# Boston Housing Market Analysis

Advanced Excel Capstone Project

Prepared by: Binyam Siyum

Date: October 15, 2025

## 1. Introduction and Project Objectives

This capstone project analyzes the Boston Housing Market dataset to identify the factors that influence the median value of owner-occupied homes (MEDV).   
The goal is to apply advanced Excel techniques to clean, visualize, and model real-world data through regression analysis and hypothesis testing.

Objectives:

• Clean and prepare data for analysis.  
• Perform Exploratory Data Analysis (EDA) using charts and summary statistics.  
• Build and interpret both simple and multiple regression models.  
• Conduct hypothesis testing to identify significant predictors.  
• Provide insights and recommendations for decision-making.

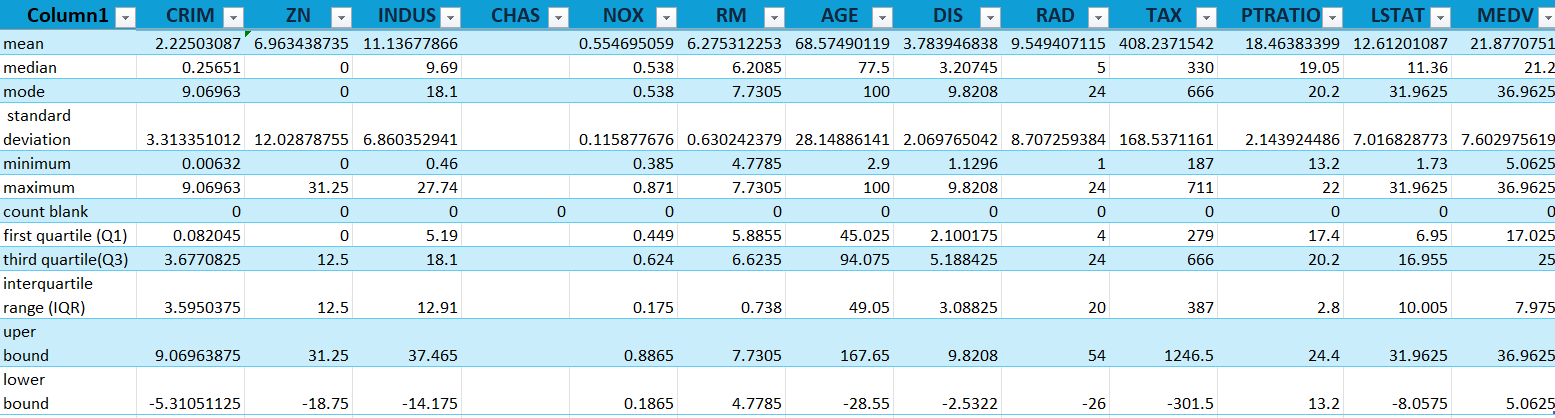
## 2. Data Preparation Summary

The dataset consists of 506 rows and 13 variables related to housing characteristics in Boston neighborhoods.

Data Cleaning Steps:

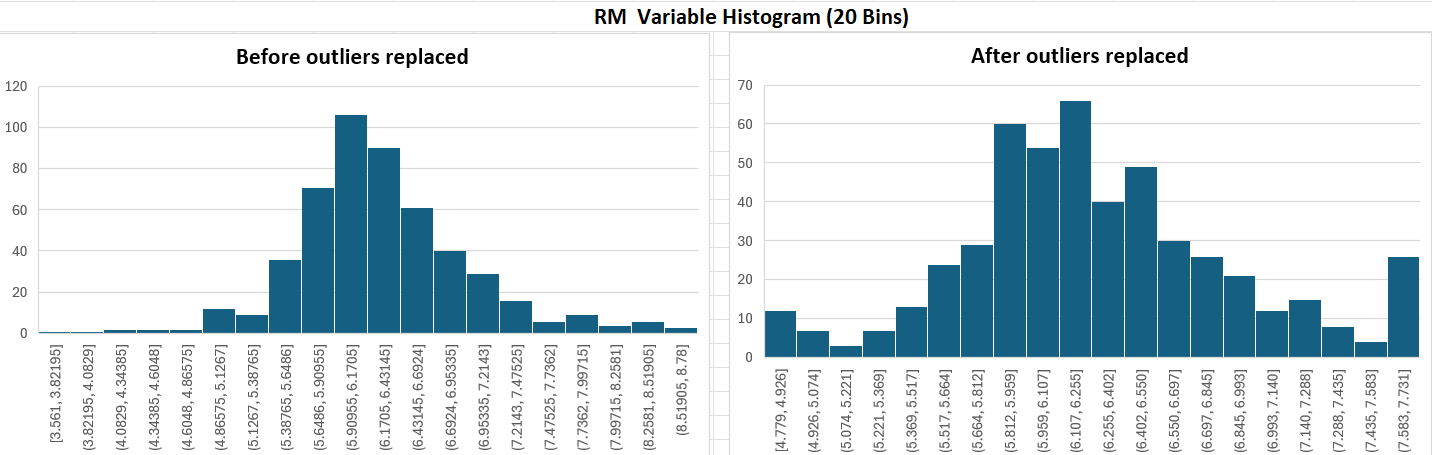
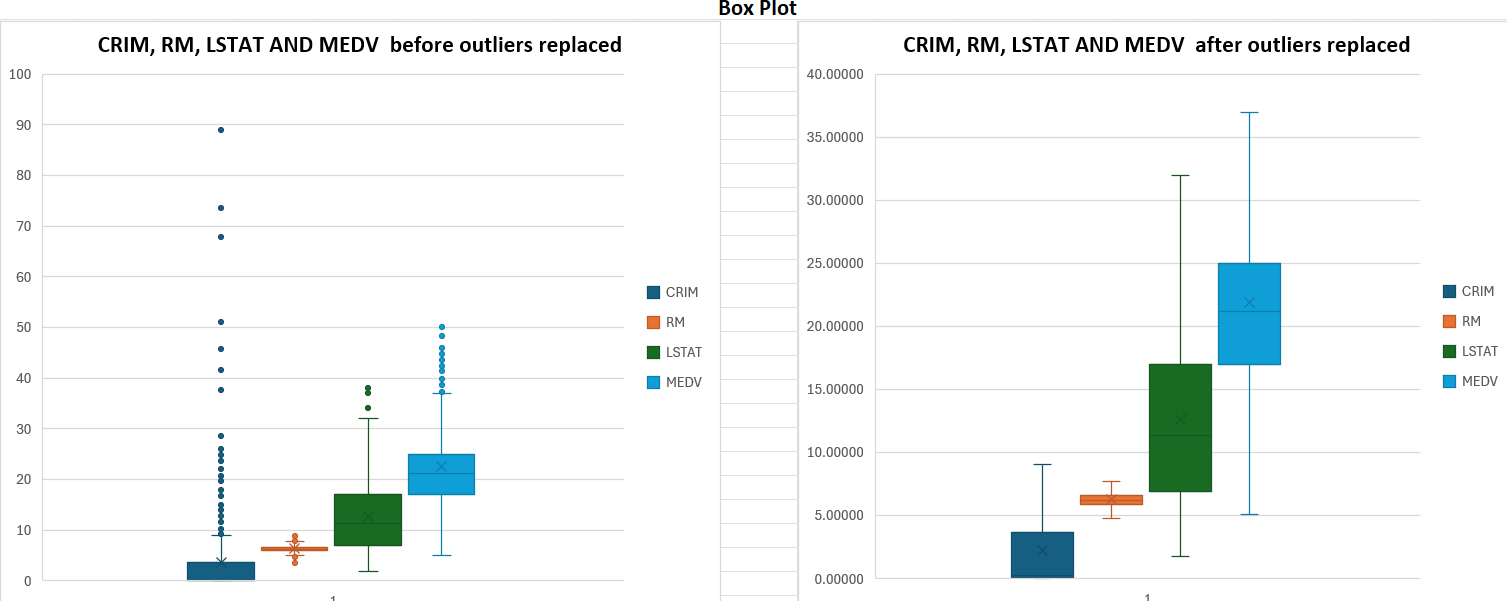
• Checked for missing values using COUNTBLANK () – no missing data found.  
• Identified outliers using IQR (Q1, Q3, and IQR bounds) and replaced values beyond limits with corresponding upper or lower bounds.  
Outliers replaced: CRIM (66), ZN (68), RM (30), DIS (1), PTRATIO (15), LSTAT (7), and MEDV (40).

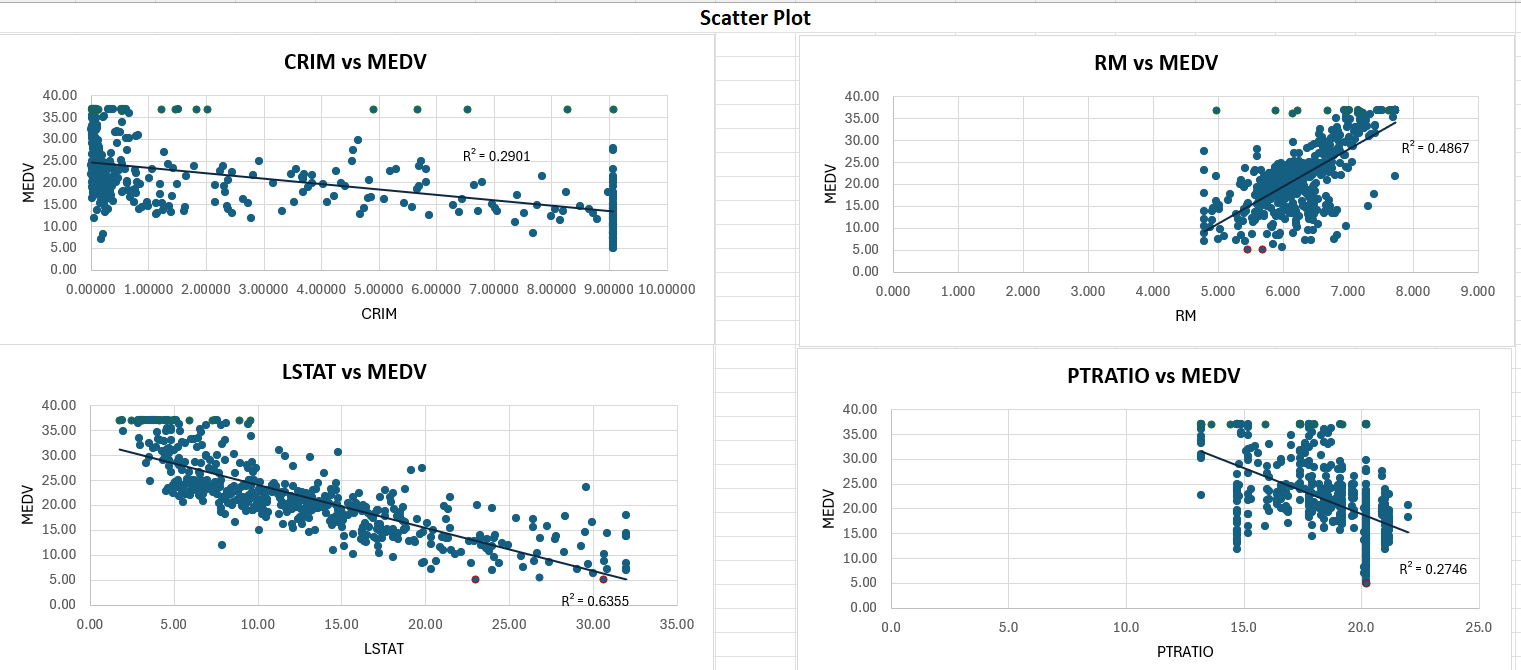
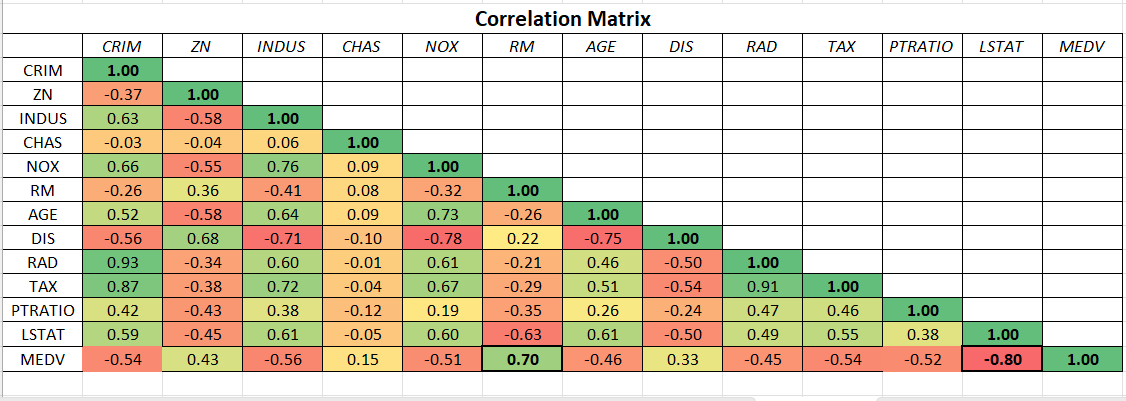
Result: A clean, consistent dataset ready for analysis with no duplicates or missing values and minimal skewness after outlier treatment.



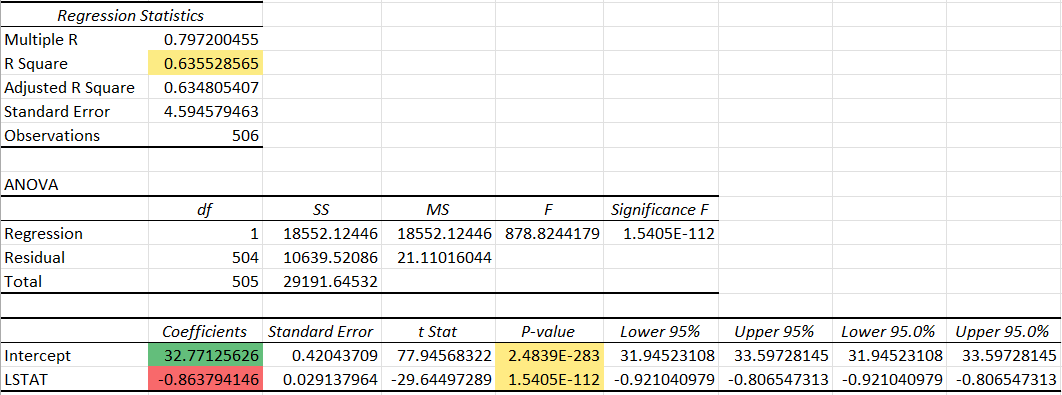
## 3. Key EDA Findings and Visualizations

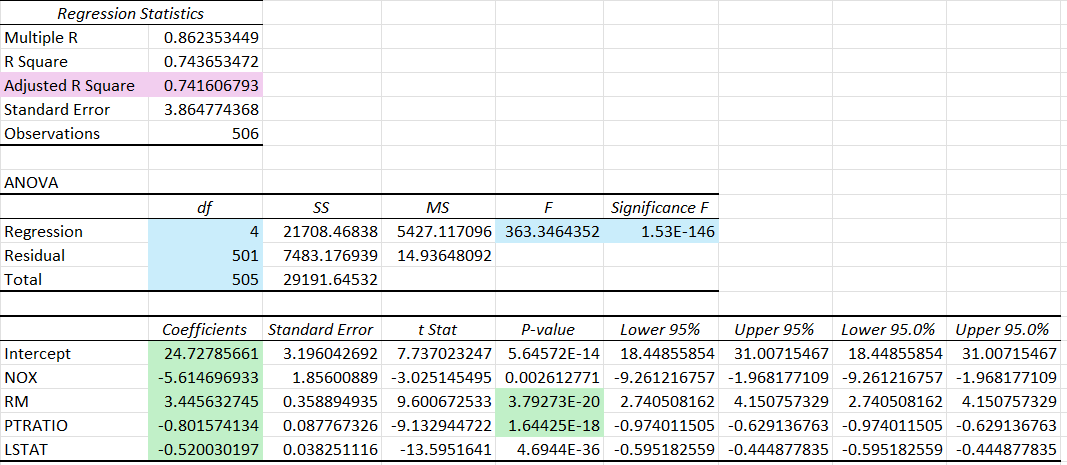
Univariate Analysis:  
• RM appears normally distributed.  
• Box plots confirmed no remaining outliers after replacement.

  
  
Bivariate Analysis:  
• RM vs MEDV: positive relationships as rooms increase, home value increases.  
• LSTAT vs MEDV: negative relationship – higher lower status percentage reduces home values.  
• PTRATIO and NOX vs MEDV: negative relationships.

  
  
Correlation Insights:  
• LSTAT and MEDV: Strong negative correlation (r = -0.80).  
• RM and MEDV: Strong positive correlation (r = 0.70).  
• RAD and TAX: Strong positive correlation (r = 0.91) → one removed to avoid multicollinearity.  


## 4. Regression Analysis Interpretation

Simple Linear Regression (LSTAT vs MEDV)  
• Equation: MEDV=32.77 - 0.86(LSTAT )  
• R² = 0.64 → 64% of variation in MEDV explained by LSTAT.  
• Adjusted R² = 0.64 (same since one predictor).  
• Coefficient (β₁ = -0.86): For each 1% increase in LSTAT, MEDV decreases by about $860.  
→ Indicates a strong negative relationship.  
  
Multiple Linear Regression (LSTAT, RM, PTRATIO, NOX vs MEDV)  
• Equation: MEDV =24.73 -5.61(NOX)+3.45(RM)-0.80(PTRATIO)-0.52(LSTAT)  
• R² = 0.74, Adjusted R² ≈ 0.74 → 74% of variation explained collectively.  
• Coefficients Interpretation:  
 - LSTAT (-0.52): 1% increase in LSTAT MEDV decreases by $520 .  
 - RM (+3.45): Each additional room increases MEDV by ~$3,450.  
 - PTRATIO (-0.80): Each unit increase in student–teacher ratio decreases MEDV by ~$800.  
 - NOX (-5.61): Each unit that rises in air pollution reduces MEDV by ~$5,610.



## 5. Hypothesis Testing and Decision Making

Hypothesis 1:  
H₀: RM has no significant effect on MEDV (β = 0)  
H₁: RM has a significant effect on MEDV (β ≠ 0)  
p-value = 3.79E-20 < 0.05 → Reject H₀ → RM significantly affects MEDV.  
  
Hypothesis 2:  
H₀: PTRATIO has no significant effect on MEDV (β = 0)  
H₁: PTRATIO has a significant effect on MEDV (β ≠ 0)  
p-value = 1.64E-18 < 0.05 → Reject H₀ → PTRATIO significantly affects MEDV.

## 6. Conclusion and Recommendations

Summary of Findings:  
• Outliers handled using IQR bounds for accurate analysis.  
• LSTAT, PTRATIO and NOX negatively impact home values.

• RM positively impacts home values.   
• The multiple regression model explains 74% of the variation in MEDV.  
  
Recommendations:  
For Homebuyers:  
• Choose neighborhoods with low NOX and LSTAT for higher property values.  
• Prefer homes with more rooms for better long-term value.  
  
For Real Estate Developers:  
• Design homes with larger rooms to increase property values.  
• Invest in environmental quality and education infrastructure.  
  
For Urban Planners and Policymakers:  
• Focus on reducing pollution and improving school systems.  
• Invest in crime reduction and safety programs to enhance community value.  
  
For Real Estate Agents:  
• Highlight “number of rooms” and “low PTRATIO areas” as selling points.  
  
Overall Conclusion:  
The Boston housing market is influenced most strongly by socioeconomic factors, air quality, and housing characteristics. Larger homes, cleaner air, and better schools consistently lead to higher median home values.