EXPLORATORY DATA ANALYSIS

Simon Bernarding

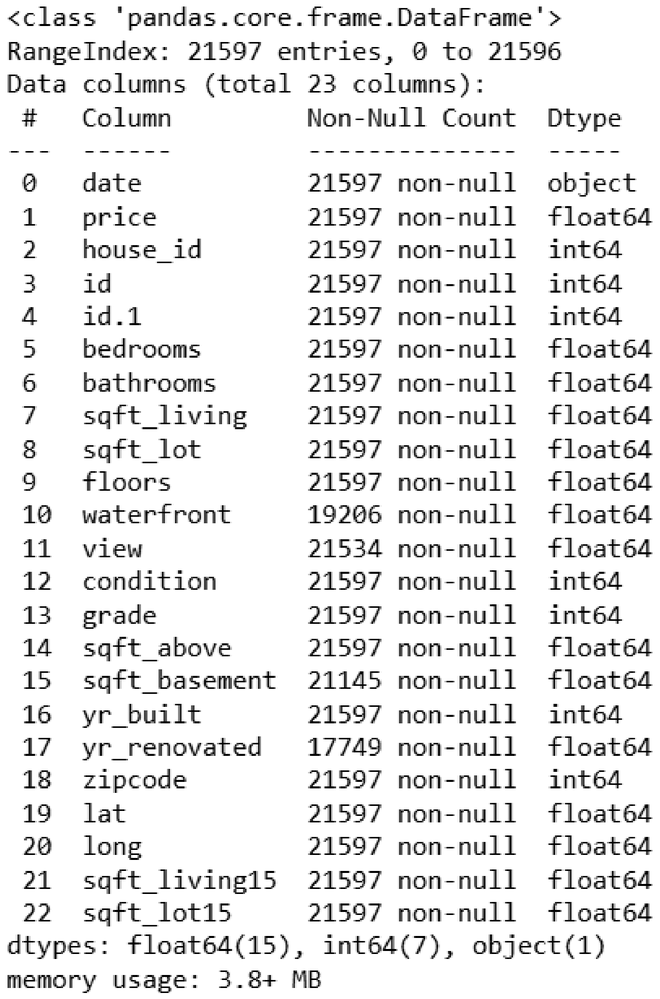
KING COUNTY (USA) HOUSING DATA (2014-15)

Client:

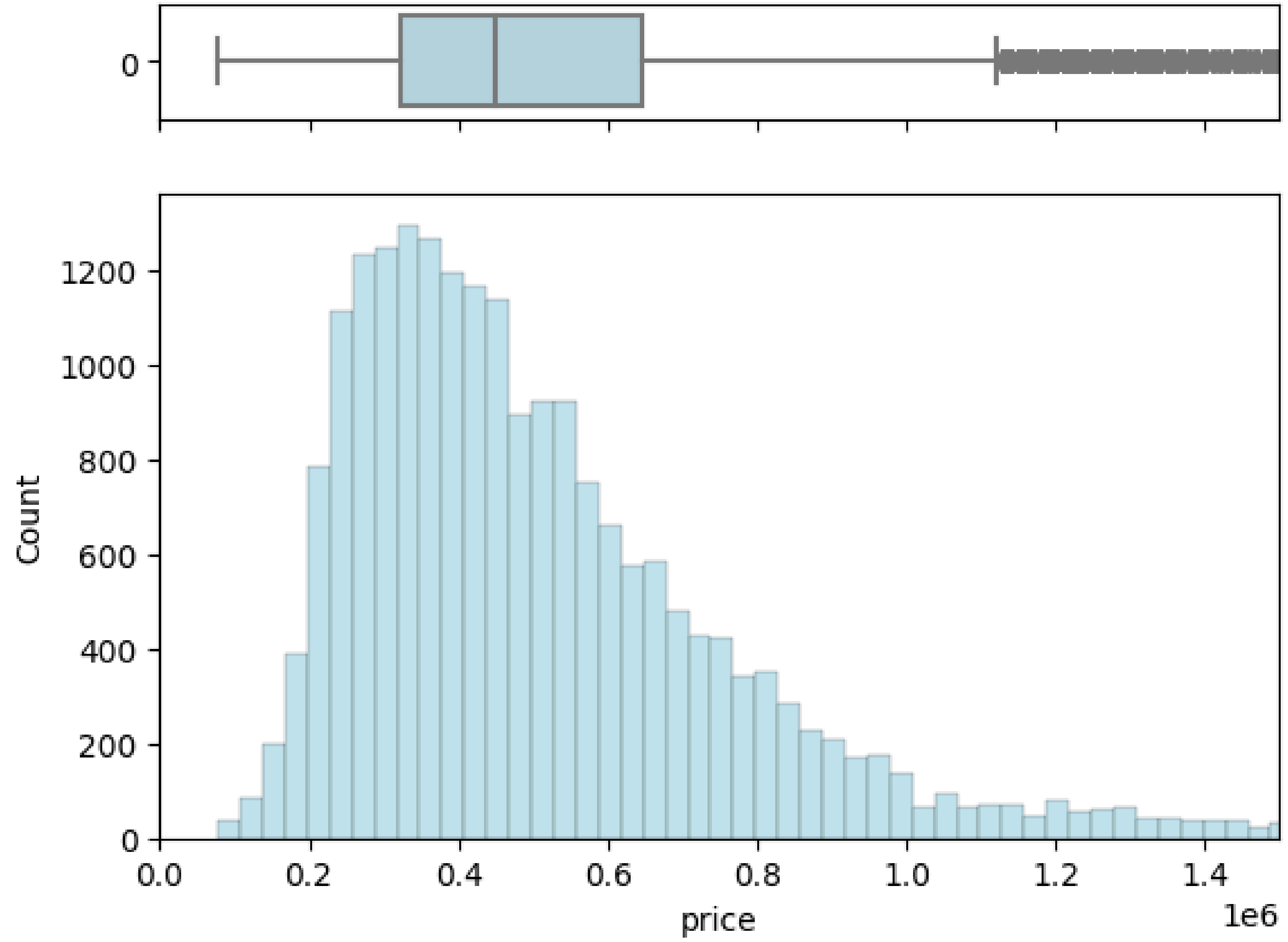
Thomas Hansen

DESCRIPTION OF DATA

- 21597 entries for data in year 2014-15
- numercial features e.g. price, sqft_living, sqft_basement
- categorical features e.g. view, condition
- geographical data: longitude, latitude



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21597 entries, 0 to 21596
Data columns (total 23 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   date           21597 non-null  object
 1   price          21597 non-null  float64
 2   house_id       21597 non-null  int64
 3   id             21597 non-null  int64
 4   id.1           21597 non-null  int64
 5   bedrooms       21597 non-null  float64
 6   bathrooms      21597 non-null  float64
 7   sqft_living    21597 non-null  float64
 8   sqft_lot       21597 non-null  float64
 9   floors         21597 non-null  float64
 10  waterfront     19206 non-null  float64
 11  view           21534 non-null  float64
 12  condition      21597 non-null  int64
 13  grade          21597 non-null  int64
 14  sqft_above     21597 non-null  float64
 15  sqft_basement  21145 non-null  float64
 16  yr_built       21597 non-null  int64
 17  yr_renovated   17749 non-null  float64
 18  zipcode        21597 non-null  int64
 19  lat            21597 non-null  float64
 20  long           21597 non-null  float64
 21  sqft_living15  21597 non-null  float64
 22  sqft_lot15     21597 non-null  float64
dtypes: float64(15), int64(7), object(1)
memory usage: 3.8+ MB
```



General statistics of price:

mean	5.402749e+05
std	3.667199e+05
min	7.800000e+04
25%	3.220000e+05
50%	**450.000 $**
75%	6.450000e+05
max	7.700000e+06

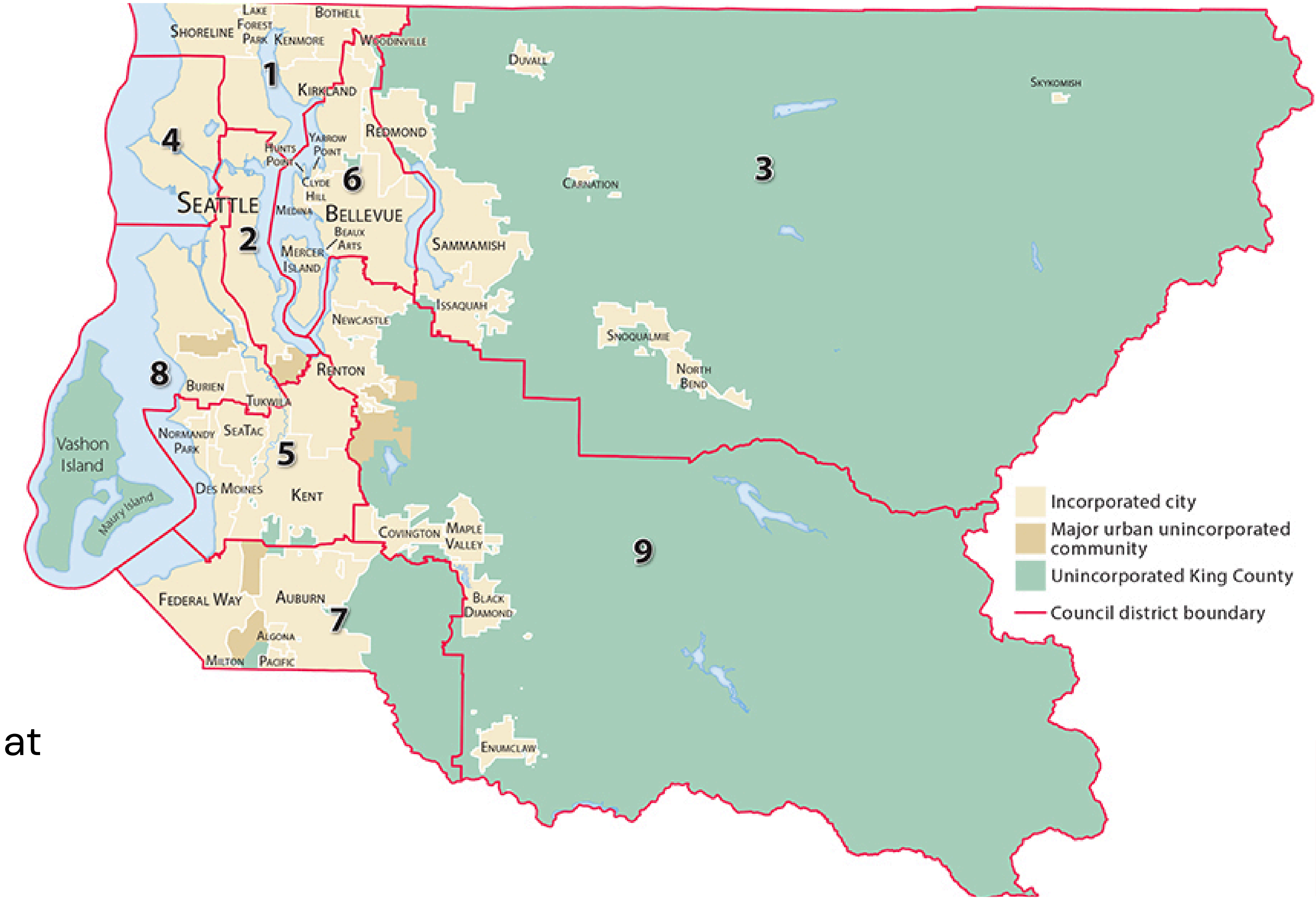
CLIENT

Description:

Buyer, 5 kids, no money, wants nice (social) neighborhood, Timing?, Location?

Assumptions:

- no money: as cheap as possible
- 5 kids: the bigger the house, the better; at least 3 or 4 bedrooms needed
- nice (social) neighborhood: not in the city center, not in the middle of nowhere
- timing: buy cheap during year

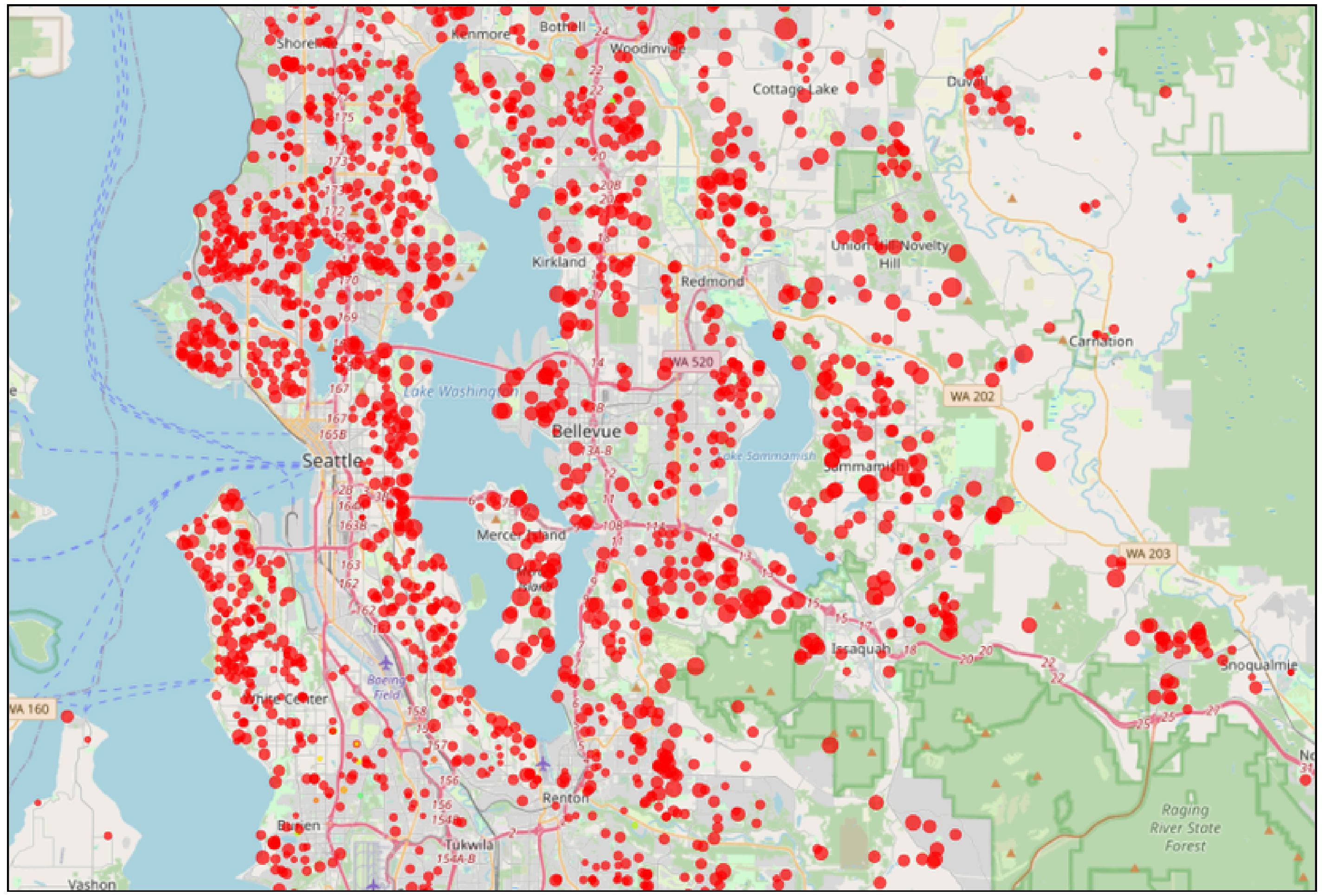


HYPOTHESIS

- Properties with waterfront have higher prices.
- Houses with lower grading are less expensive.
- Houses in the city center are more expensive.

CLEANING THE DATA

- missing values in waterfront : 11.07 %
- missing values in view : 0.29 %
- missing values in sqft_basement : 2.09 %
- missing values in yr_renovated : 17.82 %
- missing values in data frame : 1.49 %



- not located at water and not relevant for client
- **set missing values to “no waterfront”**

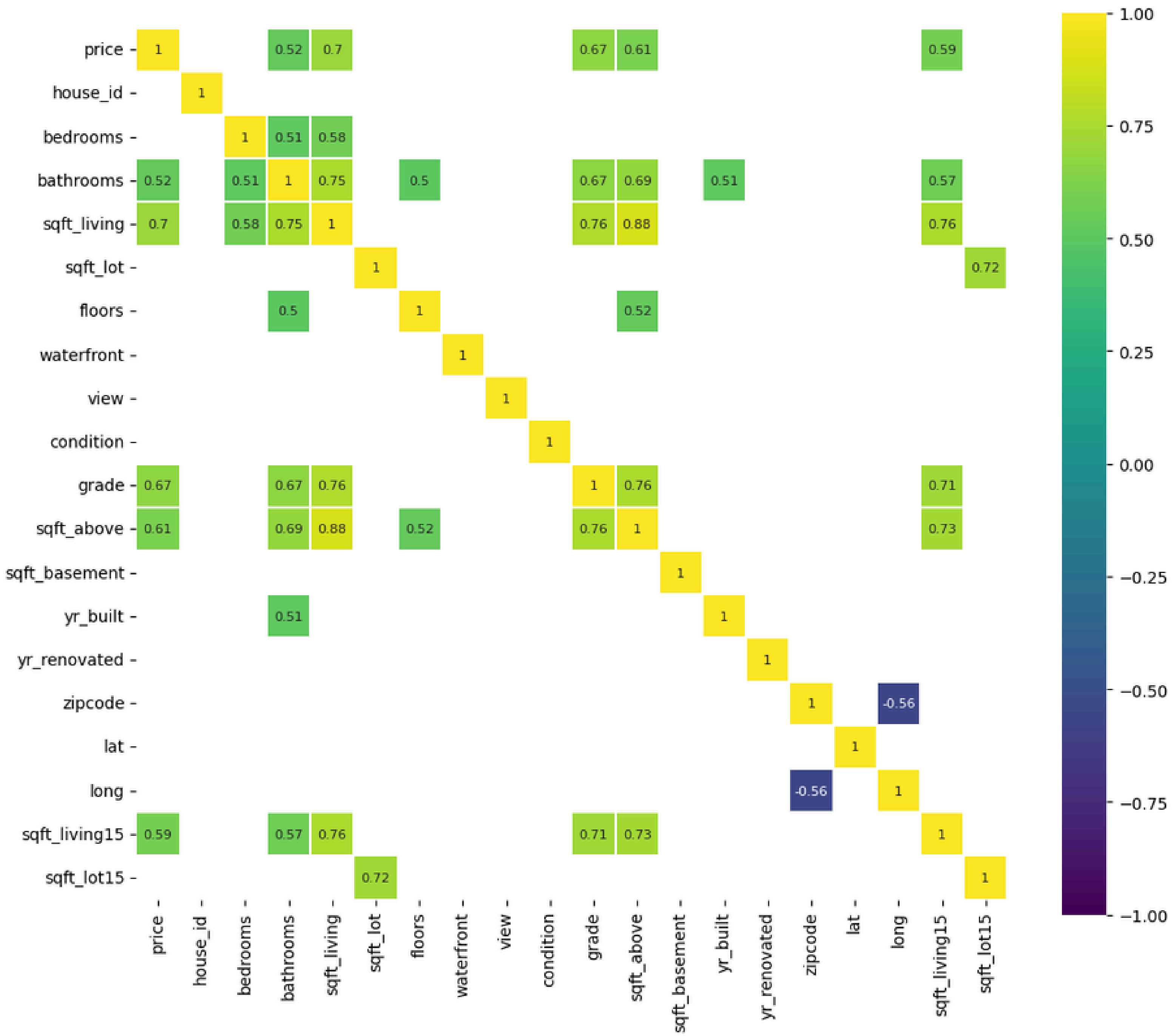
- relative small % and not relevant for client
- **drop rows**

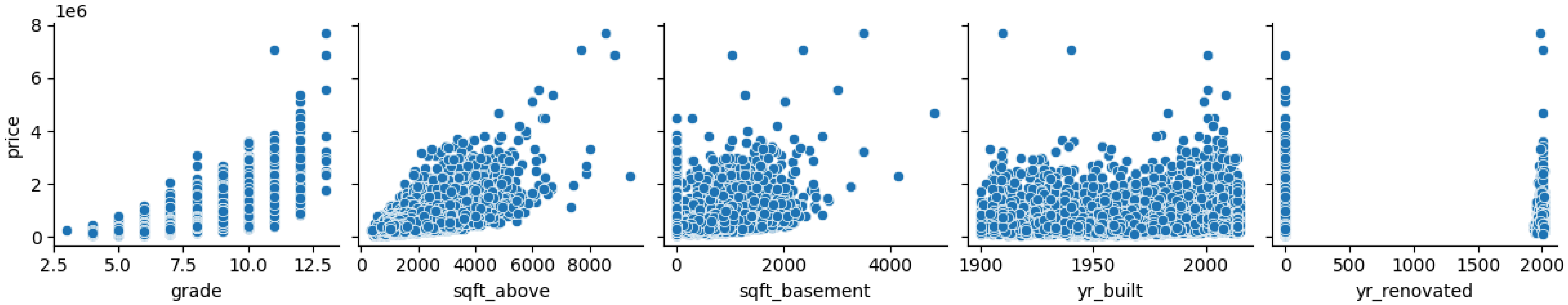
- compare missing values with the ones from houses renovated and not
- similar statistics (ie. median price) with houses not renovated
- information not relevant for client
- **set missing values to “not renovated”**

ANALYSIS

Strongly correlating parameters with price:

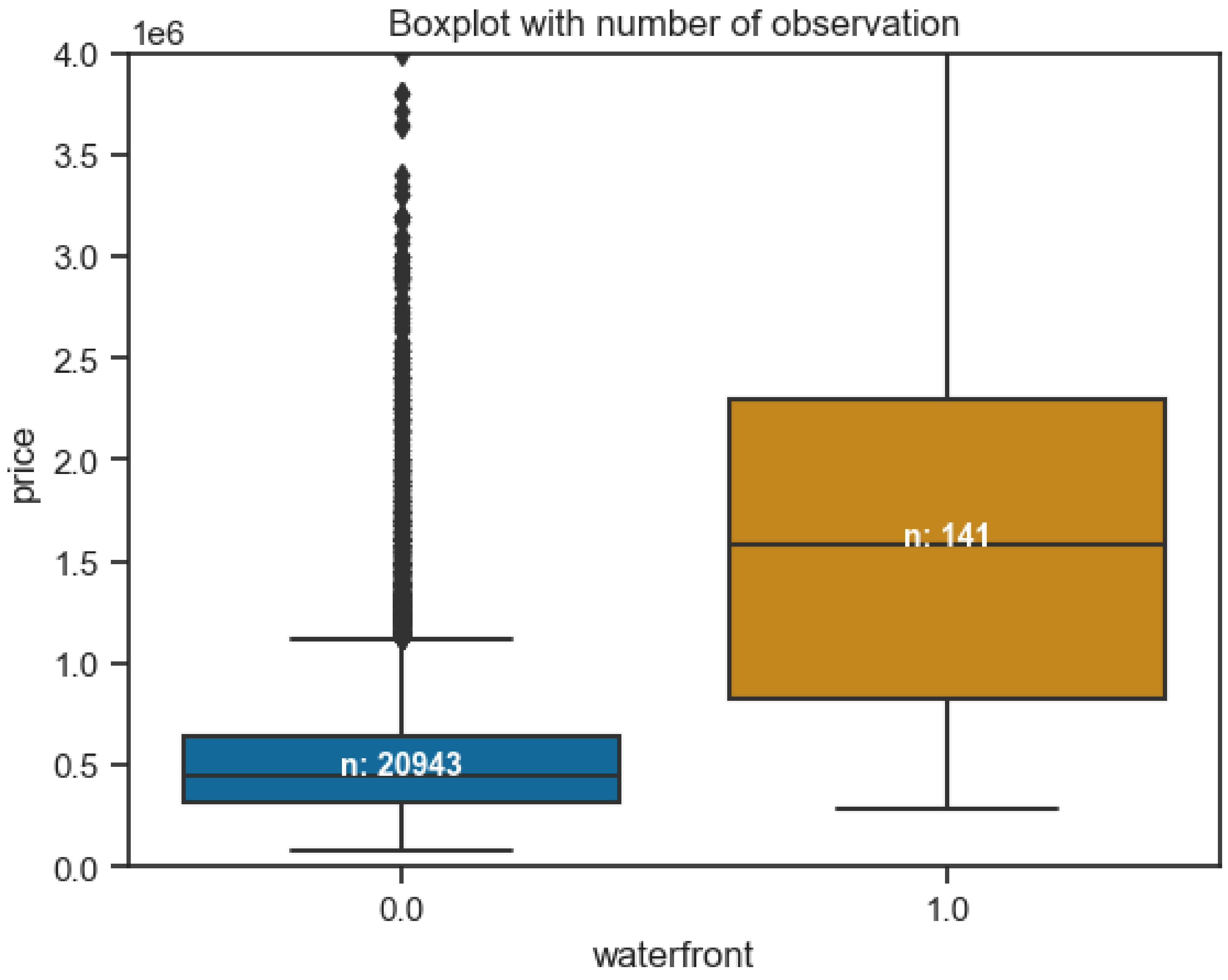
sqft_living	0.701899
grade	0.668031
sqft_above	0.605388
sqft_living15	0.586420
bathrooms	0.524849





FINDINGS

- Properties with waterfront have higher prices! ✓



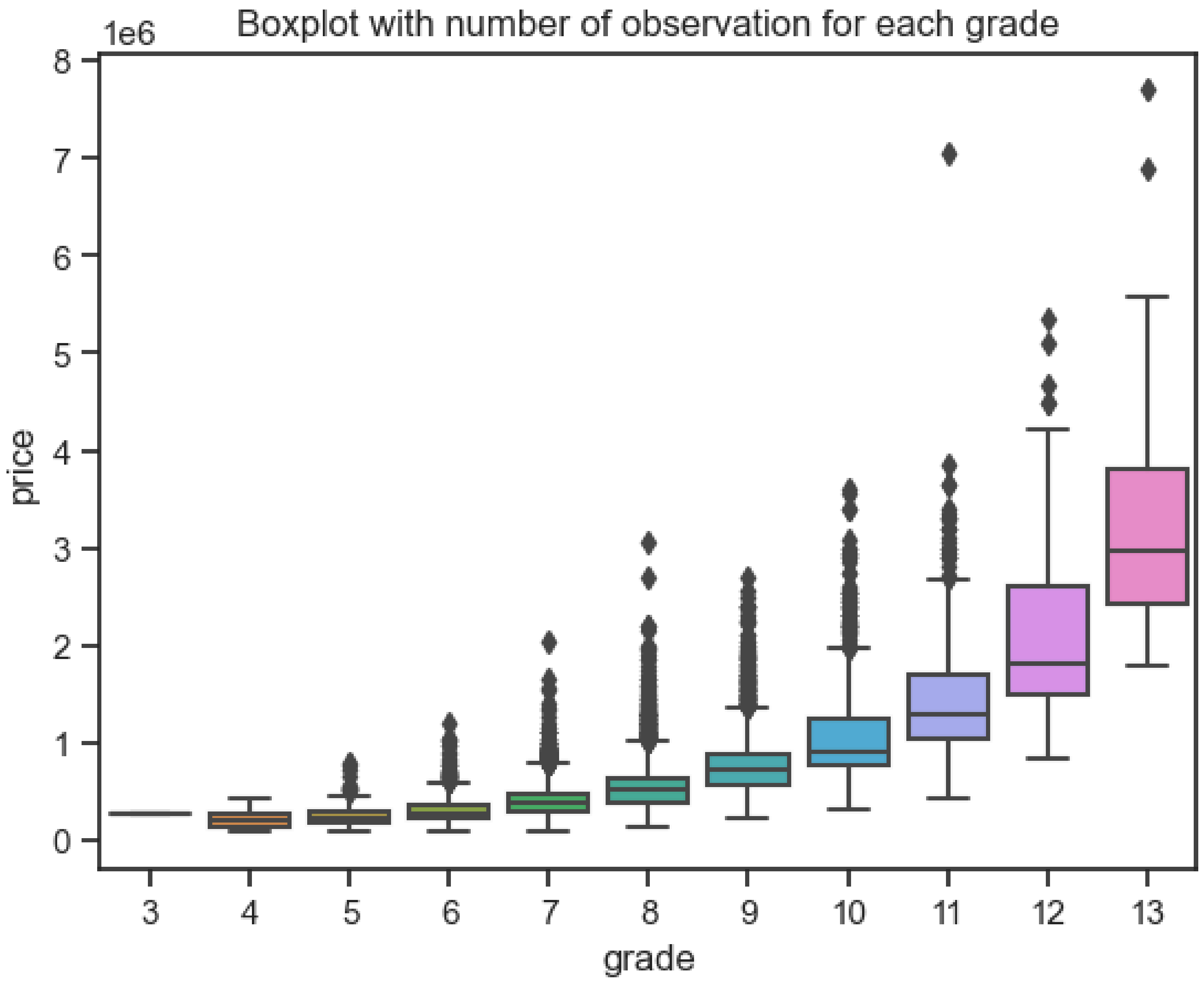
Median

without waterfront: 450.000 $

with waterfront: 1.580.000 $

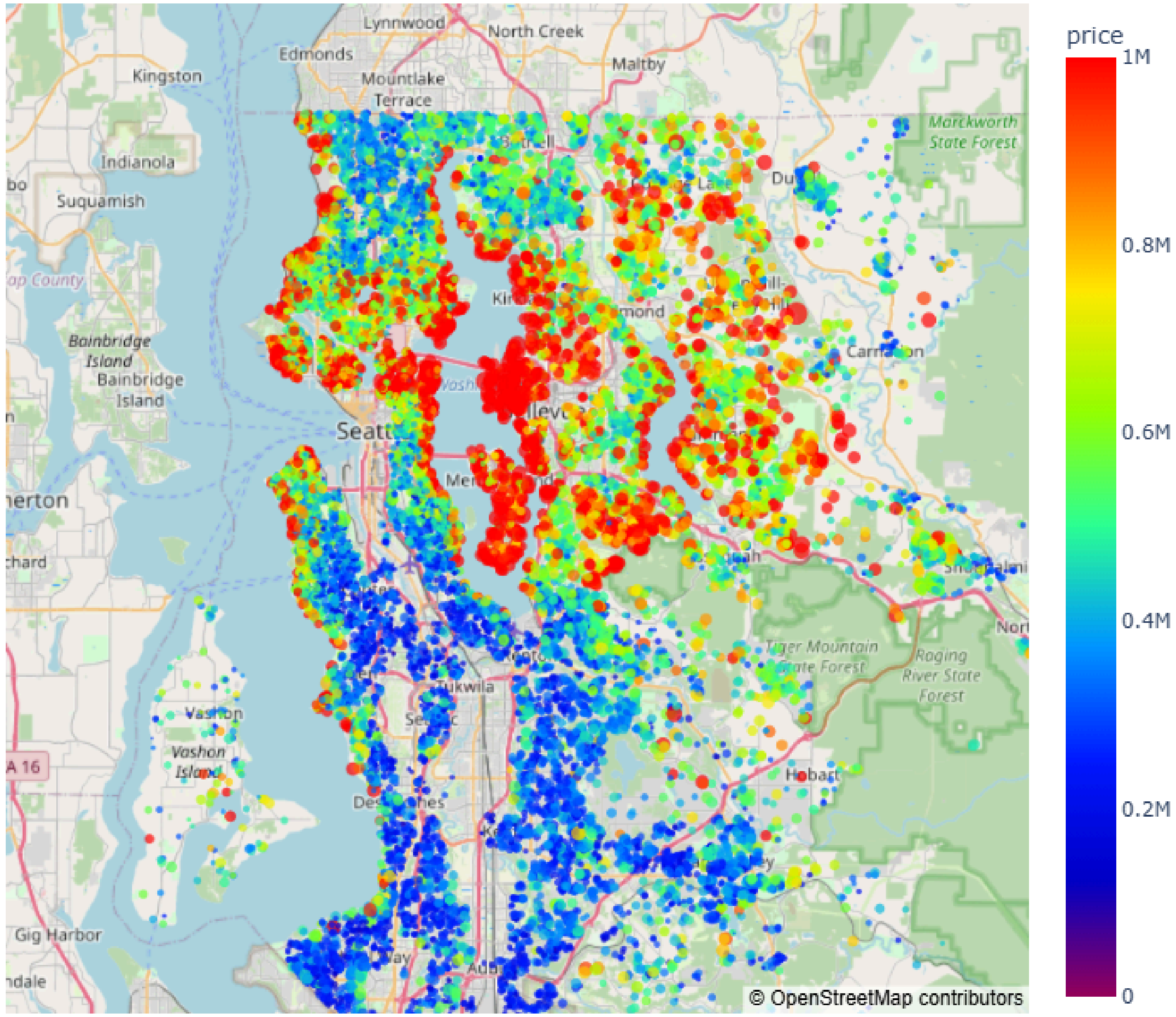
FINDINGS

- Houses with a lower grading are less expensive! ✓



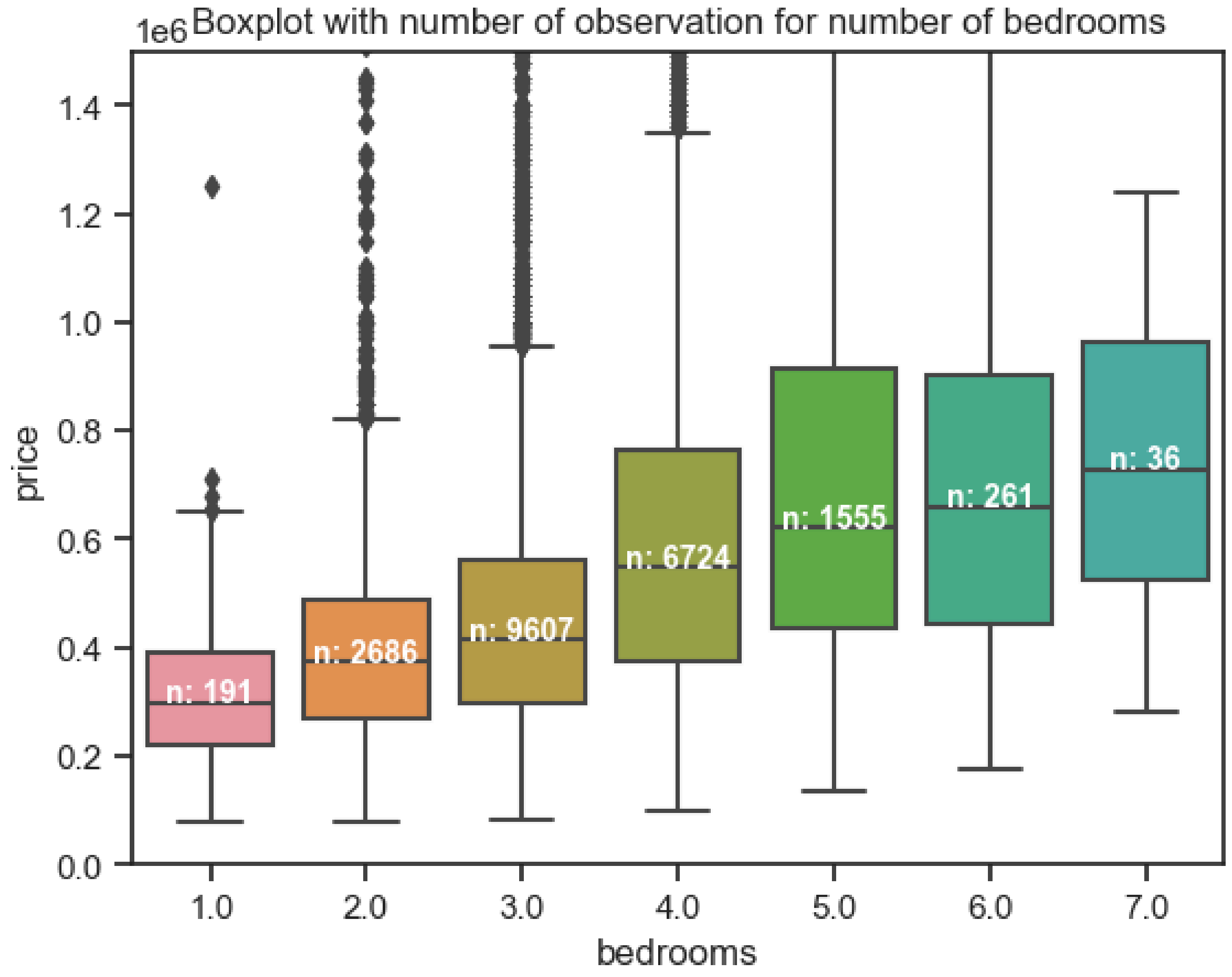
FINDINGS

- Houses in the city center are more expensive! ✓



RECOMMENDATION FOR CLIENT

- 5 kids: the bigger the house, the better; at least 4 bedrooms needed.

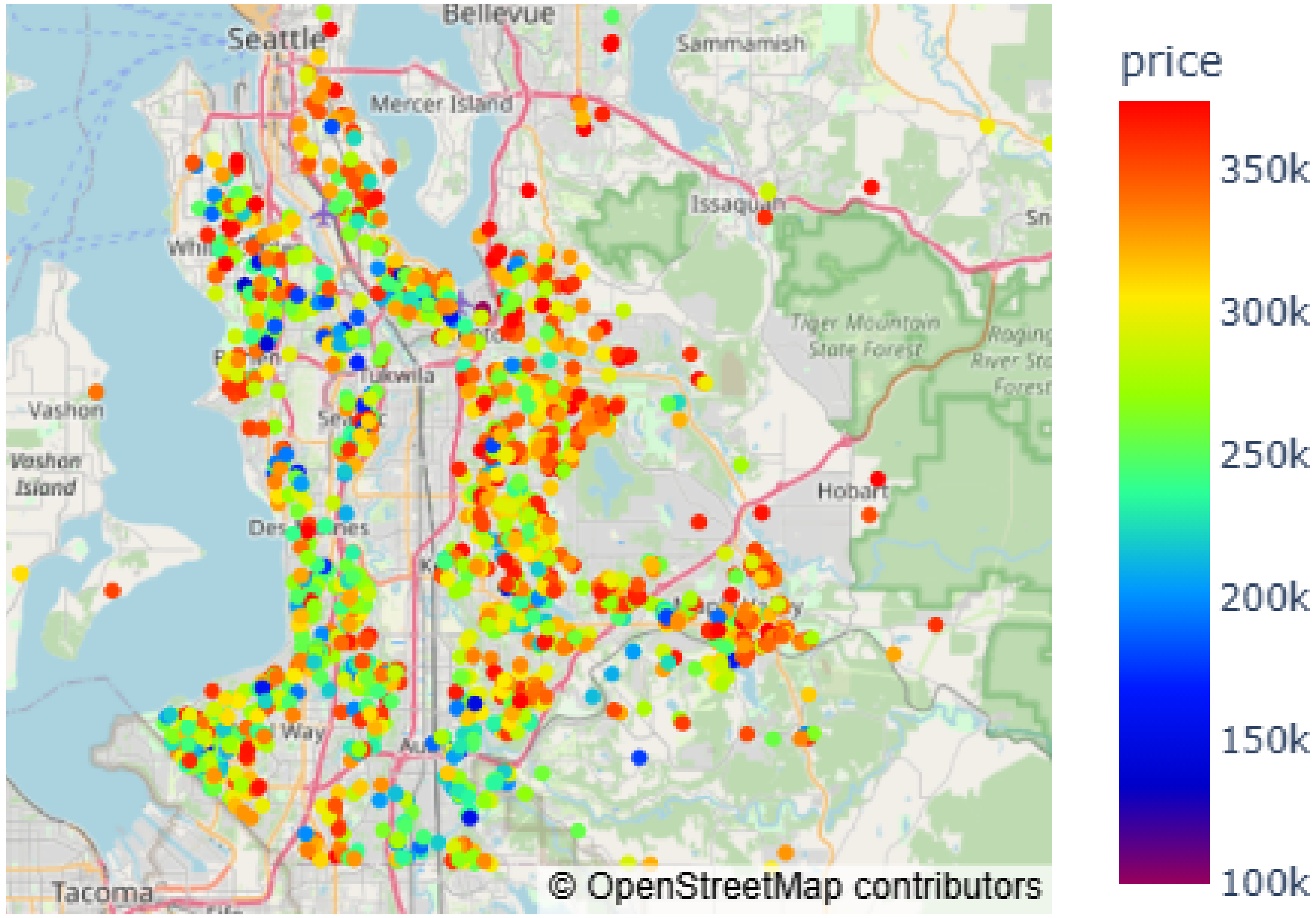

Boxplot with number of observation for number of bedrooms

bedrooms	count	min	25%	50%	75%	max
4.0	6724	100.000	**375.000 $**	549.000	765.000	4.490.000

- house with **4** bedrooms preferred
- up to **375k $** (no money)

RECOMMENDATION FOR CLIENT

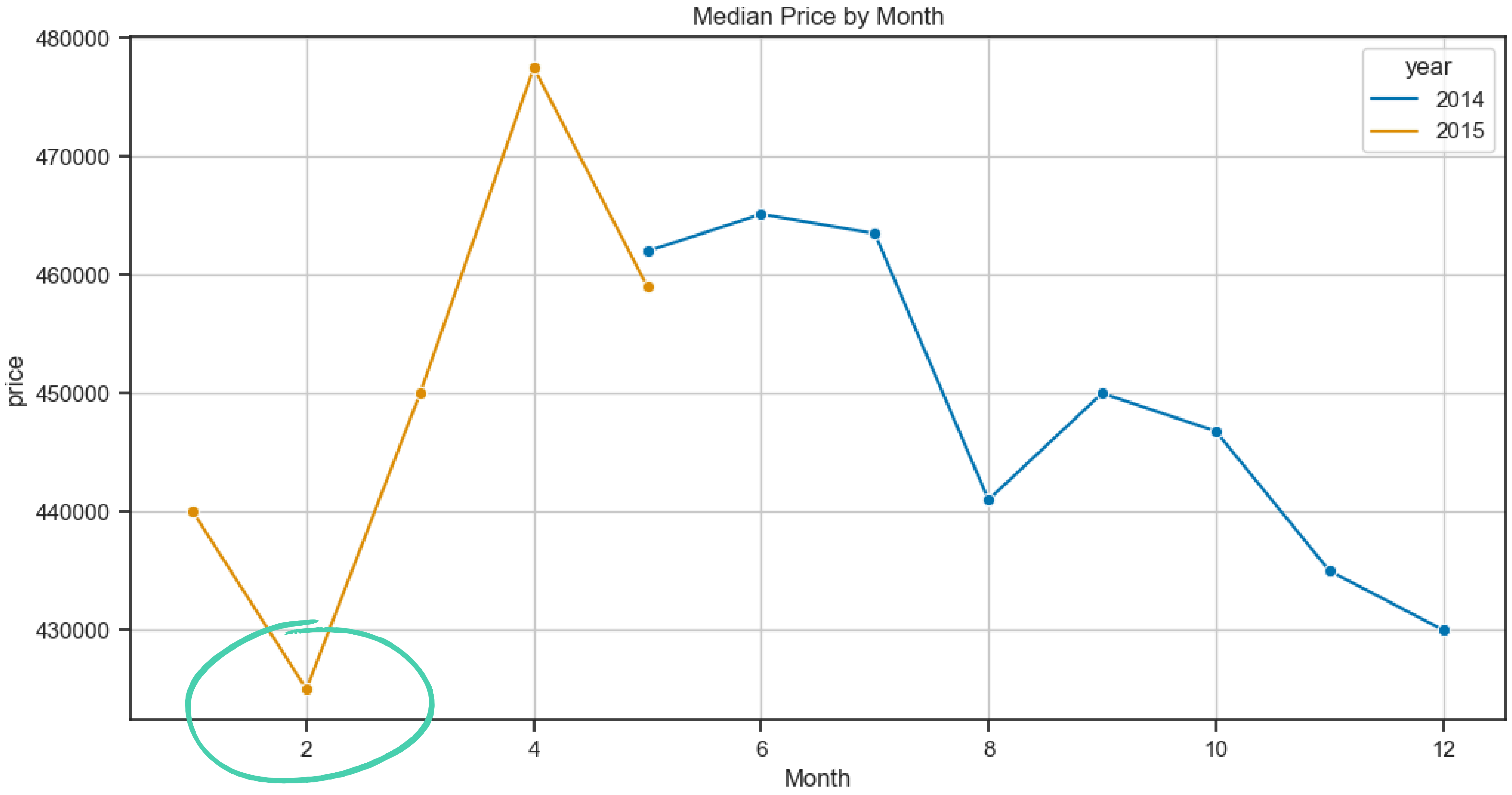
- Nice (social) neighborhood: not in the city center, not in the middle of nowhere.



- check **south and southwest** for houses

RECOMMENDATION FOR CLIENT

- Timing: buy cheap during year.



- buy house in **february**

THANK YOU