Titanic Dataset

1 Introduction

The original Titanic dataset, describing the **survival status of individual passengers** (1309) on the Titanic. The Titanic data does not contain information from the crew, but it does contain actual ages of half of the passengers. The principal source for data about Titanic passengers is the Encyclopedia Titanica.

One of the original sources is Eaton & Haas (1994) Titanic: Triumph and Tragedy, Patrick Stephens Ltd. It includes a passenger list created by many researchers (edited by Michael A. Findlay).

Figure 1: The infamous Titanic. Source: https://wallpapercave.com

Dataset basic information:

Variable	Description
`survived` (target)	0 = No, 1 = Yes
`pclass`	1 = 1st; 2 = 2nd; 3 = 3rd
`name`	First and Last Name
`sex`	Sex
`age`	Age
`sibsp`	Number of Siblings/Spouses Aboard
`parch`	Number of Parents/Children Aboard
`ticket`	Ticket Number
`fare`	Passenger Fare
`cabin`	Cabin

Variable	Description

`embarked`	Port of Embarkation C = Cherbourg; Q = Queenstown;
	S = Southampton

`body`	Body Identification Number
`boat`	Rescue Boat Number
`home.dest`	Home Destination

We use OpenML (R-Package) to download the dataset in a machine-readable format and convert it into a `data.frame`:

```r
# load the dataset from OpenML Library
d <- OpenML::getOMLDataSet(data.id = 40945)

# convert the OpenML object to a tibble (enhanced data.frame)
titanic <- d %>% dplyr::as_tibble()
skimmed_titanic <- skimr::skim(titanic)
print(titanic, width = Inf)
```

```
## # A tibble: 1,309 x 14
##    pclass survived name                                            sex       age
##     <dbl> <fct>    <chr>                                           <fct>   <dbl>
##   1      1 1        Allen, Miss. Elisabeth Walton                   female 29
##   2      1 1        Allison, Master. Hudson Trevor                  male    0.917
##   3      1 0        Allison, Miss. Helen Loraine                    female  2
##   4      1 0        Allison, Mr. Hudson Joshua Creighton            male   30
##   5      1 0        Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female 25
##   6      1 1        Anderson, Mr. Harry                             male   48
##   7      1 1        Andrews, Miss. Kornelia Theodosia               female 63
##   8      1 0        Andrews, Mr. Thomas Jr                          male   39
##   9      1 1        Appleton, Mrs. Edward Dale (Charlotte Lamson)   female 53
## 10      1 0        Artagaveytia, Mr. Ramon                         male   71
##    sibsp parch ticket    fare cabin   embarked boat   body
##    <dbl> <dbl> <chr>    <dbl> <chr>   <fct>    <chr> <dbl>
##   1     0     0 24160    211.  B5      S        2        NA
##   2     1     2 113781   152.  C22 C26 S        11       NA
##   3     1     2 113781   152.  C22 C26 S        <NA>     NA
##   4     1     2 113781   152.  C22 C26 S        <NA>    135
##   5     1     2 113781   152.  C22 C26 S        <NA>     NA
##   6     0     0 19952     26.6 E12     S        3        NA
##   7     1     0 13502     78.0 D7      S        10       NA
##   8     0     0 112050     0   A36     S        <NA>     NA
##   9     2     0 11769     51.5 C101    S        D        NA
## 10     0     0 PC 17609  49.5 <NA>    C        <NA>     22
##    home.dest
##    <chr>
##   1 St Louis, MO
##   2 Montreal, PQ / Chesterville, ON
##   3 Montreal, PQ / Chesterville, ON
##   4 Montreal, PQ / Chesterville, ON
##   5 Montreal, PQ / Chesterville, ON
##   6 New York, NY
##   7 Hudson, NY
##  8 Belfast, NI
##  9 Bayside, Queens, NY
## 10 Montevideo, Uruguay
## # ... with 1,299 more rows
```

2 Exploratory Data Analysis (EDA)

In this part, we will walk through a few characteristics of Titanic dataset using library `skimr` and `DataExplorer`.

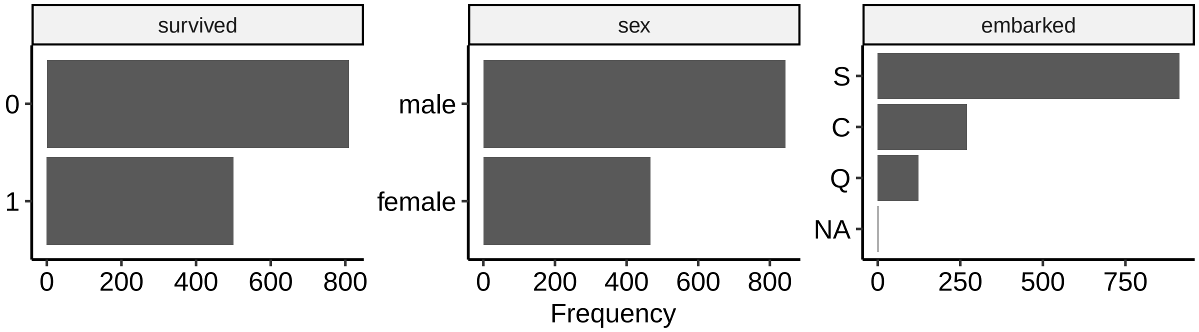
2.1 Factor variables

General statistics about factor variables from Titanic dataset:

```
skimr::partition(skimmed_titanic)$factor %>%
        knitr::kable(format = 'latex', booktabs = TRUE) %>%
        kableExtra::kable_styling(latex_options = 'HOLD_position')
```

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
survived	0	1.0000000	FALSE	2	0: 809, 1: 500
sex	0	1.0000000	FALSE	2	mal: 843, fem: 466
embarked	2	0.9984721	FALSE	3	S: 914, C: 270, Q: 123

```
titanic_factor <- titanic %>% select(where(is.factor))
DataExplorer::plot_bar(titanic_factor, ggtheme = ggpubr::theme_pubr(base_size = 10))
```



There are 3 factor variables from this dataset, namely `survived` (also the target), `sex`, and `embarked`. Both `survived` and `sex` don't have missing values, `embarked` only has 2 missing values (still 99.8% complete). The class distribution is not very imbalanced with 38% of passengers survived. Of all 1309 passengers, 36% are female and the majority of the passengers onboarded from Southampton (accounting for 70%). Looking at the `embarked` features, we may be tempted to figure out the two rows that miss this value:

```
titanic_embarked_na <- titanic %>% filter(is.na(embarked))
split(1:ncol(titanic_embarked_na), sort(rep_len(1:2, ncol(titanic_embarked_na)))) %>%
        map(~select(titanic_embarked_na, .)) %>%
        map(knitr::kable, booktabs = T) %>%
        map(kableExtra::column_spec, column = 3, width = "3cm") %>%
        map(kableExtra::kable_styling, latex_options = 'HOLD_position') %>%
        walk(print)
```

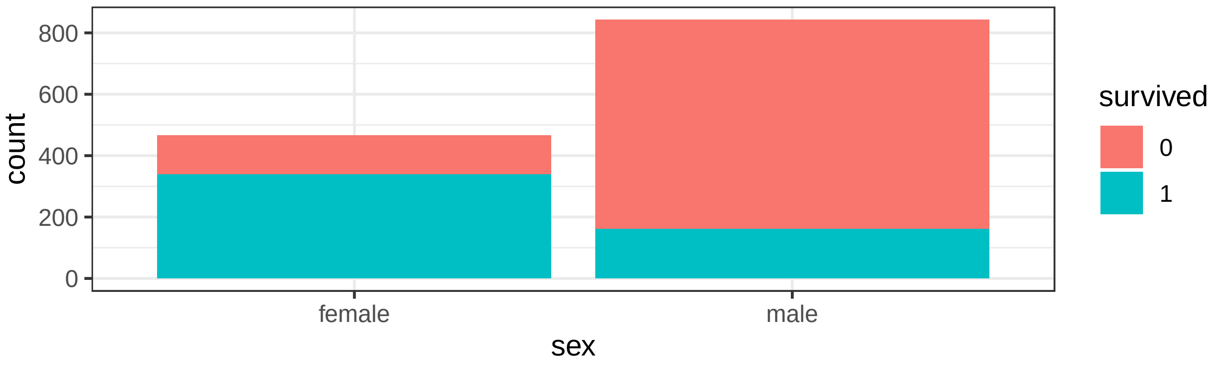
pclass	survived	name	sex	age	sibsp	parch
1	1	Icard, Miss. Amelie	female	38	0	0
1	1	Stone, Mrs. George Nelson (Martha Evelyn)	female	62	0	0

ticket	fare	cabin	embarked	boat	body	home.dest
113572	80	B28	NA	6	NA	NA
113572	80	B28	NA	6	NA	Cincinatti, OH

According to the two rows, these two passengers have the same ticket number, so it is likely that they onboarded together. Looking for the name of the two passengers, it leads to this link showing that both departed from Southampton (S).

We may also want to know if the `sex` affects the survival rate of a passenger.

```
titanic %>%
        select(sex, survived) %>%
        ggplot(mapping = aes(fill = survived, x = sex)) + geom_bar()
```



Looking from this bar plot, it suggests that `sex` does have an impact on survival rate and being a female can lead to higher survival rate. This can be an indicator that people prioritized to save women before men during the wreckage.

Next we take a brief look at the character (text) variables.

2.2 Character (text) variables

General statistics about character variables from Titanic dataset:

```
skimr::partition(skimmed_titanic)$character %>%
        knitr::kable(format = 'latex', booktabs = TRUE) %>%
        kableExtra::kable_styling(latex_options = 'HOLD_position')
```

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
name	0	1.0000000	12	82	0	1307	0
ticket	0	1.0000000	3	18	0	929	0
cabin	1014	0.2253629	1	15	0	186	0
boat	823	0.3712758	1	7	0	27	0
home.dest	564	0.5691367	5	50	0	369	0

There are 5 character variables: `name`, `ticket`, `cabin`, `boat`, and `home.dest`. `name` and `ticket` don't have any missing values, while `cabin`, `boat`, and `home.dest` only have 22.5%, 37.1% and 56.9% as complete rate, respectively.

2.3 Numerical variables

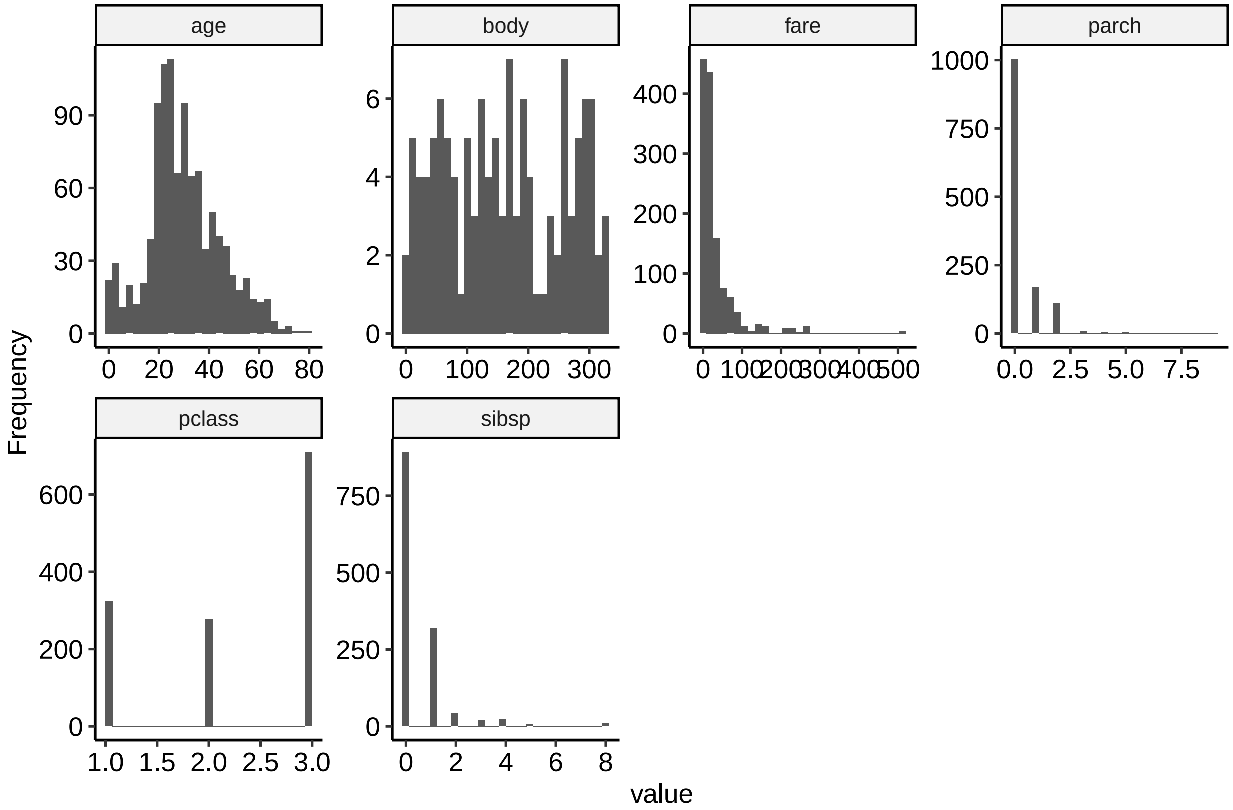
General statistics about numerical variables from Titanic dataset:

```
skimr::partition(skimmed_titanic)$numeric %>%
        knitr::kable(format = 'latex', booktabs = TRUE, digits = 2) %>%
        kableExtra::kable_styling(latex_options = 'HOLD_position')
```

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
pclass	0	1.00	2.29	0.84	1.00	2.0	3.00	3.00	3.00	
age	263	0.80	29.88	14.41	0.17	21.0	28.00	39.00	80.00	
sibsp	0	1.00	0.50	1.04	0.00	0.0	0.00	1.00	8.00	
parch	0	1.00	0.39	0.87	0.00	0.0	0.00	0.00	9.00	
fare	1	1.00	33.30	51.76	0.00	7.9	14.45	31.27	512.33	
body	1188	0.09	160.81	97.70	1.00	72.0	155.00	256.00	328.00	

Next, we plot histograms for the numerical features to get to know better their distributions.

```
titanic_numerical <- titanic %>% select(where(is.numeric))
DataExplorer::plot_histogram(titanic_numerical, ggtheme = ggpubr::theme_pubr(base_size = 10))
```



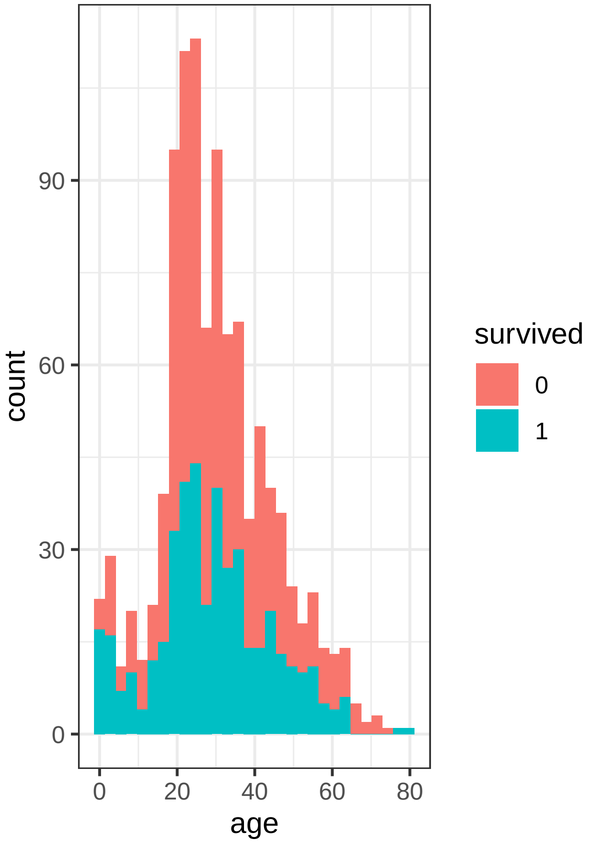
There are 6 numerical features in this dataset, i.e. `pclass`, `age`, `sibsp`, `parch`, `fare`, and `body`. Except for `body` with extremely low complete rate (only 9%), the other variables have acceptable complete rate where the worst is feature `age` with 80%. Regarding the reason for the high number of missing values for `body`, the reasons can be that at least 509 of survivors are in this list and also the remaining passengers were still not found.

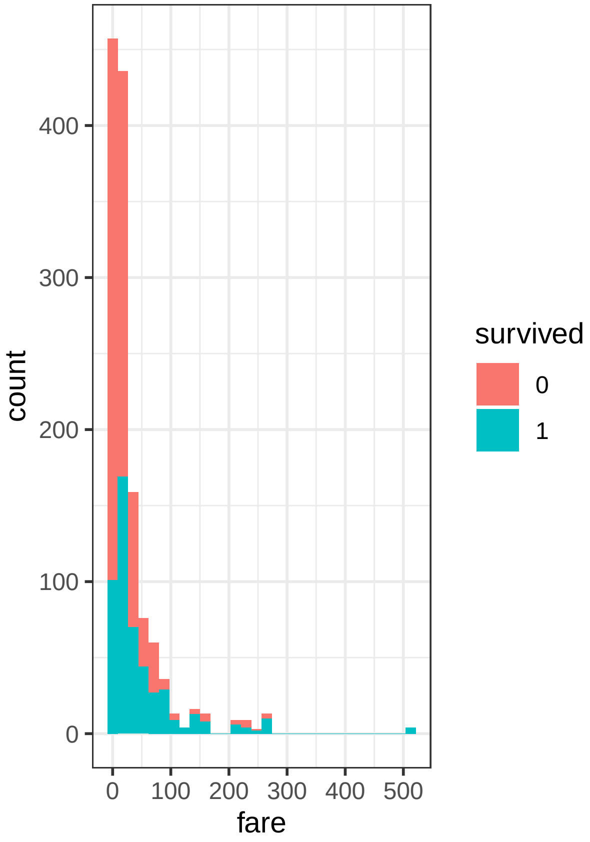
We are specifically interested in `age` and `fare` to see if they have any impact on survival rate. We can uncover that by using the segmented histograms:

```
p_age <- titanic %>%
        select(age, survived) %>%
        ggplot(mapping = aes(fill = survived, x = age)) + geom_histogram()

p_fare <- titanic %>%
        select(fare, survived) %>%
        ggplot(mapping = aes(fill = survived, x = fare)) + geom_histogram()

grid.arrange(p_age, p_fare, ncol = 2)
```

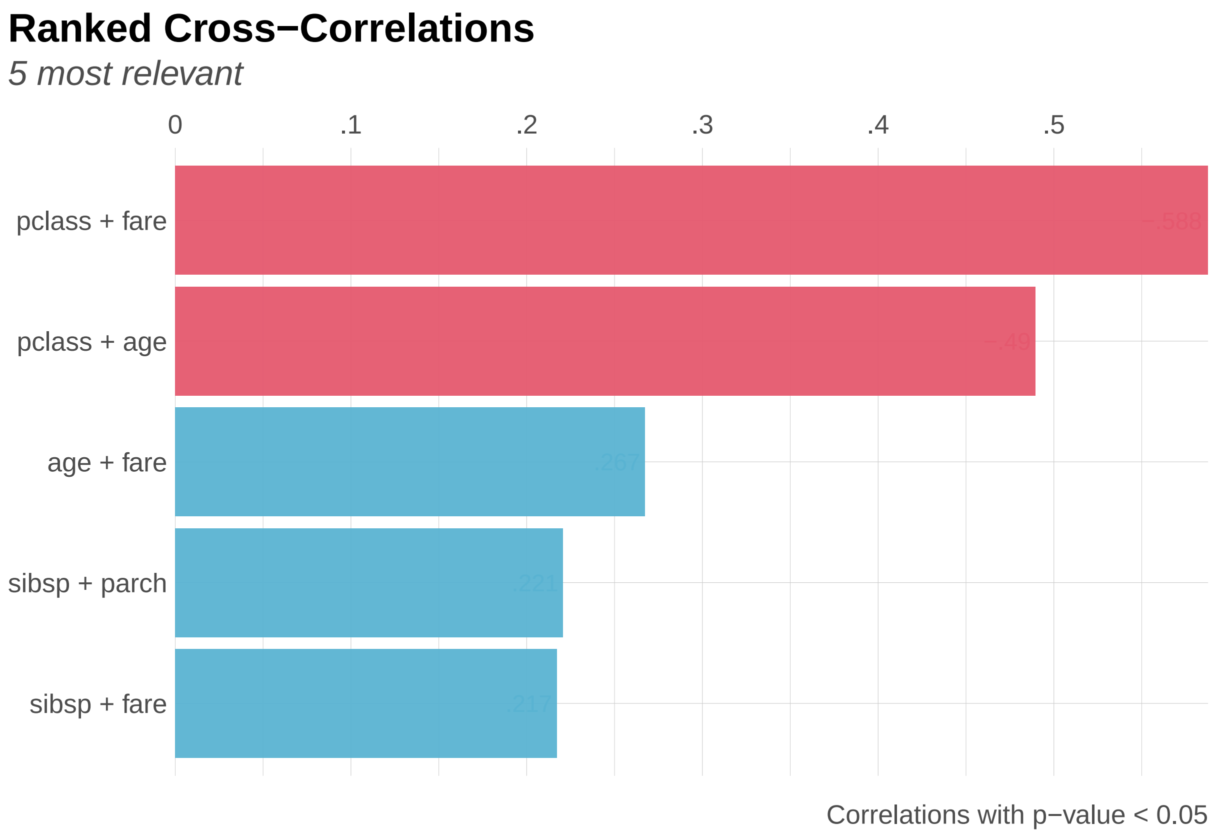




Looking at the histograms, we can see that smaller age (< 15) may increase the chance of surviving (maybe because they prioritized to save younger children first). Furthermore, the higher the fare, the better chance of getting out of the Titanic's wreckage alive. It can be due to the priority of the VIP passengers and the locations of high-class seats that helped them evacuate more easily.

After looking at the distributions, we may also look at the correlation between pairs of numerical variables:

```
corr_cross(drop_na(titanic_numerical),
  max_pvalue = 0.05, # display only significant correlations (at 5% level)
  top = 20 # display top 20 couples of variables (by correlation coefficient)
)
```

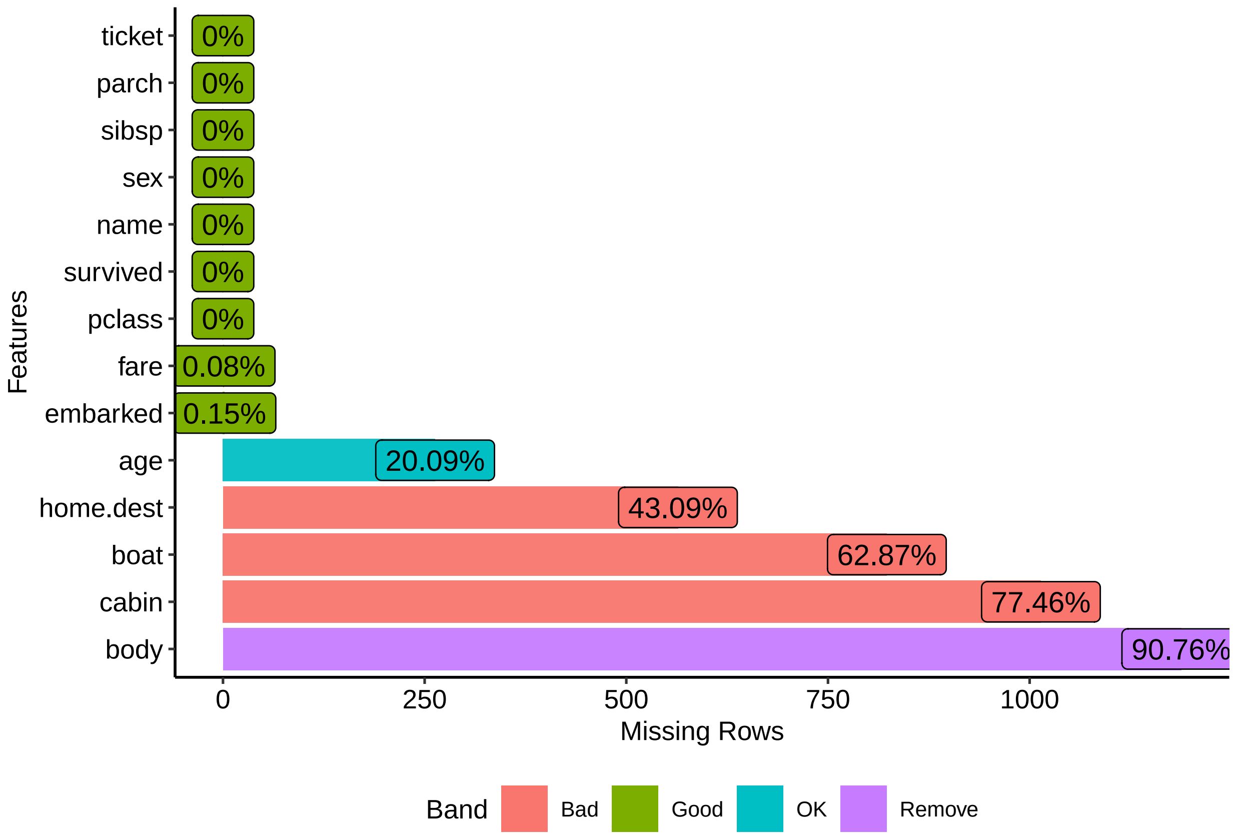


From this plot, it can be seen that there is a fairly high negative correlation between `pclass` and `fare` (roughly 0.6). This makes much sense as the higher the fare, the higher the ticket class (1 is the first class). Another notable correlation is between `pclass` and `age`, these two features are also negatively correlated. Maybe it is the case that older passengers had more money to spend on higher class tickets.

2.4 Missing data

Seems we have quite some missing observations. Let's take a closer look:

```r
DataExplorer::plot_missing(titanic, ggtheme = ggpubr::theme_pubr(base_size = 10))
```



As can be seen, `age` has acceptable missing rate. However, `home.dest`, `boat`, `cabin`, and especially `body` have too high missing rate. The other variables have very low missing rate or don't have missing values at all.

3 Data preprocessing notes

In this section, we present a few notes that can be beneficial for preprocessing the data.

3.1 Data quality assessment

From the EDA, we can see that this dataset has a lot of missing data depending on the features. Apart from that, the data seems to be well-formatted.

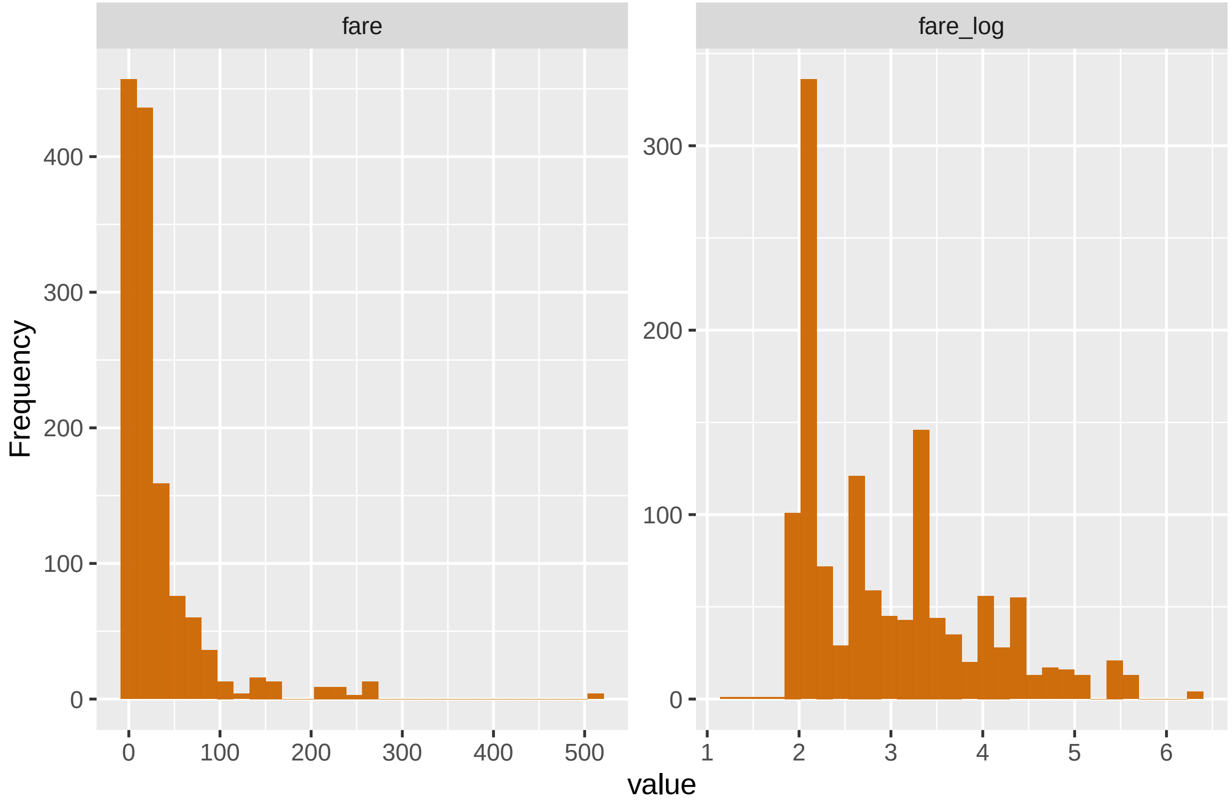
3.2 Data cleaning

Missing data needs to be dealt with in this dataset as we have a ton of them. For example, with feature `age`, we can consider filling in by using the median of age per `pclass` group (due to the observation that `pclass` has fairly high correlation with `age`).

3.3 Data transformation

The feature `fare` is highly right skewed, so it can be helpful for some models to apply `log` transformation on it.

```
fare_log <- titanic_numerical %>% mutate(fare_log = log(fare))
DataExplorer::plot_histogram(fare_log %>% select(fare, fare_log))
```


fare
fare_log
Frequency
value