Exploratory Data Analysis (EDA) and Cleaning Pima County Housing Data Set

Exploratory Data Analysis (EDA)

First imported the raw housing data csv file into Python.

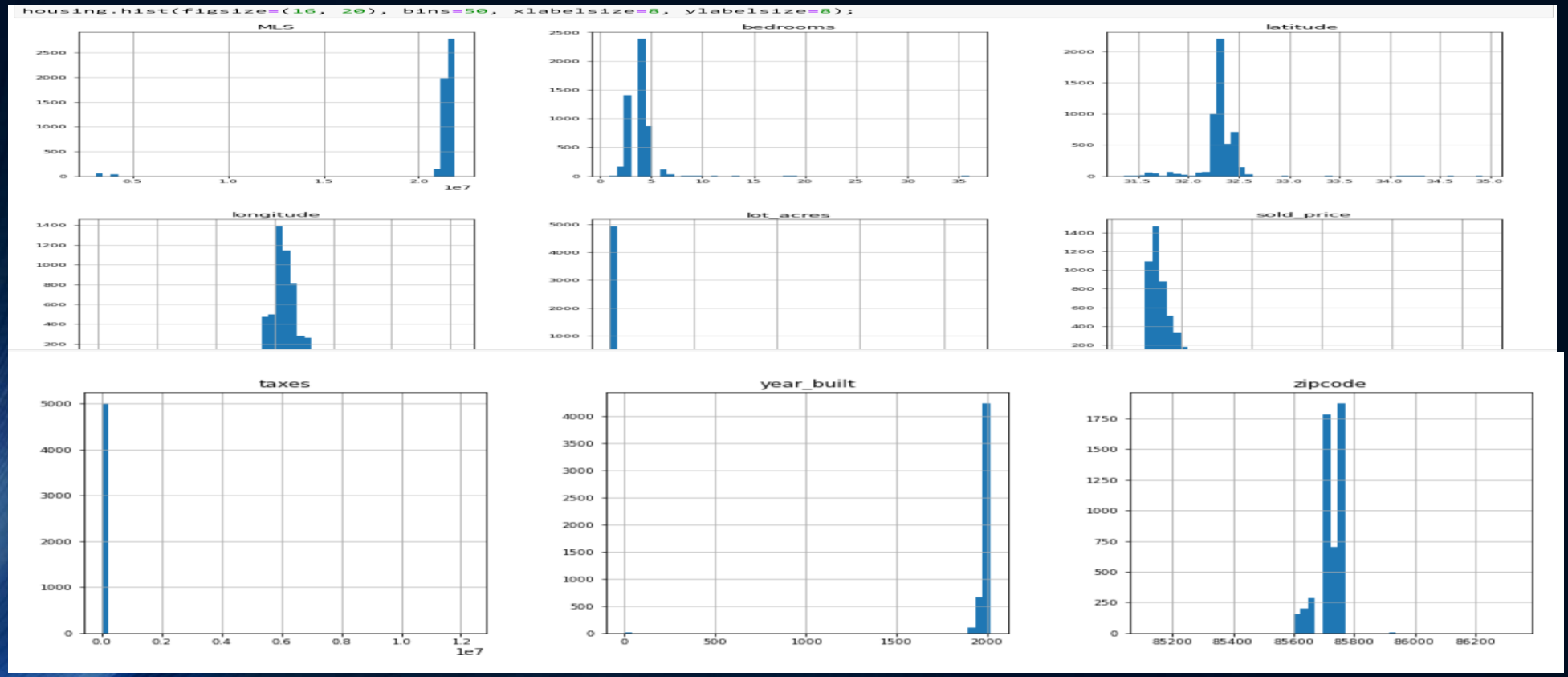
Then performed a head on the data set to view its attributes.

```
housing.head
```

```
<bound method NDFrame.head of              MLS  sold_price  zipcode   longitude   latitude  lot_acres  \
0       21530491    5300000.0    85637 -110.378200  31.356362    2154.00
1       21529082    4200000.0    85646 -111.045371  31.594213    1707.00
2        3054672    4200000.0    85646 -111.040707  31.594844    1707.00
3       21919321    4500000.0    85646 -111.035925  31.645878     636.67
4       21306357    3411450.0    85750 -110.813768  32.285162       3.21
...          ...          ...      ...         ...        ...        ...
4995    21810382     495000.0    85641 -110.661829  31.907917       4.98
4996    21908591     550000.0    85750 -110.858556  32.316373       1.42
4997    21832452     475000.0    85192 -110.755428  32.964708      12.06
4998    21900515     550000.0    85745 -111.055528  32.296871       1.01
4999     4111490     450000.0    85621 -110.913054  31.385259       4.16

          taxes  year_built  bedrooms bathrooms sqrt_ft garage  \
0       5272.00        1941        13        10   10500       0
1      10422.36        1997         2         2    7300       0
2      10482.00        1997         2         3    None    None
3       8418.58        1930         7         5    9019       4
4      15393.00        1995         4         6    6396       3
...         ...         ...       ...       ...     ...     ...
4995    2017.00        2005         5         3    3601       3
4996    4822.01        1990         4         3    2318       3
4997    1000.00        1969         3         2    1772       0
4998    5822.93        2009         4         4    3724       3
4999    2814.48        1988         4         4    4317    None

                                    kitchen_features fireplaces  \
0                  Dishwasher, Freezer, Refrigerator, Oven          6
1                            Dishwasher, Garbage Disposal           5
2                Dishwasher, Garbage Disposal, Refrigerator         5
3       Dishwasher, Double Sink, Pantry: Butler, Refri...           4
4       Dishwasher, Garbage Disposal, Refrigerator, Mi...           5
...                                                   ...         ...
4995    Dishwasher, Double Sink, Garbage Disposal, Gas...           1
4996    Dishwasher, Double Sink, Electric Range, Garba...           1
4997    Dishwasher, Electric Range, Island, Refrigerat...           0
4998    Dishwasher, Double Sink, Garbage Disposal, Gas...           1
4999    Compactor, Dishwasher, Double Sink, Island, Ap...           3

                     floor_covering   HOA
0                 Mexican Tile, Wood     0
1               Natural Stone, Other     0
2         Natural Stone, Other: Rock  None
```

Histogram of Data Set

Performed a histogram on the data set.



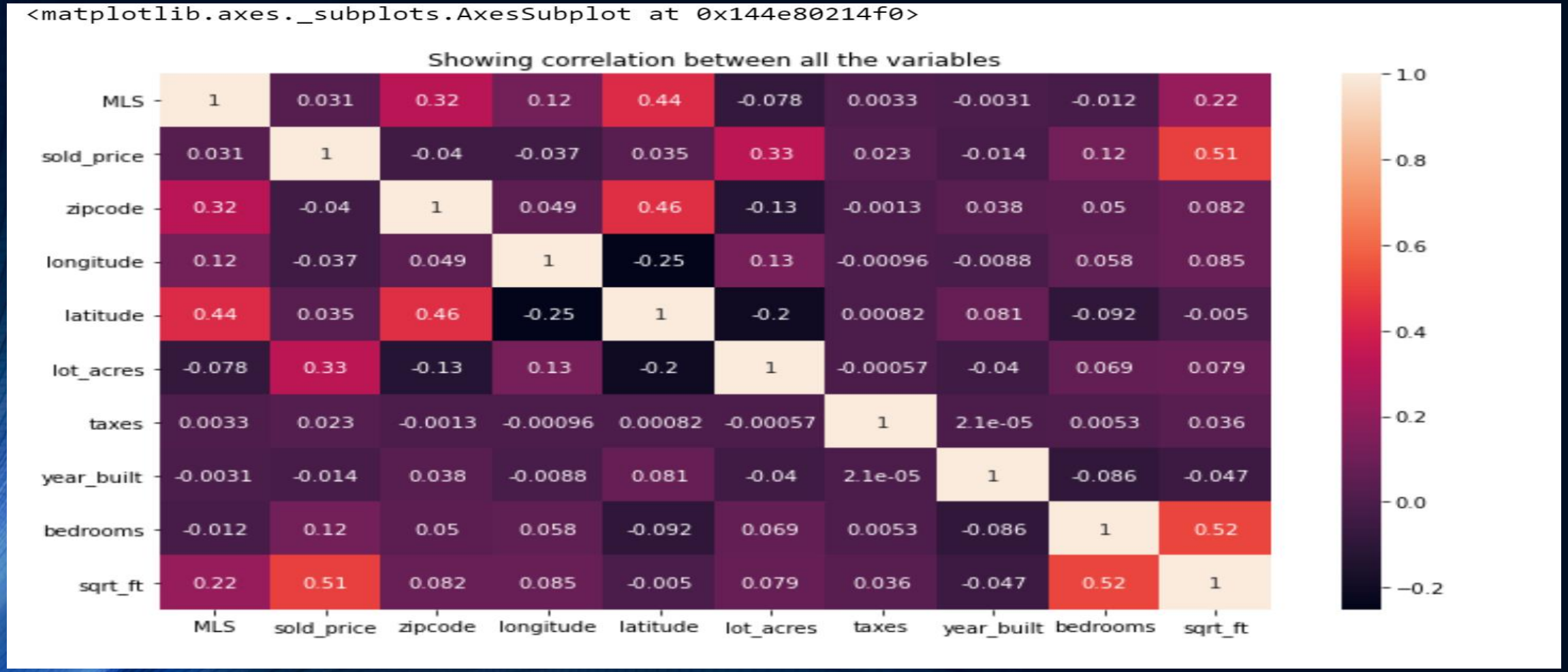
Columns with Missing Values

The lot_acres has 10 missing values.

```
housing.isnull().sum()
```

```
MLS                  0
sold_price           0
zipcode              0
longitude            0
latitude             0
lot_acres           10
taxes                0
year_built           0
bedrooms             0
bathrooms            0
sqrt_ft              0
garage               0
kitchen_features     0
fireplaces           0
floor_covering       0
HOA                  0
dtype: int64
```

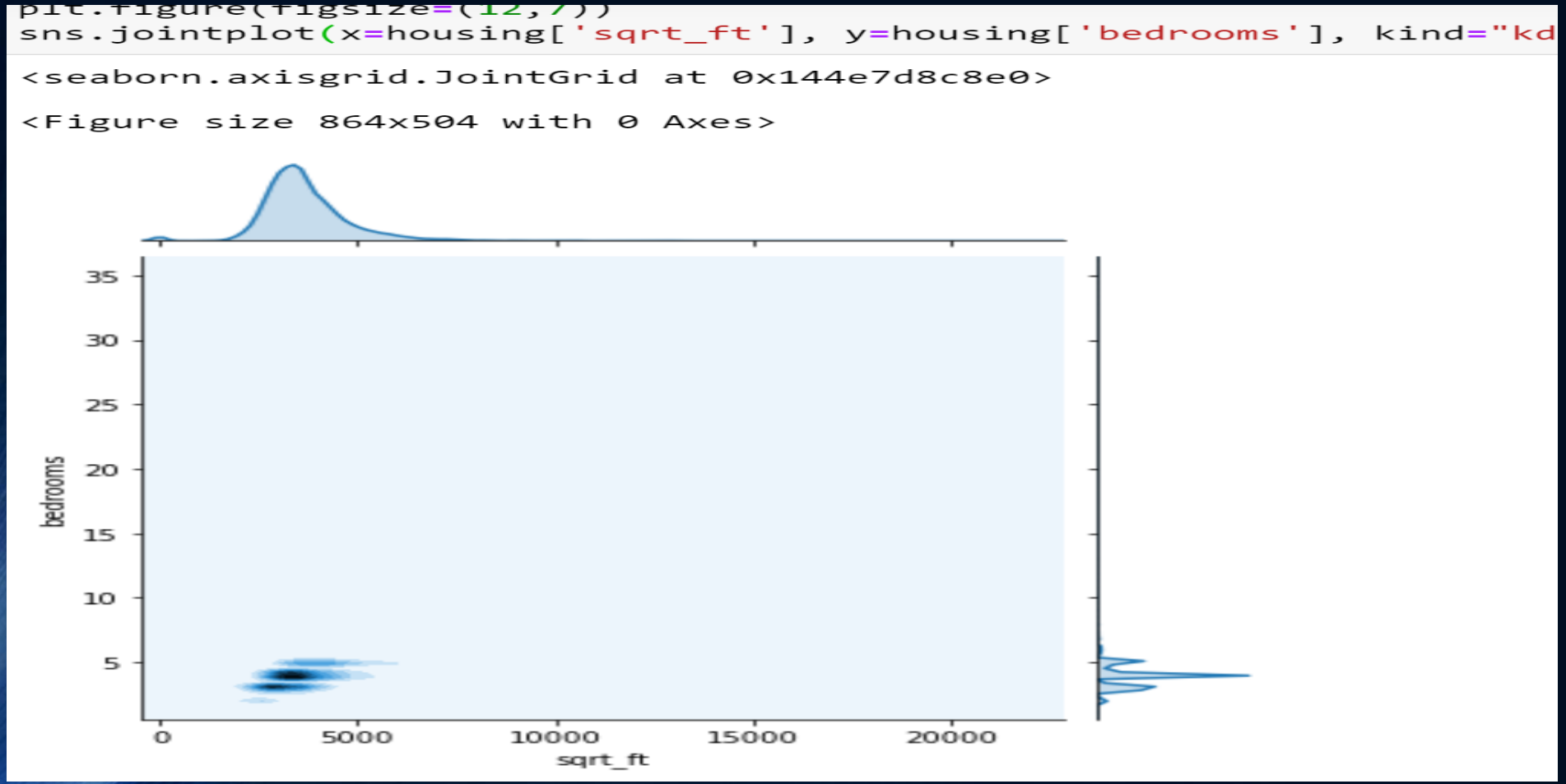
Heatmap of Missing Variables



Distribution Plot Comparing Year built Correlation with Bedrooms

```
plt.figure(figsize=(12,7))
sns.jointplot(x=housing['sqrt_ft'], y=housing['bedrooms'], kind="kd
```

```
<seaborn.axisgrid.JointGrid at 0x144e7d8c8e0>

<Figure size 864x504 with 0 Axes>
```



Dropping all Object Columns that have been One Hot Encoded

```
housing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Columns: 2633 entries, MLS to HOA_99.66
dtypes: float64(7), int64(3), object(6), uint8(2617)
memory usage: 13.1+ MB
```

Trransformed Data Set

Performed an info on it to view the structure of the file. The file has 5,000 entries.
Performed an description of the file to see its columns. Which it has 16 columns or attributes.

```
housing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Columns: 2633 entries, MLS to HOA_99.66
dtypes: float64(7), int64(3), object(6), uint8(2617)
memory usage: 13.1+ MB
```

```
housing.describe()
```

	MLS	sold_price	zipcode	longitude	latitude	lot_acres	taxes	year_built	bedrooms	sqrt_ft	...	HO
count	5.000000e+03	5.000000e+03	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03	5000.00000	5000.000000	5000.000000	...	5000.0
mean	2.127070e+07	7.746262e+05	85723.025600	-110.912107	32.308512	4.651994	9.402828e+03	1992.32800	3.933800	3674.743520	...	0.0
std	2.398508e+06	3.185556e+05	38.061712	0.120629	0.178028	51.633929	1.729385e+05	65.48614	1.245362	1181.036779	...	0.0
min	3.042851e+06	1.690000e+05	85118.000000	-112.520168	31.356362	0.000000	0.000000e+00	0.00000	1.000000	0.000000	...	0.0
25%	2.140718e+07	5.850000e+05	85718.000000	-110.979260	32.277484	0.580000	4.803605e+03	1987.00000	3.000000	3032.000000	...	0.0
50%	2.161469e+07	6.750000e+05	85737.000000	-110.923420	32.318517	0.990000	6.223760e+03	1999.00000	4.000000	3499.500000	...	0.0
75%	2.180480e+07	8.350000e+05	85749.000000	-110.859078	32.394334	1.750000	8.082830e+03	2006.00000	4.000000	4120.000000	...	0.0
max	2.192856e+07	5.300000e+06	86323.000000	-109.454637	34.927884	2154.000000	1.221508e+07	2019.00000	36.000000	22408.000000	...	1.0