Import Libraries

1
```
# Import all libraries used including the libraries that were used to determine the best model
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sn
import os
import glob
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score, explained_variance_score
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import StandardScaler
import openpyxl
import pickle
```

2
```
#set working directory to where data is stored
os.chdir(r"C:\Users\belbi\OneDrive\Documents\GCU\Capstone\Data\All Data")
```

3
```
#get filenames of all datafiles
csv_files = glob.glob('*.{}'.format('csv'))
csv_files
```

3
```
['redfin_2022-11-06-15-36-04.csv',
 'redfin_2022-11-06-15-54-57.csv',
 'redfin_2022-11-06-15-55-50.csv',
 'redfin_2022-11-06-15-56-31.csv',
 'redfin_2022-11-06-15-57-35.csv',
 'redfin_2022-11-06-15-58-16.csv',
 'redfin_2022-11-06-15-59-07.csv',
 'redfin_2022-11-06-15-59-57.csv',
 'redfin_2022-11-06-16-00-40.csv',
 'redfin_2022-11-06-16-01-03.csv',
 'redfin_2022-11-06-16-02-21.csv',
 'redfin_2022-11-06-16-04-07.csv',
 'redfin_2022-11-06-16-04-47.csv',
 'redfin_2022-11-06-16-07-10.csv',
 'redfin_2022-11-06-16-07-54.csv',
 'redfin_2022-11-06-16-08-30.csv',
 'redfin_2022-11-06-16-08-59.csv',
 'redfin_2022-11-06-16-09-39.csv',
 'redfin_2022-11-06-16-09-57.csv',
 'redfin_2022-11-06-16-10-12.csv',
 'redfin_2022-11-06-16-10-30.csv',
 'redfin_2022-11-06-16-11-00.csv',
 'redfin_2022-11-06-16-11-16.csv',
 'redfin_2022-11-06-16-11-32.csv',
 'redfin_2022-11-06-16-12-13.csv',
 'redfin_2022-11-06-16-12-24.csv',
 'redfin_2022-11-06-16-12-50.csv',
 'redfin_2022-11-06-16-13-03.csv',
 'redfin_2022-11-06-16-13-27.csv',
 'redfin_2022-11-06-16-13-41.csv',
 'redfin_2022-11-06-16-13-55.csv',
 'redfin_2022-11-06-16-14-05.csv',
 'redfin_2022-11-06-16-14-33.csv',
 'redfin_2022-11-06-16-15-04.csv',
 'redfin_2022-11-06-16-15-27.csv',
```

```
 'redfin_2022-11-06-16-15-47.csv',
 'redfin_2022-11-06-16-16-03.csv',
 'redfin_2022-11-06-16-16-52.csv',
 'redfin_2022-11-06-16-17-14.csv',
 'redfin_2022-11-06-16-17-28.csv',
 'redfin_2022-11-06-16-17-42.csv',
 'redfin_2022-11-09-17-52-28.csv',
 'redfin_2022-11-09-17-53-28.csv',
 'redfin_2022-11-09-17-54-08.csv',
 'redfin_2022-11-09-17-54-22.csv',
 'redfin_2022-11-09-17-54-53.csv',
 'redfin_2022-11-09-17-55-09.csv',
 'redfin_2022-11-09-17-56-15.csv',
 'redfin_2022-11-09-17-56-44.csv',
 'redfin_2022-11-09-17-56-56.csv',
 'redfin_2022-11-09-17-57-10.csv',
 'redfin_2022-11-09-17-58-27.csv',
 'redfin_2022-11-09-17-59-02.csv',
 'redfin_2022-11-09-17-59-25.csv',
 'redfin_2022-11-09-17-59-44.csv',
 'redfin_2022-11-09-17-59-59.csv',
 'redfin_2022-11-09-18-00-50.csv',
 'redfin_2022-11-09-18-02-05.csv',
 'redfin_2022-11-09-18-02-32.csv']
```

4

```
#build the dataframe by appending each file to the dataframe
df_append = pd.DataFrame()
for file in csv_files:
    df_temp = pd.read_csv(file)
    df_append = df_append.append(df_temp, ignore_index = True)
df_append
```

```
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
```

```
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
```

```
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1639621125.py:5: FutureWarning: The frame.append method
  df_append = df_append.append(df_temp, ignore_index = True)
```

4

	SALE TYPE	SOLD DATE	PROPERTY TYPE	ADDRESS	CITY	STATE OR PROVINCE	ZIP OR POSTAL CODE	PRICE
0	PAST SALE	NaN	Single Family Residential	1846 E Willetta St	Phoenix	AZ	85006.0	210000.0
1	PAST SALE	NaN	Single Family Residential	1515 E Brill St	Phoenix	AZ	85006.0	360000.0
2	PAST SALE	NaN	Single Family Residential	1632 E Granada Rd	Phoenix	AZ	85006.0	245000.0
3	PAST SALE	September-16-2022	Single Family Residential	1113 W PORTLAND St	Phoenix	AZ	85007.0	780000.0
4	PAST SALE	NaN	Single Family Residential	946 W Cocopah St	Phoenix	AZ	85007.0	258000.0
...	...	...	...	...	...	...	...	...
14081	PAST SALE	November-12-2021	Single Family Residential	3543 W STEINBECK Ct	Anthem	AZ	85086.0	460000.0
14082	PAST SALE	November-12-2021	Single Family Residential	34702 N 25TH Ln	Phoenix	AZ	85086.0	498000.0

	SALE TYPE	SOLD DATE	PROPERTY TYPE	ADDRESS	CITY	STATE OR PROVINCE	ZIP OR POSTAL CODE	PRICE
14083	PAST SALE	November-12-2021	Single Family Residential	3728 W MEDINAH Way	Anthem	AZ	85086.0	473000.0
14084	PAST SALE	November-10-2021	Single Family Residential	4642 W ROLLING ROCK Dr	Phoenix	AZ	85086.0	469000.0
14085	PAST SALE	November-9-2021	Single Family Residential	39829 N RIVER BEND Rd	Phoenix	AZ	85086.0	460000.0

14086 rows × 28 columns

5
```
df = df_append
```

6
```
#explore the first five rows of data
df.head()
```

6

	SALE TYPE	SOLD DATE	PROPERTY TYPE	ADDRESS	CITY	STATE OR PROVINCE	ZIP OR POSTAL CODE	PRICE	BE
0	PAST SALE	NaN	Single Family Residential	1846 E Willetta St	Phoenix	AZ	85006.0	210000.0	Na
1	PAST SALE	NaN	Single Family Residential	1515 E Brill St	Phoenix	AZ	85006.0	360000.0	Na
2	PAST SALE	NaN	Single Family Residential	1632 E Granada Rd	Phoenix	AZ	85006.0	245000.0	Na
3	PAST SALE	September-16-2022	Single Family Residential	1113 W PORTLAND St	Phoenix	AZ	85007.0	780000.0	4.0
4	PAST SALE	NaN	Single Family Residential	946 W Cocopah St	Phoenix	AZ	85007.0	258000.0	Na

5 rows × 28 columns

7
```
#view statistics on all columns in the data
df.describe(include='all')
```

7

	SALE TYPE	SOLD DATE	PROPERTY TYPE	ADDRESS	CITY	STATE OR PROVINCE	ZIP OR POSTAL CODE	
count	14086	11310	14086	14073	14073	14086	14073.000000	1.40700
unique	1	296	7	9841	8	1	NaN	NaN
top	PAST SALE	February-28-2022	Single Family Residential	3329 E Palm Ln	Phoenix	AZ	NaN	NaN
freq	14086	121	14029	9	13105	14086	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	85047.239324	3.80337
std	NaN	NaN	NaN	NaN	NaN	NaN	71.250392	7.88927
min	NaN	NaN	NaN	NaN	NaN	NaN	85003.000000	4.50000
25%	NaN	NaN	NaN	NaN	NaN	NaN	85019.000000	3.38000
50%	NaN	NaN	NaN	NaN	NaN	NaN	85032.000000	3.87000
75%	NaN	NaN	NaN	NaN	NaN	NaN	85042.000000	4.36550
max	NaN	NaN	NaN	NaN	NaN	NaN	85392.000000	1.82500

11 rows × 28 columns

8
```
#look for any null values contained in columns
df.isnull().sum()
```

8
```
SALE TYPE                        0
SOLD DATE                     2776
PROPERTY TYPE                    0
ADDRESS                         13
CITY                            13
STATE OR PROVINCE                0
ZIP OR POSTAL CODE              13
PRICE                           16
BEDS                          2714
BATHS                           72
LOCATION                      2822
SQUARE FEET                     30
LOT SIZE                        18
YEAR BUILT                      30
DAYS ON MARKET               14086
$/SQUARE FEET                   46
HOA/MONTH                    11184
STATUS                        2760
NEXT OPEN HOUSE START TIME   14086
```

```
NEXT OPEN HOUSE END TIME                                                                14086
URL (SEE https://www.redfin.com/buy-a-home/comparative-market-analysis FOR INFO ON PRICING)      0
SOURCE                                                                                  2760
MLS#                                                                                    2776
FAVORITE                                                                                   0
INTERESTED                                                                                 0
LATITUDE                                                                                   0
LONGITUDE                                                                                  0
POOL                                                                                       0
dtype: int64
```

9
```
#view the data types for each of the columns in the data
df.dtypes
```

9
```
SALE TYPE                                                                              object
SOLD DATE                                                                              object
PROPERTY TYPE                                                                          object
ADDRESS                                                                                object
CITY                                                                                   object
STATE OR PROVINCE                                                                      object
ZIP OR POSTAL CODE                                                                    float64
PRICE                                                                                 float64
BEDS                                                                                  float64
BATHS                                                                                 float64
LOCATION                                                                               object
SQUARE FEET                                                                           float64
LOT SIZE                                                                              float64
YEAR BUILT                                                                            float64
DAYS ON MARKET                                                                        float64
$/SQUARE FEET                                                                         float64
HOA/MONTH                                                                             float64
STATUS                                                                                 object
NEXT OPEN HOUSE START TIME                                                            float64
NEXT OPEN HOUSE END TIME                                                              float64
URL (SEE https://www.redfin.com/buy-a-home/comparative-market-analysis FOR INFO ON PRICING)    object
SOURCE                                                                                 object
MLS#                                                                                  float64
FAVORITE                                                                               object
INTERESTED                                                                             object
LATITUDE                                                                              float64
LONGITUDE                                                                             float64
POOL                                                                                   object
dtype: object
```

10
```
#convert the sold date to a date
date_string = df['SOLD DATE']
df['SOLDDATE'] = pd.to_datetime(date_string)
```

11
```
#removing data without MLS# as these most likely represent private sales
df = df.dropna(subset=['MLS#'],axis=0)
```

12
```
#after removing the data without MLS# determine how many columns still contain null values
df.isnull().sum()
```

12
```
SALE TYPE                                                                                  0
SOLD DATE                                                                                  0
PROPERTY TYPE                                                                              0
ADDRESS                                                                                    0
CITY                                                                                       0
STATE OR PROVINCE                                                                          0
ZIP OR POSTAL CODE                                                                         0
```

```
PRICE                                                                                  0
BEDS                                                                                   0
BATHS                                                                                  5
LOCATION                                                                              46
SQUARE FEET                                                                            0
LOT SIZE                                                                               4
YEAR BUILT                                                                             0
DAYS ON MARKET                                                                     11310
$/SQUARE FEET                                                                          0
HOA/MONTH                                                                           8408
STATUS                                                                                 0
NEXT OPEN HOUSE START TIME                                                         11310
NEXT OPEN HOUSE END TIME                                                           11310
URL (SEE https://www.redfin.com/buy-a-home/comparative-market-analysis FOR INFO ON PRICING)    0
SOURCE                                                                                 0
MLS#                                                                                   0
FAVORITE                                                                               0
INTERESTED                                                                             0
LATITUDE                                                                               0
LONGITUDE                                                                              0
POOL                                                                                   0
SOLDDATE                                                                               0
dtype: int64
```

13
```
#remove the rows of data where the bed, baths or lotsize data is missing
df = df.dropna(subset=['BEDS'],axis=0)
df = df.dropna(subset=['BATHS'],axis=0)
df = df.dropna(subset=['LOT SIZE'],axis=0)
```

14
```
#build an HOA data object based on whether there is an amount in the HOA/month column
df['HOA'] = np.where(df['HOA/MONTH'] > 0 ,1, 0)
```

15
```
#convert the Pool Y/N data to 1/0 values based on Y =1 and all else is 0
df['POOL'] = np.where(df['POOL'] == 'Y', 1,0)
```

16
```
#view the dataframe after making the changes above
df
```

16

	SALE TYPE	SOLD DATE	PROPERTY TYPE	ADDRESS	CITY	STATE OR PROVINCE	ZIP OR POSTAL CODE	PRICE
3	PAST SALE	September-16-2022	Single Family Residential	1113 W PORTLAND St	Phoenix	AZ	85007.0	780000.0
5	PAST SALE	September-9-2022	Single Family Residential	1425 E PORTLAND St	Phoenix	AZ	85006.0	385000.0
6	PAST SALE	August-8-2022	Single Family Residential	1501 W Willetta St	Phoenix	AZ	85007.0	649999.0
8	PAST SALE	August-29-2022	Single Family Residential	1515 W YAVAPAI St W	Phoenix	AZ	85007.0	350000.0

	SALE TYPE	SOLD DATE	PROPERTY TYPE	ADDRESS	CITY	STATE OR PROVINCE	ZIP OR POSTAL CODE	PRICE
9	PAST SALE	September-30-2022	Single Family Residential	810 S 3RD Ave	Phoenix	AZ	85003.0	614000.0
...	...	...	...	...	...	...	...	...
14081	PAST SALE	November-12-2021	Single Family Residential	3543 W STEINBECK Ct	Anthem	AZ	85086.0	460000.0
14082	PAST SALE	November-12-2021	Single Family Residential	34702 N 25TH Ln	Phoenix	AZ	85086.0	498000.0
14083	PAST SALE	November-12-2021	Single Family Residential	3728 W MEDINAH Way	Anthem	AZ	85086.0	473000.0
14084	PAST SALE	November-10-2021	Single Family Residential	4642 W ROLLING ROCK Dr	Phoenix	AZ	85086.0	469000.0
14085	PAST SALE	November-9-2021	Single Family Residential	39829 N RIVER BEND Rd	Phoenix	AZ	85086.0	460000.0

11301 rows × 30 columns

17
```
# select the columns to be used in the model as well as the app for analysis
#rename long column names and remove spaces in column names
data = df[['SOLDDATE','ADDRESS','CITY','STATE OR PROVINCE','ZIP OR POSTAL CODE','PRICE','BEDS','BATHS','S
data = data.rename(columns = {'ZIP OR POSTAL CODE':'ZIPCODE'})
data = data.rename(columns = {'SQUARE FEET':'SQFT'})
data = data.rename(columns = {'LOT SIZE':'LOTSIZE'})
data = data.rename(columns = {'YEAR BUILT':'YEARBUILT'})
```

18
```
#limit the dataset to sold price of less than $500k
data = data[data.PRICE <500000]
```

19
```
#Remove zipcodes captured that are not Phoenix
data = data[data.ZIPCODE < 85099]
```

20
```
#Remove Duplicates and view shape of the data
data.duplicated().sum()
data.drop_duplicates
data.shape
```

20
```
(10476, 16)
```

21
```
#Verify that there are no missing values in the new dataset
data.isnull().sum()
```

21
```
SOLDDATE             0
ADDRESS              0
CITY                 0
STATE OR PROVINCE    0
ZIPCODE              0
PRICE                0
BEDS                 0
BATHS                0
SQFT                 0
LOTSIZE              0
YEARBUILT            0
MLS#                 0
POOL                 0
HOA                  0
LATITUDE             0
LONGITUDE            0
dtype: int64
```

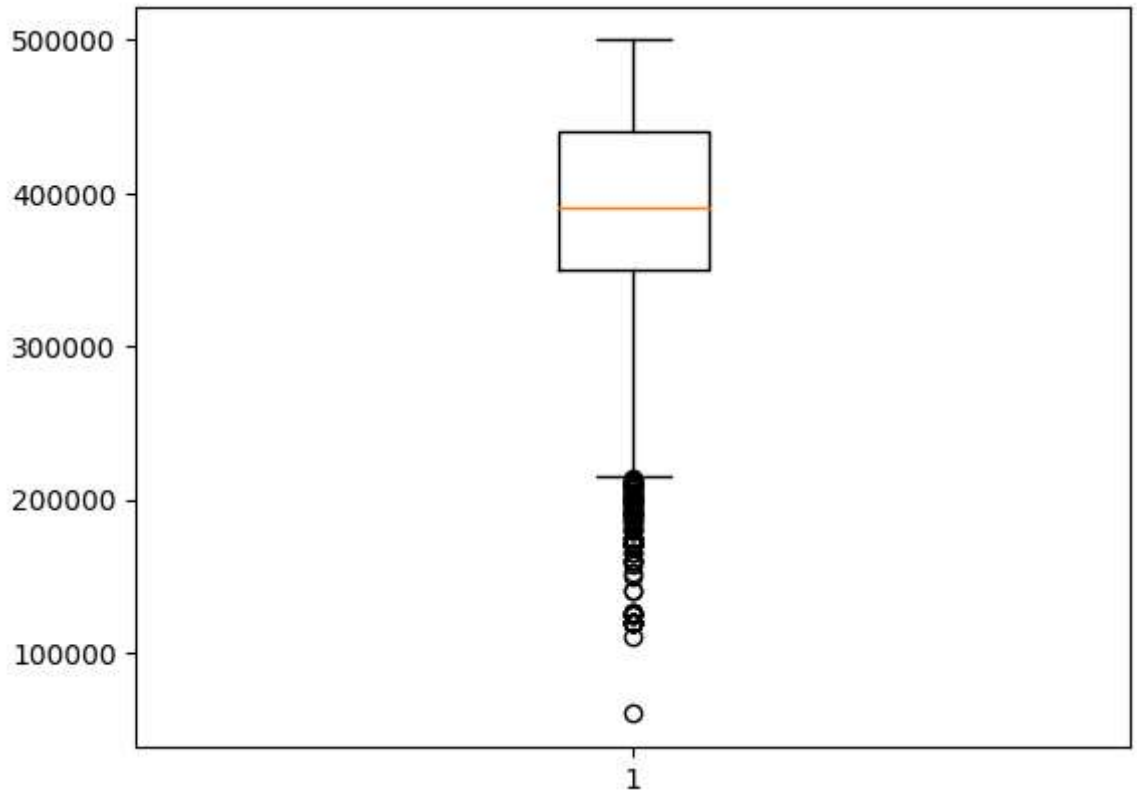
22
```
#covert zipcode to category data for one hot encoding
data['ZIPCODE']=data['ZIPCODE'].astype('int64')
data['YEARBUILT']=data['YEARBUILT'].astype('int64')
data['SQFT']=data['SQFT'].astype('int64')
data['LOTSIZE']=data['LOTSIZE'].astype('int64')
data['BEDS']=data['BEDS'].astype('int32')
```

Exploratory Data Analysis

Explore Price

23
```
#original view of price to determine outliers that need to be removed

plt.boxplot(data['PRICE'])
```

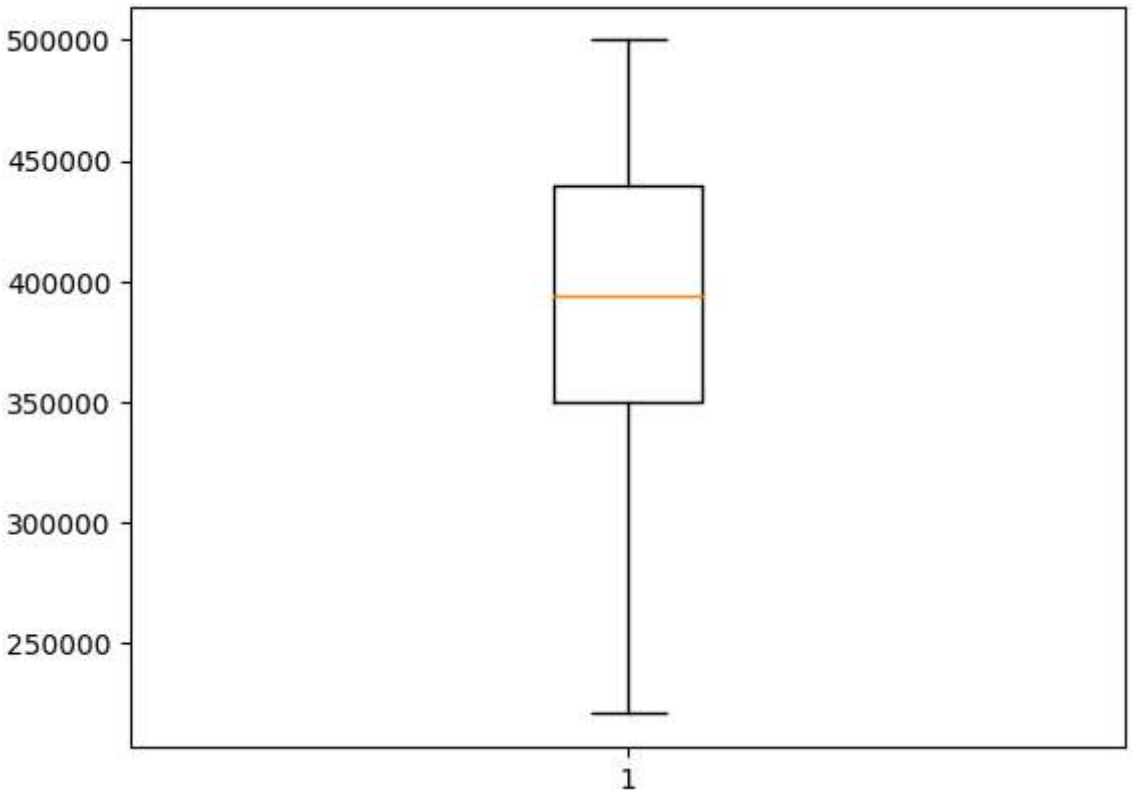
23
```
{'whiskers': [<matplotlib.lines.Line2D at 0x22fd9dc5d00>,
  <matplotlib.lines.Line2D at 0x22fd9dc5fa0>],
 'caps': [<matplotlib.lines.Line2D at 0x22fd9de3280>,
  <matplotlib.lines.Line2D at 0x22fd9de3520>],
 'boxes': [<matplotlib.lines.Line2D at 0x22fd9dc5a60>],
 'medians': [<matplotlib.lines.Line2D at 0x22fd9de37c0>],
 'fliers': [<matplotlib.lines.Line2D at 0x22fd9de3a60>],
 'means': []}
```



24
```
#remove outliers less than 220,000
data = data[data.PRICE>220000]
```

25
```
#view after removing the outliers
plt.boxplot(data['PRICE'])
```

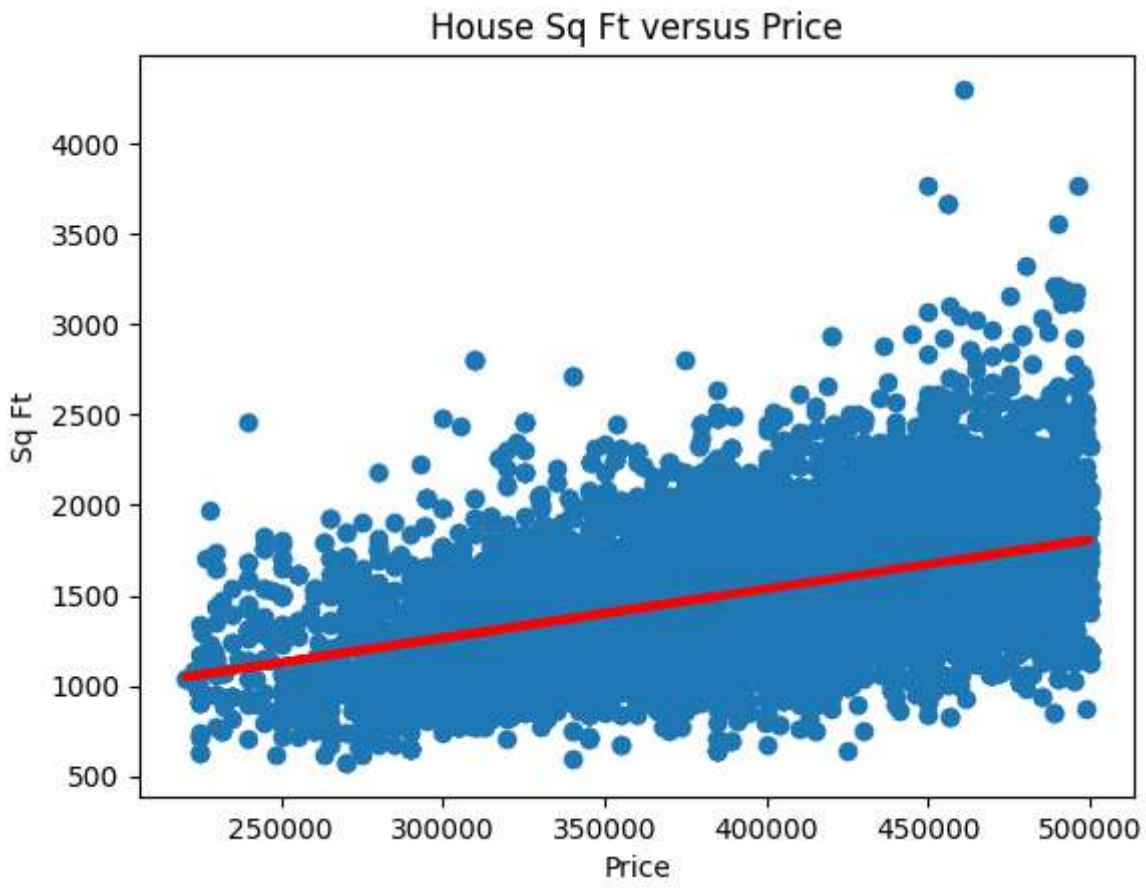
25
```
{'whiskers': [<matplotlib.lines.Line2D at 0x22fda210f40>,
  <matplotlib.lines.Line2D at 0x22fda21f220>],
 'caps': [<matplotlib.lines.Line2D at 0x22fda21f4c0>,
  <matplotlib.lines.Line2D at 0x22fda21f760>],
 'boxes': [<matplotlib.lines.Line2D at 0x22fda210ca0>],
 'medians': [<matplotlib.lines.Line2D at 0x22fda21fa00>],
 'fliers': [<matplotlib.lines.Line2D at 0x22fda21fca0>],
 'means': []}
```



Compare house size to price

```
26  x=data['PRICE']
    y=data['SQFT']
    plt.scatter(x,y)
    plt.title('House Sq Ft versus Price')
    plt.xlabel('Price')
    plt.ylabel('Sq Ft')

    z=np.polyfit(x,y,1)
    p=np.poly1d(z)
    plt.plot(x,p(x), color ='red', linewidth=3)
```

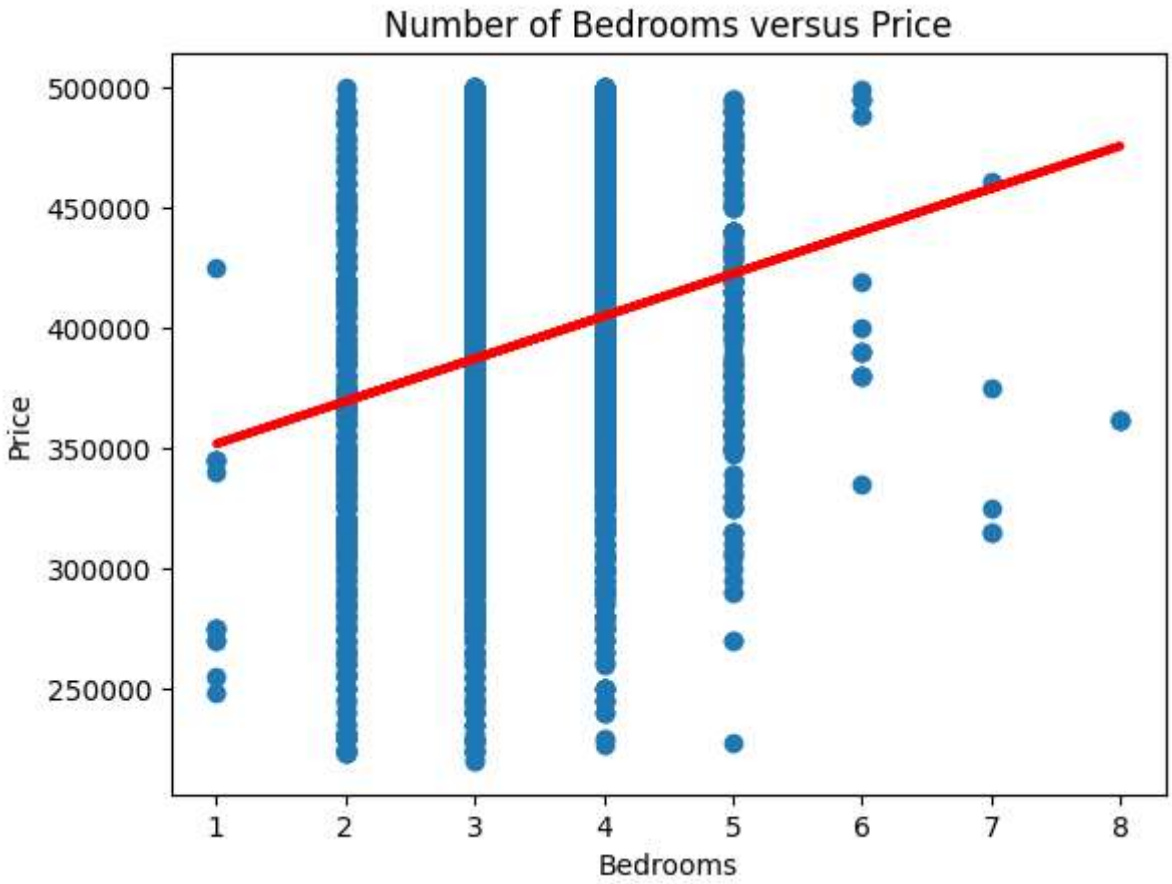
```
26  [<matplotlib.lines.Line2D at 0x22fdc554d60>]
```



Compare number of bedrooms to price

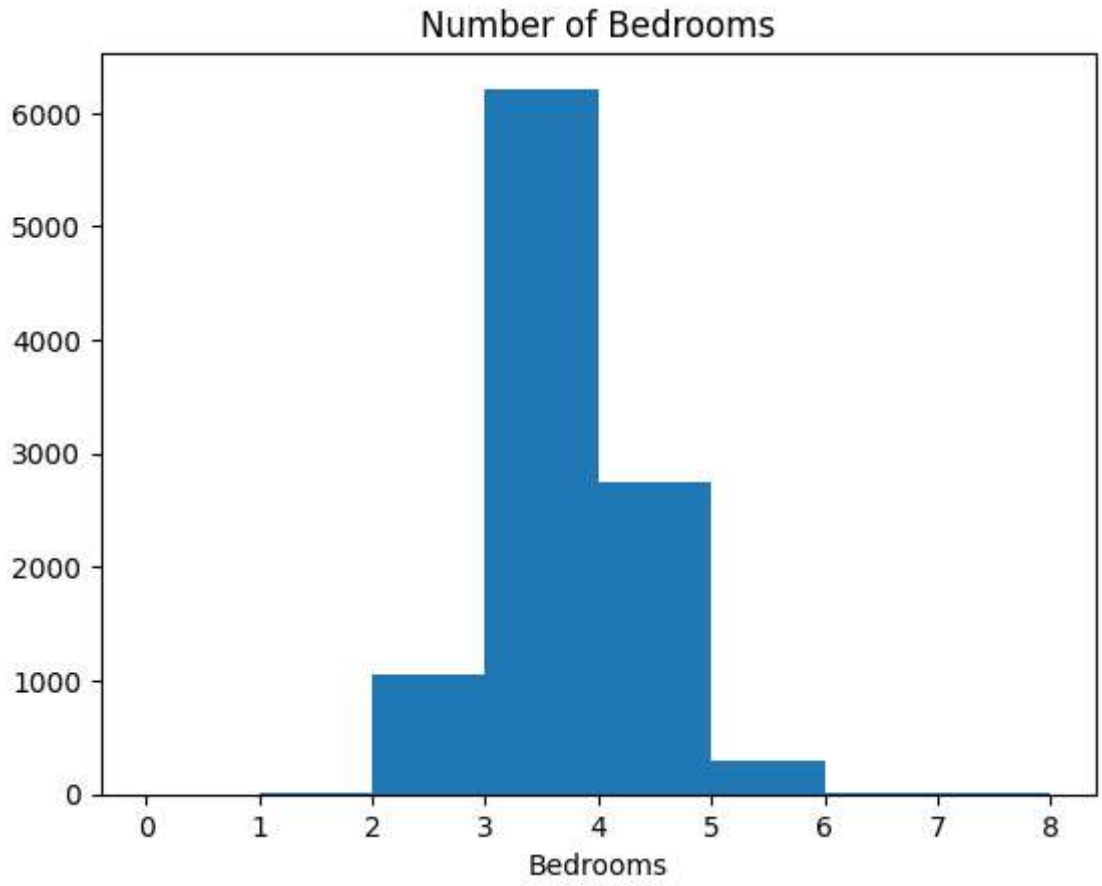
27
```
y=data['PRICE']
x=data['BEDS']
plt.scatter(x,y)
plt.title('Number of Bedrooms versus Price')
plt.xlabel('Bedrooms')
plt.ylabel('Price')

z=np.polyfit(x,y,1)
p=np.poly1d(z)
plt.plot(x,p(x), color ='red', linewidth=3)
```

27
```
[<matplotlib.lines.Line2D at 0x22fda469d60>]
```



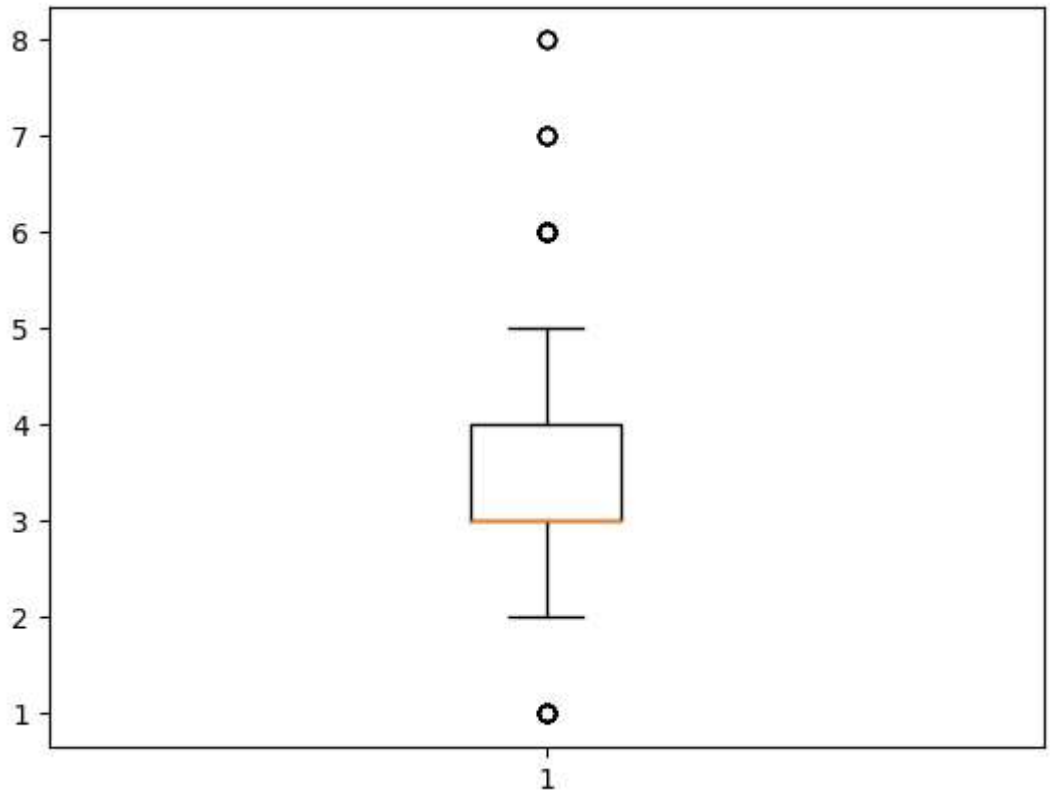
Histogram of number of Bedrooms

```
28  plt.hist(data['BEDS'], bins= [0,1,2,3,4,5,6,7,8])
    plt.title('Number of Bedrooms')
    plt.xlabel('Bedrooms')
    plt.show()
```



29
```
#View the distribution of the number of bedrooms to identify outliers
plt.boxplot(data['BEDS'])
```

29
```
{'whiskers': [<matplotlib.lines.Line2D at 0x22fdc582d00>,
  <matplotlib.lines.Line2D at 0x22fdc582fa0>],
 'caps': [<matplotlib.lines.Line2D at 0x22fdc591160>,
  <matplotlib.lines.Line2D at 0x22fdc591400>],
 'boxes': [<matplotlib.lines.Line2D at 0x22fdc582a60>],
 'medians': [<matplotlib.lines.Line2D at 0x22fdc5916a0>],
 'fliers': [<matplotlib.lines.Line2D at 0x22fdc591940>],
 'means': []}
```

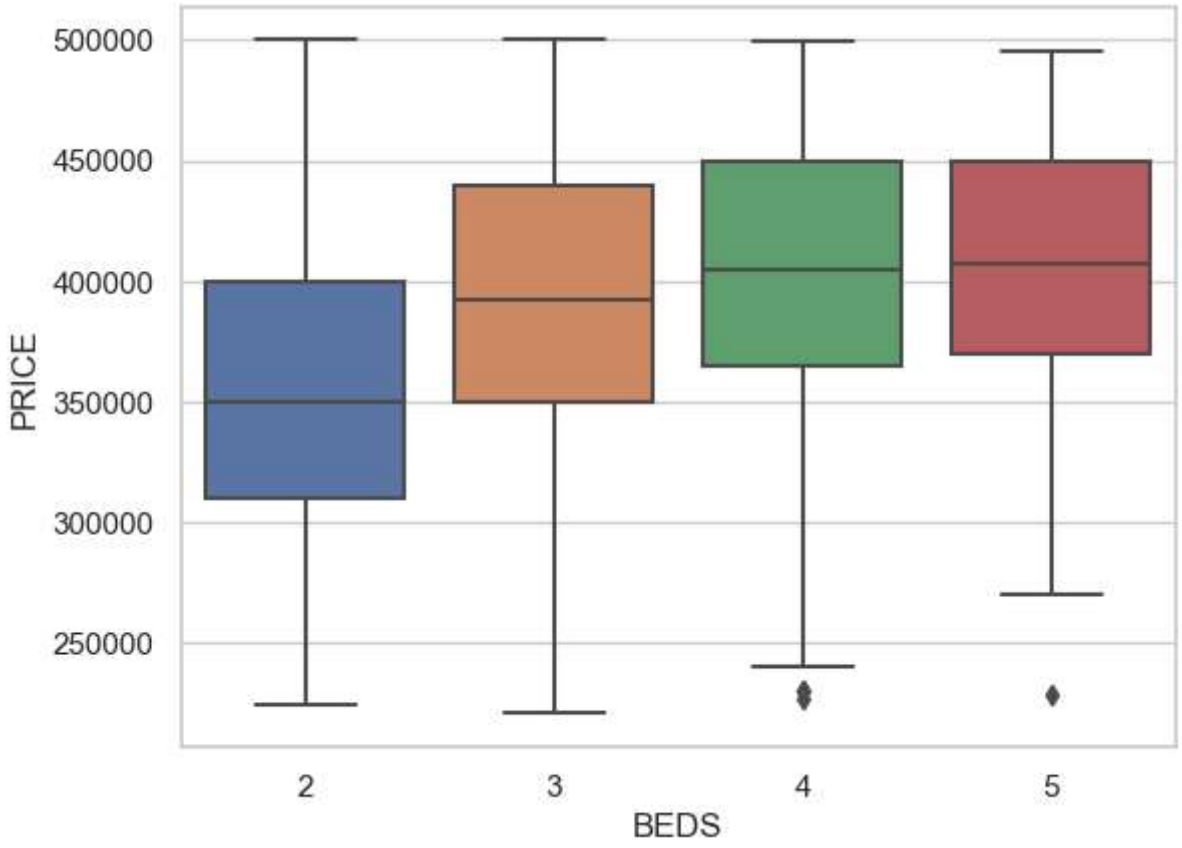



```
30  #remove the 1 bedroom and more than 5 bedrooms
    data = data[data.BEDS <6]
```

```
31  data = data[data.BEDS>1]
```

```
32  # view the beds by price to see relationship and identify outliers
    sn.set(style = 'whitegrid')
    sn.boxplot(x='BEDS',
               y='PRICE',
               data = data)
```

```
32  <AxesSubplot: xlabel='BEDS', ylabel='PRICE'>
```



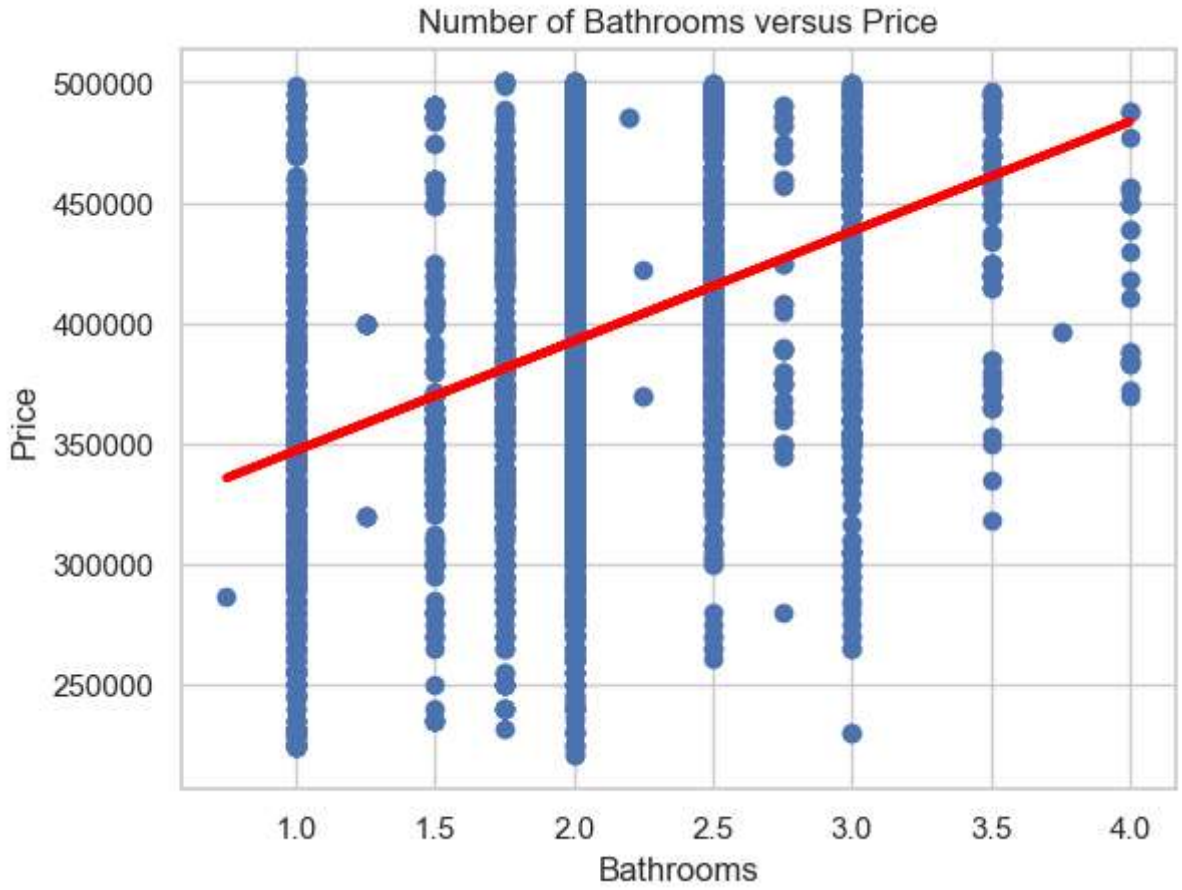
33
```
#Remove houses with 4+ bedrooms and sold for less than 250,000
dataremove = data[(data['BEDS']>=4) & (data['PRICE']< 250000)].index
data.drop(dataremove,inplace=True)
```

sn.set(style = 'whitegrid') sn.boxplot(x='BEDS', y='PRICE', data = data)

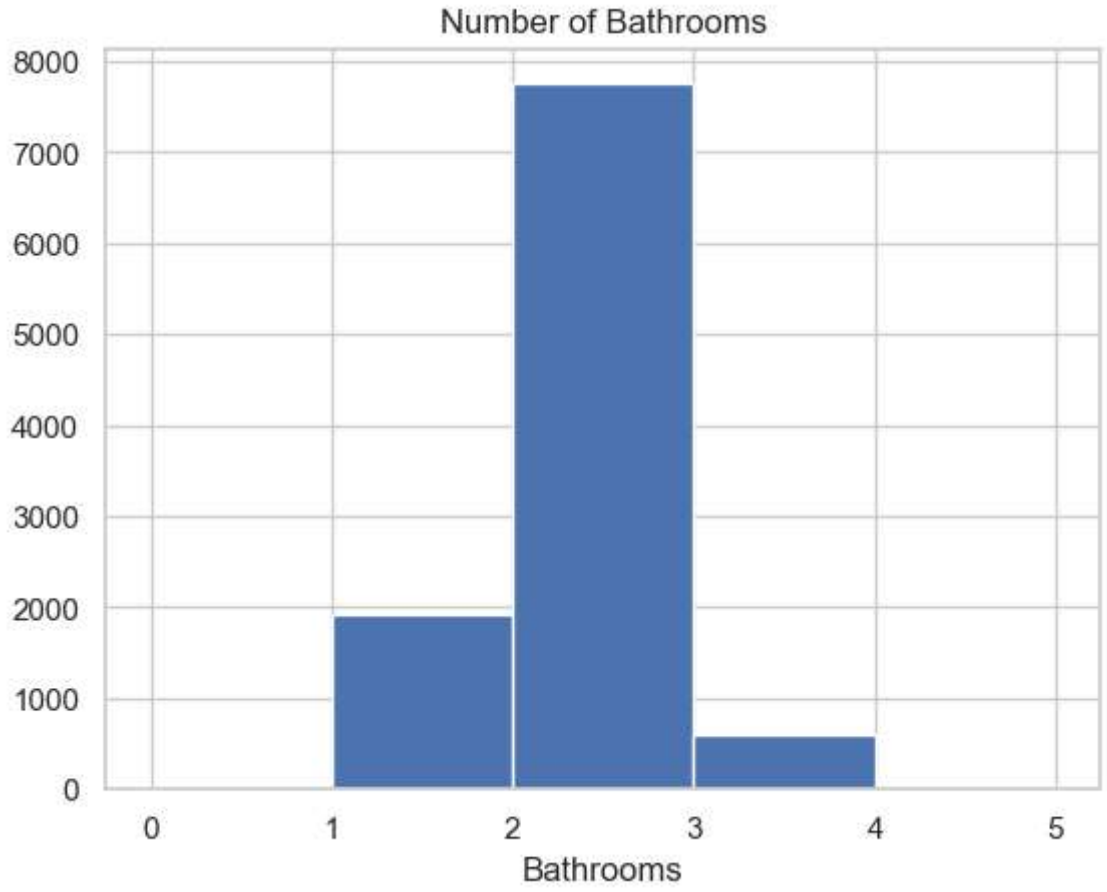
Compare number of bathrooms to price

34
```
y=data['PRICE']
x=data['BATHS']
plt.scatter(x,y)
plt.title('Number of Bathrooms versus Price')
plt.xlabel('Bathrooms')
plt.ylabel('Price')

z=np.polyfit(x,y,1)
p=np.poly1d(z)
plt.plot(x,p(x), color ='red', linewidth=3)
```

34
```
[<matplotlib.lines.Line2D at 0x22fdc925970>]
```

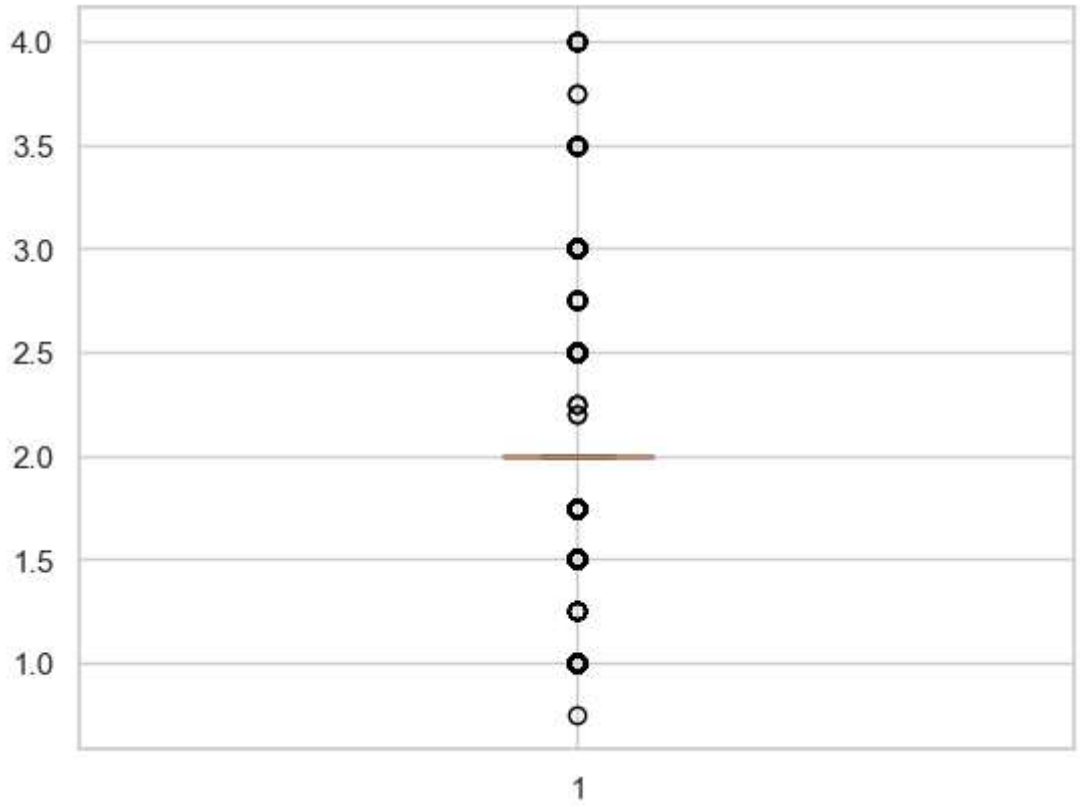



```
35  # view the distribution of bathrooms
    plt.hist(data['BATHS'], bins= [0,1,2,3,4,5])
    plt.title('Number of Bathrooms')
    plt.xlabel('Bathrooms')
    plt.show()
```



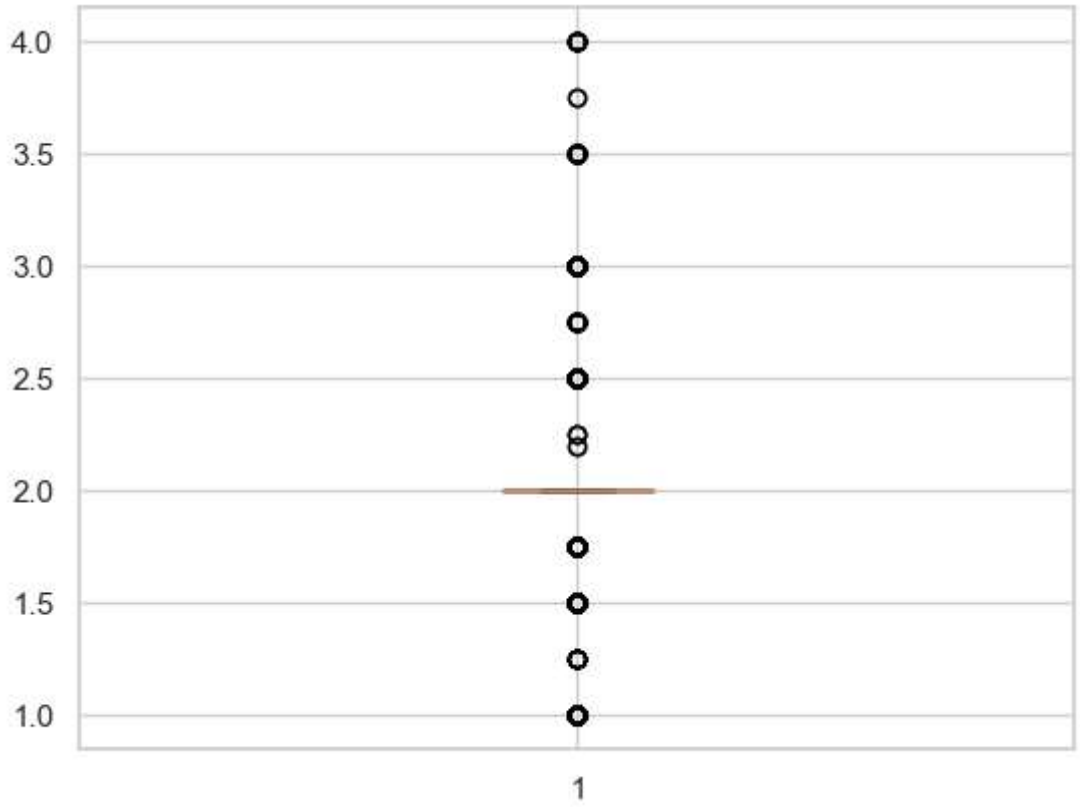
36
```
#identify outliers in the number of bathrooms
plt.boxplot(data['BATHS'])
```

36
```
{'whiskers': [<matplotlib.lines.Line2D at 0x22fdc9db970>,
  <matplotlib.lines.Line2D at 0x22fdc9d85b0>],
 'caps': [<matplotlib.lines.Line2D at 0x22fdc9d8820>,
  <matplotlib.lines.Line2D at 0x22fdc9d8fd0>],
 'boxes': [<matplotlib.lines.Line2D at 0x22fdc9dbf40>],
 'medians': [<matplotlib.lines.Line2D at 0x22fdc9d8eb0>],
 'fliers': [<matplotlib.lines.Line2D at 0x22fdc9c0040>],
 'means': []}
```



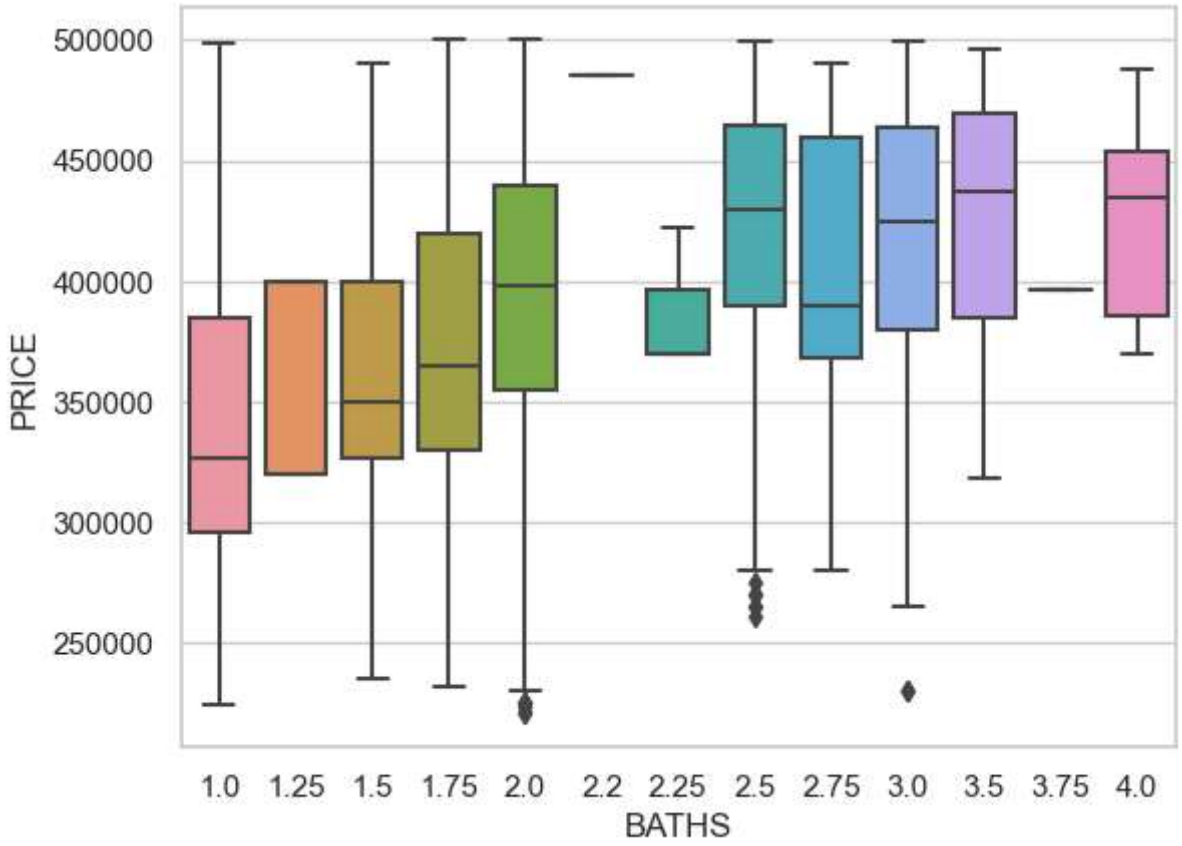

```
37  #remove less than 1 bathroom
    data = data[data.BATHS >= 1]
    plt.boxplot(data['BATHS'])
```

```
37  {'whiskers': [<matplotlib.lines.Line2D at 0x22fdc999e50>,
      <matplotlib.lines.Line2D at 0x22fdca31130>],
     'caps': [<matplotlib.lines.Line2D at 0x22fdca313d0>,
      <matplotlib.lines.Line2D at 0x22fdca31670>],
     'boxes': [<matplotlib.lines.Line2D at 0x22fdc999bb0>],
     'medians': [<matplotlib.lines.Line2D at 0x22fdca31910>],
     'fliers': [<matplotlib.lines.Line2D at 0x22fdca31bb0>],
     'means': []}
```

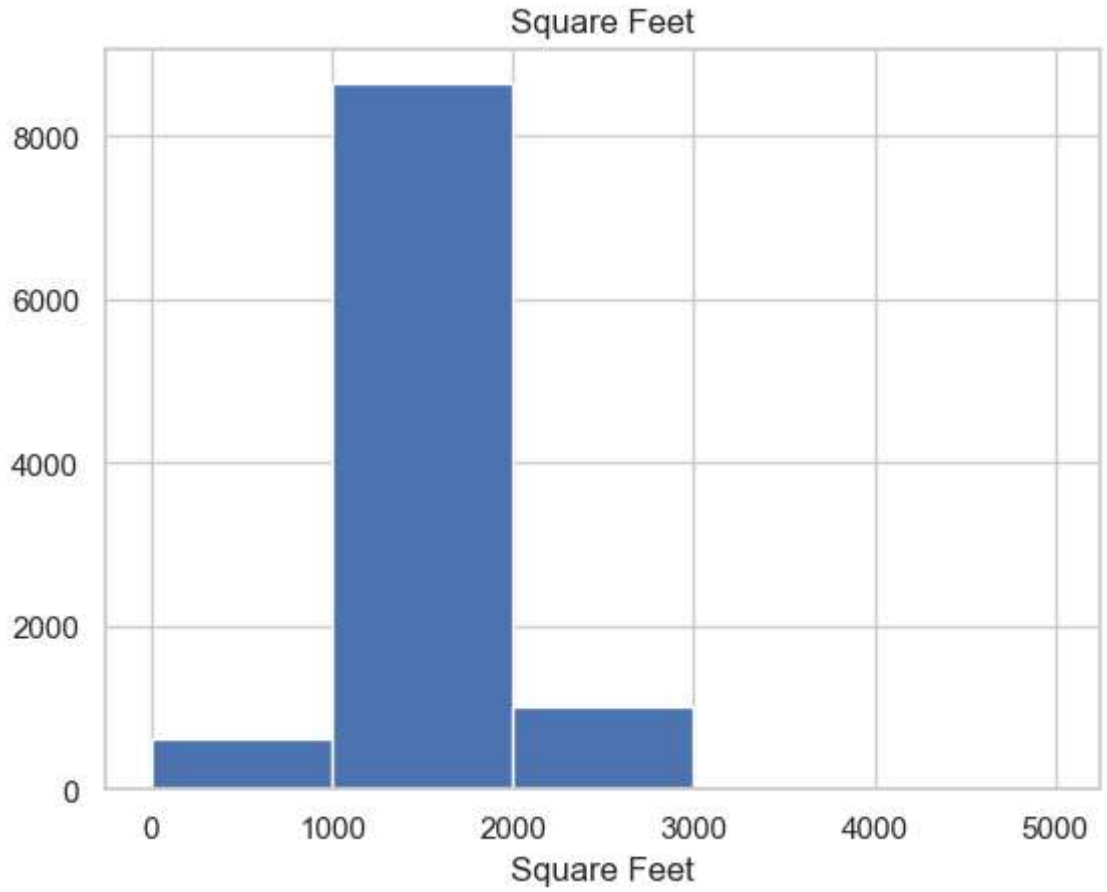



```
38  #view the relationship between bathrooms and price after removing outliers
    sn.set(style = 'whitegrid')
    sn.boxplot(x='BATHS',
               y='PRICE',
               data = data)
```

```
38  <AxesSubplot: xlabel='BATHS', ylabel='PRICE'>
```

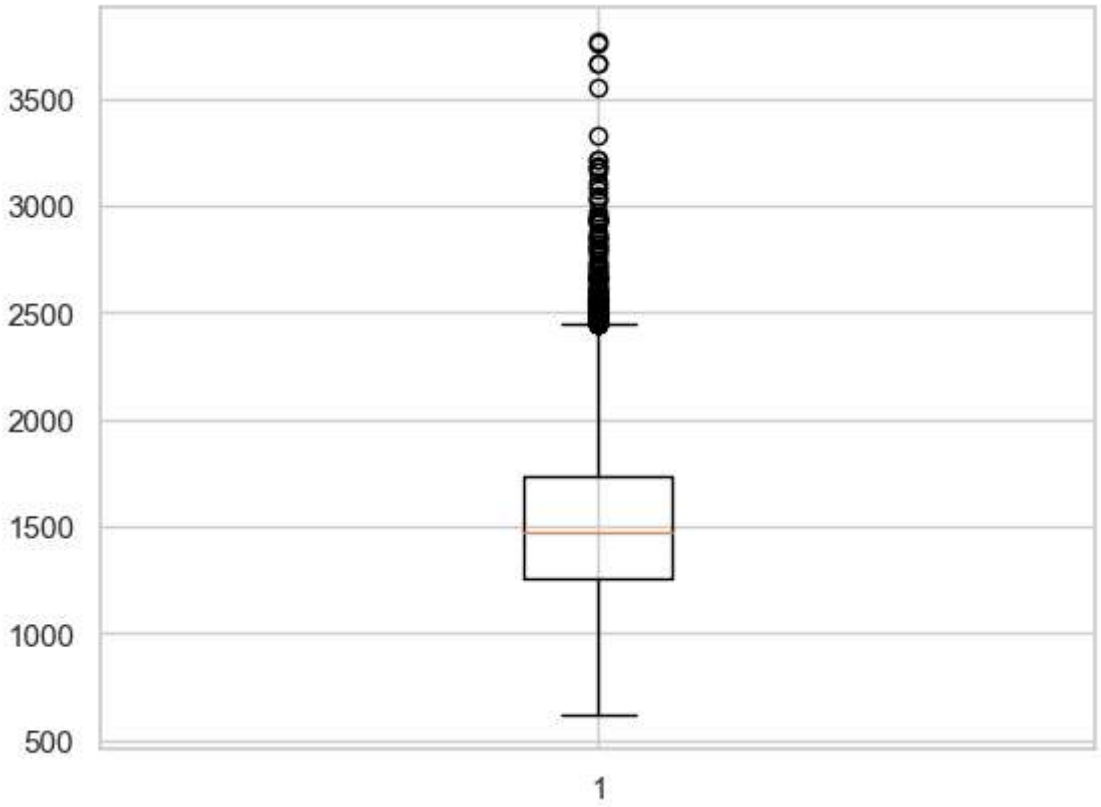



```
39  #view distribution of house square footage
    plt.hist(data['SQFT'], bins= [0,1000,2000,3000,4000,5000])
    plt.title('Square Feet')
    plt.xlabel('Square Feet')
    plt.show()
```




```
40  #view square footage as a boxplot
    plt.boxplot(data['SQFT'])
```

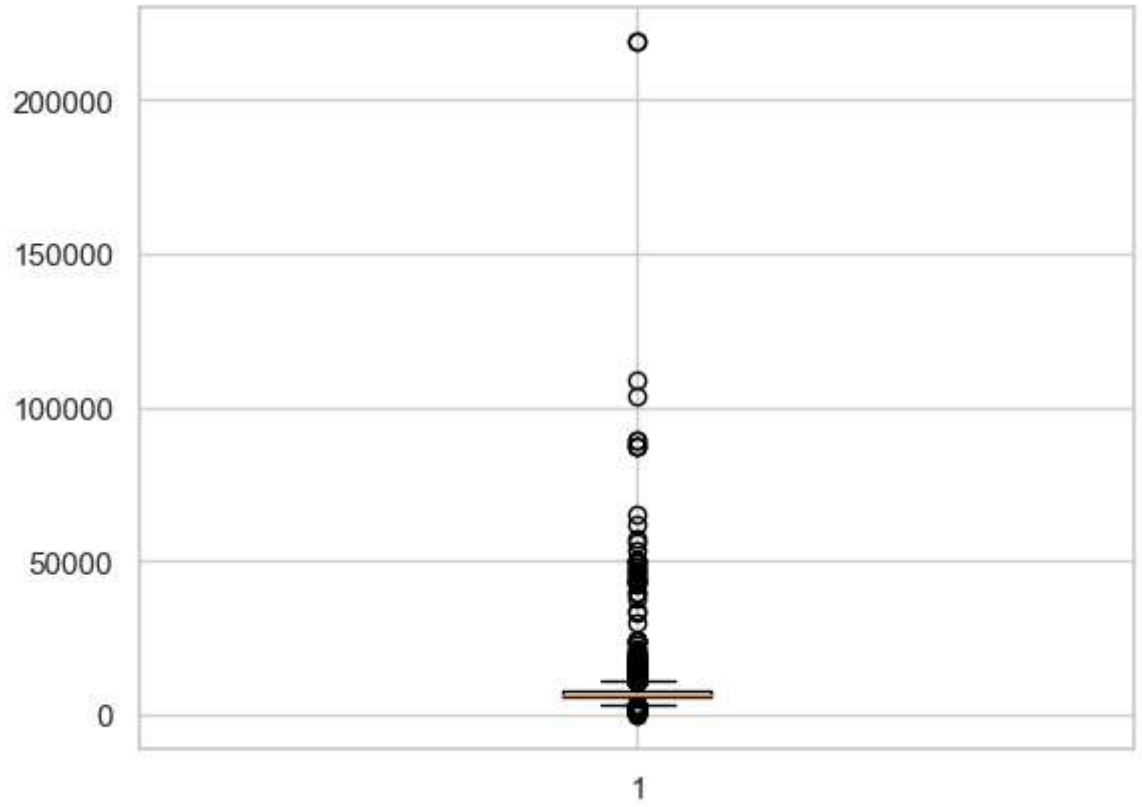
```
40  {'whiskers': [<matplotlib.lines.Line2D at 0x22fddddcc40>,
      <matplotlib.lines.Line2D at 0x22fddc31550>],
     'caps': [<matplotlib.lines.Line2D at 0x22fddc31970>,
      <matplotlib.lines.Line2D at 0x22fddc31df0>],
     'boxes': [<matplotlib.lines.Line2D at 0x22fddddc9a0>],
     'medians': [<matplotlib.lines.Line2D at 0x22fddc3b5e0>],
     'fliers': [<matplotlib.lines.Line2D at 0x22fddc3bc70>],
     'means': []}
```



Explore Lot Size

41
```
#view lot size in a boxplot to identify outliers
plt.boxplot(data['LOTSIZE'])
```

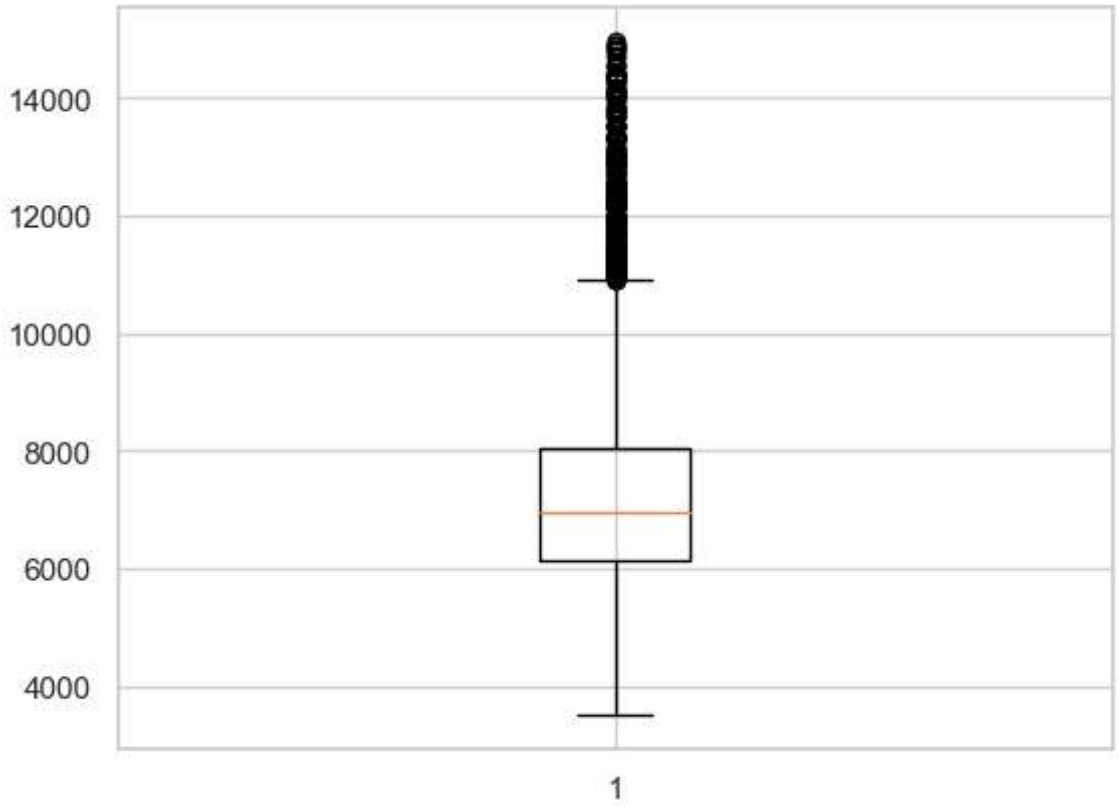
41
```
{'whiskers': [<matplotlib.lines.Line2D at 0x22fddbf8df0>,
  <matplotlib.lines.Line2D at 0x22fdddef0d0>],
 'caps': [<matplotlib.lines.Line2D at 0x22fdddef370>,
  <matplotlib.lines.Line2D at 0x22fdddef610>],
 'boxes': [<matplotlib.lines.Line2D at 0x22fddbf8b50>],
 'medians': [<matplotlib.lines.Line2D at 0x22fdddef8b0>],
 'fliers': [<matplotlib.lines.Line2D at 0x22fdddefb50>],
 'means': []}
```



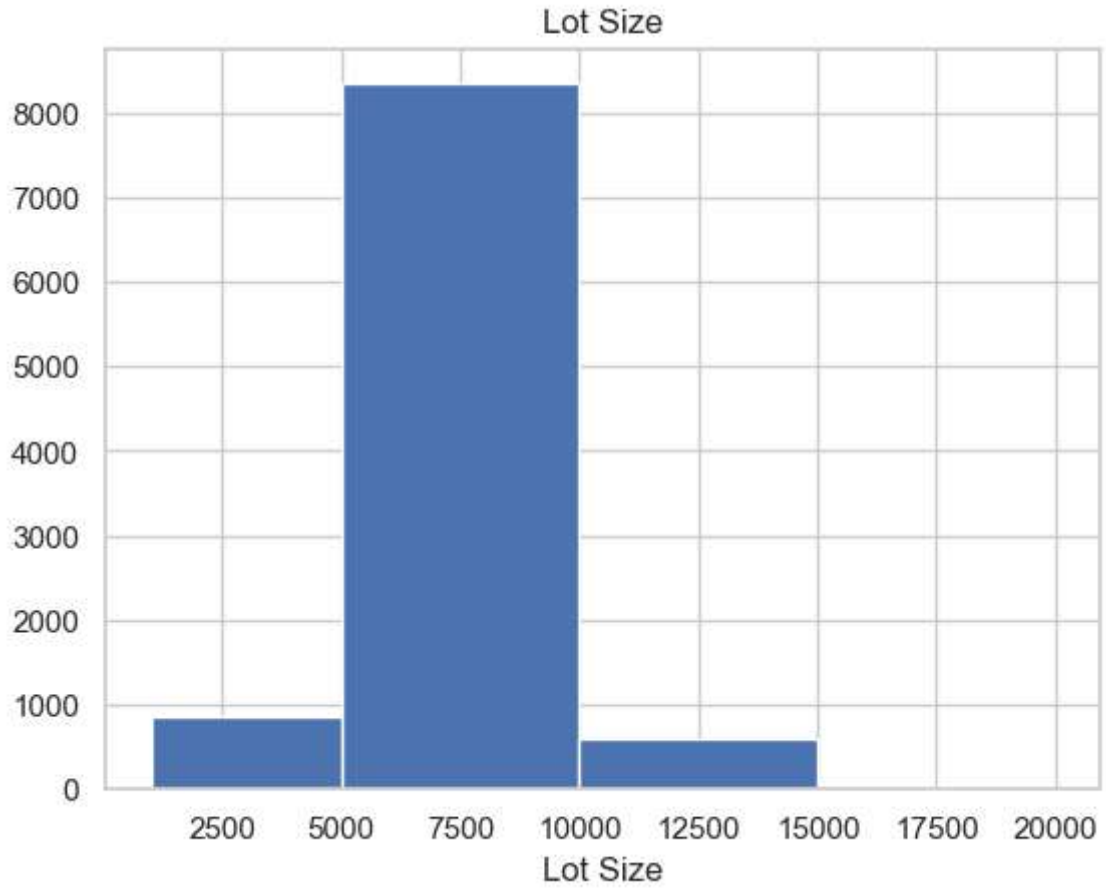
42
```
#Remove greater than 15,000sq ft and less than 3500 as unrealistic for price range
data = data[data.LOTSIZE < 15000]
data = data[data.LOTSIZE > 3500]
```

43
```
#view lotsize as a boxplot after removing outliers identified above
plt.boxplot(data['LOTSIZE'])
```

43
```
{'whiskers': [<matplotlib.lines.Line2D at 0x22fdde50640>,
  <matplotlib.lines.Line2D at 0x22fdde508e0>],
 'caps': [<matplotlib.lines.Line2D at 0x22fdde33250>,
  <matplotlib.lines.Line2D at 0x22fdde50d30>],
 'boxes': [<matplotlib.lines.Line2D at 0x22fdde503a0>],
 'medians': [<matplotlib.lines.Line2D at 0x22fdde50fd0>],
 'fliers': [<matplotlib.lines.Line2D at 0x22fdde602b0>],
 'means': []}
```

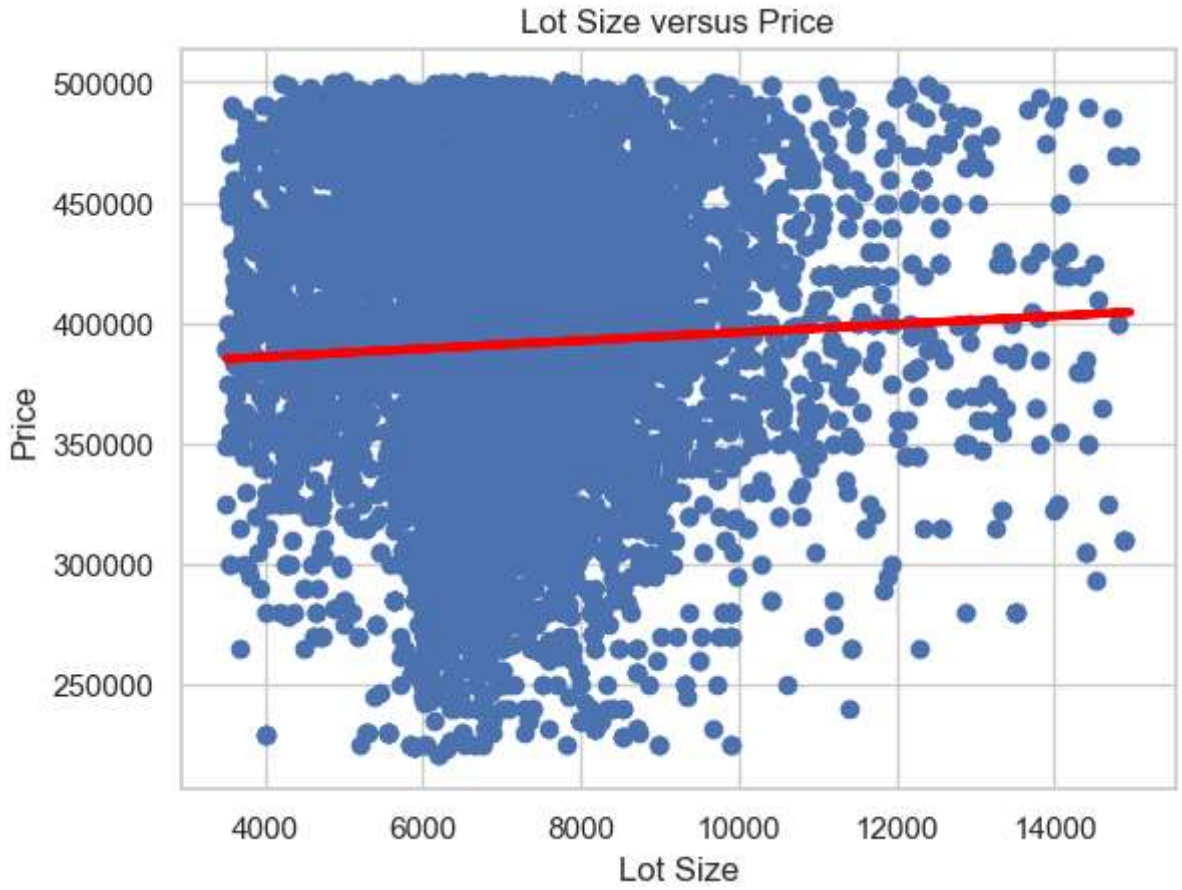



```
44  #view the distribution of houses by lotsize binning at 5000 sq ft intervals
    plt.hist(data['LOTSIZE'], bins= [1000,5000,10000,15000,20000])
    plt.title('Lot Size')
    plt.xlabel('Lot Size')
    plt.show()
```



45
```
y=data['PRICE']
x=data['LOTSIZE']
plt.scatter(x,y)
plt.title('Lot Size versus Price')
plt.xlabel('Lot Size')
plt.ylabel('Price')

z=np.polyfit(x,y,1)
p=np.poly1d(z)
plt.plot(x,p(x), color ='red', linewidth=3)
```

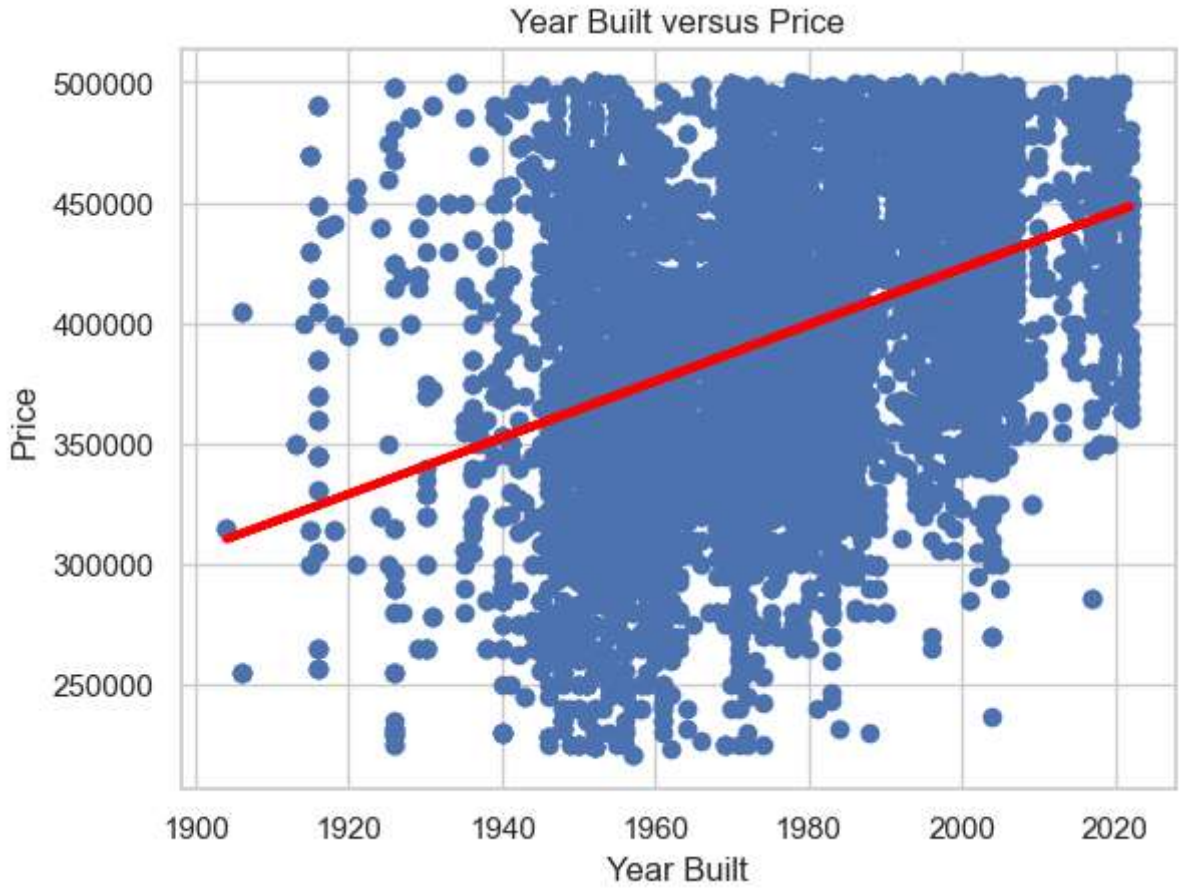
45
```
[<matplotlib.lines.Line2D at 0x22fddef51f0>]
```



Explore Year Built

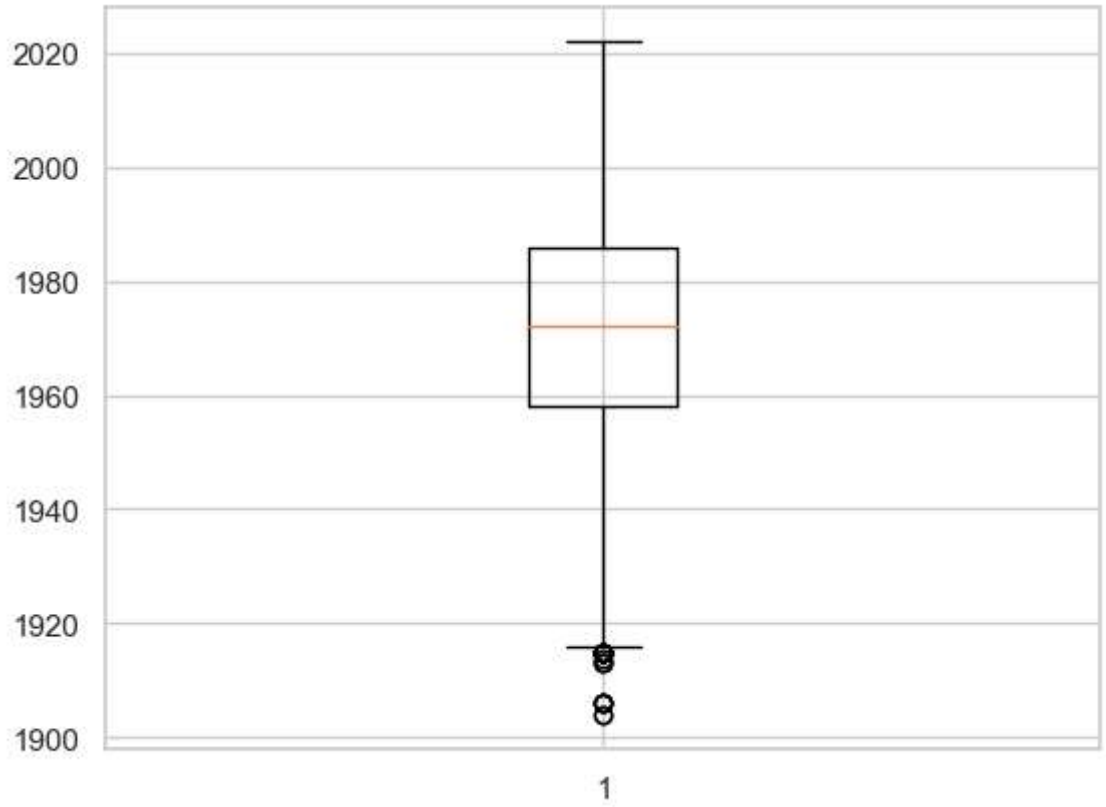
46
```
y=data['PRICE']
x=data['YEARBUILT']
plt.scatter(x,y)
plt.title('Year Built versus Price')
plt.xlabel('Year Built')
plt.ylabel('Price')

z=np.polyfit(x,y,1)
p=np.poly1d(z)
plt.plot(x,p(x), color ='red', linewidth=3)
```

46
```
[<matplotlib.lines.Line2D at 0x22fddecd130>]
```



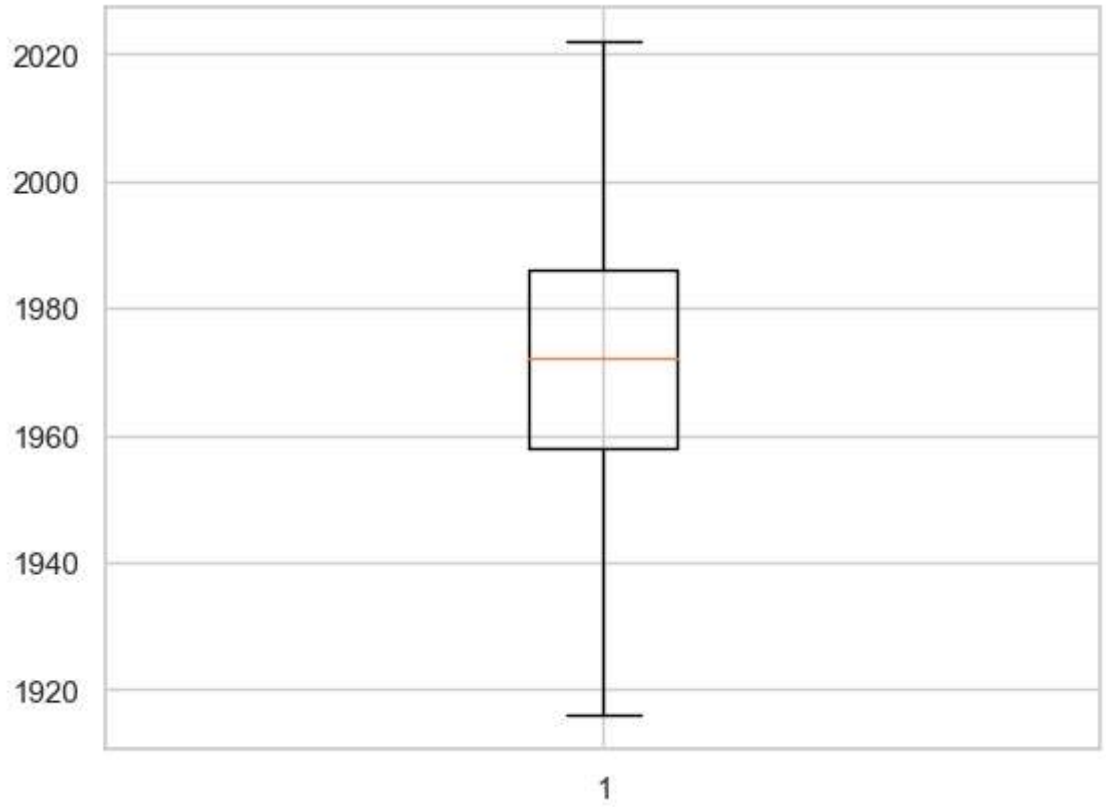
47
```
#view houses by year built to identify outliers
plt.boxplot(data['YEARBUILT'])
```

47
```
{'whiskers': [<matplotlib.lines.Line2D at 0x22fde564a00>,
  <matplotlib.lines.Line2D at 0x22fde564ca0>],
 'caps': [<matplotlib.lines.Line2D at 0x22fde564f40>,
  <matplotlib.lines.Line2D at 0x22fde2d1220>],
 'boxes': [<matplotlib.lines.Line2D at 0x22fde564730>],
 'medians': [<matplotlib.lines.Line2D at 0x22fde2d1430>],
 'fliers': [<matplotlib.lines.Line2D at 0x22fde2d16d0>],
 'means': []}
```



48
```
#Remove anything built before 1915
data = data[data.YEARBUILT > 1915]
plt.boxplot(data['YEARBUILT'])
```

48
```
{'whiskers': [<matplotlib.lines.Line2D at 0x22fde33c370>,
  <matplotlib.lines.Line2D at 0x22fde33c610>],
 'caps': [<matplotlib.lines.Line2D at 0x22fde33c8b0>,
  <matplotlib.lines.Line2D at 0x22fde33cb50>],
 'boxes': [<matplotlib.lines.Line2D at 0x22fde33c0d0>],
 'medians': [<matplotlib.lines.Line2D at 0x22fde33cdf0>],
 'fliers': [<matplotlib.lines.Line2D at 0x22fde34a0d0>],
 'means': []}
```




```
49 #Remove Sold Date, Address, City, State or Provice, MLS#, Latitude and Longitude

   model_data = data[['ZIPCODE','PRICE','BEDS','BATHS','SQFT','LOTSIZE','YEARBUILT','POOL','HOA']]
```

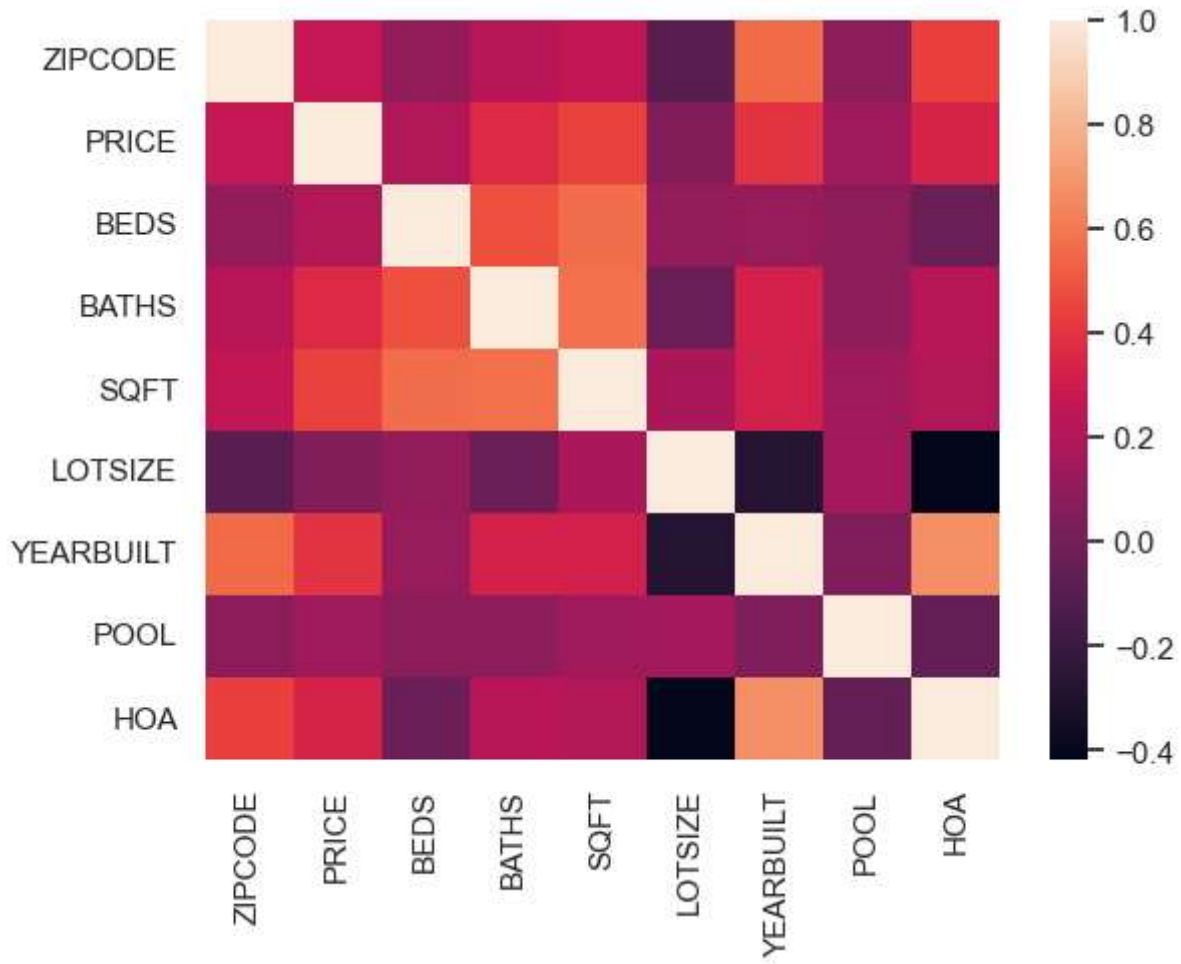
```
50 #view the first five rows of the model data and view the statistical data
   model_data.head()
   model_data.describe()
```

50

	ZIPCODE	PRICE	BEDS	BATHS	SQFT	LOTSIZE	YE
count	9788.000000	9788.000000	9788.000000	9788.000000	9788.000000	9788.000000	978
mean	85030.849816	391403.682162	3.232938	1.962725	1510.014508	7163.449734	197
std	16.180351	60204.959469	0.654691	0.453272	368.838202	1724.357887	20.
min	85003.000000	220500.000000	2.000000	1.000000	618.000000	3508.000000	191
25%	85019.000000	350000.000000	3.000000	2.000000	1256.750000	6125.000000	195
50%	85032.000000	395000.000000	3.000000	2.000000	1474.000000	6946.000000	197
75%	85041.000000	440000.000000	4.000000	2.000000	1734.000000	8031.250000	198
max	85087.000000	499999.000000	5.000000	4.000000	3770.000000	14963.000000	202

Heatmap of numerical values

```
51 ax = sn.heatmap(data = model_data.corr())
```




```
52 #correlation matrix of the model data
   corr_matrix = model_data.corr()
   corr_matrix['PRICE'].sort_values(ascending=False)
```

```
52 PRICE        1.000000
   SQFT         0.444500
   YEARBUILT    0.397496
   BATHS        0.361879
   HOA          0.338139
   ZIPCODE      0.262972
   BEDS         0.206007
   POOL         0.151689
   LOTSIZE      0.050020
   Name: PRICE, dtype: float64
```

```
53 #view data types of the model data
   model_data.dtypes
```

```
53 ZIPCODE        int64
   PRICE        float64
   BEDS           int32
   BATHS        float64
   SQFT           int64
   LOTSIZE        int64
   YEARBUILT      int64
   POOL           int32
```

```
HOA            int32
dtype: object
```

Build Models

```
54 #build the datasets removing the dependent variable PRICE from the set of independent variables
   X = model_data.drop('PRICE', axis = 1)
   y = model_data['PRICE']
```

```
55 #use one hot encoding on the feature variables

   encoder = OneHotEncoder(categories="auto", handle_unknown="ignore",sparse=False)
   X=encoder.fit_transform(X)
```

```
56 #scale the feature variables
   scaler = StandardScaler()
   X=scaler.fit_transform(X)
   X
```

```
56 array([[-0.05544728, -0.02860063,  7.20472818, ...,  2.30804781,
           0.53086999, -0.53086999],
         [-0.05544728, -0.02860063, -0.13879774, ...,  2.30804781,
           0.53086999, -0.53086999],
         [-0.05544728, -0.02860063, -0.13879774, ...,  2.30804781,
           0.53086999, -0.53086999],
         ...,
         [-0.05544728, -0.02860063, -0.13879774, ..., -0.43326659,
          -1.88370038,  1.88370038],
         [-0.05544728, -0.02860063, -0.13879774, ..., -0.43326659,
          -1.88370038,  1.88370038],
         [-0.05544728, -0.02860063, -0.13879774, ..., -0.43326659,
          -1.88370038,  1.88370038]])
```

```
57 #build the test and train datasets using 80% of the dataset for training and 20% for testing
   X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2, random_state=4)
```

After testing various models including Linear Regression, XGBoost Regression, KNeighbor Regression, Random Forest Regression and ANN, the model with the highest R^2 score and the lowest MAE was the Random Forest Regression model. After selecting the Random Forest Regessor model, GridSearchCV was used to find the best parameters.

```
58 #Using GridSearchCV found the best parameters for the random forest model
   RandFor = RandomForestRegressor(bootstrap=False, max_features=3, n_estimators=300, random_state=4)
   RandFor.fit(X_train,y_train)

   y_pred = RandFor.predict(X_test)
```

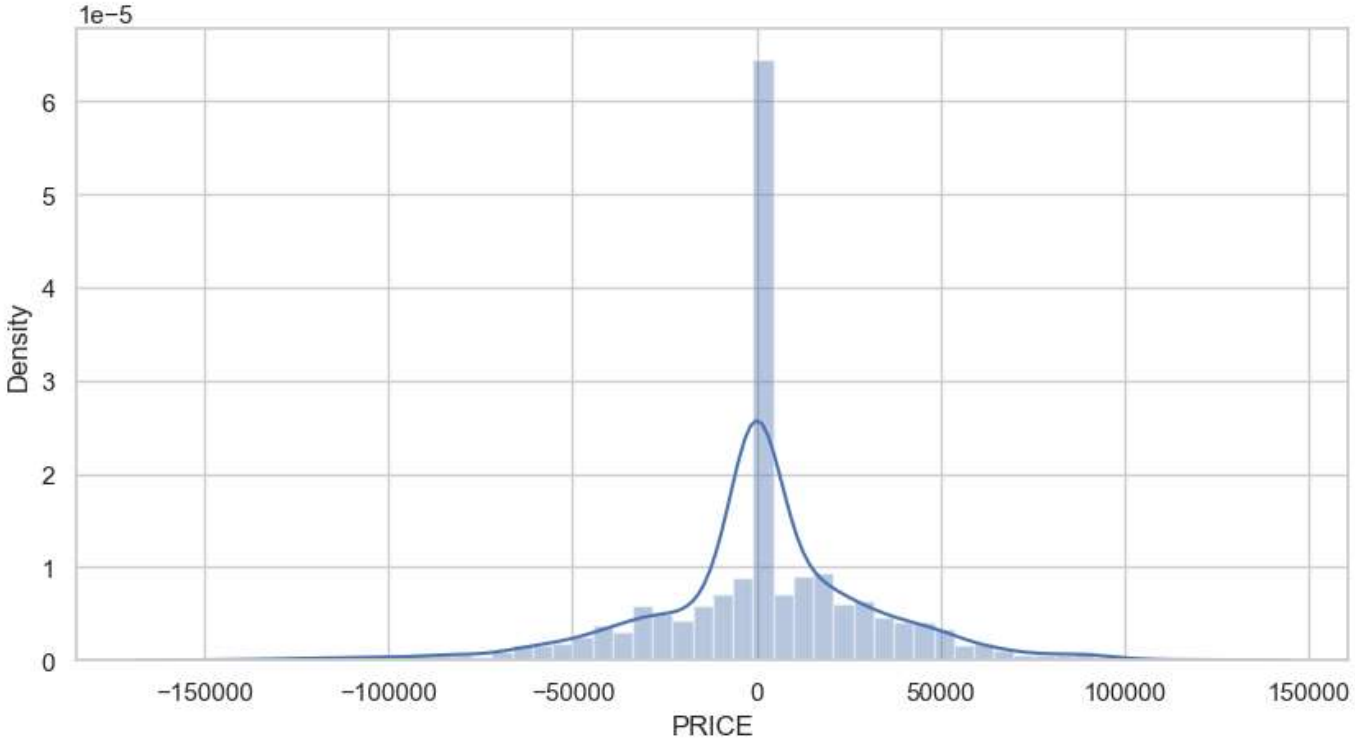
```
59 print('MAE: ', mean_absolute_error(y_test,y_pred))
   print('MSE: ', mean_squared_error(y_test,y_pred))
   print('R2: ', r2_score(y_test,y_pred))
   print('VarScore: ',explained_variance_score(y_test,y_pred) )
   print('RMSE: ', np.sqrt(mean_squared_error(y_test,y_pred)))
```

```
MAE:  20136.20126530149
MSE:  986204738.5211413
R2:  0.7296995053618887
VarScore:  0.730118413636752
RMSE:  31403.896868400603
```

60
```
fig=plt.figure(figsize=(10,5))
residuals = (y_test - y_pred)
sn.distplot(residuals)
```

```
C:\Users\belbi\AppData\Local\Temp\ipykernel_11712\1542007201.py:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sn.distplot(residuals)
```

60
```
<AxesSubplot: xlabel='PRICE', ylabel='Density'>
```



61
```
#change working directory
os.chdir(r"C:\Users\belbi\PycharmProject\Capstone")
```

62
```
## to test the model

testdata =pd.read_excel(r'C:\Users\belbi\PycharmProject\Capstone\testdata.xlsx',sheet_name='Sheet1')
datatest = encoder.transform(testdata)
```

63
```
datatest = encoder.transform(testdata)
datatest = scaler.transform(datatest)
```

```
prediction = RandFor.predict(datatest)
prediction
```

```
63 array([411074.33      , 422146.5       , 440055.76666667])
```

Save trained model

```
64 # pickle file to be used share model data with python pages used for Streamlit app

model = {"model": RandFor}
with open('trained_model.pkl','wb') as file:
    pickle.dump(model,file)
```

```
65 with open('trained_model.pkl','rb')as file:
    model = pickle.load(file)
```

```
66 # pickle file to be used to share encoder with python pages used for Streamlit app

with open('encoder.pkl' ,'wb') as f:
    pickle.dump(encoder,f)
```

```
67 with open('encoder.pkl','rb') as f:
    encoder = pickle.load(f)
```

```
68 # pickle file to be used to share standard scaler with python pages used for Streamlit app

with open('scaler.pkl' ,'wb') as f:
    pickle.dump(scaler,f)
```

```
69 with open('scaler.pkl','rb') as f:
    scaler=pickle.load(f)
```

```
70
# pickle file to be used to share data with python pages used for Streamlit app

with open('data.pkl','wb') as f:
    pickle.dump(data,f)
```

```
71 with open('data.pkl','rb') as f:
    data=pickle.load(f)
```

```
72 # pickle file to be used to share data with python pages used for Streamlit app

with open('modeldata.pkl','wb') as f:
    pickle.dump(model_data,f)
```

```
73 with open('modeldata.pkl','rb') as f:
    modeldata=pickle.load(f)
```