PREDICTING BOSTON'S HOUSING VALUATIONS: A DATA-DRIVEN MULTIPLE REGRESSION STUDY

BY
OMOGBEME ANGELA

.

AGENDA

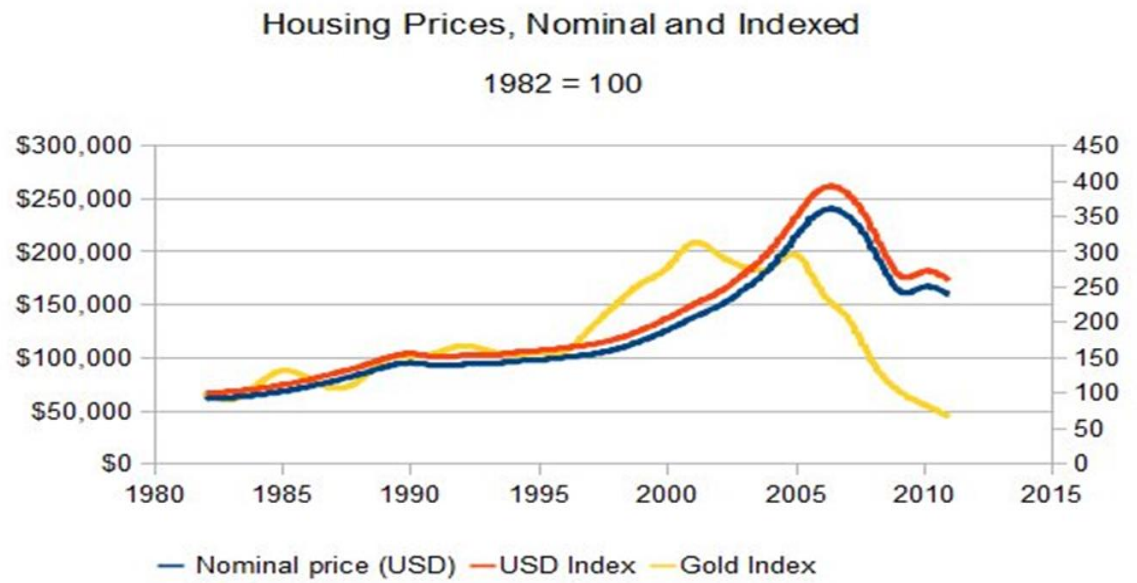
- Problem Statement and objectives
- Literature Review
- Methodology
- Findings
- Executive Summary
- Recommendations
- Appendix

Business Problem- What factors determine the worth of a house in Boston?

.

What is the significance of this project?

- To understand the factors that influence housing prices in the Boston metropolitan area.
- To build predictive models that can provide insights that are valuable to various stakeholders e. g real estate professionals, policy makers, homebuyers and sellers, researchers etc.








What insights have other researchers provided regarding similar projects?

- Property attributes and location are major contributors to the Price of a Home Hal Bundrick, September 12, 2016 .

- Jafari and Akhavian (2019) stated that a hierarchical emphasis on square footage is the primary factor and most important variable in predicting the price of a house, followed by the number of bathrooms and bedrooms.

METHODOLOGY

This study employs SAS programming techniques to address the house price prediction task, leveraging a dataset obtained from the U.S. Census Service on housing in the Boston area following the steps below.

DATA CLEANING AND ANALYSIS: HANDLING MISSING DATA, OUTLIERS, DUPLICATE VALUES ETC.

EXPLORATORY DATA ANALYSIS: SUMMARIZE THE MAIN CHARACTERISTICS OF THE DATA, GAIN INSIGHTS INTO ITS UNDERLYING STRUCTURE, DETECT PATTERNS, IDENTIFY ANOMALIES FOR FURTHER INVESTIGATION..

FEATURE SELECTION: IMPROVE MODEL PERFORMANCE, REDUCE OVERFITTING.

