

# AMES HOUSING PRICE CALCULATOR

Group 3:  
Shi Min, Wee Cheng, Rahayu,  
Matt, Courtney



# ROLES



**WE ARE DATA ANALYSTS**



**OUR AUDIENCE CONSISTS OF  
HOUSING AGENTS IN AMES**

# AGENDA

01 WHAT PROBLEM ARE WE SOLVING

02 OBSERVATIONS OF THE AMES HOUSING MARKET

03 METHODOLOGY

04 FINAL PRICING MODEL AND WHY IT IS GOOD

05 CONCLUSION AND RECOMMENDATIONS

# WHAT IS THE PROBLEM



Housing agents  
face many  
enquiries on  
housing prices

**Option 1: Respond fast with the  
mean housing price in Ames**

✗ highly inaccurate

How about Option 3, where agents can  
use a front-end calculator tool that can  
estimate the housing price from a few  
key features?

**Option 2: Dig into database and try to find the  
most similar latest transaction**

✗ takes too long

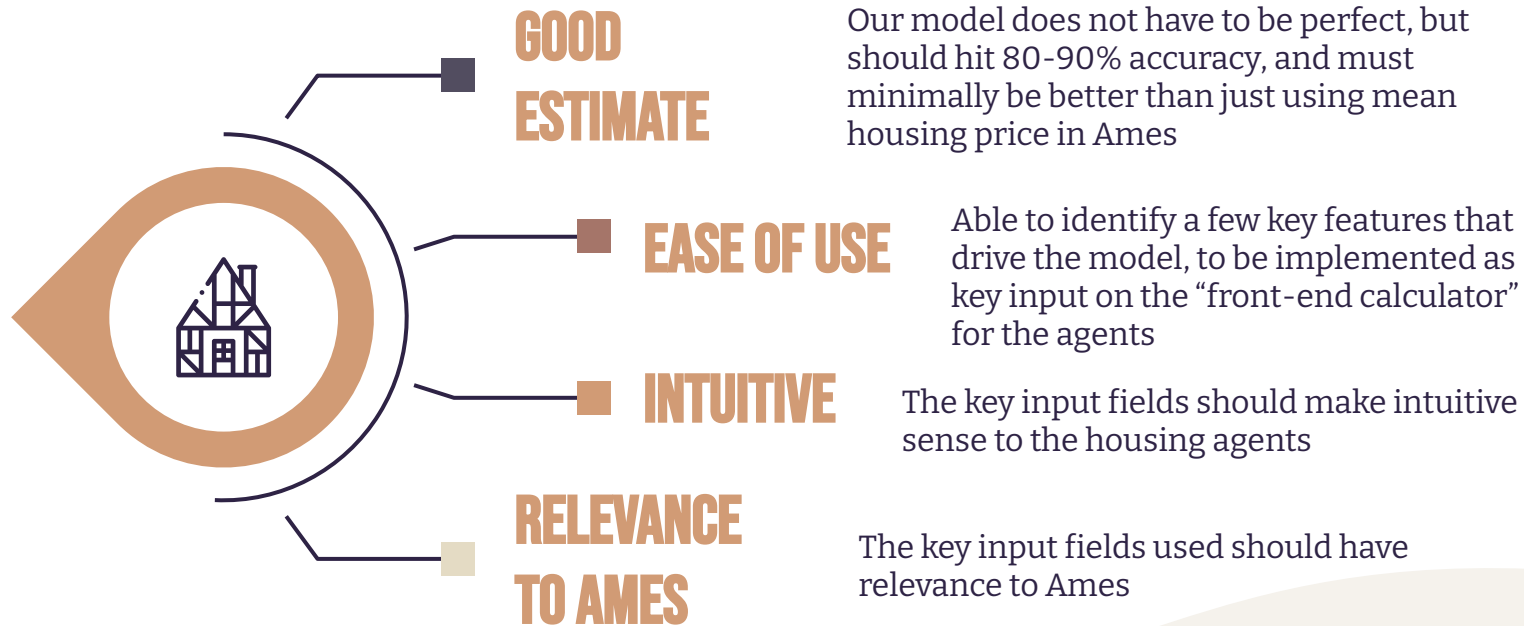
✗ might not be able to find exact match: agents are not  
sure how to adjust for the differences to better advise  
clients



# Our Objective

Help housing agents to better advise potential home buyers and sellers, by developing a tool to estimate housing prices in Ames more quickly and accurately based on a few key features of the house

# WHAT WE ARE AIMING FOR



# FUN FACTS ON AMES

**-10°C/29°C**

Lowest/Highest Temp

**25,538**

Households

**2 CARS**

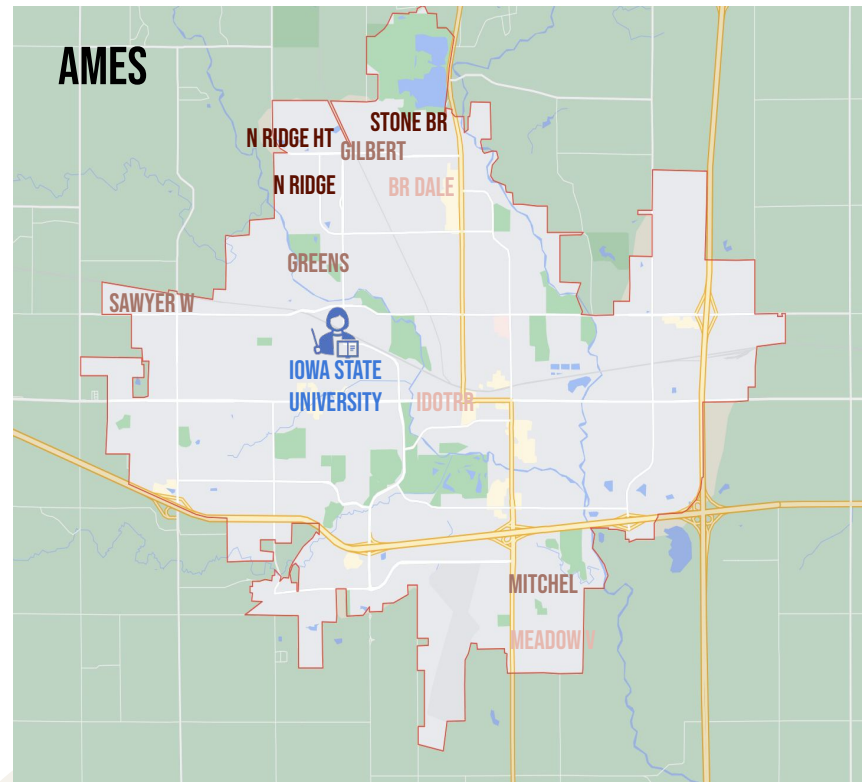
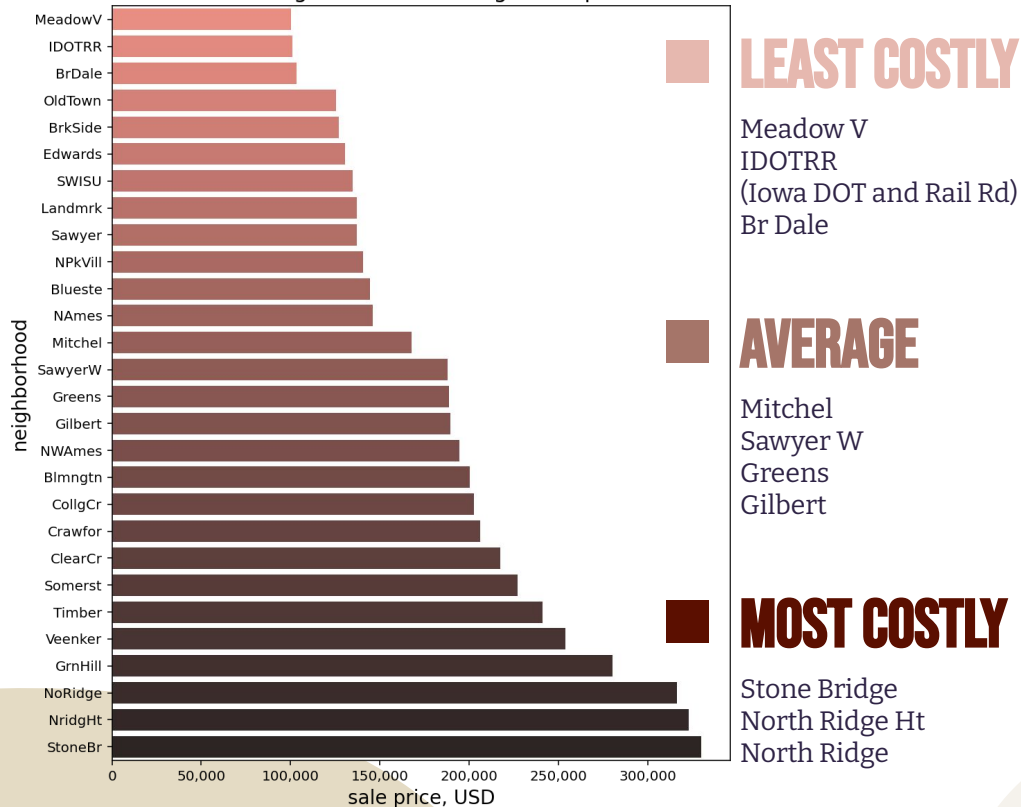
Average number of cars per  
household



# AMES HOUSING NEIGHBORHOODS

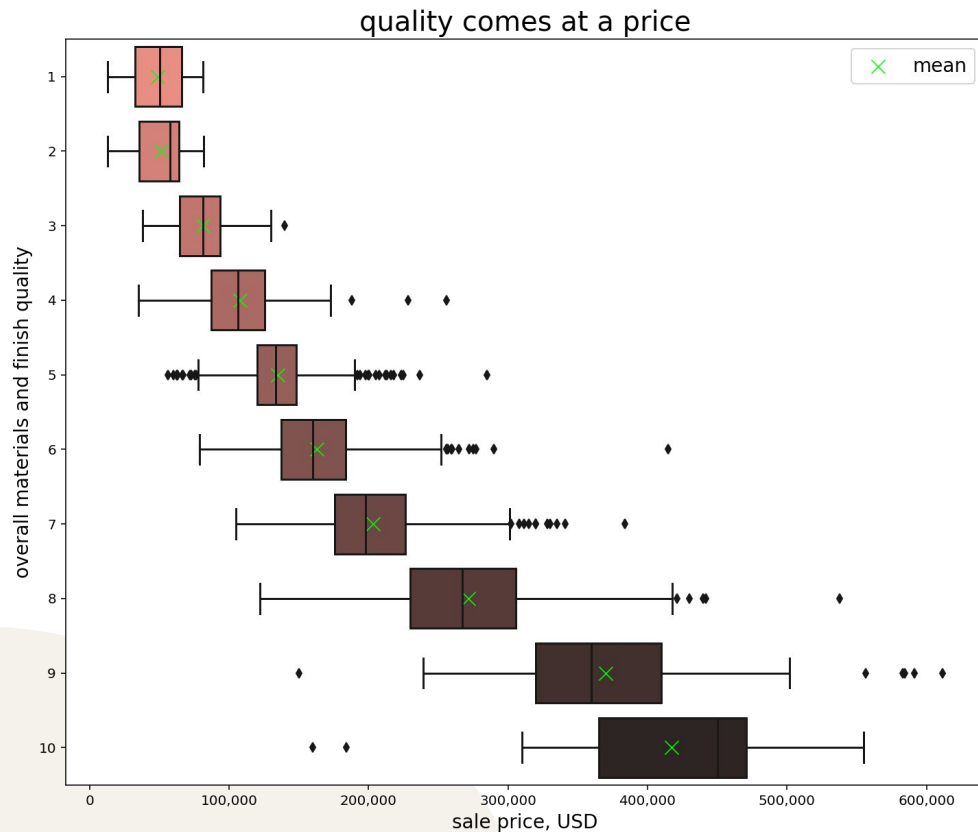


Ames neighborhood housing mean prices





# QUALITY COMES AT A PRICE



## LOW QUALITY (1-2)

~50k\$ for a house if all you need is just a shelter over your head

## AVERAGE QUALITY (5-6)

~150k\$ for an average house with average quality materials and finishing

## HIGH QUALITY (9-10)

~380k\$ for a premium house with the highest quality materials and finishing

# BIGGER HOUSE, HIGHER PRICE



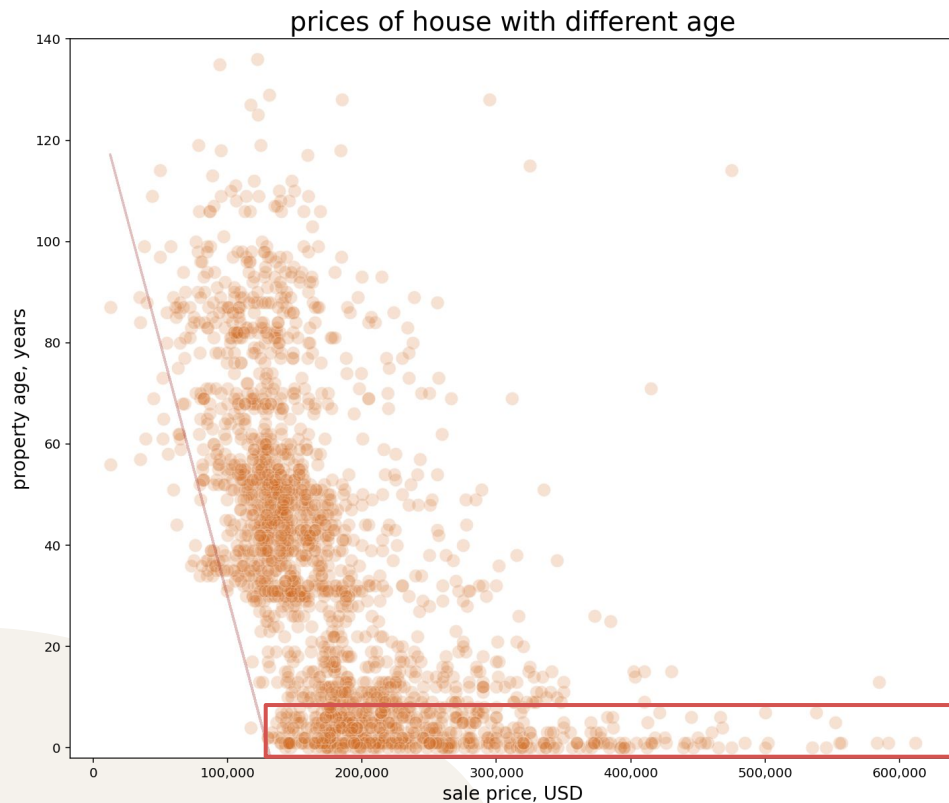
## ■ SIZE DOES MATTER

Big house naturally commands high price  
Size has a premium

## ■ MARKET FLOOR PRICE MECHANISM

There is a minimum price you can expect for a  
given house size

# HOUSE PRICES DO NOT AGE WELL



## OLD IS NOT GOLD

House prices decay with age  
While they are cheaper, older houses  
tend to be more costly to maintain

## NEW BUILD PREMIUM

You pay top dollar for something  
new and unused

# NOT ALL BASEMENTS ARE EQUAL

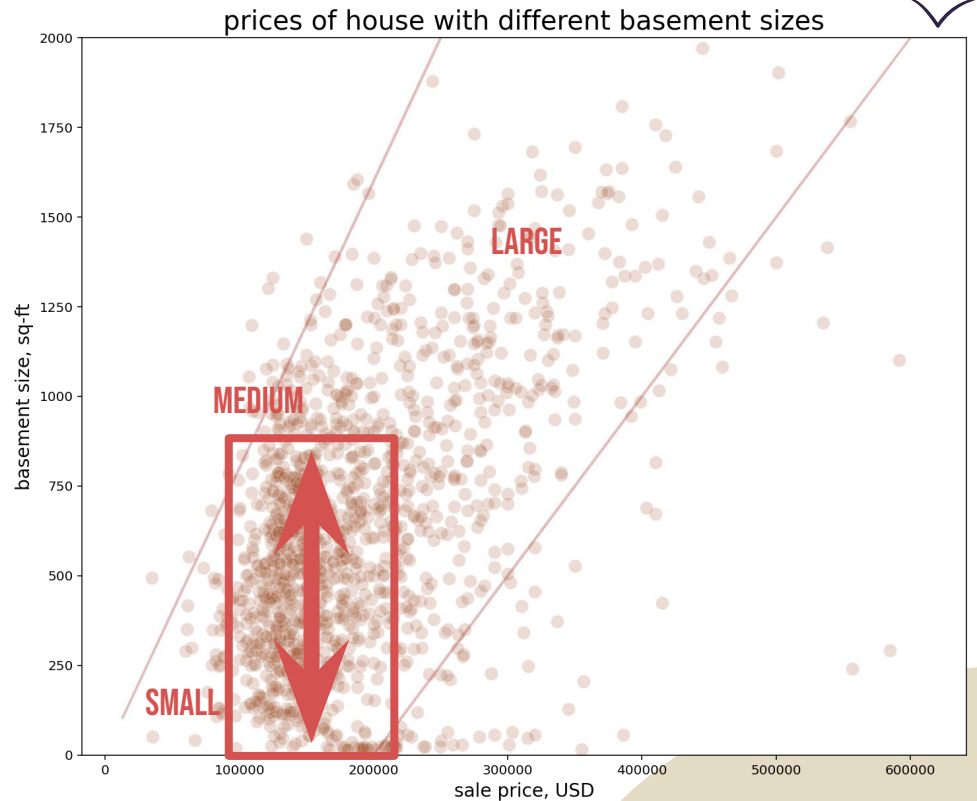
SOME ARE MORE EQUAL THAN OTHERS

## ■ SMALL AND MEDIUM

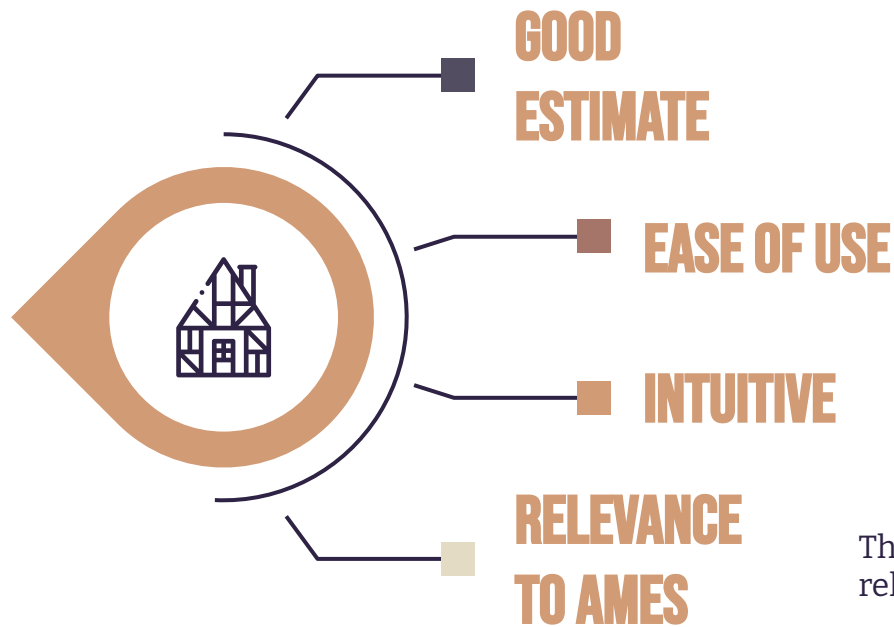
Houses with small and medium basements seems to have similar prices

## ■ LARGE MOVES THE NEEDLE

Large basements command a premium over medium and small



# RECAP OF OUR AIM



Our model does not have to be perfect, but should hit 80-90% accuracy, and must minimally be better than just using mean housing price in Ames

Able to identify a few key features that drive the model, to be implemented as key input on the “front-end calculator” for the agents

The key input fields should make intuitive sense to the housing agents

The key input fields used should have relevance to Ames

# METHODOLOGY



**Keeping in mind the need for *ease of use* and *intuitiveness*, we have done the following steps:**

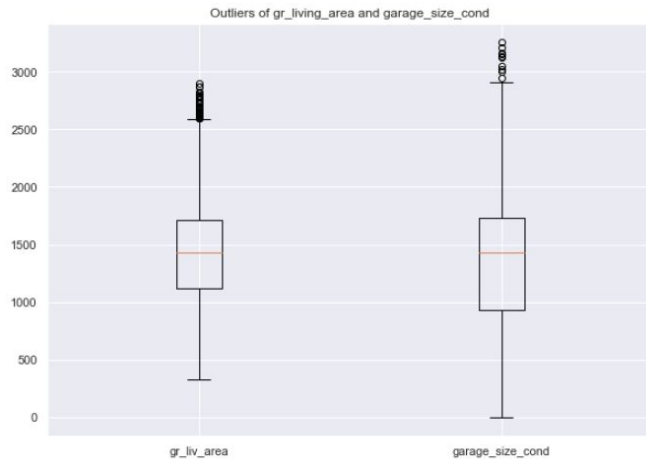
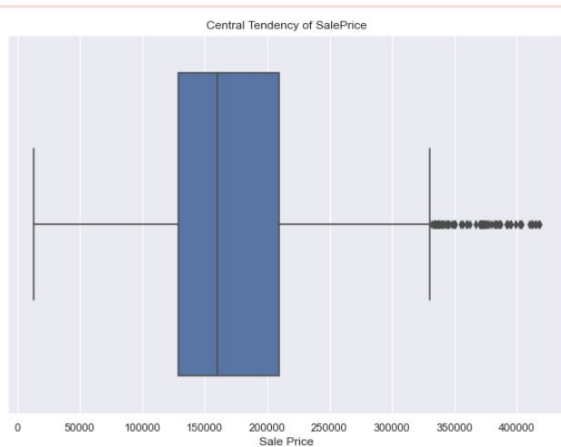
- Make sense of and group similar data fields together, out of the many many (79!!) data columns we have
- Inputs to the model should be easily interpretable, e.g. rather than input year that the house was built, we transformed the data to age of the house, so that it is immediately understandable and makes sense when used in future predictions
- Get rid of data columns that are a derivation of or closely related to other columns

# METHODOLOGY



Keeping in mind the need for a good estimate, we have done the following steps:

- Remove outliers in sale price that might distort our model and not give good estimates for the bulk of houses
- Remove outliers for the top 5 features that have strong correlation with saleprice



# METHODOLOGY




**Keeping in mind the need for features relevant to Ames, we have done the following steps:**

- Looked at the garage features in more detail and think of different combinations that house owners might be concerned about (e.g. just a big garage alone will not be attractive, house owners want garages that are both big and have good quality).
- This is because Ames' households have 2 cars each on average, hence garage will be important to them



# FINDINGS OF DATA

- ❖ Data provided consist of total 79 features with different data types
- ❖ Columns belongs to a similar feature group
- ❖ There are features with more than 50% null values!
- ❖ Missed out useful and interpretable features in the data such as property age, remodel age



```
garage_type  
garage_yr_blt  
garage_finish  
garage_cars  
garage_area  
garage_qual  
garage_cond
```

# DATA CLEAN UP INITIAL STEPS



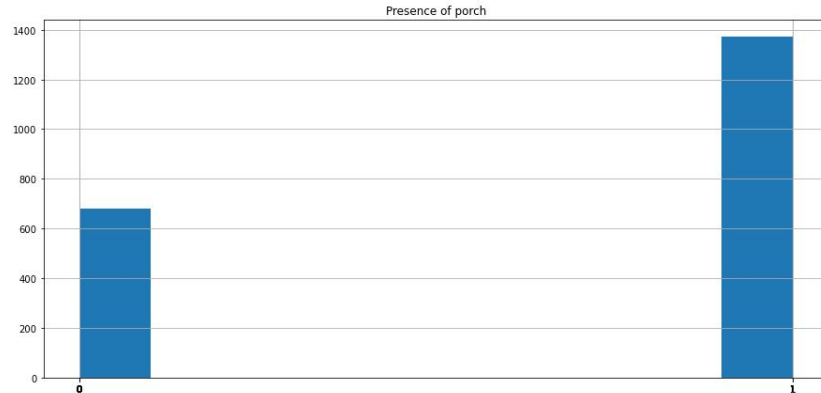
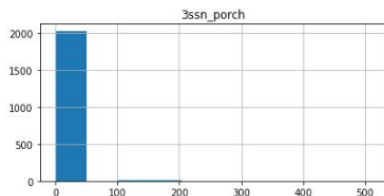
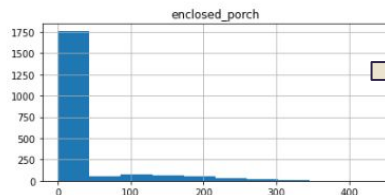
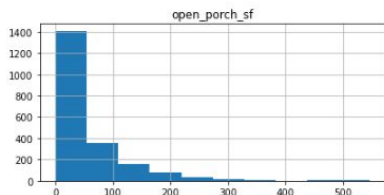
## 1. Check all the null values in the data

- a. Those columns that is integer and float type, assume it means there is no such feature and replace to 0
- b. Those that are in object type, check in data dictionary
  - i. If there is an existing “None” value for that column ,assume null means does not have that feature and replace to None
  - ii. If there “None” value doesn’t exist , check the number of null values. If more than 5% remove, else replace with mode.

# DIGGING DEEPER...

## Porch

- There are 3 porch features. Most values are 0 sq. Instead of having 3 porch features, changing it to 1 porch feature to indicate whether there is a porch or not. Categorize them to indicate if they have a porch instead



# DIGGING DEEPER...



## Garage Year built & Year Built

- Generally, garage built the same year of the property. True enough, 76% of the garage built in the data is equivalent to year built
- Remove garage year built as one of the feature

## Garage Area and Garage Condition

- Buyers usually look at the size and garage condition. Putting them together as one feature

## Remodel age and Property age

- To better interpret the data:  
Calculate property age and remodel age and include as additional features

# DIGGING DEEPER...



## 1st floor square foot, 2nd floor square foot, Ground Living Area

- There are houses that exist where there is not even a second floor
- Those houses that do not have a 2nd floor, 1st floor and ground area living area will likely be the same values
- Buyers generally look at total size of house above ground
  - 58% with houses have 0% square feet 2nd floor
  - 57.3% 1st floor = ground living area
- Remove 1st floor as a feature and change values in 2nd floor to indicate where there is it is a single storey houses or not

# DIGGING DEEPER...

## High collinearity of features

A few features that have very high collinearity( > 80%)

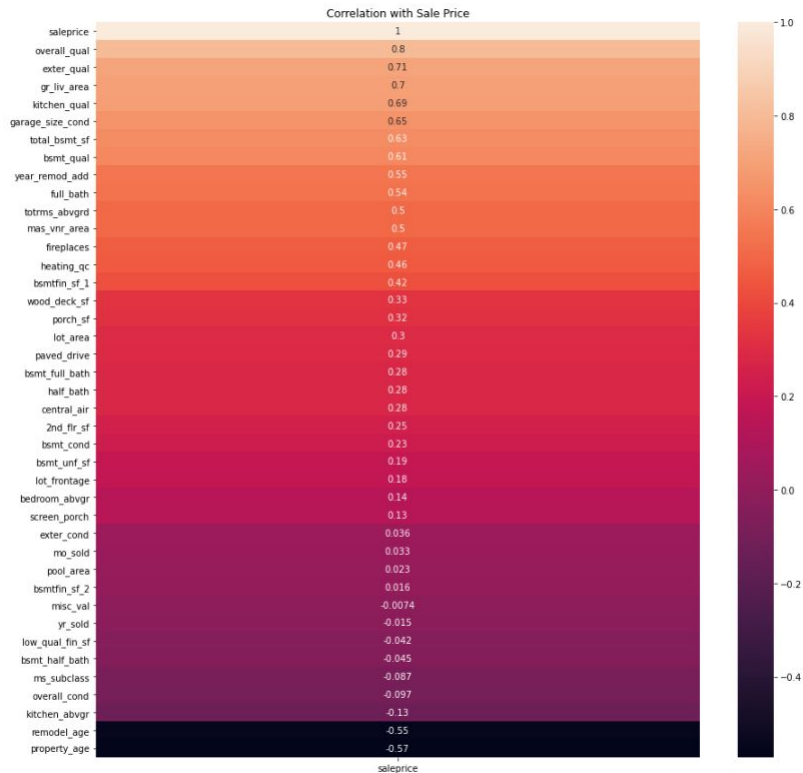
- Garage condition vs garage quality
- Garage cars vs garage area
- Pool quality vs pool area
- Remove one of the feature in each.

## Remodel year and Year built

- 65% collinearity
- But more than 50% where remodel year = year built
- Remove year built as a feature



# CORRELATION OF FEATURES WITH SALE PRICE



**Top 5 features positive correlated:** overall quality, external quality, ground living area, kitchen quality, garage size condition

**Feature that is most negatively correlated: Property age!**



# CURRENT MODEL

## Mean Price Model

Uses the average price of houses in the market to predict the price of a house



# ISSUES WITH CURRENT MODEL



## INACCURATE

High Root Mean Square Error (US\$67,981.3)



## DOES NOT PREDICT

All results are based on just the mean price of houses



## IGNORES HOUSE FEATURES

No features of the house is considered

# OUR NEW MODEL



## Supervised Learning

Machine Learning Algorithm trained on a set of current data

## Features Considered

Takes into account of features in the house

## Uses Regression Models

Creates Regression Models for a more accurate prediction

# BENEFITS OF OUR MODEL



## **Accurate and Precise**

Our model can predict House prices with accuracy and precision

## **Scalable to New Data**

Our model can use even new real-world data to expand its learning

## **Adaptable**

Our model is able to adapt to features

## **Takes Features into Account**

Features are always considered



# OUR FINAL PRICING MODEL: LASSO REGRESSION

# WHY LASSO REGRESSION



## Highest Cross Validation Score 91.7%

With a score of 91.7%, Lasso model scored the highest of the 3 models we tested

## Highest Degree of Accuracy

Lasso Model scored the highest R2 Score at 91.0% on our test score

## Lowest Root Mean Square Error (RMSE)

Low RMSE means sales price predictions are more precise

# COMPARING BETWEEN OTHER REGRESSION MODELS

Evaluation Metrics	Linear	Lasso	Ridge
Cross Value Mean	-3.88	0.917	0.913
Train R-square	0.942	0.936	0.942
Test R-square	-8.39	0.910	0.900
RMSE	16187.615	20787.274	21916.231

# COMPARING THE OLD AND THE NEW



## **More Precise**

Our model has an RMSE of US\$20,787 vs US\$67,981 on the old model

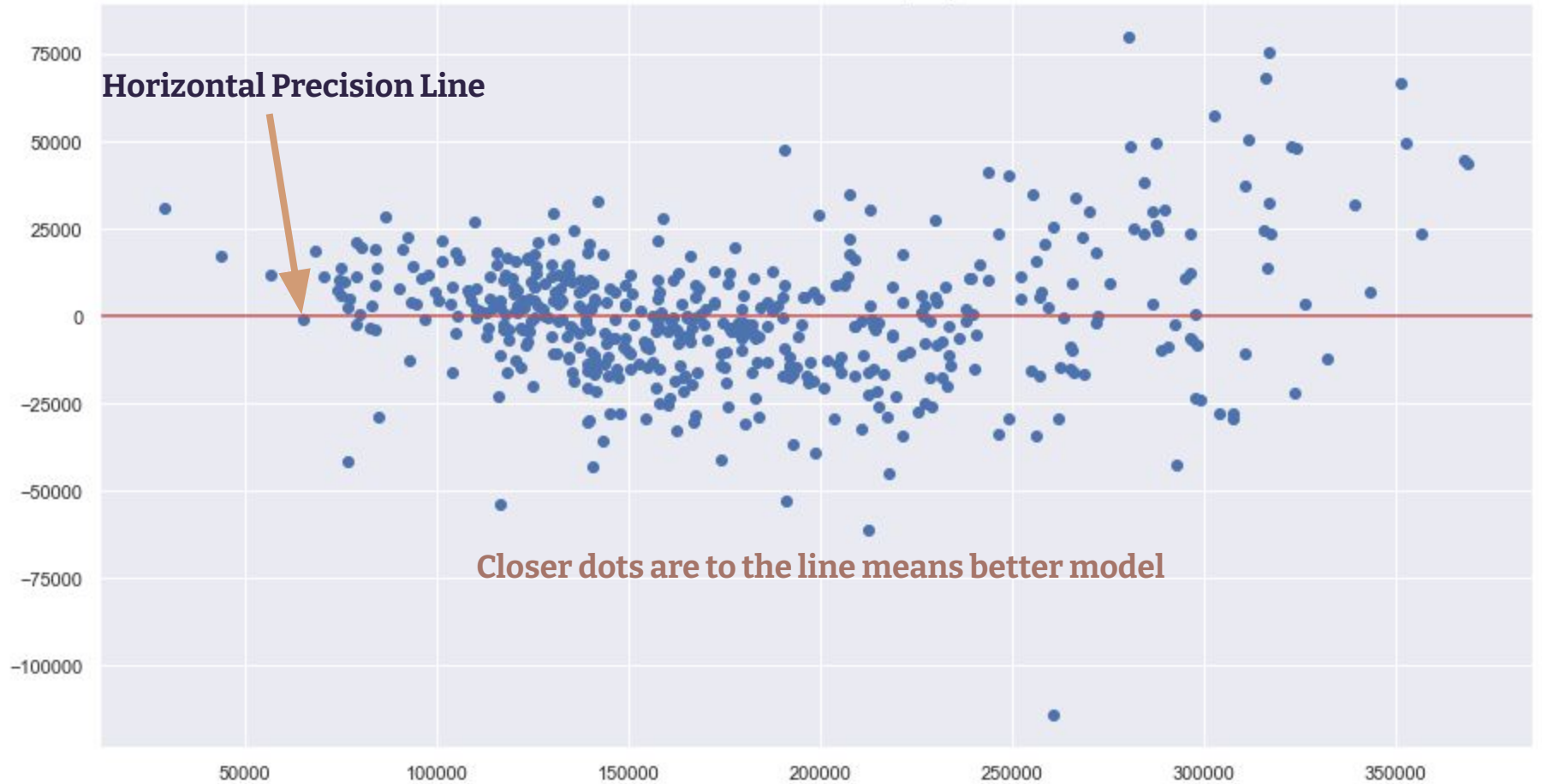
## **Takes features from housing data into account**

Our model accepts features of the house, the old model does not

## **Able to scale with current and new data**

Our model can scale with more data, the old model cannot

Residuals for best model(lasso)





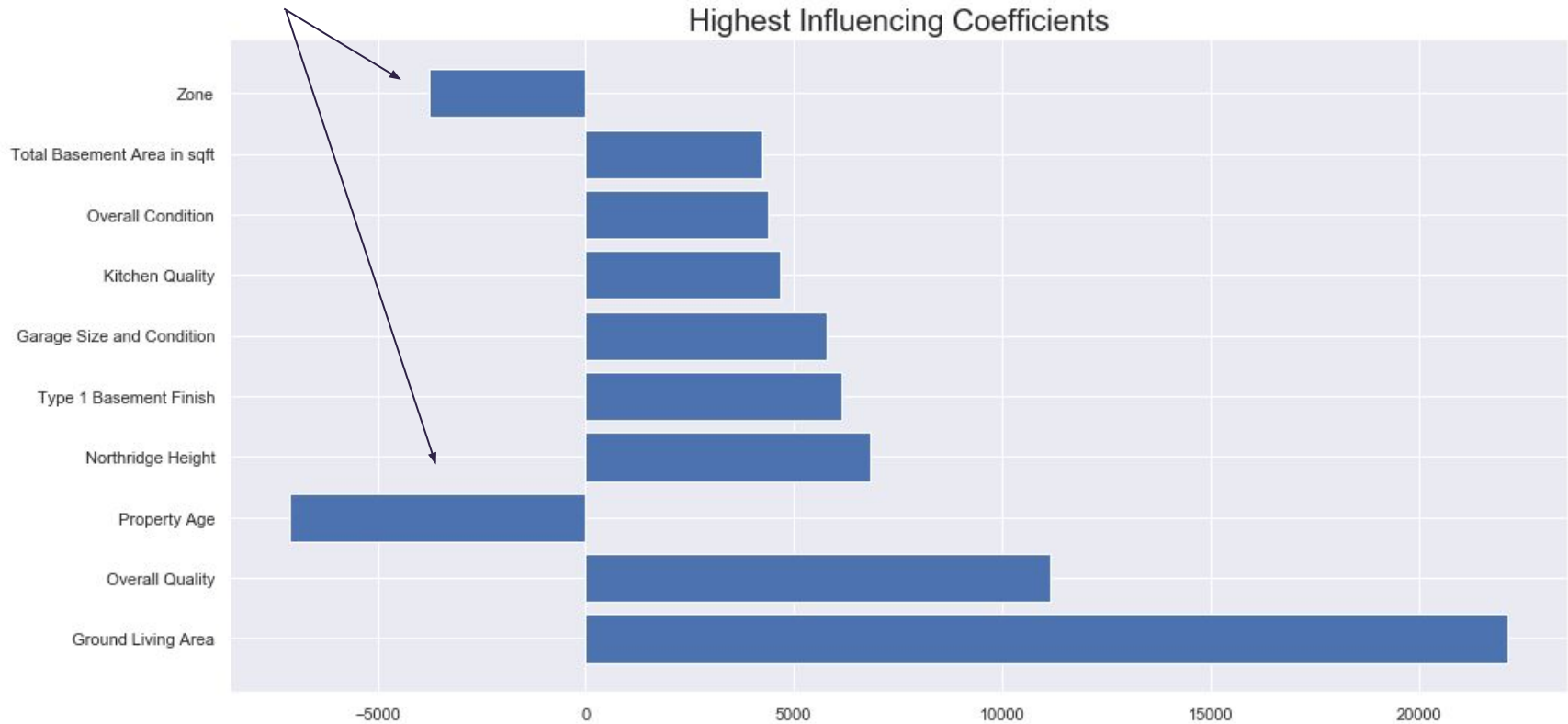
# INTERESTING FINDINGS FROM OUR MODEL



## Biggest influencing factors to House Sale Prices are:

- Ground Living Area
- Overall quality
- Property Age
- Proximity to Northridge area
- Type 1 finish basement
- Garage size and condition
- Kitchen quality
- Overall condition
- Total basement area
- House Zone

These are negatives



# MODEL VALIDATION FROM KNOWN FACTS



Northridge is one of the more expensive areas

Ames is also an Agricultural Area

Property Price depreciate with Age

Size Does Matter!

Verified with our data as well

# CONCLUSION

- Our model is 91% accurate in estimating housing prices
- 5 features identified as top sale price predictors
- **Pros:** quick, simple, more accurate
- **Cons:** oversimplified, subjective rating of feature quality

Features	
1	Size of living area (above ground)
2	Overall quality
3	Property age
4	Neighborhood
5	Rating of basement finished area

# RECOMMENDATION

Create a tool in the form of a mobile app that accepts top 5 features as inputs, that housing agents can access anytime and anywhere to estimate sale prices

- Advise home sellers on realistic listing prices
- Advise home buyers on reasonable prices to pay

# BACK-UP



# TRANSFORMATION FROM STANDARD SCALE (BACK-UP)



$$z = \frac{x - \mu}{\sigma}$$

eqn:

$$y = az_1 + bz_2 + k$$

$$y = a \left( \frac{x_1 - \mu_1}{\sigma_1} \right) + b \left( \frac{x_2 - \mu_2}{\sigma_2} \right) + k$$

$$y = \frac{a}{\sigma_1} x_1 - \frac{a\mu_1}{\sigma_1} + \frac{b}{\sigma_2} x_2 - \frac{b\mu_2}{\sigma_2} + k$$

$$y = \left( \frac{a}{\sigma_1} \right) x_1 + \left( \frac{b}{\sigma_2} \right) x_2 + \left( k - \frac{a\mu_1}{\sigma_1} - \frac{b\mu_2}{\sigma_2} \right)$$

Model coefficient interpretation:

‘a’ is the coefficient from regression post standard scaling

To get the coefficient based on original scale, we divide it by the sigma of the train dataset which was used to fit the standardscaler  
i.e. ‘a / sigma’

# MODEL COEFFICIENT ILLUSTRATION (BACK-UP)

	REFERENCE:			
	HOUSE A	HOUSE B	HOUSE C	HOUSE D
Overall quality	6	7	6	6
Above grade living area (sq-ft)	2000	2000	2100	2000
Property age	50	50	50	90
Delta value vs. A	-	+\$9,702	+\$4,813	-\$3,305
Sale price	\$250,000	\$259,702	\$254,813	\$253,305