BA222

Lot Area Causal Analysis &

Predicting Housing Prices

May 7, 2024

Jonathan Yung (U28856227),

Ethan Freshman (U63289110)

Ashley Richard (U82939847)

Arya Sukarno (U42813741)

James Rahman (U13313207)

**Data Description**

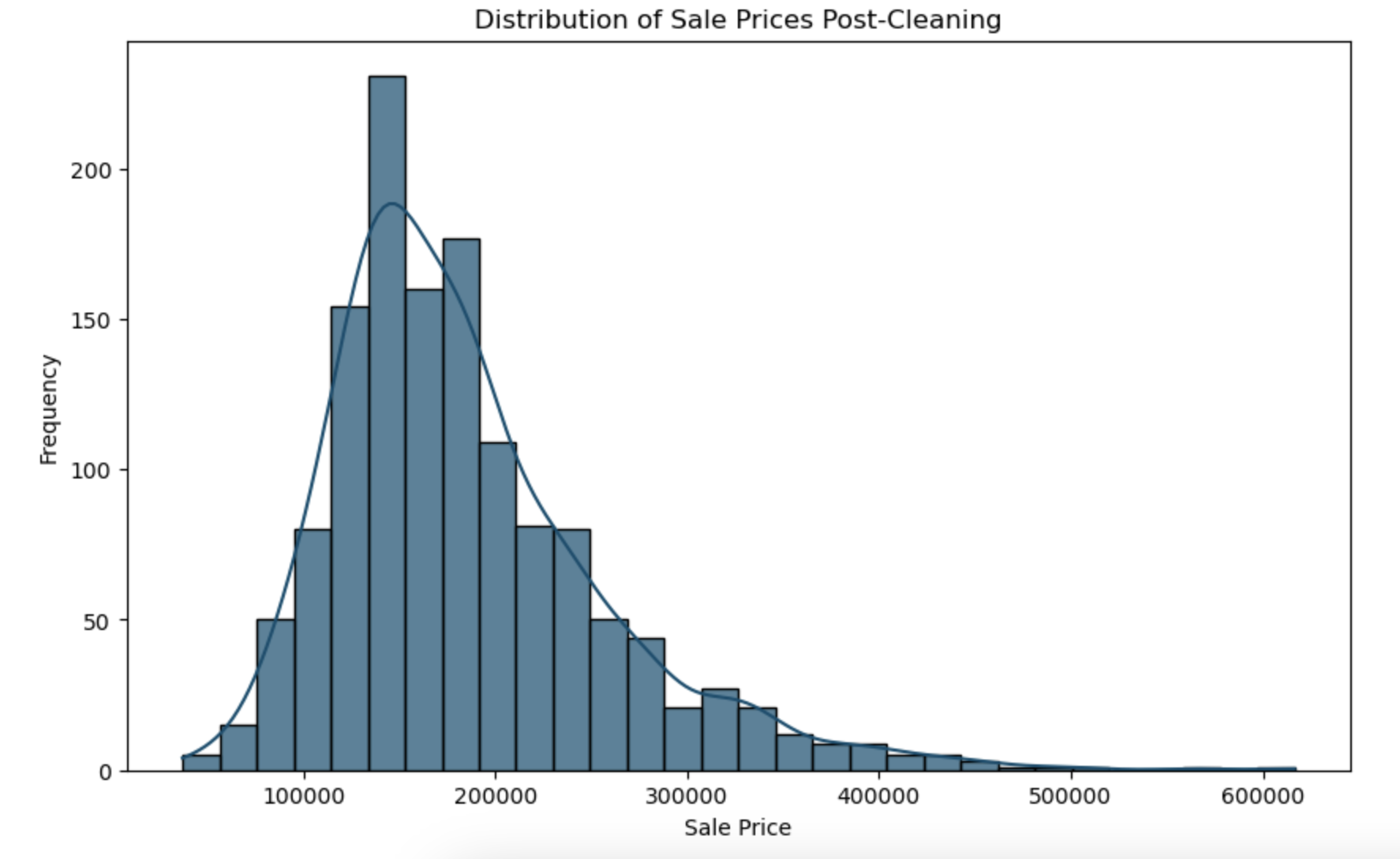
**Data Structure**

The House Price database contains analytics and descriptions of 1,460 properties sold in Ames, Iowa, from 2006 to 2010. This dataset includes 81 categorical and numerical variables, each representing unique features of sold houses, such as neighborhood, lot area, and bedrooms.

We used Python and its associated libraries to further analyze each property's variables. Our findings include a causal analysis between the Lot Area and SalesPrice. Additionally, we created regression models to see which combination of variables can produce the most accurate prediction, identifying the features that have the greatest impact on the sale price of a house.

**Data Cleaning**

To enhance the accuracy of our findings, we cleaned the data set by eliminating categorical variables missing more than 5 values and numerical variables with over 90% zeros. Additionally, we excluded the House ID since it does not reflect SalesPrice and lacks relevance for further analysis. To ensure our accuracy further, we removed outliers by implementing a function that filtered out numerical observations with a standard deviation higher than 3.5. After these adjustments, our data set was reduced to 48 variables from the original 81 and 1353 rows from 1460.



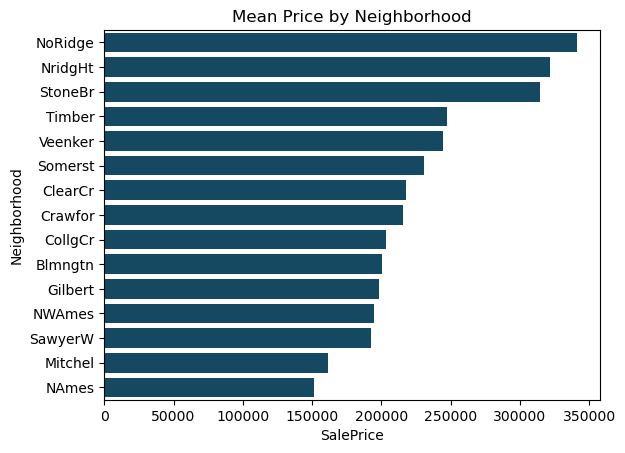
**Distribution of the SalePrice Variable**

We used the post-cleaning dataset to see the distribution of SalePrice.The SalePrice variable exhibits a mean value of $184,754.96, a median of $169,914, and a standard deviation of $72,263.40. The histogram clearly illustrates that the SalePrice variable possesses the following attributes: unimodal, asymmetric, and skewed right.

**Results of Distribution**

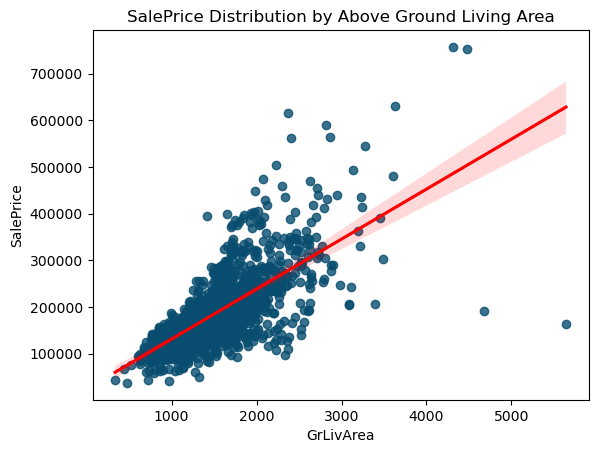
The SalesPrice distribution results

indicate most transactions fall below the $250,000 price range, as evidenced by the mean and median. When we developed our prediction model, minimizing the impact of outliers was crucial. Therefore, variables that favored high-end houses, such as PoolArea, were not ideal to include since they could skew our predictions leftwards and reduce accuracy.

**Relation of SalePrice to Other Variables** 

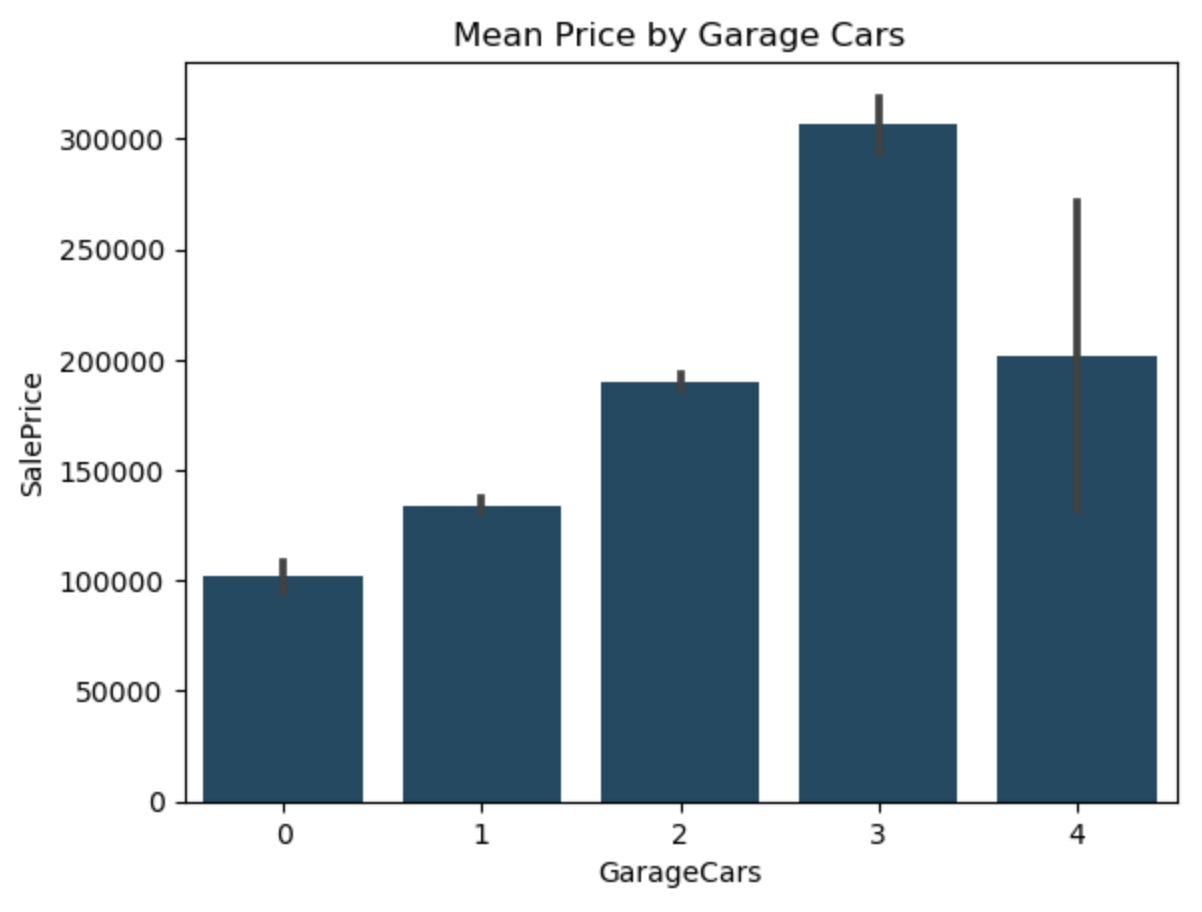
Neighborhood to SalesPrice

The Neighborhood variable is the best predictor for sales price, with an adjusted R-squared of 0.558. The average sale prices of homes can differ greatly based on the neighborhood. Looking at the regression, the neighborhoods with the lowest p-values to SalePrice are NoRidge, NridgHt, and StoneBr, which coincidentally are the neighborhoods with the highest average SalePrices, which means that Neighborhood is a good predictor for houses with higher SalePrices.



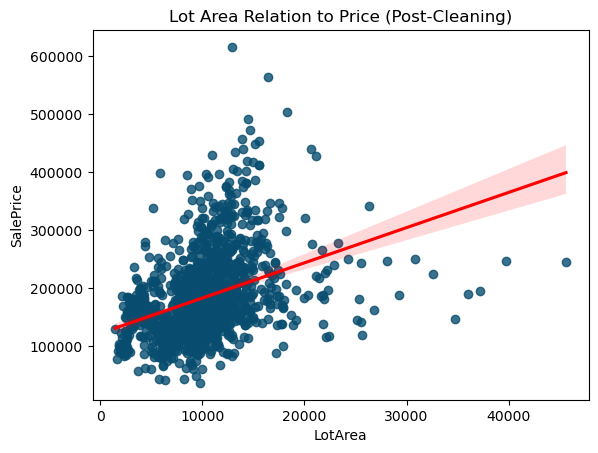
Ground Living Area to SalesPrice

The scatter plot illustrates that SalesPrice has a strong, positive correlation with the variable GrLivArea. From our observation, homes with a larger ground living area tend to have a higher sales price. The variable GrLivArea has an adjusted R-squared of 0.556.

Garage Cars to SalesPrice

The bar chart shows the relation between the SalesPrice variable and the quantity of GarageCars. Homes with 3-car garages have a significantly higher mean, which is a strong indicator of a high sales price. Every category of GarageCars is also statistically significant, all having a P-value greater or equal to 0.01. However, 4 car garages have a higher range of SalePrices. The overall Adjusted R-squared between GarageCars and SalesPrice is 0.445.

**Causal Analysis**



**Confounding Variable Process**

Our initial method to build a house-price prediction model was based on the assumption that larger Lot Areas correlate with higher prices. To verify whether LotArea directly correlates with changes in SalePrice, we first identified variables that may be confounding LotArea. Variables are considered confounding if the variable Z is statistically significant to both LotArea and SalePrice. To detect confounding variables, we compiled a list of all the potential confounders in house prices post-cleaning. We then ran a function that tested each confounder’s relationship with SalePrice and LotArea. If a numerical variable had a P-value below 0.05, it was added to our list of tested confounders. Similarly, for categorical variables, if at least one of the dummies had a P-value less than 0.05, we added the variable to the list of confounding variables.

**Checking for Omitted Variable Bias**

To check for Omitted Variable Bias (OVB), we made a function that added each confounding variable to a regression model with LotArea. We then assessed the impact of each variable on the coefficient of LotArea by adding it to the model. OVB occurs when a relevant variable is left out of a regression model, leading to biased and unreliable coefficients for the included variables. A variable is considered unaffected by OVB if there are minimal changes to the LotArea coefficient after it’s included in the regression. We created a function to filter out variables that altered the LotArea coefficient by more than ten percent. The resulting filtered data frame contained all the confounding variables without OVB.

**Regression Model with Controls**

After identifying confounding variables, we ran a regression model where these variables were controlled. Our model contained 29 control variables and produced an adjusted r-squared of 0.639, roughly a 500% increase from the LotArea univariate r-squared of 0.129. However, we recognize that our prediction model could be improved. Currently, our multivariate model only captures 63.9% of the variance in the sale price of a house, which is insufficient given the number of control variables required. This makes our model prone to overfitting and limits its applicability to the data that lacks the necessary control variables.

**Predictions**

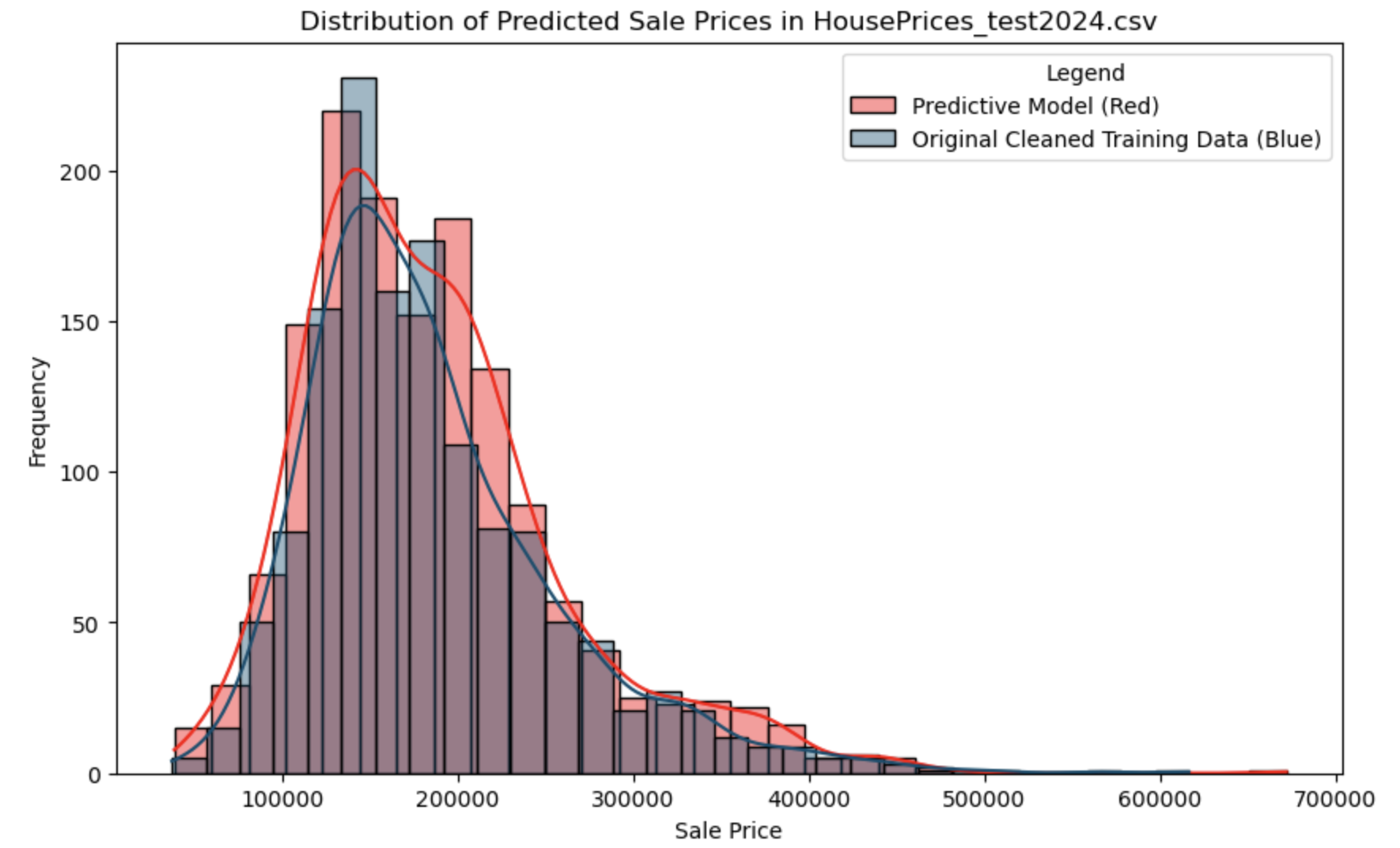
After pinpointing weaknesses in our multivariate causal model, we looked to improve accuracy using other variables in our cleaned data as LotArea was a weaker predictor. We created linear regression models for each variable to test their relationship with SalePrice, then ranked them by adjusted r-squared values to identify the strongest predictors in a new data frame.

We analyzed summary statistics on the new data frame to identify the upper quartile of adjusted r-squared values, discarding variables below this cutoff. This process selected 11 variables with the highest predictive power for SalePrice, which we used in our forward and backward regression models.

**Multivariate Regression Models**

To create a model that accurately predicts SalesPrice, we used the list of 11 variables and the stepwise regression framework. The forward regression model started with an empty model. We added variables until the adjusted R-squared values no longer increased, resulting in a model with 10 variables and an adjusted R-squared value of 0.872. Our backward regression model started with 11 variables, removing those with the highest p-values until all of our variables had a P-value below .05. This approach led us to eliminate the variables Foundation, GarageCars, and GarageArea, generating an Adjusted R-Squared of .872.

We determined that the backward regression model is optimal because it produces the same high Adjusted R-squared of .872 as the forward model but with 3 fewer variables. This makes it equally accurate but less likely to overfit, making it more adaptable. Therefore, we will be using this model to test with *HousePrices\_test2024.csv*.



**Predicting *HousePrices\_test2024.csv***

Using our backward model, we predicted house SalesPrice from *HousePrices\_test2024.csv*. The model’s line of best fit closely aligns with the original data’s distribution. Our model overestimates the number of low to medium-priced houses, likely because the backward model can't reintroduce variables it previously removed, omitting useful predictors for estimating higher-priced homes. To enhance accuracy, adding control variables specifically for higher-priced houses could improve predictions across different data sets.