**Ames Real Estate Final Project**

Wayne State University – Mike Ilitch School of Business

Professor Nabil Raad

TIS 5570-2409002

Group 6

Ethan McManus

Landon McDowell

Garrett Brady

Marisa Sterbyci

Jawad Javed

12th December 2024

**Problem Statement**

Under the context of business analytics, the first step in the business analytics process is to define and understand the business problem. Considering the complexity that lies within the many variables that make up a property listing, the ability to predict the price of homes with high accuracy is an ongoing challenge for both realtors and homeowners. Properties that rely on outdated market trends, or numbers that aren’t tailored to the local market have the ability to produce inaccuracies.

**Purpose of the Project**

For this project, our team was tasked with creating a predictive model in which we analyzed historical housing market data specifically from the city of Ames, Iowa. The goal of this project was to offer insight from the local data as to how homeowners can create listing prices for their houses based on a variety of factors collected about homes sold in the city. By analyzing the data, we hope to achieve a market price prediction model that accurately reflects the current conditions and specifications of each home, so that our team can offer actionable insights into ways homeowners can increase the value of their homes. Through our project, we hope to answer three questions:

* What is the expected selling price of a home?
* How much should homeowners invest in improving the condition of their home to increase the expected price by more than the cost of improvements?
* What factors most significantly affect the market value of a home?

**Understanding the Data**

After properly defining the business problem and identifying some of the goals this project aims to address, it is important to better understand the qualities and characteristics of the data. We will go into more detail on how we measured the data to better understand it through the data exploration portion of this report.

The data sheet we received had 2931 rows, each containing unique properties, along with 82 columns containing each property's respective variables. The variables within this dataset range from qualities and descriptions of house attributes, physical conditions and objects contained within the property, geographic information such as positioning within a neighborhood, and lastly information pertaining to the sale of a house, such as year and month sold, the condition of sale, and the overall sales price. Our initial goal was to identify some of the basic assumptions and conditions of the data so we could go forward with a better understanding of our dataset. For our assumptions, we assumed that the data from the project accurately measured real data from the city of Ames, Iowa and that there were no significant errors within the data itself. For blank values (aside from Lot\_Frontage), we assumed those to be recordings of 0 for those variables or N/A, meaning that there should be no information for those fields as they do not exist. The conditions of the dataset include variables that could remain subjective. These subjective variables include overall quality (ordinal) and overall condition (ordinal) ratings, so their ranking can vary depending on the person who measured such variables. Our insights and recommendations mainly rely on the data provided from Ames Iowa, not accounting for any economic or market trends that might have an impact on the sales price of houses within the city. Lastly, it is important to note that our exploration, analyses, recommendations and insights are all pertaining to 1 family households. Any other household type has been filtered out of this project, as we believed the scope for conducting analysis would become too wide if we were to account for all household types.

**Data Exploration and Transformation**

Once we understood some of the conditions and assumptions of the data, we created a pivot table to gather every key variable that we believed to be the most impactful for the sales price of a house. This method was not statistical, but instead, came from our initial ideas as to which variables we assumed would most significantly impact the price of a property. Essentially, we didn’t want to include variables like PID and Order in our descriptive analytics as they aren’t relevant for our findings. We also avoided including variables like the quality of a fireplace as we did not feel that the overall price of a house would be drastically affected by these specific independent variables. Essentially, if we were to account for every variable in the dataset, the complexity of our model would make interpretation rather difficult. This idea aligns with Occam’s Razor, which is a philosophical rule that prioritizes simpler theories or recommendations over the more complex ones. The following table contains the top ten variables within our descriptive statistics pivot table, sorted in descending order for correlation with the variable SalePrice.

A screenshot of a data sheet

Description automatically generated

After looking at our pivot table, it was clear that some of our data needed to be cleaned and transformed. We looked into the specific variables that had blank values and tried to understand whether those blank values mean that there is no measurable variable for that respective property (e.g. blank for Garage\_Yr\_Blt means there is no garage for that house). We wanted to differentiate those blanks from erroneous values missing from the dataset, like the variable Lot Frontage in which we filled using the mean of the column. We also created boxplots to identify columns with major outliers, although our data cleaning and transformation process regarding outliers was not thorough enough for every single key variable in our dataset. Our team also felt the need to create dummy variables to measure variables like year remodeled so we can accurately measure the impact that remodeling can have on SalePrice and YearBuilt.

Once we were done transforming our data, we wanted to go further into testing the correlation in dataset. Along with creating correlation coefficient tables within excel, we created a heatmap to visualize the relationship between our key variables.

A chart of different colors

Description automatically generated with medium confidenceThis **Correlation heatmap** provides critical insights into which factors most significantly influence house prices in the Ames housing market.

**Strong Correlation Between Overall Quality (Overall\_Qual) and SalePrice**

The Ames Real Estate Pricing Prediction project uses a Correlation Heatmap to pinpoint the factors most significantly impacting house prices in the Ames housing market. One standout relationship is the strong correlation between Overall Quality (Overall\_Qual) and SalePrice, with a correlation coefficient of 0.79. The element indicates that homes with higher construction quality and superior finishes command higher prices. According to Zhang (2023), such insights are invaluable for real estate agents, homeowners, and developers. For instance, homeowners planning renovations can prioritize areas like kitchens, bathrooms, and structural improvements. The goal of the improvements is to elevate their home’s quality score, subsequently increasing its market value. Similarly, Liu (2023) states that developers focusing on new construction can strategically invest in high-quality materials and finishes to appeal to buyers willing to pay a premium for superior craftsmanship. The correlation also highlights the importance of marketing strategies that emphasize quality features in property listings to attract higher offers. Through focusing on aspects that directly enhance the perceived value of homes, stakeholders in the Ames real estate market can make data-driven decisions that maximize returns.

The variables selected for the correlation heatmap in the Ames project are chosen for their significant influence on housing prices and their relevance to the property market. These variables are strong predictors. They demonstrate high correlations with SalePrice in housing datasets. The element makes them ideal for identifying meaningful patterns. They are also highly interpretable, which allows stakeholders such as developers, realtors, and buyers to derive actionable insights. Additionally, the selected variables provide a comprehensive view of market dynamics by emphasizing factors like quality, size, age, and selling price. In focusing on key drivers, the heatmap simplifies analysis, avoiding unnecessary complexity while highlighting actionable relationships.

**Multiple Linear Regression Analysis**

A table with numbers and a black border

Description automatically generated with medium confidenceUpon looking further into our correlation tables, we noticed that Overall\_Condition had a negative correlation with SalePrice, which wasn’t making sense to the context of properties. We ran boxplots and other tests to try and conclude why this variable was behaving this way, but we were unable to identify a reason that explained this negative correlation. Since we know through our analysis that positive correlations such as Overall\_Qual, Gr\_Liv\_Area, and Garage\_Area will see an associated increase in SalePrice as these variables increase, we want to explore the effect that each of these variables have on the specific increase in price per unit. We also wanted to include Overall\_Condition in this analysis to look further into how this variable affects the SalePrice of a house. Below is our multiple linear regression table containing our focus variables

To reach statistical significance with a confidence interval of 95%, we need to have a **p-value <= 0.05.**

From our regression table, we saw that three of the four independent variables were statistically significant with extremely low p-values, those variables being Overall\_Qual, Gr\_Liv\_Area, and Garage\_Area. After reviewing the regression table, despite the positive associated increase that Overall\_Cond has with SalePrice, we see that Overall\_Cond has a p-value that is 0.1153, which is greater than 0.05, so we could not conclude that there is a meaningful relation between these variables.

Considering the nature of the variable Overall\_Qual being measured on a scale of 1-10 (ordinal), it is difficult to conclude that each unit increase will result in a specific increase for the sale price of a property. To add, for homeowners and realtors, trying to determine specific changes that will increase overall quality by one unit is too vague, which makes acting on these insights extremely difficult. To analyze quality in a better standard that offers better insights for homeowners, we created dummy variables to measure the units from 1-10 as the following categories: 1-2 as “Bad”, 3-4 as “Somewhat Decent”, 5-6 as “Good”, 7-8 as “Very Good”, and 9-10 as “Excellent”. For our refined regression table, we used “Good” as a baseline or reference for the categories “Bad”-“Excellent” to better analyze overall quality at different price points. We also removed Overall\_Cond from the refined table as the p-value was too high and the A table with numbers and a red text

Description automatically generatedcorrelation coefficient was negative. Below is the refined regression table.

We can see that each of our independent variables have p-values < 0.05

**Evaluation**

After evaluating the effectiveness of both of our regression tables, we realized that our refined table provided insights that were easier for homeowners and realtors to act upon, where units of 1-10 are replaced for categories that measure bad-excellent. From our refined regression table, through each improvement in the overall quality of a home, we can see the sale price of a home increase significantly in respect to the category “Good” that acts as a reference of overall quality. The reason we chose “Good” as our reference category is because the variable Overall\_Qual in our dataset has a mean (average) of 6.09 and a mode (most common number) of 5.00. Considering that we hope to offer actionable insights for our homeowners, by choosing “Good” as a reference for quality, which contains the units 5-6 of our variable Overall\_Qual, we would be using the quality category that is most common for homeowners in Ames Iowa. We can also see the range from “Bad” – “Somewhat Decent” are areas for improvement. Those whose house quality is beneath the mean and mode category are missing out on potential gains from improving the overall quality of their house to reach the average of the recorded local market.

For better understanding of this increase, through our refined regression table we can look at the coefficient for the category “Very Good” which has a price increase of $46,571 compared to “Good” where the price is 0. We can then look at the next category above “Very Good” which is “Excellent”, which has a price increase of 158,950 in respect to “Good”. Lastly, for negative values, we can see that for houses beneath the mean category and within the “Bad” range, they are missing out on $48,275. So, we can conclude with our improved regression model that cost-effective improvements that raise quality seem to be quite effective for optimizing return on investment. Lastly, it is important to note that expanding the garage space and above ground living area seem to offer somewhat-desirable returns on investment. We can interpret these insights from the following data: for every unit increase of Garage\_Area (Sq Ft), the SalePrice of a house increases by $75.19, and for every unit increase of Gr\_Liv\_Area (Sq Ft), the SalePrice of a house increases by $79.18

**Implications and Limitations**

While our analysis was conducted based on the data from the Ames dataset, our biggest limitation of this project would have to be the dating of this data. We were given a dataset that contains data that is well over a decade old. Considering the age of this data, market trends and dynamics could have shifted, changing the way some of these variables behave in respect to our dependent variable. Our regression analysis, while helpful for some insight as to how certain independent variables can affect the sales price of a property, should not call for realtors and homeowners to over rely on it. There is much room for error with a model that uses data on a field such as real estate, where the price changes so fluidly with the economy and demand within it. All our insights and conclusions need to consider those limitations. We also had a time constraint, given that we were to complete our analysis of the dataset within the semester. Not every variable that was transformed or cleaned was fully explored in our analysis, such as the Year\_Remodel and dummy variables created for numerical analysis.

**Conclusion**

A graph of a bar graph

Description automatically generated with medium confidenceThe expected selling price of a property in the market on average is $184,812 with the median sale price being $160,000 and a mode of 135,000. The range could often go from as low as $12,789 to as high as $755,000. We can see the average sale price per unit of quality visualized in the chart below.

When investing in improvements for a property to increase its value, many factors will have to be considered. We saw that the greatest variable that changed the value of the home was overall quality (0.7993), along with other major variables such as the above ground living area (0.7061) and garage area (0.6478). What we also know from our data is that the average overall quality for homeowners is a 6.09. In other words, the average home in Ames Iowa has good overall quality. When considering improvements from good quality to very good quality, homeowners will see a respective increase of $46,571, and from good to excellent, they will see a respective increase of $158,950. Also, for homes with quality such as Bad or Somewhat Decent, they will incur sales losses of approximately -$48,275 or -$25,622. This loss is due to the nature of homes within the local market having an average of good overall quality, so homes that are beneath good are likely to be lower in value. It is important to consider these gains and losses for investments purposes for a few reasons.

The cost of improvements will be the target to beat when increasing the value of the property. This is a very important step as it doesn’t make sense financially to put money into the property just to then lose the money when you go to sell it, as the expenses would wind up exceeding the gains from reaching the next quality category. Second, the individual will have to use comparable homes to see the difference in price with homes that have improvements and those without. When determining the viability of an improvement, looking at the cost of the improvement and the gain in value the home will incur will determine if such an improvement will be worth making. Our recommendation for improving house value is to stick with the small details and improvements, focus on quality materials and craftsmanship, if the home price currently is below the mean. Investigating the Above Ground Living Area and Garage Area for improvements/expansion doesn’t hurt even if the current price range of the home is above the median as there might be a possible return. When exploring improvement options, it is important to keep in mind diminishing returns. When putting money into a home, people will think of it as an investment and if the people purchasing are looking to make money on their investment, they will want a property that is not the most expensive for an area as most likely that property will either drop or keep its value, and with inflation it would be a guaranteed loss. Keeping that in mind, improving the quality of the home will be the best option for most of the homes with a potential to improve other variables to increase the home’s value.

**Works Cited**

Liu, Z. (2023). Real estate price prediction based on supervised machine learning scenarios. *Highlights in Science, Engineering and Technology*, *39*, 731-737. <https://drpress.org/ojs/index.php/HSET/article/download/6637/6431>

Nashed, E., & Baron, M. (2023). Tracking residential real estate capital growth in NSW by constructing a price index from sales transactions. *Journal of Data Science and Intelligent Systems*. <http://ojs.bonviewpress.com/index.php/jdsis/article/view/1344>

Teoh, E. Z., Yau, W. C., Ong, T. S., & Connie, T. (2023). Explainable housing price prediction with determinant analysis. *International Journal of Housing Markets and Analysis*, *16*(5), 1021-1045. <https://www.emerald.com/insight/content/doi/10.1108/IJHMA-02-2022-0025/full/html>

Zhang, L. (2023). Housing price prediction using machine learning algorithm. *Journal of World Economy*, *2*(3), 18-26. <https://www.pioneerpublisher.com/jwe/article/view/392>