```
import pandas as pd
from IPython.core.display import HTML
from IPython.core.display import HTML
pd.set_option('display.max_columns', None)
```

# Bases de Dados

## DFO

- Nome: Banco de dados de mortalidade
- Base de dados: DATASUS SIM
- Localidade: Brasil/Espirito Santo
- Periodo: 2012-2016 (5 anos)

```
df_sim_es_1 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vR_MH3HjwGWD0icAnZ3LVxxKUIxYNgAE3R-m-P61g(
df_sim_es_2 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vSQAjV9c7HSRNciS8zQZFtlffe-X0V8Wcc5EAM9wqk
df_sim_es_3 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vRGEAcKThbNaB9Kfp2vdoWZIMjUaVLwDF0Cf9_534:
df_sim_es_4 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vRf-ecv5pdc0k2GR3SBdoCeZtRM6SPjSyQDFVwlUm:
df_sim_es_5 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vS9XW4Hxc0QwEENpDDXLIBgDkyaTz89AXhgYYDcw7:
df_sim_es = pd.concat([df_sim_es_1, df_sim_es_2, df_sim_es_3, df_sim_es_4, df_sim_es_5])
df_sim_es.head()
```

```
/usr/local/lib/python3.6/dist-packages/IPython/core/interactiveshell.py:2718: DtypeWarning: Columns (23) hav
  interactivity=interactivity, compiler=compiler, result=result)
```

| | ACIDTRAB | ALTCAUSA | ASSISTMED | ATESTADO | ATESTANTE | CAUSABAS | CAUSABAS_O | CAUSAMAT | CB_PRE | CIRCOE |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | NaN | 1.0 | NaN | P369/P072*P960 | 1.0 | P369 | P369 | NaN | NaN | |
| **1** | NaN | NaN | 1.0 | A419/J189/N189/E149 | 2.0 | E142 | E142 | NaN | NaN | |
| **2** | NaN | NaN | 1.0 | J969/J690/M809 | 1.0 | M809 | M809 | NaN | NaN | |
| **3** | NaN | NaN | 2.0 | T07/X950 | 3.0 | X950 | X950 | NaN | NaN | |
| **4** | NaN | NaN | 1.0 | A419/I808/N179/C169 | 5.0 | C169 | C169 | NaN | NaN | |

### DFN

- Nome: Banco de dados de nascidos vivos
- Base de dados: DATASUS SINASC
- Localidade: Brasil/Espirito Santo
- Periodo: 2012-2016

```
df_sinasc_es_1 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vQm3T1J0pWoiG-JofcT-K4gvs1KBeKuCqYIxp2
df_sinasc_es_2 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vQIyzKl_u0t_YN_stW1Nr4VHWw0XY5zdXhqi15
df_sinasc_es_3 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vTIPBsiEtG3heF7zGO6HRkVae9PAO_yJI8SVtw
df_sinasc_es_4 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vROjHq3_oX3W_j6GEVjNBoKdnoUJgE5HGIfyrW
df_sinasc_es_5 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vRsmVmV6LY0nRKPr6MeqbNmH70iM20lPQdAmq0
df_sinasc_es = pd.concat([df_sinasc_es_1, df_sinasc_es_2, df_sinasc_es_3, df_sinasc_es_4, df_sinasc_es_5])
df_sinasc_es.head()
```

| | APGAR1 | APGAR5 | CODANOMAL | CODCART | CODESTAB | CODINST | CODMUNCART | CODMUNNASC | CODMUNNATU | CODMUNRES |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 8.0 | 9.0 | NaN | NaN | 2507447.0 | MBA2910720001 | NaN | 291072 | 291630.0 | 320500 |
| **1** | 9.0 | 10.0 | NaN | NaN | 2802112.0 | MBA2913600002 | NaN | 291360 | 291360.0 | 320520 |
| **2** | 9.0 | 10.0 | NaN | NaN | 2506122.0 | MBA2922000001 | NaN | 292200 | 320405.0 | 320405 |
| **3** | 10.0 | 10.0 | NaN | NaN | 2506122.0 | MBA2922000001 | NaN | 292200 | 320501.0 | 320405 |
| **4** | 8.0 | 9.0 | NaN | NaN | 2498804.0 | MBA2922000001 | NaN | 292200 | 314430.0 | 320405 |

## Perguntas

### Quais colunas temos no SINASC?

```
df_sinasc_es.columns
```

```
Index(['APGAR1', 'APGAR5', 'CODANOMAL', 'CODCART', 'CODESTAB', 'CODINST',
       'CODMUNCART', 'CODMUNNASC', 'CODMUNNATU', 'CODMUNRES', 'CODOCUPMAE',
       'CODPAISRES', 'CODUFNATU', 'CONSPRENAT', 'CONSULTAS', 'DIFDATA',
       'DTCADASTRO', 'DTDECLARAC', 'DTNASC', 'DTNASCMAE', 'DTRECEBIM',
       'DTRECORIG', 'DTRECORIGA', 'DTREGCART', 'DTULTMENST', 'ESCMAE',
       'ESCMAE2010', 'ESCMAEAGR1', 'ESTCIVMAE', 'GESTACAO', 'GRAVIDEZ',
       'HORANASC', 'IDADEMAE', 'IDADEPAI', 'IDANOMAL', 'KOTELCHUCK', 'LOCNASC',
       'MESPRENAT', 'NATURALMAE', 'NUMERODN', 'NUMERODV', 'NUMEROLOTE',
       'NUMREGCART', 'ORIGEM', 'PARIDADE', 'PARTO', 'PESO', 'PREFIXODN',
       'QTDFILMORT', 'QTDFILVIVO', 'QTDGESTANT', 'QTDPARTCES', 'QTDPARTNOR',
       'RACACOR', 'RACACORMAE', 'RACACORN', 'SEMAGESTAC', 'SERIESCMAE', 'SEXO',
       'STCESPARTO', 'STDNEPIDEM', 'STDNNOVA', 'STTRABPART', 'TPAPRESENT',
       'TPDOCRESP', 'TPFUNCRESP', 'TPMETESTIM', 'TPNASCASSI', 'TPROBSON',
```

## Quais colunas temos no SIM?

```
df_sim_es.columns
```

```
Index(['ACIDTRAB', 'ALTCAUSA', 'ASSISTMED', 'ATESTADO', 'ATESTANTE',
       'CAUSABAS', 'CAUSABAS_O', 'CAUSAMAT', 'CB_PRE', 'CIRCOBITO', 'CIRURGIA',
       'CODCART', 'CODESTAB', 'CODIFICADO', 'CODINST', 'CODMUNCART',
       'CODMUNNATU', 'CODMUNOCOR', 'CODMUNRES', 'COMUNSVOIM', 'CRM', 'DIFDATA',
       'DTATESTADO', 'DTCADASTRO', 'DTCADINF', 'DTCADINV', 'DTCONCASO',
       'DTCONINV', 'DTINVESTIG', 'DTNASC', 'DTOBITO', 'DTRECEBIM', 'DTRECORIG',
       'DTRECORIGA', 'DTREGCART', 'ESC', 'ESC2010', 'ESCFALAGR1', 'ESCMAE',
       'ESCMAE2010', 'ESCMAEAGR1', 'ESTABDESCR', 'ESTCIV', 'EXAME', 'FONTE',
       'FONTEINV', 'FONTES', 'FONTESINF', 'GESTACAO', 'GRAVIDEZ', 'HORAOBITO',
       'IDADE', 'IDADEMAE', 'LINHAA', 'LINHAB', 'LINHAC', 'LINHAD', 'LINHAII',
       'LOCOCOR', 'MORTEPARTO', 'NATURAL', 'NECROPSIA', 'NUDIASINF',
       'NUDIASOBCO', 'NUDIASOBIN', 'NUMERODN', 'NUMERODO', 'NUMERODV',
       'NUMEROLOTE', 'NUMREGCART', 'NUMSUS', 'OBITOGRAV', 'OBITOPARTO',
       'OBITOPUERP', 'OCUP', 'OCUPMAE', 'ORIGEM', 'PARTO', 'PESO',
       'QTDFILMORT', 'QTDFILVIVO', 'RACACOR', 'SEMAGESTAC', 'SERIESCFAL',
       'SERIESCMAE', 'SEXO', 'STCODIFICA', 'STDOEPIDEM', 'STDONOVA',
       'TIPOBITO', 'TPMORTEOCO', 'TPNIVELINV', 'TPOBITOCOR', 'TPPOS',
       'TPRESGINFO', 'Unnamed: 0', 'VERSAOSCB', 'VERSAOSIST'],
      dtype='object')
```

## Quais são as colunas que em comum nos bancos de dados SINASC e SIM?

```
columns_intersection = df_sinasc_es.columns.intersection(df_sim_es.columns)
columns_intersection
```

```
Index(['CODCART', 'CODESTAB', 'CODINST', 'CODMUNCART', 'CODMUNNATU',
       'CODMUNRES', 'DIFDATA', 'DTCADASTRO', 'DTNASC', 'DTRECEBIM',
       'DTRECORIG', 'DTRECORIGA', 'DTREGCART', 'ESCMAE', 'ESCMAE2010',
       'ESCMAEAGR1', 'GESTACAO', 'GRAVIDEZ', 'IDADEMAE', 'NUMERODN',
       'NUMERODV', 'NUMEROLOTE', 'NUMREGCART', 'ORIGEM', 'PARTO', 'PESO',
       'QTDFILMORT', 'QTDFILVIVO', 'RACACOR', 'SEMAGESTAC', 'SERIESCMAE',
       'SEXO', 'Unnamed: 0', 'VERSAOSIST'],
      dtype='object')
```

## Junção das tabelas `how='left'`

```
df_merged_left = pd.merge(df_sinasc_es, df_sim_es, how='left', on = 'NUMERODN')
df_merged_left.describe()
```

| | APGAR1 | APGAR5 | CODCART_x | CODESTAB_x | CODMUNCART_x | CODMUNNASC | CODMUNNATU_x | CODMUNRES |
|---|---|---|---|---|---|---|---|---|
| count | 269923.000000 | 269978.000000 | 48.000000 | 2.732450e+05 | 54065.000000 | 273807.000000 | 266581.000000 | 273807.000( |
| mean | 8.316894 | 9.179033 | 2435.750000 | 3.107655e+06 | 320328.673079 | 320315.443323 | 317274.860601 | 320339.875 |
| std | 1.734721 | 1.599190 | 1752.429869 | 1.940048e+06 | 2199.804202 | 2195.372397 | 23782.236433 | 174.838 |
| min | 0.000000 | 0.000000 | 0.000000 | 1.250000e+02 | 110015.000000 | 110014.000000 | 110000.000000 | 320000.000( |
| 25% | 8.000000 | 9.000000 | 1446.000000 | 2.448637e+06 | 320150.000000 | 320150.000000 | 320100.000000 | 320130.000( |
| 50% | 8.000000 | 9.000000 | 2053.500000 | 2.532190e+06 | 320490.000000 | 320470.000000 | 320320.000000 | 320390.000( |
| 75% | 9.000000 | 10.000000 | 3893.000000 | 3.450198e+06 | 320520.000000 | 320520.000000 | 320520.000000 | 320500.000( |
| max | 99.000000 | 99.000000 | 6980.000000 | 9.040838e+06 | 530010.000000 | 530010.000000 | 539928.000000 | 320530.000( |

```
len(df_merged_left.index)
```

```
273807
```

## Junção das tabelas `how='inner'`

```
df_merged_inner = pd.merge(df_sinasc_es, df_sim_es, how='inner', on = 'NUMERODN')
df_merged_inner.describe()
```

| | APGAR1 | APGAR5 | CODCART_x | CODESTAB_x | CODMUNCART_x | CODMUNNASC | CODMUNNATU_x | CODMUNRES_x | |
|---|---|---|---|---|---|---|---|---|---|
| count | 1404.000000 | 1405.000000 | 0.0 | 1.424000e+03 | 257.000000 | 1438.000000 | 1415.000000 | 1438.000000 | |
| mean | 5.730057 | 7.081139 | NaN | 3.723488e+06 | 320294.817121 | 320724.509040 | 317388.977385 | 320360.959666 | 7 |
| std | 5.663572 | 5.591173 | NaN | 2.157329e+06 | 1868.600019 | 7798.646769 | 18772.723060 | 174.465849 | 2 |
| min | 0.000000 | 0.000000 | NaN | 4.520000e+02 | 292200.000000 | 291072.000000 | 150140.000000 | 320010.000000 | |
| 25% | 3.000000 | 6.000000 | NaN | 2.485572e+06 | 320490.000000 | 320240.000000 | 320060.000000 | 320150.000000 | 5 |
| 50% | 6.000000 | 8.000000 | NaN | 2.678179e+06 | 320500.000000 | 320500.000000 | 320320.000000 | 320490.000000 | 9 |
| 75% | 8.000000 | 9.000000 | NaN | 5.417139e+06 | 320530.000000 | 320520.000000 | 320530.000000 | 320501.000000 | 9 |
| max | 99.000000 | 99.000000 | NaN | 7.581467e+06 | 320530.000000 | 520870.000000 | 530010.000000 | 320530.000000 | 9 |

```
len(df_merged_inner.index)
```

```
1438
```

```
df merged outer = pd.merge(df_sinasc_es, df_sim_es, how='outer', on = 'NUMERODN', indicator=True)
df_merged_outer.describe()
```

| DFILVIVO_y | RACACOR_y | SEMAGESTAC_y | SERIESCFAL | SERIESCMAE_y | SEXO_y | STDOEPIDEM | STDONOVA | TIPOBI |
|---|---|---|---|---|---|---|---|---|
| 2542.000000 | 97139.000000 | 2627.000000 | 12051.000000 | 845.000000 | 110497.000000 | 110496.0 | 110497.000000 | 110497 |
| 3.095201 | 2.420212 | 32.647126 | 4.319061 | 4.637870 | 1.414961 | 0.0 | 0.801062 | 2 |
| 12.066113 | 1.429780 | 11.725079 | 1.880501 | 2.323685 | 0.493158 | 0.0 | 0.399203 | 0 |
| 0.000000 | 1.000000 | 0.000000 | 1.000000 | 1.000000 | 0.000000 | 0.0 | 0.000000 | 2 |

```
df_merged_outer.head()
```

| SERIESCFAL | SERIESCMAE_y | SEXO_y | STCODIFICA | STDOEPIDEM | STDONOVA | TIPOBITO | TPMORTEOCO | TPNIVELINV | TPOBITOC |
|---|---|---|---|---|---|---|---|---|---|
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | N |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | N |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | N |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | N |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | N |

```
df_merged_outer.rename(index=str, columns={"_merge": "simsim"})
```

| ERIESCFAL | SERIESCMAE_y | SEXO_y | STCODIFICA | STDOEPIDEM | STDONOVA | TIPOBITO | TPMORTEOCO | TPNIVELINV | TPOBITOCO |
|---|---|---|---|---|---|---|---|---|---|
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |

NaN NaN NaN NaN NaN NaN NaN NaN NaN Na

# Teste

NaN NaN NaN NaN NaN NaN NaN NaN NaN Na

```
import numpy as np
from scipy.stats import pearsonr
```

NaN NaN NaN NaN NaN NaN NaN NaN NaN Na

```
# Columns and target column
df.corr = df_merged_inner
target_col_name = 'df.corr'
feature_target_corr = {'','',''}

for col in df:
    if target_col_name != col:
        feature_target_corr[col + '_' + target_col_name] = \
            pearsonr(df[col], df[target_col_name])[0]

print("Feature-Target Correlations")
print(feature_target_corr)
```