UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO

BIANCA IZ OLIVEIRA LUIZE SARABIA

ANÁLISE DA QUALIDADE DOS DADOS ABERTOS DE RECIFE

RECIFE

2024

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

ENGENHARIA DA COMPUTAÇÃO

BIANCA IZ OLIVEIRA LUIZE SARABIA

ANÁLISE DA QUALIDADE DOS DADOS ABERTOS DE RECIFE

Trabalho apresentado ao Programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito para a obtenção do grau de Bacharel em Engenharia da Computação

Orientador(a): Luciano Barbosa

RECIFE

2024

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Sarabia, Bianca Iz Oliveira Luize.
Análise da qualidade dos dados abertos de Recife / Bianca Iz Oliveira Luize Sarabia. - Recife, 2024.
61 p. : il., tab.

Orientador(a): Luciano Barbosa
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Engenharia da Computação - Bacharelado, 2024.

1. dados abertos. 2. qualidade de dados. 3. dimensões de qualidade de dados. I. Barbosa, Luciano. (Orientação). II. Título.

000 CDD (22.ed.)

BIANCA IZ OLIVEIRA LUIZE SARABIA

ANÁLISE DA QUALIDADE DOS DADOS ABERTOS DE RECIFE

Trabalho apresentado ao Programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito para a obtenção do grau de Bacharel em Engenharia da Computação

Aprovado em: 18/03/2024.

BANCA EXAMINADORA

_____

Profº. Dr. Luciano Barbosa (Orientador)
Universidade Federal de Pernambuco

_____

Profº. Dr. Fernando Maciano (Examinador Interno)
Universidade Federal de Pernambuco

AGRADECIMENTOS

Primeiramente, à minha família, cujo apoio incondicional foi fundamental para que eu pudesse alcançar esta conquista. Quero destacar meus pais, Ana e Juba, e meu irmão Bruno, que estiveram ao meu lado diariamente, acompanhando de perto todo o processo e me dando forças para seguir em frente. Também desejo homenagear meus familiares que já se foram, mas que tenho a certeza de que estão presentes neste momento de celebração: Cláudio, Decrecio, Teca e Zezo.

Agradeço de coração a todos os meus amigos, que sempre estiveram ao meu lado, me incentivando, ouvindo meus problemas e compartilhando momentos inesquecíveis. Em especial, agradeço a Gabi e Mateus, cujo apoio foi indispensável ao longo dessa jornada.

Um agradecimento especial também aos meus colegas de curso, que foram companheiros fiéis em cada aula, em cada projeto, tornando a caminhada acadêmica mais leve e enriquecedora. Cada um de vocês fez parte da minha história e deixou uma marca em minha vida.

Também expresso minha profunda gratidão ao meu professor orientador, cujos conselhos, direcionamentos e apoio foram fundamentais para o desenvolvimento deste trabalho.

Por fim, mas não menos importante, agradeço a Deus, sem a sua benção nada seria possível.

Without data you're just another person with an opinion.

— William Edwards Deming

RESUMO

O Portal de Dados Abertos da Cidade do Recife disponibiliza publicamente dados produzidos pelas secretarias e órgãos de gestão municipal, possibilitando sua utilização em diversas aplicações. Portanto, é imprescindível assegurar a qualidade desses dados considerando o conceito multidimensional associado à qualidade. O foco do presente trabalho recaiu sobre os dados do tipo CSV e seus respectivos metadados, os quais estão estruturados em arquivos JSON. O objetivo principal foi avaliar a qualidade dos dados contidos nos arquivos CSV, assim como verificar a presença e qualidade dos metadados que oferecem informações adicionais sobre eles. Este processo abarcou diversas etapas, desde a coleta inicial dos dados até a análise dos resultados obtidos. Como resultado, identificaram-se problemas significativos de qualidade, tais como uma elevada incidência de dados nulos, a presença de duplicações e a existência de metadados apenas para um subconjunto dos arquivos CSV. As dimensões de qualidade afetadas incluem completude, consistência, livre de erros, representação, precisão e entendimento.

Palavras-chave: dados abertos; qualidade de dados; dimensões de qualidade de dados.

ABSTRACT

The Open Data Portal of the City of Recife publicly provides data produced by municipal departments and management bodies, enabling their use in various applications. Therefore, it is essential to ensure the quality of this data considering the multidimensional concept associated with quality. The focus of this study was on CSV data and their respective metadata, structured in JSON files. The main objective was to assess the quality of the data contained in the CSV files and to verify the presence and quality of metadata that provide additional information about them. This process encompassed several stages, from the initial data collection to the analysis of the results obtained. As a result, significant quality issues were identified, such as a high incidence of null data, the presence of duplications, and the existence of metadata only for a subset of the CSV files. The affected quality dimensions include completeness, consistency, free-of-error, representation, accuracy, and understandability.

Keywords: open data; data quality; data quality dimensions.

LISTA DE TABELAS

Tabela 1 – Princípios dos Dados Governamentais Abertos 16
Tabela 2 – Dimensões da Qualidade de Dados/Informação 17
Tabela 3 – Resumo dos resultados 54

LISTA DE FIGURAS

Figura 1 – Exemplo de arquivo JSON de metadados 20

Figura 2 – Pipeline da metodologia 23

LISTA DE GRÁFICOS

Gráfico 1 – Sucesso ou falha de leitura de arquivos CSV 37

Gráfico 2 – Separador dos arquivos CSV 37

Gráfico 3 – Quantidade de linhas dos arquivos CSV 38

Gráfico 4 – Quantidade de colunas dos arquivos CSV 38

Gráfico 5 – Alerta de preenchimento de nulos e percentual de colunas com alerta dos arquivos CSV 39

Gráfico 6 – Alerta de duplicação de linhas e percentual de linhas duplicadas dos arquivos CSV 40

Gráfico 7 – Quantidade de arquivos dos datasets 41

Gráfico 8 – Quantidade de arquivos CSV dos datasets 41

Gráfico 9 – Quantidade de arquivos JSON dos datasets 42

Gráfico 10 – Tipos de arquivos dos datasets 42

Gráfico 11 – Sucesso ou falha na leitura dos arquivos CSV dos datasets 43

Gráfico 12 – Datasets com arquivos CSV com alerta de preenchimento de nulos 44

Gráfico 13 – Datasets com arquivos CSV com alerta de duplicação de linhas 44

Gráfico 14 – Quantidade de linhas dos datasets 45

Gráfico 15 – Tipos de arquivos referenciados pelos JSON de metadados 46

Gráfico 16 – Top 10 etiquetas mais utilizadas para descrição dos arquivos CSV 47

Gráfico 17 – Nuvem de palavras das 100 etiquetas mais utilizadas para descrição dos arquivos CSV 47

Gráfico 18 – Frequência de atualização dos arquivos CSV 48

Gráfico 19 – Tipos das colunas dos arquivos CSV 49

Gráfico 20 – Descrições das colunas dos arquivos CSV 50

Gráfico 21 – Associação entre arquivos CSV e metadados 51

Gráfico 22 – Associação entre arquivos CSV referenciados pelos metadados e arquivos CSV baixados 52

Gráfico 23 – Datasets e sua cobertura de metadados 52

Gráfico 24 – Top 10 etiquetas mais utilizadas para descrição de arquivos em datasets com pelo menos um arquivo CSV sem metadados 53
Gráfico 25 – Nuvem de palavras das 100 etiquetas mais utilizadas para descrição de arquivos dos datasets com pelo menos um arquivo CSV sem metadados 54

SUMÁRIO

1 INTRODUÇÃO **14**

1.1 Objetivos **15**

1.1.1 Objetivo Geral *15*

1.1.2 Objetivos Específicos *15*

2 REFERENCIAL TEÓRICO **16**

2.1 Dados Governamentais Abertos **16**

2.2 Qualidade de Dados **17**

2.3 Avaliações Subjetivas e Objetivas de Qualidade **19**

2.4 Portal de Dados Abertos da Cidade do Recife **19**

2.5 Métricas de Qualidade **20**

2.6 Trabalhos Relacionados **21**

3 METODOLOGIA **23**

3.1 Coleta dos Dados **23**

3.2 Cálculo das Métricas **25**

3.2.1 Conjuntos de Dados *25*

3.2.2 Arquivos CSV *26*

3.2.3 Arquivos JSON *28*

3.3 Análise das Métricas **30**

3.3.1 Análises dos Arquivos CSV *30*

3.3.2 Análise dos Conjuntos de Dados *31*

3.3.3 Análise dos Metadados *33*

3.3.4 Análise da Associação dos Metadados aos Arquivos CSV *34*

4 RESULTADOS **36**

4.1 Coleta de Dados **36**

4.2 Arquivos CSV **36**

4.3 Conjuntos de Dados **40**

4.4 Metadados **45**

4.5 Associação dos Metadados aos Arquivos CSV **50**

5 CONCLUSÃO **56**

5.1 Trabalhos Futuros **57**

REFERÊNCIAS **58**

1 INTRODUÇÃO

De acordo com a definição da *Open Knowledge Foundation* [1], dados são considerados abertos quando estão disponíveis para uso, reutilização e redistribuição por qualquer pessoa, sem restrições significativas além da atribuição de autoria e do compartilhamento pela mesma licença.

Os dados abertos possuem uma variedade de aplicações. Uma aplicação frequente após o ano de 2020 foi o uso de dados relacionados à COVID-19 em conjunto com técnicas de Aprendizado de Máquina para prever diversos fatores. O trabalho [2] utilizou dados abertos dos estados de Alagoas (AL), Espírito Santo (ES) e Santa Catarina (SC) para prever a necessidade de hospitalização de pacientes com base em seus sintomas e histórico de saúde. Por sua vez, o trabalho [3] utilizou os dados abertos do Banco de Dados de Síndrome Respiratória Aguda Grave no Brasil para prever a mortalidade dos casos da doença, levando em consideração características fisiológicas e socioeconômicas.

Entretanto, é crucial ressaltar que a eficácia das aplicações está diretamente ligada à qualidade dos dados. A expressão *"garbage in, garbage out"* encapsula o princípio de que dados de entrada de baixa qualidade resultam em dados de saída pouco confiáveis [4]. Em outras palavras, a confiabilidade dos resultados de uma aplicação é limitada pela qualidade dos dados utilizados nela.

O Portal de Dados Abertos da Cidade do Recife desempenha um papel fundamental ao disponibilizar publicamente os dados produzidos pelas secretarias e órgãos de gestão municipal [5]. Essa iniciativa promove a transparência governamental e democratiza o acesso à informação, permitindo que qualquer pessoa desenvolva pesquisas, análises ou construa visualizações com base nos dados abertos disponíveis. Com isso, a sociedade pode desempenhar um papel ativo na melhoria e na resolução dos problemas do município.

No Portal, os dados são organizados utilizando as estruturas de conjuntos de dados (*datasets*) e recursos (arquivos). Os conjuntos de dados representam agrupamentos de recursos que possuem a mesma fonte ou metadados em comum. Os recursos, por sua vez, são as bases de dados, que podem ser arquivos de diversos formatos como CSV, JSON, PDF, ZIP, entre outros [6]. Os metadados associados aos conjuntos de dados ou aos arquivos geralmente são fornecidos em

arquivos de formato JSON, embora em alguns casos também possam ser disponibilizados no formato PDF.

Nesse contexto, torna-se evidente a importância de uma avaliação minuciosa de diversos aspectos de qualidade dos Dados Abertos da Cidade do Recife, a fim de garantir que suas aplicações sejam bem-sucedidas, resultando em pesquisas de alta qualidade e de impacto significativo.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo geral deste trabalho é analisar detalhadamente os dados públicos disponíveis no Portal de Dados Abertos da Cidade do Recife, com foco nos dados do tipo CSV e seus metadados associados, contidos em arquivos do tipo JSON. O propósito é avaliar a qualidade dos dados contidos nos arquivos CSV, bem como verificar a presença e a qualidade dos metadados que fornecem informações adicionais sobre eles.

1.1.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

- Realizar a coleta de todos os arquivos disponíveis no Portal, dos tipos CSV ou JSON.
- Calcular métricas para todos os arquivos do tipo CSV, visando avaliar diferentes aspectos de sua qualidade.
- Calcular métricas para todos os arquivos JSON e os metadados neles contidos, para compreender a qualidade e a consistência desses metadados.
- Calcular métricas para todos os conjuntos de dados presentes no Portal.
- Avaliar a associação entre os metadados encontrados nos arquivos JSON e os arquivos CSV coletados, por meio do cálculo de métricas específicas.
- Analisar os resultados obtidos em todas as etapas anteriores e verificar a qualidade dos dados públicos disponíveis no Portal da Cidade do Recife.

2 REFERENCIAL TEÓRICO

Neste capítulo, abordaremos o conceito de Dados Governamentais Abertos na seção 2.1. Em seguida, discutiremos o conceito de qualidade de dados e suas dimensões na seção 2.2. Na seção 2.3, exploraremos os tipos de avaliações de qualidade. Posteriormente, na seção 2.4, será aprofundado o contexto do Portal de Dados Abertos da Cidade do Recife. Na seção 2.5, cobriremos as métricas de qualidade e, por fim, na seção 2.6 revisaremos trabalhos relacionados a este tópico.

2.1 Dados Governamentais Abertos

Buscando definir os dados abertos que são provenientes de órgãos do governo, o *Open Government Working Group* [1], elaborou os oito Princípios dos Dados Governamentais Abertos. Para serem caracterizados como abertos, os dados devem satisfazer os critérios delineados na Tabela 1.

Tabela 1 – Princípios dos Dados Governamentais Abertos

Princípio	Definição
Completos	Todos os dados públicos, dados que não estão sujeitos a limitações válidas de privacidade, segurança ou privilégio, são disponibilizados.
Primários	Os dados são coletados na fonte, com o maior nível possível de granularidade, sem agregações ou modificações.
Atuais	Os dados são disponibilizados tão rápido quanto necessário para preservar o seu valor.
Acessíveis	Os dados estão disponíveis para a maior quantidade possível de usuários e aplicações.
Processáveis por Máquinas	Os dados são formatados razoavelmente para permitir um processamento automático.
Não Discriminatórios	Os dados estão disponíveis para qualquer pessoa, sem necessidade de cadastro.

Não Proprietários	Os dados são disponibilizados em um formato que nenhuma entidade tenha controle exclusivo sobre.
Livres de Licenças	Os dados não estão sujeitos a nenhuma restrição de direito autoral, patente, marcas ou segredos comerciais.

Fonte: Trabalho [7].

Nota: Tabela traduzida e organizada pelo autor com base nos conceitos informados no trabalho utilizado como fonte.

2.2 Qualidade de Dados

A definição de qualidade de dados não é unânime entre os autores. No trabalho [8], qualidade de dados é descrita como a medida do quanto as visualizações de dados em um sistema da informação correspondem aos mesmos dados na vida real. Por outro lado, no trabalho [9], qualidade ou utilidade dos dados é vista como dependente do usuário, sendo um conceito relativo que varia de acordo com as diferentes percepções e necessidades dos consumidores. Vários estudos sustentam que a qualidade de dados é um conceito multidimensional, abrangendo dimensões como precisão, completude, consistência e atualidade [10, 11, 12 e 13]. Embora essas definições tenham nuances distintas, elas se complementam. As dimensões da qualidade de dados e seus conceitos são apresentados na Tabela 2.

Tabela 2 – Dimensões da Qualidade de Dados/Informação

Dimensões	**Descrições**
Acessibilidade	O quanto as informações no site são facilmente disponíveis e podem ser baixadas.
Precisão	O quanto o conteúdo das informações no site é correto e confiável.
Adequação	O quanto o conteúdo é apropriado ao que os visitantes do site estão procurando.
Credibilidade	O quanto o conteúdo do site é verdadeiro e credível.
Completude	O quanto o conteúdo do site não possui

	partes faltantes e é suficiente.
Consistência	O quanto todas as páginas do site apresentam o mesmo formato.
Facilidade de Manipulação	O quanto o conteúdo do site é fácil de manipular.
Livre de Erros	O quanto a informação no site é correta, confiável e livre de erros.
Objetividade	O quanto o conteúdo do site é imparcial e não tendencioso.
Relevância	O quanto o conteúdo do site é aplicável, útil e relevante.
Representação	O quanto o conteúdo do site é legível, consistente e possui uma estrutura formal.
Reputação	O quanto a informação no site é altamente considerada em relação ao seu conteúdo.
Segurança	O quanto o acesso do site é restrito apropriadamente para manter sua segurança.
Velocidade	A velocidade de recuperação ou download do conteúdo do site.
Atualidade	O quanto o conteúdo do site é atualizado.
Entendimento	O quanto o conteúdo do site é fácil de entender e compreender.
Valor agregado	O quanto o conteúdo no site é benéfico e oferece vantagens a ser usado.

Fonte: Trabalho [14].

Nota: Tabela traduzida e organizada pelo autor com base nos conceitos informados no trabalho utilizado como fonte.

O estudo [15] examinou quais dimensões de qualidade estão abordadas nos princípios dos Dados Governamentais Abertos e identificou algumas lacunas. Os resultados indicaram a falta da consideração das seguintes dimensões: Credibilidade, Objetividade, Livre de Erros e Reputação. Diante disso, para este projeto, optamos por adotar uma abordagem multidimensional da qualidade de

dados, levando em consideração as diversas dimensões para a medição da qualidade.

2.3 Avaliações Subjetivas e Objetivas de Qualidade

Quando se trata da avaliação da qualidade dos dados, podemos distinguir entre dois tipos: subjetiva e objetiva. A avaliação subjetiva considera as necessidades dos consumidores dos dados, ou seja, varia de acordo com a motivação de uso e é definida por eles. Por outro lado, a avaliação objetiva pode ser de dois tipos: independente ou dependente da tarefa. A avaliação independente não leva em conta o contexto da aplicação e é genérica, podendo ser aplicada a qualquer conjunto de dados. Por outro lado, a avaliação dependente da tarefa pode incluir as regras de negócio de uma organização, empresa ou sistema onde esses dados estão inseridos variando de acordo com o contexto [16]. Este trabalho empregará avaliações objetivas independentes da tarefa, restringindo a análise a um subconjunto das dimensões de qualidade de dados disponíveis.

2.4 Portal de Dados Abertos da Cidade do Recife

Conforme discutido no Capítulo 1, os conjuntos de dados do Portal são coleções de recursos de dados (arquivos) que possuem alguma relação entre si. Esses conjuntos são organizados em grupos com o objetivo de catalogar dados de projetos ou equipes específicas, ou agrupar conjuntos relacionados a um tema comum, proporcionando aos usuários maior facilidade na busca pelos dados de interesse. Além disso, os recursos também podem ser agrupados de acordo com as organizações responsáveis por eles [17].

Os arquivos disponíveis no Portal podem adotar diferentes formatos. O formato GEOJSON é amplamente utilizado para dados espaciais, enquanto JSON e CSV são comuns para armazenar dados em geral. O formato PDF é exclusivamente empregado no Portal de Dados Abertos do Recife para metadados de recursos, porém está em processo de substituição pelo formato JSON [17]. Para este estudo, nos concentramos exclusivamente nos metadados armazenados em arquivos JSON, devido à facilidade de manipulação desse formato quando comparada a

manipulação de arquivos PDF. A estrutura básica de um arquivo JSON de metadados é ilustrada na Figura 1.

Figura 1 – Exemplo de arquivo JSON de metadados

```
{
  "metadados": {
    "cabecalho": {
      "titulo": "Conjunto de dados 1",
      "descricao": "Descrição conjunto de dados 1",
      ...,
      "frequencia_atualizacao": "Semestral",
      "etiquetas": "Etiqueta1, Etiqueta2, Etiqueta3, ..., EtiquetaN"
    },
    "recursos": [
      {
        "identificador": "IDRecurso1",
        "titulo": "Recurso 1",
        "formato": "csv",
        "descricao": "Descricao recurso 1"
      }
    ],
    "campos": [
      {
        "codigo": "Coluna 1",
        "descricao": "Descrição coluna 1",
        "tipo": "Char",
        "tamanho": 50,
        "valores_permitidos": ""
      },
      {
        "codigo": "Coluna 2",
        "descricao": "Descrição coluna 2",
        "tipo": "Num",
        "tamanho": 1,
        "valores_permitidos": "1,2,3,4,5,6"
      }
    ]
  }
}
```

2.5 Métricas de Qualidade

Para este trabalho, foram selecionadas métricas que se alinhassem com pelo menos uma das dimensões delineadas na Tabela 2. A análise da capacidade de leitura correta dos arquivos está diretamente relacionada com as dimensões de completude, pois indica se o conteúdo está completo; livre de erros, pois identifica se há erros no conteúdo; e representação, pois avalia se o conteúdo é legível.

Investigar os tipos de separadores utilizados nos arquivos CSV possibilita a identificação de padrões, o que está diretamente relacionado com a dimensão de consistência. Além disso, a análise da quantidade de linhas e colunas em cada arquivo e conjunto de dados se enquadra na dimensão de completude, assegurando

que os arquivos possuam, no mínimo, uma linha e uma coluna, fornecendo *insights* sobre a possível incompletude dos dados.

Examinar a estrutura dos conjuntos de dados em relação ao número e tipos de arquivos está associado às dimensões de completude e consistência, com o objetivo de identificar padrões e lacunas em suas composições. Analisar os metadados é pertinente à dimensão de entendimento, abrangendo a análise dos temas presentes nos dados, a tipologia das colunas e a qualidade das descrições das mesmas. Além disso, a análise de um metadado específico também se relaciona à dimensão de atualidade, que diz respeito à frequência de atualização de cada arquivo ou conjunto de dados. A capacidade de associar os metadados aos seus respectivos arquivos se enquadra nas dimensões de entendimento e completude.

Por fim, para uma análise mais aprofundada do conteúdo dos arquivos, examinar o preenchimento dos dados e a presença de duplicação de linhas, se enquadram nas dimensões de precisão e livre de erros. A avaliação do preenchimento, em particular, também se insere na dimensão de completude. Dessa forma, o presente trabalho se compromete em avaliar sete das dezessete dimensões, explorando-as em diferentes níveis de profundidade dentro do escopo do projeto.

Problemas nos dados, tais como valores ausentes e duplicatas, podem surgir em várias etapas do ciclo de vida dos dados, incluindo durante a fase de coleta e transmissão da informação. Portanto, é de extrema importância identificar esses erros precocemente, a fim de possibilitar uma correção eficaz [18]. Uma das possíveis consequências desses erros nos dados é observada nas aplicações de Aprendizado de Máquina, uma vez que esses sistemas necessitam compreender as associações entre os dados fornecidos como entrada e a propriedade que se deseja prever. Nesse contexto, a presença de valores ausentes ou de dados duplicados pode diminuir o desempenho dessas aplicações [19].

2.6 Trabalhos Relacionados

O estudo [20] investigou mais de 9000 conjuntos de dados abertos provenientes de 20 cidades distintas na América do Norte. A análise abordou

estatísticas gerais, explorando a qualidade dos dados e a viabilidade de integração entre diferentes conjuntos. Concluiu-se que os dados apresentam uma qualidade satisfatória e que há uma significativa oportunidade para integração entre múltiplas fontes. No entanto, identificaram-se áreas que requerem ajustes para otimizar a utilização dos dados, particularmente no que tange à sua disponibilização e acessibilidade.

Já o trabalho [21] teve como objetivo realizar uma análise do Portal Brasileiro de Dados Abertos, focando na avaliação da qualidade dos dados com base nos princípios dos Dados Governamentais Abertos comentados na seção 2.1 e nas dimensões de qualidade de dados comentadas na seção 2.2. Como conclusão, identificou-se que as dimensões de adequação, relevância, atualidade, facilidade de manipulação e entendimento destacam-se como pontos críticos de qualidade que demandam atenção por parte da Gestão Pública.

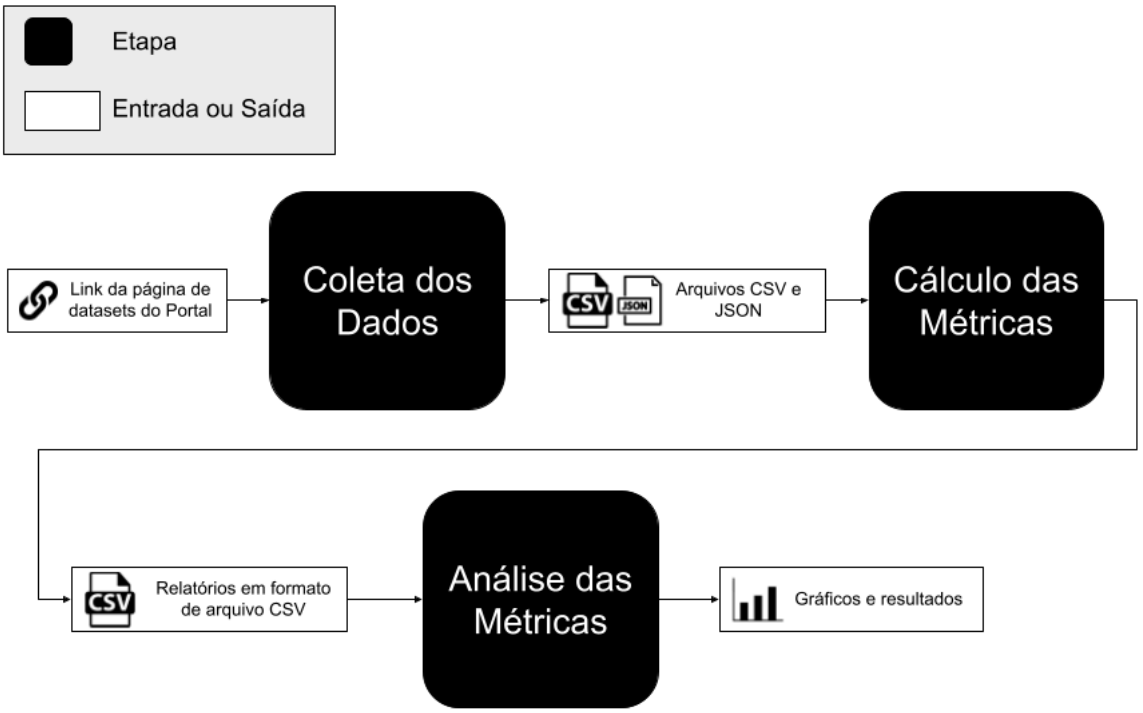
Por sua vez, o objetivo do trabalho [22] foi realizar uma análise da qualidade dos dados abertos no setor educacional brasileiro. Como resultado, foi desenvolvida uma ferramenta para auxiliar na avaliação desses dados. Além disso, foram identificados alertas relacionados à qualidade dos dados em várias categorias, incluindo licenciamento, completude, acessibilidade, processabilidade, atualidade, utilidade e legalidade. Algumas delas fazem interseção com as dimensões de qualidade.

Comparativamente aos estudos anteriores, o presente trabalho tem como propósito uma análise mais restrita, tanto no escopo dos dados, limitando-se às informações provenientes exclusivamente do município de Recife, quanto nas métricas empregadas, as quais englobam apenas algumas dimensões de qualidade em distintos níveis de profundidade.

3 METODOLOGIA

A metodologia empregada neste estudo foi guiada pela necessidade de coletar, analisar e avaliar os dados disponíveis no Portal de Dados Abertos da Cidade do Recife. Este processo envolveu uma série de etapas, desde a coleta dos dados até a análise dos resultados obtidos. Este capítulo descreve detalhadamente os métodos e técnicas empregados, fornecendo uma compreensão abrangente do processo de pesquisa. A visão geral da metodologia está apresentada na Figura 2. Na seção 3.1, descreveremos a etapa de coleta de dados; na seção 3.2, discutiremos as métricas que foram calculadas; e na seção 3.3, abordaremos o agrupamento das métricas durante a etapa de análise.

Figura 2 – *Pipeline* da metodologia



3.1 Coleta dos Dados

Para iniciar a pesquisa, o primeiro procedimento consistiu na coleta integral dos dados disponíveis no Portal de Dados Abertos de Recife. Diante da volumosa quantidade de arquivos, a realização desta etapa por meio manual mostrou-se inviável. Portanto, foi imprescindível o desenvolvimento de um script *Python* para operar um *scraper*, com o propósito de efetuar o download dos dados de maneira

automatizada. Um *scraper* é uma ferramenta computacional projetada para extrair dados de fontes online, como páginas da web. Utilizando técnicas de análise estrutural do conteúdo web, um scraper pode identificar e coletar informações específicas de forma automatizada [23].

Para a execução dessa etapa, empregou-se a biblioteca *Beautiful Soup*, uma ferramenta em *Python* reconhecida por sua capacidade de extrair dados de arquivos HTML e XML [24]. Por meio dela, tornou-se viável a navegação, pesquisa e manipulação eficaz dos elementos presentes nesses documentos, o que facilitou significativamente a coleta de informações provenientes de páginas da web.

Iniciando a coleta a partir da página de *datasets* do site, a qual organiza os links para os diversos conjuntos de dados em várias páginas numeradas, procedemos com a obtenção de todos os links disponíveis, percorrendo todas as páginas da lista. Em seguida, para cada um dos links de *dataset*, realizou-se uma segunda coleta de links, desta vez direcionada aos links de download de cada arquivo. Durante essa etapa, aplicou-se um filtro para selecionar exclusivamente os arquivos no formato CSV e JSON, que correspondiam ao foco de nossa pesquisa.

Com todos os links de download coletados, tornou-se viável enviar requisições GET para cada um deles, solicitando a transferência dos dados correspondentes. Para essa finalidade, foi utilizada a biblioteca *Requests* de *Python* [25], que é uma ferramenta amplamente utilizada para realizar requisições HTTP de forma simplificada e eficiente. Com ela, é possível interagir com recursos online enviando e recebendo dados por meio de solicitações HTTP. Ao enviar uma requisição para cada um dos links correspondentes aos arquivos, foi possível realizar o download de todos eles.

Embora o site ofereça uma API de Dados para acessar as informações, ela se apresenta disponível apenas para arquivos CSV. Além disso, seria necessário mapear o identificador de cada arquivo para solicitar o download individualmente, o que aumentaria a complexidade do *scraper*. Portanto, a fim de generalizar o processo de download e ao constatar a viabilidade de realizar o download de todos os arquivos por meio da coleta de links, optou-se por seguir essa estratégia.

3.2 Cálculo das Métricas

Com todos os dados coletados, foi possível seguir para a realização do cálculo de diversas métricas que serão utilizadas na análise de qualidade. Estas métricas serão posteriormente avaliadas, agrupadas e cruzadas entre si, com o propósito de tornar possível a realização de uma análise abrangente dos dados abertos. O objetivo final desta etapa consistiu na produção de três relatórios, os quais serão apresentados em formato CSV. Cada relatório estará relacionado a um agrupamento específico de métricas. O primeiro relatório abordará as métricas gerais dos *datasets*, enquanto o segundo relatório se concentrará nas métricas de cada arquivo CSV individual. Por fim, o terceiro relatório abordará as informações de metadados extraídas dos arquivos JSON. Nas subseções a seguir, serão detalhadas as métricas levantadas em cada um dos relatórios.

Essa etapa foi realizada na plataforma de desenvolvimento *Google Collaboratory*, ou *Colab*, de forma abreviada. O *Colab* é um produto da *Google Research* que oferece a capacidade de escrever e executar código *Python* diretamente no navegador, utilizando uma versão do *Jupyter Notebook* como serviço. Essa ferramenta é amplamente utilizada para aprendizado de máquina, análise de dados e educação, devido à sua facilidade de uso e à integração com outras ferramentas e serviços do *Google*, como o *Google Drive*, de onde os dados serão lidos [26].

