Final project: Survival analysis of cetaceans in captivity

Kerinna Good and Josh Wagner

Introduction

The `all_cetaceans2` data set used in this analysis contains information on 1,416 dolphins and whales living in captivity in the US between 1943 and 2017. Our analysis explores the relationship between cetacean survival rates in captivity and species, sex, whether animals were born in the wild or captivity, the percentage of animals' life lived in captivity, location at date of recorded status, whether animals were transferred between facilities during their lifetime, and whether the animals were transferred from foreign countries. Our research question asks, do animals that spend more of their lives in captivity live longer or shorter lives than animals that spend less of their lives in captivity? We hypothesize that animals that spend more of their lives in captivity will live shorter lives than animals that spend less of their lives in captivity. The results of our analysis could help inform best practices for holding dolphins and whales in captivity, and further could inform policies geared toward reducing cetacean fatalities in captivity.

Our data comes from an article in The Pudding published by Amber Thomas in 2017 titled "Free Willy and Flipper by the Numbers" (Thomas (2017)). Specifically, we use a data set titled `all_cetaceans2` which is a modified version of the `allCetaceanData` data set published on Data.world (Data.world (2017)). The data were originally collected from the National Marine Mammal Inventory (curated by the National Oceanic and Atmospheric Administration) and the crowd-sourced website Ceta-Base (Thomas (2017)). Starting with the `allCetaceanData` data set, we renamed several variables for clarity, mutated a new variable for status date that lists the status date for alive cetaceans as May 7, 2017 (the date at which each animal's status was evaluated) opposed to NA, mutated a new variable for each cetacean's age in years, mutated a new variable to describe the percentage of an animal's life spent in captivity, simplified the status levels to dead or alive (e.g. stillborn animals were considered dead, released animals were considered alive), simplified the acquisition levels to born into captivity, wild, or unknown (e.g. stillborn animals were considered born into captivity, rescued animals were considered wild), simplified the species variable to bottlenose or not bottlenose (to reduce the original thirty-seven species levels to two), simplified the current location variable to the ten most common locations, and removed observations with impossible birth years relative to origin year and status year.

The code book describing variables used in our analysis is below.

Variable	Class	Description
species	character	Species of animal (whale or dolphin)
sex	character	Sex of animal
acquisition2	character	Method through which an animal was brought into captivity
currentlocation2	character	Location of animal at date of recorded status
transfer	binary	Whether current location matches origin location
foreigntransfer	character	Whether an animal was transferred from outside the US
status2	binary	Binary indicator for status (dead or alive)
age	integer	Age of animal at date of recorded status
captivity	integer	Percentage of animal's life lived in captivity

We conducted exploratory data analysis on the relationship between our categorical variables of interest (species, sex, acquisition method, current location, transfer, and foreign transfer) and cetacean survival rates.

Plots of Kaplan-Meier survival time curves for acquisition method, transfer, and current location as shown below suggest that 1) cetaceans born into captivity tend to have lower survival rates at the beginning of their lives and higher survival rates at the end of their lives when compared to cetaceans captured from the wild and 2) cetaceans transferred from one facility to another tend to have lower survival rates at the beginning of their lives and higher survival rates at the end of their lives when compared to cetaceans that aren't transferred.

Figure 1

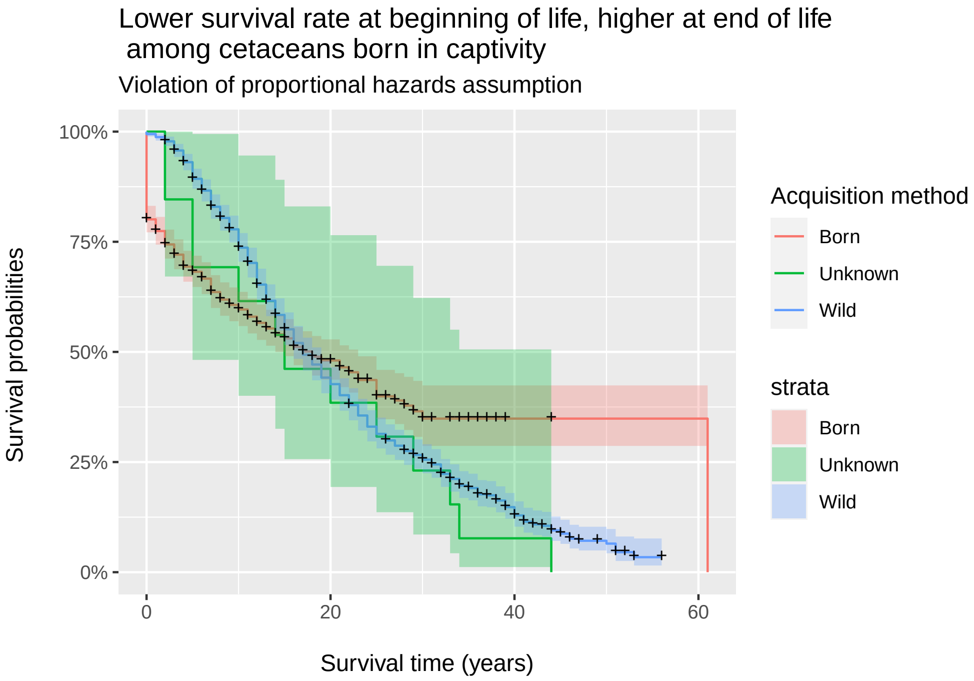


Figure 2

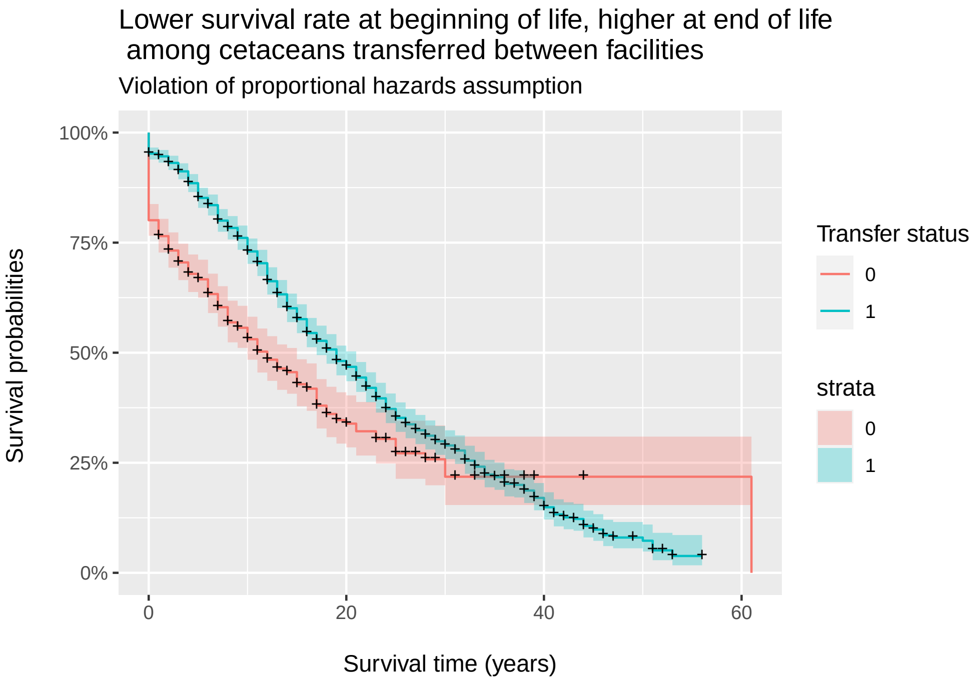
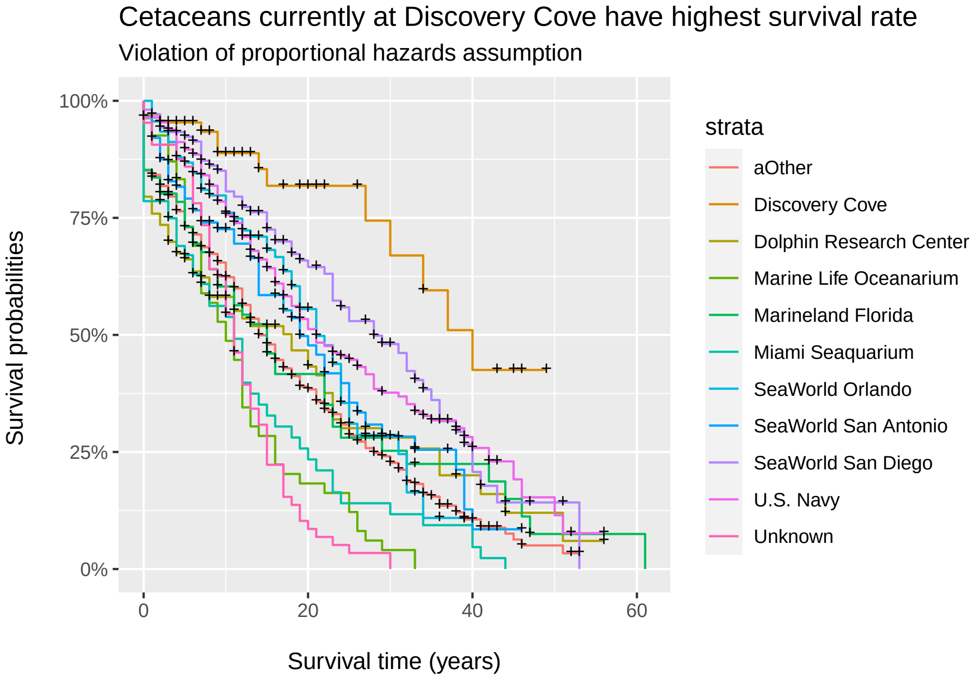


Figure 3



Code for survival visualizations from dkmathstats Website (David (n.d.)).

Methodology

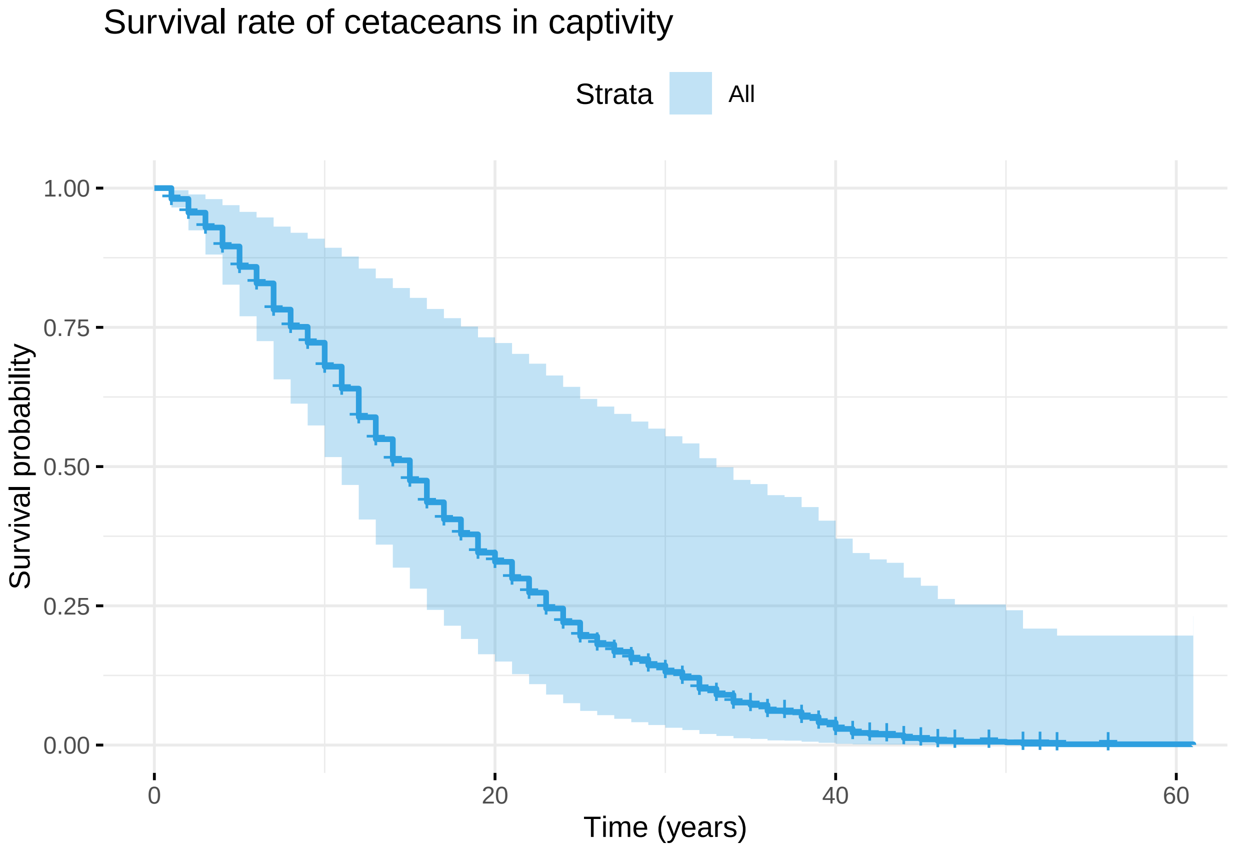
We chose to fit a multivariable Cox proportional hazards model because we want to predict survival rates of cetaceans in captivity based on a mix of categorical, binary, and numeric variables. Thus, the Cox proportional hazards model proved most appropriate in how it allows one to "evaluate simultaneously the effect of several factors on survival" (Kassambara (n.d.)). We then evaluated whether or not the assumptions of the Cox Proportional-Hazards Model were violated for our data set. First, the independence of observations assumption is shown to be satisfied because knowing something about one Dolphin does not tell you anything about another that is not controlled for in the model. Secondly, the assumption for the Cox proportional hazards model that censoring and death are independent is shown to be satisfied seeing that death is categorized separately from the other methods by which a Dolphin could be censored such as it being released. As a result, no Dolphin could be categorized as both censored and dead and thus the two phenomena are independent. Finally, the proportional hazards assumption of the Cox proportional hazards model is shown to be violated in that the hazard curves of different groups cross … However, we will proceed on in spite of this but make sure to take note of it when interpreting our results.

The outcome variable of interest in our model is lifespan, or the time between an animal's birth and death. The time variable is `age` (in years) and the event of interest is `status2`, or the status of an animal as alive (0) or dead (1). In terms of predictors variables, then, we considered all unique, informative, and usable categorical and numeric variables in the `all_cetaceans2` data set. We identified species, sex, acquisition method, current location, transfer status, foreign transfer status, and captivity as meeting the criteria for predictors. Other variables found in the `allCetaceanData` data set such as mother, father, origin location, list of transfers, and cause of death had too many variable levels to use in the model.

Results

The final model is printed below.

Predictor	Coefficient	Std. Error	Statistic	P-value
species2Bottlenose	-0.4062	0.0966	-4.2035	0.0000
sexM	0.1738	0.0751	2.3128	0.0207
sexU	0.7554	0.2674	2.8250	0.0047
acquisition2Unknown	-0.5234	0.3130	-1.6720	0.0945
acquisition2Wild	-0.3089	0.1370	-2.2545	0.0242
captivity	-2.3513	0.1502	-15.6559	0.0000
transfer	-0.4102	0.1550	-2.6467	0.0081
foreigntransferUS	0.4172	0.3285	1.2700	0.2041
currentlocation2Discovery Cove	-0.9416	0.3119	-3.0190	0.0025
currentlocation2Dolphin Research Center	-0.2323	0.1737	-1.3373	0.1811
currentlocation2Marine Life Oceanarium	0.8279	0.1617	5.1212	0.0000
currentlocation2Marineland Florida	0.0328	0.1840	0.1783	0.8585
currentlocation2Miami Seaquarium	0.5366	0.1816	2.9540	0.0031
currentlocation2SeaWorld Orlando	-0.0575	0.1568	-0.3670	0.7136
currentlocation2SeaWorld San Antonio	0.0888	0.1604	0.5535	0.5799
currentlocation2SeaWorld San Diego	-0.4571	0.1502	-3.0433	0.0023
currentlocation2U.S. Navy	-0.4717	0.1160	-4.0670	0.0000
currentlocation2Unknown	0.0937	0.2175	0.4307	0.6667



Code for visualization from Statistical tools for high-throughput data analysis (Kassambara (n.d.)).

The model output shows that the predictors for species (Bottlenose), sex (male), sex (unknown), acquisition method (wild), percentage of life spent in captivity, transfer between US facilities, current location (Discovery Cove), current location (Marine Life Oceanarium), current location (Miami Seaquarium), current location (SeaWorld San Diego), and current location (U.S. Navy) all have significant p-values at the $\alpha = 0.05$ significance level.

