



AMES HOUSING



Ames Housing Price Prediction

Using Regression Models

By: Alicia, Benjamin, Mok, Riche

Scope

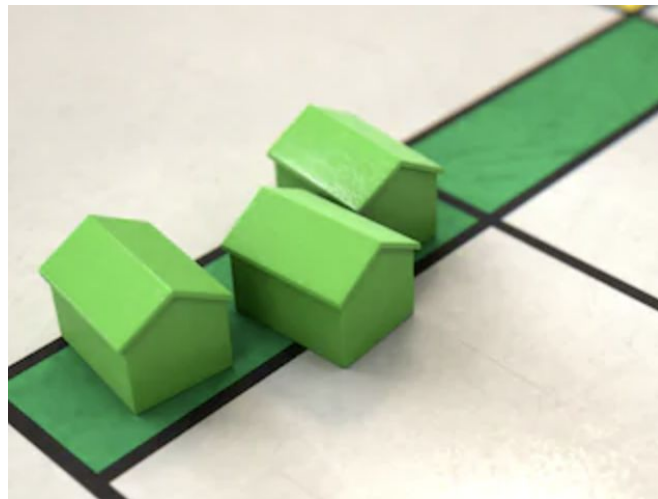
- Introduction
- Data Cleaning & Preprocessing
- EDA & Feature Engineering
- Target Engineering
- Regression Model
- Top Features
- Recommendations
- Limitations
- Conclusion

Introduction

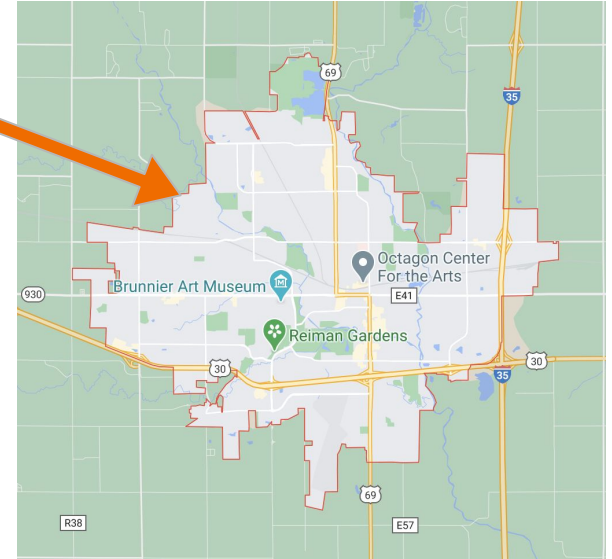
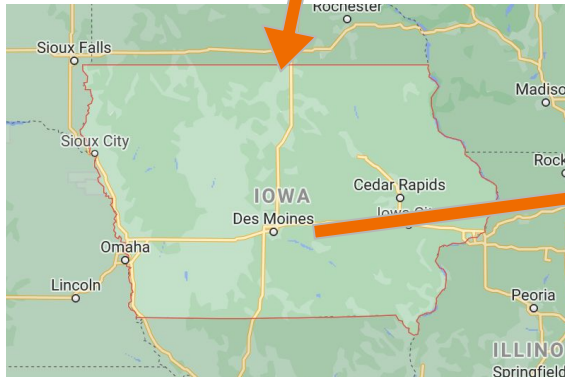
Introduction: Problem Statement

As a team of data analysts, we have been tasked by a property agency to create a **linear regression model** based on the **Ames Housing Dataset** that will predict the rough price of a house at sale.

The agency has requested that the final production model be **easy to interpret** and make use of **no more than 20 features**.



Introduction: Ames, Iowa



Introduction: Ames, Iowa



Introduction: Ames Dataset

The Ames dataset is a dataset built by **Dean De Cock**, a statistics professor from the Truman State University, who was looking for an alternative to the well-used Boston housing dataset

Some features of the dataset:

- 2051 rows with 81 columns
- Some error correction was done
- Outliers/null values were not removed



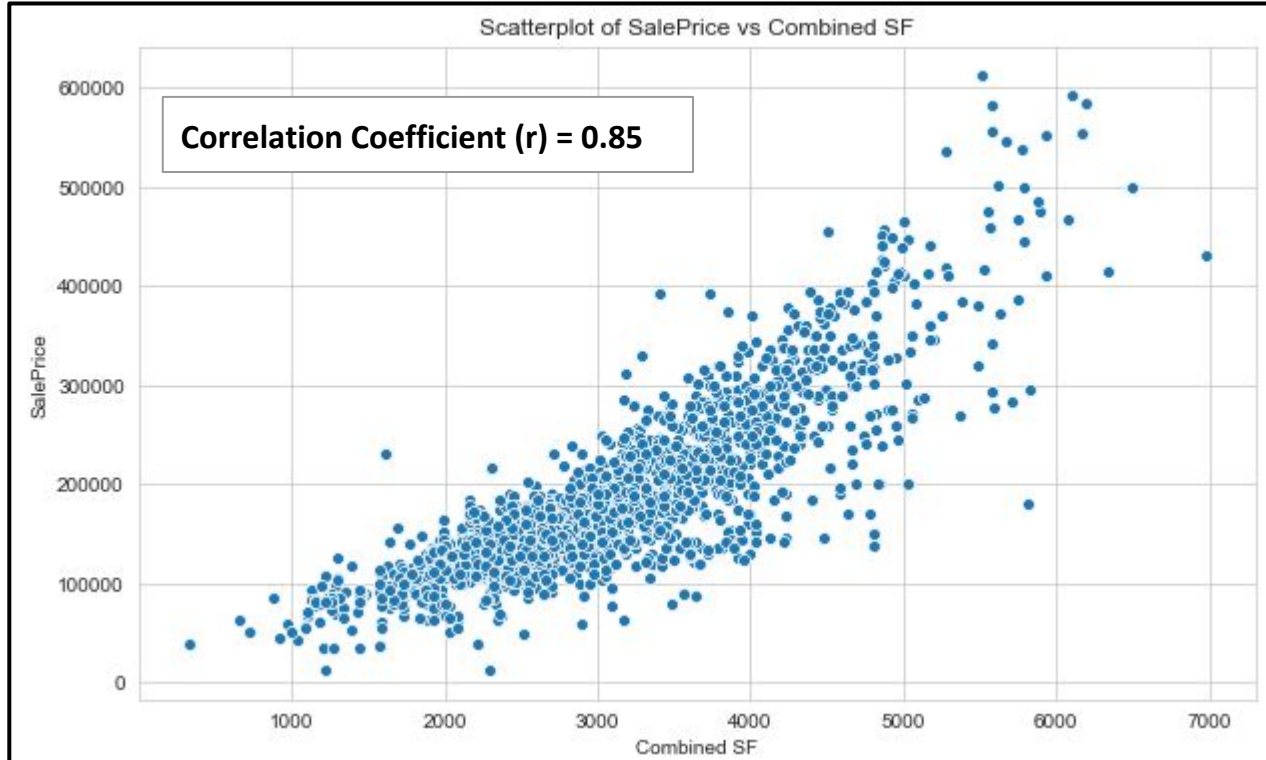
Data Cleaning & Preprocessing

Data Cleaning & Preprocessing

Data Cleaning	<u>Missing Values</u> <ul style="list-style-type: none">26/82 features contain null values - Replaced with 0 / stringMissing value area but presence with feature - replace with mean of feature <table><tr><td>poolqc</td><td>0.995612</td></tr><tr><td>miscfeature</td><td>0.968308</td></tr><tr><td>alley</td><td>0.931741</td></tr><tr><td>fence</td><td>0.804973</td></tr></table>	poolqc	0.995612	miscfeature	0.968308	alley	0.931741	fence	0.804973	<u>Eliminating Outliers</u> <ul style="list-style-type: none">Logic checkExtreme values <table><tr><th></th><th>yrsold</th><th>yearbuilt</th><th>garageyrblt</th></tr><tr><td>1699</td><td>2007</td><td>2006</td><td>2207.0</td></tr><tr><td>1885</td><td>2007</td><td>2008</td><td>2008.0</td></tr></table>		yrsold	yearbuilt	garageyrblt	1699	2007	2006	2207.0	1885	2007	2008	2008.0	<u>One Hot Encoding</u> <ul style="list-style-type: none">Getting dummies for all the Nominal featuresRated with a scale -> ordinal features were mapped with a range
	poolqc	0.995612																					
miscfeature	0.968308																						
alley	0.931741																						
fence	0.804973																						
	yrsold	yearbuilt	garageyrblt																				
1699	2007	2006	2207.0																				
1885	2007	2008	2008.0																				
Pre-processing	<u>Multicollinearity</u> <ul style="list-style-type: none">Features like Garage Cars and Garage Area <table><tr><th></th><th>garagecars</th><th>garagearea</th></tr><tr><td>0</td><td>2.0</td><td>475.0</td></tr><tr><td>1</td><td>2.0</td><td>559.0</td></tr><tr><td>2</td><td>1.0</td><td>246.0</td></tr><tr><td>3</td><td>2.0</td><td>400.0</td></tr><tr><td>4</td><td>2.0</td><td>484.0</td></tr></table>		garagecars	garagearea	0	2.0	475.0	1	2.0	559.0	2	1.0	246.0	3	2.0	400.0	4	2.0	484.0	<u>New Features</u> <ul style="list-style-type: none">Age of buildingCombining of features [total sq ft and total baths]<ul style="list-style-type: none">Total baths = Full bath + (0.5 * Half Bath)	<u>Polynomial Features</u> <ul style="list-style-type: none">Overall QualityOverall Sq Ft		
	garagecars	garagearea																					
0	2.0	475.0																					
1	2.0	559.0																					
2	1.0	246.0																					
3	2.0	400.0																					
4	2.0	484.0																					

EDA & Feature Engineering

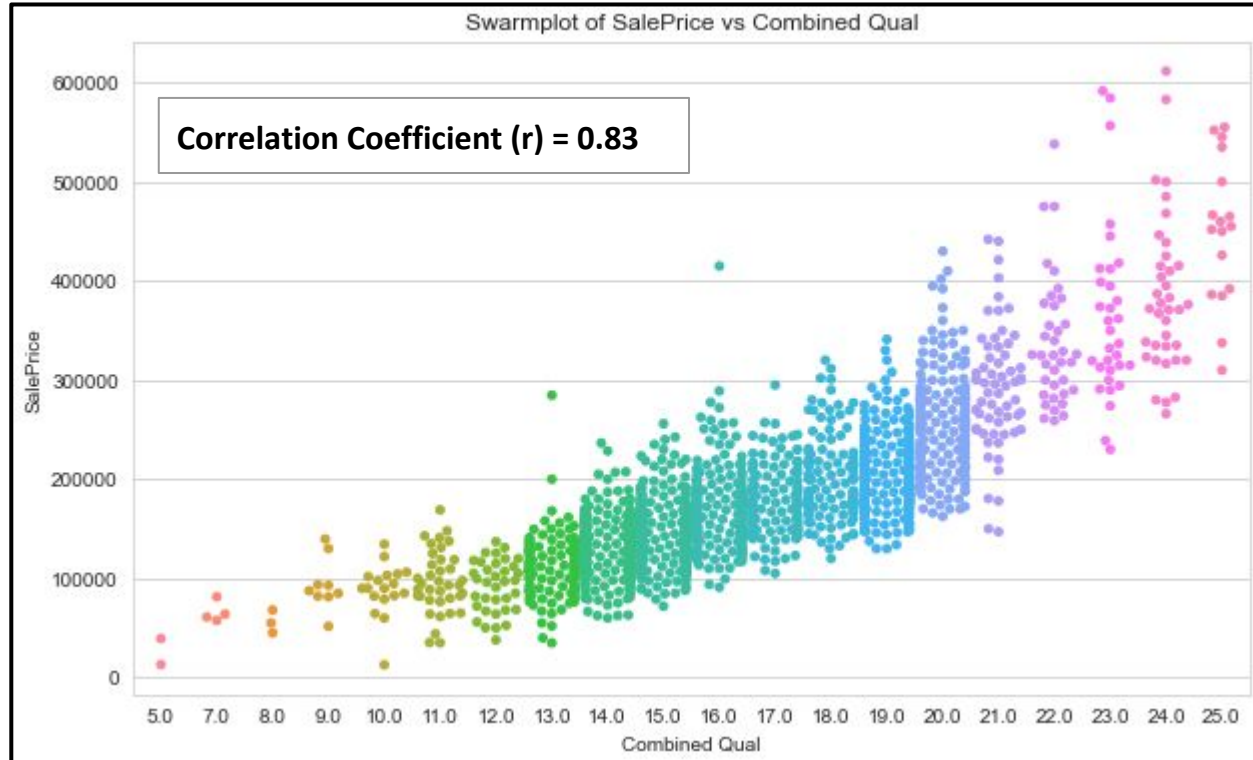
EDA & Feature Engineering: Total Area of House



Observations:

- **Strong positive relationship** between housing sale price and total area of the house
- Good indication of a linear relationship

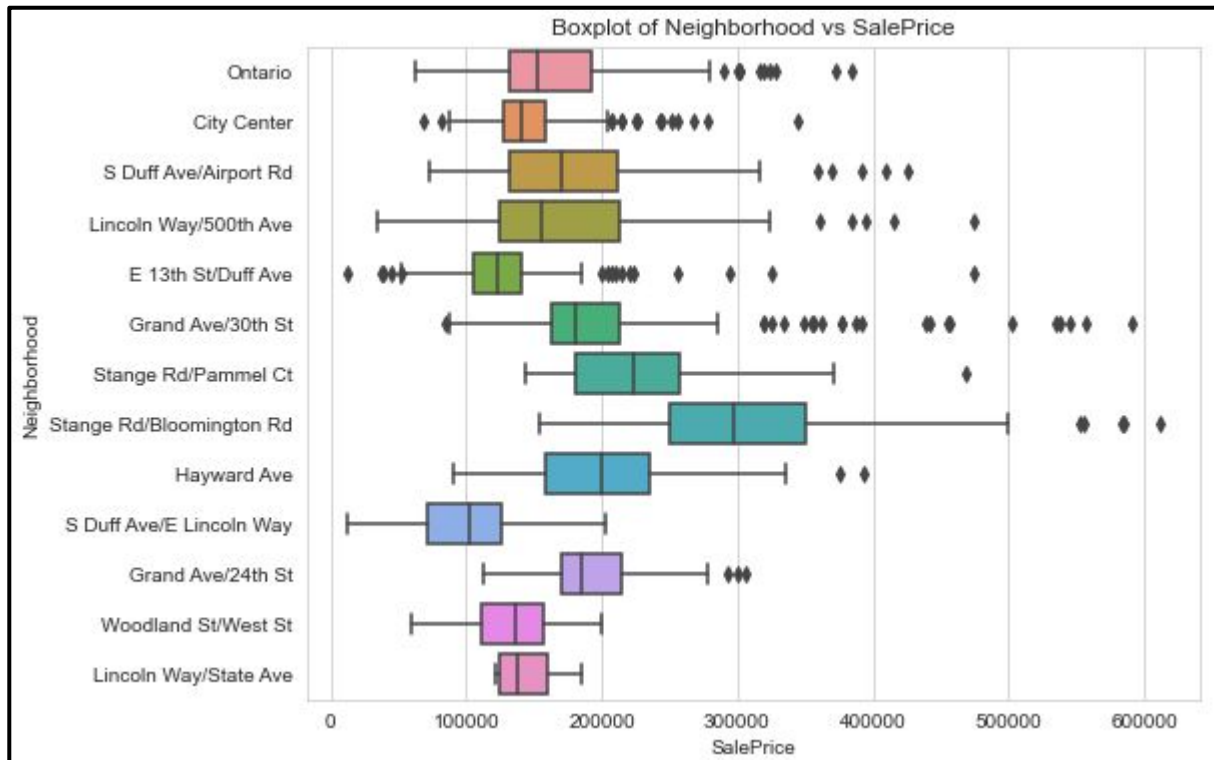
EDA & Feature Engineering: Overall Quality



Observations:

- **Strong positive relationship** between housing sale price and overall quality
- Clear linear relationship
- Largest number of houses within the overall quality range of 13-20

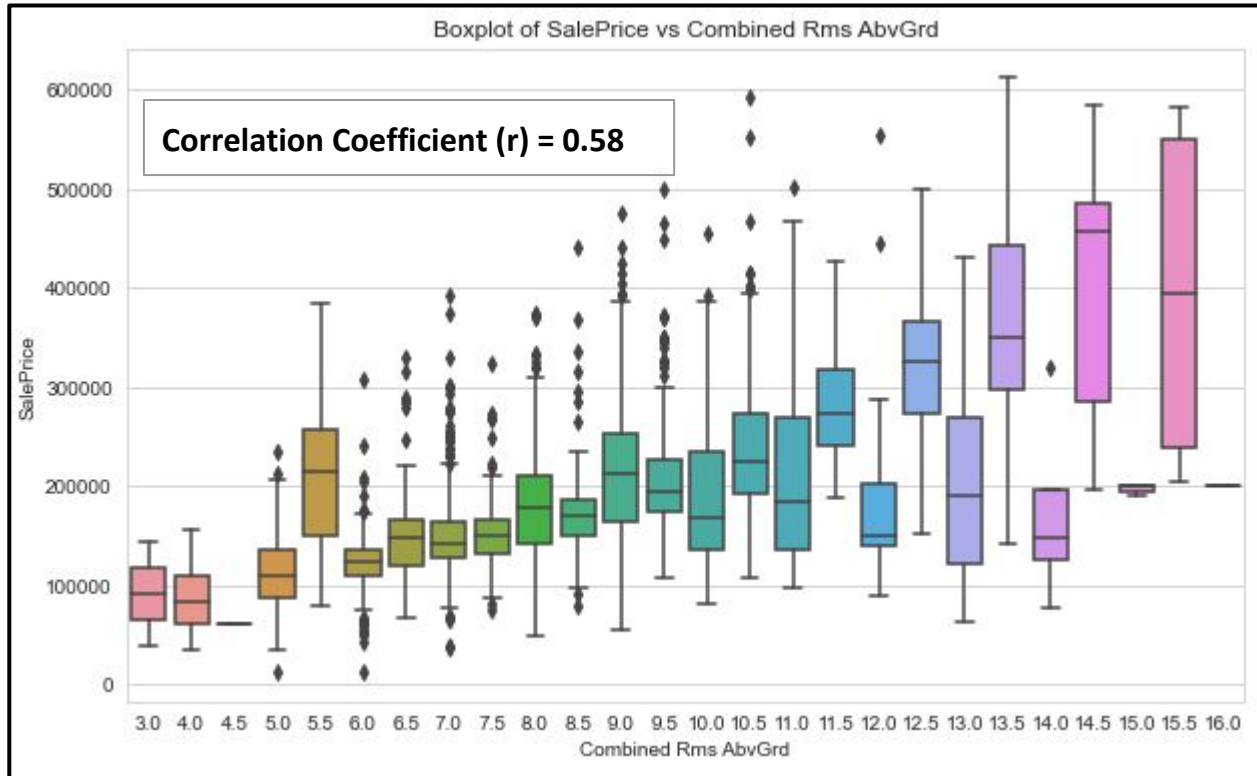
EDA & Feature Engineering: Neighborhood Regions



Observations:

- **Good variability** in the data after grouping different neighborhoods into popular Ames regions
- **Highest** median of sale price within the **Stange Rd/Bloomington Rd** region
- **Lowest** median of sale price within the **S Duff Ave/E Lincoln Way** region

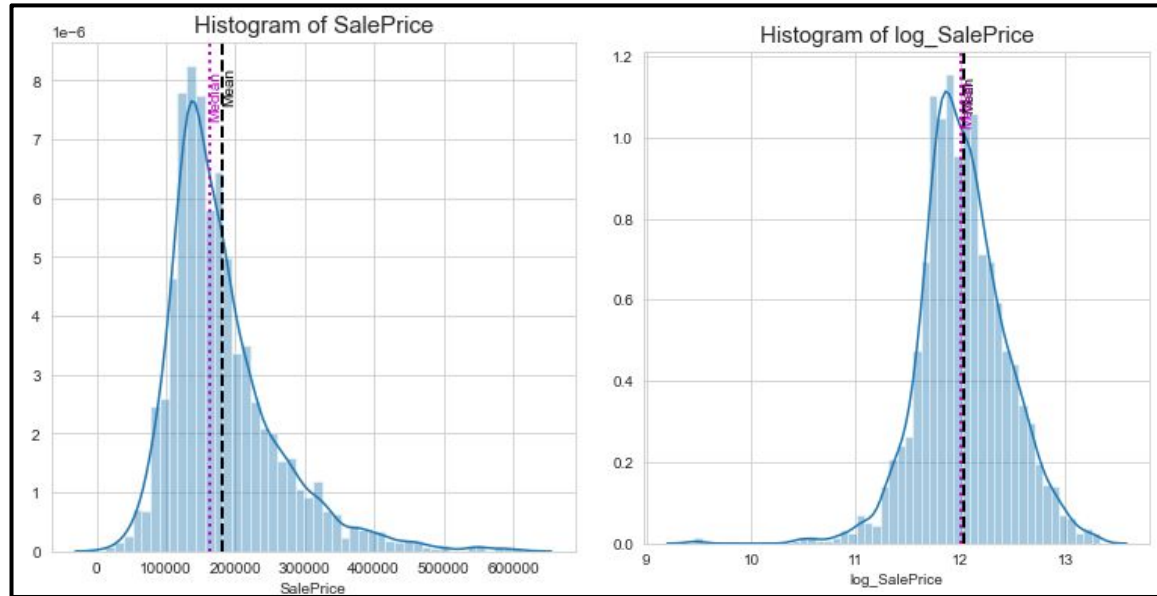
EDA & Feature Engineering: Number of Rooms



Observations:

- **Strong positive relationship** between housing sale price and total number of rooms
- Rooms include bedrooms and bathrooms
- Variability in data

Target Engineering: Log Transformation



→
Improvement in normality

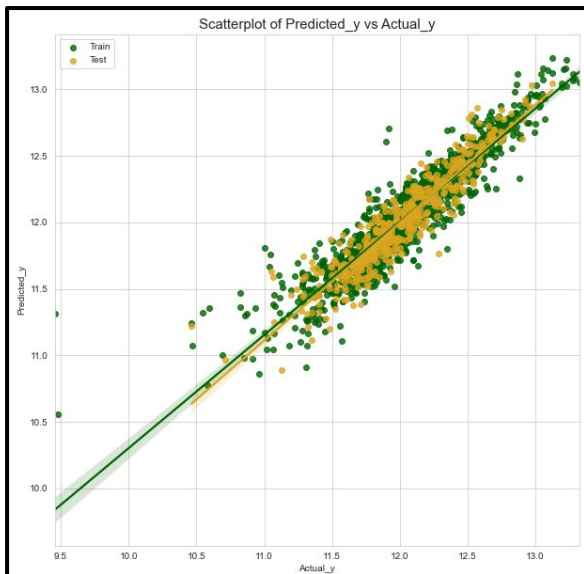
Observations:

- The distribution of target variable is **right(positive)-skewed**
- Great improvement in normality after doing a **log transformation** on the target variable
- Correcting for the **violation in normality assumption** helps to improve predictions
- Achieve a more homoscedastic model

Regression Model & Top Features

Regression Model

	model	r2_train	r2_cv_estimate	adj_r2_train	adj_r2_cv_estimate	rmse_train	rmse_cv_estimate
0	Linear Regression	0.863296	0.850527	0.861491	0.848553	0.155803	0.162230
1	Ridge Regression	0.863268	0.850658	0.861463	0.848686	0.155819	0.162156
2	Lasso Regression	0.862256	0.851643	0.860438	0.849684	0.156394	0.161574
3	ElasticNet Regression	0.862257	0.851638	0.860439	0.849679	0.156394	0.161577



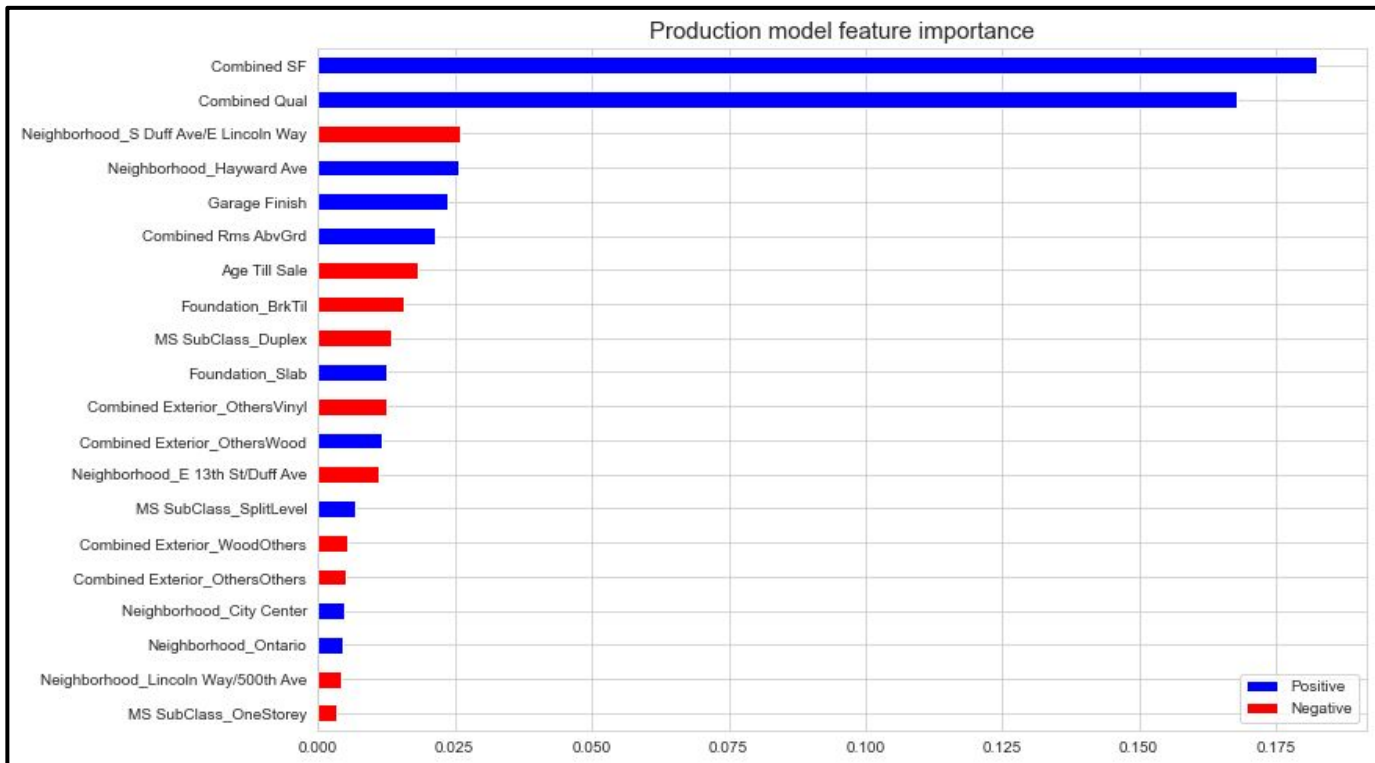
HIGHEST R^2

LOWEST
RMSE

Lasso Regression Model

- **Strong linear relationship** between the predicted and actual sale price extracted from the train data
- Minimal difference between the best-fit lines of the train data and test(hold-out) data

Top Features



Discussion and Conclusion

Recommendations

1. The more, the better!

- Add additional amenities (e.g. garage, fireplace, pool, masonry)
- Add more bedrooms, add more bathrooms

2. Quality is key

- Ensure the quality of condition of amenities

3. Location, location, location!

- Houses in highly liveable social spaces (e.g. near parks, recreation, facilities) fetch higher prices
- Less favourable environments include places near roads, petrol kiosks

Limitations

- **Dataset**

- No data on demographics
- No data on external factors such as government and external events (e.g. pandemic, disasters, subprime mortgage crisis from 2007-2010) and how it influenced decision making
- Ill defined outliers

- **Model**

- Not generalizable to other states or countries
- May not be applicable to current year
- Limited to Linear Regression
- No of features limited, might be better to have more features to select

Conclusion

Among all the variables within the dataset, **variables measuring quality condition, number of amenities or age** are better predictors.

- Higher quality rating increases price
- Additional amenities (e.g. basement/garage) increases price
- As the building ages, the price gradually decreases



Questions?



Did you know: The **mascot of Monopoly** is **Rich Uncle Pennybags**, Uncle Pennybags' full name was given as **Milburn Pennybags**. In large parts of the world he is known, additionally or exclusively, as the **Monopoly Man**, or **Mr. Monopoly**.

