ANNALS OF DATA ANALYSIS

Toward a Model for Public Policy Analysis – Lessons from Higher Education Data

Ruy de Deus e Mello Neto ruydedeus@ufc.br
Hugo Medeiros hugoavmedeiros@gmail.com

RESEARCH ARTICLE

Annals of Data Analysis
https://periodicos.ufpe.br/revistas/ADA/

Submitted on: 12.12.2025.
Accepted on: 01.08.2026.
Published on: 01.08.2026.

Abstract

This paper presents a model for public policy analysis. The model adapts common automated exploratory data analysis (autoEDA) techniques to the specific requirements of public policy, with a focus on indicators, policy planning, monitoring, and evaluation. The research followed a Design Science Research (DSR) approach, resulting in a solution implemented in R and tested with higher education data. The model consists of the following steps: acquisition and cleaning of data; assessment of data quality; exploratory data analysis guided by indicators; and reporting of results. Applying the model enabled reproducible exploratory analyses, translating raw data into indicators and segmentations that were relevant and actionable for decision makers as well as monitoring and evaluation practitioners. The findings suggest that current alternative solutions prove insufficient and require adaptation to the specific needs of decision-making and governance processes. Future research should include testing the model with data from other policy areas such as health, mobility, and other domains.

Keywords

eploratory data analysis; public policy analysis; design science research; indicators; higher education

1. Introduction

This paper reports the development of a model for public policy analysis, applied in the context of the Termo de Execução Descentralizada (Term of Decentralized Execution, TED) No. 13,738 - a research project conducted by the Federal University of Ceará (UFC) through an agreement with the Secretariat of Higher Education of the Ministry of Education (MEC) - to analyze large-scale data from the Programa Universidade para Todos (ProUni). Although the model was originally designed for analyzing data from a specific public policy on access to higher education, we argue that its steps and procedures can be applied to other contexts where the main challenge is to synthesize massive datasets into indicators.

The structure of this paper follows a variation of the framework proposed by Young and Quinn (2002), consisting of introduction, methods, results, and conclusion. This structure was chosen because it aligns both with reporting practices in public policy and with the Design Science Research (DSR) methodology, which is oriented toward identifying and solving problems.

The introduction describes the general context of public policy analysis; the activities of planning, monitoring, and evaluation, which operationalize the concrete dimensions of analysis; and the specific features of ProUni, which was analyzed within TED No. 13,738.

The methodology section introduces the methods and materials: DSR, automated exploratory data analysis (autoEDA), and the datasets used to construct the solution. Since DSR itself recognizes the possibility of adapting existing solutions, this section also outlines the guiding principles for problem-solving: the problem situation, the available alternatives, and the problem statement.

The results section presents the model itself and illustrates its application using real data. Finally, the conclusion summarizes the main findings and translates them into recommendations for practitioners interested in adopting the model, discusses its limitations, and suggests directions for future research.

.

1.1. Analysis of Public Policies

Public policies are central to government and the public sector. Yet, despite their centrality, they lack a standard definition. They are inherently polysemic, subject to multiple interpretations depending on theoretical assumptions or domains of interest.

According to Souza (2006), in her literature review of the field - including its history and evolutionamong the possible definitions there is a line of thought that emphasizes public policies as governmental action to address social problems within a field of constant dispute.

Howlett, Ramesh, and Perl (2013), in their foundational book, state that public policies are techno-political processes that occur within a limited space of constraints and trade-offs, aiming to define relevant social problems, set goals for action, and determine how best to achieve them.

Another relevant definition in the field is offered by Frey (2000). The author explores the distinctions between ‘policy’, ‘politics’, and ‘polity’, arguing that public policies may be connected to any of these dimensions: ‘polity’ refers to institutional structures; ‘politics’ to processes and disputes; and ‘policy’ to the more concrete and practical dimension of public policies.

In a more recent study, Lotta (2019) offers a definition of public policy analysis with a focus on the ‘policy’ dimension, which she conceptualizes as ‘implementation analysis’. According to the author, this type of analysis has five main elements: (1) public policies are complex; (2) they involve multiple levels and layers of decision-making; (3) they generate numerous interactions among agents; (4) they are shaped by multiple factors of influence; and (5) implementation analysis is therefore truly focused on ‘politics as it is’.

We consider this perspective more appropriate to our proposition, as it conceives public policies as disputes within a complex arena of agents interacting across multiple levels and layers to address social problems. This approach provides a more suitable theoretical foundation to operationalize research on data analysis within government, particularly in the strategic management processes of planning, monitoring, and evaluation.

1.2. Programa Universidade para Todos – ProUni

ProUni was established in 2005 as a federal strategy to expand access to higher education for low-income students, particularly those excluded from the highly competitive public sector (Francis & Tannuri-Pianto, 2012). Rooted in the goals of the National Education Plan (Law No. 10,172/2001) (Brasil, 2001), the program evolved through a sequence of legislative instruments - Bill No. 3,582/2004 (Brasil, 2004a), Provisional Measure No. 213/2004 (Brasil, 2004b), and Law No. 11,096/2005 (Brasil, 2005) - culminating in a hybrid model that combined public policy objectives with the structural interests of private higher education institutions. Its eligibility criteria were defined by income thresholds - up to 1.5 minimum wages per capita for full scholarships and up to 3 for partial scholarships - and were initially restricted to graduates of public high schools or to students who had received a full scholarship (tuition waiver) in private high schools. Scholarship modalities included full and partial (50%, and initially 25%, although the latter soon became residual). The reform introduced by Law No. 14,350/2022 (Brasil, 2022) broadened eligibility by incorporating all graduates from private high schools, provided they met the established income thresholds.

From its inception, ProUni’s design reflected both inclusionary commitments and political negotiation (Catani & Gilioli, 2005). The introduction of partial scholarships and the expansion of eligibility thresholds were critical adjustments that reconciled social objectives with institutional demands. The program also incorporated equity mechanisms, mandating proportional access for Black, Brown, and Indigenous students, as well as individuals with disabilities, alongside scholarships reserved for in-service public-school teachers (Brasil, 2005; 2022).

Financing operates through federal tax exemptions (IRPJ, CSLL, COFINS, PIS-PASEP) granted to participating private institutions that comply with scholarship quotas proportional to the number of tuition-paying students, distributed by program, time of day, and campus. Governance includes a ten-year adhesion agreement and sanctions for non-compliance, such as suspension for up to three selection cycles or definitive exclusion without prejudice to current beneficiaries. Crucially, compliance and the scale of tax relief are regulated by Normative Instruction No. 1,394, of 12 September 2013, issued by the Brazilian Federal Revenue Service (Receita Federal do Brasil, RFB) (Brasil, 2013). This regulation defines the Effective Scholarship Occupancy Ratio (POEB) and ties the scope of tax exemptions to the verified occupancy of ProUni-funded seats. In this design, the program aligns inclusion goals with institutional incentives, illustrating a negotiated state–market arrangement in Brazilian higher education.

In this sense, ProUni represents a paradigmatic case of state–market negotiation in the Brazilian higher education system: while expanding access for historically underrepresented groups, it simultaneously reinforced the central role of private institutions in absorbing unmet demand for higher education. To better address these dynamics, it is necessary to establish robust processes of planning, monitoring, and evaluation in order to identify bottlenecks and challenges within specific segments, regions, or procedures of the policy. Beyond the complexity inherent to affirmative social policies and the challenges of the strategic management cycle of planning, monitoring, and evaluation, another critical constraint lies in the need to operate with big data - integrating and analyzing diverse data sources, each characterized by its own rules and limitations.

1.3. Planning, Monitoring and Evaluating

Planning, monitoring, and evaluation are interconnected phases within the broader process of strategic management, aimed at producing more effective public policies. Planning is usually the first step, involving ex ante analysis, strategy definition, and the formulation of general guidelines for the process.

Monitoring takes place once implementation is underway. It entails paced and continuous follow-up of immediate and direct effects, generating actionable information for decision-making (Rogger & Schuster, 2023; Souza & Loreto, 2021; Vaitsman, Rodrigues, & Paes-Souza, 2006).

Finally, evaluation is a more discrete phase, intended to produce detailed and robust analyses. It applies specific techniques to minimize bias, isolate effects, and capture interactions among indicators, with the purpose of correcting ongoing actions or planning more effective interventions (Gertler et al., 2018; Instituto de Pesquisa Econômica Aplicada [IPEA], 2018a, 2018b; Souza & Loreto, 2021; Vaitsman, Rodrigues, & Paes-Souza, 2006).

Regardless of the phase, data are a core element of the strategic management process, serving as the foundation for planning, monitoring, and evaluating public policies and strategic actions. However, raw data must be translated into indicators to produce timely, robust, and consistent estimates that can effectively guide action and decision-making (OECD, 2009).

This effort has two main goals: (1) to minimize the influence of weak data and (2) to synthesize information to facilitate strategic management and decision-making. Typically, this process emphasizes two aspects: segmentation of data across social dimensions (race, gender, education, etc.) and spatial dimensions (city, administrative units, etc.); and longitudinal follow-up and analysis over time. Such practices help ensure timely and robust analysis across different subsets, allowing weaknesses and limitations to be identified, adjusted, and replanned (Cardoso Júnior, 2015).

Nevertheless, as we will discuss in more detail, existing alternatives are not primarily focused on the demands of public sector. Instead, they come from industry backgrounds, and or concentrate on assessing data quality and distribution patterns, or on problem diagnostic analysis.

2. Methodology

In this paper, we used two main methods: Design Science Research (DSR) and Exploratory Data Analysis (EDA). DSR was the method used to conduct the process of adapting and transforming EDA into a model that could be applied to government and public policy in the most automated and systematic way possible. In the following subsections, we present both methods, describe how we combined them, and detail the materials used for testing and validation.

2.1. Design Science Research (DSR)

Design Science Research (DSR) is a practical methodology for systematizing the creation of new artifacts and solutions to real-life problems, by focusing on understanding the problem, mapping the available alternatives, and devising new ways of addressing it (Rodrigues, 2018).

In an extensive literature review on the subject, Carneiro and Almeida (2019) argue that DSR is an appropriate method for problems that require adaptive and iterative solutions, since it not only focuses on practical challenges but also provides a process for updating and improving previous solutions within the same class of problems.
It is also important to highlight that some studies suggest DSR is a robust method for addressing public policies, since it is suitable for developing solutions to real problems as well as for interpreting and navigating complexity (Santos, Koerich, & Alperstedt, 2018).
According to Wiering (2009), the application of DSR can follow a four-step cycle, called the “Regulative Cycle”: (1) problem investigation, (2) solution design, (3) design validation, and (4) solution implementation. This cycle makes it possible to move from the problem to the solution through an intermediate design step, which systematizes knowledge and clarifies the intentionality of the solutions. The following figure summarizes the cycle, adapted from two original figures in the referenced article:
:

Figure 1. Summary of the Regulative Cycle.

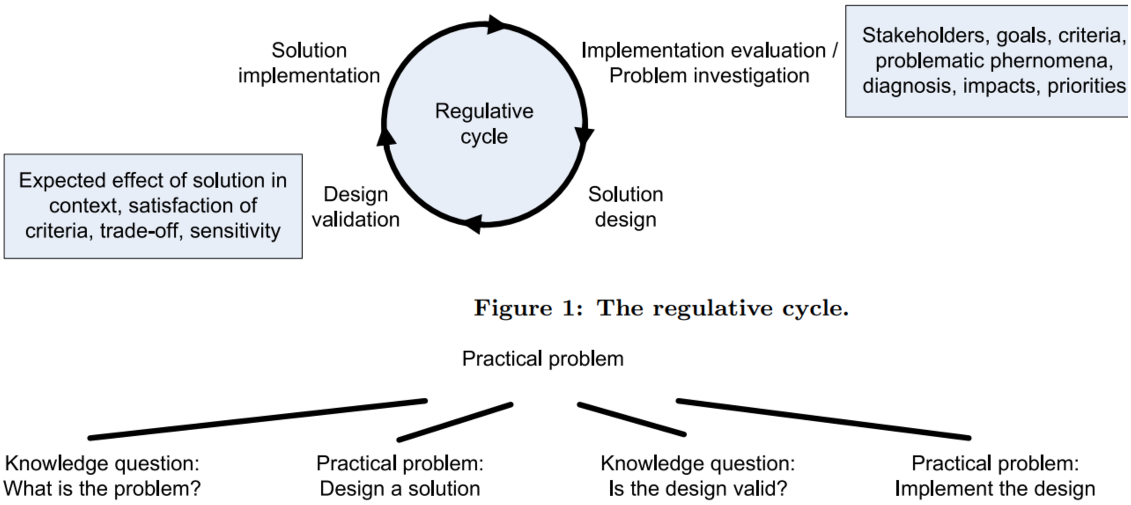


Figure 2: A simple decomposition of practical problems into subproblems.
Source: Wiering (2019).

Based on this version of the cycle, we conducted the research through the following steps: (1) problem investigation, grounded in the literature on autoEDA for public policy, resulting in a problem statement; (2) solution design, by mapping the available autoEDA solutions, since they belong to the same problem class as ours; (3) design validation, by exploring the ProUni datasets; and (4) solution implementation, carried out within TED No. 13,738. The following subsections detail each step of the procedure.

2.2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a method for analyzing data first proposed by John Tukey in the late 1970s and early 1980s. Its primary motivation was to strengthen and better define exploratory approaches. As Tukey emphasized in his seminal works (1977, 1980), statisticians at that time had made significant advances in confirmatory analysis—the type of analysis focused on inference and hypothesis testing.
According to Tukey, this progress ultimately created a scenario in which confirmatory analysis came to be regarded as more relevant and necessary. However, he argued that exploratory analysis remains the foundation from which research design and the questions that guide confirmatory analysis emerge.
Tukey’s main points were: (1) we will always rely on exploratory analysis to better understand the data, to generate ideas, to design further analyses, and—perhaps most importantly—to ask better questions of the data; and (2) it is necessary to establish procedures and define principles for exploring data, using both statistical metrics and graphical representations, to ensure a certain degree of replication and comparability.
Following Tukey, several authors contributed to the development of Exploratory Data Analysis (EDA). One of the most relevant works was by Behrens (1997), who reviewed different studies on the subject. Among other findings, the author identified that, despite variations in techniques and approaches, the EDA literature shares common ground in its attempt to understand data by searching for and representing patterns.

Another significant advance has been the automation of EDA through computer science. More recently, Staniak and Biecek (2019) mapped the landscape of automated exploratory data analysis (autoEDA) in the R language, identifying 15 different software packages.
Their mapping method was derived from the CRISP-DM process (Wirth, 2000), which relates EDA to the phases of business understanding, data understanding, and parts of data preparation.
According to the authors, autoEDA encompasses procedures for data description (e.g., identifying the variables), data validity (invalid, missing, and atypical values), data exploration (distributions and relationships), as well as data cleaning and attribute derivation.
.

Figure 2. Summary of autoEDA packages in R landscape

Task type	Task	a	aE	DE	dM	d	EPD	e	eR	fM	i	R	SE	s	v	x
Dataset	Variable types		x	x	x	x		x		x	x		x	x	x	
	Dimensions		x	x	x	x	x			x	x		x		x	
	Other info			x							x				x	
	Compare datasets	x								x	x				x	
Validity	Missing values		x	x	x	x	x	x		x	x		x	x	x	x
	Redundant col.		x		x	x		x		x			x	x	x	
	Outliers		x		x	x	x			x						
	Atypical values				x					x						x
	Level encoding				x											
Univar.	Descriptive stat.		x		x	x	x	x		x	x	x	x	x		x
	Histograms		x	x	x	x	x	x		x	x	x	x	x		
	Other dist. plots			x			x									
	Bar plots		x	x	x	x	x	x		x	x	x	x	x		
	QQ plots			x		x							x			
Bivar.	Descriptive stat.	x				x		x				x	x	x		
	Correlation matrix			x		x	x				x				x	
	1 vs each corr.		x						x	x			x			
	Time-dependency	x					x									x
	Bar plots by target		x	x		x	x	x		x		x	x			
	Num. plots by target		x			x	x	x		x			x			
	Scatter plots			x			x		x			x	x			
	Contigency tables	x				x							x	x		
	Other stats. (factor)									x	x		x			
Multivar.	PCA			x												
	Stat. models	x					x	x								
	PCP												x			
Transform.	Imputation			x		x		x								
	Scaling					x			x	x						
	Skewness					x										
	Outlier treatment					x	x	x		x						
	Binning			x		x				x						
	Merging levels			x						x						
Reporting	Reports		x	x	x	x		x				x	x			
	Saving outputs	x							x	x				x		x

Table 4: Overview of functionalities of all described packages. Package names were shortened to make the table as compact as possible. **a** denotes **arsenal**, **aE - autoEDA**, **DE - DataExplorer**, **dM - dataMaid**, **d - dlookr**, **EPD - ExPanDaR**, **e - explore**, **eR - exploreR**, **fM - funModeling**, **i - inspectdf**, **R - RtutoR**, **SE - SmartEDA**, **s - summarytools**, **v - visdat**, **x** denotes **xray**. *Num. plots by target* refers to either histogram, density, violin or box plot.

Source: Staniak & Biecek (2019).

As shown in the previous table, most available packages focus primarily on describing and evaluating datasets, while the exploratory phase itself emphasizes how variables are distributed and how they relate to one another.
Although relevant, these analyses remain insufficient for public policy activities such as planning, monitoring, and evaluation, since they neither address indicator constraints directly nor provide ways of describing data in forms more meaningful to decision makers.

2.3. Materials

To conduct the testing and validation phases of the proposal, we used three main datasets: (a) ProUni Offered Scholarships, (b) ProUni Awarded Scholarships, and (c) the Higher Education Census. With the exception of dataset (c), the other datasets were obtained through TED No. 13,738. The following table summarizes the datasets.

Table 1. Summary of datasets.

Dataset	***Description***
ProUni Offered Scholarships	Information on ProUni scholarships offered by semester and year (time), and by course and campus (unit). Indicators include type (in-person or distance education) and modality (partial or full). This dataset was used to gather basic information on scholarship supply.
ProUni Awarded Scholarships	Information on ProUni scholarships awarded by semester and year (time), and by course and campus (unit). Indicators include student trajectory (semesters attended, etc.), course fees, and scores on the Exame Nacional do Ensino Médio – ENEM (Brazilian National High School Exam). Because this dataset focuses on awarded scholarships, it provides socioeconomic information at the individual level.
Higher Education Census	Information on higher education by year (time), and by course and campus (unit). This is a very large dataset with numerous dimensions and indicators covering institutional structure, staff, and more. It was used to obtain detailed information on courses, campuses, and institutions.

Source: The authors.

2.4. The Problem

The public sector is only beginning to utilize the large volumes of data generated by public policies and public services. As a result, the data ecosystem faces a situation in which governments are major players in producing, owning, and distributing data, but not in analyzing, exploring, and making decisions based on it (Filgueiras & Lui, 2023; Giest, 2017; Nielsen, Persson, & Madsen, 2019).

A major limitation is the absence of a comprehensive framework or model for conducting automatic and systematic exploratory data analysis in the public sector. While existing models are useful for tasks such as data description and data quality assessment, they remain limited for deeper exploration, as they primarily focus on univariate, bivariate, and multivariate distribution analysis, as shown in Figure 2. By contrast, autoEDA in public policy relies primarily on indicators, which require exploration across the dimensions of time, space, and social segmentation—areas generally overlooked by conventional EDA models.

In sum, although exploratory data analysis has a well-established tradition and multiple automated tools are available, the public sector still lacks a model tailored to its specific needs. General EDA models are useful for describing data and assessing quality, but they fall short of meeting the core requirements of public policy analysis, which demand indicator-based exploration across time, space, and social segmentation. This gap underscores the need for methodological adaptations that enable governments not only to produce and distribute data but also to systematically explore and apply it for evidence-based decision-making.

2.5. Problem Statement

The problem statement was developed using the How Might We? (HMW) framework (Brown, 2020) to produce a statement that is both solvable and open enough to allow innovative solutions: How might we create an exploratory data analysis model applicable to public sector data that (1) is compatible with planning, monitoring, and evaluation indicators, and (2) establishes replicable procedures?

2.6. Alternatives

The public sector landscape of data science teams is highly diverse and flexible, employing different frameworks, models, and tools that range from data gathering to data modeling, as well as visualization and representation

(Brady, 2019). According to Medeiros et al. (2025), the most common framework used by these teams is CRISP-DM—a general-purpose set of concepts and tools originally designed for developing machine learning predictive models in industry.

Nevertheless, the main solutions available in this landscape are business intelligence (BI) tools. These tools have been applied in strategic planning, decision-making, policy design, and especially in the monitoring and evaluation of public policies, as they enable rapid and integrated analysis of data and indicators (Vaitsman, Rodrigues, & Paes-Souza, 2006). They also allow the timely identification of weaknesses and limitations for corrective action (Cardoso Júnior, 2015) by leveraging performance indicators (Costa, 2012) and data integration for decision-making (Pereira et al., 2022).

Although EDA has been an established approach since the pioneering work of John Tukey (1977, 1980), and although multiple models and tools for autoEDA exist—such as the rich ecosystem of R packages mapped by Staniak and Biecek (2019)—to the best of our knowledge, this level of discussion and development has not yet reached the public sector.

