**Capstone Project for Data Science - Tinyuka Session**

**House Price Prediction Using Machine Learning**

Table of Contents

[1.0 Introduction 2](#_Toc174703795)

[1.1 Objectives 2](#_Toc174703796)

[1.2 Problem Statement 2](#_Toc174703797)

[2. 0 Methodology 2](#_Toc174703798)

[2.1 Data Collection and Preparation 2](#_Toc174703799)

[2.1.1 Handling Missing Values: 2](#_Toc174703800)

[2.1.1 Data Cleaning 3](#_Toc174703801)

[2.2 Exploratory Data Analysis (EDA) 3](#_Toc174703802)

[2.2.1 Sale Price Distribution: 3](#_Toc174703803)

[2.2.2 Univariate Analysis: 4](#_Toc174703804)

[2.2.3 Bivariate analysis: 6](#_Toc174703805)

[2.2.4 Multivariate Analysis: 10](#_Toc174703806)

[2.3 Feature Engineering 12](#_Toc174703807)

[2.3.1 Binary Features: 12](#_Toc174703808)

[2.3.2 Time-Based Features: 14](#_Toc174703809)

[2.3.3 Average Room Area: 15](#_Toc174703810)

[2.4 Encoding and Standardization. 16](#_Toc174703811)

[2.5 Model Training and Evaluation 16](#_Toc174703812)

[3.0 Results And Discussion 18](#_Toc174703813)

[3.1 Critical Features 18](#_Toc174703814)

[3.2 Recommendations 19](#_Toc174703815)

[4.0 Conclusion 19](#_Toc174703816)

# 1.0 Introduction

The real estate market is a complex and dynamic environment where numerous factors influence property values. Accurately predicting house prices is crucial for various stakeholders, including buyers, sellers, investors, and policymakers. With the advent of advanced data science techniques and the availability of rich datasets, it is now possible to develop predictive models that can estimate house prices based on various property features.

In this capstone project, I aim to leverage data science and machine learning methodologies to build a robust model that can predict house prices with high accuracy. The project will involve a comprehensive approach, including data cleaning, exploratory data analysis (EDA), feature engineering, model training, evaluation, and interpretation. By the end of this project, I hope to create a predictive model that not only provides accurate price estimates but also offers insights into the key factors driving house prices.

## 1.1 Objectives

The project’s objective is to build a machine learning model for predicting house prices. This involves data cleaning, exploratory data analysis, feature engineering, and model training. The aim is to deliver a model that reliably forecasts property values, evaluates performance against specific metrics, and provides actionable insights for stakeholders in the real estate market.

## 1.2 Problem Statement

Accurate house price prediction is critical for real estate decision-making. This project focuses on creating a machine learning model to estimate house prices from property feature data. By employing data science techniques, the goal is to deliver a reliable predictive tool, addressing real-world challenges in pricing decisions with a focus on model accuracy and practical application.

# 2. 0 Methodology

## 2.1 Data Collection and Preparation

The dataset is the 'House Prices - Advanced Regression Techniques' Dataset from Kaggle and it contains various features describing houses in Ames, Iowa, including sales price, location, and other information. The dataset contained 79 explanatory variables describing various aspects of residential homes.

### 2.1.1 Handling Missing Values:

 Columns with more than 50% missing values were dropped. 50% is chosen as an estimation. Six features were found to have more than that namely (PoolQC with 99.5%, MiscFeature with 96.3%, Alley with 93.8%, Fence with 80.8% and MasVnrType with 59.7%)

For remaining columns, missing values were imputed using various strategies.

* FireplaceQual (47%): The missing values in this column were filled with ‘NA’ which means ‘No fireplace’ as stated in the description.
* LotFrontage (17%): Missing values in this column were filled with the median in their Neighbourhood. As it is expected that in the same neighbourhood, houses would have similar lot frontage.
* Garage Features (5%): The garage features include GarageType, GarageFinish, GarageQual, GarageCond and GarageYrBlt. GarageYrbLt has its missing values filled with 0, while all the other features missing values are filled with ‘NA’ which represents ‘No Garage’ as stated in the description.
* Basement Features (2%): The basement features include BsmtFinType2, BsmtExposure, BsmtFinType1, BsmtCond and BsmtQual. They are also filled with ‘NA’ representing ‘No basement’ as stated in the description.

### 2.1.1 Data Cleaning

The dataset was then checked for duplicates and negative values, none were found in both cases.

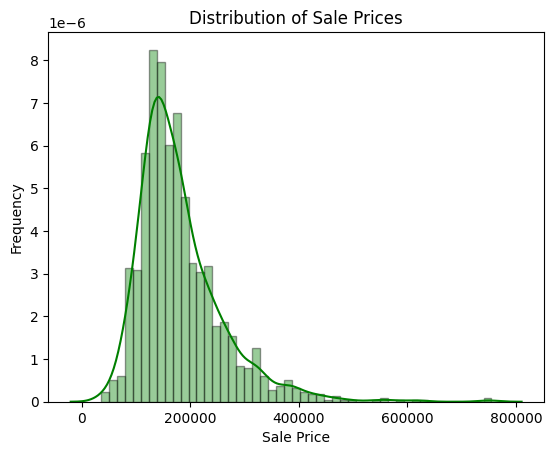
## 2.2 Exploratory Data Analysis (EDA)

Evaluation of the dataset to get an indication of what the features contain was done. Including assessment of the new dataset’s info, shape and summary of numerical and categorical features.

EDA was performed to understand the data and identify potential predictors of sale price. Key findings include:

### 2.2.1 Sale Price Distribution:

 The sale price distribution was right-skewed, indicating a higher frequency of lower-priced homes.



### 2.2.2 Univariate Analysis:

The count of the components of some other features was also inspected to see which of their components had a high number of entries.

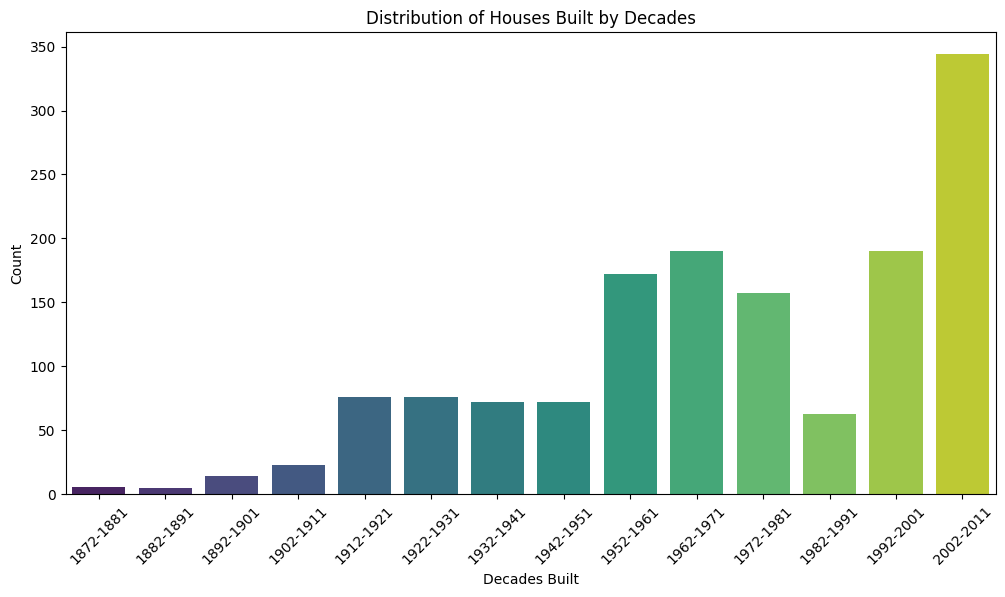
|  |  |
| --- | --- |
| **SaleCondition** | **count** |
| **Normal** | 1198 |
| **Partial** | 125 |
| **Abnorml** | 101 |
| **Family** | 20 |
| **Alloca** | 12 |
| **AdjLand** | 4 |

|  |  |
| --- | --- |
| **BldgType** | **count** |
| **1Fam** | 1220 |
| **TwnhsE** | 114 |
| **Duplex** | 52 |
| **Twnhs** | 43 |
| **2fmCon** | 31 |

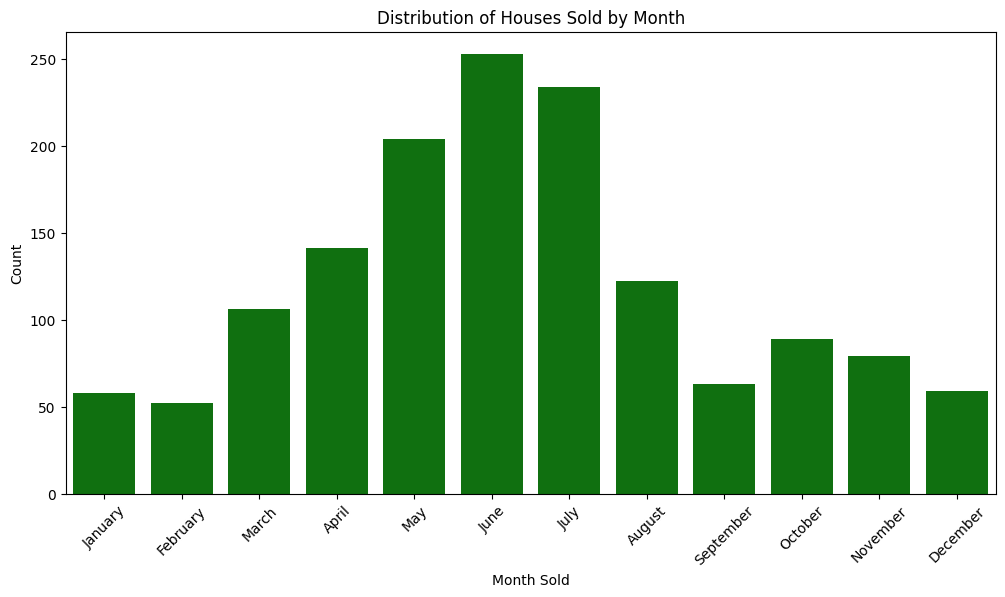
|  |  |
| --- | --- |
| **HouseStyle** | **count** |
| **1Story** | 726 |
| **2Story** | 445 |
| **1.5Fin** | 154 |
| **SLvl** | 65 |
| **SFoyer** | 37 |
| **1.5Unf** | 14 |
| **2.5Unf** | 11 |
| **2.5Fin** | 8 |

It can be seen that there is not much variety when it comes to these features, there is a large number of sales for the top two choices when compared to the others. This also shows that the simplest of the properties sell the most as most people are not looking for anything too sophisticated.

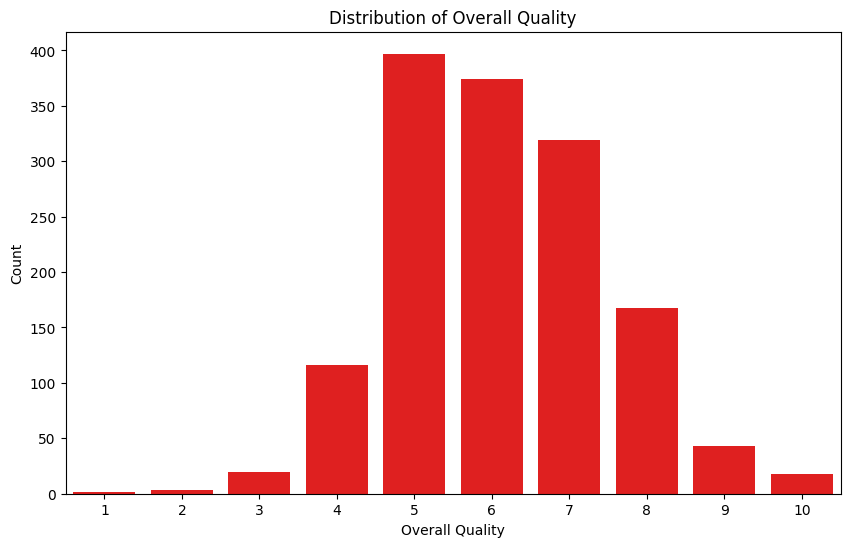
Visualizations can also be used for the analysis of some of the features like the Year Built, the months sold and overall quality.



To visualize the years-built feature, years are grouped together to form decades in other to see information better. It can be seen that almost 350 of the 1460 houses sold were built in the most recent 10-year stretch. Also, a couple houses which are centuries old were sold.



Houses are sold more in the months in middle of the year like May, June and July as can be seen.

