BSTAT Course Project AMES Housing

PROJECT REPORT

Submitted By

Sumit De(1001980218)

UNIVERSITY OF TEXAS ARLINGTON

Problem Definition:

Due to the enormous number of factors that affect pricing choices, determining the sale price of a house is frequently challenging. The majority of these parameters are frequently rather arbitrary and can range from well-known elements like the amount of bedrooms to less evident ones like basement height.

Realtors and potential home buyers who want to maximize the worth of the house they're attempting to buy or sell and prioritize the most significant aspects impacting sale price have a dilemma because of this. This is an excellent challenge for machine learning to address because it would take a long time to manually compare these elements for every house.

Research Questions:

Ames Housing dataset, which includes 82 features describing a wide range of characteristics of 2930 observations in Ames, Iowa sold between 2006 and 2010. To better understand the influence of unit sales prices in Ames Housing, we will use the dataset that was provided to answer the following questions.

- **Business Problem:**

  More than 2900 properties have information in the data set. There are more than 75 descriptive variables listed in the data dictionary. Some are non-numerical and lack a definite order since they are nominal (categorical) (Examples: Neighborhood, Type of roofing). Some are ordinal, or categorical but clearly arranged (for instance, heating quality, which can be Excellent, Good, Average, or Poor). Some are discrete, that is, numerical but spaced at regular intervals (Year Built, Number of Fireplaces). The remaining variables are continuous, which means they are numerical and can, in theory, take any value within a range (1st Floor Square Feet).

- **Constraints:**

1. Time constraints, Ames dataset needs in depth analysis to predict Sales Prices of a house by running various models and see which one is efficient.
2. The project needs to be delivered in 3 months.
3. Change in project scope affects quality of analysis.

- **Assumptions:**
  1. Demographic Economic Condition
  2. Ideal season to sell a house
  3. Impact of a pandemic on housing costs
  4. Very few of the independent variables are directly related to one another linearly.
  5. The independent variables have a conditional mean of zero, which indicates the error.
  6. For the purpose of linear regression we assume that the independent variables do not have high multiple collinearity to affect the regression analysis.

- **Limitations:**

1. There was no sufficient data from four neighborhoods.
2. Economic conditions like recession, inflation in certain years are not taken into account, in real life these factors play a crucial role in determining sales price. Inflation can vary by state. It is very difficult to predict the future price of a house. Generally, inflation has been measured over the long run, but its impact on prices is rarely consistent. For example the price of a house might go up or down depending on how much demand there is at that moment. In a recession or in hyperinflation people will dramatically change their spending habits and make more for less purchases.

- **Conditions:**

  The term demographic economic condition is used to describe the social and environmental conditions that are connected to the sales price. In general, these conditions can be influenced by a number of factors including garage area, overall quality, among other issues.

  We have seen this previously with studies that show higher socio-economic status has a positive impact on sales price but with recent economic crises it seems that housing price is also becoming a condition which has an adverse effect on economic.

SMART Objectives:

Specific:

We are expected to give the best predicted sale price for our clients by using data analysis. It is determined by using various data analytics techniques such as logistic regression, linear regression, and principal component analysis (PCA) to predict sales price of 2930 properties and 82 columns of data.

Measurable:

We use python programming and statistical techniques for our analysis.

Achievable:

Due to inflation and recession during the analysis period it was hard to predict the sales price though we had enough resources to process the dataset with some missing values.

Realistic:

We will use regression analysis to predict the sales price and we use principal component analysis (PCA)

Time-bound:

We do need to provide the clients with the best time in the year to make purchases which should benefit the client.

Part-1 : Data Preparation, Exploration and Understanding

Understanding our Data:

With regard to the Ames data, there are 82 variables and 2930 observations. The dataset displays the sales prices of homes that were constructed between 1872 and 2010 and were subsequently sold between 2006 and 2010. In the city of Ames, Iowa, the statistics were gathered from several neighborhoods.

	count	mean	std	min	25%	50%	75%	max
Order	2930.0	1.465500e+03	8.459625e+02	1.0	7.332500e+02	1.465500e+03	2.197750e+03	2.930000e+03
PID	2930.0	7.144645e+08	1.887308e+08	526301100.0	5.284770e+08	5.354536e+08	9.071811e+08	1.007100e+09
MS_SubClass	2930.0	5.738737e+01	4.263802e+01	20.0	2.000000e+01	5.000000e+01	7.000000e+01	1.900000e+02
Lot_Frontage	2930.0	6.922459e+01	2.132152e+01	21.0	6.000000e+01	6.922459e+01	7.800000e+01	3.130000e+02
Lot_Area	2930.0	1.014792e+04	7.880018e+03	1300.0	7.440250e+03	9.436500e+03	1.155525e+04	2.152450e+05
Overall_Qual	2930.0	6.094881e+00	1.411026e+00	1.0	5.000000e+00	6.000000e+00	7.000000e+00	1.000000e+01
Overall_Cond	2930.0	5.563140e+00	1.111537e+00	1.0	5.000000e+00	5.000000e+00	6.000000e+00	9.000000e+00
Year_Built	2930.0	1.971356e+03	3.024536e+01	1872.0	1.954000e+03	1.973000e+03	2.001000e+03	2.010000e+03
Year_Remod_Add	2930.0	1.984267e+03	2.086029e+01	1950.0	1.965000e+03	1.993000e+03	2.004000e+03	2.010000e+03
Mas_Vnr_Area	2907.0	1.018968e+02	1.791126e+02	0.0	0.000000e+00	0.000000e+00	1.640000e+02	1.600000e+03
BsmtFin_SF_1	2929.0	4.426296e+02	4.555908e+02	0.0	0.000000e+00	3.700000e+02	7.340000e+02	5.644000e+03
BsmtFin_SF_2	2929.0	4.972243e+01	1.691685e+02	0.0	0.000000e+00	0.000000e+00	0.000000e+00	1.526000e+03
Bsmt_Unf_SF	2929.0	5.592625e+02	4.394942e+02	0.0	2.190000e+02	4.660000e+02	8.020000e+02	2.336000e+03
Total_Bsmt_SF	2929.0	1.051615e+03	4.406151e+02	0.0	7.930000e+02	9.900000e+02	1.302000e+03	6.110000e+03
_1st_Flr_SF	2930.0	1.159558e+03	3.918909e+02	334.0	8.762500e+02	1.084000e+03	1.384000e+03	5.095000e+03
_2nd_Flr_SF	2930.0	3.354560e+02	4.283957e+02	0.0	0.000000e+00	0.000000e+00	7.037500e+02	2.065000e+03
Low_Qual_Fin_SF	2930.0	4.676792e+00	4.631051e+01	0.0	0.000000e+00	0.000000e+00	0.000000e+00	1.064000e+03
Gr_Liv_Area	2930.0	1.499690e+03	5.055089e+02	334.0	1.126000e+03	1.442000e+03	1.742750e+03	5.642000e+03
Bsmt_Full_Bath	2928.0	4.313525e-01	5.248202e-01	0.0	0.000000e+00	0.000000e+00	1.000000e+00	3.000000e+00
Bsmt_Half_Bath	2928.0	6.113388e-02	2.452536e-01	0.0	0.000000e+00	0.000000e+00	0.000000e+00	2.000000e+00
Full_Bath	2930.0	1.566553e+00	5.529406e-01	0.0	1.000000e+00	2.000000e+00	2.000000e+00	4.000000e+00

Dealing with Missing Values

We started by looking through the data to see if there were any missing values. The proportion and count of missing data for each column are shown in the table below. We made the decision to remove the columns from this project that had more than 80% of their data missing. We used the mean of that column to fill in the remaining missing values.

	Total	Percentage
Pool_QC	2917	99.556314
Misc_Feature	2824	96.382253
Alley	2732	93.242321
Fence	2358	80.477816
Fireplace_Qu	1422	48.532423
Lot_Frontage	490	16.723549
Garage_Cond	159	5.426621
Garage_Finish	159	5.426621
Garage_Yr_Blt	159	5.426621
Garage_Qual	159	5.426621
Garage_Type	157	5.358362
Bsmt_Exposure	83	2.832765
BsmtFin_Type_2	81	2.764505
Bsmt_Qual	80	2.730375
Bsmt_Cond	80	2.730375
BsmtFin_Type_1	80	2.730375
Mas_Vnr_Area	23	0.784983
Mas_Vnr_Type	23	0.784983
Bsmt_Full_Bath	2	0.068259
Bsmt_Half_Bath	2	0.068259
BsmtFin_SF_1	1	0.034130
Garage_Cars	1	0.034130
Electrical	1	0.034130
Total_Bsmt_SF	1	0.034130
Bsmt_Unf_SF	1	0.034130
BsmtFin_SF_2	1	0.034130
Garage_Area	1	0.034130

There are 34 features that have missing values. we will divide them into three groups based on the data description:

Group 1 - Categorical variables where NA means no feature:

PoolQC, MiscFeature, Alley, Fence, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, MasVnrType (15 variables).
For this group we will impute NA with 'Missing'.

Group 2 - Numerical variables where NA means no feature:

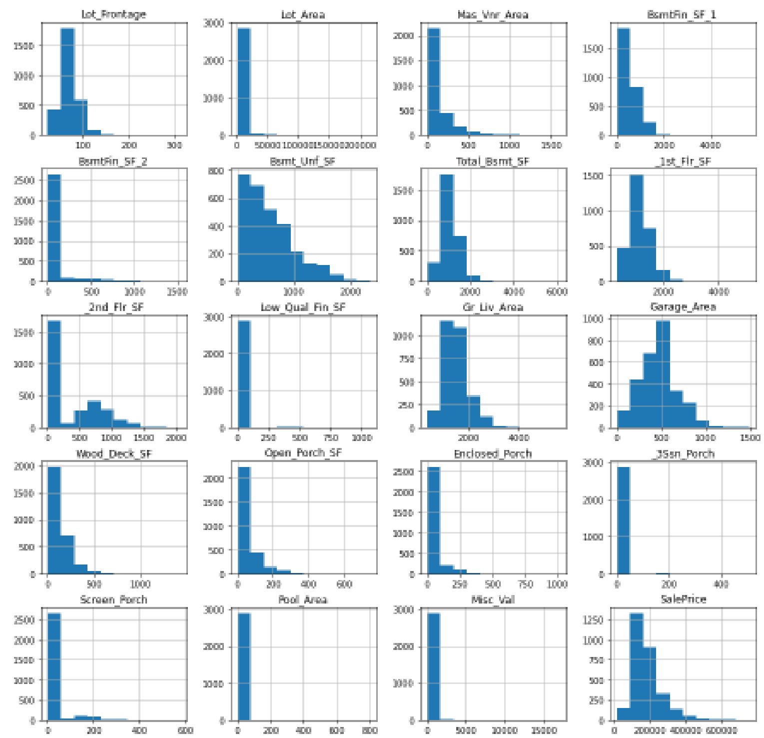
GarageArea, GarageCars, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, BsmtFullBath, BsmtHalfBath, MasVnrArea (10 variables).

Group 3 - Other variables:

Functional, MSZoning, Electrical, KitchenQual, Exterior1st, Exterior2nd, SaleType, Utilities, LotFrontage, GarageYrBlt (9 variables).

Skewness:

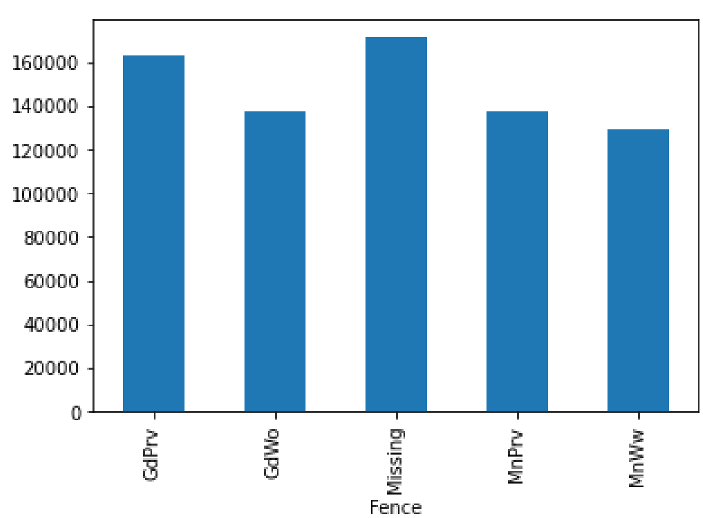
	Feature	Skewness
0	Lot_Frontage	1.648328
1	Lot_Area	12.825010
2	Mas_Vnr_Area	2.601343
3	BsmtFin_SF_1	1.423925
4	BsmtFin_SF_2	4.147500
5	Bsmt_Unf_SF	0.920388
6	Total_Bsmt_SF	1.156126
7	_1st_Flr_SF	1.467854
8	_2nd_Flr_SF	0.861897
9	Low_Qual_Fin_SF	12.095092
10	Gr_Liv_Area	1.270562
11	Garage_Area	0.242418
12	Wood_Deck_SF	1.843687
13	Open_Porch_SF	2.536328
14	Enclosed_Porch	4.006211
15	_3Ssn_Porch	11.382039
16	Screen_Porch	3.949081
17	Pool_Area	16.907100
18	Misc_Val	21.958474
19	SalePrice	1.742123



- Most of the variables in our dataset are right skewed as it has a positive skewness.
- Lot_Area, Low_Qual_Fin_SF and _3Ssn_Porch are highly positively skewed.
- Garage_Area, Bsmt_Unf_SF, _2nd_Flr_SF are the closest to normal distribution.

Relation between missing value and Sale Price

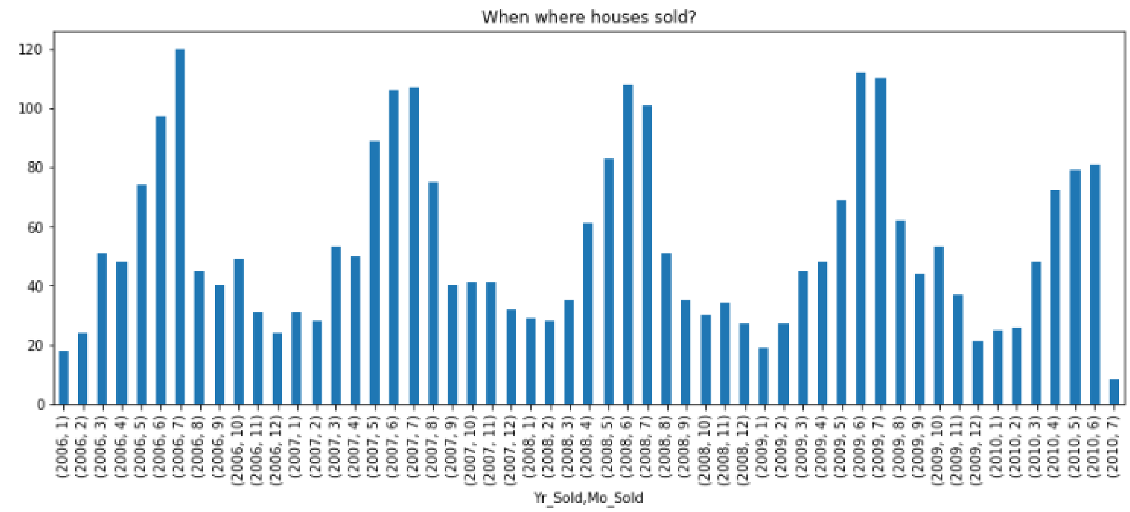
The Missing values have higher sales price in comparison to other categories



We replaced NA value as the missing value. From the graph , we can see that the missing value is higher than the other features. It does not depend on other features nor it is independent. Those missing values depend on missingness, so these missing data types are called Missing Not At Random (MNAR).

Trend in house sales by year and month

```
df.groupby(['Yr_Sold','Mo_Sold']).PID.count().plot(kind='bar', figsize=(14,5))
plt.title('When where houses sold?')
plt.show()
```



From the above graph we can observe a trend that most houses were sold between the months May-July which is the summer season. However, this isn't sufficient information to conclude if the sale price would essentially be higher in these months as the trend could be due to many reasons. One conjecture to this trend could be the fact that buyers have received tax refunds after paying their taxes and possess extra money to buy houses around this period.

Null Hypothesis:

```
#getting the count of missing values

missing = ames.isna().sum()
missing = missing[missing > 0]
percent_missing = missing*100/ ames.shape[0]

table = pd.concat([missing, percent_missing], axis = 1, keys= ["Missing values", "Percentage"])
table.sort_values(by ="Missing values", ascending = False)
```

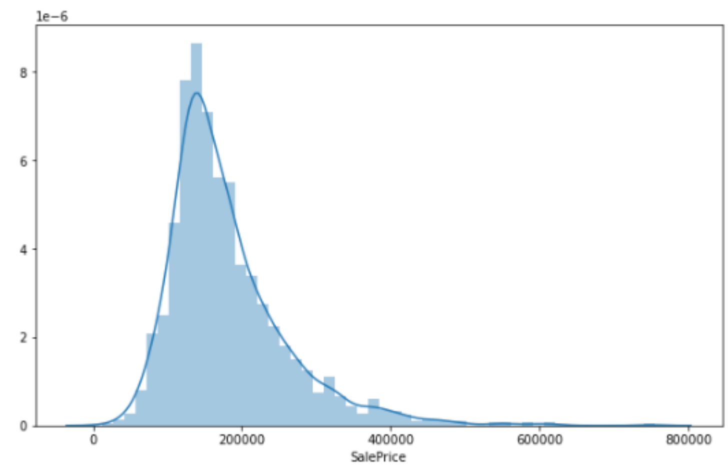
	Missing values	Percentage
Pool_QC	2917	99.56
Misc_Feature	2824	96.38
Alley	2732	93.24
Fence	2358	80.48

With 99.56% of data missing, Pool QC has the greatest percentage of missing values, although these data gaps can also indicate that the home does not also have a pool. Consequently, we must examine the values of the pool QC column in order to investigate that.

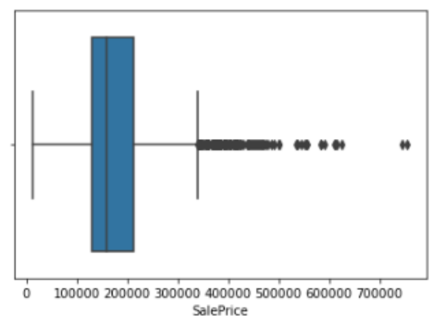
```
4 ames["Pool_Area"].value_counts()

0      2917
144       1
480       1
576       1
555       1
368       1
444       1
228       1
561       1
519       1
648       1
800       1
512       1
738       1
Name: Pool_Area, dtype: int64
```

We can see that there are 2917 entries in the pool area from the block of code above, which supports our theory that any home without a pool has a missing value in Pool QC.

Distribution of Sale Price:

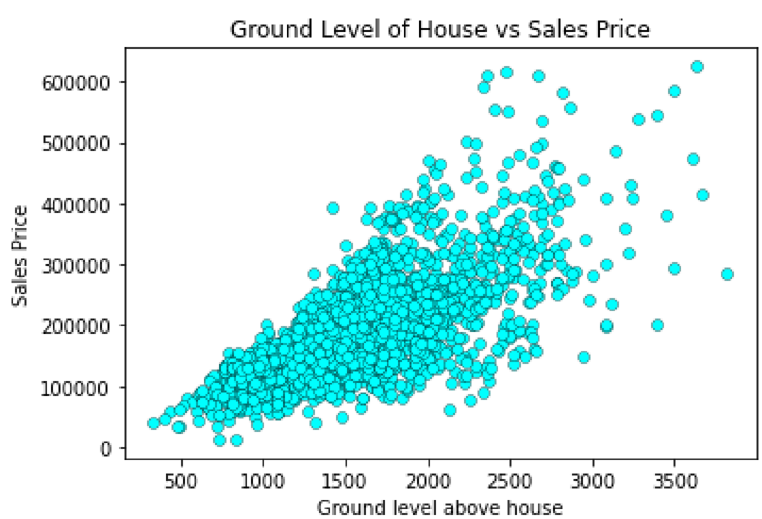



```
Out[341]: <AxesSubplot:xlabel='SalePrice'>
```



From the above graphs we can see that sale price is not normally distributed. Also houses with sale price greater than 350000 are outliers as depicted by the boxplot. The outliers are dealt with using cook's d test further in the analysis.

The below graph depicts the SalePrice after removal of outliners.



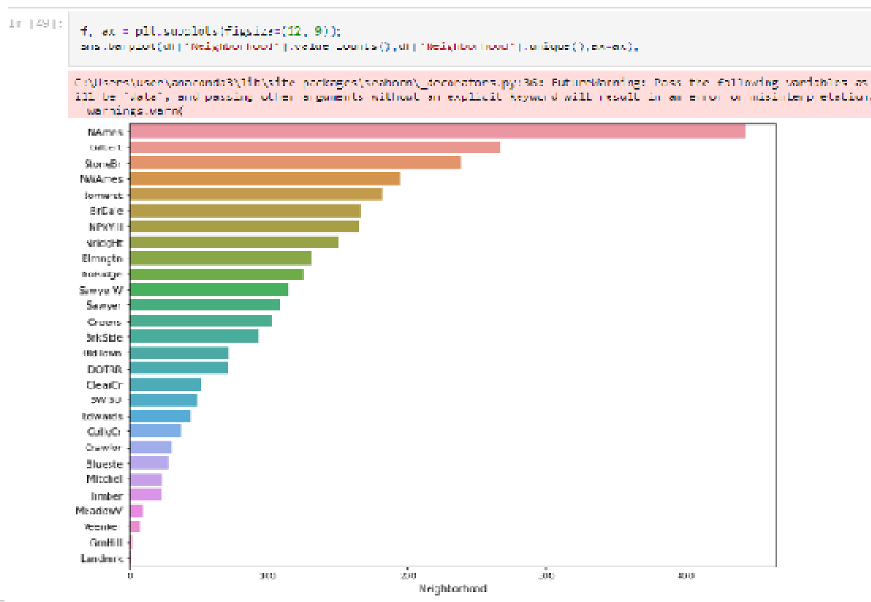
```
In [352]: print(f'The Skewness of Sale Price is:', df['SalePrice'].skew())
          print(f'The Kurtosis of Sale Price is:', df['SalePrice'].kurtosis())
```

```
The Skewness of Sale Price is: 1.7430176110528586
The Kurtosis of Sale Price is: 5.103825631594035
```

- With a skewness value of 1.74 we can say that the sale price is right skewed.
- The sale price has a kurtosis value greater than 3 indicates that it is leptokurtic.

