Anthony Ayala

8/20/23

Project 6 Appendix

## Null Analysis: What variables contain nulls and how do you propose to address null values.

I decided to take on two methods of handlings nulls to see if there is a significant difference in model selection. The first method is to create a subset of the Boston data where it drops all the nulls and be left with 13,712 rows, and the second method is to keep the Boston dataset and just fill in all the nulls with the 0's to keep all the 14,225 rows. The two variables with missing values were "yr_remod" and "land_sf". Below are pictures of code chunk outputs:

```
1 # We have 33 columns, 14255 rows, 4 floats, 14 integers, and 15 objects
2 boston.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14225 entries, 0 to 14224
Data columns (total 33 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   pid              14225 non-null  int64
 1   zipcode          14225 non-null  int64
 2   own_occ          14225 non-null  object
 3   land_sf          14221 non-null  float64
 4   yr_built         14225 non-null  int64
 5   yr_remod         13714 non-null  float64
 6   living_area      14225 non-null  int64
 7   num_floors       14225 non-null  float64
 8   structure_class  14225 non-null  object
 9   r_bldg_styl      14225 non-null  object
 10  r_roof_typ       14225 non-null  object
 11  r_ext_fin        14225 non-null  object
 12  r_total_rms      14225 non-null  int64
 13  r_bdrms          14225 non-null  int64
 14  r_full_bth       14225 non-null  int64
 15  r_half_bth       14225 non-null  int64
 16  r_bth_style      14225 non-null  object
 17  r_kitch          14225 non-null  int64
 18  r_kitch_style    14225 non-null  object
 19  r_heat_typ       14225 non-null  object
 20  r_ac             14225 non-null  object
 21  r_fplace         14225 non-null  int64
 22  r_ext_cnd        14225 non-null  object
 23  r_ovrall_cnd     14225 non-null  object
 24  r_int_cnd        14225 non-null  object
 25  r_int_fin        14225 non-null  object
 26  r_view           14225 non-null  object
 27  zip              14225 non-null  int64
 28  population       14225 non-null  int64
 29  pop_density      14225 non-null  int64
 30  median_income    14225 non-null  int64
 31  city_state       14225 non-null  object
 32  av_total         14225 non-null  float64
dtypes: float64(4), int64(14), object(15)
memory usage: 3.6+ MB
```

```
1 boston.isnull().sum()
```

```
pid                0
zipcode            0
own_occ            0
land_sf            4
yr_built           0
yr_remod         511
living_area        0
num_floors         0
structure_class    0
r_bldg_styl        0
r_roof_typ         0
r_ext_fin          0
r_total_rms        0
r_bdrms            0
r_full_bth         0
r_half_bth         0
r_bth_style        0
r_kitch            0
r_kitch_style      0
r_heat_typ         0
r_ac               0
r_fplace           0
r_ext_cnd          0
r_ovrall_cnd       0
r_int_cnd          0
r_int_fin          0
r_view             0
zip                0
population         0
pop_density        0
median_income      0
city_state         0
av_total           0
dtype: int64
```

```
1 # Instead of dropping nulls for the sake of analysis, we will rather create a subset of the boston data that filters out the nulls.
2 # We would rather not drop the nulls because we don't want to cause issues for regression
3 boston_subset = boston.dropna(axis = 0, subset=['yr_remod', 'land_sf'])
4 boston_subset
```

| | pid | zipcode | own_occ | land_sf | yr_built | yr_remod | living_area | num_floors | structure_class | r_bldg_styl | ... | r_ovrall_cnd | r_i |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 10 | 2136 | Y | 10288.00 | 1992 | 0.00 | 1681 | 1.00 | R | RR | ... | A | |
| 1 | 20 | 2132 | Y | 10148.00 | 1900 | 2016.00 | 3024 | 2.50 | R | CL | ... | G | |
| 2 | 30 | 2132 | Y | 8512.00 | 1920 | 0.00 | 1160 | 2.00 | R | CL | ... | A | |
| 3 | 40 | 2124 | Y | 3187.00 | 1900 | 2001.00 | 1868 | 2.00 | R | CL | ... | G | |
| 4 | 50 | 2136 | Y | 10088.00 | 1971 | 1975.00 | 1534 | 1.00 | R | RR | ... | A | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 14220 | 142210 | 2124 | Y | 3717.00 | 1925 | 1995.00 | 1703 | 2.00 | R | CL | ... | A | |
| 14221 | 142220 | 2132 | Y | 3895.00 | 1920 | 2004.00 | 1350 | 2.00 | R | CL | ... | A | |
| 14222 | 142230 | 2132 | Y | 4700.00 | 1928 | 0.00 | 1490 | 2.00 | R | CL | ... | A | |
| 14223 | 142240 | 2124 | Y | 5250.00 | 1925 | 0.00 | 1404 | 2.00 | R | CL | ... | A | |
| 14224 | 142250 | 2136 | Y | 5000.00 | 1945 | 0.00 | 1157 | 1.50 | R | CP | ... | A | |

13712 rows × 33 columns

Also, we performed some transformations and cleaning of the data was performed to handle the “yr_remod” column since there were a lot of homes that were not remodeled and fixing data types like “number of floors” and “year remodeled” by converting those variables to integers. This was applied to both methods. The code for that is down below:

```
boston_subset['yr_remod'].value_counts() # There are 9,057 homes that were not remodeled
```

```
0.00       9057
2015.00     347
2016.00     259
2003.00     258
2002.00     217
           ...
1949.00       1
1940.00       1
1948.00       1
1947.00       1
1951.00       1
Name: yr_remod, Length: 79, dtype: int64
```

```
# Create new column in our boston_subset data for yr_remod to determine whether a house has been remodeled. This helps with dealing with the 0's
def remod_yn(x):
 if x > 0:
  return "yes"
 else:
  return "no"
boston_subset["remod_ind"] = boston['yr_remod'].apply(remod_yn)
boston_subset["remod_ind"].value_counts()
```

```
no     9057
yes    4655
Name: remod_ind, dtype: int64
```

```
# fix datatypes
boston_subset['num_floors'] = boston_subset['num_floors'].astype(int)
boston_subset['yr_remod'] = boston_subset['yr_remod'].astype(int)

# transform
boston_subset['property_age'] = 2023 - boston_subset['yr_built']
boston_subset.head()
```

| | pid | zipcode | own_occ | land_sf | yr_built | yr_remod | living_area | num_floors | structure_class | r_bldg_styl | ... | r_int_fin | r_view | zip | population | pop_density | median_income | city_state | av_total | remod_ind | property_age |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 10 | 2136 | Y | 10288.00 | 1992 | 0 | 1681 | 1 | R | RR | ... | N | A | 2136 | 28488 | 6207 | 58890 | Hyde Park, MA | 321200.00 | no | 31 |
| 1 | 20 | 2132 | Y | 10148.00 | 1900 | 2016 | 3024 | 2 | R | CL | ... | N | G | 2132 | 36314 | 13251 | 75446 | Cambridge, MA | 845475.93 | yes | 123 |
| 2 | 30 | 2132 | Y | 8512.00 | 1920 | 0 | 1160 | 2 | R | CL | ... | N | A | 2132 | 36314 | 13251 | 75446 | Cambridge, MA | 401230.03 | no | 103 |
| 3 | 40 | 2124 | Y | 3187.00 | 1900 | 2001 | 1868 | 2 | R | CL | ... | N | F | 2124 | 47783 | 15913 | 48841 | Dorchester Center, MA | 450500.00 | yes | 123 |
| 4 | 50 | 2136 | Y | 10088.00 | 1971 | 1975 | 1534 | 1 | R | RR | ... | N | G | 2136 | 28488 | 6207 | 58890 | Hyde Park, MA | 368094.74 | yes | 52 |

5 rows × 35 columns

## Model Performance:

Model Performance was done by joining the Boston data/Boston Subset data and joining it to the model 1 prediction/model 2 predictions on a column called “pid”. Then we created a new column for both data sets called “Residual”, which was created by taking “av_total” - “pred”. This is simply a statistics calculation where you take the actual value – predicted value to then assess how good the model fits the dataset.

### Method 1 Predictions:

```
# Let's check Model 1's performance
r2 = r2_score(boston_pred1['av_total'],boston_pred1['pred'] )
mse = mean_squared_error(boston_pred1['av_total'], boston_pred1['pred'])
mae = mean_absolute_error(boston_pred1['av_total'], boston_pred1['pred'])

print("-- Linear Regression Stats for Boston House Prices -- ")
print(f'R-Square: {r2:.3f}')
print(f" - RSQUARE: approximately {r2:.1%} of the variability in the sale prices can be explained by our model.")

print(f'Root Mean Squared Error: {mse**0.5:,.2f}')
print(f" - RMSE: on average, our predictions are approximately ${mse**0.5:,.2f} away from the actual sale price")
print(f'Mean Absolute Error: {mae:,.2f}')
print(f" - MAE:  on average, the predictions made by the model are off by +/- ${mae:,.2f} from the actual")
```

```
-- Linear Regression Stats for Boston House Prices --
R-Square: 0.437
 - RSQUARE: approximately 43.7% of the variability in the sale prices can be explained by our model.
Root Mean Squared Error: 107,324.52
 - RMSE: on average, our predictions are approximately $107,324.52 away from the actual sale price
Mean Absolute Error: 76,967.08
 - MAE:  on average, the predictions made by the model are off by +/- $76,967.08 from the actual
```

```
# Let's check Model 2's performance
r2 = r2_score(boston_pred2['av_total'],boston_pred2['pred'] )
mse = mean_squared_error(boston_pred2['av_total'], boston_pred2['pred'])
mae = mean_absolute_error(boston_pred2['av_total'], boston_pred2['pred'])

print("-- Linear Regression Stats for Boston House Prices -- ")
print(f'R-Square: {r2:.3f}')
print(f" - RSQUARE: approximately {r2:.1%} of the variability in the sale prices can be explained by our model.")

print(f'Root Mean Squared Error: {mse**0.5:,.2f}')
print(f" - RMSE: on average, our predictions are approximately ${mse**0.5:,.2f} away from the actual sale price")
print(f'Mean Absolute Error: {mae:,.2f}')
print(f" - MAE:  on average, the predictions made by the model are off by +/- ${mae:,.2f} from the actual")
```

```
-- Linear Regression Stats for Boston House Prices --
R-Square: 0.947
 - RSQUARE: approximately 94.7% of the variability in the sale prices can be explained by our model.
Root Mean Squared Error: 32,857.75
 - RMSE: on average, our predictions are approximately $32,857.75 away from the actual sale price
Mean Absolute Error: 23,451.97
 - MAE:  on average, the predictions made by the model are off by +/- $23,451.97 from the actual
```

**Method 2 Predictions:**

```
# Let's check Model 1's performance
r2 = r2_score(boston_predictions1['av_total'],boston_predictions1['pred'] )
mse = mean_squared_error(boston_predictions1['av_total'], boston_predictions1['pred'])
mae = mean_absolute_error(boston_predictions1['av_total'], boston_predictions1['pred'])

print("-- Linear Regression Stats for Boston House Prices -- ")
print(f'R-Square: {r2:.3f}')
print(f" - RSQUARE: approximately {r2:.1%} of the variability in the sale prices can be explained by our model.")

print(f'Root Mean Squared Error: {mse**0.5:,.2f}')
print(f" - RMSE: on average, our predictions are approximately ${mse**0.5:,.2f} away from the actual sale price")
print(f'Mean Absolute Error: {mae:,.2f}')
print(f" - MAE:  on average, the predictions made by the model are off by +/- ${mae:,.2f} from the actual")
```

```
-- Linear Regression Stats for Boston House Prices --
R-Square: 0.436
 - RSQUARE: approximately 43.6% of the variability in the sale prices can be explained by our model.
Root Mean Squared Error: 108,148.17
 - RMSE: on average, our predictions are approximately $108,148.17 away from the actual sale price
Mean Absolute Error: 77,629.65
 - MAE:  on average, the predictions made by the model are off by +/- $77,629.65 from the actual
```

```
[99]
# Let's check Model 2's performance
r2 = r2_score(boston_predictions2['av_total'], boston_predictions2['pred'] )
mse = mean_squared_error(boston_predictions2['av_total'], boston_predictions2['pred'])
mae = mean_absolute_error(boston_predictions2['av_total'], boston_predictions2['pred'])

print("-- Linear Regression Stats for Boston House Prices -- ")
print(f'R-Square: {r2:.3f}')
print(f" - RSQUARE: approximately {r2:.1%} of the variability in the sale prices can be explained by our model.")

print(f'Root Mean Squared Error: {mse**0.5:,.2f}')
print(f" - RMSE: on average, our predictions are approximately ${mse**0.5:,.2f} away from the actual sale price")
print(f'Mean Absolute Error: {mae:,.2f}')
print(f" - MAE:  on average, the predictions made by the model are off by +/- ${mae:,.2f} from the actual")
```

```
-- Linear Regression Stats for Boston House Prices --
R-Square: 0.948
 - RSQUARE: approximately 94.8% of the variability in the sale prices can be explained by our model.
Root Mean Squared Error: 32,918.10
 - RMSE: on average, our predictions are approximately $32,918.10 away from the actual sale price
Mean Absolute Error: 23,491.86
 - MAE:  on average, the predictions made by the model are off by +/- $23,491.86 from the actual
```

**Conclusions:**

*-- Linear Regression Stats for Boston House Prices --*

- R-Square: 0.437 for Method 1
- R-Square: 0.436 for Method 2

**It seems there is not a huge difference between the methods in handling nulls for our analysis**

*-- Linear Regression Stats for Boston House Prices --*

- R-Square: 0.947 for Method 1
- R-Square: 0.948 for Method 2

Again, it seems there is not a huge difference between the methods in handling nulls for our analysis. We might suggest that keeping all the method 2 for handling nulls gives us a slightly more accurate results as the R-Squared values turn out to be higher by 0.01 percentage points, but the only downside is that the 0's that represent the nulls for our yr_remod and land_sf could cause some skew issues. In conclusion, we will say that either approach is fine, and both will reach the same results but for simplicity we will go with Method 1 (the subset) and work with correctly filled out data.

Also, we can highlight the differences in performance between the Models by going over simple statistics. In statistics for models, having a higher R-squared is great and having low error is also great. Model 2 has a higher R-squared than Model 1 and a lower error than Model 1. Model 2's Root Mean Squared Error is 32,918.10 and has Mean Absolute Error of 23,491.86, which explains on average how much the predictions are off from the actual assessed value for the property. While, Model 1 has Root Mean Squared Error of 107,324.52 and Mean Absolute Error of 76,967.08. The Root Mean Squared Difference is −74,406.42 (Model 2 MSE - Model 1 MSE) and Absolute Mean Difference is −53,475.22 (Model 2 MAE - Model 1 MAE), so this serves as greater evidence in why we chose Model 2 to be the better and more accurate Model.

## Top and bottom 10 record predictions:

An Overestimate is when the error is small, thus we use nsmallest(). An underestimate is when the error is big, thus we use nlargest(). The best prediction is when we take a look at the mean absolute error (MAE) which allows us to state on average how much the prediction made by model 2 is off by from the actual assessed value for the property. The code chunks and tables are down below:

**Overestimate Predictions of AV_TOTAL:**

```
[109] 1 # Overestimate
      2 boston_pred2.nsmallest(10, 'residual')
```

| | pid | zipcode | own_occ | land_sf | yr_built | yr_remod | living_area | num_floors | structure_class | r_bldg_styl | ... | zip | population | pop_density | median_income | city_state | av_total | remod_ind | property_age | pred | residual |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3739 | 38940 | 2130 | Y | 9000.00 | 1915 | 0 | 2594 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 657900.00 | no | 108 | 865683.06 | -207783.06 |
| 8897 | 92410 | 2131 | Y | 7280.00 | 1910 | 2007 | 2150 | 2 | R | CL | ... | 2131 | 29826 | 11505 | 66735 | Roslindale, MA | 351800.00 | yes | 113 | 520393.06 | -168593.06 |
| 11797 | 122470 | 2130 | Y | 5600.00 | 1900 | 0 | 2880 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 671200.00 | no | 123 | 838272.00 | -167072.00 |
| 6599 | 68540 | 2131 | Y | 5894.00 | 1908 | 1984 | 2043 | 2 | R | CL | ... | 2131 | 29826 | 11505 | 66735 | Roslindale, MA | 322100.00 | yes | 115 | 488623.16 | -166523.16 |
| 10243 | 106350 | 2130 | N | 6945.00 | 1890 | 0 | 2492 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 549800.00 | no | 133 | 715284.38 | -165484.38 |
| 12914 | 134090 | 2130 | Y | 5752.00 | 1910 | 2001 | 2067 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 521800.00 | yes | 113 | 687193.44 | -165393.44 |
| 510 | 5320 | 2131 | Y | 10246.00 | 1890 | 0 | 2630 | 2 | R | CL | ... | 2131 | 29826 | 11505 | 66735 | Roslindale, MA | 407800.00 | no | 133 | 572399.44 | -164599.44 |
| 10240 | 106320 | 2130 | Y | 2800.00 | 1945 | 1989 | 1572 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 313100.00 | yes | 78 | 470196.78 | -157096.78 |
| 4196 | 43710 | 2124 | Y | 4923.00 | 1960 | 2004 | 1176 | 2 | R | CL | ... | 2124 | 47783 | 15913 | 48841 | Dorchester Center, MA | 247800.00 | yes | 63 | 404759.97 | -156959.97 |
| 7237 | 75190 | 2130 | Y | 12960.00 | 1910 | 2004 | 2709 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 701400.00 | yes | 113 | 858199.50 | -156799.50 |

10 rows × 37 columns

**Underestimate Predictions of AV_TOTAL:**

```
1 # Underestimate
2 boston_pred2.nlargest(10,'residual')
```

| | pid | zipcode | own_occ | land_sf | yr_built | yr_remod | living_area | num_floors | structure_class | r_bldg_styl | ... | zip | population | pop_density | median_income | city_state | av_total | remod_ind | property_age | pred | residual |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10855 | 112750 | 2130 | Y | 3645.00 | 1960 | 0 | 1421 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 767500.00 | no | 63 | 542955.31 | 224544.69 |
| 626 | 6530 | 2130 | Y | 4377.00 | 1950 | 0 | 1584 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 732300.00 | no | 73 | 519895.88 | 212404.12 |
| 7560 | 78520 | 2130 | N | 6625.00 | 1931 | 0 | 1972 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 1011700.00 | no | 92 | 826599.88 | 185100.12 |
| 4975 | 51800 | 2130 | Y | 8731.00 | 1927 | 0 | 1770 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 944600.00 | no | 96 | 760980.69 | 183619.31 |
| 5818 | 60510 | 2130 | Y | 5966.00 | 1955 | 2015 | 1833 | 1 | R | CP | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 1062620.00 | yes | 68 | 883951.81 | 178668.19 |
| 3608 | 37590 | 2130 | Y | 5303.00 | 1965 | 0 | 1387 | 1 | R | CP | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 809300.00 | no | 58 | 642110.19 | 167189.81 |
| 1522 | 15830 | 2130 | N | 5888.00 | 1964 | 0 | 1352 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 755500.00 | no | 59 | 595222.00 | 160278.00 |
| 2525 | 26290 | 2130 | Y | 4590.00 | 1983 | 1999 | 1184 | 2 | R | SD | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 777500.00 | yes | 40 | 617846.38 | 159653.62 |
| 12523 | 130050 | 2130 | Y | 6960.00 | 1935 | 2013 | 2209 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 1060314.00 | yes | 88 | 902510.06 | 157803.94 |
| 13343 | 138490 | 2130 | Y | 4235.00 | 1910 | 2002 | 2124 | 2 | R | CL | ... | 2130 | 35401 | 10618 | 75730 | Jamaica Plain, MA | 978400.00 | yes | 113 | 821302.31 | 157097.69 |

10 rows × 37 columns

**Best Predictions of AV_TOTAL:**

```
1 # Best Prediction
2 boston_pred2['abs_residual'] = boston_pred2['residual'].abs().round(3)
3 boston_pred2.nsmallest(10, 'abs_residual')
```

| | pid | zipcode | own_occ | land_sf | yr_built | yr_remod | living_area | num_floors | structure_class | r_bldg_styl | ... | population | pop_density | median_income | city_state | av_total | remod_ind | property_age | pred | residual | abs_residual |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12666 | 131520 | 2136 | Y | 6000.00 | 1911 | 2013 | 1182 | 2 | R | CL | ... | 28488 | 6207 | 58890 | Hyde Park, MA | 332042.13 | yes | 112 | 332043.66 | -1.52 | 1.52 |
| 2051 | 21340 | 2131 | Y | 5163.00 | 1920 | 0 | 1068 | 1 | R | BW | ... | 29826 | 11505 | 66735 | Roslindale, MA | 359000.00 | no | 103 | 359010.53 | -10.53 | 10.53 |
| 46 | 480 | 2132 | Y | 5987.00 | 1960 | 0 | 972 | 1 | R | RN | ... | 36314 | 13251 | 75446 | Cambridge, MA | 344800.00 | no | 63 | 344813.38 | -13.38 | 13.38 |
| 1970 | 20530 | 2124 | Y | 2342.00 | 1890 | 1970 | 1478 | 2 | R | CL | ... | 47783 | 15913 | 48841 | Dorchester Center, MA | 285886.71 | yes | 133 | 285902.59 | -15.89 | 15.89 |
| 3698 | 38520 | 2124 | Y | 6420.00 | 1880 | 2006 | 2161 | 2 | R | VT | ... | 47783 | 15913 | 48841 | Dorchester Center, MA | 658200.00 | yes | 143 | 658181.50 | 18.50 | 18.50 |
| 8139 | 84520 | 2132 | Y | 5481.00 | 1930 | 0 | 1450 | 2 | R | CL | ... | 36314 | 13251 | 75446 | Cambridge, MA | 454388.78 | no | 93 | 454407.91 | -19.13 | 19.13 |
| 11214 | 116420 | 2131 | Y | 5464.00 | 1880 | 0 | 1799 | 2 | R | CL | ... | 29826 | 11505 | 66735 | Roslindale, MA | 451563.18 | no | 143 | 451585.12 | -21.94 | 21.94 |
| 11101 | 115280 | 2136 | Y | 5725.00 | 1884 | 0 | 1344 | 1 | R | CL | ... | 28488 | 6207 | 58890 | Hyde Park, MA | 314484.71 | no | 139 | 314506.97 | -22.25 | 22.25 |
| 1887 | 19640 | 2136 | Y | 5488.00 | 1870 | 2005 | 1046 | 2 | R | CL | ... | 28488 | 6207 | 58890 | Hyde Park, MA | 293376.69 | yes | 153 | 293345.38 | 31.31 | 31.31 |
| 9901 | 102830 | 2131 | Y | 3600.00 | 1910 | 0 | 1179 | 2 | R | CL | ... | 29826 | 11505 | 66735 | Roslindale, MA | 352856.81 | no | 113 | 352821.41 | 35.41 | 35.41 |

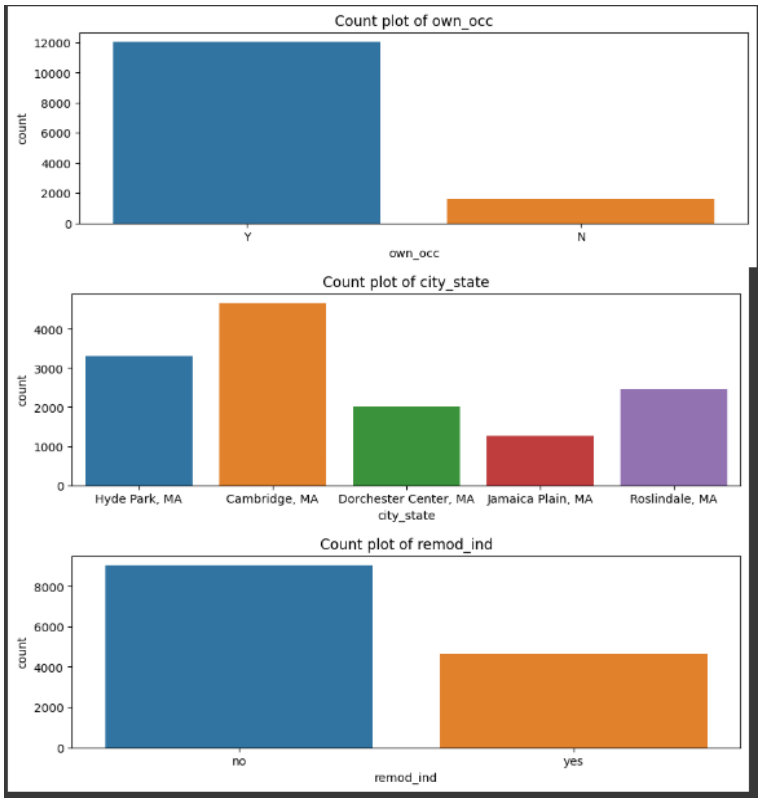
10 rows × 38 columns

## Categorical Analysis: what categorical variables are interesting and likely important.

We only selected three categorical variables as we believed they are the most important and are very useful for our analysis, especially when we have to combine these variables with AV_TOTAL. The three variables we chose were “own_occ”, “city_state” and “remod_ind”. In our analysis, we used descriptive statistics and count plots to understand frequency of these categorical variables and the unique values for these variables.

### Count Plots:

Many homes are owner occupied, were not remodeled, and Cambridge, Hyde Park and Roslindale are the top three most popular city states.




**Descriptive Statistics:**

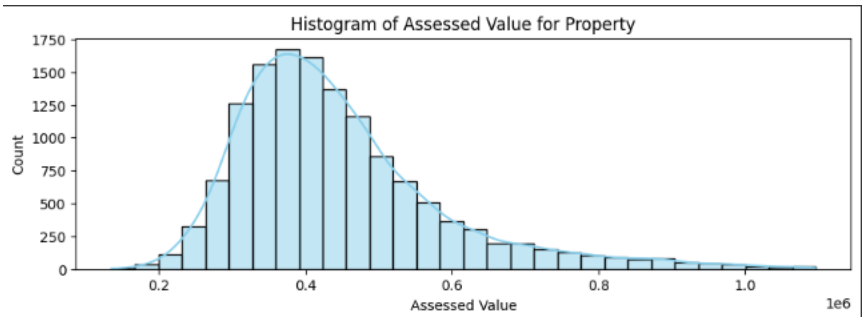
| | count | unique | top | freq |
|---|---|---|---|---|
| own_occ | 13712 | 2 | Y | 12068 |
| city_state | 13712 | 5 | Cambridge, MA | 4673 |
| remod_ind | 13712 | 2 | no | 9057 |

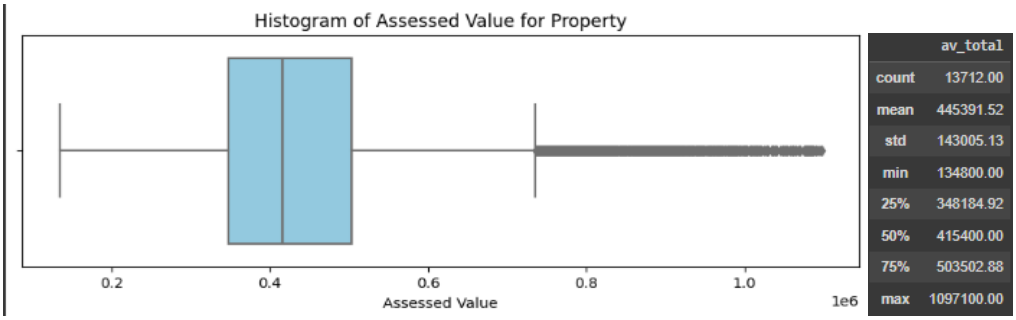
Confirms our findings that Cambridge is the most popular city state for Boston Homes.

## Numeric Analysis, Descriptive Statistics, and Histograms:

The numeric variables that we selected were based on correlation results and impact on assessed value for property. The variables we selected were living area, number of total rooms, median income, number of full bathrooms, number of fireplaces, land area in square feet, number of bedrooms, number of floors, year the home was last remodeled, number of half bathrooms, property age, population density, and zipcode.

**Target Variable – Assessed Value for Property (AV_TOTAL)**

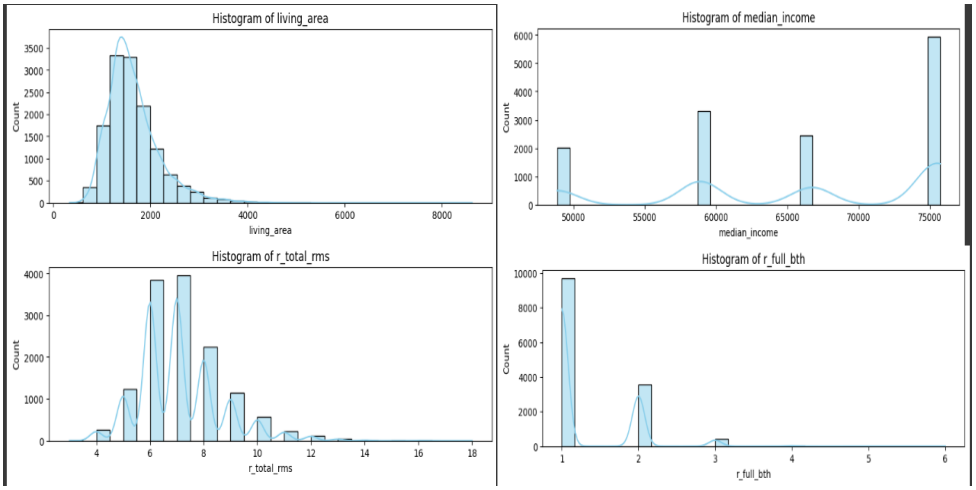


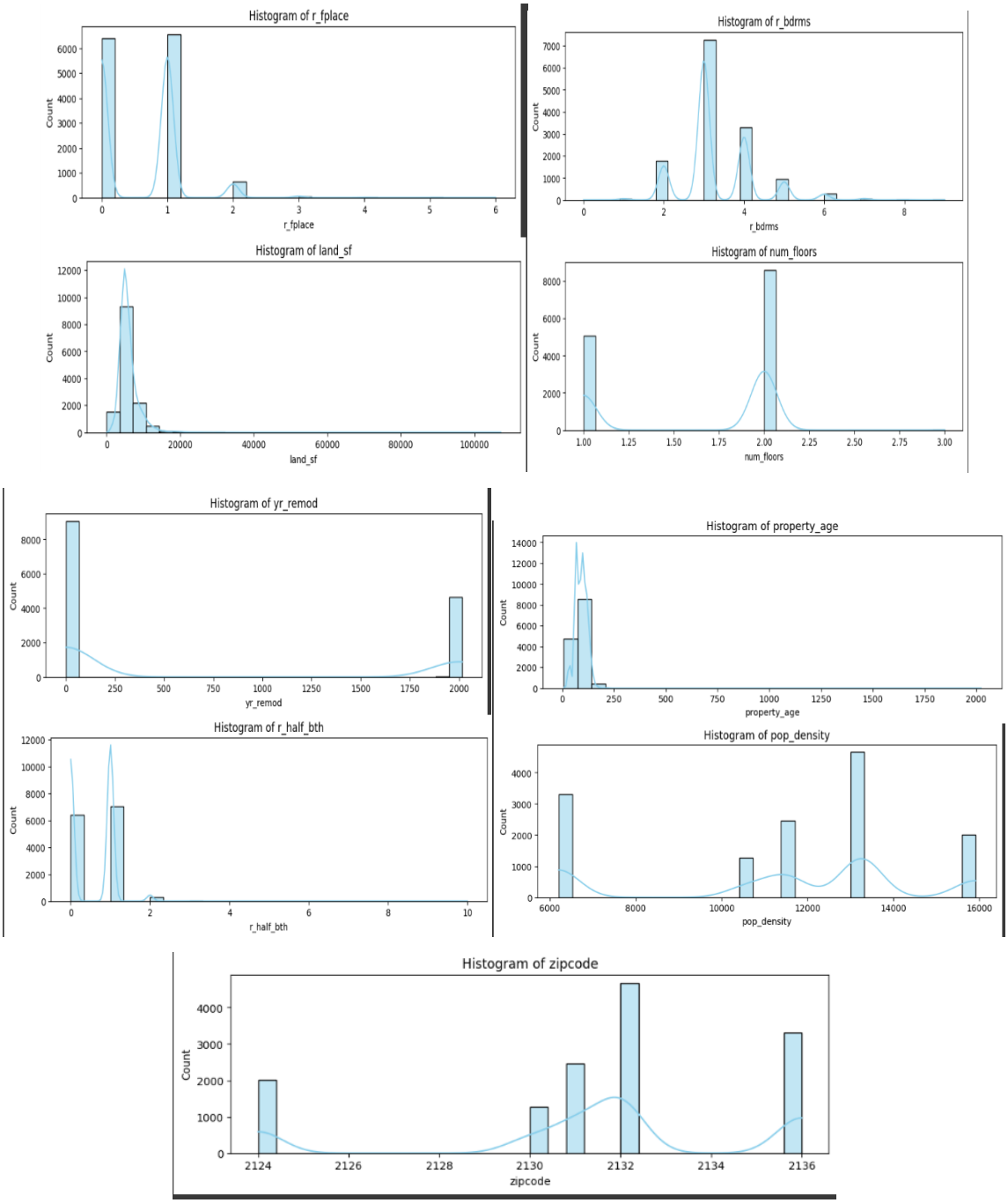




| | av_total |
|---|---|
| count | 13712.00 |
| mean | 445391.52 |
| std | 143005.13 |
| min | 134800.00 |
| 25% | 348184.92 |
| 50% | 415400.00 |
| 75% | 503502.88 |
| max | 1097100.00 |

We have a right skewed distribution that is unimodal. The median would be the best measure of center as it is not influenced by outliers/skew. However, it does seem that the majority of the Boston Homes are in the $300,000 to $500,000 price range. The box plot helps with our analysis as we can claim with evidence that there is definitely a right skewed distribution as the majority of home prices are influenced by large outliers. There are tons of outliers in our data, meaning there are a lot of expensive Boston Homes in our data. This table gives us a simple breakdown of our boxplot by giving us numerical values. Now we can accurately state 50% of Boston homes' assessed value for property ranges from $348,184.92 to $503,502.88 The mean or average Boston home price is $445,391.52. We can also state that the cheapest Boston house is $134,800.00 and the most expensive is $1,097,100.00 (excluding potential outliers). Now, we can start to investigate more the numeric features and categorical features that influence price, this is going to be the heart of our analysis and will help with our client's understanding.

## Histograms:






We can see there is a common distribution shape for many of our numeric features, which is that they are unimodal and most likely skewed (right or left, depends on what numeric feature we are referring to). Zip code and Pop Density are an exception for the not following the common distribution shape as they act more like a categorical variable/frequency count.

**Descriptive Statistics:**

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| living_area | 13712.00 | 1645.58 | 542.49 | 332.00 | 1295.00 | 1540.00 | 1884.00 | 8623.00 |
| r_total_rms | 13712.00 | 7.09 | 1.56 | 3.00 | 6.00 | 7.00 | 8.00 | 18.00 |
| median_income | 13712.00 | 66024.84 | 9724.93 | 48841.00 | 58890.00 | 66735.00 | 75446.00 | 75730.00 |
| r_full_bth | 13712.00 | 1.33 | 0.55 | 1.00 | 1.00 | 1.00 | 2.00 | 6.00 |
| r_fplace | 13712.00 | 0.59 | 0.62 | 0.00 | 0.00 | 1.00 | 1.00 | 6.00 |
| land_sf | 13712.00 | 5908.39 | 2883.10 | 0.00 | 4302.75 | 5268.50 | 6720.00 | 107158.00 |
| r_bdrms | 13712.00 | 3.33 | 0.92 | 0.00 | 3.00 | 3.00 | 4.00 | 9.00 |
| num_floors | 13712.00 | 1.63 | 0.49 | 1.00 | 1.00 | 2.00 | 2.00 | 3.00 |
| yr_remod | 13712.00 | 679.44 | 947.80 | 0.00 | 0.00 | 0.00 | 1997.00 | 2016.00 |
| r_half_bth | 13712.00 | 0.56 | 0.54 | 0.00 | 0.00 | 1.00 | 1.00 | 10.00 |
| property_age | 13712.00 | 91.32 | 32.08 | 7.00 | 70.00 | 93.00 | 113.00 | 2023.00 |
| pop_density | 13712.00 | 11386.48 | 3272.51 | 6207.00 | 10618.00 | 11505.00 | 13251.00 | 15913.00 |
| zipcode | 13712.00 | 2131.43 | 3.67 | 2124.00 | 2131.00 | 2132.00 | 2132.00 | 2136.00 |

To really understand what the shapes and what is going with the numeric features, we reference our descriptive statistics as that also serves as a far simpler version of a boxplot but also gives us accurate numbers when we need to discuss our findings.

Living Area: The average Boston Home has 1,645.58 sq ft, 50% of Boston homes have 1,540.00 sq ft, the min is 332.00 sq ft, the max is 8,623.00 sq ft, the spread in one standard deviation away from the mean is 542.49 sq ft.

Total Number of Rooms: The average Boston home has 7.09 rooms, 50% of Boston homes have 7 rooms, the min is 3 rooms, the max is 18 rooms, the spread in one standard deviation away from the mean is 1.56 rooms.

Median Income: The mean median income of the residence is $66,024.84, 50% of Boston Homes Median Income is $66,735.00, the min Median Income is $48,841.00, the max Median Income is $75,730.00, the spread of Median Income in one standard deviation away from the mean is $9,724.93.

Total Number of Full Baths: The average Boston Home has 1.33 full bathrooms, 50% of Boston homes have 1 full bathroom, the min is 1 full bathrooms, the max is 6 full bathrooms, the spread in one standard deviation away from the mean is 0.55 full bathrooms.

Total Number of Fireplaces: The average Boston Home has 0.59 fireplaces, 50% of Boston homes have 1 fireplace, the min is 0 fireplaces, the max is 6 fireplaces, the spread in one standard deviation away from the mean is 0.62 fireplaces.

Parcel's Land Area in Square Feet: The average Boston Home legal area has 5,908.39 sq ft, 50% of Boston homes have 5,268.50 sq ft, the min has 0.00 sq ft, the max has 107,158.00 sq ft, the spread in one standard deviation away from the mean is 2,883.10 sq ft.

Total Number of Bedrooms: The average Boston Home has 3.33 bedrooms, 50% of Boston homes have 3 bedrooms, the min has 0 bedrooms, the max has 9 bedrooms, the spread in one standard deviation away from the mean is 0.92 bedrooms.

Total Number of Floors: The average Boston Home has 1.63 floors, 50% of Boston homes have 2 floors, the min has 1 floor, the max has 3 floors, the spread in one standard deviation away from the mean is 0.49 floors.

Year Property was Last Remodeled: 75% of Boston Homes were remodeled in 1997 and the most recent year a Boston Home was last remodeled was in 2016. Many of the homes were not remodeled, hence the 0.

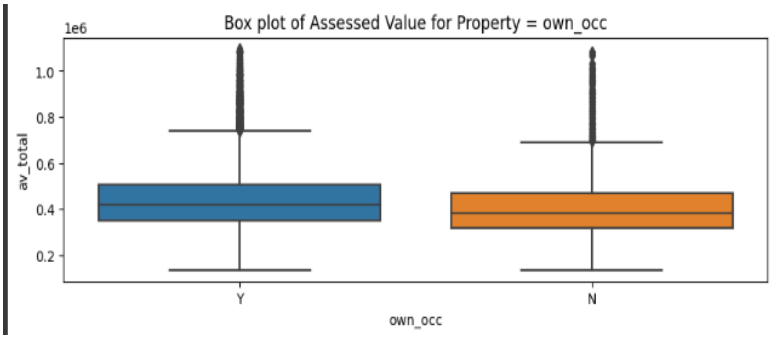
Number of Half Bathrooms: The average Boston Home has 0.56 full bathrooms, 50% of Boston homes have 1 half bathrooms, the min has 0 half bathrooms, the max has 10 half bathrooms, the spread in one standard deviation away from the mean is 0.54 half bathrooms.

Property's Age (2023 - Year Built): The average Boston Home Property Age is 91.32 years old, 50% of Boston homes' Property Age is 93 years old , the min or youngest property age is 7 years old , the spread in one standard deviation away from the mean is 32.08 years.

People Per Square Mile: The average Population Density in Boston Homes is 11,386.48 people, 50% of Boston homes have a Population Density of 11,505 people, the min Population Density is 6,207 people, the max Population Density is 15,913 people, the spread in one standard deviation away from the mean is 3,272.51 people.

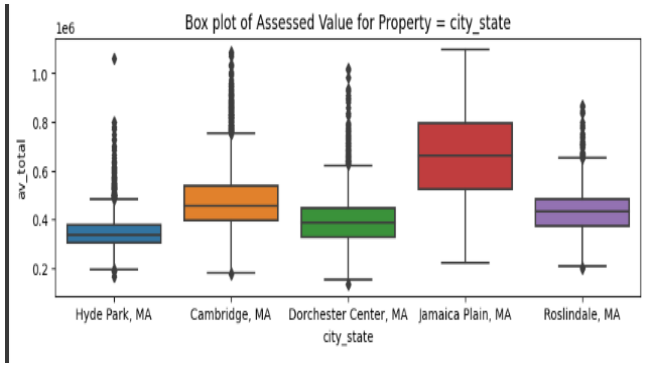
## Categorical to numeric analysis – what categorical variable combined with AV_TOTAL are likely to be useful and why? (Assess the client's beliefs)

### Box plots and Tables for Categorical Variables:




| own_occ | N | Y |
|---|---|---|
| count | 1644.00 | 12068.00 |
| mean | 417320.19 | 449215.62 |
| std | 143982.92 | 142449.89 |
| min | 136100.00 | 134800.00 |
| 25% | 320975.00 | 352193.96 |
| 50% | 384750.00 | 419500.00 |
| 75% | 469565.48 | 508001.58 |
| max | 1083537.00 | 1097100.00 |

Box Plot "Assessed Value for Property = Own_OCC", the difference in median assessed value for owner occupied homes is not that large as the homes that are owned occupied are priced at $419,500.00 vs homes that are not owner occupied are priced at $384,750.00. The difference is only $34,750 dollars which is nothing drastically different or is not a big enough difference to put emphasis on. Also, there are also a lot of outliers for owner occupied homes and homes that are not owner occupied as seen in the box plot.




| city_state | Cambridge, MA | Dorchester Center, MA | Hyde Park, MA | Jamaica Plain, MA | Roslindale, MA |
|---|---|---|---|---|---|
| count | 4673.00 | 2008.00 | 3305.00 | 1269.00 | 2457.00 |
| mean | 480877.57 | 398657.00 | 346327.12 | 667476.08 | 434646.35 |
| std | 123423.67 | 110461.41 | 69991.86 | 183192.34 | 91183.89 |
| min | 177836.69 | 134800.00 | 168300.00 | 220700.00 | 201400.00 |
| 25% | 395833.30 | 329903.14 | 303749.31 | 523560.70 | 373254.62 |
| 50% | 458079.45 | 387089.37 | 337088.47 | 659900.00 | 432396.82 |
| 75% | 538723.53 | 448625.00 | 376400.00 | 797280.32 | 485000.00 |
| max | 1086300.00 | 1019300.00 | 1060100.00 | 1097100.00 | 867600.00 |

Box Plot "Assessed Value for Property = City_State", we can determine the most expensive city states based on a measure like the median. From most expensive city state to least expensive city state, we have Jamaica Plain in 1st at $659,900.00, Cambridge in 2nd at $458,079.45, Roslindale in 3rd at $432,396.82, Dorchester in 4th at $387,089.37, and Hyde Park in 5th at $337,088.47. An interesting find is that Jamaica Plain homes are the highest priced homes and there are no outliers meaning that is consistently the most expensive city state.


| remod_ind | no | yes |
|---|---|---|
| count | 9057.00 | 4655.00 |
| mean | 417059.39 | 500515.90 |
| std | 124431.83 | 159746.04 |
| min | 134800.00 | 192500.00 |
| 25% | 334700.00 | 387779.45 |
| 50% | 394200.00 | 467900.00 |
| 75% | 469171.59 | 573350.00 |
| max | 1095200.00 | 1097100.00 |

Box Plot "Assessed Value for Property = Remod_Ind", there is a noticeably large difference in value of homes for whether they were remodeled or not, 50% the ones that were remodeled are at a high value of $467,900.00 and the 50% of homes that were not remodeled are at a lower value of$ 394,200.00. Additionally, there is an interesting story being told by the outliers for Whether a house was remodeled or not because there are a ton of outliers. For the homes that were not remodeled we make the assume the possible reason for the high value is that these homes are either new or represent old-fashioned Boston and hence the price is shot up. Also, our client is right in the sense that remodeled homes do have a higher home value, and in later plots we will explore whether the year or remodeling has an impact on home value. In other words, has a home that has been remodeled recently have a higher home value?