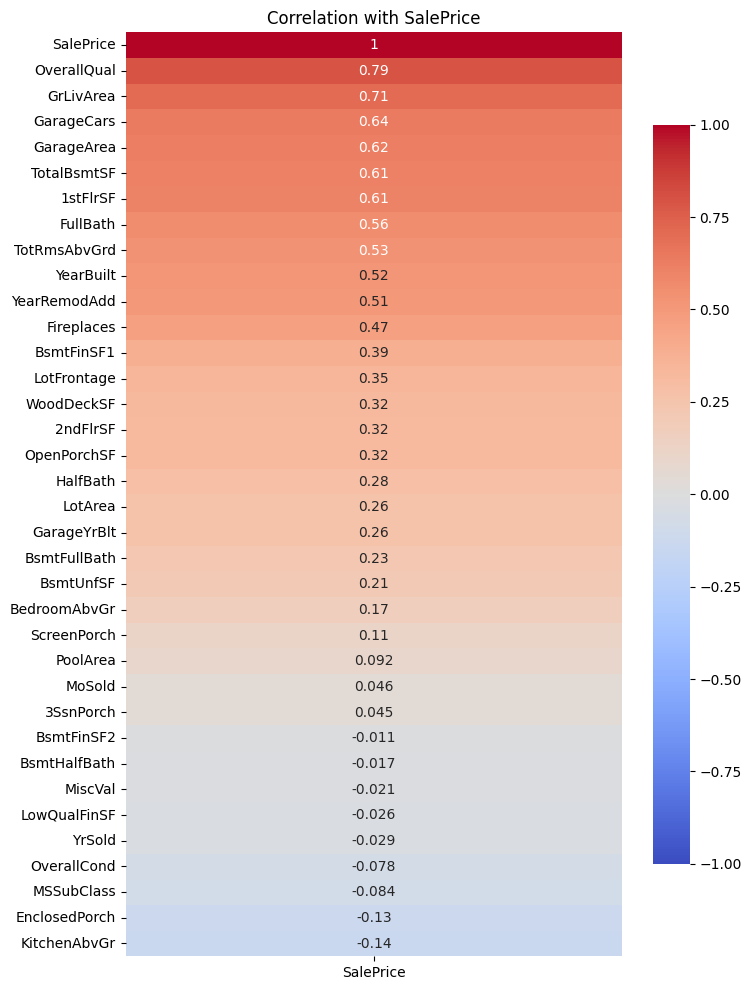


Majority of the houses sold are of average to above quality (5-7 rating). This makes sense as no one would want to buy a low rated house and highly rated house would cost a lot more.

### 2.2.3 Bivariate analysis:

This involves examination of the variables and sale price to understand the relationship between them.

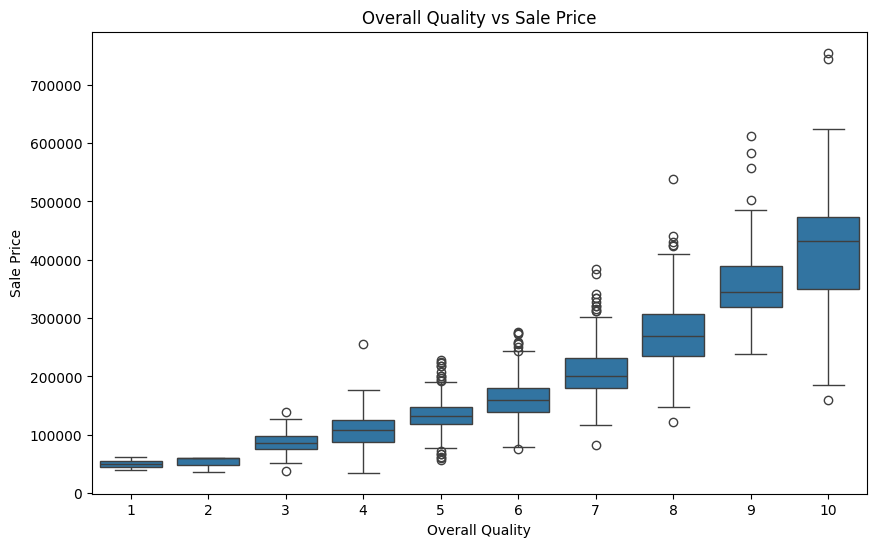
**Correlations:**  Calculation of correlation is done to give an overview of strong correlation between the numerical features and sale price.



When examining the features relationship with sale price it can be seen that some features correlate positively largely with sale price, and none correlate largely negatively. This visual can be used to determine with features to investigate their relationships with sale price further.

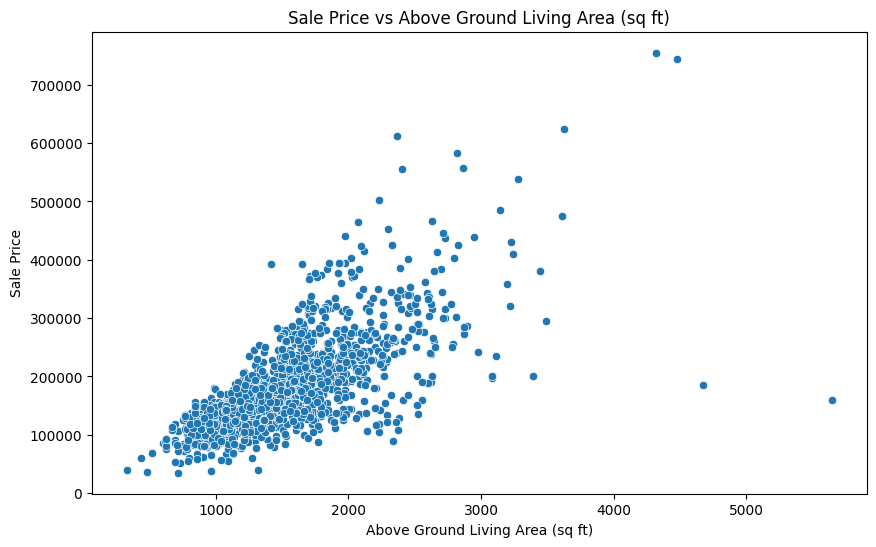
Features like 'OverallQual', 'GrLivArea', and 'GarageCars' showed strong positive correlations with sale price.

**Overall quality** correlates strongly with sale price(+0.79) as the price increase as the quality increases, plotting a box plot would help understand the relationship further.

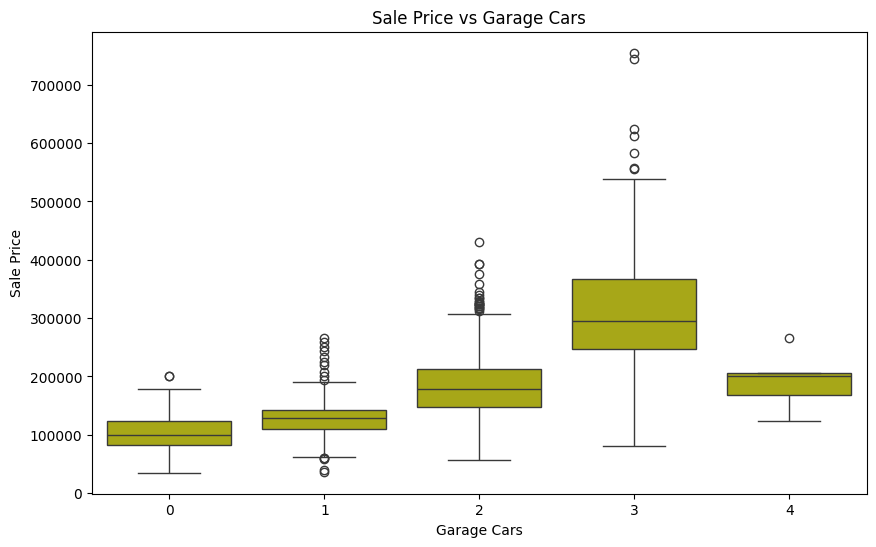


The boxplot has also shown that the range of prices also increases with increase in overall quality.

**GrLivArea** also largely correlates with sale price, we can use a scatterplot to understand this relationship.



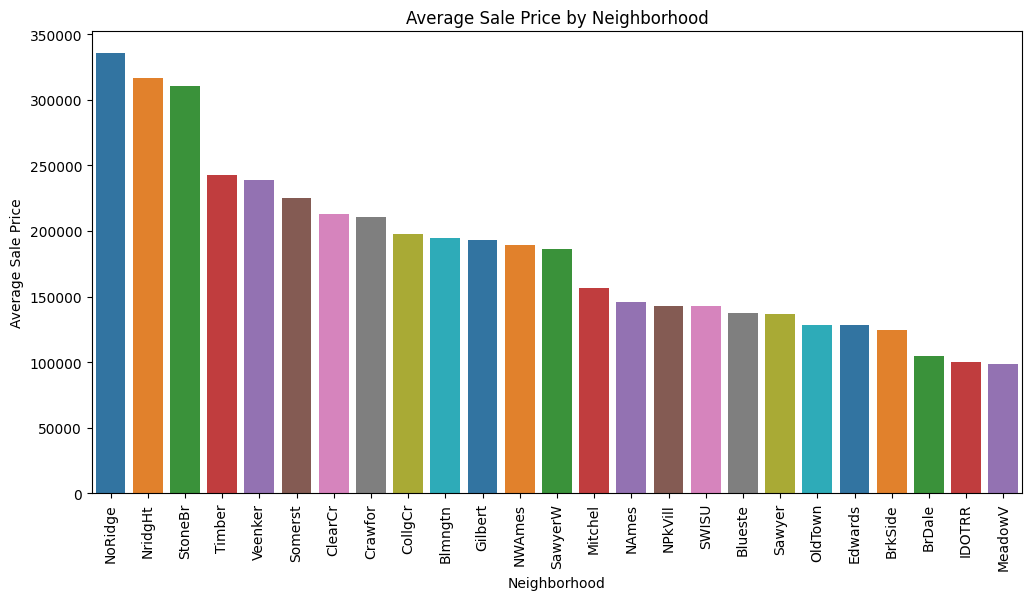
Another strong correlation (+0.71) is seen in above ground living area, showing that a larger area mostly results in a high sale price.



**Garage cars:** One more feature with a strong positive correlation with sale price is garagecars (+0.64), visualizing this shows that an increase in garage cars results in an increase in sales price but strangely this correlation peaks at a garage car number of 3 and drops at a garage car number of 4.

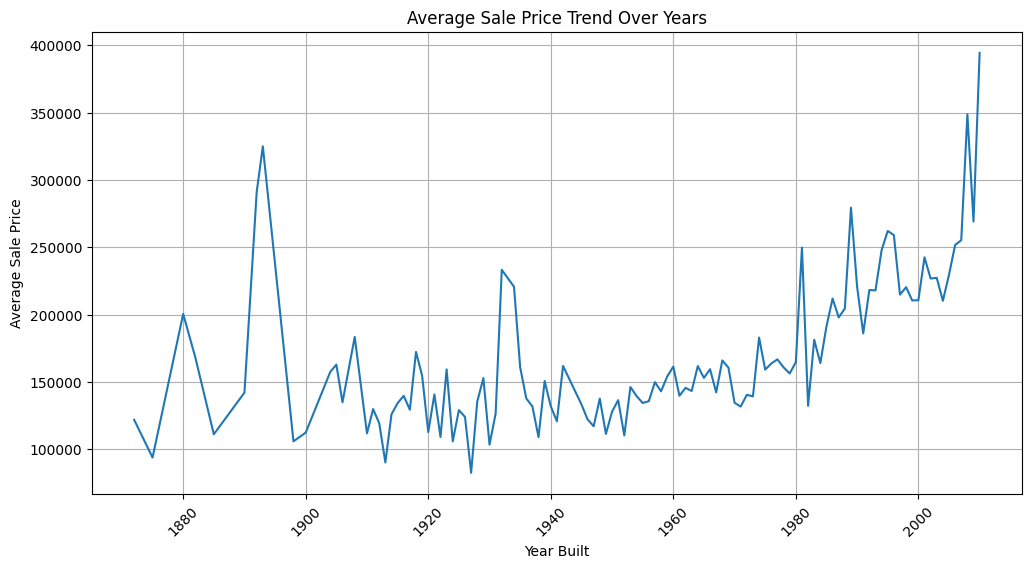
**Neighbourhood:** For non-numerical feature, the neighbourhood is assumed to be a feature that would affect the sale price as different neighbourhoods should have different types of houses which results in a varying sales price.

The average price in each neighbourhood is calculated and plotted against the sales price in a bar plot.



As expected, the neighbourhood significantly impacted average sale prices, with some neighbourhoods commanding higher prices than others. The highest average sales price is in the Northridge neighbourhood and the lowest is the Meadow Village neighbourhood.

**Years Built:** The average sales price over the years is also expected to increase as houses tend to increase when built more recently. The average sales price is plotted on a line chart to asses the trend over the years.



Average sales price is higher when the houses ae built more recently, although there are some very old houses which are expensive which could also show value in sentimental properties.

### 2.2.4 Multivariate Analysis:

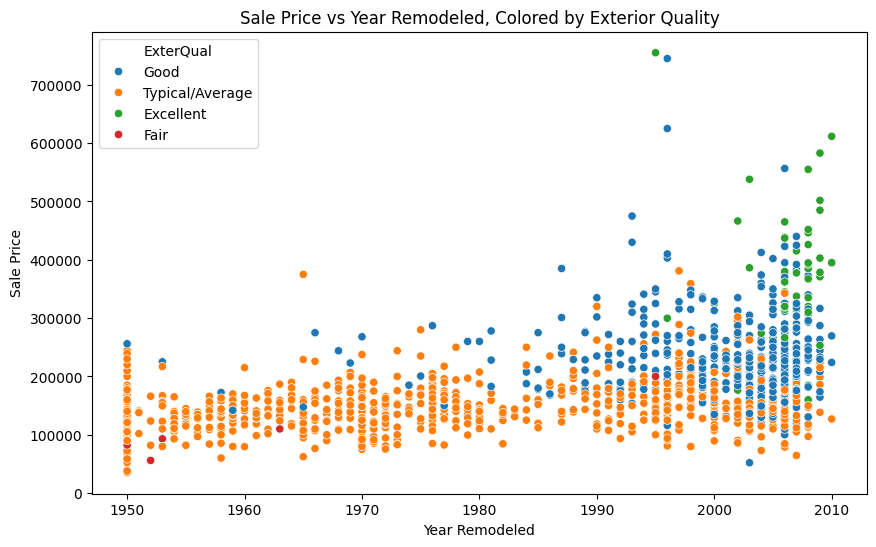
This involves comparing multiple features and their relationship with sales price in one chart or visualization to under the overall effects of combinations.

* **What is the overall quality of houses corresponding to their sales price and above ground living area?**



It can be seen that Higher overall quality and higher GrLIvArea tend to mean a larger sales price as both these features largely correlate with sales priice.

* **How does the exterior quality affect the sales price and do recently remodelled houses have higher exterior quality and sale price?**



Due to remodelling, very little to few houses are of ‘Fair’ exterior quality. Majority of the remodelled houses have an exterior quality of average meaning there was a minimum standard accepted when remodelling happened, and these houses are on the lower end of sales prices. More recently remodelled houses have better exterior quality and therefore higher sales prices.

## 2.3 Feature Engineering

New features were created to provide a different look to the dataset and potentially improve model performance:

### 2.3.1 Binary Features:

 'HasPool', 'Has2ndFloor', 'HasGarage', 'HasFireplace' were created to indicate the presence of these features.

* **Does the presence of a pool largely affect the distribution of sales price?**

A graph with a bar chart and a line of colored squares

Description automatically generated with medium confidence

The presence of a pool has a an effect on the distribution of sales price, the average sales price is higher when there is a pool but there are more expensive houses or outliers when there is no pool.

* **Does the presence of a 2nd floor or a garage largely affect the distribution of sales price?**

A graph of a graph showing a comparison between two different colored squares

Description automatically generated with medium confidence

A graph showing a garage sale

Description automatically generated

Having a second floor has the opposite effect on the distribution of sales price, there is a lower average and also a lower number of expensive houses. This is also seen to happen with having a garage.

* **Does the presence of a fireplace largely affect the distribution of sales price?**

A graph showing a fireplace and a fireplace

Description automatically generated

Having a fireplace largely affects the distribution of sales price, this understandable as a fireplace is a luxury item itself. The average cost of a house with a fireplace is higher than the average of those without.

### 2.3.2 Time-Based Features:

‘HouseAge' and 'YearsSinceRemodel' were calculated to capture the age of the house and time elapsed since remodeling.

* **Does the Age of the house at sale affect sales price?**

A graph showing a red line and blue dots

Description automatically generated

The age of the house at sale affects the sales price as the younger the house the higher the sales price tends to be. This would be as a result of newer and more expensive technologies that would go into the construction of the houses as the years go by.

* **Does the years since a remodel affect sale price?**

A graph showing a line graph

Description automatically generated with medium confidence

Similarly the years since a house has been remodelled also affects the sale price as recently remodelled houses cost more . This would also be as a result of the newer things go into the remodelling of the house.

### 2.3.3 Average Room Area:

'AvgRoomArea' was calculated by dividing the above-ground living area by the total number of rooms above ground.

* **Does the average room area affect the sales price**

A graph with a red line and blue dots

Description automatically generated

As expected, the larger the room sizes are the more expensive the sales price will be.

## 2.4 Encoding and Standardization.

Before encoding and standardization occurred, some features were selected as those that appear to affect the housing sale price the most and may help result in the preparation of the model.

The include 'SalePrice’,'SaleCondition', 'Neighborhood', 'Electrical', 'HouseStyle', 'YearBuilt', 'MonthSold\_Name', OverallQual', 'GarageCars', 'GarageArea’, ‘TotalBsmtSF', '1stFlrSF', 'AvgRoomArea', 'YearsSinceRemodel', 'GrLivArea', 'RoofStyle', 'OverallCond', 'YearRemodAdd', 'ExterQual', 'BldgType', 'Has2ndFloor', 'HasGarage', 'HasFireplace', 'HouseAge'

**Label Encoding:** Categorical features were encoded using LabelEncoder to convert them into numerical representations.

**Standardization:** Numerical features were standardized using StandardScaler to ensure they have a mean of 0 and a standard deviation of 1, improving model performance.

## 2.5 Model Training and Evaluation

To predict house prices, four machine learning algorithms were selected based on their suitability for handling the dataset and the problem at hand. These algorithms were:

1. **Linear Regression:** A simple and interpretable linear model that serves as a baseline for comparison. It helps establish a performance benchmark and identify whether more complex models are necessary.
2. **Decision Tree Regression:** This model captures nonlinear relationships and feature interactions effectively. Its visual interpretability allows for understanding how different features influence predictions.
3. **Random Forest Regression:** An ensemble method combining multiple decision trees to improve accuracy and reduce overfitting. It enhances model robustness and efficiency, especially for large datasets with complex feature interactions.
4. **Gradient Boosting Machines (GBM):** An advanced boosting technique that incrementally builds and refines models. It excels at handling complex nonlinear interactions and often outperforms other predictive models.

**Training and Initial Evaluation:** Each model was trained on the training set (80% of the data) and evaluated on the test set (20% of the data). The following metrics were used to assess model performance:

* **Root Mean Squared Error (RMSE):** Measures the average deviation of predicted values from actual values.
* **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual values.
* **R-squared (R²):** Represents the proportion of variance in the target variable explained by the model.

**Hyperparameter Tuning:** To further optimize model performance, hyperparameter tuning was performed using GridSearchCV. This technique systematically explores different combinations of hyperparameters to find the best configuration for each model. The following hyperparameters were tuned:

* **Decision Tree:** max\_depth, min\_samples\_split, min\_samples\_leaf
* **Random Forest:** n\_estimators, max\_depth, min\_samples\_split
* **Gradient Boosting:** n\_estimators, learning\_rate, max\_depth

**Refitting and Final Evaluation:** After identifying the best hyperparameters, each model was refitted on the entire training dataset to incorporate the optimal settings. The refitted models were then evaluated on the test set using the same metrics (RMSE, MAE, R²) to obtain the final performance results.

# 3.0 Results And Discussion

The following table summarizes the performance of the models after hyperparameter tuning:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **RMSE** | **MAE** | **R²** |
| Linear Regression | 33350.34 | 22657.94 | 0.8241 |
| Decision Tree | 30672.82 | 19786.24 | 0.8547 |
| Random Forest | 22046.63 | 14240.86 | 0.9183 |
| Gradient Boosting | 20419.90 | 12847.59 | 0.9287 |

The Gradient Boosting model achieved the best performance with the lowest RMSE and MAE and the highest R² value. This indicates its superior accuracy and consistency in predicting house prices compared to the other models.

This rigorous training and evaluation process ensured that the selected models were well-suited for predicting house prices and that their performance was optimized for the given dataset. The results of this process clearly indicated that the Gradient Boosting model outperformed the other models, making it the most suitable choice for this task.

## 3.1 Critical Features

A graph with blue bars

Description automatically generated

The most important features influencing house prices, as identified by the Gradient Boosting model, are:

1. **OverallQual:** Overall material and finish quality of the house.
2. **GrLivArea:** Above ground living area square footage.
3. **TotalBsmtSF:** Total basement area square footage.
4. **1stFlrSF:** First floor square footage.
5. **Neighborhood:** Location of the house.

## 3.2 Recommendations

* **Focus on Quality:** Prioritizing the overall quality of a house can significantly impact its sale price.
* **Maximize Living Space:** Larger living areas generally command higher prices.
* **Consider Location:** Neighborhood plays a crucial role in determining house prices.

# 4.0 Conclusion

This capstone project set out to develop a machine learning model capable of accurately predicting house prices using a rich dataset of residential properties in Ames, Iowa. Through a systematic process of data cleaning, exploratory data analysis (EDA), feature engineering, and model training, the project successfully achieved its objectives.

The project highlighted the importance of handling missing data, which was carefully managed to ensure the integrity of the dataset. EDA revealed key patterns and relationships between the various features and house prices, providing valuable insights into the factors that most significantly influence property values. Feature engineering further enhanced the dataset by creating new variables that captured additional dimensions of the data, such as binary indicators for specific house features and time-based attributes.

Four machine learning models were trained and evaluated: Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Machines (GBM). Among these, the Gradient Boosting model emerged as the best performer, achieving the highest accuracy with the lowest Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the highest R-squared (R²) value. This model effectively captured the complex relationships in the data and provided reliable predictions of house prices.

Critical features identified by the model included Overall Quality, Above Ground Living Area, Total Basement Area, First Floor Area, and Neighborhood, all of which played significant roles in determining house prices. The analysis underscored the importance of factors such as property quality, size, and location in real estate valuation.

The findings of this project offer actionable insights for various stakeholders in the real estate market, including buyers, sellers, and investors. By focusing on key features such as prioritizing property quality, maximizing living space, and carefully considering location, stakeholders can make more informed decisions that align with market trends.

In conclusion, the project demonstrated the power of data science and machine learning in addressing complex real-world problems, such as house price prediction. The Gradient Boosting model, with its superior performance, serves as a robust tool for estimating property values and can be further refined and applied in different real estate markets to aid in decision-making and investment strategies.