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)
  - What code do we keep? Should we show all the code?
  - How much of the EDA is finally important?
  - Which methods with which plots are we keeping?

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, based on different characteristics that are available in the dataset. It is worth mentioning that the data contain 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.

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

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

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

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

EDA

```
raw_data |>
  summary()
```

```
      id        age_upon_outcome   animal_type           breed
Min.   :    1   Length:54408       Length:54408       Length:54408
1st Qu.:19483   Class :character   Class :character   Class :character
Median :38851   Mode  :character   Mode  :character   Mode  :character
Mean   :38845
3rd Qu.:58277
Max.   :77725
   color           date_of_birth           datetime
Length:54408       Min.   :1994-01-25   Min.   :2013-10-01 09:31:00.0
Class :character   1st Qu.:2012-08-15   1st Qu.:2014-10-18 16:20:30.0
Mode  :character   Median :2014-05-12   Median :2015-11-07 16:14:00.0
```

```
                  Mean   :2013-08-24   Mean   :2015-11-23 14:39:07.2
                  3rd Qu.:2015-09-17   3rd Qu.:2016-12-13 16:24:00.0
                  Max.   :2017-12-25   Max.   :2018-02-01 18:40:00.0
    name          outcome_type           sex            spay_neuter
Length:54408      Length:54408       Length:54408       Length:54408
Class :character  Class :character   Class :character   Class :character
Mode  :character  Mode  :character   Mode  :character   Mode  :character
```

```
skim(raw_data)
```

Table 1: Data summary

Name	raw_data
Number of rows	54408
Number of columns	11
_____	
Column type frequency:	
character	8
Date	1
numeric	1
POSIXct	1
_____	
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

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date_of_birth	0	1	1994-01-25	2017-12-25	2014-05-12	5504

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
id	0	1	38845.03	22415.88	1	19482.75	38850.5	58277.25	77725	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
datetime	0	1	2013-10-01 09:31:00	2018-02-01 18:40:00	2015-11-07 16:14:00	46757

The variables `breed` and `color` have too many different combinations, so we remove them from our data, as they will not be useful for our models.

Let's see more details on the `animal_type`, `sex` and `spay_neuter`.

```
raw_data |>
  count(animal_type) |>
  gt()
```

animal_type	n
Bird	212
Cat	20561
Dog	30830
Livestock	5
Other	2800

```
raw_data |>
  count(sex) |>
  gt()
```

sex	n
Female	23931

Male	26045
Unknown	4432

```
raw_data |>
  count(spay_neuter) |>
  gt()
```

spay_neuter	n
Fixed	37013
Intact	12963
Unknown	4432

We keep these variables.

We see that we have names for 69.8% of the animals. We can create a variable for this: `has_name`, a dummy variable is there is a name or not.

```
raw_data <- raw_data |>
  mutate(has_name = !is.na(name)) # If not NA, then it has a name.
```

Also, let's treat the `animal_type` and `sex` variables as integers, so they can enter the model

```
raw_data <- raw_data |>
  mutate(integer_sex = as.integer(as.factor(sex)),
         integer_animal_type = as.integer(as.factor(animal_type)))
```

Lastly, 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.

```
raw_data <- raw_data |>
  mutate(spay_neuter = if_else(spay_neuter == "Unknown", "Intact", spay_neuter),
         integer_spay_neuter = as.integer(as.factor(spay_neuter)))
```

Now let's work with the time variable and our outcome variable.

```
raw_data |>
  count(outcome_type) |>
  gt()
```

outcome_type	n
adoption	33275
no outcome	4735
transfer	16398

We would like to model which animals were adopted, so we make a new variable `outcome` for which we take `1` for adoption and `0` for no adoption.

```
raw_data <- raw_data |>
  mutate(outcome = if_else(outcome_type == "adoption", 1, 0))
```

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 let's create our own variable

```
raw_data <- raw_data |>
  mutate(time_to_outcome = as.Date(datetime) - date_of_birth)
```

```
raw_data |>
  filter(time_to_outcome < 0) |>
  gt()
```

id	age_upon_outcome	animal_type	breed	color	date_of_birth	
14749	0 years	Cat	Domestic Shorthair Mix	Orange Tabby	17102	147
9703	0 years	Cat	Domestic Shorthair Mix	Black	16676	143
41480	0 years	Cat	Domestic Shorthair Mix	Tortie	16622	143
682	0 years	Dog	Labrador Retriever Mix	Black	16584	143
52255	0 years	Cat	Domestic Shorthair Mix	Orange Tabby	17348	149

We have some negative values, which makes no sense as the date_of_birth should come before the outcome, so we remove these observations.

```
raw_data <- raw_data |>
  filter(time_to_outcome >= 0)
```

Let's look at the encoding for sex, spay_neuter and animal_type before we go on analyzing with our clean data.

```
raw_data |>
  count(integer_animal_type, animal_type) |>
  gt()
```

integer_animal_type	animal_type	n
1	Bird	212
2	Cat	20557
3	Dog	30829
4	Livestock	5
5	Other	2800

```
raw_data |>
  count(integer_sex, sex) |>
  gt()
```

integer_sex	sex	n
1	Female	23929
2	Male	26043
3	Unknown	4431

```
raw_data |>
  count(integer_spay_neuter, spay_neuter) |>
  gt()
```

integer_spay_neuter	spay_neuter	n
1	Fixed	37011
2	Intact	17392

Now we chose the variable we will use as a clean data.

