



## **Estatística Descritiva**

Disciplina: Delineamento Experimental

Mestrado: Inteligência Artificial e Ciências de Dados

Ano Lectivo: 2025/2026

Docente: Professora Doutora Maria Manuela Melo Oliveira

Professor Doutor Nuno de Almeida Ribeiro

Relatório realizado por:

- Madalena Marques m68201  
15 de novembro 2025

## Índice

|   |           |
|---|-----------|
| <b>1.Introdução .....</b>   | <b>3</b>  |
| <b>2.Material e Métodos .....</b>   | <b>3</b>  |
| <b>3.Metodologia .....</b>  | <b>3</b>  |
| <b>4.Resultados/Discussão de Resultados .....</b>   | <b>4</b>  |
| 4.1. Classificação de variáveis .....   | 4         |
| 4.2. Medidas Estatística Descritiva .....   | 4         |
| 4.3. Dados Quantitativos .....  | 5         |
| 4.3.1.IMC (Índice Massa Corporal) .....   | 5         |
| 4.3.2. Idade .....  | 6         |
| 4.3.3. Número de cigarros por dia .....   | 6         |
| 4.4. Dados Quantitativos Ordinais .....   | 7         |
| 4.4.1. Idade em 4 categorias.....   | 7         |
| 4.5. Tabelas de Contingência entre variáveis .....  | 8         |
| 4.5.1. Tabelas de Contingência entre o género das pessoas e número de fumadores .....         | 8         |
| 4.5.2. Tabelas de Contingência entre a idade com quatro categorias e doenças associadas ..... | 8         |
| 4.5.3. Tabelas de Contingência entre género e doenças associadas ao consumo de tabaco .....   | 9         |
| 4.6. Correlações entre as variáveis .....   | 9         |
| 4.6.1. Número de pessoas que tem a doença associadas ao consumo de tabaco .....               | 9         |
| 4.6.2. Relação entre Índice Massa Corporal e número de cigarros .....                         | 10        |
| 4.6.3. Relação da Idade com o estado de saúde das pessoas inseridas neste estudo.....         | 11        |
| 4.6.4. Relação entre a Idade e o Índice de massa corporal .....                               | 12        |
| <b>5. Conclusão .....</b>   | <b>13</b> |
| <b>6. Referências Bibliográficas .....</b>  | <b>13</b> |

## 1.Introdução

O presente relatório tem como objetivo aplicar técnicas de análise estatísticas. Esta base de dados, têm como objetivo estabelecer uma relação entre as pessoas que fumam e a implicação para a sua saúde. Para tal é importante referir que o tabagismo tem como base a dependência de consumo de produtos de tabaco, causada pela presença da nicotina. Além disso, têm um risco associado para saúde humana e principal causa de morte prematura na união europeia, contribuindo para desenvolvimento de doenças nomeadamente, doenças respiratórias e cerebrais.

## 2.Material e Métodos

A base de dados *tobacco* é formada por 10 variáveis e 1000 observações, entre as quais o género, a idade, a idade com 4 categorias, índice de massa corporal, se as pessoas fumam ou não, número cigarros por dia, se têm doenças associadas e tiverem qual é doença associada e o índice de peso. Além disso, foram identificados valores omissos (NA) em diversas variáveis cujos estes valores foram convertidos e alguns foram excluídos para garantir a consistências nos resultados.

## 3.Metodologia

A análise descritiva dos dados foi feita através do programa de software estatístico *R studio* cujo output e os resultados serão apresentados ao longo do tópico 4 sobre forma de tabela de frequência e gráficos. Além disso, foram usados diversos pacotes do R para facilitar a manipulação de dados e a criação dos gráficos entre os quais destaca-se os seguintes:

- *dplyr* para manipulação de dados;
- *ggplot2* para a criação de gráficos;
- *skimr* para fornecer medidas de resumo;
- *psych* para extrair dados estatísticos. <sup>[2]</sup>

## 4.Resultados/Discussão de Resultados

### 4.1. Classificação de variáveis

As variáveis inseridas nesta base dados estão classificadas nas seguintes categorias as qualitativas, em que podem ser nominais com uma sequência não ordenável ou ordinal com uma ordem natural, e além disso também podem ser quantitativas, contínuas em que podem medir ou discretas em que se podem contar. A tabela 1 mostra as classificações das variáveis desta base de dados.

*Tabela 1 Classificação de variáveis da base dados*

| Variável                     | Tipo         | Sub-tipo |
|------------------------------|--------------|----------|
| <b>género</b>                | Qualitativa  | nominal  |
| <b>idade</b>                 | Quantitativa | continua |
| <b>Idade de 4 categorias</b> | Qualitativa  | Ordinal  |
| <b>BMI</b>                   | Quantitativa | continua |
| <b>fumadores</b>             | Qualitativa  | nominal  |
| <b>cigs.per.day</b>          | Quantitativa | continua |
| <b>diseased</b>              | Qualitativa  | nominal  |
| <b>disease</b>               | Qualitativa  | nominal  |
| <b>samp.wgts</b>             | Qualitativa  | Ordinal  |

### 4.2. Medidas Estatística Descritiva

Como observado na tabela 2, que apresenta medidas estatísticas descritivas gerais das 10 variáveis que compõem esta base dados. De realçar que género sendo uma variável qualitativa nominal não é possível aferir medidas estatísticas descritivas gerais. Com a análise da tabela 2 podemos observar que a variável idade tem maior média e o índice massa corporal.

Tabela 2- Medidas Estatísticas Descritivas da base dados

|              | vars | n    | mean     | sd     | median | trimmed | mad    | min   | max     | range  | skew  |
|--------------|------|------|----------|--------|--------|---------|--------|-------|---------|--------|-------|
| Column1      | 1    | 1000 | 500.50   | 288.82 | 500.50 | 500.50  | 370.65 | 1.00  | 1000.00 | 999.00 | 0.00  |
| gender*      | 2    | 978  | 1.50     | 0.50   | 1.50   | 1.50    | 0.74   | 1.00  | 2.00    | 1.00   | 0.00  |
| age          | 3    | 975  | 49.60    | 18.29  | 50.00  | 49.70   | 23.72  | 18.00 | 80.00   | 62.00  | -0.04 |
| age.gr*      | 4    | 975  | 2.39     | 1.05   | 2.00   | 2.36    | 1.48   | 1.00  | 4.00    | 3.00   | 0.03  |
| BMI          | 5    | 974  | 25.73    | 4.49   | 25.62  | 25.71   | 4.18   | 8.83  | 39.44   | 30.61  | 0.02  |
| smoker*      | 6    | 1000 | 1.30     | 0.46   | 1.00   | 1.25    | 0.00   | 1.00  | 2.00    | 1.00   | 0.88  |
| cigs.per.day | 7    | 965  | 6.78     | 11.88  | 0.00   | 4.12    | 0.00   | 0.00  | 40.00   | 40.00  | 1.54  |
| diseased*    | 8    | 1000 | 1.22     | 0.42   | 1.00   | 1.16    | 0.00   | 1.00  | 2.00    | 1.00   | 1.32  |
| disease*     | 9    | 222  | 6.06     | 3.68   | 6.00   | 5.90    | 4.45   | 1.00  | 13.00   | 12.00  | 0.21  |
| samp.wgts    | 10   | 1000 | 1.00     | 0.08   | 1.04   | 1.01    | 0.01   | 0.86  | 1.06    | 0.20   | -1.04 |
|              |      |      | kurtosis | se     |        |         |        |       |         |        |       |
| Column1      |      |      | -1.20    | 9.13   |        |         |        |       |         |        |       |
| gender*      |      |      | -2.00    | 0.02   |        |         |        |       |         |        |       |
| age          |      |      | -1.26    | 0.59   |        |         |        |       |         |        |       |
| age.gr*      |      |      | -1.22    | 0.03   |        |         |        |       |         |        |       |
| BMI          |      |      | 0.26     | 0.14   |        |         |        |       |         |        |       |
| smoker*      |      |      | -1.22    | 0.01   |        |         |        |       |         |        |       |
| cigs.per.day |      |      | 0.90     | 0.38   |        |         |        |       |         |        |       |
| diseased*    |      |      | -0.25    | 0.01   |        |         |        |       |         |        |       |
| disease*     |      |      | -1.03    | 0.25   |        |         |        |       |         |        |       |
| samp.wgts    |      |      | -0.90    | 0.00   |        |         |        |       |         |        |       |

