Exam PA October 17 Project Statement

IMPORTANT NOTICE – THIS IS THE OCTOBER 17, 2023 PROJECT STATEMENT. IF TODAY IS NOT OCTOBER 17, 2023, 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. For this exam there is no data file or .Rmd file provided. Neither R nor RStudio are available or required.

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 a consultant and your client has asked you to perform a study related to outcomes in university in the United States.

Your client is interested in better understanding the drivers of several key variables and developing models to predict them. These target variables include:

- *tuition prices*
- *students who are defaulting on student loans*
- *future earnings of students*
- *student loan repayment rates*
- *university admission rates*

To answer these questions, you decide to use a publicly available dataset[1] that includes aggregated data from 2,180 universities in the United States for the 2020-2021 school year.

[1] *Source: United States Department of Education*

Data Dictionary

Variable	Data Type: Range/Levels	Description
UNITID	Numeric : 100654 to 495767	ID for the institution
INSTNMH	String: N/A	Institution name
REGION	Factor: 10 levels	Region (IPEDS)
CONTROL	Factor: 3 levels (“Public”, “Private, non-profit”, ”Private, for-profit”)	Control of institution
LOCALE	Factor: 4 levels (“City”, “Suburb”, ”Town”, ”Rural”)	Locale of institution
ADMIT_TIER	Factor: 5 levels ("MOST SELECTIVE", "EXTREMELY SELECTIVE", "VERY SELECTIVE", "MODERATELY SELECTIVE", "NOT SELECTIVE")	How selective the institution is
TEST_REQ	Factor: 4 levels ("Required", "Recommended", "Neither required nor recommended", "Considered but not required")	Does the institution require standardized tests
ADM_RATE	Numeric : 0.0244 to 1.0	Admission rate
SATVRMID	Numeric: 395 to 760	Midpoint of SAT critical reading scores
SATMTMID	Numeric: 350 to 795	Midpoint of SAT math scores
SATWRMID	Numeric: 280 to 765	Midpoint of SAT writing scores
UGDS	Numeric: 2 to 109,233	Number of undergraduate certificate/degree-seeking students
SCHOOL_SIZE	Factor: 3 levels (“Small”, “Medium”, ”Large”)	The size of the university based on number of students
TUITIONFEE_IN	Numeric: 480 to 61,671	In-state tuition and fees
TUITIONFEE_OUT	Numeric: 480 to 61,671	Out-of-state tuition and fees

AVGFACSAL	Numeric: 547 to 21,143	Average faculty salary per month
PFTFAC	Numeric: 0.0339 to 1.0	Proportion of faculty that is full-time
PCTPELL	Numeric: 0.0054 to 1.0	Percentage of undergraduates who receive a Pell Grant
PCTFLOAN	Numeric: 0.0015 to 1.0	Percent of undergraduate students receiving a federal student loan
MD_EARN_WNE_P10	Numeric: 13,438 to 132,969	Median earnings of students working and not enrolled 10 years after entry
COMPL_RPY_7YR_RT	Numeric: 0.2059 to 0.9814	Seven-year repayment rate for completers
NONCOM_RPY_7YR_RT	Numeric: 0.1130 to 0.9314	Seven-year repayment rate for non-completers
GRAD_DEBT_MDN	Numeric: 2,334 to 48,148	The median debt for students who have completed
WDRAW_DEBT_MDN	Numeric: 2,352 to 24,167	The median debt for students who have not completed
COSTT4_A	Numeric: 5,663 to 81,531	Average cost of attendance
CDR3	Numeric: 0.001 to 0.357	Three-year cohort default rate
LOAN_EVER	Numeric: 0.0139 to 0.9856	Percent of students who received a federal loan while in school
AGE_ENTRY	Numeric: 17.43 to 51.60	Average age of entry into the institution
FEMALE	Numeric: 0.04156 to 0.97957	Share of female students
MARRIED	Numeric: 0.0027 to 0.8154	Share of married students
FIRST_GEN	Numeric: 0.08867 to 0.85091	Share of first-generation students
MD_FAMINC	Numeric: 1,680 to 179,864	Median family income

Task 1 (5 *points*)

Your client wants to understand the factors influencing university admission rates. Your client is interested in ensuring that the analysis has proportional representation with respect to different regions of the country (**REGION**) and population densities (**LOCALE**).

(a) (*3 points*) Describe the steps for developing a stratified sample based on your client's goals.

ANSWER:

Your client is also interested in student opinions about the university. You are given a dataset with written responses to a university satisfaction survey.

(b) (*2 points*) Discuss the advantages and disadvantages of using this kind of unstructured data in a predictive model.

ANSWER:

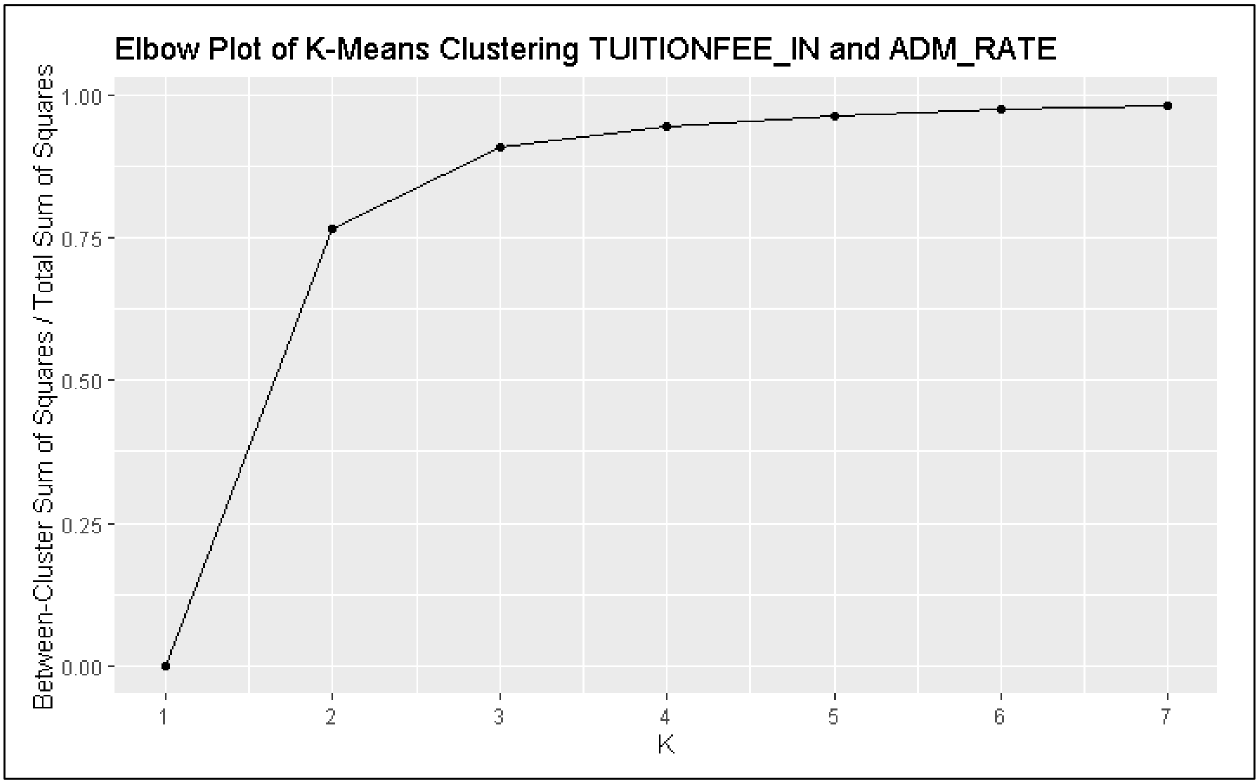
Task 2 (11 *points*)

Your assistant is interested in understanding the relationship between the features admission rate (**ADM_RATE**) and in-state tuition (**TUITIONFEE_IN**) and is considering whether to perform a K-means analysis or a hierarchical clustering analysis to better understand the relationship.

(a) (*4 points*) Describe two similarities and two differences between K-means clustering and hierarchical clustering.

ANSWER:

Your assistant prepared an elbow plot of K-means clustering using the in-state tuition (**TUITIONFEE_IN**) and admission rate (**ADM_RATE**) features, shown below.



(b) (*3 points*) Explain the tradeoff between selecting a value of K=2 and K=4. Recommend a value for K and justify your recommendation.

ANSWER:

Your assistant now wants to include more features in the K-means clustering analysis and is suggest adding the following five variables:

CONTROL	Factor: 3 levels (“Public”, “Private, non-profit”, ”Private, for-profit”)	Control of institution
ADMIT_TIER	Factor: 5 levels ("MOST SELECTIVE", "EXTREMELY SELECTIVE", "VERY SELECTIVE", "MODERATELY SELECTIVE","NOT SELECTIVE")	How selective the institution is
PFTFAC	Numeric: 0.0339 to 1.0	Proportion of faculty that is full-time
MARRIED	Numeric: 0.0027 to 0.8154	Share of married students
FIRST_GEN	Numeric: 0.08867 to 0.85091	Share of first-generation students

(c) (*4 points*) Critique your assistant’s suggestion to add these features to the K-means analysis. Include at least three considerations in your critique.

ANSWER:

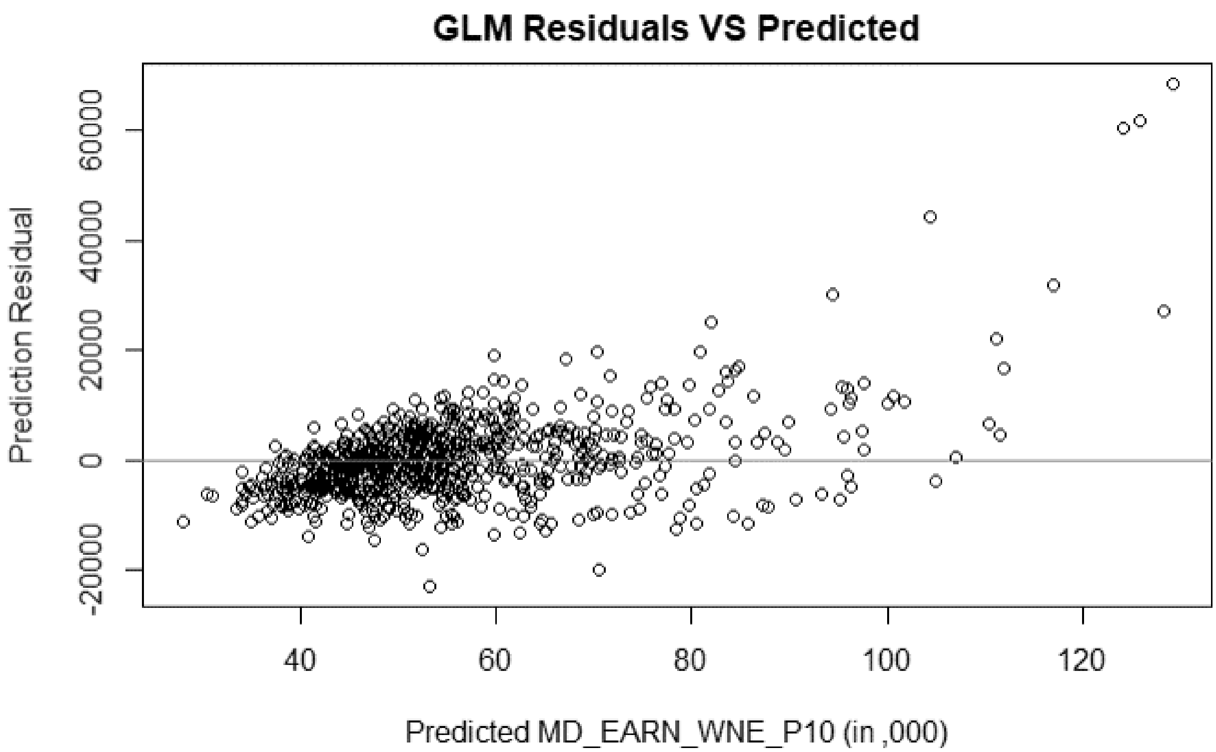
Task 3 (7 *points*)

You are investigating variables that impact the expected future earnings of students. The variable (**MD_EARN_WNE_P10**) represents the median earnings of students working and not enrolled 10 years after entry.

(a) (*3 points*) Explain three differences between fitting a normal linear regression to log(MD_EARN_WNE_P10) compared to fitting a GLM with a log link function to the unaltered MD_EARN_WNE_P10 variable.

ANSWER:

A GLM was fit on the formula MD_EARN_WNE_P10 ~ SATVRMID + SATMTMID + MD_FAMINC + AVGFACSAL. The plot of residuals vs. predicted values is as follows.



(b) (2 *points*) Analyze the residual plot.

ANSWER:

You are asked to construct a model to predict which individual students will default on their student loans.

(c) (*2 points*) Evaluate if this university dataset is appropriate for developing a model for predicting individual student loan defaults. No points will be awarded for referencing ethical issues as part of your evaluation.

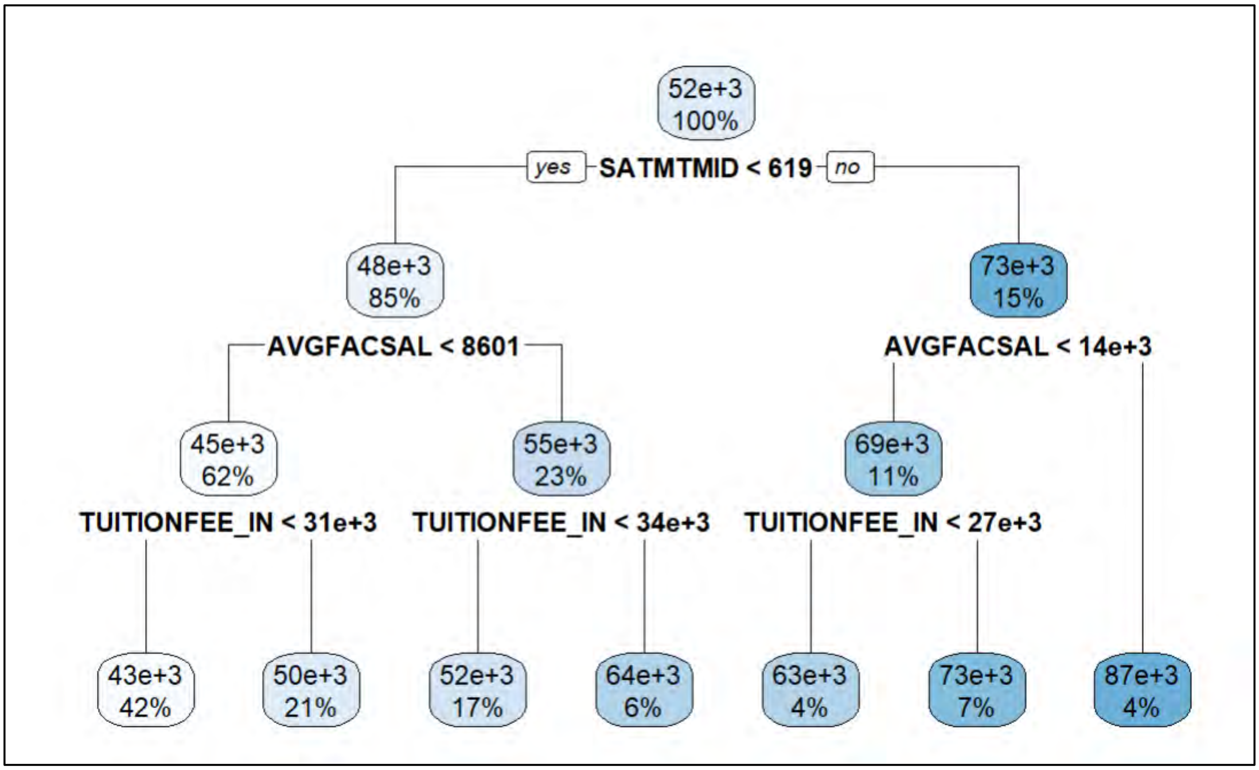
ANSWER:

Task 4 (6 *points*)

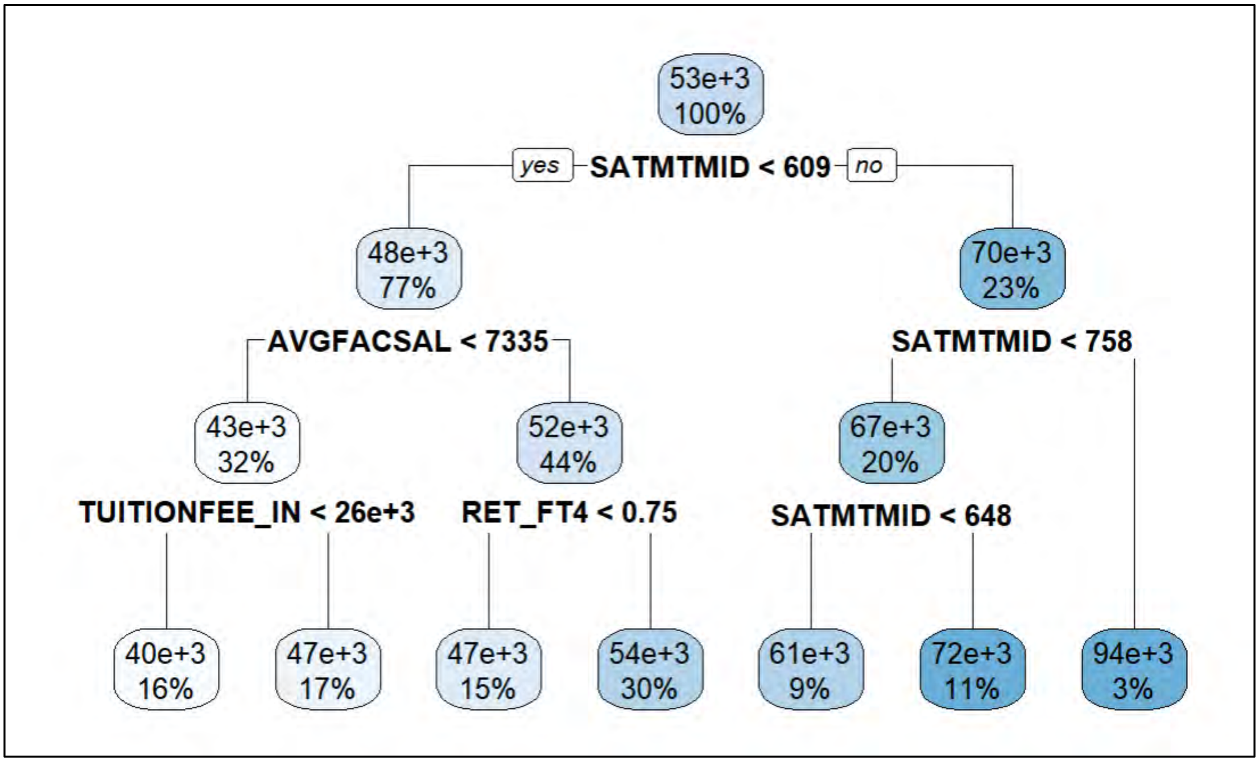
Your boss is interested in ranking the list of universities by their median earnings of students (**MD_EARN_WNE_P10**). Your Boss asks you and your Assistant to build tree-based models to predict this variable.

Your assistant creates two decision trees. Each was trained on the data using a bootstrap sample obtained without replacement. You are also provided one row from the testing data set.

Decision Tree One:



Decision Tree Two:



Note: When reading the decision trees assume all node values are rounded to the nearest thousand. For example, "63e+3" should be interpreted as 63,000.

One row of data from the Testing data set:

MD_EARN_WNE_P10	TUITIONFEE_IN	SATMTMID	SATVRMID	RET_FT4	PCTPELL	AVGFACSAL
32,084	11,068	462	485	0.6202	0.7368	7,194

(a) (*4 points*) Calculate the change in the **Absolute Error**, using the testing data row, between the first decision tree model and building a bagged model using both decision trees. State which of these two approaches yields a better result for this observation. Show all work.

ANSWER:

Your assistant is building a decision tree and does not fully understand the Complexity Parameter used in the process. You generate the Complexity Parameter (CP) table below.

	CP	nsplit	rel error	xerror	xstd
1	0.42270843	0	1.0000000	1.0026889	0.11814785
2	0.08830370	1	0.5772916	0.5988034	0.07665626
3	0.04546636	2	0.4889879	0.5240866	0.06625851
4	0.04289667	3	0.4435215	0.5359595	0.07134774
5	0.03555779	4	0.4006248	0.5138209	0.07066951
6	0.01434997	5	0.3650670	0.4381055	0.06972288
7	0.01000000	6	0.3507171	0.4673743	0.07928587

(b) (*2 points*) Interpret the Complexity Parameter table. Recommend and justify a CP value to use for the model.

ANSWER:

Task 5 (9 *points*)

Your client is interested in obtaining a deeper understanding of how tuition prices and the size of the universities are reflected in the dataset.

