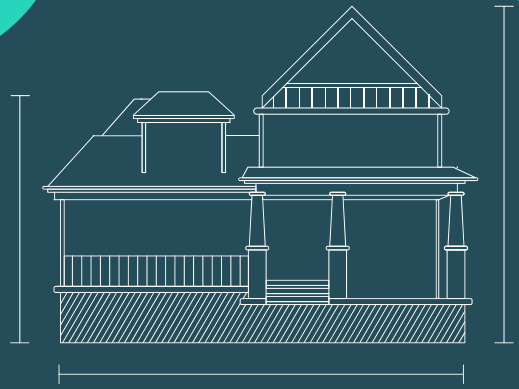
King County

Ian Kindle-Pyle
Astha Grover
Elizabeth Evans
Freda Li
Nidhi Pancholi
Dajia Bao

Who we are

Team of data scientists

01



02

Our problem statement

There is a lack of clarity on the housing price of Seattle

What we want to achieve

Predict the right price of the house in Seattle

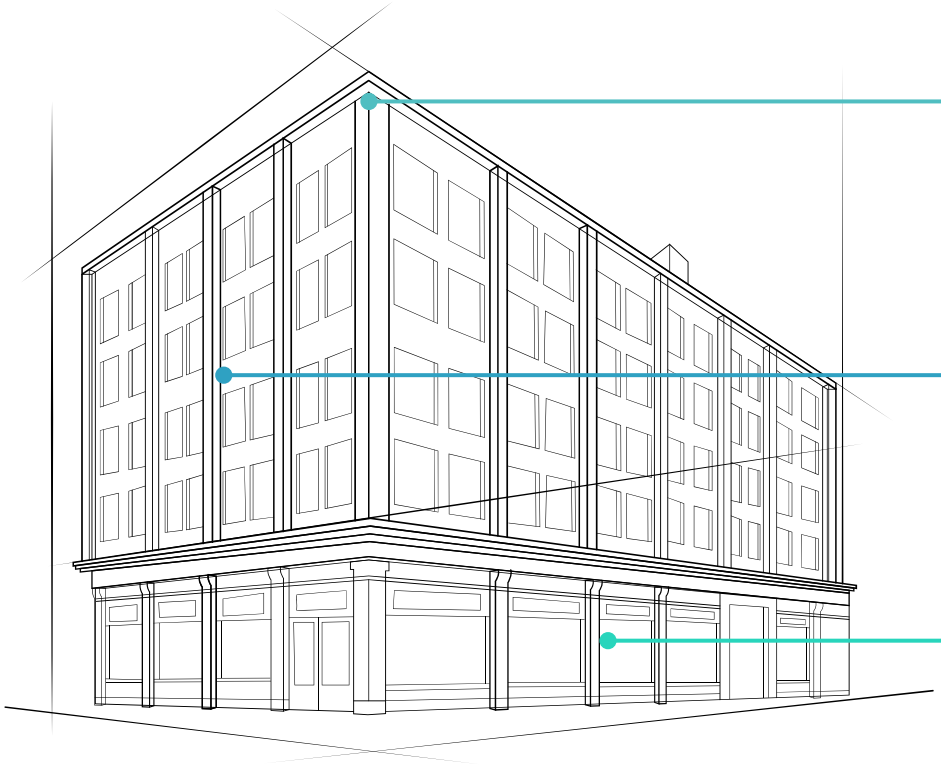
03

04

What's our future goal

Increase the scope of the project from only Seattle to Washington

Facts



There are nearly as many pieces of digital information as there are stars in the universe.

Bad data costs US businesses alone $600 billion annually.

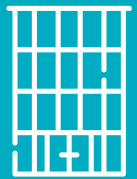
The US leads the data science market, requiring 190,000 data scientists by next year.

Outline

1

Introduction

2



Data exploration & Model selection

3

Insight & conclusions

4

Challenges & Improvements

Introduction

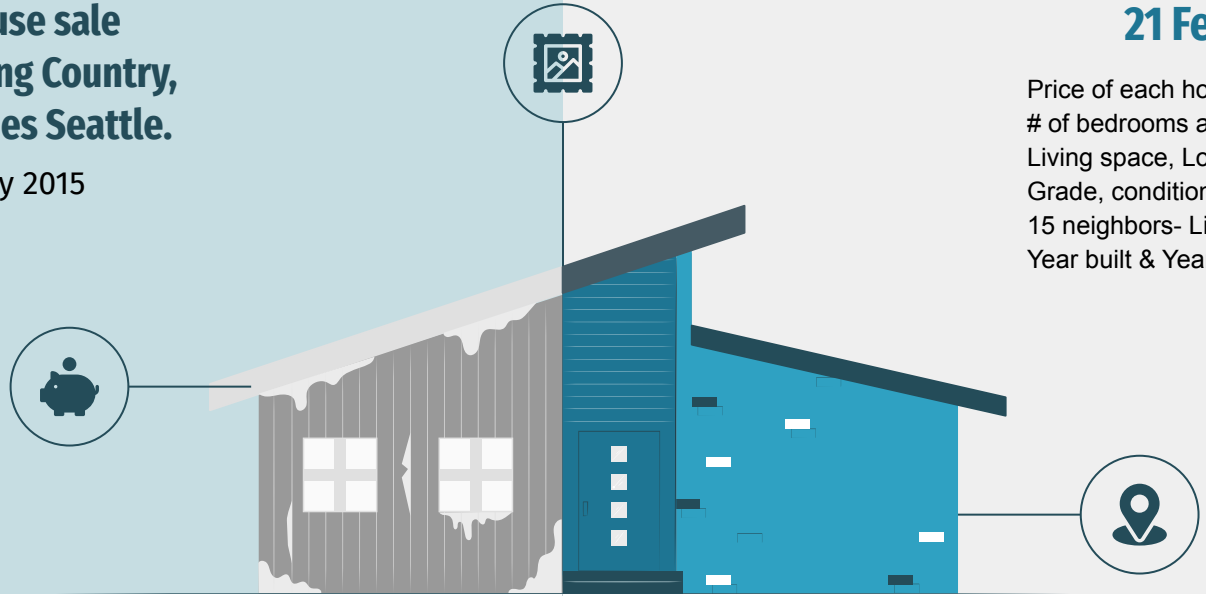
How much is this house?

Data description

01

Contains house sale prices for King Country, which includes Seattle.

May 2014 - May 2015



02

21 Features

Price of each house sold
# of bedrooms and bathrooms
Living space, Lot size
Grade, condition, view, waterfront
15 neighbors- Living space,Lot size
Year built & Year renovated

Data preprocessing

01

Removed outlier data with 33 bedrooms

02

Removed houses with zero bedrooms

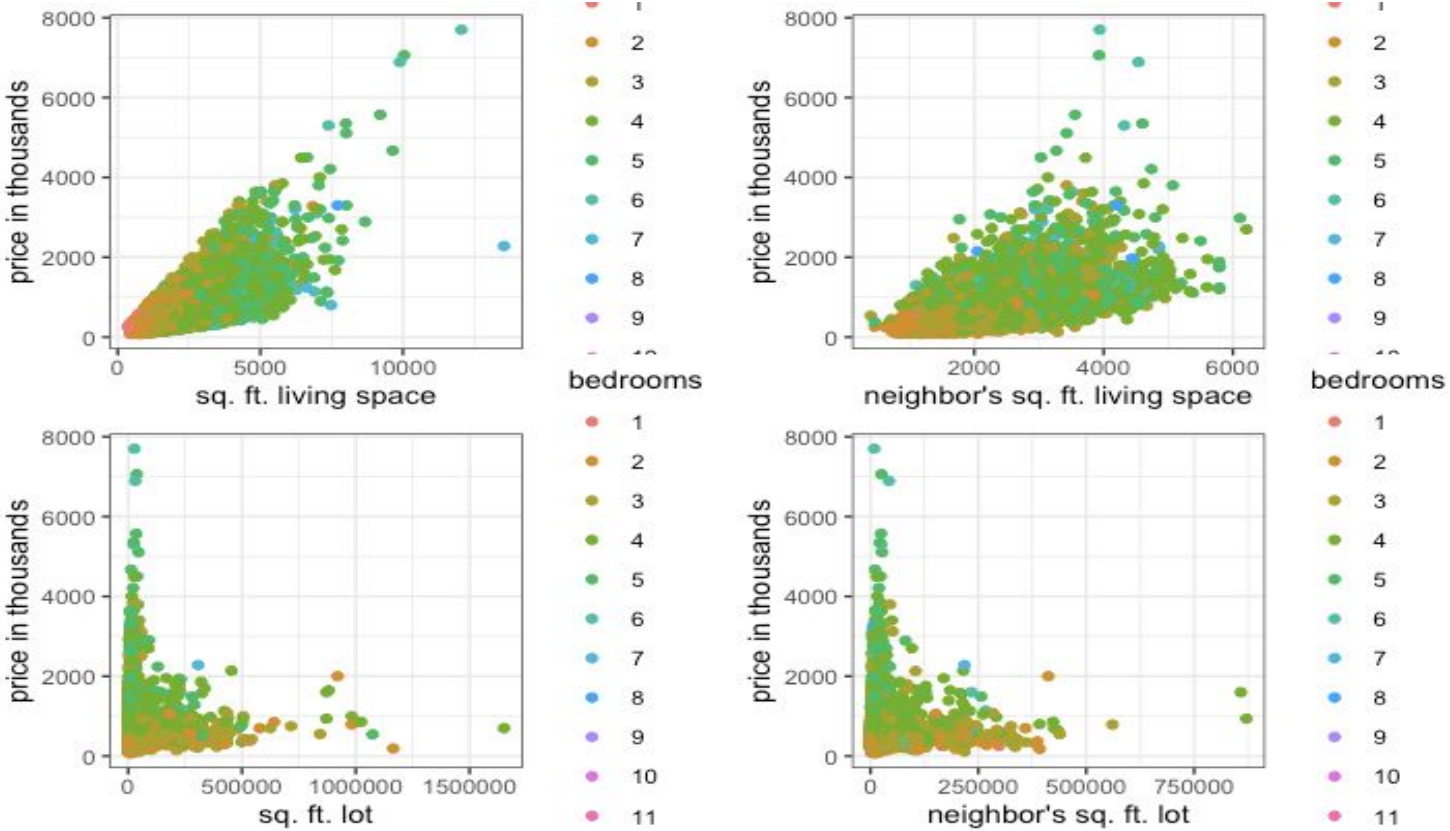
03

Removed houses with zero bathrooms

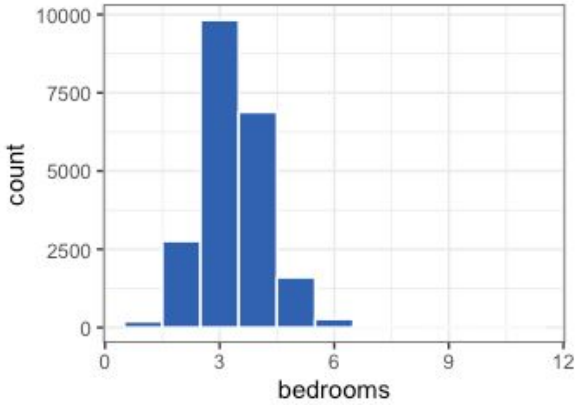
04

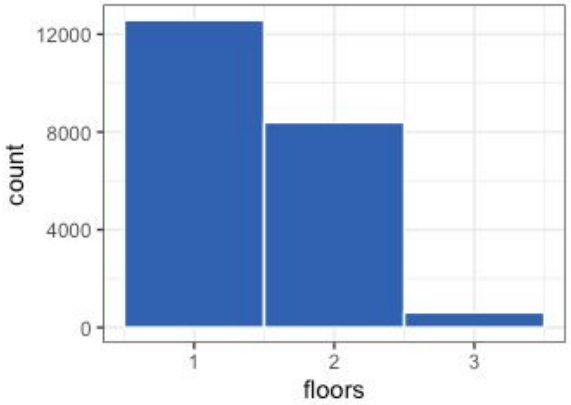
Converted dates column to date type

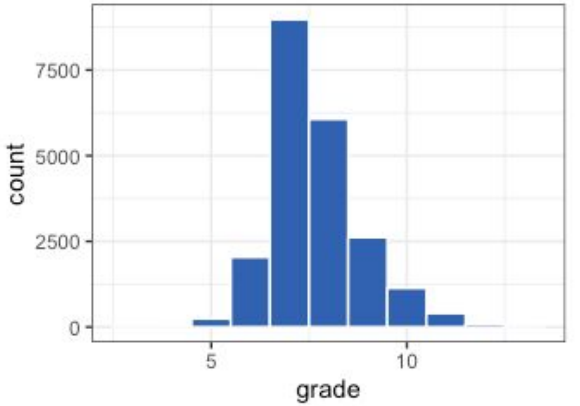
Data exploration

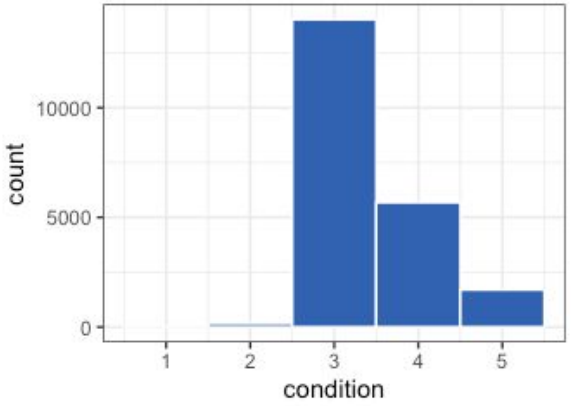


Data exploration - Frequency Distribution

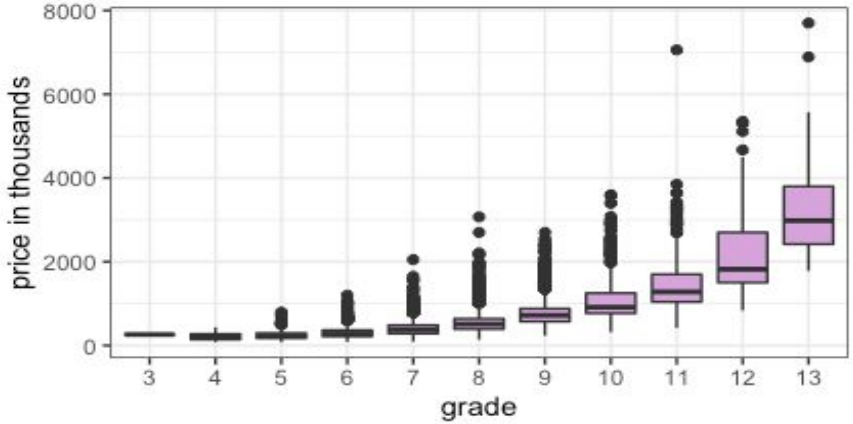


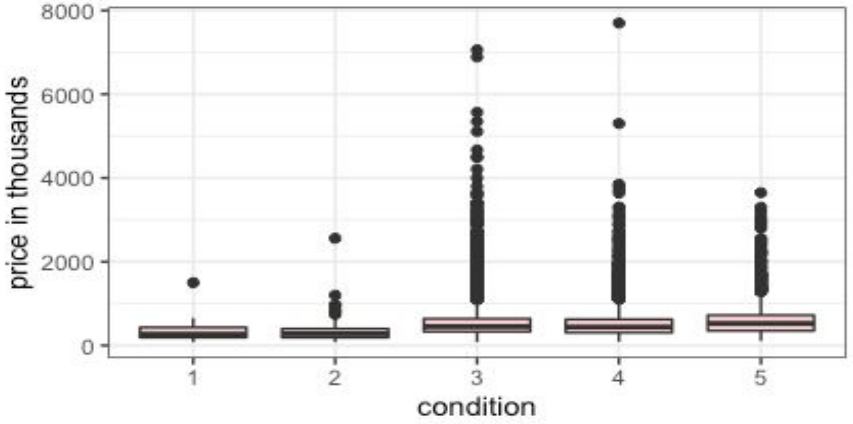


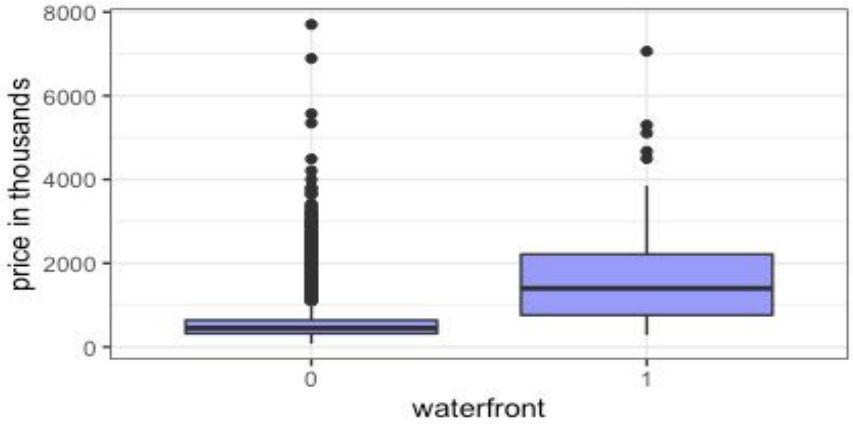


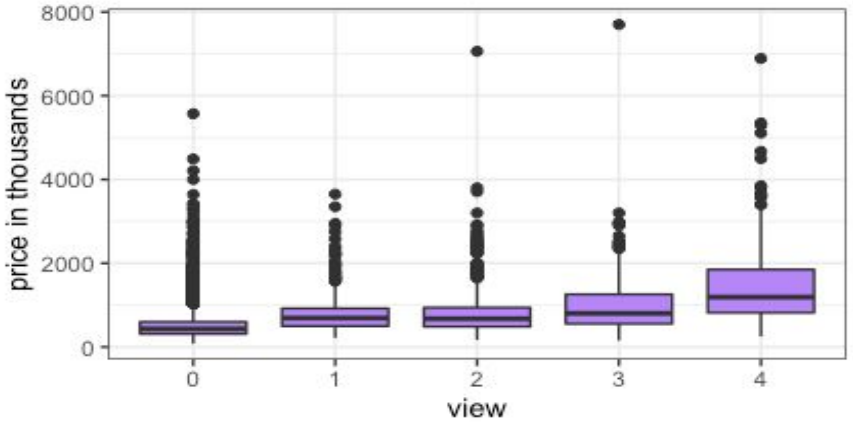


Data exploration

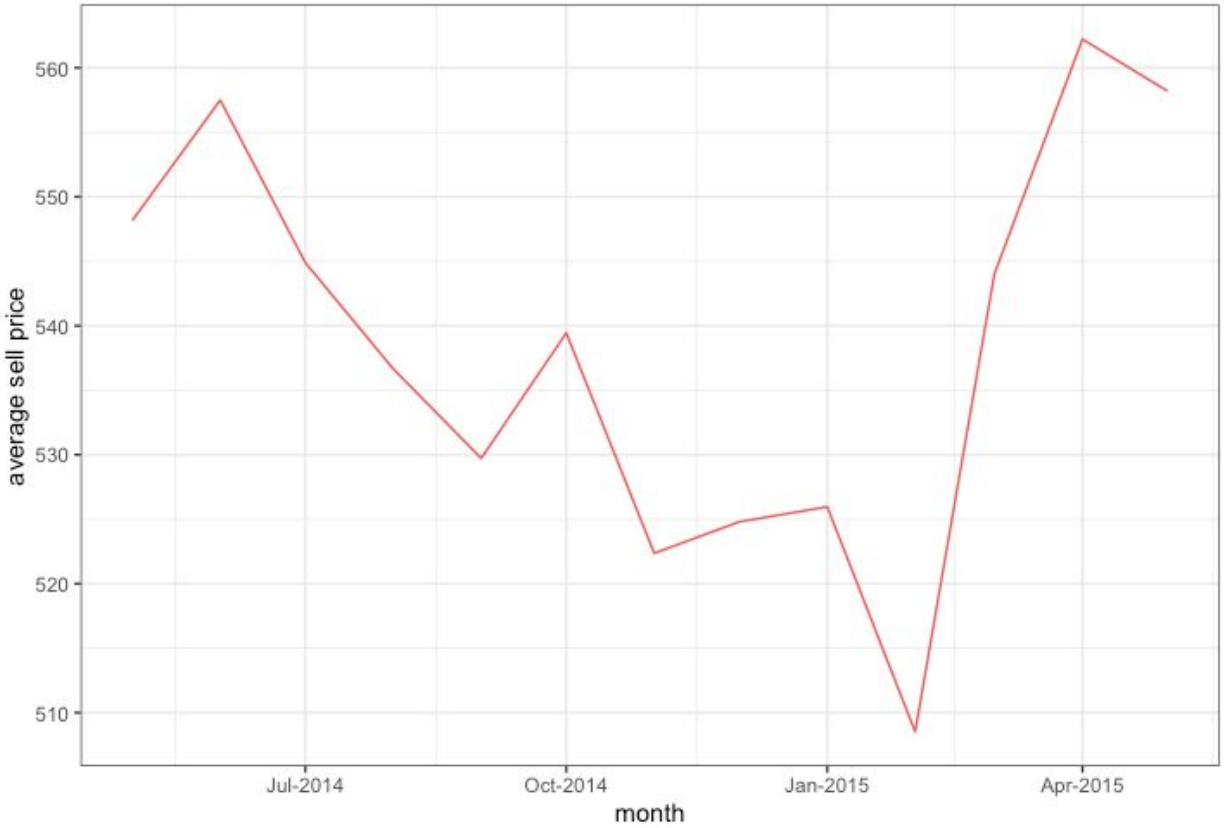




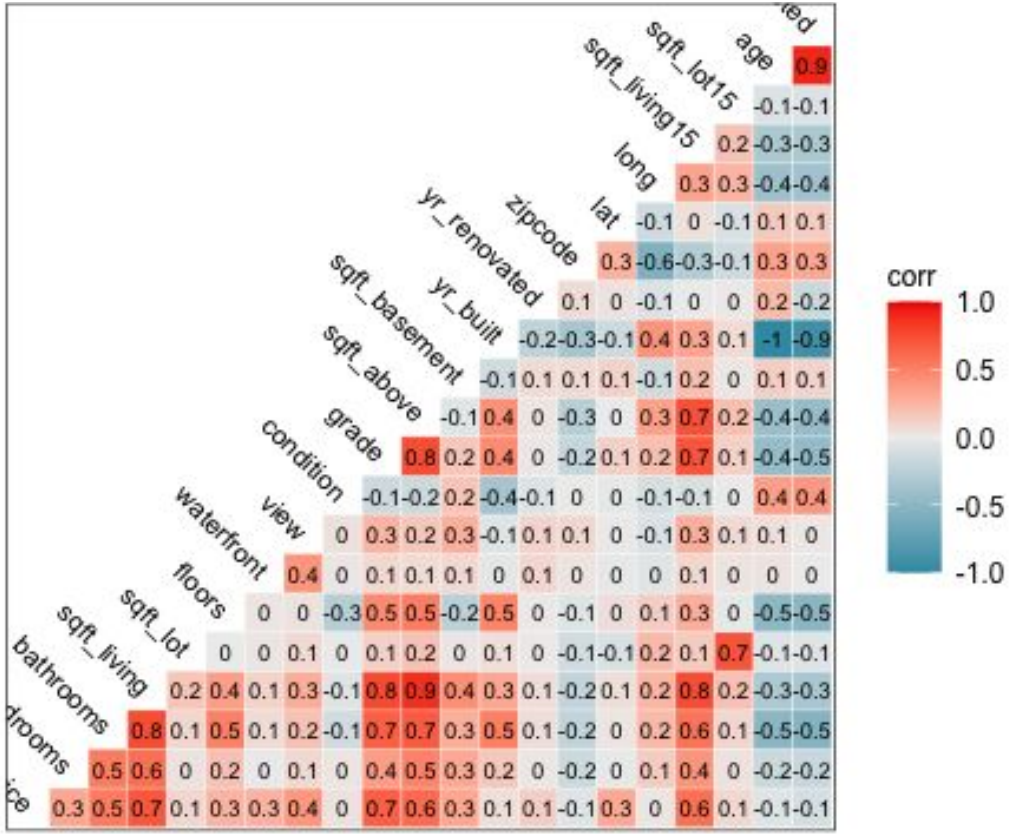




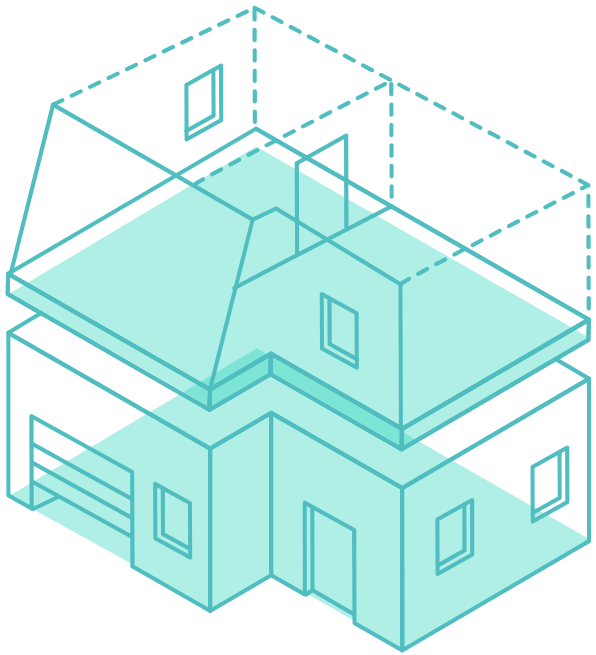
Data exploration - Average price per month



Data exploration - Correlation matrix



Model selection



Classification & Regression Trees

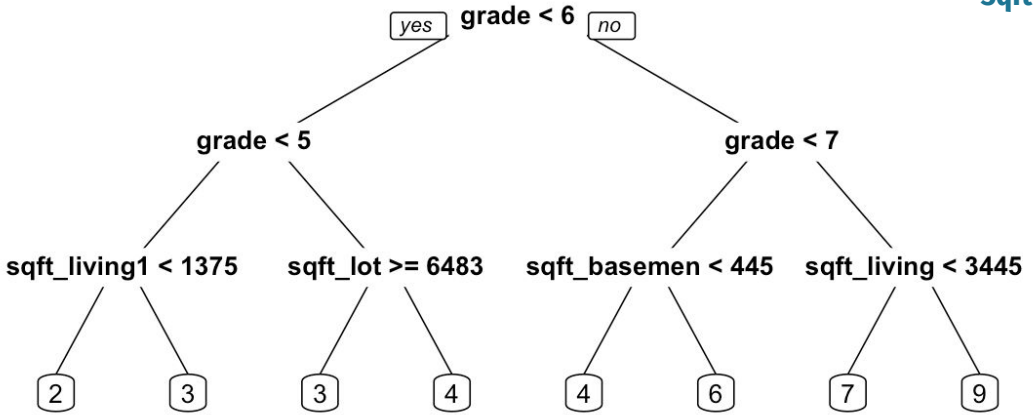
K Nearest Neighbor

K-means clustering

Multi Linear Regression

Model selection

Classification Trees

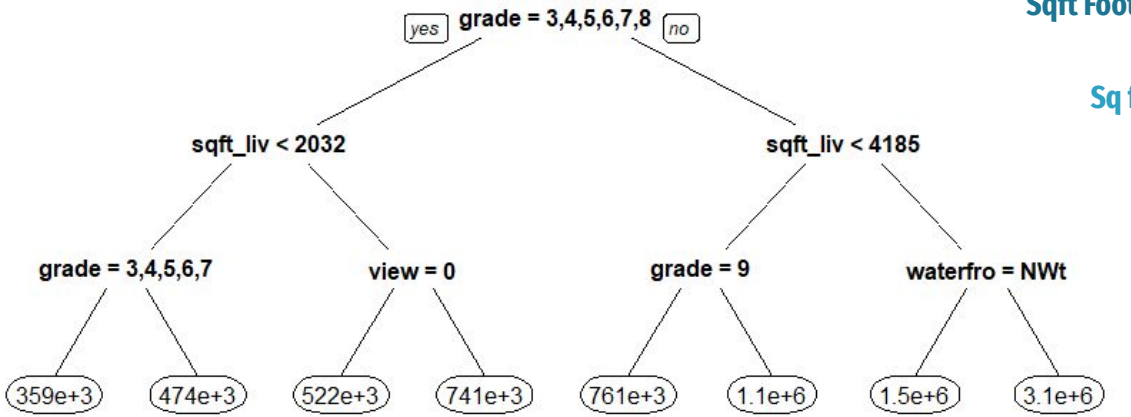


Accuracy: 4.15%

- **Bedrooms** 1
- **Bathrooms** 2
- **Sqft footage of interior space, above ground space, basement space, plot of land** 3
- **Sq footage of fifteen nearest neighbors interior and plots of land** 4
- **Waterfront** 5
- **View** 6
- **Floors** 7
- **Condition** 8

Model selection

Regression Trees



RMSE: 257242.2

1. **Bedrooms**
2. **Bathrooms**
3. **Sqft Footage of interior space, above ground space, basement space, plot of land**
4. **Sq footage of fifteen nearest neighbors interior and plots of land**
5. **Waterfront**
6. **View**
7. **Floors**
8. **Condition**

Model selection

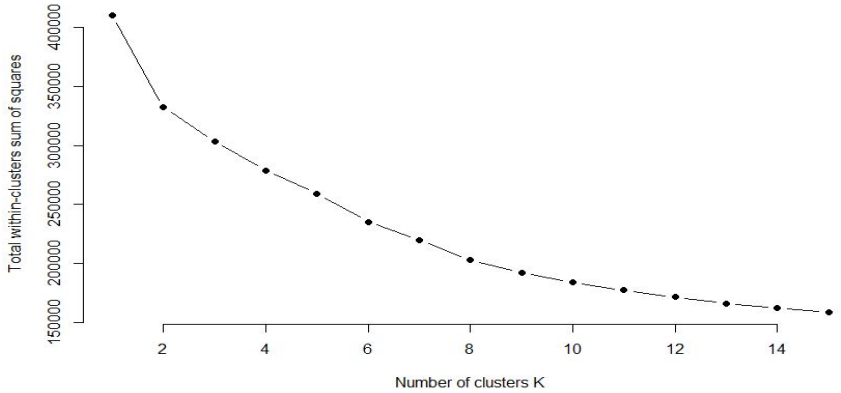
K Nearest Neighbor

- Divided price into bins, experimented to find most effective categories
- Optimal values after running the model are with price bins for $700000 and k nearest neighbors as 9. We obtain accuracy as 85.01% and sensitivity and specificity for the classes (0 and 1) having the majority of the observations is very good.
- Sensitivity for class 0 -> 0.8260, Specificity for class 0 -> 0.9332
- Sensitivity for class 1 -> 0.9142, Specificity for class 1 -> 0.7714

Price bins	K (number of nearby neighbors)	KNN result (Accuracy)
100,000	Tried with K between 1 to 20	Ranges from 40 to 45%
200,000	Tried with K between 1 to 20	Ranges from 60 to 65%
300,000	Tried with K between 1 to 20	Ranges from 70 to 75%
400,000	Tried with K between 1 to 20	Ranges from 75 to 80%
500,000	Tried with K between 1 to 20	Around 82 to 83 %
600,000	Tried with K between 1 to 20	Around 84%
700,000 and above	Tried with K between 1 to 20	Around 85%

Model selection

K-means clustering



	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
1	0.20856279	0.4263255	0.01251378	0.13917716	-0.19095431	-0.361330607	-0.08694538	0.3484534	0.8227312	-0.15678867	-0.32692211	0.8958645	-0.97378531	-0.20999497	1.0733598	0.470449	-0.8431643	-0.13341059	-0.22357464
2	0.20578496	-0.06404943	0.3489898	0.6760701	5.78310631	0.121541145	-0.08694538	0.3879655	-0.1486541	0.39239572	0.80498389	-0.1071354	0.4430092	-0.03700109	-0.6768876	-0.6408733	1.3806714	0.57405608	6.35973012
3	-0.1824159	0.03081552	0.49641389	0.04225745	-0.17365316	1.08056094	-0.08694538	-0.2238193	-0.4895375	0.2614059	0.33001391	-0.5278522	0.95019274	-0.20999497	-0.2485281	-0.1593497	0.2964195	0.07398076	-0.1888992
4	1.49695413	0.87218203	1.31214865	1.78720925	0.13752998	0.854082699	-0.08694538	0.4280801	-0.3116565	1.75721748	1.88173453	0.1756381	0.76403667	-0.15534869	-0.4009535	0.292316	0.5780647	1.71413038	0.1439053
5	-0.58594308	-0.38053857	-0.65653444	-0.67218518	-0.03990566	-0.77212593	-0.08694538	-0.2558022	0.3432528	-0.6191273	-0.53100708	-0.3953554	-0.06315534	-0.20999497	-0.8159992	-0.6081945	0.4471825	-0.52714042	-0.03331594
6	-0.45475782	-0.82744197	-1.09668561	-0.90669623	-0.20835078	-0.618025798	-0.08694538	-0.2084647	-0.1264397	-0.81631943	-0.81823046	-0.3432951	-0.91514472	-0.20999497	1.0515715	0.3071588	-0.8192894	-0.7703095	-0.23403772
7	3.02329793	-0.08940456	0.72062901	1.15940091	0.26604209	0.268821804	11.50094132	4.6318338	0.195647	0.94147446	0.79604182	0.9052352	-0.29521821	1.04345452	0.3576015	-0.1688904	-0.4833653	0.98419614	0.37635255
8	0.35702773	0.06508067	0.14280814	0.11584246	-0.06101796	0.000292033	-0.08694538	0.2595495	-0.3032229	-0.04161412	-0.01433002	0.2653476	-1.10623562	4.76196016	0.3470289	0.1653127	-0.3382354	-0.15079996	-0.09727142
9	0.07421582	0.71376077	0.33896817	0.43199305	0.04990687	-0.828609708	-0.08694538	0.0318674	0.4061868	0.0912354	-0.26220044	1.3784997	0.05669488	-0.20999497	-0.7470278	0.1172167	0.2617973	0.36300405	0.06970867

Model selection

K-means clustering

- Of the 9 clusters, cluster 7 was the most expensive with a normalized price of 3.02 and positive values throughout the selected variables. This cluster most likely corresponds to expensive homes with multiple bedrooms and ample living space. Cluster 4 had a price value of 1.49 and positive values throughout.
- Three clusters (3,5, & 6) all had negative normalized values for price and a preponderance of negative values throughout. These clusters probably correspond to lower income housing and apartments.
- Four clusters have positive normalized values for price but are significantly less than cluster 7. Clusters 1, 2, 8, and 9 have price values ranging from 0.081 to 0.46. These clusters probably represent more middle-income housing environments.

Model selection

Multi Linear Regression

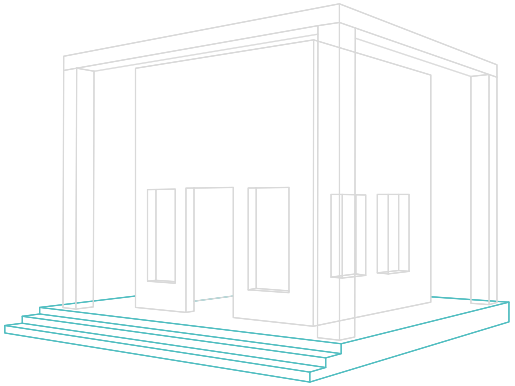
Model Number	Adjusted R-squared	RMSE	Predictors	Notes
MLR Model 1	0.7033	204826.8	All predictors	Multicollinearity +singularities
MLR Model 2	**0.7033**	**204826.8**	**Removed sqft_basement**	**Best performing; same outcomes but no errors; applied to K-clusters**
MLR Model 3	0.6615	218882.5	Predictors with correlation coefficient >0.3	Multicollinearity
MLR Model 4	0.6124	228963.8	Predictors with correlation coefficient >0.5	No errors but worse outcomes

Insights

- **Strong predictors**
  - **Bedrooms & Bathrooms**
  - **Sqrt_living & Floors**
  - **Waterfront area**
  - **Grade & View**

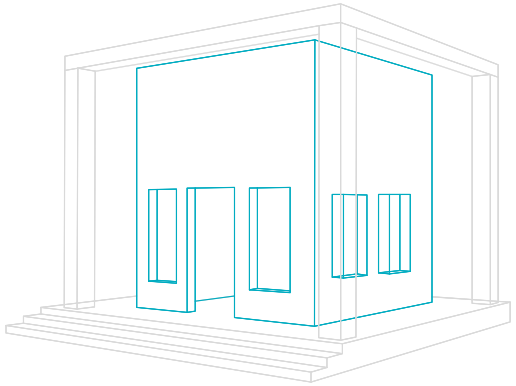
- **Locations**
  - **98039 zip code has a median sale price of $ 2.16 million whereas 98002 has a median sale price of $ 234.28 K**
  - **Clustering analysis: among 9 clusters, we observe very strong correlation within two clusters, corresponding to the most expensive cluster and the least expensive cluster (R squares are well above 0.80)**
  - **Model has highest accuracy when predicting extreme values**

Conclusions



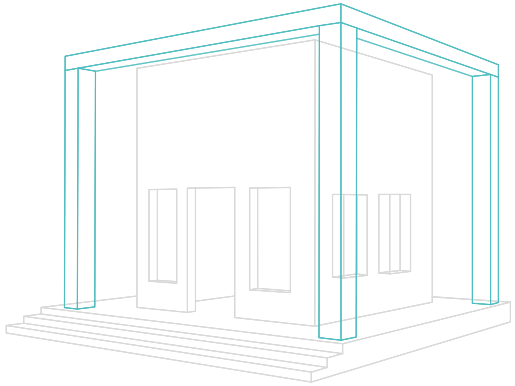
Buyers

- Evaluate the individual listing prices based on the property factors.
- Get crucial insights regarding the asking price, future appreciation of the property.



Real estate agents

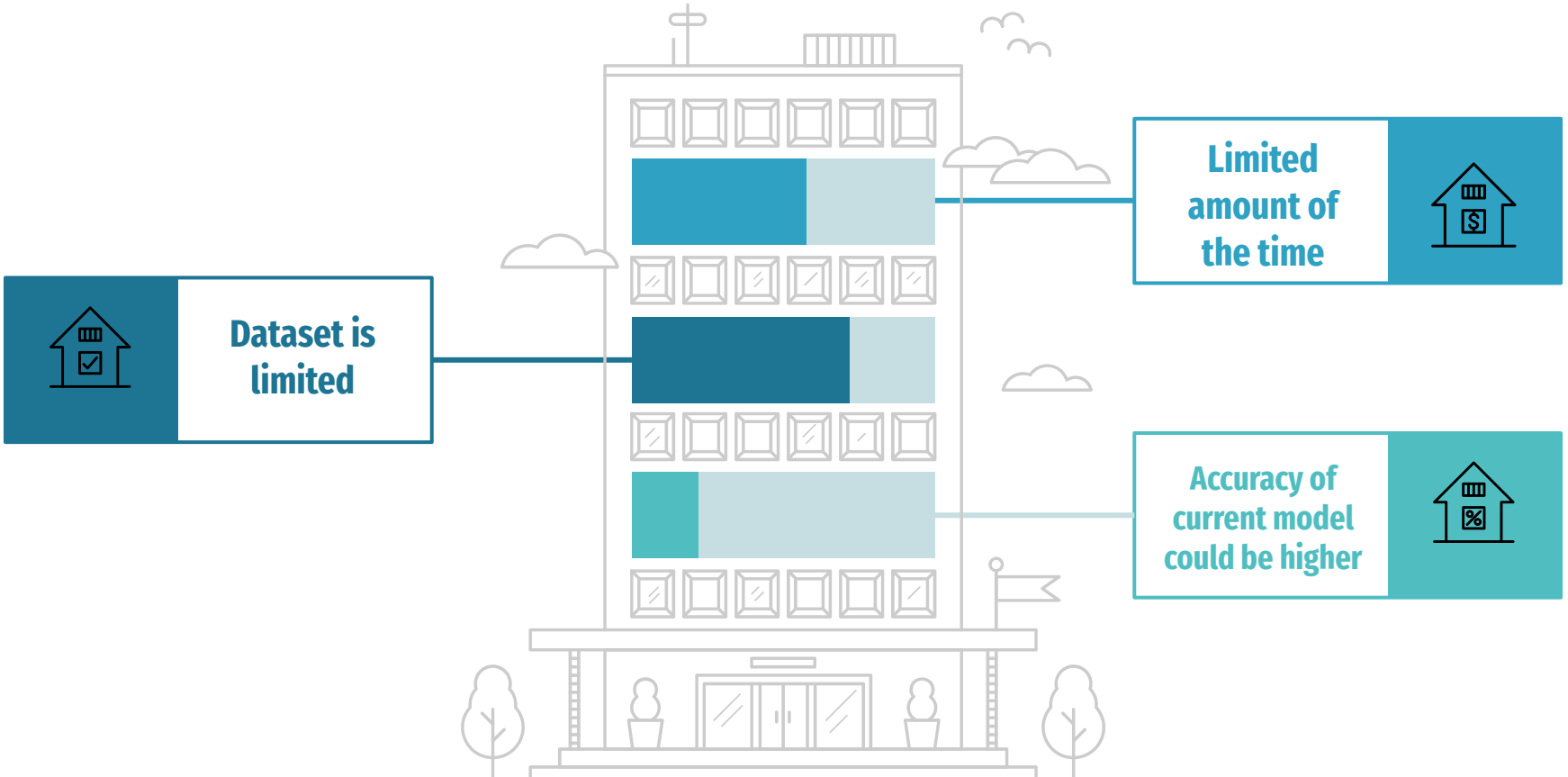
- Visualization and model prediction
- Understand the pricing trends in the King County



Real estate companies

- Better allocate resources based on the current pricing trends.

Challenges

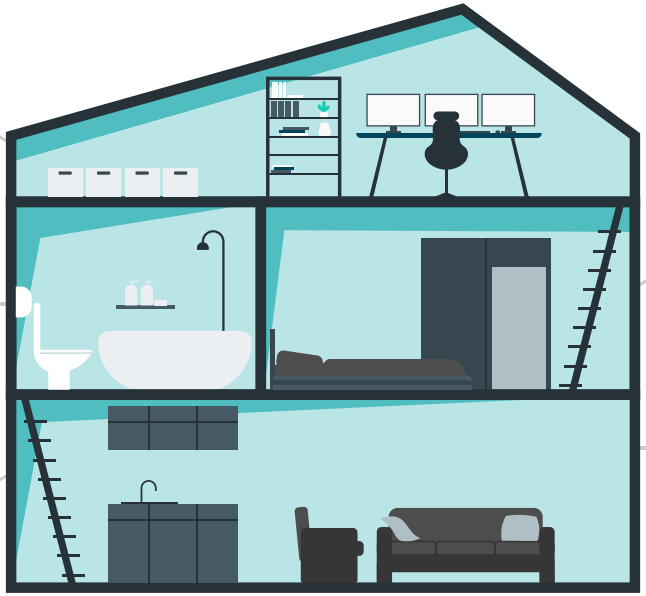

Dataset is limited
Limited amount of the time
Accuracy of current model could be higher

Recommendations for improvement

Better understand the greater Seattle area housing market

Re-training our model on a periodic basis

Further K means analysis & logistical regression



Add more crucial factors

Test the seasonality by isolating the month as a variable

Thank you!