Exam PA April 2025 Project Statement

IMPORTANT NOTICE – THIS IS THE APRIL 15, 2025, PROJECT STATEMENT. IF TODAY IS NOT APRIL 15, 2025, SEE YOUR TEST CENTER ADMINISTRATOR IMMEDIATELY.

General Information for Candidates

This examination has 11 tasks numbered 1 through 11 with a total of 70 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 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. You may use Excel for calculation for any of the tasks, but all answers must be submitted in the Word document. *If you upload the Excel document, it will not be looked at by the graders.* Neither R nor RStudio are available.

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.

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.

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.

Business Problem

You work for a consulting firm that is supporting the London Fire Brigade, which provides fire response services to the city of London in the UK. Your client is interested in understanding different aspects of their work including the frequency of different types of fire incidents, response times, and the cost of responding to calls.

Your firm is using a detailed dataset provided by the London Fire Brigade that includes all incidents between August 1, 2021 and September 30, 2024.[1] *The data set includes, along with other variables, incident-level data on the type of incident, the time for the first fire truck to arrive at the scene of the incident call, information on the time and location of the incident, and the resources used for each call.*

Note:

Some of the terms used in the variable names reference British terms including:

- Pump for fire truck.
- Northing and easting, which are comparable to latitude and longitude, respectively.
- Notional cost is measured in British pounds.

[1] *Source: London Data Store*

Data Dictionary

Variable	Data Type / Range / Example	Description
IncidentNumber	Character Example: 000008-01012018	Unique identifier for each incident.
DateOfCall	Date: 8/1/2021 – 9/30/2024	Date of the incident. MM/DD/YYYY Format.
CalYear	Numeric: 2021 – 2024	Year of the incident.
HourOfCall	Numeric: 0 – 23	Hour of the call.
IncidentGroup	Character: False Alarm, Special Service, Fire	High level incident group. This is determined after the fire response is completed.
StopCodeDescription	Character: AFA (Automatic Fire Alarm), Special Service, Primary Fire, Secondary Fire, etc.	More granular incident group classification.
PropertyCategory	Character: Dwelling, Non-Residential, etc.	High level property description.
PropertyType	Character: House - single occupancy, Secondary school, etc.	Detailed property description.
IncGeo_BoroughName	Character: Varies	Name of the borough that the fire was reported in. A borough is an administrative district in the larger city of London.
Easting_rounded	Numeric: 503,550 - 561,150	The distance eastwards of a given point. Similar to a line of longitude.
Northing_rounded	Numeric: 155,950 - 200,850	The distance northwards of a given point. Similar to a line of latitude.
IncidentStationGround	Character: Varies	The station that is responsible for the region where the fire occurred.
NumStationsWithPumpAttending	Numeric: 1 – 9	The number of different stations that sent pumps to the fire. *Pumps are the term used for fire engines.*
PumpCount	Numeric: 1 – 189	The number of fire pumps dispatched to a fire. *Pumps are the term used for fire engines.*
PumpMinutesRounded	Numeric: 60 – 60,502	The total number of minutes pumps are

		present at a fire. Two pumps present for 60 minutes each are recorded as 120. All values under 60 minutes are rounded-up to 60.
NotionalCost	Numeric: 352 – 433,598	The estimated cost to the fire department of responding to a given incident. This is calculated as the PumpMinutesRounded multiplied by a per minute cost factor. The per minute cost factor is updated in April of each year.
NumCalls	Numeric: 1 – 124	The number of distinct calls the fire department received about a given incident.
FirstPumpArriving_AttendanceTime	Numeric: 1 – 1,200	The time in seconds between when the fire is reported and when the first pump arrives at the location of the fire.

Task 1 (*6 points*)

The data dictionary outlines information from fire incident reports within London, England over four years. There are multiple ways of analyzing this data and determining what business problem should be asked (and subsequently answered).

(a) (*2 points*) Define what it means to analyze the data from the following perspectives:

i. Descriptive Analytics
ii. Predictive Analytics
iii. Prescriptive Analytics

ANSWER:

(b) (*3 points*) Develop one potential business question for each approach referencing the detail provided in the data dictionary.

i. Descriptive Analytics
ii. Predictive Analytics
iii. Prescriptive Analytics

ANSWER:

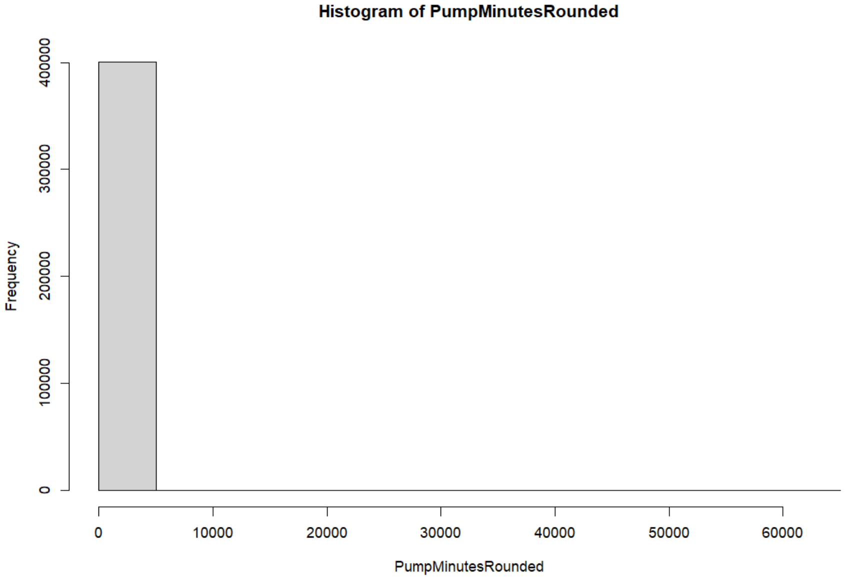
You have been hired as an analytics consultant for the fire department. The fire department is interested in understanding the past drivers of high costs.

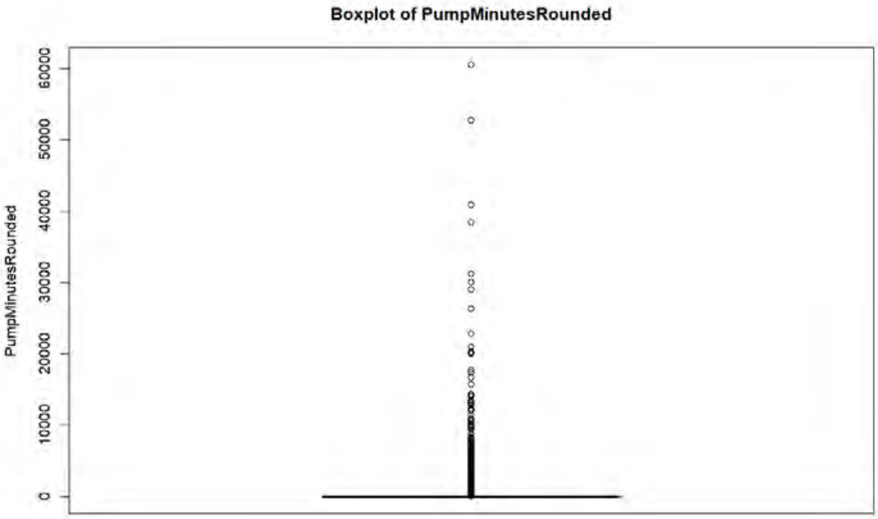
(c) (*1 points*) State whether descriptive, predictive, or prescriptive analytics applies to this business problem. Explain your decision.

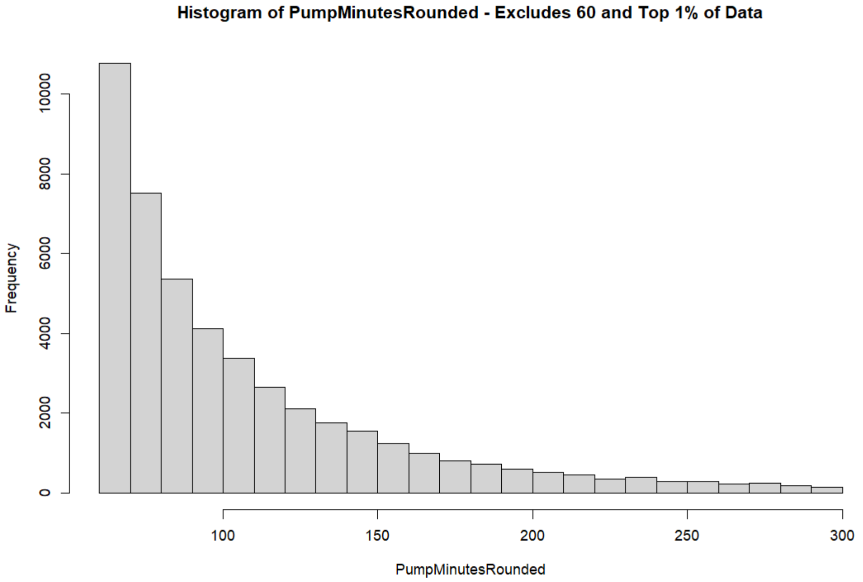
ANSWER:

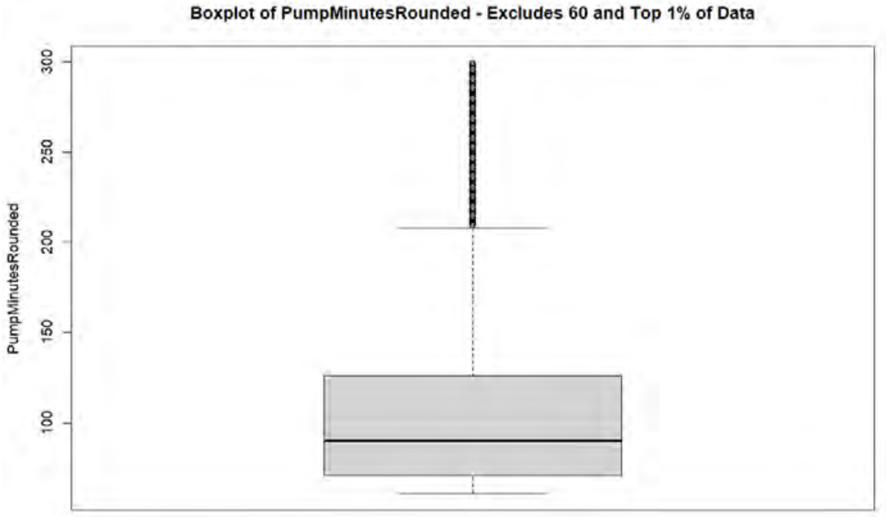
Task 2 (4 *points*)

The graphs below show two histograms and two boxplots for the number of pump minutes used per fire call (PumpMinutesRounded). The first histogram and the boxplot show all of the data. Seeing that the data is difficult to interpret, your assistant excludes the value 60, which is the minimum value for PumpMinutesRounded, and also excludes the top 1% of the data values (i.e., they only show the lower 99% of the data before removing 60) and creates the second histogram and boxplot below.









(a) (2 *points)* Describe the distribution of PumpMinutesRounded based on the graphs above and how it may impact the use of PumpMinutesRounded as an explanatory variable.

ANSWER:

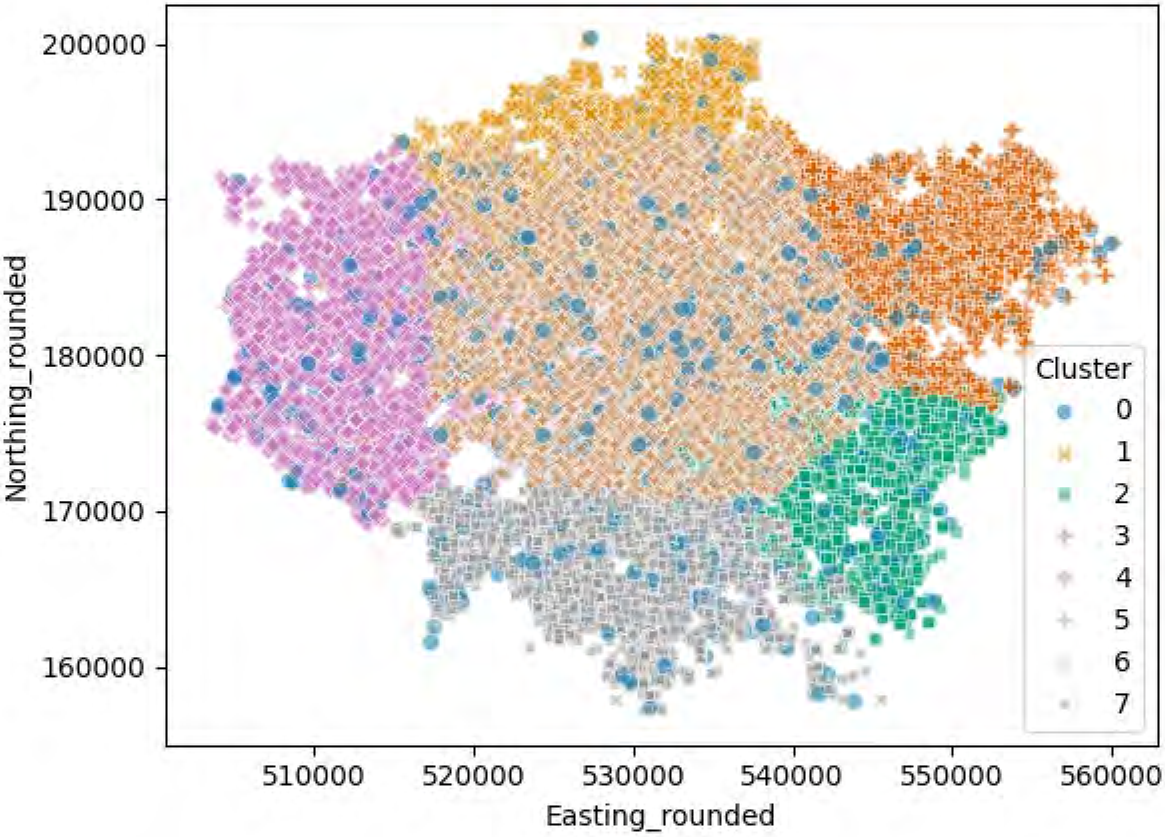
Your assistant still wants to explore the use of PumpMinutesRounded in their predictive model.

(b) (*2 points*) Recommend a potential transformation of the variable. Justify your answer.

ANSWER:

Task 3 (*6 points*)

Your assistant decides to cluster the cost of incidents using the variables: **Easting_rounded, Northing_rounded,** and **NotionalCost.** They created eight clusters**.** They provided a graph and summary statistics.



	NotionalCost	Northing_rounded	Easting_rounded
count	400,736	400,736	400,736
mean	477	180,388	530,587
std	1,651	7,436	9,748
min	352	155,950	503,550
25%	364	175,950	525,150
50%	388	180,950	530,950
75%	430	185,150	536,250
max	433,598	200,850	561,150

(a) (*2 points*) Critique the clustering approach and recommend improvements.

ANSWER:

Your boss wants to use the clusters in part (a) along with other variables to predict **NotionalCost**.

(b) (*2 points*) Critique this approach and justify your reasoning.

ANSWER:

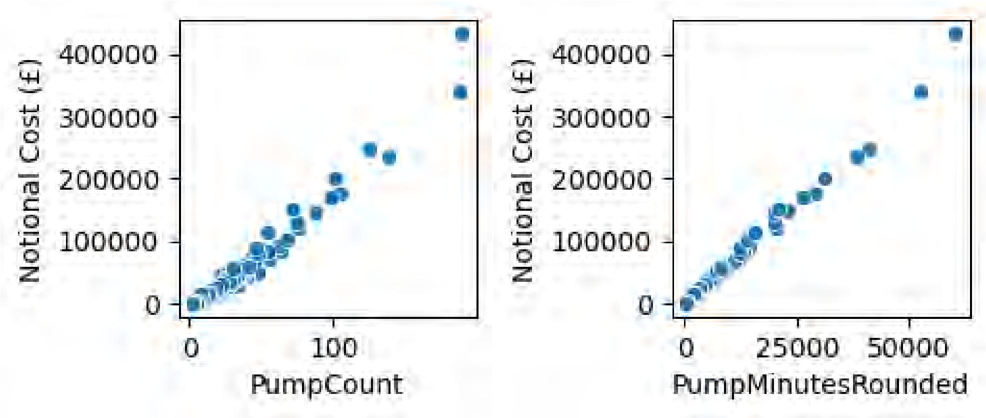
Your assistant uses k-means clustering to create clusters using the two variables **Northing_rounded** and **Easting_rounded.** Your boss proposes to use these clusters in GLM and tree-based models for the prediction of the target variable **NotionalCost** instead of using their untransformed values.

(c) (*2 points*) Compare and contrast using clustered variables versus using the variables without modification.

ANSWER:

Task 4 (5 *points*)

The fire department is interested in understanding the past drivers of high costs. You begin to perform data exploration on the numeric variables to assess whether they should be included in the model. You produce the following scatterplots:



(a) (*1 points*) Identify a potential issue with including both **PumpMinutesRounded** and **PumpCount** in a GLM.

ANSWER:

(b) (*2 points*) Recommend next steps that should be taken when evaluating whether to use both **PumpMinutesRounded** and **PumpCount** in a GLM.

ANSWER:

The fire department is interested in understanding predictors of costs at the time the call is received. You are considering the following variables as predictors of **NotionalCost**:

IncGeo_BoroughName	Character: Varies	Name of the borough that the fire was reported in. A borough is an administrative district in the larger city of London.
IncidentGroup	Character: False Alarm, Special Service, Fire	High level incident group. This is determined after the fire response is completed.
IncidentStationGround	Character: Varies	The station that is responsible for the region where the fire occurred.
PropertyCategory	Character: Dwelling, Non-Residential, etc.	High level property description.

(c) (*2 points*) Explain which of the above variables should not be used as a predictor variable in the model that predicts **NotionalCost**.

ANSWER:

Task 5 (10 *points*)

Your client, London Fire Brigade, requests assistance in analyzing London fire trends and operational efficiency. Specifically, they would like to understand how first pump attendance time (**FirstPumpArriving_AttendanceTime**) is impacted by incident types, response times, property categories, and associated costs.

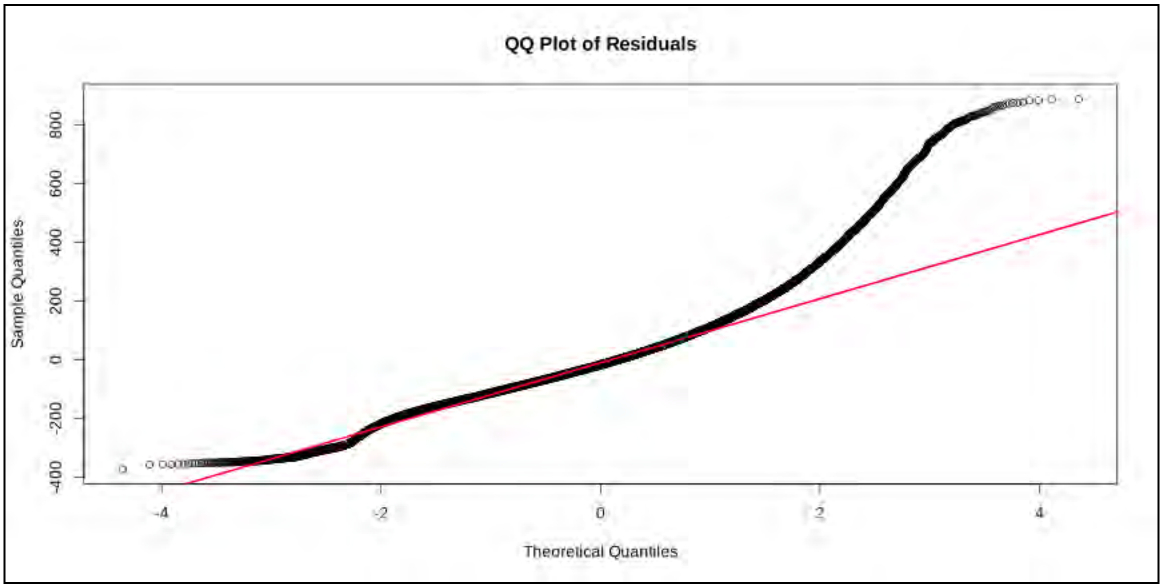
Your assistance provides you with a summary of variable **FirstPumpArriving_AttendanceTime**

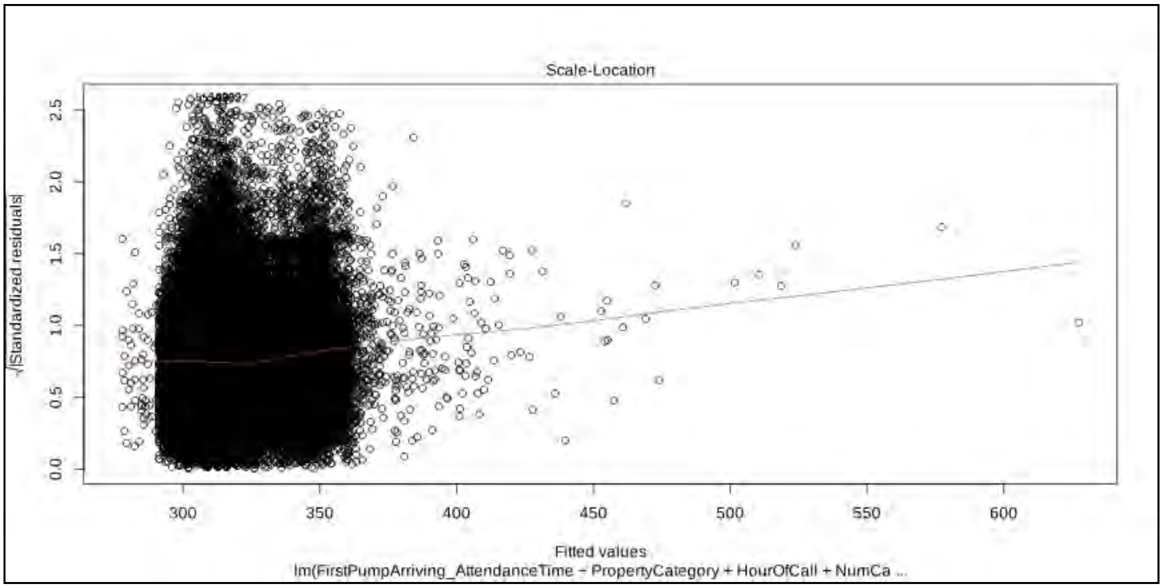
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  1.0   233.0   300.0   318.1   381.0  1200.0   21420
```

(a) (1 point) Describe the distribution of attendance time in terms of its skewness.

ANSWER:

Your assistant models the response variable FirstPumpArriving_AttendanceTime using several explanatory variables and provides you with the following model diagnostic plots:

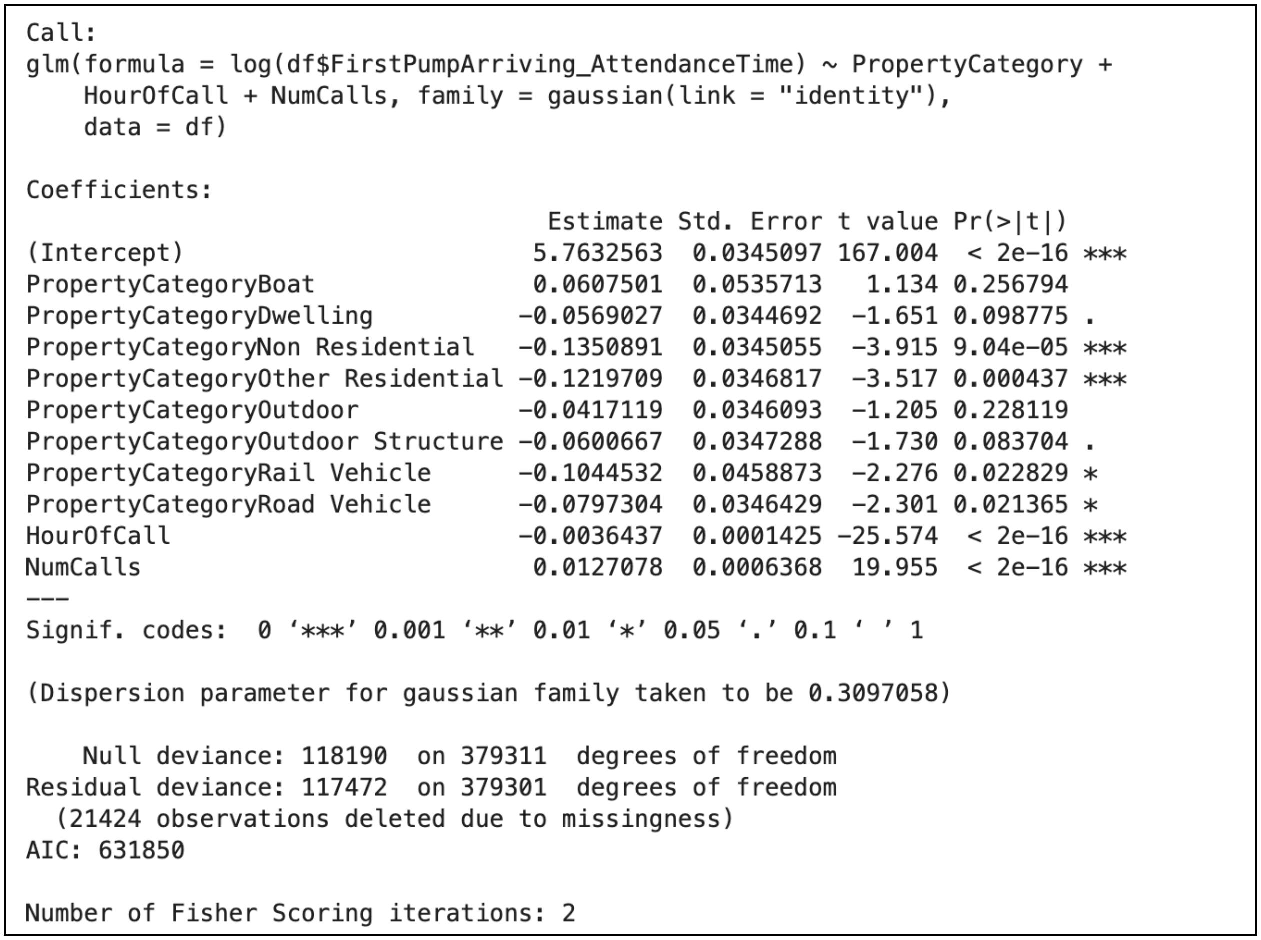




(b) (2 point) Recommend either a log transformation or the use of a log link function in modeling the response variable **FirstPumpArriving_AttendanceTime** using a GLM. Justify your choice.

ANSWER:

Your assistant builds a model to predict **FirstPumpArriving_AttendanceTime** with variables **PropertyCategory**, **HourOfCall** and **NumCalls**. You are provided with a GLM model summary.

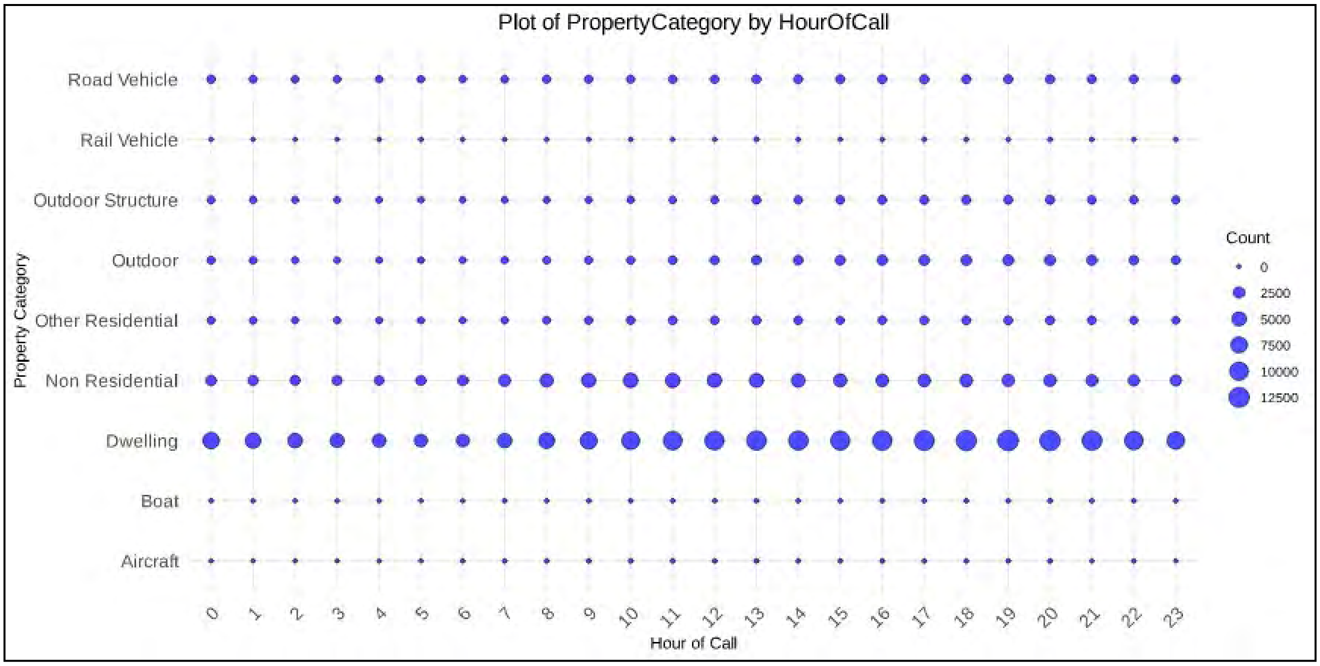


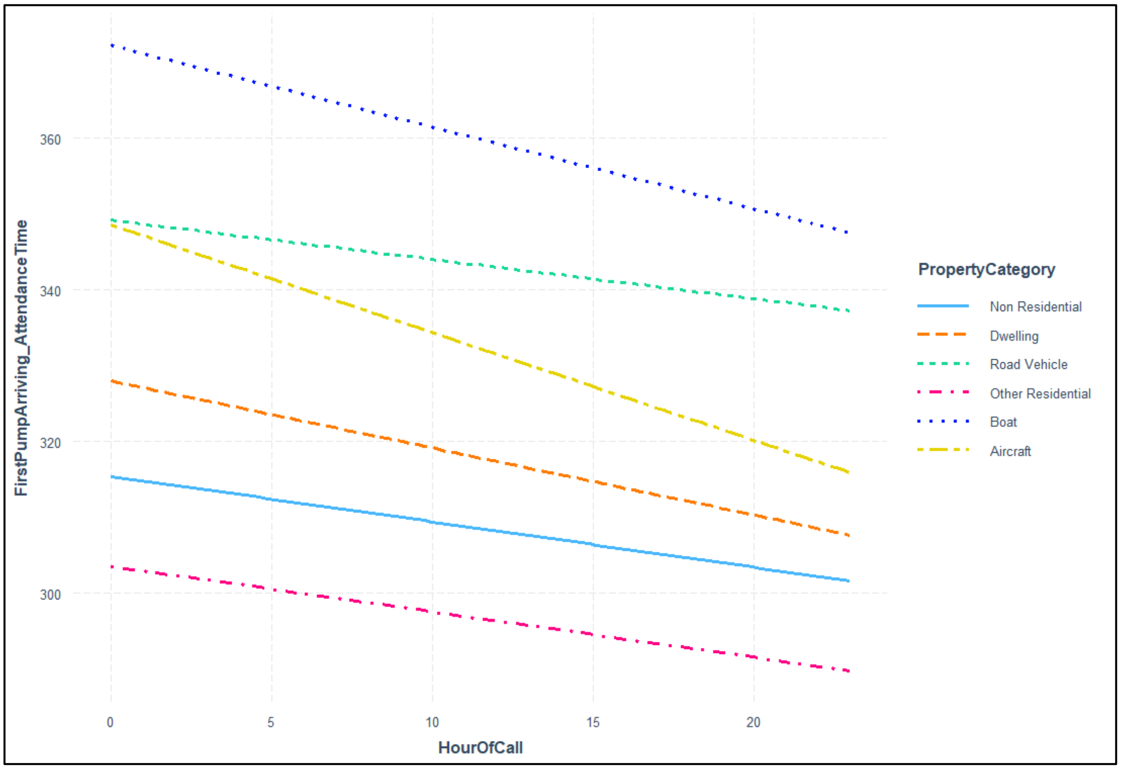
```
Call:
glm(formula = log(df$FirstPumpArriving_AttendanceTime) ~ PropertyCategory +
    HourOfCall + NumCalls, family = gaussian(link = "identity"),
    data = df)

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        5.7632563  0.0345097 167.004  < 2e-16 ***
PropertyCategoryBoat               0.0607501  0.0535713   1.134 0.256794
PropertyCategoryDwelling          -0.0569027  0.0344692  -1.651 0.098775 .
PropertyCategoryNon Residential   -0.1350891  0.0345055  -3.915 9.04e-05 ***
PropertyCategoryOther Residential -0.1219709  0.0346817  -3.517 0.000437 ***
PropertyCategoryOutdoor           -0.0417119  0.0346093  -1.205 0.228119
PropertyCategoryOutdoor Structure -0.0600667  0.0347288  -1.730 0.083704 .
PropertyCategoryRail Vehicle      -0.1044532  0.0458873  -2.276 0.022829 *
PropertyCategoryRoad Vehicle      -0.0797304  0.0346429  -2.301 0.021365 *
HourOfCall                        -0.0036437  0.0001425 -25.574  < 2e-16 ***
NumCalls                           0.0127078  0.0006368  19.955  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.3097058)

    Null deviance: 118190  on 379311  degrees of freedom
Residual deviance: 117472  on 379301  degrees of freedom
  (21424 observations deleted due to missingness)
AIC: 631850

Number of Fisher Scoring iterations: 2
```

(c) (3 points) Interpret the coefficients for **NumCalls** and **HourOfCall** on the response variable, assuming all other variables are held constant.

ANSWER:

Your boss isn't sure whether **HourOfCall** should be modeled as a categorical variable or a continuous numeric variable. They are also interested in how **HourOfCall** impacts **FirstPumpArriving_AttendanceTime** by different levels of **PropertyCategory**. You are provided with a plot of counts of Property Category by HourOfCall and an interaction plot. The interaction plot is based on fitting a GLM with **FirstPumpArriving_AttendanceTime** as the response variable and **PropertyCategory**, **HourOfCall** (as a continuous variable), and their interactions as the explanatory variables and then graphing the relationship between **HourOfCall** and **FirstPumpArriving_AttendanceTime** for different levels of **PropertyCategory**.





(d) (4 points) Discuss advantages and disadvantages of treating **HourOfCall** as a categorical variable versus a numeric variable in a GLM, including how the choice of a categorical vs. numeric variable would affect the interaction of **HourOfCall** with **PropertyCategory**.

ANSWER:

Task 6 (5 *points*)

Your assistant is starting to build a predictive model around real versus false alarms using the fire data and is looking for guidance around hyperparameters.

(a) (*4 points*) Give three examples of hyperparameters that can be used to control or limit how a single decision tree is built and for each example describe what it does within the decision tree.

ANSWER:

Your assistant is concerned about overfitting in their decision-tree model.

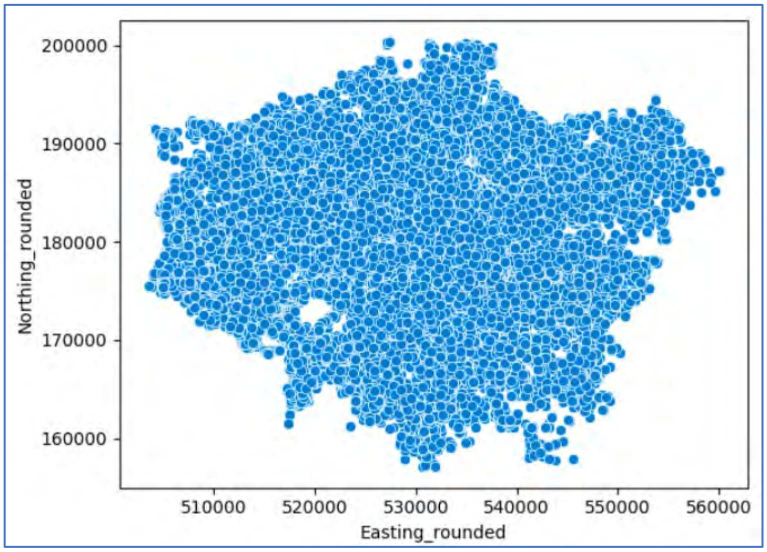
(b) (*1 points*) Choose one control parameter (hyperparameter) and recommend how it should be changed to address the overfitting concern. Justify your recommendation.

ANSWER:

Task 7 (7 *points*)

Your client has been concerned about the false alarm rate in the past years. To better allocate resources and reduce unnecessary responses, you are asked to build a model to predict the false alarm rate from historical data. Your boss suggests location (Northing and Easting) might help explain the false alarm rate. Your assistant decides to use a GLM instead of a tree-based model for this exercise.

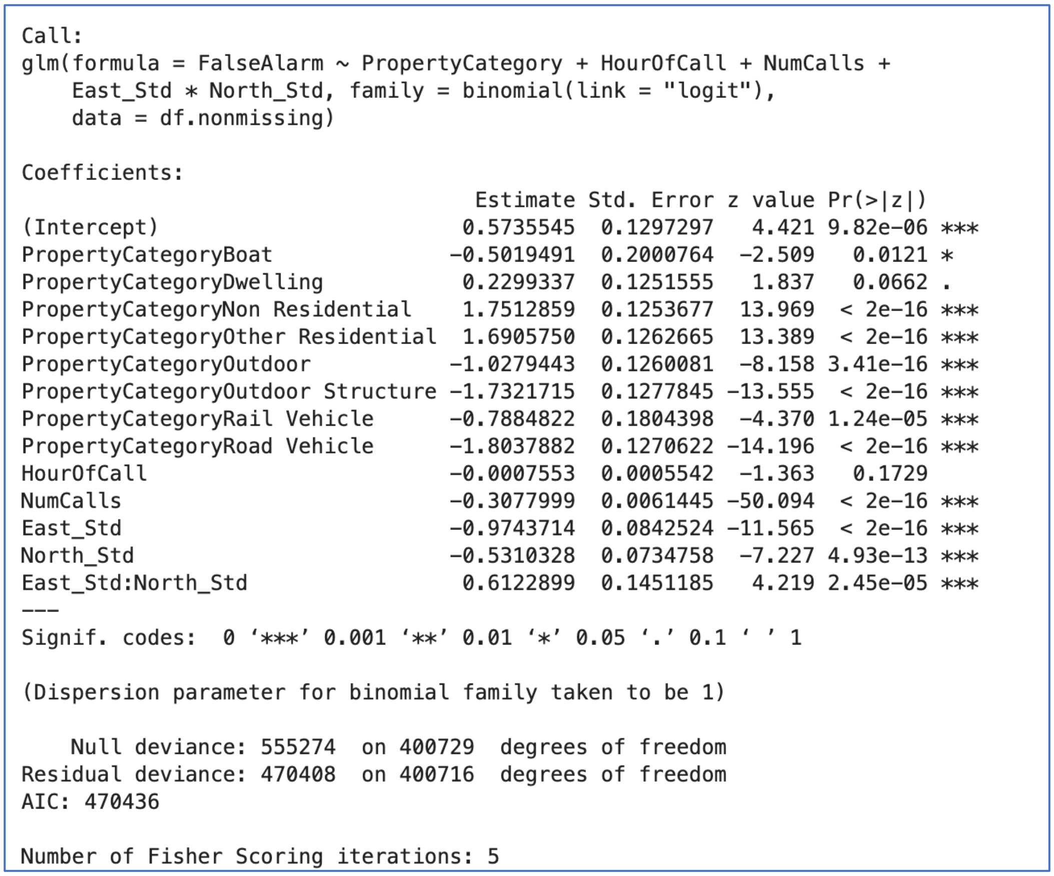
You are provided with a scatterplot of Easting and Northing.



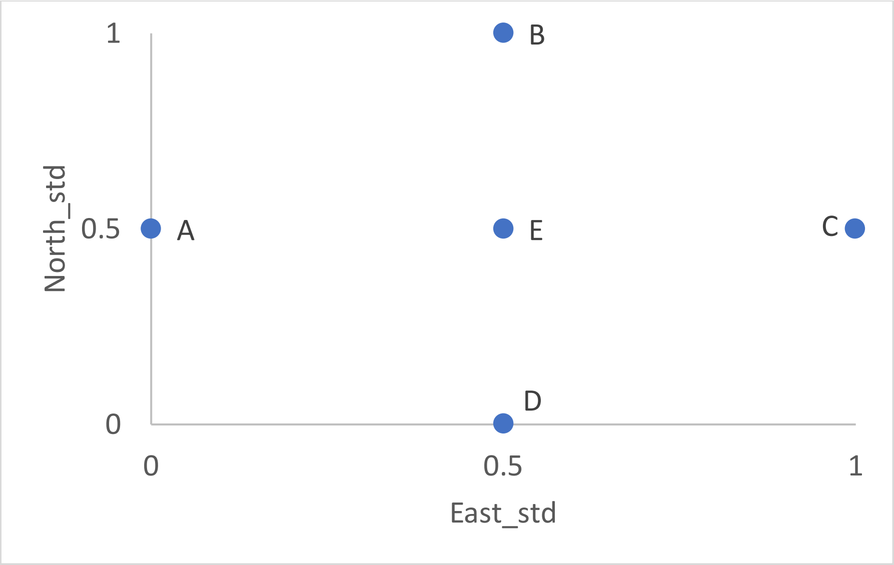
(a) (*2 points*) Explain what assumptions about the relationship between Easting and Northing and the false alarm rate are implicit in your assistant's choice to use a GLM instead of a tree-based model.

ANSWER:

Your assistant builds a GLM using logistic regression and provides you with a summary of the logistic regression model. In this model, Easting and Northing has been standardized between 0 and 1 as East_Std and North_Std, respectively.



```
Call:
glm(formula = FalseAlarm ~ PropertyCategory + HourOfCall + NumCalls +
    East_Std * North_Std, family = binomial(link = "logit"),
    data = df.nonmissing)

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        0.5735545  0.1297297   4.421 9.82e-06 ***
PropertyCategoryBoat              -0.5019491  0.2000764  -2.509   0.0121 *
PropertyCategoryDwelling           0.2299337  0.1251555   1.837   0.0662 .
PropertyCategoryNon Residential    1.7512859  0.1253677  13.969  < 2e-16 ***
PropertyCategoryOther Residential  1.6905750  0.1262665  13.389  < 2e-16 ***
PropertyCategoryOutdoor           -1.0279443  0.1260081  -8.158 3.41e-16 ***
PropertyCategoryOutdoor Structure -1.7321715  0.1277845 -13.555  < 2e-16 ***
PropertyCategoryRail Vehicle      -0.7884822  0.1804398  -4.370 1.24e-05 ***
PropertyCategoryRoad Vehicle      -1.8037882  0.1270622 -14.196  < 2e-16 ***
HourOfCall                        -0.0007553  0.0005542  -1.363   0.1729
NumCalls                          -0.3077999  0.0061445 -50.094  < 2e-16 ***
East_Std                          -0.9743714  0.0842524 -11.565  < 2e-16 ***
North_Std                         -0.5310328  0.0734758  -7.227 4.93e-13 ***
East_Std:North_Std                 0.6122899  0.1451185   4.219 2.45e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 555274  on 400729  degrees of freedom
Residual deviance: 470408  on 400716  degrees of freedom
AIC: 470436

Number of Fisher Scoring iterations: 5
```



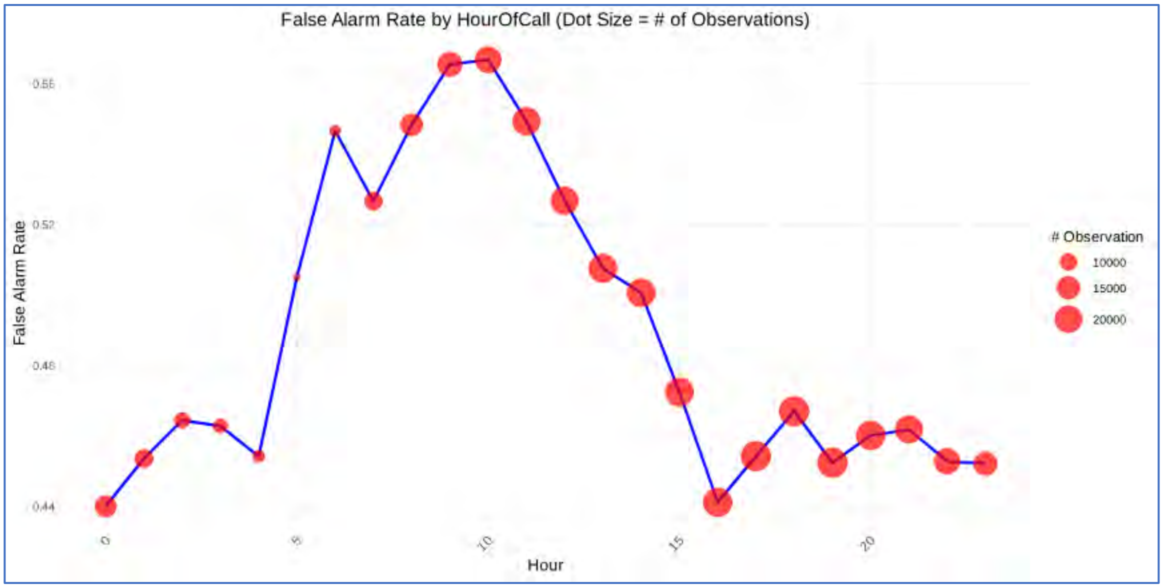
(b) (*3 points*) Identify which location on the graph above has the lowest false alarm rate, assuming other variables are constant across the city. Show your calculations.

A. The Westernmost point.

B. The Northernmost point.
C. The Easternmost point.
D. The Southernmost point.
E. The center of the city.

ANSWER:

Your assistant provides you with the graph of false alarm rate by **HourOfCall** below, and is concerned about the non-linear shape, and the coefficient of **HourOfCall** from the model summary above.



(c) (*2 points*) Explain a possible reason that the coefficient of **HourOfCall** is not statistically significant.

ANSWER:

Task 8 (3 *points*)

Your assistant is building a GLM using **NotionalCost**. They plan to transform the data by standardizing it using Min-Max scaling.

Min-Max Scaling formula: *New value = (current value-Min Value)/(MaxValue-MinValue).*

The summary statistics of the explanatory variable are:

	NotionalCost
count	400,736
mean	477
std	1,651
min	352
25%	364
50%	388
75%	430
max	433,598

(a) (*1 points*) Complete the highlighted cells in the table below for the standardized version of the NotionalCost variable, rounded to the nearest 0.1.

ANSWER:

	NotionalCost	**Standardized Variable**
count	400,736	
mean	477	
std	1,651	
min	352	
25%	364	
50%	388	
75%	430	
max	433,598	

(b) (*2 points*) Evaluate the effectiveness of the standardization approach above.

ANSWER:

Task 9 (10 *points*)

Your manager is hoping to better identify false alarms versus true alarms to better manage pump resources deployed to a call. The hope is that no or minimal resources can be deployed when a false alarm comes in – saving larger or multiple pumps for real emergencies.

(a) (*3 points)* Describe what a Receiver Operator Characteristic (ROC) curve is, what it's used for, and how it is created.

ANSWER:

Your assistant has put together a predictive model with the following confusion matrix results.

(b) *(3 points)* Calculate the AUC of the model based on the confusion matrix.

		Actual	
		False Alarm	Real Alarm
Prediction	Real Alarm	5890	12961
	False Alarm	7450	764

ANSWER:

You find out your boss outsourced the modeling request to an external consultant at the same time your assistant was developing their own model. The consultant delivered the confusion matrix for the model below.

Analyst Confusion Matrix

		Actual	
		False Alarm	Real Alarm
Prediction	Real Alarm	5890	12961
	False Alarm	7450	764

Consultant Confusion Matrix

		Actual	
		False Alarm	Real Alarm
Prediction	Real Alarm	4165	13236
	False Alarm	9175	489

(c) (*2 points*) Recommend to your boss which of the models you should present to the client. Justify your recommendation without explicitly referencing AUC.

ANSWER:

When your assistant built their confusion matrix output from their model, they used a classification threshold of 0.5, where values above the threshold are classified as real alarms.

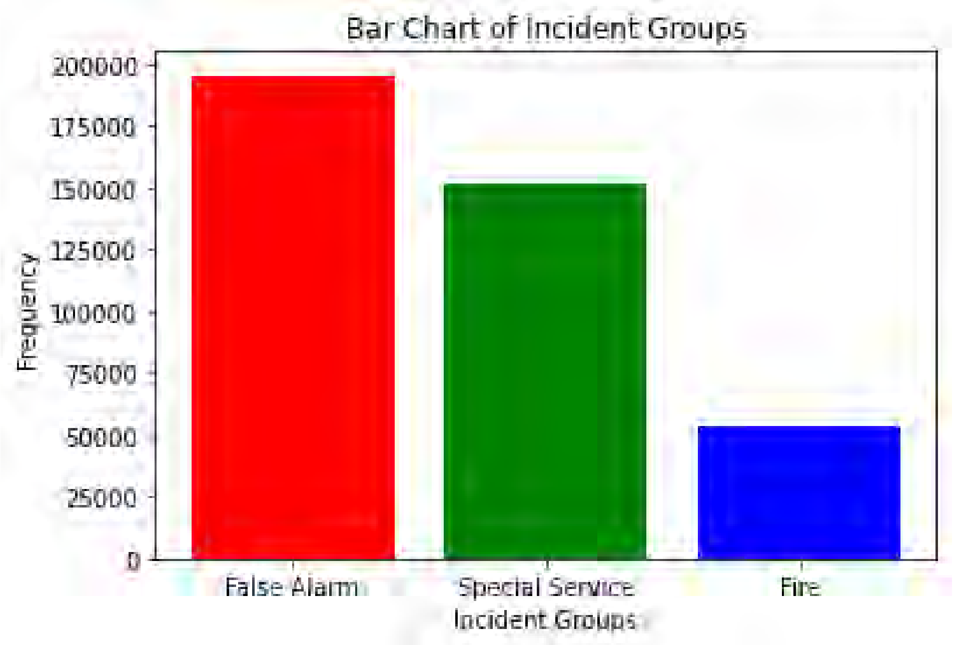
Assume the operational cost of misclassifying a real alarm as a false Alarm is significantly higher than misclassifying a false alarm as a real alarm.

(d) *(2 points)* Discuss the trade-offs of increasing or decreasing the threshold value in the context of the business problem.

ANSWER:

Task 10 (4 *points*)

You are interested in creating a tree-based model to predict fire. You begin by creating a bar chart to illustrate the proportion of each IncidentGroup.



(a) (*2 point*) Explain what makes data imbalanced and recommend one method to address the imbalance in order to improve predictive power.

ANSWER:

Your assistant recommends oversampling fire cases prior to splitting the data into training and testing sets.

(b) (*2 points*) Critique your assistant's recommendation.

ANSWER:

Task 11 (10 *points*)

(a) (*3 points*) Describe the process of backward selection in a regression model.

ANSWER:

The London Fire Brigade wants to build a GLM to predict **NotionalCost**, and have interpretable model outputs of the variables chosen. One of the explanatory variables is **PropertyType**, which has more than 200 levels.

(b) (*3 points*) Describe how each of the following four methods would treat a categorical variable with many levels.

a. Linear regression with backward selection
b. Ridge regression
c. LASSO
d. Elastic Net

ANSWER:

i. Linear regression with backward selection

ii. Ridge regression

iii. LASSO

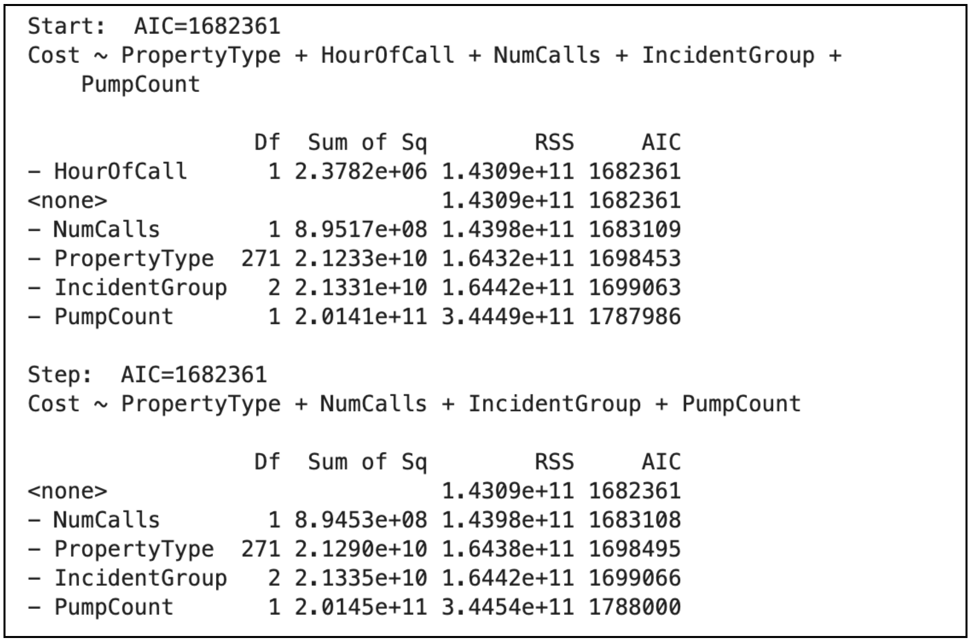
iv. Elastic Net

(c) (*2 points*) Recommend one method from part (b) above that addresses the client's request about interpretability. Justify your recommendation.

ANSWER:

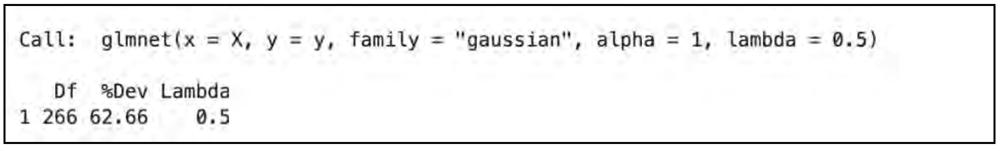
Your assistant provides you with the following modeling results:

<u>Backward Selection:</u>



```
Start:  AIC=1682361
Cost ~ PropertyType + HourOfCall + NumCalls + IncidentGroup + 
    PumpCount

                 Df  Sum of Sq        RSS     AIC
- HourOfCall      1 2.3782e+06 1.4309e+11 1682361
<none>                         1.4309e+11 1682361
- NumCalls        1 8.9517e+08 1.4398e+11 1683109
- PropertyType  271 2.1233e+10 1.6432e+11 1698453
- IncidentGroup   2 2.1331e+10 1.6442e+11 1699063
- PumpCount       1 2.0141e+11 3.4449e+11 1787986

Step:  AIC=1682361
Cost ~ PropertyType + NumCalls + IncidentGroup + PumpCount

                 Df  Sum of Sq        RSS     AIC
<none>                         1.4309e+11 1682361
- NumCalls        1 8.9453e+08 1.4398e+11 1683108
- PropertyType  271 2.1290e+10 1.6438e+11 1698495
- IncidentGroup   2 2.1335e+10 1.6442e+11 1699066
- PumpCount       1 2.0145e+11 3.4454e+11 1788000
```

Regularized Regression:



```
Call:  glmnet(x = X, y = y, family = "gaussian", alpha = 1, lambda = 0.5)

   Df  %Dev Lambda
1 266 62.66    0.5
```

Mean square errors on test dataset:

Model	MSE
Regularized Regression	1090.980
Backward Selection	1090.979

(d) (2 points) Recommend a model to your client. Justify your answer.

ANSWER: