EDA Project

Client: Jennifer Montgomery

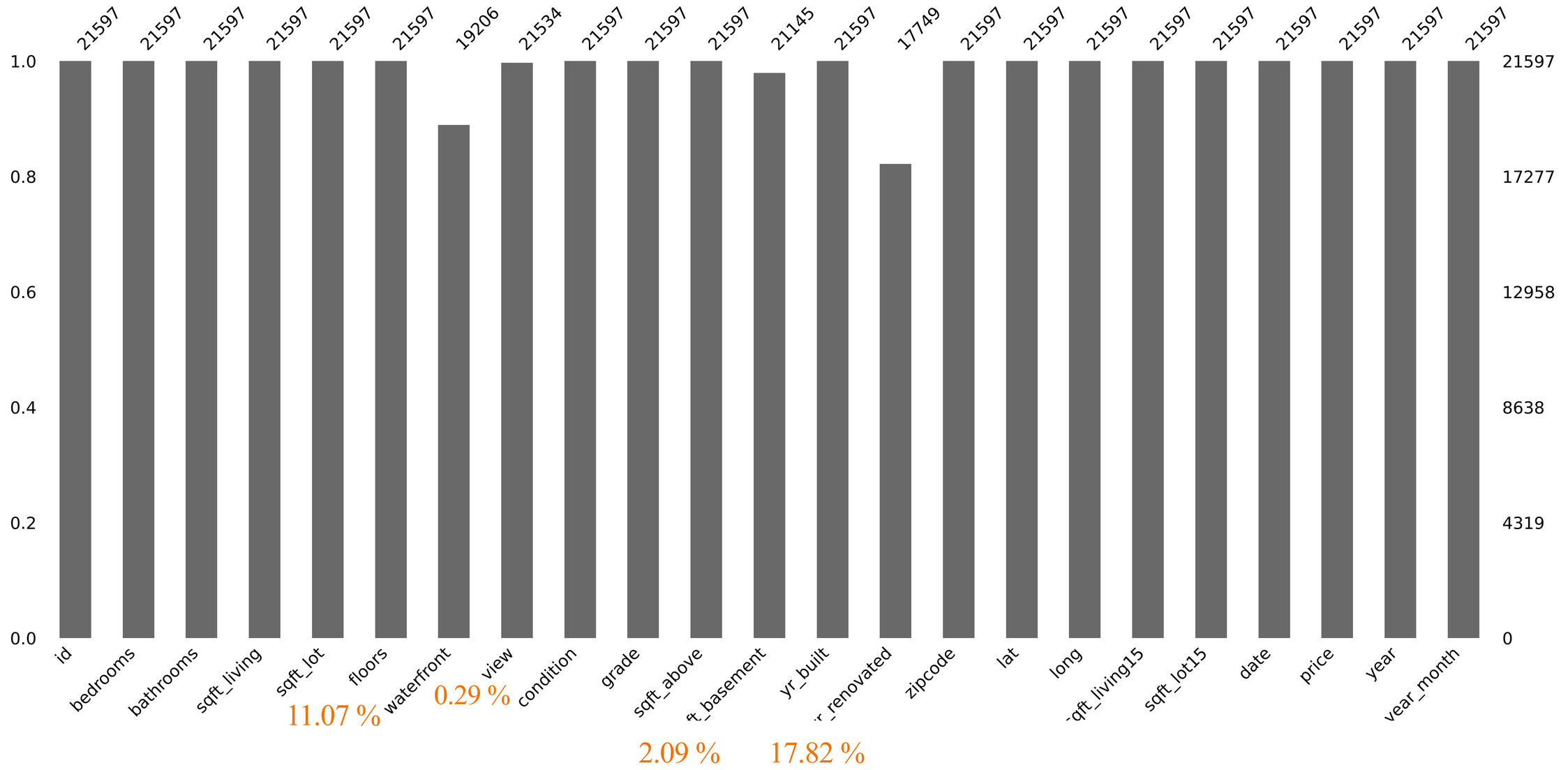
Requirements: High budget, wants to show off, timing within a month, waterfront, renovated, high grades, resell within 1 year (buyer)

Workflow

- Understanding data
- Data cleaning
- Feature engineering
- Data visualization
- Insights regarding over all data
- Data preparation for client
- Recommendation to client

Missing data

Total no of rows



- **waterfront**
- **view**
- **yr_renovated**
- **sqft_basement**

- **Over all missing values in data frame : 1.36 %**
- **No duplicate data sets**

Missing data and imputation

Column ‘year of renovation’

Column yr_renovation weird?

NaN

Zero

19900 → 1990

Fill Method

```
df_n0[["waterfront","yr_renovated","view", "sqft_basement"]] =
df_n0[["waterfront","yr_renovated","view", “sqft_basement"]].fillna(0)
df_n0[["waterfront","yr_renovated","view", "sqft_basement"]].reset_index()
df_n0['yr_renovated']=df_n0['yr_renovated'].apply(lambda x : str(x)[:4])
df_n0['yr_renovated'] = df_n0.yr_renovated.replace(‘nan',np.NaN)
df_n0['yr_renovated'] = df_n0[‘yr_renovated'].astype(float).astype('Int64')
df_n0 = df_n0.dropna().reset_index()
```

Hypothesis

Client requirements: High budget, wants to show off, timing within a month, waterfront, renovated, high grades, resell within 1 year (buyer)

- **Additional data: age of the house**

- **Additional data: age of the house renovation**

- **Additional data: price per sqft**

Feature engineering

Additional columns:

- change "date" dtype to datetime with format %Y/%m/%d.

  add two new columns: (1) year-month (2) year

- Additional column age of the house

```
df_n0['house_age'] =df_n0['date'].dt.year -df_n0['yr_built']
```

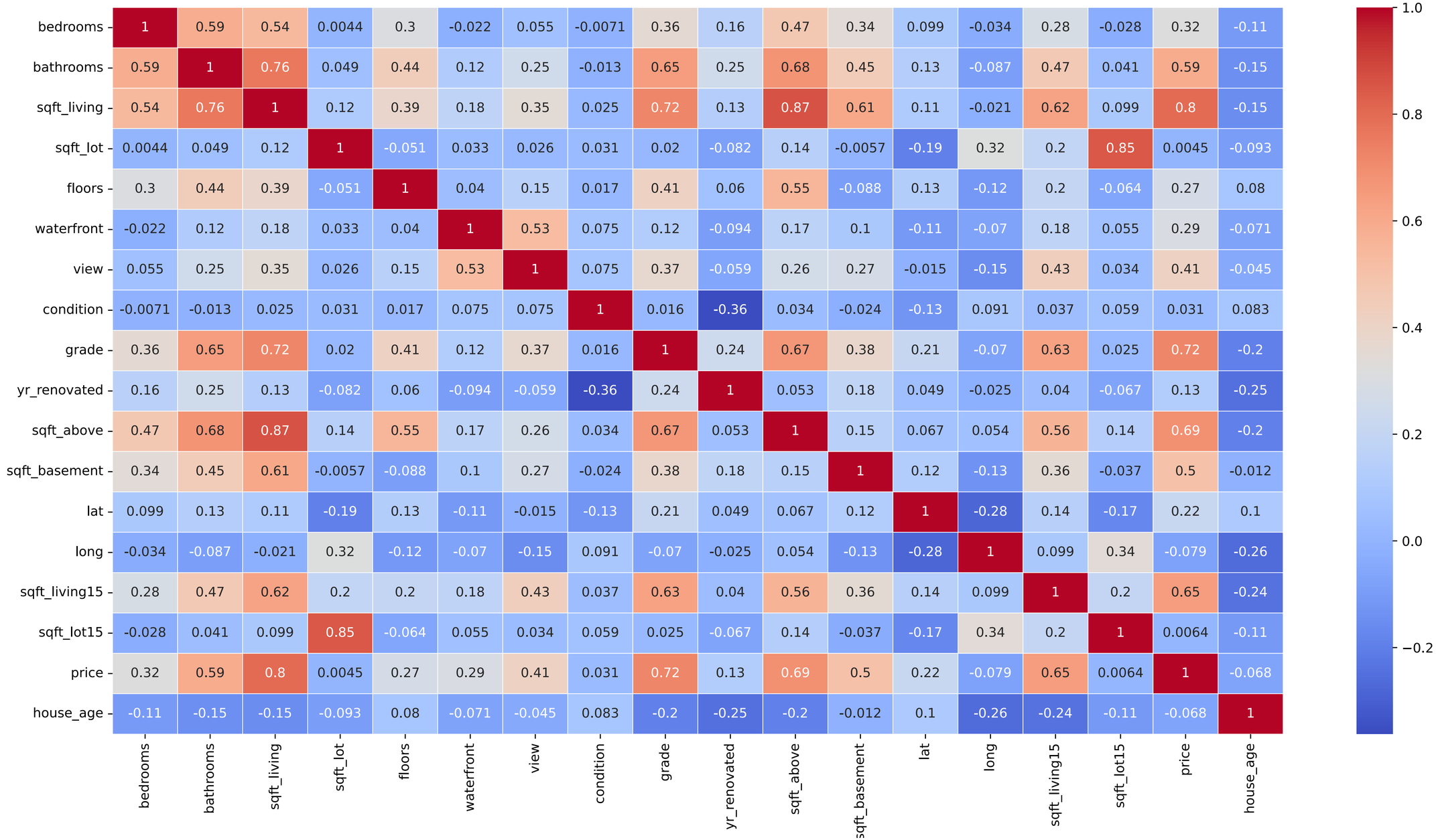
- Additional column age of the house renovation

```
df_n0['house_age_yr'] =df_n0['date'].dt.year -df_n0['yr_renovated']
```

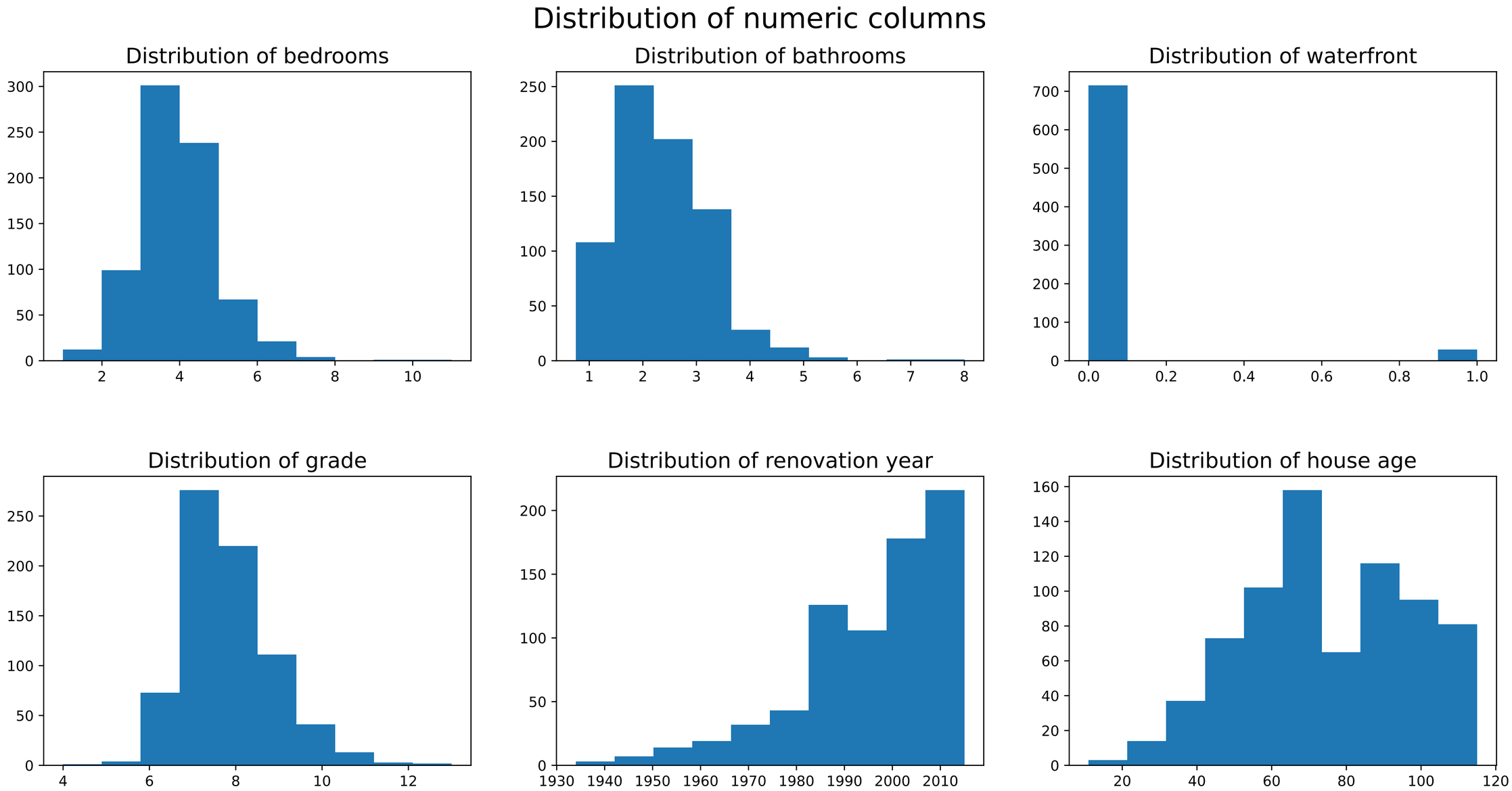
- Additional column price per sqft

```
df_n0['price_by_sqft'] = df_n0['price']/df_n0['sqft_lot']
```

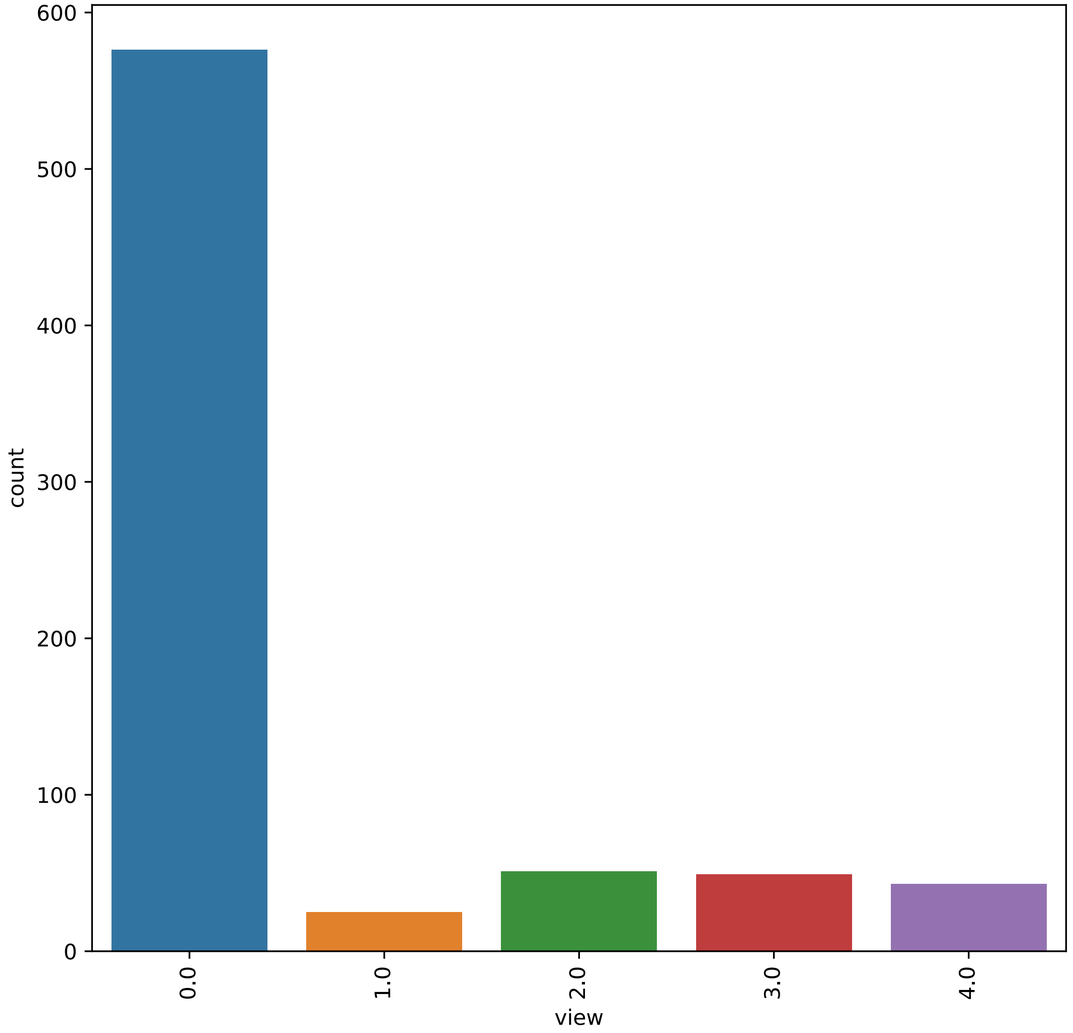
Correlation matrix



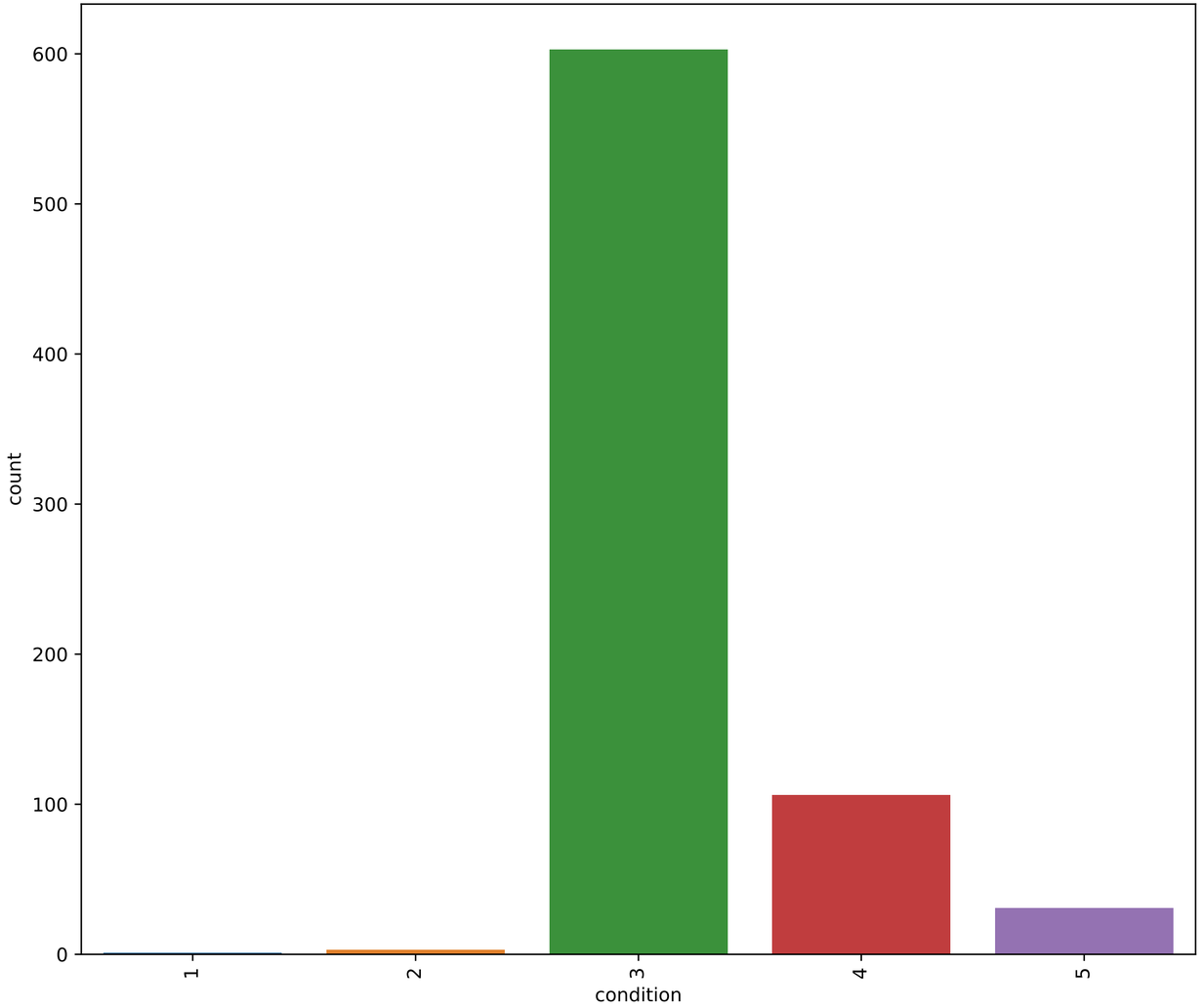
Histograms

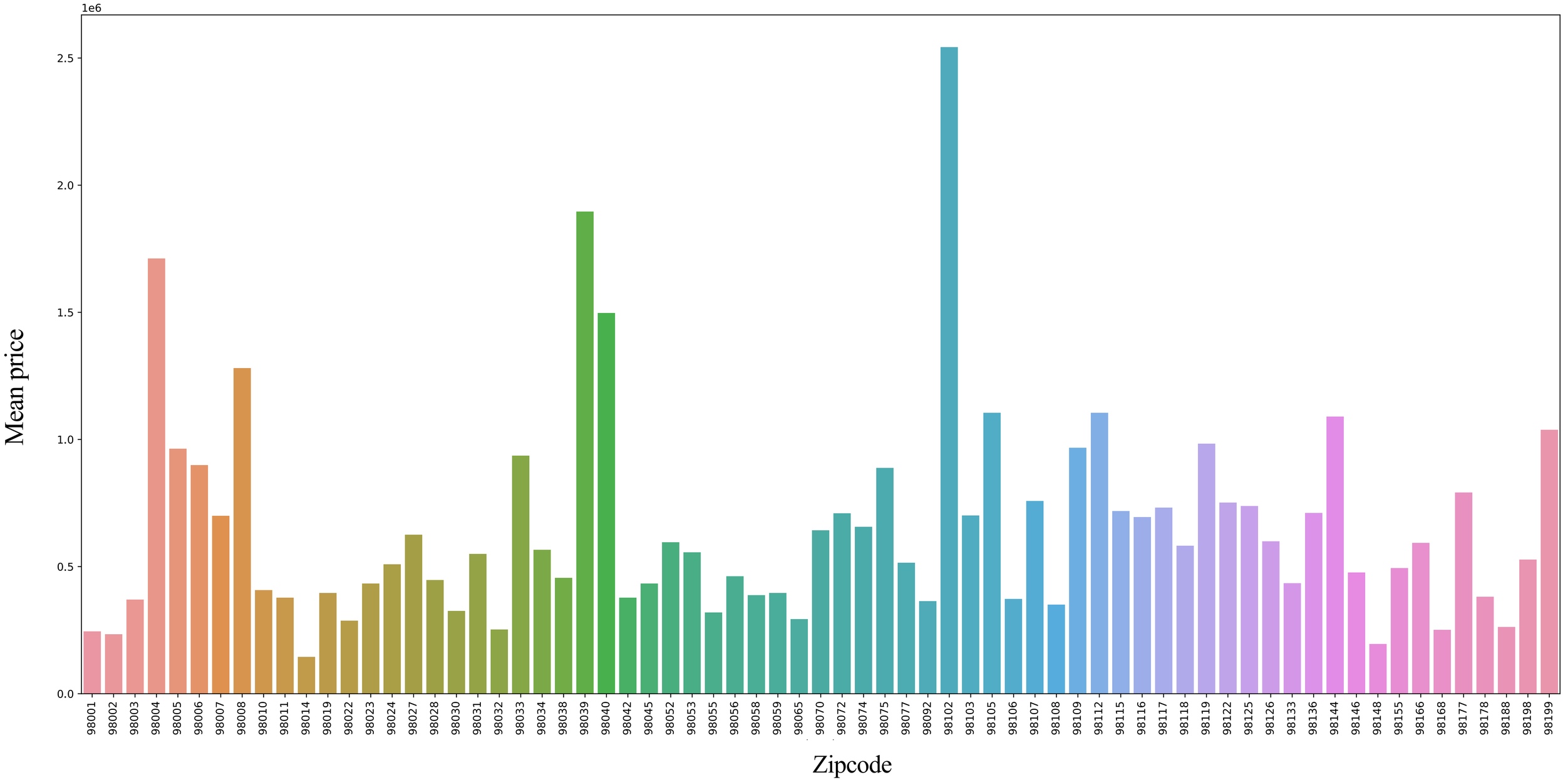

Distribution of numeric columns
Distribution of bedrooms
Distribution of bathrooms
Distribution of waterfront
Distribution of grade
Distribution of renovation year
Distribution of house age

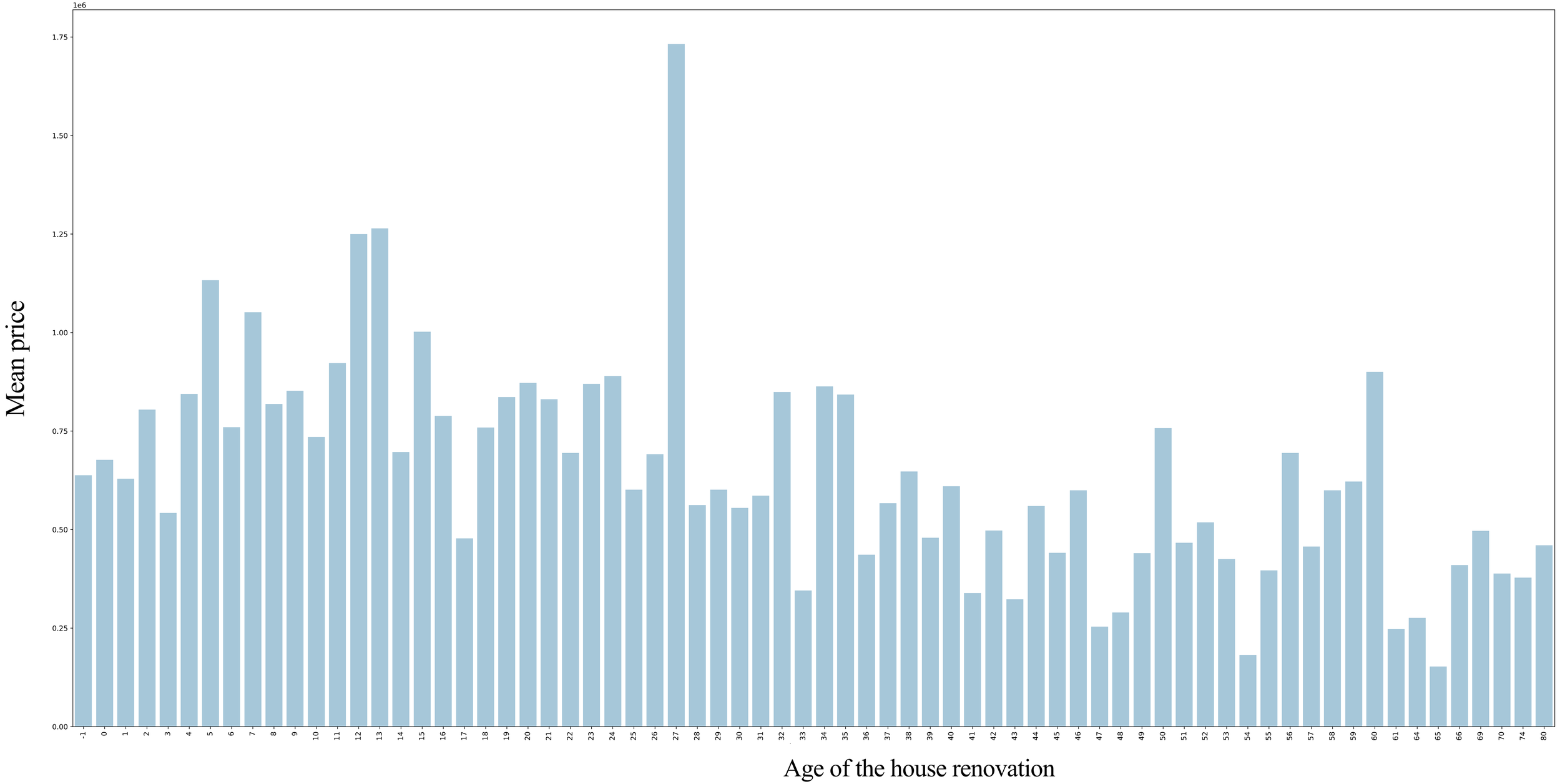
Sales count versus view



Sales count versus conditions




1e6
2.5
2.0
1.5
1.0
0.5
0.0
Mean price
98001
98002
98003
98004
98005
98006
98007
98008
98010
98011
98014
98019
98022
98023
98024
98027
98028
98030
98031
98032
98033
98034
98038
98039
98040
98042
98045
98052
98053
98055
98056
98058
98059
98065
98070
98072
98074
98075
98077
98092
98102
98103
98105
98106
98107
98108
98109
98112
98115
98116
98117
98118
98119
98122
98125
98126
98133
98136
98144
98146
98148
98155
98166
98168
98177
98178
98188
98198
98199
Zipcode


1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Mean price
-1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 64 65 66 69 70 74 80
Age of the house renovation