### 4.3. Dados Quantitativos

#### 4.3.1.IMC (Índice Massa Corporal)

Segundo OMS, índice de massa corporal (IMC), é um indicador internacional usado para identificar, rapidamente situação de déficit de peso, excesso de peso ou obesidade, usando valores de referência em que se considera que uma pessoa tem um peso baixo quando o valor de IMC dá inferior 18.5%, se IMC estiver entre 18.5% e 24.9% é normal, quando está com valores de 25% e 29.9% a pessoa tem excesso de peso e por últimos valores superiores a 30% a pessoa têm obesidade [3]. Para realizar a tabela de frequência, que está representada na tabela 3, foi usado estes valores como referência para dividir em subcategorias e no gráfico 1 está representada o histograma da distribuição do IMC por frequência.

Tabela 3- Tabela de Frequência agrupada em classes

| BMI_class     | Frequência | Porcentagem |
|---------------|------------|-------------|
| <fct>         | <int>      | <dbl>       |
| Baixo Peso    | 77         | 7.7         |
| Normal        | 372        | 37.2        |
| Sobrepeso     | 374        | 37.4        |
| Obesidade I   | 154        | 15.4        |
| Obesidade II+ | 23         | 2.3         |

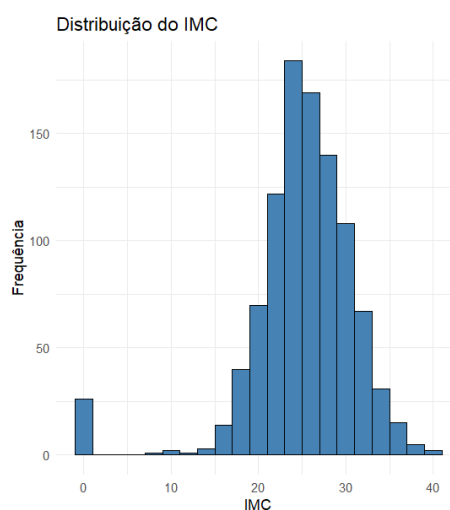


Gráfico 1- Histograma da variável IMC

Após análise da tabela 3 e gráfico 1 pode-se concluir que a maioria das pessoas inseridas neste estudo têm índice massa corporal considerado normal ou com excesso de peso segundo valores de referência.

#### 4.3.2. Idade

A faixa etária das pessoas inseridas neste estudo está compreendida entre 18 e os 80 anos e com maior incidência na faixa etária dos 60 e os 80 anos como está representado na tabela 4 em forma de tabela de frequência agrupadas em classes e no gráfico 2 em forma de gráficos de barras.

Tabela 4-Tabela de frequências agrupadas em classes

| age_class | Frequência | Porcentagem |
|-----------|------------|-------------|
| <fct>     | <int>      | <dbl>       |
| [0,20)    | 47         | 4.7         |
| [20,40)   | 313        | 31.3        |
| [40,60)   | 290        | 29          |
| [60,80)   | 340        | 34          |
| [80,100)  | 10         | 1           |

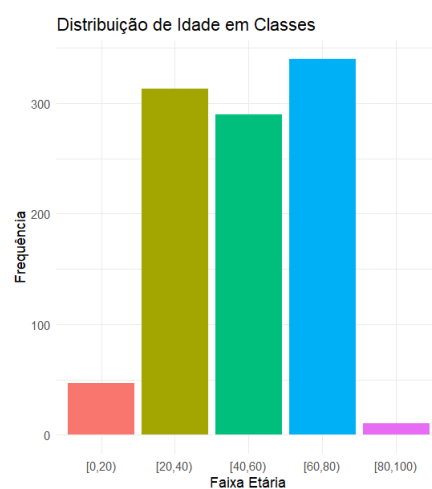


Gráfico 2- Gráfico de barras da Idade

#### 4.3.3. Número de cigarros por dia

Neste estudo foi feita a análise de quantos cigarros as pessoas consumiam por dia, em que podemos concluir que a maioria consome em média 7 cigarros por dia como está representado na tabela 5 em forma de tabela de frequência e no gráfico 3 em forma de gráfico de barras.

Tabela 5- Tabela de Frequência agrupadas por classes

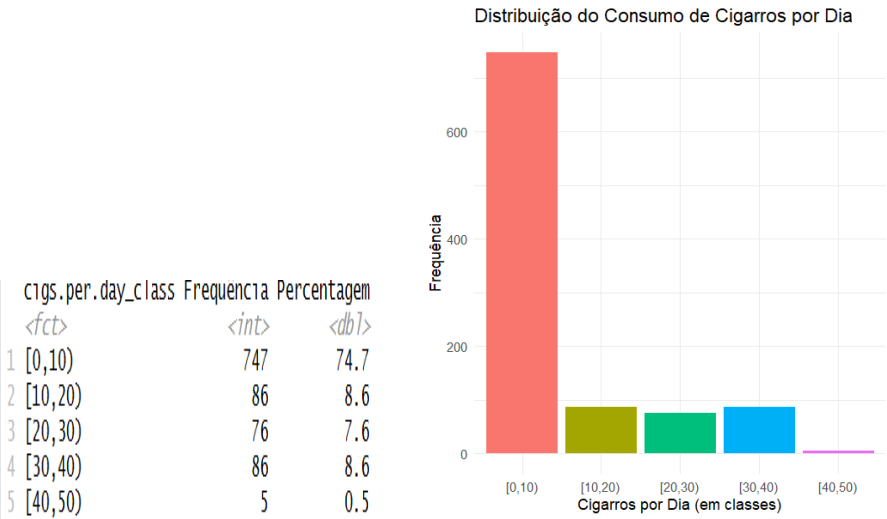


Gráfico 3-Gráfico de barras com número de cigarros por dia

4.4. Dados Quantitativos Ordinais

4.4.1. Idade em 4 categorias

A tabela 6 mostra a tabela de frequência das idades divididas em quatro faixas etárias, cujo a primeira está dentro dos 18-34 anos, a segunda categoria 35-50 anos, a terceira categoria 51-70 anos, e por último, a faixa dos maiores de 70 anos e no gráfico 4 está representada a distribuições das idades médias por grupo de faixa etária. Após análise, podemos concluir que a faixa etária dos 51-70 anos têm maior percentagem de fumadores.

Tabela 6-Tabela frequência agrupada em classes

|    |       |       |       |      |
|----|-------|-------|-------|------|
| 0  | 18-34 | 35-50 | 51-70 | 71 + |
| 25 | 258   | 241   | 317   | 159  |

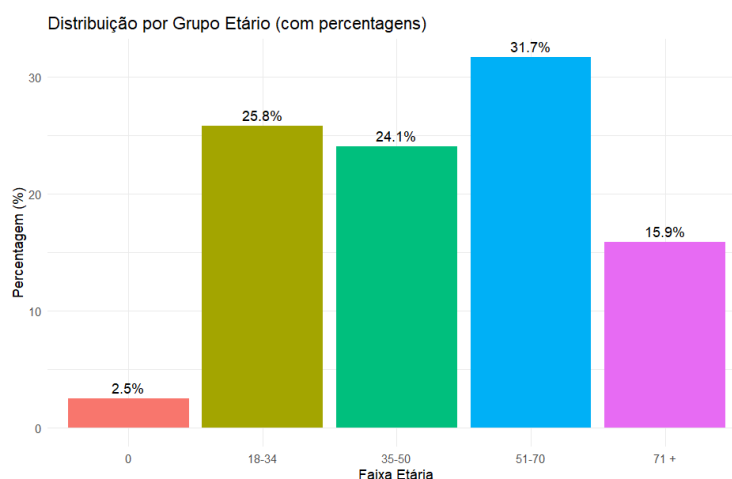


Gráfico 4- Histograma da idade em 4 categorias

## 4.5. Tabelas de Contingência entre variáveis

### 4.5.1. Tabelas de Contingência entre o género das pessoas e número de fumadores

A tabela 7 mostra a relação entre o género e número de pessoas que fumam sobre a forma de tabela de contingência, com base na análise desta tabela podemos concluir que a maioritariamente as pessoas envolvidas neste estudo não são fumadoras e sexo prevalente é masculino, ainda de referi que o sexo prevalente na população que fuma é sexo feminino.

Tabela 7-Tabela de Contingência entre género das pessoas envolvidas neste estudo e número de fumadores

|   | No  | Yes |
|---|-----|-----|
| F | 342 | 147 |
| M | 346 | 143 |

### 4.5.2. Tabelas de Contingência entre a idade com quatro categorias e doenças associadas

A tabela 8 mostra a relação entre a idade com quatro categorias e doenças associadas ao tabagismo, sobre forma de tabela de contingência, com base na análise desta tabela podemos concluir que a classe da idade onde prevalece a maioria das doenças é na faixa etária dos 51 aos 70 anos e a prevalência maior de doenças associadas ao tabagismo é a hipertensão.

*Tabela 8-Tabela de Contingência entre a idade com quatro categorias e doenças associadas*

|       | Cancer | Cholesterol | Diabetes | Digestive | Hearing | Heart | Hypertension | Hypotension | Musculoskeletal |   |
|-------|--------|-------------|----------|-----------|---------|-------|--------------|-------------|-----------------|---|
| 18-34 | 11     |             | 4        | 2         | 5       | 3     | 5            | 6           | 1               | 2 |
| 35-50 | 9      |             | 5        | 2         | 0       | 4     | 4            | 9           | 2               | 5 |
| 51-70 | 7      |             | 7        | 6         | 4       | 3     | 6            | 12          | 6               | 8 |
| 71 +  | 7      |             | 3        | 4         | 2       | 3     | 4            | 8           | 2               | 4 |

|       | Neurological | Other | Pulmonary | Vision |
|-------|--------------|-------|-----------|--------|
| 18-34 | 2            | 0     | 4         | 1      |
| 35-50 | 2            | 1     | 3         | 2      |
| 51-70 | 2            | 1     | 8         | 3      |
| 71 +  | 3            | 0     | 5         | 3      |

### 4.5.3. Tabelas de Contingência entre género e doenças associadas ao consumo de tabaco

A tabela 9 mostra a relação entre género e doenças associadas ao tabagismo, sobre a forma de tabela de contingência, com base na análise desta tabela podemos concluir que as prevalências de doenças associadas ao consumo de tabaco têm maior incidência no sexo masculino e o cancro representa a maioria das doenças associadas.

*Tabela 9-Tabela de Contingência entre género e doenças associadas ao consumo de tabaco*

|   | Cancer | Cholesterol | Diabetes | Digestive | Hearing | Heart | Hypertension | Hypotension | Musculoskeletal |
|---|--------|-------------|----------|-----------|---------|-------|--------------|-------------|-----------------|
| F | 16     | 10          | 8        | 5         | 5       | 9     | 18           | 7           | 8               |
| M | 18     | 11          | 5        | 7         | 9       | 11    | 17           | 4           | 10              |

|   | Neurological | Other | Pulmonary | Vision |
|---|--------------|-------|-----------|--------|
| F | 7            | 1     | 9         | 6      |
| M | 3            | 1     | 11        | 3      |

## 4.6. Correlações entre as variáveis

### 4.6.1. Número de pessoas que tem a doença associadas ao consumo de tabaco

Neste estudo foram avaliados as diversas doenças associadas e o número de pessoas que têm doenças associadas. As doenças associadas ao tabaco e avaliadas nesta base de dados são as seguintes o cancro, o colesterol, os diabetes, problemas de audição, coração, hipertensão, problema músculo-esquelético, neurológico, pulmões e a visão. Como já foi referido, também foi avaliado o número de pessoas com doenças, cujo resultado está representado no gráfico 5 sobre a forma de gráfico de barras. Após a análise do gráfico podemos concluir a maior parte das pessoas inseridas neste estudo não têm doenças associadas.

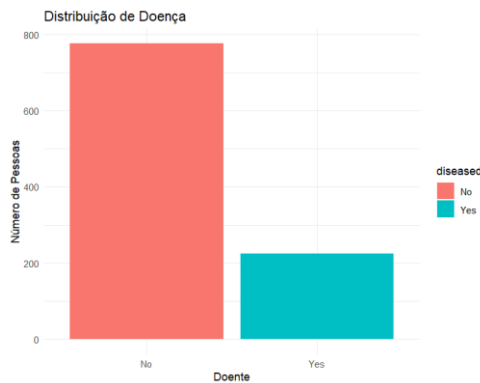


Gráfico 5- Número de pessoas que tem doença associadas

#### 4.6.2. Relação entre Índice Massa Corporal e número de cigarros

Foi testado a normalidade entre estas duas variáveis quantitativas, usando o teste de normalidade de *shapiro-wilk*, cujos resultado deste teste está representado na figura 1. Com base na análise, é possível concluir que a distribuição destas duas variáveis não é normal, ou seja, rejeita-se a normalidade, uma vez que o *p-value* deu inferior a 0.05.

Com base no resultado deste teste foi usado a correlação de *spearman* para relacionar estas duas variáveis, cujo valor 0.009624684. O que pode se concluir que não existe nenhuma relação entre índice de massa corporal e o número de cigarros por dia das pessoas que fumam. Adicionalmente foi feito um gráfico de dispersão entre o número de cigarros por dia e o índice de massa corporal, que está representado no gráfico 6.

```
> #Testar Normalidade entre estas 2 variaveis
> shapiro.test(dadost$cigs.per.day)

Shapiro-Wilk normality test

data: dadost$cigs.per.day
W = 0.6129, p-value < 2.2e-16

> shapiro.test(dadost$BMI)

Shapiro-Wilk normality test

data: dadost$BMI
W = 0.86822, p-value < 2.2e-16
```

Figura 1-Teste de shapiro-wilk para testar normalidade destas duas variáveis

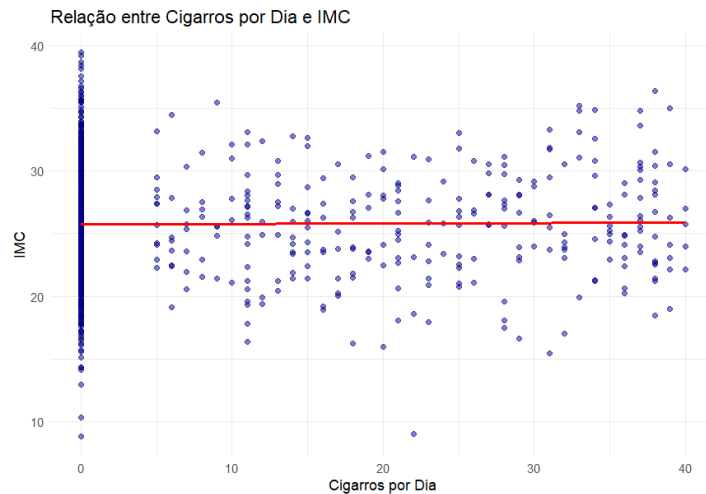


Gráfico 6-Gráfico de dispersão entre número de cigarros por dia e IMC

#### 4.6.3. Relação da Idade com o estado de saúde das pessoas inseridas neste estudo

O gráfico 7 mostra a relação da idade das pessoas inseridas neste estudo e as doenças associadas em forma de gráfico de dispersão. Em que se pode observar que as pessoas que tem doenças associadas têm uma idade avançada situada na faixa etária dos 60 aos 80 anos, além disso, têm em média de idades 51.4 anos. O que leva a concluir que a idade tem um fator relevante no aparecimento de doenças ou não. A tabela 10 mostra a relação da idade com estado de saúde das pessoas inseridas.

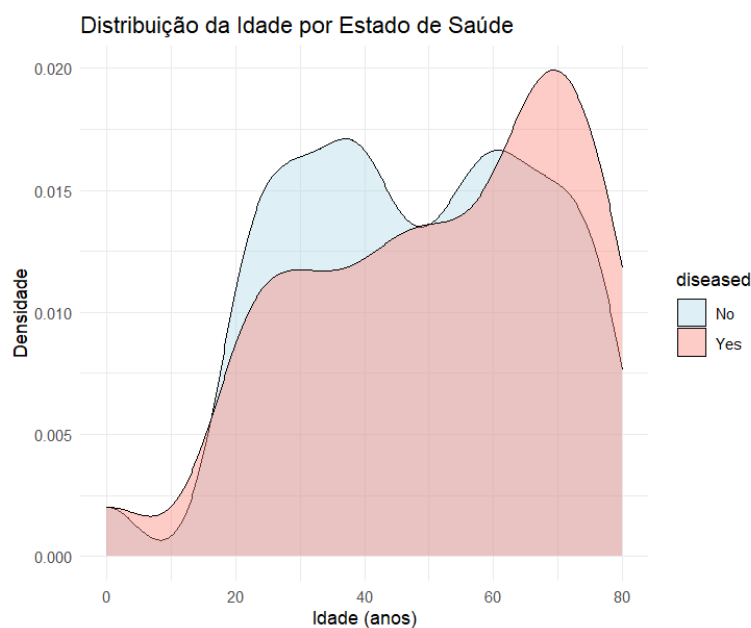


Gráfico 7- Gráfico de dispersão com relação da idade com estado de pessoas inseridas neste estudo

Tabela 10- Tabela de frequência associada à idade com estado de saúde das pessoas inseridas neste estudo

|   | diseased | Média_Idade | Mediana_Idade | Desvio_Padrão | n     |
|---|----------|-------------|---------------|---------------|-------|
|   | <fct>    | <dbl>       | <dbl>         | <dbl>         | <int> |
| 1 | No       | 47.5        | 48            | 19.3          | 776   |
| 2 | Yes      | 51.4        | 53.5          | 20.7          | 224   |

#### 4.6.4. Relação entre a Idade e o Índice de massa corporal

Foi testado a normalidade entre estas duas variáveis quantitativas, usando o teste de normalidade de *shapiro-wilk*, cujos resultado deste teste está representado na figura 2. Com base na análise, é possível concluir que a distribuição destas duas variáveis não é normal, ou seja, rejeita-se a normalidade, uma vez que o *p-value* deu inferior a 0.05.

Com base no resultado deste teste foi usado a correlação de *spearman* para relacionar estas duas variáveis, cujo valor foi 0.3021534. O que leva concluir que o índice de massa corporal tende a aumentar com idade.

Adicionalmente foi feito o gráfico de dispersão entre a idade das pessoas inseridas neste estudo e índice massa corporal das mesmas que está representado no gráfico 8.

```
> shapiro.test(dadost$age)

      Shapiro-Wilk normality test

data:  dadost$age
W = 0.96119, p-value = 1.197e-15

> shapiro.test(dadost$BMI)

      Shapiro-Wilk normality test

data:  dadost$BMI
W = 0.86822, p-value < 2.2e-16
```

Figura 2-Teste de Shapiro-wilk para testar normalidade entre estas duas variáveis

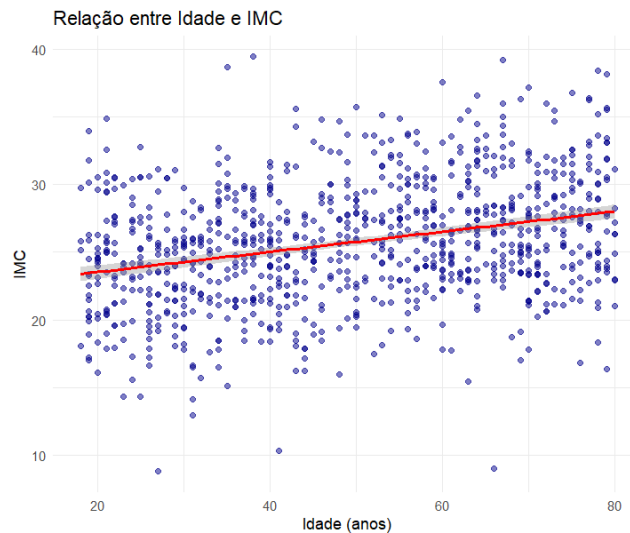


Gráfico 8- Gráfico de dispersão entre a idade e IMC

## 5. Conclusão

Ao longo desta análise descritiva foi demonstrado que o tabagismo é o fator relevante associado à ocorrência de doenças nesta base de dados. Como já foi referido anteriormente, há um risco maior associado a pessoas que fumam do que os indivíduos que não fumam em relação ao desenvolvimento de doenças associadas ao seu consumo. Para além de que, os indivíduos na faixa etária dos 60 aos 80 anos e com maior índice de massa corporal, inserido na classe do excesso de peso, contribuem também para o aumento do risco de desenvolverem doenças associadas ao tabagismo.

## 6. Referências Bibliográficas

- [1] SNS24 | Tabagismo. (n.d.). Retrieved October 24, 2025, from <https://www.sns24.gov.pt/pt/tema/dependencias/tabagismo/>
- [2] Maindonald, J. e Brown, W.J. (2003), *Data Analysis and Graphics using R*, Cambridge University Press
- [3] Body mass index (BMI). (n.d.). Retrieved October 25, 2025, from <https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/body-mass-index>