**12 - Tratamento de dados**

Fecharemos nosso treinamento de Pandas conversando sobre tratamento de dados, uma parte importante da ciência de dados. É necessário analisar e tratar as informações com as quais estamos trabalhando de modo a ganharmos confiabilidade nos resultados como cientistas de dados, por exemplo nos modelos de Machine Learning.

Continuaremos a utilizar o dataset que importamos anteriormente. Uma das formas de verificarmos rapidamente as informações de um conjunto de dados é por meio da função info().

dataset.info()COPIAR CÓDIGO

<class 'pandas.core.frame.DataFrame'>

Index: 258 entries, Jetta Variant to Macan

Data columns (total 7 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Motor 258 non-null object

1 Ano 258 non-null int64

2 Quilometragem 197 non-null float64

3 Zero\_km 258 non-null bool

4 Acessórios 258 non-null object

5 Valor 258 non-null float64

6 Km\_media 258 non-null float64

dtypes: bool(1), float64(3), int64(1), object(2)

memory usage: 24.4+ KBCOPIAR CÓDIGO

Com ela obteremos diversas informações interessantes. Por exemplo, sabemos que existem 258 linhas no conjunto e quantos valores de cada variável não são nulos. Repare que a coluna "Quilometragem" só possui 197 informações não nulas, algo que precisaremos investigar - afinal, um modelo de regressão irá eliminar todos os items com esse tipo de informação, e pode ser que com um tratamento consigamos consertar isso.

O Pandas nos oferece a possibilidade de selecionar informações nulas ou não: a função isna().

dataset.Quilometragem.isna()COPIAR CÓDIGO

Ela nos devolve uma *series* booleana identificando quais valores são ou não "NA" (nulos, vazios e assim por diante).

Nome Jetta Variant False Passat False Crossfox False DS5 True Aston Martin DB4 False ...Phantom 2013 False Cadillac Ciel concept False Classe GLK False Aston Martin DB5 False Macan False Name: Quilometragem, Length: 258, dtype: bool

Quando os valores são NA, são identificados na *series* como True. Essa pode ser uma ferramenta interessante para verificarmos como estão se comportando os nossos dados, por exemplo criando um seletor.

dataset[dataset.Quilometragem.isna()]COPIAR CÓDIGO

| **Nome** | **Motor** | **Ano** | **Quilometragem** | **Zero\_km** | **Acessórios** | **Valor** | **Km\_media** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DS5 | Motor 2.4 Turbo | 2019 | NaN | True | ['Travas elétricas', '4 X 4', 'Vidros elétrico... | 124549.07 | 0.0 |
| A5 | Motor 4.0 Turbo | 2019 | NaN | True | ['Câmbio automático', 'Câmera de estacionament... | 56445.20 | 0.0 |
| J5 | Motor V6 | 2019 | NaN | True | ['Sensor crepuscular', 'Painel digital', 'Roda... | 53183.38 | 0.0 |
| A3 | Motor 1.0 8v | 2019 | NaN | True | ['4 X 4', 'Piloto automático', 'Central multim... | 88552.39 | 0.0 |
| Série 1 M | Motor V8 | 2019 | NaN | True | ['Controle de estabilidade', 'Central multimíd... | 94564.40 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| Lamborghini Reventón | Motor 4.0 Turbo | 2019 | NaN | True | ['Controle de tração', 'Ar condicionado', 'Cen... | 67664.86 | 0.0 |
| Benni Mini | Motor V8 | 2019 | NaN | True | ['Sensor crepuscular', 'Câmbio automático', 'C... | 126247.84 | 0.0 |
| Uno | Motor Diesel V6 | 2019 | NaN | True | ['Central multimídia', 'Sensor crepuscular', '... | 128852.21 | 0.0 |
| Santa Fe | Motor 3.0 32v | 2019 | NaN | True | ['Travas elétricas', 'Ar condicionado', '4 X 4... | 129415.33 | 0.0 |
| XC60 | Motor 4.0 Turbo | 2019 | NaN | True | ['Painel digital', 'Piloto automático', 'Centr... | 77675.79 | 0.0 |

Assim conseguiremos observar algumas características. Repare que nesses dados o ano de todos so veículos é 2019, a quilometragem é NaN e o Zero\_km é True. Pensando nisso poderíamos, por exemplo, atribuir 0 à quilometragem dos carros para podermos trabalhá-los de alguma forma no conjunto.

