AI 534 Implementation Assignment 0

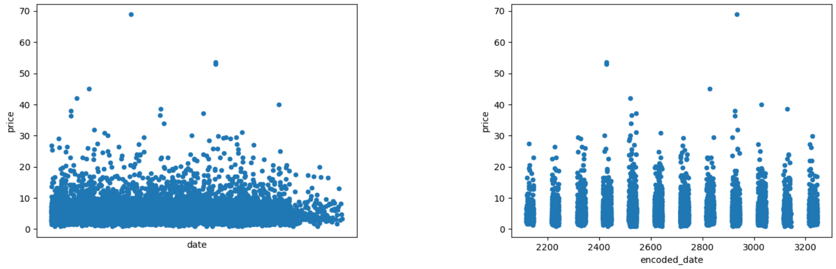
Rishab Balasubramanian

Part 2. a. I do not think that using the ID feature is a good idea as there should be no relation between the ID of the house and it's selling price. We can interchange the ID's of two houses, and still the price would not change. Hence it can be removed

b. The date of the house sale must be relevant to the selling price of the house. For example with inflation and changing housing market scenarios, the same house sold on a different date would go for a different price. Another way to use the date information is to encode it into another variable, say '$date_{encoded}$' as:

$$date_{encoded} = year + 100 * month + day \tag{1}$$

This equation ensures that for every $year > 1200$, the value of $date_{encoded}$ is unique. To test this method conider the following two figures.



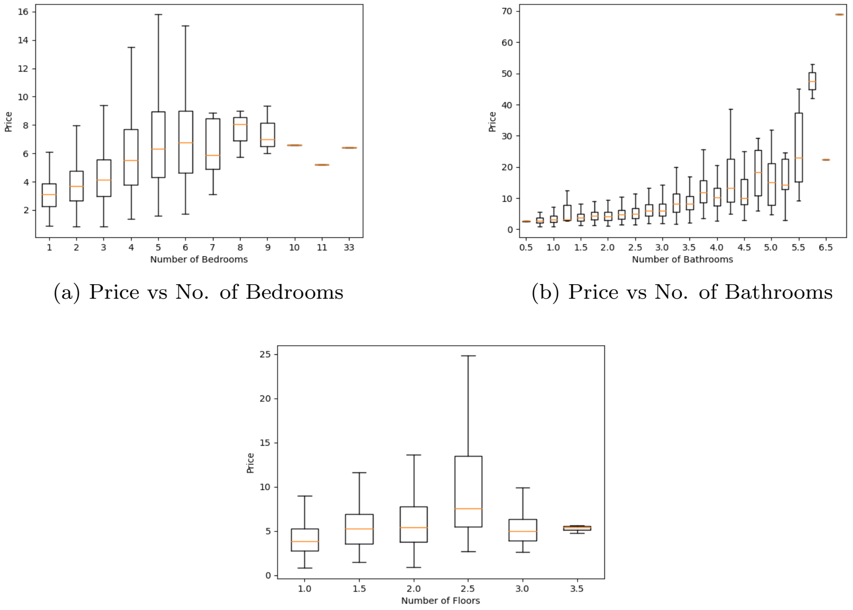
(a) Plot of the housing price vs date

(b) Plot of the housing price vs $date_{encoded}$

Figure 1: Impact of encoding date on housing price prediction

Fig. 1a shows a scatter plot of the housing price vs the date. We see that there seems to be almost no correlation between the housing price and the date. However, using the encoded date, Fig. 1b shows a much more cleaner plot, with distinct changes in the encoded date values. So by knowing the value of the encoded date, we can predict the mean and variance of housing prices.

c. Figures 2a, 2b, 2c show the variations in price with respect to each entitiy



(a) Price vs No. of Bedrooms

(b) Price vs No. of Bathrooms

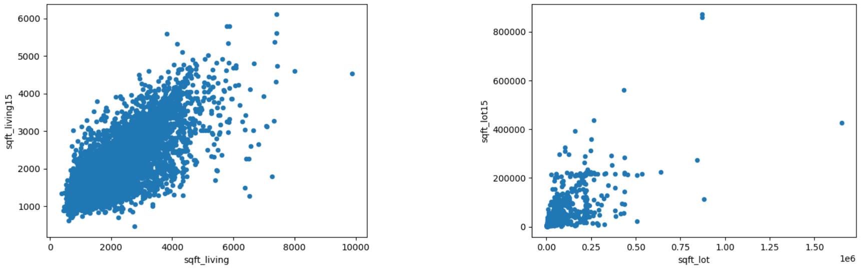
(c) Price vs No. of Floors

Figure 2: Variation of Price with: a. Bedrooms b. Bathrooms c. Floors

d. The following two tables show the output Covariance (Table. 1) and Correlation (Table. 2) matrices. Fig. 3a, 3b show the scatter plots for sqft_living15 vs sqft_living and sqft_lot15 vs sqft_lot. Although we can conclude from the correlation matrix and the output plots that there is a lot of correlation between (sqft_living15 and sqft_living), and (sqft_lot15 and sqft_lot), they are not completely the same i.e the correlation is not ±1. Therefore we can say that these features are not redundant.

Covariance Matrix				
	sqft-living	sqft-lot	sqft-living15	sqft-lot15
sqft-living	8.305303e+05	6.473942e+06	4.839029e+05	4.836731e+06
sqft-lot	6.473942e+06	1.697761e+09	4.160910e+06	8.924357e+08
sqft-living15	4.839029e+05	4.160910e+06	4.787260e+05	3.568584e+06
sqft-lot15	4.836731e+06	8.924357e+08	3.568584e+06	7.975678e+08

Correlation Matrix				
	sqft-living	sqft-lot	sqft-living15	sqft-lot15
sqft-living	1.000000	0.172406	0.767427	0.187928
sqft-lot	0.172406	1.000000	0.145951	0.766928
sqft-living15	0.767427	0.145951	1.000000	0.182629
sqft-lot15	0.187928	0.766928	0.182629	1.000000



(a) Plot of sqft_living15 vs sqft_living

(b) Plot of sqft_lot15 vs sqft_lot

Figure 3: Scatter plots