King County House Prices

EDA Project

Matthias Schmidt

2021-06-07

Data Set

King_County_House_prices_dataset.csv (kaggle.com)

21597 sales in King County between May 2014 and May 2015

19 metrics per sale:

- Status: condition, grade, view, waterfront or not
- Location: zip code, longitude, latitude
- Time: year built, year of last renovation, date of sale
- Size: nrs of bedrooms, bathrooms and floors
  - living area, areas of basement and upper floors, lot area,
  - local mean living and lot areas

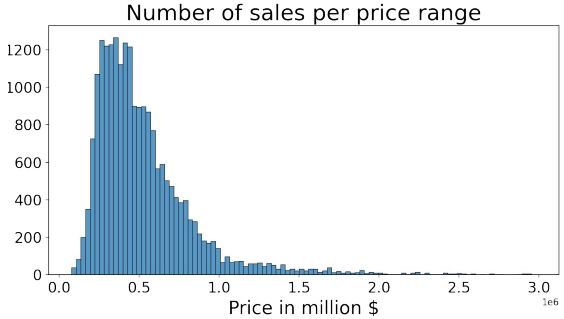
Task: Model the House Price

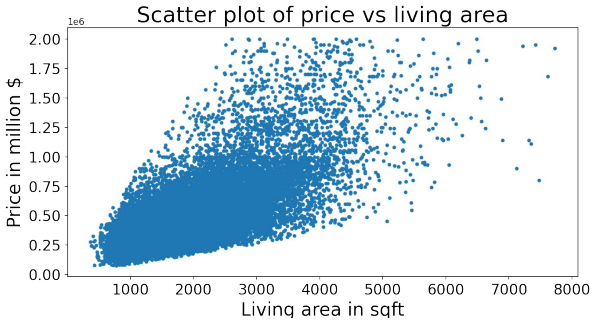
Can be used for

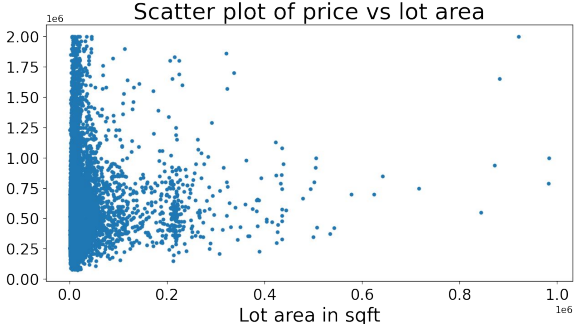
- building a price estimator for a local realtor
- building a price estimator for private sellers
- building a house configurator for home buyers (‘what to expect’
- counseling investors what and where to buy

Data Exploration

Focus on ordinary properties → price and size cutoffs





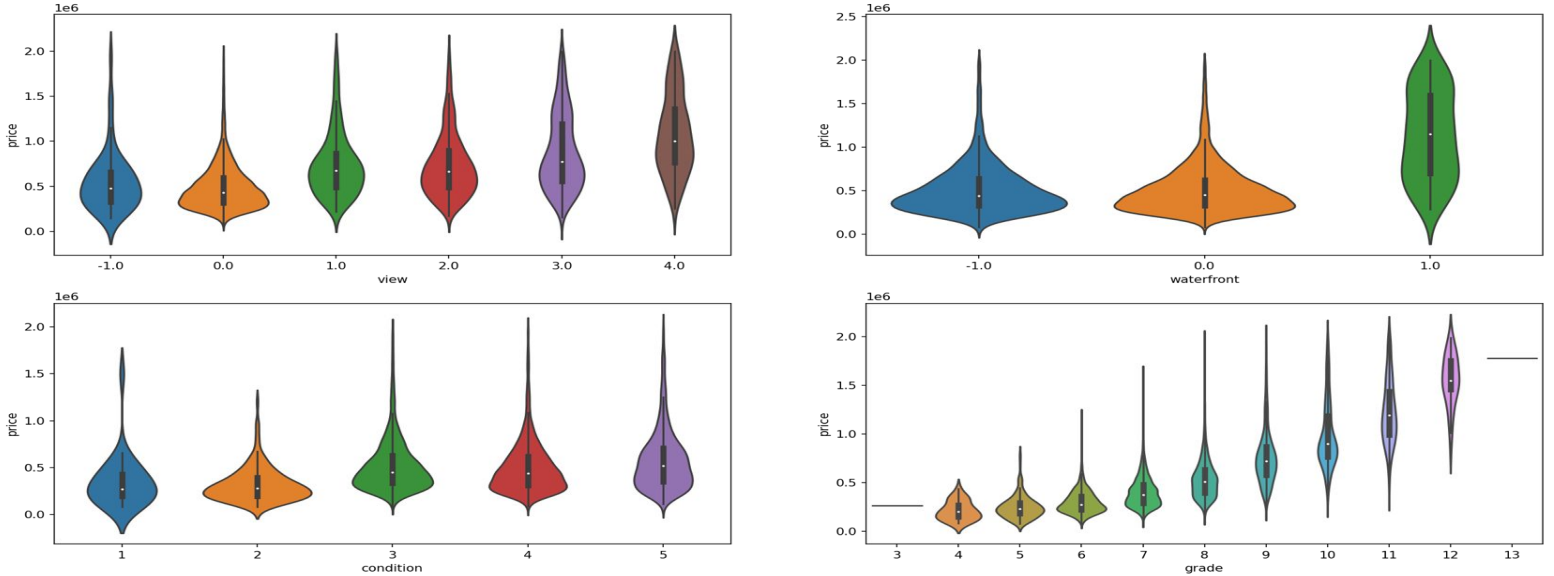


price cutoff at 2 million $	living area cutoff at 6500 sqft	lot area cutoff at 600,000 sqft

→ dropping 227 data points

Status Variables

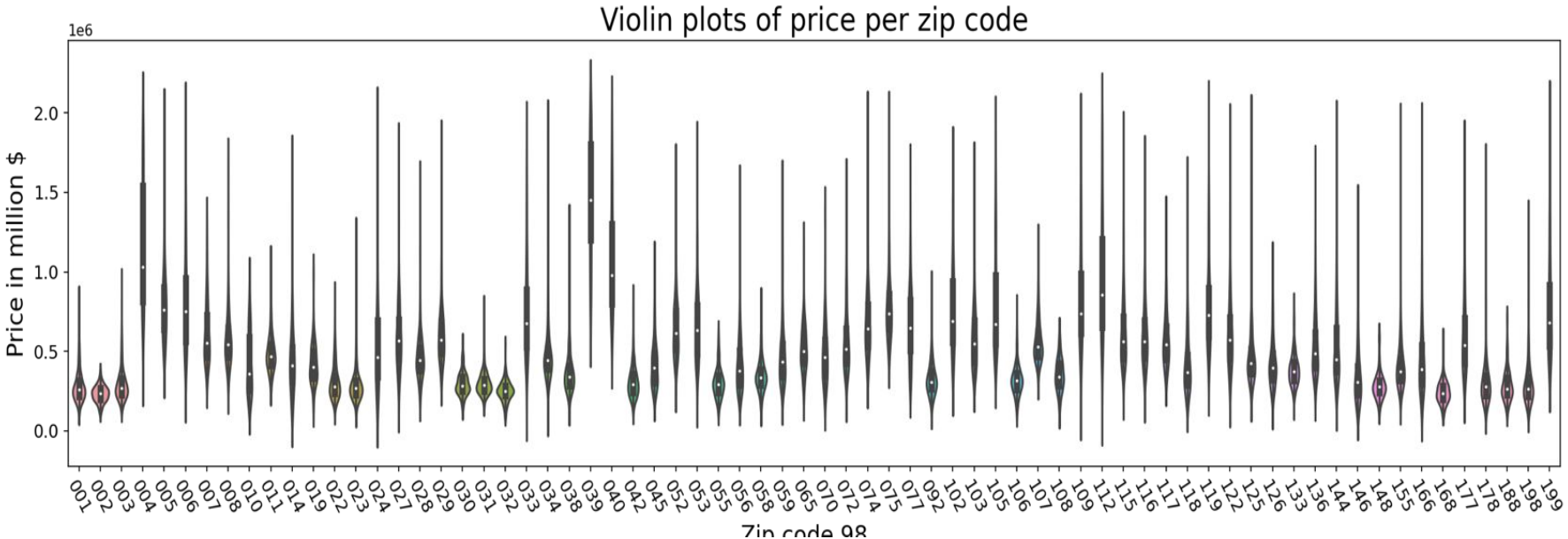
Violin plots of price vs view, waterfront, condition, and grade (-1 = n.s.):



Cleaning: remove data points with view = n.s., condition = 1, grade = 3, 13 (→ - 93)

Dummy variables for all status variables (relation to price seemingly not linear)

Location Variables

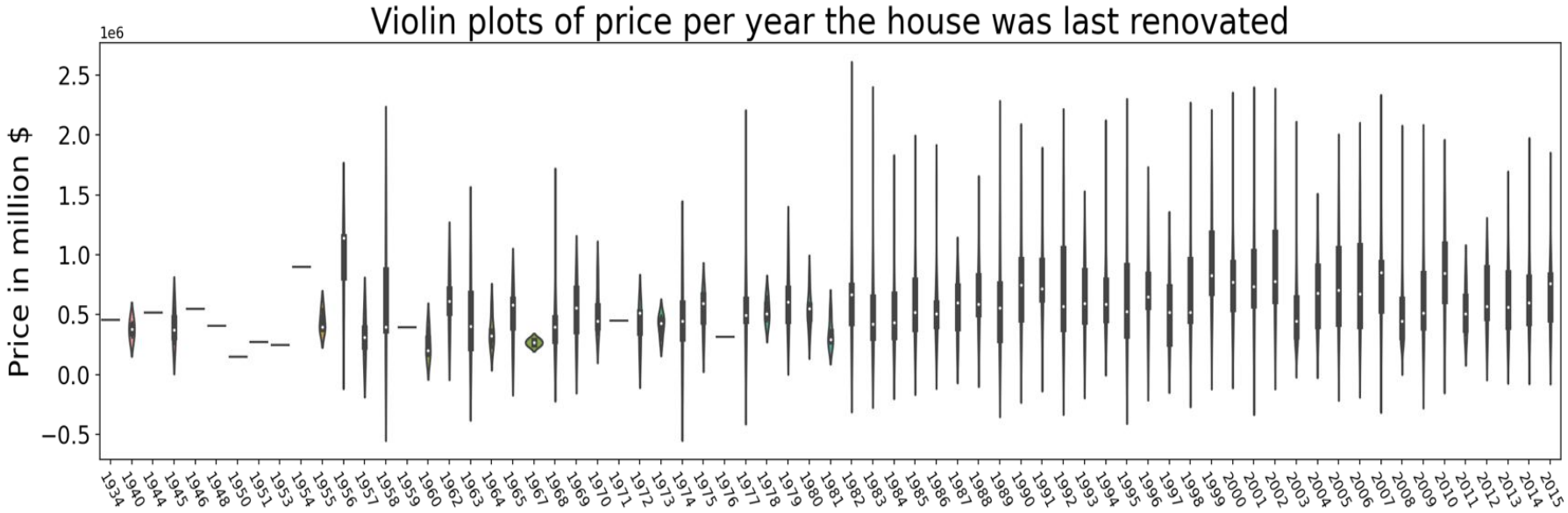


Significant dependence → replace zipcode by

location score = median price for one square foot of living area

Time Variables: Year the House Was Built

Time Variables: Year the House Was Last Renovated

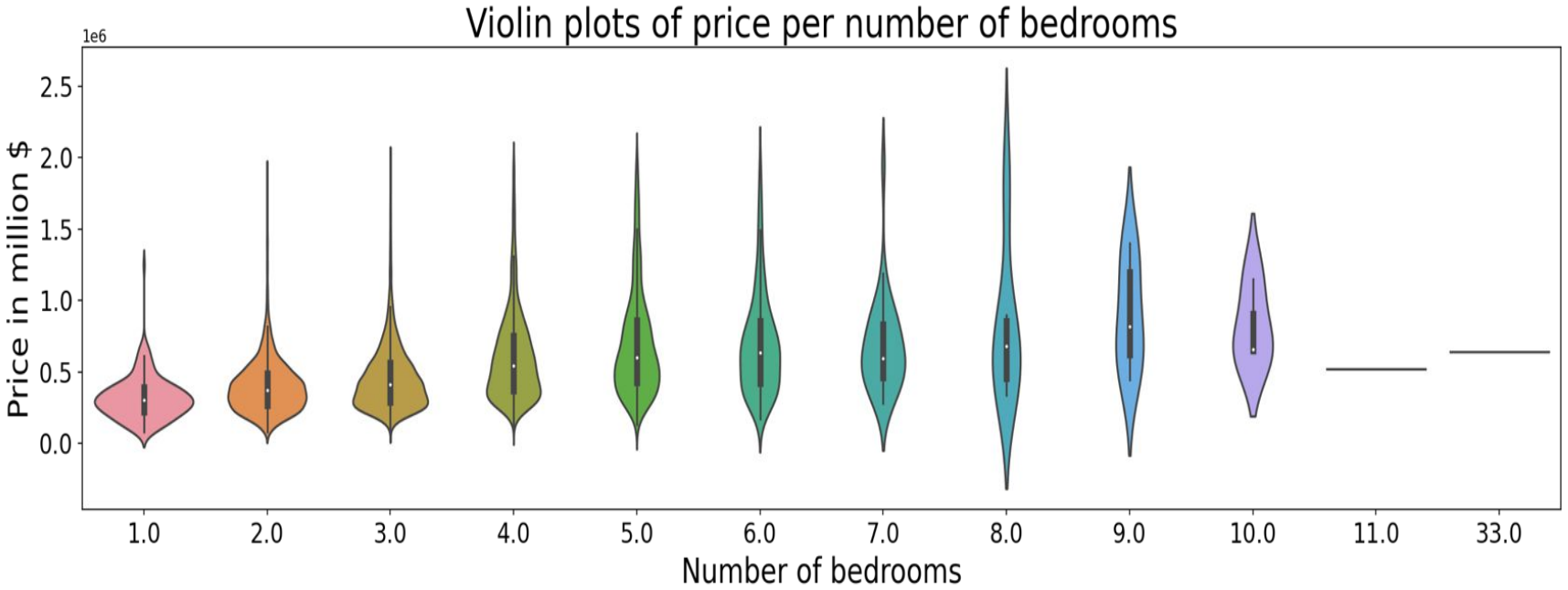


Time Variables: Date of Sale

None of the 3 time variables seems to have significant influence.

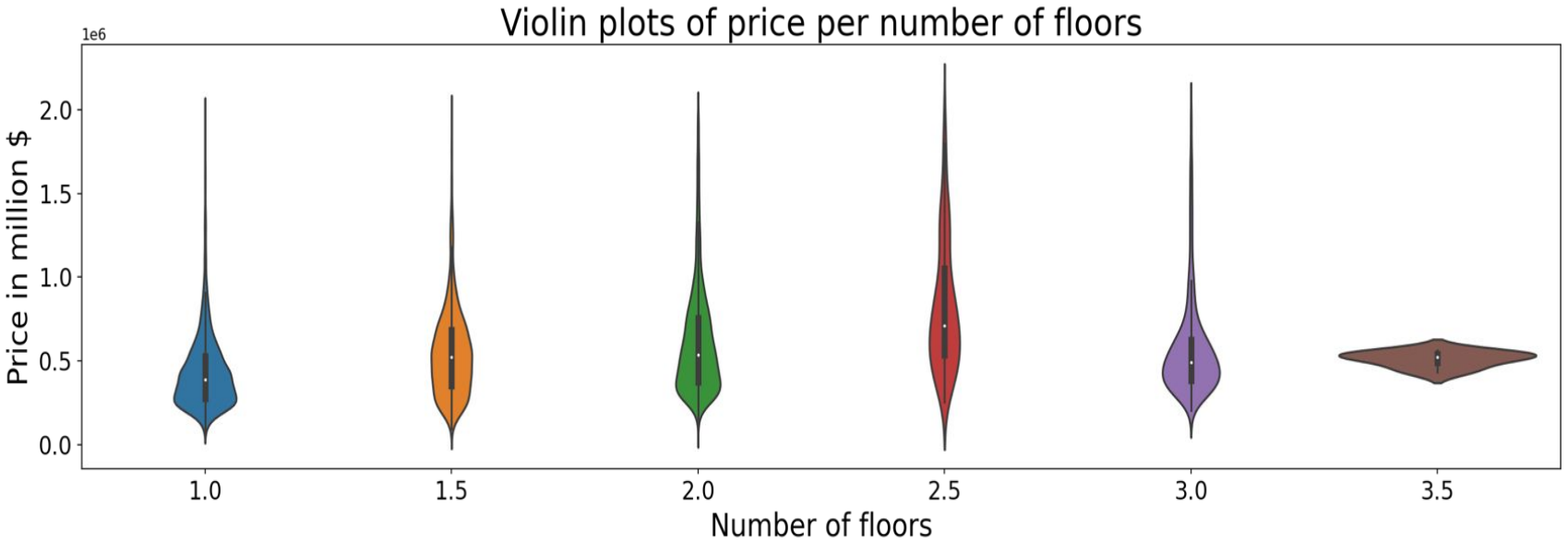
Anyway, keep as variables until tuning

Size Variables: Number of Bedrooms



Size Variables: Number of Bathrooms

Size Variables: Number of Floors



Cleaning data: remove data points with

bedrooms >= 10, bathrooms >= 5, floors > 3 (→ - 35)

Size Variables: Area

Living area A_L splits into area of basement A_B and area of upper floors A_U .

→ keep either A_L alone or both A_B and A_U as independent variables

Basement and upper floors may contribute differently to price.

→ drop A_L and keep A_B and A_U

No obvious correlations between the remaining size variables → keep all

Tuning

Linear regression of price after removing the variables with highest P value:

- 23 variables
- $R^2 = 0.836$
- $R^2_{adj} = 0.836$
- $P > |t|$ less than 0.005 except for intercept (0.079)

Removing variable with low individual price correlation does not improve result

Most significant variables

- mean local living area
- location score
- area of upper floors
- number of bathrooms
- grade

Responsible for 90% of R^2_{adj}

Supervised Learning

Split data set into training set and test set at a ratio if 1:3

Determine linear regression coefficients from training set

Build prediction function

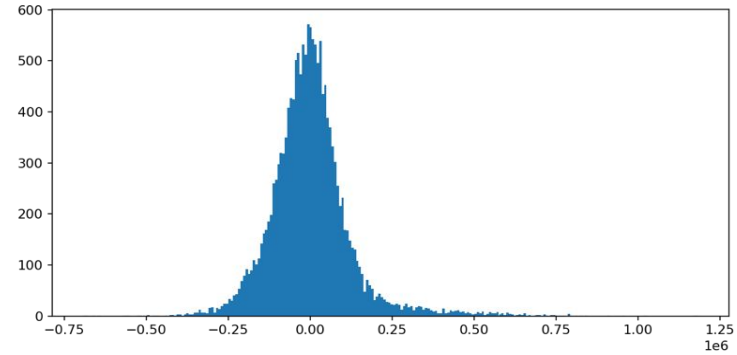
Compute residual for each data point in test set:

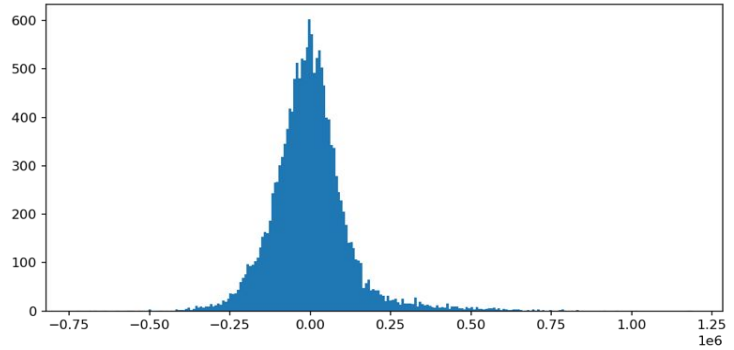
res = price - value returned by prediction function

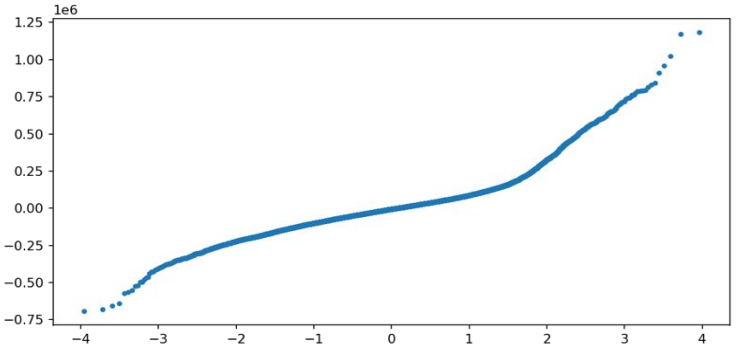
→ RMSE = 115832

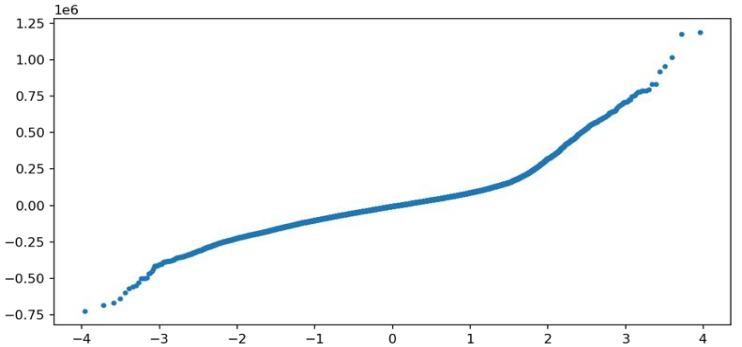
Test Normality of Residuals

Distribution plot and QQ plot of residuals: test vs trained (left), total vs total (right)









J-B test: p = 0.00 K-S test: p = 0.00