Exam PA April 14 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 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. 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

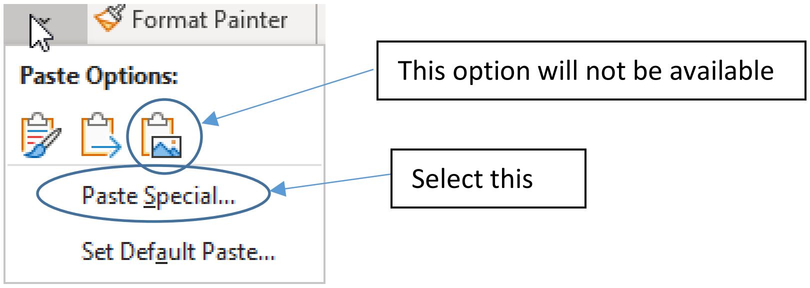
Variable Name	**Variable Values**
issue	Type of emergency event (11 categories)
vehicle	L, E indicate the two most common vehicles. X is all others.
station	1 to 8
hour	0 to 23, hours past midnight
min.past.midnight	0 to 1439
month	1 to 12
day	1 to 31
weekday	1 to 7 for Sunday to Saturday
dewpoint	a 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.

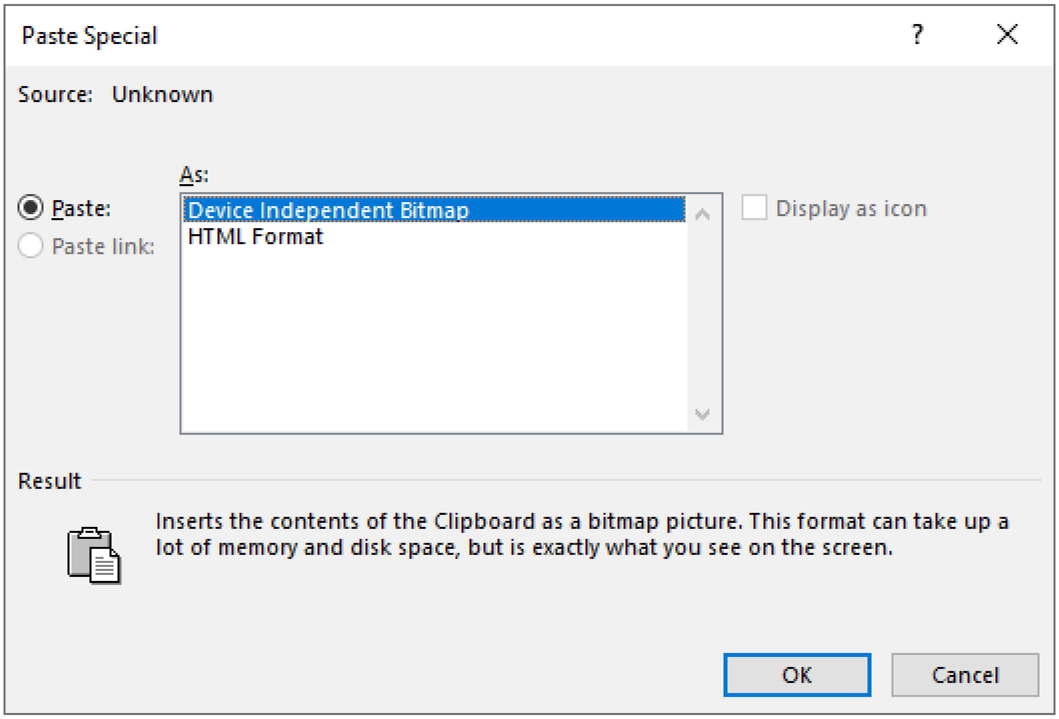
IMPORTANT NOTE: When pasting a picture from RStudio to Word, there is only one approach that will work. After right clicking on the image in RStudio and selecting “copy” the following steps need to be taken in Word. On the Home menu, click on the down arrow under “Paste” and then select “Paste Special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



This option will not be available

Select this

From this dialog box, make the indicated selection.


Paste Special

Source: Unknown

As:

● Paste:
○ Paste link:

Device Independent Bitmap
HTML Format

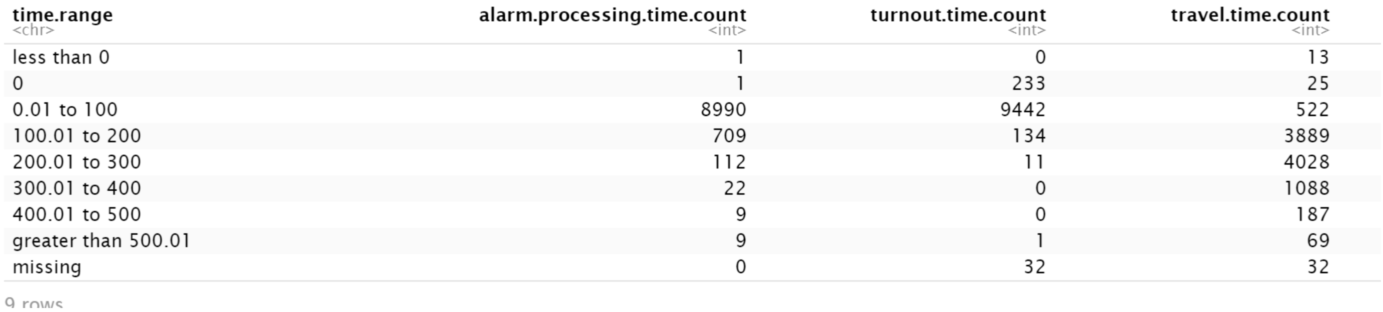
□ Display as icon

Result

Inserts the contents of the Clipboard as a bitmap picture. This format can take up a lot of memory and disk space, but is exactly what you see on the screen.

Task 1 (*10 points*)

Your assistant summarized the count of observations for each of the three response time variables (**alarm.processing.time, turnout.time**, and **travel.time**) by ranges of time, and your boss has requested that you evaluate the resulting distributions for reasonableness. Your assistant provided the table below:



time.range	alarm.processing.time.count	turnout.time.count	travel.time.count
<chr>	<int>	<int>	<int>
less than 0	1	0	13
0	1	233	25
0.01 to 100	8990	9442	522
100.01 to 200	709	134	3889
200.01 to 300	112	11	4028
300.01 to 400	22	0	1088
400.01 to 500	9	0	187
greater than 500.01	9	1	69
missing	0	32	32

9 rows

(a) (*6 points*) For each of the response time variables, using the table above:

i. Evaluate the plausibility of the zero and outlier values in the data.

ii. Discuss implications for the business problem.

Candidates struggled with this task. While most candidates were able to discuss that zero and negative values need further investigation and may be errors, few candidates discussed of potential outliers greater than 500.01. The second part of the question asked about business implications. Most candidates discussed modeling implications, with little or no connection to the business problem.

FIRST ANSWER (alarm processing time):

Evaluate the plausibility of the zero and outlier values in the data:

There are two observations with either a 0 or negative time, which are not reasonable and should either be corrected or removed from the data. The count of values greater than 500 seconds appears plausible, but we would need more information about the actual values to know whether they are all plausible. For example, a value of 700 is plausible, but a value of 10,000 probably is not plausible.

Discuss implications for the business problem:

Alarm processing time contributes approximately 1 to 2 minutes of the response time and is generally the second smallest component. If the City of Tempe has no influence over alarm processing time since dispatching is done by a regional organization, then studying it is futile. The analysis should focus on the components of response time that the client can influence.

SECOND ANSWER (turnout time):

Evaluate the plausibility of the zero and outlier values in the data:

There are 233 observations with a turnout time of 0 seconds. This sounds unreasonable unless TFMR receives the call while already in transit. It is important to find out more about the process to see how these zero values arise and determine if they are valid. The chart only indicates that the highest value is more than 500 seconds, which may or may not be plausible depending on how high it is.

Discuss implications for the business problem:

Approximately 97% of turnout times are under 100 seconds, so this is the smallest component of the response time. Despite this, it may be that turnout time is the easiest to reduce, making understanding it worthwhile. Its small size, relative to travel time, however, provides reasonable justification for first focusing efforts on understanding and aiming to reduce travel time before addressing turnout time.

THIRD ANSWER (travel time):

Evaluate the plausibility of the zero and outlier values in the data:

The 13 negative observations do not appear plausible and should either be removed or modified in some way. The 25 zero-second observations should be considered more carefully, similar to turnout time. It could be that the vehicles were already on the scene, a legitimate reason, or it could represent a failure of the measurement system.

Discuss implications for the business problem:

The travel time has the highest average time and variability of the three components of response time. Modeling this variable and understanding what contributes to long travel times may provide valuable insights.

(b) (*4 points*) Justify each of the following approaches for addressing the missing values in **turnout.time** and **travel.time**, assuming that each is being used as a predictor variable:

i. Imputing missing values based on the other variables in the data
ii. Removing the observations that contain the missing values

Overall, candidates struggled to justify imputation. Some candidates provided a comparison of different imputation approaches rather than a justification for whether imputation should be used at all. No credit was awarded for these responses.

ANSWER:

Imputing missing values based on the other variables in the data:

The only variables with missing (or "NA") values are turnout time and travel time. Imputing these values from other variables avoids having to discard those observations entirely. This is especially important if the reason for the missing values is systematic, for example, if the missing values only occur on certain types of calls where the responder is unable to record the data.

Removing the observations that contain the missing values:

Removing the observations with missing values is a simpler approach with a cost of losing any valuable information that is contained in the removed observations. Since the dataset does not contain a high number of missing values, this is less of a concern. Missing values may also be a sign of data quality issues throughout the observation and by removing them we avoid the impact that inaccurate data may have on our efforts.

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:

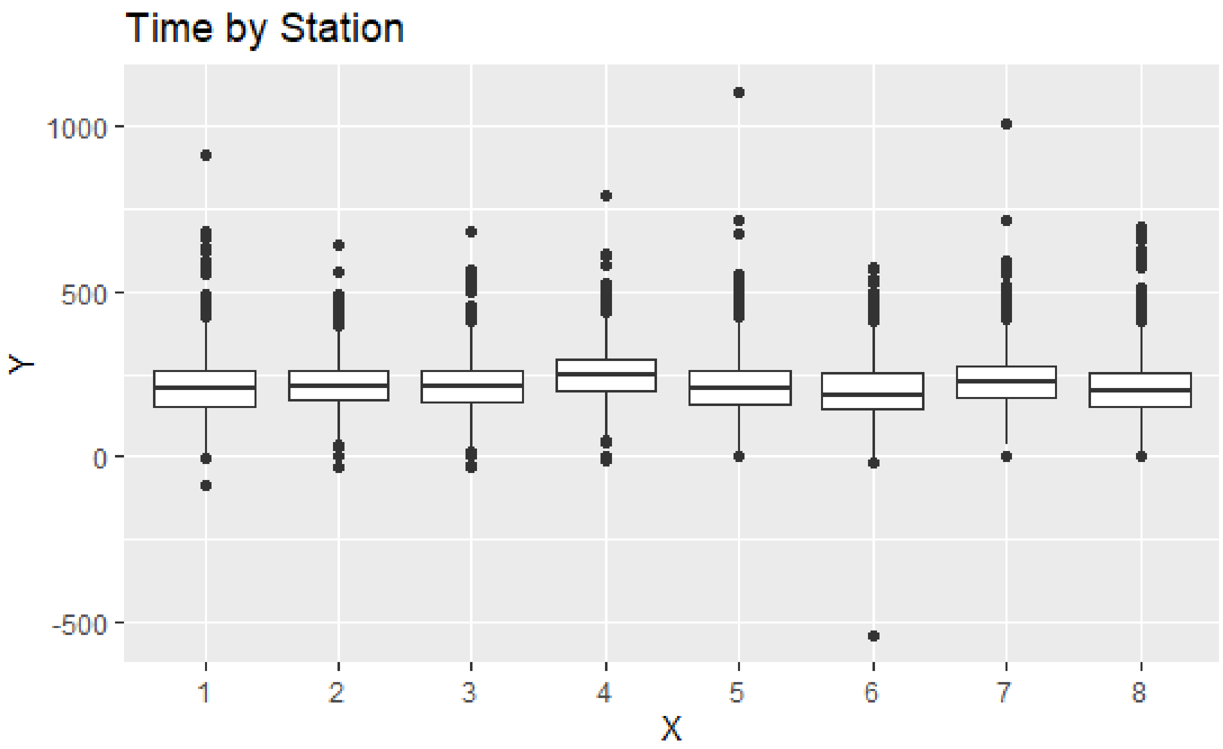
Are there any subject matter experts that you would recommend talking to prior to performing the analysis?

Question 3:

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

Task 3 (*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.

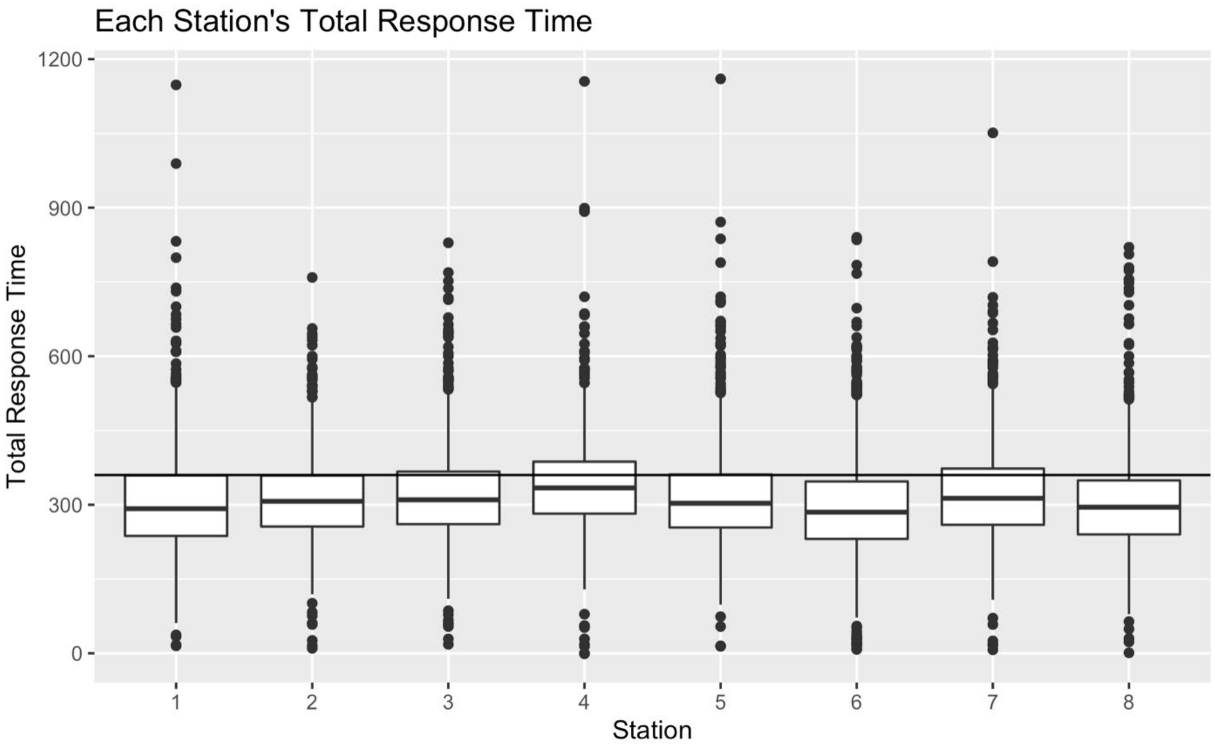
ANSWER:

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 4 (*12 points*)

Your assistant decides to investigate whether there is a relationship between response time and the frequency of calls. Your assistant is planning to model call frequency using a GLM based on the variables **min.past.midnight**, **station**, and **weekday.number** as factor variables.

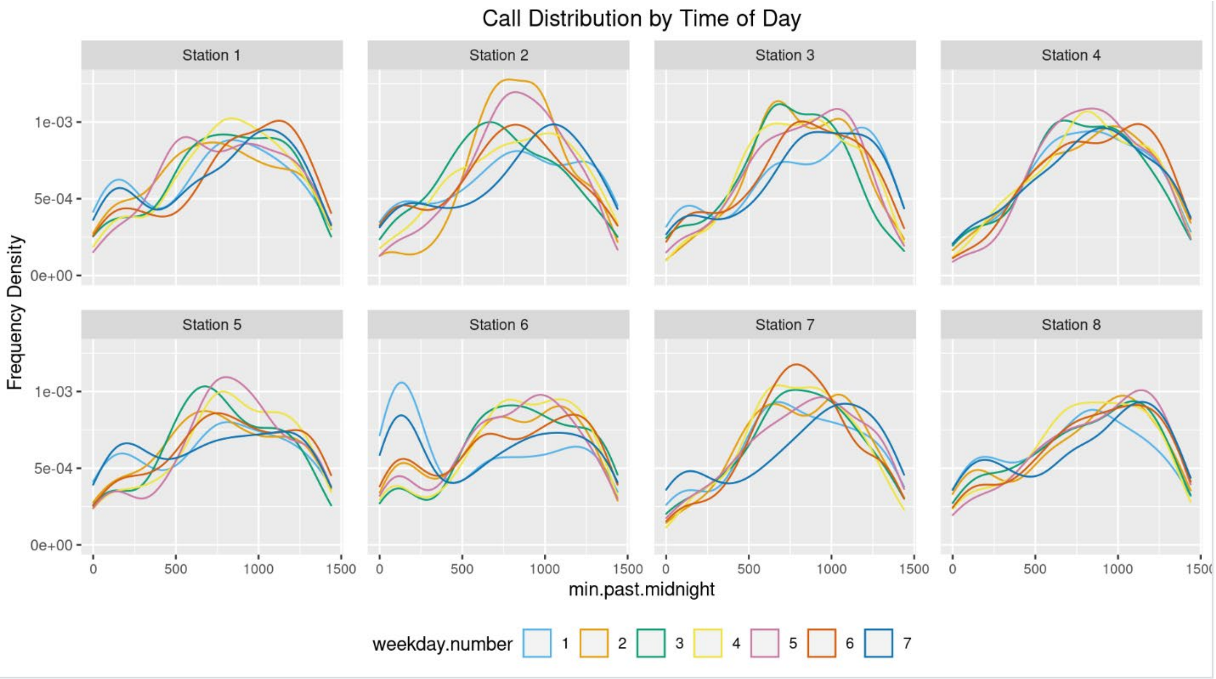
(a) (*2 points*) Describe how using these three predictor variables as factor variables will lead to a GLM with high variance.

Candidates performed well on this task overall, correctly identifying the high dimensionality problem that results from treating the three variables as factors. Only partial credit was given to generic statements about dimensionality and variance without relating to the three variables in question.

ANSWER:

The variable min.past.midnight has many different levels if used as a factor instead of numeric. It would likely lead to a model with a high variance due to the curse of dimensionality. Each level would have its own coefficient, allowing for a highly complex model that fits to both the signal and the noise in the training data. The result is a model with a high variance. The other variables, station and weekday number, could have a similar effect, though they start with more reasonable numbers of levels.

Your assistant created the graph below.



(b) (*3 points*) Recommend, separately for each of the three predictor variables, a transformation that may reduce the overall prediction error of the GLM. Briefly justify each recommendation.

A wide variety of recommendations were accepted provided they were suitable to each variable. No credit was awarded for generic responses that did not relate to these specific variables.

ANSWER:

Minutes past midnight:

Minutes past midnight could be transformed to be a two-level factor variable (call it overnight) to distinguish calls that occur between midnight and 6 am from those that occur later. Reducing the number of levels this drastically will reduce the variance from this variable, and clearly different call frequencies are seen before and after 6am for many stations.

Station:

Stations could be grouped together to create a smaller number of levels, again to reduce overfitting. One grouping could be station 6 or others since station 6 has a distinct shape that is different from the others when looking at the calls by time of day and is the only station serving ASU.

Weekday number:

Weekday could be grouped to be weekday or weekend. This makes sense considering the similarly of days within these groups as seen in the graphs, e.g., weekends exhibit a different pattern than weekdays at several stations.

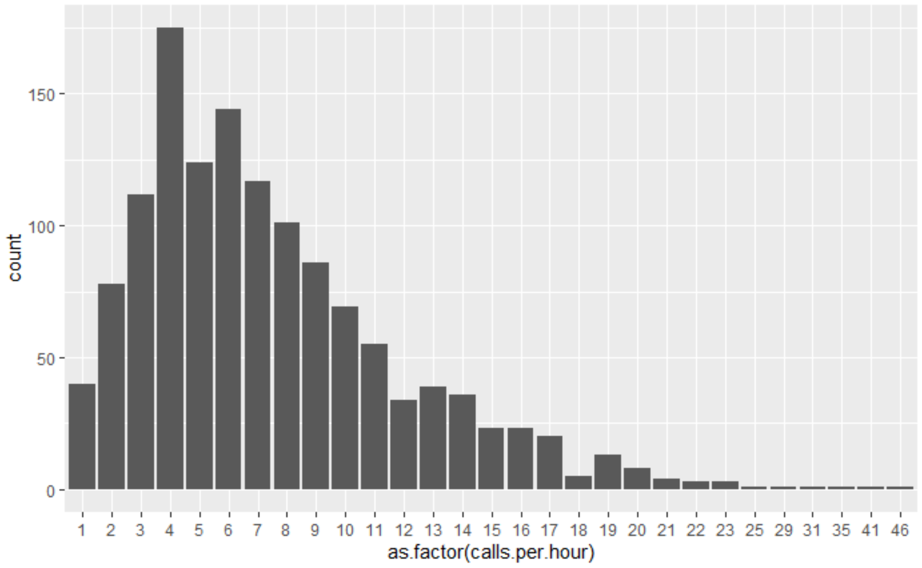
(c) (*2 points*) Recommend an interaction term given the three transformed variables in part (b). Justify your recommendation.

Candidates did well on this question overall. Most candidates were able to recommend an interaction term, but fewer candidates were able to provide a reasonable justification.

ANSWER:

The call frequency distributions are very different for station 6 vs the other stations in how the difference between overnight and non-overnight hours is less pronounced. To capture this difference in the formula for a GLM, I recommend creating an interaction between the overnight variable and station.

Your assistant decides to model claim frequency as the number of calls per hour and creates a new variable **calls.per.hour** by grouping calls according to **station**, **weekday**, and **hour**. Your assistant creates the histogram below of **calls.per.hour** and decides to model it with a GLM with Poisson family.



(d) (*3 points*) Critique both your assistant's data preparation for predicting calls per hour and the choice of the Poisson distribution.

This subtask was removed from grading due to a significant error in the wording preceding the subtask.

ANSWER:

(e) (*2 points*) Explain how manually binarizing the factor variable **station** prior to fitting a GLM can impact *p*-values in the summary output, even when the fitting function automatically binarizes factor variables.

This part of the question is to test candidates' understanding of model coefficients and the fact that the base level of the factor variable is included in the intercept coefficient. No credit was awarded for discussion of how manual binarization can facilitate dropping individual factor levels since dropping levels does not impact p-values.

ANSWER:

The glm() function internally binarizes all levels of a factor variable except for the base level. Therefore, changing the base level will cause different p-values to be calculated because these relate to the hypothesis that another variable has a different impact than the base level. If this base level is automatically selected, it could lead to cases where a coefficient may appear to be significant but is not in fact significant. Manually binarizing **station** is one way to select the base level that p-values are calculated relative to.

Task 5 (*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.

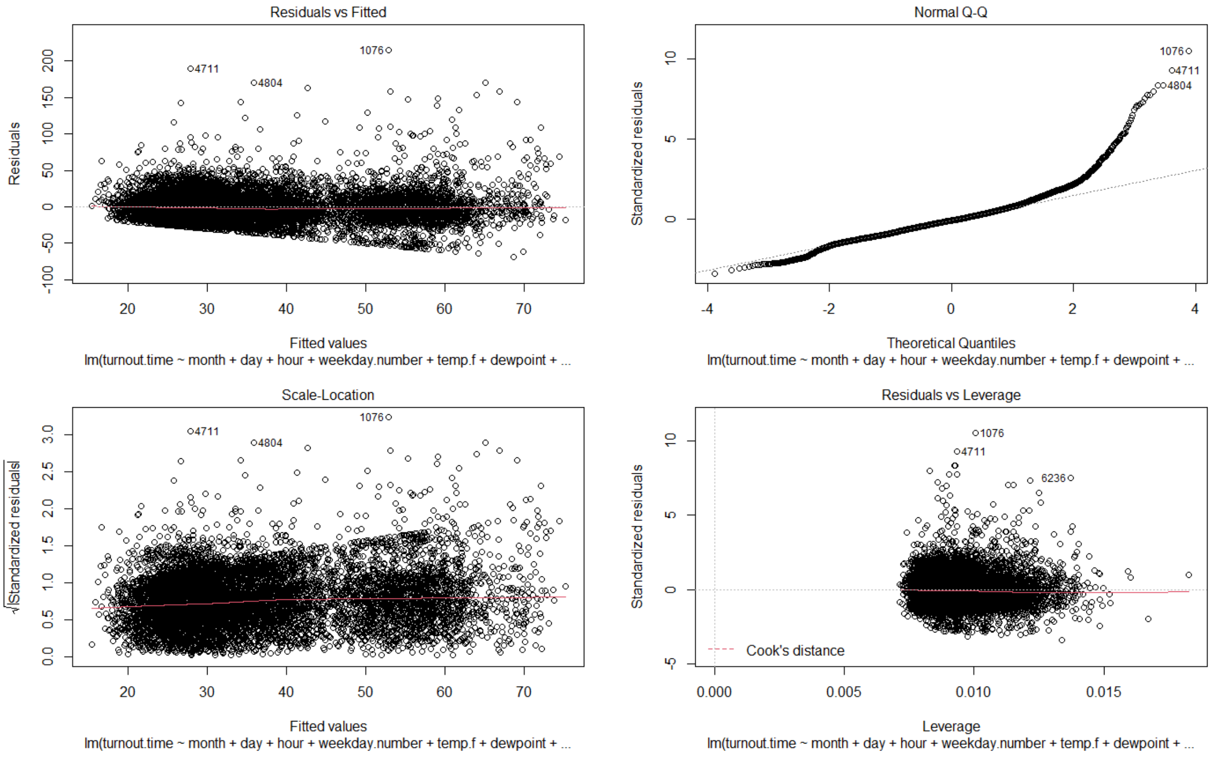
A target variable that could be modeled with the Poisson distribution is the number of calls in an hour.

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 6 (*10 points*)

Your boss wants to investigate the effect of **station** and **vehicle** on **turnout.time**. Your assistant builds one GLM with just the station variable and another GLM with both the station and vehicle variables, as seen in the .Rmd file.

(a) (*3 points*) Explain why intercepts for the two models are different.

Most candidates struggled with this task. Strong candidates made the key observation that the intercept represents the base level of all factor variables. Weaker responses tended to discuss the data, or technical details of the fitting process without relating those explanations to the intercept.

ANSWER:

The intercept is used to calculate the mean turnout time when all factors are the base level and all numeric variables are zero. The first model's intercept corresponds to the mean turnout time for the base level station across all levels of vehicle. The second model's intercept corresponds to the mean turnout time for the base level of station and the base level of vehicle. The second model's intercept is different since it only applies to the base level of vehicle whereas the first model's intercept applies across all levels of vehicle.

Your boss suggests there might be an interaction effect between **station** and **vehicle**. Your assistant has set up two models, one with an interaction term and one without. Run each of the models.

(b) (*2 points*) Compare the performance of the two models.

Candidates performed well on this task overall. Successful candidates analyzed model performance using a metric such as AIC, including a conclusion around which model performed better. Several candidates discussed model interpretability, and these discussions were not awarded any credit.

ANSWER:

The model with no interaction has an AIC of 89835 while the model with interaction has an AIC of 89767. The lower AIC of the model with interaction term indicates a fit that is good enough to justify the additional complexity from adding more explanatory variables.

(c) (*5 points*) Prepare a communication, no longer than half a page, to the city manager regarding the turnout time of station 3 compared to other stations based on the output of the model with the interaction term. Write for a general audience.

Most candidates received at least partial credit on this task. Full credit was awarded for using plain language to compare the turnout time in Station 3 to other Stations, including commentary regarding the impact vehicle had on turnout time. Common reasons for receiving partial credit were inaccurate analysis of coefficients and not writing for a general audience.

A key challenge in interpreting the interaction terms is that not all stations have all vehicles. For example, vehicle L only appears at stations 3 and 6, and the base level is set as station 6.

ANSWER:

Based on our model, station 3 has the highest turnout time among all stations, no matter which vehicle is used. When using vehicle E, this is only 3% higher than station 6, the next highest for vehicle E. However, the difference increases substantially with other vehicles. For vehicle L, turnout time for station 3 is 35% higher than it is for station 6, the only other station with this vehicle. For vehicle X, turnout time is 63% higher than that for station 4, the next highest of the seven stations using this vehicle. Consider having station 3 consult with station 6 about managing turnout time effectively where all three vehicle types are housed at the station.

Task 7 (*10 points*)

B asks A to explore the use of elastic net regression to better understand what predictor variables would be most impactful in meeting the city's goals.

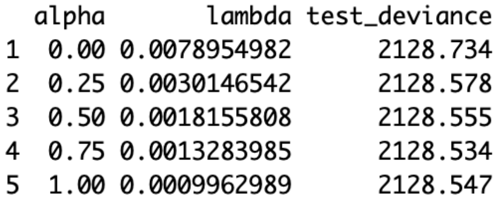
(a) (*4 points*) Explain how elastic net regression works.

Strong candidates provided both the motivation behind elastic net regression (shrinking coefficients and reducing variance) and a technical description of how elastic net regression achieves the goal. Candidates were not awarded credit for discussion of cross-validation and hyperparameter tuning, which goes beyond what this task asks for.

ANSWER:

Elastic net regression incorporates a penalty term into the loss function that is based on the magnitude of the coefficients. This penalty shrinks the coefficients, decreasing the variance in exchange for increased bias relative to a traditional GLM model. Elastic net uses a combination of the ridge and lasso penalties, which use different functions on the coefficients and have different effects in the optimization. A hyperparameter, alpha, determines the proportion of the lasso penalty, with the remainder being a ridge penalty. Elastic net allows the modeler to combine the more effective shrinkage of large coefficients from ridge with the ability to shrink coefficients all the way to zero from lasso.

Your assistant produces the following results when testing five values of alpha for an elastic net regression model.



```
  alpha       lambda test_deviance
1  0.00 0.0078954982      2128.734
2  0.25 0.0030146542      2128.578
3  0.50 0.0018155808      2128.555
4  0.75 0.0013283985      2128.534
5  1.00 0.0009962989      2128.547
```

(b) (*1 point*) Select the best elastic net model using these results. Justify the selection.

Candidates performed well overall on this task. Most candidates were able to select the optimal value. No credit was awarded to candidates who selected alpha = 0 because it is ridge regression or alpha = 1 because it is lasso since these justifications were considered arbitrary.

ANSWER:

The best model occurs at the lowest value of test deviance. The model output indicates that the lowest test deviance is at alpha = 0.75 and the corresponding lambda = 0.0013283985.

The corresponding GLM without regularization has a higher test deviance than the elastic net model.

(c) (*2 points*) Explain what the higher test deviance indicates about the GLM without regularization.

Candidates had to identify that the GLM was overfit to the test data for full credit. No credit was awarded for discussion deviance or regularization without connecting the concepts to model performance on training as opposed to test data.

ANSWER:

The higher test deviance of the GLM model indicates the model has overfit the training data more than the elastic net model has. This overfitting results in higher deviance when evaluating the model on the unseen test data.

The coefficients below are for a logistic regression with canonical link function to predict whether a response time will meet the city's goal. The **college** and **wealthy** variables are indicator variables based on **station**.

(Intercept)	0.912
dewpoint	-0.008
temp.c	0.014
college	0.530
week.end	0.118
wealthy	-0.303
AM	-0.435

(d) (*3 points*) Describe what impact station 6 has on meeting the goal compared to other stations, all else equal.

Few candidates received full credit for this task. For full credit, candidates needed to identify that both wealthy and college indicators applied to station 6 and compare the impact of the coefficients.

ANSWER:

Station 6 is both a college station and a wealthy station, so both coefficients apply. All else equal, compared to other wealthy stations (4, 5, and 7), its linear predictor is 0.530 higher; compared to the remaining stations, the linear coefficient is 0.530 - 0.303 = 0.227 higher. The linear predictor is the logit or log odds of the probability. It is easier to express these as increases in odds: Station 6 has exp(0.530) - 1 = 70% higher odds of being successful than other wealthy stations and exp(0.227) - 1 = 25% higher odds of being successful than non-wealthy stations.

Task 8 *(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:

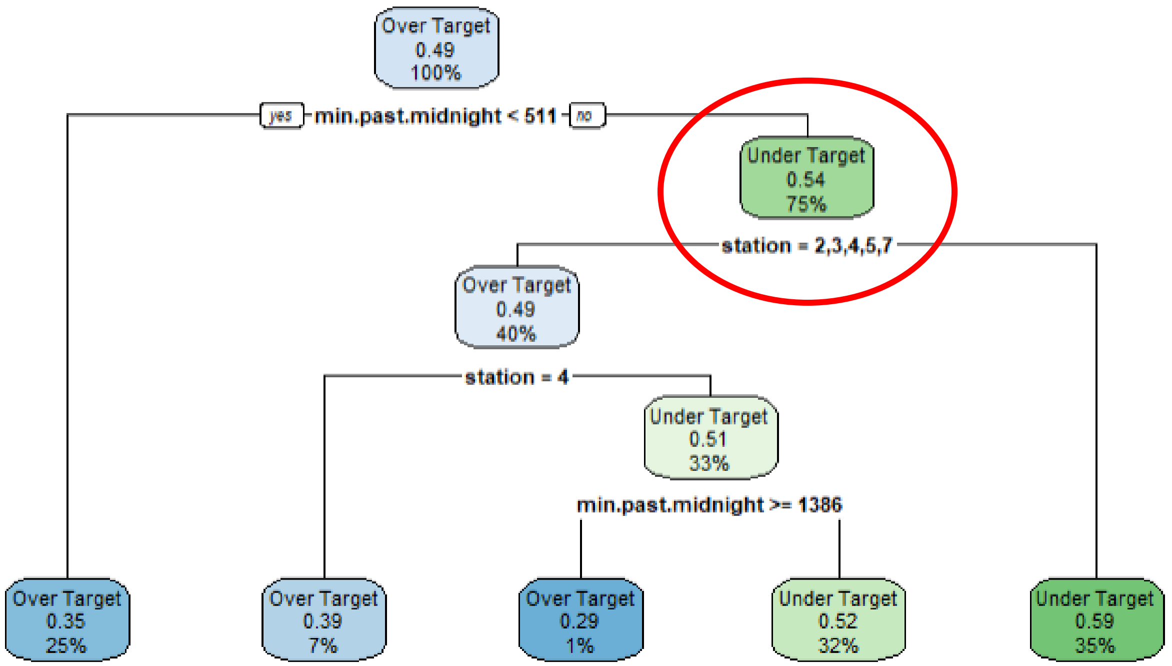
Information gain is the decrease in impurity created by a decision tree split. At each node of the decision tree, the model selects the split that results in the most information gain. Therefore, the choice of impurity measure (e.g., Gini or entropy) directly impacts information gain calculations.

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.

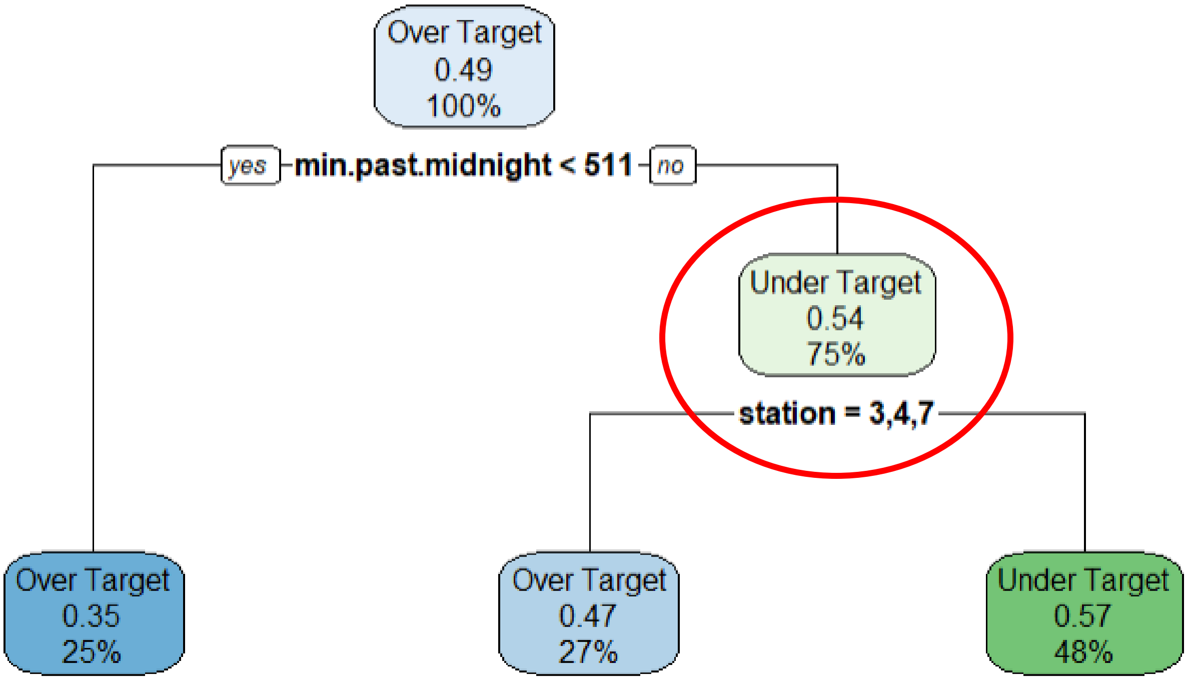
Both trees have the same first split based on min.past.midnight < 511, but the right sub-node based on specific stations (highlighted in both trees) split differently for the tree built using the Gini impurity measure compared to the tree built using the Entropy impurity measure.

(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.

Entropy decision tree



Gini decision 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	[redacted]	1992	1430	[redacted]
Under Target	3963	1270	2693	[redacted]	1921	2042	[redacted]
Total	7385	2688	4697	[redacted]	3913	3472	[redacted]
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 9 (*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 10 (*10 points*)

Your assistant creates two classification trees to assist the city of Tempe in meeting their goal. Your assistant uses the 75th percentile of **turnout.time** and **travel.time** respectively as cutoffs to define the target variables.

(a) (*4 points*) Critique three aspects of your assistant's model design.

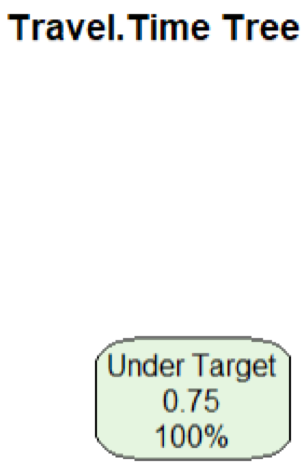
Most candidates received at least partial credit for this question. For full credit, candidates needed to provide three distinct critiques with clear explanations. The critiques in the model solution are a few examples of valid critiques that would receive full credit.

ANSWER:

This model design is not very applicable to the goal for the city of Tempe of having 90% of advanced life support (ALS) calls arrive in six minutes or less.

1. The city of Tempe's target concerns response time, which has three components: alarm processing time, turnout time, and travel time. If turnout time and travel time are modeled separately, then the alarm processing time component should also be modeled separately instead of being ignored.
2. Splitting the response time into its components may be useful for identifying variables that affect individual time components, but it will make it more difficult to identify variables that have impacts across multiple parts of the process. In addition, an observation that is at the 75th percentile of response time is not necessarily at the 75th percentile for each of the three time components. For example, it's possible for an observation to have a very long turnout time and below average travel time. The total response time should be modeled rather than each of its components separately.
3. Turnout time and travel time are numeric variables and using regression trees rather than classification trees is a better fit. By using a classification tree, the tree will make splits based on observations above and below the breakpoint, which may be less useful than a regression tree that would more easily identify the variables that have the biggest impact on time and help quantify the size of that impact.

The model plot for the classification tree created by your assistant to predict travel time is shown below.



(b) (*2 points*) Describe the different circumstances that could lead to a single-node tree.

For full credit, candidates needed to discuss model assumptions (such as high minbucket, high cp, etc.) and the possibility that the variables themselves lack predictive power. Partial credit was awarded to candidates who only discussed how model parameters could lead to a single-node tree.

ANSWER:

A single node tree is a result of the model being unable to locate any variable split that reduces the impurity of the parent node by the minimum default amount. Because of the inability to split the dataset on any variable, the average of the entire data set is applied as the estimate for all predictions. The single node can be caused by poor choice of model parameters, such as a high complexity or minbucket parameter. A single node tree can also occur because none of the independent variables have explanatory value for the target variable.

You recommend to your assistant that they adjust the complexity parameter for the classification tree predicting travel time.

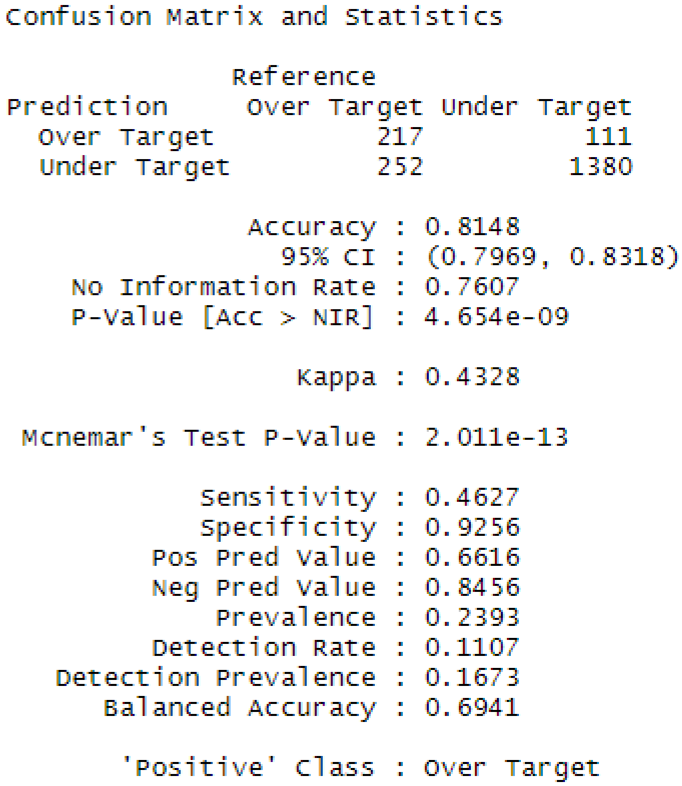
(c) (*2 points*) Explain how adjusting the complexity parameter affects the decision tree output.

For full credit, candidates needed to explain how splits are determined in terms of impurity and the complexity parameter. Many candidates only explained that increasing the complexity parameter decreases the number of splits or that setting cp = 1 creates a tree with only the root node.

ANSWER:

The complexity parameter controls the minimum level of impurity reduction for a split to be made. Reducing the complexity parameter will allow the tree to grow more leaves as it reduces the threshold for impurity reduction for a split to be made.

Your assistant provides the following confusion matrix for the classification tree predicting turnout time.



```
Confusion Matrix and Statistics

              Reference
Prediction     Over Target Under Target
  Over Target          217          111
  Under Target         252         1380

               Accuracy : 0.8148
                 95% CI : (0.7969, 0.8318)
    No Information Rate : 0.7607
    P-Value [Acc > NIR] : 4.654e-09

                  Kappa : 0.4328

 Mcnemar's Test P-Value : 2.011e-13

            Sensitivity : 0.4627
            Specificity : 0.9256
         Pos Pred Value : 0.6616
         Neg Pred Value : 0.8456
             Prevalence : 0.2393
         Detection Rate : 0.1107
   Detection Prevalence : 0.1673
      Balanced Accuracy : 0.6941

       'Positive' Class : Over Target
```

(d) (*2 points*) Interpret for your assistant the most applicable confusion matrix results for the city of Tempe.

Overall, candidates struggled to connect the confusion matrix results to the business problem. Full credit was awarded for any choice of confusion metrics with justification from the business problem. Partial credit was awarded where the metrics were correctly interpreted but with flawed rationale connecting it to the business problem.

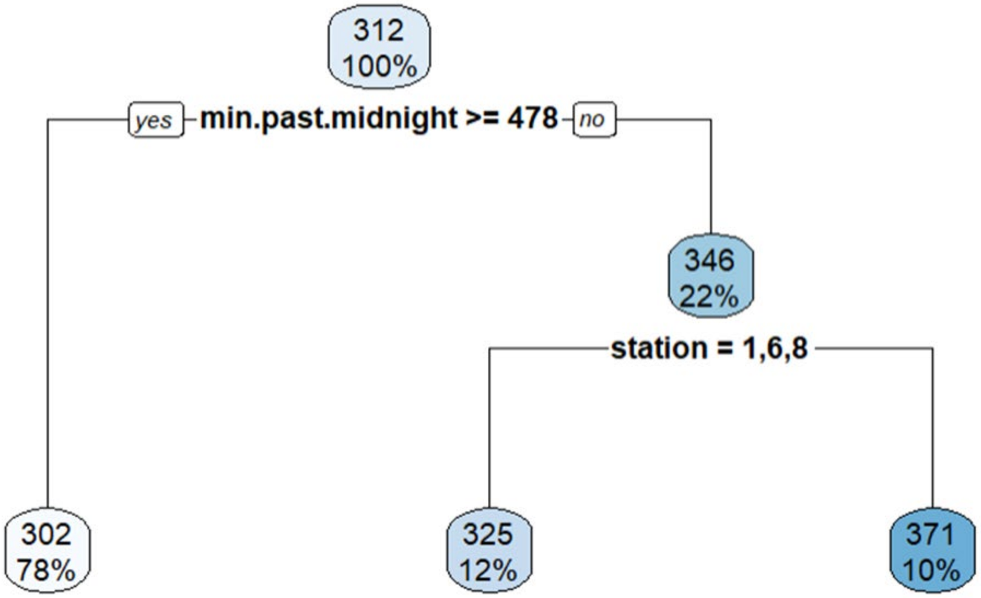
ANSWER:

The city is most interested in understanding high response times, the positive case in the output provided. Therefore, the most applicable confusion metrics will express how well the model predicts positive cases. The most helpful metric is sensitivity, which is the proportion of positive cases that the model correctly predicts.

The sensitivity of 0.4627 means that slightly less than half of the high turnout time cases are being correctly assigned. The model does not provide much insight into predicting whether a case will have high turnout times.

Task 11 (*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.