Import packages

[ ] ↳ *1 cell hidden*

Importing data

[ ] ↳ *3 cells hidden*

Locating None values

[ ] ↳ *2 cells hidden*

Correct object numerical values

[ ] ↳ *5 cells hidden*

Searching for Null/NaN values

[ ] ↳ *3 cells hidden*

Dropping unnecesary values

```
#Let's analyse the MLS column
MLS_data=df_drop["MLS"]
vecMLS=np.unique(MLS_data)
print(len(vecMLS))
print(df_drop.shape)
#We see that both of them have the same length, so the variable MLS is unnecesary for the dat
df_drop=df_drop.drop(['MLS'],axis=1)
```

```
    4973
    (4973, 16)
```

```
df_drop.head(5)
```

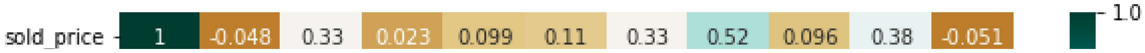
	sold_price	zipcode	longitude	latitude	lot_acres	taxes	year_built	be
0	5300000.0	85637	-1.103.782	31.356.362	2154.00	5272.00	1941	
1	4200000.0	85646	-111.045.371	31.594.213	1707.00	10422.36	1997	
2	4200000.0	85646	-111.040.707	31.594.844	1707.00	10482.00	1997	
3	4500000.0	85646	-111.035.925	31.645.878	636.67	8418.58	1930	
4	3411450.0	85750	-110.813.768	32.285.162	3.21	15393.00	1995	

Double-click (or enter) to edit

Correlated variables

```
#Getting the correlation matrix of the data frame
plt.figure(figsize=(10,5))
c= df_drop.corr()
sns.heatmap(c,cmap="BrBG",annot=True)
c
```

	sold_price	zipcode	lot_acres	taxes	year_built	bedrooms	bathr
sold_price	1.000000	-0.047941	0.332954	0.023265	0.099163	0.114050	0.32
zipcode	-0.047941	1.000000	-0.128443	-0.001697	0.014823	0.040643	-0.05
lot_acres	0.332954	-0.128443	1.000000	-0.000569	-0.067181	0.069806	0.05
taxes	0.023265	-0.001697	-0.000569	1.000000	-0.004180	0.005146	0.00
year_built	0.099163	0.014823	-0.067181	-0.004180	1.000000	-0.183764	-0.05
bedrooms	0.114050	0.040643	0.069806	0.005146	-0.183764	1.000000	0.68
bathrooms	0.326405	-0.056332	0.055510	0.008946	-0.051401	0.687501	1.00
sqrt_ft	0.524503	-0.007799	0.107511	0.037633	-0.057688	0.548193	0.66
garage	0.095537	0.092184	-0.070652	0.005666	0.322810	0.038145	0.09
fireplaces	0.384310	-0.018166	0.086382	0.022548	-0.127501	0.145279	0.22
HOA	-0.050562	-0.053586	0.087258	-0.009001	-0.305000	0.147353	0.08

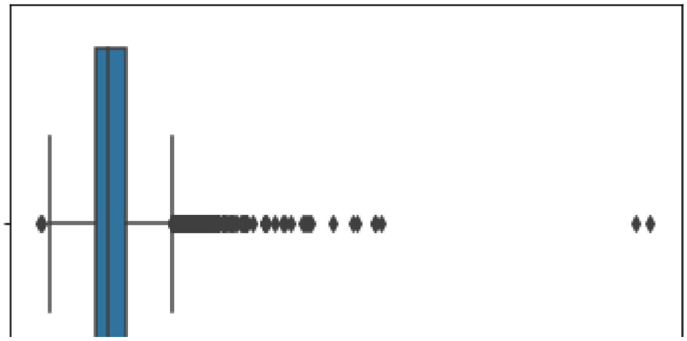


```
#There is a high correlation between the size of the house (sqrt_ft) and the sold price, as e
```

Searching for outliers

```
#Getting the box plots of all numerical variables
df=df_drop
print(df.shape)
sns.boxplot(x=df['sqrt_ft'])
#sns.boxplot(x=df['sold_price'])
#sns.boxplot(x=df['zipcode'])
#sns.boxplot(x=df['lot_acres'])
#sns.boxplot(x=df['taxes'])
#sns.boxplot(x=df['bedrooms'])
#sns.boxplot(x=df['year_built'])
#sns.boxplot(x=df['bathrooms'])
#sns.boxplot(x=df['garage'])
#sns.boxplot(x=df['fireplaces'])
plt.show()
```

```
(4973, 15)
```



```
##Getting an IQR analysis
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3 - Q1
print(IQR)
```

```
sold_price    252500.00
zipcode           32.00
lot_acres          1.17
taxes           3283.00
year_built        19.00
bedrooms           1.00
bathrooms          1.00
sqrt_ft         1084.00
garage             1.00
fireplaces         2.00
HOA             2005.00
dtype: float64
```

```
df = df[~(df['fireplaces'] > 6)]
df.shape
```

```
(4961, 15)
```

```
#Based on the plots, we eliminate the significant outliers
df = df[~(df['garage'] > 15)]
df = df[~(df['bathrooms'] == 00)]
df = df[~(df['bathrooms'] > 20)]
df = df[~(df['bedrooms'] > 15)]
df = df[~(df['taxes'] ==0)]
```

```
df.shape
```

```
(4942, 17)
```

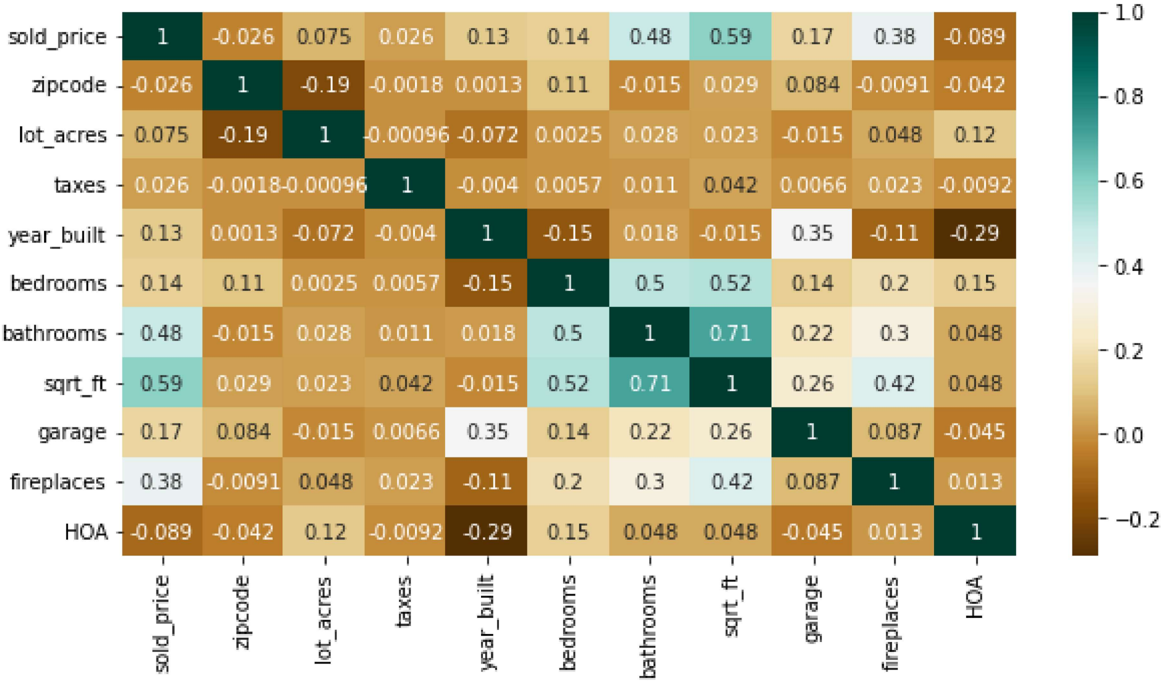
```
df = df[~(df['lot_acres'] > 500)]
df = df[~(df['lot_acres']==0)]
df = df[~(df['sold_price'] > 3000000)]
df = df[~(df['sqrt_ft'] >10000)]
```

```
#We get the final shape of the data
df.shape
```

```
(4889, 17)
```

```
#Plotting in a heat map the correlation matrix once again we notice the following beahvior
plt.figure(figsize=(10,5))
c= df.corr()
sns.heatmap(c,cmap="BrBG",annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f16e76bff50>
```



	sold_price	zipcode	lot_acres	taxes	year_built	bedrooms	bathrooms	sqrt_ft	garage	fireplaces	HOA
sold_price	1	-0.026	0.075	0.026	0.13	0.14	0.48	0.59	0.17	0.38	-0.089
zipcode	-0.026	1	-0.19	-0.0018	0.0013	0.11	-0.015	0.029	0.084	-0.0091	-0.042
lot_acres	0.075	-0.19	1	-0.00096	-0.072	0.0025	0.028	0.023	-0.015	0.048	0.12
taxes	0.026	-0.0018	-0.00096	1	-0.004	0.0057	0.011	0.042	0.0066	0.023	-0.0092
year_built	0.13	0.0013	-0.072	-0.004	1	-0.15	0.018	-0.015	0.35	-0.11	-0.29
bedrooms	0.14	0.11	0.0025	0.0057	-0.15	1	0.5	0.52	0.14	0.2	0.15
bathrooms	0.48	-0.015	0.028	0.011	0.018	0.5	1	0.71	0.22	0.3	0.048
sqrt_ft	0.59	0.029	0.023	0.042	-0.015	0.52	0.71	1	0.26	0.42	0.048
garage	0.17	0.084	-0.015	0.0066	0.35	0.14	0.22	0.26	1	0.087	-0.045
fireplaces	0.38	-0.0091	0.048	0.023	-0.11	0.2	0.3	0.42	0.087	1	0.013
HOA	-0.089	-0.042	0.12	-0.0092	-0.29	0.15	0.048	0.048	-0.045	0.013	1

Encoding Kitchen features and floor covering

[ ] ↳ *4 cells hidden*

✓ 0s completed at 5:01 PM ● ✕