# Titanic

Pedro Silvestre

2023-09-01

## 1. Introdução

Neste projeto, vamos analisar e realizar uma estatística descritiva do conjunto de dados do Titanic, que contém algumas variáveis que podem ou não explicar a probabilidade de sobrevivência de cada passageiro ao desastre.

Inicialmente, analisaremos os gráficos e tiraremos algumas conclusões sobre essas variáveis.

```
titanic <- read.csv("titanic.csv")
summary(titanic)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

```
str(titanic)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
```

```
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Conforme podemos ver, temos 891 observações e 12 variáveis, sendo que a primeira apenas enumera nossa base de dados, portanto, não será necessária para a nossa análise.

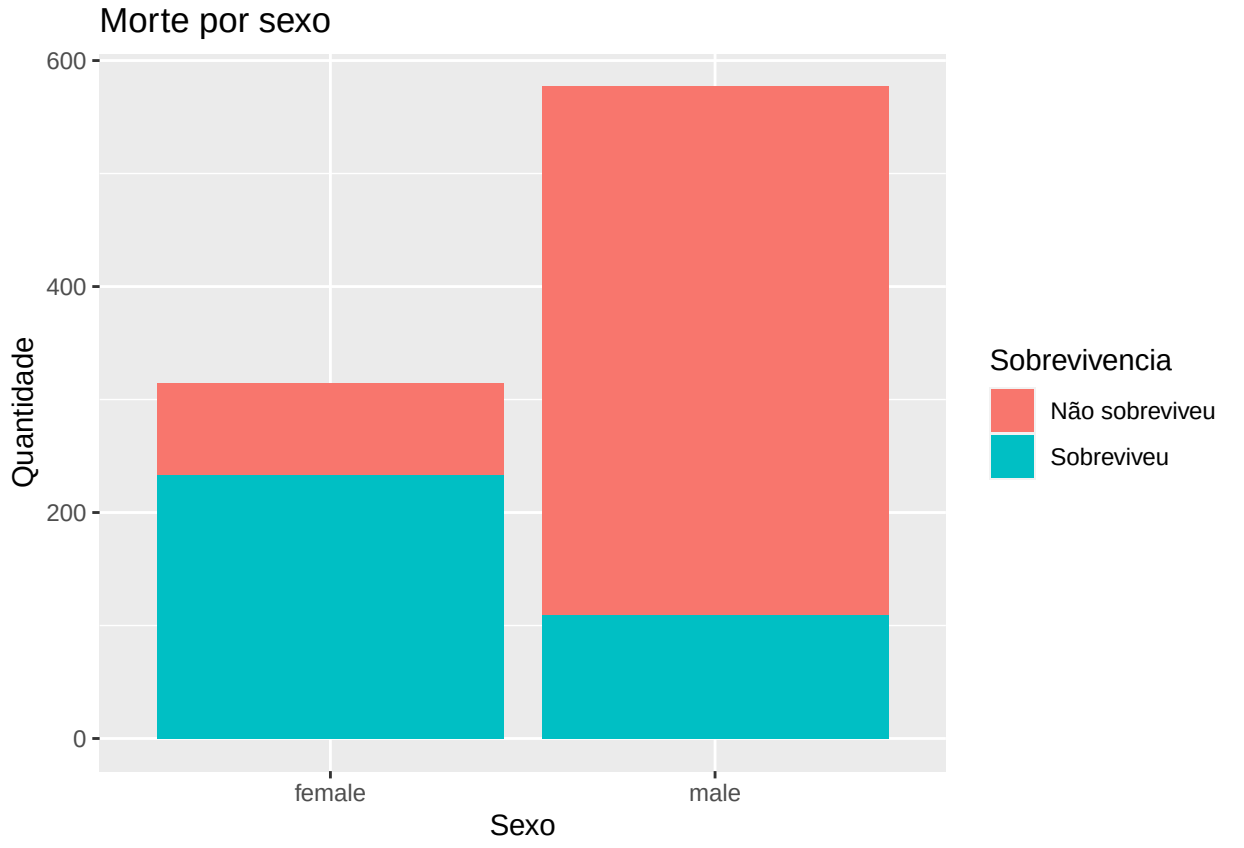
Aproveitaremos para transformar as variáveis necessárias em fatores.

```
titanic$Survived <- as.factor(titanic$Survived)
titanic$Sex <- as.factor(titanic$Sex)
titanic$Embarked <- as.factor(titanic$Embarked)
titanic$Pclass <- as.factor(titanic$Pclass)
titanic <- titanic[,-1]
```

## 2. Análise dos Dados

O primeiro gráfico que decidimos analisar foi o de mortes de acordo com o sexo:

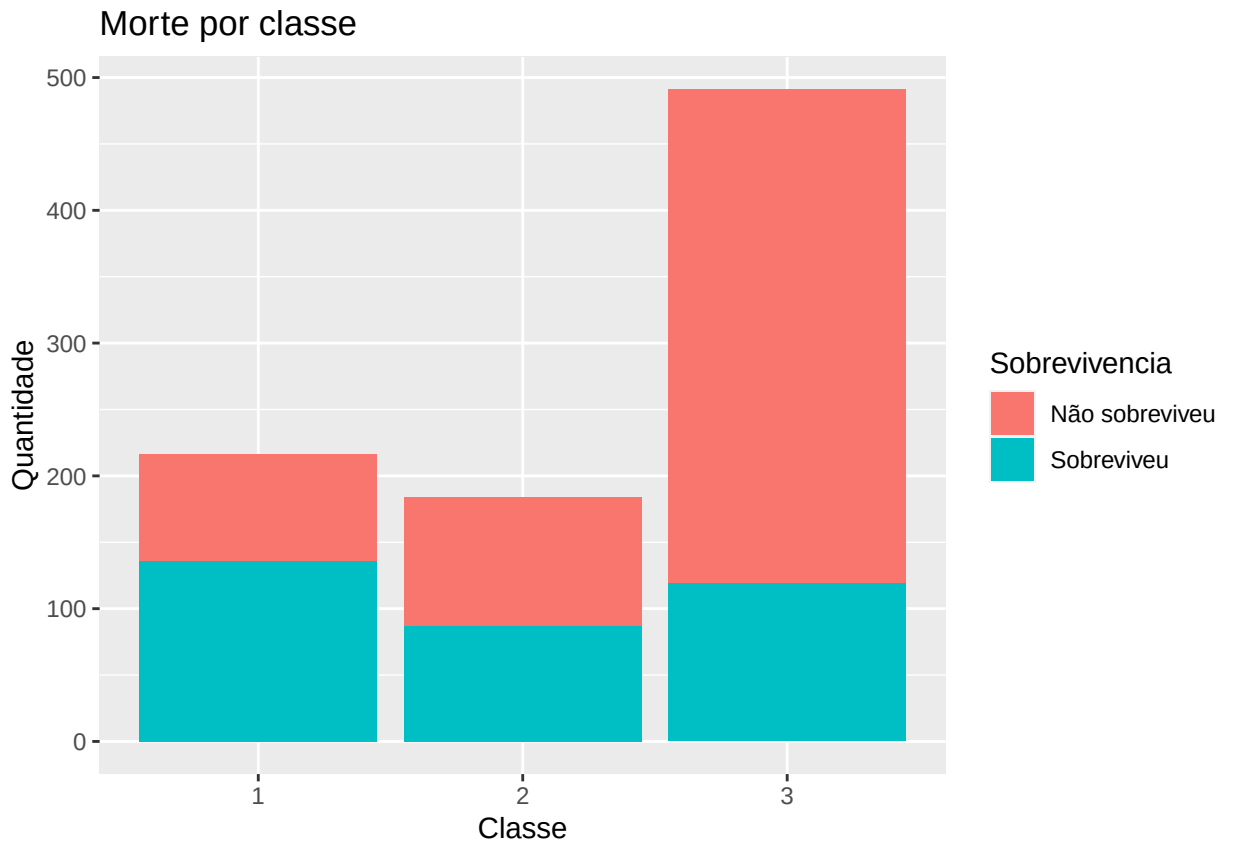
```
ggplot(data = titanic, aes(x = Sex, fill = Survived))+
  geom_bar()+
  labs(title = "Morte por sexo",
       x = "Sexo",
       y = "Quantidade")+
  scale_fill_discrete(name = "Sobrevivencia", labels = c("Não sobreviveu", "Sobreviveu"))
```



Como podemos observar, a taxa de sobrevivência das mulheres foi muito maior.

Outra variável altamente explicativa é a classe, como podemos ver abaixo:

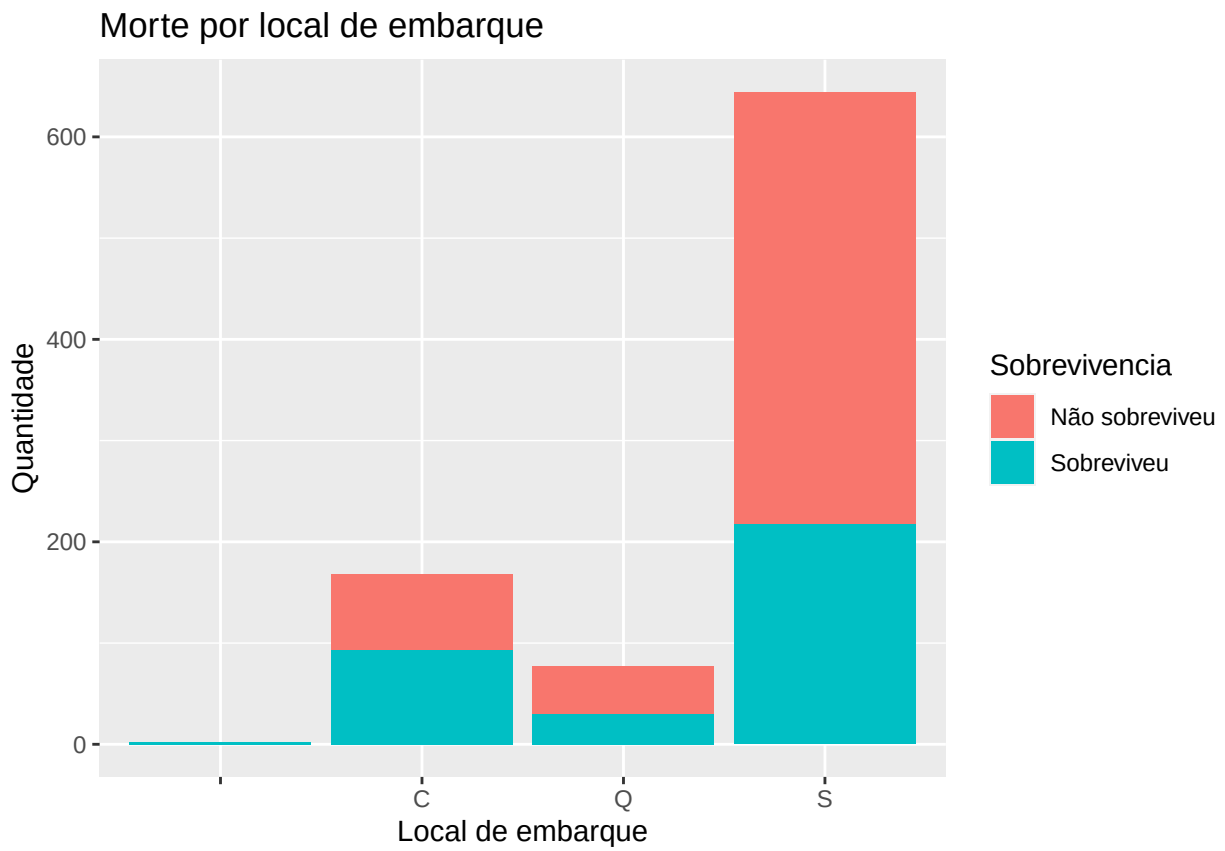
```
ggplot(data = titanic, aes(x = Pclass, fill = Survived))+
  geom_bar()+
  labs(title = "Morte por classe",
       x = "Classe",
       y = "Quantidade")+
  scale_fill_discrete(name = "Sobrevivencia", labels = c("Não sobreviveu", "Sobreviveu"))
```




Quem estava na classe 3 estava muito mais suscetível a não sobreviver.

Outra variável que podemos observar é o local de embarque:

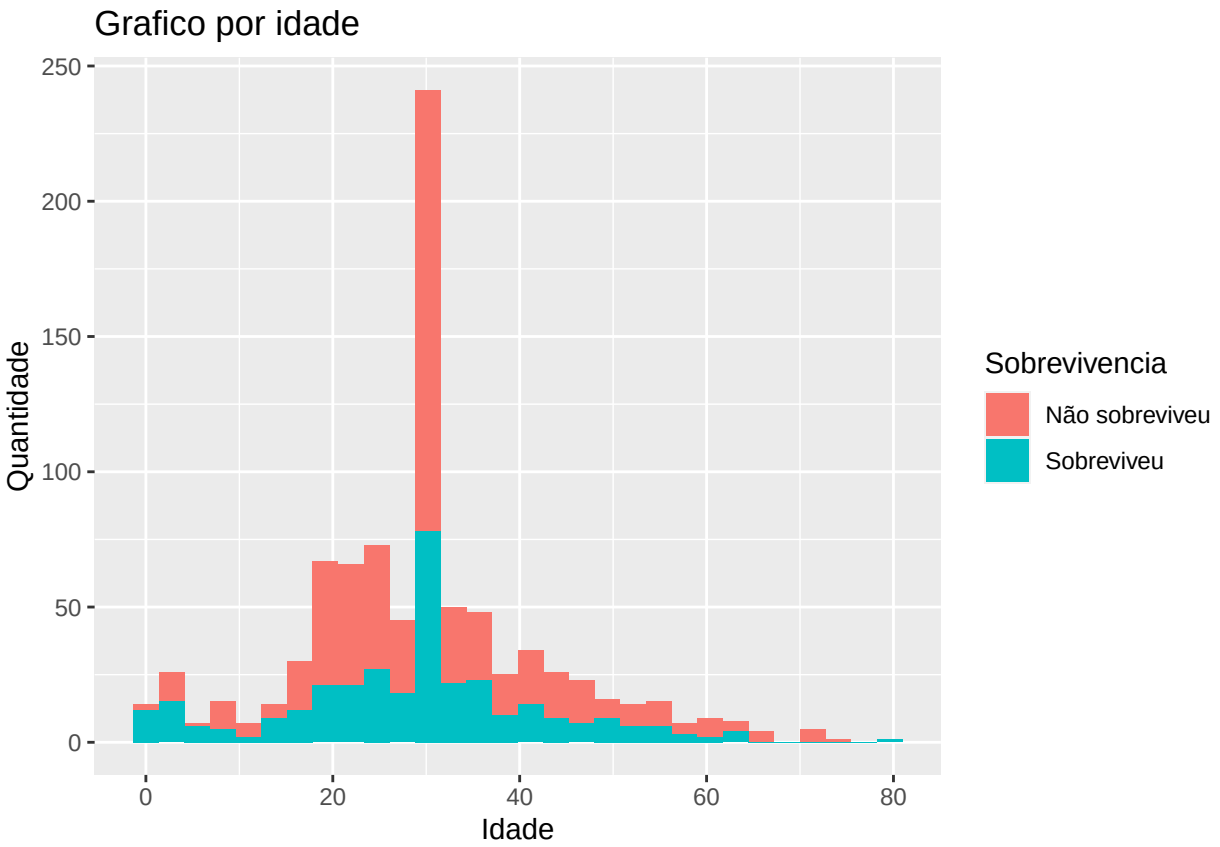
```
ggplot(data = titanic, aes(x = Embarked, fill = Survived))+
  geom_bar()+
  labs(title = "Morte por local de embarque",
       x = "Local de embarque",
       y = "Quantidade")+
  scale_fill_discrete(name = "Sobrevivencia", labels = c("Não sobreviveu", "Sobreviveu"))
```




Já a idade é uma variável com muitos dados faltantes. Como o número de dados faltantes é alto, atribuiremos a cada valor "NA" a média de idade dos restantes.

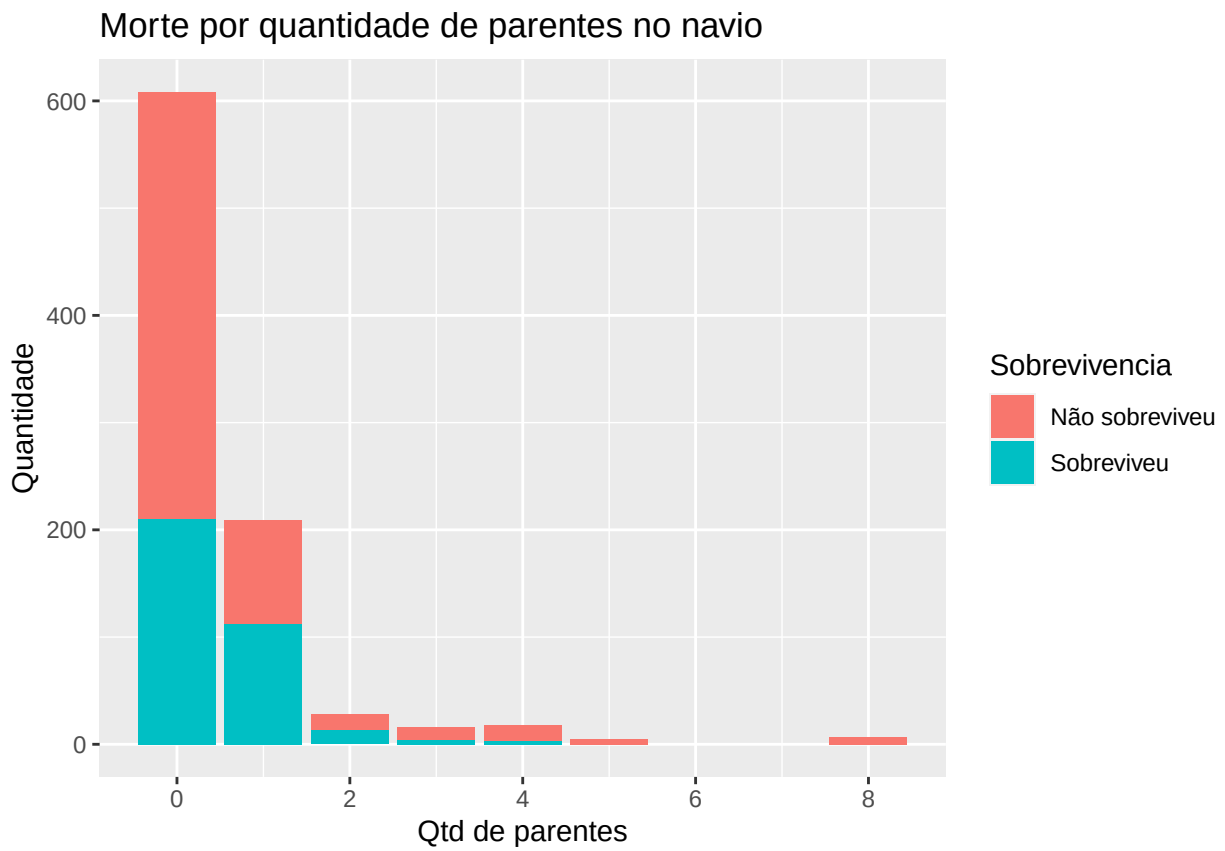
```
faltantes <- is.na(titanic$Age)

titanic[faltantes,5] <- round(mean(titanic[!faltantes,5]))

ggplot(data = titanic, aes(x = Age, fill = Survived))+
  geom_histogram()+
  labs(title = "Grafico por idade",
       x = "Idade",
       y = "Quantidade")+
  scale_fill_discrete(name = "Sobrevivencia", labels = c("Não sobreviveu", "Sobreviveu"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



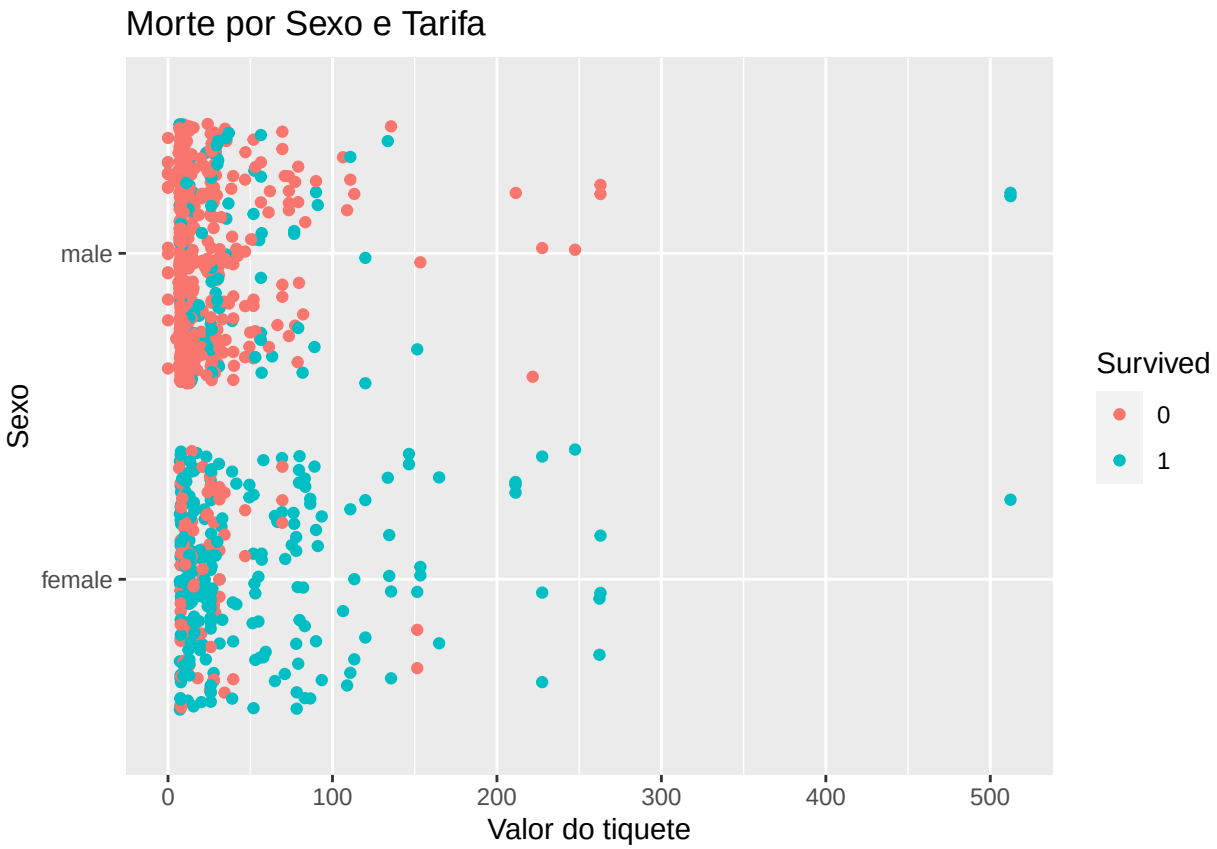

Já em relação à quantidade de parentes que embarcaram junto com cada indivíduo, obtemos o seguinte gráfico:

```
ggplot(data = titanic, aes(x = SibSp, fill = Survived))+
  geom_bar()+
  labs(title = "Morte por quantidade de parentes no navio",
       x = "Qtd de parentes",
       y = "Quantidade")+
  scale_fill_discrete(name = "Sobrevivencia", labels = c("Não sobreviveu", "Sobreviveu"))
```




Este já nos mostra que a probabilidade de sobrevivência quando se tinha mais de 2 parentes no navio era quase nula.

```
ggplot(data = titanic, aes(x = Fare, y = Sex , color = Survived))+
  geom_jitter()+
  labs(title = "Morte por Sexo e Tarifa",
       x = "Valor do tiquete",
       y = "Sexo")+
  scale_fill_discrete(name = "Sobrevivencia", labels = c("Não sobreviveu", "Sobreviveu"))
```




Outro gráfico interessante é o da tarifa cobrada por entrada, onde vemos que os que compraram os bilhetes mais caros tiveram uma probabilidade maior de sobrevivência.

## Conclusão

Pode-se concluir que a variável mais explicativa de todas é a do sexo, já que a probabilidade de sobrevivência foi muito maior para indivíduos do sexo feminino. O conjunto de variáveis também inclui outras variáveis relevantes para explicar o comportamento, como a Classe, onde a maioria das pessoas da Classe 3 não sobreviveu.

Por outro lado, variáveis como idade, devido ao grande número de dados faltantes, não contribuem muito para a nossa análise.