SME0620 - Trabalho de Pesquisa

Análise Descritiva de um Conjunto de Dados de Doenças Cardíacas

Alunos:

Giovanni dos Santos – 13695341

Luciano Gonçalves Lopes Filho - 13676520

Professor: Vicente Garibay Cancho

São Carlos, 2024

Conteúdo

1 Introdução

As doencas cardiacas são a causa mais comum de mortes do ser humano, tanto em homens quanto mulheres, segundo o *WHO Global Health Estimates-2020*. Dessa Forma, o objetivo dessa pesquisa é compreender as possíveis causas de doenças por meio das relações entre dados estatísticos. Esse conhecimento pode ser muito importante na delimitação de políticas públicas que visem reduzir as taxas de óbitos decorrentes de problemas cardíacos.

2 Metodologia

2.1 Caracterização da origem dos dados e seus significados

Os dados foram coletados no artigo *Heart Disease Dataset* na plataforma de ciencia de dados Kaggle, recomendada no documento de especificao desse trabalho. citado no final desse documento.

É importante realizar uma breve descrição de alguns dos termos que foram utilizados nessa pesquisa, em vista de esclarecer sua devida importância. Nesse âmbito, o termo "angina"refere-se a dor no peito causada pela diminuição do fluxo sanguíneo no coração. Além disso, na pesquisa foram feitos alguns exames cardíacos, como o de esforço físico e o eletrocardiograma. Dessa maneira, o termo ST refere-se a uma parte específica do traçado de um eletrocardiograma, que representa a fase inicial da repolarização ventricular. Em adição a esse segmento, temos ST-T, que relaciona-se ao segmento ST e também à onda T.

2.2 Descrição dos métodos utilizados

As ferramentas utilizadas para a obtenção dos resultados foram o LibreOffice e o Google Sheets, ambas planilhas eletrônicas similares ao Microsoft Excel. Além disso, foi também utilizada a linguagem R, na versão 4.2.2 para a determinação das medidas pedidas.

2.3 Definição das variáveis e suas categorias

As variáveis podem ser categorizadas em duas categorias principais: as qualitativas e as quantitativas. Nesse contexto, a primeira é dividida em nominais e ordinais, enquanto a segunda em contínuas e discretas. Sob essa ótica, podemos definir cada uma das variáveis utilizadas na pesquisa e encaixá-las em classificações específicas.

Com isso em mente, montamos as seguintes tabelas com as variáveis.

Tabela 1: Variáveis Qualitativas

Variável	**Subcategoria**	**Descrição**
Sexo (Sex)	Nominal	1 = macho 0 = fêmea
Tipo de dor no peito (Chest Pain Type)	Nominal	1 = Angina típico 2 = Angina atípico 3 = Dor não-anginosa 4 = assintomático
Açúcar no sangue em jejum (Fasting Blood Sugar)	Nominal	$nível > 120mg/dl$ 1 = verdadeiro, 0 = falso
Eletrocardiograma em repouso (Resting Eletrocardiogram Results)	Nominal	0 = normal 1 = ondas ST-T anormais 2 = Hipertrofia do ventrículo esquerdo
Angina decorrente de exercício (Exercise Induced Angina)	Nominal	1 = sim 0 = não
Inclinação de ST no exercício (ST slope)	Nominal	1 = crescente 2 = estática 3 = decrescente
Classe (Target)	Nominal	1 = doença cardíaca 0 = normal

Tabela 2: Variáveis Quantitativas

Variável	**Subcategoria**
Idade (Age)	Discreta
Pressão do sangue em repouso (Resting bp s)	Discreta
Colesterol no soro sanguíneo (Cholesterol)	Discreta
Máxima taxa de batimentos cardíacos atingida (Max heart rate)	Discreta
Depressão do Seguimento ST (oldpeak)	Contínua

3 Resultados

3.1 Exploração descritiva dos dados

A exploração descritiva dos dados consiste em basicamente calcular as medidas de posição e de dispersão das informações angariadas ao longo do estudo. Nesse sentido, as medidas podem ser resumidas na seguintes tabelas.

Tabela 3: Medidas - Variável "Age"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	53.72
Variância	Dispersão	87.57
Desvio Padrão	Dispersão	9.36
Mediana	Posição	54
Amplitude	Dispersão	49
Moda	Posição	54
Quantil	Posição	47 54 60

Tabela 4: Medidas - Variável "Sex"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	0.76
Variância	Dispersão	0.18
Desvio Padrão	Dispersão	0.43
Mediana	Posição	1
Amplitude	Dispersão	1
Moda	Posição	1
Quantil	Posição	1 1 1

Tabela 5: Medidas - Variável "Chest pain type"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	3.23
Variância	Dispersão	0.87
Desvio Padrão	Dispersão	0.94
Mediana	Posição	4
Amplitude	Dispersão	3
Moda	Posição	4
Quantil	Posição	3 4 4

Tabela 6: Medidas - Variável "Resting bp s"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	132.15
Variância	Dispersão	337.41
Desvio Padrão	Dispersão	18.37
Mediana	Posição	130
Amplitude	Dispersão	200
Moda	Posição	120
Quantil	Posição	120 130 140

Tabela 7: Medidas - Variável "Cholesterol"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	210.36
Variância	Dispersão	10286.12
Desvio Padrão	Dispersão	101.42
Mediana	Posição	229
Amplitude	Dispersão	603
Moda	Posição	0
Quantil	Posição	188 229 269.75

Tabela 8: Medidas - Variável "Resting ecg"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	0.69
Variância	Dispersão	0.76
Desvio Padrão	Dispersão	0.87
Mediana	Posição	0
Amplitude	Dispersão	2
Moda	Posição	0
Quantil	Posição	0 0 2

Tabela 9: Medidas - Variável "Fasting blood sugar"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	0.21
Variância	Dispersão	0.17
Desvio Padrão	Dispersão	0.41
Mediana	Posição	0
Amplitude	Dispersão	1
Moda	Posição	0
Quantil	Posição	0 0 0

Tabela 10: Medidas - Variável "Oldpeak"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	0.92
Variância	Dispersão	1.18
Desvio Padrão	Dispersão	1.09
Mediana	Posição	0.6
Amplitude	Dispersão	8.8
Moda	Posição	0
Quantil	Posição	0 0.6 1.6

Tabela 11: Medidas - Variável "Exercise Angina"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	0.39
Variância	Dispersão	0.24
Desvio Padrão	Dispersão	0.49
Mediana	Posição	0
Amplitude	Dispersão	1
Moda	Posição	0
Quantil	Posição	0 0 1

Tabela 12: Medidas - Variável "Max heart rate"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	139.73
Variância	Dispersão	651.15
Desvio Padrão	Dispersão	25.52
Mediana	Posição	140.5
Amplitude	Dispersão	142
Moda	Posição	150
Quantil	Posição	121 140.5 160

Tabela 13: Medidas - Variável "ST slope"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	1.62
Variância	Dispersão	0.37
Desvio Padrão	Dispersão	0.61
Mediana	Posição	2
Amplitude	Dispersão	3
Moda	Posição	2
Quantil	Posição	1 2 2

Tabela 14: Medidas - Variável "Target"

Medida	**Tipo de medida**	**Valor(es)**
Média/Esperança	Posição	0.53
Variância	Dispersão	0.25
Desvio Padrão	Dispersão	0.49
Mediana	Posição	1
Amplitude	Dispersão	1
Moda	Posição	1
Quantil	Posição	0 1 1

3.2 Tabelas de Frequência

A tabela de frequência consiste na análise de um determinado dado, através da frequência absoluta e relativa de valores específicos atrelados a esse dado. Dessa forma, escolhemos a variável "Chest pain type"como qualitativa e a "Age"como quantitativa. Assim, foram montadas as seguites tabelas de frequência.

Tabela 15: Tabela de Frequência - "Chest pain type"(Qualitativa)

Valores	**Frequência absoluta**	**Frequência relativa**
1	66	0.055
2	216	0.182
3	283	0.239
4	625	0.53

Tabela 16: Tabela de Frequência - "Age"(Quantitativa)

Tabela 17: Tabela 1

Valores	$\%F_i$	F_i
28	0.0008	1
29	0.0034	4
30	0.0008	1
31	0.0017	2
32	0.0042	5
33	0.0017	2
34	0.0076	9
35	0.0118	14
36	0.0050	6
37	0.0109	13
38	0.0143	17
39	0.0151	18
40	0.0134	16
41	0.0277	33
42	0.0218	26
43	0.0261	31
44	0.0244	29
45	0.0210	25
46	0.0261	31
47	0.0193	23
48	0.0319	38
49	0.0227	27
50	0.0269	32
51	0.0395	47
52	0.0395	47

Tabela 18: Tabela 2

Valores	$\%F_i$	F_i
53	0.0336	40
54	0.0563	67
55	0.0395	47
56	0.0395	47
57	0.0420	50
58	0.0487	58
59	0.0395	47
60	0.0370	44
61	0.0319	38
62	0.0387	46
63	0.0311	37
64	0.0261	31
65	0.0244	29
66	0.0160	19
67	0.0193	23
68	0.0109	13
69	0.0134	16
70	0.0092	11
71	0.0067	8
72	0.0034	4
73	0.0008	1
74	0.0067	8
75	0.0025	3
76	0.0025	3
77	0.0025	3

3.3 Representação Gráfica

Para demonstrar algumas opções de representações gráficas foi-se escolhida uma variável quantitativa, no caso a "Max heart rate". Sob essa ótica, foram feitas as seguintes representações.

Figura 1: Representação Gráfica Histograma - "Max heart rate"

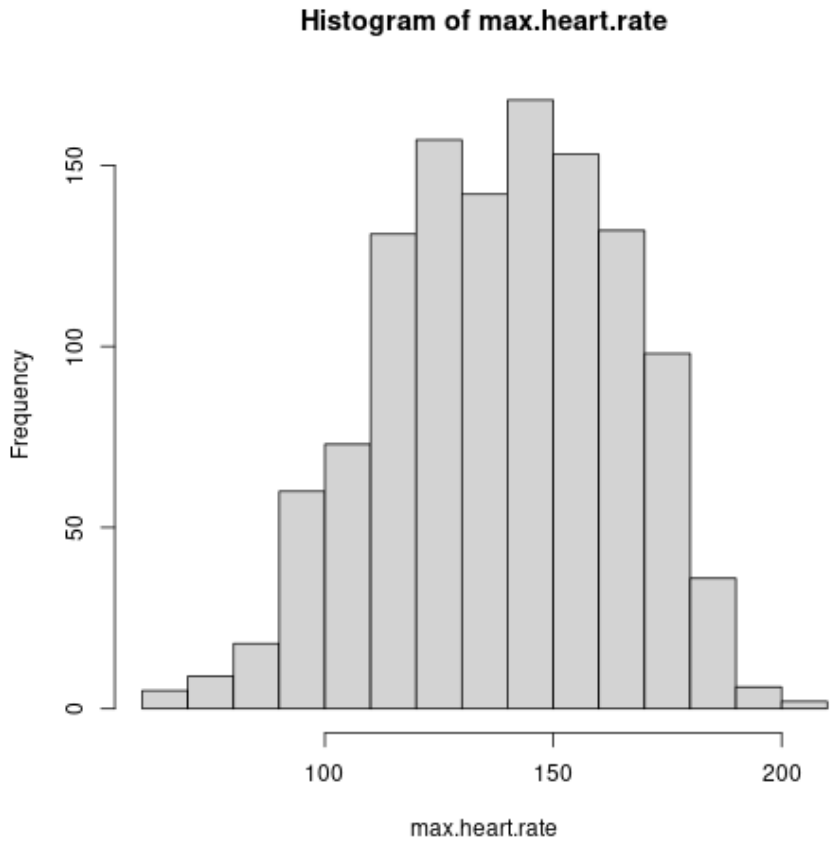
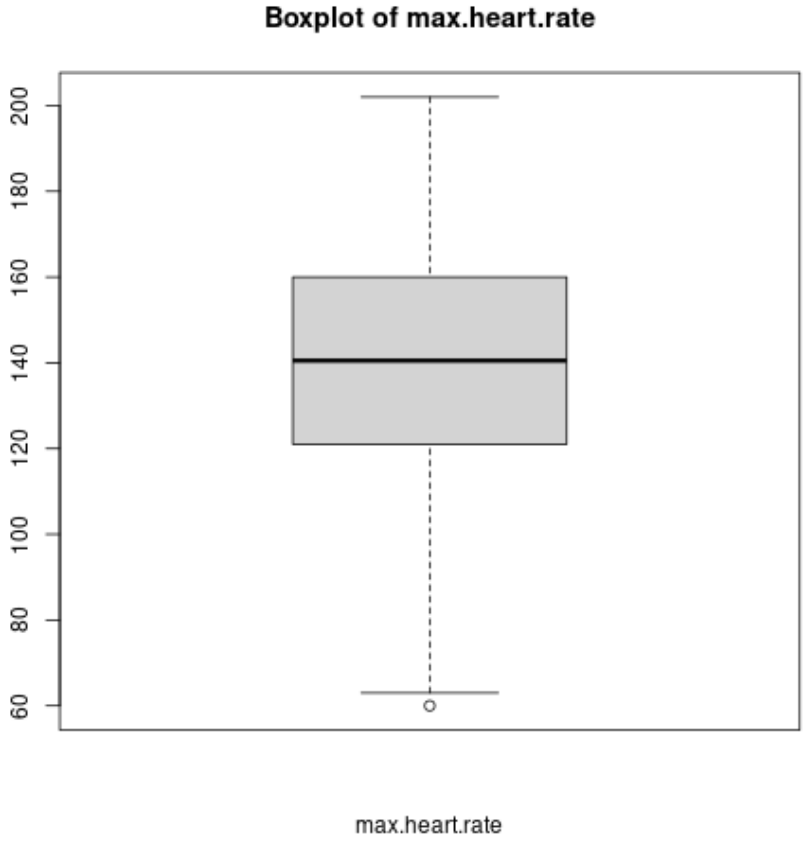


Figura 2: Representação Gráfica Boxplot - "Max heart rate"



Segundo o Histograma e o Boxplot, percebemos que a mediana de valores para o máximo batimento cardíaco é de 140BPM, e metade dos valores se encontram entre 120BPM e 160BPM, confirmado pelos quartis do boxplot. Vemos, também alguns outliers, em frequências de batimento abaixo do limite inferior de 60BPM.

3.4 Análise de duas variáveis quantitativas

Para realizar a análise e o diagrama de dispersão pedidos, foram escolhidas as variáveis "Age"e "Resting bps". Nesse sentido, foi montado o seguinte diagrama.

Figura 3: Diagrama de dispersão - "Age"X "Resting bps"

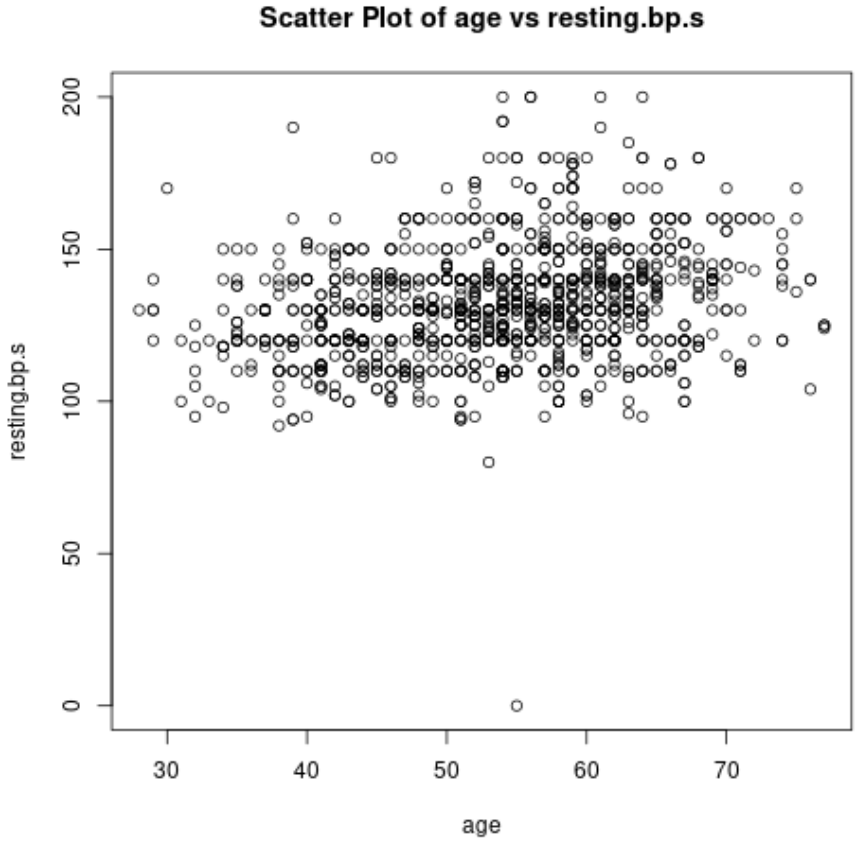


Tabela 19: Relações entre Variáveis - "Age"X "Resting bps"

Covariância	**Correlação**
44.29706	0.2576921

Com base nos dados expostos acima, depreende-se que a correlação entre Idade e Pressão sanguínea em repouso pode ser considerada desprezível, já que se enquadra no intervalo $|\rho| \leq 0,30$.

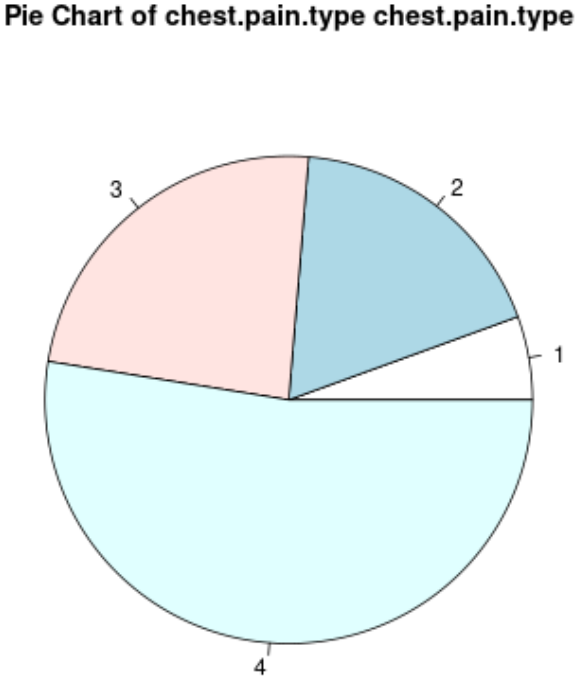
3.5 Análise de uma variável quantitativa e outra qualitativa

As medidas Resumo de todas as variáveis já foram feitas anteriormente, no tópico 3.1. Dessa forma, aqui serão apresentados apenas as representações

gráficas e suas interpretações.

No contexto explicitado, a variável qualitativa analisada será "Chest pain type", da qual foi esquematizada uma representação circular, evidenciada abaixo.

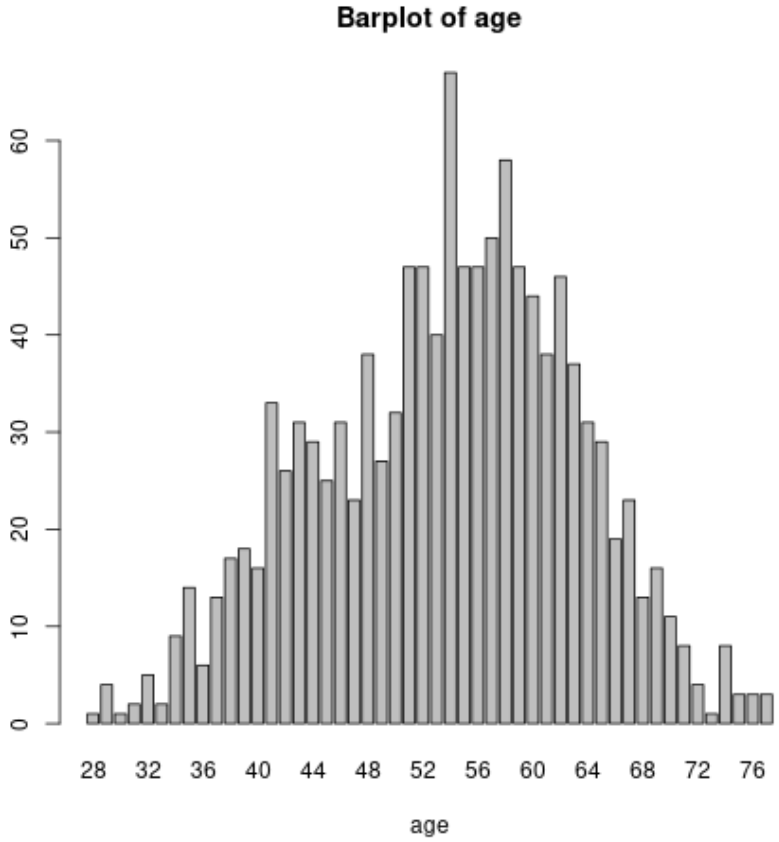
Figura 4: Representação gráfica - "Chest pain type"



Da imagem apresentada, é evidente a predominância de que a maior parte dos pacientes analisados não apresentam qualquer tipo de incômodo no peito. Além disso, nos casos em que realmente há um desconforto dolorido na região do peito normalmente não se encaixa num caso de Angina, sendo estes mais raros e representados pelos índices 1 e 2.

Em paralelo ao mostrado, temos a variável quantitativa "Age", da qual montamos a seguinte representação em barras.

Figura 5: Representação gráfica - "Age"



Da representação gráfica esquematizada acima, depreende-se que as idades dos pacientes analisados aproximam-se de uma distribuição normal, tendo como idade mais recorrente (moda) 54 anos.

4 Fontes

World Health Organization. WHO Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. Geneva: World Health Organization, 2020.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

https://www.kaggle.com/datasets/mexwell/heart-disease-dataset?select=documentation.pdf