

DigitalHouse >  
Coding School

# DATA SCIENCE

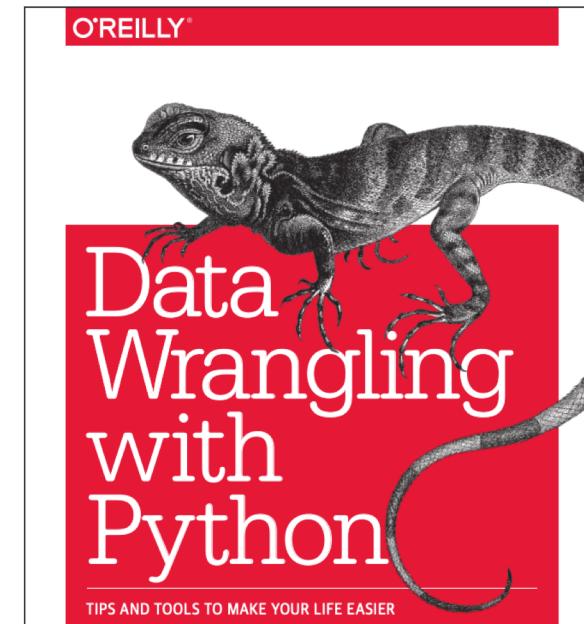
UNIDADE 1  
MÓDULO 2

Data Wrangling

- 1 Apresentação de conceitos
- 2 Limpeza e transformação de dados Prática guiada - Parte I
- 3 Variáveis categóricas e dummies, gerenciamento de strings  
Prática guiada - Parte II
- 4 TimeStamp: Gerenciamento de datas e horas no Pandas  
Prática guiada II
- 5 Prática independente.

O Data Wrangling é processo de limpeza e unificação de conjuntos de dados desordenados e complexos para facilitar o posterior acesso, exploração, análise ou modernização.

- Produzir inteligência com dados provenientes de diferentes fontes.
- Fornecer dados precisos e à mão àqueles que os necessitam.
- Reduzir o tempo gasto organizando os dados, para que possam ser utilizados.
- Permitir que cientistas de dados e analistas possam se concentrar nas análises e estudos.



Jacqueline Kazil & Katharine Jarmul  
[www.allitebooks.com](http://www.allitebooks.com)

O Data Wrangling, também conhecido como Data munging é o difícil processo de limpar, preparar e validar os dados.

- A idéia é tomar os dados em sua forma original de input e alterá-lo para um estado mais favorável, em que possamos trabalha-los e realizar análises com o mesmo.
- Limpeza.
- Transformação.
- Enriquecimento.

A limpeza é o processo inicial de abordagem dos dados, ela nos dá as primeiras informações sobre o que se passa com os dados,

- Renomeações.
- Ordenação e reordenação.
- Conversão de tipos de dados.
- Gerenciamento de dados duplicados.
- Dados faltantes ou inválidos.
- Filtragem de subconjuntos de dados.

A Transformação dos dados começa após o estágio de limpeza dos mesmos é terminado e é provável que uma alteração na forma dos dados será necessária.

- Foco na alteração da estrutura dos dados, para facilitar e canalizar as análises.
- Definição de colunas e linhas de dados, seu formato.
- **Wide Format:** Preferido em análise e planejamento de databases.
- **Long format:** Considerado como uma estruturação pobre, pois cada coluna deveria ter seu próprio tipo. Mas útil em administração de Bancos de dados.

The diagram illustrates the difference between Wide and Long data formats. On the left, a 'WIDE' table is shown with a single row for each observation. The columns are labeled 'date', 'TMAX', 'TMIN', and 'TOBS'. A bracket labeled 'observations' spans the first column, and another bracket labeled 'variables' spans the last three columns. A note says 'repeated values for date column'. On the right, a 'LONG' table is shown with three columns: 'variable names', 'datatype', and 'value'. The data is repeated for each observation, mirroring the structure of the wide table.

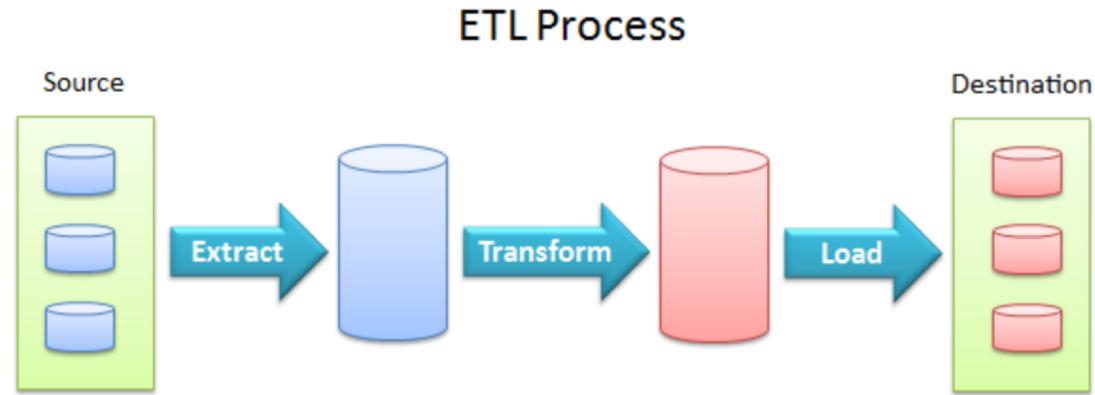
	WIDE				LONG		
observations	date	variables			variable names	datatype	value
		TMAX	TMIN	TOBS			
0	2018-10-01	21.1	8.9	13.9	0	TMAX	21.1
1	2018-10-02	23.9	13.9	17.2	1	TMIN	8.9
2	2018-10-03	25.0	15.6	16.1	2	TOBS	13.9
3	2018-10-04	22.8	11.7	11.7	3	TMAX	23.9
4	2018-10-05	23.3	11.7	18.9	4	TMIN	13.9
5	2018-10-06	20.0	13.3	16.1	5	TOBS	17.2

O enriquecimento dos dados pode se dar por fusão de um conjunto de dados a dados novos, por meio da função de apêndice, ou por meio da criação de novos dados, a partir do conjunto já existente.

- Adição de novas colunas por meio do uso de funções aplicadas à colunas pré-existentes.
- O Binning é o processo de tornar dados contínuos ou discretos em intervalos de quantizados, o que implica em valores discreteados para a coluna.
- A agregação representa operações realizadas sobre todos os dados, que resultam em valores únicos, como somas e médias.
- A re-amostragem é o processo de rearranjo dos dados em intervalos específicos.

As tarefas de **Extract, Transform and Load (ETL)**, ou no português, extrair, transformar e Lançar os dados estruturam o processo de reunir dados a partir de suas fontes e replicá-los em locais como:

- Data lakes.
- Data warehouses.
- Data marts.
- Ferramentas.



Por fim a Análise Exploratória de Dados ([AED](#)) é o processo pelo qual temos o primeiro entendimento dos dados, sumarizando suas propriedades principais e eventualmente as denotando visualmente.

Resumir suas características principais, usando com métodos visuais sempre que possível. Um modelo estatístico pode ou não ser utilizado, mas a AED serve, principalmente, para ver o que os dados podem nos dizer além da tarefa de modelagem formal ou do teste de hipóteses.

- Importar as bibliotecas pertinentes.
- Carregar os dados em dataframes.
- Checando os tipos dos dados.
- Removendo colunas indesejáveis.
- Renomeando as que forem necessárias.
- Removendo linhas duplicadas.
- Removendo valores faltantes e nulos.
- Detectando outliers.

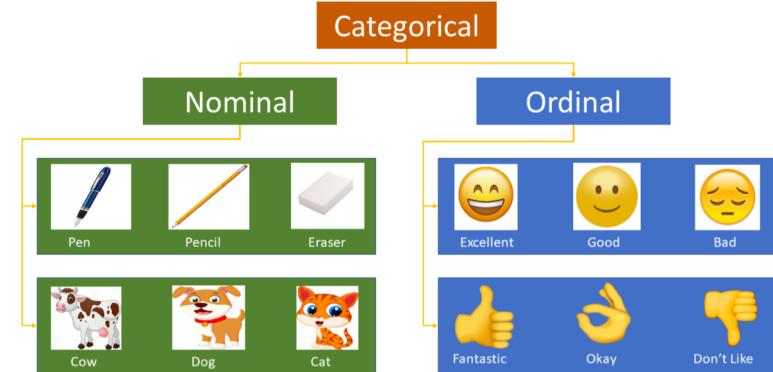
# **Prática guiada - Parte I**

## Limpeza e transformação de dados

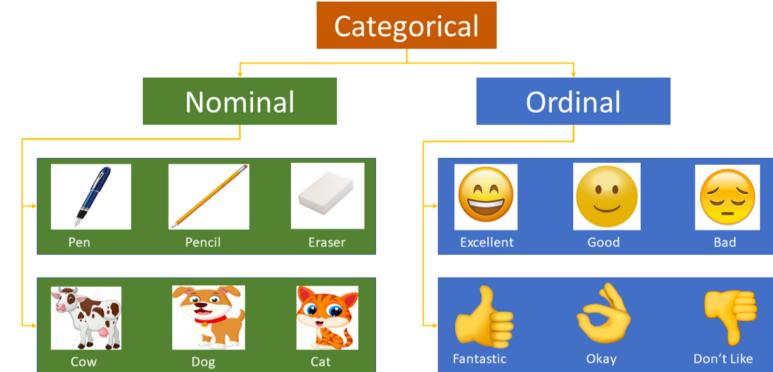
Uma **variável categórica** é uma variável que pode aceitar um número contável de valores que indicam o pertencimento a um grupo (ou categoria) com determinadas propriedades qualitativas.

- Exemplos: gênero, estado civil, grupo sanguíneo, cor do cabelo, etc.
- Uma variável categórica com dois valores possíveis é chamada de **binária** ou **dicotômica**.

- Em estatística, geralmente são atribuídos **valores ou rótulos numéricos** às variáveis categóricas:
  - Estado civil, 0 se for solteiro, 1 se for casado e 2 se for divorciado.
- Os números utilizados para rotular são arbitrários e, consequentemente, as variáveis categóricas podem ou não ser ordenadas.
  - De modo geral, o software supõe que os valores numéricos refletem quantidades algébricas e, portanto, uma ordem certa.



- A principal medida de posição é a **moda**.
  - A mediana e a média não são definidas (em geral, nenhuma operação numérica é definida).



Uma **variável dummy** é uma variável qualitativa que aceita valores 0 ou 1 para indicar a ausência ou presença de algum atributo ou efeito categórico.

- Formalmente, uma variável dummy,  $D_i$ , pode ser expressa por meio de uma **função indicadora**:

$$D_i = \mathbb{I}_A(x_i) = \begin{cases} 1 & \text{se } x_i \in A \\ 0 & \text{se } x_i \notin A \end{cases}$$

- Qual é a relação entre variáveis categóricas e variáveis dummies?
  - Variáveis categóricas podem receber um número limitado e fixo de valores possíveis, reúnem observações em grupos de categorias nominais.
  - Variáveis dummy pode receber apenas valores binários 1 e 0 para indicar a presença ou ausência de
  - Resolve o problema de interpretar as tags numéricas como um intervalo.
  - Se as categorias têm muitos valores, a dimensionalidade dos dados aumenta consideravelmente.

- Suponhamos que temos uma variável categórica, C, que registra a cidade em que reside uma amostra de habitantes do Brasil.
  - Vamos assumir que a variável pode aceitar quatro valores possíveis: São Paulo, Rio de Janeiro, Brasília e Belo Horizonte.
  - Imaginemos que temos as seguintes cinco observações:

Obs.	Cidade
1	Rio de Janeiro
2	São Paulo
3	Rio de Janeiro
4	Belo Horizonte
5	Brasília

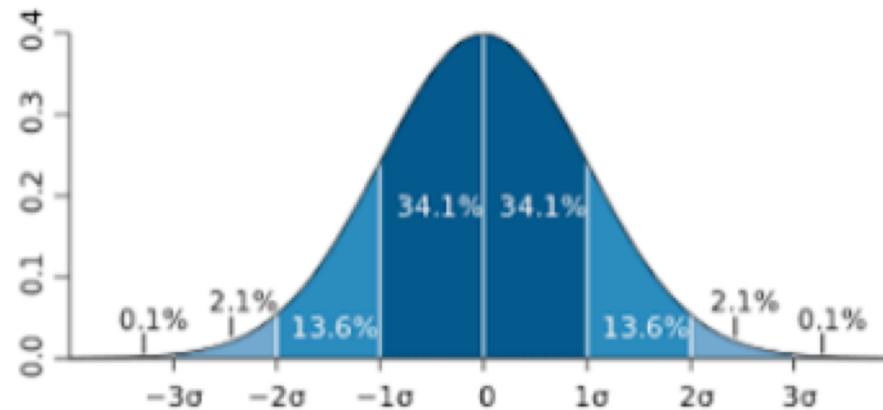
- Como alternativa, podemos expressar estas observações da variável categórica usando dummies como:
- É importante perceber que se existem k categorias, são suficientes k-1 variáveis dummies para representá-las.

Obs.	Cidade
1	Rio de Janeiro
2	São Paulo
3	Rio de Janeiro
4	Belo Horizonte
5	Brasília

Obs.	D_BA	D_C	D_R
1	0	0	1
2	1	0	0
3	0	0	1
4	0	0	0
5	0	1	0

- Recordemos que um caso muito freqüente em datasets com variáveis contínuas, é que os dados se distribuem normalmente. Uma distribuição normal pode ser caracterizada por sua média e seu desvio padrão.
- Esta estimativa nos permite esperar determinada quantidade de casos em determinados intervalos, com base na média e no desvio padrão observados.

**Como se distribui uma variável dummy?**



Ao tomar apenas dois valores, esta variável não tem uma distribuição normal, **binomial**.

- A distribuição binomial recebe um único parâmetro  $p$ , a probabilidade de ocorrência da categoria.
- Dentro da amostra, podemos calculá-lo como a **média** da variável dummy.

$$\sigma = \sqrt{\frac{p(1-p)}{n}}$$

- Em variáveis binomiais o desvio padrão  $\sigma$  é calculado como uma função da probabilidade  $p$ .
- A média e o desvio padrão são dois parâmetros dependentes.

Uma expressão regular é uma sequência de caracteres que determina um padrão de busca.

- É uma linguagem muito flexível que serve para identificar e extrair informação de um corpo de caracteres não estruturado.
- As regexes são suportadas em linguagens como Python, R, Java, SQL, Scala.
- “...a small computer language of their own...”

# Prática guiada - Parte II

# Prática independente

- Unix TimeStamp é uma maneira de acompanhar o tempo.
- O tempo é considerado como uma soma de segundos, a partir de 1 de janeiro de 1970, UTC, à 00:00:00.
- Tendo um ponto de referência fixo, é muito simples gerenciar o tempo e as datas em diferentes sistemas e arquiteturas.

## Convert date to epoch timestamp

Year	Month	Date	Hours	Minutes	Seconds	Time Zone	
2019	06 - June	07	21	34	10	Local	<button>Convert</button>

Converted epoch timestamp in seconds: 1559954050

Os dados com TimeStamps são a forma mais básica de Series de Tempo. Associam valores com pontos no tempo. No Pandas, podemos instanciar objetos da classe Timestamp

**In [8]:** pd.Timestamp(datetime(2012, 5, 1))

**Out[8]:** Timestamp('2012-05-01 00:00:00')

**In [9]:** pd.Timestamp('2012-05-01')

**Out[8]:** Timestamp('2012-05-01 00:00:00')

**In [10]:** pd.Timestamp(2012, 5, 1)

**Out[10]:** Timestamp('2012-05-01 00:00:00')

Em muitos casos, é mais natural associar variáveis a intervalos de tempo.

Nós representamos um intervalo com um objeto da classe **Period** e ele pode ser instanciado explicitamente ou inferido a partir de uma string com um determinado formato.

In [11]: pd.Period('2011-01')

Out[11]: Period('2011-01', 'M')

In [12]: pd.Period('2012-05', freq='D')

Out[12]: Period('2012-05-01', 'M')

Os objetos do tipo Timestamp e Period podem ser índices. As listas que contêm esses objetos são automaticamente convertidas em objetos do tipo **DatetimelIndex** e **PeriodIndex**, respectivamente.

In [13]: `dates = [pd.Timestamp('2012-05-01'), pd.Timestamp('2012-05-02'), pd.Timestamp('2012-05-03')]`

In [14]: `ts = pd.Series(np.random.randn(3), dates)`

In [15]: `type(ts.index)`

Out[15]: `pandas.core.indexes.datetimes.DatetimelIndex`

In [16]: `ts.index`

Out[16]: `DatetimelIndex(['2012-05-01', '2012-05-02', '2012-05-03'], dtype='datetime64[ns]', freq=None)`

Às vezes, precisamos especificar o formato que ajuda a analisar a string, por exemplo  
pd.to\_datetime('20170901 100500', format='%Y%m%d %H%M%S')

Code	Meaning	Example
%a	Weekday as locale's abbreviated name.	Mon
%A	Weekday as locale's full name.	Monday
%w	Weekday as a decimal number, where 0 is Sunday and 6 is Saturday.	1
%d	Day of the month as a zero-padded decimal number.	30
%-d	Day of the month as a decimal number. (Platform specific)	30
%b	Month as locale's abbreviated name.	Sep
%B	Month as locale's full name.	September
%m	Month as a zero-padded decimal number.	09
%-m	Month as a decimal number. (Platform specific)	9
%y	Year without century as a zero-padded decimal number.	13
%Y	Year with century as a decimal number.	2013
%H	Hour (24-hour clock) as a zero-padded decimal number.	07

# Prática guiada - Gerenciamento do tempo