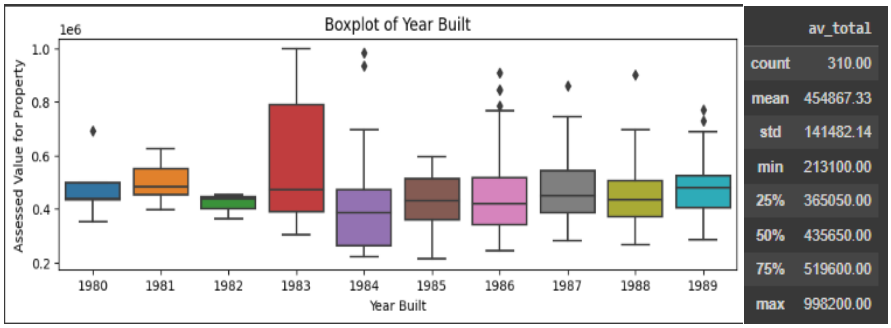
## Big Table:

| city_state | own_occ | remod_ind | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|---|
| Cambridge, MA | N | no | 300.00 | 429652.47 | 109177.30 | 177836.69 | 365850.00 | 414850.00 | 475474.17 | 906000.00 |
| | | yes | 142.00 | 497616.56 | 132814.37 | 252300.00 | 396375.00 | 468843.01 | 561324.81 | 945463.30 |
| | Y | no | 2702.00 | 457744.32 | 110285.62 | 182100.00 | 382025.00 | 437100.00 | 507843.70 | 1035900.00 |
| | | yes | 1529.00 | 530254.03 | 131111.16 | 228800.00 | 433436.58 | 508000.00 | 594700.00 | 1086300.00 |
| Dorchester Center, MA | N | no | 224.00 | 350909.64 | 89095.98 | 136100.00 | 284344.72 | 340722.08 | 401900.39 | 679100.00 |
| | | yes | 102.00 | 387466.86 | 84166.30 | 227900.00 | 331051.90 | 371579.92 | 420380.32 | 717400.00 |
| | Y | no | 1086.00 | 381080.53 | 95941.78 | 134800.00 | 319550.00 | 377102.37 | 430025.00 | 881200.00 |
| | | yes | 596.00 | 450544.32 | 126706.10 | 201200.00 | 370925.00 | 429150.00 | 508421.92 | 1019300.00 |
| Hyde Park, MA | N | no | 343.00 | 328629.29 | 59393.09 | 185372.17 | 295500.00 | 319100.00 | 353900.00 | 618400.00 |
| | | yes | 92.00 | 365367.87 | 106717.65 | 192500.00 | 311838.91 | 348482.84 | 397851.65 | 801300.00 |
| | Y | no | 2098.00 | 338344.53 | 63080.07 | 168300.00 | 300194.22 | 332157.45 | 366800.00 | 799600.00 |
| | | yes | 772.00 | 373614.78 | 78487.39 | 229500.00 | 324002.51 | 360266.89 | 404067.27 | 1060100.00 |
| Jamaica Plain, MA | N | no | 79.00 | 638814.98 | 197523.93 | 220700.00 | 468703.71 | 624266.81 | 760850.00 | 1078200.00 |
| | | yes | 83.00 | 668456.50 | 202012.42 | 358805.46 | 536184.67 | 630400.00 | 828849.12 | 1083537.00 |
| | Y | no | 572.00 | 623920.39 | 172685.81 | 258600.00 | 481572.91 | 609050.00 | 749525.00 | 1095200.00 |
| | | yes | 535.00 | 718124.13 | 176386.13 | 224000.00 | 572900.00 | 722200.00 | 842368.96 | 1097100.00 |
| Roslindale, MA | N | no | 199.00 | 398448.88 | 86376.12 | 210300.00 | 341653.62 | 397736.25 | 451950.00 | 654244.99 |
| | | yes | 80.00 | 460230.37 | 117982.76 | 236300.00 | 371285.92 | 463350.00 | 522575.00 | 867600.00 |
| | Y | no | 1454.00 | 419478.66 | 82619.08 | 201400.00 | 364949.54 | 419400.00 | 468592.84 | 844000.00 |
| | | yes | 724.00 | 472229.76 | 93014.51 | 239100.00 | 412250.00 | 466569.80 | 527395.71 | 863691.06 |

In this table, we have grouped by city_state, own_occ, and remod_ind, and selected the av_total column to get descriptive statistics. Here we can apply advanced analysis on the impact of the categorical variables on home value. This table serves as greater evidence for our conclusions of the ranking of our top 5 most expensive city states, owner occupied homes have higher

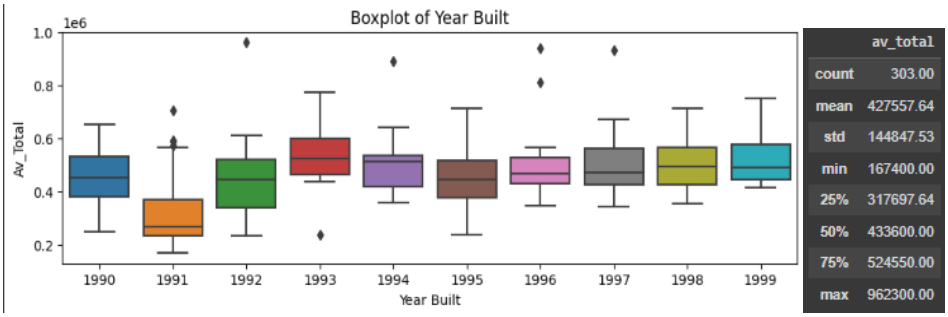
prices, and homes that were remodeled have a higher price. For example, there difference in average home value between the three categorical variables in Jamaica Plain is drastic, the average Boston Home Value in Jamaica Plain that is owner occupied and was remodeled is priced at a high of $718,124.13 while the average Boston Home in Jamaica Plain that is not owner occupied and has not been remodeled is priced at a low of $638,814.98. The difference is $79,309.05 dollars, which is close to 100,000 dollars. This table just serves as extra support for our findings, and this analysis can be done on other city states, and comparing between owner occupied homes and whether homes were remodeled or not.

### Box plots and Tables for Year Built:



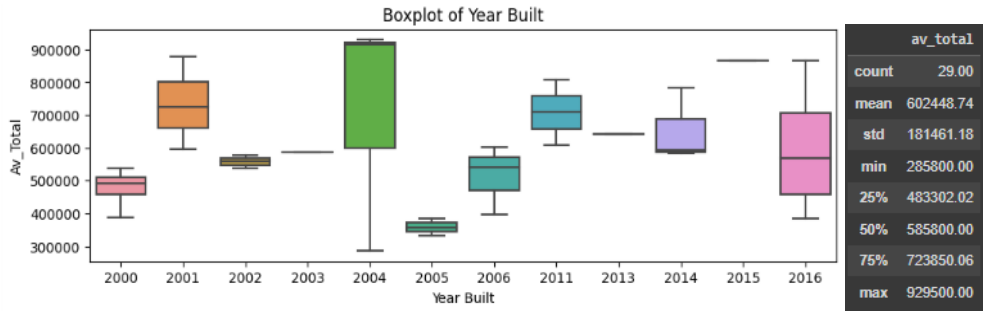

| | av_total |
|---|---|
| count | 310.00 |
| mean | 454867.33 |
| std | 141482.14 |
| min | 213100.00 |
| 25% | 365050.00 |
| 50% | 435650.00 |
| 75% | 519600.00 |
| max | 998200.00 |

Homes built in the 1980s seem to have increasing variation as the years go on, especially after 1982. 1983 stands out for the year to build homes and have super high property values. The size of the box plots gets larger, and we start to see some outliers, but to describe the home values is that the average home value was $454,867.33 and 50% of the Boston homes were $435,650.00.




| | av_total |
|---|---|
| count | 303.00 |
| mean | 427557.64 |
| std | 144847.53 |
| min | 167400.00 |
| 25% | 317697.64 |
| 50% | 433600.00 |
| 75% | 524550.00 |
| max | 962300.00 |

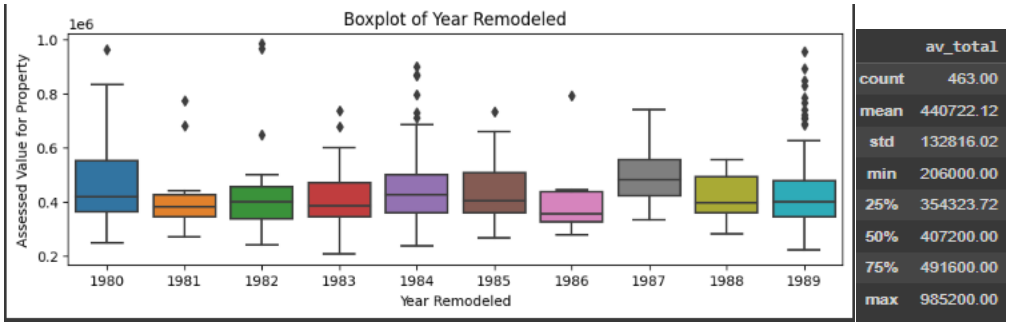
Homes built in the 1990s price ranges seem to fluctuate from 1990 to 1993 and then from 1994 to 1999 the price range seems to be relatively consistent referring to the IQR. The size of the box plots seems relatively similar and there are only a few outliers we start to see some outliers, but to describe the home values is that the average home value was $427,557.64 and 50% of the Boston homes were $433,600.00. It does seem that homes in the 1990s compared to the 1980s are slightly cheaper, but we would rather say that the homes prices in the 90s are very consistent in comparison to 1980s.




| | av_total |
|---|---|
| count | 29.00 |
| mean | 602448.74 |
| std | 181461.18 |
| min | 285800.00 |
| 25% | 483302.02 |
| 50% | 585800.00 |
| 75% | 723850.06 |
| max | 929500.00 |

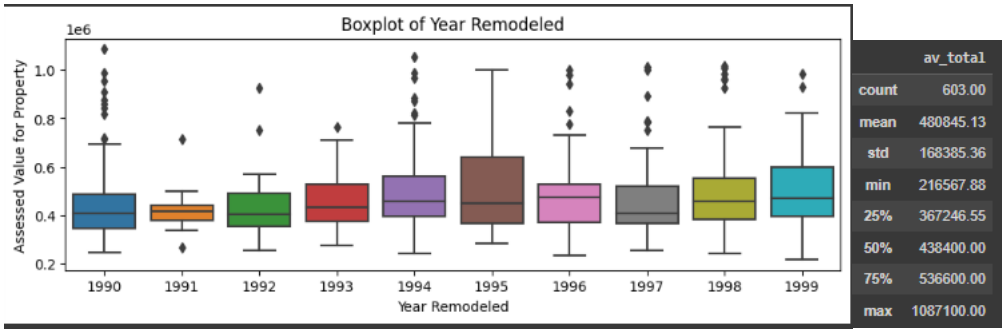
Homes built in the 2000s have lots of fluctuating variation as the years go on. We can see the increase of home values and homes built from 2005 to 2006 due to the rise of the financial crisis of 2007-08. 2013 and 2015 have expensive home prices. One thing to say about 2000s is generally the homes values did increase when the homes became recently built and home values rise due to inflation. To double check our claims, the average home value was $602,448.74 and 50% of the Boston homes were $585,800.00. Also, the Mean Home Value $602,448.74 (2000s) > $454,867.33 (1980s) > $427,557.64 (1990s) and the 50% of Home values $585,800.00 (2000s) > $435,650.00 (1980s) > $433,600.00 (1990s).

### Box plots and Tables for Year Remodeled:




| | av_total |
|---|---|
| count | 463.00 |
| mean | 440722.12 |
| std | 132816.02 |
| min | 206000.00 |
| 25% | 354323.72 |
| 50% | 407200.00 |
| 75% | 491600.00 |
| max | 985200.00 |

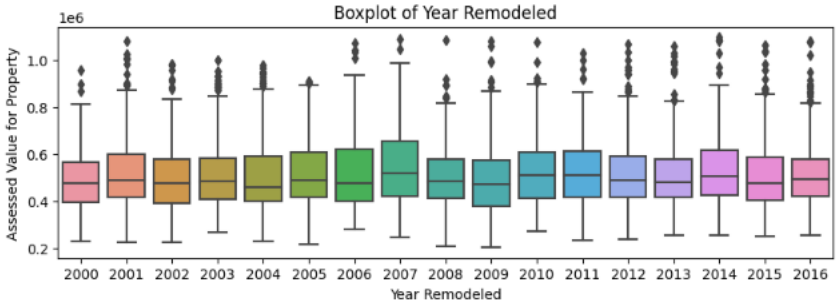
Homes Remodeled in the 1980s have relatively consistent Medians and the IQR seems to be captured by the next year as the years go on. We can point out outliers for some years like 1980, 1984, and 1989 which will influence the home value when calculating the mean. To share insights on this era, the mean homes values for a remodeled house is $440,722.12 and the 50% of remodeled homes are priced at $407,200.00.




| | av_total |
|---|---|
| count | 603.00 |
| mean | 480845.13 |
| std | 168385.36 |
| min | 216567.88 |
| 25% | 367246.55 |
| 50% | 438400.00 |
| 75% | 536600.00 |
| max | 1087100.00 |

Homes Remodeled in the 1990s have relatively consistent Medians and the IQR seems to be captured by the next year as the years go on. We see there is an increasing home value trend as the years go on, for example if we look at the years 1993 to

1995. After 1995, we get smaller IQRs but lots of outliers that'll influence the mean/average to increase. To share insights on this era, the mean homes value for a remodeled house is $480,845.13 and 50% of remodeled homes are priced at $438,400.00. Comparing Means and Median, it appears that 1990s has higher home values than 1980s: Mean $480,845.13 (1990) > $440,722.12 (1980s) and Median $438,400.00 (1990s) > $407,200.00 (1980s)




| | av_total |
|---|---|
| count | 3233.00 |
| mean | 519026.55 |
| std | 158001.79 |
| min | 201200.00 |
| 25% | 407900.00 |
| 50% | 487700.00 |
| 75% | 593000.00 |
| max | 1097100.00 |

Homes Remodeled in the 2000s have very consistent Medians and the IQR seems to be captured by the next year as the years go on. We can see that all the years have outliers, and the maxes of the box plots seem to change year after year. To share insights on this era, the mean homes value for a remodeled house is $519,026.55 and 50% of remodeled homes are priced at $593,300.00. Comparing Means and Median, it appears that 2000s has highest home values than 1980s and 1990s: Mean $519,026.55 (2000s) > $480,845.13 (1990) > $440,722.12 (1980s) and Median $593,300.00 (2000s) > $438,400.00 (1990s) > $407,200.00 (1980s).

## Pivot Tables:

| | city_state | Cambridge, MA | Dorchester Center, MA | Hyde Park, MA | Jamaica Plain, MA | Roslindale, MA |
|---|---|---|---|---|---|---|
| | living_area | | | | | |
| mean | 332 | 0.00 | 0.00 | 248400.00 | 0.00 | 0.00 |
| | 403 | 0.00 | 0.00 | 207008.20 | 0.00 | 0.00 |
| | 426 | 179022.80 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 440 | 0.00 | 0.00 | 192500.00 | 0.00 | 0.00 |
| | 517 | 0.00 | 0.00 | 0.00 | 0.00 | 236300.00 |
| | ... | ... | ... | ... | ... | ... |
| | 5017 | 0.00 | 1014600.00 | 0.00 | 0.00 | 0.00 |
| | 5156 | 0.00 | 0.00 | 0.00 | 743912.30 | 0.00 |
| | 5197 | 0.00 | 164700.00 | 0.00 | 0.00 | 0.00 |
| | 5239 | 0.00 | 980700.00 | 0.00 | 0.00 | 0.00 |
| | 8623 | 0.00 | 0.00 | 1060100.00 | 0.00 | 0.00 |

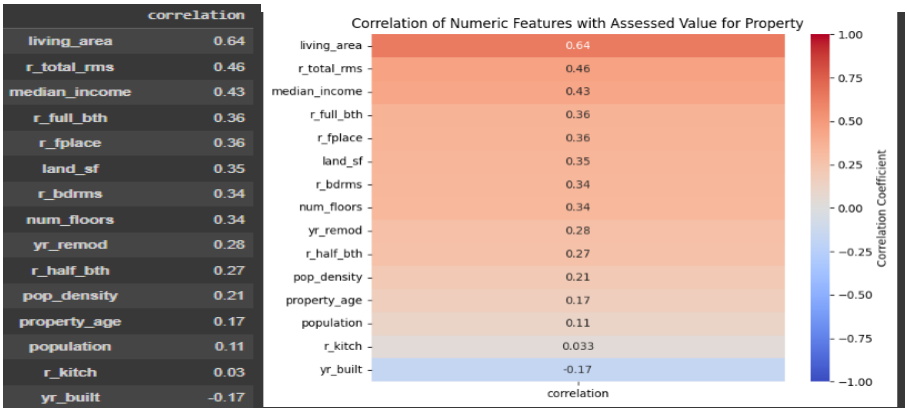
2322 rows × 5 columns

This pivot tale of indexing city states, and taking a look at the living area, and generating the home value average provides power insight in determining home value based on city state and living area. When we refer to our analysis on city state, we know that certain neighbors are more expensive than others, and with living area getting larger it can be assumed that home value increases. However, this table can show that the assumption might not always be true as we look at Hyde Park home values for a living area of 332 and 403, the smaller living area has a higher price.

## Numeric Relationships:

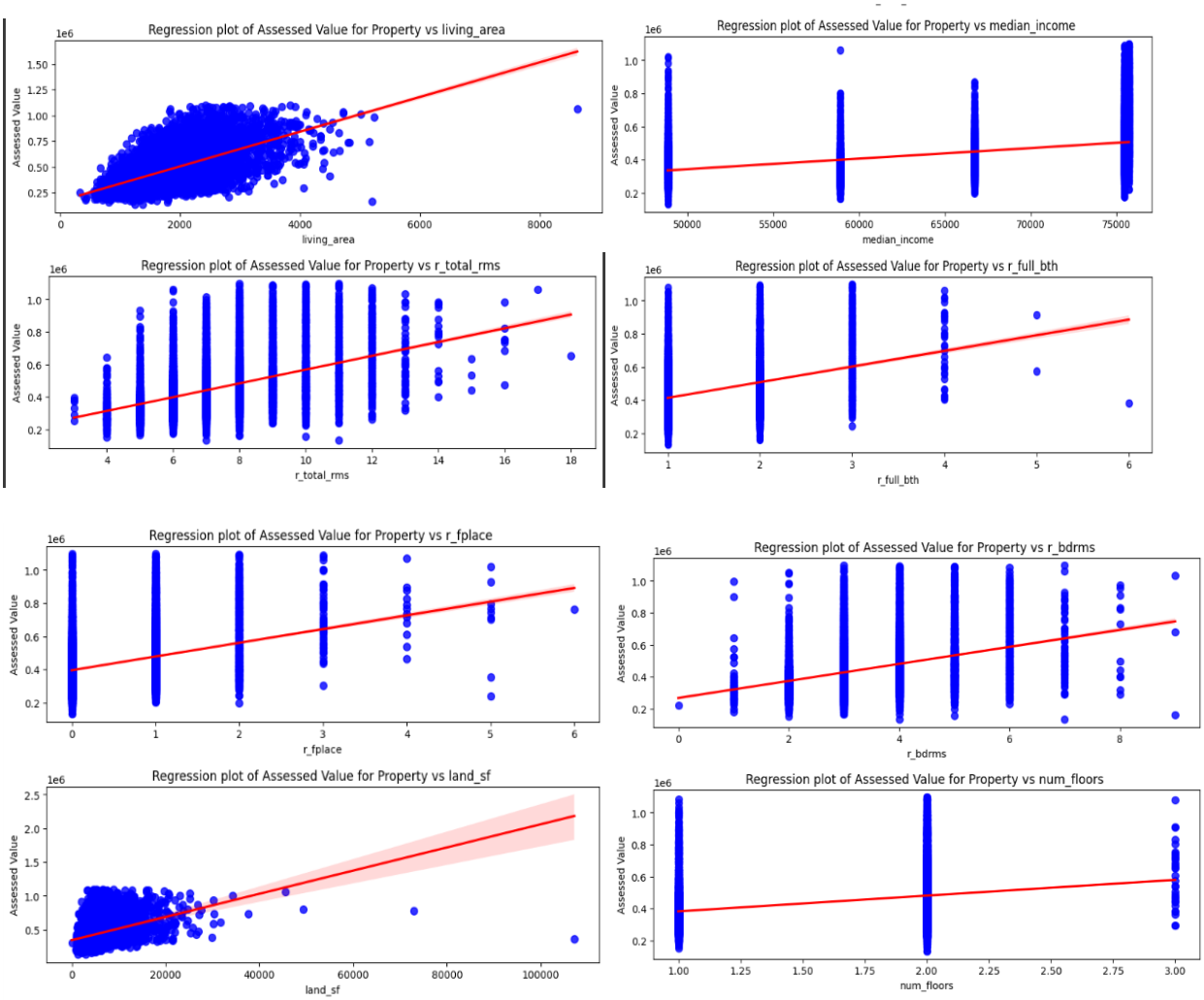
## Correlation:

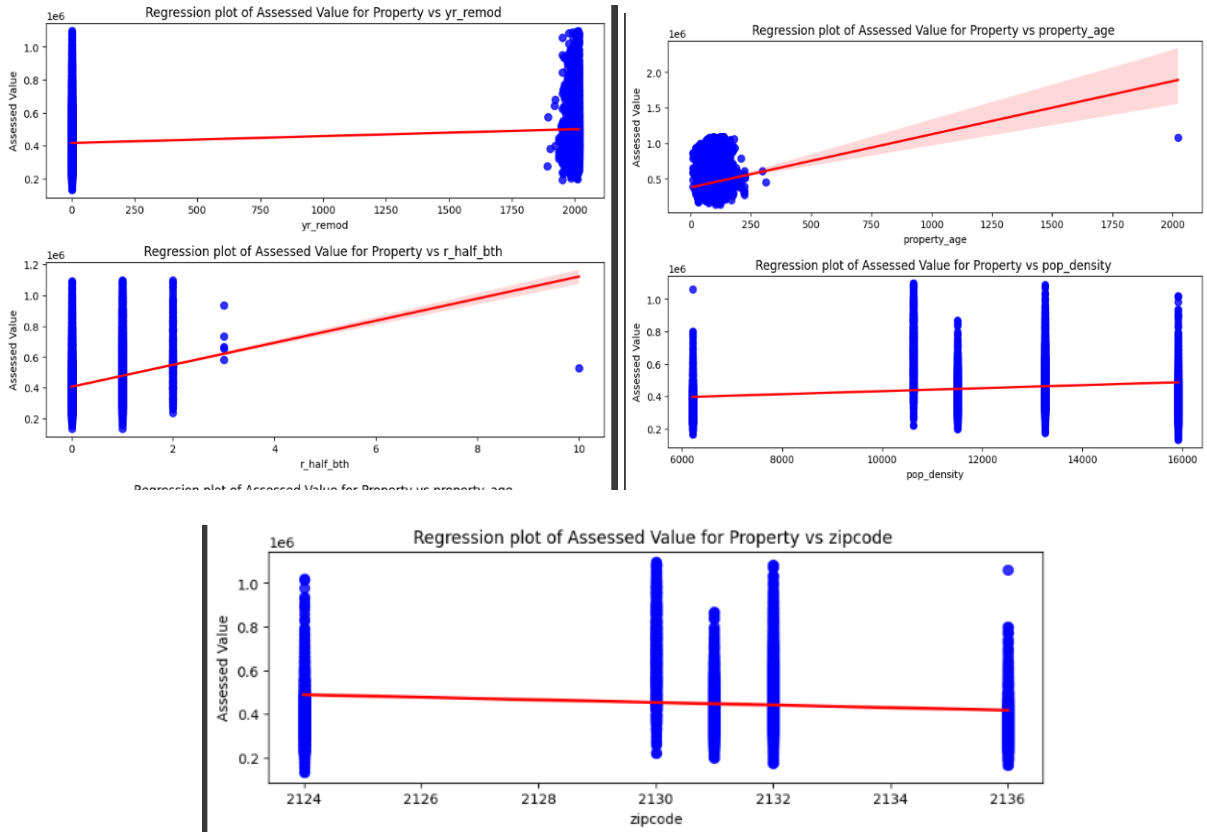
| | correlation |
|---|---|
| living_area | 0.64 |
| r_total_rms | 0.46 |
| median_income | 0.43 |
| r_full_bth | 0.36 |
| r_fplace | 0.36 |
| land_sf | 0.35 |
| r_bdrms | 0.34 |
| num_floors | 0.34 |
| yr_remod | 0.28 |
| r_half_bth | 0.27 |
| pop_density | 0.21 |
| property_age | 0.17 |
| population | 0.11 |
| r_kitch | 0.03 |
| yr_built | -0.17 |




We selected our numeric variables based on correlation values greater than 0.2 and made property age an exception.

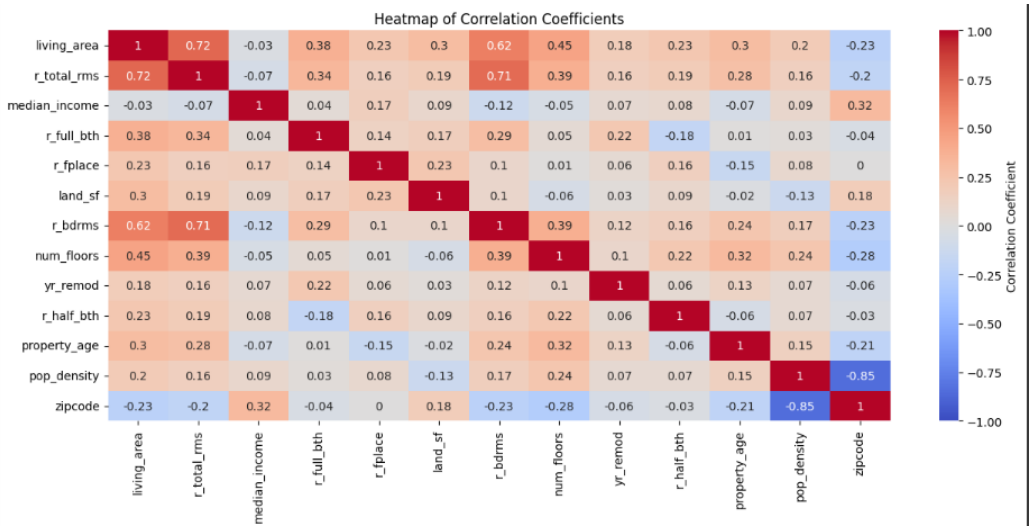
**Regression/Linear Plots:**






These graphs provide a visual of the correlation heat map and serve as a visual aid for us to see if it right to use a linear regression for the numerical variable against the target variable. Most of the plots show that there is a relatively weak positive relationship for these variables. In the future, a transformation like log, square root, quadratic or exponential could be helpful.

## Correlation Matrix:




Here, we are looking at what variables might have influence on another, which is called multicollinearity. This concept in statistics is a violation and something to be careful about when building models and assessing fit on models. Some variables that

are easily understood that have high correlations between each other are number of total rooms and living area, number of bedrooms and living area, and number of floors and living area. It's easily understood by factoring in that the living area space a home has the more likely the home is to have more space for bedroom, an additional floor, and rooms. These are examples of positive relationships.

### Scatter plots between correlated variables: