ISSN: 2237-826X
ARTIGO | PAPER

Visualização de dados para extração de conhecimento: um estudo de caso

Data visualization for knowledge extraction: a case study

Daniel Sadao Matsuba[1], Adriana Prest Mattedi[2]

[1] *Universidade Federal de Itajubá, Itajubá, Minas Gerais, Brasil. ORCID: http://orcid.org/0000-0001-8954-2436*

[2] *Universidade Federal de Itajubá, Itajubá, Minas Gerais, Brasil. ORCID: http://orcid.org/0000-0002-4605-9134*

Autor para correspondência/*Mail to:* Adriana Prest Mattedi, amattedi@unifei.edu.br

Recebido/*Submitted*: 01 de fevereiro de 2021; **Aceito/*Approved*:** 06 de abril de 2021

Resumo

Introdução: o rápido crescimento no volume de dados coletados nos últimos anos está tornando o processo de análise e extração de conhecimento cada vez mais complexo. Organizações têm tido dificuldades em agregar essas grandes quantidades de dados em análises úteis para dar suporte a suas decisões. **Objetivo:** avaliar a implantação de ferramentas de *data mining visual* em uma *startup* de seguros (*insurtech*) para *smartphones*. **Método:** foram desenvolvidos painéis dinâmicos usando o *software* Tableau. Os dados de alimentação do sistema foram divididos dois grupos ("Medidas" e "Dimensões") para cada tema de análise escolhido. Também foi feita uma pesquisa de usabilidade junto aos usuários dos painéis. **Resultados:** a agregação de diversas subtelas e informações em um mesmo painel foi importante para os usuários visualizarem novos padrões. **Conclusão:** a introdução das novas ferramentas de descoberta de conhecimento fez com que os usuários passassem a apresentar um conhecimento mais profundo sobre o tema e fazerem melhores análises sobre os padrões.

Palavras-chave: Visualização de Dados; Dashboard; Data Mining Visual.

Abstract

Introduction: *the fast growth in the volume of data collected over recent years has been making the process of knowledge extraction and analysis increasingly complex. Organizations are facing hardship in combining this large volume of data in useful analysis to support their decision-making.* ***Goal:*** *evaluate the implementation of visual data mining tools in an insurance startup (insurtech) for smartphones* ***Methods:*** *dynamic dashboards were developed using Tableau software. It divides the system's feed data into Measurements and Dimensions categories for each chosen analysis theme. It carries out a usability survey with users of the dashboards.* ***Results:*** *The aggregation of several sub-screens and information in the same dashboard was important for users to see new patterns.* ***Conclusions:*** *: The introduction of new knowledge discovery tools has enabled users to come to a deeper understanding of the topic and to make a better analysis of patterns.*

Keywords: *Data Visualization; Dashboard; Visual Data Mining.*

INTRODUÇÃO

O avanço tecnológico trouxe diversas ferramentas de captura de dados as quais permitem a coleta de um grande volume de dados a um custo relativamente baixo (Bramer, 2007; Matsunaga, Brancher, & Busto, 2014). Entretanto, muitas vezes as informações são perdidas por não se conseguir extrair conhecimentos úteis desses dados (Bramer, 2007; Keim & Ward, 2002). Neste sentido, a área de *data mining*, utilizando uma combinação de conhecimentos nas áreas de estatística, banco de dados e inteligência artificial, permite analisar grandes volumes de dados e extrair conhecimentos relevantes sobre um fenômeno (Bramer, 2007; Carvalho & Dallagassa, 2014). Ainda, o avanço tecnológico também propiciou o surgimento dos telefones móveis. Estes são essenciais no cotidiano das pessoas, seja para entretenimento, comunicação, produtividade e ferramentas de trabalho (Coutinho, 2014). A chegada no mercado destes aparelhos propiciou a criação de *startups* na área de seguro para celulares, as quais são conhecidas como *Insurtechs* (Braun & Schreiber, 2017).

Como em toda empresa nascente ainda na fase de experimentação, o sucesso destas *startups* só é possível se decisões forem tomadas com riscos calculados e, para que tal ocorra, informação e conhecimento são essenciais. Neste contexto, o objetivo deste artigo é realizar um estudo de caso sobre a implantação e análise de ferramentas utilizando técnicas de *data mining visual*. A intenção é unir a capacidade computacional com os conhecimentos humanos específicos sobre o negócio para otimizar o processo de entender padrões, tendências e comportamentos a partir de uma base de dados. O estudo de caso foi em uma *Insurtech* - *startups* na área de seguro para celulares.

Na Seção 2, são apresentados os conceitos aplicados neste artigo nas áreas de *data mining* e *data mining visual* e os estudos correlatos encontrados na literatura. A metodologia aplicada durante o desenvolvimento do artigo é descrita na seção 3, enquanto os resultados obtidos e as discussões são alocadas na seção 4. Ao final, na seção 6, são elaboradas a conclusão e as sugestões para trabalhos futuros.

FUNDAMENTAÇÃO TEÓRICA

Na área de *data mining* estudam-se formas mais eficientes de analisar grandes volumes de dados e extrair informações relevantes sobre um fenômeno. Assim, utilizando uma combinação de conhecimentos específicos, estatística, banco de dados e inteligência artificial, é possível aproveitar o potencial computacional e transformar dados em informações e, finalmente, em conhecimento (Bramer, 2007). As aplicações mais comuns do *data mining*, segundo Bramer (2007) e Keim e Ward (2002), são: geração de regras de associação, classificação e *clustering*. Geração de regras de associação é uma técnica aplicada quando se deseja entender padrões e tendências relevantes na base de dados, sem necessariamente exigir que existam relações de causalidade. Classificação e *Clustering* são técnicas aplicadas para agrupar um volume geralmente grande de informações em conjuntos homogêneos que foram agrupados por terem mesmo padrão de características e/ou comportamento (Niggemann, 2001).

A principal diferença entre os dois métodos está na forma com que eles são alimentados. Métodos de classificação recebem, além da base de dados, categorias predefinidas e suas características para que o algoritmo aloque os novos dados dentro de uma das categorias existentes. Por sua vez, no método de *clustering*, o algoritmo recebe apenas a base de dados e gera as categorias baseadas em padrões encontrados (Keim & Ward, 2002).

