Exploratory Data Analysis

King County Housing Data

table of content

self-introduction

Daniela Mueller
Passiome Ltd, Vancouver, CA

Our Passion - your new home

stakeholder introduction

Larry Sanderson

“Waterfront, limited budget, nice & isolated, but central neighborhood without kids (but got some of his own, just doesn't his kids to play with other kids .. because of germs)”

01 | central location

02 | isolated location

03 | has children

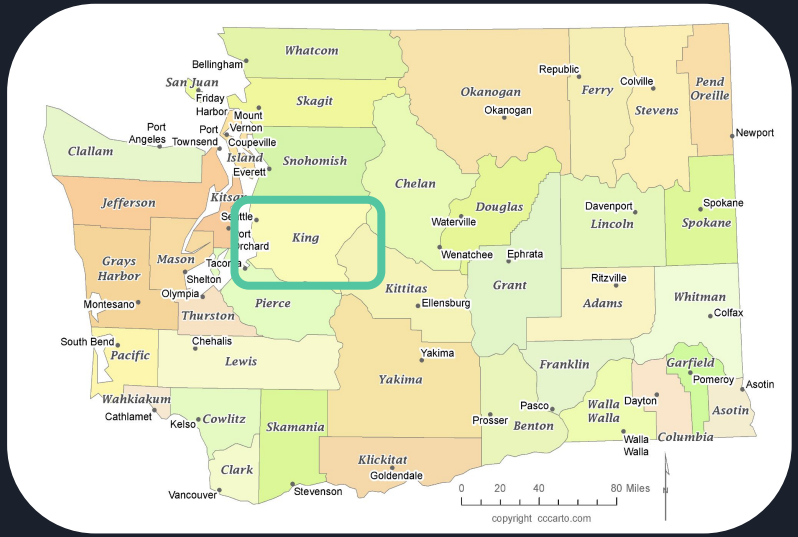
04 | waterfront

05 | limited budget

data introduction

housing data set from King County
(Washington, United States)

1. official housing data set of King County
2. ~21.600 data entries
3. contains data from may 2014 to may 2015
4. 20 features describing the housing situation*



* incl. but not limited to price, living area, bedrooms, bathrooms, floor, longitude, latitude, lot size, grade, etc.

data overview and preparation

01 | getting insight into the data

02 | exploring the data

03 | cleaning and transforming the data

04 | correlations in the data

03a | Transforming the data

01 | price_per_sqft_living

02 | price_per_sqft_loft

03 | price_group_sqft_living

04 | age

05 | age_renovation

06 | age_bin

07 | age_renovation_bin

hypotheses introduction

hypotheses overview

1. house pricing is clustered by ZIP codes / grouped by longitude & latitude

2. the age of an house impacts the official grade and the score for the overall condition

3. the number of bedrooms has a higher impact on the house pricing than the number of bathrooms

hypothesis 01

house pricing is clustered by ZIP codes / grouped by longitude & latitude

potential impact on stakeholder recommendation:

05 | limited budget

01 | central location
02 | isolated location

hypothesis 01 - analysis

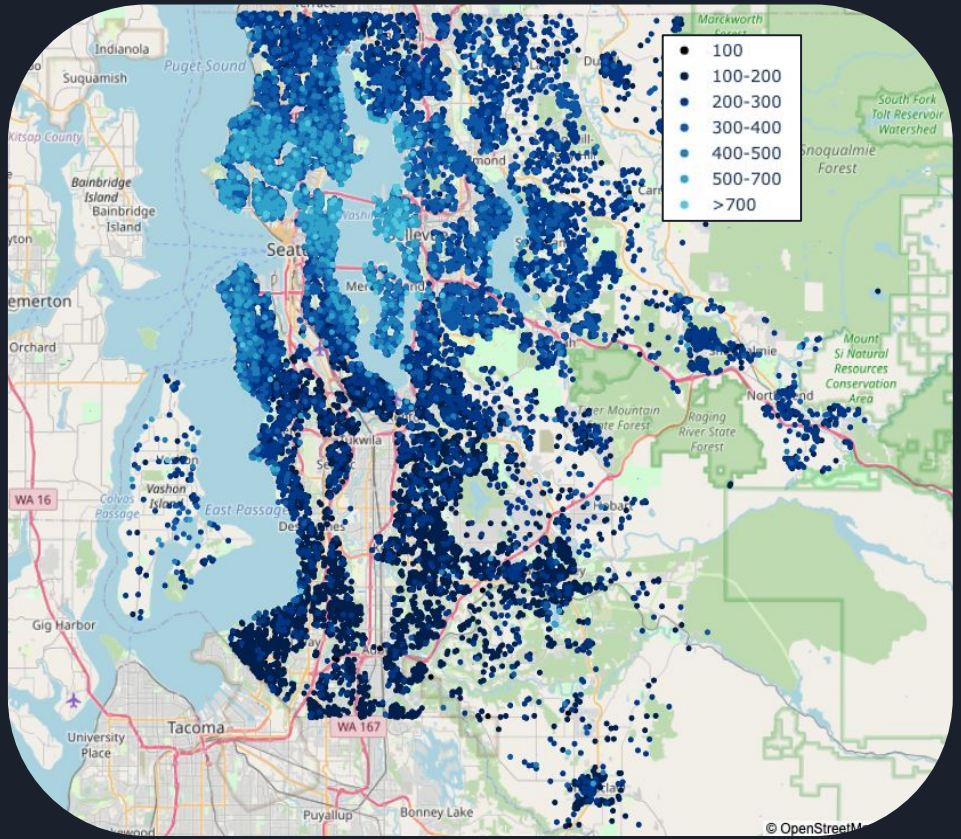
methodology:
evaluating the living area price range per sqft of all entries

findings:
01 | expensive cluster around the city center

02 | expensive cluster along the waterfront

03 | outskirts are less expensive

04 | rural areas are less expensive



hypothesis 01 - analysis

methodology:
evaluating the average living price per sqft on ZIP Code level

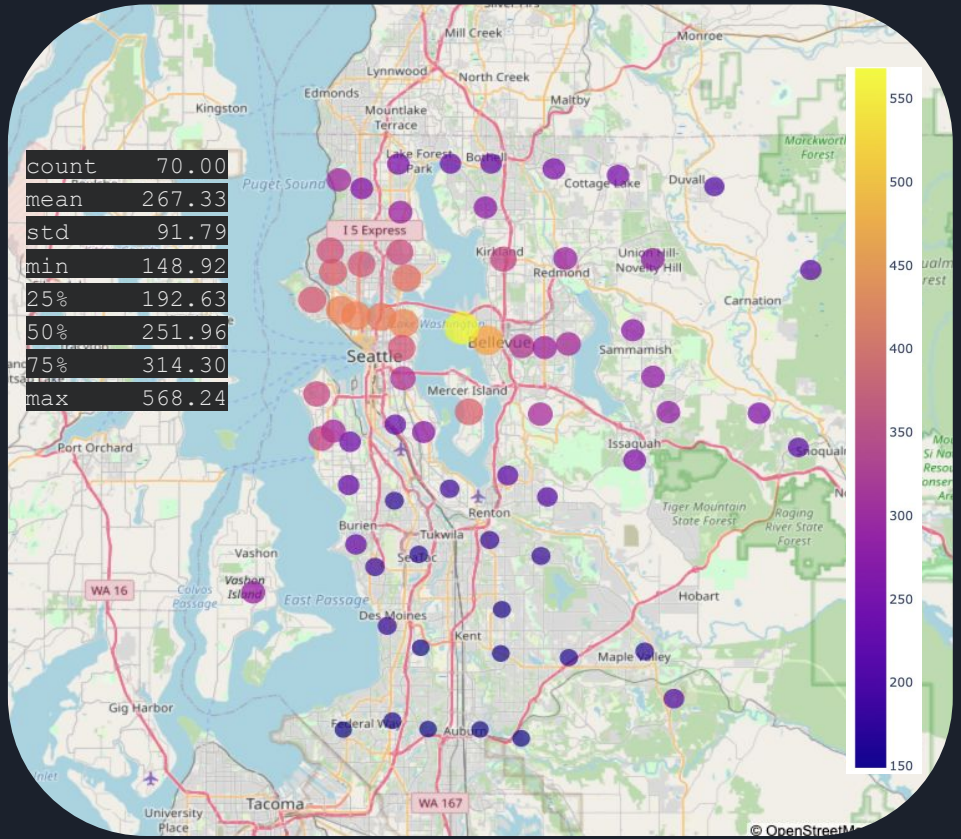
findings:
01 | Medina most expensive city, not Seattle

02 | Seattle city center expensive hotspot

03 | expensive cluster along the waterfront

04 | the further outside, the cheaper

05 | souther city district is less expensive than the countryside



CONFIRMED

hypothesis 01 - result

house pricing is clustered by ZIP codes / grouped by longitude & latitude

- *higher priced clusters in the central area*
- *higher priced clusters at the waterfront*
- *lower priced areas on the outskirts*

hypothesis 02

the age of an house impacts the official grade and

the score for the overall condition

potential impact on stakeholder recommendation:

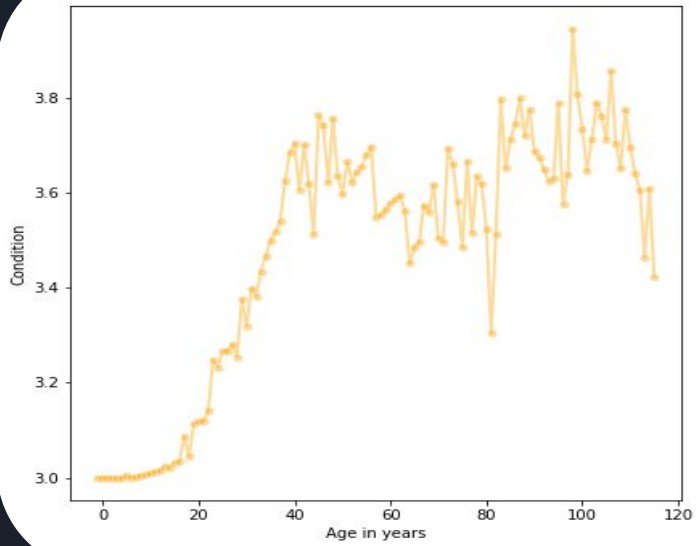
05 | limited budget

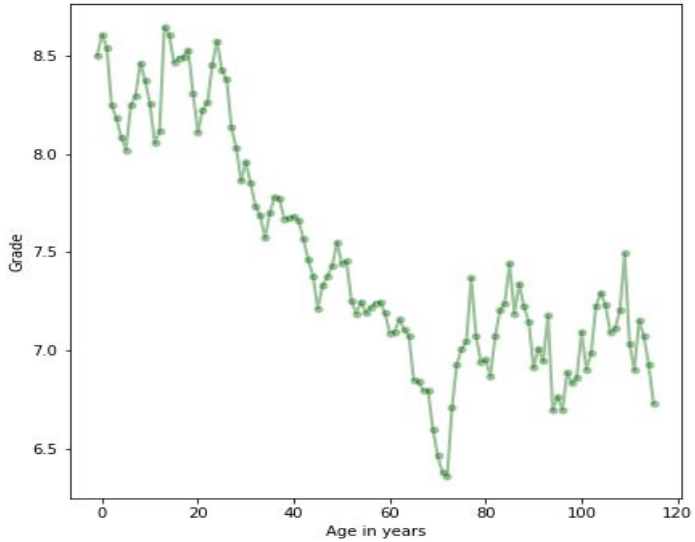
hypothesis 02 - Analysis

methodology:
evaluating the development of the condition and grade by age

findings:
01 | the condition of the houses increases with age

findings:
02 | the grade of the houses decreases with age





hypothesis 02 - Analysis

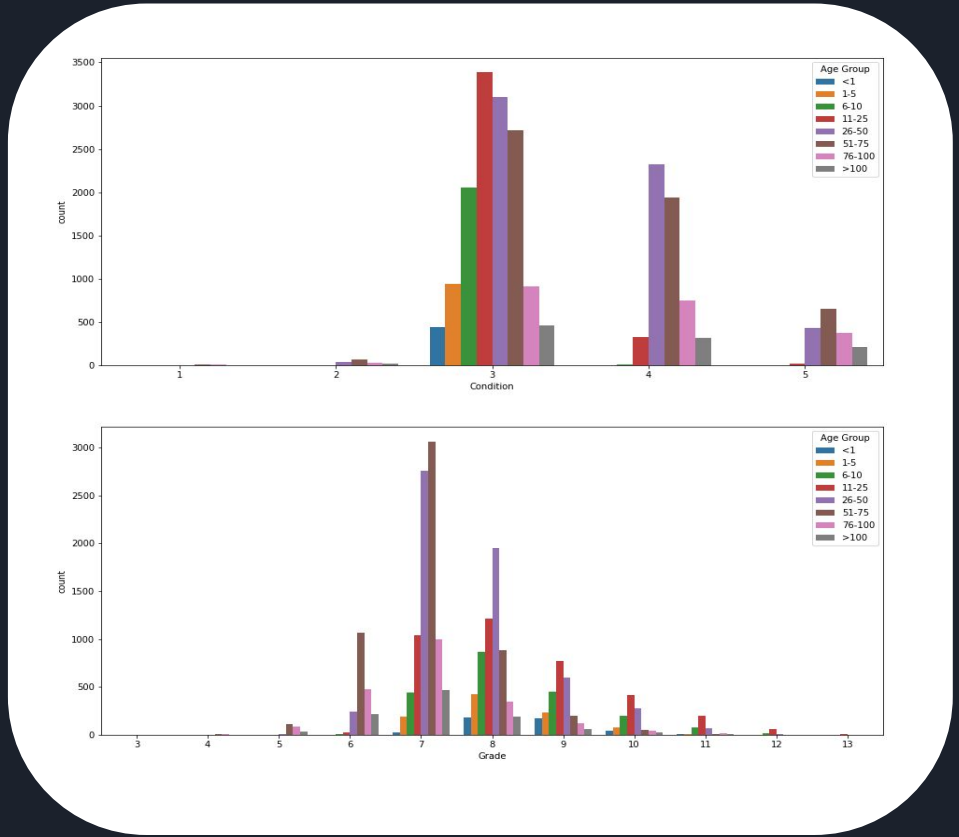
methodology:
evaluating the distribution of condition and grade by age group

findings:
01 | newer houses (under 25) are clustered around the condition 3

02 | older houses (above 25) have a condition of 3 to 5

03 | newer houses have a grade ranging from 7 and 12

04 | older houses have a grade of 6 and 9



CONFIRMED

hypothesis 02 - result

the age of an house impacts the official grade and the score for the overall condition

age impacts both grade and condition

contrary trend:

- *positive impact on condition*
- *negative impact on grade*

hypothesis 03

the number of bedrooms has a higher impact on the house pricing than the number of bathrooms

potential impact on stakeholder recommendation:

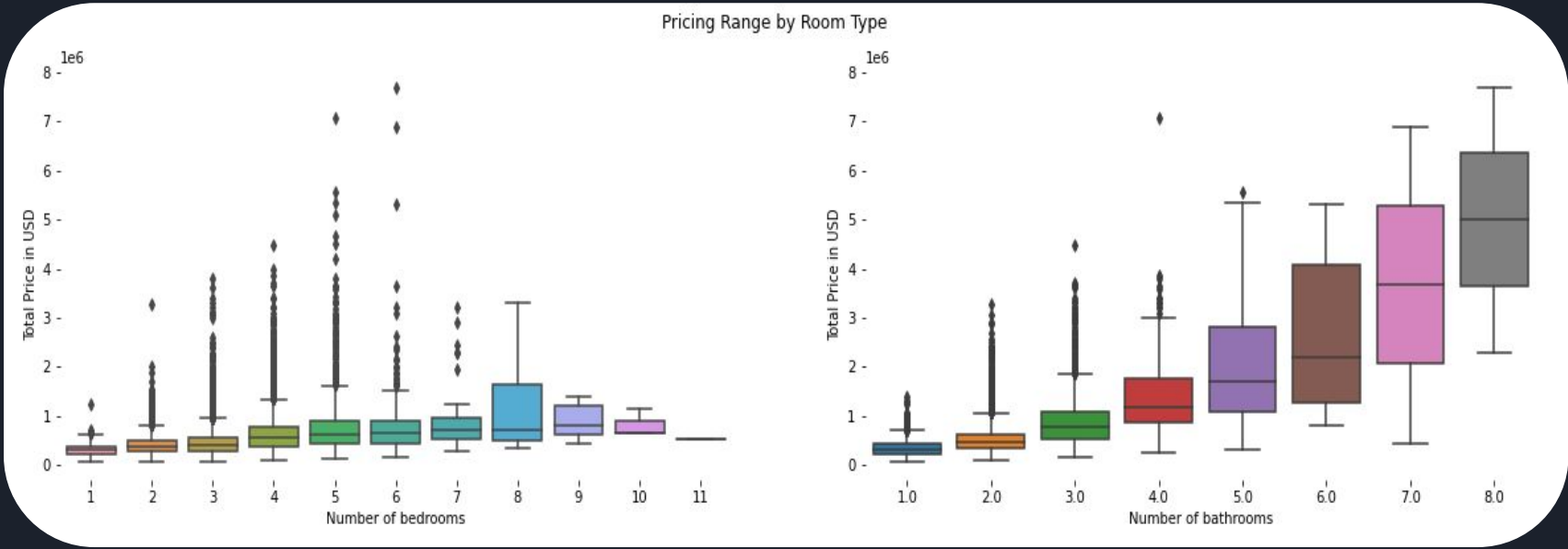
03 | has children
05 | limited budget

hypothesis 03 - Analysis

evaluating the total price development by room type

01 | total price increases not as drastically with number of bedrooms

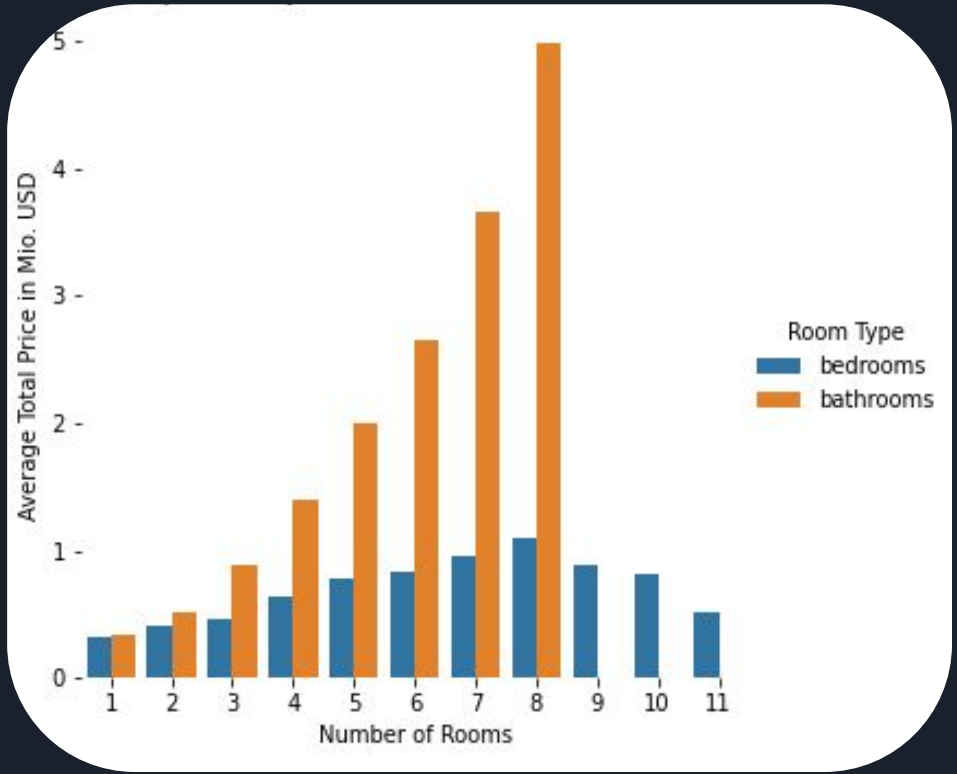
02 | increasing number of bathrooms drives the total price of house



hypothesis 03 - Analysis

comparison of average total price grouped by number of rooms

01 | average total price increases more drastically be the number of bathrooms



DENIED

hypothesis 03 - result

house pricing is clustered by ZIP codes / grouped by longitude & latitude

- *increasing amount of bathrooms leads to a higher average price*

Stakeholder situation

Larry Sanderson

“Waterfront, limited budget, nice & isolated, but central neighborhood without kids (but got some of his own, just doesn't his kids to play with other kids .. because of germs)”

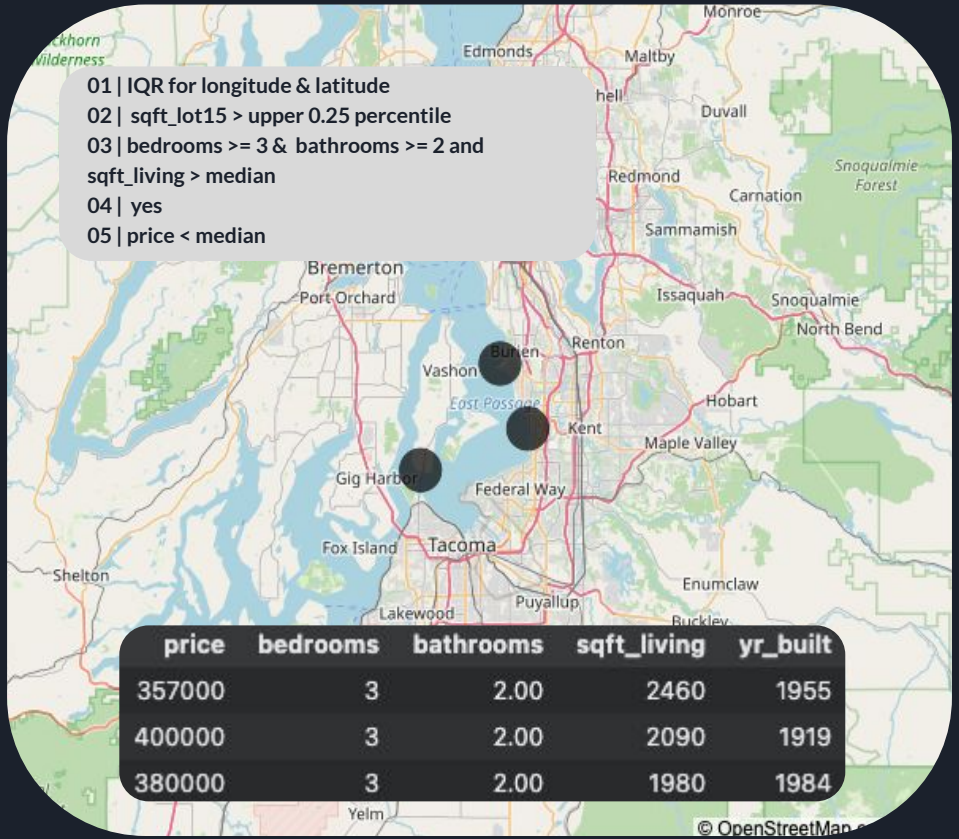
01 | central location

02 | isolated location

03 | has children

04 | waterfront

05 | limited budget



price	bedrooms	bathrooms	sqft_living	yr_built
357000	3	2.00	2460	1955
400000	3	2.00	2090	1919
380000	3	2.00	1980	1984

Stakeholder recommendations

Recommendation for adjusting the search pattern:

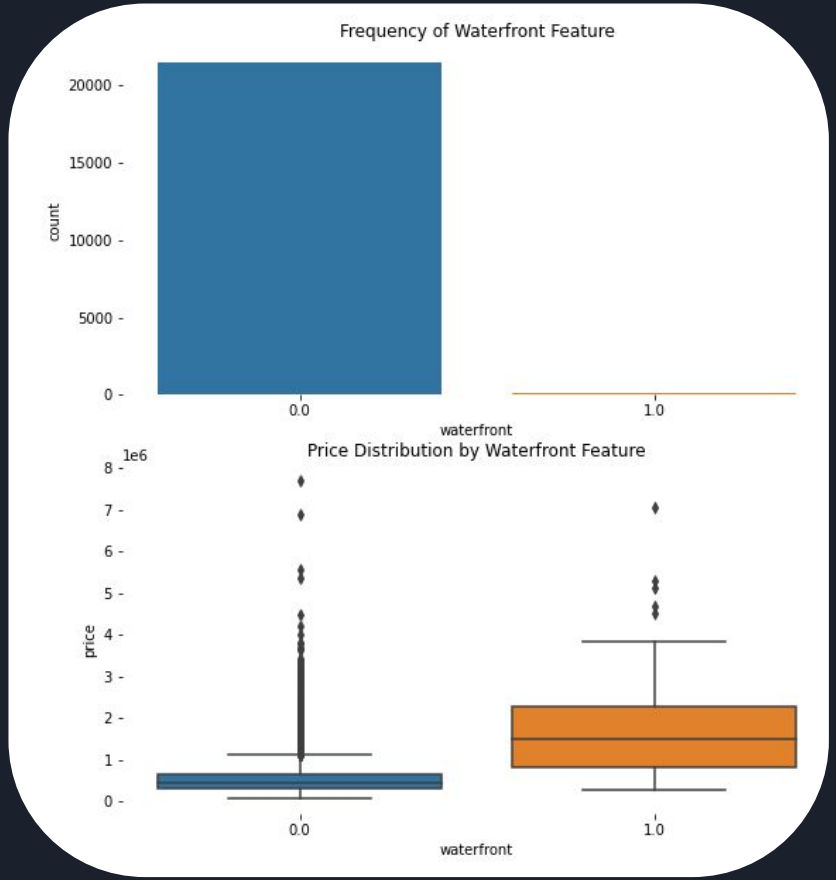
remove the waterfront feature

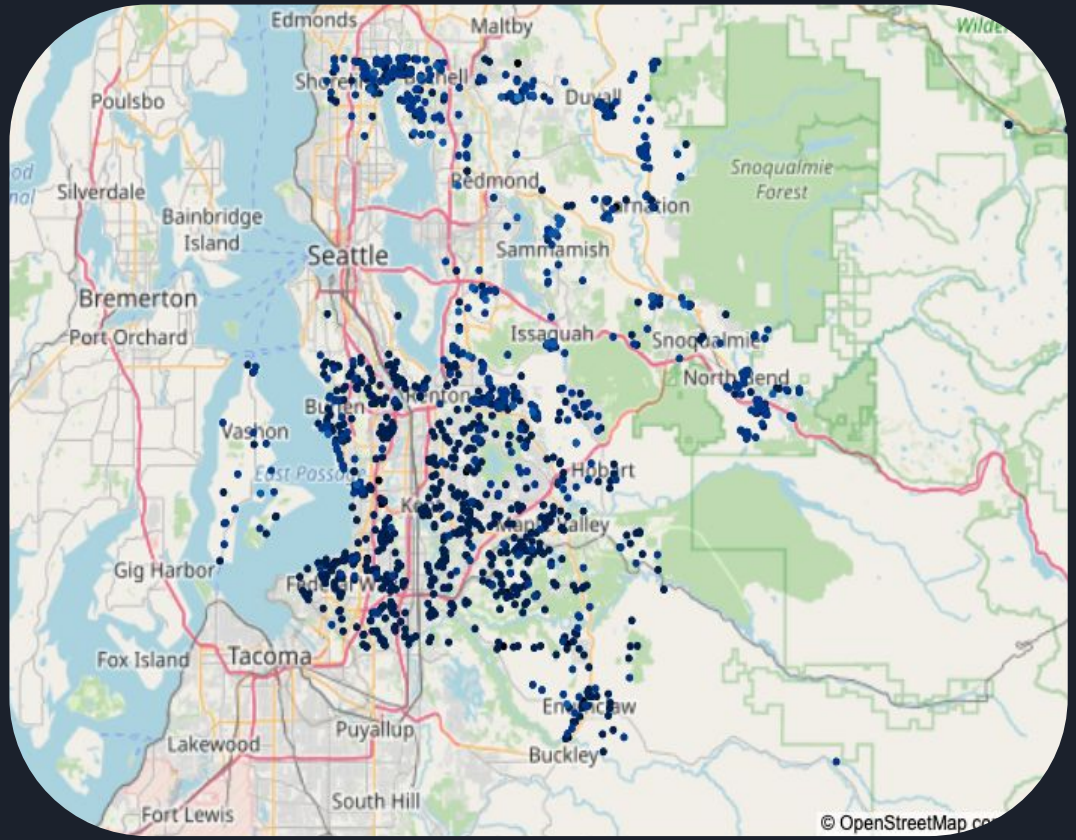
01 | central location

02 | isolated location

03 | has children

04 | limited budget





removing of the waterfront feature generates 1352 results

Stakeholder recommendations

Recommendation for adjusting the search pattern:

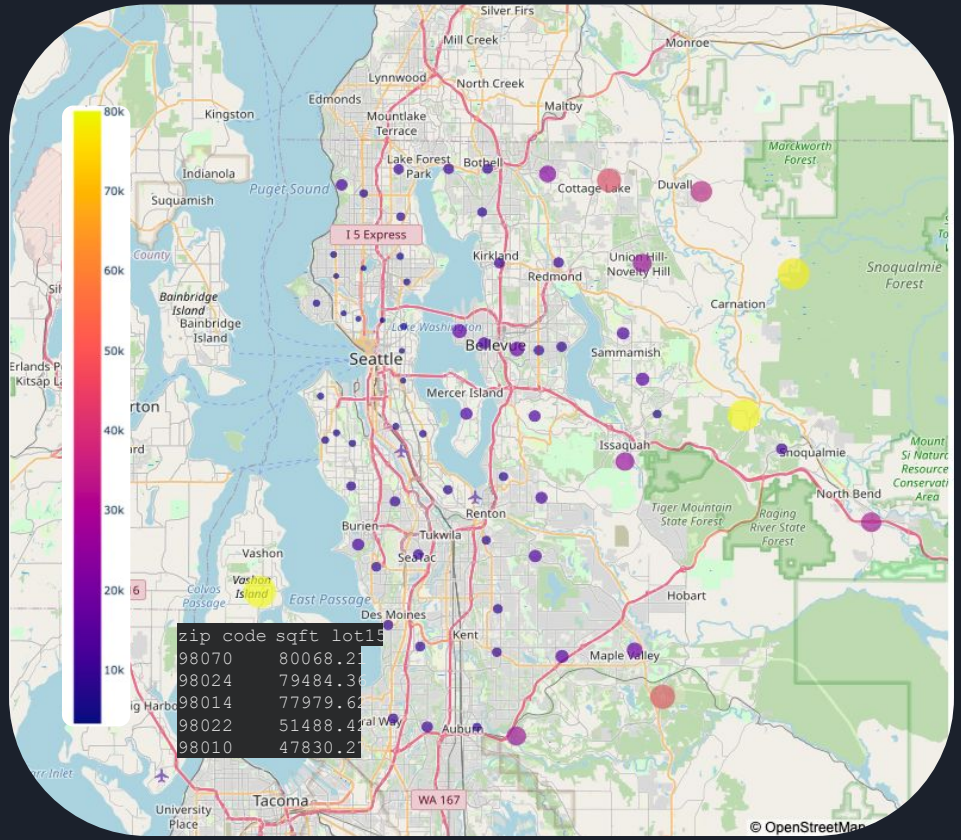
increase the requirements for isolated location

01 | central location

02 | isolated location +

03 | has children

04 | limited budget



Stakeholder recommendations

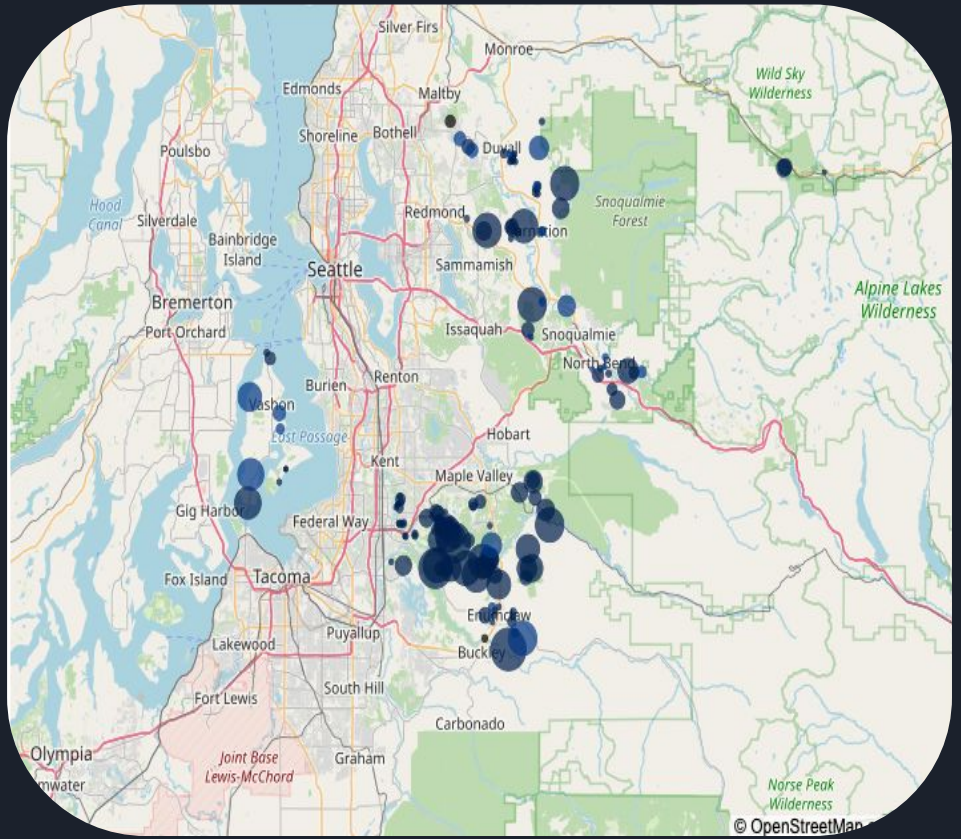
Recommendation for adjusting the search pattern:

step away from the central location

01 | isolated location

02 | has children

03 | limited budget



focusing on the top 10 zip codes in terms of sqft_lot 15 produces 94 results

Limitations & further steps

01 | get additional data to supplement the housing data set

02 | get an updated, more extensive data set for more multiple years

03 | get additional insights into the condition and grade


Thank you