Exam PA April 14 Project Statement

IMPORTANT NOTICE – THIS IS THE APRIL 14 PROJECT STATEMENT. IF TODAY IS NOT APRIL 14, SEE YOUR TEST CENTER ADMINISTRATOR IMMEDIATELY.

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

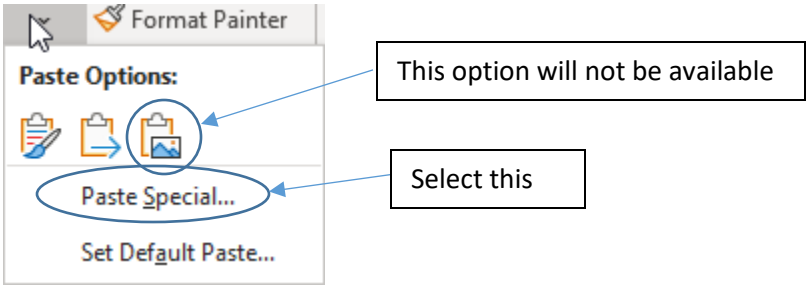
Variable Name	Variable Values
issue	Type of emergency event (11 categories)
vehicle	L, E indicate the two most common vehicles. X is all others.
station	1 to 8
hour	0 to 23, hours past midnight
min.past.midnight	0 to 1439
month	1 to 12
day	1 to 31
weekday	1 to 7 for Sunday to Saturday
dewpoint	a weather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

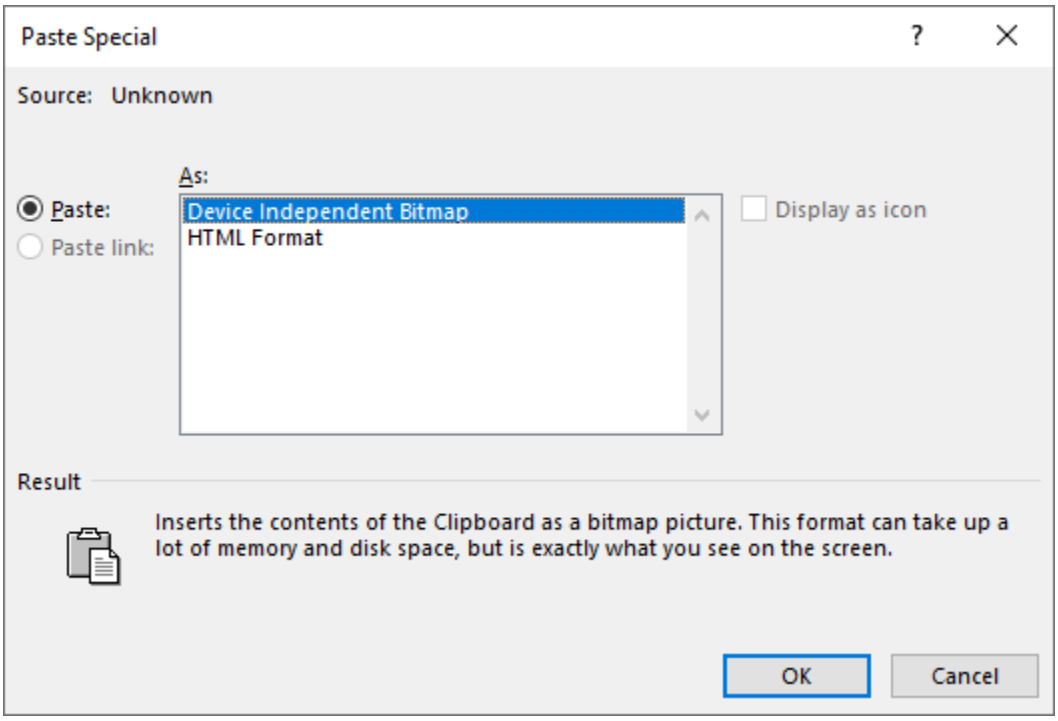
The type of medical event may not be known precisely at the time of the call, but information related to the issue variable is conveyed by the dispatcher to the workers in the vehicle.

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.



From this dialog box, make the indicated selection.



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 <chr>	alarm.processing.time.count <int>	turnout.time.count <int>	travel.time.count <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.

FIRST ANSWER (alarm processing time):

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

Discuss implications for the business problem:

SECOND ANSWER (turnout time):

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

Discuss implications for the business problem:

THIRD ANSWER (travel time):

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

Discuss implications for the business problem:

(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

ANSWER:

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

Removing the observations that contain the missing values:

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.

ANSWER:

Descriptive Modeling Objective:

Predictive Modeling Objective:

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.

ANSWER:

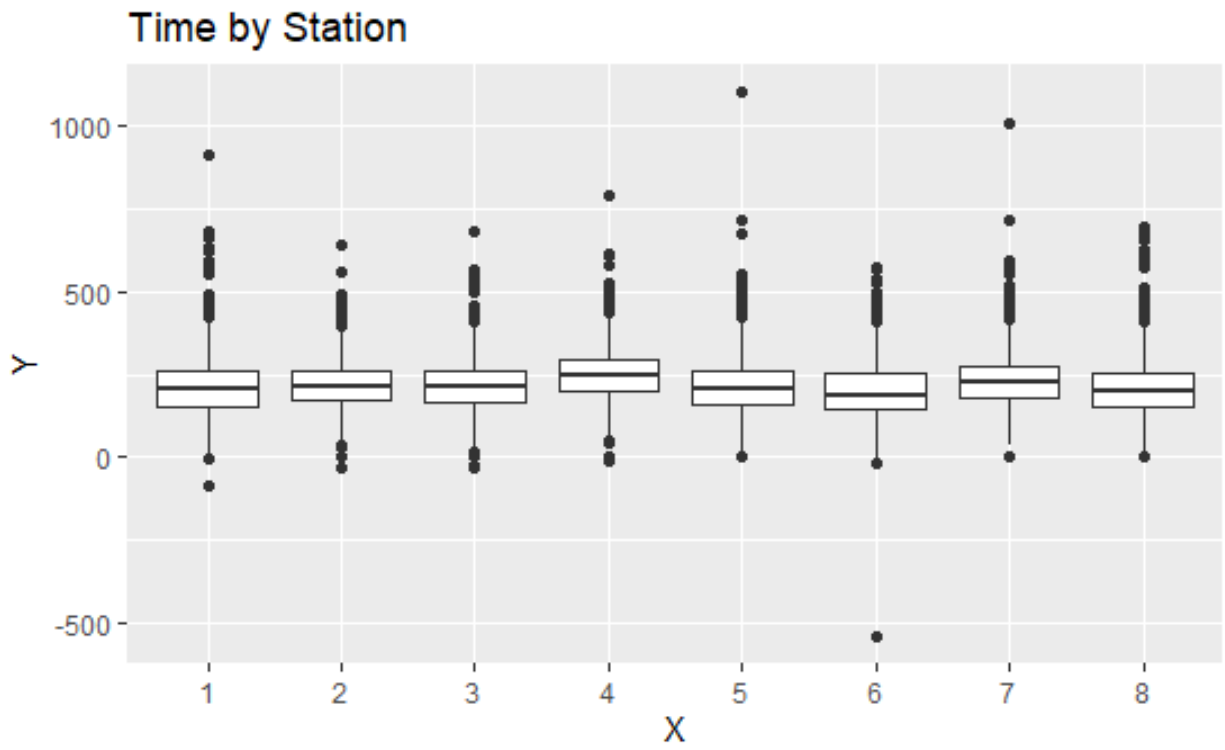
Question 1:

Question 2:

Question 3:

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**.

ANSWER:

(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.

ANSWER:

Code:

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.

ANSWER:

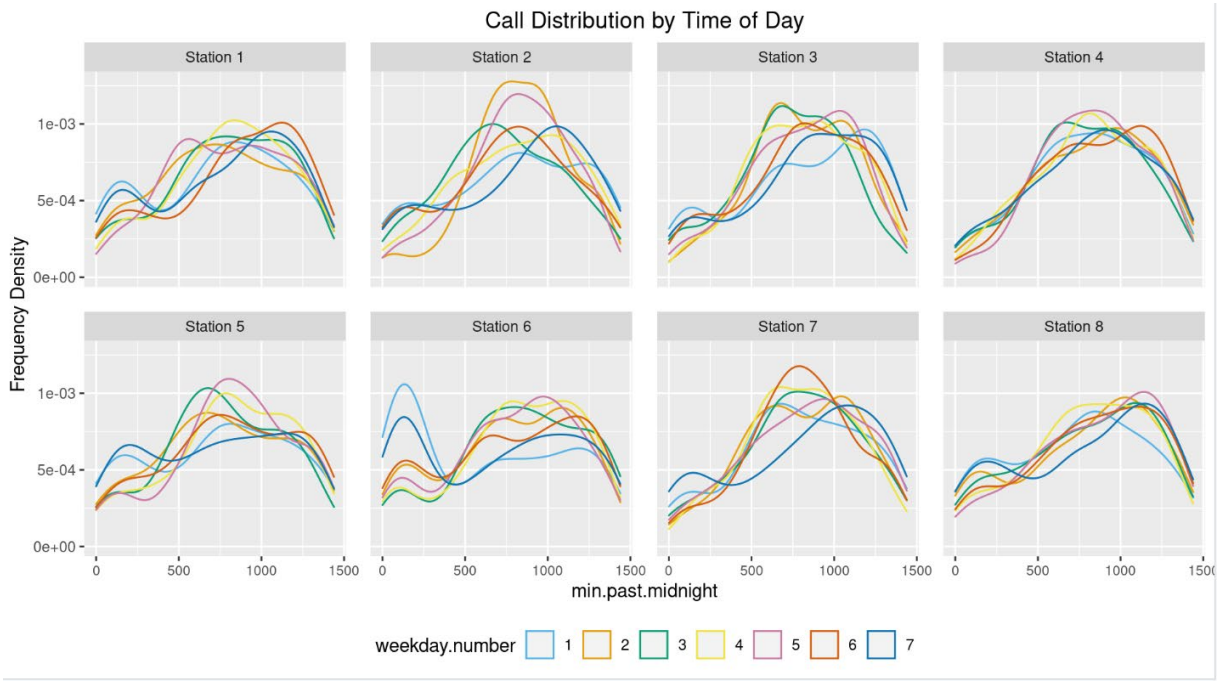
Task 4 (*12 points*)

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.

ANSWER:

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.

ANSWER:

Minutes past midnight:

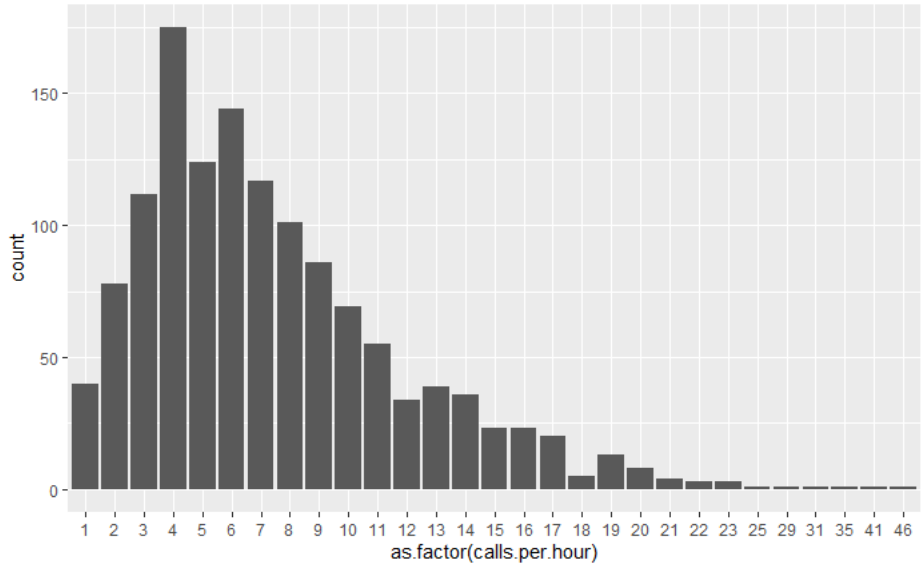
Station:

Weekday number:

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

ANSWER:

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.

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.

ANSWER:

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.

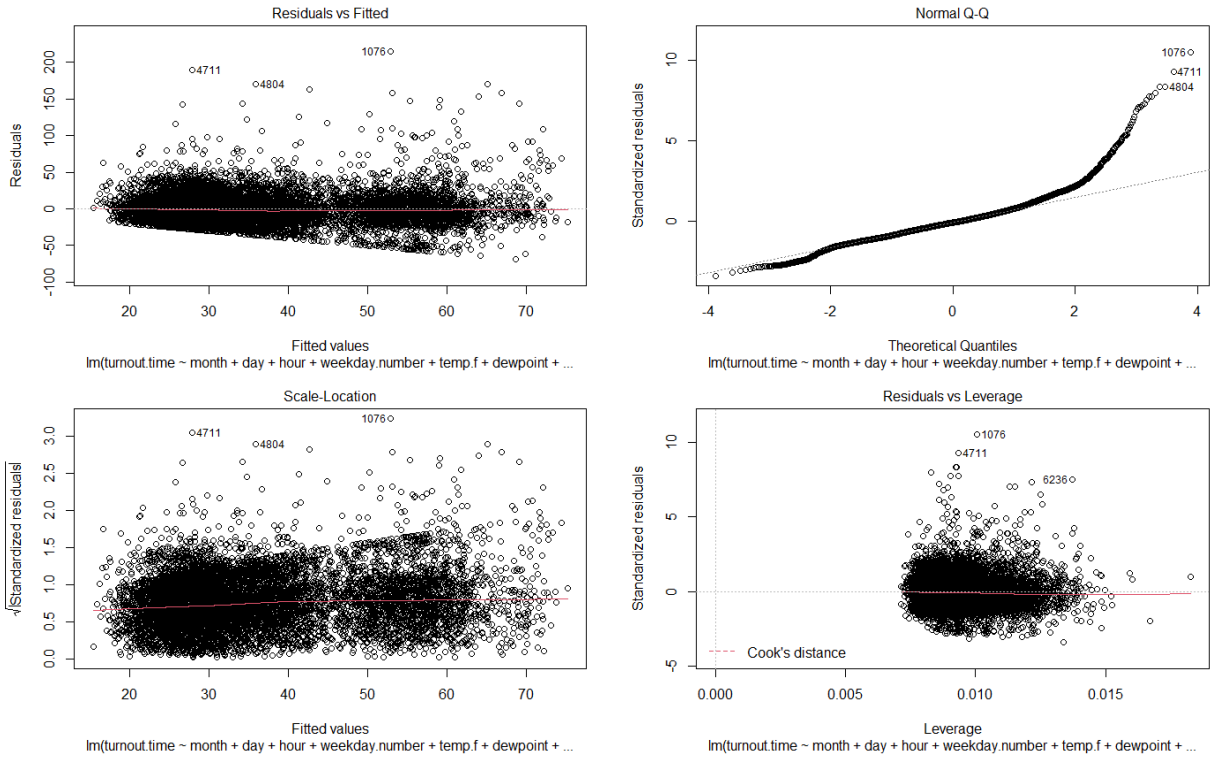
ANSWER:

Gaussian distribution:

Poisson:

Gamma distribution:

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**.

ANSWER:

First Reason:

Second Reason:

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

ANSWER:

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.

ANSWER:

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.

ANSWER:

(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.

ANSWER:

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.

ANSWER:

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.

ANSWER:

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.

ANSWER:

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.

ANSWER:

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.

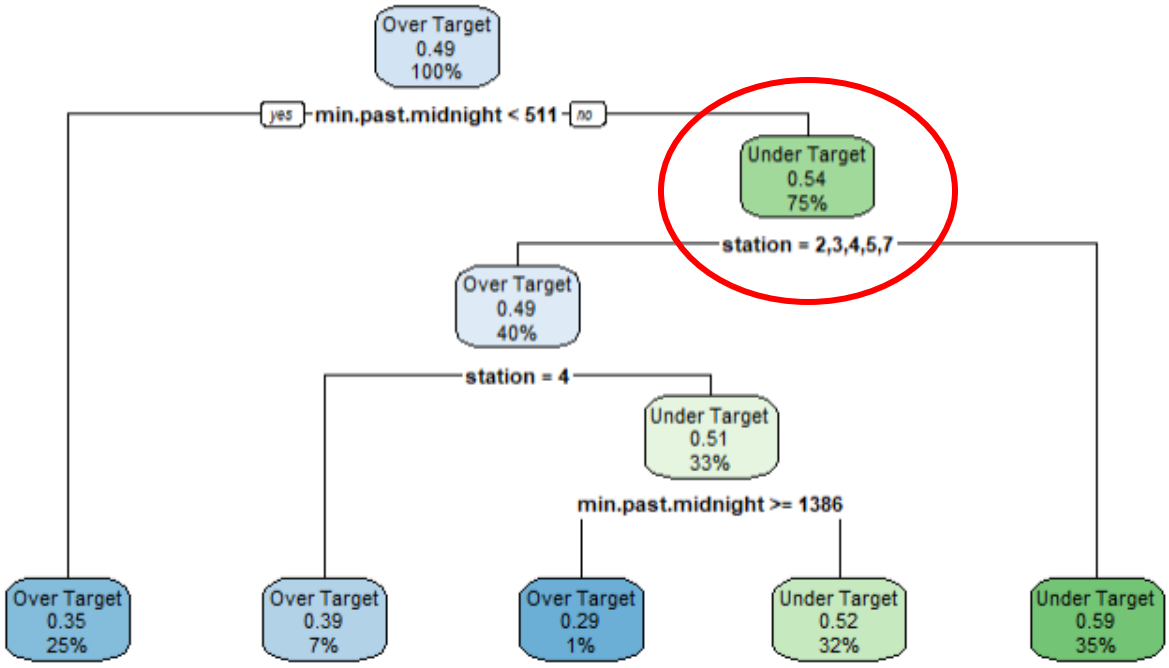
ANSWER:

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.

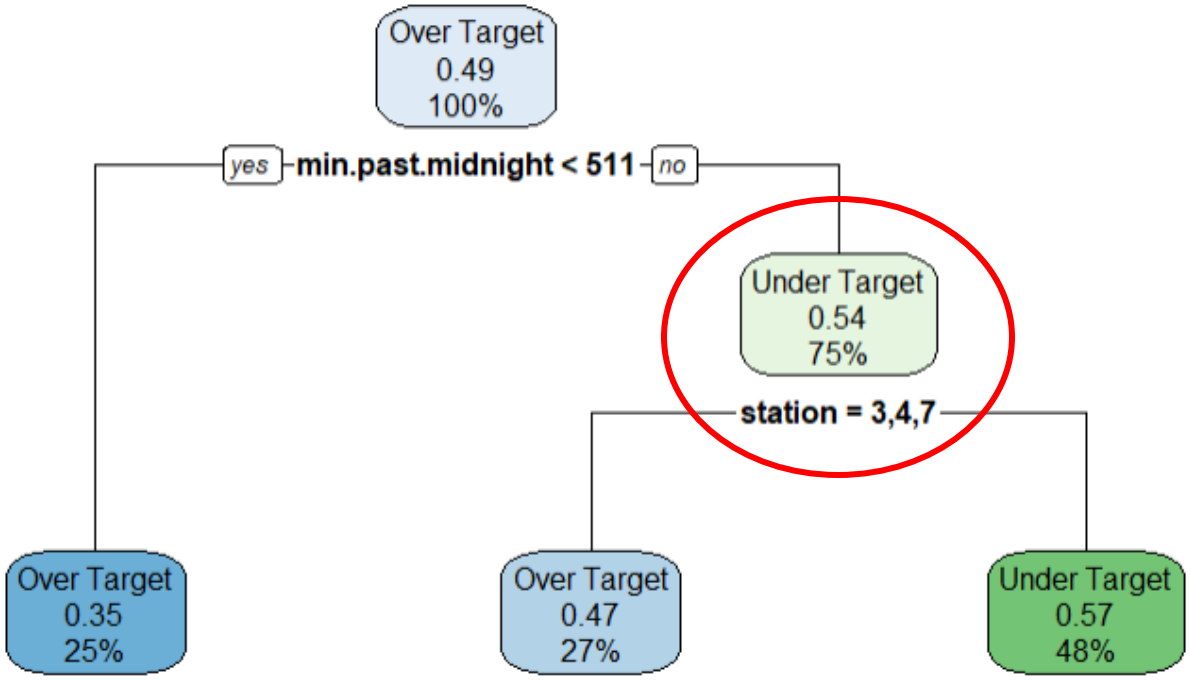
Both trees have the same first split based on min.past.midnight < 511, but the right sub-node based on specific stations (highlighted in both trees) split differently for the tree built using the Gini impurity measure compared to the tree built using the Entropy impurity measure.

(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



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		0.004712	0.499835	0.484465	0.004708
Entropy	0.996125	0.997812		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:

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.

ANSWER:

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.

ANSWER:

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

ANSWER:

Linear Models:

Tree-Based Models:

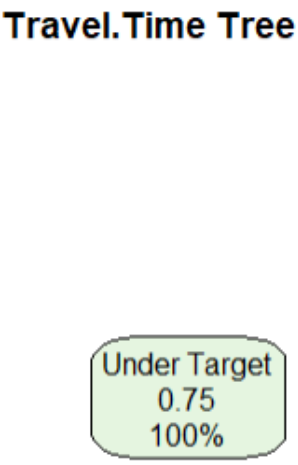
Task 10 (*10 points*)

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.

ANSWER:

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.

ANSWER:

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.

ANSWER:

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

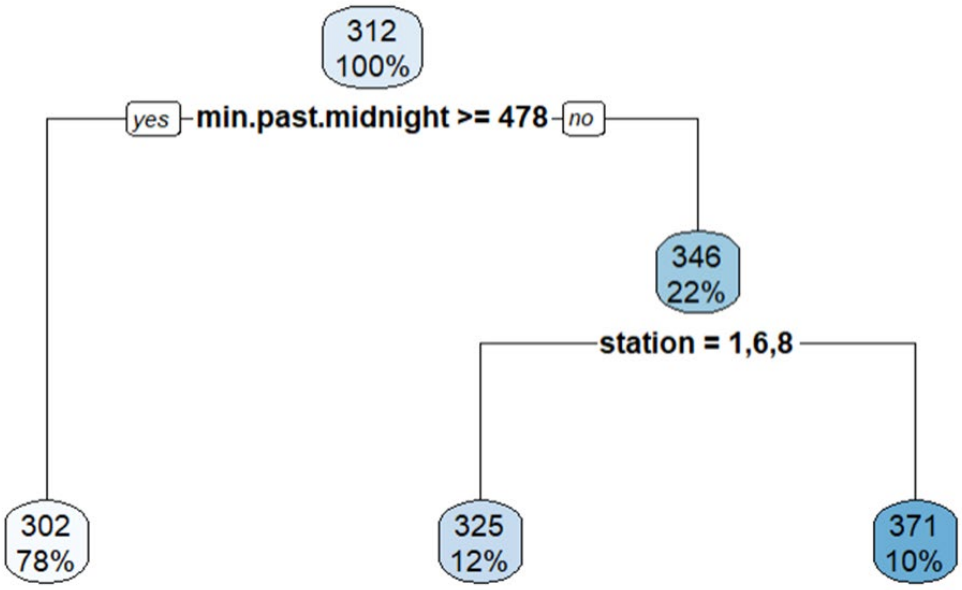
```
Confusion Matrix and Statistics

               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.

ANSWER:

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.

ANSWER:

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.

ANSWER:

(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.

ANSWER:

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.

ANSWER: