Multiple Linear Regression of Milwaukee Housing Market

Noah Choate, Eli Ritchie

Introduction

In our report, we collected our data from the City of Milwaukee's government website that consisted of recent housing sales.
With this data, we will utilize Multiple Linear Regression to predict sale prices of houses given their attributes (Year Built, Finished Sq. Ft. etc.) We aimed to define undervalued properties by finding any predicted sales prices of a property valued under the actual sale price. To finalize our report, we then identified the top undervalued properties.

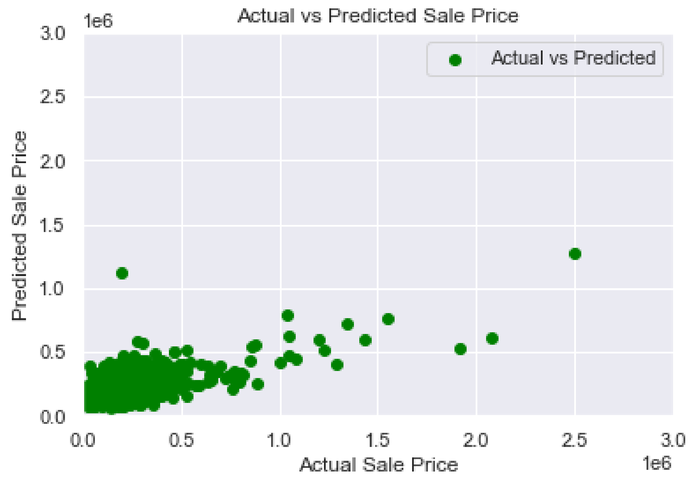
Methods

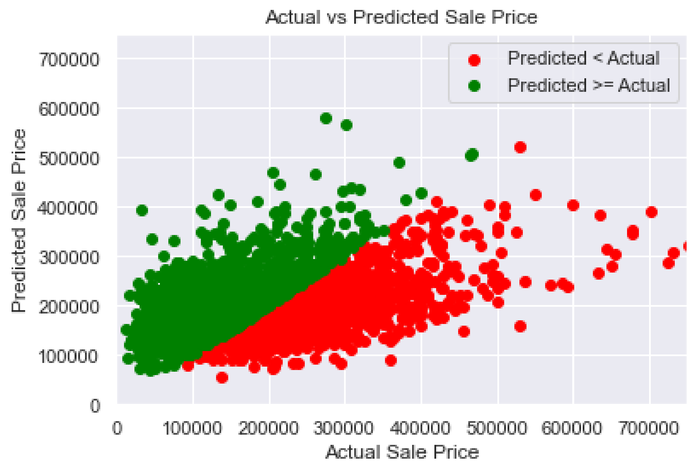
- We began our project by filtering our dataset. We filtered all properties sold to only include residential houses. We then removed columns of the data frame that were irrelevant to our model such as "Style" and "PropertyID"
- To start our Multiple Linear Regression, we identified the variables that have the highest correlation with sales price. Using the model, we compared our predicted values to actual sale values to see if there were any undervalued properties.
- We used the following variables from the given dataset for the model:

1. Year Built
2. Number of Rooms
3. Number of Bedrooms
4. Number of Bathrooms
5. Lot Size
6. Finished Square Feet

- Mathematically, we found the largest positive difference between actual sale price and predicted sale prices to identify the most undervalued properties listed on the market.

Results





Research Questions

- How accurate is the Multiple Linear Regression Model?
- Were there any additional variables not available that would aid our model?
- Who would this information be useful to?
- Are there any glaring undervalued properties from the model?

Conclusion

- The Multiple Linear Regression Model has an R^2 value = 0.285. This means that the model does a poor to moderate job at predicting sale price values of properties. This is likely due to additional information being needed to predict sales prices of a property.
- Additional variables that could aid our model would be information about the land and property itself. Data such as land value and building value could be very beneficial. There are a plethora of other variables that go into the value of a property, those of which we did not have access to in our dataset. More information regarding the properties would lead to a higher R^2 value, improving the predicting power of the model.
- This model could be useful for those entering the real estate market whether it is investing in properties or buying a house to live in. It is important to note that this model should only be a suggestion, and not a fact.
- The 5 most undervalued properties per the model are as follows:
  1. 6253 N 118th ST
  2. 4960 N Sherman BLVD
  3. 5925 N 74th ST
  4. 4459 82nd ST
  5. 4112 N 50th ST