trees. NJ (fig 2.21) has many branches to classify the data into multiple clusters of similar data.

Parameters:

Dissimilarity: Euclidean.
NJ Algorithm: Rapid Neighbor-Joining
NJ silhouette coefficient obtained: 0. 27727067

We can conclude that it does a better job in segregation than LSP but not t-SNE. Notice that changing the parameters leads to reduced silhouette coefficient values.

Conclusion:

LSP did the worst among the three. NJ even though its decent compared to other projections is not better than t-SNE. t-SNE is also a method to reduce the dimension. One of the most major differences between LSP and t-SNE is it preserves only local similarities whereas PA preserves large pairwise distance maximize variance. It takes a set of points in high dimensional data and converts it into low dimensional data [9].

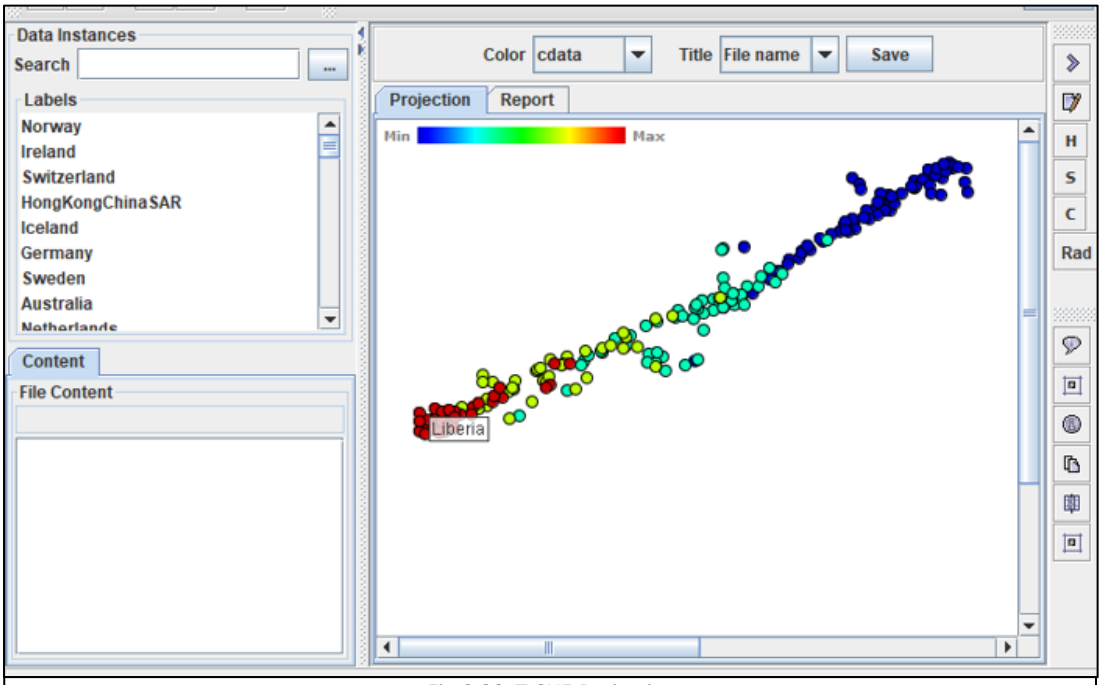

Fig 2.20. T-SNE Projection

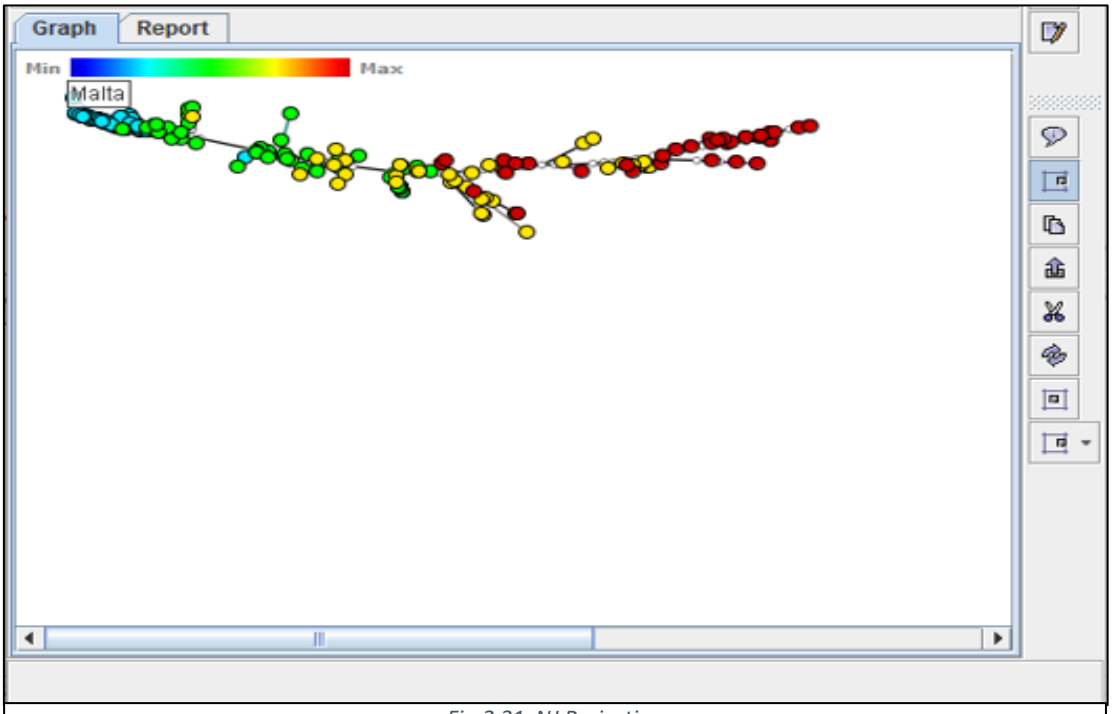

Fig 2.21. NJ Projection

Conclusion

For this task, I performed a visual report of survey data using multidimensional projection techniques. For task 1, I made use of HDR dataset and processed into a new file 'Book1.xlsx'. The new file was modified by labels, 1 for developed country and 4 for underdeveloped country. Performed visualizations on Tableau using techniques such as scatter plots, bubble charts, map plots, tree plots, bar graphs, line graphs, etc. Made use of HDI dataset based on the condition given and classified them into labels 1 to 4. I compared the attributes in the HDR dataset based on country, based on correlations and uncorrelations. I explained the expected and unexpected values and the hypothesis for the cause in their abnormal values. Next I explained some of the projection techniques used for visualizations, their advantages, and disadvantages. I made new visualizations for two old ones and explained new insights obtained from them.

For task 2, I made use of VisPipeline to perform various visualizations like t-nse, LSP, MDS, Isomaps, etc. I make conclusions based on the visual segregation done by the visualizations and based on the silhouette coefficient. I also try changing the parameters of the techniques in hopes to get better classification of the datasets. For text visualization I first convert the dataset to a points matrix using a Zip processor and a points matrix writer. Made sure to remove the stop words so it would not hinder the classification. Analyzed the Lehn's cut to get the most repeated word in the document. For image classification I performed the techniques to find the segregations.

References

[1] https://link.springer.com/chapter/10.1007/978-3-662-43591-5_9
[2] https://hdr.undp.org/en/content/inequality-adjusted-human-development-index-ihdi
[3] https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-exam
[4] http://viniciusrpb.byethost6.com/files/visualdm.pdf?i=1
[5] https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding
[6] https://en.wikipedia.org/wiki/Multidimensional_scaling
[7] https://en.wikipedia.org/wiki/Neighbor_joining
[8] https://en.wikipedia.org/wiki/Isomap
[9] https://medium.com/analytics-vidhya/pca-vs-t-sne-17bcd882bf3d