Nomes:
Lucas Mesquita de Souza - 02211044
Enzo Licatalosi de Godoy - 02211012
Giovana Rodrigues do Nascimento - 02211023
Erik Silva Pacheco - 02211013

Open Lab - Preparação Base de dados

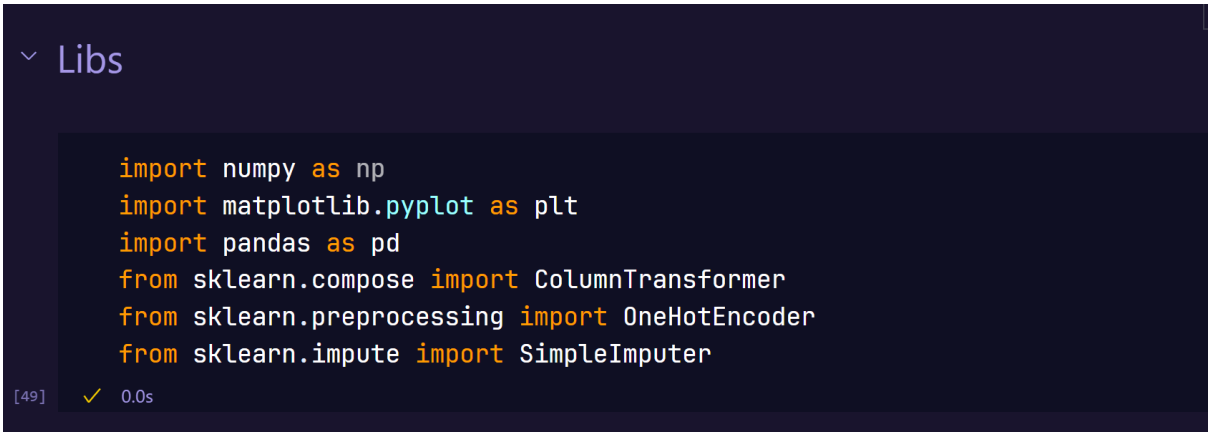
Nesse relatório iremos passar ponto a ponto sobre o processo de tratamento de dados feita na base “seguros.csv”.

Objetivo

Tratar os dados que serão utilizados pela ML pois é uma etapa crucial para garantir a qualidade, a relevância e a segurança dos dados utilizados no treinamento e na avaliação dos modelos de ML, assim trazendo uma precisão e aprendizado melhor para a ML.

Tecnologias Utilizadas

Para fazer o tratamento dos dados, usamos essas bibliotecas:



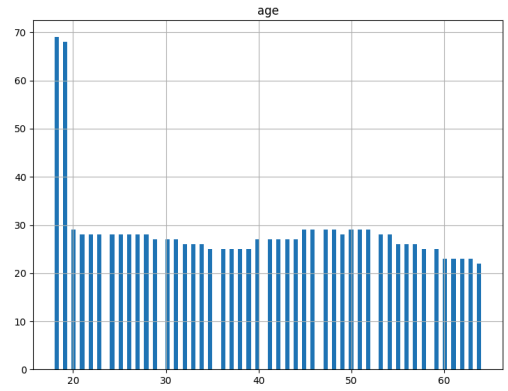
Fluxo de Trabalho

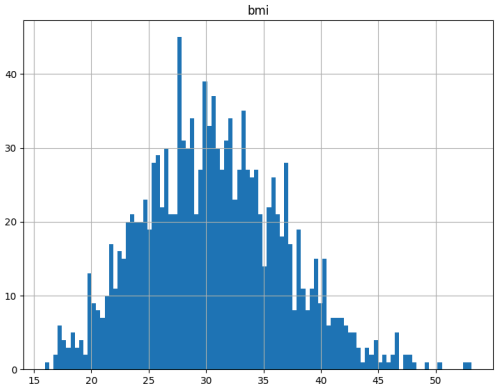
A base “seguros.csv” contém as seguintes informações

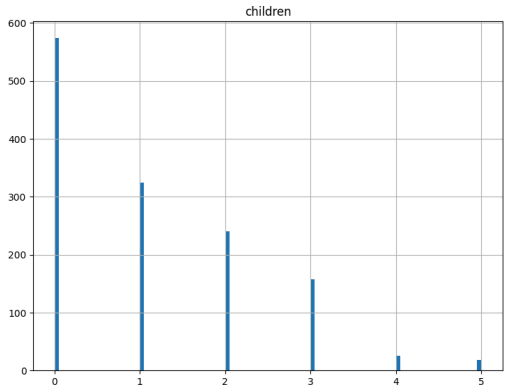
Nome Coluna	Tipo	Linhas preenchidas	Amostra
age	int64	1338	19
sex	object	1333	female
bmi	float64	1333	27.900
children	int64	1338	0
smoker	object	1338	yes

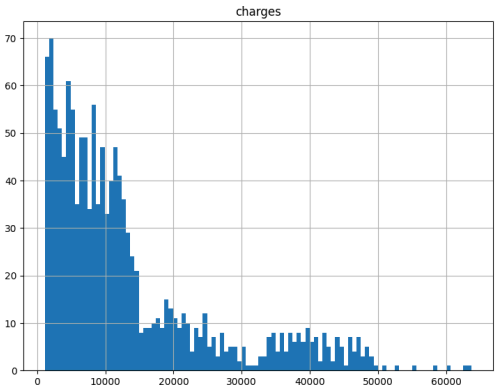
region	object	1338	southwest
charges	float64	1338	16884.92400

Como primeiro passo, plotamos os dados a fim de entender a distribuição base de cada coluna, por exemplo o gráfico de idade nos mostra um equilíbrio da distribuição de frequência de idade.





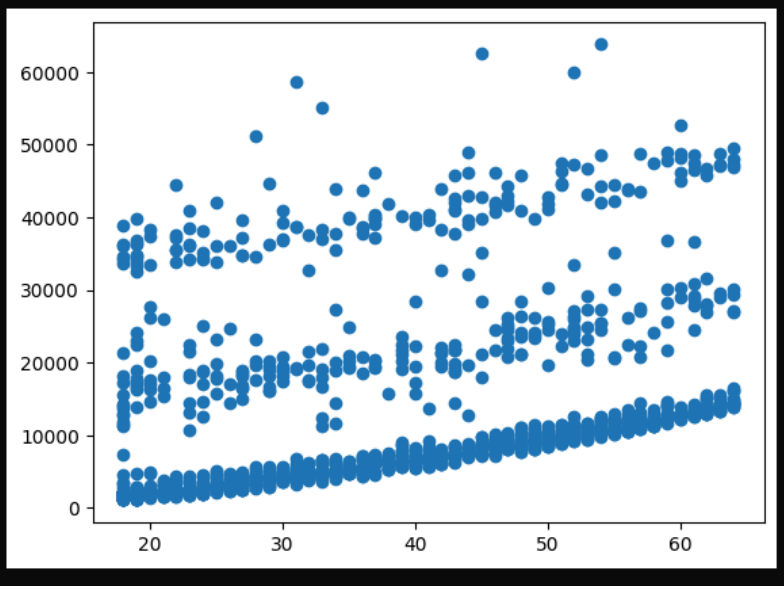




Tentamos encontrar padrões por meio da possível correlação da idade com o salário, que visualmente trouxe 3 conjuntos que podem ser usados para análise posterior

```
fig, ax = plt.subplots()
ax.scatter(df_charges['age'],df_charges['charges'],)
```

```
<matplotlib.collections.PathCollection at 0x1aad7c90f50>
```



Posteriormente, analisamos o datatype de cada coluna. Mas não foi necessária nenhuma conversão de tipagem.

Logo após, identificamos que as colunas ‘bmi’ e ‘sex’ possuíam 5 registros nulos. Depois identificamos esses registros:

```
Mostrando os registros aonde a coluna bmi.isna()
```

	age	sex	bmi	children	smoker	region	charges
25	59	female	NaN	3	no	southeast	14001.13380
441	33	female	NaN	0	yes	southwest	37079.37200
761	23	male	NaN	1	no	southwest	2416.95500
1113	28	female	NaN	3	no	northwest	5312.16985
1282	18	female	NaN	0	yes	northeast	14283.45940

```
Mostrando os registros aonde a coluna sex.isna()
```

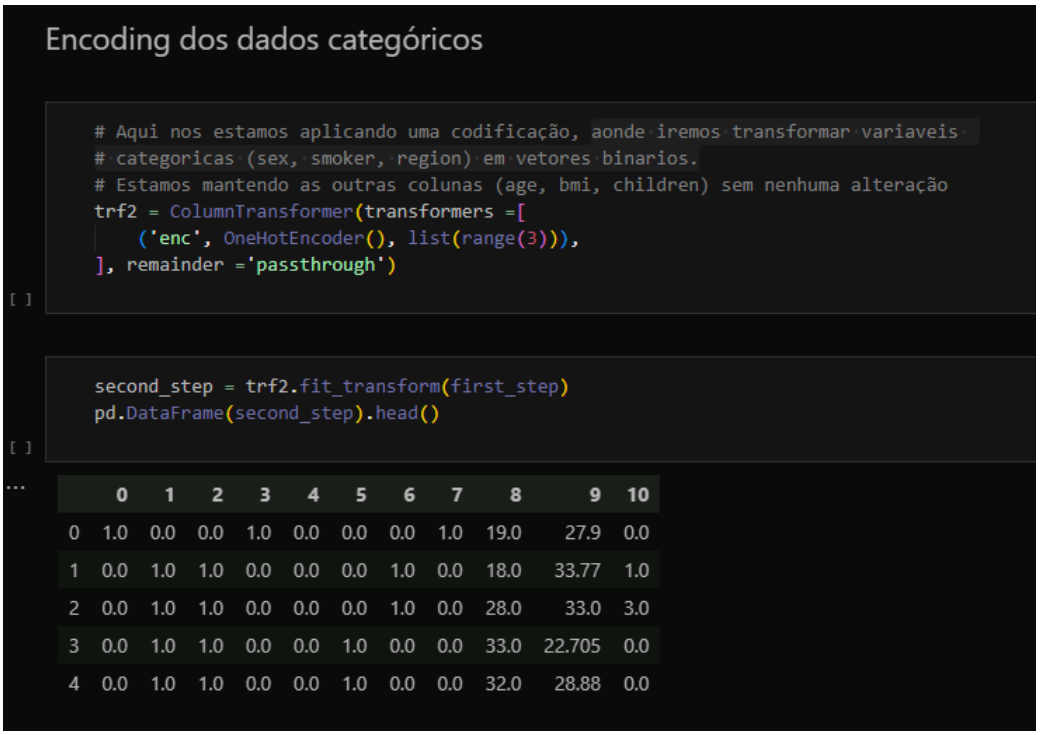
	age	sex	bmi	children	smoker	region	charges
5	31	NaN	25.740	0	no	southeast	3756.62160
141	26	NaN	32.490	1	no	northeast	3490.54910
461	42	NaN	30.000	0	yes	southwest	22144.03200
845	60	NaN	32.450	0	yes	southeast	45008.95550
1324	31	NaN	25.935	1	no	northwest	4239.89265

Antes de iniciarmos a tratativa dos dados, segmentamos as colunas em 2 partes. Sendo **y** a coluna target (charges) e **x** as colunas independentes.

Também reordenamos as colunas, a fim de agrupar inicialmente as colunas que possuem valores categóricos e posteriormente as colunas que possuem valores numéricos.

Durante a reordenação das colunas, aproveitamos para substituir os valores nulos. Para as colunas com valores categóricos o valor preenchido foi igual ao valor mais frequente dentro do DATASET. E para as colunas com valores numéricos foi preenchido pela mediana da coluna.

Após as alterações comentadas, transformamos as variáveis categóricas em vetores binários.



	0	1	2	3	4	5	6	7	8	9	10
0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	19.0	27.9	0.0
1	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	18.0	33.77	1.0
2	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	28.0	33.0	3.0
3	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	33.0	22.705	0.0
4	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	32.0	28.88	0.0