Data mining visual é uma vertente do *data mining* com enfoque em otimizar a cooperação entre sistemas computacionais e usuários. Apresentar as informações de forma a ressaltar relacionamentos ocultos nos dados e permitir que o usuário interaja com a ferramenta promove o uso da capacidade humana de detectar padrões em conjunto com um domínio de contexto e regras de negócio para extrair conhecimentos úteis, de uma forma mais eficiente (Ankerst, Ester, & Kriegel, 2000; Keim & Ward, 2002; O'Halloran, Tan, Pham, Bateman, & Vande Moere, 2018).

O trabalho em conjunto entre pessoas e computadores mitiga os problemas ocasionados ao optar-se por análises totalmente automáticas ou manuais. Algoritmos computacionais podem sofrer de falta de conhecimento de contexto e relacionamento entre conceitos, enquanto pessoas são propensas a criar padrões distorcidos como, por exemplo, vieses de confirmação. Além disso, Keim e Ward (2002) apontam que métodos visuais de descoberta de conhecimento conseguem reduzir dificuldades encontradas na análise de dados com alto nível de dispersão, exigem menos expertise do usuário em relação à estatística e algoritmos matemáticos além de resultar em ferramentas mais flexíveis e ajustáveis, graças à interação do usuário nas etapas do processo de descoberta do conhecimento. Todavia, deve-se limitar a quantidade de possibilidades de visualização e interações para o usuário, evitando que o conteúdo e manuseio da ferramenta fique desnecessariamente confuso. Uma quantidade excessiva de opções e dimensões de dados pode acarretar na redução da velocidade com que o usuário consegue tomar decisões e escolher as ações corretas (Ankerst, Keim, & Kriegel, 1996; Badjio & Poulet, 2005; Rossi, 2006). Esse fenômeno é descrito por Mitchell e Utkus (2004) como *choice overload phenomenon* (sobrecarga por excesso de escolhas).

Para reduzir estes problemas, a pesquisa em Visualização da informação (InfoVis) estuda as melhores técnicas para representação de dados e informações de forma a aumentar a compreensão destes quando se trabalha com um grande volume de dados e ainda mantê-los de forma resumida e dinâmica (Silva, Franco, Dallagranna, & Cestari, 2021). Boa parte de dados que são manipulados em organizações poderiam ser mais facilmente analisados se fossem apresentados considerando também alguns pontos básicos da visão, tais como: forma, cor, posição e movimento (Alves, 2015). Cada tipo de dado/informação demanda uma representação visual apropriada e nem sempre a seleção desta representação visual é uma tarefa fácil, sendo que uma escolha inadequada pode comprometer o processo de análise. Portanto, alguns estudos se concentram em técnicas para melhorar a visualização das informações ver[(Alves, 2015; Caetano, Ribeiro, Paula, & Mattedi, 2016)].

Shneiderman (1996) propõe que a exploração visual siga três passos para que o processo de navegação seja mais fluido: (1) visão geral inicial, (2) filtros e seleção; (3) detalhes a disposição. A visão geral situa o usuário no contexto, criando uma referência e facilitando que sejam feitas comparações futuras; os filtros e seleções são a forma com que o usuário interage com a visualização dando controle e flexibilidade à plataforma; por último, os detalhes à disposição permitem que, sempre que necessário, seja possível verificar informações específicas.

Posteriormente, Keim e Ward (2002) classificaram técnicas de visualização baseados em tipos de dados, visualizações e interações. A combinação desses três atributos pode ser feita de forma ortogonal, i.e, não há restrição entre a combinação dos atributos. Tipos de dados representam a forma como os dados sobre o fenômeno a ser estudado foram coletados e como se relacionam, podendo ser formatos de texto, imagens, numéricos, entre outros. Neste artigo, são usados dados multidimensionais nos formatos numérico e texto. Tipos de visualização são as formas escolhidas para apresentar os dados e devem ser escolhidos de forma a adequar com o estilo de análise desejada. Por fim, Tipos de interação definem a forma como o usuário poderá interagir com a visualização para desenvolver a análise desejada. Neste estudo, são usadas as técnicas de *zoom, linking* e filtros dinâmicos. *Zoom* é uma técnica de interação comumente utilizada em mapas e visualizações usando dados de geo-localização e *linking* é um artifício que permite que diferentes entidades da visualização se relacionarem, permitindo que um fenômeno seja visto por diversas óticas simultaneamente. Filtros dinâmicos são filtros que podem ser aplicados, em tempo real, sobre as visualizações de forma a limitar os dados a serem apresentados.

Dentro do tema de visualização, duas das formas mais comuns de se representar dados são através de gráficos ou tabelas. Cada uma delas tem um propósito específico e tem melhor desempenho em determinadas situações. Tabelas são mais apropriadas para exibição de dados descritivos e gráficos para dados comportamentais, quando se deseja enfatizar o comportamento em relação a mais de uma variável como vendas ao longo de uma série de temporal (DeSanctis & Jarvenpaa, 1985; Vessey, 1991). DeSanctis e Jarvenpaa (1985) também indicam que a representação tabular é mais facilmente assimilada no primeiro contato; em contrapartida, os gráficos se tornam mais eficientes conforme o usuário se acostuma com a ferramenta, mesmo que as diferenças de performance sejam pouco significativas. Em sua pesquisa, os autores concluem que, apesar das duas formas de representação de dados levarem a resultados similares, os participantes retiveram uma quantidade maior de informação ao trabalhar com gráficos, indicando que assimilaram melhor o conteúdo e contexto.

O crescimento acelerado no volume de informações geradas e uma necessidade do mercado em tomadas de decisão mais rápidas e precisas criaram a necessidade de se estudar formas mais eficientes de gerar e demonstrar conhecimento. Buscando uma forma de avaliar, de forma consistente, o desempenho de ferramentas de *data mining visual*, Marghescu, Rajanen, e Back (2004) aplicaram uma pesquisa sobre a usabilidade de *softwares* de *Self-Organazing Maps* (SOM) com 26 estudantes, em sua maioria, de um curso de sistemas de informação. Uma das observações mais relevantes foram as notas relativamente baixas nas categorias de facilidade de uso, velocidade e facilidade de aprendizado da ferramenta; apesar de os participantes, em sua maioria, terem mais afinidade com tecnologia do que a população média. O que indica que algumas ferramentas ainda são percebidas como complexas para a grande maioria dos possíveis usuários, dificultando sua difusão no mercado. Em 2000, Ankerst, Ester e Kriegel realizaram um estudo comparativo entre a eficiência de técnicas de *data mining visual*, focados na classificação de dados, com níveis variados de participação do usuário no processo de supervisão. Segundo os autores, os resultados tanto da classificação manual quanto da classificação automática (sem participação do usuário) não foram satisfatórios, implicando que a cooperação entre ambas as partes é essencial. Também é enfatizado que a participação no processo de classificação, além de auxiliar a obter melhores resultados, cria uma melhor absorção das descobertas e uma maior confiança no algoritmo por parte dos usuários. Muynarsk e Miranda (2017) desenvolveram um estudo de caso sobre a implantação de ferramentas de *Business Intelligence* (BI) em uma *startup* do setor agroflorestal. Apesar de não terem acompanhado a diferença no desempenho na organização a longo prazo, a curto prazo foram notadas melhorias significativas na velocidade de geração de relatórios. Uma das observações mais notáveis foi a dificuldade encontrada durante a implementação da ferramenta, pelo fato de os usuários não serem familiarizados com esse tipo de *software*. A solução encontrada foi, além de ajustar as interfaces, criar um sistema de motivação vinculado à taxa de comissão dos colaboradores. Colocando a qualidade de informações alimentadas como um dos resultados pelos quais seriam avaliados.

MÉTODO

A *startup* estudada está inserida no setor de seguros para *smartphones* cujo principal produto é a venda de planos de proteção (PP) com cobertura a danos estéticos ou funcionais nos quais são cobradas uma taxa mensal e uma taxa de franquia, no caso de ativação (sinistro). O estudo focou nas áreas de reincidência operacional e o acompanhamento de um novo produto chamado trade. São consideradas reincidências os casos em que o cliente, após ser indenizado, sinaliza que algum defeito persiste. Nesse caso, é aberto um novo sinistro e o atendimento segue sem nenhuma cobrança adicional ao cliente. O trade é um produto que consiste em dois estágios, *trade-in* e *trade-out*. *Trade-in* é o processo no qual são feitas parcerias com varejos de diversas regiões do país com a finalidade de oferecer a possibilidade de o cliente trocar seu aparelho atual por desconto na compra de um novo aparelho. A venda para outras empresas de aparelhos comprados, reformados ou não, é chamada de *trade-out*.

Para suprir a necessidade de informações da empresa, a proposta do projeto foi desenvolver e implementar painéis dinâmicos que permitissem aos gestores ter uma visão completa dos temas levantados (reincidência e trade), com a capacidade de *drilldown* (visualização dos dados em maior granularidade) e aplicação de técnicas de *data mining* visual a fim de facilitar o uso da ferramenta como uma plataforma de geração de descobertas de relacionamentos e comportamentos impactantes ao negócio (*insights*). Foram considerados, então, o desenvolvimento de três painéis a fim de que fosse possível: (a) mensurar o impacto financeiro causado pela taxa de reincidência e entender quais os fatores mais influentes sobre seu comportamento; (b) acompanhar a efetividade das melhorias no processo de trade através da margem de contribuição percentual (MC) do produto e sua distância da meta; (c) entender a influência de diversos fatores inerentes aos *smartphones* (memória RAM, faixa de preço etc.) no comportamento de decaimento de preços para aparelhos usados; e (d) preparação do banco de dados para reduzir duplicações, entradas de dados incorretos, inconsistências e relacionar as entidades de forma correta.

Dentro das ferramentas disponíveis para criação do painel, o Tableau (SalesForce Inc., 2020) foi o *software* escolhido por facilitar a criação de exibições dinâmicas, se conectar com múltiplos formatos de bancos de dados e ter versões para diferentes sistemas operacionais (MacOS e Windows). Neste estudo, especificamente, o serviço Tableau Online estava disponível, permitindo a visualização e edição de painéis através do navegador, além de possibilitar o agendamento de atualização de bases de dados.

O próximo passo foi levantar os dados necessários para a análise satisfatória dos temas. Os dados foram divididos em duas categorias: Medidas e Dimensões. Medidas são dados quantitativos que representam a intensidade da variável a ser avaliada enquanto que as Dimensões são características qualitativas que permitem agrupar e classificar as medidas. Após ter realinhado o conceito de reincidência com os processos atuais da empresa, alguns indicadores foram selecionados, através de reuniões com diretores e gerentes da empresa, e agrupados entre Medidas e Dimensões (Tabela 1).

Medidas/Dimensões		Descrição
Medidas	Volume de reincidentes	Quantidade de atendimentos provenientes de reincidência.
	Custo logístico	Custo relativo ao serviço prestado pelos correios e ou transportadoras.
	Custo de compra	Custo relativo à compra de um aparelho, usado para atendimento do sinistro.
	Custo de reparo	Custo relativo à mão de obra e peças usadas no atendimento do sinistro.
	Custo total	Soma dos custos logísticos, de compra e reparo.
	Custo médio unitário	Custo total dividido pelo volume de reincidentes.
	Taxa de reincidência	Volume de reincidentes atendidos dividido pelo número total de atendimentos realizados.
Dimensões	Data do atendimento	Data na qual os aparelhos são separados para atendimento do sinistro (o nome desse campo no banco de dados é *swapped_at*).
	Produto	Tipo de produto que gerou o sinistro reincidente.
	Seguradora	Seguradora responsável pelo plano relativo ao sinistro.
	Parceiro	Parceiro que vendeu o plano relativo ao sinistro.
	Loja	Loja que vendeu o plano relativo ao sinistro.
	Marca	Marca do aparelho.
	Modelo	Modelo do aparelho.
	Tipo de indenização	Indica se o aparelho utilizado para o atendimento do sinistro foi um novo, *refurb* ou reparado.

Tabela 1. Medidas e dimensões relativas à reincidência.
Fonte: Os autores (2021).

Para o tema *trade-out*, indicadores também foram selecionados e agrupados entre Medidas e Dimensões (Tabela 2). A escolha dos dados relevantes para esta análise foi feita a partir do acompanhamento das análises e preocupações reportadas pela equipe comercial.

Medidas/Dimensões		Descrição
Medidas	Quantidade de aparelhos	Contagem de aparelhos transacionados.
	Custo total de compra	Soma dos custos de aquisição dos aparelhos.
	Custo médio de compra	Custo total de compra dividido pela quantidade de aparelhos.
	Valor total de venda	Soma dos valores a que os aparelhos foram vendidos.
	Valor médio de venda	Valor total de venda dividido pela quantidade de aparelhos.
	Margem bruta total	Valor total de venda subtraído do custo total de compra.
	Margem bruta unitária	Margem bruta dividida pela quantidade de aparelhos.
	Margem líquida total	Valor total de vendas subtraído de impostos, custo de mão de obra, custos logísticos e custo total de compra.
	Margem líquida unitária	Margem líquida total dividida pela quantidade de aparelhos.
	Markup realizado	Valor total de venda dividido pelo custo total de compra.
	Markup alvo	Valor total de venda dividido pelo custo máximo de compra (calculado para atingir MC 0%).
	Distância do markup alvo	Diferença entre o markup realizado e o markup alvo.
Dimensões	Data de compra	Data da efetuação da compra do aparelho pelo parceiro.
	Data de venda	Data de registro da venda para B2B.
	Qualidade do parceiro	Qualidade da avaliação do aparelho feita pelo parceiro.
	Qualidade da empresa	Qualidade da avaliação do aparelho feita pela empresa.
	Modelo	Modelo do aparelho.

Tabela 2. Medidas e dimensões relativas ao resultado de *trade-out*.
Fonte: Os autores (2021).

Os dados referentes ao comportamento de mercado gerados pelo trade foram definidos junto ao gestor comercial e também classificados como Medidas e Dimensões (Tabela 3). A análise destes dados é importante para entender a elasticidade do produto, ou seja, a relação ideal entre variação de preços e volume de transações (Andreyeva, Long, & Brownell, 2010).

Medidas/Dimensões		Descrição
Medidas	Volume	Quantidade de aparelhos vendidos.
	Representatividade	Soma dos custos de aquisição dos aparelhos.
	Valor de venda	Quantidade de aparelhos vendidos em uma categoria específica em relação ao montante total.
Dimensões	Qualidade de venda	Qualidade da avaliação do aparelho feita pela empresa *(perfect, good, broken, unusable)*.
	Marca	Marca do aparelho.
	Modelo	Modelo do aparelho.
	Faixa de preço	Faixa de preço (em incrementos de R$500 ou R$1000) em que o preço de lançamento do modelo se encaixa.
	Família	Categorias (ou linhas) criadas pelas fabricantes onde os modelos atendem um determinado conjunto de características a fim atender melhor as necessidades de públicos específicos.
	Subfamília	Subdivisões de famílias para personalizar ainda mais a especificação do modelo com a necessidade dos usuários.

Tabela 3. Medidas e Dimensões relativas ao comportamento de mercado.
Fonte: Os autores (2021).

Com relação à criação dos painéis, a técnica de extração de conhecimento escolhida para o tema reincidência foi a geração de regras de associação em conjunto com a visualização dos dados em forma de tabelas e um gráfico composto de barras e linha. As técnicas de interação escolhidas foram *linking* e filtros dinâmicos, pois possibilitam tanto o acompanhamento de tendências de variação de custo e volume quanto o entendimento

	Perguntas	Critérios Relacionados	Simplificação
Bloco 1	1. Qual o cargo que ocupa?	Dados sobre os participantes.	-
	2. Em que área trabalha?		
	3. Qual seu objetivo em usar *dashboard*?		
	4. Qual o seu conhecimento em BI?	Conhecimento prévio sobre ferramentas similares.	Conhecimento
	5. Qual a frequência de uso de ferramentas de BI?	Frequência de uso de ferramentas similares.	Frequência
Bloco 2	6. As informações apresentadas no dashboard são claras?	Clareza dos dados apresentados.	Clareza
	7. O dashboard permitiu que você fizesse suas análises de forma rápida?	Facilidade para completar tarefas.	Velocidade
	8. O que achou dos filtros e funcionalidades aplicados no dashboard?	Facilidade de uso e eficácia das técnicas de interação aplicadas.	Filtros
	9. O dashboard permitiu que você criasse a visualização do jeito que queria?	Eficácia da ferramenta em auxiliar o usuário a cumprir seu objetivo.	Objetivo
	10. Quando cometeu algum erro, foi fácil voltar atrás?	Facilidade oferecida pela ferramenta para recuperação de erro.	Recuperabilidade
	11. O que achou do visual?	Satisfação com a estética do painel.	Visual
Bloco 3	12. No geral, o que achou do novo dashboard?	Medida de satisfação geral com a ferramenta.	Satisfação
	13. Pretende continuar usando?	Combinação entre satisfação geral e atendimento do objetivo do usuário.	Coerência
	14. Comentários adicionais.	Resposta aberta.	

Tabela 4. Relacionamento entre perguntas, critérios de avaliação e formato simplificado.
Fonte: Os autores (2021).

de padrões entre segmentos mais impactados pela reincidência. No tema de *trade-out*, dado o baixo nível de conhecimento técnico e familiaridade com interpretação de dados complexos por parte dos usuários, optou-se por desenvolver uma visualização simplificada (gráfico de barra e linha), focada em facilitar a identificação de tendências (regras de associação), mantendo poucas informações na exibição e incluindo opções para alterações de variáveis e filtros. Essas adaptações facilitam a análise de tendências sem reduzir a flexibilidade de observar os comportamentos em segmentos específicos. Por fim, no painel de comportamento de mercado, as técnicas de extração de conhecimento aplicadas foram a classificação e geração de regras de associação de forma a auxiliar na identificação de padrões de depreciação segundo características em comum entre grupos. Os dados foram agregados de forma personalizável em um gráfico de *scaterplot*, possibilitando o teste de hipóteses de sensibilidade sob óticas variadas. Para interagir com o painel, foram aplicadas técnicas de filtros dinâmicos e seletores. Os seletores permitem que o usuário escolha quais Medidas serão representadas por cor, tamanho e forma no *scaterplot*.

Por último, foi aplicada uma pesquisa com os usuários dos painéis (gestores e analistas das áreas afetadas diretamente pela reincidência e *trade*) a fim de entender o relacionamento entre facilidade de uso e experiência com análise de dados, coletar informações sobre melhorias e verificar a usabilidade das novas ferramentas de análise. Neste contexto, Marghescu et al. (2004) propõem quatro tópicos principais a serem metrificados em uma análise de desempenho de usabilidade de ferramentas visuais de mineração de dados: (a) Qualidade de uso: mede a satisfação geral do usuário ao utilizar a ferramenta; (b) Qualidade da visualização: verifica a clareza com que a ferramenta exibe as informações, usando meios para simplificar o entendimento mantendo o foco no conteúdo e não na visualização em si; (c) Qualidade da interação: mensura a velocidade de aprendizado, facilidade de uso e rapidez para completar tarefas; e (d) Qualidade da informação: analisa a veracidade e apresentação dos dados relevantes e satisfação com o período de atualização das informações apresentadas. Por fim, duas perguntas foram inclusas de forma a avaliar também a eficácia dos painéis em auxiliar o usuário a cumprir suas tarefas. O questionário, aplicado de forma virtual, é composto por catorze perguntas divididas em três blocos (Tabela 4). As questões 4 até 13 são fechadas, com notas variando de 1 a 5, e as questões 1 a 3 e 14 são questões abertas. A segunda coluna da Tabela 4 é uma breve descrição do critério a ser avaliado pela pergunta e a terceira coluna apresenta uma forma de simplificação pelo qual o atributo será chamado nas visualizações dos dados coletados.

RESULTADOS

Painel de impacto de reincidência

Na Figura 1 é apresentado o painel inicial do painel de impacto de reincidência, em que as dimensões são usadas para filtros e níveis de detalhe para as medidas levantadas. Foram criadas as seguintes hierarquias na base dados para possibilitar o *drilldown*: produto > seguradora > parceiro > loja (referenciada adiante como hierarquia de produto) e marca > modelo (referenciada adiante como hierarquia de modelo). O Tableau faz a hierarquia de datas de forma automática (ano > trimestre > mês > semana > dia). A estrutura do painel apresenta cinco partes, cada uma focando na visualização dos mesmos dados sob diferentes perspectivas. Essa disposição dos dados permite que sejam feitas análises de fatores distintos sobre a mesma base de dados filtrada, auxiliando no entendimento das relações entre comportamentos e resultados.

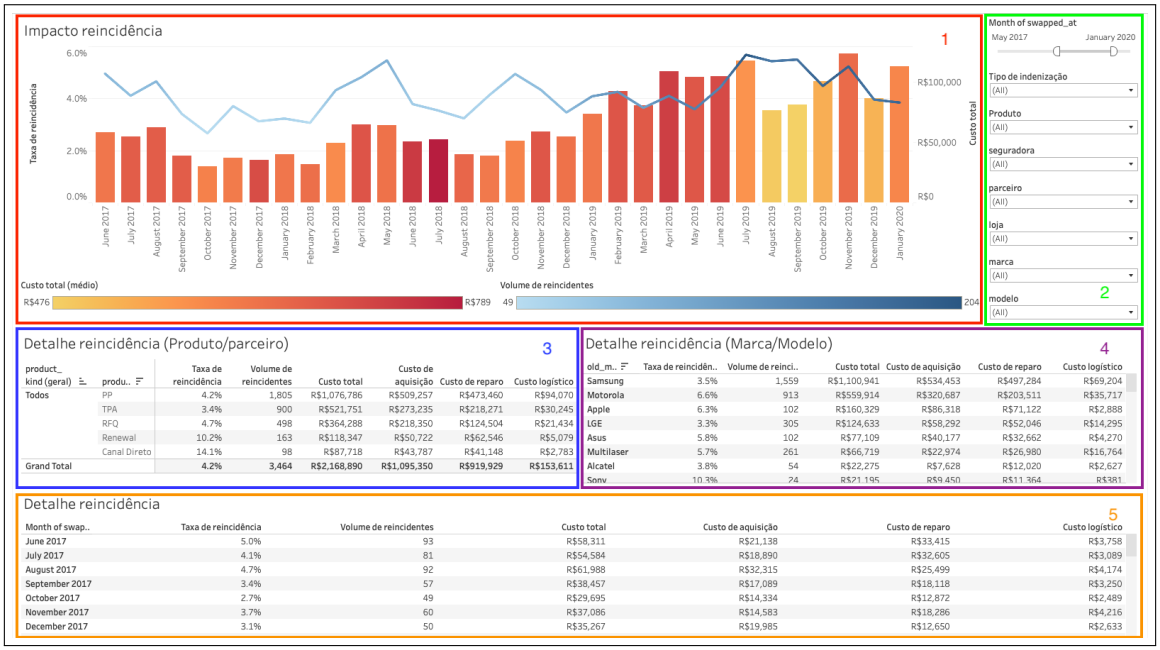

Figura 1. Estrutura da tela inicial do painel de impacto de reincidência.

A primeira parte, indicada pela borda vermelha na Figura 1, apresenta o gráfico de resultado operacional usado para acompanhar a tendência da reincidência. Consiste em um gráfico de eixo duplo no qual o eixo das abscissas é a data de atendimento dos sinistros, a altura das barras representa o custo total de reincidência no período (eixo das ordenadas à direita), a coloração das barras mostra o custo médio unitário de reincidência, a linha indica a taxa de reincidência no período (eixo das ordenadas à esquerda) e a coloração das linhas indica o volume total de reincidentes no período. Os filtros (borda verde – parte 2) permitem selecionar as dimensões levantadas para o tópico de reincidência para a base de dados das demais partes do painel. A tabela de comportamento por cliente (borda azul – parte 3) exibe as medidas no nível de detalhe da hierarquia de produto. É usada para identificar comportamentos distintos dentro de certos grupos de clientes. A tabela de comportamento por aparelho (borda roxa – parte 4) mostra as medidas no nível de detalhe da hierarquia de modelo, sendo usada para identificar comportamentos distintos dentro de certos grupos de aparelhos. A tabela de comportamento temporal (borda laranja – parte 5) exibe as medidas no nível de detalhe da hierarquia de datas, sendo usada para identificar tendências de evolução ao longo do período analisado.

Todas as tabelas apresentam as medidas de taxa e volume de reincidência, custos total, de aquisição, de reparo e logístico sendo a diferença entre elas apenas o nível de detalhe. A última linha consiste no resultado total agregado dessas medidas. Como as dimensões foram organizadas dentro de uma hierarquia, o painel permite que os níveis de detalhe sejam abertos ou colapsados dando maior controle sobre a granularidade da análise. A Figura 2 apresenta um exemplo de dados detalhados. O sinal de "+" (círculo vermelho) indica a possibilidade de interação para abrir os dados em um nível mais granular de detalhe.

Detalhe reincidência (Produto/parceiro)

product_ki..	prod	Taxa de reincid..	Volume de rein..	Custo total	Custo de aquis..	Custo de reparo	Custo logístico
Todos	PP	5.6%	420	R$241,612	R$138,392	R$101,194	R$2,026
	RFQ	4.9%	94	R$76,576	R$43,378	R$32,878	R$321
	Renewal	9.4%	61	R$50,782	R$22,353	R$28,281	R$148
	TPA	56.3%	18	R$9,855	R$375	R$9,470	R$10
	RFQ 2.0	2.8%	7	R$5,602	R$5,602	R$0	R$0
	Canal Direto	15.8%	3	R$3,849	R$3,749	R$100	R$0
Grand Total		5.8%	603	R$388,275	R$213,848	R$171,924	R$2,503

Figura 2. Detalhe do painel mostrando a tabela de comportamento por cliente aplicando *drilldown* para nível de produto.

Também foi inserida a técnica de *linking* para a seleção de dados no painel com o objetivo de facilitar as transições e prover uma experiência mais fluída de análise. Essa funcionalidade é aplicada com a seleção de qualquer componente do painel e, junto com o *drilldown*, permite que o usuário navegue pelos dados entendendo o comportamento de reincidência sob diferentes aspectos, níveis de detalhe e situações específicas.

Painel de resultado de *trade-out*

O painel principal do painel de resultado de *trade-out* (Figura 3) consiste em um gráfico com eixo duplo de exibição simples e uma série de filtros desenvolvidos para possibilitar o acompanhamento do resultado de trade-out nos formatos mais comuns para os usuários (área comercial), sem confundi-los com muitas informações e funcionalidades.

Figura 3. Estrutura do painel de resultado de *trade-out*.

O painel foi desenvolvido de forma que fosse possível a comparação de duas medidas dentro dos mesmos parâmetros. Os controles dinâmicos do painel foram divididos em quatro categorias. Os filtros de data (borda vermelha na Figura 3) foram implementados com barras deslizantes para facilitar a seleção da abrangência desejada. Para o item Quebras (borda azul), existem dois seletores: "Data"indica qual data a ser usada no eixo horizontal (as opções são data de venda ou data de compra) e "Detalhe"indica o tipo de divisão realizada (as opções são: qualidade de compra, qualidade de venda, ativos na tabela e aparelhos mais vendidos). Com relação aos Filtros gerais (borda verde), estes são listas de seleção múltipla para personalizar os dados a serem exibidos. Por fim, os Valores de exibição (borda laranja) são as opções de escolha de quais medidas a serem exibidas no gráfico, em que o valor principal representa os valores no gráfico de colunas e o valor secundário o gráfico de linha. As opções de escolha são todas as medidas listadas na Tabela 2.

As formas tradicionais de análise são por safra, em que se analisa o comportamento de grupos criados no mesmo período e por competência, medindo-se os resultados obtidos no período analisado independente de sua data de criação. No caso de *trade-out*, as análises de safra são feitas em torno da data de compra dos aparelhos e as análises de competência relativas à data de venda. Na Figura 3, é feita uma análise de safra comparando o *markup* realizado com o *markup* alvo (valor total de venda dividido pelo custo máximo de compra) ao longo de um

período de seis meses. A tela permite ao usuário perceber de forma imediata, por exemplo, que o distanciamento entre as duas medidas foi reduzido gradativamente e o produto atingiu o patamar de lucro no mês de fevereiro.

Painel de comportamento de mercado

Para fazer análises de comportamento, foi necessário desenvolver uma visualização capaz de comportar diversos tipos de dados, de forma dinâmica e apresentação de fácil entendimento. Os conceitos de muitas dimensões de informação e apresentação amigável ao usuário são, muitas vezes, conflitantes. A solução encontrada foi a inclusão de formas de organizar os dados de forma segregada, dividindo a apresentação em diversos subpainéis.

O painel principal do painel de comportamento de mercado é apresentado na Figura 4. O gráfico mostra dados em diversos formatos (*scaterplot*) em que o eixo horizontal representa o tempo decorrido (em meses) entre o lançamento do modelo do *smartphone* e a data de competência da venda e o eixo vertical ilustra a relação entre o preço realizado na venda e o preço inicial de lançamento do modelo. A divisão horizontal relativa à qualidade de venda (aparelhos bons - *Good* - e aparelhos quebrados - *Broken*) é mantida como padrão por ser uma variável recorrente, independentemente da análise. O tamanho dos formatos representa ou o volume de aparelhos vendidos ou sua representatividade em relação ao total de aparelhos vendidos no período. Ainda, os grupos 1 (borda vermelha) e 2 (borda azul) são constituídos por barras deslizantes usadas como filtro de data de venda e número de marcas com maior volume de vendas; seletores de múltipla escolha para filtrar faixas de preço e marca; e seletores relativos à formatação da visualização. A formatação da visualização pode ser personalizada nos quesitos forma, cor, tamanho e detalhe; além de criar divisões horizontais no painel. As opções disponíveis seguem as dimensões listadas na Tabela 3. No grupo 3 (borda verde), são exibidas as legendas relativas às dimensões escolhidas para representar as cores e formas do gráfico.

Figura 4. Exemplo de visualização personalizada no painel de comportamento de mercado.

A Figura 4 mostra uma visualização personalizada do painel com os seguintes parâmetros: (a) Top: 3 marcas de maior volume de vendas no trade-out; (b) Faixa de preço: intervalos de R$1,000 em relação ao preço de lançamento dos modelos; (c) Filtro faixas de preço: modelos com preços de lançamento entre R$ 1,000 e R$4,000; (d) Cor: marca dos aparelhos; (e) Forma: memória; e (f) Tamanho: representatividade. A flexibilidade na personalização dos gráficos permite mais riqueza na análise ao usuário. No exemplo da Figura 4, o usuário tem a possibilidade de perceber um relacionamento de natureza inversa entre a quantidade de memória, preço de lançamento e marca, relativos ao tempo de vida. Ao analisar esse relacionamento surgem duas hipóteses para a explicação desse comportamento: (a) as pessoas dispostas a pagar um valor mais alto por um *smartphone* são mais propensas a trocar de aparelho com mais frequência; e (b) são dois comportamentos distintos que culminam nesse resultado: as fabricantes estarem lançando modelos com preços iniciais mais caros e uma tendência do mercado de fazer a troca mais frequente de modelo.

Resultados da pesquisa de usabilidade dos painéis

A pesquisa sobre a usabilidade foi aplicada a três participantes para cada painel (impacto de reincidência, comportamento de mercado e resultado de trade-out), totalizando nove colaboradores. Os avaliadores foram

selecionados por estarem em contato direto e serem responsáveis pelos temas. Segundo Zuk, Schlesier, Neumann, Hancock, e Carpendale (2006), não existem estudos conclusivos sobre a quantidade mínima de avaliadores necessária para realizar uma análise satisfatória de usabilidade em ferramentas de visualização de informação. Entretanto, os autores também sugerem que cinco participantes podem ser considerados o número ideal e que três participantes seriam o suficiente para encontrar a maioria dos problemas de usabilidade.

Sobre o nível de conhecimento em ferramentas de BI (questão 4) na pesquisa realizada, tem-se que 78% dos participantes têm alto nível de conhecimento e os demais participantes apresentaram médio (11%) ou baixo (11%) nível de conhecimento.

A Figura 5 apresenta as notas médias obtidas para cada painel. Observa-se que o painel de comportamento de mercado obteve as avaliações mais baixas, em especial nos quesitos visual, recuperabilidade e clareza (as notas foram 2,3; 2,7 e 3,0 respectivamente). Ao entrevistar os usuários, foi concluído que a visualização em scaterplot em conjunto com a grande variedade de filtros e divisões oferecidas tornou o uso do painel muito confuso. Os usuários estavam mais concentrados em entender a ferramenta do que em analisar as informações, resultando em análises superficiais. Após algumas semanas, com os usuários acostumados à ferramenta, foram realizadas novas entrevistas e os participantes reportaram um nível de satisfação maior em relação ao painel. Pode-se concluir que técnicas mais complexas de visualização e interação, apesar de oferecerem novas possibilidades de análise, aumentam o tempo de aprendizado necessário para dominar a ferramenta. Os painéis de impacto de reincidência e resultado de trade-out obtiveram avaliações de usabilidade positivas e similares, apesar da diferença entre o nível de conhecimento dos usuários dos dois painéis. A partir dessa observação, a hipótese é que a representação das informações de forma mais simples (gráficos de barra, linha e tabelas) proporcionou aos usuários, mesmo com diferentes níveis de conhecimento, uma experiência mais rápida de aprendizado graças ao contato diário com esses formatos de visualização.

No geral, as novas ferramentas foram bem aceitas. A avaliação média entre os painéis foi de 4,3, sendo que 78% dos participantes respondeu positivamente quando questionados sobre continuar usando as ferramentas no futuro. Após entrevistas e observações sobre o uso dos painéis, foi comprovado que as maiores dificuldades encontradas foram em relação à familiaridade com a ferramenta e uma apresentação confusa do gráfico de *scaterplot*.

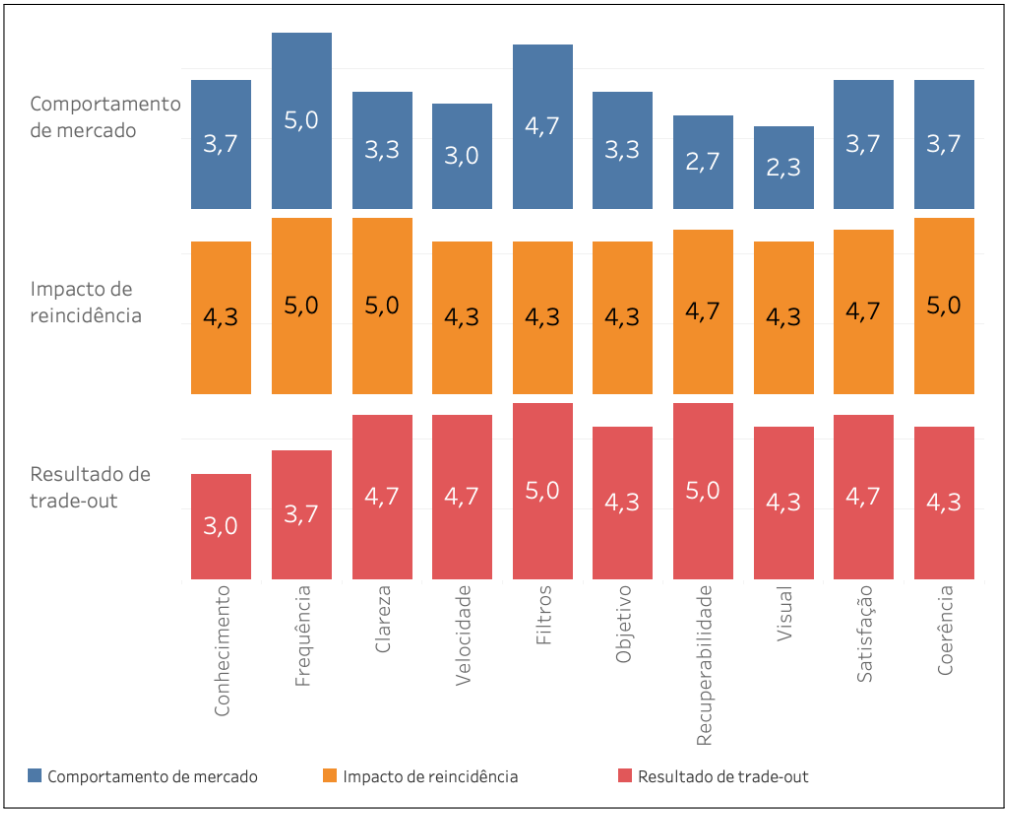

Figura 5. Resultados da pesquisa de usabilidade - Avaliação por painel.

Apesar de as dificuldades, os usuários reportaram executar tarefas de forma mais ágil após algumas semanas e que o uso de "ferramentas mais inteligentes" tornou o trabalho mais divertido, incentivou a criação de novos indicadores e motivou análises mais profundas sobre alguns comportamentos descobertos através do uso dos painéis. Ao acompanhar reuniões após a implantação das ferramentas, constatou-se que a participação dos usuários no processo de descoberta do conhecimento também estimulou uma melhor compreensão dos indicadores da empresa, efeito dos projetos executados e do negócio.

CONCLUSÃO

O objetivo deste estudo foi buscar melhores alternativas de extração de informações e geração de conhecimento em uma *startup*, através da criação e aplicação de painéis, utilizando técnicas de *data mining* visual. Com a implantação dos painéis, o tempo e esforço necessários para realizar essas tarefas foram reduzidos de forma acentuada. Com as informações sendo atualizadas de maneira confiável, frequentes e automáticas, as equipes puderam se concentrar em gerar valor ao negócio através da execução de projetos e agilidade e assertividade na tomada de decisão. A introdução das novas ferramentas de descoberta de conhecimento fez com que os usuários passassem a apresentar um conhecimento mais profundo sobre contexto das análises e curiosidade para entender novos padrões.

Apesar de possibilitarem análises mais completas, algumas formas de visualização mais complexas, como o *scaterplot*, e a aplicação de muitas funcionalidades em um mesmo painel podem tornar o painel muito confuso, afetando o tempo de aprendizado, velocidade e a satisfação do usuário ao usar a ferramenta. A solução que se provou mais eficaz foi a inclusão de diversas visualizações, com focos distintos, em um mesmo painel, como no painel de impacto de reincidência. Ao oferecer diversas perspectivas sobre os mesmos dados, relacionando-os através de técnicas como o *linking*, o usuário consegue navegar de forma fluida e focar em detalhes específicos ou observar o impacto de mudanças em diferentes situações para descobrir novos padrões, sem perder o controle da visualização.

REFERÊNCIAS

Alves, M. C. (2015). *Visualização de informação para simplificar o entendimento de indicadores sobre avaliação da ciência e tecnologia* (Dissertação de mestrado não publicada). Universidade Federal de São Carlos, São Carlos, SP, Brasil. (Dissertação de mestrado)

Andreyeva, T., Long, M. W., & Brownell, K. D. (2010). The impact of food prices on consumption: a systematic review of research on the price elasticity of demand for food. *American Journal of Public Health*, *100*(2). doi: 10.2105/AJPH.2008.151415

Ankerst, M., Ester, M., & Kriegel, H. P. (2000). Towards an effective cooperation of the user and the computer for classification. In *Proceedings of the sixth acm sigkdd international conference on knowledge discovery and data mining* (p. 179–188). Boston, US. doi: 10.1145/347090.347124

Ankerst, M., Keim, D. A., & Kriegel, H. P. (1996). Circle segments: A technique for visually exploring large multidimensional data sets. In *Proceedings of visualization'96, hot topic session.* San Francisco, US.

Badjio, F. E., & Poulet, F. (2005). Dimension reduction for visual data mining. In *International symposium on applied stochastic models and data analysis.* Brest, França.

Bramer, M. (2007). *Principles of data mining* (v. 180). London: Springer.

Braun, A., & Schreiber, F. (2017). *The current insurtech landscape: business models and disruptive potential* (v. 62). Gallen, Suiça: Institute of Insurance Economics I. VW-HSG, University of St. Gallen.

Caetano, B. P., Ribeiro, F. C., Paula, M. M. V. d., & Mattedi, A. (2016). A proposal for visualization techniques recommendation to represent survey data. In *11th iberian conference on information systems and technologies (cisti)* (p. 1–6). Chaves, Portugal. doi: 10.1109/CISTI.2016.7521633

Carvalho, D. R., & Dallagassa, M. R. (2014). Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas. *AtoZ: novas práticas em informação e conhecimento*, *3*(2). doi: 10.5380/atoz.v3i2.41340

Coutinho, G. L. (2014). *A era dos smartphones: um estudo exploratório sobre o uso dos smartphones no brasil* (Dissertação de mestrado, Universidade de Brasília, Brasília, DF, Brasil). Recuperado de https://bdm.unb.br/bitstream/10483/9405/1/2014_GustavoLeuzingerCoutinho.pdf (Monografia de graduação)

DeSanctis, G., & Jarvenpaa, S. L. (1985). An investigation of the tables versus graphs controversy in a learning environment. In *Proceedings of the 6th international conference on information systems* (p. 134–144). Indianapolis, US.

Keim, D., & Ward, M. (2002). Visual data mining techniques. In *Hand, D. and Berthold, M. (Eds.). Intelligent Data Analysis, an Introduction* (2a. ed.). Heidelberg: Springer.

Marghescu, D., Rajanen, M., & Back, B. (2004). Evaluating the quality of use of visual data-mining tools. In *Proceedings of the 11th european conference of information technology evaluation* (p. 239–250). Amsterdam, Netherlands.

Matsunaga, F. T., Brancher, J. D., & Busto, R. M. (2014). Data mining applications and techniques: a systematic review. *Revista Eletrônica Argentina-Brasil Tecnologias da Informação e da Comunicação*, *1*(2). doi: 10.5281/zenodo.59454

Mitchell, O. S., & Utkus, S. P. (2004). *Pension design and structure: New lessons from behavioral finance.* Oxford, UK: Oxford University Press.

Muynarsk, R. G., & Miranda, E. d. S. (2017). Business intelligence no agronegócio: um estudo de caso de implementação em uma startup. *Revista iPecege*, *3*(1). doi: 10.22167/r.ipecege.2017.1.75

Niggemann, O. (2001). *Visual data mining of graph-based data* (Dissertação de mestrado não publicada). Department of Mathematics and Computer Science, University of Paderborn, Germany. (Dissertação de mestrado)

O'Halloran, K. L., Tan, S., Pham, D. S., Bateman, J., & Vande Moere, A. (2018). A digital mixed methods research design: Integrating multimodal analysis with data mining and information visualization for big data analytics. *Journal of Mixed Methods Research*, *12*(1). doi: 10.1177/1558689816651015

Rossi, F. (2006). Visual data mining and machine learning. In *Proceedings of 14th european symposium on artificial neural networks* (p. 251–264). Bruges, Belgica. Recuperado de https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2006-3.pdf

SalesForce Inc. (2020). *Tableau desktop (2020.3.7).* https://www.tableau.com/products/de.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceeding of ieee symposium on visual languages* (p. 336–346). Boulder, Colorado, US. doi: 10.1109/VL.1996.545307

Silva, G. M., Franco, D. J., Dallagranna, G. J., & Cestari, J. M. A. P. (2021). Visualização da informação aplicada em dados aberto nas unidades de saúde municipais de curitiba-pr: perfil de atendimento de enfermagem. In *Coletânea especial de engenharia de produção.* Itajubá: Ed. Kreatik.

Vessey, I. (1991). Cognitive fit: A theory based analysis of the graphs versus tables literature. *Decision Sciences*, *22*(2). doi: 10.1111/j.1540-5915.1991.tb00344.x

Zuk, T., Schlesier, L., Neumann, P., Hancock, M. S., & Carpendale, S. (2006). Heuristics for information visualization evaluation. In *Proceedings 2006 avi workshop on beyond time and errors: novel evaluation methods for information visualization* (p. 1–6). Venice, Italy. doi: 10.1145/1168149.1168162

Como citar este artigo (APA):
Matsuba, D. S. & Mattedi, A. P. (2021). Visualização de dados para extração de conhecimento: um estudo de caso. *AtoZ: novas práticas em informação e conhecimento, 10*(2), 66 – 77. Recuperado de: http://dx.doi.org/10.5380/atoz.v10i2.79184