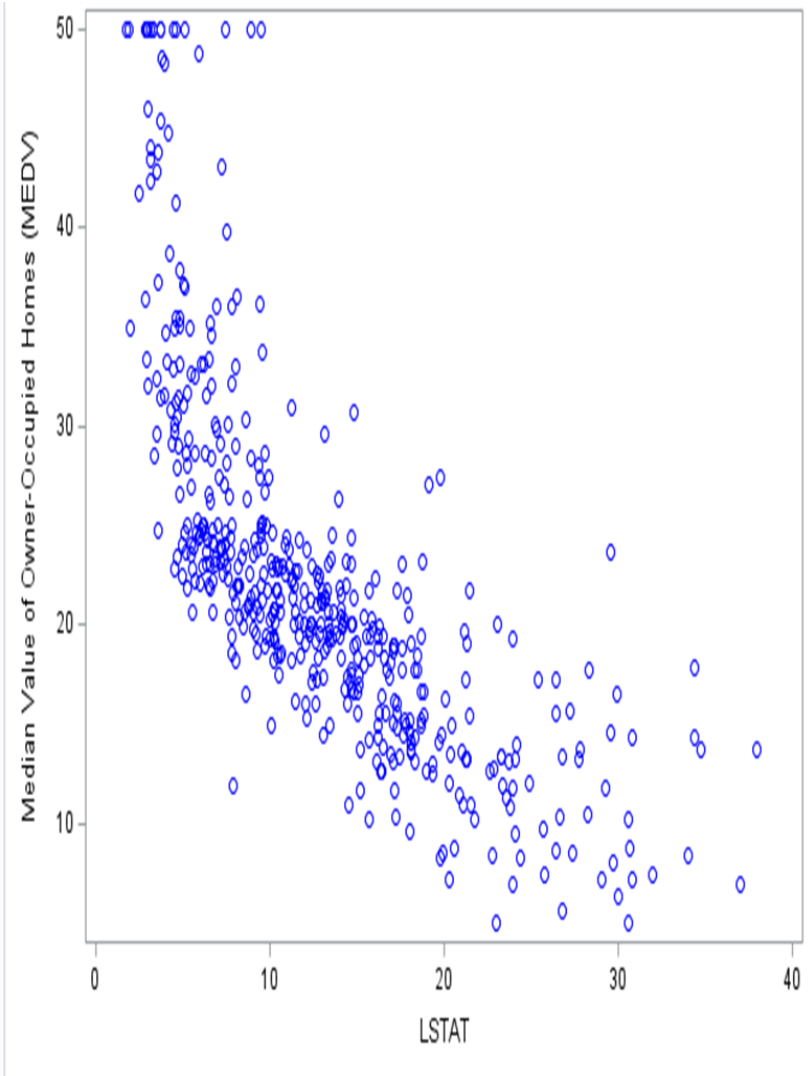
MODEL BUILDING AND EVALUATION :TRAINING, TESTING, PERFORMANCE EVALUATION

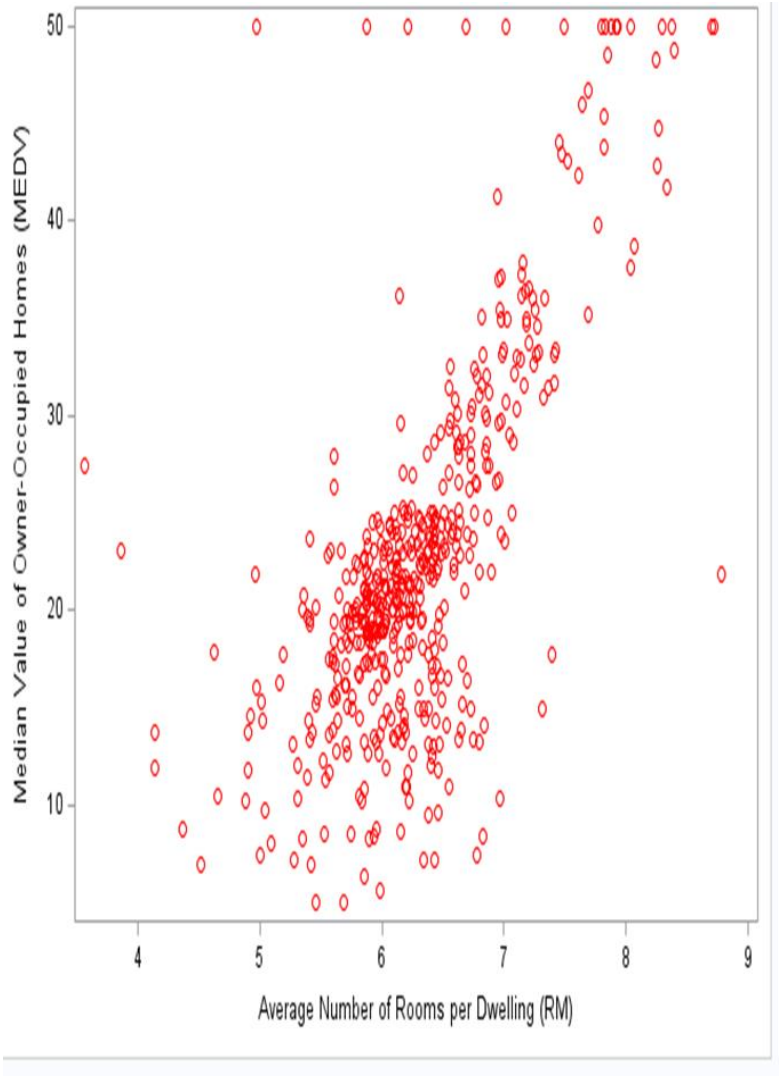
EXPLANATION OF FEATURES

- CRIM : per capita crime rate by town
- ZN : proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS : proportion of non-retail business acres per town.
- CHAS : Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX : nitric oxides concentration (parts per 10 million)
- RM : average number of rooms per dwelling
- AGE : proportion of owner-occupied units built prior to 1940
- DIS : weighted distances to five Boston employment centers
- RAD : index of accessibility to radial highways
- TAX : full-value property-tax rate per $10,000
- PTRATIO : pupil-teacher ratio by town
- B : 1000(Bk – 0.63)2 where Bk is the proportion of blacks by town
- LSTAT : % lower status of the population
- MEDV : Median value of owner-occupied homes in $1000’s

KEY FINDINGS FROM STATISTICAL ANALYSIS AND MODEL PREDICTION

SCATTERPLOTS TO VISUALIZE THE RELATIONSHIP BETWEEN "LSAT" AND "RM" AND THE PRICE OF THE HOUSE "MEDV" USING SAS PROGRAMMING

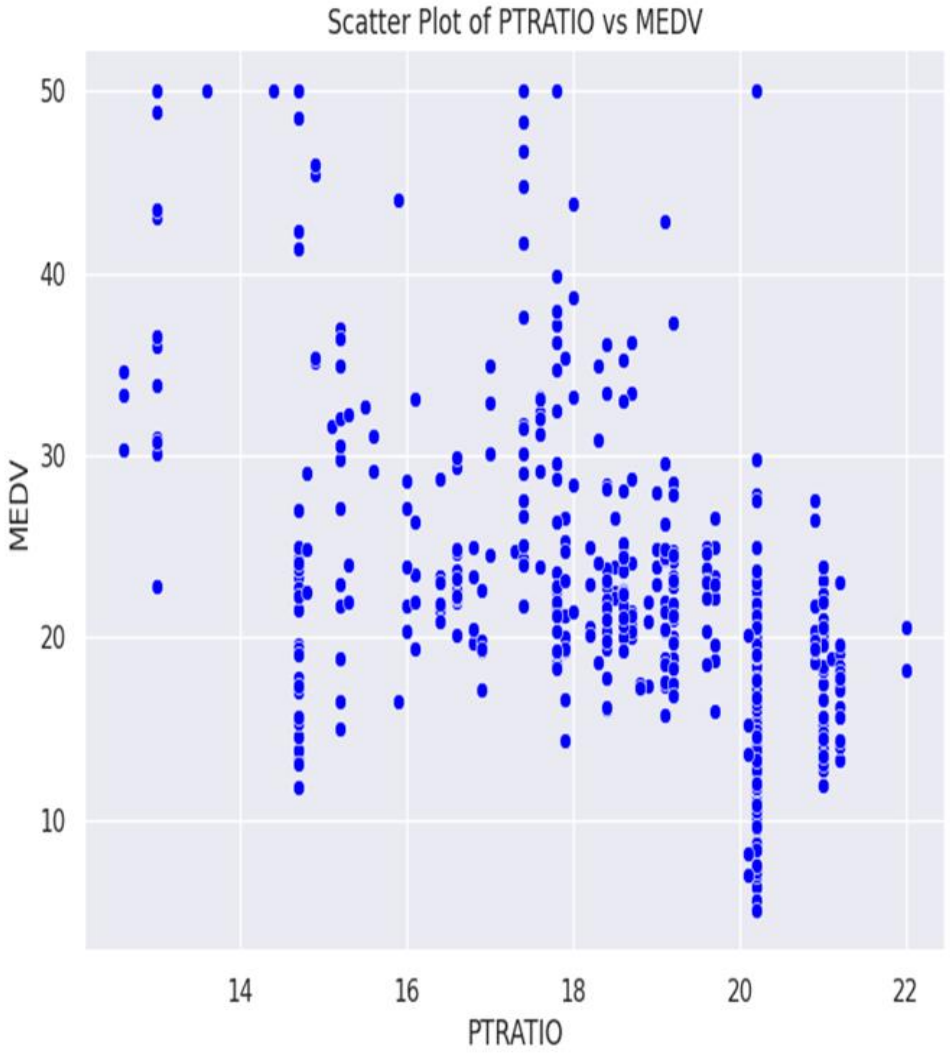


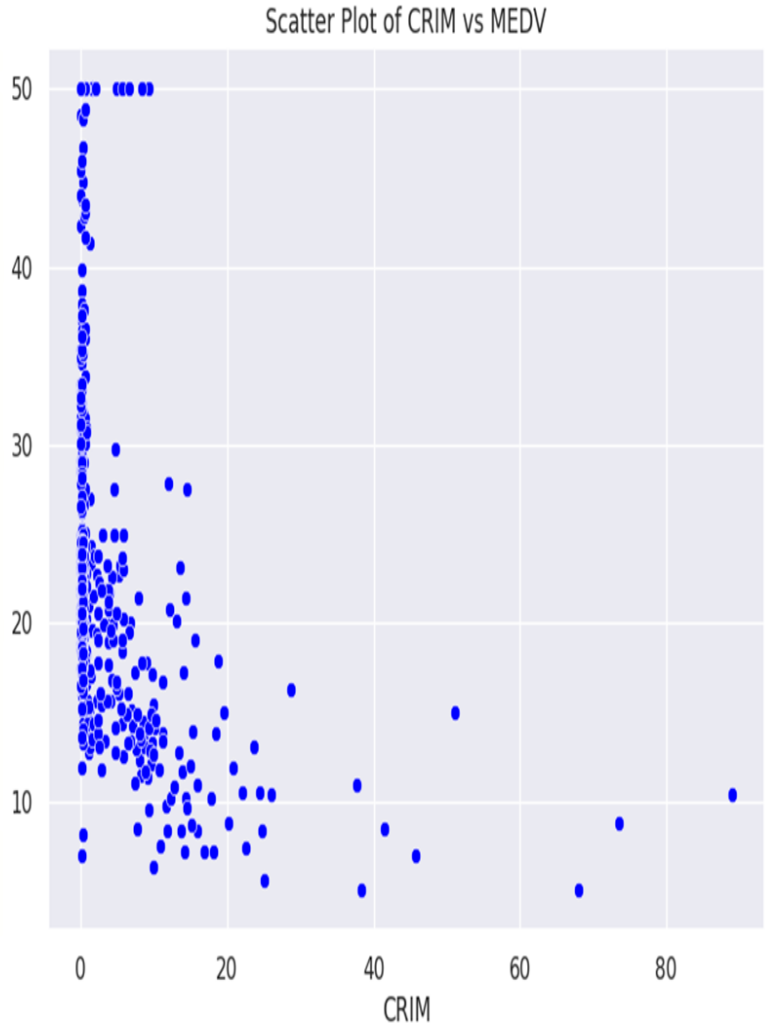


The **prices** of the house tend to decrease with an increase in LSTAT(Lower Status Of Population)

The **number of room(RM)** increases when the median value of owner-occupied homes tends to rise consistently.

