AS QUESTÕES DA ANÁLISE DE DADOS NO CONTEXTO DA CIÊNCIA DE DADOS

DATA ANALYSIS ISSUES IN THE DATA SCIENCE CONTEXT

Geyzon Ferreira da Silva Assis,
Rogério Oliveira da Silva

RESUMO

Esse artigo tem como objetivo trazer uma visão geral e objetiva a respeito da análise de dados – Data Analysis – no contexto da Ciência de dados, especificamente no que se refere ao levantamento dos questionamentos necessários para se extrair o valor adequado e esperado dos dados que se possui. Será tratadotambém o conceito de Ciência de dados e qual sua importância no contexto atual de volumes grandes de dados.

Palavras-chave: Análise de dados, Ciência de Dados, valor, dados.

ABSTRACT

This paper aims to provide a general and objective overview about Data Analysis in context of Data Science, specifically with regard to the survey of the necessary questions to extract the adequate and expected value of the data mart. We will discuss the concept of data science and its importance in the current context of large volumes of data.

Keywords: *Data Analysis, Data Science, value, data.*

INTRODUÇÃO

A correta aplicação de métodos de análise de dados é de suma importância quando se tem a intenção de ser competitivo frente a um cenário desafiador como se tem vivido atualmente, a era da informação e globalização. Um estudo realizado pela OBS (Online Business School) demostra que num período compreendido entre 2004 e 2014 foi gerada uma quantidade de dados superior a que foi gerada em toda

a história da humanidade e saber lidar com essa quantidade abismal de dados é caminho correto, embora nem sempre fácil ou intuitivo.

Tomada de decisão estratégica é o que determina o rumo de uma empresa, e possuir meios seguros para apoio a essas decisões devemser levadas em consideração para que se tenha o melhor direcionamento possível. Alguns gestores insistem em agir por meio de sua intuição, tornando o processoextremamente subjetivo, levando a uma grande fragilidade do resultado dessas decisões. Conseguir prever cenários, definir estratégias, gerar valor ao negócio e ter apoio de dados é o grande objetivo da análise de dados, verdadeiramente um grande desafio.

A IMPORTÂNCIA DOS DADOS

Em um mundo altamente competitivo, com enorme disputa de espaço num cenário tão desafiador como tem-se vivenciado, os dados têm sido amplamente considerados como verdadeiros direcionadores da melhor tomada de decisão proporcionando maior rentabilidade ao negócio, agregando segurança nesse processo.

O ativo mais importante para uma organizaçãoé asua base de dados: os cadastros em seu banco de dados, a lista de fornecedores, as informações pertinentes ao negócio, a carteira de clientes, etc. A era da informação chegou, e ter o poder dos dados ao seu favor possibilita maiores chances do negócio evoluir.

A CIÊNCIA DE DADOS

No passado, grandes empresas contratavam equipes de estatísticos, modeladores de dados e analistas para explorar e manipular os dados referentes ao negócio de forma manual, mas com o crescimento exponencial da quantidade e diversidade de dados gerados na atualidade, seja em um contexto web cada dia mais vasto ou até mesmo num ambiente estritamente corporativo onde se lida com dados referentes ao negócio, tornouessa atividade manual uma tarefa árdua, fazendo-se necessário a utilização de ferramentas de gestão especializadas. Isso foi possibilitado pelo fato dos computadores modernos terem se tornado cada vez mais poderosos em processamento, a internet se tornou onipresente e os algoritmos desenvolvidos conseguem também se conectar as bases de dados possibilitando analises mais amplas e profundas do que era possível no passado. Nesse contexto surgiu o conceito de ciência de dados, onde tem se desenvolvido técnicas

avançadas de análise de dados e manipulação, buscando sempre agregar valor nessa grande massa de informação.

“Não se pode administrar aquilo que não se pode mensurar” é uma frase atribuída a W. Edward Deming ou Peter Drucker(Brynjolfsson e McAfee 2012)que possui uma grande verdade, pois é realmente impossível fazer uma boa gestão de um negócio caso não se possua o controle, ou ao menos, a possibilidade de se medir as informações realmente relevantes ao empreendimento. Para que haja tomada de decisão de forma precisa, é necessária ela estar embasada em dados, conceito esse muito difundido nos EUA denominada *data-driven*, que engloba empresas que aplicam essa filosofia de gestão baseada em dados, procurando minimizar possíveis equívocos gerenciaise agregando valor as decisões.

Brynjolfsson e McAfee (2012) afirma que muitas empresas que se auto denominam como *data-driven* posicionadas em altos níveis competitivos em seus respectivos mercados tem conseguido em média 5% de aumento de sua produtividade e 6% mais rentáveis que seus competidores.

A ANÁLISE DE DADOS

Como o próprio nome sugere, análise de dados consiste em um método onde osdados são colecionados e organizados de tal maneira que se possa extrair informações pertinente e suficientemente ajudadoras para se alcançar um determinado fim proposto, seja para tomada de decisão estratégica, entendimento de mercado ou até mesmo predição de cenários.Possuir os dados apenas não é o suficiente pois se assim fosse, todas corporações, institutos, órgãos governamentais assim o fariam;extrair valor dos dados que se possuinão significa efetuar uma *query* SQL, por exemplo, mas vai muito além disso, usando modelos matemáticas e estatísticos avançados.

O processo de análise de dados é complexo e difícil, e não é qualquer pessoa que irá conseguir fazer isso com perfeição. É um trabalho extremante particular, comparável a uma arte, como diz Roger D. Peng (2016) em seu livro “A arte da Ciência de dados”.

Ele faz uma comparação muito interessante da execução de análise de dados com a inspiração de um compositor musical. Seria como perguntar ao compositor como ele compôs tal canção. Existem inúmeras ferramentas que poderiam ser

utilizadas nesse processo bem como o entendimento do que seria uma boa música, qual estrutura, quantos versos ou harmonia utilizada. É como se existisse um "framework abstrato" para se compor uma boa música. Mas embora se tenha ciência de toda essa teoria musical, ela não é suficiente para se compor uma música de sucesso, é necessário algo a mais.

A criatividade é fundamental e parte chave para que uma boa música seja escrita e que as pessoas queiram ouvi-la, assim como é importante para o sucesso na análise de dados. Cientistas de dados possuem a sua disposição inúmeras ferramentas para auxiliarem nesse processo, desde árvores de classificação até *Deep Learning* (Redes Neurais profundas), e isso deve ser aplicado de forma correta, ou seja, o analista deve desenvolver uma forma para, através desses subterfúgios, extrair respostas dos dados.

DEFININDO O QUESTIONAMENTO DA ANÁLISE DE DADOS

É necessário compreender o que os dados querem dizer por meio da análise, mas para que isso seja efetivo, é fundamental saber formular as perguntas corretamente para eles. Muitos questionamentos podem ser respondidos por meio da análise de dados, mas não todos eles, por mais que haja um desejo profundo de encontrar respostas, como afirma John Tukey, conhecido como o pai da análise exploratória de dados moderna, em sua icônica frase: ***"The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data."***

Por mais que seja importante possuir os dados, é fundamental saber fazer as perguntas corretas para que seja possível extrair informação suficiente para atender a demanda questionada. Algumas questões podem ser mais fáceis que outras para serem respondidas.

Fundamental também é entender que existem basicamente 6 tipos de perguntas que podem ser realizadas como explica Jeff Leek em seu artigo "*Whatisthequestion? "*publicado na revista *Science*(2015). Determinar esse ponto permite que o resultado da pesquisa esteja correto. A seguir temos uma imagem que busca auxiliar no processo de determinação dostipos corretos de perguntas que devem ser feitas mediante a necessidade observada.

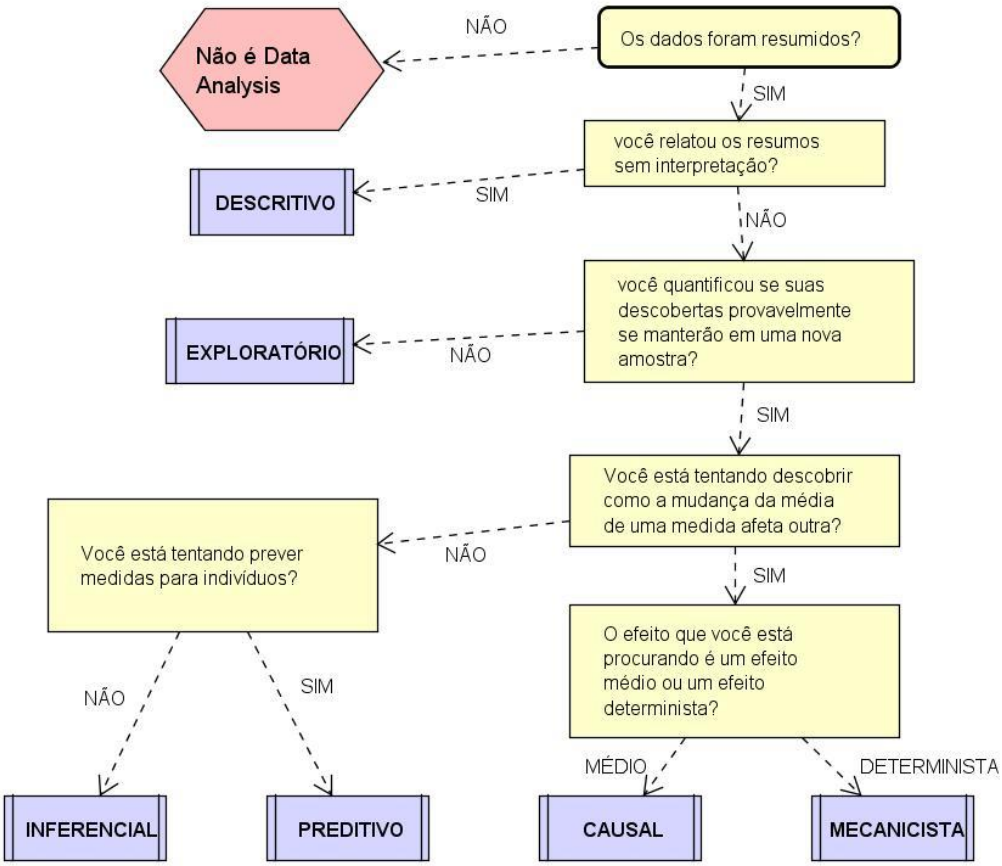

Figura 1Fluxo análise de dados - Fonte: ***The elementsofAnalyticStyle*** *(livre tradução)*

ANÁLISE DESCRITIVA

A análise descritiva de dados tem como objetivo sumarizar a pesquisa em uma lista única sem grande necessidade de interpretação. Essa fase é inicial dos estudos de análise de dados. É utilizado métodos de Estatística Descritiva para organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos.

Um exemplo citado por Jeff Leak (2015) é o censo dos EUA onde é efetuado o levantamento e uma coleção de dados dos residentes americanos, como nome, idade, sexo, raça com o objetivo estritamente descritivo. A interpretação desses dados não é um papel do *data analysis* nessa etapa.

ANÁLISE EXPLORATÓRIA

Nesse tipo de análise, é desenvolvido uma análise descritiva através de descobertas de padrões, tendências, correlações ou relacionamentos entre a medição de múltiplas variáveis para gerar novas ideias ou hipóteses. É basicamente uma “geradora de hipóteses” como afirma Peng (2016), por a ideia aqui não é testar uma hipótese, mas simplesmente aponta-la, descobri-la por meio da análise de padrões.

A análise exploratória de dados (EDA – *Exploratory Data Analysis*) é uma abordagem da análise de dados que aplica uma variedade de técnicas visando:

- Maximizar os insights dentro de uma base de dados
- Descobrir estruturas subjacentes
- Descobrir um modelo parcimonioso, ou seja, que explica os dados com uma quantidade mínima de variáveis preditoras.
- Verificar pressupostos associados a qualquer modelo ou teste de hipótese
- Criar uma lista de prováveis anomalias de dados
- Encontrar parâmetros estimados e seus intervalos ou margem de erro associados
- Extrair as variáveis mais importantes

ANÁLISE INFERENCIAL

Esse tipo de análise vai além da EDA ao quantificar se um padrão observado se manterá além da massa de dados em mãos, ou seja, ao analisar dados dessa forma deve-se verificar se aquele padrão que foi descoberto se aplicaria a uma população que ultrapassa a massa de dados em poder do analista, inferindo um comportamento ou afirmação.

ANÁLISE PREDITIVA

A análise preditiva de dados visa explorar e mensurar e buscar correlação de predições focadas para indivíduos ou unidades. Diferente da inferencial, onde a interpretação e focada numa visão generalista e populacional.

Um exemplo de análise preditiva seria o de uma empresa efetuar uma pesquisa para determinar quantas pessoas irão votar em certo candidato numa eleição. Em alguns casos a métrica usada para a predizer o resultado será intuitiva. A análise preditiva consegue mostrar o provável comportamento individual frente alguma variável, mas é geralmente é incapaz de explicar o porquê do resultado obtido.

CAUSAL

Na análise causal, o objetivo é verificar como a alteração de uma medição pode afetar o resultado de outra. Pesquisadores poderiam dessa forma determinar que certa alteração em um elemento favorece ou prejudica outro elemento.

Uma demonstração desse tipo de análise seria por exemplo o estudo se certo tipo de alimentação e uma dada parcela da população diminuiria a incidência de algum tipo de doença especificada. Observe que este tipo de análise é muito comum e ocorre frequentemente, em diversas áreas do conhecimento.

ANÁLISE MECANICISTA

O conceito de da análise mecanicista procura demonstrar que mudar uma medida ou variável sempre e exclusivamente leva a um comportamento específico, determinista em outro. O grande objetivo aqui não é apenas entender que alterando uma certa variável resultará em uma alteração em outra – como ocorre na análise causal – mas sim como essa ocorre, de forma determinista.A análise de dados mecanicista é extremamente desafiadora e raramente realizada.

CONCLUSÃO

A quantidade de informação gerada e consumida no cenário mundial contemporâneo, devido a evolução da internet e o acesso globalizado das mídias digitais exige medidas de grande importância para o aproveitamento dessa massa de dados de forma estratégica. O conceito de Ciência de Dados é cada dia mais recorrente e entender como funciona esse processo é uma grande vantagem de mercado, além de saber fazer as perguntas corretas aos dados a fim de extrair valor adequado, que é o grande segredo de uma análise de dados bem efetuada com os resultados esperados.

REFERÊNCIAS

MATSUI, E.; PENG, R. **The Art of Data Science-A Guide for Anyone Who Works with Data.**Leanpub, Victoria British Columbia, 2015.

LEEK, Jeff. **The Elements of Data Analytic Style**. Leanpub, Victoria British Columbia, 2015.

WALLER, M. A.; FAWCETT, S. E.,**Data Science, Predictive Analytics, and Big Data:** A Revolution That Will Transform Supply Chain Design and Management.Journal of business logistics, 2013.

MCAFEE, A; BRYNJOLFSSON, E. **Big Data: The Management Revolution**. HarvardBusiness Review, outubro de 2012.

REIS, Edna Afonso; REIS, Ilka Afonso**. Análise Descritiva de Dados**. Síntese numérica Estatística - Universidade Federal de Minas Gerais - Instituto de Ciências Exatas., 2002.

LEEK, jeffery t.; PENG, Roger D., **What is the question?**Science, 20 Mar 2015:Vol. 347, Issue 6228, pp. 1314-1315

What is EDA? Engineering Statistics Handbook. Disponível em <http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm> Acesso em 15 de outubro de 2017