MODELING SEATTLE HOUSING PRICES

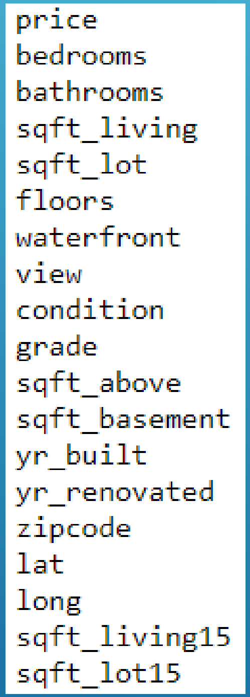
by Matthew E. Parker

INPUTS

Starting dataset:

> 21,500 house sales
19 variables per sale

variables list:

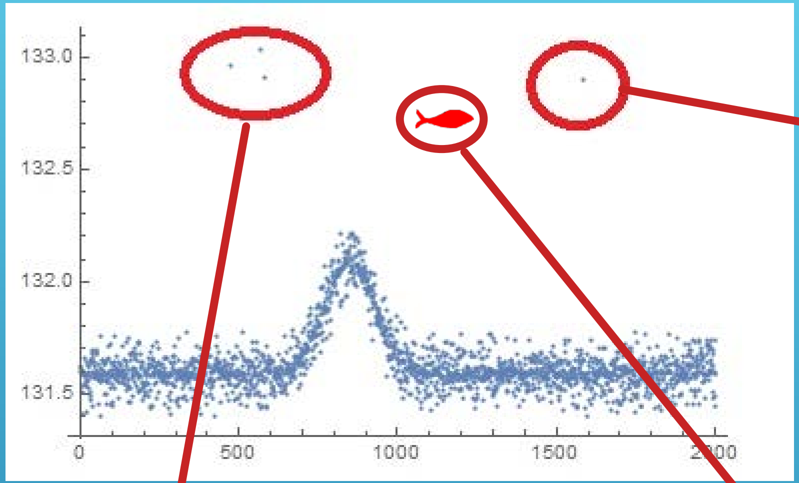


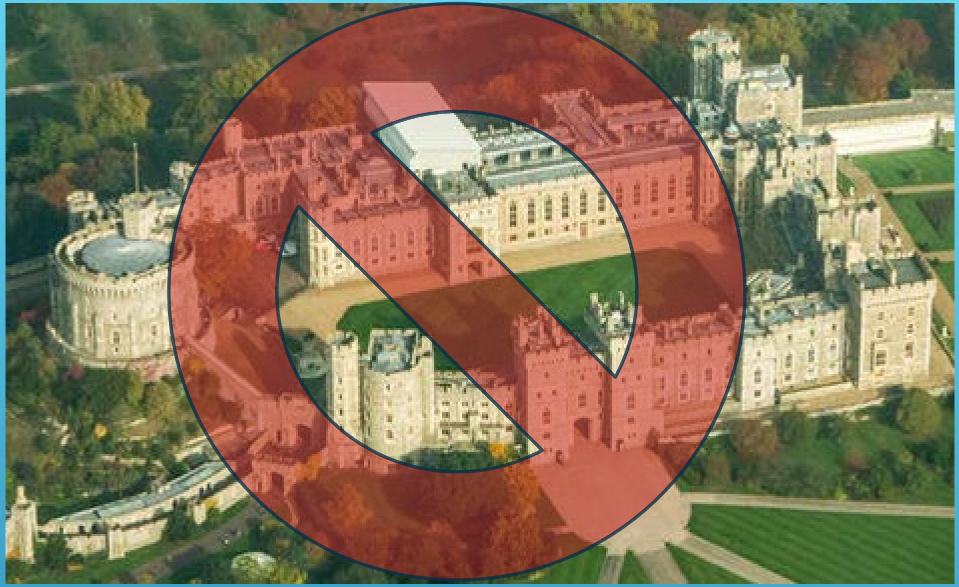
```
price
bedrooms
bathrooms
sqft_living
sqft_lot
floors
waterfront
view
condition
grade
sqft_above
sqft_basement
yr_built
yr_renovated
zipcode
lat
long
sqft_living15
sqft_lot15
```

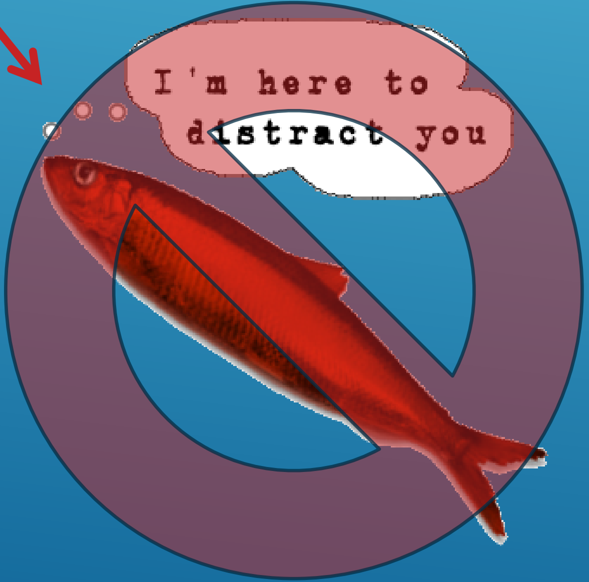
Our approach to constructing a model:

- Clean the data
- Explore and analyze the data
- Identify significant variables and build model around them
- Test and Validate model accuracy

Removing outliers from the dataset





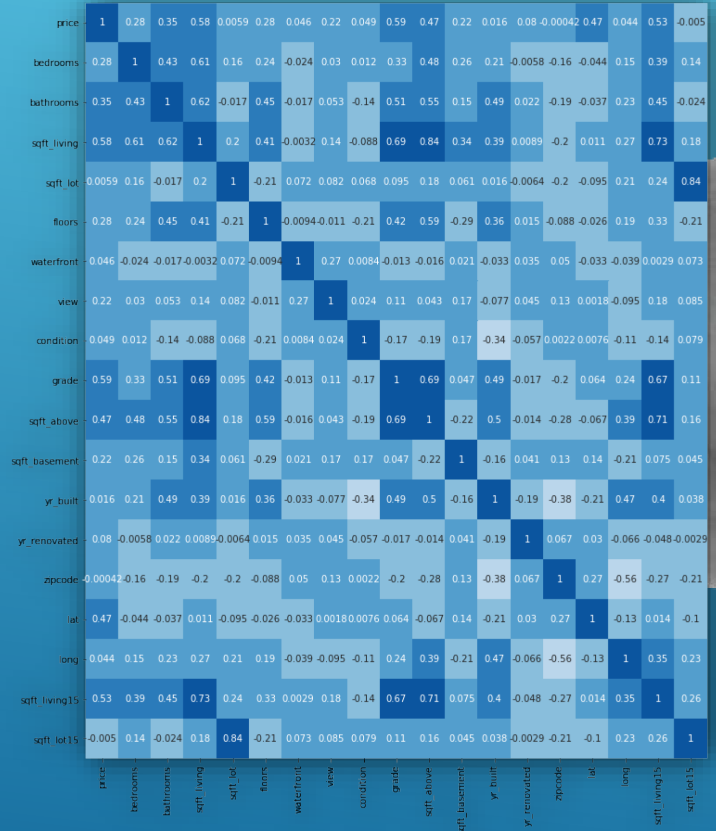


CLEANING THE DATA

EXPLORING THE DATA

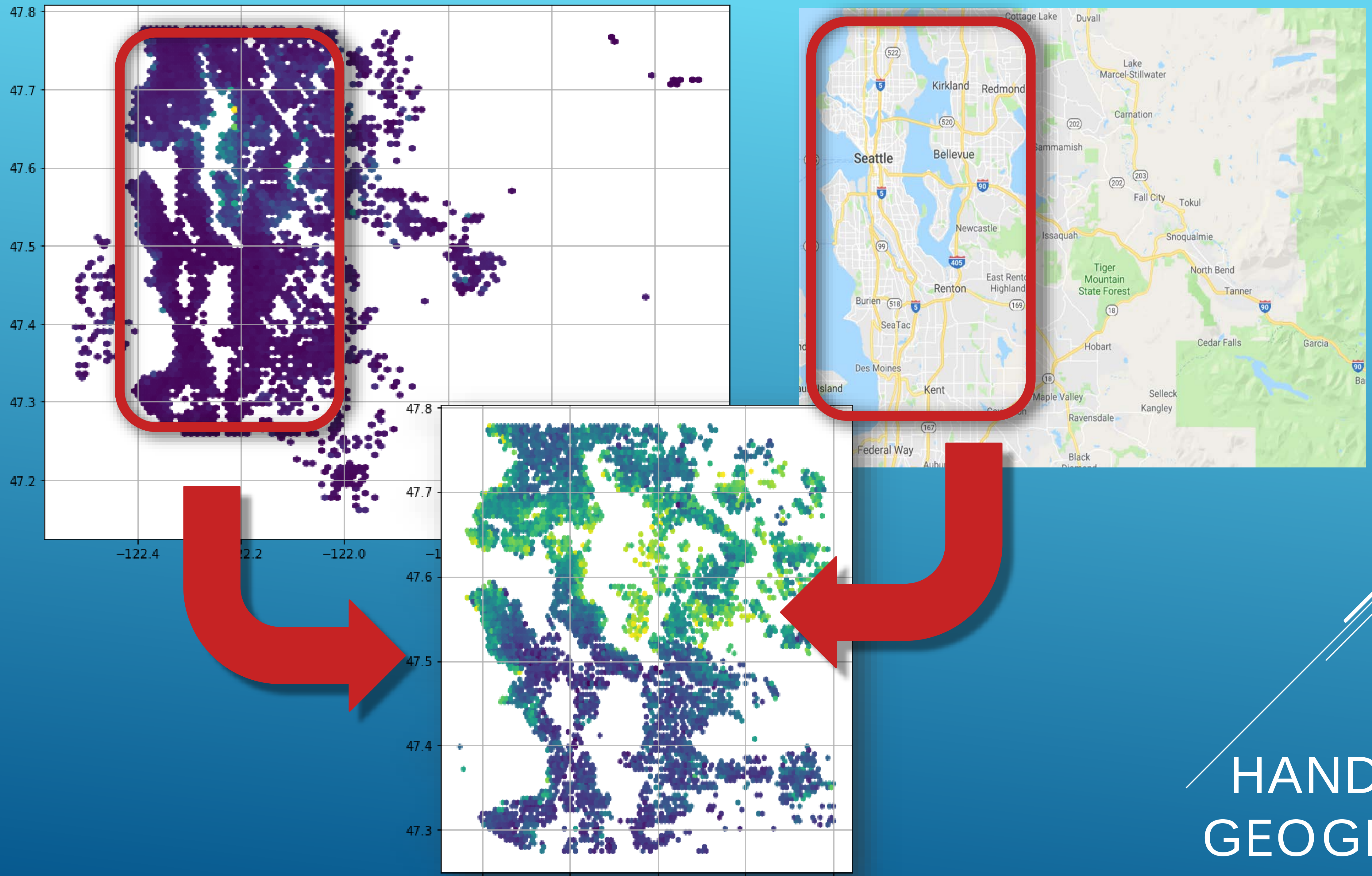
MAKING CONNECTIONS

Example of correlation matrix

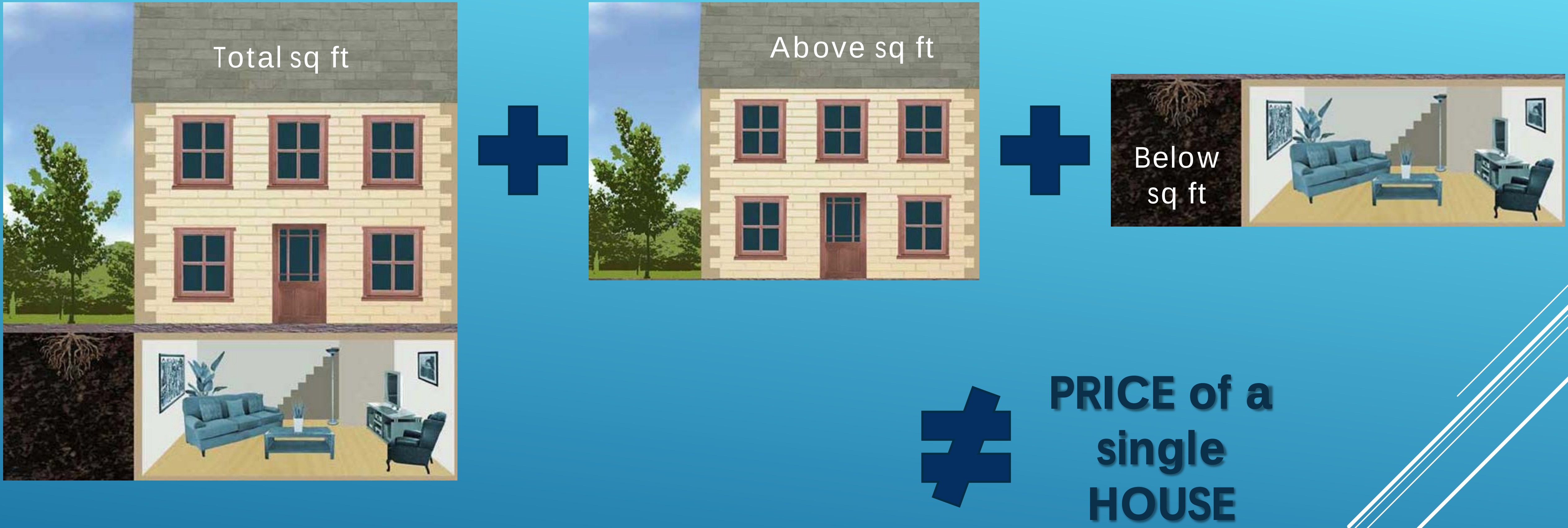


	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
price	1	0.28	0.35	0.58	0.0059	0.28	0.046	0.22	0.049	0.59	0.47	0.22	0.016	0.08	-0.00042	0.47	0.044	0.53	-0.005
bedrooms	0.28	1	0.43	0.61	0.16	0.24	-0.024	0.03	0.012	0.33	0.48	0.26	0.21	-0.0058	-0.16	-0.044	0.15	0.39	0.14
bathrooms	0.35	0.43	1	0.62	-0.017	0.45	-0.017	0.053	-0.14	0.51	0.55	0.15	0.49	0.022	-0.19	-0.037	0.23	0.45	-0.024
sqft_living	0.58	0.61	0.62	1	0.2	0.41	-0.0032	0.14	-0.088	0.69	0.84	0.34	0.39	0.0089	-0.2	0.011	0.27	0.73	0.18
sqft_lot	0.0059	0.16	-0.017	0.2	1	-0.21	0.072	0.082	0.068	0.095	0.18	0.061	0.016	-0.0064	-0.2	-0.095	0.21	0.24	0.84
floors	0.28	0.24	0.45	0.41	-0.21	1	-0.0094	-0.011	-0.21	0.42	0.59	-0.29	0.36	0.015	-0.088	-0.026	0.19	0.33	-0.21
waterfront	0.046	-0.024	-0.017	-0.0032	0.072	-0.0094	1	0.27	0.0084	-0.013	-0.016	0.021	-0.033	0.035	0.05	-0.033	-0.039	0.0029	0.073
view	0.22	0.03	0.053	0.14	0.082	-0.011	0.27	1	0.024	0.11	0.043	0.17	-0.077	0.045	0.13	0.0018	-0.095	0.18	0.085
condition	0.049	0.012	-0.14	-0.088	0.068	-0.21	0.0084	0.024	1	-0.17	-0.19	0.17	-0.34	-0.057	0.0022	0.0076	-0.11	-0.14	0.079
grade	0.59	0.33	0.51	0.69	0.095	0.42	-0.013	0.11	-0.17	1	0.69	0.047	0.49	-0.017	-0.2	0.064	0.24	0.67	0.11
sqft_above	0.47	0.48	0.55	0.84	0.18	0.59	-0.016	0.043	-0.19	0.69	1	-0.22	0.5	-0.014	-0.28	-0.067	0.39	0.71	0.16
sqft_basement	0.22	0.26	0.15	0.34	0.061	-0.29	0.021	0.17	0.17	0.047	-0.22	1	-0.16	0.041	0.13	0.14	-0.21	0.075	0.045
yr_built	0.016	0.21	0.49	0.39	0.016	0.36	-0.033	-0.077	-0.34	0.49	0.5	-0.16	1	-0.19	-0.38	-0.21	0.47	0.4	0.038
yr_renovated	0.08	-0.0058	0.022	0.0089	-0.0064	0.015	0.035	0.045	-0.057	-0.017	-0.014	0.041	-0.19	1	0.067	0.03	-0.066	-0.048	-0.0029
zipcode	-0.00042	-0.16	-0.19	-0.2	-0.2	-0.088	0.05	0.13	0.0022	-0.2	-0.28	0.13	-0.38	0.067	1	0.27	-0.56	-0.27	-0.21
lat	0.47	-0.044	-0.037	0.011	-0.095	-0.026	-0.033	0.0018	0.0076	0.064	-0.067	0.14	-0.21	0.03	0.27	1	-0.13	0.014	-0.1
long	0.044	0.15	0.23	0.27	0.21	0.19	-0.039	-0.095	-0.11	0.24	0.39	-0.21	0.47	-0.066	-0.56	-0.13	1	0.35	0.23
sqft_living15	0.53	0.39	0.45	0.73	0.24	0.33	0.0029	0.18	-0.14	0.67	0.71	0.075	0.4	-0.048	-0.27	0.014	0.35	1	0.26
sqft_lot15	-0.005	0.14	-0.024	0.18	0.84	-0.21	0.073	0.085	0.079	0.11	0.16	0.045	0.038	-0.0029	-0.21	-0.1	0.23	0.26	1

Using statistics to identify the variables with the greatest influence upon housing prices



HANDLING GEOGRAPHY



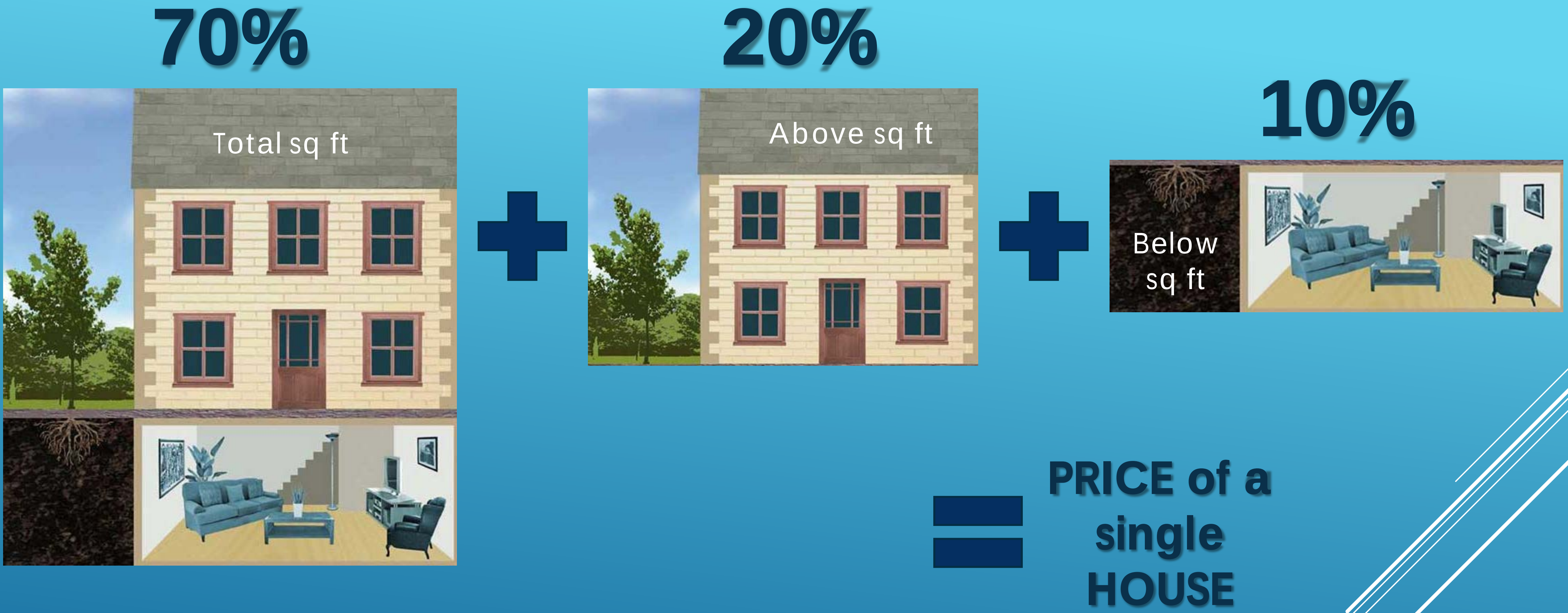
Total sq ft + Above sq ft + Below sq ft ≠ **PRICE of a single HOUSE**

Many variables are closely related, like **total sq ft**, **above-ground sq ft**, and **below-ground sq ft**.

Changing one variable can often impact another. This is bad for modeling as it produces a multiplied effect.

POTENTIAL PROBLEMS

BUILDING CUSTOM FEATURES



70%

Total sq ft

+

20%

Above sq ft

+

10%

Below sq ft

=

PRICE of a single HOUSE

We don't want to just remove variables, as the presence or absence of a basement may influence pricing.

To solve this, we can <u>build features</u> that weight the variables proportionally to their influence.

TRAINING & VALIDATING THE MODEL

Trained by 10,000 testing cycles

$$p = (1296579.6 \times f_L) + (60401.32 \times f_V) + (110.79 \times f_B) + (59144.57 \times f_G) + (111362.3 \times log_e(f_S)) - 62564620.4$$

$p =$ House price (in USD)

$f_L =$ latitude

$f_V =$ times property has been viewed

$f_B =$ square footage of basement

$f_G =$ grade given to the housing unit, based on King County grading system

$f_S =$ square footage of living space

MODEL SUMMARY

If you know a house's latitude, basement ft^2, living space ft^2, King County grade, and the number of times it has been viewed, then you can estimate it's sale price within an error margin of $126,700.^{00}.

Price =

Latitude +
Views +
Basement sq ft +
King County grade +
Total sq ft

RECOMMENDATIONS

1. If you can purchase a house for $126,700 less than the price predicted by our model, you will definitely make a profit.

2. Houses in the northern half of Kings County fetch higher prices, try to sell northern properties

3. The more times a house has been viewed, the higher it's final selling price is likely to be. Invest in advertising your properties.

FURTHER INVESTIGATION

The current model could likely be enhanced by the addition of more variables.

In particular, information on crime rates, transportation accessibility, school district ratings, etc. would be useful as these factors have in the past been shown to influence real estate pricing.

THANK YOU