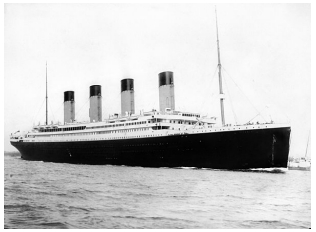
Programming for Data Analytics

Assignment 5
ggplot2

Dr. Jim Duggan,
School of Engineering & Informatics
National University of Ireland Galway.

https://twitter.com/_jimduggan

Aim



https://en.wikipedia.org/wiki/RMS_Titanic

- Prepare and explore the titanic data set
- See https://github.com/JimDuggan/CT5102
- http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets

A	B	C	D	E	F	G
pclass	**survived**	**name**	**sex**	**age**	**fare**	**embarked**
1	1	Allen, Miss. Elisabeth Walton	female	29	211.3375	S
1	1	Allison, Master. Hudson Trevor	male	0.917	151.5500	S
1	0	Allison, Miss. Helen Loraine	female	2	151.5500	S
1	0	Allison, Mr. Hudson Joshua Creighton	male	30	151.5500	S
1	0	Allison, Mrs. Hudson J C (Bessie Waldo	female	25	151.5500	S
1	1	Anderson, Mr. Harry	male	48	26.5500	S
1	1	Andrews, Miss. Kornelia Theodosia	female	63	77.9583	S
1	0	Andrews, Mr. Thomas Jr	male	39	0.0000	S
1	1	Appleton, Mrs. Edward Dale (Charlotte L	female	53	51.4792	S
1	0	Artagaveytia, Mr. Ramon	male	71	49.5042	C
1	0	Astor, Col. John Jacob	male	47	227.5250	C
1	1	Astor, Mrs. John Jacob (Madeleine Talma	female	18	227.5250	C
1	1	Aubart, Mme. Leontine Pauline	female	24	69.3000	C
1	1	Barber, Miss. Ellen "Nellie"	female	26	78.8500	S
1	1	Barkworth, Mr. Algernon Henry Wilson	male	80	30.0000	S
1	0	Baumann, Mr. John D	male		25.9250	S

Reading & Preparing the data

```
1  library(readxl)
2  library(ggplot2)
3
4  # http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets
5
6  orig_list <- data.frame(readxl::read_excel("datasets/Titanic/titanic3_assignment.xls"))
7  plist     <- orig_list
```

```
> dim(plist)
[1] 1309    7
>
> summary(plist)
     pclass         survived         name                sex                 age                fare
 Min.   :1.000   Min.   :0.000   Length:1309        Length:1309        Min.   : 0.1667   Min.   :  0.000
 1st Qu.:2.000   1st Qu.:0.000   Class :character   Class :character   1st Qu.:21.0000   1st Qu.:  7.896
 Median :3.000   Median :0.000   Mode  :character   Mode  :character   Median :28.0000   Median : 14.454
 Mean   :2.295   Mean   :0.382                                         Mean   :29.8811   Mean   : 33.295
 3rd Qu.:3.000   3rd Qu.:1.000                                         3rd Qu.:39.0000   3rd Qu.: 31.275
 Max.   :3.000   Max.   :1.000                                         Max.   :80.0000   Max.   :512.329
                                                                       NA's   :263       NA's   :1

   embarked
 Length:1309
 Class :character
 Mode  :character
```

(1) Convert survived to logical value

```
> summary(plist)
     pclass          survived          name               sex                 age               fare        
 Min.   :1.000   Mode :logical   Length:1309        Length:1309        Min.   : 0.1667   Min.   :  0.000  
 1st Qu.:2.000   FALSE:809       Class :character   Class :character   1st Qu.:21.0000   1st Qu.:  7.896  
 Median :3.000   TRUE :500       Mode  :character   Mode  :character   Median :28.0000   Median : 14.454  
 Mean   :2.295                                                         Mean   :29.8811   Mean   : 33.295  
 3rd Qu.:3.000                                                         3rd Qu.:39.0000   3rd Qu.: 31.275  
 Max.   :3.000                                                         Max.   :80.0000   Max.   :512.329  
                                                                       NA's   :263       NA's   :1        

   embarked        
 Length:1309        
 Class :character   
 Mode  :character   
```

(2) Change class to string

```
> summary(plist)
    pclass          survived          name              sex                 age               fare
Length:1309        Mode :logical   Length:1309        Length:1309        Min.   : 0.1667   Min.   :  0.000
Class :character   FALSE:809       Class :character   Class :character   1st Qu.:21.0000   1st Qu.:  7.896
Mode  :character   TRUE :500       Mode  :character   Mode  :character   Median :28.0000   Median : 14.454
                                                                         Mean   :29.8811   Mean   : 33.295
                                                                         3rd Qu.:39.0000   3rd Qu.: 31.275
                                                                         Max.   :80.0000   Max.   :512.329
                                                                         NA's   :263       NA's   :1

   embarked
Length:1309
Class :character
Mode  :character


>
> unique(plist$pclass)
[1] "First"  "Second" "Third"
```

(3) Simple imputation of age (mean of all ages)

```
> summary(plist)
    pclass          survived          name              sex                 age              fare
 Length:1309        Mode :logical   Length:1309        Length:1309        Min.   : 0.1667   Min.   :  0.000
 Class :character   FALSE:809       Class :character   Class :character   1st Qu.:22.0000   1st Qu.:  7.896
 Mode  :character   TRUE :500       Mode  :character   Mode  :character   Median :29.8811   Median : 14.454
                                                                          Mean   :29.8811   Mean   : 33.295
                                                                          3rd Qu.:35.0000   3rd Qu.: 31.275
                                                                          Max.   :80.0000   Max.   :512.329
                                                                                            NA's   :1

  embarked
 Length:1309
 Class :character
 Mode  :character
```

(4) Simple imputation of fare (mean of all fares)

```
> summary(plist)
    pclass          survived          name              sex                 age              fare        
 Length:1309        Mode :logical   Length:1309        Length:1309        Min.   : 0.1667   Min.   :  0.000  
 Class :character   FALSE:809       Class :character   Class :character   1st Qu.:22.0000   1st Qu.:  7.896  
 Mode  :character   TRUE :500       Mode  :character   Mode  :character   Median :29.8811   Median : 14.454  
                                                                          Mean   :29.8811   Mean   : 33.295  
                                                                          3rd Qu.:35.0000   3rd Qu.: 31.275  
                                                                          Max.   :80.0000   Max.   :512.329  

   embarked        
 Length:1309        
 Class :character   
 Mode  :character   
```

(5) Simple imputation of place of embarking (randomly generated) with seed of 99

```
> summary(plist)
    pclass          survived          name               sex                 age               fare
 Length:1309        Mode :logical   Length:1309        Length:1309        Min.    : 0.1667   Min.    :  0.000
 Class :character   FALSE:809       Class :character   Class :character   1st Qu.:22.0000   1st Qu.:  7.896
 Mode  :character   TRUE :500       Mode  :character   Mode  :character   Median :29.8811   Median : 14.454
                                                                          Mean   :29.8811   Mean   : 33.295
                                                                          3rd Qu.:35.0000   3rd Qu.: 31.275
                                                                          Max.   :80.0000   Max.   :512.329

  embarked
 Length:1309
 Class :character
 Mode  :character

>
> unique(plist$embarked)
[1] "S" "C" "Q"
```

(6) Create new category (age cohort)

- Child (<16), Adults (>=16 & <60) and Elderly (>=60)

```
> summary(plist)
    pclass          survived          name              sex                 age               fare
 Length:1309        Mode :logical   Length:1309        Length:1309        Min.   : 0.1667   Min.   :  0.000
 Class :character   FALSE:809       Class :character   Class :character   1st Qu.:22.0000   1st Qu.:  7.896
 Mode  :character   TRUE :500       Mode  :character   Mode  :character   Median :29.8811   Median : 14.454
                                                                          Mean   :29.8811   Mean   : 33.295
                                                                          3rd Qu.:35.0000   3rd Qu.: 31.275
                                                                          Max.   :80.0000   Max.   :512.329

   embarked          age_cohort
 Length:1309        Length:1309
 Class :character   Class :character
 Mode  :character   Mode  :character
```

(7) Put in full town origin (Queenstown (Q) replaced by Cobh)

```
> summary(plist)
    pclass          survived          name              sex                 age              fare
 Length:1309        Mode :logical   Length:1309        Length:1309        Min.   : 0.1667   Min.   :  0.000
 Class :character   FALSE:809       Class :character   Class :character   1st Qu.:22.0000   1st Qu.:  7.896
 Mode  :character   TRUE :500       Mode  :character   Mode  :character   Median :29.8811   Median : 14.454
                                                                          Mean   :29.8811   Mean   : 33.295
                                                                          3rd Qu.:35.0000   3rd Qu.: 31.275
                                                                          Max.   :80.0000   Max.   :512.329

  embarked          age_cohort
 Length:1309        Length:1309
 Class :character   Class :character
 Mode  :character   Mode  :character
```

```
> unique(plist$embarked)
[1] "Southampton" "Cherbourg"   "Cobh"
```

(8) Double check dataset

```
> head(plist)
  pclass survived                                            name    sex     age     fare    embarked age_cohort
1  First     TRUE                      Allen, Miss. Elisabeth Walton female 29.0000 211.3375 Southampton      Adult
2  First     TRUE                     Allison, Master. Hudson Trevor   male  0.9167 151.5500 Southampton      Child
3  First    FALSE                       Allison, Miss. Helen Loraine female  2.0000 151.5500 Southampton      Child
4  First    FALSE             Allison, Mr. Hudson Joshua Creighton   male 30.0000 151.5500 Southampton      Adult
5  First    FALSE Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female 25.0000 151.5500 Southampton      Adult
6  First     TRUE                                  Anderson, Mr. Harry   male 48.0000  26.5500 Southampton      Adult
```

```
> dim(plist)
[1] 1309    8
>
> table(plist$survived)

FALSE  TRUE
  809   500
> table(plist$survived,plist$age_cohort)

       Adult Child Elderly
  FALSE   732    49      28
  TRUE    422    66      12
```

```
> table(plist$survived,plist$sex)

       female male
  FALSE    127  682
  TRUE     339  161
> table(plist$survived,plist$pclass)

       First Second Third
  FALSE   123    158   528
  TRUE    200    119   181
> table(plist$survived,plist$embarked)

       Cherbourg Cobh Southampton
  FALSE       120   79         610
  TRUE        151   44         305
```

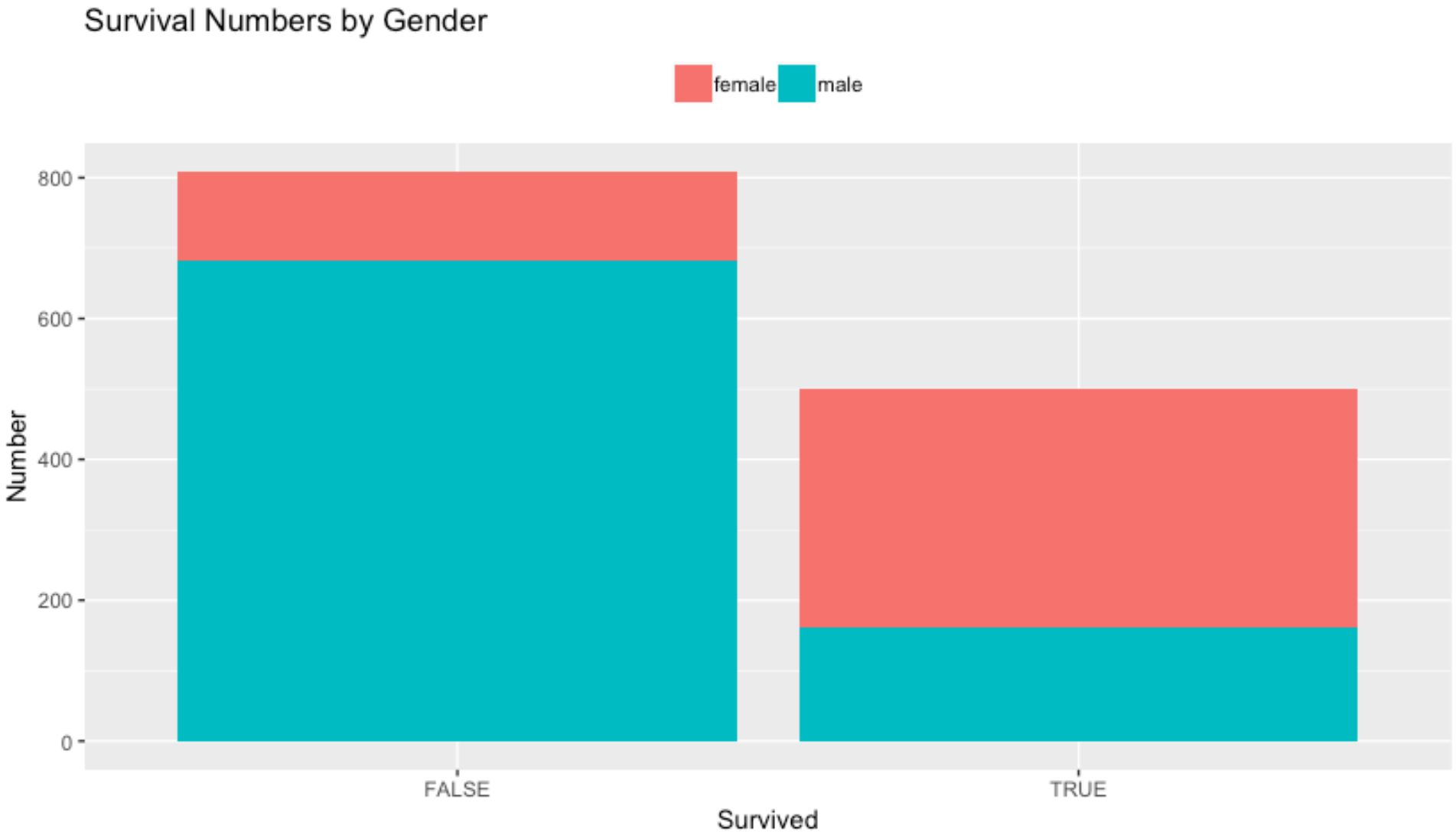
Generate Plots

- Generate the following plots.
- They must be an exact replica of what is shown.
- Some features will require some research, for example, how to hide a legend name.

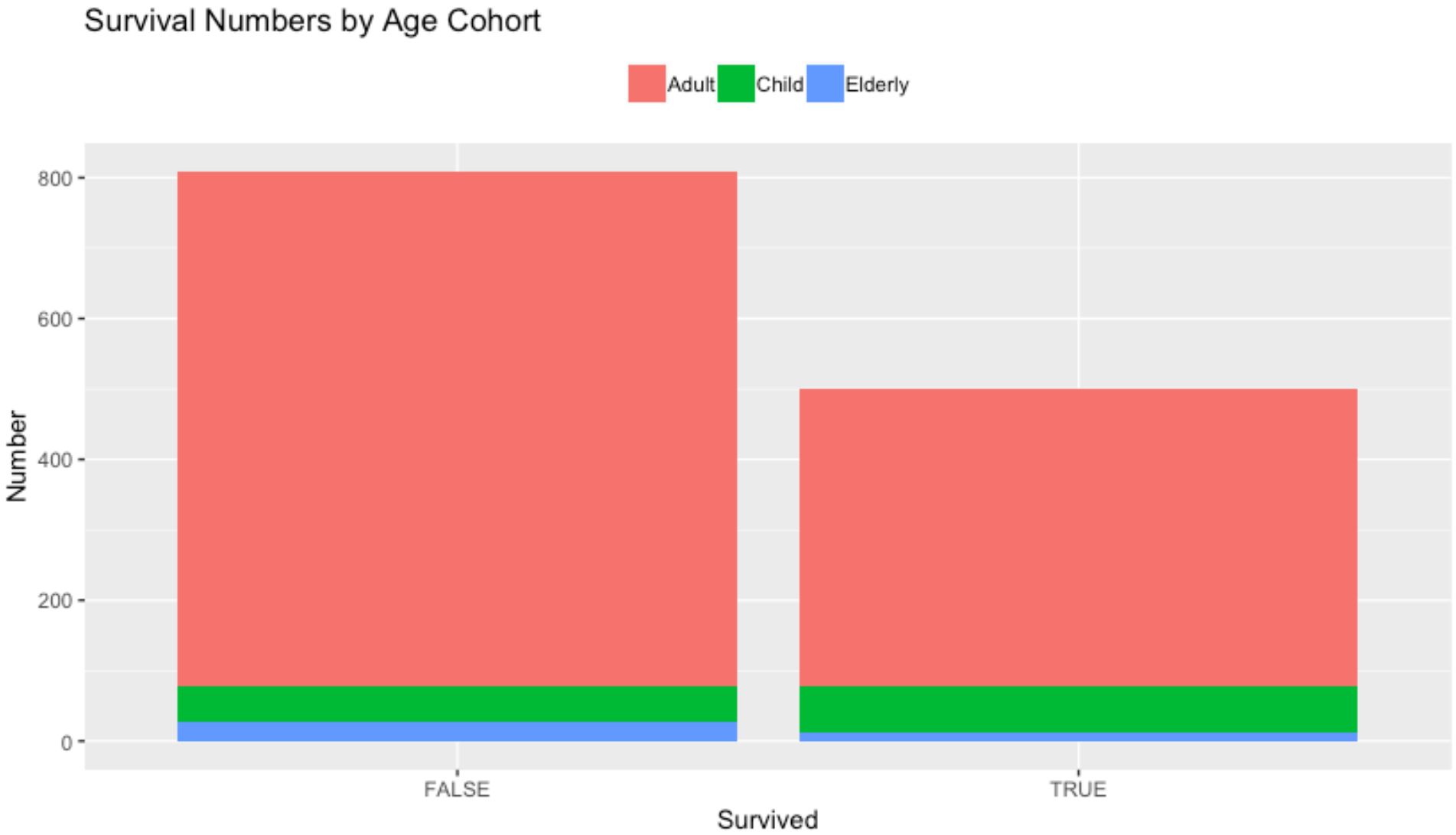
Plot 1


Survival Numbers by Travel Class
First
Second
Third
800
600
400
200
0
Number
FALSE
TRUE
Survived

Plot 2


Survival Numbers by Gender
female
male
800
600
400
200
0
Number
FALSE
TRUE
Survived

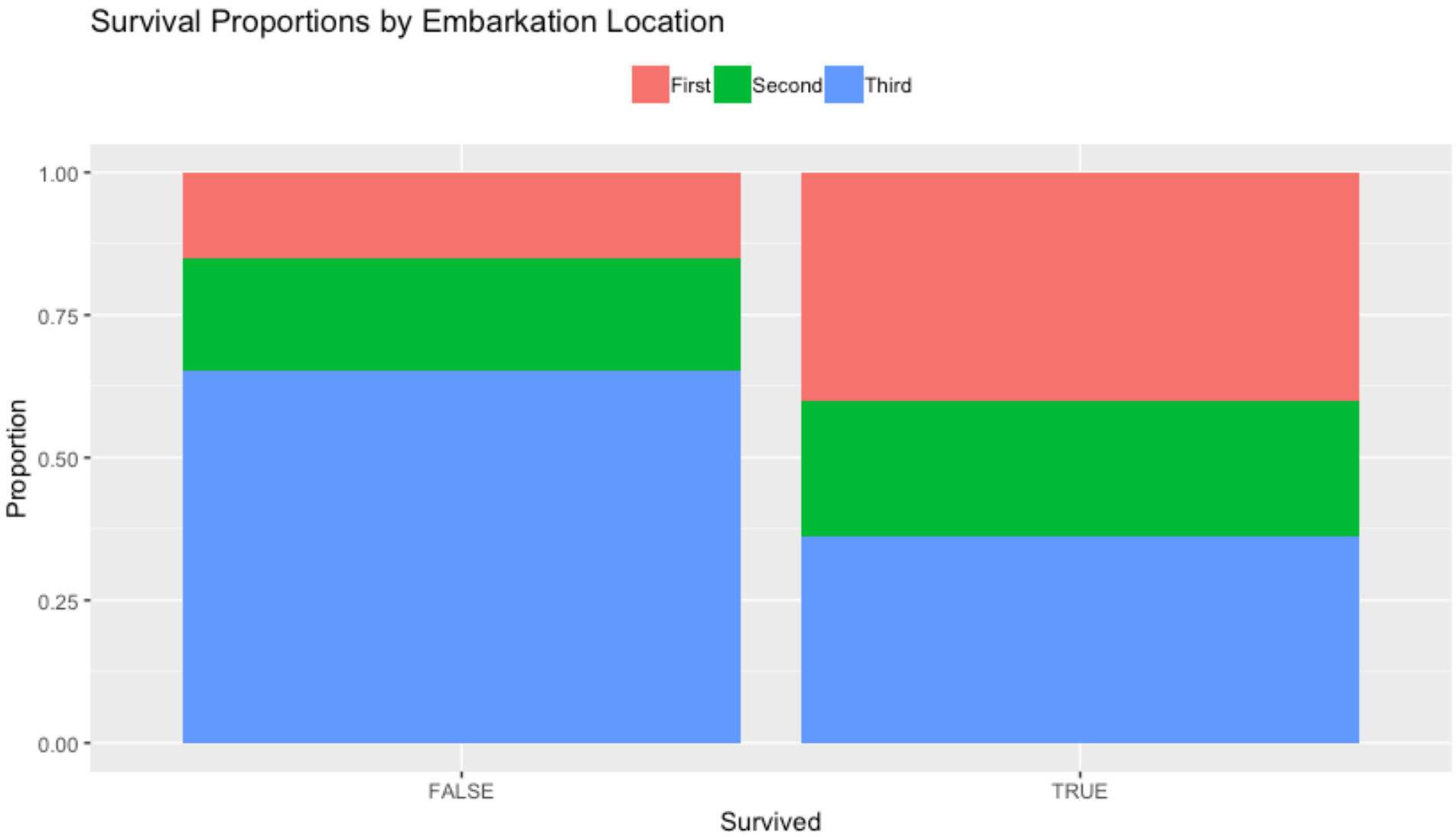
Plot 3


Survival Numbers by Age Cohort
Adult
Child
Elderly
800
600
400
200
0
Number
FALSE
TRUE
Survived

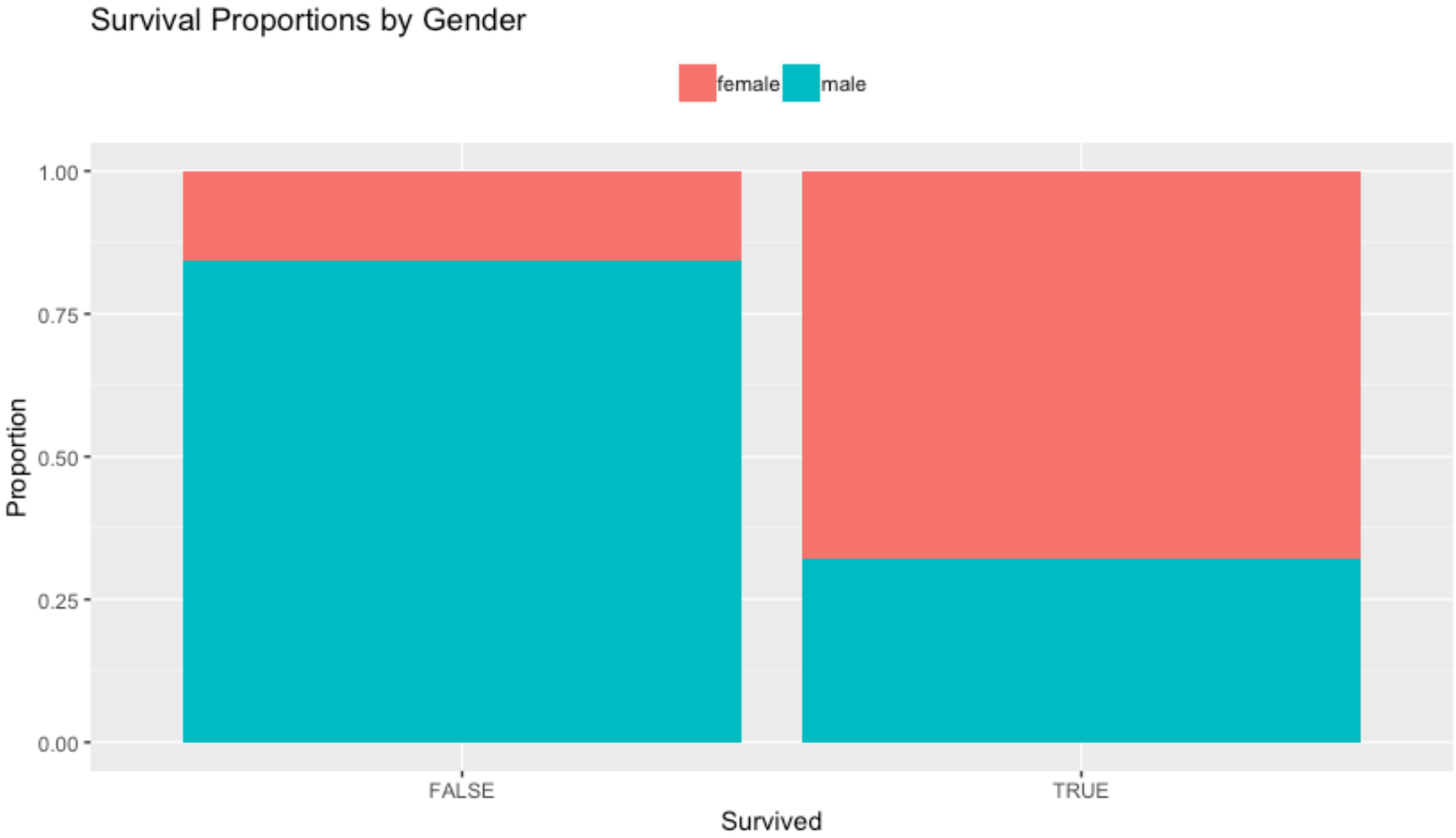
Plot 4


Survival Numbers by Embarkation Location
Cherbourg
Cobh
Southampton
800
600
400
200
0
Number
FALSE
TRUE
Survived

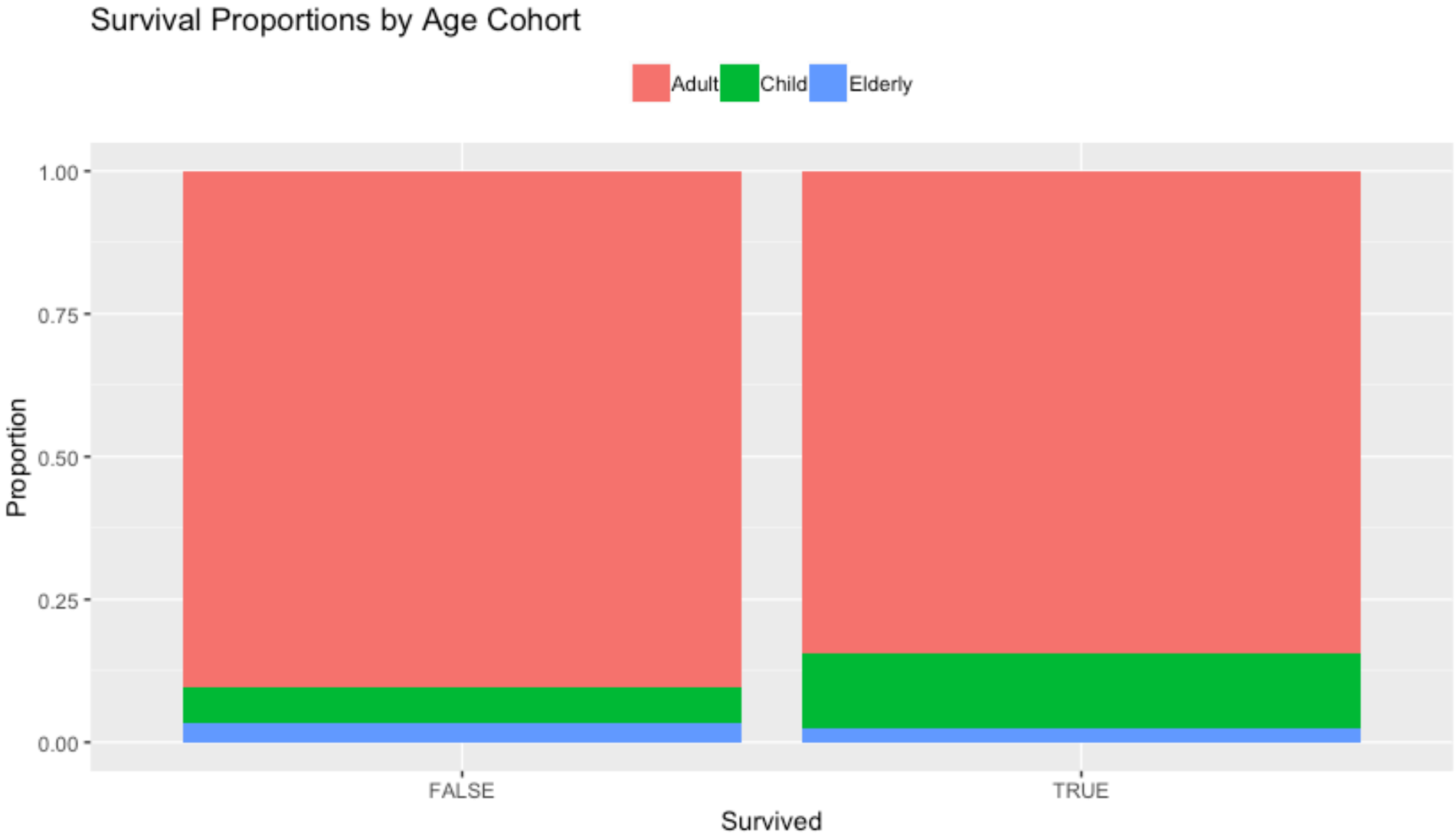
Plot 5


Survival Proportions by Embarkation Location
First
Second
Third
1.00
0.75
0.50
0.25
0.00
Proportion
FALSE
TRUE
Survived

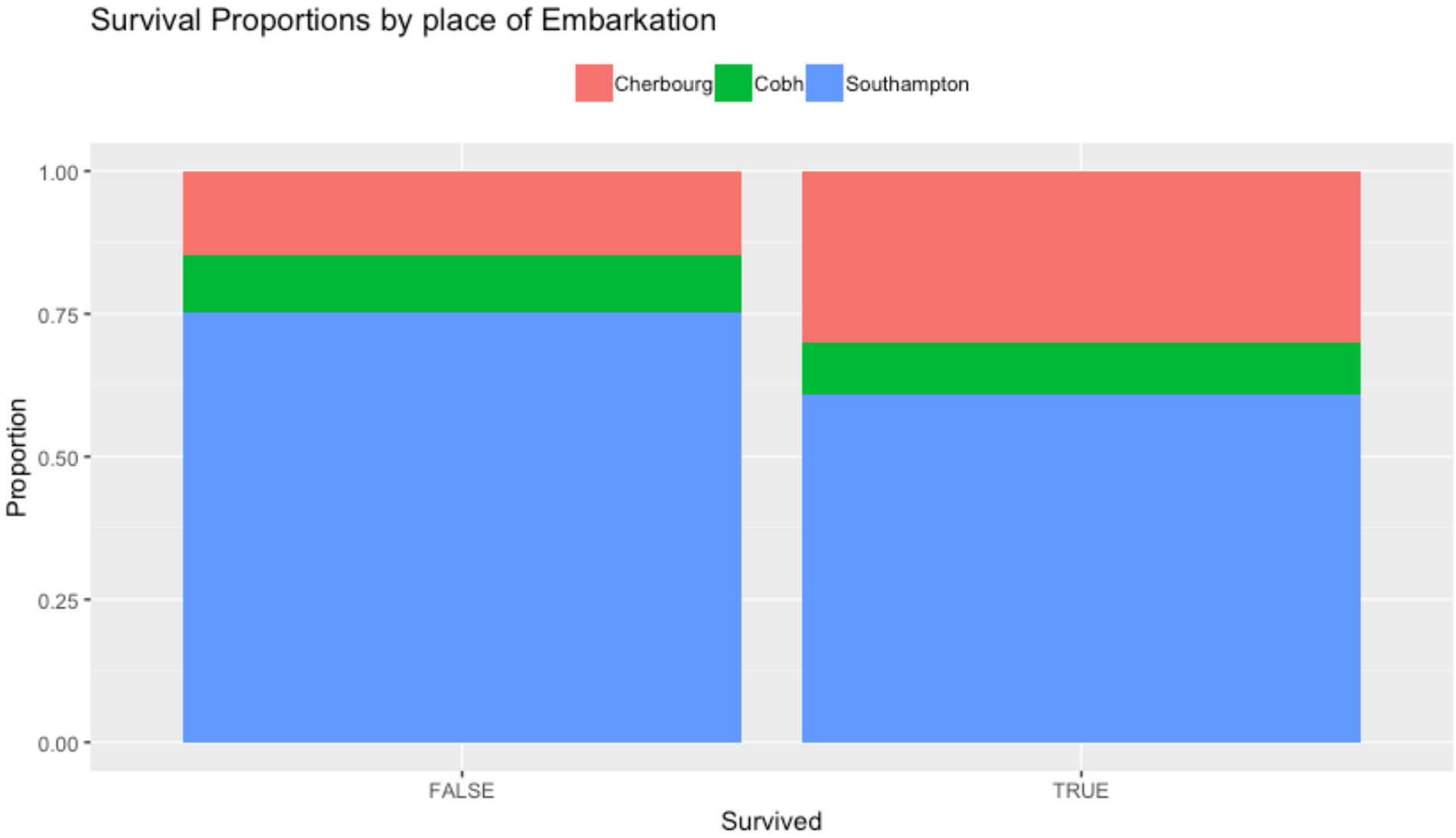
Plot 6


Survival Proportions by Gender
female
male
1.00
0.75
0.50
0.25
0.00
Proportion
FALSE
TRUE
Survived

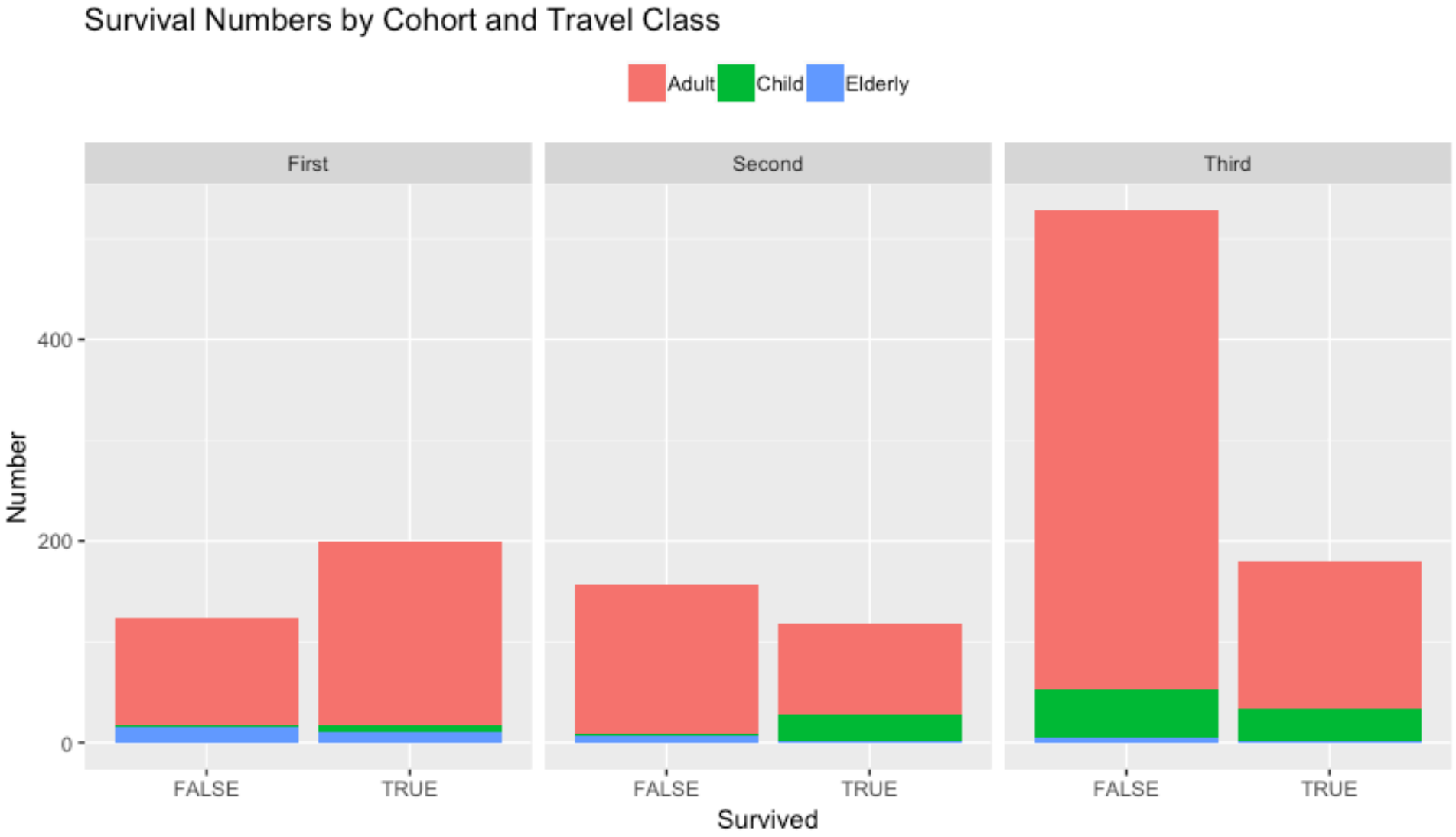
Plot 7


Survival Proportions by Age Cohort
Adult
Child
Elderly
1.00
0.75
0.50
0.25
0.00
Proportion
FALSE
TRUE
Survived

Plot 8


Survival Proportions by place of Embarkation
Cherbourg
Cobh
Southampton
1.00
0.75
0.50
0.25
0.00
Proportion
FALSE
TRUE
Survived

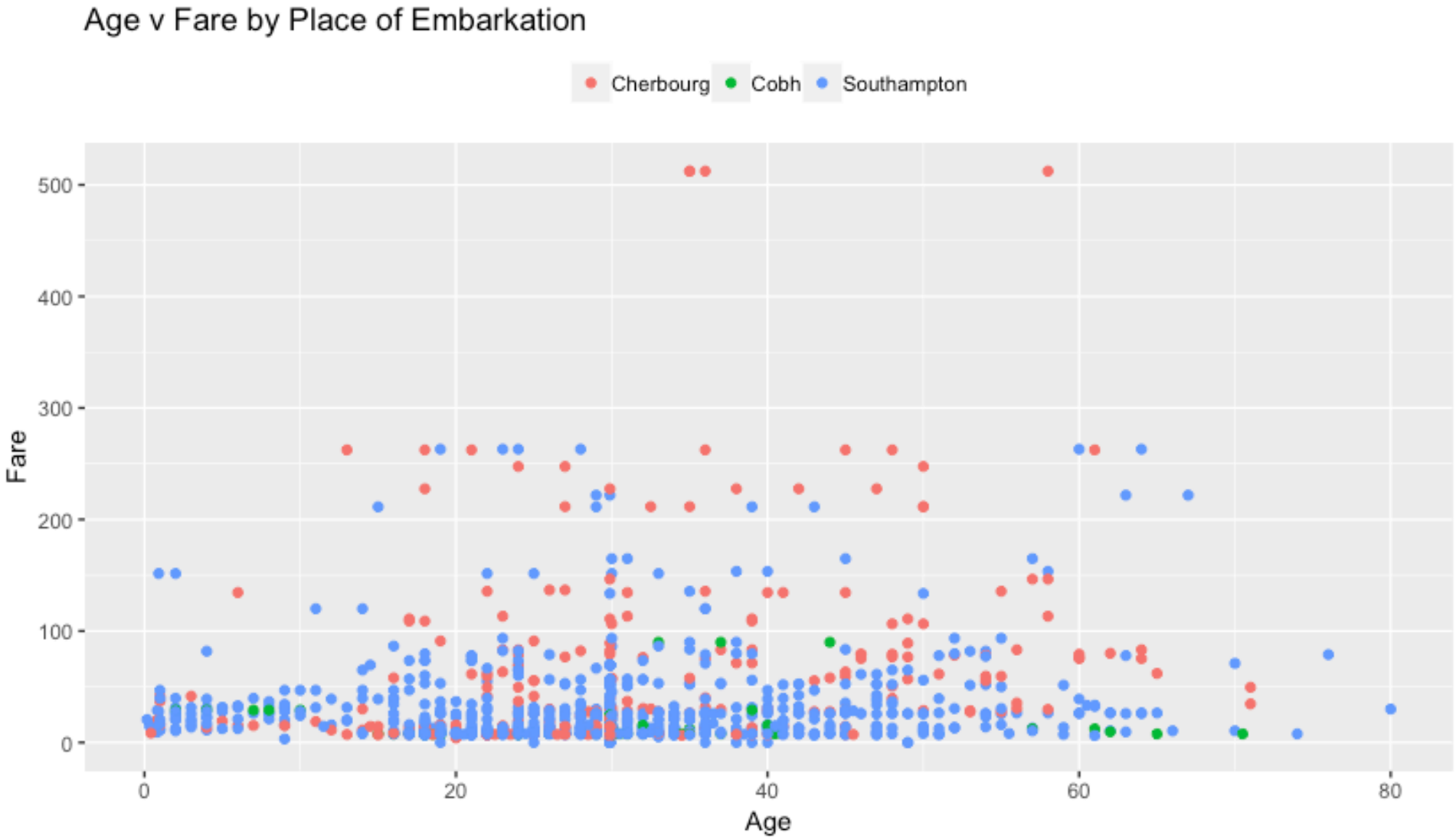
Plot 9


Survival Numbers by Cohort and Travel Class
Adult
Child
Elderly
First
Second
Third
400
200
0
Number
FALSE
TRUE
FALSE
TRUE
FALSE
TRUE
Survived

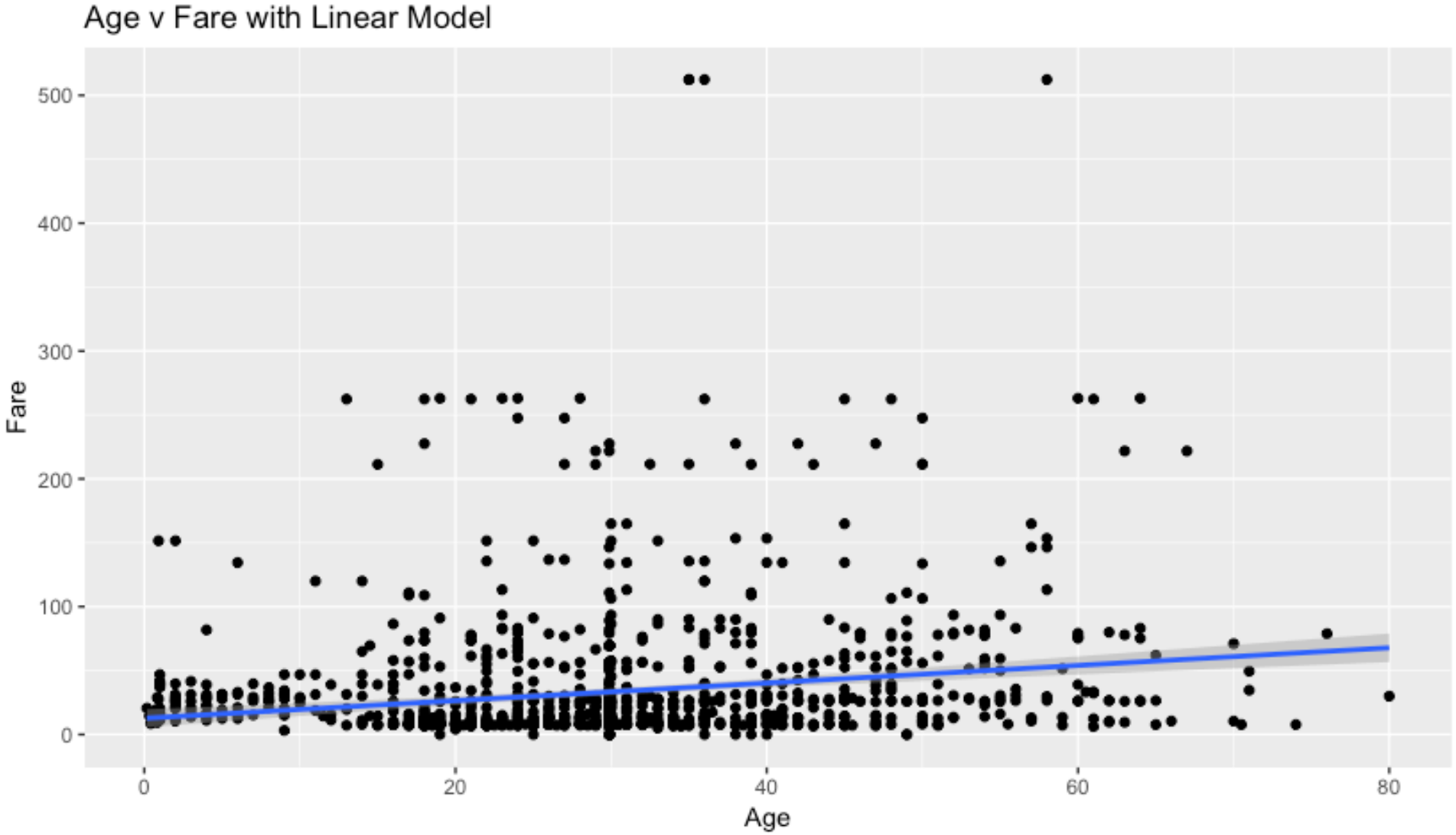
Plot 10


Survival Numbers by Gender and Travel Class
female
male
First
Second
Third
400
200
0
Number
FALSE
TRUE
FALSE
TRUE
FALSE
TRUE
Survived

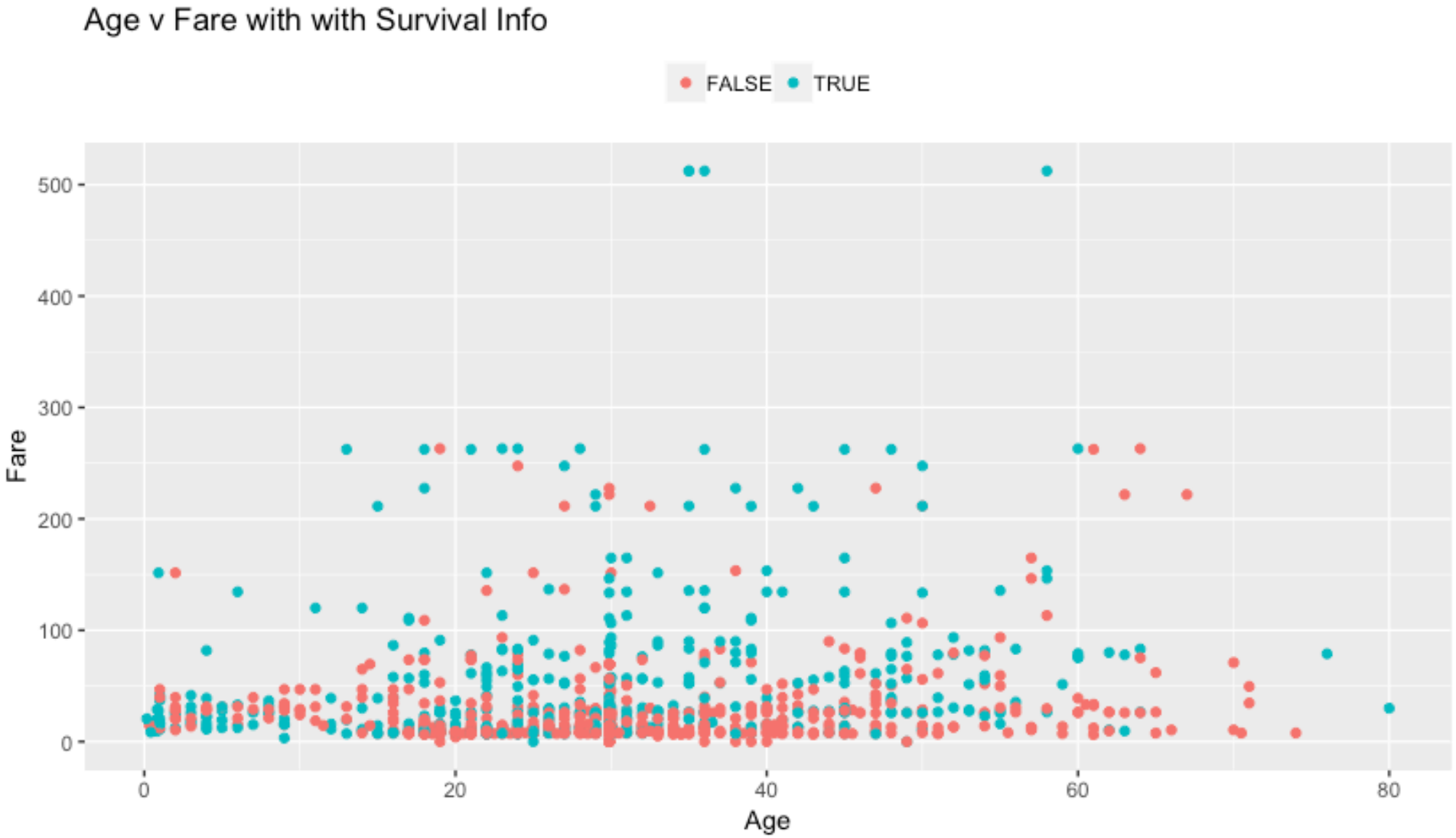
Plot 11


Age v Fare by Place of Embarkation
Cherbourg
Cobh
Southampton
Fare
500
400
300
200
100
0
0
20
40
60
80
Age

Plot 12


Age v Fare with Linear Model
Fare
500
400
300
200
100
0
0
20
40
60
80
Age

Plot 13


Age v Fare with with Survival Info
FALSE
TRUE
500
400
300
200
100
0
Fare
0
20
40
60
80
Age

Plot 14


Age v Fare By Travel Class and Point of Departure
Cherbourg
Cobh
Southampton
First
Second
Third
Fare
Age
0
100
200
300
400
500
0
20
40
60
80