Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

House Price Prediction

Supervised By:

Mr.Shubham Singhal

Submitted By:

Neeshu, 2210990603 (G-11)

Nandini, 2210990597 (G-11)

Muskan, 2210990588 (G-11)

Department of Computer Science and Engineering
Chitkara University Institute of Engineering & Technology,
Chitkara University, Punjab

ABSTRACT

Data mining is now commonly used in the real estate market. Real Estate is a clear industry in our ecosystem. The ability to extract data to extract relevant information from raw data makes it very useful to predict house prices, important housing features, and much more. Housing prices continue to change from day to day and are sometimes raised rather than based on calculations. Research has shown that fluctuations in housing prices often affect homeowners and the housing market. Literature research is done to analyze the relevant factors and the most effective models for predicting housing prices. The findings of this analysis confirmed the use of Artificial Neural Network, Support Vector Regression, and Linear Regression as the most efficient models compared to others. In addition, our findings also suggest that spatial and real estate agents are key factors in predicting house prices. This study will be of great benefit, especially to housing developers and researchers, to find the most important criteria for determining housing prices and identify the best machine learning model used to conduct research in this field.

This paper presents a data set describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values.

INDEX

1. INTRODUCTION

Thousands of houses are sold every day. There are some questions every buyer asks himself like: What is the actual price that this house deserves? Am I paying a fair price? In this paper, a machine learning model is proposed to predict a house price based on data related to the house (its size, the year it was built in, etc.). During the development and evaluation of our model, we will show the code used for each step followed by its output. This will facilitate the reproducibility of our work. In this study, Python programming language with a number of Python packages will be used.

In this report, we propose our system "House price prediction". House is one of human life's most essential needs, along with other fundamental needs such as food, water, and much more. Demand for houses grew rapidly over the years as people's living standards improved. House price prediction can be done using multiple prediction models (Machine Learning Model) such as support vector regression, artificial neural network, etc. There are many benefits that home buyers, property investors, and housebuilders can reap from the house-price model. This model will provide a lot of information and knowledge to home buyers, property investors, and housebuilders, such as the valuation of house prices in the present market, which will help them determine house prices. The target feature in this proposed model is the price of the real estate property and the independent features are: no. of bedrooms, carpet area, the floor, car parking, and lift availability. The whole implementation is done using the python programming language.

1.1 Background

The background of a house price prediction model encompasses a multifaceted understanding of real estate economics, data science methodologies, and predictive modeling techniques. In real estate economics, factors such as location, property characteristics, market trends, and economic indicators significantly influence house prices. Understanding these factors helps in selecting relevant features and building accurate prediction models.

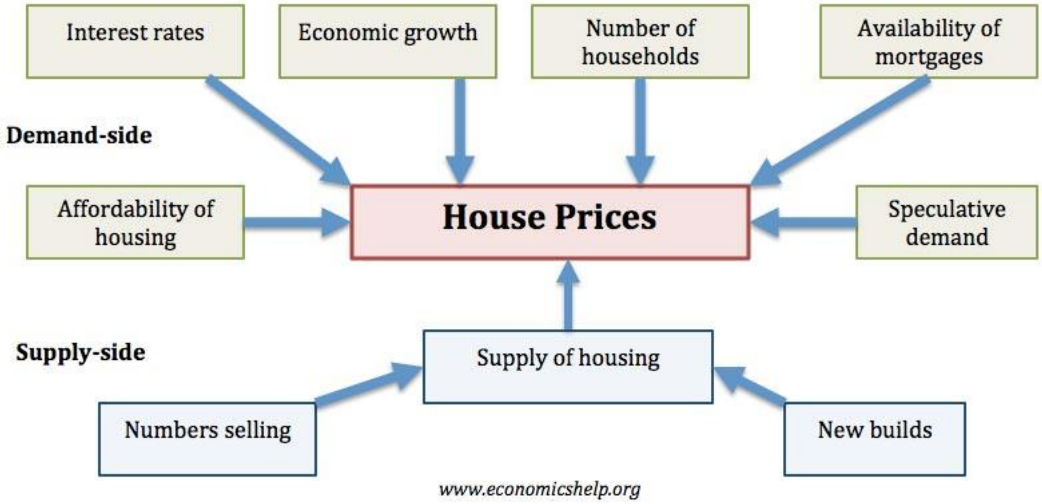


Figure 1.1.1 factors effecting house pricing

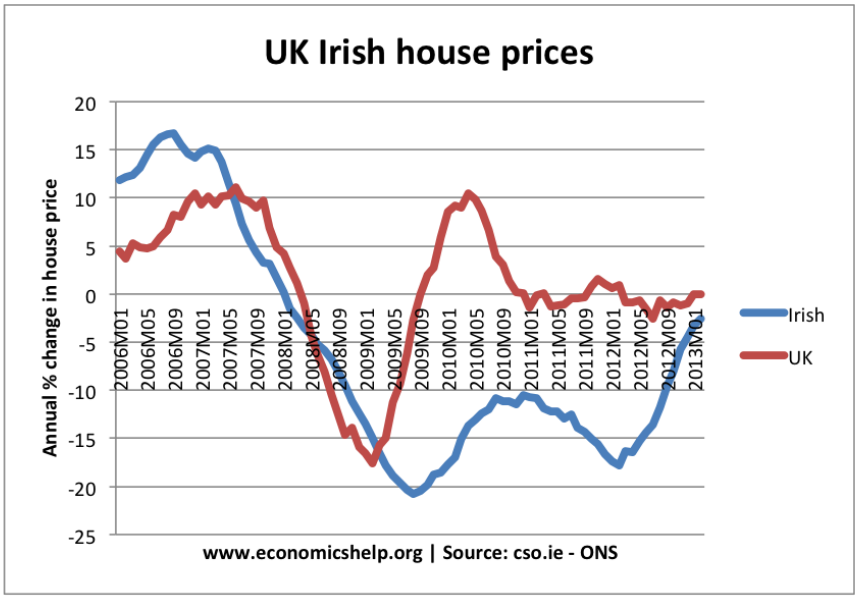
1.1.1 Economic growth: Demand for housing is dependent upon income. With higher economic growth and rising incomes, people will be able to spend more on houses; this will increase demand and push up prices.

1.1.2 Unemployment: Related to economic growth is unemployment. When unemployment is rising, fewer people will be able to afford a house. But, even the fear of unemployment may discourage people from entering the property market.

1.1.3 Interest rates: Interest rates affect the cost of monthly mortgage payments. A period of high- interest rates will increase cost of mortgage payments and will cause lower demand for buying a house.

1.1.4 Consumer confidence: Confidence is important for determining whether people want to take the risk of taking out a mortgage. In particular expectations towards the housing market is important; if people fear house prices could fall, people will defer buying.

1.1.5 Supply: A shortage of supply pushes up prices. Excess supply will cause prices to fall. For example, in the Irish property boom of 1996-2006, an estimated 700,000 new houses were built. When the property market collapsed, the market was left with a fundamental oversupply. Vacancy rates reached 15%, and with supply greater than demand, prices fell.



1.2 Objective

People looking to buy a new home tend to be more conservative with their budgets and market strategies. This project aims to analyses various parameters like average income, average area etc. and predict the house price accordingly. This application will help customers to invest in an estate without approaching an agent. To provide a better and fast way of performing operations.

To provide proper house price to the customers. To eliminate need of real estate agent to gain information regarding house prices. To provide best price to user without getting cheated. To enable user to search home as per the budget. The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and also upcoming developments future prices will be predicted. House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. We use linear regression algorithm in machine learning for predicting the house price trends.

1.2 Significance

A house price prediction model holds significant importance for stakeholders across the real estate market. It empowers homebuyers and sellers by providing accurate estimations, aiding in informed decision-making and ensuring fair transactions. Real estate investors rely on these models to identify profitable opportunities, manage risks, and optimize returns. Additionally, policymakers and economists utilize prediction models to analyze market trends, forecast developments, and formulate housing policies. Financial institutions leverage these models for risk assessment in mortgage lending, enhancing portfolio management and mitigating market risks. Real estate professionals, including agents and negotiators, utilize prediction models to guide negotiations, advocating for their clients' interests effectively. By fostering transparency, these models contribute to a more efficient and competitive real estate market, reducing information asymmetry and empowering consumers. Ultimately, the significance of a house price prediction model extends beyond individual transactions, shaping broader economic, social, and financial dynamics within the real estate ecosystem.

2.PROBLEM DEFINATION AND REQUIREMENTS

2.1 Problem statement

The general and standardized real estate characteristics are often listed separately from the asking price and general description. Because these characteristics are separately listed in a structured way, they can be easily compared across the whole range of potential houses. Because every house also has its unique characteristics, such as a particular view or type of sink, house sellers can provide a summary of all the important features of the house in the description.

All given real estate features can be considered by the potential buyers, but it is nearly impossible to provide an automated comparison of all variables due to the large diversity. This is also true in the other direction: house sellers have to estimate the value based on its features in comparison to the current market price of similar houses. The diversity of features makes it challenging to estimate an adequate market price.

Apart from providing a summary of the important features of the house, the house description is also a means of raising curiosity in the reader, or in other words persuading the person. Housing prices are an important reflection of the economy, and housing price ranges are of great interest to both buyers and sellers. In this project, house prices will be predicted given explanatory variables that cover many aspects of residential houses. The goal of this project is to create a regression model that can accurately estimate the price of the house given the features.

2.2 Hardware Requirements

The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware. A hardware requirements list is often accompanied by a hardware compatibility list, especially in case of operating systems. The minimal hardware requirements are as follows:

PROCESSOR PENTIUM IV

RAM: 8 GB

PROCESSOR: 2.4 GHZ

MAIN MEMORY: 8GB RAM

PROCESSING SPEED: 600 MHZ

HARD DISK DRIVE: 1TB

KEYBOARD :104 KEYS

2.3 Software Requirements

Software requirements deals with defining resource requirements and prerequisites that needs to be installed on a computer to provide functioning of an application. These requirements are need to be installed separately before the software is installed. The minimal software requirements are as follows:

FRONT END: PYTHON

IDE: ANACONDA

OPERATING SYSTEM: WINDOWS 10

2.4 Feasibility Features

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. The feasibility study of the proposed system is carried out. It is carried out to ensure that the proposed system is not a burden to the company. Economic feasibility

1. Economic feasibility
2. Technical feasibility
3. Social feasibility

2.4.1 Economical Feasibility

This study is generally carried out to check whether right amount of funds is invested in the model. This study is done to eliminate excess amount of money poured into a single model. It makes sure whether the model is well within the budget. It is extremely important to spend only right amount of funds to a model.

2.4.2 Technical Feasibility

It makes sure whether the technical requirements are limited to what we can offer. Any system developed should not have high demand on technical resources since it puts burden on client, It also checks the projects potential what it can do once developed.

2.4.3 Social Feasibility

It is carried out check how a system acts with other systems. It checks the level of acceptance of the system by the user. It trains the user to use the system efficiently. it is a necessity. Since a client is the final user of the system, he can criticize the system but it should be in a disciplined and meaningful manner.

2.5 Dataset

In this study, we will use a housing dataset presented by De Cock (2011). This dataset describes the sales of residential units in Ames, Iowa starting from 2006 until 2010. The dataset contains a large number of variables that are involved in determining a house price.

3.PROPOSED DESIGN / METHODOLOGY

In our house price prediction AI/ML project, various machine learning models are utilized, depending on factors such as the size and complexity of the dataset, the type of features available, and the desired level of prediction accuracy.

3.1 METHODS

3.1.1 Cleaning data

Machine learning algorithms are largely implemented to only take data that is in a numeric format as input. More than half of the columns in the Ames Housing data set are non-numerical and need to be encoded, in this case using one-hot encoding and labeling. Additionally, various columns contain some empty values that have been dealt with in different ways as described in section.

3.1.2 Encoding categorical data

Many of the variables of the data set are categorical, and take on a limited set of values. One example is the nominal variable ”Street” which represents the type of road access to the property and takes on the values ”Grvl” for gravel and ”Pave” for paved. Such categorical values can not be interpreted by conventional machine learning algorithms without preprocessing them to a numerical format. There are two types of categorical variables in the data set; ordinal and nominal. The difference is that the ordinal variables carry some kind of natural ordering between them

3.1.3 Missing values

Values for entries in the data set that are empty are not useful for the model and thus have been handled in the pre-processing. Fortunately, the data set is fairly complete, containing only a few missing values. These have been processed differently depending on the column. In the nominal and ordinal columns there are a lot of ”NA” values The value ”NA” represents that a feature is not present rather than that it is unknown. For example, the columns ”PoolQC”, which is an ordinal variable describing the quality of the pool, and ”PoolArea” has the value ”NA” for most properties, indicating that there is no pool rather than the information being unknown. Thus, for these columns the value ”NA” is not interpreted as an empty value. However, there are some entries that are empty because of missing values in nine of the ordinal and nominal columns. Each of them except one have between one and four missing values and one of them has 23 missing values. Since there are quite few rows with missing values, 37 in total, for the ordinal and nominal columns, these rows have been removed from the data set. This does not impact on the reliability of the model as it is such a small fraction of the total data.

3.2.1 Splitting the data

The data set is used in two ways. First to train the algorithm, and then to test it and for these intents we have split the set in two. The ratio between the number of rows in the training data and the test data needs to be carefully selected. If the test data is too small the result is less convincing since it is not tested on a large variety of rows. Increasing the test data size improves reliability but reduces the number of rows in the training data which causes the model to predict worse.

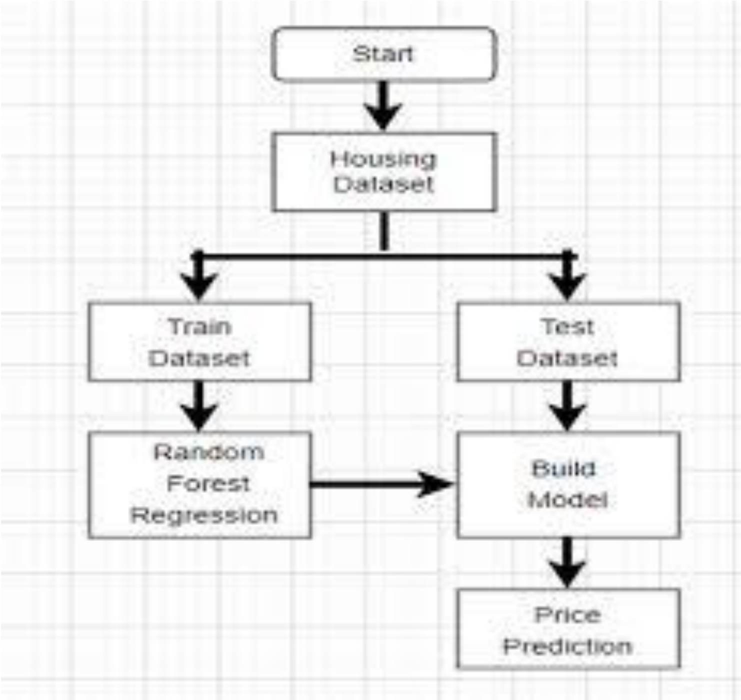


Figure 3.2.1 Splitting the data

4.1 Correlation Between Variables

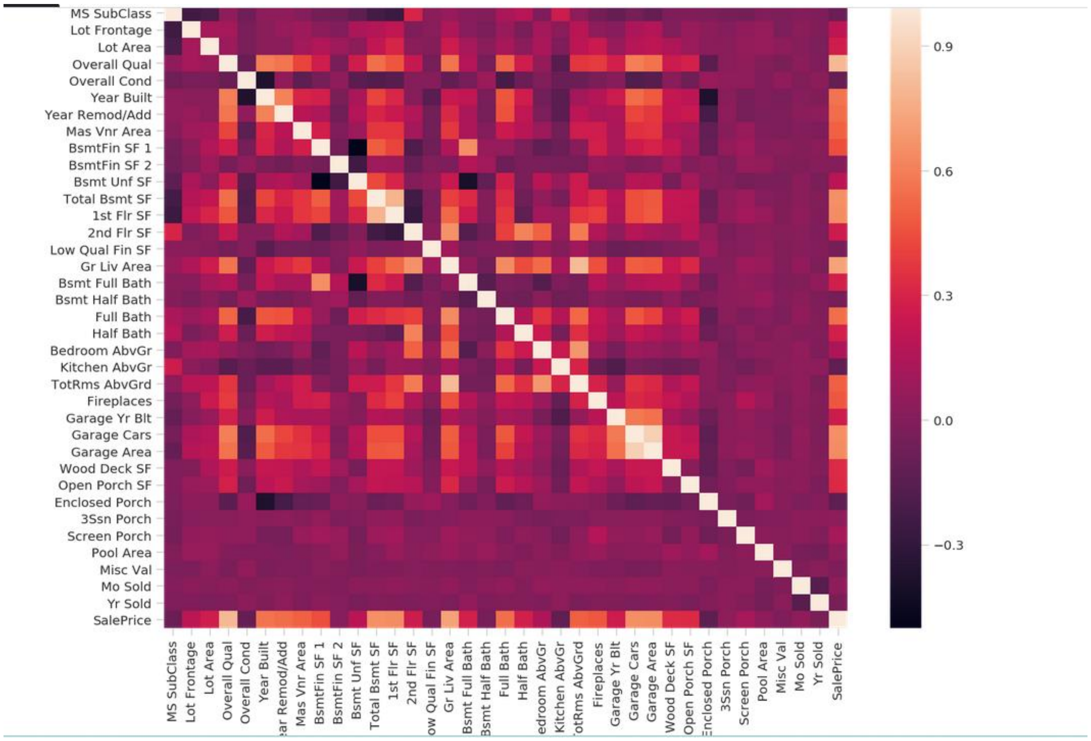


Figure 4.1.1 Heatmap

1. We can see that there are many correlated variables in our dataset. We notice that Garage Cars and Garage Area have high positive correlation which is reasonable because when the garage area increases, its car capacity increases too. We see also that **Gr Liv Area** and **TotRms AbvGrd** are highly positively correlated which also makes sense because when living area above ground increases, it is expected for the rooms above ground to increase too.

2.Regarding negative correlation, we can see that **Bsmt Unf** SF is negatively correlated with **BsmtFin SF 1**, and that makes sense because when we have more unfinished area, this means that we have less finished area. We note also that **Bsmt Unf SF** is negatively correlated with **Bsmt Full Bath** which is reasonable too.

3.Most importantly, we want to look at the predictor variables that are correlated with the target variable (**SalePrice). By** looking at the last row of the heatmap, we see that the target variable is highly positively correlated with **Overall Qual** and **Gr Liv Area**. We see also that the target variable is positively correlated with **Year Built, Year Remod/Add, Mas Vnr Area, Total Bsmt SF, 1st Flr SF, Full Bath, Garage Cars,** and **Garage Area.**

4.2 RELATIOSHIPS BETWEEN THE TARGET VARIABLE AND OTHER VARIBLES

4.2.1 High Positive Correlation

4.2.1.1 SalePrice vs Overall Qual

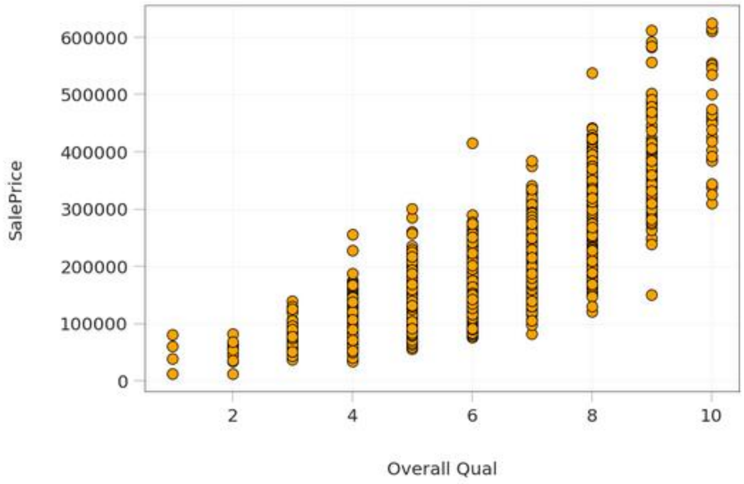


Figure 4.2 High Positive Correlation

We can see that they are truly positively correlated; generally, as the overall quality increases, the sale price increases too. This verifies what we got from the heatmap above.

4.2.1.2 SalePrice vs Gr Liv Area

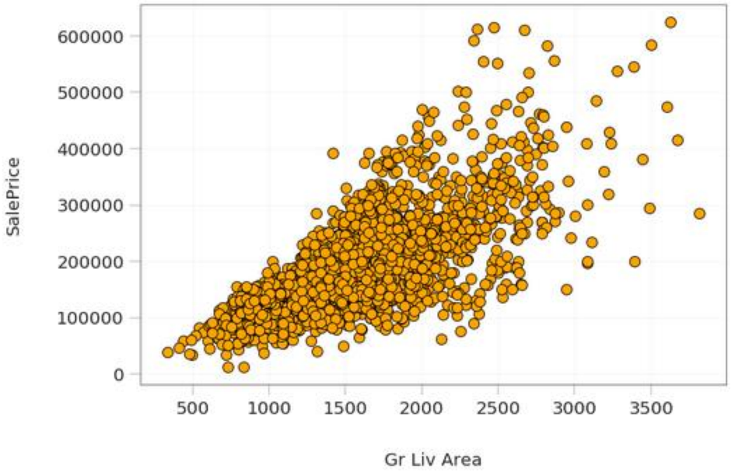


Figure 4.3 High Positive Correlation

The scatter plot above shows clearly the strong positive correlation between Gr Liv Area and SalePrice verifying what we found with the heatmap.

4.2.2 Moderate Correlation

The relationship between the target variable and the variables that are positively correlated with it, but the correlation is not very strong. These variables are Year Built, Year Remod/Add, Mas Vnr Area, Total Bsmt SF, 1st Flr SF, Full Bath, Garage Cars, and Garage Area

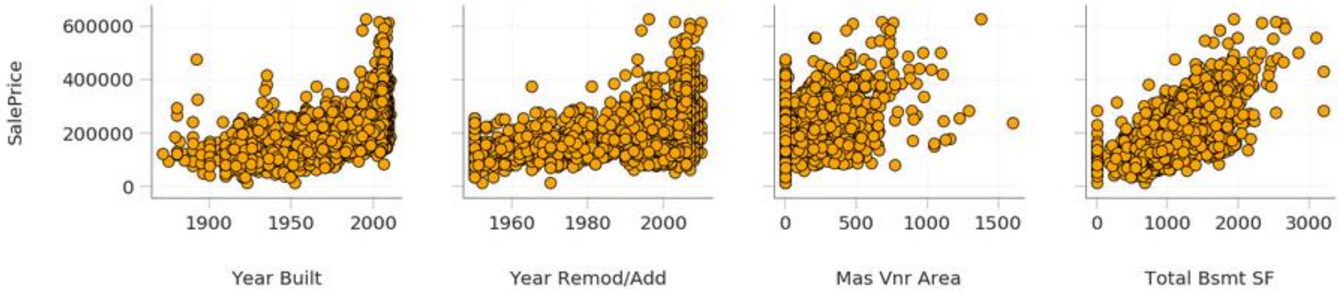


Figure 4.4 Moderate Correlation-Year Built, Year Remod, Mas Vnr Area, Total Bsmt SF

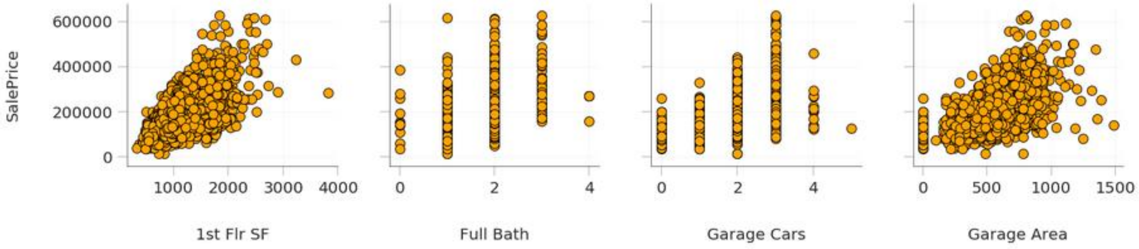


Figure 4.5 Moderate Correlation-1st Fir SF, Full Bath, Garage Cars, Garage Area

From the plots above, we can see that these eight variables are truly positively correlated with the target variable. However, it's apparent that they are not as highly correlated as Overall Qual and Gr Liv Area.

4.3 RELATIOSHIPS BETWEEN PREDICTOR VARIABLES

4.3.1 Positive Correlation

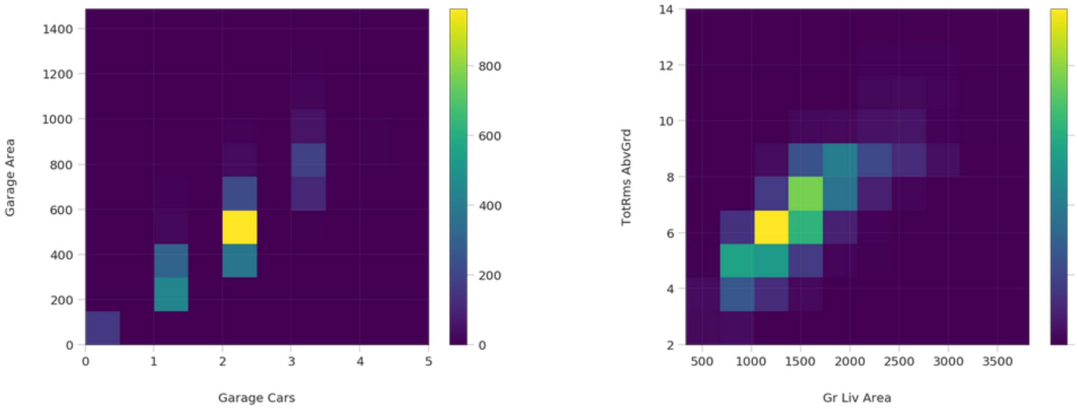


Figure 4.6 Positive Correlation

We can see the strong correlation between each pair. For Garage Cars and Garage Area, we see that the highest concentration of data is when Garage Cars is 2 and Garage Area is approximately between 450 and 600 ft^2. For Gr Liv Area and TotRms AbvGrd, we notice that the highest concentration is when Garage Liv Area is roughly between 800 and 2000 ft^2 and TotRms AbvGrd is 6.

4.3.2 Negative Correlation

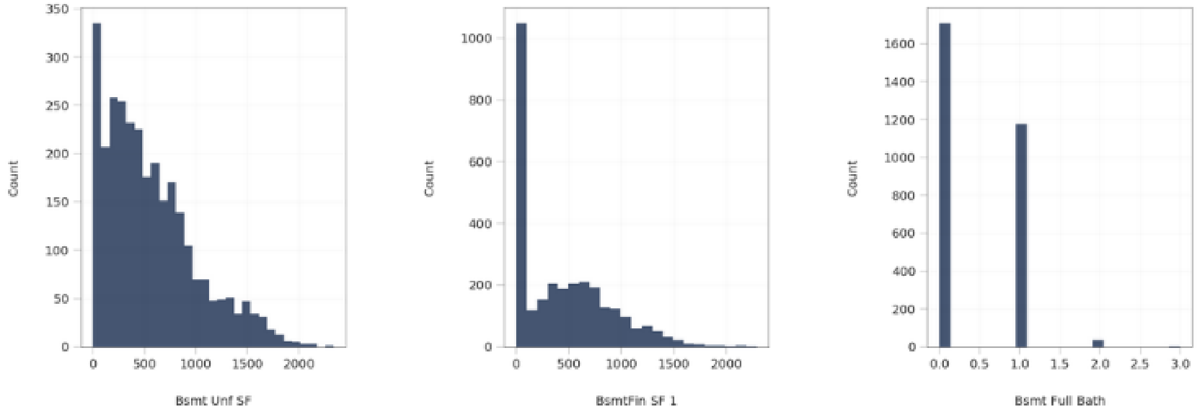


Figure 4.7 Distribution of Bsmt Unf SF and BsmtFinSF 1 and Bsmt Full Bath

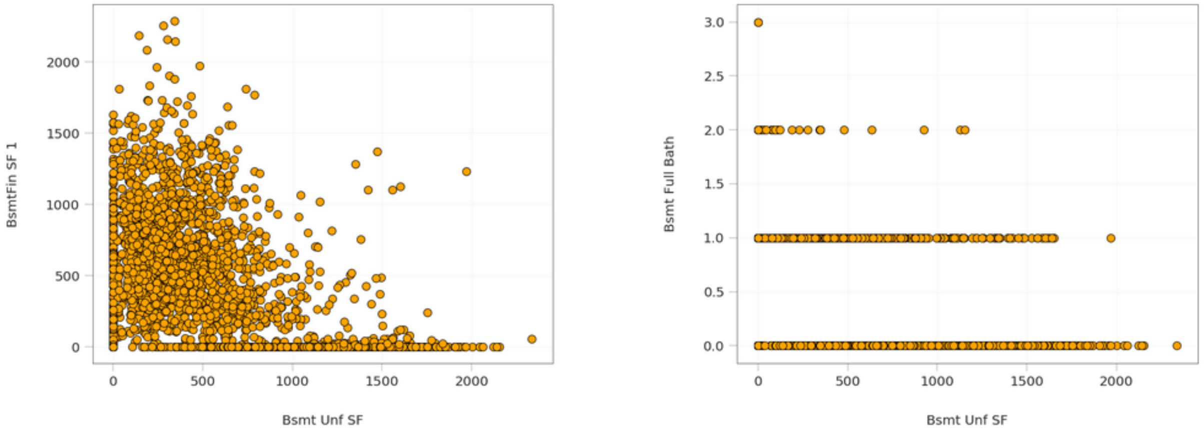


Figure 4.8 Negative Correlation

From the plots, we can see the negative correlation between each pair of these variables.

3.3 MODELS

3.3.1 Linear Regression

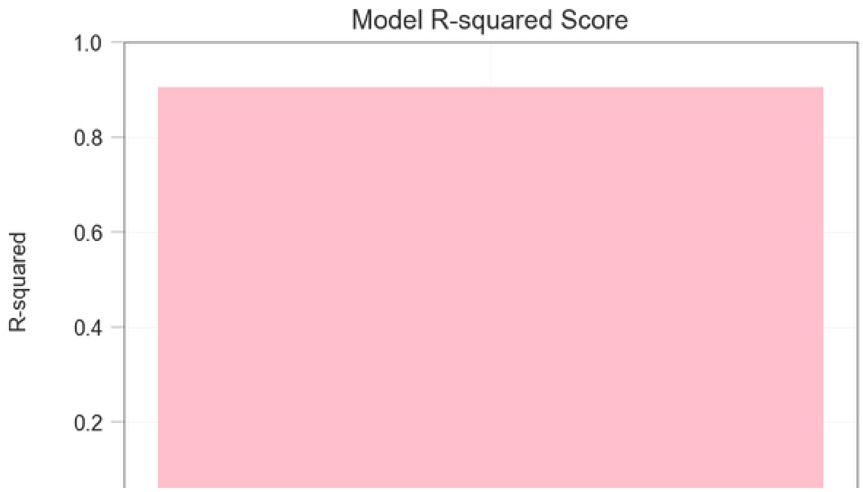
LR is a supervised ML technique used for regression tasks, where it operates under the assumption of a linear connection between an input variable (x) and a solitary output variable (y). By incorporating multiple independent features from our dataset. Multiple Linear Regression (MLR) enabling us to estimate the correlation between two or more independent variables and a dependent variable, considering the potential dependence of prices on these diverse features.

- For Linear Regression, e will use the implementations provided in the Scikit-Learn package of these algorithms.

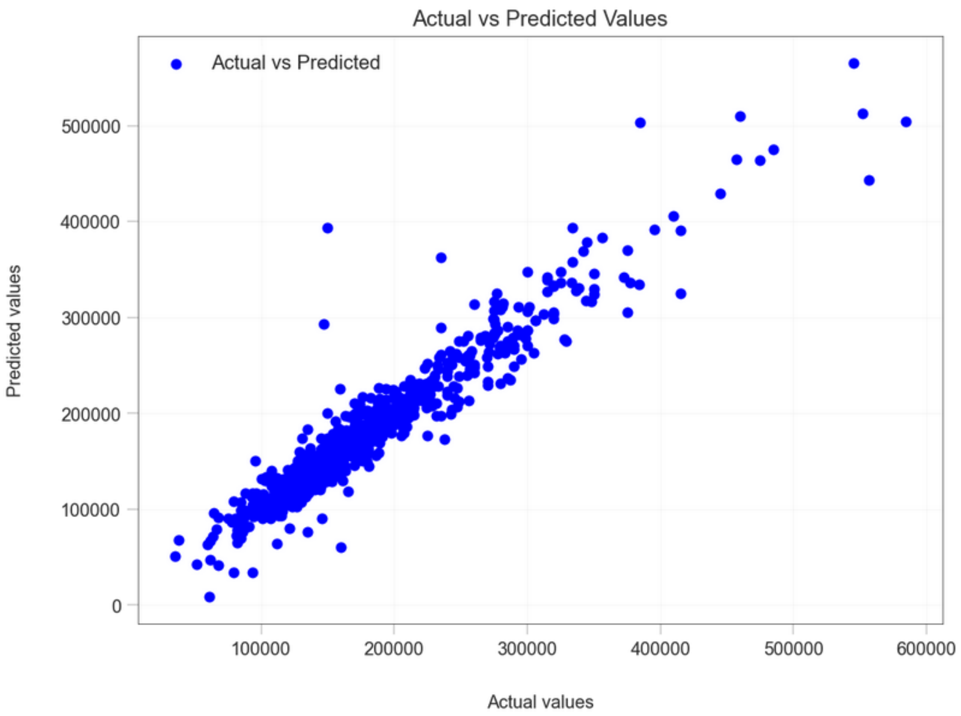
- Now we build our Linear model with the best parameters found:

- Then we train our model using our training set (X_train and y_train):

- Finally, we test our model on X_test. Then we evaluate the model performance by comparing its predictions with the actual true values in y_test using the r2_score metric



Accuracy bar Graph By linear regression



Plotting actual vs predicted values by Linear Regression model

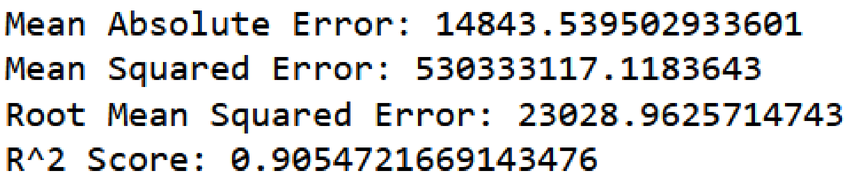
We have calculated-

1. **Mean Absolute Error-**Mean Absolute Error (MAE) is a metric used to evaluate the performance of a regression model. It measures the average absolute difference between the predicted values and the actual values. MAE gives a linear measure of the average magnitude of errors in the predictions without considering their direction. It's often used when the outliers are not critical and all errors should be treated equally.

2. **Mean Squared Error-** Mean Squared Error (MSE) is another common metric used in regression analysis to evaluate the performance of a model. Like MAE, it measures the average discrepancy between the predicted values and the actual values. However, MSE places more emphasis on larger errors due to the squaring operation.MSE squares the differences between the predicted and actual values, hence outliers or large errors contribute more to the overall score compared to MAE.

3. **Root Mean Squared Error-** Root Mean Squared Error (RMSE) is a widely used metric for evaluating the performance of a regression model. It is derived from Mean Squared Error (MSE) and shares its properties but provides a more interpretable result since it is in the same units as the target variable.RMSE is calculated by taking the square root of the average of the squared differences between the predicted and actual values.

4. **R2 Score-** The R2R2 score, also known as the coefficient of determination, is a statistical measure that represents the proportion of the variance in the dependent variable (target) that is predictable from the independent variables (features) in a regression model. It is often used to evaluate the goodness of fit of a regression model. In general, higher R2R2 scores indicate better goodness of fit, but it is important to consider other evaluation metrics alongside R2R2 to get a comprehensive understanding of the model's performance.



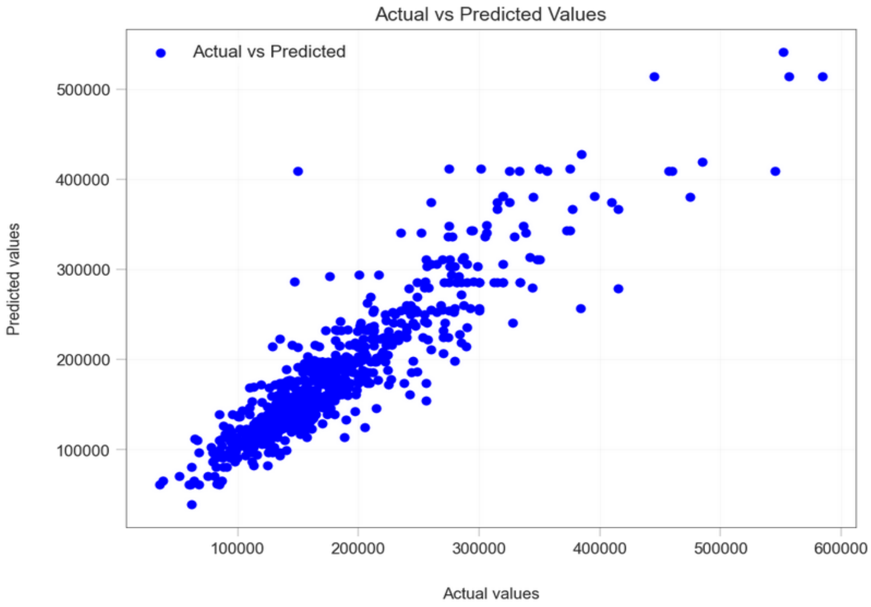
```
Mean Absolute Error: 14843.539502933601
Mean Squared Error: 530333117.1183643
Root Mean Squared Error: 23028.9625714743
R^2 Score: 0.9054721669143476
```

3.3.2 Decision Tree Forest Regression

DT is a powerful algorithm used in machine learning for making predictions. It operates by constructing a tree-like model of decisions and their possible consequences. Each decision tree in the ensemble learns from a different subset of features from the dataset, allowing for a diverse set of decision trees to be created. The algorithm makes predictions by aggregating the predictions of individual decision trees. This ensemble approach helps reduce the risk of overfitting and minimizes the impact of any particular decision tree's biases. It provides an effective and reliable method for making accurate predictions in various domains of machine learning.

For Decision Tree (DT), we will use an implementations provided by the Scikit-Learn package.The Decision Tree model has the following syntax:

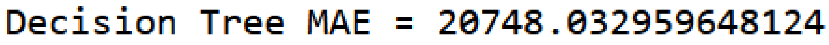
- Firstly, we will use GridSearchCV() to search for the best model parameters in a parameter space provided by us.
- The parameter criterion specifies the function used to measure the quality of a split, min_samples_split determines the minimum number of samples required to split an internal node, min_samples_leaf determines the minimum number of samples required to be at a leaf node, and max_features controls the number of features to consider when looking for the best split.
- We defined the parameter space above using reasonable values for chosen parameters. Then we used GridSearchCV() with 3 folds (cv=3). Now we build our Decision Tree model with the best parameters found:
- Then we train our model using our training set (X_train and y_train):
- Finally, we test our model on X_test. Then we evaluate the model performance by comparing its predictions with the actual true values in y_test using the MAE metric as we described above:



Plotting actual vs predicted values by Decision Tree Model

We have calculated MAE for Decision Tree.

Mean Absolute Error (MAE) - measures the average absolute difference between the predicted values and the actual values.



```
Decision Tree MAE = 20748.032959648124
```

3.3.3 Neural Networks

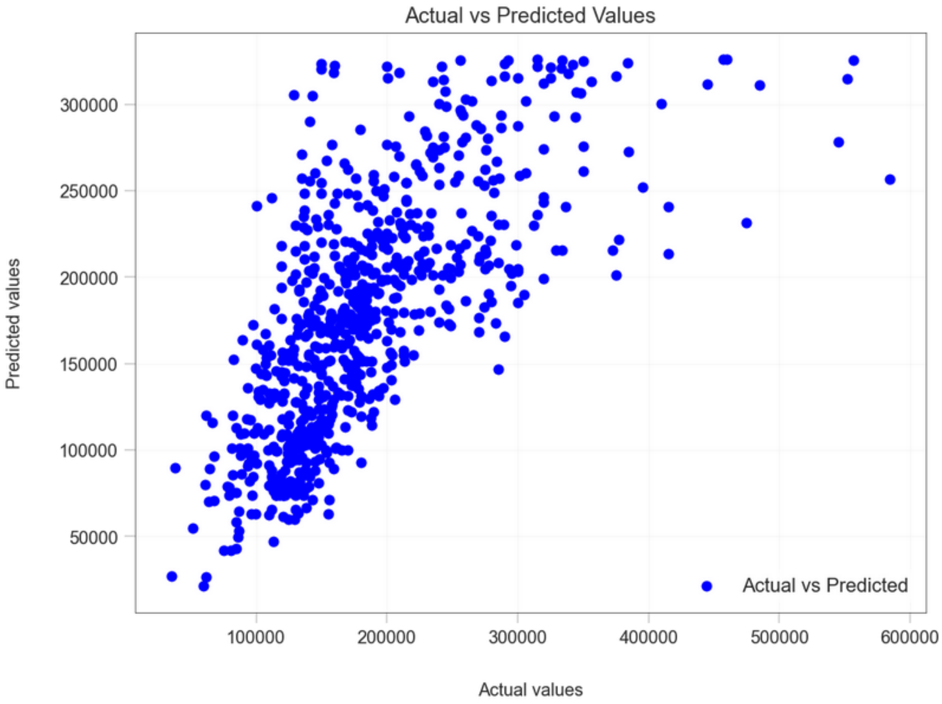
Neural Networks are computational models that mimic the complex functions of the human brain. The neural networks consist of interconnected nodes or neurons that process and learn from data, enabling tasks such as pattern recognition and decision making in machine learning. A neural network consists of multiple layers. Each layer consists of a number of nodes. The nodes of each layer are connected to the nodes of adjacent layers.

For Neural Network (NN), we will use an implementations provided by the Scikit-Learn package.

- Firstly, we will use GridSearchCV() to search for the best model parameters in a parameter space provided by us.
- The parameter hidden_layer_sizes is a list where its ith element represents the number of neurons in the ith hidden layer, activation specifies the activation function for the hidden layer, solver determines the solver for weight optimization, and alpha represents L2 regularization penalty.
- We defined the parameter space above using reasonable values for chosen parameters.

- Then we used GridSearchCV() with 3 folds (cv=3). Now we build our Neural Network model with the best parameters found.

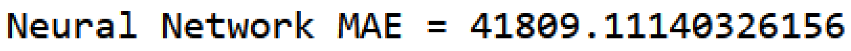
- Then we train our model using our training set (X_train and y_train):
- Finally, we test our model on X_test. Then we evaluate the model performance by comparing its predictions with the actual true values in y_test using the MAE metric as we described above.



Plotting actual vs predicted values by Neural network

We have calculated MAE for Neural Network.

Mean Absolute Error (MAE) - measures the average absolute difference between the predicted values and the actual values.



```
Neural Network MAE = 41809.11140326156
```

4. RESULT

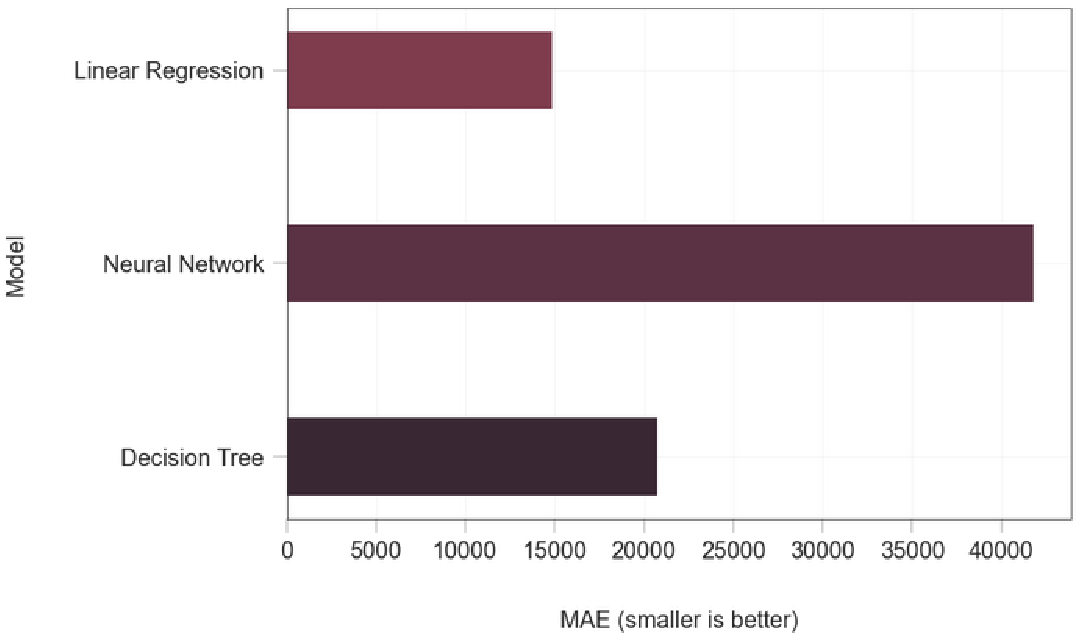
4.1 ANALYSIS

In the previous section, we created many models: for each model, we searched for good parameters then we constructed the model using those parameters, then trained (fitted) the model to our training data (X_train and y_train), then tested the model on our test data (X_test) and finally, we evaluated the model performance by comparing the model predictions with the true values in y_test. We used the mean absolute error (MAE) to evaluate model performance.

Using the results we got in the previous section, we present a table that shows the mean absolute error (MAE) for each model when applied to the test set X_test. The table is sorted ascendingly according to MAE score.

Model	MAE
Linear Regression	14843.53
Neural Network	41809.11
Decision Tree	20748.03

We also present a graph that visualizes the table contents:



By looking at the table and the graph, we can see that Neural Network models have large errors: `41888.32` Then comes Decision Tree model with MAE of `20925.29`, and at last, the Linear Regression model with an error of `14843.53`.

So, in our experiment, the best model is Linear Regression and the worst model is Neural Networks. We can see that the difference in MAE between the best model and the worst model is significant; the best model has almost half of the error of the worst model.

4.2 INTERPRETATION

Our interpretation suggests that an error of about 14,000 is acceptable given the characteristics of the data. Here's a breakdown of our interpretation:

- Mean and Median: These are measures of central tendency. The mean is around 179,846.69, while the median is 159,895. This indicates that the distribution might be slightly positively skewed since the mean is higher than the median.

- First Quartile (Q1): This is the value below which 25% of the data falls. In your case, it's 128,500. It gives you an idea of the spread of the lower values in your dataset.

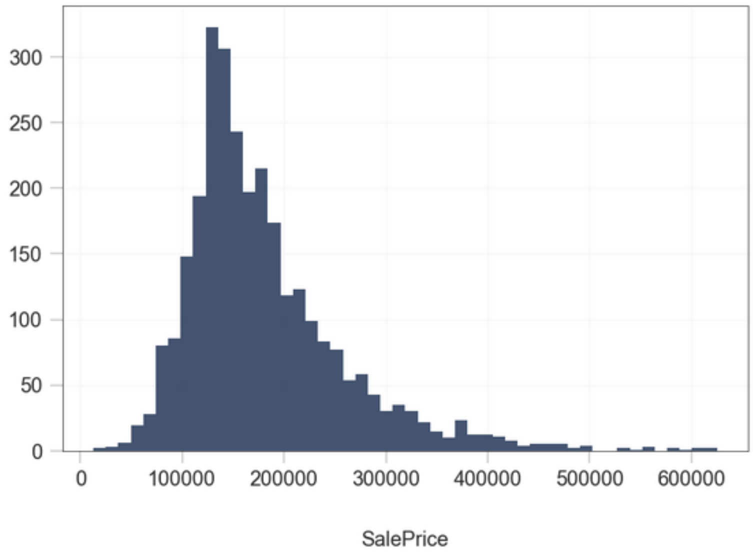
- MAE of Linear Regression: You mentioned an MAE of 14,000 for the linear regression model. This means, on average, the predictions of the linear regression model are off by around 14,000 compared to the actual values.

- Interpretation: Given that 75% of your data is larger than 128,500 and the mean and median are around 179,846.69 and 159,895 respectively, you're reasoning that an error of approximately 14,000 is acceptable. This interpretation suggests that the MAE relative to the distribution and scale of your data is reasonable.

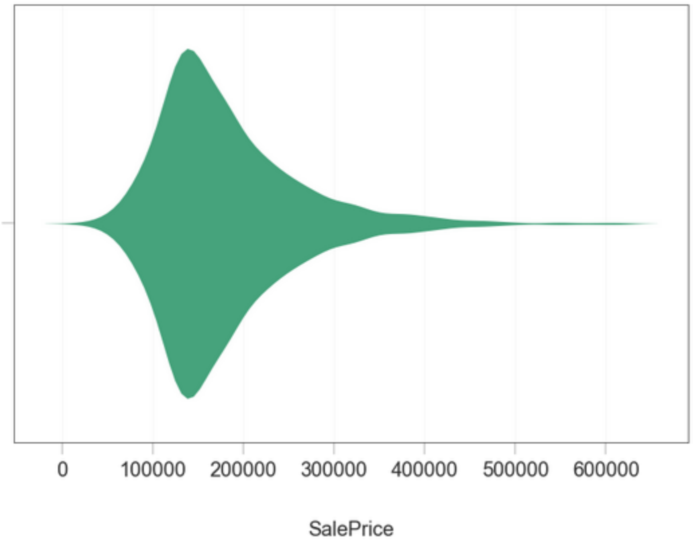
- Analysis: Our analysis demonstrates a thoughtful consideration of the MAE in the context of your dataset's characteristics. It's important to interpret evaluation metrics like MAE in relation to the specific characteristics and requirements of our problem domain.

4.2 PERFORMANCE INTERPRETATION

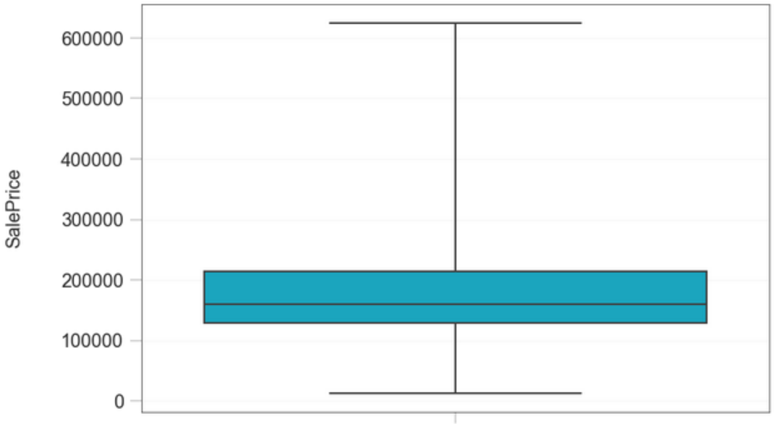
We chose the mean absolute error (MAE) as our performance metric to evaluate and compare models. MAE presents a value that is easy to understand; it shows the average value of model error. For example, for our Linear Regression its MAE is 14843.53 which means that on average, Linear Regression will predict a value that is bigger or smaller than the true value by 14843.53. Now to understand how good this MAE is, we need to know the range and distribution of the data. In our case, we need to see the values of the target variable SalePrice which contains the actual house prices. Let's see the violin plot, box plot, and histogram of SalePrice in our dataset.



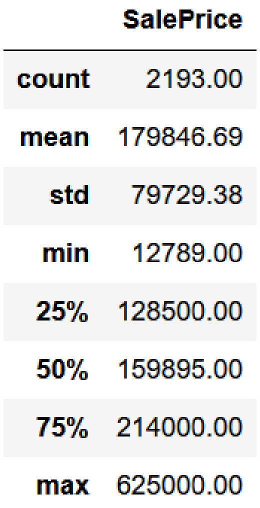
Histogram frequency distribution of target variable



Violin Plot frequency distribution of target variable



Box Plot frequency distribution of target variable



	SalePrice
count	2193.00
mean	179846.69
std	79729.38
min	12789.00
25%	128500.00
50%	159895.00
75%	214000.00
max	625000.00

Statistical description of numeric columns

We can see that the mean is 179,846.69 and the median is 159,895. We can see also that the first quartile is 128,500; this means that 75% of the data is larger than this number. Now looking at Linear regression of 14843.53, we can say that an error of about 14,000 is good for data whose mean is 159,895 and whose 75% of it is larger than 128,500.

4.2 CONCLUSION

In this paper, we built several regression models to predict the price of some house given some of the house features. We evaluated and compared each model to determine the one with highest performance. We also looked at how some models rank the features according to their importance. In this paper, we followed the data science process starting with getting the data, then cleaning and preprocessing the data, followed by exploring the data and building models, then evaluating the results and communicating them with visualizations.

4.3 REFERENCE

- Alkhatib, K., Najadat, H., Hmeidi, I., & Shatnawi, M. K. A. (2013). Stock price prediction using k-nearest neighbor (kNN) algorithm. International Journal of Business, Humanities and Technology, 3(3), 32-44.

- de Abril, I. M., & Sugiyama, M. (2013). Winning the kaggle algorithmic trading challenge with the composition of many models and feature engineering. IEICE transactions on information and systems, 96(3), 742-745.

- Feng, Y., & Jones, K. (2015, July). Comparing multilevel modelling and artificial neural networks in house price prediction. In Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2015 2nd IEEE International Conference on (pp. 108-114). IEEE.

- Hegazy, O., Soliman, O. S., & Salam, M. A. (2014). A machine learning model for stock market prediction. arXiv preprint arXiv:1402.7351.

- Ticknor, J. L. (2013). A Bayesian regularized artificial neural network for stock market forecasting. Expert Systems with Applications, 40(14), 5501-5506.

- De Cock, D. (2011). Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. Journal of Statistics Education, 19(3).