**Assignment 2**

After taking the first glance of the data structure, we split the dataset into train set (80%) and test set (20%). Then we calculated correlation between each variable and plotted the histogram for each. These are our findings.

* Age (age): Most of the properties are old and were built before 1940.
* Proximity to River Charles (chas): There very few number properties which are near the River Charles, however the correlation on the train data shows a low correlation with the median value of homes. It will be interesting to look at the correlation of the median value of homes between near the river to observe any change in pricing for the scenic view or could be negative due to fear of floods in the river?
* Crime Rate (crim): There is very less crime rate in Boston based on the data points available, but from correlation it seems that in areas where there is crime, the property median prices fall with increase in crime rate, which is expected.
* Distance from Employment Centers (dis): Most of the housing is in proximity to the employment centers, however the correlation is only 25% therefore there could be other parameters which are also impacting the pricing.
* Air Pollution (nox): Pollution is on the higher side across most properties but still has a high correlation, which might be interesting to look at in further details.
* Accessibility to radial highway (rad): Either the houses are close to the highway or are relatively far away. There is a wide range where there are no houses.
* Average rooms (rooms): The histogram is normally distributed with most properties having an average number of 6 rooms per house and has the highest correlation with the median value of houses.
* Tax (tax): The properties can be split into two large groups and are either in the range of 200 to 450 or at 700.
* Percentage of land zoned for lots (zn): Sizeable number of properties are outside the land zoned for lots.

In addition, we tried more variables, like log of age, Log of lstat, log of dis, squared age, squared lstat, and squared dis.

With the data cleaned and prepared to train the model, we tried two methods of linear regression and Ridge. First, we chose four independent variables by ranking the correlation with response variable, so we picked logDis, rooms, logAge, and lstat.

RMSE = 4.41

Cross-validation RMSE = 4.460

Using Ridge, we got a similar regression result:

RMSE = 4.18

Cross-validation RMSE = 4.459

Although the ridge regression slightly outperformed our linear regression with smaller RMSE, our observations on average are off by 21%. We would not recommend to management deploying our ridge regression model.

We made another try to take logMV as response variable, and picked rooms, logLstat, tax, indus and crim as independent variables.

Linear regression result:

RMSE = 0.0874

Cross-validation RMSE = 0.09

Ridge gave the following result:

RMSE = 0.0874

Cross-validation RMSE = 0.09

Interesting that both models gave same RMSE, so we do not have strong evidence to pick one is better than the other. But we are surprised by seeing both linear regression and ridge provided RMSE is 7% of mean value of response variable, instead of 21% in previous. Finally, we could recommend the management that both simple linear regression and more complex ridge are good to use, but we should choose log of median value of homes as the response variable, instead of median value directly. Based on the variables we selected, more rooms, fewer low-income people, less tax burden, more business, lower crime rate are all predicting a higher home value.