Insights from whole data sets

Summary

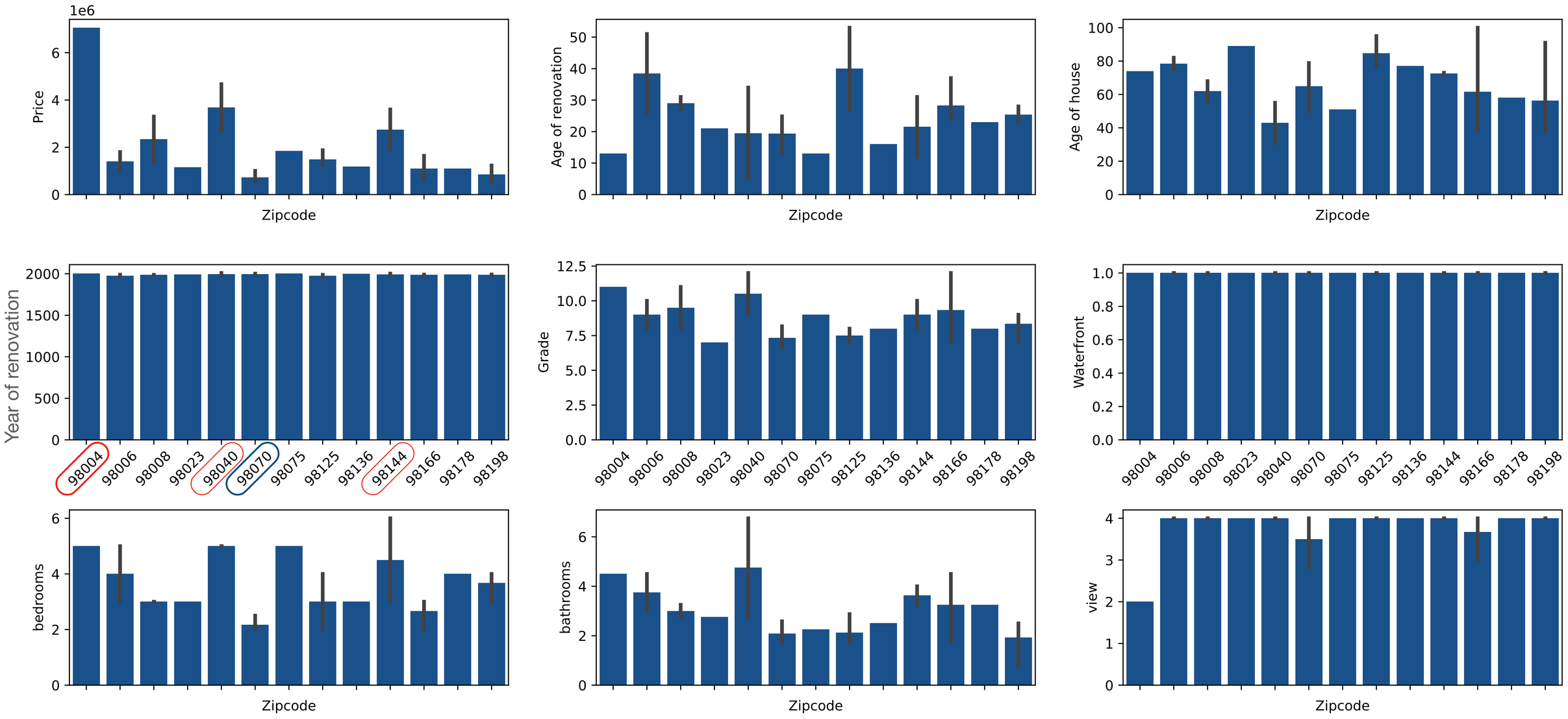
- Zipcode is an identifiers of houses
- Zipcodes 98039, 98004, 98040 - highest price . Zipcodes 98002, 98168, 98032 - lowest price
- Majority of houses sold in grade between 6 and 9
- Most of the house sold has 0 view. Buyers do not care about view
- Majority of houses doesn't have a waterfront
- Buyer prioritize 3 to 4 bedrooms and 1 to 2.5 bathrooms
- Living size is highly correlated with space of nearest 15 neighbors
- House with a waterfront has more price
- Buyer prioritize house with 2 to 3 floors. House with 2.5 floors has maximum price
- Most of the houses are in average condition
- Grade of a house is an important factor to decide the price of house
- Houses with less than one year of age has the highest median price
- The price of house is highly correlated with the living space

Final data preparation

Pseudo code:

```
df_nump = []
for ind in df_n0.index:
    if df_n0['waterfront'][ind]==1:
        df_nump.append(
         [
          ind,
          df_n0['zipcode'][ind],
          df_n0['waterfront'][ind],
          df_n0['grade'][ind],
          df_n0['yr_renovated'][ind],
          df_n0['house_age'][ind],
          df_n0['house_age_yr'][ind],
          df_n0['price'][ind],
          df_n0['bedrooms'][ind],
          df_n0['bathrooms'][ind],
          df_n0['view'][ind]]
         )
df_nump = pd.DataFrame(df_nump)
```

Recommendations to Jennifer Montgomery: Houses details versus zipcode



Thank You