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: 2001 – 2004	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

Candidates performed well on this task overall, especially on the first two definitions. Some responses could have been more concisely written. Strong candidates pointed out that descriptive analytics focuses on "what-if" scenarios.

ANSWER:

i. Descriptive analytics focuses on "what happened in the past?"
ii. Predictive analytics specifically looks at "what might happen next?"
iii. Prescriptive analytics helps answer the questions "What would happen if I do this?" or "What is the best course of action?"

(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

Most candidates performed well on this task. Partial credit was awarded for otherwise strong responses that failed to reference details in the data dictionary or struggled to articulate a reasonable business question for the analytic type of interest.

ANSWER:

i. The Easting_rounded and Northing_rounded or borough location data could be used to identify areas with more calls than others.
ii. An example of predictive analytics would be using the Easting_rounded and Northing_rounded location of the first, the station dispatching the fire pump and the HourOfCall to predict the time required for the first pump to arrive (FirstPumpArriving_AttendanceTime).
iii. A model could be used to determine, based on the number of calls (NumCalls) for a specific fire and the PropertyCategory and PropertyType, when a single pump vs. multiple pumps should be dispatched for a fire.

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.

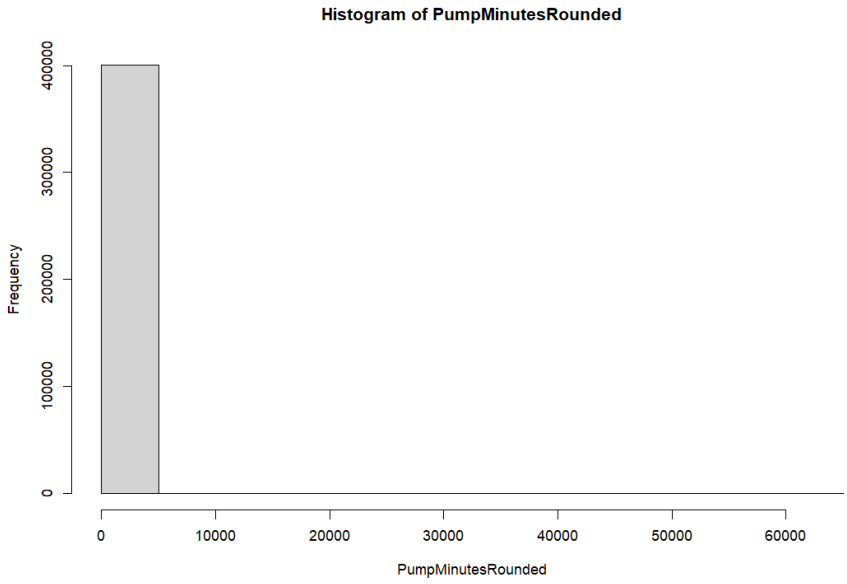
Candidate performance was strong on this task overall. Some candidates simply reiterated information from (a) and (b) without picking which type of analytics applies; these responses were not awarded any credit.

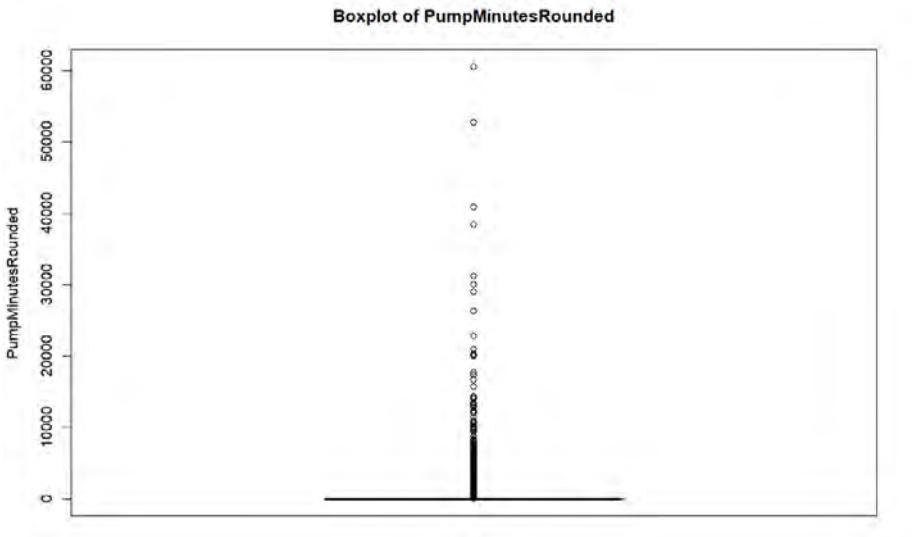
ANSWER:

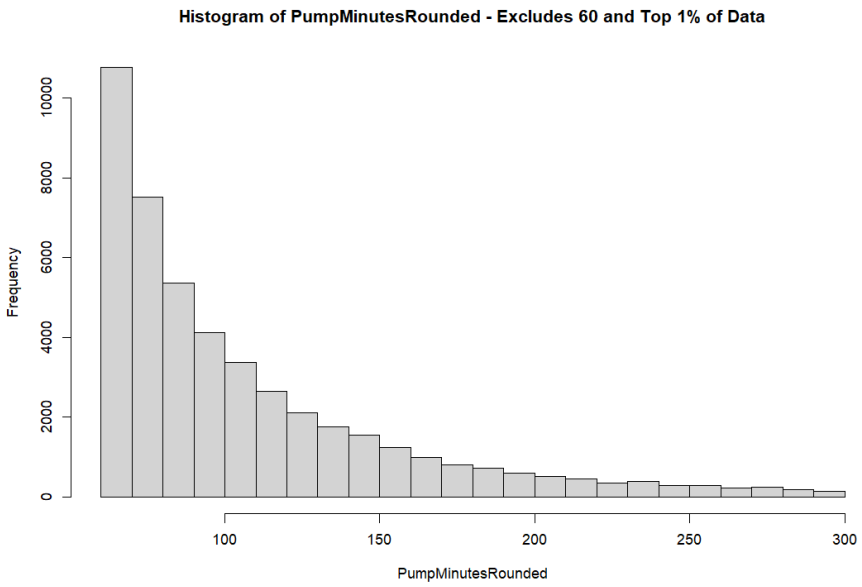
Descriptive analytics applies to this problem. This is an analysis of history – "What Happened?"

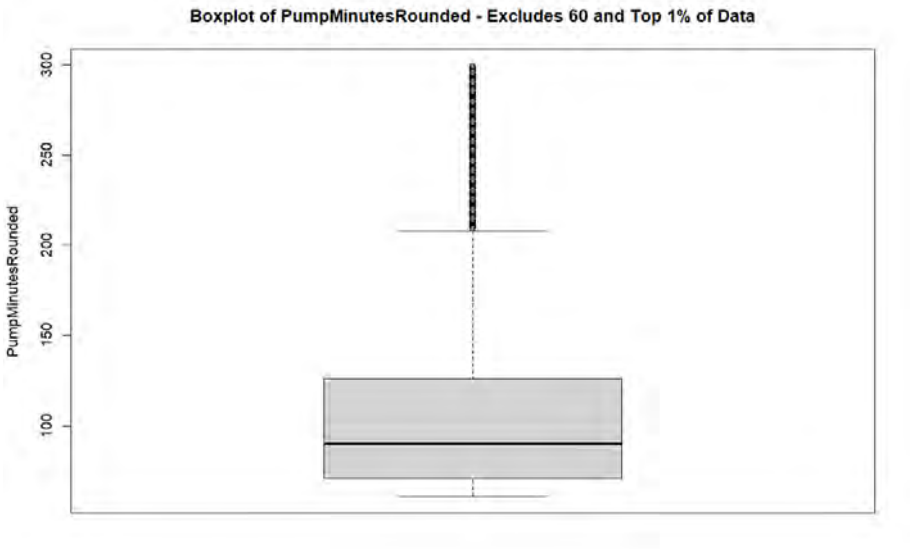
Task 2 (4 *points*)

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.

Candidate performance was mixed on this task. Most candidates gave an adequate description of the distribution of the data. Many candidates provided no description, or a very poor description of how the variable's skew and concentration at 60 impact its use as an explanatory variable.

ANSWER:

The data in the graphs above show that PumpMinutesRounded is heavily skewed and has a very large proportion of its mass at exactly 60 minutes. The combination of these characteristics means that including the unadjusted variable in models may lead to difficulty in in that it may be hard to isolate the

effect of a value of 60 versus any other value. The heavy skewness could lead to overfitting on the few very large observations.

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.

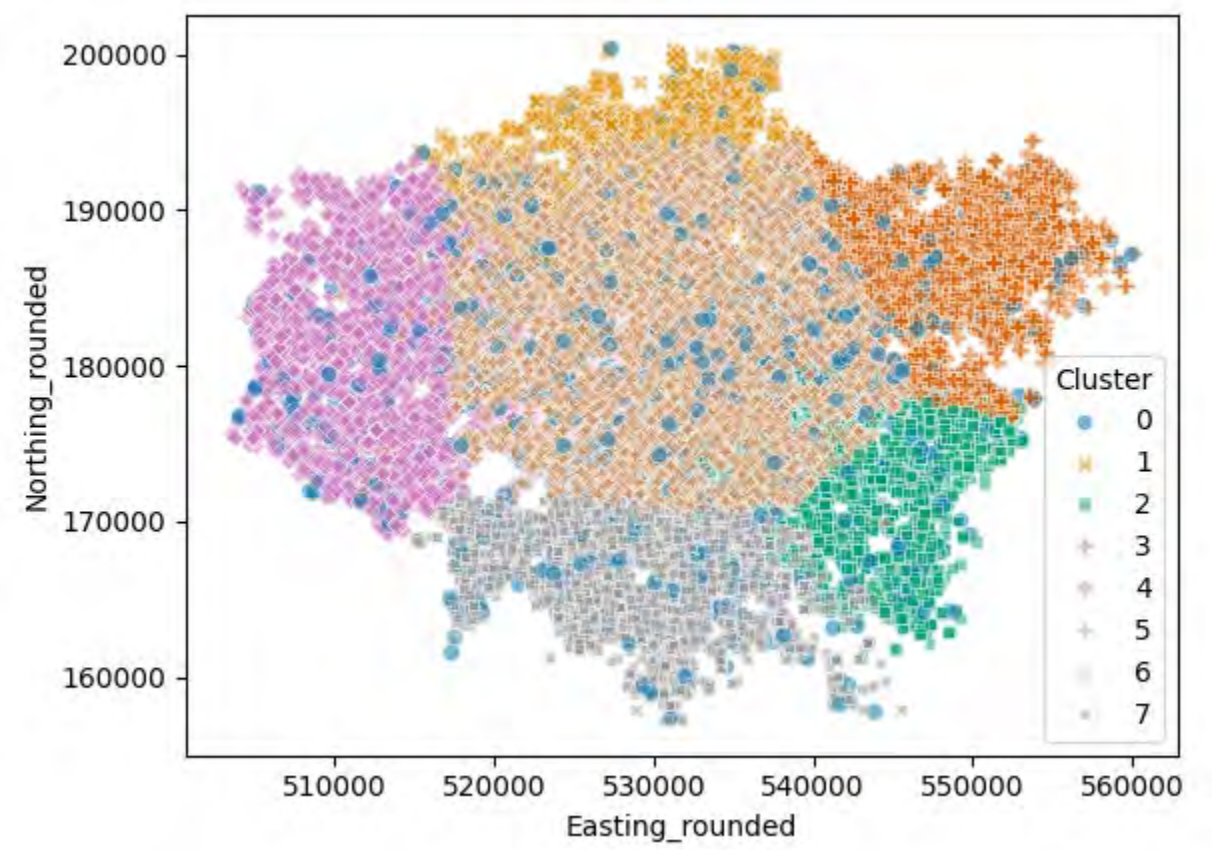
Candidates performed well on this task overall. Full credit was awarded for any transformation that improves the usability of the variable in a predictive model with justification. The most common recommendations were log and square root transformation. Some candidates recommended standardization or normalization; these recommendations do not improve the variable's usability in a predictive model and were therefore not awarded any credit.

ANSWER:
Reasonable transformations will be accepted if they make sense to help address the fact that the data is heavily skewed. Options include:

- Adding a dummy variable to indicate if a call is above or equal to the 60-minute mark. This would address the issue that the variation we're interested in is above 60-minutes.
- Applying a log or square root transformation would address the skewness and the outliers that are on the far right side of the distribution.

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.

Candidate performance was mixed on this task. Most candidates failed to discuss outliers in NotionalCost; this is a key observation required for full credit. Recommending changes to the approach with minimal or no connection to the data provided received minimal partial credit or no credit.

ANSWER:

The clustering results could be improved dramatically using normalization.

When looking at the summary statistics, we see that NotionalCost is not on a comparable basis to Northing_rounded & Easting_rounded. We can address this by scaling or normalizing all of the variables, which will generate more intuitive clusters. Also, there are outliers in the NotionalCost variable, which are skewing the results. Removing the outliers will dramatically improve the results.

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.

Candidate performance was mixed on this task. Most candidates failed to discuss target leakage introduced by including NotionalCost in the clusters; this is a key observation required for full credit. Some candidates copied or paraphrased responses from the prior task, not addressing this task in the context of a predictive modeling problem.

ANSWER:

This is not a feasible solution as we are using the target variable inside the explanatory variables (the clusters) causing data leakage issues.

You would need to remove the notional cost from the clustering and redevelop clusters without that variable.

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.

Candidate struggled with this task for a variety of reasons. Meaningful observations based on the context of the models specified (GLM and tree-based models) were required for full credit. Many responses were unclear, contradictory, or incorrect.

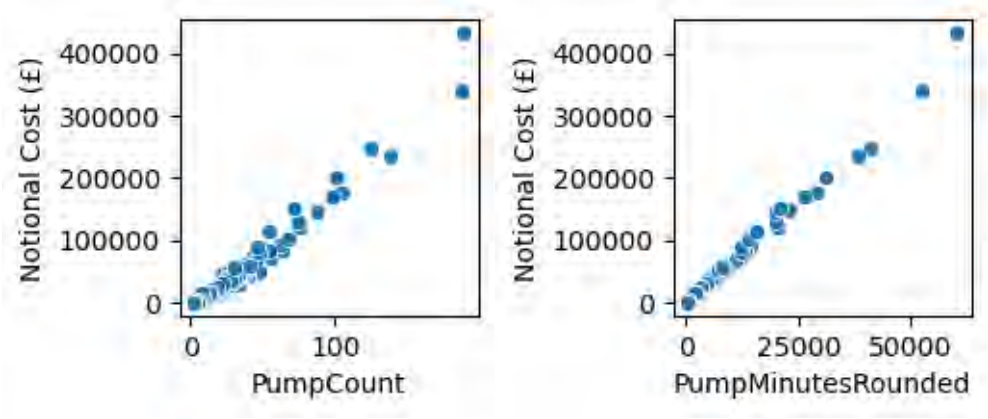
ANSWER:

Clustering variables may be more advantageous for GLMs if the target variable does not have a direct linear correlation with the original variables. This advantage may be lost when using tree-based models, which are capable of modeling non-linear effects and complex interactions.

Another consideration is that using clusters can lead to more intuitive interpretations of the model results as one can easily plot the cluster groups over a map and describe its use visually as an explanatory variable.

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.

Candidate performance was mixed on this task. Full-credit answers explained how the charts provided indicate a potential collinearity issue. Many candidates claimed that the charts indicate that an interaction exists between the two variables; this answer was considered incorrect and not awarded credit.

ANSWER:

Based on the one-way charts provided there is a clear linear relationship between NotionalCost and PumpMinutesRounded as well as between NotionalCost and PumpCount. Given the linear relationship between PumpCount and NotionalCost and PumpMinutesRounded and NotionalCost, it is possible that PumpCount and PumpMinutesRounded also have a linear relationship (i.e. collinearity). The reason collinearity is an issue in explanatory variables is that the effect of the individual variables is difficult to estimate leading to higher standard errors of coefficient estimates and making it difficult to interpret the impact of the individual variables on the response variable.

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

Candidates struggled with this task overall. Full credit was awarded for a well-justified recommendation that described how to address multicollinearity. The most common recommendations receiving full credit were selecting a variable based on statistical significance, stepwise variable selection, and regularized regression.

ANSWER:

Explore the relationship between PumpMinutesRounded and PumpCount. If there is a clear linear relationship then this would result in multicollinearity and one approach is to remove one of these variables, typically the one with the higher *p*-value when both are used. A potential way to identify this is to do a pairwise plot between PumpMinutesRounded and PumpCount.

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**.

Candidate performance was mixed on this task. Full credit was awarded for identifying IncidentGroup as an inappropriate predictor together with an explanation based on either the fact that the value is not known at the time of a call or target leakage concerns.

ANSWER:

IncidentGroup can't be used as a predictor variable in this model. This value is only known after resources have been sent to address the call. This means including IncidentGroup as a cost predictor could lead to data leakage.

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 assistant provides you with a summary of the 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
```

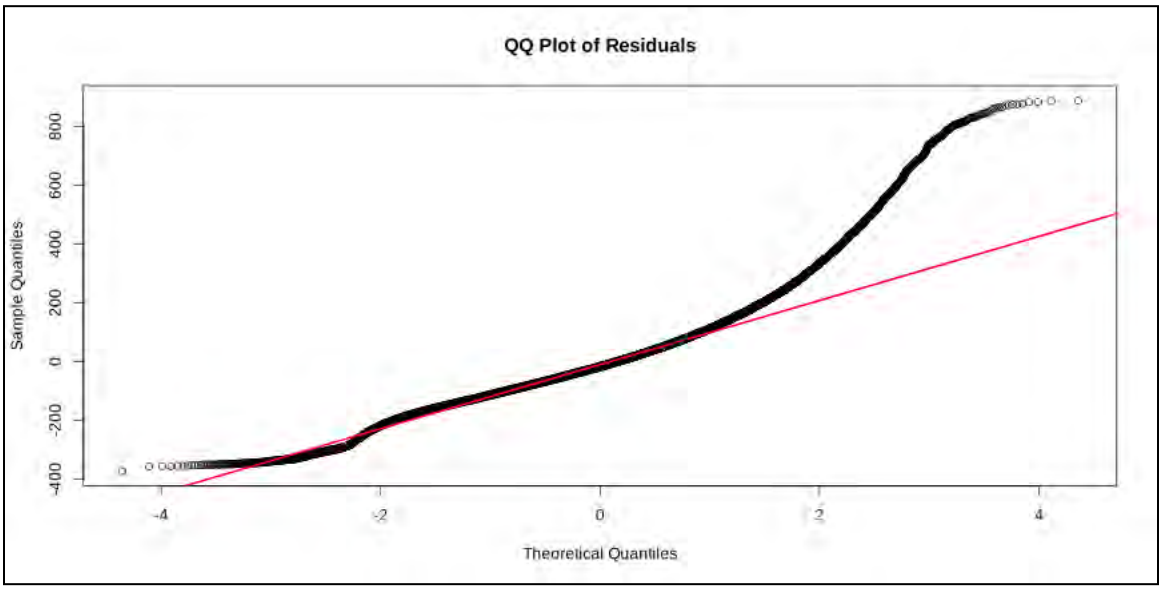
(a) (1 point) Describe the distribution of attendance time in terms of its skewness.

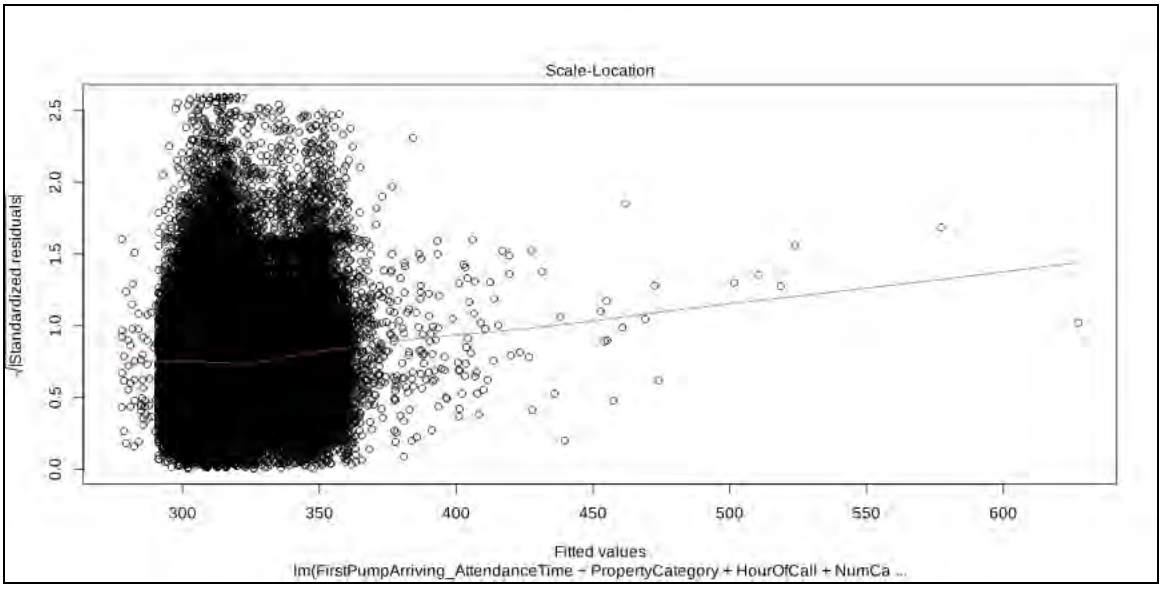
Candidates performed well on this task overall. Full credit was awarded for explaining how the information provided indicates right-skew in the variable.

ANSWER:

The mean (318.1) is greater than the median (300.0), indicating a **right-skew** (positively skewed distribution). This suggests that there are higher attendance times pulling the mean to the right.

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.

Candidate performance was mixed on this task. Either recommendation was awarded full credit provided a strong justification. Most candidates were able to demonstrate an understanding of the log transformation in their responses. However, fewer candidates were able to provide sufficient detail about the log-link to fully justify their recommendation.

ANSWER:

A Log link function is recommended in modeling of **FirstPumpArriving_AttendanceTime.**

Link functions directly model the relationship between the mean of the response variable and the predictors, which allows variance to naturally depend on the mean via the distribution. Link functions through a GLM also provides more interpretable results in terms of effects on the original scale.

Variable transformations (for example, log transformation) stabilize variance directly by transforming raw data and make the residuals closer to normal but do not explicitly address the relationship between variance and mean. Interpretation can be difficult on the model coefficients as they are interpreted as effects on the log of the response.

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.

Candidate performance was mixed on this task. Full-credit responses applied the link function correctly and interpreted the results both directionally (whether an increase in the variable increases or decreases the prediction) and in terms of magnitude.

ANSWER:

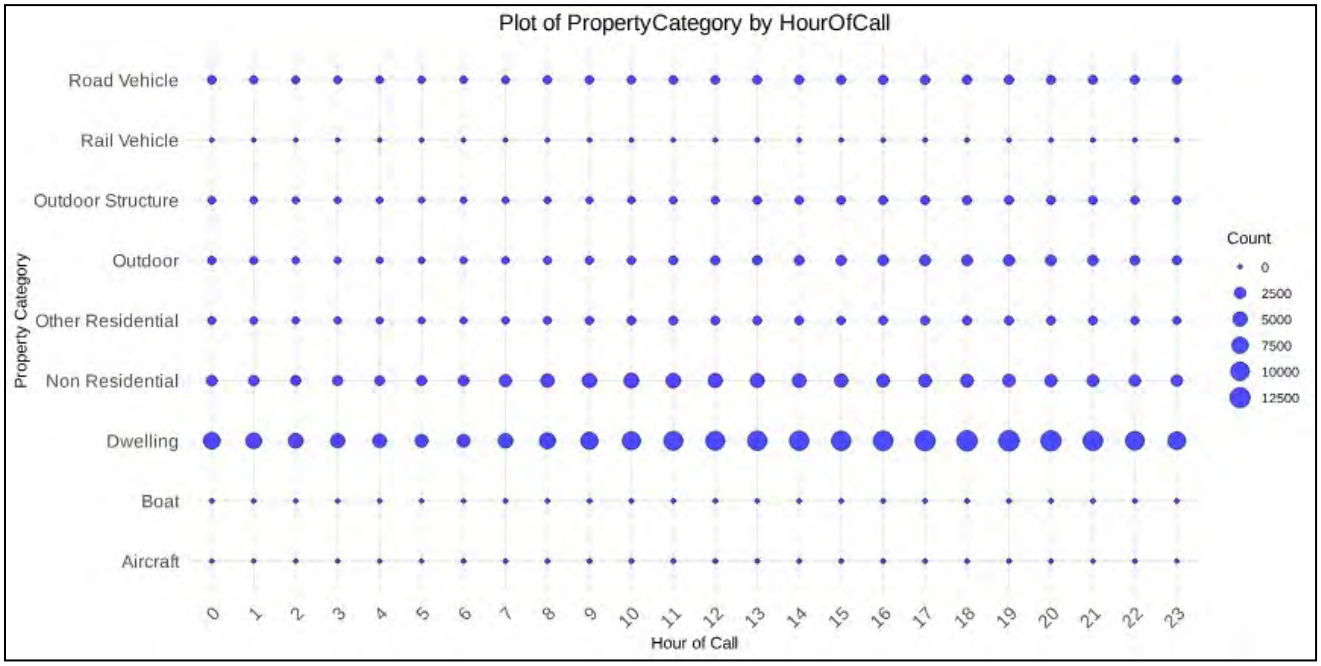
NumCalls: For each additional call received, the log of the first pump arrival time increases by 0.0127.

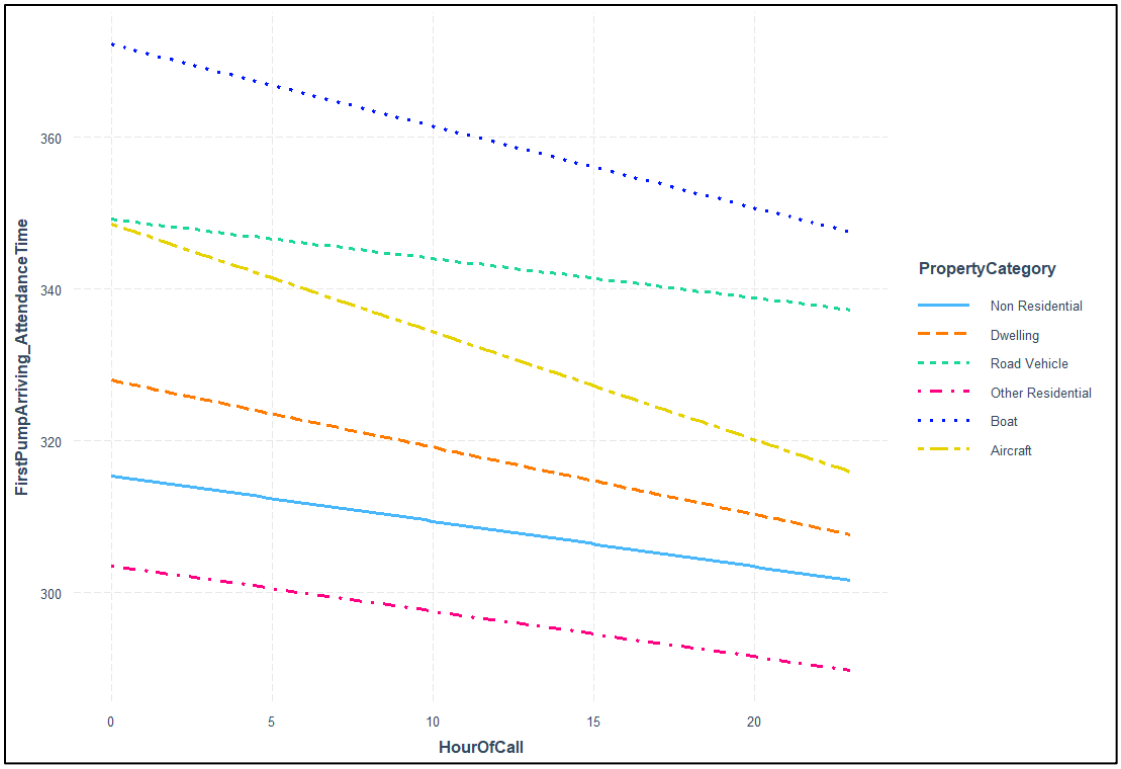
Since the response variable is log-transformed, to understand the actual impact on the raw arrival time, exponentiate the coefficient: $e^{0.0127} = 1.0128$. This means that for each additional call, the first pump arrival time increases by approximately **1.28%**, assuming all other variables are constant.

HourOfCall: For each one-hour increase in the time of the call, the log of the first pump arrival time decreases by 0.0036437

Exponentiate the coefficient: $e^{-0.0036437} = 0.9964$. This means that for each additional hour in the time of the call, the first pump arrival time decreases by approximately **0.36%**, assuming all other variables are constant.

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**.

Candidates struggled with this task overall. Full-credit responses provided at least one advantage and disadvantage for each treatment of the variable, including at least one observation pertaining to the interaction term.

ANSWER:

Advantages of treating **HourOfCall** as a numeric variable:

- If the impact of **HourOfCall** is linear, treating it as numeric provides a straightforward model.
- The model uses a single coefficient, reducing the number of parameters and avoiding overfitting.
- For the interaction with **PropertyCategory**, treating **HourOfCall** as a numeric variable allows for different slopes of the **HourOfCall** variable depending on the Property Category.

Disadvantage treating **HourOfCall** as a numeric variable:

- Missesnonlinear hour-specific patterns
- For the interaction, there must still be a linear relationship between HourOfCall and the response variable, but it can vary by each level of PropertyCategory.

Advantage treating **HourOfCall** as a categorical variable:

- Captures hour-specific effects: Each hour is modeled independently, allowing the model to capture unique patterns (e.g., rush hours, night shifts) and their impact
- It does not assume a smooth or linear relationship between **HourOfCall** and the response variable.
- For the interaction variable, this approach allows for a more flexible relationship between HourOfCall and PropertyCategory as an interaction in which certain types of dwellings can exhibit different types of non-linear relationships, as an example, there could be parts of the day where the arrival time is higher for certain property categories and lower for other property categories.

Disadvantage treating **HourOfCall** as a categorical variable:

- Modeling each hour as an independent category increases the risk of overfitting, especially with sparse data for certain hours.
- Some hours and property categories may have low sample size (as shown in the plot), leading to unreliable coefficient estimates for those hours.

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.

Candidates performed well on this task overall. Full-credit answers listed three applicable hyperparameters and described how they impact the decision tree. Candidates could use either plain English names for the hyperparameters (e.g. "complexity parameter") or parameter names used in code implementations (e.g., "cp"). The two most common errors were confusing minsplit with minbucket and struggling to describe the cp hyperparameter.

ANSWER:

The answer should cover three of four potential Control Parameters listed below:

- minsplit: This is the minimum number of observations that are need in a node before it splits. The higher this value, the fewer splits made. This simplifies the decision tree.
- minbucket: This parameter says that any resulting node after a split will need a minimum number of observations. If the split results in a node with fewer observations than this parameter the split will not be made, resulting in a simpler tree.
- cp: This is the complexity parameter metric, which defines a minimum amount of impurity reduction needed for a split to be made. Any split that does not decrease the overall lack of impurity won't be made.
- maxdepth: This parameter sets a limit on the maximum depth the tree can have and forces it not to get too complicated.

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.

Candidates performed very well on this task, with most candidates earning full credit. The most common reason for candidates receiving partial credit was lacking proper justification for their recommendation. A few candidates recommended changing a parameter in the opposite direction of what would be required to reduce overfitting.

ANSWER:

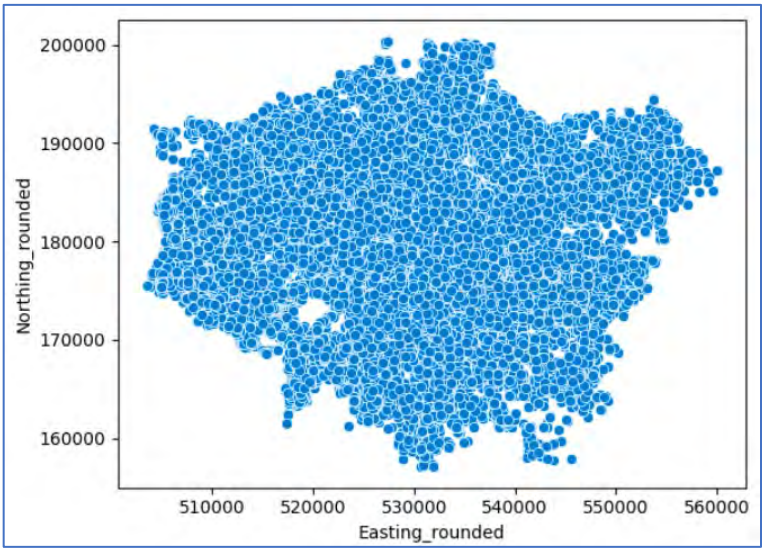
- MinSplit: Increase minimum number of samples required to split an internal node. Increasing this value forces the tree to be less sensitive to small changes in the data.
- MinBucket: Increase the minimum number of samples needed in a node after a split is made. Increasing this value forces the tree to be less sensitive to small changes in the data.

- Complexity Parameter: Increase the threshold for the minimum impurity decrease needed for a split. This ensures that only significant splits are made.
- MaxDepth: Reduce the maximum depth (number of splits) in the tree. This will prevent the tree from growing too deep and overfitting the training data.

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.

Candidates struggled with this task overall. Full credit was awarded for explaining how a GLM assumes that each of these geospatial variables has a linear relationship with the target variable. Several candidates skipped this task entirely. Some candidates claimed that the chart shows a linear relationship between Northing_rounded and Eastin_rounded, which is incorrect.

ANSWER:

Your boss is wanting to capture the relationship between a neighborhood's geographical location and the target variable. Since your assistant chose a GLM, they are assuming that the target variable is a function of a linear combination of the variables. This means that as the events originate further east or north, they will increase/decrease consistently with the explanatory variable without variation in between. The approach does not allow for the model to identify "neighborhoods" with different false alarm rates.

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.

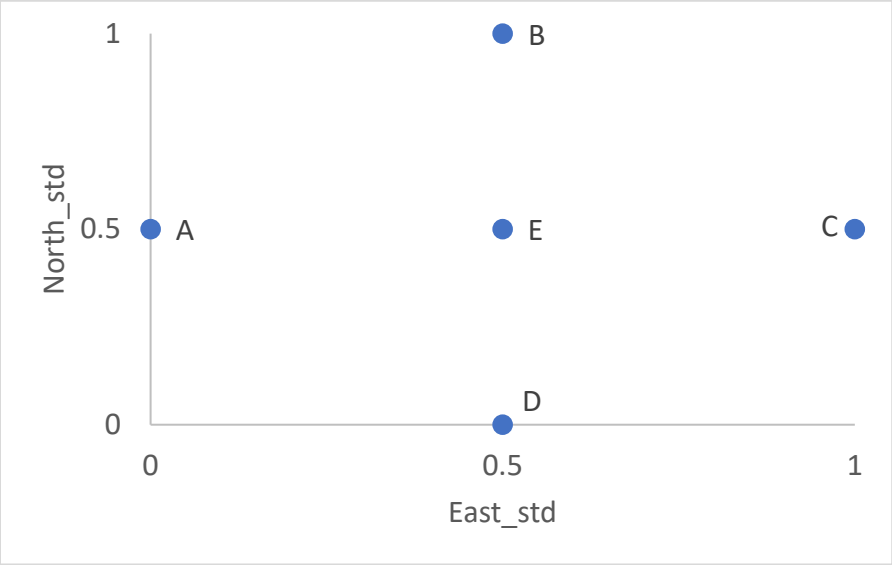
```
Call:
glm(formula = FalseAlarm ~ PropertyCategory + HourOfCall + NumCalls +
    East_Std * North_Std, family = binomial(link = "logit"),
    data = df.nonmissing)

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.

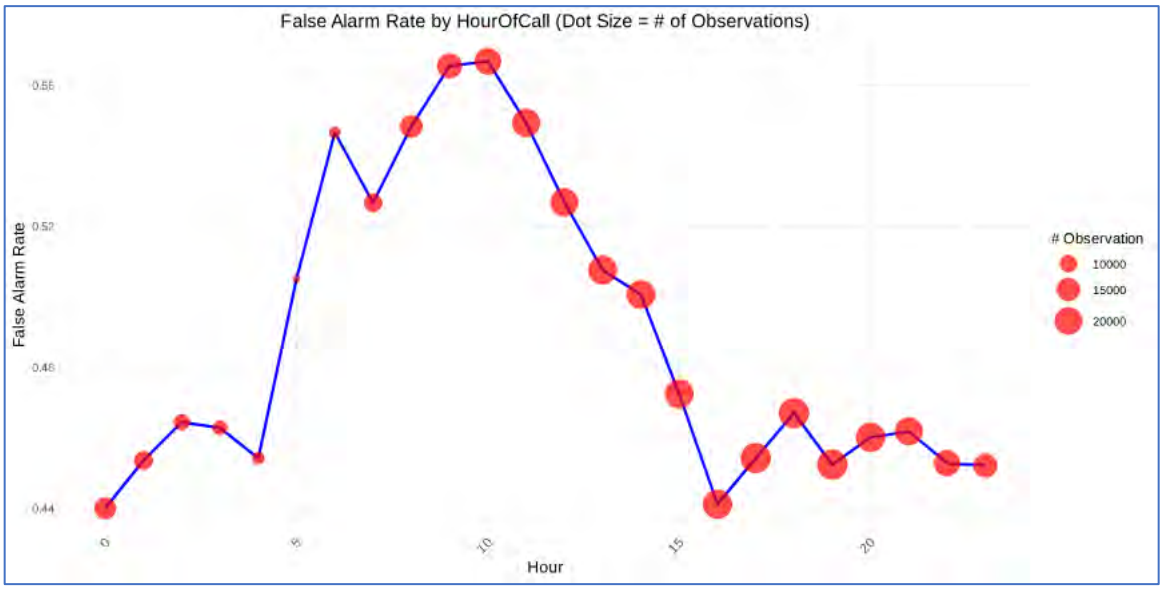
A. The Westernmost point.
B. The Northernmost point.
C. The Easternmost point.
D. The Southernmost point.
E. The center of the city.

Candidate performance was mixed on this task. There are a few valid, equivalent ways of applying min-max scaling; any such approach with correct calculations and a correct conclusion was awarded full credit. A common mistake was ignoring the interaction term; these responses were awarded partial credit if the calculation was otherwise correct.

ANSWER:

Easternmost: $-0.974 * 1 - 0.531 * 0.5 + 0.612 * 1 * 0.5 = -0.9335$
Westernmost: $-0.974 * 0 - 0.531 * 0.5 + 0.612 * 0 * 0.5 = -0.2655$
Northernmost: $-0.974 * 0.5 - 0.531 * 1 + 0.612 * 0.5 * 1 = -0.712$
Southernmost: $-0.974 * 0.5 - 0.531 * 0 + 0.612 * 0.5 * 0 = -0.487$
Center: $-0.974 * 0.5 - 0.531 * 0.5 + 0.612 * 0.5 * 0.5 = -0.5995$
Lowest false alarm location is **Easternmost**

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.

Candidate performance was mixed on this task. Responses that recognized non-linearity in the chart and accurately explained how that impacts fitting of a GLM, resulting in insignificant coefficient estimates received full credit.

ANSWER:

The model is only including linear term for HourOfCall though the plot suggests a non-linear pattern. No single slope will fit this shape well. As a result, the linear coefficient can appear small (and hence statistically insignificant) even if HourOfCall strongly influences the outcome in a non-linear way**.**

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.

Candidates performed well on this task overall. Full credit was awarded for correct calculations, including rounding. Partial credit was awarded for incorrect rounding.

ANSWER:

The denominator of the scaling formula is 433,592 – 352 = 433,246

	NotionalCost	**Standardized Variable**	**Unrounded Calculation**
count	400,736	400,736	
mean	477	0.0	(477 – 352) / 433,246
std	1,651		
min	352	0.0	(352 – 352) / 433,246
25%	364	0.0	(364 – 352) / 433,246
50%	388	0.0	(388 – 352) / 433,246
75%	430	0.0	(430 – 352) / 433,246
max	433,598	1.0	(433,246 – 352) / 433,246

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

Candidate performance was mixed on this task. Full credit responses identified that skewness or outliers are concerns with the variable, and explained why the transformation does not address these concerns.

ANSWER:

This approach can be used to transform all values to the same scale: a range between 0 and 1. Looking at the data we can see the data is right-skewed or has very large outliers. Min-Max scaling would not be an appropriate transformation since it does not address skewness. It would be more appropriate to log-transform or remove outlier values.

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.

Candidate performance was mixed on this task. Full-credit answers correctly defined ROC, including a description of each of the axes and how each point corresponds to a different cut-off value, and discussed how it is used in evaluating the performance of a classification model.

ANSWER:

An ROC curve is a way of determining how well a classification model performs. At least three points are used to graph the curve. The graph has the True Positive Rate on the y-axis (also known as the sensitivity) while the x-axis maps the False Positive Rate of the results (calculated as 1 minus the specificity). Two points that are always calculated on the graph are (0,0) and (1,1) where all observations are predicted to be either false alarms or real alarms respectively. The other points show the TPR and FPR for the confusion matrix that the model generates using a variety of cut-off values. If multiple cut-off values are used, a point can be generated from each confusion matric, leading to a smoother curve.

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

Candidates struggled with this task. Few candidates completing the entire calculation, which closely follows an example provided in the modules. Partial credit was awarded for correctly calculating sensitivity and specificity, which are relevant to the ROC calculation.

ANSWER:

The confusion matrix above generates a TPR (Sensitivity) of .944335 and a FPR (1-Specificity) of .441529. With three points on this curve: (0,0) (.441529,.944335) and (1,1) the calculation of the AUC is a matter of geometry and breaking things down into three components to calculate the area.

The first component is a triangle with a base of .441529 and a height of .944335 (area = .20848)

The second component is a rectangle with a base of (1-.441529) and a height of .944335 (area = .52738)
The final component is a triangle with a base of (1-.441529) and a height of (1-.944335) (area = .15540)

The total area under the curve is the sum of the three areas calculated above: .7514

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.

Candidates performed well on this task, with most candidates receiving full credit. A few candidates provided weak or no justification to support their recommendations, receiving partial credit. A few candidates recommended the assistant's model, citing better model interpretability or less complexity; these answers were awarded no credit since there is no support for this conclusion.

ANSWER:

A quick review of the output shows that the consultant's model is giving better results. It is correctly identifying more false alarms that are actual false alarms and more real alarms that are real alarms as compared to our analyst's initial confusion matrix.

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.

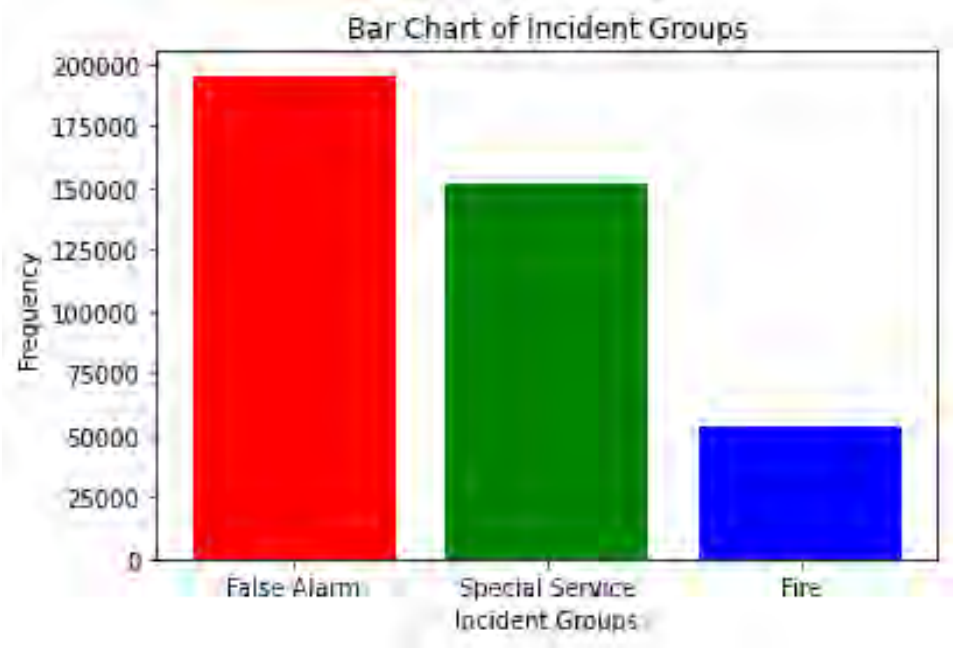
Candidates performed well on this task overall. Partial credit was awarded for only discussed lowering the threshold without addressing the other direction. A few candidates mixed up their descriptions of increasing the threshold as opposed to decreasing it; partial credit was awarded for these answers if they were otherwise consistent.

ANSWER:

When increasing the threshold, the model:
• Reduces false positives, which can be good for reducing unnecessary responses.
• Increases false negatives, meaning more true alarms get wrongly labeled as false alarms. Because this kind of misclassification has a significantly higher cost in this scenario, raising the threshold too high can become expensive if genuine emergencies are dismissed.
When lowering the threshold:
• Reduces false negatives, so fewer true alarms are missed (important when their misclassification cost is high).
• Increases false positives, meaning more alarms that turn out to be real are incorrectly flagged as false, potentially resulting in more wasted resources.

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.

Candidates performed well on this task overall. Full-credit responses correctly explained imbalanced data and made a strong recommendation. Either oversampling and undersampling was considered appropriate and awarded full credit as part of a strong recommendation.

ANSWER:

- The data is imbalanced meaning most of the target variable is concentrated in one value.
- Given this, if we utilize oversampling or undersampling we can then create a more balanced data set. Oversampling replicates rows of the data where the explanatory variable has the less frequent value. Undersampling drops some rows of the data where the explanatory variable has the more frequent value(s).

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

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

Candidates performed well on this task overall. Most candidates correctly identified the data leakage concern as part of their critique.

ANSWER:

Oversampling fire is appropriate because the data set is imbalanced. However, it is not appropriate to perform oversampling prior to splitting the data into training and testing sets.

The test set should be unaltered data to independently validate our model's performance. Oversampling before splitting the data can result in duplicates of the same observation ending up in both the train and test datasets, a form of data leakage.

Task 11 (10 *points*)

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

Candidates performed well on this task overall. Full-credit answers described the iterative nature of the backward selection procedure, the performance criteria used to evaluate the model at each iteration, and how predictors are dropped in each iteration. Many candidates failed to describe how the performance criteria is used in the algorithm.

ANSWER:

1) Start with the Full Model: Fit a regression model with all available predictors included.
2) Evaluate Predictors:
    Assess the significance of each predictor using the chosen criterion:
    P-values: Check the statistical significance of each predictor's coefficient.
    AIC/BIC: Evaluate the overall model fit and complexity with each predictor removed.
    Remove the least significant predictor.
3) Remove this predictor from the model. Refit the regression model without the removed predictor.
    Recalculate the metric used in step 2 for the reduced model.
4) Repeat the Process: Continue removing the least significant predictor in each step.
    Stop when:
    All remaining predictors are significant (if using p-values).
    Removing any additional predictors increases AIC or decreases model fit.
5) Select the final model:
    The process ends when no further improvements can be made, and the final model contains only the most significant predictors.

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.

Performance was somewhat mixed on this task. Candidates fell short of achieving full credit for a variety of reasons. Some responses mistakenly stated that binarization is required before performing the linear regression. Some responses failed to accurately describe the different ways that coefficients are treated in Ridge/Lasso/Elastic Net regularization methods.

(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

Backward selection uses sequential selection based on statistical criteria. It treats PropertyType as a single unit, requiring all dummy variables to enter or leave the model together. However, if the dummy variables are created prior to beginning the selection process, factor levels can be removed one at a time.

ii. Ridge regression

Ridge regression will keep all of the variables while reducing their influence on the final predicted model. This will not address the issues of computational efficiency and interpretability.

iii. LASSO

Lasso regularization will select the most important levels (or features after one-hot encoding). By increasing the penalty parameter (lambda) we can incrementally reduce the number of variables or levels used in the final prediction model.

iv. Elastic Net

Elastic Net, being a combination of Ridge & Lasso, will likely not remove as many levels for this problem.

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

Performance was strong overall on this task. Full credit was awarded for recommending backward selection, LASSO, or Elastic Net with a strong justification grounded in model complexity and interpretability.

ANSWER:

LASSO Regression is recommended for the following reasons:

- Can perform variable selection at the individual level category, allowing some levels of PropertyType to be retained while others are shrunk to zero
- More efficient with many categorical levels as it can identify and retain only the most influential property types

The resulting model is less complex and more interpretable because it has fewer variables.

Your assistant provides you with the following modeling results:

Backward Selection:

```
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.

Performance was strong overall on this task. Either model was considered acceptable with reasonable justification. Some responses failed to recognize the similarity between the two MSE measures, establish the connection between degrees of freedom and model simplicity, or make a clear recommendation.

ANSWER:

Regularized regression is recommended because:

1) Both models have similar MSE.
2) Regularized regression retains fewer predictors, simplifying the model while maintaining accuracy. Regularized regression has 266 degrees of freedom vs. Backward selection has 275 degrees of freedom total.