Exam PA December 13, 2021 Project Statement and Model Solution

This model solution is provided so that candidates may better prepare for future sittings of Exam PA. It includes both a sample solution, in plain text, and commentary from those grading the exam, in italics. In many cases there is a range of fully satisfactory approaches. This solution presents one such approach, with commentary on some alternatives, but there are valid alternatives not discussed here.

General Information for Candidates

This examination has 10 tasks numbered 1 through 10 with a total of 100 points. The points for each task are indicated at the beginning of the task, and the points for subtasks are shown with each subtask.

Each task pertains to the business problem (and related data file) and data dictionary described below. Additional information on the business problem may be included in specific tasks—where additional information is provided, including variations in the target variable, it applies only to that task and not to other tasks. An .Rmd file accompanies this exam and provides useful R code for importing the data and, for some tasks, additional analysis and modeling. The .Rmd file begins with starter code that reads the data file into a dataframe. This dataframe should not be altered. Where additional R code appears for a task, it will start by making a copy of this initial dataframe. This ensures a common starting point for candidates for each task and allows them to be answered in any order.

The responses to each specific subtask should be written after the subtask and the answer label, which is typically ANSWER, in this Word document. Each subtask will be graded individually, so be sure any work that addresses a given subtask is done in the space provided for that subtask. Some subtasks have multiple labels for answers where multiple items are asked for—each answer label should have an answer after it. Where code, tables, or graphs from your own work in R is required, it should be copied and pasted into this Word document.

Each task will be graded on the quality of your thought process (as documented in your submission), conclusions, and quality of the presentation. The answer should be confined to the question as set. No response to any task needs to be written as a formal report. Unless a subtask specifies otherwise, the audience for the responses is the examination grading team and technical language can be used. When "for a general audience" is specified, write for an audience **not** familiar with analytics acronyms (e.g., RMSE, GLM, etc.) or analytics concepts (e.g., log link, binarization).

Prior to uploading your Word file, it should be saved and renamed with your five-digit candidate number in the file name. It is not required to upload your .Rmd file or other files used in determining your responses, as needed items from work in R will be copied over to the Word file as specified in the subtasks.

The Word file that contains your answers must be uploaded before the five-minute upload period time expires.

Business Problem

You work at XYZ, a large actuarial consulting firm. Your boss, B, is a Fellow of the Society of Actuaries with expertise in Predictive Analytics. Outside of work, B volunteers at an animal shelter that started operating in 2019. B recently convinced the decisionmakers at XYZ to take on the shelter as a pro bono (i.e., unpaid) client and put you in charge.

Animal shelters take in unwanted and lost dogs and cats. Some animals are reclaimed by owners, typically very soon. At "No Kill" shelters like the local one the unclaimed animals are housed until someone adopts them as a pet. Before the pandemic created a surge in demand for pets that emptied the local shelter, it housed an increasing population of animals because the demand for local adoptions was less than the flow of unclaimed animals into the shelter. To avoid returning to the same unsustainable situation, the shelter plans to start a transfer program whereby some animals are transferred to partner organizations in other locations where there is high demand for adopted pets. Transfers can help a shelter place many animals at once. They are a useful tool to manage shelter population levels (as opposed to a last resort for unadoptable animals). A transfer program can only transfer animals that the partner organization agrees to accept.

B has identified the following issues that the local shelter faces:

- *Understanding the characteristics of animals included in transfer agreements would aid the local shelter in preparing to start such a program.*

- *An accurate estimate of the length of time between arrival at the shelter and placement (return to owner, adoption, or transfer) would aid the shelter in planning and budgeting. They want to estimate how long that animal will stay as each animal arrives.*

B also created a dataset[1] using public data from the Austin Animal Center (AAC) for you to use. AAC is a "No Kill" animal shelter in Austin, Texas. AAC has a robust transfer program and an excellent reputation. Your city is similar in size to Austin.

B has provided the following data dictionary and the dataset of 48,409 records derived from AAC data in a file called Exam PA Animal Shelter Data.csv*.*

[1] Adapted from Austin Animal Center Intakes (2021) and Austin Animal Center Outcomes (2021) City of Austin, Texas Open Data Portal, https://doi.org/10.26000/025.000002 and https://doi.org/10.26000/025.000001.

Data Dictionary

Variable Name	Variable Values
outcome	Adoption, Transfer, Return to Owner
stay	length of stay in days (0 to 1913)
animal	Cat, Dog
mf	Male, Female
age	age at intake in years (-0.1 to 24)
in.month	1 to 12
in.year	2013 to 2021
out.month	1 to 12
out.year	2015 to 2021
in.reason	Owner Surrender, Public Assist, Stray
in.intact	1 if able to have offspring, else 0
out.intact	1 if able to have offspring, else 0
name	many values
breed	many values
color	many values

Comments

The variables including “in.” in their names indicate conditions when an animal arrives at the shelter, and those including “out.” indicate conditions when an animal leaves the shelter, as indicated by outcome.

The data includes stays that ended in 2015 or afterwards.

Animals born at the center will have negative intake ages.

Many animals undergo a procedure during their stay that prevents offspring.

Task 1 (*6 points*)

You assess the applicability of the data for addressing the business problem.

(a) (*3 points*) Describe, for a general audience, two advantages of using data from the Austin animal shelter in addressing this business problem.

Stronger candidates clearly described two specific advantages of using data from the Austin animal shelter for the local shelter and briefly explained why. Candidates were not awarded full credit unless the advantages were specific to the business problem and the writing was appropriate for a general audience.

ANSWER:

The data from the Austin animal shelter is expected to be a good proxy of a representative sample for our local shelter since it is also a "no kill" shelter in a city similar in size to Austin.

Austin has a robust transfer program and has been operating for many years, with much of its data preceding the pandemic, providing a reliable source of data and a better opportunity to understand normal operating conditions.

(b) (*3 points*) Describe, for a general audience, two disadvantages of using data from the Austin animal shelter in addressing this business problem.

Expectations were similar to those outlined above for advantages. Disadvantages related to specific data quality issues were awarded credit only with justification tying back to the business problem

ANSWER:

The data from the Austin animal shelter may not be an appropriate proxy sample because the two shelters and local conditions may be different in some specific manner that could invalidate specific predictors, such as rates of procedures that prevent animals from having offspring.

Given that the local shelter is relatively new, it may not have the ability to collect the same data as the Austin animal shelter did on all of its cats and dogs, or the data may be collected in a substantially different manner. Therefore, the model developed from this data can't be applied to the local shelter.

Task 2 (*9 points*)

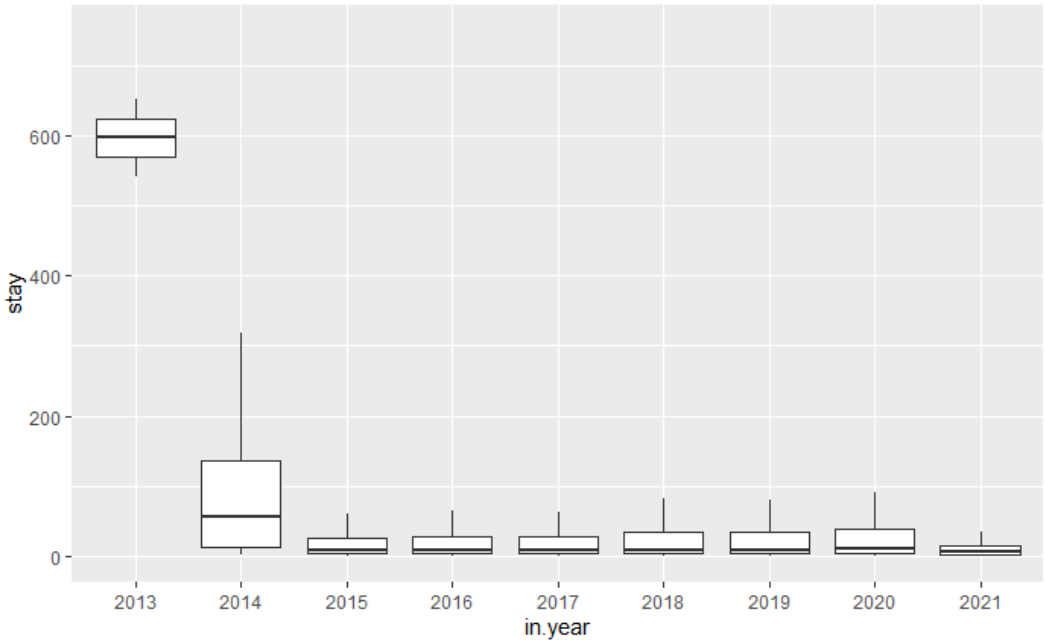
The shelter has asked for a model that predicts length of stay. Your assistant is eager to start and wants to put all the variables into a predictive model. After explaining the value of exploratory data analysis, you have your assistant make a series of graphs of **stay** against all the other variables (except **name**, **breed**, and **color**). Run the starter code and the code for task 2 in the .Rmd file to see the graphs.

