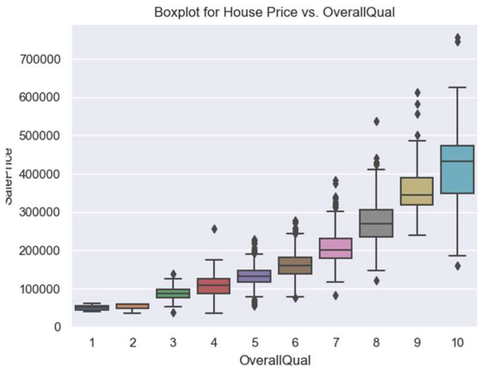
House Price Dataset

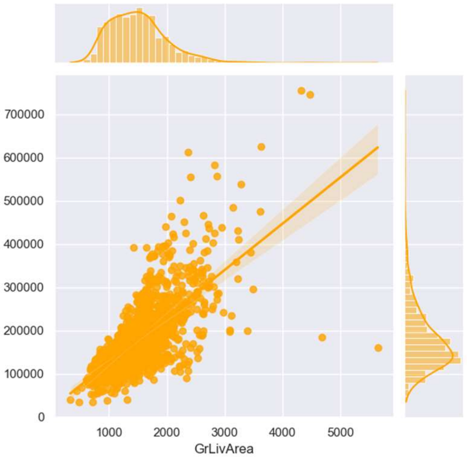
The top three variables that are highly correlated with house sale price (SalePrice) are overall material and finish quality, above ground living area, and garage car capacity.

Overall Material and Finish Quality (OverallQual)



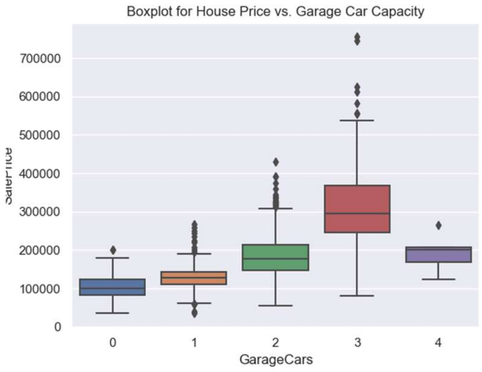
SalePrice and OverallQual has high positive linear correlation coefficient of 0.79. As shown in the boxplot, the higher the OverallQual, the higher is the SalePrice of the house—for instance that it can be observed that house with OverallQual of 10 has a minimum sale price that is higher than the maximum sale price of houses with OverallQual ratings of 1 or 2. The distribution of each OverallQual are generally symmetrical. Outliers are also present in most of the OverallQual boxplots.

Above Ground Living Area (GrLivArea)



GrLivArea variable in this dataset is measured in square feet. SalePrice and GrLivArea has moderate positive linear correlation coefficient of 0.71. The scatterplot shows several non-influential outliers between the two variables. The distribution of GrLivArea is skewed to the right. We can most likely conclude that the larger the above ground living area, the higher the sale price of the house.

Garage Car Capacity (GarageCars)



SalePrice and GarageCars has moderate positive linear correlation coefficient of 0.64. As shown in the boxplot, generally, houses with higher car capacity in their garage has higher price. Several interesting observations: houses with 3 car capacity in their garage have highest sale price in the dataset. The distribution of the data (compared to sale price of the house) for the 3-car garage is right skewed, while the 4-car garage is left skewed. Lastly, more than 50% of the 3-car capacity houses has higher prices than all of the 4-car capacity houses