fig.cap="After fitting as a survival object, our same data fits a Kaplan-Meier Curve for all shelter labradors with a horizontal line at the 50% Survivor probability which is approximately 732 days",

Data Source and Background:

The raw data included cats, dogs, birds, livestock and other. Since we were only interested in Labradors we used only a subset of the data which included 2,128 Labrador breeds and The reason there were so many Labrador breeds is because if the lab showed any signs of being mixed, the shelter employee would document it with it's suspected mix breed.

Our data is from the Austin Animal Center which is a large no-kill animal shelter. It can be found on <https://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-outcomes-and>.

SEX: We decided to exclude the rows where the sex\_upon\_outcome was “unknown” because there were only 48 rows with this, which we agreed wouldn’t make a different in our analysis, and because it would show as unknown in two of our columns, “sex” and “neuter.”

DATE\_DIFF: We got two date\_diff that were negative. We looked further into our data and saw that the date of event or censorship was before the date of entrance to the animal shelter. Since this is not possible, we assumed that there was an error in data entry and decided to exclude these two rows from our analysis.

We end up with a data frame for Labradors in the Austin Animal shelter with the columns “color,” “event,” “date\_diff,” “sex,” and “neuter.”

----PRELIMINARY ANALYSIS----

Kaplan-Meier Curves

All labs – we noticed that the confidence interval curves were really close to the actual survival curve. This proves that the data is very detailed and the information is concrete because we have enough observations to do analysis on Labradors in the Austin animal shelter.

By Sex – There is not a noticeable difference between the two curves so our pre-analysis hypothesis is that there is not going to be a significant difference. There are 3211 female observations and 3718 male observations which is pretty close.

By Fixed Status – we see a bigger difference in the two survival curves between fixed and intact, mainly that fixed has a lower survival probability over time, however this might be because there are many more observations on fixed labs than on intact labs. (fixed has about 4,000 more observations) so we will proceed with caution in our analysis because of this fact.

By fur pattern – There is a pretty much even amount of labs with mixed fur and single fur patterns. There seems to be a slight difference in survival probability between the two fur patters with mixed having a lower survival probability.

------ Research Questions-------------

On average, how much time do Labradors spend in animal shelters before they go to a home?

We will plot a Kaplan-Meier curve and find the median time that Labradors spent in this animal shelter.

Does the color, gender, or neuter status of a Labrador affect its chance for adoption?

We’ll build a Cox Proportional Hazards (Cox PH) model with the covariates neuter status, sex, and color to answer this question. We will also explore an Accelerated Failure Time (AFT) model as our extension component, to look at a comparison of survival times.

---------- Kaplan Meier -------------------

To visually see the estimated survival curves we must plot Kaplan-Meier Curves for each covariate `color`, `sex` and `neuter` on time to adoption for Labradors.

The Kaplan-Meier curves for sex in Figure 3 shows that male Labs have higher estimated survival rates since the blue male line is above the red female line. As time increases, male and female Labs have the same or similar survival rates because the lines look like they begin to overlap after the 3,000 days mark.

The Kaplan-Meier curves for neuter status in Figure 4 shows that intact Labs (labs that have not been neutered) have higher estimated survival rates than fixed labs. This means that more fixed labs are being adopted or going to homes (having the event) than labs that have not been neutered.

The Kaplan-Meier curves for fur patterns in Figure 5 show labs with single colored fur patterns have higher survival rates than those with mixed fur patterns. The gaps immediately widen after the 50% survival rate mark and begin to close again nearer to the bottom.

---------- Log-Rank Test -------------------

The log-rank test allows us to compare 2 or more survival curves, more specifically, that there is a significant difference between the survival curves. The null hypothesis that we are testing is that there is no difference in survival curves. The low p-value means we reject our null hypothesis.

After performing log-rank test for the 3 covariates (sex, fur pattern, fixed status) we see that the p-value for each log-rank test is less than alpha=0.05 which means that each of the covariates are statistically significant.

----------- Model Building -----------------------

When deciding what covariates to test in our model, we initially hypothesized that sex would not be significant. Therefore we created a separate Cox Proportional Hazards model that only controlled for sex and looked at the likelihood ratio test. However, after looking at the summary, we can see that we were wrong, and that the p-value associated with the likelihood ratio is less than $\alpha = 0.05$, which means that covariate is statistically significant for the model.

------------- Model Checking ----------------------

In order to build our Cox PH model we must check to see if the proportional hazard assumptions are met. We will do this by graphical evaluation of the log-log plots and by doing a residuals test using the cox.zph function.

When evaluating the log-log plots we want to make sure that the survival curves don’t cross and that the curves look proportional to one another.

Figure \_\_\_\_ shows the log-log plot for the sex covariate. We see that the curves cross and overlap both at the beginning and end of the curves.

--------------Accelerated Failure Time Model -----------------------

For our extension we will try to make an Accelerated Failure Time (AFT) model. The AFT model is different than the PH model in that the AFT model is used to compare survival times where the acceleration factor $\gamma$ is the key measure that describes the difference in survival times. This is analogous to the hazard ratio in the PH model. In order to obtain this acceleration factor we must first explore which parametric survival model to use because if a parametric survival model doesn’t make sense to our data, we won’t be able to create an AFT model. We analyze the $log(-log(\hat{S}))$

A parametric survival model is one in which

survival time (the outcome) is assumed to follow

a known distribution.

AFT assumption is proportional survival.

The Weibull model also has another key property:

the log(–log) of S(t) is linear with the log

of time. This allows a graphical evaluation of

the appropriateness of aWeibull model by plotting

the log negative log of the Kaplan–Meier

survival estimates against the log of time.