SCATTERPLOTS TO VISUALIZE THE RELATIONSHIP BETWEEN "CRIME RATE(CRIM)" VS PRICE(MEDV)" AND "PUPILS-TEACHER RATIO(PTRATIO) VS PRICE"



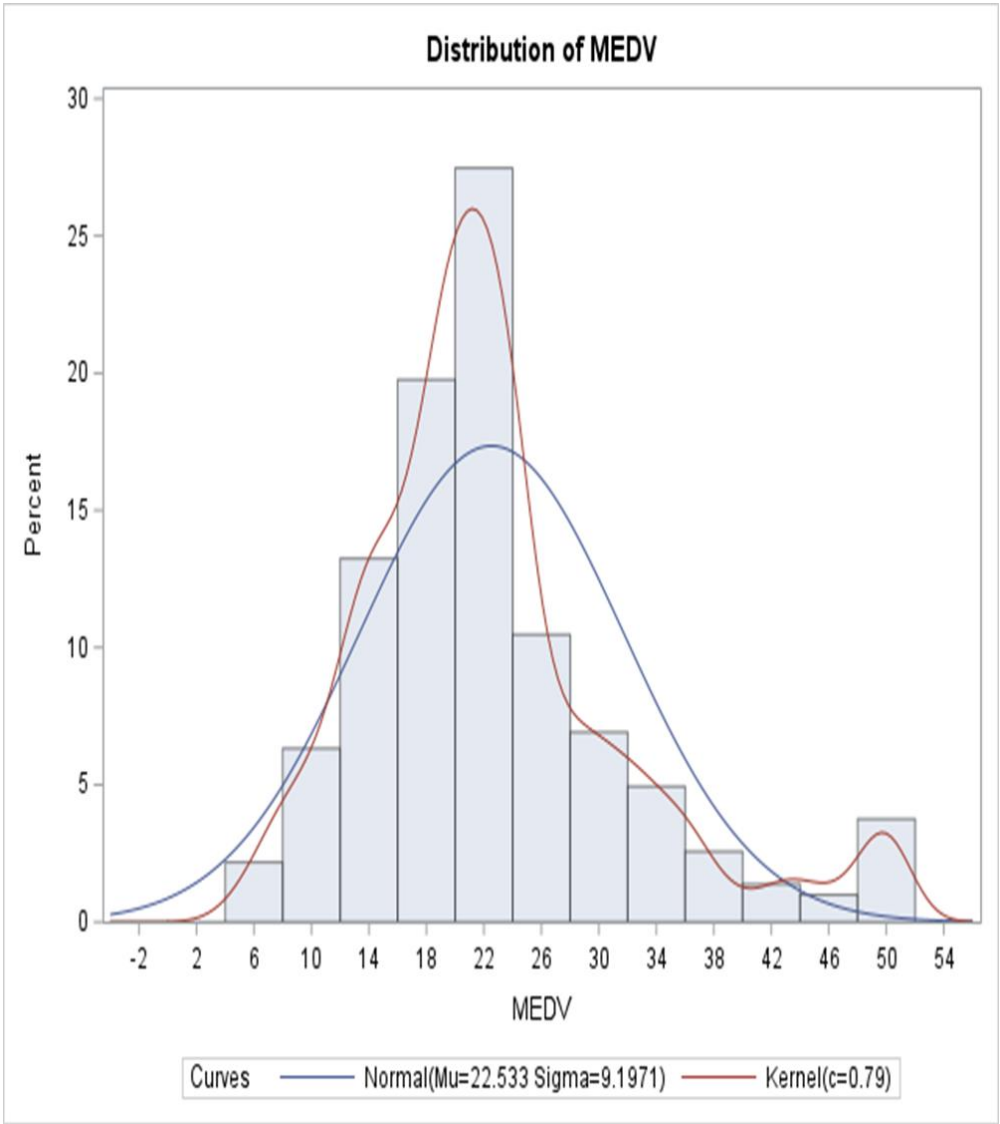


The **price(MEDV)** of house tend to decrease with an increase in **crime rate(CRIM)**

The **price(MEDV)** of house decreases as the **Pupils-Teacher Ratio(PTRatio)** increases

DISTRIBUTION OF MEDIAN VALUE OF OWNER- OCCUPIED HOMES(PRICE)

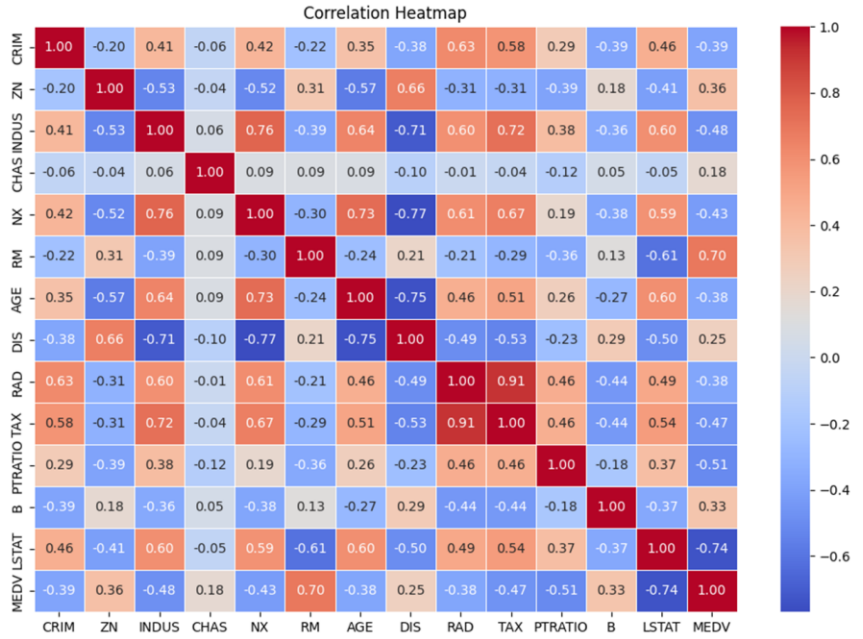
- The distribution of MEDV appears to follow a normal distribution, with just a few outliers present.



RELATIONSHIP BETWEEN INDEPENDENT VARIABLES USING HEAT MAP

- By looking at the correlation matrix we can see that RM has a strong positive correlation with MEDV (0.7) ,
- whereas LSTAT has a high negative correlation with MEDV(-0.74).

- The strong positive correlation coefficient of 0.91 between the features RAD and TAX indicates a highly linear relationship, suggesting multicollinearity, we therefore exclude one of these highly correlated features from the analysis



SUMMARY STATISTICS

DESCRIPTIVE ANALYSIS

	CRIM	ZN	INDUS	CHAS	NX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	12.653063	22.532806	
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	7.141062	9.197104	
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.730000	5.000000	
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	6.950000	17.025000	
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	11.360000	21.200000	
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	16.955000	25.000000	
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	37.970000	50.000000	

PARAMETER ESTIMATES

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	32.68006	5.68129	5.75	<.0001	21.50936	43.85076
CRIM	CRIM	1	-0.09759	0.03246	-3.01	0.0028	-0.16141	-0.03378
ZN	ZN	1	0.04890	0.01440	3.40	0.0008	0.02060	0.07721
INDUS	INDUS	1	0.03038	0.06593	0.46	0.6452	-0.09926	0.16002
CHAS	CHAS	1	2.76938	0.92517	2.99	0.0029	0.95028	4.58847
NOX	NOX	1	-17.96903	4.24286	-4.24	<.0001	-26.31144	-9.62661
RM	RM	1	4.28325	0.47071	9.10	<.0001	3.35773	5.20877
AGE	AGE	1	-0.01299	0.01446	-0.90	0.3695	-0.04142	0.01544
DIS	DIS	1	-1.45851	0.21101	-6.91	<.0001	-1.87340	-1.04362
RAD	RAD	1	0.28587	0.06930	4.13	<.0001	0.14961	0.42212
TAX	TAX	1	-0.01315	0.00396	-3.32	0.0010	-0.02092	-0.00537
PTRATIO	PTRATIO	1	-0.91458	0.14058	-6.51	<.0001	-1.19100	-0.63817
B	B	1	0.00966	0.00297	3.25	0.0013	0.00382	0.01549
LSTAT	LSTAT	1	-0.42366	0.05502	-7.70	<.0001	-0.53185	-0.31548

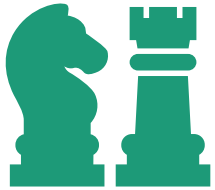
Root MSE	4.48707	R-Square	0.7671
Dependent Mean	22.35964	Adj R-Sq	0.7591
Coeff Var	20.06772		

- ZN, CHAS NOX, RM, DIS, RAD, TAX, PTRATIO, B and LSTAT have t-statistics with absolute values greater than 2.0 and are considered statistically significant.
- CRIM, ZN, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, B, LSTAT have very low p-values, they are statistically significant predictors of the outcome variable.
- INDUS, AGE: These variables have p-values above the typical threshold of 0.05, indicating that they are not statistically significant predictors in this model.
- The intercept term is found to be statistically significant, indicating a baseline value for the target variable. Notably, variables such as CRIM, ZN, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, B, and LSTAT exhibit statistically significant coefficients, suggesting their importance in influencing the target variable.

FINDINGS

The highest predictors of house price(MEDV) include Nitric Oxides Concentration (NX), Average Number of Rooms per Dwelling (RM), and Crime Rates (CRIM). These factors significantly influence the pricing dynamics in the housing market.

Root Mean Square Error (RMSE) of 4.48707 stood as a beacon of accuracy, measuring the dispersion of observed values from the regression line with a steady hand.



R-Square, a metric of goodness-of-fit, has a commendable value of 0.7671

Factors Influencing an increase in Median Home Values(MDV) in the Boston Housing Market using the positive coefficients of the predictors

FEATURES	COEFFICIENT	ECONOMIC CONSEQUENCE
RAD (index of accessibility to radial highways)	0.28	The positive coefficient of 0.28 for the index of accessibility to radial highways suggests that this variable has a **medium impact** on housing decisions
RM (average number of rooms per dwelling)	4.44	Properties with more rooms tend to have higher market values or prices and **strong impact**.
Bk is the proportion of blacks by town	0.01	The proportion of Black residents indicates that this variable has a **low negative impact** on housing price.
CHAS(Charles River)	2.78	The proximity to the Charles River has a **high positive impact** on housing decisions

Analyzing a few Factors Influencing a decrease in Median Home Values(MEDV) in the Boston Housing Market using the negative coefficients of the predictors

FEATURES	COEFFICIENT	ECONOMIC CONSEQUENCE
NX (nitric oxides concentration)	-17.2	The high impact of Nitric Oxide has a **strong negative** economic impact on the price of house
DIS (weighted distances to five Boston employment centers)	-1.45	The weighted distance to employment centers indicates that this factor has a **moderate impact** on housing decisions.
CRIM (per capita crime rate)	-0.11	The moderate negative coefficient of -0.11 implies that crime rates have a **medium impact** on housing decisions
PTRATIO (pupil-teacher ratio by town)	-0.92	The negative coefficient of -0.92 for the pupil-teacher ratio suggests that this variable has a **moderate impact** on housing decisions.

LIMITATIONS

- **Limited Geographic Scope**: The dataset is specific to the Boston metropolitan area and may not generalize well to other cities or regions with different housing market characteristics.

- **Missing Variables**: The dataset may not include all relevant features that could potentially influence housing prices, such as proximity to public transportation, school district ratings, or neighborhood amenities.

- **Temporal Analysis**: Incorporate temporal dimensions to the analysis, such as housing price trends over time, to better understand the dynamics of the housing market.

RECOMMENDATIONS

Prioritize Improving Air Quality(NX): Based on high concentration of NX, reducing air pollution could significantly boost home values in the Boston area.

Address Crime Rates through Comprehensive Strategies(CRIM): Implementing a multifaceted approach to reduce crime rates, such as community policing, social programs etc.

RECOMMENDATIONS

- **Preserve and Enhance Access to the Charles River(CHAS)**: Maintaining and improving access to the Charles River should be a key consideration for urban development plans, as proximity to this natural amenity is a highly desirable factor for homebuyers in Boston.

- **Improving the quality of local schools,** as measured by the pupil-teacher ratio (PTRATIO), should be a priority for policymakers and education authorities

- By addressing key predicting factors such as NX(,PTRATIO,CRIM and CHAS, policymakers, urban planners, and real estate professionals can work to enhance the desirability and value of the Boston housing market, ultimately benefiting both homebuyers and homeowners in the region.

LINES OF CODE

```
LIBNAME exam1 'C:\Users\aomogbe1\Desktop\exam1';

PROC IMPORT OUT=exam1.bostonhousing
            DATAFILE=  "C:\Users\aomogbe1\Desktop\exam1\bostonhousing.xls"
            DBMS=EXCEL REPLACE;
     RANGE="bostonhousing$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

proc contents data=exam1.bostonhousing;
RUN;



 /* summarystatistics /*
proc means data=exam1.bostonhousing mean std min maxn;
run;
```

LINES OF CODE

```
/* Count missing values for all variables */
proc freq data=exam1.bostonhousing;
  tables _numeric_ / missing;
run;

/* Remove duplicate observations */
proc sort data=exam1.bostonhousing_cleaned nodupkey out=exam1.bostonhousing_cleaned_nodup;
  by _all_;
run;

/* Check for outliers using PROC UNIVARIATE */
proc univariate data=exam1.bostonhousing_cleaned_nodup;
  var _numeric_;
run;

/* Clean the existing dataset */
proc stdize data=exam1.bostonhousing
        method=mean;
  var _numeric_;
run;
```

LINES OF CODE

```
/* Correlation matrix of numeric variables */
proc corr data=exam1.bostonhousing;
   var _numeric_;
run;
/* Histograms of numeric variables */
proc univariate data=exam1.bostonhousing;
   var _numeric_;
   histogram / normal kernel;
run;

/* Scatter plot of two numeric variables */
proc sgplot data=exam1.bostonhousing;
   scatter x=CRIM y=MEDV / markerattrs=(color=blue);
run;
/* Perform linear regression analysis */
proc reg data=exam1.bostonhousing;
   model MEDV = CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B LSTAT / clb;
   /* Specify the variables in the MODEL statement */
   /* CLB option requests the confidence limits for the coefficients */
run;
```

LINES OF CODE

```
/proc reg data=exam1.bostonhousing;
   model MEDV = CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B
     LSTAT;
run;


/* Model Building */


/* Split data into training and testing sets */
data housing_train housing_test;
   set exam1.bostonhousing;
   if mod(_n_, 5) = 0 then output housing_test;
   else output housing_train;
run;


/* Evaluate model performance */
proc model data=housing_test;
   score data=housing_test out=housing_pred predicted;
run;
```

LINES OF CODE

```
/* Interpretation of Results */


/* Examine model coefficients and significance */
proc reg data=exam1.bostonhousing;
  model MEDV = CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B LSTAT /
    clb;
run;


/* Assess Normality of Residuals */


/* Use PROC UNIVARIATE  to analyze the distribution of residuals
    */
proc univariate data=residuals;


  /* Specify the variable of interest */
  var residual;


  /* Create a histogram to visualize the distribution */
  histogram / normal; /* Overlay a normal distribution curve */


run;
```

THANK YOU

- **OMOGBEME ANGELA**
- **ECONOMICS DEPARTMENT**
- **UNIVERSITY OF WEST GEORGIA**
- **aomogbe1@my.westga.edu**

THANK YOU

OMOGBEME ANGELA

ECONOMICS DEPARTMENT,

UNIVERSITY OF WEST GEORGIA

aomogbe1@my.westga.edu