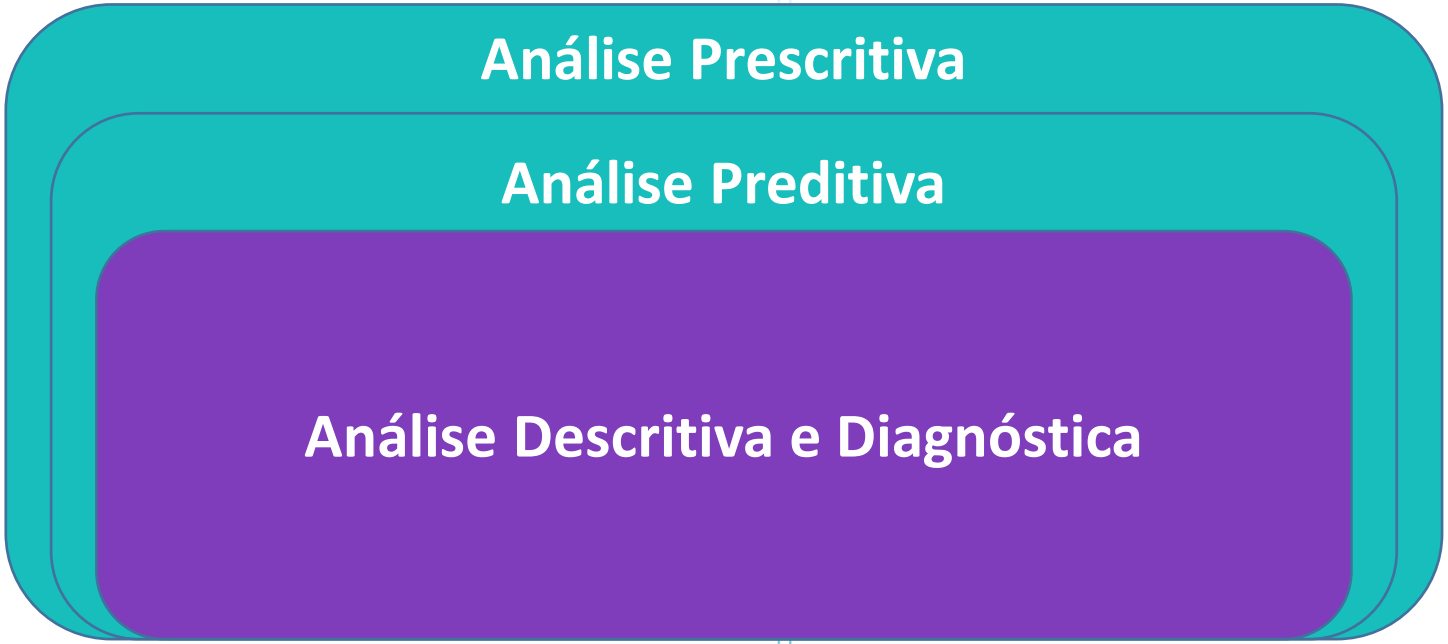
DATA DISCOVERY E ANALYTICS

Rodrigo Moravia

PUC Minas
Virtual

Análise Descritiva

Etapas de Data Discovery



Fonte: Imagem do Autor (2022)

Introdução

"A coleta de dados estatísticos tem crescido muito nos últimos anos em todas as áreas de pesquisa, especialmente com o advento dos computadores e surgimento de softwares cada vez mais sofisticados. Ao mesmo tempo, olhar uma extensa listagem de dados coletados não permite obter praticamente nenhuma conclusão, especialmente para grandes conjuntos de dados, com muitas características sendo investigadas". Reis (2002)

- A Análise Descritiva é a fase inicial deste processo de estudo dos dados coletados. Utilizamos 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.

Introdução

A descrição dos dados também tem como objetivo identificar anomalias, até mesmo resultante do registro incorreto de valores, e dados dispersos, aqueles que não seguem a tendência geral do restante do conjunto.

Mas é cada vez mais frequente a utilização destes recursos de descrição para complementar a apresentação de um fato, justificar ou referendar um argumento.

O que é?

- A análise de dados **descritiva** diz respeito a conhecer o que está acontecendo na organização e entender tendências e causas subjacentes de tais ocorrências, envolvendo a consolidação de fontes de dados e a disponibilidade de todos os dados julgados como sendo relevantes de um modo que permita a extração e a análise apropriadas de relatórios.
- Data Warehouse, onde podemos desenvolver relatórios, consultas, alertas e tendências apropriados usando ferramentas e técnicas de extração de relatórios.
- Uma tecnologia significativa que se tornou parte fundamental dessa área é a da visualização.

O que é?

- Projetos de análise de dados que ignoram tarefas de adequação de dados (algumas das etapas mais cruciais) muitas vezes acabam gerando respostas erradas para o problema certo, e essas respostas aparentemente boas, criadas sem querer, podem levar a decisões imprecisas e inoportunas.

Características

Segundo Sharda (2019), algumas das características (métricas) que definem a adequação dos dados para um estudo de análise de dados:

Confiabilidade
Acessibilidade
Riqueza de dados
Granularidade
Relevância
Precisão e consistência
Segurança e privacidade
Valor corrente/atualidade dos dados
Validade

Vantagens / Desvantagens

- **Vantagem** principal é ser um instrumento que confere **imparcialidade** a um estudo, evitando que se formem juízos de valor.
    - Também é o método mais indicado quando se deseja ter uma visão abrangente de um fenômeno e para coletar dados sobre comportamentos.

- **Desvantagem** é se a **amostra utilizada que, se mal selecionada, pode levar a respostas confusas** ou mesmo não verdadeiras, o que pode levar a tomadas de decisões incorretas.

PUC Minas
Virtual

Tipos de Variáveis

Tipos de Variáveis

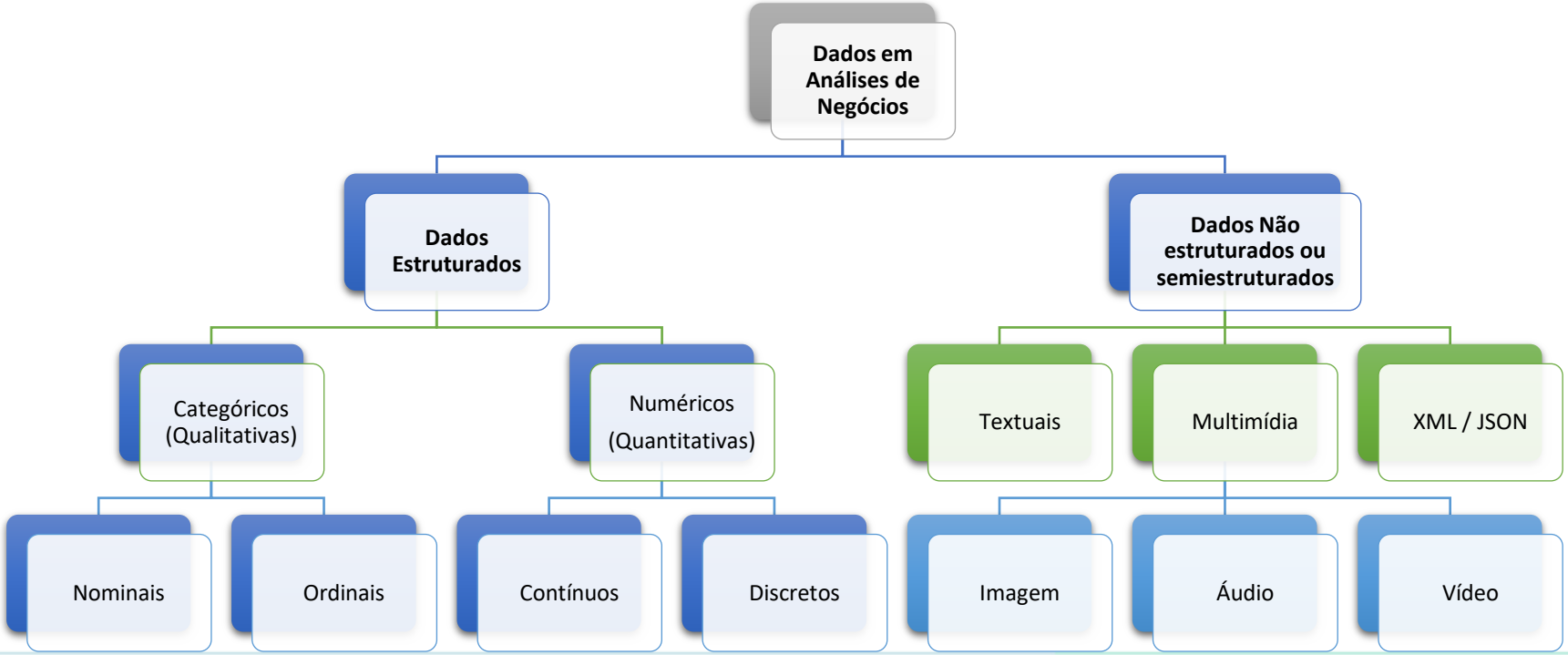
Variável é a característica de interesse que é medida em cada indivíduo da amostra ou população. Como o nome diz, seus valores variam de indivíduo para indivíduo. As variáveis podem ter valores numéricos ou não numéricos.

Levantamento Ursos de um parque nacional americano

		VARIÁVEIS →									
		Nome	Mês Obs.	Idade	Sexo	Cabeça Comp.	Cabeça Larg.	Pescoço Peri.	Altura	Tórax Peri.	Peso
ELEMENTOS ↓	1	Allen	jul	19	macho	25,4	12,7	38,1	114,3	58,4	29,5
	2	Berta	jul	19	fêmea	27,9	16,5	50,8	120,7	61,0	31,8
	3	Clyde	jul	19	macho	27,9	14,0	40,6	134,6	66,0	36,3
	4	Doc	jul	55	macho	41,9	22,9	71,1	171,5	114,3	156,2
	5	Quincy	set	81	macho	39,4	20,3	78,7	182,9	137,2	188,9
	6	Kooch	out	*	macho	40,6	20,3	81,3	195,6	132,1	196,1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	93	Sara	ago	*	fêmea	30,5	12,7	45,7	142,2	82,6	51,8
	94	Lou	ago	*	macho	30,5	14,0	38,1	129,5	61,0	37,2
	95	Molly	ago	*	fêmea	33,0	15,2	55,9	154,9	101,6	104,4
	96	Graham	jul	*	macho	30,5	10,2	44,5	149,9	72,4	58,1
	97	Jeffrey	jul	*	macho	34,3	15,2	50,8	157,5	82,6	70,8

Fonte: Reis (2002)

Taxonomia de Dados (variáveis)



Fonte: próprio autor

Taxonomia - Quantitativas

- **Dados numéricos ou Variáveis Quantitativas:** representam os valores numéricos de variáveis específicas, como por exemplo, idade, número de filhos, renda familiar total, distância viajada (em quilômetros). Os valores numéricos que representam uma variável podem ser inteiros ou reais. são as características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que fazem sentido. Podem ser contínuas ou discretas.

Taxonomia

- **Variáveis contínuas:** características mensuráveis que assumem valores em uma escala contínua (na reta real), para as quais valores não-inteiros (com casas decimais) fazem sentido. Usualmente devem ser medidas através de algum instrumento.
  - Exemplos: peso (balança), altura (régua), tempo (relógio), pressão arterial, idade.

- **Variáveis discretas:** características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros. Geralmente, são o resultado de contagens.
  - Exemplos: número de filhos, número de bactérias por litro de leite, número de cigarros fumados por dia.

Taxonomia – Variáveis Qualitativas

- **Variáveis Qualitativas (ou categóricas):** são as características que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. **Podem ser nominais ou ordinais.**

Taxonomia – Variáveis Qualitativas

- **Dados ordinais** contêm códigos atribuídos a objetos ou eventos na forma de designações, que também representam o ranking entre eles. A variável nível de crédito, por exemplo, pode ser geralmente categorizada como (1) baixo, (2) médio ou (3) alto.
    - Relações ordenadas similares podem ser vistas em variáveis como *grupo etário* (criança, jovem, meia-idade, idoso), *escolaridade* (ensino médio, superior, pós-graduação) e *estágio da doença* (inicial, intermediário, terminal).

Taxonomia – Variáveis Qualitativas

- **Dados nominais** não existe ordenação entre as categorias.

  Exemplos: sexo, cor dos olhos, fumante/não fumante, doente/sadio

Taxonomia – IMPORTANTE

- Uma variável originalmente quantitativa pode ser coletada de forma qualitativa: a variável idade, medida em anos completos, é quantitativa (contínua); mas, se for informada apenas a faixa etária (0 a 5 anos, 6 a 10 anos, etc...), é qualitativa (ordinal).

- Outro ponto importante é que nem sempre uma variável representada por números é quantitativa. O número do telefone de uma pessoa, o número da casa, o número de sua identidade. Às vezes o sexo do indivíduo é registrado na planilha de dados como 1 se macho e 2 se fêmea. Isto não significa que a variável sexo passou a ser quantitativa !

PUC Minas
Virtual

Distribuição de Frequências de uma Variável

Distribuição de Frequência

- **Variáveis Qualitativas – Nominais e Ordinais**
- No exemplo dos ursos, uma das duas variáveis qualitativas presentes é o sexo dos animais. Para organizar os dados provenientes de uma variável qualitativa, é usual fazer uma tabela de frequências com que ocorre cada um dos sexos no total dos 97 ursos observados.
- Cada categoria da variável sexo (fêmea, macho) é representada numa linha da tabela. Há uma coluna com as contagens de ursos em cada categoria (frequência absoluta) e outra com os percentuais que essas contagens representam no total de ursos (frequência relativa).

Distribuição de Frequência

- Como vimos anteriormente, as variáveis de um estudo dividem-se em quatro tipos: qualitativas (nominais e ordinais) e quantitativas (discretas e contínuas).
- Os dados gerados por esses tipos de variáveis são de naturezas diferentes e devem receber tratamentos diferentes.

Distribuição de Frequência

- Esse tipo de tabela representa a distribuição de frequências dos ursos segundo a variável sexo. Como a variável sexo é qualitativa nominal, ou seja, não há uma ordem natural em suas categorias, a ordem das linhas da tabela pode ser qualquer uma. É comum a disposição das linhas pela ordem decrescente das frequências das classes.

Distribuição de frequências dos ursos segundo sexo

Sexo	Frequência Absoluta	% (Frequência Relativa)
Fêmea	35	36,1
Macho	62	63,9
TOTAL	97	100,0

Distribuição de Frequência

- Quando a variável tabelada for do tipo qualitativa ordinal, as linhas da tabela de frequências devem ser dispostas na ordem existente para as categorias.
- Em uma tabela podemos mostrar a distribuição de frequências dos ursos segundo o mês de observação, que é uma variável qualitativa ordinal.

Distribuição de Frequência

- Nesse caso, podemos acrescentar mais duas colunas com as frequências acumuladas (absoluta e relativa), que mostram, para cada mês, a frequência de ursos observados até aquele mês.

Distribuição de frequências dos ursos segundo mês de observação

	Frequência Simples		Frequências Acumuladas	
Mês de Observação	Frequência Absoluta	% (Frequência Relativa)	Frequência Absoluta Acumulada	% (Frequência Relativa Acumulada)
Abril	8	8,2	8	8,2
Maio	6	6,2	14	14,4
Junho	6	6,2	20	20,6
Julho	11	11,3	31	32,0
Agosto	23	23,7	54	55,7
Setembro	20	20,6	74	76,3
Outubro	14	14,4	88	90,7
Novembro	9	9,3	97	100,0
TOTAL	97	100,0		

Distribuição de Frequência

- **Variáveis Quantitativas - Discretas**
- Quando estamos trabalhando com uma variável discreta que assume poucos valores, podemos dar a ela o mesmo tratamento dado às variáveis qualitativas ordinais, assumindo que cada valor é uma classe e que existe uma ordem natural nessas classes.
- Como exemplo, a tabela mostra a distribuição de frequências do número de filhos por família em uma localidade, que, nesse caso, assumiu apenas seis valores distintos.

Distribuição de Frequência

- **Variáveis Quantitativas - Discreta**
- Quando estamos trabalhando com uma variável discreta que assume poucos valores, podemos dar a ela o mesmo tratamento dado às variáveis qualitativas ordinais, assumindo que cada valor é uma classe e que existe uma ordem natural nessas classes.
- Como exemplo, a tabela mostra a distribuição de frequências do número de filhos por família em uma localidade, que, nesse caso, assumiu apenas seis valores distintos.

Distribuição de Frequência

Distribuição de frequências do número de filhos por família em uma localidade (25 lares)

Número de Filhos	Frequência Absoluta	% (Frequência Relativa)	Frequência Absoluta Acumulada
0	1	4,0	4,0
1	4	16,0	20,0
2	10	40,0	60,0
3	6	24,0	84,0
4	2	8,0	92,0
5	2	8,0	100,0
TOTAL	25	100	

Distribuição de Frequência

- **Variáveis Quantitativas - Contínua**
- Quando a variável em estudo é do tipo contínua, que assume muitos valores distintos, o agrupamento dos dados em classes será sempre necessário na construção das tabelas de frequências

Distribuição de frequências dos ursos machos segundo peso

Peso (kg)	Freqüência Absoluta	Freqüência Relativa (%)	Freq. Abs. Acumulada	Freq. Rel. Acumulada (%)
0 |- 25	3	4,8	3	4,8
25 |- 50	11	17,7	14	22,6
50 |- 75	15	24,2	29	46,8
75 |- 100	11	17,7	40	64,5
100 |- 125	3	4,8	43	69,4
125 |- 150	4	6,5	47	75,8
150 |- 175	8	12,9	55	88,7
175 |- 200	5	8,1	60	96,8
200 |- 225	1	1,6	61	98,4
225 |- 250	1	1,6	62	100
Total	62	100	-	-

Distribuição de Frequência

- Os limites das classes são representados de modo diferente daquele usado nas tabelas para variáveis discretas: o limite superior de uma classe é igual ao limite inferior da classe seguinte.
- O símbolo “|” é utilizado. Na segunda classe (25 |- 50), por exemplo, estão incluídos todos os ursos com peso de 25,0 a 49,9 kg. Os ursos que porventura pesarem exatos 50,0 kg serão incluídos na classe seguinte.
- Utiliza-se muito o gráfico de Histograma para estas situações.

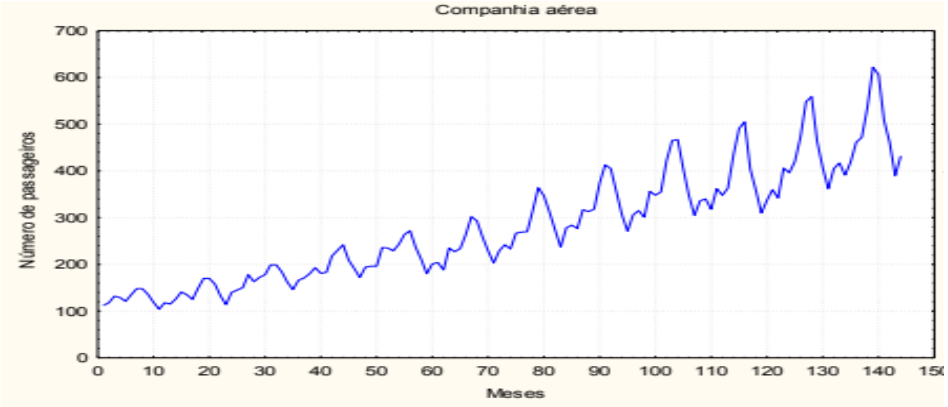
PUC Minas
Virtual

Séries Temporais

Temporais

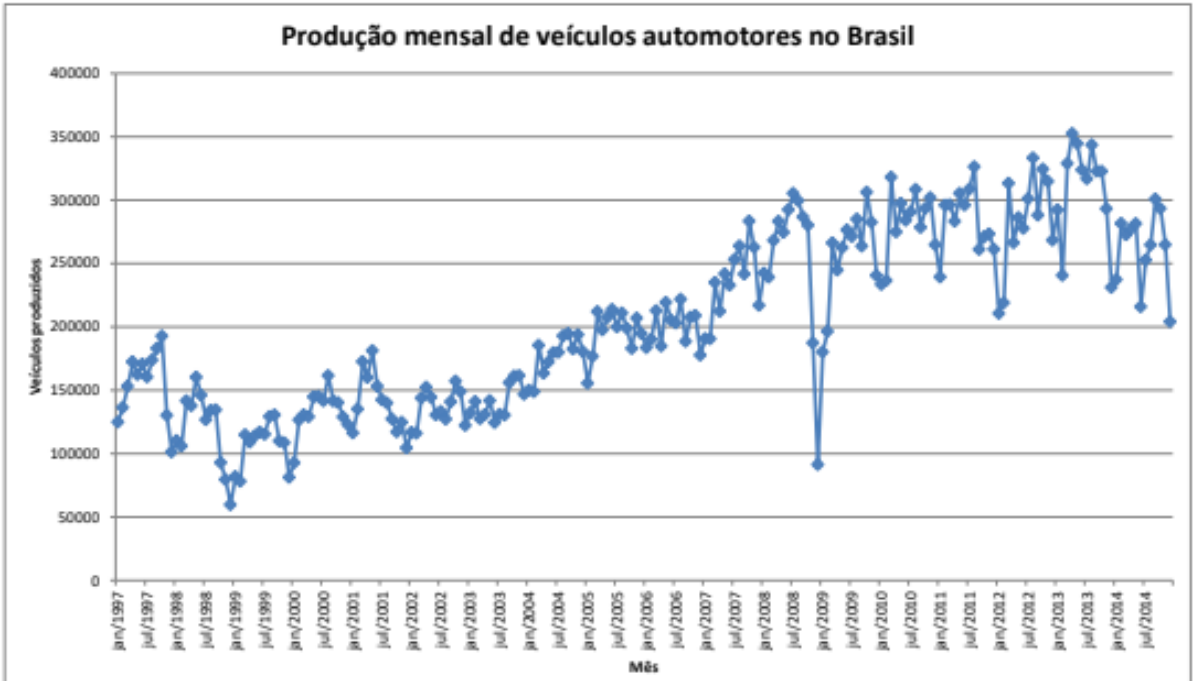
- **Série Temporal** é um conjunto de observações sobre uma variável, ordenado no tempo: diariamente (preço de ações, relatórios meteorológicos), mensalmente (taxa de desemprego, IPC), trimestralmente (PIB).
- Um dos objetivos do estudo de séries temporais é conhecer o comportamento da série ao longo do tempo (aumento, estabilidade ou declínio dos valores). Em alguns estudos, esse conhecimento pode ser usado para se fazer previsões de valores futuros com base no comportamento dos valores passados

Temporais - Exemplo



Há uma sucessão regular de "picos e vales" no número de passageiros transportados, isso deve ser causado pelas oscilações devido a feriados, períodos de férias escolares, etc., que estão geralmente relacionados às estações do ano, e que se repetem todo ano (com maior ou menor intensidade). Sobre o crescimento no número de passageiros transportados, flutuações sazonais.

Temporais


Produção mensal de veículos automotores no Brasil
Veículos produzidos
400000
350000
300000
250000
200000
150000
100000
50000
0
jan/1997
jul/1997
jan/1998
jul/1998
jan/1999
jul/1999
jan/2000
jul/2000
jan/2001
jul/2001
jan/2002
jul/2002
jan/2003
jul/2003
jan/2004
jul/2004
jan/2005
jul/2005
jan/2006
jul/2006
jan/2007
jul/2007
jan/2008
jul/2008
jan/2009
jul/2009
jan/2010
jul/2010
jan/2011
jul/2011
jan/2012
jul/2012
jan/2013
jul/2013
jan/2014
jul/2014
Mês

PUC Minas
Virtual

Etapas Análise Descritiva

Etapas para Análise Descritiva

Identificação do Problema

Coleta de Dados

Crítica dos Dados

Apresentação dos Dados

Análise e Interpretação

Fonte: próprio autor

Etapas - continuação

Identificação do Problema	Coleta de Dados	Crítica dos Dados	Apresentação dos Dados	Análise e Interpretação
Qual o problema ou dúvida que se deseja obter resposta.	Coletar as fontes de dados junto ao usuário “Chave”.	Não adianta apenas coletar os dados. Tem que haver uma criticidade pois, o dado por si só, não responde nada. Bom momento para realizar um Data Storytelling.	Projetar e construir os *Dashboards*.	Analisar e interpretar para decidir os rumos. Essa é uma forma de se inserir em uma cultura data driven, pois decisões intuitivas passam a ser a exceção, e não a regra.

Fonte: próprio autor

Conclusão

- Em empresas, corresponde à etapa introdutória da gestão de dados, pois reúne as informações que, depois, serão estudadas e transformadas em subsídios para definir os rumos do negócio.
- O foco da análise descritiva é compreender se, por trás de um ou mais fenômenos que se repetem, existem tendências ou padrões que possam ser mapeados.
- Vantagem principal é ser um instrumento que confere imparcialidade a um estudo, desde que os dados estejam corretos.

Próxima aula

- Análise Preditiva

REFERÊNCIAS

SHARDAN, R., TURBAN, Efraim - Business Intelligence e Análise de Dados para gestão de negócios – 4ª edição.2019 – Editora Bookman.

Reis, E.A., Reis I.A. (2002) - Análise Descritiva de Dados. Relatório Técnico do Departamento de Estatística da UFMG. Disponível em: www.est.ufmg.br. Acesso em 18 Jun. 2022.

INMON, W. H. Como Construir o Data Warehouse. Rio de Janeiro: Campus, 1997.

REFERÊNCIAS

Imagens: 270948935 Nome do autor: Rose Carson URL de Portfólio: https://www.shutterstock.com/g/Mosesstudio

Imagens: 1911700096 Nome do autor: JLStock Sobre: Multidisciplinary acumen in all creative designs. URL de Portfólio: https://www.shutterstock.com/g/JLStock

Imagens: 2156187219 Nome do autor: kavi designs Sobre: Hai peoples, I am a graphic designer, vector illustrator. and photographer. URL de Portfólio: https://www.shutterstock.com/g/kavi designs

PUC Minas
Virtual