

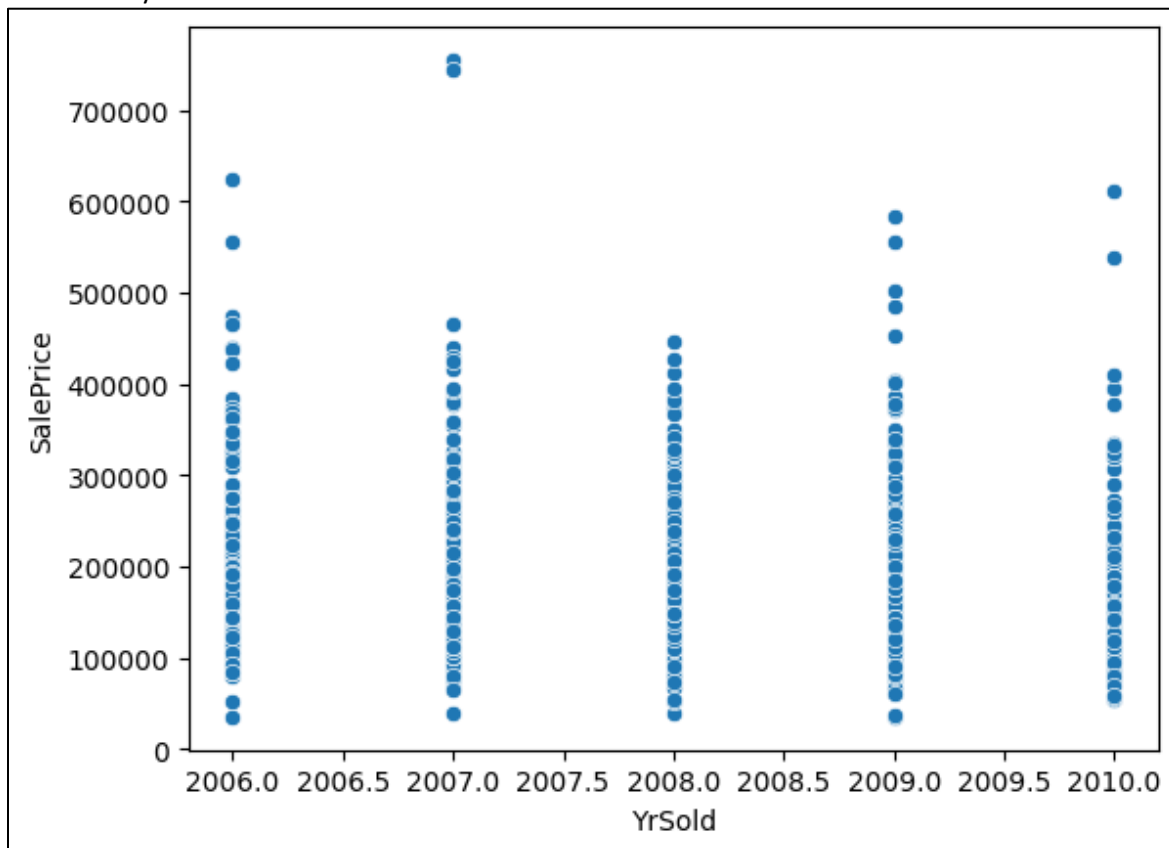
Assignment 3 Part 3

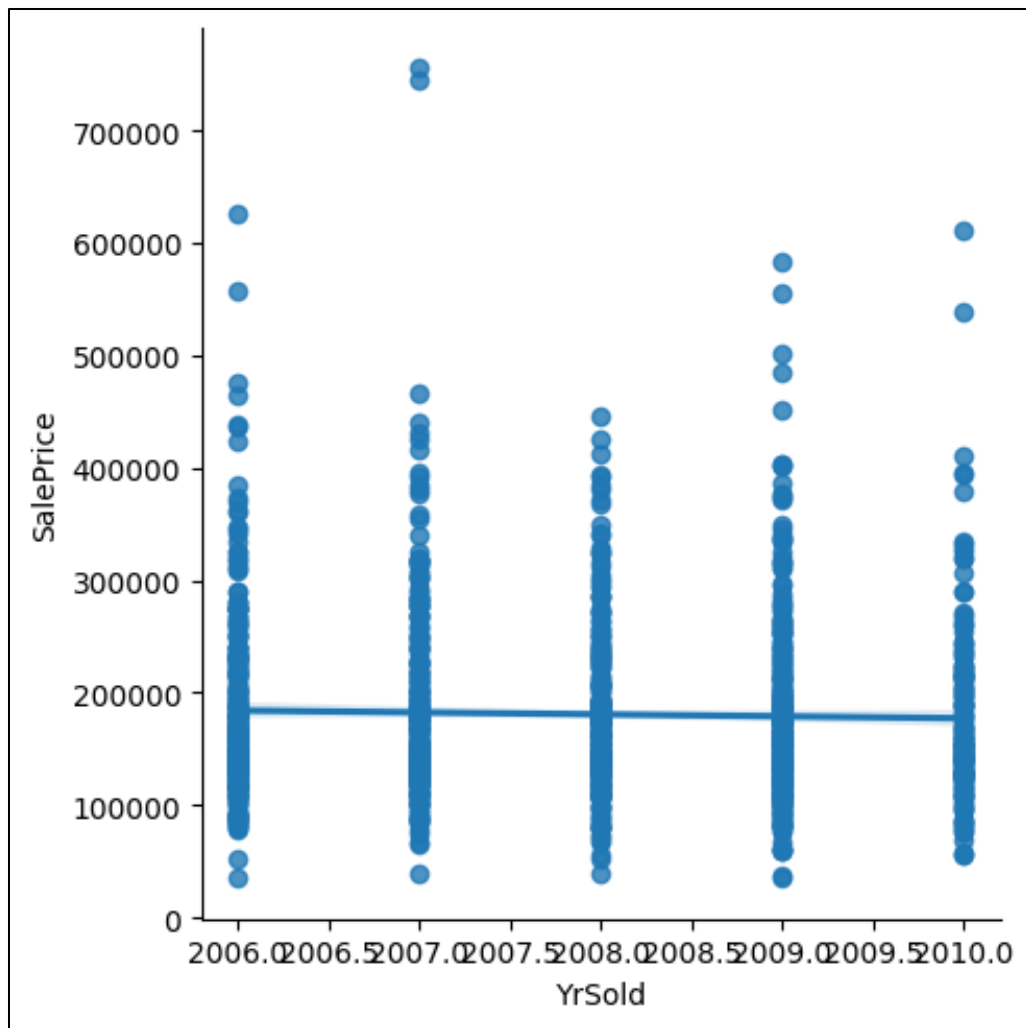
Write a summary of your findings in one page (e.g., summary statistics, plots) and submit the pdf file.

---The dataset train.csv has 81 rows and 1460 columns or data.

- (a) Among the 1460 data in the dataset, the mean value for Sale Price Variable is 180 921, while for Year Sold is 2007.82, which means to say that most garage rate is in the range of 180 000 – 200 000. Also, it is during the year 2007-2008 that garages are sold rapidly.

Here, let us check the correlation between 2 variables in the dataset, Sale Price and Year Sold. We get the coefficient result of -0.0289, which means that there is a weak relationship between these two variables. Thus, they are somehow not related to each other, any increase or decrease in price is not associated with the year sold. It is also visible in the scatterplot below. Also, the horizontal regression line does significantly mean that there is a relationship between sale price and year sold, but since the coefficient is negative, it means that there is a slight decrease in price when year increases but it's not that obvious.





- (b) Most garages have the lot area of 10 516.82 (mean value) and the selling price in within the range of 180 000-190 000. This time, we will try to check the relationship between lot area and sale price variables. It resulted in a 0.2638 correlation coefficient which means that the two variables have weak correlation. But the graph below would tell us that there is a linear relationship between these 2, which means that in an increase in lot area, there is also an increase in sale price. There are outliers in the scatterplot as seen in the figure below, thus affecting the result in correlation.

