UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Human Development Analysis through Interactive Data Visualization

Athila Quaresma Santos

Monograph - MBA in Artificial Intelligence and Big Data

ICMC USP

SÃO CARLOS

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

```
        Santos, Athila Quaresma
S237h      Human Development Analysis through Interactive
        Data Visualization / Athila Quaresma Santos;
        orientadora Rosane  Minghim. -- São Carlos, 2022.
           60 p.


           Trabalho de conclusão de curso (MBA em
        Inteligência Artificial e Big Data) -- Instituto de
        Ciências Matemáticas e de Computação, Universidade
        de São Paulo, 2022.


           1. Human Developed Index. 2.  World Bank. 3.
        Data Visualization. 4. Interactive. 5. Python Dash.
        I.  Minghim, Rosane, orient. II. Título.
```

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Athila Quaresma Santos

Human Development Analysis through Interactive Data Visualization

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence and Big Data

Advisor: Profa. Rosane Minghim

Original version

São Carlos

2022

ABSTRACT

Santos, A. Q. **Human Development Analysis through Interactive Data Visualization**. 2022. 58p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

Data visualization has been gaining momentum over the last years. The global expected market revenue for 2023 is expected to be 7.76 billion U.S. dollars. On the other hand, the amount of available data in different applications are increasing exponentially. Nevertheless, there is still lack of generic and interactive platforms with user friendly interface that uses open-source frameworks and are able to process high-dimension and high volume of data. As an alternative, we purpose a data visualization solution to study the world bank information with focus on the Human Development datasets to provide analysis of future trends and comparison capabilities. A data visualization platform based on the open-source framework Python-Dash was developed. The platform is accessible through a web interface providing easy access, connectivity, environment independence, and no need for software or hardware installation on the client side. The flexibility is provided by multiple functionalities that allows easy to select features and components. Several visualization techniques are provided based on non-supervised learning, such as correlation, scatter matrix and the 2D projection using the t-Distributed Stochastic Neighbor Embedding algorithm. The results show an easy to use and rich knowledge extraction platform. We hope that the proposed solution helps stakeholders in decision making process while providing flexibility to implement data visualization techniques applied to human development applications.

Keywords: Human Developed Index. World Bank. Data Visualization. Interactive. Python Dash.

RESUMO

Santos, A. Q. **Human Development Analysis through Interactive Data Visualization**. 2022. 58p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

A visualização de dados vem ganhando força nos últimos anos. A receita global esperada para 2023 deverá ser de $7.76 bilhões. Por outro lado, a quantidade de dados disponíveis em diferentes aplicações está aumentando exponencialmente. No entanto, ainda faltam plataformas genéricas e interativas com interfaces amigáveis e que utilizem frameworks open-source e sejam capazes de processar dados de alta dimensão e volume. Como alternativa, é proposto uma solução de visualização de dados para estudar as informações do banco mundial com foco nos conjuntos de dados de Desenvolvimento Humano para fornecer análise de tendências futuras e recursos de comparação. Foi desenvolvida uma plataforma de visualização de dados baseada no framework open-source Python-Dash. A plataforma é acessível através de uma interface web proporcionando fácil acesso, conectividade, independência de ambiente e sem necessidade de instalação de software ou hardware no lado do cliente. A flexibilidade é fornecida por múltiplas funcionalidades que permitem a fácil seleção de recursos e componentes. Diversas técnicas de visualização são fornecidas com base no aprendizado não supervisionado, como correlação, matriz de dispersão e projeção 2D usando o algoritmo t-Distributed Stochastic Neighbor Embedding. Os resultados mostram uma plataforma de extração de conhecimento rica e fácil de usar. Espera-se que a solução proposta ajude *stakeholders* no processo de tomada de decisão, ao mesmo tempo que proporcione flexibilidade para implementar técnicas de visualização de dados em aplicações de desenvolvimento humano.

Palavras-chave: Índice de Desenvolvimento Humano. Banco Mundial. Visualização de dados. Interativo. Python Dash.

LIST OF FIGURES

1 INTRODUCTION

Assessing the development of a country is a non-trivial task that depends on several parameters. Before 1990, the most accepted indicator to measure how well a country performs was the Gross Domestic Product (GDP) (KUZNETS, 1934). However, it was clear over time that this quantification alone could not provide enough information for a comprehensive development measurement (RESCE, 2021). Instead, multidimensional indicators were proposed in order to consider social, economic, environmental, and well-being factors (COSTANZA, 2015). Among them, the Human Development Index (HDI) proposed by the United Nations Development Program (UNDP) in 1990 is the most known and used by governments, NGOs and researchers (STANTON, 2007).

The HDI is based on the aggregation of three other indexes: a) Life expectancy index, b) Education index, and c) GNI index. As with other composite indicators, the HDI suffers from common issues, such as the cardinality of the aggregation role (GRECO *et al.*, 2019) or the link between development and resources (RESCE, 2021; ARROW *et al.*, 2004). To mitigate these issues, the UNDP lunches every year its Human Development Report, providing a wide range of well-being indicators spread over several country members (Pedro Conceição, 2020). This allows the analysis of different criteria and the development of new coefficients or indicators focused on specific subjects.

This research project will use data visualization techniques applied to the data provided by the Human Development Reports available publicly by the UNDP. The data comprises yearly information of more than 150 indicators spread over 189 countries since 1990 (Pedro Conceição, 2020). The developed visualizer should extract country level development information and help users to understand trends, outliers, and hidden patterns. Furthermore, it is expected that the proposed solution supports the different analyses of classification and regression commonly performed in the field through machine learning algorithms.

1.1 Motivation

In order to advance human well-being and to provide governmental policies, the growth of a country needs to be monitored over the years using several measurements. The measurement of development can also help in the evaluation of previously applied strategies or supporting the decision when implementing foreign policies. However, the challenge is to provide correct indicators and the analysis of data composed of hundreds of parameters. Although extensive work has been conducted in the literature regarding the composition of indicators (COSTANZA *et al.*, 2014), more recently the advance of data visualization (MULTIPLES, 2019) and data analytic tools, such as the ones provided by machine learning (GUO *et al.*, 2021), produces new insights into how to handle data and how to explore new hidden patterns. In addition, the full potential of machine learning is hard to obtain without human guidance (SACHA *et al.*, 2017). Therefore, the use of visual analytics combined with human-centered guidance is expected to incorporate the knowledge, insight and feedback from the final user.

In this scenario, where modern techniques are being applied to the analysis of extensive data collection, together with data visualization tools, it is necessary to provide solutions aiming to improve readability and helping to analyze massive amounts of information for data-driven decisions (TABLEAU, 2018). Important information, such as human development parameters is essential to measure the progress of individual countries, regions, or on a world level, supporting future decisions not only economically, but also related to human welfare.

Given that one of the main challenges of composite metrics is to aggregate representative data without losing any important information and that important factors can not be neglected, new methods that provide better visibility, pattern recognition and classification should be explored. In this context, it is proposed the development of a graphical representation of the available data focusing on the analysis and extraction of useful information. It is expected that the obtained results can better represent multidimensional indexes, such as the HDI.

1.2 Research Questions and Objectives

In this work, several human development indicators for different countries and regions, over several years, will be compared to a new data visualization tool. Based on the challenges currently faced by multidimensional indicators, the following research question was developed and it will be used as a guide to this project:

Q1 “Can interactive data visualization techniques in conjunction with classification and clustering algorithms provide support decision making to world policies using Human Development features?”

Given this research question, the following objectives for the development of this work are defined:

- Map data visualization techniques provided in the literature, analyzing which ones are suitable for each data parameter provided by the UNDP. This objective is related to research question Q1.
- Specify the level of importance of each parameter and how they are co-related. This objective is related to research question Q1.
- Implement Machine Learning techniques for classification and clustering of the available data. This objective is related to research question Q1.

- Integrate human-centric decisions into the visual analytics tool to provide interactive machine learning resources.
- Compare the obtained results with the state of the art, especially based on several development indexes, such as the HDI as well as current visualizers for the specific dataset, such as the one in Google Public Data Explorer (GOOGLE, 2020). This objective is related to research question Q1.

2 VISUAL ANALYTICS

Visual Analytics (VA) is a term used to describe the integration of interactive data visualization and computational analysis algorithms to provide analytical reasoning. In practice is a crossroad of multidisciplinary fields to combine machine and human intelligence. VA enables the intuitive interpretation of data by synthesizing the intrinsic knowledge in a visual aspect (SHABDIN; YA'ACOB; SJARIF, 2020).

> "*The integral combination of human knowledge, intuition, and expertise with powerful computational algorithms is the inherent strength of visual analytics and allows analysts to steer and supervise the process of analyzing and exploring large complex data effectively (CAO et al., 2018).*"

Advancements and applications in the field of VA are increasing at a fast pace over the last years boosted by the availability of data. We have witnessed a profound digitization of society in the last decades, increasing the demand for data processing, transmission and storage. Web-based applications, social, and multimedia resources enhanced by the mobile, IoT, and wireless networks are the driving forces for this transformation. This facilitates fast access to data, search, and business process applications (CHO *et al.*, 2017).

2.1 Framework for Interactivity in Visual Analytics

Several frameworks address the dynamics of interactive VA. One of the first ones is proposed by (HEER; CARD; LANDAY, 2005) and is represented in Figure 1. The framework shows how the pipeline of information works. The source data needs to be filtered in order to be represented more concisely. The VA is represented by two main components: the visualization, showing the useful information; and the graphical user interface, with the interactive display (sliders, buttons, input boxes, etc).

Although simple, this framework is being adopted successfully and adapted to include analytics models and algorithms for machine learning applications. As an example, the filtering step could be implemented by interactive user control to help the data transformation phase by removing missing, redundant or inconsistent data. Another example on the visual mapping phase, users could interact with a graphical interface to help normalize, standardize or reduce the dimensionality of the dataset.

Complementary, the semantic interaction framework is proposed by (ENDERT; FIAUX; NORTH, 2012) and is shown in Figure 2. This framework shows the pipeline of interaction between machine learning algorithms and spatial metaphor information models. This approach uses semantic interaction to infer the user analytical thinking

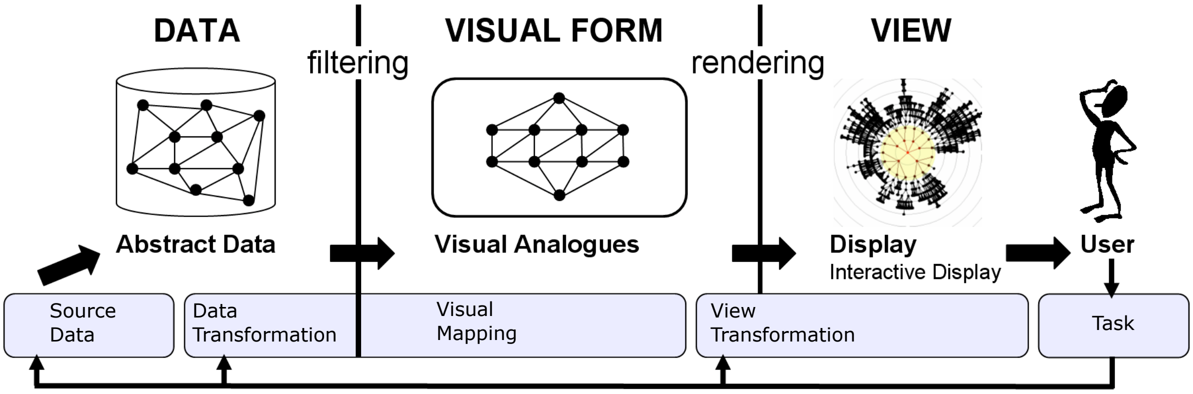


Figure 1 – The information visualization framework (Adapted from (HEER; CARD; LANDAY, 2005)).

via the graphical interface to steer the model behavior. It is an indirect co-reasoning abstraction between human interaction and intelligent machine processing capability. In this way, the user doesn't need to control the whole process directly, while providing interaction flexibility along the way.

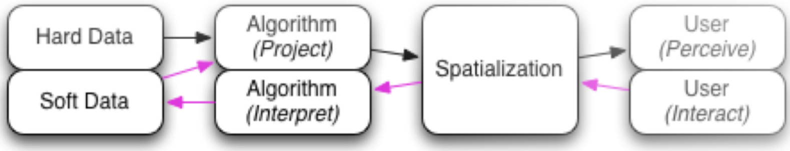


Figure 2 – Semantic interaction framework (Adapted from (ENDERT; FIAUX; NORTH, 2012)).

The spatial metaphor captures the user's interaction by translating the intentional modification of the data transformation step, represented by Figure 1, to provide coupling between human cognition and computation processing. This approach allows raw data transformation into spatial visualizing features.

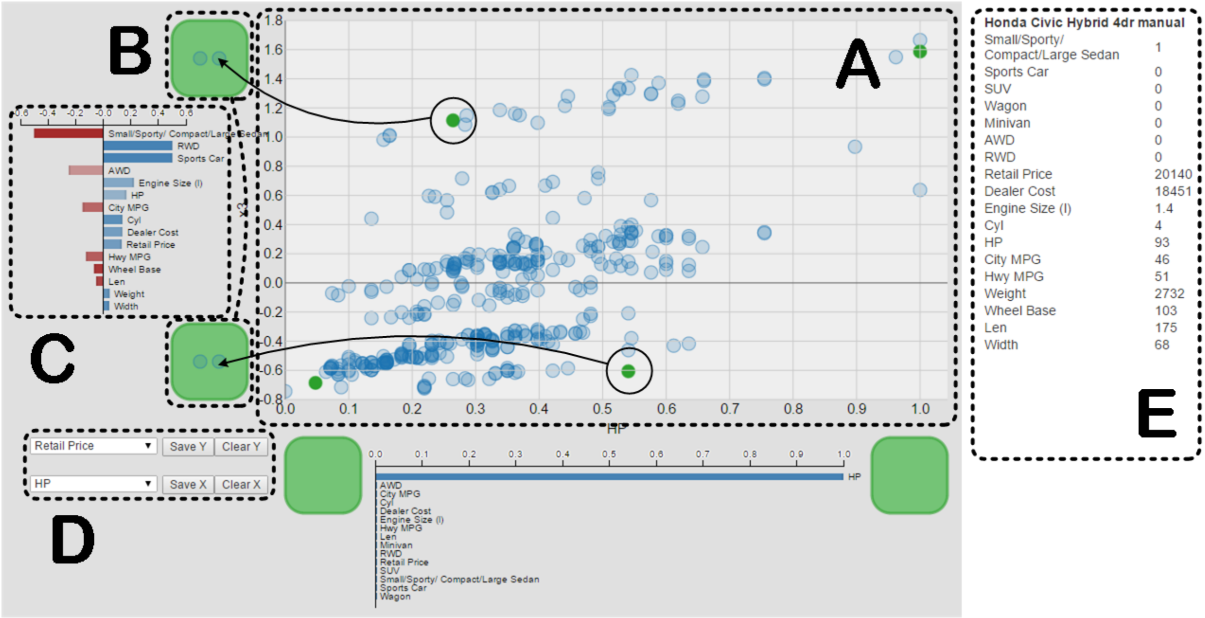
As an example, the centroid location of a k-Means algorithm used in clustering problems can be used as interactive elements provided by the users through a display visual interface. The expert knowledge from the user can be explored and inferred by the machine learning model as a visual metaphor, strengthening the bi-directional cognitive connection between the user and the spatial layout.

2.2 Dimension Reduction

Dimension reduction is a technique to reduce the space representation of input data while maintaining intrinsic meaningful information about the original data. This is important because high dimension spaces are often cognitively challenging to comprehend. Besides, data-centric applications are directly affected by the degree of dimension the input is represented (AILON; CHAZELLE, 2010). Several machine learning algorithms benefit from dimension reduction by reducing the size or complexity of the data, while maintaining intrinsic properties, such as pairwise distances (WU *et al.*, 2021).

There are two main dimension reduction classes of algorithms: 1) Data-aware and 2) data-oblivious. The former is represented by techniques that take advantage of prior information of the input. Some examples of implementation include the Principal-Component Analysis (PCA) (LASISI; ATTOH-OKINE, 2018) and the compressing sensing (GAO; SHI; CAETANO, 2012). The second class, data-oblivious, doesn't use the prior information of the data. Examples include sketches for data streams, locality sensitive hashing, and random linear mappings in Euclidean space (AILON; CHAZELLE, 2010).

Figure 3 shows an example of a visual analytics tool that uses scatterplots to represent multiple features in each axis in order to reduce the dimension of the dataset. The interface is composed of the main scatterplot area (A), the interaction panels (B, C, D), and the data detail view (E). The user can interactively select different features from the data set and group them on each axis of the scatterplot. Interaxis will calculate the weights to fit the regression function that better represents the set.



Honda Civic Hybrid 4dr manual

Small/Sporty/ Compact/Large Sedan	1
Sports Car	0
SUV	0
Wagon	0
Minivan	0
AWD	0
RWD	0
Retail Price	20140
Dealer Cost	18451
Engine Size (l)	1.4
Cyl	4
HP	93
City MPG	46
Hwy MPG	51
Weight	2732
Wheel Base	103
Len	175
Width	68

Figure 3 – InterAxis: A visual analytics tool to help interpret, define and change regression parameters (KIM *et al.*, 2016).

The PCA algorithm is one of the most used unsupervised algorithms for dimension reduction. It is a very traditional technique developed in 1933. It seeks to maximize variance and preserves large pairwise distances. There are some limitation from the PCA algorithm, which is outside the scope of this work. Another very spread algorithm is the t-Distributed Stochastic Neighbor Embedding (t-SNE) (MAATEN; HINTON, 2008). Contrary to the PCA, the t-SNE algorithm tries to maintain information about small pairwise distances. The t-SNE algorithm was designed to process high-dimensional data and provides a map containing structure of different scales.

2.3 Clustering

Clustering methods are responsible to group data entities with similar characteristics or features using unsupervised training (LI *et al.*, 2018a). It is applied to several knowledge fields whose data-driven approach is the center of the analysis. Usually, the clustering algorithms can be classified into four main classes: 1) Hierarchical, 2) partitional, 3) density-based and 4) grid-based (NAGPAL; JATAIN; GAUR, 2013).

According to (KWON *et al.*, 2018), although a large variety of clustering algorithms exist, usually they can be classified into five large categories:

- Centroid-based methods: Algorithms based on distance metric calculation. It is necessary to define a priori the number of clusters. Examples are the k-means (WANG; SUN; BAO, 2020) and Fuzzy c-mean (BLöMER; BRAUER; BUJNA, 2020) algorithms.

- Connectivity-based methods: Algorithms based on distance metric calculation and the linkage criterion for splitting and joining the aggregated clusters. Examples are the hierarchical (COHEN-ADDAD *et al.*, 2019) and agglomerate (TOKUDA; COMIN; COSTA, 2022) algorithms.

- Density based methods: Algorithms based on density calculation. The clusters are partitioned based on their density parameters. Examples are the DBSCAN (HU *et al.*, 2021) and OPTICS (ZHAO *et al.*, 2022) algorithms.

- Low Dimensional Embeddings: These algorithms need to define a priori the number of clusters and the number of dimensions that will be used to project the data attributes. As an example, there is the Spectral Clustering (BERAHMAND *et al.*, 2021) algorithm.

- Probabilistic clustering methods: Algorithms based on the probability distribution of the data hyperparameters to group the data among different clusters. Examples are the Gaussian Mixture Models (PATEL; KUSHWAHA, 2020) and the Latent Dirichlet Allocation (GROPP *et al.*, 2019).

Figure 4 shows an example of an interactive visual analytics tool. The solution allows easy parameter setting and cluster technique selection, allowing the user to integrate their own expertise by defining task-relevant constraints in the data. The graphical interface shows different perspectives of the data clustering together with the setting input parameter for optimization.

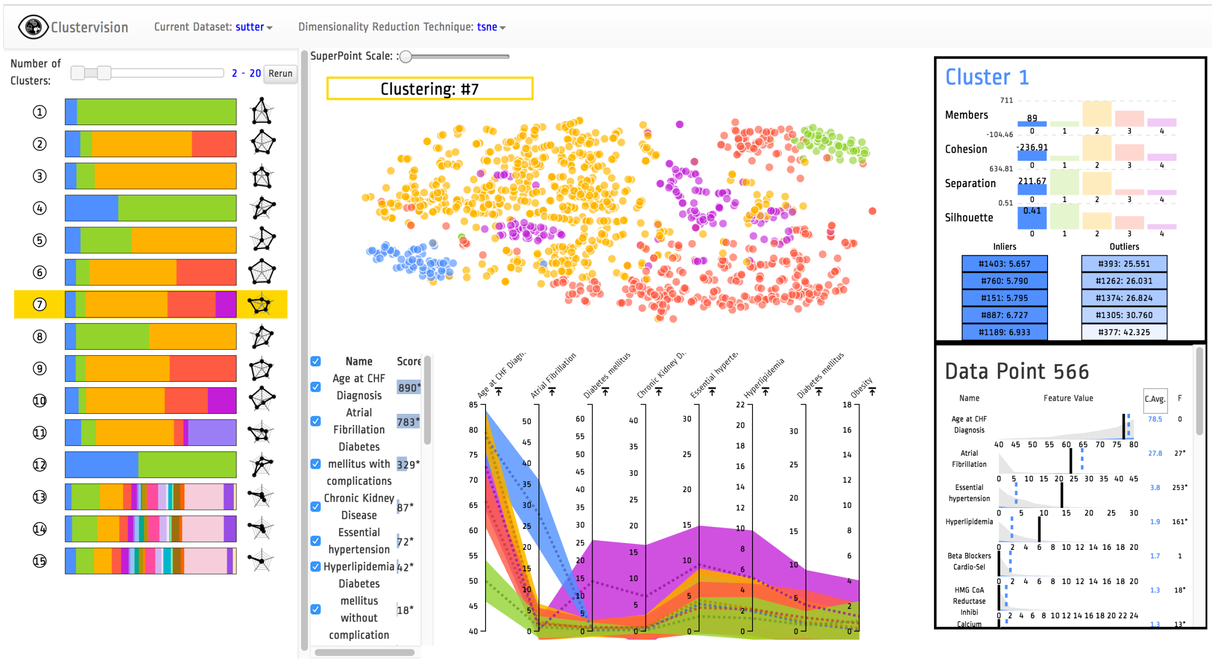


Figure 4 – Clustervision: An interactive visual analytics tool to help optimize clustering techniques and parameter selection (KWON *et al.*, 2018).

2.4 Classification

Classification is the process of assigning a categorical label y to given input instances x. The input is generally a multi-dimension attribute list grouped by a common sample. Therefore, the classification problem is to find a relation function f that maps each instance to its own category, i.e., $f(x_1, x_2) = y$ (BERNARD *et al.*, 2018). Because the labeling process of assigning a category class to each training data in advance, this process is usually referred to as a supervised learning method.

According to (ALI *et al.*, 2019), the most widely used algorithms for classification are listed below:

- K-nearest neighbors (k-NN): It is classified as a lazy learning method because no real learning is achieved during the training phase. It usually memorizes the entire training dataset.
- Decision tree (DT) and random forest: Algorithms that divide the dataset into branches. The root and internal nodes divide the dataset into several categories and represent decisions. The end nodes are leaves that represent the classes of the

classification process. Random forests are an ensemble of decision trees that needs to be ranked to obtain the best classifier.

- Support vector machines: It is a supervised learning algorithm that aims to find a class identity by separating the features of a dataset using hyperplanes. The decision surfaces use linear or non-linear kernel functions to separates positive and negative examples.

- Artificial neural networks (ANN): Algorithms based on the human brain physiology and learning process. The artificial neurons are organized in layers connected by weighted edges. The multi-layered perceptron is the simplest but most effective ANN architecture.

An example of a visual analytics tool for classification is represented in Figure 5. The interface allows users to interact with the classification process using different highlighting control and filters for search, annotators and category selection. The solution calculates statistical artifacts, such as the average change agreement over Cohen's kappa values.

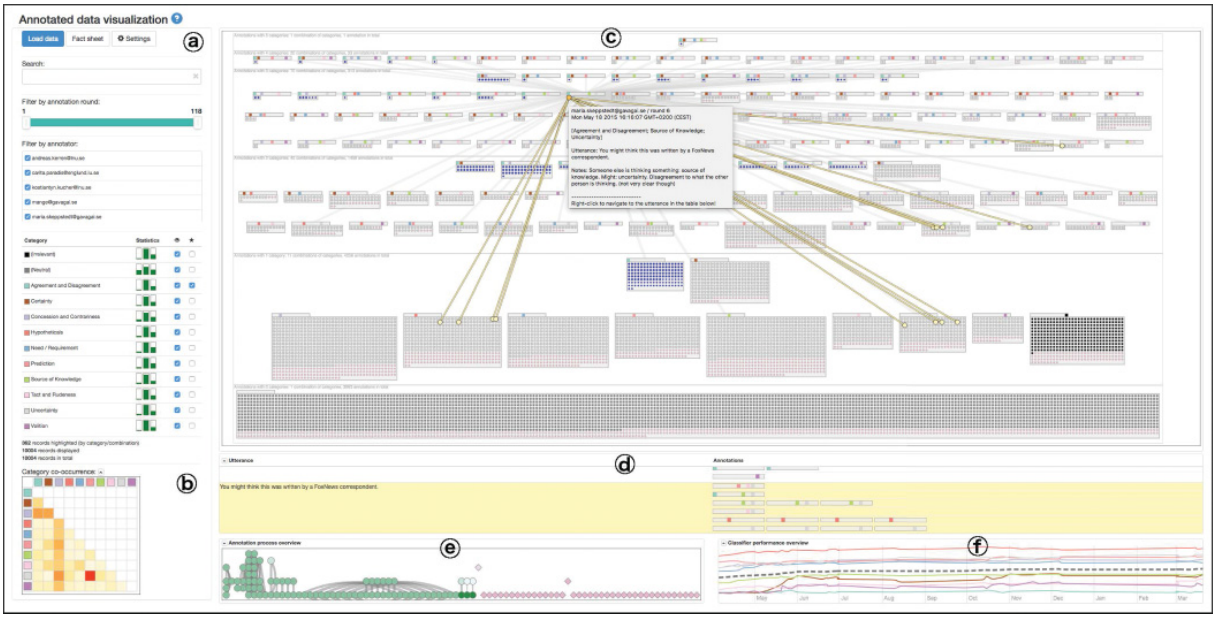


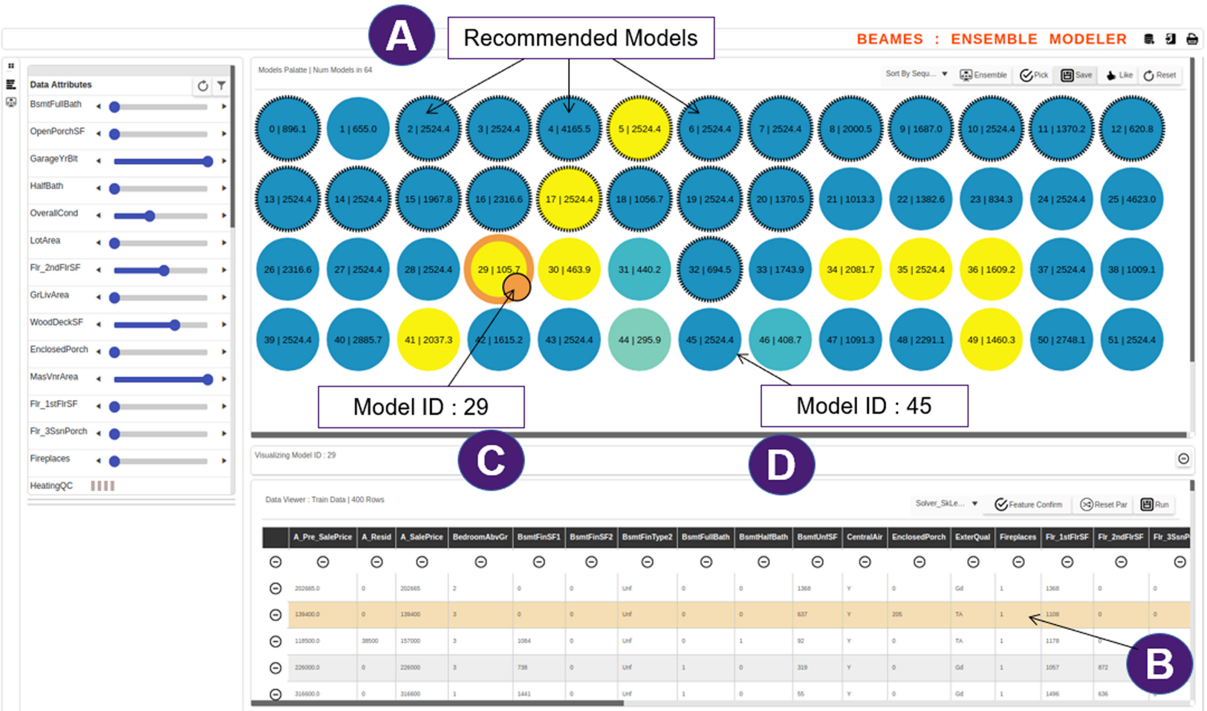
Figure 5 – ALVA: visual analytics tool to help training stance classifiers (KUCHER *et al.*, 2017).

2.5 Regression

Regression is a form of numerical labeling, where an algorithm tries to find a function f that maps an input x to a numerical label y, i.e., $f(x) = y$. Regression algorithms identify the multivariate relations within data features and ultimately find causal implications (ENDERT *et al.*, 2017). The most common regression algorithms are listed below:

- Linear Regression: The simplest form of regression that uses statistical calculations to find a linear function that represents the relationship between two continuous (quantitative) variables.
- Lasso Regression: Least Absolute Selection Shrinkage Operator (LASSO). The LASSO algorithm estimate the best set of predictors that minimize the prediction error. The shrinkage relates to the constraints for some variables to reduce their values.
- Logistic regression: Algorithm whose output is a binary result (true or false). It is used on applications such as fraud detection, credit card scoring, etc.
- Multivariate Regression: Algorithm aimed to solve linear correlations for problems with more than one independent variable (predictors) and more than one dependent variable (responses).
- Others: There are several algorithms previously discussed that are also employed to regression. For example, decision trees, random forest, support vector machine, k-NN, and neural networks.

Figure 6 shows an example of visual analytics tool designed for regression tasks. In the graphical interface, each glyph represents a regression model highlighted by different colors. The table at the bottom lists the training, test and application datasets. Several control panels are available for filtering and parameter settings.



	A_Pre_SalePrice	A_Resid	A_SalePrice	BedroomAbvGr	BsmtFinSF1	BsmtFinSF2	BsmtFinType2	BsmtFullBath	BsmtHalfBath	BsmtUnfSF	CentralAir	EnclosedPorch	ExterQual	Fireplaces	Flr_1stFlrSF	Flr_2ndFlrSF	Flr_3SsnP
⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖
⊖	202665.0	0	202665	2	0	0	Unf	0	0	1368	Y	0	Gd	1	1368	0	0
⊖	139400.0	0	139400	3	0	0	Unf	0	0	637	Y	205	TA	1	1108	0	0
⊖	118500.0	38500	157000	3	1084	0	Unf	0	1	92	Y	0	TA	1	1178	0	[illegible]
⊖	226000.0	0	226000	3	738	0	Unf	1	0	319	Y	0	Gd	1	1057	872	[illegible]
⊖	316600.0	0	316600	1	1441	0	Unf	1	0	55	Y	0	Gd	1	1496	636	[illegible]

Figure 6 – BEAMES: visual analytics tool for multi-model steering, selection, and inspection for regression tasks (DAS *et al.*, 2019).

3 STATE OF THE ART: VISUAL ANALYTICS

3.1 Dimension Reduction

(JOHANSSON; JOHANSSON, 2009) focus on the importance to preserve significant structures in the original data before applying dimensionality reduction. They propose a solution that takes into consideration the user-defined quality metrics (weights), highlighting that the identification of certain structures is task-dependent. The solution provides a visualization tool capable to explore certain structures within large multivariate data sets and demonstrated through World Bank public database case study. Later the authors use the same methodology for visual exploration and interactive analysis of microbiome populations (FERNSTAD *et al.*, 2011).

Another work to merge the concepts of visual analytics and dimension reduction is presented by (FUCHS; WASER; GROLLER, 2009). The authors are interested in investigating how personal the selection of specific parameters in an explanatory hypothesis formulation works and how this can affect the final solution of a multidimensional data problem. They propose an interactive cycle where the user controls a heuristic search algorithm to expand the search space. Properties and general information are visually represented to guide the user throughout the best generated hypotheses. A case study using data from the automotive field is used.

(STAHNKE *et al.*, 2016) implemented the concept of probing data projection to reduce multi high-dimension space information into a planar layout. They developed a web application that helps users to interact with projection techniques for comparison and result analysis, such as approximation errors, multi-level comparisons, spatial orientation and consistent design. The study is evaluated using a dataset related to OECD countries' information. They conclude by demonstrating how probing can help the visualization community to increase accuracy on error-prone dimensionality reductions.

(KIM *et al.*, 2016) presented InterAxis: a user-driven VA tool to define and change axis properties in scatter plots. The system computes a linear combination of data attributes to determine how each point should be represented in the x and y Cartesian coordinates. Later, the user can manually tune the data parameters based on the visualization. Two case studies are presented: the Car dataset and the Crime dataset. The results demonstrate the effectiveness of the proposed interaction techniques, while highlighting some limitations, such as the consideration of only linear models and the difficulty to handle sparse data. The same authors expanded a similar idea by proposing the AxiSketcher: A non-linear mapping axis tool capable of computing users' drawings over data entries (KWON *et al.*, 2017). The solution provides a simple way to explore the user's domain knowledge over high-dimensional data by calculating the new reduced representation of multidimensional attributes using non-linear functions.

3.2 Clustering

In VIS-KT (NOURASHRAFEDDIN *et al.*, 2018), a framework for interactive text-clustering is developed. The lexical double cluster algorithm is used a foundation for text clustering. Then, three unsupervised interactive versions are proposed: key term labeling, document labeling and hybrid labeling. The key term labeling algorithm presented better results and therefore was used as the core solution of a visualization tool. To evaluate the proposed solution, two case studies are used: NewsSeparate and BBC Sport. The visual interface is also evaluated from a user perspective. The obtained results and the user acceptance were widely positive.

(KWON *et al.*, 2018) built Clustervision as a generic tool to support data clustering using visual supervision. Clustervision helps the users to compare different cluster techniques and evaluate new constraints, according to the available multi-dimension data parameters. In the back-end, the tool runs a variety of clustering techniques and parameters for evaluation. The result is then ranked based on five quality metrics: Calinski-Harabaz index, Silhouette Coefficient, Davies-Boulding index, Gap Statistic and S_{Dbw}. To assess the proposed solution, a case study is applied in the medical domain to find clusters of patients with heart failure using historical records. The obtained results helped to improve the grouping of patients, showing the applicability of the proposed solution.

An impact assessment regarding outliers in k-Means and k-Medoids clustering methods is performed by (KANIKA *et al.*, 2019). The authors developed a web application to provide interactive Visual Analytics for better user insight. The idea is to alow the users to remove the outlier data points from the datasets. The known Iris dataset was used as a case study. The results show how the selection of attributes can interfere with the accuracy of the final clustering configuration.

In their work, (SHERKAT; MILIOS; MINGHIM, 2019) proposed a novel Visual Analytics for interactive document clustering. The system supports the user interaction by allowing the change of key terms related to their knowledge of the document's domain. Document cloud and temporal view are used to help users to interact with the clustering algorithm. The t-distributed stochastic neighbor embedding is combined with the Force-directed placement method to improve the distinctions among clusters. A real case study is evaluated using decision-making decisions on email document discussions. The results show proper precision, faster convergence and effective decision making.

An open-source and parameter-free method called Interactive Projection-Based Clustering (IPBC) in Visual Analytics is presented by (THRUN; PAPE; ULTSCH, 2020).

The solution provides interactive visualization resources to cluster high-dimension data and it is based on a generalized Umatrix of an arbitrary projection method. A topographic map is constructed on top of the generated scatterplot. IPBC is compared with some public available Visual Analytics tools (iPCA, Clustrophile 2, Clustervision and VISTA), using real-world experiments: Chainlink, Hepta, Tetragonula and SCADI. IPBC outperformed the compared tools in four of the analyzed cases.

3.3 Classification

In TransXplorer by (MA *et al.*, 2017), the authors developed a transfer learning process to apply the knowledge from source tasks to target tasks in the context of classification. The solution is an interactive visual system capable of providing exploratory knowledge transfer among the tasks, while allowing the incorporation of the user's experience and expertise. The visual tool is presented as a suite of visual communication for interaction that applies the proposed transfer learning methodology. Two case studies were demonstrated: the Amazon product reviews dataset and the twenty newsgroups dataset, showing the validity and efficiency of the approach.

A visual interactive labeling (VIAL) is proposed by (BERNARD *et al.*, 2018). The system combines the power of visual analytics and visual interactive techniques to support the labeling of machine learning tasks. Labeling is addressed as the common goal of assigning a label to data input. Therefore the classification task is done when having a categorical labeling process. VIAL consists of six steps: pre-processing and feature extraction, learning model, result visualization, candidate suggestion, labeling interface, and feedback interpretation. The validation is done using two real world case studies: visual interactive labeling for video classification and visual interactive labeling for similarity modeling applied for soccer players.

An important selection criterion on Machine Learning algorithms is to choose the right performance indicator. In their work, (BRZEZINSKI *et al.*, 2018) developed a classification performance measurement supported by a specialized visualization technique applicable to class imbalanced problems. The proposed interactive online visualization tool compares general properties of classification metrics using a barycentric coordinate system represented by a 3D tetrahedron. The solution compares 22 common classifier performance measures, showing the similarities and contrasts of non-parametric, internally parametric and externally parametric measurements. The results show how imbalance classification classes affect the range of indicator values and how an interactive visualization tool can help compare them.

In ALVA by (KUCHER *et al.*, 2017), a VA was developed to support users towards the exhaustive process of manual annotation. The platform focus on multi-label text classification task using machine learning and natural language process. The visual representation of multidimensional data is grouped by individual annotation items by combining label categories. The authors call this approach CatCombos (Category Combination) to help the users with the annotation and training process while improving the understanding of the different categories. A case study of notional stance categories in linguistics is evaluated with good potential for social media monitoring.

(ALI *et al.*, 2019) presented a literature review on clustering and classification for time series data in visual analytics. The authors classify over sixty papers from six perspectives: 1) data structure 2) feature and similarity based, 3) analysis techniques, 4) visualization techniques, 5) visualization tasks and interaction methods, 6) evaluation. Their analysis indicated more work related to clustering than classification. Probably because the lack of label data. However, classification papers were listed only until 2015. The survey can be used as a starting point to a future direction of visual analytics for data mining techniques.

3.4 Regression

In their work, (LI *et al.*, 2017) developed a quality evaluation method for graph layout algorithms. The method takes into consideration the user interaction by providing several graph layouts to be rated. A regression model is proposed based on readability metrics of each layout and the subjective score of the users. The regression target provides the overall quality score of a graph layout. The accuracy prediction is compared with several regression algorithms: Linear Regression, Linear kernel, Gaussian kernel, Polynomial kernel of second order and Polynomial kernel of third order. The results indicate the support vector regression of linear kernel function as more effective.

RegressionExplorer was developed by (DINGEN *et al.*, 2019) to improve interactive exploration of regression models applied to the field of clinical biostatistics. Patient data is explored through different visual models, where subsets of data can be dynamically selected. The proposed visualization method is a parallel coordinate plot system implemented as a matrix icicle plot representation. The tool supports decision making or research base formulation of hypotheses. The method is applied to two real world case studies: hypernametria and cardiac conduction disorder on critically ill patients. The data was analyzed by domain experts. Results show faster insights because RegressionExplorer provides the comparison of various model properties and covariate effects.

In order to provide user interaction on regression models and big data mining, (LI *et al.*, 2018b) developed a visual analytics tool capable of supporting feedback learning provided by field experts during the model building process. The feedback is done by allowing data selection, data labeling and data correction. The preliminary results indicate that the feedback system reduces the amount of data required to train the machine learning models. The evaluation is conducted on two case studies. The first is related to data classification, but the second is reserved for the human cognitive score prediction using a regression method. The results are promising, showing faster convergence of the result when compared with random sampling methods.

(DAS *et al.*, 2019) are the authors responsible for BEAMES: A prototype visual analytics tool that provides interactive model steering, selection and inspection of regression models. The expert users can inspect multiple learning algorithm models in order to perform accurate predictions over the analyzed dataset. The applied method consists of three interactive steps: 1) weighting of critical data instances, 2) weight feature selection, 3) model selection, and 4) building model ensembles. A case study to predict future housing pricing is used, showing the flexibility of the user interface. The same authors expanded the idea by developing a new visual analytics tool called LEGION (DAS; ENDERT, 2020) aiming to help users compare attributes and similarities among multiple regression models. The tool provides two modeling strategies: 1) tuning of hyperparameters by using TensorFlow and HP JS libraries; 2) feature engineering by implementing a boosted ridge regression model. The efficiency of these methods is evaluated using two case studies. The first is a death rate prediction that uses cancer mortality rate data from the US. The second is a stock price prediction based on the US stock price dataset. The tool provided good feedback and support decision capabilities on the selection of different trade-offs among several parameters.

4 METHODOLOGY OF VISUAL ANALYTICS APPLIED TO HUMAN DEVELOPMENT INDICATORS

The methodology presented in this chapter will describe the steps, methods and resources necessary to design a visual analytics tool with interactive functionalities, capable of analysing human development metrics and indices. Based on the challenges faced by multidimensional indicators and how independent and dependent variables are presented, this work proposes the processing of raw data and the use of visual analytics techniques to provide a more accurate picture of the importance of each parameter and how they are co-related.

4.1 Methodology pipeline

The methodology structure pipeline and organization is illustrated by Figure 7. Section 4.2 explain how the human development data was obtained. It also explain how the data is organized and the attributes, parameters and measurements that are available, according to the database provided by the United Nations Development Programme. Special attention is given to the calculation of common indicator in this field. More information regarding the database schema and the time series organized by year is available on Appendix A. In section 4.3, an overview is presented, showing how the data was manipulated by describing the processes of extraction, integration, transformation, cleaning and the selection of the most relevant features. Section 4.4 shows how the interactive visual analytics method was used and it is the main focus of this work. Finally, Section 4.5 is responsible to describe the process of transforming data values into information and, most desirable, the knowledge that can be used for experts and stakeholders.

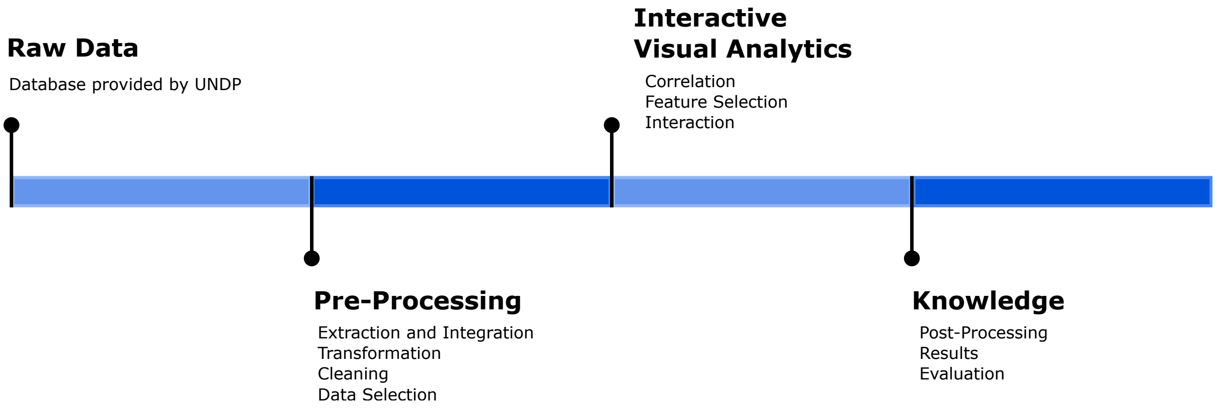


Figure 7 – Methodology pipeline.

4.2 Raw Data

All data is available on the United Nations Development Program website ((NATIONS, 2022)). Regularly, UNDP publishes its “Human Development Report”, which is a deep analysis of trends and projections for human transformation and planetary challenges, such as climate crisis, biodiversity, ocean conditions, etc. Over time, there is a gathering of country level information about indicators that measures their progress and difficulties.

The Human Development Index (HDI) is composed by tree dimensions: long and healthy life, knowledge and decent standard of living. The idea behind the dimensions is to aggregate a compound index that take into consideration different human perspectives not only economical. The HDI is the geometric mean of normalized indices for each of the three dimensions as illustrated in Figure 8.

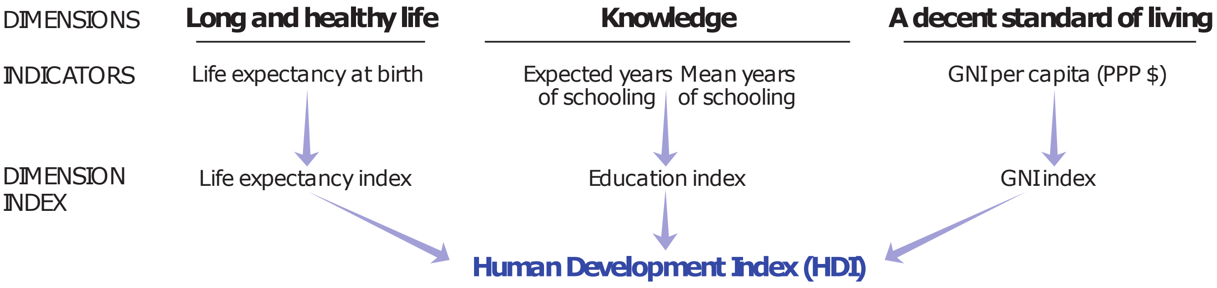


Figure 8 – Calculating the human development index(Adapted from (NATIONS, 2022)).

The dimension values are normalized between 0 and 1, according to Table 1. The minimum values acts like “natural zeros” while the maximum values are “projected targets”. The minimum value for life expectancy being 20 refers to the fact that no member country had a value lower then this limit in the 20_{th} century. A maximum life expectancy of 85 is considered a realistic target for many countries nowadays. The maximum expected years of schooling as 18 is the equivalent of achieving a master’s degree. The Gross National Income (GNI) measures the total amound of money earned by a nation and is measured by Purchasing Power Parity (PPP). The PPP unit is metric to compare economic productivity among different countries.

Table 1 – Human Development Index: dimension indices normalization

Dimension	Indicator	Minimum	Maximum
Health	Life expectancy (years)	20	85
Education	Expected years of schooling (years)	0	18
	Mean years of schooling (years)	0	15
Standard of living	GNI per capita (2017 PPP$)	100	75,000

Table 2 – Human Development Indices

	Acronym meaning	**Description**
HDI	Human Development Index	Tracks progress made by countries in improving the lives of people
GDI	Gender Development Index	Tracks gender gaps in human development achievements
IHDI	Inequality-adjusted Human Development Index	Tracks inequality across population based on HDI
GII	Gender Inequality Index	Tracks human development costs of gender inequality
PHDI	Planetary pressures adjusted Human Development Index	Tracks planetary pressures in the Anthropocene based on HDI.

The indices are then calculated by Equation 4.1. For the education dimension, an arithmetic mean is calculated between the expected years of schooling and the mean years of schooling.

$$I_{dimension} = \frac{actual\ value - minimum\ value}{maximum\ value - minimum\ value} \tag{4.1}$$

The HDI is calculated as the geometric mean of the three aforementioned dimensions as stated in Equation 4.2. If HDI needs to be aggregated by different demographic parameters, a weight needs to be taken into consideration depending on each case. For example, the life expectancy and GNI per capita can be weighted by total population or expected years of schooling can be weighted by ages 5-24.

$$HDI = \sqrt[3]{I_{Health} * I_{Education} * I_{Income}} \tag{4.2}$$

To provided fairness and inequality information, the United Nation Development Program provides other indices as shown in Table 2. Indices like the IHDI tries to correct inequalities across different demographic populations while the GDI and GII takes into consideration the impact and inequality of gender. A full list of the parameters and year availability of the data is provided in Appendix A.

4.3 Pre-processing

4.3.1 Extraction and Integration

The Human Development Report Office provides access to human development related data through two formats. The first is a database composed by CSV files comprising information about 195 countries spread over the globe. There is a different file for each of the indices provided by UNDP (HDI, GDI, IHDI, GII and PHDI). The data is categorized by year with the oldest entry being 1990 and the last compilation being 2019. The data is also categorized by each dimension that composes the final index. The data is also available through a REST API where data is in JSON format.The data can be queried by indicator id(s), year(s) and country code(s) and group by any order. Therefore, this work will explore the relationships and importance of each dimension. The Table schema of the database as well as the time series by year information for each country can be seen on Appendix A.

Using the CSV files, it is necessary to gather information about metric and dimensions in different documents. This is specially true when we need to compare data from different indices. In the same way, the API provides parameters for selection of specific data. However, if data from different index parameters is needed, we have to request two or more reading operations from the server.

4.3.2 Transformation

Transformation was not a big concern for this work because the necessary data was already collected and transformed in a common database by UNDP. Therefore the data source is available accessing an unique database that is already standardized in the same format for all indices. On the other hand, the database was not normalized on the parameter "year". The tables had one column for each year, creating a separate time series by the date of the index calculations. Therefore, a transformation was done in order to plot the data over a defined period.

4.3.3 Cleaning

The data cleaning is a process to improve the overall quality of the data set. It is responsible to treat the data to avoid errors, lack of information and noise attenuation. The biggest issue in the data collection of this work is related to lack of some fields. Some countries, for example, was not registered in the Human Development Reports at the beginning of the registering (1990). Some of the countries have new denomination because of historical events and changes in territory borders. Therefore, there are several fields with lack of relevant information.

Another issue is the synchronization between indices and parameters. Some of the indices were proposed later and have few years of historical data. The data is also collected or calculated on different time steps ranging from yearly to five years span. Some of the data used by UNDP is not easily obtained. Therefore, some estimation is done in order to evaluate the potential measurement. This add some errors to the parameters of the system, affecting directly the accuracy of the results.

4.3.4 Data selection

Usually data selection is applied on the raw data to reduce memory use, disk space or processing time. There are several techniques to achieve this, including the reduction of input instances, removing dependable and redundant variables that have directly relation with other parameters or reducing the possible values of certain attributes.

The data provided by UNDP is not extensive enough to be categorized as big data or present issues with the current power processing and volatile and non-volatile storage resources. Therefore, all data is imported to the system and the data selection is done by the final user interactively using the provided interface described in Section 4.4.

4.4 Interactive Visual Analytics

Figure 9 illustrates the proposed Interface. The solution can be provided as a web service that can be accessed everywhere or an app can be developed with the same feature to be installed locally on several operation systems. The first option is preferable because no installation is required and more visibility for this work is expected. The interface is projected to provide an user-friendly experience with intuitive commands and easy to access features.

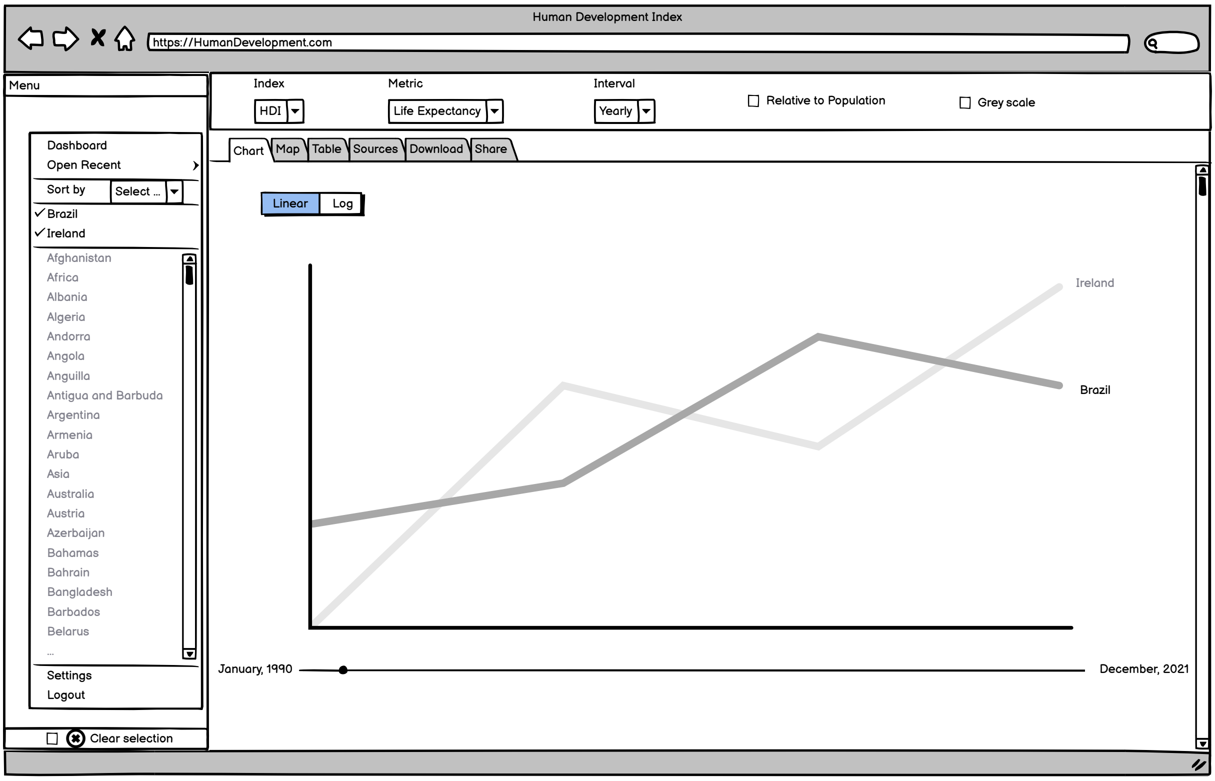


Figure 9 – Visual analytics interface.

The interface is divided in three canvas. The left vertical menu give access to the country selection based on available data provided by UNDP. It is possible to select several instances to map their respective metrics. The selection is done by clicking in the country of interested and can also be removed by the same way. A "Clear selection" option is also available, resetting the whole configuration to the initial state. The horizontal upper menu is responsible to provide the information about the group index information (HDI, GDI, IHDI, GII or PHDI), the specific dimension metric that will be used for comparison and the interval time stamp of the time series. The information can also be normalized by relative population, facilitating the comparison between different demographic scenarios. The option to plot in grey scale provides more accuracy when printing the final results in black and white. Finally, the center canvas highlights the visual aspects of the available analytics. It is possible to extract information using different representation, such as "Chart", "Map" and "Table". The option "Source" provides information of the reference source where the data was collected. The option "Download" makes possible to export the generated graph into common image formats, such as "PNG", "JPEG" or "PDF". The option "Share" is an easy resource to promote sharing the content on social medias or by email.

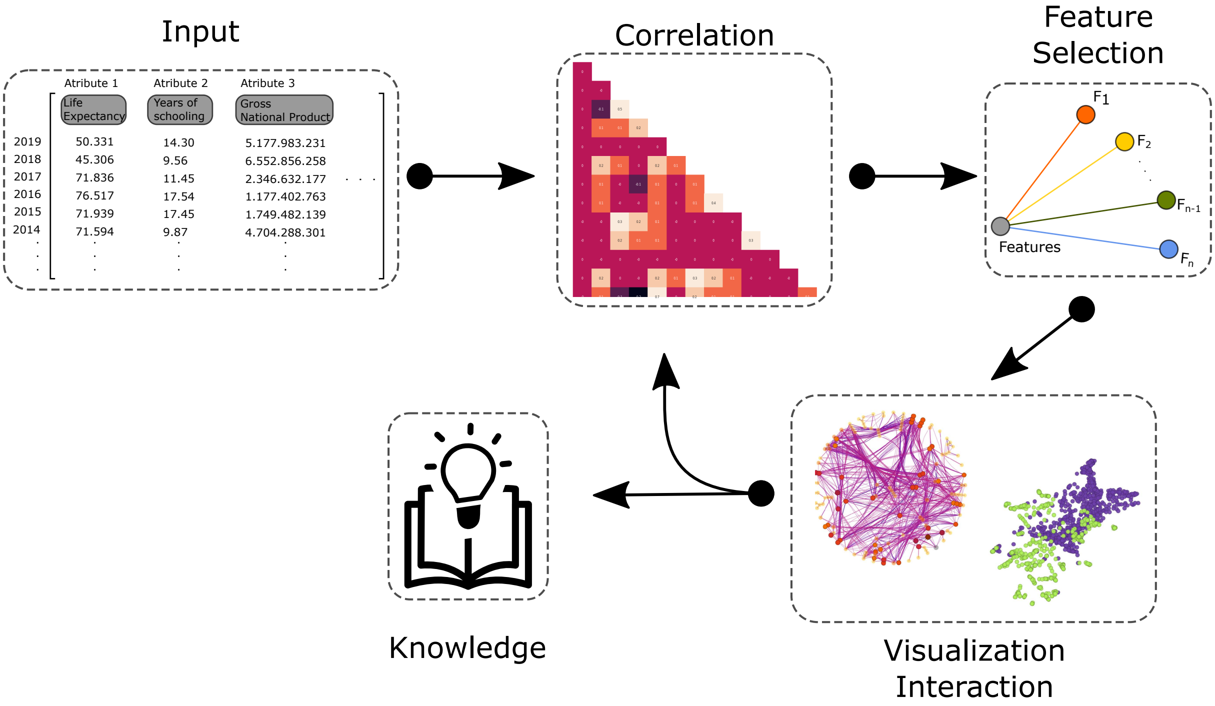
4.5 Knowledge

Figure 10 represents the analysis knowledge inference system that will be integrated into the Visual Interactive interface. After the initial pre-processing and visualization phase, the formatted input with the selected parameters can be used to test different hypothesis. The input is formatted as a matrix nxm where n is number of rows in the matrix representing each year of the time series and m is the number of columns representing the available features. The general idea is to provide a framework for analysing different assumption over a general question and use different resources to validate the proposition. For example, the following hypothesis can be suggested:

> “*"Is the Gross National Product the main metric to define life expectancy in a country?"*”

To test this hypothesis, it is possible to provide several tools that can help the final user. One of the most important statistical measurements is correlation that defines how two variables are related. The relation can assume different forms, such as linear, non-linear or even not correlated. It can also be positive, negative or neutral, depending on how the variables moves interdependently. Correlation can be measured by coefficient calculations and represented by heat map matrix when there is more than two variables. It can also be represented visually by plotting each combination pair of variables in a scatter plot.

The correlation analysis makes possible to point out dependent variables and the selection of the most important features. This is specially true for systems that have too many parameters and several data points, such as the human development data and



	Atribute 1	Atribute 2	Atribute 3
	Life Expectancy	Years of schooling	Gross National Product
2019	50.331	14.30	5.177.983.231
2018	45.306	9.56	6.552.856.258
2017	71.836	11.45	2.346.632.177
2016	76.517	17.54	1.177.402.763
2015	71.939	17.45	1.749.482.139
2014	71.594	9.87	4.704.288.301

Figure 10 – Knowledge inference system.

their time series information. To extract real real knowledge from the data, alternative representation will be used based on graphs, trees, scatter plots, etc.

The pipeline provides an interactive interface based on a matrix input composed by time series of feature related data. It works as a funnel tuning filter that selects the best features, transforming the previous input in a new representation that moves downstream. The flow can restarted in the next iteration, the user can re-evaluate the choices on previous steps. It is also the user that will define when the obtained results are satisfactory and the inferred knowledge level.

5 RESULTS

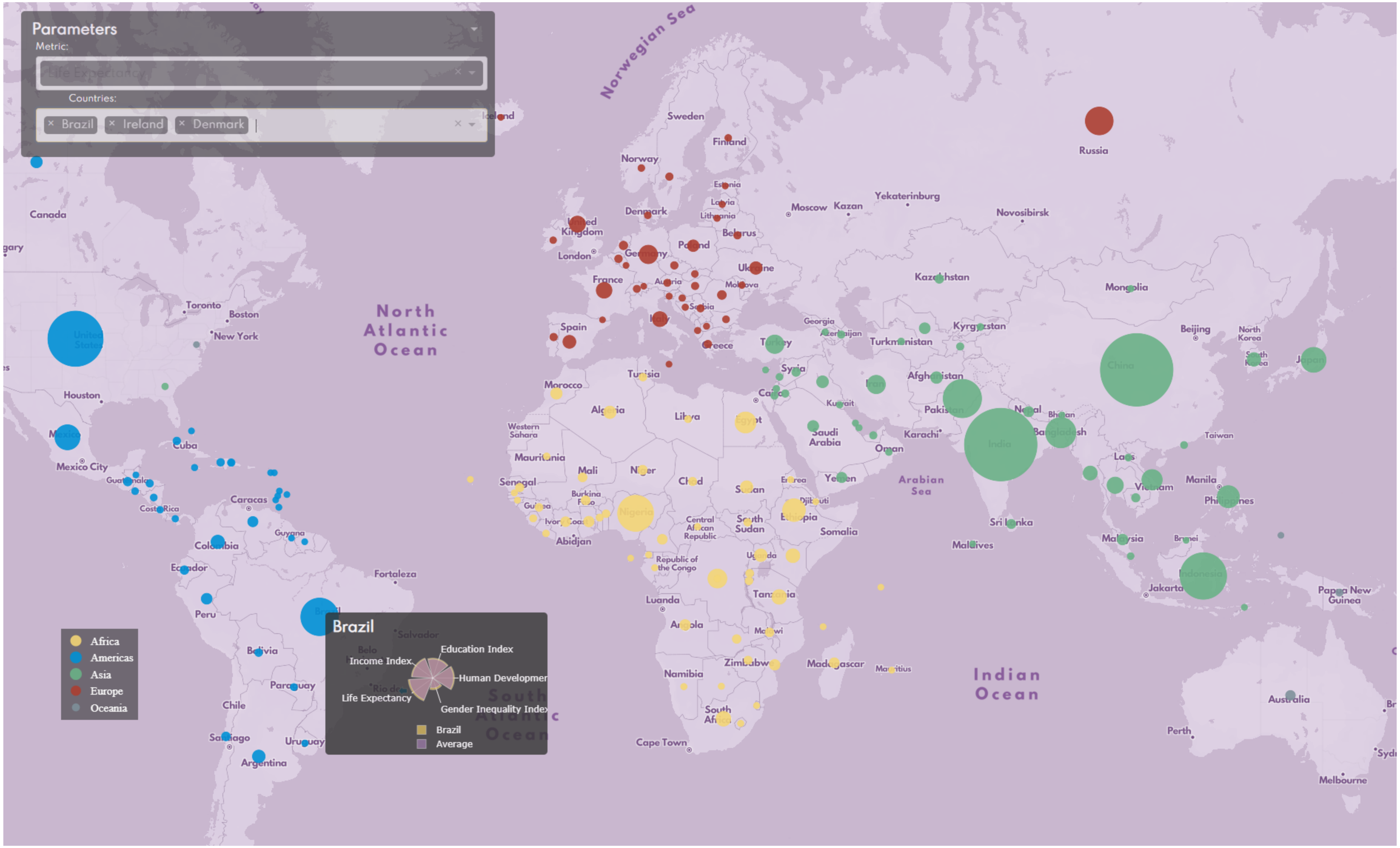
This chapter describes the results obtained when designing the interactive user web interface shown by Figure 16. The solution is a web interface that provides easy access, connectivity around the world (although it is not available online yet), environment independent (such as Windows, Linux, Mac or embedded platforms), and there is no need for software or hardware installation. The interface allows interaction with the user by providing parameter selection that will affect how data will be analyzed. It is composed by multi drop-down selection items that allows the selection of data listed in the Human Development Reports. Each measurement has its own unit or it is normalized following its own criterion. The visualization analysis provides specific data comparison, depending on its final purpose. It is also possible to specify the interval of analysis, based on the time series of the database by year of publication.

5.1 Main interface

The initial screen is a scalable background world map. The scalability is provided by a vectorial image that does not loose resolution by expanding the browser window size or applying zooming. It represents each country in the globe and their respective features for selection. Each country holds its respective relative coordinates. This is a design decision to allow easy selection access. The scalability of the map allows the zoom in and out operations. Therefore it is possible to study specific regions or areas of the map when zooming in. It is also gives the flexibility to select items from alternate areas by zooming out. The size of the items represents their population. Larger items represent very populated countries and vice-and-versa. The color of the items represents each habitable continent: *Africa, Americas, Asia, Europe, and Oceania.* It is possible to select the countries by continent.

When hover the mouse over each item, a schematic circular graph will show the information of the respective country. A selection of some basic parameters are shown: *Life Expectancy, Education index, Income Index, Gender Inequality Index, and Human Development Index.* It is also possible to compare the aforementioned measurements of the country with the world average.

There are four ways of providing selection on the main interface: 1) Double clicking in the country. This option only allows the selection of one country per click; 2) Selecting the continent. The continent option is located on the bottom left part as an interactive legend and it is indicated by name and color. All countries from the selected continent are included in the analysis. It is also possible to select multiples continents; 3) Box and Lasso selection. This option is located on the upper right corner menu. The default selection is



Parameters

Metric: Life Expectancy

Countries: × Brazil × Ireland × Denmark

Figure 11 – Visualization tool using web interface for user interaction.

the lasso, which gives flexibility to select uniform areas on the map. The box selection is similar, but the selection area needs to be rectangular; 4) Country selection by the *Parameters* menu. This option is located on the upper left corner. It is possible to list all the countries in the drop-down menu. The menu is also dynamic. It will be updated accordingly with the selected countries. It is also possible to remove previous items.

The upper left menu also provides the metric selection. Depending on the type of analysis, it is possible to list several analyzed features described in Appendix A. When the chosen country(ies) are selected together with the desired metrics, an interactive result box is shown. This interface overlaps the current map, but provides transparency over the background. So it is still possible to semi visualize the selected options when analysing the obtained results. To avoid disturbances of the transparency, when the mouse is over the result box, there is no opacity, i.e., a solid background facilitates the visualization of different graphs. An example of result analysis box is presented in Appendix B. There are five types of analysis available and presented in the result box: 1) Line Chart; 2) Correlation; 3) Scatter Matrix; 4) Bubble Chart; and 5) Projection. The following sections will describe each functionality.

5.2 Line Chart

The Line Chart visualization interface is highlighted in Figure 12. It exhibits the evolution of indexes or measurements over a period of time. It is a simple but powerful representation that shows tendencies or unexpected changes. Line Charts are also helpful to compare different curves, such as in Figure Figure 12 that shows the Human Development Index trend behaved in Brazil, Denmark and Ireland during the period between 1990 and 2019. Each country is represented by continue curve lines. To differentiate each country, an automatic color selection is used. A legend is position on the upper right corner of the panel to guide the final user on their comparison. The system is completely dynamic, i.e, every time one of the input parameters are changed, the graph will adapt to show the selected scenario, including unit and scale value. However only one of the metric features can be displayed at the same time. Other options are available to compare different metrics.

Figure 12 shows the HDI of Denmark and Ireland are more similar when compared to Brazil. In the three cases, all countries saw an improvement of the index with a slowdown in the growth rate, specially in the last years. This shows an stagnating tendency in the future with values close to 1.0 to Ireland and Denmark and 0.75 to Brazil. In this example, it is possible to analyse the situation in Brazil and specify why its HDI are not closer to 1.0. It is also possible to see that Ireland had a push of improvement after 2010, surpassing Denmark.

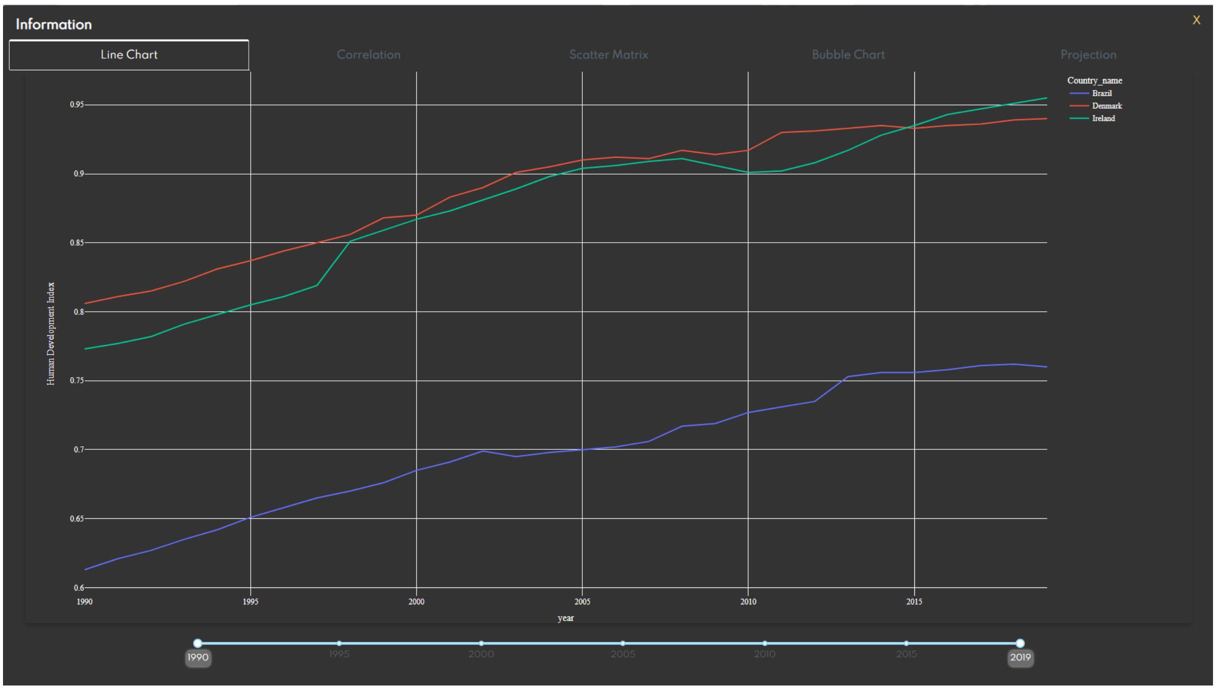


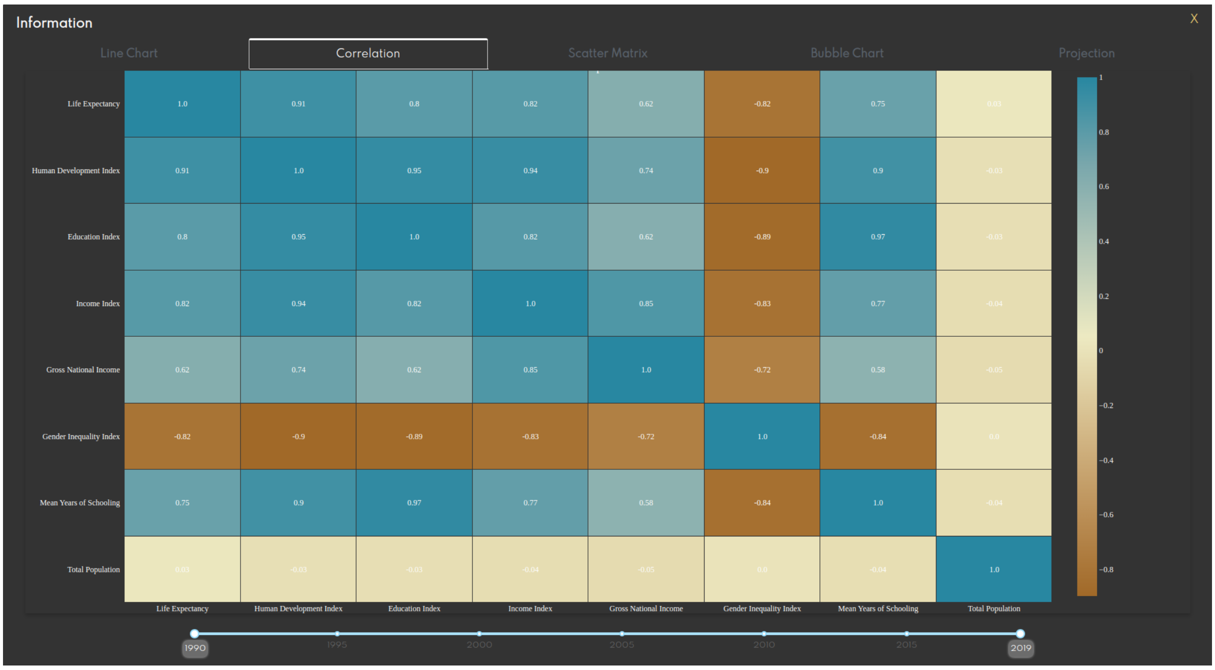
Figure 12 – Line Chart showing the evolution of the Human Development Index of Brazil, Denmark and Ireland over the period between 1990 and 2019.

5.3 Correlation Matrix

Figure 13 shows the heat map graph generated when the tab “Correlation” is selected. The correlation is represented as a matrix where lines and columns have the same parameters. The parameters selected were: *Human Development Index*, *Life Expectancy*, *Education Index*, *Income Index*, *Gross National Income*, *Gender Inequality Index*, *Life Expectancy*, *Mean Years of Schooling*, and *Total Population*.

The upper and lower triangular part of the matrix M are symmetric, i.e, the element $x_{ij} = x_{ji}$, $\forall x \in M$. The matrix diagonal items have unitary values, i.e., $x_{ii} = 1$, representing the fact that the correlation of an element with itself has the highest possible positive correlation. The heat map provides a color scheme grade positioned at the right side of the graph to help analyse large matrix where the number of columns and rows can scale drastically. The precise correlation of each pair comparison is also displayed on each cell of the matrix. The correlation value can also be negative, illustrating the scenario where two measurements are inversely proportional: when one variable becomes larger the other becomes smaller.

Some interesting results are obtained when analysing the correlation heat map presented by Figure 13. For example, the Human Development Index is highly correlated with the metrics *Life Expectancy, Education Index, and Income Index*. This is already expected, because the HDI value is not an independent metric. As shown by Equation 4.2,



	Life Expectancy	Human Development Index	Education Index	Income Index	Gross National Income	Gender Inequality Index	Mean Years of Schooling	Total Population
Life Expectancy	1.0	0.91	0.8	0.82	0.62	-0.82	0.75	0.03
Human Development Index	0.91	1.0	0.95	0.94	0.74	-0.9	0.9	-0.03
Education Index	0.8	0.95	1.0	0.82	0.62	-0.89	0.97	-0.03
Income Index	0.82	0.94	0.82	1.0	0.85	-0.83	0.77	-0.04
Gross National Income	0.62	0.74	0.62	0.85	1.0	-0.72	0.58	-0.05
Gender Inequality Index	-0.82	-0.9	-0.89	-0.83	-0.72	1.0	-0.84	0.0
Mean Years of Schooling	0.75	0.9	0.97	0.77	0.58	-0.84	1.0	-0.04
Total Population	0.03	-0.03	-0.03	-0.04	-0.05	0.0	-0.04	1.0

Figure 13 – Heat map correlation.

these are the correlated parameters used to calculate its measurement. The opposite is verified when analyzing the *Gender Inequality Index*. The correlation is inversely proportional with the other metrics, showing how important is to prioritize women's rights and autonomy for socio-economic development.

It is also clear that the *Total Population* metric is neutral when compared with the other measurements, i.e., its value is not directly or inversely proportional with the comparison variables, being around the zero value. This is specially true, because most of the index and metrics are normalized accordingly with the population of a country.

5.4 Scatter Matrix

The *Scatter matrix* tab is selected when the user is interested in visualize bivariate relationships between the selected metrics. In this way, it is possible to have an overview of relationships in only one chart. Figure 14, for example, shows a case study from 1990 to 2019 (the full period of historical data), selecting four countries: *Brazil, Ireland, Denmark*, and five metric comparison: *Human Development Index, Life Expectancy, Education Index, Income Index, and Gross National Income.*

Figure 14 shows the results of the case study. The diagonal will always represent a linear relationship because we are comparing identical variables. It is possible to compare the relationship of the *Human Development Index* with its composed variables *Life Expectancy, Education Index, and Gross National Income*. Specially for the first two, there is a visible linear relationship. The third, *Gross National Income* has a much steeper initial

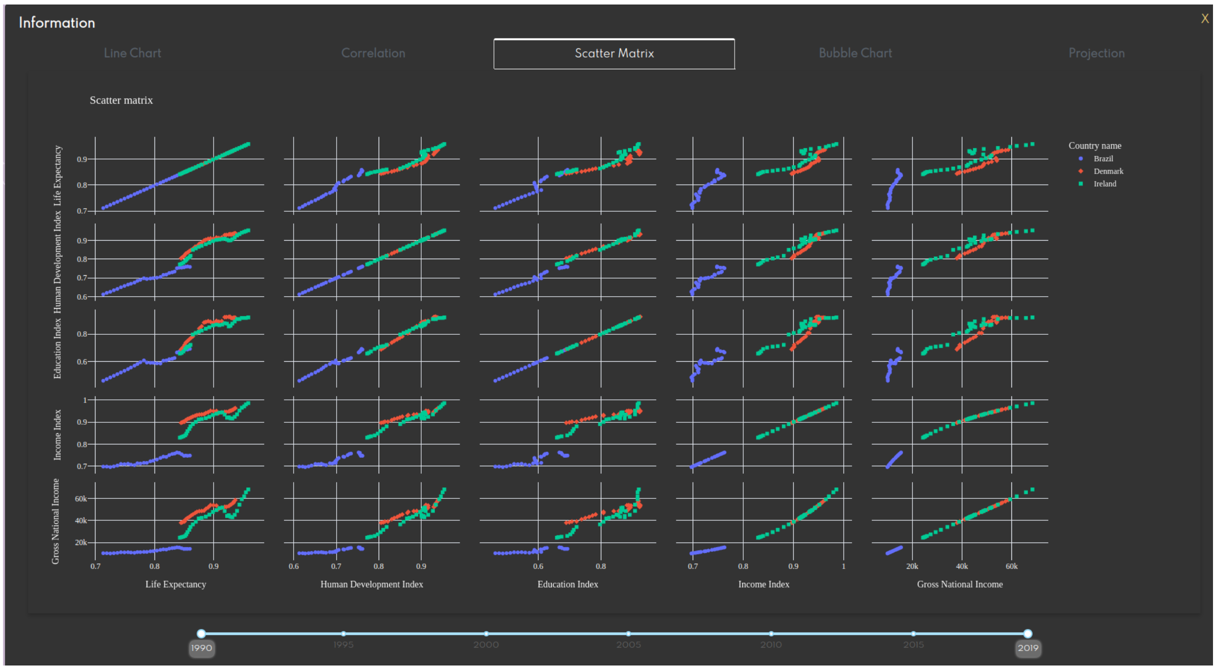


Figure 14 – Scatter matrix representing the bivariate relationship between metrics.

inclination at the beginning of the registry and shows signs of stabilization at the end with low inclination tending to constant behavior after $70,000 per capita. This is demonstrated by (KAHNEMAN; DEATON, 2014). They affirm that there is virtually no gain in human development and well being from annual income above $75,000 per capita.

If we analyze the impact of the metrics on each country, there is a clear distinction on two groups. The first is composed by developed nations: *Denmark, Ireland, and United States*. The second has only Brazil as representative. For most compared metrics, there is a gap between this two groups. It is also clear how variables such as the increase of *Gross National Income* can affect most of the development index of Brazil, such as *Life Expectancy and Education Index*. In this case, the increase is almost a vertical line, showing the potential of increasing the internal GNI.

5.5 Projection

The *Projection* tab calculates the t-Distributed Stochastic Neighbor Embedding (t-SNE). The algorithm projects high-dimension datasets in a lower dimension plan (two dimensions in this case). Therefore, it is possible to infer similarities based on the distance of each point and its neighbours, representing each selected country. Figure 15 shows the result of the t-SNE algorithm applied to all countries and all numeric features.

Each color in Figure 15 represents the continent of each country. However this is a post-processing information. The t-SNE algorithm has a non-supervised learning process, i.e., the algorithm doesn't use any previous defined labels. Only the features are analysed

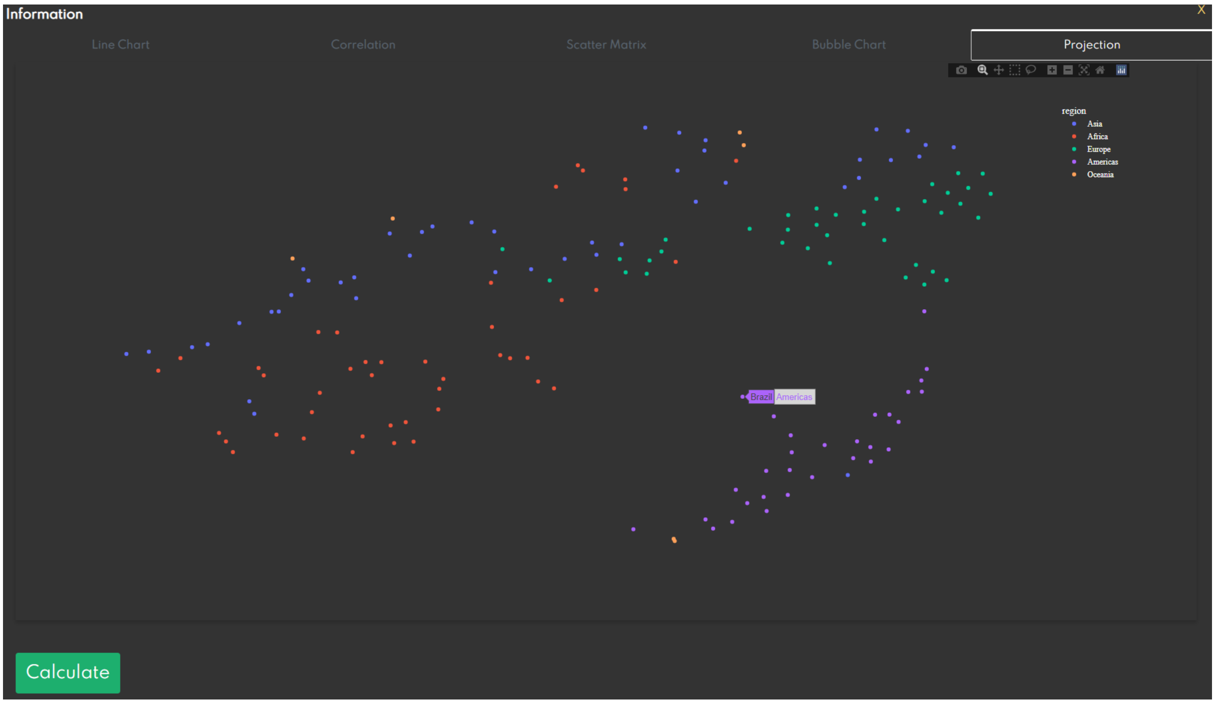


Figure 15 – t-Distributed Stochastic Neighbor Embedding (t-SNE) projection.

to project the points in a 2D plan. To avoid visual pollution, no label are shown by default. However, when hovering the mouse over the points, information about the specific country is summarized. There is no differentiation by size.

One of the common operations on non-supervised algorithms is to define clusters over the projected points. There is a clear cluster defining the America continent. However, USA is one of the points that are more distant from the others. In a clustering process USA would be closer to the cluster defined by the European points. It is interesting to see that two countries from Oceania are closer to the America cluster: Samoa and Tonga.

The cluster defined by the European countries is the one on the upper right corner. As mentioned before, the cluster includes USA. It also include nine Asian countries: *Qatar, United Arab Emirates, Kuwait, Saudi Arabia, Singapore, Brunei, Bahrain, Israel and Cyprus*. These Asian countries are also on the upper part and could be associate with its own cluster. For the other countries, it is difficult to define specific areas. The continent colors are also very diverse, including Africa, Asia, Europe and Oceania.

6 CONCLUSION

This project investigate the use of data visualization techniques applied to the world bank. Specifically, there was a focus on the Human Development Index and its internal calculation parameters. To this end, it was proposed a data visualization platform able to provide several visualizations: 1) Line Charts, 2) Correlation, 3) Scatter Matrix, 4) Bubble Chart, and 5) Projection. In order to provide flexibility, the platform was developed as a web interface in order to provide easy access, connectivity, environment independence, and no need for software or hardware installation. Finally, the interface provides several ways of filtering the search criteria process, providing a world map background to facilitate recognition of world data.

The proposed system was developed using the open-source framework Python-Dash that can be extended in the future to provide other visualization techniques. This is possible due to the modular implementation of the interface. The adopted strategy showed a trade-off between the implementation of the application using a Python solution instead of a more general approach building the web interfaces from scratch using technologies such as HTML, CSS, Javascript, NodeJs, etc.

It is clear that the Data visualization field is an extensive area and the implementation of other techniques can help further in the analysis of human development criteria. It is also vital to provide an extensive dataset with high-dimension features that support the findings and the decision making policies. The data itself provided by the world bank has several missing information, specially on older years, where countries did not make their own data available. The authors believes that it is important in the future to estimate missing data and include new feature information capable of describing the status and condition of each country.

The t-Distributed Stochastic Neighbor Embedding algorithm is an efficient implementation of a non-supervised projection that can handles high-dimensional datasets. On the other hand, it does not process categorical data and does not work well with missing information. The results obtained by the t-SNE allowed the comparison of different countries even though they were not in the same continent. Further studies is needed in order to establish the metrics responsible for the clustering decision.

Finally, the contribution of this study goes beyond pointing at which visualization technique is superior. Instead, we enrich the debate of interactive data visualization and how it can contribute to decision making process while providing flexibility to implement data visualization techniques applied to human development applications. We believe that the spreading of interactive data visualization platforms and interfaces will become more discussed on different research and development areas in the near future.

REFERENCES

AILON, N.; CHAZELLE, B. Faster dimension reduction. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 53, n. 2, p. 97–104, feb 2010. ISSN 0001-0782. Available at: https://doi-org.proxy1-bib.sdu.dk/10.1145/1646353.1646379.

ALI, M. *et al.* Clustering and classification for time series data in visual analytics: A survey. **IEEE Access**, v. 7, p. 181314–181338, 2019.

ARROW, K. *et al.* Are we consuming too much? **Journal of Economic Perspectives**, v. 18, n. 3, p. 147–172, 2004.

BERAHMAND, K. *et al.* Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding. **Computers in Biology and Medicine**, v. 138, p. 104933, 2021. ISSN 0010-4825. Available at: https://www.sciencedirect.com/science/article/pii/S0010482521007277.

BERNARD, J. *et al.* Vial: a unified process for visual interactive labeling. **The Visual Computer**, v. 34, p. 1189–1207, 09 2018.

BLöMER, J.; BRAUER, S.; BUJNA, K. A complexity theoretical study of fuzzy <i>k</i>-means. **ACM Trans. Algorithms**, Association for Computing Machinery, New York, NY, USA, v. 16, n. 4, sep 2020. ISSN 1549-6325. Available at: https://doi.org/10.1145/3409385.

BRZEZINSKI, D. *et al.* Visual-based analysis of classification measures and their properties for class imbalanced problems. **Information Sciences**, v. 462, p. 242–261, 2018. ISSN 0020-0255. Available at: https://www.sciencedirect.com/science/article/pii/S0020025518304602.

CAO, N. *et al.* Acm tist special issue on visual analytics. **ACM Trans. Intell. Syst. Technol.**, Association for Computing Machinery, New York, NY, USA, v. 10, n. 1, dec 2018. ISSN 2157-6904. Available at: https://doi-org.proxy1-bib.sdu.dk/10.1145/3277019.

CHO, K. *et al.* Economic analysis of data center cooling strategies. **Sustainable Cities and Society**, v. 31, n. Supplement C, p. 234 – 243, 2017. ISSN 2210-6707. Available at: http://www.sciencedirect.com/science/article/pii/S2210670716304899.

COHEN-ADDAD, V. *et al.* Hierarchical clustering: Objective functions and algorithms. **J. ACM**, Association for Computing Machinery, New York, NY, USA, v. 66, n. 4, jun 2019. ISSN 0004-5411. Available at: https://doi.org/10.1145/3321386.

COSTANZA, R. Time to leave GDP behind.: EBSCOhost. **Comment**, v. 505, n. NATURE, p. 2–7, 2015. Available at: http://web.a.ebscohost.com.ezproxy.royalroads.ca/ehost/pdfviewer/pdfviewer?vid=1&sid=c3a79af4-c748-4a1a-87e5-a158f249ca45%40sessionmgr4008.

COSTANZA, R. *et al.* Development: Time to leave gdp behind. **Nature News**, v. 505, n. 7483, p. 283, 2014.

DAS, S. *et al.* Beames: Interactive multimodel steering, selection, and inspection for regression tasks. **IEEE Computer Graphics and Applications**, v. 39, n. 5, p. 20–32, 2019.

DAS, S.; ENDERT, A. Legion: Visually compare modeling techniques for regression. *In*: **2020 Visualization in Data Science (VDS)**. [*S.l.: s.n.*], 2020. p. 12–21.

DINGEN, D. *et al.* Regressionexplorer: Interactive exploration of logistic regression models with subgroup analysis. **IEEE Transactions on Visualization and Computer Graphics**, v. 25, n. 1, p. 246–255, 2019.

ENDERT, A.; FIAUX, P.; NORTH, C. Semantic interaction for visual text analytics. *In*: **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2012. (CHI '12), p. 473–482. ISBN 9781450310154. Available at: https://doi-org.proxy1-bib.sdu.dk/10.1145/2207676.2207741.

ENDERT, A. *et al.* The state of the art in integrating machine learning into visual analytics. **Computer Graphics Forum**, Wiley, v. 36, n. 8, p. 458–486, Mar 2017. ISSN 0167-7055. Available at: http://dx.doi.org/10.1111/cgf.13092.

FERNSTAD, S. J. *et al.* Visual exploration of microbial populations. *In*: **2011 IEEE Symposium on Biological Data Visualization (BioVis).** [*S.l.: s.n.*], 2011. p. 127–134.

FUCHS, R.; WASER, J.; GROLLER, M. E. Visual human+machine learning. **IEEE Transactions on Visualization and Computer Graphics**, v. 15, n. 6, p. 1327–1334, 2009.

GAO, J.; SHI, Q.; CAETANO, T. S. Dimensionality reduction via compressive sensing. **Pattern Recognition Letters**, v. 33, n. 9, p. 1163–1170, 2012. ISSN 0167-8655. Available at: https://www.sciencedirect.com/science/article/pii/S0167865512000487.

GOOGLE. **World Development Indicators - Google Public Data Explorer**. 2020. Available at: https://bit.ly/3DEmnCf.

GRECO, S. *et al.* On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. **Social indicators research**, Springer, v. 141, n. 1, p. 61–94, 2019.

GROPP, C. *et al.* Clustered latent dirichlet allocation for scientific discovery. *In*: **2019 IEEE International Conference on Big Data (Big Data)**. [*S.l.: s.n.*], 2019. p. 4503–4511.

GUO, Y. *et al.* Machine learning based feature selection and knowledge reasoning for cbr system under big data. **Pattern Recognition**, v. 112, p. 107805, 2021. ISSN 0031-3203. Available at: https://www.sciencedirect.com/science/article/pii/S0031320320306087.

HEER, J.; CARD, S.; LANDAY, J. Prefuse: A toolkit for interactive information visualization. *In*: . [*S.l.: s.n.*], 2005. p. 421–430.

HU, L. *et al.* Kr-dbscan: A density-based clustering algorithm based on reverse nearest neighbor and influence space. **Expert Systems with Applications**, v. 186, p. 115763, 2021. ISSN 0957-4174. Available at: https://www.sciencedirect.com/science/article/pii/S0957417421011374.

JOHANSSON, S.; JOHANSSON, J. Interactive dimensionality reduction through user-defined combinations of quality metrics. **IEEE Transactions on Visualization and Computer Graphics**, v. 15, n. 6, p. 993–1000, 2009.

KAHNEMAN, D.; DEATON, A. High income improves evaluation of life but not emotional well-being. *In*: **Proceedings of National Academy of Sciences**. [*S.l.: s.n.*], 2014. v. 107, n. 38, p. 16489–16493.

KANIKA *et al.* Visual analytics for comparing the impact of outliers in k-means and k-medoids algorithm. *In*: **2019 Amity International Conference on Artificial Intelligence (AICAI)**. [*S.l.: s.n.*], 2019. p. 93–97.

KIM, H. *et al.* Interaxis: Steering scatterplot axes via observation-level interaction. **IEEE Transactions on Visualization and Computer Graphics**, v. 22, n. 1, p. 131–140, 2016.

KUCHER, K. *et al.* Active learning and visual analytics for stance classification with alva. **ACM Trans. Interact. Intell. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 7, n. 3, oct 2017. ISSN 2160-6455. Available at: https://doi-org.proxy1-bib.sdu.dk/10.1145/3132169.

KUZNETS, S. National income, 1929-1932. *In*: **National Income, 1929-1932**. [*S.l.: s.n.*]: NBER, 1934. p. 1–12.

KWON, B. C. *et al.* Clustervision: Visual supervision of unsupervised clustering. **IEEE Transactions on Visualization and Computer Graphics**, v. 24, n. 1, p. 142–151, 2018.

KWON, B. C. *et al.* Axisketcher: Interactive nonlinear axis mapping of visualizations through user drawings. **IEEE Transactions on Visualization and Computer Graphics**, v. 23, n. 1, p. 221–230, 2017.

LASISI, A.; ATTOH-OKINE, N. Principal components analysis and track quality index: A machine learning approach. **Transportation Research Part C: Emerging Technologies**, v. 91, p. 230–248, 2018. ISSN 0968-090X. Available at: https://www.sciencedirect.com/science/article/pii/S0968090X18304303.

LI, F. *et al.* Cluster's quality evaluation and selective clustering ensemble. **ACM Trans. Knowl. Discov. Data**, Association for Computing Machinery, New York, NY, USA, v. 12, n. 5, jun 2018. ISSN 1556-4681. Available at: https://doi-org.proxy1-bib.sdu.dk/10.1145/3211872.

LI, H. *et al.* Interactive machine learning by visualization: A small data solution. *In*: **2018 IEEE International Conference on Big Data (Big Data)**. [*S.l.: s.n.*], 2018. p. 3513–3521.

LI, J. *et al.* Overall quality evaluation of graph layouts based on regression analysis. *In*: **Proceedings of the 10th International Symposium on Visual Information Communication and Interaction**. New York, NY, USA: Association for Computing Machinery, 2017. (VINCI '17), p. 81–82. ISBN 9781450352925. Available at: https://doi-org.proxy1-bib.sdu.dk/10.1145/3105971.3105992.

MA, Y. *et al.* A visual analytical approach for transfer learning in classification. **Information Sciences**, v. 390, p. 54–69, 2017. ISSN 0020-0255. Available at: https://www.sciencedirect.com/science/article/pii/S0020025516301694.

MAATEN, L. V. D.; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, 2008. Available at: https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf.

MULTIPLES, S. **Standing on the borders of giants**. 2019. Available at: https://smallmultiples.com.au/projects/standing-on-the-borders-of-giants/.

NAGPAL, A.; JATAIN, A.; GAUR, D. Review based on data clustering algorithms. *In*: IEEE. **2013 IEEE conference on information & communication technologies**. [*S.l.: s.n.*], 2013. p. 298–303.

NATIONS, U. **Human Development Reports**. 2022. Available at: https://hdr.undp.org/data-center/documentation-and-downloads.

NOURASHRAFEDDIN, S. *et al.* A visual approach for interactive keyterm-based clustering. **ACM Trans. Interact. Intell. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 8, n. 1, feb 2018. ISSN 2160-6455. Available at: https://doi-org.proxy1-bib.sdu.dk/10.1145/3181669.

PATEL, E.; KUSHWAHA, D. S. Clustering cloud workloads: K-means vs gaussian mixture model. **Procedia Computer Science**, v. 171, p. 158–167, 2020. ISSN 1877-0509. Third International Conference on Computing and Network Communications (CoCoNet'19). Available at: https://www.sciencedirect.com/science/article/pii/S1877050920309820.

Pedro Conceição. **Human Development Report 2020 The next frontier Human development and the Anthropocene**. New York, 2020. 412 p. Available at: http://hdr.undp.org/sites/default/files/hdr2020.pdf.

RESCE, G. Wealth-adjusted human development index. **Journal of Cleaner Production**, v. 318, p. 128587, 2021. ISSN 0959-6526. Available at: https://www.sciencedirect.com/science/article/pii/S095965262102792X.

SACHA, D. *et al.* What you see is what you can change: Human-centered machine learning by interactive visualization. **Neurocomputing**, Elsevier, v. 268, p. 164–175, dec 2017. ISSN 0925-2312.

SHABDIN, N. I.; YA'ACOB, S.; SJARIF, N. N. A. Relationship types in visual analytics. *In*: **Proceedings of the 2020 6th International Conference on Computer and Technology Applications**. New York, NY, USA: Association for Computing Machinery, 2020. (ICCTA '20), p. 1–6. ISBN 9781450377492. Available at: https://doi-org.proxy1-bib.sdu.dk/10.1145/3397125.3397127.

SHERKAT, E.; MILIOS, E. E.; MINGHIM, R. A visual analytics approach for interactive document clustering. **ACM Trans. Interact. Intell. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 10, n. 1, aug 2019. ISSN 2160-6455. Available at: https://doi-org.proxy1-bib.sdu.dk/10.1145/3241380.

STAHNKE, J. *et al.* Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. **IEEE Transactions on Visualization and Computer Graphics**, v. 22, n. 1, p. 629–638, 2016.

STANTON, E. A. Human Development Index: A History. **Political Economy Research Institute Working Paper Series**, v. 127, n. February, p. 1–36, 2007.

TABLEAU. **Data visualization beginner's guide: a definition, examples, and learning resources**. 2018. Available at: https://www.tableau.com/learn/articles/data-visualization.

THRUN, M.; PAPE, F.; ULTSCH, A. Interactive machine learning tool for clustering in visual analytics. *In*: **2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)**. [*S.l.: s.n.*], 2020. p. 479–487.

TOKUDA, E. K.; COMIN, C. H.; COSTA, L. da F. Revisiting agglomerative clustering. **Physica A: Statistical Mechanics and its Applications**, v. 585, p. 126433, 2022. ISSN 0378-4371. Available at: https://www.sciencedirect.com/science/article/pii/S0378437121007068.

WANG, S.; SUN, Y.; BAO, Z. On the efficiency of k-means clustering: Evaluation, optimization, and algorithm selection. **Proc. VLDB Endow.**, VLDB Endowment, v. 14, n. 2, p. 163–175, oct 2020. ISSN 2150-8097. Available at: https://doi.org/10.14778/3425879.3425887.

WU, Y. *et al.* Enhanced clustering embedded in curvilinear distance analysis guided by pairwise constraints. **Information Sciences**, v. 556, p. 111–127, 2021. ISSN 0020-0255. Available at: https://www.sciencedirect.com/science/article/pii/S0020025520311944.

ZHAO, Y. *et al.* An independent central point optics clustering algorithm for semi-supervised outlier detection of continuous glucose measurements. **Biomedical Signal Processing and Control**, v. 71, p. 103196, 2022. ISSN 1746-8094. Available at: https://www.sciencedirect.com/science/article/pii/S174680942100793X.

APPENDIX

APPENDIX A – APPENDIX A

Table 3 – Database schema and time series by year from Human Development Report 2020

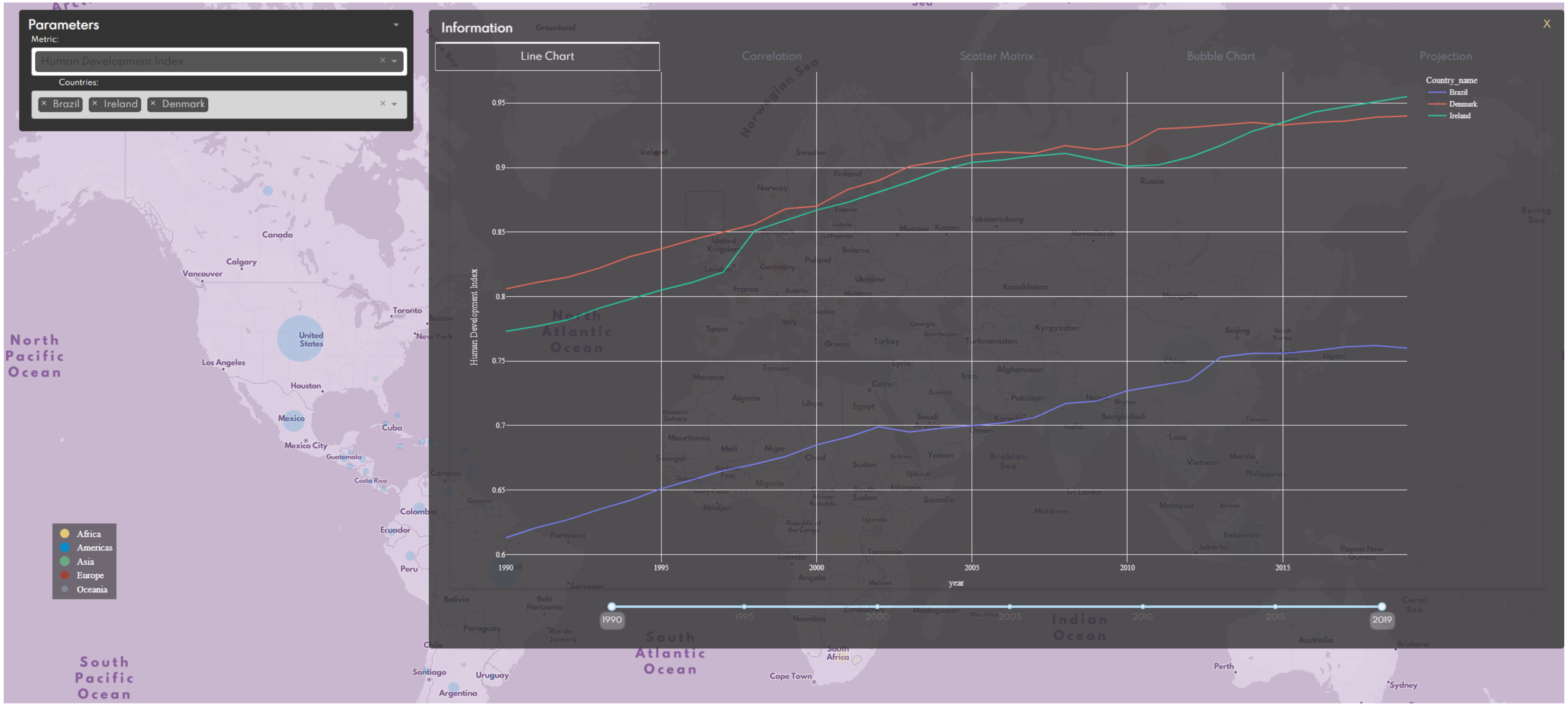
Dimension	Time series
Human Development Index (HDI)	
HDI rank	2019
Human Development Index (value)	1990-2019
Life Expectancy at Birth	1990-2019
Expected Years of Schooling (years)	1990-2019
Mean Years of Schooling (years)	1990-2019
Gross National Income Per Capita (2017 PPP$)	1990-2019
Gender Development Index (GDI)	
GDI Group	2019
Gender Development Index (value)	1995, 2000, 2005, 2010-2019
HDI female	1995, 2000, 2005, 2010-2019
Life Expectancy at Birth, female (years)	1995, 2000, 2005, 2010-2019
Expected Years of Schooling, female (years)	1995, 2000, 2005, 2010-2019
Mean Years of Schooling, female (years)	1995, 2000, 2005, 2010-2019
Gross National Income Per Capita, female (2017 PPP$)	1995, 2000, 2005, 2010-2019
HDI male	1995, 2000, 2005, 2010-2019
Life Expectancy at Birth, male (years)	1995, 2000, 2005, 2010-2019
Expected Years of Schooling, male (years)	1995, 2000, 2005, 2010-2019
Mean Years of Schooling, male (years)	1995, 2000, 2005, 2010-2019
Gross National Income Per Capita, male (2017 PPP$)	1995, 2000, 2005, 2010-2019
Inequality-adjusted Human Development Index(IHDI)	
HDI	2010-2019
Inequality-adjusted Human Development Index (value)	2010-2019
Coefficient of human inequality	2010-2019
Overall loss (%)	2010-2019
Inequality in life expectancy	2010-2019
Inequality in eduation	2010-2019
Inequality in income	2010-2019
Gender Inequality Index(GII)	
GII Rank	2019
Gender Inequality Index (value)	1995, 2000, 2005, 2010-2019
Maternal Mortality Ratio (deaths per 100,000 live births)	1995, 2000, 2005, 2010-2017

Continue on next page

Table 3 – *Continued from previous page*

Dimension	**Time series**
Adolescent Birth Rate (births per 1,000 women ages 15-19)	1995, 2000, 2005, 2010-2019
Population with at least some secondary education, female (% ages 25 and older)	1995, 2000, 2005, 2010-2019
Population with at least some secondary education, male (% ages 25 and older)	1995, 2000, 2005, 2010-2019
Share of seats in parliament, female (% held by women)	1995, 1997, 2000, 2005, 2010-2019
Share of seats in parliament, male (% held by men)	1995, 1997, 2000, 2005, 2010-2019
Labour force participation rate, female (% ages 15 and older)	1995, 2000, 2005, 2010-2019
Labour force participation rate, male (% ages 15 and older)	1995, 2000, 2005, 2010-2019
Planetary pressures–adjusted Human Development Index(PHDI)	
HDI Rank	Latest year 2019
HDI	Latest year 2019
Planetary pressures–adjusted Human Development Index (value)	Latest year 2019
Difference from HDI value (%)	Latest year 2019
Difference from HDI rank	Latest year 2019
Carbon dioxide emissions per capita (production) (tonnes)	Latest year 2019
Material footprint per capita (tonnes)	Latest year 2019

APPENDIX B – APPENDIX B



Parameters

Metric: Human Development Index

Countries: Brazil, Ireland, Denmark

Figure 16 – Box results after the selection of countries and metric.