Utilização da análise de dados como ferramenta para aprimorar a transparência no Tribunal Superior Eleitoral (TSE).

Wallynson Rodrigo Henrique da Silva, Awdren de Lima Fontão

Faculdade de computação – Universidade Federal de Mato Grosso do Sul (UFMS) – 79070-900 – Campo Grande – MS – Brazil

w.rodrigo@ufms.br

Abstract. Transparency is always a topic discussed when the subject is the Superior Electoral Court (TSE). This study and the development of a simple BI application related to data on the national electorate, is intended to demonstrate how the use of data analysis can help access court data and its understanding.

Resumo. A transparência é sempre um tema discutido quando o assunto é o Tribunal Superior Eleitoral (TSE). Com esse estudo e o desenvolvimento de uma simples aplicação de BI relacionada aos dados sobre o eleitorado nacional, pretende-se demonstrar como o uso da análise de dados pode auxiliar o acesso aos dados do tribunal e seu entendimento.

INTRODUÇÃO

É importante mencionar que esse trabalho não objetiva atacar ou descredibilizar nenhuma instituição do estado brasileiro perante a sociedade. É um trabalho técnico cujo objetivo é demonstrar como a transparência pode ser favorecida com o uso de técnicas de análise e visualização de dados.

Transparência é uma palavra de origem da língua latina e relacionada ao verbo *transparere*, que significa "mostrar a luz através, deixar a luz atravessar". No contexto do presente trabalho, ela poderia ser considerada algo como "dar acesso sem barreiras a". Ela não está explicitada entre os princípios listados no art. 37 da constituição federal do Brasil, mas está intimamente ligada ao princípio da publicidade.

Na última década, inclusive, foram criadas boas iniciativas para aumentar a transparência nos poderes públicos de todas as esferas do estado brasileiro. Pode-se citar, por exemplo, iniciativas como portais da transparência e a lei de acesso à informação.

O objetivo deste estudo é demonstrar como a aplicação de algumas técnicas de análise e visualização de dados já são capazes de facilitar o acesso e entendimento da sociedade às informações públicas.

Provost e Fawcett (2013) afirmam que

> “os últimos quinze anos testemunharam grandes investimentos em infraestrutura de negócios que têm melhorado a capacidade de coletar dados [...]. Agora , praticamente todos os aspectos dos negócios estão abertos para a coleta de dados. [...] Essa ampla disponibilidade de dados levou ao aumento do interesse em métodos para extrair informações úteis e conhecimento a partir de dados.”

Para demonstrar a importância dessa extração de dados, Provost e Fawcett (2013, p. 3) trazem o exemplo do furacão Frances, em 2004, na Flórida, Estado Unidos. Eles contam que por conta da chegada do furacão, executivos do Walmart viram que essa seria uma boa oportunidade para extraírem informações valiosas de seus dados. Eles utilizaram os trilhões de bytes de histórico de compras para tentar prever o que aconteceria. Com isso, descobriram que precisariam estar abastecidos de diversos outros produtos (e não apenas as costumeiras lanternas), e, também, que, no último furacão, a demanda por Pop-Tarts de morango foi sete vezes maior e que o produto mais vendido foi cerveja.

O trecho acima mostra como é importante buscar essa extração de informações e conhecimento, que é a análise de dados. Silva (2022) a define como

> “o processo de aplicação de técnicas estatísticas e lógicas para avaliar informações obtidas a partir de determinados processos. O principal objetivo da prática é extrair informações úteis a partir dos dados. A partir destas informações, é possível tomar decisões mais assertivas e orientadas para resultados.”

Esse processo necessita de uma metodologia para ser colocado em prática, conhecida como *pipeline*. De acordo com Biswas, Wardat e Rajan (2022), um *pipeline* de ciência de dados é “uma série de estágios de processamento que interagem com dados (geralmente aquisição, gerenciamento, análise e raciocínio)”. Dizem ainda que “um pipeline de ciência de dados pode consistir em vários estágios e conexões entre eles. Os estágios são definidos para executar tarefas específicas e conectados a outros estágios, com relações de entrada-saída.”. Esse conceito é explicitado na figura 1:

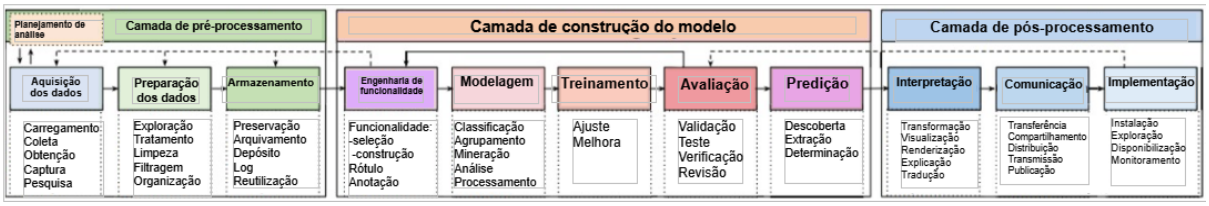

Figura 1. Conceitos em um pipeline de ciência de dados. Fonte: Biswas, Wardat e Rajan (2022)

Tem-se as camadas, os estágios e as subtarefas. As subtarefas estão listadas abaixo de cada estágio. Os estágios estão conectados com *loops* de *feedback*, indicados com setas. As setas sólidas estão sempre presentes no ciclo de vida, enquanto as setas tracejadas são opcionais. Loops de feedback distantes (por exemplo, do *deploy* à aquisição de dados) também são possíveis através do(s) estágio(s) intermediário(s).

Para o caso a que se refere o presente artigo, a camada do construção do modelo não é necessária porque não houve o uso de técnicas de *machine learning* e modelos de inteligência

artificial para se chegar ao resultado final, haja vista que o que se objetiva aqui é apenas demonstrar como técnicas de limpeza e transformação dos dados são úteis para melhorar o acesso a dados de qualidade e que facilitem o acesso a informações.

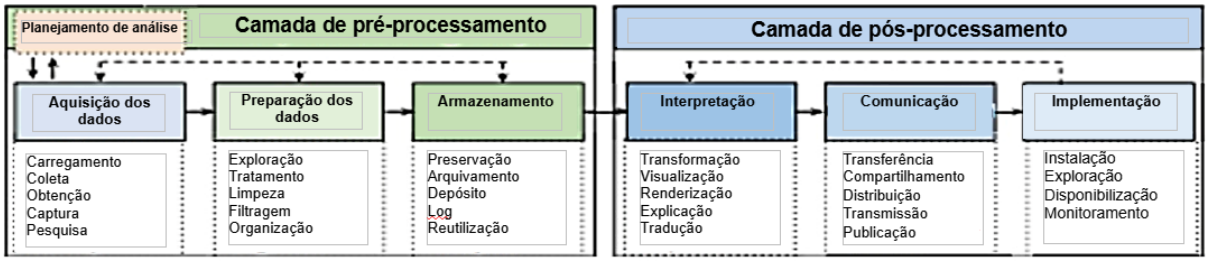

Figura 2. O pipeline utilizado neste artigo. Fonte: adaptado de Biswas, Wardat e Rajan (2022).

Seguindo a estrutura do pipeline da figura 2, este projeto passou pelos estágios de aquisição dos dados, sua preparação, armazenamento, interpretação, comunicação e implementação.

Estes estágios serão explicados a seguir:

- ***aquisição dos dados:*** os dados foram extraídos do site do TSE (Tribunal Superior Eleitoral) (https://dadosabertos.tse.jus.br/dataset/comparecimento-e-abstencao-2020). Cabe mencionar que a escolha pelos dados eleitorais de 2020 e não de 2022 se deu por conta do segundo turno. Se fosse utilizado o *dataset* de 2022 não haveria mudança nos dados mais genéricos (número de eleitores, eleitores por sexo etc.0 entre o primeiro e o segundo turnos, pelo fato de a eleição de 2022 ter tido disputa pela presidência da república, o que faz com que o segundo turno seja disputado em todo o território nacional. Já no caso do *dataset* de 2020, as eleições foram municipais, com isso, o segundo turno não ocorreu em todo o território nacional, e, por conta disso, tem-se uma diferença considerável para o primeiro turno, o que traz dados bem diferentes para análise;
- ***preparação dos dados:*** aqui, a primeira decisão a ser tomada foi qual linguagem utilizar e, por familiaridade, documentação, desempenho e suporte em fóruns e afins, a escolhida foi Python. Após a escolha da linguagem, foi necessário analisar o *dataset* e decidir como utilizá-lo. O TSE disponibiliza, junto com o csv, um pdf que explica as colunas e está na tabela 1;
- ***armazenamento:*** depois da manipulação e transformação dos dados, foram criados quatro novos *datasets*, que contém informações a respeito dos municípios e estados no primeiro e segundo turnos das eleições de 2020, que foram salvos em arquivos csv e posteriormente disponibilizados em um repositório do Github (https://github.com/wrodrigohs/tcc_si_wrodrigo);
- ***interpretação:*** tendo os *datasets* estaduais e municipais de primeiro e segundo turnos, a interpretação passou a ser facilitada por meio de diversos tipos de gráficos interativos da biblioteca *plotly express*;
- ***comunicação:*** a comunicação dos dados foi feita por meio da biblioteca *streamlit*, que permite a criação de *dashboards*;

- ***implementação:*** a biblioteca *streamlit* permite, ainda, o *deploy* de forma muito simples e com apenas alguns cliques. Basta ter o projeto em um repositório do Github e selecionar a opção *deploy* na página criada.

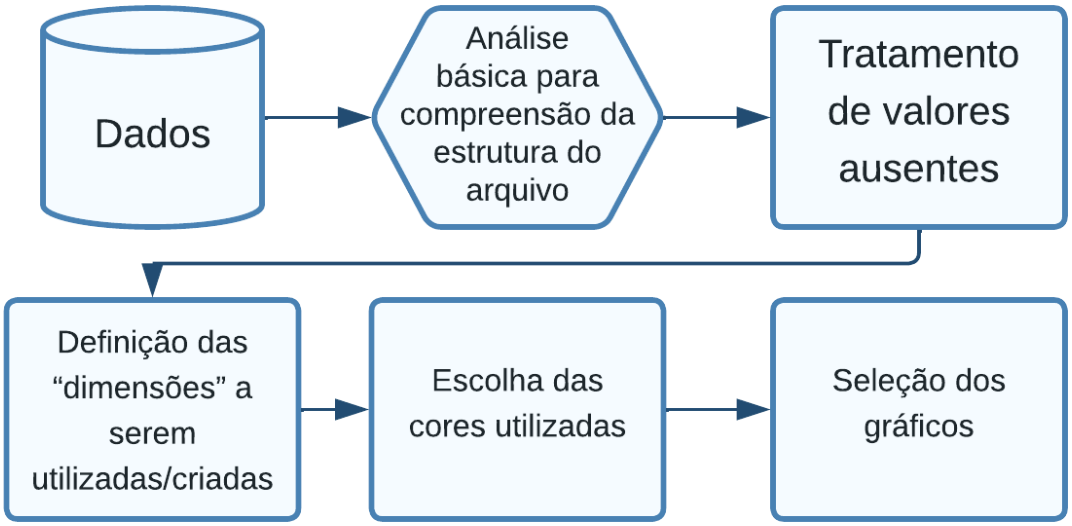

Figura 3. Metodologia utilizada para desenvolvimento do trabalho. Fonte: o autor.

DICIONÁRIO DE DADOS

O TSE disponibiliza juntamente com o arquivo .csv um arquivo .pdf que traz a explicação sobre cada coluna presente no *dataset*. O endereço para download é https://dadosabertos.tse.jus.br/dataset/comparecimento-e-abstencao-2020.

ANÁLISE INICIAL

Após entender quais as colunas e o que significam os dados presente nelas, era necessário saber qual o tamanho do *dataset (*4.770.970 linhas) e fazer o tratamento de valores ausentes..

Não foram encontrados valores nulos ou ausentes, mas há valores definidos como "não informado", "inválido", que não servem ao propósito deste estudo. De acordo com Melo (2019), as abordagens mais usuais para lidar com dados ausentes são: excluir os valores ausentes ou preenchê-los com a mediana ou média (caso seja uma variável numérica) ou com o valor mais frequente (caso seja um valor categórico). Mas ele alerta que esta é uma decisão mais radical, e deve ser feita apenas em casos em que não haverá impacto significativo no modelo. Ao eliminar uma linha inteira, você joga fora um monte de informação que poderia ser extremamente importante.".

Tendo em conta o exposto anteriormente, a decisão tomada foi de excluir esses dados, haja vista que eles representam apenas 2,97% dos dados totais. Após a análise da estrutura do

dataset e exclusão dos dados irrelevantes, o passo seguinte foi deletar algumas colunas desnecessárias para a análise.

Após toda a preparação, o trabalho poderia, enfim, começar. Segundo Nehring & Puppe (2002), quanto maiores as possibilidades de obter subconjuntos a partir de um conjunto de dados, maiores serão as potencialidades de tratamento e de elaboração de resultados. Partindo desse conceito, foram criados outros dois *datasets* (um para o 1º turno e outro para o 2º turno), que originaram os *datasets* df_estados_1turno_2020, df_estados_2turno_2020, df_municipios_1turno_2020 e df_municipios_2turno_2020. Essa possibilidade de analisar os dados por essas quatro óticas é entendida como granularidade. Braga e Menezes (2015) explicam que o termo granularidade originou-se da palavra grão. Metaforicamente, quanto maior a quantidade de 'grãos' de um sistema, maior será a sua granularidade.

Mas, para criar esses conjuntos de dados, foi preciso analisar e definir quais informações poderiam ser extraídas dos "*datasets* fonte" porque eles, a despeito de serem arquivos com muitos dados, traziam poucas informações. A figura 4 demonstra isso.

NM_MUNICIPIO	DS_GENERO	DS_ESTADO_CIVIL	DS_FAIXA_ETARIA	DS_GRAU_ESCOLARIDADE	QT_APTOS
SÃO PAULO	MASCULINO	CASADO	75 a 79 anos	ENSINO FUNDAMENTAL INCOMPLETO	269
SÃO PAULO	MASCULINO	CASADO	85 a 89 anos	ENSINO FUNDAMENTAL INCOMPLETO	78
SÃO PAULO	MASCULINO	CASADO	95 a 99 anos	ENSINO MÉDIO COMPLETO	2
SÃO PAULO	MASCULINO	VIÚVO	30 a 34 anos	ANALFABETO	1
SÃO PAULO	MASCULINO	VIÚVO	40 a 44 anos	ENSINO MÉDIO INCOMPLETO	2
SÃO PAULO	MASCULINO	VIÚVO	70 a 74 anos	ENSINO FUNDAMENTAL COMPLETO	6
SÃO PAULO	MASCULINO	VIÚVO	95 a 99 anos	ANALFABETO	1
SÃO PAULO	MASCULINO	SEPARADO JUDICIALMENTE	35 a 39 anos	ENSINO FUNDAMENTAL INCOMPLETO	4
SÃO PAULO	MASCULINO	SEPARADO JUDICIALMENTE	40 a 44 anos	SUPERIOR INCOMPLETO	3
SÃO PAULO	MASCULINO	SEPARADO JUDICIALMENTE	45 a 49 anos	ENSINO MÉDIO COMPLETO	24

Figura 4. *Printscreen* de uma parte do *dataset* original. Fonte: o autor.

Na imagem se vê que o arquivo traz diversas linhas sobre um mesmo município, dificultando ao usuário do arquivo uma leitura clara dos dados para extrair informações.

Após análise das colunas e das informações que elas trazem ficou definido que os *datasets* finais teriam apenas uma linha por município (ou estado) trazendo as informações de forma mais simplificada e acessível. Com isso foi possível trazer mais claramente as seguintes informações, por município e por estado: eleitores aptos, comparecimento, abstenção (de eleitores com e sem deficiência), o número de eleitores aptos e por sexo (em números absolutos e percentuais), o número de eleitores facultativos aptos e por sexo (em números absolutos e percentuais), o número de eleitores aptos divididos por estado civil (solteiro, casado, divorciado, viúvo e separado judicialmente) por sexo (em números absolutos e percentuais).

Essa definição de quais informações extrair dos dados "brutos" é importante porque, segundo Beck e Libert (2018), falando sobre *data quality*, dados úteis são raros. Eles afirmam, também, que:

> "para usar *machine learning* dessa maneira [rápido, melhor compreendido, mais escalável e menos propenso a erros], o sistema não é alimentado com qualquer dado conhecido, de qualquer campo. Ele deve ser alimentado com um conjunto de

conhecimentos cuidadosamente escolhidos, com a esperança de que o sistema possa aprender e, talvez, estender o conhecimento que as pessoas já têm”

Ronald (2012) afirma que a qualidade dos dados pode ser melhor descrita como dados que atendem aos critérios definidos por uma empresa ou organização e que esses critérios podem depender da empresa ou organização.

Categoria	Descrição
Acurácia	O grau em que os dados testados estão de acordo com uma fonte padrão, que é considerada correta. Na verdade, é uma contagem do número de vezes que um atributo estava incorreto.
Integridade	Os dados não são excluídos inadvertidamente e as alterações feitas são aplicadas. Você pode medir a integridade rastreando o número de problemas registrados ou registros afetados de um call center.
Seguro	Usuários e sistemas que não deveriam ter direitos de acesso são devidamente controlados. Você pode medir isso registrando incidentes, rastreando o número de indivíduos com acesso de gravação ou medindo a conformidade com os padrões de segurança por função.
Completude	O grau em que todos os valores de dados necessários aparecem no registro de dados ou objeto de negócios. Você pode medir isso rastreando a porcentagem de registros com um ou mais atributos obrigatórios ausentes.
Validade	O grau em que os dados testados estão em conformidade com os requisitos de validação de dados. Esta é uma contagem do número de regras em conformidade.
Pontualidade	O grau em que os dados são fornecidos no momento exigido ou especificado. Você pode acompanhar isso comparando a disponibilidade planejada versus a real.
Cobertura	O grau em que a amostra de dados representa com precisão toda a população a ser medida. Isso pode ser testado por meio de estatísticas para determinar se a população da amostra é uma amostra verdadeiramente representativa.
Sem redundância	O grau em que existem várias cópias dos mesmos dados exatos em vários bancos de dados. Para medir isso, você pode rastrear o número de cópias desse registro em toda a sua paisagem. Uma medida mais importante, no entanto, é rastrear o número de aplicativos que criam esses dados, Sistemas de Registro e Sistemas de Destino.
Sem duplicatas	O grau em que há registros duplicados para o mesmo item. Você pode acompanhar isso executando um teste para contar o número de itens duplicados com diferentes identificadores de item.
Relevância	Os dados devem estar vinculados a um processo de negócios, métrica ou objetivo. Isso pode ser difícil de rastrear, mas você pode rastrear atributos em um registro de dados mestre para determinar qual processo ou documento comercial requer esse atributo para determinar a relevância.
Acessibilidade	A facilidade de uso para os consumidores de dados acessarem e consumirem os dados. Existem vários métodos para medir isso, incluindo determinar a velocidade de acesso, pontos de acesso ou até mesmo a qualidade da pesquisa.
Disponibilidade	O grau em que os dados estão disponíveis para acesso aos usuários. Isso pode ser medido pelos tempos em que os dados está disponível para ser usado, planejado versus real.
Consistência	A capacidade de usar ou interpretar os dados com precisão em vários domínios, por exemplo, várias descrições de produto para o mesmo produto em vários aplicativos. Isso pode ser medido verificando o mesmo ID de item em vários sistemas ou domínios quanto a inconsistências ou o mesmo item com diferentes identificadores em vários domínios ou aplicativos.

Figura 5. Aspectos de qualidade dos dados. Fonte: Huang, Lee e Yang (1999)

Ainda segundo Ronald (2012), na prática da vida real, tende-se a adotar uma abordagem mais pragmática e declarar apenas alguns desses aspectos de qualidade aplicáveis aos dados em seu ambiente. Aqui foram observados os aspectos da cobertura, sem duplicatas, relevância, acessibilidade, disponibilidade e consistência.

A qualidade dos dados é fundamental para o processo de ETL (extração, transformação e carregamento, em tradução do inglês) feito no *dataset* original.

PROCESSO DE EXTRAÇÃO, TRANSFORMAÇÃO E CARREGAMENTO (ETL)

Figura 6. Etapas do processo de ETL

Atualmente existem várias ferramentas gratuitas com capacidade para lidar com um grande volume de dados e fazer seu processamento, mas, como já foi dito, optou-se por utilizar apenas *Python* para fazer tanto o processamento (ou transformação) dos dados quanto a sua visualização por conta da familiaridade com a linguagem, sua simplicidade, desempenho e por não haver uma quantidade tão grande de dados com os quais lidar. Estas tecnologias, atualmente, possuem suas arquiteturas bem definidas, principalmente quando focamos nos sistemas de informações gerenciais (Bazzotti e Garcia, 2005) ou o *Business Intelligence* (BI).

A arquitetura de uma aplicação de BI, segundo Kimball et al. (2013) e Inmon (2005), é composta por:

- ***data source***: são as fontes de dados, que podem ser arquivos csv, planilhas do excel, bancos de dados relacionais, sistemas transacionais ou quaisquer outras fontes de dados;
- ***extract, transform and load* (ETL)**: é o processo que extrai dados das fontes e realiza transformações, que podem ser uma agregação, limpeza ou apenas uma adaptação nos dados para que eles estejam de acordo com o que se objetiva fazer com eles;
- ***staging area***: é uma área temporária que armazena os dados antes das transformações, utilizada para que o processo ETL não consuma diretamente a fonte de dados de origem;
- ***operational data store* (ODS)**: é uma área de integração, contém os dados das fontes de dados, porém já transformados e com algumas validações básicas;
- ***visualização de dados:*** é a etapa que possibilita extrair informações dos dados utilizados. Pode-se ser resumida a construção de *dashboards* (painéis gerenciais), que apresentam os indicadores por meio de gráficos, tabelas com marcadores de performance, mapas entre outros recursos.

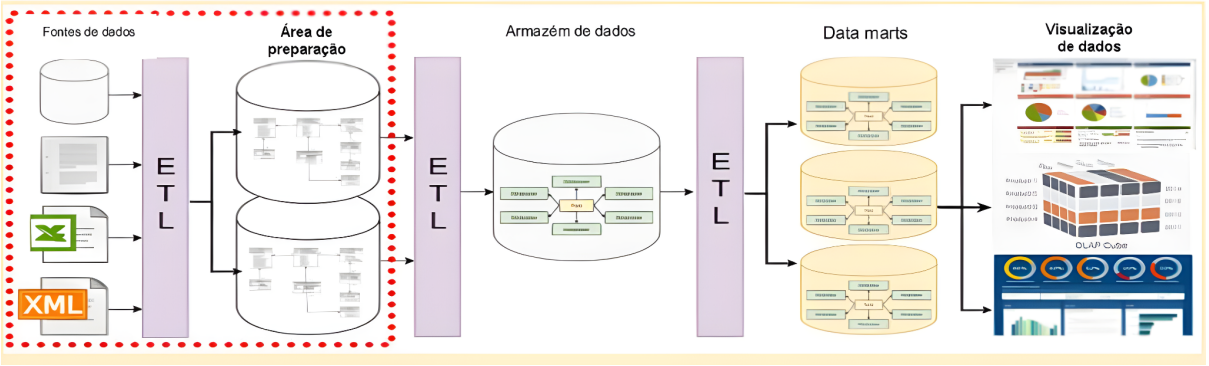

Figura 7. Modelo tradicional de construção de aplicações BI. Fonte: Adaptado de Ferreira (2015)

O *data marts* são apenas pequenos DWs, um aperfeiçoamento feito por Ralph Kimball no modelo de Inmon no início dos anos de 1990, de acordo com Kempe (2012).

Da arquitetura tradicional foram utilizadas apenas as etapas de *data source*, etl, *staging area* e visualização de dados na aplicação de *BI* deste estudo.

TRANSFORMAÇÃO DOS DADOS

Após a importação do *dataset* do TSE foram feitas algumas análises iniciais (mencionadas no começo deste texto), deleção de algumas colunas e de linhas com valores "inválido" e "não informado".

Em seguida foram criados *datasets* para o 1º e 2º turnos, e 1 que será utilizado para armazenar as transformações realizadas.

Esse *dataset* foi criado, a princípio, apenas com as colunas estado e município e excluindo as linhas duplicadas. As outras colunas foram criadas da seguinte maneira:

- As colunas quantitativas (aptos, comparecimento, abstenção, eleitorado masculino e feminino etc.) foram criadas selecionando-se o município e/ou estado e extraindo a quantidade de aptos, comparecimento, abstenção etc. em números absolutos e percentuais.;
- As colunas categóricas escolaridade (analfabeto, lê e escreve, fundamental incompleto e completo, médio incompleto e completo, superior incompleto e completo), estado civil (solteiro, casado, divorciado, viúvo e separado judicialmente) e faixa etária (de 16 anos a 100 anos ou mais) foram tratadas de outra maneira. Foi criado um *dataframe* temporário para pivotar essas categorias e preenchê-las com a quantidade de aptos (absoluta e percentual);
- Eleitorado facultativo: foram somados os valores das colunas 16_anos, 17_anos, 65_69_anos, 70_74_anos, 75_79_anos, 80_84_anos, 85_89_anos, 90_94_anos, 95_99_anos e 100_anos em números absolutos e percentuais para o sexo masculino e feminino;.

Esse processo foi realizado 4 vezes, para criar os *datasets* segmentados por turno e granularidade (municipal e estadual).

VISUALIZAÇÃO DE DADOS

A parte final deste trabalho é o "extrair informações úteis" citado por Provost e Fawcett (2013). Essa extração se dá pela visualização dos dados.

A visualização foi feita seguindo os princípios da psicologia da Gestalt. BOCK (2006) diz que Gestalt é um termo alemão de difícil tradução. O termo mais próximo em português seria forma ou configuração. A Gestalt, segundo Murta, Mont'Alvão e Kosminsky (2020), estuda como se organizam, se estruturam, ou se ordenam as formas psicologicamente percebidas. Os

autores desta escola afirmam que uma forma não é vista sozinha, mas nas suas relações com outras.

Ela é regida por 7 princípios, que são:

- ***Proximidade:*** coisas que estão próximas parecem ser mais relacionadas entre si do que se estivessem distantes;
- ***Similaridade:*** elementos parecidos são percebidos como parte do mesmo grupo e tendo a mesma função;
- ***Continuidade:*** elementos posicionados em uma linha ou curva são percebidos como mais relacionados do que se não estivessem dispostos desta forma;
- ***Fechamento:*** a memória é utilizada para converter objetos complexos em formas simples e/ou já conhecidas;
- ***Figura-fundo:*** nossa percepção instintivamente percebe objetos como estando ou *à frente* ou *ao fundo*;
- ***Região comum:*** quando objetos são posicionados dentro da mesma região fechada estes são percebidos como parte do mesmo grupo;
- ***Ponto focal:*** qualquer elemento que se destacar visualmente vai capturar e prender a atenção de quem está vendo.

Como pode-se ver na imagem a seguir, os princípios acima foram seguidos na elaboração do *dashboard.* As primeiras informações, por exemplo, ficam na parte de cima, em destaque, para já serem vistas pelo usuário.

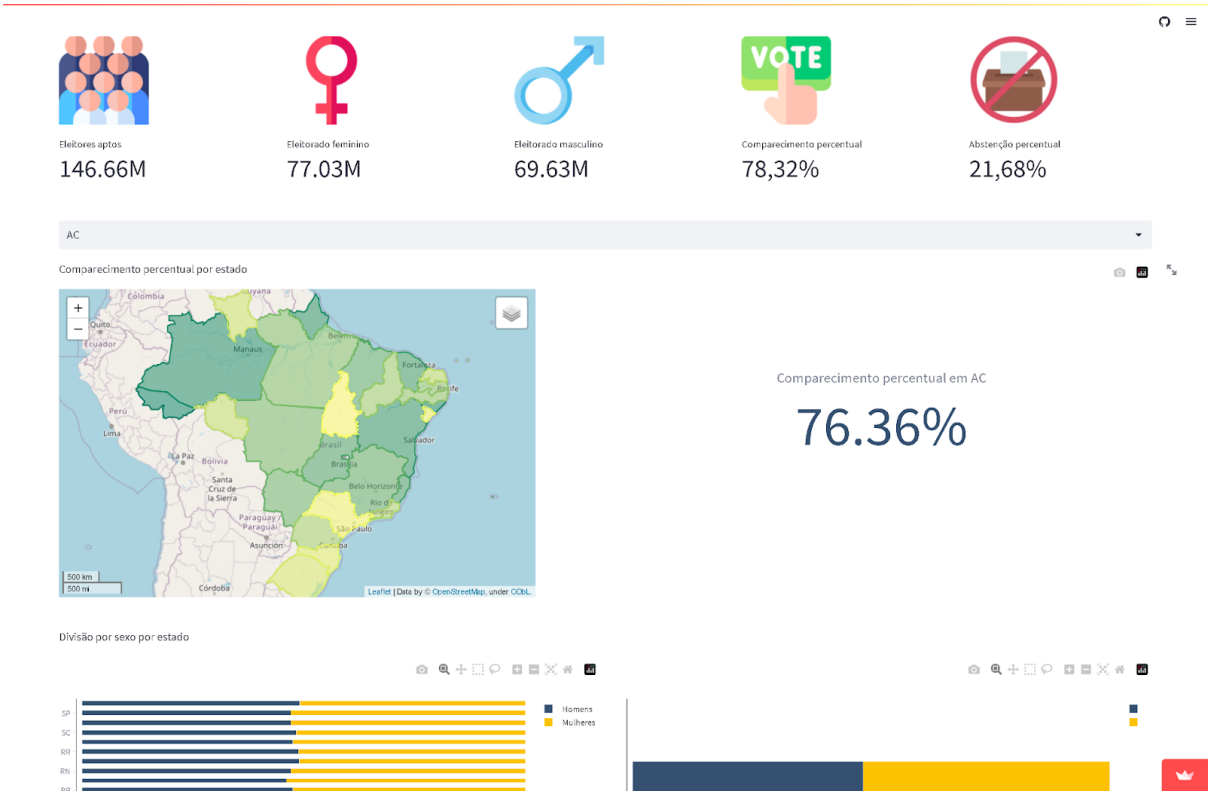

Figura 8. *Printscreen* da aplicação desenvolvida para a visualização dos dados. Fonte: o autor.

Figura 9. *Printscreen* da aplicação desenvolvida para a visualização dos dados. Fonte: o autor.

Figura 10. *Printscreen* da aplicação desenvolvida para a visualização dos dados. Fonte: o autor.

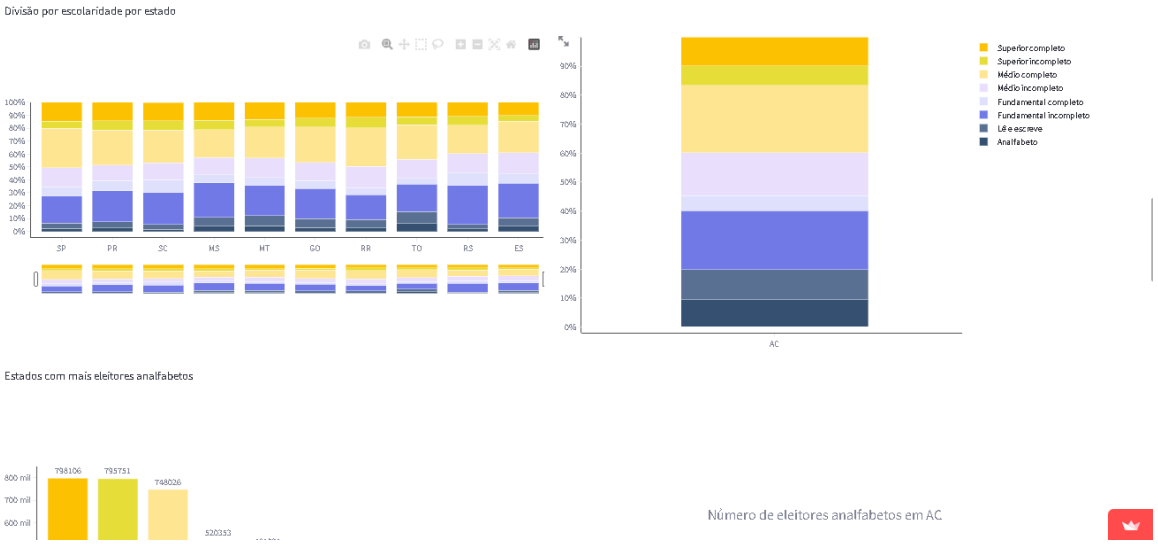

Figura 11. *Printscreen* da aplicação desenvolvida para a visualização dos dados. Fonte: o autor.

Na figura 8 pode-se observar os princípios do ponto focal e continuidade aplicados no *dashboard*. O mapa e o texto informativo sobre o percentual de comparecimento são entendidos como elementos associados.

Já nas figuras 9, 10 e 11 nota-se a utilização de similaridade e região comum.

As cores utilizadas foram escolhidas tomando como base a logomarca do TSE.

A escolha dos gráficos segue as orientações de Knaflic(2018, p. 36) quando ela diz que existem muitos gráficos diferentes e outros tipos de exibições de informação visuais, mas alguns poucos funcionarão para a maioria das suas necessidades.

Knaflic (2018, p. 47) diz que "quando você tem apenas um número ou dois para compartilhar, um simples texto pode ser uma ótima maneira de comunicar". Sobre o gráfico de barras, ela diz

> "às vezes os gráficos de barras são evitados por serem comuns. Isso é um erro. Os gráficos de barras devem ser aproveitados porque são comuns, pois isso significa menos curva de aprendizado para seu público. Em vez de usar o poder do cérebro para tentar ler o gráfico, seu público o utiliza para decidir qual informação extrair da apresentação."

Sobre os gráficos de barras empilhadas, Knaflic (2018, p. 51) afirma que "os casos de uso para gráficos de barras verticais empilhadas são mais limitados. Eles se destinam a permitir a comparação de totais entre categorias e ainda para visualizar as partes subcomponentes dentro de determinada categoria."

Por último, sobre o gráfico de barras horizontais, Knaflic (2018, p. 54) comenta

> "O gráfico de barras horizontais é particularmente útil se seus nomes de categoria são longos, pois o texto é escrito da esquerda para a direita, conforme a maioria lê, facilitando a leitura do gráfico para seu público. Além disso, graças ao modo como normalmente processamos informações - começando à esquerda e em cima, e fazendo

> Zs com nossos olhos na tela ou página -, a estrutura do gráfico de barras horizontais é tal que nossos olhos atingem os nomes de categoria antes dos dados. Isso significa que, ao chegarmos aos dados, já sabemos o que eles representam."

CONCLUSÃO E TRABALHOS FUTUROS

A análise de dados pode ser uma ferramenta poderosa para promover a transparência e a eficiência em organizações governamentais, como o Tribunal Superior Eleitoral. Ao utilizar técnicas de *business intelligence* e outras metodologias de análise de dados, é possível obter *insights* valiosos sobre o eleitorado nacional. Além disso, a transparência no TSE é fundamental para garantir a integridade do processo eleitoral e a confiança do público na democracia brasileira.

Este trabalho pode ser utilizado como um ponto de partida para futuras pesquisas sobre a aplicação de análise de dados no Tribunal Superior Eleitoral e em outras organizações governamentais. Algumas possíveis áreas de pesquisa incluem:

- explorar outras técnicas de análise de dados que podem ser aplicadas no TSE, como aprendizado de máquina e mineração de dados;
- investigar como a análise de dados pode ser usada para melhorar a segurança do processo eleitoral e prevenir fraudes;
- estudar como a análise de dados pode ser aplicada em outras áreas do governo, como saúde, educação e segurança pública, por exemplo;
- desenvolver ferramentas e plataformas de visualização de dados para tornar as informações mais acessíveis e compreensíveis para o público em geral;
- criação de um banco de dados OLAP para estudo desses dados. Isso pode ser útil para permitir que se analise os dados de diferentes perspectivas e níveis de granularidade, facilitando a identificação de tendências e padrões;
- esse trabalho pode ser facilmente alterado para aumentar a granularidade e fazer análises por zonas eleitorais;
- partidos políticos e candidatos podem utilizar os arquivos gerados nesse trabalho para análise e entendimento do eleitorado, a fim de identificar padrões de eleitores e desenvolver estratégias de campanha mais eficazes.
- ao cruzar com os dados sobre as votações pode-se identificar tendências e preferências do eleitorado. Isso pode ajudar os políticos a desenvolver plataformas mais relevantes e a se comunicar de maneira mais eficaz com seus eleitores.

REFERÊNCIAS

Achari, S,. Hadoop Essentials: Delve into the key concepts of Hadoop and get a thorough understanding of the Hadoop ecosystem. Packt Publishing, Birmingham, UK, pp 34. 2015.

Beck, M. and Libert, B. Disponível em: https://sloanreview.mit.edu/article/the-machine-learning-race-is-really-a-data-race/. Acesso em: 01 jun. 2023. 2018.

Biswas, S., Wardat, M., e Rajan, H. The Art and Practice of Data Science Pipelines: A Comprehensive Study of Data Science Pipelines In Theory, In-The-Small, and In-The-Large. In: IEEE/ACM INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING (ICSE), 44th, 2022.

Bock, A. M. Psicologias. Uma introdução ao estudo de psicologia. São Paulo: Saraiva, 2006. pág. 50-57.

Braga, J.; Menezes, L. Introdução aos Objetos de Aprendizagem. In: BRAGA, J. (Org.). Objetos de aprendizagem volume 1: introdução e fundamentos. Santo André: UFABC, 2015. Disponível em: http://netel.ufabc.edu.br//cursosinternos/ntme/wp-content/uploads/2015/09/FundamentosEaD_Unidade6.pdf . Acesso em: 07 jun. 2023.

Hadoop, A. What Is Apache Hadoop? – Disponível em: http://hadoop.apache.org/#What+Is+Apache+Hadoop%3F. Acesso em 06 jun. 2023.

Inmon, W. H. Building the Data Warehouse, 4º Edition. Wiley Publishing, Inc., 2005.

Kempe. S. A Short History of Data Warehousing. Disponível em: http://www.dataversity.net/a-short-history-of-datawarehousing/ . Acesso em: 06 jun. 2023. 2012

Kimball, R., R, M., The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition. John Wiley & Sons, Inc., 2013.

Knaflic, C. N. (2018). Storytelling com dados: um guia sobre visualização de dados para profissionais de negócios. (2ª Ed.). Alta Book.

Melo, C. Como Tratar Dados Ausentes com Pandas. Disponível em: https://sigmoidal.ai/como-tratar-dados-ausentes-com-pandas/. Acesso em: 06 jun. 2023.

Murta, G., Mont'Alvão, C, e Kosminsky, D. Visualização de Dados – Uma análise do redesign de visualizações da COVID-19. In: Colóquio internacional de design. 2020.

Nehring, K. e Puppe, C.. A Theory of diversity. Econometrica, v.70, n.3, may., 2002. p.1155–1198.

Provost, F. e Fawcett, T. Data Science para negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados. (1ª Ed). Alta Book.

Ronald, A. J. Disponível em: https://www.compact.nl/en/articles/data-quality-assessment/. Acesso em: 01 jun. 2023. 2012.

Silva, D. da. Conheça os 4 tipos de análise de dados para criar estratégias certeiras. Disponível em: https://www.zendesk.com.br/blog/tipos-analise-de-dados/. Acesso em: 04 jun. 2023. 2022.