Avaliação Final – ECD – Questões 3 e 4

Na base de dados comorbidades.csv, são apresentados dados reais de uma amostra obtida do seade-R (Fonte dos dados originais: https://github.com/seade-R/dados-covid-sp).

Essa base de dados contém as seguintes informações sobre pacientes que foram internados com diagnóstico de COVID-19 entre fevereiro de 2020 e maio de 2021:

Identificação do paciente

Município

Código do IBGE

Idade

Sexo (1: feminino, 0: masculino)

Óbito (1: sim, 0: não)

Comorbidades: asma, cardiopatia, diabetes, doença hematológica, doença renal, doença hepática, doença neurológica, imunodepressão, obesidade, outros fatores de risco, pneumopatia, puérpera, síndrome de down (para cada uma delas 1: presente, 0: ausente)

As observações com dados faltantes foram excluídas da base original para esta análise específica, considerando que essa exclusão não afeta a representatividade da amostra.

Questão 3

Descreva por meio de gráficos a associação entre idade e óbito, e repita para sexo e óbito. Considere então as comorbidades: asma, cardiopatia, diabetes, doença renal e obesidade e investigue a associação de cada uma delas com a variável óbito. Note que algumas variáveis estão codificadas em 0 e 1 mas são qualitativas. Fique atento à forma de analisar a associação entre duas variáveis qualitativas. Comente os resultados obtidos, observando que não é possível concluir estatisticamente se existe ou não associação entre as variáveis apenas com análise exploratória.

Resposta:

Inicialmente vamos verificar algumas informações gerais sobre a base de dados.

Quantidade de dados: (1182, 18)

Temos 1182 observações e 18 atributos ou variáveis.

Dessas 18 variáveis, temos interesse para esse estudo em um número menor delas, a saber:

Variáveis preditoras: 'idade', 'sexo', 'asma', 'cardiopatia', 'diabetes', 'doenca_renal', 'obesidade'

Target = 'obito'

Na base de dados, 36% dos casos são de óbito e 64% de não óbito.

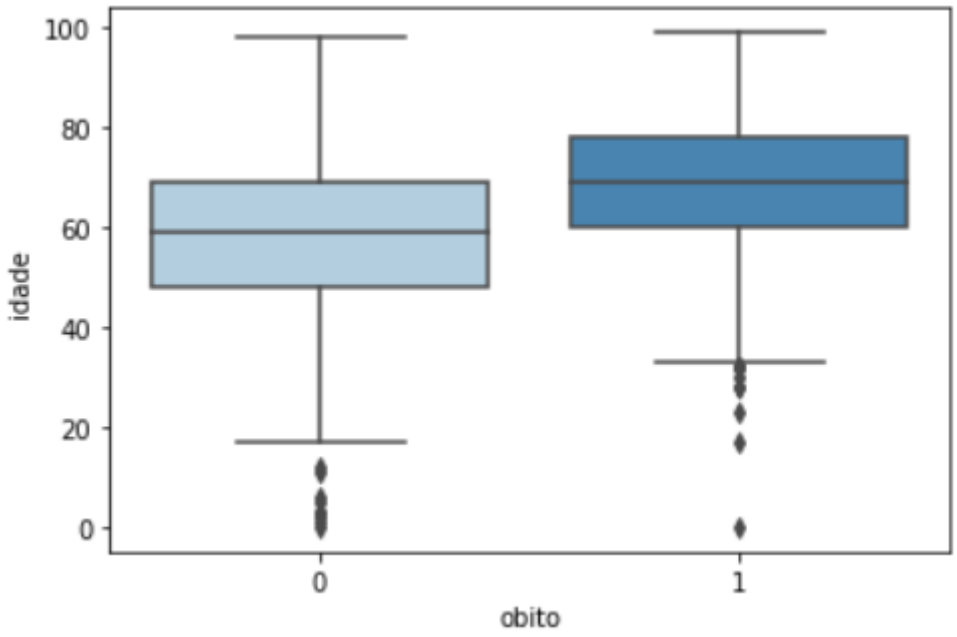
Vamos agora fazer a análise exploratória da associação entre óbito e as variáveis preditoras.

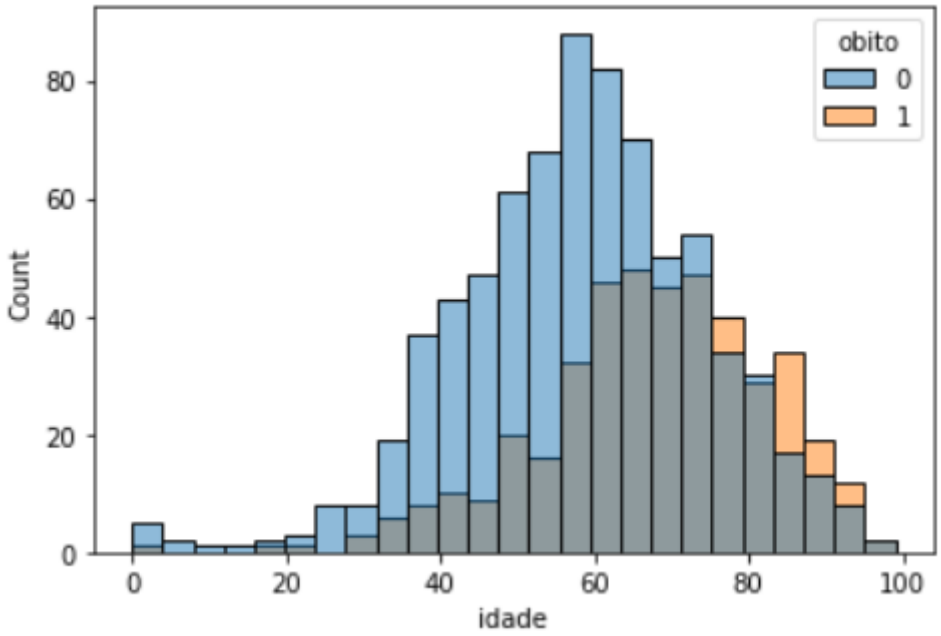
Associação entre óbito e idade:

A mediana para óbito ('obito'=1) é maior do que a mediana para não óbito ('obito'=0).

É uma pista de que quanto maior a idade maior o risco de óbito.

Entretanto, o IQ (intervalo interquartil) de ambos têm uma faixa de idades comum, indicando que uma análise mais aprofundada se faz necessária.



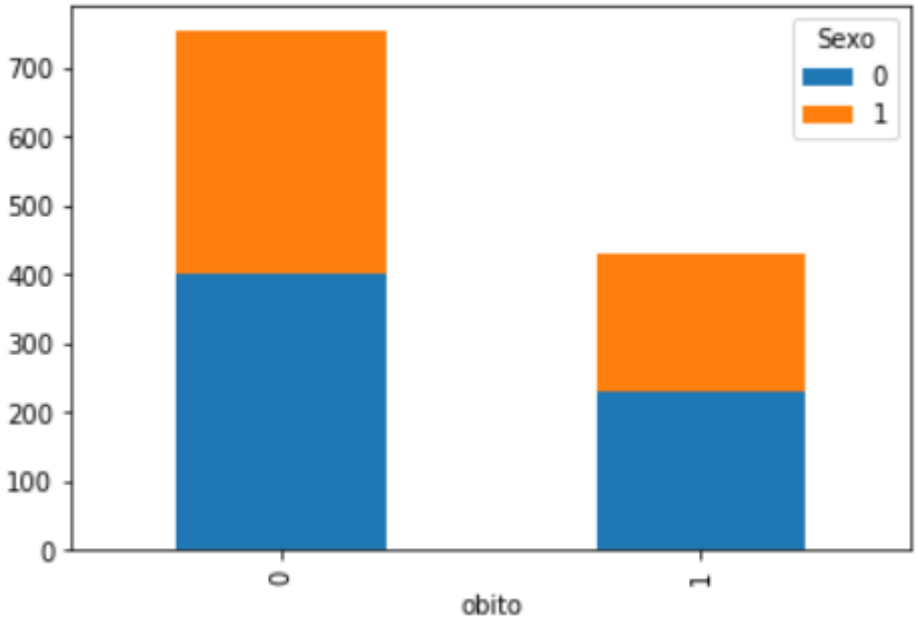


Associação entre óbito e sexo:

Uma vez infectado, a taxa de óbito para ambos os sexos é muito similar, próximo a 36%, a mesma do total da base de dados, sendo uma pista de que não há uma correlação entre esses fatores.

sexo	0	1
obito		
0	400	353
1	232	197

sexo	0	1
obito		
0	0.633	0.642
1	0.367	0.358



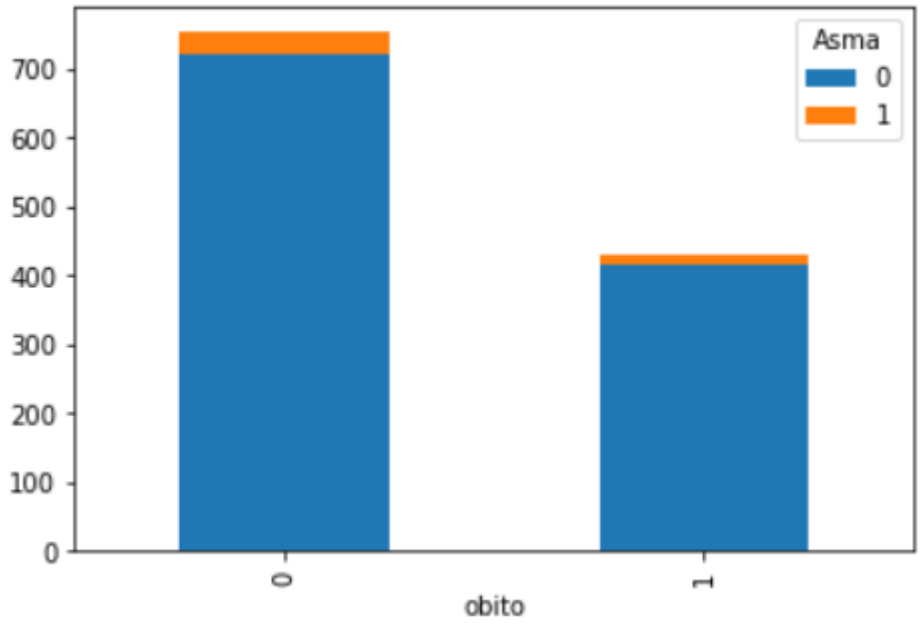
Associação entre óbito e asma:

Não há um aumento na taxa de morte para os portadores de asma uma vez infectados pelo vírus da covid-19, pelo contrário, há uma redução de 36,6 para 27,9%.

Porém, há um número muito pequeno de observações de casos de asma, o que pode ocasionar distorções nas conclusões.

asma	0	1
obito		
0	722	31
1	417	12

asma	0	1
obito		
0	0.634	0.721
1	0.366	0.279

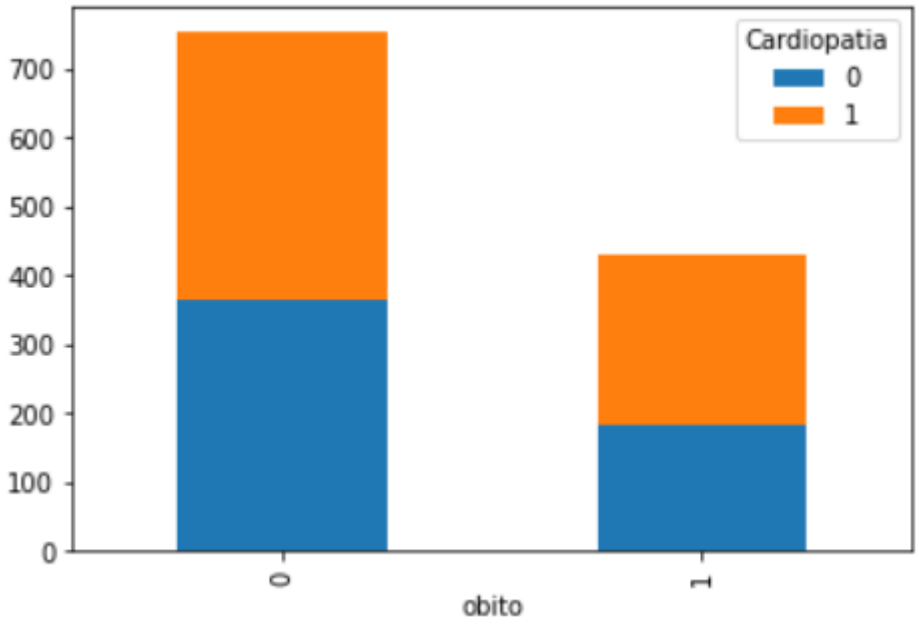


Associação entre óbito e cardiopatia:

Há um aumento na taxa de morte para os portadores de cardiopatia de 33,3 para 38,9% uma vez infectados pelo vírus da covid-19.

cardiopatia	0	1
obito		
0	366	387
1	183	246

cardiopatia	0	1
obito		
0	0.667	0.611
1	0.333	0.389

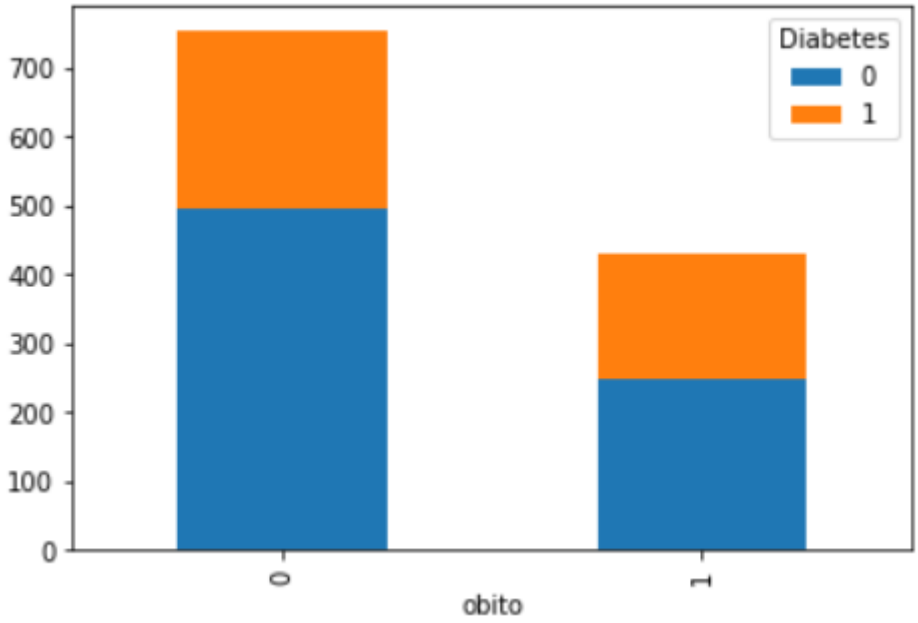


Associação entre óbito e diabetes:

Há um aumento na taxa de morte para os portadores de diabetes de 33,3 para 41,4% uma vez infectados pelo vírus da covid-19.

diabetes	0	1
obito		
0	497	256
1	248	181

diabetes	0	1
obito		
0	0.667	0.586
1	0.333	0.414



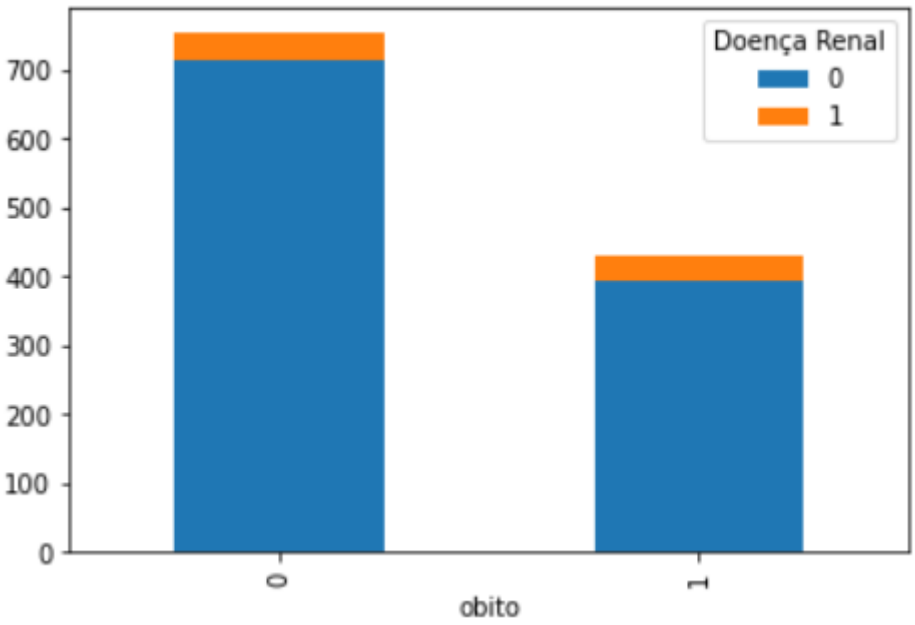
Associação entre óbito e doença renal:

Há um aumento na taxa de morte para os portadores de doença renal de 35,6 para 47,9% uma vez infectados pelo vírus da covid-19.

Porém, há um número muito pequeno de observações de casos de doença renal, o que pode ocasionar distorções nas conclusões.

doenca_renal	0	1
obito		
0	716	37
1	395	34

doenca_renal	0	1
obito		
0	0.644	0.521
1	0.356	0.479

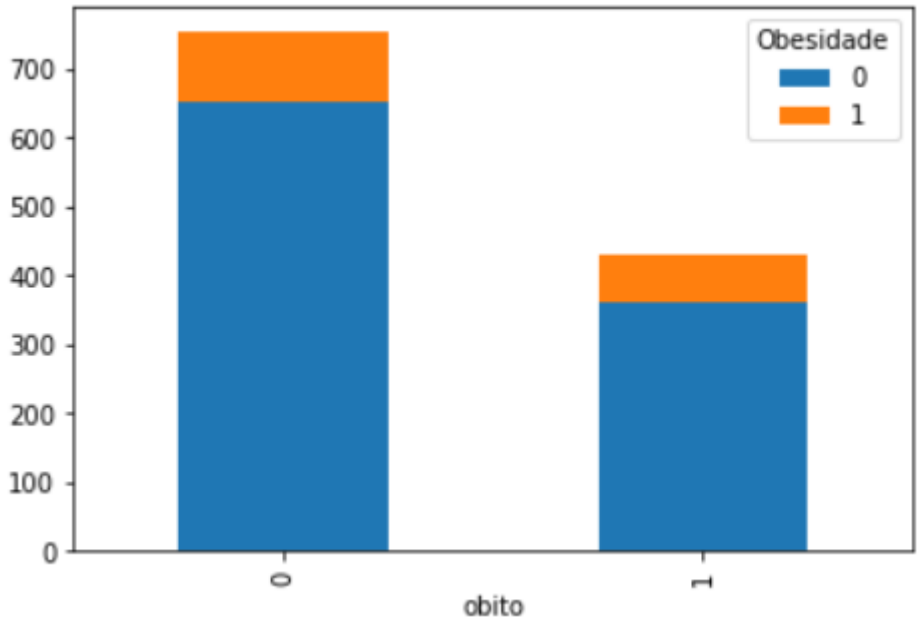


Associação entre óbito e obesidade:

Há um aumento na taxa de morte para os portadores de obesidade de 35,7 para 40,0% uma vez infectados pelo vírus da covid-19.

obesidade	0	1
obito		
0	651	102
1	361	68

obesidade	0	1
obito		
0	0.643	0.6
1	0.357	0.4



Conclusão da Análise Exploratória das Associações entre óbito e atributos selecionados:

A análise descritiva ou análise exploratória de dados (AED) efetuada acima tem como objetivos básicos:

. explorar os dados para descobrir ou identificar aspectos ou padrões de maior interesse,

. representar os dados de forma a destacar ou chamar a atenção para aspectos ou padrões que podem ou não se confirmar inferencialmente.

Com base nas análises de visualização e exploração de dados, parece existir associação entre 'óbito' e as variáveis 'idade', 'cardiopatia', 'diabetes', 'doenca_renal', 'obesidade' e 'asma', sendo que essa última apresenta uma possível correlação negativa.

Da mesma forma essa análise indicou que o risco de 'óbito' possivelmente não esteja associada ao 'sexo'.

Embora sejam indicativos baseados em dados, não é possível concluir estatisticamente se existe ou não associação entre as variáveis apenas com análise exploratória, sendo necessária uma análise confirmatória através de análises inferenciais.

Questão 4

Ajuste um modelo de regressão logística com intercepto, considerando como variáveis preditoras o sexo, a idade e as comorbidades asma, cardiopatia, diabetes, doenca_renal, obesidade. Descreva o impacto da presença das duas comorbidades com maior significância marginal utilizando a razão de chances. Os resultados confirmam a investigação inicial da Questão 3? Comente sobre os resultados obtidos. Para este item, considere que não é necessário fazer a seleção de atributos ou dividir a base em treinamento e teste.

Resposta:

Vamos usar o modelo de Regressão Logística com intercepto pela biblioteca statsmodels e o método GLM (Generalized Linear Model) ajustando-o a uma família binomial por termos um caso de resposta binária.

Conforme o enunciado, não faremos seleção de atributos, ou seja, não será usado o método backward, forward ou stepwise para a seleção dos mesmos.

Da mesma forma, a base de dados não será dividida entre treinamento e teste.

O sumário estatístico está apresentado abaixo:

Generalized Linear Model Regression Results

Dep. Variable:	obito	No. Observations:	1182
Model:	GLM	Df Residuals:	1174
Model Family:	Binomial	Df Model:	7
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-714.90
Date:	Wed, 16 Jun 2021	Deviance:	1429.8
Time:	08:20:55	Pearson chi2:	1.21e+03
No. Iterations:	4		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.4709	0.317	-10.948	0.000	-4.092	-2.850
idade	0.0442	0.005	9.396	0.000	0.035	0.053
sexo	-0.1438	0.129	-1.114	0.265	-0.397	0.109
asma	-0.3582	0.366	-0.980	0.327	-1.075	0.358
cardiopatia	-0.0642	0.131	-0.490	0.624	-0.321	0.192
diabetes	0.2654	0.130	2.035	0.042	0.010	0.521
doenca_renal	0.4241	0.262	1.621	0.105	-0.089	0.937
obesidade	0.6619	0.186	3.553	0.000	0.297	1.027

Comorbidades de maior significância marginal:

A significância marginal é dada pelos valores da coluna ‘P>|z|’ que representa o p-value do teste de hipóteses.

	P>|z|
const	0.000
idade	0.000
sexo	0.265
asma	0.327
cardiopatia	0.624
diabetes	0.042
doenca_renal	0.105
obesidade	0.000

Quanto menor o valor de ‘P>|z|’, maior a significância marginal.

Excluindo as não-comorbidades da análise, chegamos a que as 2 comorbidades com maior significância marginal são por ordem de importância: obesidade e diabetes.

O impacto da presença dessas duas comorbidades pode ser analisada usando a razão de chances (em inglês OR: Opportunity Ratio).

OR = exp (coef estimado da variável)

<u>Obesidade:</u>

Coeficiente: 0.6619

Razão de Chances: 1.9385

Aumento de percentual na chance de vir a óbito: 93.85 %

<u>Diabetes:</u>

Coeficiente: 0.2654

Razão de Chances: 1.3040

Aumento de percentual na chance de vir a óbito: 30.40 %

Conclusão da Análise Confirmatória das Associações entre óbito e atributos selecionados:

Na análise exploratória de dados (AED), identificamos como possíveis associações com 'obito' as variáveis: 'idade', 'cardiopatia', 'diabetes', 'doenca_renal', 'obesidade' e 'asma', e sem associação a variável 'sexo'.

Num nível de confiança de 95%, ou seja, p-value 0.050, concluímos nessa análise confirmatória:

p-value < 0.050 (maior significância marginal): 'idade', 'obesidade', 'diabetes'

p-value > 0.050 (menor significância marginal): 'sexo', 'cardiopatia', 'doenca_renal', 'asma'

Portanto, 3 variáveis tiveram mudança de classificação: 'cardiopatia', 'doenca_renal', 'asma'

As demais tiveram sua associação ou não associação confirmadas nessa última análise.