
Phase 2 Project

By: Tommy Quan-Duc Phung

Overview

Client: Zillow Home Group Inc.

Source: King County House Sales



Parameters Used:

- **Dependent** - Price
- **Independent** - Square Feet of Living, Grade, and House Age
 - Extra categorical variables were used

Business Problem



- What makes an expensive house?
- Is the house estimate accurate?
- Could we predict the price of a house?

Objective

Zillow Estimation Tool: Zestimate

- Company's Algorithm
- Median Error Rate - **2.4% to 7.49%**

Objective:

Create a model to incorporate with or replace currently placed algorithm in hopes to improve accuracy for seller and buyers.

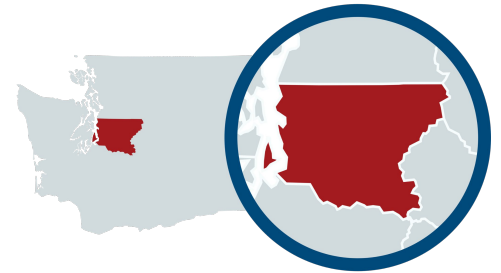


Data

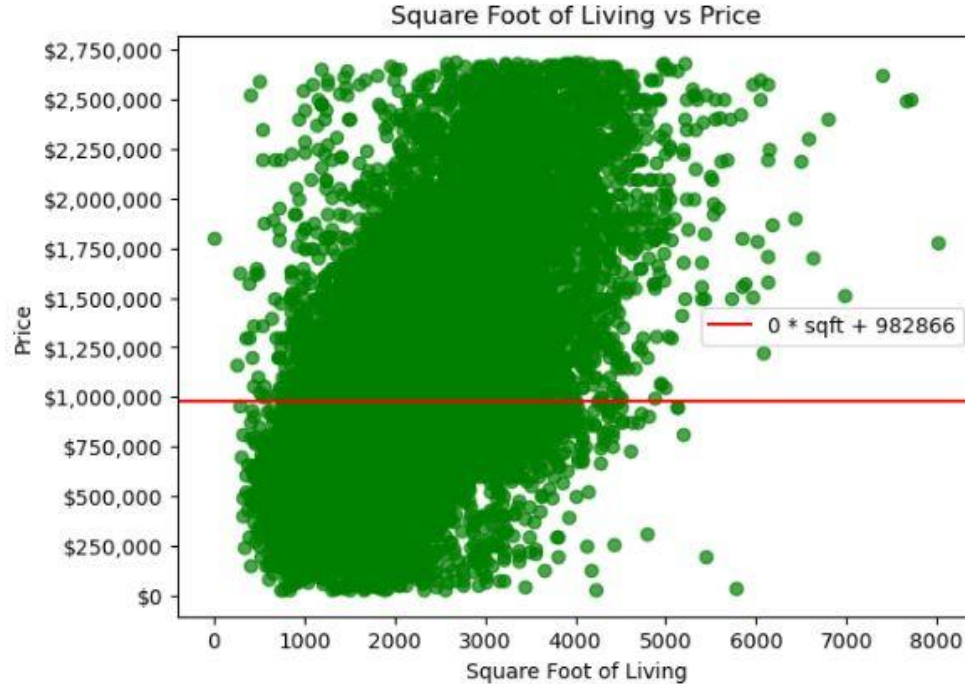
- **30154** samples (3%)
- **25** columns or attributes
- **Numerical** and **Categorical** data types



Analysis should only be applied to houses
in King County, Washington



Baseline Model - Intercept-Only Model



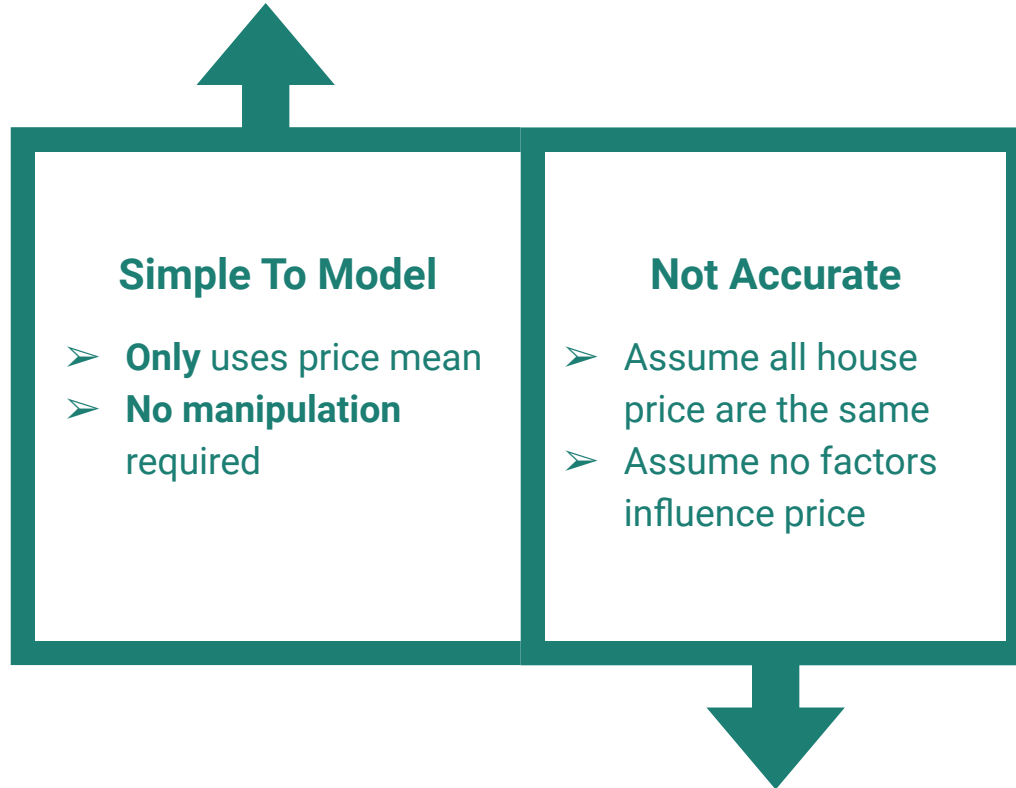
$$Y = 982,866$$

Interpretation:

All houses are priced at
\$982,866.

Fig 1. Square foot of living vs Price with mean price

Baseline Model - Limitation



Method

A series of modeling and progression plots:

1. **Highest correlation** - Single Variable
2. **Second Highest** - 2 Variables
3. ... - Multi-Variable

Added parameters:

- **Interaction Terms** - Top Two Highest
- **House Age** - Modify Year Built

Why Linear Regression

1. Explains the **relationship** between two variables.
2. Can be used to **predict prices** with given variables.

Line of Best Fit:

Price = (sqft_living cost * num sqft) + constant

Model Overview

36.2%

Model 1:
+ Square Feet of Living



- Only 1 parameter
- Relatively low house cost

41.8%

Model 2:
+ Grade



- Two parameters with a negative house cost.
- High grade influence.

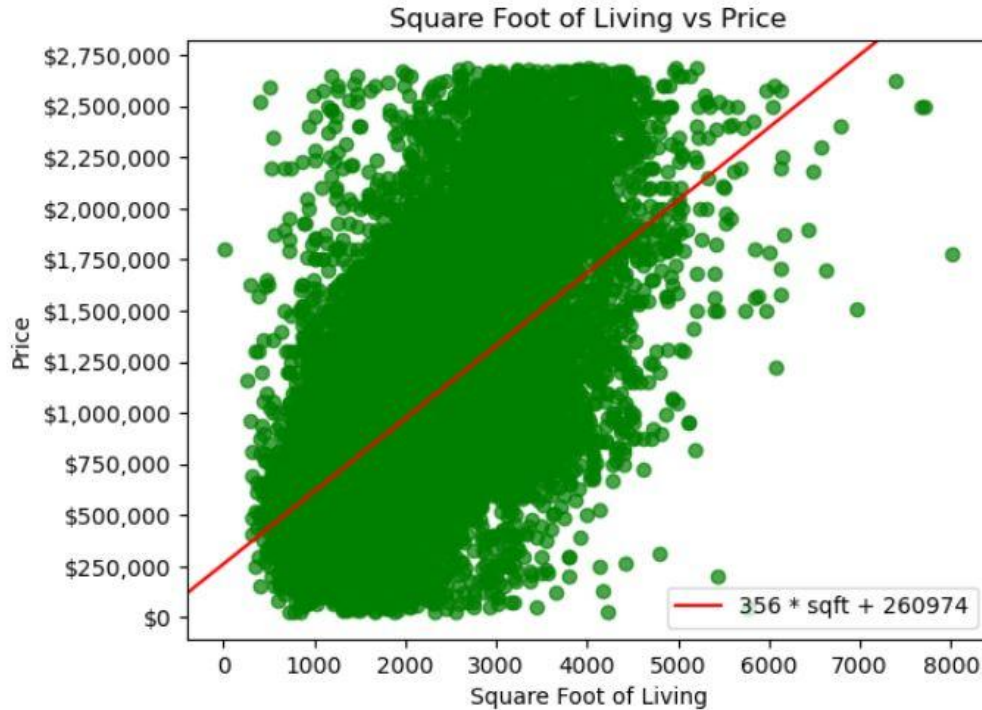
46.4%

Model 3:
+ House Age



- A larger negative starting value.
- Lower cost per square foot.
- Bigger grade influence.

Model 1



Interpretation -

➤ **\$356** per square foot

Price = \$261,000 +
(356 * square footage)

Fig 2. Square foot of living vs Price with line of best fit

Model 1 Limitations

Simple To Interpret

- **Single** Independent Variable
- **Logical** starting price

Only Explains 36.2% of Variance

- Not Considering **Other parameters**
- "A house with no living cost \$261,000"

Multiple Linear Regression

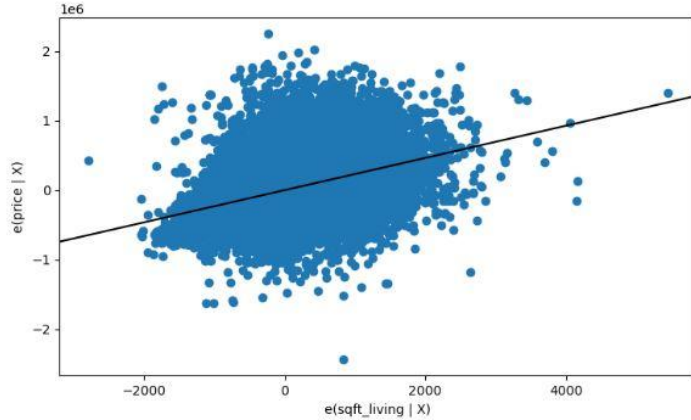
Uses:

- Multiple Variables
- Increase **complexity**

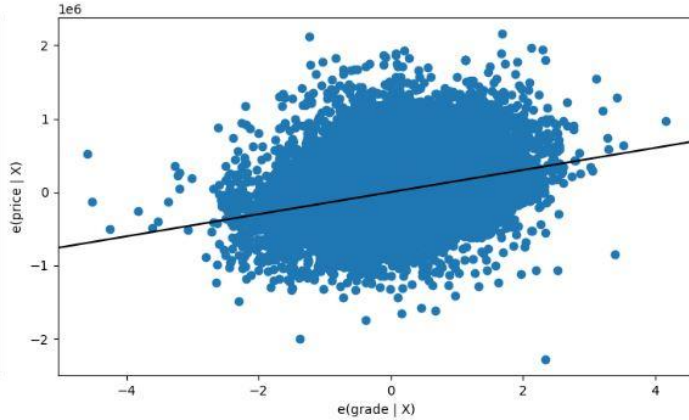
Partial Regression Plots

- Plot values not explained by model against one another
- Shows **benefit** of adding the variable in the model

Model 2



Partial Regression QQ Plot - Sqft_living

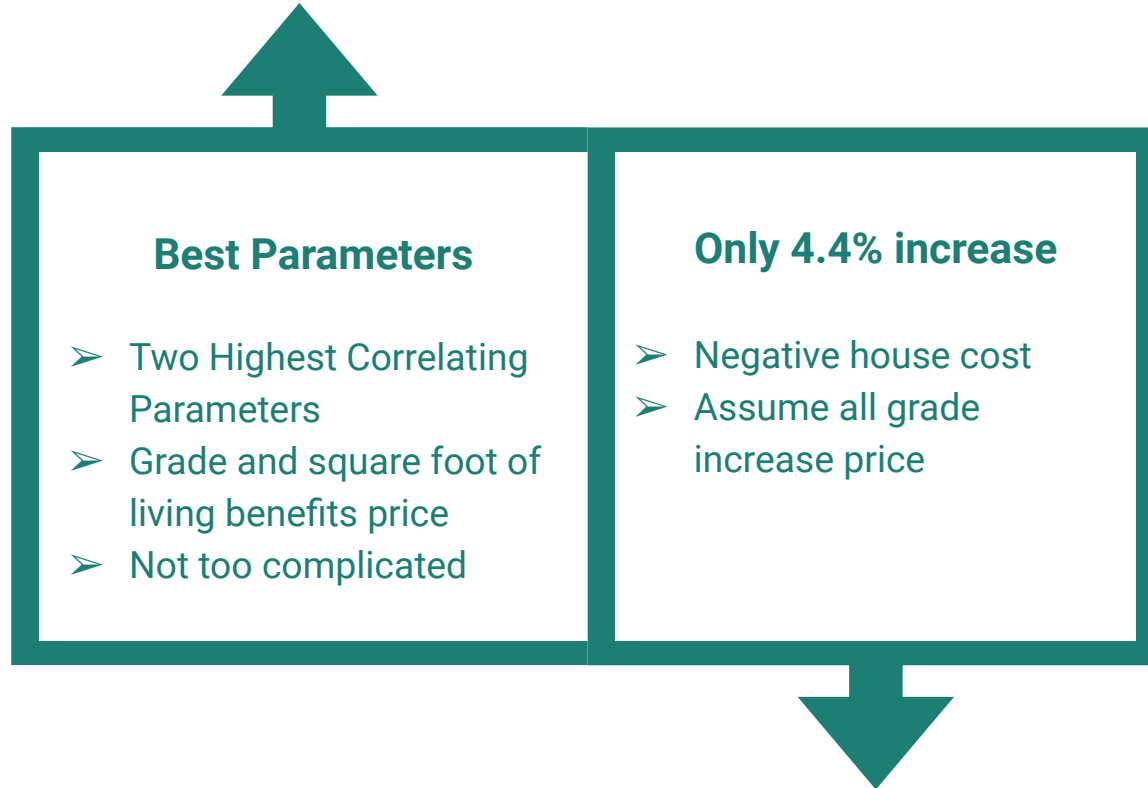


Partial Regression QQ Plot - Grade

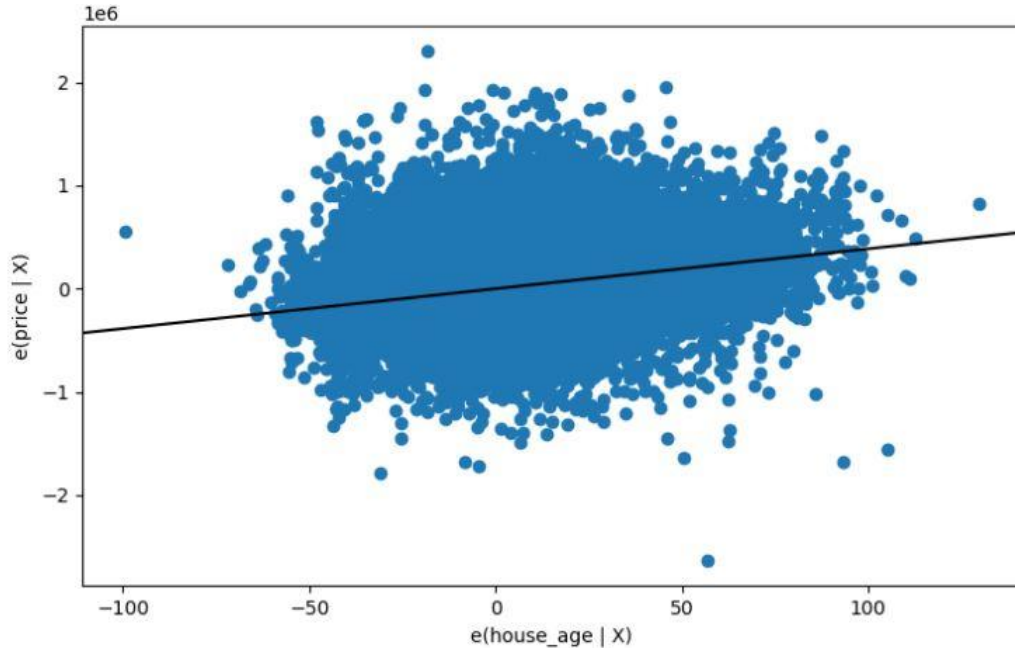
Interpretation -

- **\$232** per square foot of living
- **\$150,700** per grade (1-13)

Model 2 Limitations



Model 3

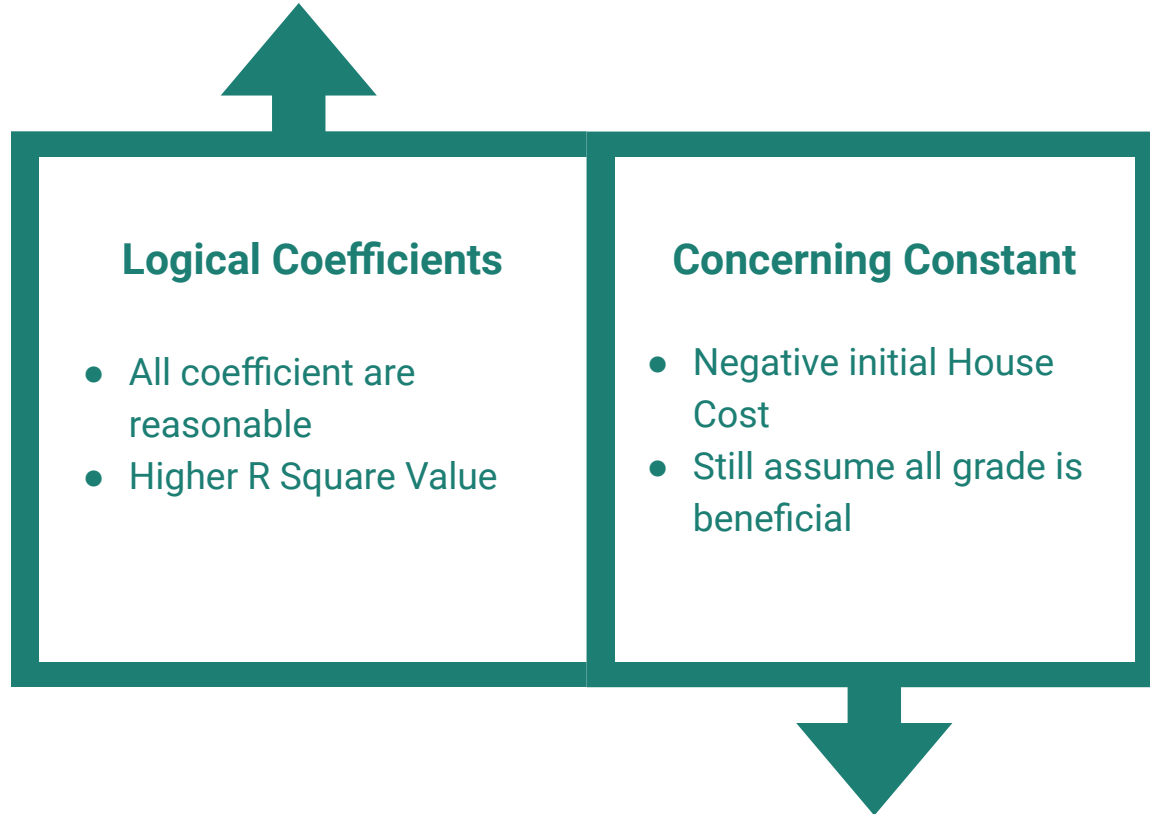


Interpretation -

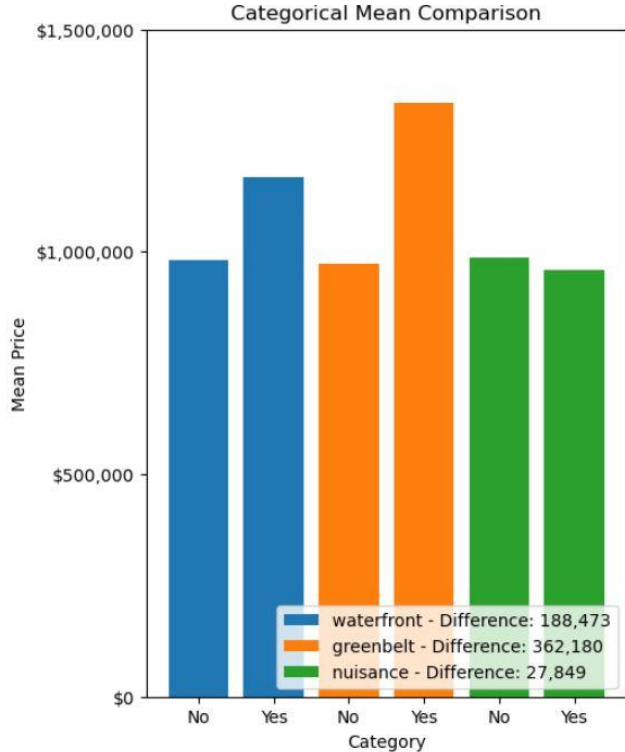
- **\$221** per square foot
- **\$216,400** per grade
- **\$3867** per age of house.

The more parameters added,
the larger the constant
becomes

Model 3 Limitations



Other Categorical Data



Method: Grouped by category with mean taken for each sub-category

Outcome:

No significant difference in Nuisance

Include: Waterfront and Greenbelt

Exclude: Nuisance

Final Model Reference

The model parameters are in reference to the following:

- No Greenbelt
- No view
- Average Condition
- Oil Heat Source
- Private Sewer System
- Building Grade of 7

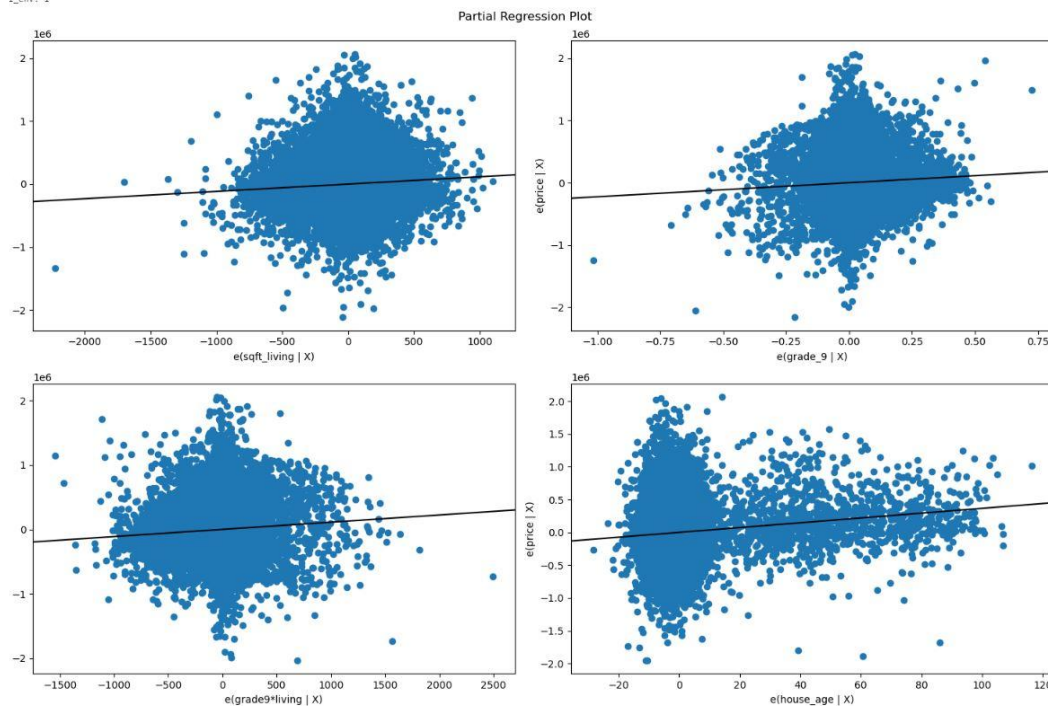
Final Model Key Coefficients 49.7%

1. **Sqft_living** - \$116 per square feet
2. **Grade** (1-13) - Depends on grade
3. **Greenbelt** - \$126,900 if on a greenbelt
4. **house_age**- \$3673 per year *
5. **Interaction Terms** (9) - \$112 per square feet if grade is 9 *

*** *Created variables added to the final model***

Other parameters didn't influence price heavily.

Final Model Regression Plots



Slight Positive Linear Relationship

- Not as strong as model 2
- House Age has clear relationship

Legend:

Top Left: Sqft_living

Top Right: Grade 9

Bottom Left: Sqft_living * Grade 9

Bottom Right: House Age

Grade Interpretation

Formula:

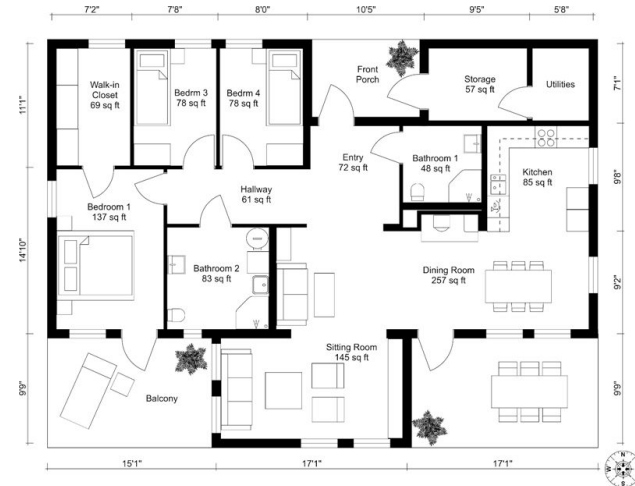
Price = Constant + grade + (interaction * square foot of living) + others*

- Grades below 7 are **negative**
- Grades above 7 are **positive**.
- Grade **doesn't match** their interaction terms.
 - (Positive Grade, Negative Living)
- **Less influence** the closer to Grade 7

Continuous Interpretation

In general, **all parameters** involving square foot of an area are **positive, excluding square foot of garage.**

- Garage = **-73** per square foot
- Patio = **63** per square foot
- Basement = **61** per square foot



Other Categorical Parameters

Greenbelt - Yes = \$126,900

View -

- Average = \$61,690
- Good = \$72,050
- Excellent = \$288,600

Condition -

- Good = 42,480
- Very Good = 106,700

In general, any parameters that is **better than the reference**, will **increase** house value.

Conclusion

- **More Parameters** → Better modeling
- **Negative Parameter** → More Positive Constant
- **Strongest Effect on Price:** Square living and Grade
- **Not all parameters are useful** such as sqft_above, nuisance, etc.

Recommendations: Zillow

Zillow: Use model to improve Zestimate.

➤ Improve Accuracy

Final Model: 49.7%

Can help explain half of the variance for Zestimate.

Recommendations: Buyer

Buyers: Inform on expensive and inexpensive aspect of a house.

➤ Price Checking

Expensive Attributes:

- View > Good
- Condition > Good
- Grade < 9

Inexpensive Attributes:

- Nuisances

Recommendations: Seller

Sellers: Renovate or improve aspects to increase price.

- Increase Home Values

Positive Attributes:

- Basement
- Building Grade
- Patio

Negative Attributes:

- Garage

Next Step

1. **More interaction terms**

- Greenbelt and Square Foot of Living

2. **More outside interaction**

- Schools, parks, crime rate

3. **Economic Status**

- Recession, Pandemic

Question?

Email: phungtommy109@gmail.com

Github: @Tommyphung1