PREDICTIVE MODELLING FOR REAL ESTATE

(PREDICTING THE SALES PRICE OF THE HOUSES)

Candidate No: 220298193, Group No. 10

BNM861 – Data Mining and Web Analytics

Table of Contents

PREDICTIVE MODELLING FOR REAL ESTATE -PREDICTING SALES PRICE

INTRODUCTION

HOUSING BUBBLE 2.0. These words are making round in the US housing market. Rising inflation and higher mortgage rates have made this bubble pop-up. The housing market condition of USA at present are quite similar as they were in back 2007. All the stakeholders want to understand the similarity of the situations and the sales price of the houses will be the biggest factor in analysing the market trend. Thus, a predictive model needs to be created which predicts the sales price of the houses. The dataset has been provided and the model will be created using the DATA MINING TECHNIQUES. The provided dataset gives a detailed information on several houses in the **Ames city, IOWA, USA** that **were sold between 2006 and 2010. The predicted sales price will help to compare the market conditions and will help the stakeholders to take wise decisions. The project will be completed in 4 stages.**

STAGE 1: BUSINESS UNDERSTANDING PHASE

i) AIM

The project's primary objective is to build an **EFFICIENT** predictive model that would forecast the **SALES PRICE** of the houses based on a certain set of predictors.

ii) UNDERSTANDING THE PROBLEM AND RECOGNISING STAKEHOLDERS

In 2023, the US housing market is at a turning point, with divergent predictions for its future. Everyone, including the stakeholders like **HOMEOWNERS, HOMEBUYERS, REAL ESTATE BROKERS, MORTGAGE LENDERS, AND BUILDERS**, is speculating about the possibility of a **housing market collapse and attempting to anticipate if the US housing market would crash the way it did in 2007**.

iii) STAKEHOLDERS AND THEIR CONCERNS

a) **HOMEOWNERS**: Home sellers would like to understand the market trend in order to decide whether or not they should delay their decision to market the property now and if they should lower the asking price of their houses.

b) **HOMEBUYERS**: Owing to a rise in interest and mortgage rates, homeowners are considering whether now is the ideal time to buy a home and if they will be able to make ends meet if they do.

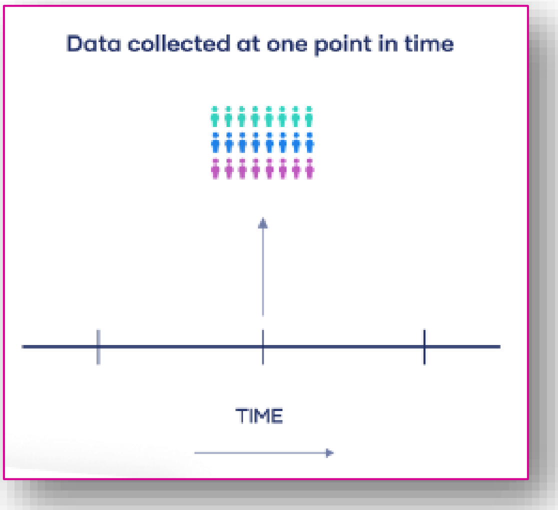
c) **MORTGAGE SECTOR**: The mortgage industry is seeking to modify its action plans to protect itself from the losses as it doesn't want to repeat the catastrophes of 2007.

All the stakeholders would thus be interested in the predicted Sales Price of the properties in order to comprehend the housing market trend and develop plans appropriately.

STAGE 2: DATA UNDERSTANDING PHASE: DATASET AND VISUALISATION

i) TYPE OF THE DATA

- The modelling process makes use of **“Multivariate Cross-Sectional Data”.**
- As contrast to time series data, cross-sectional data is a type of data that is acquired by observing several subjects at once or over a short period of time.
- The dataset provides detailed information on several houses in the **Ames city, IOWA, USA** that **were sold between 2006 and 2010.**



ii) DIMENSIONS OF THE DATASET

- The number of **records (rows)** and number of **features/predictors(columns)** refers to the dimensions of the dataset.
- The dimensions of the CW dataset are **1144 X 79** i.e., there are 1144 records, where each observation has 79 features.

The modelling process will make use of only **10 features/predictors** which are chosen after an initial modelling on the dataset. The initial modelling included performing **“Feature Importance”** and **“Correlation Analysis”** for the Numerical Variables.

Nodes	Importance
HeatingQC	0.0078
GarageType	0.0078
MSZoning	0.0078
GarageQual	0.0078
YearRemodAdd	0.0166
MSSubClass	0.0289
GrLivArea	0.0508
TotalBsmtSF	0.0646
OverallQual	0.6883

iii) VARIABLES; TYPES, DEFINITIONS AND THEIR ROLES

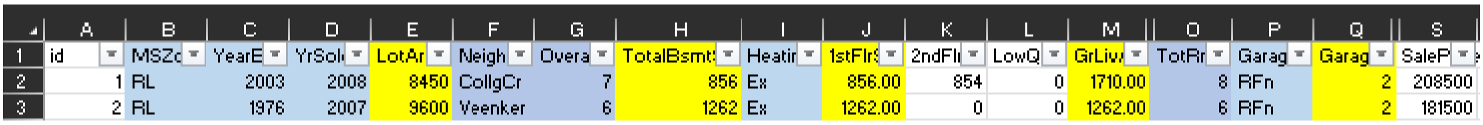
- The variables considered are both **Quantitative (Numerical)** and **Qualitative (Categorical)** in nature, where a numeric variable has values are numbers, and a categorical variable tends to categorize items into a few groups.
- The **Numerical** variables are further categorized as **Discrete** (No meaningful value between two consecutive real values) and **Continuous** (Can assume an infinite number of real values within a given interval).
- The **Categorical** variables are further classified as **Nominal** (Values are just labels, without any order) and **Ordinal** (Values are defined by an order relation between the categories).

- **"Sales Price"** is the **Target variable(T)** in the modelling process. As, the name suggests, the values of this variable will be predicted based on other variables which are known as **"Predictors(P)"** or **"Predictor Variables".**

The variables considered are:

Variable Names and Types	Definitions and their roles
a) **Ms Zoning: Nominal Variable. (P)**	It identifies the general zoning classification of the sale. The prices of the houses may vary as per their zones.
b) **YearBuilt and Yr Sold: Categorical (P)**	These two variables will help in **determining the age** of the houses. **Feature Engineering** will be done to find the age of the houses as it is the most critical factor which determines the house price.
c) **Lot Area: Continuous Variable(P)**	It is the total area of a property i.e. the area where the house is built on.
d) **Overall Quality: Ordinal Variable (P)**	It refers to the overall material and finish of the house.
e) **Total BsmtSF: Continuous Variable (P)**	It refers to the total square feet of basement area. The basement area is not included in the total floor area and is calculated separately.
f) **Heating Quality: Ordinal Variable (P)**	It refers to the heating quality of the house.
g) **1stFlrSF, 2nd FlrSf, LowQualFinSF, GrLIvArea : Continuous Variable (P)**	These **4 variables** refer to the floor area of the house and thus will be aggregated to find the **"Total Floor Area"**.
h) **TotRmsAbvGrd: Discrete Variable (P)**	It refers to the total number of rooms in the house.
i) **GarageCars: Discrete Variable (P)**	It refers to the size of garage in cars capacity.
j) **GarageFinish: Nominal Variable (P)**	It refers to the finish of the garage.
k) **Sales Price: Continuous Variable (T)**	**It refers to the selling price of the house. The model will be used to predict the sales price of the houses based on all the above predictors.**

iv) VERBAL REPRESENTATION



	A	B	C	D	E	F	G	H	I	J	K	L	M	O	P	Q	S
1	id	MSZc	YearE	YrSol	LotAr	Neigh	Overa	TotalBsmt	Heatir	1stFlr	2ndFl	LowQ	GrLiv	TotRr	Garag	Garag	SaleP
2	1	RL	2003	2008	8450	CollgCr	7	856	Ex	856.00	854	0	1710.00	8	RFn	2	208500
3	2	RL	1976	2007	9600	Veenker	6	1262	Ex	1262.00	0	0	1262.00	6	RFn	2	181500

- The first record belongs to the house in **RL zone** i.e. a Low-Density Residential Area, which was built in **2003** and sold in the year **2008**. It had a lot area of **8450 square feet** and was in the **College Creek** Area. The area of the first floor, second floor and ground floor were 856 square feet,854 square feet & 1710 square feet respectively. It aggregates to a total floor area of around **3500 square feet,** with **8** rooms and a **900 square feet basement**. The overall quality of house was good. It also had an excellent heating quality along with an unfinished **garage** with a capacity to park **2** cars. The house was sold for **$208,500.**

v) LEVEL OF THE RECORDS

The level of the records is the Sales Price of the houses. There are as many rows as the Sales Price of the houses.

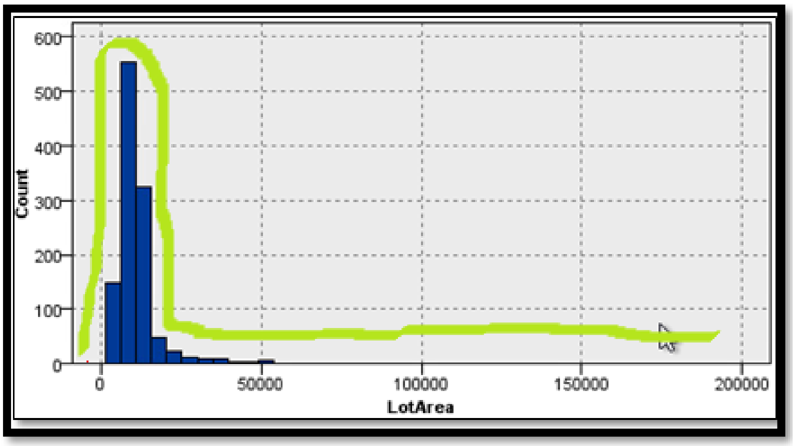
vi) UNIVARIATE VISUALISATION

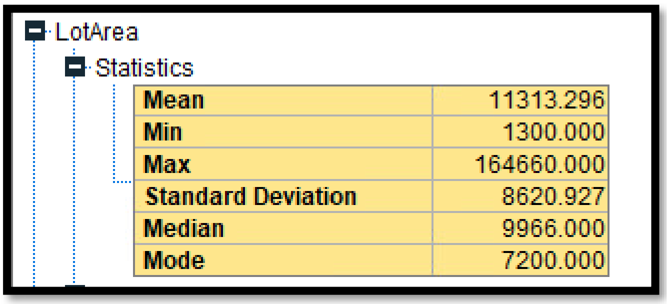
- Univariate Visualisation and Analysis of the dataset will not only help us in **understanding the distribution of the variables** but will also help us in **uncovering the hidden trends, patterns, and anomalies.**
- It is very important to check the types of the variables and convert them into their **desired types and the role of each variable should be inputted correctly.**
- The visualisation and analysis will be done in 2 parts:

a) EXPLORING NUMERICAL VARIABLES

- **Histograms** will be plotted for analysing the distribution of the variables i.e. are they following **Normal Distribution** or not.
- They will also throw light on the **skewness** of the distribution, where if the tail of the curve is stretched towards right-hand side, the distribution will be **right** or **positively skewed**, else it will be **negatively skewed**.
- The measures of centrality will give the insight on **mean** i.e. that value of the **variable around which other variables are scattered, mode**; the **most frequent** occurring value of the variable, **median** and **Standard Deviation** indicating the **spread of the values** in the distribution.
- As, only 10 important predictors must be taken into account for modelling, 4 different variables corresponding to the floor area were aggregated to form one variable. This will be further explained in **"FEATURE ENGINEERING"**

LOT AREA





LotArea

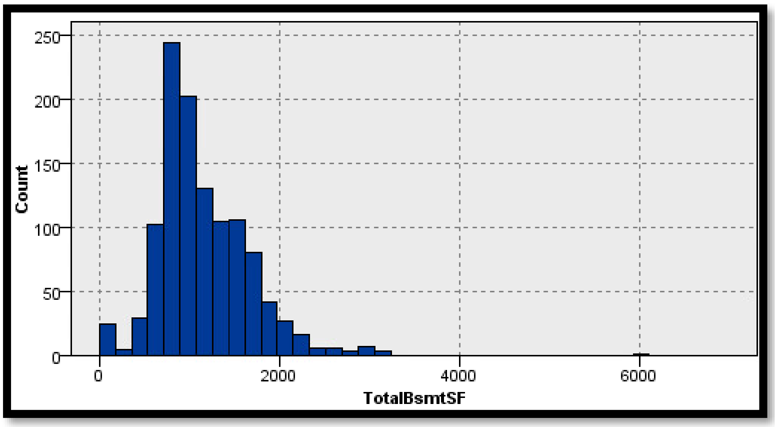
Statistics

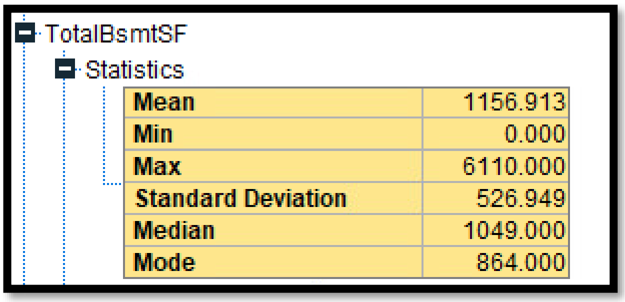
Mean	11313.296
Min	1300.000
Max	164660.000
Standard Deviation	8620.927
Median	9966.000
Mode	7200.000

OBSERVATIONS

- The histogram points out to the variable being highly positively skewed, with only a handful number of houses having lot area more than 30,000 square feet.
- The average lot area for houses is 11,313 square feet.
- **The high standard deviation points out to a wide spread in the lot areas values i.e. the values of the lot area are too much distant from its mean and are not clustered around it.**

TOTAL BsmtSF (BASEMENT AREA)





TotalBsmtSF

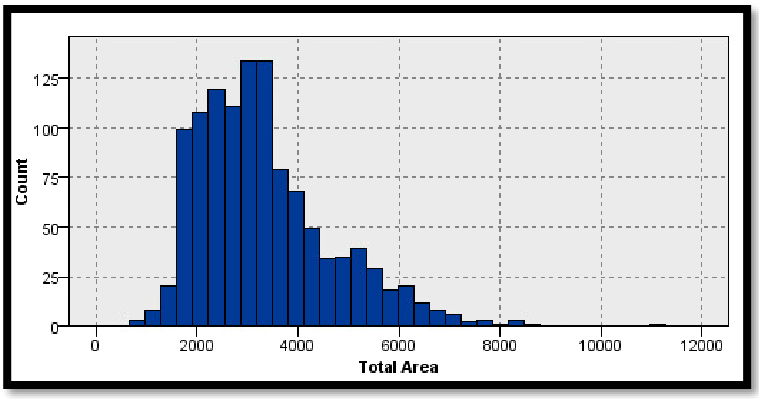
Statistics

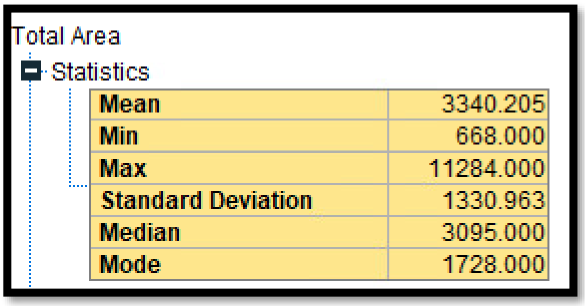
Mean	1156.913
Min	0.000
Max	6110.000
Standard Deviation	526.949
Median	1049.000
Mode	864.000

OBSERVATIONS

- There were certain houses with no basement. The average basement area of the houses was 1156 square feet.
- Alike, the lot area, the distribution is **positively skewed** (right skewed, with a skewness measure of 1.44).
- There are some houses which have a large basement area, **almost three times the average basement area and will be treated as outliers.**

TOTAL AREA





Total Area

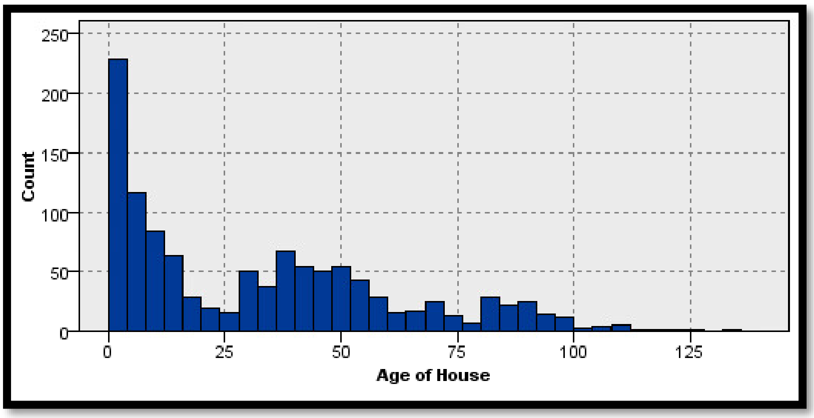
Statistics

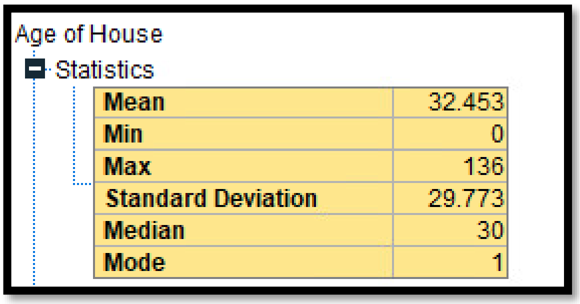
Mean	3340.205
Min	668.000
Max	11284.000
Standard Deviation	1330.963
Median	3095.000
Mode	1728.000

OBSERVATIONS

- The average total floor area of the houses was 3340 square feet.
- Alike, the lot and basement area, the distribution is **positively skewed** (right skewed, with a skewness measure of 1.126).
- The median of the **Total area** is 3095 square feet. It is interesting to observe that **there were 50% of the houses in a small range of 668-3000 square feet and another 50% of the houses had total floor area in a widespread range of 3095 -11284 square feet.**

AGE OF HOUSE





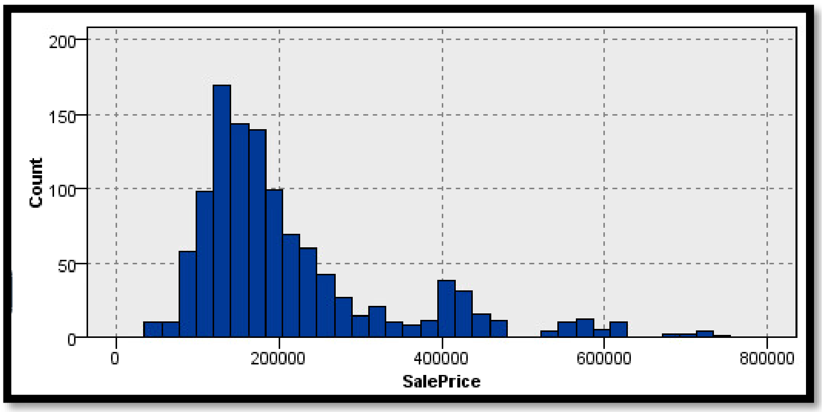
Age of House

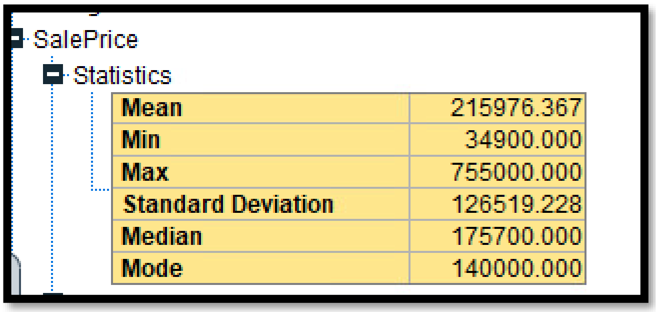
Statistics

Mean	32.453
Min	0
Max	136
Standard Deviation	29.773
Median	30
Mode	1

OBSERVATIONS

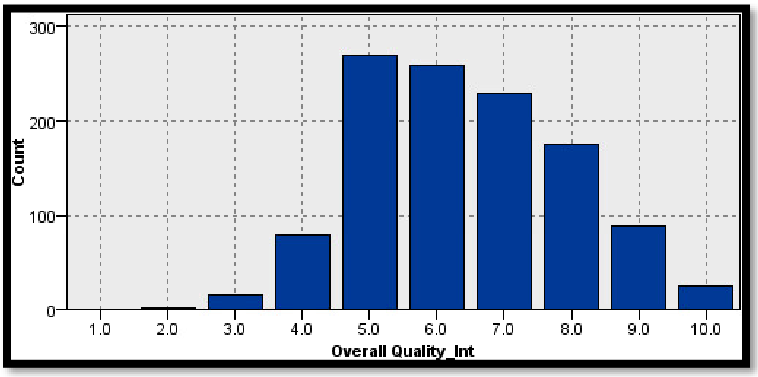
- The average age of the houses sold was 32 years.
- It is interesting to observe that the maximum no.of houses that were sold **(MODE)** were only 1 year old.

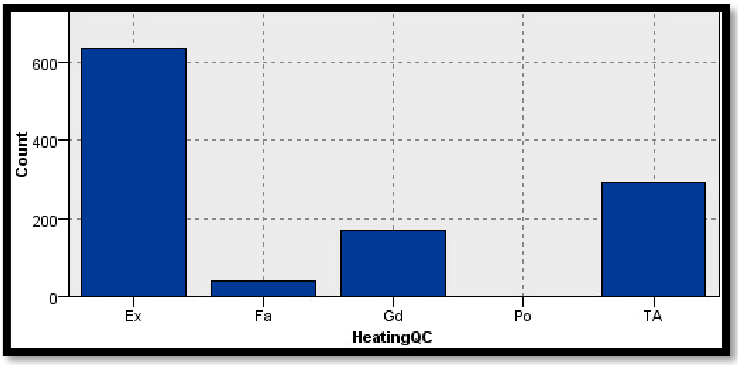
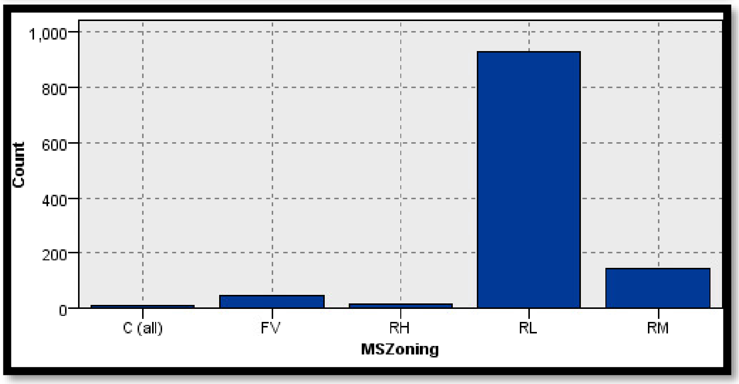
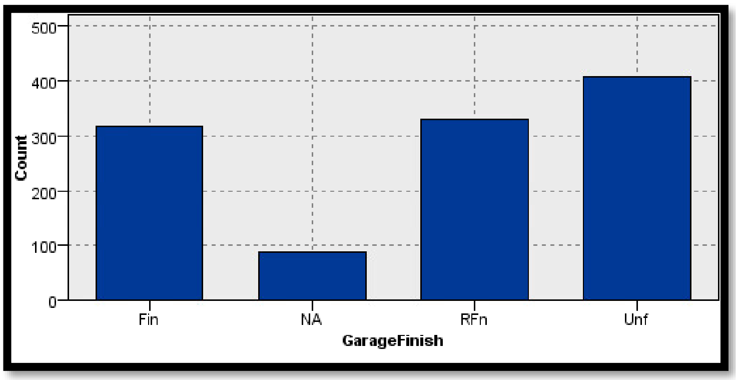
<table>
<tr><th>SALES PRICE (TARGET VARIABLE)</th><th>OBSERVATIONS</th></tr>
<tr><td>



SalePrice
Statistics
<table>
<tr><td><b>Mean</b></td><td>215976.367</td></tr>
<tr><td><b>Min</b></td><td>34900.000</td></tr>
<tr><td><b>Max</b></td><td>755000.000</td></tr>
<tr><td><b>Standard Deviation</b></td><td>126519.228</td></tr>
<tr><td><b>Median</b></td><td>175700.000</td></tr>
<tr><td><b>Mode</b></td><td>140000.000</td></tr>
</table>
</td><td>
- The average sale price of the houses is **$ 215,976.367 with the maximum sale price as $755,000. It clearly indicates that the Sales Price values are widely spread around the mean.**
- The maximum number of houses **were sold for less than $ 200,000.**
- The variable doesn't **follow Normal Distribution and is positively skewed.**
</td></tr>
</table>

b) EXPLORING CATEGORICAL VARIABLES

Bar Graphs will be plotted to visualise each data category in a frequency distribution.

OVERALL QUALITY	OBSERVATIONS
	It can be observed that most of the houses sold were of **medium quality**. It throws a **light on the possibility that the sales price of the houses must have been increasing with an increase in the overall quality.**

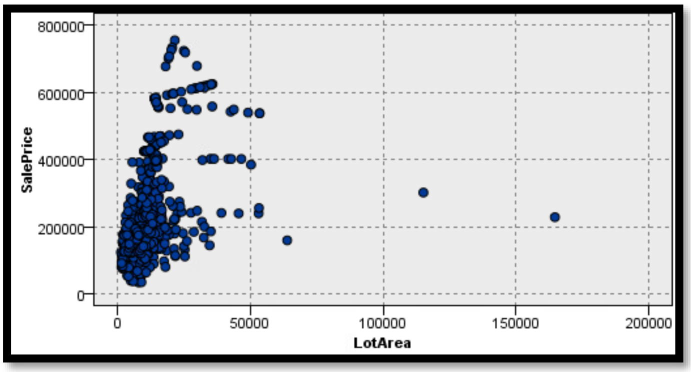
HEATING QUALITY,	OBSERVATIONS
  	It can be observed that the maximum number of houses sold were of excellent heat quality. The most interesting thing to observe here is that the houses with unfinished garage were the most preferred one The Bar Graph illustrates that the maximum number of houses sold belonged to the RL zone i.e. Residential Low Density zone.

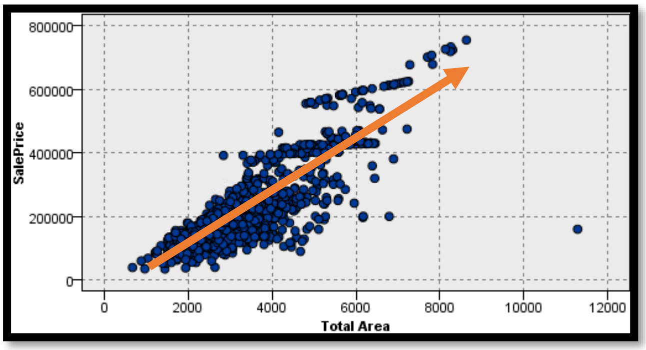
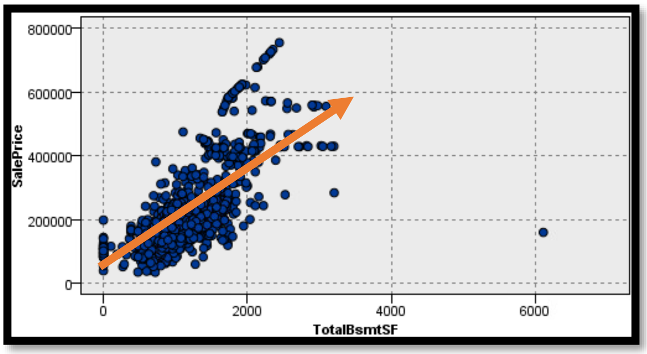
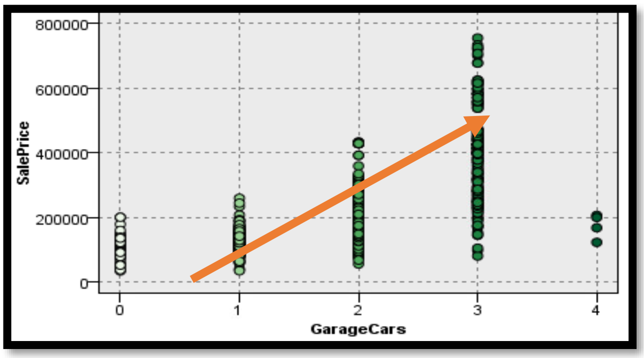
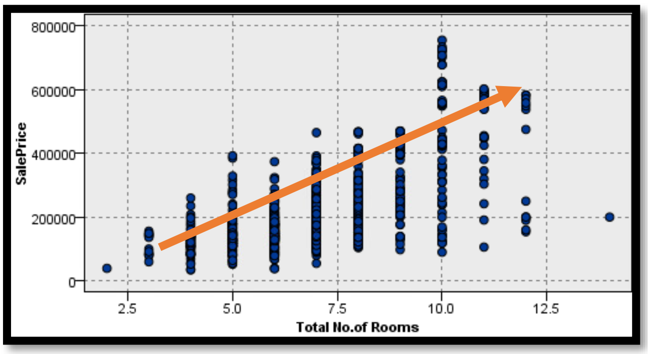
c) BIVARIATE VISUALISATION AND ANALYSIS

Bivariate Visualisation and Analysis will reveal the increasing / decreasing trend of the **Sales Price** in accordance with the predictor variables, thus revealing the **predictive power** of the features.

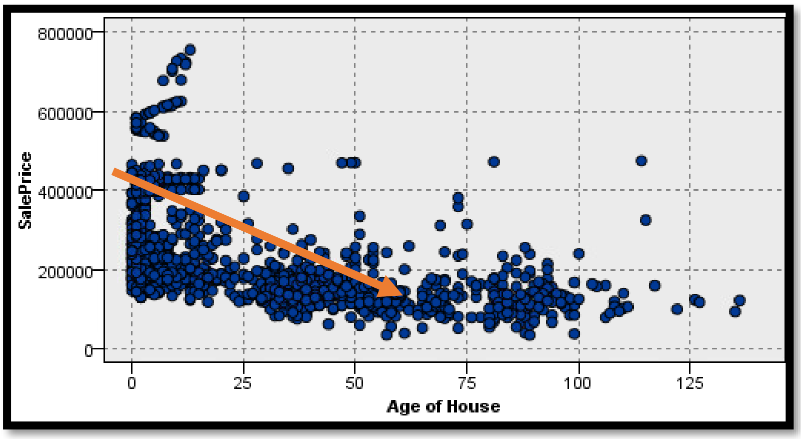
a) PLOTTING SALES PRICE VS NUMERICAL PREDICTORS

Scatter plots will be plotted between the **numerical variables** (taken on X-axis) and the **SALES PRICE (taken on Y-axis)**.

LOT AREA AND SALES PRICE	OBSERVATIONS
	There is **no** evidence of any **linear relationship** between the **LOT AREA AND SALES PRICE.**

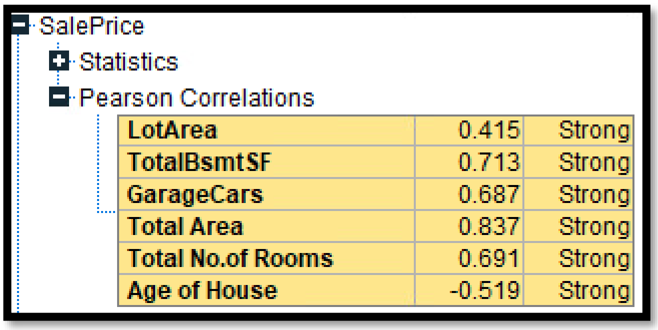
TOTAL AREA 	TOTAL BASEMENT AREA 
GARAGE CAPACITY 	TOTAL NO.OF ROOMS 
There is a **positive correlation** between all these 4 **variables and Sales Price .** With an increase in these variables, the sales price will also increase.	

AGE OF HOUSE | **OBSERVATIONS**



As the Age of House is increasing, the <span>SALES PRICE</span> is **decreasing**. It can also be observed that the **newer houses were in high demand and some of them were sold for exorbitantly high prices.**

CORRELATION MATRIX: | **OBSERVATIONS**



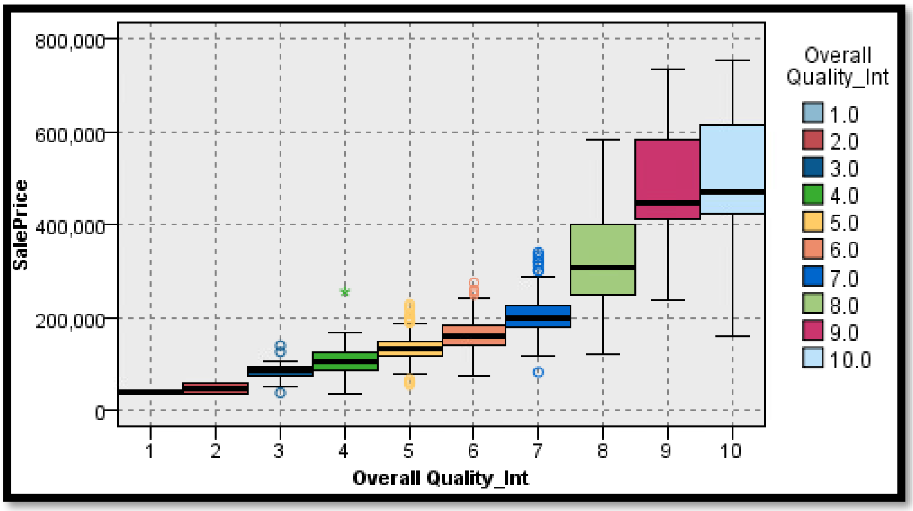
SalePrice
Statistics
Pearson Correlations

LotArea	0.415	Strong
TotalBsmtSF	0.713	Strong
GarageCars	0.687	Strong
Total Area	0.837	Strong
Total No.of Rooms	0.691	Strong
Age of House	-0.519	Strong

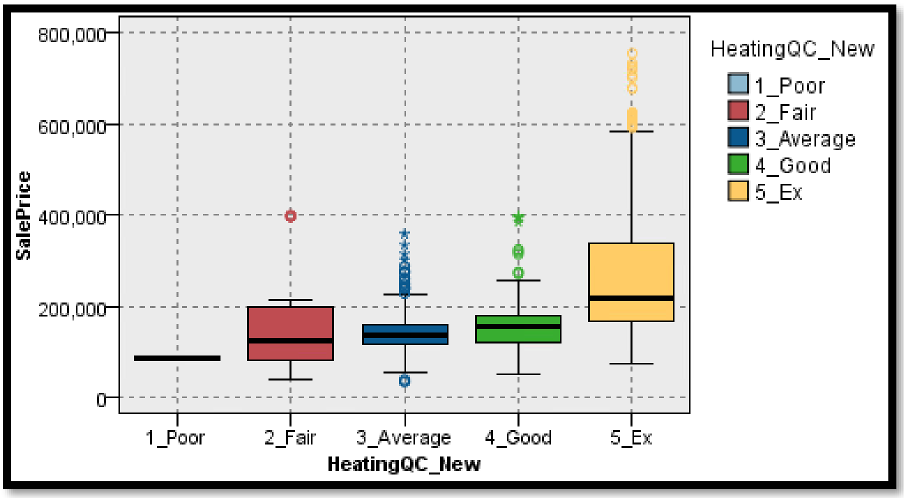
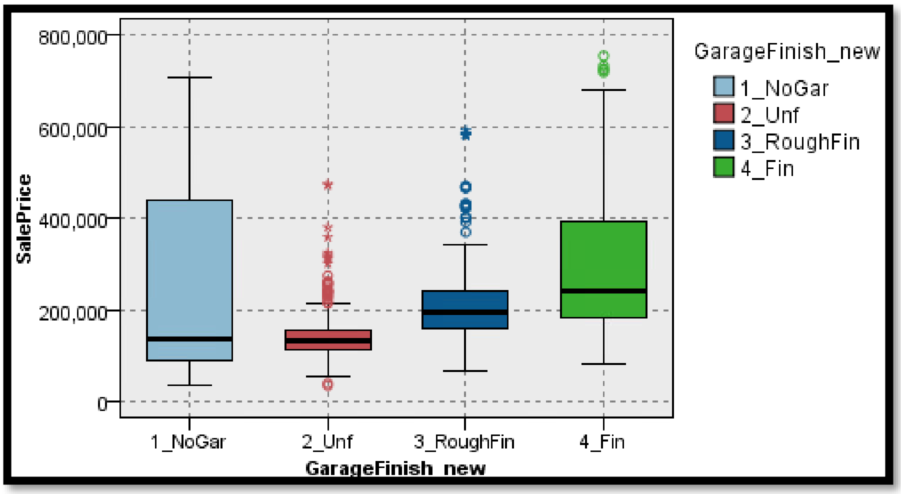
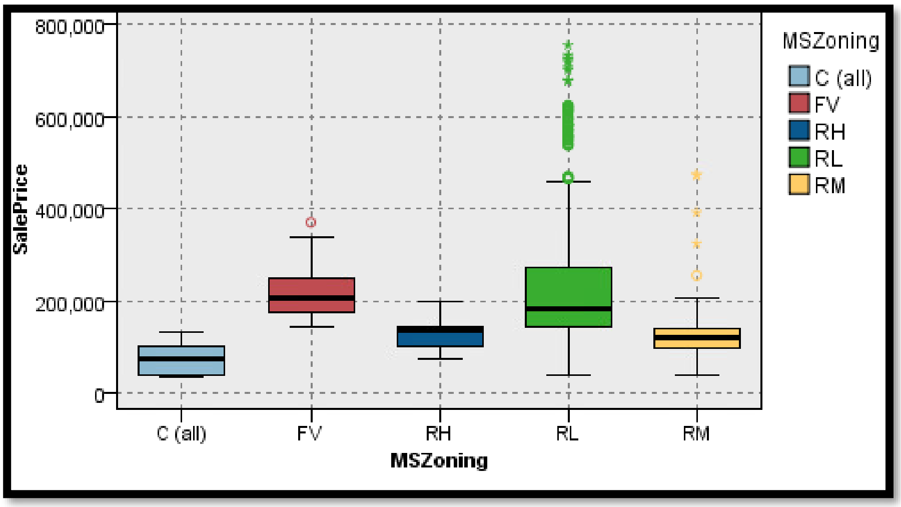
The Correlation Matrix between the Sales Price and all the other Numerical Predictor shows that all of them are **strongly correlated** with Sales Price, thus all of them being **IMPORTANT PREDICTORS**.

b) **PLOTTING SALES PRICE VS CATEGORICAL PREDICTORS**

BOX PLOTS will be plotted between the **categorical variables (taken on X-axis)** and the **SALES PRICE(taken on Y-axis)**



The Box Plot clearly indicates the strong predictive power of this feature. There is a clear increase in the prices of the houses with an increase in the overall quality.

	There is a marginal increase in the Median sale price of the houses with an improvement in the Heating Quality.
	The prices show an upward trend with the improvement in the finish of the Garage. However, the median Sale Price of the houses with No Garage and Unfinished garage are almost same.
	No trend can be seen in the median prices. MsZoning seems a Weak Predictor.

c) DATA QUALITY ASSESSMENT AND TREATMENT

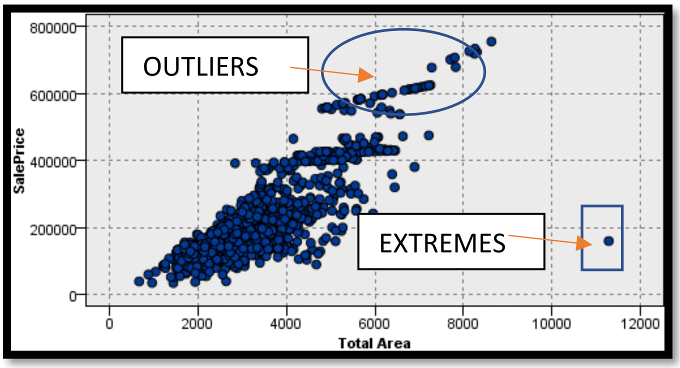
" **If the input data is seriously flawed, no amount of statistical massaging will produce a meaningful result".**

(Guttag, John.V, 2017).

This relates to the **concept of GIGO; "GARBAGE IN GARBAGE OUT".** The overriding objective of minimizing GIGO can be achieved by Data Cleaning which involves with dealing of duplicate records, missing values and outliers.

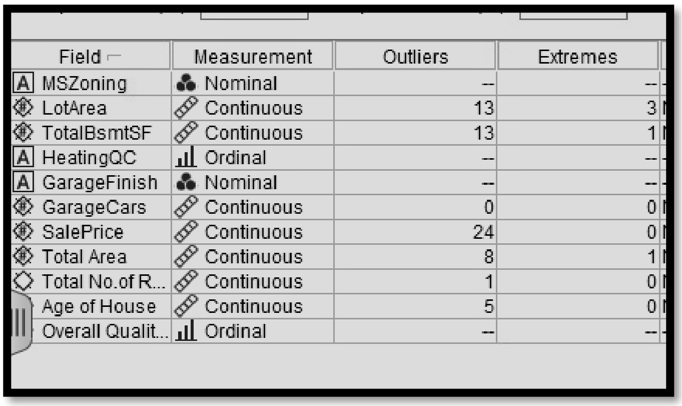
a) OUTLIERS AND EXTREMES

- The observations that lie at an abnormal distance from the other values in a distribution are called Outliers and Extremes.



- Outliers follow the distribution of the datapoints but are quite distant from the cluster whereas Extremes doesn't follow the distribution and are remotely placed.
- The values which are beyond Q1-1.5* IQ and Q3+1.5*IQ are considered as outliers or any value which is more than 2 S.D. and 3 S.D. are considered as an outlier and extreme respectively.

  (Here, Q1= First Quartile, Q3=Third Quartile , IQ=Interquartile Range and S.D. = Standard Deviation)



Field	Measurement	Outliers	Extremes
MSZoning	Nominal	--	--
LotArea	Continuous	13	3
TotalBsmtSF	Continuous	13	1
HeatingQC	Ordinal	--	--
GarageFinish	Nominal	--	--
GarageCars	Continuous	0	0
SalePrice	Continuous	24	0
Total Area	Continuous	8	1
Total No.of R...	Continuous	1	0
Age of House	Continuous	5	0
Overall Qualit...	Ordinal	--	--

- The above table shows the number of outliers and extremes in the dataset.

TREATMENT: Outliers are not removed from the dataset as it is quite possible that **they are representing the true state of the dataset** and it is necessary for the model to learn from the data set. However, as they are skewing the dataset, **LOG TRANSFORMATION** can be done to reduce the skewness.

b) MISSING VALUES

- Missing values in the dataset occurs when the data is not stored for certain features.
- There were 5 missing values in the Masonry Veneer area column. However, this variable is not being used for the modelling purpose. Else, these values would have been imputed using mean/median.

c) OTHER DATA QUALITY ISSUES

- **The columns "Year Built", "Year Sold", "Total Rooms Above Grd", " Overall Quality" had decimal values which were not correct**. Since, there were around 100 cells with decimal values for each type, they **were not considered as noise**. Rather , they were rounded off .

STAGE 3 : MODELLING PHASE

i) PREDICTIVE MODELLING FORMULATION

a) WHAT:

- The goal is to build a model to predict the **SALES PRICE** for the houses based on a number of predictors.

b) HOW:

1) Here, as we have a **PRESPECIFIED TARGET & HISTORICAL DATA**, the model will be built using the **SUPERVISED METHODS**,
2) Using Data Partitioning, the dataset will be divided into 2 parts**;** **TRAINING AND TESTING SET .**
3) **T**he model will be **TRAINED** and **BUILT** on the **TRAINING SET** i.e. it will search, **LEARN** and analyse the patterns and associations between the predictor variables and the target variable.
4) The model will then **USE THIS LEARNING TO PREDICT THE PRICES** for any new learning instance.
5) The **QUALITY OF LEARNING** will be evaluated on the **TESTING SET.**

ii) IS DATA PARTIONING REQUIRED?

- YES, Data Partitioning is required as it splits the original dataset before creating the model (on TRAINING SET) so that there is **"NEW"** (TESTING SET) data available to **assess the MODEL'S PERFORMANCE**.
- It also assists in choosing **a model from a set of PREDICTIVE MODELS** and helps in **reducing the overfitting** of the data.

iii) TYPE OF PROBLEM

- Supervised learning models are generally used for either **REGRESSION** (PREDICTING **CONTINUOUS OUTCOMES**) or **CLASSIFICATION**(CLASSIFYING AN OBJECT) problems.
- Here, the purpose is to build a model to predict the **SALES PRICE** which is a **CONTINUOUS NUMERICAL VARIABLE**, thus the problem at hand is a **REGRESSION PROBLEM.**

iv) PERFORMANCE METRIC

- A performance metric is used to assess the quality of prediction (quality of learning) by a model.
- A regression model will be of a good quality if its prediction matches up against the actual values. There are three types of Performance Metrics for Regression:

  a) Mean absolute error (MAE)

  b) Root mean square error **(RMSE)**

  c) R^2
- **METRIC CHOICE FOR THE PROBLEM AND RATIONALE**

- The metric choice for this problem is **RMSE**.
- The RMSE or Root Mean Squared Error is the average root-squared difference between the real value and the predicted value.

$RMSE = \sqrt{\frac{1}{N}\sum(Y - Y^{\wedge})^2}$, where Y = Actual value, $Y^{\wedge}$ = Predicted Value

- RMSE has been chosen for 2 main reasons:
  - Firstly, as it gives **more weightage to large errors and outliers,** it will help in raising the alarm if the sales price **has been overestimated or underestimated too much**. In such cases, **RMSE will be too large** and will give us an opportunity to **go back in the dataset, analyse those datapoints which are contributing to the error & perform some corrective measures.**
  - Secondly, it will also help in understanding the **overperformance and underperformance** of the model.

v) FEATURE ENGINEERING

- Feature engineering is the process of **identifying, modifying, and converting unprocessed data into features** that may be utilised in supervised learning.
- A new feature **"AGE OF HOUSE"** was created from 2 features in the dataset; " YEAR BUILT" AND "YEAR SOLD" . (Age = Year Sold – Year Built)
- **"TOTAL AREA "** was created by merging 4 features ; " LOWQLFINSF" , "1STFLRSF", " GRLIVAR" AND " 2NDFLRSF".
- **DUMMY VARIABLES** were created for all the **CATEGORICAL VARIABLES.**

vi) MODEL CREATION USING DIFFERENT SUPERVISED METHODS

a) BASELINE MODEL

- Baseline Model or the Naïve Model is the **LEAST SOPHISTICATED PREDICTIVE MODEL.**
- The performance of a baseline model is used as a benchmark or a reference point to compare the performance of other complex models.
- For regression problems with a target called y, the **naïve model prediction is MEAN**(y) for any observation on the training dataset, regardless of all the features.
- The same prediction will be used for the testing set as well.

b) LINEAR REGRESSION MODEL

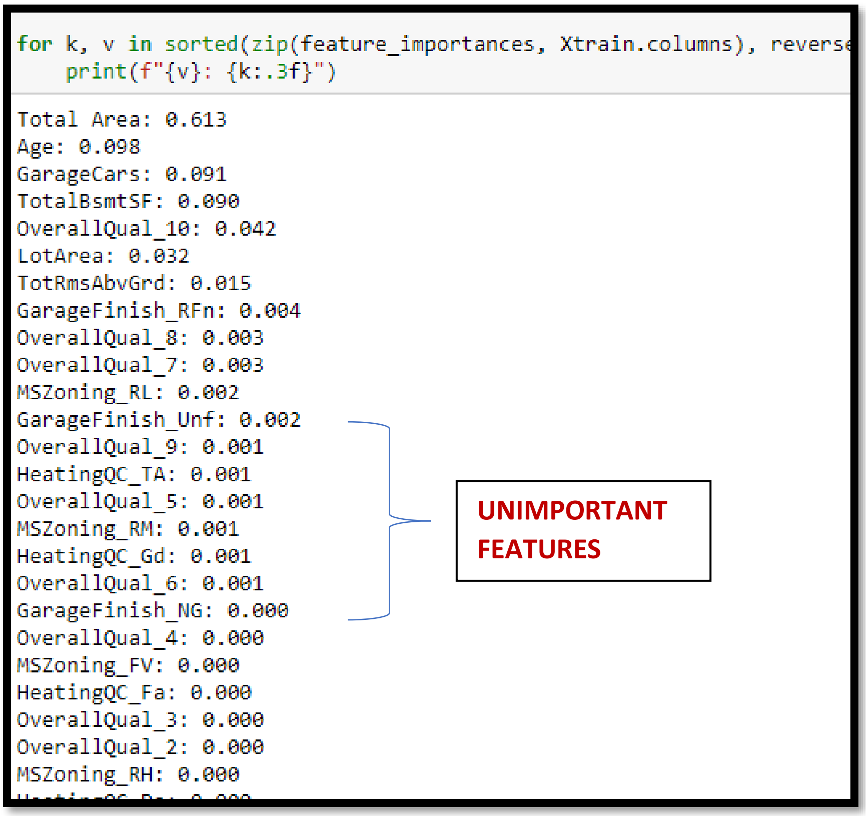
- Linear regression model tries to find the best fit linear line between the independent and dependent variables such that the values of intercept and coefficients are optimized and the error is minimized, where

Error = Actual Values – Predicted Values

c) DECISION TREES

- A decision tree uses node splitting to make predictions. Node splitting breaks a node into many sub nodes, ensuring that the new nodes are PURE NODES.
- During training, splitting is carried out repeatedly and recursively until only homogeneous nodes are left.
- **Decision trees were used to understand the important features in the dataset. All the unimportant features were dropped from the dataset.**
- **HYPERPARAMETERS:**

It is used to **control the learning process** and is provided by the user . It can’t be estimated from the data. **Hyperparameter Tuning** will be done for both **C and R trees and Random Forests.**



```
for k, v in sorted(zip(feature_importances, Xtrain.columns), reverse
    print(f"{v}: {k:.3f}")
```

```
Total Area: 0.613
Age: 0.098
GarageCars: 0.091
TotalBsmtSF: 0.090
OverallQual_10: 0.042
LotArea: 0.032
TotRmsAbvGrd: 0.015
GarageFinish_RFn: 0.004
OverallQual_8: 0.003
OverallQual_7: 0.003
MSZoning_RL: 0.002
GarageFinish_Unf: 0.002
OverallQual_9: 0.001
HeatingQC_TA: 0.001
OverallQual_5: 0.001
MSZoning_RM: 0.001
HeatingQC_Gd: 0.001
OverallQual_6: 0.001
GarageFinish_NG: 0.000
OverallQual_4: 0.000
MSZoning_FV: 0.000
HeatingQC_Fa: 0.000
OverallQual_3: 0.000
OverallQual_2: 0.000
MSZoning_RH: 0.000
```

UNIMPORTANT FEATURES

d) RANDOM FORESTS

- It is an ensemble method based on the ideas of BOOTSTRAPPING AND BAGGING.
- It uses a collection of trees of the same problem, get predictions from all of them and then average all of them using **MEAN**

e) SUMMARY TABLE FOR MODELS PERFORMANCE

MODEL	HYPERPARAMETERS	TRAINING SET	TESTING SET
BASELINE	-		126880.42
LINEAR REGRESSION	-	41520.61	34782.45
C AND R TREES	Max Depth = 5	29315.96	39032.66
	Max Depth=6	23996.69	35260.71
	Max Depth = 7	18785.08	33936.77
	Max Depth=8	15389.78	32910.34
RANDOM FORESTS	n-estimators = 10, Max Depth =5	26501.04	32687.94
	n-estimators = 15, Max Depth =6	20199.42	29134.88
	n-estimators = 20, Max Depth =8	17395.02	29279.89
	n-estimators = 30, Max Depth =8	16241.28	29114.51

(Note: RMSE has been used as a performance metric for all the models)

f) REVIEWING MODELS PERFORMANCE

- The performance of the models is increasing with the complexity of the models.

- However, the model is performing better on a training set as compared to the testing set, which is bringing out the fact that the model is **OVERFITTING THE DATA.**

g) ERROR COST ANALYSIS

- Two types of Errors are involved in a Regression Problem.
  1) **OVER-ESTIMATION** – The model shouldn't overestimate the sales price because it would be detrimental for both home buyers and sellers. Overestimating prices can result in losses for home buyers since they are not only wind-up spending more but they additionally increase their mortgages. It is very likely for home sellers that their properties won't sell as a result of the exorbitant pricing
  2) **UNDER ESTIMATION** – For house sellers the model should not underestimate values since doing so would result them receiving less revenue than they are entitled to.

     The **worst errors** would be one involving **OVERESTIMATION** for both ; **HOMEOWNERS AND HOMEBUYERS** since it would make homebuyers put off their plans to buy a house. Moreover, a homeowner may be compelled to decrease prices since his residence may fall under the category of being overpriced.

STAGE 4: CONCLUSION AND RECOMMENDATIONS

i) OBSERVATIONS

A model to predict the sales price of the houses was created based on a set of 10 predictors. Some of the important observations are:

- Total area and age of the house are the 2 important factors determining the sales price of the houses.
- There was a significant variation in the LOT AREA sizes and ages of the houses.
- There were a considerable number of houses which were priced quite high as compared to the other houses.

ii) RECOMMENDATIONS AND FUTURE PROJECTS

- The dataset had some outliers that were left in since they accurately reflected the status of the data and were not deleted. The fact that the outliers are having a significant influence on the models and should be handled accordingly is highlighted by the increased RMSE values.
- Remodelling should be done by removing the outliers to check if there is some improvement in models' performance.
- Moreover, even on the training set, the complex RANDOM FORESTS model's RMSE is still excessively high. A future project can entail developing a new model by including more features and data.