Prediction model of

HOUSE PRICE IN AMES

By utilizing Python to analyze the provided dataset and constructing a predictive model for property prices, the goal is to gain insights into the business environment using data. Additionally, the aim is to address the following four questions as part of the analysis:

1. How does time affect price of house?
2. What is the distribution regarding type of deal?
3. When is the most profitable time to have a good deal?
4. Which are the top 05 factors has a significant impact on the price?

At the end, the analysis aims to provide valuable insights aligned with the problem statements. Its objective is to identify and address the key challenges and determine the critical factors that can drive profitability and unlock growth opportunities for the business.

Table of content

A. DATASET **4**

B. DATA CLEANING **4**

1. Null / Not a number (NaN) value cleaning: 4

2. Outliers 5

3. Dummy value: 5

C. PREDICTION MODEL **5**

1. Model Construction 5

2. Model Evaluation 6

D. EXPLORATORY DATA ANALYSIS **7**

About the dataset 7

1. How does time affect price of house? 7

2. What is the distribution regarding type of deal? 9

3. When is the most profitable time to have a good deal? 10

4. Which are the top 05 factors has a significant impact on the price? 10

E. KEY TAKEAWAYS **11**

F. REFERENCE **11**

G. APPENDIX **11**

Appendix 1: Number of null values in each feature (in order training and testing set) 11

Appendix 2: Outlier 12

Appendix 3: Categorical variables 14

Appendix 4: Residual summary statistic 14

Appendix 5: Normal Probability Plot of Residuals. 14

Appendix 6: Sale price changes by year. 15

Appendix 7: Sale price by number of year built and year remodelled 15

Appendix 8: Good deal and worst deal for house price 16

Appendix 9: Profiling house of good and worst deal. 16

Appendix 10: Good deal performance by each month of a year. 16

Appendix 11: Correlation between sale price and the house features 17

A. DATASET

The Master dataset is a combination of Train and Test dataset which constructed by 2920 observations and 159 feature [1]. In detail, Train and Test dataset both share 79 features except 'Sale Price' which is only available in Train dataset, namely:

Id, MSSubClass, MSZoning, LotFrontage, LotArea, Street, Alley, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, OverallQual, OverallCond, YearBuilt, YearRemodAdd, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, MasVnrArea, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinSF1, BsmtFinType2, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, Heating, HeatingQC, CentralAir, Electrical, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr', KitchenAbvGr, KitchenQual, TotRmsAbvGrd, Functional, Fireplaces, FireplaceQu, GarageType, GarageYrBlt, GarageFinish, GarageCars, GarageArea, GarageQual, GarageCond, PavedDrive, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, PoolQC, Fence, MiscFeature, MiscVal, MoSold, YrSold, SaleType, SaleCondition.

In general, the data is collected in Ames city of the US, with the sales record and feature of the properties during 2006 to 2010. Besides, there is a code book that give a detail for the features going along with the working files.

B. DATA CLEANING

Feature Engineering is performed fro training and testing dataset separately. After investigating the dataset, there are 03 tasks required to done include:

1. Null / Not a number (NaN) value cleaning:

Status: There are 18 features containing null values in which Alley, PoolQC, Fence,

MiscFeatures have above 50% of missing value. [Appendix 1]

Action:

- Drop columns with missing values above 50%
- For Categorical variables: replaced the missing values by MODE
- Numerical variables: replace the missing values by MEDIAN

Method: Via Python - Apply function:

- df.isnull().sum()

- df.drop ([‘col name’], inplace=True)
- df[ ].fillna(df[ ].mean/mode)

2. Outliers

Status: There are 16 continuous feature that are being affected by the outliers. Most of the feature have the positive skew. [Appendix 2]

Action: Normalise the data to have a standard Gaussian distribution as well as limit outlier.

Method: Via Python – apply df[ ] = df[ ].apply(lambda x: np.log(x)).

3. Dummy value:

Status: There are 39 categorical features. [Appendix 3]

Action: Transforming each unique value of a categorical variable into a new binary column.

Method: Via Python - apply pd.get_dummies(df[ ], drop_first=True, prefix=).

C. PREDICTION MODEL

1. Model Construction

Preparation: Assign dependent and independent variable for training and testing set in which the independent is always sale price.

Modelling: In order to find the most optimal model for this dataset, there are 04 pipeline created for comparison including Linear Regression, Decision Tree, Random Forest, K Nearest Neighbour. The selection is based on the cross-validation scores of each model.

Pipeline	Cross-validation score with 20 folds
Linear Regression	-0.148271
Decision Tree	-0.268281
Random Forrest	-0.179844
K Nearest Neigbor	-0.214996

Table 1: Cross-validation score 04 pipelines

Fitting: Basing on Table 1, the model Linear Regression is the most reliable model for this dataset. Therefore, carry out the prediction by this selected model with testing data.

2. Model Evaluation

R-squared and Root Mean Squared Error (RSME) score

Criteria	Score
R-squared	0.875715874090264
RSME	0.140774341951192

Table 2: Metric to evaluate Linear Regression model.

A value of 0.876 for the R-squared score indicates that approximately 87.67% of the variance in the sale price can be explained by the model's predictors. A higher R-squared score indicates a better fit of the model to the data.

The RMSE value of 0.141 implies that, on average, the model's predictions deviate from the actual sale prices by approximately 0.141 units.

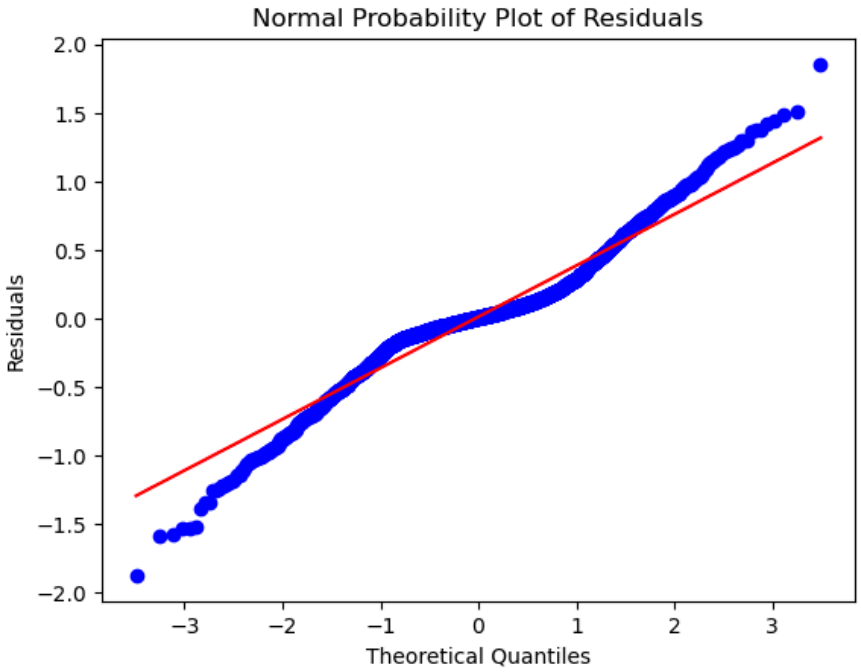
Based on these scores, the model seems to perform well. It has a relatively high R2 score, indicating a good amount of variance in sale prices being explained by the model's predictors. Additionally, the RMSE is relatively low, suggesting that the model's predictions are close to the actual sale prices in the training data.

Residual analysis

Summary statistic [Appendix 4]

- Mean residual is very close to zero (0.013195), suggesting that, on average, the model slightly overestimates the house prices.
- Standard deviation (0.384087) represents the average deviation of the residuals from the mean. It indicates that the residuals have a moderate amount of dispersion, with some predictions deviating from the mean by a significant amount.
- The minimum and maximum values of the residuals (-1.872258 and 1.853430, respectively) indicate the range of the deviations between the predicted and actual house prices.
- The quartiles (25%, 50%, and 75%) provide information about the spread of the residuals. The median (50%) residual is close to zero (0.007470), indicating that half of the predictions have positive residuals, while the other half has negative residuals. The interquartile range (75% - 25%) is 0.280130, indicating that the middle 50% of the residuals fall within this range.

Residual distribution [Appendix 5]



It can be seen as the dots on the normal probability plot of residuals closely adhere to a straight line which suggests that the residuals are approximately normally distributed. This indicates that the model's assumptions regarding the distribution of errors are met, and the model is capturing the underlying patterns in the data effectively.

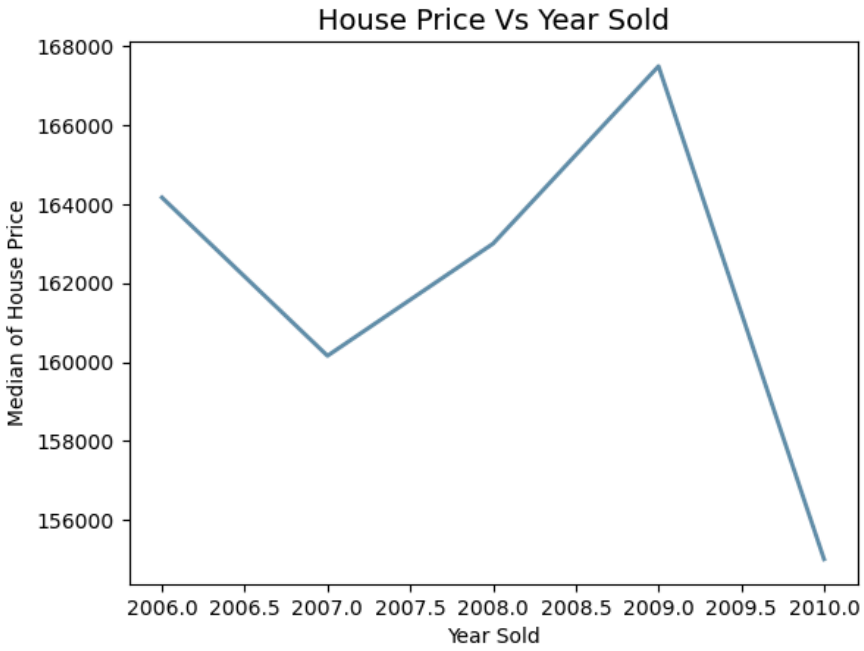
D. EXPLORATORY DATA ANALYSIS

About the dataset

This EDA is performed on the final dataset where the prediction is completed.

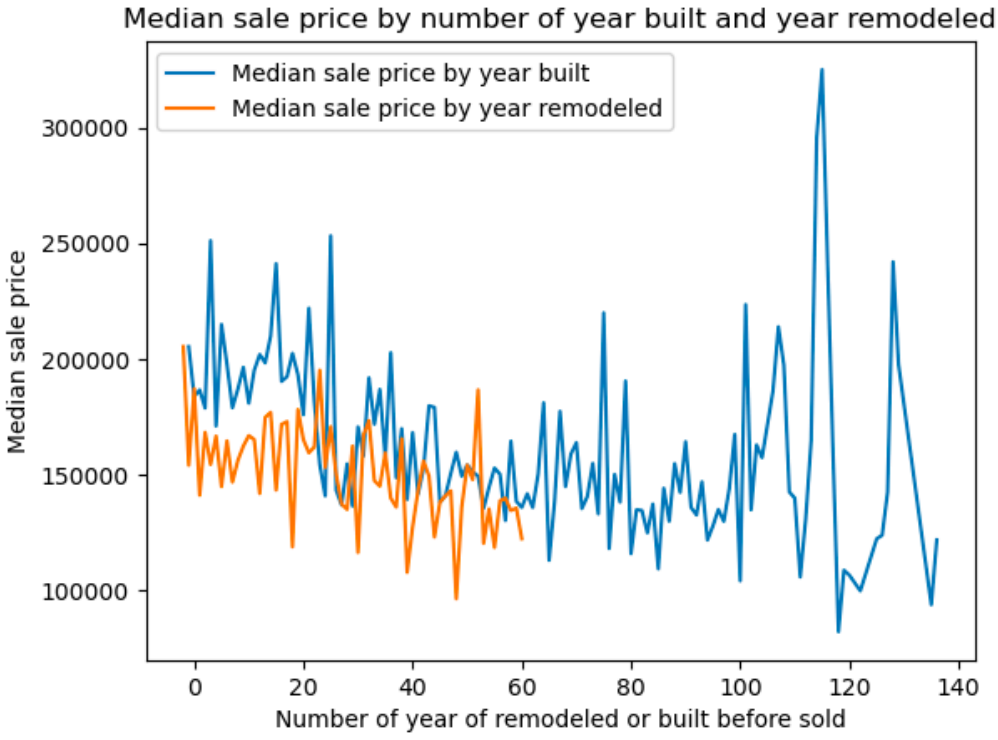
1. How does time affect price of house?

The chart is created by the price median of 2920 properties from 2006 and 2010 [Appendix 6]. The price from 2006 has a downward trend for one year before reaching to the peak, which is about $168,000 in 2009. Since then, it keeps dropping below $156,000 in 2010.



Appendix 6: Sale price changes by year.

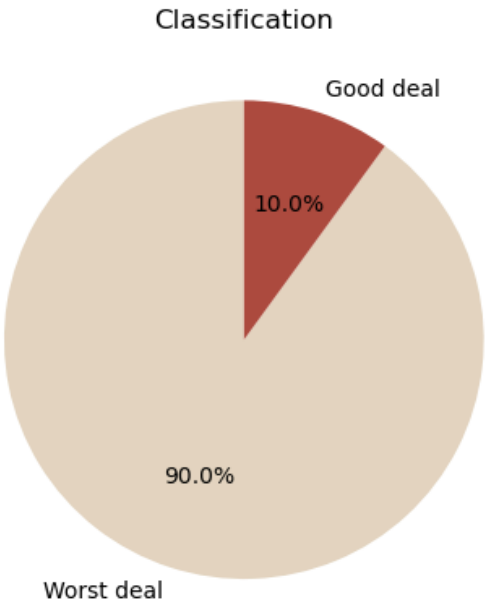
On the other hand, the price is also under the influence of the number of built or remodelled year. According to the chart, it can be understood as the newer the property is, the higher price it has [Appendix 7].



Appendix 7: Sale price by number of year built and year remodelled.

2. What is the distribution regarding type of deal?

The classification is to categorise the actual price whether it is a good or worst deal basing on the predicted value. If the gap between actual price and predicted one is not over 0.02, it is labeled “Good deal” or otherwise. In another word, the price is said to be undervalued or overpriced when the the gap between actual price and predicted one, as known as the absolute residual, is greater than 0.02 which is approximately a 10th percentile.



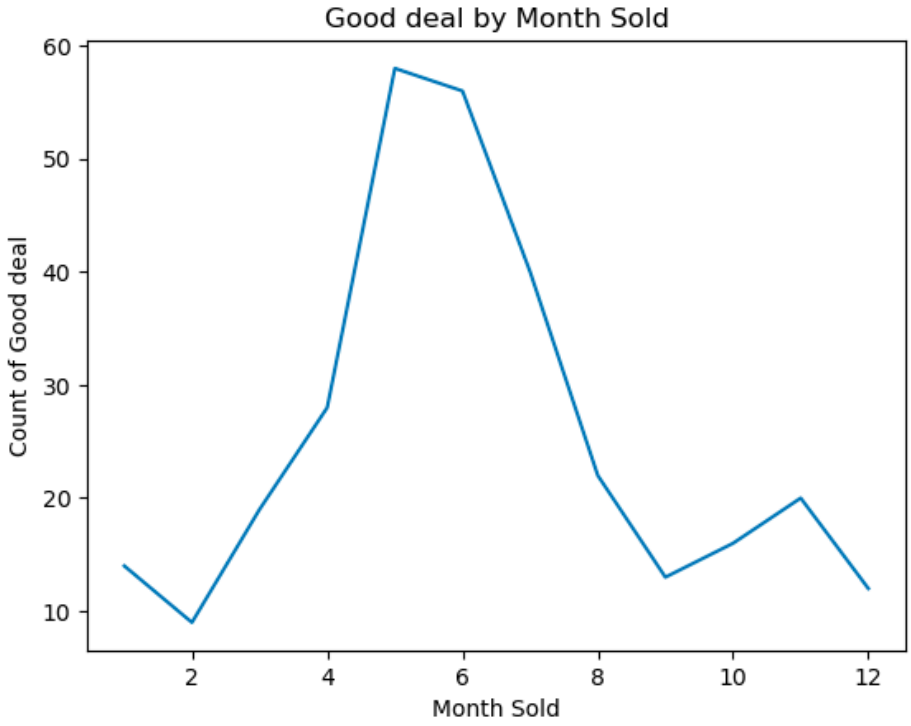
Appendix 8: Good deal and worst deal for house price.

The most favourable deals can be identified by certain characteristics, such as large houses with above average qualities, an average of 2 car parking spaces, and 2 bathrooms. These houses are priced, on average, 2% below the overall mean SalePrice of $180,070 and approximately 0.01% above the predictions made by the model [Appendix 9].

Conversely, the least desirable deals are associated with smaller houses of lower quality, averaging 2 car parking spaces and 2 bathrooms. These houses tend to be priced, on average, 0.2% higher than the overall mean SalePrice of $180,070 and roughly 2% above the model's predictions [Appendix 9].

3. When is the most profitable time to have a good deal?

Basing on the chart, the most ideal time to sale the house between may and July when the high season falls in June [Appendix 10].



Appendix 10: Good deal performance by each month of a year.

4. Which are the top 05 factors has a significant impact on the price?

Basing on the correlation between sale price and the sub-feature [Appendix 11] , it can be seen as the house is able to be highly priced if it can meet one of those feature:

- Locate in Stone Brook neighbour hood.
- The local resident should have low density.
- The property must a built-in garage.
- The height of basement should be below 70 inches.
- The exterior material does not need to be expensive.

E. KEY TAKEAWAYS

The analysis focuses on key aspects price and its correlation with time and the house profile including

Time: The price vacillates with time. Mostly the house has the high price as it has been built or done renovating. On the other hand, the price is on high peak in the middle of the year.

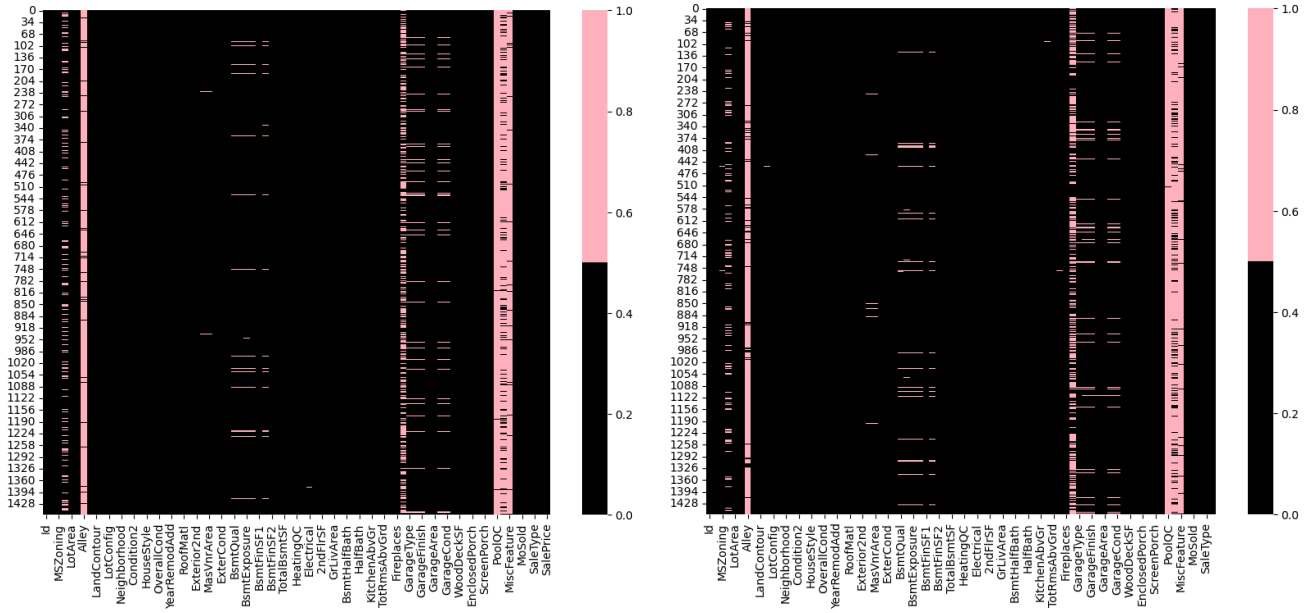
House profile: The best deals are characterised by larger, higher-quality houses with more amenities, priced below the average and model predictions. On the other hand, the worst deals involve smaller, lower-quality houses with fewer amenities, priced above the average and model predictions. Besides, the geographic advantage is one of the most important factor for pricing a property.

F. REFERENCE

[1] Montoya, A., “House Prices - Advanced Regression Techniques”, Kaggle, 2016, accessed on 10 May 2023, Available at: <https://kaggle.com/competitions/house-prices-advanced-regression-techniques>

G. APPENDIX

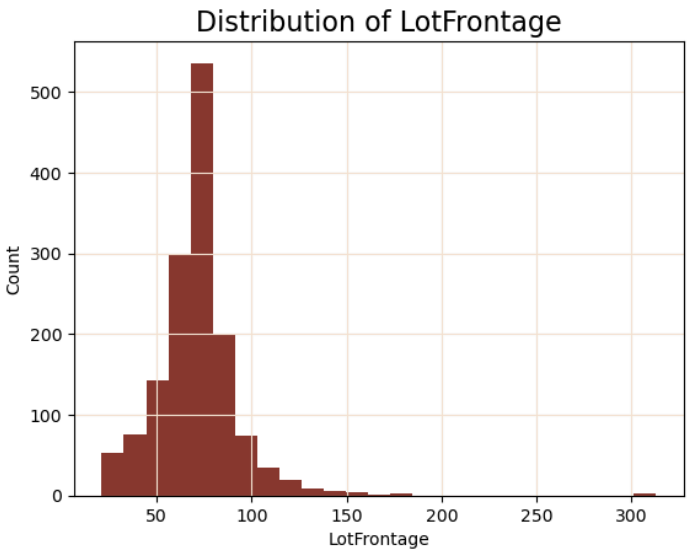
Appendix 1: Number of null values in each feature (in order training and testing set)

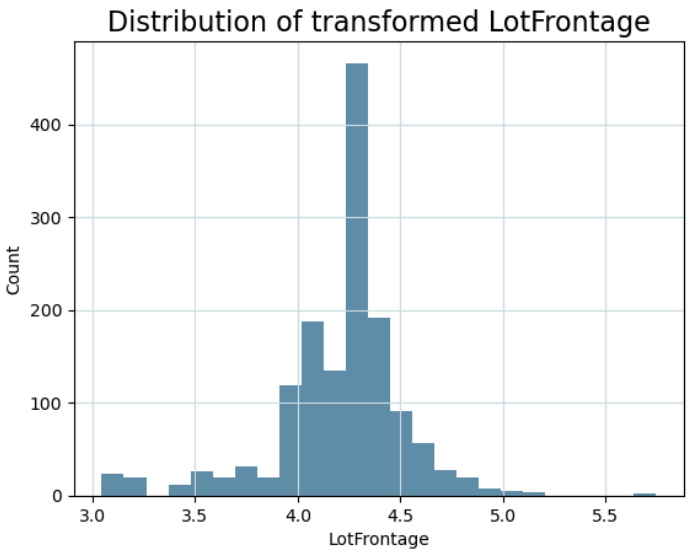


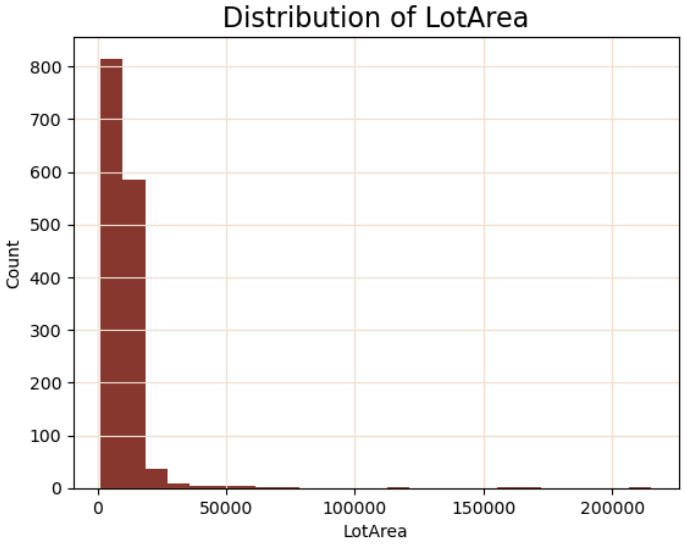
```
LotFrontage has 0.1774 % missing values.
Alley has 0.9377 % missing values.
MasVnrType has 0.0055 % missing values.
MasVnrArea has 0.0055 % missing values.
BsmtQual has 0.0253 % missing values.
BsmtCond has 0.0253 % missing values.
BsmtExposure has 0.026 % missing values.
BsmtFinType1 has 0.0253 % missing values.
BsmtFinType2 has 0.026 % missing values.
FireplaceQu has 0.4726 % missing values.
GarageType has 0.0555 % missing values.
GarageYrBlt has 0.0555 % missing values.
GarageFinish has 0.0555 % missing values.
GarageQual has 0.0555 % missing values.
GarageCond has 0.0555 % missing values.
PoolQC has 0.9952 % missing values.
Fence has 0.8075 % missing values.
MiscFeature has 0.963 % missing values.
```

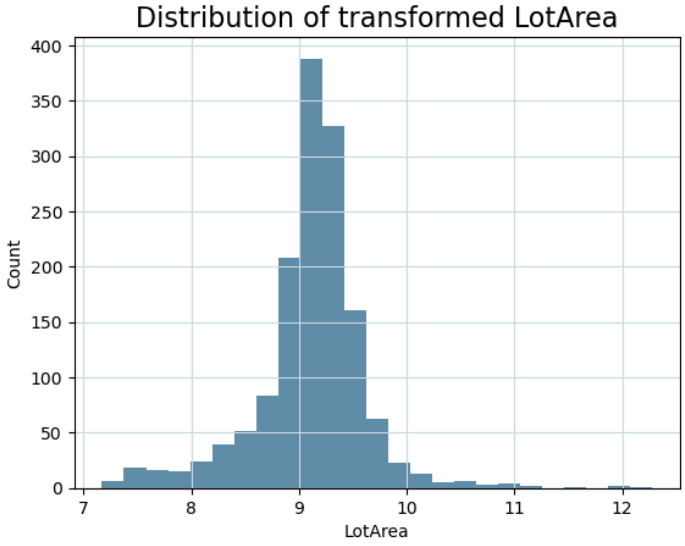
```
MSZoning has 0.0027 % missing values.
LotFrontage has 0.1555 % missing values.
Alley has 0.9267 % missing values.
Utilities has 0.0014 % missing values.
MasVnrType has 0.011 % missing values.
MasVnrArea has 0.0103 % missing values.
BsmtQual has 0.0301 % missing values.
BsmtCond has 0.0308 % missing values.
BsmtExposure has 0.0301 % missing values.
BsmtFinType1 has 0.0288 % missing values.
BsmtFinType2 has 0.0288 % missing values.
BsmtFullBath has 0.0014 % missing values.
BsmtHalfBath has 0.0014 % missing values.
Functional has 0.0014 % missing values.
FireplaceQu has 0.5 % missing values.
GarageType has 0.0521 % missing values.
GarageYrBlt has 0.0534 % missing values.
GarageFinish has 0.0534 % missing values.
GarageQual has 0.0534 % missing values.
GarageCond has 0.0534 % missing values.
PoolQC has 0.9979 % missing values.
Fence has 0.8014 % missing values.
MiscFeature has 0.9651 % missing values.
```

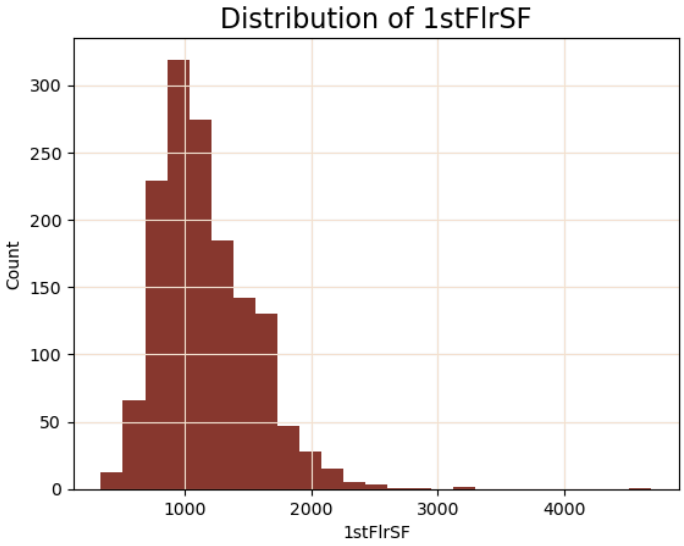
Appendix 2: Outlier

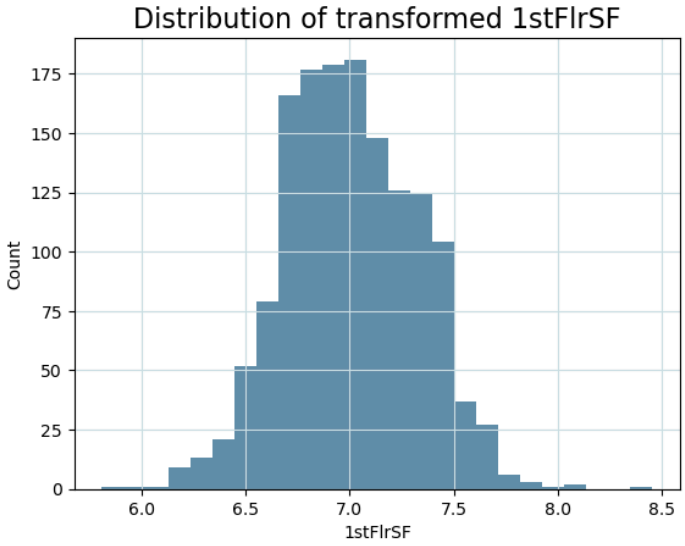


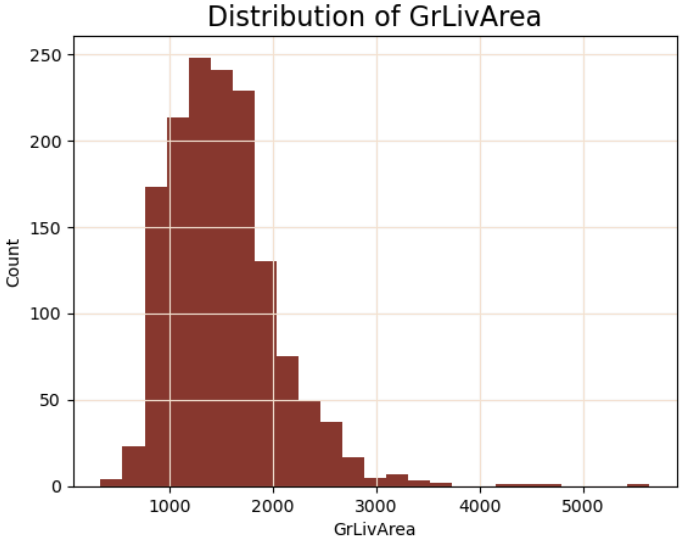


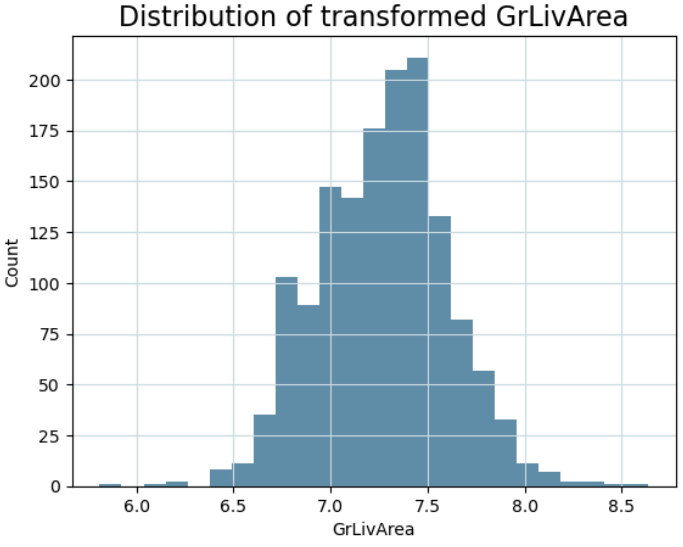

Distribution of LotArea
Count
LotArea
0
100
200
300
400
500
600
700
800
0
50000
100000
150000
200000


Distribution of transformed LotArea
Count
LotArea
0
50
100
150
200
250
300
350
400
7
8
9
10
11
12


Distribution of 1stFlrSF
Count
1stFlrSF
0
50
100
150
200
250
300
1000
2000
3000
4000


Distribution of transformed 1stFlrSF
Count
1stFlrSF
0
25
50
75
100
125
150
175
6.0
6.5
7.0
7.5
8.0
8.5


Distribution of GrLivArea
Count
GrLivArea
0
50
100
150
200
250
1000
2000
3000
4000
5000


Distribution of transformed GrLivArea
Count
GrLivArea
0
25
50
75
100
125
150
175
200
6.0
6.5
7.0
7.5
8.0
8.5

Appendix 3: Categorical variables

```
+ MSZoning has number of unique values are 5.
+ Street has number of unique values are 2.
+ LotShape has number of unique values are 4.
+ LandContour has number of unique values are 4.
+ Utilities has number of unique values are 2.
+ LotConfig has number of unique values are 5.
+ LandSlope has number of unique values are 3.
+ Neighborhood has number of unique values are 25.
+ Condition1 has number of unique values are 9.
+ Condition2 has number of unique values are 8.
+ BldgType has number of unique values are 5.
+ HouseStyle has number of unique values are 8.
+ RoofStyle has number of unique values are 6.
+ RoofMatl has number of unique values are 8.
+ Exterior1st has number of unique values are 15.
+ Exterior2nd has number of unique values are 16.
+ MasVnrType has number of unique values are 4.
+ ExterQual has number of unique values are 4.
+ ExterCond has number of unique values are 5.
+ Foundation has number of unique values are 6.
```

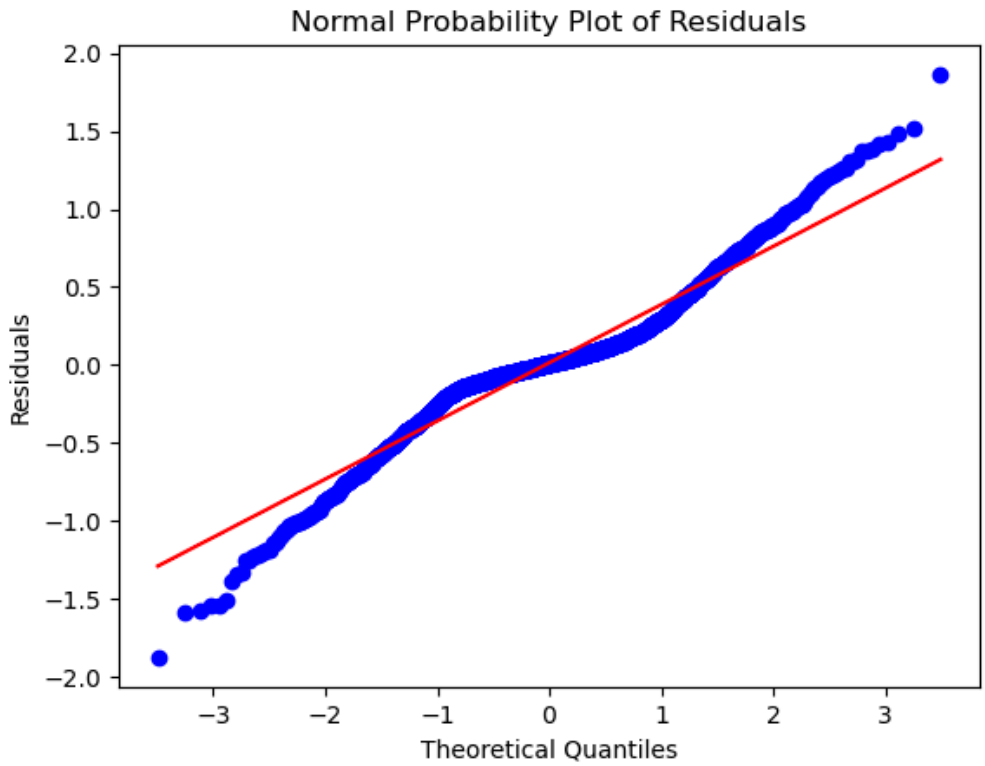
```
+ BsmtQual has number of unique values are 4.
+ BsmtCond has number of unique values are 4.
+ BsmtExposure has number of unique values are 4.
+ BsmtFinType1 has number of unique values are 6.
+ BsmtFinType2 has number of unique values are 6.
+ Heating has number of unique values are 6.
+ HeatingQC has number of unique values are 5.
+ CentralAir has number of unique values are 2.
+ Electrical has number of unique values are 5.
+ KitchenQual has number of unique values are 4.
+ Functional has number of unique values are 7.
+ FireplaceQu has number of unique values are 5.
+ GarageType has number of unique values are 6.
+ GarageFinish has number of unique values are 3.
+ GarageQual has number of unique values are 5.
+ GarageCond has number of unique values are 5.
+ PavedDrive has number of unique values are 3.
+ SaleType has number of unique values are 9.
+ SaleCondition has number of unique values are 6.
```

Appendix 4: Residual summary statistic

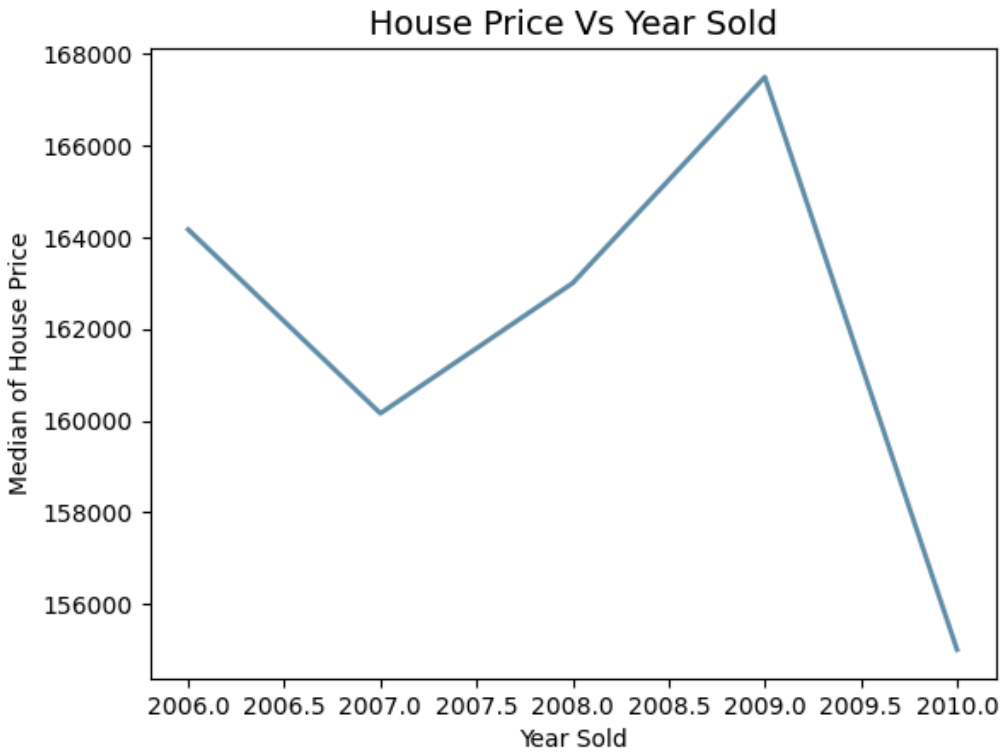
```
: df12['Residual'].describe()
```

```
: count    2920.000000
  mean        0.013175
  std         0.384338
  min        -1.878586
  25%        -0.122395
  50%         0.008032
  75%         0.153025
  max         1.864447
  Name: Residual, dtype: float64
```

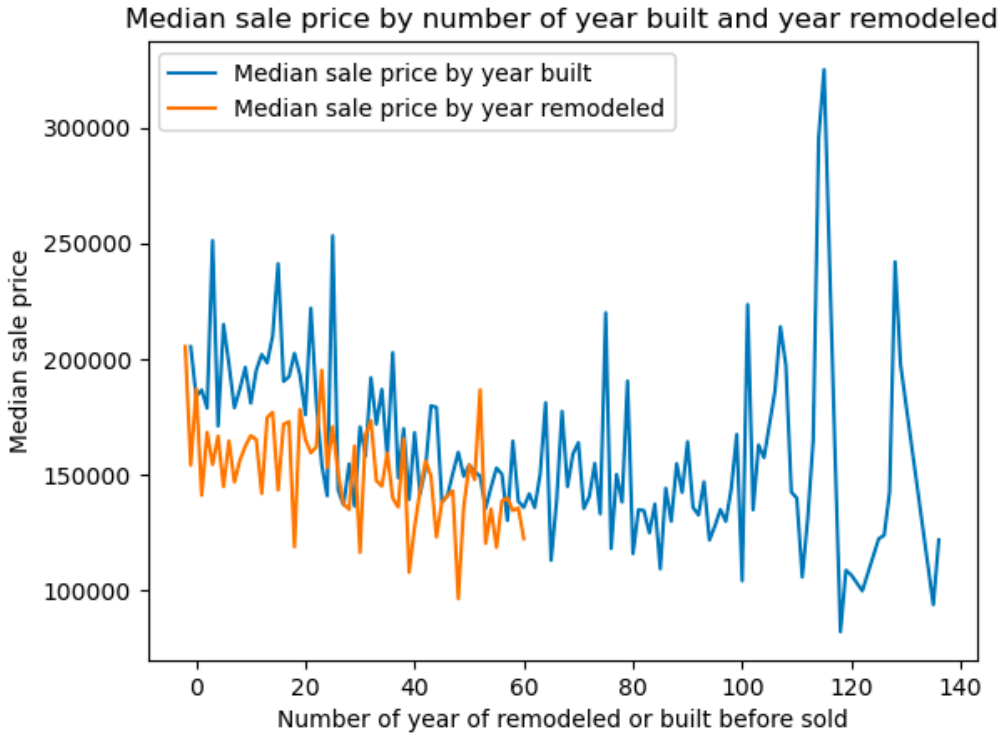
Appendix 5: Normal Probability Plot of Residuals.



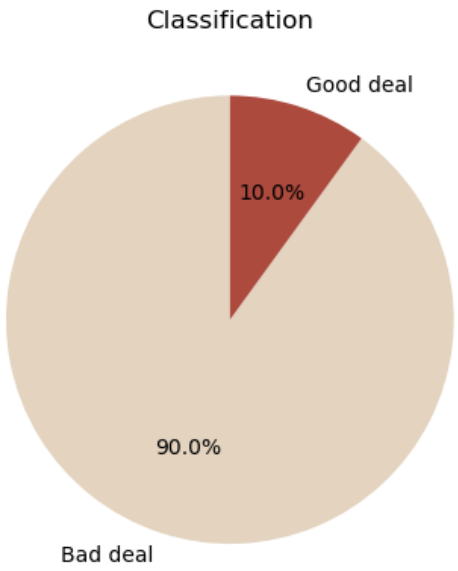
Appendix 6: Sale price changes by year.



Appendix 7: Sale price by number of year built and year remodelled



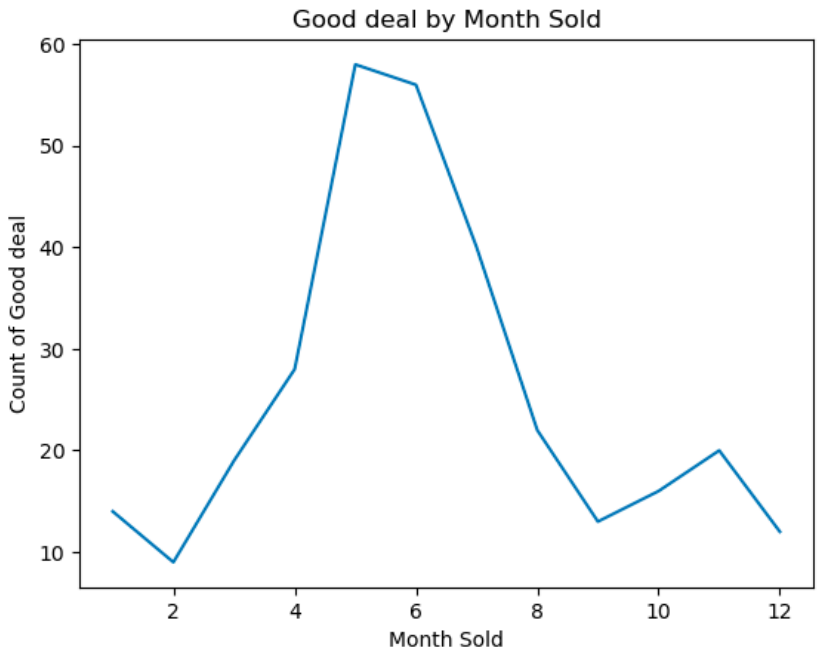
Appendix 8: Good deal and worst deal for house price



Appendix 9: Profiling house of good and worst deal.

Classification	OverallQual	GrLivArea	SalePrice	Predicted sale price	GarageCars	GarageArea	TotalBsmtSF	FullBath
Good deal	6.10	7.27	176008.66	175782.45	1.81	485.38	1037.16	1.56
Bad deal	6.09	7.26	180547.88	177075.55	1.76	471.47	1053.47	1.57

Appendix 10: Good deal performance by each month of a year.



Appendix 11: Correlation between sale price and the house features

```
SalePrice               1.000000
MSZoning_RL             0.065408
BsmtQual_TA             0.050788
GarageType_BuiltIn      0.050375
ExterCond_Po            0.049404
Neighborhood_StoneBr    0.048446
HeatingQC_Gd            0.046920
Exterior1st_HdBoard     0.046812
Neighborhood_NAmes      0.045910
FireplaceQu_TA          0.044938
Name: SalePrice, dtype: float64
```

END OF REPORT.