**4.2.3.1 Importação e preparação dos dados**

Após o compartilhamento dos dados coletados pelos investigadores, houve necessidade de normalização e padronização dos mesmos. Tal processo é definido como limpeza, preparação e tratamento de dados, uma das etapas do processo de mineração de dados (Hai, Kamber e Pei, 2011), correspondendo a fase de pré-processamento (ou *arrumar*) da Figura NNN. Ainda na etapa inicial de pré-processamento dos *dados brutos* e com o objetivo de seguir os objetivos e fundamentar com a análise dos dados a resposta à pergunta de pesquisa desta tese, algumas colunas das planilhas originais foram eleitas, logo, não foram utilizadas todas as variáveis (ou colunas), dos trabalhos apurados do Quadro nnn.

Assim, para esta investigação foram selecionados os campos: nome do pesquisador; país onde foi aplicado o questionário; gênero e idade do respondente; e o estilo de uso do espaço virtual; variáveis estas que foram considerados relevantes para análise e interpretação.

De acordo com Wickham (2014), 80% do processo de análise de dados é consumido no processo de limpeza e preparação dos dados. Segundo o autor, a preparação de dados não é apenas uma primeira etapa, mas deve ser repetida muitas vezes ao longo da análise, à medida que novos problemas surgem ou novos dados são coletados e acrescentados. Apesar da quantidade de tempo que leva, ainda existem poucos trabalhos de pesquisa sobre como limpar bem os dados, processo também definido como *higienização*. (da SILVA, 2021).

Parte do desafio é a amplitude de atividades que a higienização de dados engloba atividades, tais como: verificação de valores discrepantes; análise das datas informadas; atribuição de valores ausentes; dentre outras define da Silva (2021).

Para lidar com tal dificuldade, esta tese se concentrou, primeiramente em um aspecto importante da limpeza de dados, organizando os dados compartilhados e estruturando o conjunto de dados para facilitar a posterior análise.

Para tanto, a pesquisadora utilizou-se do apoio oferecido pelos recursos humanos e técnicos do Laboratório de Inteligência de Dados – LinDa, grupo multidisciplinar de pesquisa da UNESP-Bauru que tem como finalidade apoiar e desenvolver pesquisas relacionadas à “ciência de dados” e, em particular, estudar estruturas de dados e suas aplicações onde seja possível extrair "inteligência".

Inicialmente, os dados recebidos foram compilados pela pesquisadora em um *workbook* utilizando o aplicativo de computador (software) Excel® da Microsoft®, constituído de diferentes tabelas formadas com todas as colunas e linhas das planilhas originais (*worksheets*) compartilhadas. Uma amostra do conteúdo das planilhas de trabalho em Excel® com os dados brutos, ou *dados não numericamente organizados*, é mostrada na Figura zzz.

Figura zzz. Dados não numericamente organizados.

|  |  |
| --- | --- |
| Tabela  Descrição gerada automaticamente | Interface gráfica do usuário, Aplicativo, Tabela, Excel  Descrição gerada automaticamente |
| Interface gráfica do usuário, Aplicativo, Tabela, Excel  Descrição gerada automaticamente | Interface gráfica do usuário, Aplicativo, Tabela, Excel  Descrição gerada automaticamente |

Fonte: Elaborado pela pesquisadora (2022).

Como pode ser observado na Figura zzz, algumas tabelas recebidas continham os dados primários recolhidos nas pesquisas diretamente dos questionários CEUEVE aplicados, outros foram enviados com dados faltantes e outros já pré-processados com a decodificação dos estilos de uso.

A próxima etapa do processo se consistiu na *preparação dos dados* recebidos, ainda utilizando os recursos do software Excel®. Segundo Wickham (2014), a preparação de dados é o processo de reunir, combinar, estruturar e organizar os dados brutos para que possam ser usados nos procedimentos de AED e também para os processos de visualização desses dados. Os componentes da preparação de dados incluem, de acordo com da Silva (2021): o pré-processamento de dados; documentação dos procedimentos utilizados; criação dos dicionários de dados; limpeza; validação e transformação; e envolve também buscar nas fontes externas originais dos dados a complementação dos dados faltantes (*missing data*).

Para finalizar esta etapa, em algumas das investigações recebidas a pesquisadora retornou o contato com os pesquisadores dos trabalhos selecionados para solicitar o envio dos dados faltantes bem como orientação e apoio no preenchimento desses campos, uma tarefa manual e criteriosa que demandou várias semanas.

Após esta etapa inicial, os dados compartilhados foram compilados e geraram uma versão de trabalho do conjunto de dados em formato Excel®. Uma amostra do *dataset* pode ser visualizada na Figura aaa.

Figura aaa. Amostra do *dataset* inicial de trabalho

Interface gráfica do usuário, Aplicativo, Tabela, Excel

Descrição gerada automaticamente

Fonte: Elaborado pela pesquisadora (2022).

**4.2.3.2 Tratamento dos dados**

Segundo da Silva (2021), um dos principais objetivos da preparação de dados é garantir que os dados brutos que estão sendo preparados para processamento e análise sejam precisos e consistentes para que os resultados dos aplicativos de análise sejam válidos.

Como visto na Figura zzz, os dados geralmente são criados com valores ausentes, imprecisões ou outros erros, e conjuntos de dados separados geralmente têm formatos diferentes que precisam ser reconciliados quando combinados. Corrigir erros de dados, validar a qualidade dos dados e consolidar conjuntos de dados são grandes partes dos projetos de preparação de dados, afirma Wickham (2014).

Após a fase inicial de preparação de dados, a próxima etapa envolve encontrar dados relevantes para garantir que os aplicativos de análise forneçam informações significativas e *insights* acionáveis para a tomada de decisões.

Os dados nesta nova etapa geralmente são enriquecidos e otimizados para torná-los mais informativos e úteis - por exemplo, combinando conjuntos de dados internos e externos -, criando novos campos de dados, eliminando valores discrepantes e abordando conjuntos de dados desequilibrados que podem distorcer os resultados da análise. Esta fase, de acordo com Wickham (2014), da Silva (2021) e Askham (2022), pode ser definida como **tratamento dos dados**.

Portanto, na fase de tratamento, os dados são estruturados e organizados para estar de acordo com os requisitos da ***Análise Exploratória de Dados*** (**AED**). Além de serem estruturados, afirma da Silva (2021), os dados normalmente devem ser transformados em um formato padronizado e utilizável, como por exemplo, a criação de novas variáveis ou colunas que agreguem valores das existentes. O *enriquecimento* de dados, por outro lado, aprimora e otimiza os *datasets* conforme necessário, por meio de medidas como aumentar e adicionar dados, define Askham (2022).

*Datasets*, são definidos por Wickham e Grolemund (2017) como grupos de dados relacionados compostos por uma ou mais tabelas, mas que podem ser manipuladas como uma unidade por um computador.

Na etapa de tratamento, rotinas automatizadas desenvolvidas em um ambiente de desenvolvimento integrado são executadas nos dados para transformar, organizar e estruturar e validar sua consistência, integridade e precisão (WICKHAM , 2014). Os dados assim preparados são então armazenados em um repositório e usados diretamente para a AED ou então disponibilizados para que outros usuários possam também utilizá-los.

Para a realização das etapas de preparação, transformação, organização e estruturação serão utilizados nesta tese os recursos da **Linguagem R,** um *software livre* de programação voltada ao tratamento, análise e visualização de dados. A linguagem é largamente utilizada por acadêmicos, estatísticos e analistas de dados para realizar projetos de AED. Foi criada originalmente no departamento de Estatística da Universidade de Auckland, Nova Zelândia e sua manutenção é realizada por uma comunidade de colaboradores voluntários que contribuem para melhoria e otimização da linguagem e com a expansão das suas funcionalidades básicas por meio de bibliotecas, ou *packages* residentes no repositório CRAN-R (The Comprehensive R Archive Network)[[1]](#footnote-1) (LINGUAGEM R**,** 2022**).**

Com o objetivo de facilitar e agilizar o processo de desenvolvimento dos cálculos estatísticos e a confecção dos gráficos na linguagem R apresentados nesta tese, utilizou-se também o Ambiente de Desenvolvimento Integrado (do inglês *Integrated Development Environment* – IDE) **RStudio**, um programa de computador de fonte aberta com características e ferramentas de apoio ao desenvolvimento acelerado de software (LINGUAGEM R**,** 2022**).**

Portanto, como tarefa inicial para a organização e estruturação dos dados, no Quadro yyy é possível visualizar o dicionário de dados com os nomes e as definições das variáveis utilizadas na AED desta tese.

Um glossário ou dicionário de dados, de acordo com Askham (2022), é utilizado para armazenar um conjunto limitado de metadados[[2]](#footnote-2) concentrados nos nomes e definições relacionadas aos dados físicos e objetos relacionados.

Quadro yyy: Dicionário de dados.

|  |  |  |  |
| --- | --- | --- | --- |
| Nome da variável | Tipo de variável | Conteúdo | Descrição |
| **pesquisador** | texto |  | Nome do(s) pesquisador(es) responsável(is) pela pesquisa |
| **data.ano** | data | dd/mm/aaaa | Dia, mês e ano em que o questionário foi aplicado |
| **pais** | texto | Brasil; Espanha; México ou Portugal | Países onde o questionário CEUEV foi aplicado |
| **genero** | texto | Feminino; Masculino ou NR | Gênero do inquirido |
| **idade** | texto | de 11 a 14 anos; de 14 a 17 anos; de 17 a 20 anos;  de 20 a 30 anos; de 30 a 40 anos; de 40 a 50 anos;  de 50 a 60 anos; de 60 a 70 anos; acima de 70 anos. | Faixa etária dos respondentes |
| **estilo** | texto | **Estilo de Uso A**: Uso Participativo no Espaço Virtual;  **Estilo de Uso B**: Busca e Pesquisa no Espaço Virtual;  **Estilo de Uso C**: Estruturação e Planejamento no Espaço Virtual;  **Estilo de Uso D**: Ação Concreta e Produção no Espaço Virtual;  **Não Identificado** | Estilos de uso do espaço virtual definidos em BARROS (2013) |

Fonte: Elaborado pela pesquisadora (2022).

De acordo com Silva, Peres e Boscarioli (2016) a AED compreende, em conceitos estatísticos a análise descritiva, uma ferramenta capaz de descrever ou resumir dados, mostrando aspectos importantes do conjunto de dados (em inglês *datasets*) como o tipo de distribuição associada aos valores mais representativos do conjunto, e permitindo criar visualizações referentes à tais aspectos.

De forma geral, a análise exploratória é uma abordagem à análise de conjuntos de dados de modo a resumir suas características principais, frequentemente com métodos visuais. Um modelo estatístico pode ou não ser usado, mas primariamente a AED tem como objetivo observar o que os dados podem nos dizer além da modelagem formal ou do processo de teste de hipóteses (WICKHAM e GROLEMUND, 2017).

**4.2.3.3 Transformação dos dados**

**REFERÊNCIAS**

**WICKHAM**, H. Tidy Data. Journal of Statistical Software, [S. l.], v. 59, n. 10, p. 1–23, 2014. DOI: 10.18637/jss.v059.i10. Disponível em: https://www.jstatsoft.org/index.php/jss/article/view/v059i10. Acesso em: 26 jul. 2022.

**da SILVA**, D. Higienização de dados: o que é e 7 boas práticas para realizá-la. Blog da Zendesk, 11 ago 2021. Disponível em: <https://www.zendesk.com.br/blog/higienizacao-de-dados/>. Acesso em: 26 jun. 2022.

**LINGUAGEM R**. In: WIKIPÉDIA, a enciclopédia livre. Flórida: Wikimedia Foundation, 2022. Disponível em: <https://pt.wikipedia.org/w/index.php?title=R\_(linguagem\_de\_programa%C3%A7%C3%A3o)&oldid=63783140>. Acesso em: 13 jun. 2022.

ASKHAM. N. Data Glossary or Data Dictionary? The Data Governance Coach. 14 nov. 2017. Disponível em: <https://www.nicolaaskham.com/blog/2017/11/8/what-is-a-data-glossary-and-how-is-it-different-from-a-data-dictionary>. Acesso em 08 mai. 2022.

BARROS, D. M. V. Diseño y aplicación del Cuestionario Estilo de Uso Espacio Virtual (CEUEV). In CUE, J. L. G, Estilos de aprendizaje y otras perspectivas pedagógicas del siglo XXI. México: Colegio de Postgraduados, 2013. P 149-164.

HAN, J.; KAMBER, M.; PEI, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. ISBN 978-0-12-381479-1.

KOTU, V. Data Science: Concepts and Practice. Cambridge: Morgan Kaufmann, 2019.

METADADOS. In: WIKIPÉDIA, a enciclopédia livre. Flórida: Wikimedia Foundation, 2022. Disponível em: <<https://pt.wikipedia.org/w/index.php?title=Metadados&oldid=63544067>>. Acesso em: 08 mai. 2022.

SILVA, Leandro Augusto da, PERES, Sarajane Marques, BOSCARIOLI, Clodis. Introdução à Mineração de Dados com Aplicações em R. Rio de Janeiro: Elsevier, 2016.

TYAGI, Amit Kumar. Data Science and Data Analytics: Opportunities and Challenges. Flórida: CRC Press, 2022.

WICKHAM, Hadley, GROLEMUND, Garret. R for Data Science: Imort, Tidy, Transform, Visualize and Model Data. Sebastopol: O'Relly Media, 2017.

1. <https://cran.r-project.org/>. [↑](#footnote-ref-1)
2. Metadados, ou Metainformação, são dados sobre outros dados. Um item de um metadado pode dizer do que se trata aquele dado, geralmente uma informação inteligível por um computador. Os metadados facilitam o entendimento dos relacionamentos e a utilidade das informações dos dados. METADADOS (2022). [↑](#footnote-ref-2)