EDA Project Housing

Finding a House for Larry Sanders

The Data

King County House Prices Dataset

- Houses sold in King County, WA, between May 2014 and May 2015
- 21597 observations
- 21 characteristics/columns

Client: Buyer Larry Sanders

Wishes:

- Waterfront
- Limited budget
- Nice house
- Isolated house
- Central neighborhood
- Neighborhood without kids
- Has children
- (germaphobe)

A closer look at the Data

- No duplicates
- Outlier 33 bedrooms removed
- Other extreme values remain
- Missing values in columns waterfront, view, sqft_basement and yr_renovated
  - Only waterfront relevant to client's wishes
  - Assume NaN houses are those where unknown
- Some columns not relevant for client's wishes
- Information in md, on kaggle, King County standards clear up meaning of view, grade, condition

Hypotheses

- Houses located at the waterfront have a higher price
- The higher the number of bedrooms, the higher the price
- The better the state of the house (grade, condition), the higher the price

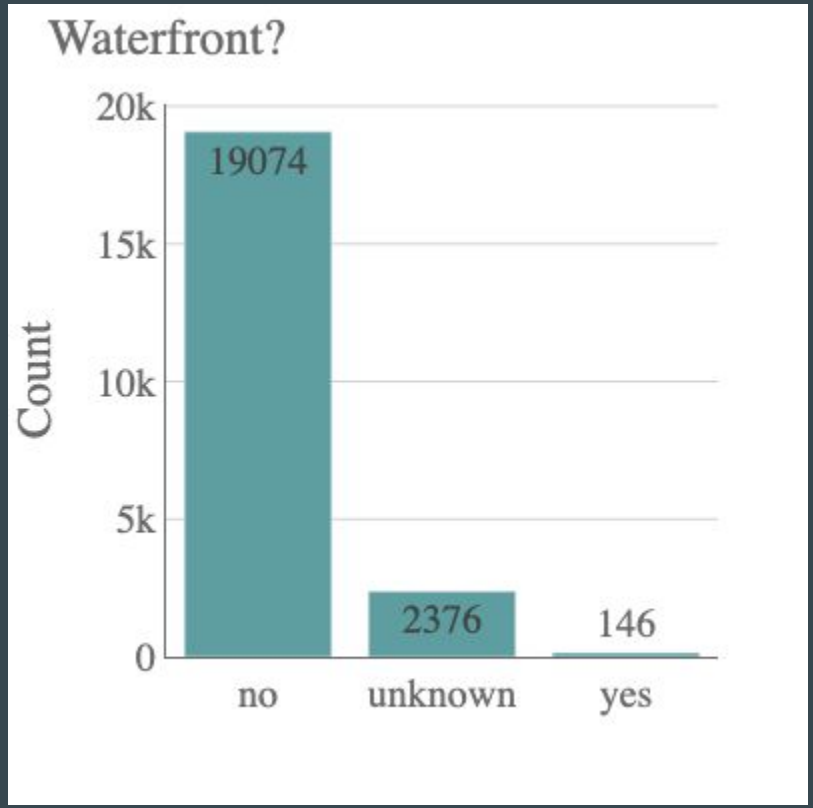
Hypothesis: Houses located at the waterfront have a higher price

Only 0.67 % of the houses are waterfront houses.

Correlation of 0.28

Insignificant?

	Waterfront	No waterfront
Mean Price	1717214.73	532636.36
Median Price	1510000.00	450000.00



Hypothesis: Houses located at the waterfront have a higher price

For houses with sqft_living $\in$ [4000, 6000]:

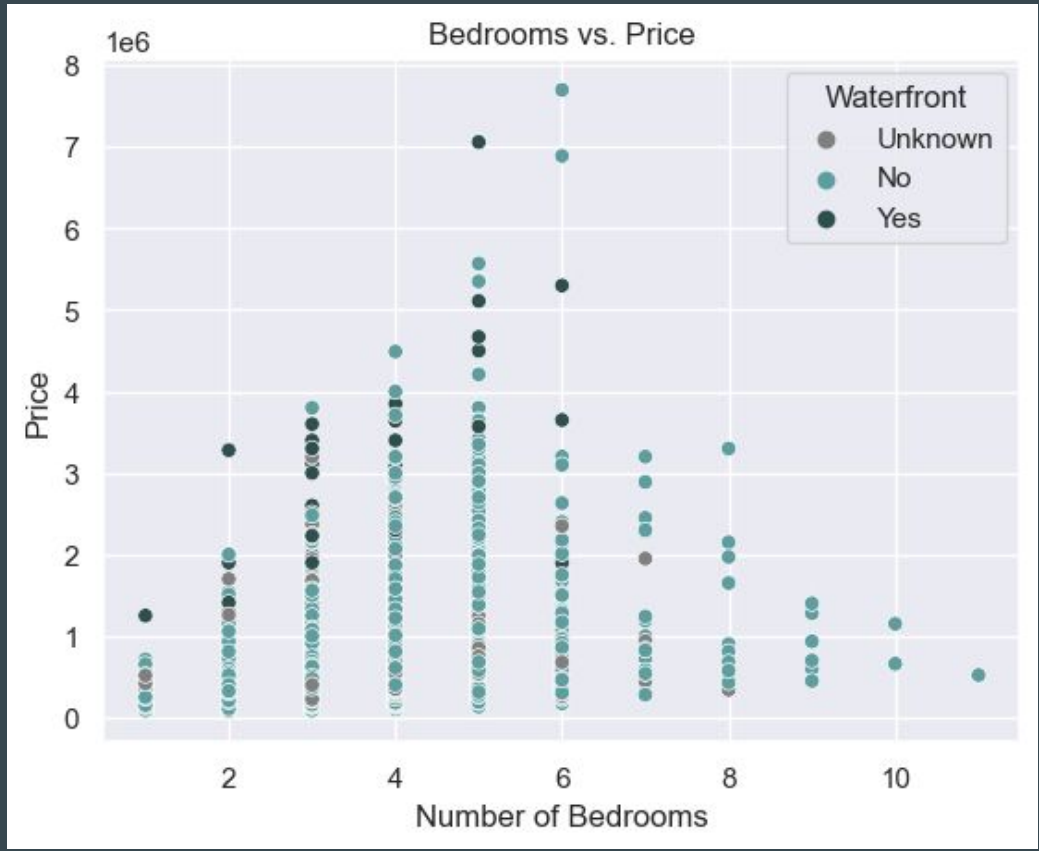
Corr(Price, waterfront) = 0.39

Hypothesis confirmed

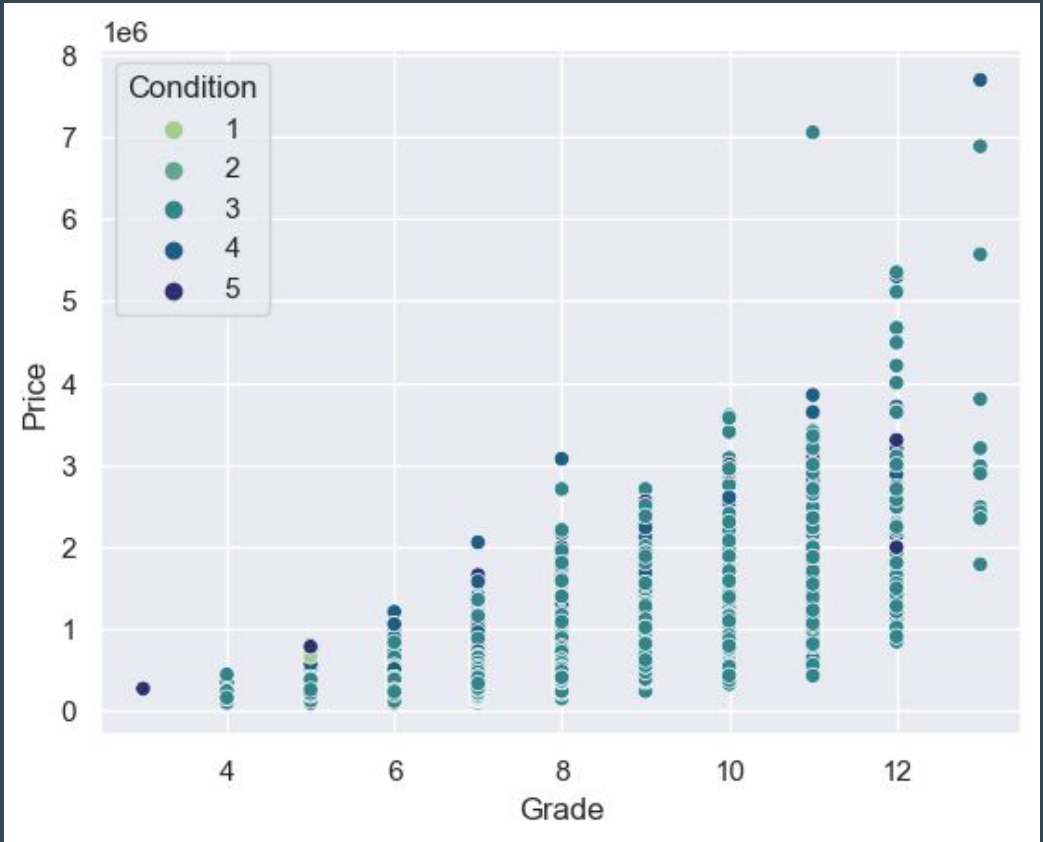
Hypothesis: The more bedrooms, the higher the price

- Correlation of 0.32
- Distribution unexpected for ≥ 6 bedrooms
- Further investigation of the larger houses needed
- Higher for waterfront houses (corr 0.51)

Hypothesis confirmed



Hypothesis: The better the state of the house, the higher the price



- Price and grade are strongly correlated (0.67)
- Price and condition are unexpectedly weakly correlated (0.04)
- Hypothesis holds true for grade, rejected for condition

A Recommendation for Larry

Additional Data Collected

- King County Population Density per zip code, based on 2010 census
- School sites in King County from King County GIS Open Data

- Measures for centrality and children free neighborhoods:

  Transform the data, engineer new features ‘population density rank’ and ‘schools per zip code’ of house observations

How to parametrize Larry's wishes?

- Waterfront: = 1
- Limited budget: not larger than median
- Nice house: condition 3 and up, grade 7 and up (= median)
- Isolated house: sufficiently large lot size, lot size of the 15 nearest neighbors. Above median
- Central neighborhood: Top 15 zip codes with the highest population density
- Neighborhood without kids: zip codes with at most 2 schools
- Has children: at least 2 bedrooms

A Wish Function for Larry

A function to the rescue!

- 7 wishes, 7 criteria. Each criterion satisfied counts for 1
  - If 2 variables determine the criterion, each accounts for 0.5
- Maximum possible outcome per house is 7

- Maximum value attained is 5

What now?

- Maximum value of points attained by houses sold is 5.
- A house fulfilling all of Larry's wishes should have 7 points.

  Not a single house meeting all of Larry's criteria has been sold.

- Relax the criteria
- Imaginary talk with my client
  :-)

Which criteria to relax?

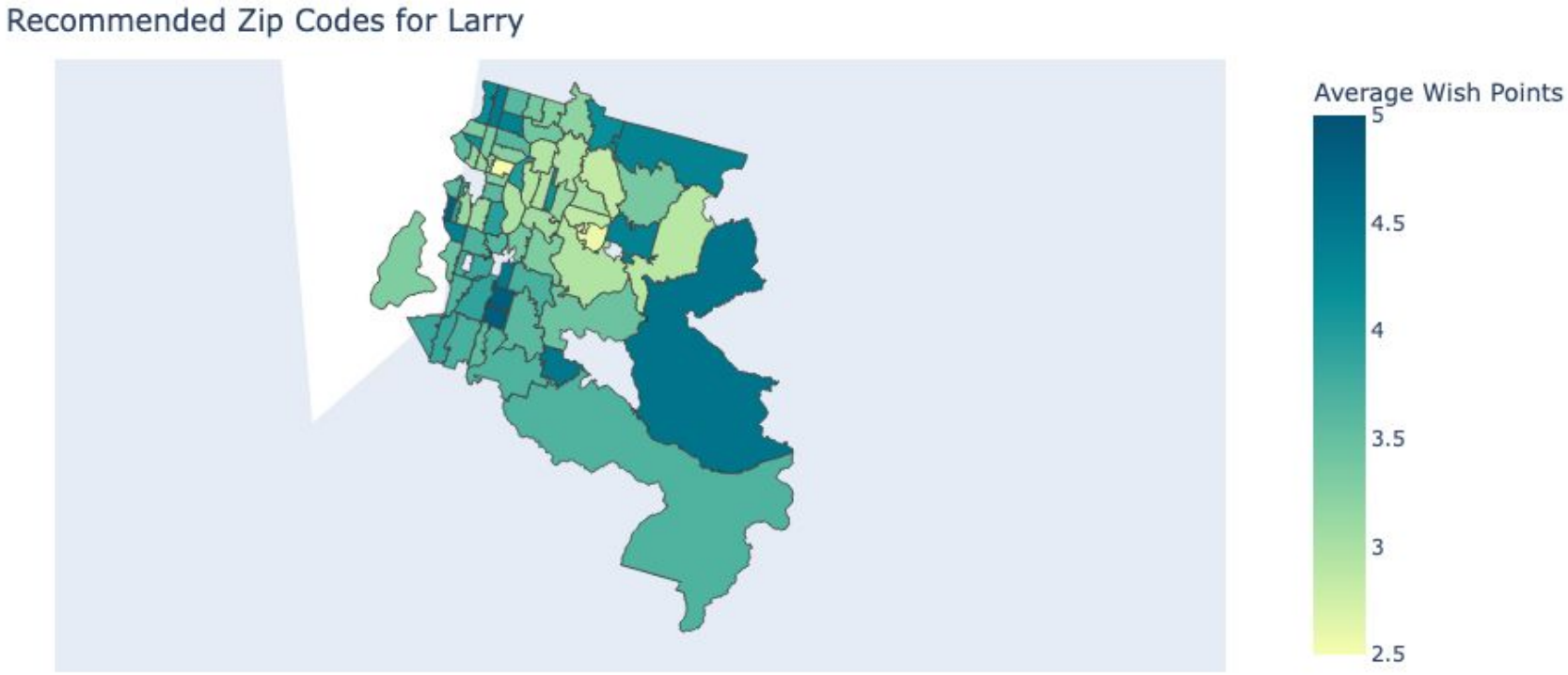
- Waterfront: omit entirely
- Limited budget: not larger than median
- Nice house: condition 3 and up, grade 7 and up (= median)
- Isolated house: sufficiently large lot size, lot size of the 15 nearest neighbors both above 30th percentile
- Central neighborhood: in top 25 zip codes with the highest population density
- Neighborhood without kids: no more than 4 schools in zip code
- Has children: at least 2 bedrooms

Recommendations for Larry

1. Don’t restrict yourself to waterfront housing!
2. If all other criteria should be satisfied: Zip code 98136! All 17 Houses with maximal possible 6 wish points are there
3. Top 10 zip codes with highest average “wish fulfilling property”

Zip codes	Average wish points
98030	4.86
98031	4.84
98136	4.83
98045	4.54
98010	4.50
98055	4.47
98133	4.46
98146	4.39
98125	4.36
98024	4.35

Recommendations for Larry



Future Work

- Add weight options or ranks to the criteria in the wish function
- Nicer recommendation plot
- Build dashboard where you can change the weights/ranks
- Neighborhoods without kids:
  - Match school data to time frame sold: School data is from 2021
  - Find data on area per zip code, compute “school density” for more precise ranking
  - Find/understand census data on age groups per zip code
- Alternative strategy concerning “central” houses
- Add renovation status as a factor for “nice” houses

Thank you!