Benjamin Mather

5.2 Final Project Report

20280827

<https://github.com/bpm33/Ames-Housing-Analysis>

Contents

[Dashboard Figure Appendix 2](#_Toc207190026)

[Page 1: Exploratory Data Analysis: Key Market Trends (2006 – 2010) 2](#_Toc207190027)

[Page 2: Predictive Model Performance and Comparison 6](#_Toc207190028)

[Page 3: Feature Importance Analysis 10](#_Toc207190029)

[Page 4: Scenario Analysis 14](#_Toc207190030)

[Project Report: Predictive Modeling of Residential Property Values in Ames, Iowa 19](#_Toc207190031)

[Dataset Selection and Problem Definition. 19](#_Toc207190032)

[Data Cleaning, EDA, and Preprocessing. 19](#_Toc207190033)

[Exploratory Data Analysis 19](#_Toc207190034)

[Handling missing values 19](#_Toc207190035)

[Outliers 20](#_Toc207190036)

[Preprocessing Pipeline 20](#_Toc207190037)

[Baseline Model Statistics and Neural Network Implementation 21](#_Toc207190038)

[Baseline Model Performance Metrics 21](#_Toc207190039)

[Neural Network Implementation and Performance 21](#_Toc207190040)

[Neural Network Performance Metrics 21](#_Toc207190041)

[Dimensionality Reduction 22](#_Toc207190042)

[Feature Importance 22](#_Toc207190043)

[Performance Over Time 22](#_Toc207190044)

[Insights from the scenario analysis. 22](#_Toc207190045)

[Luxury Upgrade 23](#_Toc207190046)

[Fixer Upper 23](#_Toc207190047)

[Green Renovation 23](#_Toc207190048)

[General Insights 23](#_Toc207190049)

[Actionable Recommendations 23](#_Toc207190050)

# Dashboard Figure Appendix

## Page 1: Exploratory Data Analysis: Key Market Trends (2006 – 2010)

The first page of the dashboard is dedicated to Exploratory Data Analysis. Its purpose is to provide a high-level overview of the data and its underlying trends and relationships. This page allows for a preliminary assessment of key features and their connection to home prices, as well as an initial visual discernment of the effects of the housing crisis over time.

**Figures:**

* 1.1 – Year selector
  + A screenshot of a slider

    AI-generated content may be incorrect.
  + This slicer allows users to
  + filter the data by year, enabling the examination of key metrics and a comparison of market conditions before and after the housing crisis.
* 1.2 – Average sales price
  + A green and black sign with black text

    AI-generated content may be incorrect.
  + The target of the analysis. Reflects selections made on the slicer and graphs. Shows the average sale price unless an individual sale is selected on fig. 1.8, which will then show the individual sale price.
* 1.3 – Number of home sales
  + A green and black sign with black text

    AI-generated content may be incorrect.
  + This card tracks the total number of home sales, revealing a slight decline in volume over the years.
* 1.4 – Average overall quality
  + A close-up of a sign

    AI-generated content may be incorrect.
  + Shows the average overall quality for homes sold. Reflects selections made on slicer and graphs. I hypothesized that this would be a key feature in predicting sale price.
* 1.5– Average above ground living area
  + A green and black sign with black text

    AI-generated content may be incorrect.
  + Shows the average above ground living area in square feet. Reflects selections made on slicer and graphs. I hypothesized that this would be a key feature in predicting sale price.
* 1.6 – Average Home Sale Price Over Time
  + A graph showing a line of sales

    AI-generated content may be incorrect.
  + This line chart provides a visual depiction of the housing crisis by showing the average home sale price trend over time, which noticeably drops after 2009. This visual is not affected by slicer selections, providing a constant baseline for comparison.
* 1.7 – Distribution of Overall Quality
  + A graph of orange bars

    AI-generated content may be incorrect.
  + A bar chart showing the distribution of home quality, ranked on a scale 1 through 10 with 10 being the highest quality home. There appears to be a normal distribution of home quality, with a majority falling within quality 5 through 7. The distribution remains more or less stable over the years.
* 1.8 – Sale Price by Above Ground Living Area
  + A screen shot of a map

    AI-generated content may be incorrect.
  + This scatter plot shows a strong positive correlation between above ground living area and sale price. It also highlights several outliers, most notably a very large home that sold for a relatively low price. These outliers are important as they could influence the predictive power of the model and warrant further investigation. The correlation also appears to weaken in later years, which could be due to a decrease in the number of large, expensive homes sold after 2008.

## Page 2: Predictive Model Performance and Comparison

The second page of the dashboard is dedicated to evaluating the performance of the machine learning models. Its purpose is to compare the predictive accuracy and reliability of the Ridge Regression model and the Neural Network. By visualizing key metrics like Mean Absolute Error (MAE) and R-squared (R2 ), this page provides a clear substantiation for the final model selection.

**Figures:**

* 2.1 – Model Filter
  + A close-up of a sign

    AI-generated content may be incorrect.
  + This slicer allows users to select either model to see its individual performance metrics
* 2.2 – Mean Absolute Error
  + A green and black sign with black text

    AI-generated content may be incorrect.
  + A key performance metric displaying the average dollar amount by which the predictions are off. While both models had relatively low error (much lower than the standard deviation), the ridge regression model performed better with a lower mean error.
* 2.3 – Sale Price Standard Deviation
  + A green and black sign with black text

    AI-generated content may be incorrect.
  + The standard deviation of the target, Sale Price, for comparison against the mean absolute error.
* 2.4 – R2 Score
  + A close up of numbers

    AI-generated content may be incorrect.
  + A key performance metric indicating the percentage of variation in Sale Price that is explained by the model’s features. Both models performed well, but the ridge regression model was slightly better at 91% of variance explained.
* 2.5 – Neural Network Actual vs Predicted Sale Price
  + A graph showing a network

    AI-generated content may be incorrect.
  + A scatter plot comparing actual prices against predicted prices made by the neural network model. A perfect model would have all points along the trendline. The model gets less accurate as homes become more expensive. There is one extreme outlier and may be related to the large yet inexpensive home as the model predicted a very high price, but the actual price was quite low.
* 2.6 – Ridge Regressor Actual vs Predicted Sale Price
  + A graph showing the difference between a price and a price

    AI-generated content may be incorrect.
  + A scatter plot comparing actual prices against predicted prices made by the ridge regression model. A perfect model would have all points along the trendline. The model gets less accurate as homes become more expensive. There is one extreme outlier and may be related to the large yet inexpensive home as the model predicted a very high price, but the actual price was quite low.
* 2.7 - Average MAE by Year and Model
  + A graph showing the average of the year

    AI-generated content may be incorrect.
  + A dual line chart showing the mean absolute error of both models over time. This visualization demonstrates that the housing crisis did not negatively impact the model’s predictive power. While it may appear to have improved after the crisis, this may be due to a greater abundance of training data in later years.
* 2.8 – Average R2 by Year and Model
  + A graph showing the average r2 by year and model

    AI-generated content may be incorrect.
  + A dual line chart showing the R squared score of both models over time. This visualization demonstrates that the housing crisis did not negatively impact the model’s predictive power. While it may appear to have improved after the crisis, this may be due to a greater abundance of training data in later years.
* 2.9 – Mean Absolute Error Comparison by Model
  + A graph with green and blue bars

    AI-generated content may be incorrect.
  + A side by side comparison of the mean absolute error of both models indicating that the ridge regression model performed better.
* 2.10 – R2 Score Comparison by Model
  + A graph with a green and blue bar

    AI-generated content may be incorrect.
  + A side by side comparison of the R squared score of both models indicating that the ridge regression model performed better.

## Page 3: Feature Importance Analysis

This page is designed to provide actionable insights into which home features have the most significant impact on sale price. By focusing on the coefficients from the Ridge Regression model, this section provides a clear and interpretable breakdown of the key positive and negative drivers of home value. The interactive elements allow users to explore the data and focus on the most pertinent features or any feature that interests them.

**Figures:**

* 3.1 – Coefficient Range Adjuster
  + A green box with black text

    AI-generated content may be incorrect.
  + A slider that allows users to select different ranges of coefficients, filtering figures 3.5 and 3.6. This enables filtering of features by coefficient to examine only those with negative or positive correlation only, for example.
* 3.2 – Feature Selector
  + A screenshot of a computer

    AI-generated content may be incorrect.
  + A slicer allowing users to select certain feature(s) for examination of their coefficients, filtering figures 3.5 and 3.6. I could select above ground living area and overall quality to see how my initial feature importance hypothesis held up.
* 3.3 – Top 5 Positive Correlation
  + A screenshot of a computer

    AI-generated content may be incorrect.
  + This card provides a quick, at-a-glance summary of the features with the strongest positive correlations to SalePrice.
* 3.4 – Top 5 Negative Correlation
  + A screenshot of a computer

    AI-generated content may be incorrect.
  + This card provides a quick, at-a-glance summary of the features with the strongest negative correlations to SalePrice.
* 3.5 – All Features By Importance
  + A screenshot of a computer

    AI-generated content may be incorrect.
  + A bar chart displaying all of the features available, sorted in descending order by coefficient. Bars are color coded with the legend indicating green is strong positive correlation and red is strong negative correlation. The strongest impacts were negative with clay roof having a strong negative correlation with sales price, indicating perhaps a very undesirable feature. The strongest positive correlation being a particular neighborhood, Green Hill. While my hypothesis wasn’t far off (above ground living area came in the top 5 strongest positive features) I am not surprised at this result. As they say, location, location, location. Unfortunately, there are three miscellaneous features occupying top positive and negative spots and this doesn’t tell us much without further examination of these features and perhaps even removal if determined to be necessary.
* 3.6 – All Feature Coefficients
  + A screenshot of a computer

    AI-generated content may be incorrect.
  + A matrix listing all features and their coefficients. Coefficients are color coded similar to the bar chart, creating a heat map effect that also directly displays the coefficient of every feature.

## Page 4: Scenario Analysis

The purpose of this page is to make the predictive model's findings tangible and actionable for a non-technical audience. It simulates the potential financial impact of various home renovations by allowing users to explore hypothetical scenarios. This not only demonstrates the model's utility but also empowers homeowners to make data-driven decisions regarding buying, selling, or renovating.

**Figures:**

* 4.1 – Scenario Filter
  + A screenshot of a screen

    AI-generated content may be incorrect.
  + This slicer enables users to switch between scenarios for direct comparison of the predicted outcomes.
* 4.2 – Year Sold Slider
  + A green and black text

    AI-generated content may be incorrect.
  + This interactive slider lets users select a specific year from 2006 to 2010. By adjusting the year, users can analyze how the market's performance during the housing crisis affected the outcomes of each renovation scenario.
* 4.3 – KPI Summary
  + A screenshot of a phone

    AI-generated content may be incorrect.
  + This card provides a high-level summary of key performance indicators for the selected scenario and year. It instantly shows the average predicted price gain and the average original sales price, offering a quick delineation of the financial impact.
* 4.4 – Average Price Over Time
  + A graph showing the price of a sold

    AI-generated content may be incorrect.
  + This dual-line chart depicts the average predicted sales price over time compared to the average original sales price. It visually represents the price trend for homes in the selected scenario, revealing how the renovations' value changes with market fluctuations.
* 4.5 – Average Gain Per Scenario
  + A graph of a graph with different colored bars

    AI-generated content may be incorrect.
  + A bar chart that visually compares the average price gain across all three scenarios. This figure makes it easy to quickly identify which type of renovation, on average, provides the greatest return on investment based on the model's predictions.
* 4.6 – Impact on Original High vs. Low Value Homes
  + A screen shot of a graph

    AI-generated content may be incorrect.
  + This scatter plot illustrates the inverse relationship between a home’s original value and the predicted price gain from renovations. It provides a crucial insight that, according to the model, more expensive homes tend to see a smaller percentage increase in value from improvements, and in some cases, may even lose value.

# Project Report: Predictive Modeling of Residential Property Values in Ames, Iowa

## Dataset Selection and Problem Definition.

The project’s goal is to address the challenge of accurately assessing residential property values. The chosen dataset, the Ames, Iowa Housing Data for Regression Models, provided a comprehensive set of 2,930 home sales from 2006 to 2010. This period was particularly pertinent due to the onset of the 2008 housing crisis, allowing for an analysis of how a local market behaved during a time of national economic upheaval.

The project was guided by three key research questions:

1. Predictive Accuracy: How accurately can machine learning models predict residential property values and will a volatile market affect the accuracy?
2. Influential Features: What property and neighborhood features are most significant in influencing home values?
3. Actionable Crisis Insights: How can data-driven insights be communicated through visualizations to empower the average home buyer or seller in making more robust real estate decisions?

## Data Cleaning, EDA, and Preprocessing.

### Exploratory Data Analysis

The initial exploratory data analysis (EDA) revealed a rich and complex dataset. The raw data contained 82 columns and 2,930 rows, with numerous features exhibiting missing values and a high degree of skewness. The dashboard’s EDA page provides a high level overview of these trends, including a line chart that depicts the drop in average home sale prices beginning after 2007.

### Handling missing values

* A significant number of features had missing data. These features were imputed with either the median (for numerical data) or a ‘None’ value (for categorical data).
* List of features missing values:
  + Pool QC
  + Misc Feature
  + Alley
  + Fence
  + Mas Vnr Type
  + Fireplace Qu
  + Lot Frontage
  + Garage Yr Blt
  + Garage Finish
  + Garage Qual
  + Garage Cond
  + Garage Type
  + Bsmt Exposure
  + BsmtFin Type 2
  + Bsmt Qual
  + Bsmt Cond
  + BsmtFin Type 1
  + Mas Vnr Area
  + Bsmt Full Bath
  + Bsmt Half Bath
  + BsmtFin SF 1
  + BsmtFin SF 2
  + Bsmt Unf SF
  + Total Bsmt SF
  + Electrical
  + Garage Cars
  + Garage Area

### Outliers

* Outliers were identified, but intentionally not removed, as they likely represented unique, high-value properties rather than data entry errors. The neural network’s architecture was deemed robust enough to handle these values effectively.

### Preprocessing Pipeline

* Identifier columns ‘Order’ and ‘PID’ were dropped
* ‘MS Subclass’, while having a number value, was identified as categorical and flagged for one-hot encoding
* A detailed pipeline was created to automate the preprocessing steps for both the baseline and neural network models. It included:
  + Log transformation for features with high skewness to normalize their distribution.
    - High skewness was defined as having a skewness threshold greater than 0.75
  + Standardization of numerical features to ensure they were on a similar scale, which is crucial for the performance of a neural network.
  + One-hot encoding for all categorical features, which significantly increased the dimensionality of the dataset from 82 columns to 330 columns.

## Baseline Model Statistics and Neural Network Implementation

A Ridge Regression model was selected as the baseline model. It is effective in handling the high dimensionality that resulted from one-hot encoding.

### Baseline Model Performance Metrics

* The Ridge Regression model performed exceptionally well.
  + Mean Absolute Error (MAE): $14,276.62
    - This error is significantly less than the target variable's standard deviation, indicating a high degree of predictive accuracy.
  + R-Squared (R2): 0.9060
    - Clear indication that a simple, linear model was already capable of explaining over 90% of the variance in home prices.

## Neural Network Implementation and Performance

A sequential, feed-forward neural network was built with a clear architecture: an input layer matching the 330 features, followed by three dense hidden layers with a decreasing number of neurons (128, 64, 32), and a single-neuron linear output layer. ReLU activation functions were used to capture complex, non-linear relationships.

While the neural network’s training and validation loss were very low, its final predictive power was less accurate than the simpler Ridge Regression model. This could indicate that the underlying relationships in the data were more linear than initially assumed, or that the neural network required more meticulous (meaning precise and with great attention to detail) hyperparameter tuning to surpass the baseline.

### Neural Network Performance Metrics

* Final Training Loss: 0.0060
* Final Validation loss: 0.0208
  + Training and validation loss indicates the model learns well on seen and unseen data.
* Mean Absolute Error (MAE): $18,706.93
* R-Squared (R2): 0.8225

## Dimensionality Reduction

Principal Component Analysis (PCA) was performed to understand feature importance and test if reducing the dimensionality would improve performance. The analysis confirmed that a home’s overall quality, size, and newness were significant drivers of price. However, when the models were re-trained on the reduced feature set, their performance was drastically degraded, confirming that the full feature set was necessary for the most accurate predictions.

## Feature Importance

The Ridge Regression model’s coefficients provided a more lucid view of feature importance.

* Strongest Positive Correlation
  + Neighborhood\_GrnHill
  + The Neighborhood of Green Hill had the highest positive impact on home sale price, reinforcing the classic real estate aphorism, “location, location, location.” Since a house cannot be moved, this was not a viable option for adjustment in future scenarios.
* Strongest Negative Correlation
  + Roof Matl\_ClyTile
    - A clay tile roof was identified as a highly undesirable feature and presented a good opportunity for feature adjustment in the scenario analysis.
* Unclear features
  + Several miscellaneous features had high coefficients, but their meaning is less clear without further investigation into the raw dataset.

## Performance Over Time

Running the models on a yearly basis to track their performance confirmed that they remained resilient during the 2008 housing crisis. The evaluation metrics actually improved over time, a finding likely due to the increasing amount of training data available in later years. The models did not show a decline in predictive power during the crisis, which is a key finding for crisis preparedness.

## Insights from the scenario analysis.

To make the findings more actionable for non-technical audiences, a scenario analysis was performed. This involved generating three hypothetical scenarios, simulating users making improvements to homes based on the models' feature importance analysis.

### Luxury Upgrade

* Targeted average homes ranging in Overall Quality (5 – 7) and takes three top positively correlated features and changes them to the most desirable options, simulating home improvement that spares no expense.
* This scenario displayed the highest gain in Sale Price, with the average home sale increasing by $50,000

### Fixer Upper

* Targeted lower quality homes, ranging in Overall Quality (1 – 6) and takes three top negatively correlated features and changes them to a more neutral, and assumably more affordable option, simulating a scenario where improvements are restricted by a lower budget.
* While this scenario still displayed positive improvements to sales price, it resulted in the lowest average change with the average increase being just under $4,000.

### Green Renovation

* Targeting all homes in the dataset, this scenario was designed to test the impact of eco-friendly upgrades on home value, irrespective of their feature importance rankings. Features relating to heating and cooling were upgraded to more modern and efficient options.
* This scenario displayed slightly better results than the ‘Fixer Upper’ scenario, with an average price gain of roughly $6, 000.

### General Insights

All scenarios displayed an increase in average home sale prices over time, generally mirroring the price decline shown in the EDA. A key insight from this analysis was the inverse relationship between a home's original sale price and the gains from improvements. The more expensive a home was originally, the less was gained after improvements, with some higher-priced homes even losing value. This insight is perspicuous in the dashboard’s visualizations, where the predicted price changes show a negative trend relative to the original price.

## Actionable Recommendations

Based on the model’s performance and the scenario analysis, the following actionable recommendations can be made:

1. Property Valuation:
   1. The ridge regression model is a reliable and accurate tool for assessing a home’s value, explaining over 90% of the variance in sale prices. It can be confidently used as a robust, data-driven alternative to traditional valuation methods.
2. Feature Importance for Homeowners:
   1. To increase a home’s value, focus on the property’s overall quality and size. The scenario analysis revealed that strategic renovations and upgrades to features like living area, garage quality and roof material have a quantifiable positive impact on home value for the average home.
3. Crisis Preparedness:
   1. The Ames housing market experienced a steady decline in sale prices from 2007 to 2010. However, the predictive power of the models remained durable, with evaluation metrics improving over time. This suggests that a home’s core features remain the most significant drivers of value, even during market volatility. Homeowners can therefore view investments in these key features as a sound strategy to maintain or even increase a home’s value through periods of economic uncertainty.