> (*3 points for each decision*) Explain to your assistant three important modeling decisions you can make based on one or more of the graphs. Copy and paste the graph(s) used in each response below. Do not apply these modeling decisions for future tasks.

Candidates generally performed well on this task. Responses based on in.reason, in.intact, animal type, gender were all accepted. Very few candidates addressed the issue of age binning. Partial credit was awarded for comments on how graphs looked visually while failing to give a recommendation on modeling decisions. No credit was awarded for the use of "out" variables as the use of those variables constituted target leakage.

FIRST ANSWER:

Graph(s):

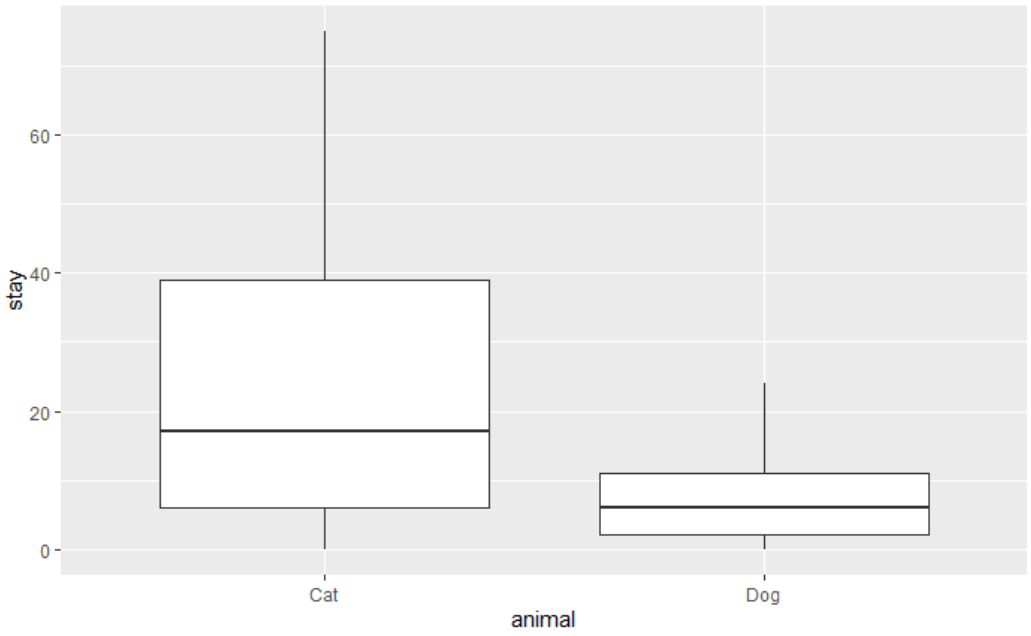


Explanation:

The data includes stays that ended in 2015 or later. For stays that started in 2013 or 2014, only those stays that lasted until 2015 or later are included in the dataset while shorter stays are excluded, leading to skewed expected stays for 2013 and 2014 arrivals, as seen in the graph above. Because **stay** will be estimated based on arrival data, remove all 2013 and 2014 arrivals to prevent a skewed prediction.

SECOND ANSWER:

Graph(s):

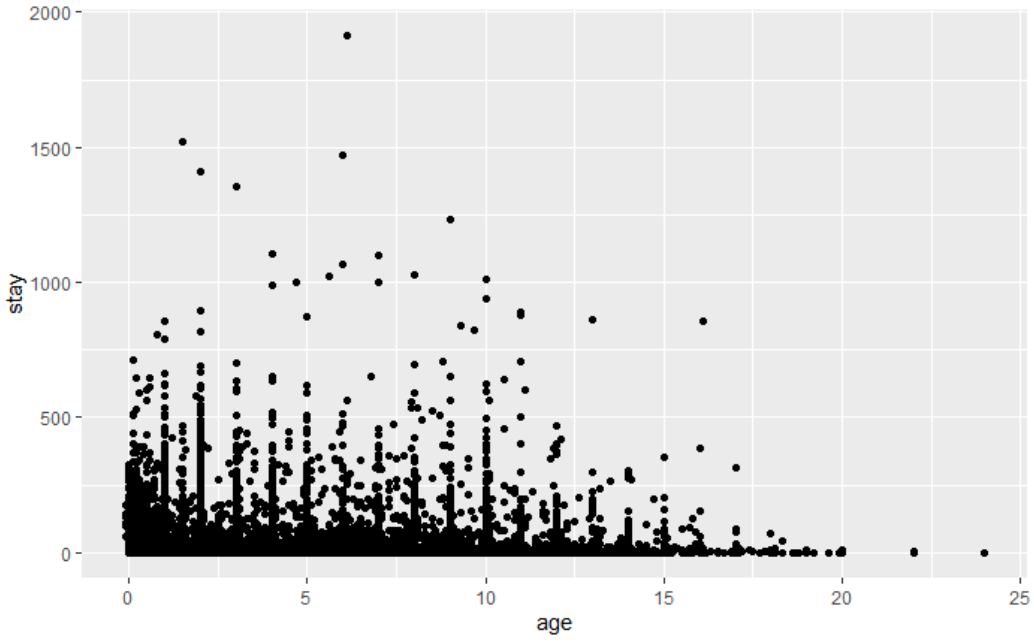


Explanation:

The boxplots, including the one above by animal, show that **stay** is a skewed variable, with upper percentiles being further away from the median length of stay than lower percentiles. A skewed target variable can affect how much prediction errors, particularly those of outliers in the upper tail, influence the fitting of the predictive model. Based on the type of model used, transform the **stay** or choose a loss function that works well with skewed distributions to improve the model fit.

THIRD ANSWER:

Graph(s):



Explanation:

The clumped nature of **age** indicates that the variable has a misleading appearance of precision. The lack of precision is not surprising in that ages are being assigned to stray dogs–certainly educated guesses are involved, and such guesses tend to be round numbers. Develop rounded and binned versions of age as possible predictors so that round-number guessing has less influence on the predictive model results.

Task 3 (*14 points*)

Your assistant notices several instances in **name** where weight is provided instead of a name and creates a new variable called **name.weight** for such instances. You explore the new variable.

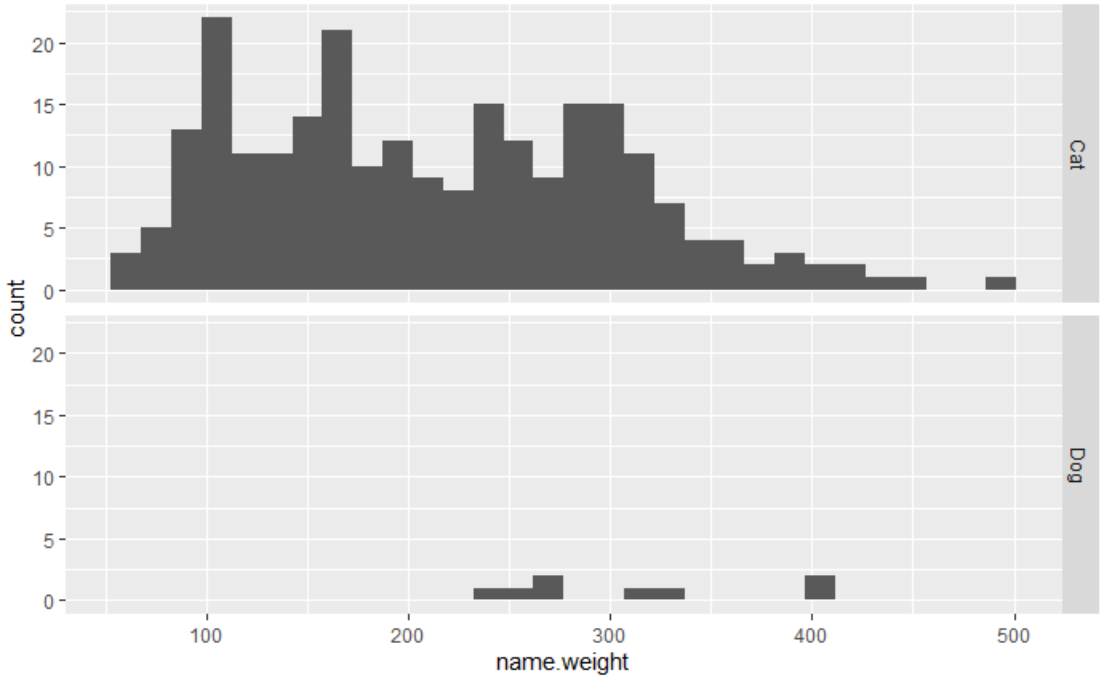
(a) (*4 points per observation*) Provide two observations based on bivariate analysis of **name.weight**. Include a graph supporting each of your observations in your answers below.

The goal of this subtask is for the candidate to provide bivariate analysis of name.weight with each of two other variables and to provide reasonable observations for each based on an informative and supportive graph. Many candidates were successful in completing this subtask and received full credit. Common observations include (but are not limited to) weights being used instead of names for transferred, young, cats, and in the most recent years. Another observation is that all 251 instances of weight were for animal transfers, and none for adoptions or returns to owner.

Providing two observations for the same bivariate analysis of two variables received partial credit. Common mistakes include (but are not limited to) providing single variable analysis such as a density plot or quantiles for name.weight and creating a regression for the limited observations and using it to indicate whether weight is a predictive variable.

FIRST ANSWER:

Graph:

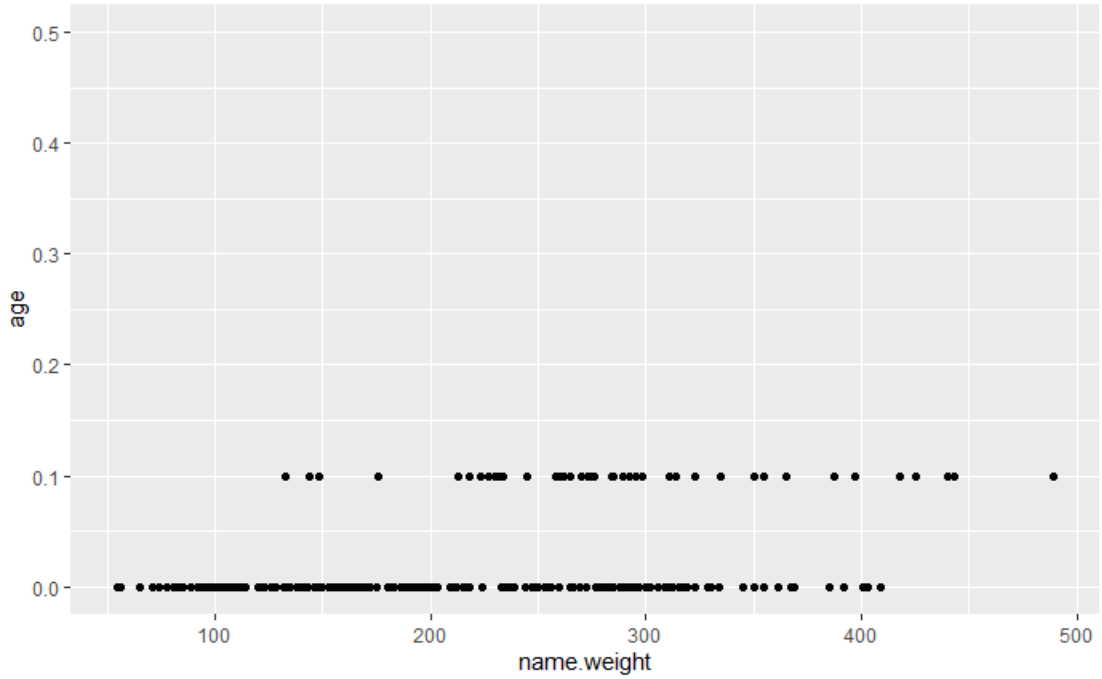


Observation:

In general, few names are weights, being only 251 observations out of 48,409 records. As seen above, the overwhelming majority of these weights as names are cats, and only 8 are dogs.

SECOND ANSWER:

Graph:



Observation:

No animal with weight is older than 0.1 years old, and most are 0.0 years old. Together with the prior observation, weights are assigned primary to kittens.

(b) (*3 points*) Describe two advantages of collecting animal weight within the context of the business problem.

Quality responses clearly explained two advantages specific to the business problem. Candidates generally did well on this subtask. Some candidates failed to receive full credit as they listed two identical advantages, such as both relating to improving predictive power. Another advantage is that weight is a good indication of health given average weight for a given breed and age.

ANSWER:

Animal weight is already tracked so minimal effort would be needed to include the data in the future.

Animal weight may capture some information from **breed** while reducing dimensionality and being less subjective, leading to more robust modeling.

(c) (*3 points*) Describe two disadvantages of collecting animal weight within the context of the business problem.

Quality responses clearly explained two disadvantages specific to the business problem. Candidates generally did well on this subtask. Some candidates failed to receive full credit as they listed two identical

disadvantages, such as both relating to the presence of missing values. Another disadvantage is that it is unclear that weight will improve predictions materially.

ANSWER:

Historical data for weight is not available in the Austin animal shelter data and weight data would be time-consuming to input for the local animal shelter because of incomplete records and manual processes, so a model including animal weight cannot be built for some time.

The animal shelter may not track animal weight immediately at intake or in all situations, and weight may vary during a stay, all reducing the reliability of the factor.

Task 4 (*11 points*)

Your assistant notes that it may be a good idea to develop a feature for a generalized linear model (GLM) based on whether the pet is a purebred or of a mixed breed.

(a) (*2 points*) Explain the challenges the **breed** variable presents when considering it as a predictor variable in a GLM.

Well prepared candidates mentioned the high number of levels or high dimensionality and the impacts of high dimensionality on GLM (e.g. overfitting and/or increased computational burden). Many candidates lost points even though they mentioned the high levels/dimensionality because no commentary on the impact to GLM was provided.

No credit was given for responses pertaining to the ethics of collecting this data or other comments that were not relevant to the breed when considered in the context of GLM.

Many candidates incorrectly identified the number of levels as 7 (based on top 7 outputted in certain R summary views) and others assumed there was an 'other' category that existed (again misinterpreting certain R summary views). While these candidates did not lose points on these explicitly, it often led to major issues in parts (b) and (c) of this task.

As with most subtasks, a full-credit answer can be brief, but many candidates gave a longer than needed response.

ANSWER:

The factor variable **breed** has 1727 levels, contributing a very large number of dimensions. As a predictor variable in a GLM, binarization would lead to 1726 separate coefficients to be fitted, many with very low exposures. Fitting the GLM would require long runtime and would overfit the low-exposure levels.

Your assistant creates a new variable, **mix**, based on **breed**, that indicates whether an animal is a mixed breed or purebred (not mixed breed). Run the code provided by your assistant and perform bivariate analysis of the new **mix** variable with other variables.

(b) (*5 points*) Describe an important finding that causes **mix** not to be usable as a predictive variable when using of all the data. Include a table or graph that illustrates the finding.

The best candidates recognized the change in mix of purebred and mixed breed animals by year and then explained why this finding may impact the ability of breed to be a predictor variable.

*Many candidates did not point out why the mix variable was not **usable** but rather why the mix variable was not expected to be **useful**. These candidates failed to receive full points but could receive more points for noting to the similarity of the distribution of purebreds and mixed breed animals on the target stay variable. The graphs supporting this position were expected to be well displayed with titles and axis, and the data was transformed to clearly show the similarities. The better commentary also mentioned the means, medians, and variability in support of their position.*

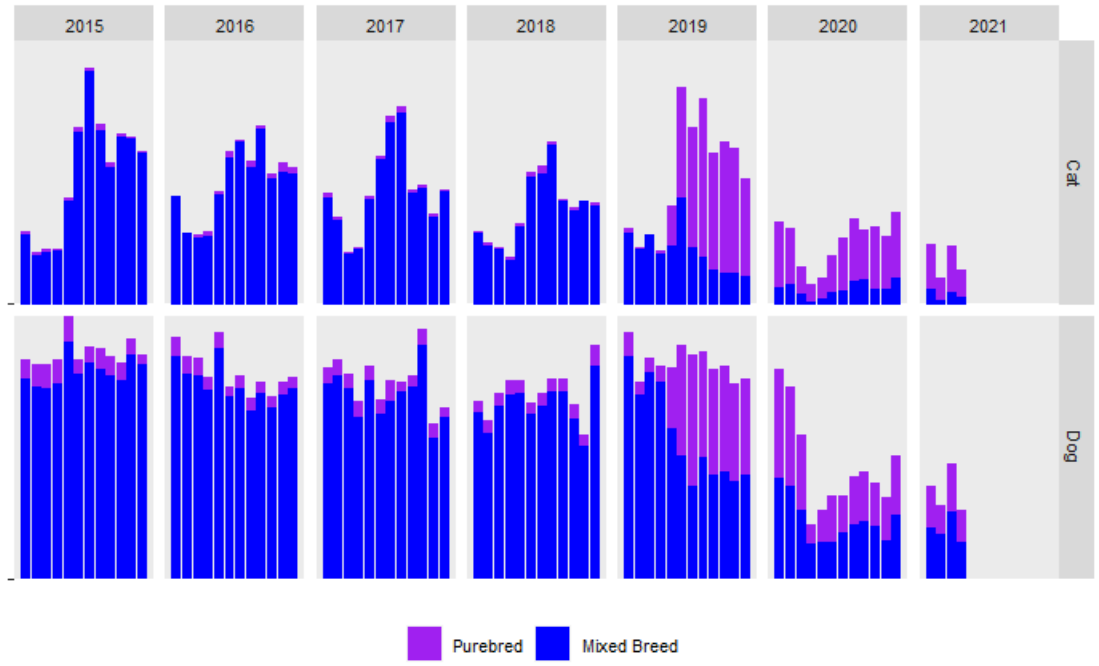
Candidates lost points if the axis for stay boxplots were not well scaled, due to difficulty of interpreting the results. Many candidates assumed these poorly scaled boxplots meant the target mean was 0 for both pure and mix, when it was actually around 27.

Little to no credit was awarded for responses which chose a reason that did not answer the question, did not include graphs that supported their position, or had graphs and commentary that were not clearly explained and left a lot to be interpreted by the grader, e.g., the reader does not know what "true" or "false" is in the data.

*Many candidates did not have any form of **bivariate** analysis completed, instead focusing on the unbalanced count of TRUE/FALSE in the mix variable. This was not considered a valid issue as both levels had significant numbers of observations and would not require oversampling/undersampling, which is commonly done on the target variable. Additionally, some candidates incorrectly compared counts by different variables like in.reason and assume the larger count of pure strays was indicative of correlation. This was incorrect as the distributions between in.reason and mix were all similar.*

ANSWER:

Table or graph: see next page



Important finding:

In the middle of 2019, the proportion of purebred animals suddenly increases, from a small percentage of animals to roughly half of dogs and more than half of cats. The most likely explanation is that how breeds were assigned changed at this time, rather than a sudden change in the animals arriving at the shelter. For example, the same dog or cat may have been categorized differently in 2020 compared to 2018. Because **mix** is not consistently defined over time, it cannot be used as predictor variable when using all of the data.

(c) (*4 points*) Describe two additional approaches for using the information in the original **breed** variable for a GLM.

The best candidates provided two approaches to reduce the levels that gave specific suggestions on how to use the information in the original breed variable. Examples included categorizing by size, reducing the number of levels by name, etc.

The question specifically requested answers using the original breed variable. Full credit was not given if the modeling methods were mentioned without any reference to the breed characteristics such as the examples mentioned above.

Many candidates lost points for suggesting modeling methods to be applied to breed, rather than adjusting the breed variable itself. PCA and clustering were provided as two common options, although neither are useful on a categorical non-hierarchical variable. Binarization and stepwise selection was also mentioned by candidates, but this option would result in recursive use of the target variable and does not address issues of data sparsity by level or computational considerations.

Additionally, some candidates suggested turning the breed variable into a numeric variable but lost points if no suggestions were made to deal with the lack of hierarchical classification present in the variable.

ANSWER:

One additional approach for using the information in **breed** is to isolate the most common breeds or words used in those breed names and group them by the appearance of those breeds or words. For instance, cats could be reduced to shorthair and not shorthair and dogs to breeds that appear at least 1000 times. This would reduce the number of levels to a usable number for a GLM without requiring additional knowledge but may easily group differently perceived animals together.

Another additional approach would be to supplement the data with information from other sources about the relative size or temperament of certain breeds. For instance, new predictors for small/medium/large animals and aggressive/passive animals could be built based on **breed** and the original variable dropped for a sharp reduction in the number of levels making it usable for a GLM, though this mapping requires additional knowledge and may be time-consuming to create.

Task 5 (*9 points*)

You ask your assistant to conduct principal component analysis (PCA) on some of the numerical variables in the data with the intention of creating features for the target of **stay**. Your assistant returns with the PCA coding as shown in the .Rmd file and suggests that you only keep the first principal component of the PCA to minimize dimensionality. Run your assistant's code on the data and inspect the output. Then do the following:

(a) (*2 points*) Assess the tradeoff between the assistant's suggestion, to keep only the first principal component, and including additional principal components.

Points were deducted if candidates did not adequately discuss the tradeoff (improved information retention vs. dimensionality, improved information retention vs. explain ability). Candidates who only mentioned one side of the tradeoff received partial credit. Some candidates showed the PCA output from R and/or recommended adding one or two principal components, but this was not required to receive credit. Being an assessment, a specific recommendation was not required for full credit.

ANSWER:

From the summary of the PCA, the first principal component only explains 45% of the variance that exists among the five variables included in the PCA. The second principal component explains almost as much of the variance, 42%, as the first and should also be included. If the third component were included, 99% of the variance would be explained while still reducing the number of dimensions by two, but it is less valuable than the first two principal components. The last two components add almost no value and should not be included.

(b) (*4 points*) Critique your assistant's PCA and make improvements to address the issues discussed. Include your revised PCA code (but not the output) in the space below.

Points were deducted for failing to suggest scaling the variables. Other critiques and improvements,e.g., PCA is most helpful for analyzing numeric variables or valid discussion of including or excluding certain variables, were also awarded partial credit.

ANSWER:

Critique:

The assistant's code did not scale the variables before running the principal component analysis. As a result, the variables **in.month** and **age** dominate the first principal component because they has the greatest variance among the variables included. Also, the assistant's code uses numbers to select the predictors, making it hard to see what it included in the PCA. The variable **out.intact** should not be included as that may not be known until **stay**, the variable to be predicted is known.

Revised code:

```
shelter.pca <- data.all.task5[, c("age", "in.month", "in.year", "in.intact")]
shelter.pca[2:4] <- lapply(shelter.pca[2:4], as.numeric)
s.pca.scaled <- prcomp(shelter.pca, center = TRUE, scale. = TRUE)
```

(c) (*3 points*) Interpret the loadings of the first two principal components of your improved PCA and discuss whether they would be sensible features for predicting **stay**. Include the loadings in the space below.

To receive full credit for interpretation, candidates needed to recognize that the magnitude (not the absolute sign) of loading coefficients indicates the differentiation in each principal component. For full credit, both sides of the contrast should be discussed, e.g., for PC1, old and unable to have offspring versus young and able to have offspring.

No credit was awarded for conflating PCA loadings (R function: pca$loadings) with the importance of components (R function: pca$importance).

*No credit was awarded for writing out equations such as "PC1 = -0.70 * age + …". These equations are correct but they are not an interpretation of the loadings.*

To receive full credit on the sensibility of features, candidates should opine on whether the PCs make sense from an interpretability perspective, not solely on the proportion of variance explained or other quantitative analysis.

ANSWER:

Loadings:

	PC1	PC2	PC3	PC4
age	-0.70111940	-0.01558925	-0.10549152	0.705024898
in.month	0.08814502	0.70358926	-0.70511514	-0.002290782
in.year	0.07944446	-0.71004635	-0.69844476	-0.041202852
in.intact	0.70310072	-0.02352233	0.06212169	0.707980911

Interpretation:

For PC1, large and opposite weights (positive and negative are arbitrary here) are given to **age** and **in.intact**, so it most strongly distinguishes younger animals still intact (kittens and puppies) from older animals not able to have offspring, a sensible distinction.

For PC2, large and opposite weights are given to **in.month** and **in.year**, so PC2 most strongly distinguishes animals arriving in early months of later years from those arriving in later months of earlier years. This distinction makes little sense in the context of predicting **stay**.

Task 6 (*10 points*)

To reduce costly long-term stays, the animal shelter plans to display with each cat and dog available for adoption the number of days the pet has been available for adoption and the typical time to adoption for that sort of pet.

Believing that just using average or median times for the typical time across all pets would be inadequate for this purpose, B applies a generalized linear model (GLM) using a Poisson distribution with log link function on the public dataset to predict **stay** + 1 (allowing the model performance measures described below to work on stays of zero days). Five such GLM models, differing only by their predictors, are set up.

Rather than rely on a single fitting for each model, each model is fit 200 times to directly observe how well that model will predict unseen data. Each fitting is called a trial for that model. For each of the 200 trials on a given model (set of predictors), the training data is a 20% random sample of adopted pets that came into AAC before 2020 (using **in.year**) and the test data was consistently all adopted pets from 2020 onwards.

The model performance is first measured individually for each record in the test data, calculating both the variance of the predicted results and the square of the bias (the difference between the average predicted and actual result). The predicted and actual results used are not **stay** + 1 but log(**stay** + 1), or the level of the linear predictor, so that the variance and average are less affected by the skewness of **stay**. Then, for each model, these record-level results are averaged across all test data records.

The process of fitting 200 trials, measuring performance for each record, and calculating the average performance over all records is repeated for each of the five models, each using the same random samples for the 200 trials, with the following results:

Model formula	Mean Variance	Mean Squared Bias
stay ~ age	0.0010	1.5805
stay ~ animal	0.0011	1.4856
stay ~ age + animal	0.0016	1.4786
stay ~ age + animal + mf	0.0021	1.4792
stay ~ age + animal + black	0.0019	1.4767

The **black** predictor is derived from the **color** variable, being 1 if “Black” and 0 otherwise. Your assistant comes to you wanting to better understand B’s methodology.

(a) (*2 points*) Explain what the variance and bias values indicate about the relative quality of predictions when comparing predictive models.

Most candidates were able to identify the difference between bias and variance but some failed to relate these concepts to the quality of the predictions or accuracy. Some candidates described the change in bias and variance of the actual model outputs in the exercise, not the general theory. Successful candidates clearly differentiated that variance is focused on the training data and that the reduction in bias is what leads to high quality predictions.

ANSWER:

The variance figures indicate how much the predictions vary depending on the training data used. As more predictors are used, the variance increases because the model more precisely fits the training data for each trial and becomes less generalized. The bias figures indicate how close expected predictions and actual results are on unseen data. Generally, as more predictors are used, the bias decreases as more accurate predictions are made.

(b) (*2 points*) Calculate, for the first model listed, the typical errors up or down from the true value due separately to variance and bias for predictions of **stay** + 1.

Well prepared candidates were able to recognize that the variance and squared bias in the table were calculated at the linear model level and the nature of the log link required exponentiation. However, many candidates seemed unsure of what the question was driving at leaving the question blank or simply entering the variance and squared bias from the table.

ANSWER:

The variance and squared bias figures above are calculated at the level of the linear model. For the predictions of **stay** + 1 itself, this means, using the results of the first model, that the typical error due to variance is a exp(0.0010^0.5) = 103% factor up or down from the predicted value and the typical error due to bias is a exp(1.5805^0.5) = 352% factor up or down.

(c) (*3 points*) State two reasons why bias, as calculated here, may not always decrease with additional degrees of freedom, as seen with the model that adds **mf**.

Many candidates were able to identify that non-random splits between the test and the train datasets could lead to bias not being reduced as the calculation occurs on the test set. Very few candidates identified the distinction that minimizing squared bias is not the same as minimizing the objective function in the case of a Poisson GLM. Other responses that were accepted include adding variables with no predictive power or substantial collinearity between variables.

ANSWER:

One reason the bias calculation may not always decrease is that it is calculated on test data whereas the models are trained on training data—to the extent these are different, particularly when not split randomly, adding predictors may not improve the accuracy of the predictions. Another reason is that, in this case of fitting a Poisson GLM, minimizing the squared bias given the training data is not the same as minimizing the objective function.

(d) (*3 points*) State which predictors should be selected based on the above data, putting them in order from more to less predictive. Explain your selection and ranking.

Successful candidates were able to order the importance of the predictors correctly and justify the ordering using the composite of variance and squared bias as the determining factor. Most candidates tended to get the order right but some didn't explicitly note that mf should not be included in the model.

Some candidates used either just variance or squared bias as the determining characteristic for ranking and exclusion. No credit was awarded for ranking entire models instead of model predictors.

ANSWER:

Predictors to select from more to less predictive:

animal, **age**, **black**

Explanation:

Model formula	Variance + Squared Bias
stay ~ age	1.5815
stay ~ animal	1.4867
stay ~ age + animal	1.4803
stay ~ age + animal + mf	1.4814
stay ~ age + animal + black	1.4786

To select the best model, the sum of the variance and squared bias errors should be used. Where this sum is lowered by including a variable, the variable should be included, and the greater the reduction, the more predictive the variable. On their own, **animal** has a lower sum error than **age**, making **animal** more predictive than **age**, which subsequently reduced the sum error by 1.4867 - 1.4803 = 0.0064 once **animal** was already included in the model formula. As **black** reduced the sum error by a further 1.4803 - 1.4786 = 0.0017, it is less important than **age**, establishing the order.

Task 7 (*8 points*)

The animal shelter has asked for a simple model to identify animals that can be transferred. Your assistant builds decision trees for this purpose and performs cost complexity pruning on the tree.

(a) (*3 points*) Explain, for a general audience, what cost complexity pruning does.

Well prepared candidates recognized that a complex decision tree is likely to perform drastically differently on different datasets and discussed both decision trees and how cost complexity pruning reduces the size/number of nodes while adding sufficient predictive power. No points were awarded for stating "complexity pruning" prunes a tree without explaining complexity pruning.

ANSWER:

In general, cost complexity pruning is part of a two-step approach to building a tree model. The first step is to build a large, complex decision tree, which is essentially a flow chart for deciding whether to try to transfer an animal. Because this flow chart may be too specific to the data used to create it and not generalize well to new data, a second step called pruning is taken. Pruning reduces the size and complexity of the initial flow chart to a more useful one. That it is called cost complexity pruning has to do with the technical tradeoff being made between how simple the flow chart is compared to how well it distinguishes whether animals can be transferred or not.

Your assistant produces the table below with the error rate, sensitivity, and specificity for the training and test data on the unpruned and pruned trees:

	ErrorRate	Sensitivity	Specificity
unpruned.train	0.219	0.107	0.984
pruned.train	0.225	0.052	0.992
unpruned.test	0.232	0.073	0.974
pruned.test	0.229	0.039	0.988

(b) (*3 points*) Recommend whether the unpruned or pruned tree would be <u>preferable</u> to use. Justify your recommendation in the context of the business problem.

Points were deducted where a recommendation was not clearly stated, as the question called for a recommendation which should be clear. Many candidates lost points because they explained in detail what sensitivity, specificity and error rates are without clearly stating which one is important and which approach the candidate is recommending.

In the context of the question the use of test data is the relevant metric; many candidates lost points for recommending training data or not recommending a test metric at all.

ANSWER:

Of these three metrics, the shelter should be most interested in sensitivity, which measures what percentage of transfers the model captures. Between train and test data, the test data should be used as it better simulates how the animal shelter will use the model. The unpruned tree performs better (higher) on the sensitivity metric on the test data, so it would be preferable to use among the two. I recommend using the unpruned tree.

(c) (*2 points*) Explain why neither the unpruned nor the pruned tree will be a satisfactory model for the local animal shelter to use.

Well prepared candidates noticed that the tree models identified less than 10% of the animals which were eventually transferred. Many candidates attempted to answer a different question than was asked.

ANSWER:

The unpruned tree, the preferable of the two trees, will not be satisfactory for animal shelter use because, when tested on unseen data, identified less than 10% of the animals that were eventually transferred. This is exceptionally poor model performance, enough so that the animal shelter would likely be more successful identifying potential animals to transfer by other means than to apply this predictive model.

Task 8 (*13 points*)

Your assistant seeks to fit a boosted tree to identify animals that can be transferred.

Your assistant wants to set eta as high as possible to reduce runtime.

(a) (*3 points*) Explain why you disagree with your assistant about setting eta as high as possible.

Many candidates were able to link eta with learning speed, but many failed to be able to explain the inherent tradeoff. No points were awarded for candidates who mistakenly think eta as being like a complexity parameter or mention unrelated decision tree splits.

ANSWER:

In boosting algorithms, which work by iteratively fitting a model the residuals of a prior learner, eta, also called the learning rate or shrinkage parameter, slows down the model fitting process so that the residuals from the prior learner do not have too large an influence on the final model. With eta at its maximum of 1, each model iteration is the prior learner plus the model fitting its residuals. While this will run quickly, it will be prone to high variance, overfitting the training data and not generalizing well to unseen data. Setting eta to less than 1 slows down the fitting process by only adding eta times the model fitting the residuals to form the next learner and will substantially reduce the variance.

Your assistant then needs help setting eta.

(b) (*3 points*) Explain cross validation and how it can be used to set the eta hyperparameter.

Both a general explanation of cross validation and an explanation of its specific application in setting a hyperparameter were needed to receive full credit. Many candidates were able to explain the general procedure of cross validation but failed the link this to how to set the eta hyperparameter.

Many candidates missed the explanation about how cross-validation error calculated. Less prepared candidates mistook cross validation as being equivalent to a general data partition, just dividing the data to training and testing sets.

ANSWER:

Cross validation divides the available data into multiple folds for a series of model fitting runs. In each model fitting run, the model is fit to data from all but one fold, and a test metric is calculated on the data not used in the fitting for that run. Across the model fitting runs, each fold is used as test data exactly once. The average test metric across the runs is the result of the cross validation.

To use cross validation to set the eta hyperparameter, a series of reasonable values for eta would be chosen beforehand. Then, for each value of eta, cross validation would be performed with each model fitting run using the same eta. The result is one average test metric result from each cross validation for each value of eta. The value of eta with the superior test metric, some measure of predictive power on unseen data, would be chosen for subsequent predictive modeling work.

The assistant uses the eta found via cross validation and calculates an AUC of 0.628 for the boosted tree, claiming that this demonstrates excellent model performance.

(c) (*3 points*) Explain, for a general audience, what AUC represents.

A full-credit explanation, despite being for a general audience, considers how AUC represents not just a single set of classification predictions but the spectrum of such predictions. Points were deducted for stating a high AUC leads to a better model without any justification.

ANSWER:

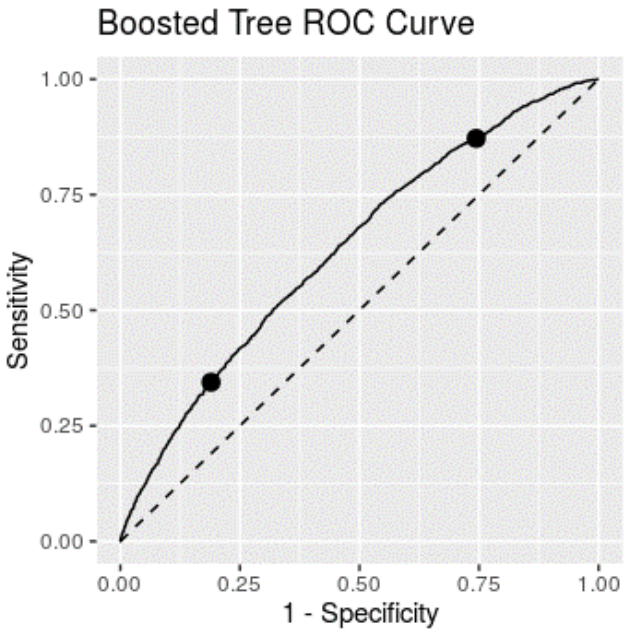
AUC, which stands for "area under the curve," is a number that represents how well a particular classification model (as it is called) distinguishes one outcome from another. The number combines the efficacy of many possible interpretations of the classification model, each interpretation leaning more heavily towards one outcome or another. When comparing the AUC of two models, a higher AUC indicates a better predictive model. As a baseline, a model that randomly guessed one outcome or another based on how common each one is, akin to a coin flip, has an AUC of 0.5, so only models with AUC above 0.5 are desirable.

(d) (*2 points*) Critique your assistant's assessment of the model performance.

Well prepared candidates realized the AUC of 0.628 needs to be compared relative to another model and were able to relate the task back to the business problem.

ANSWER:

While an AUC of 0.628 does indicate that the boosted tree is expected to perform better than random guessing (an AUC of 0.5), it only indicates that and does not by itself indicate an excellent model. The performance of the boosted tree at a particular cutoff point needs to be evaluated against other candidate models using a test metric more specifically aligned to the business problem to establish its excellence.



You present the above ROC curve to B, who then circles two points on the curve to investigate further.

(e) (*2 points*) Explain how the same model can be used to produce both points on the ROC curve.

Well prepared candidates recognized the ROC curve is a spectrum and were able to articulate how each point could be generated by the same model. Points were deducted if candidates defined sensitivity and specificity without linking back to how the same model can produce the same points. Some incorrectly thought the different points were produced by the same model with different hyperparameters.

ANSWER:

The fitted boosted tree can be used to produce, for each record being tested, a probability that the animal will be transferred. Then, the modeler separately decides the lowest probability, called the cutoff value, that will convert probabilities equal to and higher than that cutoff value to a prediction of transfer. Moving from left to right on the ROC curve represents all possibilities of the cutoff value, from 1 to 0, before this decision has been made. The two points represent using different cutoff values, one lower and one higher, to determine class predictions from the probability predictions of the same predictive model.

Task 9 (*9 points*)

Your assistant creates a GLM called **glm_start** on training data to predict whether an animal is adopted and then runs the drop1 function on it. Refer to the assistant's code in the .Rmd file.

(a) (*3 points*) Create a new model called **glm_drop** based on the results of the drop1 function. Justify your predictor variables based solely on the drop1 results. Include the code that creates **glm_drop** in the space below.

Well prepared candidates correctly understood the drop1 function was evaluating the AIC of several models fit after dropping one variable. The model which dropped in.month resulted in the model with the lowest AIC and should be selected.

Full points were frequently achieved with short and effective answers – one or two sentences.

Common point deductions included (but were not limited to):

- *Thinking that a higher AIC is better;*
- *Recognizing that lower AIC is better, but incorrectly interpreting the drop1 results (e.g. dropping in.reason because "it has the highest AIC, so removing would reduce the AIC", or interpreting drop1 as consecutive, "so the model with only in.month remaining has the lowest AIC")*
- *Ignoring AIC – using p-values and df to justify dropping a predictor.*

ANSWER:

Code to create glm_drop:

```
glm_drop <- glm(formula = adoption ~ animal + mf + age + in.reason +
in.intact,
  data = df_train,
  family = binomial(link = "logit")
)
```

Justify your predictor variables:

The drop1 function compares the AIC of keeping all the predictor variables to that from keeping all but the specified predictor variable. Dropping a variable always increases the deviance but, in the case of **in.month**, that increase is not enough to justify the addition of 11 degrees of freedom, as indicated by the lower AIC of 47,148 compared to 47,156 for the model with all predictor variables. All the other AIC values are higher than 47,156, so only **in.month** should be dropped, as shown in the code above.

Your assistant also creates a model using LASSO, creating a model called **glm_lasso**.

(b) (*3 points*) Contrast the two methods, drop1 and LASSO, for selecting predictor variables.

Many candidates struggled with this subtask. Candidates were asked to contrast the two modeling methods; however, some candidates contrasted the two modeling outputs. For example, they noted that LASSO removed in.month, mf, and in.intact, while drop1 only removed in.month. Such answers earn zero points.

Some candidates only described one method (LASSO), without mentioning drop1 (or vice versa).

To earn full credit, candidates needed to include both LASSO and drop1 methods with respect to model differences. Accepted differences included (but were not limited to):

- *Drop a single variable per iteration and risk being stuck at a locally optimal solution (drop1) vs. assess all predictors in concert (LASSO)*
- *Dropping entire predictors vs. reducing coefficients (potentially to zero, effectively dropping the predictor)*
- *Differences between penalties on coefficients to dropping predictors*
- *Automatic binarization of categorical variables and ability to drop levels of that variable (LASSO) vs. dropping entire categorical variables (drop1) – unless binarized before running drop1*
- *Manual process to remove a predictor (drop1) vs. automated (LASSO).*

The question asked candidates to contrast methods. Higher-scoring candidates made a direct contrast, where lower-scoring candidates defined LASSO and drop1 but did not directly point out the differences between the two methods.

ANSWER:

Drop1 shows the AIC impact from individually removing each predictor variable. The modeler removes the predictor that produces the largest drop in AIC, and then iterates until no more predictors should be removed.

LASSO uses a penalty in the optimization function that penalizes large coefficients in the model. As a result, the coefficients are pushed towards zero, and can be set to zero, effectively removing the predictor.

The differences include:

- Drop1 requires the modeler to manually remove the predictor. LASSO automatically removes predictors.
- Drop1 removes the entire categorical variable. LASSO binarizes categorical variables and can remove individual levels.
- Drop1 removes one predictor at a time. LASSO assesses all predictors in a single model fitting.

(c) (*3 points*) Recommend whether to use **glm_drop** or **glm_lasso** based on AUC results and which predictors the models use. Justify your recommendation. Display the AUC results used in the space below.

Either model could be recommended for full credit as long as it had a clear, multi-faceted justification. Well prepared candidates acknowledged predictive performance was one element of a model recommendation, while less well prepared candidates made a recommendation solely on quantitative model performance.

Common mistakes included incorrect interpretation of AUC and stating that one model is more interpretable without an explanation.

ANSWER:

AUC Results:

Model	Test AUC
glm_drop	0.7103
glm_lasso	0.6960

Recommendation and Justification:

I recommend the **glm_lasso** model.

glm_lasso is a simpler model. It only considers whether the animal is a cat or dog, its age, and whether it arrived via Public Assist. **glm_drop** includes three additional predictors, making it more cumbersome to explain to a non-technical audience at the animal shelter.

The higher AUC suggests that **glm_drop** is slightly better at classifying adoptions in this train/test data partition. However, the **glm_lasso** model has fewer predictors, protecting against overfitting and adding confidence that the model performance will be stable with unseen data.

The interpretability and robustness of **glm_lasso** outweigh the slight decrease in predictive performance.

Task 10 (*11 points*)

An accurate estimate of the length of stay would aid the shelter in planning and budgeting. The shelter is uncertain whether the effects of the pandemic will continue into 2022. Therefore, they would like to predict **stay** under two different assumptions:

1) Activity returns to pre-pandemic levels (prior to March 2020)
2) Activity remains at the levels seen since March 2020

Your assistant creates a feature called **pandemic** to distinguish data as described above and wants to build a GLM including this feature but is unsure what distribution to use.

Your assistant says that because we only care about the predicted mean, the choice of distribution does not matter as long as its domain includes all values of the target variable.

(a) (*3 points*) Explain why your assistant's statement is incorrect.

There are several ways to describe the concept that the distribution strongly affects the model outcome due to how prediction errors are measured. Talking about deviance or loglikelihood, referring to the error or loss function being optimized, or providing a clear counterexample to the assistant's statement are all valid approaches. Exceptional candidates explained why the assistant's statement was incorrect by referring to the GLM model fitting process and how a distribution with a shape that did not closely align to the target variable would result in large contributions to model deviance or loglikelihood.

Most candidates received at least partial credit. Candidates did a good job of explaining that the assistant's statement was incorrect because you would also want your distribution to match the shape of the target variable, especially for a skewed variable. Candidates also identified that you would typically want the distribution to match the target variable in terms of whether it is discrete or continuous.

Many candidates earned credit by pointing out that a distribution's domain could include all values of the target variable but still be inappropriate for modeling the target variable as it could allow for impossible values (e.g., using the normal distribution for a target variable that is strictly non-negative). Some candidates correctly identified differences in variance between different distributions that happen to have the same mean as potential issues or possible non-convergence of the model due to incorrectly selected distribution.

The most common cause of low or no-credit responses were vague or very brief answers that only touched on a single point or did not clearly describe why the statement was wrong.

ANSWER:

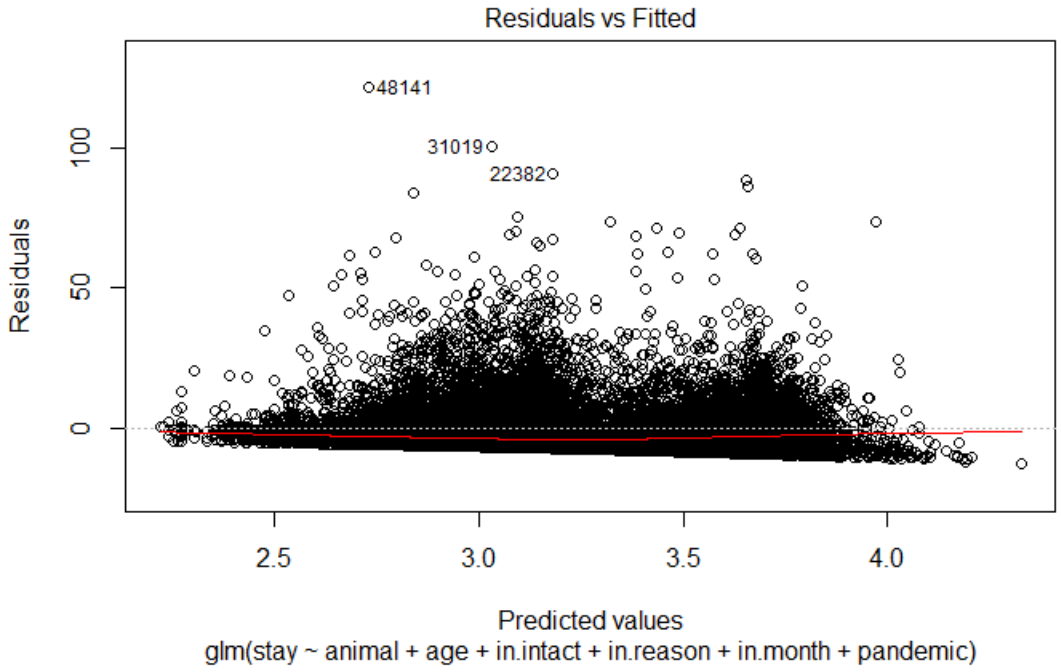
My assistant is incorrect because two models fit by two different distributions on the same data can make substantially different predictions, even if the mean of all predictions happen to be the same.

Different choices of distribution for a GLM mean different choices of how prediction errors are measured on the data being fit to the GLM. Where one distribution may find a large difference between actual and predicted values for a particular record only improbable and assign it a moderately large deviance (the transformed measure of likelihood whose sum across records is being minimized when fitting a GLM), another distribution may find that difference nearly impossible and assign it an extremely large deviance which ultimately leads the iterative fitting process of the GLM to lean more towards

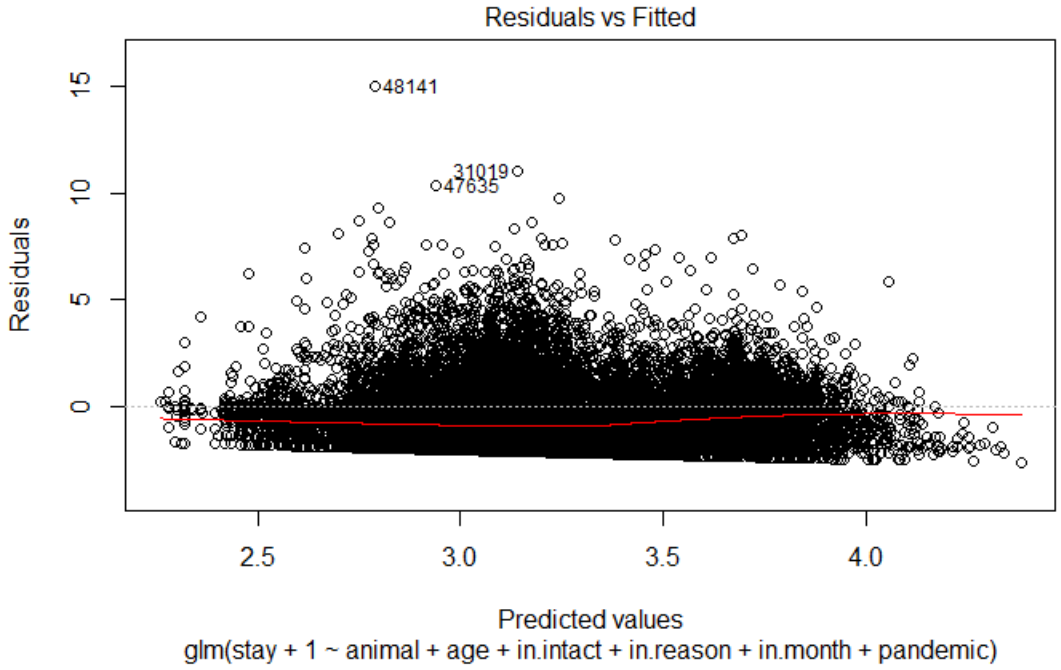
reducing the impact of this record. For example, a distribution where the variance is proportional to the mean will better tolerate larger prediction errors occurring with larger predictions than one whose variance is independent of the mean.

Your assistant runs two models with different distributions, one using Poisson and one using gamma, using the log link function for each, and presents you the deviance residual plots below:

Poisson:



Gamma:



(b) (*3 points*) Describe what the deviance residual plot for the gamma distribution shows about the model fit.

Candidates generally did a good job of identifying ways in which the gamma deviance residuals plot potentially violated assumptions such as: exhibiting more variability in residuals above 0 than below 0, the two humps in the graph which may violate the homoscedasticity assumption, resemblance to a normal distribution, and the smoothed red line declining below 0 in the range from 2.5 to 3.5 for predicted values.

Candidates received only partial credit for brief answers that only commented on one part of the deviance residual plot. Vague or very brief answers which only identified the gamma model as a good fit or poor fit but did not explain why the model fit was good or poor received no credit.

Candidates generally did a poor job of demonstrating an understanding of deviance residuals. Many candidates incorrectly described the residuals in the deviance residual plots as the difference between actual and predicted value, which is true for linear regression residuals but not for deviance residuals. Deviance residuals are a generalized version of residuals for GLMs that compares the model's loglikelihood to the maximum achievable loglikelihood under the perfectly fit saturated model.

Many candidates misinterpreted the meaning of positive vs. negative residuals. For linear regressions, the residual is defined as the actual value minus the predicted value. This means that the large positive outliers are underestimates (i.e., the predictive value is much less than the actual value), not overestimates. For GLMs, the deviance residual is calculated by multiplying the sign of the residual by the difference of model loglikelihood and saturated loglikelihood, and so positive deviance residuals are also underestimates.

Responses that compared the model fit of the gamma plot to the Poisson plot tended to receive little credit for this section as the intent was to assess the meaning of the deviance residual plot for the gamma distribution, not to compare the plots.

Exceptional candidates described the meaning of the deviance residual plot by demonstrating an understanding of deviance residuals, identifying the key assumptions that should be reviewed to assess the fit of a model using the plot, and commented on how the gamma model met or did not meet those model assumptions.

ANSWER:

The deviance residual plot for the gamma distribution shows how much each individual record contributed to the deviance, which was minimized when fitting the GLM. Each circle is plotted at the log of the predicted length of stay (plus one, as **stay** has zero values and gamma can only be evaluated at strictly positive values), and the smoothed red line shows typical deviance residuals at different lengths of stay.

In general, predictions for longer stays were slightly poorer, on a multiplicative scale, than predictions for shorter stays. The highest circles indicate actual stays that were most unlikely (longer or shorter) given the predicted stay. These tend to be more less common with the more extreme predictions, indicating the model is capturing breadth of the distribution of stays fairly well.

(c) (*2 points*) State which distribution is preferable. Explain your choice.

The key observation with these similar looking deviance residual plots is the smaller scale for the gamma distribution. This is an absolute scale when fitting to the same data—smaller deviance is better.

Candidates that identified the large difference in the scale of the residuals and recommended the gamma distribution earned full credit.

Many candidates recommended the gamma distribution and provided a strong justification unrelated to the scale of the residuals, such as the AIC of the gamma distribution being lower or that it violated fewer of the assumptions for the distribution of the deviance residual plot. These answers received partial credit

Candidates that recommended the Poisson distribution because it matched the structure of the stay variable as discrete counts earned partial credit.

Responses that recommended the Poisson distribution because the red line representing the average residuals for each predicted value was closer to 0 than for the Gamma distribution were not accepted and earned no credit. This is because the scale on the Poisson graph is nearly 10x the scale on the gamma graph and if the Poisson distribution was plotted on the scale of the gamma graph, the average residual line would be further from 0 for the Poisson distribution.

ANSWER:

The gamma distribution is preferable, because the range of its deviance is on a much smaller scale, indicating that its predictions are more plausible, at least for the actual data it was fitted to. It is not

possible to say which model has the lower total deviance based on these plots, but the gamma distribution has substantially less variation in its goodness of fit.

(d) (*3 point*s) Fit a GLM on all data using your preferred distribution and its model formula as seen in the deviance residual plots above. Provide the code that fits the GLM (but not the output) in the space below. State the coefficient for **pandemic** and calculate the effect of being in a pandemic on length of stay based on the GLM.

Candidates generally did the best on this part of the question, with many candidates earning full credit.

The most common error on the coding section was using a different model formula than the one specified in the deviance residual plots above (e.g., a Poisson model using stay + 1 as the target variable or including the mf (Male/Female) variable in the model formula. A few candidates were unable to run the model due to errors in their code, which made them unable to derive the pandemic coefficient or the effect of the pandemic on the length of stay. A typical cause was attempting to fit a gamma distribution on stay rather than stay+1. Candidates who chose gamma in part (c) and then fit a Poisson in part (d) because of this still received credit for part (d).

Most candidates correctly identified the coefficient for the pandemic variable. In the case that a candidate mis-specified the model formula, they were still able to earn credit for correctly identifying the pandemic coefficient in the model output. A few candidates mistakenly reported the p-value for pandemic rather than the coefficient.

Most candidates correctly identified the effect of being in a pandemic on length of stay as a <u>percentage</u> increase in length of stay equal to e^(coefficient for pandemic) − 1. Candidates also received credit for correctly identifying it as a multiplicative factor.

The most common error on interpreting the effect of being in a pandemic was to interpret the effect as an addition to the length of stay rather than a percent increase or multiplicative factor.

ANSWER:

Code to fit GLM:

```
glm_pandemic <- glm(
  formula = stay + 1 ~ animal + age + in.intact  + in.reason + in.intact +
in.month + pandemic,
  data = data.all.task10,
  family = Gamma(link = "log")
)
```

Coefficient for pandemic:

0.050667

Effect of being in a pandemic on length of stay:

Due to log link, the effect of being in a pandemic is a exp(0.050667) = 105% multiplier on **stay** + 1.