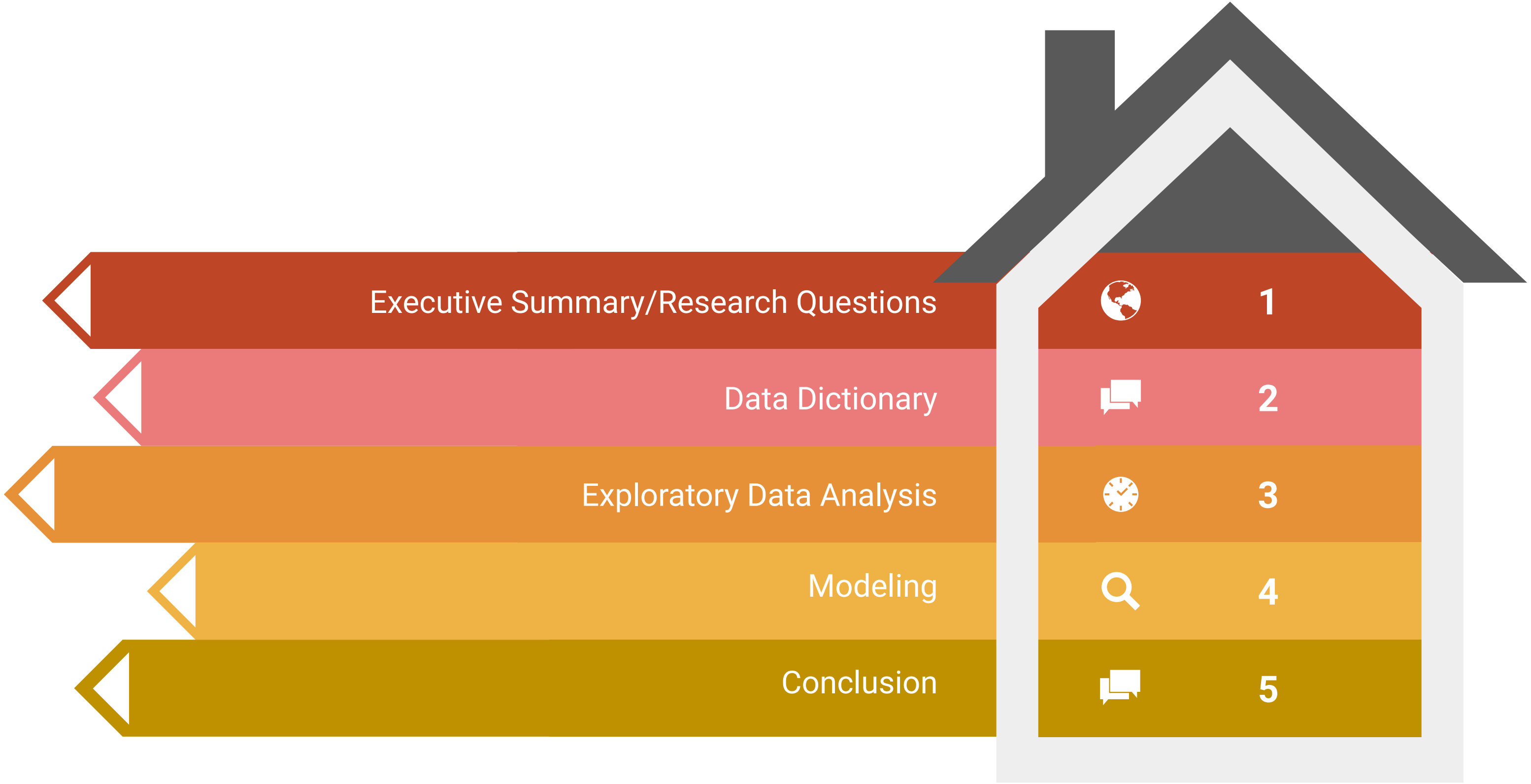
Statistical Modeling of House Prices

Tianshu Fan, Jaehee Jeong, Lisa Kaunitz

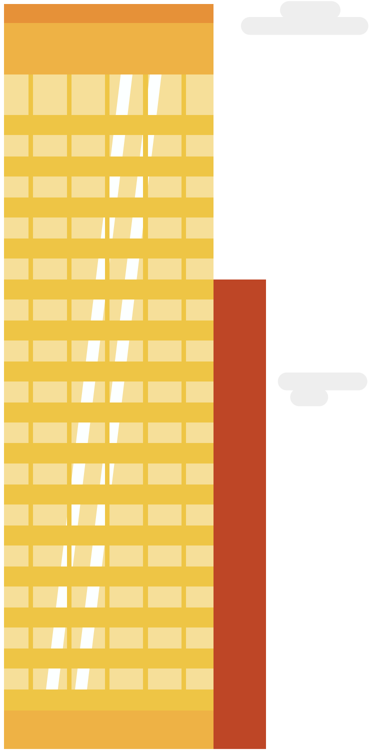
STAT 412 | FALL 2021

Content



- Executive Summary/Research Questions 1
- Data Dictionary 2
- Exploratory Data Analysis 3
- Modeling 4
- Conclusion 5

1. Executive Summary/Research Questions



Executive Summary

Goal: Find significant variables that affect housing price and fit models to predict prices for King County, WA.

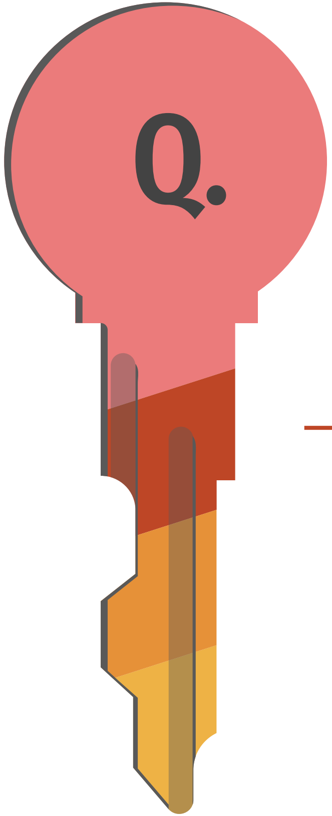
What we did:

- Explored the dataset and transformed certain features
- Look at feature importances
- Built various models

Findings:

- Sqft of living space, quality of house, location matter most
- Random Forest performed best

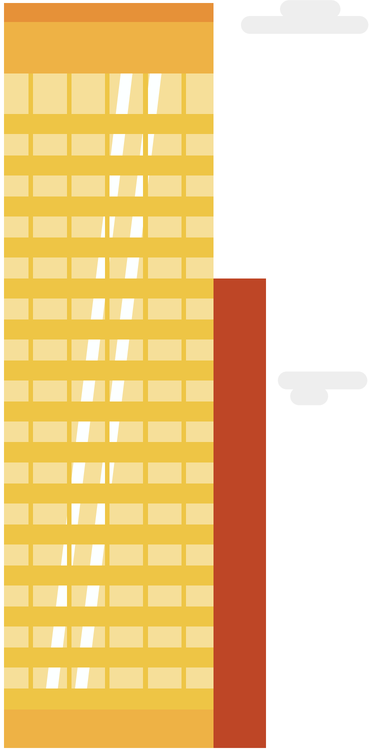
Research Questions



1 **Significant variables** What variables are significant in predicting the price of house

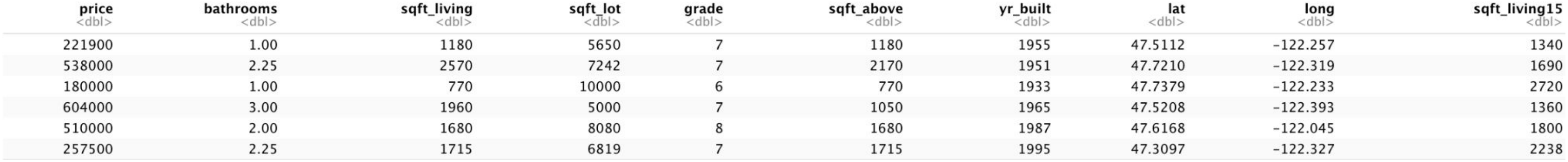
2 **Model Selection** Which model best predicts the price

2. Data Dictionary



Overview of the Data

- Our dataset contains house sale prices for King County, WA from 2014-05-02 to 2015-05-24
- 21,613 observations and 21 variables
- The variables describe housing features, rather than features about the population
- Source: Kaggle



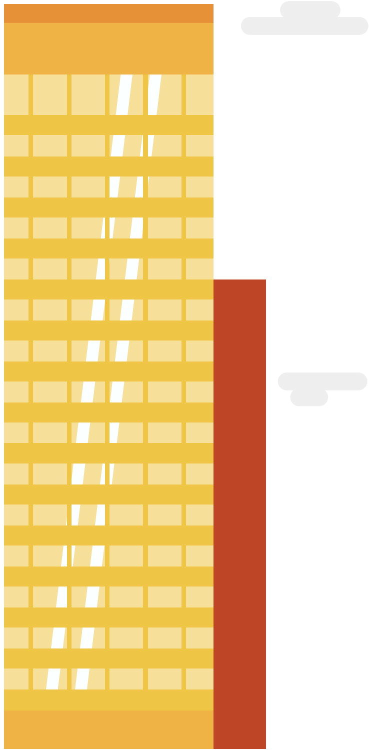
price	bathrooms	sqft_living	sqft_lot	grade	sqft_above	yr_built	lat	long	sqft_living15
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
221900	1.00	1180	5650	7	1180	1955	47.5112	-122.257	1340
538000	2.25	2570	7242	7	2170	1951	47.7210	-122.319	1690
180000	1.00	770	10000	6	770	1933	47.7379	-122.233	2720
604000	3.00	1960	5000	7	1050	1965	47.5208	-122.393	1360
510000	2.00	1680	8080	8	1680	1987	47.6168	-122.045	1800
257500	2.25	1715	6819	7	1715	1995	47.3097	-122.327	2238

Data Dictionary

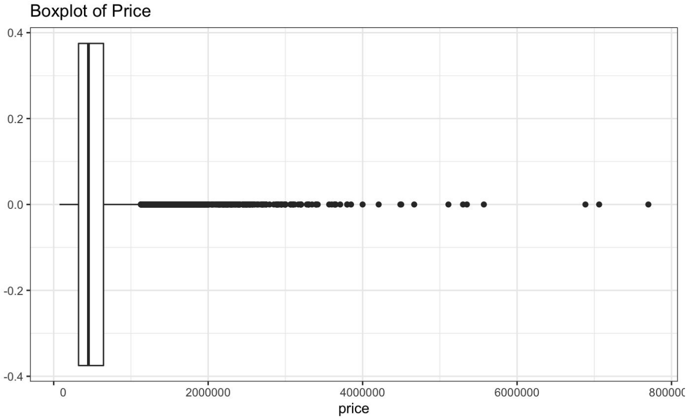
Column	Data Type	Description
id	num	Unique ID for each home sold
date	date	Date of the house sale between 2014-05-02 to 2015-05-24
price	num	Price of each home sold
bedrooms	int	Number of bedrooms
bathrooms	num	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
sqft_living	int	Area of the house interior living space measured in square feet
sqft_lot	int	Area of the land space measured in square feet
floors	num	Number of floors
waterfront	int	A indicator variable for whether the house was overlooking the waterfront or not
view	int	An index from 0 to 4 of how good the view of the property was
condition	int	An index from 1 to 5 on the condition of the house

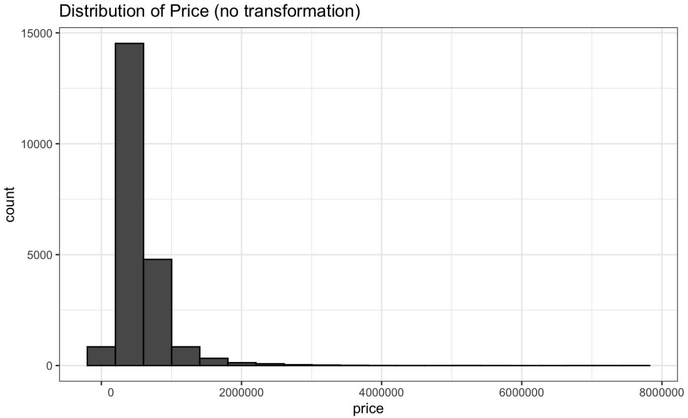
Column	Data Type	Description
grade	int	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design
sqft_above	int	Area of the interior housing space that is above ground level measured in square feet
sqft_basement	int	Area of the interior housing space that is below ground level measured in square feet
sqft_basement_yesno	boolean	Whether the house has a basement or not
yr_built	int	The year the house was initially built
yr_renovated	int	The year of the house’s last renovation
zipcode	int	What zipcode area the house is in
lat	num	Latitude
long	num	Longitude
sqft_living15	int	Average of the area of interior housing living space for the nearest 15 neighbors measured in square feet
sqft_lot15	int	Average of the area of the land lots of the nearest 15 neighbors measure in square feet

3. Exploratory Data Analysis



Dealing with Response: Price





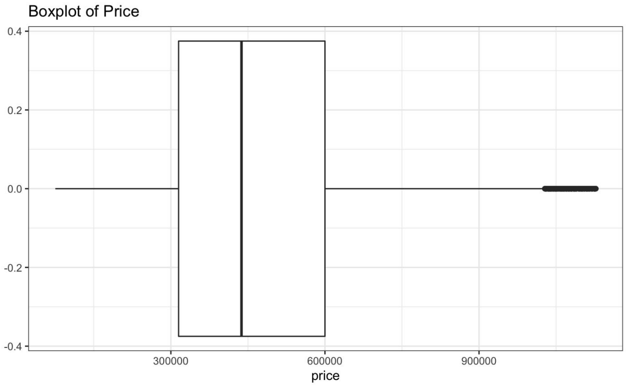
Original Dataset:

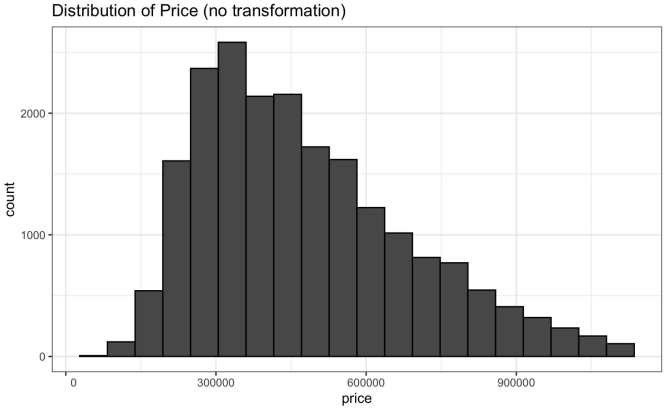
- The distribution is heavily skewed.
- most of the price below 1 million

Removing outliers

- Keep data points within 1.5*IQR
  - IQR = Q3 - Q1
  - Lower fence: Q1 - 1.5 IQR
  - Upper fence: Q3 + 1.5 IQR

Dealing with Response: Price



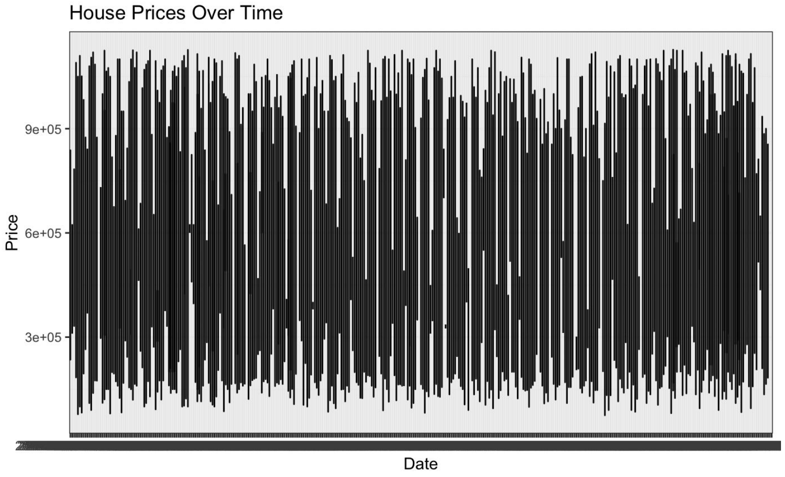


After we dropped outliers:

- 20194 out of 21613 kept (93.4%)
- Considered log-transform
    - But want to keep interpretability

Drop features

Features	Reason
date	No pattern
id	No meaning
zipcode	Represented by latitude and longitude
sqft_basement	Represented by sqft_basement_yes no

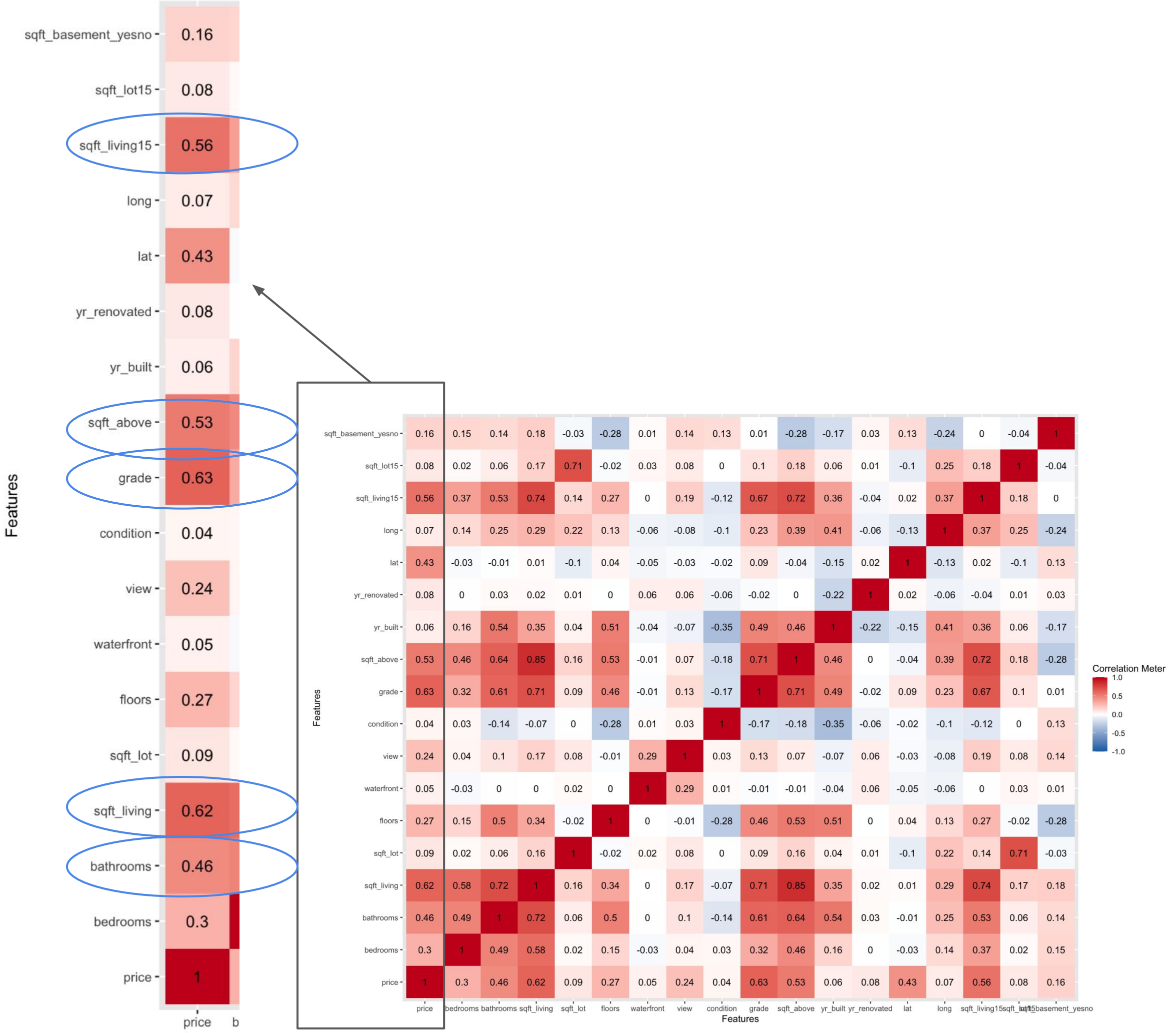


Feature Selection

Method 1:

Checking Correlation with Price

- ○ Sqft_living
- ○ Grade
- ○ Sqft_above
- ○ Sqft_living15
- ○ Bathrooms



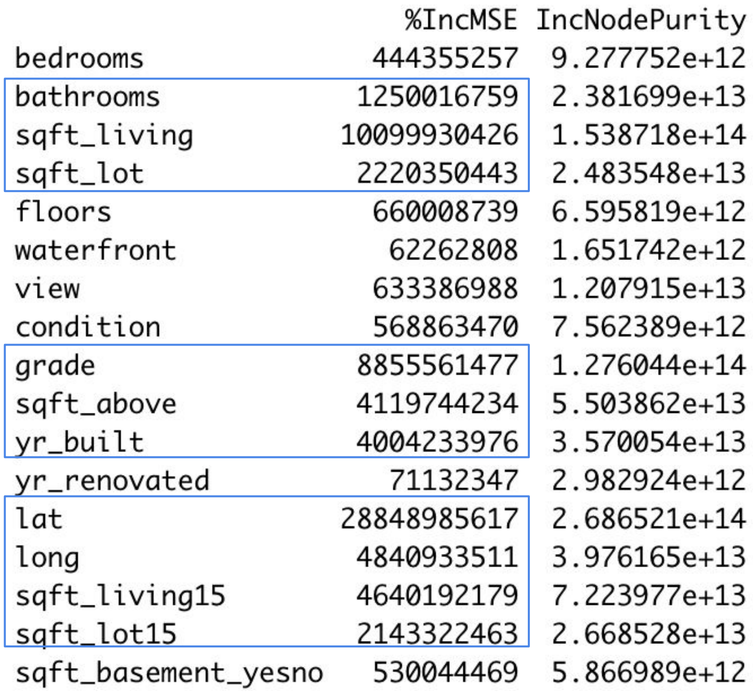
Features	price
sqft_basement_yesno	0.16
sqft_lot15	0.08
sqft_living15	0.56
long	0.07
lat	0.43
yr_renovated	0.08
yr_built	0.06
sqft_above	0.53
grade	0.63
condition	0.04
view	0.24
waterfront	0.05
floors	0.27
sqft_lot	0.09
sqft_living	0.62
bathrooms	0.46
bedrooms	0.3
price	1

Features	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	yr_built	yr_renovated	lat	long	sqft_living15	sqft_lot15	sqft_basement_yesno
sqft_basement_yesno	0.16	0.15	0.14	0.18	-0.03	-0.28	0.01	0.14	0.13	0.01	-0.28	-0.17	0.03	0.13	-0.24	0	-0.04	1
sqft_lot15	0.08	0.02	0.06	0.17	0.71	-0.02	0.03	0.08	0	0.1	0.18	0.06	0.01	-0.1	0.25	0.18	1	-0.04
sqft_living15	0.56	0.37	0.53	0.74	0.14	0.27	0	0.19	-0.12	0.67	0.72	0.36	-0.04	0.02	0.37	1	0.18	0
long	0.07	0.14	0.25	0.29	0.22	0.13	-0.06	-0.08	-0.1	0.23	0.39	0.41	-0.06	-0.13	1	0.37	0.25	-0.24
lat	0.43	-0.03	-0.01	0.01	-0.1	0.04	-0.05	-0.03	-0.02	0.09	-0.04	-0.15	0.02	1	-0.13	0.02	-0.1	0.13
yr_renovated	0.08	0	0.03	0.02	0.01	0	0.06	0.06	-0.06	-0.02	0	-0.22	1	0.02	-0.06	-0.04	0.01	0.03
yr_built	0.06	0.16	0.54	0.35	0.04	0.51	-0.04	-0.07	-0.35	0.49	0.46	1	-0.22	-0.15	0.41	0.36	0.06	-0.17
sqft_above	0.53	0.46	0.64	0.85	0.16	0.53	-0.01	0.07	-0.18	0.71	1	0.46	0	-0.04	0.39	0.72	0.18	-0.28
grade	0.63	0.32	0.61	0.71	0.09	0.46	-0.01	0.13	-0.17	1	0.71	0.49	-0.02	0.09	0.23	0.67	0.1	0.01
condition	0.04	0.03	-0.14	-0.07	0	-0.28	0.01	0.03	1	-0.17	-0.18	-0.35	-0.06	-0.02	-0.1	-0.12	0	0.13
view	0.24	0.04	0.1	0.17	0.08	-0.01	0.29	1	0.03	0.13	0.07	-0.07	0.06	-0.03	-0.08	0.19	0.08	0.14
waterfront	0.05	-0.03	0	0	0.02	0	1	0.29	0.01	-0.01	-0.01	-0.04	0.06	-0.05	-0.06	0	0.03	0.01
floors	0.27	0.15	0.5	0.34	-0.02	1	0	-0.01	-0.28	0.46	0.53	0.51	0	0.04	0.13	0.27	-0.02	-0.28
sqft_lot	0.09	0.02	0.06	0.16	1	-0.02	0.02	0.08	0	0.09	0.16	0.04	0.01	-0.1	0.22	0.14	0.71	-0.03
sqft_living	0.62	0.58	0.72	1	0.16	0.34	0	0.17	-0.07	0.71	0.85	0.35	0.02	0.01	0.29	0.74	0.17	0.18
bathrooms	0.46	0.49	1	0.72	0.06	0.5	0	0.1	-0.14	0.61	0.64	0.54	0.03	-0.01	0.25	0.53	0.06	0.14
bedrooms	0.3	1	0.49	0.58	0.02	0.15	-0.03	0.04	0.03	0.32	0.46	0.16	0	-0.03	0.14	0.37	0.02	0.15
price	1	0.3	0.46	0.62	0.09	0.27	0.05	0.24	0.04	0.63	0.53	0.06	0.08	0.43	0.07	0.56	0.08	0.16

Feature Selection

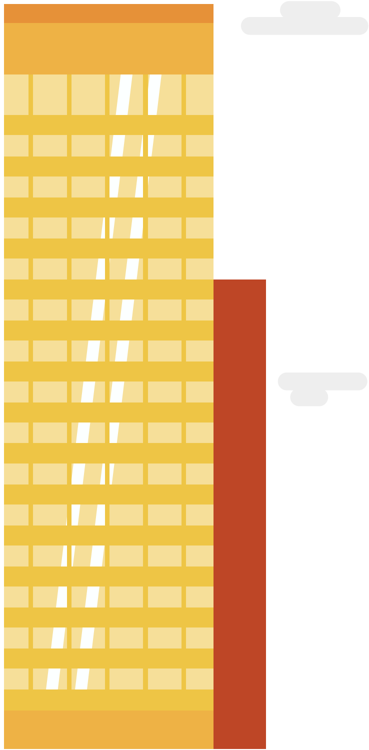
Method 2: **Utilizing the importance feature of the random forest model**

- ○ Lat
- ○ Sqft_living
- ○ Grade
- ○ Sqft_living15
- ○ Sqft_above
- ○ Long
- ○ Yr_built
- ○ Sqft_lot15
- ○ sqft_lot
- ○ Bathrooms



```
                        %IncMSE IncNodePurity
bedrooms              444355257  9.277752e+12
bathrooms            1250016759  2.381699e+13
sqft_living         10099930426  1.538718e+14
sqft_lot             2220350443  2.483548e+13
floors                660008739  6.595819e+12
waterfront             62262808  1.651742e+12
view                  633386988  1.207915e+13
condition             568863470  7.562389e+12
grade                8855561477  1.276044e+14
sqft_above           4119744234  5.503862e+13
yr_built             4004233976  3.570054e+13
yr_renovated           71132347  2.982924e+12
lat                 28848985617  2.686521e+14
long                 4840933511  3.976165e+13
sqft_living15        4640192179  7.223977e+13
sqft_lot15           2143322463  2.668528e+13
sqft_basement_yesno   530044469  5.866989e+12
```

4. Modeling



Model Results

Using the top 10 variables by feature importance, we achieved the following results

Model	RMSE
Simple Linear Regression	120,550
Step Regression	120,550
PCR	153,961
Lasso	120,550
Ridge	120,550
Random Forest	84,023

Simple Linear Regression

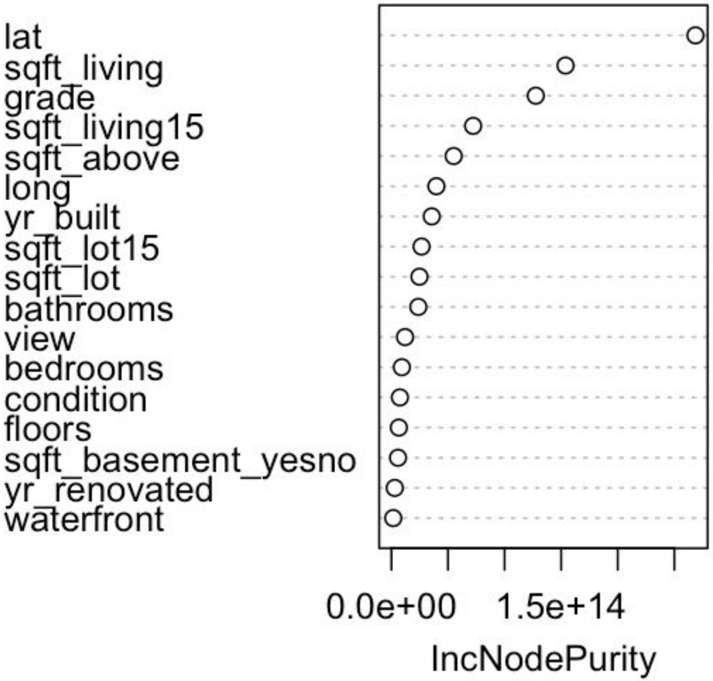
All else held constant...

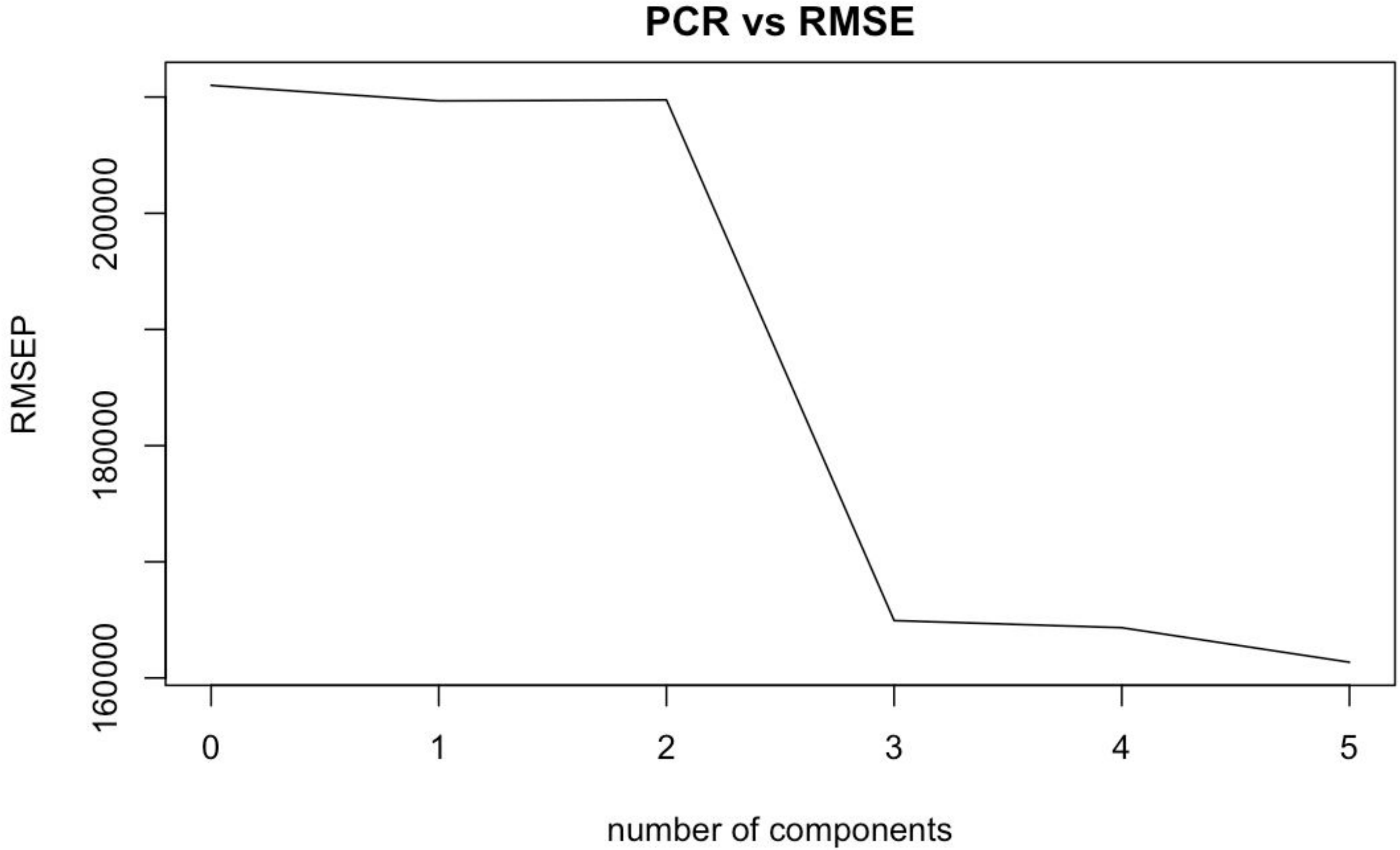
- 1 unit increase in the quality or “grade” of the home results in a price increase of $77,660

- For each increase of 100 sqft average of the homes in your neighborhood, will result in a price increase of about $5000

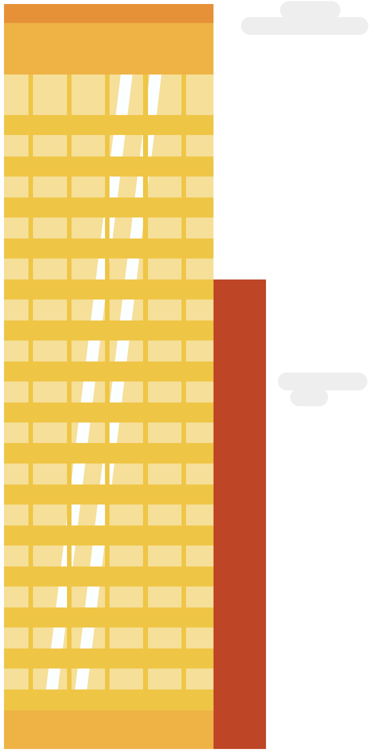
- Each added bathroom will increase the price of the home value by about $3000

Random Forest Regression




PCR vs RMSE
RMSEP
1600000
1800000
2000000
0
1
2
3
4
5
number of components

5. Conclusion

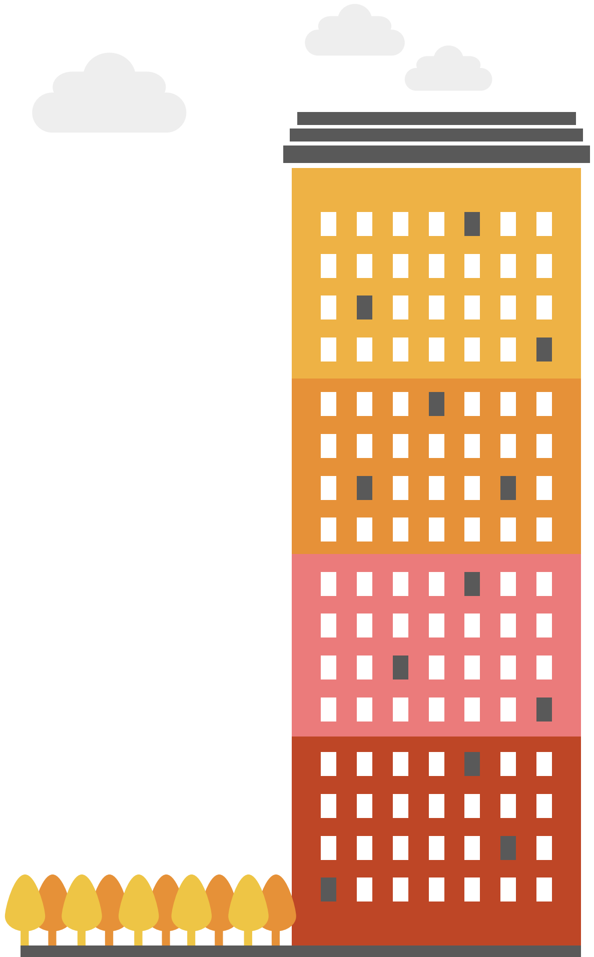


Conclusion

Location: Homes more West will have a higher value

Selected model: Random forest model

Important variables: Lat, Sqft_living, Grade, Sqft_living15, Sqft_above, Long, Yr_built, Sqft_lot15, sqft_lot Bathrooms



Shortcomings

- Hard to interpret the selected model
- Error is still too large
- Limited range of house prices

Future Recommendations

Data Acquisition

- Getting the population in the area
- Checking effects of the population
- Having more than 1 year of data

Thank you
Questions?