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

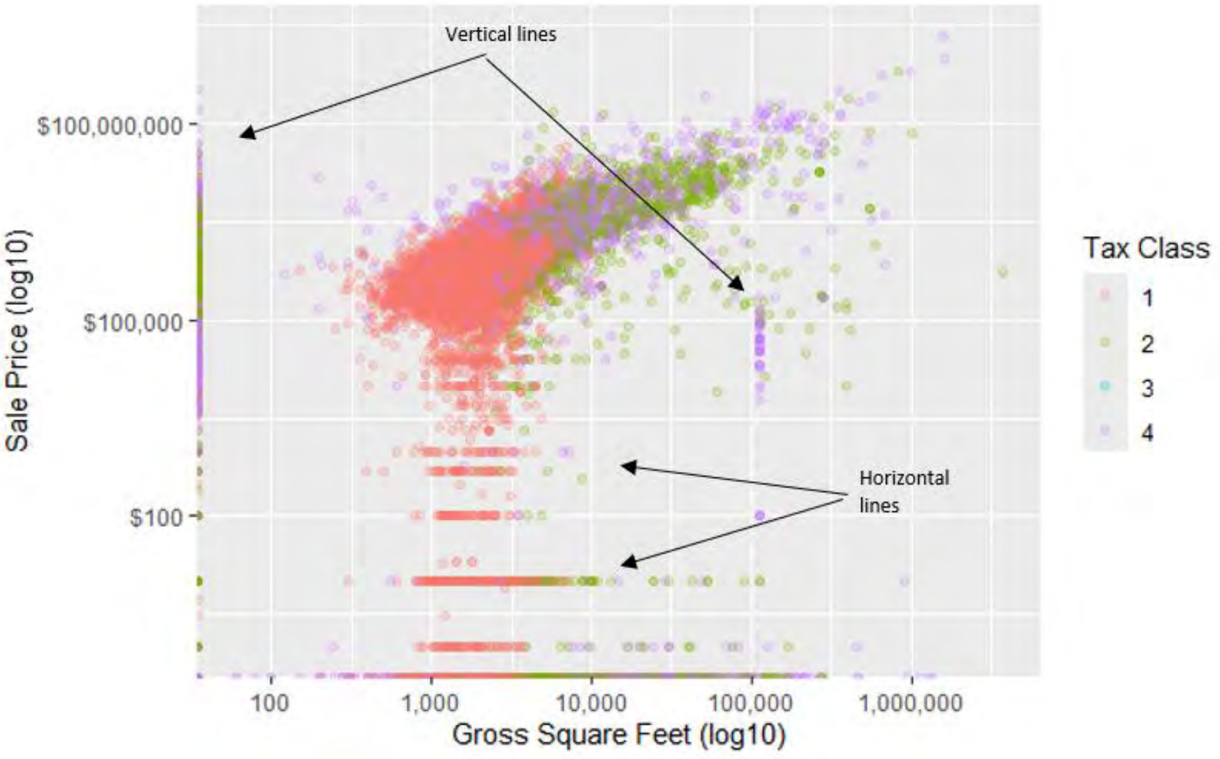
Data Dictionary

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.

ANSWER:

(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.

ANSWER:

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).

ANSWER:

1.
2.
3.

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.

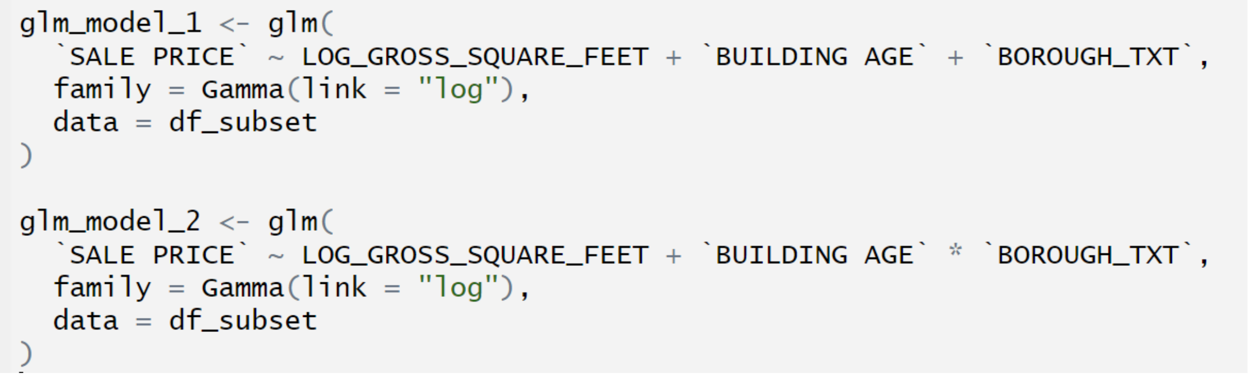
ANSWER:

i.
ii.
iii.

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:

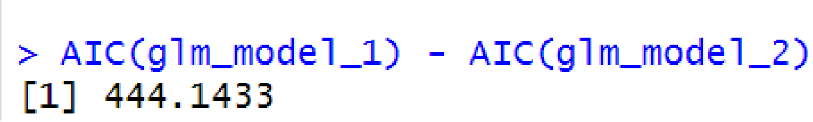


```
glm_model_1 <- glm(
  `SALE PRICE` ~ LOG_GROSS_SQUARE_FEET + `BUILDING AGE` + `BOROUGH_TXT`,
  family = Gamma(link = "log"),
  data = df_subset
)

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.

ANSWER:

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.

ANSWER:

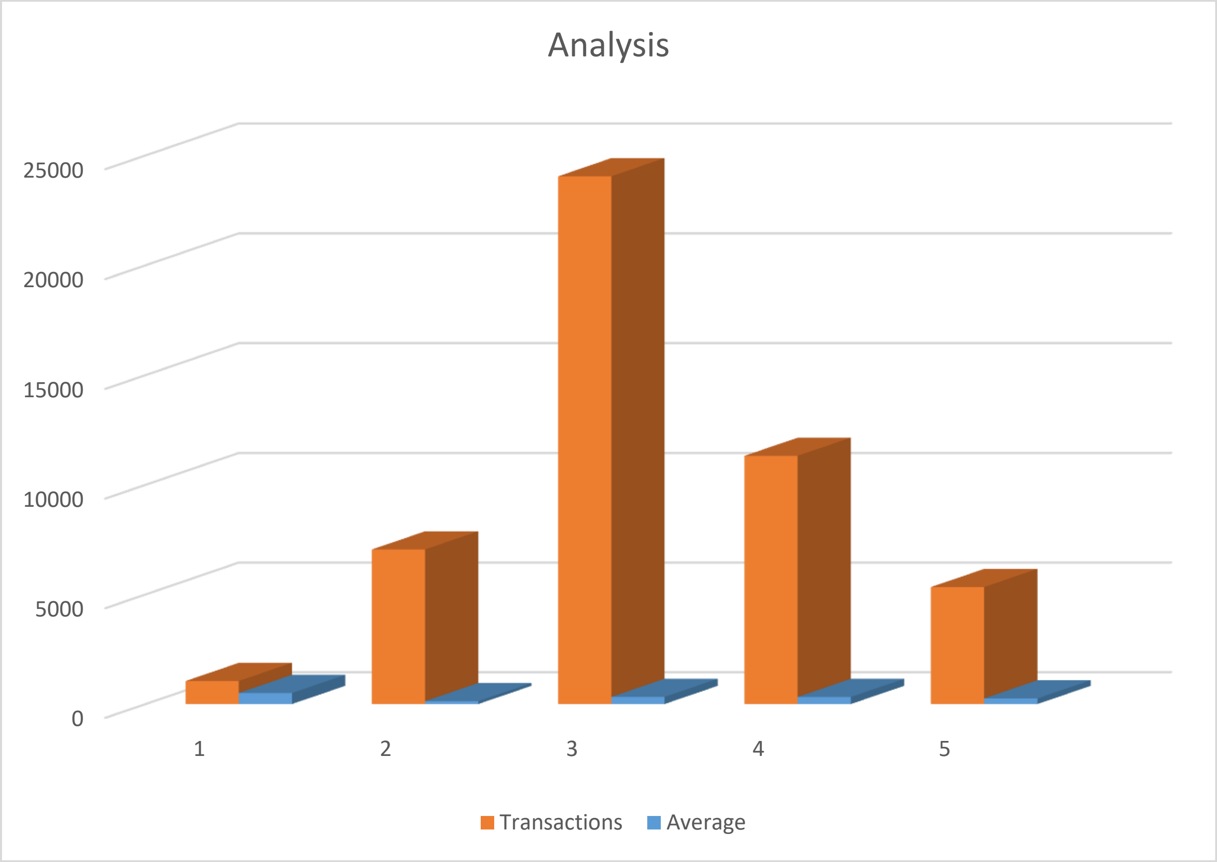
Task 4 (8 *points*)

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.

ANSWER:

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.

ANSWER:

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.

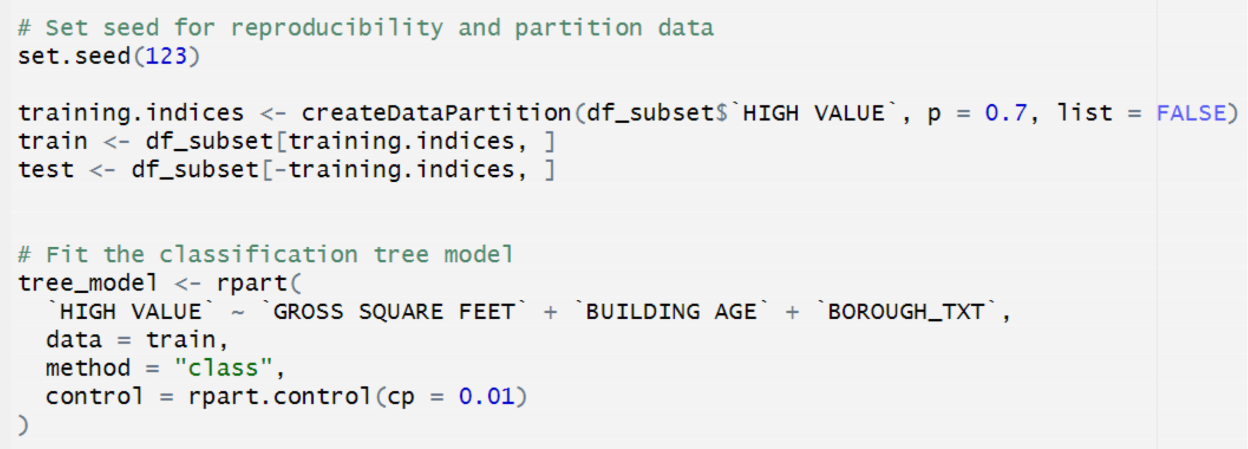
ANSWER:

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

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.

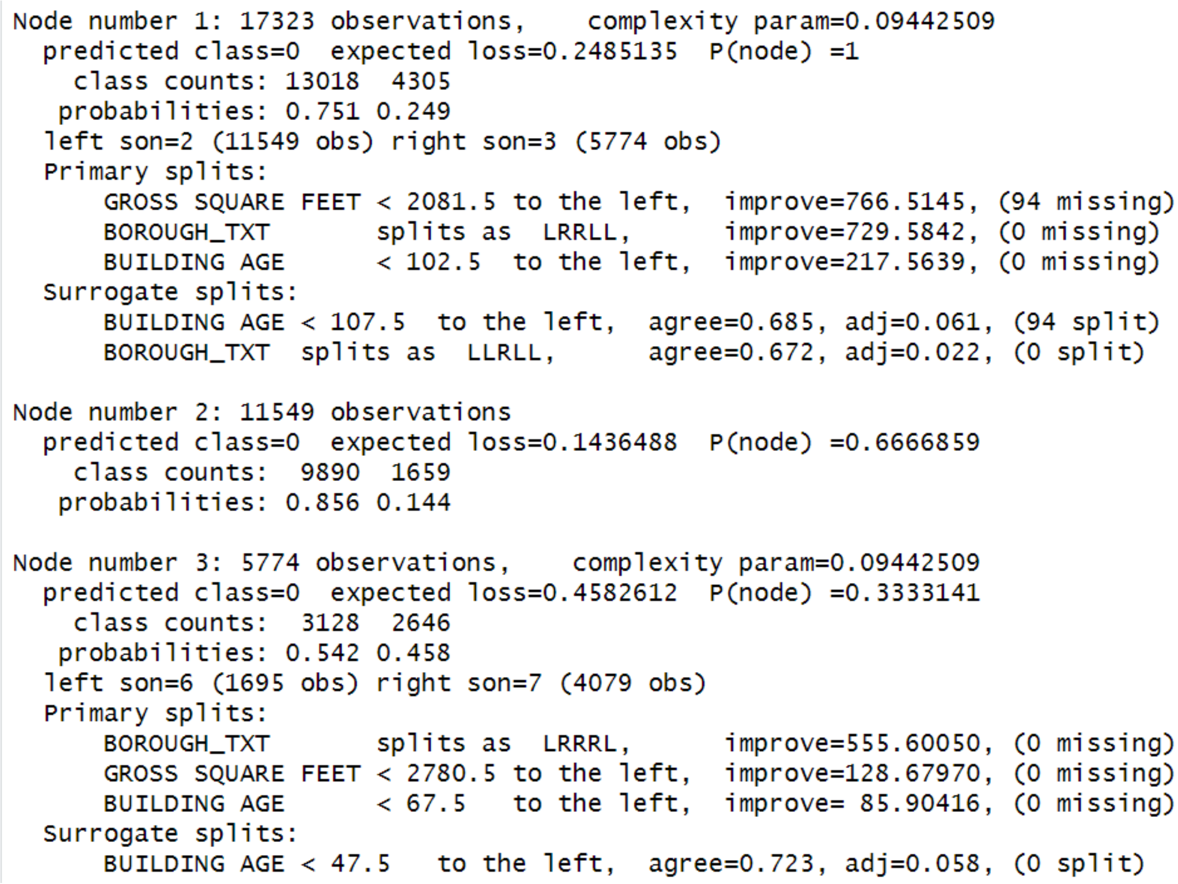


```
# Set seed for reproducibility and partition data
set.seed(123)

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?

ANSWER:

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)
```

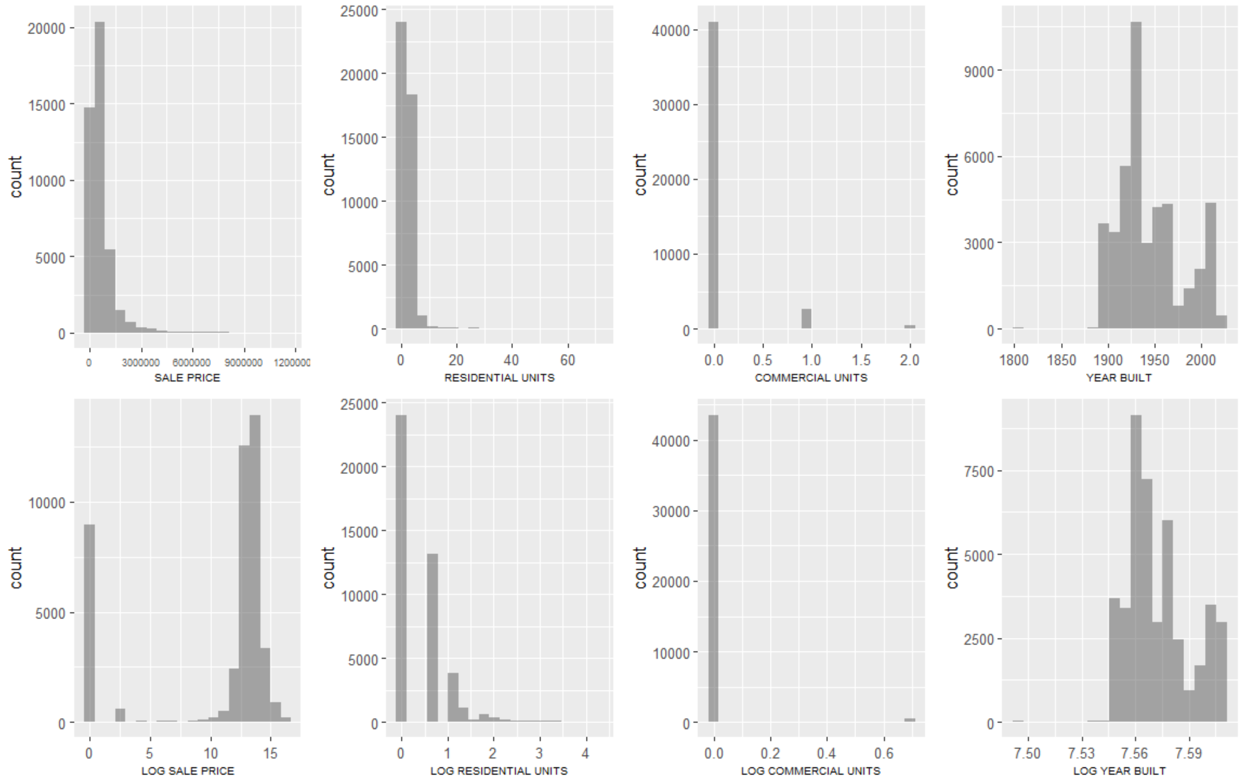
(b) (*2 points*) Explain how the model handles the missing observations on Node number 1.

ANSWER:

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.

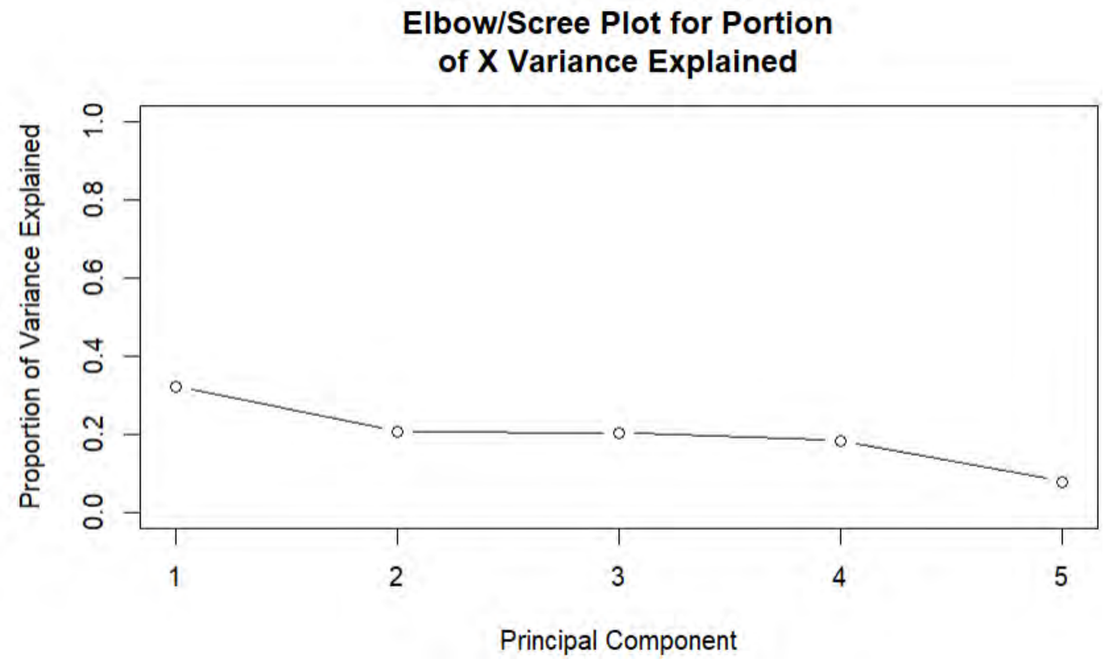
For several quantitative variables, the following histograms were produced, including histograms for the log version of each variable (in all cases values below 1 of the raw variables have been forced to equal 1 before taking the log):

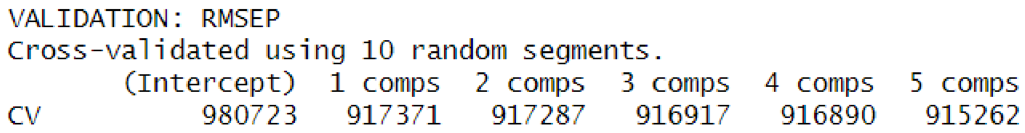


(a) (*2 points*) Identify the variables for which a log transformation appears to be helpful; in each case explain why it would be helpful.

ANSWER:

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:



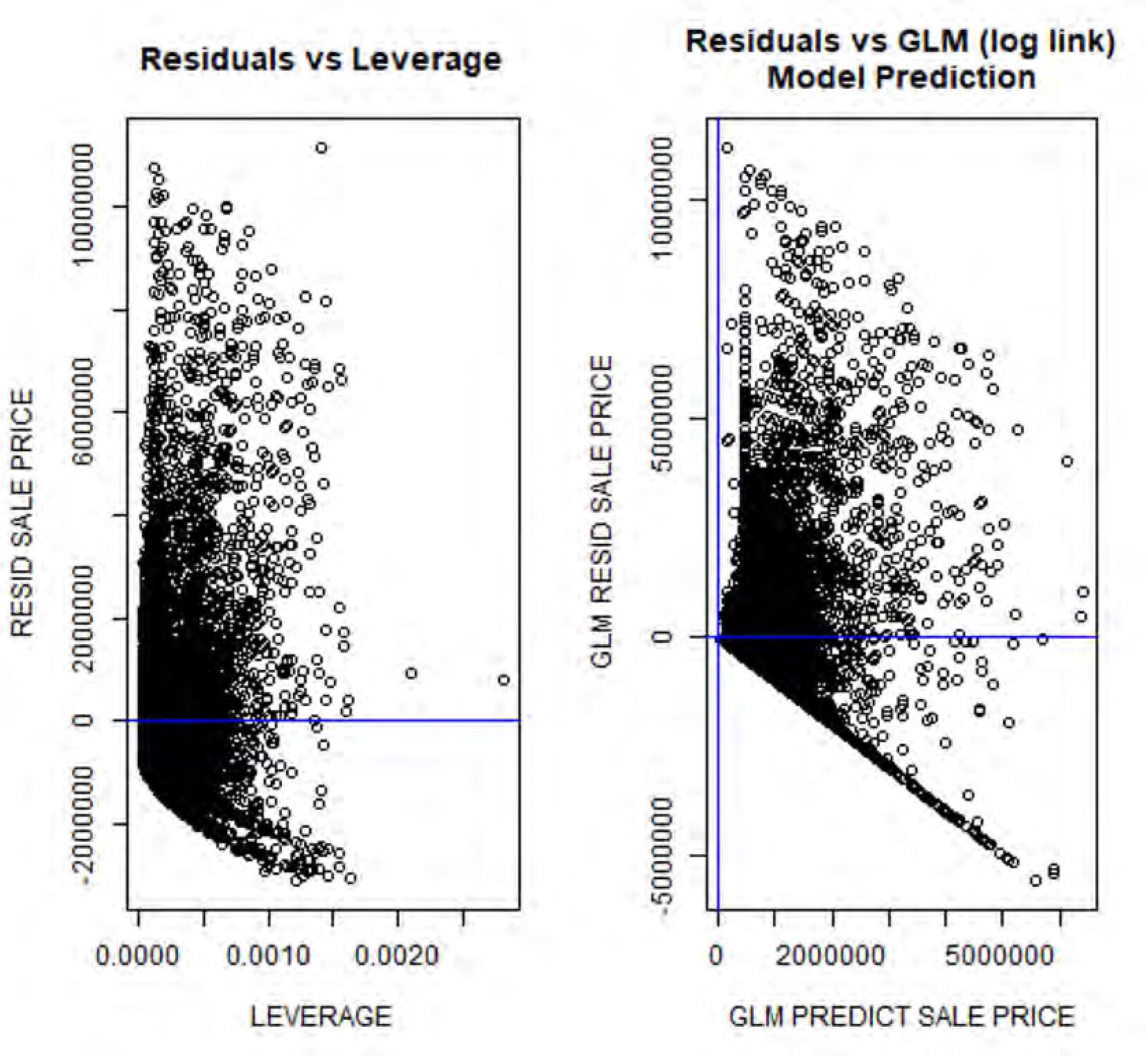


```
VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV          980723   917371   917287   916917   916890   915262
```

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

ANSWER:

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.

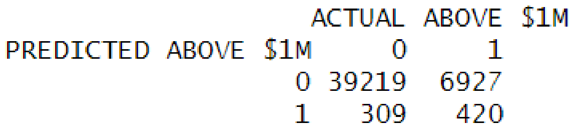
(c) (*3 points*) Describe what you would look for on residual and leverage plots and identify any concerns based on each of the plots above.

ANSWER:

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.

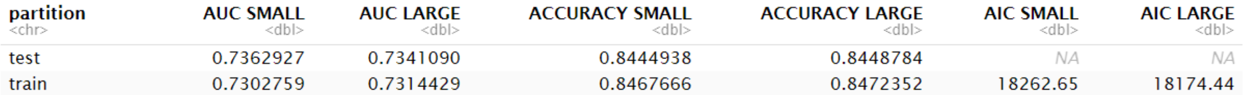
ANSWER:

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.

ANSWER:

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	AUC SMALL	AUC LARGE	ACCURACY SMALL	ACCURACY LARGE	AIC SMALL	AIC LARGE
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<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.

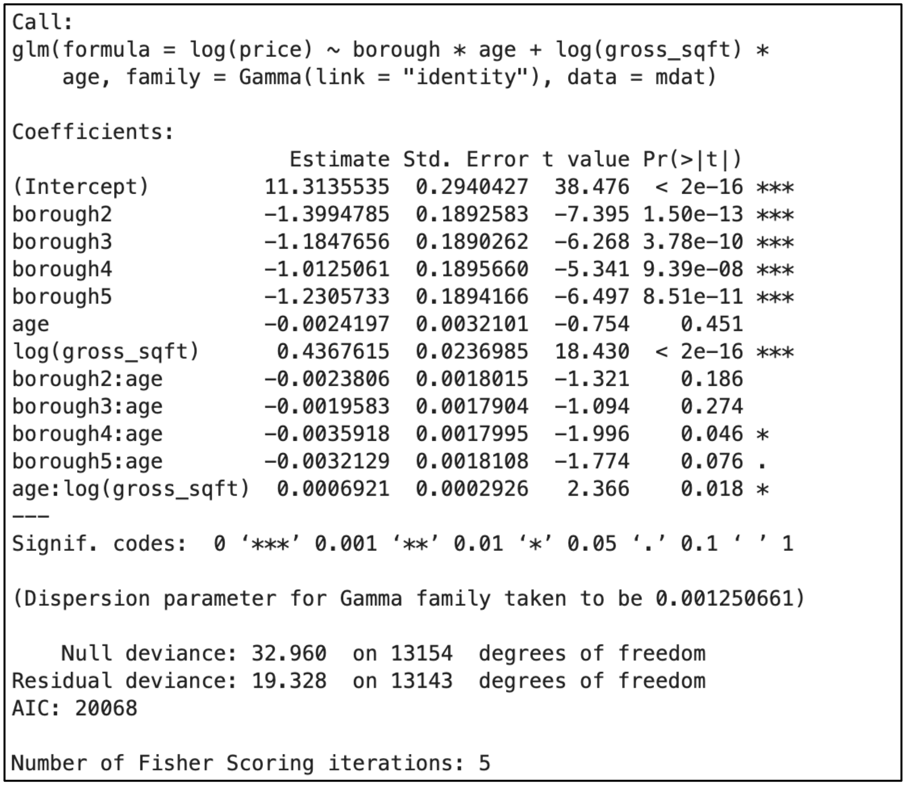
ANSWER:

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.

ANSWER:

	Price Increase with Age (Yes/No)
borough 1	
borough 2	
borough 3	
borough 4	
borough 5	

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.

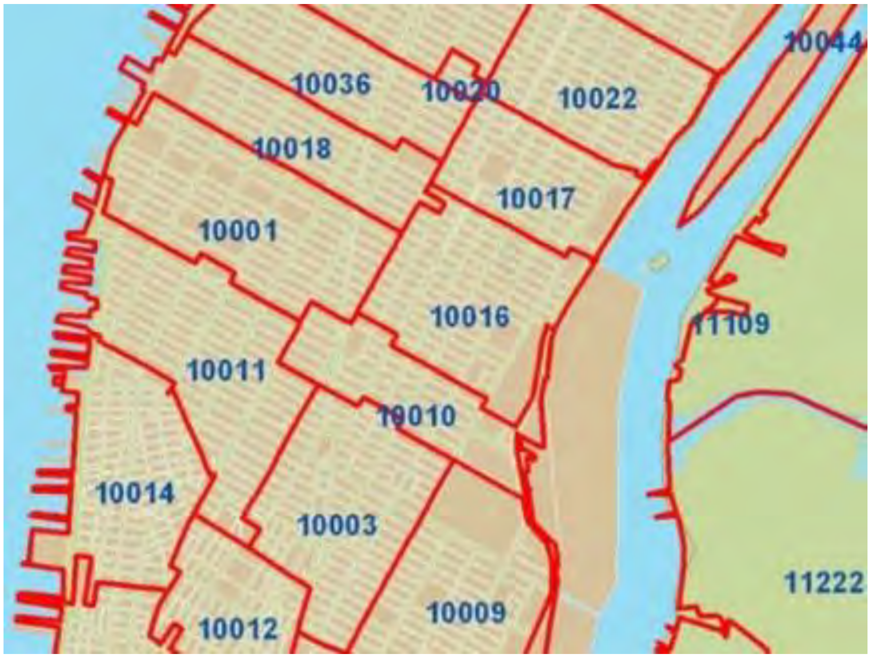
(a) *(1 point)* Provide one reason why using this as a factor variable could be problematic.

ANSWER:

(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.

ANSWER:

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.

ANSWER:

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

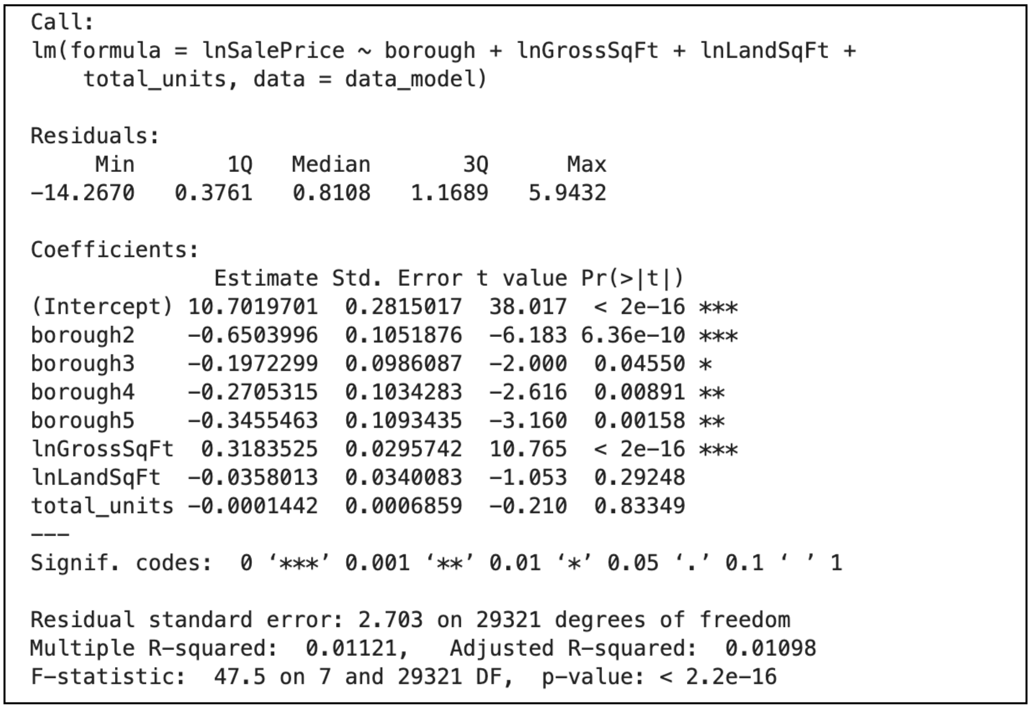
(d) (*1 point*) Describe one problem that arises from applying K-means clustering to the variables **Borough** and **Sale Price**.

ANSWER:

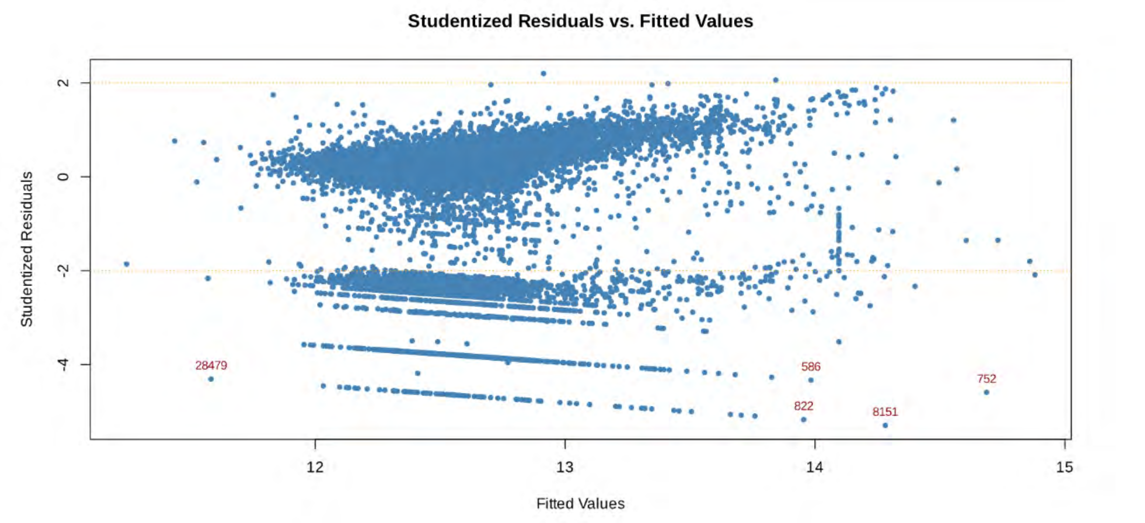
Task 10 (4 *points*)

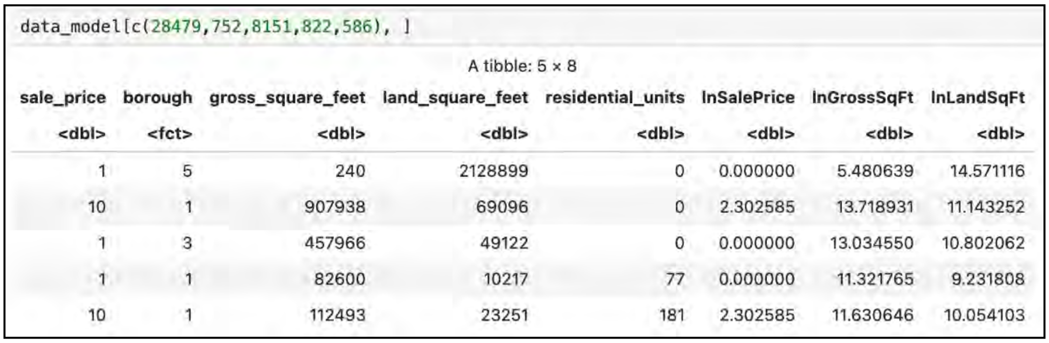
Your client is acquiring a portfolio of NYC properties and wants to use the NYC Rolling Sales dataset to build a linear model that predicts log (**Sale Price**) per **Gross Square Foot** using variables including **Borough**, **Year Built**, **Total Units**, **Land Square Feet** and **Gross Square Feet**. 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.

ANSWER:

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.

ANSWER:

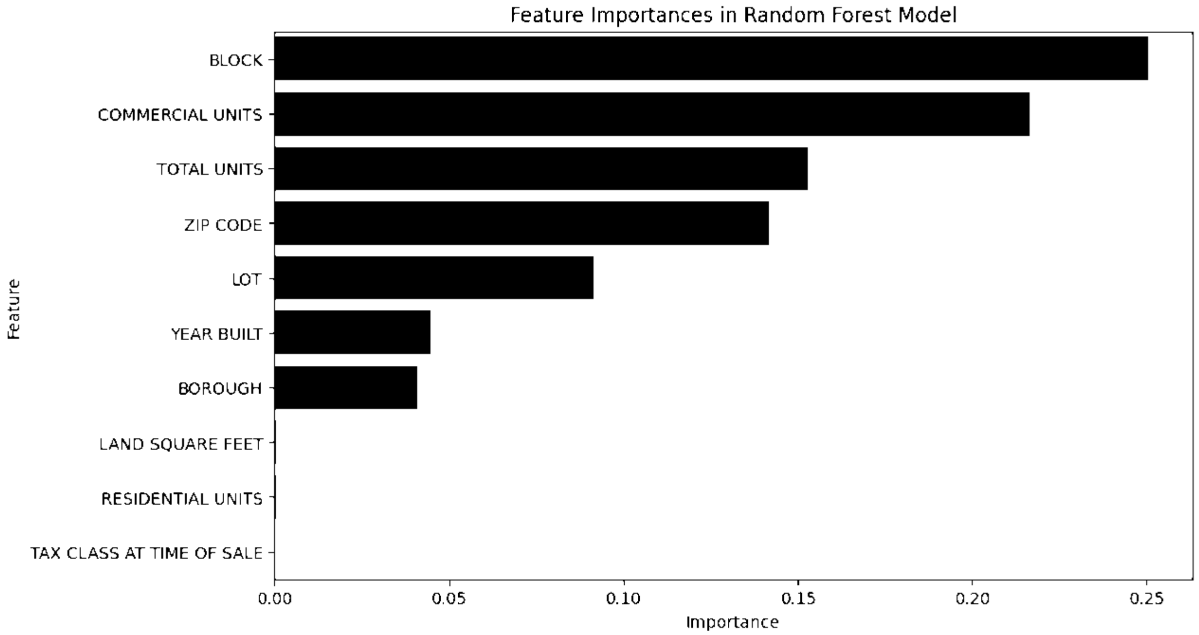
Task 11 (7 *points*)

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.

ANSWER:

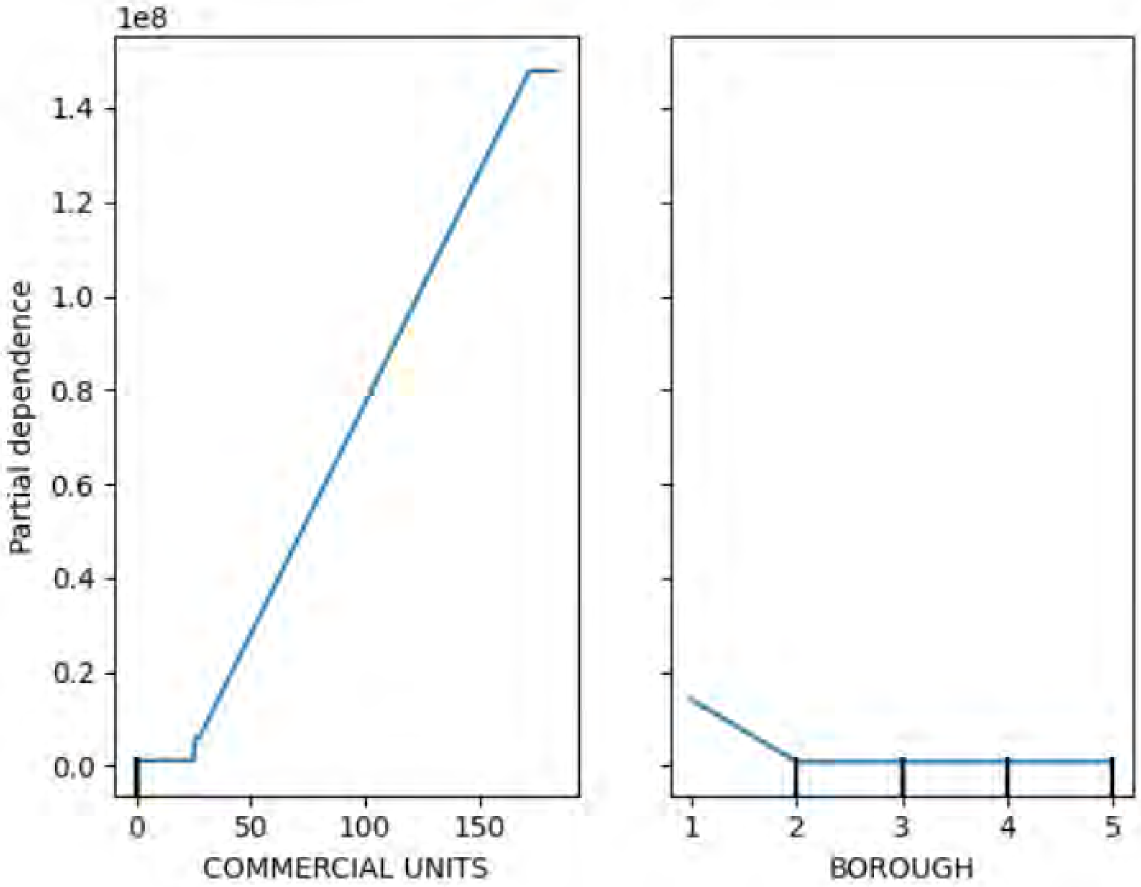
Your assistant has built a random forest and produced a feature importance chart:



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

ANSWER:

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.

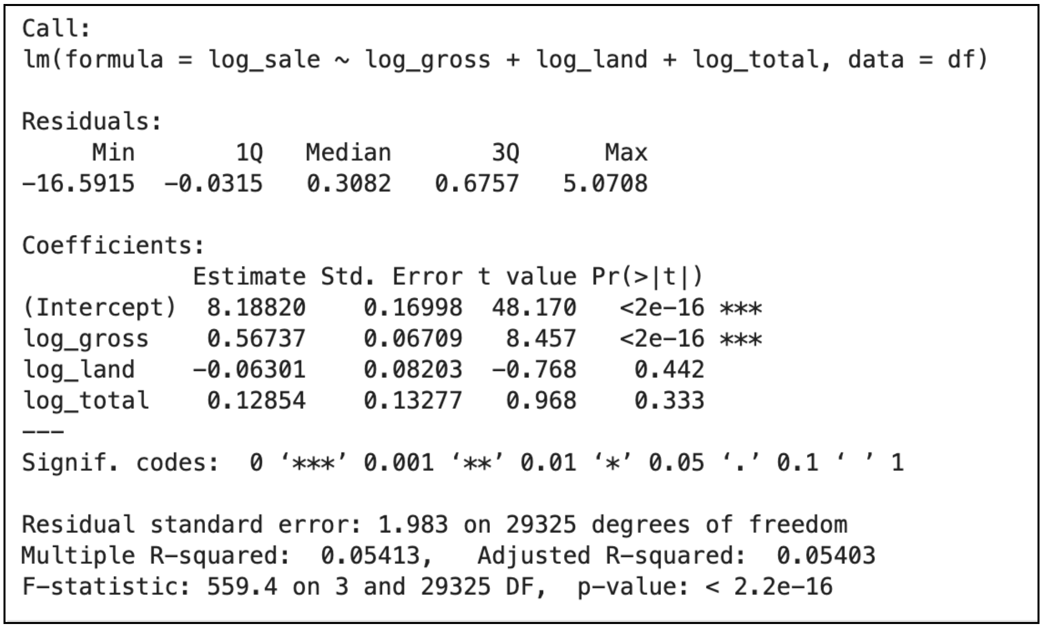
ANSWER:

Task 12 (9 *points*)

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

ANSWER:

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
```

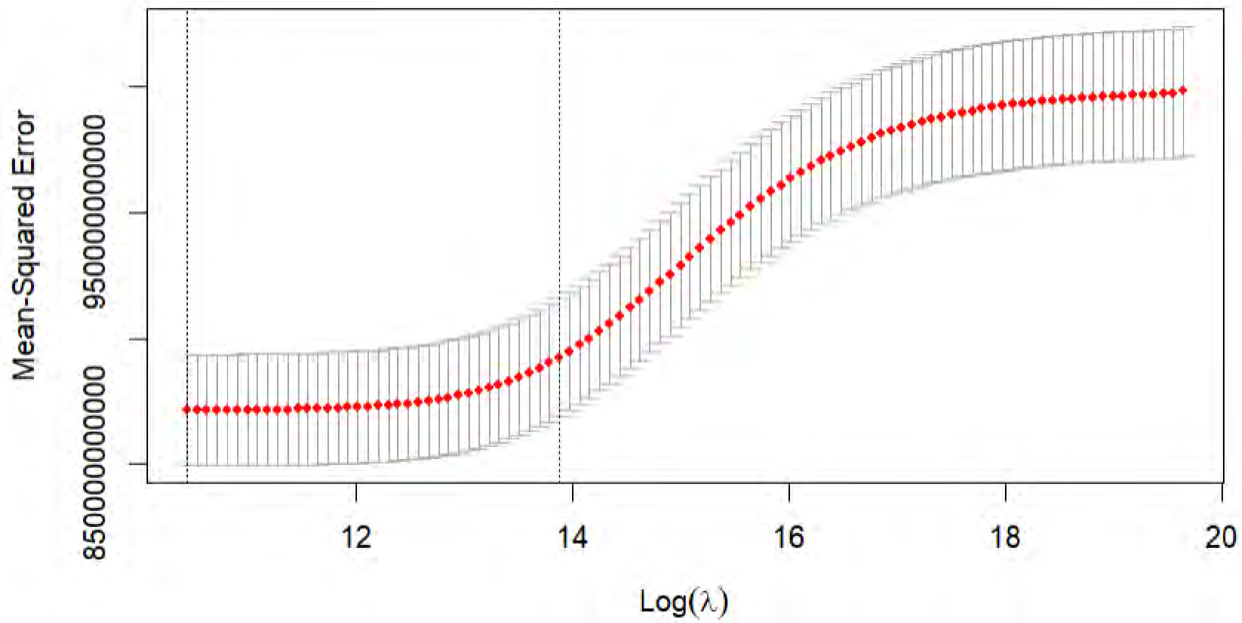
You are provided with the variance-inflation factor for each of the three predictors:

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.

ANSWER:

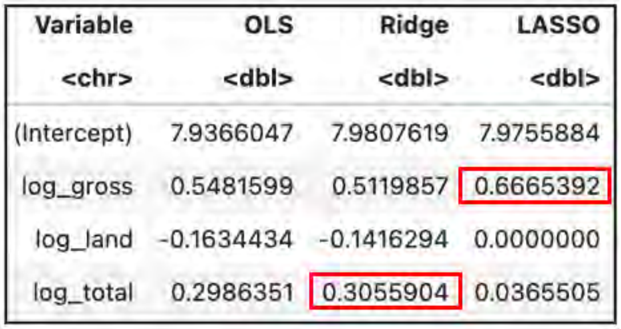
Your assistant tries to apply ridge regression to this problem, and provides you with the following plot showing cross validated mean squared error against the log of the ridge penalty parameter:



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

ANSWER:

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 <chr>	OLS <dbl>	Ridge <dbl>	LASSO <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.

ANSWER: