EDA-Project

Houses in King County, WA, USA

by Kilian Gedat

Table of *contents*

The Data

- for this data I was asked to work with the King County House Data dataset
- the time span was 3 days and it is my first EDA project
- the data has 22 columns and 21.597 observations

```
#import the data to a pandas dataframe (IMPORTANT to put the 'eda.')
query_string = """SELECT
    kchs.date,
    kchs.price,
    kchs.id AS sale_id,
    kchd.id AS house_id,
    kchd.bedrooms,
    kchd.bathrooms,
    kchd.sqft_living,
    kchd.sqft_lot,
    kchd.floors,
    kchd.waterfront,
    kchd.view,
    kchd.condition,
    kchd.grade,
    kchd.sqft_above,
    kchd.sqft_basement,
    kchd.yr_built,
    kchd.yr_renovated,
    kchd.zipcode,
    kchd.lat,
    kchd.long,
    kchd.sqft_living15,
    kchd.sqft_lot15
FROM
    eda.king_county_house_sales kchs
LEFT JOIN eda.king_county_house_details kchd ON
    kchd.id = kchs.house_id ;"""
df = pd.read_sql(query_string, db)
```

Data Overview

	date	price	sale_id	house_id	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	...	sqft_above	is_renovated	sqft_basement	yr_built	yr_renovated	zipcode	lat
0	2014-10-13	221900.0	1	7129300520	3.0	1.00	1180.0	5650.0	1.0	False	...	1180.0	False	0.0	1955	1955	98178	47.5112
1	2014-12-09	538000.0	2	6414100192	3.0	2.25	2570.0	7242.0	2.0	False	...	2170.0	True	400.0	1951	1991	98125	47.7210
2	2015-02-25	180000.0	3	5631500400	2.0	1.00	770.0	10000.0	1.0	False	...	770.0	False	0.0	1933	1933	98028	47.7379
3	2014-12-09	604000.0	4	2487200875	4.0	3.00	1960.0	5000.0	1.0	False	...	1050.0	False	910.0	1965	1965	98136	47.5208
4	2015-02-18	510000.0	5	1954400510	3.0	2.00	1680.0	8080.0	1.0	False	...	1680.0	False	0.0	1987	1987	98074	47.6168

- modified the data by dropping: “grade, date, view, sale_id, floors, bathrooms”
- added new columns: “is renovated” and “has_basement”
- cleaned the column “yr_renovated”
- cleaned “NaN” values in “sqft_basement” , ”yr_renovated” and “waterfront”
- “yr_renovated” the “NaN” replaced with the “yr_built”

long	sqft_living15	sqft_lot15
-122.257	1340.0	5650.0
-122.319	1690.0	7639.0
-122.233	2720.0	8062.0
-122.393	1360.0	5000.0
-122.045	1800.0	7503.0

Client

Larry Sanders

Fictive picture of imaginary Larry Sanders:
https://ew.com/thmb/hFXksDb27EuxBK0OrlPbD7FDQuo=/1500x0/filters:no_upscale():max_bytes(150000):strip_icc()/gettyimages-986900
34-e3b12140d94b413b8a34da3972367338.jpg

- Buyer
- wants a house on the waterfront
- has limited budget
- central, but isolated location
- no kids as neighbors because of the germs

Hypotheses - Insight

1. New houses are less likely to have basements.

2. Since houses with waterfront access are rare they are more expensive than houses without waterfront access with the same criteria.

3. A house that is renovated in 1990 is more worth than a house build in 1990.

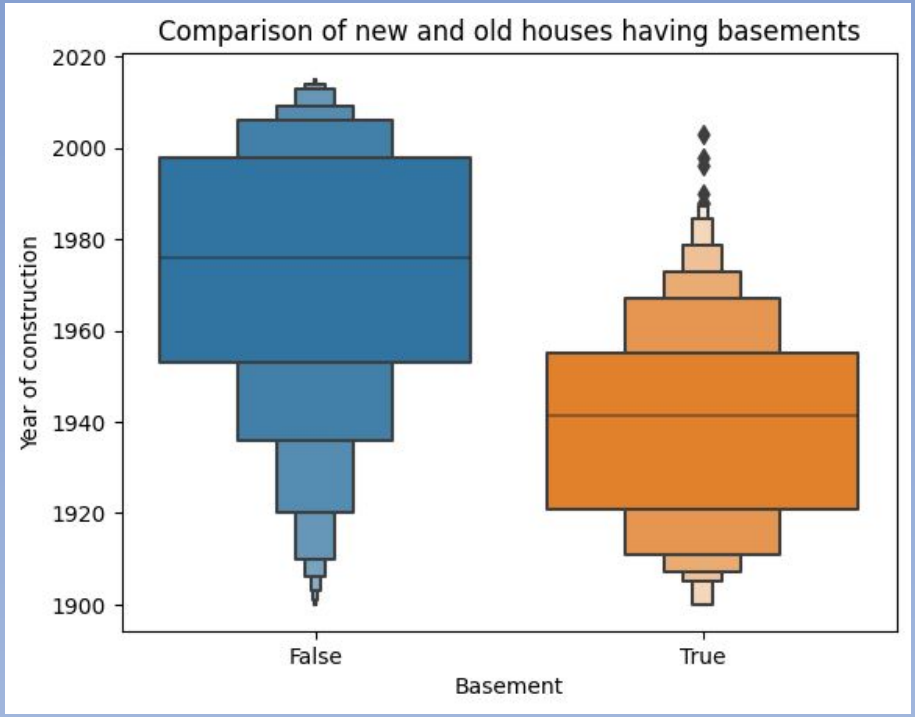
Hypotheses - Client

1. All houses with waterfront access are expensive and require high budget.

2. All waterfront houses are on the countryside rather than central.

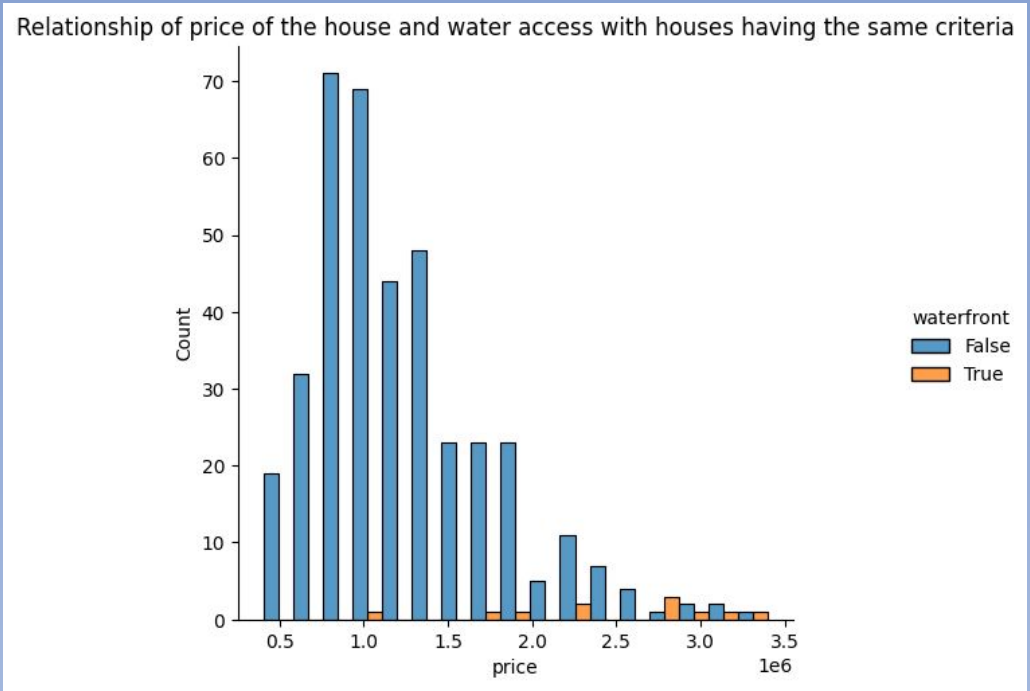
3. Waterfront houses are big enough to be isolated from neighbors in any kind.

Hypothesis 1: New houses are less likely to have basements.



Insights

Hypothesis 2: Since houses with waterfront access are rare they are more expensive than houses without waterfront access with the same criteria.



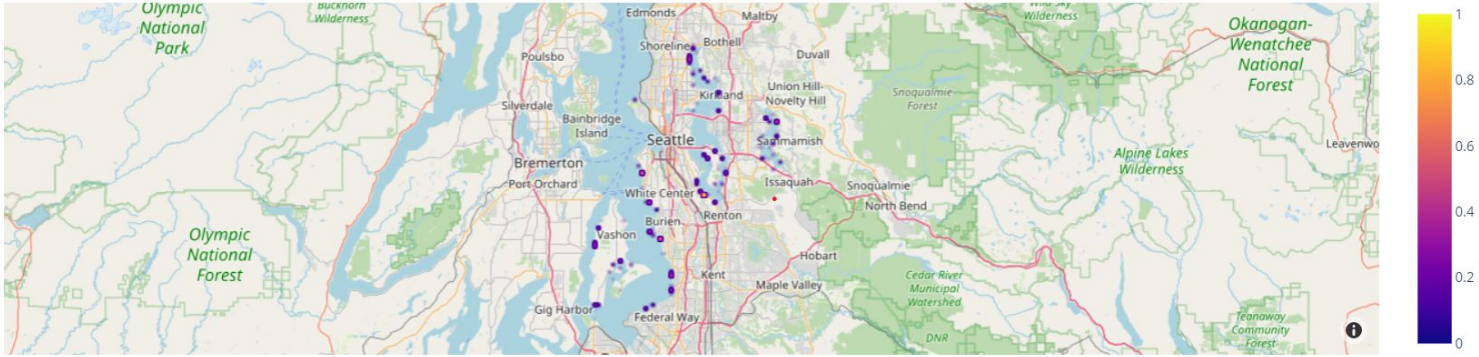
Additional info:
→ houses have between 3 and 5 bedrooms
→ and between 4000 and 4500 sq ft living space (because they have a high correlation to the price

Insights

Hypothesis 3: A house that is renovated in 1990 is more worth than a house built in 1990.

- I couldn't solve that code and it was not geographical
- so I went to see, where the waterfront lots are located

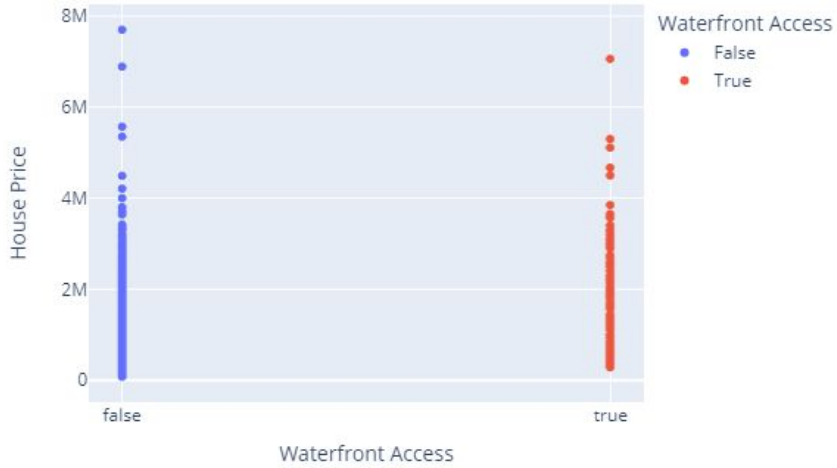
Location of the Waterfront Lots

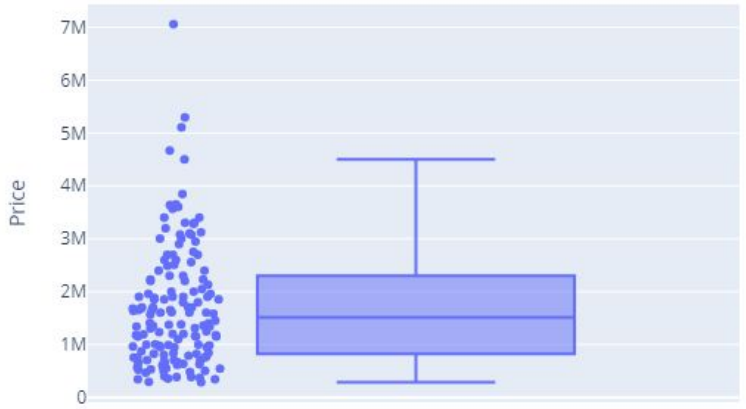


Insights

Hypothesis 1: All houses with waterfront access are expensive and require high budget.

wants a house on the waterfront





Client

Generally there are expensive houses with or without waterfront

Price range between $285.000 - $7.100.000

Hypothesis 1: All houses with waterfront access are expensive and require high budget.

Has limited budget

Assuming that $ 500.000 for a house is limited budget for a Dad with kids

Client

Hypothesis 2: All waterfront houses are on the countryside rather than central.

Central location

Lowest price for a house in a central zip code with waterfront access is $658.000

Larry Sanders has to spend at least $658.000 with his criteria.

ℹ️ ***Info***: Having the criteria: waterfront access, limited budget and central location = No data

Client

Hypothesis 3: Waterfront houses are big enough to be isolated from neighbors in any kind.

Isolated and no kids as neighbors

Smallest lot of the given houses would be 14.244 sq ft which equals to 1323 qm = isolated

the square feet of the living area of the 15 closest neighbors from the $ 658.000 house is 1820 sqt which equals to ~ 170qm = no conclusion about kids as neighbors

Client

Conclusion and recommendation for the client

- looking at all the criteria that Larry Sanders has given
- there will be no perfect fit for him from the houses listed
- I would pick for him the house with house_id *7936500221* that costs $US 658.000 as it fulfills the criterias being central, having a waterfront, the least biggest houses close from the lot and a huge lot itself for the kids to play on by themselves

Future work

- Analyze outliers
- Use font from coding : `Courier new`
-