King County

Price Prediction

Business Problem

Horizon Home (HH) - Real Estate Agency

- Build a model which will help HH to make informed investment decisions
- Make data driven conclusion about housing prices

METHODOLOGY

OBTAIN & ANALYSE DATASET

INVESTIGATE FEATURES

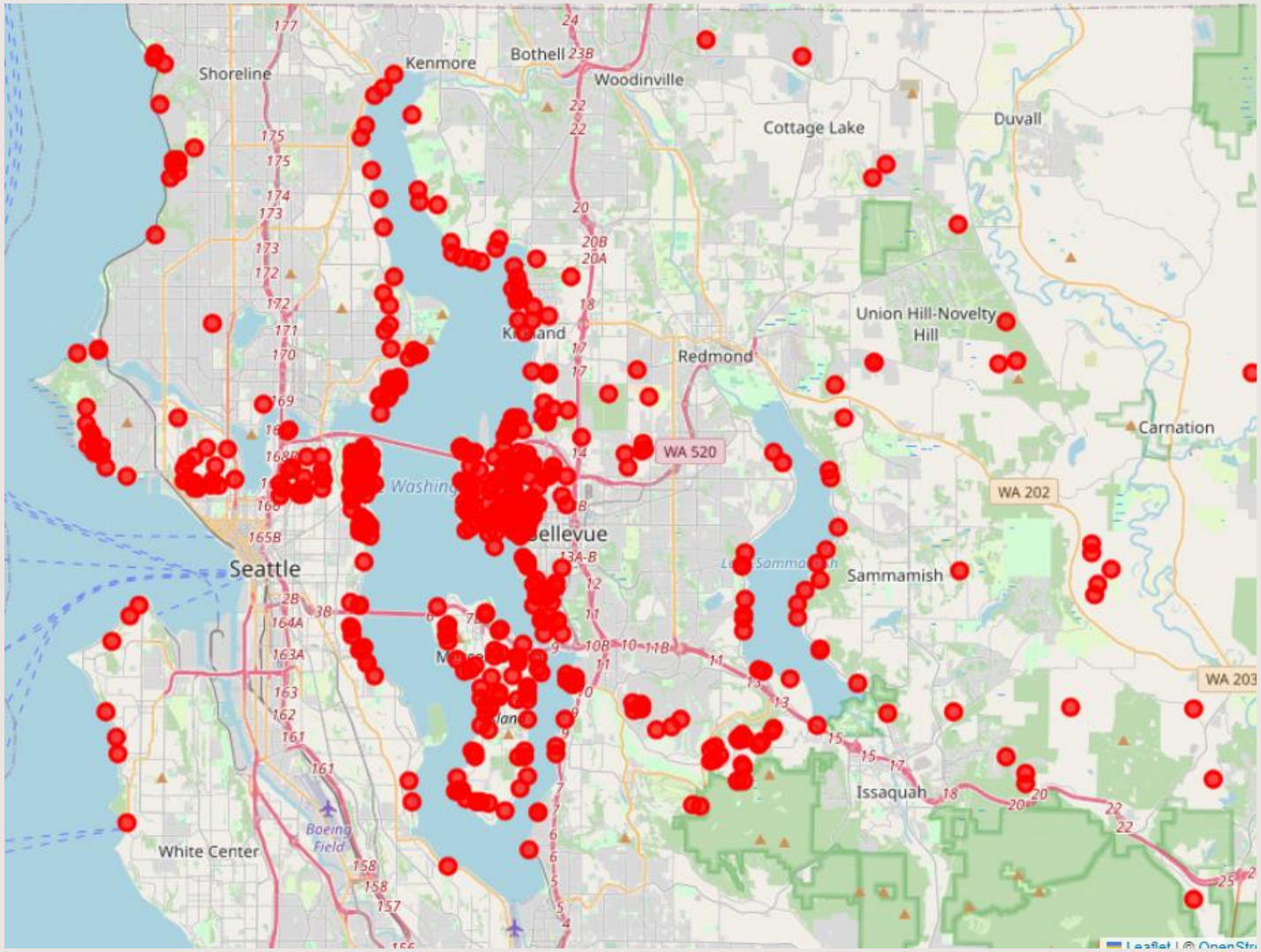
BUILD PREDICTIVE MODEL

CONCLUSION & RECOMMENDATIONS

OBTAIN & ANALYSE DATASET

Data scrubbing:

1. Investigating the names of the columns, dataset shape and size.
2. Removing incomplete, incorrect, inaccurately formatted or repeated data.
3. Adding and replacing columns.
4. Removing columns we no longer need for our model.



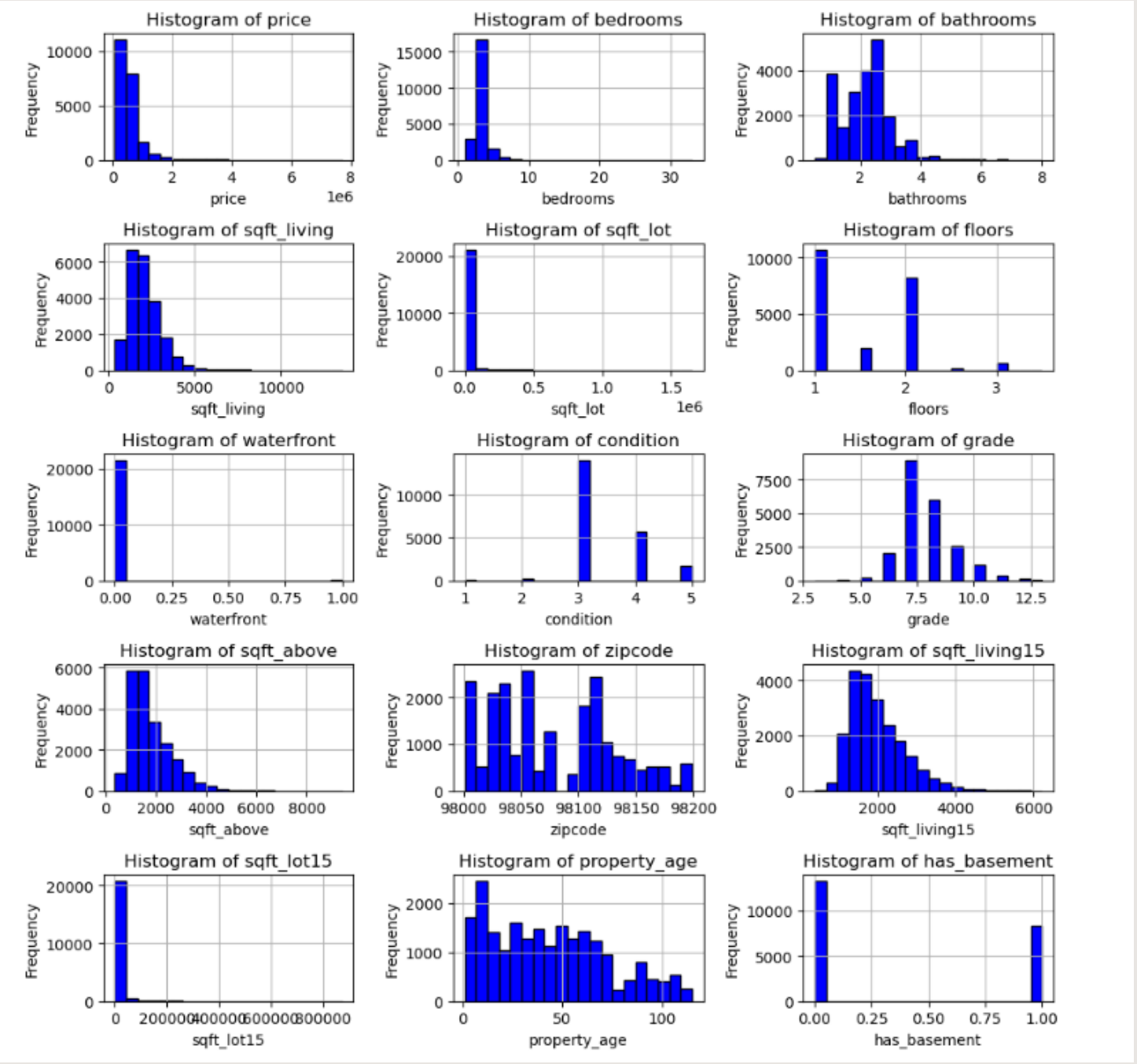
OBTAIN & ANALYSE DATASET

- The most expensive properties are located around Lake Washington near cities Bellevue, Medina, Seattle and Mercer Island and those near water complimenting feature “waterfront”

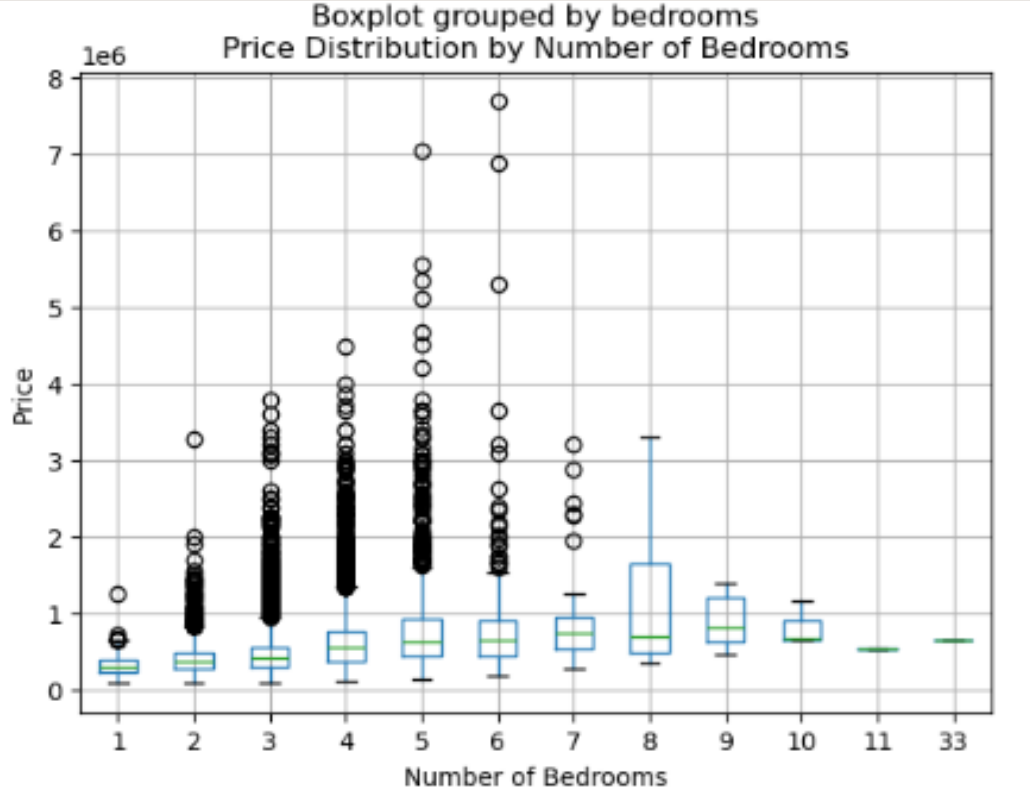
INVESTIGATE FEATURES

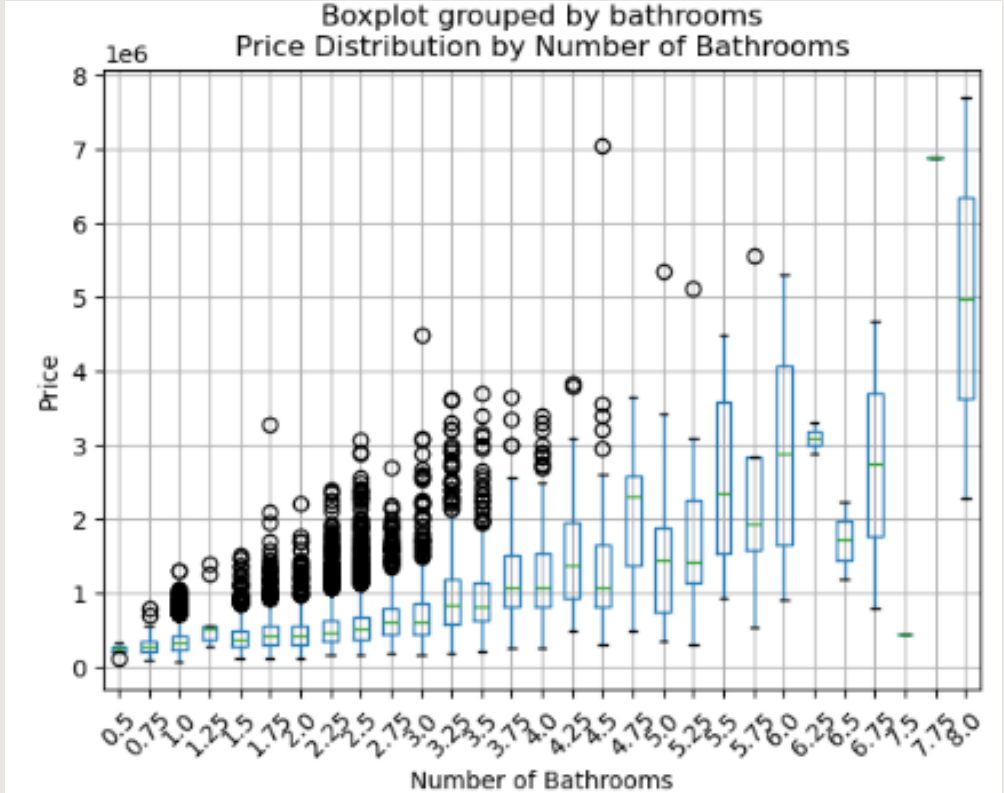
Distribution Histogram

- Bedrooms
- Bathrooms

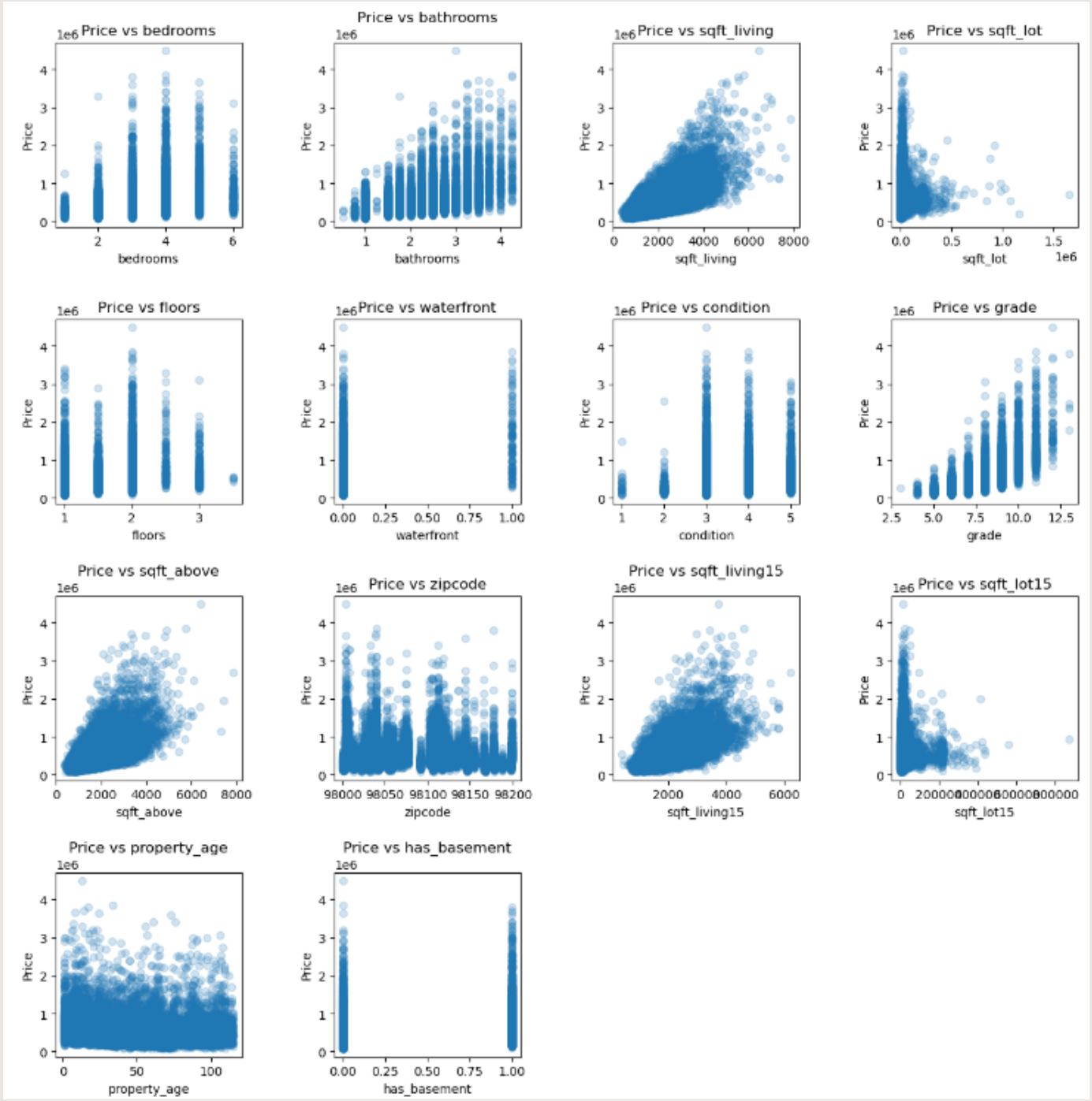


INVESTIGATE FEATURES



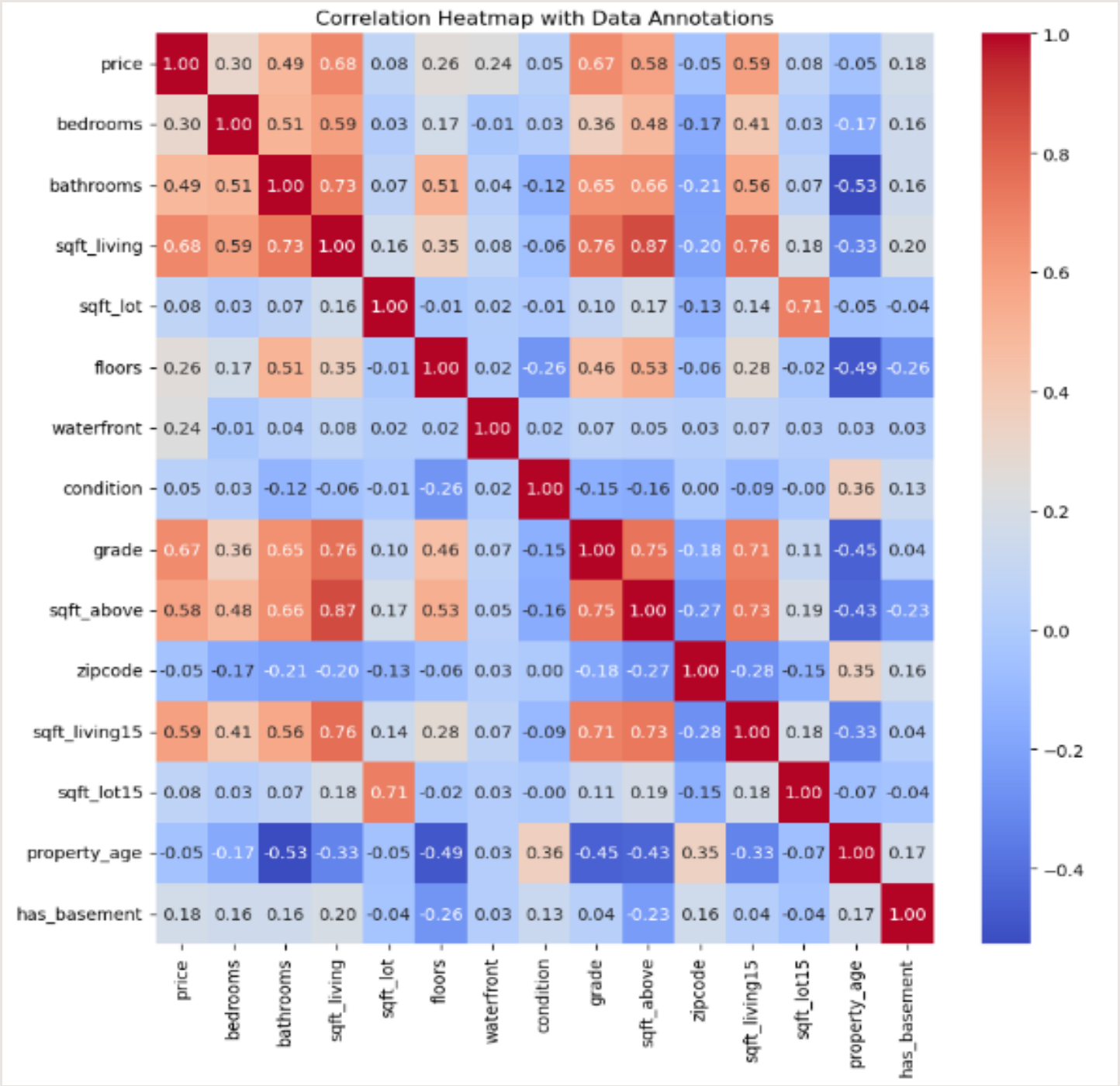


Getting rid of outliers using Z-score



Closer look required to “sqft_lot” and “sqft_lot15” features since the linearity is not obvious

INVESTIGATE FEATURES



INVESTIGATE FEATURES

pairs	cc
(sqft_living, sqft_above)	0.866811
(sqft_living15, sqft_living)	0.762733
(grade, sqft_living)	0.755437
(grade, sqft_above)	0.746896
(sqft_above, sqft_living15)	0.733247
(sqft_living, bathrooms)	0.732038
(sqft_lot, sqft_lot15)	0.710701
(grade, sqft_living15)	0.706767

- Signs of the Multicollinearity.
- Some of the features removed from the model.

BUILD PREDICTION MODEL

OLS Regression Results

Dep. Variable:	price	R-squared:	0.633
Model:	OLS	Adj. R-squared:	0.633
Method:	Least Squares	F-statistic:	3349.
Date:	Mon, 13 May 2024	Prob (F-statistic):	0.00
Time:	13:20:53	Log-Likelihood:	-2.9087e+05
No. Observations:	21351	AIC:	5.818e+05
Df Residuals:	21339	BIC:	5.819e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.005e+06	2.82e+06	-0.356	0.722	-6.53e+06	4.52e+06
bedrooms	-3.827e+04	2041.391	-18.747	0.000	-4.23e+04	-3.43e+04
bathrooms	3.702e+04	3385.885	10.935	0.000	3.04e+04	4.37e+04
sqft_living	151.4208	3.177	47.662	0.000	145.194	157.648
sqft_lot	-0.1439	0.034	-4.179	0.000	-0.211	-0.076
floors	3.383e+04	3441.935	9.830	0.000	2.71e+04	4.06e+04
waterfront	6.548e+05	1.76e+04	37.227	0.000	6.2e+05	6.89e+05
condition	2.074e+04	2298.963	9.021	0.000	1.62e+04	2.52e+04
grade	1.35e+05	2004.808	67.339	0.000	1.31e+05	1.39e+05
zipcode	-0.4592	28.758	-0.016	0.987	-56.827	55.908
property_age	3625.3536	66.058	54.882	0.000	3495.876	3754.831
has_basement	2.146e+04	3271.112	6.560	0.000	1.5e+04	2.79e+04

Omnibus:	10883.474	Durbin-Watson:	1.969
Prob(Omnibus):	0.000	Jarque-Bera (JB):	186226.582
Skew:	2.049	Prob(JB):	0.00
Kurtosis:	16.876	Cond. No.	2.05e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.05e+08. This might indicate that there are strong multicollinearity or other numerical problems.

Steps to improve model

- R-squared 0.63.
- Zipcode P-value is insignificant (>0.05), we replaced zipcode with the actual city name.
- Created dummy variables for categorical variables.
- Did Log transformation for some of the numerical variables.

BUILD PREDICTION MODEL

OLS Regression Results

Dep. Variable:	price	R-squared:	0.725
Model:	OLS	Adj. R-squared:	0.724
Method:	Least Squares	F-statistic:	1001.
Date:	Mon, 13 May 2024	Prob (F-statistic):	0.00
Time:	13:21:33	Log-Likelihood:	-2.8780e+05
No. Observations:	21351	AIC:	5.757e+05
Df Residuals:	21294	BIC:	5.762e+05
Df Model:	56		
Covariance Type:	nonrobust		

Omnibus:	11385.155	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	229258.033
Skew:	2.124	Prob(JB):	0.00
Kurtosis:	18.481	Cond. No.	97.0

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Training Data Metrics:
MSE: 31075912377.38278
RMSE: 176283.61346813486
MAE: 115223.00707397386

Test Data Metrics:
MSE: 25481386575.611668
RMSE: 159628.9026950059
MAE: 110731.4216906077

Final Model Key Features

- R-squared - 0.725.
- F-statistic – 1001.
- Prob (F-statistic) – 0.00.
- No. of conditions - 97.0.
- P-values are significant.
- RMSE on Train data is slightly higher than RMSE on Test data (within 10% bracket – model is not overfitted.
- Training and Test Error metrics are reasonable and smaller than standard deviation.

const	5.195e+05	7511.609	69.164	0.000	5.05e+05	5.34e+05
waterfront	6.595e+05	1.55e+04	42.588	0.000	6.29e+05	6.9e+05
bed_3	-3.79e+04	4182.512	-9.062	0.000	-4.61e+04	-2.97e+04
bed_4	-5.735e+04	4969.154	-11.542	0.000	-6.71e+04	-4.76e+04
bed_5	-6.931e+04	6689.404	-10.361	0.000	-8.24e+04	-5.62e+04
bed_6	-9.888e+04	1.26e+04	-7.854	0.000	-1.24e+05	-7.42e+04
bath_1.5	-2.245e+04	5301.720	-4.234	0.000	-3.28e+04	-1.21e+04
bath_1.75	-1.993e+04	4267.897	-4.671	0.000	-2.83e+04	-1.16e+04
bath_2.0	-1.478e+04	4850.884	-3.047	0.002	-2.43e+04	-5273.858
bath_2.25	-1.751e+04	5087.848	-3.441	0.001	-2.75e+04	-7536.649
bath_2.5	-2.436e+04	4483.235	-5.434	0.000	-3.31e+04	-1.56e+04
bath_3.0	1.965e+04	7404.483	2.653	0.008	5131.692	3.42e+04
bath_3.25	9.998e+04	8479.636	11.790	0.000	8.34e+04	1.17e+05
bath_3.5	5.493e+04	8069.238	6.807	0.000	3.91e+04	7.07e+04
bath_3.75	1.918e+05	1.5e+04	12.766	0.000	1.62e+05	2.21e+05
bath_4.0	1.973e+05	1.65e+04	11.981	0.000	1.65e+05	2.3e+05
bath_4.25	3.368e+05	2.09e+04	16.149	0.000	2.96e+05	3.78e+05
floor_1.5	4.899e+04	4531.544	10.811	0.000	4.01e+04	5.79e+04
floor_2.0	1.689e+04	3753.364	4.500	0.000	9532.707	2.42e+04
floor_2.5	1.384e+05	1.46e+04	9.502	0.000	1.1e+05	1.67e+05
floor_3.0	2.527e+04	8588.132	2.942	0.003	8434.850	4.21e+04
grade_4	-1.797e+05	3.44e+04	-5.230	0.000	-2.47e+05	-1.12e+05
grade_5	-2.478e+05	1.3e+04	-18.990	0.000	-2.73e+05	-2.22e+05
grade_6	-2.719e+05	6967.127	-39.023	0.000	-2.86e+05	-2.58e+05
grade_7	-2.192e+05	5008.158	-43.768	0.000	-2.29e+05	-2.09e+05
grade_8	-1.406e+05	4401.510	-31.938	0.000	-1.49e+05	-1.32e+05
grade_10	1.613e+05	6506.825	24.790	0.000	1.49e+05	1.74e+05
grade_11	4.063e+05	1.06e+04	38.205	0.000	3.85e+05	4.27e+05
grade_12	7.75e+05	2.27e+04	34.146	0.000	7.31e+05	8.2e+05
grade_13	1.403e+06	7.79e+04	18.017	0.000	1.25e+06	1.56e+06
cond_4	3.168e+04	3022.697	10.481	0.000	2.58e+04	3.76e+04
cond_5	8.869e+04	4738.908	18.716	0.000	7.94e+04	9.8e+04
city_Bellevue	3.499e+05	6309.049	55.453	0.000	3.37e+05	3.62e+05
city_Black Diamond	7.918e+04	1.79e+04	4.419	0.000	4.41e+04	1.14e+05
city_Bothell	1.218e+05	1.3e+04	9.360	0.000	9.63e+04	1.47e+05
city_Carnation	1.102e+05	1.64e+04	6.734	0.000	7.81e+04	1.42e+05
city_Duvall	1.002e+05	1.32e+04	7.589	0.000	7.43e+04	1.26e+05
city_Enumclaw	2.936e+04	1.21e+04	2.432	0.015	5692.908	5.3e+04
city_Fall City	1.761e+05	2.03e+04	8.671	0.000	1.36e+05	2.16e+05
city_Federal Way	-3.033e+04	7333.691	-4.135	0.000	-4.47e+04	-1.6e+04
city_Issaquah	1.515e+05	7594.598	19.947	0.000	1.37e+05	1.66e+05
city_Kenmore	1.243e+05	1.1e+04	11.277	0.000	1.03e+05	1.46e+05
city_Kirkland	2.548e+05	6819.843	37.363	0.000	2.41e+05	2.68e+05
city_Maple Valley	4.203e+04	8153.775	5.155	0.000	2.6e+04	5.8e+04
city_Medina	1.091e+06	2.78e+04	39.228	0.000	1.04e+06	1.15e+06
city_Mercer Island	4.581e+05	1.16e+04	39.541	0.000	4.35e+05	4.81e+05
city_North Bend	1.145e+05	1.24e+04	9.212	0.000	9.01e+04	1.39e+05
city_Redmond	2.122e+05	6845.779	30.991	0.000	1.99e+05	2.26e+05
city_Renton	5.614e+04	5792.094	9.693	0.000	4.48e+04	6.75e+04
city_Sammamish	1.482e+05	7499.557	19.760	0.000	1.33e+05	1.63e+05
city_Seattle	2.154e+05	4690.308	45.930	0.000	2.06e+05	2.25e+05
city_Snoqualmie	1.149e+05	1.08e+04	10.678	0.000	9.38e+04	1.36e+05
city_Vashon	5.399e+04	1.7e+04	3.184	0.001	2.08e+04	8.72e+04
city_Woodinville	1.28e+05	9080.999	14.092	0.000	1.1e+05	1.46e+05
liv	1.107e+05	2420.518	45.753	0.000	1.06e+05	1.15e+05
lot	-7562.8688	1686.591	-4.484	0.000	-1.09e+04	-4257.022
prop_age	2.436e+04	1810.741	13.455	0.000	2.08e+04	2.79e+04

CONCLUSION

Key features with strong influence

- Waterfront house gets us increase of $659,500 in house price.
- Higher Grade brings significant house price increase from $160k for grade 10 to $1.4 mil for grade 13.
- House in Bellevue would get additional $350k as well as house in Medina would get us over $1 mil.
- Over 3 bathrooms will get us from $20k to $336k increase in average.
- Number of bedrooms is not representative, however each additional sqft living will get us $111k increase.

RECOMMENDATIONS

- Invest in the waterfront houses.
- Invest in the houses with the higher grade, especially those with the grade over 10.
- Invest in the houses near water such as Bellevue and Medina.
- Invest in properties with 3 and more bathrooms.
- Invest into the houses with the higher sqft_living (not necessarily higher number of bedrooms).

THANK YOU

Email: grigorenko.jane@gmail.com

GitHub: https://github.com/JaneGrig

LinkedIn: linkedin.com/in/evgeniya-jane-grigorenko-47692211b