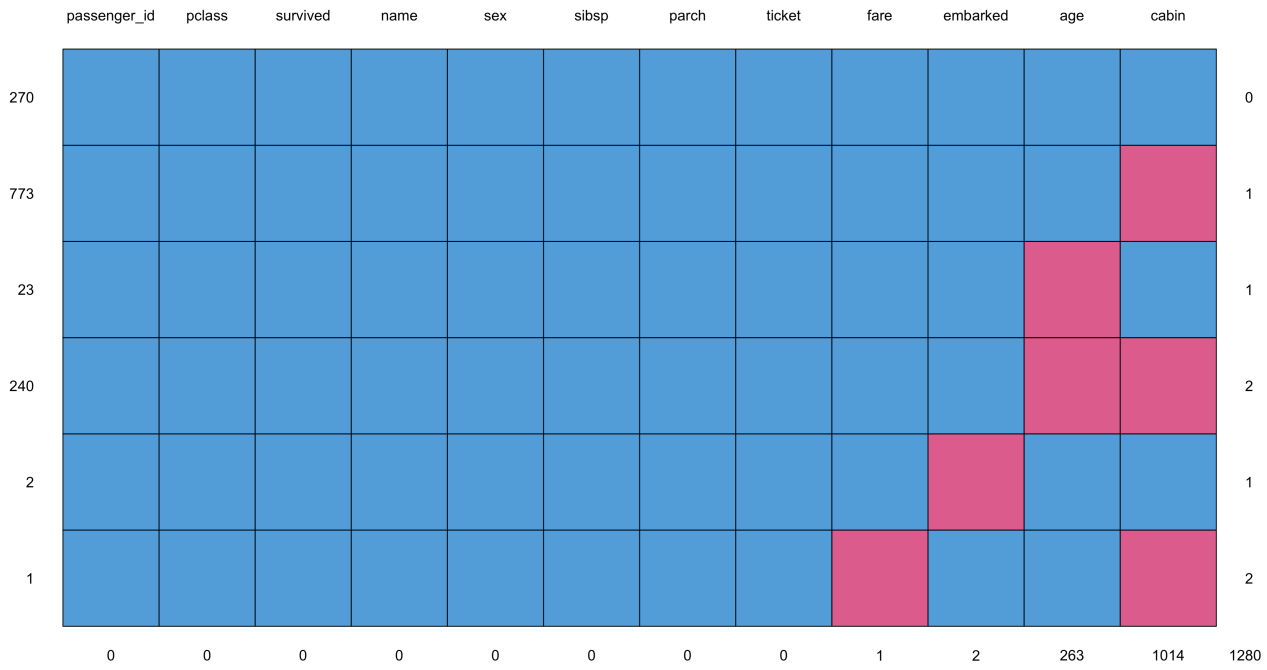
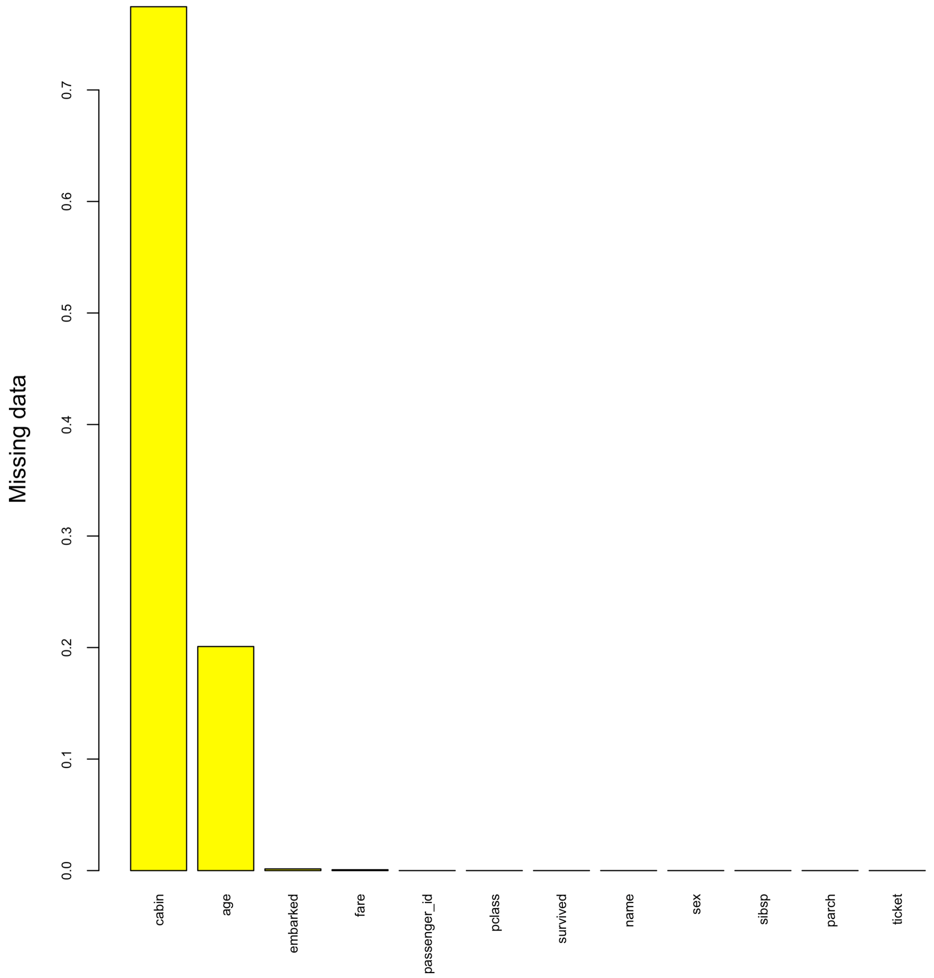
1. Analiza braków danych

#age missing 263 times
#fare missing once
#cabin missing 1014 times
#embarked missing twice



	passenger_id	pclass	survived	name	sex	sibsp	parch	ticket	fare	embarked	age	cabin	
270	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	0
773	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	1
23	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	1
240	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	2
2	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	1
1	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟥	2
	0	0	0	0	0	0	0	0	1	2	263	1014	1280



Age missing % around 20% of dataset observations number

Cabin missing % over 77% of dataset observations number

Fare and Embarked missing % below 1% of dataset observations number

1.2 Analiza struktury dla zmiennych: sex, pclass, embarked

SEX

- no missing values
- Male : 843 | 64,4 %
- Female : 466 | 35,6 %

PCLASS

- no missing values
- 1st class : 709 people | 54,16%
- 2nd class : 277 people | 21,16%
- 3rd class : 323 people | 24,68%

EMBARKED

- 2 values missing
- Southampton : 914 | 69,82%
- Queenstown : 123 | 9,4%
- Cherbourg : 270 | 20,63%
- NA - Unspecified : 2 | 0,15%

2. Częstości dla zmiennej objaśnianej survived

SURVIVED: 500 | 38,2%
NOT SURVIVED: 809 | 61,8%

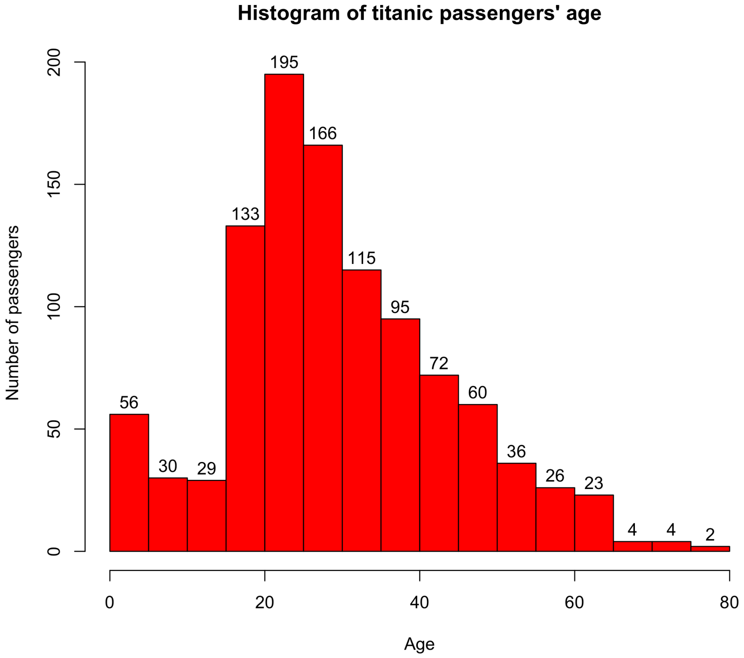
3. Braki danych w poszczególnych kolumnach

- w punkcie 1

4. Statystyki opisowe dla zmiennych age i fare

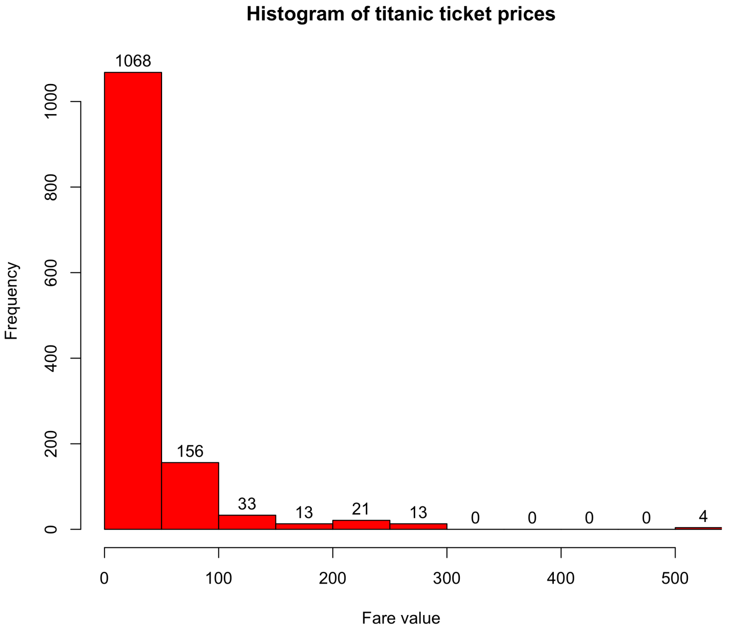
AGE

#not applying imputations -> 263 values missing
#min - 0.1667 -> Youngest person was a baby younger than a year
#max - 80.00 -> Oldest person was 80 years old
#avg - 29.88 -> The average of age among titanic passengers was 29.88 years
#med - 28.00 -> The median of age among titanic passengers was 28.00 years
#Q1 - 21.00 -> 25% of titanic passengers were in the age <= 21.00 years and 75% of passengers were in the age >= 21.00
#Q3 - 39.00 -> 75% of titanic passengers were in the age <= 39.00 years and 25% of passengers were in the age >= 39.00



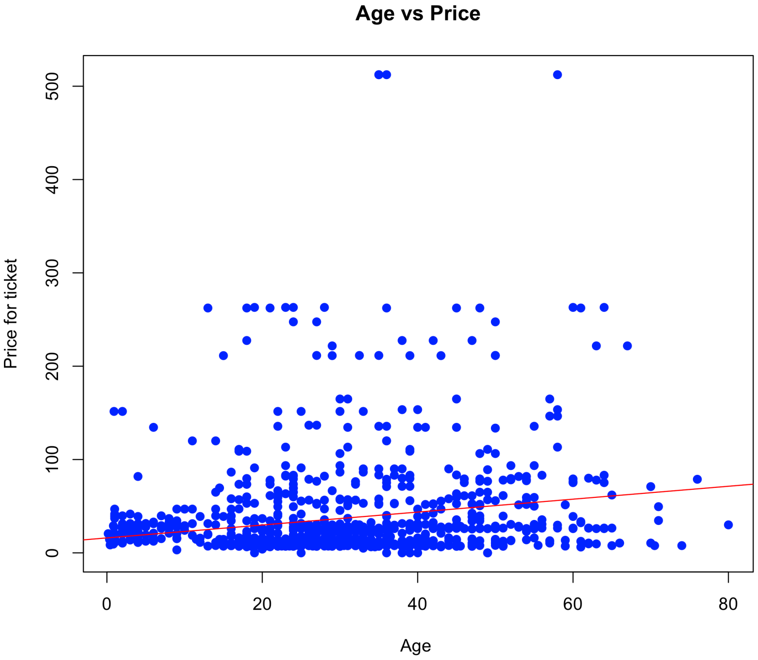
FARE

#not applying imputations -> 1 value missing
#min - 0.000 -> We can quess -> some people were traveling for free (invitation, reward etc.)
#max - 512.329 -> Value of most expensive ticket was 512.32 in unspecified currency
#avg - 33.295 -> The average of ticket prices paid by the passengers was 33.29 in uspecified currency
#med - 14.454 -> The median value of ticket price paid by the passengers was 14.45 in unspecified currency
#Q1 - 7.896 -> 25% of titanic passengers had to pay a price <= 7.89 and 75% of them had to pay >= 7.89
#Q3 - 31.275 -> 75% of titanic passengers had to pay a price <= 31.27 and 25% of them had to pay >= 31.27



5. Wykres rozrzutu age vs fare + badanie korelacji

#H0: correlation = 0
#H1: correlation != 0
#p-val -> <0.05
#=> H1 -> cor = 0.1787394 -> weak positive correlation.
#Indicates a weak linear relationship between the variables -> the age does not define the price of the ticket



6. Podział na training/test set => w kodzie

7. Porównanie struktury survived w obu zbiorach

TRAINING SET:
#0 - not survived - 55 people | 30,56%
#1 - survived - 125 people | 69,44%

TEST SET:
#0 - not survived - 35 people | 38,89%
#1 - survived - 55 people | 61,11%

8. Zastąpienie braków dla zmiennej objaśniającej embarked => w kodzie

9 i 10. Braki danych dla zmiennej objaśniającej age => w kodzie

A -> uzupełniono wartością średnią całego zbioru
B -> Wykorzystanie metody imputacji HMISC z opcją median

11. Oszacowanie modelu logitowego GLM

Według współczynnika istotności -> zmienne age, sex, pclass są istotne do dalszej analizy

12,13,14,15