ASSIGNMENT 02

**Name: Gayatri Krishna**

**Reg No: 21BDA16**

DATASET - House Sales Analysis for King County

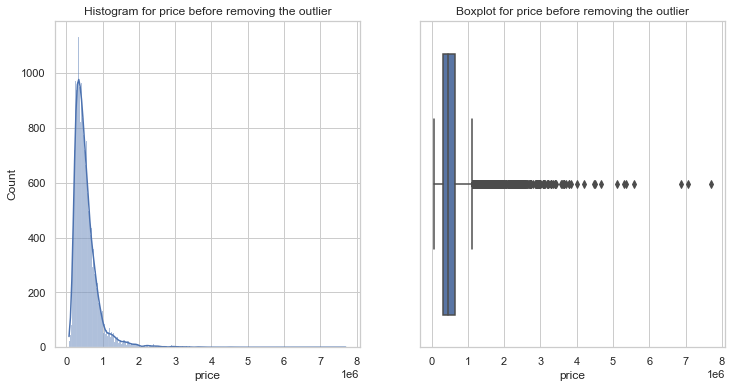
The dataset is about the house sales analysis for King Country. The dataset is

collected from Kaggle (<https://www.kaggle.com/datasets/harlfoxem/housesalesprediction>).

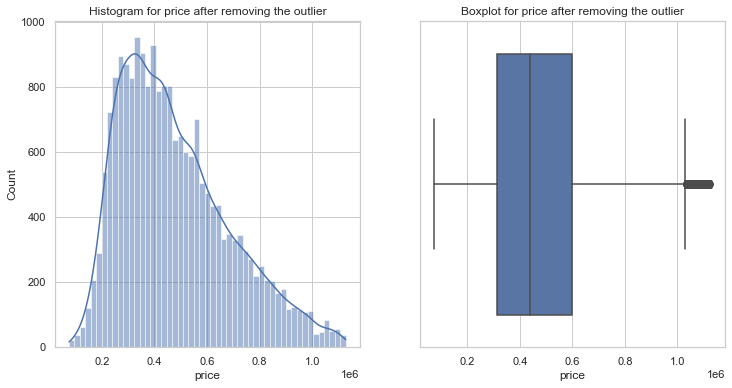
* The dataset consists of 21613 rows and 21 columns. The dataset has no missing missing values. But after cleaning the data the datast has 20467 rows and 21 columns.
* This dataset contains house sale prices for King County. It has homes sold between May 2014 and May 2015.

|  |  |  |  |
| --- | --- | --- | --- |
| **COLUMNS** | **MEANINGS** | **COLUMNS** | **MEANINGS** |
| Id | a notation for a house | yr\_built | Built Year |
| date | Date house was sold | yr\_renovated | Year when house was renovated |
| price | Price is prediction target | zipcode | zip code |
| bedrooms | Number of Bedrooms/ | lat | Latitude coordinate |
| bathrooms | Number of bathrooms | long | Longitude coordinate |
| sqft\_living | square footage of the home |
| sqft\_lot | square footage of the lot |
| floors | Total floors (levels) in house |
| waterfront | House which has a view to a waterfront |
| view | Has been viewed |
| condition | How good the condition is Overall |
| grade | overall grade given to the  housing unit, based on King County grading system |
| sqft\_above | square footage of house apart from basement |
| sqft\_basement | square footage of the basement |

* The dataset has a lot of categorical variables. The below flow chart represents the categorical variables.
* The maximum and minimum price of the houses are $7700000 and $75000 and the mean price is $ 540088.
* Here we are considering our target variable as “price”. We see from the below plots that the price variable is positively skewed. There’s many outliers from the price data.

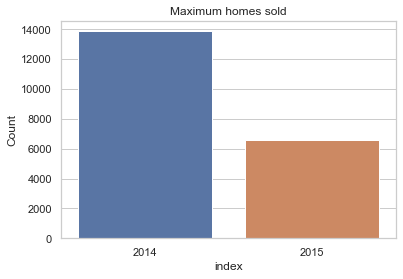
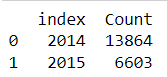


* So, we have to remove these outliers and for this we created a function. Inter Quartile Range approach to finding the outliers. To define the outlier base value is defined above and below datasets normal range namely Upper and Lower bounds, define the upper and the lower bound (1.5\*IQR value is considered). After all the outlier are removed the dataset observations are reduced and the plot for the price variable looks similar to a normal distribution curve(bell shaped).

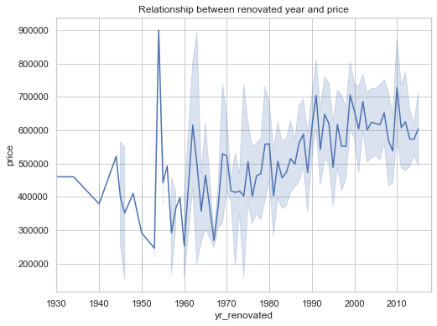


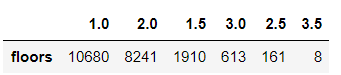
* **Exploratory data analysis:**

1. The dataset consists of data recorded from May 2014 to May 2015. Using a bar plot, we can find out which year had Maximum sales.



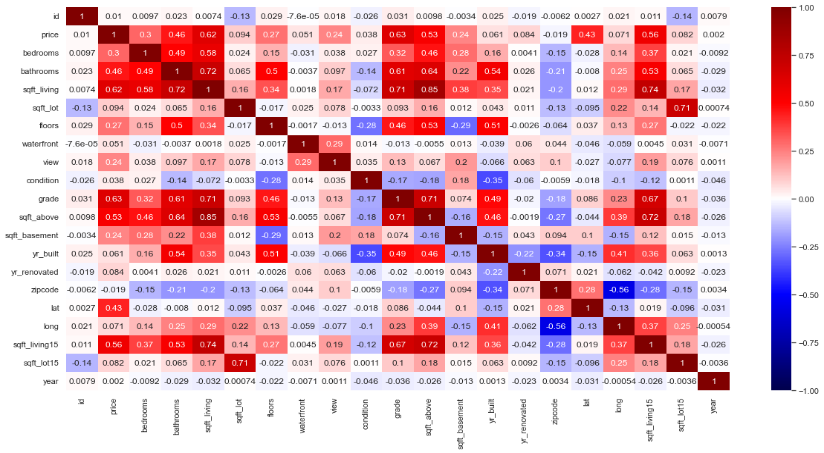
Clearly 2014 has maximum home sales. The total sales in these 2 years were 20467 homes of which 13864 sales alone happened in the year 2014.

1. The seaborn library also provides lineplot. With the lineplot we can visulise the relation between the renovated year and the price of the house. From the plot we see that there is no relation between these two variables. The correlation between the is very very small (ie) 0.08
2. The below data frame show the count the number of houses with unique floor values. It is clear that the least is for hoses with 3.5 floors. Most of the houses have only 1 level.



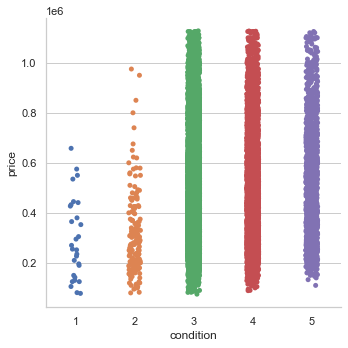
1. **Correlation**:

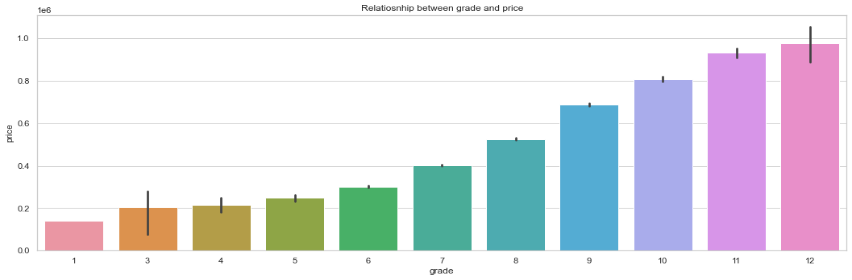
Correlation is a statistical method of to understand the relationship between two variables. With the help seaborn package we can construct a heat map.

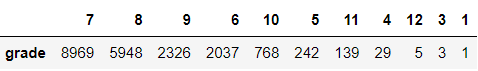


From the heat map we see that the diagonal values is 1 (ie) correlation of variable with itself is a perfect positive correlation. With the help of various shades of colours ranging from dark red to dark blue we can know the degree of correlation.

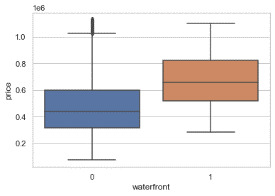
1. **View and Quality Features**:

We know that the variable condition is a categorical variable. We can find out if the condition of the house affects the price. With the help of a bar chart we can conclude that among the 5 categories in condition variable the price of the house is more for condition 5 followed by condition 3.

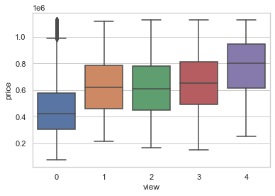




The above plot is a bar char that shows which grade has the highest rating. Grade of a house plays an important role in deciding the price of the house. The highest grade is 12. The indiviual count of the grade is obtained and shown below and we see that grade with 7 has maximum count with mean price $ 432473.0 followed by grade 8 with mean price of $ 538748.

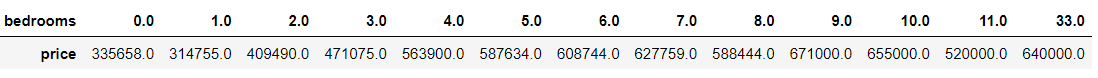


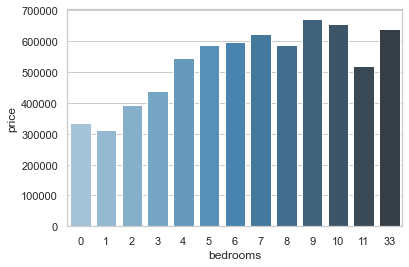
The price of a home is affected by its proximity to the water. The median price of residences with waterfronts versus those without waterfronts differs significantly.

We can see the same pattern from a distance, but the pricing is not dramatically different, but it does have an impact on the house price (positive effects).

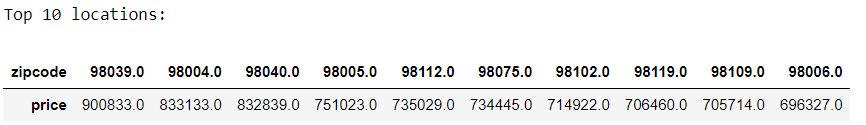
1. **Price & Bedrooms**:

The variable bedroom is a categorical variable that tells us the number of bedrooms in a house. The dataset consists of houses that have no bedrooms. The average price of the houses that is

grouped with number of bedrooms is found and also the same is plotted on a bar chart.

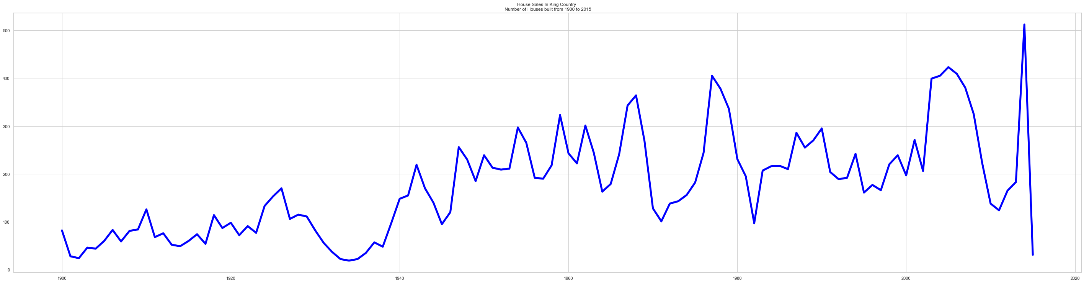


1. **Price & Zip Codes**:

Fixing the price of a house depends on many factors. One major factor is the location (zip code). If the location is a prime location obviously the price of the house also increases. The top 10 locations are displayed below. Location with zip code 98039 has the highest price.

1. **House Sales**:

We have houses that are constructed from 1900 to 2015. From below chart, we gain some insight that buildings have been built are increasing (has increasing trend) from 1900 - 2015. Although there are fluctuations in the amount of buildings construction each year.



* **Simple Linear Regression Model:**

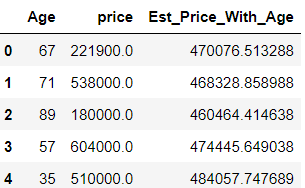
Regression is a statistical method that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

In our dataset we have a variable yr\_built. This variable tells us the year the house was built. We can create a column called Age that is basically the difference between the current year (ie) 2022 and the year the house was built.

We wish to find if the age of the house has any effect on the price of the house. So first it self we see that the correlation between the two variable is negative (very week). We fit a SLM for these two variables and we get the regression equation as:

**Price(Y) = 499349.722 + (-436.9135 \*Age)**

1. We can say that when the age of house is zero (ie) constructed now then the price of the house is $499349.722.
2. The slope value is negative (ie) -436.9135. This means that as age of the house increases the price drops.
3. We also find the value of R square. Here the value of R2 is 0.373857%. This means that only 0.37% of information about the price of the house is determined by age, which is really a small amount of information. Rest 99.66% information is explained by the variables mentioned in the data or those variables that are not present in the dataset.
4. The estimated price of the house for first-five observations is:



* **Multiple Linear Regression Model:**

For fitting the multiple regression model the independent variables x that are taken are: sqft\_living(x1), grade(x2), sqft\_above(x3), sqft\_living15(x4), bathrooms(x5), view(x6) and sqft\_basement(x7). The independent variable here is again price.

We wish to find if these independent variables will affect the price of the house.

The multiple linear regression model is:

Price(Y) = **-302402.997 + (58.62\*x1) +(77889.82\*x2) +( 7.34\*x3) + (40.02\*x4) + (-1.6232\*x5) +(37381.36\*x6)+(51.28\*x7)**

1. We can say that when the all the independent variables are zero then the price of the house drops by $302402.997.
2. Most of the slope values are positive (ie) if there is a 1 unit increase in x1, x2, x3, x4, x6 or x7 then price also increases by 58.62, 77889.82, 7.34, 40.02, 37381.36 or 51.28 units. But the slope of x5 is negative which means that if there is a 1 unit increase in x5 then price of house drops by 1.6232 units.
3. We also find the value of R square. Here the value of R2 is 49.061%. This means that only 49.061% of information about the price of the house is determined by these independent variables. Rest 50.93% information is explained by the variables that may or may not be present in the dataset.

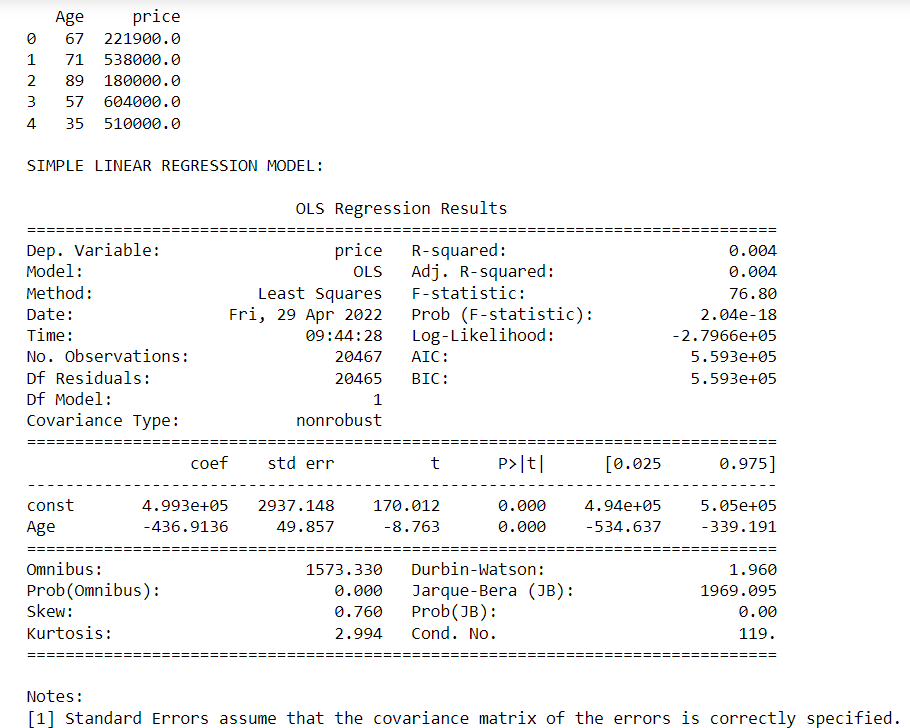
* **Assumptions of Regression Model**:

The assumption of a regression model are:

1. Linear relationship
2. Multivariate normality
3. No or little multicollinearity
4. No auto-correlation
5. Homoscedasticity

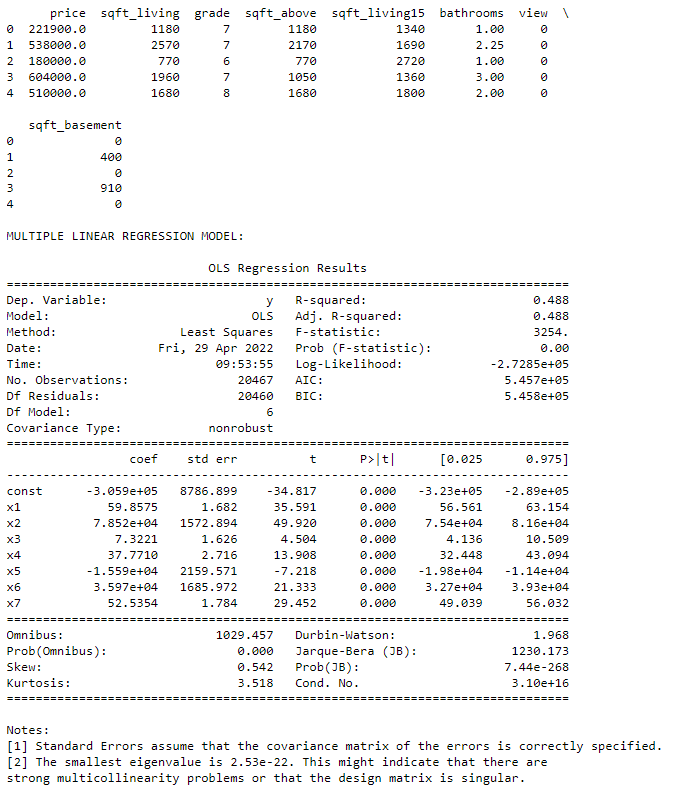
But there is definitely multicollinearity among independent variables. In order to determine which variables to drop as a result of multicollinearity, we can use the variance inflation factor (VIF). The variance inflation factor measures the amount of multicollinearity among independent variables.

* **Statsmodels Package:**

1. For Simple Linear Regression Model:

* We can say that when the age of house is zero (ie) constructed now then the price of the house is $499349.722.
* The slope value is negative (ie) -436.9135. This means that as age of the house increases the price drops.
* We also find the value of R square. Here the value of R2 is 0.4%. This means that only 0.4% of information about the price of the house is determined by age, which is really a small amount of information. Rest 99.66% information is explained by the variables mentioned in the data or those variables that are not present in the dataset.

1. For Multiple Linear Regression Model:



* For fitting the multiple regression model the independent variables x that are taken are: sqft\_living(x1), grade(x2), sqft\_above(x3), sqft\_living15(x4), bathrooms(x5), view(x6) and sqft\_basement(x7). The independent variable here is again price.
* We can say that when the all the independent variables are zero then the price of the house drops by $302402.997.
* Clearly the coefficients of the independent variables are mentioned.
* The R2 value is 48.8%. This means that only 48.8% of information about the price of the house is determined by these independent variables. Rest information is explained by the variables that may or may not be present in the dataset.
* **Difference Between sklearn and Statsmodel libraries:**

The way the two libraries handle constants is a significant distinction. Statsmodels' OLS class offers a function that adds a constant to a given array, while Scikit-learn allows the user to choose whether or not to add a constant via an argument.

ANSWER THE FOLLOWING -

1. **What are the assumptions of linear regression?**

Regression is a type of Machine learning which helps in finding the relationship between independent and dependent variable. A linear regression model is predicated on four assumptions:

* Linearity refers to the relationship between X and the mean of Y.
* Homoscedasticity: For every value of X, the variance of the residual is the same.
* Independent observations: Observations are not reliant on one another.
* Normality: Y is normally distributed for any fixed value of X.

1. **How can we evaluate a Regression model? Define each metric and its interpretation.**

In regression, there are three main measures for evaluating models:

* R Square/Adjusted R Square.
* Mean Square Error (MSE)
* Root Mean Square Error (RMSE)
* Mean Absolute Error (MAE)

MAE is a straightforward metric that determines the absolute difference between actual and projected values.

MSE is a widely used and straightforward statistic that accounts for a small change in mean absolute error. Finding the squared difference between the actual and anticipated value is defined as mean squared error.

The R Square score is a metric that measures the performance of your model, not the loss in terms of how many wells it performed. It is also known as Coefficient of Determination or sometimes also known as Goodness of fit.

1. **Can R squared be negative?**

* R squared isn't always the square of anything, therefore it can be negative without breaking any mathematical principles. Only when the chosen model does not follow the data trend does R squared become negative.
* For equations without a constant term, it is possible to generate a negative R-square. Because R-square is defined as the proportion of variation explained by the fit, it is negative if the fit is worse than fitting a horizontal line.

1. **What is dummy variable trap?**

* The label encoding process can be used to convert categorical information to numerical attributes (label encoding assigns a unique integer to each category of data).
* However, this approach is not suitable on its own; as a result, In regression models, one hot encoding is employed after label encoding.
* This allows us to construct additional attributes based on the number of classes in the category attribute; for example, if the categorical attribute has n categories, n new attributes will be formed. Dummy Variables are the attributes that are created. As a result, in regression models, dummy variables serve as "proxy" variables for categorical data.
* These fake variables will be constructed using one-hot encoding, with each attribute having a value of 0 or 1, indicating whether it is present or not.
* The Dummy variable trap is a situation in which multiple qualities are highly correlated (Multicollinear), and one predicts the value of others. When categorical data is handled using one-hot encoding, one dummy variable (attribute) can be predicted using other dummy variables.

1. **Is One Hot Encoding different from Dummy Variables?**

Encoding categorical variables can be done in two ways. Let's say there are n values in a category variable. It is converted into n variables using one-hot encoding, and n-1 variables using dummy encoding. If we have k categorical variables with n values each. Hot encoding produces kn variables, whereas dummy encoding produces kn-k variables.

1. **How is polynomial regression different from linear regression?**

* Polynomial regression is a type of linear regression in which additional polynomial factors are added to linear regression to convert it to polynomial regression due to the non-linear relationship between the dependent and independent variables.
* Polynomial provides the best approximation of the relationship between the dependent and independent variable

1. **Interpret the screenshot below from the notebook we discussed in class today:**



* The regression predictions exactly fit the data if the R2 is 1. When the model fits the data worse than the poorest possible least-squares predictor, R2 values outside the range 0 to 1 occur (equivalent to a horizontal hyperplane at a height equal to the mean of the observed data).
* As you add more parameters, R2 continues to rise. R2=1 denotes a perfect fit. That is, you've explained all of the variance that there is to explain. In ordinary least squares (OLS) regression (the most typical type), your coefficients are already optimised to maximise the degree of model fit (R2) for your variables and all linear transforms of your variables.
* You can always get R2=1 if you have a number of predicting variables equal to the number of observations, or if you've estimated an intercept the number of observations - 1.

1. **We saw Sweetviz as an Automated EDA option. What are the other options? Try a few of them and share which one did you find the best.**

The 4 differed types of automated EDA are:

1. dtale
2. pandas profiling
3. sweetviz
4. autoviz

D-Tale is one of the best EDA libraries available, in my opinion. Based on the level of customizations accessible within the library, it outshines the competition. In other words, it achieves the goal of exploratory data analysis by allowing you to delve into the minutiae of your dataset. It has a code export capability that allows you to regenerate/recreate any plot or analysis you produced during your exploration.