Danielle Senechal

MAT-374: Data Analytics

Hom­­ework 4

April 14th, 2020

Part 1

1. Based off the scatterplot matrices, it looks like the variables with the strongest correlation to sale price would be year built.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

1. There is no correlation between sale price and lot area, total basement square footage, full bath, bedroom above garage, fireplaces, and car garages. There is very small correlation between sale price and pool area, and medium correlation between sale price and year built.

A screen shot of a computer

Description automatically generated

1. The equation for the linear regression is: ŷ = 1375x – 2530308. The y-intercept is -2530308, and this means that for a house built in the year 0, the sale price would be $-2530308. This would be extrapolation since no house in the dataset was built in the year 0, and 0 is not included on either axis of the linear regression.

A close up of a map

Description automatically generated

1. Multiple R-squared: 0.2734

Since R2 = 0. 2734, this means that 27.34% of the variability in sale price is accounted for by the year the house was built.

1. The typical distance between a predicted value of sale price and the actual value of sale price is $67740. This means that for each value the model predicts it is about $67740 off from the actual value.
2. This model would not be the best predictor of sale price. This is because only 27.34% of the variability in sale price is accounted for by build year, meaning 72.66% of variability in sale price comes from other factors. The model also has an average residual error of $67740, which means for every prediction it makes, it averages $67740 more than the actual value. These two reasons show that the year the house was built is not a good predictor of sale price.

Part 2

1. houses$CentralAir has been reassigned to 1 = Y and 0 = N.

A close up of a screen

Description automatically generated

1. Scatterplot matrices with central air included.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. The highest R2 is in the variables GarageCars, TotalBsmtSF, and YearBuilt when compared to SalePrice.
2. The equation for the mulitiple regression model based on the variables chosen by the stepwise regression:

SalePrice = 40956.04(GarageCars) + 68.98(TotalBsmtSF) + 438.09(YearBuilt) - 827989.16

1. The adjusted R2 for this model is 0.5667. This means that 56.67% of the variability in sales prices is due to the three variables chosen by the stepwise selection.

The typical distance between the predicted value of the sale price and the actual sale price of the house is $52290.

1. The variance inflation factor statistic for GarageCar is 1.535064. For TotalBsmtSF, it is 1.288428, and for YearBuilt, the VIF statistic is 1.470473.

The VIF statistics all do not show multicollinearity. They are all less than two, which is an acceptably low value.