08

Faça o que eu fiz na aula

A nossa estratégia para cada usuário será a seguinte: mostrar as informações qualitativas da primeira e última sessão, e a soma dos valores quantitativos. Como primeiro passo, vamos criar um Dataframe com o fullVisitorId e com a última sessão que temos para aquele usuário no site.

```
visitas_ultima = df.groupby('fullVisitorId',as_index=False)
visitas_ultima = visitas_ultima['visitNumber'].max()
visitas_ultima.head()
```

	fullVisitorId	visitNumber
0	0002365800130207040	1
1	0010286039787739137	1
2	0011056874471185769	1
3	0014443856125569702	1
4	0017260116665815114	6

Agora vamos salvar um DataFrame com os valores únicos entre usuário e visita.

```
usuarios_visitas_unicos = df.drop_duplicates(subset=['fullVisitorId','visitNumber'])
```

Agora podemos recuperar para cada última sessão de usuário, os valores para aquela sessão.

```
visitas = pd.merge(visitas_ultima,usuarios_visitas_unicos,left_on=['fullVisitorId','visitNumber'],
                   right_on=['fullVisitorId','visitNumber'],how='left')
```

Assim como fizemos para a última sessão, vamos criar um DataFrame com o usuário e a sua primeira sessão no site:

```
visitas_primeira = df.groupby('fullVisitorId',as_index=False)
visitas_primeira = visitas_primeira['visitNumber'].min()
visitas_primeira.set_index('fullVisitorId',inplace=True)
```

	fullVisitorId	visitNumber
0	0002365800130207040	1
1	0010286039787739137	1
2	0011056874471185769	1
3	0014443856125569702	1
4	0017260116665815114	6

Agora vamos trazer todas as informações sobre a primeira visita para o dataframe “visitas”.

```
visitas = visitas.join(visitas_primeira,how='left',on='fullVisitorId',rsuffix='_primeira')
visitas = pd.merge(visitas,usuarios_visitas_unicos,left_on=['fullVisitorId','visitNumber_primeira'],
                   right_on=['fullVisitorId','visitNumber'],how='left', suffixes=['_ultima','_primeira'])
visitas.head()
```

Exclusão das colunas primeira e última referentes às variáveis quantitativas.

```
for coluna in quant:
    visitas.drop(coluna + '_ultima',axis=1,inplace=True)
    visitas.drop(coluna + '_primeira',axis=1,inplace=True)
```

Exclusão das colunas de ids.

```
ids = ['sessionId_ultima','visitId_ultima','sessionId_primeira', 'visitId_primeira']
visitas.drop(ids,axis=1,inplace=True)
```

Exclusão das variáveis geográficas.

```
visitas.drop(geo,axis=1,inplace=True)
```

Join das variáveis quantitativas ao dataframe visitas.

```
visitas = pd.merge(visitas,df_quant,left_on=['fullVisitorId'],
                   right_on=['fullVisitorId'],how='left')
```

Criação da variável diferença de tempo entre primeira e última visita.

```
visitas['tempo_dif'] = visitas.visitStartTime_ultima - visitas.visitStartTime_primeira
visits = df.groupby('fullVisitorId',as_index=False).count().visitNumber.values
visitas['visits'] = visits
```

Criação da variável total de visitas por usuário.

```
visits = df.groupby('fullVisitorId',as_index=False).count().visitNumber.values
visitas['visits'] = visits
```

Criação de variável com informações sobre ano, mês e dia dos acessos.

```
visitas['ano_ultima'] = pd.to_numeric([data[0:4] for data in visitas.date_ultima])
visitas['mes_ultima'] = pd.to_numeric([data[4:6] for data in visitas.date_ultima])
visitas['dia_ultima'] = pd.to_numeric([data[6:8] for data in visitas.date_ultima])
visitas['ano_primeira'] = pd.to_numeric([data[0:4] for data in visitas.date_primeira])
visitas['mes_primeira'] = pd.to_numeric([data[4:6] for data in visitas.date_primeira])
visitas['dia_primeira'] = pd.to_numeric([data[6:8] for data in visitas.date_primeira])
```

Ao final de todas as transformações, o visitas.head() e visitas.dtypes, deverá ser:

```
visitas.head()
```

	fullVisitorId	visitNumber_ultima	channelGrouping_ultima	date_ultima	visitStartTime_ultima	browser_ultima	deviceCategory_ultima
0	0002365800130207040	1	Social	20160904	1472974804	Edge	desktop
1	0010286039787739137	1	Organic Search	20160928	1475084026	Chrome	desktop
2	0011056874471185769	1	Social	20161205	1480996024	Chrome	desktop
3	0014443856125569702	1	Social	20161002	1475423502	Opera	desktop
4	0017260116665815114	6	Direct	20170420	1492707286	Safari	desktop

5 rows × 54 columns

```
visitas.dtypes
```

```
fullVisitorId               object
visitNumber_ultima           int64
channelGrouping_ultima      object
date_ultima                 object
visitStartTime_ultima        int64
browser_ultima              object
deviceCategory_ultima       object
isMobile_ultima               bool
operatingSystem_ultima      object
city_ultima                 object
continent_ultima            object
country_ultima              object
metro_ultima                object
networkDomain_ultima        object
region_ultima               object
subContinent_ultima         object
adContent_ultima            object
campaign_ultima             object
campaignCode_ultima         object
isTrueDirect_ultima         object
keyword_ultima              object
medium_ultima               object
referralPath_ultima         object
source_ultima               object
visitNumber_primeira         int64
channelGrouping_primeira    object
date_primeira               object
visitNumber_primeira         int64
visitStartTime_primeira      int64
browser_primeira            object
deviceCategory_primeira     object
isMobile_primeira             bool
operatingSystem_primeira    object
adContent_primeira          object
campaign_primeira           object
campaignCode_primeira       object
isTrueDirect_primeira       object
keyword_primeira            object
medium_primeira             object
referralPath_primeira       object
source_primeira             object
bounces                    float64
hits                         int64
newVisits                  float64
pageviews                    int64
transactionRevenue         float64
tempo_dif                    int64
visits                       int64
ano_ultima                   int64
mes_ultima                   int64
dia_ultima                   int64
ano_primeira                 int64
mes_primeira                 int64
dia_primeira                 int64
```