Atlântico Academy Bootcamp

Barbara Oliveira
Carlos André S. Monteiro
Carlos Augusto
Hugo Silveira Sousa
Larissa A. B.
Luiz Henrique
Matheus Amorim Constancio
Rafael Galdino da Silva

Professor: Madson Dias
Monitor: Aislan S. F.

ANÁLISE EXPLORATÓRIA DE DADOS

Sumário

Configurações iniciais

- Carregamento necessário dos pacotes
- Coleta dos dados
- Pré-descrição dos dados
- Conecimento dos dados em si
- Sonho de Todos os Cientistas de Dados!!

Processo conhecido como:

1. Pré-tratamento
2. Visualização de Dados!

Aqui vai um exemplo! Representação Gráfica dos Dados

Carregamento necessários dos pacotes

Importações das bibliotecas

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from pathlib import Path
from IPython.display import display, HTML
import numpy as np
from src.data import visualize, prepare
```

Definição das cores dos gráficos

```
colors = ["20B2AA", "B22028"]
sns.set_theme(style = "whitegrid")
sns.set_palette(sns.color_palette(colors))
```

Variáveis, Significado e Tipos

```
data_path = Path('../data/external/dicionario.csv')
df_dict = pd.read_csv(data_path, sep=';') #Obtendo o dataset
df_dict
```

Variáveis, Significado e Tipos

Variáveis

(GENDER, AGE,SMOKING, YELLOW_FINGERS, ANXIETY, PEER_PRESSURE, CHRONIC, DISEASE, FATIGUE, ALLERGY, ..., SWALLOWING, DIFFICULTY, CHEST PAIN, LUNG_CANCER).

Significado

GENDER:Indica o genêro do paciente, AGE: Indica a idade do paciente, SMOKING: Indica se o paciente é fumante ...

Tipos

Nominal, Discreta

Grupos e Variáveis

Sintomas

(SMOKING, YELLOW_FINGERS, ANXIETY, PEER_PRESSURE, CHRONIC DISEASE, FATIGUE, ALLERGY, WHEEZING, ALCOHOL CONSUMING, COUGHING, SHORTNESS OF BREATH, SWALLOWING DIFFICULTY, CHEST PAIN).

Identificação

(GENDER, AGE)

Nota: LUNG_CANCER não foi incluída em nenhum grupo pois é a variável alvo.

Conjunto de Dados

```
data_path = Path('../data/raw/data.csv')
df = (
        pd
        .read_csv(data_path)
        .rename(columns='FATIGUE ': 'FATIGUE', 'ALLERGY ': 'AL-
LERGY')
        #Para manter a consistência do nome das colunas no
        #dicionário de dados
)        #Obtendo o dataset
df
```

Conjunto dos Dados

CHRO-DISE	FATIG	ALLERGY	...	CHST_PAIN	LNG_CANCR
1	2	1	...	2	YES
2	2	2	...	2	YES
1	2	1	...	2	NO
1	1	1	...	2	NO
1	1	1	...	1	NO
.	.	.	...	.	.
.	.	.	...	.	.
.	.	.	...	.	.
1	2	2	...	2	YES
1	2	2	...	1	YES

Tabela: Conjunto de dados.

Obtendo Informações dos dados

df.info()

#	Column	Non-Null	Count	Dtype
0	GENDER	309	non-null	object
1	AGE	309	non-null	int64
2	SMOKING	309	non-null	int64
3	YELLOW_FINGERS	309	non-null	int64
4	ANXIETY	309	non-null	int64
.	.	.	...	.
.				
.	.	.	...	.
.				
.	.	.	...	.
.				
14	CHEST PAIN	309	non-null	int64
15	LUNG_CANCER	309	non-null	object

Tabela: dtypes: int64(14), object(2)
memory usage: 38.8+ KB

Verificação de dados faltantes

df.isnull().sum()

Variável	Dados Faltantes
GENDER	0
AGE	0
SMOKING	0
YELLOW_FINGERS	0
ANXIETY	0
.	.
.	.
.	.
CHEST PAIN	0
LUNG_CANCER	0

Tabela: Tabela de Dados faltantes

Verificação de preenchimento das linhas do dataframe

```
uniques = 'Variavel': df_dict['Variavel'], 'Valores': []
for index, row in df_dict.iterrows():
uniques['Valores'].append(df[row['Variavel']].unique())
df_uniques = pd.DataFrame.from_dict(uniques)
display(HTML(df_uniques.to_html())) #Para impedir que os valores
#de idade fiquem cortados
```

Verificação de dados faltantes

df.isnull().sum()

Variável	Valores
GENDER	[M, F]
AGE	[69, 74, 59, 63, 75, 52, 51, 68, 53, 61, 72, 60, 58, 48, 57, 44, 64, 21, 65, 55, 62, 56, 67, 77, 70, 54, 49, 73, 47, 71, 66, 76, 78, 81, 79, 38, 39, 87, 46]
SMOKING	[1, 2]
YELLOW_FINGERS	[2, 1]
.	.
.	.
.	.
CHEST PAIN	[2, 1]
LUNG_CANCER	[YES, NO]

Tabela: Tabela de Dados faltantes

Mapeamento e tradução de variáveis

```
#Traduzindo as variáveis para o português
df.replace('YES': 'Sim', 'NO': 'Não', inplace=True)
df.replace(1: 'Não', 2: 'Sim', inplace=True)
df.head()
```

Mapeamento e tradução de variáveis

	GENDER	AGE	SMOKING	...	LUNG_CANCER
0	M	69	Não	...	SIM
1	M	74	Sim	...	SIM
2	F	59	Não	...	NÃO
3	M	63	Sim	...	NÃO
4	F	63	Não	...	NÃO

Tabela: Tradução dos valores das variáveis

Traduzindo as variáveis

dict_columns = {'GENDER': 'Gênero', 'AGE': 'Idade', 'SMOKING': 'Fumante', 'YELLOW_FINGERS': 'Dedos amarelados', 'ANXIETY': 'Ansiedade', 'PEER_PRESSURE' : 'Pressão grupal', 'CHRONIC DISEASE' : 'Doença crônica', 'FATIGUE' : 'Fadiga', 'ALLERGY' : 'Alergia', 'WHEEZING' : 'Pieira', 'ALCOHOL CONSUMING' : 'Consumo alcoólico', 'COUGHING' : 'Tosse', 'SHORTNESS OF BREATH' : 'Falta de ar', 'SWALLOWING DIFFICULTY' : 'Dificuldade de ingestão', 'CHEST PAIN' : 'Dor torácica', 'LUNG_CANCER' : 'Câncer pulmonar'}
pd.DataFrame(dict_columns.items(), columns=['Variável', 'Tradução'])

Traduzindo as variáveis

	Variável	Tradução
0	GENDER	Gênero
1	AGE	Idade
2	SMOKING	Fumante
3	YELLOW_FINGERS	Dedos amarelados
.	.	.
.	.	.
.	.	.
13	SWALLOWING DIFFICULTY	Dificuldade de ingestão
14	CHEST PAIN	Dor torácica
15	LUNG_CANCER	Câncer pulmonar

Tabela: Variáveis traduzidas

Perguntas de partida e hipóteses

- Qual a distribuição de gênero dos pacientes?
- Qual a quantidade de casos confirmados de cancer de pulmão?
- Qual a distribuição de idade dos pacientes?
- Qual a correlação entre os sintomas dos pacientes?
- Existem dados fora do padrão?

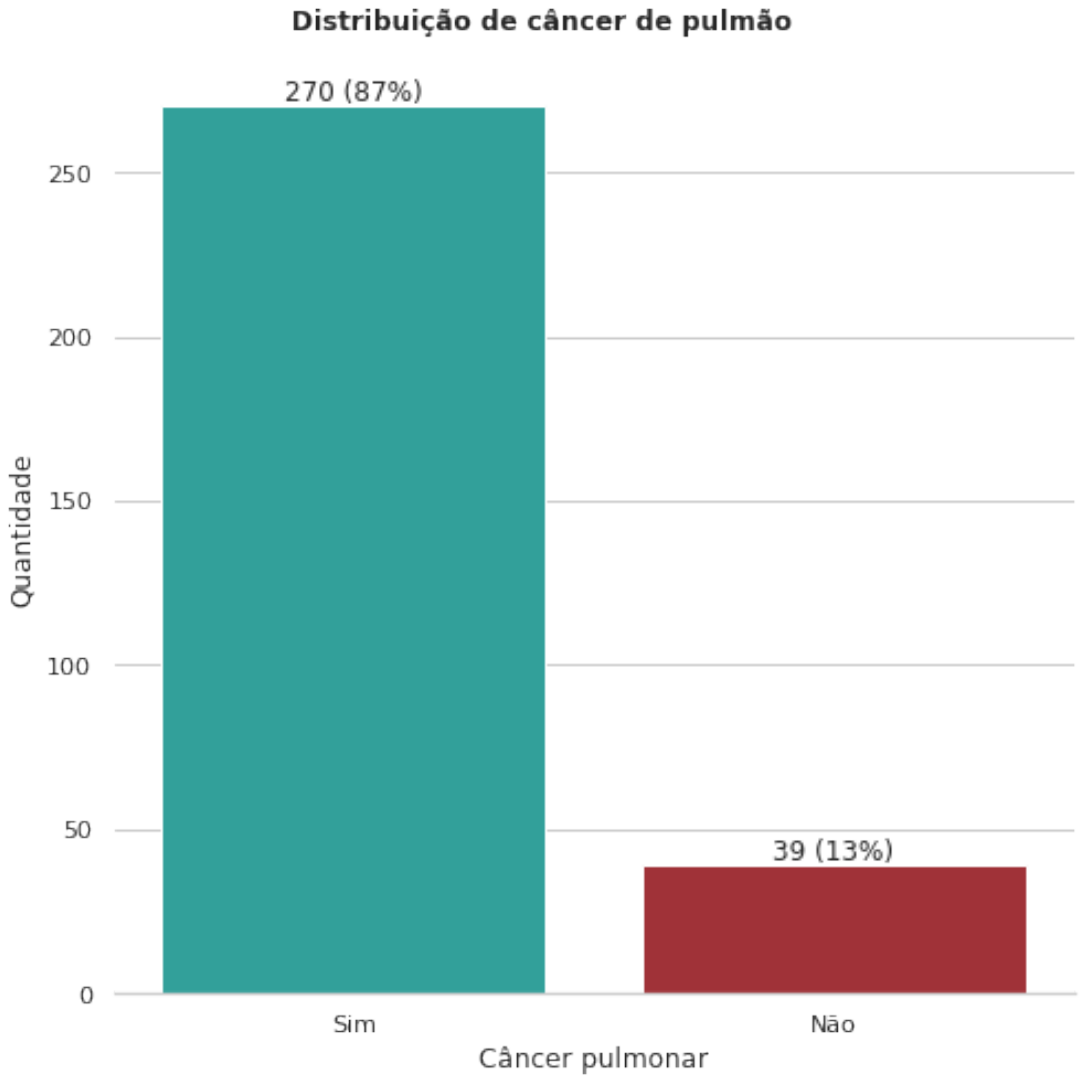
Casos confirmados de câncer de pulmão

```
plt.figure(figsize=(7,7))
plt.suptitle('Distribuição de câncer de pulmão', fontweight='bold')
visualize.variable_dist_count(df, 'LUNG_CANCER')
plt.tight_layout()
plt.show()
```

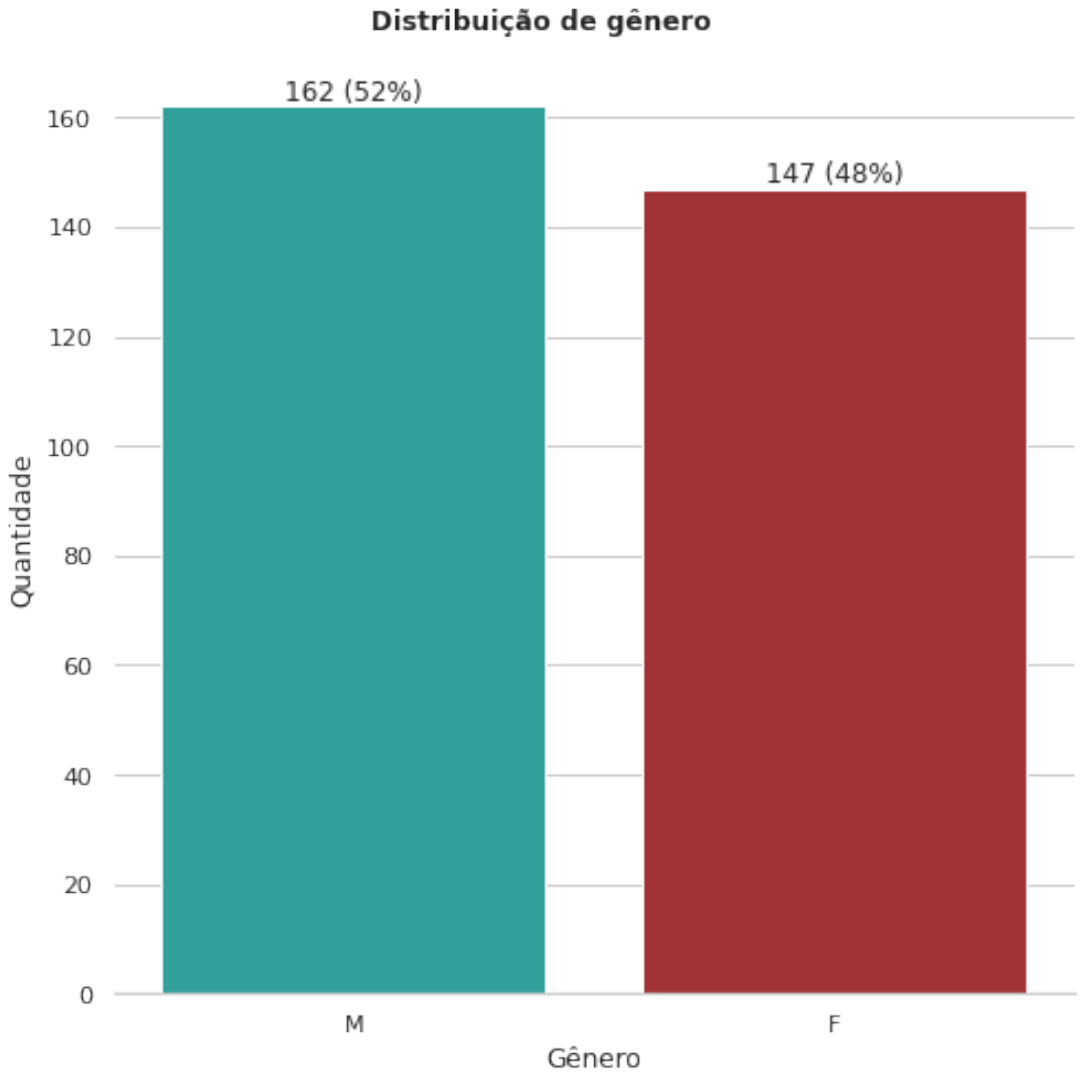
Casos confirmados de câncer de pulmão

```
plt.figure(figsize=(7,7))
plt.suptitle('Distribuição de câncer de pulmão', fontweight='bold')
visualize.variable_dist_count(df, 'LUNG_CANCER')
plt.tight_layout()
plt.show()
```

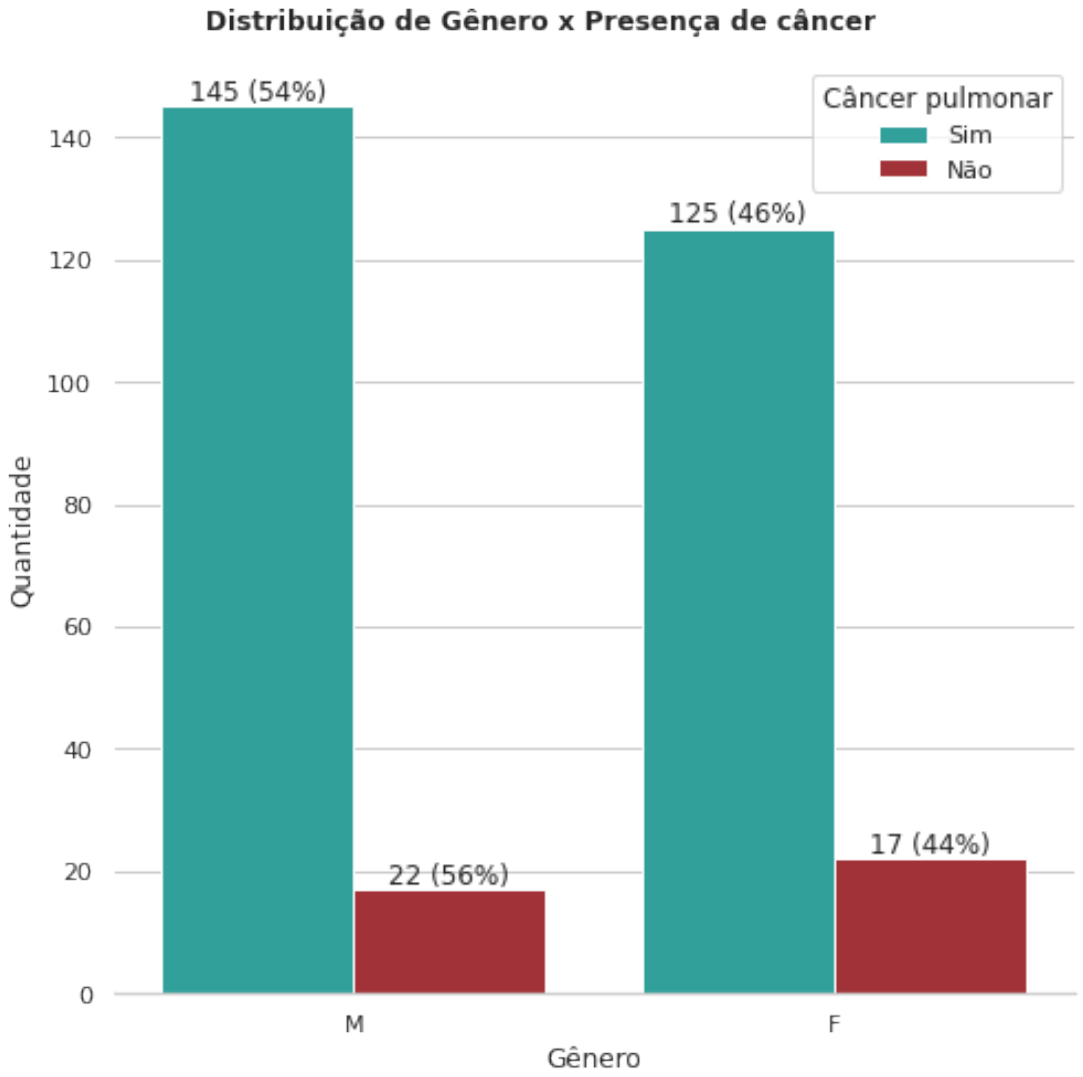
Casos confirmados de câncer de pulmão



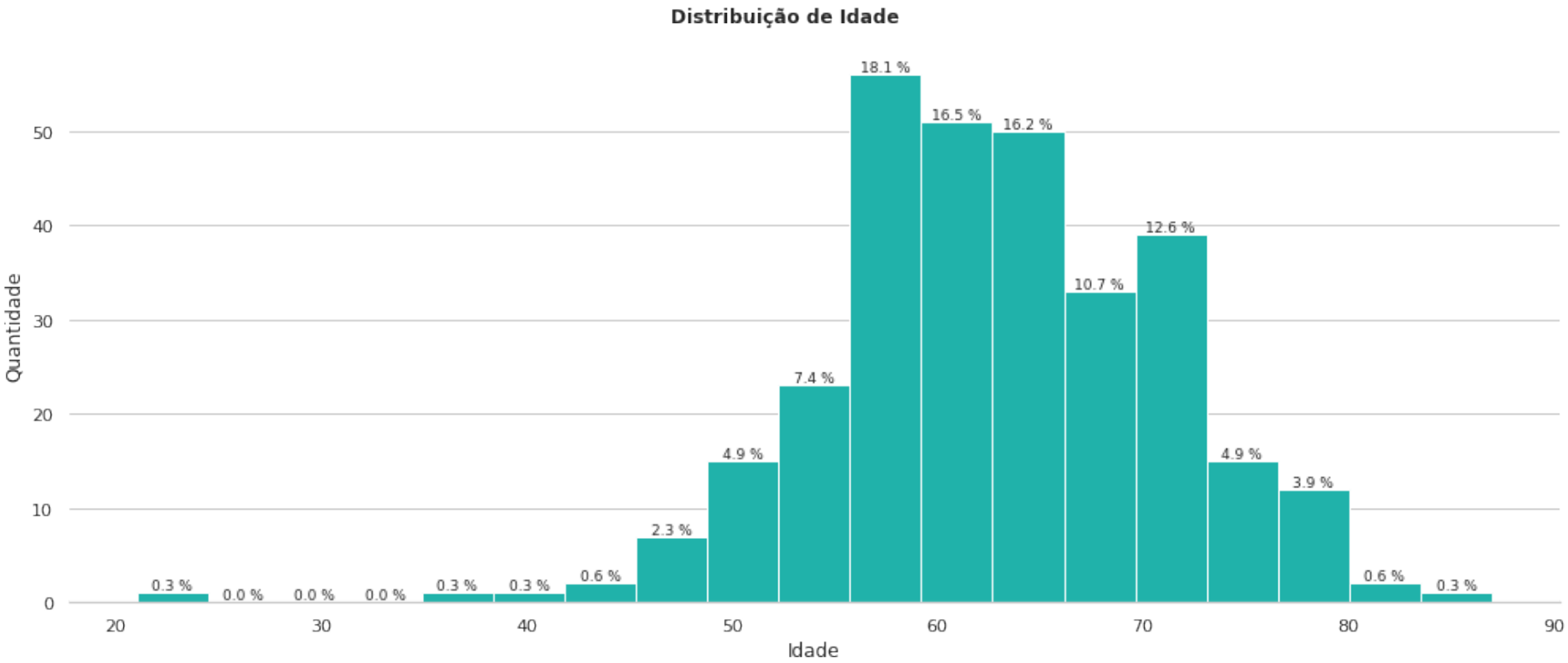
Distribuição de gênero



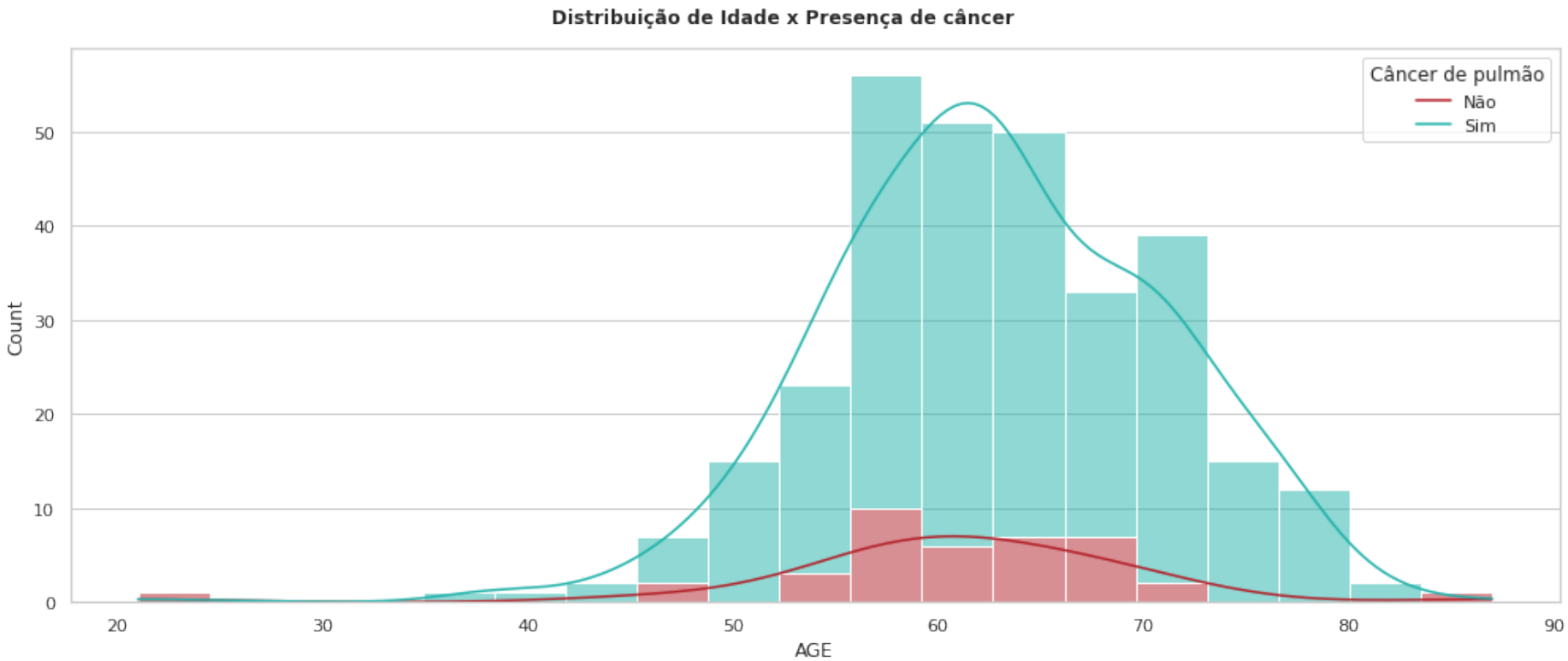
Distribuição de gênero X presença de Câncer



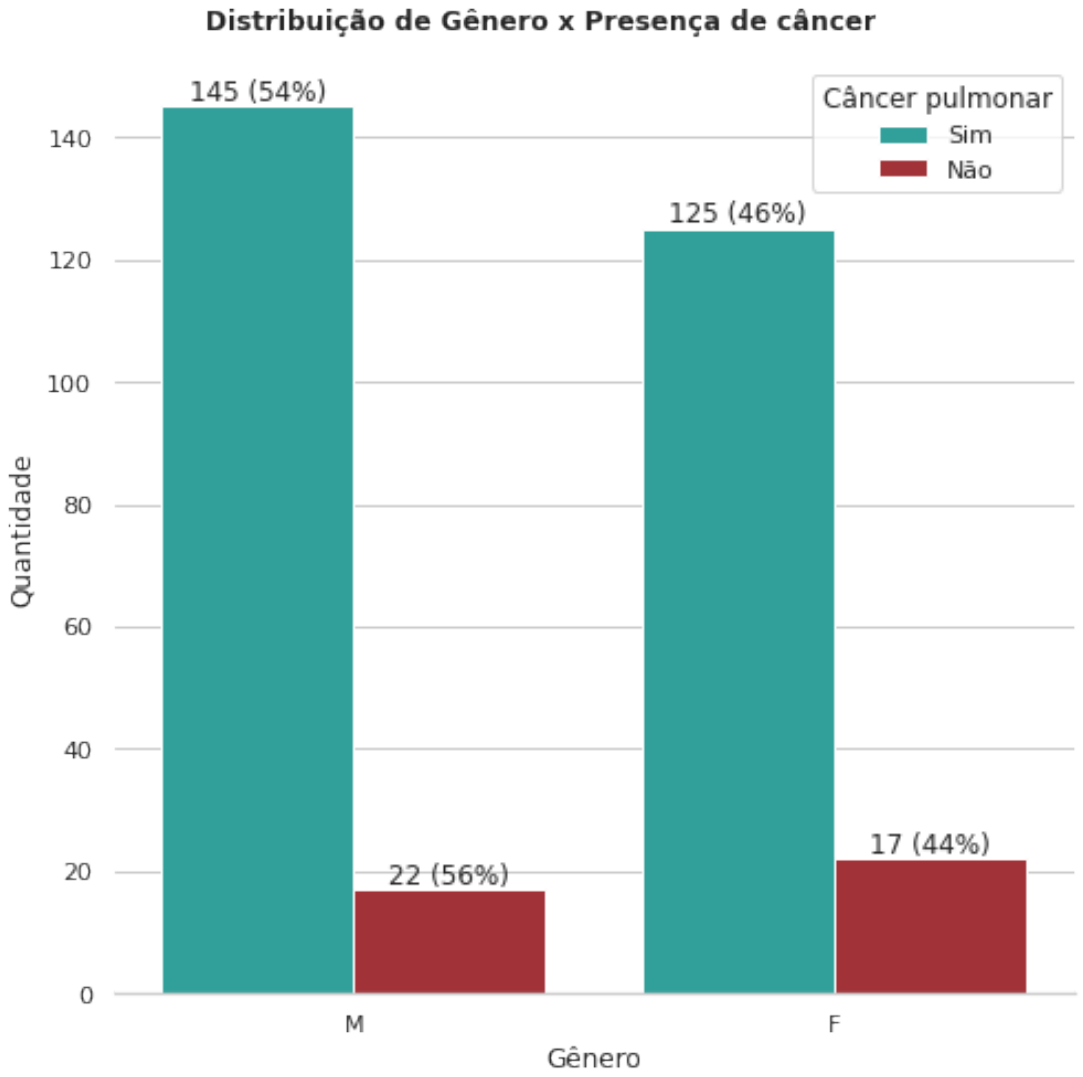
