Housing Price Analysis

By: Charles Dillaway

Agenda

Some Topics For This Presentation:

- Client Background
- Project Purpose
- Data Processing
- Exploratory Data Analysis
- Machine Learning Predictions
- Recommendations/Conclusion

Background

The Ames Planning & Housing Department is the local housing authority for Ames, Iowa. There responsibilities usually involve offering supplemental housing for low-income individuals and to ensure that houses within the local area meet a high standard. They have given me two tasks.

1. To make general analysis on the housing market within Ames, Iowa.

2. To Make a machine learning model that can accurately forecast the cost of future houses within the area.

Goals

1. Uncover patterns within the dataset to understand the market demand of houses and their features in Ames, Iowa

2. Prepare the dataset for modeling through data-processing

3. Develop the machine learning model to drive accurate predictions of housing prices within Ames, Iowa

Purpose

My Analysis will show:

1. How the housing market of Ames, Iowa differs from that of other regions within the United States.

2. Why understanding the market differences within the town of Ames will help us drive more accurate predictions.

3. How the developed machine learning model will help all sorts of industries understand the housing market within Ames

Data Processing

Data Overview

Shape of the dataframe: 2930 Rows, 82 Columns

Total Missing Values: 15749 NaN values

Number of Object Columns: 43 object Columns

Key Columns for the dataset…

- Sale Price (target Variable)
- Overall Qual (Strong Correlation with Target Variable)
- Garage Area (highlights an interesting demand in the community)

Handling Object Values

There were many Object Values, which are values made from text and not numbers, I had to handle these before I did modeling because:

- Machine learning models cannot handle text values properly, as computers can really only read numerical values

- Dropping them would work, but changing the object values to numbers meant that I had a larger pool of features when building the model

Missing Values

In order to handle missing values, I decided to drop the rows that had missing values. I decided to do this for a couple of reasons:

1. Having missing values within a machine learning model isn’t possible.

2. Imputing the missing values with the median, while good practice, isn’t the most practical with machine learning, as it is synthetic data.

3. Synthetic data within machine learning could lead to misleading model performance, as it would’ve developed a prediction from an estimate, not real data.

Outlier Handling

In this dataset, I decided to do some outlier handling, I did outlier handling because…

1. I Planned on using Regression Plots, and the presence of outliers in those would have affected the regression line.

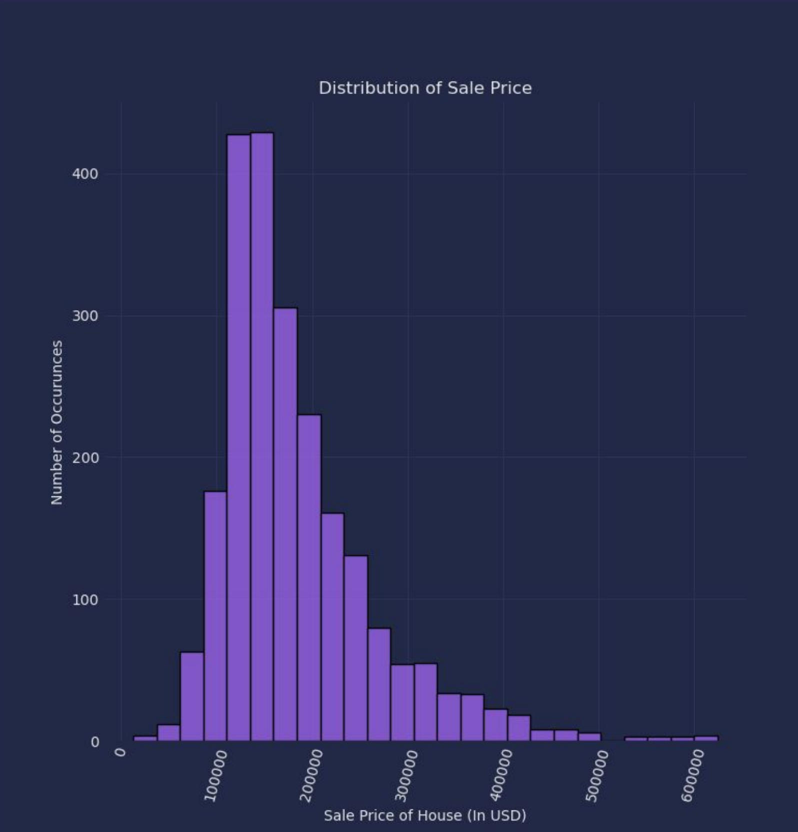
2. Outliers can make correlation coefficients higher or lower than a cleaned set, which can lead to misleading results

Exploratory Data Analysis

Visualization 1

A histogram that shows the distributions of housing prices in Ames, Iowa.

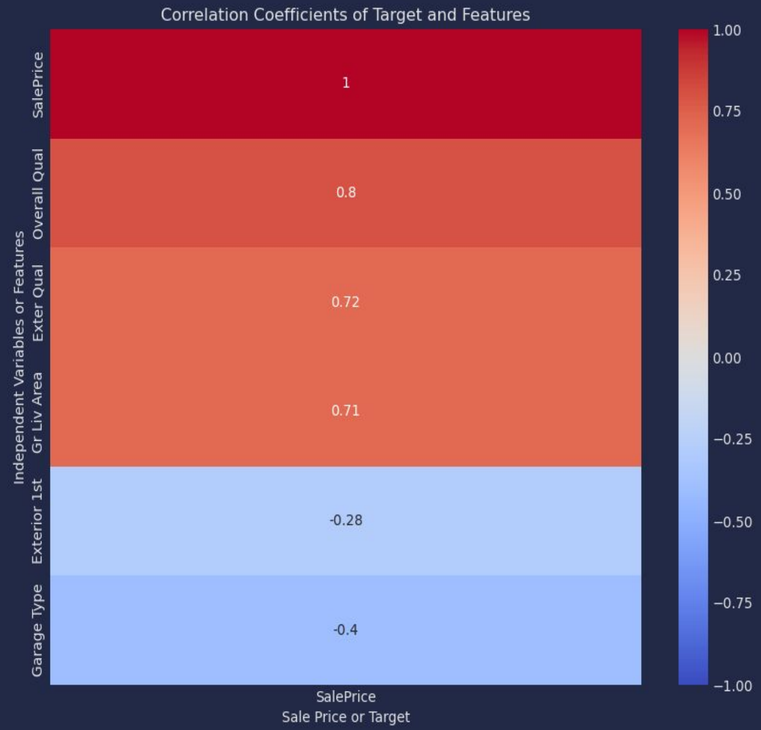
- Most houses range from being 100,000 to 200,000 dollars



Visualization 2

A heatmap that shows the correlation coefficients of each feature by the target variable.

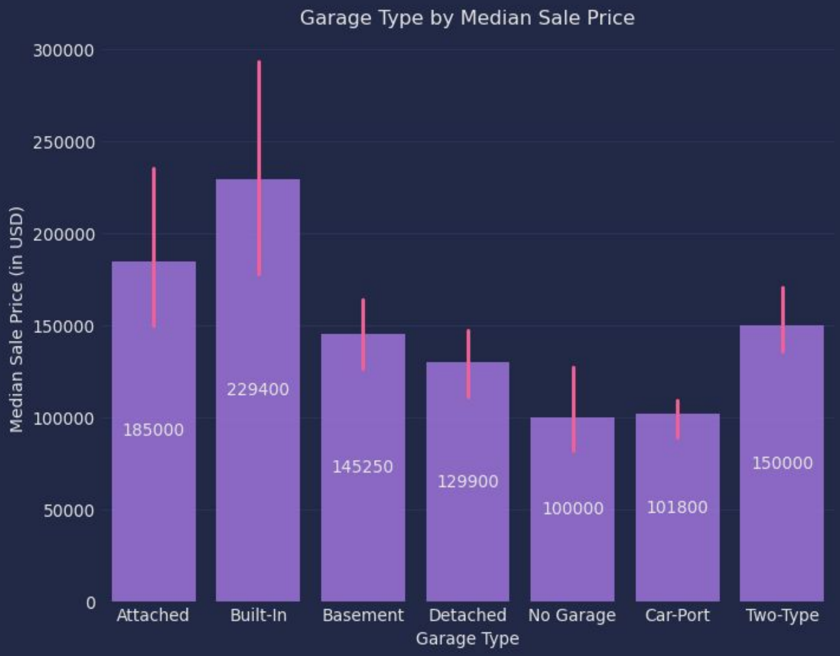
- Example, as the actual one was to large.
- Important as it will give us ideas on features to use for the model



Visualization 3

A bar plot that shows the sale price by the type of garage.

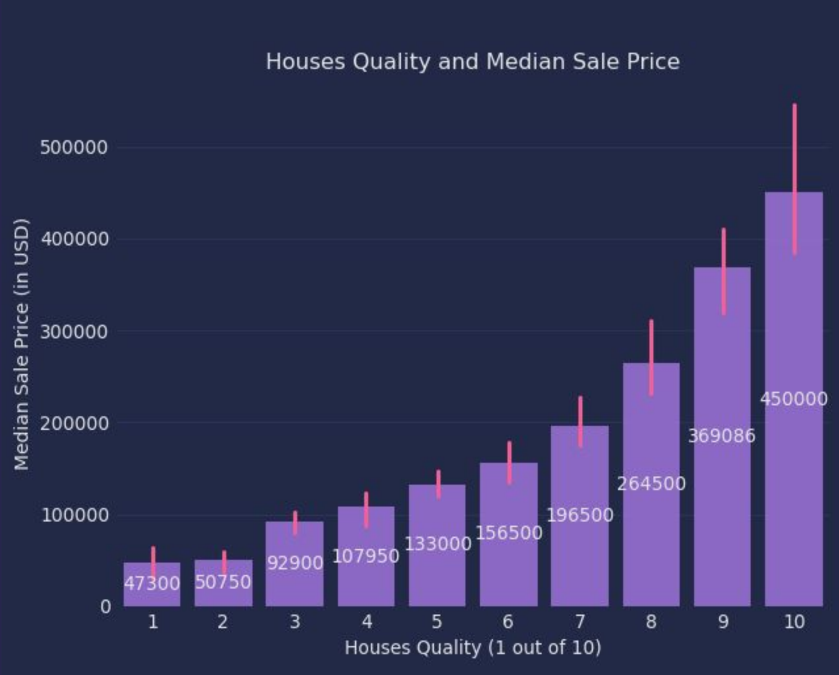
- Built-In garages have the most expensive median sale price
- A house with no garage has the least expensive median sale price
- The error bars in pink represent the IQR, it shows the dispersion of the middle 50% of the data



Visualization 4

A bar plot that shows the sale price by the overall quality of the house.

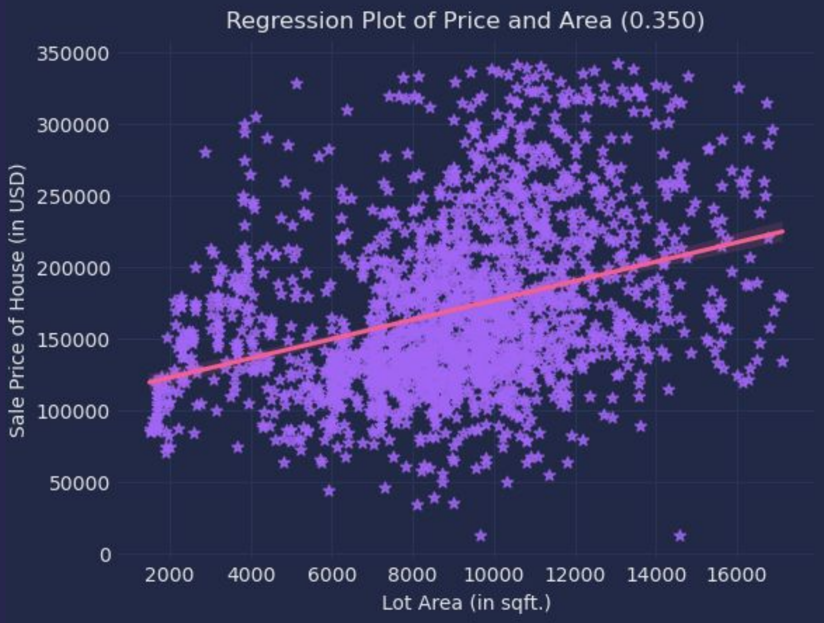
- A house rating of 10 has a median sale price of 450,000 dollars.
- A house rating of 1 has a median sale price of 47,300 dollars.
- As the rating goes up, the sale price will likely go up as well



Visualization 5

A regression plot that shows the correlation between the sale price and the Lot area

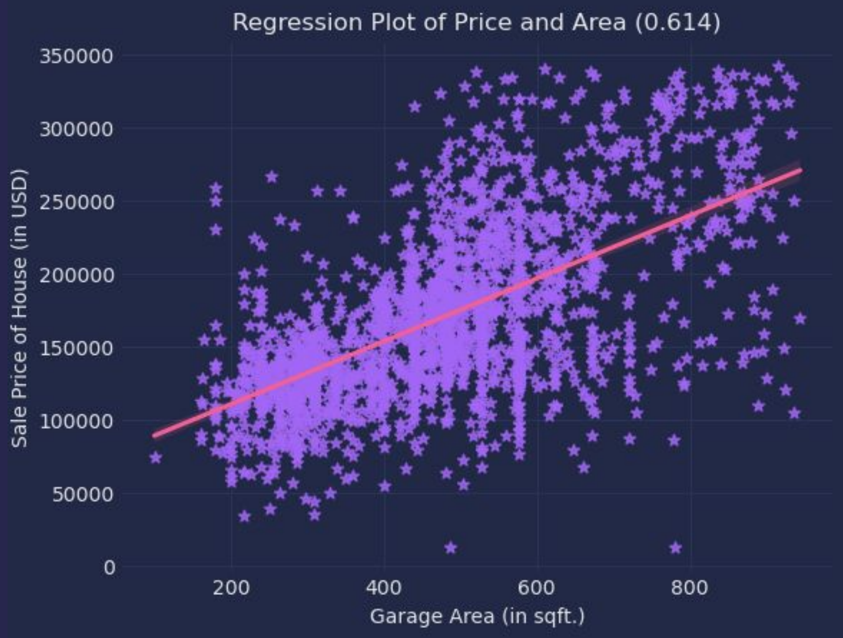
- Correlation Coefficient of 0.35 implies weak to moderate correlation
- An increase in Lot Area doesn’t affect the Sale Price strongly



Visualization 6

A regression plot that shows the correlation between the sale price by the area of the garage.

- Correlation Coefficient of 0.614 implies moderate to strong correlation.
- An increase in the Garage Area likely will lead to an increase in Sale Price.

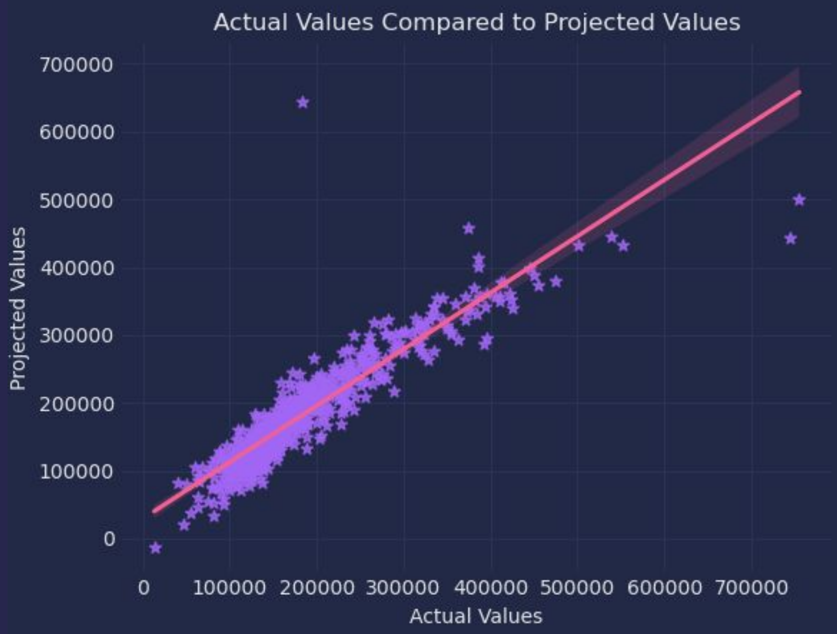


Projections

Visualization of Model

A regression plot that shows the Actual Values compared to the ones the model Projected.

- Showcases the accuracy of the model
- A couple of outliers here and there, mostly towards the end of the regression line



Model Performance

The Model I chose was linear regression due to its simplicity as well as it’s interpretability.

- R Squared Score: 83.2% of the Dependent Variable is explained by the Independent variables
- RMSE Score: 34861.47 means that we can expect an error of around 35,000 dollars everytime our model makes a prediction
- Baseline Score: 85065.49 against our RMSE shows us that our model makes predictions with less of an error than the mean error, meaning our model is effective

Closing Statements

Closing Statement

My Analysis of the housing price within Ames, Iowa has shown and given:

1. How aspects of the Ames Housing affect the price of the house, And how understanding that can cause an increase or decrease in the price of housing for the Ames, Iowa area.

2. A Model that the Ames Planning & Housing Department could use in order to develop accurate predictions on the cost of houses, to assess whether the houses are too expensive

In Conclusion

In this presentation, we have discussed:

- The Clients Needs
- Purpose of Analysis
- Data Processing
- Exploratory Data Analysis
- Housing Price Projections
- Closing Statement

Questions?