EDA Project - King County

•••

P. McRae

OverView

1. Task and Data Description
2. Price - features
3. Scatterplot - relations
4. Insights
5. Categorical characteristics
6. Ordinary Linear Regression

1. Task and DataDescription

Task:

- Analyse the King County housing prices.
- Find insights
- Train a OLS model to predict sales prices

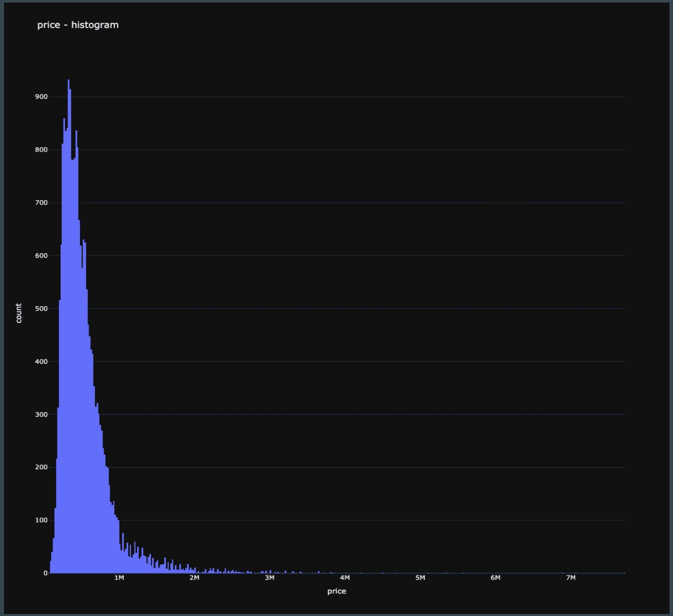
1. Task and DataDescription

- shape:
  - 21597 observations
  - 21 columns
- nan values:
  - waterfront: 2376
  - view 63
  - yr_renovated 3842

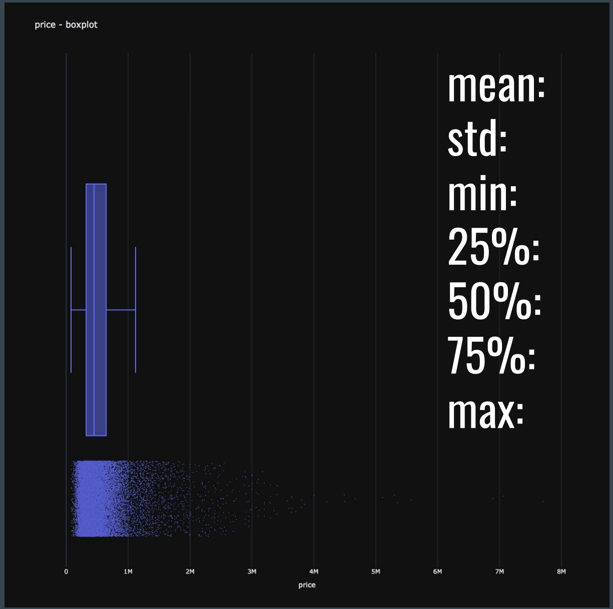
column names:

- id
- price
- bedrooms
- bathrooms
- sqft_living
- sqft_lot
- floors
- waterfront
- view
- condition
- grade
- sqft_above
- sqft_basement
- yr_built
- yr_renovated
- zipcode
- lat
- long
- sqft_living15
- sqft_lot15

2. Price - features



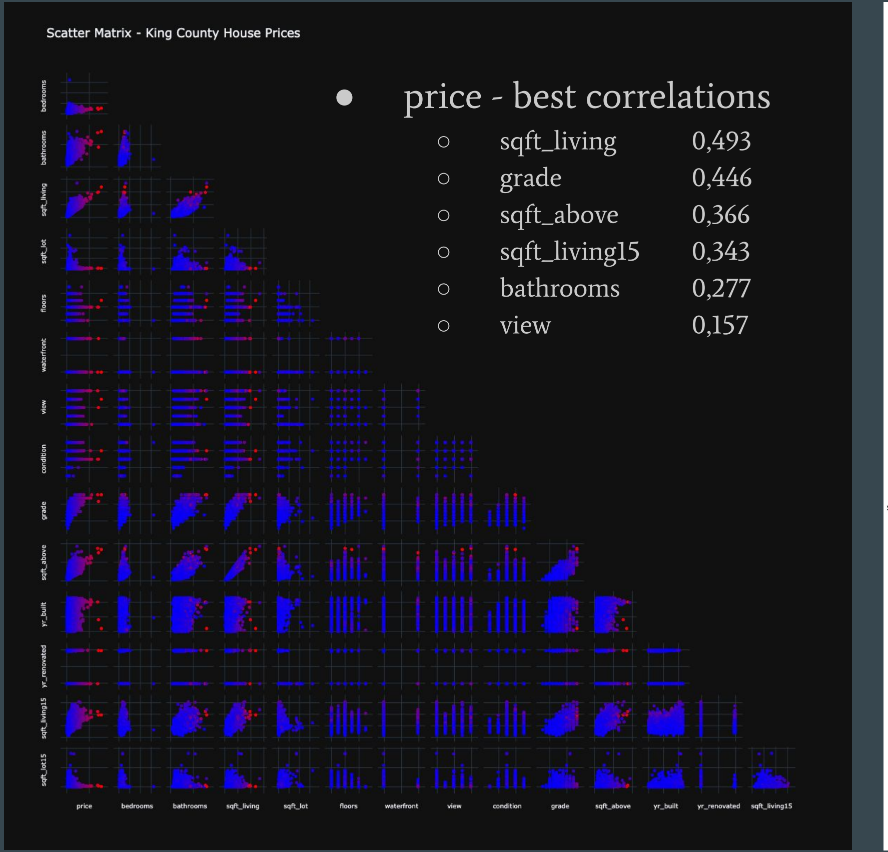
price - histogram



price - boxplot

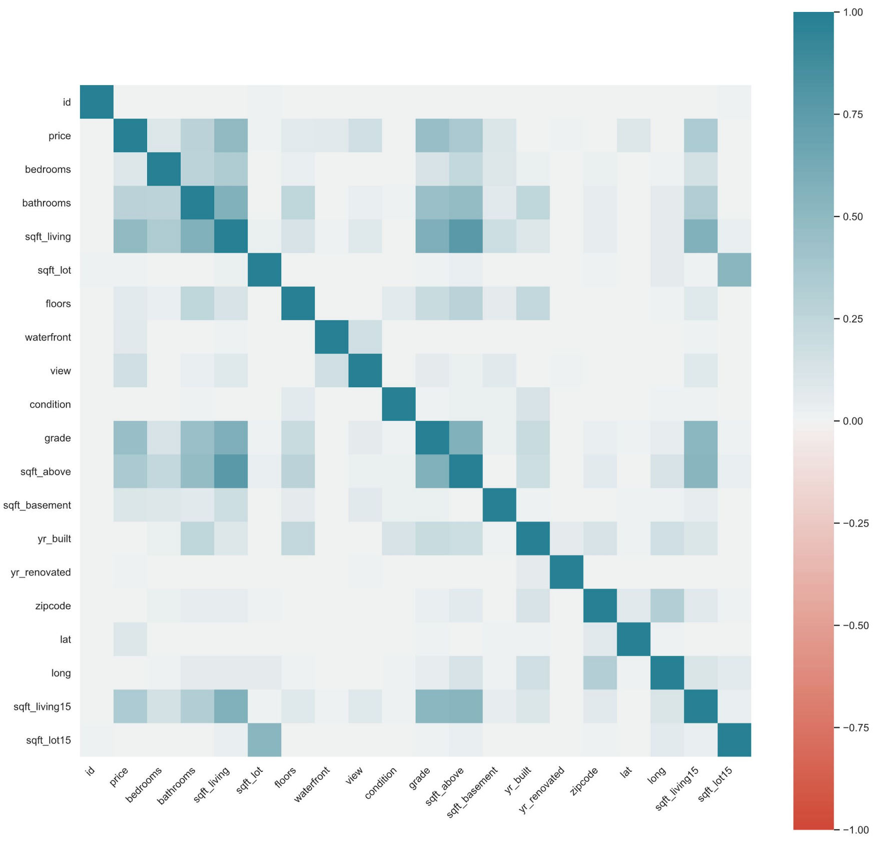
mean:	540296	5,4e5
std:	367368	3,6e5
min:	78000	7,8e4
25%:	322000	3,2e5
50%:	450000	4,5e5
75%:	645000	6,5e5
max:	7700000	7,7e6

3. Scatterplot relations



- price - best correlations

○ sqft_living	0,493
○ grade	0,446
○ sqft_above	0,366
○ sqft_living15	0,343
○ bathrooms	0,277
○ view	0,157



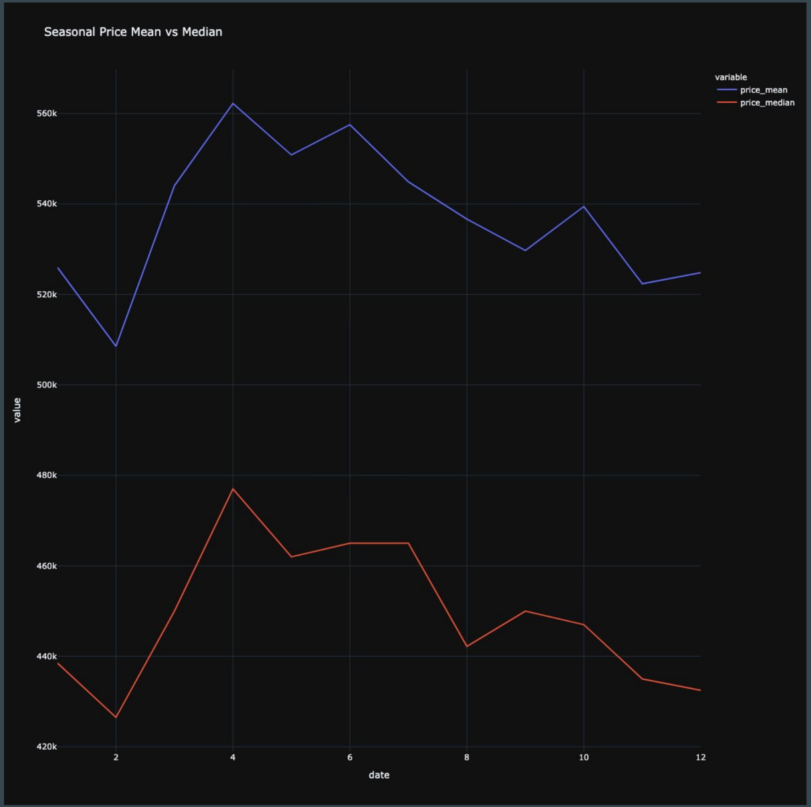
4. Insights

I. price - best correlations
   - sqft_living 0,493
   - grade 0,446

II. grade
   - sqft_living15 0,714

III. sqft_lot
   - sqft_lot15 0,718

4. Insights

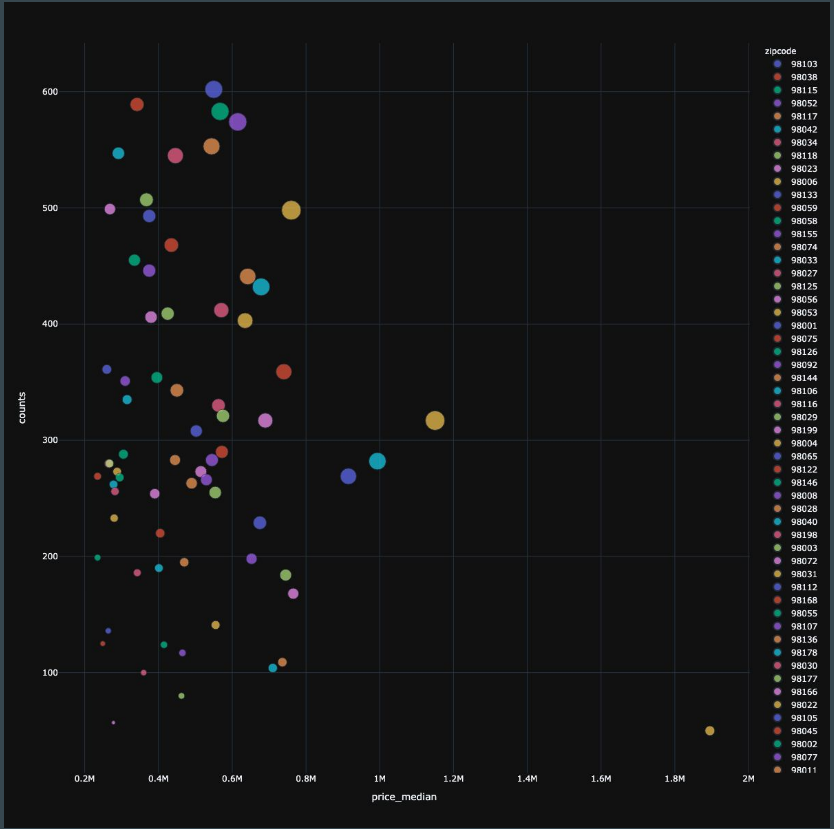


Comparison of sales prices over the seasons.

-> Lowest prices at the beginning and end of the year

-> Highest prices in spring and early summer

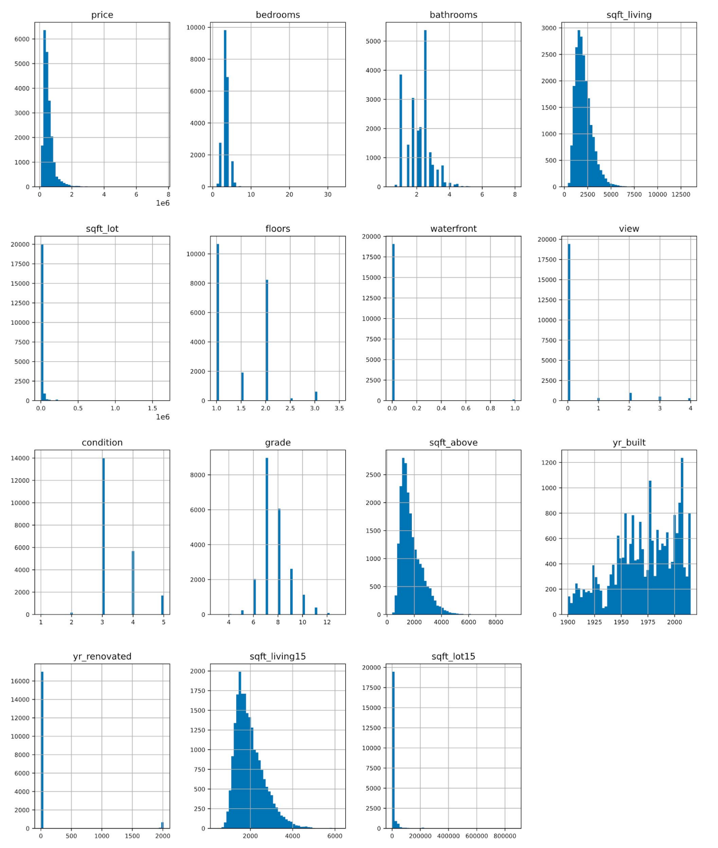
4. Insights



Comparison sales prices to the amount of sales in a zipcode area. Size stands for the cumulated sum of sales for that zipcode

-> slight trend: the higher the median, the more sales in that zipcode area

5. Categorical characteristics



- date -> category (turn into months only)
- floors -> category
- bathrooms -> category
- bedrooms -> category
- waterfront -> category (will be dropped)
- view -> category
- condition -> category
- grade -> category
- zipcode -> category

5. Ordinary Linear Regression

drop columns:

[id, waterfront, lat, long, yr_renovated]

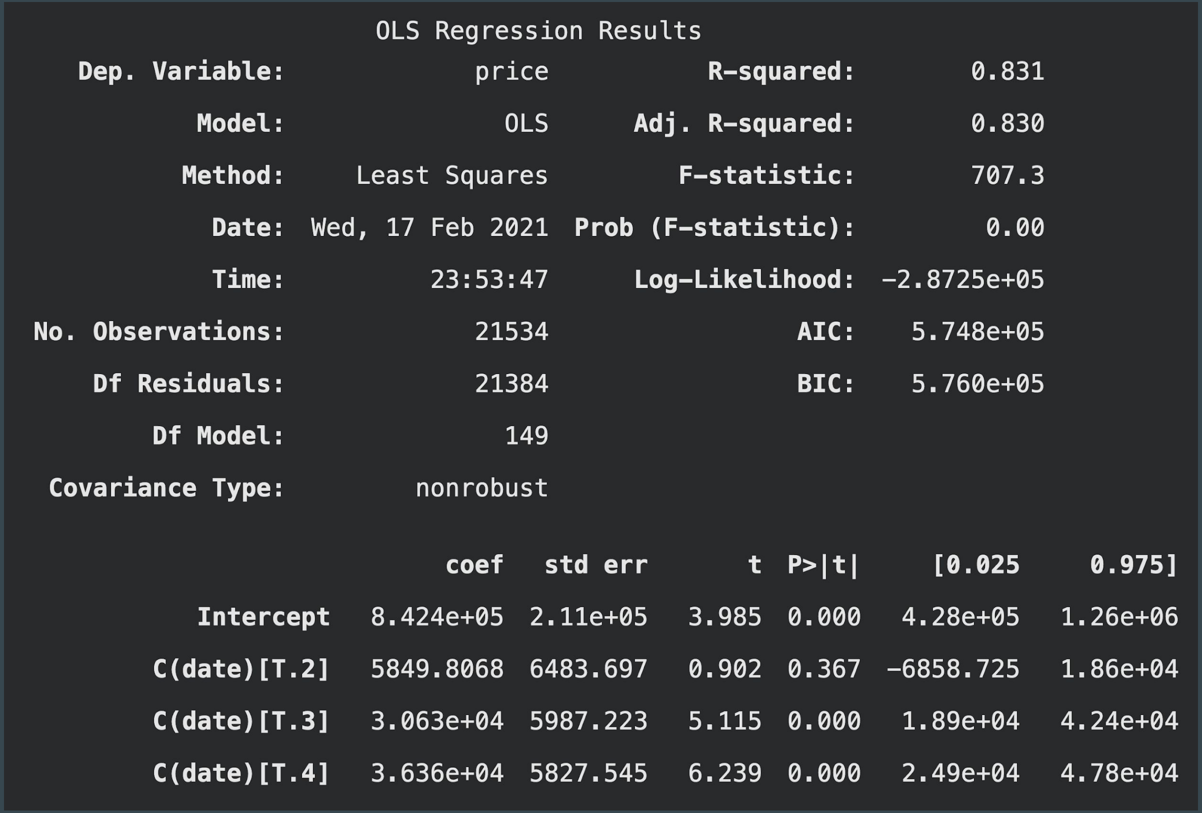
num columns:

[sqft_living, sqft_lot, sqft_above, sqft_basement, yr_built, sqft_living15, sqft_lot15]

cat columns:

[date, bedrooms, bathrooms, floors, view, condition, grade, zipcode]

5. Ordinary Linear Regression



```
                    OLS Regression Results
    Dep. Variable:            price          R-squared:       0.831
            Model:              OLS      Adj. R-squared:      0.830
           Method:    Least Squares         F-statistic:      707.3
             Date: Wed, 17 Feb 2021 Prob (F-statistic):        0.00
             Time:         23:53:47      Log-Likelihood: -2.8725e+05
No. Observations:            21534                 AIC:   5.748e+05
     Df Residuals:           21384                 BIC:   5.760e+05
         Df Model:             149
 Covariance Type:        nonrobust

                    coef    std err       t  P>|t|     [0.025     0.975]
     Intercept  8.424e+05  2.11e+05   3.985  0.000   4.28e+05   1.26e+06
  C(date)[T.2]  5849.8068  6483.697   0.902  0.367  -6858.725   1.86e+04
  C(date)[T.3]  3.063e+04  5987.223   5.115  0.000   1.89e+04   4.24e+04
  C(date)[T.4]  3.636e+04  5827.545   6.239  0.000   2.49e+04   4.78e+04
```

END