Dados Principais

Rafael Alves Cardoso - RM 360124

Silvio Cezer Saczuck - RM 360204

Luciano Giles Soares - RM 359834

William Judice Yizima - RM 360214

Luiz Ricardo Zinsly Calmon - RM359894

GitHub do TCC1 : https://github.com/WilliamYizima/POS-4IADT

Vídeo da apresentação : https://youtu.be/s73Bl4_o6Oc

Escopo

Você é um(a) profissional encarregado(a) de desenvolver um modelo preditivo de regressão para prever o valor dos custos médicos individuais cobrados pelo seguro de saúde.

Dataset

https://osf.io/7u5gy

Variáveis

Variáveis do dataset:

- 'idade'= Idade do segurado
- 'sexo'= Sexo do segurado (Feminino ou Masculino)
- 'imc'= Indice de Massa Corporal - Pela OMS -> 18.5 até 24.9
- 'filhos'= Quantidade de filhos (5 ou + = 5)
- 'fumante'= Se é fumante. Não possui tempo de consumo nem quantidade.
- 'regiao'= Como o atendimento era central, foi dividido em 5 regiões.
- 'encargos'= Valor, em dólares americanos dos encargos anuais
- 'amigos'= Número de pessoas com as quais se relaciona amistosamente.
- 'inimigos'= Número de pessoas do convivio geradoras de stress
- 'chips'= Consumo de produtos industrializados
- 'fritas'= Consumo de frituras e saturados

- 'miojo'= Consumos de refeições rápidas
- 'comportamento'= Modelo adotado de vida (stress, sono, excesso, etc)

```
esse dataset tem 1338 linhas e 13 colunas
```

Exploração de Dados

- Carregamento da base de dados
- Exploração das características;
- Analise estatísticas descritivas e visualize distribuições relevantes.

Amostra inicial dos dados

Out[2]:

	idade	sexo	imc	filhos	fumante	regiao	encargos	amigos	in
0	19	feminino	NaN	0	sim	sudoeste	16884.92400	5	
1	18	masculino	33.770	1	nao	sudeste	1725.55230	0	
2	28	masculino	33.000	3	nao	sudeste	4449.46200	9	
3	33	masculino	22.705	0	nao	noroeste	21984.47061	5	
4	32	masculino	28.880	0	nao	noroeste	3866.85520	4	

Informação sobre o dataset

```
esse dataset tem 1338 linhas e 13 colunas
```

Na imagem abaixo é possível visualizar as colunas e quais são os tipos de dados armazenados.

Um comentário interessante é que temos algumas variáveis categóricas que possivelmente deverão ser transformada para que as 'features'(x) sejam aceitas pelos modelos escolhidos.

Por exemplo, região(localidade) e fumante(possivelmente será transformado em 0 e 1).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   idade          1338 non-null   int64
 1   sexo           1338 non-null   object
 2   imc            1327 non-null   float64
 3   filhos         1338 non-null   int64
 4   fumante        1338 non-null   object
 5   regiao         1338 non-null   object
 6   encargos       1338 non-null   float64
 7   amigos         1338 non-null   int64
 8   inimigos       1338 non-null   int64
 9   chips          1338 non-null   int64
 10  fritas         1338 non-null   int64
 11  miojo          1338 non-null   int64
 12  comportamento  1338 non-null   int64
dtypes: float64(2), int64(8), object(3)
memory usage: 136.0+ KB
```

Abaixo é possível visualizar as análises descritivas estatísticas do conjunto.

Essa visão overview é interessante, pois é possível visualizar alguns pontos de atenção, como:

- idade negativa (uma idade menor que 0 é um absurdo para nosso contexto)
- idade maior que 110 anos(pode ser considerado um erro de preenchimento ou um absurdo para nosso contexto)
- IMC é 0 (um cálculo de altura e peso que traz o valor 0, não existe)

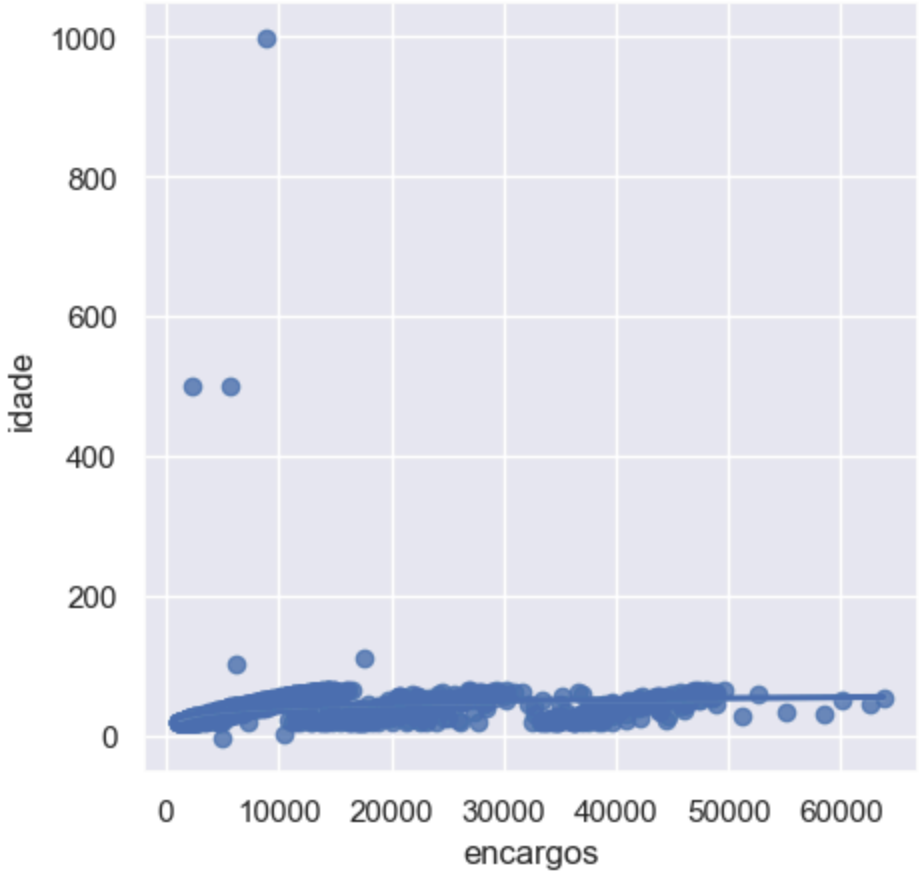
Out[5]:

	idade	imc	filhos	encargos	amigos	i
count	1338.000000	1327.000000	1338.000000	1338.000000	1338.000000	1338
mean	40.670404	30.667841	1.094918	13270.422265	4.933483	4
std	34.784427	6.103216	1.205493	12110.011237	3.198855	3
min	-3.000000	15.960000	0.000000	1121.873900	0.000000	0
25%	26.250000	26.302500	0.000000	4740.287150	2.000000	2
50%	39.000000	30.400000	1.000000	9382.033000	5.000000	5
75%	51.000000	34.687500	2.000000	16639.912515	8.000000	8
max	999.000000	53.130000	5.000000	63770.428010	10.000000	10

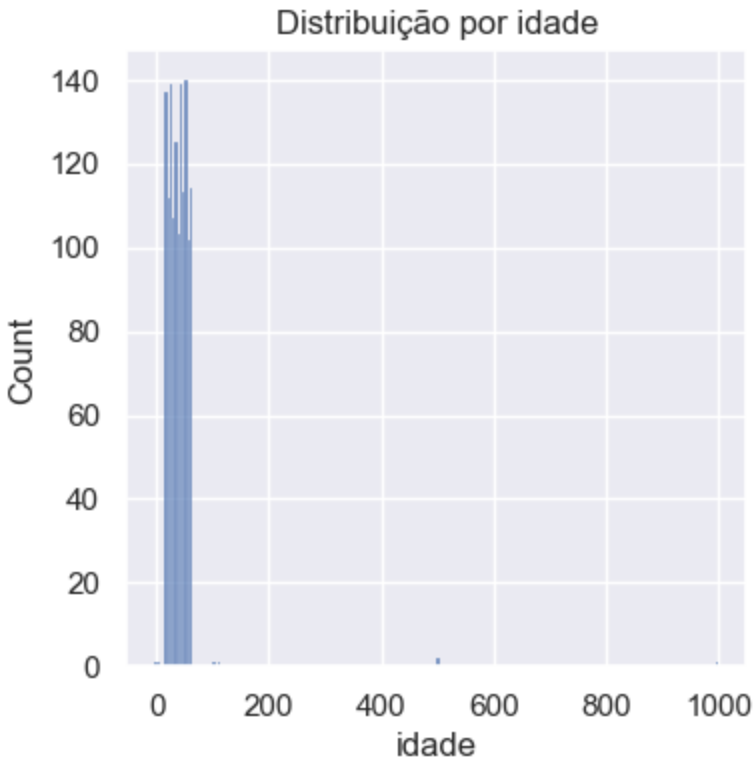
Visualização de dados

Abaixo a visualização gráfica de dados inicial conjunto (posteriormente iremos tratar os dados e ajustando alguns pontos)

Encargos x Idade

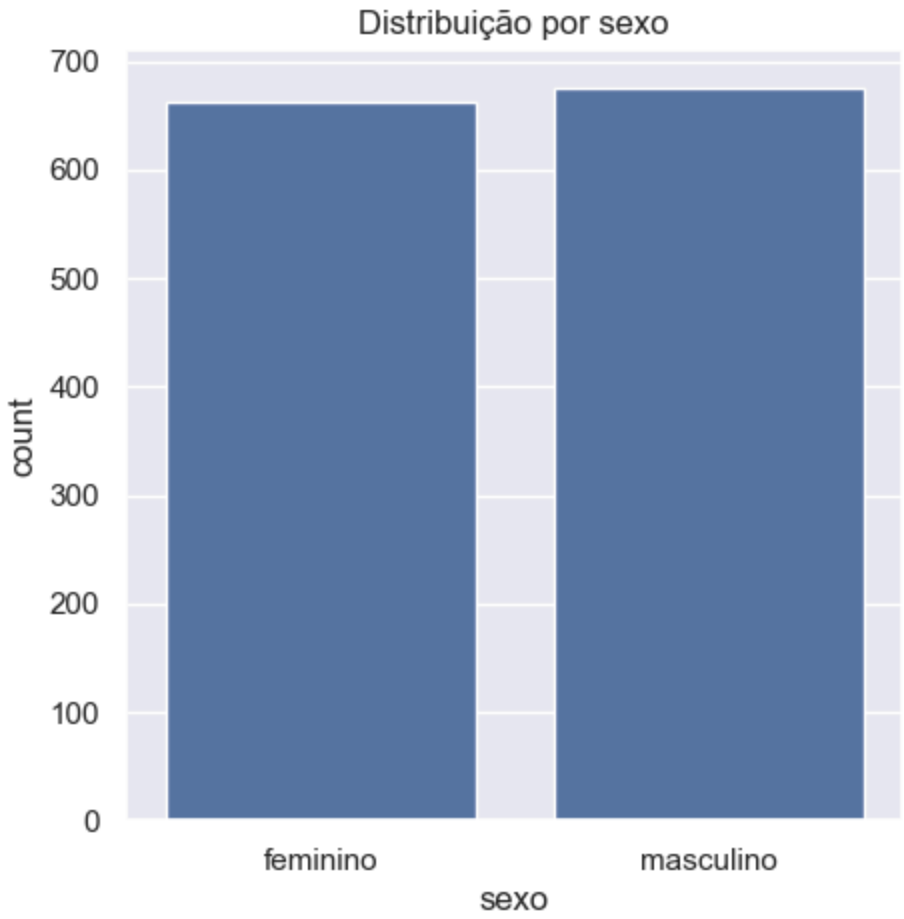


Idade

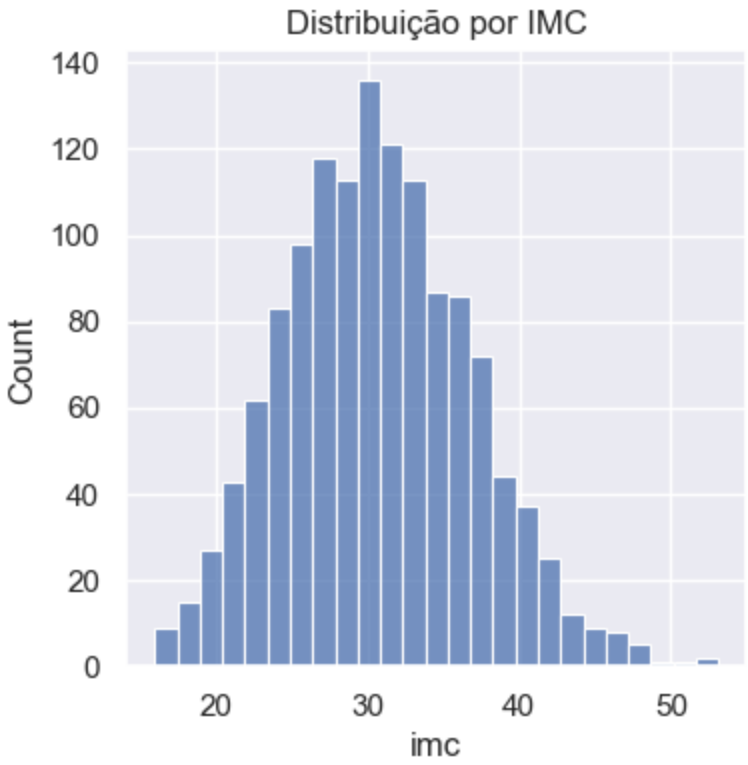


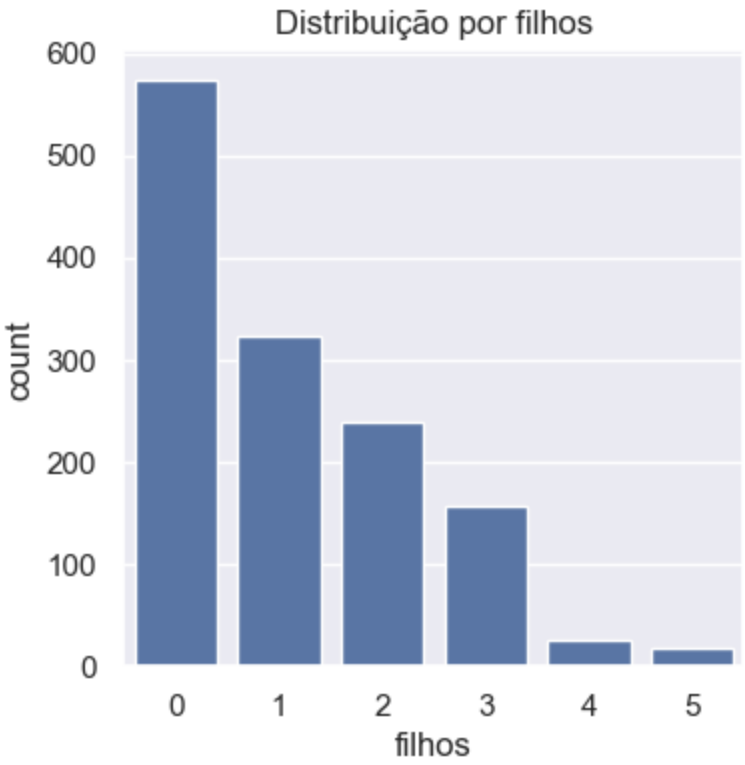
sexo

Out[9]:
```
masculino    676
feminino     662
Name: sexo, dtype: int64
```


Distribuição por sexo
700
600
500
400
300
200
100
0
count
feminino
masculino
sexo

IMC


Distribuição por IMC
140
120
100
80
60
40
20
0
Count
20
30
40
50
imc

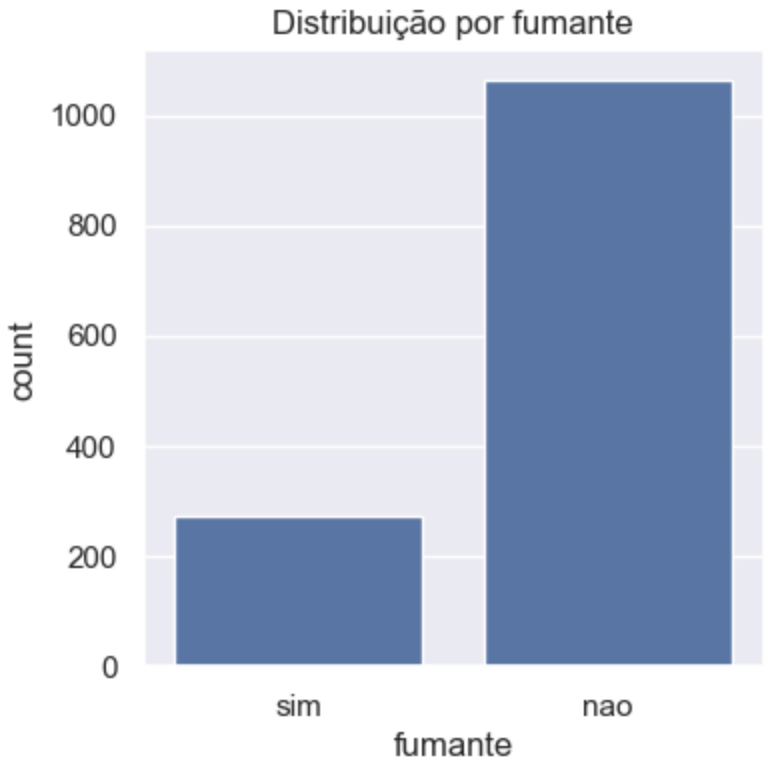


Filhos

Out[13]:
```
0    574
1    324
2    240
3    157
4     25
5     18
Name: filhos, dtype: int64
```

Fumantes

Out[14]:
```
nao    1064
sim     274
Name: fumante, dtype: int64
```

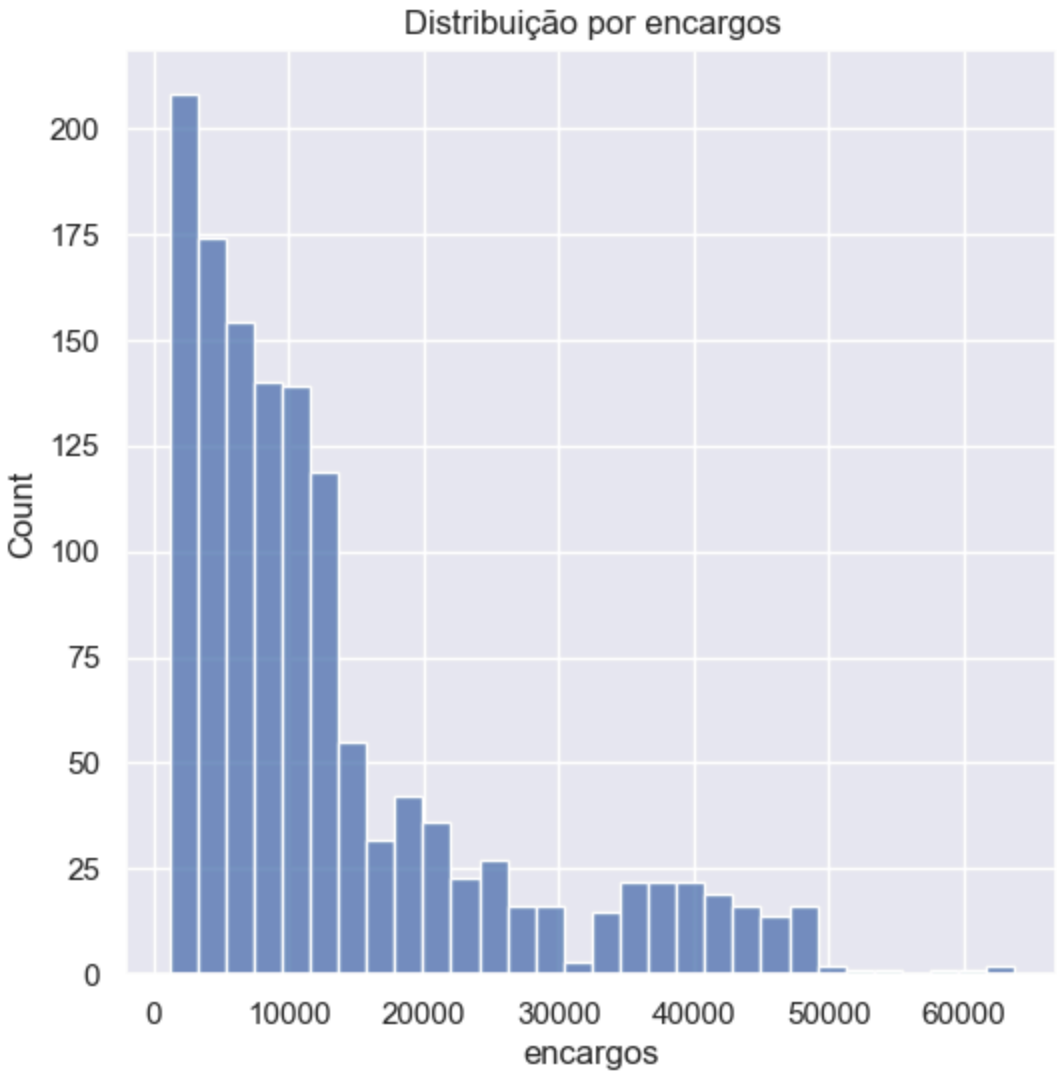


Região

Out[16]:
```
sudeste     360
sudoeste    324
noroeste    324
nordeste    320
centro       10
Name: regiao, dtype: int64
```


Distribuição por região
350
300
250
200
150
100
50
0
count
sudoeste
sudeste
noroeste
nordeste
centro
regiao

Encargos



Pré-Processamento de dados

- Realização de limpeza
- Visualização da correlação
- Transformação de variáveis categóricas em formatos para a modelagem

Correlação

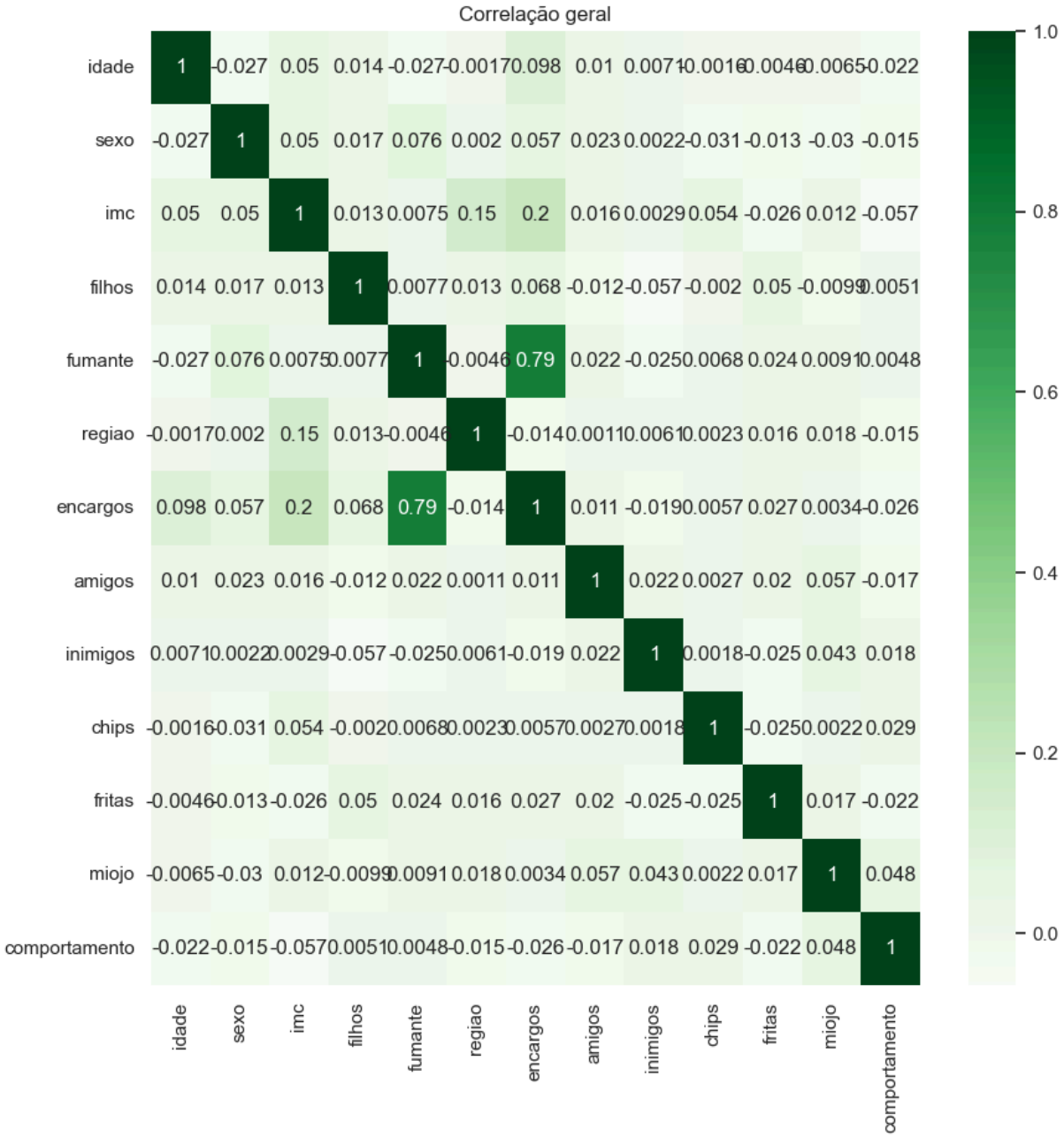
Mapa de correlação inicial (sem tratamento de dados).

Nossa variável target (y) são os Encargos, portanto, queremos saber quais variáveis do nosso dataset (x) trazem maior mudança para os Encargos.

Visualizando o gráfico de correlação abaixo, podemos dizer que:

- fumante-> 0.79

- imc ->0.2 Essas variáveis são as que mais alteram com uma relação positiva os Encargos (quanto mais fumante e maior o IMC, maior o encargo)



Limpeza e tratamento de dados

Como visto anteriormente, existem dados nulos em IMC que não fazem sentido para nós. Existem dados em idades que também não fazem sentido Existem algumas variáveis categóricas que não serão possíveis inserir no modelo (ele não aceita 'strings' para fazer o cálculo). Portanto será necessário:

- retirar valores nulos
- retirar 'outliers'

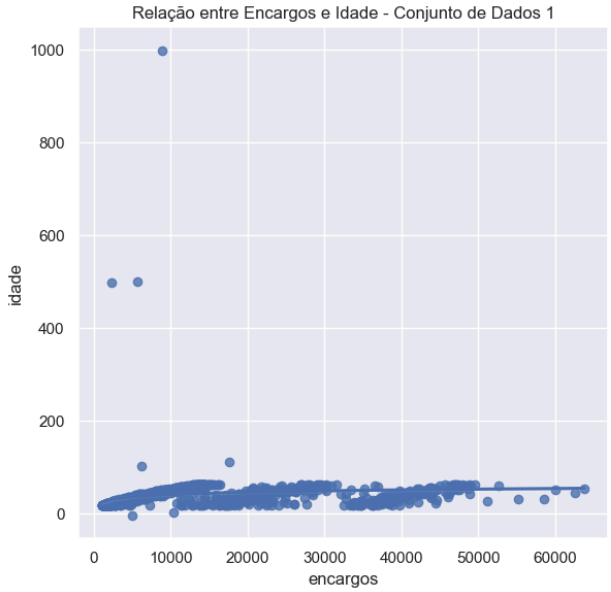
- transformar algumas variáveis categóricas importantes (como região, conforme visto acima) para que seja possível inserir no modelo

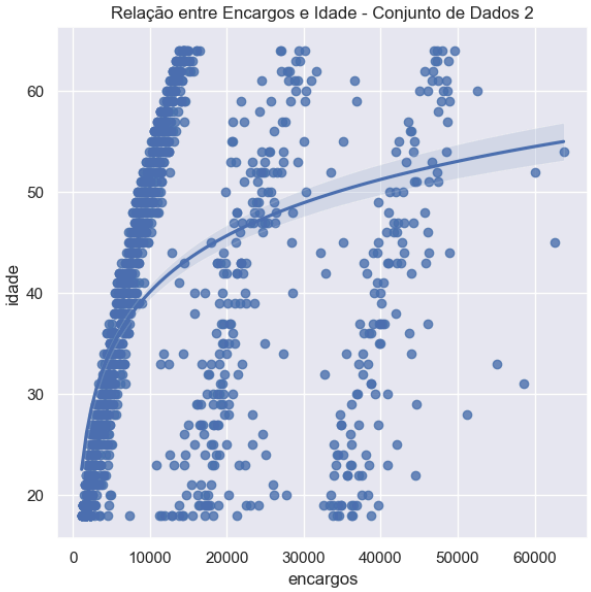
Para facilitar a visualização, será mostrado os dados antes e depois dessa transformação.

A idéia dessa é etapa é que o modelo consiga explicar nossa base de dados de forma mais genérica, e não se baseando em outliers

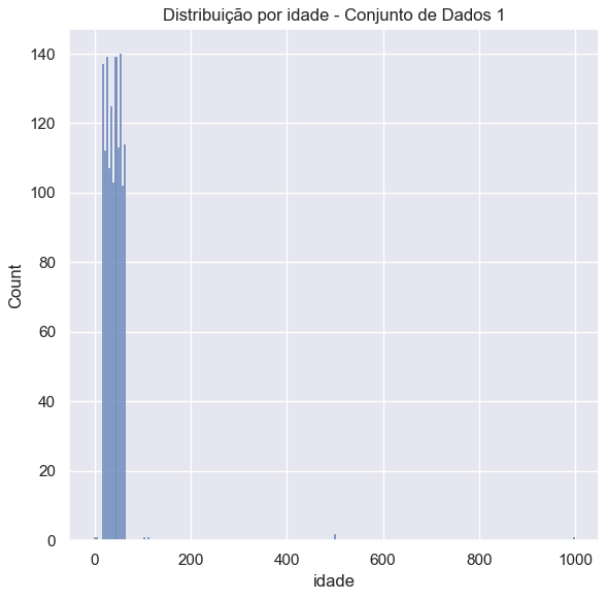
Encargos x Idade

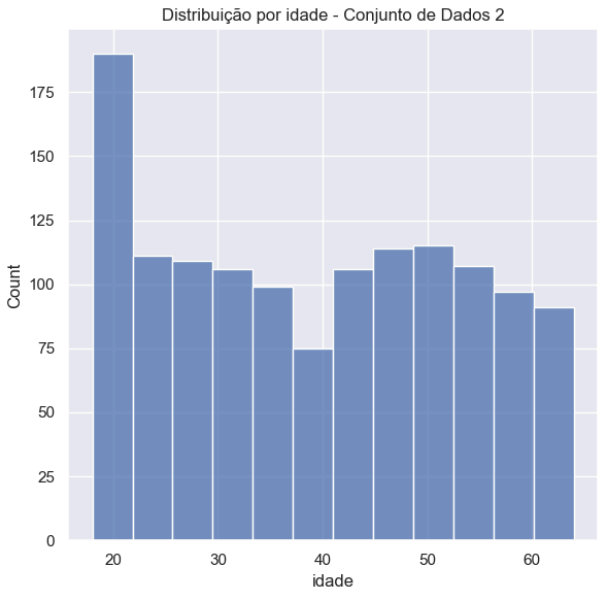
É possível ver a diferença do 'esticamento' de dados e sua distribuição.



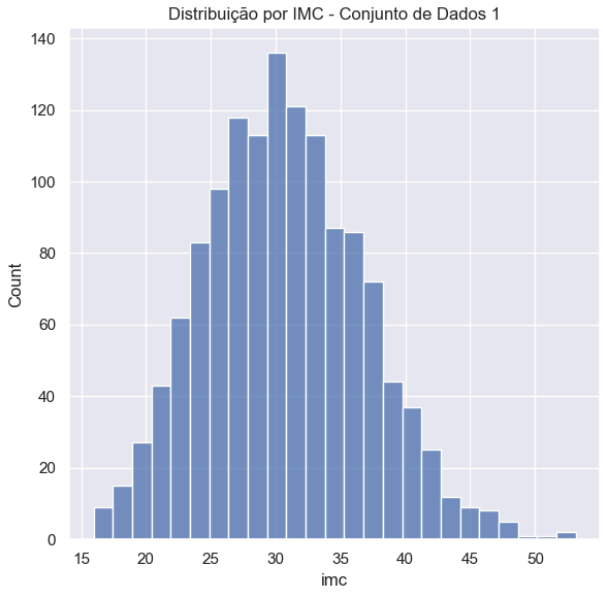


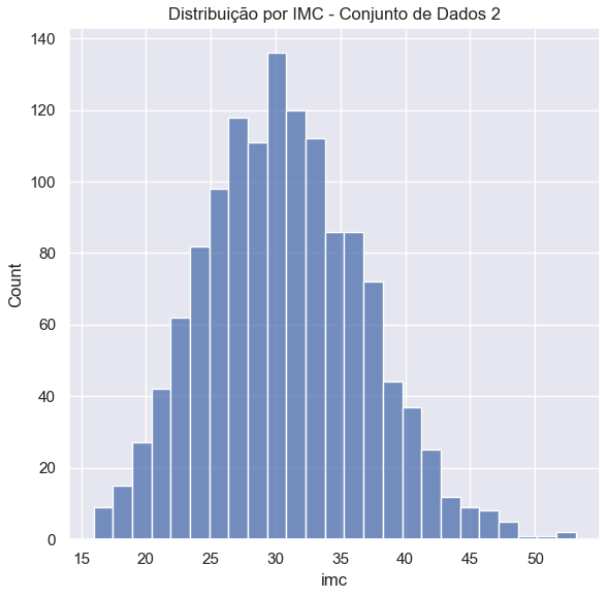
Idade





IMC



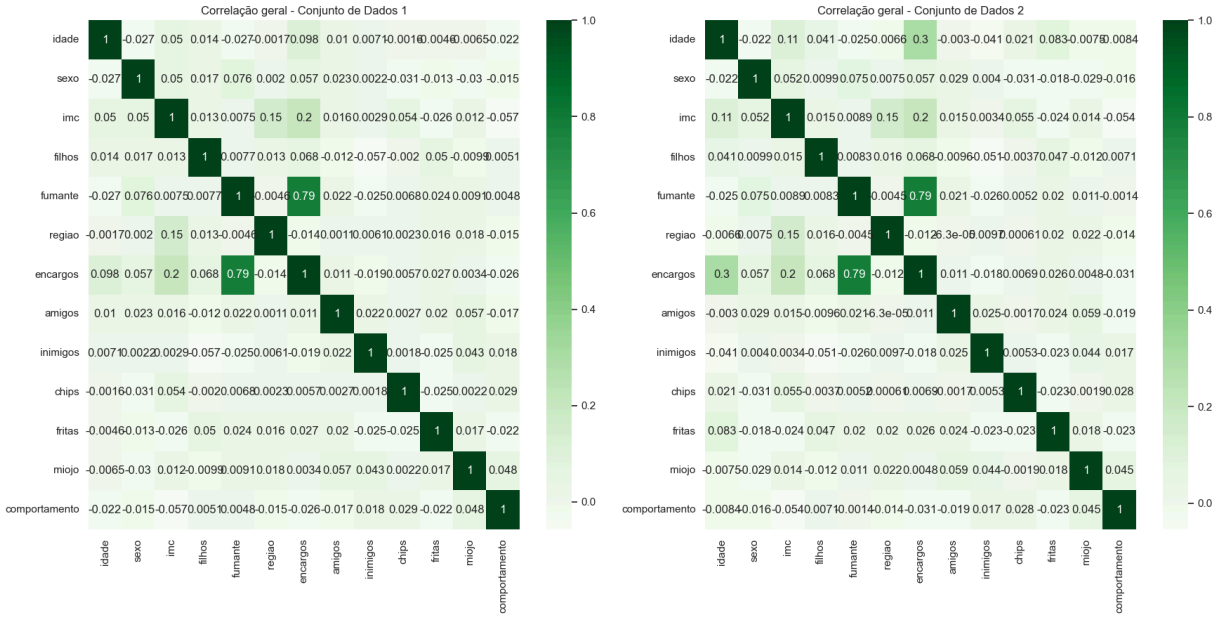


`Utilizando LabelEncoder para transformar as variáveis categóricas em "número s aceito pelo modelo", legenda:`

Mapa de Correlação

Com a limpeza e tratamento de dados já feito, podemos validar novamente o mapa de correlação.

Veja que foi alterado uma feature (idade), que é de extrema importância. Assim como Região e Fumante, a idade impacta de forma positiva com um peso alto (0.3) em relação aos encargos



Modelagem

- Crie um modelo preditivo de regressão utilizando uma técnica à sua escolha (por exemplo: Regressão Linear, Árvores de Decisão etc);
- Divida o conjunto de dados em conjuntos de treinamento e teste.

y -> target, o que queremos buscar = Encargos

x -> features, outras variáveis

Modelo Regressão Linear

Trecho de código que separa treino e teste

```
#separa para treino 80% e teste 20%
X_train, X_test, Y_train, Y_test = train_test_split(X, Y,
test_size=0.2, random_state=2)
```

```
R**2 treino :  0.7635651088920332
R^2 teste :  0.7067024745910114

Resultados iniciais:
```

Out[31]: 'Dados de teste tiveram uma acerto de 71.0% e os dados de treino tiveram 76.0%'

MAE, MSE, RMSE

MAE (Erro Médio Absoluto - Mean Absolute Error)

O Erro Médio Absoluto (MAE) mede a média dos valores absolutos das diferenças entre os valores reais (Y) e os valores previstos (y_pred). Um MAE mais baixo indica que o modelo está fazendo previsões mais próximas dos valores reais.

MSE (Erro Quadrático Médio - Mean Squared Error)

O Erro Quadrático Médio (MSE) calcula a média dos quadrados dos erros (diferenças entre valores reais e previstos). Penaliza erros maiores de forma mais significativa por elevar as diferenças ao quadrado.

RMSE (Raiz do Erro Quadrático Médio - Root Mean Squared Error)

O RMSE é simplesmente a raiz quadrada do MSE. Ele traz o erro para a mesma escala dos valores de Y. Assim como o MSE, mas em uma escala mais fácil de interpretar.

Essas métricas são usadas para avaliar a performance de modelos de regressão. Quanto menores esses valores, melhor o modelo está prevendo os resultados.

Out[33]: 'REGRESSAO LINEAR-> MAE: 4023.41'

Out[34]: 'REGRESSAO LINEAR-> MSE: 34487296.24'

Out[35]: 'REGRESSAO LINEAR-> RMSE: 5872.59 '

Modelo Regressão(Ordinary Least Squares)

Visualização de modelo

Out[39]:

OLS Regression Results

Dep. Variable:	encargos	**R-squared:**	0.752
Model:	OLS	**Adj. R-squared:**	0.750
Method:	Least Squares	**F-statistic:**	330.6
Date:	Mon, 20 Jan 2025	**Prob (F-statistic):**	0.00
Time:	23:29:17	**Log-Likelihood:**	-13367.
No. Observations:	1320	**AIC:**	2.676e+04
Df Residuals:	1307	**BIC:**	2.683e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercepto	-1.043e+04	1239.936	-8.409	0.000	-1.29e+04	-7994.114
idade	258.2526	12.064	21.407	0.000	234.585	281.920
imc	328.2205	28.038	11.706	0.000	273.215	383.226
sexo	-148.1266	336.842	-0.440	0.660	-808.936	512.683
filhos	492.4399	139.345	3.534	0.000	219.076	765.804
fumante	2.393e+04	416.048	57.523	0.000	2.31e+04	2.47e+04
regiao	-347.7333	151.044	-2.302	0.021	-644.048	-51.419
amigos	-25.3049	52.558	-0.481	0.630	-128.412	77.802
inimigos	65.2124	53.527	1.218	0.223	-39.795	170.220
chips	-76.4843	84.604	-0.904	0.366	-242.459	89.490
fritas	-80.2982	85.397	-0.940	0.347	-247.828	87.231
miojo	-11.3392	85.943	-0.132	0.895	-179.941	157.262
comportamento	-118.1083	84.460	-1.398	0.162	-283.800	47.584

Omnibus:	281.659	**Durbin-Watson:**	2.103
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	648.666
Skew:	1.169	**Prob(JB):**	1.39e-141
Kurtosis:	5.516	**Cond. No.**	389.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Out[42]: 'Ordinary Least Squares -> MAE: 4187.230790681157'

Out[43]: 'Ordinary Least Squares -> MSE: 36608518.18116458'

Out[44]: 'Ordinary Least Squares -> RMSE: 6050.497349901459'

MAE

Em média, o modelo erra cerca de 4187.23 unidades em cada previsão. O MAE mede o erro de forma linear, considerando a média dos erros absolutos, o que significa que cada erro tem o mesmo peso.

MSE indica que existem erros grandes no modelo, pois penaliza erros maiores de forma mais severa. RMSE

O RMSE é a raiz do MSE e indica que, em média, o erro do modelo é de aproximadamente 6050.50 unidades

SVR (Support Vector Regressor)

```
Utilizando 'cross val score:'
Scores em cada fold: [-0.11623809 -0.10865375 -0.08022862 -0.10961705 -0.105
69786]
A média
Média do Score: -0.104087071920266
```

O dataset é o já ajustado(retirado outliers, etc)

Foi visualizado que o score é negativo, portanto o modelo não conseguiu prever realmente(abaixo da média)

Out[48]: 'SVR -> MAE: 7700.5256717802085'

Out[49]: 'SVR -> MSE: 151348820.90702936'

Out[50]: 'SVR -> RMSE: 12302.390861415084'

Observações e validação estatística

Observação dos resultados:

O Erro Médio Absoluto (MAE) mede a média dos valores absolutos das diferenças entre os valores reais (Y) e os valores previstos (y_pred).

Um MAE mais baixo indica que o modelo está fazendo previsões mais próximas dos valores reais.

O menor valor é da regressão linear

Out[51]: 'REGRESSAO LINEAR-> MAE: 4023.41'

Out[52]: 'Ordinary Least Squares -> MAE: 4187.230790681157'

Out[53]: 'SVR -> MAE: 7700.5256717802085'

MSE indica que existem erros grandes no modelo, pois penaliza erros maiores de forma mais severa.

Out[54]: 'REGRESSAO LINEAR-> MSE: 34487296.24'

Out[55]: 'Ordinary Least Squares -> MSE: 36608518.18116458'

Out[56]: 'SVR -> MSE: 151348820.90702936'

O RMSE é a raiz do MSE e indica que, em média, o erro do modelo.

Out[57]: 'REGRESSAO LINEAR-> RMSE: 5872.59 '

Out[58]: 'Ordinary Least Squares -> RMSE: 6050.497349901459'

Out[59]: 'SVR -> RMSE: 12302.390861415084'

Se baseando nessas métricas, o modelo de regressão linear se saiu melhor