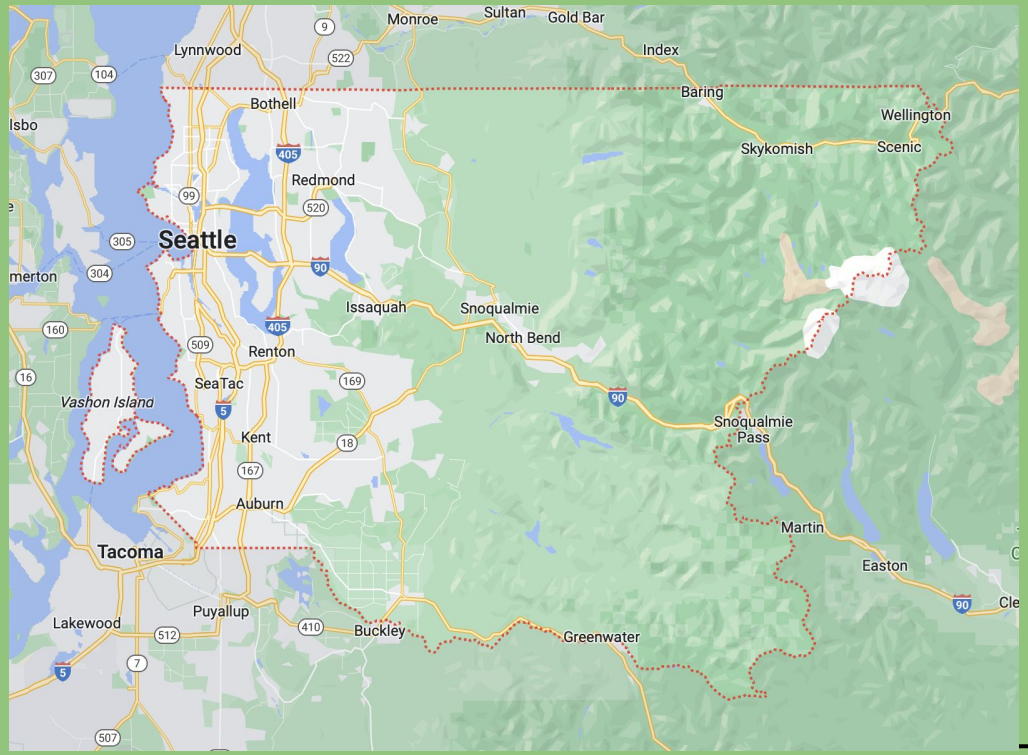
Exploratory Data Analysis

of the King County Housing dataset

Overview of the dataset

- 21597 observations
- 20 features
- target variable: price at which a house was sold
- houses in King County, USA
- sales between May 2014 and May 2015
- mostly around Seattle area
- unique sales id’s
- 190 houses are duplicate

King County


Monroe
Sultan
Gold Bar
Lynnwood
Index
Bothell
Baring
Wellington
Skykomish
Scenic
Redmond
Seattle
Issaquah
Snoqualmie
North Bend
Renton
SeaTac
Vashon Island
Kent
Snoqualmie Pass
Auburn
Martin
Tacoma
Easton
Puyallup
Lakewood
Buckley
Greenwater

The Client

a buyer who wishes to buy a house in King County with these characteristics:

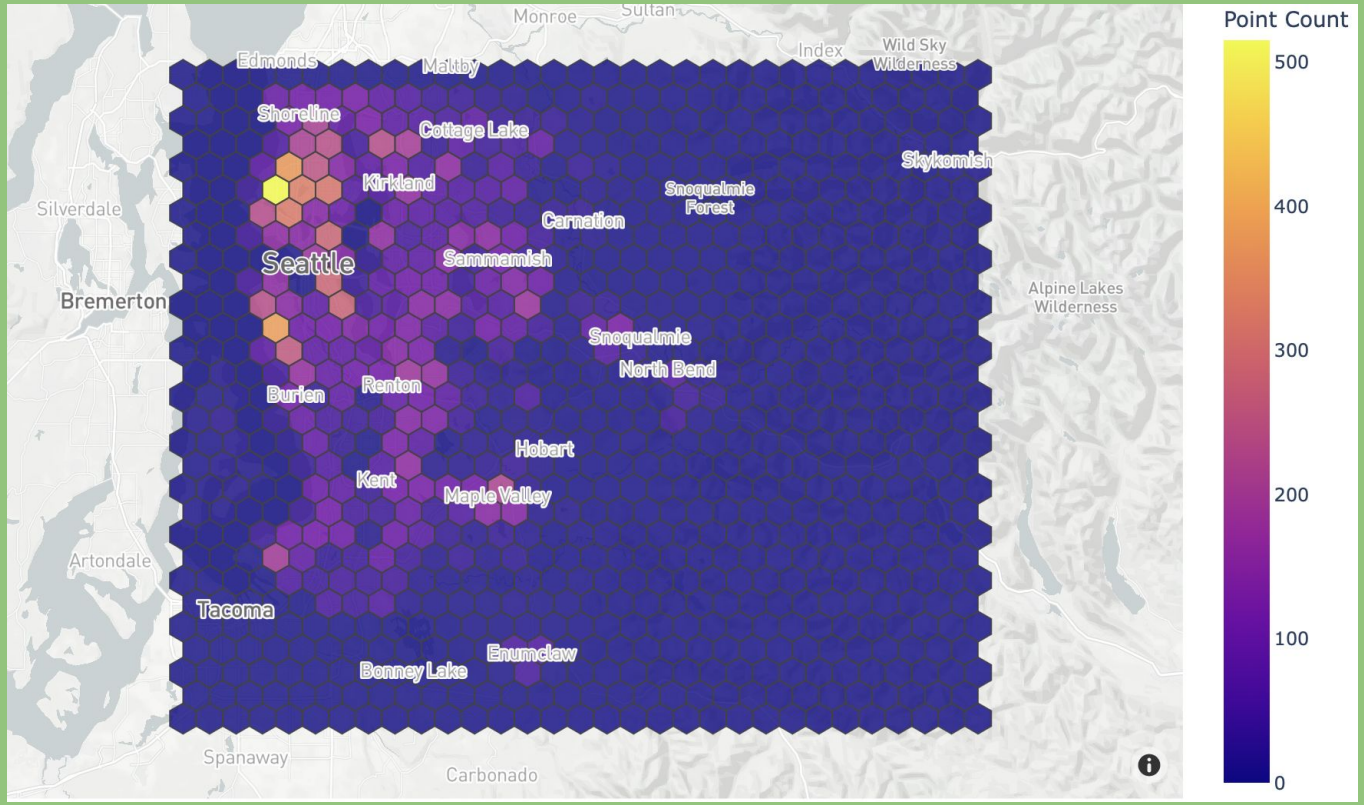
- located in a lively area in the city center or nearby
- mid range price

wants to know what time of the year is a good time to buy such a house

Quality of the data

- missing data in a number of columns
- no indication of inaccurate measurements
- all values are plausible (with a few exceptions)
- outliers in a number of variables

A closer look at the data

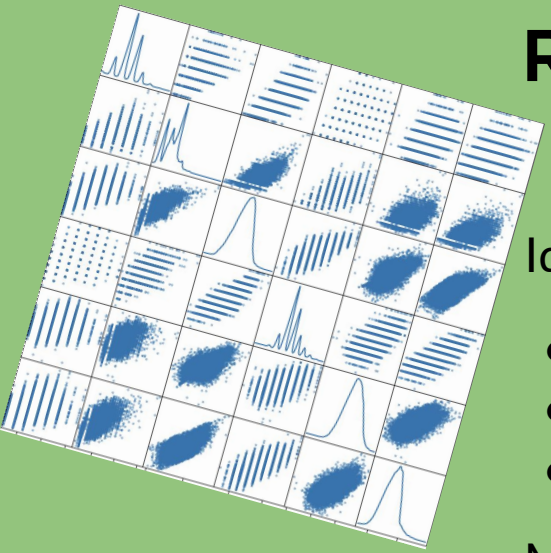


Goals

- draw insights from the data
- give advice to the client

Research Questions

- Are there strong correlations between certain features and the price?
- Do the housing prices in the northern part of the county differ from the southern regions of King County?
- Is the proportion of luxury houses dependent on the zipcode?
- When is the best time to buy a house for the client?
- In which zip-codes does the client has a better chance to find a house according to their specifications?



Relationships in the data

Identified strong correlation between the **price** and

- size of the living area
- grade denoting the class (luxury, average, bad condition)
- and a number of other variables

Methods used:

- Pearson correlation coefficient
- Hypothesis testing of the significance of the correlation coefficient

Difference if prices

- Identified a considerable difference in prices between north and south

- Methodology used:

  two sample t-test (p-value close to zero)

Proportion of luxury houses in various zip-codes

- based the definition of of a luxury house on the **grade** variable (grading system clarification)
- proportions of luxury houses range from 0% to 47%
- confirmed the difference with the chi-squared test

zip code	%	town
98039	47	Medina
98075	41	Sammamish

The best time of the year for the client to buy a house

- defined “mid range price” based on the IQR
- defined “lively neighborhood near the city center” as being located within 3 km of the city centers of the top 5 largest cities in King County:
  - Seattle
  - Bellevue
  - Kent
  - Renton
  - Federal Way

- used the Haversine distance function to determine the distance based on the latitude and longitude given in the data

The best time of the year for the client to buy a house

- Time series plot of the filtered out data didn't show a clear seasonal fluctuation in price
- After smoothing the data (with **numpy.correlate**) a peak and a drop in prices became visible

The best time of the year for the client to buy a house

- Best time to buy a house: December and January
- Expect high prices in: October

zipcode
98122
98144
98112
98102
98109
98119
98056
98055
98059
98031

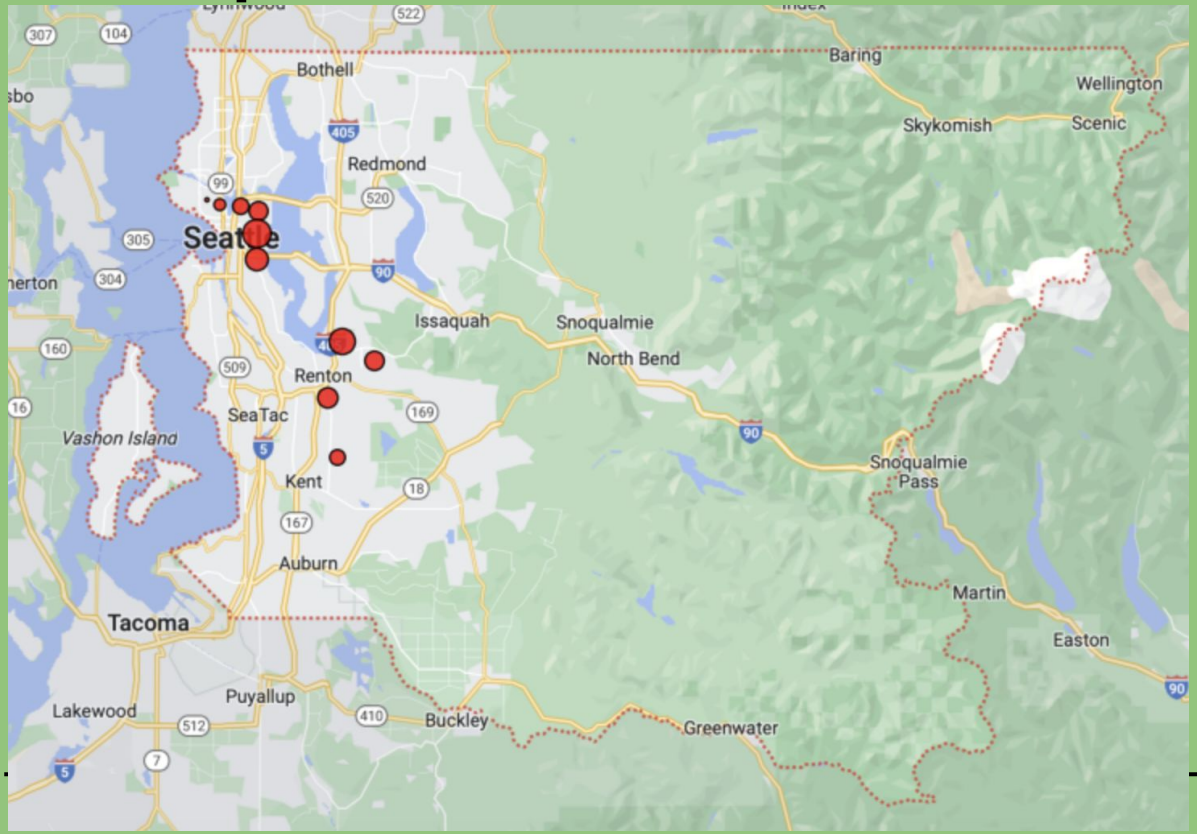
Best zip-codes for the client

- contain a large number of houses with a price lying in the mid-range
- located near the city center
- “lively” neighbourhood inferred from the feature denoting the total living area of the nearest 15 houses
- low price of a square meter (engineered feature)

Method:

rating the zip-codes through sorting on multiple features with a prior binning of those features

Best zip-codes for the client



Impact and future applications

- advising potential buyers
- building a (linear) model predicting house prices

Jupyter notebook

repository