O *Jupyter Notebook* [27], por sua vez, é uma aplicação web de código aberto que possibilita a criação e compartilhamento de documentos interativos. Esses documentos podem conter código executável, equações, visualizações e texto, proporcionando uma ferramenta completa para análise de dados.

3.2.1 Conjuntos de Dados

Para cada um dos *datasets*, foram registradas três métricas específicas no relatório: a quantidade total de arquivos, a quantidade de arquivos no formato CSV e a quantidade de arquivos no formato JSON. Devido à organização em pastas estabelecida durante o processo de download, a obtenção dessas métricas demandou apenas a listagem de diretórios utilizando a biblioteca *OS* [28]. Este módulo integrado do *Python* oferece métodos para interação com o sistema

operacional, incluindo a criação e gerenciamento de arquivos e diretórios, manipulação de variáveis de ambiente, entre outros recursos.

A determinação da quantidade total de arquivos baseou-se no tamanho da lista resultante do método *listdir*. Para contabilizar os arquivos no formato CSV, foram considerados aqueles cujos nomes terminam com a extensão ".csv"; de forma análoga, os arquivos JSON foram identificados pela terminação ".json". Adicionalmente, foram criadas duas colunas de *flags* para indicar a presença ou ausência de arquivos de cada tipo. Essas *flags* foram desenvolvidas com o intuito de facilitar a análise dos resultados em uma etapa seguinte do projeto.

O relatório desenvolvido nesta etapa consiste em múltiplas linhas, em que cada linha corresponde a um conjunto de dados. A primeira coluna do relatório é destinada ao nome do conjunto de dados e cada coluna subsequente representa uma das métricas mencionadas anteriormente nesta subseção.

3.2.2 Arquivos CSV

Para a leitura dos arquivos CSV, foi utilizada a biblioteca *Pandas* [29], uma ferramenta de análise e manipulação de dados de código aberto poderosa e flexível, construída sobre a linguagem de programação *Python*. No entanto, antes da execução da leitura, tornou-se necessário identificar automaticamente o separador utilizado nos arquivos CSV, a fim de garantir uma leitura bem sucedida. O separador é o caractere utilizado para delimitar os campos em uma linha do arquivo CSV. Embora a vírgula (,) seja o separador mais comum, outros caracteres, como ponto e vírgula (;), tabulação (\t), espaço ( ), barra vertical (|), entre outros, também podem ser utilizados.

Para essa finalidade, utilizamos o módulo *csv* do *Python* [30], o qual oferece funcionalidades para leitura e gravação de arquivos CSV. A identificação do separador foi realizada por meio da classe *Sniffer*. Esta classe tem como finalidade deduzir o formato do arquivo CSV, fornecendo diversas informações úteis, como a presença de aspas em torno dos valores, a existência de um cabeçalho, entre outras. No entanto, para o nosso contexto, foi necessária apenas a identificação do separador. Essa informação foi uma das métricas que decidimos registrar sobre cada um dos arquivos, levando em consideração a possibilidade de variação desse

elemento. Assim, no relatório, é especificado o separador utilizado em cada arquivo individualmente.

Durante o processo de leitura, foi possível identificar dois problemas nos arquivos: erro de parser e arquivos vazios. O erro de parser indicava que o arquivo não estava em conformidade com o padrão CSV; por exemplo, quando uma das linhas do CSV continha mais campos do que o cabeçalho (*header*). O cabeçalho corresponde à primeira linha do arquivo, na qual estão determinados os nomes de cada uma das colunas. O problema de arquivos vazios se refere a arquivos que não continham nenhum conteúdo, que estavam totalmente vazios, sem nenhum caractere. Para cada tipo de erro de leitura, foi associada uma coluna de *flag*, na qual é marcada como verdadeira para os casos que se enquadram nele e falso caso contrário.

Para os arquivos que não apresentaram nenhum dos problemas mencionados anteriormente, o próximo passo consistia em verificar se eles continham alguma linha de dados ou se possuíam apenas o cabeçalho. Essa informação era registrada por meio de uma coluna de *flag* no relatório, marcando como verdadeiros os casos que se enquadram nessa condição. Adicionalmente, também foram registradas no relatório a quantidade de linhas e de colunas de cada arquivo.

As próximas métricas levantadas dizem respeito ao preenchimento de cada coluna. Para cada coluna, foi calculada a porcentagem de campos não preenchidos. Em seguida, um limiar de 50% foi estabelecido. Se alguma coluna apresentasse uma porcentagem de campos não preenchidos maior que esse limiar, no relatório uma *flag* era marcada como verdadeira, indicando a presença de uma coluna com alerta de preenchimento no respectivo arquivo CSV. Além disso, no relatório, são armazenados o percentual de campos nulos de cada coluna, a quantidade de colunas com problema de preenchimento e o nome das colunas que apresentam esse problema.

Por fim, as métricas calculadas foram direcionadas à análise de dados duplicados no arquivo, especificamente a verificação da existência de linhas exatamente iguais umas às outras. Para isso, foi realizada a remoção de linhas duplicadas e recalculada a quantidade total de linhas. A diferença entre a quantidade de linhas antes e depois da remoção de duplicados revelou a quantidade de linhas

duplicadas. Se essa quantidade fosse maior que zero, uma *flag* no relatório era marcada como verdadeira, indicando a ocorrência desse problema. As quantidades de linhas distintas e duplicadas também foram registradas no relatório.

O relatório construído nesta etapa é constituído por várias linhas, em que cada linha representa um arquivo de um conjunto de dados específico. A primeira coluna do relatório é dedicada ao nome do conjunto de dados, a segunda coluna corresponde ao nome do arquivo, e cada coluna subsequente representa uma das métricas mencionadas anteriormente nesta subseção.

3.2.3 Arquivos JSON

Para cada arquivo JSON, foi utilizada a biblioteca *json* para sua leitura [31]. Esta biblioteca tem como finalidade interpretar JSON a partir de strings ou arquivos, retornando um dicionário ou uma lista *Python*. Durante esse processo, foi realizada uma avaliação para determinar se o JSON representava o metadado de outro arquivo ou se consistia em uma lista de dicionários, caracterizando um conjunto de dados. No relatório desta etapa, uma coluna de *flag* foi criada para indicar se o JSON representava metadados, sendo marcada como verdadeira quando essa condição era atendida.

Realizamos os passos seguintes apenas para os arquivos que representavam metadados, os quais fornecem diversas informações sobre um ou mais arquivos do mesmo conjunto de dados. A primeira informação extraída do arquivo JSON foram as etiquetas, um campo que contém uma lista de palavras relacionadas aos arquivos que ele faz referência, indicando as principais características dos dados contidos neles. Além disso, foi extraído o campo que continha a informação da frequência de atualização, que indica com que frequência os arquivos referenciados sofrem atualizações. Essa informação pode ter os seguintes valores: SEMANAL, MENSAL, TRIMESTRAL, SEMESTRAL, ANUAL, FINALIZADO ou nulo. Ambas as informações mencionadas foram incluídas no relatório.

A próxima informação coletada foi o tipo de cada uma das colunas. Colunas de texto podem ser representadas no arquivo JSON pelos tipos CHAR, TEXT ou VARCHAR. Enquanto colunas numéricas podem ser identificadas por NUM, NUMBER, NUMER ou NUMEBR, evidenciando erros de digitação na construção do

JSON. Colunas do tipo data são caracterizadas por DATE, DATA ou por uma *string* que contenha DATA ou DATE em sua composição. Além disso, existem colunas que consideramos como outros tipos, tais como JSON, MULTIPOLYGON e TEMPO. Para nossas análises, agrupamos as colunas nesses quatro tipos distintos. No relatório, armazenamos listas contendo as colunas de cada tipo, bem como a quantidade de colunas de cada tipo.

Outra informação que mapeamos foi a descrição de cada uma das colunas. Realizamos uma classificação das descrições em: nulas (“NÃO POSSUI”), iguais ao nome da coluna (“IGUAIS AO NOME”), menores do que o nome da coluna (“MENORES QUE O NOME”) e maiores do que o nome da coluna (“MAIORES QUE O NOME”). Essa classificação nos permitirá posteriormente avaliar a qualidade de cada uma dessas descrições de forma mais precisa. No relatório, armazenamos listas contendo as colunas de cada classificação, bem como a quantidade de colunas em cada categoria.

Por último, há um campo no arquivo JSON que consiste em uma lista de todos os arquivos aos quais esses metadados pertencem, juntamente com os tipos desses arquivos referenciados. Com base nisso, mapeamos essas informações no relatório para facilitar a identificação de qual arquivo CSV está sendo referenciado por qual arquivo JSON de metadados.

O relatório elaborado nesta etapa é composto por várias linhas, onde cada linha representa um arquivo específico, referenciado por um arquivo JSON. A primeira coluna do relatório corresponde ao nome do conjunto de dados, enquanto a segunda coluna indica o nome do arquivo JSON. Cada coluna subsequente representa uma das métricas mencionadas anteriormente nesta seção, com exceção das duas últimas. A penúltima coluna exibe o nome do arquivo referenciado pelo JSON, enquanto a última indica o seu tipo. Em resumo, para cada arquivo JSON, serão geradas N linhas, onde cada linha representa um arquivo ao qual ele faz referência, juntamente com seus metadados.

3.3 Análise das Métricas

Com os relatórios construídos, definimos nossa abordagem de análise das métricas em quatro pilares distintos listados abaixo. Todas serão detalhadas nas seguintes subseções.

- Análise dos arquivos CSV;
- Análise dos conjuntos de dados;
- Análise dos metadados presentes nos arquivos JSON que se referem aos arquivos CSV;
- Associação dos metadados aos arquivos CSV.

Para facilitar a compreensão e avançar na análise das métricas, criamos categorias para algumas delas. Essa abordagem foi aplicada às métricas de quantidade, como número de arquivos, colunas, linhas, entre outras, transformando esses valores numéricos em categorias. Por exemplo, para analisar a distribuição dos arquivos com base no número de linhas, utilizamos a estratégia de criar categorias como "menos de 10 linhas", "entre 10 e 100 linhas" e assim por diante. Para determinar essas categorias, analisamos os valores mínimos e máximos e dividimos esse intervalo em quatro ou mais grupos. Essa metodologia nos permite gerar gráficos de distribuição e facilita a interpretação dos resultados.

Para criar os gráficos, utilizamos a biblioteca *Matplotlib* [32]. Esta é uma ferramenta que permite a criação de visualizações estáticas, animadas e interativas em *Python*. Além disso, ela possui uma integração forte com a biblioteca *Pandas*, que já estamos utilizando, o que facilitou muito o desenvolvimento. Essa integração simplifica a criação de gráficos a partir de objetos do *Pandas*, tornando o processo mais eficiente e intuitivo. Essa etapa, de maneira análoga à etapa de cálculo das métricas, também foi conduzida no *Google Colab*.

3.3.1 Análises dos Arquivos CSV

Com o relatório dos arquivos CSV, conduzimos diversas análises. Primeiramente, avaliamos a quantidade de arquivos lidos com sucesso e os que apresentaram falhas na leitura. Em seguida, focando nos arquivos que foram lidos com sucesso, realizamos uma análise da distribuição da quantidade de arquivos por tipo de separador.

Também realizamos a análise da quantidade de linhas de cada arquivo. Dado que essa métrica é uma variável numérica sujeita a grande variação, optamos por criar seis categorias, seguindo a estratégia mencionada no início desta seção. As seis categorias definidas foram as seguintes: "0 LINHAS", "ENTRE 1 E 1K LINHAS", "ENTRE 1K E 10K LINHAS", "ENTRE 10K E 100K LINHAS", "ENTRE 100K E 1M LINHAS" e "1M OU MAIS LINHAS". Com essas categorias estabelecidas, levantamos a quantidade de arquivos em cada grupo.

Analogamente, para analisar a quantidade de colunas de cada arquivo, adotamos a mesma estratégia. Definimos seis categorias: "0 COLUNAS", "ENTRE 1 E 10 COLUNAS", "ENTRE 11 E 50 COLUNAS", "ENTRE 51 E 200 COLUNAS", "ENTRE 201 E 1K COLUNAS" e "1K OU MAIS COLUNAS". Após categorizar os arquivos de acordo com o número de colunas, levantamos a quantidade de arquivos em cada grupo.

Para a análise de preenchimento, investigamos a quantidade de arquivos que apresentavam alguma coluna com alerta de preenchimento nulo. Além disso, calculamos, para cada arquivo, o percentual de colunas com alerta de nulo. Esses percentuais foram agrupados em cinco categorias: "MENOS QUE 25%", "ENTRE 25 E 50%", "ENTRE 50 E 75%", "ENTRE 75 E 100%" e "100%". Posteriormente, levantamos a quantidade de arquivos em cada grupo mencionado acima.

Por fim, para a análise de linhas duplicadas, investigamos a quantidade de arquivos que continham linhas duplicadas. Além disso, calculamos, para cada arquivo, o percentual de linhas duplicadas. Esses percentuais foram agrupados em cinco categorias: "MENOS QUE 25%", "ENTRE 25 E 50%", "ENTRE 50 E 75%", "ENTRE 75 E 100%" e "100%". Posteriormente, também levantamos a quantidade de arquivos em cada classificação mencionada acima.

3.3.2 Análise dos Conjuntos de Dados

Com base no relatório dos *datasets*, conduzimos diversas análises. Primeiramente, calculamos a quantidade total de *datasets*, representada pela contagem de linhas no relatório. Em seguida, determinamos o número total de arquivos baixados durante a coleta, esse valor é o resultado da soma da quantidade de arquivos por conjunto de dados. Por fim, realizamos a contagem dos arquivos de

cada tipo, CSV e JSON, somando as respectivas quantidades por *dataset* encontradas no relatório.

Para proporcionar uma melhor compreensão do tamanho dos conjuntos de dados em relação à quantidade de arquivos, foram criadas categorias com faixas de quantidades de arquivos. Essas categorias permitiram analisar a distribuição dos conjuntos de dados. Para a quantidade total de arquivos, foram definidas quatro categorias: “ENTRE 1 E 10 ARQUIVOS”, “ENTRE 11 E 20 ARQUIVOS”, “ENTRE 21 E 30 ARQUIVOS” e “31 OU MAIS ARQUIVOS”. Da mesma forma, para a quantidade total de arquivos CSV, foram estabelecidas quatro categorias: “0 ARQUIVOS”, “ENTRE 1 E 10 ARQUIVOS”, “ENTRE 11 E 20 ARQUIVOS” e “21 OU MAIS ARQUIVOS”. Para a quantidade total de arquivos JSON, foram delineadas cinco categorias: “0 ARQUIVOS”, “ENTRE 1 E 3 ARQUIVOS”, “ENTRE 4 E 6 ARQUIVOS”, “ENTRE 7 E 9 ARQUIVOS” e “10 OU MAIS ARQUIVOS”. Para os três casos acima, foram construídos gráficos de distribuição da quantidade de *datasets* por categoria.

Além disso, na análise da quantidade de arquivos, também investigamos o número de conjuntos de dados que incluem tanto arquivos CSV quanto arquivos JSON, os conjuntos que possuem exclusivamente arquivos JSON e aqueles que contêm apenas arquivos CSV. Essa análise permite compreender a distribuição e a interseção entre diferentes tipos de arquivos nos conjuntos de dados investigados.

Com o relatório dos arquivos CSV, realizamos agrupamentos por *dataset* e analisamos diversas métricas associadas a cada arquivo CSV. Uma dessas métricas é o erro de leitura, que nos permite identificar se houve algum problema ao ler o arquivo. A partir disso, podemos determinar a quantidade de conjuntos de dados que possuem pelo menos um arquivo CSV com erro de leitura. Com base nesta análise, identificamos os conjuntos de dados que possuem arquivos CSV, porém nenhum deles foi lido com sucesso, indicando possíveis problemas na integridade ou na estrutura dos dados nesses conjuntos específicos.

Além disso, realizamos agrupamentos semelhantes para as análises de alerta de preenchimento das colunas nos arquivos CSV. Nesse sentido, levantamos a quantidade de conjuntos de dados que possuem pelo menos um arquivo CSV com esse tipo de alerta. Fazemos o mesmo para as análises de linhas duplicadas. Levantamos a quantidade de *datasets* que possuem pelo menos um arquivo CSV com o alerta de duplicação.

Por último, conduzimos uma análise adicional para obter uma visão do tamanho dos conjuntos de dados em relação à quantidade de linhas. Para determinar a quantidade de linhas de um conjunto de dados específico, realizamos a soma das quantidades de linhas de cada um dos arquivos CSV que o compõem. Com base nessa abordagem, criamos categorias com faixas de quantidades de linhas para classificar cada conjunto de dados. As seis categorias estabelecidas são as seguintes: "0 LINHAS", "ENTRE 1 E 1K LINHAS", "ENTRE 1K E 10K LINHAS", "ENTRE 10K E 100K LINHAS", "ENTRE 100K E 1M LINHAS" e "1M OU MAIS LINHAS". Essa categorização nos permite entender melhor a distribuição e o tamanho dos conjuntos de dados em termos de quantidade de linhas.

3.3.3 Análise dos Metadados

Com base no relatório dos arquivos JSON, foi possível conduzir uma série de análises. A primeira delas consistiu em uma análise para distinguir a quantidade dos arquivos JSON que são metadados e aqueles que não o são. Além disso, calculamos a quantidade de arquivos JSON que são metadados de arquivos CSV, a quantidade total de arquivos referenciados pelos arquivos JSON e a quantidade total de arquivos do tipo CSV referenciados por esses JSONs.

Para as análises a seguir, focamos exclusivamente nos metadados que referenciam arquivos do tipo CSV, permitindo-nos desconsiderar os metadados de outros tipos de arquivos, como GEOJSONs, por exemplo. Essas análises fornecem um entendimento abrangente dos arquivos CSV baixados durante a fase de coleta, examinando diversos aspectos desses arquivos, oferecendo percepções sobre sua estrutura, conteúdo e qualidade.

Uma análise importante foi realizada nas etiquetas utilizadas para descrever os arquivos, com o objetivo de levantar a distribuição das 10 palavras mais frequentemente utilizadas para descrever os dados públicos. Essa análise também resultou na geração de uma nuvem de palavras (*wordcloud*), que é uma representação visual das palavras mais utilizadas, onde o tamanho de cada palavra é proporcional à sua frequência de uso nos metadados. A *wordcloud* foi construída com as 100 palavras mais frequentes, proporcionando uma visão intuitiva e representativa do vocabulário utilizado na descrição dos arquivos.

Também realizamos uma análise das frequências de atualização dos arquivos CSV, visando compreender o intervalo de tempo em que os dados são atualizados e disponibilizados no portal. Investigamos quantos dos dados disponíveis no portal não serão mais atualizados, bem como quantos deles não possuem informações sobre atualizações futuras.

Também examinamos os tipos das colunas dos dados abertos de Recife. Fizemos um levantamento da quantidade total de colunas de cada tipo, fornecendo uma visão sobre os tipos de dados mais predominantes no portal. Por último, investigamos a qualidade das descrições de cada coluna. Para isso, calculamos a distribuição da quantidade total de colunas por categoria, como discutido em uma seção anterior. O objetivo foi analisar a qualidade geral das descrições das colunas dos dados, oferecendo insights sobre a clareza das informações disponibilizadas.

3.3.4 Análise da Associação dos Metadados aos Arquivos CSV

Para analisar quais arquivos CSV possuem metadados associados a eles, realizamos um cruzamento entre os relatórios de arquivos CSV e arquivos JSON. Nesse processo, verificamos se o nome do arquivo referenciado pelo JSON é idêntico ao nome do arquivo CSV e se os *datasets* são os mesmos. Além disso, consideramos a condição de que o formato do arquivo referenciado pelo JSON deve ser CSV. Isso nos permitiu analisar várias métricas, como a quantidade de arquivos CSV que foram mapeados e aqueles que não foram. Outra análise realizada foi a contagem de arquivos CSV referenciados pelos JSONs que corresponderam a algum CSV baixado e a quantidade que não tiveram correspondência.

Analisando os *datasets*, levantamos a quantidade de conjuntos que tiveram todos os seus arquivos CSV mapeados em metadados. Também identificamos a quantidade dos que tiveram alguns arquivos mapeados e outros não, além da quantidade dos que não possuíam nenhum arquivo CSV mapeado para metadados. Essa análise nos permite entender a integridade e a consistência dos metadados em relação aos *datasets* e seus arquivos associados.

Além disso, conduzimos uma análise apenas dos conjuntos de dados que possuem algum arquivo CSV não mapeado em metadados. Ao examinar as etiquetas de todos os arquivos JSON presentes em cada *dataset* deste subconjunto, levantamos a distribuição das dez palavras mais frequentes. Em seguida,

construímos uma nuvem de palavras com as cem palavras mais presentes, seguindo uma abordagem análoga à discutida em uma subseção anterior. Essa análise nos proporciona insights adicionais sobre os conjuntos de dados que possuem arquivos CSV não mapeados em metadados, nos permitindo identificar possíveis áreas de melhoria nos dados do portal.

4 RESULTADOS

Neste capítulo, são apresentados e discutidos os resultados obtidos a partir das etapas detalhadas no Capítulo 3. Os resultados da etapa de coleta dos dados serão abordados na seção 4.1. Em seguida, na seção 4.2, serão apresentadas as análises realizadas por arquivo CSV. Posteriormente, na seção 4.3, serão discutidos os resultados das análises por conjunto de dados. As análises dos metadados serão tratadas na seção 4.4, enquanto as associações dos metadados aos arquivos CSV serão apresentadas na seção 4.5.

4.1 Coleta de Dados

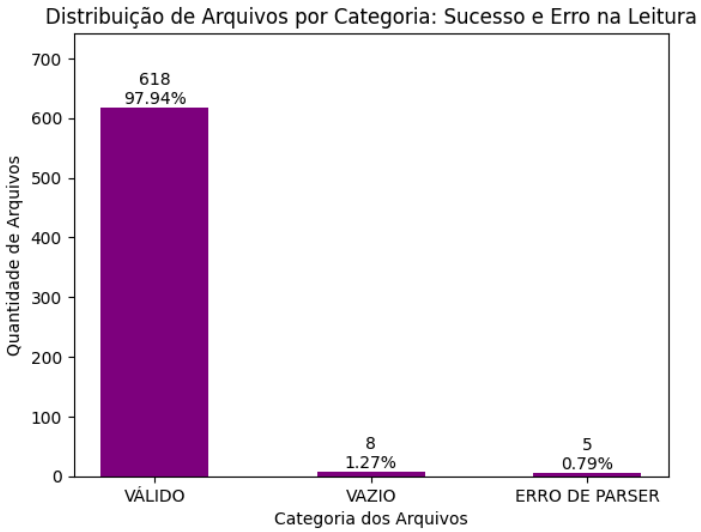
Durante a execução do código de coleta de dados detalhado na seção 3.1, havia 155 conjuntos de dados disponíveis no Portal de Dados Abertos da Cidade do Recife. Após a conclusão dessa etapa, foram geradas 151 pastas correspondentes aos conjuntos de dados. Isso se deve ao fato de que 4 desses conjuntos não continham arquivos nos formatos CSV ou JSON. Os conjuntos de dados sem arquivos coletados foram os seguintes: Cobertura da terra; Espécies de aves presentes no Recife; Faixas marginais dos recursos hídricos; Unidades protegidas do Recife.

Dos 151 *datasets*, um total de 909 arquivos dos tipos de interesse da pesquisa, CSV ou JSON, foram baixados. Essas 151 pastas e 909 arquivos totalizam 11,1GB de armazenamento. Atualmente, o Portal apresenta mais de 200 conjuntos de dados disponíveis, pois novos dados foram incluídos no portal após a conclusão da nossa pesquisa. Dessa forma, nosso projeto abrange a análise de cerca de 70% dos dados disponíveis hoje no Portal.

4.2 Arquivos CSV

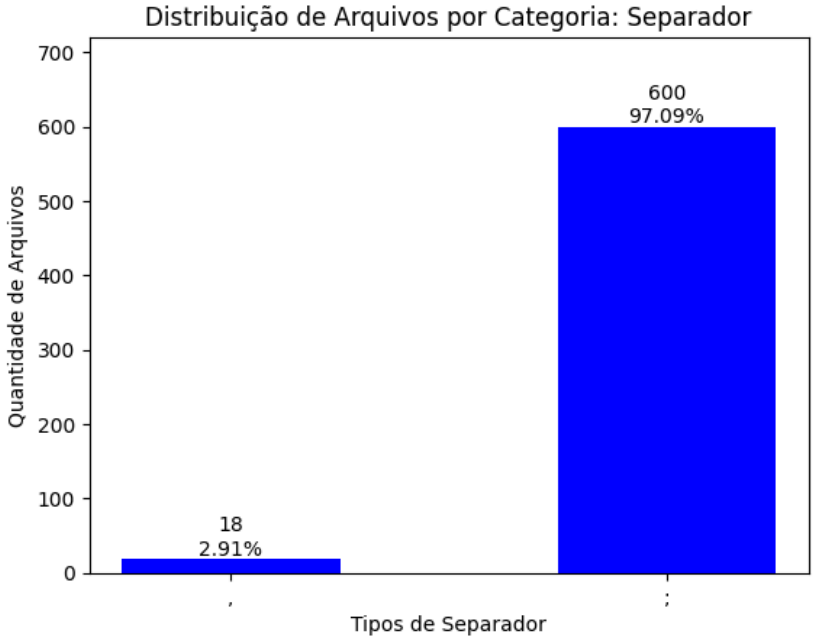
Em relação ao sucesso ou falha durante a leitura dos arquivos CSV, os resultados estão apresentados no Gráfico 1. Observa-se que a grande maioria deles não apresentou problemas, porém houve alguns casos em que ocorreram falhas na leitura. Isso destaca a necessidade de atenção por parte do portal, pois ainda existem arquivos defeituosos disponíveis que precisam ser ajustados.

Gráfico 1 – Sucesso ou falha de leitura de arquivos CSV



Para os arquivos lidos com sucesso, obtivemos a distribuição por tipo de separador conforme apresentado no Gráfico 2. Observamos que o tipo mais utilizado foi o ponto e vírgula, embora tenha havido alguns casos em que o separador utilizado foi a vírgula. Portanto, existe uma variação nesse parâmetro, o que evidencia a falta de padronização nesse aspecto.

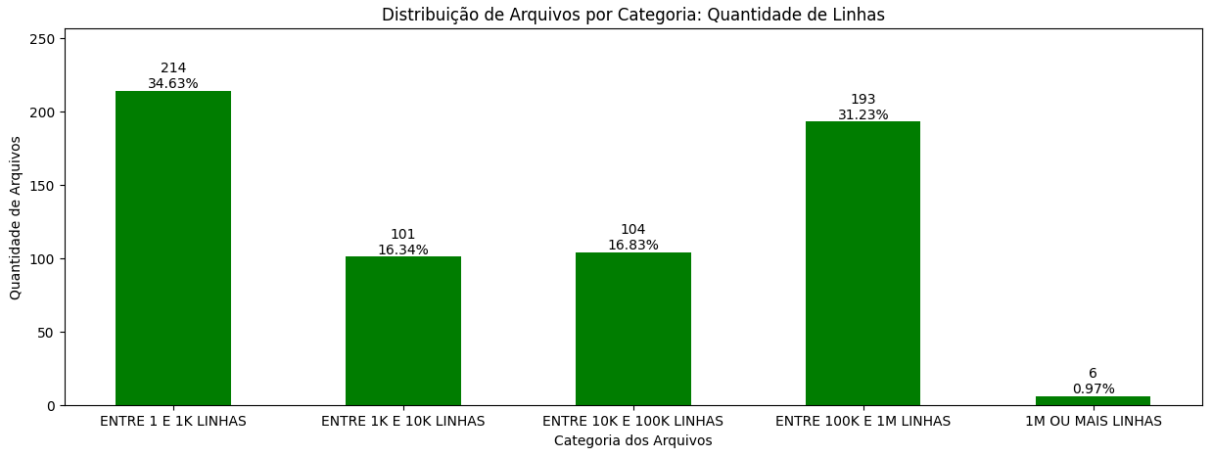
Gráfico 2 – Separador dos arquivos CSV



O resultado da análise de tamanho dos arquivos por quantidade de linhas está representado no Gráfico 3. É perceptível que a maioria dos arquivos possui um número relativamente pequeno de linhas, com 34,63% deles contendo entre 1 e 1000 linhas. No entanto, também é observado que uma proporção significativa de

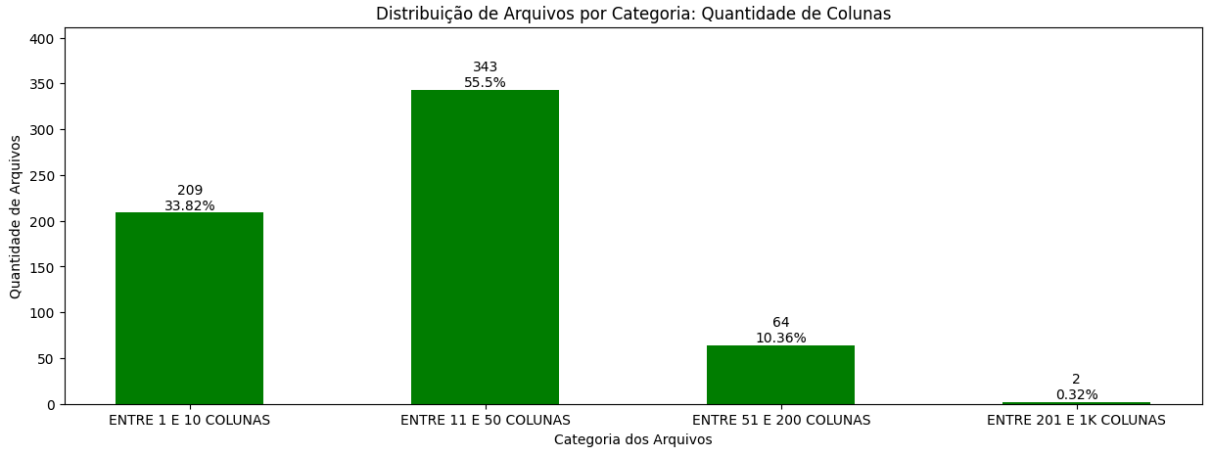
arquivos possui um número considerável de linhas, com 31,23% deles contendo entre 100,000 e 1,000,000 linhas. Isso sugere uma variação no tamanho dos arquivos, que pode variar desde arquivos relativamente pequenos até arquivos medianos, embora a maioria se limite a um milhão de linhas.

Gráfico 3 – Quantidade de linhas dos arquivos CSV



O resultado da análise de tamanho dos arquivos por quantidade de colunas está representado no Gráfico 4. Pode-se concluir que a maioria dos arquivos possui um número moderado de colunas, contendo até no máximo 50 colunas. Vale ressaltar que uma parcela significativa possui um número mais reduzido de colunas, com 33,82% deles contendo no máximo 10. Dessa forma, fica evidente que há apenas uma pequena proporção de arquivos com um número mais elevado de colunas, indicando em geral uma baixa dimensionalidade dos dados.

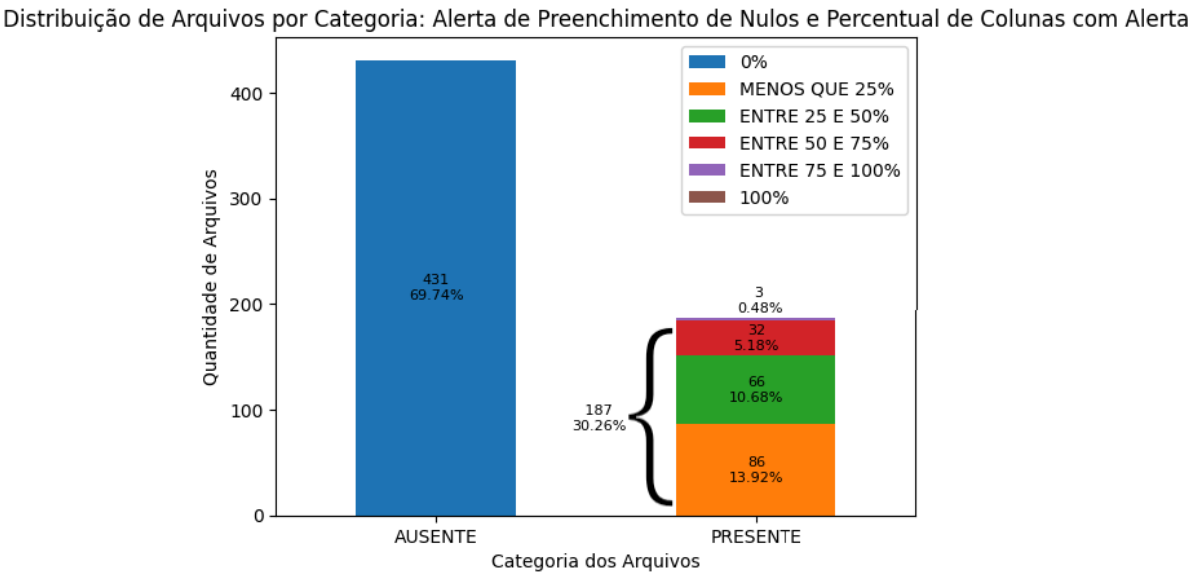
Gráfico 4 – Quantidade de colunas dos arquivos CSV



Os resultados referentes ao alerta de preenchimento das colunas nos arquivos são apresentados no Gráfico 5. Observa-se que uma parcela significativa dos arquivos possui colunas com mais de 50% dos seus campos preenchidos com

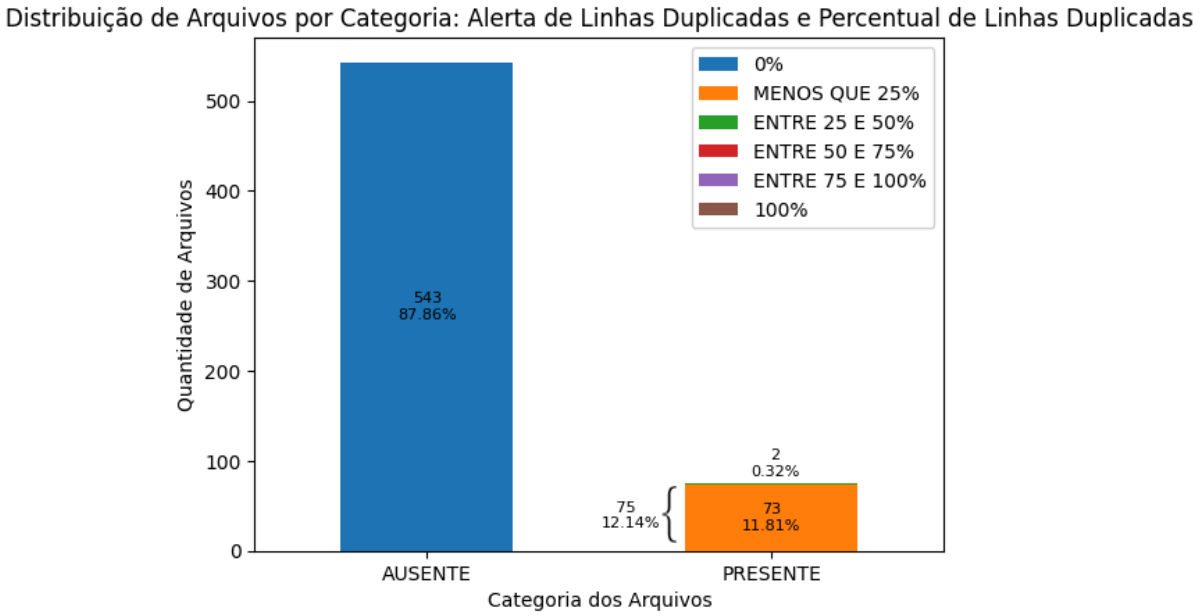
valores nulos. Ao analisarmos mais detalhadamente, percebemos que muitos desses arquivos têm mais de 25% de suas colunas afetadas por esse alerta, o que destaca um ponto crítico em relação à qualidade desses dados.

Gráfico 5 – Alerta de preenchimento de nulos e percentual de colunas com alerta dos arquivos CSV



Os resultados relacionados ao alerta de duplicação de linhas nos arquivos são apresentados no Gráfico 6. Observa-se que apenas uma pequena parcela dos arquivos possui linhas duplicadas em seu conteúdo. Ao analisarmos mais detalhadamente, percebemos que o impacto dessa duplicação é relativamente baixo, ocorrendo na maioria dos casos em menos de 25% dos dados. No entanto, mesmo sendo um alerta de baixo impacto, destaca-se a importância de garantir a unicidade dos dados.

Gráfico 6 – Alerta de duplicação de linhas e percentual de linhas duplicadas dos arquivos CSV



Dessa forma, observa-se que os dados analisados na pesquisa apresentam, em diferentes proporções, os seguintes problemas de qualidade: falha na leitura, alto percentual de valores nulos em colunas, duplicação de linhas e falta de padronização no separador utilizado nos arquivos CSV.

4.3 Conjuntos de Dados

Dos 151 conjuntos de dados baixados, havia um total de 909 arquivos, sendo 631 deles arquivos do tipo CSV e 278 arquivos do tipo JSON. Esses números destacam uma discrepância significativa na quantidade de arquivos CSV em relação aos arquivos JSON, levantando a suspeita inicial de que talvez não tenhamos metadados disponíveis para todos os arquivos CSV.

A análise da quantidade de arquivos por *dataset* está presente nos Gráficos 7, 8, 9 e 10. Observa-se que a maioria dos conjuntos de dados possui no máximo 10 arquivos. Entre os *datasets* que contêm arquivos CSV, a maioria possui entre 1 e 10 arquivos, enquanto para aqueles que incluem arquivos JSON, a maioria possui entre 1 e 3. É importante ressaltar que existem 5 *datasets* sem arquivo CSV e 26 *datasets* sem arquivo JSON. No geral, a maioria dos *datasets* inclui ambos os tipos de arquivos.

Gráfico 7 – Quantidade de arquivos dos *datasets*

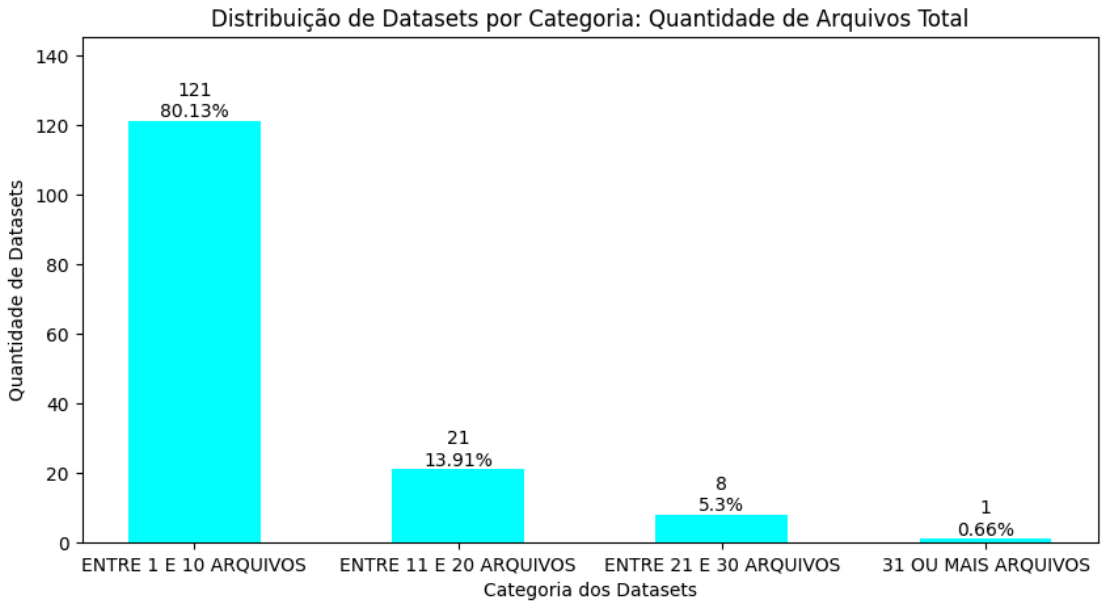


Gráfico 8 – Quantidade de arquivos CSV dos *datasets*

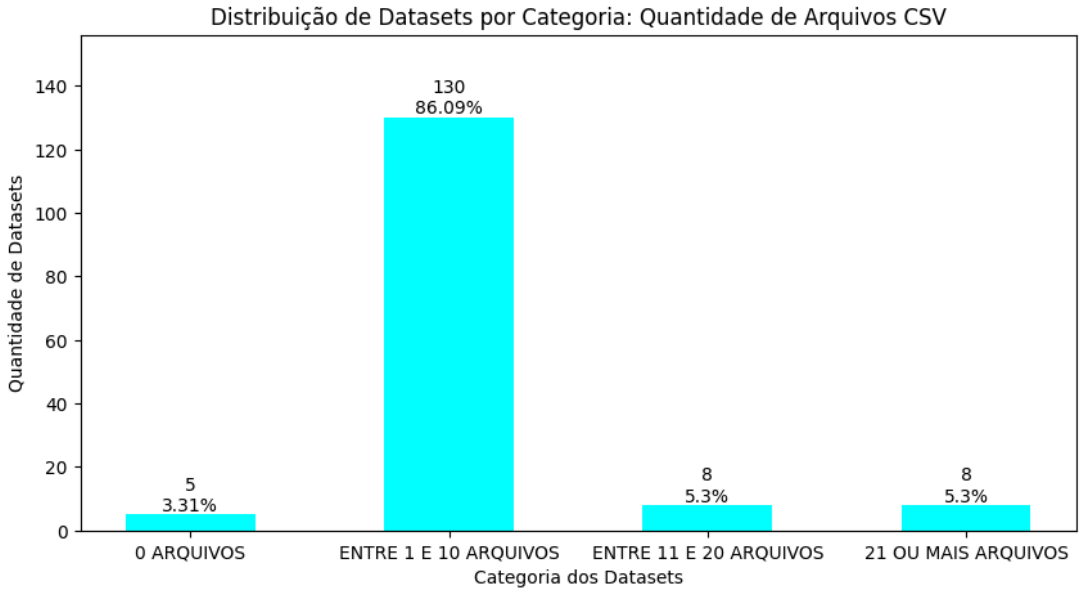


Gráfico 9 – Quantidade de arquivos JSON dos *datasets*

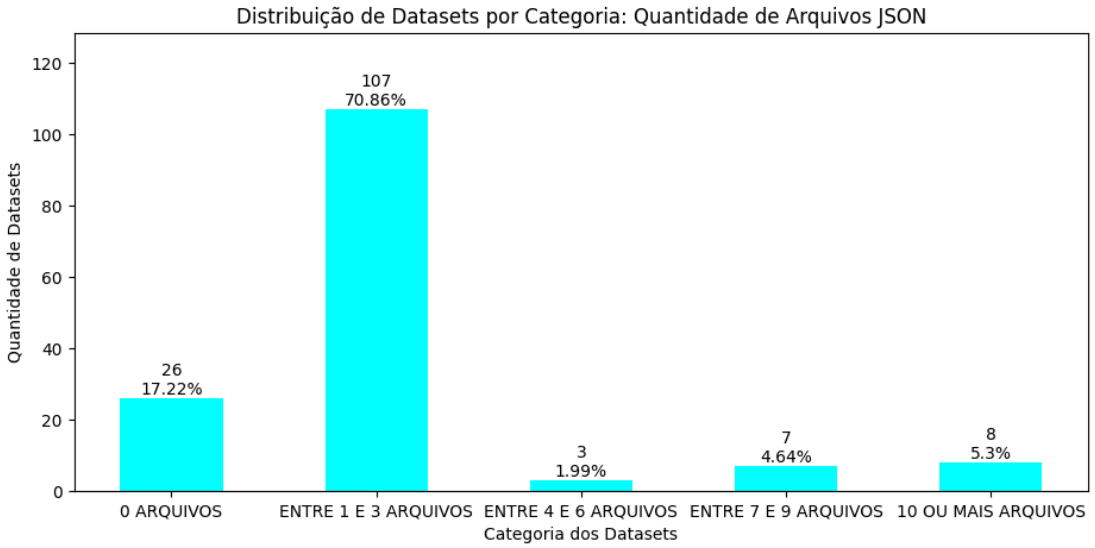
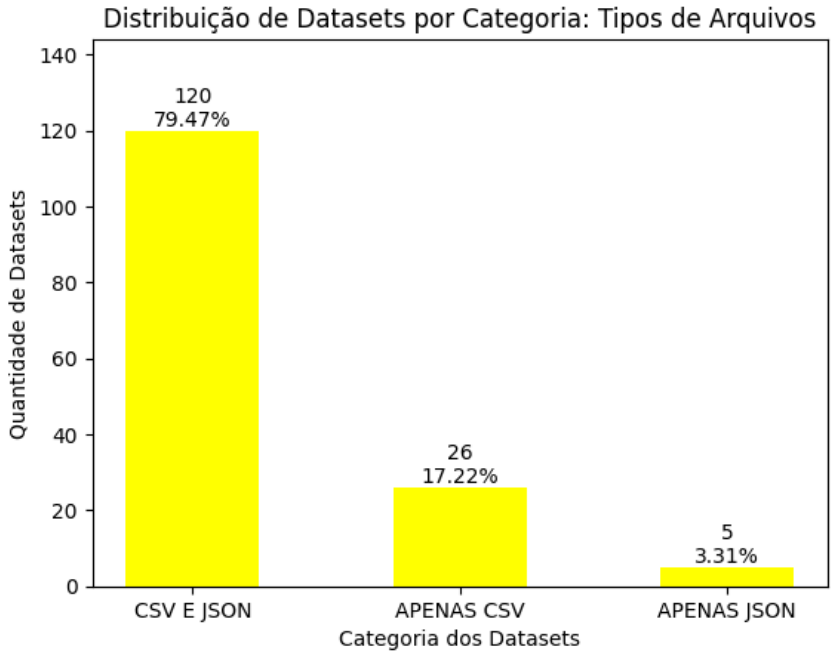
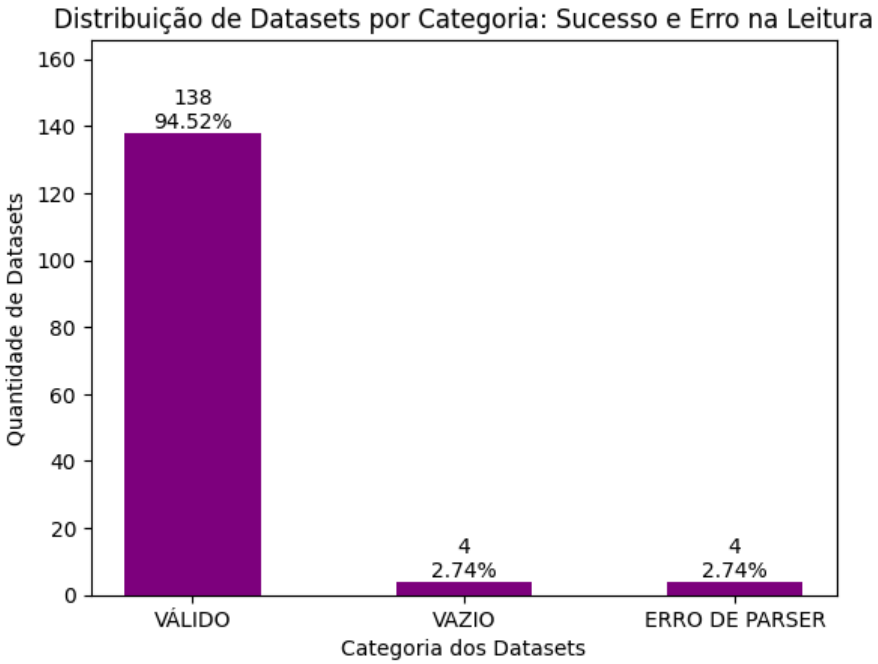


Gráfico 10 – Tipos de arquivos dos *datasets*



As métricas a seguir levam em consideração apenas os *datasets* que possuem pelo menos 1 arquivo CSV. Iniciando os resultados das análises de *dataset* em relação às suas métricas por arquivos CSV, observamos a distribuição dos *datasets* que apresentaram alguma falha na leitura no Gráfico 11. Percebe-se que há apenas 8 *datasets* com problemas de leitura, e nenhum deles apresenta ambos os tipos de erros de leitura.

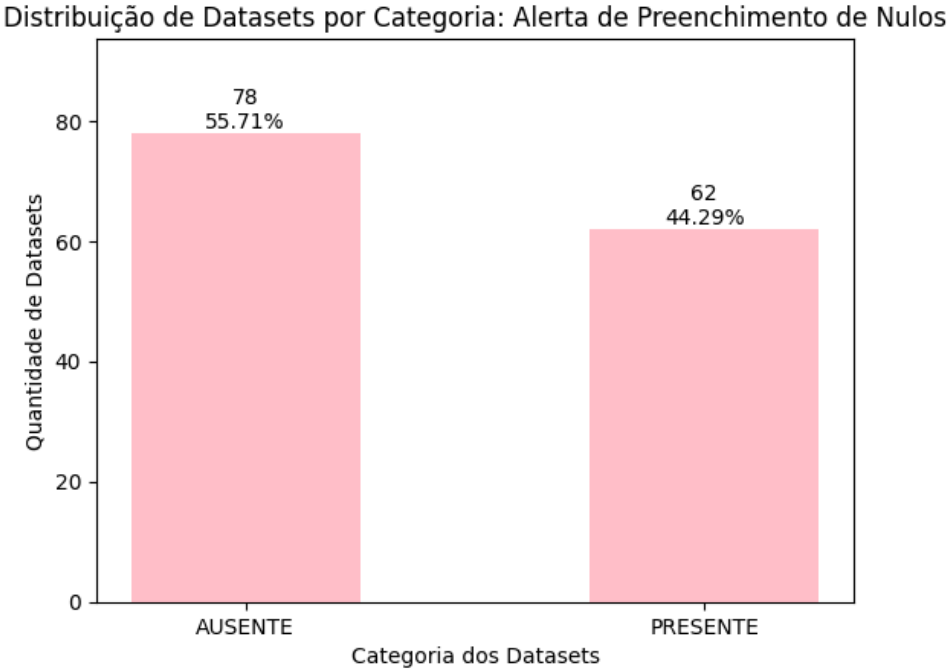
Gráfico 11 – Sucesso ou falha na leitura dos arquivos CSV dos *datasets*



Mapeamos também os *datasets* que possuem todos os seus arquivos CSV apresentando erro de leitura, são eles: Academia Recife; Contratos; Gestão de dados abertos; Licitações; Serviço de atendimento móvel de urgência (SAMU) 2022; Serviço de atendimento móvel de urgência (SAMU) 2023.

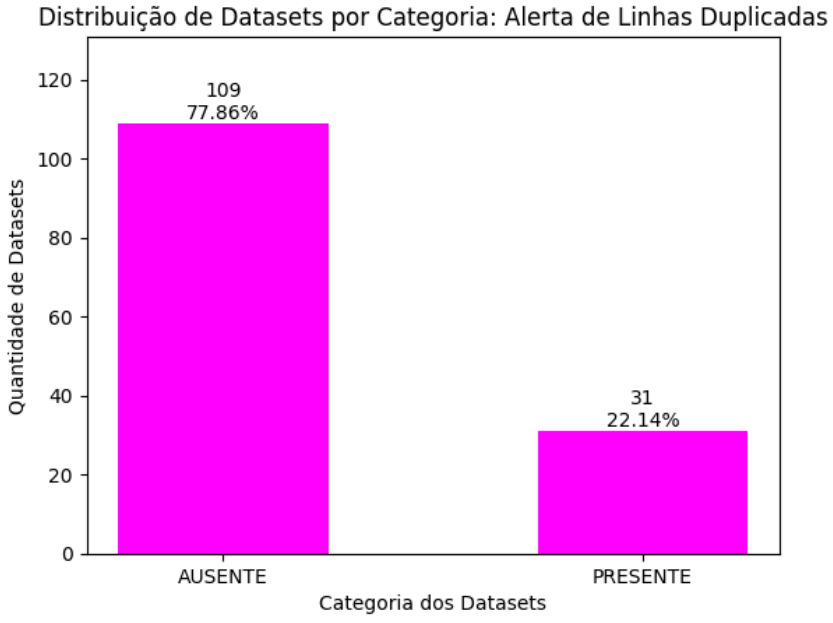
Os resultados referentes aos problemas de preenchimento nos *datasets* estão apresentados no Gráfico 12. Ao generalizarmos esse alerta para os conjuntos de dados, observamos que a divisão é quase equitativa entre aqueles que apresentam ou não esse tipo de problema, o que reforça a preocupação com a qualidade desses dados em relação ao preenchimento.

Gráfico 12 – *Datasets* com arquivos CSV com alerta de preenchimento de nulos



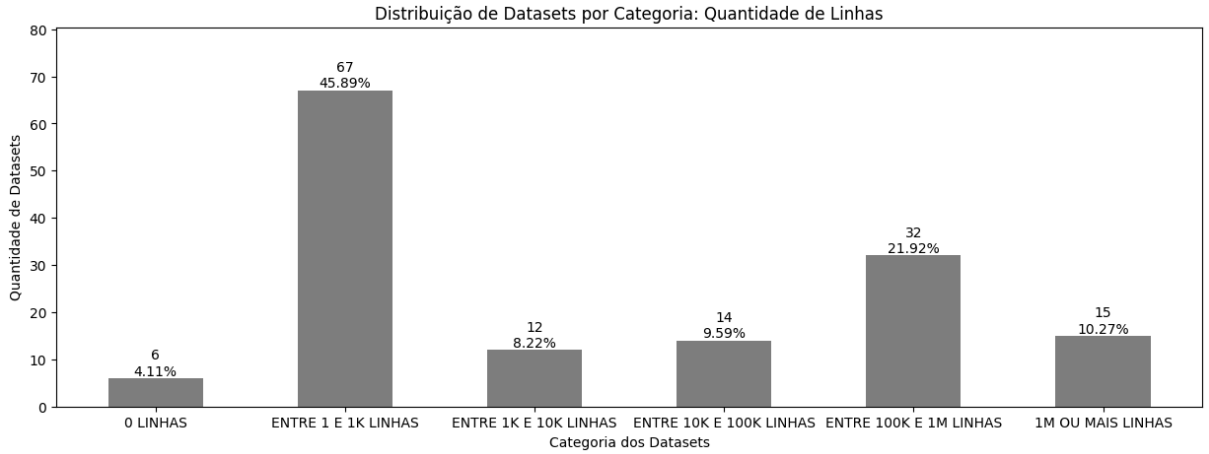
Os resultados referentes aos problemas de duplicação de linhas nos *datasets* estão apresentados no Gráfico 13. Ao generalizarmos por *dataset*, observa-se que uma porcentagem considerável dos *datasets* possui pelo menos um arquivo CSV que apresenta esse tipo de problema. Essa constatação revela que uma parcela considerável dos conjuntos de dados não mantém a unicidade dos registros, o que pode comprometer a integridade e a confiabilidade das informações.

Gráfico 13 – *Datasets* com arquivos CSV com alerta de duplicação de linhas



O resultado da análise da quantidade de linhas por *dataset* está apresentado no Gráfico 14. Os 6 *datasets* que possuem 0 linhas são os datasets que contêm arquivos CSV, mas nenhum deles foi lido com sucesso listados anteriormente. Além disso, é evidente que a maioria dos *datasets* se mantém com até mil linhas. Vale ressaltar também a existência de *datasets* com mais de 1 milhão de linhas, sendo os 5 maiores os seguintes: Registro das infrações de trânsito, Perfil das pessoas vacinadas COVID 19, Velocidade das vias - Quantitativo por velocidade média - 2022, Servidores e Velocidade das vias - Quantitativo por velocidade média - 2019. Esta variação no número de linhas sugere uma diversidade significativa nos tamanhos dos conjuntos de dados analisados.

Gráfico 14 – Quantidade de linhas dos *datasets*



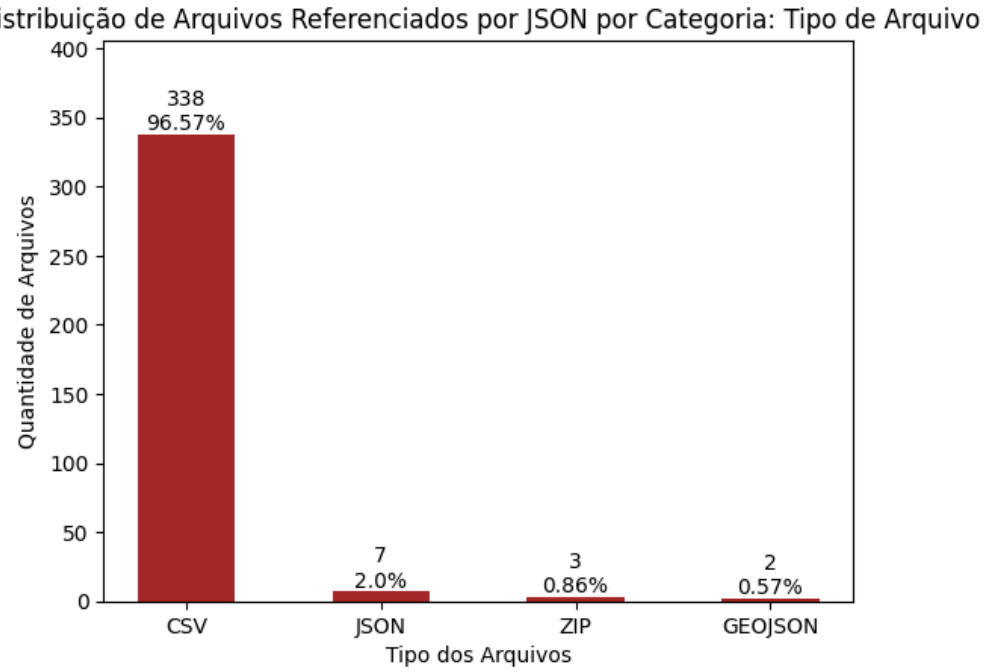
Com isso, observamos uma disparidade na quantidade de arquivos de cada tipo, apesar da maioria dos *datasets* conter ambos os tipos. Embora poucos *datasets* tenham apresentado erro de leitura, os alertas de preenchimento e duplicação mostraram-se bastante presentes.

4.4 Metadados

Durante a fase de coleta, foram obtidos 278 arquivos JSON, dos quais 276 são arquivos de metadados, enquanto apenas 2 contêm dados no formato JSON. Esses 276 referenciam um total de 350 arquivos, aos quais os metadados pertencem. No entanto, dos 276 arquivos de metadados, 264 referem-se a arquivos CSV, enquanto os outros 12 fazem referência a arquivos de outro tipo. Esses 264 arquivos JSON estão referenciando 338 arquivos CSV distintos. Mesmo que conseguíssemos relacionar todas as 338 referências a algum dos arquivos CSV

baixados, ainda restariam arquivos CSV sem metadados. Portanto, fica evidente que os metadados para todos os arquivos não estão disponíveis no portal. O Gráfico 15 demonstra quais tipos de arquivos são referenciados pelos arquivos JSONs de metadados.

Gráfico 15 – Tipos de arquivos referenciados pelos JSON de metadados



Os resultados apresentados a seguir se restringem aos metadados que fazem referência aos arquivos CSV. A análise das etiquetas dos arquivos CSV é demonstrada nos Gráficos 16 e 17, permitindo a observação dos temas mais recorrentes nos dados disponíveis no portal. É evidente o considerável enfoque na área da educação, uma vez que três das dez palavras mais frequentes são "EDUCAÇÃO", "ESCOLAR" e "CENSO ESCOLAR". Além disso, destaca-se a relevância da saúde, visto que "SAÚDE" e "URGÊNCIA" ficaram entre as palavras mais utilizadas. Observa-se ainda um foco significativo no trânsito, com "MOBILIDADE", "CTTU" e "TRÂNSITO" entre as palavras mais recorrentes.

Gráfico 16 – Top 10 etiquetas mais utilizadas para descrição dos arquivos CSV

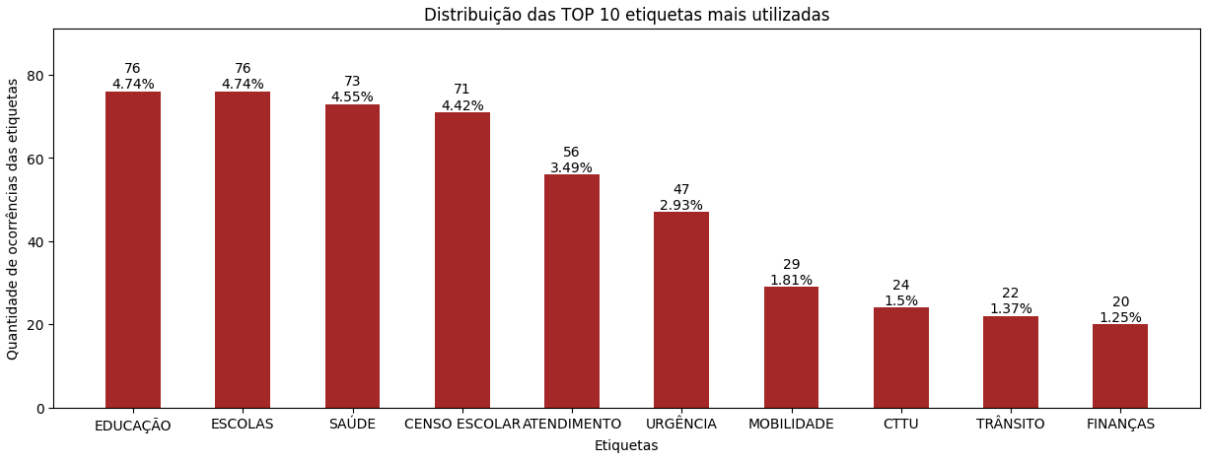
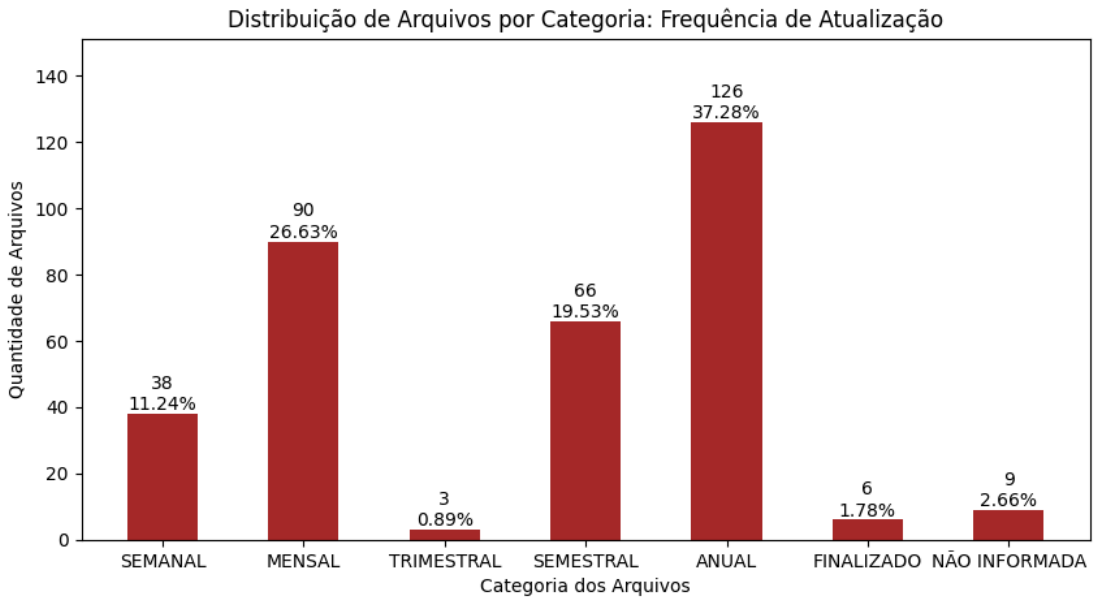


Gráfico 17 – Nuvem de palavras das 100 etiquetas mais utilizadas para descrição dos arquivos CSV

Os resultados referentes à frequência de atualização dos dados do Portal são apresentados no Gráfico 18. Nota-se que a maioria dos dados possui uma atualização anual, enquanto uma parcela considerável é atualizada mensalmente. Uma pequena proporção dos arquivos é atualizada semanalmente e pertence aos seguintes conjuntos de dados: Despesas orçamentárias, Gestão de dados abertos e Receitas. Essa diversidade na periodicidade de atualização dos dados reflete uma variabilidade na disponibilidade de dados atualizados dentro do portal. Destaca-se que apenas 1,78% dos arquivos não receberão atualizações futuras, o que sugere uma constância na manutenção e atualização dos conjuntos de dados. No entanto, é

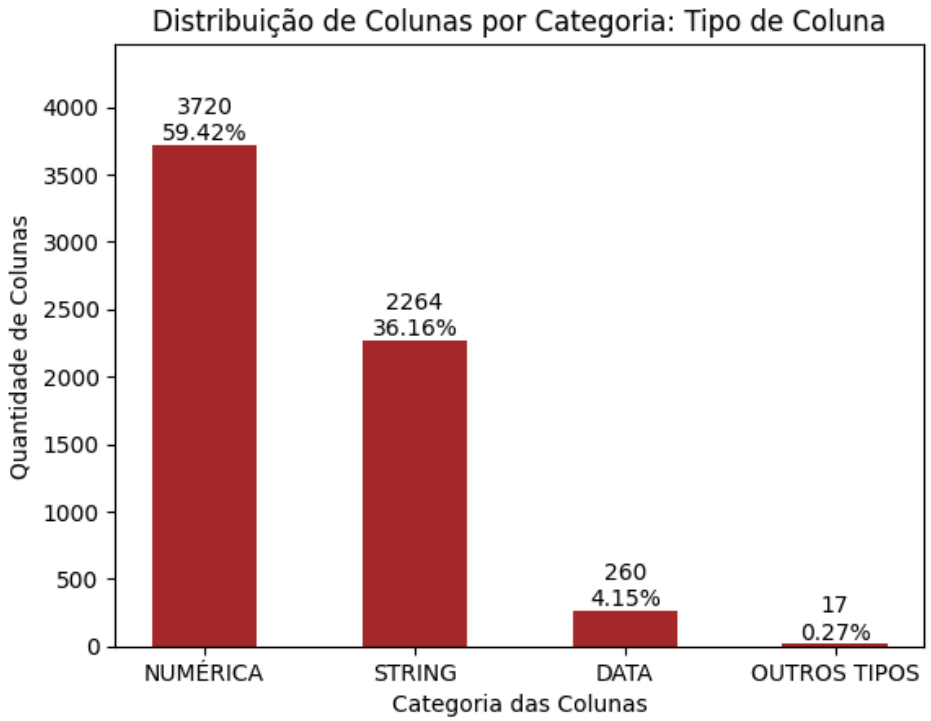
importante ressaltar que 2,66% dos arquivos não tiveram sua frequência de atualização informada, apesar de possuírem metadados associados, indicando uma lacuna na disponibilidade dessas informações.

Gráfico 18 – Frequência de atualização dos arquivos CSV



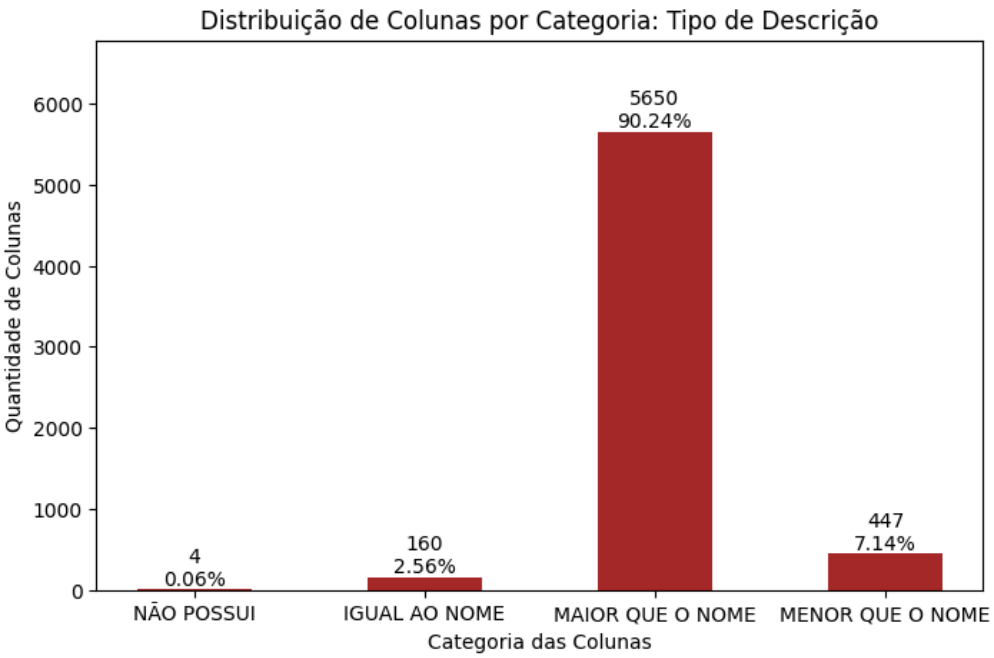
Para compreendermos os tipos de dados predominantes disponíveis no portal, é essencial avaliar os resultados apresentados no Gráfico 19. Observa-se que a maioria das colunas é do tipo numérico, seguido das colunas de *string*, com uma presença menos significativa de colunas de data. A predominância de colunas numéricas indica uma forte presença de dados quantitativos, o que pode ser valioso para aplicações estatísticas e tomada de decisões informadas. Já a escassez de colunas de data pode sugerir uma dificuldade em determinar a referência temporal das informações disponíveis.

Gráfico 19 – Tipos das colunas dos arquivos CSV



Por fim, o resultado da análise da qualidade das descrições das colunas dos arquivos CSV está apresentado no Gráfico 20. Observa-se que quase todas as colunas (próximo de 100%) possuíam alguma descrição associada. Além disso, a maioria dessas descrições é mais extensa do que o próprio nome da coluna, sugerindo a presença de informações adicionais sobre os dados contidos, o que pode oferecer uma maior clareza e utilidade na interpretação dos dados.

Gráfico 20 – Descrições das colunas dos arquivos CSV

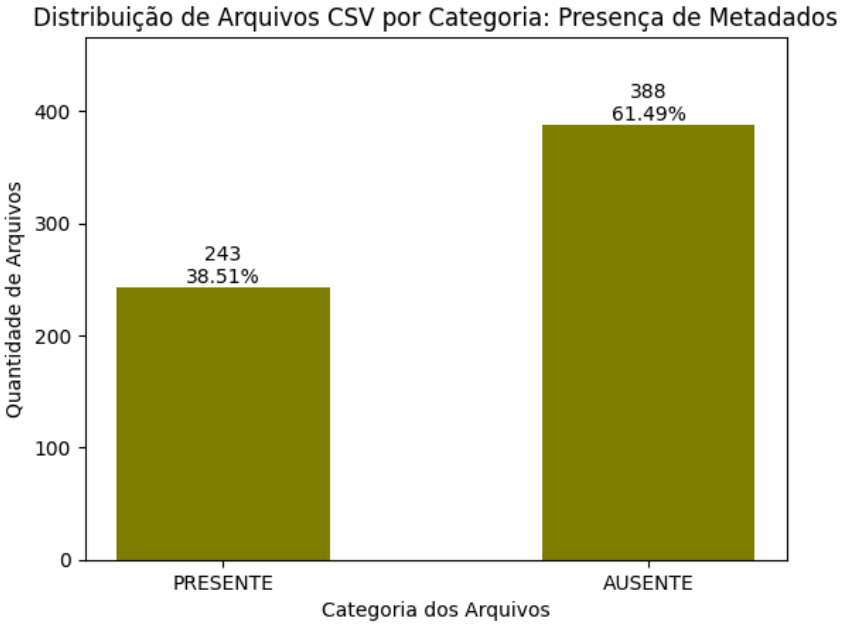


Assim, torna-se evidente que não é viável encontrar metadados para todos os arquivos CSVs baixados. No entanto, os dados coletados permitiram a identificação dos principais temas presentes nos conjuntos de dados, como Educação, Saúde e Trânsito. Observou-se também que os dados possuem uma frequência de atualização bem definida na maioria dos casos, o que representa uma característica relevante. Além disso, ao analisar os tipos e as descrições das colunas, foi possível obter uma compreensão mais detalhada do perfil dos dados disponíveis no portal.

4.5 Associação dos Metadados aos Arquivos CSV

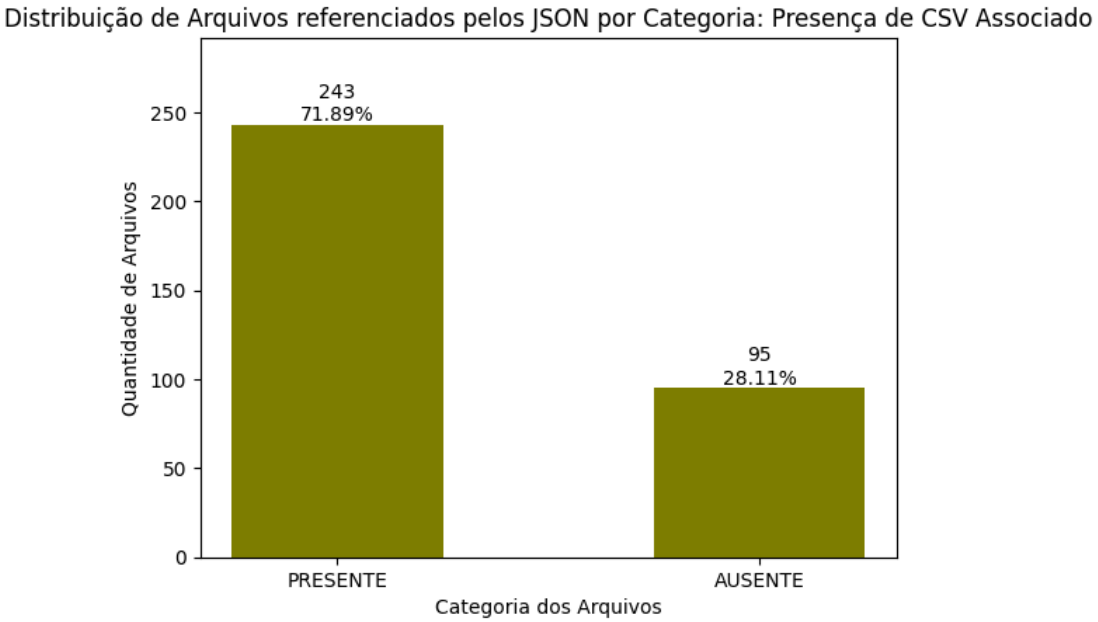
Os resultados relativos à associação dos metadados presentes nos arquivos JSON aos arquivos CSV coletados estão representados no Gráfico 21. Verifica-se que nossa suspeita discutida na seção 4.3 se confirmou: não foi viável encontrar os metadados de todos os arquivos CSV. Na verdade, isso não ocorreu para a maioria deles. Isso levanta um ponto de alerta significativo, pois a falta de metadados dificulta a compreensão dos dados e torna mais desafiador identificar possíveis aplicações.

Gráfico 21 – Associação entre arquivos CSV e metadados



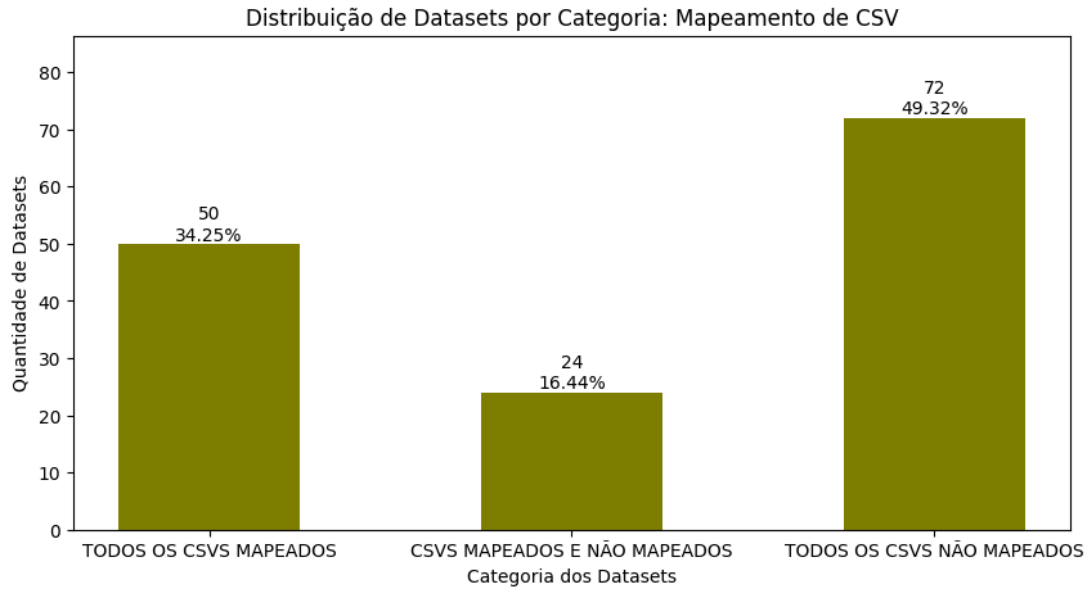
Agora, ao analisarmos os resultados pelo caminho inverso, a quantidade de arquivos CSV referenciados pelos JSONs que conseguiram ou não ser associados a um dos arquivos baixados está apresentada no Gráfico 22. A considerável quantidade de correspondências que não foram bem-sucedidas ocorre porque, em alguns casos, o nome do arquivo referenciado pelo JSON não corresponde exatamente ao nome do arquivo CSV baixado. Portanto, para obter uma correspondência precisa, seria necessário aplicar algum método de similaridade de *strings* e associar os metadados ao arquivo que possua um nome semelhante ao que está sendo referenciado. Esse é um ponto de melhoria importante para o Portal, de forma a facilitar a identificação de metadados.

Gráfico 22 – Associação entre arquivos CSV referenciados pelos metadados e arquivos CSV baixados



O resultado da quantidade de *datasets* em que todos, alguns ou nenhum dos seus arquivos CSV foram associados a metadados é apresentado no Gráfico 23. Observamos que quase metade dos *datasets* não teve nenhum dos seus arquivos CSV associados a metadados, o que sugere a falta de informações adicionais para uma parcela significativa dos dados. Esse cenário reforça a necessidade de o portal priorizar a disponibilização de metadados para todos os arquivos, visando melhorar a qualidade e a utilidade dos conjuntos de dados disponíveis.

Gráfico 23 – *Datasets* e sua cobertura de metadados



Focando na análise dos *datasets* que possuem pelo menos um arquivo CSV não associado a metadados, os resultados das etiquetas mais frequentes estão representados nos Gráficos 24 e 25. Observa-se que seis das dez palavras mais frequentes mencionadas na seção 4.4 se repetem, enquanto surgem quatro novas palavras: DESPESAS, ALINEA, RECEITA e UNIDADE. As palavras que permanecem são: ESCOLAS, EDUCAÇÃO, CENSO ESCOLAR, SAÚDE, TRÂNSITO e FINANÇAS. Notavelmente, as palavras ATENDIMENTO, URGENCIA, MOBILIDADE e CTTU não aparecem mais. Isso sugere que a falta de metadados está concentrada principalmente na área de Educação, representada pelas três palavras mais frequentes.

Gráfico 24 – Top 10 etiquetas mais utilizadas para descrição de arquivos em *datasets* com pelo menos um arquivo CSV sem metadados

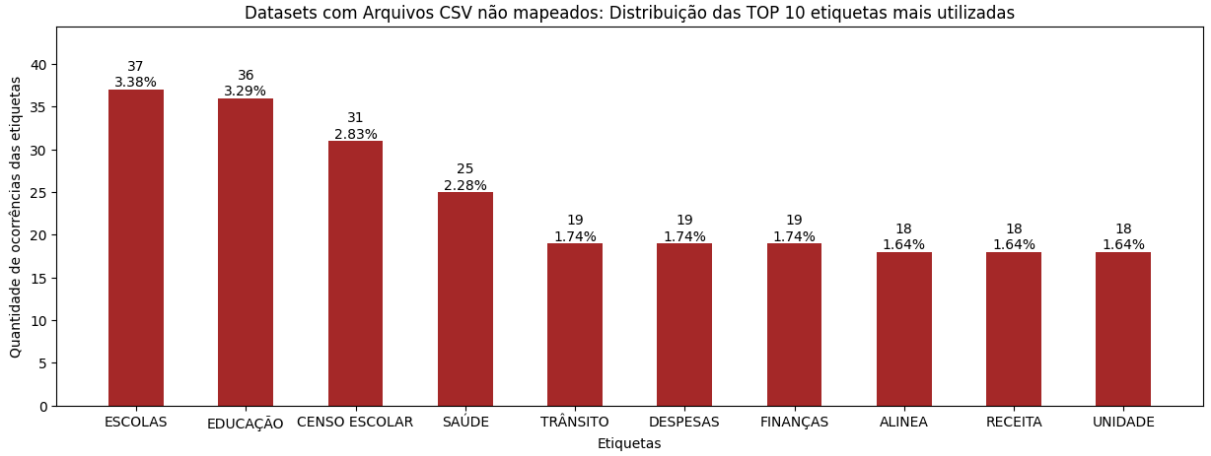


Gráfico 25 – Nuvem de palavras das 100 etiquetas mais utilizadas para descrição de arquivos dos *datasets* com pelo menos um arquivo CSV sem metadados

Com base nessas análises, podemos concluir que não foi viável encontrar metadados para todos os arquivos CSV baixados, evidenciando uma lacuna significativa na disponibilidade dessas informações no portal. Essa questão destaca-se como uma área de melhoria crucial para garantir a qualidade dos dados disponíveis. Além disso, a inconsistência nos nomes dos arquivos CSV baixados e dos arquivos CSV referenciados pelo JSON complicou ainda mais a associação entre eles, contribuindo para a ausência de metadados em muitos dos conjuntos de dados. Essa falta de informações essenciais compromete a compreensão e a utilidade dos dados disponíveis.

Um resumo dos resultados encontrados neste capítulo está apresentado na Tabela 3, com o objetivo de fornecer uma visão geral das descobertas e conclusões obtidas.

Tabela 3 – Resumo dos resultados

Característica	Dimensões Impactadas	Impacto
Ausência de arquivo JSON com metadados em parte dos conjuntos de dados.	Completude, consistência e entendimento.	Negativo.

Erros de leitura em arquivos CSV.	Completude, livre de erros e representação.	Negativo.
Ausência de padrão na escolha do separador de campos nos arquivos CSV.	Consistência.	Negativo.
Arquivos CSV com linhas duplicadas.	Precisão e livre de erros.	Negativo.
Arquivos CSV com alta presença de campos nulos.	Precisão, completude e livre de erros.	Negativo.
Ausência de metadados para todos os arquivos CSV.	Entendimento.	Negativo.
Atualização regular dos arquivos.	Atualidade.	Positivo.
Identificação dos temas mais abrangentes: Educação, Saúde e Trânsito.	Entendimento.	Positivo.
Identificação da predominância de dados numéricos.	Entendimento.	Positivo.
Presença de descrições para as colunas.	Entendimento.	Positivo.

5 CONCLUSÃO

O principal objetivo deste estudo foi avaliar a qualidade dos dados disponíveis no Portal de Dados Abertos da Cidade do Recife, com foco nos arquivos CSV e seus metadados associados. A pesquisa foi motivada pela relevância de dados de qualidade na tomada de decisões, mitigação de riscos, análises estatísticas e outras iniciativas importantes para pesquisadores e gestores públicos.

Durante o processo de coleta, foram obtidos todos os conjuntos de dados e arquivos de interesse disponíveis no site naquele período. Subsequentemente, o site foi atualizado com novos dados, os quais não foram abrangidos pela pesquisa. No entanto, a coleta inicial contemplou uma quantidade significativa de dados, totalizando mais de 10GB, proporcionando uma base substancial para a análise e elaboração de conclusões.

Ao analisarmos os *datasets*, observamos que nem todos possuíam arquivos JSON com metadados, evidenciando uma lacuna na plataforma nesse aspecto. Isso gera alertas nas dimensões de qualidade de completude e consistência. Ao examinarmos os arquivos CSV individualmente, identificamos a ocorrência de problemas de leitura, embora em uma proporção reduzida. Esses problemas se enquadram nas dimensões de completude, livre de erros e representação. Além disso, observamos uma falta de padronização na escolha do separador de campos nos arquivos, levantando mais um alerta para a dimensão consistência. Os problemas de duplicação eram mais comuns, porém ainda afetavam apenas uma pequena parte dos dados. No entanto, o alerta mais significativo em relação à qualidade recaiu sobre a frequente presença de campos não preenchidos, levantando preocupações nas dimensões de precisão, completude e livre de erros. Todos esses são pontos de melhoria a serem considerados.

Com a análise dos metadados, conseguimos identificar os temas mais abrangentes nos conjuntos de dados disponíveis, destacando-se Educação, Saúde e Trânsito. Concluímos que a maioria dos metadados indicava uma programação regular de atualização dos arquivos, o que reflete positivamente na dimensão de qualidade atualidade. Observamos também uma predominância de dados numéricos, sugerindo possíveis aplicações estatísticas. Além disso, as descrições das colunas, em sua maioria, forneciam informações adicionais além do simples

nome da coluna, indicando uma qualidade satisfatória dos metadados encontrados, contribuindo positivamente para a dimensão entendimento.

No entanto, é crucial ressaltar a ausência de metadados para todos os arquivos CSV, juntamente com a dificuldade em associar integralmente os arquivos referenciados pelos metadados aos arquivos coletados, devido às possíveis disparidades entre os nomes dos arquivos coletados e os nomes referenciados pelos JSONs. Tal lacuna representa uma desvantagem considerável no que concerne à dimensão de entendimento. Este é um aspecto de extrema relevância que deve ser priorizado pelo município de Recife, visando aprimorar a compreensão dos dados disponibilizados.

Portanto, conclui-se que os dados públicos do Município de Recife apresentam deficiências de qualidade que requerem atenção e correção por parte do órgão responsável. Os alertas identificados abrangeram diversas dimensões de qualidade, incluindo precisão, completude, livre de erros, representação, consistência e entendimento. A abordagem destas questões possibilitará uma utilização mais eficaz e abrangente desses dados, beneficiando múltiplos setores e facilitando sua aplicação em diversos contextos.

5.1 Trabalhos Futuros

Como próximos passos, visando alcançar uma associação mais precisa entre os metadados existentes e os arquivos CSV, uma estratégia sugerida é modificar o *scraper* para mapear os códigos identificadores, usados internamente pelo Portal, vinculados a cada um dos arquivos. Isso permitirá realizar a correspondência por identificador, em vez do nome do arquivo, supondo que o identificador seja universal dentro do sistema de gerenciamento dos dados do Portal, ao contrário do nome que pode variar.

Além disso, uma extensão do projeto pode conter o estudo e a implementação de técnicas de tratamento de dados destinadas a corrigir alguns dos problemas de qualidade identificados, como a presença de campos nulos e a duplicação de linhas. Dessa forma, seria possível aplicar tais técnicas nos arquivos que apresentam esses problemas e fornecer os dados tratados para futuras aplicações.

REFERÊNCIAS

1 OPEN KNOWLEDGE FOUNDATION. Open Definition. 2015. Página inicial. Disponível em: <https://opendefinition.org/>. Acesso em: 5 fev. 2024.

2 MIRANDA, Igor et al. Machine learning prediction of hospitalization due to COVID-19 based on self-reported symptoms: A study for Brazil. In: **2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)**. IEEE, 2021. p. 1-5.

3 WOLLENSTEIN-BETECH, Salomon et al. Physiological and socioeconomic characteristics predict COVID-19 mortality and resource utilization in Brazil. **PloS one**, v. 15, n. 10, p. e0240346, 2020.

4 KILKENNY, Monique F.; ROBINSON, Kerin M. Data quality: “Garbage in–garbage out”. **Health Information Management Journal**, v. 47, n. 3, p. 103-105, 2018.

5 PREFEITURA DO RECIFE. Portal de Dados Abertos da Prefeitura da Cidade do Recife. [S.l]. Sobre o Portal. Disponível em: <http://dados.recife.pe.gov.br/about>. Acesso em: 5 fev. 2024

6 PREFEITURA DO RECIFE. Portal de Dados Abertos da Prefeitura da Cidade do Recife. [S.l]. Perguntas frequentes. Disponível em: <http://dados.recife.pe.gov.br/perguntas-frequentes>. Acesso em: 5 fev. 2024.

7 MALAMUD, C. et al. The Annotated 8 Principles of Open Government Data. 2007. The 8 Principles of Open Government Data. Disponível em: <https://opengovdata.org/>. Acesso em: 5 fev. 2024.

8 PARSSIAN, Amir; SARKAR, Sumit; JACOB, Varghese S. Assessing data quality for information products: impact of selection, projection, and Cartesian product. **Management Science**, v. 50, n. 7, p. 967-982, 2004.

9 STRONG, Diane M.; LEE, Yang W.; WANG, Richard Y. Data quality in context. **Communications of the ACM**, v. 40, n. 5, p. 103-110, 1997.

10 BALLOU, Donald P.; PAZER, Harold L. Modeling data and process quality in multi-input, multi-output information systems. **Management science**, v. 31, n. 2, p. 150-162, 1985.

11 HUANG, Kuan-Tse; LEE, Yang W.; WANG, Richard Y. **Quality information and knowledge.** Prentice Hall PTR, 1998.

12 REDMAN, Thomas C. **Data quality for the information age.** Artech House, Inc., 1997.

13 WANG, Richard Y.; STRONG, Diane M. Beyond accuracy: What data quality means to data consumers. **Journal of management information systems**, v. 12, n. 4, p. 5-33, 1996.

14 PARKER, Michelle B. et al. An evaluation of Information quality frameworks for the World Wide Web. 2006.

15 ARAUJO, Marcelo Henrique de et al. Dados governamentais abertos: uma análise sob a ótica das dimensões de qualidade da informação. **Anais**, 2012.

16 PIPINO, Leo L.; LEE, Yang W.; WANG, Richard Y. Data quality assessment. **Communications of the ACM**, v. 45, n. 4, p. 211-218, 2002.

17 PREFEITURA DO RECIFE. Portal de Dados Abertos da Prefeitura da Cidade do Recife. 2023. Sistemática de Atualização de Dados do Portal de Dados Abertos do Recife. Disponível em: <http://dados.recife.pe.gov.br/base/doc/Sistematica_de_Atualizacao_de_Dados_do_Portal_de_Dados_Abertos_da_Prefeitura_do_Recife-v1.pdf>; Acesso em: 7 fev. 2024.

18 AZEROUAL, Otmane; ABUOSBA, Mohammad. Improving the data quality in the research information systems. **arXiv preprint arXiv:1901.07388**, 2019.

19 ROSTALSKI, Frauke et al. **Künstliche Intelligenz: Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?**. Mohr Siebeck, 2022.

20 BARBOSA, Luciano et al. Structured open urban data: understanding the landscape. **Big data**, v. 2, n. 3, p. 144-154, 2014.

21 SILVA, Ambrozina de Abreu Pereira; MONTEIRO, Doraliza Auxiliadora Abranches; DE OLIVEIRA REIS, Anderson. Qualidade da Informação dos dados governamentais abertos: análise do portal de dados abertos brasileiro. **Revista Gestão em Análise**, v. 9, n. 1, p. 31-47, 2020.

22 MARETTO, Ciro X.; BARCELLOS, Monalessa P. Uma Avaliação de Dados Abertos Educacionais Brasileiros: Qualidade, Privacidade e Learning Analytics. In: **Anais do XXXIV Simpósio Brasileiro de Informática na Educação**. SBC, 2023. p. 1341-1350.

23 KHDER, Moaiad Ahmad. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. **International Journal of Advances in Soft Computing & Its Applications**, v. 13, n. 3, 2021.

24 RICHARDSON, Leonard. Beautiful soup documentation. c2004-2015. Disponível em: <https://beautiful-soup-4.readthedocs.io/en/latest/#>. Acesso em: 9 fev. 2024.

25 REITZ, Kenneth. Requests: HTTP for Humans. [S.l]. Disponível em: <https://requests.readthedocs.io/en/latest/>. Acesso em: 9 fev. 2024.

26 GOOGLE. Damos-lhe as boas-vindas ao Colab!. [S.l]. Disponível em: <https://colab.research.google.com/>. Acesso em: 9 fev. 2024.

27 JUPYTER TEAM. Jupyter Notebook: The Classic Notebook Interface. c2024. Disponível em: <https://jupyter.org/>. Acesso em: 9 fev. 2024.

28 PYTHON SOFTWARE FOUNDATION. OS - Miscellaneous operating system interfaces. c2001-2024. Disponível em: <https://docs.python.org/3/library/os.html>. Acesso em: 9 fev. 2024.

29 NUMFOCUS, INC. Pandas documentation. c2024. Disponível em: <https://pandas.pydata.org/docs/>. Acesso em: 15 fev. 2024.

30 PYTHON SOFTWARE FOUNDATION. csv - CSV File Reading and Writing. c2001-2024. Disponível em: <https://docs.python.org/3/library/csv.html>. Acesso em: 15 fev. 2024.

31 PYTHON SOFTWARE FOUNDATION. json - JSON encoder and decoder. c2001-2024. Disponível em: <https://docs.python.org/3/library/json.html>. Acesso em: 15 fev. 2024.

32 MATPLOTLIB DEVELOPMENT TEAM. Matplotlib: Visualization with Python. c2012-2023. Disponível em: <https://matplotlib.org/>. Acesso em: 15 fev. 2024.