****

We could also keep the raw character variables as was done in class

****

```
data <- raw_data |>
  select(id, animal_type = integer_animal_type, sex = integer_sex,
```

```
        spay_neuter = integer_spay_neuter,
        has_name, outcome, time_to_outcome) |>
  mutate(id = as.character(id))
```

Statistical Analysis

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

Multivariate Cox regression

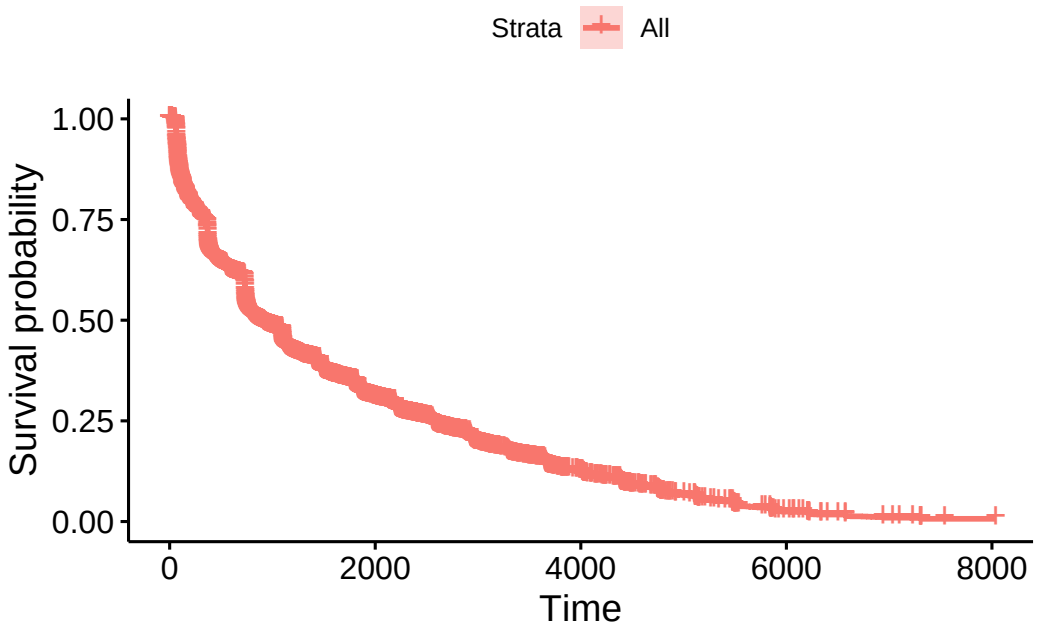
```
cox_model <- coxph(Surv(time_to_outcome, outcome) ~ animal_type + sex + spay_neuter, data
summary(cox_model)
```

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

  n= 54403, number of events= 33273

                coef exp(coef) se(coef)      z Pr(>|z|)
animal_type -0.26469   0.76744  0.01068 -24.79  < 2e-16 ***
sex         -0.06903   0.93330  0.01057  -6.53 6.59e-11 ***
spay_neuter -0.74474   0.47486  0.01825 -40.81  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
animal_type    0.7674      1.303    0.7515    0.7837
sex            0.9333      1.071    0.9142    0.9528
spay_neuter    0.4749      2.106    0.4582    0.4922

Concordance= 0.615  (se = 0.002 )
Likelihood ratio test= 3193  on 3 df,   p=<2e-16
Wald test            = 2601  on 3 df,   p=<2e-16
```

```
Score (logrank) test = 2676  on 3 df,   p=<2e-16
```

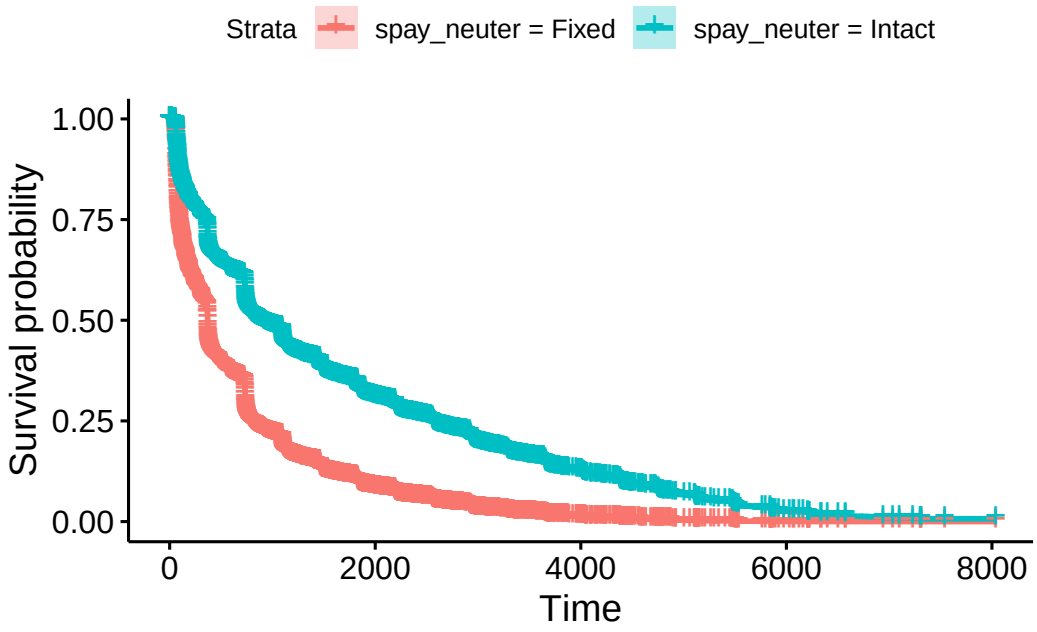
Let's see our predicted survival proportion for the whole data.

```
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(spay_neuter = c(1, 2),
                             animal_type = c(1, 1),
                             sex = c(1, 1))
# 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