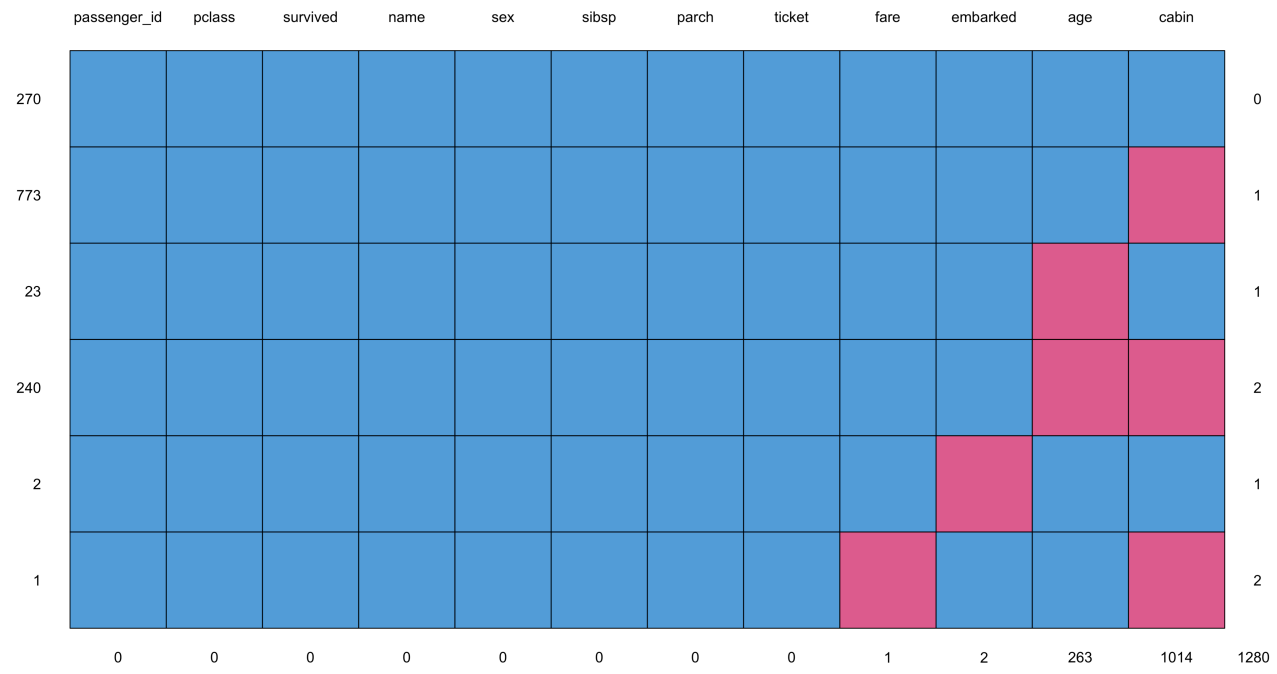
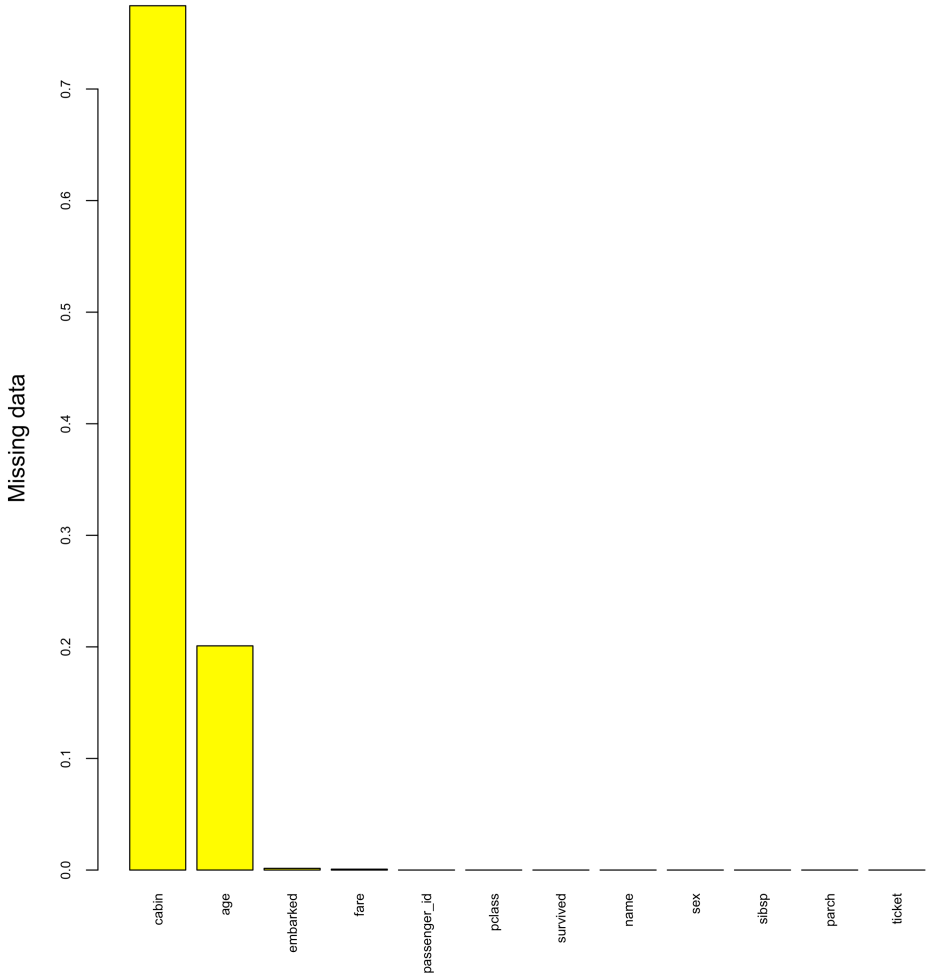
1. Analiza braków danych

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





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%

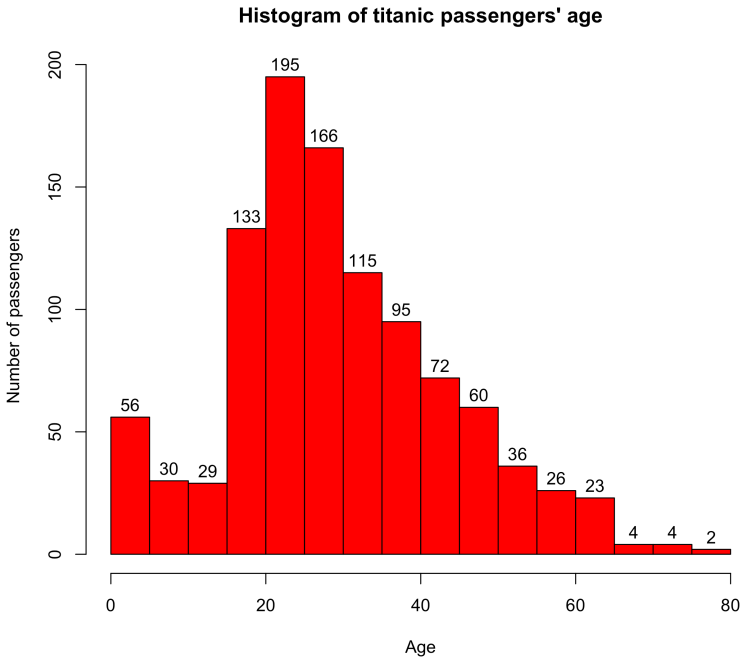
3. Braki danych w poszczególnych kolumnach

- w punkcie 1

4. Statystyki opisowe dla zmiennych age i fare

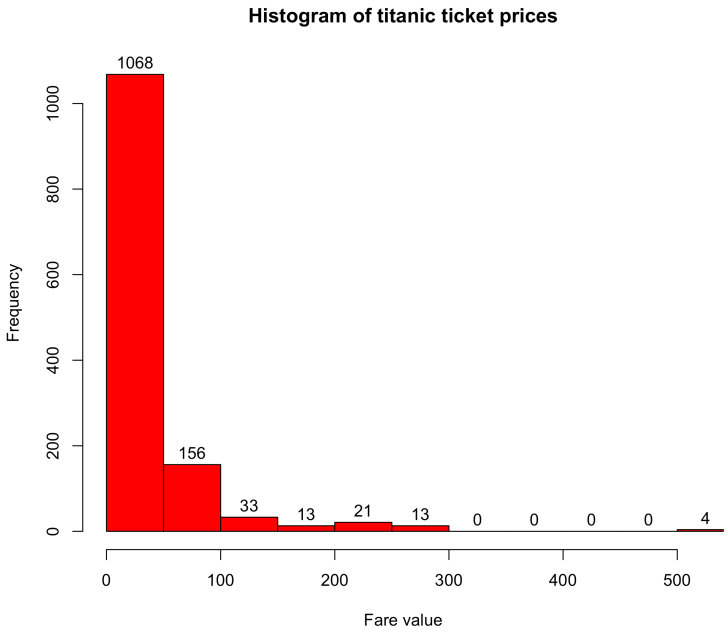
AGE

#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



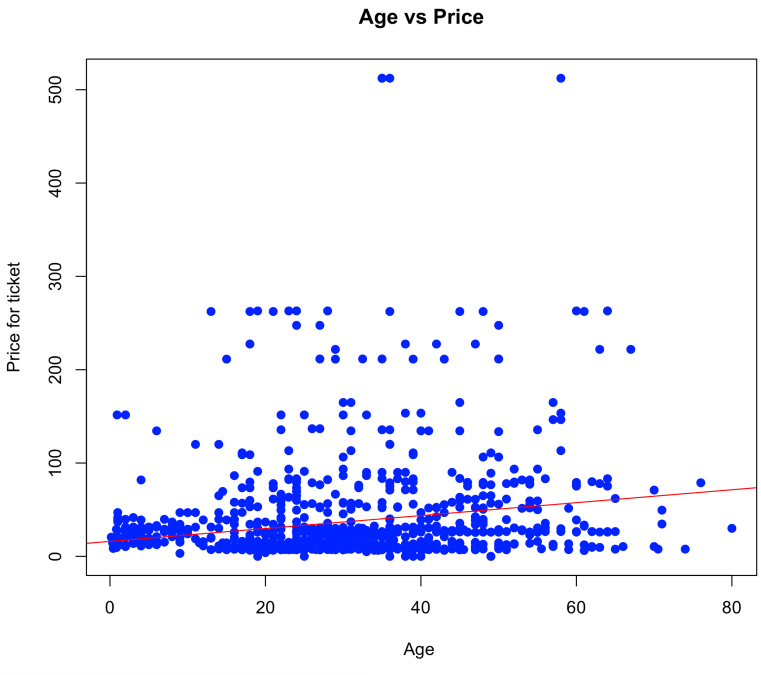
FARE

#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 & 12 Oszacowanie modelu regresji logistycznej GLM

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

#model formula
#hasSurvived = 3.496 - 0.03(age) - 2.44(the patient is male) - 1.0596(travelingBy2ndClass) - 1.879115(travelingBy3rdClass) - 0.595(GotOnBoarnInPortQ) -0.4122(GotOnBoardInPortS)

#significant factors - sex, pclass, age, embarked is not significant -> not taking into the account during further analysis

#model formula
#hasSurvived = 3.496 - 0.03(age) - 2.44(the patient is male) - 1.0596(travelingBy2ndClass) - 1.879115(travelingBy3rdClass) - 0.595(GotOnBoarnInPortQ) -0.4122(GotOnBoardInPortS)

Interpretacje:

#Along with increase of age by 1 year the chances of survival were decreasing by 3% ceteris paribus
#Changes of survival for men were decreased by 244% than in case of female ceteris paribus
#Chances of survival for passengers traveling in the second class were lower by 105.9% than for passengers traveling in first class ceteris paribus
#Chances of survival for pasengers traveling in third class were lower by 187,9% than for passengers traveling first class ceteris paribus

AIC - 844.1
ACC - 79,4% -> model ma 79,4 % skuteczności ogólnej w prognozowaniu
SENS - 70,5% -> w 70.5% model poprawnie sklasyfikował osoby które przeżyły jako te, które rzeczywiście przeżyły
SPEC - 84,2% -> w 70.5% model poprawnie sklasyfikował osoby które nie przeżyły jako te, które rzeczywiście nie przeżyły
AUC - 84,6% -> model cechuje się 84.6 % stopniem trafnego prognozowania
GINIE - 69,24% -> model jest idealny w 69.24%
F1 Score - 70,9% -> współczynnik F1 precyzji prognozy wynosi 79,9%

13. Oszacować random forest => w kodzie

#results
#OOB estimate of error rate: 20.62%
#Confusion matrix:
0   1 class.error
#0 482 45 0.08538899
#1 135 211 0.39017341

14. Interpretacja Random Forest

ACC - 81,19% -> model ma 81,19% skuteczności ogólnej w prognozowaniu

SENS - 81,0% > w 81% model poprawnie sklasyfikował osoby które przeżyły jako te, które rzeczywiście przeżyły
SPEC - 81,25%-> w 81,25% model poprawnie sklasyfikował osoby które nie przeżyły jako te, które rzeczywiście nie przeżyły
AUC - 68,72% -> model cechuje się 68,72% stopniem trafnego prognozowania
GINIE - 69,59%-> model jest idealny w 69,59%
F1 Score - 69,59% -> współczynnik F1 precyzji prognozy wynosi 69,59%

15. Porównanie modelu GLM + Random Forest

#Będę mógł przejść do tego dopiero w momencie, kiedy uda mi się dokończyć pełne statystyki dla GLM i Random Forest

GLM

AIC - 844.1
ACC - 79,4% overall success in prediction
SENS - 70,5% success in predicting real survived case
SPEC - 84,2% success in predicting real not survived cases
AUC - 84,6% rate of successful classification
GINIE - 69,24% degree of ideality of the model
F1 Score - 70,9% measure of a test's accuracy

RF

ACC - 81,19% overall success in prediction
SENS - 81,0% success in predicting real survived case
SPEC - 81,25% success in predicting real not survived cases
AUC - 68,72% rate of successful classification
GINIE - 69,59% degree of ideality of the model
F1 Score - 69,59% measure of a test's accuracy

Reguły decyzyjne

AIC - MIN. => -
ACC - MAX => RF
SENS - MAX => RF
SPEC - MAX => GLM
AUC - MAX => GLM
GINIE - MAX => RF
F1 SCORE - MAX => GLM

Decyzja:

Oba modele są dość zbliżone jeżeli chodzi o skuteczność predykcji. Jednakże mając na uwadze stopień zdolności predykcji modeli zarówno przypadków osób które rzeczywiście przeżyły oraz przypadków które rzeczywiście nie przeżyły => model Random Forest wykazuje się lepszym % skuteczności i w obu przypadkach jest on bardzo zbliżony. Model GLM prognozuje natomiast nieco lepiej przypadki osób, które rzeczywiście nie przeżyły. W związku z tym najlepszym wyborem byłby model Random Forest

Wnioski: W celu poprawy skuteczności prognozy można rozważyć eliminację niektórych wartości odstających, które np. widoczne były na wykresie rozrzutu