Name: Anasua Das

ID: 120220427

Module: CS6426

Report Title: Assignment1 Report Year:2022

INTRODUCTION:

Data in the 21st Century is like Oil in the 18th Century: an immensely, untapped valuable asset. (TOONDERS)

Standing in the 21st century, we are living in a world controlled and performed by data.

Data is everywhere and with the progress of digitization, data is not only collected and stored in computers drives or as paper documents; it's all over the web and highly agile in nature. Data is created and transformed in every fraction of seconds which puts some direction to each and every unit of our ever changing world. We can use this data in medical, F&B, textile, Retail, Banking and Finance, EdTech and all sorts of industries that we have in mind. In this course, one of the greatest challenges with structure of data that we scrape from the web and then analyzing and understanding the meaning behind such data in terms of their dimensional complexity. Data obtained from various sources come in various formats which we can classify as structured or unstructured. In addition to that, dealing with multi-dimensional datasets with typically more than two attributes tend to be more challenging and also create a lot of problems. Here, different data visualization techniques come to rescue. Data visualization provides a number of techniques which help to reduce the dimensionality of the multidimensional datasets and get rid of the problems that were cropping up due to the multidimensionality.

The first half of the report tries to understand Human Development Report Data and evaluate different parameters of the data, their underlying potential and their impact on Human Development Index of each country across the world.

The second part of the report aims to analyse different projection methods to analyse different types of image and text datasets. We evaluate accuracy and efficiency of different projection techniques like e t-distributed Stochastic Neighborhood embedding, Multidimensional Scaling, Least Squared Projection.

Data Description:

Following data have been used at different stages of this report generation.

HDR dataset: Collected from canvas, updated during the lab session. Refer to [HDR_HY_prepared]

The Human Development Report (HDR) is an annual Human Development Index report published by the Human Development Report Office of the United Nations Development Programme (UNDP).[1]

Each HDR report for a particular year presents an updated set of indices, including the Human Development Index (HDI), which is a measure of average achievement in the basic dimensions of human development across countries, and a compendium of key development statistics relevant to the report theme. (Wikipedia, Human Development Report) (William G. Bowen)

Corel dataset: type: Corel data set is a vector space composed of image features, collection of photographs and drawings which contains 10 labels and 1000 items.

CBR dataset: It's a vector space model extracted from 600 + documents. It holds certain number of article abstracts.

Two other datasets of my choice: Medical12Classes and News_from_Lecture are used to evaluate performance of certain projection techniques.

HDI.csv -Data Courtesy: From a fellow student: Anjana Prem Kumar

HDR2020_DAT- Data Courtesy: From a fellow student: Madhu Mohan

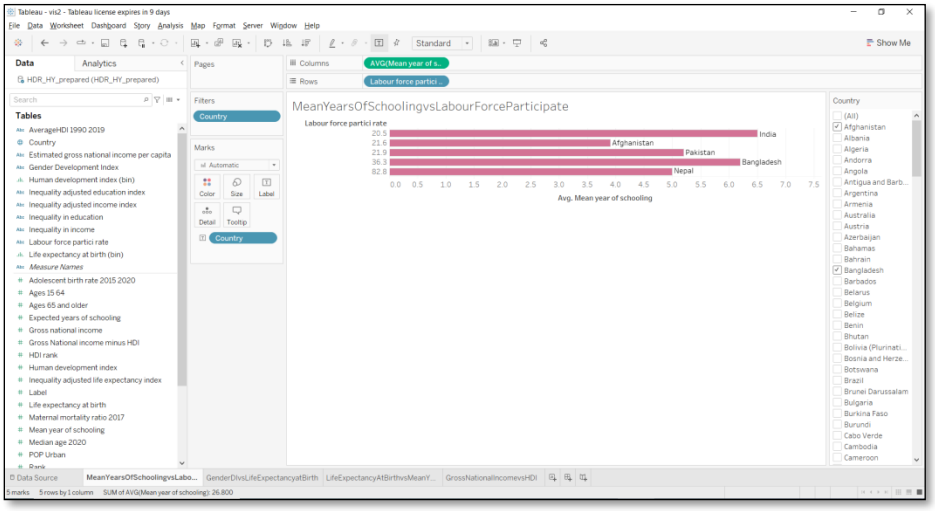
Used visualization tools like: Tableau and DataVispipeline for the transformations and operations on the dataset to extract meaningful visualization of their different features.

Task 1:

Free Exploration:

Mean Year of Schooling and Labour Force Participation Rate

Figure I.



Description of *Figure I:*

Two variables are taken into consideration: Mean Years of Schooling & Labour Force Participation Rate for five neighbor countries of my interest: India, Afghanistan, Pakistan, Bangladesh, and Nepal

Expected: If mean years of schooling increases, Labor Force participation rate should decrease.

Observed: We can see deviation from the expected result.

India:

Average Mean Years of Schooling (MYOS) = 6.5 units, Labour Force Participation Rate (LBFPR)= 20.5

Bangladesh:

Average Mean Years of Schooling is 6.2 units, Labor Force participation Rate= 36.3

Pakistan:

Average Mean Years of Schooling is 5.2 units, Labor Force participation Rate= 21.9

Nepal:

Average Mean Years of Schooling is 5.0 units, Labor Force participation Rate= 82.8

Afghanistan:

Average Mean Years of Schooling is 3.9 units, Labor Force participation Rate= 21.6

India: Shows Maximum MYOS, Minimum LBFPR

Bangladesh: Shows more LBFPR compared to MYOS

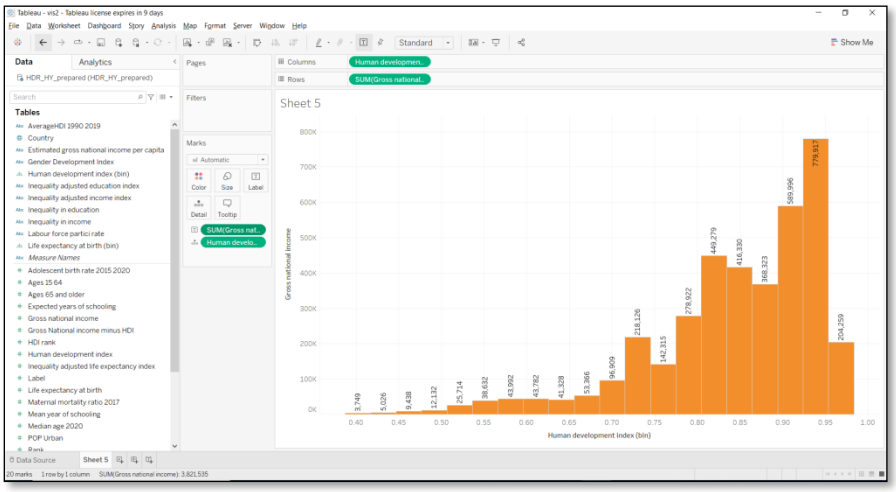
Afghanistan: Shows less LBFPR compared to MYOS

Nepal: Maximum LBFPR but third out of 5 in terms of MYOS

Hypothesis of uniformly decreasing LBFPR with increasing MYOS fails.

GNI Vs HDI

Figure II



Description of *Figure II.*

Variables considered: Gross National Income and Human Development Index

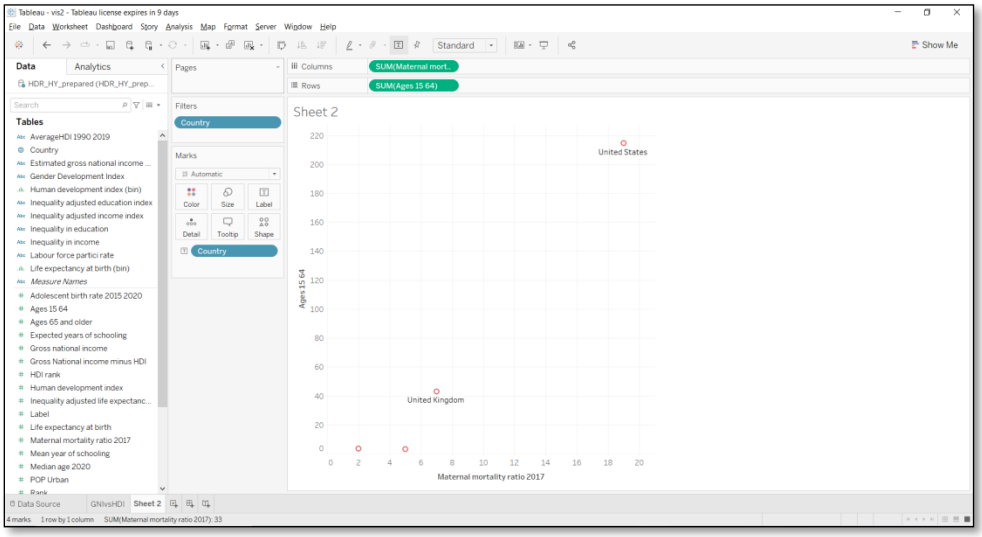
Expected: If Gross National Income increases, HDI also expected to increase

Observed: When Gross National Income is maximum at 7, 79,917, HDI lies between 0.925 – 0.95

Overall HDI increases to 0.95 with the increase of GNI with a few outliers present but from HDI range 0.95 and 1.00 Gross National income suddenly drops to a lower value.

Age 15 64 vs Maternity Mortality Ratio

Figure III



Description of *Figure III*

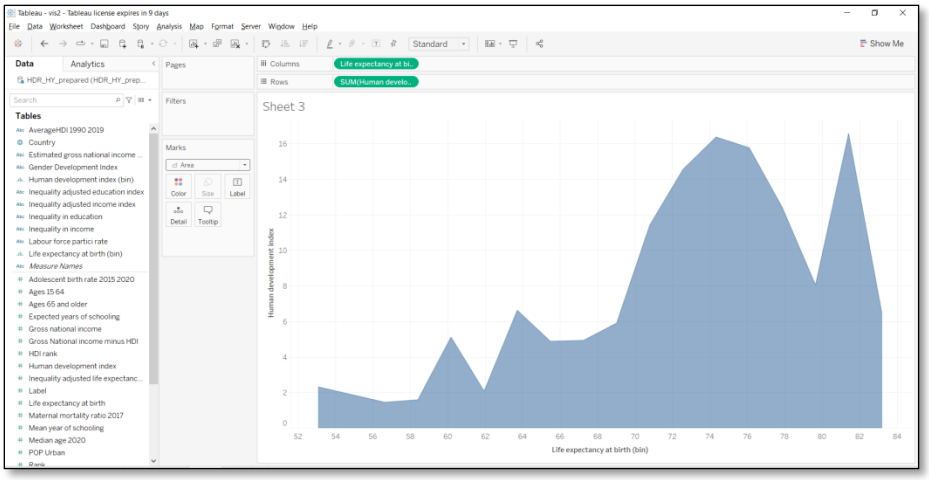
Variables Considered: Age 15 64 and Maternal Mortality Ratio 2017

Hypothesis: Norway, Ireland, United Kigdom, United States being one of the most developed countries, Maternal Mortatlity Ratio is expected to be minimum to nil.

Observed: The scatter plot shows that for the age group: 15-64, or Norway's and Ireland's Maternal Mortatlity Ratio is 0 whereas United Kingdom shows total of just above 40 maternal demise and United States shows the maximum Maternal Mortatlity Ratio count among these four countries at around 218 count.

Life Expectancy at Birth vs Human development Index

Figure IV



Description of *Figure IV:*

Variables taken: Life Expectancy at Birth (LEAB) and Human Development Index

Expected: Higher HDI with Higher Life Expectancy at Birth

Observed: The Area Curve shows that HDI with 16.57 (max) shows highest life expectancy at birth 81.

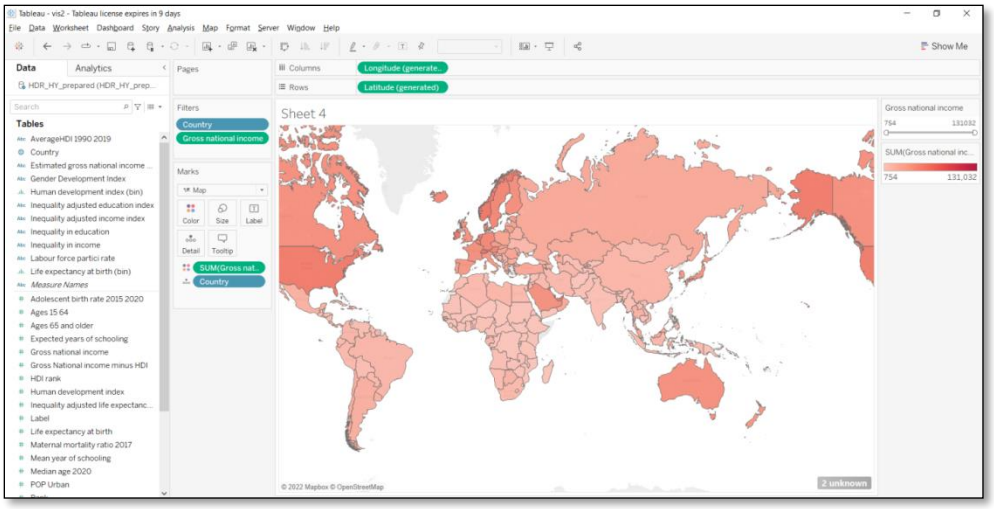
The area curve shows that the hypothesis is not maintained throughout and HDI dops to lower values for certain cases even if the life expectancy is large.

Which indicates that there must be effect of some other parameters on HDI.

Specific Observations:

Countries vs Gross National Income

Figure V



Description of *Figure V*

Variables considered: Countries and their Gross National Income.

With the increasing GNI the shade increases from light red to dark red.

Figure VI

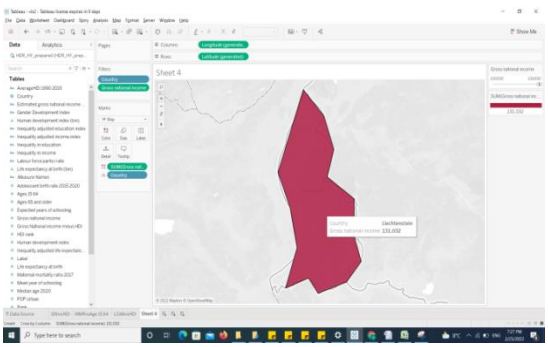
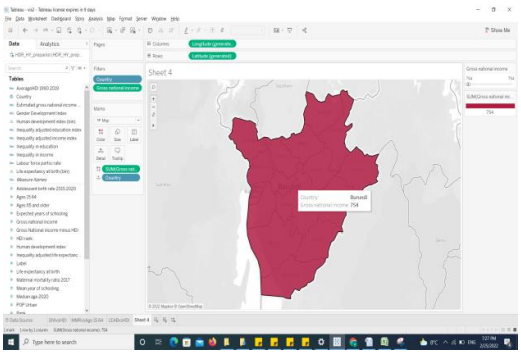


Figure VII



Country with Max Gross National Income:

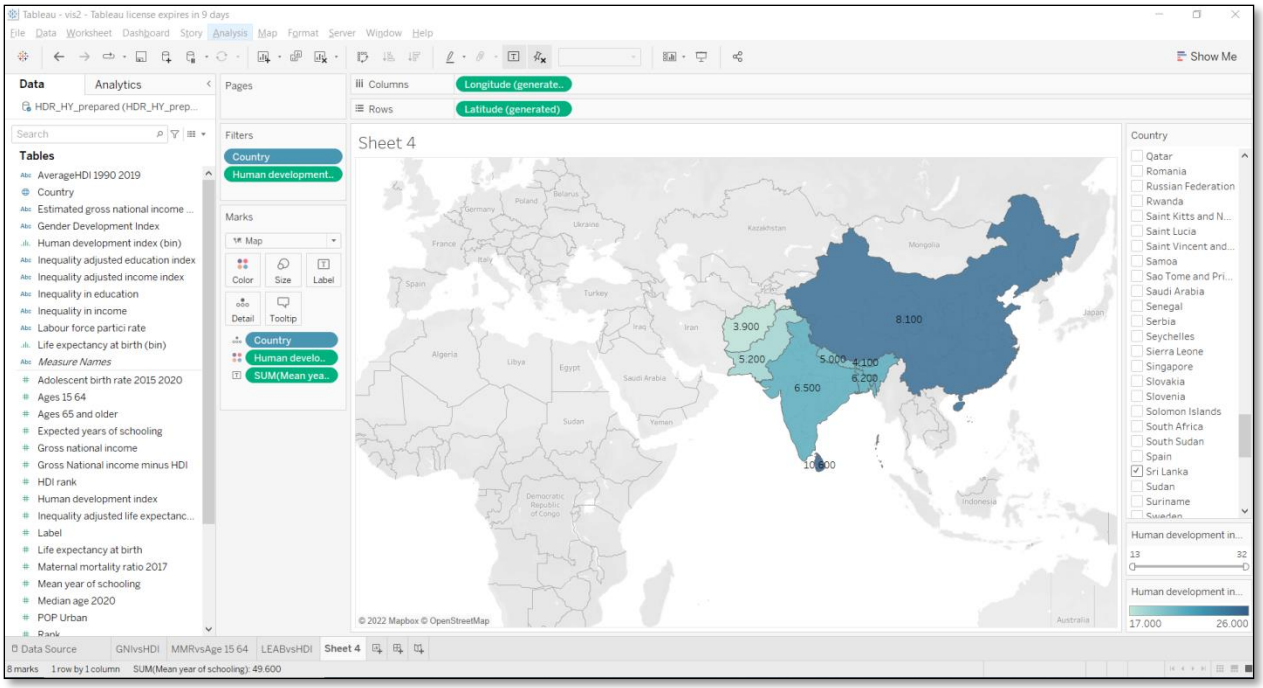
Liechtenstein with GNI: 1,31,032

Country with min Gross National Income:

Burundi with GNI: 754

Mean Years of Schooling vs Human Development Index

Figure VIII



Description of *Figure VIII:*

Here countries taken: Neighbouring Countries: India, Nepal, Bhutan, Bangladesh,Pakistan, Afghanistan, China and Sri Lanka

Their Mean Years of Schooling and HDI parameters are taken into consideration.

Hypothesis: With increasing Mean Years of Schooling, HDI should increase

Where,

Country Name	**Mean Years of Schooling**	**HDI**	**Comment**
Sri Lanka	10.6	0.77	Highest MYOS, highest HDI
China	8.1	0.74	2nd Highest MYOS,2nd Highest HDI
India	6.5	0.63	3rd Highest MYOS,3rd Highest HDI
Bangladesh	6.2	0.63	MYOS slightly lower than INDIA but equal HDI
Pakistan	5.2	0.54	4th highest MYOS, 4th Highest HDI
Nepal	5	0.60	MYOS is lower than Pakistan but higher HDI #some other parameter might have played role

			for increased HDI
Bhutan	4.1	0.63	MYOS is lower than Pakistan and Nepal but higher HDI than both #some other parameter might have played role for increased HDI
Afghanistan	3.9	0.51	Lowest MYOS, lowest HDI

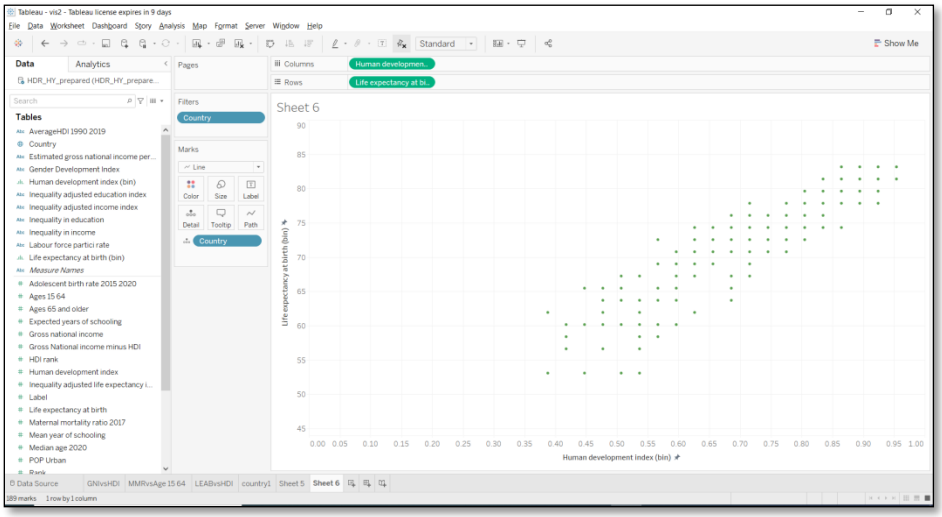
For Nepal and Bhutan, HDI is higher even though MYOS is less.

From sources:

"Today, conditions in Bhutan are very different from what they were in the sixties. Over 90% of the population has access to primary health care. Gross primary enrolment date has reached 72%. Life expectancy at birth has gone up to 66 years." (Programme, 2020)

"As per the report, between 1990 and 2018, Nepal's HDI value increased from 0.380 to 0.579, an increase of 52.6 per cent, basically driven by **increased life expectancy of people and years of schooling**. (UNDP)

Figure IX



Description of *Figure IX:*

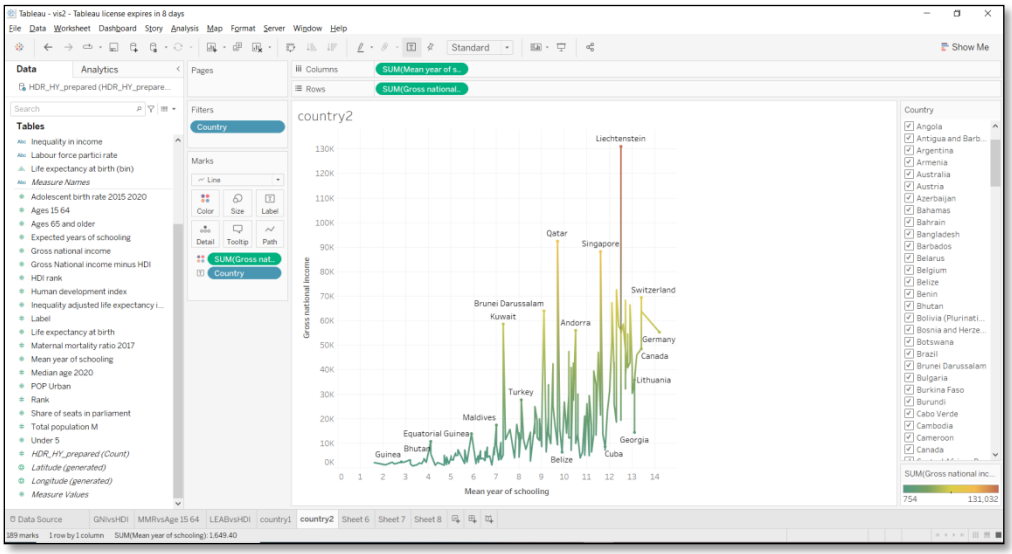
Considering all the countries we plot 'Life Expectancy at Birth' against HDI and take a scatter plot.

The result of the scatter plot shows that the two variables are highly correlated and they are having a strong linear relationship.

Low LEAB contributes to low HDI and higher value of LEAB contributes to higher HDI**.**

Mean Years of Schooling Vs Gross National Income

Figure X



Description of *Figure X:*

Variables considered: Mean Years of Schooling(MYOS) and Gross National Income (GNI) of all the countries

When the variables are plotted against eachother we can see a gradually increasing curve which shows MYOS is positive contributor of GNI. If MYOS increases, GNI should also increases.

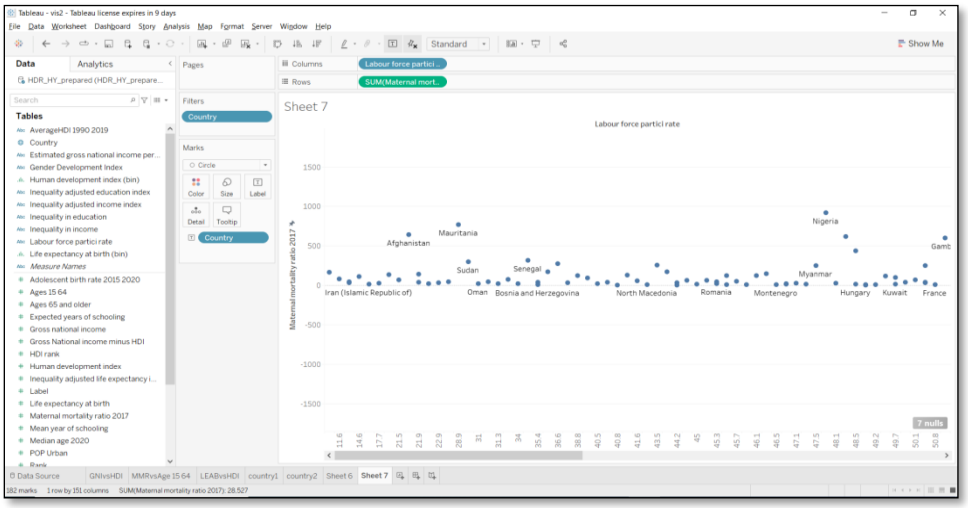
There are some sharp peaks and dips in the curve, they are the outliers.

Kuwait, Singapore, Qatar, Brunnei Darussalam, are having higher GNI given their MYOS

Zimbabwe, Uzbekistan,Tajikistan, Marshall Islands are a few of those which are having lower GNI given their MYOS

Maternal Mortality Ratio vs Labor Force Participate

Figure XI



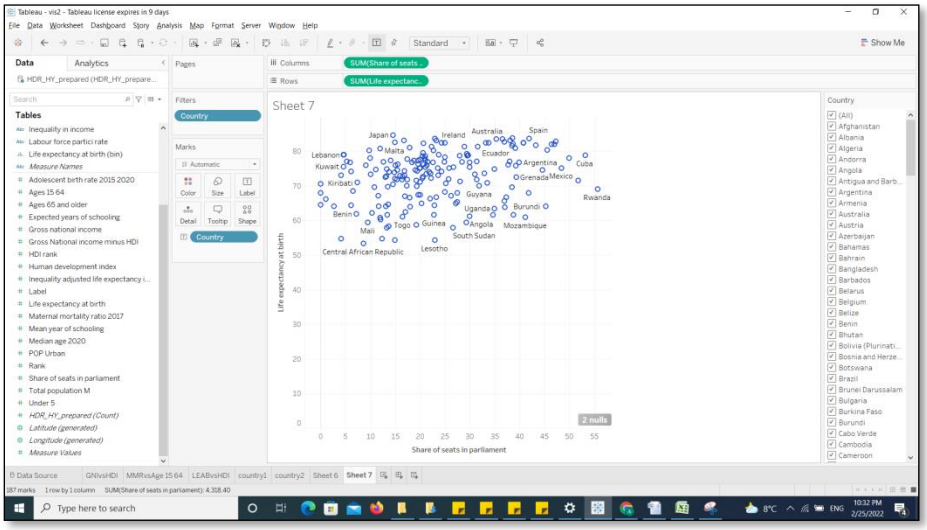
Description of *Figure XI:*

Parameters considered: Maternal Mortality Ratio and Labour Force Participate for all the countries.

These two parameters don't show much correlation between them and are randomly distributed.

Share of Seats in Parliament vs Life Expectancy at Birth

Figure XII



Description of *Figure XII:*

Parameters: Share of seats in Parliament and Life expectancy at Birth

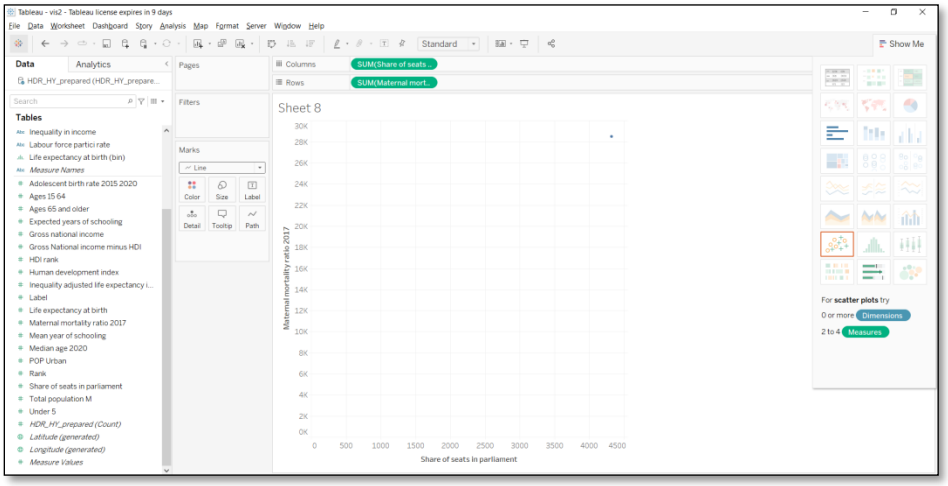
Hypothesis: The given parameters are not dependent on each other or uncorrelated while The former parameter indicates The proportion of seats held by women in national parliaments is the number of seats held by women members in single or lower chambers of national parliaments, expressed as a percentage of all occupied seats; it is derived by dividing the total number of seats occupied by women by the total number of seats in parliament where as the later indicates to the expected life of an infant at the time of birth.

But the scatter plot here shows some correlation between them which violates the hypothesis.

Shows a potential correlation between them.

Share of seats in Parliament vs Maternal mortality

Figure XIII



Description of *Figure XIII:*

Parameters considered: Share of seats in Parliament and Maternal mortality Ratio for all the countries.

Hypothesis: The two parameters are not correlated.

These parameters seem unrelated to each other as the scatter plot fails to show any promising argument to fail the hypothesis

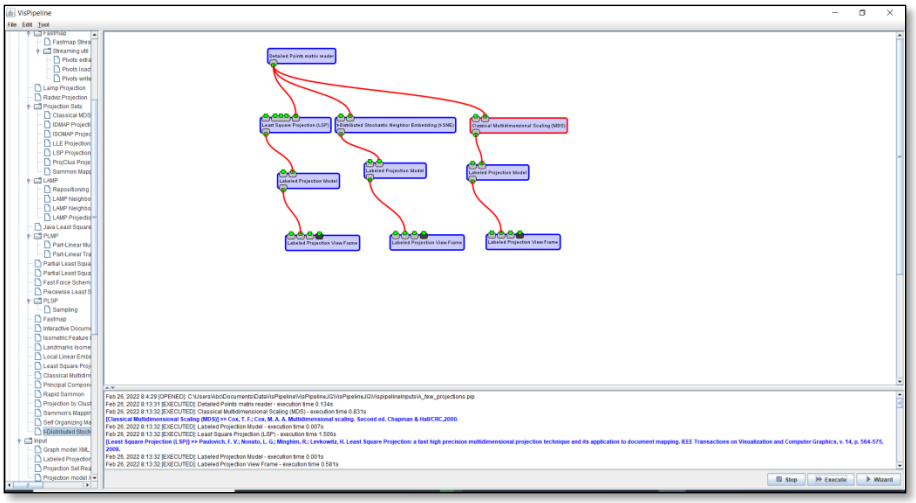
Task 2.

EMPLOYING PROJECTIONS TO VISUALISE MULTIDIMENSIONAL DATA

Dataset taken: Image Corel dataset. Description given in the Data Description part.

Projections taken: Least Square Projection(LSP), t-SNE, Multidimensional scaling (MDS)

Figure XIV



Visualization of LSP with Default Parameters:

Figure XV

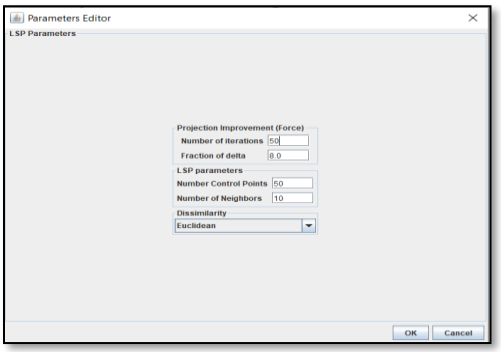
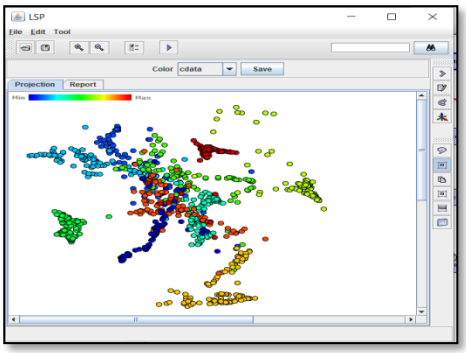


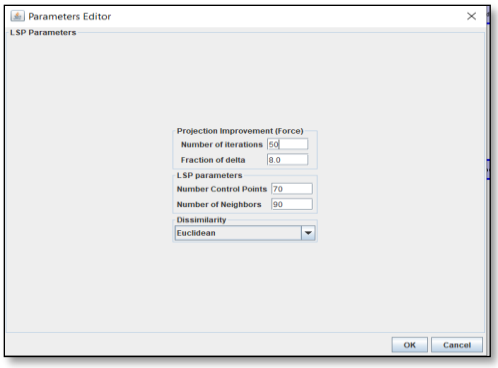
Figure XVI

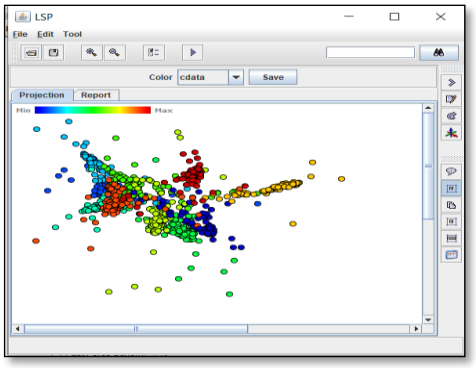


Visualization of LSP with Changed Parameters:

Figure XVII

Figure XVIII





Please refer to Table: 1 for the comparative discussion of the above visualizations:

Table 1

Projections	Parameters	Datasets Corel Dataset	Original Silhouette Coefficient	Projection silhouette Coefficient
LSP	Default Parameters	We get proper classification for majority of the datapoints but a few are overlapped at the center.Clusters of same coloured datapoints are found together. Data points are not too much scatterd all over the projection plane It'sa good projection as projection silhouette coefficient is greater than base silhouette coefficient.	0.159	0.287
	Changed Parameters	We get to see same colors together but the different sets of colors are lying side by side.Here we see that a few datapoints have moved away and dispersed randomly on the projection plane. This also turns out to be a good projection but inferior to the projection obtained by default parameters. Silhouette coefficient slightly drops from 0.287 to 0.236. There's a marginal loss in the changed parameter	0.159	0.236

Visualization of t-SNE with Default Parameters:

Figure XXIX

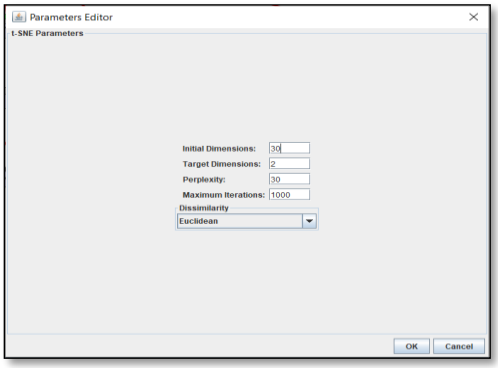
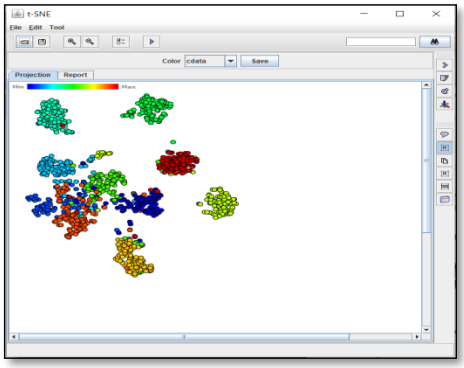


Figure XX



Visualization of t-SNE with Changed Parameters:

Figure XXI

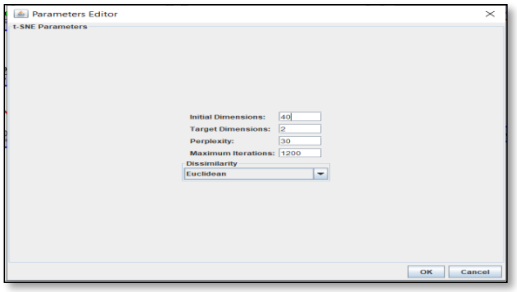
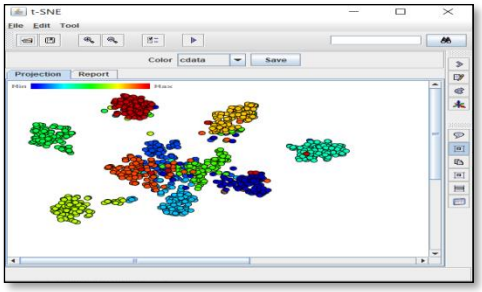


Figure XXII



Please refer to Table: 2 for the comparative discussion of the above visualizations:

Table:2

Sl.No	Projections	Parameters	Datasets Corel Dataset	Original Silhouette Coefficient	Projection silhouette Coefficient
	t-SNE	Default Parameters	We get to see a proper classification of datapoints. Silhoette coefficient shows a higher value than the original silhoette coefficient which signifies a good projection	0.159	0.4628
		Changed Parameters	There's no much change with the new parameters for tsne. Datapoints are well classified and shows proper segregation of images.	0.159	0.463

Visualization of MDS with Default Parameters:

Figure XXIII

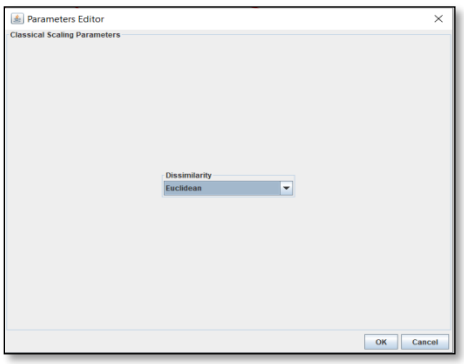
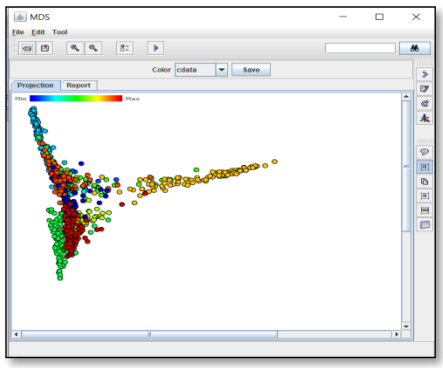


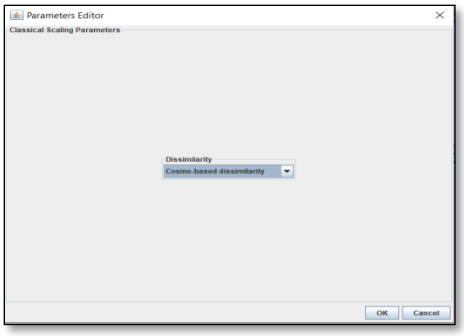
Figure XXIV

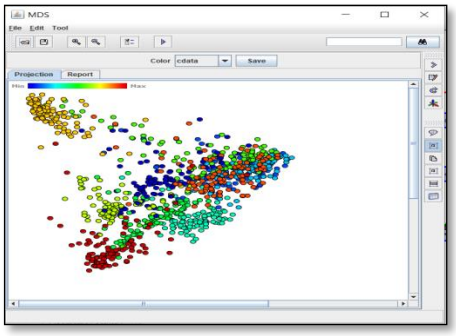


Visualization of MDS with Changed Parameters:

Figure XXIV

Figure XXIV





Please refer to Table: 3 for the comparative discussion of the above visualizations:

Table 3:

Sl.No	Projections	Parameters	Datasets		
			Corel Dataset	Original Silhouette Coefficient	Projection silhouette Coefficient
	MDS	Default Parameters	Poor segregation of the datapoints and overlapping is found. Not a good projection as the projection silhoette coefficient value is way less than the original value.	0.159	0.0523
		Changed Parameters	There's slight improvement for projection silhoette coefficient with changed parameters for MDS and reaches closer to original value. Even though overlapping is present, a few clusters are also developed.	0.159	0.096

Dataset taken: CBR dataset. Description given in the Data Description part.

Projections taken: Least Square Projection(LSP), t-SNE, Multidimensional scaling (MDS)

Visualization of LSP with Default Parameters:

Figure XXV

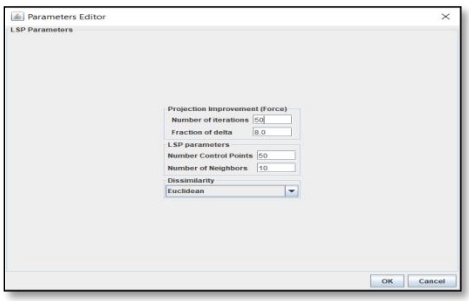
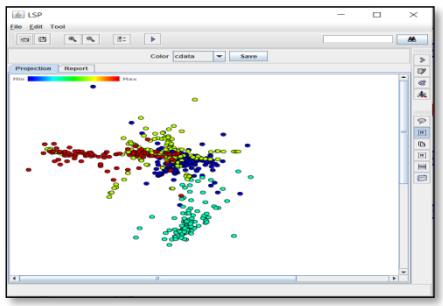


Figure XXVI



Visualization of LSP with Changed Parameters:

Figure XXVI

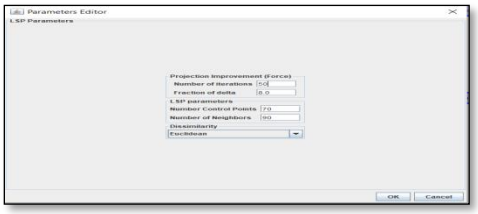
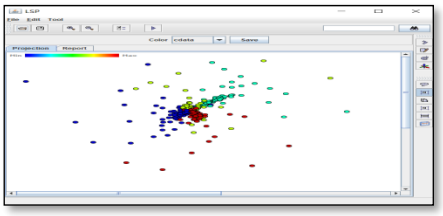


Figure XXVII



Please refer to Table: 4 for the comparative discussion of the above visualizations:

Table:4

Projections	Parameters	Dataset CBR dataset	Original Silhouette Coefficient	Projection Silhouette Coefficient
LSP	Default Parameters	LSP Projection Silhouette coefficient with default parameters justifies for a good projection. Similar type of texts are clustered together and represented by separate coloured datapoints. However, datapoints overlap with each other in the center. It's expected that segregation task would be more difficult in this region	0.0095	0.275
	Changed Parameters	Projection with changed parameters showa good projection overall as the segregation of datapoints is done properly. However datapointsare too much scattered over the projection pane .Silhouette Coefficient value also drops in the projection with changed parameters.	0.0095	0.221

Visualization of t-SNE with Default Parameters:

Figure XXVIII

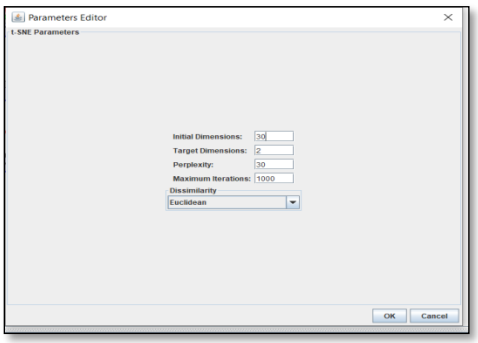
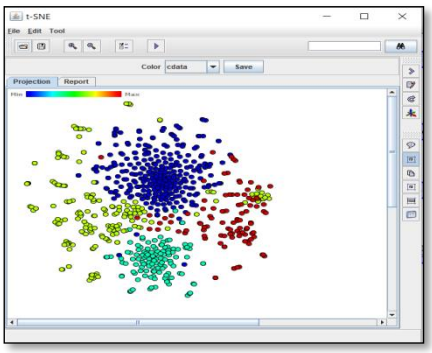


Figure XXVIII



Visualization of t-SNE with Changed Parameters:

Figure XXIX

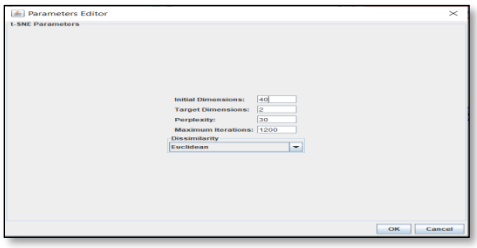


Figure XXX

Please refer to Table: 5 for the comparative discussion of the above visualizations:

Table:5

Projections	Parameters	Dataset CBR dataset	Original Silhouette Coefficient	Projection Silhoette Coefficient
t-SNE	Default Parameters	We get to see segregation of datapoints represented by same color for a cluster.There's a tendency of data to get scatteredness in the outwards direction.	0.0095	0.2688
	Changed Parameters	There's no much change in t-SNE projection with changed parameters.A bit of overlapping is also seen	0.0095	0.268

Visualization of MDS with Default Parameters:

Figure XXXII

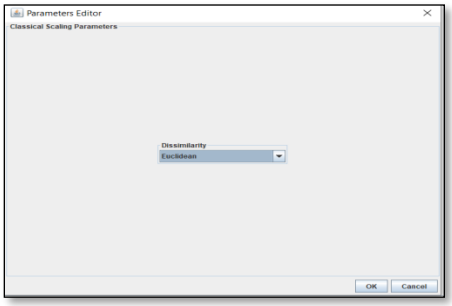
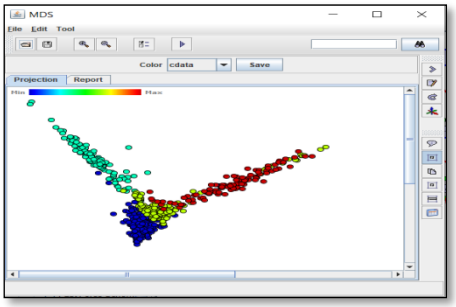


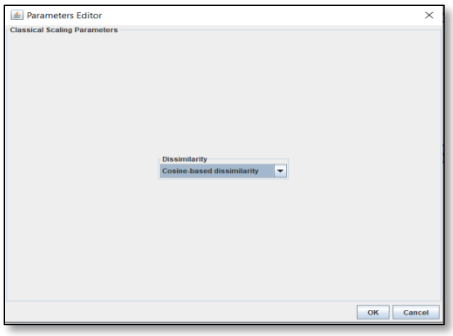
Figure XXXIII

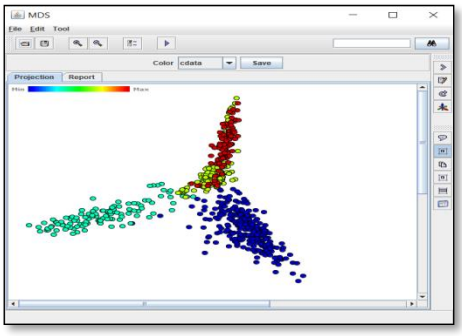


Visualization of MDS with Changed Parameters:

Figure XXXIV

Figure XXXIV





Please refer to Table: 6 for the comparative discussion of the above visualizations:

Table 6:

Projections	Parameters	Dataset CBR dataset	Original Silhouette Coefficient	Projection Silhouette Coefficient
MDS	Default Parameters	For CBR dataset, with default parameters we can see proper classification of datapoints. Projection Silhouette coefficient also shows higher value than that of original one.	0.0095	0.388
	Changed Parameters	There's an improvement in projection silhouette coefficient for the changed parameters of MDS. The segregation of data gets better.That's a gain.	0.0095	0.4547

3. Two other datasets taken: Medical12classes dataset, News_from_Lecture Dataset

Projections performed: Least Square Projection(LSP), t-SNE, Multidimensional scaling (MDS)

Medical12Classes Dataset

Visualization of LSP with Default Parameters:

Figure XXXV

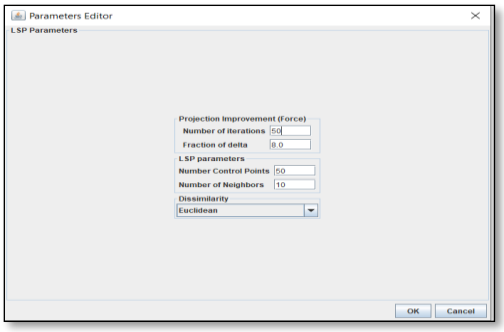
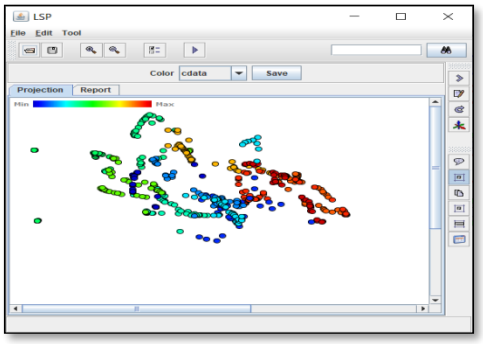


Figure XXXVI



Visualization of LSP with Changed Parameters:

Figure XXXVII

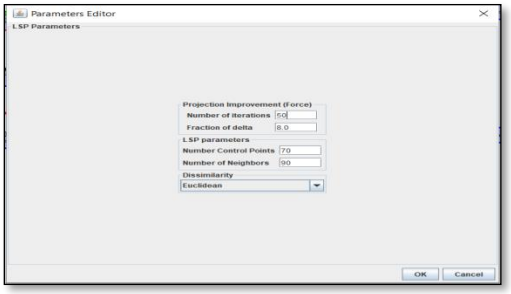
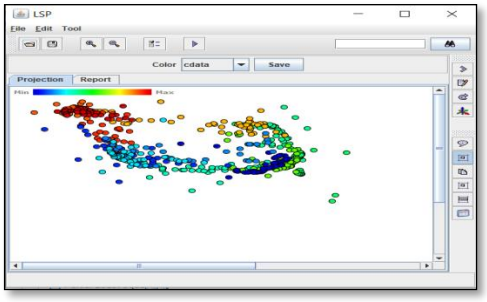


Figure XXXVIII



Please refer to Table: 7 for the comparative discussion of the above visualizations:

Table 7:

Projections	Parameters	Dataset Medical Classes	Original Silhouette Coefficient	Projection silhoutte Coefficient
LSP	Default Parameters	For Medical dataset with the defaulted parameters we get a bad projection as the same colours are not properly separated and have not formed clusters. Projection silhoette coefficient is also less than the original silhoette coefficient.	0.1177	0.0653
	Changed Parameters	By changing the parameters the coefficient drops below 0 which signifies that the data which formed the clusters may be incorrect. Here's loss is shown .	0.1177	-0.032

Visualization of t-SNE with Default Parameters:

Figure XXXIX

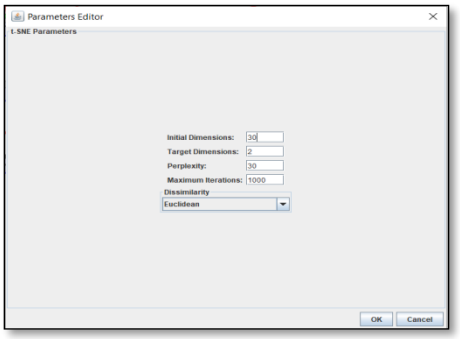
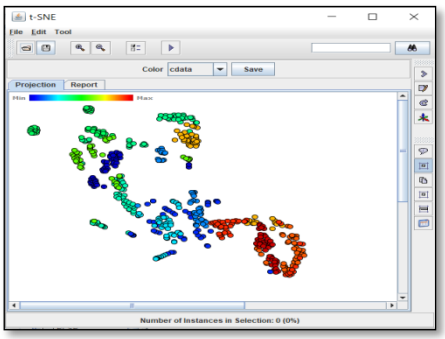


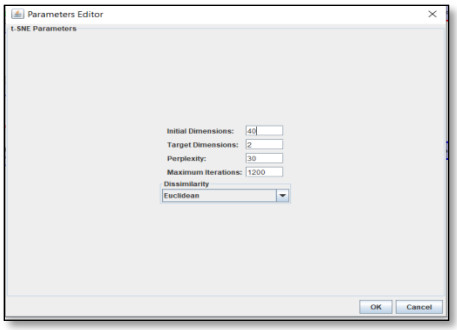
Figure XXXX

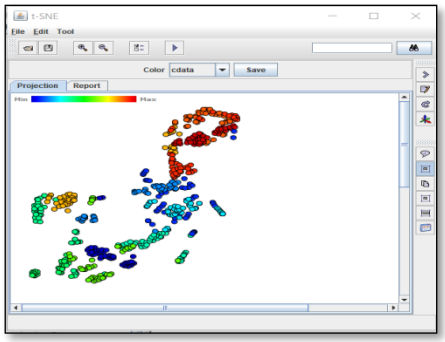


Visualization of t-SNE with Changed Parameters:

Figure XXXXI

Figure XXXXII





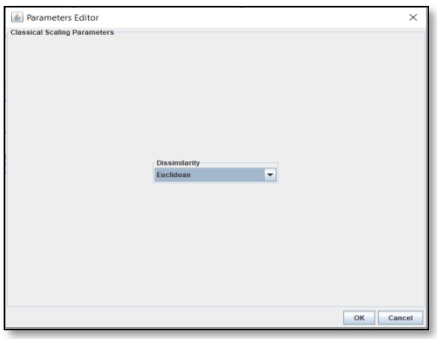
Please refer to Table: 8 for the comparative discussion of the above visualizations:

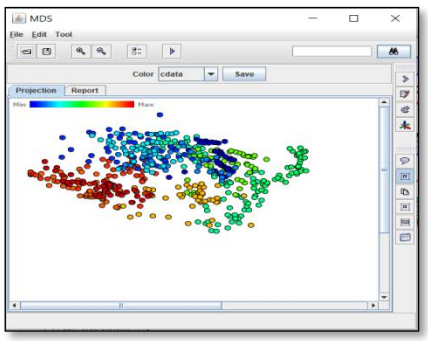
Table:8

Projections	Parameters	Dataset Medical Classes	Original Silhouette Coefficient	Projection silhoutte Coefficient
t-SNE	Default Parameters	There's no proper segregation of data and clusters are not found.Projection is not good as per standard as projection silhoette coefficient value is less than the original value	0.1177	0.0628
	Changed Parameters	Projection improves with changed parameters and projection silhoette coefficient value gets a slight leap and becomes more than the original silhoette coefficient value. This signifies gain in the segregation of datapoints.	0.1177	0.136

Visualization of MDS with Default Parameters:

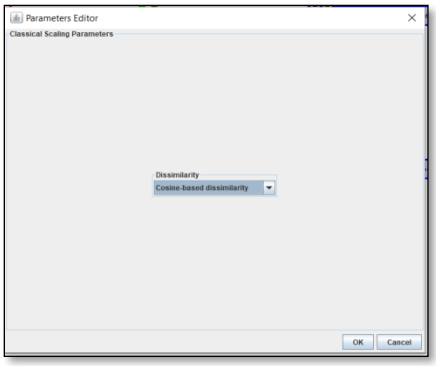
Figure XXXXIII ***Figure XXXXIV***

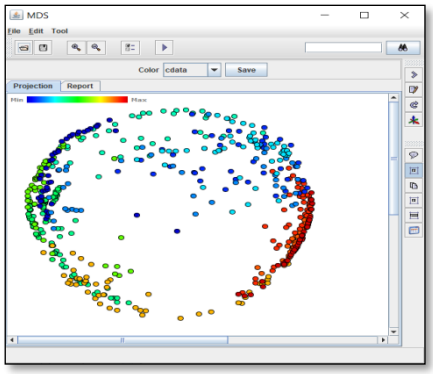




Visualization of MDS with Changed Parameters:

Figure XXXXV ***Figure XXXXVI***





Please refer to Table: 9 for the comparative discussion of the above visualizations:

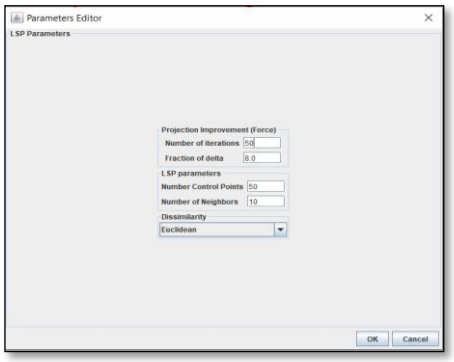
Table:9

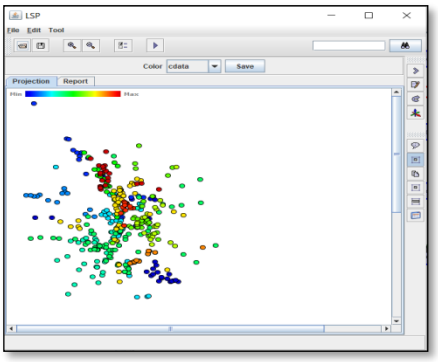
Projections	Parameters	Dataset Medical Classes	Original Silhouette Coefficient	Projection silhoutte Coefficient
MDS	Default Parameters	The projection doesn't show a good classification of data. Overlapping is clear. Silhoette coefficient also lies below the baseline.	0.1177	0.0347
	Changed Parameters	Projection deteriorates wih the change in MDS parameters and we see a loss in segregation.Projection silhoette coefficient drops to negative value and shows a poor projection of data segregation	0.1177	-0.071

4. News_from_Lecture Dataset

Visualization of LSP with Default Parameters:

Figure XXXXVII ***Figure XXXXVIII***





Visualization of LSP with Changed Parameters:

Figure XXXXIX

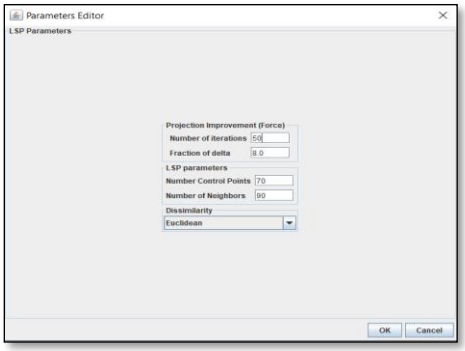
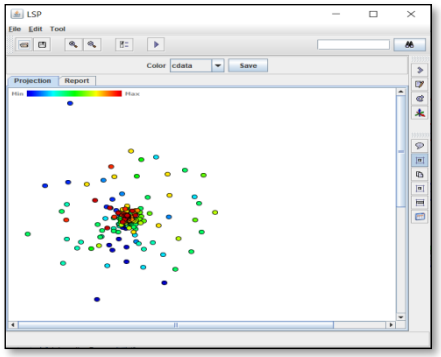


Figure XXXXX



Please refer to Table: 10 for the comparative discussion of the above visualizations:

Table 10

Projections	Parameters	Datasets News from Lecture	Original Silhouette Coefficient	Projection silhouette Coefficient
LSP	Default Parameters	LSP Projection for News_from_Lecture dataset is quite poor.The projection silhouette coefficient drops to negative value. Data points are not properly segregated.	0.0307	-0.1416
	Changed Parameters	Changed Parameters made the projection worse as there is further drop in silhouette coefficient to -0.366	0.0307	-0.366

Visualization of t-SNE with Default Parameters:

Figure XXXXXI

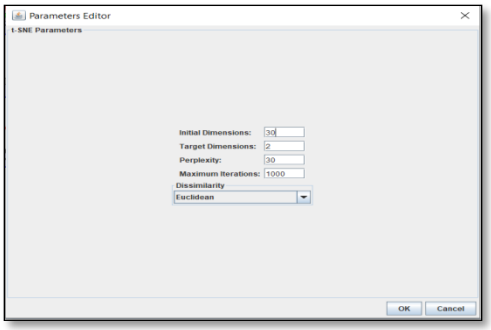
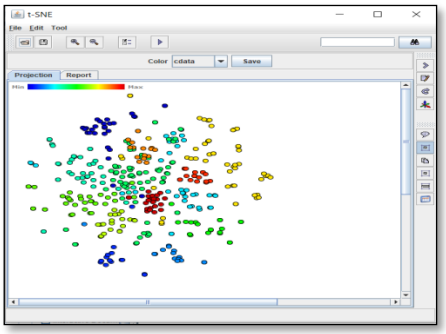
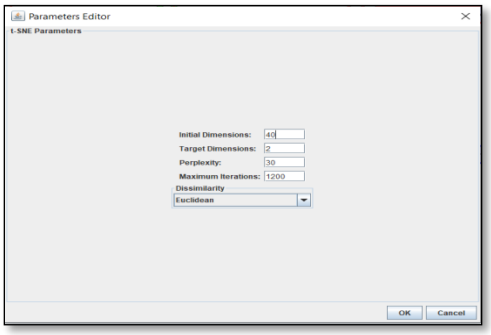


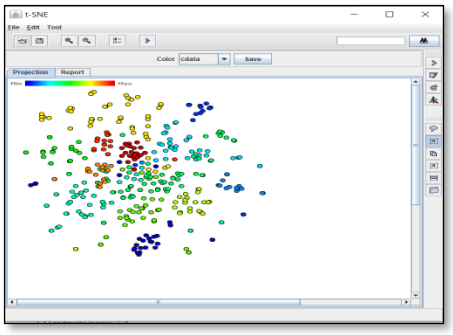
Figure XXXXXII



Visualization of t-SNE with Changed Parameters:

Figure XXXXXIII ***Figure XXXXXIV***





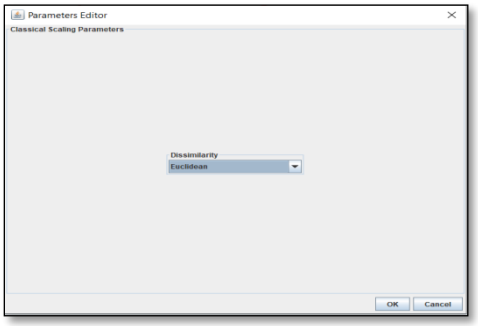
Please refer to Table: 11 for the comparative discussion of the above visualizations:

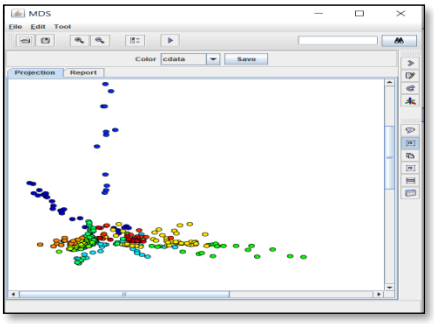
Table 11:

Projections	Parameters	Datasets News from Lecture	Original Silhouette Coefficient	Projection silhouette Coefficient
t-SNE	Default Parameters	Datapoints don't get clustered but similar coloured points are lying to together,so can see a mild classification along with overlapping in certain cases.Projection Silhouette coefficient shows greatervalue from the original value.So can say a good projection. Original coefficient itself shows a low value for t-sne	0.0307	0.104
	Changed Parameters	Projection deteriorates with the change in parameters as the silhoette coefficient drops for the new projection.	0.0307	0.0431

Visualization of MDS with Default Parameters:

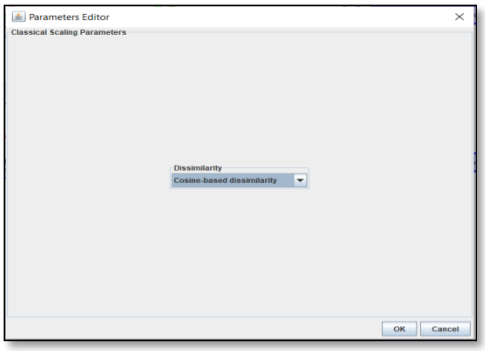
Figure XXXXXV ***Figure XXXXXVI***

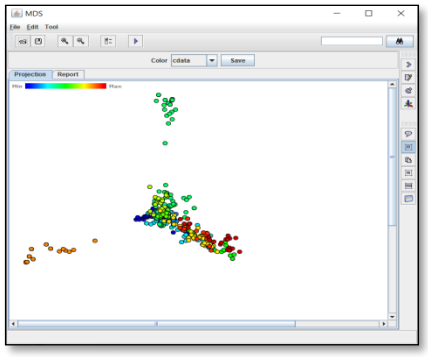




Visualization of MDS with Changed Parameters:

Figure XXXXXVII ***Figure XXXXXVIII***





Please refer to Table: 12 for the comparative discussion of the above visualizations:

Table 12:

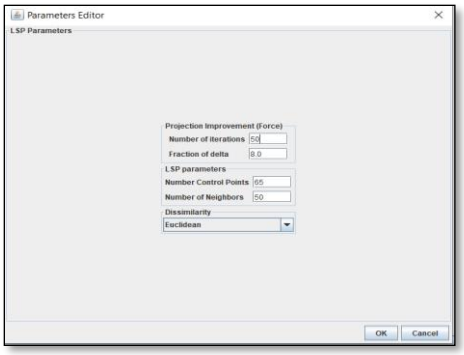
Projections	Parameters	Datasets News from Lecture	Original Silhouette Coefficient	Projection silhouette Coefficient
MDS	Default Parameters		0.0307	-0.111
	Changed Parameters		0.0307	-0.1104

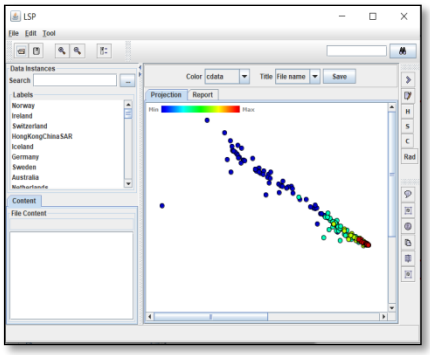
c) Dataset taken: HDR2020_transformed

Projections performed: Least Square Projection(LSP), t-Distributed Stochastic Neighbor Embedding (t-SNE) , ISOMAP

Visualization of LSP :

Figure XXXXXIX ***Figure XXXXXX***





The Silhouette coefficient for the original data is 0.17955607 whereas for the projected data it increases to 0.24507892 which implies that the data segregation has improved after performing LSP projection technique. So, LSP can be performed on HDR2020 data.

Visualization of t-SNE :

Figure XXXXXXI

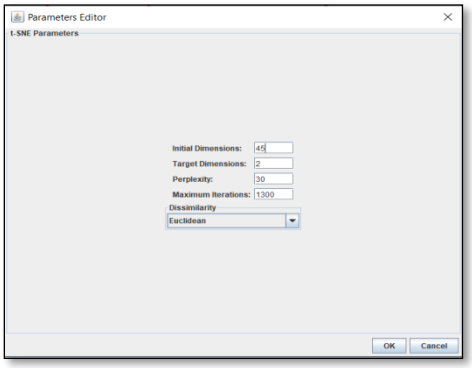
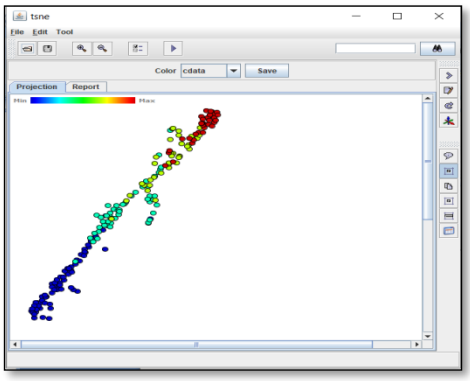


Figure XXXXXXII



The Silhouette coefficient for the original data is 0.17955607 whereas for the projected data it is increased to 0.36358464 which is more than double .Hence t-SNE performed really well for this dataset. So, t-SNE is highly recommended for this type of data set.

ISOMAP:

Figure XXXXXXIII

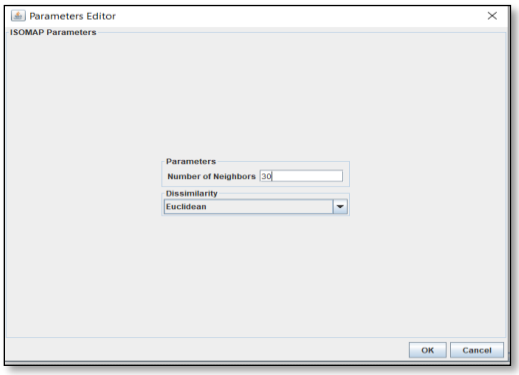
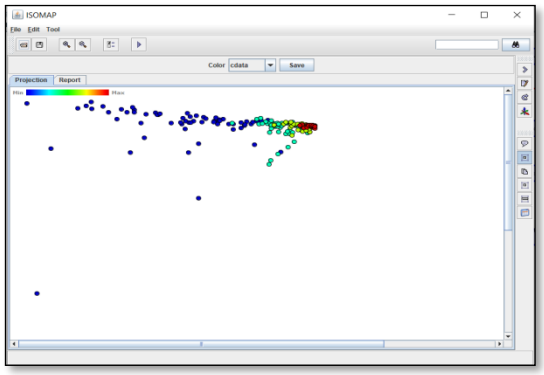


Figure XXXXXXII



The Silhouette coefficient for the original data is 0.17955607 whereas for the projected data it is increased to 0.17811044 which is less than the original value. This shows ISOMAP did not perform well for this dataset and is not recommended.

Conclusion

The first part of the report helped us to realize some interesting relationships among data and features of the various countries of the world which was not understandable in the tabular format. Our visualization task through Tableau helped to find out some hidden patterns and relationships in the dataset in colorful graphical format.

I reached to a realization that not only gross national income but how important is lifestyle, health system, educational system and infrastructure are important for a country's wellbeing and development.

The second part of the report gives a deep understanding of how different projection techniques play different roles for different types of data. Same projection technique cannot be used for all types of data and the performance of the projection technique is determined by Silhouette coefficient, one of the significant parameters to determine a technique's accuracy and efficiency for a particular dataset.

As a conclusion, it can be said that further data visualization and projection techniques can help us in getting more useful patterns underlying in the data which will ultimately enhance quality of life and resources for an overall betterment. (ourworldindata.org; Damian Clarke, 2016)

Bibliography

ourworldindata.org. (n.d.). Retrieved from Human Development Index: https://ourworldindata.org/human-development-index

Programme, U. N. (2020). *Human Development Reports Nepal.* UNDP.

TOONDERS, J. (n.d.). Data Is the New Oil of the Digital Economy.

UNDP. *HDR 2020 Report Bhutan.* UNDP.

Wikipedia. *bhutan national human development.*

Wikipedia. (n.d.). *Human Development Report*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Human_Development_Report

William G. Bowen, T. A. (n.d.). JSTOR. *Educational Attainment and Labor Force Participation* .