Ames Home Prices Project Report – Group 1

In today's world the prediction of assets due to market fluctuations has become very difficult. House sales mostly depend on various factors such as neighborhood, city, utilities, space but ultimately selling price is the most crucial factors to be considered.

Ames Real Estate Associates" (AREA) is a company which helps customers to predict this crucial factor to sell their house. As a business perspective just predicting selling prices is not enough but finding out the time and customer expectations is needed. Providing customers better understanding of the housing market and making them aware of the condition of their house is necessary.

Business Problem:

As a data analyst, we will be determining sales prices for housing data of customers by a real estate company called "Ames Real Estate Associates" (AREA) to help them improve their ability to predict house prices and identify the factors that are likely to increase the price of homes that their customers want to sell. Developing a model which correctly predicts the value or selling price for your asset (house) considering all the characteristics and factors is important. There are many factors which affect the pricing of the house, factors like last sold price, total area, garage space, floors, material used, etc.

The company wants to predict at what price a customer can sell his house and what variables or characteristics can affect the selling price. Analyze the data to maximize the selling price and what can be improved to get the expected selling price for the house i.e., giving customer the correct insights from the analysis of their house.

Motivation:

The business problem helps us to identify the dependent and independent variables, the real challenge is to select the most impacting variables. Understanding the data and identifying variables using frequency analysis in SAS, then using data analysis techniques and implementing them. Converting our assumptions into practice is using regression techniques is needed here.

The data given is mostly a mix of categorical and nominal data where we need to find the variables associated with sales price having a significant correlation between them. Manually trying to weigh these factors against each other for every house would be extremely time-consuming, which makes this a great problem for data modelling to solve.

As our basis of modelling, we will sort the data where housing features are important and selling price is a dependent variable (predicted variable) and others are independent variable (where we assume the impact on dependent variable).

Constraints:

Not considering qualitative data in regression model, Quantitative factors such as neighborhood and house type can be factored in the regression model to localize the prediction if required.

Assumptions:

Previous selling data of the house is correct, and houses are considered to have the construction completed. We have assumed that in order to increase the quality of the house by one score (the newer investment made on renovating the house) should be equal to 5% of the average sales price of that neighborhood

Limitations:

Ames being a comprehensive housing dataset out there, which describes the sale of individual residential property. The data set contains 2931 observations and many variables (nominal, ordinal, discrete and continuous) involved in home values. These variables are therefore both objective/quantitative and a bit more subjective/qualitative. These large number of continuous variables give us many opportunities to differentiate themselves as they consider various methods of using such as combining the variables. Not considering all possible variables due to limited time and prioritizing other variables depending on our understanding of the problem.

Conditions:

We are removing the extreme outliers for variables like sales price, area related variables such as lot_area, garage area, front area, etc. Rest of the conditions are explained in Analysis phase.

Operational Definitions:

1. The sales price is the price of the house sells at before realtor commissions are taken out.
2. Sale date and month, which should represent the date when the sale contract was signed.
3. Overall condition is the condition of the house which considers every factor and gives score accordingly.

   Rest of the definitions are explained as an when we are using the variables.

SMART objectives:

Specific: As a data analyst we need to correctly analyze the dataset and formulate a modelling technique which helps customers to sell their houses at an accurate selling price interval. We have the dataset collected by Ames housing group, we will be using this dataset divided into test and train for our modelling.

Measurable: This goal can be achieved by eliminating the outliers, exploring the data and preparing the dataset by identifying the null values or blank fields. Also understanding the factors that influence the variance in the predictors.

Achievable: We will achieve this by using SAS, python and excel as our tools for different phases of analysis.
Relevant: As a data analyst testing the dataset and correctly predicting the results is a challenge but our assumptions are relevant that we can fulfill this demand with CALC method.
Time-Bound: The target is to analyze the customer situation and all the factors and predict the selling price. Considering certain recommendations this can be completed within the time frame provided.

List of questions to be answered:

1. What is the expected selling price of my home?
2. What factors influence the price of my home?
3. Which factors are more important than others?
4. How much should I invest in improving the condition of my home in order to increase the expected price by more than the cost of improvements?
5. Which homes should I compare my house to?
6. When is the best time of the year to sell my home?

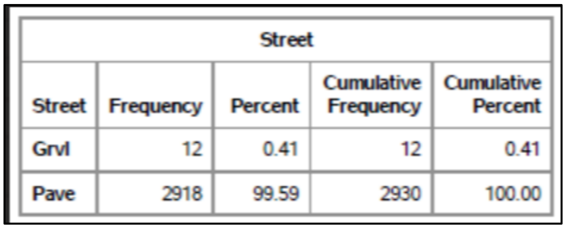
Data Preparation, exploration, and Understanding:

We have total 82 variables and >2500 records to be analyzed, for that we read the data dictionary and understood the data variables and categorized it into different dataset into qualitative, quantitative, categorical, continuous, discrete variables. This took a long time as the dataset is huge.

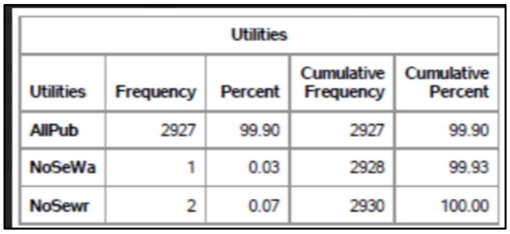
To filter unstructured, inconsistent, and disordered data we started with removing unnecessary data and outliers. By using the 'FREQ' in SAS we have figured out the unstructured variables, *Street* has 99.59 percent single value or null values which will not have much impact in analysis phase, similarly for *land contour, utilities, land slope, blg_type, roof_style,* etc. We will be explaining the requirement of the variables in regression modelling by looking at the p-values and identifying the variables which are not useful depending upon their significance.

Using appropriate patterns for refining all the data and filling the empty space for data flow need to be done for the unstructured data. *E.g.: Alley has 93.24 null values; garage is absent for most of the houses and has 157 null values and therefore we identified all the further garage observations for these 157 rows replaced these rows values as 'no garage'.*

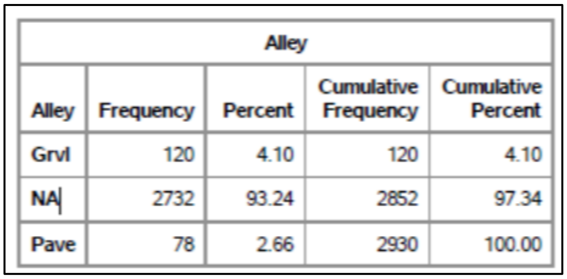
Find below the results of 'FREQ' to determine this,



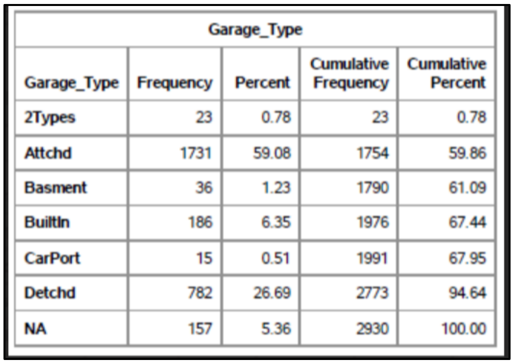
Street				
Street	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grvl	12	0.41	12	0.41
Pave	2918	99.59	2930	100.00



Utilities				
Utilities	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AllPub	2927	99.90	2927	99.90
NoSeWa	1	0.03	2928	99.93
NoSewr	2	0.07	2930	100.00

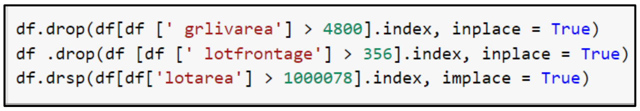


Alley				
Alley	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grvl	120	4.10	120	4.10
NA	2732	93.24	2852	97.34
Pave	78	2.66	2930	100.00



Garage_Type				
Garage_Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2Types	23	0.78	23	0.78
Attchd	1731	59.08	1754	59.86
Basment	36	1.23	1790	61.09
BuiltIn	186	6.35	1976	67.44
CarPort	15	0.51	1991	67.95
Detchd	782	26.69	2773	94.64
NA	157	5.36	2930	100.00

Outliers can be identified by box and whiskers plot and can be removed. The values which will drastically affect the model and may change the prediction can be eliminated in python by dropping theses values. Example for some variables –



```
df.drop(df[df [' grlivarea'] > 4800].index, inplace = True)
df .drop(df [df [' lotfrontage'] > 356].index, inplace = True)
df.drsp(df[df['lotarea'] > 1000078].index, implace = True)
```

Analysis:

We have explained 3 techniques in this project such as Multiple Regression, Cluster analysis and correlation matrix.

1. Multiple Regression:

a. Approach:

We have used multiple regression model to predict sales price using all the quantitative predictors present in the Ames data. The initial predictors are mentioned below:

[Lot_Frontage Order Lot_Area Year_Built BsmtFin_SF_1 Year_Remod_Add BsmtFin_SF_2 Bsmt_Full_ BathBsmt_Unf_SF Bsmt_Half_Bath Total_Bsmt_ SFFull_Bath _1st_Flr_SF Half_Bath _2nd_Flr_SF Bedroom_AbvGr Low_Qual_Fin_SF Kitchen_AbvGr Gr_Liv_Area TotRms_AbvGrd Garage_Area Fireplaces Wood_Deck_SF Garage_Yr_Blt Open_Porch_SF Garage_Cars Enclosed_Porch Mo_Sold _3Ssn_Porch Yr_Sold Screen_Porch Misc_Val]

Before starting the analysis, the data set was divided into test and train data, The train data was 70% of the complete data that was randomly selected out of the complete data set, this was the data that the model was built on. The test data is used to check the validity and claims of the model, which consist of remaining 30% of the complete data.

The approach was to create a model that provides maximum explanation of the variation with minimum possible and only significant predictors to keep the model parsimonious.

b. Modelling:

Post using the above predictors the regression model was revised by removing the insignificant variables. The significant was determined based on its P value from the regression result. Please find an example below.

Q. What factors influence the price of my home?

Predictors such as *Full_Bath, Half_Bath and Low_Qual_Fin_SF* are insignificant, and we have a 0.05 significance threshold that they don't not satisfy and therefore are removed from the actual Model. The variable such as *Gr_Liv_Area and Total_Bsmt_SF* are already covered as the individual split variables shown in the below screen capture.

Please find the SAS code for the regression model and data creation and regression model results in the below file:

Ames Regression model.sas

sales price prediction.xlsx

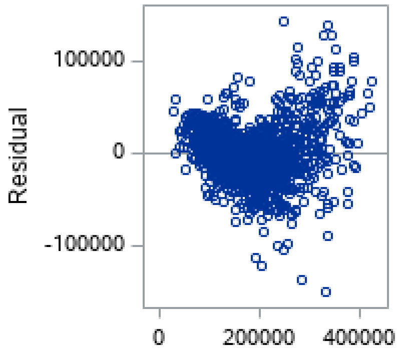
Initial Regression.pdf

Variable Impact:

Initially we started with all the quantitative variables (32) and model our regression around. We used the stepwise selection method in order analyze the contribution or difference the adding of each variable made to the regression the report for stepwise selection process is below: -

Stepwise selection for Quantitative vari

This report tells us that the best model was with 15 variables and all other variables were not significant at 0.005 level. These variables also created a highly asymmetrical residual graph.

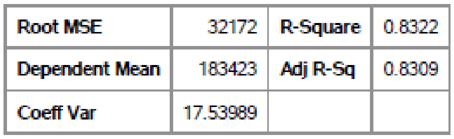


We also saw that the addition of the variables in step gave us an R2 increase from 0.49 to 0.77. all the methods like R2, Adjusted R2, AIC, BIC, SBC supported that model with 15 variables was the best model (all these graphs are present in the above report). Each predicted values and its residual with the prediction limit is also present in the report. Therefore, we found that removing the variables helped us reduce the difference between the R2 and ADJ R2 and helps the model to be less complex that only consider the useful variables and reduces unnecessary noise. *We can also infer that the model selected by stepwise selection are more dominant over the others from Sale price prediction*

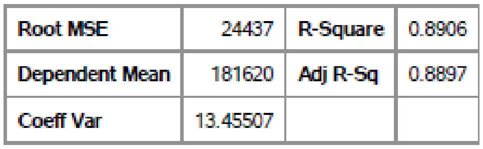
Outlier Impact:

Post the analysis of the initial model we found that there were some data points with massive residuals that were creating asymmetry in the residuals, we then removed those values for sales price more than 500000, this removed around 45 values. But this provided us with much more symmetric residual plot and the explained variation percentage or **R2** increased from 0.82 to 0.89, The adjusted R2 values also went up to a 0.889 and the Cook's distance was drastically reduced as seen below

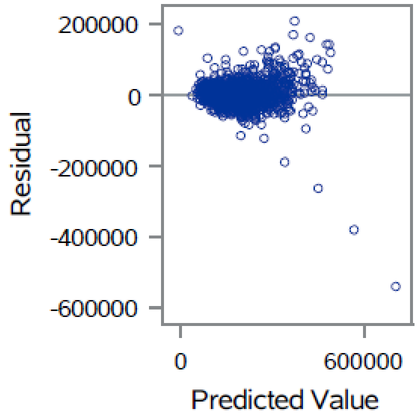
Regression post outlier removal.pdf

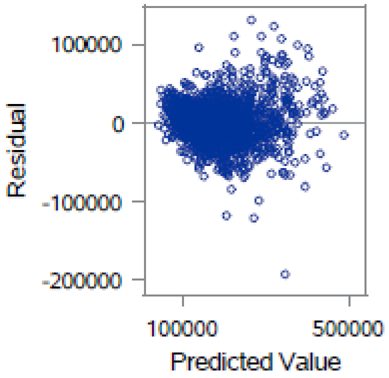


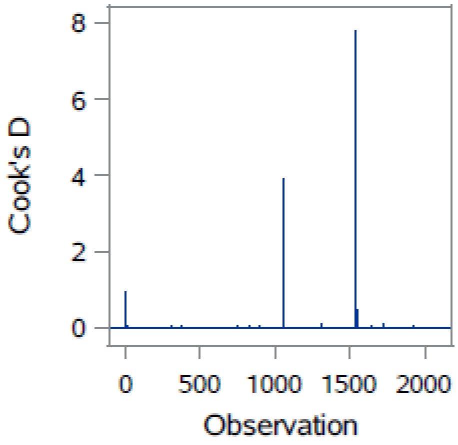
Root MSE	32172	R-Square	0.8322
Dependent Mean	183423	Adj R-Sq	0.8309
Coeff Var	17.53989		

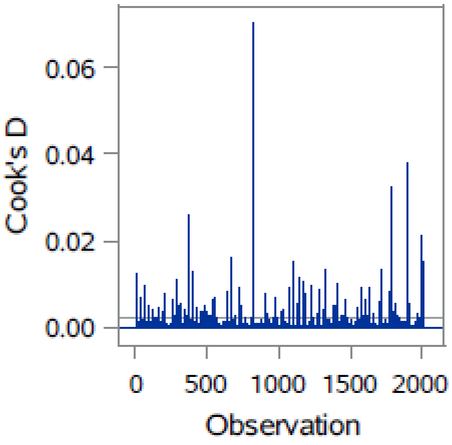


Root MSE	24437	R-Square	0.8906
Dependent Mean	181620	Adj R-Sq	0.8897
Coeff Var	13.45507		









Key Observations:

The model gives us an RMS value of 24437 that implies that the there is an average error of the said amount when the model tries to predict the sales price.

One key observation from the residual plot is that the variance in residuals does not follow a specific order i.e., increasing, decreasing or constant, the variance of residual increase in between with lowers values at both ends, this could be as there is much more data for those middle sales price values.

The Normal plot for residuals we get supports our assumption of the residuals being in normal distribution.

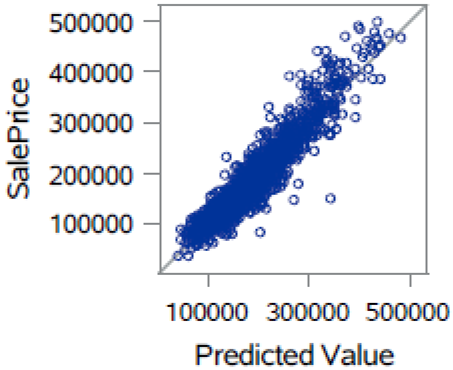
Our Assumption on the independence of the residuals holds true as the sales price itself is an independent variable and so the same hold true for the residuals.

Conclusion:

Q. What is the expected selling price of my home?

We found that the model explains 82.9% of the variation and upon running the test data through the model the predicted values had an average difference of approximately 11% that can be translated as the model predicts the price with +-11% accuracy. Going by the report we have an average error of $24437, This can be considered high but does provide us with the business ballpark number. Please find the list of actual and predicted sales price above.

The variation or asymmetry of the residual could be due to presence of outliers that shifts the symmetry. The effect of each individual variable can be seen in the modelling report above. The distribution of the residual is also approximately normal that can be seen in the model report as well. The predicted vs actual sales price graph is shown below:



Based on the model we can conclude that multiple regression model can be used to predict the sales price on the houses if we already know the quantitative factors for that house

Improvement:

Considering only the quantitative factors could limit the application of the model.

We have taken this approach to evade level heavy variables such as Neighborhood which has 28 levels or building materials. But the above created model can be improved to provide more localized or house specific results by including the categorical filters. For example, using the above regression model for only houses in North Ames Neighborhood with 1 story house.

2. Cluster Method:

a. Approach:

Rather than giving a pinpoint number for a house based on the quantitative numbers, we thought it would be beneficial for the customer to have a range of estimated price so that the customer can individually gauge the price for His/her property. Therefore, we used the Ames data to create different clusters or groups of sales prices and are trying to predict which cluster the house of our customer falls into.

b. Modelling:

This modelling is based on the idea that the house of similar characteristics will fetch similar sale price in the market that is a general assumption (there might be some factors affecting this rule such as the market situation or the year the house is being sold. For example, a house might get difference price per pandemic or recession and a different price post such events). We classified every 11 groups which equally distribute the population according to the sale price distribution. The data was then used in SAS to create the cluster group with the help of the below characteristics.

[Lot_Frontage Order Lot_Area Year_Built BsmtFin_SF_1 Year_Remod_Add BsmtFin_SF_2 Bsmt_Full_Bath Bsmt_Unf_SF Bsmt_Half_Bath Total_Bsmt_SF Full_Bath _1st_Flr_SF Half_Bath _2nd_Flr_SF Bedroom_AbvGr Low_Qual_Fin_SF Kitchen_AbvGr Gr_Liv_Area TotRms_AbvGrd Garage_Area Fireplaces Wood_Deck_SF Garage_Yr_Blt Open_Porch_SF Garage_Cars Enclosed_Porch Mo_Sold _3Ssn_Porch Yr_Sold Screen_Porch Misc_Val]

Please find the SAS code used in the modelling in the file below:

cluster.sas

Here we can see that there are multiple misclassifications and almost no classification in the groups 10 and 11. The calculated eigen values and cluster results are present in the below file.

cluster-results (1).pdf

We created 15 Cluster groups and found the below classification

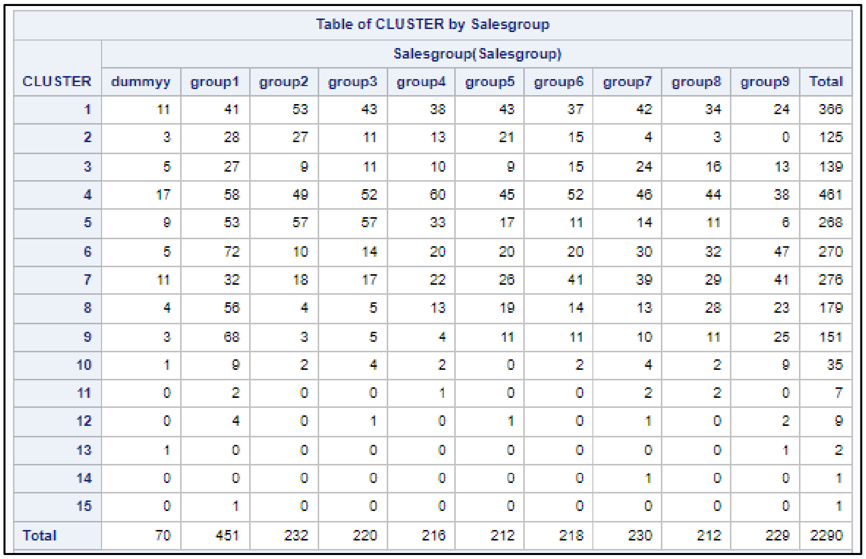


Table of CLUSTER by Salesgroup

	Salesgroup(Salesgroup)										
CLUSTER	dummyy	group1	group2	group3	group4	group5	group6	group7	group8	group9	Total
1	11	41	53	43	38	43	37	42	34	24	366
2	3	28	27	11	13	21	15	4	3	0	125
3	5	27	9	11	10	9	15	24	16	13	139
4	17	58	49	52	60	45	52	46	44	38	461
5	9	53	57	57	33	17	11	14	11	6	268
6	5	72	10	14	20	20	20	30	32	47	270
7	11	32	18	17	22	26	41	39	29	41	276
8	4	56	4	5	13	19	14	13	28	23	179
9	3	68	3	5	4	11	11	10	11	25	151
10	1	9	2	4	2	0	2	4	2	9	35
11	0	2	0	0	1	0	0	2	2	0	7
12	0	4	0	1	0	1	0	1	0	2	9
13	1	0	0	0	0	0	0	0	0	1	2
14	0	0	0	0	0	0	0	1	0	0	1
15	0	1	0	0	0	0	0	0	0	0	1
Total	70	451	232	220	216	212	218	230	212	229	2290

Conclusion:

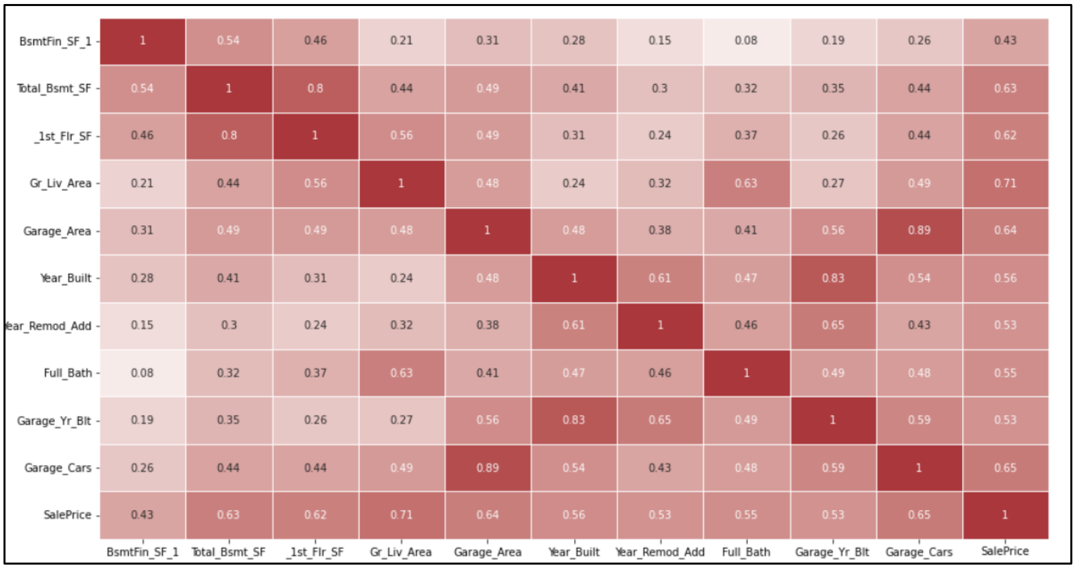
The randomness of the formed cluster and a huge number of misclassifications leads us to believe that the quantitative characteristics that we used do not have the power to clearly distinguish between the sales group based on the price. We however found that increasing the cluster to a larger number (350+) provides us with a much clearer classification but the, but this defeated the purpose of the classification and might end up in comparing a customer’s house to the exact house with same characteristics from the data.

Improvement:

This model can also be improved by factoring in the qualitative feature such as building story, condition or neighborhood. This can be done by training the model with only limited observation with the required characteristics as a filter.

3. Correlation matrix:

In order to estimate the influence of variables we are to the correlation approach to see which predictors are more corelated to the dependent variable. The values in the cells represent the strength of the relationship. You can use correlation heatmaps to identify possible links between variables and to gauge how strong these relationships are. We have cleaned the data by removing null values from the dataset. Null values have been replaced by the average of rest of the data in the specific column.



We have performed a correlation analysis on the numerical data from the given dataset. Upon analysis, we found that the following variables *[BsmtFin_SF_1, Total_Bsmt_SF, _1st_Flr_SF, Gr_Liv_Area, Garage_Area, Year_Built, Year_Remod_Add, Full_Bath,* Garage_Yr_Blt, Garage_Cars] have a high correlation value compared to other variables.

Q. *Which factors are more important than others?*

Through this, we can conclude that the above-mentioned variables affect the sales price more compared to the other variables in the dataset.

4. Data Life Cycle:

The data analysis process is a set of steps required to make sense of the available data. However, each step is equally important to ensure that the data is analyzed correctly and provides valuable and actionable insights. Let's look at the five essential steps that make up the data analysis process.

Here the problem statement is to help customers to predict the selling price of their homes also help them to help increase their selling prices by evaluating certain parameters. The ways to increase the selling price can be renovating the house or some conditions can be improved if and if the renovating cost is not greater than selling price. Factors such as *garage type/condition, completed floors, conditioning (central air), interior, electrical, heating, fireplace, etc.*

We can also provide an estimated increase require for these factors by clustering according to neighborhood and calculating average selling price. The analysis is required to correlate these factors with selling price and identify how can they be used to predict the correct selling price along with that the percentage difference between given selling price and predicted selling price can be calculated by regression model.

80% of the data scientist's time is spent on cleaning data than generating insights. We must identify and eliminate duplicate data, anomalies, and other inconsistencies that can skew analysis to produce accurate results. Performing data analysis is one of the last phases, one can do it through data mining. Data mining techniques such as clustering analysis, anomaly detection, and association rule mining can reveal hidden patterns in data that were previously invisible. We applied predictive analytics, here the selling price is affected by the month of the year in which you are selling the house. Also, how old houses will be sold in future can be predicted by corelating the selling price with year built and month sold variables.

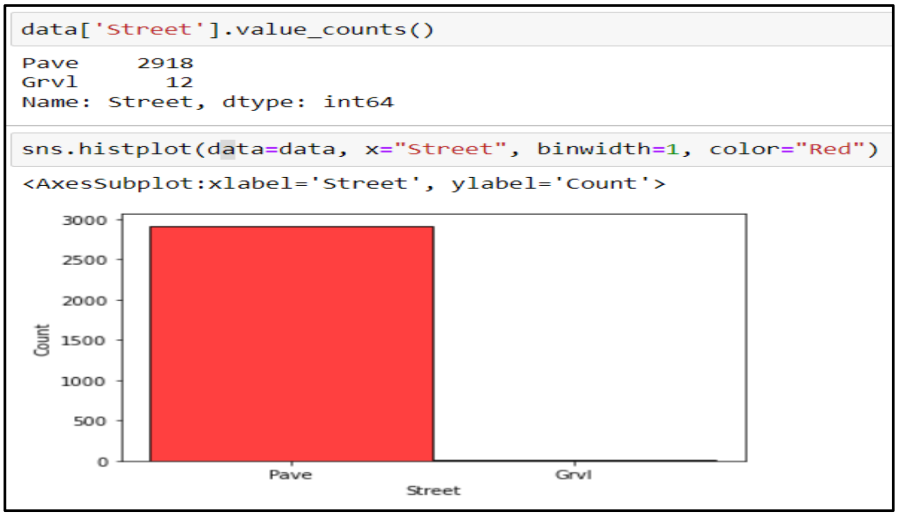
The final step is to interpret the results of the data analysis. This section is essential because this is how a business will gain real value from the previous four steps. Interpreting the results of data analysis will validate why you are doing it, even if it is not 100% conclusive. Modelling techniques used are multiple regression analysis analyze data by modeling the change in one variable caused by another are used in here., cluster analysis, correlation matrix along with techniques explained are imbalanced data, kurtosis and data life cycle.

Q. Which homes should I compare my house to?

As explained above, clustering the houses in same neighborhood and determining the similar characteristics like floor, area, utilities, area, etc. can be used to compare the houses.

5. Imbalanced Data:

The dataset provided have imbalance data for many variables. The variable "Street" has two types of observations namely Pave and Grvl. The Fig showed that this variable is highly imbalance.



```
data['Street'].value_counts()
```

```
Pave     2918
Grvl       12
Name: Street, dtype: int64
```

```
sns.histplot(data=data, x="Street", binwidth=1, color="Red")
```

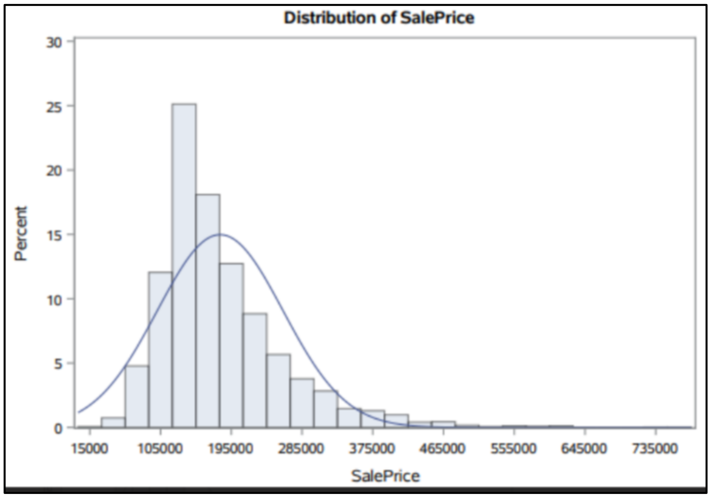
```
<AxesSubplot:xlabel='Street', ylabel='Count'>
```

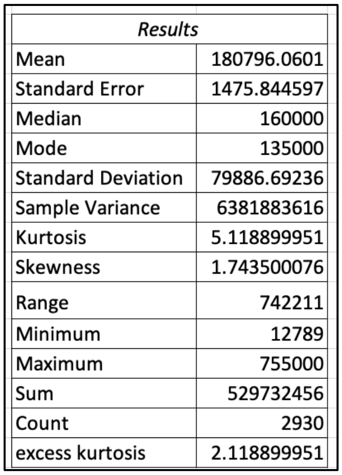
Here we are using the variable Street as dependent variable. So, when we use this data set to make a prediction model and try to find accuracy of the model there is high chance that the model chooses majority data as true and predict all the data points as Pave. So, if models' accuracy we will be getting is nearly 99% but it may not be able to predict a single Grvl variable correctly. Here the Pave is majority class with count 2918 and Grvl is minority class with count 12 in the dataset.

To avoid this type of problem we used resampling technique to get balance dataset. We use python programming for analysis. In that we have use pandas, seaborn, sklearn, etc libraries. There are many techniques like resampling, Synthetic Minority Oversampling Technique (SMOTE), Balanced Bagging Classifier, Threshold moving etc., for imbalance dataset. In resampling there are two methods one is oversampling, where we usually over sample the minority class and another one is under sampling, where the majority class is under sampled. But here if we use under sampling, we get total counts in the dataset will be only 24, which is very small sample size to build a model and further analysis. So, we used over sampling technique. Here we can use any number to over sample the minority class. But both the majority and minority class equal we oversample minority class with 2918 counts. So, in the balance dataset we get both the class in same frequency.

6. **Kurtosis:**

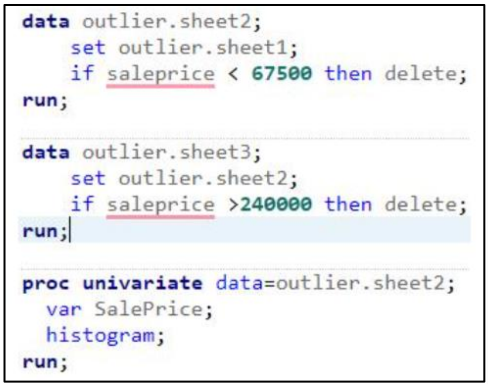
Kurtosis can be used to detect outliers in our data. It provides the overall level of outliers that are present. In order to understand the distribution of the dependent variable we run various descriptive measures like mean, median, mode, kurtosis, skewness, range, standard error.





Results	
Mean	180796.0601
Standard Error	1475.844597
Median	160000
Mode	135000
Standard Deviation	79886.69236
Sample Variance	6381883616
Kurtosis	5.118899951
Skewness	1.743500076
Range	742211
Minimum	12789
Maximum	755000
Sum	529732456
Count	2930
excess kurtosis	2.118899951

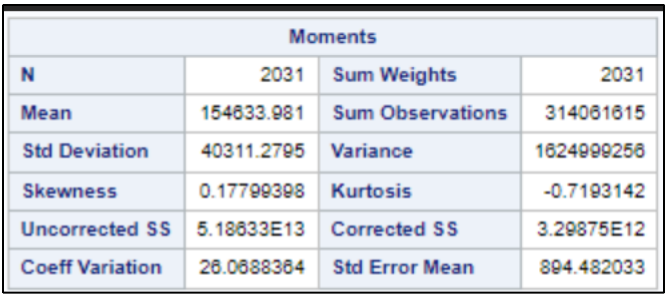
From the results we can say that our data is highly skewed, by using graph we can infer that our data is high positive skewed with long right tail and, we can conclude that it is a Leptokurtic high peak as the degree of peak is greater than 3. The outliers are causing the skewness and hence we need to handle these outliers. The outliers are eliminated using criteria,



```
data outlier.sheet2;
    set outlier.sheet1;
    if saleprice < 67500 then delete;
run;

data outlier.sheet3;
    set outlier.sheet2;
    if saleprice >240000 then delete;
run;

proc univariate data=outlier.sheet2;
  var SalePrice;
  histogram;
run;
```



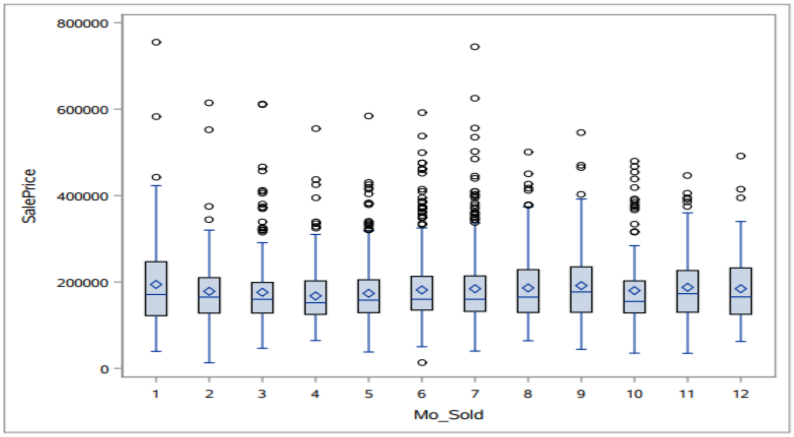
Moments			
N	2031	Sum Weights	2031
Mean	154633.981	Sum Observations	314061615
Std Deviation	40311.2795	Variance	1624999256
Skewness	0.17799398	Kurtosis	-0.7193142
Uncorrected SS	5.18633E13	Corrected SS	3.29875E12
Coeff Variation	26.0688364	Std Error Mean	894.482033

The snippet shows that skewness is reduced but we are losing 30% of the observations hence we decided not to eliminate all the outliers but only the extreme ones.

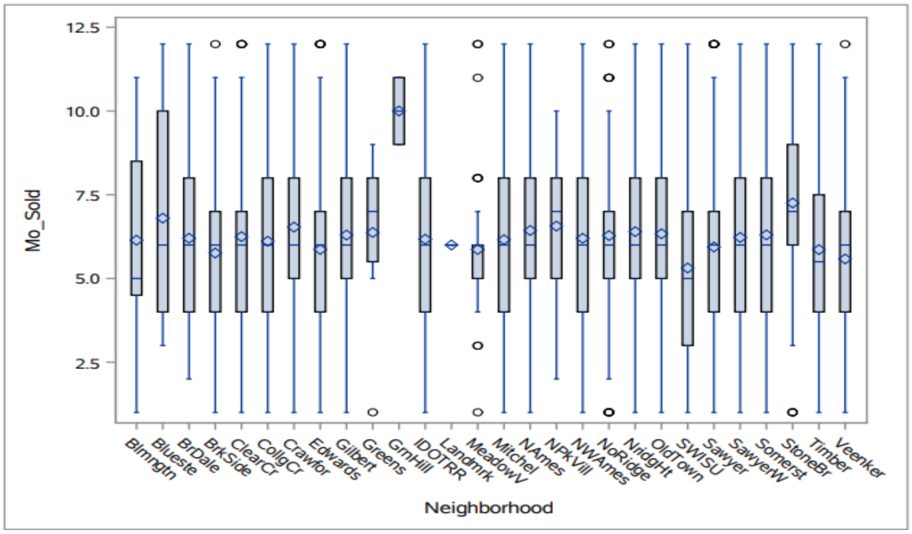
Recommendations:

Q. When is the best time of the year to sell my home?

According to our data please see the box plot below:



We can see that there is a very slight change in the average sales price based on the month in which the sale is made. This shows that the sale price is not favored by any specific month. However even if the change is very less months 1, 6, 7, 9 have the highest sale sales price average and also , the neighborhoods that have almost the 40 % of overall sales like *Names, Edwards, CollgeCr, Old town* have more sales in 6 – 8 months , you can see the box plot below, This concludes that June, July and August are the months one needs to sell to get the best sales price , and this conclusion is be well supported in the data as these moths have the highest number of houses sold. (*mo_sold* is a discrete variable but inference can be taken from the box plot even if non integer months have no observations.)



Q. How much should I invest in improving the condition of my home to increase the expected price by more than the cost of improvements?

We for this question take into consideration the role of overall quality of a house as a major deciding factor of sales price, we work under the assumption that to increase the overall quality of the house by ‘one’ unit the owner must invest approximately 5 % of the average selling price of the house in of that neighborhood.

For an example: A house with the overall quality scores as ‘6’ in *NAmes* neighborhood (avg of *NAmes* is 145007) yields an average sales price of $157000 to increase the overall quality score to a ‘7’ the owner needs to invest 10% of 145007 i.e., $14500 that basically ups the house price as 157000+14500 = 161500 with a score of ‘7’ and a score of 7 yields an average of $192000. This assumption can be used to calculate if the investment is worth having or not. The above criteria are very specific and might not be always correct as there are assumption involved, on a more general note in order to increase the house value the factors influencing the sales prices should be changes to the better therefore addition of features such as pool, fireplace, garage or renovation of certain things like kitchen and severe damages should help in drastically improving the selling price.

Group 1: Team Members

Prachi Manoj Mahajan	1002078060
Mehul Anupam Hivlekar	1002046379
Visesh Sagar Veeramraju	1002057402
Mohan Surya Varma Addala	1002028636
Anurag Jadhav	1002094933
Chaitanya Krishna Kuppuru	1002070545