Thus, none of the current alternatives is sufficient to enable automated and systematic analysis of public policies, as detailed in the following summary table

.

Table 2. Summary of alternatives.

Alternative	***Limitation in face of the problem statement***
CRISP-DM	Focus on machine learning for industry. Its main purpose is to create predictive models rather than provide broader analyses of public policy data.
Business Intelligence	Focus on creating dashboards for descriptive and diagnostic purposes. Although widely used—especially for monitoring—it falls short in terms of data quality and more detailed analysis.
autoEDA	Focus on creating automated exploratory data analysis and covering several relevant steps to understand data. However, it does not take into account the specificities of public policy analysis. Nevertheless, due to its characteristics, it was chosen as the solution to be further developed.

Source: The authors.

3. Results

In this section, we present the model along with some snapshots of its application using the datasets mentioned in the methodology. As noted earlier, the alternative chosen as the base model to be improved was autoEDA, since it most closely resembles the proposed solution and its desired characteristics. The first point presented is the adaptation of the traditional autoEDA steps table—based on Staniak & Biecek (2019).

It is important to emphasize that the cases presented below are primarily based on publicly available databases. To illustrate the procedure we followed, we used only aggregated information from the ProUni Offered Scholarships. In other words, no individual-level data provided by SESu/MEC are included; instead, we rely primarily on data from the Higher Education Census.

Table 3. Summary of changes on better available alternative.

Task	***autoEDA***	***Public Policy Analysis Model***
Dataset	Variable types, dimensions, other information, and dataset comparison.	Changes • *Focus dimensions on indicators, separating them from other dimensions; and* • *Directly compare datasets to identify the level of granularity (time and unit) that allows them to be joined.*
Validity	Missing Values, Redundant Cols., Outliers, Atypical Values, Level Encoding.	No changes.

Univar., Bivar. and Multivar.	Descriptive Stats., Histograms, Other Dist. Plots, Bar Plots, Qq Plots, Correlation Matrix, 1 vs Each Corr., Time-Dependency, Bar Plots by Target, Num. Plots by Target, Scatter Plots, Contingency Tables, Other Stats.	Distribution statistics and plots are less useful than: • *Correlation statistics and scatter plots, to analyze possible patterns and interactions*; • *Contingency tables and plots by target, focusing on how indicators interact with dimensions of interest for public policy;* • *Time-dependency analyses, aimed at identifying indicator trends and segmented time series*. Plots by target with map representations, since most decisions are made using targets such as cities, states, and regions. ABC curves or similar techniques, which are particularly useful in large datasets for targeted prioritization.
Transform.	Imputation, Scaling, Skewness, Outlier Treatment, Binning	The transformation stage is possibly the most suitable for modifications, since transformation tasks tend to focus on preparing data for confirmatory and predictive analysis. However, because we are working with indicators, techniques such as population-based rates and proportional ratios are more useful for addressing transformation problems, as they align more closely with the unit characteristics of the dataset.
Reporting	Reports, Saving Outputs.	No changes.

Source: The authors, based on the original table at Staniak & Biecek (2019).

As the previous table makes clear, the most relevant suggested changes are: (1) focus dataset analysis on indicators and on identifying the granularity level; (2) use targeted prioritization on large datasets; (3) use population-based rates and proportional ratios in transformation tasks. To better illustrate these changes, we now provide snapshots of applications on the datasets. We start with dataset quality analysis focused on indicators, showing that separating indicators from contextual socioeconomic dimensions may lead to a better understanding of the dataset.

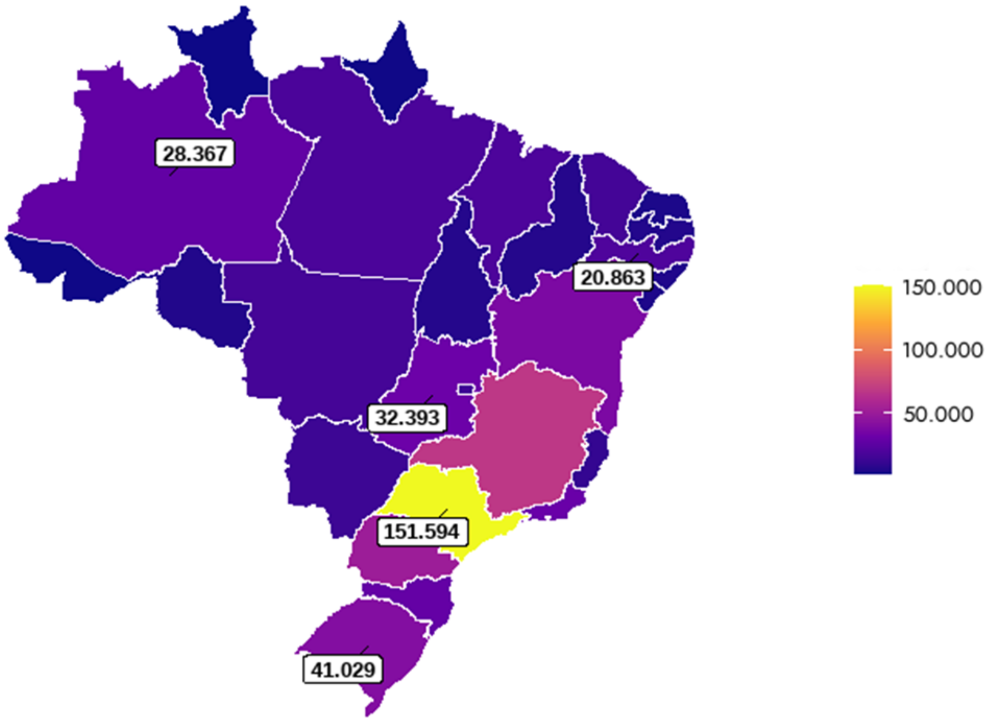
Table 4. Dataset quality analysis with focus on indicators.

