15.071: The Analytics Edge **Spring 2021**
Homework 1: Linear Regression and CART

Out: February 19; Due: March 3, 11:59pm.
Please post the assignment in pdf format with file name "Lastname_15071_HW1.pdf".
For each question, please include the main `R` *commands that you used in your submission.*

All of the `R` *code in this document is also provided in the file "HW1_UsefulCode.R". Please use the .R file if you have difficulty copying code from the pdf document.*

Problem 1: Predicting Housing Prices in Ames, Iowa

In this problem, we consider the Ames, Iowa Housing Prices dataset, which describes sales of 2,838 properties in the town of Ames, Iowa from 2006 to 2010[1].

You will work with the dataset provided in the **`AmesSales.csv`** file. This file has been pre-processed to simplify the analysis. We started with a partially processed set available on Github[2] and selected a few of the most relevant variables to include in our analysis.

The file contains 12 variables, described below. The first variable is the property's sale price—which we aim to predict. The other variables describe the property details in quantitative terms (square footage, number of rooms, date of construction). There is one categorical variable, **`BldgType`**, which describes different types of homes (e.g. townhouse, duplex, etc.).

- SalePrice - the property's sale price (dollars)
- TotalRooms: Total number of rooms
- Bedrooms: # bedrooms
- FullBath: Full bathrooms
- HalfBath: Half baths
- LivArea: Ground living area (sq. feet)
- Fireplaces: Number of fireplaces
- GarageArea: Size of garage (sq. feet)
- PoolArea: Size of pool (sq. feet)
- YearBuilt: Original construction date
- YearSold: Year Sold
- BldgType: Type of dwelling

Run the following commands to read in the data and make sure **`SalePrice`** is encoded as a numeric variable and not a factor variable.

```
ames = read.csv("AmesSales.csv")
ames$SalePrice = as.numeric(ames$SalePrice)
```

1a Use the summary() and hist() function on SalePrice to view the distribution of the dependent variable. What do you notice about its distribution? Does this agree with your intuition? [6 points]

1b We will now develop regression models to predict housing prices in Ames, Iowa. First, split the dataset into a training set (70%) and a test set (30%). Please use the code below to ensure that you get the same split as the TA's solution.

```
library(caret)
RNGkind(sample.kind = "Rounding")
set.seed(310)
idx = createDataPartition(ames$SalePrice, p = 0.70, list = FALSE)
train = ames[idx,]
test = ames[-idx,]
```

[1] De Cock, Dean. "Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project." *Journal of Statistics Education* 19.3 (2011).
[2] `https://github.com/mikearango/DATS_Final`

[You can check whether your split is correct by evaluating `mean(train$SalePrice)`. The answer should be 178635.7. If you get a different answer, please see the note at the end of this homework.]

Train a linear regression model to predict SalePrice using all the other variables. What is the R^2 for this model? [6 points]

1c `R` offers a number of useful statistics and plots to help assess the quality of regression models and identify potential problems with the model. For linear regression models, you can access a significant amount of information by plotting your regression model:

```
par(mfrow=c(2,2)) #create two rows and columns in the plot window
plot(model)
par(mfrow=c(1,1)) #reset the plot window to default settings
```

The output from this code is shown in Figure 1.

In this exercise, we will focus on identifying outliers. Outliers can sometimes have a disproportionate effect on the regression model and lead to poor estimates of the model coefficients. In the regression diagnostic plots for this model, three extreme outliers are highlighted: observations "1451", "2114", and "2115"[3]. These points have very large negative residuals (plotted on the y-axis), which means that the regression model does not fit these rows well (the predicted prices are significantly higher than the actual prices). In the last plot we see that these observations also have higher-than-average *leverage*, which is a measure of how much these observations influence the regression equations[4]. This combination of factors means that we should definitely investigate these observations further!

First, let's print out these three observations and compare them to the rest of the training data.

```
# Outliers
outliers=c("1451", "2114", "2115")
print(train[outliers,])
# Data summary
summary(train)
```

Which column in the dataset has a significant discrepancy between the outliers and the other observations? Based on the discrepancies, do you think the outlier values are realistic? What are some potential explanations for why these rows are so different? [8 points]

1d When we find outliers that are also very influential in the model, we can consider the following options:

- Remove the observations from the model. This is appropriate if we believe the outliers are due to data errors, or if the outliers are substantially different from the target population that we are trying to model (for example, we could probably justify excluding a professional athlete from a model that uses data on physical exercise habits).
- Treat the outlier values as missing data. This should only be done if we believe the outlier values are due to data errors. Depending on the type of application, we may be able to use imputation or other methods to avoid discarding the entire observation from our dataset.
- Keep the outliers in our model. If we don't believe that there is anything inherently wrong with the outlier, then it may reflect something important about the system we are trying to model.

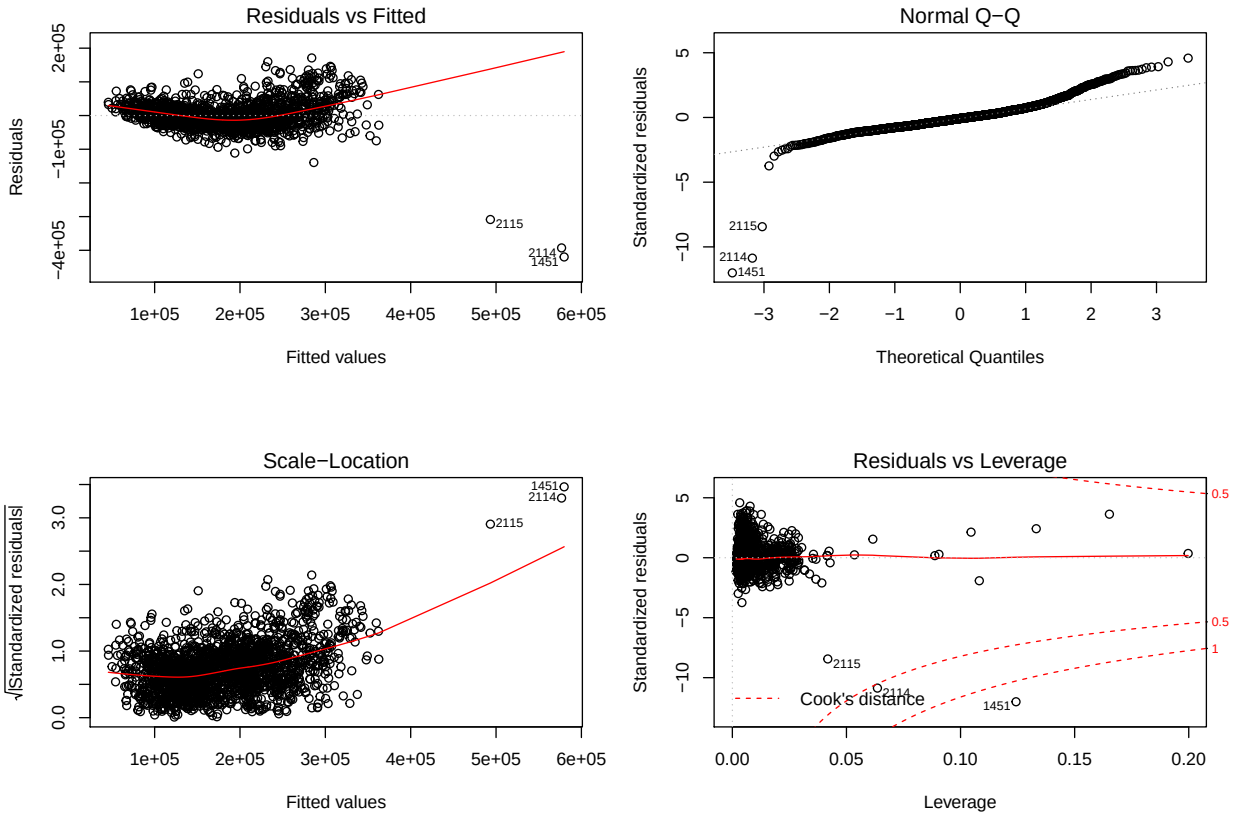
In this case, we will remove the outliers from our dataset. You can use this code to create a new training set without the outliers:

```
train2 = ames[setdiff(idx,outliers), ]
```

[3]NB: The outlier labels refer to row names, not row numbers. The row names match the position of the outlier observations in the initial dataset, but not in the training data (because the training data is missing all the rows in the testing data).

[4]If you'd like to read more about leverage statistics, this source may be helpful: `http://www.mit.edu/~6.s085/notes/lecture4.pdf`

Figure 1: Regression diagnostic plots: each circle on these plots represents an observation (row) in the training data, and the y-axis shows different versions of the residuals (the difference between the observed SalePrice and the model's prediction). `R` has automatically labeled some of the outlier observations in each plot—this is addressed in question 1c.
Advanced topics: If you are familiar with regression diagnostics, you might notice other potential problems in these graphs: (1) there is a non-linear trend in the residuals (the red trendline in plot 1 indicates that larger fitted values tend to have more positive residuals), and (2) the variance or 'spread' of the residuals is not even (the residuals are very close together on the left of plot 1, and more spread out on the right). These trends can indicate that some of our model assumptions are violated, but we will not deal with these advanced topics in this assignment.



Train a new regression model without the outlier observations. What is the new R^2? How does the fit of this model compare to the previous one? [6 points]

1e Do the regression coefficients in this model make sense? Are there any model coefficients that have the wrong sign? Is there anything surprising about the statistical significance (or lack thereof) of certain coefficients? Check for correlations among the numerical variables in the training data and comment on the consequences of these correlations in the model. [10 points]

1f Suppose that a fireplace installation company in Iowa looks at your model and decides to run an advertising campaign with the following claim: *"a model developed at MIT shows that installing a fireplace will raise the value of your house by more than $10,000!"* Explain why regression models cannot be used to support this type of claim. Can you suggest an alternative way of phrasing this claim so that it is supported by the data? [6 points]

1g Train one final regression model using the following variables:

```
SalePrice ~ BldgType+YearBuilt+Fireplaces+GarageArea+PoolArea+LivArea
```

Remember to remove the outliers! (Use **train2**.)

Evaluate the R^2 and compare the fit of this model to the previous model. Calculate and report the OSR^2 (on the test set) for both models.

Which model would you use to predict house prices? Which model would you use to analyze the relationship between different features and SalePrice? [10 points]

Problem 2: Regression trees

2a Train a CART model on the training set without outliers (**train2**). Plot an image of your tree using the **prp()** function and include it in your solutions. Which variables seem important? Does the tree seem sensible? [8 points]

2b Make predictions and calculate the R^2 for your tree model on the training set. How does it compare to the last regression model? [6 points]

2c Let's fit a bigger tree to see if we can obtain more detailed predictions. Train a new model with the following settings:

```
tree.model2 = rpart(SalePrice ~., data = train2, control = rpart.control(cp=0.0001))
```

This tree will be large and may take a long time to plot! If you cannot see the tree structure properly in R Studio, try exporting the plot to a pdf file and zooming in.

Since this tree is so large, we'll use a shortcut to analyze the different variables in the tree. Run the following code to extract a variable importance plot:

```
barplot(tree.model2$variable.importance, cex.names=.7)
```

This plot shows the relative contributions of each variable to the final tree model (you might need to click the "Zoom" button to see all of the variable names). Based on this plot, what are the two most important features in predicting SalePrice? Is this variable importance plot similar to the variables identified as significant in the linear regression model? (You don't need to discuss every variable, just highlight any differences that you think are important.) [8 points]

2d Make predictions on the test set, using the tree obtained in question 2a (with the default value of **cp=0.01**) and your second model obtained in question 2c. Report their OSR^2 for the test data. What are some of the pros and cons of these two tree models? [8 points]

2e How does the best tree model compare with the linear regression models? Which model would you be more likely to use if you were buying or selling a property in Ames, Iowa? [8 points]

2f We have seen that the complexity parameter (cp) can have a very significant effect on the output of tree models. Can you suggest a data-driven way to choose the best value for this parameter in order to improve the fit of the model? (You don't need to implement this, just describe it in words.) [6 points]

2g Explain why it is important to split our data into training and testing sets, and how we use each dataset. Based on your answer, do you think there are any problems with

- the way we have chosen the "best tree" in 2d?
- comparing the OSR^2 of our "best tree" to the OSR^2 of the linear regression model we constructed in 1g?

[4 points]

Notes on question 1b

In all our homework assignments we will try to ensure that every student uses the same train/test split for building and evaluating models. Different splits can lead to different model output, which may make it more difficult for you to answer subsequent questions. It is also much easier to get help with coding problems if you are using the same data as the TAs.

Most students should get the same train/test split by running the code provided in question 1b. However, certain students may get different results due to differences in their operating system or package versions. If this happens, we ask that you use the following alternative option to create your train/test datasets:

```
amesSplit = readRDS("amesSplitBackup.RDS")
idx = which(amesSplit$partition=="train",)
train = ames[idx,]
test = ames[-idx,]
```