Predictive Analytics Assignment

Name: MadhuSudhan Ashwath
Student Number: 19203116

❖ Exploratory Data Analysis:

1. Using a boxplot, histogram and summary. Describe the distribution of the sales price of the houses.

    - We can see that **sales price** mostly lays between **242k euros to 336 euros**.
    - The **highest** sales price of the house is **450k euros**.
    - The **lowest** sales price of the house is **155.5k euros**.
    - **Mean** is towards the **right of median**, that represents the **distribution is positively skewed** when seen in boxplot and histogram.
    - About **25%** of the data has the sale price under **242.8.**
    - About **50%** of the data has the sale price under **276.0** and **75%** of sales price fall under **336.8.**

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.

- **Price v/s Bedroom:**
    - It is seen that the average sales **price** tends to **decrease** as the number of **bedrooms increases**.
    - The price vary is **less** when the house has **2 and 6** bedrooms and **high** variance when the house has **3, 4 and 5** bedrooms.
    - We can see some **outliers** when houses have **4 bedrooms**.
    - On an average we can see that price of houses with **2 bedrooms** is **larger** than any and houses with **6 bedrooms** are **lower** than any.

- **Price v/s Bathroom:**
    - Overall as the number of **Bathroom increases** from 2, the **Price** goes on **increasing**.
    - **Maximum** Sale Price is for the House with **3 bathrooms**.
    - **Minimum** Sale price is for the House with **2 Bathrooms**.
    - The distribution of the Sale Price is **negatively skewed** for Bath1.1 and **positively skewed** for Bath2, Bath3, and Bath3.1.

- **Price v/s Garage:**
  - Overall as the number of **car capacity** in the house **increases** the **sale price** of the house seems to **increasing gradually**.
  - House with **2 car** capacity has the **maximum** sale Price.
  - House with **no car** capacity has the **lowest median** Sale Price.
  - House with **1 car** capacity has the **lowest** Sale Price.

- **Price v/s School:**
  - Summary shows that more **houses are near** St. Mary's School followed by St. Louis, Notre Dame and High school.
  - The houses **near** High School shows **higher** price.
  - The highest variance in prices when the houses are near **Alexander School** and **High School** and the distribution is **positively skewed**.
  - The houses near **St Mary's** also show **high variance** in price and have a **normal distribution**.

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

- It's seen there is **good correlation** of house price of **0.2014 and 0.244** with **Size and Lot** variable respectively.
- **Year** has a **poor correlation** with the house price with the value of **0.154**.
- From the pairs plot and correlation table we can say that the **correlation between the price and numerical variables are pretty low**.

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.

- Price = β0 + β1Lot + β2Size + β3Year + β4Bath + β5Bed + β6Garage + β7School + ERROR.

- model <- lm(house$Price ~ house$Lot + house$Size + house$Year + house$Bath + house$Bed + house$Garage + house$School, data = house)

2. Interpret the estimate of the intercept term β0.

   - The estimated intercept term **β0 = 376.1016.**
   - The p-value is **1.36e-09** we can see that the value can be approximately be equal to zero, so the value is significant.

3. Interpret the estimate of βsize the parameter associated with floor size (Size).

   - The estimated value of **βsize = 59.4503.**
   - By the value of βsize we can say that for every **1 unit increase** in the parameter size, the price value will increase by **59.4503**.
   - The p-value of size is **0.04501** which is less than alpha value, so the value is significant.
   - The change in parameter of size **vary** in a scale of **0 – 0.04501.**

4. Interpret the estimate of βBath1.1 the parameter associated with one and a half bathrooms.

   - The estimated value of **βBath1.1 = 135.8983.**
   - By the value of βBath1.1 we can say that **addition of 1 bathroom** the price value will increase by **135.8983**.
   - The p-value of size is **0.00779** which is less than alpha value, so the value is significant.
   - The change in parameter of Bath1.1 **vary** in a scale of **0 – 0.00779.**

5. Discuss and interpret the effect the predictor variable bed on the expected value of the house prices.

   - It is seen that as the number of **bedrooms increases** for 2 to 6 the **price goes on decreasing**.
   - We can see that as we **increase** bedrooms to **3,4,5,6** the **price decreases** by **228.10, 238.26, 237.61, 255.02 euros** respectively, when compared with houses with 2 Bedrooms.

6. List the predictor variables that are significantly contributing to the expected value of the house prices

   - The p-values in the model which is less than 0.05 (alpha value) significantly affect the expected value of Price.
   - Below is a table of all the values :

No	**Variable**	**Price**
1	Lot	0.0029
2	Size	0.04501
3	Bath1.1	0.00779
4	Bed3	0.00211
5	Bed4	0.00177
6	Bed5	0.00299
7	Bed6	0.00543
8	Garage3	0.01193
9	High School	0.00334
10	Notre Dame school	0.02730

7. For each predictor variable what is the value that will lead to the largest expected value of the house prices.

   - Below is a table of all the values :

No	Predictor Variable	Value/Level	Max Price
1	Lot	7.013	458.6
2	Size	0.925	431.1
3	Bed	Bed 3	450
4	Bath	Bath3	450
5	Garage	Garage2	450
6	School	School High	450
7	Year	35.59	395.9

8. For each predictor variable what is the value that will lead to the lowest expected value of the house prices.

   - Below is a table of all the values :

No	Predictor Variable	Value/Level	Lowest Price
1	Lot	-2.986	343.1
2	Size	-0.53	344.5
3	Bed	Bed 4	155.5
4	Bath	Bath2	155.5
5	Garage	Garage1	155.5
6	School	School High	Alex
7	Year	-64.4	340

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.

   - **Residuals:** The **difference** between the **observed value** and the **Predicted Value.**
   - By the **residual plot** we can see that points are **randomly distributed** around the **horizontal axis**, hence we can say that the model **doesn't** have a **non- linear relationship** and the model will be a **good fit**.
   - We can see that **Mean and Sum** value are **near** equal to **Zero**.

10. Interpret the Adjusted R-squared value.

   - **Adjusted R-Squared Value:**
     - Value that helps us to **predict how good** is the model based on the **number of explanatory variables** of the model.
     - Higher the Adjusted R-Squared Value, **Higher the accuracy**.
     - Here it helps in understanding the **variation in the Price** over the **predicted variables**.
   - The **Accuracy** of the model created is **51.25%**.

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.

- **Hypothesis:**
  - Used to infer the result of a hypothesis performed on sample data from a larger data set.
  - This test tells the user whether or not his primary hypothesis is true.
  - **H0** -> All the predicted variables are **zero**.
  - **H1** -> At least one of the predicted variable is **Non-Zero**, **Reject hypothesis test.**

- **F-Statistics:**
  - Tells about the **group of variables are jointly significant**.
  - **F-Value** of the model is **4.942** with **20 and 55 Degrees of Freedom**.
- The **p-value** of the model is **1.265e-06**, as it is **less than 0.05**(alpha value) we can **reject** the hypothesis test.
- Hence we can say that at least **one variable** 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.

- **Type 1 Anova :**
  - Used to predict which variable on **addition** to the model will have **significant impact** on the model.

- **Hypothesis Test:**
  - **H0 ->** Mean value with in the group is same.
  - **H1 ->** At least one of the mean value in the group is different.
- From the anova table it is clear that, there is at least one predictor variable which is **significant rejecting the null hypothesis**.
- As we see the **F-Statistics** value of **Year** is **1.645** which has a huge difference when compared to the value in f-table.
- The **P-value** of the **Year** is **0.10565** which is less than 0.05(alpha value). Hence it **isn’t a significant parameter**.

2. Which predictor variable does the type 1 anova table suggest you should remove the regression analysis.

   - The parameter **Year can be removed** from the model as it doesn't contribute more to the model.

3. Compute a type 2 anova table comparing the full model with all predictor variables to the the reduced model with the suggessted 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.

   - **Hypothesis Test:**
     - **H0 ->** βYear equal to Zero.
     - **H1 ->** βYear not equal to Zero.
   - We can see that the **F-statistics** value is **2.7064** and the **p-value** is **0.1057** which is **less** that **0.05**(alpha) thus we **can't reject** the null **hypothesis** test as the parameter year is **not contributing** significantly to the model when compared to other parameter's.
   - Hence it can be considered as the **non-significant parameter.**

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?

   - We can see that the size, bath, lot, high school and Notre Dame School variables show a **positive linear relationship**, bed and Garage3 has a **negative linear relationship**.
   - Year, Garage1, Garage2 and rest of the School parameter's doesn't exhibited **good linear relationship** when compared to other parameter's.
   - Non Linearity can cause beta estimates to become biased and provide inconsistent estimates.

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?

- **By hypothesis of Durbin Watson test:**
  - H0 -> No autocorrelation in the model.
  - H1 -> Autocorrelation in the model.
  - As per the test, the **test statistic** value should be in between **1.5 and 2.5** and corresponding **p-value** should be above **0.05**.
  - In our model **test statistics** value is **1.587** which is the range with a **p-value** of **0.044** which is below 0.05, thus we can **reject hypothesis** and **there exists autocorrelation** in the model.
- There may be some errors due to **repeated measurement** of the parameter's on different time intervals, **multiple measurement** of same parameter. This may result in **structure dependence, non-constant variance and biased outliner**.
- This issue can be rectified be using **mixed effect model.**

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.

- **GVIF value** of all the regression is **less than 5** hence they are said to be **independent** of each other. Since certain parameters has more than **1 degrees of freedom** we have considered the **GVIF^ (1/ (2*DF))** value for the **interpretation**.
- **Problems of Multi-collinearity:**
  - If there exists a **strong correlation** between the **two predictor** variables then their **Beta** becomes **unstable**, the estimate of the Beta will **strongly** depend on the **other predictors** that are included in the model.
  - If the **predictors** are **correlated** then we **cannot interpret** the regression coefficients.
- **Improvement Measures:**
  - **Removal** of **highly correlated** predictors from the model.
  - Use Partial Least Square Regression, Principal Component Analysis, and Ridge regression.

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.

   - By the observation of **studentized residual v/s fitted value plot**, we can see the **zero conditionality** since all the points are lined up **against** the zero and the **band** which they lie around shows that they have **constant variance**.
   - By the observation of **studentized residual v/s predicator variables**, all plots shows heteroscedasticity since there is a **constant distribution** of the variance across zero.
   - By the observation of **boxplot** the **median** is almost **same** for various categories hence it shows heteroscedasticity.
   - **Standard errors** are **biased or distorted** due to the effect of heteroscedasticity.
   - Thus to correct this we can use **Weighted Least Squares technique.**

5. Check the Normality assumption by interpreting the histogram and quantilequantile 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.

   - **Histogram** shows a bell shape curve which seems **normally distributed** around the **mean** and also the Q-Q plot shows that **errors** are **normally distributed** as it tracks the line over the most part of the plot.
   - Critical values of f-test and t-test can **go wrong** due to **Non-normality- effects**.
   - Using transformation of response/predictor variables, or interaction model, or building a different model, our model can be **improved or corrected.**

❖ 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 is any leverage points.

   ➢ **Leverage Point:**
   - The leverage measures the amount by which the predictive value would change if the observation was shifted one unit in the y-axis.
   - The leverage always takes the values between 0 and 1.
   - Point with zero leverage has no effect on the regression model.
   - If a point has equal to 1 the line must follow the point perfectly.

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.

   ➢ **Influence Point:**
   - An influential point is the one if removed from the data would significantly change the fit.
   - An influential point may either be an outlier or have large leverage, or both, but it will tend to have at least one of those properties.
   - High leverages cases are potentially influential and should be examined for their influence.

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 is any outliers. Deal with the outliers if any are identified.

   ➢ **Outlier:**
   - 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 regression model and may change the model equation which may result in wrong prediction of model.
   - Outliner can be found by estimating the model with other data apart from the outliner.

❖ 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.

   - **Confidence Intervals [CI]:**
     - A confidence interval is a type of interval estimate, computed from the statistics of the observed data that might contain the true value of an unknown parameter.
     - Confidence intervals consist of a range of potential values of the unknown parameter.

   - **Predication Intervals [PI]:**
     - Prediction interval is an estimate of an interval in which a future observation will fall, with a certain probability, given what has already been observed.
     - Prediction intervals are often used in regression analysis.

   - **By the observing the graph we can say that this model doesn't have any outlier and provides the best estimate for house price compared to the models with outliers.**