The Goal of the project is to Predict the sale price with given features like bedrooms, bathrooms, lot size etc. and to minimize the difference between the actual sale price and predicted sale price.

**(a) What kind of data processing did you do and why?**

First imported the data. The Dataset (including training and testing) Contains 5031 observations with 109 features.

Out of 109 features, I dropped the columns which are not important like phone number, agent name etc. I took the subset of data set with 33 features which are important for the prediction.

Most of the columns are Categorical very few are numeric variables. First checked for the null values. Most of the them have null values. Deleted some observation which have very less percentage of null values. So, we can impute Null values with median or mean values for numeric variables or we can do regression also by sub setting the dataset. I chose to impute it with median values.

For the categorical columns imputed null values with “missing data”. Separated the text data and found the unique values in every column and created separate columns. Imputed values “1” or “0“by searching the keyword in every row of the derived column. for some categorical variable used dummy encoding and appended it column wise. After the data is clean deleted the rows with duplicates. Age having 3 outliers, so capped them to maximum (2017) of that column. Did log transformation to the price columns used. Deleted the columns used for creating distinct feature columns.

**b) What is the metric (e.g. MSE, accuracy, precision, R-square etc.) that you optimized for? Why did you choose this metric?**

The metric I chose is MAPE (mean absolute percentage error) as the goal of the project is to reduce the difference between actual and predicted values. So, with this we can check how much percentage actuals are different from predicted. As the MSE gives the total error not the error. so, my metric is MAPE.

**c) Summarize your findings?**

In Conclusion, the dataset with many features it is very important to do the Feature Engineering. If all the features are important, it is better to keep all. The model will learn more from the training then it predicts correctly. After splitting the data into training and testing, applied linear regression and Random forest regression for predicting the sales price. Both models gave similar results. Did the Delog by taking exponential of the test and predicted data. The Optimization Metric MAPE is 4.5% for both the models. It means the actual value is 4.5% different from predicted value.