Titanic

Thiago Pires

Introduction

This article presents some analyzes with information about the passengers of the famous ship Titanic that tragically wrecked in 1912.

Read data

```r
require(dplyr)
require(magrittr)
require(titanic)
train <- titanic_train
```

About the attributes

The variables used are:

- Sex: sex
- Age: age
- Pclass: Passenger Class
- Survived: Passenger Survival Indicator

Initial plan for data exploration

Univariate investigation to lead after a multiple analysis. I will use plots and supervised modeling in the analysis.

Manipulation (Actions taken for data cleaning and feature engineering)

```r
# tables and figures numbers
require(captioner)
fig_nums <- captioner(prefix = "Figure")
table_nums <- captioner(prefix = "Table")
# character to factor
train %<>%
  mutate(Survived = factor(x = Survived, labels = c("no", "yes")),
         Pclass = factor(x = Pclass, labels = c("1st", "2nd", "3rd")),
         Sex = factor(x = Sex))
```

Univariate analysis

My three hypothesis about this data are:

- Relation between sex and survived
- Relation between pclass and survived
- Relation between age and survived

Ticket class (Pclass)

There were more survivors in first class than in the second and third class (Figure 1).

```
require(ggplot2)
train %>% count(Pclass, Survived) %>%
  group_by(Pclass) %>%
  mutate(total = sum(n)) %>%
  ggplot(aes(Pclass, n * 100/total, fill = Survived)) + geom_col() +
  theme_light() +
  labs(x = "Ticket class", y = "%", fill = "Survived",
       caption = fig_nums("pclass", "Percentual distribution of survivors according to ticket class"))
```

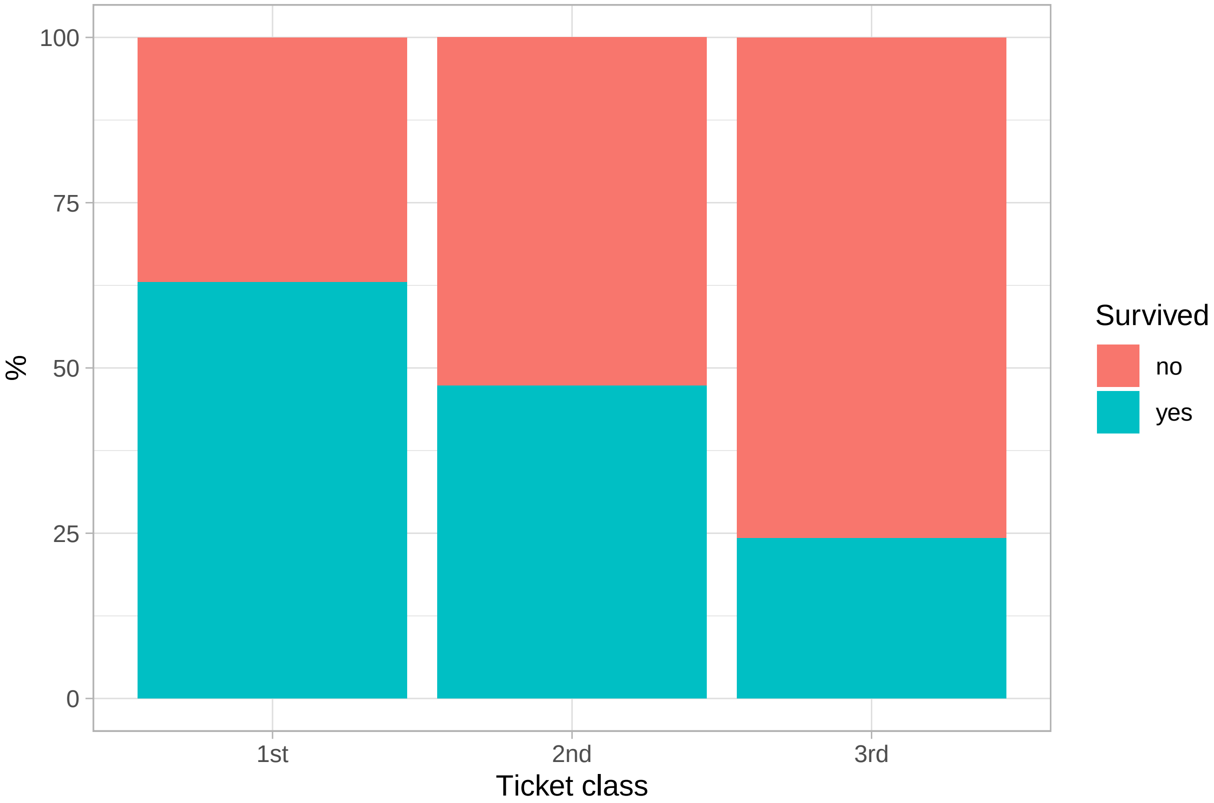

Figure 1: Percentual distribution of survivors according to ticket class

The model shows the odds to survive in the third class is 81.18 (`1 - exp(-1.67039)`) percent lower than in the first class (Table 1).

```
require(knitr)
require(kableExtra)
model <-
  glm(Survived ~ Pclass, family = binomial, data = train)
model %>%
  summary() %$% coefficients %>%
  kable(caption = table_nums("pclass", "Model estimation"))
```

Sex

There were more female survivors than male (Figure 2).

Table 1: Table 1: Model estimation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.5306283	0.1409005	3.765979	0.0001659
Pclass2nd	-0.6394311	0.2040992	-3.132943	0.0017306
Pclass3rd	-1.6703986	0.1759104	-9.495738	0.0000000

```
train %>% count(Sex, Survived) %>%
  group_by(Sex) %>%
  mutate(total = sum(n)) %>%
  ggplot(aes(Sex, n * 100/total, fill = Survived)) + geom_col() +
  theme_light() +
  labs(x = "Sex", y = "%", fill = "Survived",
       caption = fig_nums("sex", "Percentual distribution of survivors according to sex"))
```

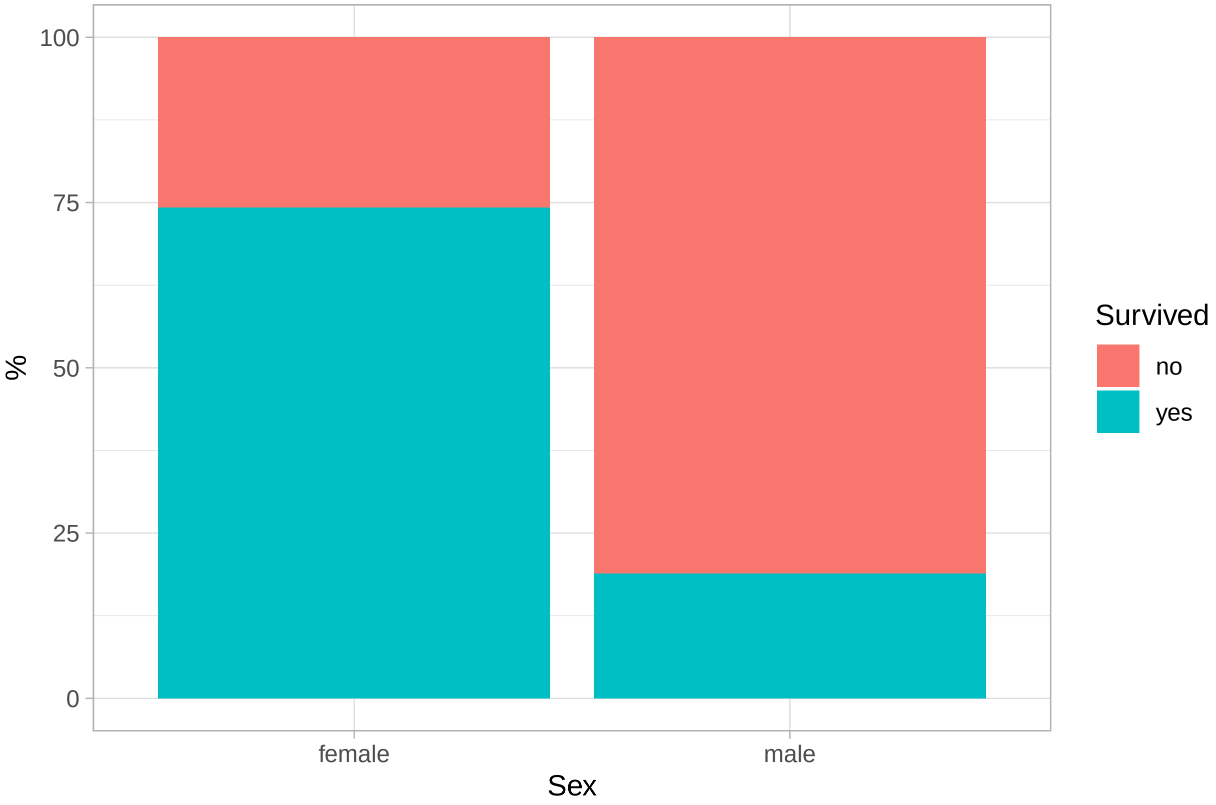

Figure 2: Percentual distribution of survivors according to sex

The model shows the odds to survive for male is 91.9 (`1 - exp(-2.513710)`) percent lower than for female (Table 2).

```
model <-
  glm(Survived ~ Sex, family = binomial, data = train)
model %>%
  summary() %$% coefficients %>%
  kable(caption = table_nums("sex", "Estimativas do modelo"))
```

Table 2: Table 2: Estimativas do modelo

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.056589	0.1289864	8.191477	0
Sexmale	-2.513710	0.1671782	-15.036107	0

Table 3: Table 3: Model estimation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.4122472	0.5867893	5.815115	0.0000000
Pclass2nd	-0.9555114	0.7247579	-1.318387	0.1873741
Pclass3rd	-3.4122472	0.6099995	-5.593852	0.0000000
Sexmale	-3.9493901	0.6160608	-6.410715	0.0000000
Pclass2nd:Sexmale	-0.1849918	0.7939117	-0.233013	0.8157513
Pclass3rd:Sexmale	2.0957553	0.6572051	3.188891	0.0014282

Interaction between Sex and Ticket Class

The interaction between sex and ticket class was significant (Table 3), showing that the differences between the probability of survival between men and women is greater when the ticket class improves (Figure 3).

```
model <-
  glm(Survived ~ Pclass*Sex, family = binomial, data = train)
model %>%
  summary() %$% coefficients %>%
  kable(caption = table_nums("pclass_sex", "Model estimation"))
```

```
# table with all possibilities
newdata <-
  expand.grid(Pclass = c("1st", "2nd", "3rd"),
              Sex = c("male", "female")) %>% as_tibble()
# predictions (probability)
newdata %<>%
  mutate(Pihat = model %>% predict(newdata = newdata, type = "response"))
newdata %>%
  ggplot(aes(Sex, Pihat, group = Pclass, colour = Pclass)) + geom_line() + geom_point() +
  theme_light() +
  labs(x = "Sex", y = expression(pi(Survived == yes)), colour = "Ticket Class",
       caption = fig_nums("pclass_sex", "Interaction between sex and ticket class"))
```

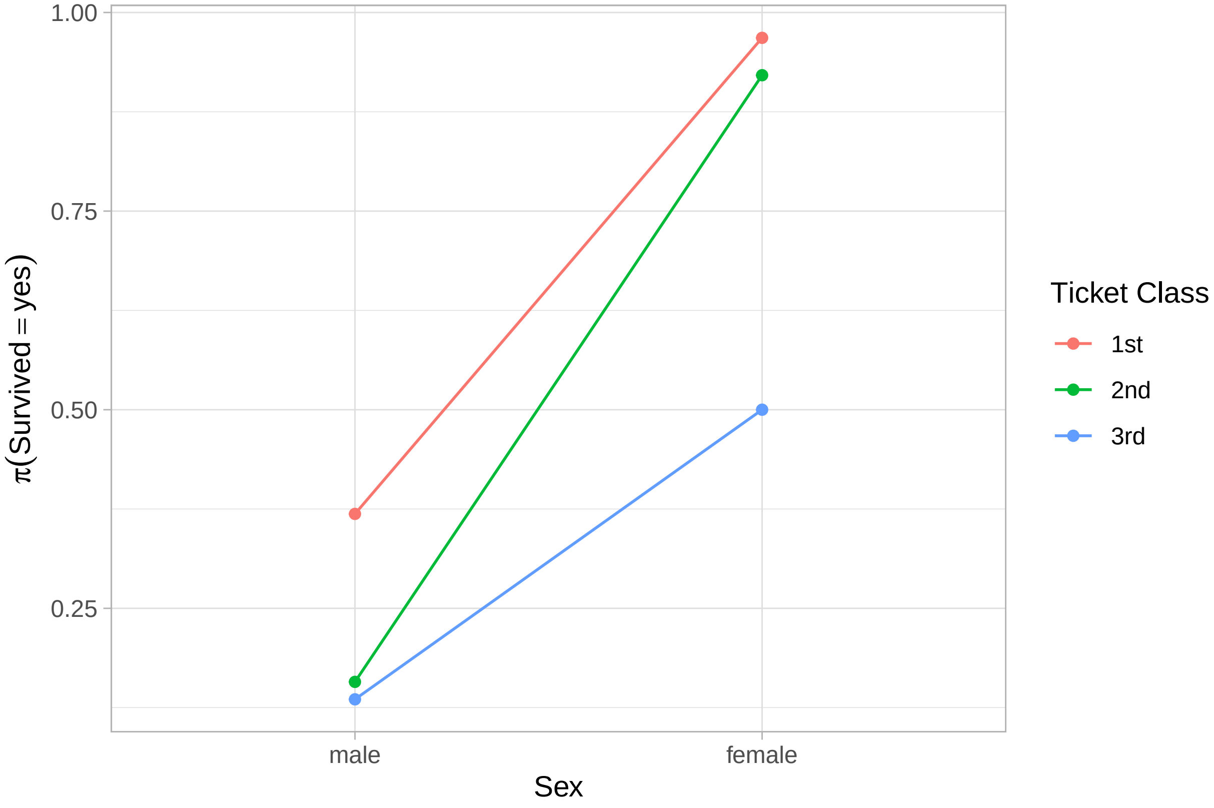

Figure 3: Interaction between sex and ticket class

Age

There is no difference between the distribution of age according to survived status (Figure 4).

```
train %>%
  ggplot(aes(Survived, Age)) + geom_boxplot() +
  theme_light() +
  labs(caption = fig_nums("age", "Distribution of age according to survived status"))
```

Table 4: Table 4: Model estimation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.0567236	0.1735804	-0.3267859	0.7438298
Age	-0.0109635	0.0053299	-2.0569560	0.0396905

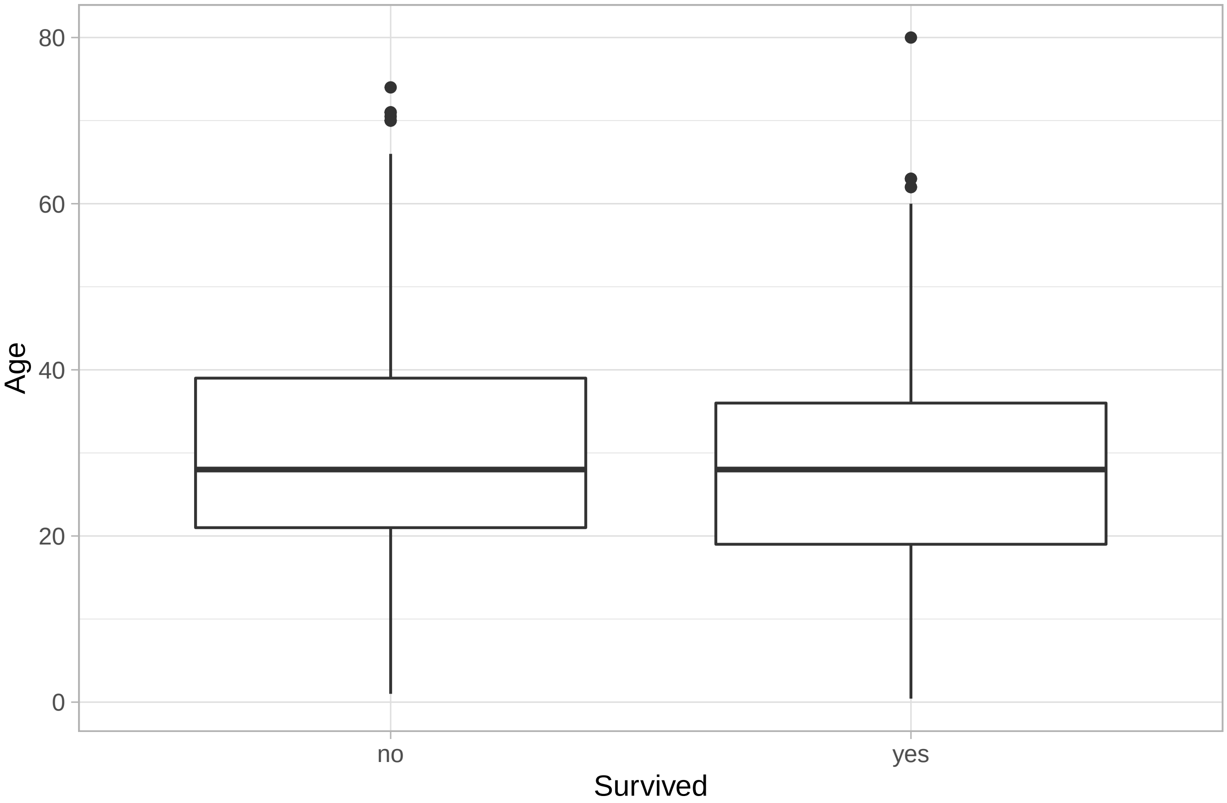

Figure 4: Distribution of age according to survived status

Fitting a simple model is showed a little age effect on survival probability (Table 4). The increase of one year in age decreases the odds of survival by 1.09 (`1 - exp(-0.0109635)`) percent.

```
model <-
  glm(Survived ~ Age, family = binomial, data = train)
model %>%
  summary() %$% coefficients %>%
  kable(caption = table_nums("age", "Model estimation"))
```

Interaction between Age and Ticket Class

Below, note that there is no interaction between age and ticket class (p-values > .05)

```
model <-
  glm(Survived ~ Pclass*Age, family = binomial, data = train)
model %>%
  summary() %$% coefficients %>%
  kable(caption = table_nums("pclass_age", "Model estimation"))
```

Table 5: Table 5: Model estimation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.2425170	0.4912807	4.5646345	0.0000050
Pclass2nd	-1.0532520	0.6325989	-1.6649604	0.0959207
Pclass3rd	-2.4071609	0.5648872	-4.2613124	0.0000203
Age	-0.0404384	0.0114339	-3.5367100	0.0004051
Pclass2nd:Age	-0.0023603	0.0168729	-0.1398882	0.8887483
Pclass3rd:Age	-0.0017200	0.0160496	-0.1071703	0.9146539

Suggestions for next steps in analyzing this data

- Compare the logistic regression with another algorithms
- Chose another variables to input in the model

The variables choose there were a good quality, but age that was some missing (N = 177). To age could apply a missing imputation, this approach could be a next step to analyse too.