Homework 3

Instructions

Due: 5:00pm on Wednesday, September 29th

1. Add your name between the quotation marks on the author line in the YAML above.
2. Compose your answer to each problem between the bars of red stars.
3. Commit your changes frequently.
4. Be sure to knit your .Rmd to a .pdf file.
5. Push both the .pdf and the .Rmd file to your repo on the class organization before the deadline.

Regression Competition

Your objective is to build a multiple regression model **using the tools we have discussed in class** that predicts the price of a house in Ames, Iowa as a function of other characteristics of the house. You will construct your model using a training data set with information on 66 variables recorded for 1808 houses. I've held back the data on 600 other houses which will serve as the test data set for assessing the predictive accuracy of your model.

You should record your answers in this .Rmd file. However, you are encouraged to use a separate .Rmd file for scratchwork. The assignment is divided into several **Components** to help organize your work. Put all work you want graded between the bars of red stars in the corresponding section.

Grading

Your final score on this assignment will be based 40% on the accuracy of your model predictions on the test data (as measured by root MSE in the original response variable) and 60% on the quality and depth of your explanations and analysis.

I *do not* expect you to find the absolute best model for this data set. In fact, it is entirely possible to earn top marks on this assignment with a model of mediocre accuracy, provided you submit insightful analysis based on topics we have investigated in our course.

You will be graded as much on your discussion of what *didn't work* and why, as what did.

The Data

The data set `house_train` can be found in the hw_3 repo and can be loaded by running the following code.

```
house<-read_csv("house_train.csv")
```

Additionally, the `data_description.txt` file in the same repo gives a full description of the variables appearing in the data set.

There is one special column of note:

- `Sale_Price` is your response variable and should not be included as a predictor.

Components

Data Exploration

In this section, you should perform preliminary data exploration and analysis. This data set is a bit too large to do a full investigation of each variable, so select several quantitative and categorical variables you think may be useful (at least 3 of each). Look at each variable's individual distribution, along with the joint distribution of this variable and the response (i.e. using a scatterplot for quantitative variables and side-by-side boxplots / histograms for categorical variables.)

Do these relationships look strong/weak? Linear/non-linear? Does it seem like a transformation would be useful? Comment on any other relationships you find in the data.

Model Building

In this section, you should build a series of MLR models (at least 3) of varying complexity and that use a variety of the tools we have studied thus far. Explain why you choose to implement various features in each model. Each model **must** have at least 6 variables and can include **at most** one interaction term.

Model Selection

In this section, compare the results of your models. Which models seemed to perform better or worse? Why? Be sure to consider the effect your choices may have on MSE based on the Bias-Variance trade-off.

Your model

Identify the model you feel will be most accurate in predicting `Sale_Price` on the test set. Your model **must** have at least 5 variables and can include **at most** one interaction term.

The following three functions will help me assess your model accuracy. Copy the following templates and modify to create R functions for your model. Be sure to change the name of the functions to your own first and last names.

These functions should be self-contained, so include any packages you need or data processing you use. I will input the training data and run in a separate .Rmd, so it is important it can stand alone.

```
#This function performs data processing. Anything you do to the training set must be repeated on the te
# I will apply this function to the test and training data

FirstName_LastName_processing <- function(my_data){
  library(tidyverse) ## Load whatever packages you need
  processed_data <- my_data %>% mutate(Sale_Price = Sale_Price) ## Include all relevant processing step
  processed_data ## returns the processed data as output
}
```

```
#This function creates your linear model. I will apply it to the results of FirstName_LastName_processi

FirstName_LastName_model <- function(training_data){
  library(tidyverse)     ## Load whatever packages you need
```

```
  my_mod <- lm(Sale_Price ~ 1, data = training_data)    ## Create your model. Replace 1 with your actua
  my_mod      ##return your model as output
}
```

```
# This function makes predictions for the Sale_Price of houses.
# I will apply it to the results of FirstName_LastName_processing(house_test) and FirstName_LastName_mo
# If you performed any transformations on the response variable, you must transform the predicted value

FirstName_LastName_predictions <- function(model, test_data){
  library(tidyverse)     ## Load whatever packages you need
  my_preds <- predict(model, test_data)   ## Make predictions based on your model. Don't change this li
  my_preds<- my_predictions*1 ## Transform your model predictions back to the original units for Sale_P
  my_preds      ##return your predictions as output
}
```

To verify that your functions are working as desired, open a new .Rmd file, load the `house_train` data, copy the code for your 3 functions over to the new .Rmd, and then run the following code:

```
A <- FirstName_LastName_processing(house_train) # Change to your First and Last Name
library(dplyr)
B <- sample_n(house_train, size = 10) # This creates a test set of 10 observations
mod <- FirstName_LastName_model(A) # Change to your First and Last Name
FirstName_LastName_predictions(mod,B) # Change to your First and Last Name
```

If everything is working correctly, the result of the code should be 10 predicted sale prices.

Model Diagnostics

In this section, perform diagnostics on the model you selected. Does it appear your model satisfies the MLR modeling assumptions? If not, comment on the effect this might have on the accuracy of your model.

Model Interpretation

Interpret some of the coefficients in your model. Are there any coefficients or predictors that are surprising?

Conclusions

Discuss some limitations of your methods and your model. What are some ways you could improve your model if you had more **time**? Identify one variable **not** in the data set you feel could be an important predictor of `SalePrice`. How confident are you in the accuracy of your model?