

Análise de dados e Pandas

Um pouco sobre boas práticas de análise e onde o pandas ajuda

Sobre o que vamos falar...

1. Meta análise sobre análise de dados
2. Tidy data
3. Levando tudo para a máquina

Introdução

Quem sou eu?

- Sou economista pela PUC-SP e atualmente estudo estatística no IME-USP.
- Trabalhei com análises e pareceres econômicos na GO-Associados;
- Fui analista de produto e BI no enjoei.com.br;
- Atualmente sou analista de dados na Umoe Bioenergy.

1. Meta análise sobre análise de dados

Xkcd obrigatória...

MANY META-ANALYSIS STUDIES INCLUDE THE PHRASE "WE SEARCHED MEDLINE, EMBASE, AND COCHRANE FOR STUDIES..."

THIS HAS LED TO META-META-ANALYSES COMPARING META-ANALYSIS METHODS.

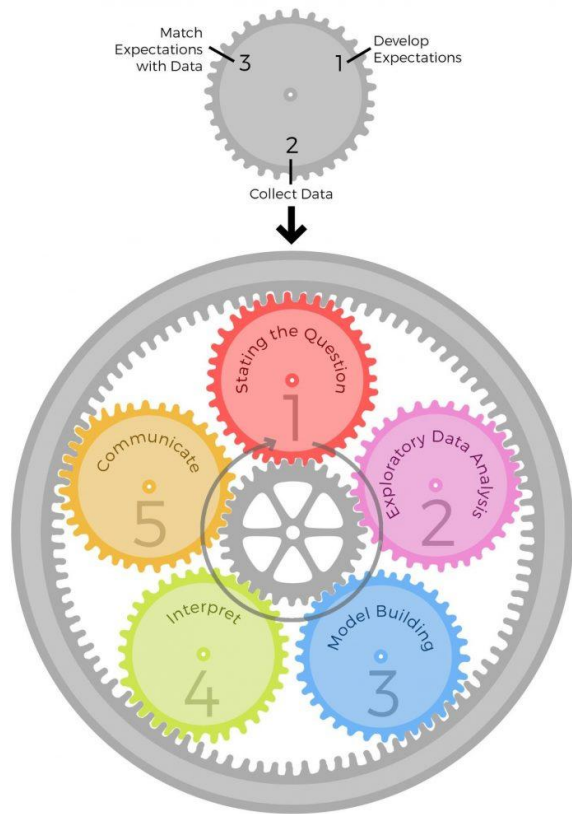
e.g. M SAMPSON (2003), PL ROYLE (2005)
E LEE (2011), AR LEMESHOW (2005)

WE PERFORMED A META-META-META-ANALYSIS OF THESE META-META-ANALYSES.

METHODS: WE SEARCHED MEDLINE, EMBASE, AND COCHRANE FOR THE PHRASE "WE SEARCHED MEDLINE, EMBASE, AND COCHRANE FOR THE PHRASE "LIFE SEARCHED MEDLINE EMBASE AND

LIFE GOAL #28: GET A PAPER REJECTED WITH THE COMMENT "TOO META"

O ciclo da análise de dados...



Antes de colocar a mão na massa:

Confronto entre expectativas + realidade dos dados;
Coleta de dados;

A análise em si:

1. Explicitando e refinando sua pergunta ←
2. Explorando os dados (EDA) ←
3. Modelagem
4. Interpretação dos resultados
5. Comunicando os resultados ←

Perguntas sobre perguntas...

Tipos de pergunta:

1. Descritiva
2. Exploratória
3. Inferencial
4. Preditiva
5. Causal
6. Mecânica

O que faz uma pergunta ser boa:

1. Interessante/necessária
2. Não respondida
3. Plausível
4. Respondível
5. Especificidade

Um pequeno “causo” “fictício”...



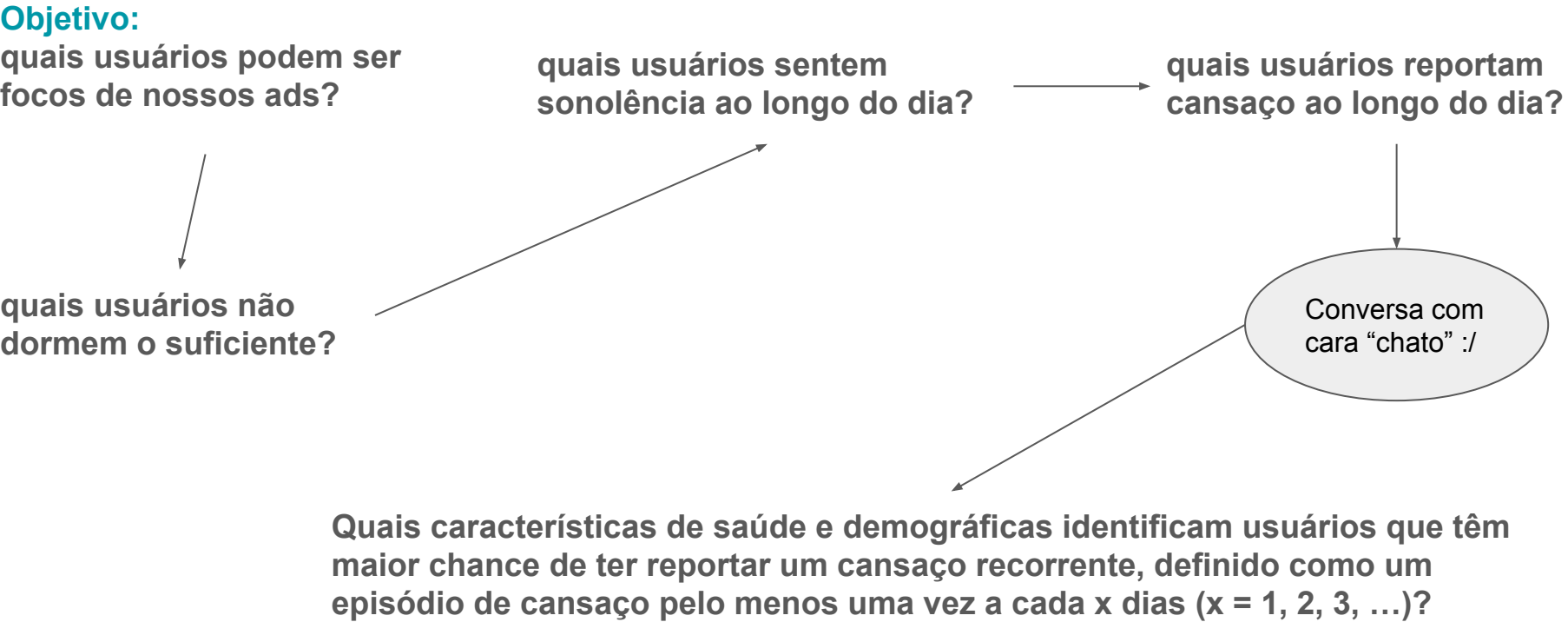
Contexto...

Uma empresa que faz aparelhos de tracking de saúde (fitbits ou coisa parecida) quer lançar um novo produto para trackear o sono & aplicativo para com dashboards e dicas para melhoria do sono que já está quase pronto.

O boss pergunta, quais usuários podem ser focos de nossos ads?

Os dados...

Demográficos...	Passos por dia	Número de degraus no dia	Horas de cansaço inputadas	Horas de sono detectadas	Horas trabalhando
-----------------	----------------	--------------------------	----------------------------	--------------------------	-------------------



COLA

- 1. Interessante/necessária
- 2. Não respondida
- 3. Plausível
- 4. Respondível
- 5. Especificidade

2. Tidy data

Quantas variáveis a seguinte tabela mostra?

	comprou	Comprou produtos	gênero	Não comprou	Não comprou produtos
	sim		M	30	
Homem	sim		F 30	44	87
	não		M	87	
Mulher	não		F 44	70	70

Definindo o tidy data

1. Cada coluna é uma variável
2. Cada observação é uma linha
3. Cada unidade observacional é uma tabela

Rápido glossário

Armazenamento	Significado
tabela/arquivo	Conjunto de dados
linhas	observações
colunas	variáveis
célula	valor

Mas o tidy data não é desambíguo...

A definição de sua pergunta ainda é importante!

Qual está certo?

id	x	y
1	22.19	24.05
2	19.82	21.98
3	19.81	21.19
4	17.45	19.43

OU

id	variável	valor
1	x	22.19
2	x	19.82
3	x	19.81
4	x	17.45
1	y	24.05
2	y	21.98
3	y	21.19
4	y	19.43

Várias maneiras de dados não estarem tidy...

- Títulos das colunas não são variáveis e sim valores.
- Mais de uma variável em uma coluna.
- Variáveis presentes nas colunas e nas linhas.
- Muitas unidades observacionais na mesma tabela.
- A mesma unidade observacional está em várias tabelas.

**Happy families are all alike; every unhappy family is unhappy in its own way
— Leon Tolstoy**

Ao notebook!!



3. Levando tudo para a máquina

Meus features favoritos do pandas

1. Encadeamento de métodos
2. Construtores poderosos
3. Multi-índices ✖
4. Manipulação de texto + datas + séries de tempo ✖
5. Ferramentas de manipulação de tabelas ✔
6. Modularidade, flexibilidade e velocidade

```
empurar_bruxa(  
    find(  
        trilha(  
            trilha(passear(joao_maria, 'bosque'), type = 'paões'),  
            status = Null, reason = 'passaros'),  
            bruxa = True, evil = True  
        ),  
        where = 'fogao'  
    )  
)
```

```
joao_maria = JoaoEMaria()  
(joao_maria.passear('bosque')  
    .trilha(type = 'paões')  
    .trilha(status = Null, reason = 'passaros')  
    .find('casa_de_doces', bruxa = True, evil = True)  
    .empurar_bruxa(where = 'fogao')  
)
```

E você vai precisar de:

- [assign](#) (0.16.0)
- [pipe](#) (0.16.2)
- [rename](#) (0.18.0)
- funções lambda
- group_by
- where/mask/query para filtros
- map
- Métodos de janela: `pd.rolling_*` & `pd.expanding_*`

Construtores poderosos...

`pandas.DataFrame.from_records:`

```
In [4]: record = [(1, 2., 'Hello'), (2, 3., 'World')]
```

```
In [5]: pd.DataFrame.from_records(record)
```

```
Out[5]:
```

	0	1	2
0	1	2.0	Hello
1	2	3.0	World

`pandas.DataFrame.from_dict:`

```
In [8]: dc = {'coluna_1': [0, 1], 'coluna_2': [5, 6]}
```

```
In [9]: pd.DataFrame.from_dict(dc)
```

```
Out[9]:
```

	coluna_1	coluna_2
0	0	5
1	1	6

pandas.DataFrame.from_items:

```
In [10]: itens = [('A', [1, 2, 3]), ('B', [4, 5, 6])]
```

```
In [11]: pd.DataFrame.from_items(itens)
```

```
Out[11]:
```

	A	B
0	1	4
1	2	5
2	3	6

```
In [12]: pd.DataFrame.from_items(itens, orient='index',  
...:                             columns=['coluna_1', 'coluna_2', 'coluna_3'])
```

```
Out[12]:
```

	coluna_1	coluna_2	coluna_3
A	1	2	3
B	4	5	6

Lidando com muitos arquivos...

Partindo de:

```
files = glob.glob('weather/*.csv')
columns = ['station', 'date', 'tmpf', 'relh', 'sped', 'mslp',
           'p01i', 'vsby', 'gust_mph', 'skyc1', 'skyc2', 'skyc3']
```

Jeito pythonico:

```
# iniciar um DataFrame vazio, como se fosse uma lista
```

```
weather = pd.DataFrame(columns=columns)
```

```
for fp in files:
```

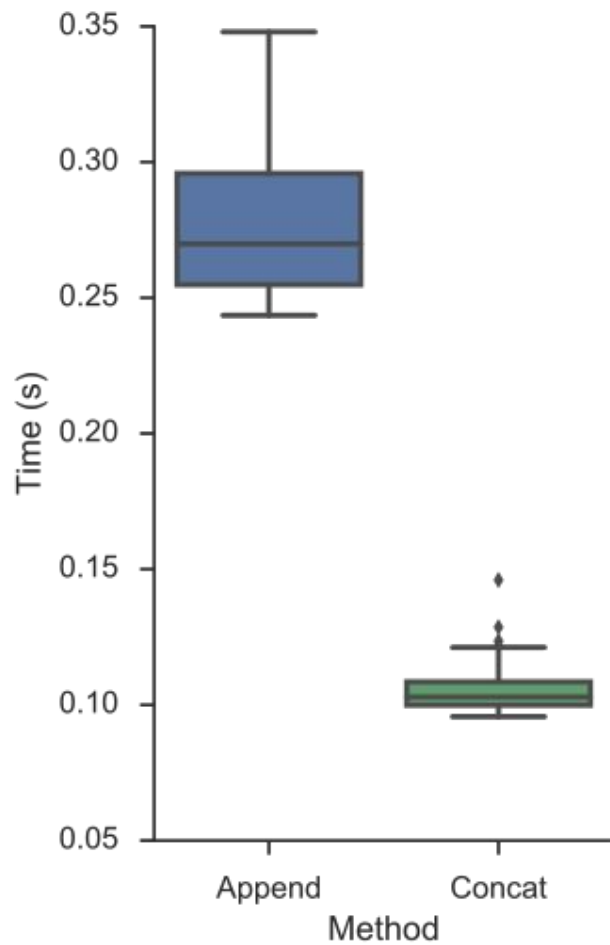
```
    city = pd.read_csv(fp, columns=columns)
```

```
    weather.append(city)
```

Jeito pandorable:

```
weather_dfs = [pd.read_csv(fp, names=columns) for fp in files]
```

```
weather = pd.concat(weather_dfs)
```



Modularidade, flexibilidade e velocidade...

Para facilitar a análise, separar o código em funções/scripts/modulos:

`make_dataset.py`

`build_features.py`

`predict_model.py`

`train_model.py`

`visualize.py`

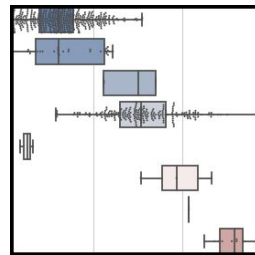
Ferramentas interessantes...

Cookiecutter Data Science

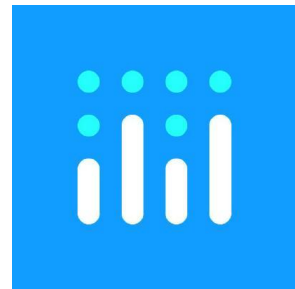
A logical, reasonably standardized, but flexible project structure for doing and sharing data science work.



Apache
Airflow




Seaborn



Dash by plotly



\$ click_ 

Alguns
exemplos!!



Referências...

Art of Data Science por Roger Peng - <https://leanpub.com/artofdatascience>

Tidy Data por Hadley Wickham - <http://vita.had.co.nz/papers/tidy-data.html>

Tidy Data vignette - <https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html>

Modern Pandas por Tom Augspurger - <https://tomaugspurger.github.io/modern-1.html>

Python for data analysis por William McKinney - <http://shop.oreilly.com/product/0636920023784.do>

Reproducible Data Analysis in Jupyter por Jake Vanderplas - www.youtube.com/channel/UCscdxGKSj4hOaVFYvslW1-g/

Data Science Cookiecutter por Driven Data - <http://drivendata.github.io/cookiecutter-data-science/>

