Exam PA April 12 Project Statement

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 13 tasks numbered 1 through 13 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. If any part of your exam was answered in French, also include "French" in the file name. Please keep the exam date as part of 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

Your boss, B, recently started a consulting firm, PA Consultants, specializing in predictive analytics. You and your assistant, A, are the only other employees. B informs you that the City Manager of Tempe has hired your firm to understand why Tempe is not meeting one of its goals and what steps should be taken to achieve the goal.

Tempe is a small city of about 200,000 residents next to the larger city of Phoenix in Arizona, USA. Tempe has a desert climate and is the home of Arizona State University (ASU). ASU has over 50,000 students.

The City of Tempe wants to respond to emergency calls for help that require advanced life support (ALS) in six minutes or less for 90% of such calls. Such arrivals increase the probability of good outcomes for the person in need of ALS. Unfortunately, only 75% of ALS calls have response times of six minutes or less and efforts to increase the percentage to 90% have not had any effect. Efforts consisted of disseminating the metric and goals to the personnel involved. Your tasks are to understand the hindrances to achieving the ALS goal and to recommend steps that will allow Tempe to realize its goal. B emphasizes the need to understand the issues and data involved even if they are not directly related to the performance goal. You sense B would welcome hearing of any additional projects to pitch to the City of Tempe or to ASU.

The response time has three components:

- *The alarm processing time is the time from when the emergency phone call is answered until the Tempe Fire Medical Rescue Department (TFMR) is notified. This part of the process is handled by a regional dispatching organization that also classifies the calls as ALS.*
- *The turnout time is the time from when TFMR receives notification of the ALS call until the firefighter/medics enter their vehicle.*
- *The third component is travel time, during which the vehicle travels to the site of the ALS emergency.*

B directs you to use a dataset[1] of public data that includes all the 2018 ALS calls for Tempe and some weather variables. B has provided the following data dictionary and the dataset of 9,853 records in a file called Exam PA Tempe ALS Data.csv*.*

[1] Adapted from 1.01 ALS Response Time" (2018) by City of Tempe, AZ is licensed under Creative Commons — Attribution 2.0 Generic — CC BY 2.0. Weather data is from the Arizona Meteorological Network (AZMET).

Data Dictionary

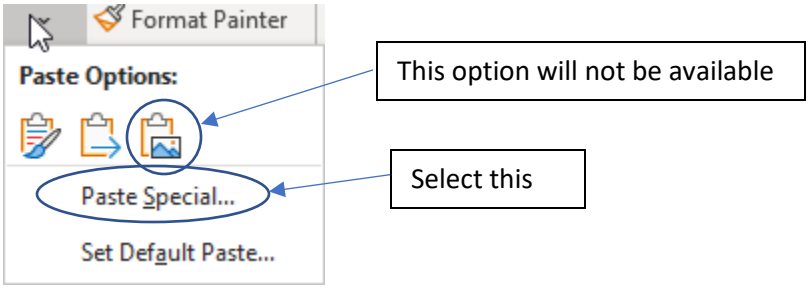
Variable Name	Variable Values
issue	Type of emergency event (11 categories)
vehicle	L, E indicate the two most common vehicles. X is all others.
station	1 to 8
hour	0 to 23, hours past midnight
min.past.midnight	0 to 1439
month	1 to 12
day	1 to 31
weekday.number	1 to 7 for Sunday to Saturday
dewpoint	a weather value that incorporates humidity
temp.f	hourly temperature (degrees Fahrenheit)
temp.c	hourly temperature (degrees Celsius)
alarm.processing.time	seconds from answered call until TFMR notification
turnout.time	seconds from TFMR notification until vehicle travels
travel.time	seconds of travel to the site of the emergency
response.time	sum of the above three values

Comments

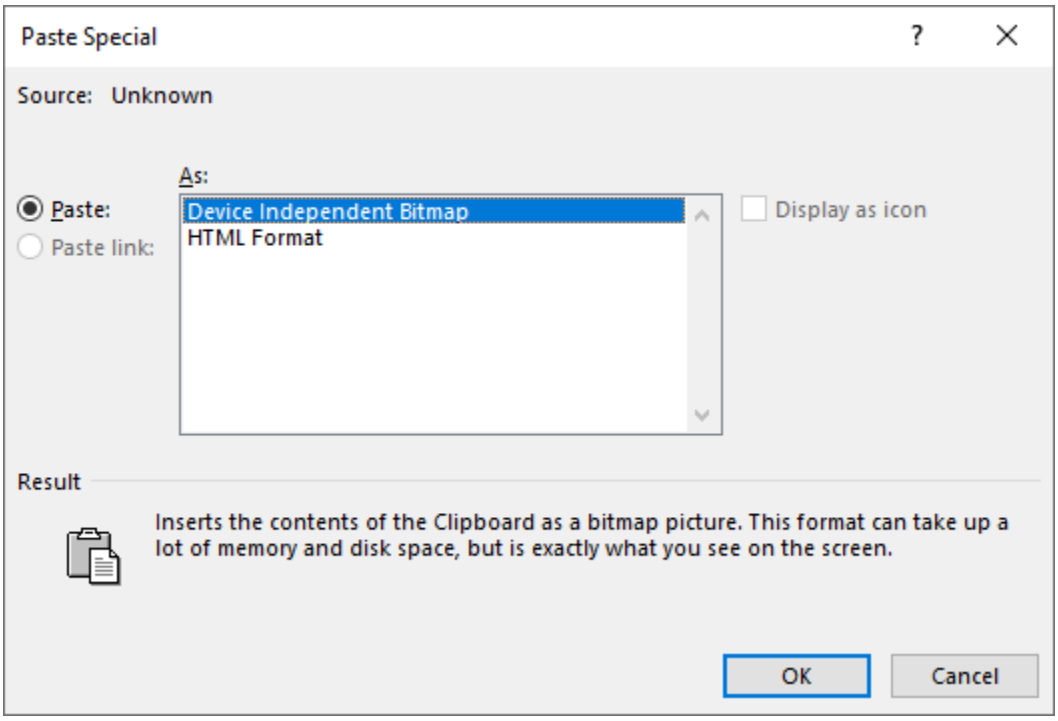
The type of medical event may not be known precisely at the time of the call, but information related to the issue variable is conveyed by the dispatcher to the workers in the vehicle.

Station 6 serves ASU. Stations 4-7 serve wealthier areas than the others.

IMPORTANT NOTE: When pasting a picture from RStudio to Word, there is only one approach that will work. After right clicking on the image in RStudio and selecting "copy" the following steps need to be taken in Word. On the Home menu, click on the down arrow under "Paste" and then select "Paste Special …" From the list of options, select "Device Independent Bitmap." The following images indicate these steps.

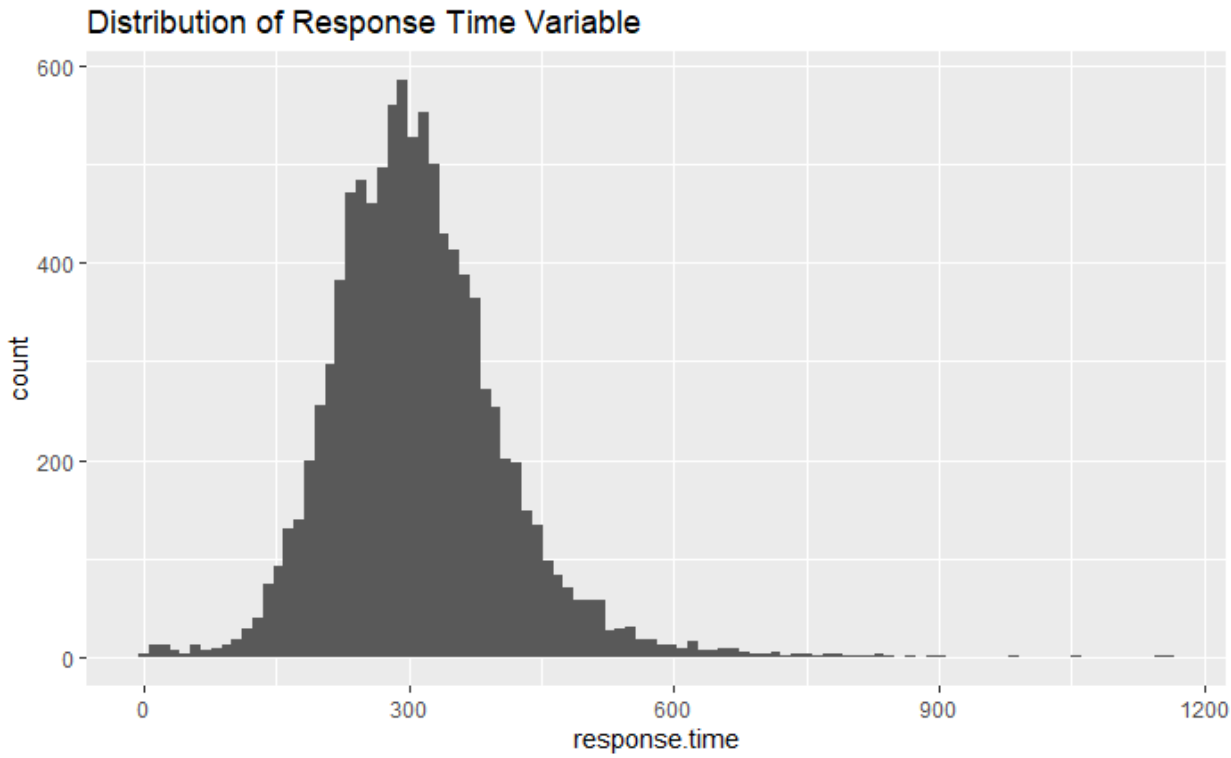


From this dialog box, make the indicated selection.



Task 1 (*8 points*)

You asked your assistant to explore the **response.time** variable with the understanding that it will be used as the target variable in a GLM. Your assistant produced the graph below titled "Distribution of Response Time Variable" and suggests you consider a transformation to the **response.time** variable.



(a) (*3 points*) Recommend whether a transformation should be applied to the **response.time** variable in the GLM given the business problem. Justify your reasoning.

For full credit, candidates needed to supplement a technical interpretation of the graph with rationale derived from the business problem. Many candidates based their recommendations solely on technical interpretations of the distribution graphs.

One common recommendation was to apply a log transformation to correct for right skewness. This recommendation was only awarded full credit if the candidate discussed why it was preferred over alternative ways that a GLM can handle skewness (for example, through use of a Gamma or inverse Gaussian distribution).

ANSWER:

I do not recommend any transformation to be applied to response time. Based on the histogram, the variable appears to be nearly normally distributed with a mean near 300 seconds and almost all observations between 0 and 600 seconds. Target variables with a normal distribution are easily handled by GLMs. There seems to be slight right skewness but not enough to warrant applying a transformation.

Moreover, the City of Tempe will be interested in the factors driving long response times. A transformation could reduce the leverage that long response times have in fitting the GLM.

Your assistant warns that some outliers in **response.time** may be skewing its distribution.

(b) (*5 points*) Discuss the outliers in the **response.time** variable with respect to each of the following:

i. The plausibility of the outliers
ii. The goal to reduce response time below 6 minutes for 90% of calls
iii. Fitting a GLM that predicts **response.time**

Many candidates did not provide any evaluation of the outliers to support their conclusion around plausibility.

ANSWER:

The plausibility of the outliers:

Outliers should be evaluated for whether they represent plausible values or some data issue. The minimum value of response time is -1, which seems to be a data error. A response time of 0 also seems to be a data error. The maximum value is 1,160 seconds, which is about 19 minutes. This seems plausible as a long response time may be caused by an accident, heavy traffic, or other issues.

The goal to reduce response time below 6 minutes for 90% of calls:

The City of Tempe wants to respond to emergency calls for help that require advanced life support (ALS) in six minutes or less for 90% of such calls. Currently, only 75% of ALS calls have response times of 6 minutes or less. High outliers contribute to the trips in which response time exceeded 6 minutes. Understanding some common characteristics of such outliers may be a quick way to identify some of the likely drivers of a longer response time. The City of Tempe's goal is to reduce the percentage of trips with long response time, so the actual magnitude of the outliers beyond the 90th percentile does not impact the goal.

Fitting a GLM that predicts response.time:

High outliers increase the mean and standard deviation of the response time variable, which may increase the standard error of a GLM on the response time. Outliers may produce high leverage on a GLM. This leverage can have a strong effect when fitting high dimensional categorical variables. The levels that contain the outliers may have much higher coefficients than other levels, which may indicate a poor fit unless there is an underlying reason why one would expect outliers in those particular levels.

Task 2 (*7 points*)

Your Boss, B, would like to educate the client on types of modeling objectives.

(a) (*4 points*) Explain descriptive and predictive modeling objectives. Write for a general audience. Include an example of how each type of objective could be applied to this business problem.

Some candidates distinguished descriptive modeling as focusing on the past and predictive modeling as focusing on the future.

Some candidates failed to emphasize that descriptive modeling helps identify the key relationships and patterns between the variables in the dataset.

A significant proportion of candidates discussed predictive modeling as focused on predicting the future outcomes without any discussion of the required accuracy of the prediction.

ANSWER:

Descriptive Modeling Objective:

For a descriptive modeling project, the primary goal is to understand the relationships between conditions and outcomes. The Tempe ALS project is primarily a descriptive analysis project since reducing longer response times requires Tempe to take action on the key factors that impact response time. Tempe must understand the relationship between various factors and components of response time to decide how to manage them.

Predictive Modeling Objective:

Prediction refers to projects where the primary goal is the accuracy of the predictions from the final model. Interpretability of the model and the understanding of how input variables impact outcomes may not be as important. Implementing a model to predict response time at the time of the initial ALS call could possibly be used by the dispatch team to help manage communication with the emergency caller or consider alternative stations for response.

B would like to clarify the deliverable from PA Consultants.

(b) (*3 points)* Propose three questions for the City of Tempe that will help clarify the business objective.

Candidates performed poorly overall on this task. Many of the questions that candidates produced had no connection, or a very weak connection, to the business objective. Questions around the completeness of the dataset were not awarded credit.

ANSWER:

Question 1:

Do you have any initial hypothesis or intuition that might explain potential variation in the response times?

Question 2:

Are there any subject matter experts that you would recommend talking to prior to performing the analysis?

Question 3:

Over the span of the data collection period, where there any notable changes or events that we should know about?

Task 3 (*7 points*)

(a) (*3 points*) Describe the "curse of dimensionality" and how it can lead to problems in a GLM.

Candidates performed well on this task overall. Only partial credit was awarded for candidates who did not define dimensionality. Some candidates received partial credit for only discussing the curse of dimensionality generally without any description of how it impacts GLMs specifically.

ANSWER:

The curse of dimensionality refers to the problem created when the number of explanatory variables, or the number of levels in explanatory factor variables, is large compared to the volume of data (i.e., the number of observations).

In a GLM, each additional level in a factor variable results in the creation of an additional binary explanatory variable. Having many variables or variable levels can result in model complexity that is greater than that of the underlying process being modeled. This extra complexity ends up fitting the idiosyncratic noise in the training data.

(b) (*4 points)* Recommend a distinct improvement on each of two high dimensional variables in the ALS data to reduce granularity and likely improve predictive power. Justify your two improvements.

Full credit responses required identifying two high granularity variables and recommending methods for reducing their granularity. Only partial credit was awarded for identifying a high granularity variable but not making a recommendation or recommending that the variable be removed entirely.

ANSWER:

Improvement 1:

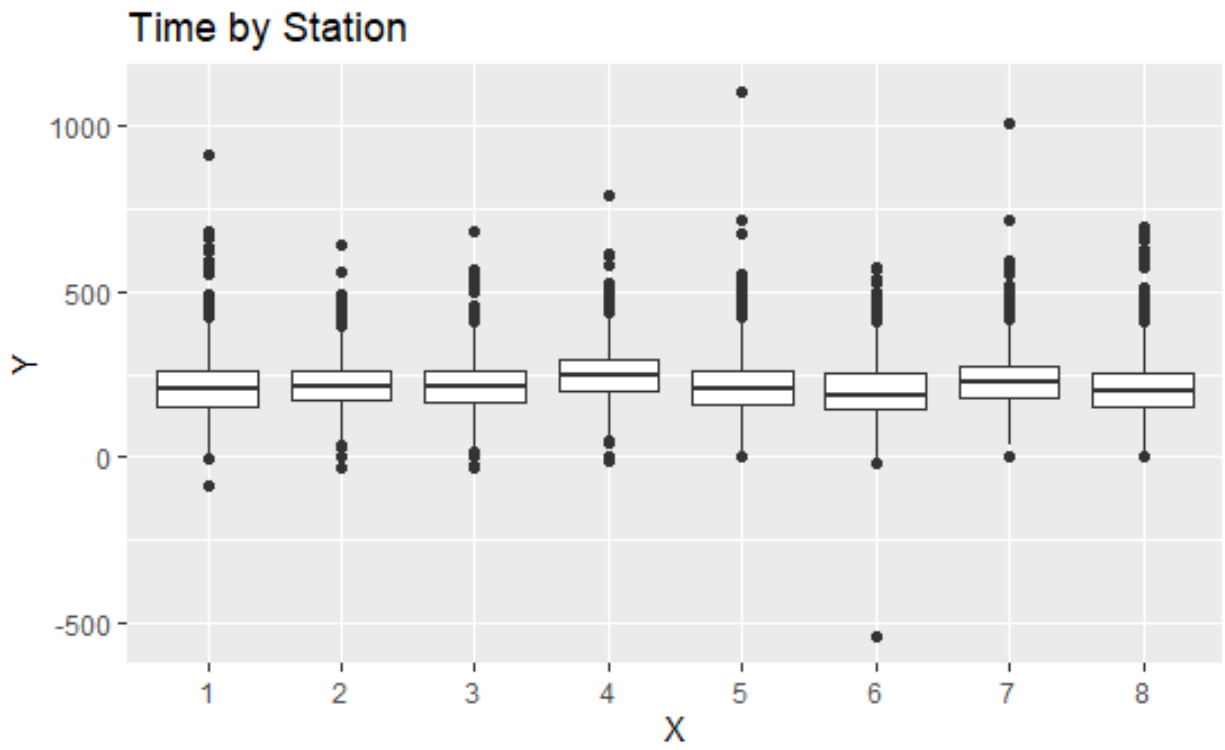
I recommend generating a feature from the hour variable that groups hours into interpretable levels. Possible levels could be morning, afternoon, and evening. Small differences between adjacent hours are not likely to be predictive, and judicious grouping of hours can capture the important time effects with fewer dimensions.

Improvement 2:

I recommend transforming the day of month into a variable that contains levels for weekdays, weekends, and holidays. The day of the month is not obviously related to the target variable or business problem. As is, day of month likely would likely lead to overfitting without any increase in predictive power. This transformation would reduce the variable's dimension from 31 to 3, thus reducing the potential for overfitting.

Task 4 (*9 points*)

Your boss, B, has asked you to use data visualization techniques to better understand the distributions of response time or its components by station.



(a) (*3 points)* Describe strengths and weaknesses of the graph above, which was created by your assistant to depict **travel.time**.

Most candidates focused on what was observable in the graph, rather than the appropriateness of the choice of a box plot itself. Several candidates identified the title and axis labels as weaknesses but did not discuss why the ones provided are weak. These answers were not awarded full credit.

ANSWER:

The assistant's use of a boxplot is a good choice for insight into the distributions of a continuous variable (travel.time) across a categorical variable (station). The representation of the interquartile range is particularly useful as this business problem is about the percentage of observations below a certain value. However, the assistant gave general labels to the X and Y axis where a better approach would have been to assign the labels "Station" and "Travel Time" to the X and Y axis respectively. Also, while the assistant did provide a title, it is ambiguous as to which variable the graph depicts as there are four time variables in the dataset. Labeling the graph as "Travel Time by Station" would have been better.

(b) (*4 points)* Create an informative boxplot of **response.time** by **station** that B can include in a report to the city manager. Include a horizontal line at 360 seconds. Paste the code used to create the graph and the image of the graph below.

Most candidates earned full credit for this task. The most common reason for reduced credit was graphing travel time instead of response time.

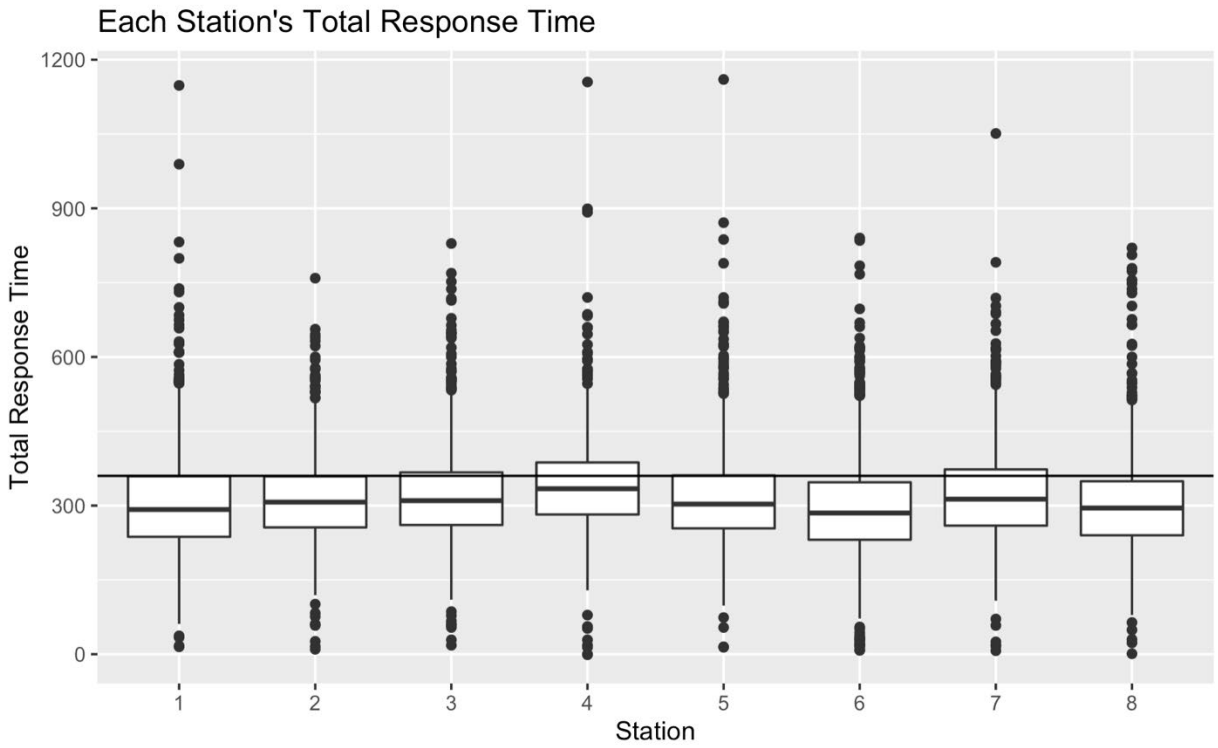
ANSWER:

Code:

```
p <- ggplot(data.all.task4, aes(x = station, y = response.time))

p + geom_boxplot() + xlab("Station") + ylab("Total Response Time") + labs (title = "Each Station's Total Response Time") + geom_hline(yintercept = 360)
```

Graph:



(c) (*2 points)* Compare the outliers in travel time and response time between the assistant's chart in part (a) and the chart you produced in (b) and describe what is surprising.

Most candidates were able to point out the lack of negative values in the chart from (b). However, very few candidates were able to explain the underlying cause. Simply hinting at a data issue was not sufficient for cull credit.

ANSWER:

The chart that my assistant created in part (a) represented the distributions of travel.time by station whereas the graph produced in part (b) depicts distributions of total response.time by station. Total response time is the sum of alarm time, travel time, and turnout time. Travel time is the largest

contributor to total response time and the distributions are very similar with station #4 having the highest 75th percentiles. In the graph of travel time in (a) there is an extremely negative value around -500 for station #6. However, in the plot of total response time in (b) no such negative values exist. For this to happen, one of the other variables (alarm time or turnout time) must have been extremely large to offset it.

Task 5 (*6 points*)

When fitting a GLM, some numeric variables can be modeled as factor variables.

(a) (*3 points*) State three reasons to convert a numeric variable to a factor variable when fitting a GLM.

Candidates generally did well on this task. A common error was to state that converting a numeric variable to a factor variable reduces dimensionality.

ANSWER:

Reason 1:

Converting to a factor variable allows each level of the factor to be examined separately. For example, a regularized regression can distinguish which levels of the factor are significant.

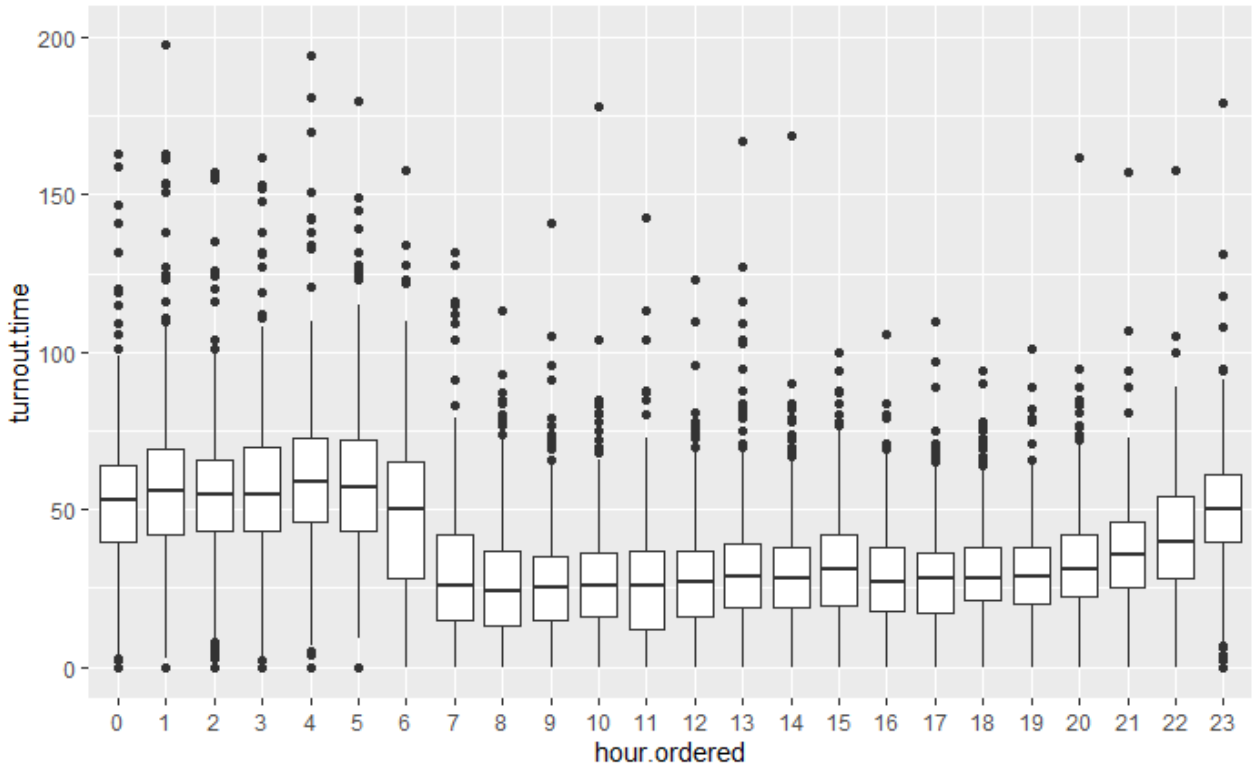
Reason 2:

A factor variable may be more interpretable than numeric. For example, it may be simpler to convey discrete age group comparisons than a general linear effect of age.

Reason 3:

Converting to a factor variable allows for more complex relationships to the target variable since separate coefficients are estimated for each level.

B wants you to build a GLM to find the predictors of turnout time. Your assistant removed the extreme outliers and prepared the graph below.



(b) (*3 points*) Recommend a specific transformation of the **hour** variable that will enhance the predictive power and interpretability of the GLM. Justify your recommendation.

Most candidates did well on this task. A common full credit response was to suggest reducing the factor variable to two or three dimensions based on time of day, justified by the reduced dimensionality, improved predictive power, and better interpretability.

No credit was awarded for recommending log or square root transformations. These transformations would not be appropriate given the range of hour.ordered [0, 23] and since hour.ordered does not have a monotonic relationship with the target variable.

ANSWER:

I recommend grouping *hour.ordered* by time of day to produce a less granular factor feature. I specifically recommend the following two groups:

- Day (hours 7-22), as the base level
- Night (hours 23 and 0-6)

These groups have similar turnout times within each group, and the night turnout times are noticeably higher. The two groups capture most of the variability of turnout time by hour. Since the model will have one hour-related coefficient instead of 23, the model will be less susceptible to overfitting to noise in the training data. This will aid the predictive power of the model when it is used on new data. The grouping also improves interpretability of the model.

Task 6 (*7 points*)

(a) (*3 points*) In the context of a GLM, do the following for each of the Gaussian, Poisson, and Gamma distributions:

i. State the domain of the distribution function.

ii. State a target variable that is appropriate for the distribution. The target variable does NOT need to be from the dataset you are provided but should relate to the problem statement.

No credit was awarded for generic examples of target variables that were not related to the problem statement, such as claim counts or claim amounts.

Some candidates correctly identified Poisson's domain as non-negative integers but incorrectly stated any integer variable could be modeled as Poisson. Credit was only given for count variables. For example, it is not appropriate to model an integer day of the week variable as Poisson. The number of calls could be Poisson.

To receive full credit for modeling any of the time variables from the dataset as Gamma distributed, candidates needed to demonstrate knowledge that zero and negative values need special treatment.

Many candidates incorrectly stated that the domain of Gamma includes 0, or that the domain of Poisson does not include 0.

ANSWER:

Gaussian distribution:

The domain is all real values.

A target variable that could be modeled with the Gaussian distribution is the response time.

Poisson:

The domain is non-negative integers.

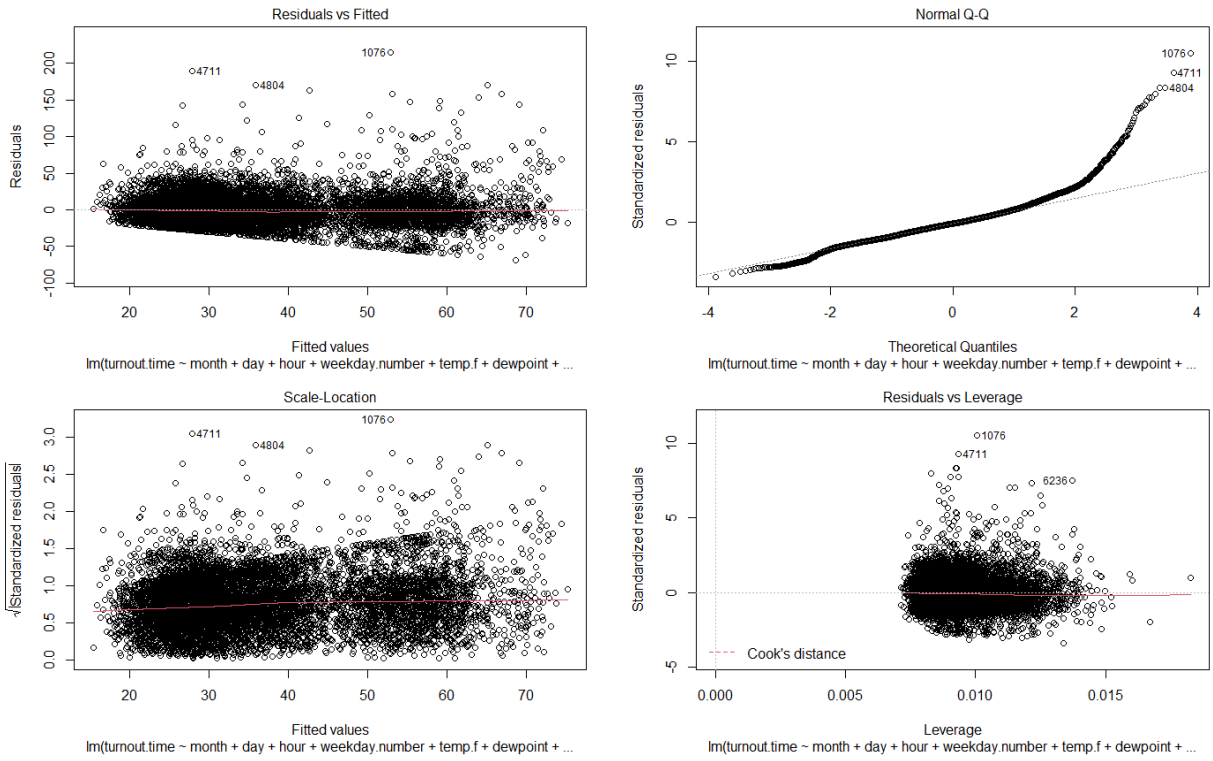
A target variable that could be modeled with the Poisson distribution is the number of calls in an hour.

Gamma distribution:

The domain is positive real values.

A target variable that could be modeled with the Gamma distribution is the turnout time plus one second since it is continuous and positive.

Your assistant runs an ordinary least squares (OLS) model to model **turnout.time**. Review the diagnostic plots below.



(b) (*2 points*) Explain two reasons why OLS is not a good choice to model **turnout.time**.

For full credit, candidates needed to make an observation based on the plots provided and explain why that observation indicates OLS is not a good choice for modeling turnout.time. Most candidates correctly identified issues with OLS using the Residuals vs Fitted and Q-Q plots, but credit was also awarded for referencing the other plots if correctly interpreted. No credit was awarded for general observations about OLS assumptions without referencing the diagnostic plots.

ANSWER:

First Reason:

OLS is not a good choice because the constant variance assumption for the residuals is not met. The fitted vs residuals plot shows that variability increases as the fitted values increase, particularly for the negative residuals.

Second Reason:

The Q-Q plot of residuals indicates that they are not normally distributed, violating another OLS assumption. In particular, there are too many extreme high values.

(c) (*2 points*) Recommend a transformation to **turnout.time** that will improve the residuals when fitting an OLS model. Justify your recommendation.

Full credit was awarded to candidates who recommended a valid transformation, explained any data handling needed in conjunction with that transformation, and justified their recommendation. Most candidates recommended a log transformation, which required an explanation for how to account for values of zero in the data for full credit. No credit was awarded to candidates who recommended removing outliers, as this is not considered as a data transformation.

ANSWER:

I recommend a log transformation to turnout.time. The transformation will shrink the large values relative to the smaller values. This should reduce the phenomenon where the residuals grew in variability as the fitted values increased in the OLS using the transformed target compared to those of the OLS on the untransformed target variable. In doing this, a small positive value should be added to turnout.time to make the log operation feasible.

Task 7 (*5 points*)

B directs you to create a Gaussian GLM model with a log link for **turnout.time**. Use the model as built in the .Rmd file.

(a) (*3 points*) Interpret the **hour9** and **temp.f** coefficients and their impacts on the target variable.

Candidate performance was mixed on this task. To receive full credit candidates needed to interpret the coefficients correctly and calculate the impacts accurately. Common errors were not exponentiating the coefficients and not recognizing that the hour9 variable needs to the interpreted relative to the base factor level.

ANSWER:

The hour9 coefficient is -7.174e-01. The coefficient is negative, indicating that turnout time is lower at 9 a.m. than it is at the base hour, midnight. Since the log link function was used, turnout time at 9 a.m. will be e^(-0.7174) = 48.8% of the turnout time at midnight, all else equal.

The temp.f coefficient is 1.094e-03. This means that for additional degree Fahrenheit, the turnout time is multiplied by e^0.001094 = 1.001095. This is an imperceptible change.

(b) (*2 points*) Recommend two variables for further investigation based on the output of the model. Justify your recommendations.

Candidates generally performed well on this task. Candidates did not receive credit for suggesting investigation of insignificant variables.

ANSWER:

Two variables for further investigation are vehicle L and station 3. The coefficients for these variables are positive and significant, indicating they increase turnout.time. I recommend that the city could investigate why vehicle L appears to be slower and what processes station 3 is using which contribute to increased turnout.time.

Task 8 (*11 points*)

On B's direction, A ran a classification GLM with the target variable set to whether the city meets its response time goal. Run the R code provided and analyze the output, focusing on the drop1 test.

(a) (*2 points*) Explain, for the top row of the drop1 test, how the AIC of 8764.1 is calculated from the deviance of 8736.1.

Full credit responses both described the formula for AIC and applied that formula to the situation at hand. Some candidates had trouble explaining how the penalty term was determined.

ANSWER:

AIC is defined as the deviance plus a penalty of 2 times the number of parameters/degrees of freedom. For this situation the deviance is 8736.1 and there are 14 degrees of freedom; 13 are shown in the output and there is an additional degree for the intercept. Twice this amount is added as a penalty to deviance to form the AIC, resulting in a calculated value of 8764.1 = 8736.1 + 14*2.

(b) (*2 points*) Explain how the results of the drop1 function suggest that only the **vehicle** variable be dropped.

Candidates generally performed well on this task. Effective responses mentioned that a lower AIC suggests a more effective model and identified that vehicle is the only variable that can be removed to lower the AIC.

ANSWER:

The AIC column from the model output indicates that dropping vehicle from the model results in a more parsimonious model with a lower AIC (full model AIC of 8764.1 vs. 8763.6 for the model with vehicle dropped). Removing additional variables does not further reduce AIC. Therefore, only vehicle should be dropped.

(c) (*1 point*) Identify a limitation of the drop1 function as shown by **vehicle**.

The key limitation shown by this example is that the drop1 functions drops all levels of a factor variable at once. No credit was awarded for identifying general weaknesses of drop1 that were not shown in this example, in particular that drop1 removes one variable at a time.

ANSWER:

A limitation of the drop1 function is that it drops all levels of a factor variable at once. The **vehicle** variable has 3 levels. The drop1 function will not tell us whether levels of vehicle have predictive power if considered individually.

In addition, the city manager also wanted straightforward explanations about the impact of several predictor variables on the program meeting its goal. In particular,

- What impact does a station serving a wealthy area have on response time?
- What impact does a station serving a college campus have on response time?
- What impact does a weekend or weekday have on response time?

You ask A to modify the predictor variables for answering the city manager's questions. Run the code provided by A in the .Rmd file.

(d) (*6 points*) Write a brief report (no more than a half page) based on the summary output to address the manager's questions. Write for a general audience.

Candidate performance was mixed on this task. The best candidates demonstrated understanding of the model output and logit transformation as well as explaining in simple terms how the 3 variables impact the response time goal. Common mistakes included incorrect interpretation of the link function, not providing impacts in terms of probability or odds, and incorrectly interpreting the target variable as response time rather than whether the response time goal was under 6 minutes.

ANSWER:

The odds of meeting the goal (response time under six minutes) are 23% lower for a call from a wealthy area compared to a call from a non-wealthy area, assuming all other characteristics of the two calls (such as time of day) are equal. Investigating the components of the response time, such as travel time, may give insight into whether the response time difference is due to distance or other factors.

The odds of meeting the response time goal for calls from ASU are 66% higher than for a non-college call. Since ASU is in the wealthy area, the correct interpretation is that the odds of meeting the response time goal for calls from ASU are approximately 66% higher than for a non-college call from a wealthy area.

To compare odds of meeting the goal for a call from ASU to the odds for a non-college call from a non-wealthy call, then both effects must be considered. The two effects are in different directions, so the wealthy area effect will partially offset the ASU effect. When the impacts are combined, a call from ASU has 27% higher odds of making the goal than does a call from a non-wealthy area.

The odds of receiving a response in under six minutes for a weekend call is 18% higher compared to weekday calls, all else equal.

Task 9 *(10 points)*

Your assistant decides to build a classification tree and notices that the structure of the tree is slightly different when Gini is used as the measure of impurity compared to when entropy is used.

(a) (*2 points*) Explain how measures of impurity are related to information gain in a decision tree.

Many candidates struggled with this task. Rather than describing the relationship between impurity and information gain, some candidates described differences between Gini and entropy, and some described the decision tree recursive splitting without mention of either impurity or information gain. These candidates received no credit. Partial credit was awarded for candidates who described the relationship between impurity and information gain but did describe how they are used in a decision tree.

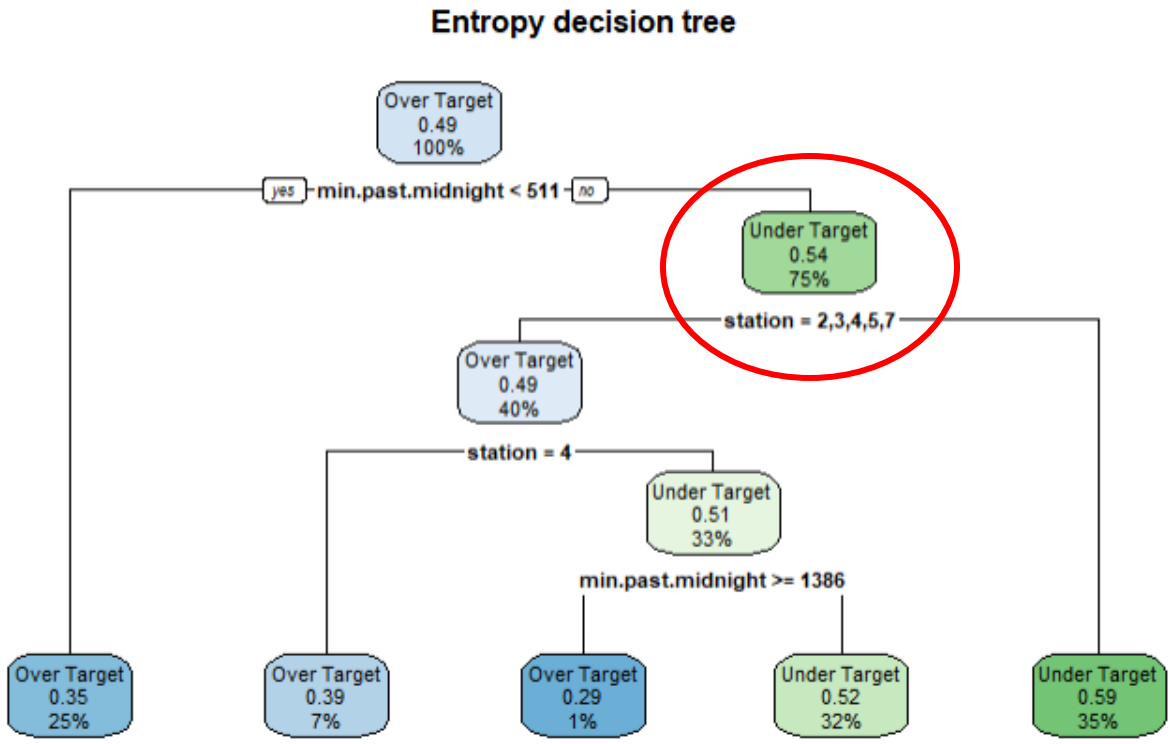
ANSWER:

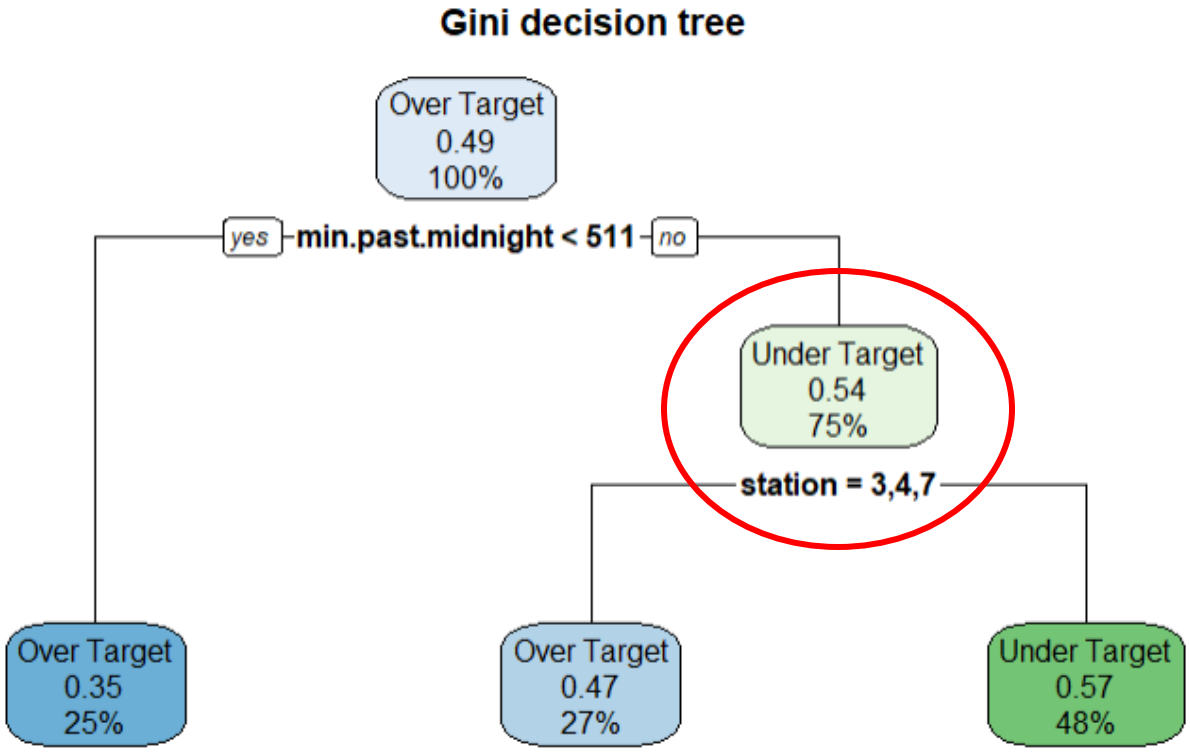
Information gain is the decrease in impurity created by a decision tree split. At each node of the decision tree, the model selects the split that results in the most information gain. Therefore, the choice of impurity measure (e.g., Gini or entropy) directly impacts information gain calculations.

The assistant creates two classification decision trees to identify the important variables, one using entropy as a measure of impurity and the other using Gini. See the tree diagrams below.

Both trees have the same first split based on min.past.midnight < 511, but the right sub-node based on specific stations (highlighted in both trees) split differently for the tree built using the Gini impurity measure compared to the tree built using the Entropy impurity measure.

(b) (*5 points*) Complete the missing values in the chart below to calculate the Gini impurity measure and Entropy impurity measure for the split chosen by the Gini Tree. Round all answers to 6 decimal places. Also, explain how the choice of Gini vs. Entropy as an impurity measure resulted in different splits in the tree.





Most candidates struggled with this task, with many candidates providing no response. Partial credit was available for candidates showing their work and providing the formulas for Gini and entropy.

Successful candidates were able to identify how the differences in information gain led to different splits in the decision trees based on the selected impurity measure. Candidates did not receive credit for making observations about the total number of splits or nodes in each tree.

ANSWER:

Chart with two highlighted cells to complete:

		Gini Tree Node Split			Entropy Tree Node Split		
	Primary Node	**Left Node**	**Right Node**	**Information Gain**	**Left Node**	**Right Node**	**Information Gain**
Over Target	3422	1418	2004		1992	1430	
Under Target	3963	1270	2693		1921	2042	
Total	7385	2688	4697		3913	3472	
Gini	0.497317	0.498484	0.489241	0.004712	0.499835	0.484465	0.004708
Entropy	0.996125	0.997812	0.984422	0.006829	0.999762	0.977470	0.006843

How the choice of Gini vs. Entropy as an impurity measure resulted in different splits in the tree:

The choice of impurity measure leads to different calculations of information gain for each tree. Using the Gini impurity measure, the information gain on the split chosen by the Gini Tree was higher than the information gain of the split chosen by the Entropy Tree: 0.004712 vs. 0.004708. However, for the entropy impurity measure the information gain of the split chosen by the Entropy Tree was higher: 0.006843 vs. 0.006829.

B is interested in a more accurate tree-based model but is concerned about the model variance.

(c) (*3 points*) Recommend whether to use a random forest or a gradient boosting machine given B's concern. Justify your recommendation.

Candidates generally performed well on this question, demonstrating excellent knowledge of both ensemble methods. Candidates that performed poorly on this question failed to provide a recommendation or did not sufficiently acknowledge B's concern around variance minimization.

ANSWER:

Given B's concern regarding variance, I recommend a random forest, which tends to do well in reducing variance while having a similar bias to that of a basic tree model. The variance reduction arises from the use of many small trees and sampling of the data (bagging). Both practices hinder overfitting to the idiosyncrasies of the training data, and hence keep the variance low.

Gradient boosting machines use the same underlying training data at each step. This is very effective at reducing bias but is very sensitive to the training data (high variance).

Task 10 (*6 points*)

Your assistant, A, builds a decision tree to investigate which variables have a significant impact on response time. The variable **day**, when used as a categorical variable, is deemed important by the tree-based model. A knows from experience, and from testing other models, that **day** is not actually a significant variable.

(a) (*2 points*) Explain why a decision tree model may emphasize **day**, when used as a categorical variable, despite it not being an important variable.

This task tested candidates' ability to recognize and explain why decision trees overfit to factor variables with many levels. Most candidates were able to identify the large number of levels as a concern. However, fewer candidates were able to explain why decision trees tend to select variables with many levels.

ANSWER:

Because day of month is coded as a categorical variable, the number of levels is 31. This means the number of ways to split day of the month into two groups is very large, making it likely that the tree will find spurious splits that happens to produce information gain for that particular training data.

Decision trees tend to create splits on categorical variables with many levels because it is easier to choose a split where the information gain is large. However, splitting on these variables will likely to lead to overfitting.

(b) (*4 points*) Describe the handling of categorical variables in linear models and tree-based models.

Most candidates received partial credit for this question. For the section on explaining how categorical variables are handled in linear models, almost all candidates recognized that binarization is necessary. Strong candidates provided more detailed descriptions, including a discussion of how the base level is determined, how to interpret the binarized coefficients against the base level, or that there are n-1 dummy variables produced (where n is the total number of levels).

For the section explaining how categorical variables are handled in decision tree models, most candidates recognized that binarization was not required. Strong candidates explained why binarization is not required (trees consider all possible groupings of the levels as potential splits), or how decision trees can split on a single categorical variable multiple times.

ANSWER:

Linear Models:

Linear models fit a coefficient for each level of a categorical variable except the base level. The coefficient for each level represents the impact relative to the base level of the variable. This is equivalent to "one-hot" encoding, which creates multiple new variables with value of 1 for observations at that level of the categorical variable and 0 otherwise.

Tree-Based Models:

Decision trees split the levels of a categorical variable into groups. The more levels the variable has, the more potential ways to split the category into groups. The decision tree algorithm will identify which variables to split and into which groups based on maximizing information gain. Decision trees naturally allow for interactions between categorical variables based on how the tree is created. For instance, a leaf node could have two or more parent nodes that split based on categorical variables, which would represent the interactions of those categorical variables. The tree may also split on the same variable more than once in the tree.

Task 11 (*9 points*)

Your assistant creates a new variable called **post.alarm.time** that equals the sum of only **turnout.time** and **travel.time**.

(a) (*2 points*) Assess using **post.alarm.time** instead of **response.time** as the target variable in the context of the business problem.

This task tested the candidates' ability to connect the new variable to the original business problem. Strong candidates recognized that using the new variable would require adjusting the original 6-minute target response time.

ANSWER:

Excluding alarm.processing.time removes processing time that is outside of TFMR's control from the problem, allowing the model to focus on what variables contribute to longer turnout.time and travel.time. Since the new variable would be lower than the old variable, we would also need to revisit the original goal of 90% of ALS calls having response times of six minutes or less. A target of less than six minutes would be appropriate since there will always be some alarm.processing.time.

Your assistant creates three random forest classification models to predict whether **post.alarm.time** is in the highest 10% of the original observations using three adjusted data sets, including one that oversamples the top decile of observations. A summary of the three training data sets is below:

Data Set	Target	Predictors	**hour** and **day** variables	Rows	With Oversampling
Df.Train.1	Post.Alarm	11	Included	7,846	No
Df.Train.2	Post.Alarm	9	Excluded	7,846	No
Df.Train.3	Post.Alarm	9	Excluded	11,808	Yes

(b) (*3 points*) Describe one benefit that each data set may have for creating a random forest model.

Candidates performed well on this task overall. Several candidates correctly identified benefits but did not describe what they were or how they pertain to the task. Strong candidates demonstrated understanding of how oversampling can help with imbalanced data.

ANSWER:

Df.Train.1:

Df.Train.1 includes all the predictor variables and observations from the original set. This data set allows for the random forest model to look at all potential variables in our data sets so we can assess, based on model results, which variables are most important for our predictions.

Df.Train.2:

The day and hour variables are factor variables with many levels and may only have slightly more information than the weekday/month variables. Given the random forest model's greediness with many-level factors variables and the lack of large potential information gain from these variables, removing them will allow us to help our random forest avoid overfitting.

Df.Train.3:

Df.Train.3 uses oversampling to create a more even balance for our target classes. This should allow the model to focus more on accurately predicting the observations in the highest 10%, which is the group we are most interested in. The size of the dataset is still relatively small so computing issues should not impact the oversampled dataset.

Your assistant provides the three corresponding sets of model outputs, shown in the table below:

Model Name	Train Dataset Observations	Test Dataset Observations	Model AUC on Train Dataset	Model AUC on Test Dataset	mtry Value	nodesize Value
rf.1	7846	1961	0.9997425	0.5587454	3	1
rf.2	7846	1961	0.9979403	0.5447259	3	1
rf.3	11808	1961	0.9995766	0.5633839	3	1

(c) (*2 points*) Explain what led to the large difference in AUC values between the train and test datasets.

To earn full credit, candidates needed to either explain why overfitting may occur (explaining differences between training and test dataset) or provide an explanation of what AUC captures and how to interpret this.

ANSWER:

The AUC outputs above show a significant decline in predictive power from the training dataset to the test dataset. An AUC of 1.0 indicates the model perfectly predicts the data while an AUC of 0.5 indicates the model performs as well as random chance. In this case, the models perform very well on the training data and very poorly on the test data. This indicates that the models are overfit to noise in the training datasets. The likely cause is that the chosen hyperparameters create very deep trees, and therefore the hyperparameters need to be adjusted to create simpler trees which are less prone to overfitting.

(d) (*2 points*) Recommend an adjustment to **either** mtry **or** nodesize to address the decline in AUC between the train and test datasets. Justify your recommendation.

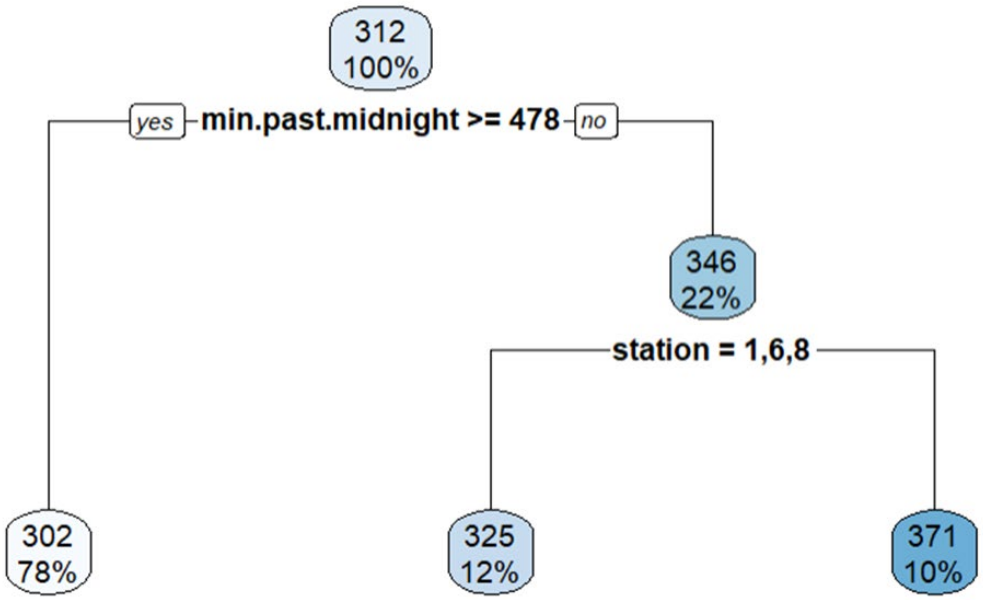
To earn full credit, the candidate needed to explain what the chosen parameter controlled in the model, in addition to why it would help lower the test AUC as part of their recommendation. While some candidates referenced and defined these parameters in Part c, the candidate was required to explain the hyperparameter in this part alongside their recommendation.

ANSWER:

I recommend increasing nodesize, the minimum number of observations contained in a final node. It had been set to 1, the default for classification trees. Increasing nodesize will reduce complexity by limiting tree growth when the data on which the split is based becomes sparse. Reducing complexity will reduce the overfitting that is causing the decline in AUC.

Task 12 (*9 points*)

Your assistant, A, creates a simple regression tree to better understand the drivers of response time and concludes, based on the regression tree below, that the only important variables for a decision tree model are **minutes.past.midnight** and **station**.



(a) (*2 points*) Critique A's conclusion that the other variables are not important.

Although responses were varied, candidates performed well on this task overall. Successful candidates included a statement that the two predictors may have been most important for this particular training data, model type, and set of parameters, but that other predictors may also be important. For full credit, the response required a description of supplemental analysis that would better support A's conclusion, for example changing the complexity parameter, increasing maxdepth, or comparing to other trained predictive models.

Some candidates identified the limited complexity in the tree but did not provide a clear critique of the statement that other variables are not important. Simply stating that the tree was too simple is not enough to earn full credit. Several candidates misinterpreted the inequality for min.past.midnight as "less than or equal", or did not continue to use the right branch as "No" for the split on station.

ANSWER:

The choice of model parameters influences which and how many splits the model makes. For example, changing the complexity parameter will impact which variables are used in the decision tree. There is not one specific parameter value that is correct for all decision trees and any parameter choice comes with trade-offs between model complexity and model accuracy. While this is the tree model produced by the default specifications, that does not imply that variables not included in this model are irrelevant, only that they were not used in this particular model. A should conduct further analysis to support the selected model parameters.

The city manager reviews the tree and points out that the left two nodes add up to 90% of the data and are both less than the 360-second target. The city manager states that this means the response time is six minutes or less for 90% of calls and the City of Tempe has reached their goal.

(b) (*2 points*) Explain for a general audience why this interpretation is not correct.

Many candidates conveyed that the number in each node reflects the average response time but does not include variance or guarantee that all observations within the node are below the target time. The strongest candidates used language that was clear and interpretable and avoided technical terms. A common response provided an example of data that would average below 360 seconds but included outliers above the target or referenced the root node and how the manager's interpretation would mean 100% of observations met the target which is clearly incorrect.

Several candidates incorrectly stated that this is a classification tree with the percentages representing the probability of an observation being below 360 seconds. Another common mistake was stating that terminal nodes cannot be combined, which is not correct in this context.

ANSWER:

The numbers that the city manager is pointing out are averages and do not give any information about the spread of response times. Even though the averages for 90% of the observations are below six minutes, that does not mean that 90% of the observations that led to those predictions are below six minutes. For example, if the average of all response times were five minutes, it could still be the case that 50% of calls lasted for three minutes and 50% of calls lasted for seven minutes, which would not meet the response time goal.

(c) (*2 points*) Interpret the meaning, for a general audience, of the right-most node. Include a description of what each of the splits leading to that node means.

This tested the candidate's ability to interpret and describe the decision tree's splits to a non-technical audience. Most candidates were able to define the underlying observations included in the right node and accurately describe the numbers presented.

A common mistake was not elaborating further on what each split determines/accomplishes leading up to the rightmost node. Full marks were awarded when each of the criteria for station and min.past.midnight were clearly described in order based on the tree hierarchy or followed if-else logic to arrive at the terminal node, not simply stated.

Candidates were expected to describe each of the numbers included in the node as the average response time of the training data (or alternatively the resulting predicted response time) and the portion of training data that makes up that node. Some candidates who did not correctly identify the tree as a regression tree described the listed percentage as node purity which is incorrect.

ANSWER:

The top split represents when min.past.midnight is greater or less than 478, roughly 8:00 a.m. The right split is observations before 8:00 a.m. The right sub-split differentiates between response times coming from stations 1, 6, 8 and stations 2, 3, 4, 5, 7.

The right-most prediction represents calls coming between midnight and 8:00 a.m. at stations 2, 3, 4, 5, and 7. About 10% of all observations fall into this category and these observations have the highest average response time of 371 seconds.

B has asked you to build a random forest to understand which predictors are the most important for achieving the City of Tempe's goal of reducing ALS response times.

(d) (*3 points*) Describe both the challenge of interpreting a random forest model and a method to identify which predictors from a random forest model the City of Tempe should focus on. Do not build a random forest model.

This tested a candidate's ability to clearly explain why a random forest cannot be interpreted like a single tree, as well as provide a comprehensive description of method to identify high priority predictors. Most candidates received partial credit for part d, however few candidates received full credit.

Nearly all candidates were able to identify the complexity of interpreting random forests due to the large number of trees involved and give some indication that visualization is near-impossible. Credit was also given for candidates who appropriately described the process for building a random forest and that averaging the results of independent trees built from randomly available features at each split leads to ambiguous relationships to predictors.

Candidates struggled to identify a method to identify predictors on which to focus. Partial credit was given for appropriate mention of variable importance or partial dependence. Exceptional candidates not only identified variable importance or partial dependence, but additionally provided a high-level description of how it is calculated and how it would help the city prioritize. Credit was not given for candidates who identified variable importance simply as "importance" or described variable importance plots without correct identification or defining importance. A handful of candidates included bootstrapping, cross validation, and hyperparameter tuning, however these relate to building and refining random forest performance, and do not identify specific predictors on which to focus.

ANSWER:

A random forest is difficult to interpret because, unlike a decision tree where the splits and the impact of those splits can be observed, a random forest is made up of the aggregated results of hundreds or thousands of decision trees. Directly observing the component decision trees is generally uninterpretable or in some cases not possible.

Variable importance is a measure of how much a predictor contributes to the overall fit of the model. This can be used to rank which predictors are most important in the model. It is calculated by aggregating across all trees in the random forest the reductions in error that all splits on a selected variable produce. Variable importance cannot be used to draw inference as to what is causing model results but can identify which variables cause the largest reduction in model error on the training data.

Task 13 (*6 points*)

B has asked you to build a decision tree to better understand how to achieve the City of Tempe advanced life support (ALS) goal. B asks you to create a new target variable where the target variable is a categorical variable with a value of 1 if the response time is 6 minutes or less and a value of 0 otherwise.

You ask A to fit a decision tree using cost complexity pruning.

(a) (*3 points*) Explain how cost-complexity pruning works, including how complexity is optimized.

For full credit, candidates needed to cover the cost-complexity pruning process starting from building a large tree all the way through complexity parameter optimization. Partial credit was awarded for only covering a subset of the process, for example focusing solely on decision tree pruning.

ANSWER:

Cost-complexity pruning involves growing a large tree and then pruning it back by dropping splits that do not reduce the model error by a fixed value determined by the complexity parameter. We can use cross validation to optimize the complexity parameter, which is the process repeatedly training models and testing models on different folds of the data. This is done for different values for the complexity parameter, and the one with the lowest cross validation error is selected as the optimal choice. We then prune back our trained tree using the complexity parameter from the cross validation.

A tells you that re-running the model produces different results. To ensure consistent results, you ask A to set a random seed prior to running the model.

(b) (*1 point*) State why changing the random seed would affect the tree constructed using cost-complexity pruning.

Successful candidates correctly identified where randomization is required in cross validation. Candidates that performed poorly failed to communicate the reasons why randomization occurred.

ANSWER:

Each time the cost-complexity pruning algorithm is run, the splits of data used in the cross validation are randomly assigned. This makes the calculation of the optimal complexity parameter dependent on the seed selected. Therefore, changing the seed can result in different pruned trees.

You ask A to prepare a confusion matrix for the decision tree model. A produces the following:

```
Confusion Matrix and Statistics

              Reference
Prediction    Over_Target  Under_Target
Over_Target            87            77
Under_Target          419          1377

               Accuracy : 0.7469
                 95% CI : (0.7271, 0.7661)
    No Information Rate : 0.7418
    P-Value [Acc > NIR] : 0.3131
```

```
                 Kappa : 0.1526

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.17194
            Specificity : 0.94704
         Pos Pred Value : 0.53049
         Neg Pred Value : 0.76670
             Prevalence : 0.25816
         Detection Rate : 0.04439
   Detection Prevalence : 0.08367
      Balanced Accuracy : 0.55949

       'Positive' Class : 0
```

(c) (*2 points*) Recommend a measure relevant to the business problem and assess whether this is a useful model. Justify your recommendation.

Successful candidates determined that model was not useful and were able to justify this by correlating this to one of the following measures: Accuracy, Balanced Accuracy, Sensitivity, or Positive Predictive Value (Precision). Candidates that performed poorly either failed to mention a specific measure or stated that specificity was a relevant measure, thereby recommending use of the model.

ANSWER:

Tempe is most concerned about the results that are over target, since those are causing them to miss their target of 90% of responses being under 360 seconds. Therefore, I recommend assessing the model with the sensitivity measure. Unfortunately, it is 17% (87/506), indicating that the model does a poor job predicting the results that are above target, and hence it is not a useful model for the business problem.