INSTITUTO FEDERAL
Norte de Minas Gerais

Introdução a Sistemas Inteligentes

Análise Exploratória

Parte 2

Profª. Suzana Mota

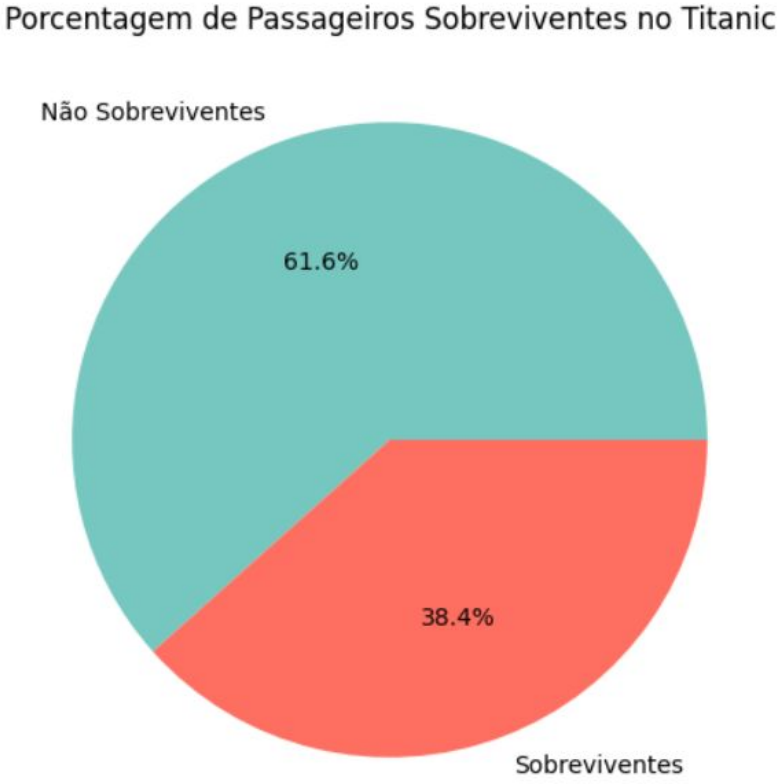
Análise Exploratória de Dados

https://colab.research.google.com/drive/1DY8bVKkhkoBpAsp2yUe2qQW9e9LLhNcz?usp=sharing

Analisando dados do Titanic

Hello World de Data Science!

Análise Exploratória dos Dados



Análise Exploratória dos Dados

Seja curioso e faça perguntas:

- ~~Qual foi a porcentagem dos passageiros sobreviventes?~~
- Qual era a faixa etária dos passageiros que estavam no Titanic?
- Houveram mais crianças ou mais adultos que sobreviveram?
- O gênero influenciou na sobrevivência?
- E deixe a curiosidade fluir, para fazer outras perguntas!

Análise Exploratória dos Dados

Overview de dados

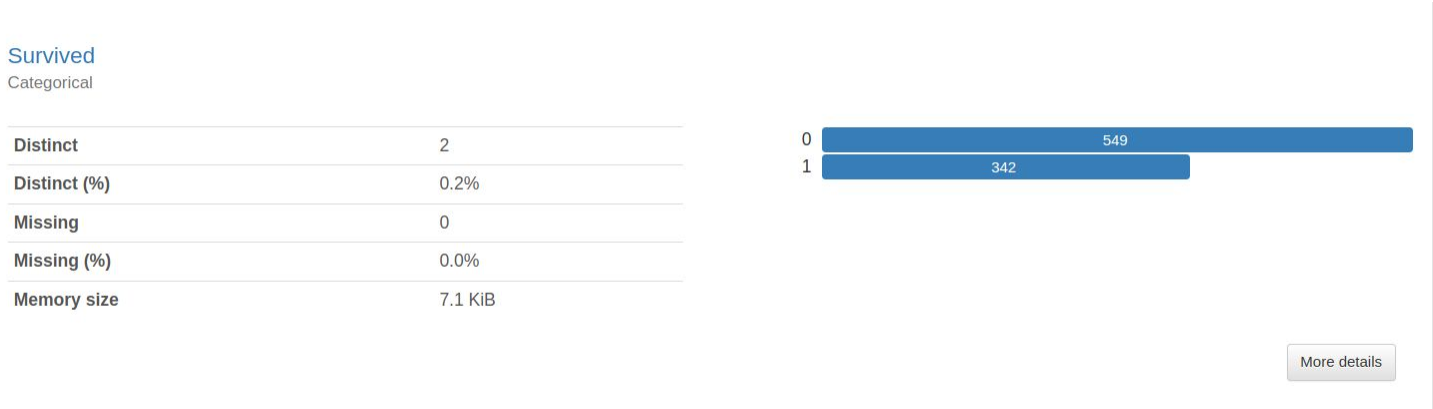
PassengerId = ID do Passageiro
Survived = Sobreviveu (1 = Sim, 0 = Não)
Pclass = Classe do Bilhete (1ª, 2ª, 3ª Classe)
Name = Nome
Sex = Sexo
Age = Idade
SibSp = Número de Irmãos/Cônjuges a Bordo
Parch = Número de Pais/Filhos a Bordo
Ticket = Bilhete
Fare = Tarifa (preço do bilhete)
Cabin = Cabine
Embarked = Embarcou (Porto de Embarque: C = Cherbourg, Q = Queenstown, S = Southampton)

Análise Exploratória dos Dados

Overview de dados

```
!pip install ydata_profiling
from ydata_profiling import ProfileReport

profile  = ProfileReport(df)
profile
```



Mediana x Média

- **O que fazer com os valores vazios?**

A mediana é considerada uma medida melhor que a média em certos contextos, especialmente quando os dados contêm outliers (valores extremos) ou são assimétricos (não seguem uma distribuição simétrica).

Mediana x Média

Qual medida é mais representativa?

1.000 1.200 1.400 1.500 1.700 2.000 10.000

Média: R$2.685

Situações:

Notas
Altura
Peso

Mediana: R$1.500

Situações:

Renda
Preço com alta variação
Idade

Comandos Básicos

df.head()

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	P
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	ST 3
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	

df.info()

```
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
```

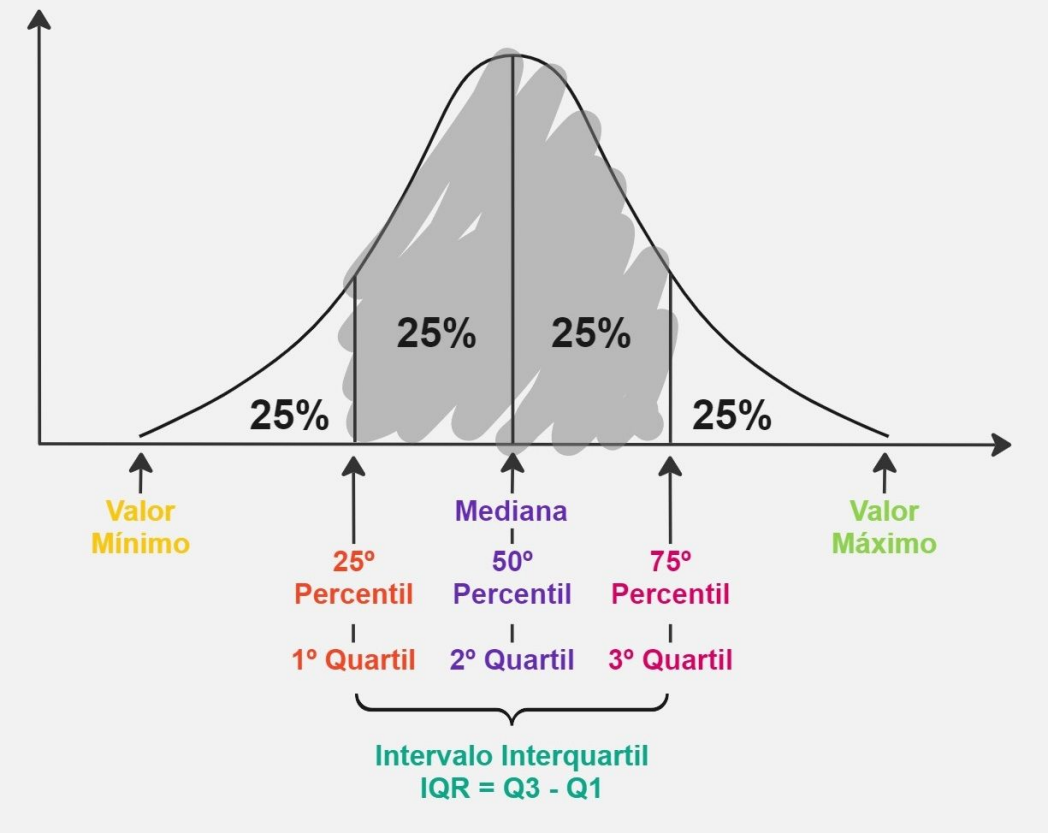
Comandos Básicos

df.describe()

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Comandos Básicos

df.describe()



Selecionando os dados

Selecionando uma coluna específica (Nome)

df['Name'].head()

	Name
0	Braund, Mr. Owen Harris
1	Cumings, Mrs. John Bradley (Florence Briggs Th...
2	Heikkinen, Miss. Laina
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)
4	Allen, Mr. William Henry

Selecionando os dados

Selecionando múltiplas colunas (Nome, Idade, Sobrevivente)

df[['Name', 'Age', 'Survived']].head()

	Name	Age	Survived
0	Braund, Mr. Owen Harris	22.0	0
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38.0	1
2	Heikkinen, Miss. Laina	26.0	1
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35.0	1
4	Allen, Mr. William Henry	35.0	0

Filtrando os dados

Filtrando os passageiros que sobreviveram Filtro Booleano

df[df['Survived'] == 1].head()

```
df['Survived'] == 1
```

	Survived
0	False
1	True
2	True
3	True
4	False

```
df[df['Survived'] == 1].head()
```

	PassengerId	Survived	Pclass	
1	2	1	1	C
2	3	1	3	
3	4	1	1	
8	9	1	3	J

Filtrando os dados

Filtrando os passageiros que sobreviveram Query

df.query('Survived==1').head()

```
df.query('Survived==1').head()
```

	PassengerId	Survived	Pclass
1	2	1	1
2	3	1	3
3	4	1	1
8	9	1	3
9	10	1	2

Lidando com Dados Faltantes

Verificando onde há valores faltantes

df.isnull().sum()

	0
PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	0
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687

Lidando com Dados Faltantes

```
# Substituir valores faltantes na coluna 'Age' pela mediana
df['Age'].fillna(df['Age'].median(), inplace=True)
```

```
# Substituir valores faltantes na coluna 'Age' pela mediana,
sem inplace = True
df['Age'] = df['Age'].fillna(df['Age'].median())
```

```
df['Age'].describe()
```

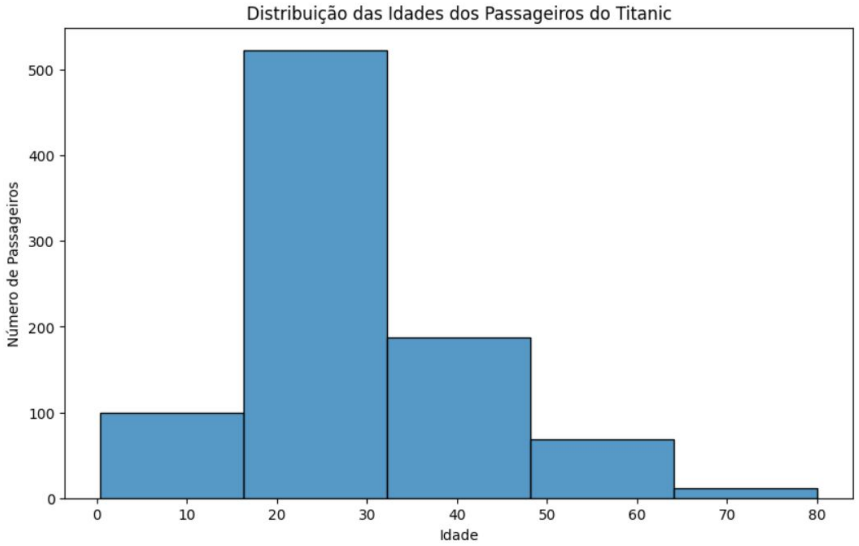
	Age
count	891.000000
mean	29.361582
std	13.019697
min	0.420000
25%	22.000000
50%	28.000000
75%	35.000000
max	80.000000

Visualização de Dados

Histograma

Um histograma é um gráfico que representa a **distribuição de um conjunto de dados**, dividindo-os em intervalos (bins) e contando quantas observações caem em cada intervalo.

```
plot_histogram(df, 'Age', bins=5, title='Distribuição das Idades dos Passageiros do Titanic', xlabel='Idade', ylabel='Número de Passageiros')
```



Bin 1: 0 a 16 anos
Bin 2: 16 a 32 anos
Bin 3: 32 a 48 anos
Bin 4: 48 a 64 anos
Bin 5: 64 a 80 anos

Histograma

```
[31] plot_histogram(df, 'Age', bins=8, title='Distribuição das Idades dos Passageiros do Titanic', xlabel='Idade', ylabel='Número de Passageiros')
```

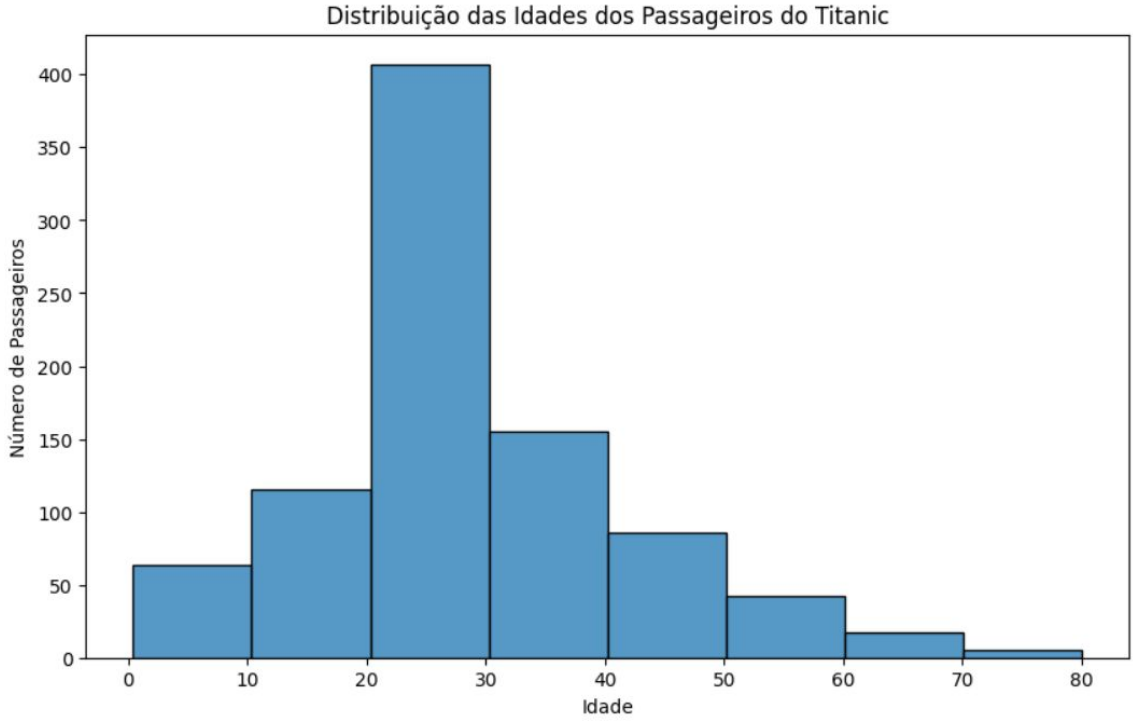
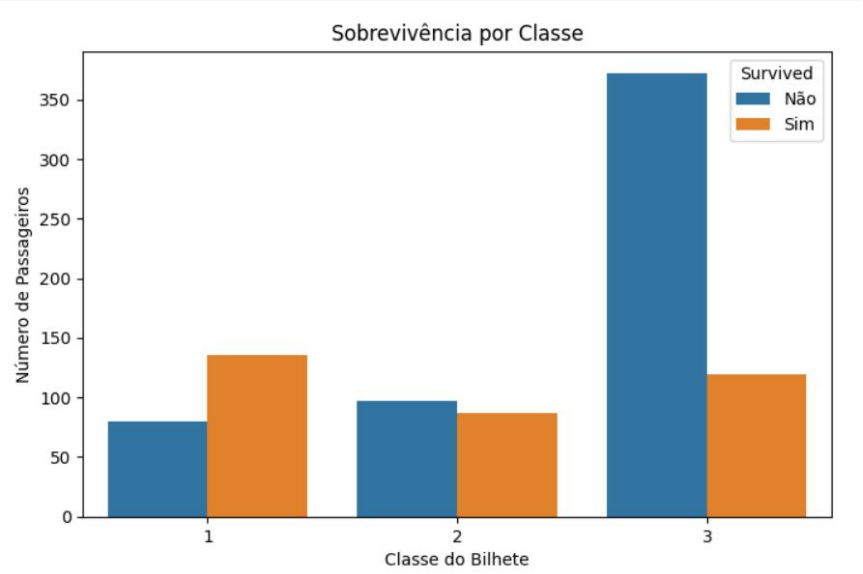
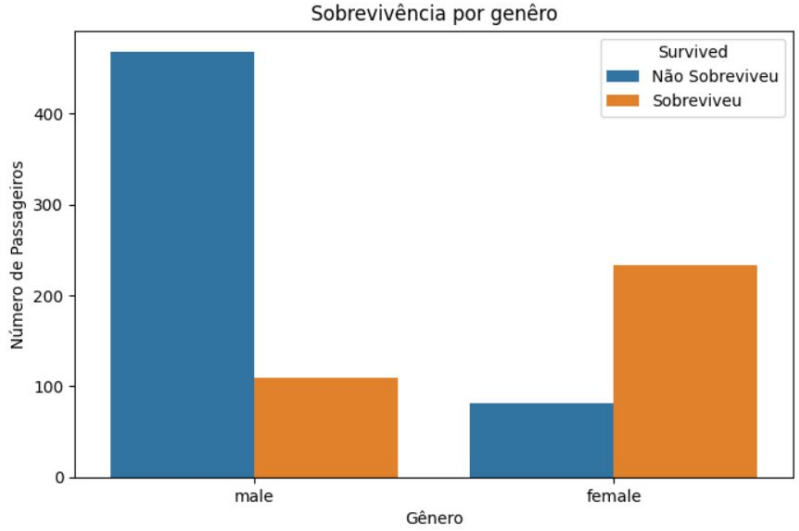


Gráfico de Barras Comparativo

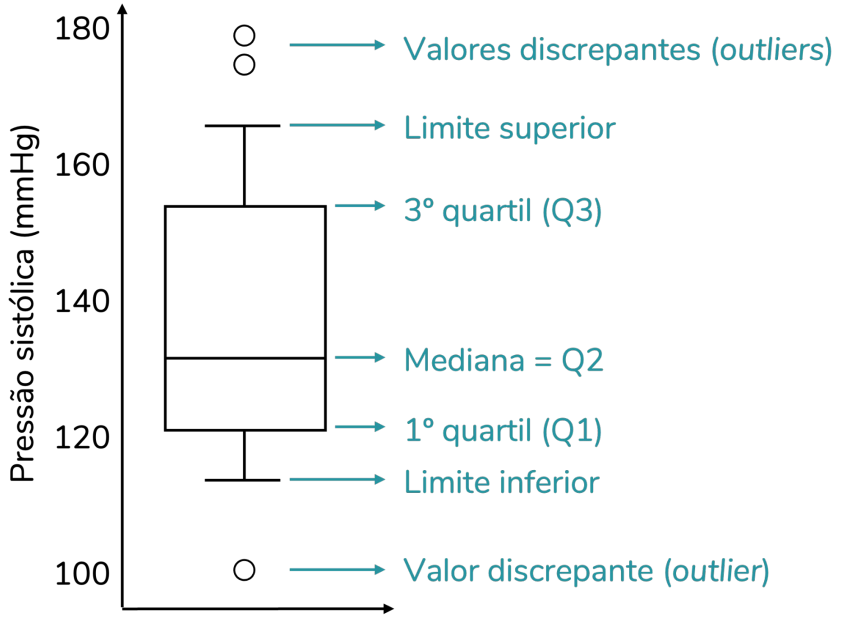
Os gráficos de barras comparativos permitem que duas ou mais séries de dados sejam comparadas lado a lado para facilitar a análise e interpretação.





Boxplot

O box plot, também conhecido como diagrama de caixa, é uma representação gráfica que permite visualizar a distribuição de um conjunto de dados através de seus quartis, além de identificar a presença de outliers.



Boxplot

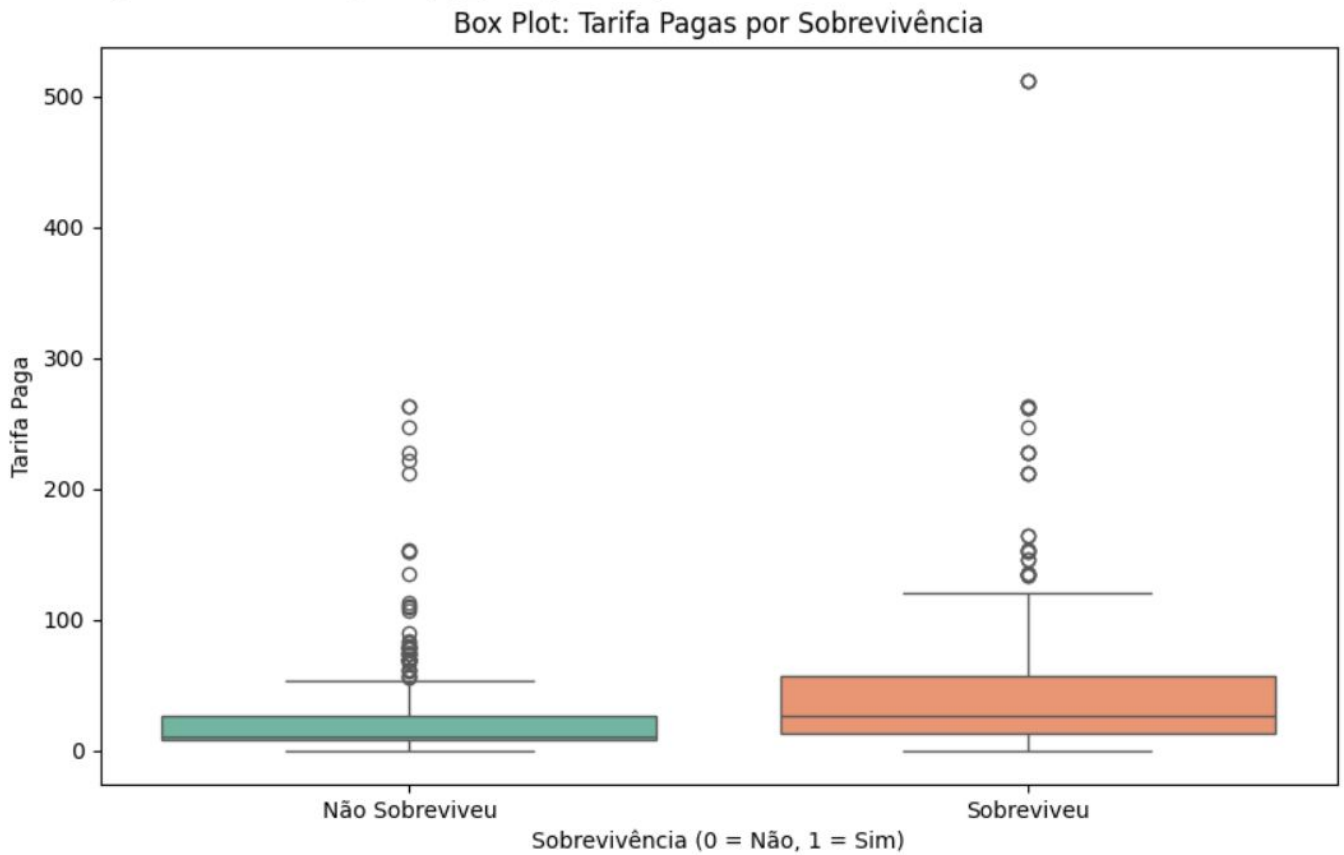
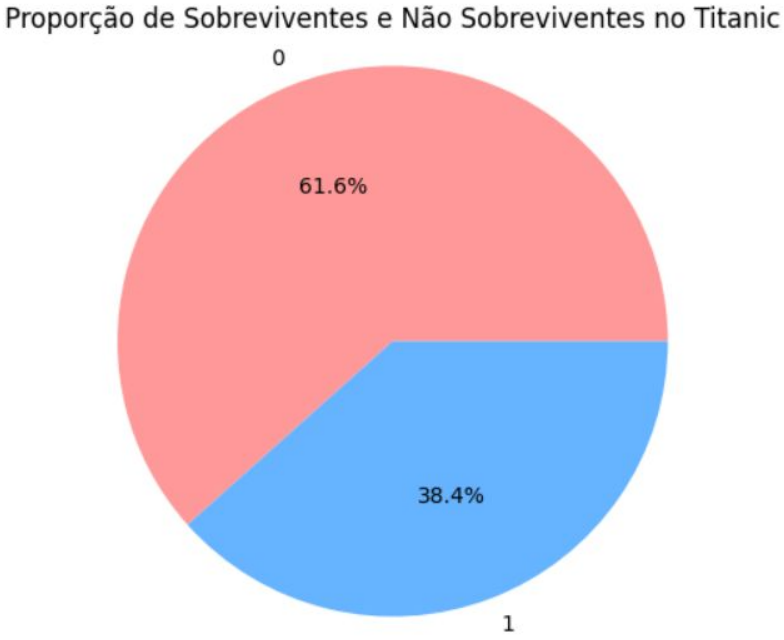

Box Plot: Tarifa Pagas por Sobrevivência
Tarifa Paga
500
400
300
200
100
0
Não Sobreviveu
Sobreviveu
Sobrevivência (0 = Não, 1 = Sim)

Gráfico de Pizza



É uma representação gráfica que mostra a proporção de diferentes categorias em um conjunto de dados.

SEMPRE UTILIZE COM OS VALORES EM PORCENTAGEM BEM DESCRITOS!