Indicator	*Full Presential Scholarships*	*Full Remote Scholarships*	*Partial Presential Scholarships*	*Partial Remote Scholarships*
Missing	0	0	0	0
% Complete	1	1	1	1
Standardized sd	1,777081	1,49277	2,016476	0,358218
sd	5,162231	5,069922	12,73062	3,920121
Min.	0	0	0	0
Q1	0	0	0	0
Med.	0	1	0	0
Q3	2	2	0	0
Max.	352	1.100	1.478	1.467
Hist.	▇▁▁▂	▇▁▁▂	▇▁▁▂	▇▁▁▂

Source: The authors.

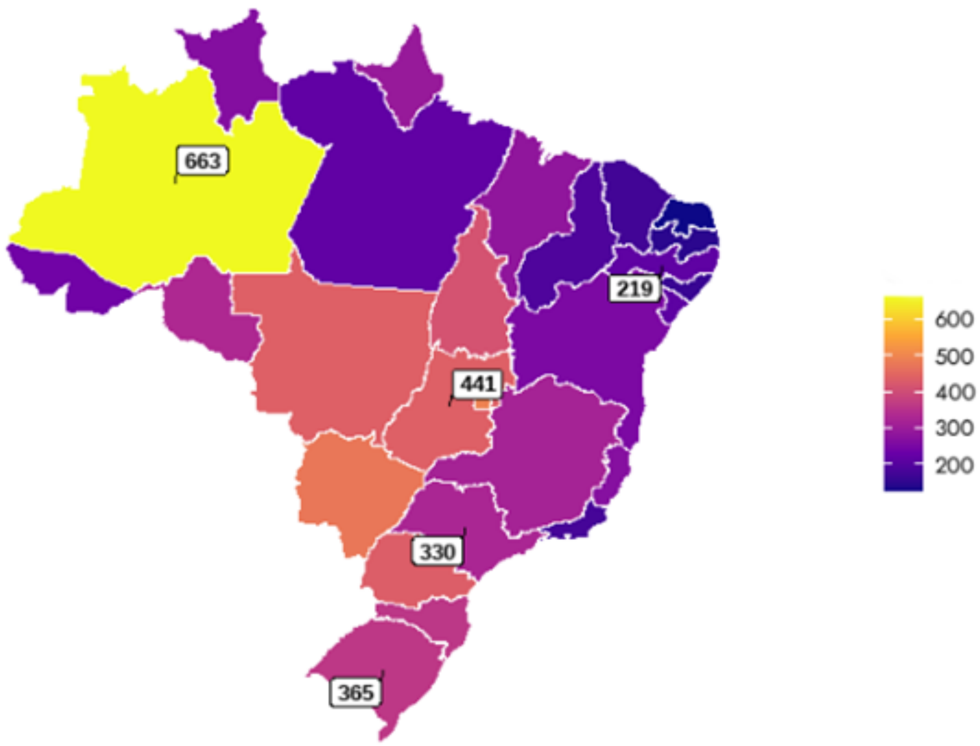
The second relevant snap is a set of maps that implement suggested changes in a) Univar., Bivar. and Multivar. tasks, by using state as target to better represent and understand indicators distribution; and b) in Transform. tasks, with both population-based rate and proportional ratio.

Figure 3. State map of offered scholarships.



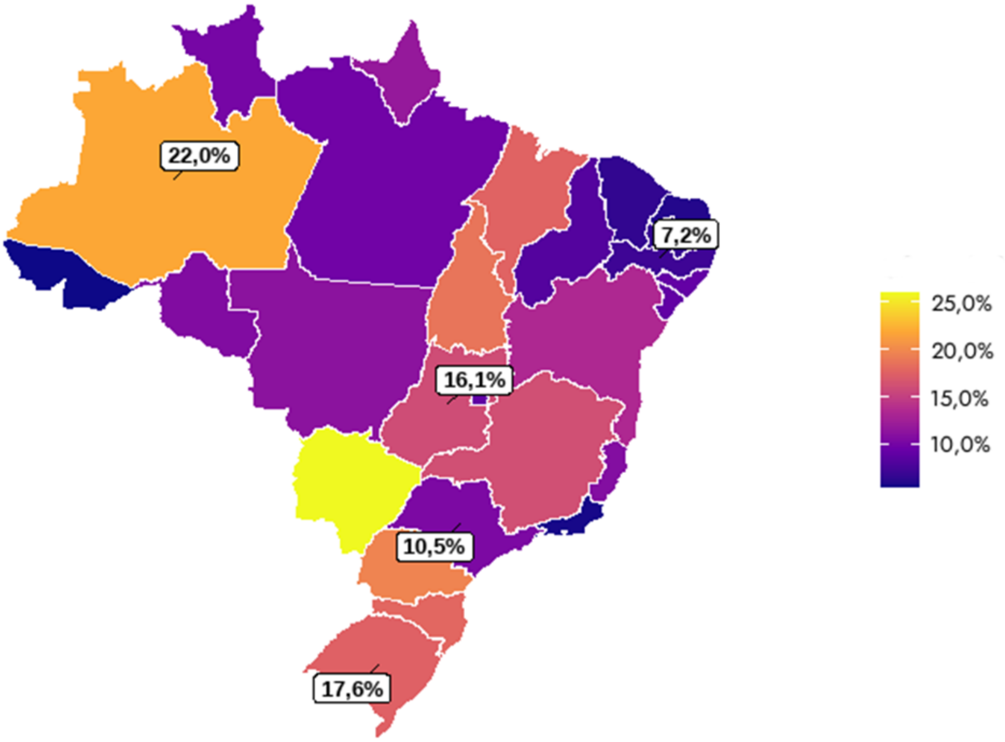
Source: The authors.

Figure 4. State map of offered scholarships (Rate per 100,000 population).



Source: The authors.

Figure 5. State map of offered scholarships (% of total private higher education enrollment).



Source: The authors.

As we can see in the maps, the use of population-based rates and proportional ratios to transform the indicators produces visualizations that avoid simply highlighting the states offering the largest absolute number of scholarships. Instead, they provide different insights, allowing us to identify states that grant a high number of scholarships relative to their populations, as well as states where scholarships represent a large share of total available spots in private higher education. Finally, the transformation tasks can be carried out with techniques more closely aligned to the characteristics of public policies, avoiding unnecessarily complex methods that may add difficulty to data analysis and storytelling.

4. Conclusion

This paper presents a model for public policy analysis that adapts common automated exploratory data analysis (autoEDA) techniques to the specific requirements of public policy, focusing on indicators, policy planning, monitoring, and evaluation, and is developed through a design science research approach. approach.

The proposed model enhances standard autoEDA tasks by adapting them to the constraints of public policy analysis, particularly in the steps of dataset quality assessment, target analysis in univariate, bivariate, and multivariate contexts, and the use of more appropriate transformations such as population-based rates and proportional ratios.

An important limitation of the model is that it was applied to a single dataset from a specific higher education affirmative action policy, and therefore needs to be tested with data from other policy areas such as health, mobility, and other domains—or even with additional educational datasets.

Another limitation is that we were not able to incorporate the analysis into automated business intelligence tools and workflows, which would add value by making the process more dynamic and suitable for drill-down analysis. However, we believe this is a strong start toward constructing a novel model that adapts autoEDA to the context of public policy analysis, which, with further refinement and development, could eventually be formatted into new packages for the business intelligence and autoEDA landscape.

References

1. Behrens, J. T. (1997). Principles and procedures of Exploratory Data Analysis. *Psychological Methods*, 2(2), 131-160.
2. Brady, H. E. (2019). The challenge of big data and data science. *Annual Review of Political Science*, 22(1), 297-323.
3. Brasil. (2004a). Projeto de Lei nº 3.582, de 18 de maio de 2004. Dispõe sobre a instituição do Programa Universidade para Todos – PROUNI, e dá outras providências. Câmara dos Deputados
4. Brasil. (2004b). Medida Provisória nº 213, de 10 de setembro de 2004. Institui o Programa Universidade para Todos – PROUNI, regula a atuação de entidades beneficentes de assistência social no ensino superior, e dá outras providências. Diário Oficial da União.
5. Brasil. (2005). Lei nº 11.096, de 13 de janeiro de 2005. Institui o Programa Universidade para Todos (PROUNI), regula a atuação de entidades beneficentes de assistência social no ensino superior; altera a Lei nº 10.891, de 9 de julho de 2004, e dá outras providências. Diário Oficial da União.
6. Brasil. (2013). Instrução Normativa RFB nº 1.394, de 12 de setembro de 2013. Dispõe sobre a isenção do Imposto sobre a Renda e de contribuições aplicável às instituições que aderirem ao Programa Universidade para Todos. Diário Oficial da União.
7. Brasil. (2022). Lei nº 14.350, de 25 de maio de 2022. Altera as Leis nº 11.096, de 13 de janeiro de 2005, e 11.128, de 28 de junho de 2005, e a Lei Complementar nº 187, de 16 de dezembro de 2021, para aperfeiçoar a sistemática de operação do Programa Universidade para Todos (PROUNI). Diário Oficial da União
8. Brown, T. (2020). *Design thinking: Uma metodologia poderosa para decretar o fim das velhas ideias*. Elsevier.
9. Cardoso Júnior, J. C. (2015). *Monitoramento estratégico de políticas públicas: requisitos tecnopolíticos, proposta metodológica e implicações práticas para a alta administração pública brasileira*. IPEA.
10. Carneiro, L., & Almeida, M. (2019). Design Science: Research in a Theorical Field. *Brazilian Journal of Information Studies*: Research Trends, 13(3), 68-80.
11. Catani, A. M., & Gilioli, R. de S. P. (2005). O PROUNI na encruzilhada: entre a cidadania e a privatização. Linhas Críticas, 11(20), 55–68.
12. Costa, S. A. R. da. (2012). *Sistema de Business Intelligence como suporte à gestão estratégica* (Dissertação de mestrado, Universidade do Minho, Escola de Engenharia). Universidade do Minho.
13. De Toni, J., & Dorneles, R. (2022). *Ciência de dados em políticas públicas: uma experiência de Formação*. ENAP.
14. Filgueiras, F., & Lui, L. (2023). Designing data governance in Brazil: An institutional analysis. *Policy Design and Practice*, 6(1), 41-56.
15. Francis, A., & Tannuri-Pianto, M. (2012). The Redistributive Equity of Affirmative Action: Exploring the Role of Race, Socioeconomic Status, and Gender in College Admissions. *Economics of Education Review*, 31(1), 45- 55.
16. Frey, K. (2000). Políticas Públicas: um debate conceitual e reflexões referentes à prática da análise de políticas públicas no Brasil. *Planejamento e Políticas Públicas*, 21(1), 212-259.
17. Gertler, P., Martinez, S., Rawlings, L. B., Premand, P., & Vermeersch, C. M. J. (2018). *Avaliação de Impacto na Prática*. Washington: Banco Internacional para Reconstrução e Desenvolvimento / Banco Mundial.
18. Giest, S. (2017). Big Data for Policymaking: Fad or Fast Track?. *Policy Sciences,* 50 (3), 367–382.
19. Greco, S., Ishizaka, A., Tasiou, & M., Torrisi, G. (2019). On the methodological framework of composite indices: a review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, 141(1), 61–94.
20. Guimarães, J. R. S., & Jannuzzi, P. M. (2004). Indicadores Sintéticos no processo de formulação e avaliação de políticas públicas – limites e legitimidades. *Encontro Nacional de Estudos Populacionais*, 14.

21. Howlett, M., Ramesh, M., & Perl, A. (2013). *Política Pública: seus ciclos e subsistemas*. Elsevier.

22. Instituto de Pesquisa Econômica Aplicada – IPEA. (2018). *Avaliação de políticas públicas: guia prática de análise ex ante – Volume 1*. IPEA.

23. Instituto de Pesquisa Econômica Aplicada – IPEA. (2018). *Avaliação de políticas públicas: guia prática de análise ex post – Volume 2*. IPEA.

24. Jannuzzi, P. M. (2002). Considerações sobre o uso, mau uso e abuso dos indicadores sociais na formulação e avaliação de políticas públicas municipais. *RAP*, 36(1), 51-72.

25. Kusek, J. Z., & Rist, R. C. (2004). *Ten steps to a results-based monitoring and evaluation system*. Washington: Banco Mundial.

26. Lotta, G. (2019). Teoria e análises sobre implantação de políticas públicas no Brasil. In G. Lotta (Org.), A *política pública como ela é: Contribuições dos estudos sobre implementação para a análise de políticas públicas*, 11-38. Enap.

27. Nardo, M. et al. *Handbook on constructing composite indicators: methodology and user guide*. OCDE.

28. Nielsen, O.B., Persson, J.S., & Madsen, S. (2019). Why Governing Data Is Difficult: Findings from Danish Local Government. In: Elbanna, A., Dwivedi, Y., Bunker, D., Wastell, D. (eds) Smart Working, Living and Organising. TDIT 2018. IFIP *Advances in Information and Communication Technology*, vol 533. Springer.

29. Pereira, J. G., Koch, L. L., Krause, M. G., Dandolini, G. A., & Souza, J. A. de. (2022). Aplicação do Business Intelligence no setor público. RECIMA21 – Revista Científica Multidisciplinar, 3(8), 1-15.

30. Rodrigues, D. R. (2018). Design Science Research como caminho metodológico para disciplinas e projetos de Design da Informação. *Brazilian Journal of Information Design*, 15(1), 111-124.

31. Rogger, D., & Schuster, C. (2023). *The Government Analytics Handbook: Leveraging Data to Strengthen Public Administration*. World Bank.

32. Saltelli, A. Composite indicators between analysis and advocacy. (2007). *Social Indicators Research*, 81(1), 65–77.

33. Santos, G., Koerich, G., & Alperstedt, G. A contribuição da design research para a resolução de problemas complexos na administração pública. *RAP*, 52(5), 956-970.

34. Souza, C. (2006). Políticas Públicas: uma revisão de literatura. *Sociologias*, 8(16), 20-45.

35. Souza, G., & Loreto, M. (2021). Avaliação e monitoramento de políticas públicas: produção acadêmica em periódicos científicos brasileiros (2010 a 2020). *Meta: Avaliação*, 13(40), 649–673.

36. Staniak, M., & Biecek, P. (2019). The Landscape of R Packages for Automated Exploratory Data Analysis. *The R Journal,* 11(2), 347-369.

37. Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company.

38. Tukey, J. W. (1980). We Need Both Exploratory and Confirmatory. *The American Statistician,* 34(1), 23-25.

39. Vaitsman, J., Rodrigues, R. W. S., & Paes-Sousa, R. (2006). O sistema de avaliação e monitoramento das políticas e programas sociais: A experiência do Ministério do Desenvolvimento Social e Combate à Fome do Brasil (Policy Papers, 17). UNESCO.

40. Vasconcelos Medeiros, H. A., Wanderley, A. L., de Carvalho Bezerra, D., Gomes de Amorim Filho, C. A., Zimmerle da Nóbrega, R., & de Moraes Ferreira, F. G. (2025). Government Data Science Teams: A Framework for Implementing Strategic Monitoring Solutions. *Conference on Digital Government Research*, 26.

41. Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.

42. Wieringa, R. (2009). Design science as nested problem solving. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, 1-12.

Young, E., & Quinn, L. (2002). *Writing effective public policy papers: A guide for policy advisers in Central and Easte*