In [143…]
```
import pandas as pd
from pandas_profiling import ProfileReport
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb
```

In [144…]
```
df = pd.read_csv("./Data/raw_house_data.csv")
```

In [145…]
```
df.head()
```

Out[145…]

	MLS	sold_price	zipcode	longitude	latitude	lot_acres	taxes	year_built	bedrooms	bathrooms	sqrt_ft	garage
0	21530491	5300000.0	85637	-110.378200	31.356362	2154.00	5272.00	1941	13	10	10500	
1	21529082	4200000.0	85646	-111.045371	31.594213	1707.00	10422.36	1997	2	2	7300	
2	3054672	4200000.0	85646	-111.040707	31.594844	1707.00	10482.00	1997	2	3	None	Non
3	21919321	4500000.0	85646	-111.035925	31.645878	636.67	8418.58	1930	7	5	9019	
4	21306357	3411450.0	85750	-110.813768	32.285162	3.21	15393.00	1995	4	6	6396	

In [146…]
```
df.shape
```

Out[146…]
```
(5000, 16)
```

In [147…]
```
df.describe()
```

Out[147…]

	MLS	sold_price	zipcode	longitude	latitude	lot_acres	taxes	year_built	bedroom
count	5.000000e+03	5.000000e+03	5000.000000	5000.000000	5000.000000	4990.000000	5.000000e+03	5000.00000	5000.0000
mean	2.127070e+07	7.746262e+05	85723.025600	-110.912107	32.308512	4.661317	9.402828e+03	1992.32800	3.9338
std	2.398508e+06	3.185556e+05	38.061712	0.120629	0.178028	51.685230	1.729385e+05	65.48614	1.2453
min	3.042851e+06	1.690000e+05	85118.000000	-112.520168	31.356362	0.000000	0.000000e+00	0.00000	1.0000
25%	2.140718e+07	5.850000e+05	85718.000000	-110.979260	32.277484	0.580000	4.803605e+03	1987.00000	3.0000
50%	2.161469e+07	6.750000e+05	85737.000000	-110.923420	32.318517	0.990000	6.223760e+03	1999.00000	4.0000
75%	2.180480e+07	8.350000e+05	85749.000000	-110.859078	32.394334	1.757500	8.082830e+03	2006.00000	4.0000
max	2.192856e+07	5.300000e+06	86323.000000	-109.454637	34.927884	2154.000000	1.221508e+07	2019.00000	36.0000

In [148…]
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   MLS               5000 non-null   int64  
 1   sold_price        5000 non-null   float64
 2   zipcode           5000 non-null   int64  
 3   longitude         5000 non-null   float64
 4   latitude          5000 non-null   float64
 5   lot_acres         4990 non-null   float64
 6   taxes             5000 non-null   float64
 7   year_built        5000 non-null   int64  
 8   bedrooms          5000 non-null   int64  
 9   bathrooms         5000 non-null   object 
 10  sqrt_ft           5000 non-null   object 
 11  garage            5000 non-null   object 
 12  kitchen_features  5000 non-null   object 
```

```
 13  fireplaces        5000 non-null   object
 14  floor_covering    5000 non-null   object
 15  HOA               5000 non-null   object
dtypes: float64(5), int64(4), object(7)
memory usage: 625.1+ KB
```

In [149…
```
df["kitchen_features"] = df["kitchen_features"].astype("category")
```

In [150…
```
df["floor_covering"] = df["floor_covering"].astype("category")
```

kitchen_features and floor_covering could be encoded with TF-IDF, word2vec, or fasttext

Imputation

In [151…
```
len(df[df["lot_acres"]==0]) / len(df)
```

Out[151… 0.007

In [152…
```
len(df[df["taxes"]==0]) / len(df)
```

Out[152… 0.0044

In [153…
```
len(df[df["year_built"]==0]) / len(df)
```

Out[153… 0.001

In [154…
```
for i in df.columns:
    print(i)
    print(df[df[i]=="None"].count()/len(df))
```

```
MLS
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
sold_price
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
```

```
zipcode
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
longitude
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
latitude
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
lot_acres
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
taxes
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
```

```
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
year_built
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
bedrooms
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
bathrooms
MLS                 0.0012
sold_price          0.0012
zipcode             0.0012
longitude           0.0012
latitude            0.0012
lot_acres           0.0012
taxes               0.0012
year_built          0.0012
bedrooms            0.0012
bathrooms           0.0012
sqrt_ft             0.0012
garage              0.0012
kitchen_features    0.0012
fireplaces          0.0012
floor_covering      0.0012
HOA                 0.0012
dtype: float64
sqrt_ft
MLS                 0.0112
sold_price          0.0112
zipcode             0.0112
longitude           0.0112
latitude            0.0112
lot_acres           0.0092
taxes               0.0112
year_built          0.0112
bedrooms            0.0112
bathrooms           0.0112
sqrt_ft             0.0112
garage              0.0112
kitchen_features    0.0112
fireplaces          0.0112
floor_covering      0.0112
HOA                 0.0112
dtype: float64
garage
MLS                 0.0014
sold_price          0.0014
zipcode             0.0014
```

