Análise de custos com cuidados médicos

Vinícius H. P. Cavalcanti

PERGUNTAS

- Gênero, índice de massa corporal (imc), quantidade de filhos, se é fumante ou não explicam os gastos?
- Se explicam, quanto explicam ? Todas as variáveis explicam 100% dos gastos?
- Existe diferença dos custos entre regiões para a variável mais explicativa ?

SOBRE O CONJUNTO DE DADOS

- Foi obtido na plataforma Kaggle
- É utilizado no livro `Machine Learning with R` por Brett Lantz
- Segundo o autor do livro, foi desenvolvido com base no censo demográfico dos EUA
- Possui 1338 observações
- Possui 7 colunas, sendo essas:
    - idade
    - gênero
    - índice de massa corporal
    - quantidade de filhos
    - fumante?
    - região
    - custos

METODOLOGIAS

Estatística descritiva & Análise exploratória

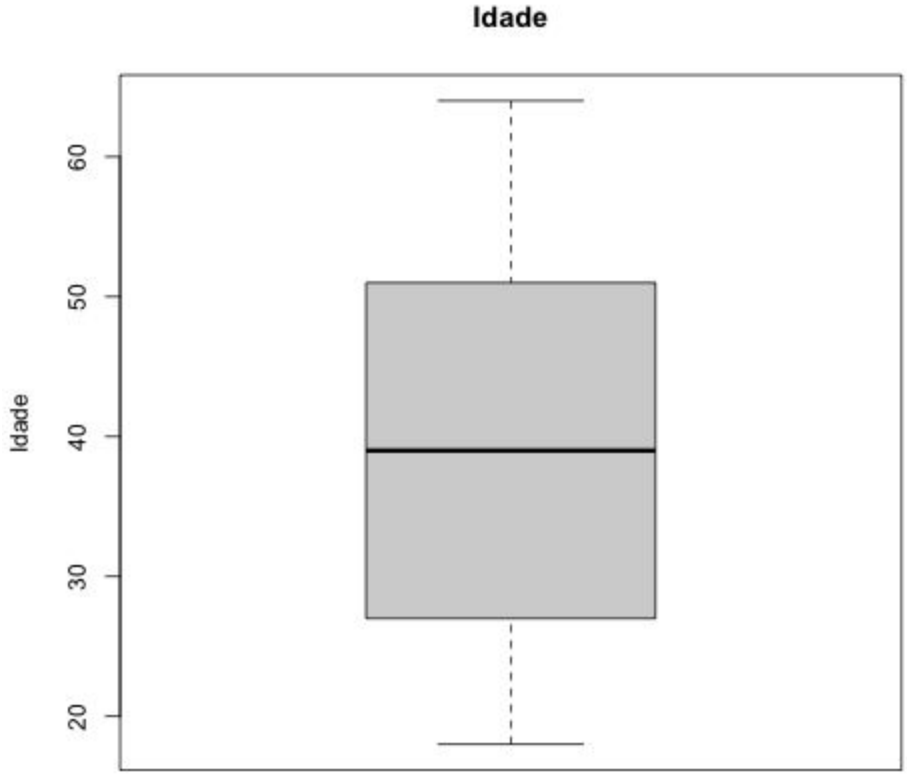
- Quais os tipos de dados presentes (numérico ou categórico) ?
- Para os valores numéricos, qual a média? moda ? mediana? os dados variam?
- Para os valores categóricos, quantos valores distintos existem?
- Existem valores fora da curva?
- Visualmente, qual a relação entre custo e gênero, se é fumante ou nao, região e quantidade de filho?

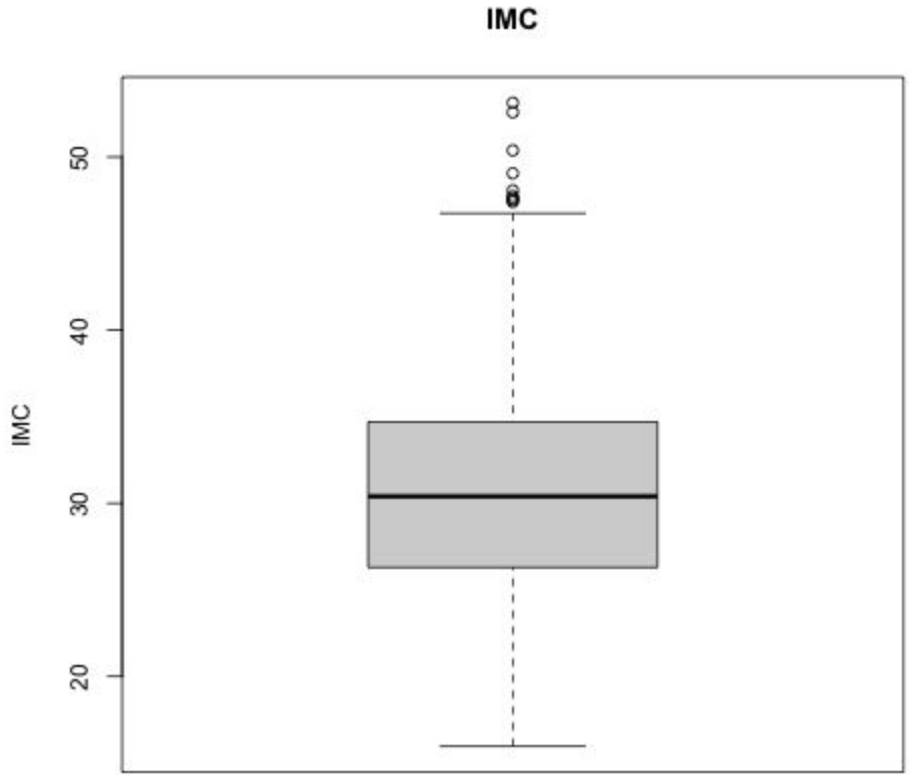
Inferência estatística

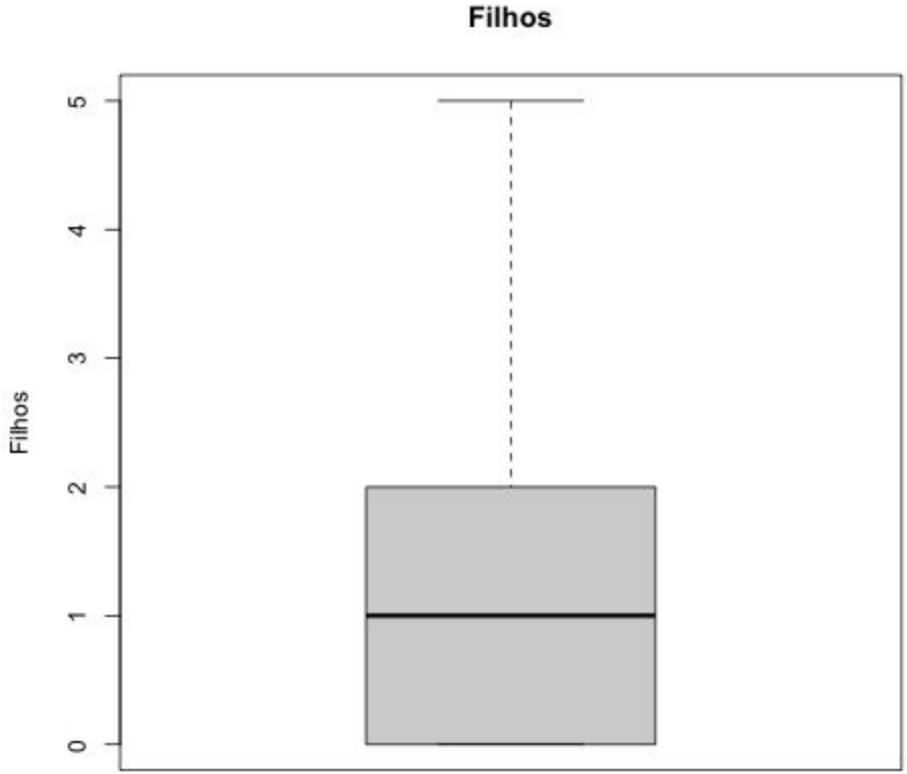
- Regressão linear múltipla
  - as variáveis explicam o custo?
  - qual a variável mais explicativa?
  - as variáveis explicam 100% dos custos?
- Análise de variância
  - entre a variável mais explicativa, existe diferença entre região ou por quantidade de filhos?

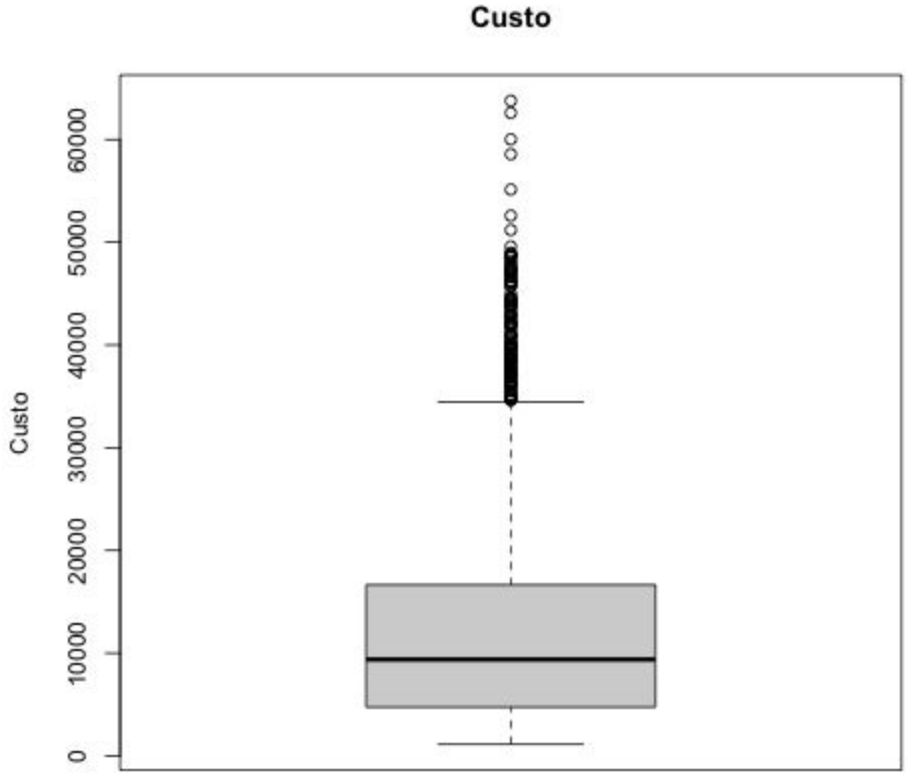
RESULTADOS E ANÁLISE

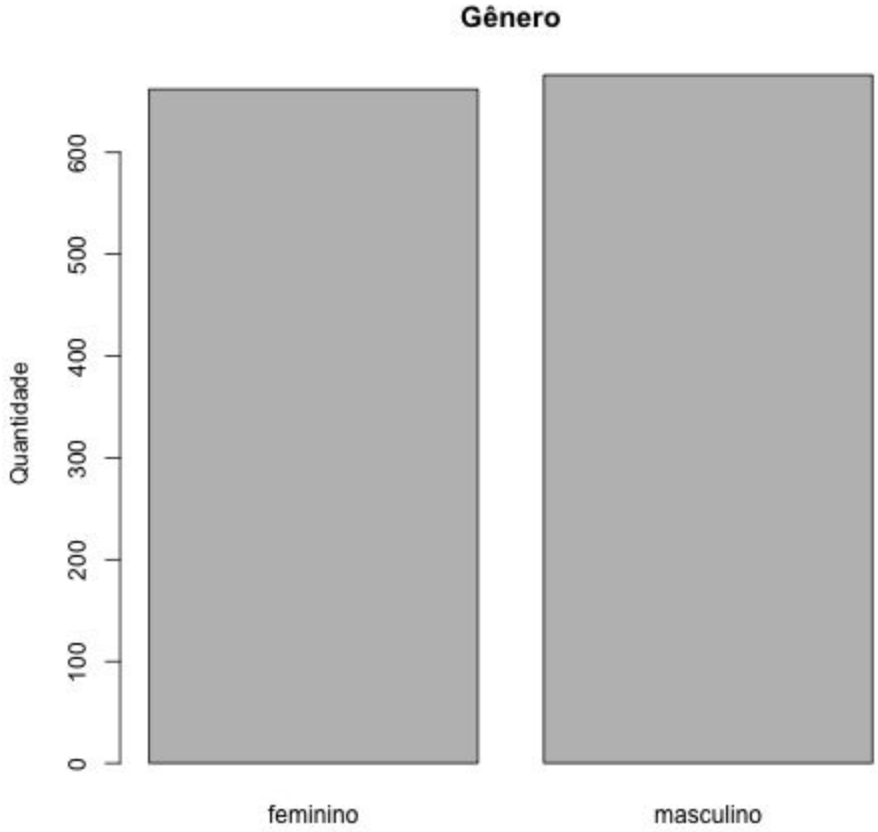
Estatística descritiva

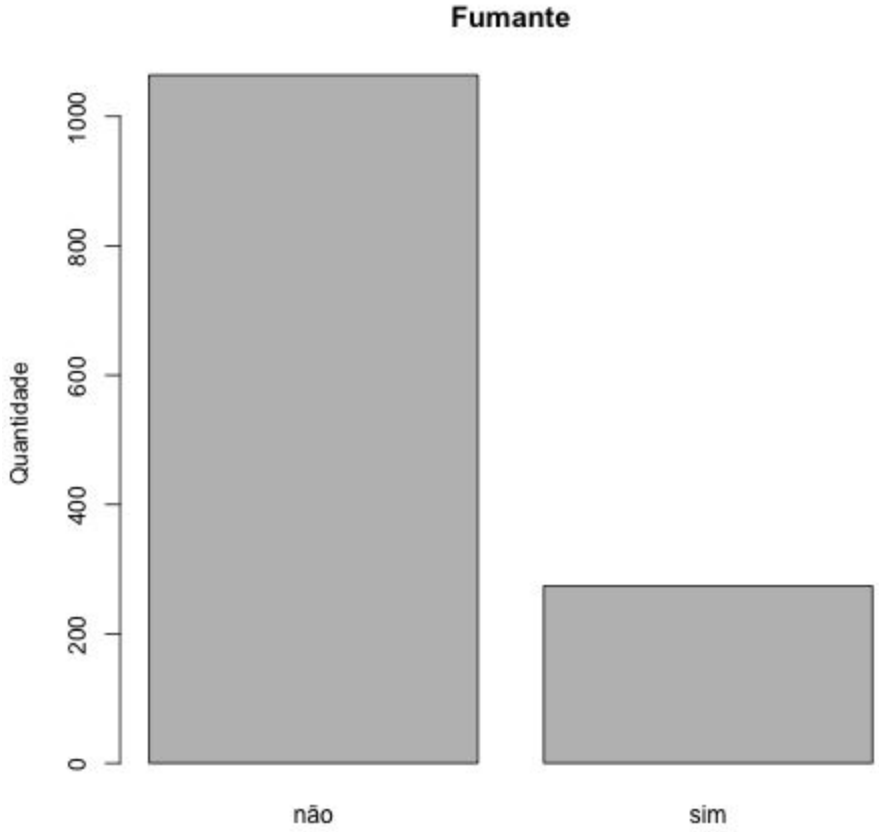

Idade
Idade
20
30
40
50
60

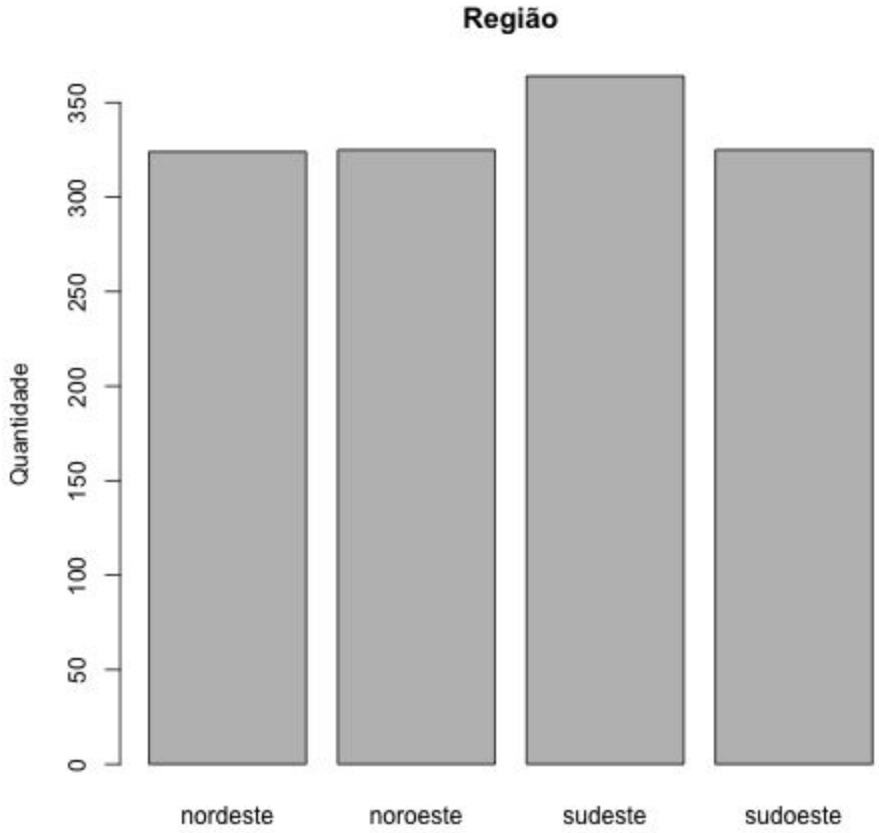

IMC
IMC
50
40
30
20


Filhos
Filhos
0
1
2
3
4
5


Custo
Custo
0
10000
20000
30000
40000
50000
60000


Gênero
Quantidade
0
100
200
300
400
500
600
feminino
masculino

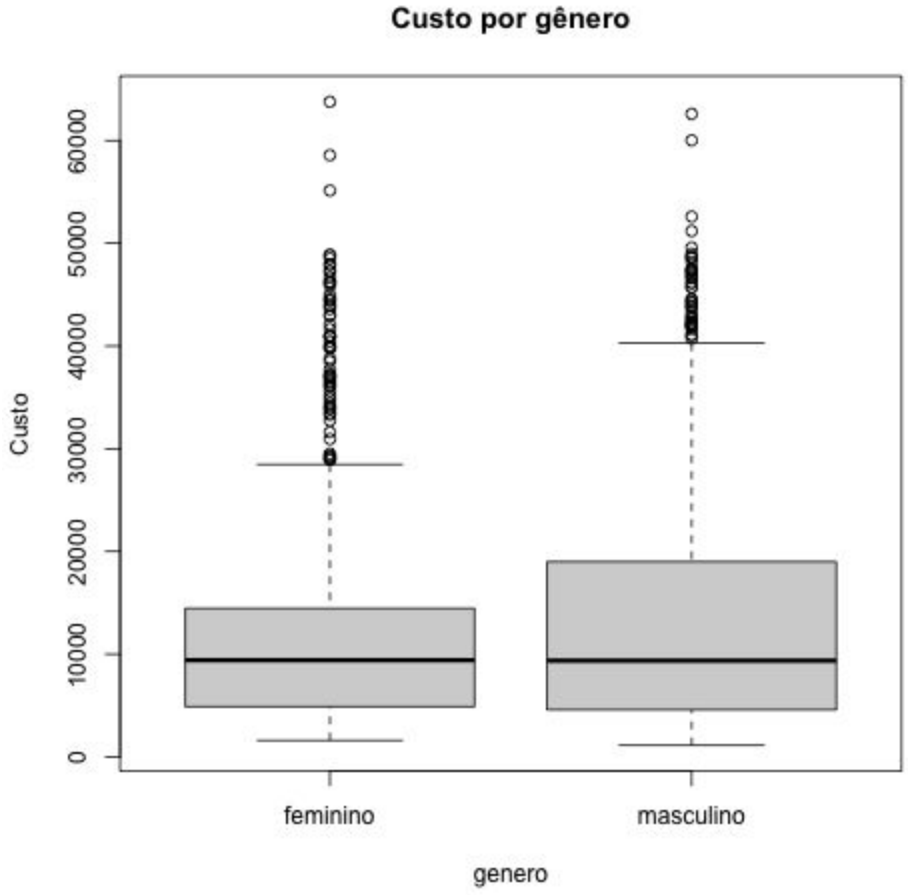

Fumante
Quantidade
0
200
400
600
800
1000
não
sim

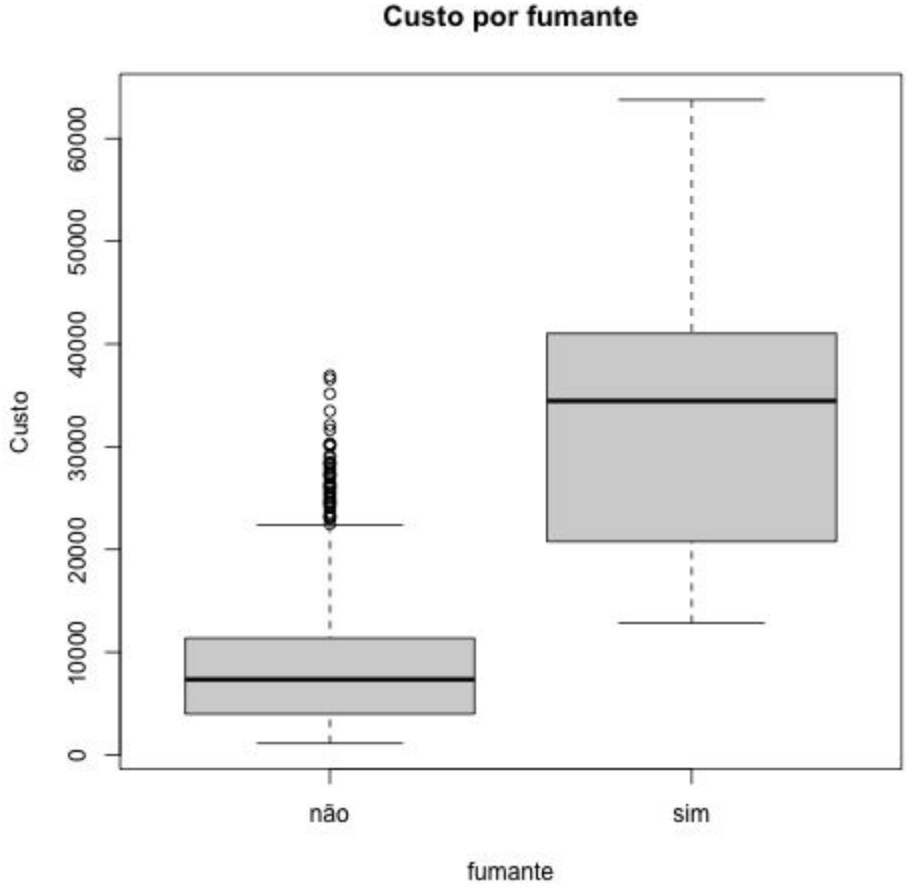

Região
Quantidade
0
50
100
150
200
250
300
350
nordeste
noroeste
sudeste
sudoeste

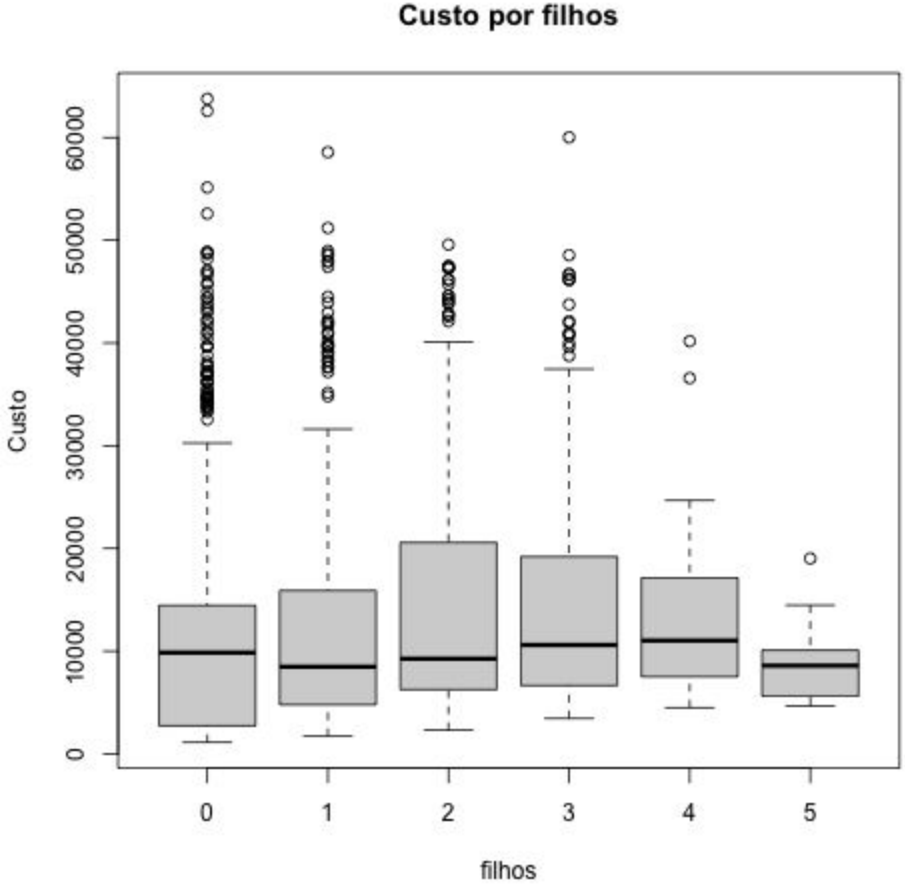
Observações vazias

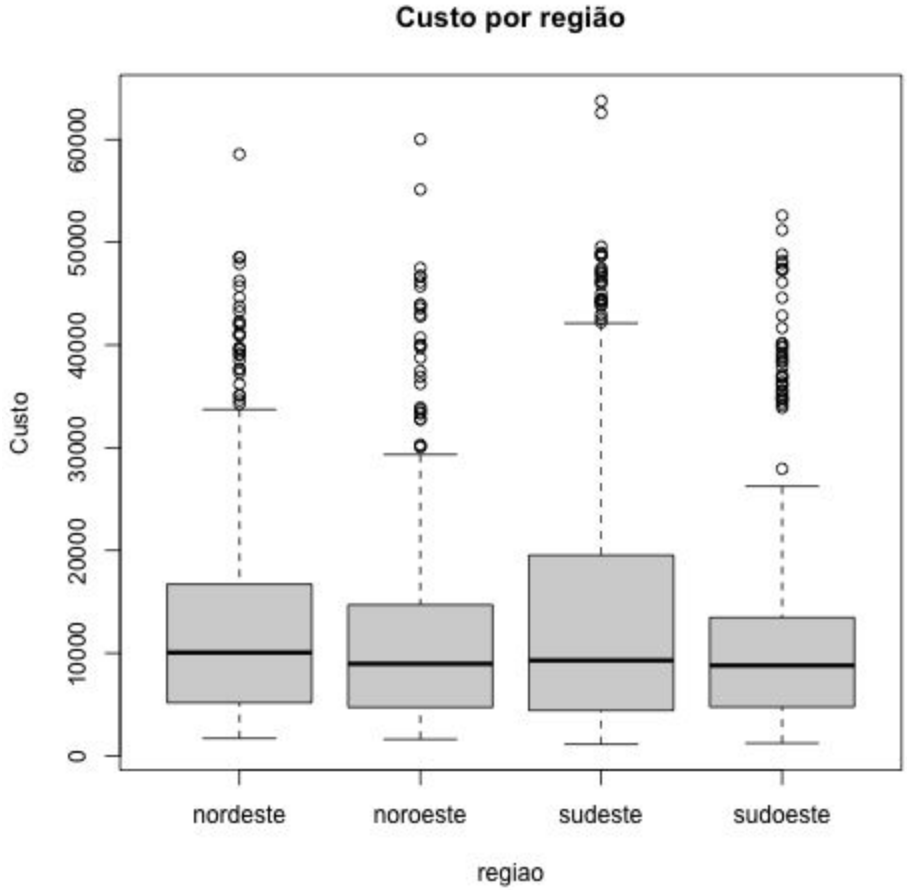
idade	0
gênero	0
imc	0
filhos	0
fumante	0
região	0
custo	0

Análise exploratória

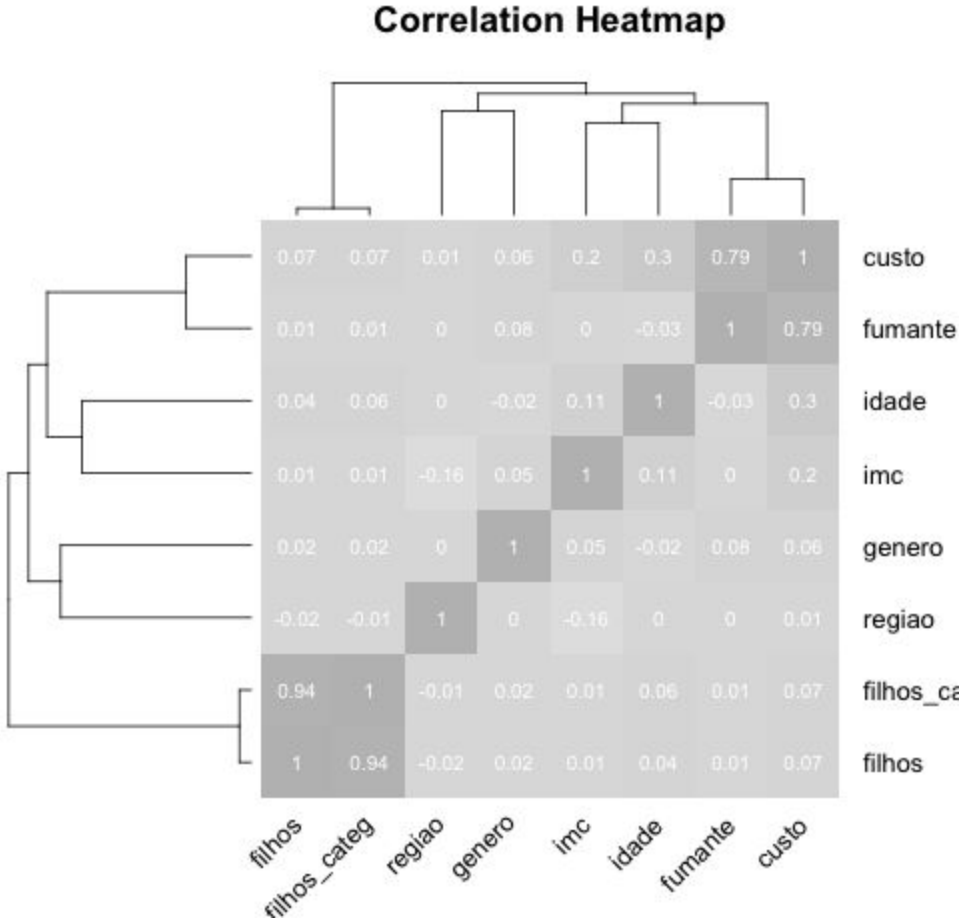

Custo por gênero
Custo
0
10000
20000
30000
40000
50000
60000
feminino
masculino
genero


Custo por fumante
Custo
0
10000
20000
30000
40000
50000
60000
não
sim
fumante


Custo por filhos
Custo
0
10000
20000
30000
40000
50000
60000
0
1
2
3
4
5
filhos


Custo por região
Custo
0
10000
20000
30000
40000
50000
60000
nordeste
noroeste
sudeste
sudoeste
regiao

Inferência Estatística


Correlation Heatmap
0.07 0.07 0.01 0.06 0.2 0.3 0.79 1
0.01 0.01 0 0.08 0 -0.03 1 0.79
0.04 0.06 0 -0.02 0.11 1 -0.03 0.3
0.01 0.01 -0.16 0.05 1 0.11 0 0.2
0.02 0.02 0 1 0.05 -0.02 0.08 0.06
-0.02 -0.01 1 0 -0.16 0 0 0.01
0.94 1 -0.01 0.02 0.01 0.06 0.01 0.07
1 0.94 -0.02 0.02 0.01 0.04 0.01 0.07
custo
fumante
idade
imc
genero
regiao
filhos_ca
filhos
filhos
filhos_categ
regiao
genero
imc
idade
fumante
custo

lm(formula = custo ~ idade + imc + filhos + fumante, data = numeric_df)

	Estimativa	Erro	T-Valor	P-Valor
Intercepto	-12102.77	941.98	-12.848	<2e-16
Idade	257.85	11.90	21.675	<2e-16
IMC	321.85	27.38	11.756	<2e-16
Filhos	473.50	137.79	3.436	6,08E-04
Fumante	23811.40	411.22	57.904	2,00E-16

Erro Padrão Residual	6068	
R^2 Múltiplo	0,75	
R^2 Ajustado	0,75	
Estatística F	998,1	p < 2.2e-16

lm(formula = custo ~ fumante, data = numeric_df)

	Estimativa	Erro	T-Valor	P-Valor
Intercepto	8434.3	229.0	36.83	<2e-16
Fumante	23616.0	506.1	46.66	<2e-16

Erro Padrão Residual	7470	
R^2 Múltiplo	0,62	
R^2 Ajustado	0,62	
Estatística F	998,1	p < 2.2e-16

Shapiro-Wilk fumante$custo | nao_fumante$custo

	W	P-Valor
Fumantes	0,94	3,63E-06
Não fumantes	0,87286	2,20E-16

ANOVA custo para fumantes e não fumantes por região

	P-Valor
Fumantes	0,0205
Não fumantes	0,0917

Teste de Tukey

	diff	lwr	upr	p adj
noroeste-nordeste	518	-4.766	5.803	0,99
sudeste-nordeste	5.171	428	9.914	0,03
sudoeste-nordeste	2.596	-2.689	7.880	0,58
sudeste-noroeste	4.653	-297	9.603	0,07
sudoeste-noroeste	2.077	-3.394	7.548	0,76
sudoeste-sudeste	-2.576	-7.526	2.374	0,54

CONSIDERAÇÕES FINAIS

- A variável que mais explica o custo é a fumante
  - Somente a variável fumante explica 62% do custo
  - Onde se a pessoa é fumante existe um aumento de $23811.40
- Para fumantes existe uma diferença significativa somente para a região sudeste-nordeste