Predictive Analytics Assignment

Aniket Guha Roy

ID: 19200164

Exploratory Data Analysis:

1. **Boxplot**:

* Gives us an idea of the distribution of the sales price to be right or positively skewed.
* Most of the (i.e. around 50%) data is spread around 242.8 (1st quad) and 336.8 (3rd quad).
* Median is around 275 (i.e. 276) which is slightly inclined towards bottom.
* There are no outliers for the distribution and the maximum value is 450 and minimum value is little over 150 (i.e. around 155).

**Histogram:**

* This further strengthens our inference of the distribution of the sales price to be slightly right or positively skewed which is evident from the slight longer tail extending to the right.

**Summary** :

* Provides us with the mean and median which indicates that since mean is greater than median, we ought to have a slight positive skewed distribution.
* It also gives us an idea of 50% of the data lies between 242.8 (1st quad) and 336.8 (3rd quad).

1. We didn’t have any variable as categorical, so we converted the following variables which were either numerical or integer into factors :

bedrooms, garages, garage size and school

**Bedrooms:**

* From the summary and box plot, we infer that house with 2 bedrooms had a higher price than other categories and house with 6 bedrooms had lowest price.
* It also tells us that variation in price for 2 bedrooms (ignoring 6 bedrooms since it has only one data) was the lowest among all categories and almost entire price range was between 300 and 350.

Although difference between 3rd quartile and 1st quartile is larger for distribution of 5 bedrooms but since its tails quite short, Variation for 3 bedrooms was the largest with longest tails out of all the box plots .

* We can also see few outliers for 4 bedrooms, which mean some of the houses went for significantly high prices.

**Bathroom:**

* From the summary and box plot, we infer that house with 3 bathrooms had a higher price (median) than other categories and house with 2 bathrooms had lowest price.
* Houses with 1.1,3.1 bathrooms are negatively skewed and with 2,2.1 and 3 bathrooms are positively skewed. House with 1 Bathroom is normally distributed.
* 2 Bathroom has the maximum variation with the longest tails out of all categories, but 1.1 Bathroom has the maximum variation between 1st quantile and 3rd quantile

**Garage:**

* From the summary and box plot, we infer that house with 3 Garages had a higher price (median) than other categories and house with 0 garages had lowest price(median)
* Houses with 0,1,2 Garages are positively skewed. House with 3 Garages is normally distributed.
* House with 0 garage has one outlier , house with 2 garages has the maximum variation.

1. From the summary, pair wise plot and correlation , the relationships between the response sales price and each of the numeric predictor variables are as follows:

**Price ~ Size:**

* The correlation coefficient is on a lower side-i.e. 0.2014378 indicating a kind of positive relationship with price of the house.
* We can derive the same from the pair plot where we can see mostly a scattered distribution but on a closer analysis, the second plot on the second row indicates that with increase in Size (X-axis), Price (y-axis) also increases with few exceptions.
* Relatively wide confidence interval from -0.02516037 to 0.40834813 indicates the dispersion is high.
* We assume a p-value more than 0.05 is statistically insignificant.

P-value : 0.081>0.05 :  indicates weak evidence against the null hypothesis.

**Price ~ Lot:**

* The correlation coefficient is on a lower side-i.e. 0.2442323 but is the highest of all the other numerical variables.
* This indicates stronger positive relationship with price of the house amongst all the other numerical variables.
* The pair plot predicts an increase in Price with increase in Lot size.
* Relatively wide confidence interval from -0.02516037 to 0.40834813 indicates the dispersion is high.
* We assume a p-value more than 0.05 is statistically insignificant.

P-value : 0.03349 < 0.05 :  indicates strong evidence against the null hypothesis.

**Price ~ Year:**

* The correlation coefficient is quite low -i.e. 0.151248 and is the lowest of all the other numerical variables.
* This explains the scattered distribution in the third plot of the first row, which still hold a positive correlation with price of the house.
* Wide confidence interval from -0.07389863 to 0.36683347 indicates the dispersion is quite high.
* We assume a p-value more than 0.05 is statistically insignificant.

P-value : 0.1837>0.05 :  High p- value depicts a weak evidence against the null hypothesis.

Regression Model:

1. Equation for the Model: Y = + Bsize Xsize+Blot Xlot+Bbath1.1 Xbath1.1+ Bbath2 Xbath2+ Bbath2.1 Xbath2.1+Bbath3 Xbath3+Bbath3.1 Xbath3.1+Bbed3 Xbed3+Bbed4 Xbed4+ Bbed5 Xbed5+Bbed6 Xbed6+Byear Xyear+Bgarage1 Xgarage1+ Bgarage2 Xgarage2+ Bgarage3 Xgarage3+ Bhigh Xhigh+ BNotre XNotre+ BSL XSL+ BSM XSM+ Bstrat Xstrat+ ε
2. The estimate of intercept term Beta0 : -884.3531 which means if all other explanatory variables part of the model is equal to 0, then Beta0 represents the Price.

But this value seems to be exceptional considering the price of a house to be negative . Moreover B0, having negative p value proves that this variable is statistically insignificant.

1. The estimate of intercept term associated with floor size : 59.4503

This means 1 unit increase of floor size will result in a change of 59.4503 for the house when all other explanatory variables are constant.

The p-value if 0.045< 0.05 so we can that the variable is statistically significant.

1. The estimate of intercept term associated with Bath1.1 : 135.8983

This means 1 unit increase of Bath1.1 will result in a change by a factor of 135.8983 for the house when all other explanatory variables are constant.

The p-value if 0.007< 0.05 so we can say that the variable is statistically significant.

1. The predictor variable Bed has a negative impact on the response variable Price which means as the number of bedrooms increases the price of the house decreases. The estimates -228.1052, -238.2609, -237.6155, -255.0211 of houses with 3, 4 , 5 and 6 bedrooms shows a strong correlation but negatively. The p-values ranging from 0.0017 to 0.0054 for all the categories of Bed signifies statistical importance of the variable
2. The predictor variables significantly contributing to the expected value arranged as per increasing p-value are :

* Bed4
* Bed3
* Lot
* Bed5
* SchoolHigh
* Bed6
* Bath1.1

1. Largest Expected Value:

Below Predictor Variables have been considered along B0 to predict the largest Expected value:

* 1. Bo = -884.3531
  2. Size: BSize - 59.4503 ; largest value – 2.896
  3. Lot: Blot - 11.7701 ; largest value - 11
  4. Bath : BBath1.1 - 135.8983 ; Largest value - 3
  5. Bed : BBed: – 228.1052 ; Largest value - 3
  6. Year : BYear: 0.5567 ; Largest value – 2005
  7. Garage : BGarage2 : 18.2435 ; Largest value -2
  8. School High : BSchoolHigh - 113.2774 ; Largest value -High

Largest value of the above predictors has been considered below:

Equation:

59.4503(2.896)

+ 11.7701(11) + 135.8983(1.1) – 228.1052(3) + 0.5567(2005) +18.2435(2)+113.2774(High)

1. Lowest Expected Value:

Below Predictor Variables has been considered along B0 to predict the lowest Expected value:

1. Bo : -884.3531
2. BSize : 59.4503 ; lowest value: 1.44
3. Blot : 11.7701 ; lowest value: 1
4. BBath2 :73.9317 ; lowest value: 2
5. BBed6: – 255.0211 ; lowest value: 6
6. BYear: 0.5567 ; lowest value: 2005
7. BGarage3 : -209.9038 ; lowest value: 2
8. BSchoolSt Louis : 9.0367 ; lowest value: StLouis

Lowest value of the above predictors has been considered below:

Equation:

59.4503(1.44)

+ 11.7701(1) + 135.8983(2) – 228.1052(6) + 0.5567(2005) - 209.9038 (2) + 9.0367(StLouis)

1. From the summary of the plot, we find that mean is 0 and median 0.173 indicating the distribution of the residual is negatively skewed.

The residual vs fitted plot indicates that we see that there is no distinctive pattern and the residuals are fairly scattered around the regression line. The positive values of the residuals mean that prediction was too low and negative values mean the prediction was too high. Since the red line follows the regression line, we infer that the model is good fit and is linear in nature.

Normal Q-Q plot shows that the quantiles of residual follow normality and the distribution is normal

Scale Location plot indicates that there is constant variance along the regression line. So, homoscedasticity is maintained.

We also see that the points 25, 30 and 44 are outliers.

Residuals and Leverage indicate that there are no influential points since no point lies beyond cook’s distance.

There are 3 leverage points 21,47 and 74 as can be deduced from the graph.

10. We use adjusted R-squared to compare the goodness-of-fit for regression models that contain differing numbers of independent variables. In comparison to R- squared, adjusted R—squared only increases if the new term improves the performance of the model. So, we can say with adjusted R-square value of 0.5125 , a little over 50 % of the variation of the error due to significant predictor variables can be explained by the model.

11. F statistic of 4.942 shows that the coefficients in the model are significantly important in predicting the model. Thus, it depicts a strong evidence against the null hypothesis.

Null Hypothesis: All the regression coefficients are equal to zero

Alternate Hypothesis: At least one of the coefficients is significant.

In addition to that, p-value = 1.265 \* 10^-6 is also significantly less. So, we reject the null hypothesis and hence, discard the F-value result.

**ANOVA:**

1. ANOVA type 1:

Df Sum Sq Mean Sq F value Pr(>F)

Size 1 11078 11077.7 6.2426 0.015489 \*

Lot 1 15232 15232.5 8.5839 0.004929 \*\*

Bath 5 36824 7364.7 4.1502 0.002861 \*\*

Bed 4 25502 6375.4 3.5927 0.011310 \*

Year 1 554 554.4 0.3124 0.578474

Garage 3 16101 5367.1 3.0245 0.037179 \*

School 5 70112 14022.4 7.9020 1.153e-05 \*\*\*

The Null hypothesis is testing whether the coefficient of the predictor variable is 0

We see that the p value for the respective f statistic of all the variables except for Year are less than 0.5. The “Year ” having very less F- statistic of 0.3214 and P-value of 0.57 > 0.5. Hence, we cannot reject the null hypothesis and derive that it is not statistically significant, and the coefficient can be termed as 0.

ANOVA type 1 table indicate that the variable “Year” having higher P-value should be removed since this is statistically insignificant.

1. ANOVA type 2:

Res.Df RSS Df Sum of Sq F Pr(>F)

1 55 97599

2 56 102402 -1 -4802.6 2.7064 0.1057

Anova type 2 null hypothesis states that pairs of the coefficients terms are equal to 0.

We create another model (fit 2) without Year variable and find that p-value comes as 0.105>0.05 . Hence, we cannot reject the null hypothesis and infer that Year variable is not significant.

**Diagnostics:**

1. From the added -variable plot, we see that almost all the predictor variables maintain a linear relationship with the response variable (Price) apart from few exceptions such as Garage 1, Garage 2 and School St Louis, School Stratford, that almost maintain a linear relation but are very close to 0.

Bed 3, Bed 4 and Bed 5 have steep slopes indicating strong influence on house price but inversely proportional. Among the positive steep slopes, we have Bath 1.1 have the strongest influence in the positive direction

Component Residual plot

A significant difference between the residual line and the component line indicates that the predictor does not have a linear relationship with the dependent variable.

All the plots seem to suggest a linear relationship with response variable (Price) since the component line almost traces the residual line

Effects of non-linearity on regression model:

* Non-Linearity opens door to huge number of possible forms. Consequently, nonlinear regression can fit an enormous variety of curves. However, because there are so many options, we may need to conduct considerable research to determine which functional form provides the best fit for our data.
* Results in biased estimates and inconsistency in nature

To correct non-linearity, we might apply appropriate transformations on trial and error basis such as:

* sqrt(x)
* 1/x
* log(x)
* n^x.

The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical regression analysis.

Null Hypothesis: The error terms are not auto correlated

Alternate Hypothesis: The error terms are corelated.

The Durbin-Watson statistic will always have a value between 0 and 4.

Values from our model : DW = 1.6142, p-value = 0.02035

Since our p-value comes 0.02 < 0.05, this indicates in favor of alternate hypothesis and conclude that there is auto-correlation between the residuals. Hence all the residuals can’t be termed as independent and identically distributed.

The common types of violations are:

* Data collected from particular sub group.
* Time series data
* Heteroskedasticity

Effects of dependent variables:

* Non constant variance
* Structural dependence between the residuals
* The parameters can be estimated but not interpreted correctly
* Outliers can cause inefficiency of the model and can lead to biasness of the parameters.

Steps to improve the model:

* Use mixed effective model

1. From the VIF table we see that the values of GVIF^(1/(2\*Df)

less than 5 (Threshold value), so we can safely say that there is no multicollinearity.

The correlation plot also gives an idea that the correlations between the variables is very less

Effect of multicollinearity :

* It can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model.
* The result is that the coefficient estimates are unstable and difficult to interpret.
* Multicollinearity reduces the precision of the estimate coefficients, which weakens the statistical [power](https://statisticsbyjim.com/glossary/power/) of your regression model. We might not be able to trust the p-values to identify independent variables that are statistically significant.
* It can make choosing the correct predictors to include more difficult.

To correct the model:

The variance inflation [factor](https://statisticsbyjim.com/glossary/factors/) (VIF) identifies correlation between independent variables and the strength of that correlation.

Statistical software calculates a VIF for each independent variable. VIFs start at 1 and have no upper limit. A value of 1 indicates that there is no correlation between this independent variable and any others. VIFs between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures. VIFs greater than 5 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable

1. From the plots, we infer that the model has almost a Zero conditional mean and homoscedastic distribution barring few higher values at the end.

Studentisized residuals ~ Fitted values: The distribution follows a constant variation around 0 and thus maintains homoskedasticity

Studentisized residuals ~ Size: The distribution follows a constant variation for all the data and maintains homoskedasticity.

Studentisized residuals ~ Lot: The distribution follows a constant variation for all the data and maintains homoskedasticity.

Studentisized residuals ~ Year

The distribution has a constant variance especially from 1960 to 2000. There are few values at the start which is due to sample values collected.

Effects of heteroscedasticity:

* While heteroscedasticity does not cause bias in the [coefficient](https://statisticsbyjim.com/glossary/regression-coefficient/) [estimates](https://statisticsbyjim.com/glossary/estimator/), it does make them less precise. Lower precision increases the likelihood that the coefficient estimates are further from the correct population value.
* Heteroscedasticity tends to produce p-values that are smaller than they should be. This effect occurs because heteroscedasticity increases the variance of the coefficient estimates but the OLS procedure does not detect this increase. Consequently, OLS calculates the t-values and F-values using an underestimated amount of variance. This problem can lead you to conclude that a model term is statistically significant when it is not significant.

Steps to improve the model:

* Re-build the model with new predictors
* If the data are heteroscedastic, a non-linear data transformation or addition of a quadratic term might fix the problem.

1. The histogram of the studentized residuals gives us an of the distribution to be almost normal like a bell curve with a slight inclination towards negative skewness.

Normal Q-Q plot shows that the quantiles are distributed normally along the line. The few observations seem to tail off at the start which is normal due to the sample size. There is one possible outlier since one point is quite far from the line.

Effects of non-normality

* T and F statistics would be measured incorrectly

To improve the model:

* Remove outliers
* We can take Transformations
* Fit other distributions
* We can use non- parametric tests

**Leverage, Inﬂuence and Outliers:**

1. **A Leverage point is an observation** made at extreme or outlying values of the independent variables such that the lack of neighboring observations means that the fitted regression model will pass close to that particular observation is a measure of how far an observation on the predictor variable  from the mean of the predictor variable.

A data point that has an x-value far from the mean of the x- values is said to have high leverage.

Leverage points exist between 0 and 1. A point with zero leverage has no effect on the regression line and a point with 1 leverage will be followed by the regression line

Effects of leverage point:

* Usually have a small residual
* Impact the regression line
* Often strengthens the correlation and R^2 value
* Can be misleading

Leverage points have been identified from the leverage point:

While, leverage plots graph we identify the following points:

**Price ~ Size:** Points 20, 44, 30 and 76 are identified as leverage points .

**Price ~ Lot:** Points 74, 41,44 are identified as leverage points

**Price ~ Bath** : Points 37, 30, 7 are identified as leverage points

**Price ~ Bed:** Points 4, 37,44 are identified as leverage points

**Price ~ Year:** Points 15, 30,44 are identified as leverage points

**Price ~ Garage:** Points 37,44,4 are identified as leverage points

1. Influential Point: A data point is influential if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results. Outliers and high leverage data points have the potential to be influential, but we generally must investigate further to determine whether they are influential.

An influential point is an [outlier](https://stattrek.com/Help/Glossary.aspx?Target=Outlier) that greatly affects the slope of the regression line. One way to test the influence of an outlier is to compute the regression equation with and without the outlier.

Effects:

* It has a large effect on the slope of the regression line
* Usually extreme values but can be low
* It affects the [coefficient of determination](https://stattrek.com/Help/Glossary.aspx?Target=Coefficient%20of%20determination) to be bigger

From the influential plot, we see points 47, 21, 44, and 30 are over the cook’s distance and hence are influential points.

1. Outlier: Data points that diverge in a big way from the overall pattern are called **outlier**.

* It could have an extreme X value compared to other data points.
* It could have an extreme Y value compared to other data points.
* It could have extreme X and Y values.
* It might be distant from the rest of the data, even without extreme X or Y values.

To correct the outliers, we can try out the following ways:

* It’s possible that this is a measurement or data entry error, where the outlier is just wrong, in which case we can delete it.
* It’s possible that what appears to be a legitimate outlier and hence we can consider [transforming the variable](http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#transform) if one of your variables has an asymmetric distribution
* Applying log transformation is the most common method.

From the outlier test we get to know there is one outlier (44) and from the outlier and leverage diagnostics plot, we see there are four outliers 25, 30, 44, 74. But since the cook’s distance for the outliers are much less than 1 as is evident form cook’s D bar plot, so we don’t need to remove them currently and can be investigated further.

**Expected Value, CI and PI:**

A prediction interval reflects the uncertainty around a single value, while a confidence interval reflects the uncertainty around the mean prediction values. Thus, a prediction interval will be generally much wider than a confidence interval for the same value.

We see that the fitted values are within the confidence interval and prediction intervals. Hence, we can say that the model is a good fit or good estimate of the housing prices.