2023

Data Mining and Web Analytics Coursework

Group Members

1. Anandu Karunakaran (220241328)
2. Aravind Kalissery Vijayakumar (220241454)
3. Hydher Senan Pandi Kadavath (220242831)
4. Shon Mathew (220241029)
5. Aiswarya Shyni Suresh (220135245)
6. Abin Simon (220141453)

Anandu Karunakaran

220241328

3/24/2023

Contents

Introduction

Real estate is an industry that revolves around buying, selling, leasing, and renting buildings and so on. Accurately forecasting the market's property sales prices has become more and more important recently. In the real world, predicting the accurate price of a property is difficult as there are a lot of variables that can affect its price. The objective of this report is to predict the selling price of properties using a selection of 10 relevant predictors from the CW dataset, which has a total of 79 possible predictors. To improve the analysis's practicality and make data interpretation easier, the number of predictors must be decreased.

The report will help in determining which factors are key in determining the price of a property.

Major Stakeholders

Property buyers and sellers

- **Investors in the real estate market**
- **Real estate brokers**
- **Property developers**
- **Financial Organisations**
- **Government**

Why does this problem matter to each stakeholder?

- **Property buyers and Sellers:** They are the people who are interested in buying or selling the property and can use the predictions to evaluate the value of the property to avoid overpaying to an extent or can negotiate for a better price.
- **Investors in the real estate market**: People who are interested in investing in the market can use the predictions and evaluate each feature of the data to know where to invest their money.
- **Real estate brokers**: Accurate sale price predictions help brokers give correct valuations of properties to their clients, increasing their credibility and improving their chances of closing deals. They can also use the sales predictions to pitch their clients with a price which will also include their commission.
- **Property developers**: They can plan their construction and development plan by assessing the price to understand the market which can help them in getting huge return on investment.
- **Financial Organisations**: They can use the models price predictions to assess the risk of lending money.
- **Government:** In order to regulate and guard the market from fraud, government bodies can use this model.

Verbal presentation

A regular shaped house in Iowa DOT and Railroad neighbourhood is the cheapest house in the entire dataset with a below average quality and built in 1920.

Level of data

Each record in the data talks about price of the house and various features of that house. It has various numerical and categorical features.

Data set and Visualization.

Type of the Data set

The given dataset is a cross sectional dataset as it gives information of different properties and their prices.

Dimension of the data set

The given dataset has 79 predictors and a target variable out of which we take 10 predictors for the modelling purpose. The dataset contains 1145 rows or observations.

Variables, definitions, their types, and their roles

- OverallQual
  - Definition- Rates the overall material and finish of the house. It ranks the quality of the house from 1 to 10.
  - Type- Categorical data - Ordinal Data
  - Role- Predictor
- GrLivArea
  - Definition- Above grade (ground) living area square feet.
  - Type- Continuous Data
  - Role- Predictor
- Neighborhood
  - Definition- It depicts the location of the property within the Ames city limit.
  - Type- Categorical Data – Nominal data
  - Role- Predictor
- MoSold
  - Definition- It is the month in which the given house was sold.
  - Type- Categorical data – Nominal data
  - Role- Predictor
- LotShape
  - Definition- It gives the general shape of the given property.
  - Type- Categorical data – Nominal data
  - Role- Predictor
- MasVnrArea
  - Definition- Masonry veneer area in square feet
  - Type- Continuous data

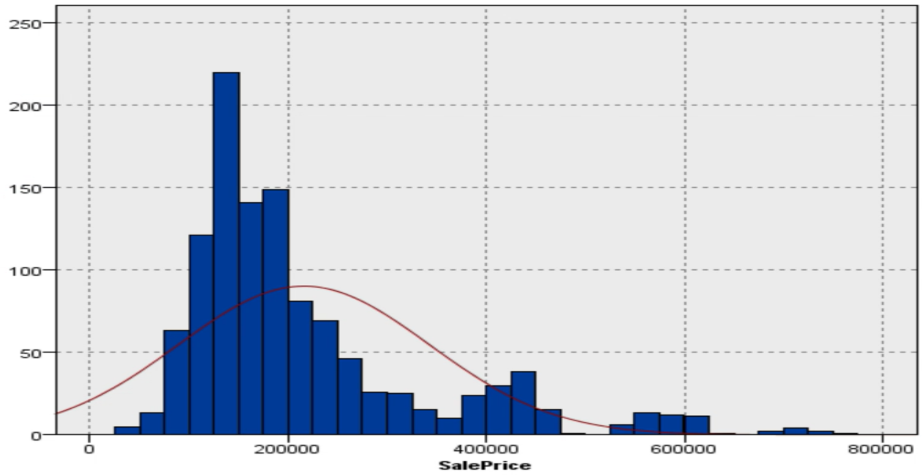
  - Role- Predictor
- BedroomAbvGr
  - Definition- Number of bedrooms above grade in the property. This does not include the basement room.
  - Type- Continuous data
  - Role- Predictor
- BsmtUnfSF
  - Definition- Unfinished square feet area in basement of the given property
  - Type- Continuous data
  - Role- Predictor
- LotArea
  - Definition- Lot area of the property in square feet
  - Type- Continuous data
  - Role- Predictor
- Age
  - Definition- Age variable is obtained from the year the house was built.
  - Type- Continuous data
    Role- Predictor

- SalePrice
  - Definition- Sale price of the house.
  - Type- Continuous data
  - Role- Target

Visualisation

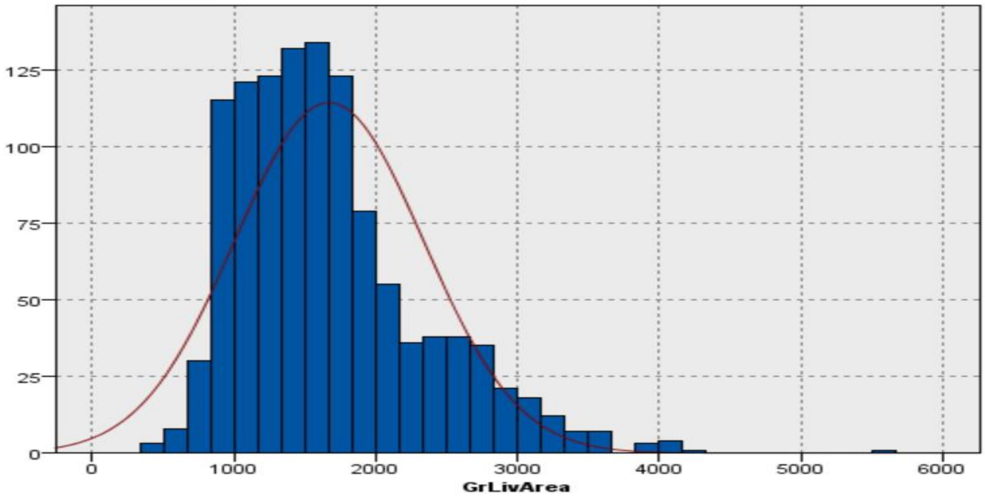
Uni Variate Visualisation

Distribution of sales price



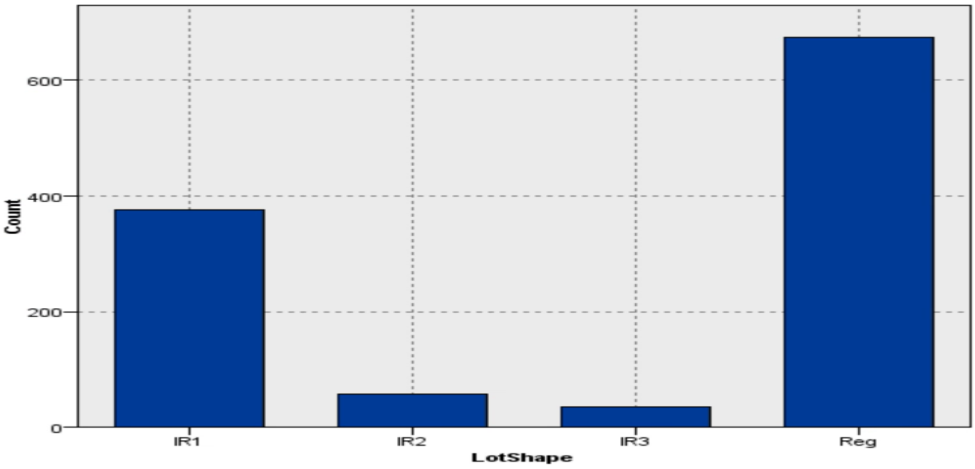
The above histogram shows the distribution of the target variable price. From the graph, it can be seen that majority house prices are concentrated between 100000 and 200000 .

Distribution of Above ground living area



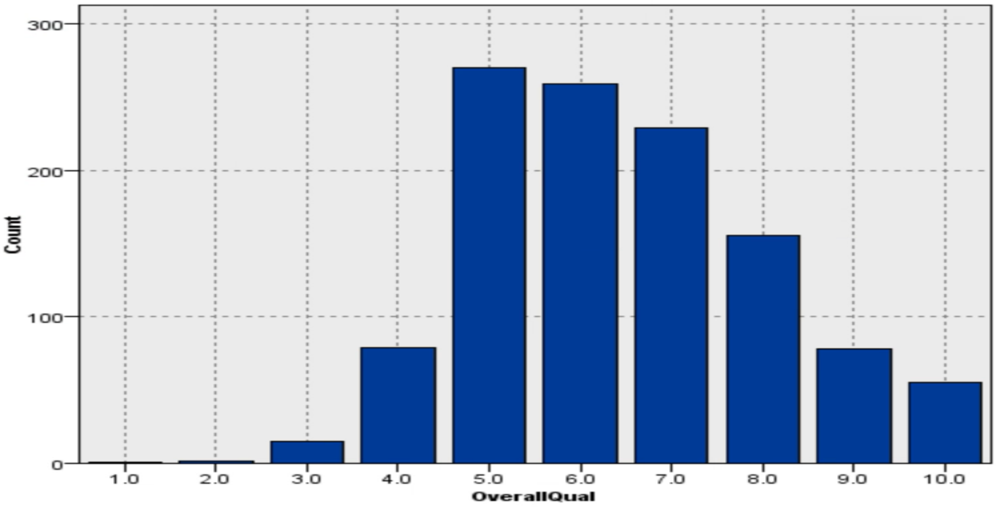
This is a distribution of the variable above ground living area, and it has a distribution similar to a normal distribution. Most houses have a ground living area between 1000 and 2000 square feet.

Lot shape



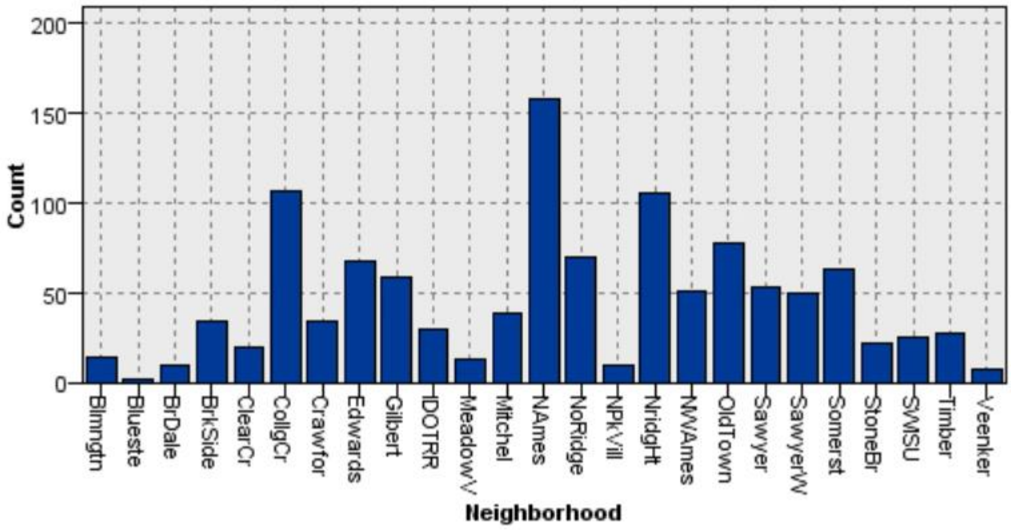
From the above bar graph of Lot shape, it can be seen that most houses in the area are of regular shape. More than 600 of the total number of houses are regular shaped ones.

Overall Quality



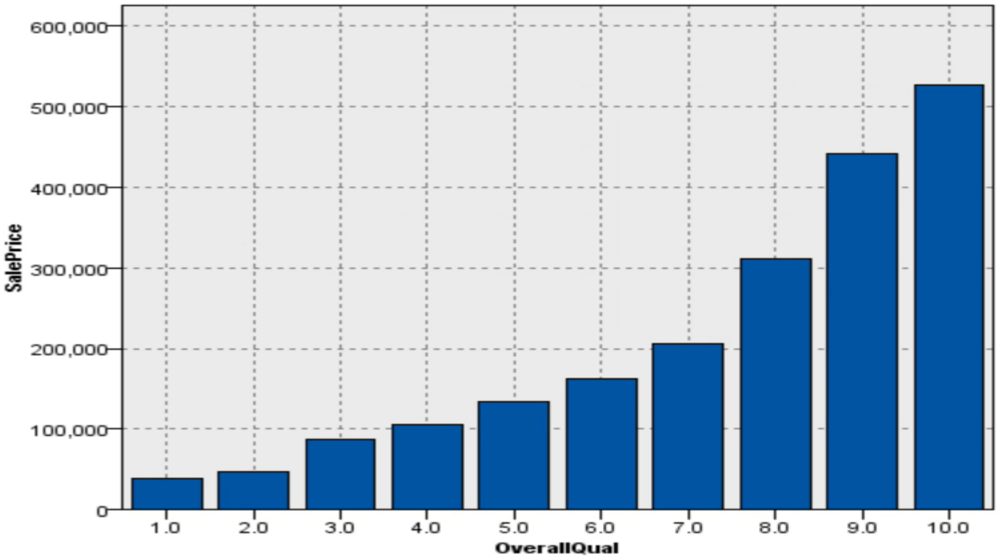
From the above graph, it can be seen that majority of the houses have an overall quality between 5 and 7.

Neighbourhood



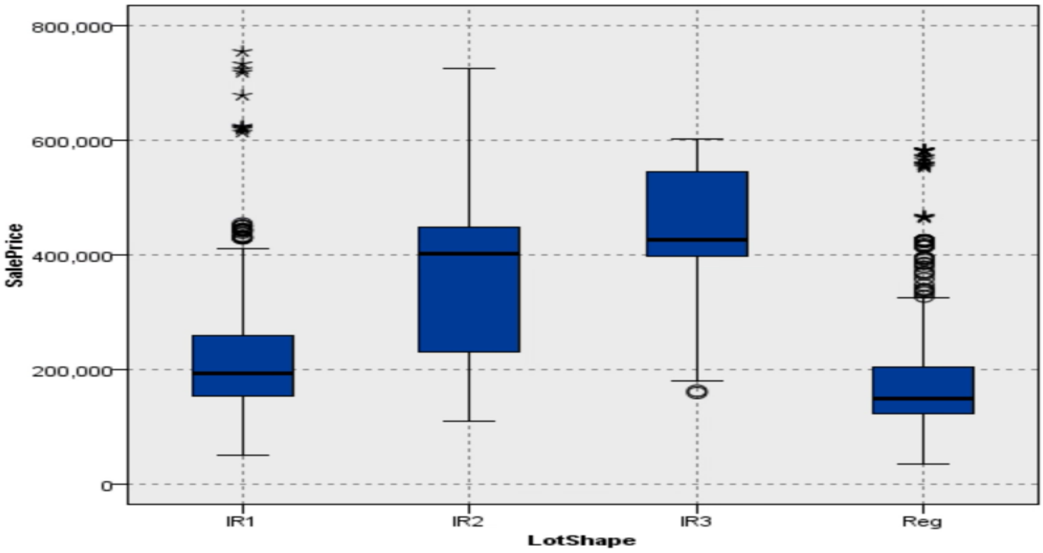
From the above bar plot of neighbourhood, it can be seen that highest number of houses are in the neighbourhood named **NAmes** followed by **Collgcr** and **NridHt**

Bi Variate Visualisation
Overall Quality vs Sales price



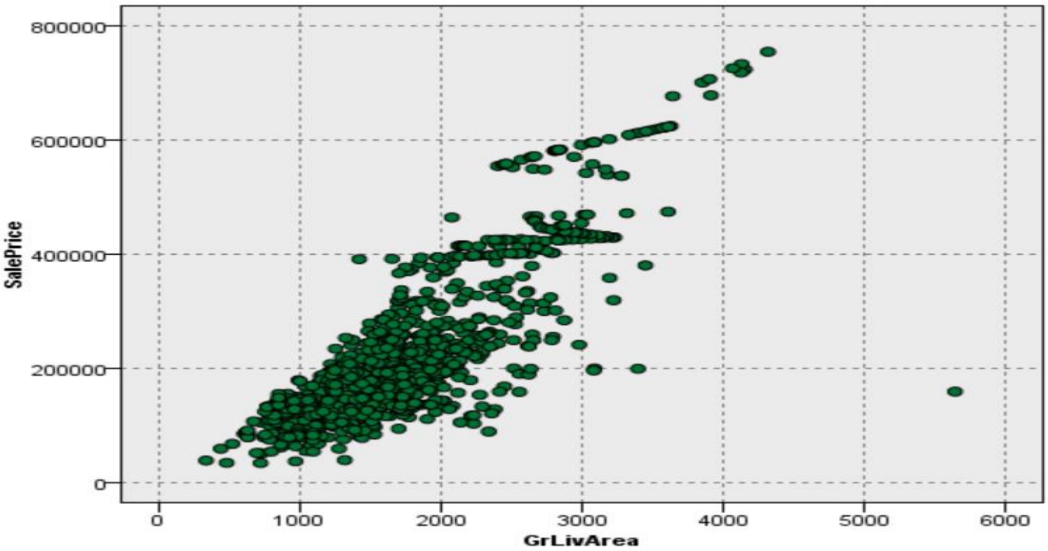
The above graph depicts the relationship between Sales price and Overall quality. From the graph it can be inferred that price of the house tends to increase with the rise in quality of the house. Overall quality has the potential to be a strong predictor of price.

Lot shape vs Sales price



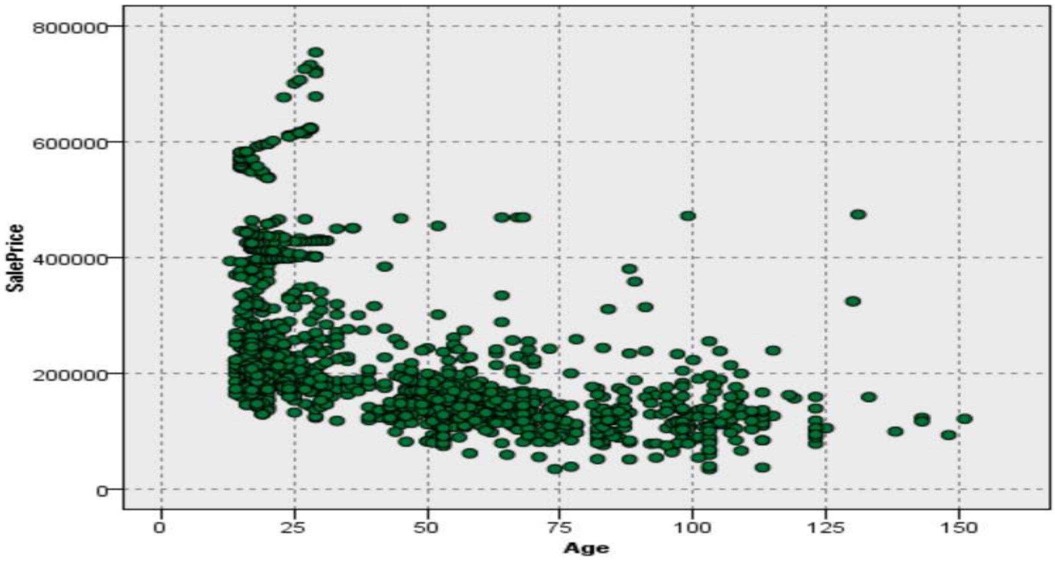
The relationship between Lot Shape and the Selling price is depicted on the graph. The highest median price in the data is found in IR3 (Irregular). Reg (Regular) and IR1(Slightly irregular) have some extreme values and outliers which may be due to the effect of other variable on the selling price of the house.

Ground living area vs Sales Price



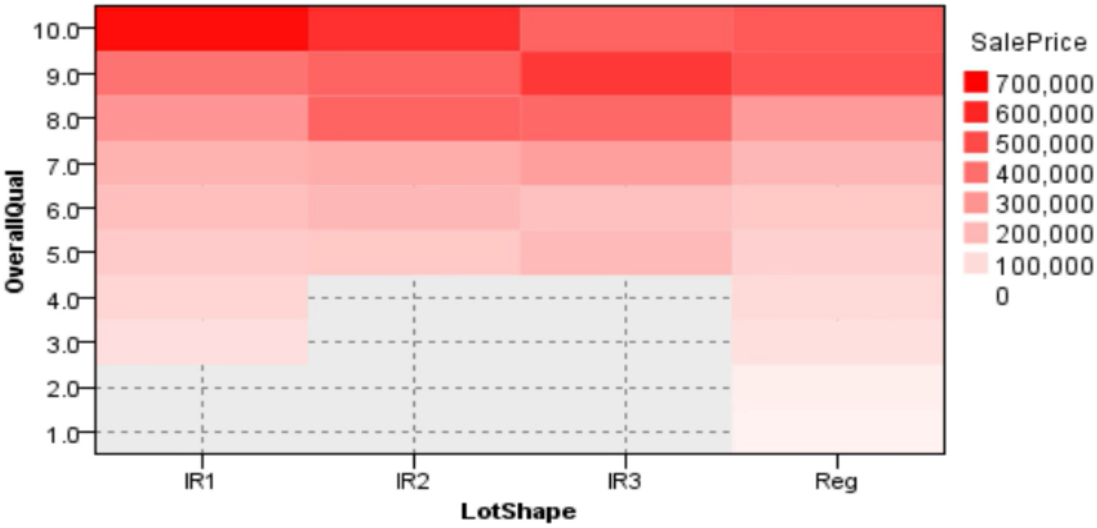
The above scatter plot shows the relation between two continuous variables Sales price and Above ground living area. The two graphs have a positive linear relationship. That is, with the increase in Above ground living area, selling price of the house tends to increase. As it has a strong relationship, it can be a a strong predictor of the target variable.

Age vs Sales Price



The scatter plot shows the relationship between Age of the house and Sales price. It can be seen that Age and selling price has an inverse relationship. That is, with the increase in Age of the house, selling price tends to decrease.

Lot Shape and Overall quality vs Sales Price



From the above Heatmap it can be seen that houses with the highest price have overall quality of 9 or 10 and slightly irregular shape.

Data Quality Assessment and treatment

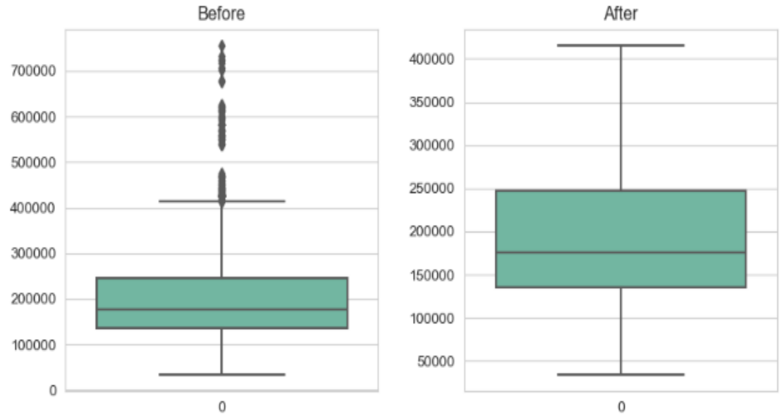
Outliers and extremes

Extreme values and outliers mean observations that vary highly from the rest of the observations in a dataset.

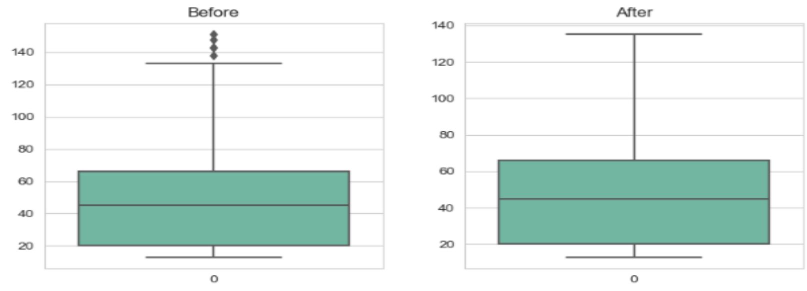
An observation that deviates significantly from the variable’s median is referred to as an outlier. An outlier, for instance, may be a house that is valued much higher or cheaper than other homes with comparable qualities in a dataset of housing prices. On the other hand, extreme values are observations that deviate significantly from a variable's normal distribution of values. Depending on how distant they are from the data's median value, these extreme numbers may or may not be considered outliers.

Outliers in this dataset is treated using the method of **Inter quartile range capping** and is done using python.

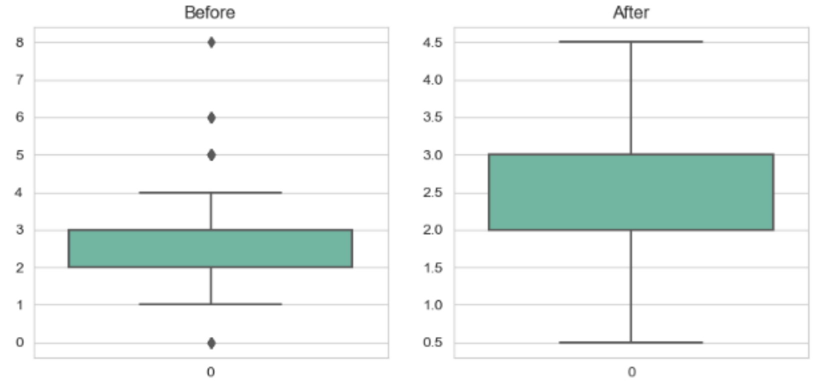
Sales price



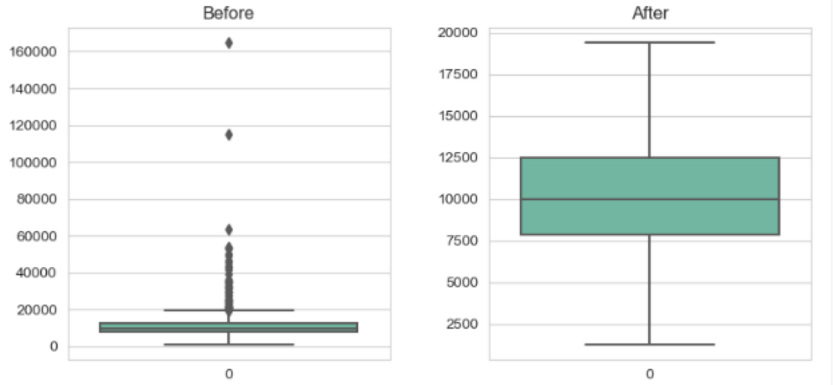
Age



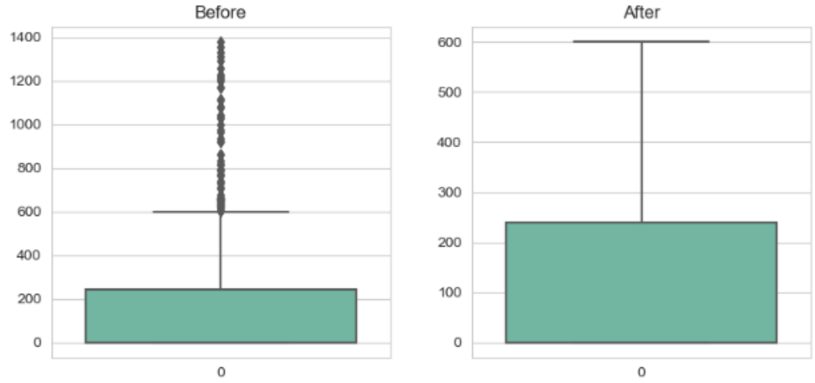
BedroomAbvGr



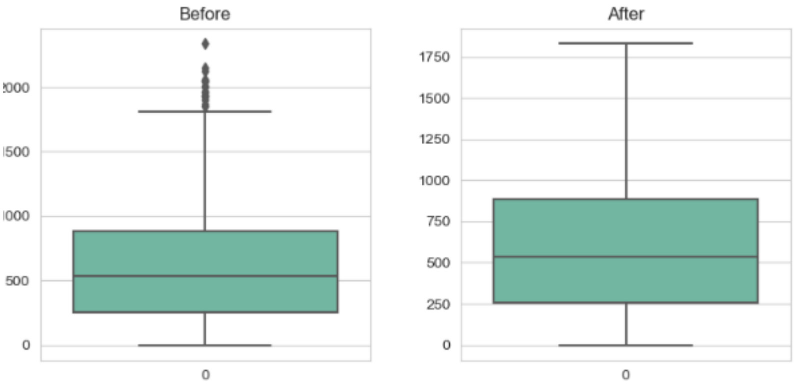
Lot area



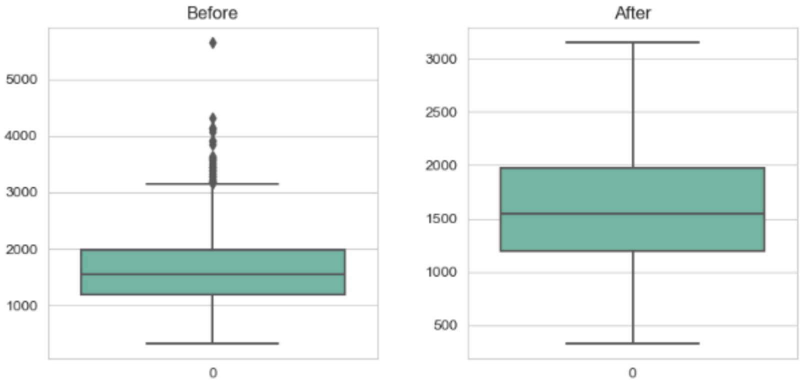
MasvnrArea



BsmtUmnfSF

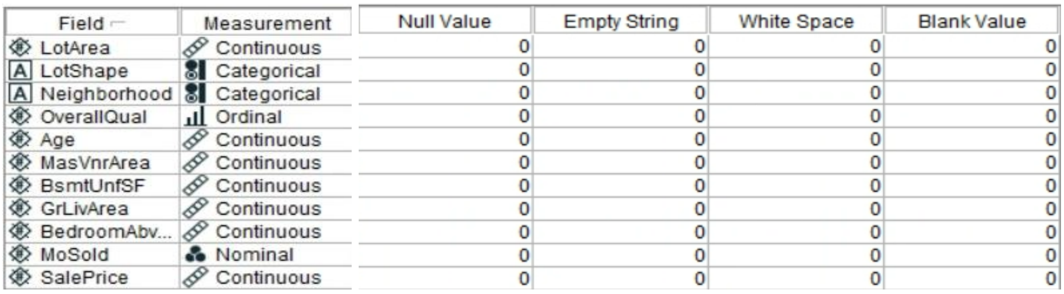


GrLivArea



Missing values

In the dataset we are working with, there are no missing values to be observed.



Field	Measurement	Null Value	Empty String	White Space	Blank Value
LotArea	Continuous	0	0	0	0
LotShape	Categorical	0	0	0	0
Neighborhood	Categorical	0	0	0	0
OverallQual	Ordinal	0	0	0	0
Age	Continuous	0	0	0	0
MasVnrArea	Continuous	0	0	0	0
BsmtUnfSF	Continuous	0	0	0	0
GrLivArea	Continuous	0	0	0	0
BedroomAbv...	Continuous	0	0	0	0
MoSold	Nominal	0	0	0	0
SalePrice	Continuous	0	0	0	0

Predictive model formulation

The problem we are dealing with is regarding the real estate market in Ames city. The main objective is to forecast the selling price of houses using various features of the house. We have a dataset with a total of 79 predictors or features of the house and we need to build a predictive model that accurately predicts the price of each house with the least possible error.

Incorporating all 79 predictors will lead to a very complicated model and hence it is necessary to filter out the variables. Therefore, we choose 10 variables after an initial modelling.

The predictive models we are using to address the issue are Decision tree, Linear regression, and neural network.

Type of the problem

As we are predicting the selling price of houses which is a numeric continuous variable, this is a regression problem.

We try to relationship between the target variable and predictors with the aim to predict the price future houses with maximum level of accuracy.

Data Partitioning

Data partitioning is one of the most important steps in predictive modelling. In this step we split the main dataset in to two parts namely training and testing set. We split the data in such a way that training set will have major proportion of the data.

Train data is used to train the model so that it learns from the data to catch patterns and find relationship between predictors and target variable.

After training the model, we use the test data to assess performance.

Reasons of splitting the data

- The main reason of splitting the dataset is to assess the performance of the model. Once the model is ready after training, we evaluate its performance on the testing data which acts as an unseen data to the model, assessing performance in this way will help us to understand how well the model works on future unseen data.
- Issue of overfitting: Overfitting occurs when the model performs well on the training data but does not generalise well on a new data. We can use the testing test to see if our model has issue of overfitting. After training the model, comparing the result with that of testing set will allow the user to understand if the model has over fitting issue.

Performance Metrices

Performance metrices are certain metrices that are used to evaluate how well a predictive model performs. There are various performance metrices that can be used which will depend on the type of problem we are dealing with.

As the problem we are dealing with is a regression problem, there are mainly three performance metrices namely, R square value, Mean Absolute Error, Root Mean Square Error.

R square value: This metric refers to the proportion of variation in the dependent variable explained by the independent variable or predictors. If there are more multiple independent variables, it is better to use adjusted R square value as it also takes into account the number of independent variables as well. Value of R square lies between 0 and 1. Higher value of R square means that the model explains the variability to a higher extend and hence has a better goodness of fit.

Mean Absolute Error: This metric measures the average of absolute difference between actual value and predicted value off the target variable.

Root Mean Square Error: This metric measures the root of average of squared difference between actual value and predicted value of the target variable. RMSE gives more weight to larger errors as it squares the errors before taking average.

Baseline modelling (Naïve model)

Baseline model is the least sophisticated model of the dataset which is used to compare performance with other more sophisticated models.

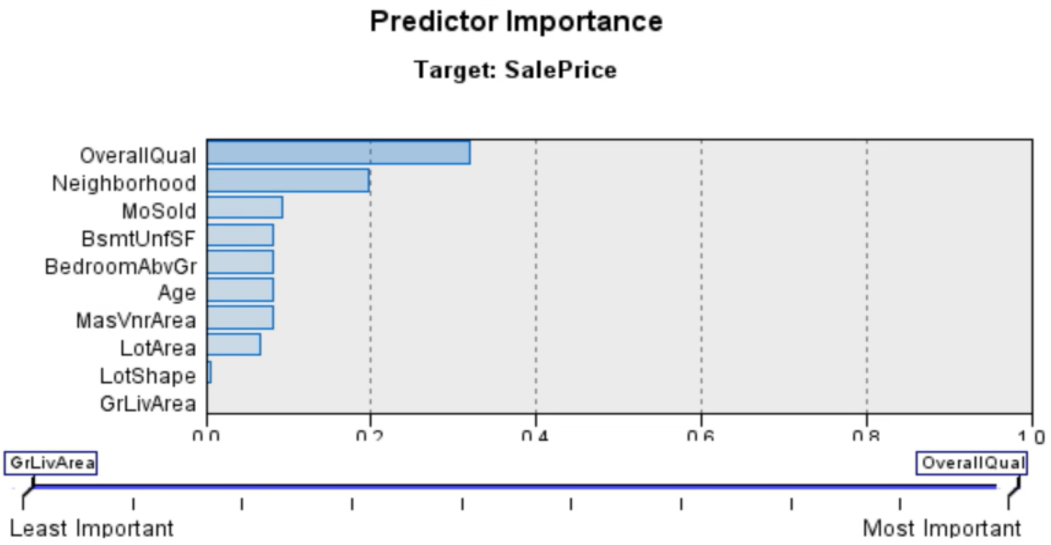
The mean of the dataset was 206053.9 and MAE of the naïve model obtained is 80186 which acts as the baseline for all the future models to be build.

Performance on the testing set.

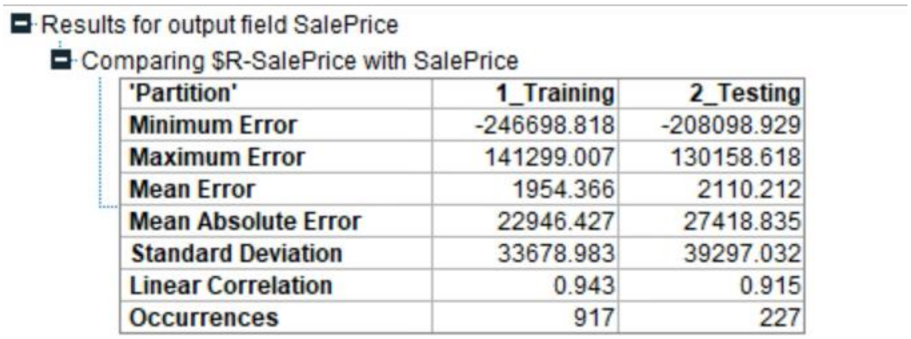
After building the naïve on the data set and testing its performance on testing set, we get An MAE of **77755.4**

Predictive Modelling

Decision Tree Modelling



From the above Predictor importance graph, we can see that overall quality is the predictor that has the highest potential to predict selling price. In a real-life scenario this case is true. Neighbourhood of the house also has high prediction importance which can be due to the fact buyers of the house does consider this the neighbourhood of the house before buying the house. Other predictors have almost importance predictor importance according to the model.



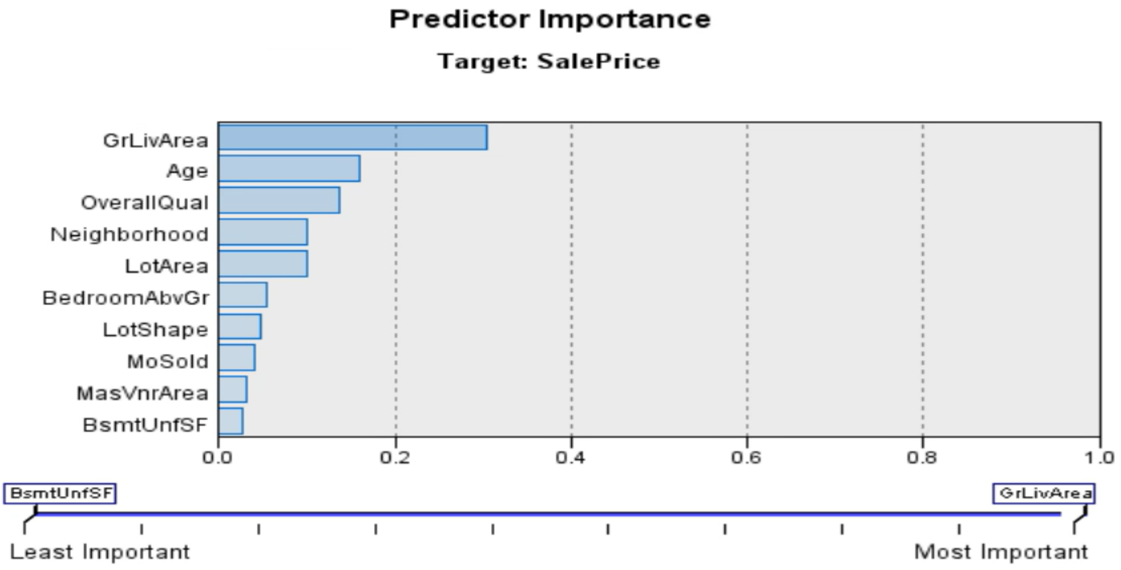
Results for output field SalePrice

Comparing $R-SalePrice with SalePrice

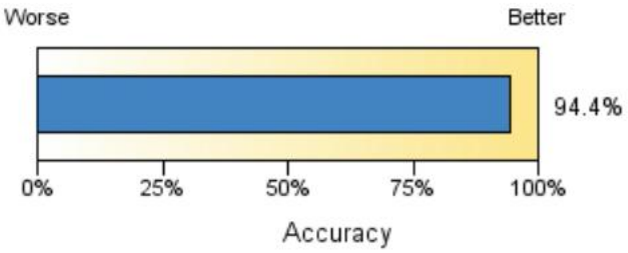
'Partition'	1_Training	2_Testing
Minimum Error	-246698.818	-208098.929
Maximum Error	141299.007	130158.618
Mean Error	1954.366	2110.212
Mean Absolute Error	22946.427	27418.835
Standard Deviation	33678.983	39297.032
Linear Correlation	0.943	0.915
Occurrences	917	227

The decision tree has Mean Absolute error of 22946.427 for training set and 27418.835 for testing set.

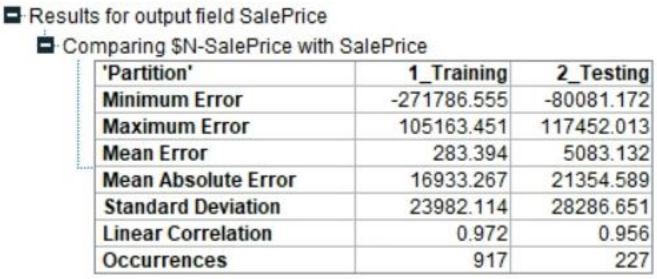
Neural Network Modelling



From the above graph of predictor importance of neural network it can be seen that Abouve ground living area has the highest predictor importance in predicting sales price followed by age of the house and overall quality. Customers will look at the living area size before offering a price for the house and hence it can be an important feature.



It can also be seen that this model has an accuracy of 94.4%.



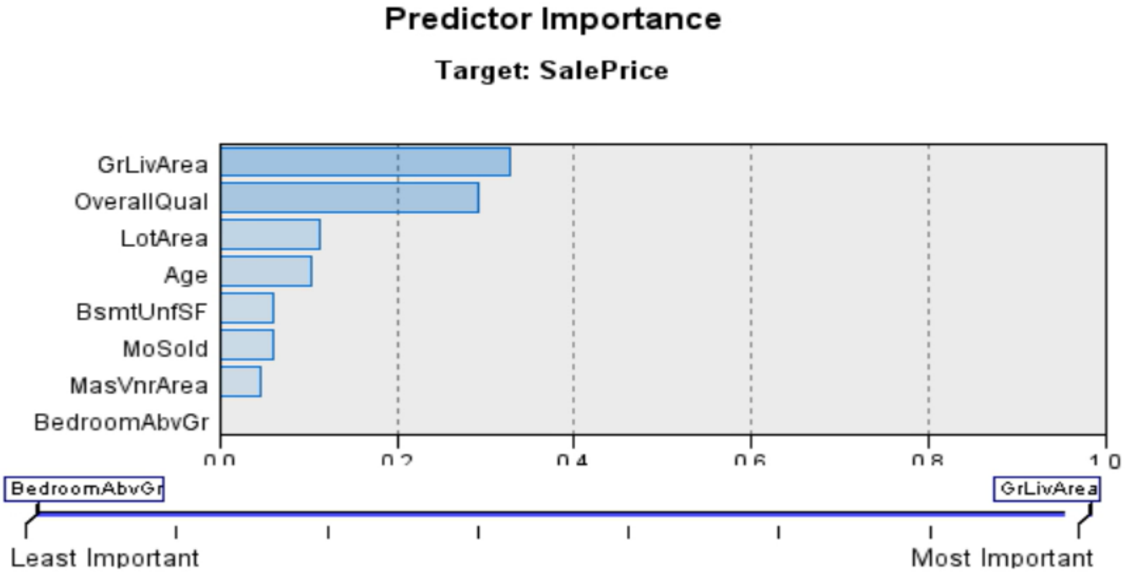
Results for output field SalePrice

Comparing $N-SalePrice with SalePrice

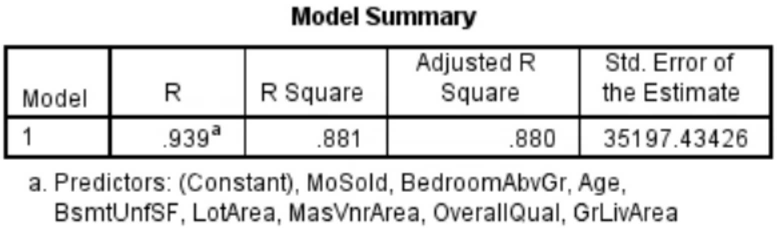
'Partition'	1_Training	2_Testing
Minimum Error	-271786.555	-80081.172
Maximum Error	105163.451	117452.013
Mean Error	283.394	5083.132
Mean Absolute Error	16933.267	21354.589
Standard Deviation	23982.114	28286.651
Linear Correlation	0.972	0.956
Occurrences	917	227

The model has a MAE of 16933.26 and 21354.58 for training and testing respectively.

Linear Regression Modelling



The predictor importance graph of regression model shows a similar predictor importance graph as that of neural network and decision tree except for the fact that it has lot area as an important predictor.

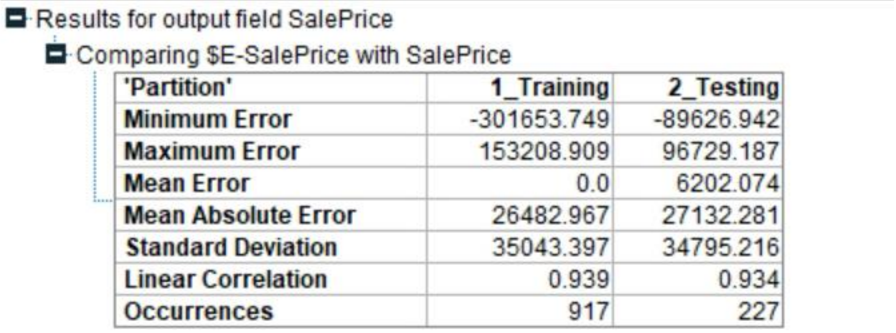


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.939[a]	.881	.880	35197.43426

a. Predictors: (Constant), MoSold, BedroomAbvGr, Age, BsmtUnfSF, LotArea, MasVnrArea, OverallQual, GrLivArea

Adjusted R square value of 88 % means that 88 percent of variability in target variable(sales price) is explained by the predictor.



Results for output field SalePrice

Comparing $E-SalePrice with SalePrice

'Partition'	1_Training	2_Testing
Minimum Error	-301653.749	-89626.942
Maximum Error	153208.909	96729.187
Mean Error	0.0	6202.074
Mean Absolute Error	26482.967	27132.281
Standard Deviation	35043.397	34795.216
Linear Correlation	0.939	0.934
Occurrences	917	227

The model has a MAE of 26482.967 and 27132.28 for training and testing set respectively.

Feature Engineering

The one feature engineering used in building this model is calculating the age of house based on the year in which it was built. This was useful and necessary as we cannot directly use year as a useful variable but if we convert it into age, we can use it as a continuous variable and is a potential key predictor.

Effort to improve the model.

Model	Hyperparameter	Error
Decision tree	Default	Training set = 22946.427 Testing set = 27418.835
Decision tree	Maximum tree depth = 8	Training set = 22020.95 Testing set = 26109.81
Decision tree	Maximum tree depth = 6 Parent branch =3.2 Child branch = 0.5	Training set= 19967.374 Testing set = 25725.76
Decision tree	Maximum tree depth = 10 Parent branch =2.2 Child branch = 1.0	Training set= 22020.95 Testing set = 26109.81
Linear Regression	Default	Training set = 26482.967 Testing set = 27132.28
Neural network	Default	Training set = 16933.26 Testing set = 21354.58
Neural network	Hidden layer 1 = 2 Hidden layer 2 = 1	Training set = 21190.61 Testing set = 22034.76
Neural network	Hidden layer 1 = 2 Hidden layer 2 = 1	Training set = 22036.67 Testing set = 24638.11
Neural network	Hidden layer 1 = 2 Hidden layer 2 = 1	Training set = 18131.65 Testing set = 19709.96

The three main modelling techniques used for predicting sales price are Decision tree, linear regression, and neural network.

After doing the modelling on default parameters and changing hypermeters based on trial-and-error method we can see that for decision tree model the best result was obtained with hypermeters set at

maximum tree depth = 16, parent branch =3.2 and child branch = 0.5

MAE Training set= 19967.374

MAE Testing set = 25725.76

Linear regression does not have any parameters and there is only one model.

Best model obtained from neural network have hypermeters set at

Hidden layer1 = 2 and Hidden layer2 = 2

MAE Training set = 18131.65

MAE Testing set = 19709.96

This is the best model out of the 10-model made.

Models used.

Decision Tree

It is a supervised learning method used for both classification and regression problems. A root node, branches, internal nodes, and leaf nodes make the tree.

It is a tree like structure that gives possible outcomes based on given inputs. The tree tries to create homogeneous subsets based on the given inputs. It starts at root node and splits in to two or more branches which gives possibility of each outcome until it reaches a decision. At the end leaf nodes shows all the possible final outcomes.

Neural Network

Neural network is a machine learning algorithm that are arranged in layers to process input data into predictions or classifications as output.

As the network being trained patterns of information are sent into it via the input units, which then cause the layers of hidden units to be triggered, which then bring the patterns of information to the output units. The inputs are multiplied by the weights of the connections they travel along as they are received by each unit from the units to its left.

Linear Regression

Linear regression is type of supervised machine learning method used to find the linear relationship between a target variable and a set of predictors or independent variables. Finding the line of best fit that minimises the difference between the predicted values and the actual values is the objective of linear regression.

Error cost analysis

- **Property buyers and Sellers:** Error cost analysis can help people who are looking to buy and sell their property to analyse the model's error while they use a predictive model to predict the sale price of a property. The buyer can lose out on the chance to own the property for a lesser price and at the same time sellers can face lose if the model underestimates the property. The buyer can overpay for the property if the model overestimates the sale price and sellers can profit from it.
- **Investors in real estate market**: Apart from the error cost of property buyers and sellers, investors can also learn more about the possible expenses of long-term property holding by using error cost analysis. An investor may pay extra for maintenance and other expenses if the model overestimates the sale price, and the property stays on the market longer than expected.
- **Real estate brokers**: Error cost analysis can help brokers assess the risk of recommending a property to their clients. Convincing client to buy an overestimated property can affect the brokers credibility and conversely providing the client with an underestimated property can improve his commission and business.
- **Property developers**: Property developers can use error cost analysis to understand the models fault before using the model with confidence. For predicting price, the developer can lose out on the chance to sell the property for more money if the model, for instance, underestimates the sale price. In contrast, if the model overestimates the sale price, the developer can wind up spending too much money on the property and not getting the projected return.
- **Financial Organisations**: They can over lend if the model over estimate the price of a property and the same time there is a chance that they will lend less money if the model underestimates the price which will cost them their profit
- **Government**: Government may miss out on revenue in the form of tax if the model underestimates price and conversely get more revenue if the model overestimates price

Which error is the worst and why

As this is a business problem, both kinds of error costs are almost equally important. While overestimating or underestimating the price is beneficial to one party, it becomes worse for the opposite party in that particular business transaction.

Conclusion

After finishing the report, we were able to predict the selling price of houses. The report mainly contains a visualisation part which helps us to understand how various factors affect the price of a house. From looking at the graphs in the report, it can be seen that age, overall quality of the house, ground living area and lot area have a huge impact on the price of the house. After analysing the results of each models build, it can be seen that neural network is the best model that predict the price of house more accurately. Hence using this model will help minimise the error.

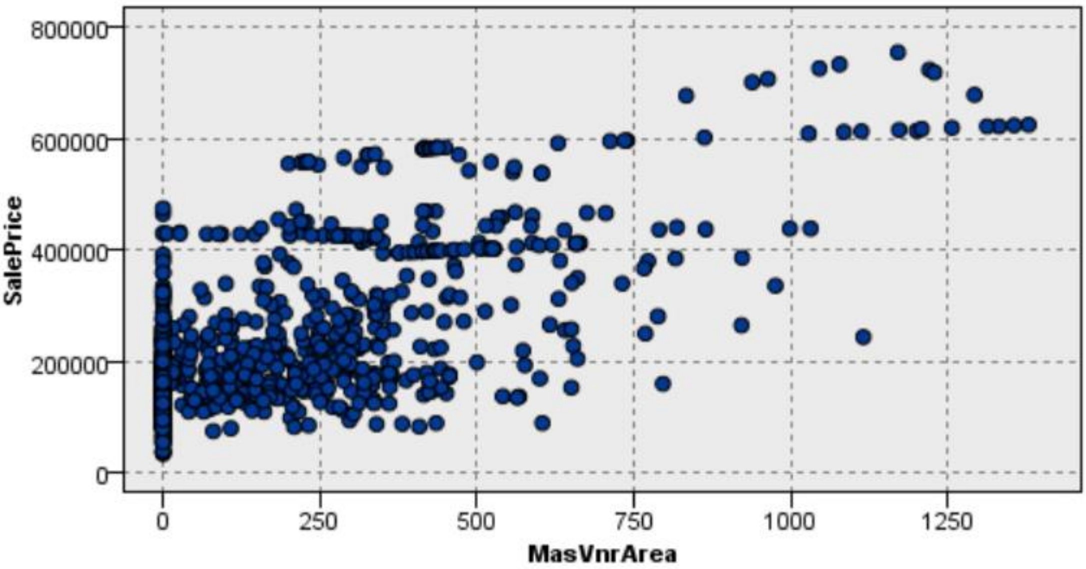
Steps to improve the performance of the model

- More data which means more information for the model to learn from and hence better model.
- Ensemble method: Combining various models which will improve the predictability.

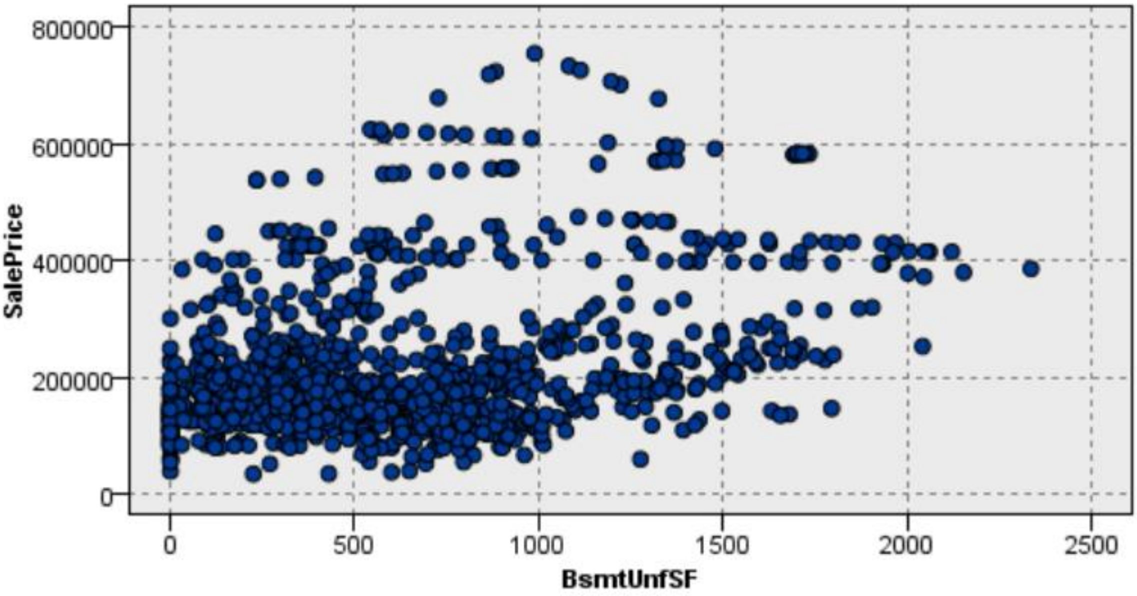
Appendix

Additional Bi variate and Uni variate visualisations

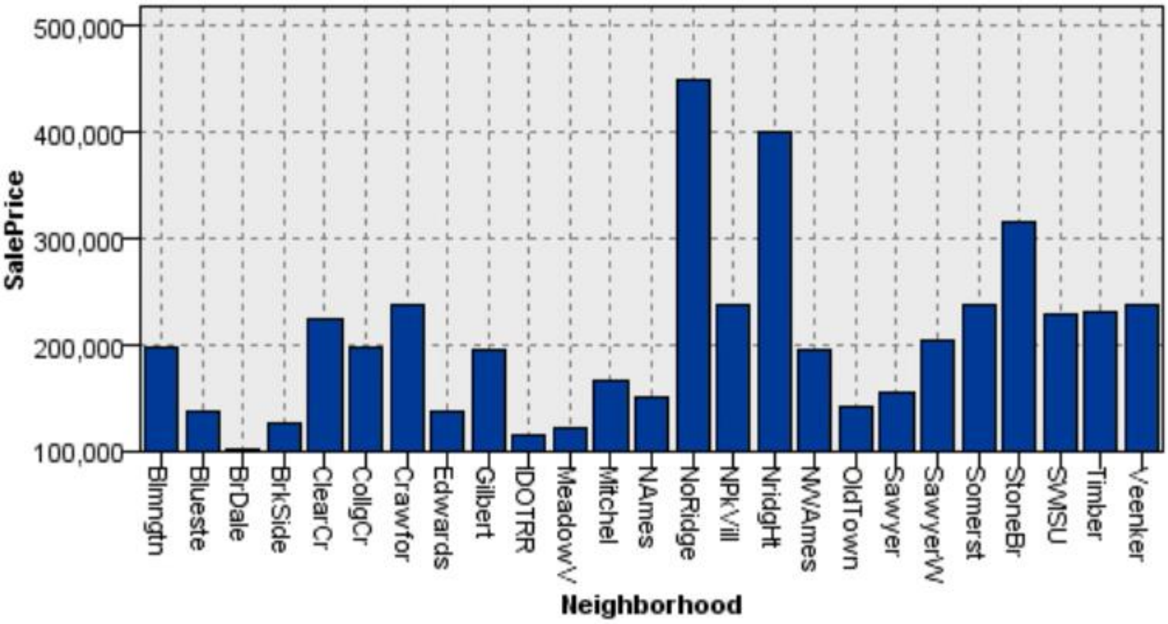
MasVnrArea vs Sales Price



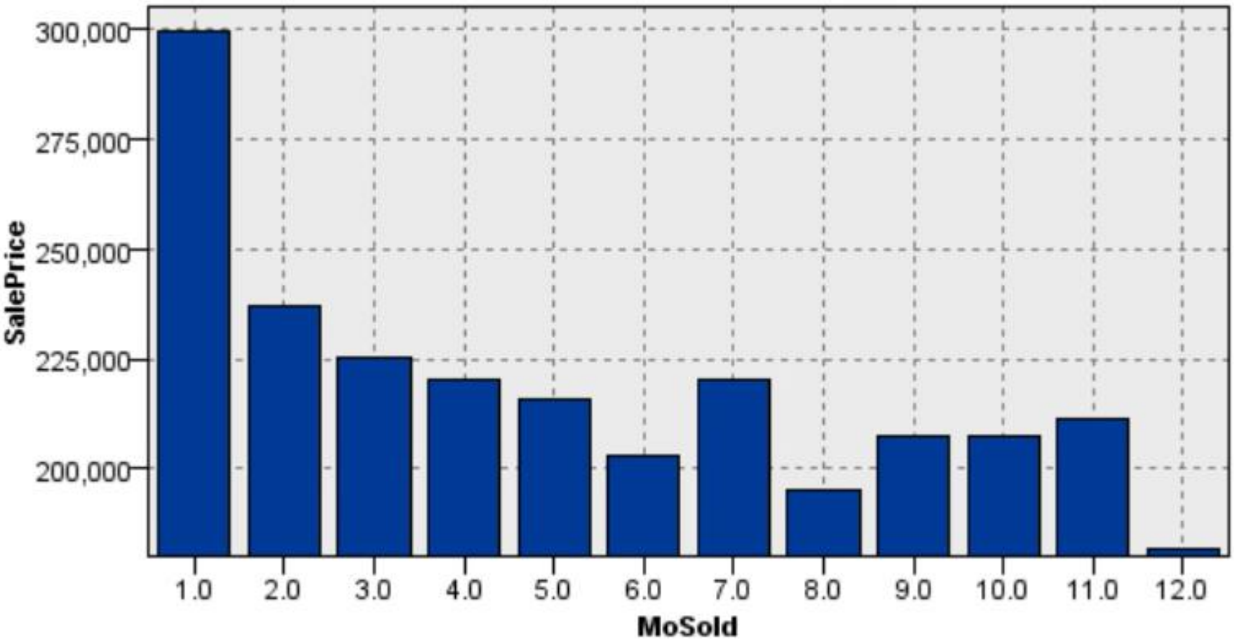
BsmtUnfsF vs Sales Price



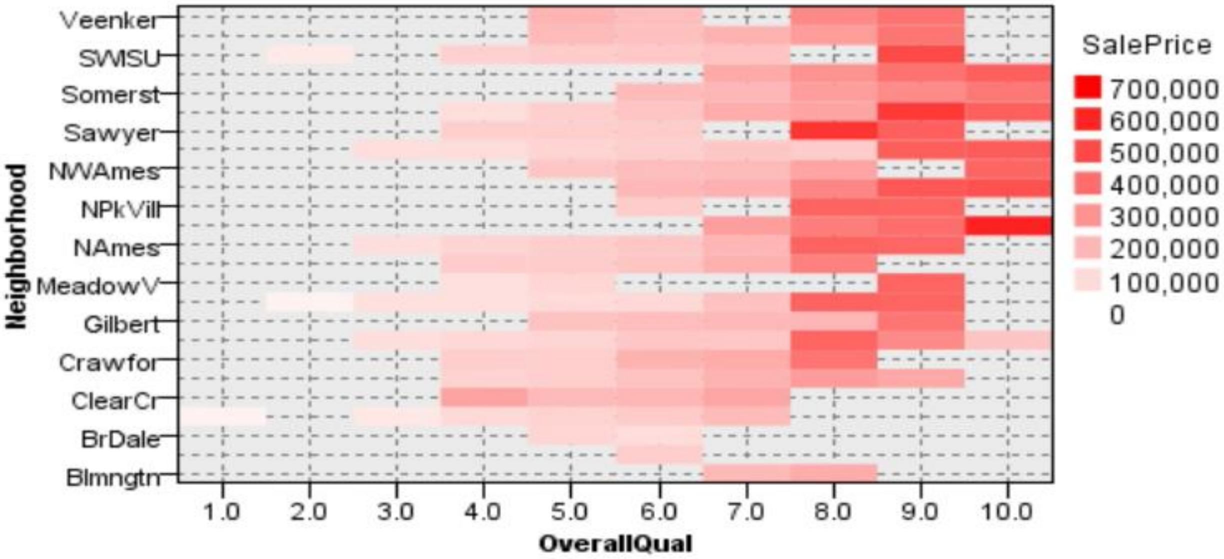
Neighbourhood vs Sales Price



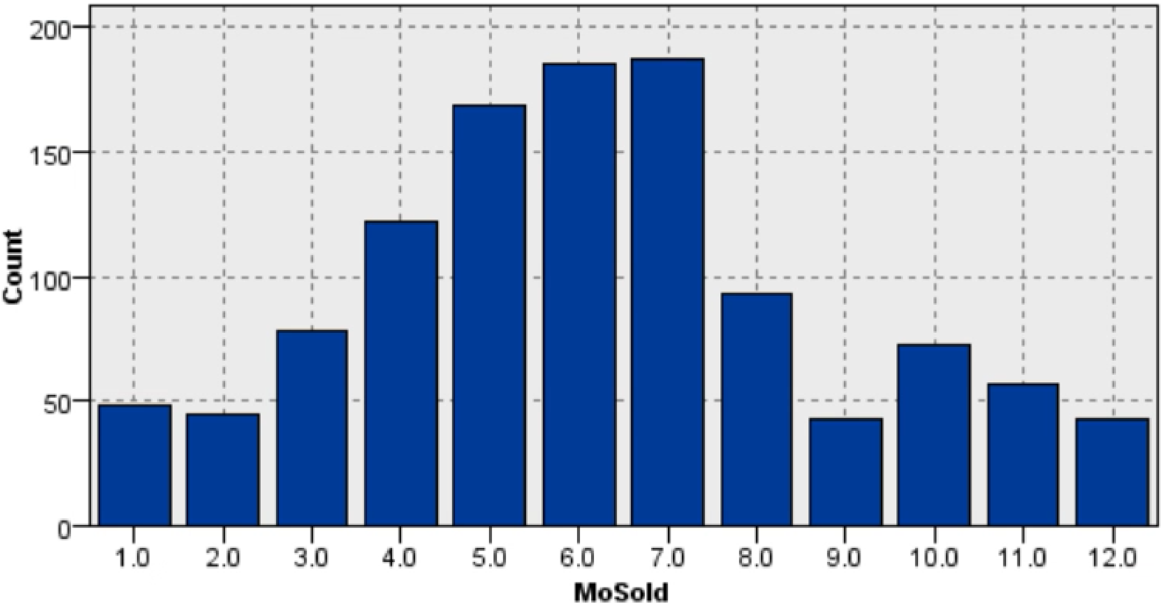
Month sold vs Sales Price



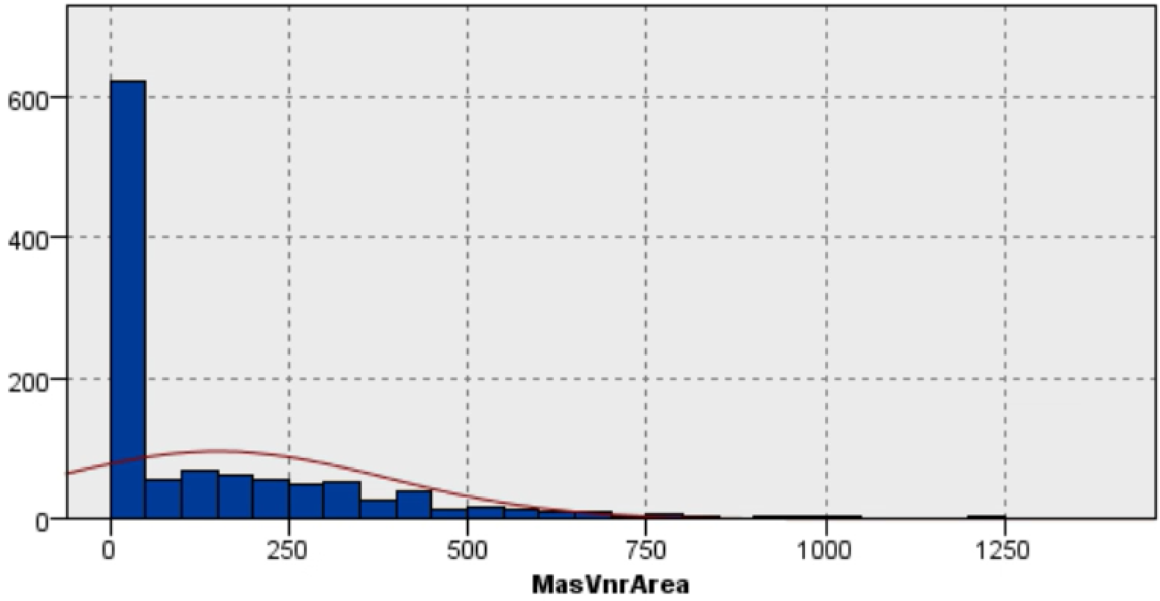
Overall quality and Neighbourhood vs Sales Price



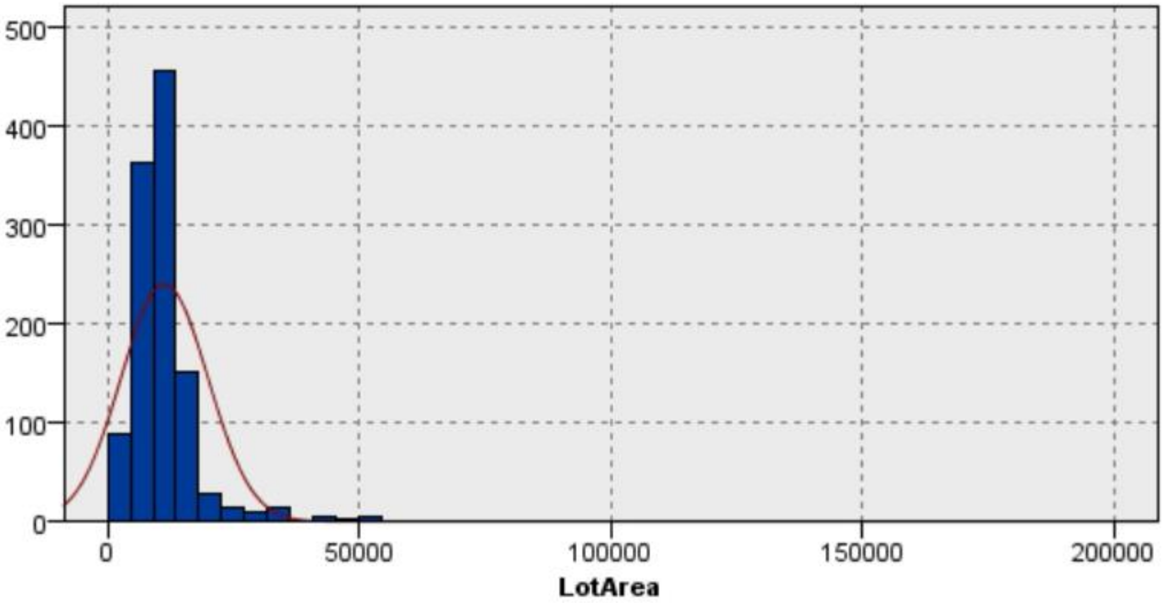
Month Sold



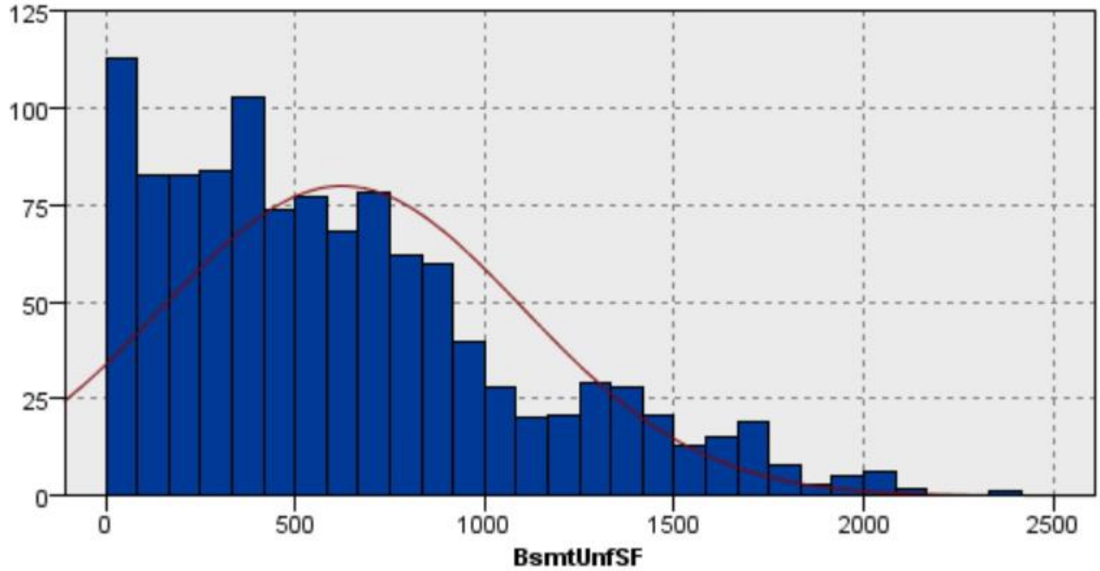
MasVnrArea



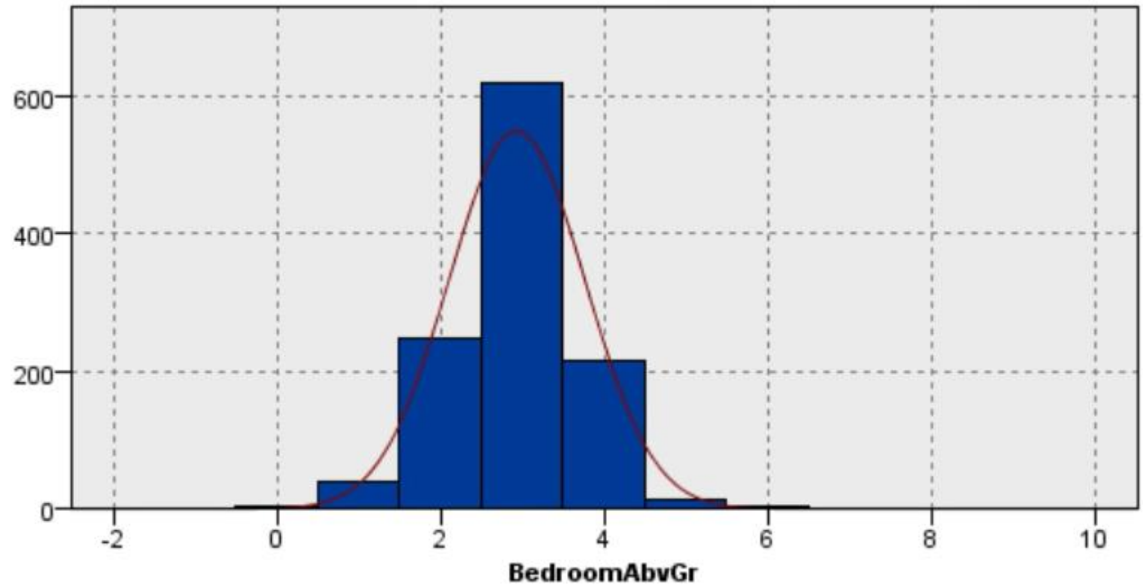
Lot Area



BsmtUnfsF



BedroomAbvGr


BedroomAbvGr