Ban210 – Final Assessment Report
By:- Laxmikant Mukkawar
Student Id:- 168554210

Prediction of housing prices

1. Introduction

The dataset contains information regarding various aspects of houses. The information includes and is not limited to characteristics such as house style, quality, condition, built year, no of bedrooms, bathrooms, fireplaces, area of living room, garage style, foundation type, etc. A key variable here is the "SalePrice" variable (selling price of a house). It is understood from the dataset that all other variables have the possibility to impact the selling price of a house. Hence, for analytic purposes, "SalePrice" variable has been selected as the dependent variable and all other variables as independent variables.

2. Objective

The main objective of the analysis is to make use of the available dataset to make predictions of housing prices considering the impact of several other variables as mentioned above. Other aims of the analysis is to extract meaningful relationships between variables and to determine if any key variables impact the selling price more than others.

3. Analysis

Initially, an exploratory analysis was carried out on the dataset using Python to better understand the information and to obtain a holistic view of the dataset. It was observed that there were 300 entries spanning over 32 variables.

3.1 Understanding data types

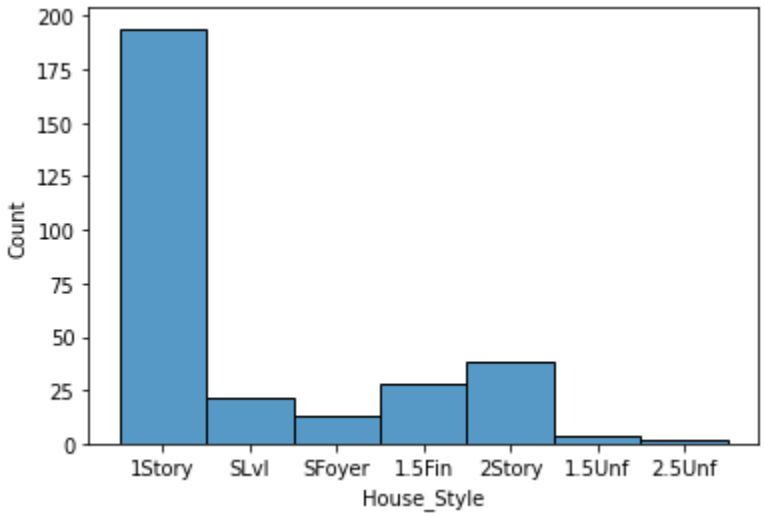
.info() and **.columns** functions indicated that the data was in 3 different datatypes; float64 (3 variables), int64 (21 variables), and object (8 variables). It also showed that the variable or field named "score" did not have any values. Meanwhile, the majority of the variables had all 300 entries while some of them were seen to lack a few entries.

3.2 Distribution of data within variables

Subsequently, plots were executed, using the matplotlib library, to understand the distribution of some of the variables. The idea behind this was to determine the nature of those variables and gain a deeper understanding of the contents of each variable. Some examples of the plots generated are shown below:

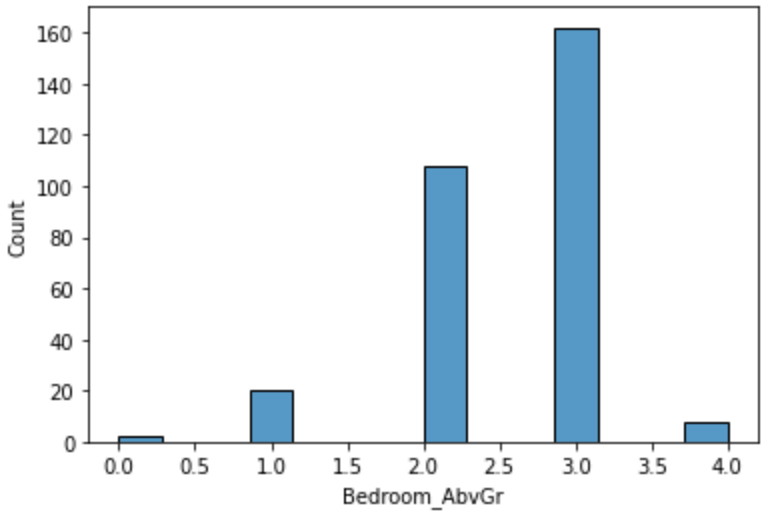
Graph 01 – Housing styles

This plot illustrates that, by far, the majority of the house in the dataset were single-storied. The second place was for 2 story houses, followed by one and a half stories with the second level finished.



Graph 02 – No. of bedrooms

The above graph indicates that the dataset consisted of houses with bedroom numbers ranging from 0 to 4, with most houses having 3 bedrooms followed by 2 bedrooms. Out of the 300 entries, 270 had either 2 or 3 bedrooms.



3.2 Removing unnecessary variables

Next, the different variables were analyzed to determine which variables needed to be considered for analysis. It was identified that variables “PID” and “Score” were redundant. PID did would not add any analytical meaning to the data due to the fact that it is just an ID that was created to identify the entry. Variable Score did not contain any values. Hence, these two variables were removed from the dataset.

3.3 Dealing with null values

.isnull() variables was used to reconfirm the number of missing values in each variable. It showed that Garage_Type_2, Masonry_Veneer, and Lot_Shape_2 had 32, 2, and 1 missing values respectively. The entries pertaining to the null values were then removed from the dataset in order to obtain a final analysis-ready dataset with no missing or null values. This resulted in a dataset containing a total of 266 entries after the removal of null entries.

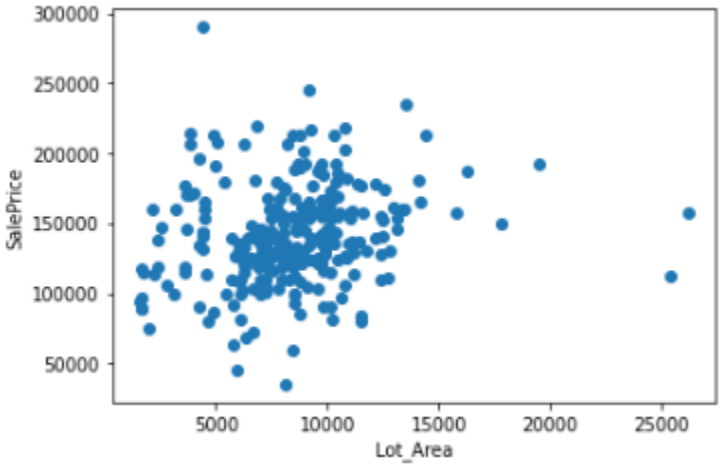
```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   PID              266 non-null    int64
 1   Lot_Area         266 non-null    int64
 2   House_Style      266 non-null    object
 3   Overall_Qual     266 non-null    int64
 4   Overall_Cond     266 non-null    int64
 5   Year_Built       266 non-null    int64
 6   Heating_QC       266 non-null    object
 7   Central_Air      266 non-null    object
 8   Gr_Liv_Area      266 non-null    int64
 9   Bedroom_AbvGr    266 non-null    int64
 10  Fireplaces       266 non-null    int64
 11  Garage_Area      266 non-null    int64
 12  Mo_Sold          266 non-null    int64
 13  Yr_Sold          266 non-null    int64
 14  SalePrice        266 non-null    int64
 15  Basement_Area    266 non-null    int64
 16  Full_Bathroom    266 non-null    int64
 17  Half_Bathroom    266 non-null    int64
 18  Total_Bathroom   266 non-null    float64
 19  Deck_Porch_Area  266 non-null    int64
 20  Age_Sold         266 non-null    int64
 21  Season_Sold      266 non-null    int64
 22  Garage_Type_2    266 non-null    object
 23  Foundation_2     266 non-null    object
 24  Masonry_Veneer   266 non-null    object
 25  Lot_Shape_2      266 non-null    object
 26  House_Style2     266 non-null    object
 27  Overall_Qual2    266 non-null    int64
 28  Overall_Cond2    266 non-null    int64
 29  Log_Price        266 non-null    float64
 30  Bonus            266 non-null    int64
dtypes: float64(2), int64(21), object(8)
```

3.4 Visualizing relationships between dependent and independent variables

Scatter plots were generated in order to visualize any relationships between the dependent variable and independent variables. A few examples are indicated below:

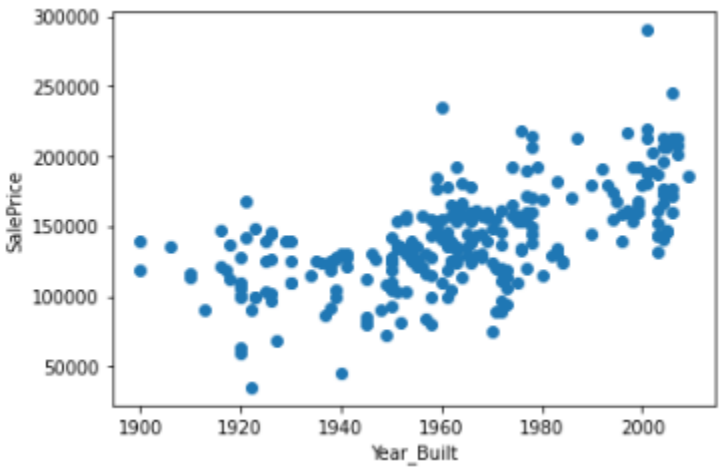
Graph 03 – Lot area (land area) vs. Sale price

The graph suggests that the price point of a house varies extensively regardless of the lot area. For example, if you consider a lot area of 5000, the price varies from around 75,000 to almost 300,000. Although 300,000 is possibly an outlier in the dataset, the next largest is around 220,000 which is still a considerable range in the same lot area. This is the case for many values of lot area. This characteristic suggests that there are definitely other factors that determine the price point and lot area does not indicate a clear visual correlation with the selling price. The right-most side of the graph too indicates two possible outliers where the lot area is extremely high but has average selling prices.



Graph 04 – Year of built vs. Sale price

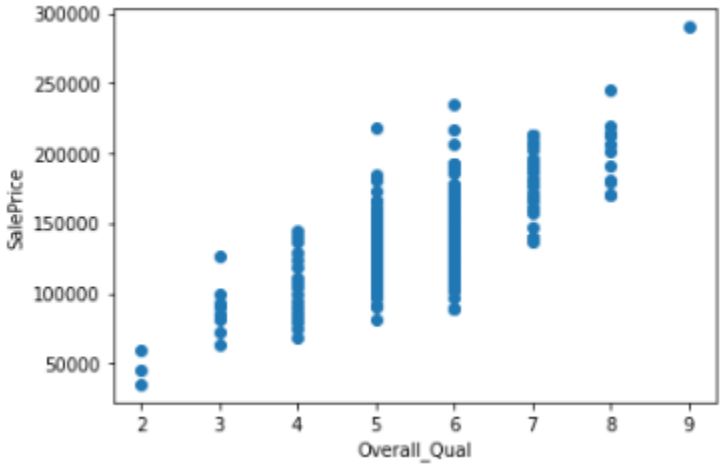
This graph illustrates a correlation between the year of built and the selling price. There is a general trend where the price increase as age of a house decreases. This analysis further conforms the common knowledge that new houses are generally more expensive than older houses.



Graph 05 – overall quality vs. Sale price

The graph shows that the overall quality of a house has been rated on a discrete scale possibly ranging from 0 to 10, even though there have been no ratings at 0, 1, or 10. However, if the average sale prices at

each rating are considered, there is a clear trend that indicates a strong positive correlation between rated quality and sale price.



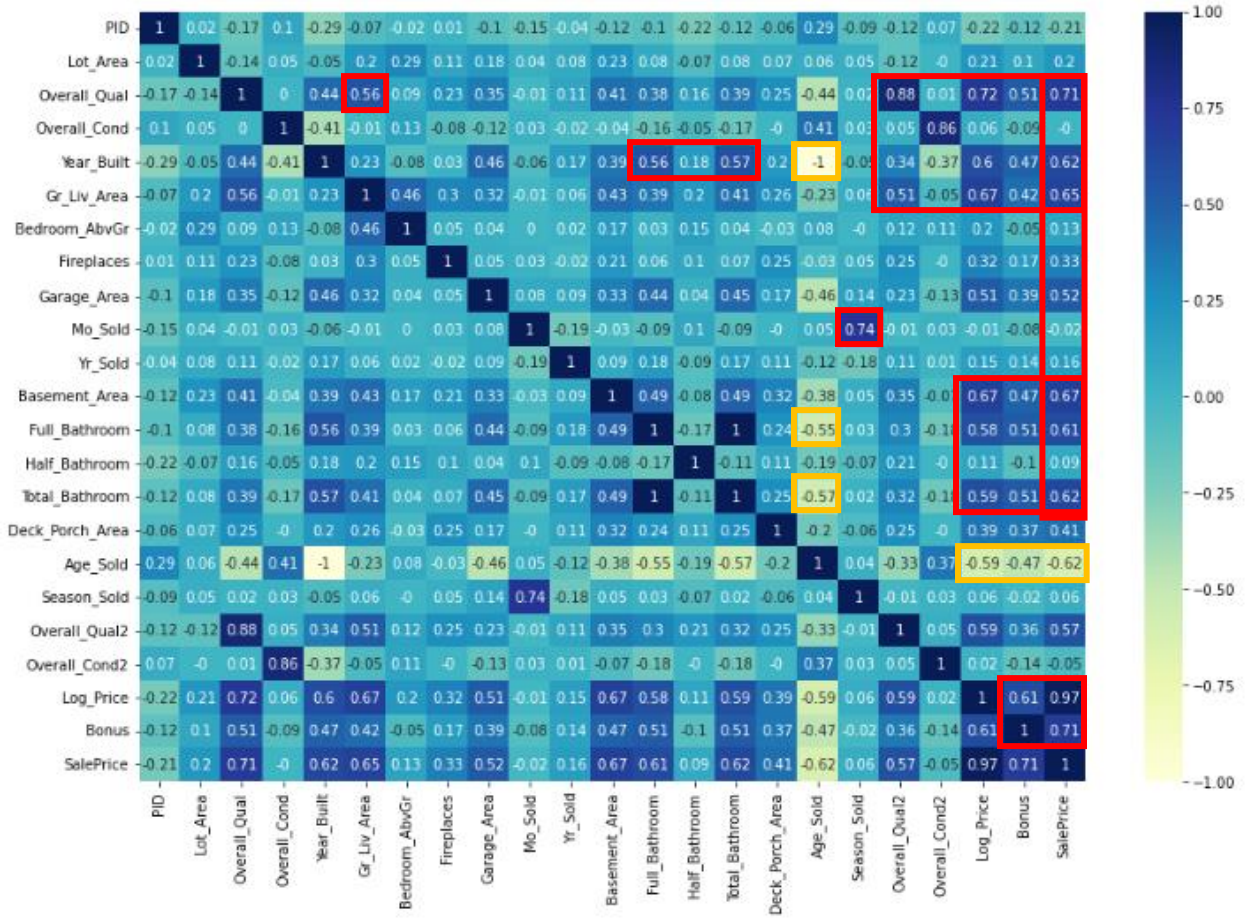
3.5 Converting categorical values into numerical values

As discussed previously as well, some of the variables were listed as categorical. Since it is difficult to carry out further analysis using categorical values, it was necessary to convert the categorical variables to numerical ones so that it enables detailed analysis. “Housing_Style”, “Heating_QC”, “Central_Air”, “Garage_Type_2”, “Foundation_2”, “Masonry_Veneer”, “Lot_Shape_2” and “House_Style2” were identified to be categorical. These unique values in the above categorical variables were then converted into numerical values as indicated below:

House_Style_vec	hous1	House_Stylevec	Heating_QCvec	Central_Airvec	Garage_Type_2vec	Foundation_2vec	Masonry_Veneervec	Lot_Shape_2vec	House_Style2vec
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1	0	0
1	1	1	0	0	0	0	0	1	1
0	0	0	1	0	0	1	0	0	0

3.6 Multicollinearity

A heat map of collinearity between all variables was constructed as shown below. This provided an overview of which variables were highly correlated and which showed low correlation.



The red highlighted areas in the heat-map indicate moderate to high positive correlation while those highlighted in amber indicate moderate to high negative correlation.

3.6.1 Key medium to high positive correlation pairs

- Overall_Qual vs. Overall_Qual2 (88%)
- Overall_Cond vs. Overall_Cond2 (86%)
- Overall_Qual vs. Log_Price (72%)
- Overall_Qual vs. Sale_Price (71%)
- Bonus vs. Sale_Price (71%)
- Log_Price vs. Sale_Price (97%)
- Total_Bathroom vs. Sale_Price (62%)
- Full_Bathroom vs. Sale_Price (61%)
- Basement_Area vs. Sale_Price (67%)
- Gr_Liv_Area vs. Sale_Price (65%)
- Year_Built vs. Sale_Price (62%)

- Mo_Sold vs. Season_Sold (74%)
- Garage_Area vs. Sale_Price (52%)

The key takeaway from the above analysis is that number of bathrooms, area of basement, year of built, size of living area and overall quality has a bigger positive impact on sale price compared to the other variables. It was interesting to note that the overall condition did not indicate any correlation to the sale price. The other positive correlation pairs were as expected.

3.6.2 Key medium to high negative correlation pairs

- Year_Built vs. Age_Sold (100%)
- Full_Bathroom vs. Age_Sold (55%)
- Total_Bathroom vs. Age_Sold (57%)
- Age_Sold vs. Sale_Price (62%)
- Age_Sold vs. Log_Price (59%)

The key takeaway here is that the age of the house has a strong negative correlation with the sale price. Year of build and age shows a 100% negative correlation as expected.

4. Prediction Model

4.1 Model 01 – Multilinear Regression

The multilinear regression model was selected as the first prediction model due to the vast number of independent variables that showed varying levels of correlation with the dependent variable.

In order to carry out prediction using linear regression, it was necessary to train the dataset. For this purpose, the dataset was randomly split into two sets; a training set and a test set with allocations of 80% and 20% respectively. A major proportion was set to the training set as it needed to undergo thorough learning of the data.

The model indicated an overall match of 97.6% with the training set and an overall match of 96.2% with the test set. These statistics are excellent indicators that the multilinear regression model that was used is a success and is able to correctly predict the sale price of houses.

```
Model train accuracy: 97.564%
Model test accuracy: 96.237%
```

4.2 Model 02 - Ridge Regression

Ridge regression was selected as the next predictive model. Ridge regression is used when we have multicollinearity in the model and in our dataset we have multicollinearity between some variables.

The model has an accuracy of 96.5% on the Train dataset and 94.4% on the test dataset.

```
Model train accuracy: 96.565%
Model test accuracy: 94.408%
```

5. Conclusion and Recommendation

In conclusion, it can be stated that not all of the variables have individual correlations with the sale price. For example, variables such as overall condition, lot area, and fireplaces do not need to be focused on in order to sell a house. Since the overall condition has no impact on the sale price, it is not even necessary for a seller to include that aspect or invest in either calculating or improving it. Comparatively, they can focus more on characteristics such as basement and garage area, the number of bathrooms, bedrooms, and living area to market houses.

It can be beneficial to invest in improving the outlook of the living area as it has a 56% positive correlation with the overall quality. Since overall quality has a 71% positive correlation with the sale price, it would be worthwhile investing in ways to further improve quality. From the analysis, it was observed that the age of the house has a major impact on the sale price. While the age itself cannot be changed, the importance of maintenance is highlighted here.

The highest frequency of sales has taken place in the months of April to July (summer) and the frequencies are comparatively low in the months of November to February (winter). However, the highest sale prices have been recorded in September. This could be an indication that people may pay an extra price to lock in a deal just before the start of winter.

I, Laxmikant Mukkawar, declare that the attached assignment is my own work in accordance with the Seneca Academic Policy. I have not copied any part of this assignment, manually or electronically, from any other source including websites, unless specified as references. I have not distributed my work to other students.