AMES HOUSE SALE PRICE PREDICTION

BSTAT 5325 – FINAL PROJECT REPORT

Team Members

Visrutha Balashanmugam
Jay Vikas Warke
Shashank Chhajed
Kamalkumar Kannan
Venkata Satya Bollina
Tu Vi Hoang Le

Modelling code link: https://drive.google.com/file/d/1TE-VQEbGtyAiNWlYchs73A9zKyJWnL65/view?usp=sharing
EDA code link: https://colab.research.google.com/drive/1TtcIfPV7neraA9H5RhiaOYXv8upEF7Vx?usp=sharing

BUSINESS PROBLEM

Abstract

In today's world of fluctuating markets and prices, it is essential to know the peak and downfall periods of a commodity or an asset. Real estate is one of the booming areas which is and has always been in demand for years. Our client is one of the major real estate companies called AMES REAL ESTATE ASSOCIATES (AMES), who would like to build an application that would help them generate a report on the factors that affect house prices. AMES is known for its expertise and approach to helping its clients sell their houses. The main objective of this project is not only to give an insight into the prices of houses but also to generate a report that would give various suggestions to the customers in the best possible time to sell the house. AMES will also give comparable factors that would let customers know why they cannot sell their houses according to their convenience, since the market is unpredictable, and would do the best they can so that the customers can make a profit. Stakeholders such as investors, buyers, builders, architects, etc. will benefit from this application.

Business problem

For millions of Americans, real estate has been and still is the core of establishing wealth and a significant bridge in the distribution of goods, services, and income. A study shows that real estate drives the US economy by contributing up to 17% of the GDP. Numerous factors affect house sales in the market which include: availability of the house, location, neighborhood comps, house size and usable space, conditioning, etc. One of the most significant factors is the price of the house. The price of the house sold is a deal breaker, every house can be sold when priced properly.

Ames Real Estate Associates (AREA) is a successful real estate company based in Ames, Iowa that helps its clients by providing them with information on the price range of the house in the market as well as the qualities of the house that should be enhanced to increase its selling price. The key to any business's success is figuring out its customers' needs and anticipating when they will make a purchase. As data analysts of a consulting firm, it is our responsibility to help AREA automate this process by building data models to predict the trends and patterns to derive useful information to maximize the price of the houses sold.

The main objective of this project is not only to provide a predictive model for houses' prices but also to generate a report that would offer additional insights on various questions that could be helpful for the customers. The dataset that is used for this purpose is the housing dataset in Ames that describes the property sales in Iowa from 2006-2010 consisting of several variables that establish the residential sale prices. The dataset contains about 2930 records of 82 distinctive features of the house. In this project, we will analyze the real-world dataset and provide AREA with a report on how the properties can be priced based on the features of the house versus its comparable properties. With this project, AREA will be able to provide customers with comparable reasons and evidence on the best possible selling prices for them.

Constraints

- The report is to be completed and turned in on December 6th, 2022.
- A team of six analysts/scientists (students) will be working together on the project
- The model is constructed using quantitative data and one or two qualitative data have been taken into consideration based on the model.
- Any improvement recommended should not exceed more than 5% of the sales price of the property.

Assumptions

- The data were collected and entered in the excel sheet accurately.
- This model is being constructed right after the data is collected. So we assume that we are in 2011.

- The assumptions for linear regression are true in this model. Linear regression assumptions – constant variant assumption, constant assumption, independence assumption, linear relationship between variables.

Limitations

- The data that is taken into consideration has been limited to one-story and two-story houses present in Ames. The other house styles have been removed.
- There are multiple home features/variables with empty cells such as garage year built, lot frontage, etc. which have been removed based on their influence on the dataset.
- There are a large number of variables in the dataset (82) which includes numerical, categorical, ordinal, and nominal variables. Only some variables are considered based on the model and other have been removed because of the time limit.

Conditions

- The results of this report should only be applied to houses in Ames, Iowa since the dataset included houses in Ames, Iowa
- The result and model do not take into consideration any of the external factors that affect the house price such as great recession in 2008. Hence the model is not suitable for predicting prices in 2022.

Operational definitions

- Sale price is the price that a buyer pays for the house which will include the realtor commission and tax. Hence the amount that is given to the owner will be less than the value presented.
- Overall quality of the house considers all the physical conditions and dwellings of the house such as the material used, replacement of mechanical systems, maintenance, etc., and is rated by a professional real estate appraiser.

SMART objectives

Specific	Who: Real estate organization, management, project team What: House sale price prediction
Measurable	The data must be prepared and the analysis has to be done to predict the best selling price. This is documented by the team members and submitted to the management.
Attainable	Justification: We have enough data that can be analyzed and the sale price can be predicted. The project members and resources present will help to achieve this objective.
Relevant	The data is relevant and it can modified to fulfill the objective of the project before the time frame.
Time Bound	The deadline is 6th December. A working model should be completed by the end of 3 months.

Research Questions

The main goal of this project is to predict the price of the properties in AMES and to understand the effect of the various features on the same. Some of the questions that will help us to achieve the goal are as follows:

1. What is the expected sale price of a property in Ames?
2. What features in the data mainly affect the price of houses in Ames, Iowa?
3. Which of the features that help to predict the house price are more important in comparison to the others?
4. How can one improve the sales price of the house in Ames? What features need to be improved so that the expected price can be increased?
5. What are the homes that I can compare my house to, so I can improve the sale price? When is the best time to sell the house in Ames?

DATA PREPARATION AND UNDERSTANDING

The AMES Housing dataset describes the housing properties that is present in Ames, Iowa. The dataset reports information on property sales in Iowa from 2006-2010 consisting of several variables that establish the residential sale price. The dataset contains about 2930 records of 82 distinctive features of the house. The features of the dataset include the physical property measurements in addition to computation factors utilized in the assessment of properties by realtors in Ames. They included a mixture of nominal, ordinal, continuous, and discrete variables used in the calculation of assessed values.

Discrete variables: Around 15 discrete variables are present that represents the frequency of the units that are present in the house such as number of bedrooms, kitchen, bathrooms, etc. It also contains data on the number of cars that could be fit in the garage as well as the date of remodelling (if done) of the house.

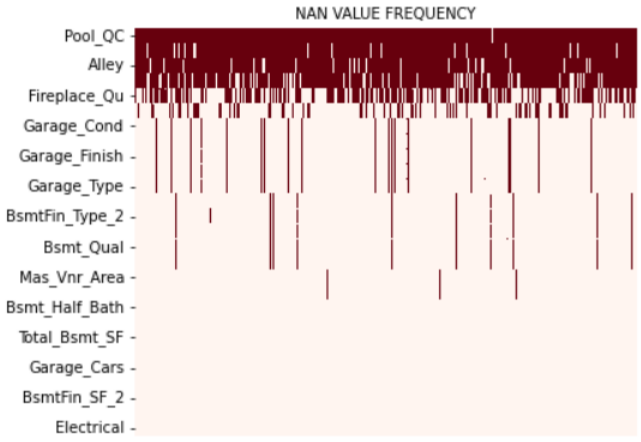
Categorical variables: When we look at the data, it can be observed that around 20 continuous variables are related to the dimensions of the properties. Apart from the most common dimension features like Total lot area, Garage Area, it also includes other area measurements area of the main living room, total basement area and also about the porch area which is further categorised based on its types. The categorical data contain a wide range of data, with a minimum of 2 to a maximum of 28 classes which are STREET (gravel/paved) and NEIGHBOURHOOD ('NAmes', 'Gilbert', 'StoneBr', 'NWAmes', etc.) respectively. The nominal data in the dataset express details on the garage properties, materials used and other external conditions. The ordinal data is used to rate the modules of the property.

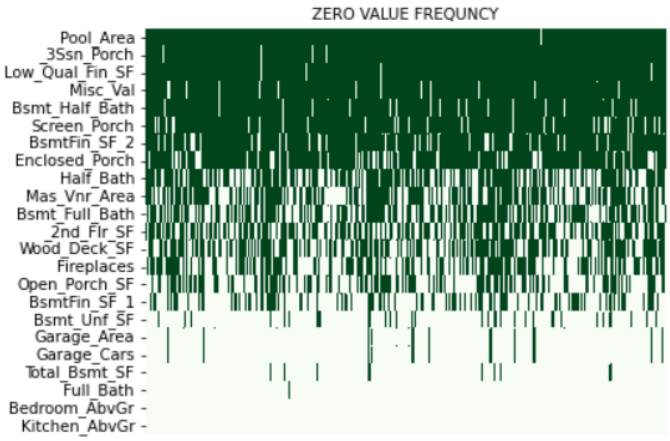
Learning about Null Values:

Based on the description of data, it can observed in the median values that features such as 'Enclosed Porch', 'Pool Area', 'Fence', 'Pool QC' have a large number of NAN/zero values. A large number of missing values can also be seen in variables such as 'Garage_Yr_Built' (114 values missing) and Lot_frontage (330 values missing).

Hence we are visualising the zero/NAN values with the help of heat-maps. By comparing the heat-maps, we can see that some trends

FEATURES	NAN/ZERO COUNT
Pool_QC	2042 (NAN)
Pool_Area	2042 (0)
Misc_Feature	1986 (NAN)
Misc_Val	1986 (0)

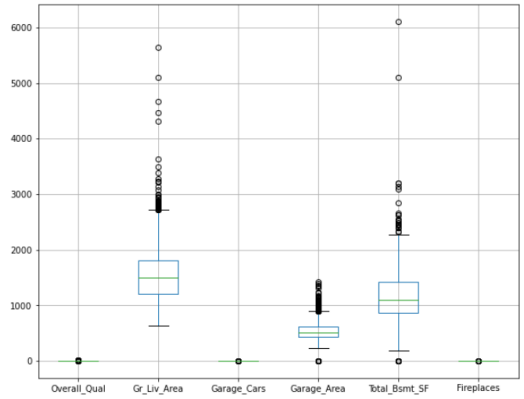




From this we can come to a conclusion that there is consistency between Null values in ordinal variables and zero in discrete variables. In this project, the null values have been removed from the dataset. The ordinal variables are given numbers based on the ranking of the attributes.

Outlier Analysis and Elimination

The outlier of the data is any observation that is abnormal and is significantly different from the other values. In the outlier analysis we first find them using box blots and IQR and to tackle them the outliers in data were converted into null values and the null values have been removed from the dataset.




```
In [88]:
for x in ['Gr_Liv_Area']:
    q75,q25 = np.percentile(data.loc[:,x],[75,25])
    intr_qr = q75-q25

    max = q75+(1.5*intr_qr)
    min = q25-(1.5*intr_qr)

    data.loc[data[x] < min,x] = np.nan
    data.loc[data[x] > max,x] = np.nan
```

```
In [90]:
data['Gr_Liv_Area'].isnull().sum()
```

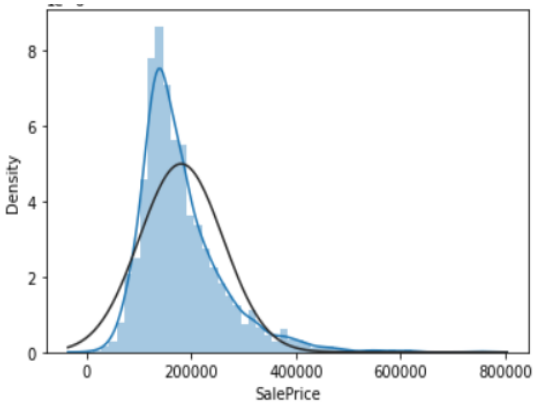
```
Out[90]: 47
```

```
In [91]:
data = data.dropna(axis = 0)undefined
```

```
In [92]:
data['Gr_Liv_Area'].isnull().sum()
```

```
Out[92]: 0
```

DATA EXPLORATION



Distribution of the target variable: '`SalePrice`' - **Skewness and Kurtosis**

Skewness: It is a measure of the lack of symmetry in the distribution of data. The data is said to be symmetric when both the right and left part of the distribution looks same from the mid-point.

Kurtosis: It is the measure of if the tail of the distribution curve is light-tailed or heavy tailed relative to the normal distribution. High kurtosis is indicated by the heavy tails on the distribution and low kurtosis will tend to have a thinner tail.

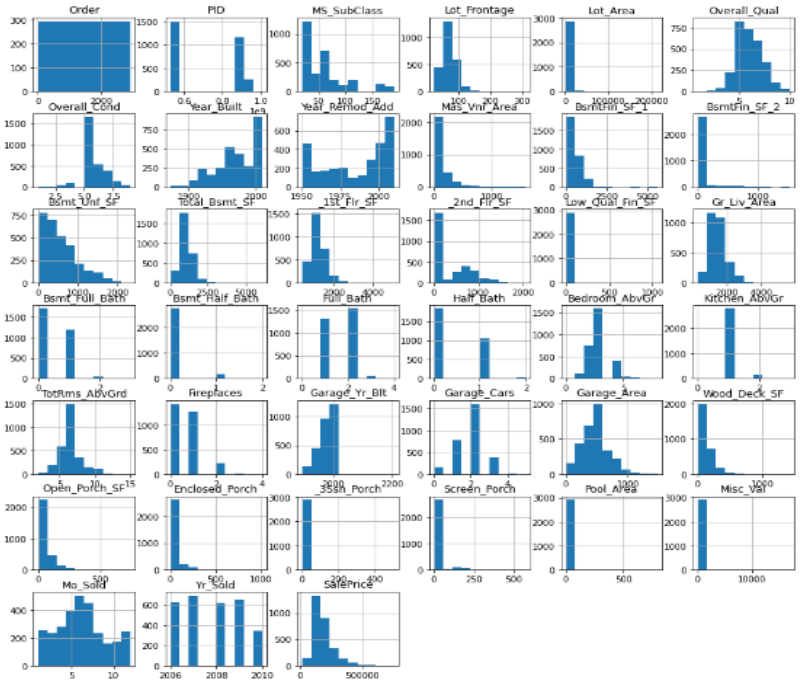
Histogram is one of the most effective techniques to present both knew and kurtosis.
Observation from the histogram:

- It can be observed from the distribution on the left is that the target variable is not normally distributed. It is skewed to the right. In order to normalize this distribution, we can consider logging the target variable.
- The value of kurtosis for the target variable is found to be `5.118899951130896`. As seen in the histogram, a variable with a high kurtosis has a thin bell shaped curve with a sharp peak and heavy tails. This means that the data will be seen more on the distribution tails and values are closer to the mean of the data.

```
count      2930.000000
mean     180796.060068
std       79886.692357
min       12789.000000
25%      129500.000000
50%      160000.000000
75%      213500.000000
max      755000.000000
```

Distribution of Numerical variables

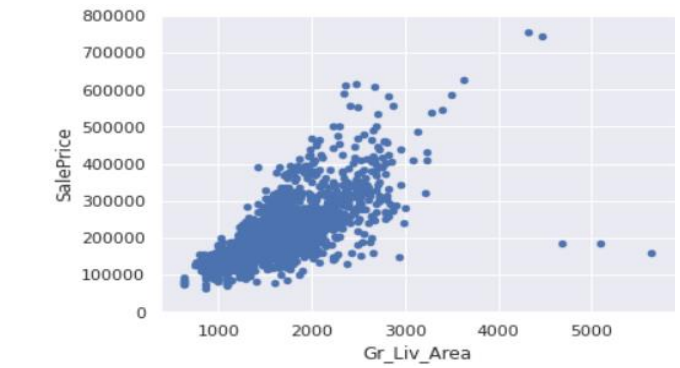
The histogram plot for all the numerical data can be used to check if the distribution is normal or not. As observed, most of the numeric features do not have a normal distribution. Yr_sold and Mo_Sold do not have any relationship with the SalePrice. There are a lot of extreme values present in come of the variable such as Gr_Liv_Area. There is very less variance in data such as Pool_Area, Misc_Val, Porch, etc.

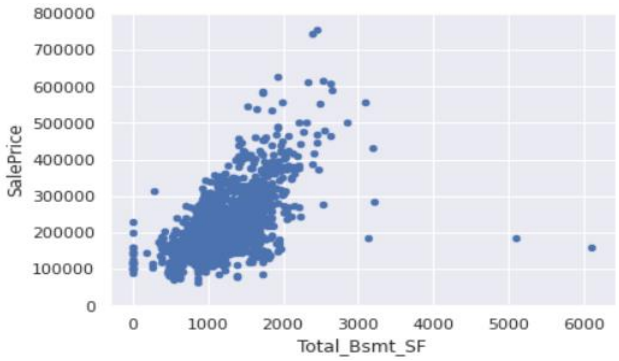


Based on correlation analysis (explained in detail in the later section) of all the variables in the dataset, high correlation can be seen in the following

```
['Overall_Qual', 'Total_Bsmt_SF',
'_1st_Flr_SF', 'Gr_Liv_Area',
'Garage_Cars', 'Garage_Area',
'SalePrice']
```

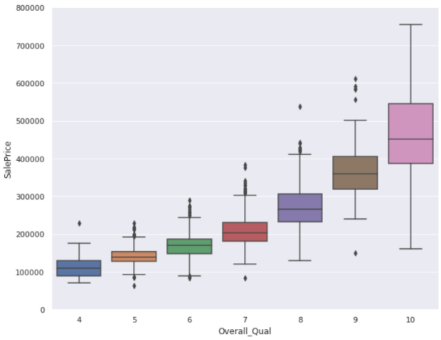
EDA with highly correlated numeric variable: ‘GrLivArea’ and ‘TotalBsmtSF’:





It is evident that GrLivArea andSalePrice are linearly related in a positive manner. Hence it can be concluded that if one increases, so does the other. Similarly, SalePrice is linearly related to TotalBsmtSF as well but it has a higher slope. Hence the SalePrice increase with increase in TotalBsmtSF but not in a linear manner.

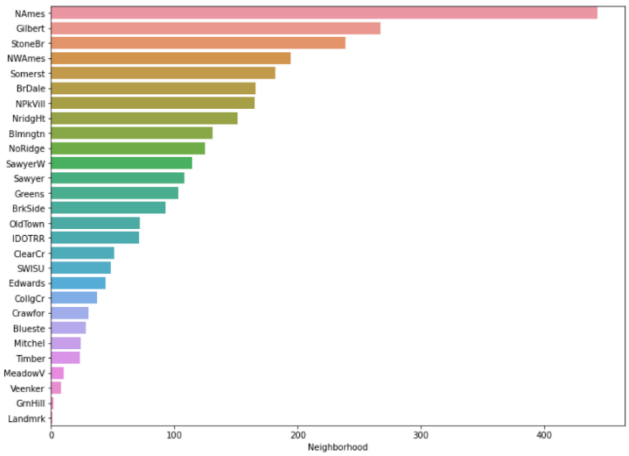
EDA with highly correlated categorical variable: ‘OverallQual’:



It is evident from the box plot that SalePrice increases with increase in OverallQual.

Imbalanced data

Imbalanced data is caused by the difference in the distribution of classes in the data. It usually occurs when the elements of once class is greater than the other elements. This means that the model will be biased towards that class of the variable. This might be a cause for Type-II error. One such variable in this dataset is ‘Neighborhood’. This feature has 28 classes denoting the various areas present in Ames. In this the label with the highest count is ‘Names’ with 443 and the lowest is ‘Landmrk’ with just 1 element.



	Count Neighborhood	Percentage Neighborhood	Average SalePrice
NAmes	443	15.119454	145097.349887
CollgCr	267	9.112628	201803.434457
OldTown	239	8.156997	123991.891213
Edwards	194	6.621160	130843.381443
Somerst	182	6.211604	229707.324176
NridgHt	166	5.665529	322018.265060
Gilbert	165	5.631399	190646.575758
Sawyer	151	5.153584	136751.152318
NWAmes	131	4.470990	188406.908397
SawyerW	125	4.266212	184070.184000
Mitchel	114	3.890785	162226.631579
BrkSide	108	3.686007	124756.250000
Crawfor	103	3.515358	207550.834951
IDOTRR	93	3.174061	103752.903226

Timber	72	2.457338	246599.541667
NoRidge	71	2.423208	330319.126761
StoneBr	51	1.740614	324229.196078
SWISU	48	1.638225	135071.937500
ClearCr	44	1.501706	208662.090909
MeadowV	37	1.262799	95756.486486
BrDale	30	1.023891	105608.333333
Blmngtn	28	0.955631	196661.678571
Veenker	24	0.819113	248314.583333
NPkVill	23	0.784983	140710.869565
Blueste	10	0.341297	143590.000000
Greens	8	0.273038	193531.250000
GrnHill	2	0.068259	280000.000000
Landmrk	1	0.034130	137000.000000

ANALYSIS

For analysis, the processed data is split into train and test data on which the model is built on. Here, we have taken a split ratio of 70/30 – 70% for train data and 30% for test data respectively. The train set is used to train the model and the test data is used to evaluate the model fit.

```
1 X_train, X_test,y_train,y_test = train_test_split(train,test,test_size=0.3,random_state=2)
```

Linear-Regression:

Linear regression is a supervised learning approach used to predict the value of the dependent variable based on the values of the independent variable. In this method, a linear equation is found that best describes the correlation of the independent variable with the dependent variable by fitting a straight line. In this method, quantitative variables are used to predict the target variable: 'SalePrice'. The quantitative predictor variables that are used in this model are: 'Lot_Frontage','Lot_Area', 'House_Style','Overall_Qual','Year_Built','Year_Remod_Add','Mas_Vnr_Area','BsmtFin_SF_1','Bsmt_Unf_SF','Total_Bsmt_SF','_1st_Flr_SF','_2nd_Flr_SF','Gr_Liv_Area','Bsmt_Full_Bath','Full_Bath','Half_Bath','Bedroom_AbvGr','TotRms_AbvGrd','Fireplaces','Garage_Yr_Blt','Garage_Cars','Garage_Area','Wood_Deck_SF','Open_Porch_SF','Enclosed_Porch','Screen_Porch', 'Pool_Area'

Implementation and Results: The model was trained multiple times to obtain the best score. The accuracy of the model after training turned out to be 83%. The model is not over-fitted and has a good prediction rate.

Actual VS Predicted Value Plot

```
#MODEL SCORE
print("The model score is:")
regr.score(X_test, y_test)
```

```
The model score is:
0.8302359269818637
```

Prediction Interval

```
In [11]:
def get_prediction_interval(prediction, y_test, test_predictions, pi=.95):
#get standard deviation of y_test
    sum_errs = np.sum((y_test - test_predictions)**2)
    stdev = np.sqrt(1 / (len(y_test) - 2) * sum_errs)
#get interval from standard deviation
    one_minus_pi = 1 - pi
    ppf_lookup = 1 - (one_minus_pi / 2)
    z_score = stats.norm.ppf(ppf_lookup)
    interval = z_score * stdev
#generate prediction interval lower and upper bound
    lower, upper = prediction - interval, prediction + interval
    return lower, prediction, upper
get_prediction_interval(predictions[0], y_test, predictions)

Out[11]: (248391.4575530908, 317636.8553547489, 386882.25315640704)
```

The prediction interval has been calculate by using standard deviation. First we calculate the standard deviation of the target test variable and calculate the interval from the standard deviation. Using this interval, the upper and lower bound are calculated and tabulated. One of the interval is shown in the above figure.

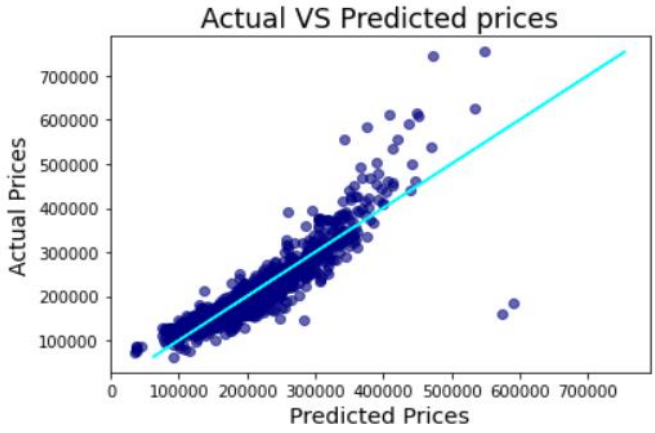
Residual Plot: The residual standard error of the model is computed. Residual values are obtained by the difference between an obtained value of the target variable and predicted value that is fit in the regression model. The fitted values are then plotted against the actual values.

```
#residual standard error
print(np.sqrt(mlr.mse_resid))

const = sm.add_constant(X_train)
model =sm.OLS(y_train, const)
result = model.fit()
result

#Residual value = observed value - fitted value
print("THE RESIDUAL VALUES ARE:")
residulas = y_train -result.fittedvalues
print(residulas)
```

```
40800.776074012894
197    -24210.110939
905    -11228.838150
549     -9872.685418
128     -1950.807510
727     20687.597558
           ...
466     19547.230653
299      3990.973343
493     -9086.846980
527    -31491.832847
1192    21381.862358
Length: 945, dtype: float64
```



Reasons for using Linear Regression:

- It is a simple model that is used to relationship between the variables and also is easy to interpret the results.
- It is used to understand about the strength of relationship between the predictors using values such as R-squared, adjusted R-squared.
- It can be used to determine what variables are influence the data better than the others.

Q. What is the expected sale price of a property in Ames?
According to the model, the selling price of the house is predicted with an 83% accuracy. Hence, the model describes about 83% of the variability in the data. Since the R-squared value is high, the model has a good predictive power. The expected selling price of the house will be approximately have a threshold of $\pm$11% of the actual price given in the data.

Correlation:

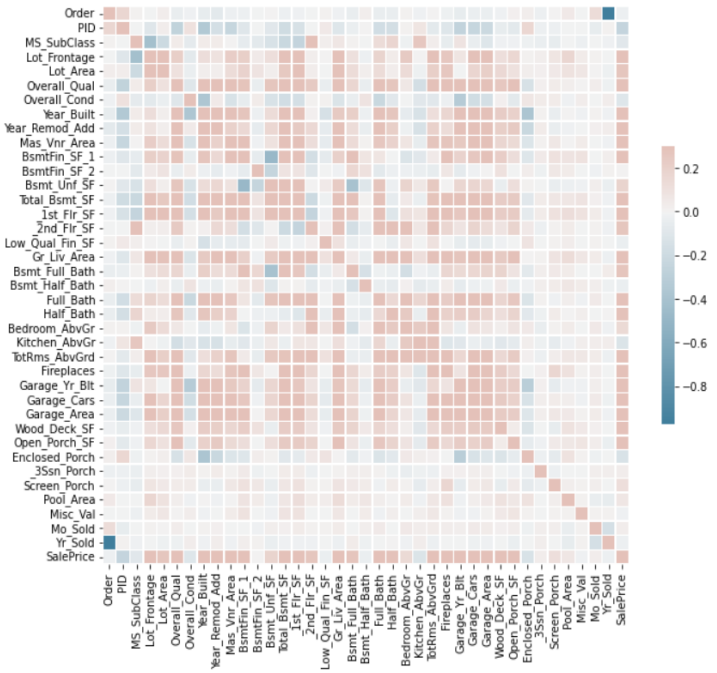
Correlation is a measure that is used to determine the relationship between any two quantitative or categorical variables. Correlation analysis is used to compute the magnitude of the relationship between two random variables. The correlation of the variables are represented in the form of a correlation matrix where the rows and columns coverages to the value that represents the correlation between the two variables. For visualization purpose, we display the correlation matrix as heat-map. Heat-maps represent correlation of two discrete variables in a monochromatic scale.

Reasons to use Correlation:

- It is one of the easiest methods to find the relationship between the variables which will, in turn, help us to predict its future trends.
- It helps to gather insights about the future and plan things ahead, in case of business decisions. It helps to disclose relationship between variables without declaring anything about its cause and effect.

Implementation and Result

The correlation between the quantitative variables are obtained and are represented in the form of a heat-map. It helps to determine if one variable will have a positive or negative effect on the other variable. The heat-map representation of the variables is shown below.



Based on the analysis, the variables with high correlation was extracted. Here, any numerical variable with a correlation greater than 0.6 was considered as highly positive correlated and less than -0.2 were considered highly negative correlated

High +VE correlation	High –VE correlation
'Overall_Qual', 'Total_Bsmt_SF', '_1st_Flr_SF', 'Gr_Liv_Area', 'Garage_Cars', 'Garage_Area.	'MS_SubClass', 'Overall_Cond', 'Low_Qual_Fin_SF', 'Bsmt_Half_Bath', 'Kitchen_AbvGr', 'Enclosed_Porch', 'Misc_Val', 'Yr_Sold'

The plot below represents the correlation heat-map of the highly correlated variables. Among the highly correlated variables, 'GrLivArea' & 'TotalBsmtSF' was quantitative variables and 'YearBuilt' and 'OverallQual' are categorical variables. All the variables have a linear relationship with the target variable: 'SalePrice'

From the heat-map, it can be observed that some of the independent variables, like 'Overall_Qual', External_Qual', are highly correlated to one another. This statistical concept is known as **multi-collinearity.** Multi-collinearity among the independent features of a dataset can be an issue since it will reduce the statistical significance of the variable and also reduces the reliability of the model.

Q. What features in the data mainly affect the price of the houses in Ames, Iowa?

The features that are highly correlated with the data will influence the target variable. According to this dataset, the influential features are:

- `'Overall_Qual','Total_Bsmt_SF','_1st_Flr_SF','Gr_Liv_Area','Garage_Cars', 'Garage_Area'`. As these features increase, 'SalePrice' also increases.
- `'MS_SubClass','Overall_Cond','Low_Qual_Fin_SF','Bsmt_Half_Bath','Kitchen_AbvGr', 'Enclosed_Porch','Misc_Val','Yr_Sold'`. As these features increase, 'SalePrice' declines.

Factor-Analysis:

One of the ways to overcome multi-collinearity is to dimensionality reduction. By introducing dimension reduction, redundant features can be removed by grouping of variables that are highly correlated among themselves and uncorrelated with other clusters. Factor analysis is a method that is used to reduce down an abundant number of variables into a smaller number of factors. In Factor Analysis, we are attempting to reduce the total features to 15 variables that influences the data model. According to this method, these 15 factors are said to have true correlation with the predictor variables. Hence, the data with original variables are transformed to data containing n-factors which, in this case, is 15.

Implementation and Results:

```
1 #FACTOR ANALYSIS - REDUCING PREDICTOR VARIABLES INTO 15 FACTORS
2 fa = FactorAnalysis(n_components=15, random_state=2)
3 transformed=fa.fit_transform(train)
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0.177982	1.709448	-0.866543	0.101443	0.308633	-1.037653	-0.419861	-0.107213	0.224711	-1.480305	-0.357834	-0.331345	-0.329432	-0.174871	1.039683
1	0.364638	0.793300	-0.533829	0.802488	-1.120627	-1.793980	-0.089915	-1.148511	-1.072991	1.898426	-0.519256	-0.193910	0.056142	-1.055885	-0.463901
2	-0.348774	-1.001630	-0.634930	0.695525	0.215091	-0.781771	-0.668549	-0.135607	-0.106407	0.005010	-0.243431	-0.172366	0.228433	-0.570731	1.004433
3	0.363799	-1.258873	-1.510683	-0.897895	-0.898060	0.076326	0.086278	0.148778	0.225672	0.548506	-0.294133	-0.086542	-0.115767	0.609475	0.193104
4	0.018914	2.143944	-0.872185	0.110781	0.789533	0.803231	-0.001263	0.038721	0.448345	0.339537	-0.668014	-0.314032	-0.362100	-0.135241	1.330168
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1346	-0.263810	1.232975	0.299237	1.144825	-0.496543	1.266633	-0.756333	0.844262	-0.127422	-1.085396	-0.431354	-0.257626	-0.347132	1.317904	-1.161675
1347	-0.235187	-1.050912	-0.539614	0.536270	0.975218	-0.894836	-0.642795	-0.183022	-0.141186	-0.840769	-0.076612	0.058668	0.440997	1.313360	-2.307413
1348	0.096768	0.101469	0.114313	1.135356	0.399880	1.633766	-0.746446	-0.027759	0.270933	-0.780153	-0.222809	-0.078204	-0.047955	1.308692	-1.186894
1349	0.200618	-0.077848	-0.289669	0.977189	0.711128	1.923767	-0.772772	0.123924	0.313903	-0.079640	-0.373426	-0.043527	-0.020894	1.245976	-1.480669
1350	1.287172	1.074335	1.003431	1.019645	-0.426487	-0.646267	2.628762	-1.682876	0.083918	-0.457115	0.313038	0.070625	0.357674	4.253018	0.005626

1351 rows × 15 columns

Applying Linear Regression on transformed data from Factor Analysis:

We do the train-test split as 70/30 and apply Linear Regression model on the data obtained from Factor Analysis. We plot the scatter plot for the actual VS predicted sales price using this factored data.

```
X_train, X_test,y_train,y_test = train_test_split(transformed,y,test_size=0.3,random_state=2)
```


```
1 regr_FA.score(X_train,y_train)
```

```
0.7312313743875155
```

```
1 regr_FA.score(X_test,y_test)
```

```
0.8004395497221262
```

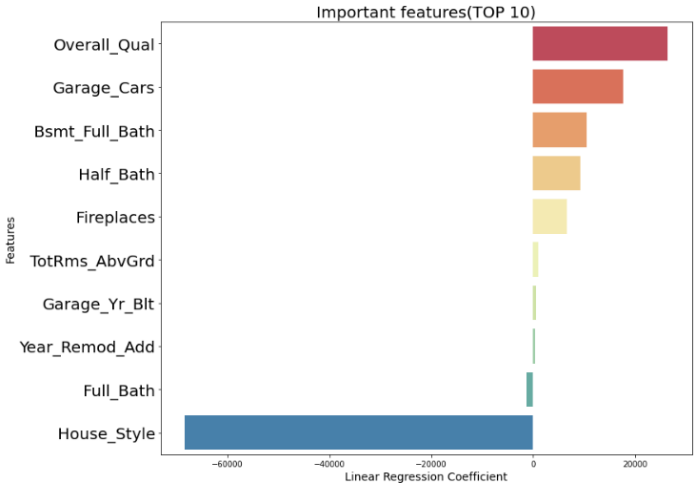
The accuracy of the model is seen to be 80%. The accuracy is calculated with the Explained variance score which is given by the formula:

$$Explained\ variance(y, y') = 1 - \frac{Var(y - y')}{Var(y)}$$

where y is the actual target output, y' is the predicted target output. $Var(y - y')$ and $Var(y)$ are respectively, the variance of the predictor errors and actual values. Although the score is 0.10 lesser than the Linear Regression score before Factor Analysis, it is highly desirable since 0.800 is closer to 1.0, indicating better squares of standard deviation errors.

Feature Importance

It is the process of assigning score to the features in the data that help to predict the target variable through the model. In this case, the input features include 'Overall_Qual', 'TotalBsmtSF', 'Garage_Area','Garage_Cars', etc. and scores are given to these variable relative to the prediction of the 'SalePrice' variable. In linear regression, the feature importance is given based on the `coeff_` value of regression. A plot has been made for the top 10 features that play an important role in predicting the sale price.



Q. Which of the features that help to predict the house price are more important in comparison to the others?

According to this dataset for the linear regression model, the 10 most important features include

- Overall_Qual
- Garage_Cars
- Bsmt_Full_Bath
- Half_Bath
- Fireplaces
- TotRms_AbvGed
- Garage_Yr_Blt
- Year_Remod_Add
- Full_Bath
- House_Style

The first 5 increase the price of the house and the last price lowers the price of the house.

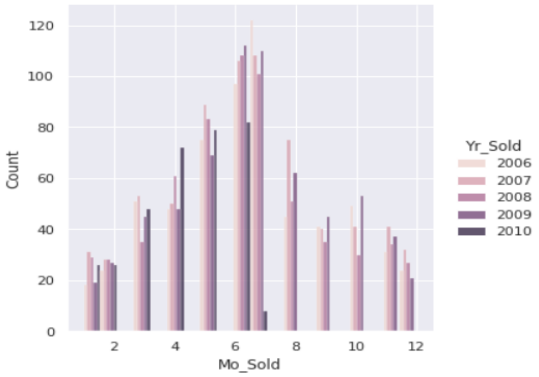
Data life Cycle

- **Problem definition:** We begin the cycle of this project by going through the problem definition and bringing up points that the clients would like as an outcome of this project. The data and its dictionary were compared to learn about the data and bring about descriptive analytics of the data.
- **Data Preparation:** In this phase, the data searched through for null values and single values. If the data has null values, we simply discard them. The outlier in the data is removed using the IQR and the points that lie below the low-bound are converted to null and are then discarded.
- **Analysis:** Correlation finds the strength of relationship between 2 variables. Using this technique, we can find the highest correlated variables with the sales price. We use Factor Analysis next to extract 15 factors that influences the data and using output provided by this technique, we perform Linear Regression as the columns have a linear relationship and Linear Regression is used to predict continuous values. We base our models using these techniques.
- **Action and Value Management:** For a client with a house like the ones we have built the prediction model on, we can recommend the approximate sales price and what factors can be improved so that the house may be sold for a better price. For example, if a house has an overall quality of 0.5, this is a low value while considering the similar houses that are sold for much higher prices. We can also recommend improvements to the house for about 10-15% of the sales price, if the improved factor improves the sales price significantly.
- **Conclusion and Interpretation:** Based on the models and analytics involved, from the EDA and factors used in training the models, we can say that certain factors like overall quality, rooms above grade, total basement SF, etc. are more important than other factors that are important like the neighbourhood affecting the sales price. We can deduce that the overall accuracy achieved for predicting sales is 83%. This score can provide decent customer satisfaction. We can also interpret from the data that the busiest months are in Spring but, the most profitable month is seen to be September according to the data.

RECOMMENDATION AND CONCLUSION

Recommendation

Best time of the year to sell the house The data of the number of houses sold each month from 2006-2010 has been analyzed and visualized into bar plots.



It can be observed that the number of houses are sold between the months of May to August. The peak sale is seen to be in July 2006. If the owner would like to sell the house as soon as possible, we would recommend them to sell it during the spring and summer of the year. However, if the customer wants to sell the house at the highest price, we would recommend them to look at the trend of the selling price which is greatest during September.

Q. When is the best time to sell the house in Ames?
Based on the analysis, we would recommend them to sell it during spring or summer of the, especially in the months of May, June, July, August.

Q. How can one improve the sales price of the house in Ames? What features need to be improved so that the expected price can be increased?
Based on the analysis of the dataset, here are some suggestions to improve the price of the homes.

- Try to improve the overall quality of the house by routine maintenance of its utilities like electrical plumbing etc.
- Garage area is one of the highly correlated features, hence try to increase the garage area to fit in more cars.
- Having a bathroom in the basement can also increase the price of the houses.
- We observed houses made of cement and brick are sold at a higher price, so material greatly influences the price.
- If the house has a large number of bedrooms, it can be renovated into multi-purpose rooms.

Q. What are the homes that I can compare my house to, so I can improve the sale price?
When comparing clients house, we can take multiple factors into account which range from house style to the building material most of the comparison can be done using the factors that are categorical. So we pick similar type of houses and then compare them based on the house size, and using the sale price of the similar houses we can recommend the expected sales price or the improvements for the client's house.

Improvements and Enhancement

Here are some of our suggestions on how we can improve this model and prediction.

- We have not taken into consideration a number of categorical. We could try to improve the it by using the categorical variable into consideration for the prediction. It would be better to use both numeric and categorical data for modelling like using an ensemble model.
- We could take into consideration the type of land the house belongs to like if it is an agricultural land, Commercial space, etc. Although this feature did not much affect the price, in real life, it is an important factor.
- We could try to extract more predominant features from the dataset to get better predictive scores. For example, we could create new variables by combining two or more variables like combining neighborhood and total area of the property, which could probably give better results.

Conclusion

Data analytics improved the real estate companies' decision making process and eliminated the need to constantly study all the various factors that affect the price of a house. By getting familiar with the data we were able to figure out which data we should consider and which data can be eliminated making the data cleaner and effective for modelling. Using Factor Analysis, we were able to reduce the dimensions and find out the columns that influence the data while taking into consideration the underlying factors, Correlation provided us with the variables that have the highest correlation with the sales price of a house, Linear regression is performed as it's a really good technique to use when the variables used to predict show linear relationship and this technique is used to predict continuous values. The data analysts did a really good job of landing a prediction score of 83%, a lot more factors engineering is required to increase the predictions score, with these data analytics we can highly improve the customer satisfaction.