Querying a data frame

January 25, 2022

1 Querying a data frame

[5]:
```
#vamos começar com um exemplo
import pandas as pd
df = pd.read_csv('resources/week-1/datasets/Admission_Predict.csv', index_col =␣
↪0)
df.head()
```

[5]:
```
             GRE Score  TOEFL Score  University Rating  SOP  LOR   CGPA  \
Serial No.
1                  337          118                  4  4.5   4.5  9.65
2                  324          107                  4  4.0   4.5  8.87
3                  316          104                  3  3.0   3.5  8.00
4                  322          110                  3  3.5   2.5  8.67
5                  314          103                  2  2.0   3.0  8.21

             Research  Chance of Admit
Serial No.
1                   1             0.92
2                   1             0.76
3                   1             0.72
4                   1             0.80
5                   0             0.65
```

[6]:
```
df.columns = [x.lower().strip() for x in df.columns]
```

[7]:
```
df.head()
```

[7]:
```
             gre score  toefl score  university rating  sop  lor  cgpa  \
Serial No.
1                  337          118                  4  4.5  4.5  9.65
2                  324          107                  4  4.0  4.5  8.87
3                  316          104                  3  3.0  3.5  8.00
4                  322          110                  3  3.5  2.5  8.67
5                  314          103                  2  2.0  3.0  8.21

             research  chance of admit
Serial No.
```

```
1                  1              0.92
2                  1              0.76
3                  1              0.72
4                  1              0.80
5                  0              0.65
```

[8]:
```
#mascaras booleanas são criadas aplicando operadores diretamente nos objetos␣
 ↪dataframe ou series do pandas
#por exemplo, no dataset graduate_admission poderíamos estar interessados em␣
 ↪ver apenas estudantes
#que tem chance de admissão maior que 0.7

#para construir uma mascara booleana para essa query, queremos projetar a␣
 ↪chance de admissão usando o operador
#indexado e aplicando o operador maior que > como comparação do valor de 0.7.
#a série resultante é indexada onde o valor de cada célular é Verdadeiro ou␣
 ↪Falso dependendo se o estudane
#tem chance de admissão maior que 0.7

admit_mask = df['chance of admit'] > 0.7
admit_mask.head()
```

[8]:
```
Serial No.
1     True
2     True
3     True
4     True
5    False
Name: chance of admit, dtype: bool
```

[9]:
```
#e agora oque fazemos com o resultado da mascara booleana?
#nós colocamos por cima dos dados para 'esconder' os dado que nós não queremos,␣
 ↪que são representados como Falso
#isso é feito usando a função .where() no dataframe original

df.where(admit_mask).head()
```

[9]:
```
             gre score  toefl score  university rating  sop  lor  cgpa  \
Serial No.
1                337.0        118.0                4.0  4.5  4.5  9.65
2                324.0        107.0                4.0  4.0  4.5  8.87
3                316.0        104.0                3.0  3.0  3.5  8.00
4                322.0        110.0                3.0  3.5  2.5  8.67
5                  NaN          NaN                NaN  NaN  NaN   NaN

             research  chance of admit
Serial No.
1                 1.0             0.92
2                 1.0             0.76
```

```
3                1.0             0.72
4                1.0             0.80
5                NaN              NaN
```

[10]:
```
#notamos que os valores representados por True aparecem normais, os que são
 ↪False vêm acompanhado com NaN
#se não quisermos esses dados NaN usamos a função dropna()

df.where(admit_mask).dropna().head()
```

[10]:
```
             gre score  toefl score  university rating  sop  lor  cgpa  \
Serial No.
1                337.0        118.0                4.0  4.5  4.5  9.65
2                324.0        107.0                4.0  4.0  4.5  8.87
3                316.0        104.0                3.0  3.0  3.5  8.00
4                322.0        110.0                3.0  3.5  2.5  8.67
6                330.0        115.0                5.0  4.5  3.0  9.34

             research  chance of admit
Serial No.
1                 1.0             0.92
2                 1.0             0.76
3                 1.0             0.72
4                 1.0             0.80
6                 1.0             0.90
```

[11]:
```
#o where é pratico mas não muito usado, daí os devs do pandas criaram outro
 ↪operador para fazer isso

df[df['chance of admit'] > 0.7].head()
```

[11]:
```
             gre score  toefl score  university rating  sop  lor  cgpa  \
Serial No.
1                  337          118                  4  4.5  4.5  9.65
2                  324          107                  4  4.0  4.5  8.87
3                  316          104                  3  3.0  3.5  8.00
4                  322          110                  3  3.5  2.5  8.67
6                  330          115                  5  4.5  3.0  9.34

             research  chance of admit
Serial No.
1                   1             0.92
2                   1             0.76
3                   1             0.72
4                   1             0.80
6                   1             0.90
```

[ ]:

[ ]:

[ ]:

[ ]: