5303 Project 4

Due by Friday Nov. 1th, 2024

Note: You need to submit two files for this assignment, a pdf report, and the original code file in R. Missing the original code file will have 50% reduction of your score. If you submit it one day late, you will have 50% reduction of your score.

Useful Libraries reference:
MASS, Ameshousing, tidyverse, caret, glmnet

Download and use the Diamond.csv file for this project

1. Please view the dataset, extract the target variable ('price'), the features dataset.
2. Apply some necessary preprocessing to the dataset. List all the steps you do and explain it.
3. Split the data into training and testing datasets.
4. Build a linear regression model. Interpret the results you get. Calculate the MSE for testing dataset. Name it by MSE_lm and save it.
5. Build a Lasso regression model. Interpret the results you get. Calculate the MSE for testing dataset. Name it by MSE_lasso and save it. (Hint: use cv.glmnet to get the best lambda)
6. Build a Ridge regression model. Interpret the results you get. Calculate the MSE for testing dataset. Name it by MSE_ridge and save it. (Hint: use cv.glmnet to get the best lambda)
7. Compare MSE_lm, MSE_lasso and MSE_ridge. Make your conclusion.
8. Plot three pictures for the predictions of the testing dataset for all models you obtained in 3 and 4. Show some observations your get.

Bonus: Use the dataset AmesHousing.
Follow the steps below then you can download the data.

```
install.packages("AmesHousing")
library(AmesHousing)
data <- make_ames()
```

Go through the question 1-7 for this new dataset AmesHousing. Check the MSE of testing dataset for all models. If MSEs are very big, provide some methods to deal with it. If your MSEs are not big, list all the variables you have in each model. The label for this dataset is ‘Sale_Price’.