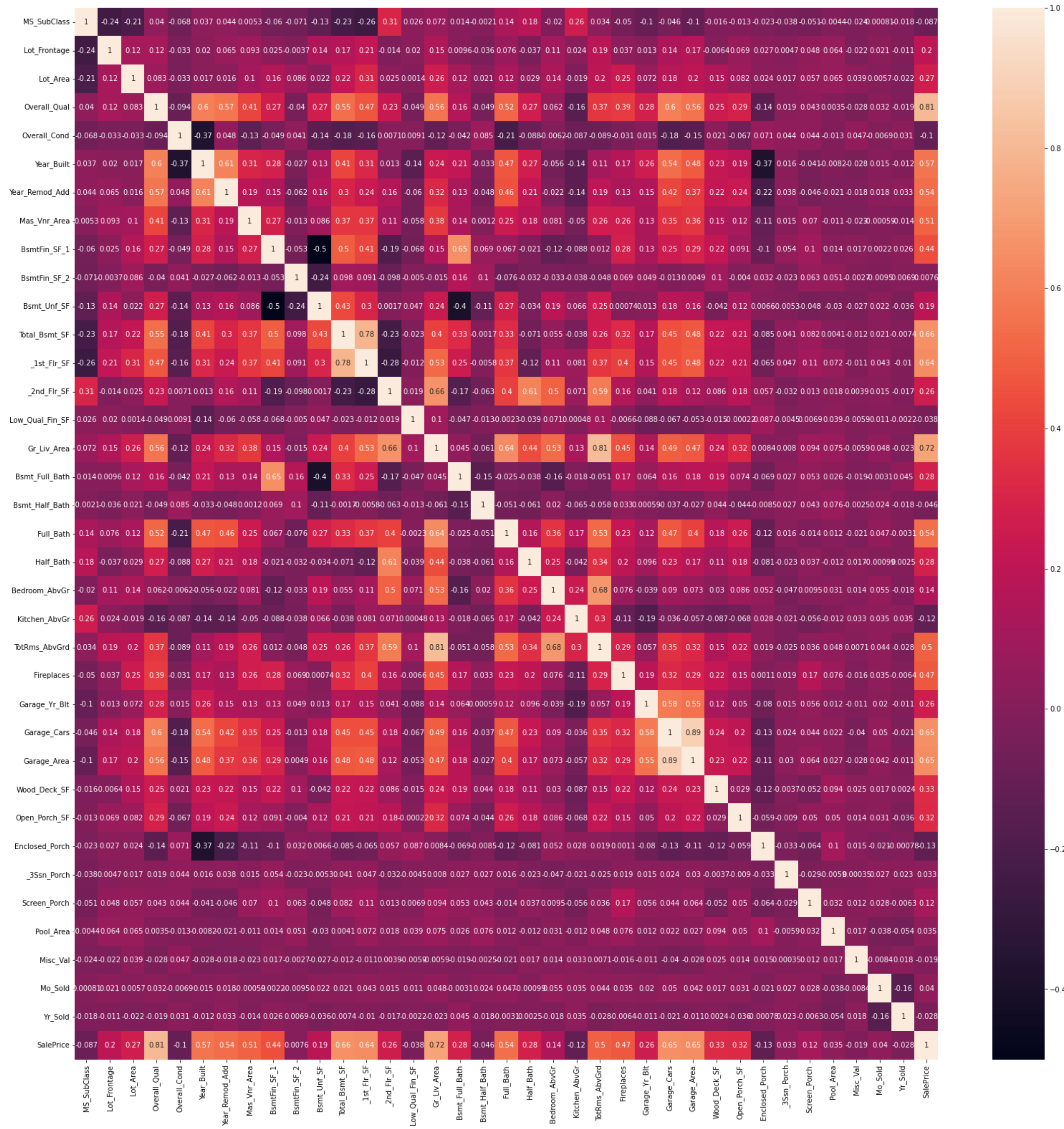
Neighborhood

	Count	Percentage
NAmes	443	15.119454
CollgCr	267	9.112628
OldTown	239	8.156997
Edwards	194	6.621160
Somerst	182	6.211604
NridgHt	166	5.665529
Gilbert	165	5.631399
Sawyer	151	5.153584
NWAmes	131	4.470990
SawyerW	125	4.266212
Mitchel	114	3.890785
BrkSide	108	3.686007
Crawfor	103	3.515358
IDOTRR	93	3.174061
Timber	72	2.457338
NoRidge	71	2.423208
StoneBr	51	1.740614
SWISU	48	1.638225
ClearCr	44	1.501706
MeadowV	37	1.262799
Blmngtn	28	0.955631
Veenker	24	0.819113
NPkVill	23	0.784983
Blueste	10	0.341297
Greens	8	0.273038
GrnHill	2	0.068259
Landmrk	1	0.034130



- It was observed that most of the houses in the data belonged to the NAmes, CollgCr & OldTown.
- Since only a mere number of houses belonged to the Landmark (0.03%), GrnHill(0.06%) and Greens(0.27%) neighborhood we conclude that there isn't sufficient data for these neighborhood's for us to provide any prediction.
- We thereby drop those observations.

Correlation Heat Map:



A statistical measure known as correlation expresses how linearly related two variables are to one another. It's a common technique for describing straightforward connections without explicitly stating cause and effect. From the above correlation we can see that there is a high correction between Garage Cars and Garage Area which indicates, as the Garage Area increases its car capacity increases as well.

Feature Selection

The number of categorical features in the dataset is very high and is difficult to perform analysis. Moreover, with more features it would also take the model more time to train. With more and more organizations moving to the cloud this would not be cost beneficial for organizations as they would have to pay more for cloud services.

```
from sklearn.model_selection import train_test_split
X_train, y_train, X_test, y_test= train_test_split(X,y,test_size=0.25,random_state=100)
```

```
from mlxtend.feature_selection import SequentialFeatureSelector as sfs
from sklearn.linear_model import LinearRegression

lreg = LinearRegression()
sfs1 = sfs(lreg, k_features=4, forward=True, verbose=2, scoring='r2')

sfs1 = sfs1.fit(X, y)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    0.0s remaining:    0.0s
[Parallel(n_jobs=1)]: Done 228 out of 228 | elapsed:    2.1s finished

[2022-12-06 17:38:01] Features: 1/4 -- score: 0.34460378989866847[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    0.0s remaining:    0.0s
[Parallel(n_jobs=1)]: Done 227 out of 227 | elapsed:    2.6s finished

[2022-12-06 17:38:04] Features: 2/4 -- score: 0.5119594379716824[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    0.0s remaining:    0.0s
[Parallel(n_jobs=1)]: Done 226 out of 226 | elapsed:    2.7s finished

[2022-12-06 17:38:07] Features: 3/4 -- score: 0.5507320745676669[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    0.0s remaining:    0.0s
[Parallel(n_jobs=1)]: Done 225 out of 225 | elapsed:    2.9s finished

[2022-12-06 17:38:10] Features: 4/4 -- score: 0.5889142716098247
```

```
feat_names = list(sfs1.k_feature_names_)
print(feat_names)
len(feat_names)
```

```
['Exter_Qual_TA', 'Bsmt_Qual_Ex', 'Bsmt_Qual_Gd', 'Kitchen_Qual_Ex']
4
```

In order to select the 4 most important categorical features we opted for the forward selection method. The R square scoring method was used to check if we get better explained variance in each iteration. The results show that the R^2 value increases from 0.344 to 0.588 in 4 folds. From this analysis we can conclude the Exter_Qual, Bsmt_Qual & Kitchen_Qual are 3 important important features we can use

Part-2 : Analysis:

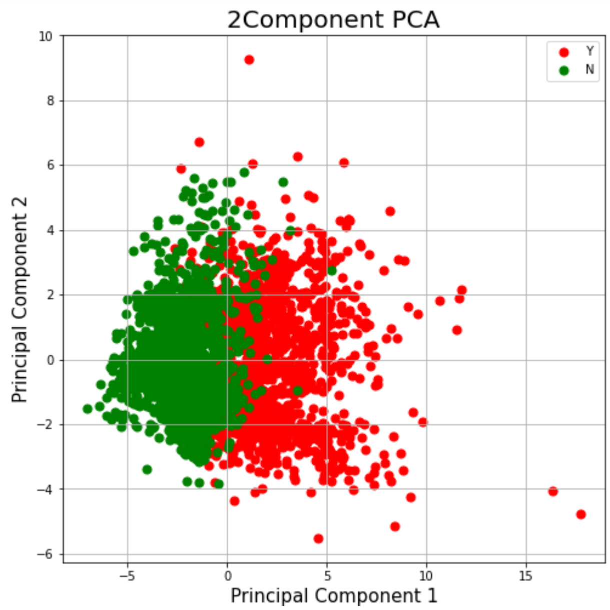
The Ames dataset's intuitiveness is one of its benefits. We all know what the main characteristics of the housing market would be, so it's not particularly complicated to understand. Generally speaking, the number of rooms, the quality of the kitchen or other remodeling, and the size of the lot are the factors that have the greatest impact on a home's price, according to a stranger. Neighborhood. economy as a whole.

Modeling

Principal Component Analysis

Two Component PCA

The 2 component PCA is used mainly to visualize multidimensional data. However, the trade off is that there is a lot of variance lost when we converge multi dimensional into only 2 principal components.



	principal component 1	principal component 2	SalePrice > 160000
0	1.440707	-1.614244	Y
1	-2.413093	-1.784224	N
2	-0.318280	-1.156350	Y
3	3.130048	-1.029130	Y
4	0.931526	0.288805	Y

Variance Explained:

array([0.20689964, 0.08339449])

Variance explained by two components is only 28.3%. Which shows that a lot of information is lost by converting multi dimensional data into two principal components.

6 Component PCA

	principal component 1	principal component 2	principal component 3	principal component 4	principal component 5	principal component 6	SalePrice > 160000
0	1.440512	-1.619379	3.045861	-2.303518	1.714278	-0.241770	Y
1	-2.413306	-1.777833	0.120563	-1.728939	1.549481	-0.777520	N
2	-0.318814	-1.118856	2.474123	-2.792766	1.076172	0.662210	Y
3	3.129856	-1.033513	1.860602	-1.871084	2.407550	0.825049	Y
4	0.931483	0.288697	-1.596381	-2.681560	0.389814	-0.450497	Y

For 6 Component PCA the explained variance is 49.3%

Linear Regression

Why Linear Regression?

It determines the nature of the relationship between the variables. It's really simple to learn how to use this strategy. Regularization decreases overfitting, it lowers the complexity to prevent overfitting. Additionally, it decides how strong the predictors will be.

Linear Regression Assumptions

Five fundamental assumptions of the linear regression model include: linear relationship, multivariate normality, lack of or minimal multicollinearity, lack of autocorrelation, and homoscedasticity (same variance).

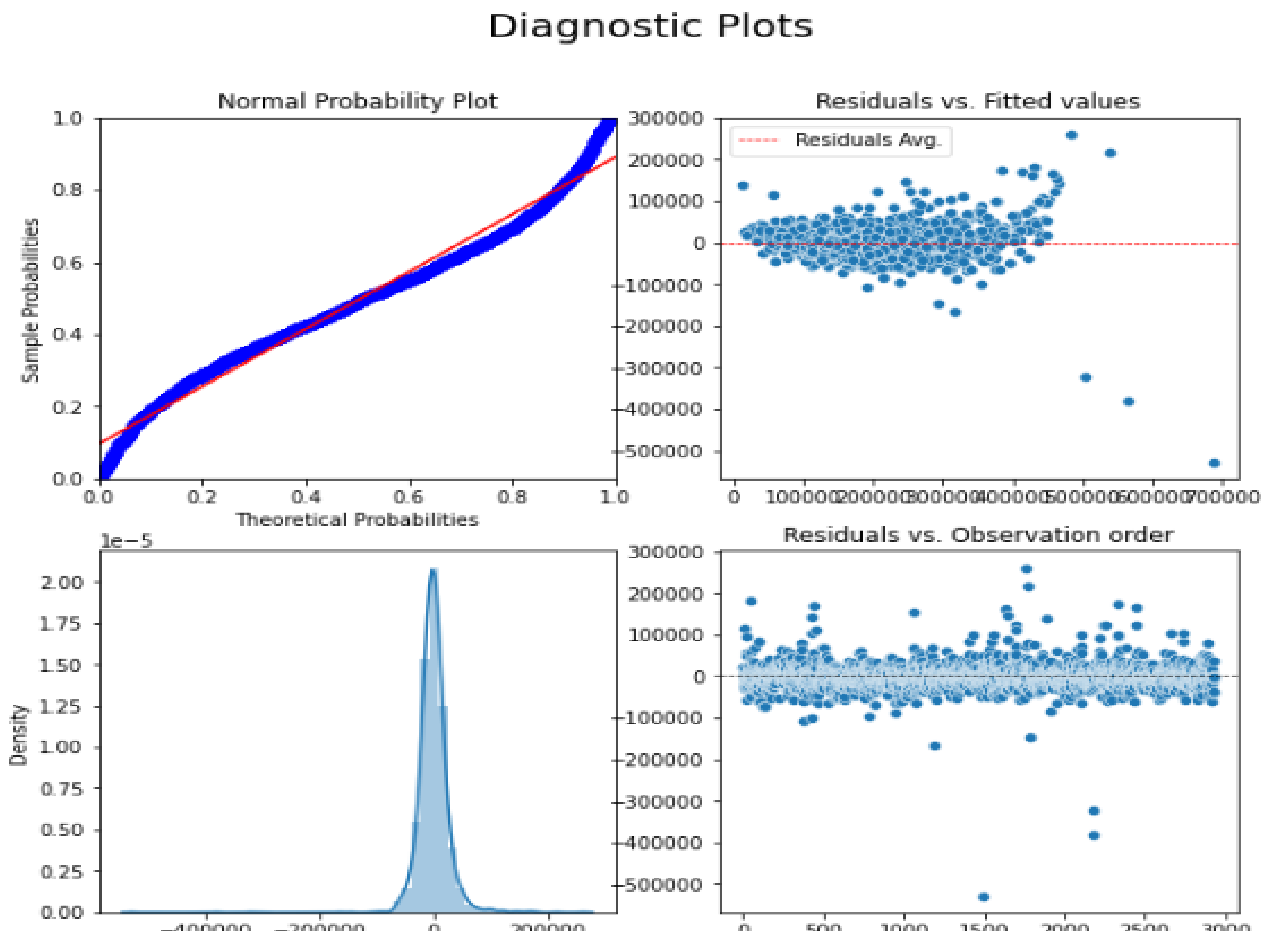
Taking into account 52 independent variables (this includes categorical variables which were converted to dummies) and Sales Price as the dependent we performed the Ordinary Least Square regression and obtained the following results.

As may be observed, the majority of the explanatory variables in the OLS model are statistically significant at the 86% level (i.e. *p-value* < 5%), as indicated by the individual t-statistics and associated p-values.

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.868
Model:	OLS	Adj. R-squared:	0.866
Method:	Least Squares	F-statistic:	395.7
Date:	Thu, 01 Dec 2022	Prob (F-statistic):	0.00
Time:	13:56:45	Log-Likelihood:	-34262.
No. Observations:	2930	AIC:	6.862e+04
Df Residuals:	2881	BIC:	6.892e+04
Df Model:	48		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.832e+06	5.03e+06	0.762	0.446	-6.03e+06	1.37e+07
Order	-3.0125	5.732	-0.526	0.599	-14.251	8.226
PID	-6.271e-06	5.63e-06	-1.113	0.266	-1.73e-05	4.78e-06
MS_SubClass	-149.0977	16.462	-9.057	0.000	-181.377	-116.818
Lot_Frontage	-29.9477	32.465	-0.922	0.356	-93.604	33.709
Lot_Area	0.5371	0.079	6.833	0.000	0.383	0.691
Overall_Qual	1.146e+04	749.644	15.284	0.000	9987.530	1.29e+04
Overall_Cond	5601.8658	617.705	9.069	0.000	4390.677	6813.054
Year_Built	279.3543	43.590	6.409	0.000	193.884	364.825
Year_Remod_Add	72.3315	42.937	1.685	0.092	-11.859	156.522
Mas_Vnr_Area	18.8864	3.708	5.094	0.000	11.617	26.156
BsmtFin_SF_1	9.0681	1.523	5.956	0.000	6.083	12.054
BsmtFin_SF_2	5.2540	2.605	2.017	0.044	0.147	10.361
Bsmt_Unf_SF	-2.0296	1.499	-1.354	0.176	-4.969	0.909
Total_Bsmt_SF	12.2921	2.410	5.100	0.000	7.566	17.018



Residuals vs Fitted values

- It is a scatter plot of residuals on the y axis and fitted values (estimated responses) on the x axis. The plot is used to detect non-linearity, unequal error variances, and outliers**.**
- A model's performance can be measured using a variety of charts. The graph displays residuals in comparison to fitted values (between the actual and predicted values). One must examine the trend line in order to determine whether the residuals are in accordance with the presumption that they are normally distributed with a mean equal to zero. It must closely resemble the plot's y = 0 line (Cotton). It is mostly true in this instance. The majority of the points appear to be plotted near to y = 0, with a few spots deviating further from the trend line.
- The residuals vs. fitted values scatter plot shows that the average of residuals, red line, is very close to zero, so the linearity assumption is satisfied
- However the vertical distribution is not so constant by which we can conclude that the variance is not constant
- The residuals are left skewed and we can conclude that residuals do not follow a normal distribution.
- The residual vs order plots show that there are outliers in our data which can influence our regression line.

Dealing with Outliers

- We performed the Cook's D and DFFITS test to find those observations that are extreme in nature and influence the regression line. Any distance above 0.5 in the Cook's D test is considered to be an outlier which negatively impacts our regression analysis.
- After analyzing we found out that the observations at indices 210,1498 and 2180 have cook's distance of more than 0.5.

Cooks_d and DFFITS test for checking if outliers are influential parameters

In [22]:
```
trial.iloc[[result.resid.sort_values().head(1).index[0]]]
```

Out[22]:

	Order	PID	MS_SubClass	Lot_Frontage	Lot_Area	Overall_Qual	Overall_Cond	Year_Built	Year_Remod_Add	Mas_Vnr_Area	BsmtFin_SF_1	BsmtFin_SF_2	Bsmt_Unf_SF	Total_Bsmt_SF	_1st_Flr_
1498	1499	908154235	60	313.0	63887	10	5	2008	2008	796.0	5644.0	0.0	466.0	6110.0	46

In [23]:
```
k = trial.shape[1] - 1 #No. of predictors
n = trial.shape[0] #No. of observations
diffits_ref = 3*np.sqrt((k+2)/(n-k-2)) #Reference value of DIFFITS
influence = result.get_influence()
print(diffits_ref)
```

```
0.40718264445294244
```

In [24]:
```
influence.summary_frame().loc[(influence.summary_frame(
).cooks_d > 0.5), influence.summary_frame().columns[-6:]]
```

Out[24]:

	cooks_d	standard_resid	hat_diag	dffits_internal	student_resid	dffits
210	2.771877e+09	-1.629110	1.000000	-383287.709326	-1.629110	-383287.709792
1498	2.206988e+00	-20.444594	0.218656	-10.815284	-22.107558	-11.695000
2180	1.931442e+00	-13.668993	0.353955	-10.117631	-14.132570	-10.460765

- We removed these influential parameters and performed the OLS test again.

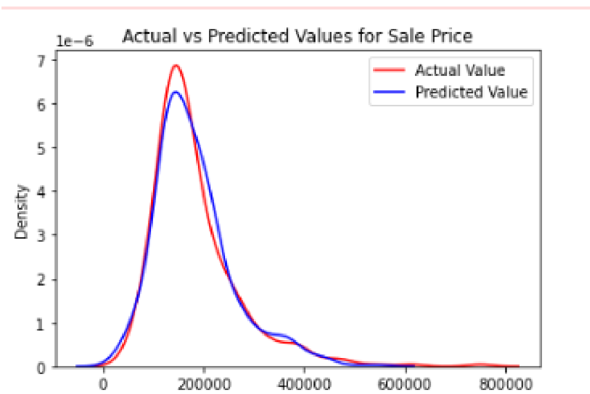
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SalePrice   R-squared:                       0.898
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     526.5
Date:                Thu, 01 Dec 2022   Prob (F-statistic):               0.00
Time:                        14:00:03   Log-Likelihood:                -33869.
No. Observations:                2928   AIC:                         6.784e+04
Df Residuals:                    2879   BIC:                         6.813e+04
Df Model:                          48
Covariance Type:            nonrobust
```

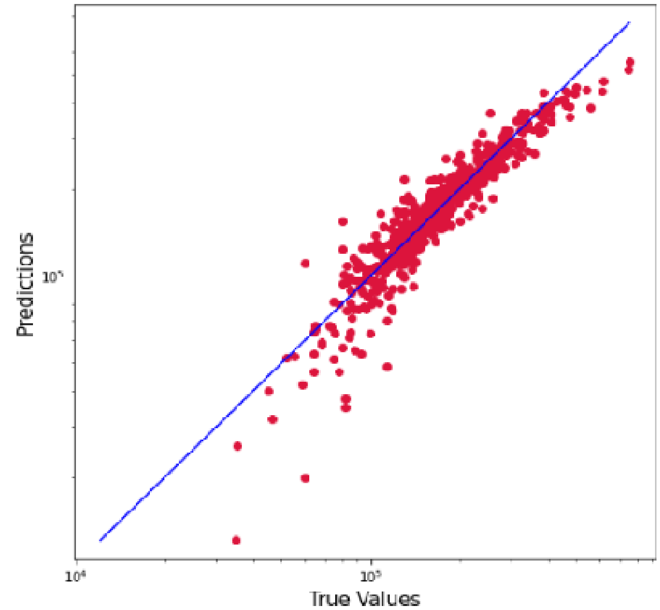
- Upon removal of outliers we fetched a R^2 value of 89.8% which was 3% better than the model containing outliers.

Prediction:

We have used 75% of our data for training and 25% for testing. We obtained a R^2 score of 0.899. From this we conclude that 89.9% of the variation in Sales Price was explained by our model.

- From the density distribution plot we can see that the predicted sale price follows the actual values to a fair extent.
- From the scatter plot it can be observed that most of the predicted values are close to the straight line. By this it's clear that the distance between actual and predicted values is less.





Logistic Regression

- Logistic Regression is used when the dependent variable is binary in nature. An additional column was added to the dataset with the condition that if the sale price was greater than 180000 it was 'Yes' otherwise 'No'. We can now perform logistic regression to tell our clients if their house would sell for more than $180000 or not.
- We split the data into 75% training and 25% testing .

```
In [84]: from sklearn.metrics import classification_report
         print (classification_report(y_test, y_pred_test))

                  precision    recall  f1-score   support

              No       0.83      0.77      0.80       460
             Yes       0.65      0.72      0.69       270

        accuracy                           0.75       730
       macro avg       0.74      0.75      0.74       730
    weighted avg       0.76      0.75      0.76       730
```

- Accuracy is not a good metric for imbalance data so we choose f1 score. F1 score is a harmonic mean of precision and recall.
- The f1 score range is from 0 to 1.
- **F1 Score** : This value is calculated as:

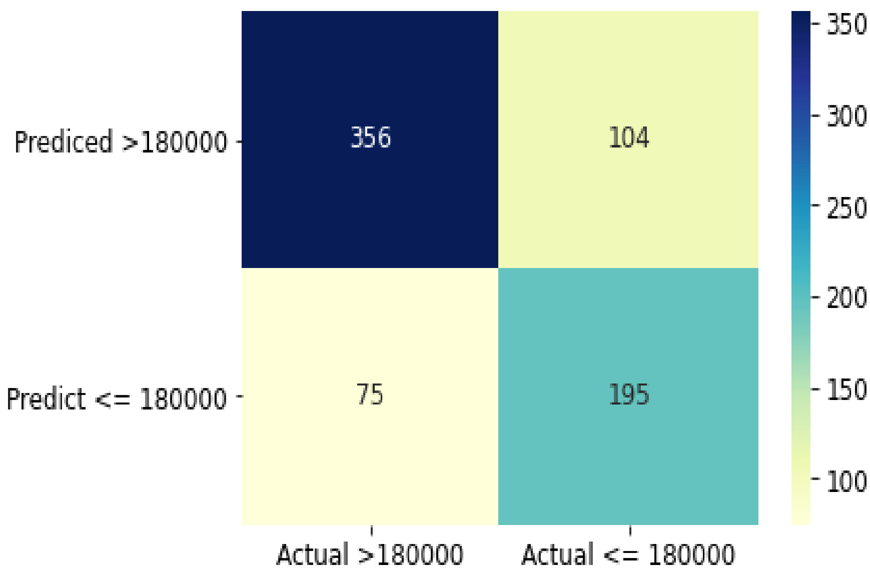
  F1 score = 2* (precision*recall) / (precision+recall)

  F1 score = (.65 * .72) / (.65 +. 72)

  F1 score = 0.69
- Our f1 score of 0.69 depicts that the houses which are predicted is more than 180000, is fairly accurate as the value is closer to 1 .

```
cm_matrix = pd.DataFrame(data=cm, columns=['Actual >180000', 'Actual <= 180000'],
                                 index=['Prediced >180000', 'Predict <= 180000'])

sns.heatmap(cm_matrix, annot=True, fmt='d', cmap='YlGnBu')
```

```
<AxesSubplot:>
```



From above confusion matrix :

- 356 houses were predicted to have sales price > 180000 and the prediction was correct (TP)
- 104 houses were predicted to have sales price > 180000 but the actual price was <=180000 (FP)
- 75 houses were predicted to have sales price <= 180000 but the actual price was >180000 (FN)
- 195 houses were predicted to have sales price <= 180000 and the actual price was <= 180000 (TN)

Sensitivity: Is the True Positive Rate (TPR). The value is 0.64
Specificity: Is the True Negative Rate (TNR) i.e for all times a house was predicted to have sale price lesser than180000 how many were true. For our model the value is 0.652

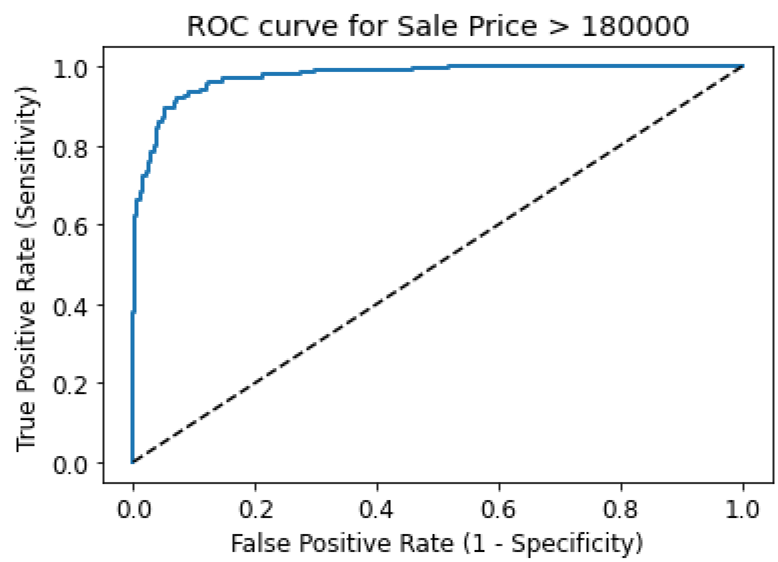
1- Specificity: number of times houses had sale price lesser than 180000 and model predicted it to be greater than 180000.

Type I and Type II errors:

- Type I error occurs when we reject the null hypothesis which is actually true (False Positive). Type II error occurs when we fail to reject the null hypothesis while it is actually false (False Negative).
- For our business problem we would prefer Type II errors over Type I because it would satisfy the customer more that his/her house was sold for a price more than the initial predicted value. Type I error is one which we would want to stay away from as we don't want to dissatisfy our clients by raising their hopes and not living up to it.

ROC Curve:

- The ROC curve is the plot between Sensitivity vs 1-Specificity.
- The straight line depicts a threshold of 0.5. This indicates the probability of an event happening by chance.
- The ROC curve shows that the threshold curve is above the 0.5 line which means the probability of prediction is more than that of a coin toss.
- The curve also maximizes the area under the curve.



Conclusion:

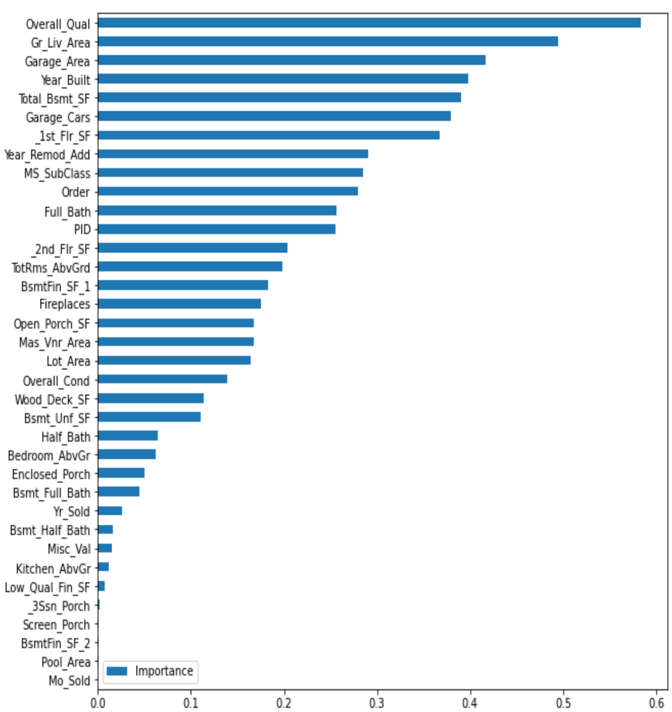
By analyzing the data collected in the Ames, Iowa real estate market, we created a model that can help future sellers price their homes in the market to sell quickly while still generating a profit. In this analysis we have used linear regression to determine the price and study how the data are related to one another. Regularization reduces overfitting by bringing down the complexity to do so. It also determines how reliable the predictors will be as well as. Also utilizing logical regression, we could determine the threshold prices for our clients. The most important factors when determining the price, as determined by our analysis, are the year built, excellent kitchen and basement quality, the square footage of both the basement and first floor, the square footage of both above-grade living area and garage area, and the overall quality (as determined by material and finish) of the home. Because our model is based on these variables, we believe it to be a useful tool for real estate agents to utilize in the Ames, Iowa market.

Q-1] What is the expected selling price of my home?

After thourough analysis of the ames data our model predicts the range of sale price to be between $12088 and $554360 with a the average sale price a house being $181065 with a standard deviation of $277477.

Q-2] What factors influence the price of my home?

Overall Quality, Ground Floor Living Area ,Garage Area, Basement square foot and Year Built influence the sale price of the house.



	Importance
Overall_Qual	0.582958
Gr_Liv_Area	0.494807
Garage_Area	0.417028
Year_Built	0.397811
Total_Bsmt_SF	0.389964
Garage_Cars	0.379305
_1st_Flr_SF	0.366936
Year_Remod_Add	0.290117
MS_SubClass	0.285333
Order	0.279882
Full_Bath	0.256331
PID	0.255116
_2nd_Flr_SF	0.203404
TotRms_AbvGrd	0.198636
BsmtFin_SF_1	0.182439
Fireplaces	0.174942
Open_Porch_SF	0.168103
Mas_Vnr_Area	0.167075
Lot_Area	0.164303
Overall_Cond	0.139350
Wood_Deck_SF	0.113665

```
In [82]: mutual_info_regression(X,y)

Out[82]: array([0.2808791 , 0.26110144, 0.29120166, 0.16869292,
               0.12988955, 0.39251628, 0.29754578, 0.18173966,
               0.01219243, 0.10981941, 0.39189659, 0.36878562,
               0.01058295, 0.49382987, 0.03595756, 0.01468358,
               0.06944315, 0.04050694, 0.02117925, 0.19374289,
               0.38349104, 0.41609456, 0.10938874, 0.18108894,
               0.00159466, 0.00977393, 0.01028445, 0.00184914,
               0.02905659])
```

The mutual information measures the dependency between two variables. A value close to one depicts high dependency between the two variables and value close to 0 means the variables are closer to being independent.

If X and Y are two random variables:

$I(X;Y) = H(X) - H(X|Y)$

$I(X;Y)$ = Mutual Information of X and Y;

$H(X)$ = Entropy of X

$H(X|Y)$= Entropy of X given Y

Q-3] Which factors are more important than others?

```
high_corr = ames.select_dtypes(include = [np.number])
corr = high_corr.corr()
print("correlation with respect to saleprice")
print(corr["SalePrice"].sort_values(ascending = False)[:6], "\n")

#displaying the top 5 correlation
```

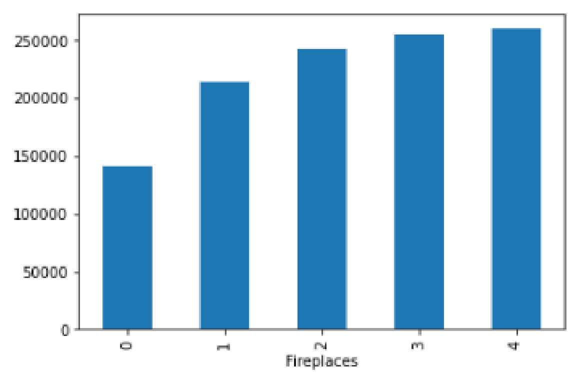
```
correlation with respect to saleprice
SalePrice          1.00
Overall_Qual       0.81
Gr_Liv_Area        0.72
Total_Bsmt_SF      0.66
Garage_Cars        0.65
Garage_Area        0.65
Name: SalePrice, dtype: float64
```

Q-4] How much should I invest in improving the condition of my home in order to increase the expected sale price by more than the cost of imporvements?

1.Fireplace:

```
df.groupby('Fireplaces')['SalePrice'].mean().plot.bar()
```

<AxesSubplot:xlabel='Fireplaces'>




```
print("Upgrade from 0 to 1:",213556.001570-141195.772152)
print("Upgrade from 0 to 2:",242316.162896-141195.772152)
print("Upgrade from 0 to 3:",255820.833333-141195.772152)
```

```
Upgrade from 0 to 1: 72360.229418
Upgrade from 0 to 2: 101120.390744
Upgrade from 0 to 3: 114625.061181
```

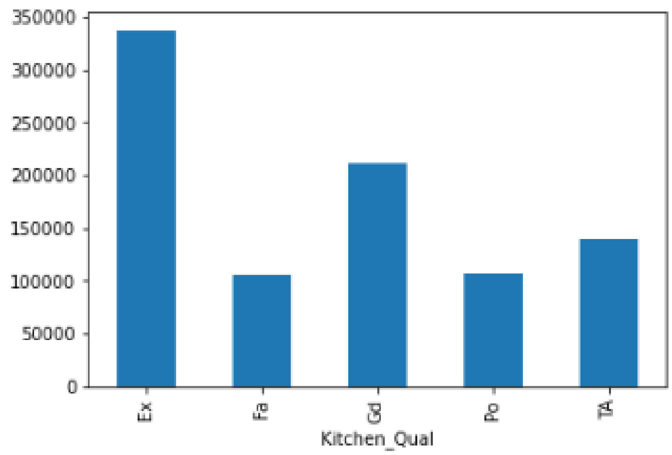
Fireplaces	SalePrice
0	141195.772152
1	213556.001570
2	242316.162896
3	255820.833333
4	260000.000000

The analysis shows that the houses with no fireplace has a low mean sale's price. Upgrading these houses to have 1 fireplace would increase the sale price by $72360. As an analyst we suggest our customers having no fireplace to upgrade to one as the return on investment is more than the cost of investment.

2.Kitchen Quality:

Out[11]:

Kitchen_Qual	SalePrice
Ex	337339.341463
Fa	105907.042857
Gd	210835.582759
Po	107500.000000
TA	139549.947791

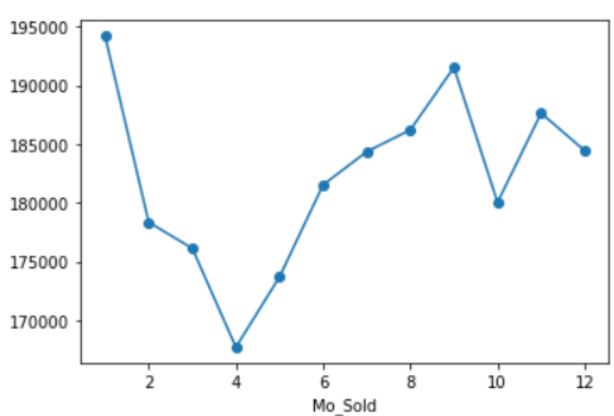


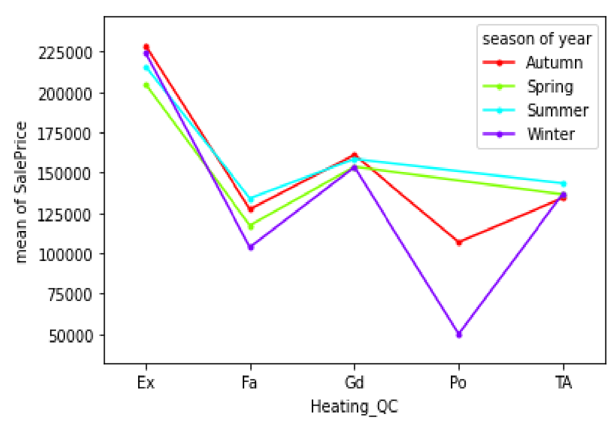
As an analyst we suggest our clients with houses having fair and poor kitchen quality to renovate their kitchen. By upgrading the kitchen quality to "good" results in a 49.76% increase in the sales price of the house which is quite significant. Kitchen quality happens to be a significant feature in our model so we would strongly recommend renovation to customers.

Q-5] Which homes should I compare my house to?

One must compare their houses to those which have better kitchen quality, exterior quality, basement quality in the same neighborhood as these are the three important features for our model They should look to renovate these parameters for them to fetch a better sales price .

Q-6] When is the best time of the year to sell a home?





There are two factors we as analysts would take into account before suggesting what is best for our clients. The month of January fetches the best price. However, our analysis also shows that in the summer season the demand for houses is more. It would be easier for a seller to find a buyer in these months. All in all january, may, june & july are the best months of the year to sell a house.

Recommendations:

According to our concept, a person wishing to raise the value of their home might do the following actions:

- They make an effort to improve their home's interior and external qualities by remodeling.
- If employing a hardboard exterior, change to a cement or brick exterior.
- Expand the garage to accommodate more than one automobile.
- Ensure that only one indoor kitchen is built (if there are multiple kitchens).
- Reduce the number of bedrooms in the house, or transform the ones that are already there into multipurpose spaces (if the house has more than three bedrooms).

Given that each city tends to differ significantly in terms of external characteristics like geographical features, seasonal weather, or the specific city's economic climate, this model may not be applicable to other cities even though it generalizes well to the city of Ames.

Another thing to keep in mind is that this model does not account for home price increases. Housing costs in the US have been rising gradually year over year since the financial crisis ended in 2008. For our program to accurately estimate the current housing prices in Ames, it would require extensive retraining.

References:

- City of Ames Iowa (2002). Ames City Assessor Homepage. Retrieved March 24, 2011 from http://www.cityofames.org/assessor/
- Pardoe, I. (2008). “Modeling home prices using realtor data”, Journal of Statistics Education, Volume 16, Number 2. http://www.amstat.org/publications/jse/v16n2/datasets.pardoe.html
- Cotton, R. Introduction to Regression in R. Datacamp, Retrieved from: https://app.datacamp.com/learn/courses/introduction-to-regression-in-r
- Rego, F. (2015). Quick Guide: Interpreting Simple Linear Model Output in R. Retrieved from: https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R
- Statistics How To. (2016) Durbin Watson Test & Test Statistics. Retrieved from: https://www.statisticshowto.com/durbin-watson-test-coefficient/
- STHDA (2018). Best Subsets Regression Essentials in R. Retrieved from: http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/
- Williams, C. (2020). How to Create a Correlation Matrix with Too Many Variables in R. Towards Data science. Retrieved from: https://towardsdatascience.com/how-to-create-a-correlation-matrix-with-too-many-variables-309cc0c0a57
- Yobero, C. (2016). Methods for Detecting and Resolving Heteroskedasticity. Retrieved from: https://rpubs.com/cyobero/187387