Survival Analysis Project

Project outline:

- Introduction
- Prepare data
- Short EDA
- Perform analysis with multiple methods (4-5-6)
- Chose 2-3
- Chose what to keep (maximum 9 pages)

Introduction

In terms of survival analysis, the objective of this project is to determine the "risk" that pets in an animal shelter have of being adopted (or even duration in the shelter), based on different characteristics that are available in the chosen data set. It is worth mentioning that the data contains different outcomes for the animals: adoption, transfer, or no outcome. However, in our case we will only use a dichotomous output indicating whether it was adopted or not. The data ends on February 2, 2018, after which date it is not known what happens to the animals.

We will also analyze how the following variables, when possible to include them, may influence the adoption or non-adoption of the animals.

Data dictionary:

- id: unique id for each animal
- age_upon_outcome: age of the animal when the outcome was determined
- animal_type: cat, dog, or … something else
- breed: breed of the animal
- color: color of the animal
- date_of_birth: date of birth of the animal
- datetime: date and time when the outcome was determined
- name: name of the animal

- outcome_type: there are three possible outcomes: adoption, transfer, no outcome (euthanized, died)
- sex: sex of the animal
- spay_neuter: whether the animal was spayed or neutered: intact or fixed

Data preparation

Load necessary libraries and data

```r
# Packages for data preparation
library(dplyr)
library(gtsummary)
library(readr)
library(skimr)
library(gt)

# Packages for analyzing survival data
library(survival)
library(survminer)

raw_data <- read_csv("data/train.csv")
```

EDA

At a quick glance at the data, we observe that in the case of categorical variables we will have to make modifications in order to be able to work with them, for example in the case of breed, color and name we have too many different types or levels, which would be impossible to use them all. The variable name, on the other hand, has a significant number of null values.

Table 1: Data summary

Name	raw_data
Number of rows	54408
Number of columns	11
Column type frequency:	
character	8
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
age_upon_outcome	0	1.0	4	9	0	46	0
animal_type	0	1.0	3	9	0	5	0
breed	0	1.0	3	54	0	1812	0
color	0	1.0	3	27	0	475	0
name	16433	0.7	1	12	0	11826	0
outcome_type	0	1.0	8	10	0	3	0
sex	0	1.0	4	7	0	3	0
spay_neuter	0	1.0	5	7	0	3	0

In order to be able to work with this data, these will be the first modifications we will make:

- Since the variables `breed` and `color` have too many different combinations, we will remove them from our data, they will not be useful for our models.
- We have names for 69.8% of the animals, so in order to use it, we will create a dummy variable `has_name`, a variable that indicates that if there is a name or not.
- Also, we decided to treat the `animal_type` and `sex` variables as factors, so they can enter the model.
- For the duration data, we have the `age_upon_outcome` which is the difference between `date_of_birth` and `datetime` but this is not numeric, so we create our own variable `time_to_outcome`.
- We will work with `spay_neuter` as our grouping variable. We would like to verify if there are differences in adoption times if the animal is spay neutered or not. If we have an `Unknown` in the variable, we will encode it as `Intact`, as knowing for sure how the animal has been handled is important.
- We keep `animal_type`, `sex` and `spay_neuter` as they are, the only modification in this case will be that for `animal_type`, since there are only 5 observations in Livestock, we will join this category to Other. In summary we have a majority of dogs and cats, in sex variable 44% of female, 48% of male and 8% classified as Unknown. In the case of spaying and neutering, we have 68% fixed, 24% intact and 8% classified as Unknown.

```r
raw_data <- raw_data |>
  mutate(has_name = !is.na(name), # If not NA, then it has a name.
         sex = as.factor(sex),
         time_to_outcome = as.Date(datetime) - date_of_birth,
         spay_neuter = as.factor(if_else(spay_neuter == "Unknown", "Intact", spay_neuter))
         animal_type = as.factor(if_else(animal_type == "Livestock","Other", animal_type))
```

```
           )
```

In the case of the output variable, where the possibilities are adopted, unadopted or unknown, we will create a new variable `outcome` for which we take `1` for adoption and `0` for no adoption or unknown.

outcome	n	prop
0	21133	38.8
1	33275	61.2

Finally, since we got some negative values for our variable `time_to_outcome` , which makes no sense as the `date_of_birth` should come before the outcome, we will remove these five observations (5).

Now we chose the variable we will use as a clean data , check that they have the desired formats and do some previous analysis on the variables.

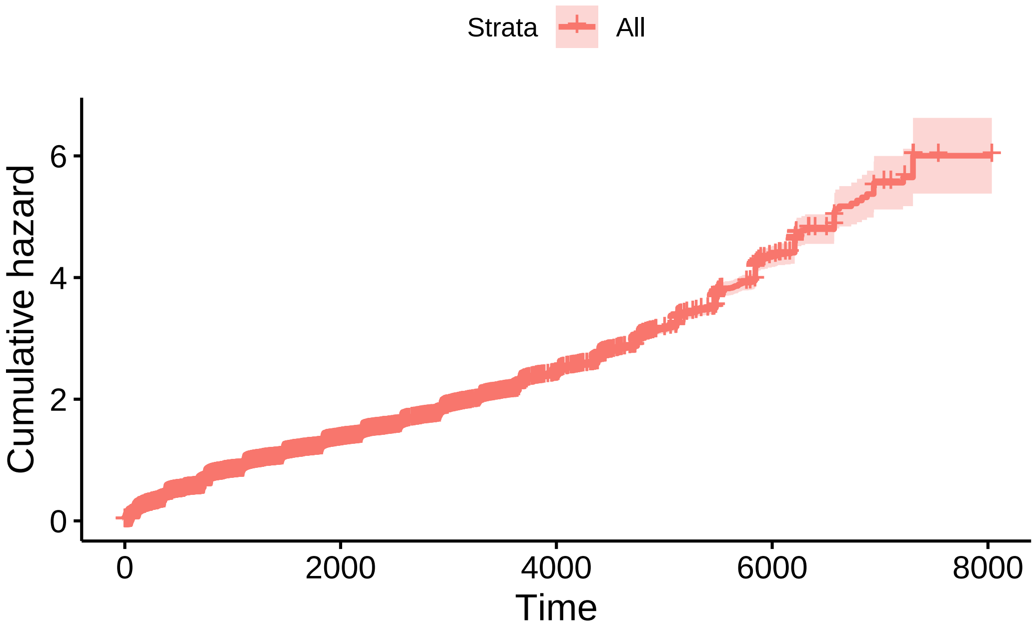
```
Rows: 54,403
Columns: 7
$ id              <chr> "1265", "24053", "4785", "65439", "45732", "38636", "5~
$ time_to_outcome <drtn> 837 days, 366 days, 63 days, 738 days, 555 days, 736 ~
$ animal_type     <fct> Cat, Other, Dog, Cat, Dog, Dog, Cat, Dog, Dog, Dog, Ca~
$ has_name        <lgl> TRUE, FALSE, FALSE, TRUE, TRUE, TRUE, FALSE, TRUE, TRU~
$ sex             <fct> Male, Unknown, Female, Female, Female, Male, Male, Mal~
$ spay_neuter      <fct> Fixed, Intact, Fixed, Fixed, Fixed, Fixed, Fixed, Fixe~
$ outcome         <dbl> 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, ~
```

	0	1	**Total**
spay__neuter			
Fixed	7,146 (19%)	29,865 (81%)	37,011 (100%)
Intact	13,984 (80%)	3,408 (20%)	17,392 (100%)
Total	21,130 (39%)	33,273 (61%)	54,403 (100%)

Within the animals, there is not a difference that seems to be important according to sex, at least not proportionally. In the case of animals with names, more than 70% have been adopted against 26% in the case of those without names. Something similar happens with the animals that have been spayed or neutered, the proportion of adopted animals in this case exceeds 80%, while in the group of those that are not known or have not undergone surgery, the percentage of adopted animals is barely 20%. Lastly, dogs are not only the largest number of animals but also have the highest adoption rate, while among cats and birds there is practically half and half between adoptees and non-adoptees.[1]

Statistical Analysis

The next graph is the survival curve and the cumulative hazard function, for this data:



Example options given by instructions:

- nonparametric estimation of survival for one or more groups
- nonparametric comparison of 2 or more groups
- semi-parametric Cox regression

