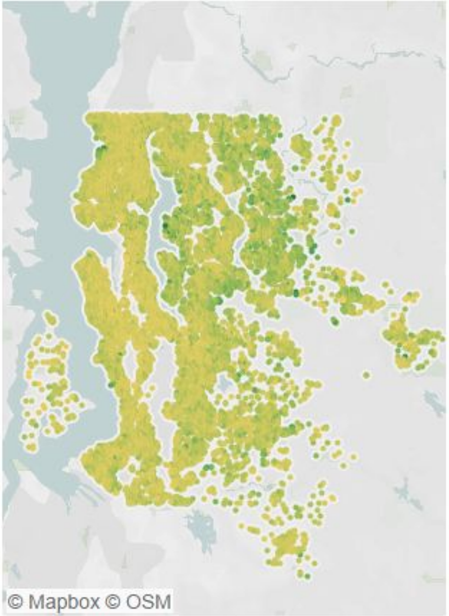
IRON REGRESSION

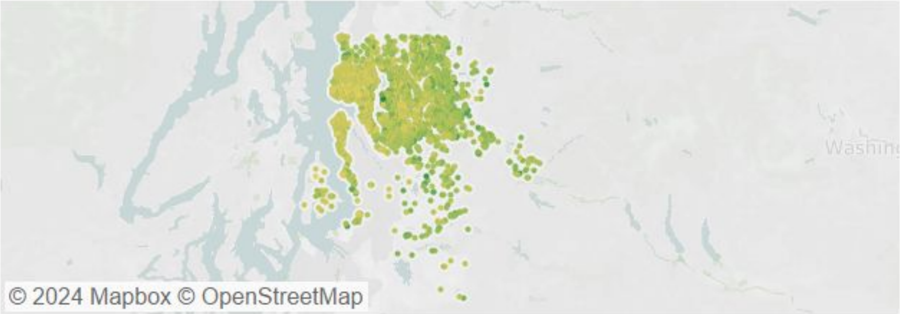
Paula, Enrique, Yasmine, Rubén

House Prices

Square Footage of the Houses



Square Footage of the Houses (Price>$647K)



Houses with a Waterfront View

House Prices

Square Footage of the Houses

Some steps for EDA

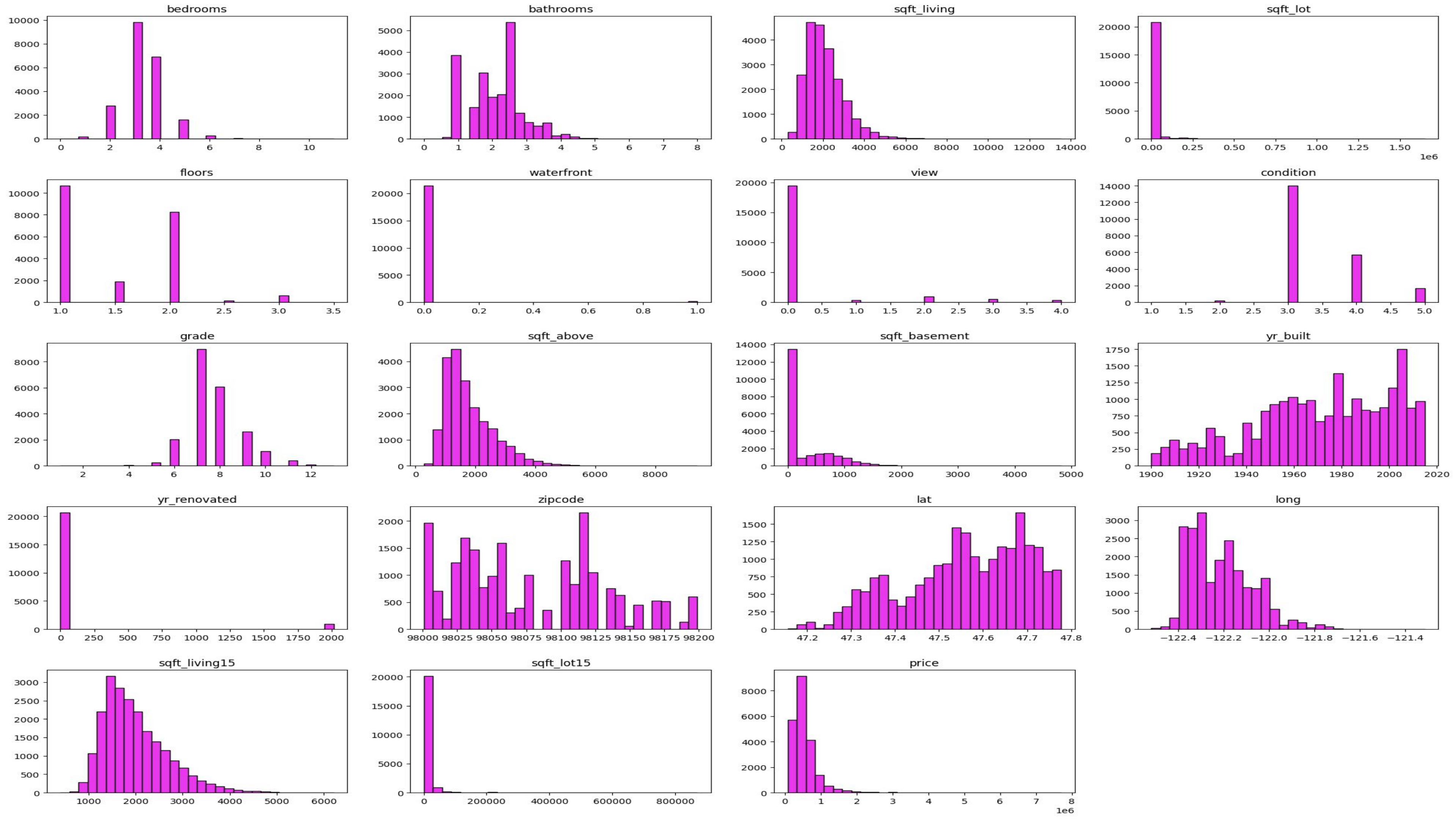
- Libraries imported: Pandas, Numpy, Matplotlyb, Seaborn
- Cleaning
- Descriptive statistics

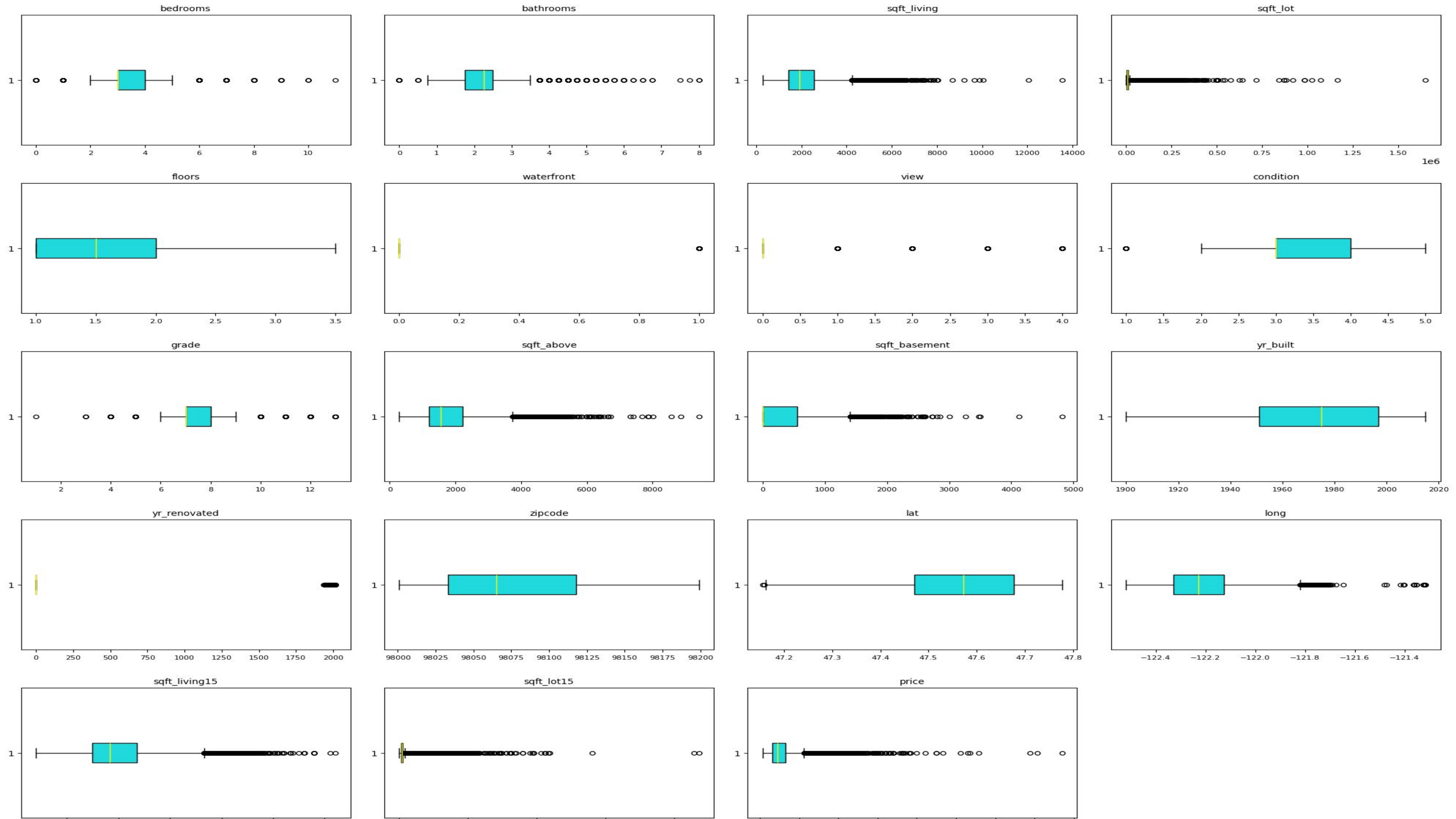
	bedrooms	floors	waterfront	view	condition	grade
id						
7.129301e+09	3.0	1.0	0.0	0.0	3.0	7.0
6.414100e+09	3.0	2.0	0.0	0.0	3.0	7.0
5.631500e+09	2.0	1.0	0.0	0.0	3.0	6.0
2.487201e+09	4.0	1.0	0.0	0.0	5.0	7.0
1.954401e+09	3.0	1.0	0.0	0.0	3.0	8.0
...	...	...	...	...	...	...
6.600060e+09	4.0	2.0	0.0	0.0	3.0	8.0
1.523300e+09	2.0	2.0	0.0	0.0	3.0	7.0
2.913101e+08	3.0	2.0	0.0	0.0	3.0	8.0
1.523300e+09	2.0	2.0	0.0	0.0	3.0	7.0

sqft_total
2.161300e+04
1.718687e+04
1.423000e+03
7.035000e+03
9.575000e+03
1.300000e+04
1.652659e+06
4.158908e+04

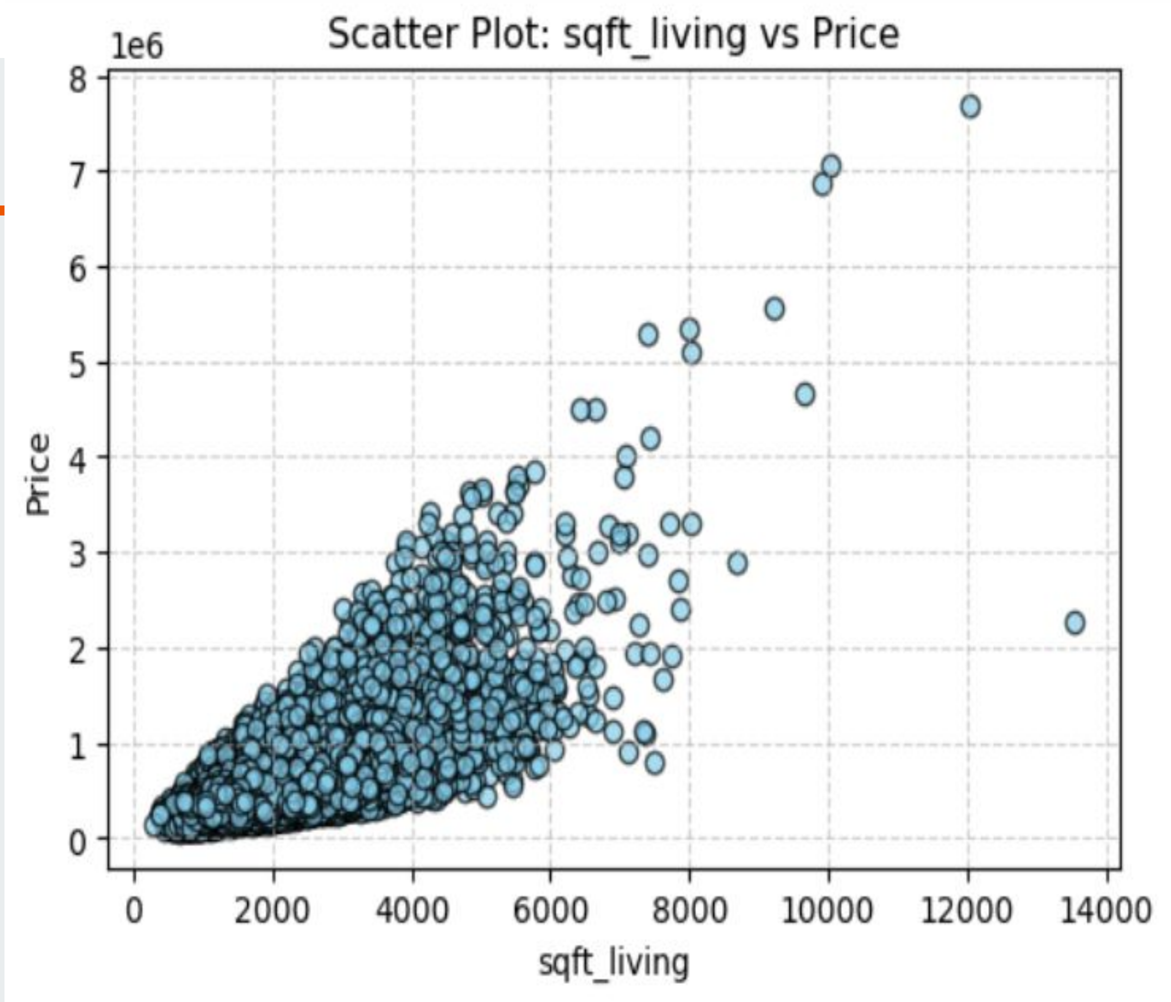
sqft_total	lat_area	long_area
6830.0	north area	west
9812.0	north area	west
10770.0	north area	west
6960.0	north area	west
9760.0	north area	west

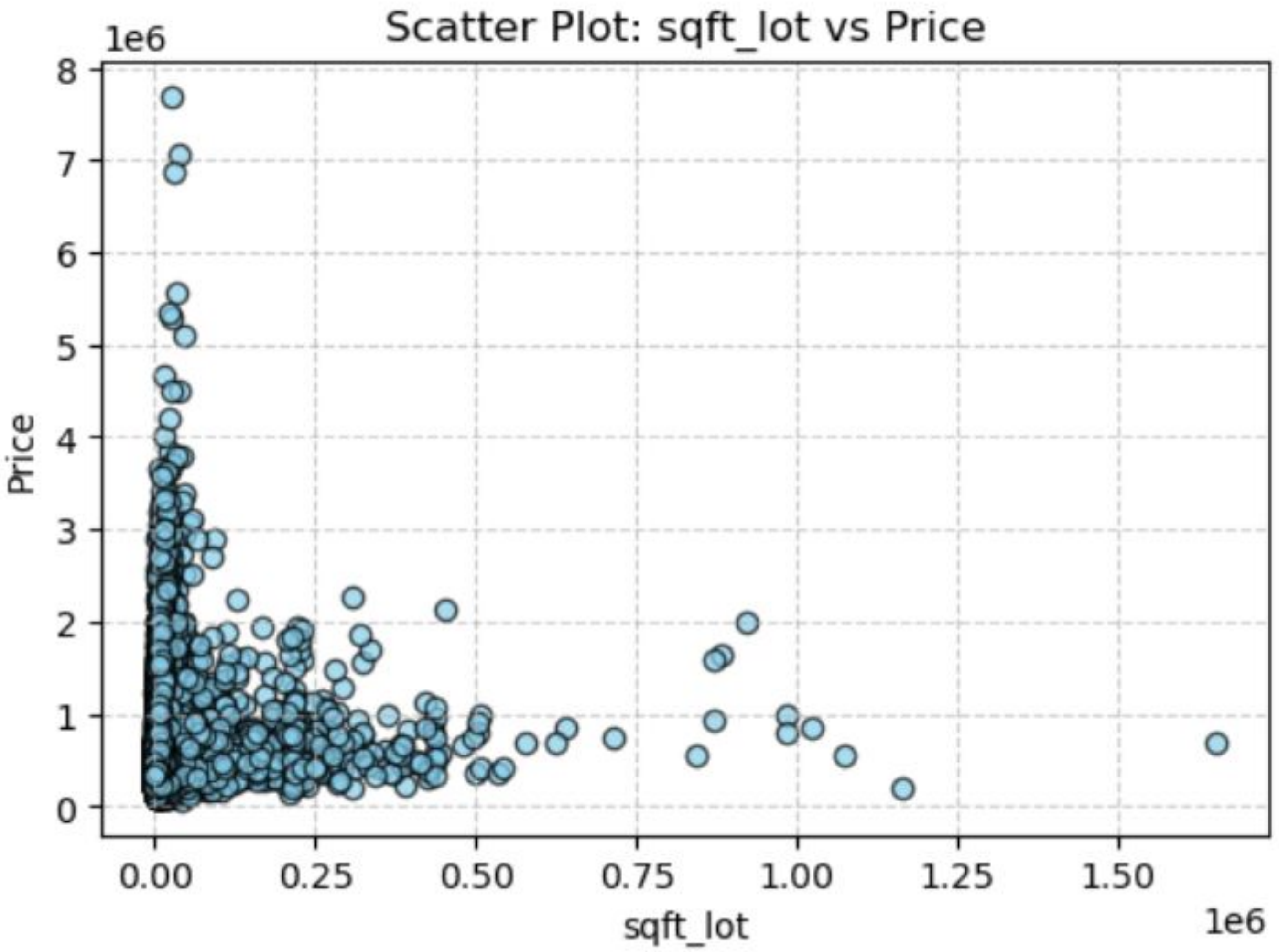
	count	mean	min	25%	50%	75%	max	std
date	21613	2014-10-29 04:38:01.959931648	2014-05-02 00:00:00	2014-07-22 00:00:00	2014-10-16 00:00:00	2015-02-17 00:00:00	2015-05-27 00:00:00	NaN
bedrooms	21613.00	3.37	0.00	3.00	3.00	4.00	33.00	0.93
bathrooms	21613.00	2.11	0.00	1.75	2.25	2.50	8.00	0.77
sqft_living	21613.00	2079.90	290.00	1427.00	1910.00	2550.00	13540.00	918.44
sqft_lot	21613.00	15106.97	520.00	5040.00	7618.00	10688.00	1651359.00	41420.51
floors	21613.00	1.49	1.00	1.00	1.50	2.00	3.50	0.54
waterfront	21613.00	0.01	0.00	0.00	0.00	0.00	1.00	0.09
view	21613.00	0.23	0.00	0.00	0.00	0.00	4.00	0.77
condition	21613.00	3.41	1.00	3.00	3.00	4.00	5.00	0.65
grade	21613.00	7.66	1.00	7.00	7.00	8.00	13.00	1.18
sqft_above	21613.00	1788.39	290.00	1190.00	1560.00	2210.00	9410.00	828.09
sqft_basement	21613.00	291.51	0.00	0.00	0.00	560.00	4820.00	442.58
yr_built	21613.00	1971.01	1900.00	1951.00	1975.00	1997.00	2015.00	29.37
yr_renovated	21613.00	84.40	0.00	0.00	0.00	0.00	2015.00	401.68
zipcode	21613.00	98077.94	98001.00	98033.00	98065.00	98118.00	98199.00	53.51
lat	21613.00	47.56	47.16	47.47	47.57	47.68	47.78	0.14
long	21613.00	-122.21	-122.52	-122.33	-122.23	-122.12	-121.31	0.14
sqft_living15	21613.00	1986.55	399.00	1490.00	1840.00	2360.00	6210.00	685.39
sqft_lot15	21613.00	12768.46	651.00	5100.00	7620.00	10083.00	871200.00	27304.18
price	21613.00	540088.14	75000.00	321950.00	450000.00	645000.00	7700000.00	367127.20

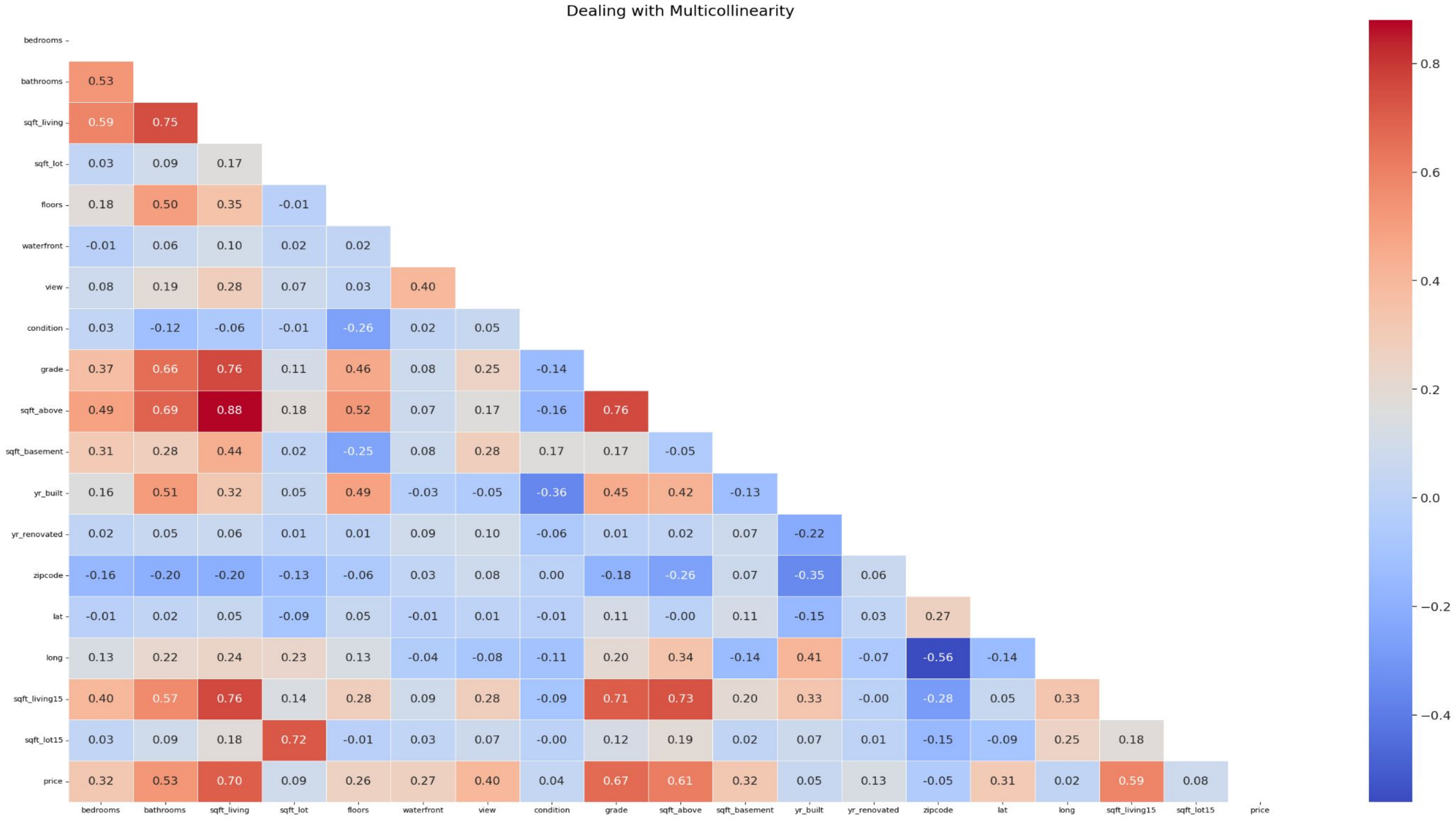

bedrooms
0 2 4 6 8 10
0 2000 4000 6000 8000 10000
bathrooms
0 1 2 3 4 5 6 7 8
0 1000 2000 3000 4000 5000
sqft_living
0 2000 4000 6000 8000 10000 12000 14000
0 1000 2000 3000 4000
sqft_lot
0.00 0.25 0.50 0.75 1.00 1.25 1.50
1e6
0 5000 10000 15000 20000
floors
1.0 1.5 2.0 2.5 3.0 3.5
0 2000 4000 6000 8000 10000
waterfront
0.0 0.2 0.4 0.6 0.8 1.0
0 5000 10000 15000 20000
view
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
0 5000 10000 15000 20000
condition
1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
0 2000 4000 6000 8000 10000 12000 14000
grade
2 4 6 8 10 12
0 2000 4000 6000 8000
sqft_above
0 2000 4000 6000 8000
0 1000 2000 3000 4000
sqft_basement
0 1000 2000 3000 4000 5000
0 2000 4000 6000 8000 10000 12000 14000
yr_built
1900 1920 1940 1960 1980 2000 2020
0 250 500 750 1000 1250 1500 1750
yr_renovated
0 250 500 750 1000 1250 1500 1750 2000
0 5000 10000 15000 20000
zipcode
98000 98025 98050 98075 98100 98125 98150 98175 98200
0 500 1000 1500 2000
lat
47.2 47.3 47.4 47.5 47.6 47.7 47.8
0 250 500 750 1000 1250 1500
long
−122.4 −122.2 −122.0 −121.8 −121.6 −121.4
0 500 1000 1500 2000 2500 3000
sqft_living15
1000 2000 3000 4000 5000 6000
0 500 1000 1500 2000 2500 3000
sqft_lot15
0 200000 400000 600000 800000
0 5000 10000 15000 20000
price
0 1 2 3 4 5 6 7 8
1e6
0 2000 4000 6000 8000


bedrooms
0 2 4 6 8 10
bathrooms
0 1 2 3 4 5 6 7 8
sqft_living
0 2000 4000 6000 8000 10000 12000 14000
sqft_lot
0.00 0.25 0.50 0.75 1.00 1.25 1.50
1e6
floors
1.0 1.5 2.0 2.5 3.0 3.5
waterfront
0.0 0.2 0.4 0.6 0.8 1.0
view
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
condition
1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
grade
2 4 6 8 10 12
sqft_above
0 2000 4000 6000 8000
sqft_basement
0 1000 2000 3000 4000 5000
yr_built
1900 1920 1940 1960 1980 2000 2020
yr_renovated
0 250 500 750 1000 1250 1500 1750 2000
zipcode
98000 98025 98050 98075 98100 98125 98150 98175 98200
lat
47.2 47.3 47.4 47.5 47.6 47.7 47.8
long
−122.4 −122.2 −122.0 −121.8 −121.6 −121.4
sqft_living15
sqft_lot15
price
1

```
price            1.000000
sqft_living      0.702035
grade            0.667434
sqft_above       0.605567
sqft_living15    0.585379
bathrooms        0.525138
view             0.397293
sqft_basement    0.323816
bedrooms         0.308350
lat              0.307003
waterfront       0.266369
floors           0.256794
yr_renovated     0.126434
sqft_lot         0.089661
sqft_lot15       0.082447
yr_built         0.054012
condition        0.036362
long             0.021626
zipcode         -0.053203
dtype: float64
```


Scatter Plot: sqft_living vs Price
1e6
8
7
6
5
4
3
2
1
0
Price
0
2000
4000
6000
8000
10000
12000
14000
sqft_living

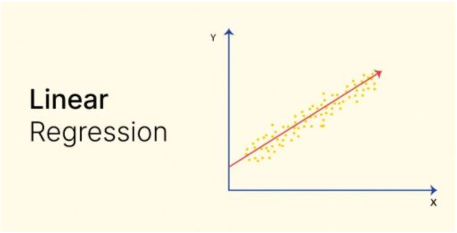

Scatter Plot: sqft_lot vs Price
1e6
8
7
6
5
4
3
2
1
0
Price
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1e6
sqft_lot


Dealing with Multicollinearity
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
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Choosing Models for Implementation

We tested different models such as:

- **Linear Regression**: For simple linear relationships.
- **Ridge**: For data with multicollinearity.
- **XGBoost**: For complex data and large volumes.
- **Random Forest**: For data with outliers.

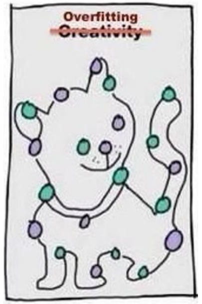


First Model

Bedrooms	Bathrooms	Sqft Living	Sqft Lot	Floors	Waterfront	View	Condition	Grade
Number of bedrooms	Number of bathrooms	Interior size (sqft)	Land size (sqft)	Number of floors	Waterfront view (yes/no)	Number of views	House condition	House grade

Sqft Above	Sqft Basement	Year Built	Year Renovated	Zipcode	Lat/Long	Sqft Living 15	Sqft Lot 15	Price
Above-ground size (sqft)	Basement size (sqft)	Year built	Year renovated (if any)	Postal code	Coordinates	Neighbors' interior size (sqft)	Neighbors' land size (sqft)	Sale price

- **Total**: 19 Columns.

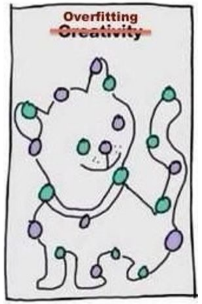


First Model

Bedrooms	Bathrooms	Sqft Living	Sqft Lot	Floors	Waterfront	View	Condition	Grade
Number of bedrooms	Number of bathrooms	Interior size (sqft)	Land size (sqft)	Number of floors	Waterfront view (yes/no)	Number of views	House condition	House grade

Sqft Above	Sqft Basement	Year Built	Year Renovated	Zipcode	Lat/Long	Sqft Living 15	Sqft Lot 15	Price
Above-ground size (sqft)	Basement size (sqft)	Year built	Year renovated (if any)	Postal code	Coordinates	Neighbors' interior size (sqft)	Neighbors' land size (sqft)	Sale price

- **Total**: 10 Columns.

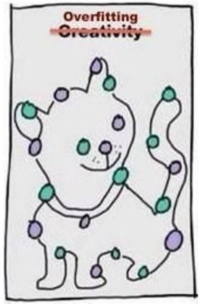


First Model

Bedrooms	Bathrooms	Sqft Living	Sqft Lot	Floors	Waterfront	View	Condition	Grade
Number of bedrooms	Number of bathrooms	Interior size (sqft)	Land size (sqft)	Number of floors	Waterfront view (yes/no)	Number of views	House condition	House grade

Sqft Above	Sqft Basement	Year Built	Year Renovated	Zipcode	Lat/Long	Sqft Living 15	Sqft Lot 15	Price
Above-ground size (sqft)	Basement size (sqft)	Year built	Year renovated (if any)	Postal code	Coordinates	Neighbors' interior size (sqft)	Neighbors' land size (sqft)	Sale price

- **Total**: 19 Columns.



EVALUATION RESULTS, Feature Set 1

- **R^2 :** Coefficient of Determination
- **RMSE :** Mean Squared Error
- **MSE :** Root Mean Squared Error
- **MAE :** Mean Absolute Error

Model	R^2	RMSE	MSE	MAE
Linear Regression	0.691	196918.261	38776801706.383	124866.389
Ridge	0.691	196916.746	38776204959.441	124841.919
XGBoost	0.876	124580.981	15520420795.834	67180.355
Random Forest	0.869	128456.214	16500998930.655	67568.693

EVALUATION RESULTS, Feature Set 1

- **R^2 :** Coefficient of Determination
- **RMSE :** Mean Squared Error
- **MSE :** Root Mean Squared Error
- **MAE :** Mean Absolute Error

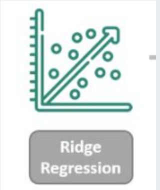
Model	R^2	RMSE	MSE	MAE
Linear Regression	0.691	196918.261	38776801706.383	124866.389
Ridge	0.691	196916.746	38776204959.441	124841.919
XGBoost	0.876	124580.981	15520420795.834	67180.355
Random Forest	0.869	128456.214	16500998930.655	67568.693

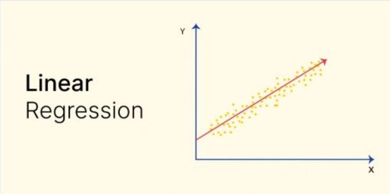
EVALUATION RESULTS, Feature Set 1

- **R^2 :** Coefficient of Determination
- **RMSE :** Mean Squared Error
- **MSE :** Root Mean Squared Error
- **MAE :** Mean Absolute Error

Model	R²	RMSE	MSE	MAE
Linear Regression	0.691	196918.261	38776801706.383	124866.389
Ridge	0.691	196916.746	38776204959.441	124841.919
XGBoost	0.876	124580.981	15520420795.834	67180.355
Random Forest	0.869	128456.214	16500998930.655	67568.693

ANALYSIS OF RESULTS, Feature Set 1





Linear Regression and Ridge:

These models are simple, explain the data well (R^2 of 0.7), and don't focus too much on small details in the data.

Model	R²	RMSE	MSE	MAE
Linear Regression	0.691	196918.261	38776801706.383	124866.389
Ridge	0.691	196916.746	38776204959.441	124841.919

- **R^2 :** Coefficient of Determination
- **RMSE :** Root Mean Squared Error
- **MSE :** Mean Squared Error
- **MAE :** Mean Absolute Error

ANALYSIS OF RESULTS, Feature Set 1

XGBoost and Random Forest showed:

The best performance with high R^2 and low RMSE, but they are more likely to overfit, because we used 19 columns, meaning they may work well on the training data but not on new data

Model	R^2	RMSE	MSE	MAE
XGBoost	0.876	124580.981	15520420795.834	67180.355
Random Forest	0.869	128456.214	16500998930.655	67568.693

- **R^2 :** Coefficient of Determination
- **RMSE :** Root Mean Squared Error
- **MSE :** Mean Squared Error
- **MAE :** Mean Absolute Error

Creating New Columns

Transform sqft_basement:

We will create a binary where::

1 Indicates the presence of a basement.
0 Indicates no basement.

New column for the house's age:

New column, year_house, that calculates the age of the house by subtracting yr_built from the current year.

This will allow us to track the age of the house at any given time using the system date.

Model, Feature Set 5

Feature	(1) Bedrooms	(2) Bathrooms	(3) Sqft Living	(4) Waterfront	(5) Grade	(6) Sqft Basement	(7) Zipcode	(8) Year House	(9) Price
Description	Number of bedrooms	Number of bathrooms	Interior size (sqft)	(1 = yes, 0 = no)	House grade	Basement (1 = yes, 0 = no)	Postal code	Age of the house (in years)	Sale price

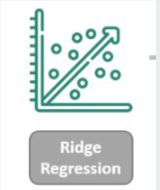
Total: 9 Columns.

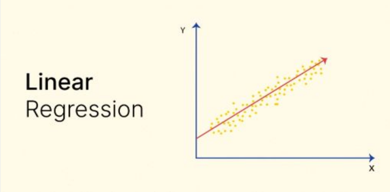
EVALUATION RESULTS, Feature Set 5

- **R^2 :** Coefficient of Determination
- **RMSE :** Mean Squared Error
- **MSE :** Root Mean Squared Error
- **MAE :** Mean Absolute Error

Model	R^2	RMSE	MSE	MAE
Linear Regression	0.645	231080.192	53398055024.996	143349.439
Ridge	0.645	231111.259	53412413882.207	143355.879
XGBoost	0.845	152817.093	23353064045.665	81071.314
Random Forest	0.830	159672.300	25495243316.720	89617.403

ANALYSIS OF RESULTS, Feature Set 5





Linear Regression and Ridge:

R^2: 0.645 (model has a weaker fit to the data). than the previous model (R^2 = 0.7).

Errors (RMSE, MSE, MAE): Increased significantly compared to the previous model with 19 Columns.

(Worse results)

Model	R²	RMSE	MSE	MAE
Linear Regression	0.645	231080.192	53398055024.996	143349.439
Ridge	0.645	231111.259	53412413882.207	143355.879
Linear Regression	0.691	196918.261	38776801706.383	124866.389
Ridge	0.691	196916.746	38776204959.441	124841.919

Previous Model

- **R^2 :** Coefficient of Determination
- **RMSE :** Root Mean Squared Error
- **MSE :** Mean Squared Error
- **MAE :** Mean Absolute Error

ANALYSIS OF RESULTS, Feature Set 5

XGBoost and Random Forest

XGBoost: R^2 = 0.845, RMSE = 152817.093 (best performance).

Random Forest: R^2 = 0.83, RMSE = 159672.3 (second-best performance).

(Better results)

Previous Model

Model	R^2	RMSE	MSE	MAE
XGBoost	0.876	124580.981	15520420795.834	67180.355
Random Forest	0.869	128456.214	16500998930.655	67568.693
XGBoost	0.845	152817.093	23353064045.665	81071.314
Random Forest	0.830	159672.300	25495243316.720	89617.403

- **R^2 :** Coefficient of Determination
- **RMSE :** Root Mean Squared Error
- **MSE :** Mean Squared Error
- **MAE :** Mean Absolute Error

CONCLUSIONS

Conclusion: "Even with fewer features, **XGBoost** and **Random Forest** perform much better than the linear models,

XGBoost being the most precise results."

Bedrooms	Bathrooms	Sqft Living	Sqft Lot	Floors	Waterfront	View	Condition	Grade
Number of bedrooms	Number of bathrooms	Interior size (sqft)	Land size (sqft)	Number of floors	Waterfront view (yes/no)	Number of views	House condition	House grade

Sqft Above	Sqft Basement	Year Built	Year Renovated	Zipcode	Lat/Long	Sqft Living 15	Sqft Lot 15	Price
Above-ground size (sqft)	Basement size (sqft)	Year built	Year renovated (if any)	Postal code	Coordinates	Neighbors' interior size (sqft)	Neighbors' land size (sqft)	Sale price

Total: 19 Columns.

Feature	(1) Bedrooms	(2) Bathrooms	(3) Sqft Living	(4) Waterfront	(5) Grade	(6) Sqft Basement	(7) Zipcode	(8) Year House	(9) Price
Description	Number of bedrooms	Number of bathrooms	Interior size (sqft)	(1 = yes, 0 = no)	House grade	Basement (1 = yes, 0 = no)	Postal code	Age of the house (in years)	Sale price

Total: 8 Columns.

Steps (Trying to) Improve the Models

Different sets of features:

1. Only **high correlation** with target
2. AND **no multicollinearity**
3. AND **only continuous**

⇨ BUT: **No improvements**, only deterioration in all models

Steps (Trying to) Improve the Models

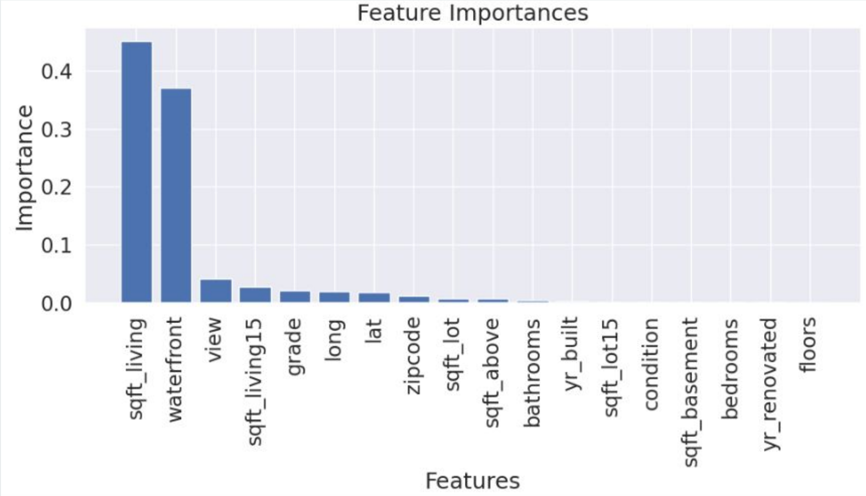
Scaling Train through Normalization:

- Slightly better metrics and **best model so far**:

```
XGBoost with Feature Set 1
 and Scaled
R2 =  0.87
RMSE =  136676.5
MSE =  18680465894.55
MAE =  71714.48
```

Steps (Trying to) Improve the Models

- **Oversampling** “waterfront”
- **Feature Importance Check**: Two more sets of features



Steps (Trying to) Improve the Models

Going back to the start:

- **Removing outliers** from all features before modelling

Much **improved metrics** (MSE, MAE and RSME):

```
XGBoost with
no Outliers in Train
R2 = 0.87
RMSE = 72,731.09
MSE = 5,289,810,875.94
MAE = 48,567.98
```

Summary & High Priced Houses

- Best model **without outliers** but is the **price too high**?

- Side show: **high priced houses**

Challenges and Future Analysis

Challenges

- Overfitting ?!?

Future Analysis

- Location: Geocoding?
- Some houses were sold more than once = more EDA might be insightful
- Modifying the target

THANK YOU!