Predictive Analytics Assignment

Student No – 19200078

Name – Kshitij Baranwal

Exploratory Data Analysis:

1) By boxplot, histogram and summary we can say that, sales price is a multi-modal distribution and also it is positively skewed since mean is greater than median (285.8 > 276). The minimum value is 155.5 thousand and maximum value is 450 thousand.
2) After converting categorical data to factors we get following observations:

Price & Bedroom:

The distribution is symmetric for bed 3 and 4. For 2 and 5 data is negatively skewed. For 6, we have only one value and this value is the lowest median amongst all. Bed 4 has some outliers.

Price & Bathroom:

The distribution is symmetric for bath 1 otherwise it is skewed for all other bath. For 2,2.1 and 3 data is positively skewed and for 1.1 and 3.1 data is negatively skewed.

Price & Garage:

The distribution is symmetric for bath 3 otherwise it is positively skewed for all other garage. 0 garage has the lowest median and one outlier and 3 has the highest median. We can observe as number of garage increases price increases.

Price & School:

The distribution is symmetric for St Louis and St Mary's, positively skewed for Alexandra, High School and Stratford and negatively skewed for Notre Dame. St Louis and St Mary's have outliers. The price of house near Alexandra is the cheapest amongst all.

3) From the summary, correlation and the pairs plots, we can observe that there is weak correlation between the sales price and the numeric predictor variables and the highest being with the numeric variable Lot with the value 0.2442.

Regression Model:

1) The equation for the multiple linear regression model is:

Price = $\beta 0 + \beta 1$ Lot + $\beta 2$ Size + $\beta 3$ Year + $\beta 4$ Bath1.1 + $\beta 5$ Bath2 + $\beta 6$ Bath2.1 + $\beta 7$ Bath3 + $\beta 8$ Bath3.1 + $\beta 9$ Bed3 + $\beta 10$ Bed4 + $\beta 11$ Bed5 + $\beta 12$ Bed6 + $\beta 13$ Garage1 + $\beta 14$ Garage2 + $\beta 15$ Garage3 + $\beta 16$ HighSchool + $\beta 17$ NotreDameSchool + $\beta 18$ StLouisSchool + $\beta 19$ StMarysSchool + $\beta 20$ StratfordSchool + ε

2) The intercept value is the expected price of a house built during the mean year with 1 bath, 2 beds and 0 garage, which is near Alexandra college belonging to the mean lot size category, having mean floor size is 376.1016.

3) The estimate of βsize parameter associated with floor size (i.e. β2) is 59.4503. This indicates that if the floor size increases by 1 unit then the price of the house increases by 59.4503.

4) βBath1.1 the parameter associated with one and a half bathrooms is 135.898. There is an increase of 135.8983 thousand euros in the price if the single family wants house with 1 and half bath instead of 1 bath.

5) Summary tells us the estimates of the parameters associated with 3,4,5 and 6 bedrooms (i.e. β9, β10, β11, β12) equal to -228.1052, -238.2609, -237.6155, -255.0211 respectively.

   Interpretation:
   For 3 Bed: The expected price during the mean year with 1 bath, 3 bed, 0 garage near Alexandra belonging to the mean lot size category, having mean floor size is 147.9964.
   For 4 Bed: The expected price during the mean year with 1 bath, 4 bed, 0 garage near Alexandra belonging to the mean lot size category, having mean floor size is 137.8407.
   For 5 Bed: The expected price during the mean year with 1 bath, 5 bed, 0 garage near Alexandra belonging to the mean lot size category, having mean floor size is 138.4861.
   For 6 Bed: The expected price during the mean year with 1 bath, 6 bed, 0 garage near Alexandra belonging to the mean lot size category, having mean floor size is 121.0805.

6) The predictor variables (having p-value <0.05): Lot, Size, Bath, Bed, Garage, School

7) Values that will lead to the largest expected value of the house prices:
   Size: 0.925605
   Lot: 7.01316 1
   Year: 35.59211
   Bath: 1.1
   Bed: 2
   Garage: 2
   School: High School

8) Values that will lead to the lowest expected value of the house prices:
   Size: -0.530395
   Lot: -2.98684
   Year: 64.40789
   Bath: 1
   Bed: 6
   Garage: 3
   School: Alexandra College

9) The most of the residuals fall in -50 and 50 and are evenly spread around 0. Therefore, we can say that there is huge difference in observed and estimated value of response variables.

10) Adjusted R-squared = 0.5125

This tells us that our model is a good fit.

11) Hypothesis:
H0: β1=β2=β3=β4=β5=β6=β7=β8=β9=β10=β11=β12=β13=β14=β15=β16=β17=β18=β19= β20=0
HA: At least one of the βk is not zero.
F-statistic = 4.942.
P-value = 1.265*e^ (-6)
Here, p-value < 0.05. Therefore, we reject H0.
Thus, we have enough evidence to say that at least one of the β's is non-zero.

ANOVA:

1) Hypothesis 1:
H0: β1=0
HA: β1 not equal to 0.
F-statistic = 9.1767 and p-value = 0.003729 which is < 0.05
Hence, we reject H0 and conclude the β1 is not equal to 0 i.e. the variable Lot is significant.

Hypothesis 2:
H0: β2= 0
HA: β2 not equal to 0.
F-statistic = 5.6498 and p-value = 0.020964 which is < 0.05
Hence, we reject H0 and conclude that β2 is not equal to 0 i.e. the variable Size is significant.

Hypothesis 3:
H0: β3= 0
HA: β3 not equal to 0.
F-statistic = 2.6715 and p-value = 0.107872 which is > 0.05
Hence, we do not reject H0 and conclude that β3=0 i.e. the variable Year is insignificant.

Hypothesis 4:
H0: βk = 0 for k=4, 5,6,7,8.
HA: At least one of the βk is not equal to 0 for k=4, 5,6,7,8.
F-statistic = 4.2760 and p-value = 0.002345 which is < 0.05
Hence, we reject H0 and conclude that at least one of the βk is not equal to 0 for k=4, 5,6,7,8.

Hypothesis 5:
H0: βk = 0 for k=9,10,11,12.
HA: At least one of the βk is not equal to 0 for k=9,10,11,12.
F-statistic = 2.8458 and p-value = 0.032393 which is < 0.05
Hence, we reject H0 and conclude that at least one of the βk is not equal to 0 for k=9,19,11,12.

Hypothesis 6:
H0: βk = 0 for k=13,14,15.
HA: At least one of the βk is not equal to 0 for k=13,1,4,15.
F-statistic = 3.0245 and p-value = 0.037179 which is < 0.05
Hence, we reject H0 and conclude that at least one of the βk is not equal to 0 for k=13,14,15.

Hypothesis 7:
H0: βk = 0 for k=16,17,18,19,20.
HA: At least one of the βk is not equal to 0 for k=16,17,18,19,20.
F-statistic = 7.9020 and p-value = 1.153*e^ (-5) which is < 0.05
Hence, we reject H0 and conclude that at least one of the βk is not equal to 0 for k=16,17,18,19,20.

2)'Year' is the predictor variable suggested by Type 1 ANOVA that we should remove from our model.

3) Hypothesis:

H0: β3=0
HA: β3 not equal to 0.
F-statistic = 2.7064 and p-value = 0.1057 which is > 0.05
Hence, we do not reject H0 and conclude that β3 = 0 and the variable Year is insignificant. Thus, we get the same result as that of type 1 ANOVA.

Diagnostics:

1) By performing variable plots, we see Price has linear relation with all predictor variables.
   By performing component-plus-residual plots, we see that Year has somewhat non-linear relation with Price.
   The presence of non-linearity brings in biasedness and inconsistency in the estimates of the parameters and it can be dealt by using transformations, polynomials and splines in the model.

2) By reading the data description, we can say that random/iid sample assumption might not be successfully met by our model as we are using the variable Year in our model.
   Hypothesis for Durbin – Watson test:
   H0: ρ = 0
   HA: ρ not equal to 0.
   Test statistics = 1.614157 and p-value = 0.04 which is < 0.05.
   Here, H0 should be rejected and we can conclude that ρ is not equal to 0 and that autocorrelation exists.
   The test statistic is between 0 and 2. Therefore, there is positive autocorrelation.

The two common violations of the random/iid sample assumption are heteroskedasticity and bias due to outlying observations. We can deal with this with the help of time series analysis and the use of mixed effect model.

3) Correlation matrix for the 3 numeric variables:

	Size	Lot	Year
Size	1.00000000	0.04079199	0.17656934
Lot	0.04079199	1.00000000	-0.03933975
Year	0.17656934	-0.03933975	1.00000000

From above, no two variables are strongly correlated with each other.

Observing the variance inflation factors of which none are greater than 4, we conclude that multicollinearity is absent.

If multicollinearity exists in the model, then the estimates of the parameters cannot receive a reliable interpretation.

In the presence of multicollinearity, we can either remove one of the highly correlated predictor variables or Partial Least Square Regression, Principal Component Analysis or Ridge regression.

4) Observing the studentized residuals vs fitted values plot, we see a symmetric pattern in the spread of the data below and above 0 value. Therefore, there is no heteroskedasticity. The plots of studentized residuals vs the predictor variables suggest the same.
In the presence of heteroskedasticity, the standard errors are biased.
In order to deal with heteroskedasticity, one can make use of the Weighted Least Squares Method.

5) The histogram obtained is symmetric and Q-Q plot the points on a straight line which means the normality assumption is met by out model. The non – normality will lead to the wrong critical values for the t and F tests.
To deal with non – normality, one can make use of transformations, interactions or a completely different model.

Leverage, Influence and Outliers:

1) A leverage point is one with an unusual X value. It affects the model summary statistics but has little effect on the estimates of the regression coefficients. High leverage points have the potential to affect the fit of the model.
Leverage points of this model:

1, 2, 3, 4, 5, 6, 7, 9, 15, 20, 21, 22, 28, 31, 32, 33, 34, 35, 36, 37, 39, 41, 42, 43, 44, 46, 47, 49, 50, 51 52, 54, 56, 57, 58, 64, 66, 69, 71, 72, 73, 74, 76.

2) An influential point is the one whose removal from the dataset would cause a large change in the fit. An influential point may or may not be an outlier and may or may not have large leverage but it will tend to have one of these two properties. High leverage cases are potentially influential and should be examined for their influence.

   Influence points of this model:
   21, 30, 44, 47, etc.
3) An outlier is an observation where the response does not correspond to the model fitted to the bulk of the data. Outliers might affect the estimation of the regression coefficients. In order to detect if an observation is an outlier, the best way is to drop the particular observation and then find the estimates of the model.

   After performing outlier test and the outlier and leverage diagnostics, we get to know there are no outliers in this model.

Expected Value, CI and PI:

By observing the plot, fitted values, Confidence Intervals and the Prediction Intervals we can predict that all values are inside this interval. Therefore, the model provides good estimate of house prices.