Exam PA October 2024 Project Statement

IMPORTANT NOTICE – THIS IS THE OCTOBER 15, 2024, PROJECT STATEMENT. IF TODAY IS NOT OCTOBER 15, 2024, SEE YOUR TEST CENTER ADMINISTRATOR IMMEDIATELY.

General Information for Candidates

This examination has 10 tasks numbered 1 through 10 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. This exam includes an Excel data file with information for Task 5. 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.

The Word file that contains your answers must be uploaded before the five-minute upload period time expires.

Business Problem

You are working for a firm that consults on energy use in the United States. Your firm serves a wide range of clients including energy producers, municipal governments, developers, and building owners. Your clients are interested in using data to understand patterns of existing energy use and to predict energy usage in the future.

Your firm is currently focused on the Chicago market and will use data from the city of Chicago[1] that looks at detailed energy use at the census block[2]-level in residential, commercial, and industrial buildings. The energy data has also been enriched with weather data.[3]

Notes on the data set:

Energy data is collected at the census block level for specific building types (residential, commercial, and industrial).

U.S. census data is organized at different levels of granularity with the census block being the most granular. A census block within an urban area typically represents a single city block.

The weather data is captured at a daily level and is collected from Chicago's Midway airport and used for the entire city.

[1] *Source: City of Chicago – Chicago Data Portal*

[2] A census block is the smallest area used by the U.S. Census Bureau. In a city, it is typically an area bounded by four streets with no streets running through it. They will be referred to as simply "blocks" in this assessment.

[3] *Source: National Oceanic and Atmospheric Administration – National Centers for Environmental Information*

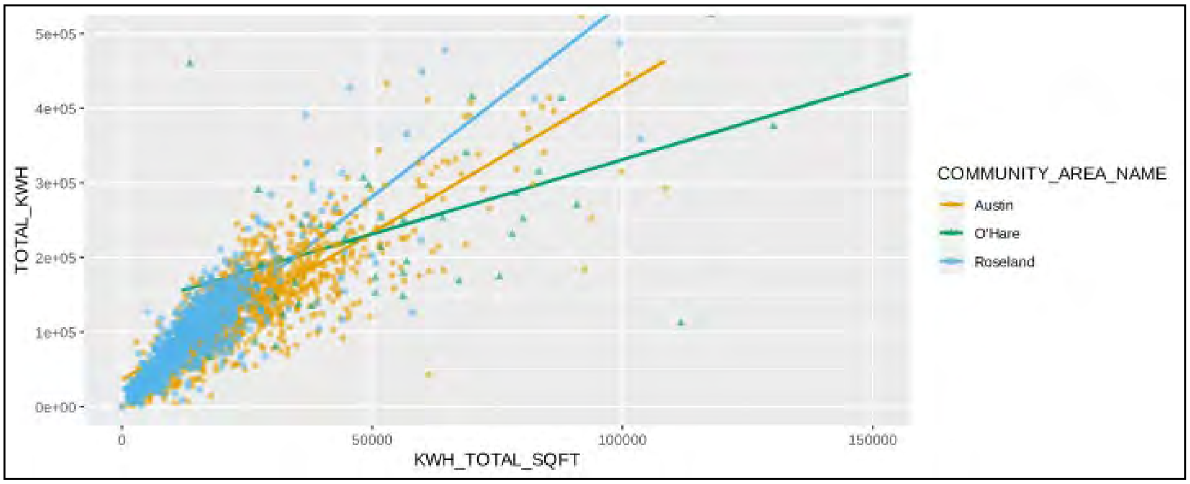
Data Dictionary

Variable	Data Type / Range / Example	Description
COMMUNITY_AREA_NAME	Character: Roseland, O'Hare, Austin, etc.	Name of the neighborhood in Chicago. Each neighborhood comprises many census blocks.
BLOCK_HOUSING_OCCUPIED_UNITS	Numeric	Number of occupied housing units on the block
KWH_TOTAL_SQFT	Numeric: 600 - 5453937	Total square footage for a census block and building type
BUILDING_AGE or AVG_BLDG_AGE	Numeric: 0 - 158	Average age of the buildings in a census block
STORIES	Numeric: 1 - 110	Average number of stories for buildings in a census block
BUILDING_TYPE	Character: Commercial, Residential, Industrial	The type of building based on what it is used for, possible values are Commercial, Residential, and Industrial
AVERAGE_HOUSEHOLD_SIZE	Numeric: 1.28-4.37	Average household size. Measured at the census tract level.
OVER_AGE_65	Numeric: 0.23-0.45	Proportion of the population in the census tract over age 65.
GAS_ACCOUNT	Numeric: 1 - 383	The number of different gas accounts associated with a census block
ELECTRICITY_ACCOUNTS	Numeric: 1 - 1904	The number of different electricity account associated with a census block
KWH_JAN, KWH_FEB, …	Numeric: Varies	Total KWH usage for that census block and building type for a month. 12 variables, one for each month. KWH is a unit of measurement for electricity.
TOTAL_KWH	Numeric: 312 - 72,070,053	Total KWH usage for that census block and building type for a year.
THERM_JAN, THERM_FEB, …	Numeric: Varies	Total therm usage for that census block and building type for a month. 12 variables, one for each month. A therm is a unit of measurement for natural gas.
TOTAL_THERM	Numeric: 28 - 1,940,742	Total therm usage for that census block and building type for a year.
THERMS_PER_SQFT	Numeric	Total therm usage divided by total square footage for a census block and building type.

KWH_PER_SQFT	Numeric	Total KWH usage divided by total square footage for a census block and building type.
THERMS_PER_ACCOUNT	Numeric: 0 - 43,222	Total therm usage divided by the number of gas accounts.
TMAX_FAHRENHEIT	Numeric: 12 - 95	Max temperature recorded during the day in degrees Fahrenheit.
TMIN_FAHRENHEIT	Numeric: 0 - 79	Min temperature recorded during the day in degrees Fahrenheit.
PRECIPITATION_INCHES	Numeric: 0 - 4.77	Inches of precipitation during the day.
SNOW_FALL_INCHES	Numeric: 0 - 8.8	New snowfall during the day.
SNOW_DEPTH_INCHES	Numeric: 0 - 9	Snow depth reported as 7 am each day.

Task 1 – (6 *points*)

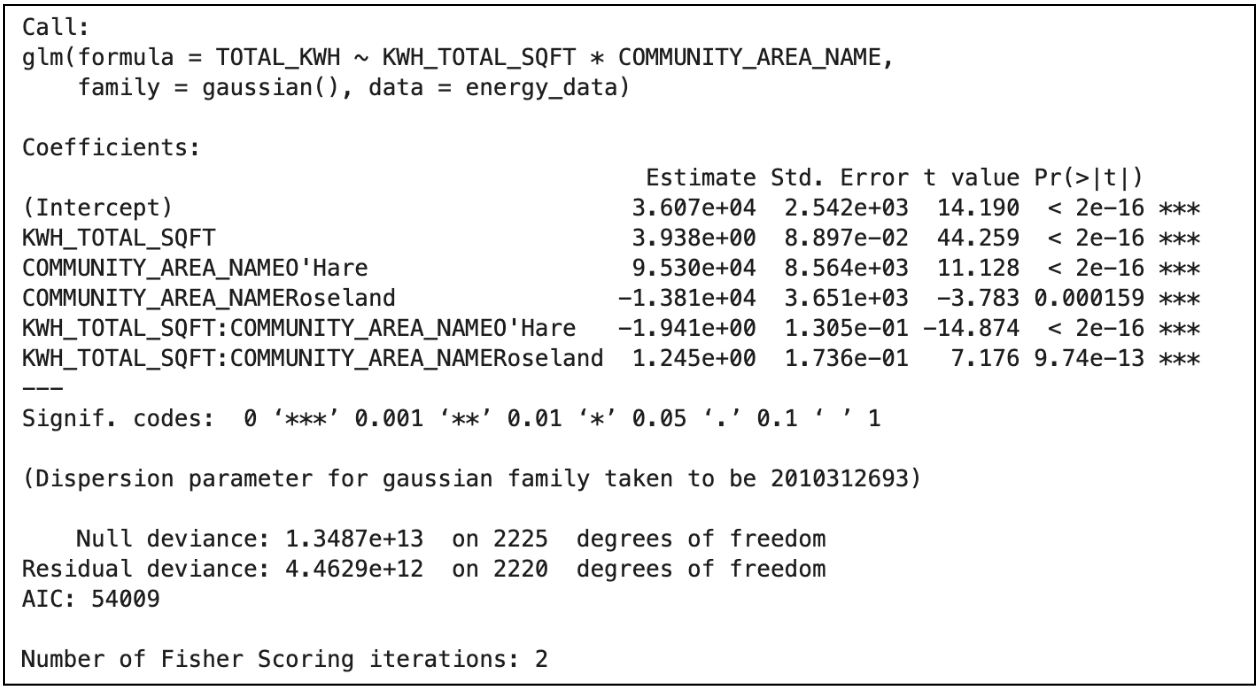
Your client wants to understand the relationship between electricity consumption and total square footage in residential buildings from different community areas, in particular Austin, Roseland and O'Hare. Your assistant created the graph below filtering only residential building energy use across the three community areas. The lines represent ordinary linear fit within each community area. Your manager suggests that adding an interaction between COMMUNITY_AREA_NAME and KWH_TOTAL_SQFT is necessary to capture the slope differences.



(a) (2 points) Assess your manager's suggestion that adding this interaction can capture the slope differences.

ANSWER:

Your assistant built a GLM with interaction and provided you with the model summary.



```
Call:
glm(formula = TOTAL_KWH ~ KWH_TOTAL_SQFT * COMMUNITY_AREA_NAME, 
    family = gaussian(), data = energy_data)

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                3.607e+04  2.542e+03  14.190  < 2e-16 ***
KWH_TOTAL_SQFT                             3.938e+00  8.897e-02  44.259  < 2e-16 ***
COMMUNITY_AREA_NAMEO'Hare                  9.530e+04  8.564e+03  11.128  < 2e-16 ***
COMMUNITY_AREA_NAMERoseland               -1.381e+04  3.651e+03  -3.783 0.000159 ***
KWH_TOTAL_SQFT:COMMUNITY_AREA_NAMEO'Hare  -1.941e+00  1.305e-01 -14.874  < 2e-16 ***
KWH_TOTAL_SQFT:COMMUNITY_AREA_NAMERoseland  1.245e+00  1.736e-01   7.176 9.74e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2010312693)

    Null deviance: 1.3487e+13  on 2225  degrees of freedom
Residual deviance: 4.4629e+12  on 2220  degrees of freedom
AIC: 54009

Number of Fisher Scoring iterations: 2
```

(b) (2 points) Calculate the impact of a 10 square foot increase on total energy usage for each community based on the model above:

a. O'Hare
b. Roseland
c. Austin

Show your work.

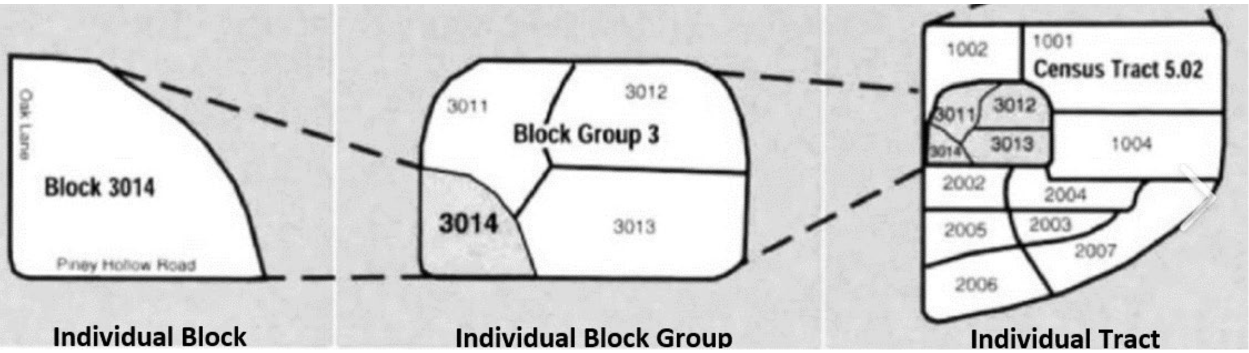
ANSWER:

(c) (2 points) Recommend and justify one additional variable to include in the modeling that could help explain the variation in electricity consumption per square foot across different communities.

ANSWER:

Task 2 – (6 *points*)

The relationship between a census block and census tract is illustrated below. Census blocks are a geographical subset of a tract.



Your manager wants you to predict residential energy usage in a block for the entire year. Some of the predictor variables are stored at a more aggregate granularity than the block level (e.g., tract and city level detail). Given this, your manager asks you to remove all tract level and city level variables as they will not provide predictive power.

(a) (*2 points*) Critique your manager's recommendation to remove city-level data.

ANSWER:

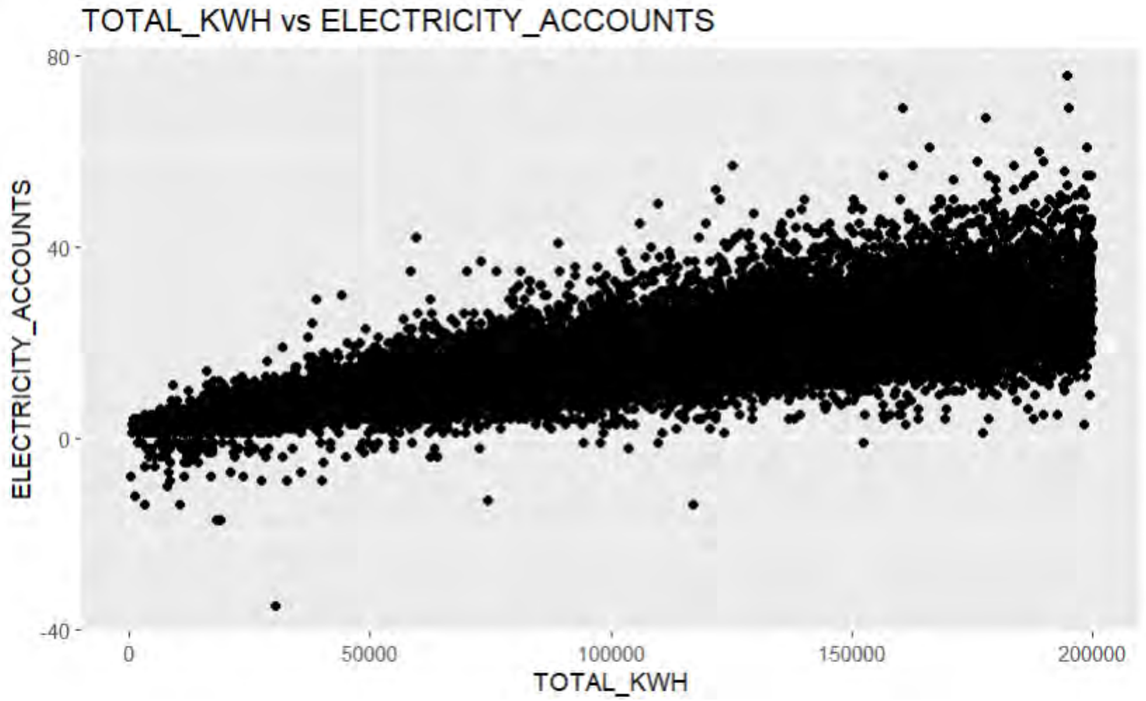
(b) (2 *points*) Critique your manager's recommendation to remove tract level data.

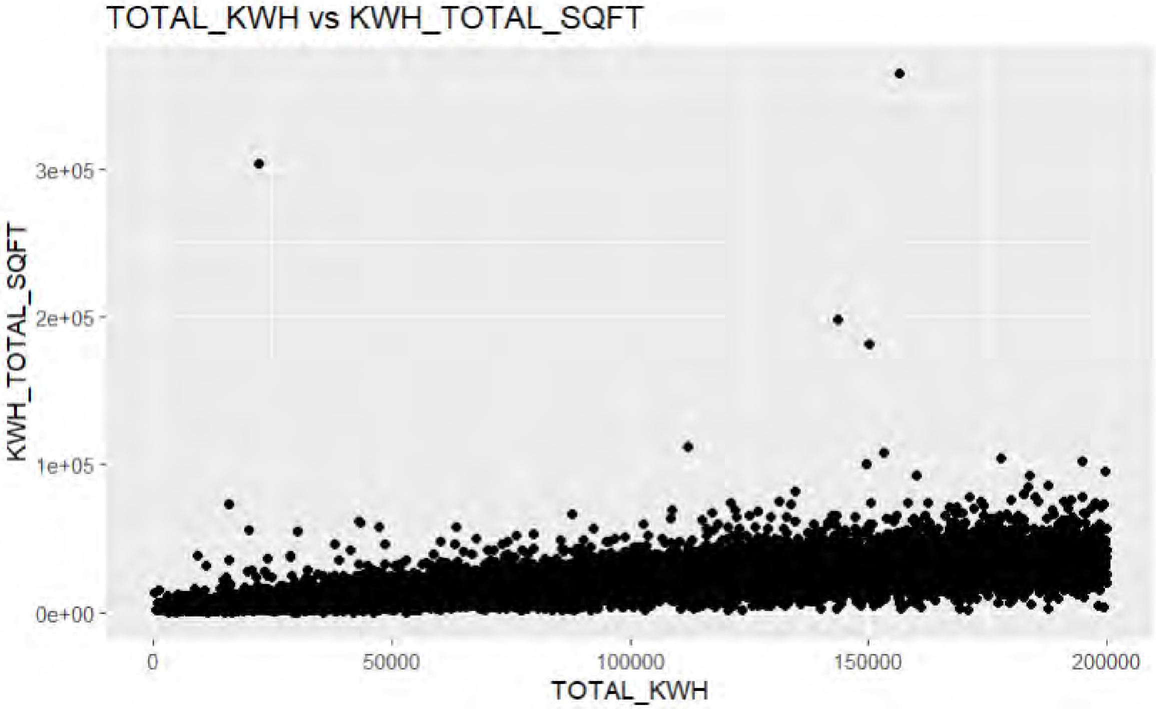
ANSWER:

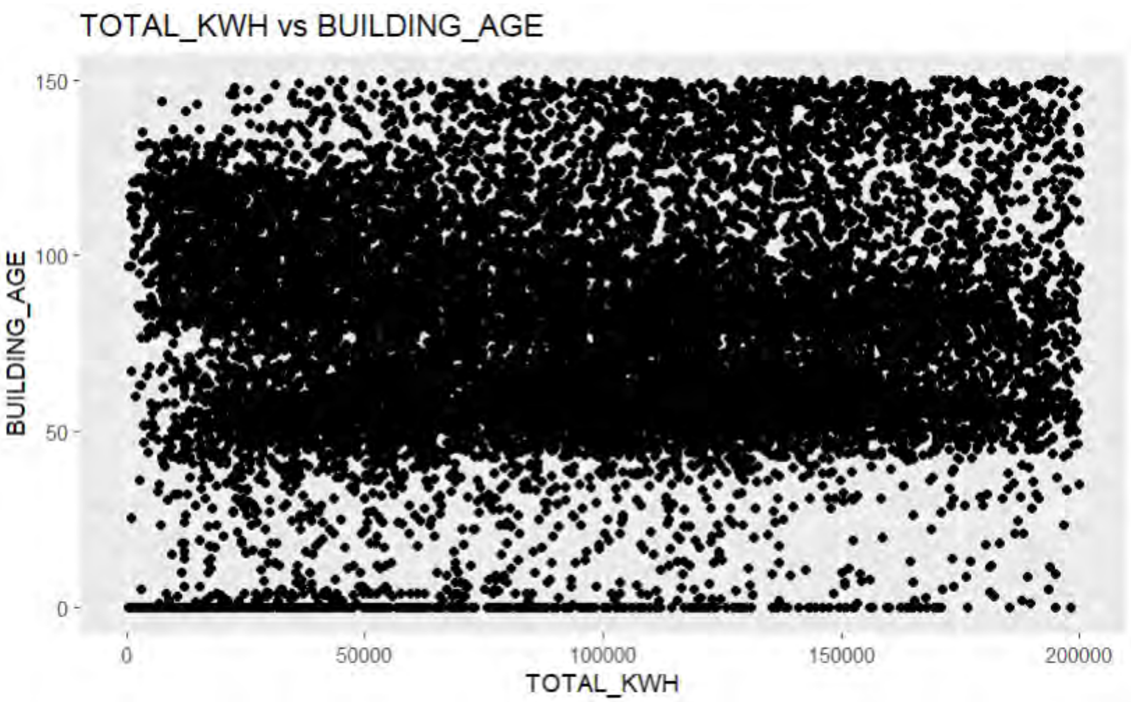
Your assistant wants to model the target variable **TOTAL_KHW** using the formula:

TOTAL_KWH ~ ELECTRICITY_ACCOUNTS + KWH_TOTAL_SQFT + BUILDING_AGE

They have also provided thee charts:


TOTAL_KWH vs ELECTRICITY_ACCOUNTS
ELECTRICITY_ACCOUNTS
80
40
0
-40
0
50000
100000
150000
200000
TOTAL_KWH


TOTAL_KWH vs KWH_TOTAL_SQFT
KWH_TOTAL_SQFT
3e+05
2e+05
1e+05
0e+00
0
50000
100000
150000
200000
TOTAL_KWH



(c) (*2 points*) Recommend whether a tree-based model or a generalized linear model would be a better choice for predicting the target variable. Justify your recommendation.

ANSWER:

Task 3 – (8 *points*)

Your manager would like to identify and predict the highest energy usage blocks. Your assistant created an initial classification model that classifies values as True if they exceed a certain threshold or False otherwise. The output of the model's confusion matrix is shown below.

		ACTUAL	
		False	True
PREDICTION	True	4	660
	False	21493	616

(a) (4 *points*) Calculate the following values from the confusion matrix, and interpret the significance of the results.

a. Accuracy
b. Sensitivity
c. Precision

ANSWER:

Your manager is interested in a marketing promotion targeting high energy utilization homes. Your manager's goal is to identify more of the high utilization homes in the prediction model even at the cost of misclassifying some low energy utilization homes as high energy utilization.

(b) (*2 points*) Recommend how to change the cutoff value of the model to achieve the desired objective. Explain the directional impact of the change.

ANSWER:

Your manager observes that the data set used to train the model is unbalanced.

(c) (*2 points*) Recommend and explain one method to improve model performance by creating a more balanced data set.

ANSWER:

Task 4 – (5 *points*)

Your manager would like you to build three tree-based models and tune their hyperparameters. The three models will be a decision tree, random forest, and boosted tree.

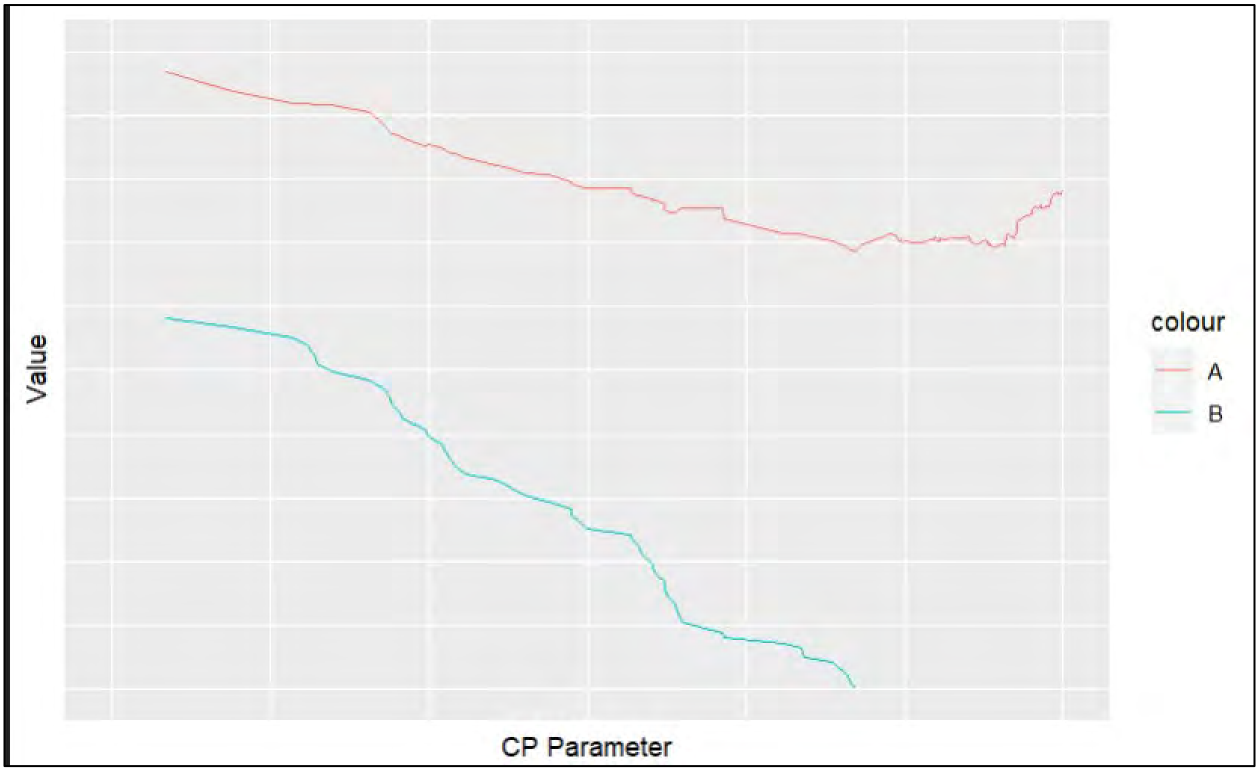
(a) (1 point) List one common hyperparameter that could be tuned for all three models.

ANSWER:

(b) (2 points) Describe a unique parameter for tuning a random forest model and a unique parameter for tuning a boosted tree.

ANSWER:

Your assistant creates a single decision tree and is trying to tune the hyperparameters, specifically the Complexity Parameter (cp). Creating a CP Table, they show the cp values decreasing across the x-axis from left to right and the corresponding values of the error terms for "xerror" and "relerror."



(c) (*2 points*) Identify which line corresponds to the xerror (cross-validation error) versus the relerror (relative error) term and provide a rationale. Explain the behavior of each.

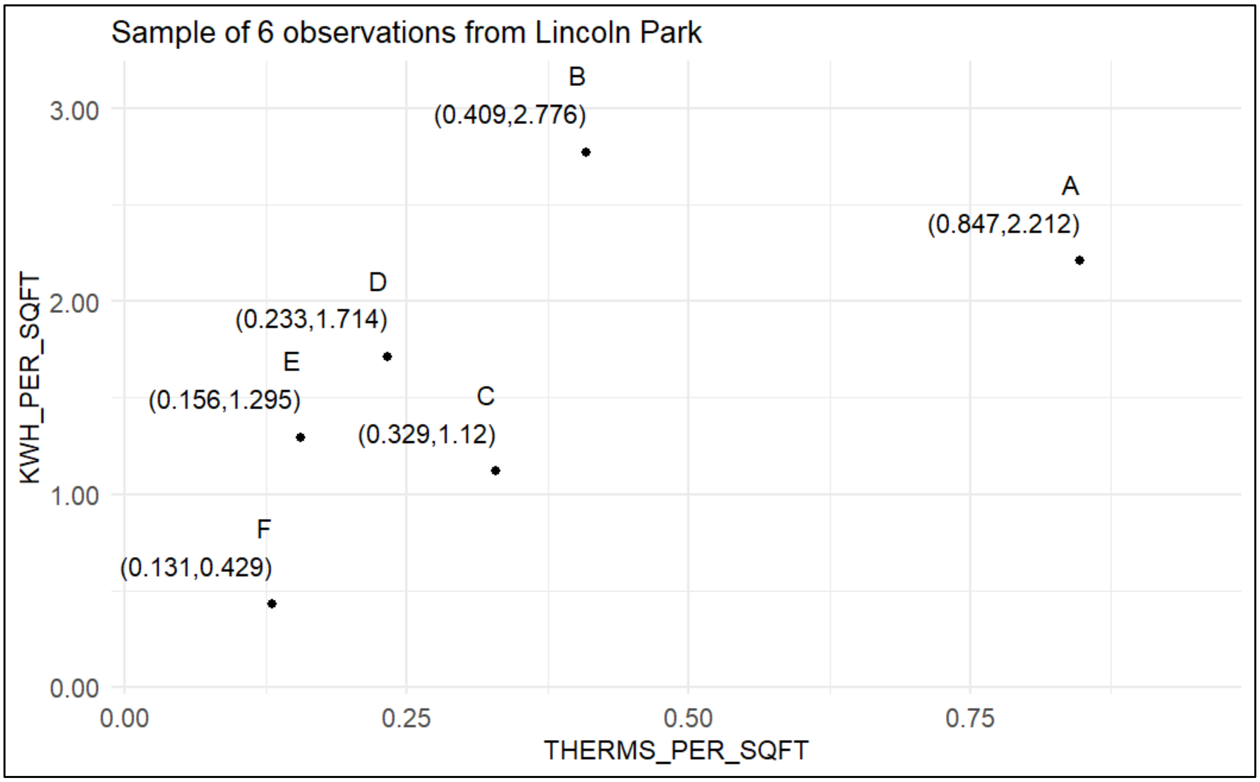
ANSWER:

Task 5 – (7 *points*)

Note: The data for this task is provided in the available Excel file on the "5" tab. You may use it for your calculations, but if the file is uploaded it will not be looked at by the graders. All your work must be shown in this Word document.

Your manager wants to better understand how hierarchical clustering works and asks you to create a dendrogram using the single linkage method on a subset of the energy data. Your assistant selects two variables from the dataset: THERMS_PER_SQFT and KWH_PER_SQFT, scales them, and chooses six scaled observations from the Lincoln Park community area.

	THERMS_ PER_SQFT	KWH_ PER_SQFT
A	0.847	2.212
B	0.409	2.776
C	0.329	1.120
D	0.233	1.714
E	0.156	1.295
F	0.131	0.429



(a) (2 points) Complete the distance matrix below by calculating the Euclidian distances between the missing pairs of observations and enter them into the table below. Round to two decimal places.

ANSWER:

	A	B	C	D	E	F
A	0.00	X	X	X	X	X
B		0.00	X	X	X	X
C	1.21	1.66	0.00	X	X	X
D		1.08	0.60	0.00	X	X
E	1.15	1.50	0.25	0.43	0.00	X
F	1.92	2.36		1.29	0.87	0.00

Based on the distance matrix, the first cluster formed is with observations C and E.

(b) (1 point) Complete the updated distance matrix using single linkage.

ANSWER:

	A	B	C,E	D	F
A	0.00	X	X	X	X
B	0.71	0.00	X	X	X
C,E			0.00	X	X
D	0.79	1.08	0.43	0.00	X
F	1.92	2.36	0.72	1.29	0.00

(c) (2 points) Complete the distance matrices using the tables below to provide the information needed to construct the dendrogram.

ANSWER:

	0.00	X	x	x
		0.00	X	x
			0.00	X
				0.00

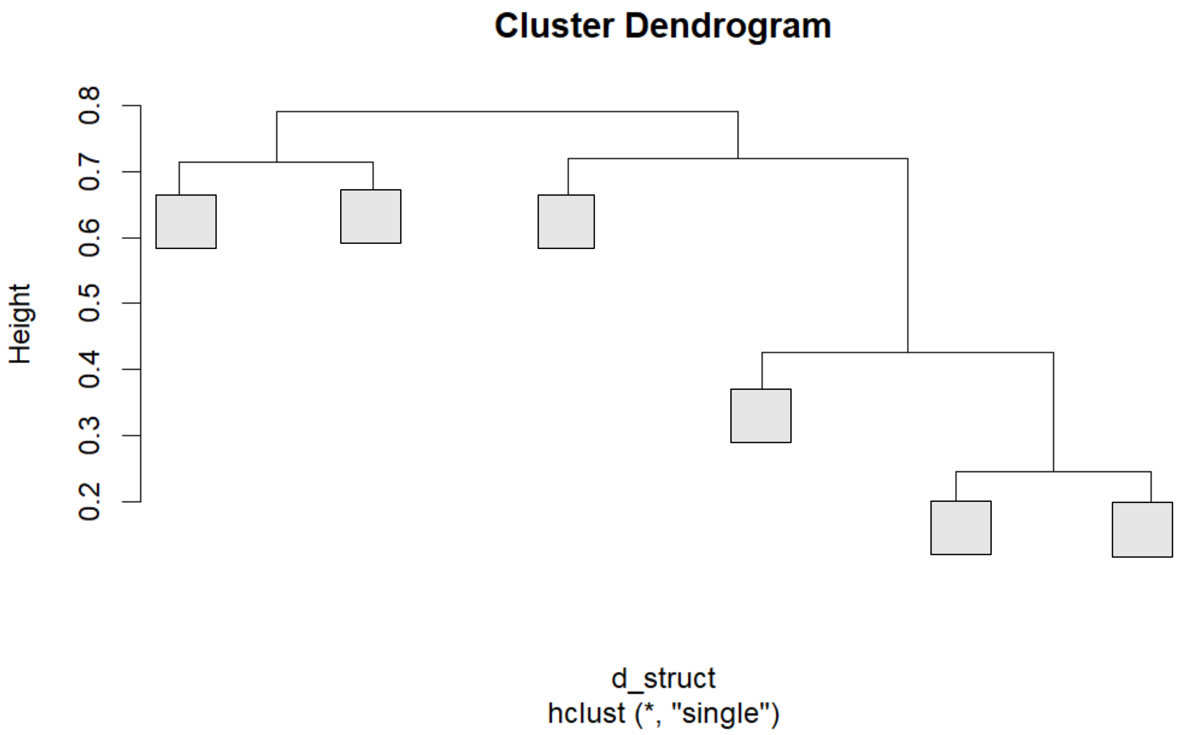
	0.00	X	x
		0.00	X
			0.00

	0.00	X
		0.00

(d) (2 points) Complete the dendrogram below by labeling the observations.

[Note: please click into each node to label it or type the node labels from left to right.]

ANSWER:



Task 6 – (12 *points*)

(a) (2 *points*) Describe the differences between using weights and offsets in an ordinary least squares model.

ANSWER:

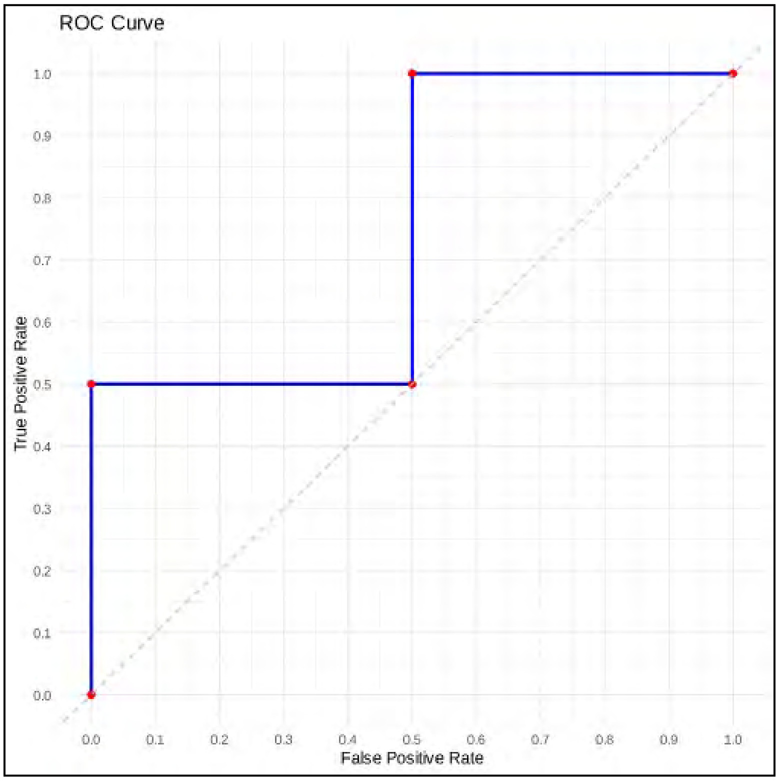
(b) (3 points) Explain the differences between using a variable as a weight versus a predictor variable in the context of a generalized linear model (GLM).

ANSWER:

(c) (2 points) Compare and contrast ROC and AUC in the context of model performance evaluation.

ANSWER:

You are provided with an ROC curve below:



(d) (1 points) Calculate AUC. Show your work.

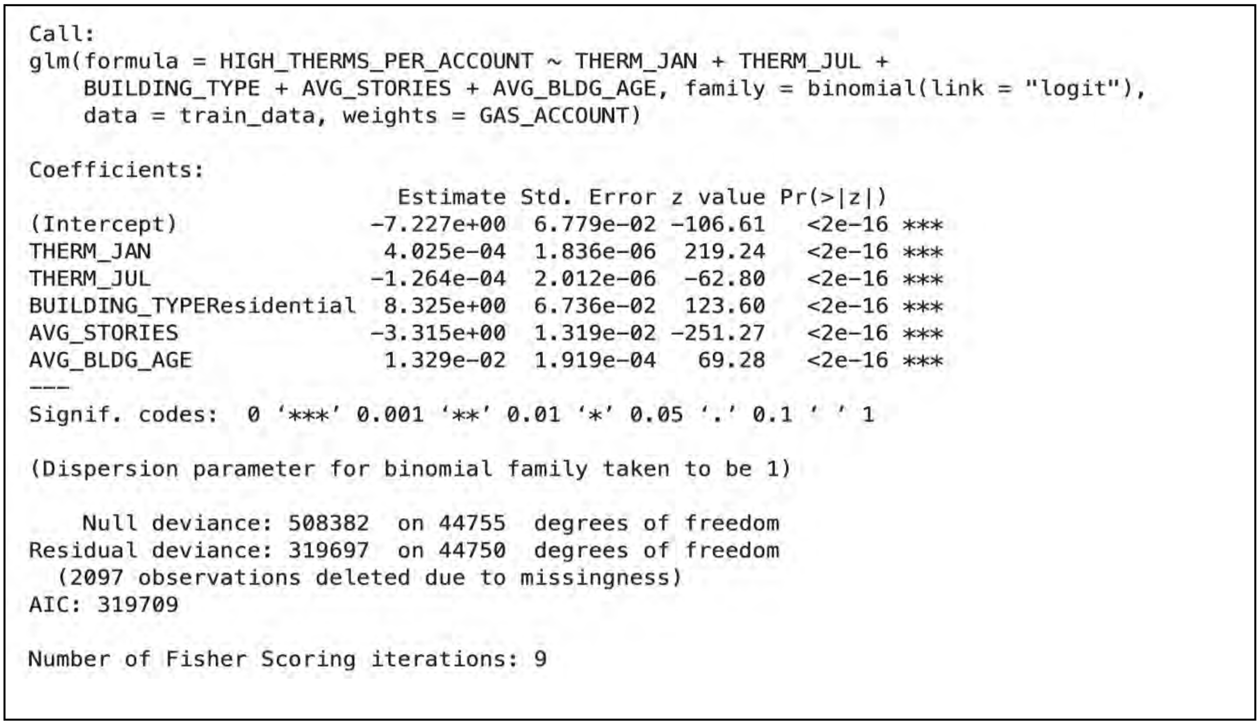
ANSWER:

Your client wants to understand the factors influencing per account natural gas usage (THERMS_PER_ACCOUNT). Your assistant prepares the data on total natural gas usage (TOTAL_THERMS), the number of gas accounts (GAS_ACCOUNT), and other variables such as natural gas usage in January (THERM_JAN), natural gas usage in July (THERM_JUL), and building type (BUILDING_TYPE). Your manager suggests building a logistic regression model to identify high energy users, and use ROC and AUC as evaluation metrics.

You consider using GAS_ACCOUNT as a weight variable in a generalized linear model (GLM) to better understand its impact on high natural gas usage.

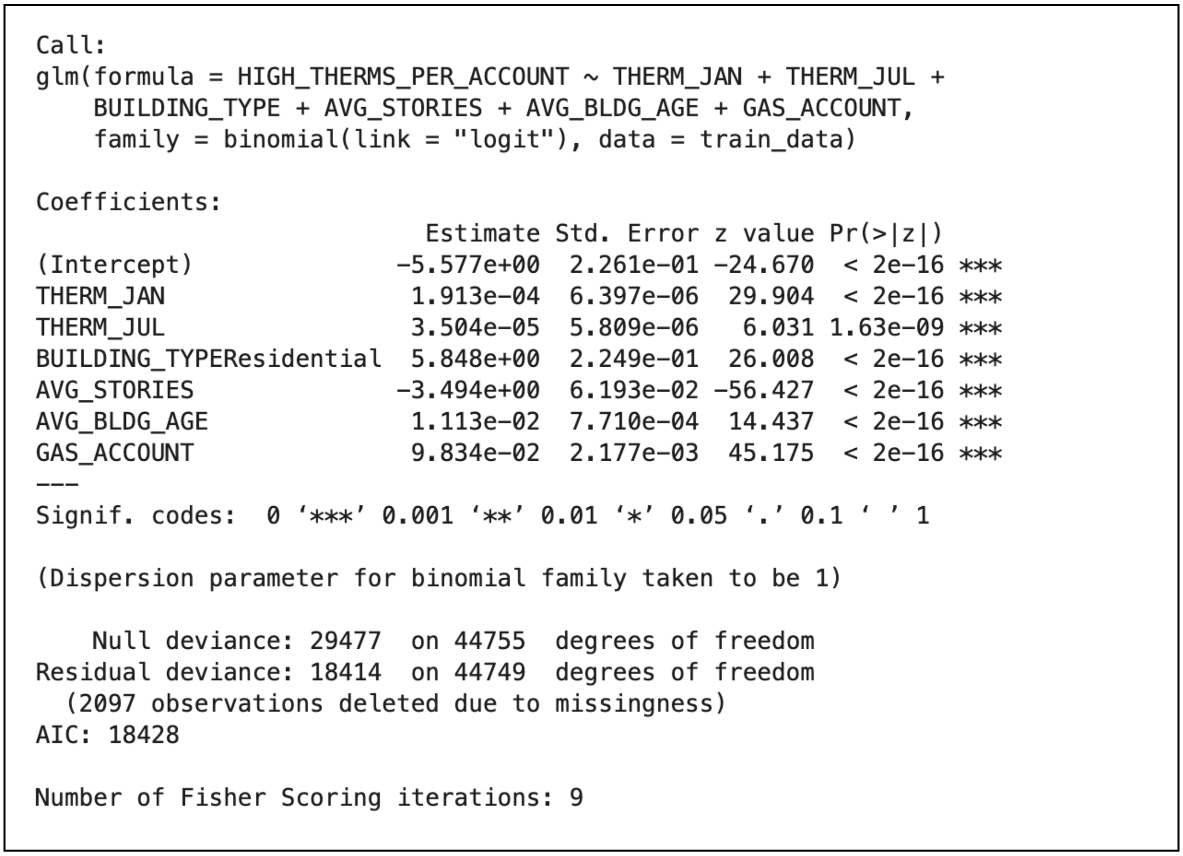
Your assistant creates a binary variable HIGH_THERMS_PER_ACCOUNT to identify high natural gas accounts, and builds two GLMs, one model uses GAS_ACCOUNT as a weight, but not as a variable, and the other is unweighted and includes GAS_ACCOUNT as a variable. You are provided with model summaries and ROC curves for the 2 models.

Model 1:

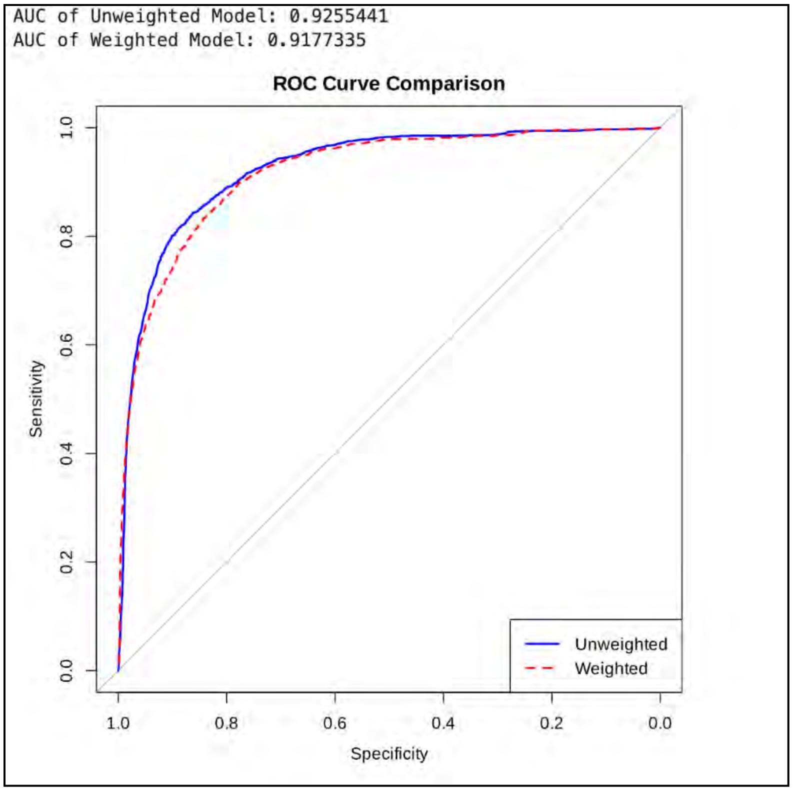


```
Call:
glm(formula = HIGH_THERMS_PER_ACCOUNT ~ THERM_JAN + THERM_JUL +
    BUILDING_TYPE + AVG_STORIES + AVG_BLDG_AGE, family = binomial(link = "logit"),
    data = train_data, weights = GAS_ACCOUNT)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -7.227e+00  6.779e-02 -106.61   <2e-16 ***
THERM_JAN                 4.025e-04  1.836e-06  219.24   <2e-16 ***
THERM_JUL                -1.264e-04  2.012e-06  -62.80   <2e-16 ***
BUILDING_TYPEResidential  8.325e+00  6.736e-02  123.60   <2e-16 ***
AVG_STORIES              -3.315e+00  1.319e-02 -251.27   <2e-16 ***
AVG_BLDG_AGE              1.329e-02  1.919e-04   69.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 508382  on 44755  degrees of freedom
Residual deviance: 319697  on 44750  degrees of freedom
  (2097 observations deleted due to missingness)
AIC: 319709

Number of Fisher Scoring iterations: 9
```

Model 2:



```
Call:
glm(formula = HIGH_THERMS_PER_ACCOUNT ~ THERM_JAN + THERM_JUL +
    BUILDING_TYPE + AVG_STORIES + AVG_BLDG_AGE + GAS_ACCOUNT,
    family = binomial(link = "logit"), data = train_data)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -5.577e+00  2.261e-01 -24.670  < 2e-16 ***
THERM_JAN                 1.913e-04  6.397e-06  29.904  < 2e-16 ***
THERM_JUL                 3.504e-05  5.809e-06   6.031 1.63e-09 ***
BUILDING_TYPEResidential  5.848e+00  2.249e-01  26.008  < 2e-16 ***
AVG_STORIES              -3.494e+00  6.193e-02 -56.427  < 2e-16 ***
AVG_BLDG_AGE              1.113e-02  7.710e-04  14.437  < 2e-16 ***
GAS_ACCOUNT               9.834e-02  2.177e-03  45.175  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29477  on 44755  degrees of freedom
Residual deviance: 18414  on 44749  degrees of freedom
  (2097 observations deleted due to missingness)
AIC: 18428

Number of Fisher Scoring iterations: 9
```



```
AUC of Unweighted Model: 0.9255441
AUC of Weighted Model: 0.9177335
```

(e) (3 points) Interpret the model results in the context of AIC and AUC for each model.

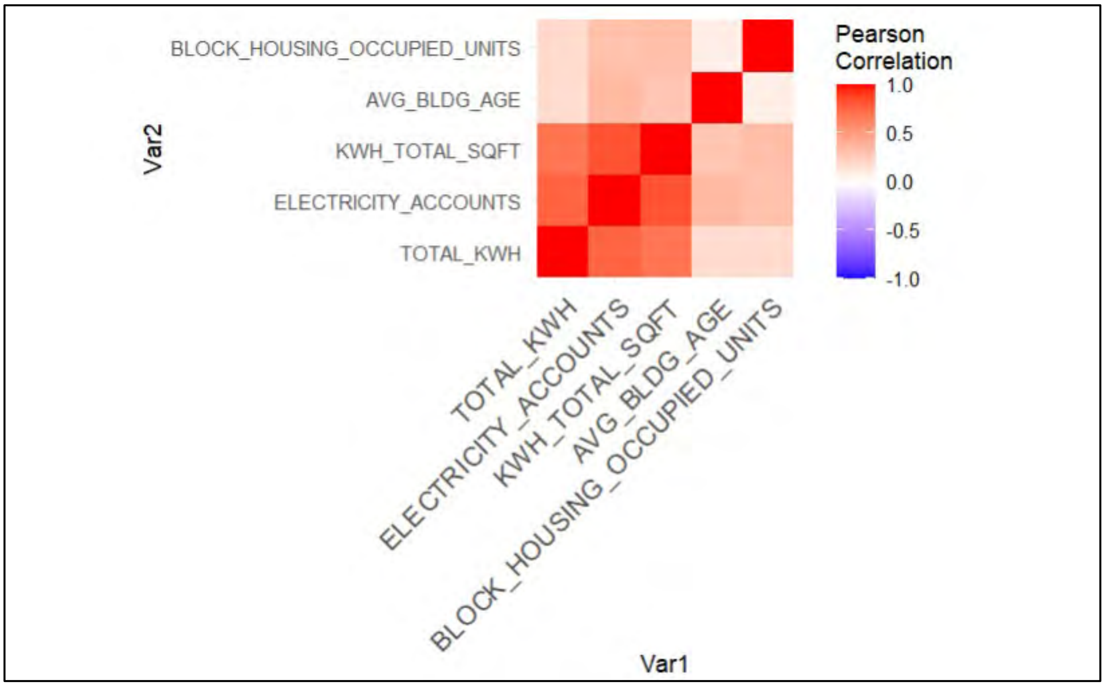
ANSWER:

(f) (1 points) Recommend a model to your client. Justify your answer.

ANSWER:

Task 7 – (3 *points*)

Your assistant is modeling the TOTAL_KWH of a block as the target variable. They want to examine the relationships between the variables being considered and has built a correlation heat map below.



(a) (*2 points*) Interpret the graphic and identify which independent variables exhibit collinearity.

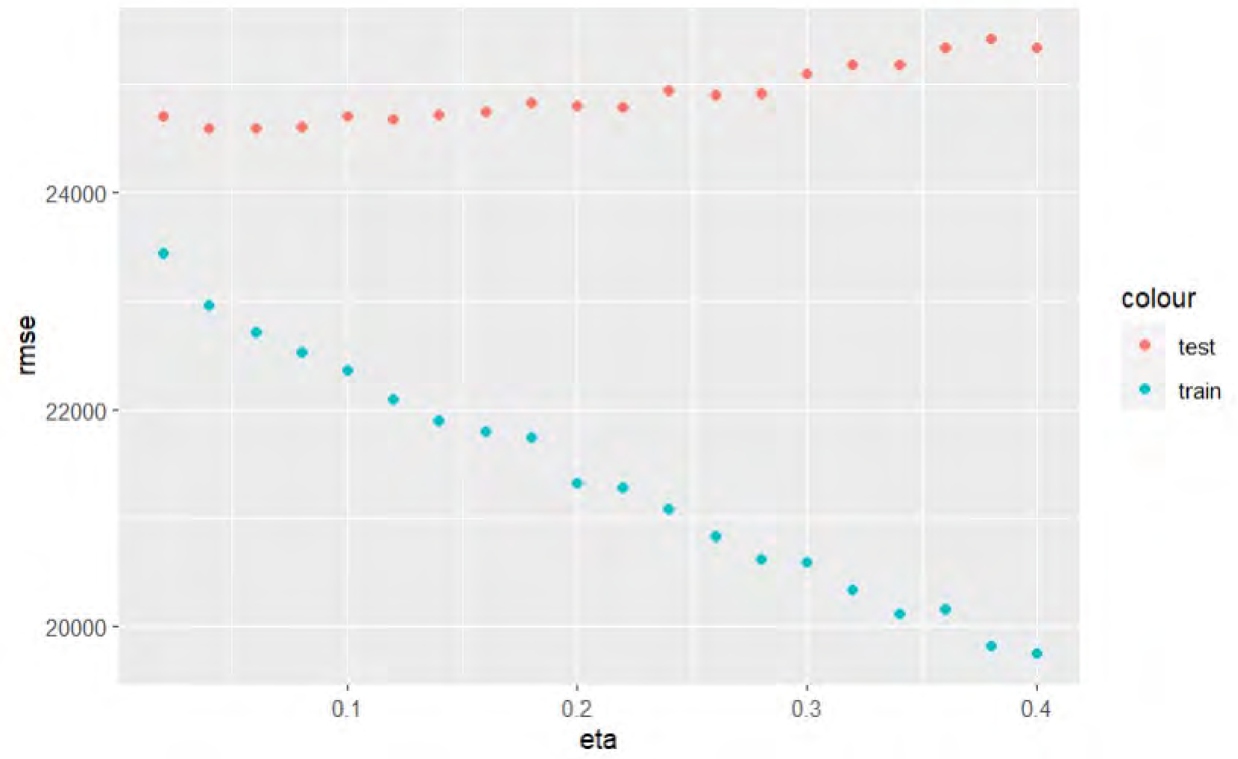
ANSWER:

(b) (*1 point*) Recommend and justify one enhancement to improve the chart.

ANSWER:

Task 8 – (4 *points*)

Your assistant decides to build a boosted tree and is tuning the model hyperparameter for an optimal learning rate e. Your assistant has divided the data into a training set with 80% of the data and a testing set with 20% of the data. They notice the model evaluation metric (RMSE) on the train set (without cross validation) and test set behave differently when changing the learning rate.



(a) (*2 points*) Explain why increasing the learning rate would result in such a difference in performance on the train vs. test set for a boosted tree.

ANSWER:

Your assistant would like to set a value for the learning rate based on the results from the test set in the chart above.

(b) (*2 points*) Critique the assistant's proposed method of hyperparameter tuning. Recommend and justify an alternative approach.

ANSWER:

Task 9 – (8 *points*)

Your manager is working on a project for the city of Chicago to measure the impact of weather and climate on energy use. Your manager asks your assistant to prepare a graphic overview of monthly weather patterns.

Your assistant isn't sure how best to aggregate the weather variables, which are captured on a daily basis, into monthly variables. The description of each of the weather variables is copied below from the data dictionary.

- TMAX_FAHRENHEIT: Max temperature recorded during the day.
- TMIN_FAHRENHEIT: Min temperature recorded during the day.
- PRECIPITATION_INCHES: Inches of precipitation during the day.
- SNOW_FALL_INCHES: New snowfall during the day.
- SNOW_DEPTH_INCHES: Snow depth reported at 7 am each day.

Your manager wants to be able to use these monthly weather variables in linear models and for the interpretation of the variables to be intuitive.

(a) (*2 points*) Recommend whether each of the following weather variables should be aggregated based on taking the average of the daily values or the sum of the daily values; briefly justify your recommendation.

ANSWER:

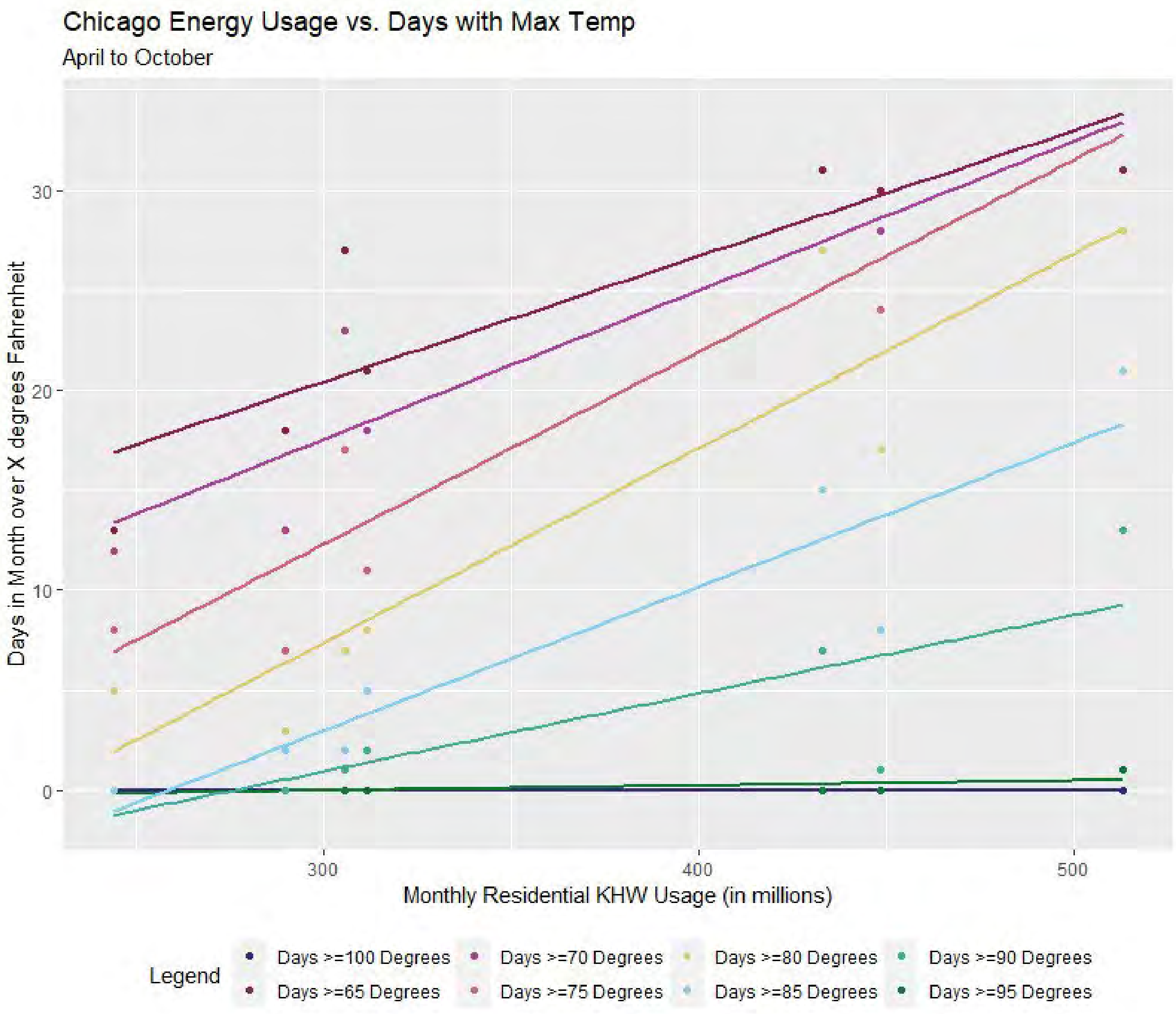
TMAX_FAHRENHEIT
Recommended Monthly Aggregation:

Justification:

SNOW_DEPTH_INCHES
Recommended Monthly Aggregation:

Justification:

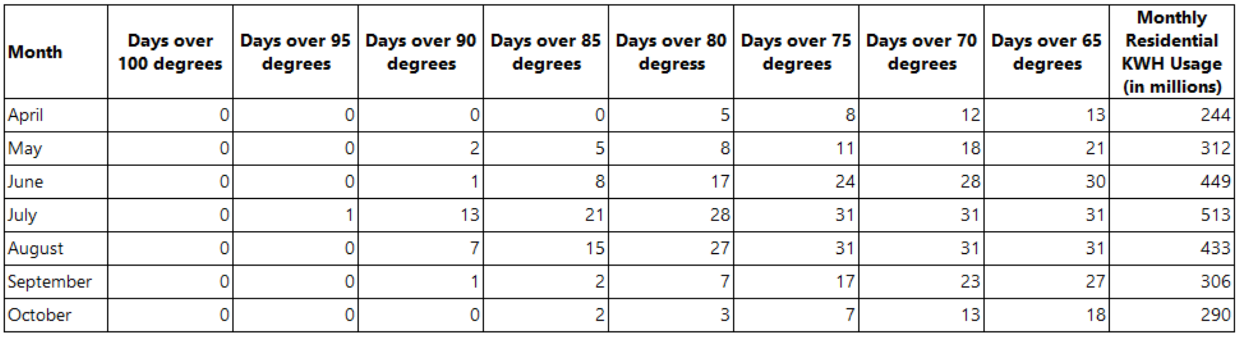
Your assistant observes that residential energy usage in terms of kilowatt hours peaks in April-October and hypothesizes that this is due to air conditioning usage and higher temperatures. Your manager agrees, but isn't sure whether the relationship is due to the average temperature during the month or the number of days in a month that the temperature exceeds a specific level. Your assistant graphs the number of days that the maximum temperature exceeds different levels in each month against monthly residential Kilowatt hours for April through October.



(b) (*2 points*) Briefly summarize the relationship between monthly residential KWH usage and days in a month exceeding a specific temperature based on the graph above.

ANSWER:

Your assistant also provides the table of values for the chart above.



Month	Days over 100 degrees	Days over 95 degrees	Days over 90 degrees	Days over 85 degrees	Days over 80 degress	Days over 75 degrees	Days over 70 degrees	Days over 65 degrees	Monthly Residential KWH Usage (in millions)
April	0	0	0	0	5	8	12	13	244
May	0	0	2	5	8	11	18	21	312
June	0	0	1	8	17	24	28	30	449
July	0	1	13	21	28	31	31	31	513
August	0	0	7	15	27	31	31	31	433
September	0	0	1	2	7	17	23	27	306
October	0	0	0	2	3	7	13	18	290

(c) (*2 points*) Recommend and justify one improvement in your assistant's analytical approach and one improvement in your assistant's graph design. Your answer should be based on the graph and table provided.

ANSWER:

Your manager is interested in a deeper understanding of the impact of weather and climate on energy usage in Chicago. You are concerned that the current data isn't sufficient to support some of your manager's questions. For each question below:

(d) (*2 points*) Explain whether or not the analysis can be supported by the current data. If the data is available, state which variables you would use. If the data is not sufficient, state the additional data you would need to collect.

ANSWER:

Which decade of residential building age is most efficient in terms of energy use during June, July, and August?

Do you have sufficient data to perform the analysis?

If so, what variables would you use? If not, what additional data would you need to request?

Do communities in Chicago nearer to Lake Michigan experience lower average temperatures and correspondingly require less energy during June, July, and August?

Do you have sufficient data to perform the analysis?

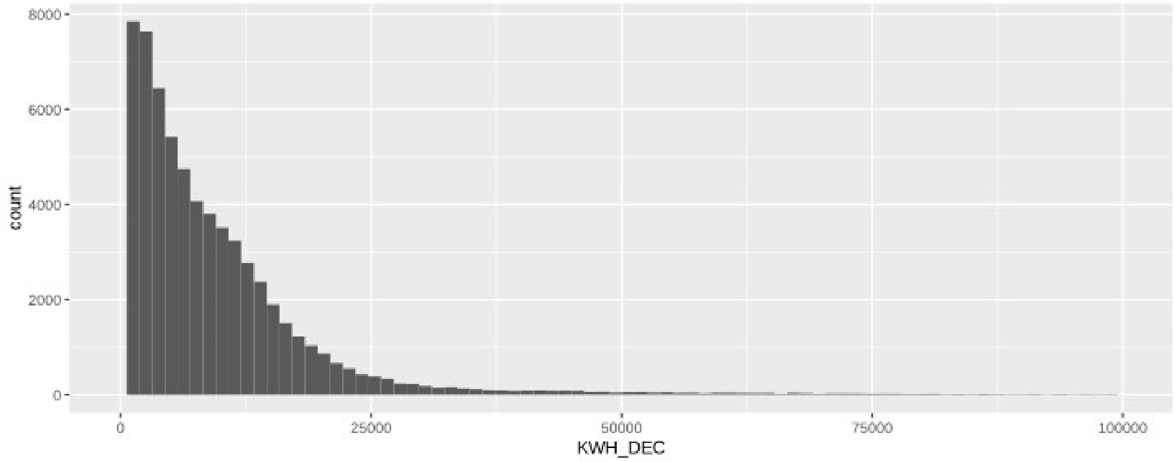
If so, what variables would you use? If not, what additional data would you need to request?

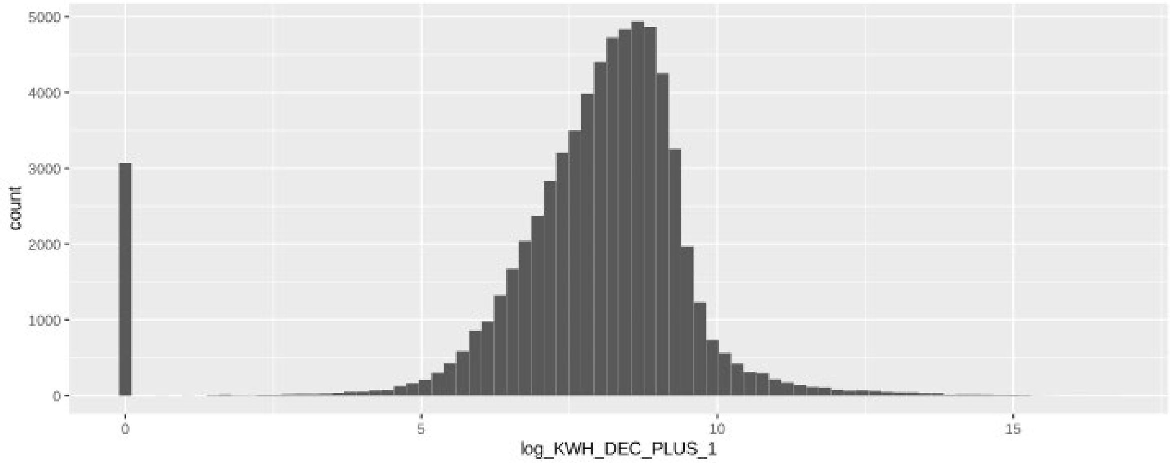
Task 10 – (11 *points*)

(a) (2 point) Describe the key assumptions of the generalized linear model (GLM).

ANSWER:

Your client wants to use the previous months' electricity usage to predict December's utilization by building type. Your manager suggests making transformations of the KWH variables. Your assistant took December utilization (KWH_DEC) and made a histogram without transformation, and with log(KWH_DEC+1) transformation. Your assistant suggests to use the log(KWH_DEC+1) transformation.

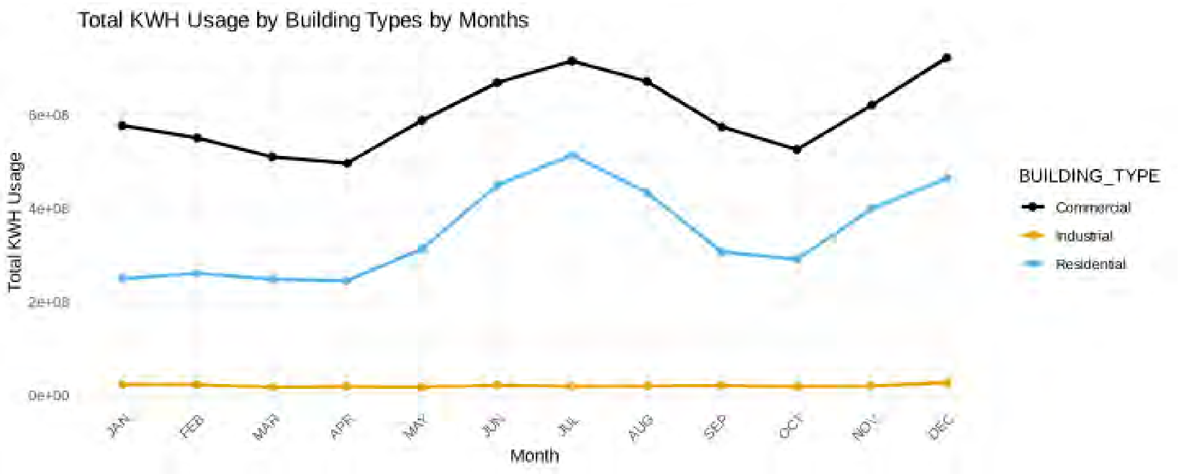




(b) (3 points) Explain the pros and cons of your assistant's suggested variable transformation.

ANSWER:

Your assistant provides you with the total KWH usage by building type by month plot below and points out the cyclic pattern for commercial and residential building types. Your assistant suggests applying seasonality to the entire dataset.



(c) (2 points) Justify your assistant's suggestion and recommend a way to model this cyclic pattern.

ANSWER:

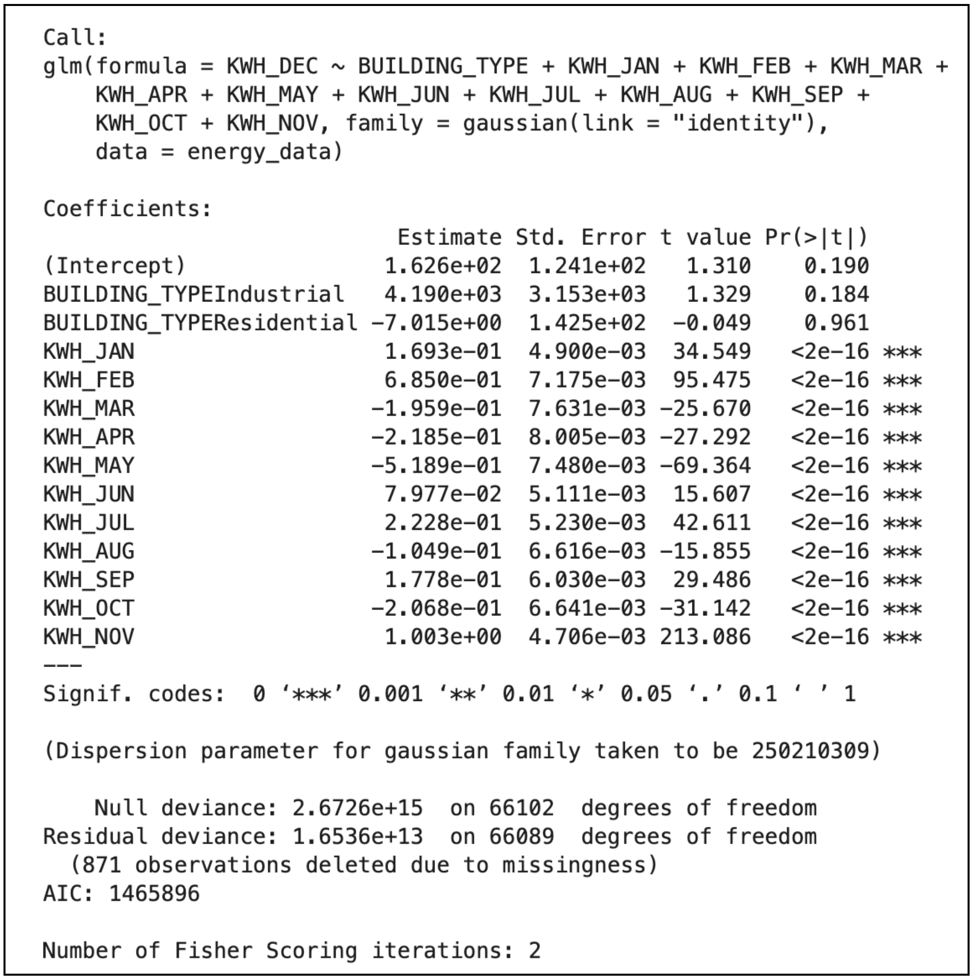
Your assistant has built two models to predict December electricity usage using months and BUILDING_TYPE variables.

Model 1: Each row corresponds to an observation for a specific year. The columns include variables for different months (e.g., KWH_JAN, KWH_FEB, ..., KWH_NOV).

Model 2: Each row corresponds to an observation at a specific time point (e.g., month). Separate columns are used for the Month variable and the electricity usage (e.g., KWH).

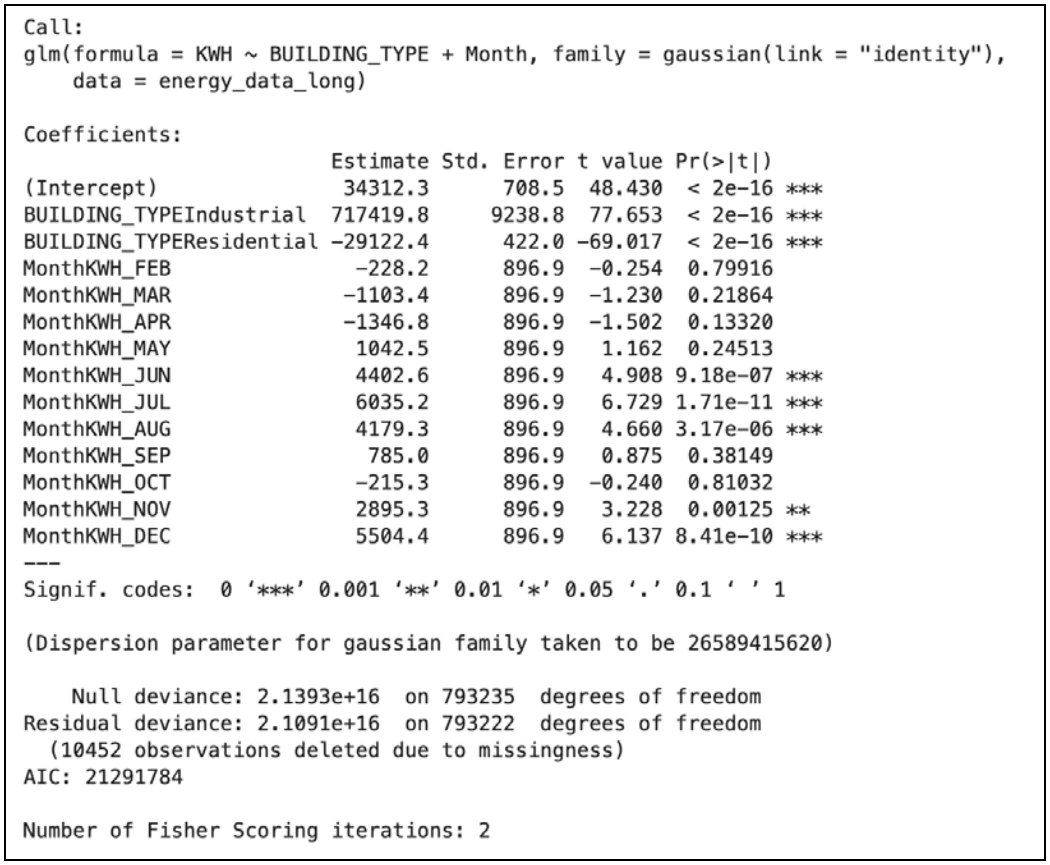
You are provided with the model summary output. Your assistant also pointed out that BUILDING_TYPE is statistically significant in Model 2 but not Model 1.

Model 1:



```
Call:
glm(formula = KWH_DEC ~ BUILDING_TYPE + KWH_JAN + KWH_FEB + KWH_MAR +
    KWH_APR + KWH_MAY + KWH_JUN + KWH_JUL + KWH_AUG + KWH_SEP +
    KWH_OCT + KWH_NOV, family = gaussian(link = "identity"),
    data = energy_data)

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.626e+02  1.241e+02   1.310    0.190
BUILDING_TYPEIndustrial   4.190e+03  3.153e+03   1.329    0.184
BUILDING_TYPEResidential -7.015e+00  1.425e+02  -0.049    0.961
KWH_JAN                   1.693e-01  4.900e-03  34.549   <2e-16 ***
KWH_FEB                   6.850e-01  7.175e-03  95.475   <2e-16 ***
KWH_MAR                  -1.959e-01  7.631e-03 -25.670   <2e-16 ***
KWH_APR                  -2.185e-01  8.005e-03 -27.292   <2e-16 ***
KWH_MAY                  -5.189e-01  7.480e-03 -69.364   <2e-16 ***
KWH_JUN                   7.977e-02  5.111e-03  15.607   <2e-16 ***
KWH_JUL                   2.228e-01  5.230e-03  42.611   <2e-16 ***
KWH_AUG                  -1.049e-01  6.616e-03 -15.855   <2e-16 ***
KWH_SEP                   1.778e-01  6.030e-03  29.486   <2e-16 ***
KWH_OCT                  -2.068e-01  6.641e-03 -31.142   <2e-16 ***
KWH_NOV                   1.003e+00  4.706e-03 213.086   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 250210309)

    Null deviance: 2.6726e+15  on 66102  degrees of freedom
Residual deviance: 1.6536e+13  on 66089  degrees of freedom
  (871 observations deleted due to missingness)
AIC: 1465896

Number of Fisher Scoring iterations: 2
```

Model 2:



```
Call:
glm(formula = KWH ~ BUILDING_TYPE + Month, family = gaussian(link = "identity"),
    data = energy_data_long)

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               34312.3      708.5  48.430  < 2e-16 ***
BUILDING_TYPEIndustrial  717419.8     9238.8  77.653  < 2e-16 ***
BUILDING_TYPEResidential -29122.4      422.0 -69.017  < 2e-16 ***
MonthKWH_FEB               -228.2      896.9  -0.254  0.79916    
MonthKWH_MAR              -1103.4      896.9  -1.230  0.21864    
MonthKWH_APR              -1346.8      896.9  -1.502  0.13320    
MonthKWH_MAY               1042.5      896.9   1.162  0.24513    
MonthKWH_JUN               4402.6      896.9   4.908 9.18e-07 ***
MonthKWH_JUL               6035.2      896.9   6.729 1.71e-11 ***
MonthKWH_AUG               4179.3      896.9   4.660 3.17e-06 ***
MonthKWH_SEP                785.0      896.9   0.875  0.38149    
MonthKWH_OCT               -215.3      896.9  -0.240  0.81032    
MonthKWH_NOV               2895.3      896.9   3.228  0.00125 ** 
MonthKWH_DEC               5504.4      896.9   6.137 8.41e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 26589415620)

    Null deviance: 2.1393e+16  on 793235  degrees of freedom
Residual deviance: 2.1091e+16  on 793222  degrees of freedom
  (10452 observations deleted due to missingness)
AIC: 21291784

Number of Fisher Scoring iterations: 2
```

(d) (2 points) Explain the reason that BUILDING_TYPE is statistically significant in Model 2 but not Model 1 from the model summary outputs.

ANSWER:

(e) (2 point) Recommend a model to your manager and justify your answer.

ANSWER: