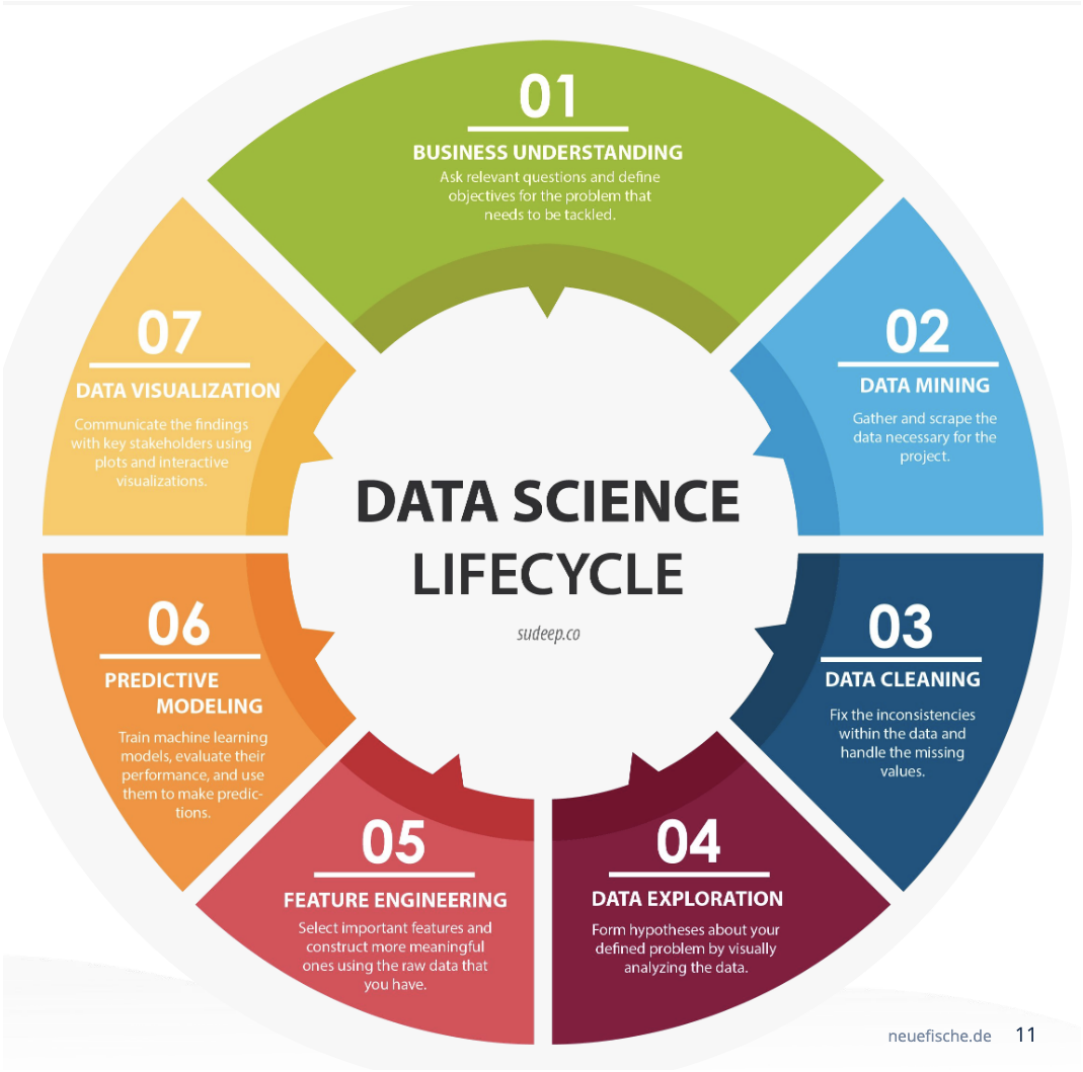
1st Project: Regression

by Silas Mederer
at neue fische GmbH

What you can expect

1. Overviewing data
2. Client descriptions and request
3. Regression models

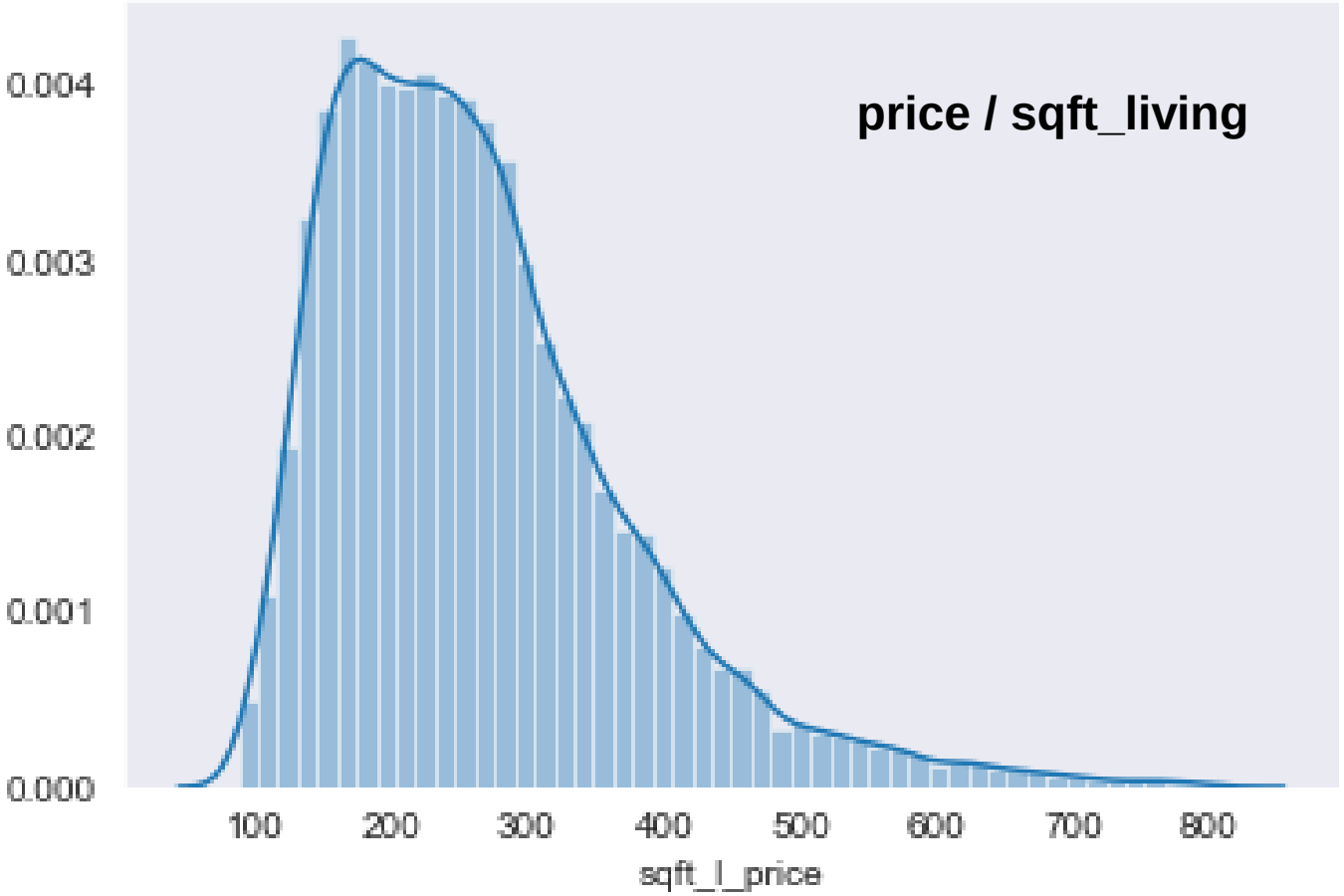
4. Linear regression with bad results
5. Future projects


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co
neuefische.de 11

1. Overviewing data

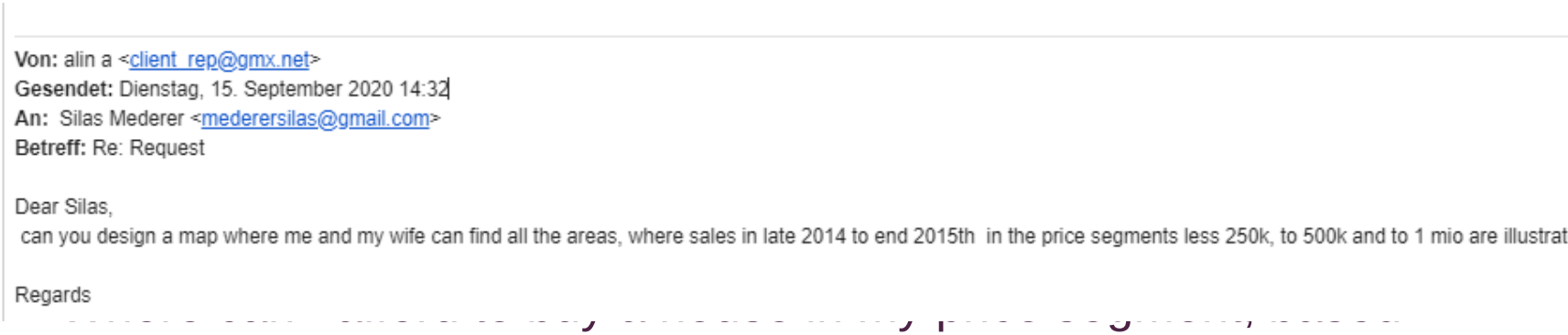
- find missing
- fillna()
- find outliners & deal with them
  - only bedrooms < 10 if the ratio was less than 3 to 1 bathroom

1. Overviewing data – adding feature



2. Client descriptions and request

- Client status groups:
  - institutional developers and wealthy clients



Von: alin a <client_rep@gmx.net>
Gesendet: Dienstag, 15. September 2020 14:32
An: Silas Mederer <mederersilas@gmail.com>
Betreff: Re: Request

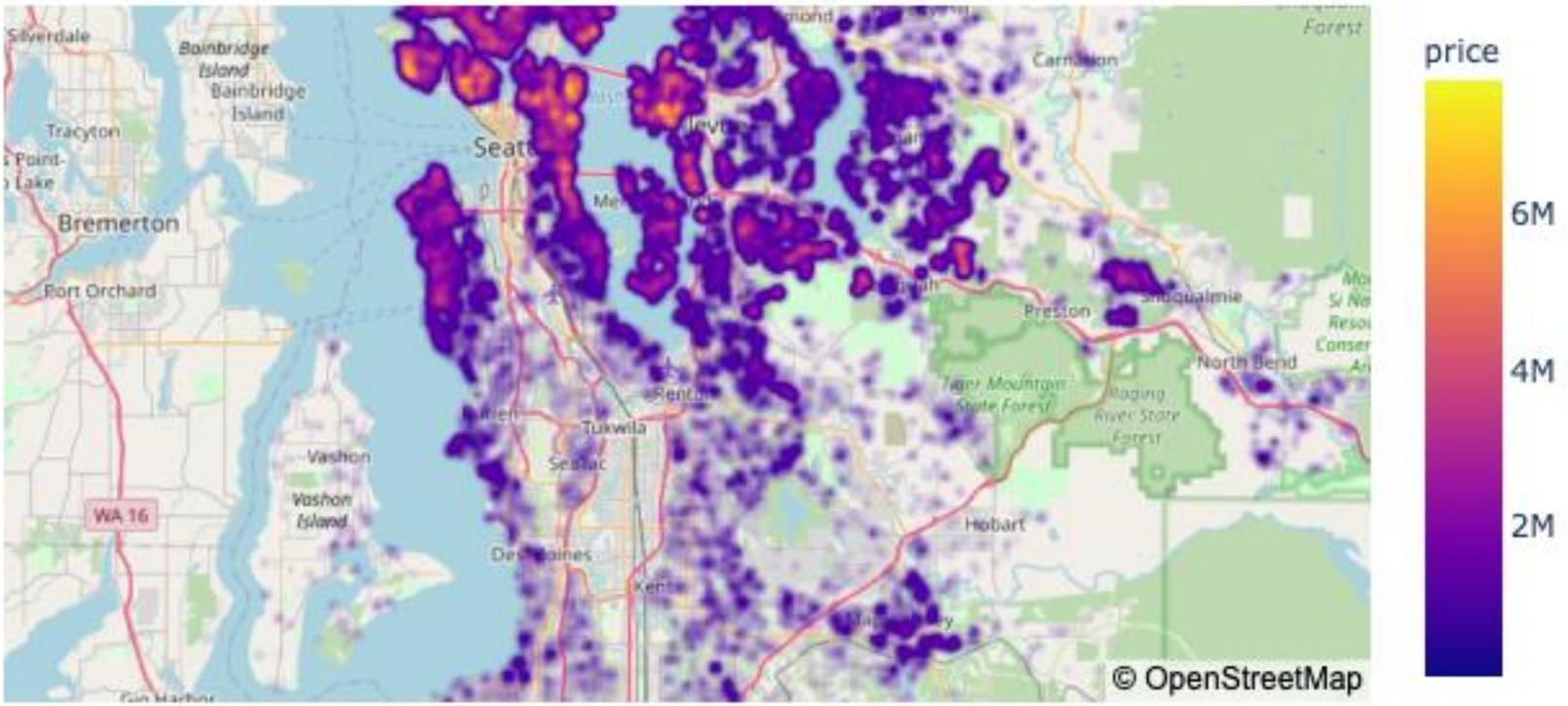
Dear Silas,
can you design a map where me and my wife can find all the areas, where sales in late 2014 to end 2015th in the price segments less 250k, to 500k and to 1 mio are illustrat

Regards

- F
  - on previous sales?
  - We want a map!

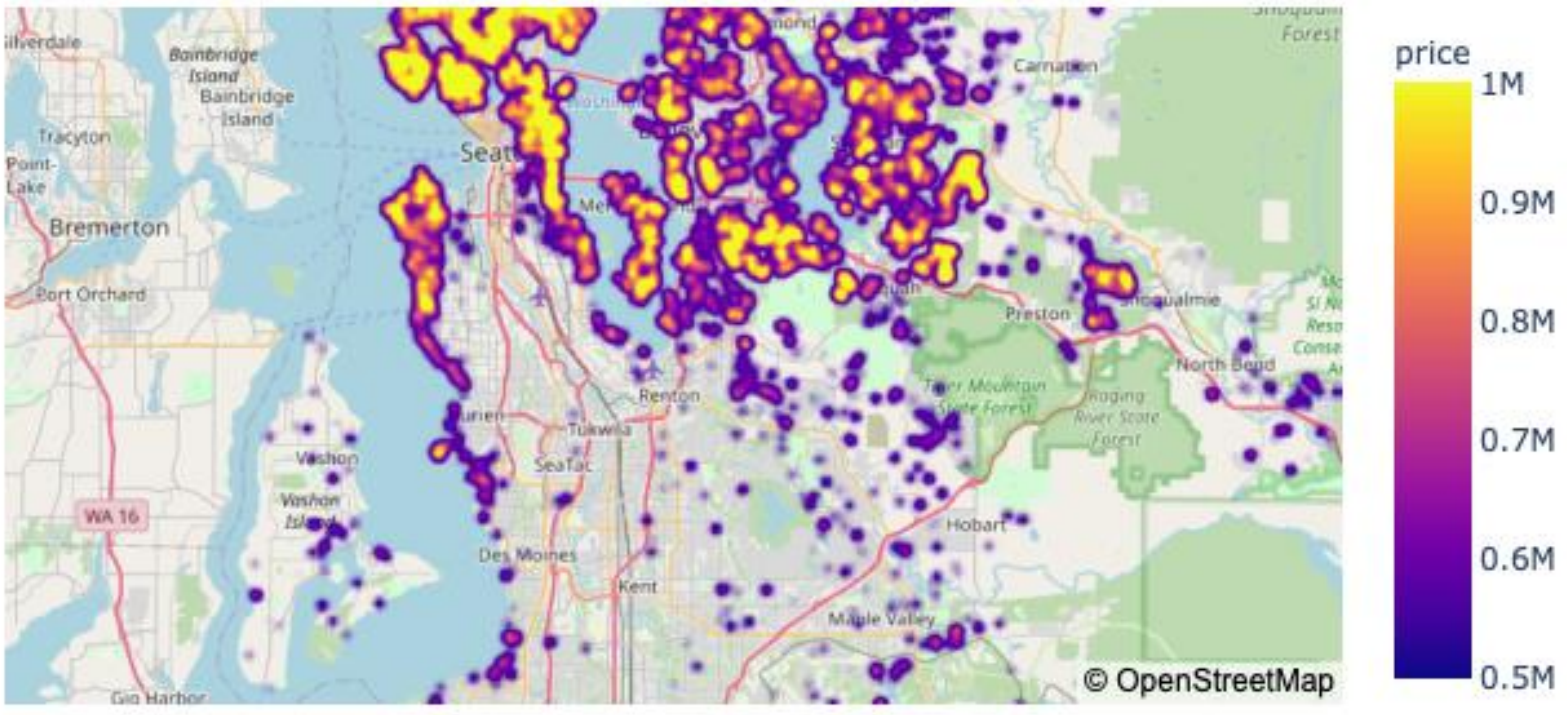
2. Client descriptions and request

Real Estate Regional Cluster



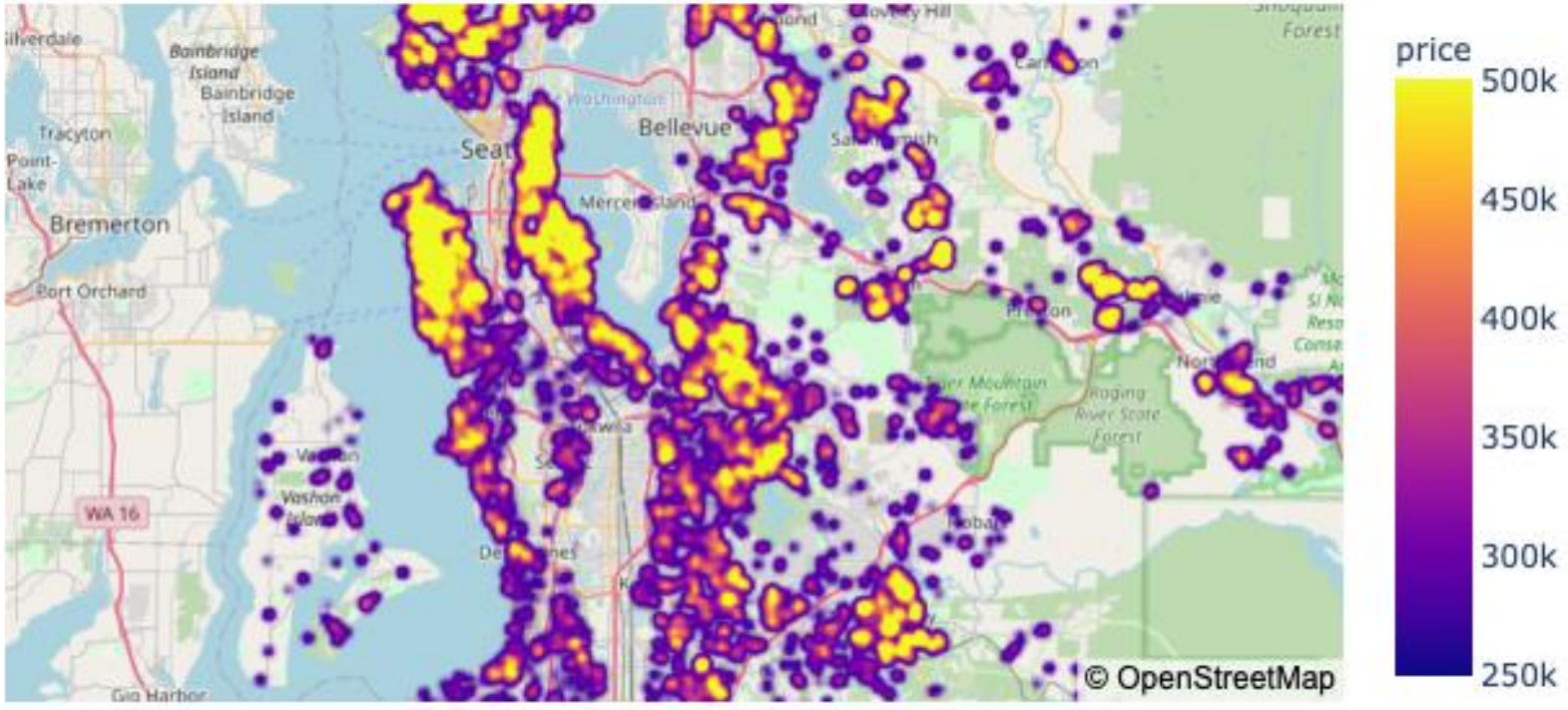
2. Client descriptions and request

Real Estate Regional Cluster Price less 1 Mio



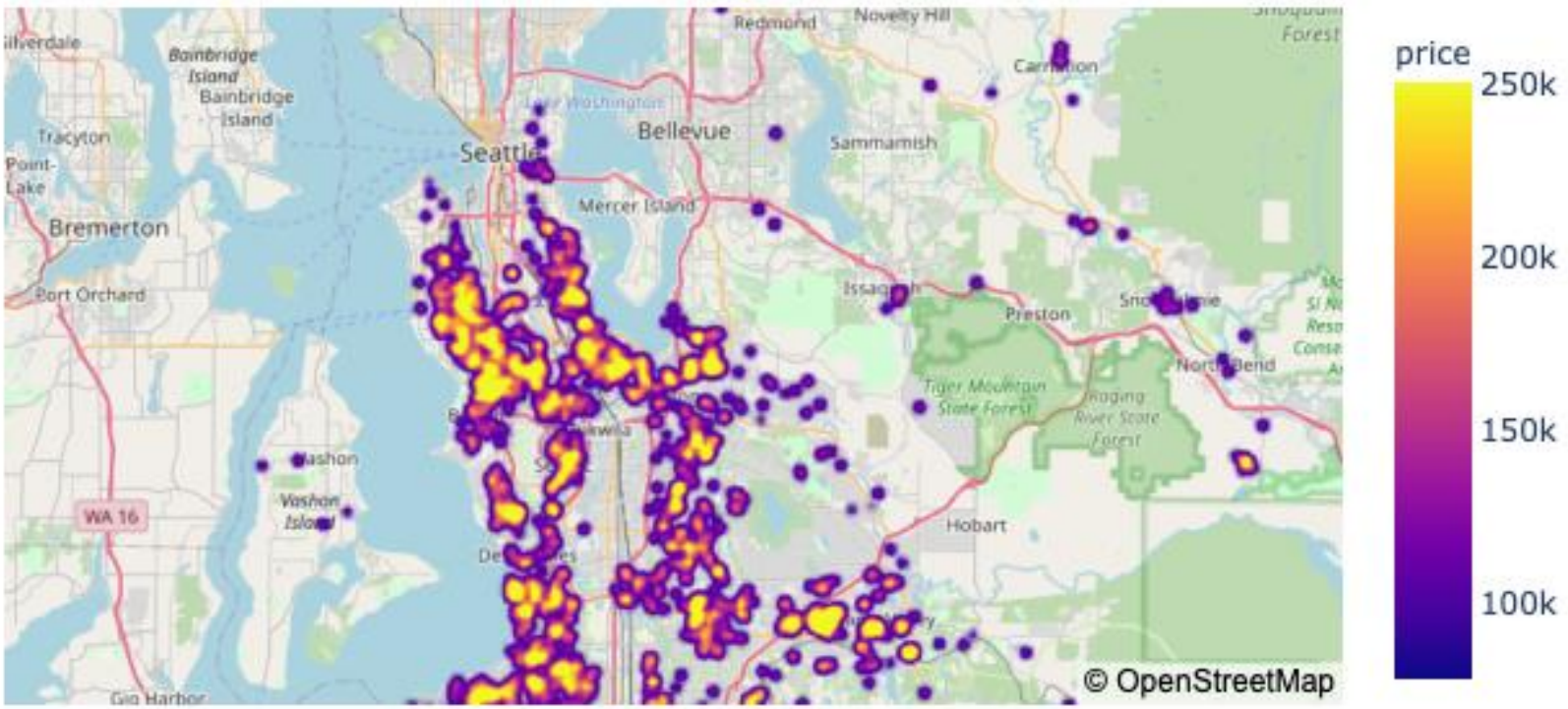
2. Client descriptions and request

Real Estate Regional Cluster Price less 500k



2. Client descriptions and request

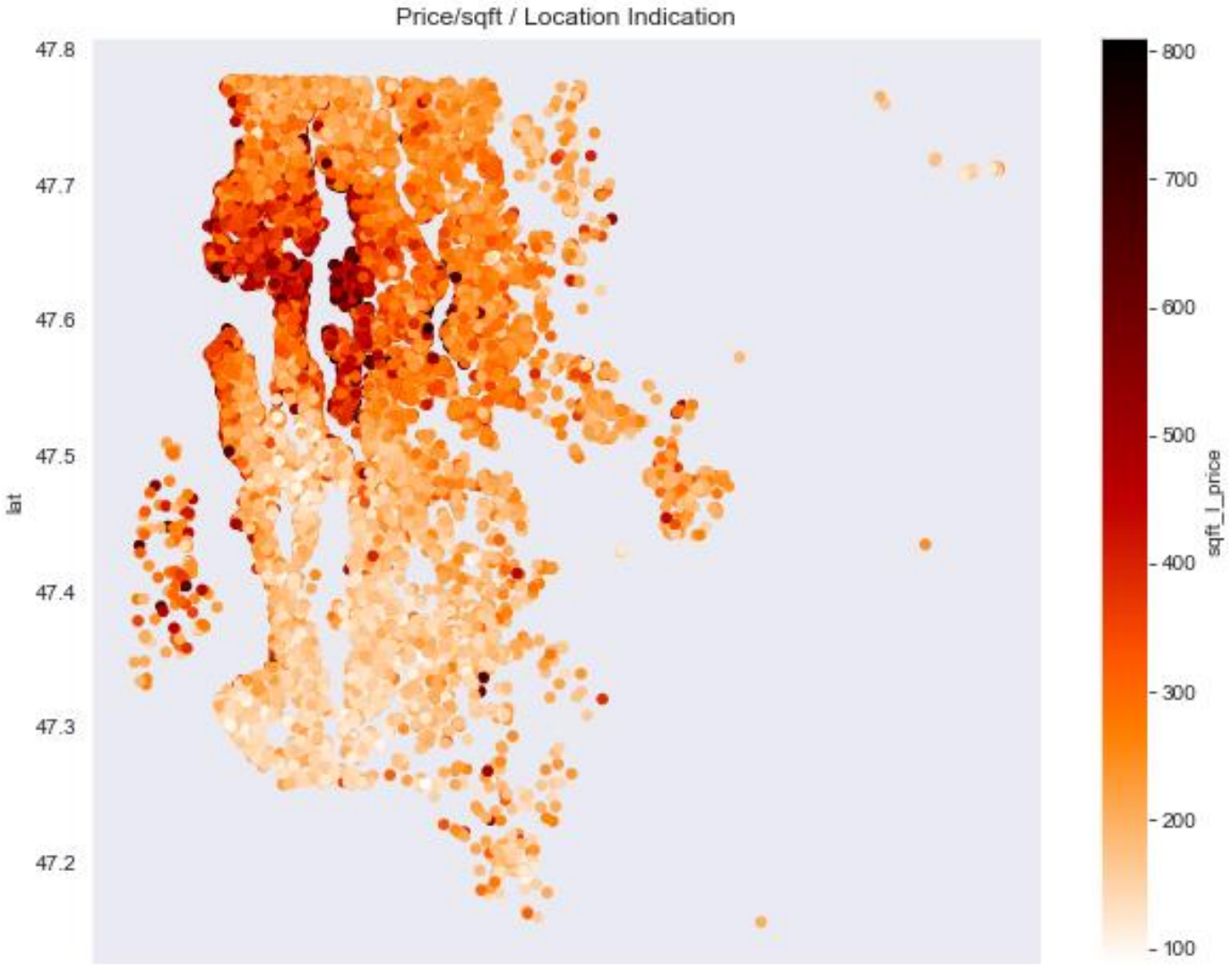
Real Estate Regional Cluster less 250k



2. Client descriptions and request

The more you go to the south the cheaper it gets

2. Client descriptions and request



3. Regression Model

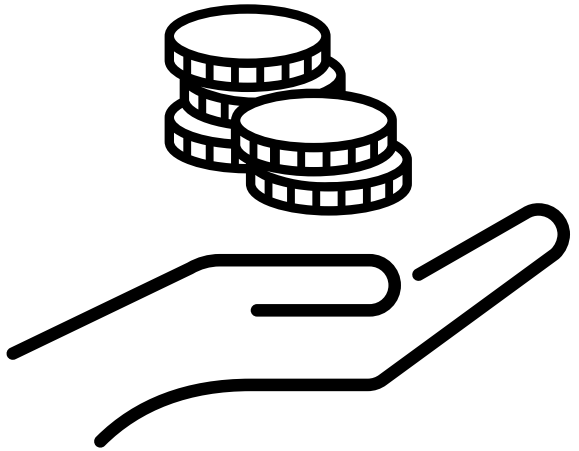
less 1 Mio R 54%

In [31]:
```
# less 1mio
model = 'price ~ sqft_living + bathrooms + bedrooms + waterfront + yr_built + condition + grade + sqft_living15 + sqft_lot15'
smf.ols(formula=model, data=df_less_1mio).fit().summary()
```

Out[31]:

Dep. Variable:	price	**R-squared:**	0.549
Model:	OLS	**Adj. R-squared:**	0.549
Method:	Least Squares	**F-statistic:**	2720.
Date:	Thu, 17 Sep 2020	**Prob (F-statistic):**	0.00
Time:	07:39:57	**Log-Likelihood:**	-2.6600e+05
No. Observations:	20137	**AIC:**	5.320e+05
Df Residuals:	20127	**BIC:**	5.321e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.452e+06	8.25e+04	53.951	0.000	4.29e+06	4.61e+06
sqft_living	61.2185	2.552	23.991	0.000	56.217	66.220
bathrooms	4.236e+04	2172.432	19.500	0.000	3.81e+04	4.66e+04
bedrooms	-1.681e+04	1378.878	-12.192	0.000	-1.95e+04	-1.41e+04
waterfront	1.698e+05	1.87e+04	9.066	0.000	1.33e+05	2.07e+05
yr_built	-2525.3528	42.614	-59.260	0.000	-2608.881	-2441.825
condition	1.368e+04	1547.673	8.841	0.000	1.06e+04	1.67e+04
grade	9.344e+04	1457.136	64.126	0.000	9.06e+04	9.63e+04
sqft_living15	50.3715	2.424	20.781	0.000	45.620	55.123
sqft_lot15	-0.1550	0.036	-4.277	0.000	-0.226	-0.084



No bonus for this

3. Regression Model

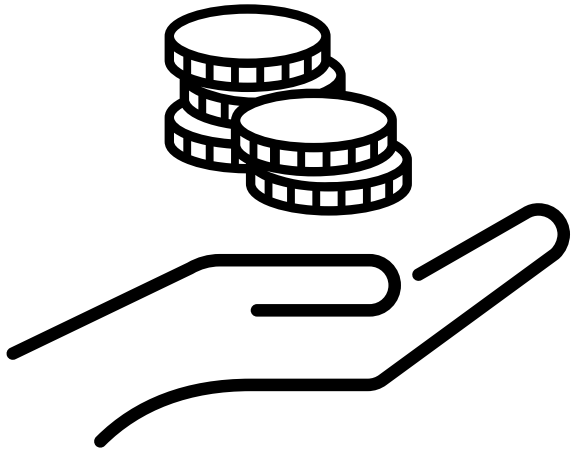
Over all 65%

In [28]:
```
# over all
model = 'price ~ sqft_living + bathrooms + bedrooms + waterfront + yr_built + condition + grade + sqft_living15 + sqft_lot15'
smf.ols(formula=model, data=df).fit().summary()
```

Out[28]:

Dep. Variable:	price	**R-squared:**	0.648
Model:	OLS	**Adj. R-squared:**	0.648
Method:	Least Squares	**F-statistic:**	4409.
Date:	Thu, 17 Sep 2020	**Prob (F-statistic):**	0.00
Time:	07:39:57	**Log-Likelihood:**	-2.9610e+05
No. Observations:	21595	**AIC:**	5.922e+05
Df Residuals:	21585	**BIC:**	5.923e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.59e+06	1.27e+05	51.722	0.000	6.34e+06	6.84e+06
sqft_living	168.3836	3.629	46.400	0.000	161.271	175.497
bathrooms	5.835e+04	3366.491	17.333	0.000	5.18e+04	6.5e+04
bedrooms	-4.768e+04	2133.938	-22.345	0.000	-5.19e+04	-4.35e+04
waterfront	7.53e+05	1.83e+04	41.113	0.000	7.17e+05	7.89e+05
yr_built	-3774.7484	65.280	-57.824	0.000	-3902.702	-3646.795
condition	1.757e+04	2460.444	7.141	0.000	1.27e+04	2.24e+04
grade	1.257e+05	2224.164	56.530	0.000	1.21e+05	1.3e+05
sqft_living15	28.7164	3.529	8.136	0.000	21.799	35.634
sqft_lot15	-0.5908	0.056	-10.544	0.000	-0.701	-0.481



No bonus for this

3. Regression Model

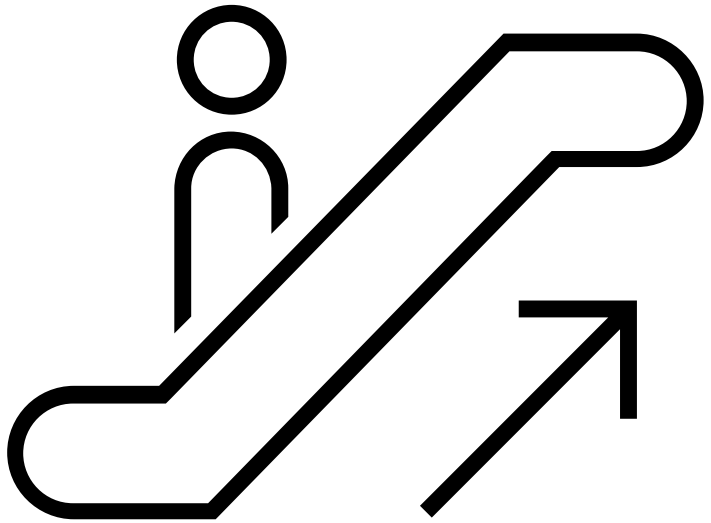
Price sqft lower Mean R 76%

In [32]:
```
# better price
model = 'price ~ sqft_living + bathrooms + bedrooms + waterfront + yr_built + condition + grade + sqft_living15 + sqft_lot15'
smf.ols(formula=model, data=df_better_pricesqft).fit().summary()
```

Out[32]:

Dep. Variable:	price	**R-squared:**	0.761
Model:	OLS	**Adj. R-squared:**	0.761
Method:	Least Squares	**F-statistic:**	4396.
Date:	Thu, 17 Sep 2020	**Prob (F-statistic):**	0.00
Time:	07:39:57	**Log-Likelihood:**	-1.6009e+05
No. Observations:	12429	**AIC:**	3.202e+05
Df Residuals:	12419	**BIC:**	3.203e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.406e+06	9.11e+04	15.428	0.000	1.23e+06	1.58e+06
sqft_living	141.9033	2.200	64.503	0.000	137.591	146.215
bathrooms	1.336e+04	2016.045	6.628	0.000	9411.475	1.73e+04
bedrooms	-9536.2299	1256.931	-7.587	0.000	-1.2e+04	-7072.450
waterfront	-906.2210	3.37e+04	-0.027	0.979	-6.69e+04	6.51e+04
yr_built	-871.1596	47.118	-18.489	0.000	-963.519	-778.800
condition	6564.2311	1500.943	4.373	0.000	3622.150	9506.312
grade	4.504e+04	1404.643	32.064	0.000	4.23e+04	4.78e+04
sqft_living15	31.7251	2.245	14.131	0.000	27.324	36.126
sqft_lot15	0.1253	0.028	4.410	0.000	0.070	0.181



Getting better nine weeks to go

4. Linear regression with bad results

=

5. Future projects

Thanks for your attention

any questions or critiques

special Thanks to Tjade, JJ, Silvia and Lars who supported me in the making