Distribuição por Idade



Distribuição por Idade X presença de Câncer



Distribuição de gênero X presença de Câncer

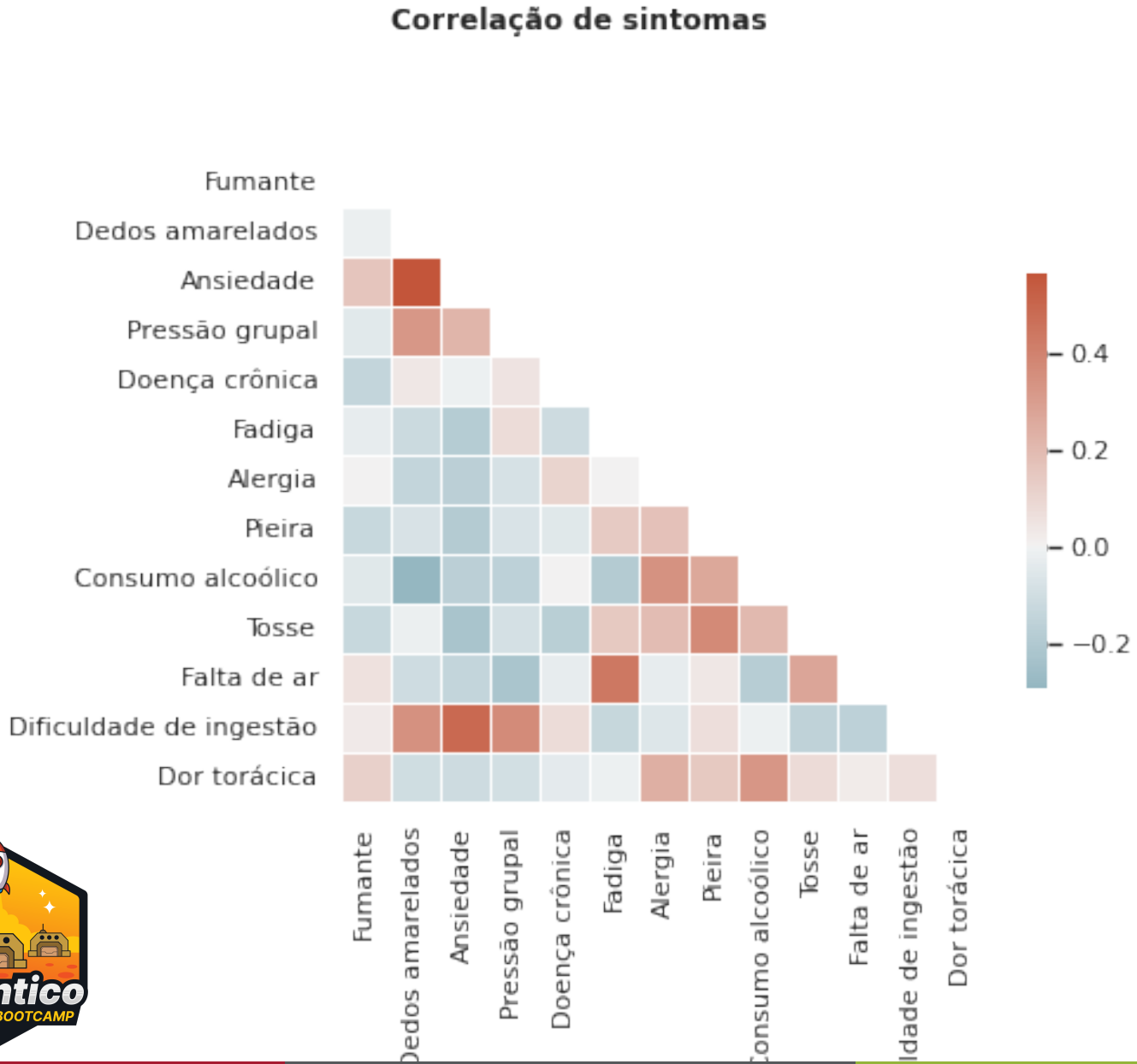


Análise de sintomas

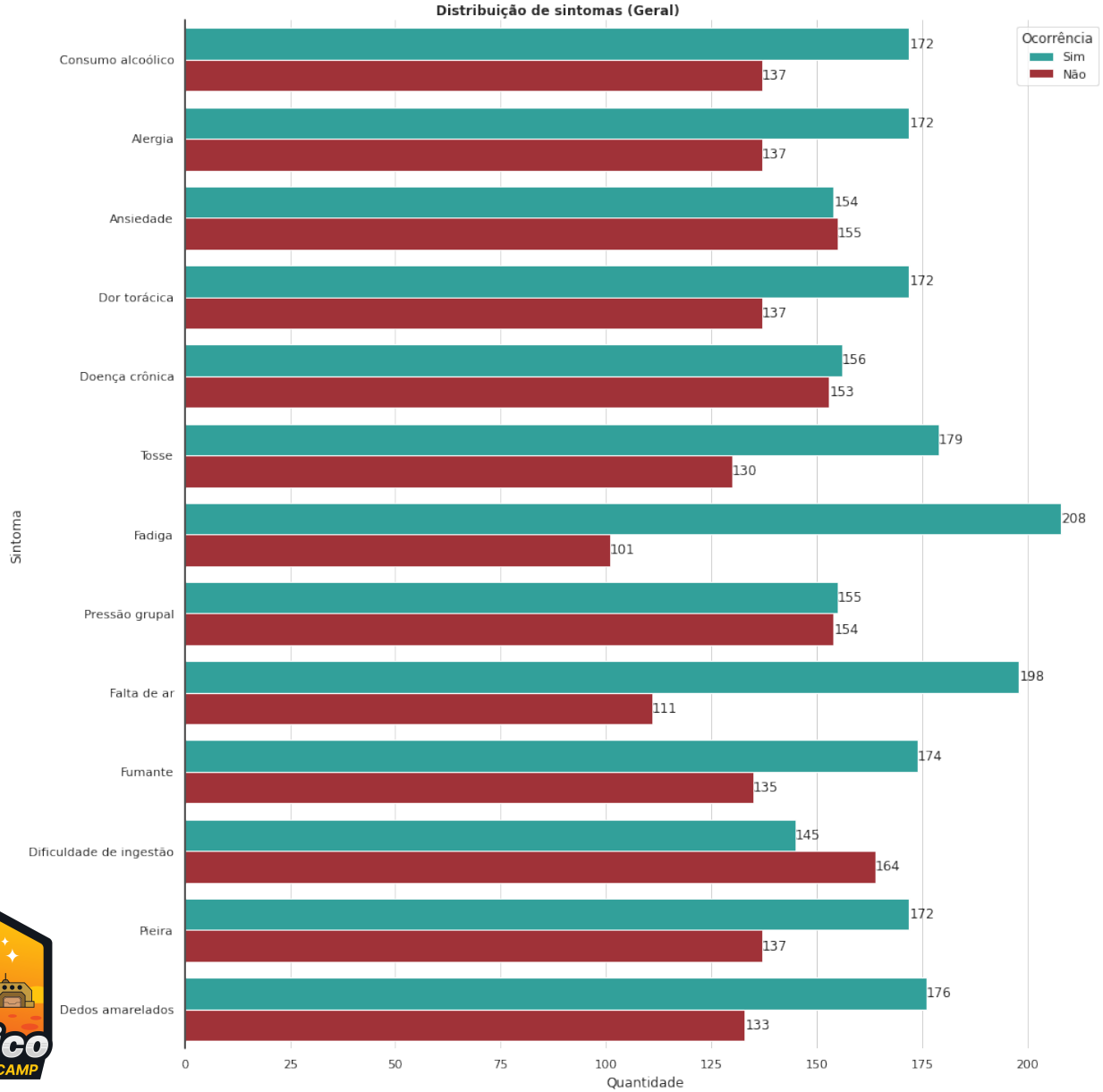
Analisamos a presença de sintomas e buscamos encontrar:

- correlação entre sintomas e a existência de câncer de pulmão

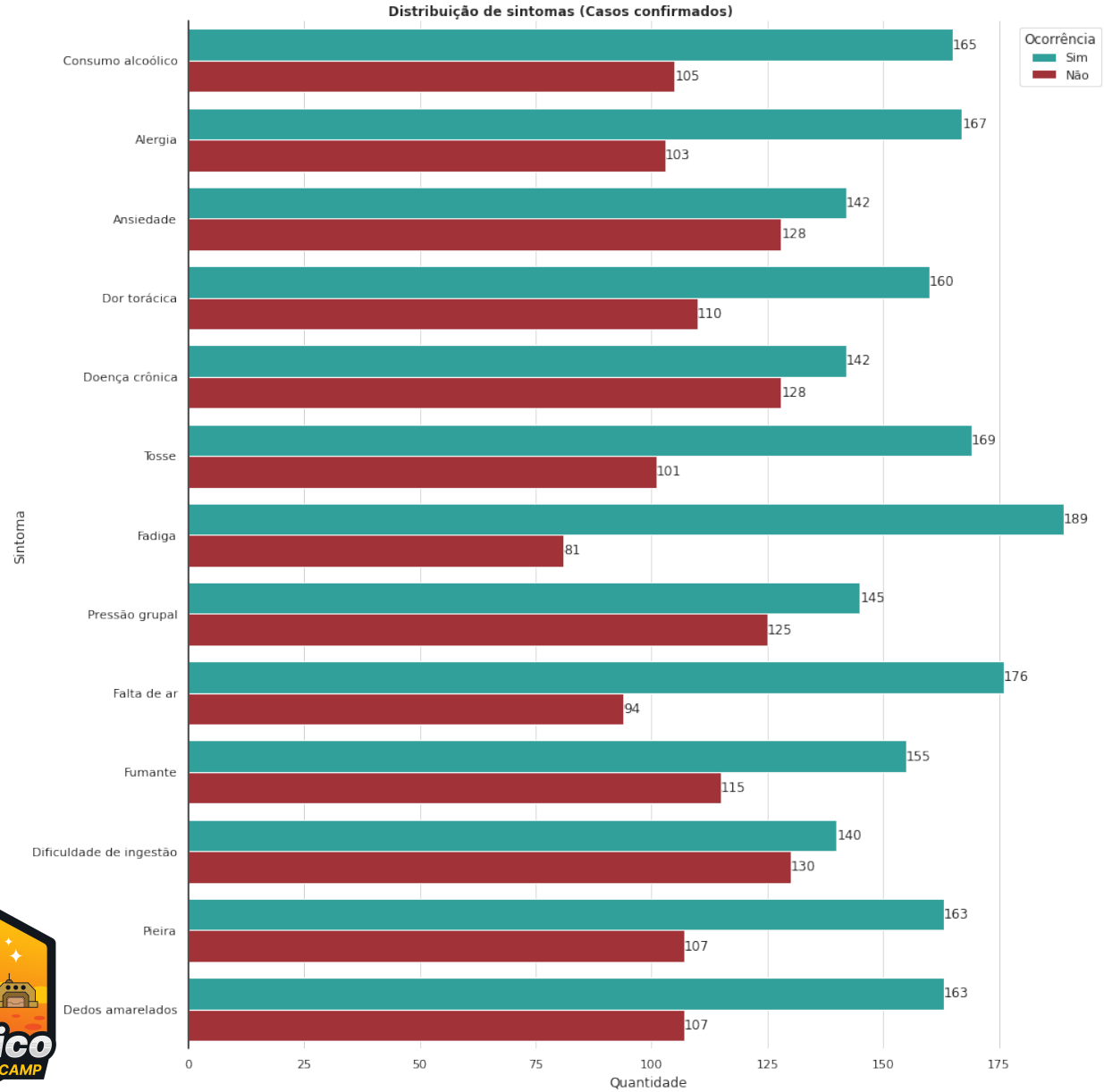
Correlação de Sintomas



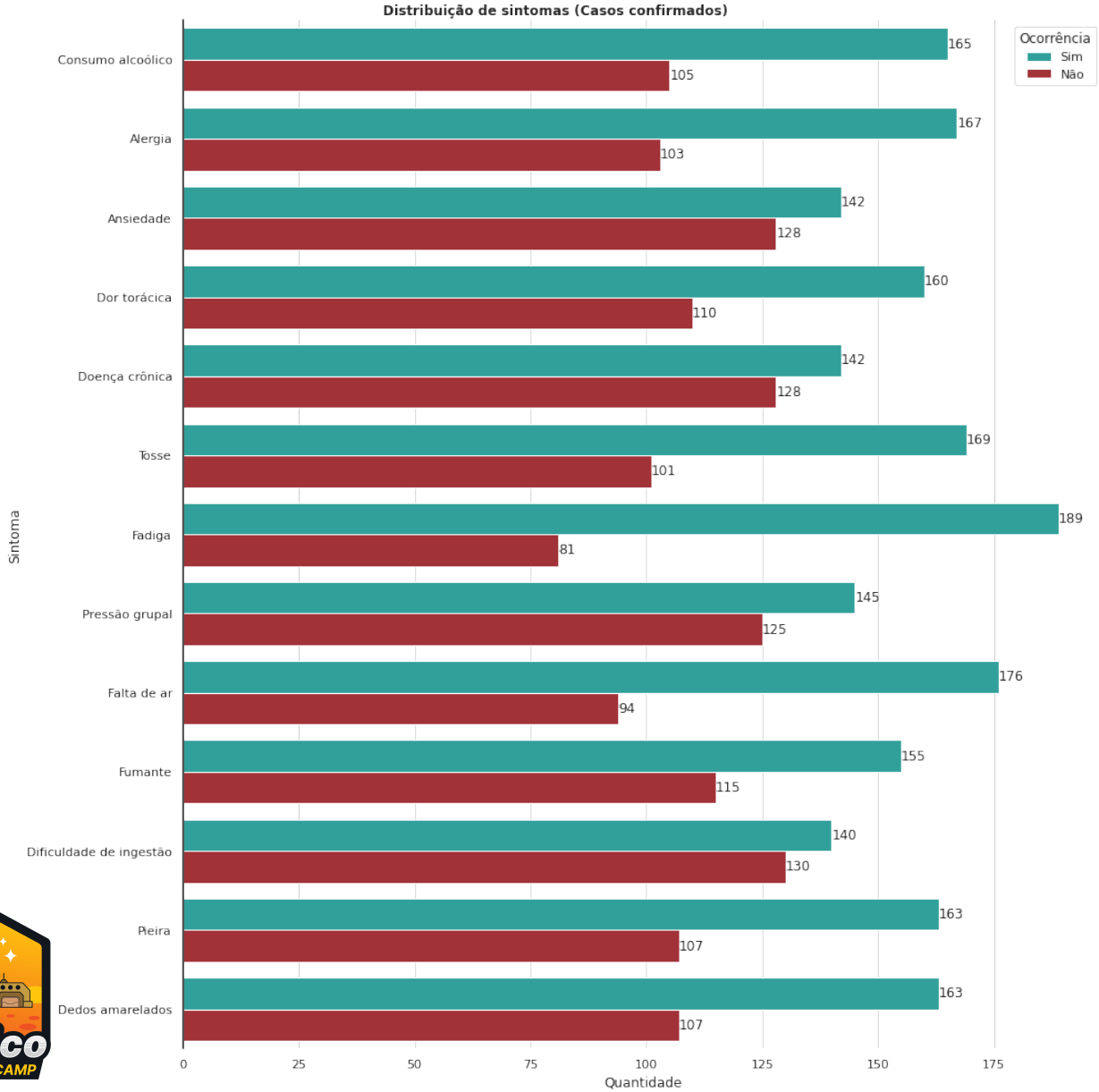
Ocorrência de sintomas



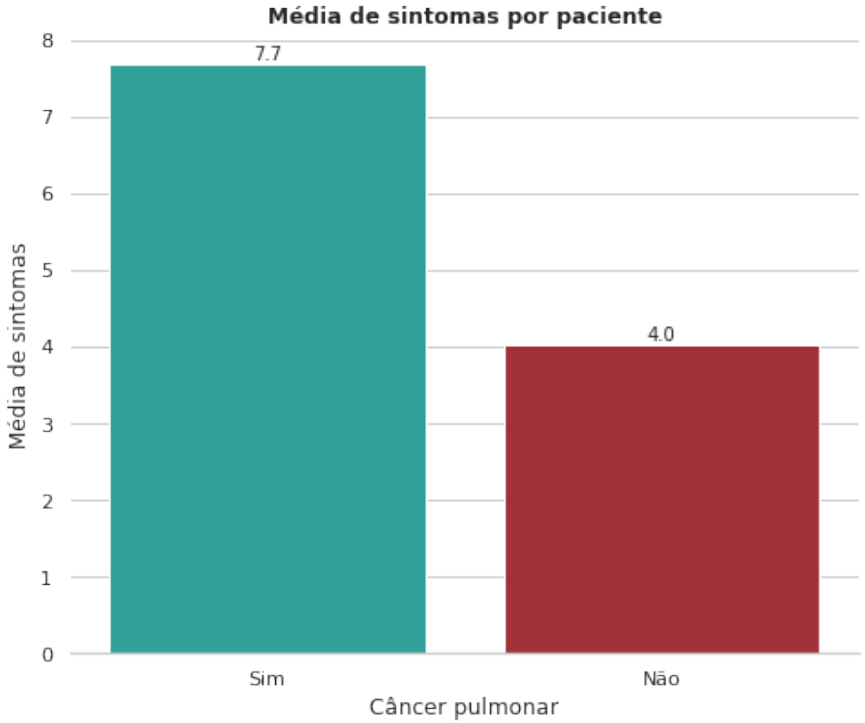
Distribuição de sintomas em casos confirmados

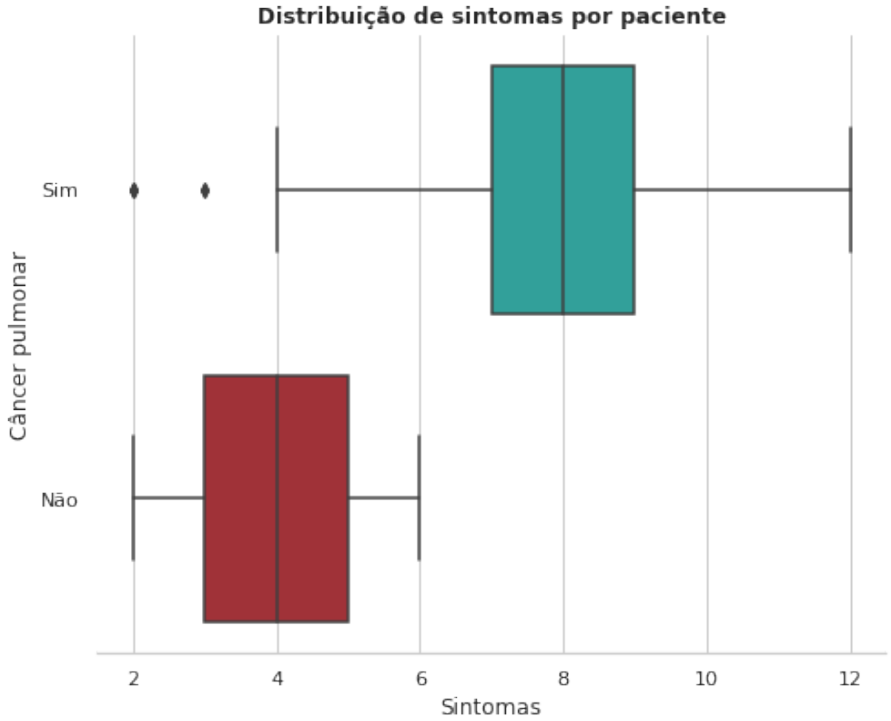


Distribuição de sintomas em casos confirmados



Dados estatísticos sobre sintomas





Dados estatísticos sobre sintomas

Câncer pulmonar	Média	Mediana	Mínimo	Máximo
Sim	7,7	8,0	2	12
Não	4,0	4,0	2	6

Tabela: Estatística Descritiva

Dados estatísticos sobre sintomas

	GENDER	AGE	SMOKING	...	LUNG_CANCER	Sintomas
187	M	55	Sim	...	SIM	2
190	F	69	Não	...	SIM	2
193	F	64	Não	...	SIM	2

Tabela: Estatística Descritiva

Conclusão

Notamos então que há 3 pacientes com 2 sintomas que possuem câncer de pulmão. O que pode ser um indicativo de que esses pacientes podem ter descoberto a doença em estágio inicial, ou que a doença pode estar causando nestes indivíduos outros sintomas que não foram registrados.

No geral, concluimos que há casos fora do padrão, e que nem sempre a ausência de sintomas significa que o paciente não possui câncer de pulmão.

Além disso, também podemos concluir que a presença dos sintomas apresentados pode influenciar na presença de câncer de pulmão. Ademais podemos afirmar que não há sintomas que sejam exclusivos de pacientes com câncer de pulmão, assim como não há como determinar apenas um (ou um grupo) de sintomas que sejam suficientes para provar que uma pessoa possui câncer de pulmão.

FIM