```
longitude           0.0014
latitude            0.0014
lot_acres           0.0014
taxes               0.0014
year_built          0.0014
bedrooms            0.0014
bathrooms           0.0014
sqrt_ft             0.0014
garage              0.0014
kitchen_features    0.0014
fireplaces          0.0014
floor_covering      0.0014
HOA                 0.0014
dtype: float64
kitchen_features
MLS                 0.0066
sold_price          0.0066
zipcode             0.0066
longitude           0.0066
latitude            0.0066
lot_acres           0.0066
taxes               0.0066
year_built          0.0066
bedrooms            0.0066
bathrooms           0.0066
sqrt_ft             0.0066
garage              0.0066
kitchen_features    0.0066
fireplaces          0.0066
floor_covering      0.0066
HOA                 0.0066
dtype: float64
fireplaces
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
floor_covering
MLS                 0.0002
sold_price          0.0002
zipcode             0.0002
longitude           0.0002
latitude            0.0002
lot_acres           0.0002
taxes               0.0002
year_built          0.0002
bedrooms            0.0002
bathrooms           0.0002
sqrt_ft             0.0002
garage              0.0002
kitchen_features    0.0002
fireplaces          0.0002
floor_covering      0.0002
HOA                 0.0002
dtype: float64
HOA
MLS                 0.1124
sold_price          0.1124
zipcode             0.1124
longitude           0.1124
latitude            0.1124
lot_acres           0.1118
taxes               0.1124
year_built          0.1124
bedrooms            0.1124
bathrooms           0.1124
sqrt_ft             0.1124
garage              0.1124
kitchen_features    0.1124
fireplaces          0.1124
```

```
floor_covering      0.1124
HOA                 0.1124
dtype: float64
```

In [155…

```
for i in df.columns:
    print(i)
    print(df[df[i] == 0].count()/len(df))
#lot_acres, taxes, year_built
```

```
MLS
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
sold_price
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
zipcode
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
longitude
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
```

```
latitude
MLS                0.0
sold_price         0.0
zipcode            0.0
longitude          0.0
latitude           0.0
lot_acres          0.0
taxes              0.0
year_built         0.0
bedrooms           0.0
bathrooms          0.0
sqrt_ft            0.0
garage             0.0
kitchen_features    0.0
fireplaces         0.0
floor_covering     0.0
HOA                0.0
dtype: float64
lot_acres
MLS                0.007
sold_price         0.007
zipcode            0.007
longitude          0.007
latitude           0.007
lot_acres          0.007
taxes              0.007
year_built         0.007
bedrooms           0.007
bathrooms          0.007
sqrt_ft            0.007
garage             0.007
kitchen_features    0.007
fireplaces         0.007
floor_covering     0.007
HOA                0.007
dtype: float64
taxes
MLS                0.0044
sold_price         0.0044
zipcode            0.0044
longitude          0.0044
latitude           0.0044
lot_acres          0.0044
taxes              0.0044
year_built         0.0044
bedrooms           0.0044
bathrooms          0.0044
sqrt_ft            0.0044
garage             0.0044
kitchen_features    0.0044
fireplaces         0.0044
floor_covering     0.0044
HOA                0.0044
dtype: float64
year_built
MLS                0.001
sold_price         0.001
zipcode            0.001
longitude          0.001
latitude           0.001
lot_acres          0.001
taxes              0.001
year_built         0.001
bedrooms           0.001
bathrooms          0.001
sqrt_ft            0.001
garage             0.001
kitchen_features    0.001
fireplaces         0.001
floor_covering     0.001
HOA                0.001
dtype: float64
bedrooms
MLS                0.0
sold_price         0.0
zipcode            0.0
longitude          0.0
latitude           0.0
lot_acres          0.0
taxes              0.0
year_built         0.0
bedrooms           0.0
bathrooms          0.0
```

```
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
bathrooms
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
sqrt_ft
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
garage
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
kitchen_features
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
fireplaces
MLS                 0.0
sold_price          0.0
zipcode             0.0
```

```
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
floor_covering
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
HOA
MLS                 0.0
sold_price          0.0
zipcode             0.0
longitude           0.0
latitude            0.0
lot_acres           0.0
taxes               0.0
year_built          0.0
bedrooms            0.0
bathrooms           0.0
sqrt_ft             0.0
garage              0.0
kitchen_features    0.0
fireplaces          0.0
floor_covering      0.0
HOA                 0.0
dtype: float64
```

In [156…
```
df['HOA'].replace('None', 0, inplace=True)
```

In [157…
```
df['bathrooms'].replace('None', 0, inplace=True)
```

In [158…
```
df['garage'].replace('None', 0, inplace=True)
```

In [159…
```
df["sqrt_ft"].replace('None', 0, inplace=True)
```

In [160…
```
df["sqrt_ft"] = df["sqrt_ft"].astype('float')
```

In [161…
```
df["sqrt_ft"] = df.groupby(['bedrooms']).transform(lambda x: x.replace(0, x.mean()))['sqrt_ft']
```

In [162…
```
df["year_built"].value_counts()
```

Out[162…
```
2006    247
2007    236
2002    235
2005    230
2004    202
       ...
1907      1
1921      1
```