Nonparametric methods for censored data

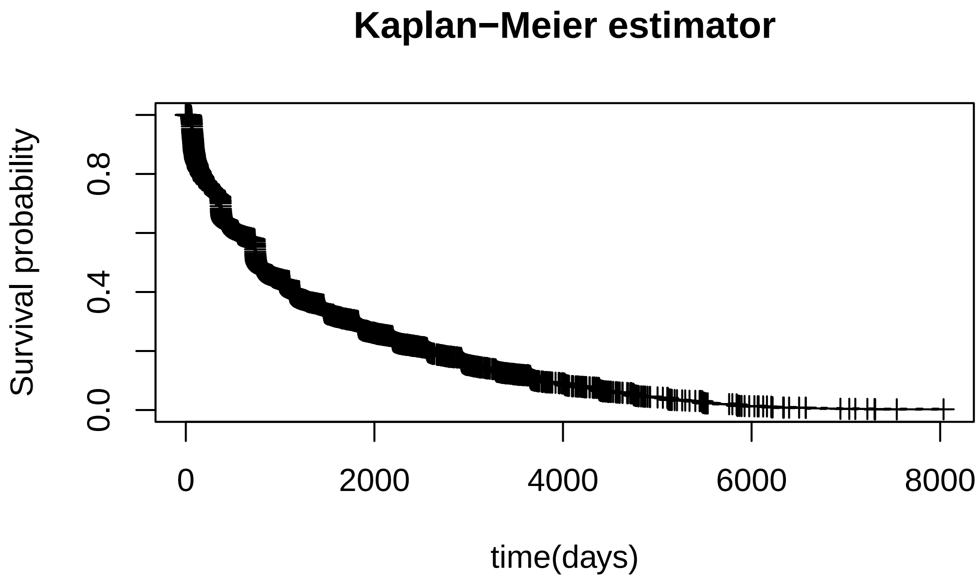
Using the Kaplan-Meyer estimator (KM) methodology we can see that the median survival time is 749 days (in this particular case the time from birth to adoption, and it's about 2 years). If we estimate the probability of not being adopted when the animal is (at least) one year old, we see that this value is 71.9%.

[1] The cross tables for the variables `sex`, `animal_type` and `has_name`, see appendix 1.

```
KM <- survfit(Surv(time_to_outcome, outcome) ~ 1 , data = data[,-1])
KM
```

```
Call: survfit(formula = Surv(time_to_outcome, outcome) ~ 1, data = data[,
    -1])

         n events median 0.95LCL 0.95UCL
[1,] 54403  33273    749     745     754
```

```
plot(KM, mark.time = TRUE, main = "Kaplan-Meier estimator",
     ylab = 'Survival probability',
     xlab = 'time(days)')
```



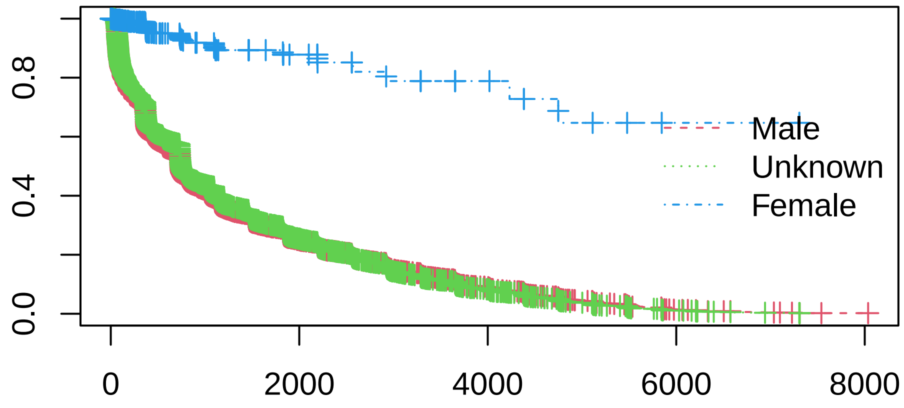
```
#estimated probability of not being adopted for (at least) 1 year
summary(KM, time = 365)
```

```
Call: survfit(formula = Surv(time_to_outcome, outcome) ~ 1, data = data[,
    -1])

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  365  31417   13167    0.719 0.00208        0.715        0.723
```

Nonparametric comparison of groups

Even if we do not have two groups in the dataset, for example with different treatments, we want to use this technique to analyze, for example, whether or not being part of the females group contributes to being adopted more quickly. So, the null hypothesis or question that we are going to try to answer is if the survival curves generated for the groups are the same, and the alternative hypothesis is that they are different. We will perform the logrank test.

```
Call: survfit(formula = Surv(time_to_outcome, outcome) ~ sex, data = data)

                n events median 0.95LCL 0.95UCL
sex=Female  23929  15767    738     737     741
sex=Male    26043  17410    743     741     748
sex=Unknown  4431     96     NA    4748      NA
```



```
Call:
survdiff(formula = Surv(time_to_outcome, outcome) ~ sex, data = data)

                N Observed Expected (O-E)^2/E (O-E)^2/V
sex=Female  23929    15767    15100      29.4      54.2
sex=Male    26043    17410    16985      10.6      21.8
sex=Unknown  4431       96     1187    1003.1    1061.5
```

```
 Chisq= 1064  on 2 degrees of freedom, p= <2e-16 
```

As we can see both in the graph and in the logrank test, there is statistical evidence to reject the null hypothesis of equality between curves, and this is due to the fact that although there do not seem to be differences between male and female, in those animals for which the sex is not determined, there do seem to be differences in the time from birth to adoption.

In the `has_name` we get also a difference between the curves and it looks like the adoption goes faster for those animal that don't have a name at the moment of adoption. We can perform the same process with other categorical variables[2].

Multivariate Cox regression

```r
cox_model <- coxph(Surv(time_to_outcome, outcome) ~ animal_type + has_name + sex + spay_ne
summary(cox_model)
```

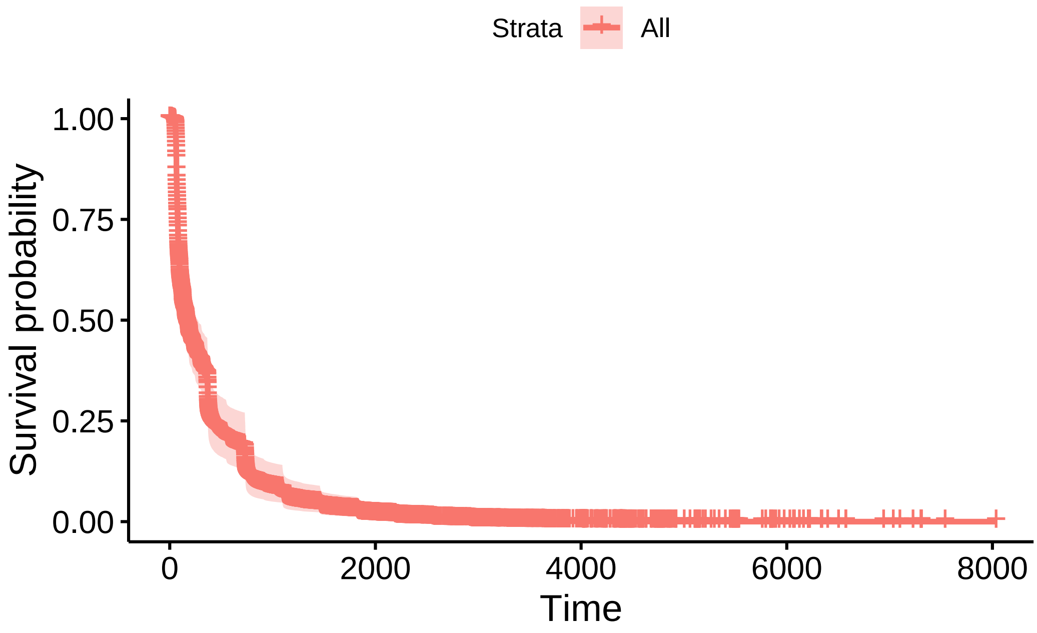
```
Call:
coxph(formula = Surv(time_to_outcome, outcome) ~ animal_type + 
    has_name + sex + spay_neuter, data = data)

  n= 54403, number of events= 33273 

                      coef exp(coef) se(coef)       z Pr(>|z|)    
animal_typeCat    -0.73413   0.47992  0.11031  -6.655 2.82e-11 ***
animal_typeDog    -0.94597   0.38830  0.11013  -8.590  < 2e-16 ***
animal_typeOther  -1.33023   0.26442  0.13528  -9.833  < 2e-16 ***
has_nameTRUE      -0.09051   0.91347  0.01751  -5.168 2.37e-07 ***
sexMale           -0.01344   0.98665  0.01102  -1.220    0.223    
sexUnknown        -1.86478   0.15493  0.11532 -16.170  < 2e-16 ***
spay_neuterIntact -0.67203   0.51067  0.01905 -35.282  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
animal_typeCat       0.4799      2.084    0.3866    0.5957
animal_typeDog       0.3883      2.575    0.3129    0.4819
animal_typeOther     0.2644      3.782    0.2028    0.3447
has_nameTRUE         0.9135      1.095    0.8826    0.9454
```

[2] for further information look at Appendix 2

```
sexMale                 0.9866      1.014     0.9656     1.0082
sexUnknown              0.1549      6.455     0.1236     0.1942
spay_neuterIntact       0.5107      1.958     0.4920     0.5301

Concordance= 0.629  (se = 0.002 )
Likelihood ratio test= 3686  on 7 df,   p=<2e-16
Wald test            = 2285  on 7 df,   p=<2e-16
Score (logrank) test = 2804  on 7 df,   p=<2e-16
```

Let's see our predicted survival proportion for the whole data.

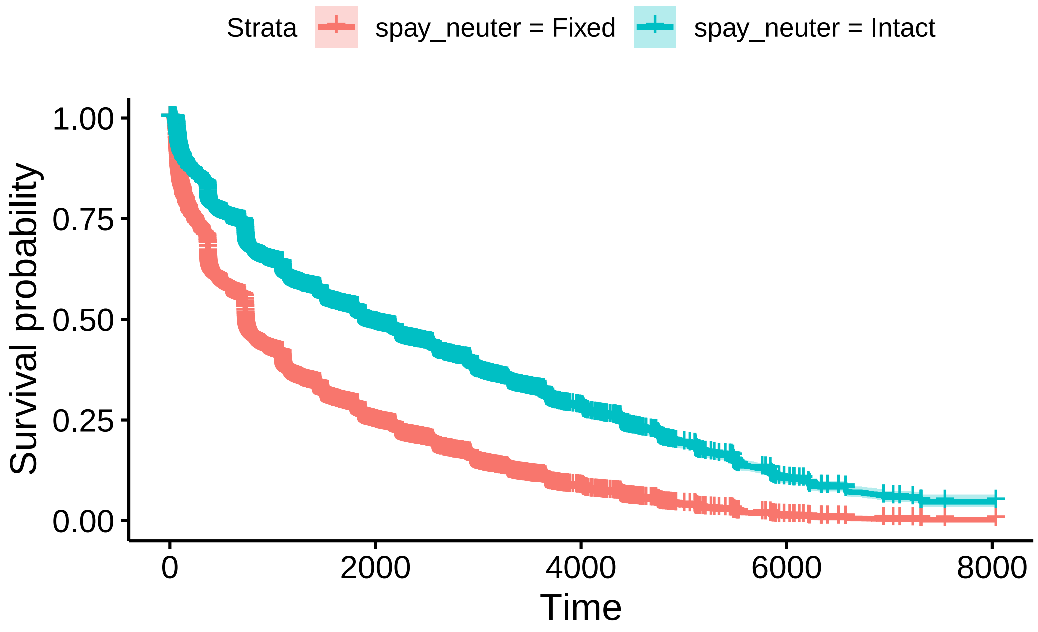
```
ggsurvplot(survfit(cox_model), data = data)
```



Now, we can verify how this survival changes depending on the `spay_neuter` variable.

```
# Prediction
spay_neutered_data <- tibble(animal_type = c('Dog', 'Dog'),
                             has_name = c(TRUE,TRUE),
                             sex = c("Male", "Male"),
                             spay_neuter = c('Fixed', 'Intact'))
```

```
# Fit a prediction
fit <- survfit(cox_model, newdata = spay_neutered_data)

ggsurvplot(fit, data = data, conf.int = TRUE,
            legend.labs = c("spay_neuter = Fixed", "spay_neuter = Intact"))
```



As expected, animal which were spay neutered, have shorter survival time, that means they were adopted faster.

Appendix

Appendix 1: cross tables

```
#TODO: add tables names
tbl_cross(data, row = has_name, col = outcome,percent = c("row"), digits = 0)|>
  bold_labels()
```

	0	1	**Total**
has_name			
FALSE	12,228 (74%)	4,203 (26%)	16,431 (100%)
TRUE	8,902 (23%)	29,070 (77%)	37,972 (100%)
Total	21,130 (39%)	33,273 (61%)	54,403 (100%)

```
tbl_cross(data, row = sex, col = outcome,percent = c("col"), digits = 0)|>
  bold_labels()
```

	0	1	**Total**
sex			
Female	8,162 (39%)	15,767 (47%)	23,929 (44%)
Male	8,633 (41%)	17,410 (52%)	26,043 (48%)
Unknown	4,335 (21%)	96 (0%)	4,431 (8%)
Total	21,130 (100%)	33,273 (100%)	54,403 (100%)

```
tbl_cross(data, row = animal_type, col = outcome,percent = c("row"), digits = 0)|>
  bold_labels()
```

	0	1	**Total**
animal_type			
Bird	126 (59%)	86 (41%)	212 (100%)
Cat	10,582 (51%)	9,975 (49%)	20,557 (100%)
Dog	7,785 (25%)	23,044 (75%)	30,829 (100%)
Other	2,637 (94%)	168 (6%)	2,805 (100%)
Total	21,130 (39%)	33,273 (61%)	54,403 (100%)

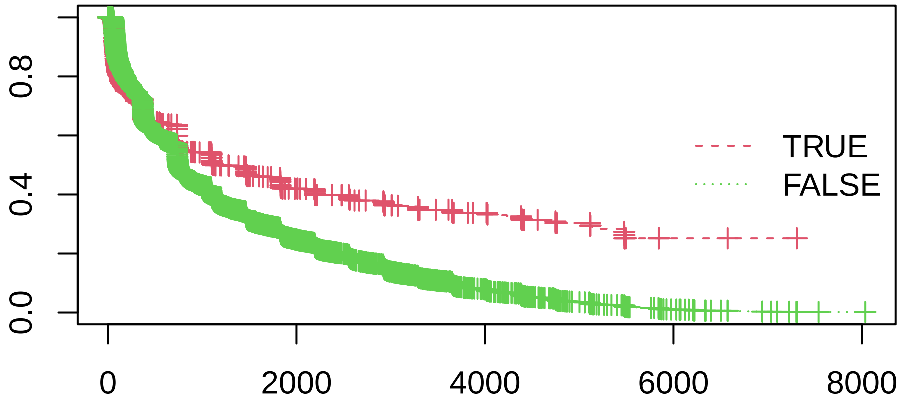
Appendix 2: Nonparametric comparison of groups

Variable `has_name`

```
has_name_comp <- survfit(Surv(time_to_outcome, outcome) ~ has_name, data = data)

has_name_comp
```

```
Call: survfit(formula = Surv(time_to_outcome, outcome) ~ has_name,
    data = data)

                   n events median 0.95LCL 0.95UCL
has_name=FALSE 16431   4203   1125    1101    1465
has_name=TRUE  37972  29070    744     741     747
```

```
plot(has_name_comp,mark.time = T, lty = 2:3,col=2:3)
legend("right", legend=unique(data$has_name), col=2:3, lty=2:3, horiz=FALSE,  bty='n')
```



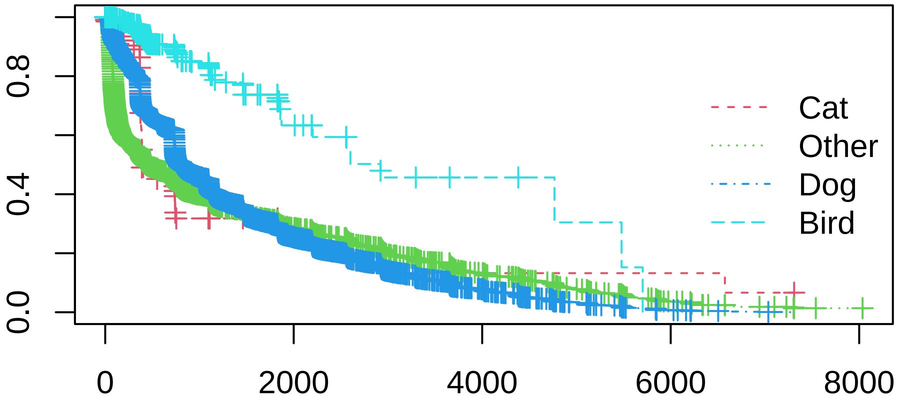
Variable `animal_type`

```
type_comp <- survfit(Surv(time_to_outcome, outcome) ~ animal_type, data = data)
type_comp
```

```
Call: survfit(formula = Surv(time_to_outcome, outcome) ~ animal_type,
    data = data)

                      n events median 0.95LCL 0.95UCL
animal_type=Bird    212     86    388     388     740
animal_type=Cat   20557   9975    436     410     491
animal_type=Dog   30829  23044    782     770     797
```

```
animal_type=Other  2805    168   2922    2202      NA
```

```
plot(type_comp,mark.time = T, lty = 2:6,col=2:6)
legend("right", legend=unique(data$animal_type), col=2:6, lty=2:6, horiz=FALSE,  bty='n')
```



```
# The logrank test
survdiff(Surv(time_to_outcome, outcome) ~ animal_type, data = data)
```

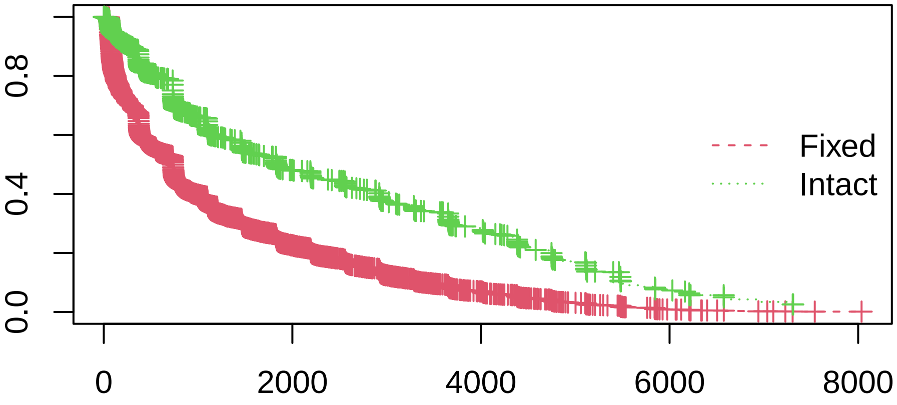
```
Call:
survdiff(formula = Surv(time_to_outcome, outcome) ~ animal_type,
    data = data)

                      N Observed Expected (O-E)^2/E (O-E)^2/V
animal_type=Bird    212       86     92.7     0.483     0.488
animal_type=Cat   20557     9975   8378.3   304.280   413.016
animal_type=Dog   30829    23044  23706.0    18.488    65.587
animal_type=Other  2805      168   1096.0   785.715   829.464

 Chisq= 1127  on 3 degrees of freedom, p= <2e-16
```

Variable spay_neuter

```
spay_neuter_comp <- survfit(Surv(time_to_outcome, outcome) ~ spay_neuter, data = data)
spay_neuter_comp
```

```
Call: survfit(formula = Surv(time_to_outcome, outcome) ~ spay_neuter,
    data = data)

                        n events median 0.95LCL 0.95UCL
spay_neuter=Fixed   37011  29865    735     734     735
spay_neuter=Intact 17392   3408   1831    1827    1898
```

```
plot(spay_neuter_comp,mark.time = T, lty = 2:3,col=2:3)
legend("right", legend=unique(data$spay_neuter), col=2:3, lty=2:3, horiz=FALSE,  bty='n')
```



```
# The logrank test
survdiff(Surv(time_to_outcome, outcome) ~ spay_neuter, data = data)
```

```
Call:
survdiff(formula = Surv(time_to_outcome, outcome) ~ spay_neuter,
    data = data)
```

```
                        N Observed Expected (O-E)^2/E (O-E)^2/V
spay_neuter=Fixed  37011    29865    26499       427      2131
spay_neuter=Intact 17392     3408     6774      1672      2131

 Chisq= 2131  on 1 degrees of freedom, p= <2e-16
```