Exam PA October 2025 Project Statement

IMPORTANT NOTICE – THIS IS THE OCTOBER 14, 2025, PROJECT STATEMENT. IF TODAY IS NOT OCTOBER 14, 2025, SEE YOUR TEST CENTER ADMINISTRATOR IMMEDIATELY.

General Information for Candidates

This examination has 12 tasks numbered 1 through 12 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. We recommend that you read the business problem and data dictionary to learn additional context about each task. 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 are working for a consulting firm that advises several clients on real estate. You are currently working on the New York City (NYC) market. Your clients are interested in a range of goals including understanding the drivers of real estate transaction prices, crafting strategies for identifying properties to buy and sell, and predicting which properties will exceed a given price threshold.

You have access to a data set of all NYC property sales data for sales completed between September 2016 and August 2017[1]*. Your data includes pricing for both commercial and real estate transactions. The data set includes location details for the properties being sold including which borough of New York the property is in and more granular neighborhood, block, lot, zip code, and address data. The data set also includes information on the characteristics of the property such as square footage, year built, and the number of units.*

We recommend that you review the data dictionary to see additional information about each variable.

[1] NYC Open Data

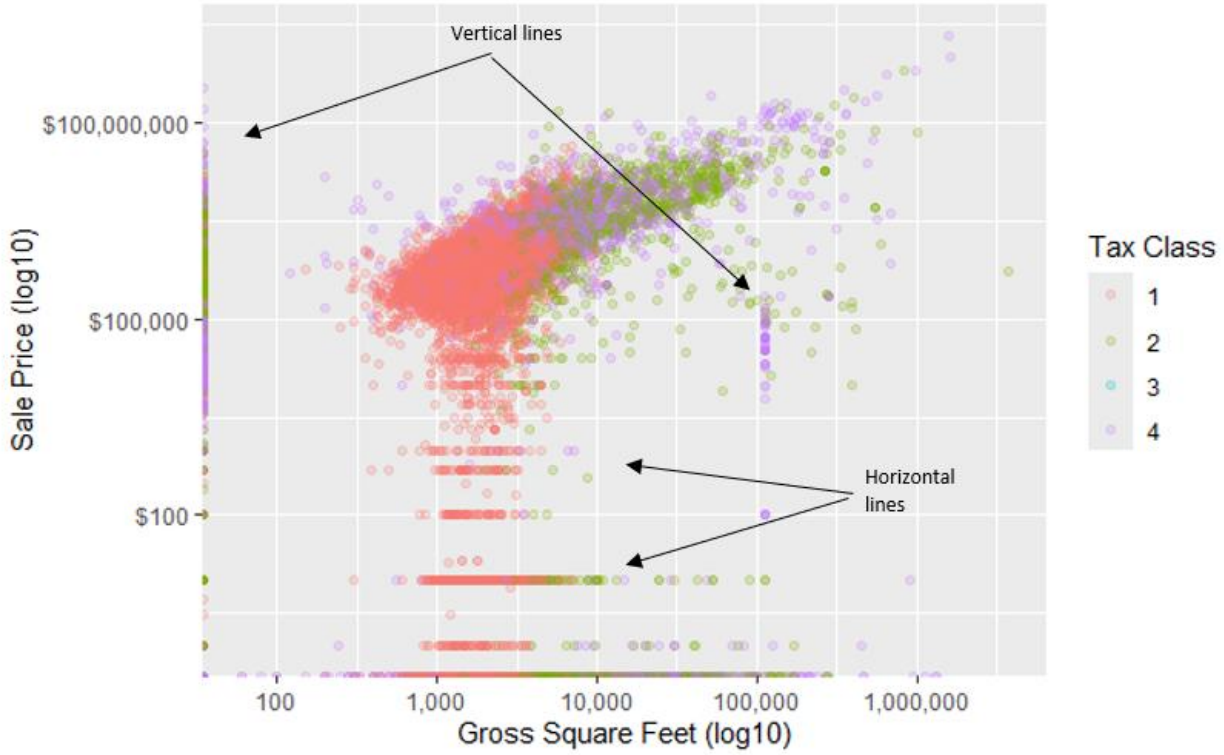
Data Dictionary

Variable	Data Type / Range / Example	Description
BOROUGH	Character Values are: 1, 2, 3, 4, 5	Borough of NYC that the property is located in. 1 - Manhattan 2 - Bronx 3 - Brooklyn 4 - Queens 5 - Staten Island
NEIGHBORHOOD	Character 254 unique values Example: Chelsea	The neighborhood the property is located in. Typically contained entirely within a borough with two exceptions.
BLOCK	Character Examples: 392, 790, 7351	Unique code representing a tax block that subdivides a borough. The combination of a block and lot code represents a unique property.
LOT	Character Examples: 6, 153, 1301	Unique code representing a lot within a tax block. The combination of a block and lot code represents a unique property. Lot codes can repeat across different blocks.
ZIP CODE	Character Examples: 10009, 10011	The ZIP code for the property. (This is equivalent to postal codes as used in other countries.)
ADDRESS	Character Example: 153 Avenue B	Street address of the property.
APARTMENT NUMBER	Character Example: 7E	Apartment number for the property (if listed).
TAX CLASS AT TIME OF SALE	Character Values are: 1, 2, 3, 4	Code representing the class for the type of building. Class 1 is residential property up to three units. Class 2 is residential property with more than three units. Class 3 is utility buildings. Class 4 is commercial and industrial properties.
RESIDENTIAL UNITS	Numeric Range: 0-1,844	Number of residential units in the property.
COMMERCIAL UNITS	Numeric Range: 0-2,261	Number of commercial units in the property.
TOTAL UNITS	Numeric Range: 0-2,261	Number of total units in the property.

GROSS SQUARE FEET	Numeric Range: 0-3,750,565	The total area of all the floors of a building as measured from the exterior surfaces of the outside walls of the building, including the land area and space within any building or structure on property.
LAND SQUARE FEET	Numeric Range: 0-4,252,327	The land area of the property in square feet.
YEAR BUILT	Numeric Range: 0-2017	Year the structure on the property was built. (Zeroes represent missing data.)
BUILDING AGE	Numeric Range: 0-2017	The difference, in years, between the sale year and the year built. (These are all whole numbers.)
SALE PRICE	Numeric Range: $0-$2,210,000,000	Price paid for the property. A $0 sale indicates that there was a transfer of ownership without a cash consideration. There can be several reasons for a $0 sale including transfers of ownership from parents to children.
SALE DATE	Date Range: September 1st, 2016, to August 31st, 2017	Date the property sold.

Task 1 (4 *points*)

Your assistant produces the following plot to investigate the relationship between **Sale Price** and **Gross Square Feet** by **Tax Class at Time of Sale**.



(a) (*2 points*) Interpret the meaning of the horizontal and vertical lines of data points.

Most candidates struggled with this task. Full-credit responses described what both the vertical and horizontal lines indicate about the underlying data, including the vertical line at 0.

ANSWER:

The vertical line at the left edge of the plot represents values of 0 square feet in the data, which are likely missing data. The vertical lines inside the plot most likely represent placeholder, estimated, or rounded values instead of exact numbers. The horizontal lines within the plot are cases where there are multiple instances of the same value for sale price (e.g., Sale Price = $100 across a wide range of property sales with different values for gross square footage); these are also likely placeholder, estimated, or rounded values.

(b) (*2 points*) Recommend how to handle the values representing horizonal and vertical lines in modeling the impact of **Gross Square Feet** on **Sales Price**. Justify your recommendation.

Candidate performance was mixed on this task. Most full-credit responses recommended either treating the observations as outliers and removing them from the dataset or adding a categorical variable to the model to capture unique behavior that may occur for thse datapoints. Many candidates provided a recommendation with weak justification or no justification at all, receiving partial or no credit.

ANSWER:

I recommend identifying and removing records with 0 or likely placeholder. They may not represent the true relationship between the variables, and including them will bias the coefficient estimates.

Task 2 (4 *points*)

You work for an actuarial consulting firm and have been approached by a real estate investment organization to analyze recent NYC sales data to help them make some purchase decisions. They've asked for the following:

1. A recommendation on whether to convert current properties from Commercial use to Residential.
2. A valuation of properties under purchase consideration given historical valuations.
3. A summary of the recent sales experience.

(a) (*1 point)* Categorize how each deliverable ties to work done within the general area of Predictive Analytics (descriptive, predictive, or prescriptive analytics).

Candidates performed very well on this task. A common error was the inability to correctly differentiate between prescriptive and predictive analytics. Full points were awarded for correctly naming the type of analytics. Partial credit was awarded for detailed descriptions of the types of analytics problems even if the final classifications provided were mismatched.

ANSWER:

1. This is a prescriptive analytics analysis because it answers a "what if?" question
2. This is a predictive analytics analysis because it is used to estimate what a property is worth.
3. This is a descriptive analytics analysis because it is used to understand what has happened.

Your investment client is presented with the opportunity to purchase some buildings to tear down for new construction. They would like to maximize the value of the new construction.

(b) (*3 points)*
i. Identify what type of analytic question this represents.
ii. Identify current data that would be useful for this analysis.
iii. Propose additional data outside of what was provided that you would want to include as well.

Candidates performed well on this task. A common error was failure to include Price or Sale Price variables in the analysis; these would be important factors for evaluating the value of any construction project. Candidates must also clearly identify the need for prescriptive analysis here, as the business problem requires recommending an appropriate course of action. Most full-credit responses to iii proposed variables related to construction costs related variables, but other numerical or categorical variables were accepted if they were relevant to the question.

ANSWER:

i. This is a prescriptive analysis task.

ii. The current data relevant to this task would be the price per square foot for different types of buildings. This would be relevant both for the price you could sell a new building for and the cost of acquiring buildings that would be replaced.

iii. Demolition and construction costs represent data that are not included in the current data set, but that would be relevant in estimating the value of tearing down and replacing buildings.

Task 3 (*4 points*)

Your client is interested in predicting **SALE PRICE** for residential properties in New York City, df_subset represents only residential properties. The analysis focuses on **LOG_GROSS_SQUARE_FEET** (natural logarithm of gross square feet), **BUILDING AGE** (in years), and **BOROUGH_TXT** (a categorical variable representing the borough for the property). Note that 'BUILDING AGE' * 'BOROUGH_TXT' in the formula below will include both variables and their interaction.

Two Generalized Linear Models (GLMs) are fitted in R using a gamma distribution with a log link for SALE PRICE:

```
glm_model_1 <- glm(
  `SALE PRICE` ~ LOG_GROSS_SQUARE_FEET + `BUILDING AGE` + `BOROUGH_TXT`,
  family = Gamma(link = "log"),
  data = df_subset
)

glm_model_2 <- glm(
  `SALE PRICE` ~ LOG_GROSS_SQUARE_FEET + `BUILDING AGE` * `BOROUGH_TXT`,
  family = Gamma(link = "log"),
  data = df_subset
)
```

(a) (*2 points*) Describe a benefit of each model.

Candidates performed well on this task overall. To earn full credit, candidates needed to correctly differentiate the advantages of the model without interaction term (Model 1) versus including it (Model 2). Model 1's key strength is simplicity or ease of interpretation. Model 2 is expected to be superior in goodness of fit because it allows the effect of the building age variable to vary across different levels of the borough variable.

ANSWER:

In glm_model_1, the coefficient for BUILDING AGE represents a single, constant effect of building age on the log of expected SALE PRICE across all boroughs. The model assumes that a one-year increase in building age has the same impact on SALE PRICE regardless of the borough. The benefit of this model is the simplicity of the interpretation and the ability to isolate the impact of each variable.

In glm_model_2, the inclusion of the interaction term BUILDING AGE * BOROUGH_TXT allows the effect of BUILDING AGE on SALE PRICE to vary for each borough. Each borough will have its own unique slope for BUILDING AGE, which is a combination of the main effect of BUILDING AGE (for the reference borough, Bronx) and the interaction term for that borough. The benefit of this model is the increased flexibility in allowing for different slopes for building age in each borough.

You compare the AIC of glm_model_1 and glm_model_2

```
> AIC(glm_model_1) - AIC(glm_model_2)
[1] 444.1433
```

(b) (2 points) Explain what the Akaike Information Criterion (AIC) is telling you about the difference between the model fits. Recommend and justify which model to use.

Candidates performed well on this task overall. Full credit required the correct recommendation of Model 2 justified by its superior AIC value, together with an explanation of how AIC works, including a description of the penalty parameter. Some candidates failed to describe the penalty parameter and only described AIC as a measure of goodness-of-fit; these responses received partial credit.

ANSWER:

I recommend glm_model_2 since it has a lower AIC value. Lower AIC indicates a model with a better balance between model fit (measured using log-likelihood) and complexity (measured as 2 times the number of parameters).

The positive AIC difference (444.1443) tells us that the inclusion of the interaction terms between the variables BUILDING AGE and BOROUGH_TXT significantly improves the model's fit, even after penalizing for the additional parameters introduced by these terms. This indicates that allowing the effect of building age to vary by borough provides a more powerful model.

Task 4 (8 *points*)

Your assistant has identified supplemental information for each sale that documents written notes taken by each real estate agent describing the property sale including background on the sellers, impressions of the neighborhood, and a description of the property.

(a) (*2 points)* Identify what kind of data this is and give one advantage and one disadvantage of including this information in your model.

Candidates performed well on this task overall. Most candidates provided responses similar to the model solution. Some candidates stated that qualitative or personal information is inappropriate to use; credit was not awarded for these responses as they cannot be justified for the given business problem.

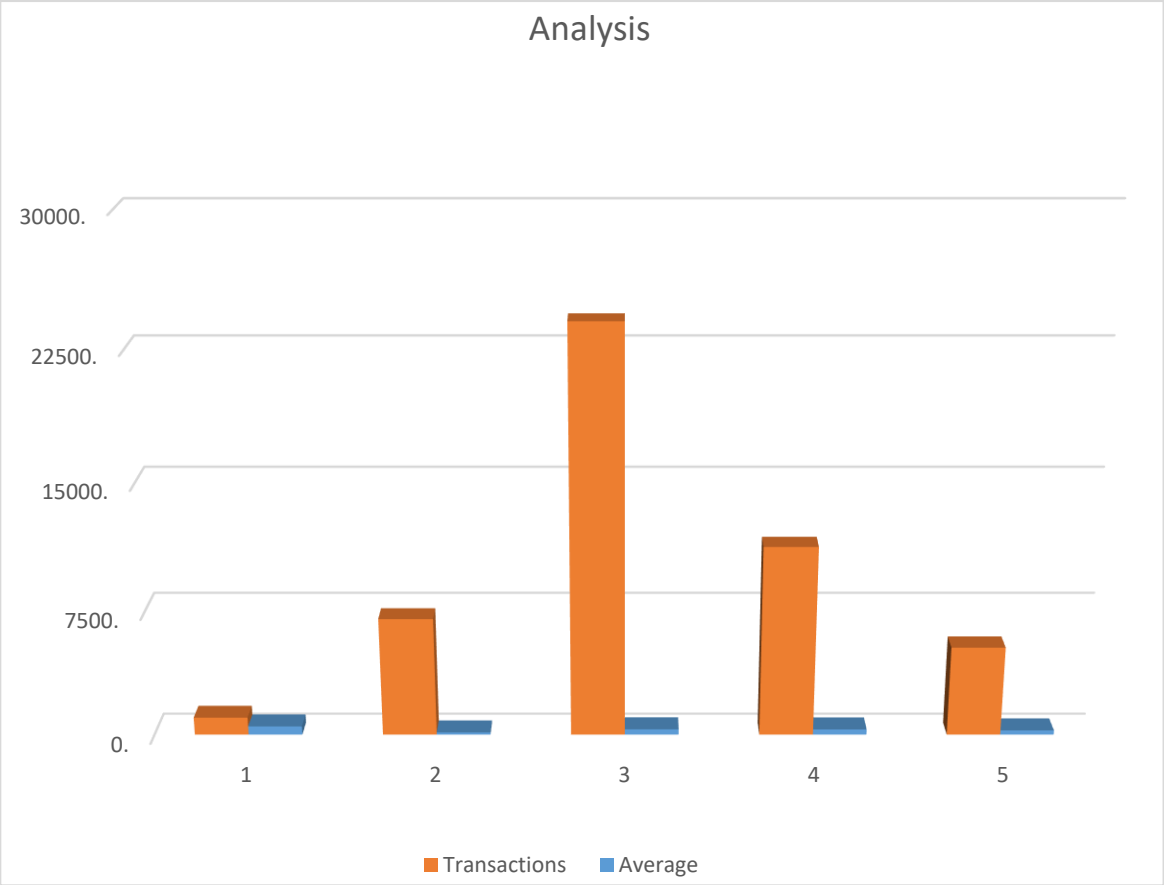
ANSWER:

This is unstructured data.

Advantage: Unstructured data includes information that cannot be stored in a tabular format. Using this unstructured data gives insights and qualitative information that cannot be included in a structured dataset, e.g. insights on the sellers and background around the type of sale.

Disadvantage: Unstructured data requires more complex methods to process for input into a predictive model. It can also be more time-consuming and resource-intensive to analyze unstructured data.

Your assistant started his analysis of the sales data by summarizing the number of different transactions by NYC Borough along with the average sales dollar per square foot and created the following graph.



(b) (*3 points)* Identify three issues with the graph; recommend a way of addressing each issue that you identify.

Candidates performed well on this task overall. Full-credit responses addressed problems with the way the graph presents the data. Some candidates described modeling challenges that may need to be addressed given the data provided in the graph; these responses were considered off-topic and not awarded credit. Three valid responses were required for full credit; any combination of issues in the model solution or other valid issues, like using a secondary y-axis were accepted.

ANSWER:

- The variable names are not descriptive. I recommend renaming the variables to "Number of Transactions" and "Average Sales Dollar per Square Foot."
- The horizontal axis labels are numeric but represent categorical Borough data. I recommend using Borough names (Manhattan, Bronx, etc.) instead of the numeric encodings.

- **Number of Transactions** and **Average Sales Dollar per Square Foot** are on different scales, making the visualization ineffective for understanding **Average Sales Dollar per Square Foot**. I recommend using two different charts, one for each of these variables.
- 3-dimensional effects are unnecessary and make it harder to read the bar heights. I recommend using a two-dimensional graph instead.

Your assistant is digging further into three variables and wants to use some bivariate graphing techniques including scatterplots, stacked histograms, and box plots to try to gather insights from the data.

BOROUGH	Character Values are: 1, 2, 3, 4, 5	Borough of NYC that the property is located in. 1 - Manhattan 2 - Bronx 3 - Brooklyn 4 - Queens 5 - Staten Island
GROSS SQUARE FEET	Numeric Range: 0-3,750,565	The total area of all the floors of a building as measured from the exterior surfaces of the outside walls of the building, including the land area and space within any building or structure on property.
SALE PRICE	Numeric Range: $0-$2,210,000,000	Price paid for the property. A $0 sale indicates that there was a transfer of ownership without a cash consideration. There can be several reasons for a $0 sale including transfers of ownership from parents to children.

(c) (*3 points)* Complete the table below.

i. Identify an appropriate bivariate graph for each pair of variables.

ii. Explain what kind of insight could be gained from each graph.

Candidates performed well overall on this task. A common poor answer was recommending stacked histograms.

ANSWER:

Bivariate Pair	**Borough, Gross Square Feet**	**Sale Price, Gross Square Feet**	**Borough, Sale Price**
Bivariate Graph			
Insight Gained			

Bivariate Pair	**Borough,** **Gross Square Feet**	**Sale Price,** **Gross Square Feet**	**Borough,** **Sale Price**
Bivariate Graph	Box Plots	Scatterplot	Box Plots
Insight Gained	Comparative distribution of size of square footage for sold properties between various boroughs	Look for any linear relationships, correlation, or patterns between Sale price and Square Feet	Look at the relative distribution of sale prices between the different boroughs

Task 5 (4 *points*)

You are an actuarial analyst reviewing a model built by a colleague. The goal of the model is to identify "high value" residential properties, defined as those with a **Sale Price** greater than $1,000,000.
Your analyst creates a variable called **High Value** and calculates the proportion of each class.

Next, your colleague splits the data into test and training data and builds a classification tree.

```
# Set seed for reproducibility and partition data
set.seed(123)

training.indices <- createDataPartition(df_subset$`HIGH VALUE`, p = 0.7, list = FALSE)
train <- df_subset[training.indices, ]
test <- df_subset[-training.indices, ]


# Fit the classification tree model
tree_model <- rpart(
  `HIGH VALUE` ~ `GROSS SQUARE FEET` + `BUILDING AGE` + `BOROUGH_TXT`,
  data = train,
  method = "class",
  control = rpart.control(cp = 0.01)
)
```

(a) (*2 points*) Explain the role of the hyperparameter cp and the likely effect of decreasing the cp value?

Candidates performed well on this task overall. A common error that was awarded partial credit was confusing increases in cp with decreases. Some candidates explained the role of hyperparameters generally, not addressing cp specifically; these responses were awarded minimal partial credit.

ANSWER:

Cp stands for Complexity Parameter. It is a threshold used to control the size of the tree. A split must decrease the impurity measure by at least the threshold represented by the cp. Decreasing the cp value would relax this constraint, likely resulting in a larger, more complex tree with more splits and leaf nodes.

Your analyst runs a summary table of the decision tree, and the results are depicted below:

```
Node number 1: 17323 observations,    complexity param=0.09442509
  predicted class=0  expected loss=0.2485135  P(node) =1
    class counts: 13018  4305
   probabilities: 0.751 0.249
  left son=2 (11549 obs) right son=3 (5774 obs)
  Primary splits:
      GROSS SQUARE FEET < 2081.5 to the left,  improve=766.5145, (94 missing)
      BOROUGH_TXT       splits as  LRRLL,      improve=729.5842, (0 missing)
      BUILDING AGE      < 102.5  to the left,  improve=217.5639, (0 missing)
  Surrogate splits:
      BUILDING AGE < 107.5  to the left,  agree=0.685, adj=0.061, (94 split)
      BOROUGH_TXT  splits as  LLRLL,      agree=0.672, adj=0.022, (0 split)

Node number 2: 11549 observations
  predicted class=0  expected loss=0.1436488  P(node) =0.6666859
    class counts:  9890  1659
   probabilities: 0.856 0.144

Node number 3: 5774 observations,    complexity param=0.09442509
  predicted class=0  expected loss=0.4582612  P(node) =0.3333141
    class counts:  3128  2646
   probabilities: 0.542 0.458
  left son=6 (1695 obs) right son=7 (4079 obs)
  Primary splits:
      BOROUGH_TXT       splits as  LRRRL,      improve=555.60050, (0 missing)
      GROSS SQUARE FEET < 2780.5 to the left,  improve=128.67970, (0 missing)
      BUILDING AGE      < 67.5   to the left,  improve= 85.90416, (0 missing)
  Surrogate splits:
      BUILDING AGE < 47.5   to the left,  agree=0.723, adj=0.058, (0 split)
```

(b) (*2 points*) Explain how the model handles the missing observations on Node number 1.

Performance was mixed on this task, with many with many candidates leaving the task blank. Most candidates who did provide an answer correctly described how surrogate splits are created and applied in this example, receiving full credit.

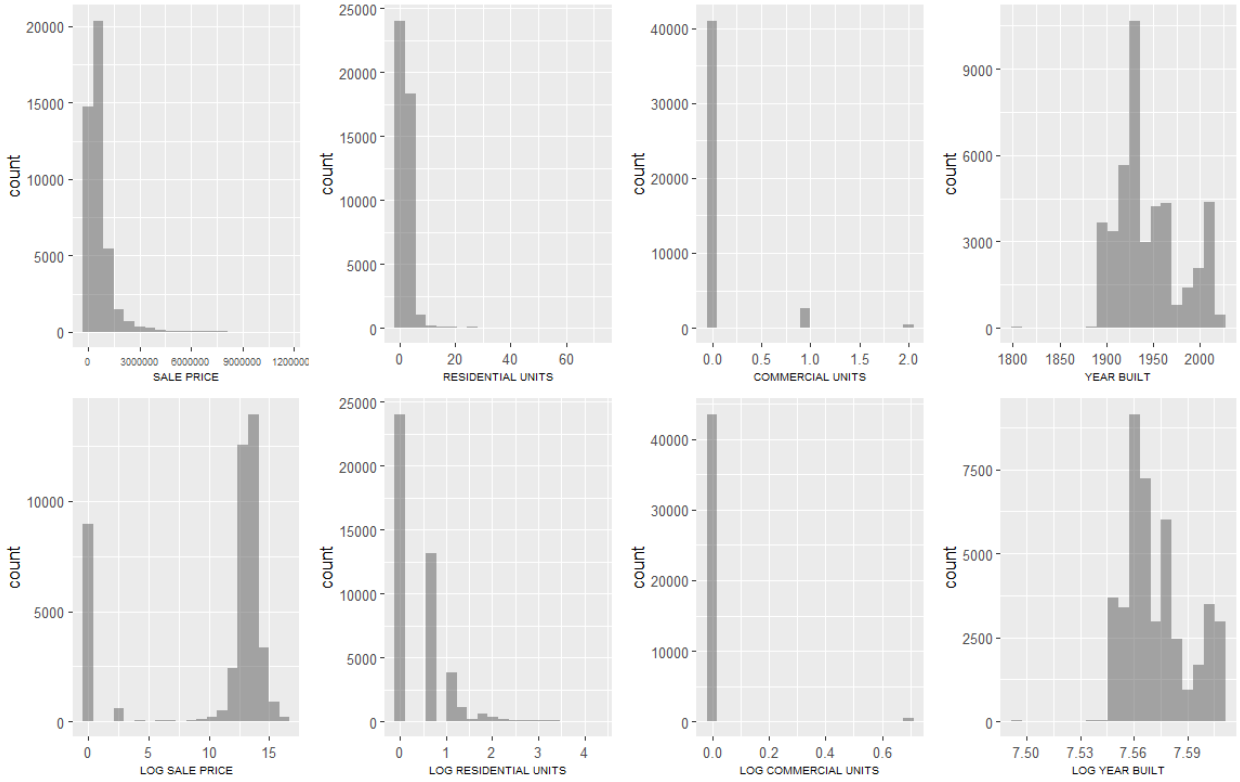
ANSWER:

94 observations in the training data had a missing value for GROSS SQUARE FEET and could not be evaluated by the primary split rule. The model handles these by using the best surrogate split. According to the output, the first surrogate is BUILDING AGE < 107.5. The (94 split) next to this rule confirms that this is the rule used to classify all 94 of those specific observations.

Task 6 (8 *points*)

Your client has asked you to build a regression model predicting the **Sale Price** for a wide variety of properties in New York City, to help identify properties that may have been sold for unreasonably low prices. Before building a model, your manager suggests examining some of the available variables to see whether transformations should be applied first.

For several quantitative variables, the following histograms were produced, including histograms for the log version of each variable (in all cases values below 1 of the raw variables have been forced to equal 1 before taking the log):



(a) (*2 points*) Identify the variables for which a log transformation appears to be helpful; in each case explain why it would be helpful.

Candidates performed very well on this task. Full-credit answer identified that SALE PRICE is where a log transformation is most promising. Partial credit was awarded to candidates who chose RESIDENTIAL UNITS but failed to provide sufficient justification. Some candidates stated that a log transformation is inappropriate for all variables since 0 may be in the domain of the variables; these responses received minimal credit since this may only apply to some of the variables and can be easily avoided by adding a constant inside the log transformation.

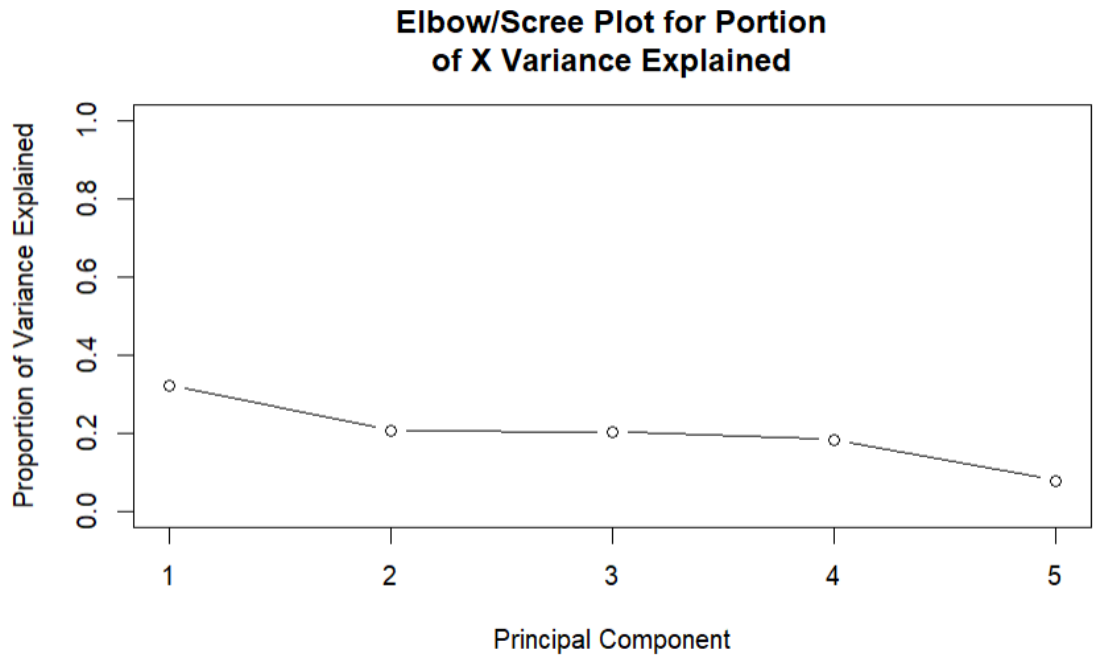
ANSWER:

The log transformation seems primarily to be helpful for the SALE PRICE itself, which appears very skewed. Zero values can be addressed by including a constant in the log transformation.

While a log transformation seems to help a little with RESIDENTIAL UNITS, it still looks very skewed and it may be more helpful to perform some other transformations instead.

The log transformation does not appear useful for commercial units or year built.

After modifying and standardizing five of the numeric variables (**Residential Units**, **Commercial Units**, **Land Square Feet**, **Gross Square Feet**, and **Year Built**), you ask your assistant to try applying principal components regression to them, with the following results:




```
VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV          980723   917371   917287   916917   916890   915262
```

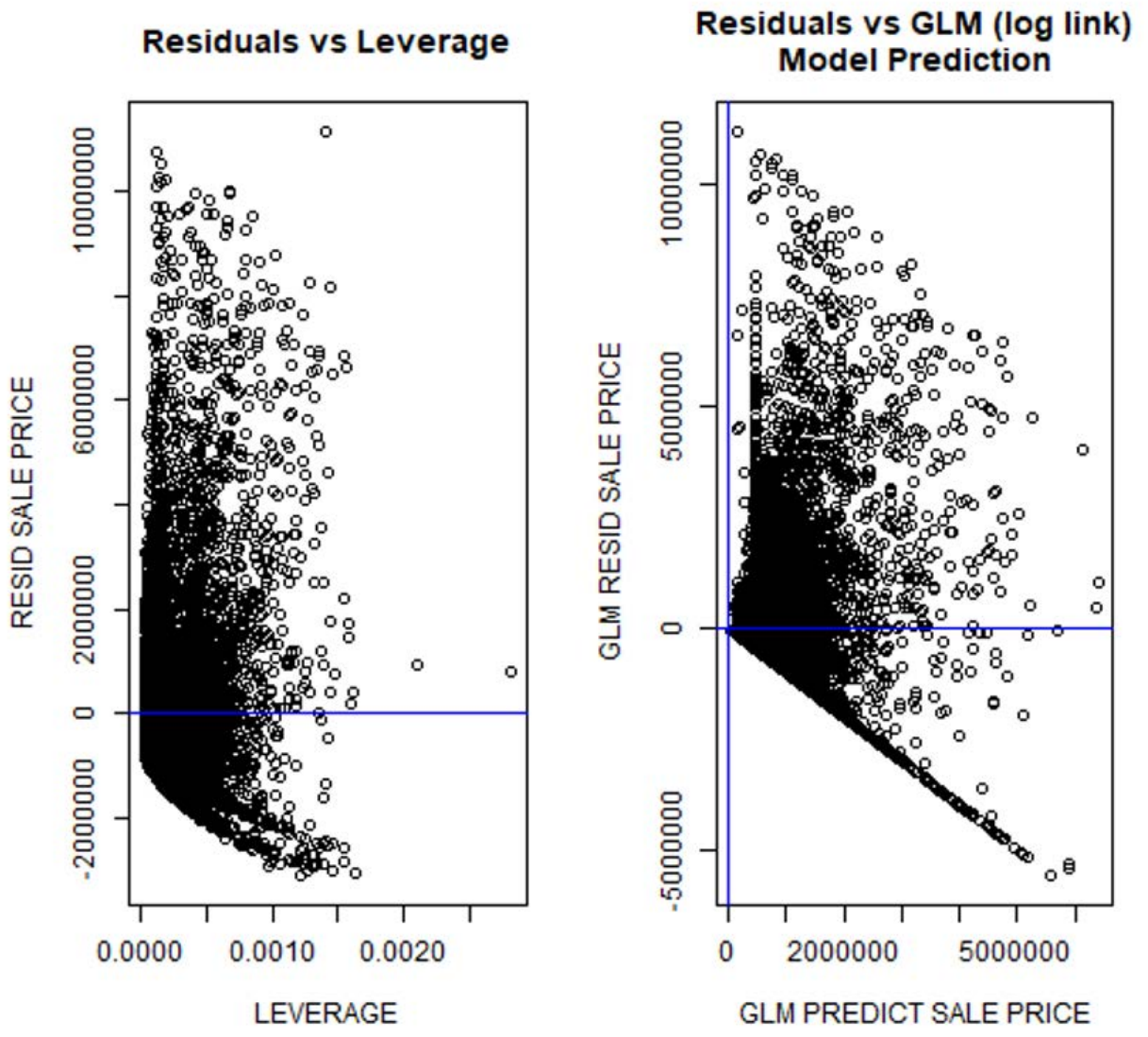
(b) (*3 points*) Recommend whether principal components should be used for this model, and if so, how many components. Justify your recommendation.

Candidates performed well on this task overall. The most common partial-credit responses recommended PCA with a generic or otherwise weak justification. A few candidates mistakenly interpreted the elbow plot as if this was a cluster analysis problem rather than a PCA analysis; these responses were awarded no credit.

ANSWER:

I do not recommend using PCA in this scenario. PCA is most valuable where there are many correlated variables, and a large proportion of the total variance is explained using just a few of the components. In this case, all the components, except for the last one, seem to explain a reasonable proportion of the variation. The gain of using fewer variables through the use of PCA in this case would not seem to outweigh the loss of explainability.

You decide to train a GLM with a log link function without applying ridge regression or using PCA. Your model produces the following diagnostic plots:



Recall from the histograms that **Sale Price** can never be negative.

(c) (*3 points*) Describe what you would look for on residual and leverage plots and identify any concerns based on each of the plots above.

Candidates performance was mixed on this task. Full-credit answers included descriptions on both residual and leverage plots and listed proper concerns on each plot. Most candidates addressed the residual plot by discussing the mean at 0, heteroskedacitiy, and distribution. Fewer candidates effectively discussed the leverage plot, with very few candidates discussing leverage.

ANSWER:

The desired distribution of the residual plots is to be symmetrical and centered at zero with no clear trend in the distribution of residuals across predictions. The residual plot exhibits a much higher distribution above 0 than below 0, this is due to the sales price being non-negative.

The leverage plot can be used to identify outliers with high leverage, meaning they have a large impact on the model coefficient estimates. There are a few large outliers in the leverage plot. I recommend reviewing these observations to determine whether they represent data issues and whether they should be removed from the dataset.

Task 7 (8 *points*)

You have been engaged for tax assessment purposes to predict which properties *should have sold* for at least $1,000,000, regardless of the actual **Sale Price**.

You split your data into a 50% training set and a 50% testing set not used to train the model. You start by fitting a simple logistic regression model with a single predictor. Using a predicted probability threshold of 50%, the resulting confusion matrix is:

		ACTUAL ABOVE $1M	
PREDICTED ABOVE $1M		0	1
	0	39219	6927
	1	309	420

(a) (*3 points*) Calculate the accuracy, the sensitivity, and the precision of this model.

Candidates performed well on this task overall, with most candidates receiving full credit. The most common error was mixing up the positive and negative cases; these responses received partial credit. Descriptions of the metrics, like those in the model solution, were not required for full credit, but partial credit was awarded for providing these when candidates made calculation errors.

ANSWER:

Accuracy is the proportion of observations where the model prediction is correct:

Accuracy = (39219 + 420) / (39219 + 420 + 6927 + 309) = 84.56%.

Sensitivity is the proportion of actual positives that are correctly classified by the model:

Sensitivity = 420 / (420 + 6927) = 5.72%.

Precision, or true positive rate, is the proportion predicted positives that are correctly classified by the model:

Precision = 420 / (420 + 309) = 57.61%.

In flagging properties as potentially being worth $1M or more, your boss would prefer a model that captures more of the properties that truly would sell for $1M or more (more true positives).

(b) (*2 points*) Explain how the model can be adjusted to increase the number of true positives.

Most candidates received full credit on this task. Many candidates provided an answer around changing the threshold, similar to the model solution; these responses received full credit provided they had a

sound explanation for how this impacts the number of positive and negative predictions. Some candidates recommended over/undersampling; these answers received full credit, provided a strong explanation was given. Although the model solution presents two different explanations, only one was required for full credit.

ANSWER:

The model can be adjusted simply by changing the probability threshold that is used to flag an observation as predicted positive or predicted negative. For example, instead of using a 50% threshold, which is typically the default, we could predict an observation to be positive if the probability is > 40%.

Alternatively, oversampling could be used to increase the number of positive response in the training dataset. This forces the model to fit better to the positive cases, which will improve predictive power on positive observations.

You fit a second model with several predictors and compare the two models using the following statistics. The statistics labeled SMALL are for the model with one predictor, which those labeled LARGE are for the larger model. The model was trained only on the observations in the training set, statistics for both the training and test sets are presented below.

partition <chr>	AUC SMALL <dbl>	AUC LARGE <dbl>	ACCURACY SMALL <dbl>	ACCURACY LARGE <dbl>	AIC SMALL <dbl>	AIC LARGE <dbl>
test	0.7362927	0.7341090	0.8444938	0.8448784	*NA*	*NA*
train	0.7302759	0.7314429	0.8467666	0.8472352	18262.65	18174.44

(c) (*3 points*) Explain how these statistics can be used to evaluate model performance and discuss a limitation of each.

Candidates struggled with this task. Full credit was awarded for an explanation of each statistic and some discussion of how to interpret train vs test statistics for each statistic.

ANSWER:

The AUC measures the tradeoff between the true positive rate and the false positive rate at different thresholds for classification. A larger number is better with 1 representing perfect classification and 0.5 representing random classification. One limitation of AUC is that it is calculated across all probability thresholds. If we want to understand performance of classification at a set threshold, other metrics will be more informative.

Accuracy represents the number of true positives and true negatives divided by the total number of predictions in a classification problem; a higher model accuracy indicates better model performance. One limitation of accuracy is the calculation assigns the same weight to true positives and true negatives. These may have different importance in context of the business problem.

AIC is a statistic used to measure model performance, but with a penalty for model complexity. For AIC a lower value indicates better model performance. One limitation of AIC is that it is typically measured on

training data and used for model selection purposes but is only a proxy for the actual performance of models on test data. The model with the lowest AIC may not perform best on unseen test data.

Task 8 (4 *points*)

Your assistant has fit a model on a subset of the total data using the following variables:

- **borough**: factor encoding Borough
- **age**: calculated as Sale Date – Year Built (same as **Building Age**)
- **gross_sqft**: Gross Square Feet

You are provided with the following output from their R code:

```
Call:
glm(formula = log(price) ~ borough * age + log(gross_sqft) *
    age, family = Gamma(link = "identity"), data = mdat)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         11.3135535  0.2940427  38.476  < 2e-16 ***
borough2            -1.3994785  0.1892583  -7.395 1.50e-13 ***
borough3            -1.1847656  0.1890262  -6.268 3.78e-10 ***
borough4            -1.0125061  0.1895660  -5.341 9.39e-08 ***
borough5            -1.2305733  0.1894166  -6.497 8.51e-11 ***
age                 -0.0024197  0.0032101  -0.754    0.451
log(gross_sqft)      0.4367615  0.0236985  18.430  < 2e-16 ***
borough2:age        -0.0023806  0.0018015  -1.321    0.186
borough3:age        -0.0019583  0.0017904  -1.094    0.274
borough4:age        -0.0035918  0.0017995  -1.996    0.046 *
borough5:age        -0.0032129  0.0018108  -1.774    0.076 .
age:log(gross_sqft)  0.0006921  0.0002926   2.366    0.018 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.001250661)

    Null deviance: 32.960  on 13154  degrees of freedom
Residual deviance: 19.328  on 13143  degrees of freedom
AIC: 20068

Number of Fisher Scoring iterations: 5
```

(a) (*4 points*) Identify the boroughs for which the model predicts that property value increases as age increases, given the gross square feet of the property is 2,500. Show your work.

Most candidates struggled with this task. Full credit required the identification of the correct boroughs, showing the supporting calculations. Many candidates failed to properly capture all coefficients relating to age in their calculations. Other errors included calculation mistakes and failing to substitute the 2,500 constant provided for gross square feet.

ANSWER:

From the summary output provided, the relationship of predicted property value with age, borough, and log(gross_sqft). Plugging in log(2500) = 0.005415022, the relevant coefficient calculations are summarized in the table below:

	age	borough:age	age: log(2500)	Sum of age coefficients	Price Increase with Age

borough 1	-0.0024197	0	0.005415022	0.002995322	TRUE
borough 2	-0.0024197	-0.0023806	0.005415022	0.000614722	TRUE
borough 3	-0.0024197	-0.0019583	0.005415022	0.001037022	TRUE
borough 4	-0.0024197	-0.0035918	0.005415022	-0.000596478	FALSE
borough 5	-0.0024197	-0.0032129	0.005415022	-0.000217578	FALSE

Since only boroughs 1, 2, and 3 have positive coefficients, predicted property values for those boroughs increase with age.

Task 9 (6 *points*)

Your boss wants to predict the **Sale Price** of properties with a decision tree model. Your assistant is investigating which variables are appropriate for use in building this model. Four sample rows of the variable **Address** have been output below:

Record ID	ADDRESS
101	153 AVENUE B
234	154 EAST 7th STREET
384	164 EAST 10th STREET
452	210 AVENUE B
……	…….

Your assistant has discovered this is a character variable and is unique for each record in the data set. They want to change this field to a factor variable and use it in the decision tree.

(a) *(1 point)* Provide one reason why using this as a factor variable could be problematic.

Candidates performed very well on this task, and most candidates received full credit. Very few candidates addressed the issue around making predictions on values for address that were not in the training data. Most full-credit responses resembled the alternative reason in the model solution. Some responses received partial credit for being overly generic; many of these responses were around interpretability concerns or overfitting concerns in a decision tree model.

ANSWER:

One problem is that the variable is very granular with many unique values. When the model is run on test or future data, there will be levels of the address variable that were not in the training data, and the model will not be able to make predictions. Hence, this field cannot be relied on for making predictions on unseen data.

Alternative issue: Another potential issue is that the model will struggle to fit to such granular data, with very few observations at each level. Conclusions based on how the model treats levels of this variable would likely be spurious and due to model overfitting rather than a true relationship between the variables.

(b) *(2 points)* Recommend and Justify one data transformation that can be applied to **Address** to make it more useable in the decision tree model. Give one example of the data transformation applied to one of the sample rows above in your answer.

Candidates performed well on this task overall. Full-credit answers included three components: a recommendation, proper justification, and a feasible example of data transformation. No credit was awarded for answers lacking any justification.

ANSWER:

One data transformation is to extract the Street name as a new factor variable. This new variable is an improvement because it will remove the problem of having unique or very few value for each observation. There will be significantly fewer levels, which allows us to generalize the **Address** field, which can then be more effectively used in the prediction model.

An example of this would be to transform "153 AVENUE B" along with "210 AVENUE B" into "AVENUE B" as a single level as they are both on that street.

Your assistant notices **ZIP Code** can be a numeric field, and that the sequential numbers many times are geographically adjacent to one another, however this is not always the case. Some ZIP codes that are adjacent to one another have non-sequential numbers.

(c) (*2 points*) Compare and Contrast using **ZIP Code** as a Numeric value in a GLM vs. a Tree Based Model.

Candidate performance was mixed on this task. Full-credit answers discussed how GLMs treat numeric variables as opposed to how tree models treat numeric variables and addressed the implications for zip code. Partial credit was awarded for general explanations of linear vs non-linear relationships in the models but failed to make a connection to the ZIP Code variable.

ANSWER:

For ZIP Code to be appropriate to use as a numeric variable, especially in a GLM, there should be a meaning to the ordering and the scale of the numbers. While it may be true that the numbers close to each other sometimes share similar traits, is certainly is not the case that ZIP code 10000 is in any meaningful way half as much as ZIP code 20000. Also, ZIP codes that border one-another do not necessarily have values directly sequential to one another.

Using this variable in a tree-based model may produce better results than using it in a GLM model. Tree based models natively handle non-linear relationships and can carve out segments of zip codes that are numerically close to each other and have predictive value.

Your assistant would like to use K-means clustering using the dimensions **Borough** and **Sale Price.** They convert **Borough** to a numeric value by using label encoding as shown in the table below. Your assistant plans to standardize/normalize both variables before clustering**.**

Borough	Encoded Number
Queens	1
Brooklyn	2
Bronx	3
Staten Island	4
Manhattan	5

(d) (*1 points*) Describe one problem that arises from applying K-means clustering to the variables **Borough** and **Sale Price**.

Candidate performance was mixed on this task. Full-credit answers addressed how the distance measure is not meaningful for factor variables encoded this way, leading to poor model results. A few candidates discussed target leakage issues in a GLM; no credit was awarded for these responses as they were considered off-topic.

ANSWER:

The clusters will be nonsensical, as K-means clustering uses Euclidian Distance to calculate the similarity of two points. Hence once the data is normalized, it will tend to group Queens/Brooklyn as similar to each other and dissimilar to Manhattan. This is solely an artifact of the way the data is encoded rather than reflecting a meaningful similarity between the levels of the variable.

Task 10 (4 *points*)

Your client is acquiring a portfolio of NYC properties and wants to use the NYC Rolling Sales dataset to build a linear model that predicts log (**Sale Price**) per **Gross Square Foot** using variables including **Borough**, **Year Built**, **Total Units**, **Land Square Feet** and **Gross Square Feet**. Your assistant inspects the data set and points out there are missing values, they decide to remove all rows with missing values, build a GLM model and provide you with the model summary, diagnostic plot and 5 selected observations.

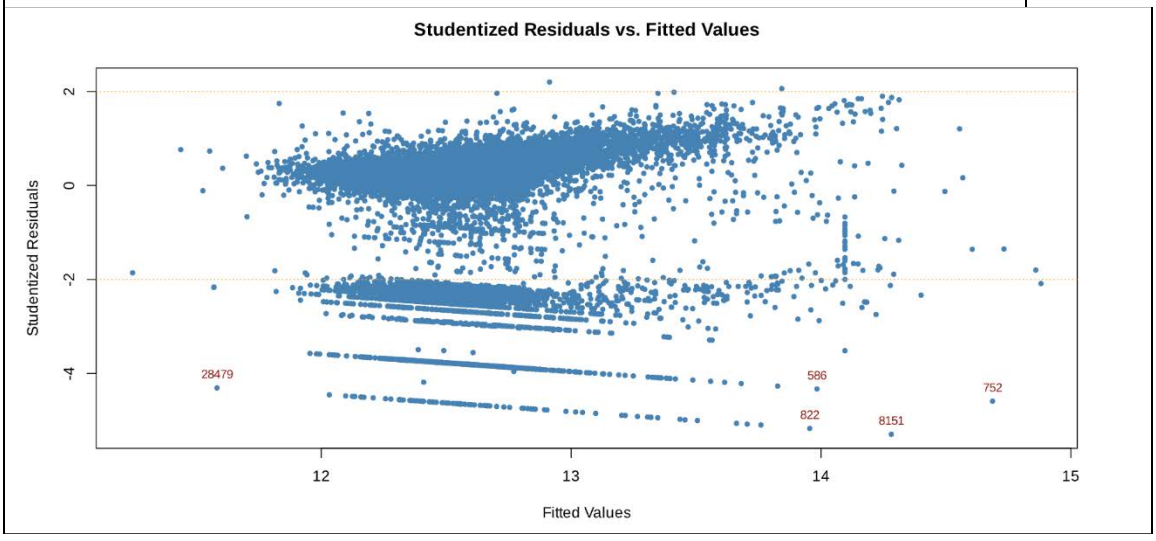
Note that log transformations are applied to variables **Sale Price**, **Gross Square Feet** and **Land Square Feet** variables.

```
Call:
lm(formula = lnSalePrice ~ borough + lnGrossSqFt + lnLandSqFt +
    total_units, data = data_model)

Residuals:
     Min       1Q   Median       3Q      Max
-14.2670   0.3761   0.8108   1.1689   5.9432

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.7019701  0.2815017  38.017  < 2e-16 ***
borough2    -0.6503996  0.1051876  -6.183 6.36e-10 ***
borough3    -0.1972299  0.0986087  -2.000  0.04550 *
borough4    -0.2705315  0.1034283  -2.616  0.00891 **
borough5    -0.3455463  0.1093435  -3.160  0.00158 **
lnGrossSqFt  0.3183525  0.0295742  10.765  < 2e-16 ***
lnLandSqFt  -0.0358013  0.0340083  -1.053  0.29248
total_units -0.0001442  0.0006859  -0.210  0.83349
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.703 on 29321 degrees of freedom
Multiple R-squared:  0.01121,   Adjusted R-squared:  0.01098
F-statistic:  47.5 on 7 and 29321 DF,  p-value: < 2.2e-16
```



```
data_model[c(28479,752,8151,822,586), ]
```

A tibble: 5 × 8

sale_price	borough	gross_square_feet	land_square_feet	residential_units	lnSalePrice	lnGrossSqFt	lnLandSqFt
<dbl>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	5	240	2128899	0	0.000000	5.480639	14.571116
10	1	907938	69096	0	2.302585	13.718931	11.143252
1	3	457966	49122	0	0.000000	13.034550	10.802062
1	1	82600	10217	77	0.000000	11.321765	9.231808
10	1	112493	23251	181	2.302585	11.630646	10.054103

(a) (*2 points*) Explain how the Studentized Residual is calculated and its advantage over ordinary residuals.

Most candidates struggled this question. Full-credit responses included both a complete explanation of the calculation and its advantage over raw residuals. Some candidates partially explained how the studentized residual is calculated, but few were able to get the complete correct answer. Many candidates gave an incorrect definition which received zero credit. Few candidates were able to explain the advantage of using studentized residual, although some did mention the reduced impact of outliers and the benefit of being on the same scale.

ANSWER:

Studentized residuals scale the raw residuals by the estimated standard errors, adjusting for each point's leverage. This puts all residuals on a common, approximately t-distributed, scale, so you can directly compare extremeness across observations. Raw residuals ignore leverage and let a single outlier inflate the standard error.

Your assistant builds another GLM removing one outlier observation that corresponds to a very large property with extremely high **Land Square** but a low **Sale Price** that is much lower than predicted by the original model.

(b) (*2 points*) Discuss the directional change of the model coefficient for **LnLandSqFt** after removing this observation.

Candidate performance was mixed on this task. Full credit was awarded for discussing how. Most candidates correctly stated the directional change, but fewer candidates were able to clearly describe why this is the case.

ANSWER:

A very large land_square_feet property with low sale_price will have a negative residual. This pushes the coefficient of LnLandSqFt downward. Removing this outlier will increase the coefficient estimate of LnLandSqFt.

Task 11 (7 *points*)

Your assistant has built a random forest to predict the sale price of properties in New York. Your boss would like to better interpret the variables used in the model.

(a) (*2 points*) Compare and contrast the purpose of feature importance vs. partial dependent plots as methods to interpret the inner workings of the model.

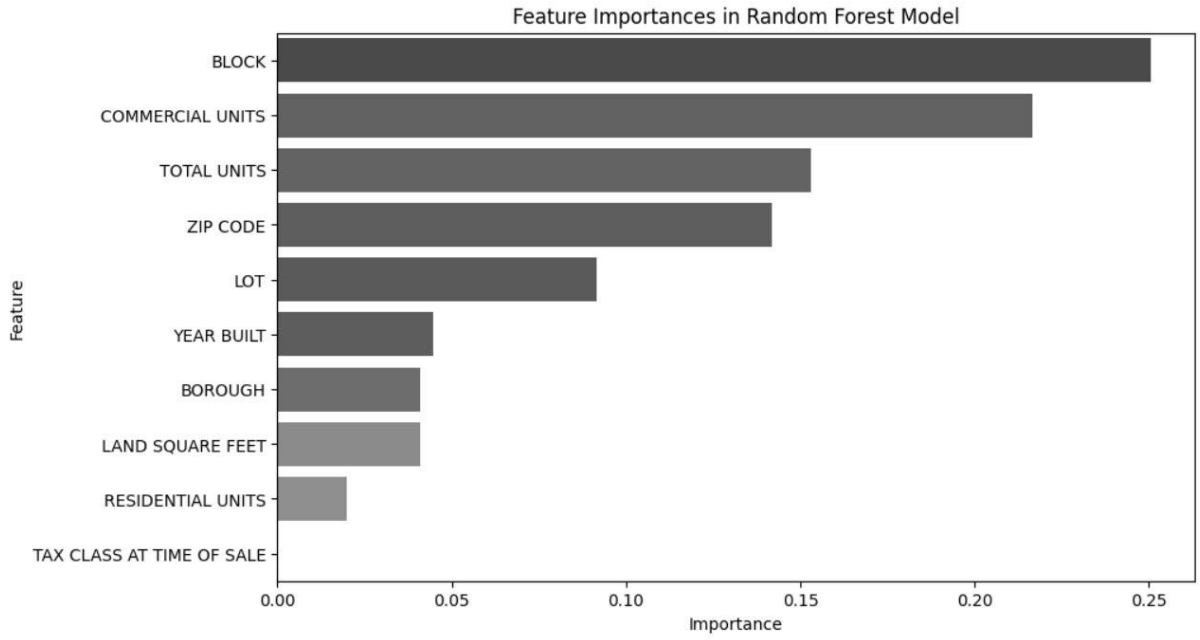
Candidate performance was mixed on this task, with most candidates receiv ing partial credit. Candidates received varying amounts of partial credit for omitting a comparison or making incorrect statements about each type of plot.

Answer:

Feature importance and partial dependence plots are both methods used to interpret "black box" models like random forests.

Feature importance plots show the contribution of each feature in the model structure. They can be used to rank which features are most important in fitting the model structure. A partial dependence plot shows the impact that a specific variable has on the final prediction, without considering the model structure.

Your assistant has built a random forest and produced a feature importance chart:



(b) (*2 points*) Describe the graph and interpret its meaning.

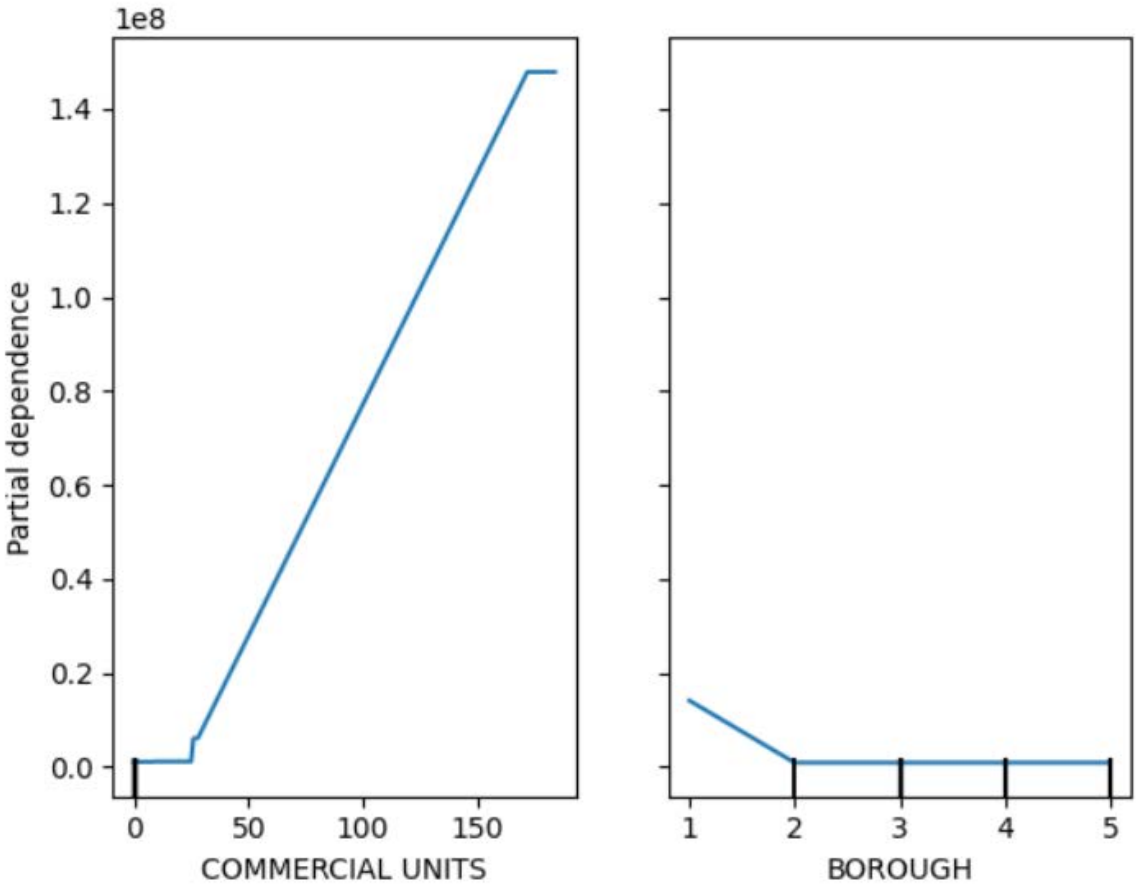
Most candidates performed well on this task. The most common reason for partial credit was simply describing the graph, and not reasoning through the meaning of values in the chart.

ANSWER:

The feature importance plot analyzes the structure of the model, ranking the contribution of each feature. Features higher up in the list resulted were more important in splitting the nodes. The feature importance values in this chart sum to 1 to indicate the relative magnitude of each contribution of each element in an easy to interpret format.

Here we see that **BLOCK** is very important for the random forest model in predicting final sale price. This might be due to outliers in certain blocks, that the model is able to carve out. Next we have Commercial Units which seems to indicate that commercial buildings may have outsized effects on the final prediction value.

Your assistant continues by building partial dependence plots on the variables **Commercial Units** and **Borough**. Note that both variables are encoded as numeric values in the model.



(c) (*3 points*) Describe a partial dependence plot graph and interpret the specific meaning of the graphs shown above.

Candidates struggled with this task overall. For full credit, responses needed to address the interpretation of the predictor in each chart and recognize the problem with borough being encoded as a number.

ANSWER:

The partial dependence plot shows the relative impact that each variable has on the final predicted value. As the value of the variable increases along the x-axis, the relative impact to the final predicted value will be shown on the y-axis.

From the first plot we see that as commercial units in the property increase the sale value increases monotonically. It also seems to cap out at around 175 units where additional commercial units are not adding to the property's sale price.

In the second plot borough is encoded as a numeric variable, making the plot somewhat difficult to interpret. We can see that borough 1 (Manhattan) seems to have the most positive impact on price.

Task 12 (9 *points*)

(a) (*2 points*) Explain how high multicollinearity affects coefficient estimates and their standard errors in a GLM.

Candidate performance was mixed on this task. Most candidates correctly identified the impact to standard errors. However, many candidates were not able to explain how this affects the coefficient estimates.

ANSWER:

High multicollinearity inflates standard errors for the variables and results in unstable GLM coefficient estimates. The impacts interpretability of magnitude, direction, and statistical significance of coefficient estimates.

As a consultant for an urban planning analytics firm, you've been tasked with analyzing historical property sale data. The goal is to predict sale prices based on property attributes. The client is particularly concerned about multicollinearity (including **Gross Square Feet**, **Land Square Feet**, and **Total Square Feet** – the sum of **Gross Square Feet** and **Land Square Feet**) and wants you to explore GLMs and penalized regression to address overfitting and collinearity.

```
Call:
lm(formula = log_sale ~ log_gross + log_land + log_total, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-16.5915  -0.0315   0.3082   0.6757   5.0708

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.18820    0.16998  48.170   <2e-16 ***
log_gross    0.56737    0.06709   8.457   <2e-16 ***
log_land    -0.06301    0.08203  -0.768    0.442
log_total    0.12854    0.13277   0.968    0.333
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.983 on 29325 degrees of freedom
Multiple R-squared:  0.05413,   Adjusted R-squared:  0.05403
F-statistic: 559.4 on 3 and 29325 DF,  p-value: < 2.2e-16
```

You are provided with the variance-inflation factor for each of the three predictors:

log_gross	log_land	log_total
18.73878	16.46342	48.29336

(b) (*2 points*) Define variance-inflation factor and interpret the results from the table above.

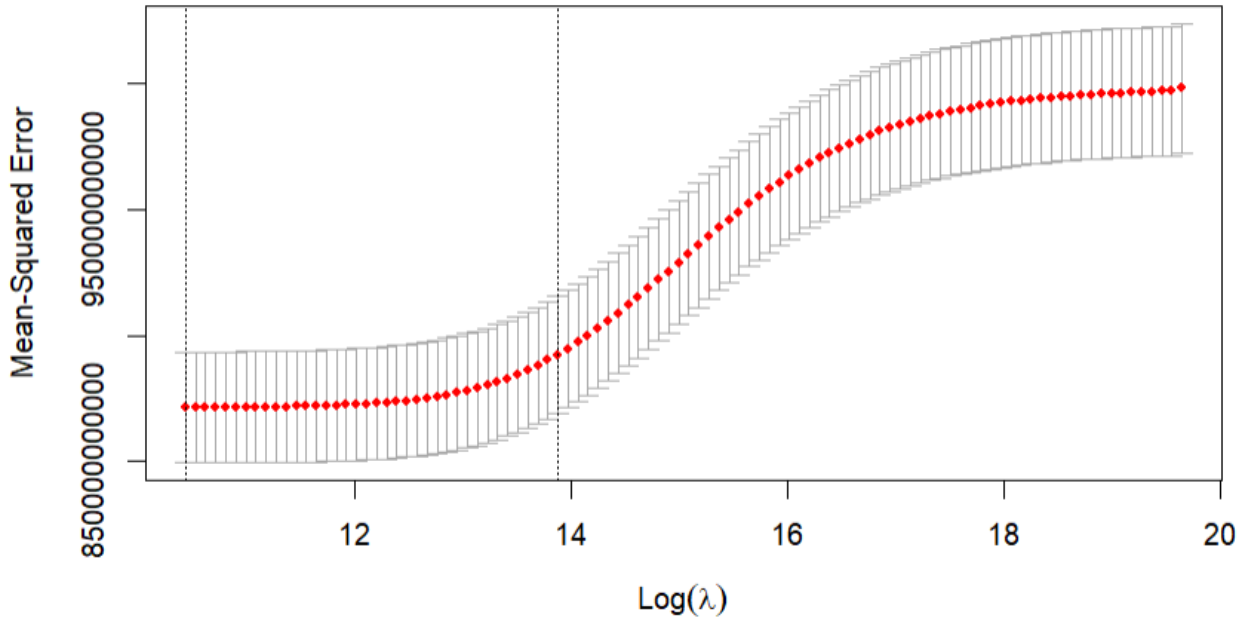
Candidates struggled with this task. Very few candidates could correctly define variance-inflation factor. Similarly, few candidates recognized that all three variables have high multicollinearity due to having a VIF > 5-10. Many candidates observed that log_total has the highest VIF, but few candidates were able to interpret this. Some candidates confused VIF with variance importance plots.

ANSWER:

Variance inflation factor is the ratio of the variance of a predictor when fitting the full model divided by the variance of the same predictor if fit on a standalone model.

All 3 variables have VIF>>10, meaning high multicollinearity exists among these predictors.

Your assistant tries to apply ridge regression to this problem, and provides you with the following plot showing cross validated mean squared error against the log of the ridge penalty parameter:



(c) (*2 points*) Describe how a ridge regression model changes as λ increases. Your description should outline how to interpret a model with $\lambda = 0$ and how to interpret a model with a very large value of λ.

Although many candidates provided full-credit responses to this task, performance was mixed overall. To receive full credit the candidate needed to correctly state what happens for both extremes of lambda. Many candidates provided wrong explanations of one or both extremes. Some candidates stated that a large lambda was 1, which indicated the candidate may have been confusing lambda with the L1/L2 penalty of ridge vs lasso regression.

ANSWER:

As the penalty parameter gets larger, the coefficients get pushed toward 0 until the model consists of only an intercept. As the penalty parameter gets smaller, the coefficients are allowed to be larger and at zero becomes identical to an unpenalized GLM.

Your assistant provides you with a comparison of model coefficients from 3 GLM models: OLS, Ridge regression and LASSO. Your assistant is concerned that some coefficients (highlighted in the table) increase from the OLS model when penalized models are supposed to shrink them.

Variable	OLS	Ridge	LASSO
<chr>	<dbl>	<dbl>	<dbl>
(Intercept)	7.9366047	7.9807619	7.9755884
log_gross	0.5481599	0.5119857	0.6665392
log_land	-0.1634434	-0.1416294	0.0000000
log_total	0.2986351	0.3055904	0.0365505

(d) (*3 points*)

i. Explain the reason that **log_land** coefficient is 0 in the LASSO model.
ii. Explain the reason that the 2 highlighted coefficients in the Ridge and LASSO models increase from the OLS model.

Candidate performance was mixed on this task. The majority of candidates answered I correctly. To receive full credit on part ii, the candidate needed to state why each model had a different impact on the highlighted coefficients due to multicollinearity.

ANSWER:

i. LASSO's penalty function allows for coefficients to be reduced to exactly to 0, thus performing variable selection.

ii. When regularized regression models encounter multicollinearity, they will tend to allocate the coefficient into stronger predictors while shrinking/removing other coefficients.
Ridge regression doesn't zero out coefficients. It balances weights, shrinking log_gross and log_land closer to zero, and shifting the coefficient to stronger variable log_total.
Similarly, LASSO sets log_land to 0 and log_total very close to 0, and reallocates the explanatory power to log_gross.