Predictive Analytics Assignment

Name: Tanay Umesh Sawant **Student no: 19203264**

Note: From the beginning of the interpretation, I have removed bed 6 of category bed from the data.

Exploratory Data Analysis:

1. **Using a boxplot, histogram and summary. Describe the distribution of the sales price of the houses.**
   By looking at the boxplot, histogram and summary, we can come up to the conclusion that the distribution is skewed, i.e. positively skewed and multi modal.
   The minimum value is 155.5 thousand and maximum value is 450 thousand.

2. **Convert all the categorical variables to factors. Using the summary and a boxplot describe how sales prices vary with respect to the number of bedrooms, bathrooms, garage size and school.**
   After converting the categorical data to factors:
   Price v/s Bedroom:
   - We can see from the boxplot that for houses with 2 beds and 5 beds, the data is skewed and for 3 beds and 4 beds, it is symmetric with a couple of outliers for bedroom with 4 beds.
   - The highest median price is for a house with 2 beds, nearing about 350 thousand euros and the lowest median price is for a house with 4 beds which is just above 250 thousand euros.

   Price v/s Bathrooms:

   - We can observe that the median price for a house with different bath levels lie in the range of 250 thousand euros to 350 thousand euros.
   - The data is skewed for almost all the houses with different levels of baths except it is symmetric for a house with 1 bath.

   Price v/s Garage:

   - We can observe that median price of the house keeps on increasing as the number of garages increase.
   - Data is only symmetric for houses with 3 garages and for rest it is skewed with an outlier for house with 0 garage.

   Price v/s School:

   - We can observe that a house near Alexandra school is the cheapest and a house near School Notre Dame is at the higher side with the highest being near school High
   - There are outliers in the data for houses near St Louis and St Marys School.

3. **Using the summary, correlation and the pairs plots discuss the relationship between the response sales price and each of the numeric predictor variables.**

From the correlation table we can observe that there is no high correlation between the sales price and the numeric predictor variables and the highest being with the numeric variable Lot with the value 0.2381. This can be observed from the pairs plot as well.

Regression Model:

1. **Fit a multiple linear regression model to the data with sales price as the response and size, lot, bath, bed, year, garage and school as the predictor variables. Write down the equation for this model.**
   After rescaling the numeric variables, the equation is:
   Price= βo+β1 Lot1 + β2 size1 + β3 Year1+ β4 Bath1.1 + β5 Bath 2 + β6 Bath 2.1 + β7 Bath 3 + β8 Bath 3.1 + β9 Bed3 + β10 Bed 4 + β11 Bed 5 + β12 Garage 1 + β13 Garage 2 + β14 Garage 3 + β15 SchoolHigh + β16 SchoolNotreDame + β17 SchoolStLouis + β18 SchoolStMarys + β19 SchoolStratford +ε
2. **Interpret the estimate of the intercept term β0.**
   Intercept value is the expected mean value when all the independent X variables are 0. The intercept can be interpreted as, for mean lot size, floor size and during mean year, a house with 1 bath, 2 beds and 0 garage near Alexandra school, the price is **376.8455** thousand euros.
3. **Interpret the estimate of βsize the parameter associated with floor size (Size).**
   The βsize1 can be estimated as there is an increase of **59.4503** thousand euros on an average if we change the floor size by 1 unit.
4. **Interpret the estimate of βBath1.1 the parameter associated with one and a half bathrooms.**
   There is an increase of **135.8983** thousand euros in the price if the single family wants a house with 1 and half bath instead of 1 bath.
5. **Discuss and interpret the effect the predictor variable bed on the expected value of the house prices.**
   According to the model we can see that if the families prefer a house with number of beds more than 2, the price of the house keeps on decreasing as the number of beds keep on increasing. It is observed that the maximum number of houses are with 2 and 3 beds.
   This means that a 2 or 3 bed house is the most preferred by the families.
6. **List the predictor variables that are significantly contributing to the expected value of the house prices**
   The predictor variables that are significantly contributing are:
   a) Lot size b) Bath c) Bed d) School e) size f) Garage
7. **For each predictor variable what is the value that will lead to the largest expected value of the house prices.**
   The values that will lead to the largest expected house prices for each predictor variables are as follows:
   For Lot1: Maximum value is **6.98667** (after rescaling)
   For size1: Maximum value is **0.924507** (after rescaling)
   For Year1: Maximum value is **34.9333** (after rescaling)
   For Bath: The value is **135.8983** corresponding to Bath 1.1 i.e. a house with 1 and half baths.

For Bed: The value is included in the intercept i.e. of **Bed 2.** i.e. a house with 2 beds.
For Garage: The value is **18.2435** corresponding to Garage2 i.e. a house with 2 garages.
For School: The value is **113.2774** corresponding to school High i.e. a house near SchoolHigh.

8. **For each predictor variable what is the value that will lead to the lowest expected value of the house prices.**
   The values that will lead to the lowest expected house prices for each predictor variables are as follows:
   For Lot1: Minimum value is **-3.0133** (after rescaling)
   For size1: Minimum value is **-0.531493** (after rescaling)
   For Year1: Minimum value is **-65.06667** (after rescaling)
   For Bath: The value is included in the intercept i.e. of **Bath 1**. i.e. a house with 1 bath.
   For Bed: The value is **-238.2609** corresponding to Bed 4 i.e. a house with 4 beds.
   For Garage: The value is **-209.9038** corresponding to Garage 3 i.e. a house with 3 garages.
   For School: The value is included in the intercept corresponding to **Alexandra college.**

9. **By looking at the information about the residuals in the summary and by plotting the residuals do you think this is a good model of the expected value of the house prices.**
   Looking at the information on residuals in the summary and plotting them we see that the residuals are spread evenly around 0 but most of them lie in the range of -50, 50 which indicate that there is large difference in the observed and the estimated value of the response variable.
10. **Interpret the Adjusted R-squared value.**
    Adjusted R square value is: 0.517. It suggests that **51.7%** variation in Price is explained by the predictor variables after penalising for all the predictor variables in the model indicating that our model is a good fit.

11. **Interpret the F-statistic in the output in the summary of the regression model. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.**
    To test H0: $\beta1= \beta2= \beta3= \beta4= \beta5= \beta6= \beta7= \beta8= \beta9=\beta10=\beta11=\beta12=\beta13=\beta14=\beta15=\beta16=\beta17=\beta18=\beta19=0$
    v/s H1: at least one of the βj is non-zero.
    The F-test statistic is: 5.169
    The p value is 8.613e-07 which is < 0.05 and hence we reject H0.
    We have enough evidence to say that at least one of the βj's is non-zero.

ANOVA:

1. **Compute the type 1 anova table. Interpret the output. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.**

   The test statistic is:

   Hypothesis 1 (Lot1):
   To test H0: β1=0
   v/s H1: β1 is not equal to zero.
   The F-statistic is: 8.6852
   The p value is 0.004699 which is < 0.05 and hence we reject H0.
   We have enough evidence to say that Lot1 is significant.

   Hypothesis 2 (size1):
   To test H0: β2=0
   v/s H1: β2 is not equal to zero.
   The F-statistic is: 5.5877
   The p value is 0.021647 which is < 0.05 and hence we reject H0.
   We have enough evidence to say that size1 is significant.

   Hypothesis 3 (Year1):
   To test H0: β3=0
   v/s H1: β3 is not equal to zero.
   The F-statistic is: 2.6092
   The p value is 0.111971 which is > 0.05 and hence we do not reject H0.
   We have enough evidence to say that **Year1 is insignificant.**

   Hypothesis 4 (Bath):
   To test H0: βk=0 k=4,5,5,6,7,8
   v/s H1: βk is not equal to zero. k=4,5,6,7,8
   The F-statistic is: 4.3979
   The p value is 0.001936 which is < 0.05 and hence we reject H0.
   We have enough evidence to say that Bath is significant.

   Hypothesis 5 (Bed):
   To test H0: βk=0 k=9,10,11
   v/s H1: βk is not equal to zero. k=9,10,11
   The F-statistic is: 3.5855
   The p value is 0.019331 which is < 0.05 and hence we reject H0.
   We have enough evidence to say that Bed is significant.

   Hypothesis 6 (Garage):
   To test H0: βk=0 k=12,13,14
   v/s H1: βk is not equal to zero. k=12,13,14
   The F-statistic is: 3.0245
   The p value is 0.037179 which is < 0.05 and hence we reject H0.
   We have enough evidence to say that Garage is significant.

Hypothesis 7 (School):
To test H0: βk=0 k=15,16,17,18,19
v/s H1: βk is not equal to zero. k=15,16,17,18,19
The F-statistic is: 7.9020
The p value is 1.153e-05 which is < 0.05 and hence we reject H0.
We have enough evidence to say that School significant.

2. **Which predictor variable does the type 1 anova table suggest you should remove the regression analysis.**
   Year1 is the predictor variable that the anova table suggests to be removed as it is insignificant.

3. **Compute a type 2 anova table comparing the full model with all predictor variables to the reduced model with the suggested predictor variable identified in the previous question removed. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.**
   To test H0: β3=0
   v/s H1: β3 is not equal to zero.
   The F-statistic is: 2.7064
   The p value is 0.1057 which is > 0.05 and hence we do not reject H0.
   We have enough evidence to say that **Year1 is insignificant.**

Diagnostics:

1. **Check the linearity assumption by interpreting the added variable plots and component-plus-residual plots. What effect would non-linearity have on the regression model and how might you correct or improve the model in the presence of non-linearity?**

   Added Variables plot: we can observe that there exists a linear relationship with all the predictor variables individually given that the other predictor variables are in the model.

   We observe that the Year1, Bath 2, Bath 2.1, Bath 3.1, Garage1, Garage2, School St Louis, School St Marys, School Stratford are insignificant as their distribution is near 0 which matches with our summary for the model.

   Component plus residual plots : This helps us to narrow down our conclusion with multiple lines where the blue line indicates the best fit for the data and the pink line is what our fitting is and based on it we can observe that there is not so strong linear relationship between sales price and year and it's the best with the size. For categorical data too we can observe difference in the boxplot as the medians in each category are at different positions.

   Non-linearity would cause biased and inconsistent estimates and this can be corrected by using transformation, polynomials and spline.

2. **Check the random/i.i.d. sample assumption by carefully reading the data description and computing the Durbin Watson test (state the hypothesis of the test, the test statistic and p-value and the conclusion in the context of the problem). What are the two common violations of the random/i.i.d. sample assumption? What effect would dependant samples have on the regression model and how might you correct or improve the model in the presence of dependant samples?**

   Based on the data description, our random/i.i.d. sample assumption would be false due to the inclusion of variable Year in our model.

   Durbin-Watson Test:

   To test H0: There is no autocorrelation
   v/s H1: not H0.
   Test statistic: 1.5976722.
   p value: $0.036 < 0.05$
   The null hypothesis of no autocorrelation is rejected. There is a positive autocorrelation.
   The observations cannot be classed as independent.

   Violations are bias/inefficiency due to outliers and heteroskedasticity which can be corrected by using mixed effect models and time series analysis.

3. **Check the collinearity assumption by interpreting the correlation and variance inflation factors. What effect would multicollinearity have on the regression model and how might you correct or improve the model in the presence of multicollinearity.**

   VIF:

   Based on the GVIF values, we can see that all of our GVIF values are closer to 1, indicating there is no correlation among the predictor variables and that the variance of the βj's is not inflated at all.
   If there exists multi-collinearity in our model, we cannot interpret regression coefficients as the predictor variables would be highly correlated among themselves and this can be corrected by removing the highly correlated predictors in the model or by using a Partial
   Least Square Regression (PLS), Principal Components Analysis, Ridge Regression i.e. use a method that cuts no. of predictor to a smaller set of uncorrelated components.

4. **Check the zero conditional mean and homoscedasticity assumption by interpreting the studentized residuals vrs fitted values plots and the studentized residuals vrs predictor variable plots. What effect would heteroscedasticity have on the regression model and how might you correct or improve the model in the presence of heteroscedasticity.**

   We observe that the data is evenly spread above and below 0 indicating that there is no heteroscedasticity which can be verified with the plot.
   If there exists heteroscedasticity in the model, the standard errors are biased which can be corrected by using Weighted Least Square Method.

5. **Check the Normality assumption by interpreting the histogram and quantile-quantile plot of the studentized residuals. What effect would non-normality have on the regression model and how might you correct or improve the model in the presence of non-normality**

   From both the plots we can conclude that our normality condition is satisfied as histogram is symmetric and we observe a 45-degree line indicating the distributions are equal.

   Non-Normality would cause a major problem by giving us the critical values of t and F tests wrong which can be corrected by using transformations, interactions or a different model.

Leverage, Influence and Outliers:

1. **What is a leverage point? What effect would a leverage point have on the regression model? Use the leverage values and the leverage plots to see if there are any leverage points.**

   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.

   Observing leverage plots and the leverage values, the following are the leverage points:

   1, 2, 3, 4, 5, 6, 7, 9, 15, 20, 21, 22, 28, 31, 32, 33, 34, 35, 36, 38, 40, 41, 42, 43, 45, 46, 48, 49, 50, 51, 53, 55, 56, 57, 63, 65, 68, 70, 71, 72, 73, 75.

2. **What is an influential point? What effect would an influential point have on the regression model? Use the influence plot to see if there is any influence points.**

   An influential point has an unusual Y- value along with an unusual X value. It influences the regression model in its direction which may not be good for the model. 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. Some of the influential points from our graph are 47,44,30,21.

3. **What is an outlier? What effect would an outlier have on the regression model? How would you correct for outliers? Use the outlier test and outlier and leverage diagnostics plot to see if there are any outliers. Deal with the outliers if any are identified.**

   An outlier is an observation where the response does not correspond to the model fitted for the bulk of the data. They might affect the estimation of regression coefficients.

   The best way to detect an outlier is to estimate the model from the rest of the data. For our data, based on the outlier test, we can see that there is no studentized residual with Bonferroni $p < 0.05$ indicating that there is no outlier in our data.

Expected Value, CI and PI:

1. **Plot the observed house prices, their expected vale (fitted value), confidence intervals (in red) and prediction intervals (in blue). Looking at this plot is this model providing a good estimate of the house prices.**
   Based on the plotting, we can conclude that most of the values are lying inside the interval indicating that the model is a good fit.