# Real Estate Prediction Model

By Hakeem Lawrence

Dataset link: https://www.kaggle.com/code/shreayan98c/boston-house-price-prediction/data

## Importing Modules

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from xgboost import XGBRegressor
from sklearn import metrics
```

## Importing the dataset

In [2]:
```
housing = pd.read_csv('housing.csv')
housing.head()
```

Out[2]:

| | 0.00632 | 18 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.09 | 1 | 296 | 15.3 | 396.9 | 4.98 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.02731 | 0.0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 | 396.90 | 9.14 | 21.6 |
| 1 | 0.02729 | 0.0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 | 392.83 | 4.03 | 34.7 |
| 2 | 0.03237 | 0.0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 | 394.63 | 2.94 | 33.4 |
| 3 | 0.06905 | 0.0 | 2.18 | 0 | 0.458 | 7.147 | 54.2 | 6.0622 | 3 | 222 | 18.7 | 396.90 | 5.33 | 36.2 |
| 4 | 0.02985 | 0.0 | 2.18 | 0 | 0.458 | 6.430 | 58.7 | 6.0622 | 3 | 222 | 18.7 | 394.12 | 5.21 | 28.7 |

## Turning header row into first row

In [3]:
```
housing = pd.concat([housing.columns.to_frame().T, housing], ignore_index=True)
```

## Replacing column names with acronyms from data dictionary

In [4]:
```
housing.columns = ['CRIM', 'ZN','INDUS','CHAS','NOX','RM','AGE','DIS','RAD','TAX','PTRAT
```

In [5]:
```
housing.head()
```

Out[5]:

| | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | Target Price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.00632 | 18 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.09 | 1 | 296 | 15.3 | 396.9 | 4.98 | 24 |
| 1 | 0.02731 | 0.0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 | 396.9 | 9.14 | 21.6 |
| 2 | 0.02729 | 0.0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 | 392.83 | 4.03 | 34.7 |
| 3 | 0.03237 | 0.0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 | 394.63 | 2.94 | 33.4 |
| 4 | 0.06905 | 0.0 | 2.18 | 0 | 0.458 | 7.147 | 54.2 | 6.0622 | 3 | 222 | 18.7 | 396.9 | 5.33 | 36.2 |

## Checking for missing values and duplicated rows

In [6]:
```
housing.isna().sum().sum()
```

Out[6]:
```
0
```

In [7]:
```
housing.duplicated().sum()
```

Out[7]:
```
0
```

In [8]:
```
housing.shape
```

Out[8]:
```
(506, 14)
```

Inspecting the dataframe

In [9]:
```
housing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   CRIM          506 non-null    object
 1   ZN            506 non-null    object
 2   INDUS         506 non-null    object
 3   CHAS          506 non-null    object
 4   NOX           506 non-null    object
 5   RM            506 non-null    object
 6   AGE           506 non-null    object
 7   DIS           506 non-null    object
 8   RAD           506 non-null    object
 9   TAX           506 non-null    object
 10  PTRATIO       506 non-null    object
 11  B             506 non-null    object
 12  LSTAT         506 non-null    object
 13  Target Price  506 non-null    object
dtypes: object(14)
memory usage: 55.5+ KB
```

## Changing object datatypes to floats

In [11]:
```
housing = housing.astype(float)
```

In [13]:
```
housing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   CRIM          506 non-null    float64
 1   ZN            506 non-null    float64
 2   INDUS         506 non-null    float64
 3   CHAS          506 non-null    float64
 4   NOX           506 non-null    float64
 5   RM            506 non-null    float64
 6   AGE           506 non-null    float64
 7   DIS           506 non-null    float64
 8   RAD           506 non-null    float64
```

```
 9   TAX            506 non-null    float64
 10  PTRATIO        506 non-null    float64
 11  B              506 non-null    float64
 12  LSTAT          506 non-null    float64
 13  Target Price   506 non-null    float64
dtypes: float64(14)
memory usage: 55.5 KB
```

## Running descriptive statistics on the dataframe

In [14]:
```
housing.describe()
```

Out[14]:

| | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.000000 | 506.00000C |
| **mean** | 3.613524 | 11.363636 | 11.136779 | 0.069170 | 0.554695 | 6.284634 | 68.574901 | 3.795043 | 9.549407 |
| **std** | 8.601545 | 23.322453 | 6.860353 | 0.253994 | 0.115878 | 0.702617 | 28.148861 | 2.105710 | 8.70725S |
| **min** | 0.006320 | 0.000000 | 0.460000 | 0.000000 | 0.385000 | 3.561000 | 2.900000 | 1.129600 | 1.00000C |
| **25%** | 0.082045 | 0.000000 | 5.190000 | 0.000000 | 0.449000 | 5.885500 | 45.025000 | 2.100175 | 4.00000C |
| **50%** | 0.256510 | 0.000000 | 9.690000 | 0.000000 | 0.538000 | 6.208500 | 77.500000 | 3.207450 | 5.00000C |
| **75%** | 3.677083 | 12.500000 | 18.100000 | 0.000000 | 0.624000 | 6.623500 | 94.075000 | 5.188425 | 24.00000C |
| **max** | 88.976200 | 100.000000 | 27.740000 | 1.000000 | 0.871000 | 8.780000 | 100.000000 | 12.126500 | 24.00000C |

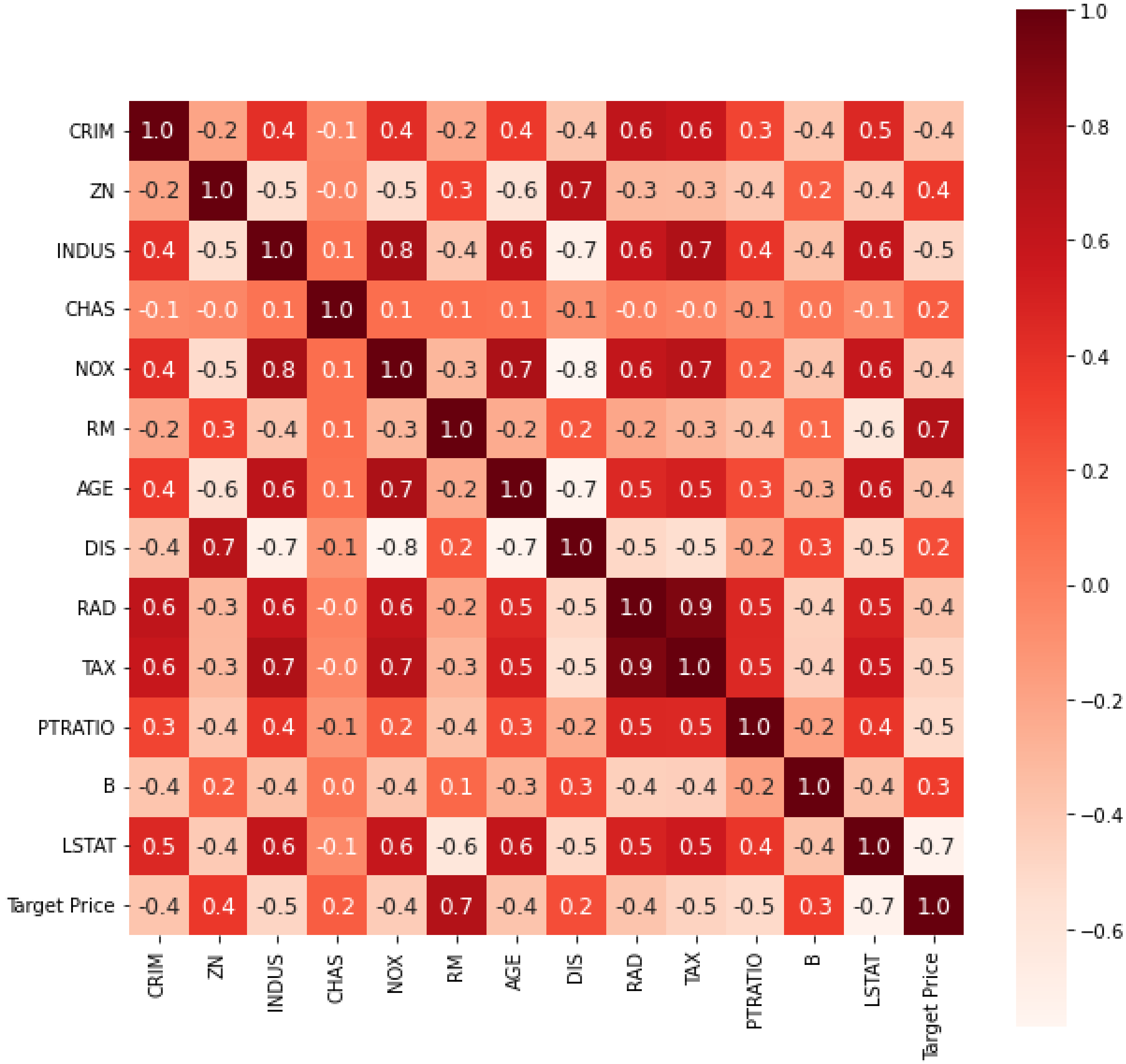
## Understanding Correlation between variables

In [15]:
```
correlation = housing.corr()
```

In [19]:
```
plt.figure(figsize = (10,10))
sns.heatmap(correlation, cbar=True, square=True, fmt='.1f', annot=True, annot_kws={'size
```

Out[19]:

```
<AxesSubplot:>
```




## Creating data model

In [22]:
```
X = housing.drop(['Target Price'], axis=1)
Y = housing['Target Price']

# print(X)
# print(Y)
```

In [23]:
```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, random_state
```

In [24]:
```
print(X.shape, X_train.shape, X_test.shape)
```

```
(506, 13) (404, 13) (102, 13)
```

## Training the model

In [30]:
```
model = XGBRegressor()
model.fit(X_train, Y_train)
```

Out[30]:

```
▾ XGBRegressor
XGBRegressor(base_score=0.5, booster='gbtree', callbacks=None,
             colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
             early_stopping_rounds=None, enable_categorical=False,
             eval_metric=None, gamma=0, gpu_id=-1, grow_policy='depthwise',
             importance_type=None, interaction_constraints='',
             learning_rate=0.300000012, max_bin=256, max_cat_to_onehot=4,
             max_delta_step=0, max_depth=6, max_leaves=0, min_child_weight=1,
             missing=nan, monotone_constraints='()', n_estimators=100, n_jobs=0,
             num_parallel_tree=1, predictor='auto', random_state=0, reg_alpha=0,
             reg_lambda=1, ...)
```

### Prediction for training data

In [32]:
```
train_data_predict = model.predict(X_train)
```

In [34]:
```
print(train_data_predict) #values are in thousands
```

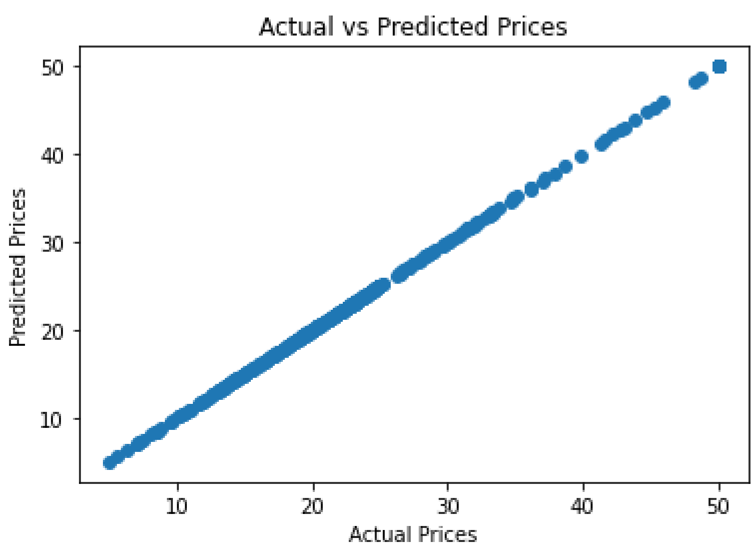
```
[23.147501  20.99463   20.090284  34.69053   13.903663  13.510157
 21.998634  15.1940975 10.899711  22.709627  13.832816   5.592794
 29.810236  49.99096   34.89215   20.607384  23.351097  19.23555
 32.695698  19.641418  26.991022   8.401829  46.00729   21.708961
 27.062933  19.321356  19.288303  24.809872  22.61626   31.70493
 18.542515   8.697379  17.395294  23.700663  13.304856  10.492197
 12.688369  25.016556  19.67495   14.902088  24.193798  25.007143
 14.900281  16.995798  15.6009035 12.699232  24.51537   14.999952
 50.00104   17.525454  21.184624  31.998049  15.613355  22.89754
 19.325378  18.717896  23.301125  37.222923  30.09486   33.102703
 21.00072   49.999332  13.405827   5.0280113 16.492886   8.405072
 28.64328   19.499939  20.586452  45.402164  39.79833   33.407326
 19.83506   33.406372  25.271482  50.001534  12.521657  17.457413
 18.61758   22.602625  50.002117  23.801117  23.317268  23.087355
 41.700035  16.119293  31.620516  36.069206   7.0022025 20.3827
 19.996452  11.986318  25.023014  49.970123  37.881588  23.123034
 41.292133  17.596548  16.305374  30.034231  22.860699  19.810343
 17.098848  18.898268  18.96717   22.606049  23.141363  33.183487
 15.010934  11.693824  18.78828   20.80524   17.99983   19.68991
 50.00332   17.207317  16.404053  17.520426  14.593481  33.110855
 14.508482  43.821655  34.939106  20.381636  14.655634   8.094332
 11.7662115 11.846876  18.69599    6.314154  23.983706  13.084503
 19.603905  49.989143  22.300608  18.930315  31.197134  20.69645
 32.21111   36.15102   14.240763  15.698188  49.99381   20.423601
 16.184978  13.409128  50.01321   31.602146  12.271495  19.219482
 29.794909  31.536846  22.798779  10.189648  24.08648   23.710463
 21.991894  13.802495  28.420696  33.181534  13.105958  18.988266
 26.576572  36.967175  30.794083  22.77071   10.201246  22.213818
 24.483162  36.178806  23.09194   20.097307  19.470194  10.786644
 22.671095  19.502405  20.109184   9.611871  42.799637  48.794792
 13.097208  20.28583   24.793974  14.110478  21.701134  22.217012
 33.003544  21.11041   25.00658   19.122992  32.398567  13.605098
 15.1145315 23.088867  27.474783  19.364998  26.487135  27.499458
 28.697094  21.21718   18.703201  26.775208  14.010719  21.692347
 18.372562  43.11582   29.081839  20.289959  23.680176  18.308306
 17.204844  18.320065  24.393475  26.396057  19.094141  13.3019905
 22.15311   22.185797   8.516214  18.894428  21.792608  19.331121
 18.197924   7.5006843 22.406403  20.004215  14.412416  22.503702
 28.53306   21.591028  13.810223  20.497831  21.898977  23.104464
 49.99585   16.242056  30.294561  50.001595  17.771557  19.053703
 10.399217  20.378187  16.49973   17.183376  16.70228   19.495337
 30.507633  28.98067   19.528809  23.148346  24.391027   9.521643
 23.886024  49.995125  21.167099  22.597813  19.965279  13.4072275
```

```
 19.948694  17.087479  12.738807  23.00453   15.222122  20.604322
 26.207253  18.09243   24.090246  14.105     21.689667  20.08065
 25.010437  27.874954  22.92366   18.509727  22.190847  24.004797
 14.788686  19.89675   24.39812   17.796036  24.556297  31.970308
 17.774675  23.356768  16.134794  13.009915  10.98219   24.28906
 15.56895   35.209793  19.605724  42.301712   8.797891  24.400295
 14.086652  15.408639  17.301126  22.127419  23.09363   44.79579
 17.776684  31.50014   22.835577  16.888603  23.925127  12.097476
 38.685944  21.388391  15.98878   23.912495  11.909485  24.960499
  7.2018585 24.696215  18.201897  22.489008  23.03332   24.260433
 17.101519  17.805563  13.493165  27.105328  13.311978  21.913465
 20.00738   15.405392  16.595737  22.301016  24.708412  21.422579
 22.878702  29.606575  21.877811  19.900253  29.605219  23.407152
 13.781474  24.454706  11.897682   7.2203646 20.521074   9.725295
 48.30087   25.19501   11.688618  17.404732  14.480284  28.618876
 19.397131  22.468653   7.0117908 20.602013  22.970919  19.719397
 23.693787  25.048244  27.977154  13.393578  14.513882  20.309145
 19.306028  24.095829  14.894031  26.382381  33.298378  23.61644
 24.591206  18.514652  20.900269  10.406055  23.303423  13.092017
 24.675085  22.582184  20.502762  16.820635  10.220605  33.81239
 18.608067  49.999187  23.775583  23.909609  21.192276  18.805798
  8.502987  21.50807   23.204473  21.012218  16.611097  28.100965
 21.193024  28.419638  14.294126  49.99958   30.988504  24.991066
 21.433628  18.975573  28.991457  15.206939  22.817244  21.765755
 19.915497  23.7961   ]
```

In [41]:
```
#R squared error

score_1 = metrics.r2_score(Y_train, train_data_predict)

# Mean Absolute Error

score_2 = metrics.mean_absolute_error(Y_train, train_data_predict)

print('R squared error:',score_1,'\n','Mean Absolute Error:', score_2)
```

```
R squared error: 0.9999948236320982
 Mean Absolute Error: 0.0145848437110976
```

In [46]:
```
plt.scatter(Y_train, train_data_predict)
plt.xlabel("Actual Prices")
plt.ylabel('Predicted Prices')
plt.title('Actual vs Predicted Prices')
plt.show()
```




## Predictions for test data

In [44]:
```
test_data_predict = model.predict(X_test)
```

In [45]:
```
score_1 = metrics.r2_score(Y_test, test_data_predict)

# Mean Absolute Error

score_2 = metrics.mean_absolute_error(Y_test, test_data_predict)

print('R squared error:',score_1,'\n','Mean Absolute Error:', score_2)
```

```
R squared error: 0.8711660369151691
 Mean Absolute Error: 2.2834744154238233
```