Homework 2

1. **First, I imported the two datasets namely**- house_train.csv and house_test.csv in R. Here, house_train.csv consists of the whole data along with the price2013 column and house_test.csv consists of less no. of records without price2013 column where I'll be predicting the house prices for year 2013 based on my regression model.

2. **Performing Simple Linear Regression**
   To predict the home prices for year 2013 based on the state information only, I have run a simple linear regression model with dummy state variable i.e. categorical variable on the house_train.csv dataset. On doing this, I got the following results-
   a. **The intercept is 281730.0 and it is significant as it predicts the average price of the house which exists in state "AK".** All the other regression coefficients are related to this as to calculate the average price of any other state, we need to add this intercept value to the specific state's coefficient.
   b. We get this information from our regression model as the intercept is the value of the dependent variable in the regression equation when the value of all independent variables is zero. Here, we have all states as variable apart from "AK" and hence the intercept denotes the price of the houses from state "AK".
   c. **Based on the regression coefficients I found out that the state "DC" has the most expensive average homes and the state "WV" has the least expensive average homes.**
   d. To get this information, I looked at the state which had the highest regression coefficient and the state which had the lowest regression coefficient.
   e. **The average price of homes in state "WV" is 98423.1 while the average price of homes in state "DC" is 514288.9**
   f. To get the average price of the home, I considered each state in the regression equation one by one and calculated the sum of the intercept and the regression coefficient of the state to get the average price of the house in that state.

3. **Performing Multiple Linear Regression**
   To predict the 2013 home prices based on state and county information, I have run a multiple linear regression model with both the categorical variables i.e. state and county. On doing this we get the following results-
   a. The county with the highest regression coefficient is "Pitkin" and the one with the lowest is "Calaveras".

4. **Building a regressor that best predicts average home values in this dataset.**
   a. To build the best predictor, I ran multiple linear regression model with different variable combinations. I tried to include all those variables that have a significant impact on the home prices. I kept trying until I reached to a point when I got a good value of adjusted R-squared which was very near to 1. Also, the dataset had a large no. of records and therefore a good regression model must have an F-statistic greater than 1. I kept checking my score on kaggle by uploading the prediction file there. The house_test.csv file had 6 extra counties which were not present in house_training.csv file and

therefore I included these 6 counties in the training file with zero values in the related columns. By doing this change, I ran the regression model and got the best possible predictor.

b. **My best kaggle score is 44316.2**
c. **My kaggle user name is Saurabh Thakrani.**

5. **P(White ball) = (P(white ball/Bag1)*P(Bag1))+(P(white ball/Bag2)*P(Bag2))**

$$= \frac{2}{3} * \frac{1}{2} + \frac{3}{4} * \frac{1}{2}$$

$$= \frac{17}{24}$$

6. **OSG = Opposing team scores first goal**
   SSG = Self team scores first goal

   P(Win) = P(Win/SSG)*P(SSG) + P(Win/OSG)*P(OSG)
   = 0.6*0.3 + 0.1*0.7
   = 0.25

   Therefore, there are 25% of the games won, if the team scores the first goal 30% times.