```
1910      1
1914      1
1927      1
Name: year_built, Length: 112, dtype: int64
```

In [163…
```
df["year_built"].replace(0, 2006, inplace=True)
```

In [164…
```
df["taxes"] = df.groupby(['bedrooms']).transform(lambda x: x.replace(0, x.mean()))['taxes']
```

In [165…
```
df["lot_acres"] = df.groupby(['bedrooms']).transform(lambda x: x.replace(0, x.mean()))['lot_acres']
```

In [166…
```
df["bathrooms"] = df["bathrooms"].astype('float')
```

In [167…
```
df["garage"] = df["garage"].astype('float')
```

In [168…
```
df["HOA"] = df['HOA'].str.replace(',', '', regex=False)
df["HOA"] = df["HOA"].astype('float')
```

In [169…
```
df["kitchen_features"].value_counts()
```

Out[169…
```
Dishwasher, Garbage Disposal, Refrigerator, Microwave, Oven
1719
Dishwasher, Garbage Disposal, Microwave, Oven
270
Compactor, Dishwasher, Garbage Disposal, Refrigerator, Microwave, Oven
189
Dishwasher, Garbage Disposal, Refrigerator, Oven
181
Dishwasher, Freezer, Garbage Disposal, Refrigerator, Microwave, Oven
127

...
Dishwasher, Double Sink, Garbage Disposal, Pantry: Walk-In, Refrigerator, Appliance Color: Stainless, Countertops
: Granite          1
Dishwasher, Double Sink, Garbage Disposal, Pantry: Walk-In, Refrigerator, Appliance Color: Stainless, Microwave:
SS                 1
Dishwasher, Double Sink, Garbage Disposal, Pantry: Walk-In, Refrigerator, Wet Bar, Appliance Color: Other
1
Dishwasher, Double Sink, Garbage Disposal, Refrigerator, Appliance Color: Stainless, Countertops: Granite
1
# of Ovens: 1, Dishwasher, Freezer, Garbage Disposal, Gas Range, Refrigerator, Appliance Color: Stainless, Counte
rtops: Slate       1
Name: kitchen_features, Length: 1872, dtype: int64
```

In [170…
```
df["floor_covering"].value_counts()
```

Out[170…
```
Carpet, Ceramic Tile                       1235
Carpet, Natural Stone                       579
Carpet, Ceramic Tile, Wood                  258
Ceramic Tile                                247
Concrete                                    242
                                           ...
Carpet, Other: brick                          1
Carpet, Other: Travertine Tile                1
Carpet, Other: Travertine & slate             1
Natural Stone, Other: Travertine & Slate      1
Wood, Other: porcelain tile                   1
Name: floor_covering, Length: 311, dtype: int64
```

Outlier Detection

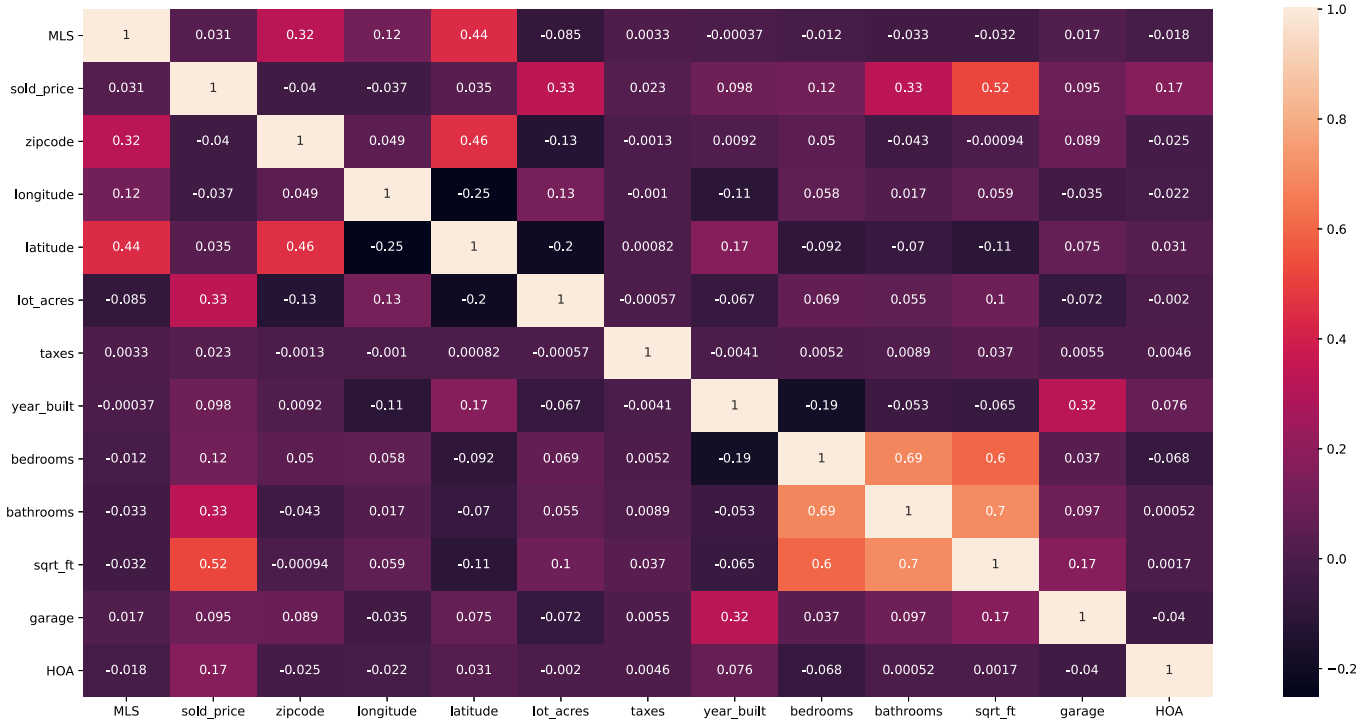
In [171…
```
def outliers(df):
    for k, v in df.items():
        q1 = v.quantile(0.25)
        q3 = v.quantile(0.75)
```

```
        iqr = q3 - q1
        v_col = v[(v <= q1 - 1.5 * iqr) | (v >= q3 + 1.5 * iqr)]
        percent = v_col.shape[0] * 100.0 / df.shape[0]
        print("Column %s outliers = %.2f%%" % (k, percent))
outliers(df[["sold_price", "lot_acres", "taxes", "sqrt_ft", "HOA"]])
```

```
Column sold_price outliers = 7.90%
Column lot_acres outliers = 10.46%
Column taxes outliers = 5.54%
Column sqrt_ft outliers = 4.76%
Column HOA outliers = 1.90%
```

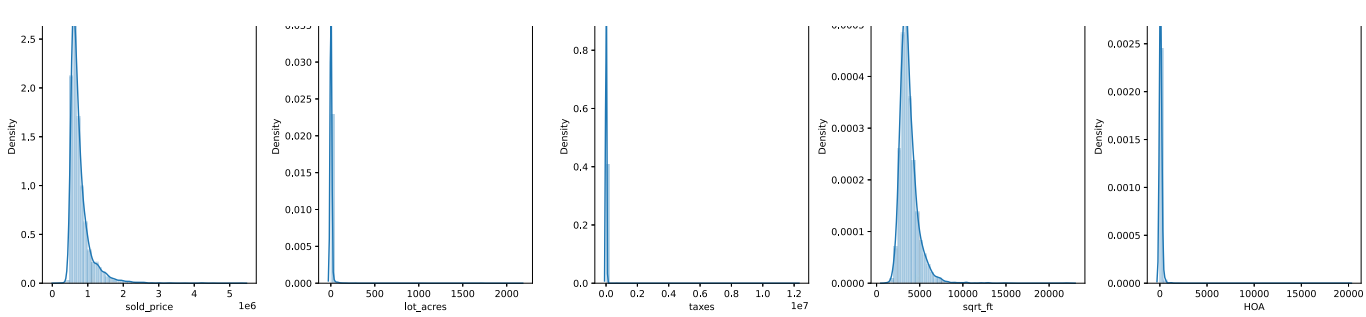
In [172…

```
plt.figure(figsize=(20, 10))
sb.heatmap(df.corr(),  annot=True);
```



In [173…

```
fig, axs = plt.subplots(ncols=5, nrows=1, figsize=(20, 5))
index = 0
axs = axs.flatten()
for k,v in df[["sold_price", "lot_acres", "taxes", "sqrt_ft", "HOA"]].items():
        sb.distplot(v, ax=axs[index])
        index += 1
plt.tight_layout(pad=0.4, w_pad=0.5, h_pad=5.0)
```



In [174…
```
df["sold_price"] = np.log1p(df["sold_price"])
```

In [175…
```
df["lot_acres"] = np.log1p(df["lot_acres"])
```

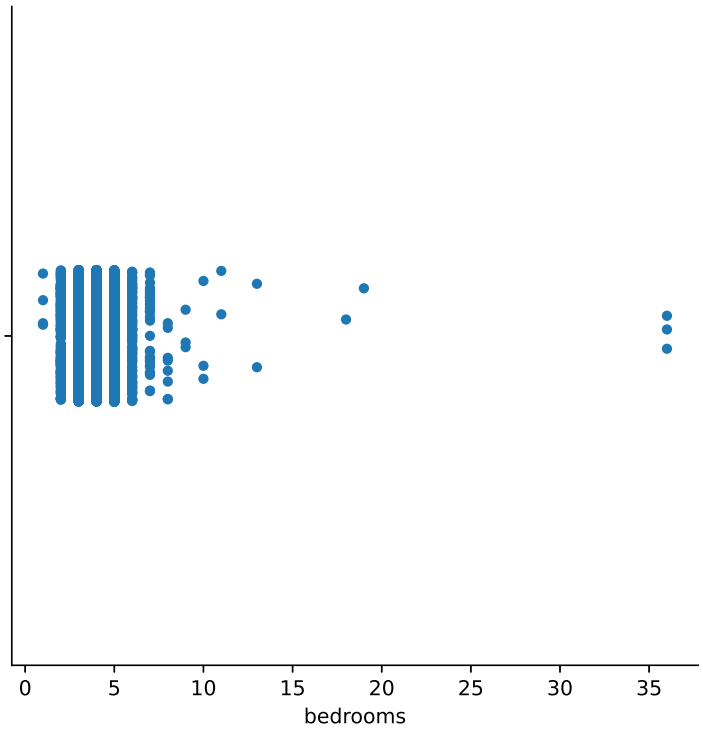
In [176…
```
df["taxes"] = np.log1p(df["taxes"])
```

In [177…
```
df["sqrt_ft"] = np.log1p(df["sqrt_ft"])
```

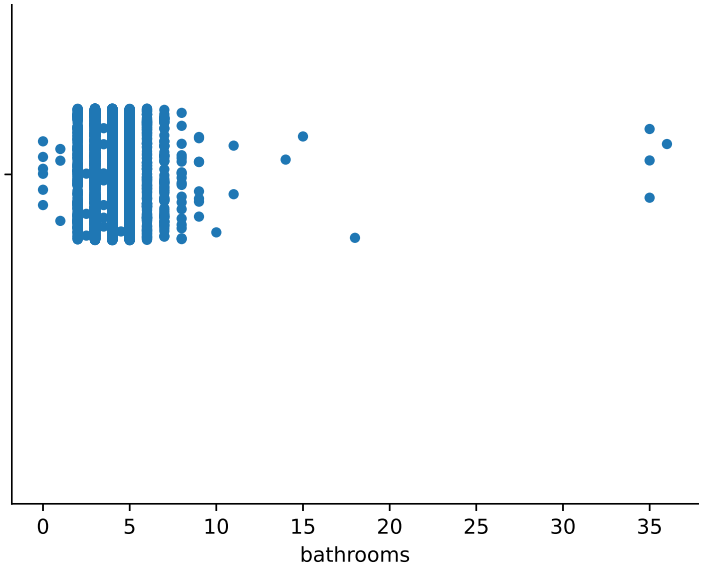
In [178…
```
df["HOA"] = np.log1p(df["HOA"])
```

In [179…
```
sb.catplot(x="bedrooms", data=df);["sold_price", "lot_acres", "taxes", "sqrt_ft", "HOA"]
```

Out[179…
```
['sold_price', 'lot_acres', 'taxes', 'sqrt_ft', 'HOA']
```



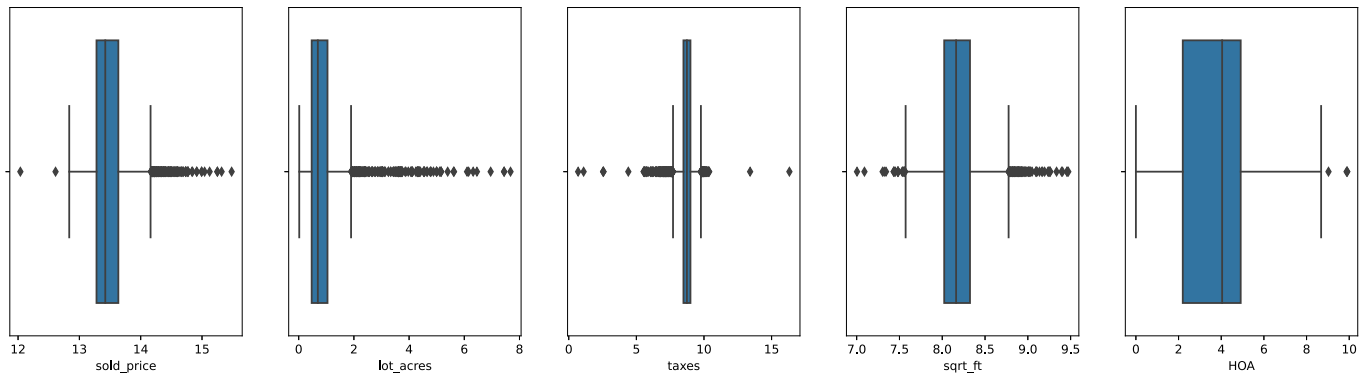
In [180…
```
sb.catplot(x="bathrooms", data=df);
```



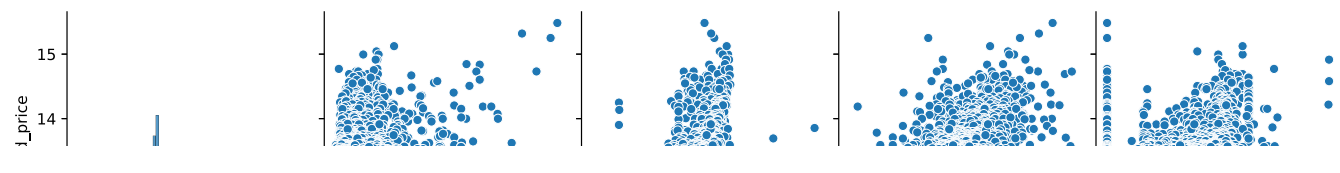
In [181…
```
df = df[df["bedrooms"] < 30]
```

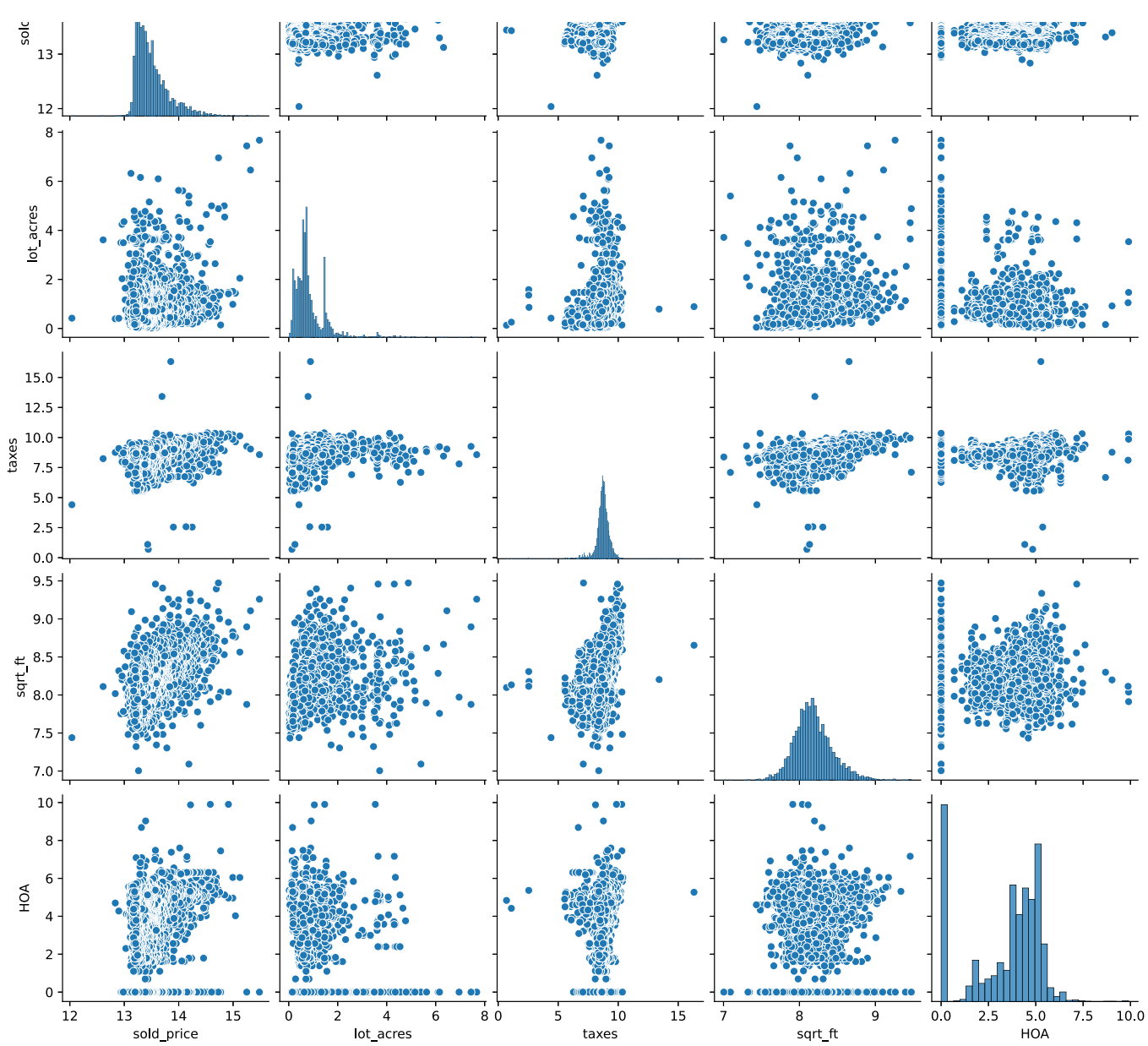
In [182…
```
df = df[df["bathrooms"] < 30]
```

In [183…
```
fig, axs = plt.subplots(ncols=5, nrows=1, figsize=(20, 5))
index = 0
axs = axs.flatten()
for k,v in df[["sold_price", "lot_acres", "taxes", "sqrt_ft", "HOA"]].items():
    sb.boxplot(x=k, data=df, ax=axs[index])
    plt.xlabel(k)
    index += 1
```



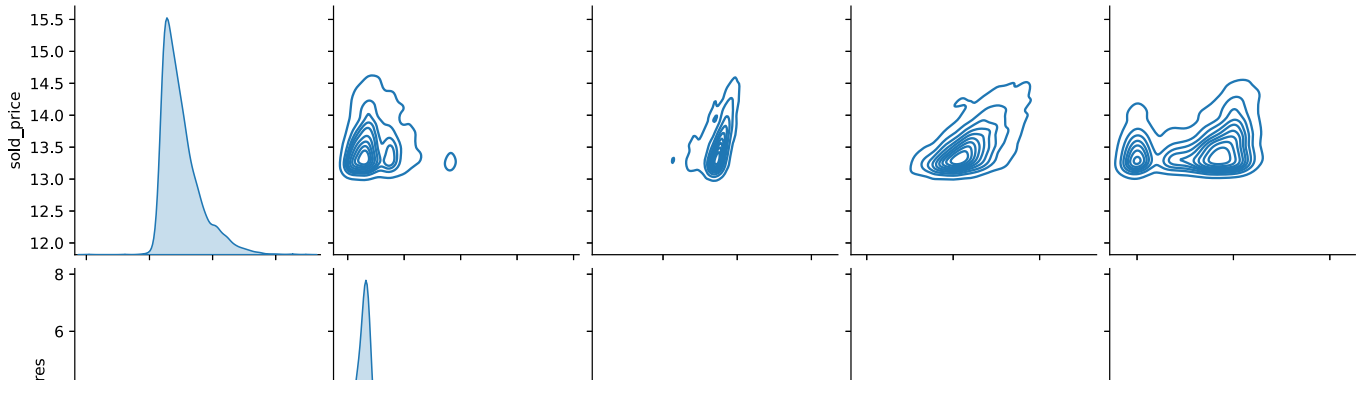
In [184…
```
sb.pairplot(data = df[["sold_price", "lot_acres", "taxes", "sqrt_ft", "HOA"]]);
```

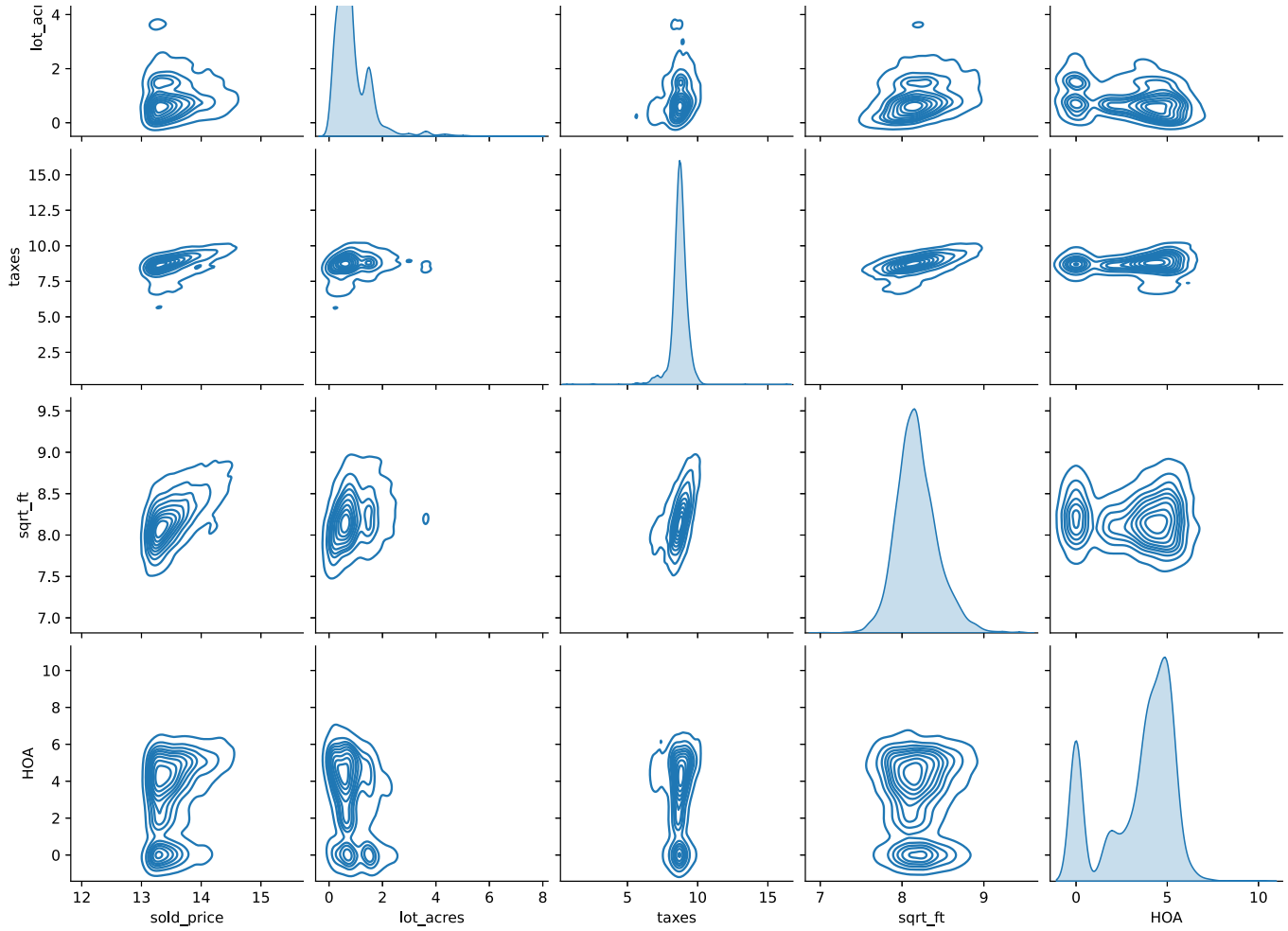




In [185…

```
sb.pairplot(data = df[["sold_price", "lot_acres", "taxes", "sqrt_ft", "HOA"]], kind="kde");
```





In [186…

```
profile = ProfileReport(df, title="Pandas Profiling Report")
profile
```

```
Summarize dataset: 100%|██████████| 30/30 [00:33<00:00,  1.11s/it, Completed]
Generate report structure: 100%|██████████| 1/1 [00:06<00:00,  6.77s/it]
Render HTML: 100%|██████████| 1/1 [00:05<00:00,  5.50s/it]
```

Overview

Dataset statistics

Number of variables	17
Number of observations	4996
Missing cells	569
Missing cells (%)	0.7%
Duplicate rows	0
Duplicate rows (%)	0.0%

Variable types

Numeric	14
Categorical	3

Total size in memory	712.2 KiB
Average record size in memory	146.0 B

Warnings

`kitchen_features` has a high cardinality: 1869 distinct values	**High cardinality**
`floor_covering` has a high cardinality: 311 distinct values	**High cardinality**
`HOA` has 559 (11.2%) missing values	**Missing**
`df_index` is uniformly distributed	**Uniform**
`df_index` has unique values	**Unique**
`MLS` has unique values	**Unique**

Out[186…

In [187…
```
# May want to drop MLS, as it is some kind of ID
```

In [188…
```
df[df["HOA"].isna() == True].count()
```

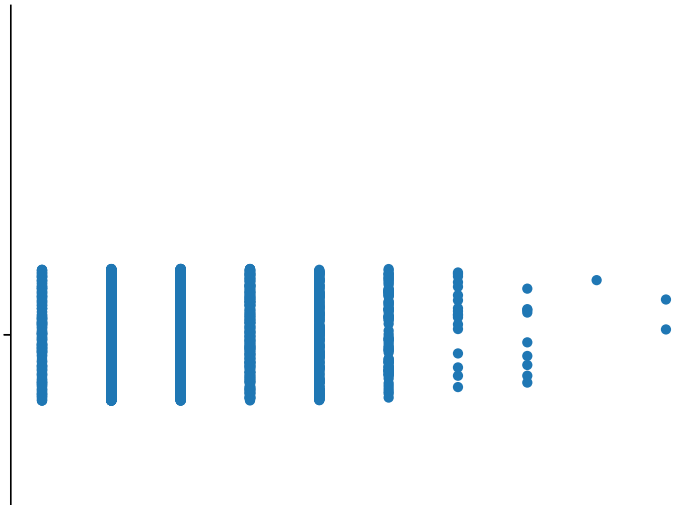
Out[188…
```
MLS                559
sold_price         559
zipcode            559
longitude          559
latitude           559
lot_acres          556
taxes              559
year_built         559
bedrooms           559
bathrooms          559
sqrt_ft            559
garage             559
kitchen_features   559
fireplaces         559
floor_covering     559
HOA                  0
dtype: int64
```

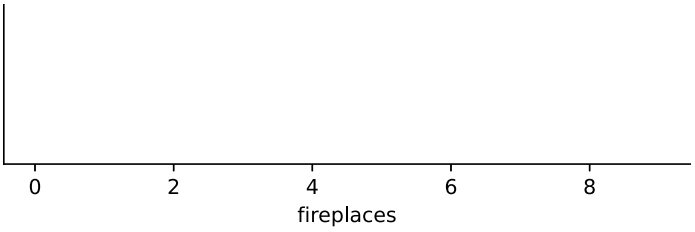
In [189…
```
df["HOA"] = df["HOA"].fillna(0)
```

In [190…
```
df["fireplaces"] = df["fireplaces"].replace(' ', 0)
```

In [191…
```
df["fireplaces"] = df["fireplaces"].astype("int")
```

In [192…
```
sb.catplot(x="fireplaces", data=df);
```





In [193…
```
df = df[df["fireplaces"] < 8]
```

In [194…
```
outliers(df[["sold_price", "lot_acres", "taxes", "sqrt_ft", "HOA"]])
```

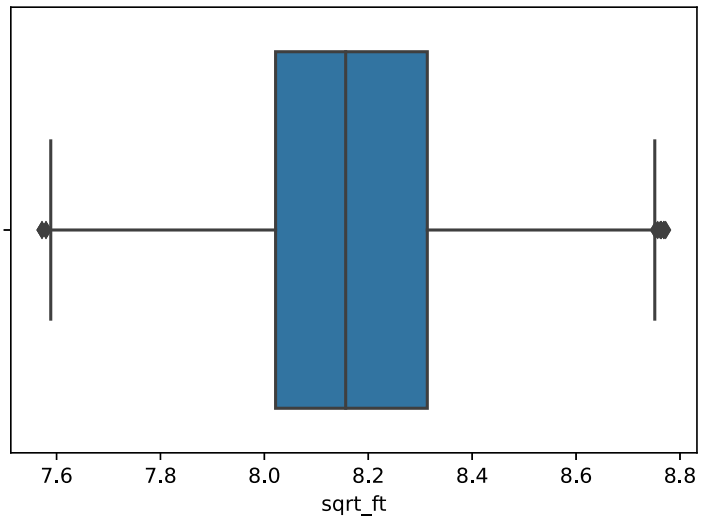
```
Column sold_price outliers = 4.47%
Column lot_acres outliers = 5.41%
Column taxes outliers = 6.91%
Column sqrt_ft outliers = 2.54%
Column HOA outliers = 0.00%
```

In [195…
```
q1 = df["sqrt_ft"].quantile(0.25)
q3 = df["sqrt_ft"].quantile(0.75)
iqr = q3 - q1
df = df[df["sqrt_ft"] >= q1 - 1.5 * iqr]
df = df[df["sqrt_ft"] <= q3 + 1.5 * iqr]
```

In [196…
```
outliers(df[["sold_price", "lot_acres", "taxes", "sqrt_ft", "HOA"]])
```

```
Column sold_price outliers = 4.03%
Column lot_acres outliers = 5.36%
Column taxes outliers = 6.41%
Column sqrt_ft outliers = 0.43%
Column HOA outliers = 0.00%
```

In [197…
```
sb.boxplot(x="sqrt_ft", data=df);
```



In [198…
```
len(df)/5000 * 100
```

Out[198… 97.32

In [ ]: