Titanic

2024-08-21

The data-set

Data Dictionary

Variable
survival: Survival 0 = No, 1 = Yes

pclass: Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd

sex: Sex

Age: Age in years

sibsp: # of siblings / spouses aboard the Titanic

parch: # of parents / children aboard the Titanic

ticket: Ticket number

fare: Passenger fare

cabin :Cabin number

embarked: Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton

pclass: A proxy for socio-economic status (SES)

1st = Upper 2nd = Middle 3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way…

Sibling = brother, sister, stepbrother, stepsister

Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way…

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

Some children travelled only with a nanny, therefore parch=0 for them.

TRAIN DATA

PassengerId <int>	Survived <int>	Pclass <int>	Name <chr>	Sex <chr>
1	0	3	Braund, Mr. Owen Harris	male
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female
3	1	3	Heikkinen, Miss. Laina	female
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female
5	0	3	Allen, Mr. William Henry	male
6	0	3	Moran, Mr. James	male
7	0	1	McCarthy, Mr. Timothy J	male
8	0	3	Palsson, Master. Gosta Leonard	male
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female

1-10 of 891 rows | 1-5 of 12 columns   Previous 1 2 3 4 5 6 ... 90 Next

Dimension of the data-set

```
[1] 891  12
```

Variables in the data-set

```
'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)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
 $ 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" ...
```

Column Names

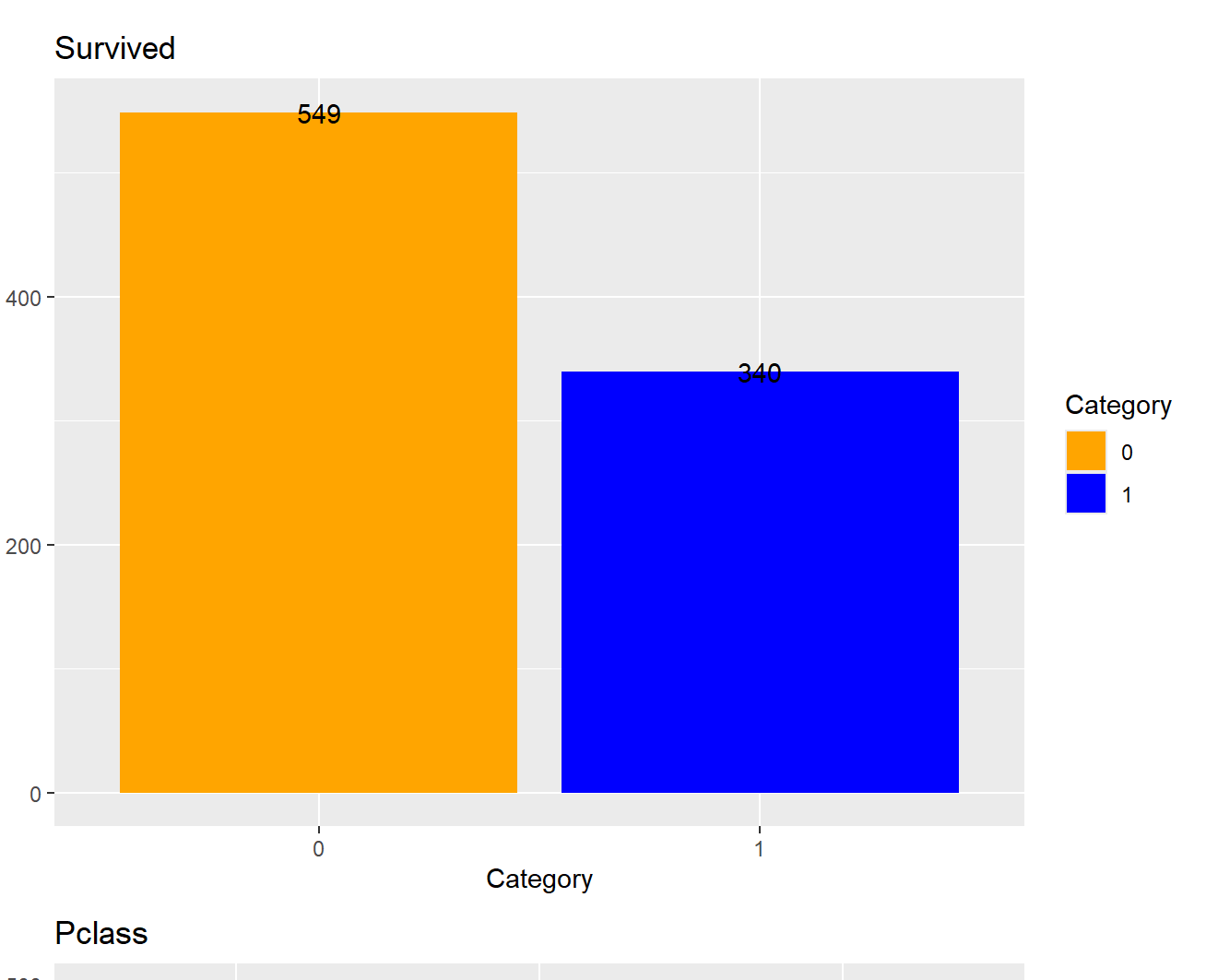
```
 [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
 [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
[11] "Cabin"       "Embarked"
```

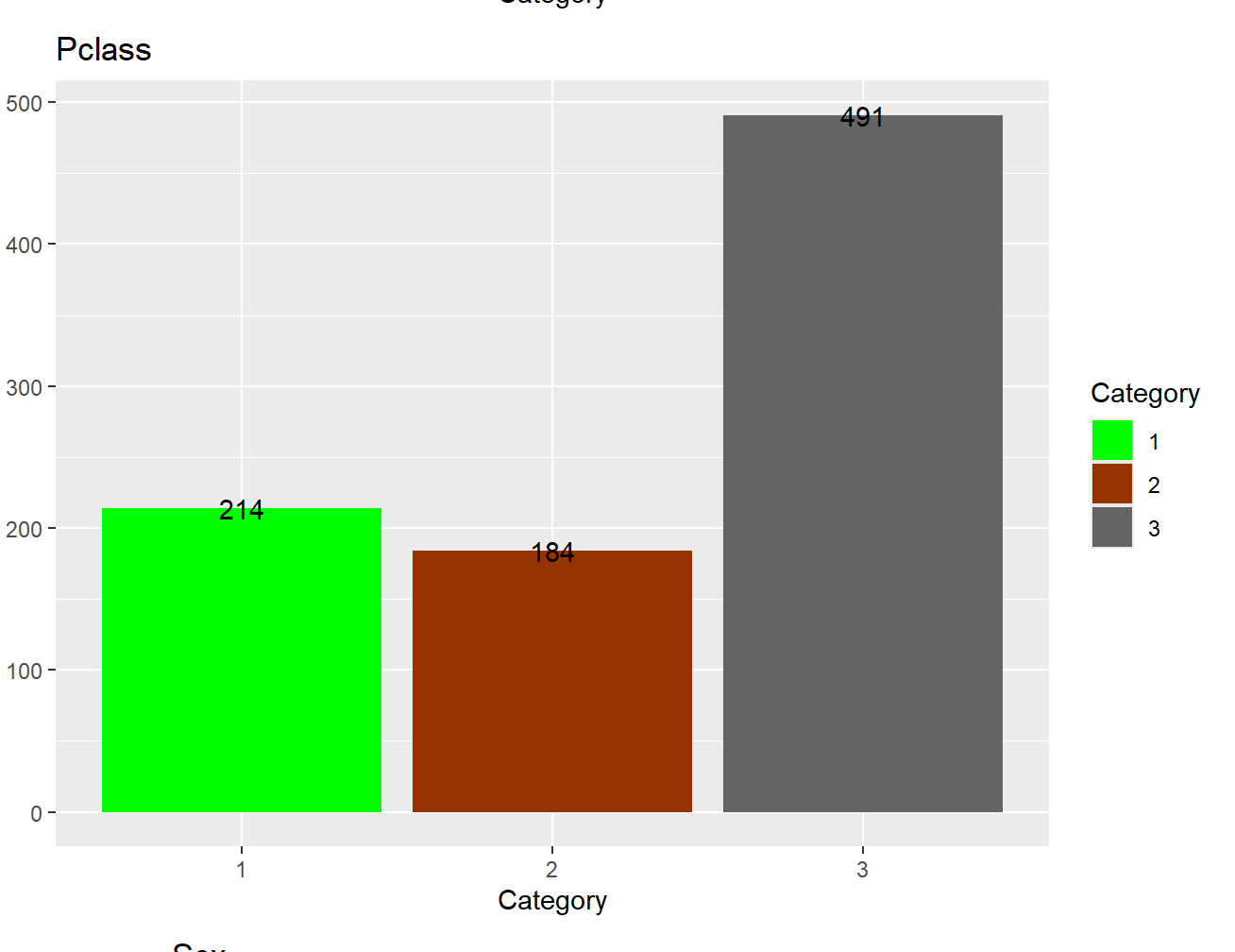
We remove the 1st,4th,9th and 11th column from the data-set

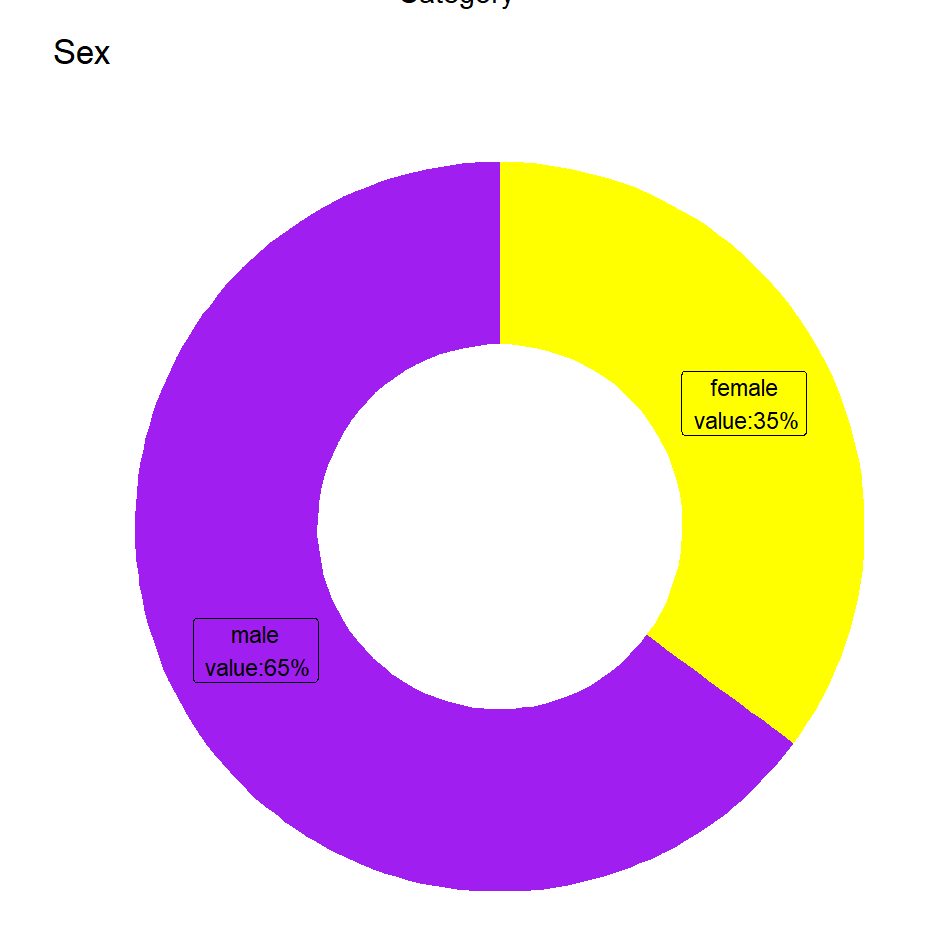
We replace the NA values in the Age column with the mean of that column and convert the character variables into factor variables.

In the dataset, we remove the empty cells from the 'Embark' column

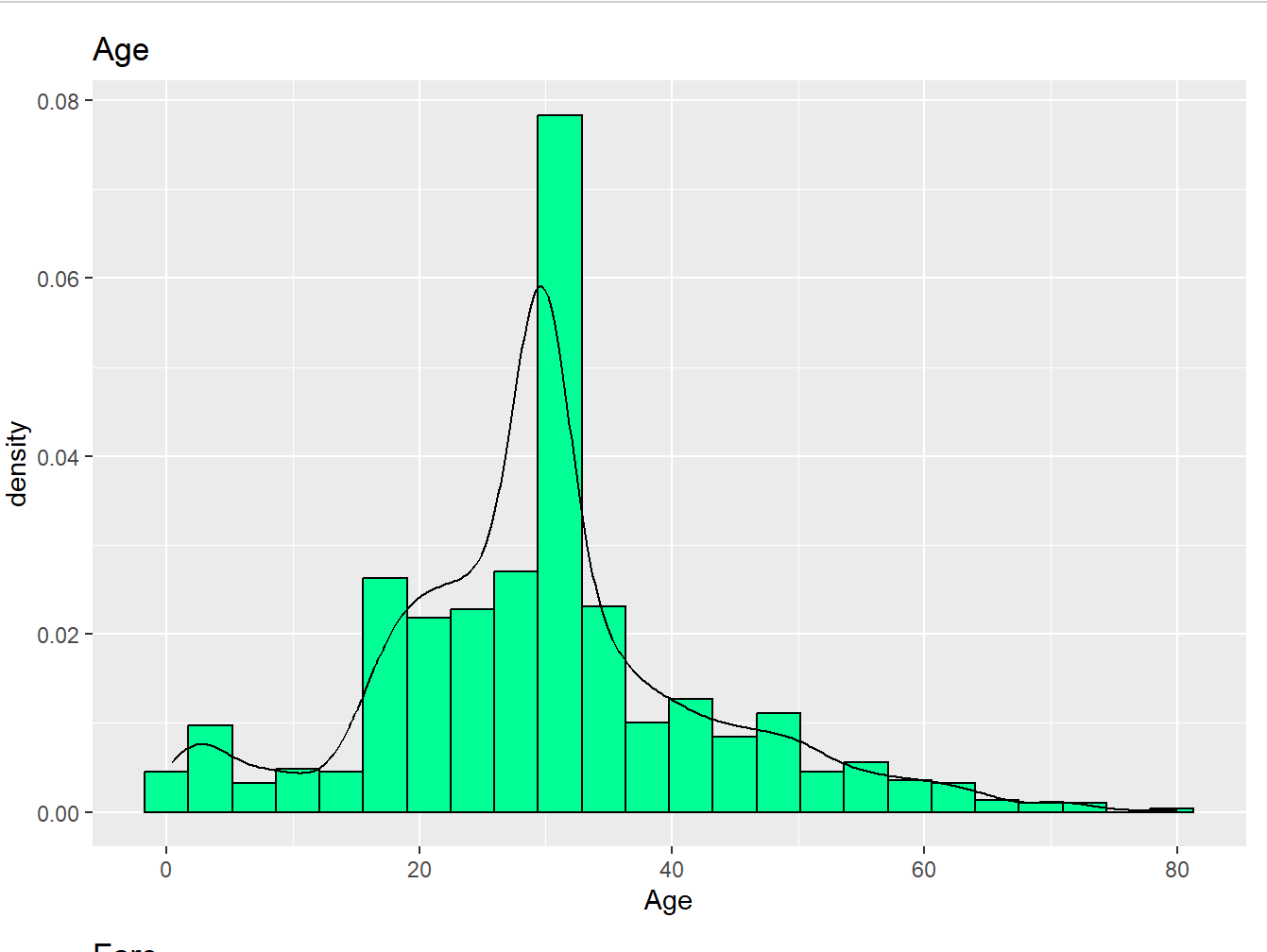
PLOTS

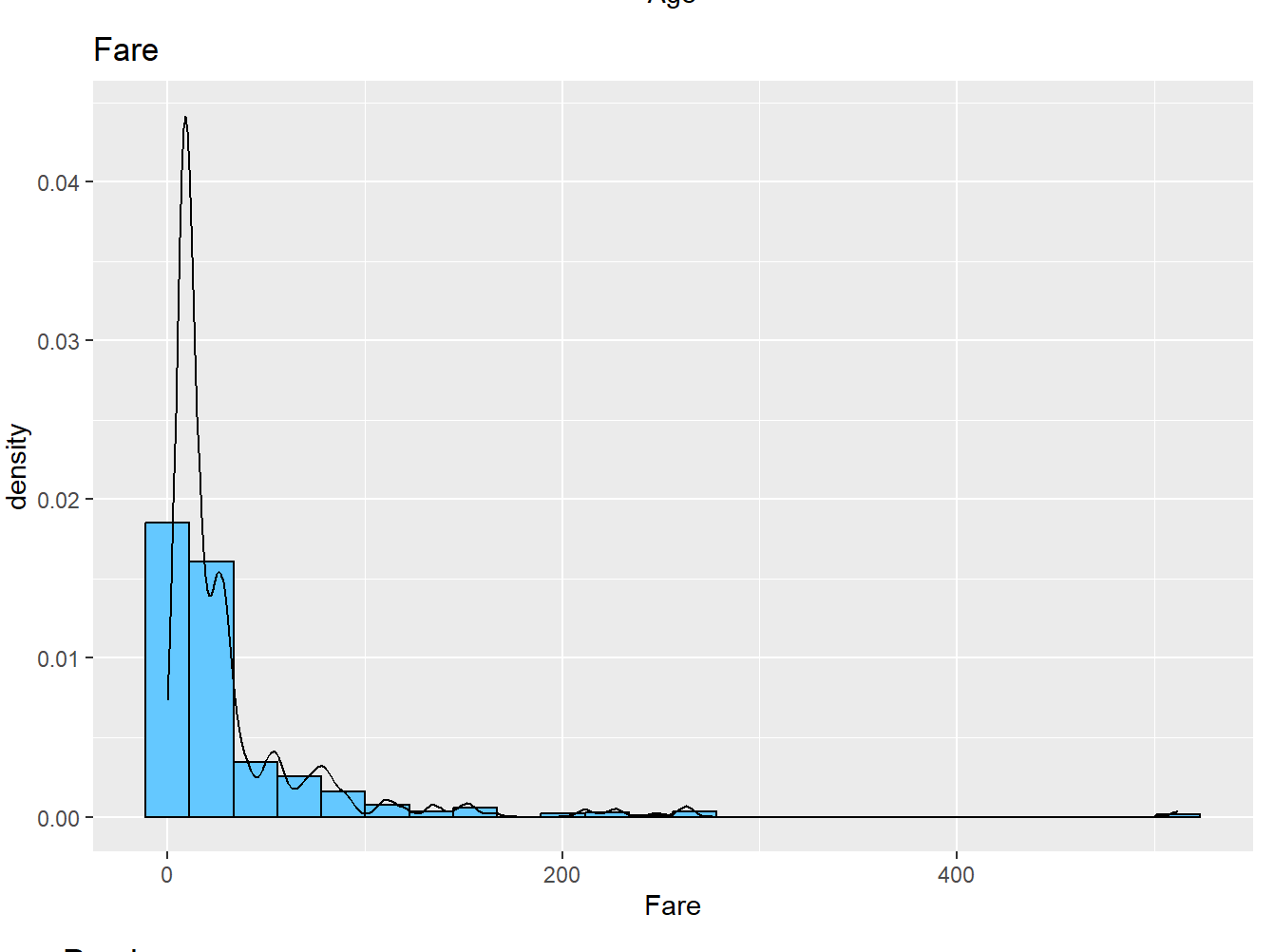


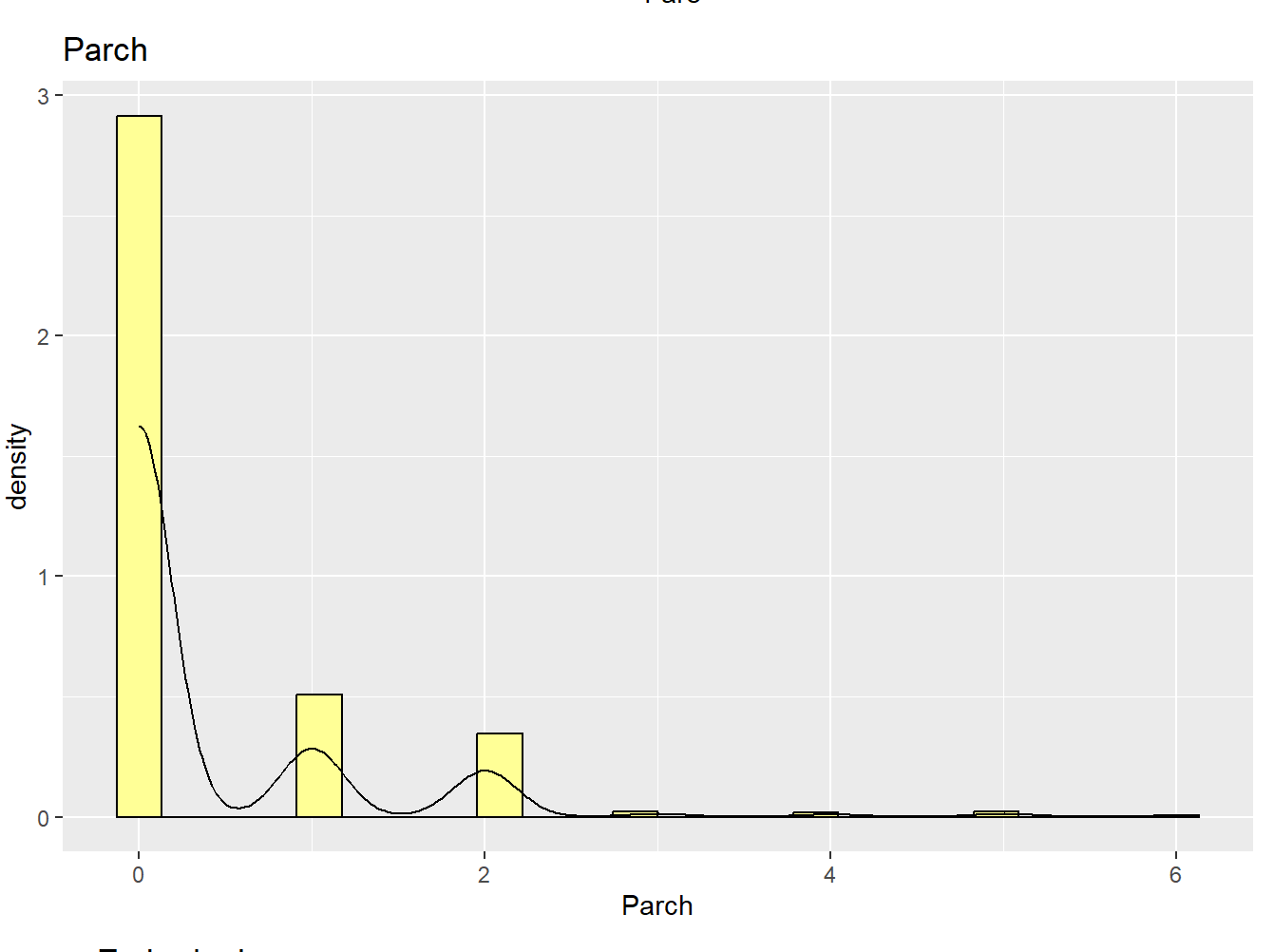


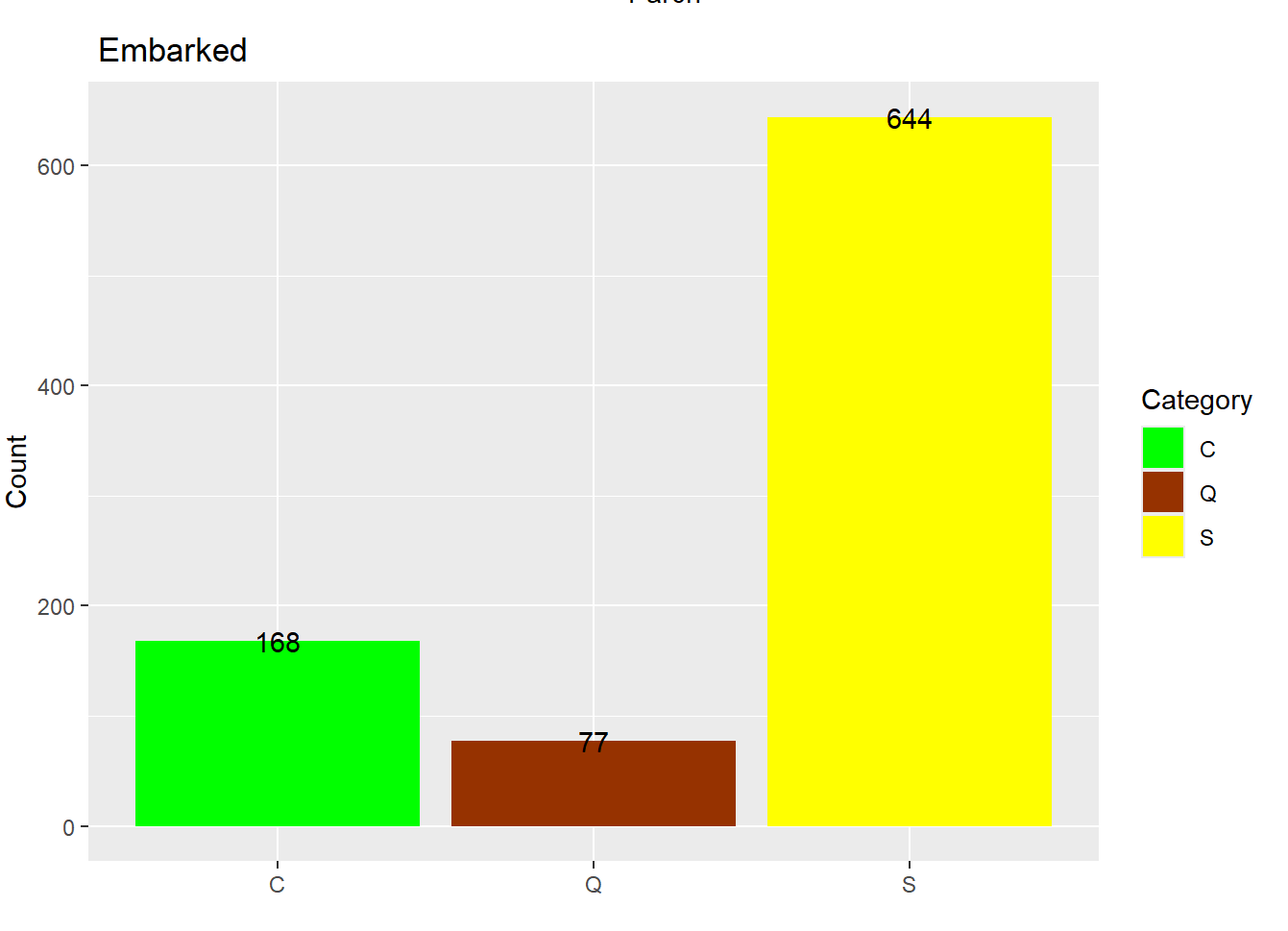



```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```









Since our response variable is binary, taking only the values 0 and 1, we use a logistic regression model.

```

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

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  4.102784   0.476303   8.614  < 2e-16 ***
Pclass2     -0.924047   0.297882  -3.102  0.00192 ** 
Pclass3     -2.149626   0.297749  -7.220 5.21e-13 ***
Sexmale     -2.709611   0.201336 -13.458  < 2e-16 ***
Age         -0.039320   0.007888  -4.984 6.21e-07 ***
SibSp       -0.322143   0.109545  -2.941  0.00327 ** 
Parch       -0.095061   0.119028  -0.799  0.42450    
Fare         0.002261   0.002462   0.918  0.35842    
EmbarkedQ   -0.029839   0.381534  -0.078  0.93766    
EmbarkedS   -0.445754   0.239730  -1.859  0.06297 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.82  on 888  degrees of freedom
Residual deviance:  783.74  on 879  degrees of freedom
AIC: 803.74

Number of Fisher Scoring iterations: 5
```

We can see that 'Pclass2', 'Pclass3', 'SexMale', 'Age', and 'SibSp' are statistically significant, as their p-values are less than 0.05.

To improve the fit of our model we use AIC backward method.

Backward elimination based on the Akaike Information Criterion(AIC)

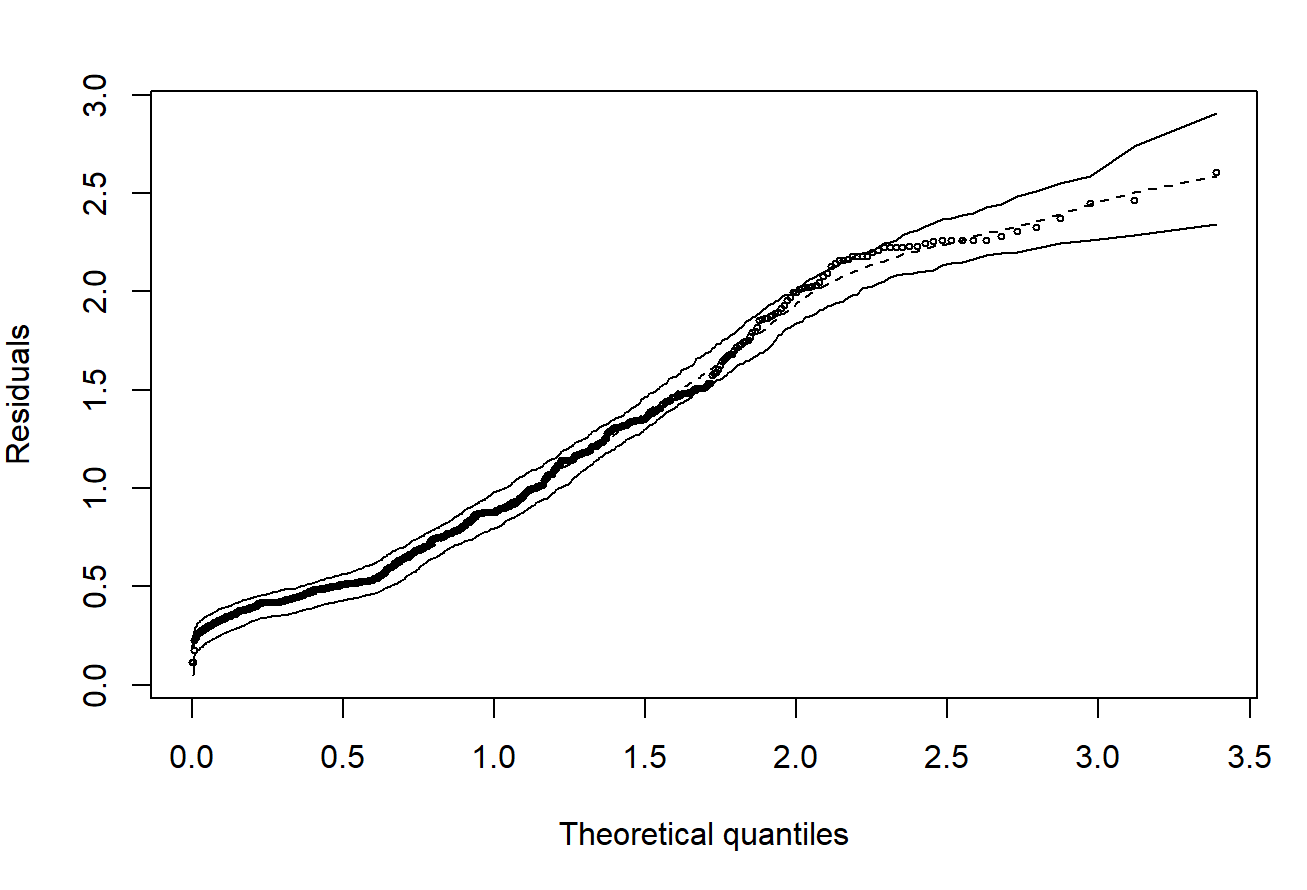
```
Start:  AIC=803.74
Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked

           Df Deviance     AIC
- Parch     1   784.38  802.38
- Fare      1   784.65  802.65
<none>          783.74  803.74
- Embarked  2   788.16  804.16
- SibSp     1   793.85  811.85
- Age       1   810.58  828.58
- Pclass    2   843.57  859.57
- Sex       1  1010.80 1028.80

Step:  AIC=802.38
Survived ~ Pclass + Sex + Age + SibSp + Fare + Embarked

           Df Deviance     AIC
- Fare      1   785.03  801.03
<none>          784.38  802.38
- Embarked  2   789.08  803.08
- SibSp     1   797.02  813.02
- Age       1   811.03  827.03
- Pclass    2   847.53  861.53
- Sex       1  1016.06 1032.06

Step:  AIC=801.03
Survived ~ Pclass + Sex + Age + SibSp + Embarked

           Df Deviance     AIC
<none>          785.03  801.03
- Embarked  2   790.30  802.30
- SibSp     1   797.02  811.02
- Age       1   812.43  826.43
- Pclass    2   882.25  894.25
- Sex       1  1022.23 1036.23
```

We have selected the variables 'Embarked', 'SibSp', 'Age', 'Pclass', and 'Sex' for our regression model.

Half-Normal Probability (hnp) Plot

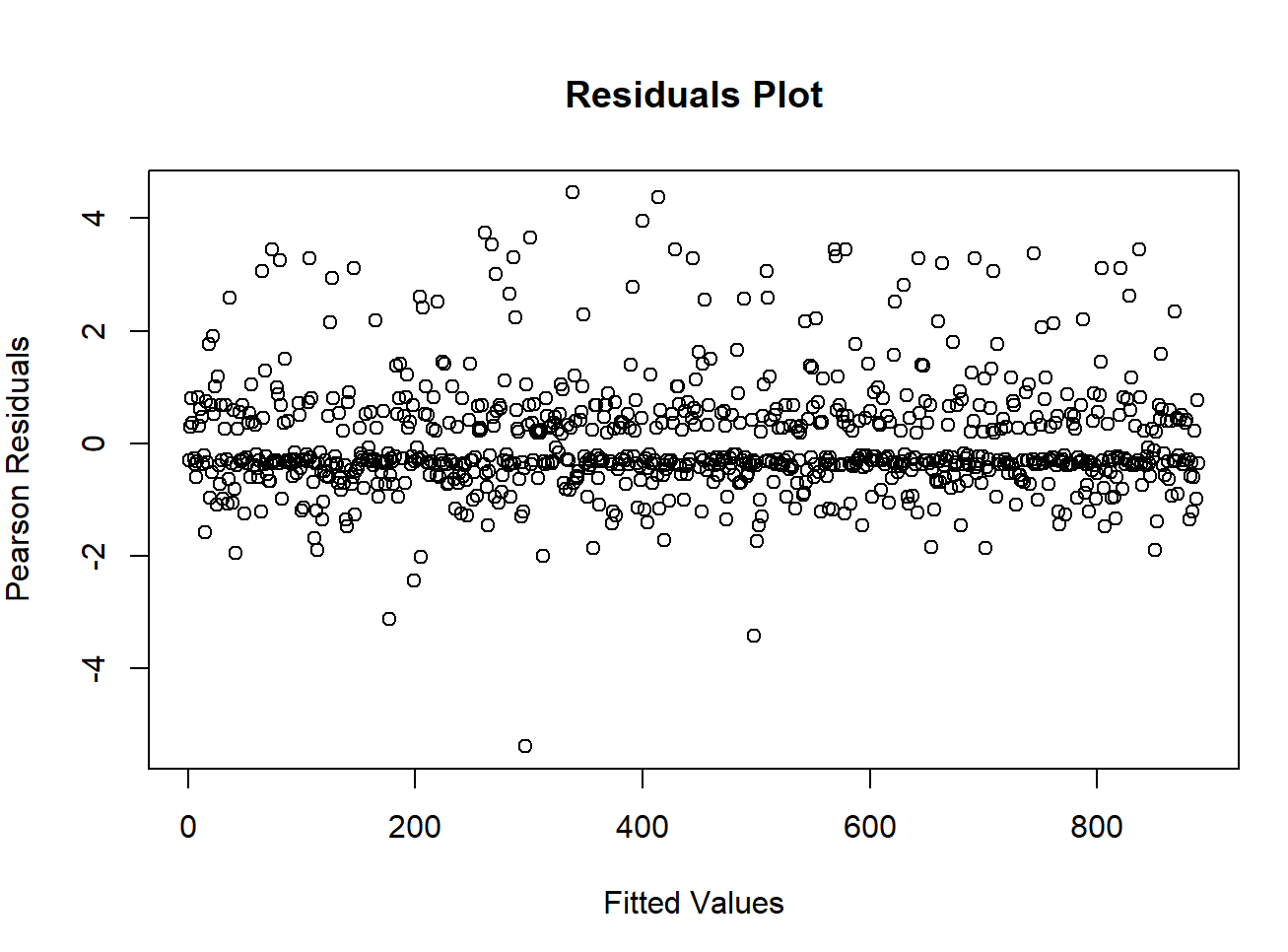
```
Loading required package: MASS
```

```
Binomial model
```



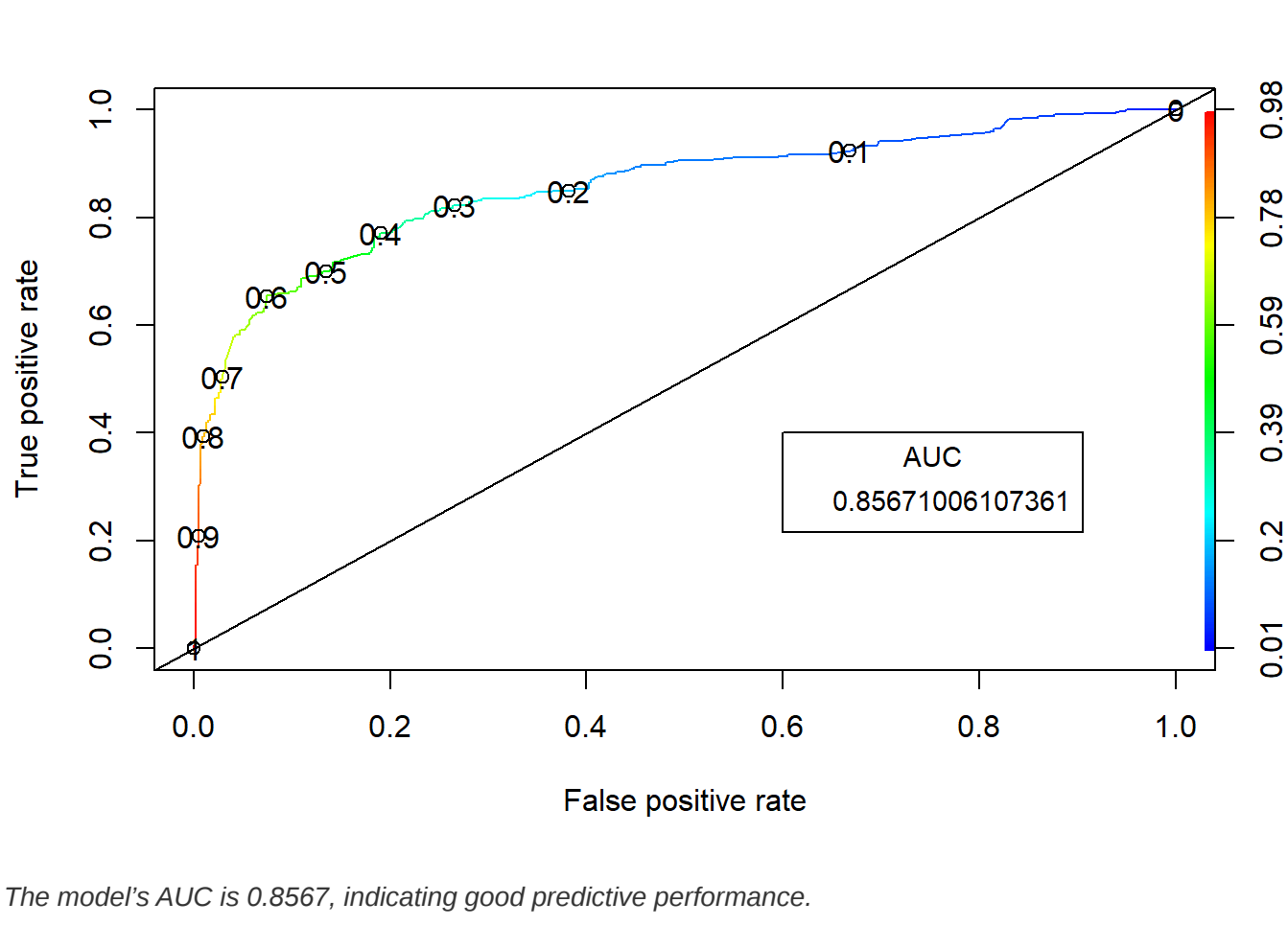
Most of the residuals fall within these confidence bands, suggesting that the model is capturing the majority of the variability in the data accurately.

Residual vs Fitted Plot



Receiver Operator Characteristic (ROC) Curve and Area Under the Curve (AUC)

```
Loading required package: lattice
```



The model's AUC is 0.8567, indicating good predictive performance.

TEST DATA

PassengerId <int>	Pclass <int>	Name <chr>	Sex <chr>	Age <dbl>	Sib… <int>
892	3	Kelly, Mr. James	male	34.50	0
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.00	1
894	2	Myles, Mr. Thomas Francis	male	62.00	0
895	3	Wirz, Mr. Albert	male	27.00	0
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.00	1
897	3	Svensson, Mr. Johan Cervin	male	14.00	0
898	3	Connolly, Miss. Kate	female	30.00	0
899	2	Caldwell, Mr. Albert Francis	male	26.00	1
900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18.00	0
901	3	Davies, Mr. John Samuel	male	21.00	2

1-10 of 418 rows | 1-6 of 11 columns   Previous 1 2 3 4 5 6 ... 42 Next

PREDICTION

	PassengerId <int>	Survived <dbl>
1	892	0
2	893	0
3	894	0
4	895	0
5	896	0
6	897	0
7	898	1
8	899	0
9	900	1
10	901	0

1-10 of 418 rows   Previous 1 2 3 4 5 6 ... 42 Next