Titanic - Kaggle - Some exploration

Explore for classroom use

Bruno Fischer Colonimos

23 mai 2018

Contents

---

1 Preliminary code

1.1 Libraries and auxiliary code (install before running)

(Not fully echoed here)

1.1.1 code config parameters

```
# adjust Gggplot2 theme and color palette
bwtheme <- TRUE
specialpalette <- FALSE
showarning <- FALSE
datmis <- "Ukn"
```

1.1.2 Required packages (install before running)

"caret", "ggplot2", pander

2 Data

2.1 Get it

The data is downloaded from Kaggle (https://www.kaggle.com/c/titanic) and saved. It is loaded here from disk.

- The data head is shown in table 1

Table 1: A glimpse of the data (continued below)

PassengerId	Survived	Pclass	Name	Sex	Age
1	0	3	Braund, Mr. Owen Harris	male	22
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38
3	1	3	Heikkinen, Miss. Laina	female	26
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35
5	0	3	Allen, Mr. William Henry	male	35
6	0	3	Moran, Mr. James	male	NA

SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	A/5 21171	7.25		S
1	0	PC 17599	71.28	C85	C
0	0	STON/O2. 3101282	7.925		S
1	0	113803	53.1	C123	S
0	0	373450	8.05		S
0	0	330877	8.458		Q

2.2 Variables organization and work to do

Table 3: Data organization

Variables	type	Values	Treatment
Demographic variables			
Sex	String	female ; male	Make factor
Age	numeric		many NA's
Agestat	factor	enfant/ado/adulte	new:
Family context			
SibSp	numeric		Combined ->
Parch	numeric		Combined ->
Famly	numeric		new: SibSp + Parch

Variables	type	Values	Treatment
Relationship to ship			
Pclass		1,2,3	Make factor
ticket		ticket number	not used
Fare	numeric		
Cabin		Cabin nbr	modified -> Deck
Deck	String		new (from Cabin)

Embark	String	C = Cherbourg,
		Q = Queenstown,
		S = Southampton
			make Factor

Survival			
survived	binary	0/1	Make factor

- Quick dataframe summary: table 4

Table 4: Data summaries (continued below)

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	NA
3rd Qu.:668.5	3rd Qu.:1.0000	3rd Qu.:3.000	NA
Max. :891.0	Max. :1.0000	Max. :3.000	NA
NA	NA	NA	NA

Table 5: Table continues below

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
NA	Mean :29.70	Mean :0.523	Mean :0.3816
NA	3rd Qu.:38.00	3rd Qu.:1.000	3rd Qu.:0.0000
NA	Max. :80.00	Max. :8.000	Max. :6.0000
NA	NA's :177	NA	NA

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
NA	Mean : 32.20	NA	NA
NA	3rd Qu.: 31.00	NA	NA
NA	Max. :512.33	NA	NA
NA	NA	NA	NA

2.3 Data modifications

- Make `Survived` , `Pclass` and `Embark` factors,
- Create `Famly` = `SibSp + Parch`
- Create `Sex.Pclass`
- Substitute missing values with Ukn in variable `Embarked`
- added variables
  - `Agestat` = age status : “child”, “teen”, “adult” (cutoff ages = 12, 18)
  - `Title` = civility.
  - `Letticket` = Ticket number begins with letters (yes = 1, no = 0)
  - `Hascabin` = is the cabin number known? (yes = 1, no = 0)
  - `Deck`= if the cabin is known, the first letter is the Deck (T, A, B, C. . .), otherwise “Ukn”

Table 7: Before modifications, Number of missing values (continued below)

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	0	0	0	0	177	0	0	0

Fare	Cabin	Embarked
0	0	0

3 Data Analysis

3.1 Passenger Identity

3.1.1 Sex and Age

- Gender

Table 9: Gender distribution

	female	male
Frequency	314	577
Rel.Frequency	0.352	0.648

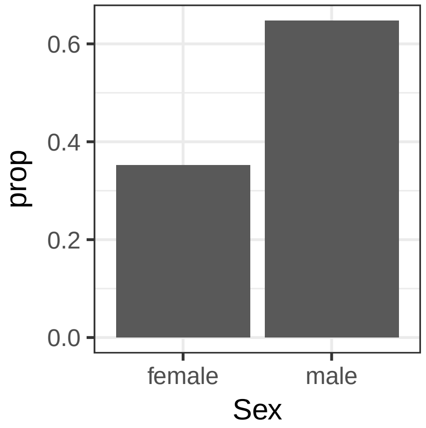


Figure 1: Gender distribution

Figure 1 shows the gender distribution

- Age :

Table 10: Age distribution

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.42	20.12	28	29.7	38	80	177

- Gender and age

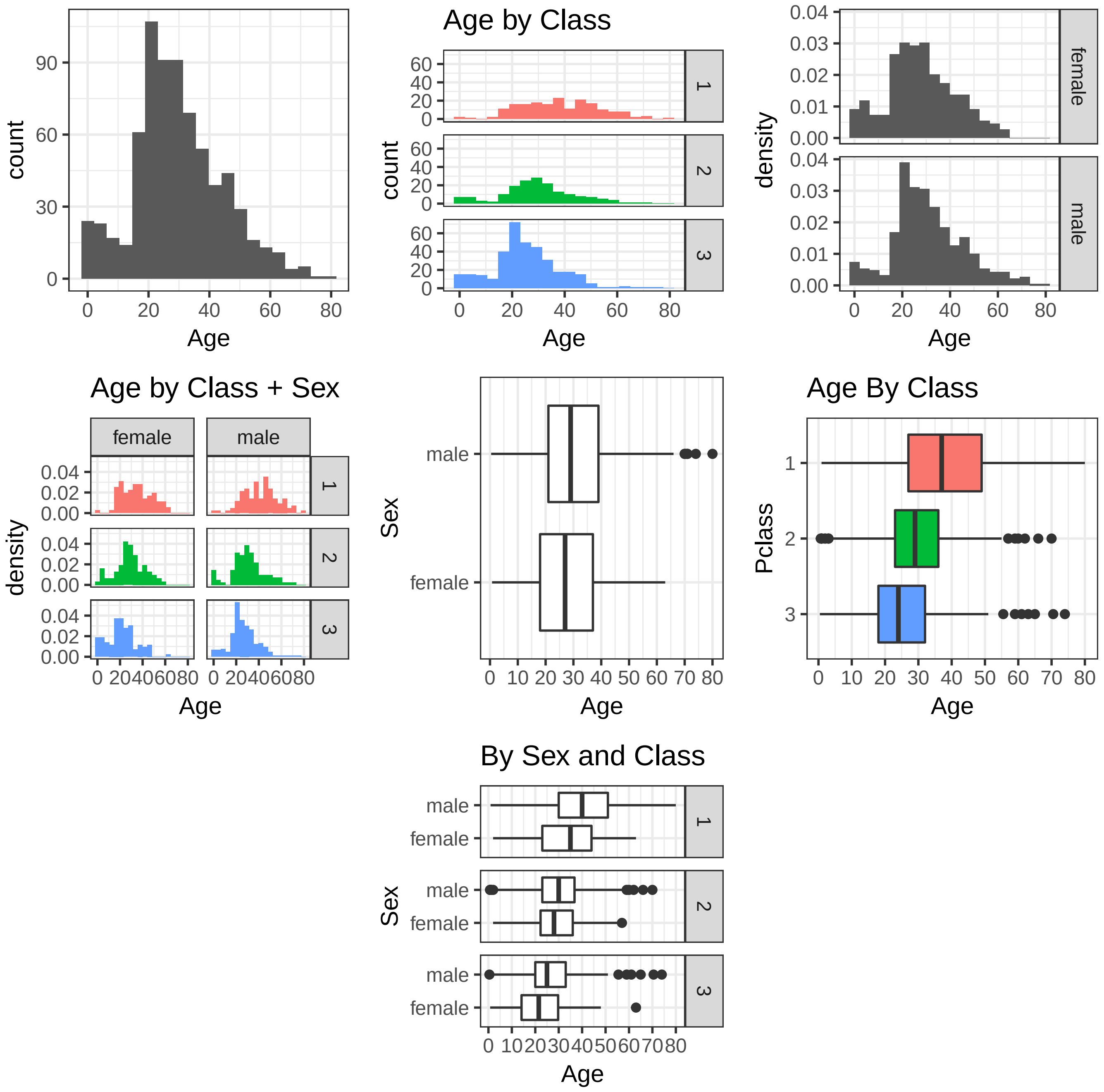


Figure 2: Age and Gender

Figure 2 shows the age distributions of both genders

- Age Status :

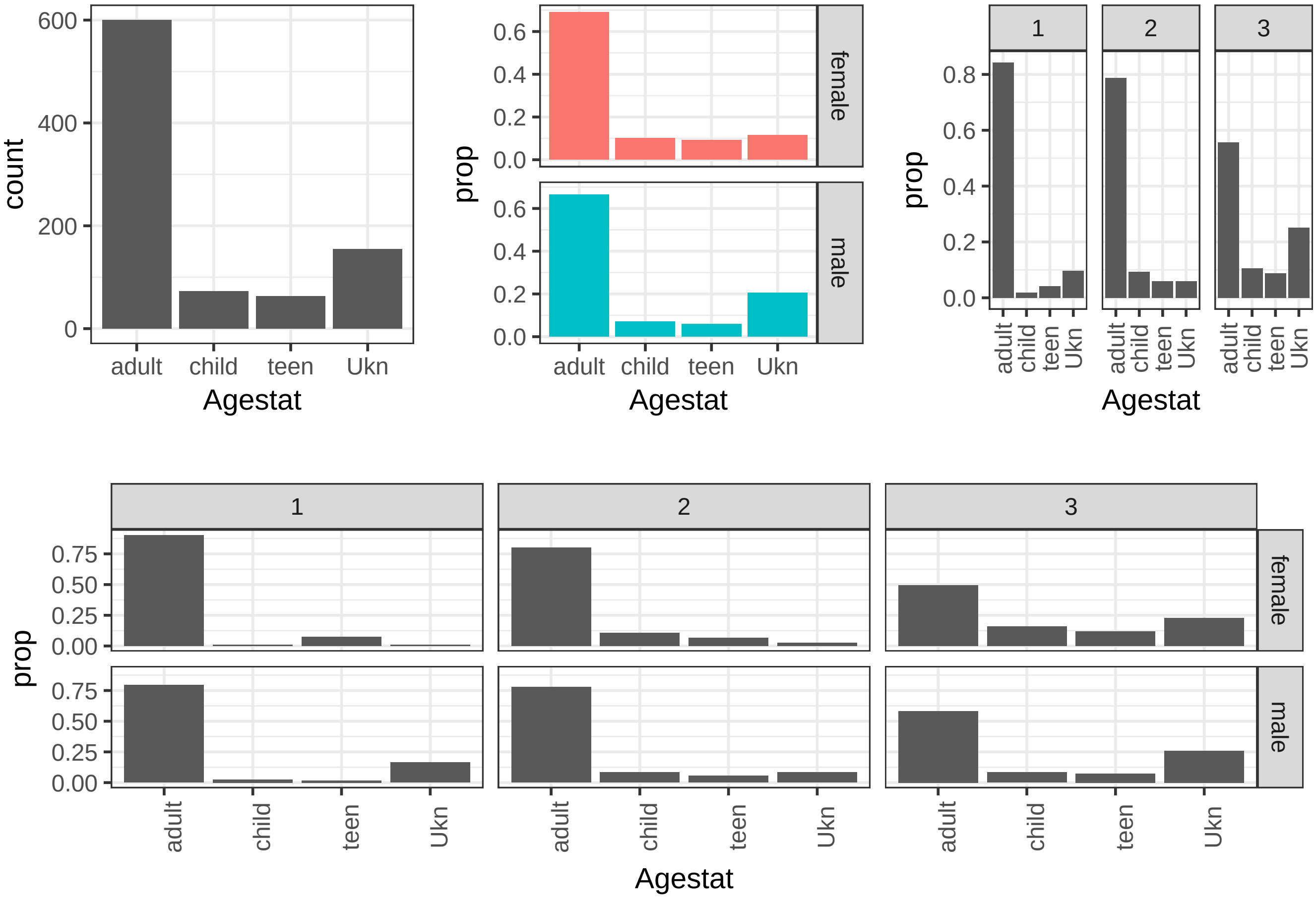


Figure 3: Age status by sex and class

3.2 Family members on board

Table 11: Family members: correlation of the variables

	SibSp	Parch	Famly
SibSp	1	0.4148	0.8907
Parch	0.4148	1	0.7831
Famly	0.8907	0.7831	1

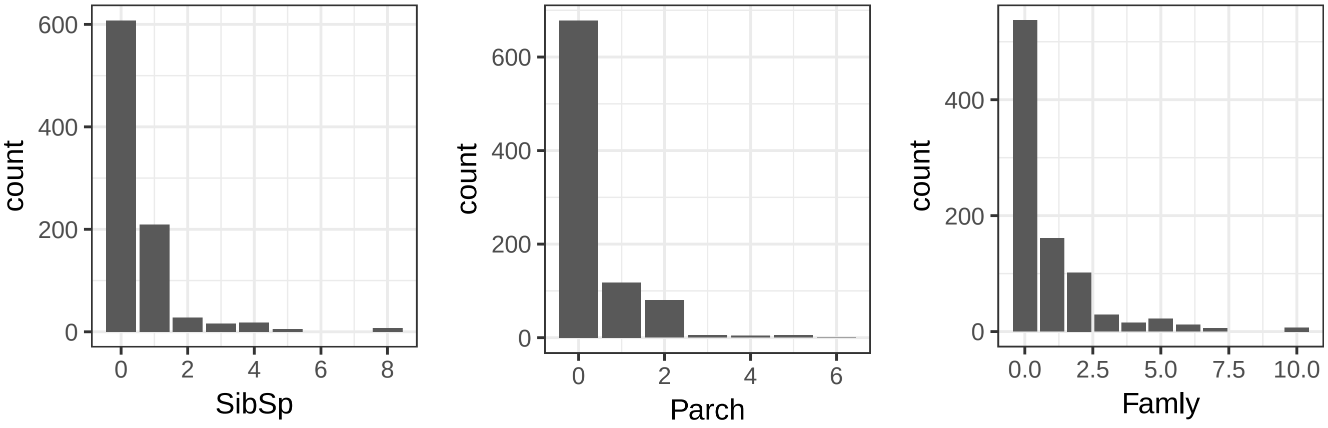


Figure 4: Family members on board

Figure 4 shows the Family size variablesand their correlation. Famly seems to convey most of the information.

3.3 Passenger class

figure 5 Shows that the third class accounts for about 55% of the passengers.

Table 12: Passenger Class (Pclass)

	1	2	3
frequency	216	184	491
rel.frequency	0.2424	0.2065	0.5511

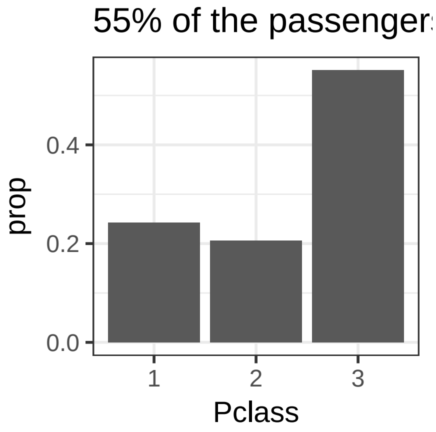


Figure 5: Passenger Classes Distribution

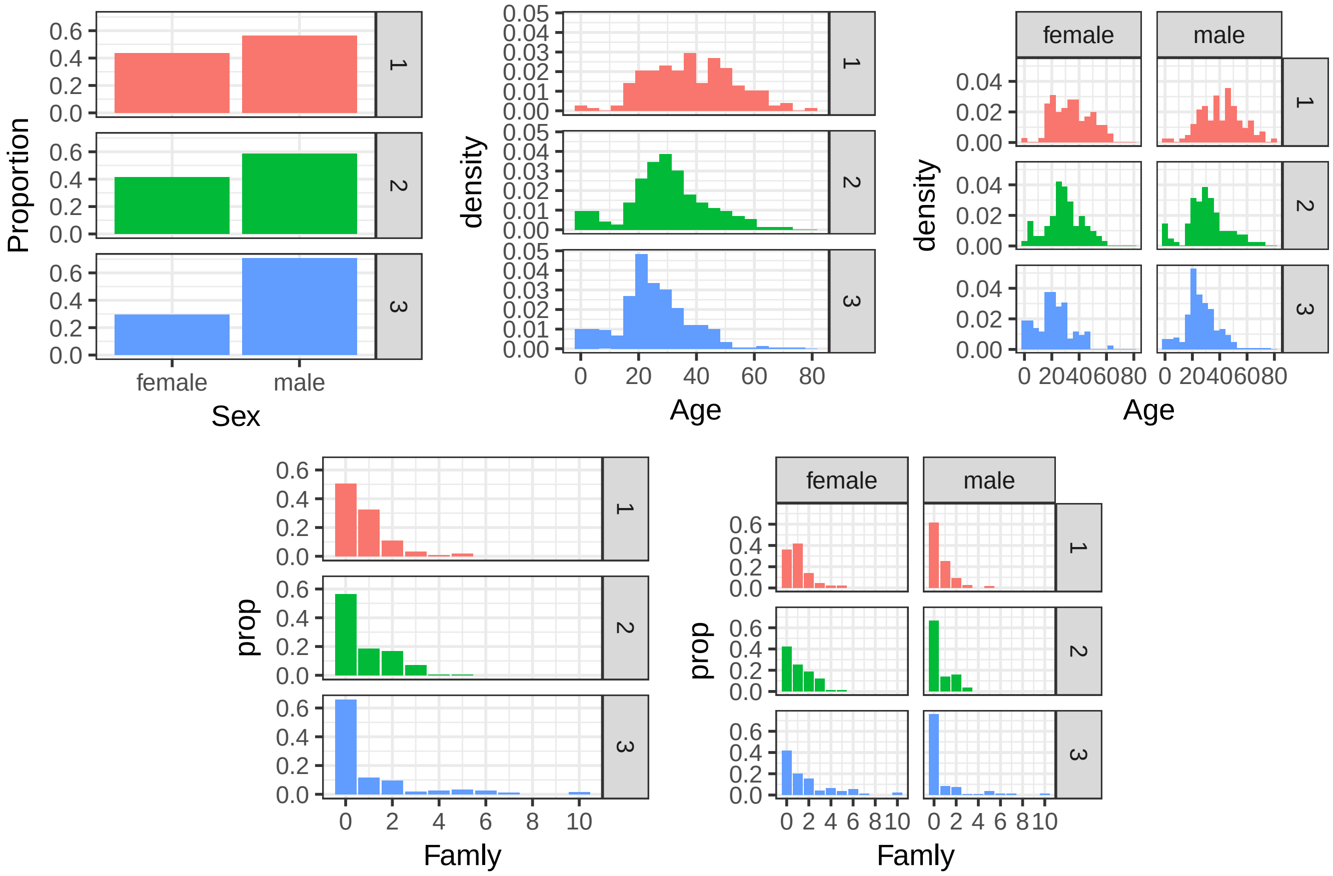
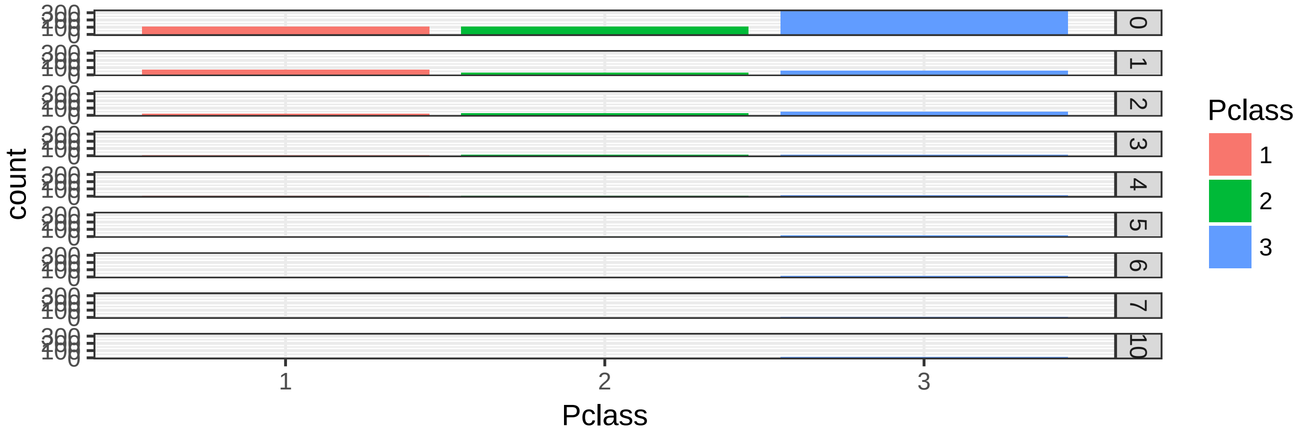


Figure 6: Passenger demographics by class

Figure 6. Compare with reference : figure 5



3.4 Deck

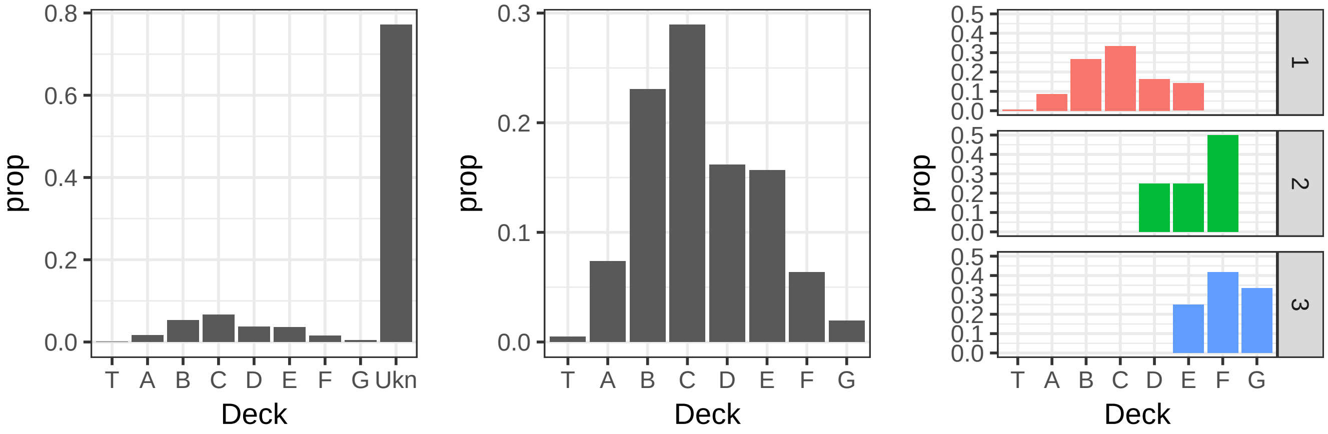


Figure 7: Distribution of decks

Figure ‘r .ref(“fig:”, “Deck”) shows a very partial distribution of decks: the deck of nearly 80% of the passengers is unknown. However, the third graph reveals the strong link between the passencger class and the deck.

3.5 Embarkation point

The passengers could embark at Southampton (S) , Cherbourg or Cobh, alias Queenstown (Q)

Table 13: Embarkation port

	S	C	Q
frequency	644	168	77
Rfreq	0.72	0.19	0.087

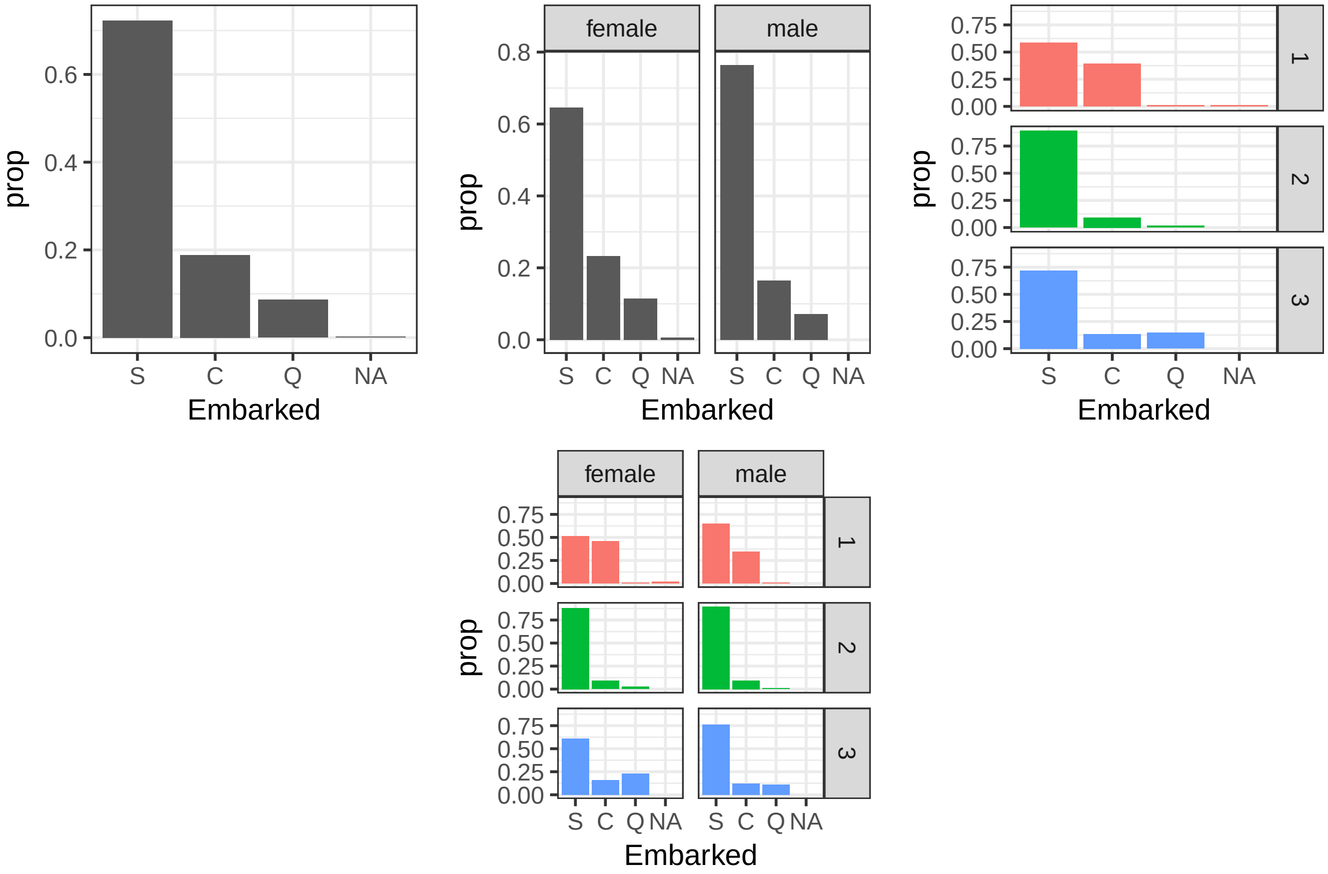


Figure 8: Embarkation port (C = Cherbourg, Q = Queenstown, S = Southampton)

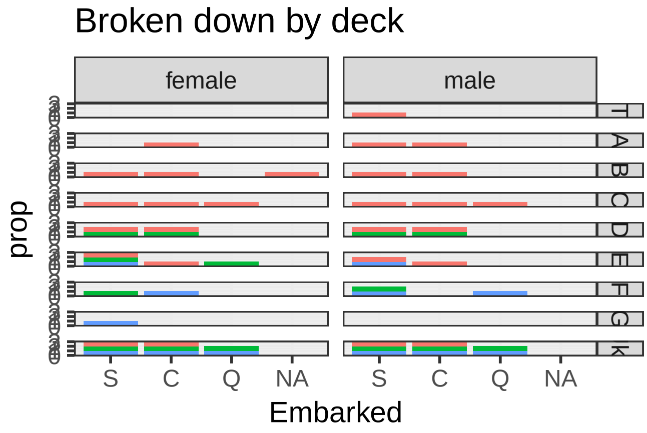


Figure 8 shows that Southampton was the major embarkation point (72.4409449% of the passengers). However, this is not as true for the women, particularly for the women with a first-class ticket.(only about 50% of them embarked at Southampton)

3.6 The Fare

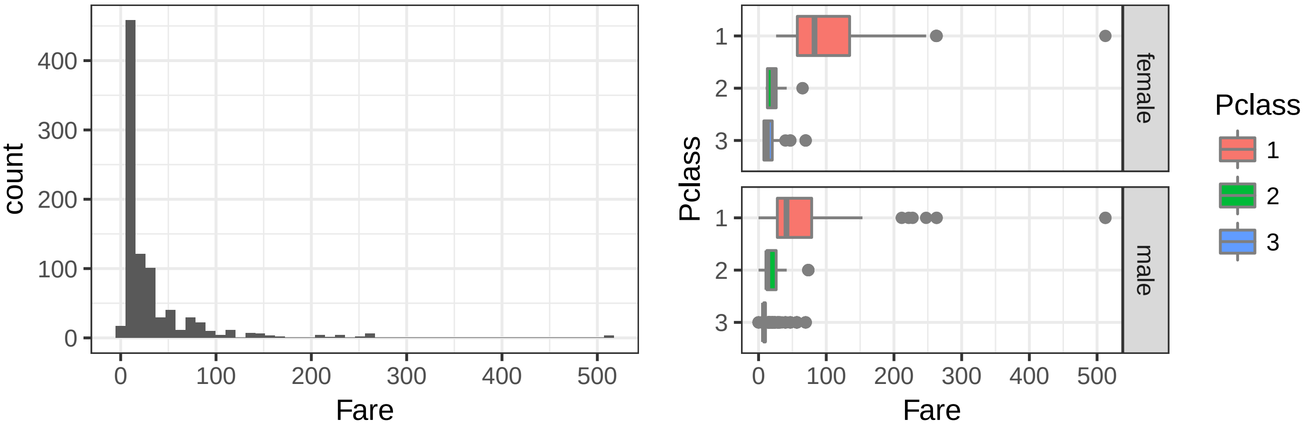


Figure 9: Fare by class and Sex

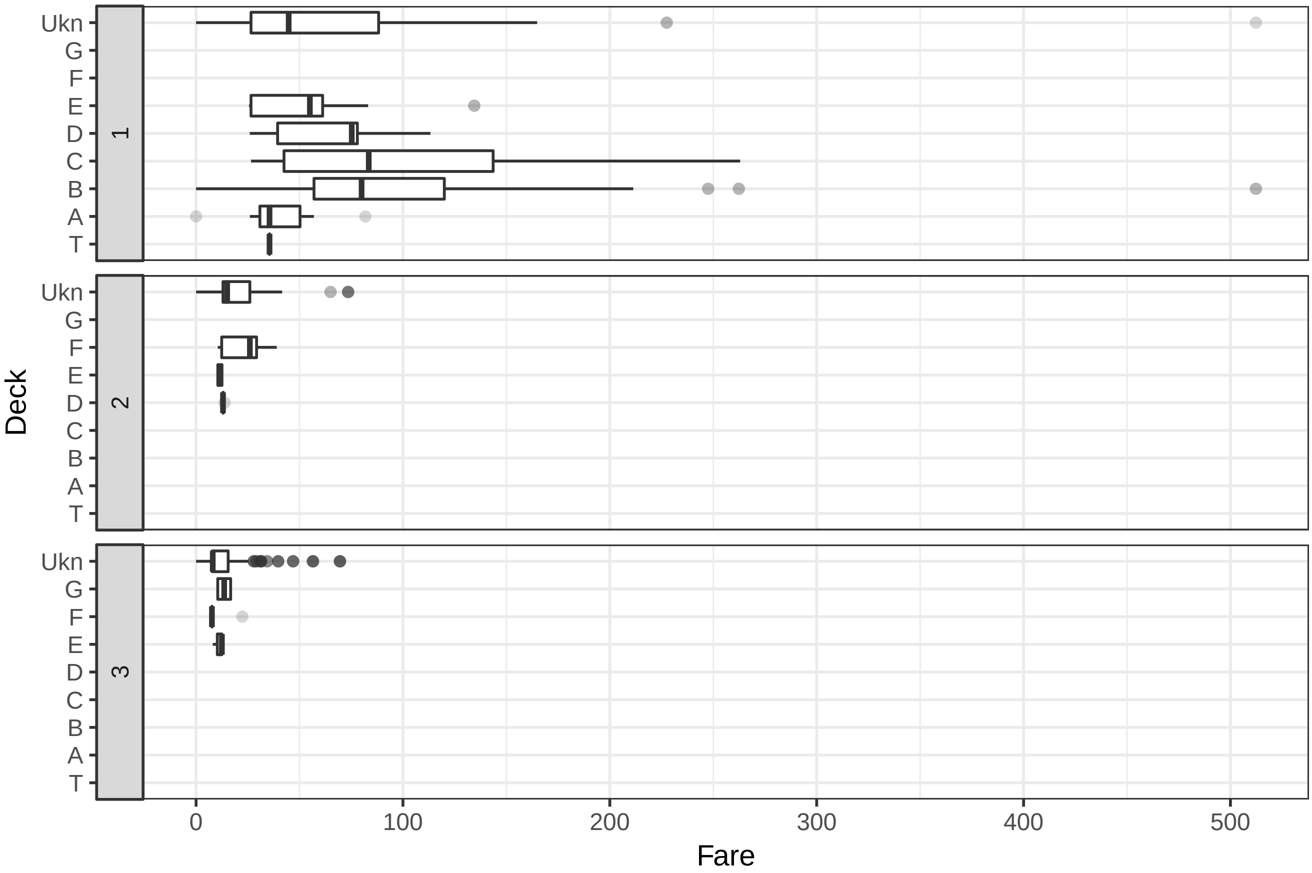


Figure 10: fare by deck and class

3.7 Survival

3.7.1 overall

Table 14: Suvival distributiond

No	Yes
0.62	0.38

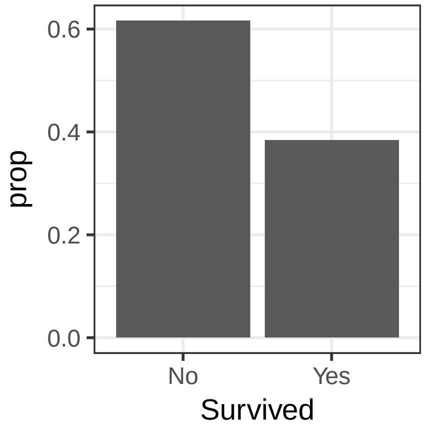


Figure 11: Overall survival

3.7.2 By category

Table 15: Survival by Sex

	female	male
No	0.26	0.81
Yes	0.74	0.19

Table 16: Survival by Passenger class

	1	2	3
No	0.37	0.53	0.76
Yes	0.63	0.47	0.24

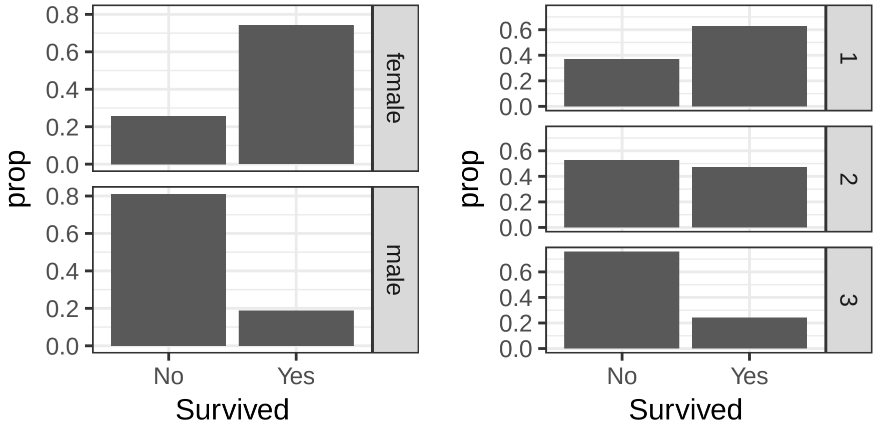


Figure 12: Survival by sex or Passenger Class

- Class and sex

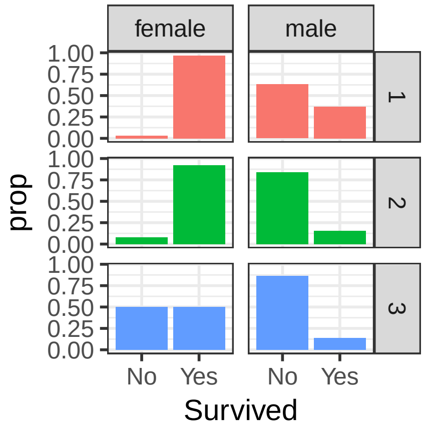


Figure 13: Survival by Sex and Class

- Class , sex, Port of embarkation

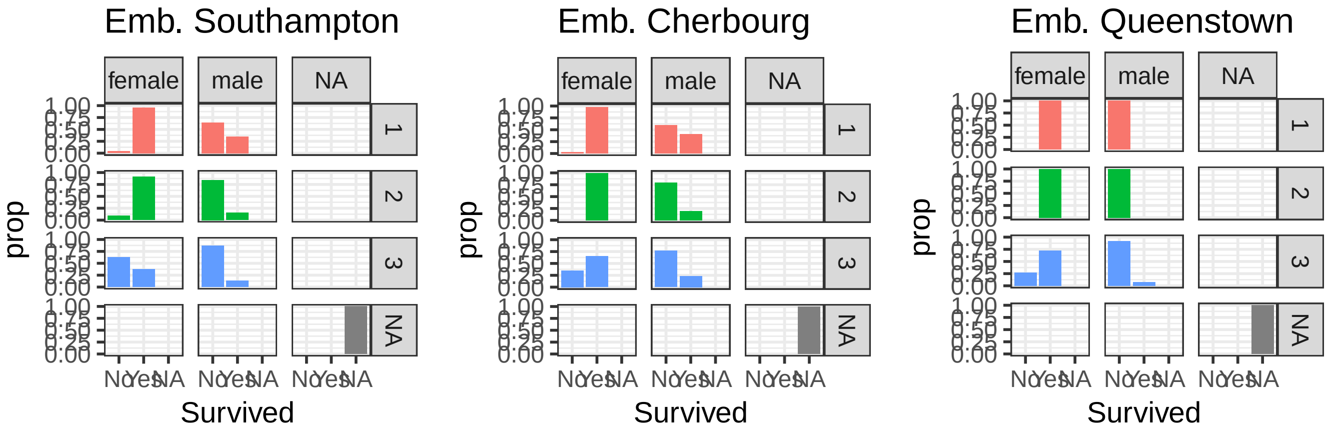


Figure 14: Survival by Sex and Class / Embarked

Class and Age see figure 15

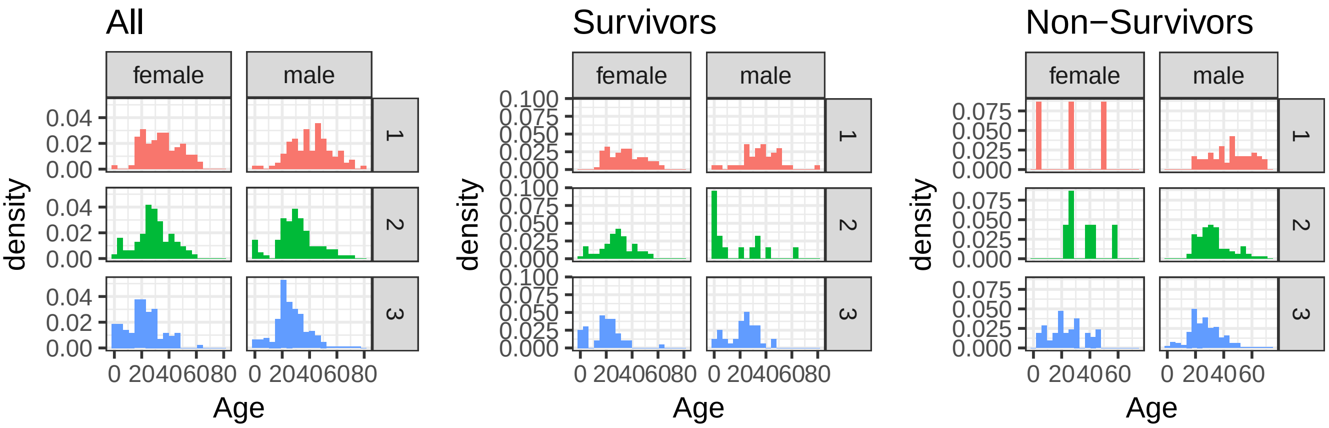


Figure 15: Age by Survival and Class

- Class and Agestat, Agestat and sex

Table 17: Distribution of age-status vs sex and Pclass

		1	2	3
adult	female	85	61	71
	male	97	84	202
child	female	1	8	23
	male	3	9	29
teen	female	7	5	17
	male	2	6	26
Ukn	female	1	2	33
	male	20	9	90

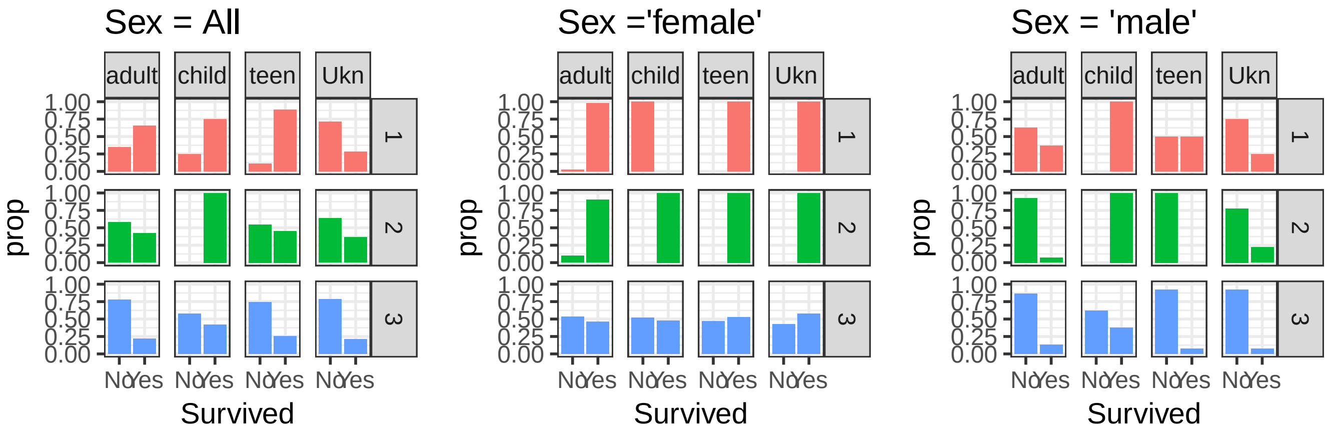


Figure 16: Survival by age status, Class and sex

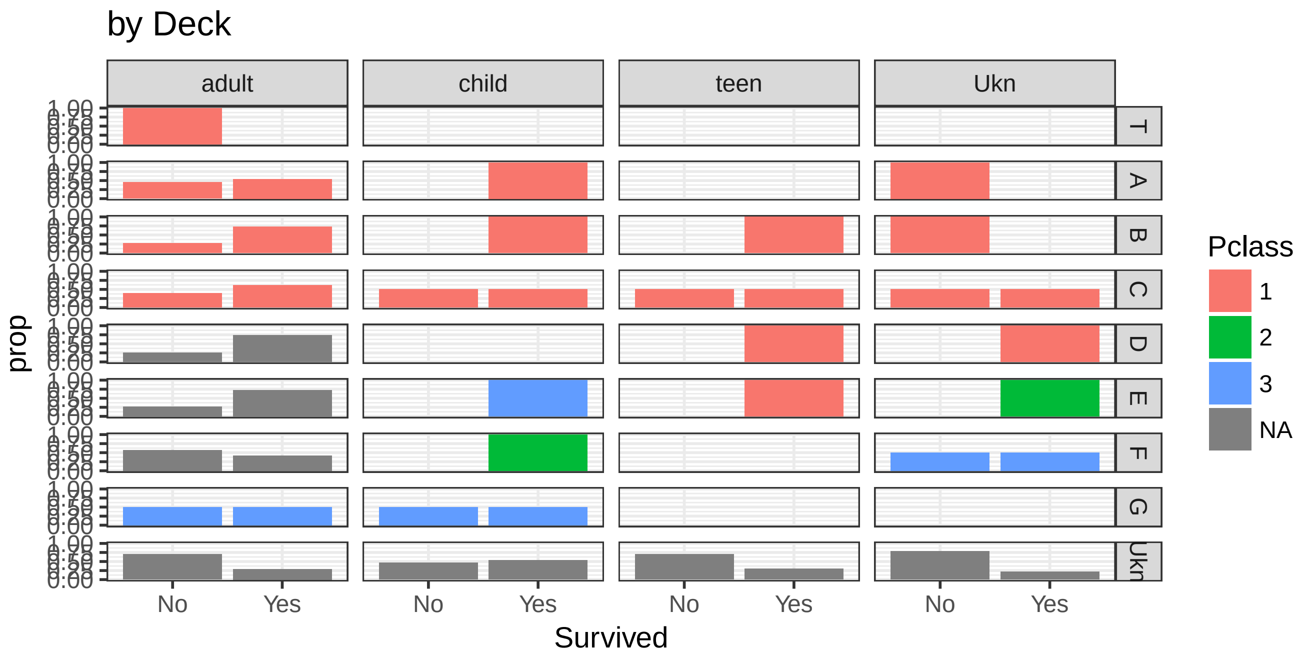


Figure 17: Survival by Deck, age and class

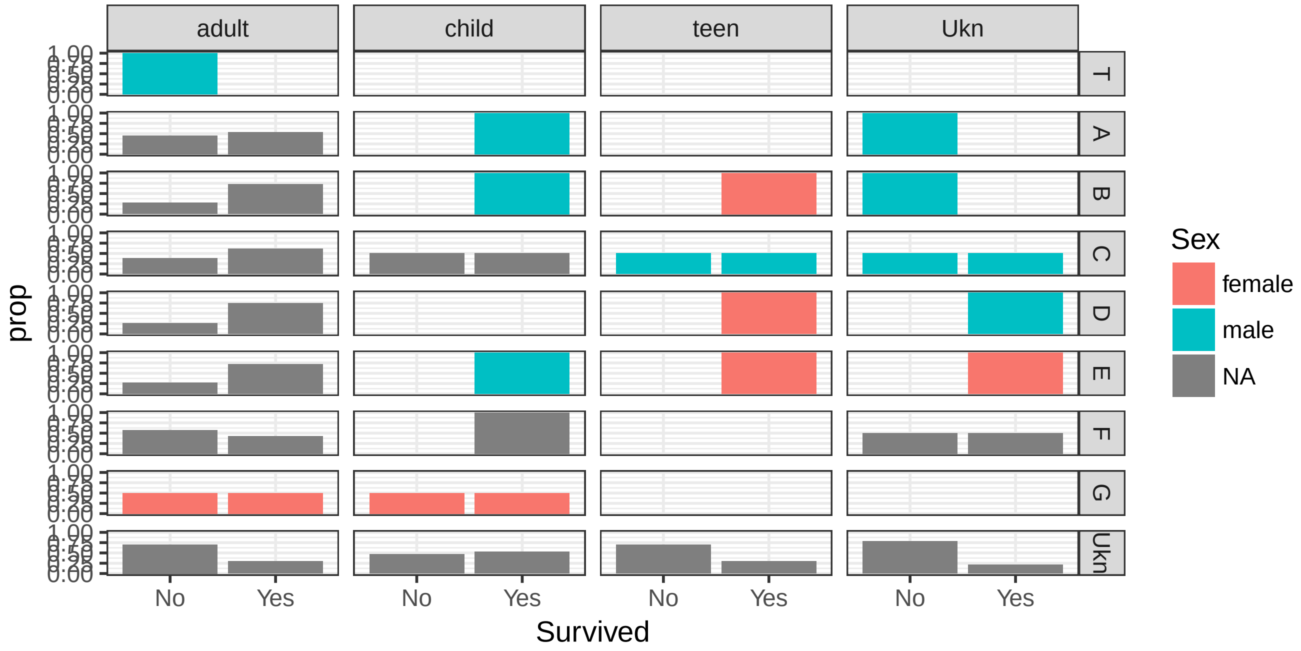


Figure 18: Survival by Deck, Age and Sex