King County House Prices Exploratory Data Analyses

Stakeholder : Bonnie Brown (Seller)

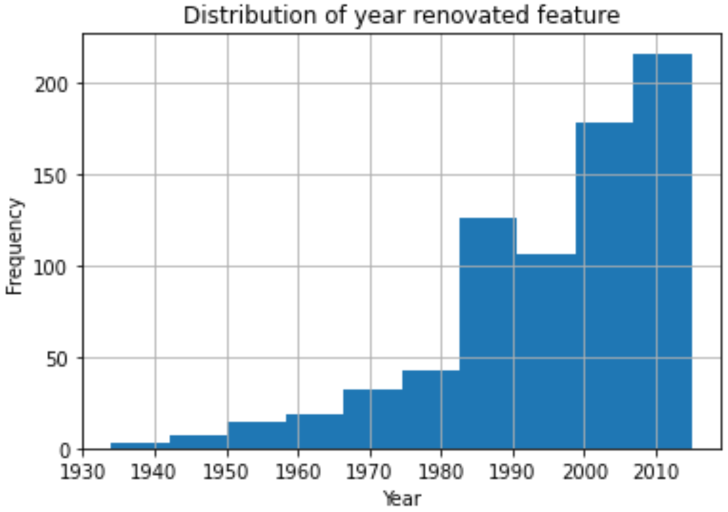
He as a house and wants to move soon (timing?), but wants high profit in middle class NH (neighborhood)

Hypothesis:

- House prices are high in city center
- waterfront Houses are expensive
- Best time for selling a house in mid of the year (Jun –July)
- After renovation the price will be increased

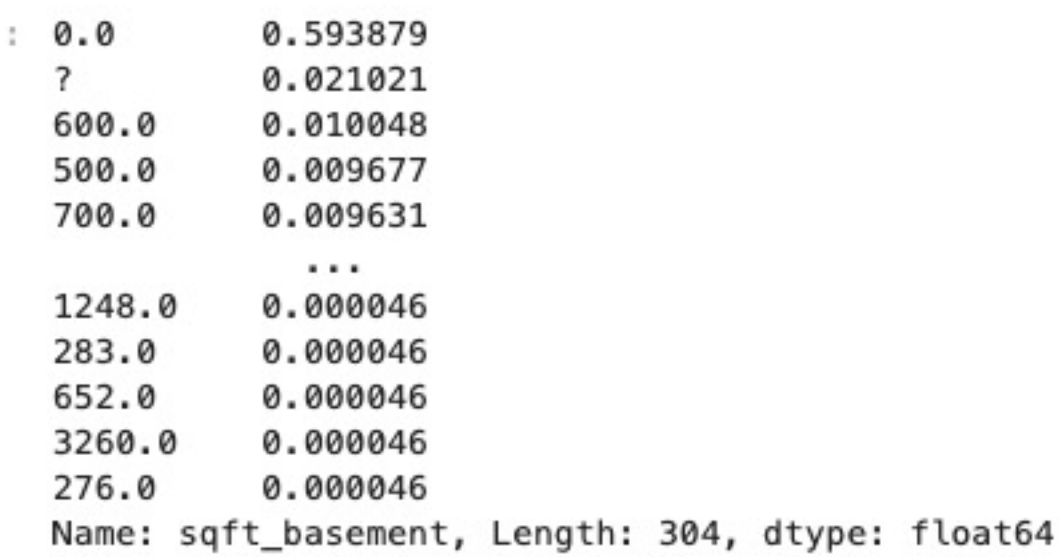
Data Cleaning

- **date** feature is converted to DateTime type
- **NaN** values are replaced:
  - waterfront 99% of the values are 0 therefore all NAN values in this data series are replaced by 0.
  - view 90% of the values are with 0's and this data is categorical variable data. In this analyses this data will be with 0's and 1's i.e. 0 = not viewed, 1 = viewed. Therefore all NAN values are replaced by 0's
    new feature = has_veiwed
  - yr_renovated this is also kind of categorical data like renovated = 1, not renovated = 0 in last 15 years. All renovations before 1990 are obsolete and assigned to 0. Similarly all NaNs are replaced with 0.
    - New feature = has_renovated

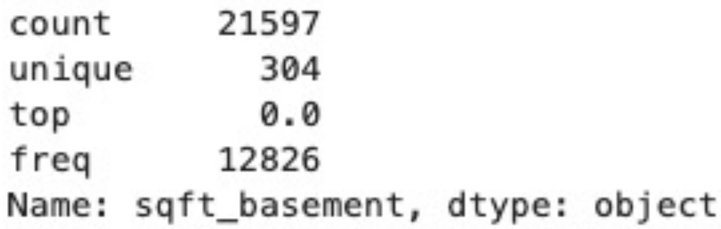


Data Cleaning

- The following analyses shows that sqft_basement feature is identified as object instead of float since it has `?` character in the dataset and 59% of the houses haven’t basements.
- For quick analyses this feature also considered as categorical variable feature like house has basements = 1 or not = 0
- A new feature has_basement feature created and dropped sqft_basement feature



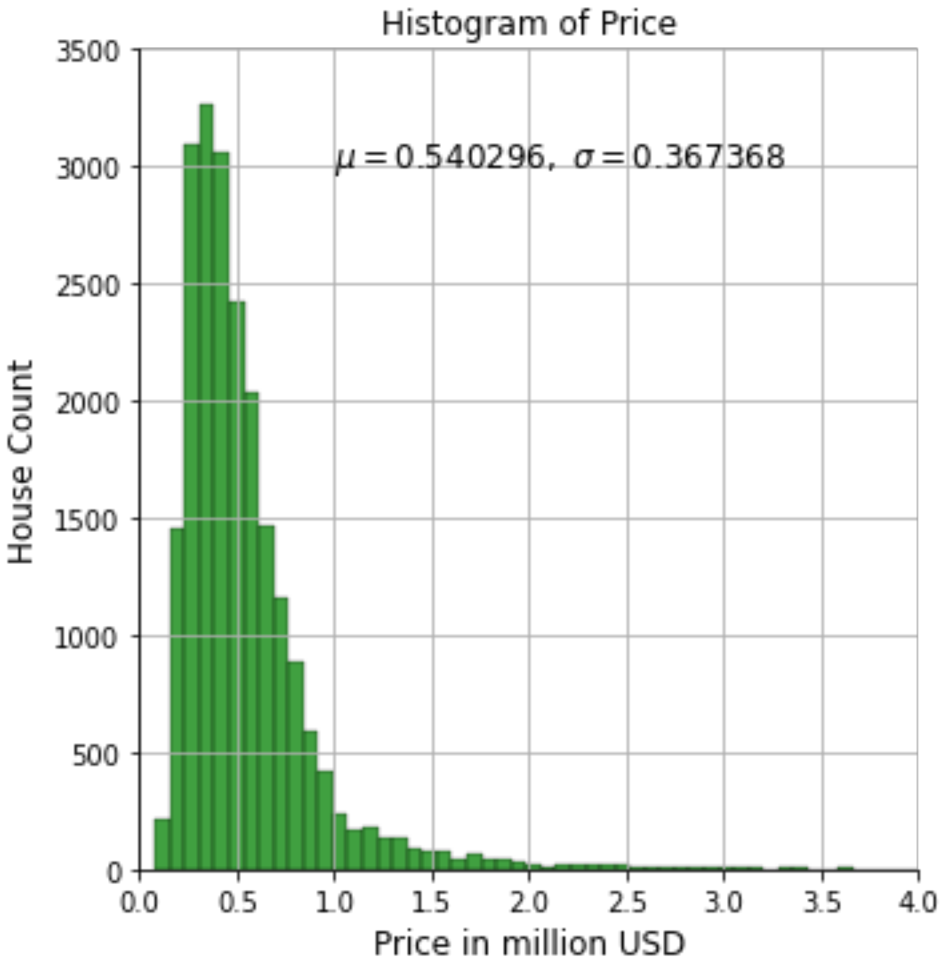
```
: 0.0       0.593879
  ?         0.021021
  600.0     0.010048
  500.0     0.009677
  700.0     0.009631
                ...
  1248.0    0.000046
  283.0     0.000046
  652.0     0.000046
  3260.0    0.000046
  276.0     0.000046
  Name: sqft_basement, Length: 304, dtype: float64
```

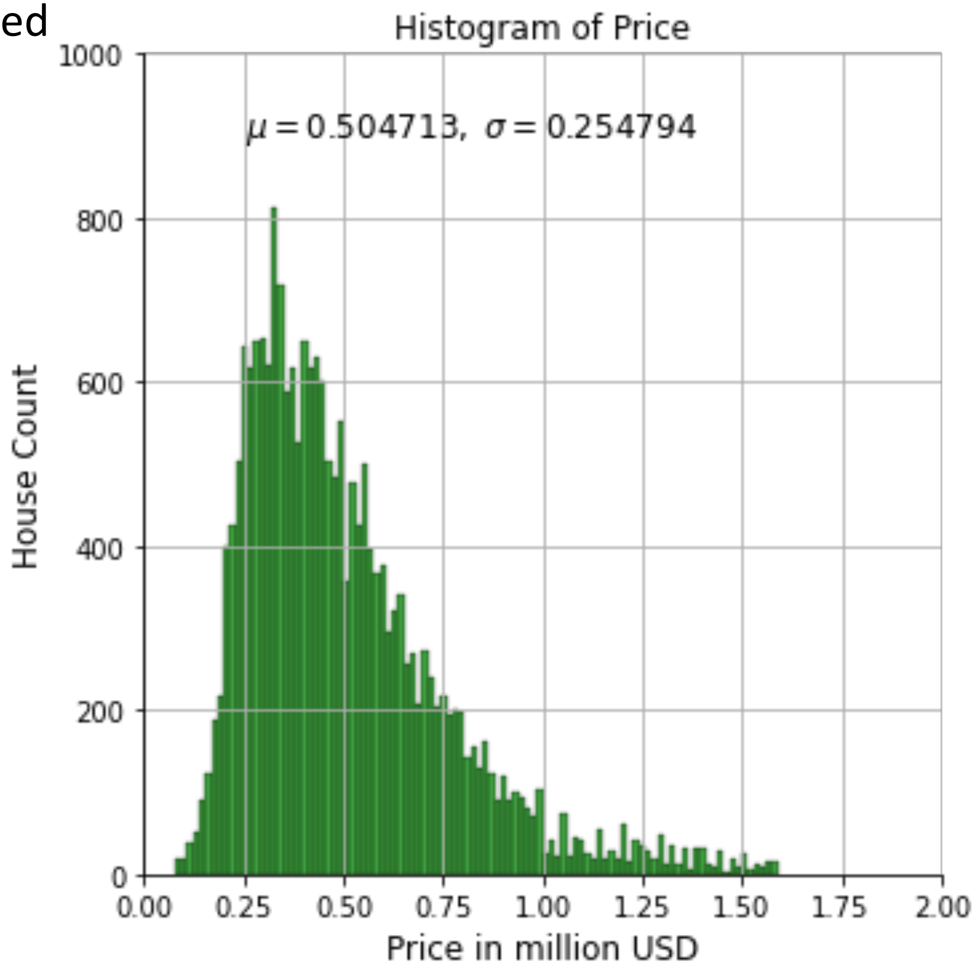


```
count     21597
unique      304
top         0.0
freq      12826
Name: sqft_basement, dtype: object
```

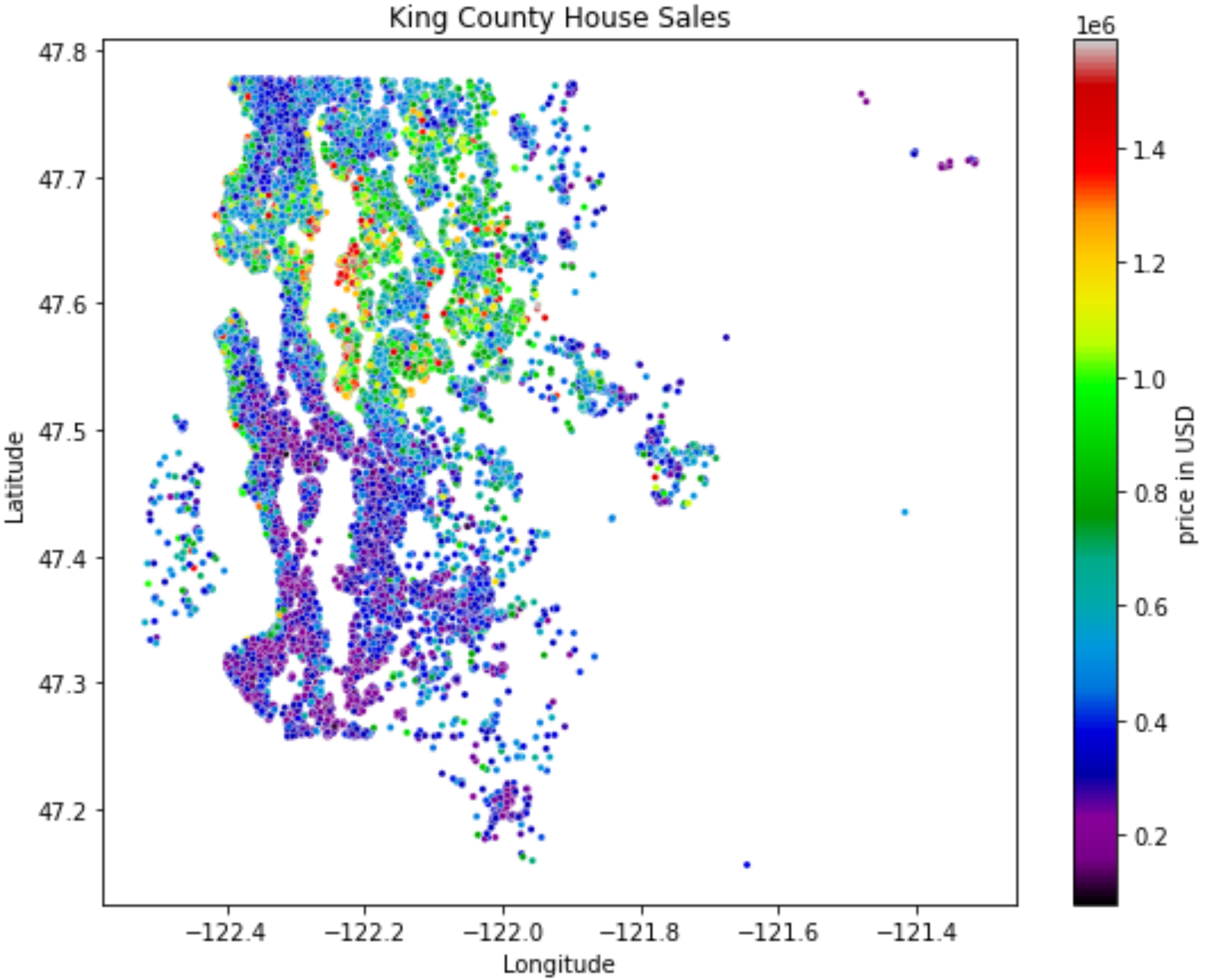
Analyses of price distribution

From price distribution graph, the houses with price above $1.6m are very few and can be considered as outliers. In this project for simplification the data of houses above 1.6m USD are removed

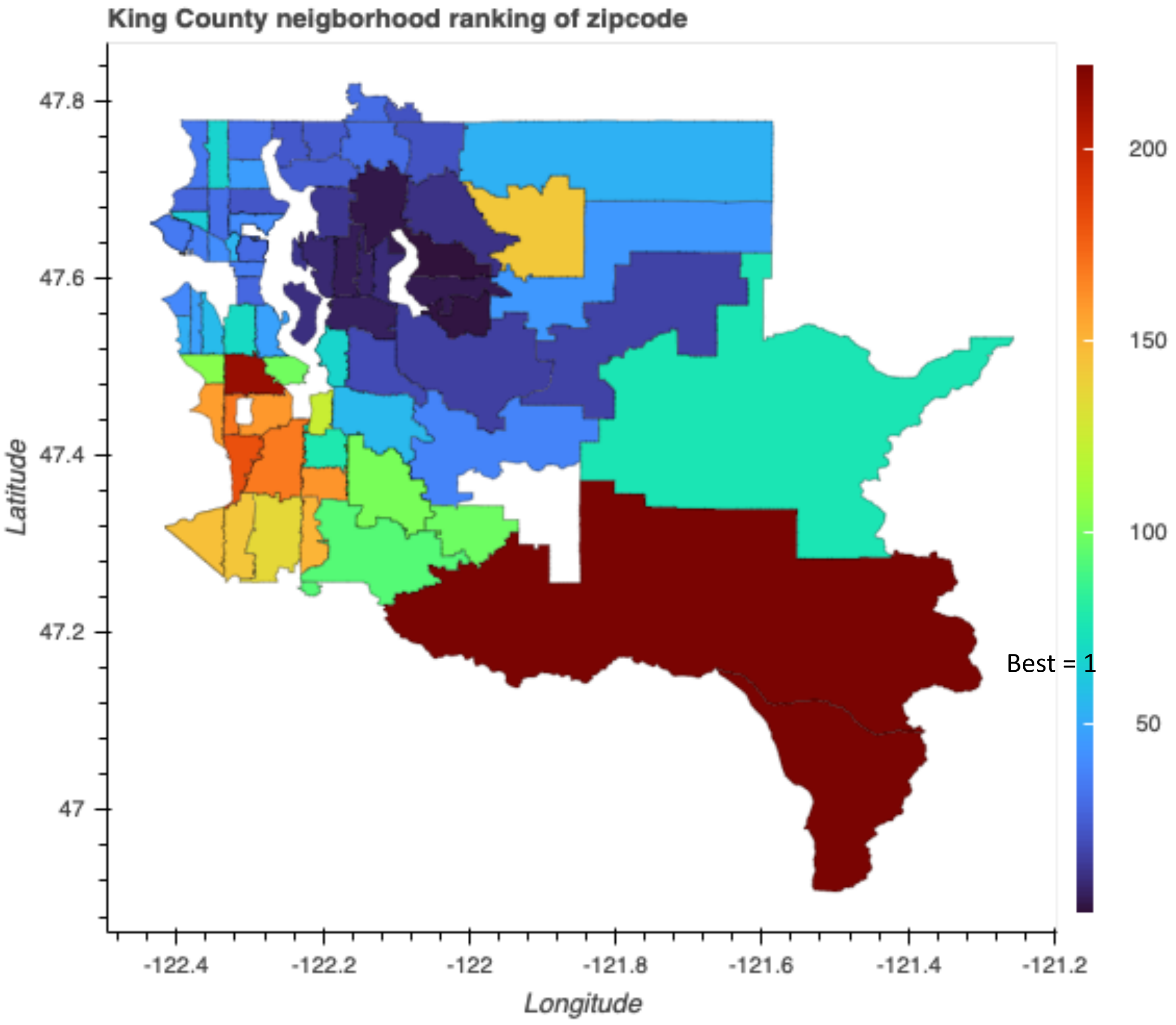




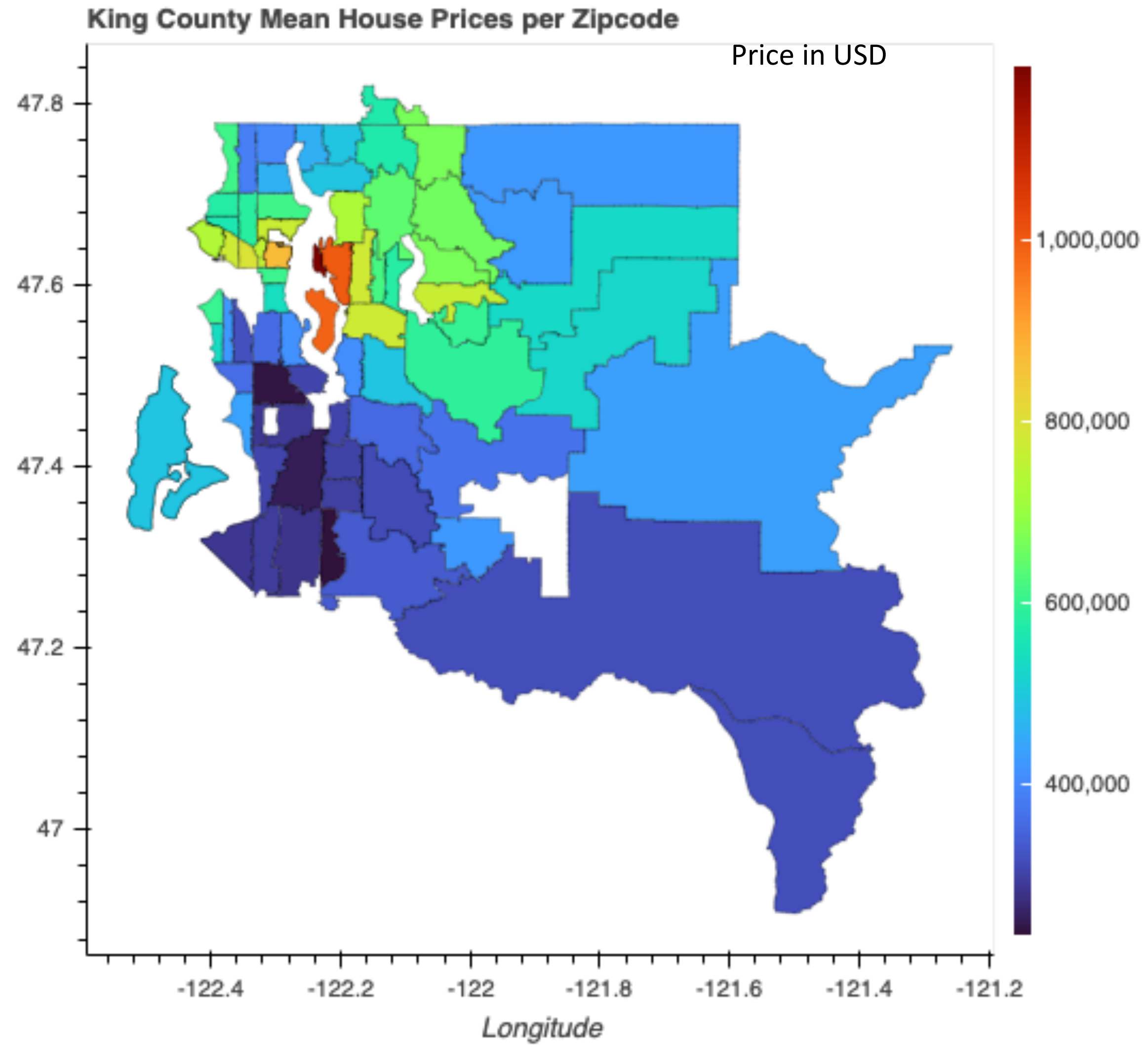
House price distribution



Analyses of price distribution and neighborhood



King County neigborhood ranking of zipcode

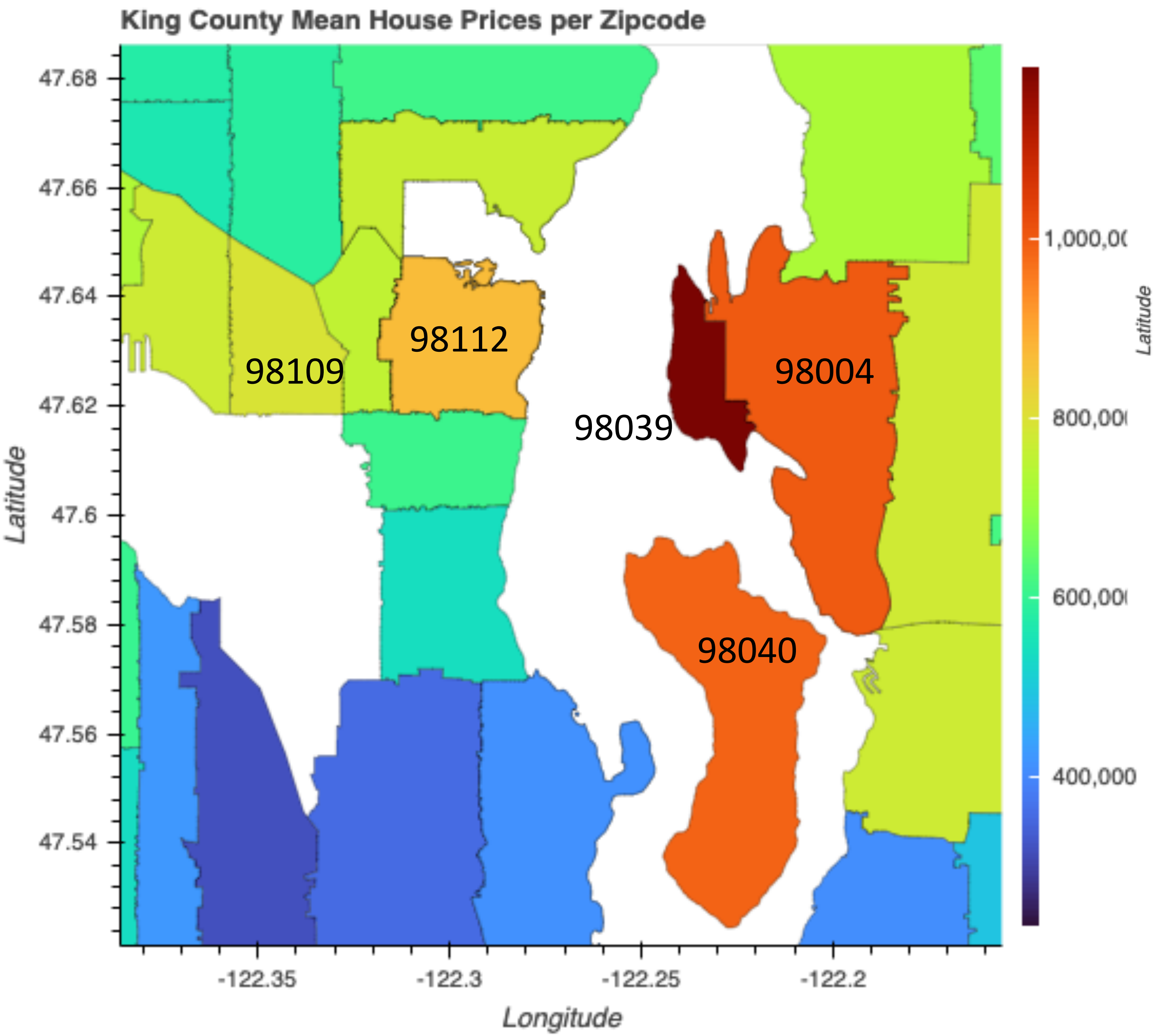


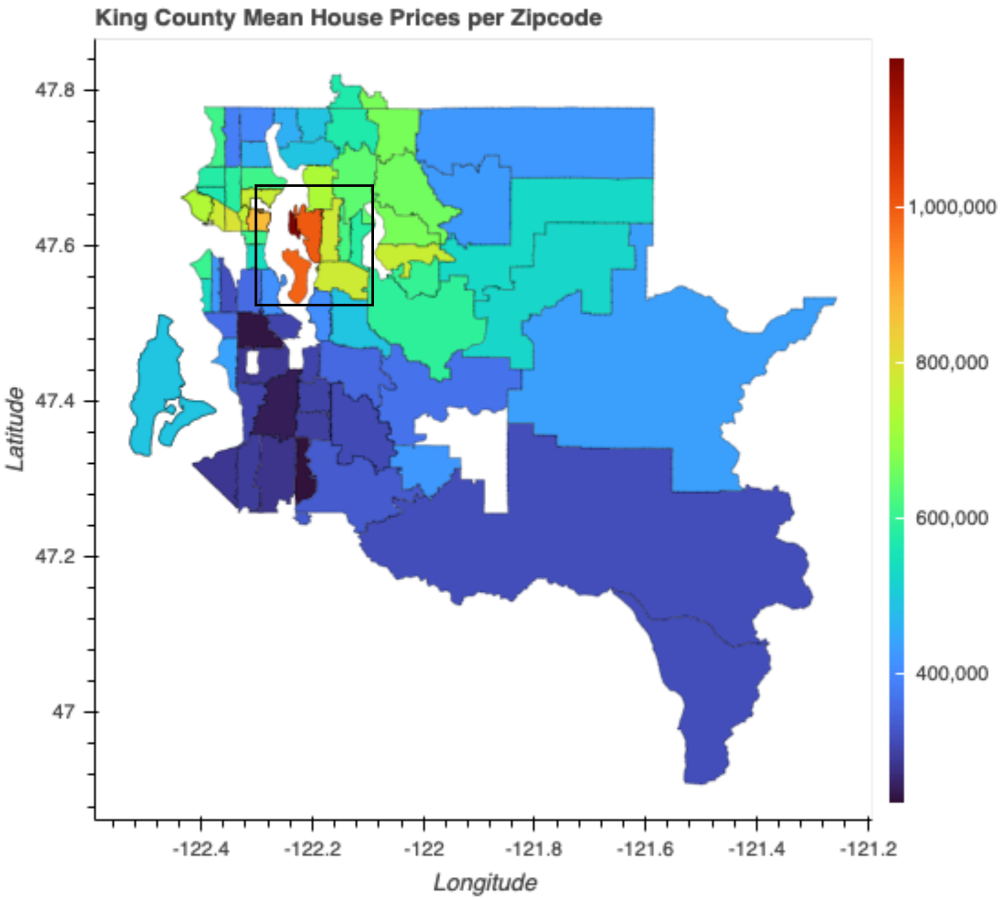
King County Mean House Prices per Zipcode

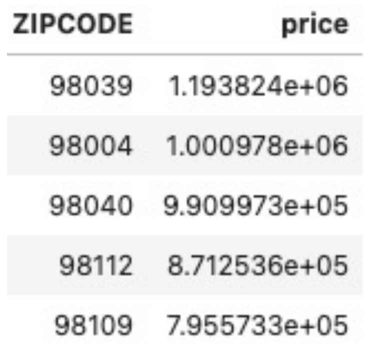
Price in USD

houses above 1.6mUSD are not considered in this analyses

Max mean price







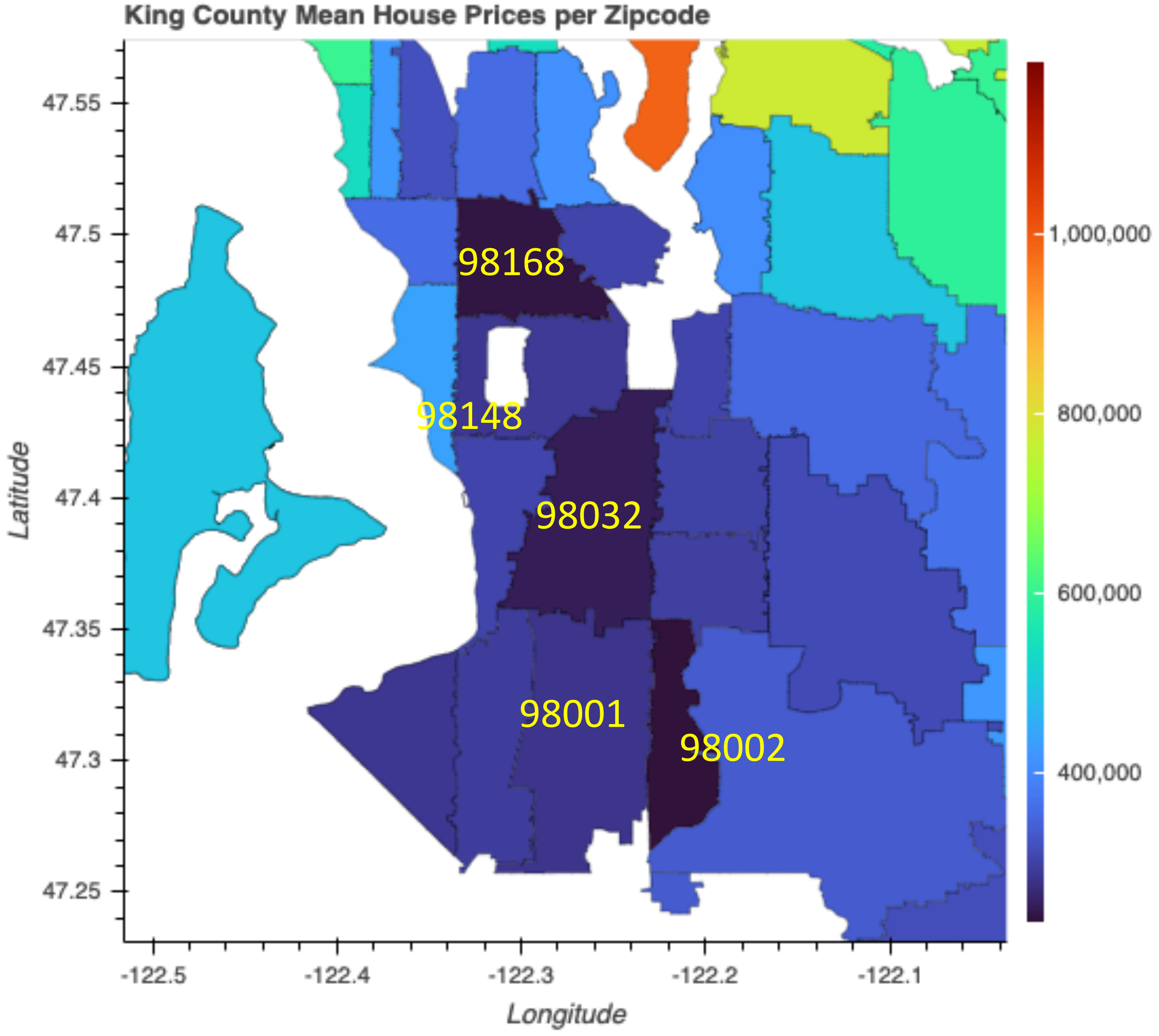
ZIPCODE	price
98039	1.193824e+06
98004	1.000978e+06
98040	9.909973e+05
98112	8.712536e+05
98109	7.955733e+05

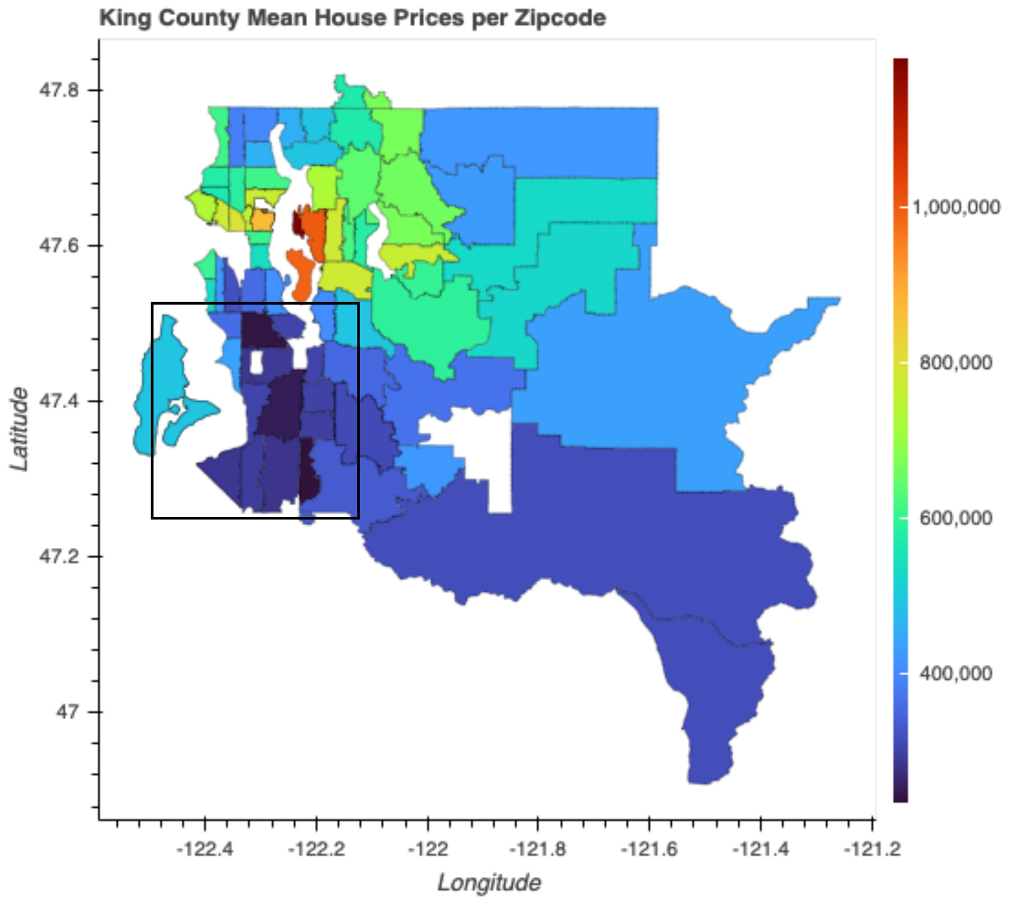
Top 5 places:

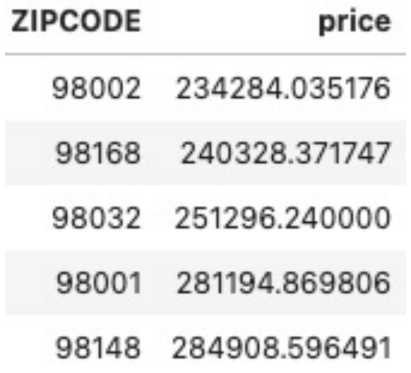
- Medina
- Clyde Hill /
- Mercer Island
- Madison Park
- Kenmore Air Harbor

houses above 1.6mUSD are not considered in this analyses

Min mean price







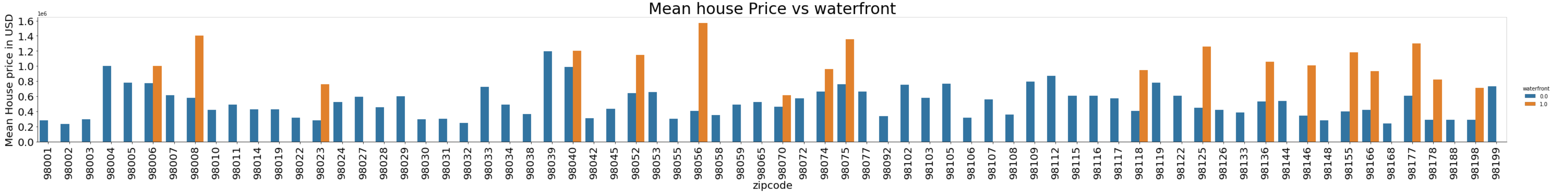
ZIPCODE	price
98002	234284.035176
98168	240328.371747
98032	251296.240000
98001	281194.869806
98148	284908.596491

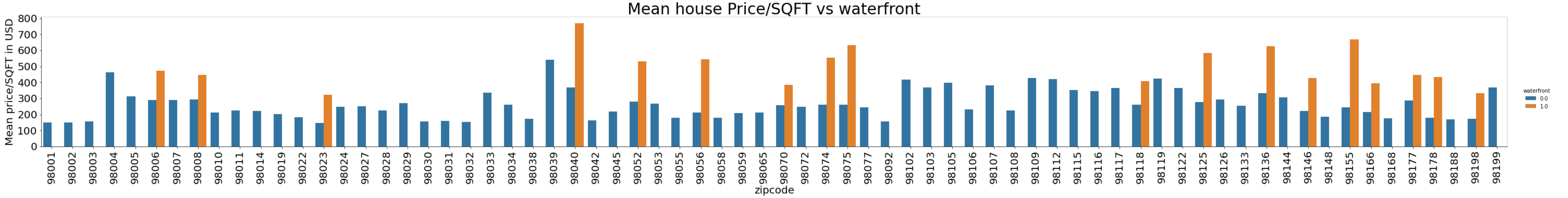
Bottom 5 places:

- South Auburn
- Burien
- Kent
- Auburn
- Burien

houses above 1.6mUSD are not considered in this analyses

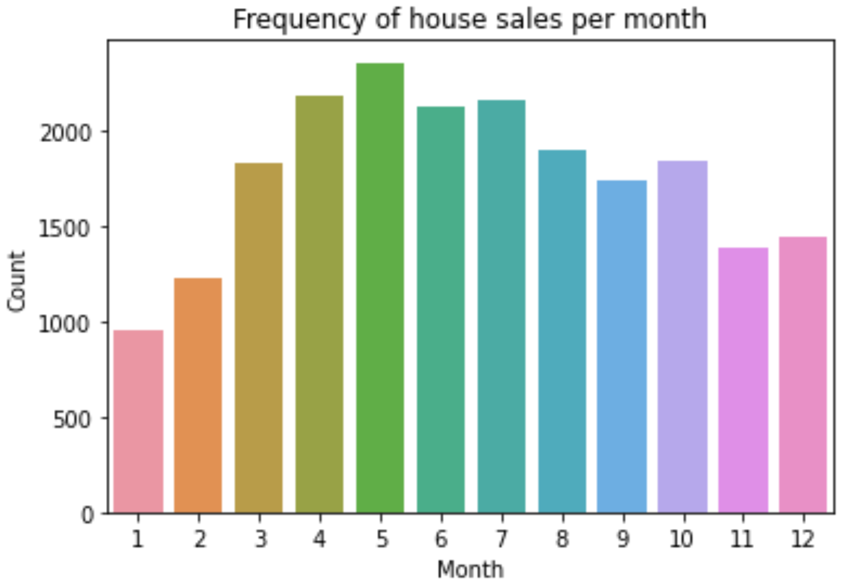
Sales, price vs waterfront

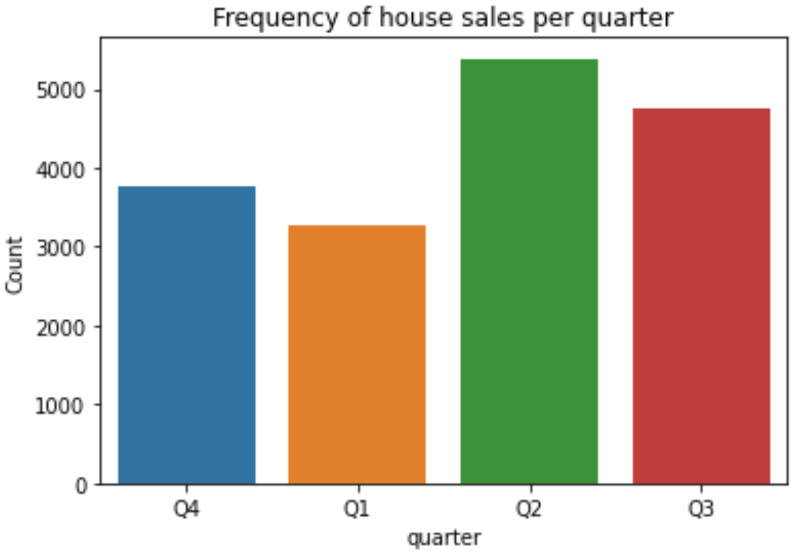

Mean house Price vs waterfront

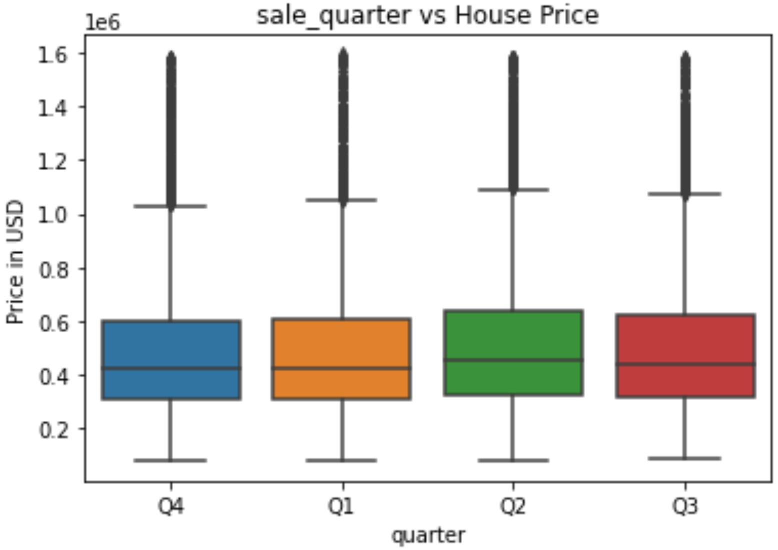

Mean house Price/SQFT vs waterfront

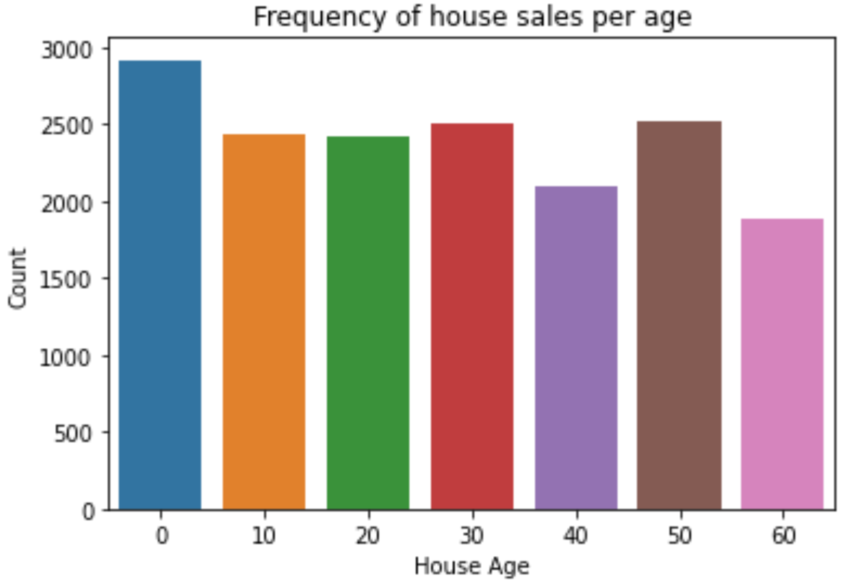
houses above 1.6mUSD are not considered in this analyses

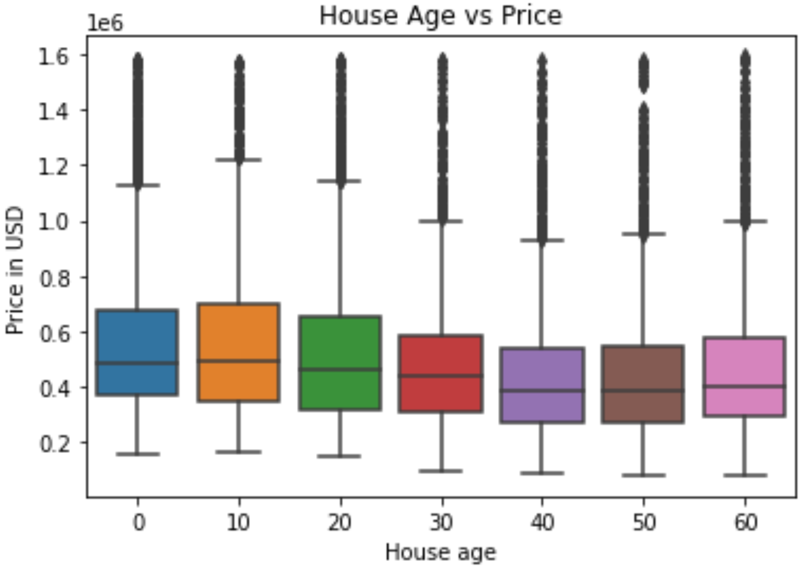
House sales analyses


Frequency of house sales per month


Frequency of house sales per quarter

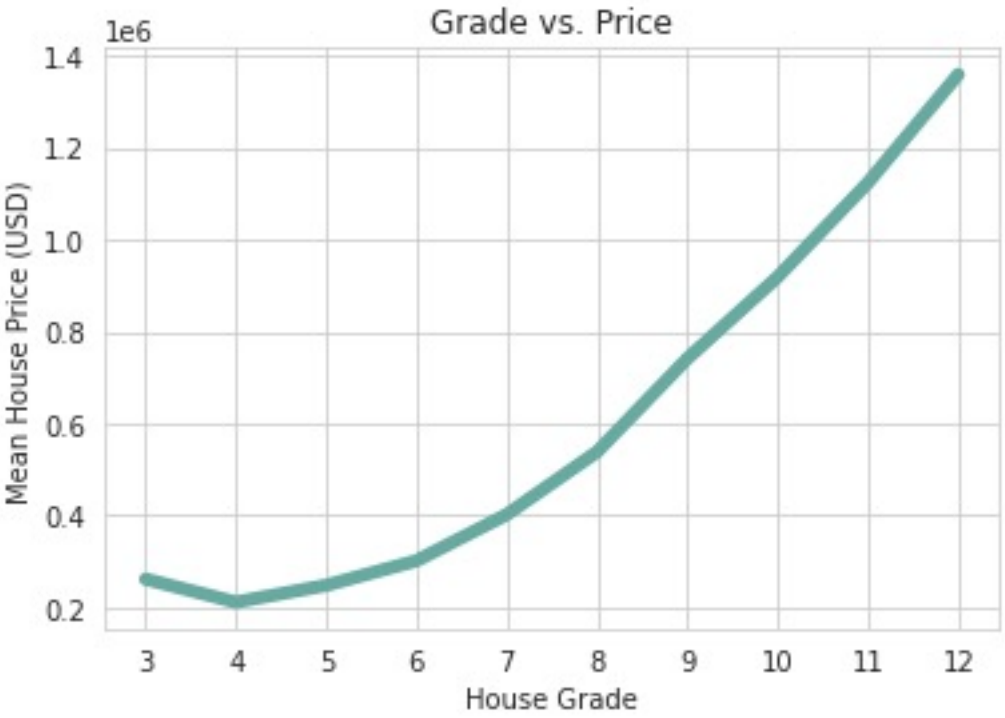

sale_quarter vs House Price

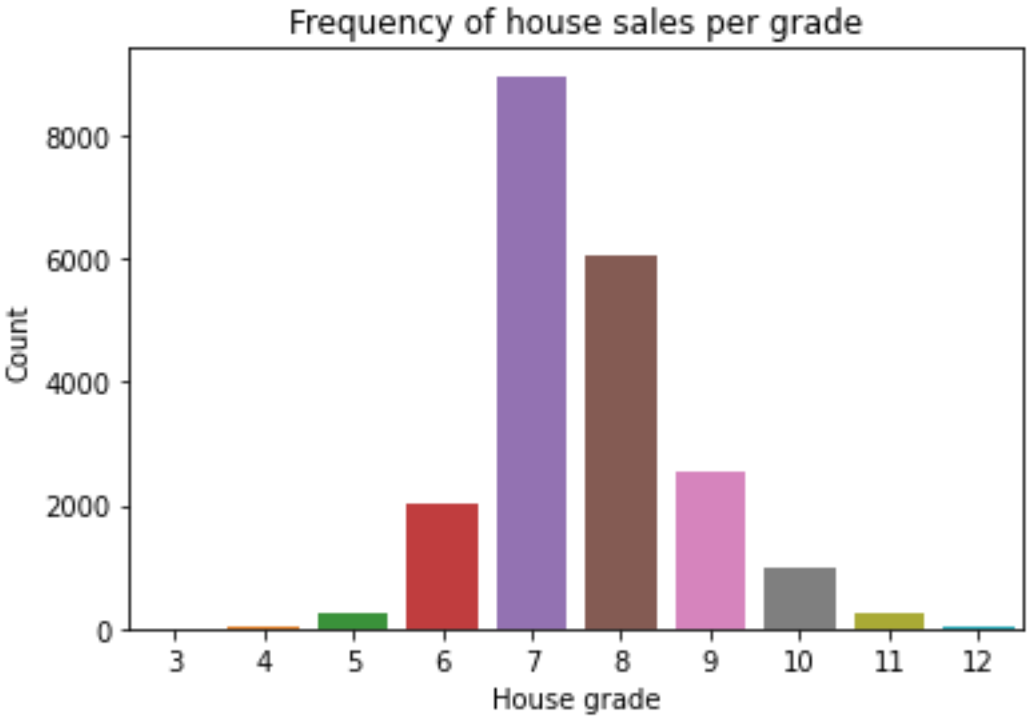

Frequency of house sales per age


House Age vs Price

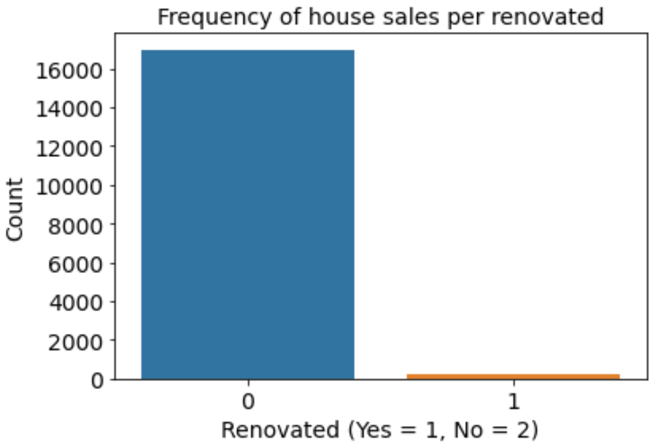
houses above 1.6mUSD are not considered in this analyses

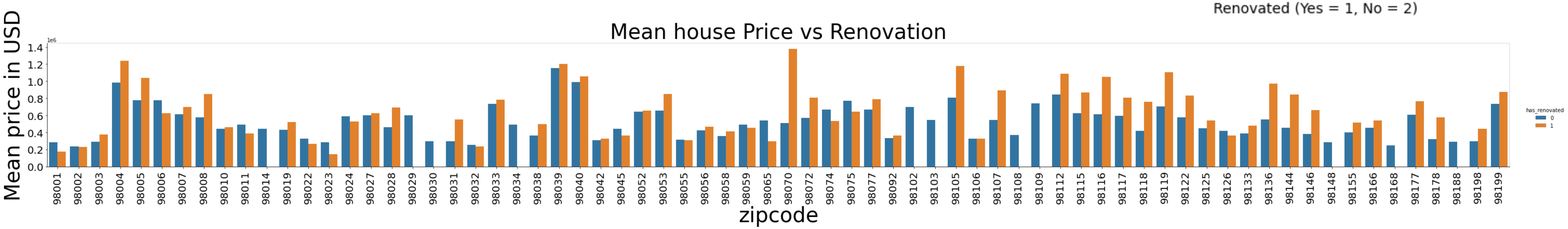
House sales per grade

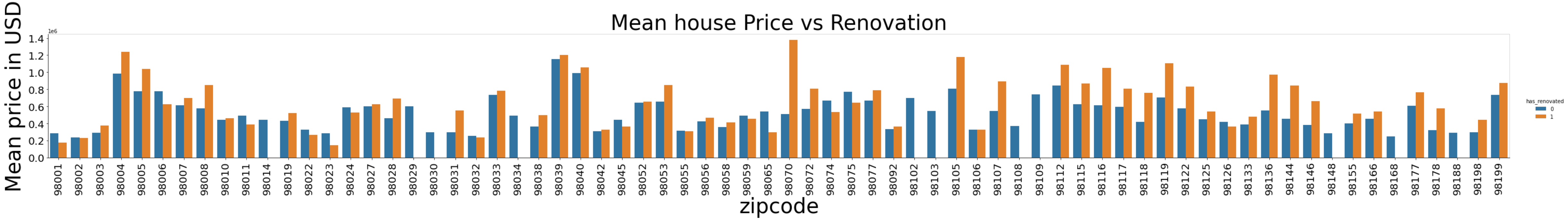




Renovation vs sales

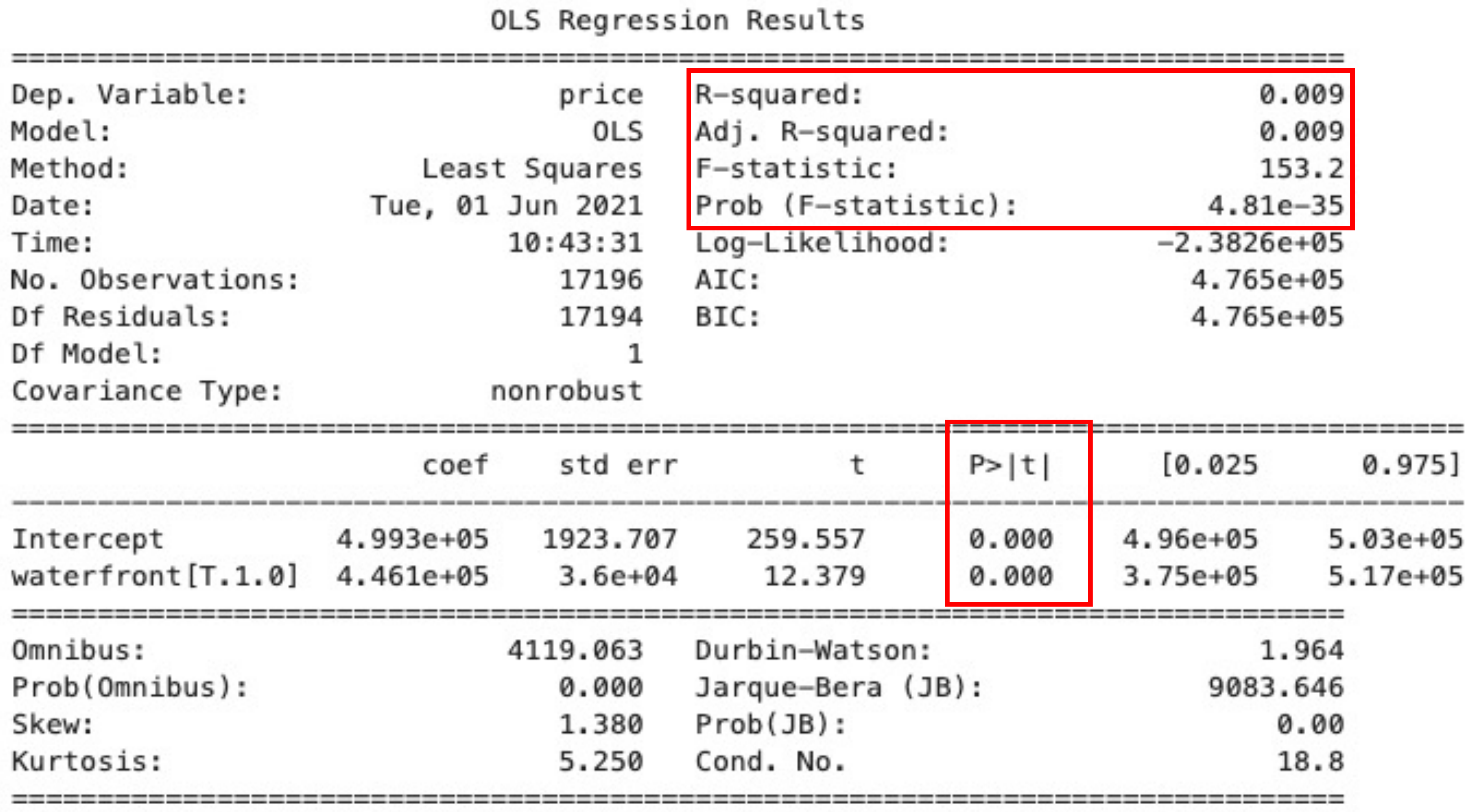






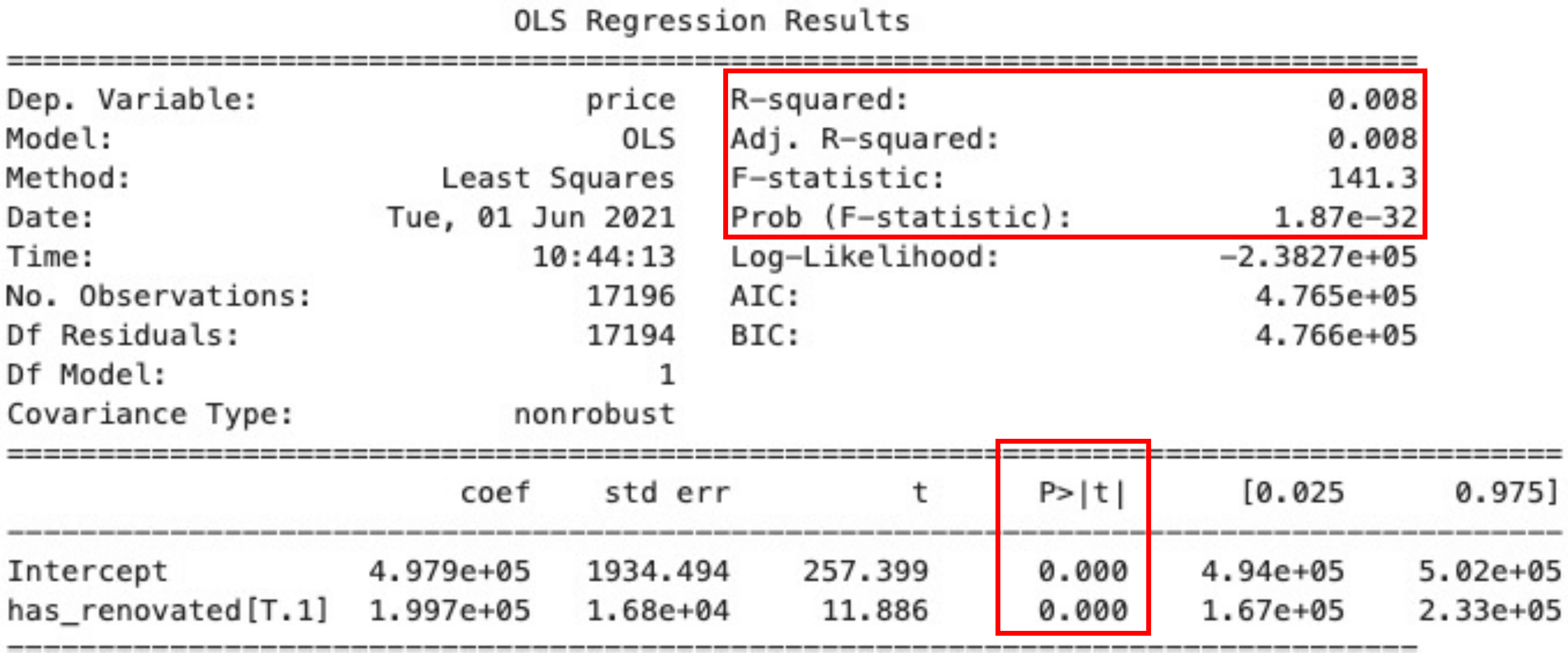
Post 1990renovations are obsolated, houses above 1.6mUSD are not considered in this analyses

Price vs waterfront



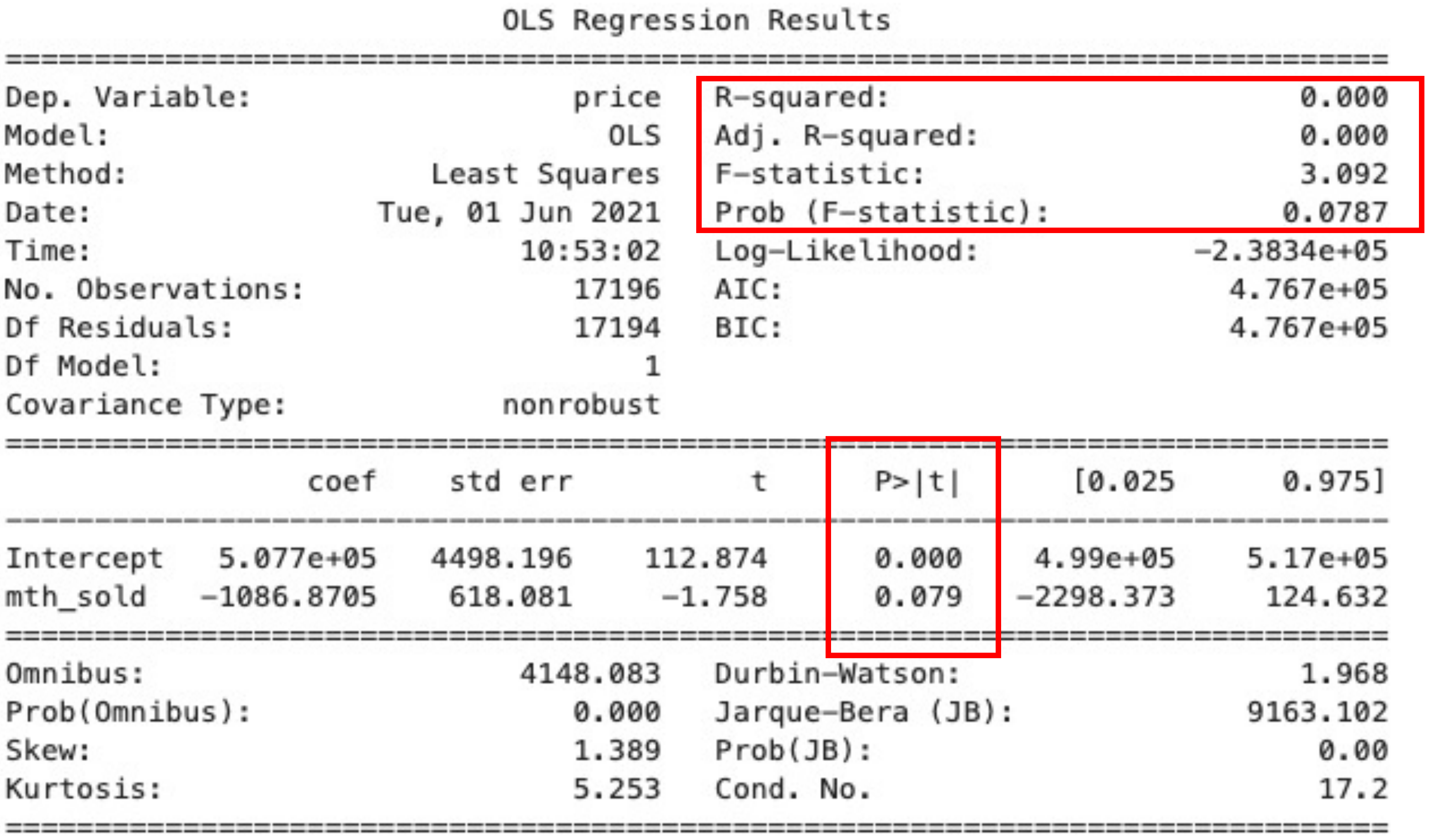
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.009
Model:                            OLS   Adj. R-squared:                  0.009
Method:                 Least Squares   F-statistic:                     153.2
Date:                Tue, 01 Jun 2021   Prob (F-statistic):           4.81e-35
Time:                        10:43:31   Log-Likelihood:            -2.3826e+05
No. Observations:               17196   AIC:                         4.765e+05
Df Residuals:                   17194   BIC:                         4.765e+05
Df Model:                           1
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept          4.993e+05   1923.707    259.557      0.000    4.96e+05    5.03e+05
waterfront[T.1.0]  4.461e+05    3.6e+04     12.379      0.000    3.75e+05    5.17e+05
==============================================================================
Omnibus:                     4119.063   Durbin-Watson:                   1.964
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9083.646
Skew:                           1.380   Prob(JB):                         0.00
Kurtosis:                       5.250   Cond. No.                         18.8
==============================================================================
```

Price vs renovation



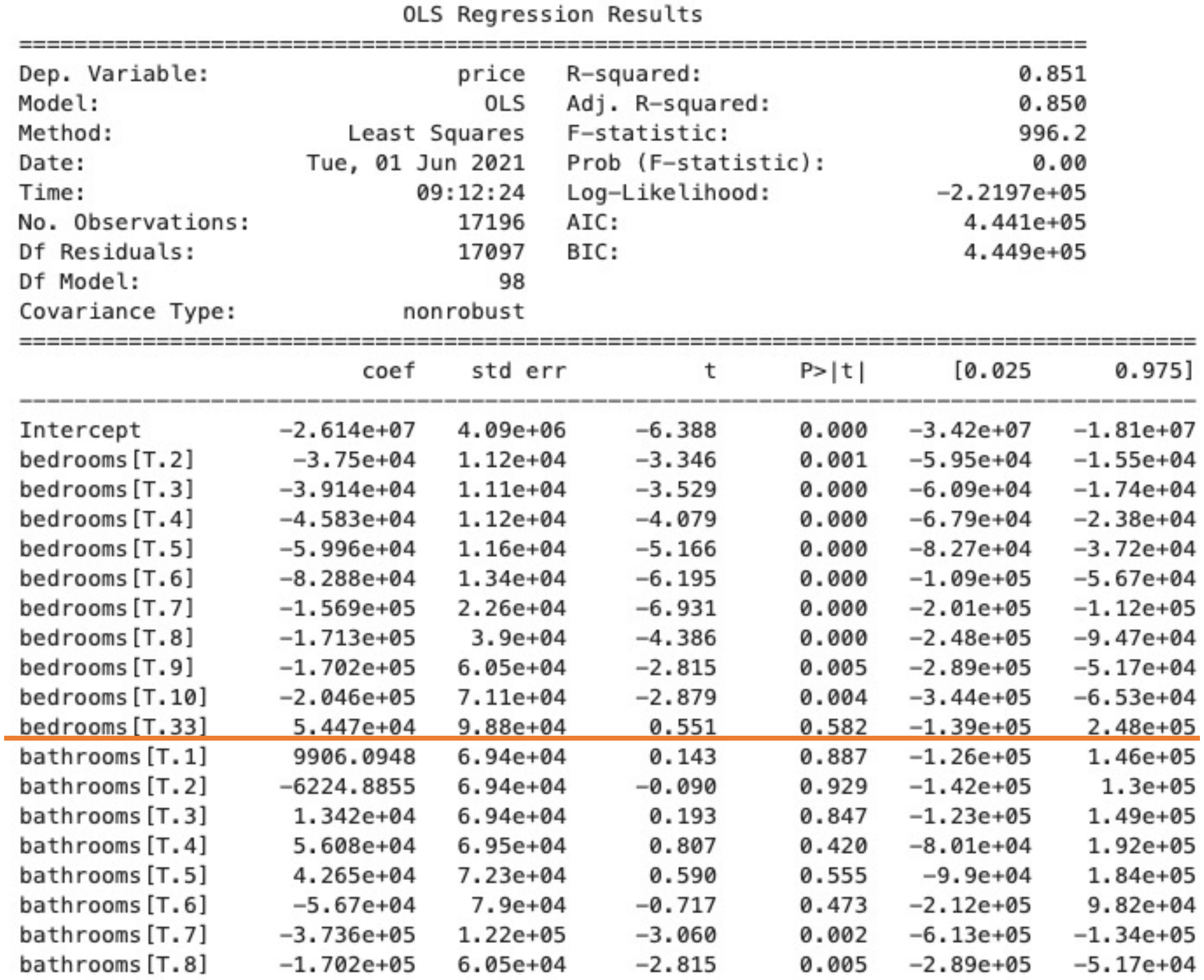
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.008
Model:                            OLS   Adj. R-squared:                  0.008
Method:                 Least Squares   F-statistic:                     141.3
Date:                Tue, 01 Jun 2021   Prob (F-statistic):           1.87e-32
Time:                        10:44:13   Log-Likelihood:            -2.3827e+05
No. Observations:               17196   AIC:                         4.765e+05
Df Residuals:                   17194   BIC:                         4.766e+05
Df Model:                           1
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept           4.979e+05   1934.494    257.399      0.000    4.94e+05    5.02e+05
has_renovated[T.1]  1.997e+05    1.68e+04     11.886      0.000    1.67e+05    2.33e+05
==============================================================================
```

Price vs month



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     3.092
Date:                Tue, 01 Jun 2021   Prob (F-statistic):             0.0787
Time:                        10:53:02   Log-Likelihood:            -2.3834e+05
No. Observations:               17196   AIC:                         4.767e+05
Df Residuals:                   17194   BIC:                         4.767e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   5.077e+05   4498.196    112.874      0.000    4.99e+05    5.17e+05
mth_sold   -1086.8705    618.081     -1.758      0.079   -2298.373     124.632
==============================================================================
Omnibus:                     4148.083   Durbin-Watson:                   1.968
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9163.102
Skew:                           1.389   Prob(JB):                         0.00
Kurtosis:                       5.253   Cond. No.                         17.2
==============================================================================
```

Price vs all features (incl. dummies)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.851
Model:                            OLS   Adj. R-squared:                  0.850
Method:                 Least Squares   F-statistic:                     996.2
Date:                Tue, 01 Jun 2021   Prob (F-statistic):               0.00
Time:                        09:12:24   Log-Likelihood:            -2.2197e+05
No. Observations:               17196   AIC:                         4.441e+05
Df Residuals:                   17097   BIC:                         4.449e+05
Df Model:                          98
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept          -2.614e+07   4.09e+06     -6.388      0.000   -3.42e+07   -1.81e+07
bedrooms[T.2]       -3.75e+04   1.12e+04     -3.346      0.001   -5.95e+04   -1.55e+04
bedrooms[T.3]      -3.914e+04   1.11e+04     -3.529      0.000   -6.09e+04   -1.74e+04
bedrooms[T.4]      -4.583e+04   1.12e+04     -4.079      0.000   -6.79e+04   -2.38e+04
bedrooms[T.5]      -5.996e+04   1.16e+04     -5.166      0.000   -8.27e+04   -3.72e+04
bedrooms[T.6]      -8.288e+04   1.34e+04     -6.195      0.000   -1.09e+05   -5.67e+04
bedrooms[T.7]      -1.569e+05   2.26e+04     -6.931      0.000   -2.01e+05   -1.12e+05
bedrooms[T.8]      -1.713e+05    3.9e+04     -4.386      0.000   -2.48e+05   -9.47e+04
bedrooms[T.9]      -1.702e+05   6.05e+04     -2.815      0.005   -2.89e+05   -5.17e+04
bedrooms[T.10]     -2.046e+05   7.11e+04     -2.879      0.004   -3.44e+05   -6.53e+04
bedrooms[T.33]      5.447e+04   9.88e+04      0.551      0.582   -1.39e+05    2.48e+05
bathrooms[T.1]      9906.0948   6.94e+04      0.143      0.887   -1.26e+05    1.46e+05
bathrooms[T.2]     -6224.8855   6.94e+04     -0.090      0.929   -1.42e+05     1.3e+05
bathrooms[T.3]      1.342e+04   6.94e+04      0.193      0.847   -1.23e+05    1.49e+05
bathrooms[T.4]      5.608e+04   6.95e+04      0.807      0.420   -8.01e+04    1.92e+05
bathrooms[T.5]      4.265e+04   7.23e+04      0.590      0.555    -9.9e+04    1.84e+05
bathrooms[T.6]      -5.67e+04    7.9e+04     -0.717      0.473   -2.12e+05    9.82e+04
bathrooms[T.7]     -3.736e+05   1.22e+05     -3.060      0.002   -6.13e+05   -1.34e+05
bathrooms[T.8]     -1.702e+05   6.05e+04     -2.815      0.005   -2.89e+05   -5.17e+04
```