O Pandas inclusive nos disponibiliza uma forma de preenchermos esses valores, que é a função fillna(0).

dataset.fillna(0)COPIAR CÓDIGO

Ao fazermos isso, teremos como retorno nosso dataset com os valores NA preenchidos como 0.

| **Nome** | **Motor** | **Ano** | **Quilometragem** | **Zero\_km** | **Acessórios** | **Valor** | **Km\_media** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Jetta Variant | Motor 4.0 Turbo | 2003 | 44410.0 | False | ['Rodas de liga', 'Travas elétricas', 'Piloto ... | 88078.64 | -1981.003962 |
| Passat | Motor Diesel | 1991 | 5712.0 | False | ['Central multimídia', 'Teto panorâmico', 'Fre... | 106161.94 | -1988.170877 |
| Crossfox | Motor Diesel V8 | 1990 | 37123.0 | False | ['Piloto automático', 'Controle de estabilidad... | 72832.16 | -1971.613175 |
| DS5 | Motor 2.4 Turbo | 2019 | 0.0 | True | ['Travas elétricas', '4 X 4', 'Vidros elétrico... | 124549.07 | 0.000000 |
| Aston Martin DB4 | Motor 2.4 Turbo | 2006 | 25757.0 | False | ['Rodas de liga', '4 X 4', 'Central multimídia... | 92612.10 | -1993.242694 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| Phantom 2013 | Motor V8 | 2014 | 27505.0 | False | ['Controle de estabilidade', 'Piloto automátic... | 51759.58 | -2000.376919 |
| Cadillac Ciel concept | Motor V8 | 1991 | 29981.0 | False | ['Bancos de couro', 'Painel digital', 'Sensor ... | 51667.06 | -1976.150570 |
| Classe GLK | Motor 5.0 V8 Bi-Turbo | 2002 | 52637.0 | False | ['Rodas de liga', 'Controle de tração', 'Câmbi... | 68934.03 | -1975.929173 |
| Aston Martin DB5 | Motor Diesel | 1996 | 7685.0 | False | ['Ar condicionado', '4 X 4', 'Câmbio automátic... | 122110.90 | -1992.193660 |
| Macan | Motor Diesel V6 | 1992 | 50188.0 | False | ['Central multimídia', 'Teto panorâmico', 'Vid... | 90381.47 | -1967.142150 |

Entretanto, essa alteração não refletirá no nosso dataset original, já que apenas modificamos a visualização dele. Para mantermos tal alteração, será necessário passarmos mais um parâmetro para a função fillna(), o inplace, que deverá ser definido como True.

O parâmetro inplace também é utilizado em diversas outras funções do Pandas.

dataset.fillna(0, inplace = True)COPIAR CÓDIGO

Feito isso, apesar de não recebermos nenhum retorno, a alteração terá sido concluída. Podemos visualizá-la exibindo os cinco primeiros itens do conjunto:

dataset.head()COPIAR CÓDIGO

| **Nome** | **Motor** | **Ano** | **Quilometragem** | **Zero\_km** | **Acessórios** | **Valor** | **Km\_media** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Jetta Variant | Motor 4.0 Turbo | 2003 | 44410.0 | False | ['Rodas de liga', 'Travas elétricas', 'Piloto ... | 88078.64 | -1981.003962 |
| Passat | Motor Diesel | 1991 | 5712.0 | False | ['Central multimídia', 'Teto panorâmico', 'Fre... | 106161.94 | -1988.170877 |
| Crossfox | Motor Diesel V8 | 1990 | 37123.0 | False | ['Piloto automático', 'Controle de estabilidad... | 72832.16 | -1971.613175 |
| DS5 | Motor 2.4 Turbo | 2019 | 0.0 | True | ['Travas elétricas', '4 X 4', 'Vidros elétrico... | 124549.07 | 0.000000 |
| Aston Martin DB4 | Motor 2.4 Turbo | 2006 | 25757.0 | False | ['Rodas de liga', '4 X 4', 'Central multimídia... | 92612.10 | -1993.242694 |

Ou selecionando todos os veículos que possuem True como valor de "Zero\_km".

dataset.query('Zero\_km == True')COPIAR CÓDIGO

| **Nome** | **Motor** | **Ano** | **Quilometragem** | **Zero\_km** | **Acessórios** | **Valor** | **Km\_media** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DS5 | Motor 2.4 Turbo | 2019 | 0.0 | True | ['Travas elétricas', '4 X 4', 'Vidros elétrico... | 124549.07 | 0.0 |
| A5 | Motor 4.0 Turbo | 2019 | 0.0 | True | ['Câmbio automático', 'Câmera de estacionament... | 56445.20 | 0.0 |
| J5 | Motor V6 | 2019 | 0.0 | True | ['Sensor crepuscular', 'Painel digital', 'Roda... | 53183.38 | 0.0 |
| A3 | Motor 1.0 8v | 2019 | 0.0 | True | ['4 X 4', 'Piloto automático', 'Central multim... | 88552.39 | 0.0 |
| Série 1 M | Motor V8 | 2019 | 0.0 | True | ['Controle de estabilidade', 'Central multimíd... | 94564.40 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| Lamborghini Reventón | Motor 4.0 Turbo | 2019 | 0.0 | True | ['Controle de tração', 'Ar condicionado', 'Cen... | 67664.86 | 0.0 |
| Benni Mini | Motor V8 | 2019 | 0.0 | True | ['Sensor crepuscular', 'Câmbio automático', 'C... | 126247.84 | 0.0 |
| Uno | Motor Diesel V6 | 2019 | 0.0 | True | ['Central multimídia', 'Sensor crepuscular', '... | 128852.21 | 0.0 |
| Santa Fe | Motor 3.0 32v | 2019 | 0.0 | True | ['Travas elétricas', 'Ar condicionado', '4 X 4... | 129415.33 | 0.0 |
| XC60 | Motor 4.0 Turbo | 2019 | 0.0 | True | ['Painel digital', 'Piloto automático', 'Centr... | 77675.79 | 0.0 |

Agora se fizermos algum modelo, os veículos 0 km não serão eliminados da amostra, já que não estão mais com valores nulos. Antes de continuarmos, importaremos novamente o conjunto db.csv de modo a retornarmos ao seu estado original.

dataset = pd.read\_csv('db.csv', sep = ';')COPIAR CÓDIGO

Fizemos isso para demonstrar outra ferramenta do Pandas que nos ajuda a eliminar os valores nulos de um dataframe, chamada dropna(). Chamaremos essa função a partir do nosso dataset e passaremos como parâmetro subset, que recebe a variável na qual queremos queremos buscar os valores nulos que serão eliminados.

dataset.dropna(subset = ['Quilometragem'])COPIAR CÓDIGO

É necessário prestar atenção, pois todos os registros que contêm esse valor serão eliminados completamente (ou seja, toda a linha será removida). Da mesma forma que na função fillna(), podemos usar o parâmetro inplace = True para que a alteração seja permanente.

dataset.dropna(subset = ['Quilometragem'], inplace = True)COPIAR CÓDIGO

|  | **Nome** | **Motor** | **Ano** | **Quilometragem** | **Zero\_km** | **Acessórios** | **Valor** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | Jetta Variant | Motor 4.0 Turbo | 2003 | 44410.0 | False | ['Rodas de liga', 'Travas elétricas', 'Piloto ... | 88078.64 |
| 1 | Passat | Motor Diesel | 1991 | 5712.0 | False | ['Central multimídia', 'Teto panorâmico', 'Fre... | 106161.94 |
| 2 | Crossfox | Motor Diesel V8 | 1990 | 37123.0 | False | ['Piloto automático', 'Controle de estabilidad... | 72832.16 |
| 4 | Aston Martin DB4 | Motor 2.4 Turbo | 2006 | 25757.0 | False | ['Rodas de liga', '4 X 4', 'Central multimídia... | 92612.10 |
| 5 | Palio Weekend | Motor 1.8 16v | 2012 | 10728.0 | False | ['Sensor de estacionamento', 'Teto panorâmico'... | 97497.73 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 253 | Phantom 2013 | Motor V8 | 2014 | 27505.0 | False | ['Controle de estabilidade', 'Piloto automátic... | 51759.58 |
| 254 | Cadillac Ciel concept | Motor V8 | 1991 | 29981.0 | False | ['Bancos de couro', 'Painel digital', 'Sensor ... | 51667.06 |
| 255 | Classe GLK | Motor 5.0 V8 Bi-Turbo | 2002 | 52637.0 | False | ['Rodas de liga', 'Controle de tração', 'Câmbi... | 68934.03 |
| 256 | Aston Martin DB5 | Motor Diesel | 1996 | 7685.0 | False | ['Ar condicionado', '4 X 4', 'Câmbio automátic... | 122110.90 |
| 257 | Macan | Motor Diesel V6 | 1992 | 50188.0 | False | ['Central multimídia', 'Teto panorâmico', 'Vid... | 90381.47 |

Feito isso, nosso conjunto passará a ter somente 197 registros, diferente dos 258 anteriores. Aprendemos então duas abordagens diferentes para lidarmos com dados problemáticos nos nossos conjuntos: preencher os registros com alguma informação que faça mais sentido ou seja útil, ou simplesmente eliminá-los.

Assim encerramos nossa introdução ao Pandas. Agora você tem diversas ferramentas para seus projetos de ciência de dados, e pode continuar os seus estudos em cursos mais avançados de Pandas, Python, modelagens, dentre outros assuntos. No próximo vídeo faremos uma recapitulação do conteúdo abordado nesse treinamento.