## Introduction

The real estate market serves as a vital component of the economy, with housing prices influenced by a wide range of factors including property characteristics, location, and market dynamics. This project focuses on analyzing and predicting housing prices in Ames, Iowa, leveraging a robust dataset and machine learning techniques to explore the relationships between key property features and sale prices.

The dataset utilized for this project contains detailed information on 2,930 residential properties in Ames, encompassing 81 features. These features cover a broad spectrum, including physical attributes such as living area and basement size, qualitative measures like overall construction quality, and locational characteristics such as neighborhood. The diversity and comprehensiveness of the dataset make it an excellent resource for building predictive models and uncovering insights into housing price determinants. The dataset's relatively controlled environment, with consistent pricing trends and a variety of property types, allows for a deep dive into both localized and generalizable patterns in the mid-sized housing market.

The primary goal of this project is to accurately predict housing prices using machine learning models while simultaneously identifying the key factors that drive these prices. Multiple models were applied, including Multiple Linear Regression, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest, to understand how different techniques perform in capturing the complex relationships between features and sale prices. Each model offers unique strengths, from capturing linear relationships to detecting non-linear interactions and ranking feature importance.

Our analysis revealed several key insights. Features such as Overall Quality, Above Ground Living Area, and Neighborhood emerged as the strongest predictors of sale price. Neighborhood, in particular, demonstrated significant influence, with certain areas commanding consistently higher property values. The models performed well, with the Random Forest model delivering the best accuracy and interpretability, highlighting its ability to capture both complex interactions and feature contributions.

By integrating data-driven analysis and predictive modeling, this project not only provides accurate price predictions but also offers actionable insights into the factors shaping housing prices in Ames. These findings are invaluable for stakeholders such as homeowners, buyers, real estate professionals, and financial institutions, enabling them to make informed decisions and strategies based on a deeper understanding of market dynamics. This analysis sets a foundation for further exploration into mid-sized housing markets, demonstrating the value of machine learning in real estate valuation.

## Data Description

The Ames Housing dataset provides a comprehensive record of 2,930 residential property sales in Ames, Iowa, between 2006 and 2010, containing 81 features that cover structural, locational, and qualitative property attributes. The stable yet diverse housing market in Ames makes it an ideal case study for analyzing property valuation, particularly in mid-sized markets where trends are consistent and insights are generalizable. The dataset's features can be categorized into Property Characteristics, such as 1st Floor Square Footage (1st Flr SF) and Garage Area, which reflect living and storage spaces highly valued by buyers; Location Features, such as Neighborhood, which influence desirability, accessibility, and market value; Quality Indicators, including Overall Quality (Overall Qual), a 1-10 rating of material and construction standards; and Additional Amenities, such as Garage Cars and Total Basement Square Footage (Total Bsmt SF), which enhance functional utility and market appeal.

Six key predictors of SalePrice were identified based on high correlation values: Overall Qual, Garage Area, Garage Cars, 1st Flr SF, Total Bsmt SF, and Neighborhood. These features capture the structural and locational factors that most strongly influence property prices. For instance, Overall Qual emerged as the strongest predictor (correlation coefficient 0.80), underscoring the importance of material and construction quality, while Neighborhood highlighted locational disparities in pricing. Exploratory Data Analysis (EDA) provided further insights into these relationships. The Monthly Average Sale Prices graph showed significant price fluctuations between 2006 and 2010, peaking above $220,000 in early 2006 before dropping to approximately $140,000 by mid-2010, reflecting the prolonged impact of the 2008 financial crisis. Similarly, the Year-Over-Year Price Change graph revealed volatility, with sharp declines of -10% to -20% during 2008-2009, brief recoveries in 2009, and a negative trend persisting into 2010. Seasonal trends were evident in the Average Sale Price by Month chart, where prices peaked in January (~$200,000) and showed slight declines in spring, followed by recoveries during summer and a minor dip in November and December.

Scatter plots further illuminated the relationships between key features and SalePrice. The scatter plot of Garage Area (correlation 0.64) revealed that properties with garages between 400 and 800 sq. ft. sold for $100,000-$300,000, while those with larger garages exceeding 1,000 sq. ft. often surpassed $500,000. Total Bsmt SF (correlation 0.63) showed similar trends, with basements in the 500-2,000 sq. ft. range contributing to prices between $100,000 and $300,000, while larger basements were characteristic of luxury homes exceeding $600,000. Likewise, 1st Flr SF (correlation 0.62) highlighted that homes with floor spaces exceeding 2,500 sq. ft. consistently commanded prices above $400,000, particularly when paired with high Overall Qual ratings. Violin plots illustrated the concentration of lower-quality homes (rated 1-4) within the $50,000-$150,000 range, while higher-quality homes (rated 8-10) predominantly exceeded $300,000. The analysis of SalePrice by Neighborhood further confirmed local disparities, with neighborhoods like North Ridge and Stone Brook averaging prices above $300,000, whereas others like Meadow Village and Briardale remained below $150,000.

To ensure robust analysis, missing data and outliers were carefully handled. Median imputation addressed missing values for key numerical variables like Garage Area, Total Bsmt SF, and 1st Flr SF, while frequent category imputation handled missing values in categorical variables like Neighborhood. Features with excessive missing data or limited relevance were excluded to streamline the analysis. These measures preserved data integrity and enhanced the predictive power of the models. The focus on six key predictors—Garage Cars, Garage Area, Total Bsmt SF, 1st Flr SF, Overall Qual, and Neighborhood—allowed the creation of accurate and interpretable models. These findings highlight the dominant role of quality, usable space, and location in property valuation, particularly in mid-sized, stable markets like Ames.

In conclusion, the Ames Housing dataset revealed valuable insights into the determinants of residential property values. Despite market challenges during 2006-2010, structural features such as 1st Flr SF and Garage Area, combined with location and quality indicators like Neighborhood and Overall Qual, consistently emerged as strong predictors of SalePrice. Monthly and year-over-year trends highlighted the impact of macroeconomic factors, while EDA confirmed robust relationships between key features and property values. These insights provide homeowners, buyers, and real estate professionals with actionable knowledge about the Ames housing market, offering broader implications for similar markets.

## Models and Methods

To predict housing sale prices in Ames, Iowa, I explored a variety of machine learning models, leveraging the comprehensive Ames Housing dataset. This dataset, with its 82 features capturing both numerical and categorical attributes, offered a rich foundation for building and testing predictive models. My goal was to compare different approaches to find the most effective one for this task while maintaining a focus on accuracy, interpretability, and reproducibility.

I began by setting up my development environment, using essential Python libraries like NumPy and Pandas for data manipulation, Matplotlib and Seaborn for visualizations, and Scikit-learn for model building and evaluation. This foundational setup allowed for a streamlined workflow from preprocessing to model evaluation.

For modeling, I selected several regression techniques that represent a range of strengths and methodologies. I started with Linear Regression for its simplicity and ease of interpretation. Next, I added the K-Nearest Neighbors (KNN) Regressor, a non-parametric model that relies on feature similarity to make predictions. I also implemented the Decision Tree Regressor, known for its clear decision rules, and the Random Forest Regressor, a robust ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting. This diverse selection allowed me to compare models with different complexities and capabilities.

Preprocessing was a critical step due to the dataset's complexity and diversity. Using Scikit-learn’s tools, I built a preprocessing pipeline that standardized numerical features with

StandardScaler, encoded categorical features like neighborhood using OneHotEncoder, and handled missing values with SimpleImputer. To manage these transformations efficiently, I used Pipeline and ColumnTransformer, which ensured a clean and reproducible workflow.

To evaluate model performance, I adopted a rigorous approach using cross-validation. This technique allowed me to assess how each model performed on multiple subsets of the data, providing a more reliable measure than a single train-test split. For optimization, I applied GridSearchCV, which systematically tested various hyperparameter combinations to find the best configuration for each model. Key evaluation metrics included mean squared error (MSE) to quantify prediction error and R-squared ($R^2$) to measure the proportion of variance explained by the models.

Understanding model behavior was another priority. I used permutation importance to identify the features that had the most significant impact on predictions, enhancing interpretability. For decision tree models, I visualized their structure using plot_tree, which provided a clear representation of how predictions were made. These tools made the models not just accurate but also understandable.

By combining simple models like Linear Regression with more advanced approaches like Random Forests, I could balance interpretability and predictive power. Random Forests, in particular, stood out for their ability to handle complex feature interactions while maintaining robustness. Overall, this systematic approach—from preprocessing and feature engineering to model selection and interpretation—enabled me to develop a reliable framework for predicting housing prices, offering valuable insights into the factors driving property values in Ames.

## Results and Interpretation

This analysis of the Ames housing dataset involved implementing several predictive models to understand their strengths, weaknesses, and ability to capture the underlying relationships in the data. The models included a baseline predictor, multiple regression, k-nearest neighbors (KNN), a decision tree, and a random forest. Each model's results provide valuable insights into its performance, interpretability, and suitability for predicting house prices.

### Baseline Model

The baseline model served as a starting point for performance comparison. It used a simple approach of predicting the mean sale price for all observations, resulting in a high mean squared error (MSE) of approximately **637,995,983.41**. This substantial error reflects the model's inability to account for variations in house prices, making it unsuitable for practical use. However, the baseline established a benchmark for evaluating the effectiveness of more complex models. Its simplicity and lack of feature dependencies highlighted the importance of incorporating relevant predictors and sophisticated algorithms.

### Multiple Regression Model

The multiple regression model leveraged both categorical and numerical variables, including **Garage Area**, **Total Basement SF**, and **Neighborhood**, to predict house prices. After preprocessing the data with techniques like one-hot encoding and scaling, the model achieved strong results, with training and testing R2R^2R2 values of **0.857** and **0.848**, respectively, and an

MSE of **151,989,495.20** (training) and **156,158,235.88** (testing). This model demonstrated significant improvement over the baseline by capturing linear relationships and feature interactions. However, its reliance on linear assumptions limited its ability to model nonlinear dependencies, particularly for high-value properties. Residual patterns indicated that while the model performed well overall, it struggled to address outliers effectively.

### K-Nearest Neighbors (KNN) Model

The KNN model provided a proximity-based approach, capturing localized variations in housing prices based on neighborhood characteristics. By tuning the number of neighbors (kkk) via cross-validation, the model achieved a testing R2R^2R2 of **0.732** and a training R2R^2R2 of **0.785**. While the KNN model effectively captured localized patterns, it struggled with high-dimensional data and rare feature combinations. Additionally, the computational cost of finding nearest neighbors for larger datasets presented scalability challenges. Despite these limitations, the KNN model performed well in identifying neighborhood-specific pricing dynamics, showcasing its strength in understanding localized trends.

### Decision Tree Model

The decision tree model introduced interpretability by visually representing the decision-making process. With training and testing R2R^2R2 values of 0.734 and 0.711, respectively, the decision tree highlighted the importance of features like Overall Qual and Gr Liv Area. The residual analysis showed that errors were centered around zero, but increased for high-value properties, indicating heteroscedasticity. While the hierarchical decision structure provided clear insights into feature importance, the binary splitting mechanism limited its ability to capture complex interactions. As a result, the decision tree offered simplicity and interpretability but underperformed in terms of predictive accuracy compared to ensemble models.

### Random Forest Model

The random forest model, an ensemble method combining multiple decision trees, emerged as the most robust and accurate predictor. After hyperparameter tuning, the random forest achieved training and testing R2R^2R2 values of 0.885 and 0.831, with an MSE of 145,332,733.50 (training) and 183,719,733.02 (testing). It effectively captured nonlinear relationships and generalized well, demonstrating resilience to overfitting. Key features like Overall Qual and Gr Liv Area were consistently identified as critical determinants of house prices. However, residual analysis revealed slightly higher errors for luxury properties, suggesting that additional feature engineering or ensemble techniques like gradient boosting might further enhance performance.

All advanced models outperformed the baseline significantly, highlighting the value of incorporating features and using sophisticated algorithms. The multiple regression model excelled in capturing linear relationships, while KNN effectively modeled localized dynamics. The decision tree provided clear interpretability but struggled with complex interactions. Ultimately, the random forest stood out for its balance of accuracy, generalization, and flexibility.

Residual patterns across models revealed increasing errors for high-value properties, indicating a need for specialized handling of outliers and heteroscedasticity. Furthermore, while ensemble models like random forests performed best overall, their computational demands and relative opacity in decision-making warrant consideration in real-world applications.

## Conclusion

The exploration of the Ames housing dataset through multiple predictive models has provided valuable insights into the relationship between various housing features and their influence on sale prices. The models ranged from a simple baseline predictor to sophisticated algorithms such as random forests, each offering unique advantages and limitations. Together, they not only revealed the dataset's structure but also highlighted areas for improvement and opportunities for further analysis.

The baseline model, while rudimentary, served as a benchmark to gauge the efficacy of more advanced methods. By predicting the mean sale price for all observations, the baseline established the need for models that could incorporate meaningful features to reduce predictive errors. The high mean squared error (MSE) of over 637 million underscored the model's inability to explain variability in the data, setting the stage for more refined approaches.

The multiple regression model introduced feature engineering and preprocessing techniques such as scaling and one-hot encoding. By leveraging features like Garage Area, Total Basement SF, and Neighborhood, the model demonstrated the predictive power of linear relationships. Its testing R2R^2R2 value of 0.848 and relatively low MSE highlighted its strength in handling numerical and categorical features. However, residual analysis showed that the model struggled with outliers and nonlinear interactions, suggesting limitations in its ability to generalize complex relationships.

The K-nearest neighbors (KNN) model excelled at capturing localized patterns in the data, particularly neighborhood-specific pricing trends. By tuning the kkk parameter, the model achieved a testing R2R^2R2 of 0.732, demonstrating its utility in identifying spatial pricing dynamics. However, KNN's performance was hindered by its sensitivity to high-dimensional data and computational inefficiency for large datasets. This model highlighted the importance of feature selection and dimensionality reduction for optimizing proximity-based methods.

The decision tree model provided interpretability, enabling a visual representation of feature importance and decision rules. Its hierarchical structure revealed key insights, such as the dominance of Overall Qual and Gr Liv Area in price determination. With a testing R2R^2R2 of 0.711, the decision tree balanced simplicity with predictive power. However, the binary splitting mechanism limited its ability to capture complex, nonlinear interactions, as evidenced by heteroscedasticity in residual patterns and increased errors for luxury properties.

The random forest model emerged as the most robust and accurate predictor. By aggregating decisions from multiple trees, it achieved a testing R^2 of 0.831 and the lowest MSE among the models. The ensemble method captured nonlinear relationships effectively and generalized well, demonstrating resilience to overfitting. Feature importance analysis consistently identified Overall Qual and Gr Liv Area as critical predictors, reinforcing their significance in housing valuation. Nonetheless, higher errors for high-value properties suggested potential for further improvement through additional feature engineering or complementary ensemble techniques.

## Next Steps

To enhance the predictive capabilities of my models and gain deeper insights into the Ames housing market, incorporating additional contextual, economic, and structural features would be invaluable. These enhancements would allow for a more comprehensive understanding of the diverse factors that influence housing prices, enabling the models to better capture the intricacies of the market and improve their predictive accuracy.

A critical area of improvement would be the inclusion of detailed neighborhood characteristics. Data on school quality ratings, crime rates, and proximity to essential public amenities like parks, grocery stores, and hospitals could provide a more nuanced understanding of how location impacts property values. These factors often drive buyer preferences, and their integration into the analysis would allow the models to better differentiate between similar properties in different neighborhoods.

Another key addition would be historical price trends. By incorporating data on how property values have evolved over time across different neighborhoods, the models could account for temporal fluctuations and broader economic conditions such as interest rate changes and inflation. This would situate current housing prices within a larger market context and allow the models to predict long-term value trajectories more effectively.

Details on property renovations and maintenance history could also significantly enhance the models. Information on recent upgrades to key areas like kitchens and bathrooms or the addition of energy-efficient appliances would quantify the value added by these improvements. Such data would allow the models to capture differences in buyer interest and pricing for similarly sized properties that differ in quality or condition.

The growing emphasis on sustainability makes the inclusion of energy efficiency and green features particularly important. Data on features such as solar panels, energy-efficient HVAC systems, and other sustainable upgrades would help the models assess how environmentally friendly homes influence buyer preferences and pricing. These features increasingly serve as selling points for modern buyers and could add another dimension to the analysis.

Seasonal market dynamics represent another area worth exploring. Housing prices often fluctuate based on the time of year, with spring generally seeing more buyer activity and competition compared to winter. Including this information would enable the models to better predict how seasonal trends impact property valuations and buyer behavior.

Accessibility factors, such as commuting times to employment hubs, proximity to major highways, and the availability of public transportation, would further enrich the analysis. Buyers frequently prioritize convenience and reduced commute times, and these elements significantly influence property desirability. Adding such data would enhance the models' ability to understand location-based premiums.

Demographic and economic indicators like median household income, population growth, and employment rates could provide valuable context about buyer demographics and affordability in specific neighborhoods. These factors would allow the models to assess broader market pressures and economic conditions that influence property demand and pricing.