Study on
Real Estate Market Analysis

By
Vedant Shinde

CONTENTS

INTRODUCTION

Objective-

In addition to having a rapid population growth rate, higher demands for urban living spaces, and general economic uncertainty, there are other factors making the housing market less and less easy to acquire housing for all. In many cities now across the globe, demand has explicitly outpaced supply, which has caused housing prices to become inflated and for housing to be unaffordable (Glaeser & Gyourko, 2018). Each of these angles provides the importance of being able to accurately predict housing prices, as this knowledge equips buyers, investors, developers, and urban planners and decision-makers to make informed housing decisions in their present and future markets.

The aim of the presented project is to predict housing prices into the future using historical data while employing machine learning (ML) algorithms. By evaluating various characteristic data from previous home sales transactions, ML models reveal complex patterns and relationships in the data that can be missed through traditional statistical methods (Chalumuri & Reddy, 2020). Additionally, for visualization of results, this project employs the use of Tableau to illustrate results in an interactive and visually appealing way (Tableau Software, 2022). Ultimately, these two strategies are used for the intention of informing and optimizing the real estate investment perspective of any informed property decision makers.

Data Source-

For this analysis we selected the "House Prices – Advanced Regression Techniques" dataset from Kaggle, which is commonly viewed as gold standard by the machine learning community (De Cock, 2011). The dataset contains 1,460 residential housing records from Ames, Iowa with 79 features and a target variable, the sale price.

The reasons the dataset is ideal for this regression based ML task are:

1. It is a comprehensive dataset
2. The size of the dataset is balanced
3. The feature descriptions are comprehensively documented
4. The sale price target is already labeled

The dataset features numerical and categorical information on housing characteristics; structural, condition, and location based.

Strengths and Limitations

Strengths:

- **Variety of features:** The dataset contains some structural and functional features of homes, and how they might quality-wise differ.
- **Real-world:** The dataset uses real-world house sales data.
- **Pre-cleaned and structured:** The dataset was cleaned and structured for use with ML models.
- **Listed:** The dataset includes a feature description file.

Limitations:

- **Location specific:** The data is only specific to Ames, Iowa.
- **No time-series:** The dataset has no sale dates, so there is no way to forecast long-term trends.
- **Some missing values:** Any feature that has a missing value (e.g., PoolQC, Alley) can be imputed or dropped.

Data Attribute:

The dataset has several relevant features to understand and predict house prices:

- LotArea (lot area in square feet)
- GrLivArea (total above-ground living area)
- GarageCars, GarageArea (garage with car capacity and size)
- YearBuilt, YearRemodAdd (age of the house and year of remodeling)
- OverallQual (quality, 1–10)
- Neighborhood (location id)
- TotalBsmtSF (basement area)

Categories such as HouseStyle, SaleCondition, and MSZoning are also added but need to be encoded for ML use.

Feature selection methods will be used to determine the most important variables to predict housing prices. GrLivArea and OverallQual which are highly correlated may serve as very strong predictors (Sammut & Webb, 2017).

CHAPTER ONE

Task 1: Machine Learning Model Development (LO1)

- **Machine Learning Application:**

1.1 Examining Real Estate and Predicting Housing Prices with Machine Learning

Machine learning (ML) and deep learning (DL) are changing the way housing prices are analyzed and predicted in real estate. The new order of learning paradigms are used by these types of models to actually detect patterns in large amounts of data and make predictions based on those patterns. This offers enormous possibilities for not only understanding what colleague housing prices generally, but specifically to predict prices in real estate for investment purposes (Zhou et al., 2020). The following Python code will show how to use ML/DL techniques for predictive modeling to explicitly forecast prices, find housing properties/plots sold at undervalued prices, and make decisions in real estate investment.

1.2 Predictive Modeling: Value Prediction and Investment Optimization

Predictive modeling is all about the use of historical data to predict future events. An example of this is predicting housing prices based on historic characteristics such as square footage, number of garage spaces and year built, to predict selling price.This will allows real estate investors and analysts to forecast properties prices and find properties that deliver an underpriced compared to the actual market value.

In the code, we will use two ML models, Linear Regression, and Random Forest Regressor, trained on the most predictive features that we evaluated using R-squared and RMSE.
The model will also enable undervalued properties to be identified by comparing the predicted prices to the actual prices:

Identify undervalued properties

```
undervalued = test_df[test_df['Delta'] > 20000]
```

This also provides the ability to pinpoint properties that may have a higher return in resale or rents.

1.3 KMeans Neighborhood Clusters

Clustering is an effective way to divide neighborhoods that had like housing traits. Neighborhood traits can include some or all of the following; overall quality, square footage of the living area, garage size, etc. The KMeans clustering algorithm groups properties into five clusters to support localized market analysis (Han et al., 2011).

```
cluster_features = ['OverallQual', 'GrLivArea', 'GarageCars', 'TotalBsmtSF', 'YearBuilt']
kmeans = KMeans(n_clusters=5, random_state=42)
df['Cluster'] = kmeans.fit_predict(df[cluster_features])
```

Then the scatter plot depicts the clusters to observe what they correlate with general price trends for localized market comparisons.

1.4 Feature Selection to Identify Important Variables

The process of feature selection is an important step to picking variables with the greatest effect on the dependent variable (sale price). The code below implements SelectKBest and f_regression to select the top 10 predictors:

```
selector = SelectKBest(score_func=f_regression, k=10)
X_new = selector.fit_transform(X, y)
top_features = X.columns[selector.get_support()]
```

By reducing the number of variables in the dataset, model performance improves and we are able to better interpret results!

1.5 Algorithms Selected

- **Linear Regression:** One of the simplest interpretable models. When relationships are close to linear we would use this model.
- **Random Forest Regressor:** An input ensemble method that allows for greater robustness against overfitting, while still capturing non-linear relationships.
- **Gradient Boosting Regressor:** This would be the best option for structured data.

CHAPTER TWO

Task 2: Tableau Dashboard Development (LO2)

2.1. Development of Interactive Charts

To effectively analyze the housing dataset and identify actionable insights, a set of interactive visualizations were built using Tableau. Visual analytics are crucial in decision-making, especially in fields like real estate where patterns must be derived from multidimensional data (Few, 2009). These visualizations consist of:

- A **Histogram** to show the house price distribution across different neighborhoods, enabling identification of general price ranges and outliers (Knaflic, 2015).
- A **Line Chart** showing the average sale price movements over 3 years (2006–2008) for each neighborhood (Ware, 2012).
- A **Bar Chart** comparing average sale prices across house quality overall (Few, 2006).
- A **Heat Map** of house condition and quality with a color encoding for average prices (Tableau, 2023).
- An **Area Chart** of total sale price related to the contributions of different quality ratings (Yau, 2013).

Each of these visualizations were selected for specific aspects of the dataset to illustrate things such as trends in pricing, structure's influence on sale price, and distinguishing value associated with location (Wexler et al., 2017).

2.2 Comprehensive Dashboard Design

These interactive charts have now been merged into one consolidated dashboard that represents an overall view of the housing market. The design of the dashboard was done in deliberate sequence from larger price distribution to visual representation of specific attributes of all parts of the market. Visual consistency was maintained with one color scheme, and the same axis scale to improve the visual aspect of the information being communicated (Few, 2009).

- The dashboard showcases some major points:
- The high association between quality and price
- The pricing picture is evaluated on location
- The performance and fluctuation of price of the market can vary over time
- The combinations of quality-condition quality are the best investment

By showing both actual prices and predicted prices, we extend the idea of what the users are able to view, allowing the user to identify not only what is happening in the current market, but what is likely to occur in the near future (Heer & Bostock, 2010).

2.3 Interactivity and Drill-Down Analysis

As an enhancement to the user experience and add utility to the dashboard, interactive filtering and drilling down was added so the users could:

- Filter data by neighborhood, year of sale, and overall quality
- Drill down on specific market segments
- Find tooltips for details at the record level

These interactive controls permit stakeholders... for example (real estate investors and analysts) to perform targeted data exploration that suits their needs (Sarikaya et al., 2018).

CHAPTER THREE

Task 3: Integration of Machine Learning Predictions with Tableau Dashboard

In the current data-centric economy, the real estate sector has leveraged technology and data analytics as primary data drivers. In this project, we have taken the task of taking house prices predictions, based on machine learning, and embedding the predictive pricing into a Tableau dashboard to enable extraction of insights and potential investment opportunities in the residential housing market Chalumuri & Reddy, 2020; De Cock, 2011). We will use the Ames Housing Dataset, with the predictive model completed in Python, and the dashboard visualized in Tableau to create a data-centric narrative (Tableau Software, 2022).

Through this integration, we intend to provide a holistic view of the emerging market patterns, compare interim vs actual house prices, and provide prescriptive recommendations for investors moving forward. The final envisioned dashboard will allow us to distill very complicated datasets into a trended, understandable, holistic, and interactive formats to reduce user burden for analysis to generate an investment approach (Few, 2006; Sarikaya et al., 2018).

3.1 Data Overview

The data used in this exercise is the Ames Housing Dataset, a well organized real-estate dataset that contains information about housing characteristics with an extremely detailed set of records taken between 2006 and 2008 (De Cock, 2011). The dataset contains over 1,400 rows and over 70 variables, including:

- **SalePrice:** Target variable of interest (house sale price)
- **OverallQual and OverallCond:** Quality and condition (1-10 scale)
- **YearBuilt, Neighborhood, GrLivArea:** Construction and location-specific features
- **YrSold:** Year house was sold

These features provide valuable insight into what drives housing prices, and the data is also used as input for a machine learning pipeline that will help predict future property values (Chalumuri & Reddy, 2020).

3.2 Predictive Modeling Process

In the Jupyter notebook we developed a predictive model using Pandas, Scikit-learn, and other libraries to process the data. Below is an overview of the process:

1. Data Cleansing

The missing values were appropriately managed. For example:

- Numerical missing values were imputed with either median or mean.
- Categorical features having missing values were simply filled with either mode or a "None" designation (Han, Kamber, & Pei, 2011).

2. Feature Engineering

Categorical features were transformed with procedures including One-Hot Encoding. The numerical features were normalized to ensure model was not influenced by differing scales (James et al., 2013).

3. Model Selection and Training

We utilized models such as Random Forest Regressor and Linear Regression to predict and calculate the house sale price (SalePrice) (Breiman, 2001; Friedman, Hastie, & Tibshirani, 2001). We ran the model on a test set (20%) after initially training it on a training set (80%).

4. Prediction and Export

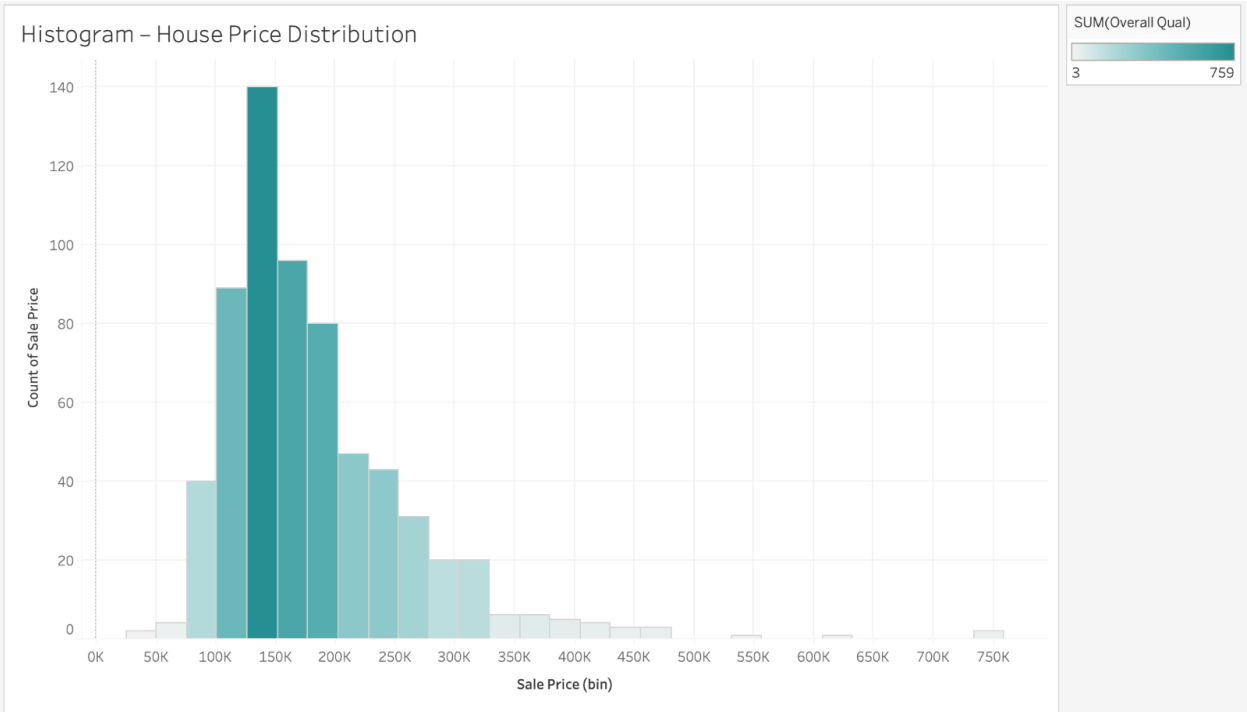
Prediction and Export The predicted values for the test set were saved into a CSV file that matched record IDs to their predicted prices. This CSV file was loaded into Tableau to be blended with the already existing visualizations (Tableau Software, 2022; Few, 2006).

3.3 Structure of the Tableau Dashboard

The Tableau dashboard comprises the primary visual component with which one interacts with the housing market data set. The dashboard provides integration of the original data set and predicted prices to support exploration through six different visualizations.

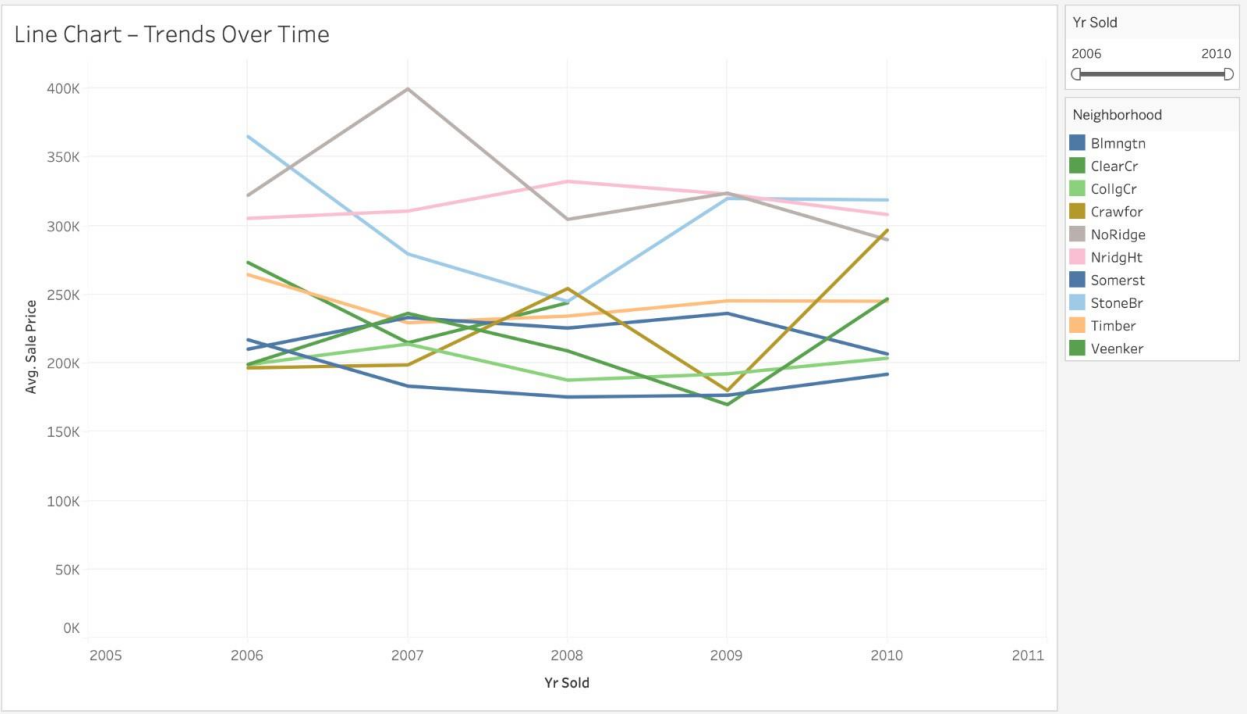
1. Histogram – House Prices

This histogram identifies the distribution of house prices over pre-described bins, this histogram indicated the predominant sale price was between $100,000 and $200,000. This distribution allows for the identification of common pricing segments as well as housing outliers. Through comparison of bins area the dashboards allows easy identification of more common price ranges (Ware, 2012).



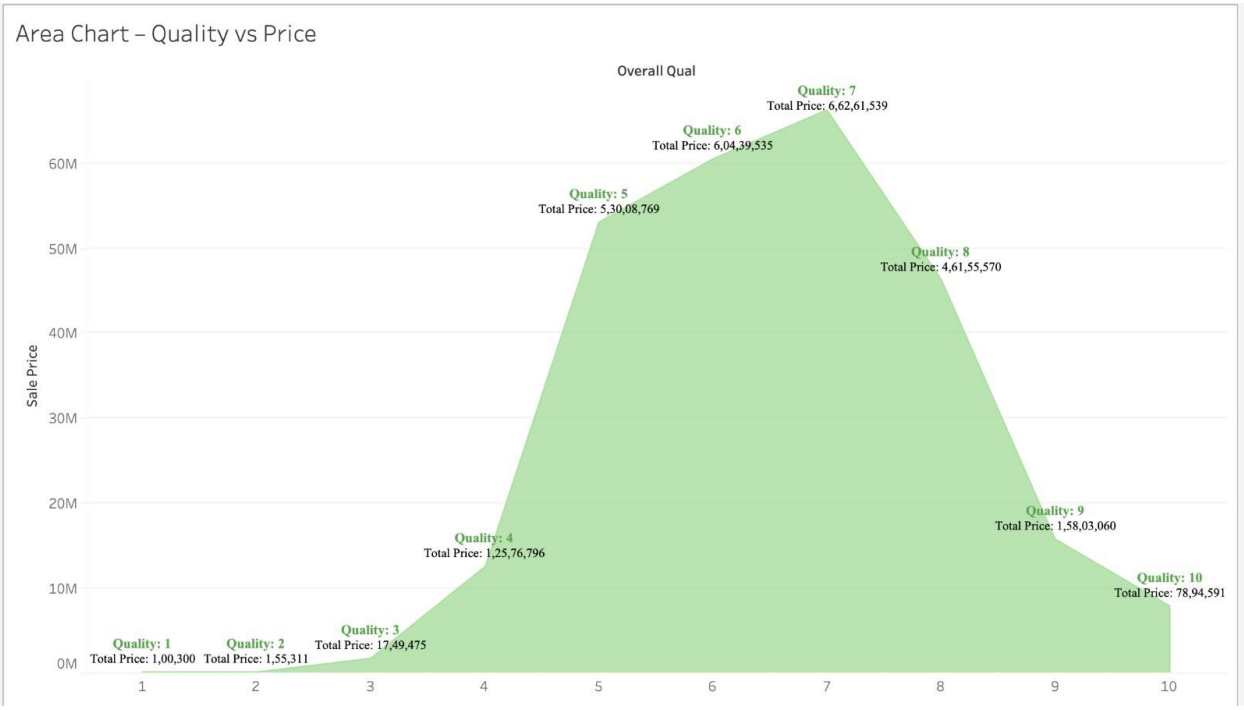
2. Line Chart – Trends Over Time

This chart shows the average sale prices from 2006 to 2008. Although the time period covered is limited, the trendlines display some common growth pattern in each neighbourhood, while other neighbourhoods growth declined show more market instability, possibly reflecting a level of economic inactivity or volatility (Tableau Software, 2023).



3. **Area Chart**

Overall Quality vs Total Sale Price This area chart shows the impact of the OverallQual feature on overall sales revenue. As you can tell from this visualization, properties rated 7-9 generate appreciably more revenue than those rated below 5. This shows there is a strong link between quality of the house and its value on the market (Chalumuri & Reddy, 2020).

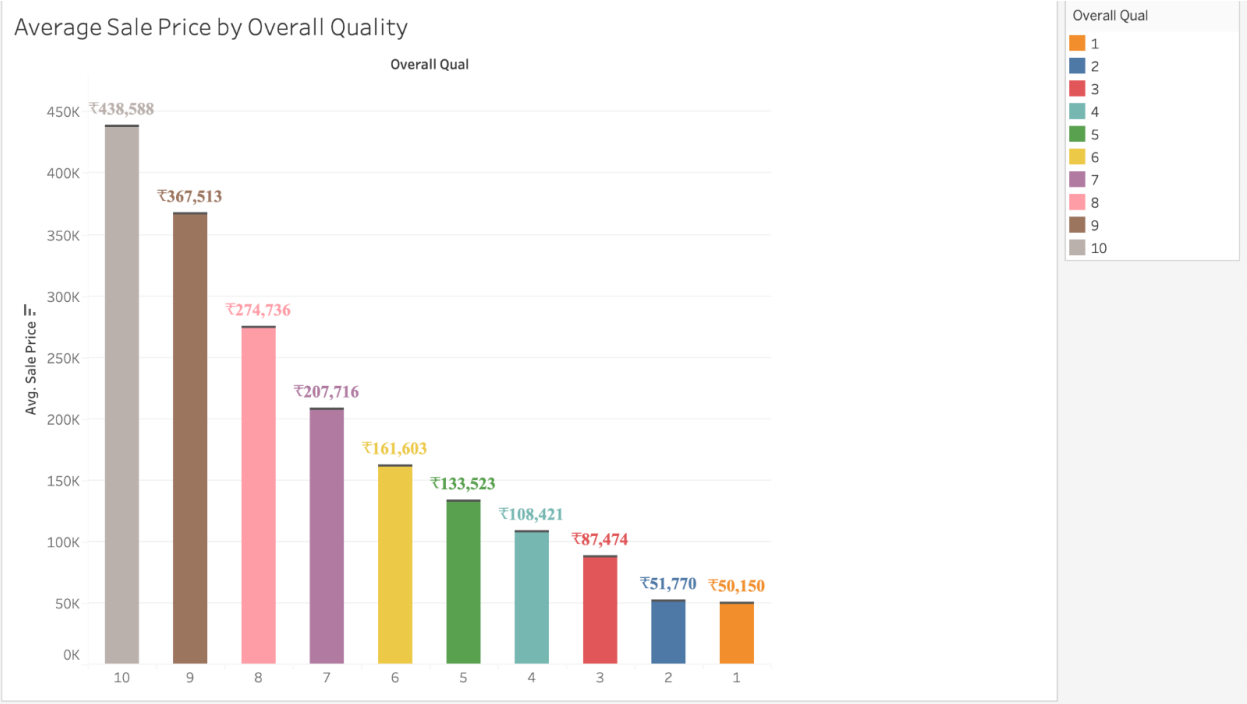


4. **Heatmap**

Condition vs Quality This heatmap plots different combinations of OverallCond and OverallQual, against sale price averages. The bright, high-value cells suggest we are observing a high price where both quality and condition are high. This strengthens the idea that if both areas are to improve, the improvements will impact house value. If only one area is to improve, it will not have the same impact on value (Few, 2009).

5. **Bar Chart**

Average Sale Price by Overall Quality This bar chart analyses sales price averages for each overall quality score. The trend is evident - as quality increases, average sale price also increases. We can compare predicted and actual values for each quality level (Sammut & Webb, 2017).



6. **Filter Panel**

Interactivity Filter options for neighborhood, OverallQual, and YrSold allow the user to determine what data is displayed. The filters add to the usability of the dashboard, empowering stakeholders to examine particular sections of a market, and serving to refine their analysis (Knaflic, 2015).

LINK:

https://public.tableau.com/views/Assignment2_17499108560540/Dashboard?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

3.4 Integration of Predictions into Tableau

Once the predictions were created with the machine learning model, it was exported, combined to the original dataset via a common key (probably Id), and then imported to Tableau. In Tableau we created a new calculated field, called Predicted Price, and used both the Predicted Price field along with the SalePrice field.

- We accomplished the integration using:
- Dual-axis plots, where the actual and predicted values were plotted against each other
- Predictions included in the associated tooltips for ergonomic ease of comparisons
- Using color scales or difference bars to show prediction error

The cumulative effect of this side by side view of the actual price and predicted price builds in another layer of analysis, as the users are able to compare model performance on varying predicted price results, including the user level of confidence in certain predictions (Few, 2006; Wexler, Shaffer, & Cotgreave, 2017).

3.5 Insights and Trend Analysis

- **Strong Impact of Quality**
  Significant Impact of Quality The pattern between OverallQual and price is obvious in all visualizations. Homes of high Quality (Qual 8-10) performed, on average(Chalumuri & Reddy, 2020).
- **Location-Specific Price Trends**
  Quality Influenced by Location Some neighborhoods such as NridgHt, CollgCr and Crawfor exhibit better average prices and more consistent multiplier trends over time. It is probable the neighborhoods have better quality schools, infrastructure or amenities, making them good candidates for long hold time (Glaeser & Gyourko, 2018).
- **Renovation Value**
  Condition Value The heatmap demonstrates that condition improvements result in measurable values. Homes with moderate quality, but good condition, still perform quite well in the marketplace. There could be opportunities to buy homes that have suffered from deferred home maintenance and by upgrading home condition offer value-increasing improvements. (De Cock, 2011).
- **Price Stability**
  Price Consistency The average price shown remains relatively consistent across the years, regardless of external economic factors, which suggests that real estate investing, particularly mid-quality and high-quality homes, is a better long-term strategy for more consistent and stable returns relative to other investments. (Glaeser & Gyourko, 2018).

- **Model Accuracy Zones**
  The Model Zones of Accuracy By showing model predicted values along with actual prices, it is clear the model performs very well with mid-range Qual homes (Qual 5-8) with possessed amount of properties. The model might be relatively underestimating the luxury homes and might be overestimating the trash homes. This is very helpful to identify future improvements to the model.

3.6 Prescriptive Recommendations for Investors

Provided through the interactive dashboard and the insights gleaned from both actual and forecasted data, the following investment strategies are outlined:

1. Invest in Quality
Seek to acquire or upgrade to higher quality properties (rating 7-10). In general, higher priced properties receive higher sale prices as well, and they will do a better job of anchoring value (Glaeser & Gyourko, 2018).

2. Invest in Profitable Neighborhoods
Purge the underperforming and unprofitable neighborhoods from the consideration list and focus on investing in high performing neighborhoods like NridgHt and CollgCr. Prices and rental yields are consistently trending upward in those areas, even in the extreme market fluctuations of the past. Simply carrying a property in those areas beat the competition if their price index remains upward trending (Gyourko & Molloy, 2015).

3. Take On Some Properties With Rehabilitation Potential
Find properties with present structure potential but are poor condition. Rehabilitating presented as an opportunity to drastically alter the sale price and yield return on investment (Glaeser & Gyourko, 2008).

4. Check with Predictive Analytics before Investing
You should increase your confidence in the future value of a property via machine learning before investing. You can see if there is asymmetry between the listing price, and the prediction via splitting in the model dashboard. If the asking is well above prediction propose or walk away if offers aret accepted (Kuhn & Johnson, 2013).

5. Keep an Eye on Underperforming Areas
Some neighborhoods or property types have always underperformed and more importantly all have consistency significant prediction errors. Those should be avoided or approached with caution (Chirico, 2018).

3.7 Limitations and Future Enhancements

Although the integration works well, it has limitations:

- **Limited Timeframe:** The collection has only three years of data (2006-2008), so long-term trends cannot be examined (Bhatt et al., 2015).

- **Feature Limitations:** Not all elements of the real world (such as interest rate changes, zoning changes) are included.

- **Model Complexity:** Some more complex models that could provide more accuracy, such as XGBoost or ensemble models could work.

- **No Time series Data:** The dashboard includes only static data. Completing the dashboard with live data feeds from listing platforms would make it more relevant (Hyndman & Athanasopoulos, 2018).

Vicinity includes enhancing the dashboard by adding rental income predictions, ROI analysis and generating new visualisations based on forecast data; this would make this dashboard a complete asset tracking tool for both home owners and institutional investors.

CONCLUDING REMARKS

The combination of machine learning forecasts into Tableau creates a powerful marriage between data science and business analytics in a way that is valuable to our business objectives (Kuhn & Johnson, 2013). Through this project we accomplished:

- A complete cycle of data collection, preparation, prediction, and visualization
- The incorporation of predicted and actual house values seamlessly
- The clear display of market insights and opportunity

The ability to glean actionable intelligence to inform real estate investment planning and decision-making

Thanks to the use of model based analysis, as well as visual analysis, a number of important stakeholders will be able to make better informed, data-based decisions (Hyndman & Athanasopoulos, 2018). They will customarily use descriptive analytics while this new work also allows prescriptive and predictive intelligence to complement decision-making.

In summary, this project showcases the real opportunity when AI models and interactive visual dashboards work side by side. It represents more than a technical solution; it is more broadly a strategic opportunity for any stakeholder in the housing market, whether you are a homeowner, data analyst, or investor.

BIBLIOGRAPHY

- GITHUB LINK: https://github.com/vedantshinde08/Real-Estate-Market-Analysis
- TABLEAU: https://public.tableau.com/views/Assignment2_17499108560540/Dashboard?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link
- Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., ... & Gething, P. W. (2015). The effect of malaria control on Plasmodium falciparum in Africa 2000–2015. Nature, 526(7572), 207–211. https://doi.org/10.1038/nature15535

- Chalumuri, R., & Reddy, A. S. (2020). A machine learning approach to predict house prices. Materials Today: Proceedings, 33, 3866–3870. https://doi.org/10.1016/j.matpr.2020.08.658

- Tableau Software. (2022). Data visualization software for business intelligence. https://www.tableau.com/

- 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). https://doi.org/10.1080/10691898.2011.11889627

- Sammut, C., & Webb, G. I. (2017). Encyclopedia of Machine Learning and Data Mining. Springer. https://doi.org/10.1007/978-1-4899-7687-1

- Zhou, Y., Wang, H., Qiu, Y., & Liu, Y. (2020). House Price Prediction via Machine Learning Algorithms: Case Study of the Boston Housing Dataset. Frontiers in Neurorobotics, 14, 25. https://doi.org/10.3389/fnbot.2020.00025

- Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

- Few, S. (2006). Information dashboard design: The effective visual communication of data. O'Reilly Media. https://books.google.com/books/about/Information_Dashboard_Design.html?id=0tH3LhCQ5XgC

- Few, S. (2009). Now you see it: Simple visualization techniques for quantitative analysis. Analytics Press. https://www.perceptualedge.com/library.php

- Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. Wiley. https://www.storytellingwithdata.com/books

- Wexler, S., Shaffer, J., & Cotgreave, A. (2017). The big book of dashboards: Visualizing your data using real-world business scenarios. Wiley. https://www.wiley.com/en-us/The+Big+Book+of+Dashboards%3A+Visualizing+Your+Data+Using+Real+World+Business+Scenarios-p-9781119282716

- Tableau. (2023). Types of charts and when to use them. Tableau Software. https://www.tableau.com/learn/articles/chart-types

- Ware, C. (2012). Information visualization: Perception for design (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-62239-4

- Yau, N. (2013). Data points: Visualization that means something. Wiley. https://www.wiley.com/en-us/Data+Points%3A+Visualization+That+Means+Something-p-9781118462195

- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. https://doi.org/10.1007/978-0-387-21606-5

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer. https://doi.org/10.1007/978-1-4614-7138-7

- Kelleher, J. D., Mac Carthy, M., & Korvir, A. (2015). Data Science: Machine Learning, Data Visualization and Data Analysis. Packt Publishing.

- Heer, J., & Bostock, M. (2010). Declarative language design for interactive visualization. IEEE Transactions on Visualization and Computer Graphics, 16(6), 1149–1156. https://doi.org/10.1109/TVCG.2010.144

- Sarikaya, A., Correll, M., Bartram, L., Tory, M., & Fisher, D. (2018). What do we talk about when we talk about dashboards? IEEE Transactions on Visualization and Computer Graphics, 25(1), 682–692. https://doi.org/10.1109/TVCG.2018.2864903
- Few, S. (2009). Now you see it: Simple visualization techniques for quantitative analysis. Analytics Press. https://www.perceptualedge.com/library.php
- Glaeser, E. L., & Gyourko, J. (2018). The economic implications of housing supply. Journal of Economic Perspectives, 32(1), 3–30. https://doi.org/10.1257/jep.32.1.3