Housing Project

Submitted by:

Aditya Kissen Mitra

ACKNOWLEDGMENT

I have taken help from google.com and kaggle.com. Few mentioned sites-

https://www.kaggle.com/code/mgmarques/houses-prices-complete-solution/notebook

https://www.geeksforgeeks.org/

INTRODUCTION

- Business Problem Framing

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies.

- Conceptual Background of the Domain Problem

In this project, we will develop and evaluate the performance and the predictive power of a model trained and tested on data collected from a US-based housing company named Surprise Housing. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

- Review of Literature

  By doing the analysis I get to know few points-

  More than 3 parking cars and more than 900 of area are outliers, since a few numbers of their observations. Although there is a relationship between them, most likely with a smaller number of parking spaces, there may be more garage area for other purposes

  Prices are affected by the neighbourhood, yes, if more similar more they attract. But we will delve a little and see how the year and month of the sale also has great influence on the price variation and confirm the seasonality.

  The seasonality does have some effect

- Motivation for the Problem Undertaken

  The main objective is to predict the Selling Price of a house taking consideration of many factors.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

1. The data have 19 features with nulls, five of them area categorical and with more than 47% of missing ration.
2. Features high skewed right, heavy-tailed distribution, and with high correlation to Sales Price. It is important to treat them
3. Features skewed, heavy-tailed distribution, and with good correlation to Sales Price. It is important to treat them
4. Features high skewed, heavy-tailed distribution, and with low correlation to Sales Price. Maybe we can drop these features, or just use they with other to create a new more important feature.
5. Total Rooms above Ground and Living Area
6. We found that, the interaction between the two features- Total Rooms above Ground and Living Area did not present a better correlation than that already seen in the living area, include it improves to 0.74 with the cut of the outliers.
7. We can note that more than 3 parking cars and more than 900 of area are outliers, since a few number of their observations. Although there is a relationship between them, most likely with a smaller number of parking spaces, there may be more garage area for other purposes, reason why the correlation between them is 0.88 and not 1.
8. The area by car is little useful, but contrary to common sense the multiplication of the area by the number of vacancies yes is. In the division we lose the magnitude and we have to maintain one or another functionality to recover it. With the multiplication we solve the problem of 1 parking space of 10 square feet against another of 10 with 1 square feet each. We could still improve the correlation by 0.06, already considering the exclusion of only 4 outliers.

9. Full Bath has the largest correlation with Sale Price. The others individually, these feature is not very important.
10. Porch features have low correlation with price, and by the graphics we see all most has low base and high variance, being a high risk to end complex models and fall into over fit.
11. For Slope of property and Lot area It is interesting to note that the slope has a low correlation, but as an expected negative. On the other hand, the lot size does not present such a significant correlation, contrary to the interaction between these two characteristics, which is better and also allow us to identify some outliers. Let's take a look at the effect of removing the outliers.
12. The seasonality does have some effect, but of course we draw this conclusion based on the graphs is precipitated if not erroneous, given that even having restricted the views still exist houses with different characteristics in the same neighbourhood. However, this is sufficient to understand that the timing of the sale matters, so the model will probably have to take this into account, or this will be part of the residual errors.

- Data Pre-processing Done

  For cleaning of Data I used techniques like mean, median and mode methods based on continuous data and discrete data. I fill the nan data and drop few of them which are not necessary for model building like ID.

- Data Inputs- Logic- Output Relationships

  Most of the data was of Object type.

  Nulls: The data have 19 features with nulls, five of them are categorical and with more than 47% of missing ration. They are candidates to drop or use them to create another more interesting feature:

- PoolQC
- MiscFeature

- Alley
- Fence
- FireplaceQu

Features high skewed right, heavy-tailed distribution, and with high correlation to Sales Price. It is important to treat them:

- TotalBsmtSF
- 1stFlrSF
- GrLivArea

Features skewed, heavy-tailed distribution, and with good correlation to Sales Price. It is important to treat them (boxcox 1p transformation, Robustscaler, and drop some outliers):

- LotArea
- KitchenAbvGr
- ScreenPorch
- EnclosedPorch
- MasVnrArea
- OpenPorchSF
- LotFrontage
- BsmtFinSF1
- WoodDeckSF
- MSSubClass

Features high skewed, heavy-tailed distribution, and with low correlation to Sales Price. Maybe we can drop these features, or just use them with other to create a new more important feature:

- MiscVal
- TSsnPorch
- LowQualFinSF
- BsmtFinSF2
- BsmtHalfBa

Transform from Year Feature to Age, 2011 - Year feature, or YEAR(TODAY()) - Year Feature

- YearRemodAdd:
- YearBuilt
- GarageYrBlt
- YrSold

- Hardware and Software Requirements and Tools Used

Software used- Jupiter notebook
Language used- Python

Important libraries used-

Numpy- To perform mathematical functions.
Pandas- To do import the dataset, do analysis and manipulate data.
Seaborn- To visualize the data based on matplotlib
Scipy- To check the skewness, normalization of data
Statsmodels- for pipeline and VIF
Sklearn- To build model for machine learning

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

Nulls: The data have 19 features with nulls, five of them are categorical and with more than 47% of missing ration.

Most of the Data are of object data type

For categorical data I used mode function to fill out missing data and for continuous data I used mean to fill the missing data.

I dropped ID column as it is of no use in model building and PoolQC as it is of no use while predicting since prediction sheet containing the column with all null values.

- Testing of Identified Approaches (Algorithms)

  For training and testing I divide the train data in 30% test and

  70%train data.

  Models used-

  Linear Regression

  Ridge Regression

  Gradient boosting

  K-Nearest Neighbour

  Random Forest Regression

  Decision Tree

- Run and Evaluate selected models

  Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

  Linear Regression-

Linear Regression

```
model = LinearRegression()
model.fit(X_train, y_train)
print('Train Score : {:.4f}'.format(model.score(X_train, y_train)))
print('Test Score : {:.4f}'.format(model.score(X_test, y_test)))
```

```
Train Score : 0.8668
Test Score : 0.8322
```

Ridge-

Ridge

```
model = Ridge()
model.fit(X_train, y_train)
print('Train Score : {:.4f}'.format(model.score(X_train, y_train)))
print('Test Score : {:.4f}'.format(model.score(X_test, y_test)))
```

```
Train Score : 0.8668
Test Score : 0.8324
```

Gradient-

Gradient Boost

```
model = GradientBoostingRegressor()
model.fit(X_train, y_train)
print('Train Score : {:.4f}'.format(model.score(X_train, y_train)))
print('Test Score : {:.4f}'.format(model.score(X_test, y_test)))
```

```
Train Score : 0.9749
Test Score : 0.8917
```

K-Nearest Neighbour-

K-Nearest Neighbors

```
model = KNeighborsRegressor()
model.fit(X_train, y_train)
print('Train Score : {:.4f}'.format(model.score(X_train, y_train)))
print('Test Score : {:.4f}'.format(model.score(X_test, y_test)))
```

```
Train Score : 0.8384
Test Score : 0.7889
```

Decision Tree-

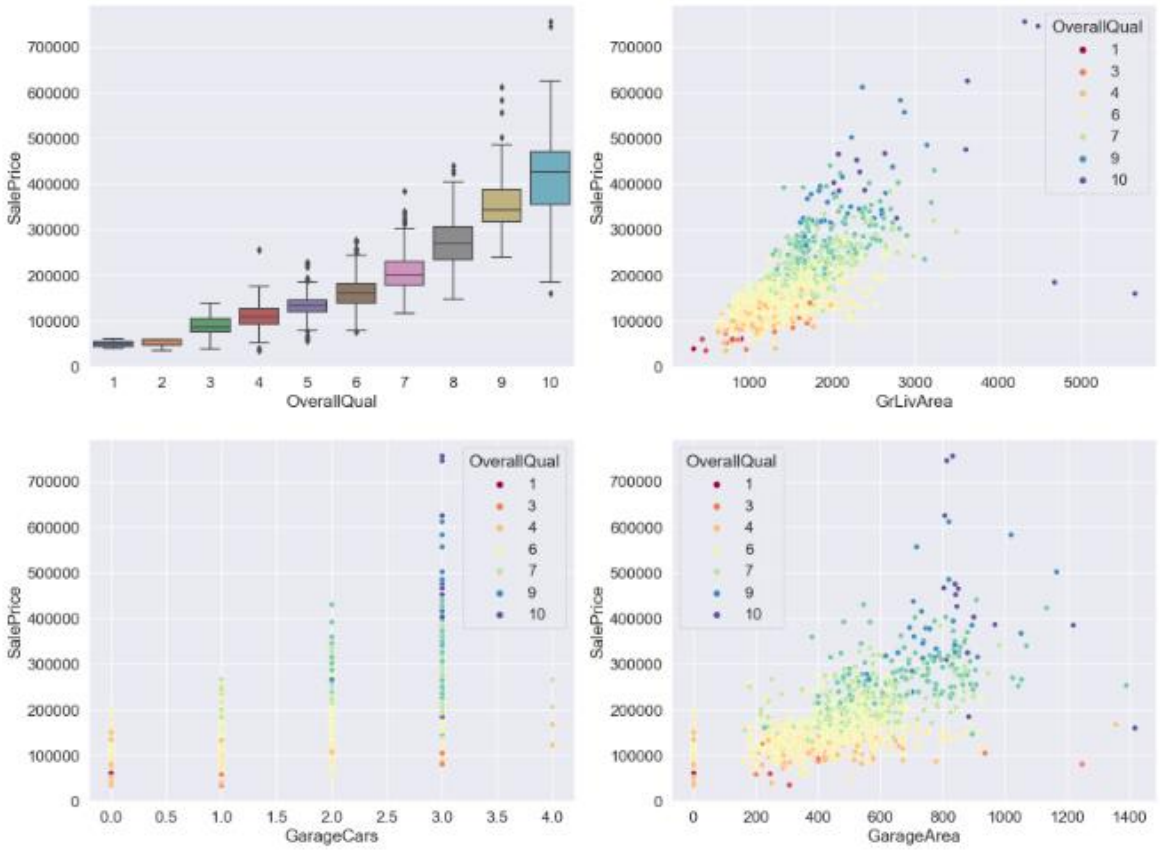
Decision Tree

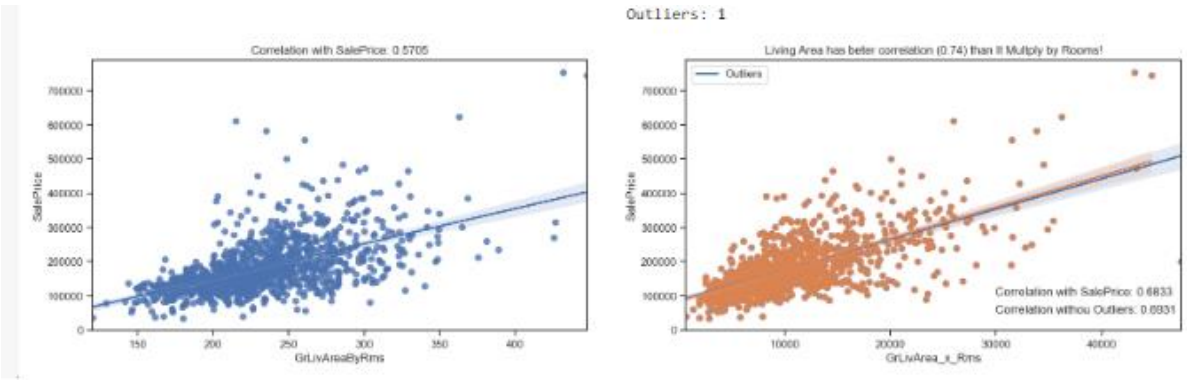
```
model = DecisionTreeRegressor()

model.fit(X_train, y_train)
print('Train Score : {:.4f}'.format(model.score(X_train, y_train)))
print('Test Score : {:.4f}'.format(model.score(X_test, y_test)))
```

```
Train Score : 1.0000
Test Score : 0.8040
```

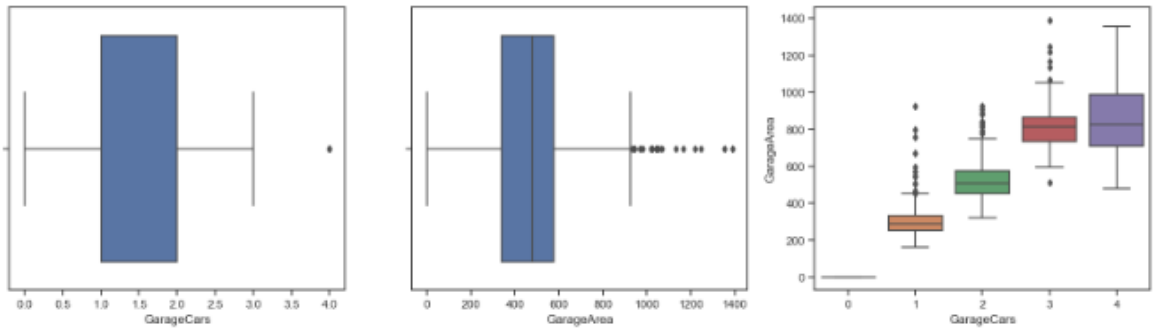
- Key Metrics for success in solving problem under consideration

  I used standard scalar for scaling the dataset before testing.
- Visualizations

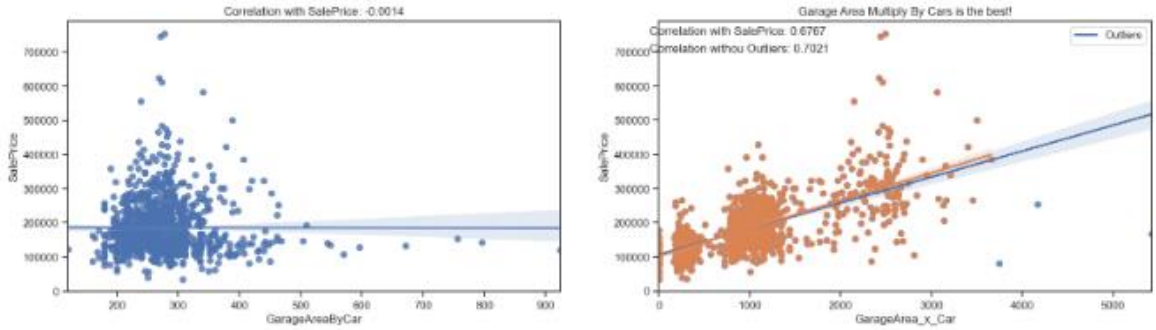




As we can see, the interaction between the two features did not present a better correlation than that already seen in the living area, include it improves to 0.74 with the cut of the outliers. On the other hand, the multiplication not only demonstrated the living area outliers already identified, but it still emphasized another. If the strategy is to drop the TotRmsAbvGrd, we should also exclude this additional outlier

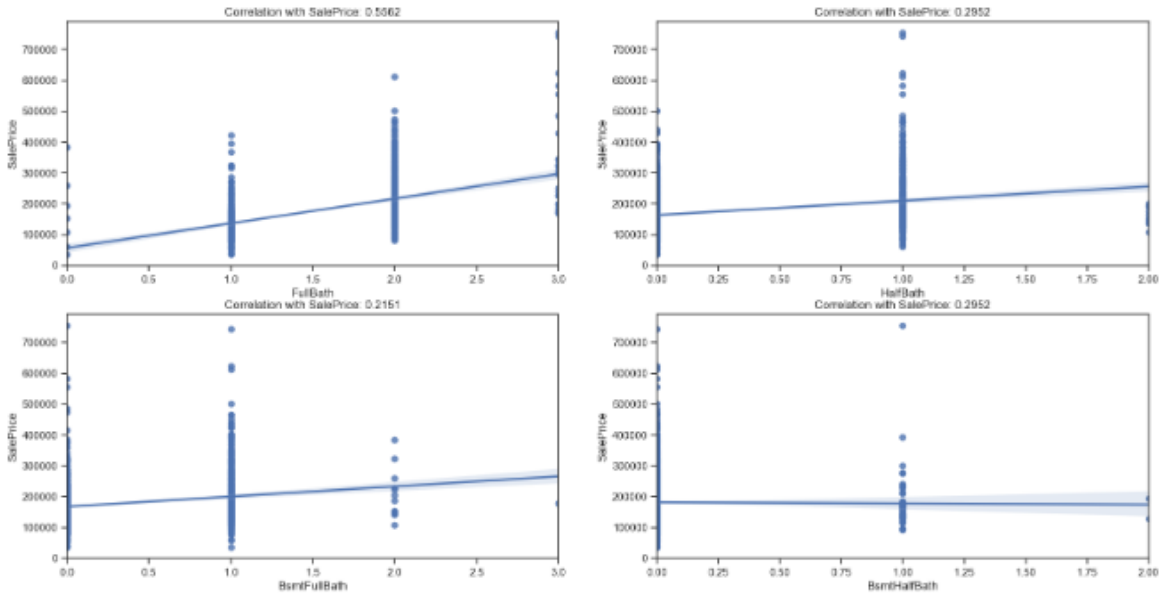


Observation- From the boxplot above, we can note that more than 3 parking cars and more than 900 of area are outliers, since a few number of their observations. Although there is a relationship between them, most likely with a smaller number of parking spaces, there may be more garage area for other purposes, reason why the correlation between them is 0.88 and not 1

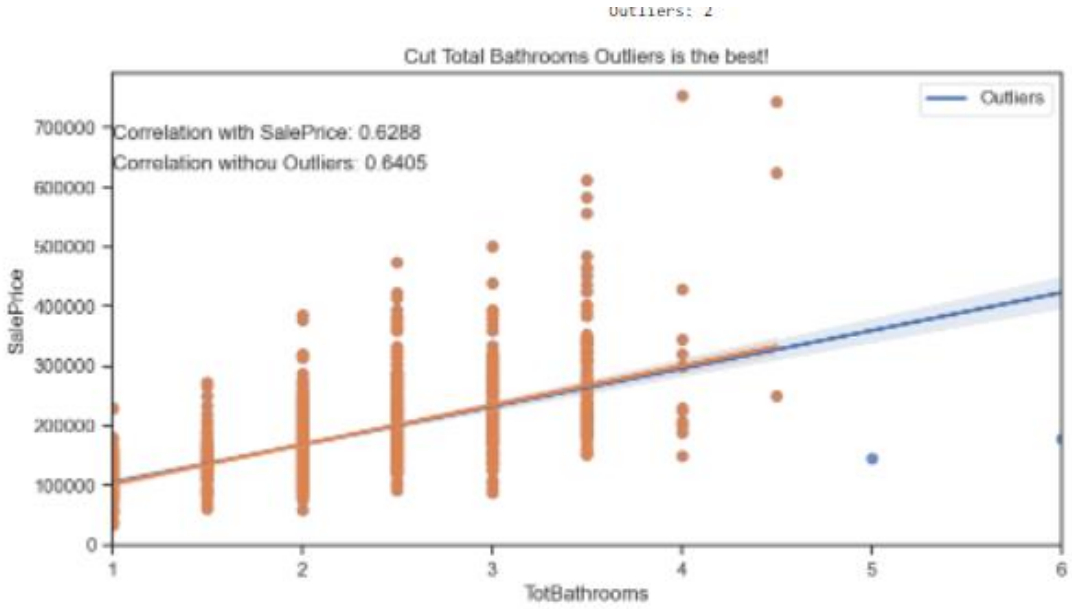


As can be seen the area by car is little useful, but contrary to common sense the multiplication of the area by the number of vacancies yes is. In the division we lose the magnitude and we have to maintain one or another functionality to recover it. With the multiplication we solve the problem of 1 parking space of 10 square feet against another of 10 with 1 square feet each. We could still improve the correlation by 0.06, already considering the exclusion of only 4 outliers.

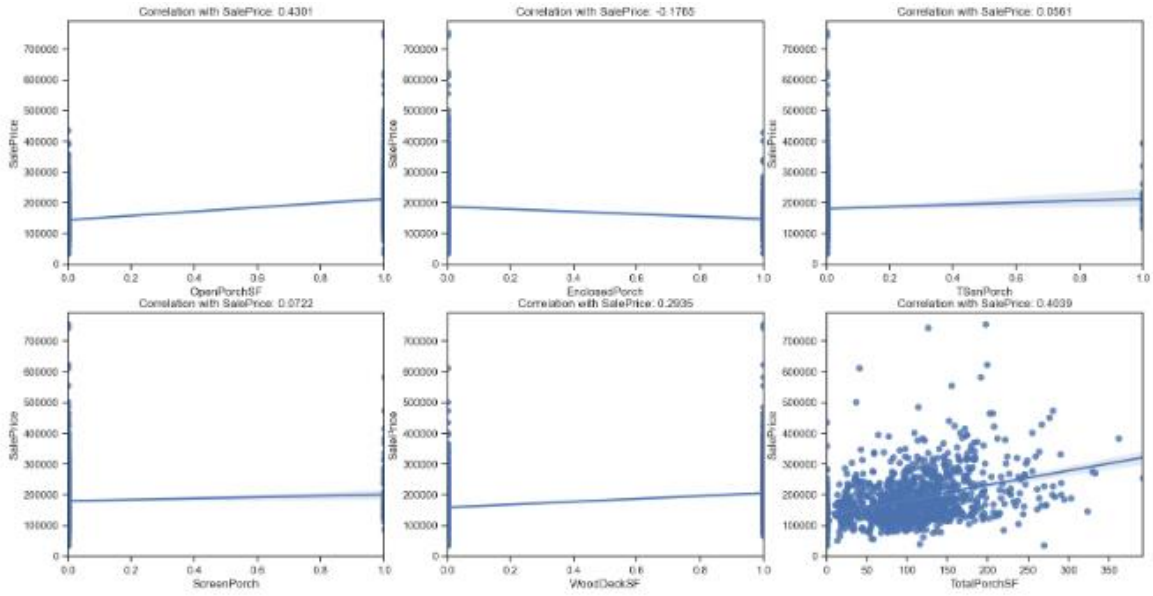
The identification of the outliers was facilitated, note that before we would have a greater number of outliers, since the respective of each features alone are not coincident.



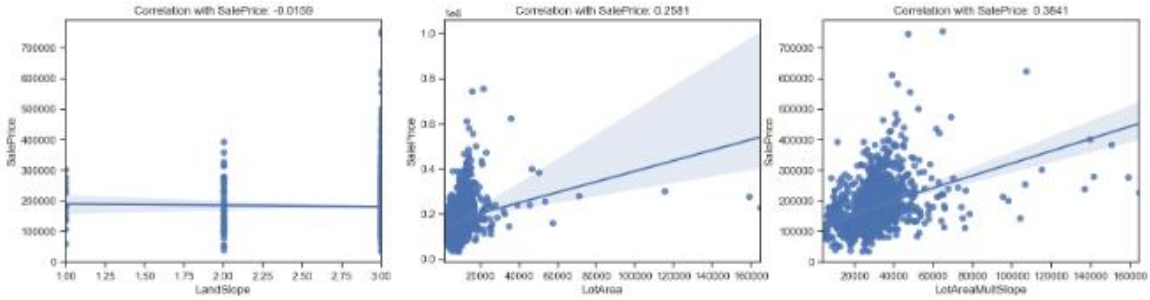
FullBath has the largest correlation with Sale Price between than. The others individually, these features are not very important.



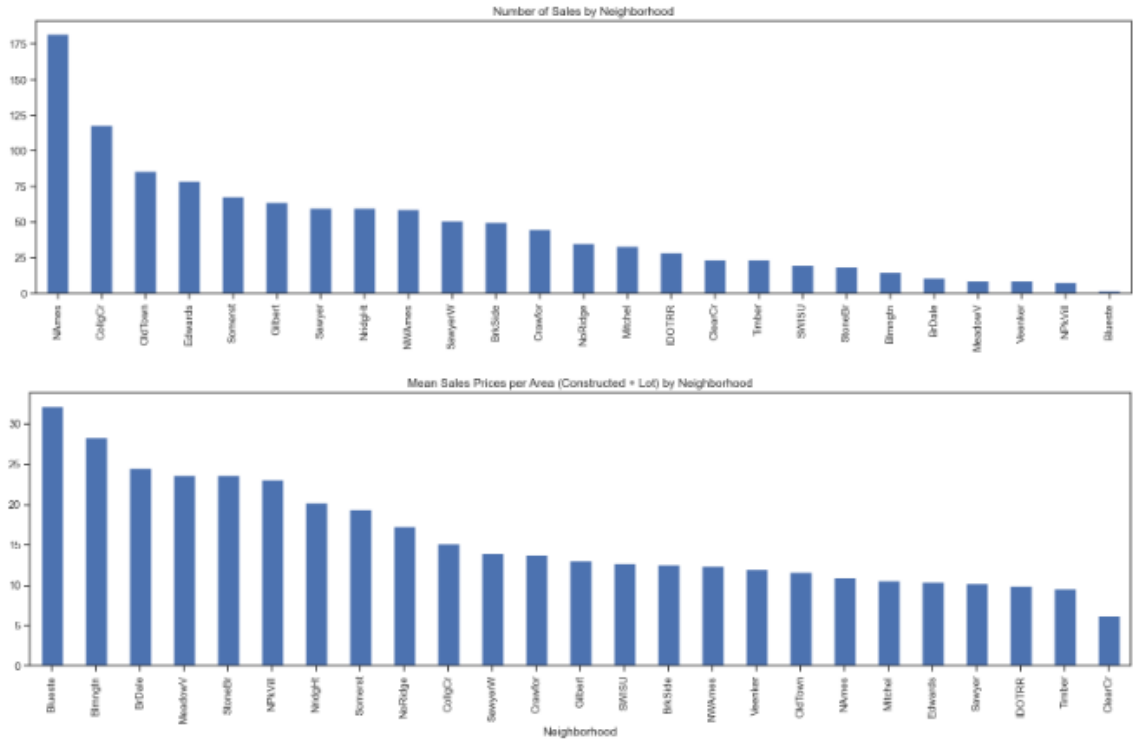
So, with our best predictor, we can cut only two outliers, use it and substitute all others bath features with a existence indicator.



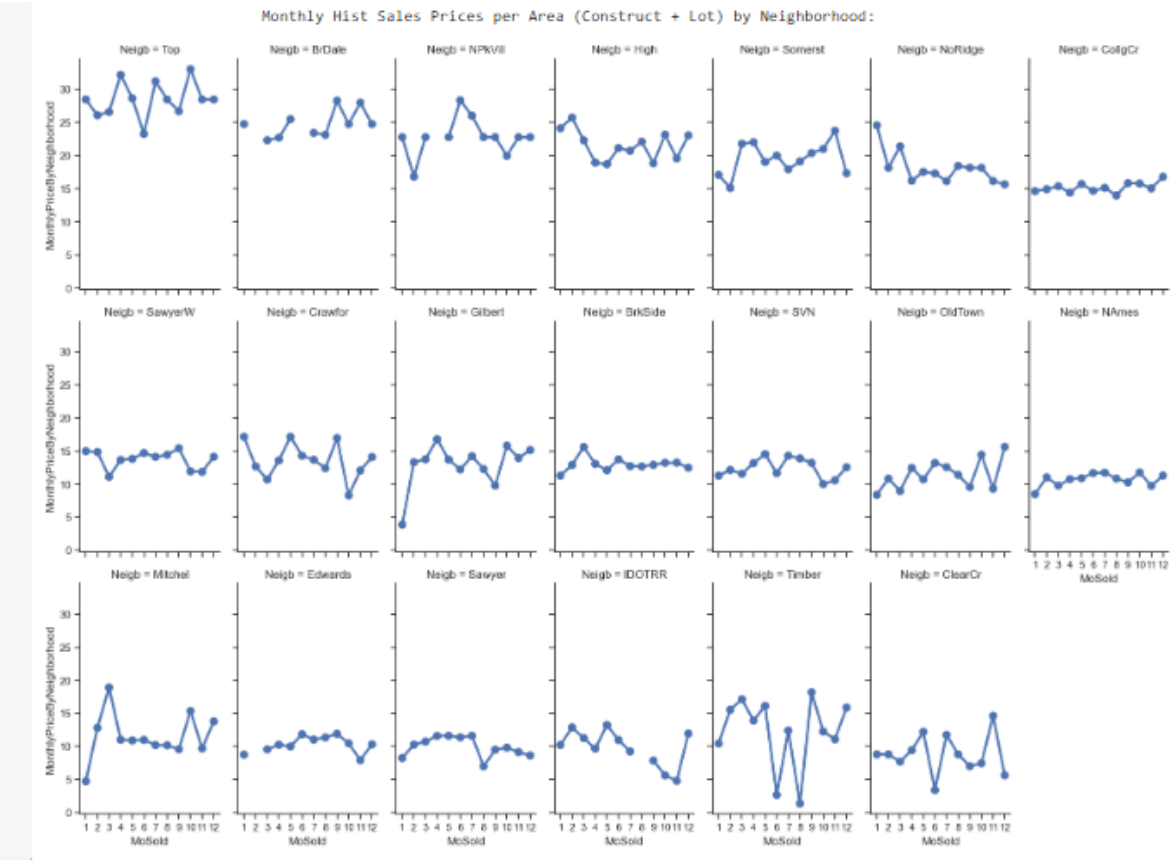
As we have seen, porch features have low correlation with price, and by the graphics we see all most has low bas and high variance, being a high risk to end complex models and fall into over fit.



It is interesting to note that the slope has a low correlation, but as an expected negative. On the other hand, the lot size does not present such a significant correlation, contrary to the interaction between these two characteristics, which is better and also allow us to identify some outliers. Let's take a look at the effect of removing the outliers



As we can see prices are affected by the neighbourhood, yes, if more similar more they attract. But we will delve a little and see how the year and month of the sale also has great influence on the price variation and confirm the seasonality.



As we expected, the seasonality does have some effect, but of course we draw this conclusion based only on the above graphs is precipitated if not erroneous, given that even having restricted the views still exist houses with different characteristics in the same neighbourhood.
However, this is sufficient to understand that the timing of the sale matters, so the model will probably have to take this into account, or this will be part of the residual errors.

- Interpretation of the Results

As described above after the visualization and pre-processing and model building I get the accuracy of 89% and I choose Gradient boosting model for prediction as it provided the highest accuracy.

CONCLUSION

- Key Findings and Conclusions of the Study

  The data have 19 features with nulls, five of them area categorical and with more than 47% of missing ration.

  Features high skewed right, heavy-tailed distribution, and with high correlation to Sales Price. It is important to treat them

  Features skewed, heavy-tailed distribution, and with good correlation to Sales Price. It is important to treat them

  Features high skewed, heavy-tailed distribution, and with low correlation to Sales Price. Maybe we can drop these features, or just use they with other to create a new more important feature.

  Total Rooms above Ground and Living Area

  We found that, the interaction between the two features- Total Rooms above Ground and Living Area did not present a better correlation than that already seen in the living area, include it improves to 0.74 with the cut of the outliers.

  We can note that more than 3 parking cars and more than 900 of area are outliers, since a few number of their observations. Although there is a relationship between them, most likely with a smaller number of parking spaces, there may be more garage area for other purposes, reason why the correlation between them is 0.88 and not 1.

  The area by car is little useful, but contrary to common sense the multiplication of the area by the number of vacancies yes is. In the division we lose the magnitude and we have to maintain one or another functionality to recover it. With the multiplication we solve the problem of 1 parking space of 10 square feet against another of 10 with 1 square feet each. We could still improve the correlation by 0.06, already considering the exclusion of only 4 outliers.

Full Bath has the largest correlation with Sale Price. The others individually, these feature is not very important.

Porch features have low correlation with price, and by the graphics we see all most has low base and high variance, being a high risk to end complex models and fall into over fit.

For Slope of property and Lot area It is interesting to note that the slope has a low correlation, but as an expected negative. On the other hand, the lot size does not present such a significant correlation, contrary to the interaction between these two characteristics, which is better and also allow us to identify some outliers. Let's take a look at the effect of removing the outliers.

The seasonality does have some effect, but of course we draw this conclusion based on the graphs is precipitated if not erroneous, given that even having restricted the views still exist houses with different characteristics in the same neighbourhood. However, this is sufficient to understand that the timing of the sale matters, so the model will probably have to take this into account, or this will be part of the residual errors.

- Limitations of this work and Scope for Future Work

  For further studies I can work on removing more outliers and try to increase the accuracy more by hyper tuning more.