**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Inteligência Artificial e Aprendizado de Máquina**

**Vinicio Silva Lima**

**PREDIÇÃO DE INSUFICIÊNCIA CARDÍACA**

Belo Horizonte

Junho de 2022

**Vinicio Silva Lima**

**PREDIÇĀO DE INSUFICIÊNCIA CARDÍACA**

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Inteligência Artificial e Aprendizado de Máquina, como requisito parcial à obtenção do título de *Especialista*.

Belo Horizonte

Junho de 2022 **SUMÁRIO**

[**1. Introdução**](#_heading=h.gjdgxs) **4**

[**2. Descrição do Problema e da Solução Proposta**](#_heading=h.30j0zll) **5**

[**3. Canvas Analítico**](#_heading=h.1fob9te) **6**

[**Figura 1: Canvas Análitico**](#_heading=h.3znysh7) **6**

[**4. Coleta de Dados**](#_heading=h.tyjcwt) **6**

[**5. Processamento/Tratamento de Dados**](#_heading=h.3dy6vkm) **8**

[**6. Análise e Exploração dos Dados**](#_heading=h.1t3h5sf) **12**

[**7. Preparação dos Dados para os Modelos de Aprendizado de Máquina**](#_heading=h.4d34og8) **23**

[**8. Aplicação de Modelos de Aprendizado de Máquina**](#_heading=h.2s8eyo1) **25**

[**9. Discussão dos Resultados**](#_heading=h.17dp8vu) **25**

[**10. Conclusão**](#_heading=h.3rdcrjn) **26**

[**11. Links**](#_heading=h.26in1rg) **26**

[**12. Referências**](#_heading=h.35nkun2) **26**

# 1. Introdução

O ser humano tem a habilidade de aprender com os problemas e tomar decisões a partir de experiências já vividas. Quando nos referimos a computação, a Inteligência Artificial consegue, por meio de algoritmos complexos de Aprendizado de Máquina, encontrar padrões matemáticos para inferir resultados, de acordo com a necessidade das pessoas. Hoje, com o avanço da Inteligência Artificial, é possível realizar predições se um paciente tem uma pré-disposição a sofrer um ataque cardiovascular baseando-se em dados, tais como: idade, pressão arterial, colesterol, entre outros dados. Apesar disso, por incrível que pareça, ainda sofremos frequentemente com baixas relacionadas a problemas de insuficiência cardíaca, até mesmo em países mais desenvolvidos.

A insuficiência cardíaca aflige ou mata um em cada dois adultos nos Estados Unidos e em outros países desenvolvidos. (Go AS, Mozaffarian D, Roger VL, et al. 2013). Uma das causas dos ataques cardiovasculares é o acúmulo de placas de colesterol nas paredes internas de artérias, conhecido como aterosclerose. Normalmente o desenvolvimento da doença é silencioso e pode desenvolver-se na adolescência ou até mesmo na infância. Quando começa a apresentar sintomas, causa listras esbranquiçadas no revestimento interno das artérias que, com o passar do tempo, transformam-se em bolsas de colesterol que podem inchar-se, dificultando a passagem do fluxo sanguíneo na parede arterial, causando desconforto no peito. Em casos de rompimento das placas de colesterol, pode causar coágulos que, se muito grandes, podem impedir o fluxo sanguíneo e resultar em um ataque cardíaco ou um derrame. (Harvard T.H Chan School of Public Health, 2022).

A prática de atividade física regular reduz o risco de doenças cardíacas, diabetes, acidente vascular cerebral, pressão alta, osteoporose e até mesmo alguns tipos de câncer. (Harvard T.H Chan School of Public Health, 2022). Apesar disso, à medida que vamos envelhecendo, temos uma diminuição de atividade física, nos deixando mais expostos a doenças cardiovasculares (Matthews CE, George SM, Moore SC, et al. 2012).

Com base nisso, entendendo a importância deste tema, este trabalho tem como objetivo aplicar técnicas de Aprendizado de Máquina em uma base de dados de Insuficiência Cardíaca encontrada na plataforma de desafios de Ciência de Dados *Kaggle.* Por meio de modelos de classificação, podemos predizer se os pacientes encontrados no *dataset* possuem insuficiência cardíaca ou não, apresentando todo o processo de desenvolvimento do estudo como um todo e, desta forma, oferecer métricas de classificação efetivas para avaliar os modelos desenvolvidos.

## 2. Descrição do Problema e da Solução Proposta

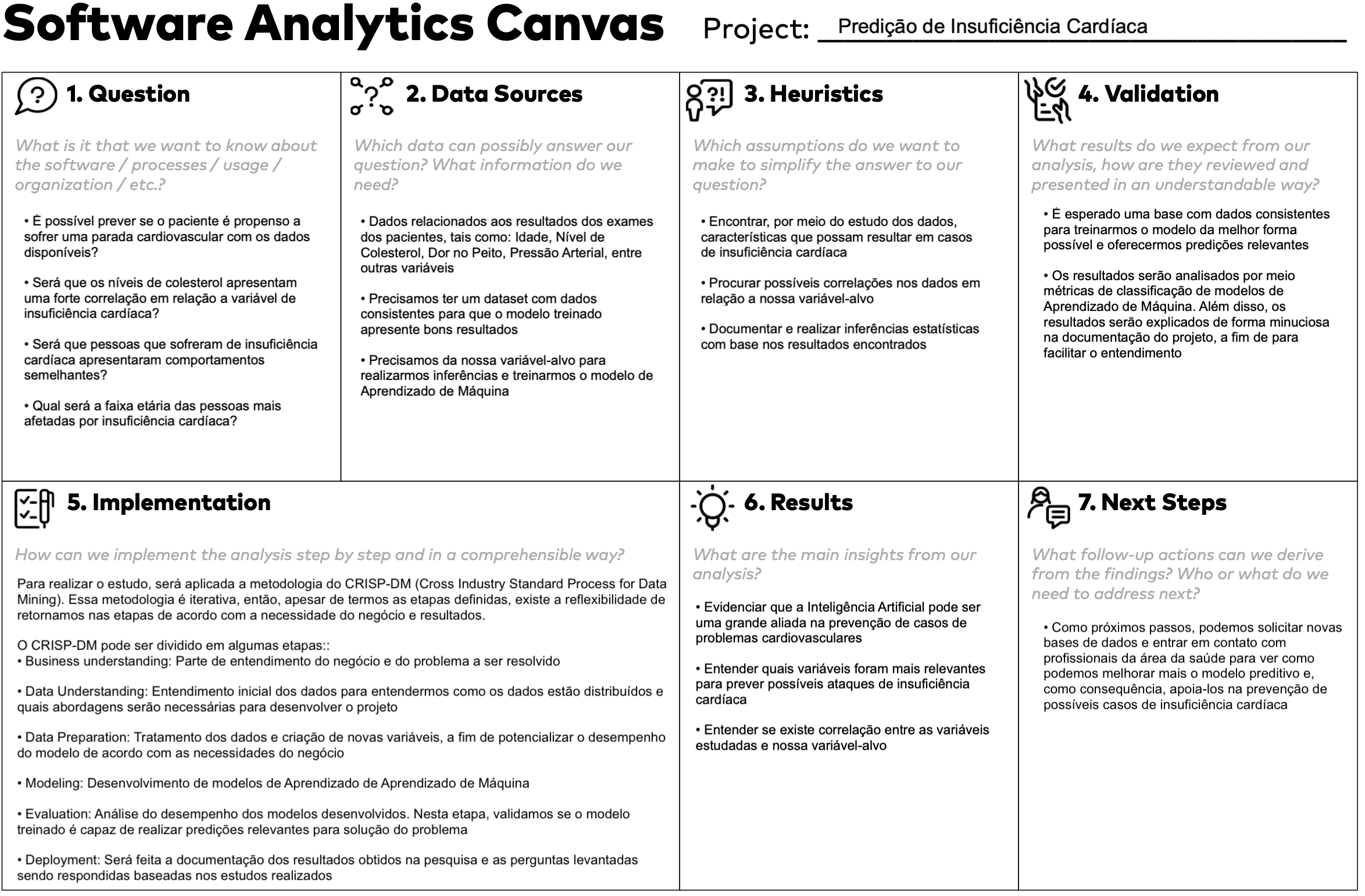
Doenças cardiovasculares levam cerca de 17,9 milhões de vidas a cada ano, representando 31% de todas as mortes em todo mundo. Grande parte desses ataques de insuficiência cardíaca estão relacionados a problemas de acúmulo de placas de colesterol nas artérias. Independentemente do tamanho deste armazenamento de colesterol nas artérias, esses pacientes estão sujeitos a rompimentos inesperados nesses vasos, causando dores no peito e até mesmo derrames.

Não existe uma idade específica para um paciente ter o risco de sofrer de um ataque cardíaco. Infelizmente, muitas dessas vítimas sofrem de hipertensão, diabetes ou até mesmo hiperlipidemia.

Para evitar mais baixas de tantas pessoas, é importante que o diagnóstico de pacientes propensos a sofrerem de ataques cardíacos seja feito com antecedência. Neste requisito, modelos de Aprendizado de Máquina podem apoiar na detecção precoce de possíveis vítimas de doenças cardiovasculares, já que conseguem aprender com os padrões encontrados nas bases de dados e predizer se o paciente corre risco de ter um ataque cardíaco ou não.

Para resolver problemas de classificação, os modelos são comumente de Árvores de Decisão ou de Regressāo Logística, por exemplo. Com as predições feitas, os modelos sāo avaliados através de métricas de classicaçāo (Recall, Precision, F1-Score, etc) e os resultados podem ser compartilhados com uma equipe de médicos, a fim de mostrar que a Inteligência Artificial pode ser uma grande aliada na prevenção de possíveis ataques cardiovasculares.

# 3. Canvas Analítico



# Figura 1: Canvas Análitico

# 

# 4. Coleta de Dados

A base de dados *Heart Failure Prediction Dataset* foi extraída na plataforma Ka*ggle* por meio do seguinte link: <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction> . A extração foi feita no dia 20/06/2020.

A base é composta por *datasets* de pacientes dos Estados Unidos, Hungria, França e Suíça. Não foi fornecida a data de extração dos dados e a forma de como a base foi obtida.

| **Nome do dataset: Heart Failure Prediction Dataset**  **Descrição:** *Dataset* com informações de pacientes americanos, suíços, franceses, húngaros com ou sem histórico de insuficiência cardíaca. O objetivo é analisar as variáveis e desenvolver um modelo capaz de realizar predições de possíveis pacientes propensos a sofrer de insuficiência cardiovascular.  **Data de publicaçāo:** 10/2022  **Link:** *https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction* | | |
| --- | --- | --- |
| **Nome do Atributo** | **Descrição** | **Tipo** |
| Age | Idade do paciente | Inteiro |
| Sex | Sexo do paciente | String |
| ChestPainType | Tipo de dor no peito | String |
| Cholesterol | Colesterol total | Inteiro |
| ExerciseAngina | Dor no peito proveniente do exercício | String |
| FastingBS | Glicose no sangue | Inteiro |
| MaxHR | Máxima de batimento cardíaco | Inteiro |
| Oldpeak | Depressāo ST induzida por exercício relativamente sossegado | Float |
| RestingECG | Conclusāo do Eletrocardiograma | String |
| ST\_Slope | Taxa de frequência cardíaca | String |
| HeartDisease | Insuficiência Cardíaca (variável-alvo) | Inteiro |

# 5. Processamento/Tratamento de Dados

Para realizar a análise, foi utilizado o *Jupyter Notebook* e as bibliotecas pandas, numpy e category\_encoders

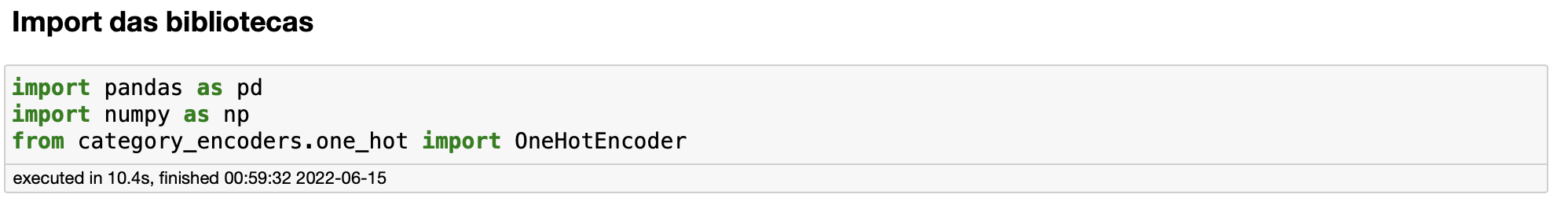


Figura 2: import das bibliotecas

Para começar a análise, precisamos entender como os nossos dados estāo distribuídos, conforme imagem abaixo:

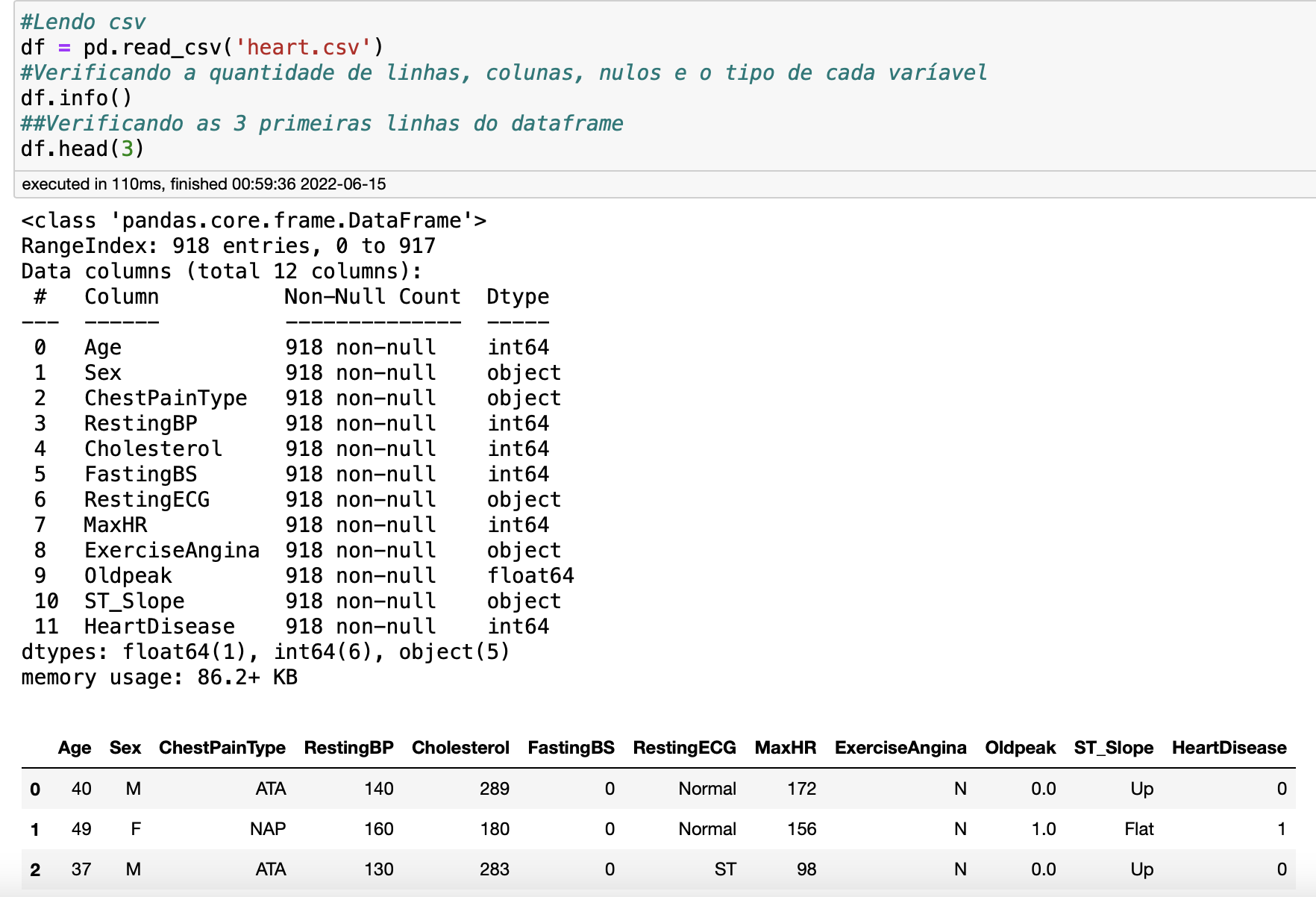


Figura 3: Leitura e entendimento inicial dos dados

Após importarmos o csv, podemos verificar que nossa base não possui variáveis nulas. Como temos atributos do tipo *object*, vamos precisar separar o nosso *dataframe* em duas partes para trabalharmos com as variáveis numéricas separadamente.

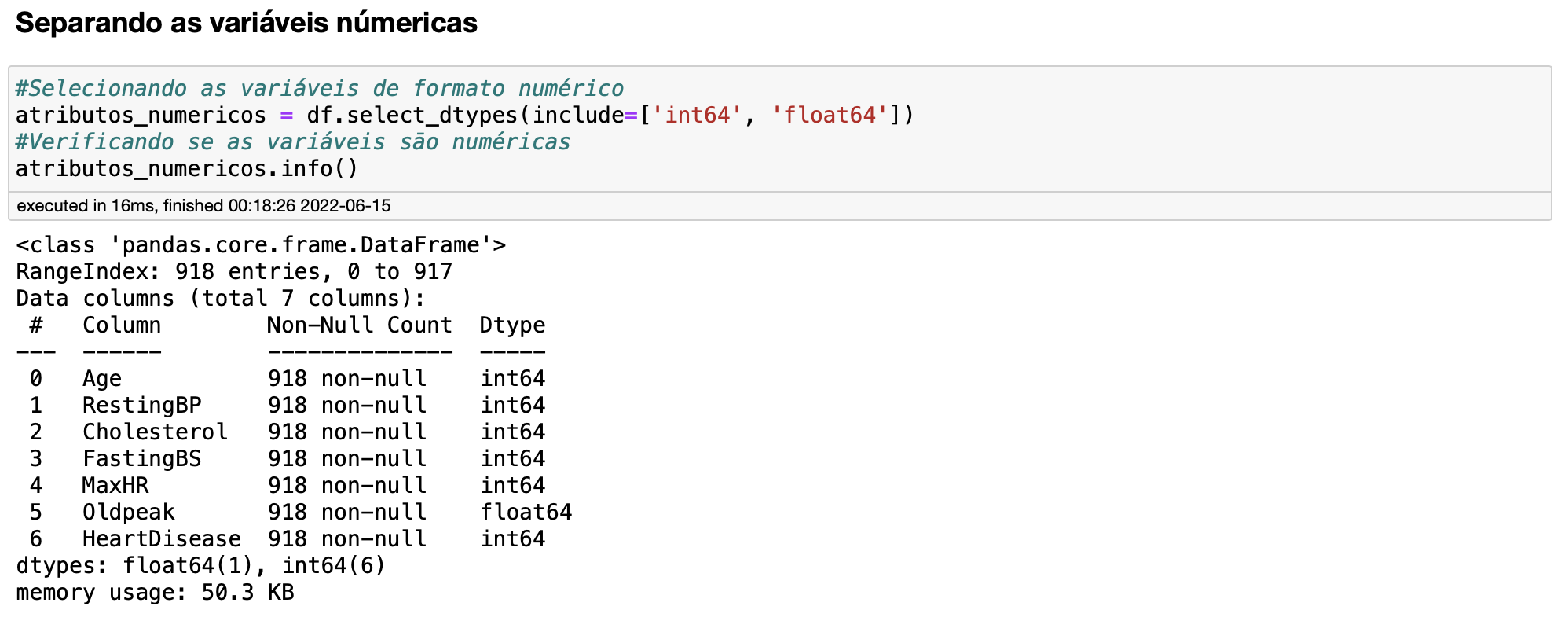


Figura 4: Separaçāo das variáveis numéricas

Para separar as variáveis numéricas foi utilizada a função *select\_dtypes* da biblioteca *pandas* e foi feito um filtro para selecionar as que possuem valores inteiros e *float.* Com o resultado atribuído à variável “atributos\_numericos”, foi utilizada a função *info* para conferir os resultados.

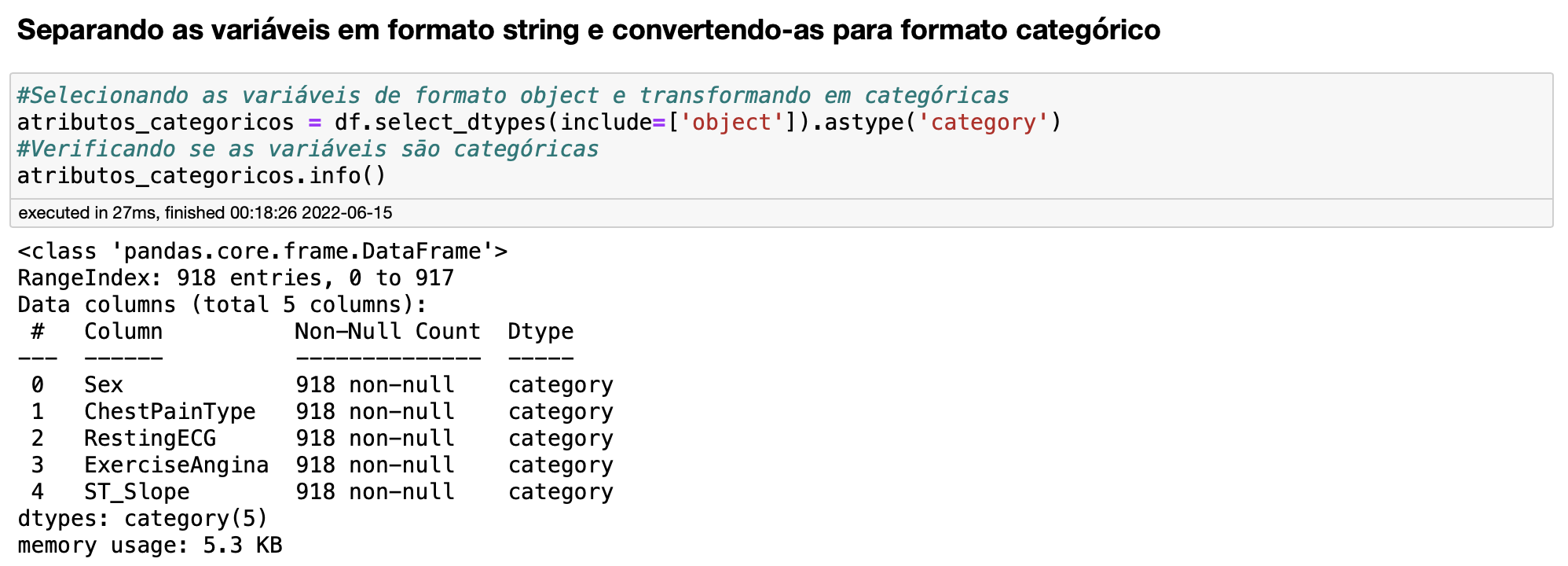


Figura 5: Separaçāo das variáveis categóricas

No caso das variáveis do tipo objeto foi feito uma conversão para formato categórico, pois temos atributos como “*Sex*” que apresenta resultados provenientes de categorias, como masculino e feminino. Após a atribuição do resultado à variável “atributos\_categoricos”, realizamos a checagem se o tipo das variáveis foram convertidos para *object*.

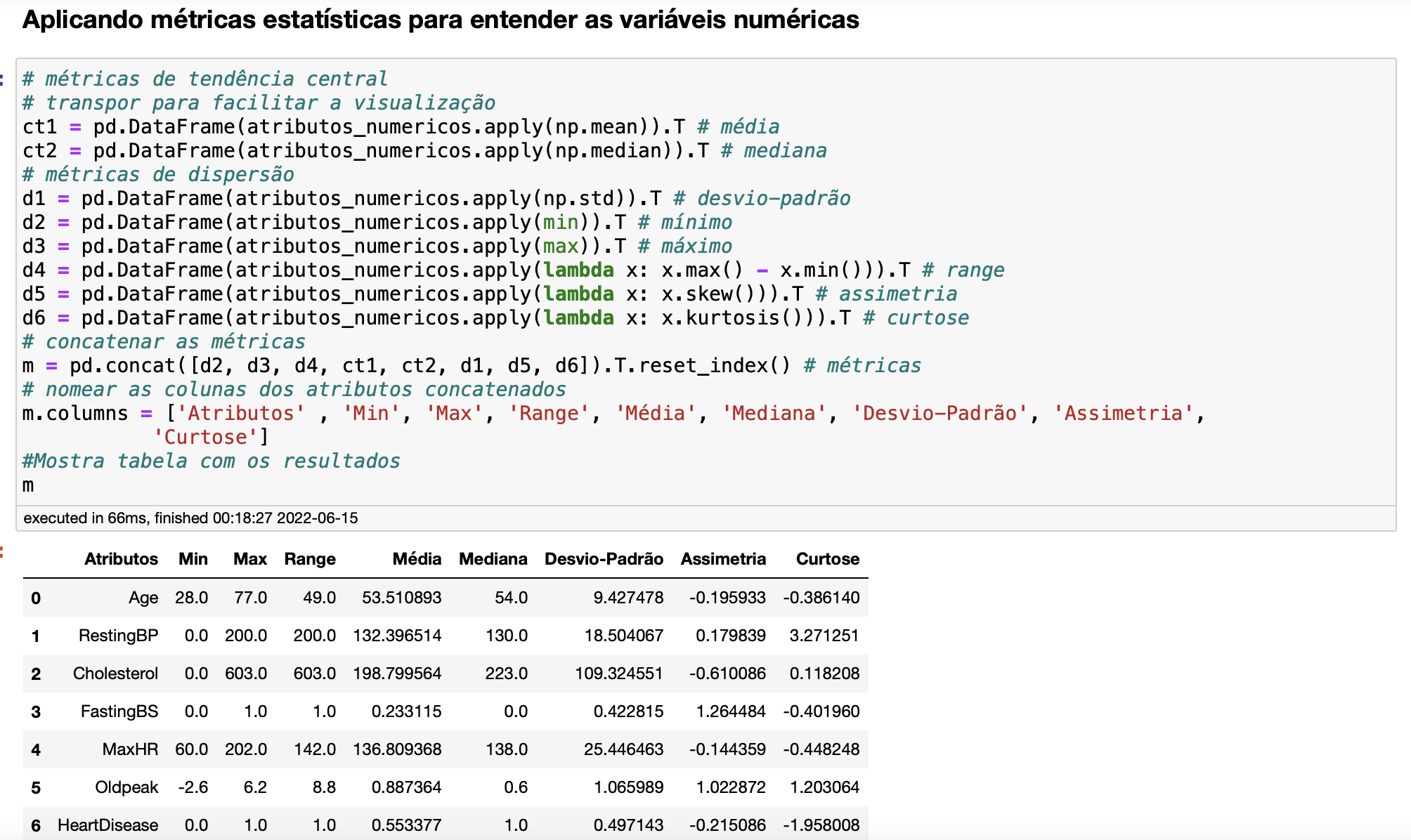


Figura 6: Métricas estatísticas para entendimento de variáveis numéricas

Utilizando nossa variável “atributos\_numericos” fizemos a declaração de alguns atributos e adicionamos medidas de dispersāo para entendermos melhor os dados que estamos trabalhando. No caso, podemos observar que a variável de *Cholesterol* apresenta um desvio-padrão bem alto e um valor mínimo de 0, o que pode indicar possíveis valores inválidos na base.

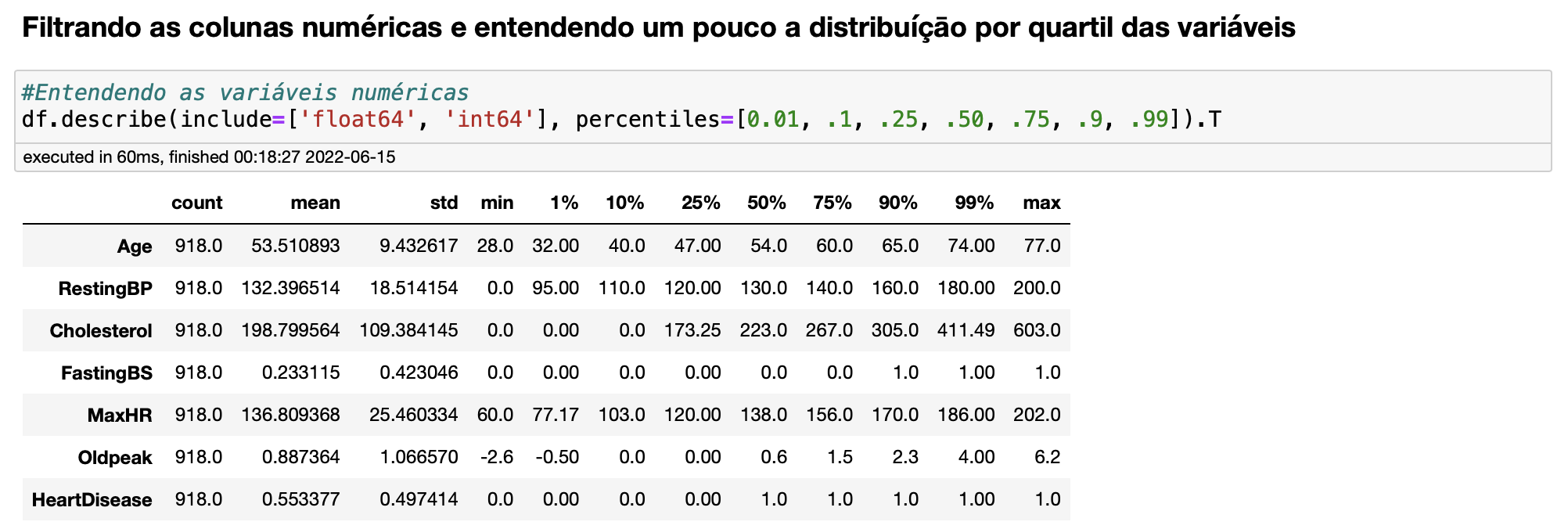


Figura 7: Analisando as variáveis com quebras por quartil

Para nos aprofundarmos um pouco mais nas variáveis, foi aplicada a função *describe* com quebras por quartis para entendermos melhor como está a dispersão dos atributos. Com base nisso, podemos observar que a variável de *Cholesterol* provavelmente possui *outliers*, já que o valor máximo encontrado está muito acima da média.

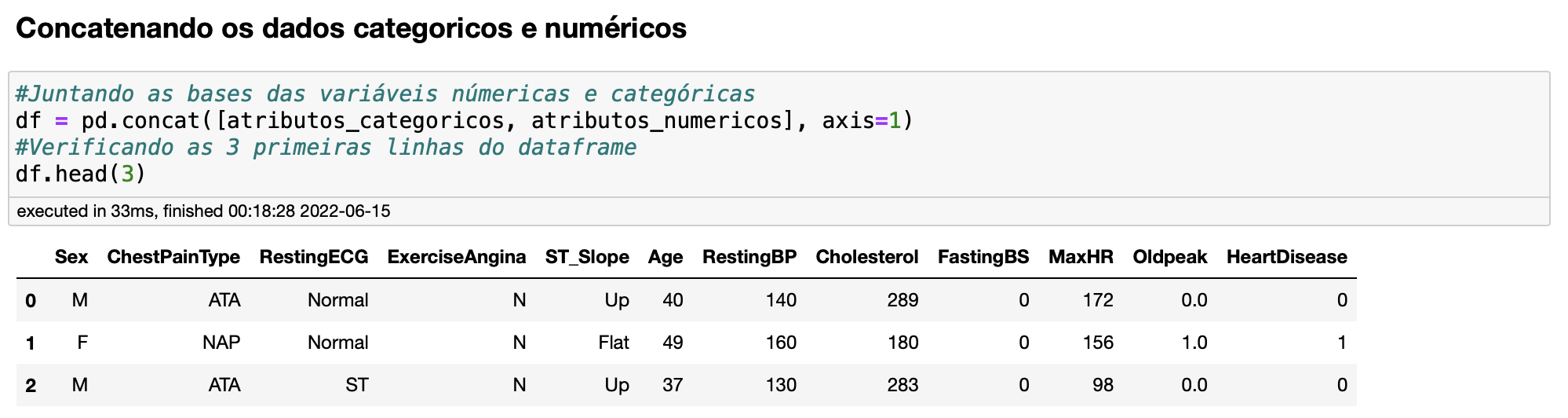


Figura 8: Junçāo de dataframes

Utilizando as variáveis “atributos\_categoricos” e “atributos\_numericos” podemos criar um novo *dataframe* concatenando as duas bases. Isso pode ser feito agora sem muitos impeditivos, já que o tipo dos atributos foram tratados anteriormente.

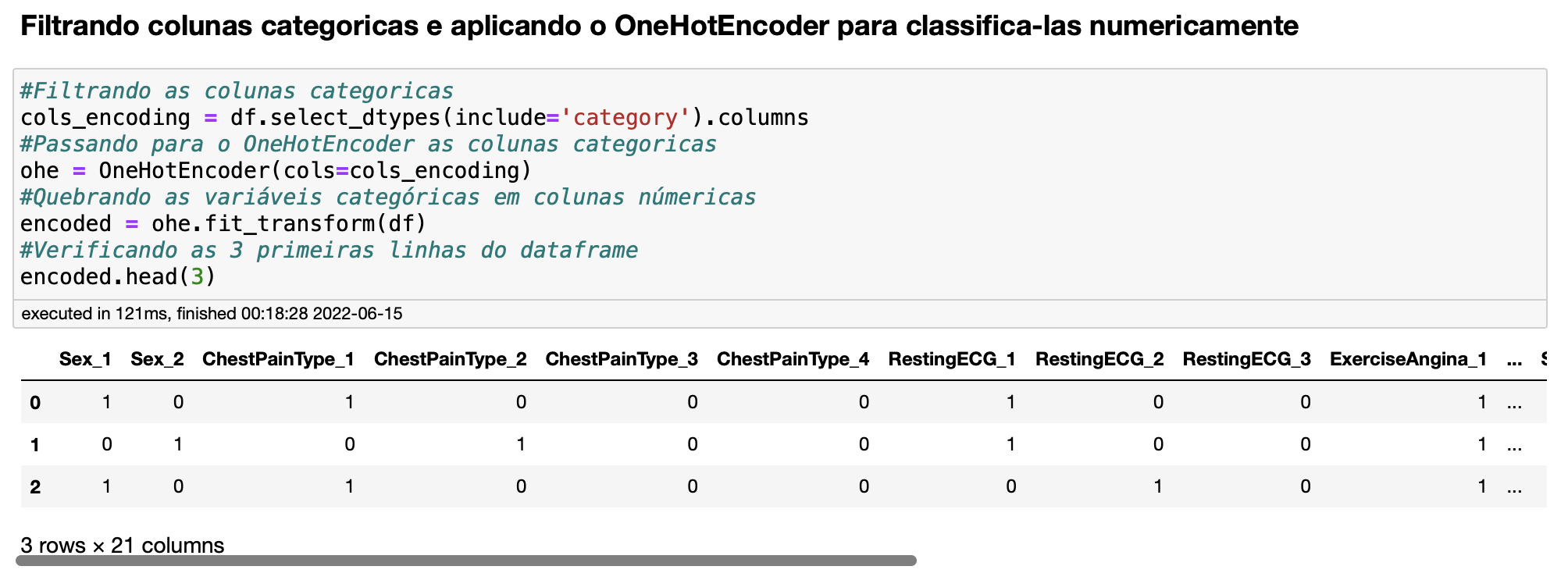


Figura 9: Aplicando o OneHotEncoder

Com o nosso novo *dataframe*, precisamos separar os valores categóricos novamente e atribuirmos a uma variável temporária chamada “cols\_encoding”. Feito isso, passamos as variáveis para o *OneHotEnconder* e aplicamos a transformação com o nosso método *fit*, quebrando nossas colunas de acordo com o número de categorias presentes em cada coluna. No caso de “ChestPainType”, existiam 4 categorias, então foram geradas quatro colunas distintas para separá-las.

# 6. Análise e Exploração dos Dados

Nesta etapa vamos dar continuidade a parte de análise descritiva das variáveis e entender melhor o comportamento das variáveis em relação ao nosso *target*. Para facilitar um pouco, vamos iniciar entendendo a distribuições de frequência de forma univariada e depois multivariada. Dependendo dos nossos resultados, vamos aplicar testes de hipótese, a fim de confirmar ou entender se as distribuições são gaussianas, por exemplo. Desta forma, vamos conseguir entender quais tipos de tratamentos vamos precisar realizar e quais variáveis são mais adequadas para realizar um modelo preditivo mais assertivo.

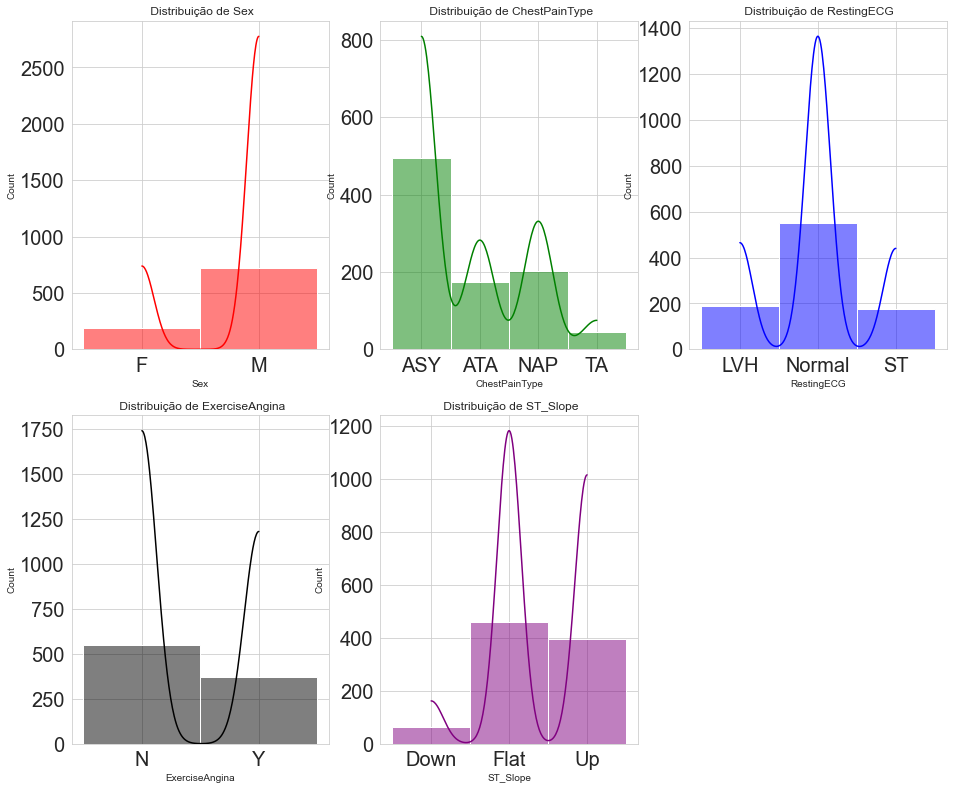


Figura 10: Distribuiçāo variáveis categóricas

Analisando as variáveis categóricas, podemos ver que a maioria dos nossos pacientes da base sāo do sexo masculino com base na nossa variável sex. Quando analisamos a variável de *ChestPainType*, temos uma predominância de pacientes assintomáticos, o que talvez seja um problema para descobrirmos se o paciente tem risco ou nāo de sofrer um ataque cardíaco, por isso precisamos verificar as outras variáveis.

*RestingECG* apresentou que a maioria dos pacientes não tiveram nenhuma alteraçāo nos exames de eletrocardiograma. Apesar disso, a variável *ExerciseAngina* possui uma distribuição relativamente equilibrada, demonstrando que nem todos os pacientes apresentam desconforto cardíaco.

*ST\_Slope* mostra que a maioria dos pacientes apresentaram aceleração nos batimentos cardíacos quando expostas a atividades físicas, enquanto uma minoria nāo teve uma resposta muito positiva.

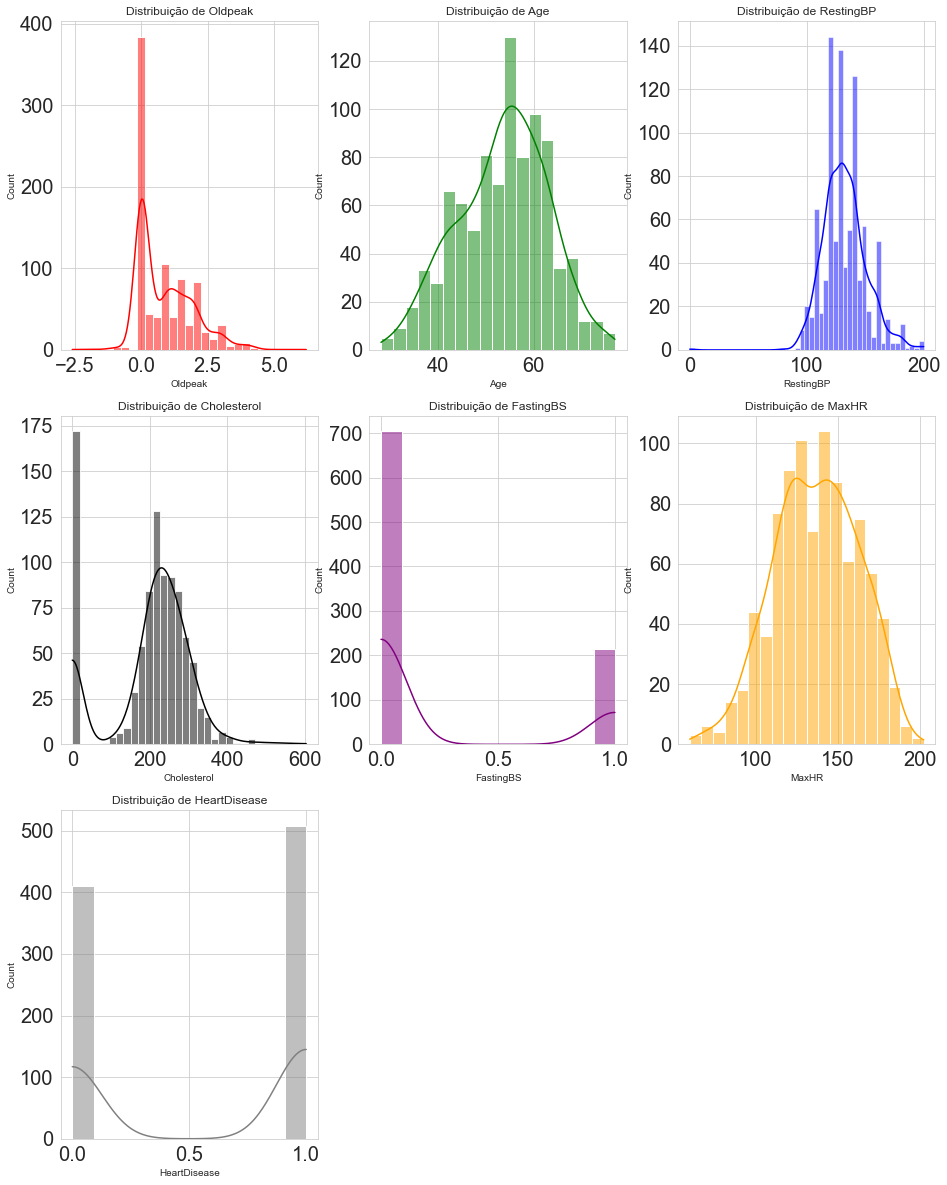


Figura 11: Distribuiçāo variáveis numéricas

Partindo para as variáveis numéricas, aparentemente não temos nenhuma distribuição normal gaussiana, mas é importante que comprovemos posteriormente aplicando testes de normalidade. Outra informaçāo bem importante é a distribuiçāo de *HeartDisease* (nosso *target*) que possui uma quantidade de pacientes com e sem insuficiência cardiaca bastante balanceada, o que acaba nos poupando tempo posteriormente com abordagens de oversampling e undersampling para balancear a nossa variável alvo.

Em relaçāo ao atributo *Cholesterol*, temos alguns valores zerados que provavelmente sāo indicativos de valores inválidos, conforme tínhamos visto anteriormente na *Figura 7: Analisando as variáveis com quebras por quartil*. O mesmo parece acontecer com *Oldpeak*, com apenas alguns valores negativos.

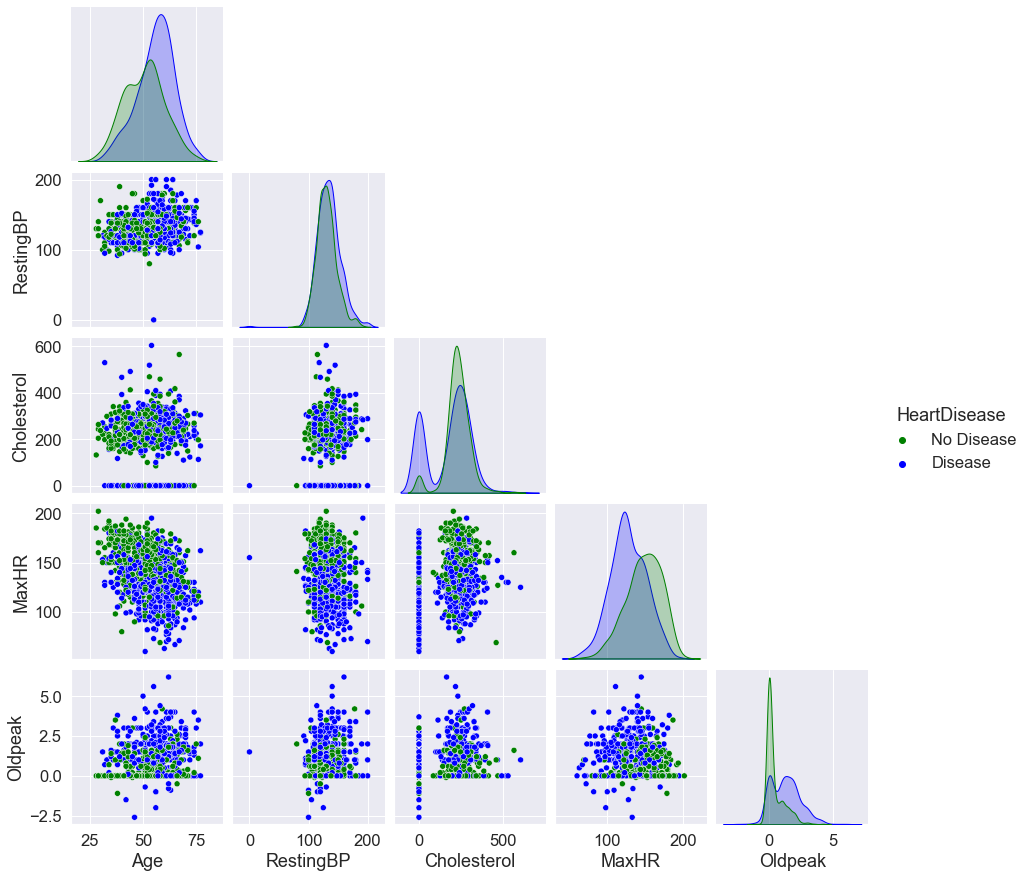


Figura 12: Distribuiçāo variáveis categóricas usando *pairplot*

Utilizando o *scatterplot*, podemos observar o relacionamento entre as variáveis numéricas em relaçāo a variável de *HeartDisease*, sendo a cor azul representada pela incidência de insuficiência cardíaca e a cor verde em casos de ausência de incidências. Apesar do *scatterplot* ser uma boa ferramenta para entendermos um pouco melhor a correlação das variáveis em alguns casos, nāo foi possível tirar muitas conclusões observando os gráficos, já que os dados estāo distribuídos de forma bastante uniforme.

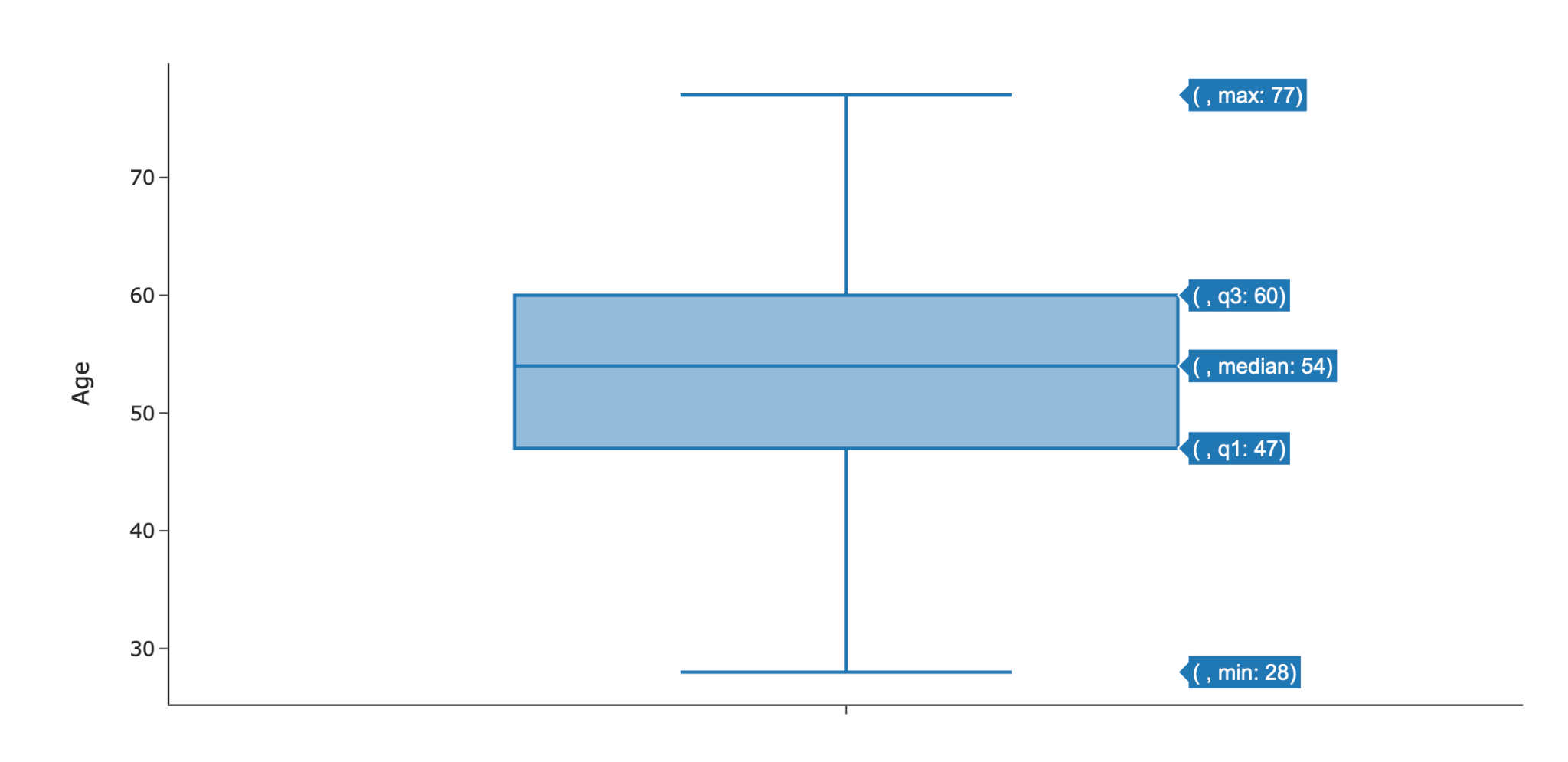


Figura 13: Boxplot variável *Age*

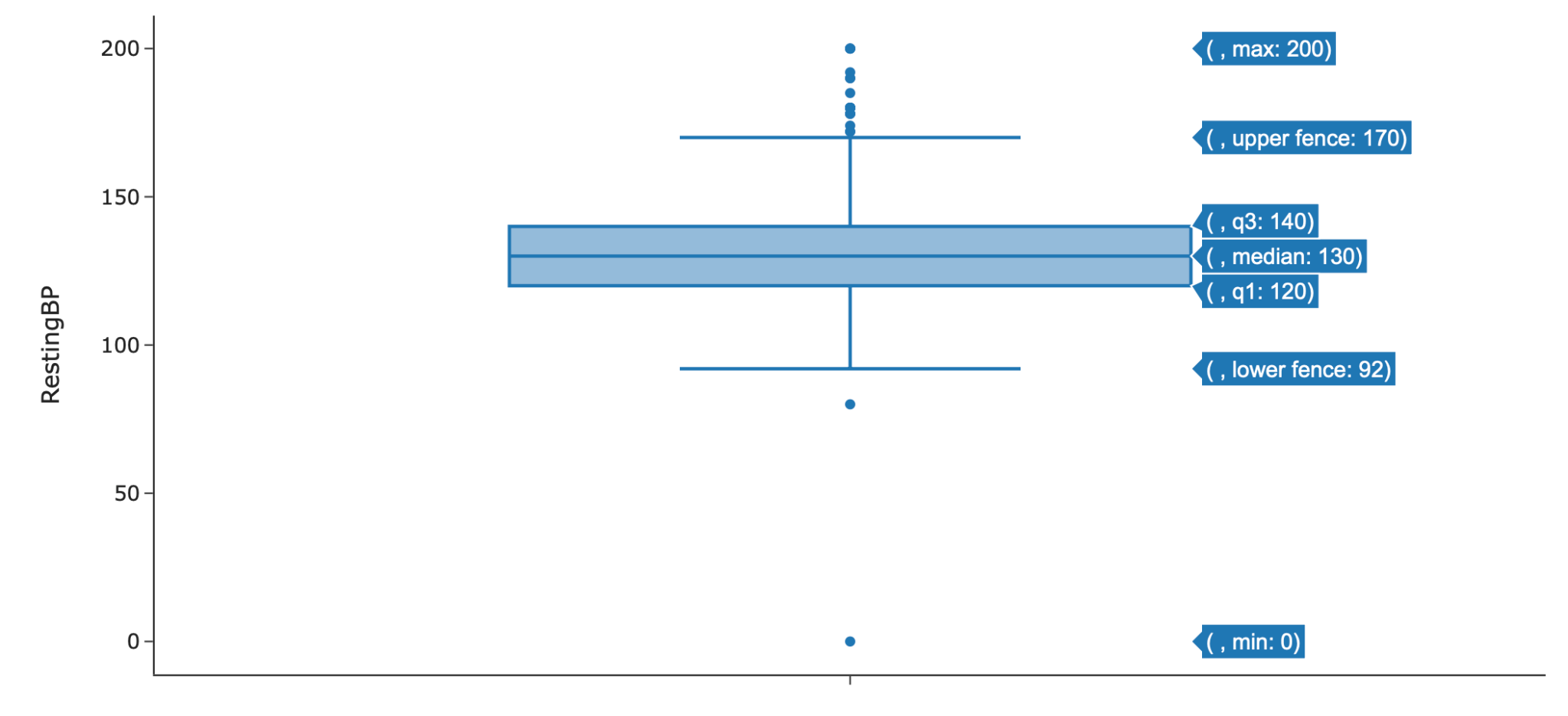


Figura 14: Boxplot variável *RestingBP*

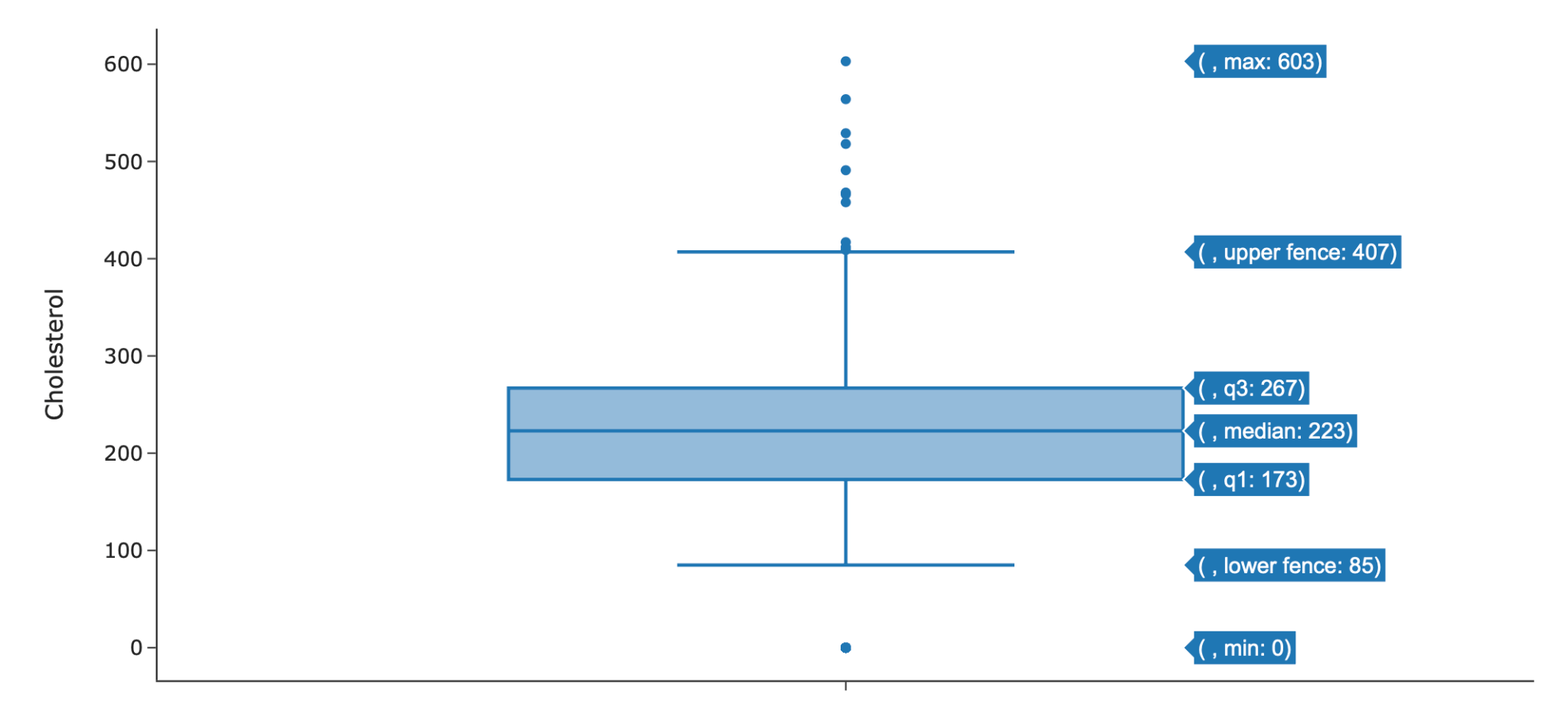


Figura 15: Boxplot variável *Cholesterol*

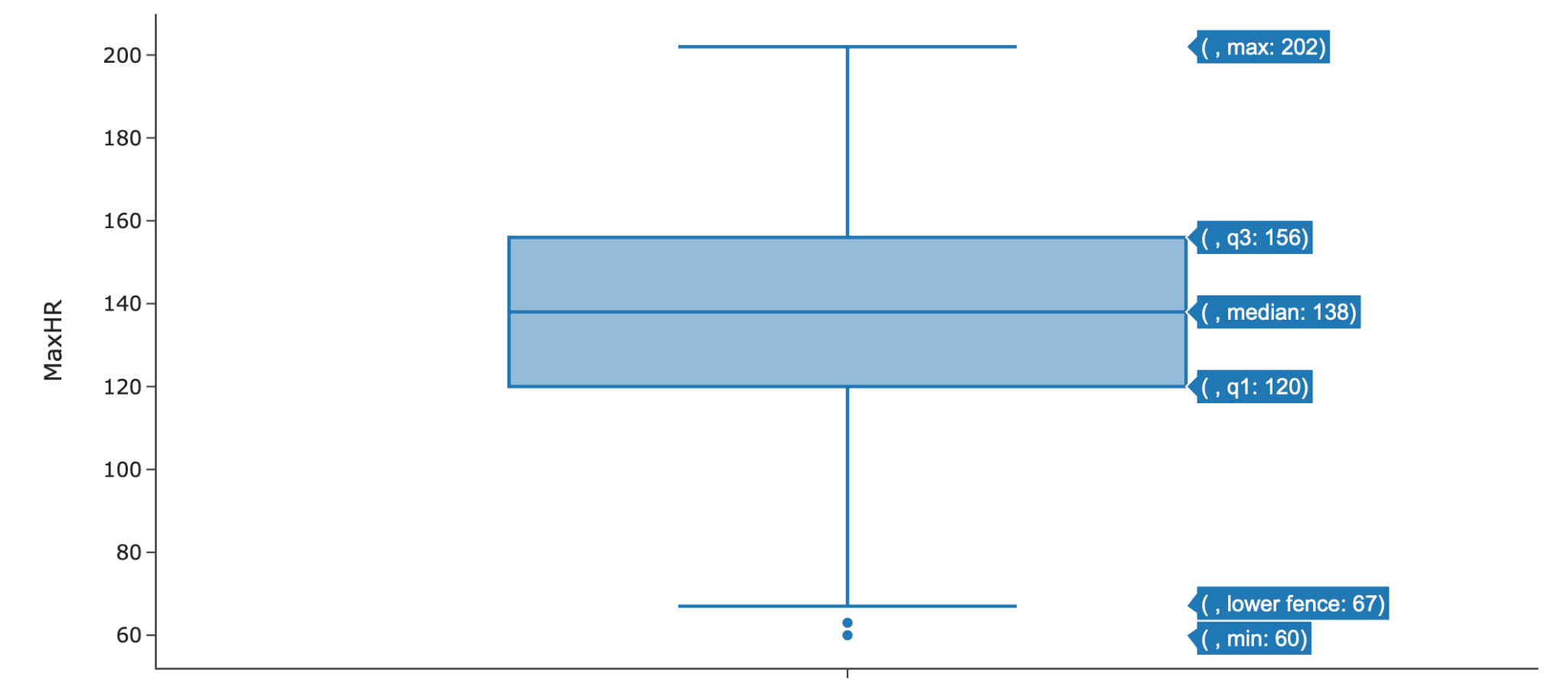


Figura 16: Boxplot variável *MaxHR*

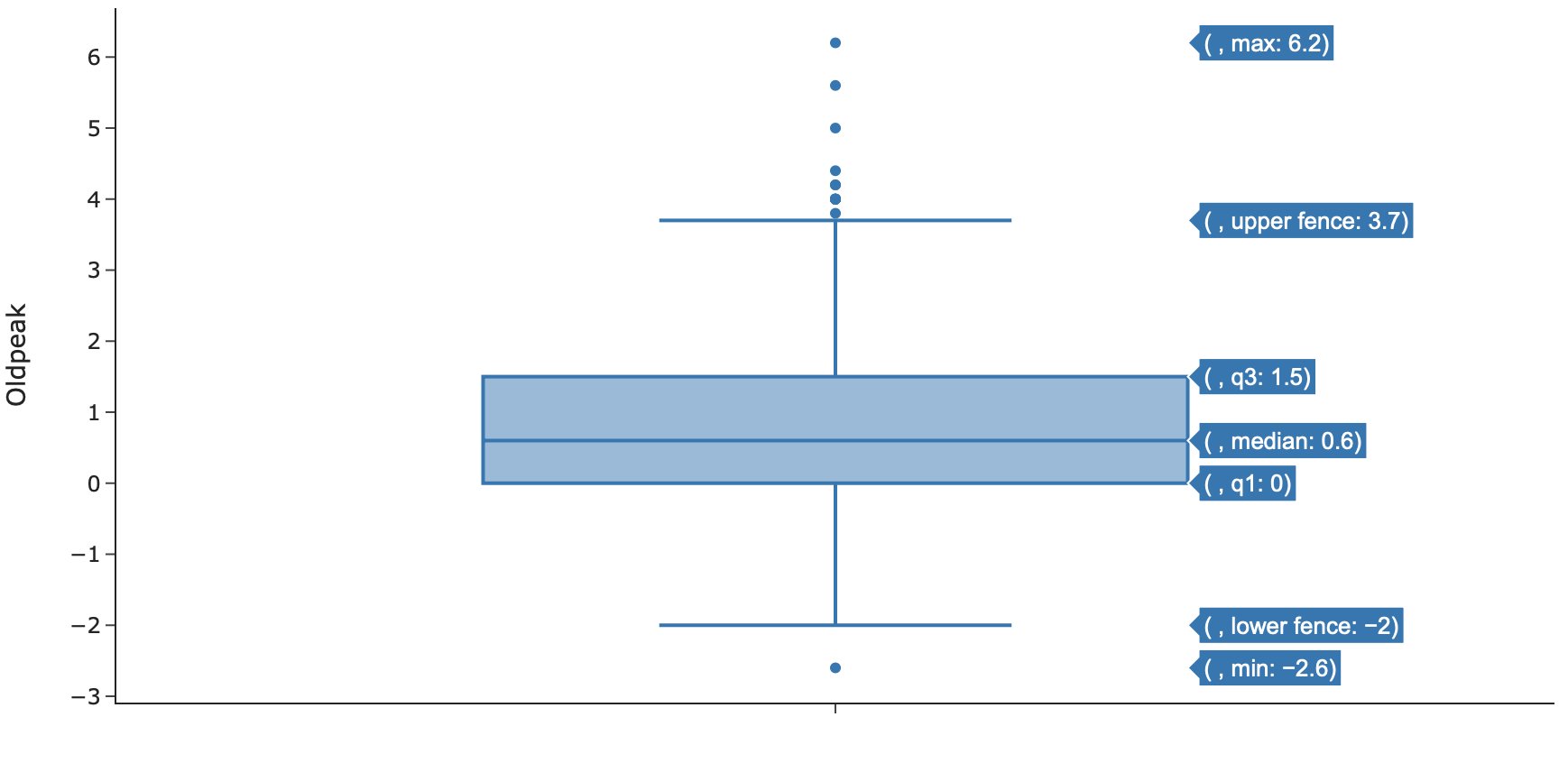


Figura 17: *Boxplot* variável *Oldpeak*

Analisando os boxplots, temos algumas variáveis com outliers, tais como: *RestingBP*, *Cholesterol, MaxHR* e *Oldpeak*. Esses outliers podem gerar ruídos nos modelos, prejudicando as predições, já que o modelo irá aprender com o comportamento apresentado nos dados. Iremos remover os valores inválidos na etapa 7 deste trabalho.



Figura 18: *Boxplot* variável *Oldpeak*

Observando as distribuições da Figura 11, conseguimos entender que os atributos não possuem uma distribuição gaussiana. Mesmo assim, é importante aplicarmos testes estatísticos, a fim de provarmos matematicamente que nossas conclusões tem fundamento. Além disso, quando vamos realizar o processo de *Normalizing* ou *Scaling* precisamos ter certeza dessas informações para não nos equivocarmos durante a modelagem.

Para realizar este teste de normalidade, foi utilizado o teste de *Kolmogorov-Smirnov* onde sāo levantadas duas hipóteses:

* H0: A variável está normalmente distribuída.
* H1: A variável não se encontra na forma normal.

Para que a hipótese alternativa (H1) seja validada, a hipótese nula (H0) deve ser rejeitada. Para isso foi aplicado um valor de alpha 5% para critério de avaliação. Neste caso, para que a hipótese nula seja rejeitada, o *p-value* deve ser menor do que o alpha estabelecido.

Segue os resultados abaixo:

| **Variável** | **P-value** | **Resultado** |
| --- | --- | --- |
| Age | 0 | Hipótese H0 rejeitada |
| Cholesterol | 0 | Hipótese H0 rejeitada |
| RestingBP | 0 | Hipótese H0 rejeitada |
| MaxHR | 0 | Hipótese H0 rejeitada |
| OldPeak | 4.2004655072318E-200 | Hipótese H0 rejeitada |

Tabela 1: Teste de *Kolmogorov-Smirnov*

Com base na tabela, é podemos afirmar que todas as variáveis contínuas tiveram a hipótese nula (H0) rejeitada. Logo, estatisticamente, podemos afirmar que nāo sāo distribuiçōes gaussianas.

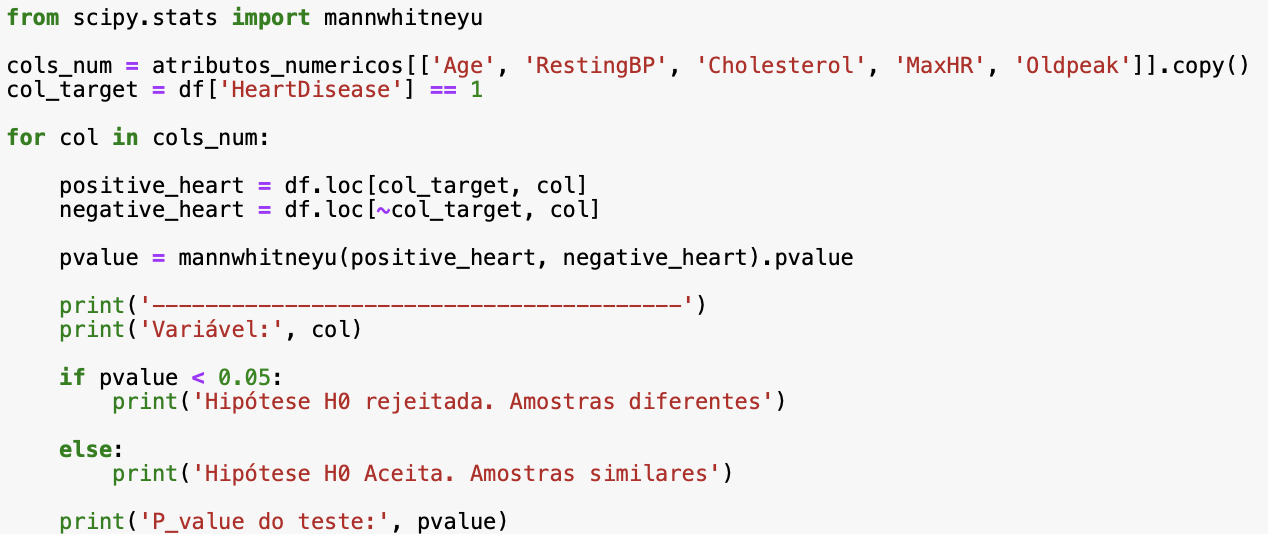


Figura 19: Teste de *Mann-Whitney*

Para analisar se existe mais de uma distribuição para a mesma população, iremos fazer o teste de Mann-Whitney. Este teste trabalha com dois grupos de amostras e tem como objetivo medir o grau de similaridade entre elas.

As hipóteses que compõem o teste são as seguintes:

* H0: As amostras são semelhantes e suas variações são decorrentes da aleatoriedade.
* H1: As amostras se demonstram diferentes e suas variações se diferenciam de acordo com o acontecimento do fato estudado.

| **Variável** | **P-value** | **Resultado** |
| --- | --- | --- |
| Age | 9.02847069514623E-19 | Hipótese H0 rejeitada |
| Cholesterol | 1.14015618177246E-05 | Hipótese H0 rejeitada |
| RestingBP | 0.000282403774686077 | Hipótese H0 rejeitada |
| MaxHR | 7.53179435979887E-35 | Hipótese H0 rejeitada |
| OldPeak | 3.3839225217194E-37 | Hipótese H0 rejeitada |

Tabela 2: Teste de *Mann-Whitney*

De acordo com o teste de *Mann-Whitney*, todas as variáveis testadas tiveram a sua hipótese nula (H0) rejeitada. Com base nisso, podemos afirmar que as amostras são estatisticamente divergentes.

Para as variáveis qualitativas. Em adição, será feito o teste de *Qui-Quadrado* para alpha < 0.05, visando testar as proporções das variáveis. Para o teste, temos:

* H0 as proporçōes são iguais;
* H1 as proporções não são iguais.

| **Variável** | **P-value** | **Resultado** |
| --- | --- | --- |
| ChestPainType@NAP | 1.86E-10 | Hipótese H0 rejeitada |
| ChestPainType@ASY | 8.63E-55 | Hipótese H0 rejeitada |
| ChestPainType@ATA | 0.131576751228143 | H0 aceita |
| ChestPainType@TA | 3.38E-37 | Hipótese H0 rejeitada |
| ExerciseAngina@N | 2.91E-50 | Hipótese H0 rejeitada |
| ExerciseAngina@Y | 2.91E-50 | Hipótese H0 rejeitada |
| RestingECG@LVH | 0.809528258440575 | H0 aceita |
| RestingECG@Normal | 0.0067906242525704 | Hipótese H0 rejeitada |
| RestingECG@ST | 0.00250729009841889 | Hipótese H0 rejeitada |
| Sex@F | 4.60E-20 | Hipótese H0 rejeitada |
| Sex@M | 4.60E-20 | Hipótese H0 rejeitada |
| ST\_Slope@Down | 0.000342178577264425 | Hipótese H0 rejeitada |
| ST\_Slope@Flat | 8.91E-63 | Hipótese H0 rejeitada |
| ST\_Slope@Up | 1.03-78 | Hipótese H0 rejeitada |

Tabela 3: Teste de *Qui-Quadrado*

Pode-se avaliar pelos gráficos e também pelos resultados do teste *Qui-Quadrado*, que todas as variáveis possuem uma diferença estatisticamente significativa em suas proporções, com exceção de *ChestPainType@ATA* e *RestingECG@LVH*.

Para analisarmos melhor a correlação entre as variáveis, utilizamos o *heatmap*. O método utilizado foi o *spearman*, já que nossas variáveis nāo seguem a distribuição gaussiana.

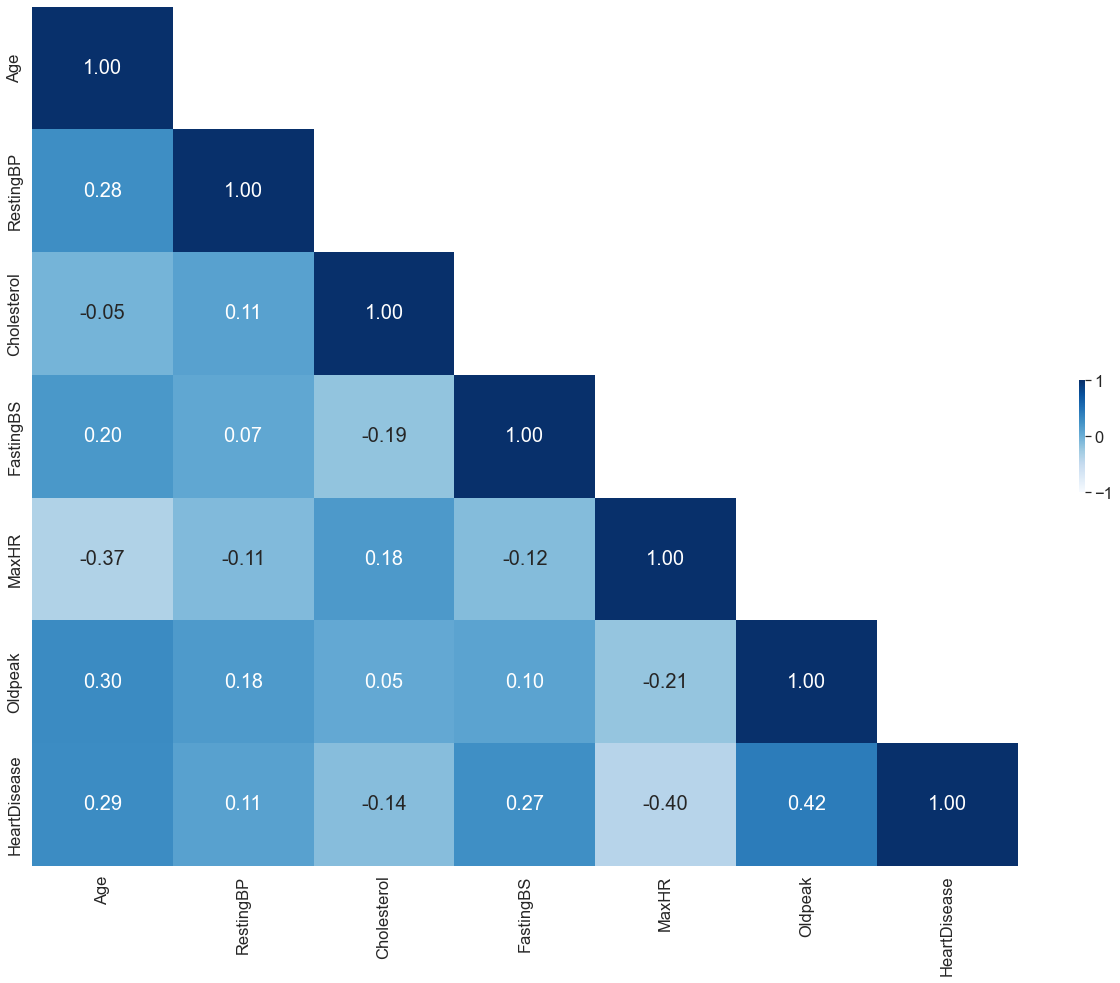


Figura 20: *Heatmap*

As variáveis que possuem mais correlaçāo em relaçāo ao nosso target sāo *MaxHR* e *Oldpeak*. As outras variáveis possuem correlaçāo abaixo de 50%, o que é considerado uma correlaçāo fraca.

# 7. Preparação dos Dados para os Modelos de Aprendizado de Máquina

Nesta etapa, vamos realizar a preparação dos dados e remover os dados inválidos. Além disso, vamos criar as variáveis para começar os testes com os modelos preditivos.

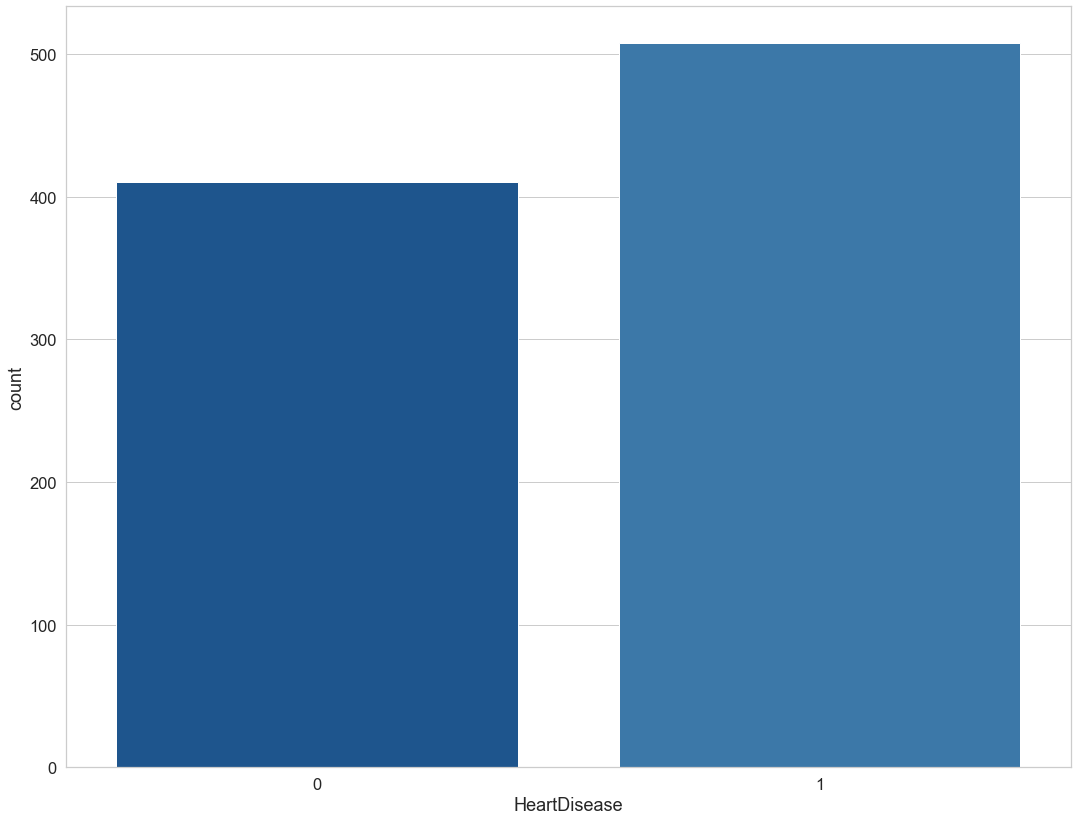


Figura 21: Verificando o balanceamento da variável *target*

Com base no gráfico acima, podemos afirmar que a nossa variável target encontra-se balanceada. Neste caso, nāo será necessário aplicar nenhuma técnica de balanceamento como *oversampling* ou *undersampling*.

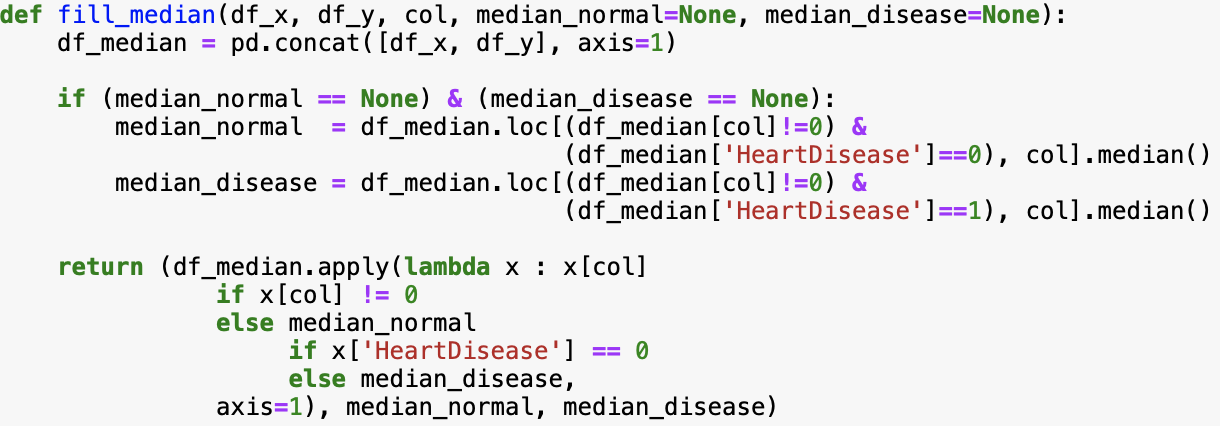


Figura 22: Método para preencher os outliers com a mediana

Como verificamos que existiam muitos dados zerados para a variável de Cholesterol, para nāo perdemos muitos dados, vamos substituir os dados zerados pela mediana. Desta forma, conseguimos continuar com a mesma quantidade de registros na base.

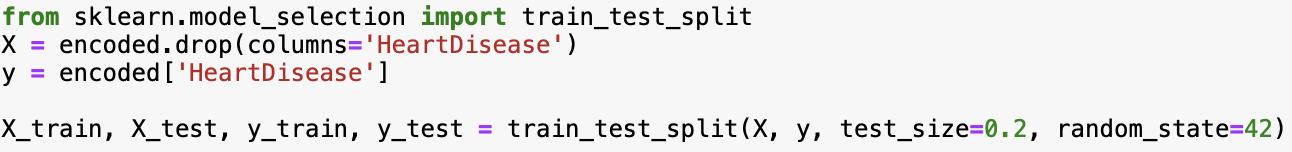


Figura 23: Separaçāo variáveis treino e teste

Nesta última parte, passamos para o método de *train\_test\_split* as variáveis tratadas anteriormente com o *OneHotEncoder* para que os atributos categóricos, em conjunto com as features de valores contínuos, sejam aplicados corretamente no modelo preditivo.

# 8. Aplicação de Modelos de Aprendizado de Máquina

Nesta seção você deve apresentar os modelos de Aprendizado de Máquina desenvolvidos no trabalho. Mostre partes do código-fonte para ilustrar a implementação de cada modelo, além do pipeline completo do processo. A escolha dos modelos deve ser adequada ao problema proposto. Embora possa ser considerado o uso de ferramentas como Weka, Knime e Orange, por exemplo, encoraja-se a implementação com linguagens como Python ou R. É importante testar mais de um tipo de algoritmo, para que resultados distintos possam ser comparados. Por exemplo, se o trabalho trata de uma classificação, modelos como Árvores de Decisão, Redes Neurais Artificiais e Support Vector Machine poderiam ser utilizados. Além disso, devem ser escolhidas e implementadas as métricas adequadas ao problema proposto, bem como os seus resultados apresentados.

# 9. Discussão dos Resultados

Nesta seção você deve relatar os resultados alcançados ao final do trabalho. Mostre os resultados das métricas adotadas, seja através de gráficos, tabelas, dentre outros, que permitam a validação do seu trabalho.

# 10. Conclusão

Nesta seção você deve apresentar um fechamento para o trabalho. É importante apresentar um breve resumo do trabalho, resgatando o problema, como foi tratado e os resultados obtidos, bem como as limitações e perspectivas (trabalhos futuros).

# 11. Links

Nesta seção você pode disponibilizar *links* para repositórios, como é o caso do GitHub, onde podem ser encontrados o seu projeto, códigos-fonte, vídeos demonstrativos, dentre outros.

# 

# 12. Referências

Harvard Health Publishing Harvard Medical School. **Heart Health**. 2022. Disponível em: <<https://www.health.harvard.edu/topics/heart-health>> Acesso em 02/06/2022

Harvard T.H. Chan School of Public Health. **Heart Disease**. 2022. Disponível em: <<https://www.hsph.harvard.edu/nutritionsource/disease-prevention/cardiovascular-disease/>> Acesso em 02/06/2022

GO AS, Mozaffarian D, Roger VL, et al. **Heart disease and stroke statistics**, 2013. National Library of Medicine. Disponível em: <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5408511/>> Acesso em 02/06/2022;

Matthews CE, George SM, Moore SC, et al. 2012. **Amount of time spent in sedentary behaviors and cause-specific mortality in US adults**. National Library of Medicine. Disponível em: <<https://pubmed.ncbi.nlm.nih.gov/22218159/>> Acesso em 02/06/2022