Trilha 6
Bruna Matos

Problema 1

a) Fazer a preparação dos dados para serem utilizados na análise, considerando que serão utilizadas apenas as variáveis Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked

i) Do conjunto de dados original, você deve selecionar um subconjunto apenas com as variáveis indicadas acima e a variável PassengerId.

```
# selecionando apenas as colunas solicitadas
# Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked
train_frame = data.frame(titanic_train[,c("Survived", "Pclass", "Sex", "Age", "SibSp", "Parch", "Fare", "Embarked","PassengerId")])
head(train_frame)
```

A data.frame: 6 × 9

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	PassengerId
	<int>	<int>	<ord>	<dbl>	<int>	<int>	<dbl>	<ord>	<int>
1	0	3	male	29	0	0	7.8542	S	2
2	1	1	female	37	1	0	90.0000	Q	3
3	0	2	male	36	0	0	13.0000	S	4
4	1	1	female	NA	0	1	55.0000	S	5
5	0	3	male	30	0	0	8.0500	S	6
6	0	3	male	21	1	0	6.4958	S	10

ii) Você deve atribuir um valor para os NAs na variável Age. Utilize algum critério razoável, por exemplo, o valor médio.

```
# Você deve  atribuir um valor para os NAs na variável Age.
#str(train_frame, strict.width = "wrap")
media_age = mean(train_frame$Age)
media_age
media_age_se_na = mean(train_frame$Age, na.rm = TRUE)
media_age_se_na
# Estou com elementeo NA que está dando problema no meu valor de media
train_frame$Age[is.na(train_frame$Age)]<-media_age_se_na
head(train_frame$Age)
```

```
<NA>
29.9936254520167
29 · 37 · 36 · 29.9936254520167 · 30 · 21
```

iii) Você deve remover as linhas onde ainda estiverem faltando dados, depois de atribuir o valor para os NAs de Age. Poucas linhas estarão ainda com dados faltantes.

```
# Você deve remover as linhas onde ainda estiverem faltando dados
train_frame_2 = na.omit(train_frame)
str(train_frame_2)
```

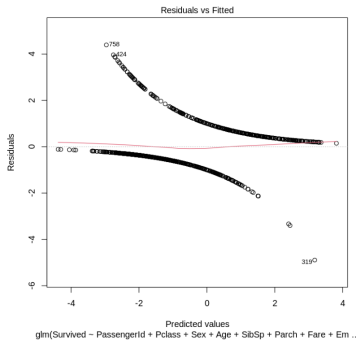
```
'data.frame':   889 obs. of  9 variables:
 $ Survived   : int  0 1 0 1 0 0 1 0 0 1 ...
 $ Pclass     : int  3 1 2 1 3 3 1 3 3 2 ...
 $ Sex        : Ord.factor w/ 2 levels "female"<"male": 2 1 2 1 2 2 2 2 2 2 ...
 $ Age        : num  29 37 36 30 30 ...
 $ SibSp      : int  0 1 0 0 0 1 1 0 0 0 ...
 $ Parch      : int  0 0 0 1 0 0 1 0 0 0 ...
 $ Fare       : num  7.85 90 13 55 8.05 ...
 $ Embarked   : Ord.factor w/ 3 levels "C"<"Q"<"S": 3 2 3 3 3 3 3 3 3 3 ...
 $ PassengerId: int  2 3 4 5 6 10 11 12 13 15 ...
 - attr(*, "na.action")= 'omit' Named int [1:2] 135 316
  ..- attr(*, "names")= chr [1:2] "135" "316"
```

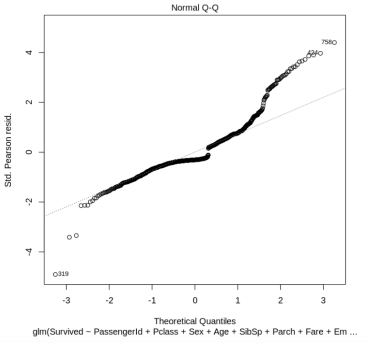
b) Você deve criar um modelo onde Survived será uma função das demais variáveis

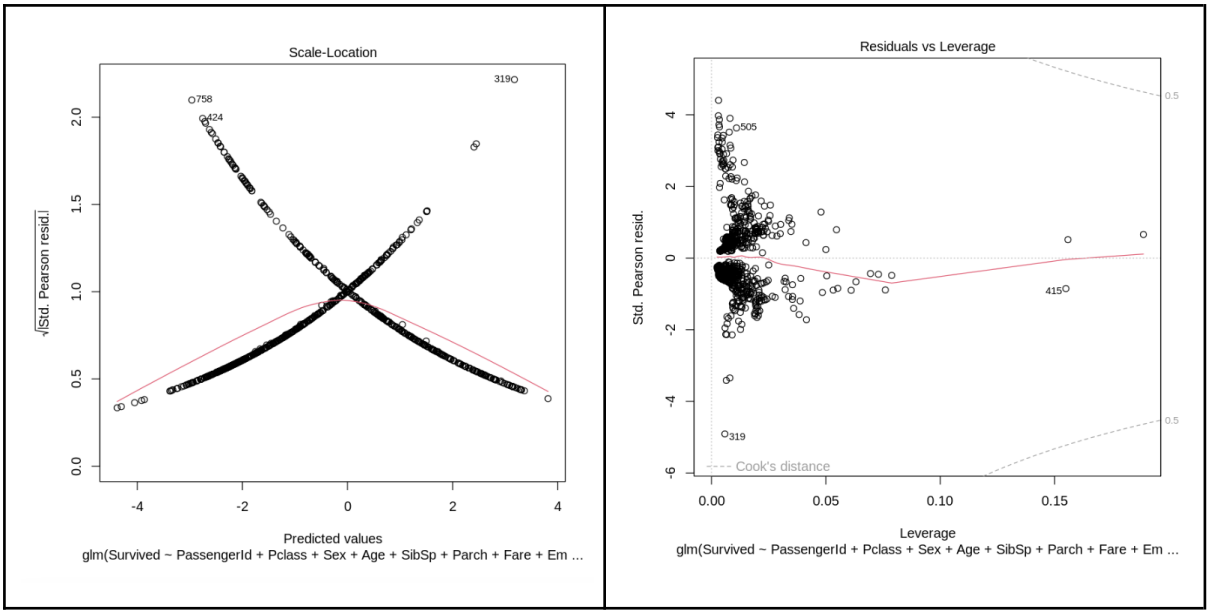
```
Call:
glm(formula = Survived ~ PassengerId + Pclass + Sex + Age + SibSp +
    Parch + Fare + Embarked, family = binomial, data = train_frame_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5364  -0.6484  -0.4239   0.6593   2.4542

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.2653669  0.5257234   6.211 5.26e-10 ***
PassengerId -0.0004932  0.0002363  -2.087   0.0369 *
Pclass      -0.9592678  0.1412381  -6.792 1.11e-11 ***
Sex.L       -1.8800665  0.1403031 -13.400  < 2e-16 ***
Age         -0.0303198  0.0073321  -4.135 3.55e-05 ***
SibSp       -0.2285917  0.1035452  -2.208   0.0273 *
Parch       -0.0874034  0.1013408  -0.862   0.3884
Fare         0.0026262  0.0022531   1.166   0.2438
Embarked.L  -0.3197643  0.1594599  -2.005   0.0449 *
Embarked.Q   0.1621650  0.2629128   0.617   0.5374
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1177.91  on 888  degrees of freedom
Residual deviance:  809.79  on 879  degrees of freedom
AIC: 829.79
```

c) Faça as análises do modelo, verificando a significância estatística das variáveis (e seus parâmetros ajustados), gráficos diagnósticos, etc.

```
Number of Fisher Scoring iterations: 5
```







Analisando os valores de Pr podemos dizer que as variáveis Parch, Fare, Embarked.Q não fazem uma contribuição significativa para o modelo.Ainda analisando os gráficos de dispersão e Q-Q percebemos que não temos uma reta no de dispersão com os pontos localizados de forma coerente e no Q-Q eles não estão em cima da reta que também não está a 45 graus.

d) Atualize o modelo como consequência da análise realizada no item anterior.

Removendo Parch

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5581  -0.6482  -0.4201   0.6875   2.4633

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.2971536  0.5233025   6.301 2.96e-10 ***
PassengerId -0.0004916  0.0002362  -2.082   0.0374 *
Pclass      -0.9781560  0.1392470  -7.025 2.15e-12 ***
Sex.L       -1.8537708  0.1363859 -13.592  < 2e-16 ***
Age         -0.0302377  0.0073232  -4.129 3.64e-05 ***
SibSp       -0.2475921  0.1016167  -2.437   0.0148 *
Fare         0.0021315  0.0021280   1.002   0.3165
Embarked.L  -0.3284219  0.1591008  -2.064   0.0390 *
Embarked.Q   0.1313794  0.2597847   0.506   0.6131
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1177.91  on 888  degrees of freedom
Residual deviance:  810.55  on 880  degrees of freedom
AIC: 828.55
```

Removendo Fare

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4739  -0.6501  -0.4180   0.6616   2.4695

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.5292444  0.4777239   7.388 1.49e-13 ***
PassengerId -0.0004905  0.0002359  -2.079   0.0376 *
Pclass      -1.0426064  0.1231231  -8.468  < 2e-16 ***
Sex.L       -1.8810196  0.1400352 -13.432  < 2e-16 ***
Age         -0.0304601  0.0073151  -4.164 3.13e-05 ***
SibSp       -0.2089870  0.1017334  -2.054   0.0400 *
Parch       -0.0567297  0.0974976  -0.582   0.5607
Embarked.L  -0.3588024  0.1558706  -2.302   0.0213 *
Embarked.Q   0.1592924  0.2627911   0.606   0.5444
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1177.91  on 888  degrees of freedom
Residual deviance:  811.28  on 880  degrees of freedom
AIC: 829.28
```

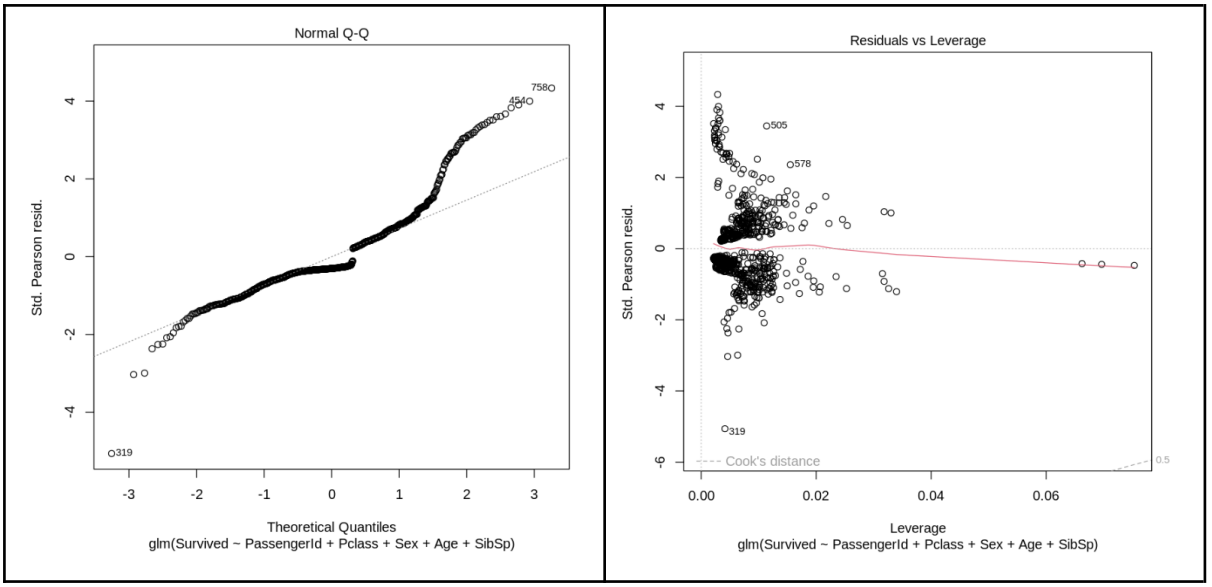
Remove Embarked

```
Call:
glm(formula = Survived ~ PassengerId + Pclass + Sex + Age + SibSp +
    Parch + Fare, family = binomial, data = train_frame_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6092  -0.6500  -0.4265   0.6653   2.4264

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.2316971  0.4978582   6.491 8.52e-11 ***
PassengerId -0.0005029  0.0002356  -2.135   0.0328 *
Pclass      -0.9781281  0.1365689  -7.162 7.94e-13 ***
Sex.L       -1.8801617  0.1369584 -13.728  < 2e-16 ***
Age         -0.0301831  0.0072766  -4.148 3.35e-05 ***
SibSp       -0.2427895  0.1026109  -2.366   0.0180 *
Parch       -0.0967103  0.0995959  -0.971   0.3315
Fare         0.0034439  0.0022167   1.554   0.1203
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Removendo as três

```
Call:
glm(formula = Survived ~ PassengerId + Pclass + Sex + Age + SibSp,
    family = binomial, data = train_frame_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5600  -0.6591  -0.4192   0.6523   2.4419

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.5435970  0.4503378   7.869 3.58e-15 ***
PassengerId -0.0004998  0.0002349  -2.128   0.0334 *
Pclass      -1.0914865  0.1175731  -9.283  < 2e-16 ***
Sex.L       -1.8676684  0.1335704 -13.983  < 2e-16 ***
Age         -0.0301062  0.0072448  -4.156 3.24e-05 ***
SibSp       -0.2371829  0.0973233  -2.437   0.0148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1177.91  on 888  degrees of freedom
Residual deviance:  816.94  on 883  degrees of freedom
AIC: 828.94
```

Pudemos perceber uma melhora no modelo, visto que não temos mais nenhuma variável com Pr > 5%.
Porém ao analisarmos os gráficos de Resíduos e Q-Q abaixo, vemos que houve uma piora.



e) Faça as previsões da variável Survived na base de dados de teste utilizando o modelo refinado, e prepare um arquivo CSV para submissão que contenha apenas duas colunas: PassengerId, Survived.

f) Submeta seu arquivo e também o script R com todas as análises realizadas, até a criação do arquivo CSV.