(a) (*3 points*) Suggest two numerical variables from the Data Dictionary for this analysis and describe two univariate technique that can be used to explore them.

ANSWER:

(b) (*2 points*) Suggest a categorical variable from the Data Dictionary for this analysis and describe a univariate technique to explore the variable.

ANSWER:

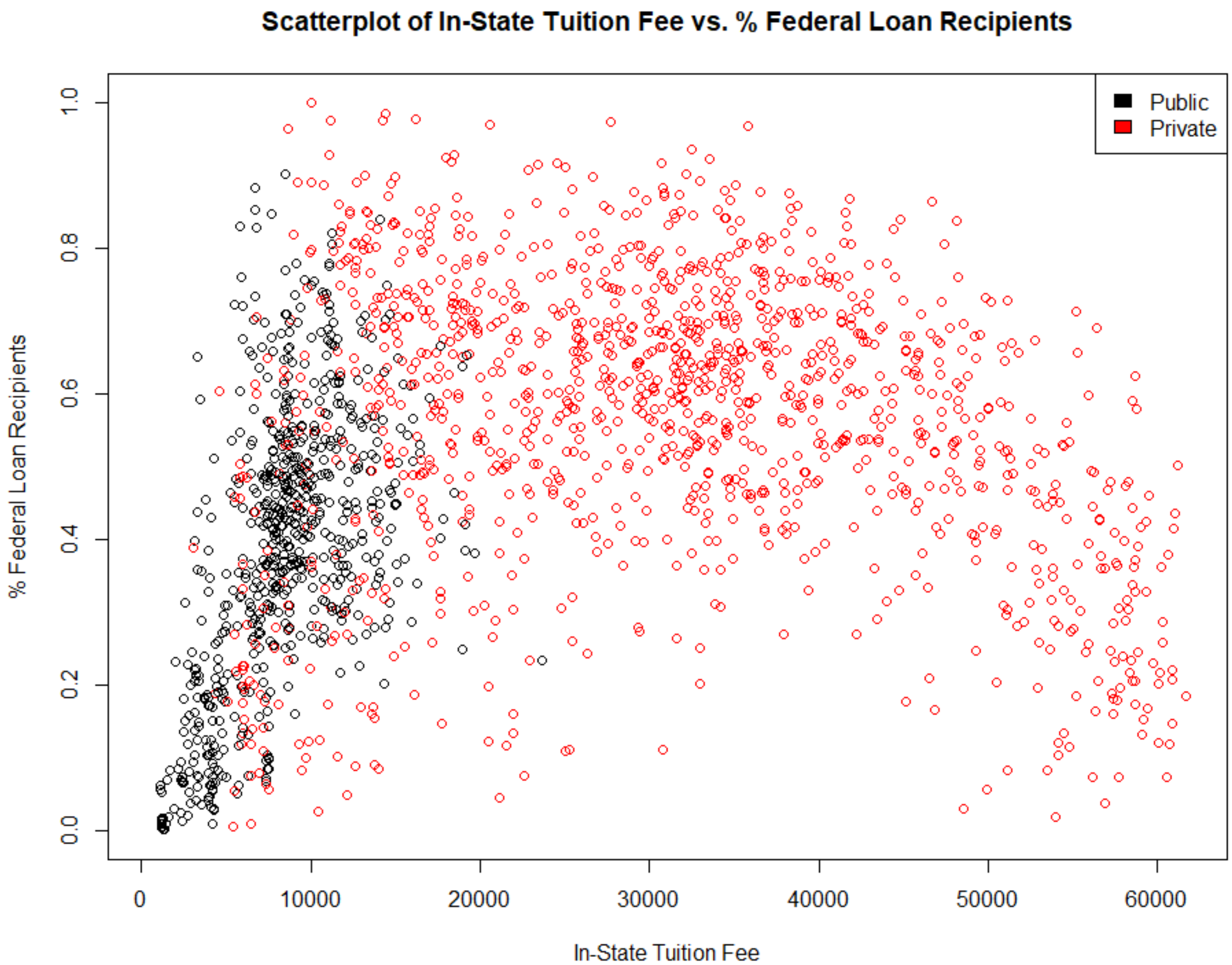
Your client is interested in obtaining a deeper understanding of the relationship between tuition prices and the university's size.

(c) *(2 points)* Describe a bivariate visualization that can be applied to understand the relationship between a numeric variable and a categorical variable.

ANSWER:

Your client is interested in the relationship between tuition prices and the percentage of students receiving a federal loan.

A plot is provided between TUITIONFEE_IN and PCTFLOAN variables where the data is split between public and private universities.

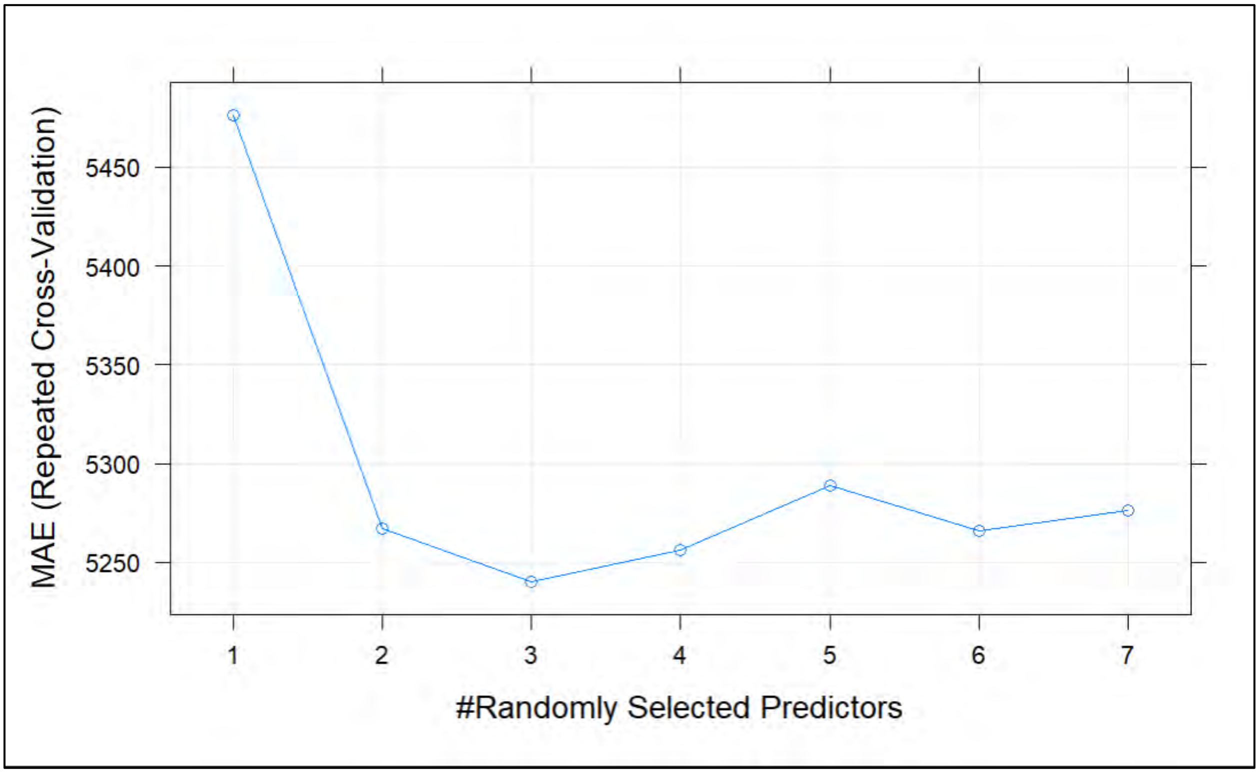


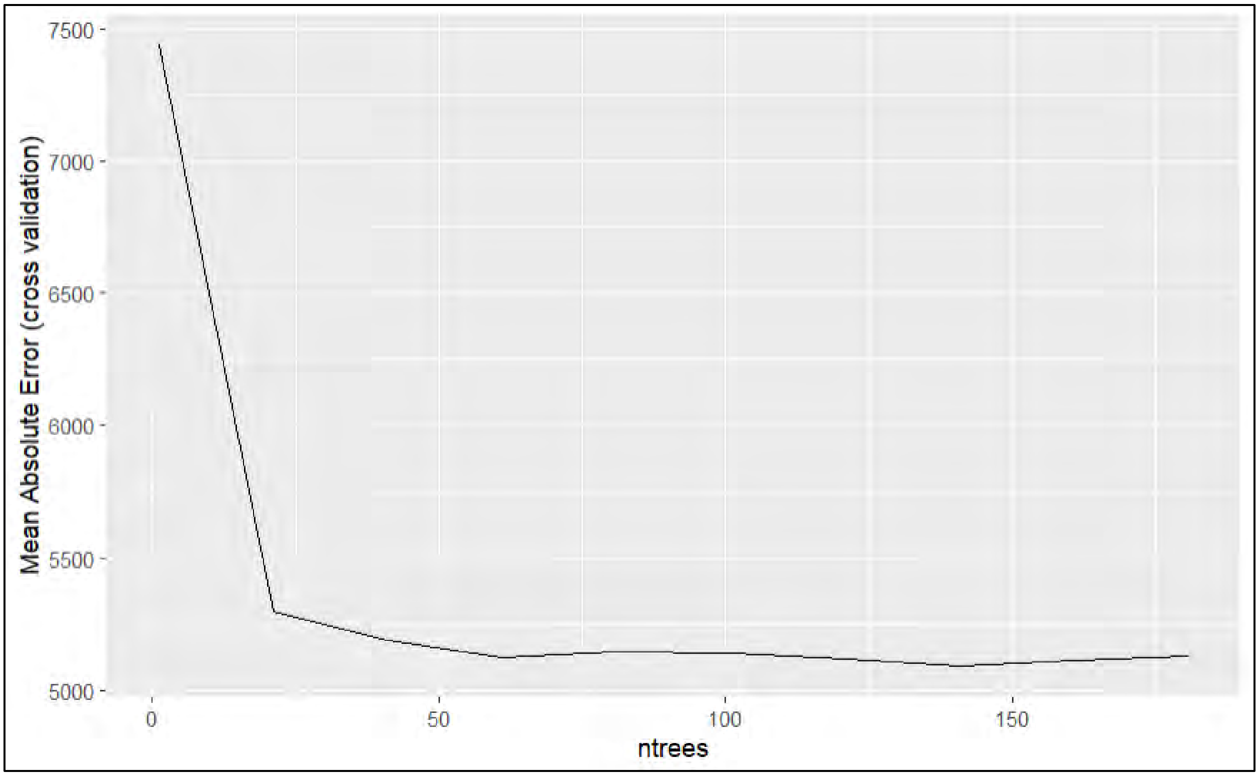
(d) *(2 points)* Interpret the plot above.

ANSWER:

Task 6 (2 *points*)

Your assistant is building a random forest model. They are asking for help with hyperparameter tuning and selection. Your assistant has provided the following output:





(a) (*2 points*) Recommend a value for each parameter in the random forest. Justify your recommendation.

ANSWER:

Parameter	**Value**
Mtry	
Ntrees	

Task 7 (8 *points*)

You ask your assistant to create a generalized linear model using variables in the dataset to predict an institution's 7-year loan repayment rate for students who completed their degree (**COMPL_RPY_7YR_RT**). Your assistant creates both an ordinary linear model and generalized linear model with a log link function and gamma distribution.

Each model has the following output:

Linear Model 1:

```
Call:
lm(formula = COMPL_RPY_7YR_RT ~ PCTPELL + CONTROL.Public + CONTROL.Private..for.profit,
    data = df.train.input1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62042 -0.04041  0.01292  0.05528  0.27690

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.008946   0.005817 173.438  < 2e-16 ***
PCTPELL                     -0.549204   0.013952 -39.365  < 2e-16 ***
CONTROL.Public              -0.034650   0.004567  -7.587 5.59e-14 ***
CONTROL.Private..for.profit -0.105873   0.010227 -10.353  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Generalized Linear Model 1:

```
Call:
glm(formula = COMPL_RPY_7YR_RT ~ PCTPELL + CONTROL.Public + CONTROL.Private..for.profit,
    family = Gamma(link = "log"), data = df.train.input1)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.09102  -0.05371   0.01451   0.06942   0.42317

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.060133   0.008155   7.373 2.68e-13 ***
PCTPELL                     -0.770396   0.019559 -39.389  < 2e-16 ***
CONTROL.Public              -0.040110   0.006403  -6.264 4.82e-10 ***
CONTROL.Private..for.profit -0.153759   0.014337 -10.725  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The **CONTROL** variable has three levels: Public, Private non-profit, and Private for-profit.

(a) (*2 points*) Describe the impact that each of the three levels of the **CONTROL** variable has on the 7-year loan repayment rate in **Linear Model 1**.

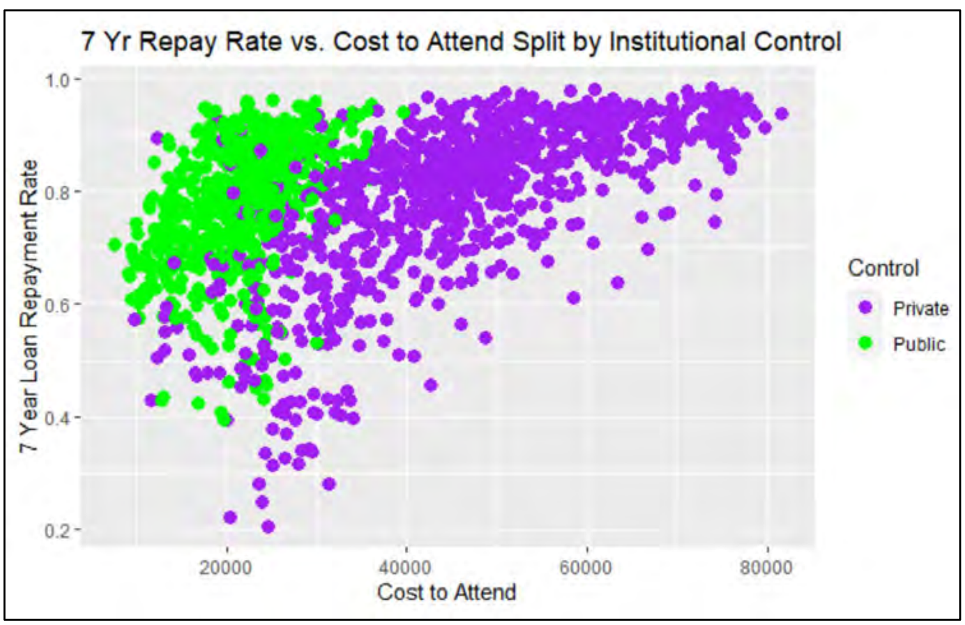
ANSWER:

(b) (*3 points*) Calculate the model's predicted 7-year loan repayment rate for each scenario below and show your work:

Answer:

Model	Scenario	Predicted 7-year loan repayment rate
Linear Model 1	Public institution with 100% of undergrads receiving Pell grants.	
Generalized Linear Model 1	Private for-profit institution with 50% of undergrads receiving Pell grants.	

Your assistant wants to add a cost of attendance variable to the model. You hypothesize that because public universities usually have a lower cost of attendance than private universities, the cost of attendance for public and private schools may have a different relationship to loan repayment rates. You create the following graph:



Based on the graph, you ask your assistant to create an interaction variable between public schools and cost of attendance. Your assistant adds the cost of attendance variable (**COSTT4_A**) and the interaction variable (**public_costt4a** = COSTT4_A * CONTROL) to the GLM and produces the following output:

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -8.779e-02  2.141e-02  -4.101 4.33e-05 ***
PCTPELL                   -6.579e-01  2.311e-02 -28.471  < 2e-16 ***
COSTT4_A                   2.264e-06  3.195e-07   7.087 2.06e-12 ***
CONTROL.Public            -1.223e-01  2.359e-02  -5.183 2.47e-07 ***
CONTROL.Private..for.profit -1.396e-01  1.406e-02  -9.929  < 2e-16 ***
public_costt4a             6.549e-06  8.781e-07   7.458 1.45e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Your assistant suggests including the new interaction variable in the model.

(c) (3 *points*) Critique your assistant's suggestion using the information above.

ANSWER:

Task 8 (8 *points*)

Your assistant performed a principal components analysis (PCA) on the three features with midpoints of SAT scores (**SATVRMID**, **SATMTMID**, **SATWRMID**). The output of the PCA is shown below.

```
[1] "SATVRMID = Midpoint of SAT scores at the institution (critical reading)"
[1] "SATMTMID = Midpoint of SAT scores at the institution (math)"
[1] "SATWRMID = Midpoint of SAT scores at the institution (writing)"
[1]
[1] "Summary of PCA on SAT scores"
Importance of components:
                          PC1     PC2     PC3
Standard deviation     1.6983 0.27724 0.19730
Proportion of Variance 0.9614 0.02562 0.01298
Cumulative Proportion  0.9614 0.98702 1.00000
[1]
[1] "Loadings of Principal Components"
               PC1        PC2        PC3
SATVRMID 0.5800841 -0.3098180  0.7533360
SATMTMID 0.5783739 -0.4945688 -0.6487568
SATWRMID 0.5735731  0.8120434 -0.1077009
```

Note that the SAT scores were standardized for the PCA analysis, and that institutions with missing SAT scores in the data were excluded.

(a) (*4 points*) Interpret standard deviation and proportion of variance in the output. Discuss the implications.

ANSWER:

(b) (*4 points*) Interpret the "Loadings of Principal Components" for PC1 and PC2 and describe the relationship among the three SAT features.

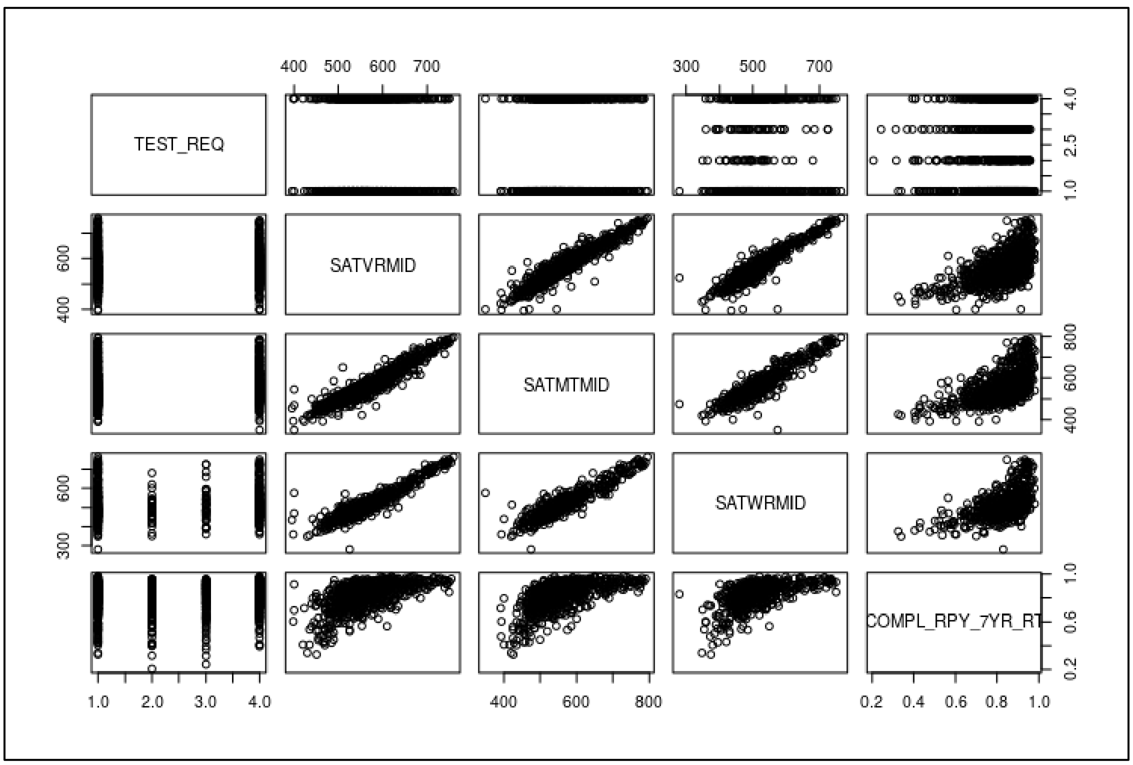
ANSWER:

Task 9 (7 *points*)

Your client is working with a national bank to provide student loans to universities and wants to have a better understanding of how certain variables impact loan repayment after graduation (**COMPL_RPY_7YR_RT**), particularly the variables **TEST_REQ**, **SATVRMID**, **SATMTMID** and **SATWRMID**.

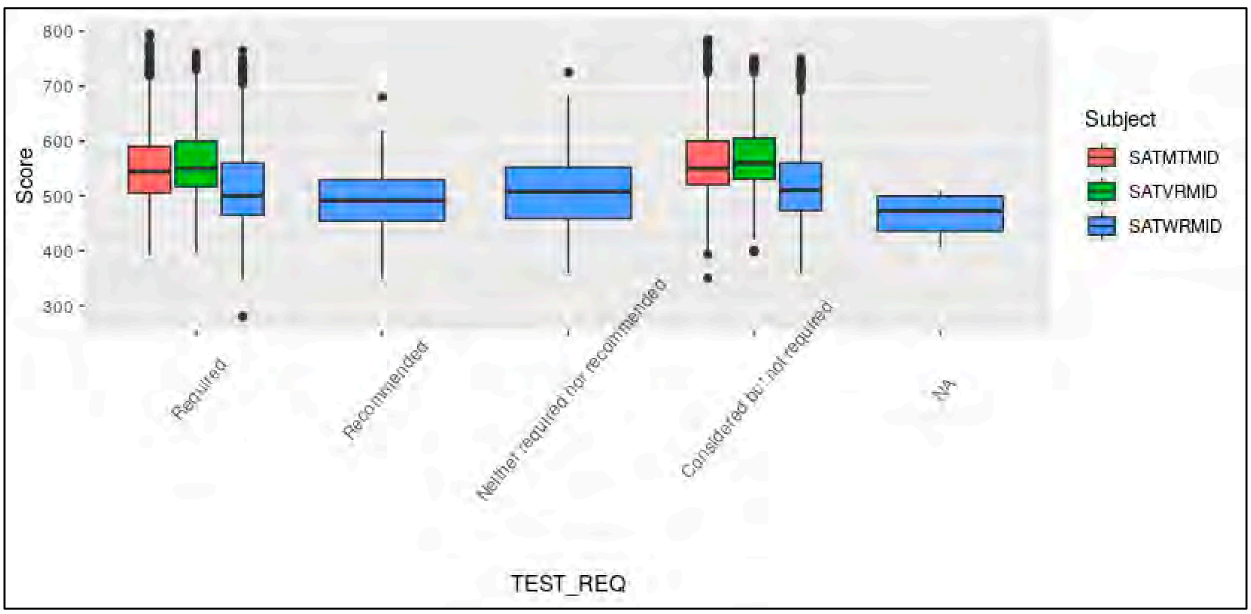
Your assistant provides you the following exploratory data analyses.

```
                                  TEST_REQ      SATVRMID        SATMTMID        SATWRMID
Required                             :601   Min.   :395.0   Min.   :350.0   Min.   :280.0
Recommended                          :209   1st Qu.:524.5   1st Qu.:512.8   1st Qu.:465.0
Neither required nor recommended:274   Median :555.0   Median :545.0   Median :500.0
Considered but not required          :575   Mean   :565.4   Mean   :560.7   Mean   :519.2
NA's                                 :521   3rd Qu.:600.0   3rd Qu.:595.0   3rd Qu.:560.0
                                            Max.   :760.0   Max.   :795.0   Max.   :765.0
                                            NA's   :1128    NA's   :1128    NA's   :1501
COMPL_RPY_7YR_RT
Min.   :0.2059
1st Qu.:0.7118
Median :0.8105
Mean   :0.7817
3rd Qu.:0.8811
Max.   :0.9814
NA's   :542
```



The table below shows each combination of the TEST_REQ with SATVRMID, SATMTMID, SATWRMID, and COMPL_RPY_7YR_RT variables. It shows N: the total number of data points with that value of TEST_REQ, n_miss: the number of times the specified variable is missing, and pct_miss: n_miss / N.

	TEST_REQ	variable	N	n_miss	pct_miss
1:	<NA>	SATVRMID	521	521	100.000000
2:	<NA>	SATMTMID	521	521	100.000000
3:	<NA>	SATWRMID	521	511	98.080614
4:	<NA>	COMPL_RPY_7YR_RT	521	221	42.418426
5:	Required	SATWRMID	601	292	48.585691
6:	Required	COMPL_RPY_7YR_RT	601	78	12.978369
7:	Required	SATVRMID	601	33	5.490849
8:	Required	SATMTMID	601	32	5.324459
9:	Recommended	SATVRMID	209	209	100.000000
10:	Recommended	SATMTMID	209	209	100.000000
11:	Recommended	SATWRMID	209	171	81.818182
12:	Recommended	COMPL_RPY_7YR_RT	209	32	15.311005
13:	Neither required nor recommended	SATVRMID	274	274	100.000000
14:	Neither required nor recommended	SATMTMID	274	274	100.000000
15:	Neither required nor recommended	SATWRMID	274	234	85.401460
16:	Neither required nor recommended	COMPL_RPY_7YR_RT	274	148	54.014599
17:	Considered but not required	SATWRMID	575	293	50.956522
18:	Considered but not required	SATMTMID	575	92	16.000000
19:	Considered but not required	SATVRMID	575	91	15.826087
20:	Considered but not required	COMPL_RPY_7YR_RT	575	63	10.956522



Your assistant used these variables to create an OLS model. You are provided with the model summary:

```
Call:
lm(formula = COMPL_RPY_7YR_RT ~ TEST_REQ + SATVRMID + SATMTMID +
    SATWRMID, data = data %>% select(sel_cols))

Residuals:
     Min       1Q   Median       3Q      Max
-0.36939 -0.04131  0.01136  0.05480  0.16010

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         3.726e-01  3.782e-02   9.852  < 2e-16 ***
TEST_REQConsidered but not required 3.525e-02  7.544e-03   4.673 3.86e-06 ***
SATVRMID                            3.274e-05  2.324e-04   0.141   0.8880
SATMTMID                            3.427e-04  1.941e-04   1.766   0.0781 .
SATWRMID                            4.317e-04  1.516e-04   2.847   0.0046 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08313 on 486 degrees of freedom
  (1689 observations deleted due to missingness)
Multiple R-squared:  0.3773,    Adjusted R-squared:  0.3722
F-statistic: 73.62 on 4 and 486 DF,  p-value: < 2.2e-16
```

(a) (*3 points*) Describe 3 weaknesses of your assistant's model with regard to model choice and data issues.

ANSWER:

(b) (*2 points*) Explain the reason that variable **TEST_REQ** has only one level shown in the model output.

ANSWER:

Your assistant replaces the missing values of the SAT scores with the mean of each score and runs two models with different interaction terms:

- Model 1 with the following three interaction terms: TEST_REQ * SATWRMID, TEST_REQ * SATVRMID, and TEST_REQ * SATMTMID
- Model 2 with one interaction term: TEST_REQ * SATWRMID

You are provided with model output:

Model 1:

```
Call:
lm(formula = COMPL_RPY_7YR_RT ~ TEST_REQ + SATVRMID + SATMTMID +
    SATWRMID + TEST_REQ * SATVRMID + TEST_REQ * SATMTMID + TEST_REQ *
    SATWRMID, data = data_fill)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35931 -0.04353  0.00711  0.05360  0.18535

Coefficients: (4 not defined because of singularities)
                                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   3.406e-01  5.002e-02   6.809  2.6e-11 ***
TEST_REQRecommended                          -2.996e-02  1.454e-01  -0.206  0.83678
TEST_REQNeither required nor recommended      2.228e-02  1.487e-01   0.150  0.88101
TEST_REQConsidered but not required           1.131e-01  7.465e-02   1.515  0.13043
SATVRMID                                      3.662e-05  3.352e-04   0.109  0.91305
SATMTMID                                      2.575e-04  2.798e-04   0.920  0.35790
SATWRMID                                      5.824e-04  1.890e-04   3.081  0.00217 **
TEST_REQRecommended:SATVRMID                         NA         NA      NA       NA
TEST_REQNeither required nor recommended:SATVRMID    NA         NA      NA       NA
TEST_REQConsidered but not required:SATVRMID  6.570e-05  4.594e-04   0.143  0.88634
TEST_REQRecommended:SATMTMID                         NA         NA      NA       NA
TEST_REQNeither required nor recommended:SATMTMID    NA         NA      NA       NA
TEST_REQConsidered but not required:SATMTMID  1.834e-04  3.834e-04   0.478  0.63260
TEST_REQRecommended:SATWRMID                  1.480e-04  2.930e-04   0.505  0.61362
TEST_REQNeither required nor recommended:SATWRMID 5.162e-05  2.943e-04   0.175  0.86084
TEST_REQConsidered but not required:SATWRMID -4.217e-04  3.055e-04  -1.380  0.16805
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08134 on 547 degrees of freedom
  (1621 observations deleted due to missingness)
Multiple R-squared:  0.3782,    Adjusted R-squared:  0.3657
F-statistic: 30.25 on 11 and 547 DF,  p-value: < 2.2e-16
```

Model 2:

```
Call:
lm(formula = COMPL_RPY_7YR_RT ~ TEST_REQ + SATWRMID + TEST_REQ *
    SATWRMID, data = data_fill)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36243 -0.04519  0.00920  0.05430  0.22260

Coefficients:
                                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   3.710e-01  3.075e-02  12.066  < 2e-16 ***
TEST_REQRecommended                           1.002e-01  1.155e-01   0.867  0.38612
TEST_REQNeither required nor recommended      1.525e-01  1.198e-01   1.273  0.20370
TEST_REQConsidered but not required           1.267e-01  4.885e-02   2.593  0.00975 **
SATWRMID                                      8.460e-04  5.862e-05  14.432  < 2e-16 ***
TEST_REQRecommended:SATWRMID                 -1.156e-04  2.330e-04  -0.496  0.62011
TEST_REQNeither required nor recommended:SATWRMID -2.120e-04  2.347e-04  -0.903  0.36688
TEST_REQConsidered but not required:SATWRMID -1.777e-04  9.346e-05  -1.902  0.05769 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08195 on 581 degrees of freedom
  (1591 observations deleted due to missingness)
Multiple R-squared:  0.3655,    Adjusted R-squared:  0.3579
F-statistic: 47.81 on 7 and 581 DF,  p-value: < 2.2e-16
```

(c) (*2 points*) Explain the reason that some interaction coefficients in Model 1 are NAs.

ANSWER:

Task 10 (7 *points*)

Looking at the target variable showing median earnings of students (**MD_EARN_WNE_P10**) your assistant notices outliers in the data and is concerned about poor model fit.

(a) (*2 points*) Explain why tree-based models are resilient to outliers in predictor variables.

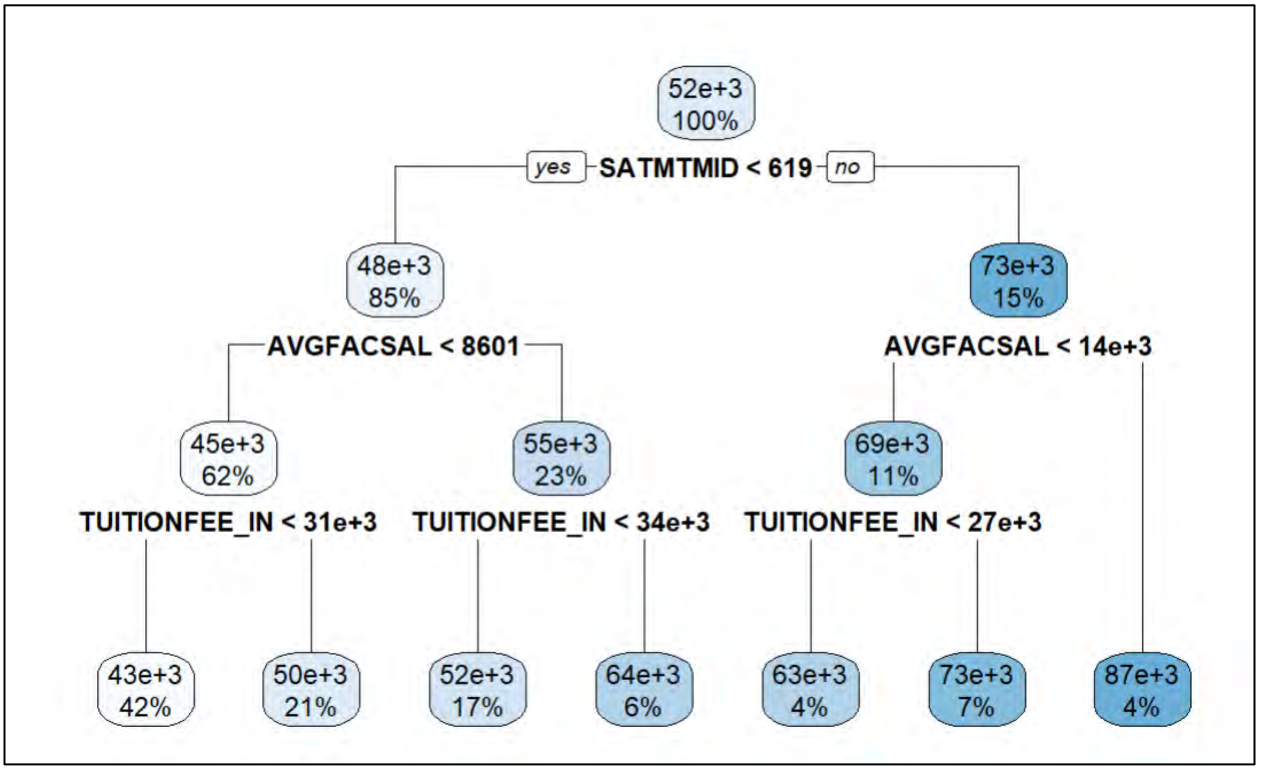
ANSWER:

Your assistant asks which of the two metrics, **Root Mean Square Error** (RMSE) or **Mean Absolute Error** (MAE), to use in evaluating model performance. Your assistant wants to use a metric that is more robust to outliers in the target variable.

(b) *(2 points)* Recommend which metric to use and justify your recommendation.

ANSWER:

Your assistant built the tree below with median earnings (**MD_EARN_WNE_P10**) as the target variable and provided two data points pulled from the test data.



Note: When reading the decision tree assume all node values are rounded to the nearest thousand. For example, "63e+3" should be interpreted as 63,000.

	SATMTMID	AVGFACSAL	TUITIONFEE_IN	MD_EARN_WNE_P10
Data Point 1	630	9,000	35,000	120,000
Data Point 2	630	9,000	35,000	200,000

(c) *(3 points)* Calculate the RMSE and MAE for the test data above using the tree model. Show your work.

ANSWER: