EDA Project:
Seattle House Sales

Client : Jennifer Montgomery

- high budget
- wants to show off
- waterfront
- renovated
- high grades
- (buy within a month / resell within 1 year)

my assumptions:

- main purpose: showing off
- not profit-driven, even reselling at loss would be ok
- focus on: sqft living/lot, #bedrooms, grades

  → comparing to neighbours

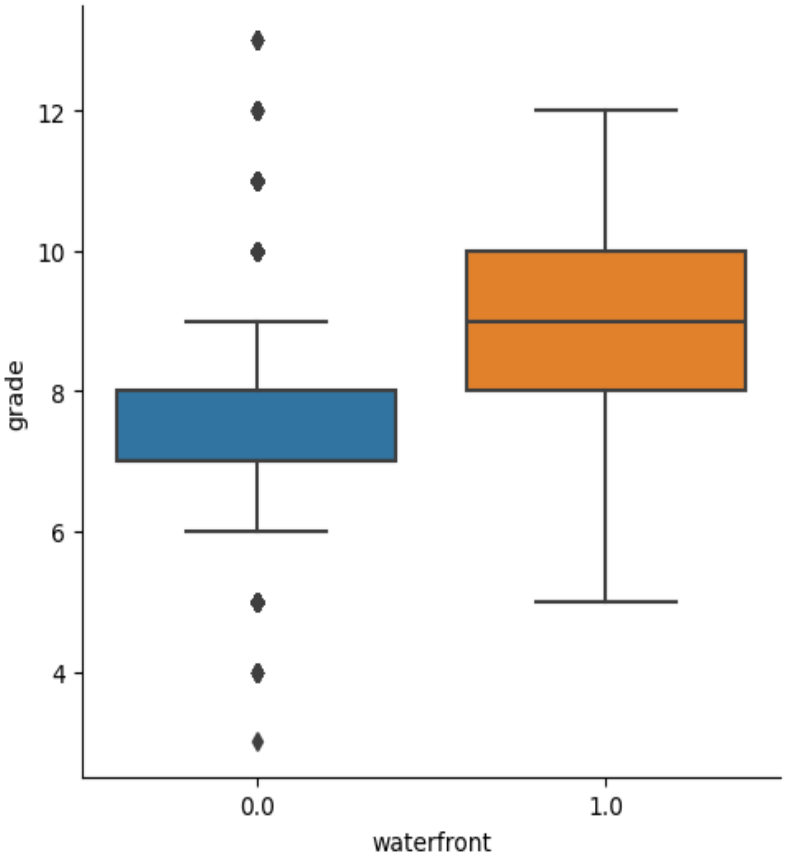
Hypotheses:

1. houses with waterfront have higher grades
   and are more expensive

2. houses with high sqft_living:
   have nighbours with high sqft_living

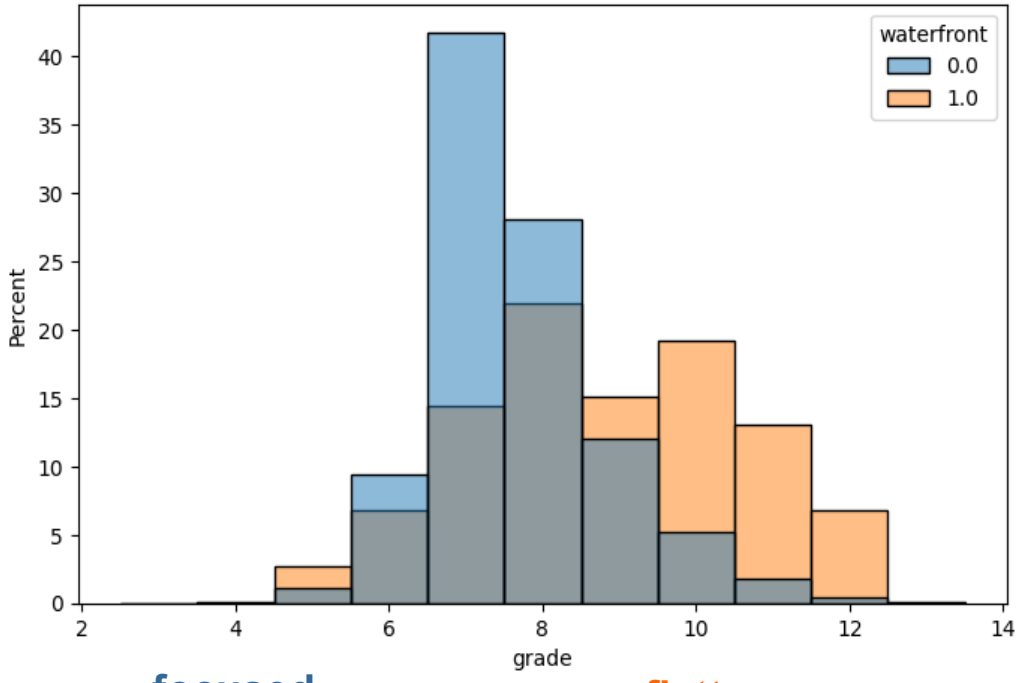
3. houses with higher sqft_living than their neighbours
   are much more expensive

Hypothesis 1.1.
Houses with waterfront have higher grades



mean:
7.7 +/- 1.2

mean:
8.8 +/- 1.8



- focused
~ 70%: 7 & 8

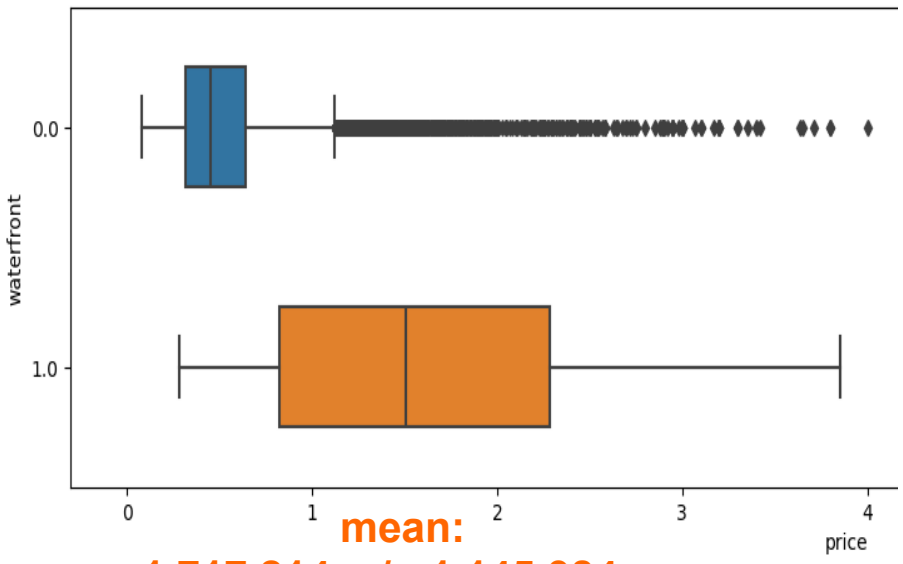
- flatter
- spread out over
7 - 11

significant? yes!

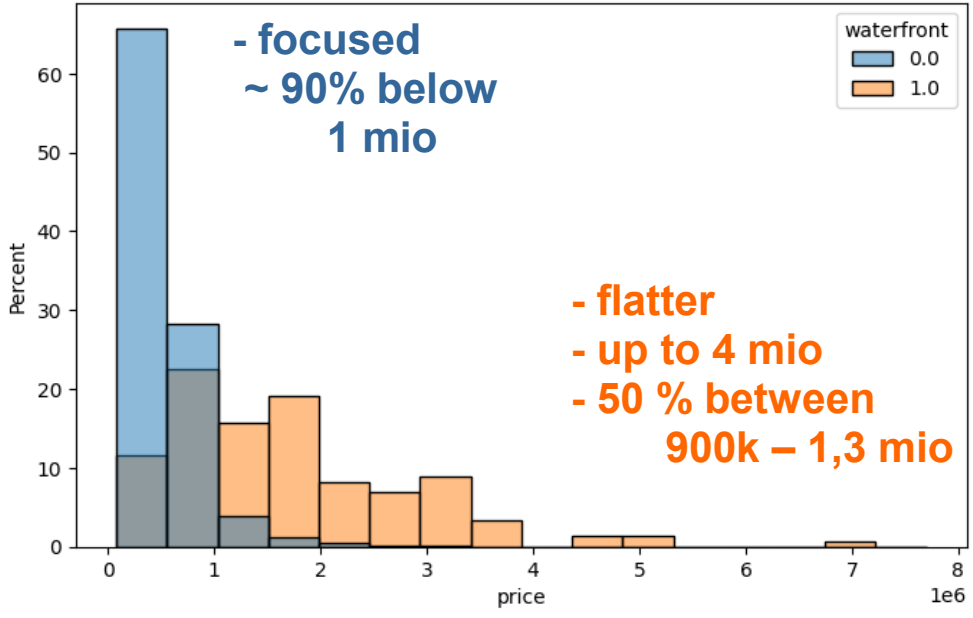
one-sided t-test: H0 : μ_wf < = μ_nwf H1: μ_wf > μ_nwf
t-statistic: 12.21 , p-value: 1.73 e-34, alpha = 0,05
reject H0: waterfront houses have significantly higher grades

Hypothesis 1.2.
Houses with waterfront are more expensive

mean:
532 286 +/- 342 808



mean:
1 717 214 +/- 1 145 384

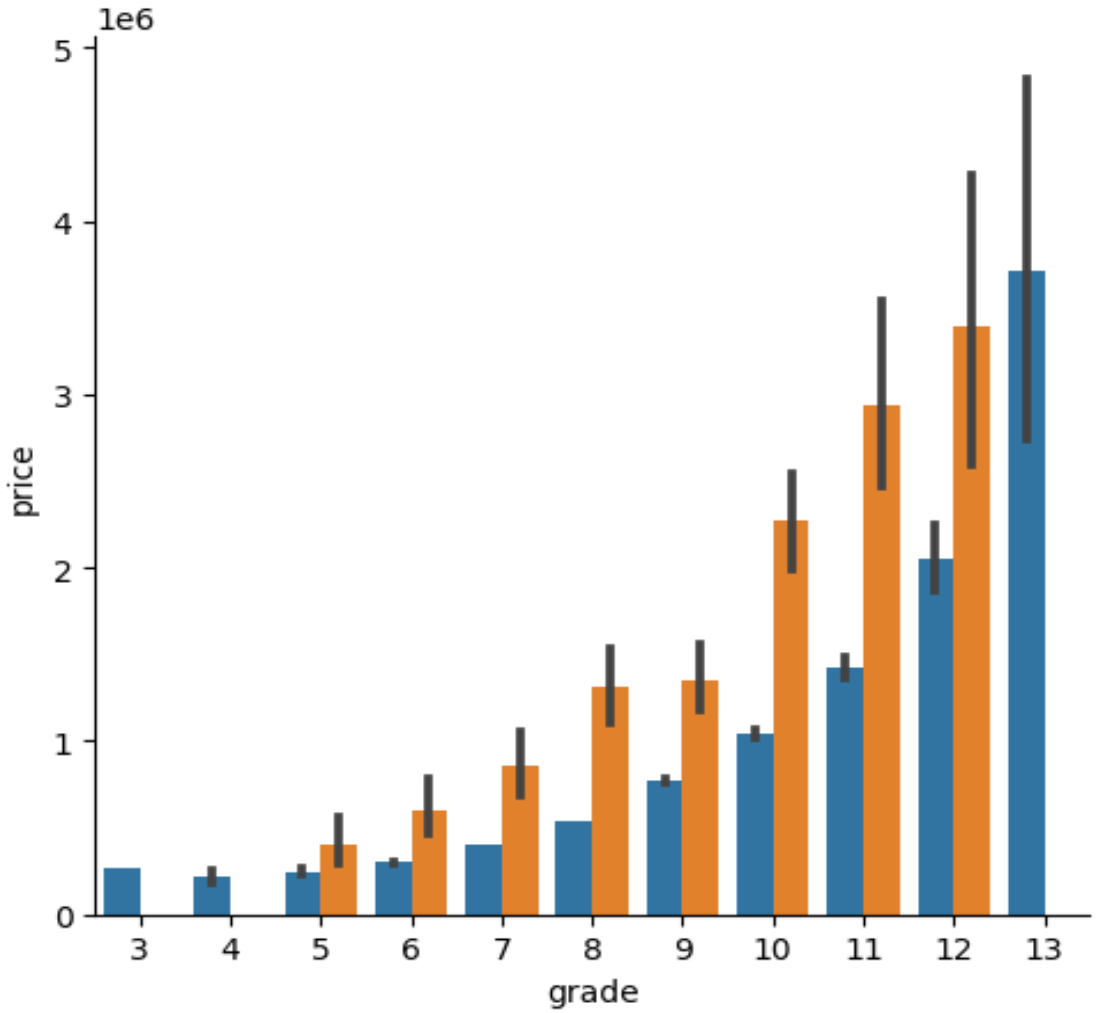


significant? yes!

one-sided t-test: H0 : μ_wf < = μ_nwf H1: μ_wf > μ_nwf
t-statistic: 40.27, p-value: 0.0 alpha = 0,05
reject H0: waterfront houses are significantly more expensive

Hypothesis 1

Houses with waterfront have higher grades and are more expensive



WF houses:
higher prices in each gradeing category

grade-11 no WF
=
grade-9 WF

Hypotheses:

1. **houses with waterfront have higher grades and are more expensive** 
2. houses with high sqft_living:
   have nighbours with high sqft_living
3. houses with higher sqft_living than their neighbours
   are much more expensive

Hypothesis 2
Houses with high sqft_living have nighbours with high sqft_living

sqft_living ↔ sqft_living15

correlation coefficient (pearson): **0,76**

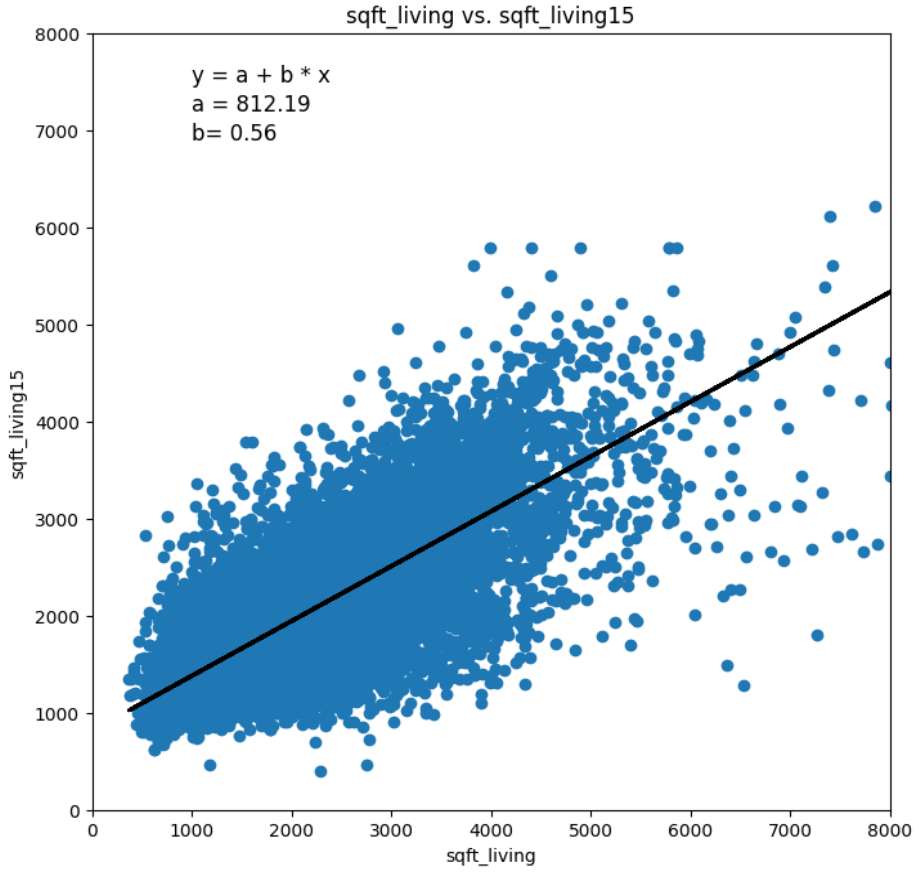
expected a higher correlation?

significant? yes!

H0 : no correlation H1: correlation
H0 rejected: correlation is statistically significant

→ positive correlation!

$$b = corr \cdot \frac{s_{sqft\ living15}}{s_{sqft\ living}}$$



Hypotheses:

1. houses with waterfront have higher grades and are more expensive 
2. **houses with high sqft_living have nighbours with high sqft_living** 
3. houses with higher sqft_living than their neighbours are much more expensive

Hypothesis 3

houses with higher sqft_living than their neighbours are much more expensive

my definition: → factor 2

Group „normal houses“

sqft_living < = 2 * sqft_living15

mean price: 532049 +/- 343897

Group „showoff houses“

sqft_living > 2 * sqft_living15

mean price: 1016086 +/- 932761

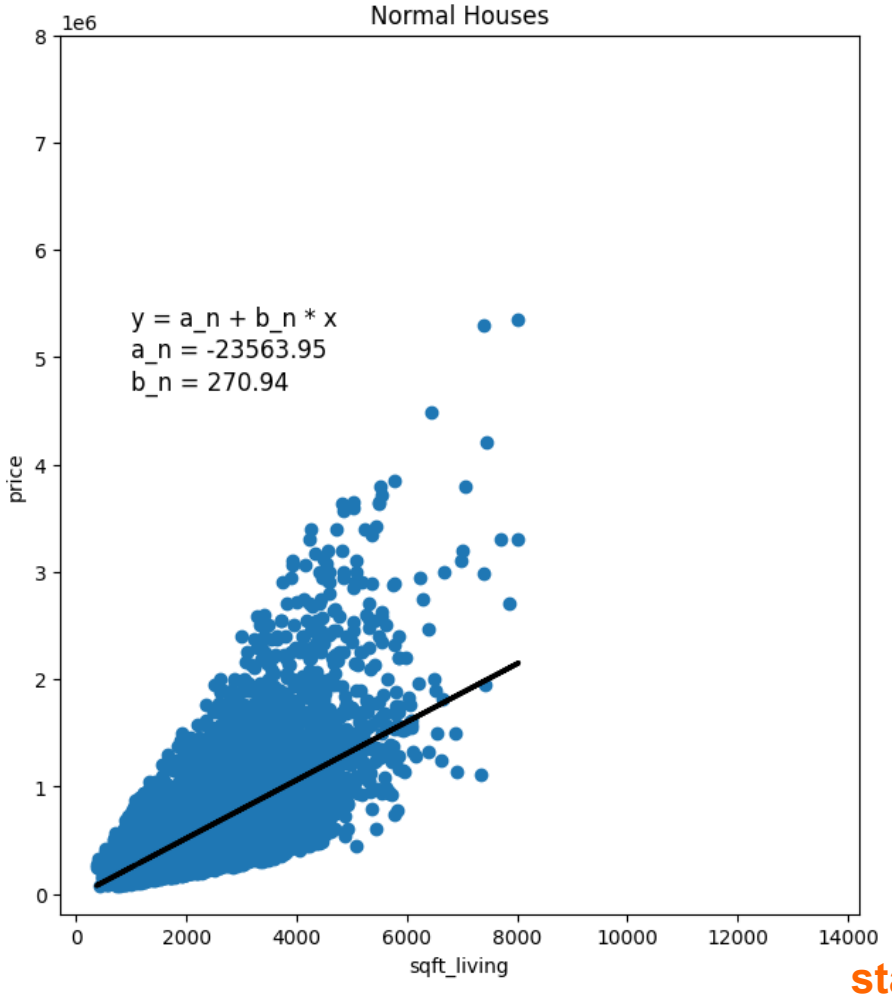
significant? yes!

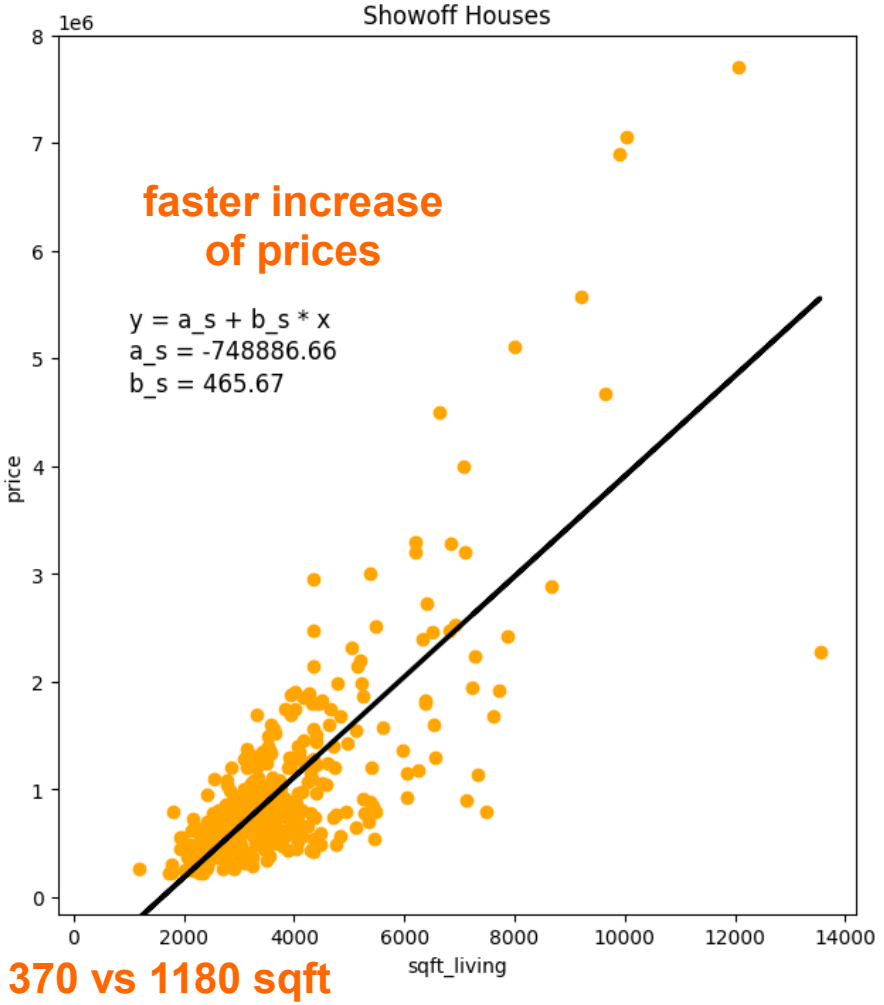
one-sided t-test: H0 : μ_s <= μ_n H1: μ_s > μ_n
alpha = 0,05 t-statistic: 9.94 p-value: 4.28 e-21
reject H0: showoff houses are significantly more expensive

Hypothesis 3

houses with higher sqft_living than their neighbours are much more expensive

my definition: → factor 2





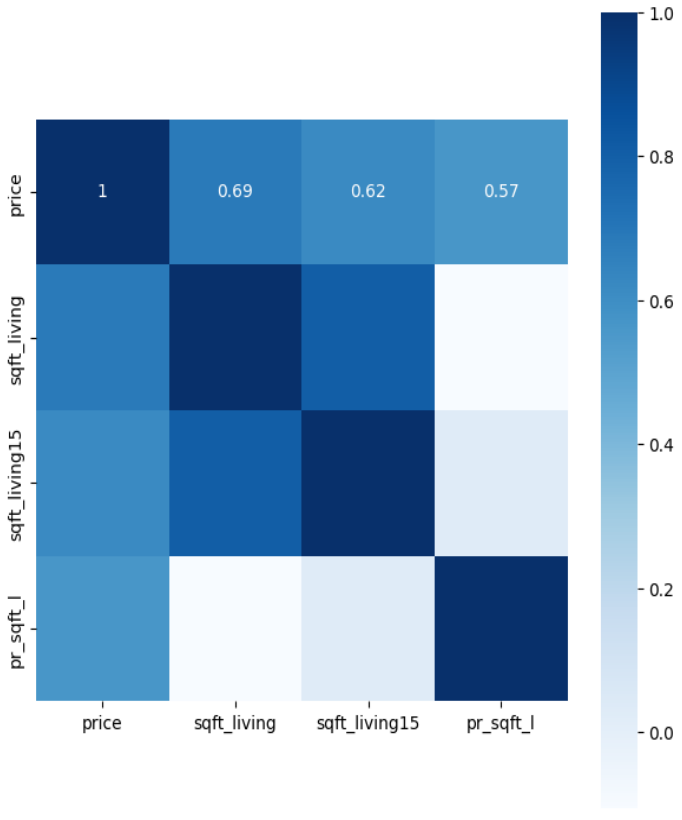
start: 370 vs 1180 sqft

Hypothesis 3

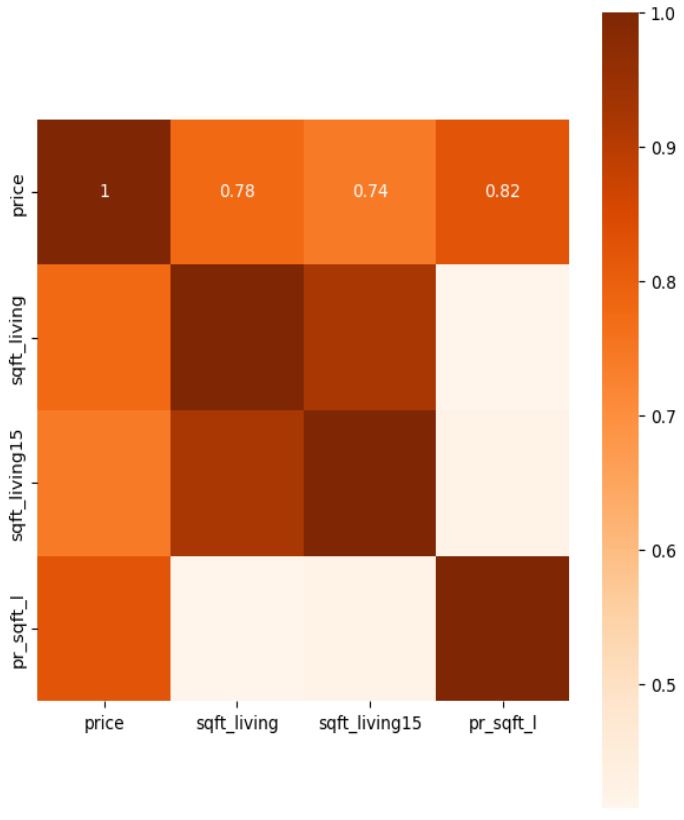
houses with higher sqft_living than their neighbours are much more expensive

my definition: → factor 2

0,69



0,78



Hypothesis 3

houses with higher sqft_living than their neighbours are much more expensive

my definition: → factor 2

what about

price per sqft_living

??

higher prices?	**even a difference?**	**lower price?**
no!	**yes!**	**yes!**
H0 : μ_s <= μ_n H1: μ_s > μ_n t-statistic: -2.77 p-value: 0.997 not reject H0: showoff houses do not have significantly higher prices per sqft_living	#H0: μ_n = μ_s H1 : μ_n != μ_s t-statistic: -2.77 p-value:0.00573 reject H0: significant difference in price per sqft_living	H0 : μ_s >= μ_n H1: μ_s < μ_n t-statistic: -2.77 p-value: 0.0028 reject H0: showoff houses have significantly lower prices per sqft_living

Hypothesis 3

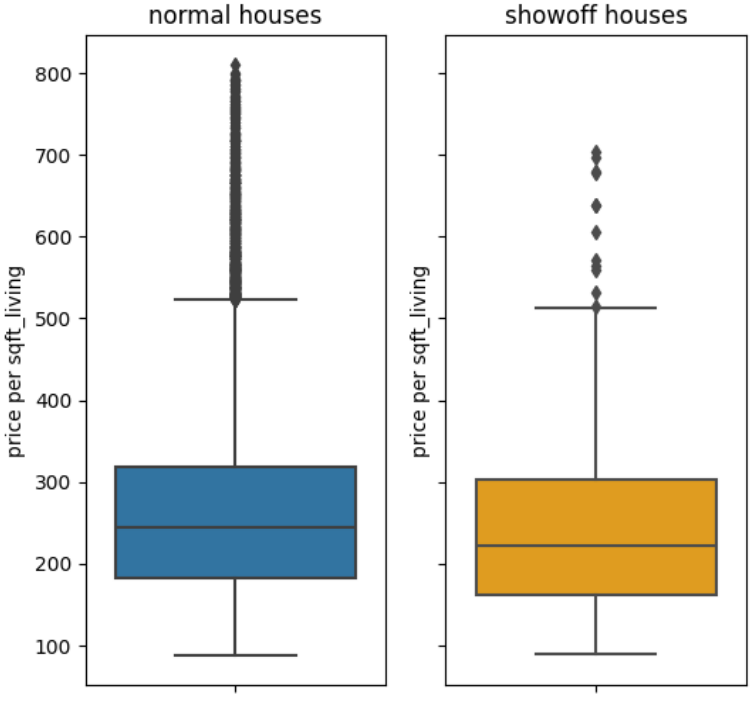
houses with higher sqft_living than their neighbours are much more expensive

my definition: → factor 2

what about

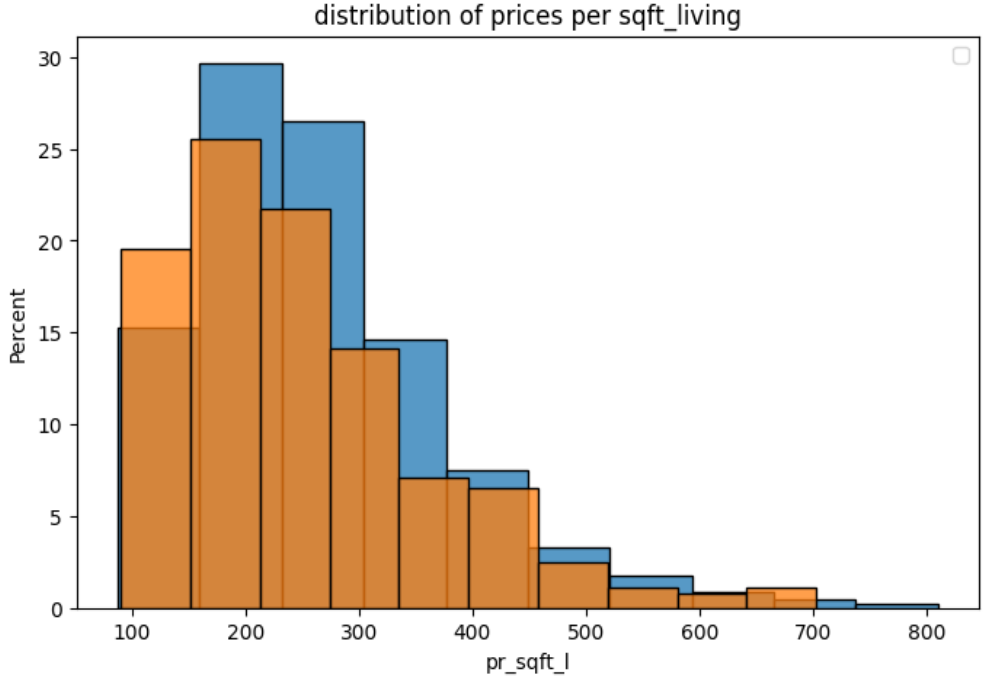
price per sqft_living

??



mean:
264 +/- 110

mean:
248 +/- 118



Hypothesis 3

houses with higher sqft_living than their neighbours are much more expensive

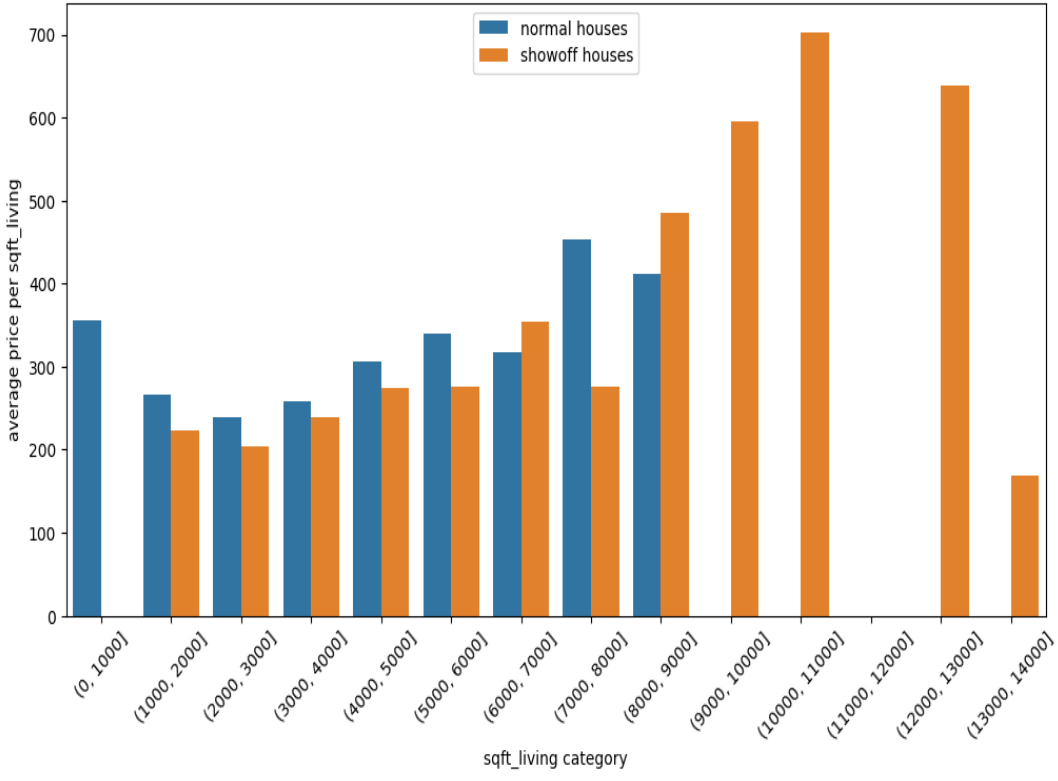
my definition: → factor 2

what about

price per sqft_living

??

outliers: after certain sqft, thery are always > 2x sqft



possible reason for lower price per sqft_living:

- because of the lower overall price of the 15 neighbours with only half as much sqft_living
- prices in the neighbourhood ?
- sqft_lot / sqft_lot15 ?

Hypotheses:

1. houses with waterfront have higher grades and are more expensive 

2. houses with high sqft_living have nighbours with high sqft_living 

3. **houses with higher sqft_living than their neighbours are much more expensive**

 **overall price**

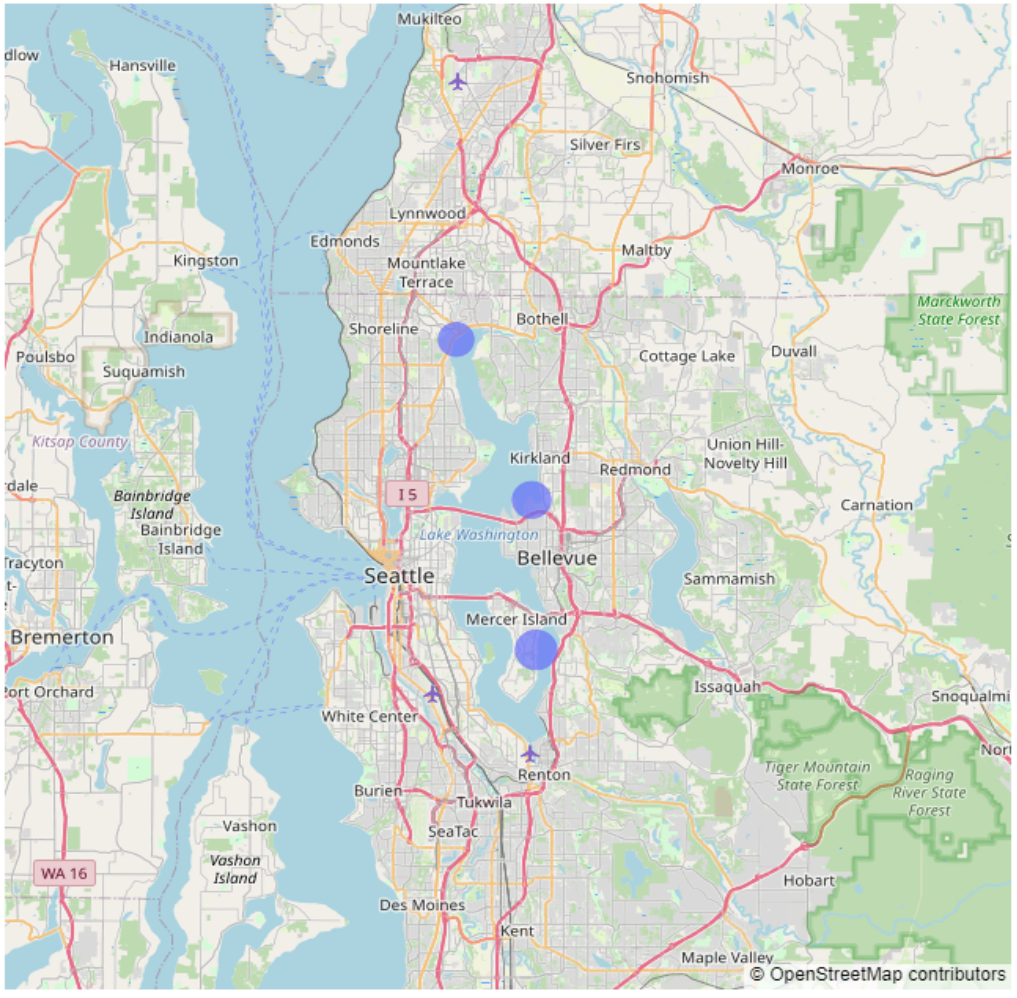
 **price per sqft_living**

Recommendations

Jennifer Montgomery

- high budget
- wants to show off → sqft_living > 2 * sqft_living15
- waterfront
- renovated → built or renovated after 2000
- high grades → better than 75% quantile

Recommendations



Object 1

price:	4 500 000
bedrooms:	5
bathrooms:	5.5
sqft_liv:	6 640
sqft_lot:	40 014
grade:	12
yr_built:	2004

Object 2

price:	7 060 000
bedrooms:	5
bathrooms:	4.5
sqft_liv:	10 040
sqft_lot:	37 325
grade:	11
yr_built:	1940
yr_renov:	2001

Object 3

price:	4 670 000
bedrooms:	5
bathrooms:	6.75
sqft_liv:	9 640
sqft_lot:	13 068
grade:	12
yr_built:	1983
yr_renov:	2009