The p-value for the `captivity` variable is notably small at less than $2e^{-16}$. (Add interpretation of captivity variable here).

What other variables do we want to interpret? Acquisition2 (Wild)? Transfer?

Discussion

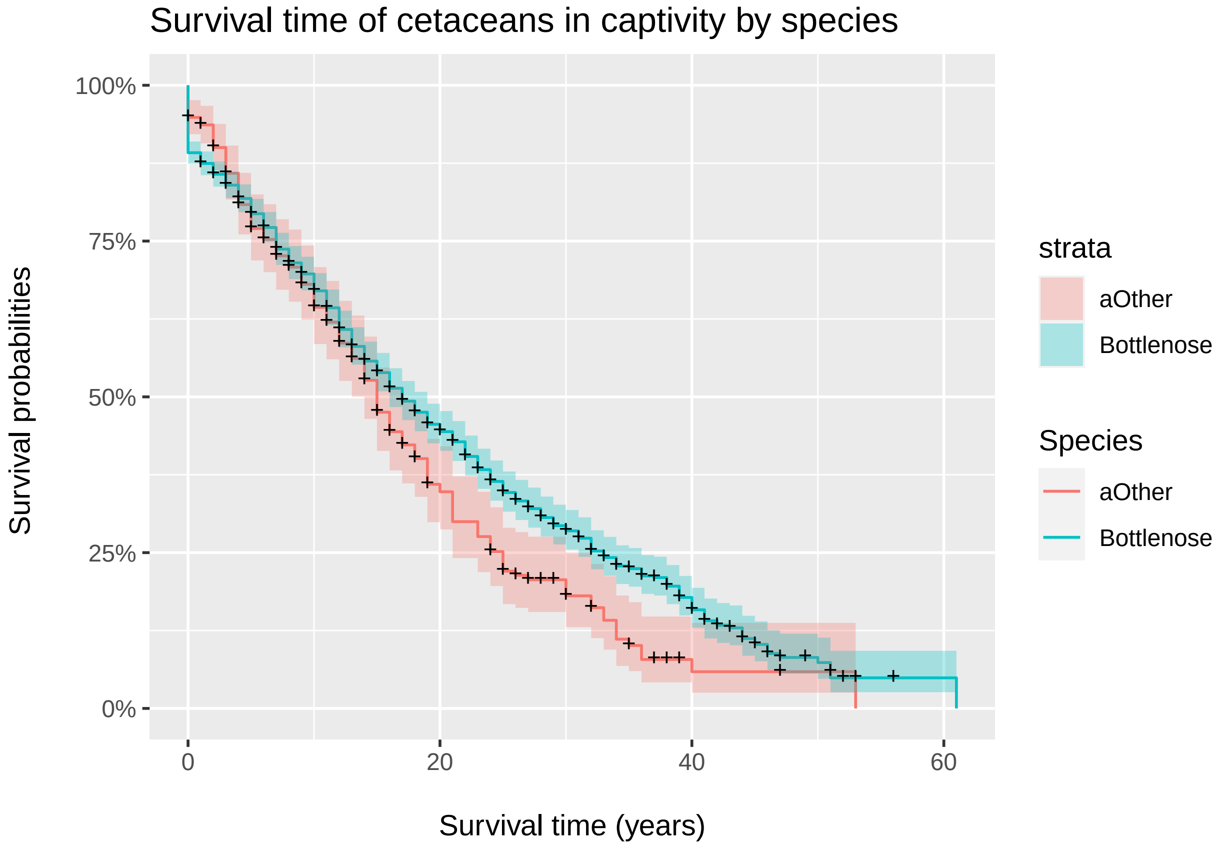
Our model suggests that the longer a cetacean lives in captivity, the less likely it is to survive (not sure if this is our actual conclusion).

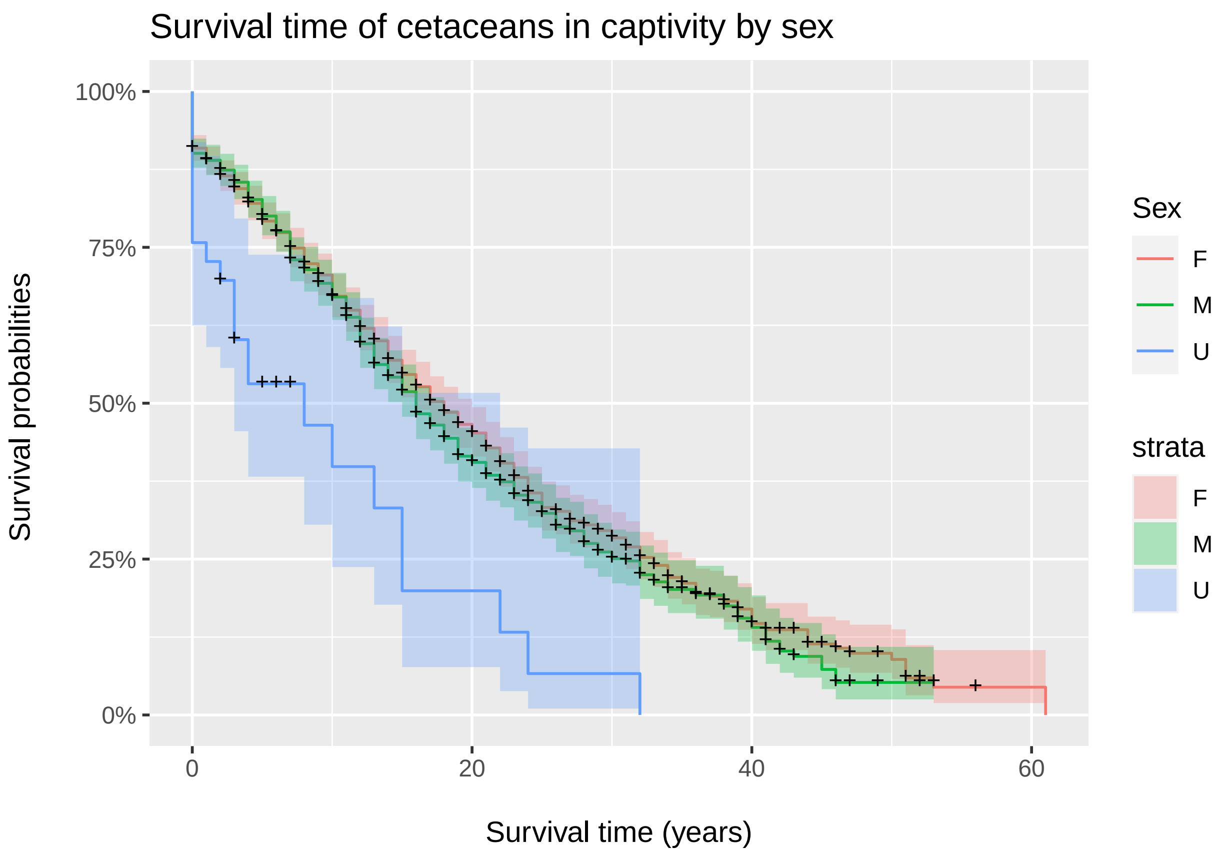
When initially considering how to select our predictor variables, we considered using a variable selection technique such as LASSO regression or all subset selection. However, in the case of LASSO regression we found that using LASSO as a means of selecting variables for a Cox model was relatively new and involved statistical principles which were beyond the scope of our current understanding (Tibshirani (1997)). Similarly, in researching the feasibility of using all subset selection, we found that the answer to which variable selection technique is best for the Cox proportional hazards model is still up for debate (Ekman (2017), Fan and Li (2002), Petersson and Sehlstedt (2018)). Thus, to avoid misusing the appropriate variable selection technique for our project, we decided to instead use all of the unique, informative, and usable predictor variables from our data set in our model. As a result, it is possible that a more effective model could be derived if future researchers were to beforehand use one of the aforementioned variable selection techniques.

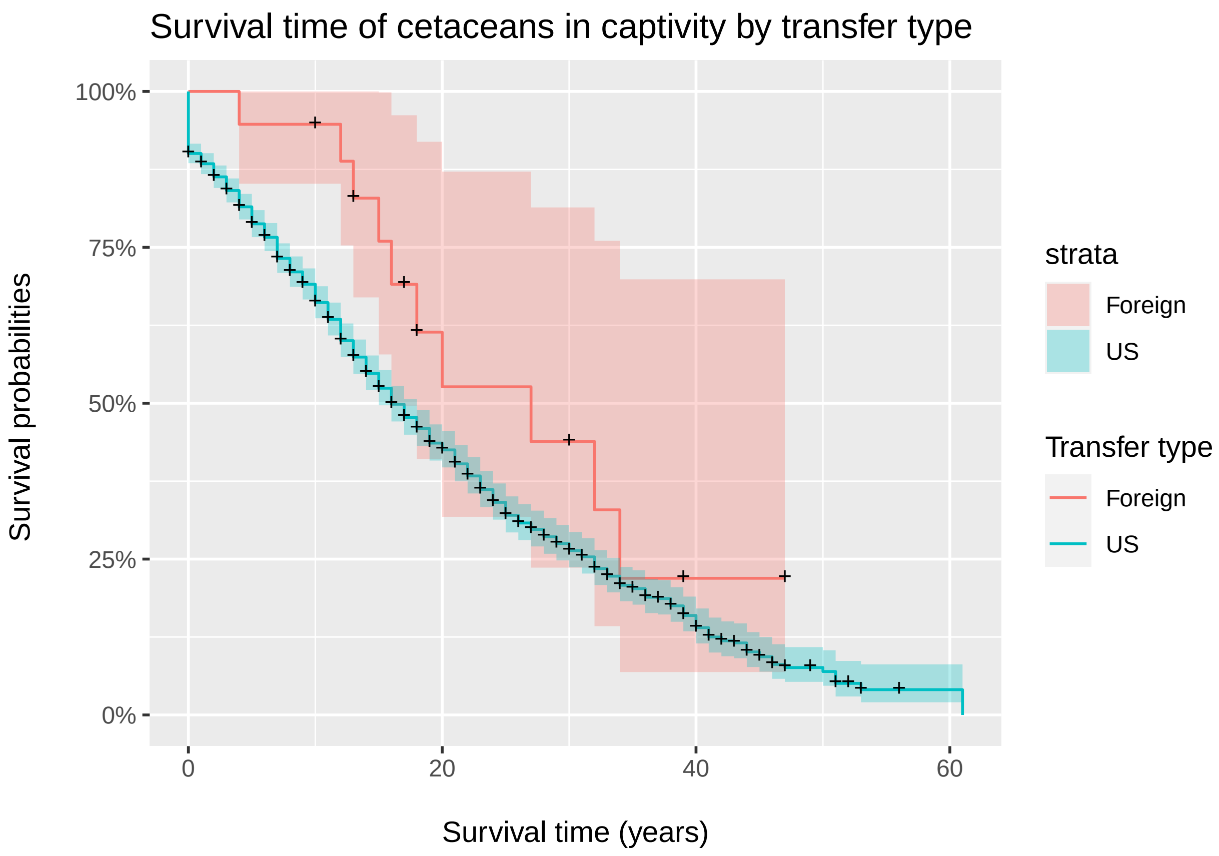
Additionally, as mentioned in the methodology section, not all of the assumptions for the Cox Proportional Hazards Model were shown to be satisfied. Subsequently, the interpretations and conclusions we are able to draw from our model are inherently limited. Future work could be done with the data set so that the assumptions are not violated and more rigorous analysis can be performed.

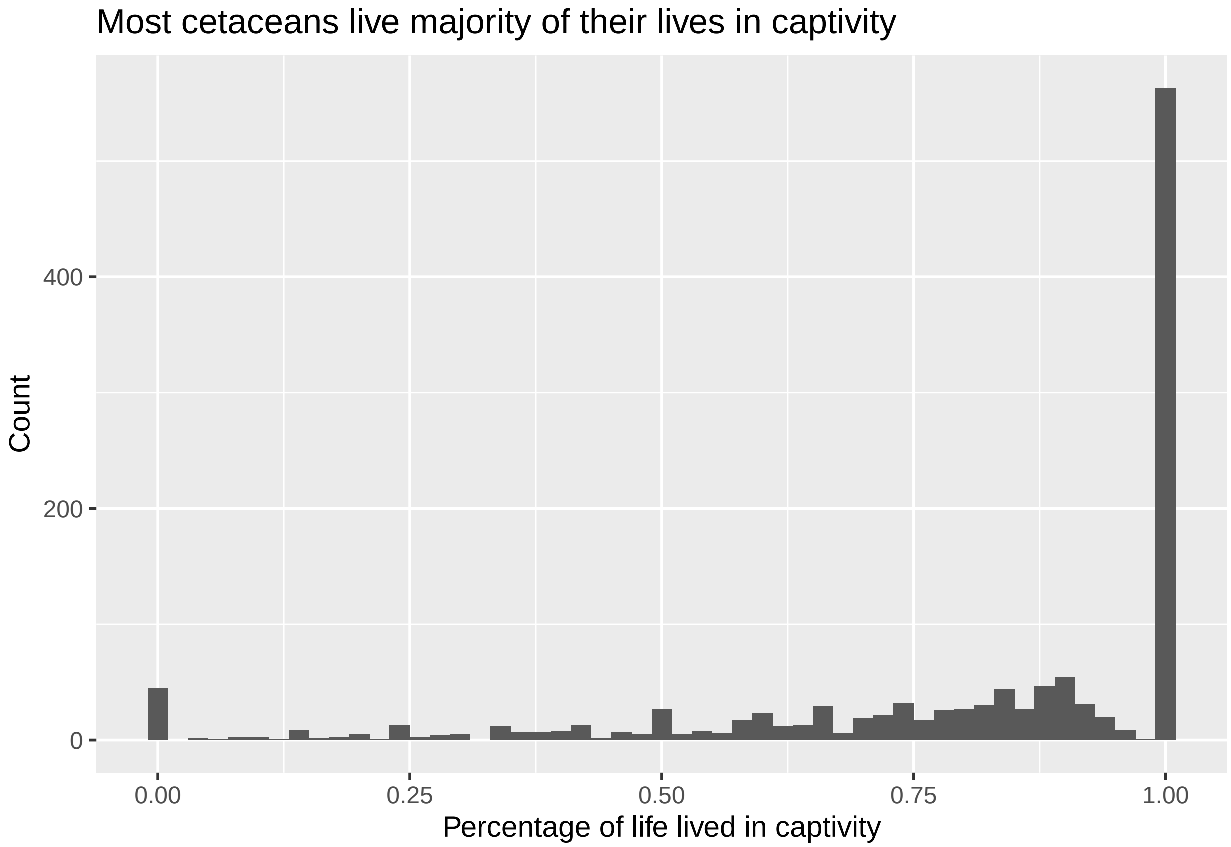
Appendix

Maybe include these figures as an appendix? Not sure.




Survival time of cetaceans in captivity by sex
Survival probabilities
100%
75%
50%
25%
0%
0
20
40
60
Survival time (years)
Sex
F
M
U
strata
F
M
U


Survival time of cetaceans in captivity by transfer type
Survival probabilities
100%
75%
50%
25%
0%
0
20
40
60
Survival time (years)
strata
Foreign
US
Transfer type
Foreign
US



References

Data.world. 2017. “Captive Whales and Dolphins in the US (1938 - 2017).” Data.world. https://data.world/the-pudding/cetaceans/activity.

David. n.d. “Plotting Kaplan-Meier Survival Times Curves in r with Ggplot2.” dkmathstats Website. https://dk81.github.io/dkmathstats_site/index.html.

Ekman, Anna. 2017. “Variable Selection for the Cox Proportional Hazards Model: A Simulation Study Comparing the Stepwise, Lasso and Bootstrap Approach.” PhD thesis, Umeå University. http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-130521.

Fan, Jianqing, and Runze Li. 2002. “Variable Selection for Cox’s Proportional Hazards Model and Frailty Model.” *The Annals of Statistics* 30 (1): 74–99. https://doi.org/10.1214/aos/1015362185.

Kassambara, Alboukadel. n.d. “Cox Proportional-Hazards Model.” Statistical tools for high-throughput data analysis. http://www.sthda.com/english/wiki/cox-proportional-hazards-model.

Petersson, Simon, and Klas Sehlstedt. 2018. “Variable Selection Techniques for the Cox Proportional Hazards Model: A Comparative Study.” PhD thesis, University of Gothenburg School of Business. https://core.ac.uk/download/pdf/152600668.pdf.

Thomas, Amber. 2017. “Free Willy and Flipper by the Numbers.” https://pudding.cool/2017/07/cetaceans/.

Tibshirani, R. 1997. “The Lasso Method for Variable Selection in the Cox Model.” *Statistics in Medicine* 16 (4): 385–95.