Machine Learning with the Titanic Data Set

Alex McIntosh

8/19/2022

Contents

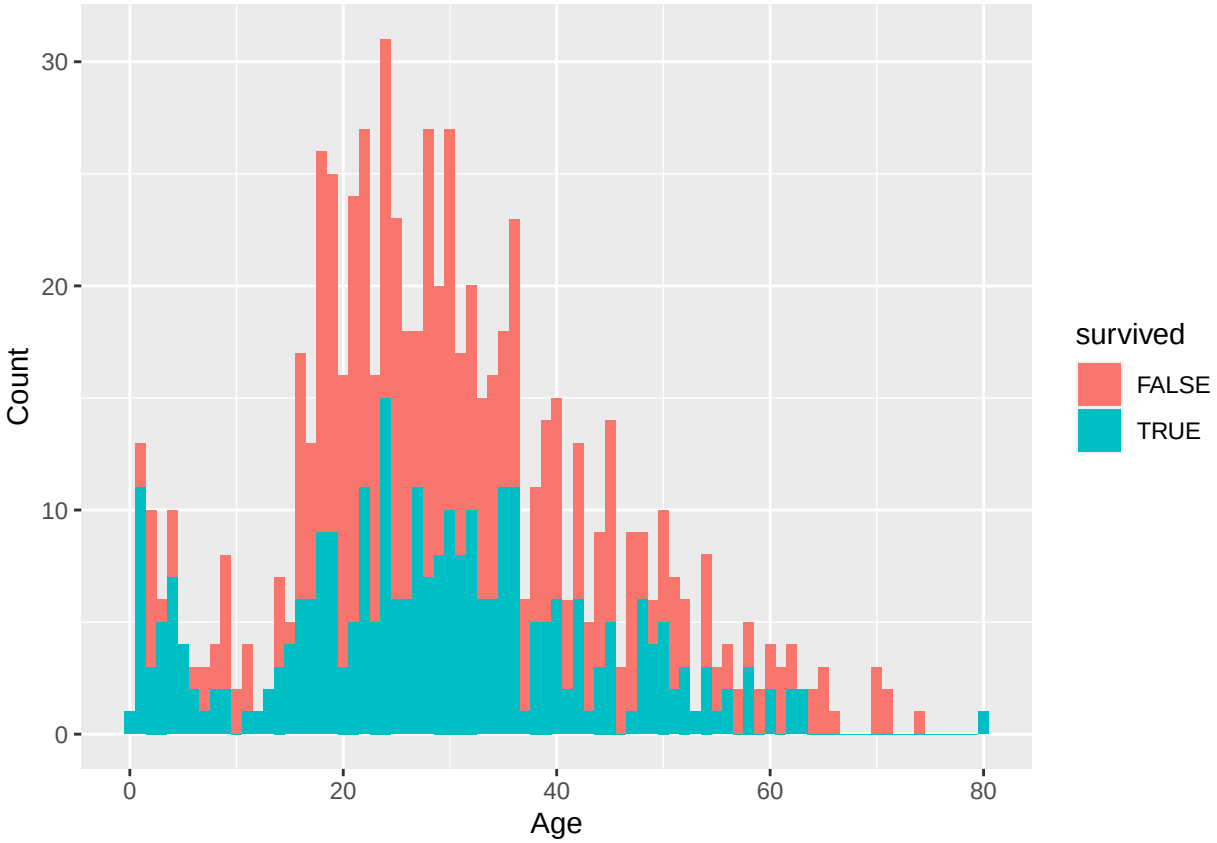
1 Cover Page

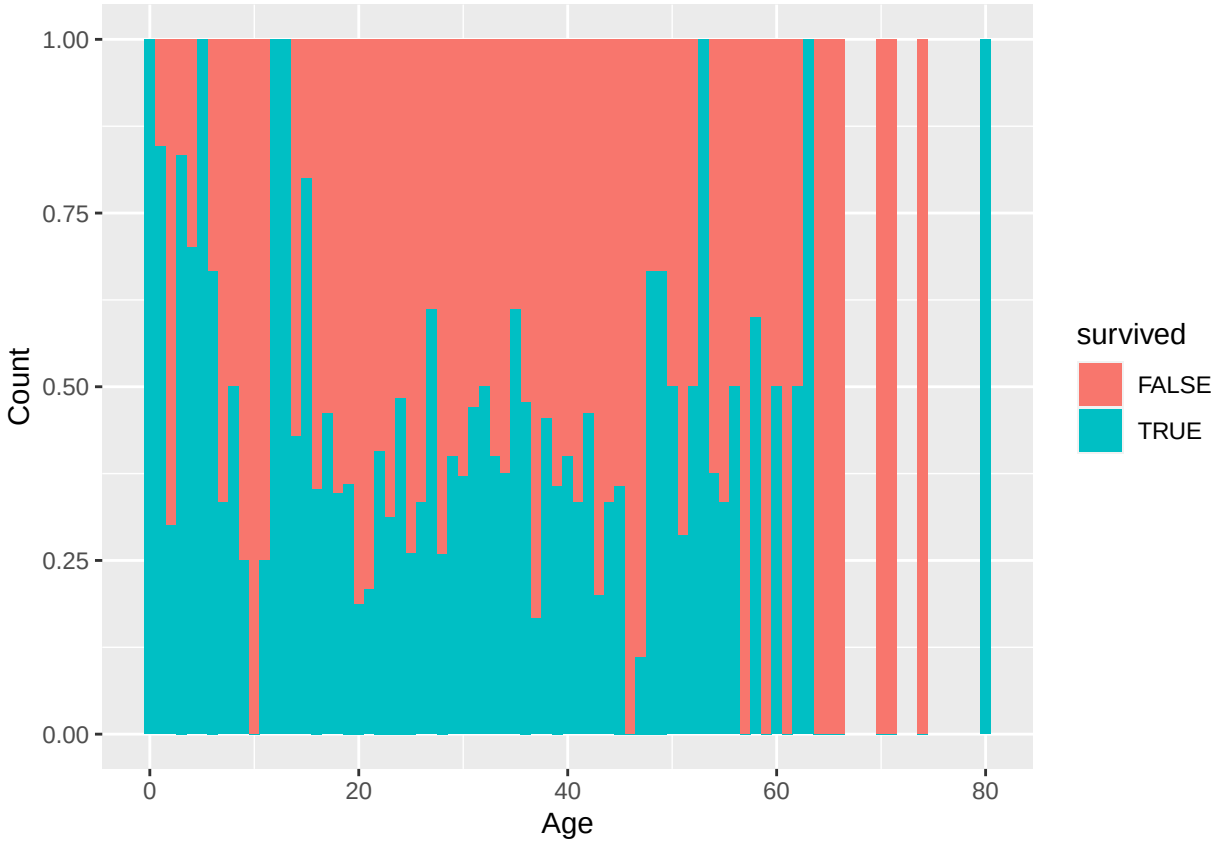
2 Table of Contents

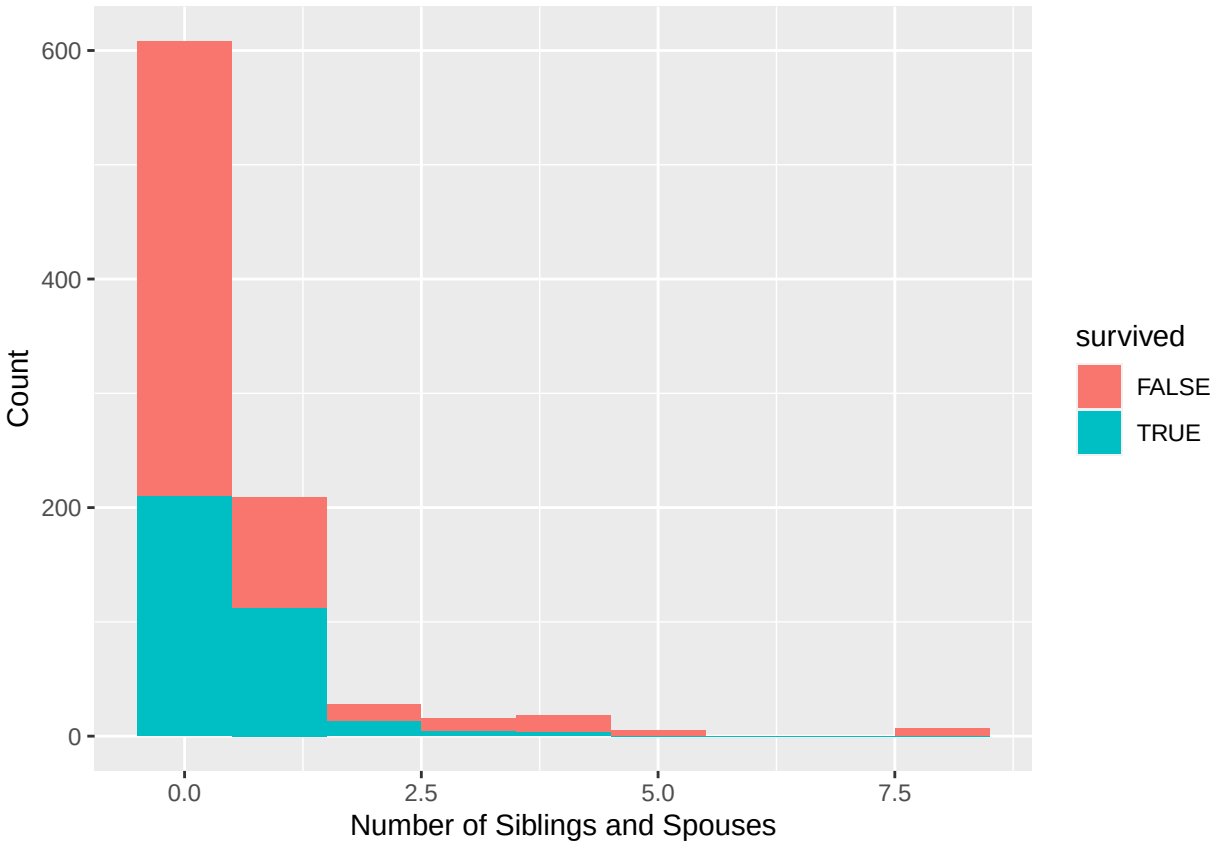
3 Executive Summary / Abstract

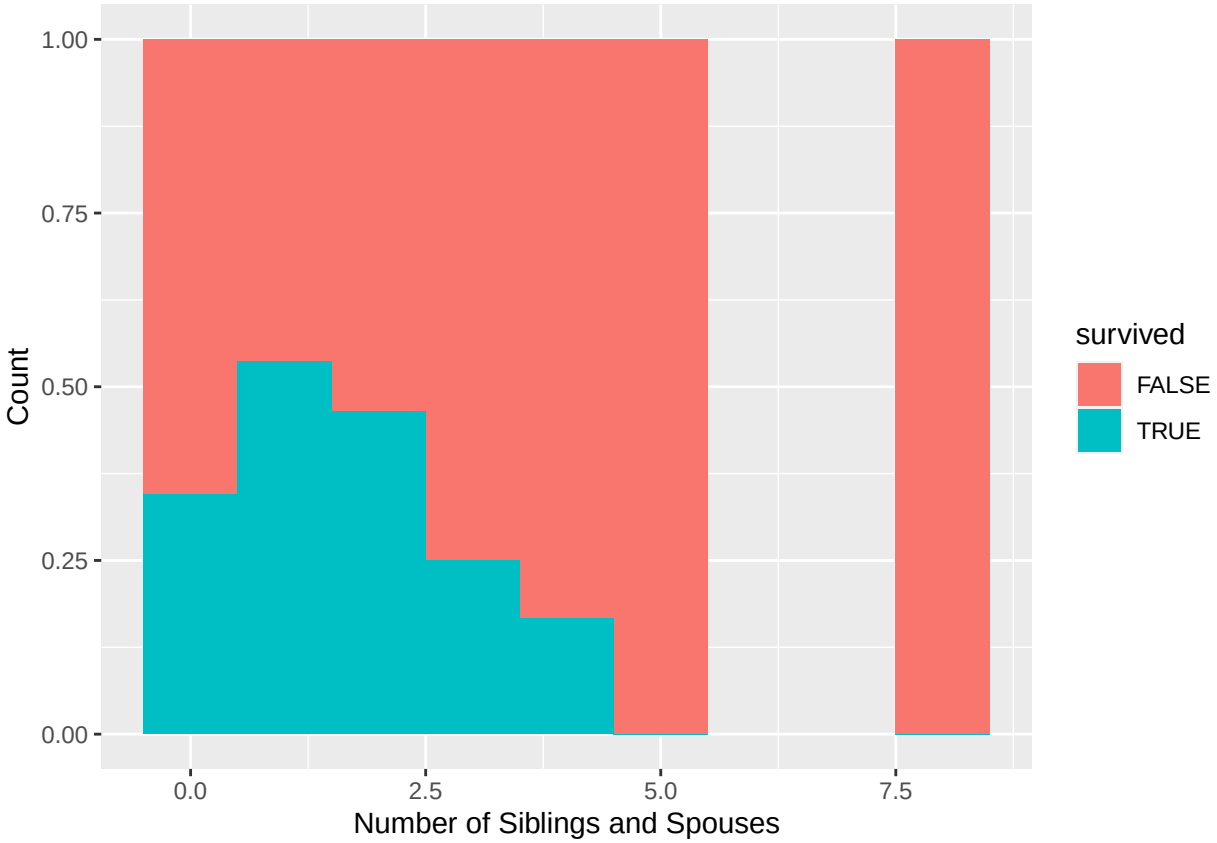
4 Methodology

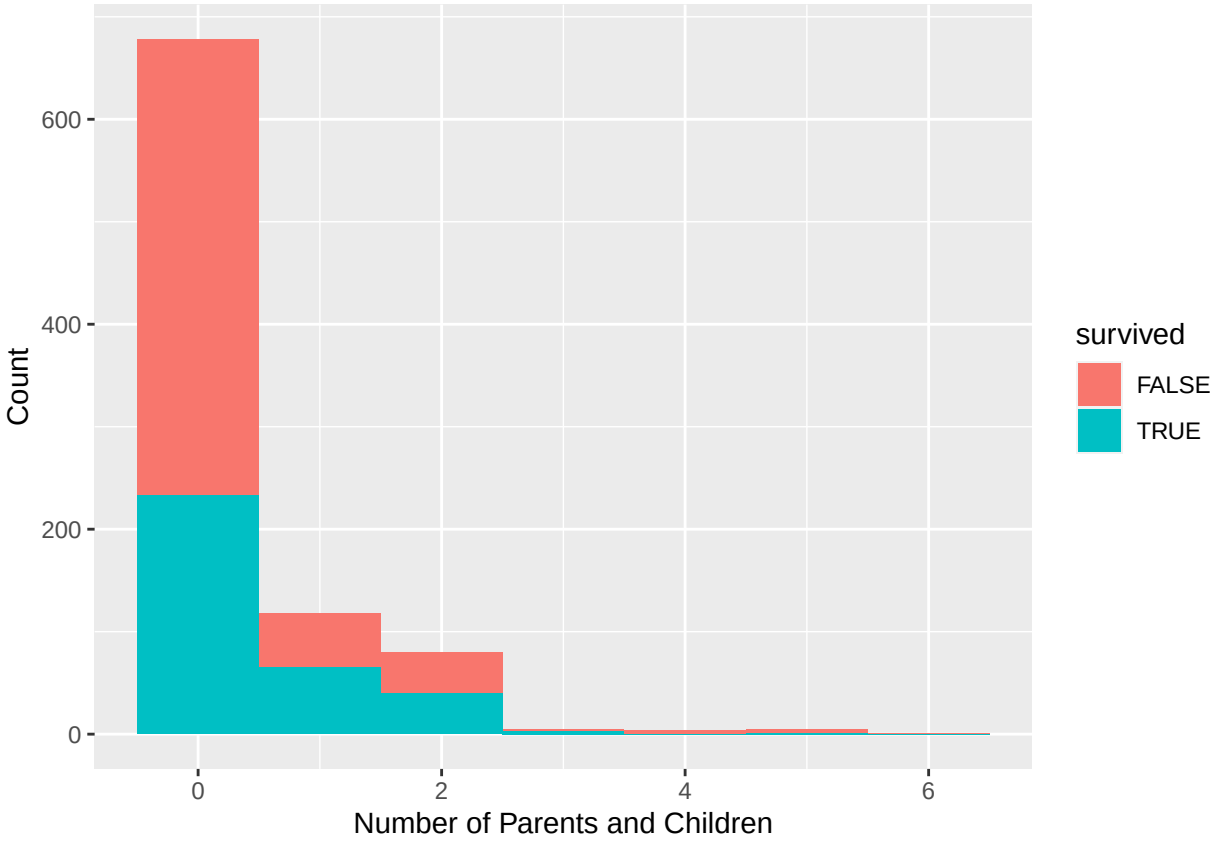
```
## # A tibble: 5 x 5
##   Variable   Min   Max    Mean      SD
##   <chr>    <dbl> <dbl>   <dbl>   <dbl>
## 1 id         1     891  446     257.
## 2 age        0.42   80   29.7    14.5
## 3 sib_sp     0       8    0.523   1.10
## 4 par_ch     0       6    0.382   0.806
## 5 fare       0     512.  32.2    49.7
```

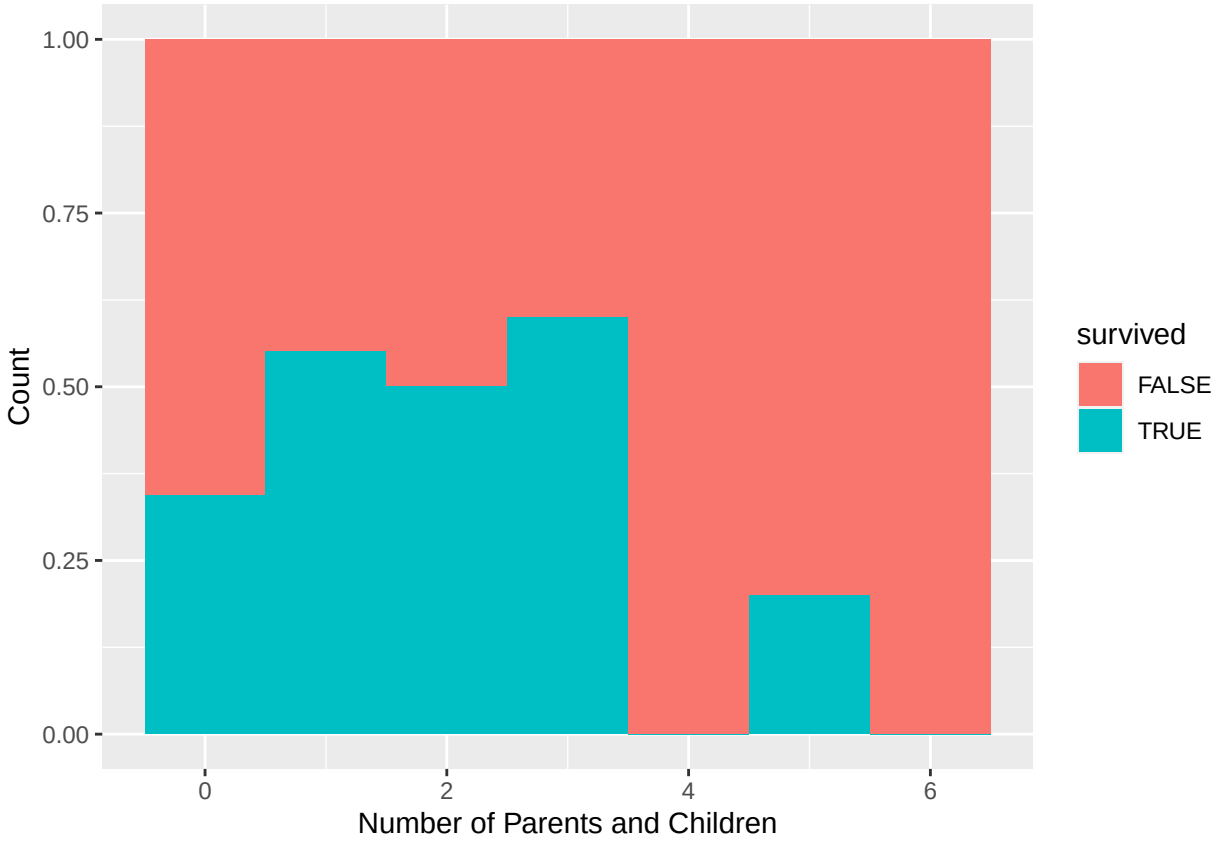

30
20
10
0
Count
0
20
40
60
80
Age
survived
FALSE
TRUE

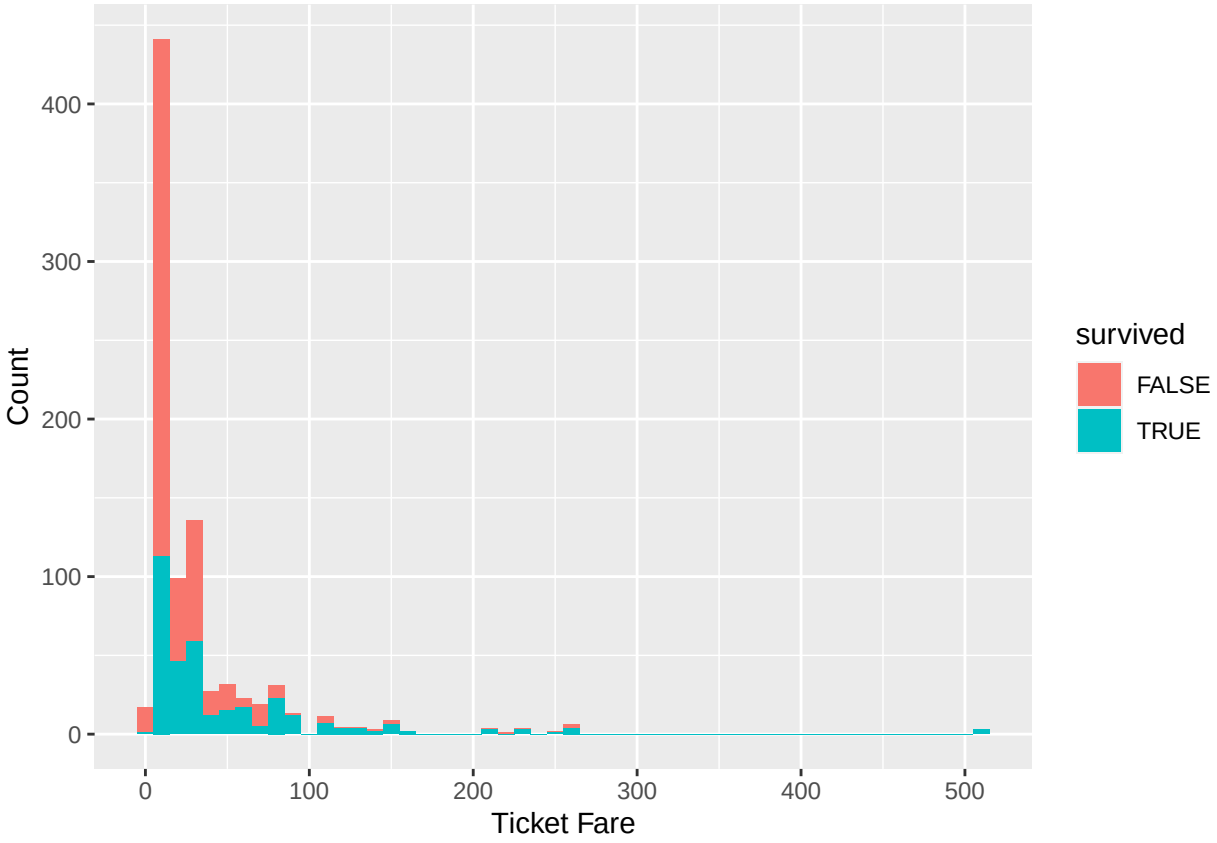

1.00
0.75
0.50
0.25
0.00
Count
0
20
40
60
80
Age
survived
FALSE
TRUE

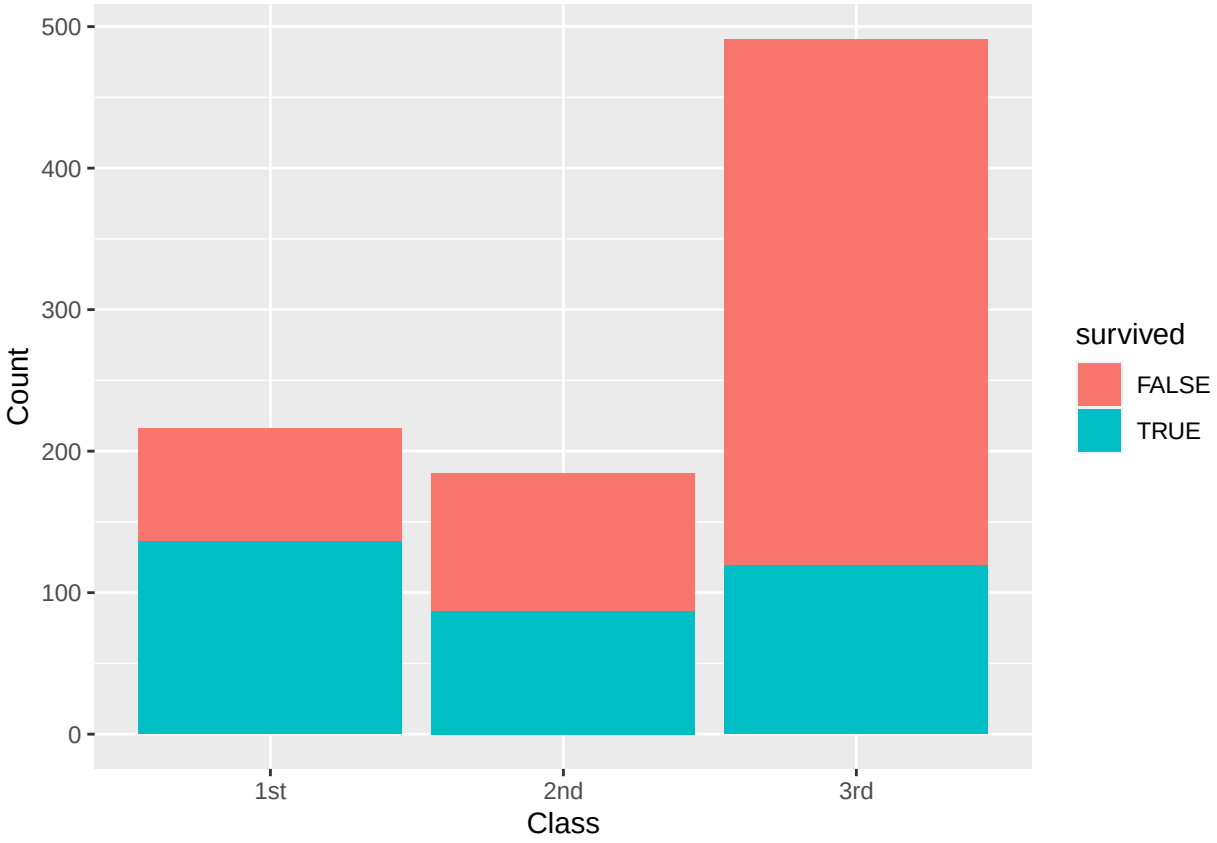

600
400
200
0
Count
0.0
2.5
5.0
7.5
Number of Siblings and Spouses
survived
FALSE
TRUE

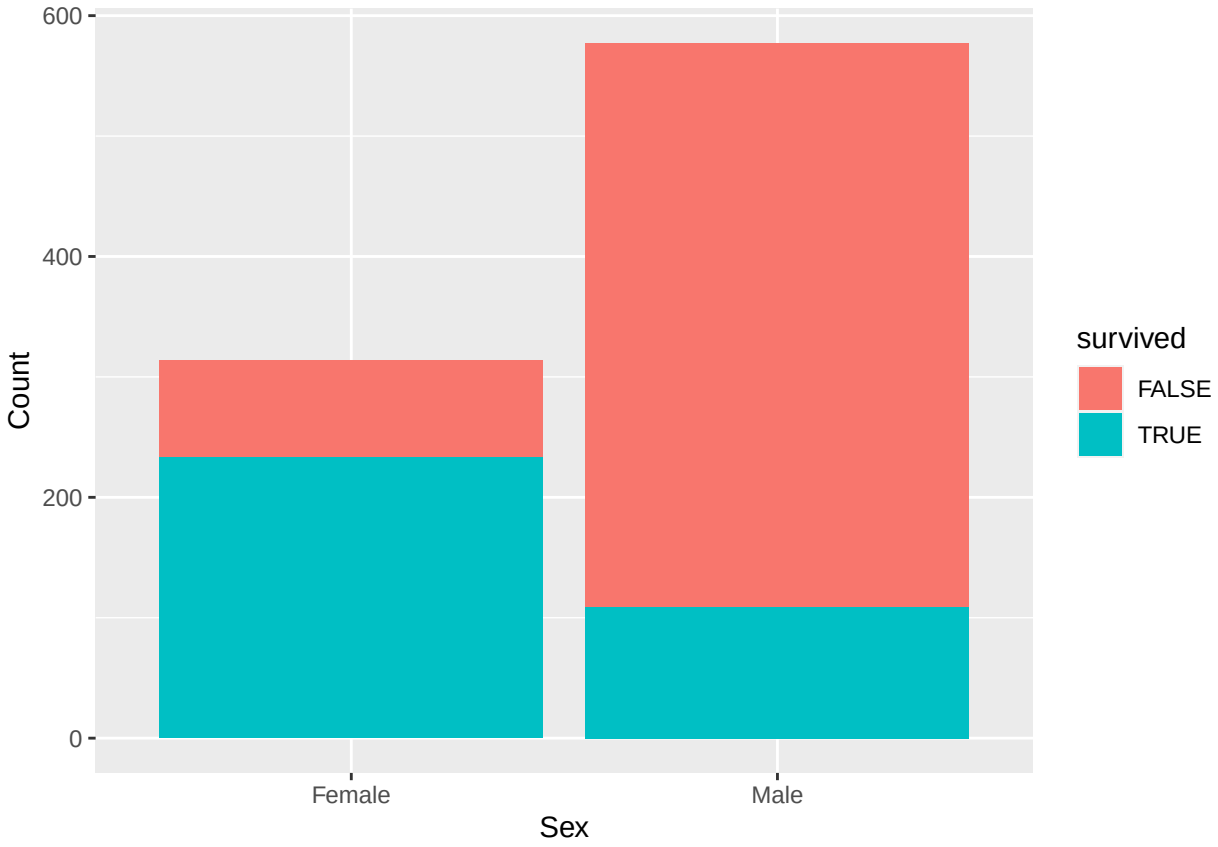

1.00
0.75
0.50
0.25
0.00
Count
0.0
2.5
5.0
7.5
Number of Siblings and Spouses
survived
FALSE
TRUE

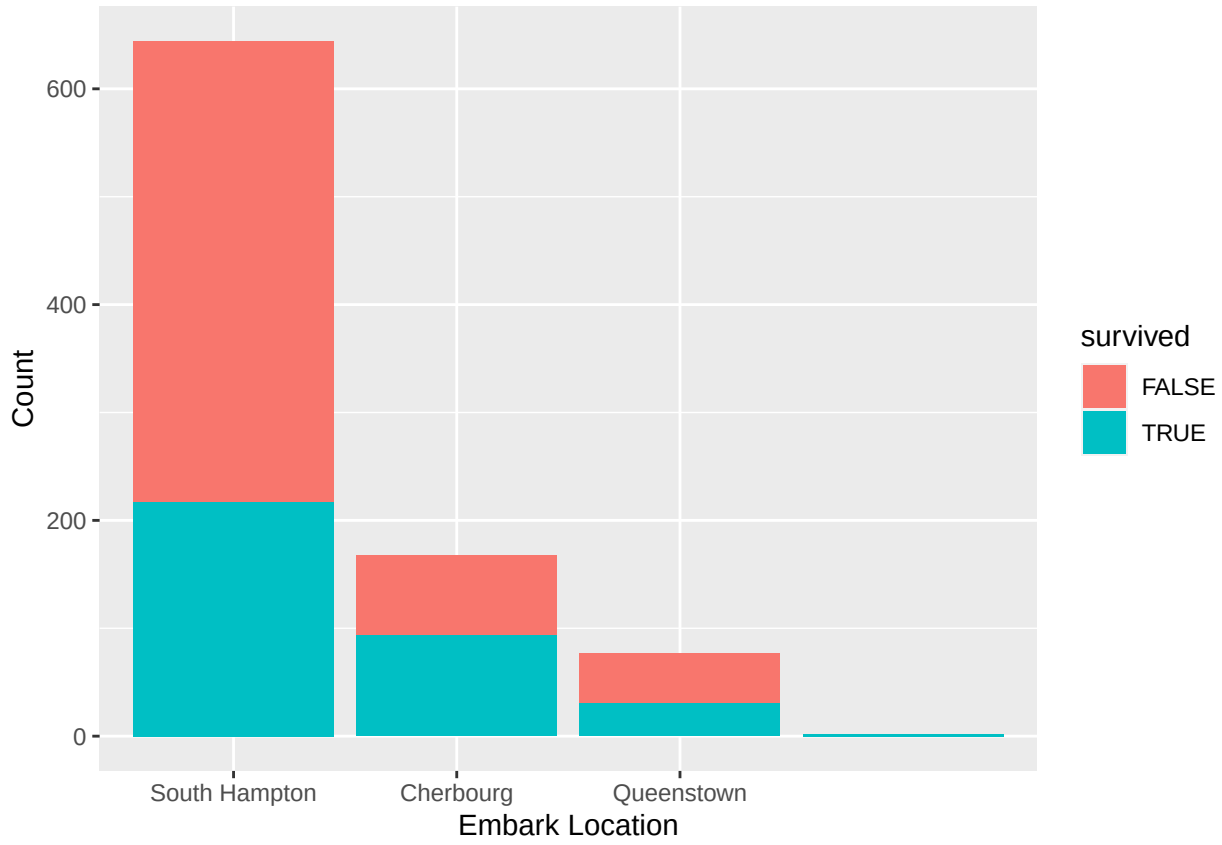

600
400
200
0
Count
0
2
4
6
Number of Parents and Children
survived
FALSE
TRUE


1.00
0.75
0.50
0.25
0.00
Count
0
2
4
6
Number of Parents and Children
survived
FALSE
TRUE

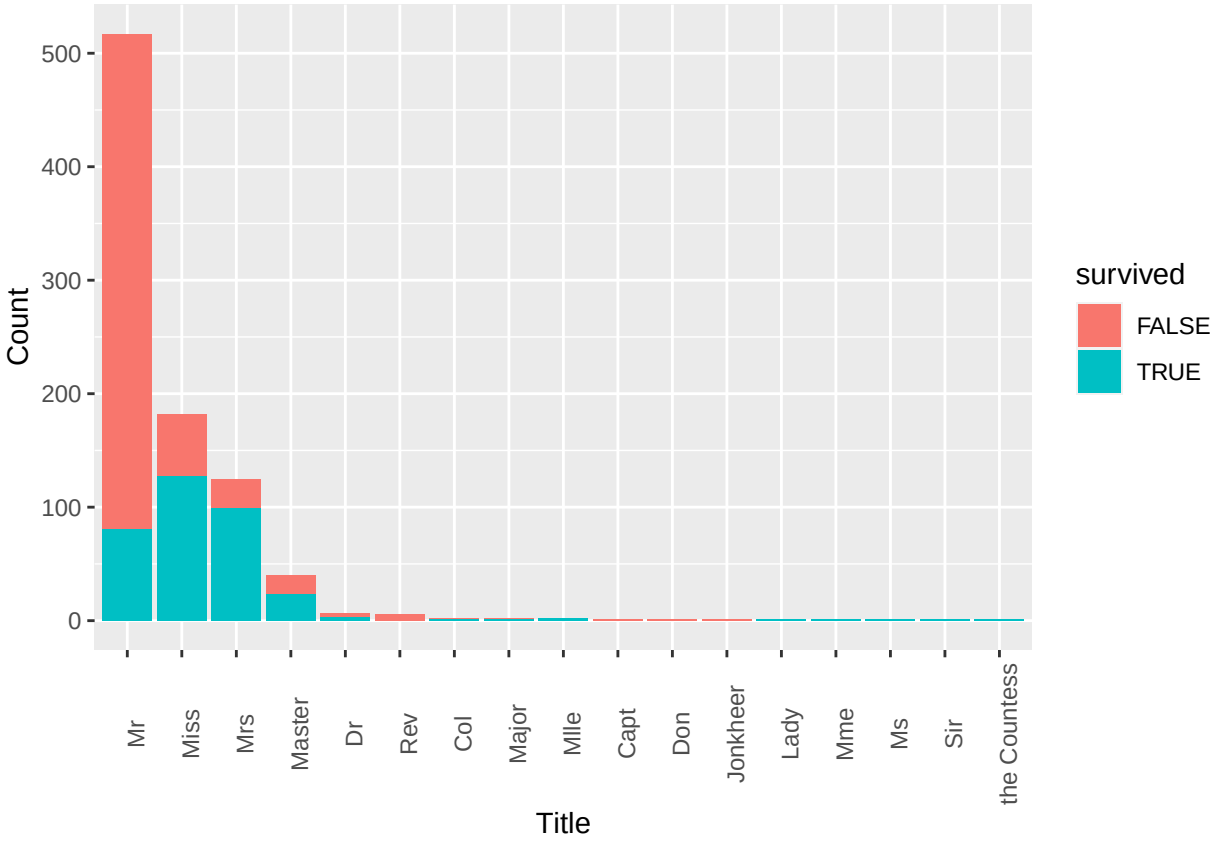

Count
400
300
200
100
0
0
100
200
300
400
500
Ticket Fare
survived
FALSE
TRUE

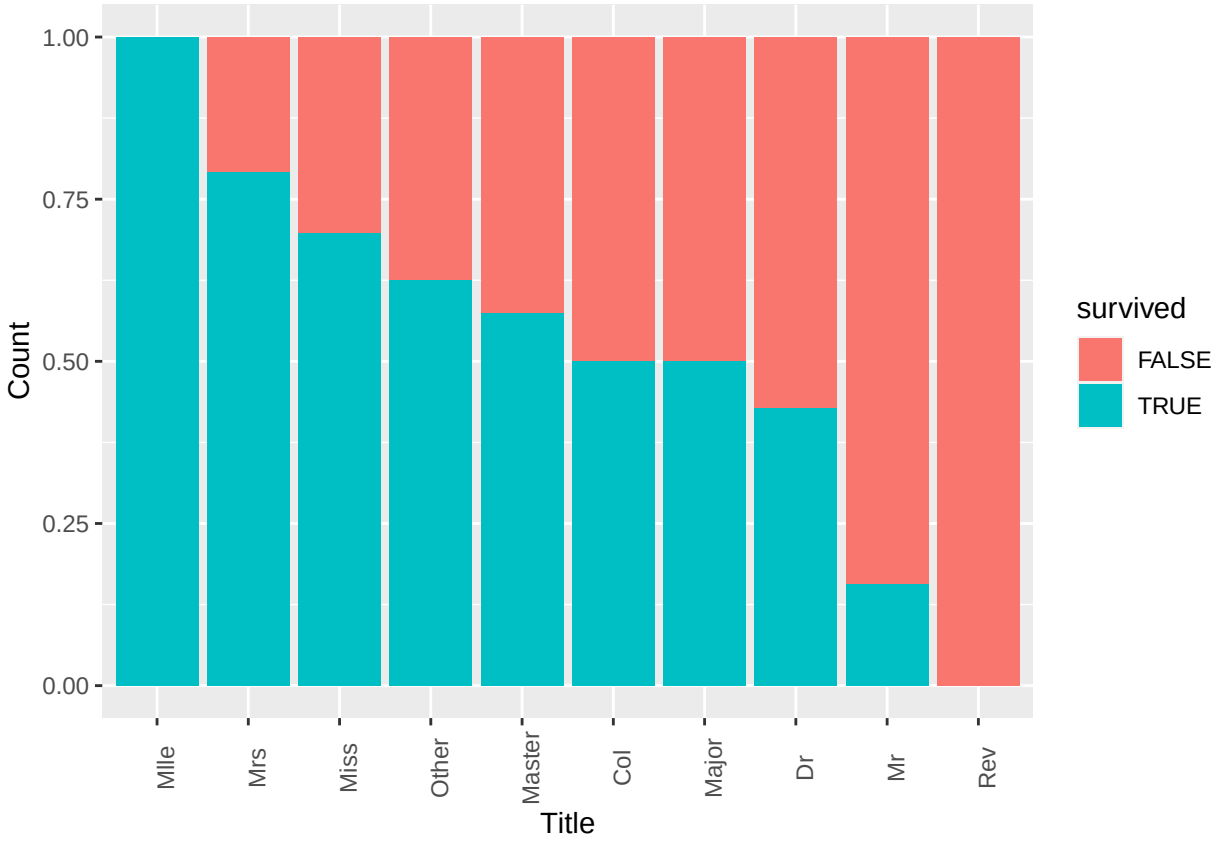

500
400
300
200
100
0
Count
1st
2nd
3rd
Class
survived
FALSE
TRUE

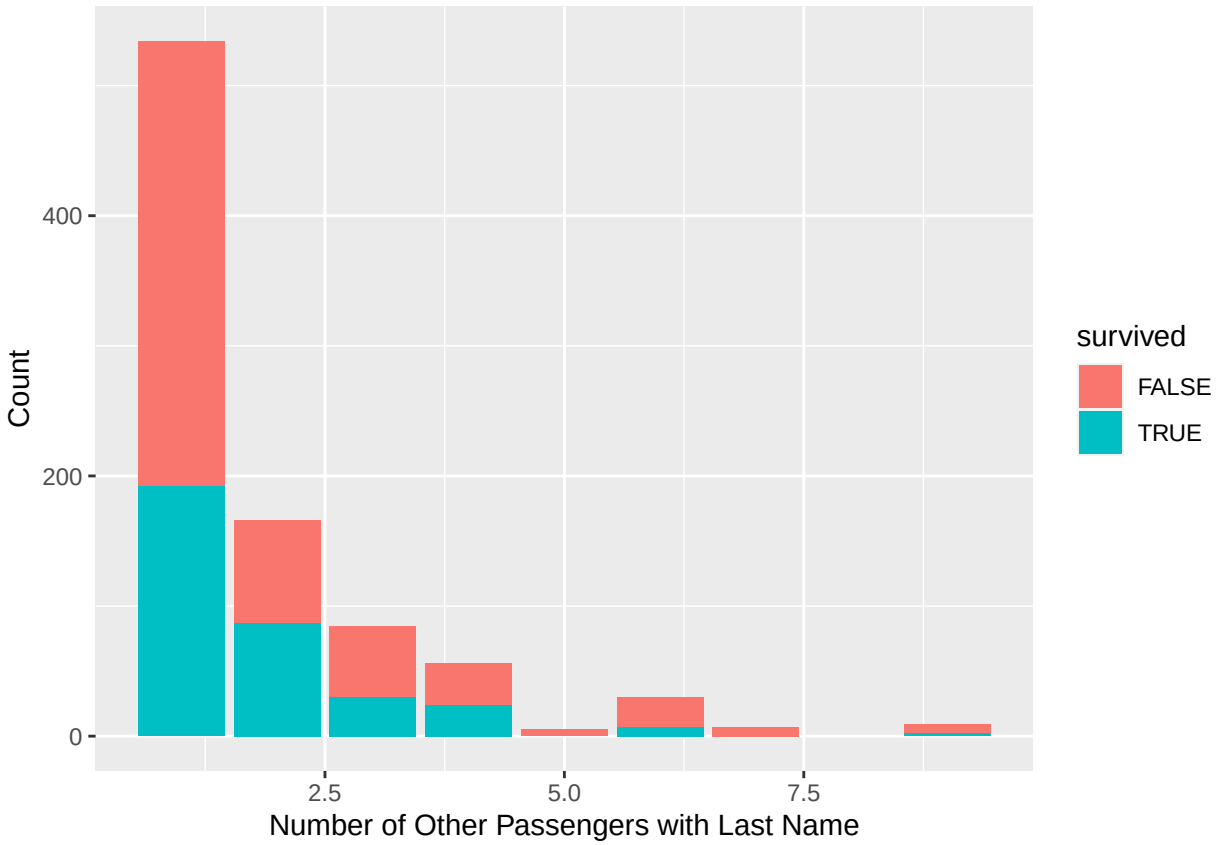

600
400
200
0
Count
Female
Male
Sex
survived
FALSE
TRUE

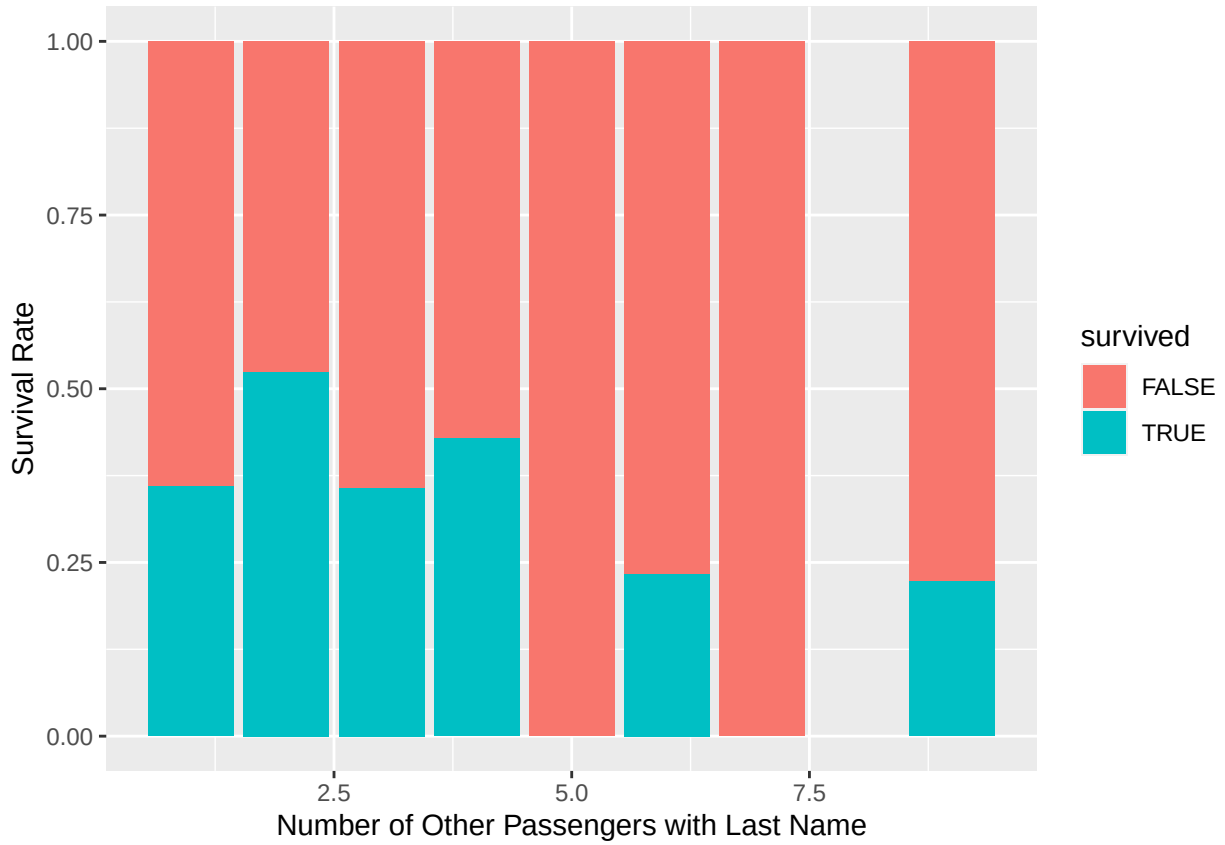



```
## # A tibble: 6 x 4
##      id last_name title first_name
##   <int> <chr>     <chr> <chr>
## 1     1 Braund    Mr    Owen
## 2     2 Cumings   Mrs   John
## 3     3 Heikkinen Miss  Laina
## 4     4 Futrelle  Mrs   Jacques
## 5     5 Allen     Mr    William
## 6     6 Moran     Mr    James
```

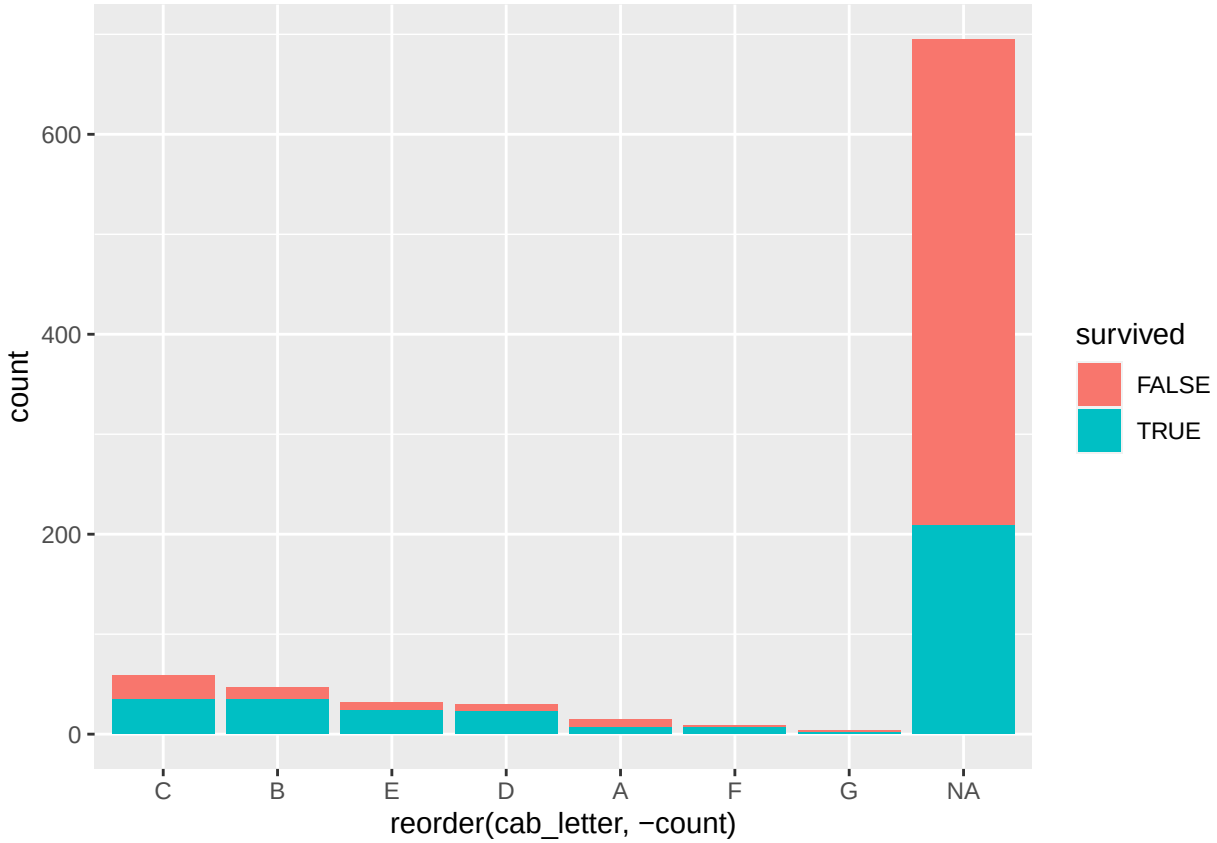

Count
500
400
300
200
100
0
Mr
Miss
Mrs
Master
Dr
Rev
Col
Major
Mlle
Capt
Don
Jonkheer
Lady
Mme
Ms
Sir
the Countess
Title
survived
FALSE
TRUE

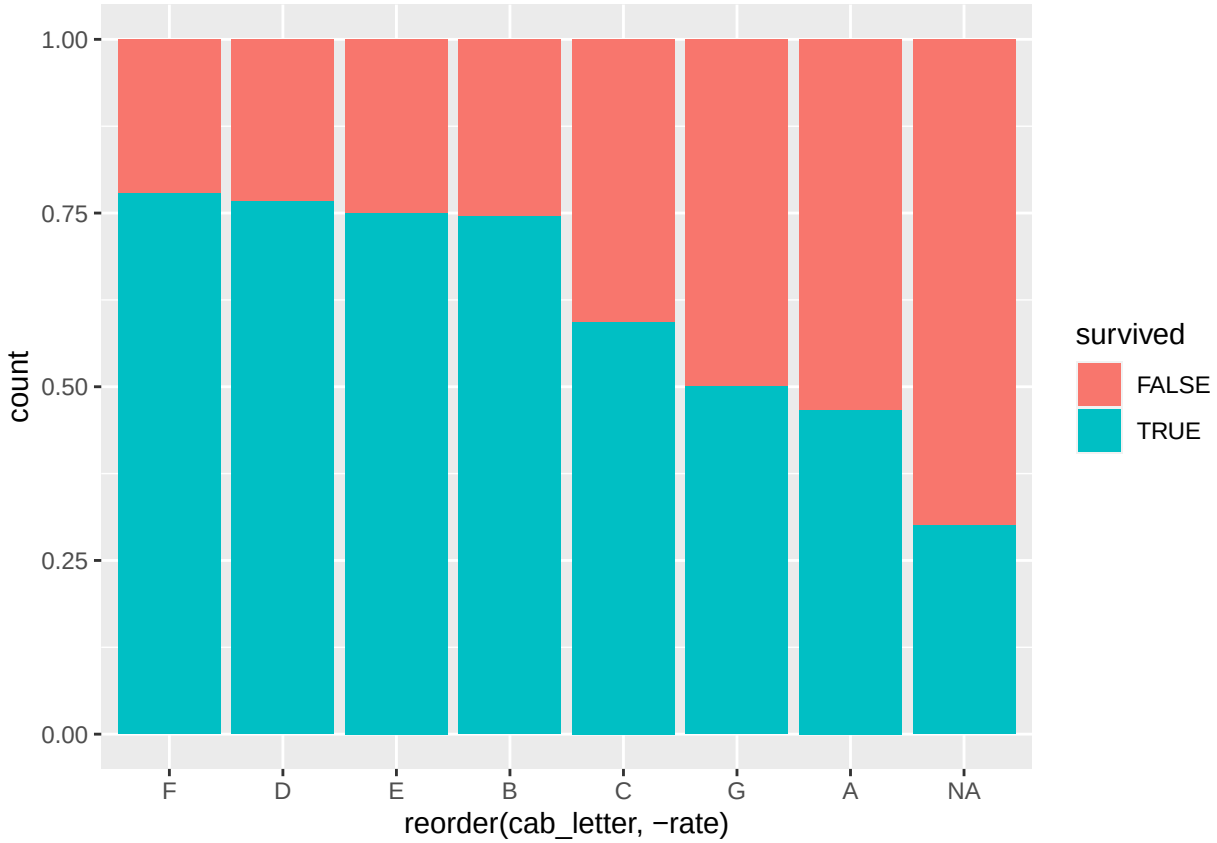

1.00
0.75
0.50
0.25
0.00
Count
Mlle
Mrs
Miss
Other
Master
Col
Major
Dr
Mr
Rev
Title
survived
FALSE
TRUE

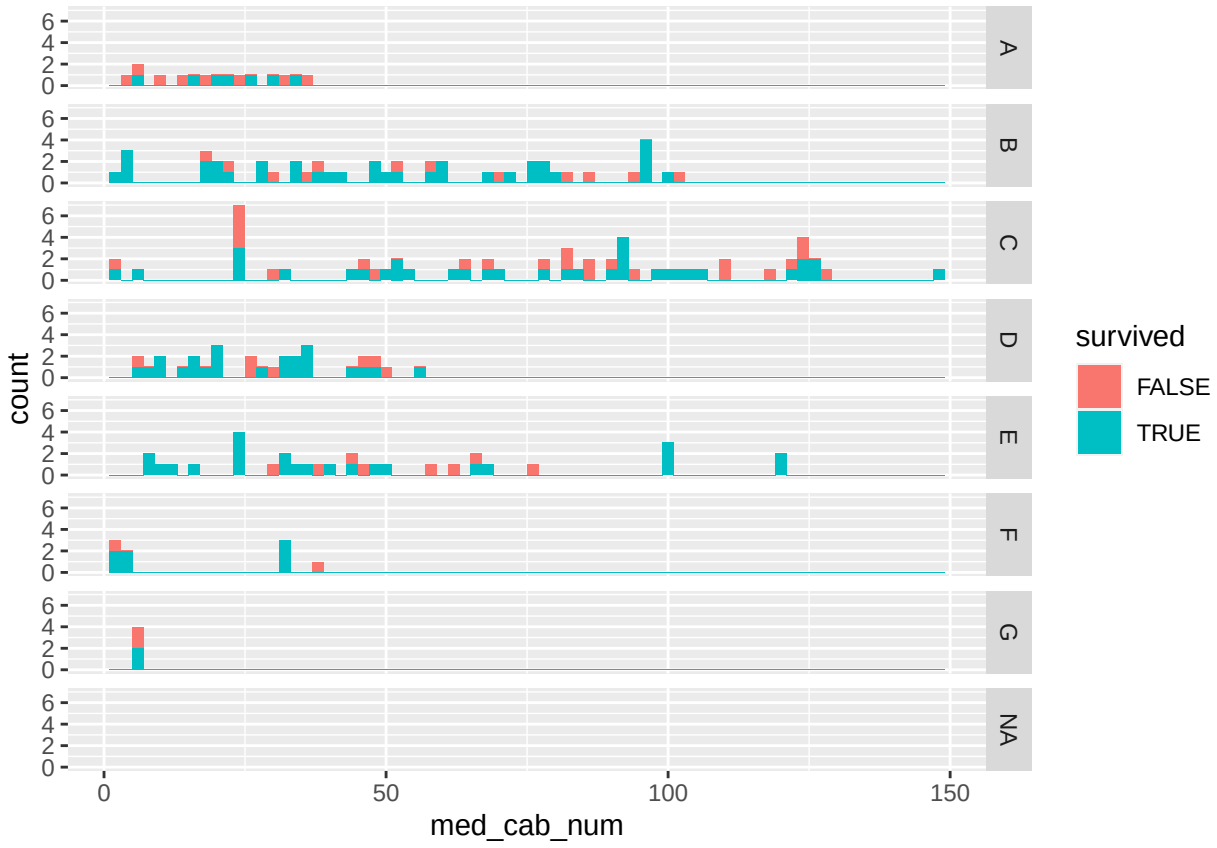

Count
400
200
0
2.5
5.0
7.5
Number of Other Passengers with Last Name
survived
FALSE
TRUE




```
## # A tibble: 891 x 4
## # Rowwise:
##       id cab_letter med_cab_num num_cab
##    <int> <chr>            <dbl>   <int>
##  1     1 <NA>                NA       0
##  2     2 C                   85       1
##  3     3 <NA>                NA       0
##  4     4 C                  123       1
##  5     5 <NA>                NA       0
##  6     6 <NA>                NA       0
##  7     7 E                   46       1
##  8     8 <NA>                NA       0
##  9     9 <NA>                NA       0
## 10    10 <NA>                NA       0
## # ... with 881 more rows
```

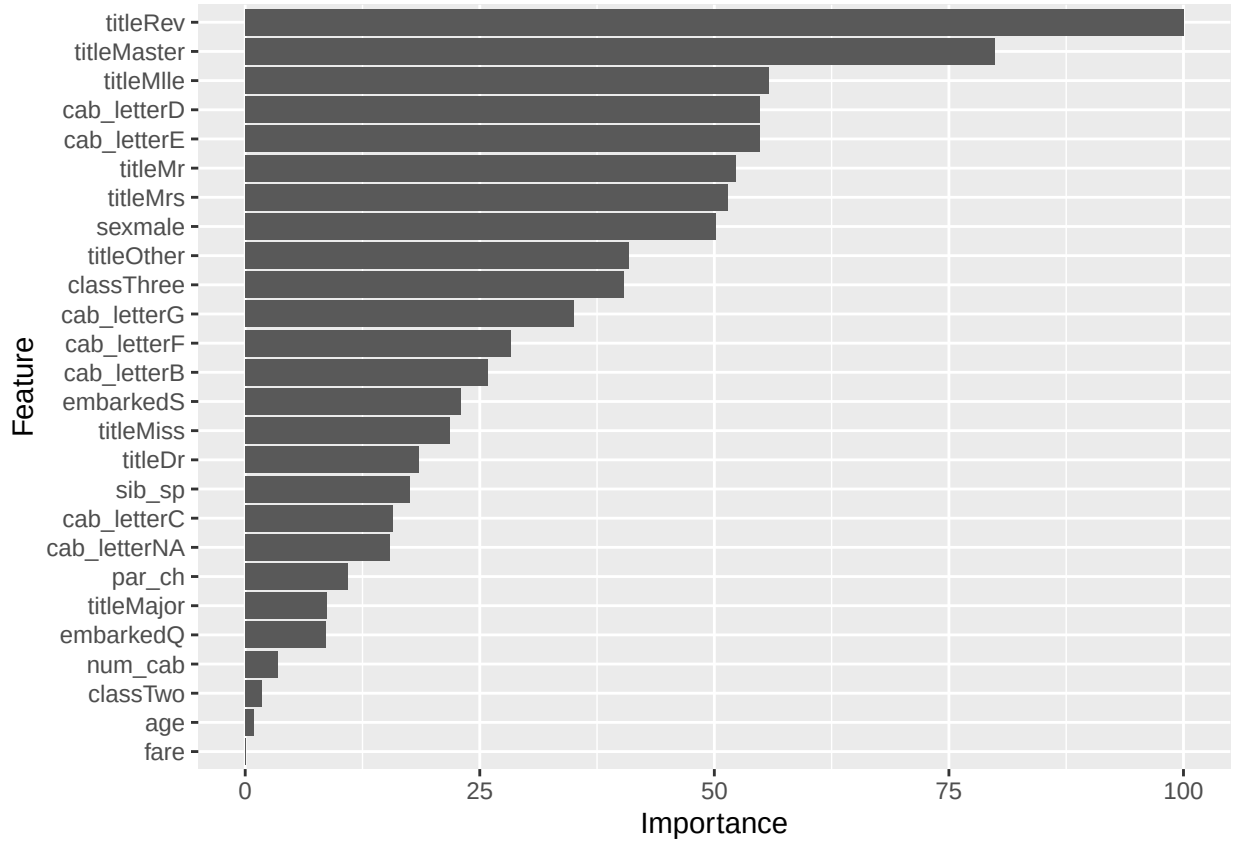

600
400
200
0
count
C
B
E
D
A
F
G
NA
reorder(cab_letter, −count)
survived
FALSE
TRUE

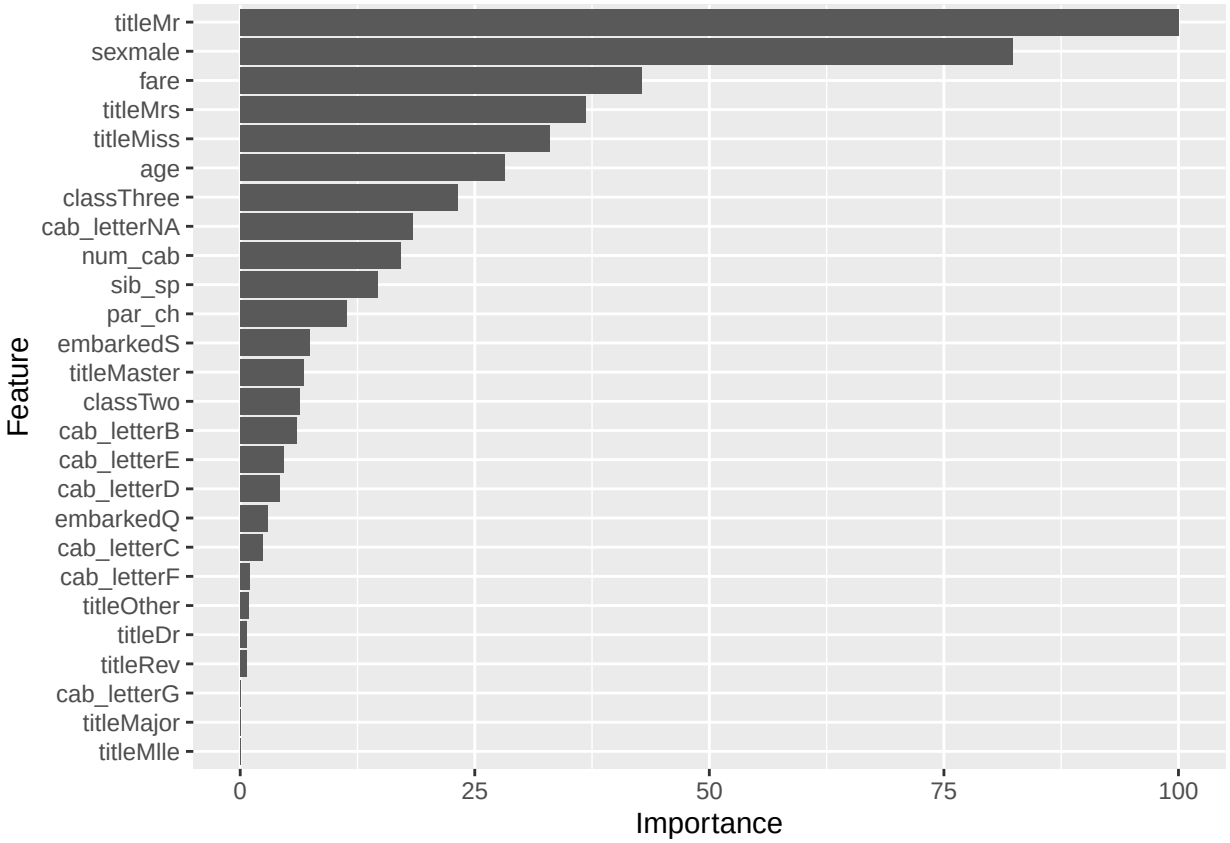

1.00
0.75
0.50
0.25
0.00
count
F
D
E
B
C
G
A
NA
reorder(cab_letter, −rate)
survived
FALSE
TRUE

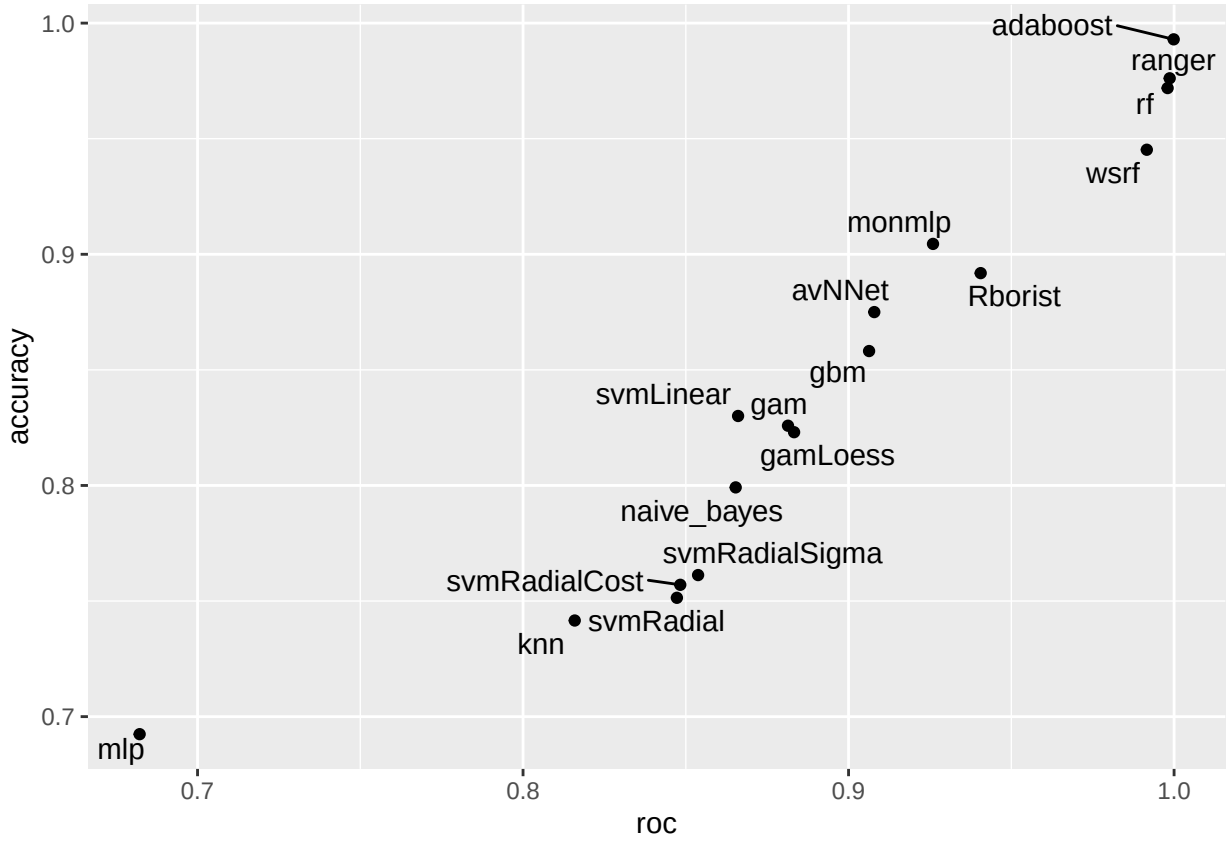



```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##        id  survived     class      name       sex       age    sib_sp    par_ch
##        ""        ""        ""        ""        ""     "pmm"        ""        ""
##    ticket      fare     cabin  embarked
##        ""        ""        "" "polyreg"
## PredictorMatrix:
##          id survived class name sex age sib_sp par_ch ticket fare cabin
## id        0        1     1    0   1   1      1      1      0    1     0
## survived  1        0     1    0   1   1      1      1      0    1     0
## class     1        1     0    0   1   1      1      1      0    1     0
## name      1        1     1    0   1   1      1      1      0    1     0
## sex       1        1     1    0   0   1      1      1      0    1     0
## age       1        1     1    0   1   0      1      1      0    1     0
##          embarked
## id              1
## survived        1
## class           1
## name            1
## sex             1
## age             1
## Number of logged events:  3
##   it im dep     meth    out
## 1  0  0     constant   name
## 2  0  0     constant ticket
```

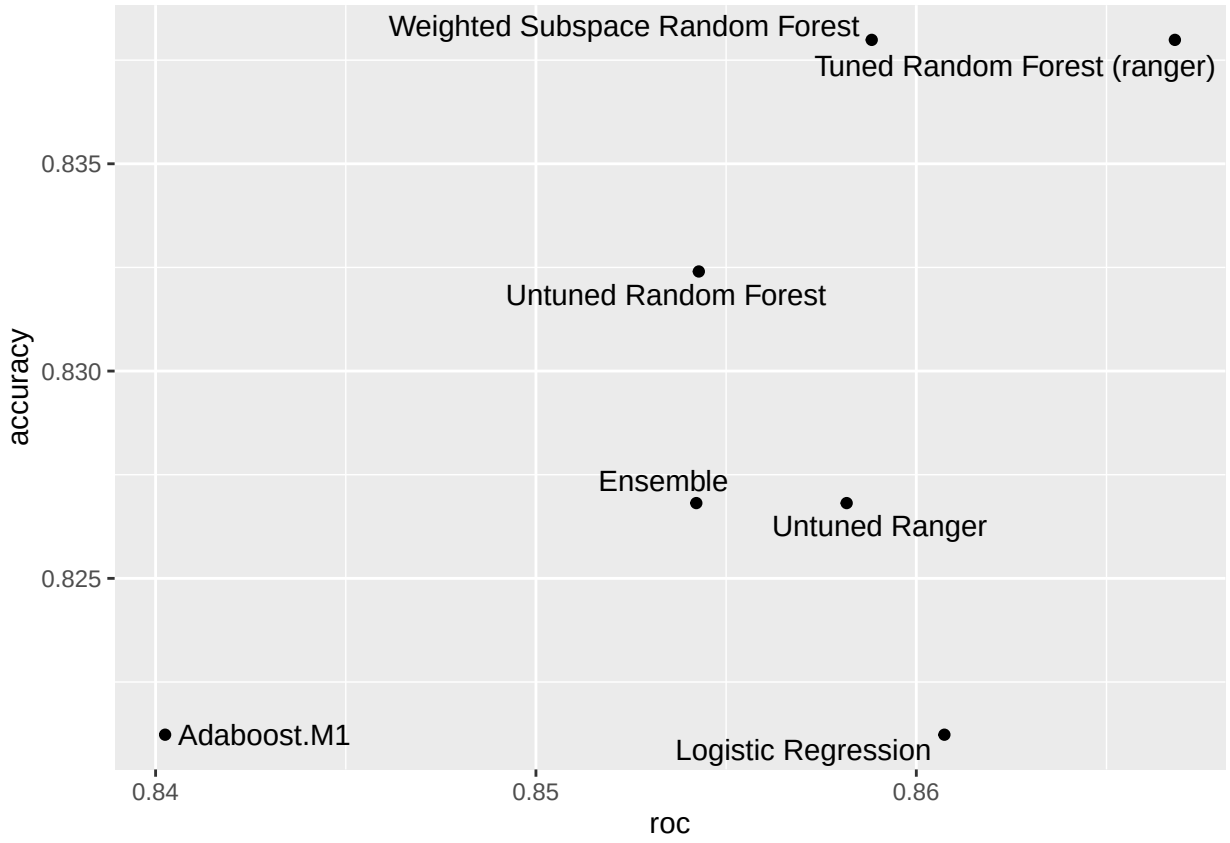
```
## 3  0  0     constant  cabin
```




titleMr
sexmale
fare
titleMrs
titleMiss
age
classThree
cab_letterNA
num_cab
sib_sp
par_ch
embarkedS
titleMaster
classTwo
cab_letterB
cab_letterE
cab_letterD
embarkedQ
cab_letterC
cab_letterF
titleOther
titleDr
titleRev
cab_letterG
titleMajor
titleMlle
Feature
0
25
50
75
100
Importance


1.0
0.9
0.8
0.7
accuracy
0.7
0.8
0.9
1.0
roc
adaboost
ranger
rf
wsrf
monmlp
Rborist
avNNet
gbm
svmLinear
gam
gamLoess
naive_bayes
svmRadialSigma
svmRadialCost
svmRadial
knn
mlp

5 Results




```
## # A tibble: 7 x 3
##   method                            roc accuracy
##   <chr>                           <dbl>    <dbl>
## 1 Tuned Random Forest (ranger)    0.867    0.838
## 2 Logistic Regression             0.861    0.821
## 3 Weighted Subspace Random Forest 0.859    0.838
## 4 Untuned Ranger                  0.858    0.827
## 5 Untuned Random Forest           0.854    0.832
## 6 Ensemble                        0.854    0.827
## 7 Adaboost.M1                     0.840    0.821
```

```
## # A tibble: 6 x 12
##      id class sex      age sib_sp par_ch  fare embarked cab_letter num_cab title
##   <int> <fct> <fct>  <dbl>  <int>  <int> <dbl> <fct>    <fct>        <int> <fct>
## 1   892 Three male    34.5      0      0  7.83 Q        <NA>             0 Mr
## 2   893 Three female  47        1      0  7    S        <NA>             0 Mrs
## 3   894 Two   male    62        0      0  9.69 Q        <NA>             0 Mr
## 4   895 Three male    27        0      0  8.66 S        <NA>             0 Mr
## 5   896 Three female  22        1      1 12.3  S        <NA>             0 Mrs
## 6   897 Three male    14        0      0  9.22 S        <NA>             0 Mr
## # ... with 1 more variable: survived <lgl>
```

Score of 0.75837, which in this case was determined by accuracy. This means that I correctly predicted survival for 317 of the 418 passengers in the test set. 0.75837 is significantly lower than the some of the accuracies reached during model development. This is a clear indication that the methods were over-trained

on the training data set.

6 Discussion

7 Conclusion

8 Acknowledgements

9 References

10 Appendices (if needed)