Exam PA December 14, 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 11 tasks numbered 1 through 11 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 (*9 points*)

(a) (*3 points*) Explain two reasons why predicting **stay** is not sufficient for addressing planning and budgeting for the animal shelter.

Candidates generally performed well on this subtask. To receive full credit, candidates needed to address the business problem in their response.

ANSWER:

The target variable **stay** cannot be used by itself to determine the number of animals at any time in the shelter, only how long animals stay once they have arrived. Also, additional interpretation of our predicted results, such as how the number of animals translates into specific expenses, will be required to solve the direct business issue for the animal shelter.

(b) (*1 point*) Identify the target variable for a second predictive model (in addition to the model predicting **stay**) that could be developed using the given data to provide a more complete picture for planning and budgeting.

Candidates generally performed well on this subtask. This is an open question hence acceptable answers are not limited to the model solution. Another acceptable answer was the way in which animals leave the shelter (Adoption, Transfer, Return to Owner).

ANSWER:

The number of animals arriving each month.

(c) (*5 points*) Write a problem statement, written for a general audience, that incorporates both models for the purpose of planning and budgeting.

Few candidates received full credit on this subtask. Well prepared candidates noted the models should be assessed and updated for the purposes of planning and budgeting. Points were deducted for failing to incorporate both models and failing to write for a general audience.

ANSWER:

The local animal shelter wants to estimate the number of animals at the shelter at least one month in advance for budgeting and planning purposes. Two predictive models will be developed to support this estimate. The first model will estimate the number of animals arriving at the shelter in the upcoming month given arrivals in recent months. The second model will estimate the length of stay for animals that have already arrived. Predictions from these models along with assumptions about the composition of animals arriving in the next month will be used for the estimate. Data from a similar shelter, the Austin Animal Center, will be used to train the initial predictive models, but subsequent models will be trained on a periodic basis using data from the local animal shelter. The accuracy of each model will be monitored based on a performance metric yet to be developed in consultation with the local animal shelter.

Task 2 (*7 points*)

Run the starter code and code for this task in the .Rmd file to see an overview of the data.

(a) (*4 points*) Recommend two elements to add to the data dictionary to improve it. Justify your recommendations.

Most candidates suggested adding two variables instead of two elements to the data dictionary. Partial credit was awarded for suggesting new variables. Accepted responses included (but were not limited to) adding descriptions to variables, cleaning up data, adding statistics such as mean, median, and percentiles to data dictionary, and distinguishing numerical and factor variables.

ANSWER:

The data dictionary could include documentation on what cleaning or validation has been performed compared to the original data source so that users could understand what other data may be available and how the current data differs from the source data.

The data dictionary could include contact information for the person or group responsible for the data so that questions or concerns with the data may be addressed.

(b) (*3 points*) Describe two reasons why **name** cannot be used in a predictive model for the animal shelter.

Candidates generally performed well on this task. Only partial credit was awarded when failing to justify any reasons given. Accepted responses included (but were not limited to) name having a large number of factor levels, unseen names being encountered in the future, and the questionable link between name and length of stay.

ANSWER:

First, if **name** is treated as a factor variable, with each unique name treated as an independent predictor, then the large number of levels, over a thousand, will lead to overfitting of the predictive model.

Second, any model trained on previously seen animal names cannot be applied when a previously unseen animal name is encountered. While this could be treated as “missing”, there is little confidence that the presence of a new name would be similar to prior situations where the name was deemed “missing”.

Task 3 (*11 points*)

Your boss, B, has asked you to develop a simple visualization to get a sense of the impact of the COVID-19 pandemic on monthly cat adoptions in 2020. B is interested only in cats that were ultimately adopted and is only interested in total adoptions per month, not details such as breed, color, etc. Your assistant has completed some initial data work to summarize the monthly cat adoption data B asked about but is not sure how to create graphics using R.

Use the assistant's data to complete the items below to assist B with the visualization. As B is interested in a quick, understandable visualization, complex formatting or color schemes are not required.

(a) (*3 points*) Create a graph that shows the number of cat adoptions per month in 2020. Paste the code used to create the graph as well as the graph below.

Many candidates used default axis labeling and did not include a title to the graph. Clear axis labeling is always useful when preparing a graph to make sure it is clear and easily understood. A title is not always necessary if the axis labeling is sufficiently detailed but is generally a good idea to include as well. This graph is focused on 2020 data, so a reference to 2020 should appear somewhere in the graph, either in the title or in the x-axis label.

Most candidates used a bar graph for this question, which was a good choice because the y-axis scale automatically starts at 0 for bar graphs. The y-axis in this case is a count variable comparing absolute counts by month so the y-axis should start at 0. Most candidates who did not use a bar graph used a dot plot instead, which does not automatically start at 0 on the y-axis skewing the relative size of the counts making the graph misleading. Well prepared candidates recognized this and overrode the default y-axis scale to begin with 0.

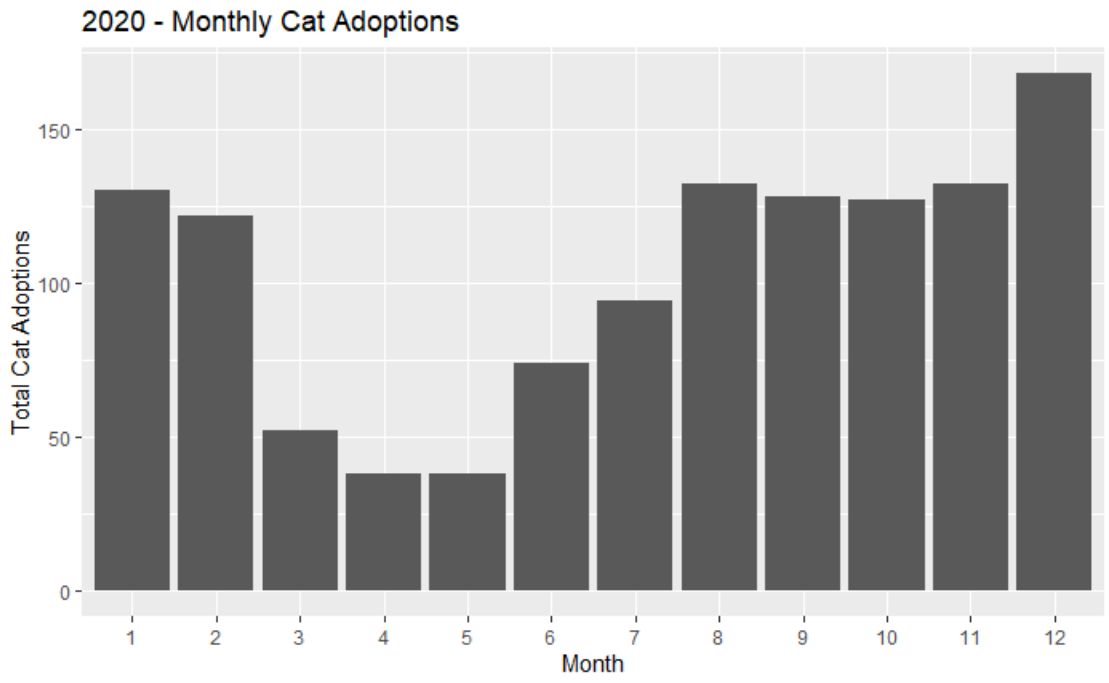
Some candidates faced issues copying graphs to the response template. In this case points were awarded based on the R code provided in the response.

ANSWER:

Code:

```
p <- ggplot(data = df_temp, mapping = aes(x=out.month, y =
Total_Cat_Adoptions))
p + geom_col() + labs(x = "Month", y = "Total Cat Adoptions", title = "2020 -
Monthly Cat Adoptions")
```

Graph:



(b) (*2 points*) Explain to your assistant why the vertical axis on a bar graph should include zero while the axis for a scatterplot or line graph may or may not include zero.

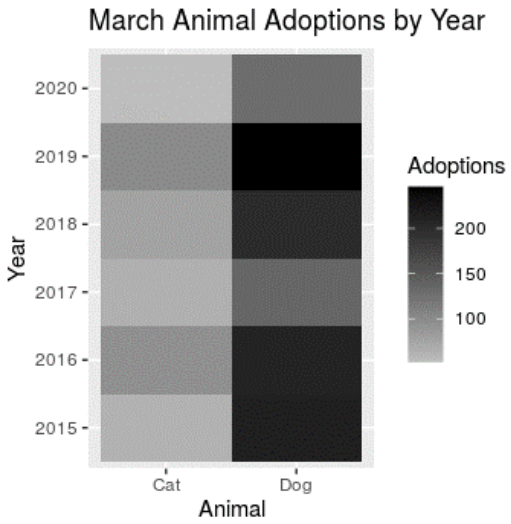
Many candidates included accurate descriptions of bar graphs showing that they understood their purpose and structure but did not adequately explain why the y-axis must start at zero. Many responses included an explanation that bar graphs should have a common base to accurately compare bars but did not explain what exactly that means. Well prepared candidates noted that failing to use a base other than zero skews the interpretability of the bars' heights.

A lot of candidates also explained that zero is needed on the y-axis because a count variable can potentially be zero. While true, this is not the reason why bar graphs in particular must start at zero, so no credit was given for this justification alone.

Some candidates interpreted the question as asking why the x-axis should start at zero, which was not given credit due to a misunderstanding of the terminology and structure of a bar graph.

ANSWER:

A bar graph should begin at zero so that the relative areas of the bars are proportional to the values they represent. A dot plot or line graph may or may not start at zero depending on what the designer intends. As it is the position relative to a labeled axis that indicates a specific value rather than its length or area, there is no explicit need for the height of a point or line above the x-axis to be directly proportional to its value.



B identifies a large decrease in cat adoptions starting in March 2020 and asks the assistant to create a graph comparing cat adoptions each March from 2015 through 2020 to assess whether adoptions are always low in March or if this change is due to the pandemic. The assistant produces the graph above using some R code found online.

(c) (*3 points*) Critique two aspects the assistant's graph.

Most candidates identified that the choice of heat map (in particular the shades of grey) was not clear and difficult to interpret and that the inclusion of dogs was not necessary and made the graph harder to understand for cats.

Other accepted critiques that many candidates made included the following:

- *The choice of heat map would have been improved if the actual adoption values were included in the graph as well*
- *Year displayed vertically is less clear than displayed horizontally*
- *Only showing March values does not give a full understanding of how March compares to other months in the year*

ANSWER:

First, a heatmap is an inappropriate choice for this task as the differences in the relative colors are difficult to discern.

In addition, the chart also includes irrelevant data on dog adoptions that skew the colors used to represent different values and draws attention away from the key issue of cat adoptions.

(d) (*3 points*) Recommend (but do not create) a graph that would be more effective in helping B make the desired assessment. Justify your recommendation.

Most candidates suggested using a bar graph and provided a decent justification for why this is a better choice than the heat map used above. Stronger candidates also suggested the data be limited to cats

only. Some candidates did suggest that separate bars be used for cats versus dogs, which is a good suggestion, but to receive credit for this suggestion, an explanation is also needed that showing separate bars will allow the assistant to isolate the trends for cat adoptions only.

Other responses that received credit included:

- *Bar graphs by month from 2015 to 2020, with March highlighted a different color for ease of review*
- *Line graphs by month with different lines and colors for each year to review seasonality*
- *Using March adoptions as a percent of available animals instead of absolute adoptions*
- *Using March adoptions relative to the average of January and February adoptions of that year*
- *Using March adoptions relative to the average of non-March months of that year*
- *Using a stacked bar graph where the x-axis as month and the stacked amounts are years*

Some candidates suggested using a boxplot, which was not applicable to this problem and received no credit. The data is summarized by year already so there is no distribution within each year to show for this problem.

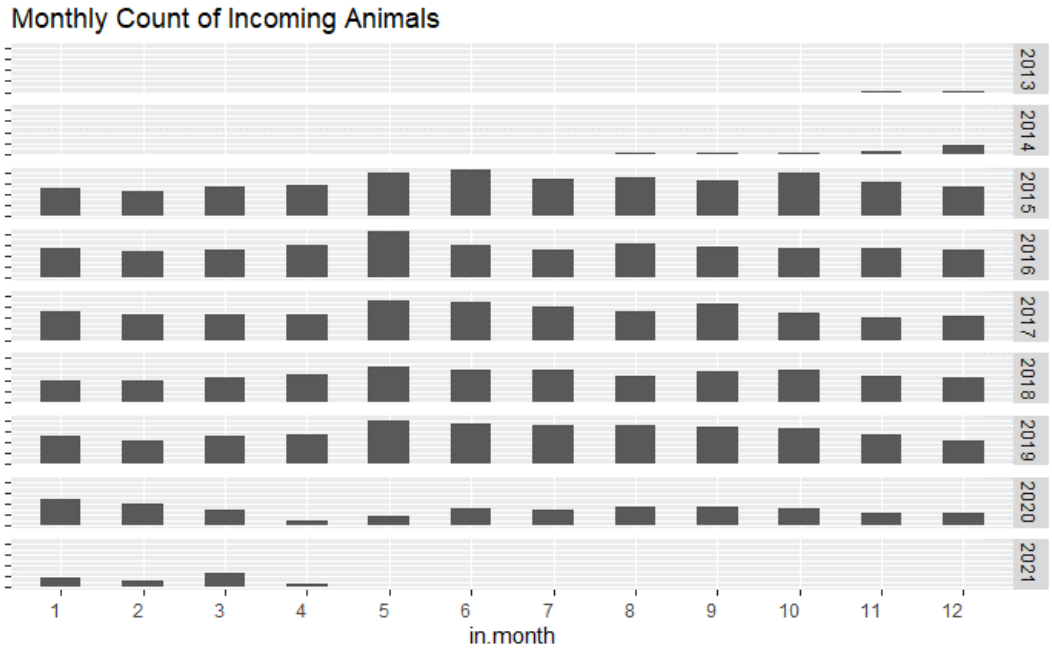
Some candidates included suggestions in 3c above but did not carry those suggestions to this problem as well. Candidates who referred their response in 3c when answering 3d received credit for those suggestions but candidates who did not reference 3c did not receive credit. The exam instructions state that each question should be considered independent from others.

ANSWER:

To better show the March 2020 cat adoptions compared to previous years, I recommend a bar chart where, for each month, the ratio of the cat adoptions for that month in 2020 compared to the average of cat adoptions for that month from 2015 through 2019 is displayed. This would allow B to more easily determine if the lower adoption figure seen in March 2020 is specific to March adoptions every year or unique to the onset of the pandemic in 2020.

Task 4 (*7 points*)

B is uncertain whether the entire range of dates should be used or only a subset after seeing the following graph:



(a) (*3 points*) Describe, for a general audience, the effect on predictions for **stay** from including the earliest arrivals in the dataset.

Candidates did not perform well on this subtask. Most candidates were able to identify the incompleteness of datasets in 2013 and 2014 and provide responses without technical terms or concepts. However, many struggled to successfully identify the fact that the animals leaving before 2015 are not included. Some candidates failed to conclude that Stay would be longer by including the 2013 and 2014 arrivals.

ANSWER:

The data provided by the Austin animal shelter includes outcomes from 2015 onwards. This includes all animals that arrived at the shelter in 2015 but only some animals that arrived in 2013 or 2014, specifically those whose had long enough stays to not leave the shelter until 2015. The animals that arrived in 2013 or 2014 with shorter stays, leaving before 2015 are not included.

The prediction for how long an animal stays will be made when an animal arrives, but if the 2013 and 2014 arrivals are included when developing this prediction, the predicted stays will be longer than is appropriate because only the longer stays and not the shorter stays of these arrivals would be included in the data.

(b) (*3 points*) Describe, for a general audience, one advantage and one disadvantage of including data from the pandemic era, March 2020 and forward.

*Candidates performed reasonably well on this subtask. Some candidates identified an advantage and disadvantage of including pandemic-era data but failed to make a connection to the business problem of predicting **Stay**. For example, candidates discussed the number of incoming animals during pandemic but did not point out the relationship to the **Stay** variable.*

ANSWER:

One advantage of including pandemic-era data is that it provides more data that can be directly compared with that of the local shelter, including whether the pandemic has similar effects for each shelter. Without this comparison, it would be harder to assert that a model built using the Austin data is appropriate for the local shelter.

One disadvantage of including pandemic-era data is that some of the time periods, particularly the onset of the pandemic, include conditions that are unlikely to be repeated in future periods, such as the surge of adoptions that emptied out the shelter. The predictions for normal operating conditions will be less accurate if these unusual time periods are included.

Your assistant wishes to use the "month" columns as predictors for length of stay in a GLM.

(c) (*1 point*) Identify the issue with using both **in.month** and **out.month** to predict **stay**.

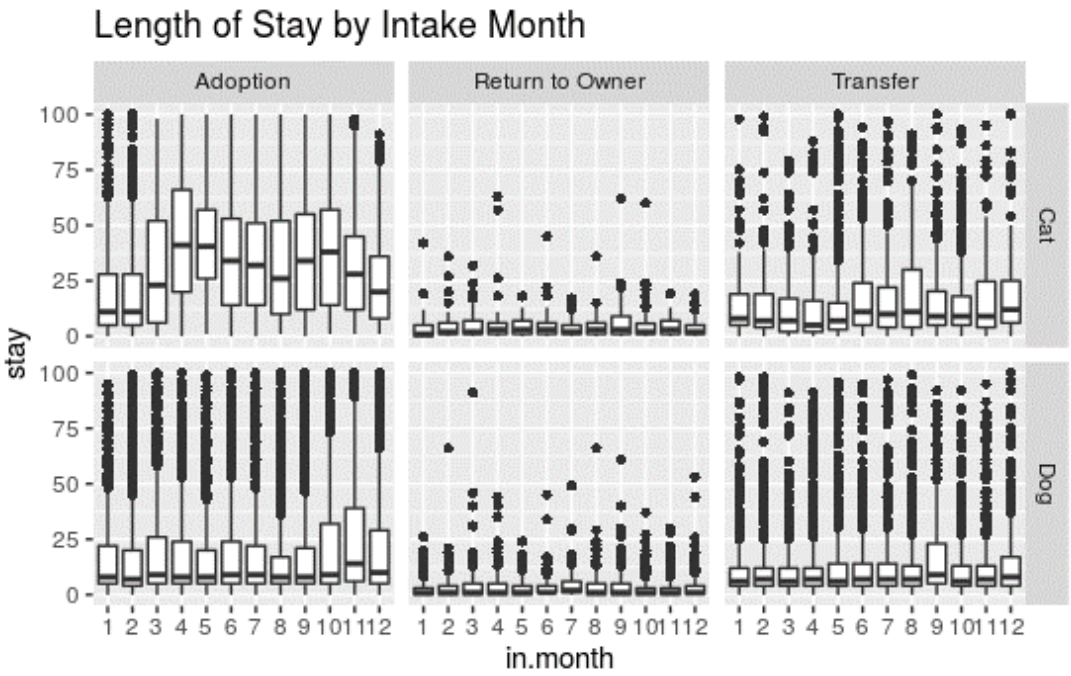
*Most candidates failed to identify the data timing issue that out.month is unknown Upon animal arrival. Many candidates incorrectly discussed whether **Stay** can be calculated using these two variables, the collinearity issue, or the increased model complexity due to multiple levels of the variables.*

ANSWER:

Once **out.month** is known, then **stay** is also known and no longer needs to be predicted—including it in the predictive model is target leakage.

Task 5 (*12 points*)

B asks for explanations about the shelter operations and predictive modeling implications based on the graph below.



(*4 points for each observation*) Describe three observations from the graph that are important for both modeling and shelter operations. Discuss both the impact on shelter operations and the predictive modeling implications. Refer to the graph features that back up your observations.

To get a full credit for a particular observation, the solutions should include all four required components and clearly 1) state an observation that is important for both shelter operations and modeling; 2) refer to the graph features supporting the observation; 3) discuss the impact on shelter operations; and 4) discuss predictive modeling implications.

Observations: *Most candidates were able to identify three important observations. Exceptional candidates stated the observations clearly concisely.*

Credit was given only once for observations that were similar in nature, for example stating that cats stay longer before being adopted and then stating that cats stay longer before being transferred would only receive credit once. Another weak point were observations not being presented clearly, including when many observations were presented at once.

No credit was given for observations that were not important for shelter operations or modeling, for example focusing on one particular month or outlier.

Another common error was that candidates incorrectly interpreted data as being counts of animals rather than the length of stay. This error was penalized only once and credit was given for other components if they were consistent in making this error.

Partial credit was given for observations stating that there were outliers.

Graph features: *To get full credit, candidates were expected to refer to certain features of the graphs such as height of whiskers, height of the median line, dots outside the boxes, etc. Partial credit was given for only mentioning a particular chart (e.g. top-left, or "Dogs Transfer", etc.) along with clearly stating the observation. Many candidates failed to refer to the graph features supporting their observations and many candidates did not even refer to any particular graph or sets of graphs, losing credit.*

Shelter operations: *Many candidates did well in identifying impacts on shelter operations, providing details, and specifying actionable insights for the shelter. However, many candidates lost points for either failing to identify the impact on shelter operation or not providing details. Simply stating that the observation was important for shelter operations was not enough to receive credit. No credit was given if the impact on shelter operations was not consistent with the observation.*

Modeling Implications: *Candidates generally did poorly in this part as it required candidates to connect the observation with how the models will be impacted. For example, many candidates observed skewness of variable* ***stay*** *and mentioned that it will impact the predictive models but did not explain the impact more fully.*

A common mistake was to suggest outcome as a potentially predictive variable, as this is not known for new data and cannot be used as a predictor. However, strong candidates pointed this fact out and suggested having a separate model to identify the outcome first and then using the predicted outcome as an input in a separate model to predict the length of stay.

FIRST OBSERVATION ANSWER:

Observation: The distribution of stay is right skewed.

Graph features: Looking at Adoption Dog, Transfer Cat, and Transfer Dog charts, the inner quartiles are asymmetric, and outliers are heavy on the right tail.

Shelter: The right skewness means potentially long stay for certain animals. It would be important for the shelter to prepare more food, water, and shelter space for the longer stay animals

Modeling: The observations in the long right tail will be high leverage points in a regression model, leading to a poorer fit in other part of the distribution of stay.

SECOND OBSERVATION ANSWER:

Observation: Cats exhibit strong seasonality.

Graph features: This can be seen in the adoption charts (the leftmost column of graphs) with higher median adoptions in months 3-11, that is not seen when comparing months for dogs.

Shelter: Adoption efforts for cats may require seasonally varying resources, such as food and care supplies.

Modeling: To capture this seasonality we need a type of predictive model that naturally captures interactions, such as tree-based models, or explicitly include an interaction term between **in.month** and **animal**.

THIRD OBSERVATION ANSWER:

Observation: Animals are returned to owners very quickly compared to other outcomes.

Graph features: This can be seen by the lower and smaller box plots in the middle column compared to the other columns

Shelter: The focus should be on predicting relatively longer stays associated with transfers and adoptions, which the shelter had more control over. Those predicted to be returned to owners will not require much resources.

Modeling: we can try to predict Return to Owner outcome first using a standalone classification model based on characteristics at arrival, such as presence of collar or overall health, then use separate models for predicting the length of stay based on the predicted outcome. This model setup could be more accurate than trying to predict **stay** across all outcomes.

Task 6 (*6 points*)

B proposes fitting a classification model to distinguish 0-day stays and overnight stays and asks you to explore the data.

(a) (*2 points*) Identify two significant differences between 0-day stays and overnight stays using the assistant's code in the .Rmd file.

No justification of the significance of the differences was needed—the differences merely needed to be identified.

FIRST ANSWER:

About 4/5 of animals with 0-day stays are Return to Owner for **outcome**, but less than 1/5 of animals with overnight stays are Return to Owner.

SECOND ANSWER:

The proportions of **in.intact** and **out.intact** are nearly identical for 0-day stays but very different for overnight stays.

(b) (*4 points*) Assess, for each of the differences you identified, whether it could be used to predict **stay** in this business problem.

While the practicality of predicting 0-day stays is itself questionable, the responses needed to address specifically the viability of applying the observed differences assuming the model could be used immediately upon arrival.

FIRST ANSWER:

The **outcome** variable cannot be used directly to predict **stay** became once the outcome is known, the length of stay will be known and will not need to be predicted. However, if there are strong correlates among the usable predictor variables to **outcome** (for instance, Return to Owner does not occur often after Owner Surrender in **in.reason**), those could help predict whether an animal is expected to stay at least the night.

SECOND ANSWER:

The **out.intact** variable cannot be used directly to predict **stay** because once it is known, the length of stay will be known and will not need to be predicted. However, the difference does suggest further consideration of the **in.intact** variable. Animals arriving not intact should leave not intact, and animals arriving intact may be deemed need to stay overnight for a procedure, affecting the length of stay.

Task 7 (*8 points*)

You ask your assistant to try unsupervised learning techniques for exploring and better understanding the data. You assistant does k-means clustering on **age** and **stay** and plots results based on choices of *k* from 1 to 6, as shown in the .Rmd file. Run your assistant's code on the data and inspect the output. Then do the following:

(a) (*3 points*) Recommend what *k* should be. Justify your recommendation.

Most candidates provided a generic description of how the variation should be less within the cluster and more between the clusters, but few candidates were able to associate this to drop in between sum of squares / total sum of squares ratio. Points were deducted if candidates failed to justify their selection of the number of clusters, k.

ANSWER:

Visually, there are not distinct clusters, but k-means clustering can be evaluated by the relative proportion of the sum of squares between cluster means (between sum of squares or bss) to the same plus the sums of squares within each cluster, or total sum of squares (tss). The proportions are as follows:

k	bss/tss	Gain
1	0%	
2	37%	37%
3	65%	28%
4	74%	9%
5	82%	8%
6	86%	4%

Adding a dimension by adding another cluster can be justified by the additional variation explained. The second and third clusters add 37% and 28% respectively, but the additional clusters add no more than 10% to this proportion. Thus, k = 3 is recommended as a good balance between gaining information and not adding too many dimensions.

(b) (*2 points*) Assess your assistant's claim that these specific clusters can help with the prediction of **outcome**.

To receive full credit, the data timing issue needed to be recognized, regardless of the quality of the clusters. Well prepared candidates referred to the plots and how the clusters were not clearly distinguished. Few candidates were able to make the connection between stay and outcome.

ANSWER:

Using **stay** to predict **outcome** will not be helpful because once the length of the animal's stay at the shelter is known, the outcome will be known at the same time. That **stay** is embedded in a cluster does not change this.

(c) (*3 points*) Describe two practical differences between hierarchical clustering and k-means clustering. Do not implement hierarchical clustering.

Practical differences are those which matter in practice, affecting how the modeling technique is used. Technical differences not impacting practical use could not receive full credit. Most candidate did well in identifying the differences. *Candidates who just described the two methods and their working principles received minimal partial credit.*

ANSWER:

In k-means clustering, the number of clusters is chosen first and then the clusters are determined, but in hierarchical clustering, the number of clusters is chosen after creating the clusters.

In k-means clustering, the clusters determined can vary based on the random centers chosen at the beginning of the algorithm, but in hierarchical clustering, the same set of nested clusters is created without randomness.

Task 8 (*13 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 Gaussian distribution with identity link function on the public dataset to predict **stay**. 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 for **stay** and the square of the bias (the difference between the average predicted result and actual result for **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 ~ animal*age	7.4	2799.2
stay ~ animal*age_1	3.9	2791.3
stay ~ animal*age_3	10.1	2735.4
stay ~ animal*age_5	17.1	2760.0
stay ~ animal*age_10	47.2	2787.4

The **age** predictor is the original numeric variable while the other four predictors involving age, e.g. **age_1**, are categorical variables using age rounded down to the nearest integer, with the lowest age group always being “Under Age 1” and the highest age group being the age indicated and above. The levels for each categorical variable can be seen in the .Rmd code. 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**.

Partial credit was given to candidates who gave either the square root of variance or square root of bias, but both were needed to justify full points. Partial credit was given if the candidate knew a square root was involved but applied it incorrectly. Many candidates left this subtask blank.

ANSWER:

The variance and squared bias figures above relate to the predicted and actual lengths of stay for adopted pets. For the predictions of **stay**, using the results of the first model, the typical error due to variance is a 7.4^0.5 = 2.7 difference up or down from the predicted value and the typical error due to bias is a 2799.2^0.5 = 52.9 difference up or down.

(c) (*2 points*) Explain how the mechanics of fitting the underlying linear model causes the variance to be higher for the model using **age** than it is for the model using **age_1** even though they have the same number of coefficients.

Well prepared candidates recognized that understanding the data is critical to application of predictive analytics. Partial credit was awarded if the candidate explained that outliers in a numeric variable could cause higher variance for age.

Stating that age is numeric while age_1 is categorical was not enough to get any points. Neither was stating that age had a wide range of values. Even as a numeric variable with a wide range of values, age could have been centered around 0-1 which would have resulted in very similar models.

ANSWER:

The model using **age** fits, separately for cats and dogs, a linear coefficient that is affected by how far away from zero each observation of **age** is. Most observations of **age** are close to zero and relatively few are far from zero, as shown by this summary for the variable in the training data:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -0.10    0.20    0.80    1.76    2.00   19.00
```

When random 20% samples are chosen from this training data, there can be significant variation in the more sparsely distributed older ages, and these older ages have an outsized impact when fitting a linear model, leading to more variation in the coefficient for **age** and hence more variance for the model. For the categorical **age_1** variable, the distribution of ages sampled for ages 1 and above for each of cats

and dogs does not matter, and the fitting of the coefficient is more stable as a result, giving the model less variance.

(d) (*2 points*) Explain why bias (as calculated here) may not always decrease with additional degrees of freedom, as seen with the model that uses **age_5** compared to that which uses **age_3**.

Partial credit was given if the candidate explained that the model is trained on the training data while bias is measured on the test set. Many candidates failed to recognize that the training data was a 20% random sample before 2020 but the testing data was 2020 onwards. Partial credit was given if the candidate explained that the age_5 variable may not have predictive power. Candidates stating that there may not be many observations in the additional age cohort when using the age_5 variable were not given partial credit unless they related to how that impacted the training data.

ANSWER:

The bias calculation may not always decrease when adding predictors because 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 as in B's process, adding predictors may not improve the accuracy of the predictions.

(e) (*2 points*) Recommend which GLM should be used based on the above results. Justify your recommendation.

*It was not clear to some candidates that the question was directed to making a recommendation based on the 5 models in the table above, and instead responded with a model of their own. Successful candidates correctly answered that stay ~ animal*age_3 model should be used since this had the lowest sum of variance and squared bias.*

Partial credit was given if the candidate identified the correct model but justified based on only lowest bias. Some commentary on balancing bias with variance was needed for full credit.

ANSWER:

Model formula	**Variance + Squared Bias**
stay ~ animal*age	2806.4
stay ~ animal*age_1	2795.2
stay ~ animal*age_3	2745.5
stay ~ animal*age_5	2777.1
stay ~ animal*age_10	2834.6

To select the best model, the sum of the variance and squared bias errors should be used. The calculations above show the model fitting **age_3** should be used, as it has the lowest sum. The notable decrease in bias compared to the simpler model with **age_1** more than makes up for the increased variance.

(f) (*3 points*) Calculate, after fitting the recommended GLM to all available data, a complete list of predicted stays based on animal and age for a general audience to use. Include the code used to fit the model (but not its output) in the space below.

Many candidates did not provide a complete list of all possible combinations of animal and age or including model output.

No credit was given for full explanations of the intercept and coefficients as part of an explanation for how to calculate the predicted stay. The question asked to see a table of predicted stays, not an explanation of how to calculate it.

ANSWER:

Code used to fit model:

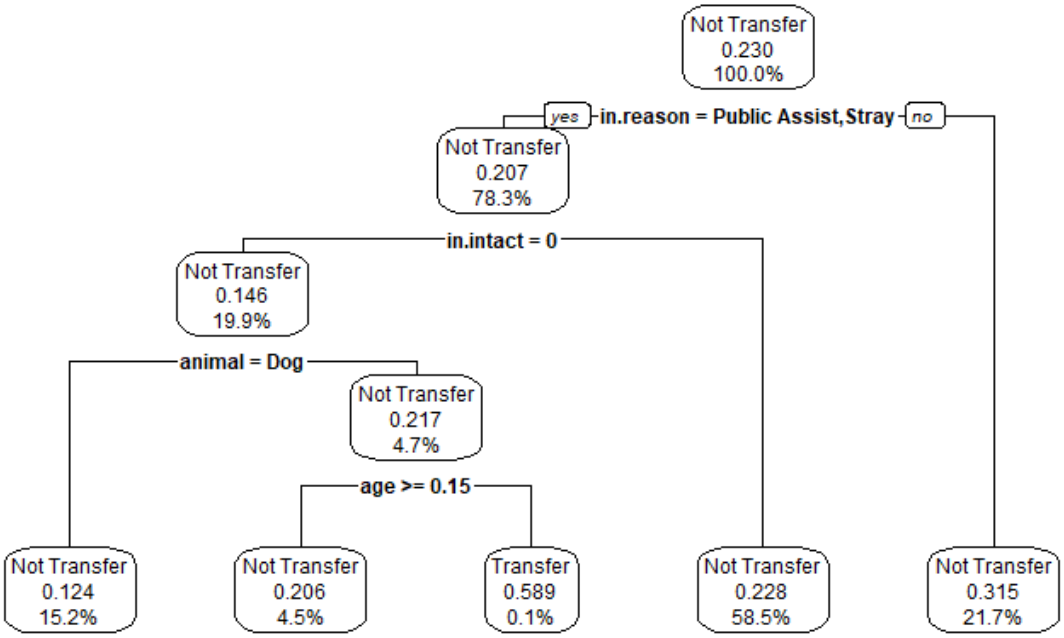
```
chosen_glm <- glm(stay ~ animal*age_3, gaussian, data.adopt)
```

List of predicted stays:

Age (rounded down)	Cat	Dog
Under Age 1	45	22
Age 1	44	29
Age 2	49	37
Age 3 and Up	63	55

Task 9 (*9 points*)

Your assistant builds a decision tree to predict whether the **outcome** is "Transfer" or not, resulting in the following tree.



B suggests that you include a new feature called kitten (**animal** = cat, **age** < 0.5) for your decision tree.

(a) (3 points) Without creating a new decision tree, assess whether this variable will materially improve the decision tree.

Most candidates were able to recognize that the new split would not be helpful and justified their answer stating that only few observations would be impacted. Few candidates noted how interactions do not typically improve decision trees.

ANSWER:

The variable "kitten" is an interaction variable combining **animal** and **age**. While some model types (e.g. GLM) do not have a natural mechanism for handling interactions and require new features to be manually created and fed to the model, decision trees can detect and model interactions without any manual intervention.

This tree already has a split for cats at age 0.15 which results in very similar behavior to a "kitten" variable. Also, the model selected the 0.15 value for the of the split by optimizing an impurity measure, meaning that splitting at 0.15 will result in a better model than 0.5. Therefore, adding the kitten variable will not materially improve the decision tree.

Due to high costs associated with transfers, the animal shelter would rather underpredict than overpredict the "Transfer" outcome while still successfully identifying "Transfer" outcomes. Your assistant produces the following confusion matrix from R, where TRUE corresponds to "Transfer", using a cutoff of 0.25:

```
          Reference
Prediction TRUE FALSE
     TRUE   680  1392
     FALSE 1574  6034
```

(b) (3 points) Recommend an evaluation metric based on the confusion matrix which supports the animal shelter's goal. Calculate the metric using the confusion matrix above. Justify your recommendation.

Most candidates incorrectly identified specificity or sensitivity as the metric best supporting the animal shelter's goal while few candidates correctly identified precision. Using the term positive predictive value also earned full credit. Specificity fails to measure successfully identifying transfer outcomes. Sensitivity emphasizes reducing how much actual transfers are not predicted, which the shelter is OK with.

ANSWER:

Based on the animal shelter's business needs, I recommend using precision as an evaluation metric. Precision is the number of true positives divided by the number of predicted positives. A high value means a greater portion of animals identified for transfer will be likely to be transferred. Using a cutoff of 0.25, the model precision is 680/(680 + 1392) = 33%.

(c) (3 points) Explain, for a general audience, why increasing the cutoff from 0.25 to 0.26 does not impact the confusion matrix from your assistant's decision tree. Identify the smallest cutoff greater than 0.25 (rounded to 2 decimal places) that will result in a different value of the evaluation metric.

Overall candidates did will on this subtask. *Candidates were able to explain deciison trees well and identify the lowest cutoff.*

ANSWER:

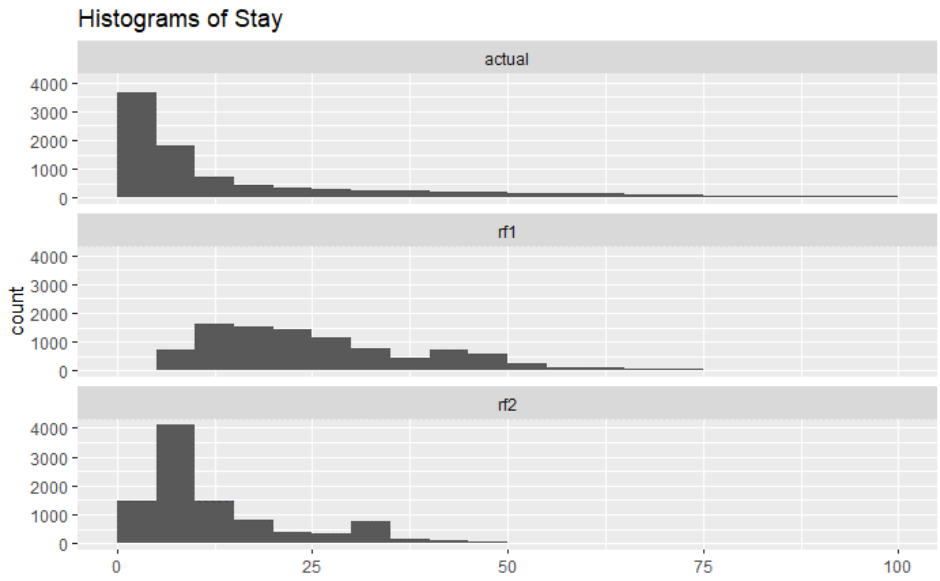
When using a decision tree, one follows a series of if-then decisions, like a flow chart, until reaching a final "leaf" which contains the observed probability of a transfer occurring. In this decision tree, there are five leaves, meaning the model is only able to predict one of five different probabilities. The five transfer probabilities this tree can predict, ranked from lowest to highest, are 0.124, 0.206, 0.228, 0.315, and 0.589.

The cutoff determines how to convert the probabilities into predicted outcomes. A cutoff of 0.25 means that the two predictions below 0.25 are treated as not transfers, and the values above 0.25 are treated as transfers. All cutoffs between 0.228 and 0.315 will result in the same classifications: three non-transfers and two transfers. Therefore, the lowest cutoff greater than 0.25 that will change the model and its evaluation metric is 0.32, which predicts owner surrenders (the remaining reason animals come into the shelter) as not transfers.

Task 10 (*9 points*)

B would like you to use a random forest to predict **stay**. B notes that it is important to the animal shelter that the predictions be equally likely to be too high or too low.

Your assistant fits two random forests to predict the length of stay. The first, rf1, predicts **stay** without a log transformation; the second, rf2, predicts log(**stay** + 1). The histograms below present the distribution of the actual **stay** variable followed by the distributions of the predicted values of **stay** from the two models.



(a) (*3 points*) Explain why the histograms from appropriately fitted decision tree models would likely not be as smooth as those produced by a random forest, as above.

Well prepared candidates articulated how a random forest is an average of a number of decision trees. Weaker candidates provided responses related to the graphs themselves or focused on the log transform producing a smoother graph, with neither response receiving credit.

ANSWER:

A decision tree regression model produces at most as many possible predictions as it has terminal nodes/leaves. A well fit decision tree will have been limited in the number of terminal nodes/leaves, to prevent overfitting. A histogram of predicted values will have at most the same number of bins with non-zero values, while certain bins within the range of the target variable will have no predictions, creating a "step" like pattern that will not appear smooth.

In a random forest, the prediction is averaged over the predictions of many individual trees. Each of these trees has randomly chosen fitting data and predictors considered at each split. The averaging of these uncorrelated trees smooths out prediction values and the corresponding histograms.

(b) (*3 points*) Explain why rf2 produces more predictions of shorter stays than rf1.

Most candidates recognized how a log transformation of the target variable led to extreme values being less heavily weighted. Well prepared candidates also noted how a log-transformed model will produce more even predictions across all outcomes.

ANSWER:

The distribution of **stay** is right-skewed. When rf1 is minimizing the mean squared error of **stay** in choosing an optimal split, it can reduce those squared errors by choosing a split favoring the very highest values. The other side of that split will average together the low to moderately high values for its prediction.

When rf2 is minimizing the mean squared error on log(**stay** + 1), the target variable is less right-skewed due to the non-linear log transformation; the optimal splits do not favor the very highest values over the low to moderate values as much. The result is a model with more balanced predictions over all data points, which increases the number of shorter stay predictions.

Your assistant states that the random forest without the log transformation is preferable since the algorithm preserves the underlying distribution; i.e., predicted values more closely follow the distribution of the original dependent variable.

(c) (*3 points*) Assess your assistant's reasoning, including consideration of how the animal shelter will use the model.

Well prepared candidates provided a response relative to the business problem, noting how overemphasizing high values of stay was undesirable from the business problem perspective. Points were lost if the candidate did not incorporate the business problem into their response.

ANSWER:

From a statistical perspective, the assistant's reasoning is accurate: rf1 predicts the untransformed target variable; it will typically perform better since it is optimizing to an error that is on the same scale as the evaluation metric.

In this case, however, the business context given by B requires the predictions be equally likely to be too high or too low. From the graphs above, it is more likely that rf1 will predict fewer lower values of **stay** and predict more higher values of **stay**. While not perfect, rf2 does a better job than rf1 at modeling most of the distribution of stay. The assistant's reasoning does not take this into consideration.

Task 11 (*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.