Exam PA April 12 Project Statement

IMPORTANT NOTICE – THIS IS THE APRIL 12 PROJECT STATEMENT. IF TODAY IS NOT APRIL 12, SEE YOUR TEST CENTER ADMINISTRATOR IMMEDIATELY.

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.
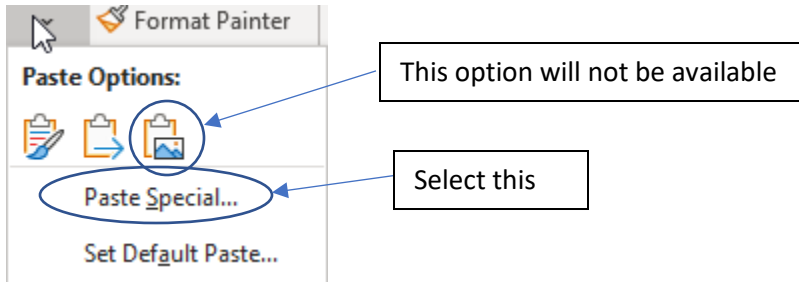
From this dialog box, make the indicated selection.
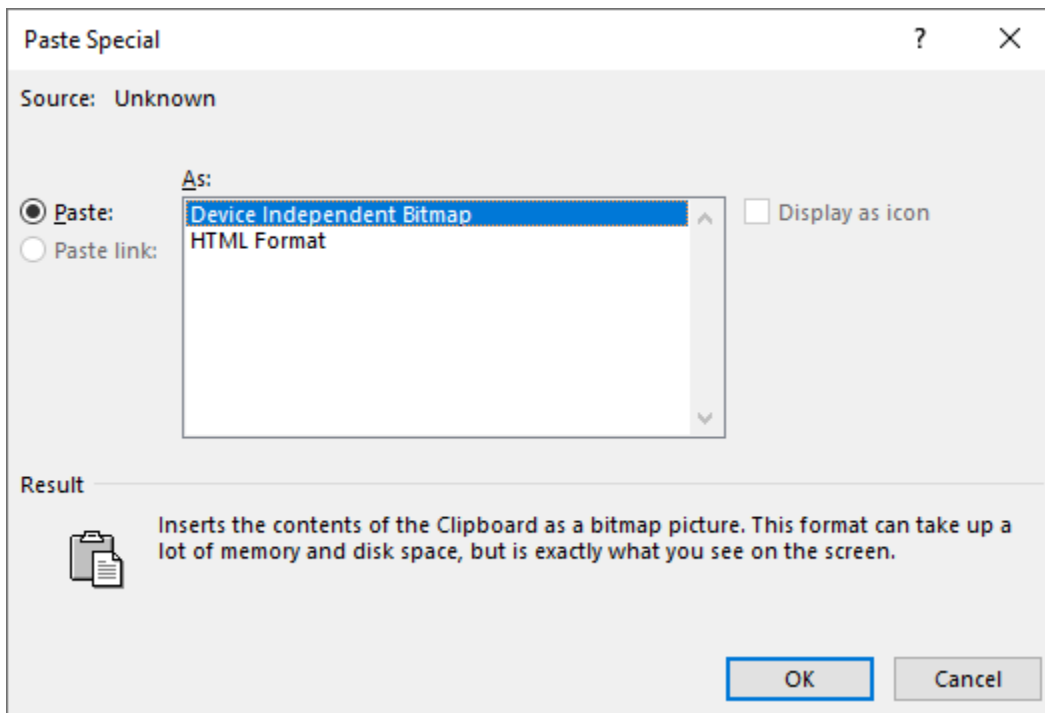
Task 1 (*8 points*)

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.

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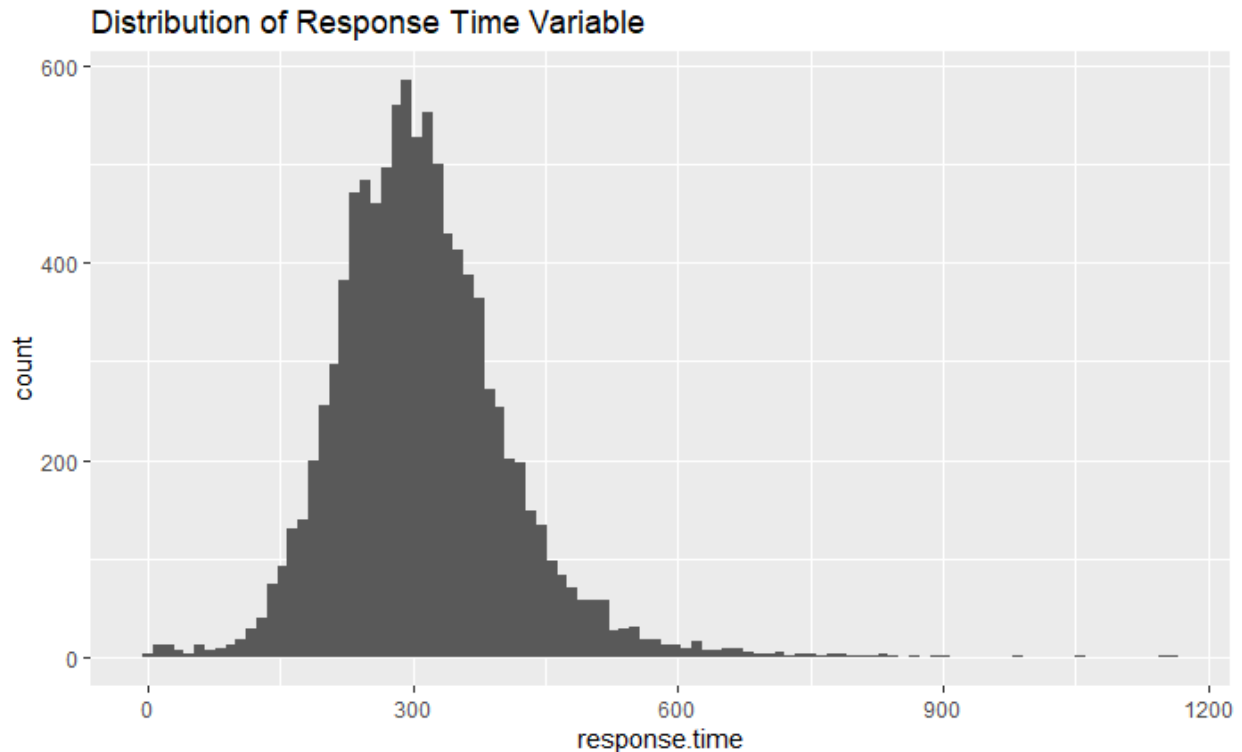
(a)   (*3 points*) Recommend whether a transformation should be applied to the **response.time** variable in the GLM given the business problem. Justify your reasoning.

ANSWER:

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**

ANSWER:

The plausibility of the outliers:

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

Fitting a GLM that predicts response.time:

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 (*7 points*)

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

ANSWER:

(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.

ANSWER:

Improvement 1:

Improvement 2:

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.
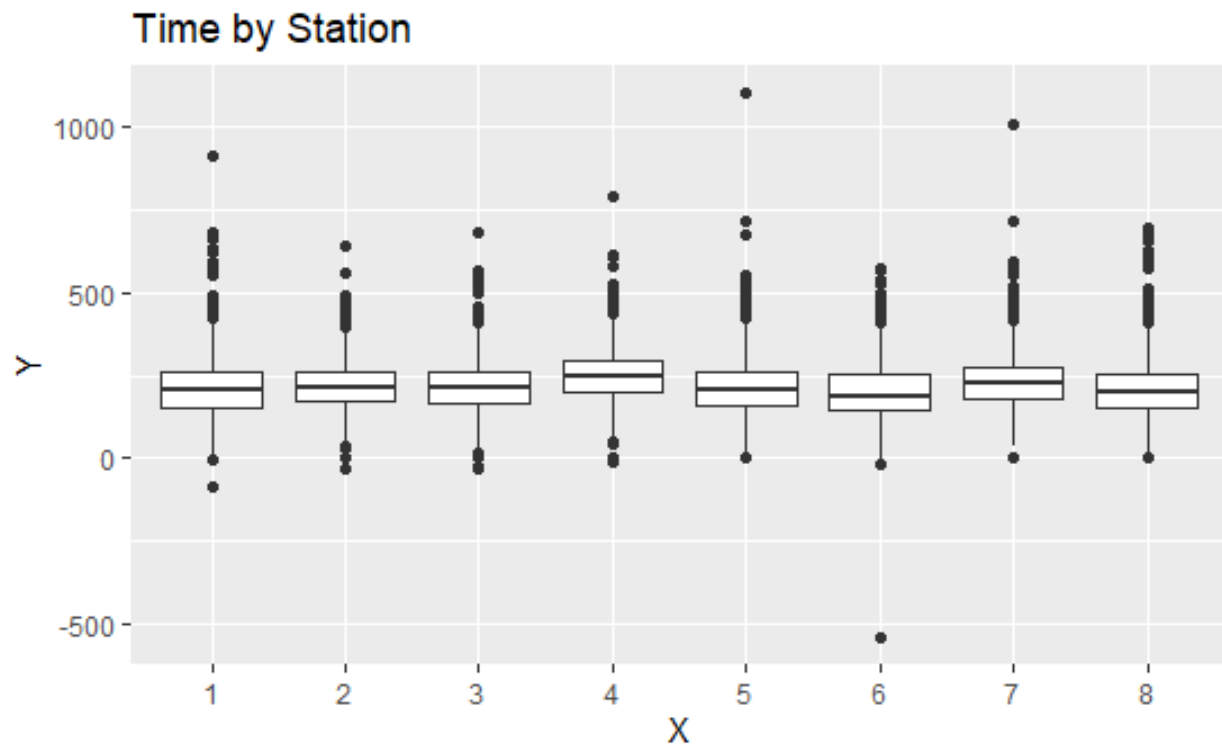
Time by Station

(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 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.

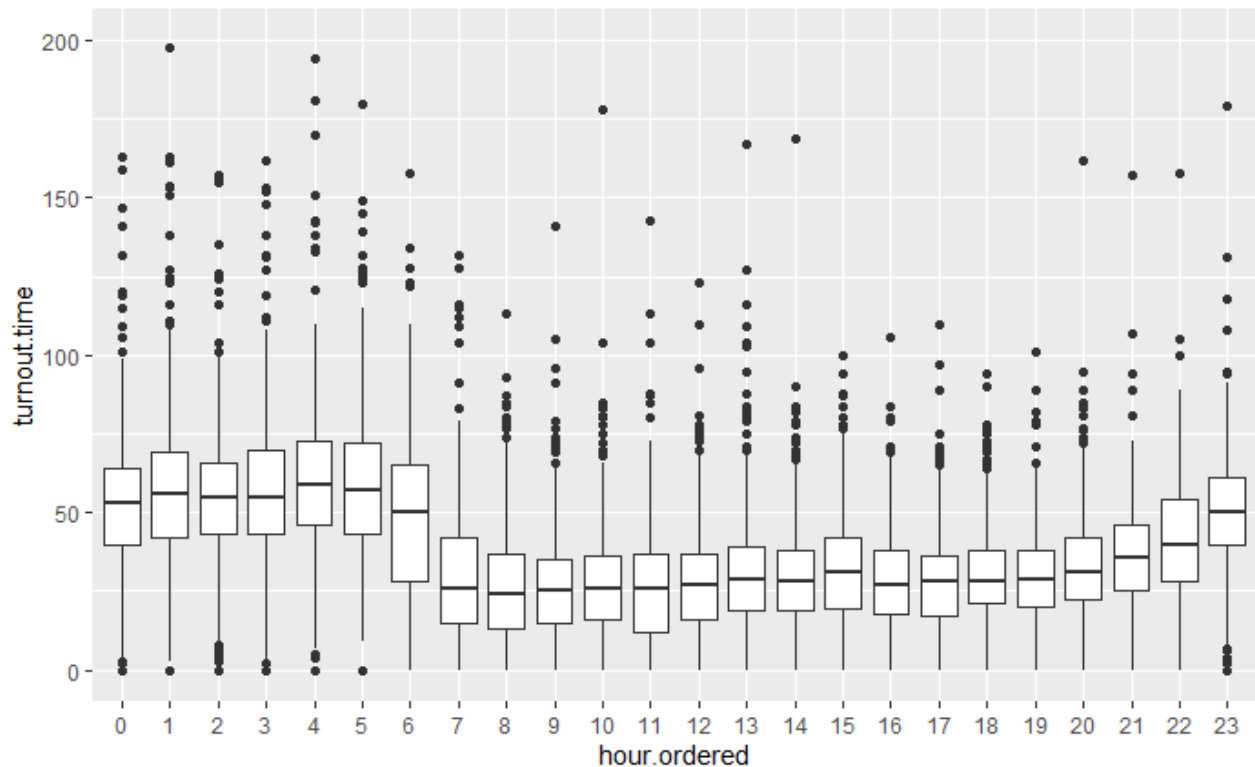**ANSWER:**

Reason 1:

Reason 2:

Reason 3:

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.

ANSWER:

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.

ANSWER:

Gaussian distribution:

Poisson:

Gamma distribution:

Your assistant runs an ordinary least squares (OLS) model to model **turnout.time**. Review the diagnostic plots below.
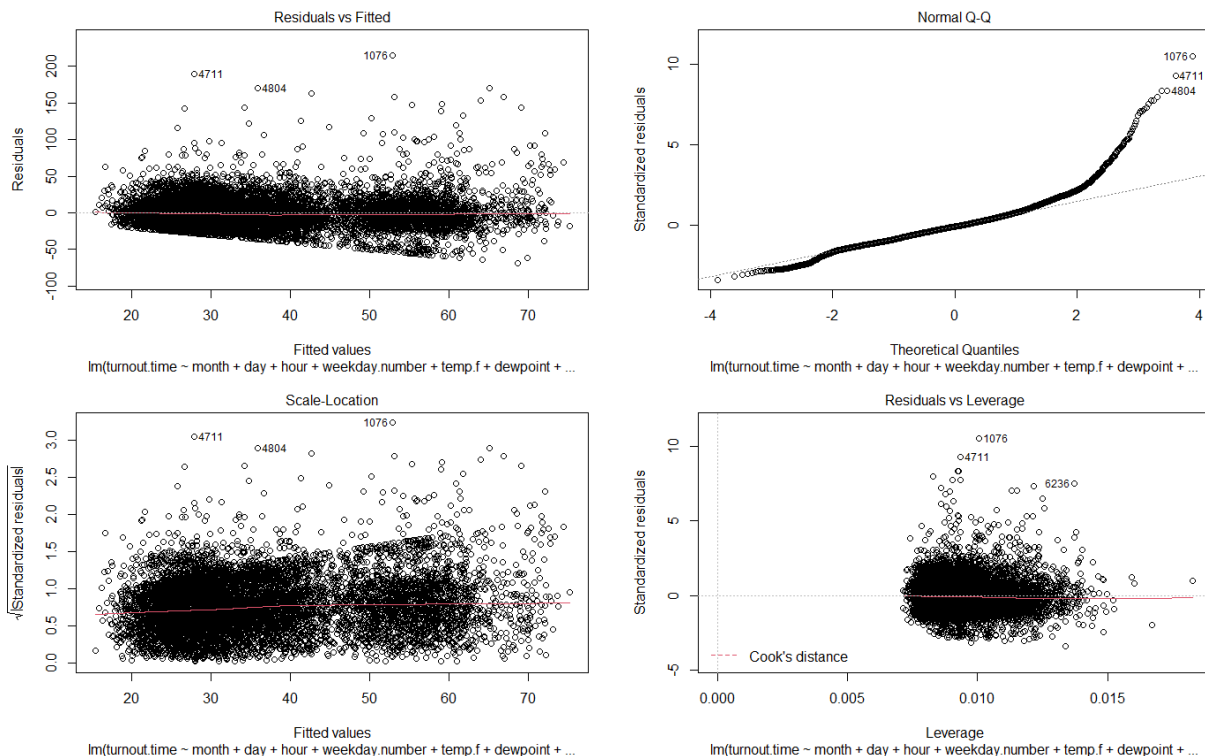
(b)   (*2 points*) Explain two reasons why OLS is not a good choice to model **turnout.time**.

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 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.

ANSWER:

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

ANSWER:

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.

ANSWER:

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

ANSWER:

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

ANSWER:

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.

ANSWER:

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.

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


```
                        Over Target
                           0.49
                           100%
              yes ─ min.past.midnight < 511 ─ no
                                                Under Target
                                                   0.54
                                                   75%
                                    station = 2,3,4,5,7
                           Over Target
                              0.49
                              40%
                        station = 4
                                    Under Target
                                       0.51
                                       33%
                              min.past.midnight >= 1386

Over Target      Over Target      Over Target      Under Target      Under Target
   0.35             0.39             0.29             0.52              0.59
   25%              7%               1%               32%               35%
```

Gini decision tree


```
                    Over Target
                       0.49
                       100%
          yes ─ min.past.midnight < 511 ─ no
                                          Under Target
                                             0.54
                                             75%
                                    station = 3,4,7

Over Target          Over Target          Under Target
   0.35                 0.47                  0.57
   25%                  27%                   48%
```
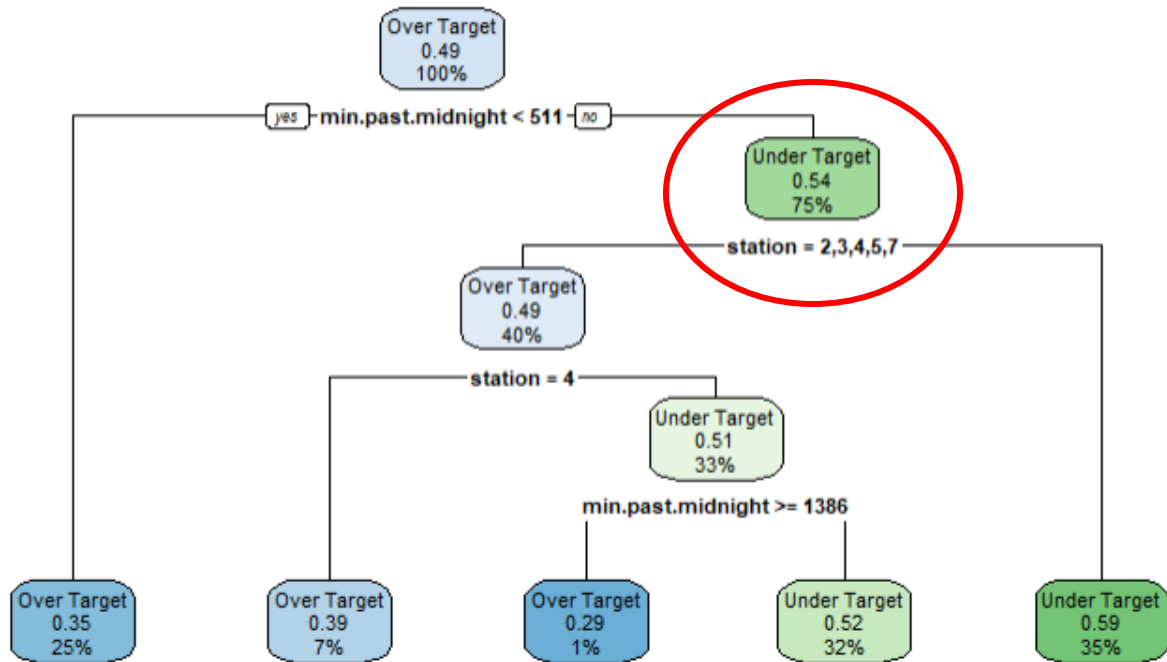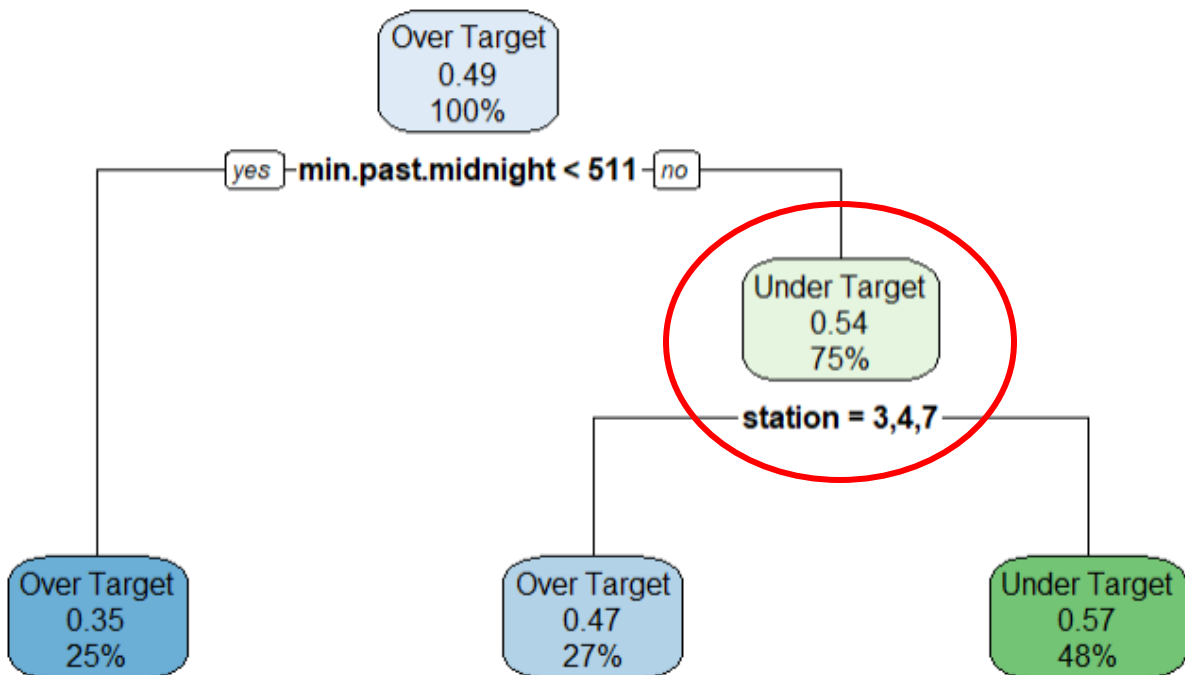
ANSWER:

Chart with two highlighted cells to complete:

	Primary Node	Gini Tree Node Split			Entropy Tree Node Split		
		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 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.

ANSWER:

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

ANSWER:

Linear Models:

Tree-Based Models:

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.

ANSWER:

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.

ANSWER:

Df.Train.1:

Df.Train.2:

Df.Train.3:

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.
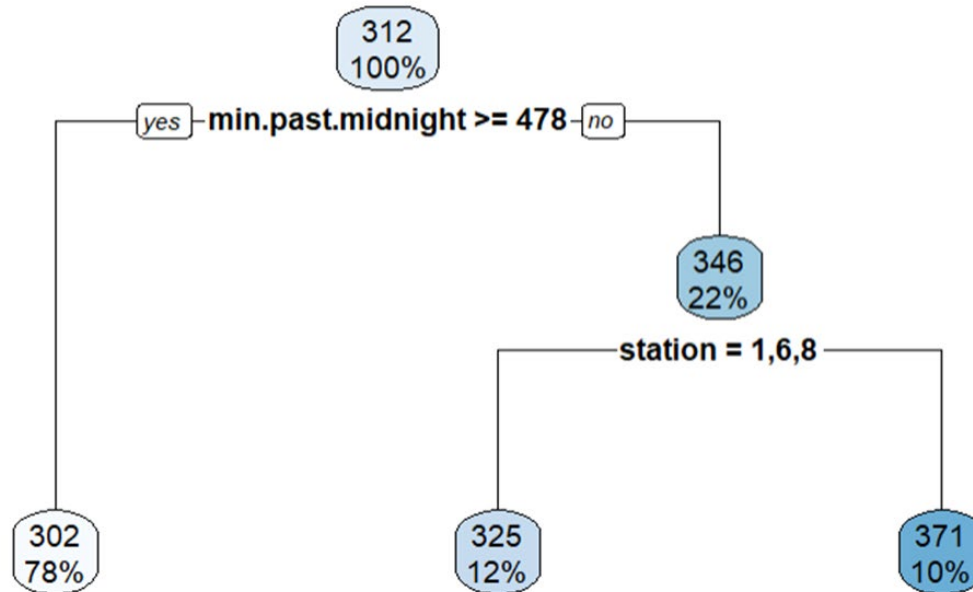
ANSWER:

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

ANSWER:

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.

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:

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.

ANSWER:

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.

ANSWER:

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.

ANSWER: