09

Faça o que eu fiz na aula

```
import pandas as pd
```

Ao ler a nossa base de dados vamos alterar o tipo de 3 colunas por boas práticas:

Date: Por ser uma data no formato ano/mês/dia, vamos converter para texto e assim conseguir separar os caracteres.

fullVisitorID e VisitId: Esses campos são ID's que podem possuir zeros a esquerda, se mantermos como número, os zeros a esquerda serão eliminados, alterando assim o valor do campo.

Para corrigir o formato dessas colunas, vamos ler a base passando o parâmetro "dtype" especificando o tipo de dados para essas três colunas.

```
df = pd.read_csv('train.csv', dtype={'date':object,'fullVisitorId':object,'VisitId':object})
```

Ao ler a base a saída do head deve ser:

```
df.head()
```

	channelGrouping	date	device	fullVisitorId	geoNetwork	sessionId	socialEngagementType	totals
0	Organic Search	20160902	{"browser": "Chrome", "browserVersion": "not a...	9674781571160116268	{"continent": "Asia", "subContinent": "Southea...	9674781571160116268_1472804607	Not Socially Engaged	{"visits": "1", "hits": "1", "pageviews": "1",...
1	Organic Search	20160902	{"browser": "Chrome", "browserVersion": "not a...	8590648239310839049	{"continent": "Europe", "subContinent": "Easte...	8590648239310839049_1472835928	Not Socially Engaged	{"visits": "1", "hits": "1", "pageviews": "1",...
2	Affiliates	20160902	{"browser": "Chrome", "browserVersion": "not a...	9772828344252850532	{"continent": "Americas", "subContinent": "Sou...	9772828344252850532_1472856802	Not Socially Engaged	{"visits": "1", "hits": "1", "pageviews": "1",...
3	Organic Search	20160902	{"browser": "Safari", "browserVersion": "not a...	1350700416054916432	{"continent": "Americas", "subContinent": "Nor...	1350700416054916432_1472879649	Not Socially Engaged	{"visits": "1", "hits": "1", "pageviews": "1",...
4	Organic Search	20160902	{"browser": "Safari", "browserVersion": "not a...	1350700416054916432	{"continent": "Americas", "subContinent": "Nor...	1350700416054916432_1472829671	Not Socially Engaged	{"visits": "1", "hits": "5", "pageviews": "4",...

Para manipular os dados no formato de dicionários, vamos utilizar a biblioteca "json".

```
import json
```

Primeiramente, vamos identificar as colunas que contém dicionários:

```
dicionarios = ['device','geoNetwork','trafficSource','totals']
```

Nesta parte do código, para cada coluna que definimos na variável dicionário, vamos carregar cada uma das linhas com o método json.loads. Estas linhas são alocadas dentro de uma lista que será convertida em um DataFrame. Ao fazer a

conversão a biblioteca pandas vai separar automaticamente cada chave do json em uma coluna do Dataframe.

Após isso basta unir o nosso DataFrame original(df) ao novo DataFrame gerado através da função "join".

```
for coluna in dicionarios:
    df = df.join(
        pd.DataFrame([json.loads(linha) for linha in df[coluna]])
    )
```

Após finalizarmos a transformação podemos excluir as colunas originais dos dicionários:

```
df.drop(dicionarios, axis=1, inplace=True)
```

```
df.head()
```

	channelGrouping	date	fullVisitorId	sessionId	socialEngagementType	visitId	visitNumber	visitStartTime
0	Organic Search	20160902	9674781571160116268	9674781571160116268_1472804607	Not Socially Engaged	1472804607	1	1472804607
1	Organic Search	20160902	8590648239310839049	8590648239310839049_1472835928	Not Socially Engaged	1472835928	1	1472835928
2	Affiliates	20160902	9772828344252850532	9772828344252850532_1472856802	Not Socially Engaged	1472856802	1	1472856802
3	Organic Search	20160902	1350700416054916432	1350700416054916432_1472879649	Not Socially Engaged	1472879649	2	1472879649
4	Organic Search	20160902	1350700416054916432	1350700416054916432_1472829671	Not Socially Engaged	1472829671	1	1472829671

Como vimos no curso, a coluna não possui informação relevante, então podemos deletá-la.

```
df.drop('adwordsClickInfo',axis=1,inplace=True)
```

Agora vamos criar um loop para identificar a quantidade de valores únicos por coluna e excluir as que tiver apenas 1 valor.

```
coluna_na = []
for coluna in df.columns:
    print(coluna + ': ' + str(len(df[coluna].unique())))
    if len(df[coluna].unique()) == 1:
        coluna_na.append(coluna)

df.drop(coluna_na,axis=1,inplace=True)
```

Ao final da exclusão o head deve ser:

```
df.head()
```

	channelGrouping	date	fullVisitorId	sessionId	visitId	visitNumber	visitStartTime	browser	deviceCategory
0	Organic Search	20160902	9674781571160116268	9674781571160116268_1472804607	1472804607	1	1472804607	Chrome	desktop
1	Organic Search	20160902	8590648239310839049	8590648239310839049_1472835928	1472835928	1	1472835928	Chrome	desktop
2	Affiliates	20160902	9772828344252850532	9772828344252850532_1472856802	1472856802	1	1472856802	Chrome	desktop
3	Organic Search	20160902	1350700416054916432	1350700416054916432_1472879649	1472879649	2	1472879649	Safari	mobile
4	Organic Search	20160902	1350700416054916432	1350700416054916432_1472829671	1472829671	1	1472829671	Safari	mobile

E o tamanho da sua base deve ser:

```
df.shape
```

(12283, 31)