Business Intelligence and Data Analytics

Module 7: Portfolio Project

Adam S. Cubbage

Colorado State University-Global

25SD-MIS581-1: Business Intelligence and Data Analytics

Dr. Morad

July 24, 2025

Contents

0.0 Abstract

This project examines how neighborhood characteristics—specifically crime rate, socioeconomic status, and education quality—affect median housing values using the Boston Housing dataset (Altaf, 2022), which includes census tract data from the 1970 U.S. Census. A total of 466 complete observations were analyzed using multiple statistical methods. A multiple linear regression model explained 61% of the variance in median home values, indicating a strong model fit. Welch's ANOVA revealed statistically significant differences in median home values across school quality tiers, while a Spearman rank correlation confirmed a negative monotonic relationship between crime rate and housing value. All null hypotheses were rejected, supporting the influence of the selected predictors. Outliers were retained to preserve distributional integrity and ensure analytical completeness. The findings align with themes identified in the literature and emphasize the impact of structural and socioeconomic conditions on real estate outcomes. Recommendations include reintroducing excluded sensitive features for future study, investing in equitable educational resources, supporting community-based crime reduction, reinforcing affordable housing protections, expanding geographic and longitudinal scope, analyzing interaction effects, and applying causal inference techniques.

Business Intelligence and Data Analytics

Module 7: Portfolio Project

1.0 Introduction

This paper presents the finished version of the Portfolio Project. It integrates revised content from prior modules along with new material developed for the Module 7 assignment (Morad, 2025). The project centers on evaluating how crime rates, socioeconomic status, and education quality influence median housing values using the Boston Housing dataset (Altaf, 2022). Quantitative methods including regression, ANOVA, and correlation were applied to explore these relationships.

This final draft presents a comprehensive overview of the study, including a dataset summary, clearly defined research questions and hypotheses, an integrative literature review, and a detailed research design. It further addresses the methodology, limitations, and ethical considerations, followed by analysis findings and actionable recommendations.

2.0 Objectives

The goal of this project was to determine the impact of per capita crime rate by town (CRIM), percentage of lower-status population (LSTAT), and pupil–teacher ratio by town (PTRATIO) on the median value of owner-occupied homes in USD 1000s (MEDV). MEDV serves as a proxy for housing affordability and market value, while PTRATIO is used as an indicator of education quality. This analysis follows the Six-Step Research Process outlined by Polonsky and Waller (2018):

1. Problem Definition:
    - Housing values are influenced in part by neighborhood characteristics that may benefit from targeted social programs and policy interventions; however, it remains unclear which of these factors should be prioritized based on their relative impact.

2. Research Objectives:
    - Determine the impact that CRIM, LSTAT, and PTRATIO have on MEDV.
3. Research Design:
    - Use multiple linear regression to evaluate the combined influence of CRIM, LSTAT, and PTRATIO on MEDV.
    - Use Welch's ANOVA to compare mean home values across categorized PTRATIO groups representing school quality tiers.
    - Use Spearman Rank Correlation to test the strength and direction of a monotonic relationship between MEDV and LOG_CRIM (log-transformed CRIM).
4. Data Gathering:
    - The Boston Housing dataset (Altaf, 2022) includes census tract-level data from Boston. It was sourced from Kaggle, originally made available via the UCI Repository.
5. Data Analysis and Interpretation:

   - Descriptive statistics were performed to summarize key variables (see ***Appendix A),*** while predictive statistical analyses were conducted to test the research hypotheses (see ***Appendices B-D***).

6. Presenting the Results:
   - Results are presented in Section 10.0, Findings, which details the statistical outcomes for each research question.

When applying the SMART framework, the project objectives follow this:

- Specific:
  - Analyze the influence of CRIM, LSTAT, and PTRATIO on MEDV.
- Measurable:
  - Use regression, ANOVA, and correlation analysis to quantify statistical significance of variable relationships.
- Achievable:
  - Analysis was conducted using Jupyter Notebook, part of the Anaconda platform (Anaconda, Inc., 2024; Project Jupyter, 2024), following best practices outlined in McKinney (n.d.) and official Python documentation (Python Software Foundation, 2024).
- Relevant:
  - Findings support data-driven decisions in housing and urban policy.
- Time-bound:

- Analysis and reporting phases of the project have been completed within the established timeframe.

3.0 Study Overview

This section provides the finalized dataset summary. While descriptive analyses are not discussed in detail here, both the code and results have been preserved in ***Appendix A.*** In-body figures display only partial output, omitting the underlying code. In contrast, the appendix offers a more complete view by presenting both the analytical code and its corresponding results.

Once predictive analysis was completed, key statistical results were included directly in the body of the paper (excluding code), with a separate appendix created to display the relevant code alongside full output. This approach ensures transparency and reproducibility while maintaining clarity in the main narrative.

This is an observational study that applies predictive statistical analysis in a non-experimental context to explain variation in median housing values (MEDV). The analysis is designed to identify and quantify the statistical significance of selected predictors, crime rate, socioeconomic status, and education quality. Methods include multiple linear regression, Welch’s ANOVA, and Spearman rank correlation. It is important to note that while the regression and correlation analyses may reveal statistically significant associations, they do not imply causation.

3.1 Dataset Summary

- Dataset: Boston Housing
  - Rationale: The Boston Housing dataset was selected due to its public availability and inclusion of a unique combination of variables that directly align with the study's research questions—specifically, the impact of crime rate, education quality, and socioeconomic status on median home values. By integrating these variables into a single, cohesive source, the dataset eliminates the need for cross-dataset merging, thereby enhancing consistency and reducing preprocessing complexity. Additionally, it offers a valuable snapshot of neighborhood conditions in the early 1970s, prior to the widespread adoption of the internet, capturing a pivotal moment of social and policy transformation in urban America.
- Sample Size: initially 14 variables across 506 observations, reduced to four key variables across 466 complete observations (see ***Appendix A***)
  - Note: Outliers were retained to preserve the integrity of the distributions and ensure the analysis remained accurate, relevant, and representative of the full dataset.
- Variables:
  - MEDV: Median value of owner-occupied homes in USD 1,000's
  - PTRATIO: Pupil–teacher ratio, (categorized to reflect school quality tiers)
  - LOG_CRIM: Log-transformed per capita crime rate
    - Note: CRIM was highly right-skewed and log-transformed to improve distributional symmetry.

- LSTAT: Percentage of lower-status population (used as a proxy for neighborhood-level economic disadvantage)

During the data cleaning and variable reduction phase, the dataset was refined to emphasize socially relevant, policy-informative neighborhood characteristics while maintaining model simplicity. Several features were removed, these variables, while potentially predictive, either overlapped with structural or demographic proxies or represented hyper-localized and geographically unique attributes that reduce generalizability.

This targeted reduction supports a framework centered on socioeconomic equity and focuses the analysis on crime, education quality, and economic status. The retained variables offer actionable insight into urban planning and housing policy, making the results more generalizable across diverse urban settings without reinforcing demographic stereotypes or including proxies that could bias interpretation.

4.0 Research Questions and Hypotheses

This section presents research questions (RQs) along with the statistical analysis methods used to test them. Each RQ outlines the dataset, key variables, hypothesis statements, rationale for the selected method, visualizations used to illustrate findings, and any limitations. The order of the questions has been strategically arranged so that the multivariate regression (RQ1) appears first in analysis execution, allowing residual plots and diagnostics from that model to help verify assumptions before proceeding with the ANOVA and correlation tests in RQ2 and RQ3.

4.1. RQ1: Predicting Housing Prices — Question and Hypotheses

1. Research Question 1:
   - How well can neighborhood safety, education quality, and socioeconomic status collectively predict housing prices in Boston?
2. Null Hypothesis (H_0):
   - The combined predictors (CRIM, PTRATIO, and LSTAT [percentage of lower status of the population]) do not significantly explain the variance in median housing prices.
3. Alternate Hypothesis (H_1):
   - The combined predictors significantly explain the variance in median housing prices.

4.2 RQ2: Education Quality Effect — Question and Hypotheses

4. Research Question 2:
   - Is there a relationship between the quality of local schools and home values?
5. Null Hypothesis (H_0):
   - There is no significant difference in median housing prices across different PTRATIO (pupil–teacher ratio) groups.

6. Alternate Hypothesis (H_1):
   - At least one PTRATIO group has a significantly different median housing price.

4.3 RQ3: Crime Rate Influence — Question and Hypotheses

7. Research Question 3:
   - Does the crime rate in a neighborhood significantly influence median housing prices?
8. Null Hypothesis (H_0):
   - There is no significant monotonic relationship between crime rate and median housing prices.
9. Alternate Hypothesis (H_1):
   - There is a significant monotonic relationship between crime rate and median housing prices.

5.0 Literature Review

The literature review suggests that while many factors influence housing values, four key neighborhood themes consistently emerge: school quality, crime rates, community perceptions, and attitudes toward affordable housing developments. These factors not only shape the desirability of a neighborhood but are also embedded in policy decisions and public investment. Additionally, the Boston Housing dataset is particularly valuable for this research because it reflects census tract conditions recorded just four years after the establishment of the U.S.

Department of Housing and Urban Development (HUD). This context provides a unique snapshot of urban development before the implementation of major housing policy interventions. Notably, the Section 8 housing voucher program was introduced in 1974, four years after this dataset was collected, making this dataset a useful baseline for studying pre-intervention neighborhood dynamics.

5.1 Crime Rate — Housing Value

The literature consistently shows that higher neighborhood crime rates are associated with reduced housing values. Among the variables influencing homebuyer decisions, crime emerged as one of the most prominent concerns. A significant proportion of crime occurs after school hours, suggesting that juveniles are frequently involved. Timmer et al. (2024) highlight that a youth's likelihood of engaging in criminal activity is influenced not only by the neighborhood where they live but also by the neighborhood where their school is located, as students spend much of their time there. The authors stress the importance of structural context in crime prevention, stating, “It is imperative that juvenile justice policy grows beyond punishing and rehabilitating individual youth to consider more comprehensively the role of social structure in generating crime” (p. 225). This finding suggests that including crime rate as a predictor variable of housing values is highly relevant. These findings underscore the importance of including neighborhood crime rate as a key predictor variable in housing value models.

5.2 Education Quality — Economic Development

School quality emerged as another consistent theme in literature related to housing values. In a study examining the effects of school district rezoning (consolidation), researchers found a strong link between educational quality and home prices. Collins and Kaplan (2017) reported that “a one standard deviation increase in school quality increases predicted housing prices by about 3 percent… homes rezoned to a municipal district experienced a 5–7 percent increase in price, holding school quality constant” (p. 632). These results indicate that both actual and perceived changes in educational opportunities can significantly influence market value. This evidence strongly supports the inclusion of education-related metrics, such as pupil–teacher ratios, as relevant predictor variables in housing value analyses.

5.3. Socioeconomic Status — Neighborhood Impact

While HUD’s Section 8 program provides rental assistance to tenants, another important program, the Low-Income Housing Tax Credit (LIHTC), offers tax incentives to developers to build housing with rent caps based on area median income (AMI). Many prospective homeowners avoid buying properties near LIHTC sites out of concern that their home values will decline. However, empirical findings challenge this assumption. Diamond and McQuade (2019) concluded that, “In lower-income areas, house prices appreciate substantially over the long run in response to the introduction of affordable housing projects… [and] LIHTC development leads to a reduction in both violent and property crime” (p. 1114). These results indicate that well-managed affordable housing can contribute positively to neighborhood conditions and long-term property values.

5.4 Support Summary — Implications

The findings from the literature review strongly support the inclusion of crime rate (CRIM), school quality (PTRATIO), and socioeconomic status (LSTAT) as key predictor variables for housing value (MEDV). These factors were consistently cited across studies and may even interact with one another, given their interconnected roles in shaping neighborhood conditions. The recurring prominence of these themes suggests that each variable could also serve as a dependent variable in future research exploring the broader effects of public policy, education, and community safety. For example, Carlson et al. (2012) found that the Section 8 voucher program generated substantial net benefits by improving housing stability, enhancing educational outcomes, reducing crime, and lowering public costs through decreased homelessness. Further analysis of these relationships could inform the design and expansion of social programs, as well as support public education campaigns aimed at reducing stigma and misconceptions surrounding affordable housing initiatives.

6.0 Research Design

This section outlines the statistical analyses selected to address each of the three research questions (RQs). For each RQ, the corresponding method is described in detail, including the variables involved, the rationale for choosing the statistical test, the test equation, and the visualization(s) used to support interpretation of results. Rationale for each visualization is also provided, ensuring transparency in how analytical decisions support the study's objectives.

6.1 RQ1: Predicting Housing Prices — Multivariate Analysis

10. RQ1 Variables:
    - MEDV (median value of owner-occupied homes in USD 1,000's)
    - PTRATIO (pupil–teacher ratio, categorized)
    - LOG_CRIM (log-transformed per capita crime rate)
    - LSTAT (Lower Status Population %)
11. RQ1 Test Type:
    - Multiple Linear Regression
    - Equation:
        - $MEDV=\beta_0+\beta_1 \cdot LOG\_CRIM+\beta_2 \cdot PTRATIO+\beta_3 \cdot LSTAT+\epsilon$
        - Where:
            - β_0: Intercept
            - β_1: Coefficient for LOG_CRIM
            - β_2: Coefficient for PTRATIO
            - β_3: Coefficient for LSTAT
            - ϵ: Error term (residual)
    - Test Rationale:
        - Determine the influence of multiple variables on median home value.
12. RQ1 Visualization:
    - Coefficient Plot and Residual Plots
    - Visual Rationale:
        - Display direction and strength of each predictor. The Residual Plots for evaluating model assumptions.

6.2 RQ2: Education Quality Effect — Welch's ANOVA

13. RQ2 Variables:
    - MEDV
    - PTRATIO (school quality proxy)

14. RQ2 Test Type:
    - Welch's ANOVA (with Levene's Test)
    - Equation:
        - $MEDV_{ij\,=}\,\mu+\alpha_i+\epsilon_{ij}$
        - Where:
            - $MEDV_{ij}$: House value observation $_j$ in group $_i$
            - μ: Overall mean house value
            - α_i: Effect of $_i$-th school quality group tier
            - ϵ_{ij}: Random error
    - Test Rationale:
        - Comparing the mean home values across multiple school quality tiers to determine if significant differences exist.

15. RQ2 Visualization:
    - Least Squares Means (LS-Means) Plot
    - Pairwise Comparisons Plot
    - Visual Rationale:

    - Displays least squares means of MEDV by tier and highlight statistically significant differences across categories.

6.3 RQ3: Crime Rate Influence — Spearman Rank Correlation

16. RQ3 Variables:
    - MEDV
    - LOG_CRIM (Log-transformed crime rate)
17. RQ3 Test Type:
    - Spearman Rank Correlation
    - Equation:
        - ρ_s = Spearman(Rank(LOG_CRIM),Rank(MEDV))
        - Where:
            - ρ_s: Spearman's rank correlation coefficient
    - Test Rationale:
        - Determine the strength and direction of a relationship (monotonic) between a MEDV and LOG_CRIM.
18. RQ3 Visualization:
    - Boxplot
    - Bar Chart (grouped LOG_CRIM tiers)
    - Visual Rationale:
        - Display distribution and spread of MEDV across LOG_CRIM tiers. A boxplot for summarizing MEDV means across tiers.

7.0 Limitations

19. RQ1 Limitation:
    - This model relies on key assumptions, including linear relationships, independent residuals, constant variance (homoscedasticity), normally distributed residuals, and the absence of multicollinearity among predictors. Violations of these assumptions may affect the model's accuracy and interpretability. Residual plots were used to assess whether these assumptions are reasonably satisfied.
20. RQ2 Limitation:
    - Categorizing PTRATIO introduces a simplified structure that may reduce the precision of continuous data, especially if group cutoffs are not grounded in external standards.
21. RQ3 Limitation:
    - Grouping LOG_CRIM into categories may obscure subtler patterns found in the continuous data. While useful for comparison, this simplification can reduce granularity.

While planning helped structure a clear and focused analysis, limitations related to model assumptions, potential confounders, and variable simplification remain. The selected predictors, crime rate, education quality, and poverty, are likely interrelated. For instance, neighborhoods with higher poverty levels may also experience elevated crime rates and reduced school quality.

Although each variable contributes unique policy-relevant insight, their shared variance must be considered when interpreting the model's results and drawing conclusions about their individual effects.

8.0 Ethical Considerations

The Boston Housing dataset has long served as a foundational resource in academic and professional research analyzing factors that influence housing markets, economic development, and educational quality. While widely used, prior discourse has highlighted that some variables in this dataset may oversimplify complex neighborhood dynamics, potentially leading to misinterpretation or misuse (Scikit-learn developers, 2021). Such attributes could inadvertently reinforce stereotypes or be exploited in discriminatory practices involving real estate, zoning, lending, or policy decisions.

To mitigate these concerns, sensitive variables were included in this project only when they could be ethically justified as relevant to the research focus, specifically, the relationship between housing values and neighborhood characteristics such as crime rates, socioeconomic status, and education quality. No personally identifiable information (PII) is present in the dataset, and all records are anonymized and publicly available, reducing privacy risks and supporting responsible, ethical analysis.

During preprocessing, the crime rate variable was log-transformed to address skewness, and statistical assumptions were evaluated to ensure methodological soundness. Analytical methods, such as Welch's ANOVA and Spearman Rank Correlation, were selected to account for unequal variances, non-normal distributions, and monotonic relationships, where appropriate. Interpretation was purposefully framed to avoid demographic generalizations, focusing instead on structural factors with implications for public resource allocation.

Finally, all steps were transparently documented with annotated screenshots to ensure reproducibility and accountability. By centering the analysis on socioeconomic indicators and excluding identity-based traits, this project upholds ethical best practices while contributing meaningful insights to support equitable housing policy, public safety planning, and education reform.

9.0 Summary and Next Steps

The project so far demonstrates how the Six-Step Research Process and the SMART framework were applied to structure the analysis of how neighborhood characteristics influence housing values using the Boston Housing dataset. The dataset has been described, research questions and hypotheses have been outlined, the research design has been detailed, and a supporting literature review has been presented.

The next step in this project was to perform the statistical analyses described using Jupyter Notebook. The results were used to test each hypothesis by determining whether to reject or fail to reject the corresponding null hypotheses. Statistically significant relationships were identified and interpreted, providing a foundation for evidence-based recommendations related to housing development, public policy, and urban planning.

10.0 Findings

This section presents an overview of the results for each analysis conducted in the study. Key data visualizations are included within the corresponding subsections below. For full details, including the Python code, libraries used, extended charts, and annotated visuals, please refer to the Appendices. The complete multivariate regression analysis for RQ1: Predicting Housing Prices is in ***Appendix B***. The Welch's ANOVA results for RQ2: Education Quality Effect are provided in ***Appendix C***, and the Spearman Rank Correlation analysis for RQ3: Crime Rate Influence is documented in ***Appendix D***.

10.1 RQ1: Predicting Housing Prices — Findings

The multivariate linear regression (MLR) analysis was conducted to evaluate how well three predictors—LSTAT (percentage of lower-status population), PTRATIO (pupil–teacher ratio), and LOG_CRIM (log-transformed crime rate)—explain the variation in MEDV (median home value in USD 1,000s). The model accounted for approximately 61% of the variance in housing prices ($R^2 = 0.610$), indicating moderately strong explanatory power.

Two predictors, LSTAT ($p < 0.001$) and PTRATIO ($p < 0.001$), were found to be statistically significant negative predictors of home values. In contrast, LOG_CRIM ($p = 0.505$) was not statistically significant. This lack of significance may be due to multicollinearity or overlapping variance with LSTAT, which is a stronger socioeconomic indicator. Nonetheless, LOG_CRIM was retained in the model based on theoretical relevance to neighborhood safety.

The Residuals vs. Fitted Values plot (see ***Figure 1***) suggests mild violations of linearity and homoscedasticity, particularly at the tails, but not to a degree that undermines model validity. The Coefficient Plot with 95% Confidence Intervals (see ***Figure 2***) visually reinforces the statistical findings: the intervals for LSTAT and PTRATIO are tight and do not cross zero, whereas the interval for LOG_CRIM spans zero, confirming its non-significance. The Q-Q Plot of Residuals (see ***Figure 3***) reveals minor deviations from normality in the tails, suggesting a few outliers but not severe skewness.

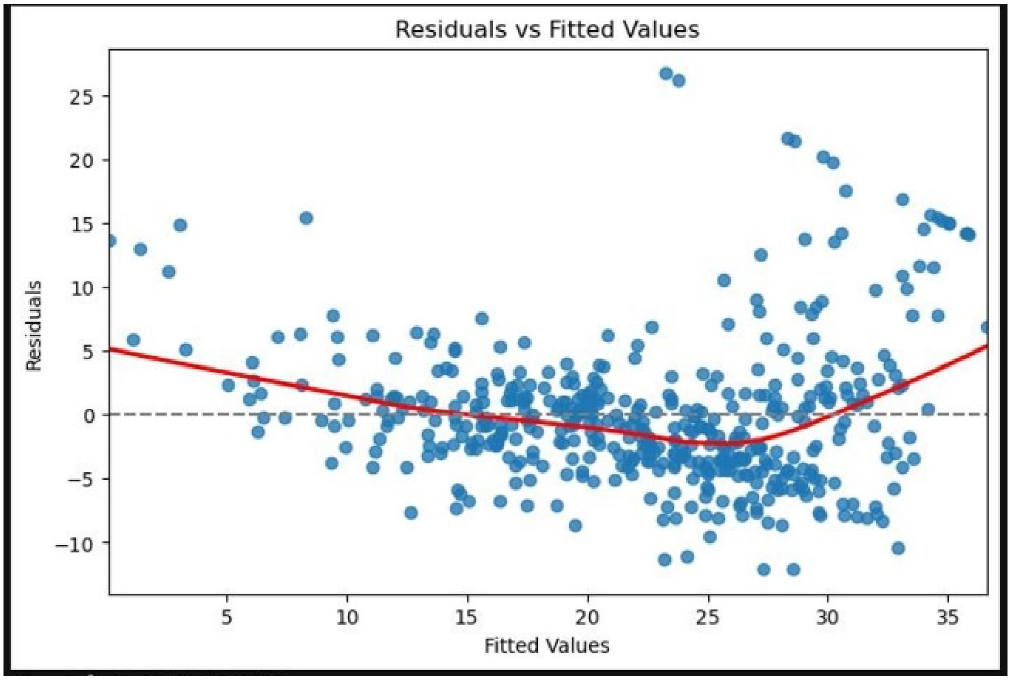


***Figure 1*. Q1 Residual Plot.** Residuals vs. fitted values with LOWESS curve.

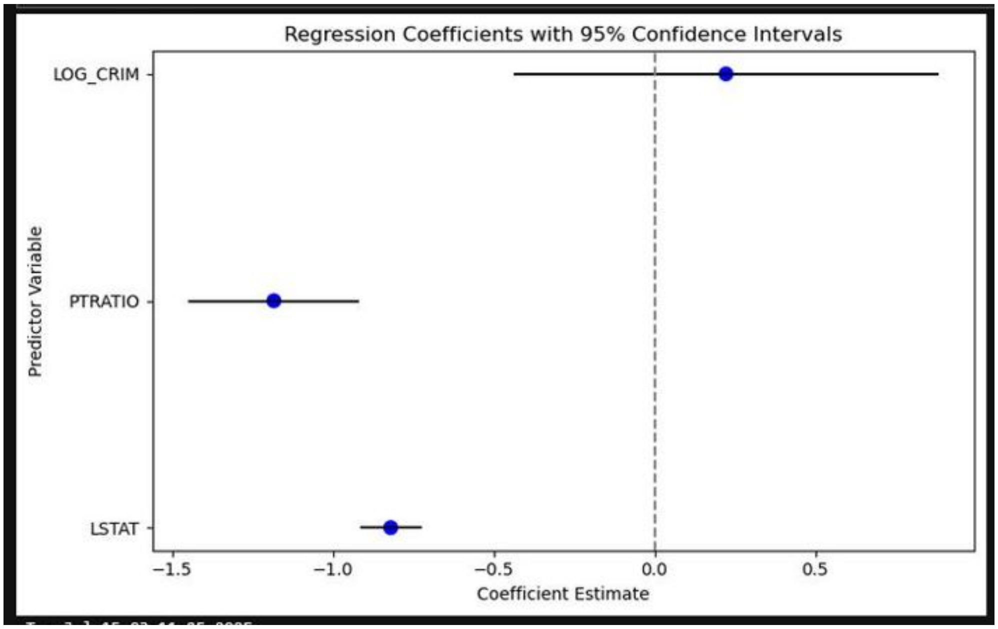


***Figure 2*. Q1 Coefficient Plot.** Coefficient estimates and 95% CIs for model predictors.

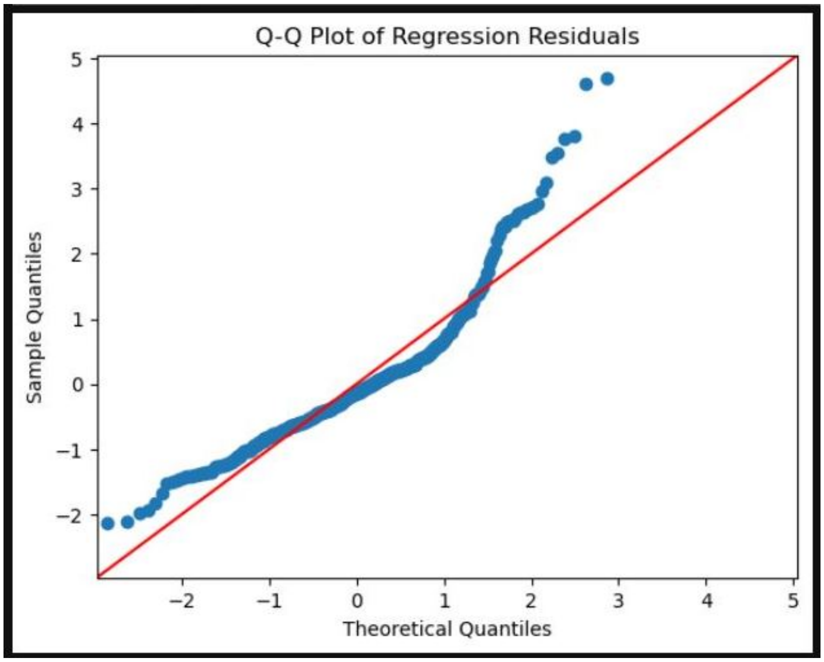


***Figure 3*. Q1 Q-Q Plot of Residuals.** Normal Q-Q plot for regression residuals.

Despite minor violations of regression assumptions, the model demonstrates strong performance and supports the hypothesis that socioeconomic status and education quality are significant predictors of housing values in Boston. Practically speaking, higher student-to-teacher ratios (indicating lower education quality) and greater concentrations of socioeconomically disadvantaged populations are both associated with lower median home values. Specifically, for every one-unit increase in PTRATIO, the median home value (MEDV) is expected to decrease by approximately $1,184, and for every one-percentage-point increase in LSTAT, MEDV is expected to decrease by roughly $820, assuming all other predictors remain constant.

These results support the Alternate Hypothesis (H_1), confirming that the combined predictors significantly explain the variance in median housing prices. While LOG_CRIM was not independently significant, the strong significance of PTRATIO and LSTAT, even when controlling for LOG_CRIM, indicates that education quality, socioeconomic status, and neighborhood safety collectively influence median home values in Boston.

10.2 RQ2: Education Quality Effect — Findings

To evaluate whether education quality impacts median housing values, the continuous variable PTRATIO (pupil–teacher ratio) was categorized into three tiers—Low, Medium, and High—representing varying levels of school quality. Preliminary assumption checks were conducted to determine the appropriate test. Group sizes were found to be uneven (Low = 158, Medium = 253, High = 55), and Levene's Test for homogeneity of variances indicated a significant violation (Levene's statistic ≈ 15.07, p ≈ 4.56e–07). Because these assumptions are critical to traditional ANOVA, the Welch's ANOVA, a more robust alternative for handling unequal variances and sample sizes, was used.

The results revealed a statistically significant difference in MEDV (median home value) across the PTRATIO tiers (F ≈ 61.41, p ≈ 2.22e–24), supporting the hypothesis that education quality is associated with housing values. As illustrated in the LS-Means Plot: MEDV by PTRATIO Tier (see ***Figure 4***), homes in Low PTRATIO neighborhoods (better school quality) had a mean MEDV of approximately $28,000, while Medium and High PTRATIO tiers had lower mean values of $20,000 and $18,000, respectively.

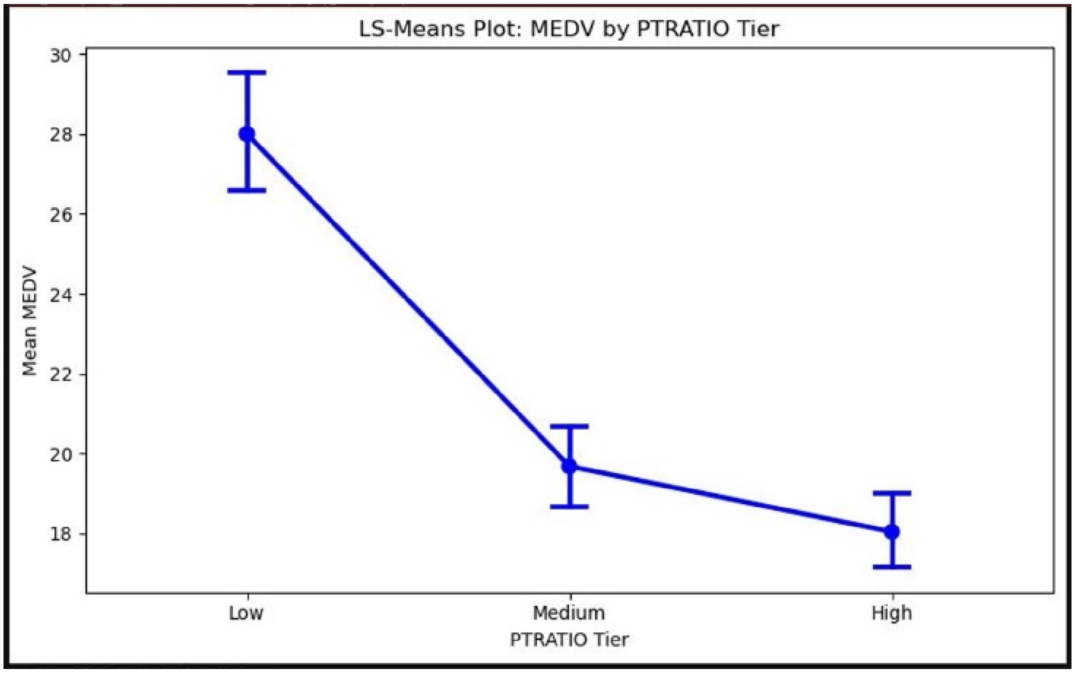


***Figure 4*. RQ2 LS-Means Plot.** Estimated MEDV means by PTRATIO tier with 95% CIs.

The 95% Confidence Intervals indicate that the differences between Low vs. Medium and Low vs. High are statistically significant, while the difference between Medium vs. High is comparatively smaller and potentially not significant.

The Pairwise Mean Differences in MEDV by PTRATIO Tier plot (see ***Figure 5***) confirms these findings. The mean difference between Low and Medium tiers is approximately $8,000, and between Low and High tiers, it is about $10,000—both statistically significant. However, the difference between Medium and High tiers is only around $2,000, with confidence intervals suggesting this comparison may not reach significance.

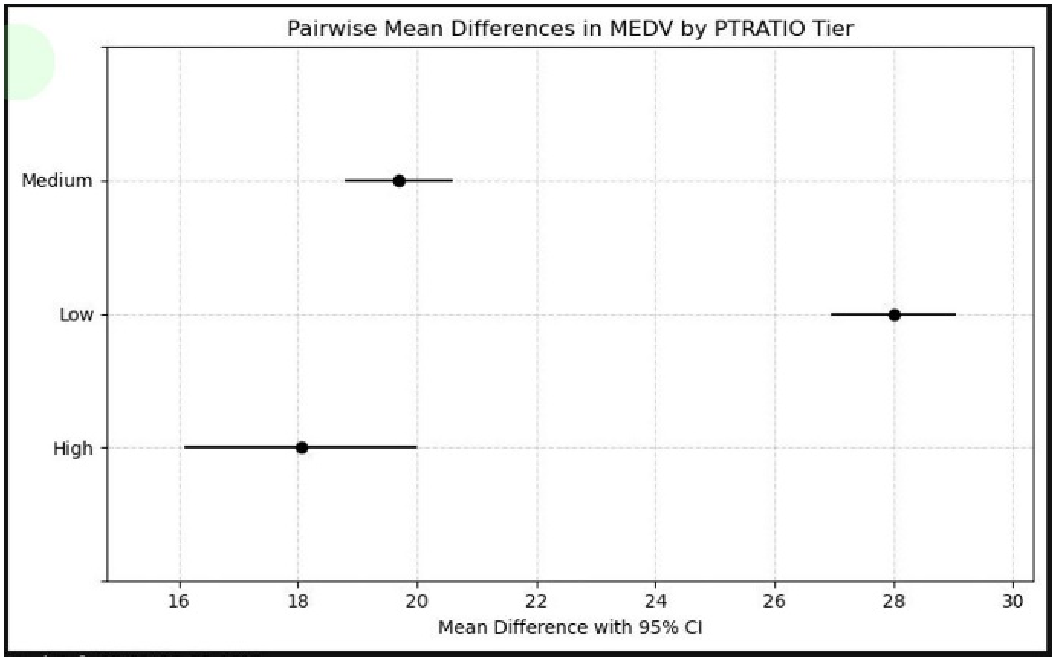


***Figure 5*. RQ2 Pairwise Comparisons Plot.** Tukey HSD mean differences in MEDV across tiers.

These results support the alternate hypothesis (H_1), confirming that at least one PTRATIO group differs significantly in terms of median housing price. Overall, the findings highlight education quality as a significant driver of home values in Boston's housing market. Specifically, better education quality, reflected by lower pupil–teacher ratios, is strongly associated with higher home values, while poorer education quality is linked to lower home values. It is also important to consider that areas with higher PTRATIOs may reflect smaller tax bases, fewer taxpayers, or communities in lower income brackets, which can limit the financial resources available to support public education. This interplay between local economic capacity and school quality may further contribute to disparities in housing values across neighborhoods.

10.3 RQ3: Crime Rate Influence — Findings

The Spearman Rank Correlation was selected to assess the monotonic relationship between crime rate and housing values. As a non-parametric test, Spearman's correlation is appropriate for evaluating relationships when variables, such as LOG_CRIM (log-transformed per capita crime rate), do not meet assumptions of normality and linearity. The test results (Spearman's rho $\rho = -0.5732$, $p \approx 4.73\text{e-}43$) indicate a strong and statistically significant negative monotonic relationship between crime rate and median housing values. In other words, as crime rates increase, housing prices tend to decrease.

To support the interpretation, LOG_CRIM was categorized into three tiers—Low, Medium, and High—for visualization (see ***Figure 6***). The Low Crime Tier exhibited the highest median MEDV (approximately $24,000), with greater variability (Interquartile range or IQR, around $21,000 to $31,000) and a wider spread (from roughly $11,000 to $45,000). Several high-end outliers extended this range, suggesting that higher-value homes tend to be in neighborhoods with lower crime rates.

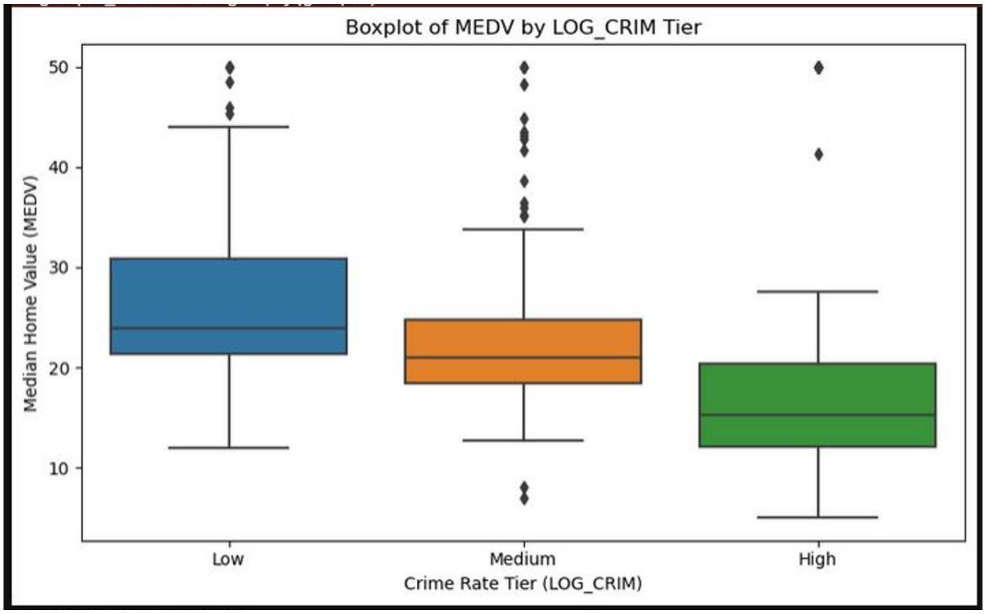


***Figure 6*. RQ3 Boxplot.** Distribution of MEDV across LOG_CRIM tiers.

The Medium Crime Tier showed a lower median MEDV (approximately $21,000) with moderate variability (IQR around $19,000 to $25,000) and a spread from approximately $11,000 to $33,000. This group also included more extreme outliers, reflecting moderate home values in neighborhoods with average crime rates.

In In contrast, the High Crime Tier exhibited the lowest median MEDV, approximately $16,000, with greater variability (IQR around $11,000 to $20,000) and a narrower overall spread, ranging from about $4,000 to $28,000. While a few high-end outliers were observed, most values clustered toward the lower end of the distribution, reinforcing the conclusion that higher crime rates are generally associated with lower housing values.

The Boxplot of MEDV by LOG_CRIM Tier (see ***Figure 7***) visually reinforces the downward trend in housing values as crime rates increase. The Low Crime Tier displays a mean MEDV near $27,000, followed by the Medium Crime Tier at approximately $23,000, and the High Crime Tier at about $18,000. These patterns align with the strong negative monotonic relationship identified in the Spearman Rank Correlation analysis. Together, the statistical and visual evidence support the alternate hypothesis (H_1), confirming that crime rate is significantly and inversely associated with median housing prices.

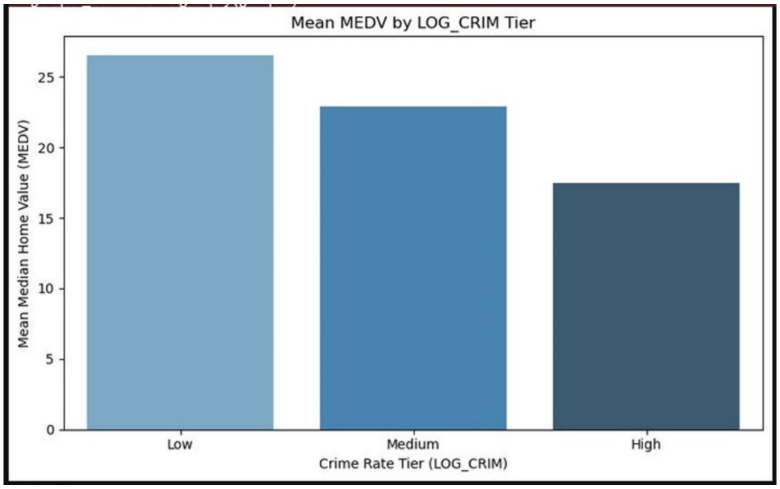


***Figure 7*. RQ3 Bar Chart.** Mean MEDV by LOG_CRIM tier highlighting trend.

11.0 Conclusion

This project confirmed that key neighborhood characteristics, socioeconomic status, education quality, and crime rate, each have a statistically significant impact on housing values. The multivariate linear regression (MLR) model showed that both education quality, represented by PTRATIO (pupil–teacher ratio), and socioeconomic status, represented by LSTAT (lower-status population percentage), were significant predictors of median housing price (MEDV), even when controlling for LOG_CRIM (log-transformed per capita crime rate).

Welch’s ANOVA further validated the influence of education quality, revealing a significant difference in median home values across PTRATIO tiers, with a potential value gap of up to $10,000 (unadjusted for inflation, based on 1970s housing values). While LOG_CRIM was not statistically significant in the MLR model, the Spearman Rank Correlation test identified a strong negative monotonic relationship between crime rate and housing values, suggesting that crime rate exerts a measurable influence, likely driven in part by indirect factors such as public perception, neighborhood reputation, and investment risk.

These findings align with the literature, which emphasizes the interconnected roles of education, crime, and socioeconomic structure in shaping housing markets. Studies reviewed highlighted that both objective measures (e.g., class size) and public perception of neighborhood conditions can significantly affect property values. Furthermore, the literature suggests that variables like housing value could also be studied as predictors of community development, crime reduction, and educational opportunity. Such reciprocal analyses could help inform future interventions, such as equitable school zoning or expanded low-income housing initiatives, though ethical considerations would be essential.

By focusing on structural rather than identity-based factors, this study avoided ethical pitfalls while maintaining transparency and reproducibility through detailed documentation. The integration of literature-based themes strengthened the argument for policy solutions targeting neighborhood equity. A well-educated, economically stable community may help reduce poverty and lower crime rates, ultimately reinforcing a positive feedback loop that elevates housing values and community well-being.

12.0 Recommendations

1. Invest in Equitable Education Resources:
   - Revise school zoning policies and funding structures to reduce class sizes, particularly in underserved neighborhoods. Consider consolidating smaller schools into larger, better-resourced districts where appropriate. Additionally, investing in remote learning technologies and AI-supported educational programs can help expand access to high-quality instruction across diverse communities.
2. Implement Community-Centered Crime Reduction Programs:
   - Support localized crime prevention strategies that prioritize neighborhood safety and youth engagement. Initiatives such as community policing, neighborhood watch programs, and after-school activities can foster trust, reduce juvenile crime, and generate long-term social and economic benefits.
3. Support Affordable Housing with Structural Protections:
   - Adopt policies that expand access to affordable housing, such as the Low-Income Housing Tax Credit (LIHTC) and Section 8, while embedding safeguards to promote long-term community stability. Public education

campaigns to reduce stigma, zoning incentives for inclusive development, and protections against displacement can help reduce resistance, foster economic integration, and support sustainable neighborhood growth.

4. Expand Longitudinal Research:
    - This dataset offers a valuable baseline for understanding neighborhood conditions prior to major policy interventions such as LIHTC and Section 8. Future research comparing later census tracts, after the implementation of these programs, could reveal how key features interact over time using time series analysis. Additionally, this dataset serves as a foundation for examining the influence of broader societal shifts, such as the adoption of the internet and artificial intelligence, allowing for comparisons before, during, and after these technological transformations.
5. Geographic Comparison:
    - Extend data collection to include a broader range of locations, thereby increasing the sample size and enhancing generalizability. Additionally, organize the data into homogeneous subsets to enable comparisons among areas with similar geographic characteristics (e.g., metropolitan areas, micropolitan regions, suburbs, rural communities, or culturally distinct zones).
6. Interaction Effects:
    - Investigate how variables such as socioeconomic status, crime rate, and education quality interact with one another and how these interactions jointly influence median housing values. Exploring these combined effects may reveal nuanced patterns not captured by examining each factor in isolation.

7. Causal Inference:
    - Move beyond correlation analysis by employing quasi-experimental methods to evaluate the causal impact of policy changes on neighborhood dynamics over time or apply instrumental variable techniques to address potential endogeneity in key predictors.
8. Reintroduce Excluded Variables:
    - Consider reintroducing previously excluded variables to explore potential patterns of socioeconomic or demographic stratification that may underlie observed disparities in neighborhood outcomes. This approach could provide deeper insight into structural influences on housing values and related policy implications.

References

Altaf, M. (2022). *Boston Housing Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/altavish/boston-housing-dataset

Anaconda, Inc. (2024). Anaconda Web App [Web-based application]. https://anaconda.com/app/

Carlson, D., Haveman, R., Kaplan, T., & Wolfe, B. (2012). *The Benefits and Costs of the Section 8 Housing Subsidy Program: A framework and estimates of first-year effects*. Journal of Policy Analysis and Management, 31(1), 61–85. https://doi.org/10.1002/pam.20624

Collins, C. A., & Kaplan, E. K. (2017). *Capitalization of School Quality in Housing Prices: Evidence From Boundary Changes in Shelby County, Tennessee*. American Economic Review: Papers & Proceedings, 107(5), 628–632. https://doi.org/10.1257/aer.p20171129

Diamond, R., & McQuade, T. (2019). *Who Wants Affordable Housing in Their Backyard? An Equilibrium Analysis of Low-income Property Development*. Journal of Political Economy, 127(3), 1063–1117. https://doi.org/10.1086/701354

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2024). NumPy user guide. NumPy Developers. https://numpy.org/doc/stable/user/index.html#user

McKinney, W. (n.d.). *Python for data analysis: Data wrangling with pandas, NumPy &*

Jupyter (3rd ed.). O'Reilly Media. https://platform.virdocs.com/read/2237633/423/#/4/2

Morad, O. (2025, July). *Module 7: Portfolio Project*. 25SD-MIS581. Canvas Student.

Polonsky, M. J., & Waller, D. S. (2018). *Designing and managing a research project: A business student's guide* (4th ed.). SAGE Publications.

Project Jupyter. (2024). Jupyter Notebook [Computer software]. https://jupyter.org

Python Software Foundation. (2024). Python 3 documentation (Version 3.13.5) [Documentation]. https://docs.python.org/3/contents.html

Scikit-learn developers. (2021). *sklearn.datasets.load_boston* [Documentation]. Scikit-learn v1.0. https://scikit-learn.org/1.0/modules/generated/sklearn.datasets.load_boston.html

Seabold, S., & Perktold, J. (2024). Statsmodels user guide (Version 0.14.4). Statsmodels. https://www.statsmodels.org/stable/user-guide.html

The Matplotlib Development Team. (2024). Quick start guide (Version 3.8) [Documentation]. Matplotlib. https://matplotlib.org/stable/users/explain/quick_start.html

The Pandas Development Team. (2024). User guide (Version 2.2.2) [Documentation]. pandas. https://pandas.pydata.org/docs/user_guide/index.html

Timmer, A., Lautenschlager, R., Antonaccio, O., Botchkovar, E. V., & Hughes, L. A.

(2024). *When Your School is in a 'Rough' Neighborhood: What Can Shield Youth from Crime and Delinquency?* American Journal of Criminal Justice, 49(1), 201–229. https://doi.org/10.1007/s12103-023-09748-2

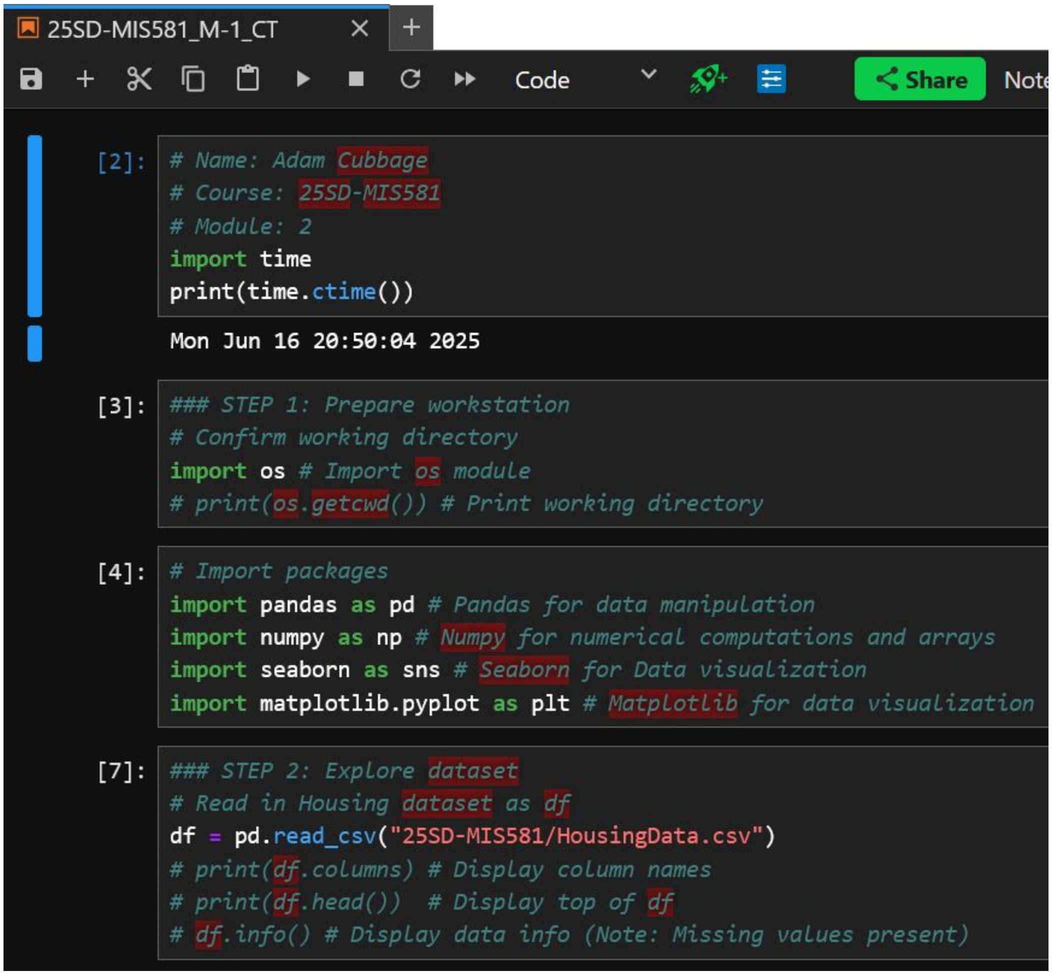
Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van der Walt, S. J. (2024). scipy.stats — Statistical functions (Version 1.13.0) [Documentation]. SciPy. https://docs.scipy.org/doc/scipy/reference/stats.html#

Waskom, M. (2024). Seaborn tutorial (Version 0.12.2) [Documentation]. Seaborn. https://seaborn.pydata.org/tutorial.html

Appendices

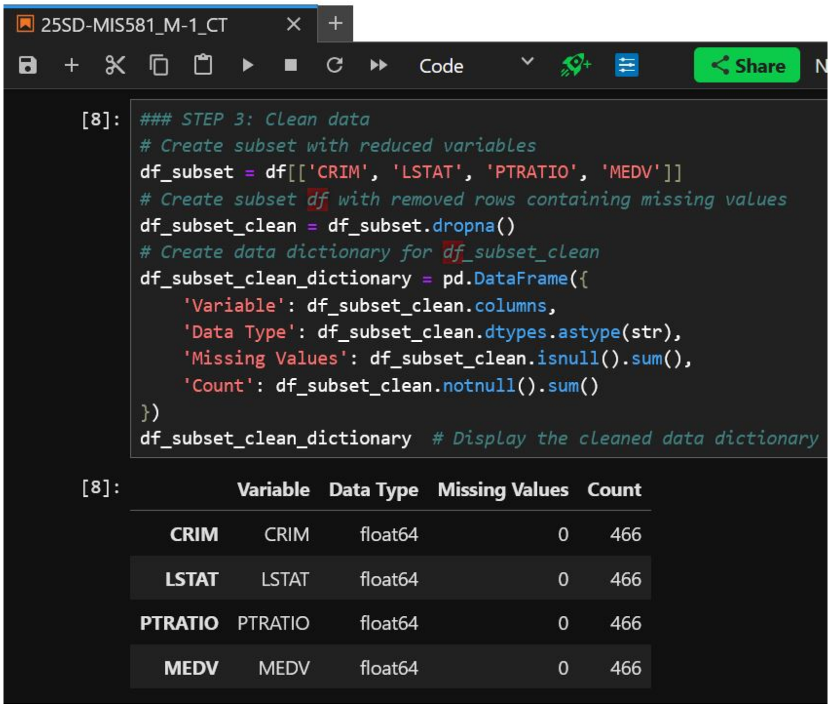
Appendix A: Boston Housing Dataset — Descriptive Analysis Script

Note: Descriptive analysis was conducted in accordance with best practices outlined in the official documentation of Python (Python Software Foundation, 2024) and its supporting libraries, including Matplotlib (The Matplotlib Development Team, 2024), Pandas (The Pandas Development Team, 2024), NumPy (Harris et al., 2024), SciPy (Virtanen et al., 2024), Seaborn (Waskom, 2024), and StatsModels (Seabold & Perktold, 2024), which were referenced to guide and validate each step of the analytical process.



```
[2]: # Name: Adam Cubbage
     # Course: 25SD-MIS581
     # Module: 2
     import time
     print(time.ctime())

     Mon Jun 16 20:50:04 2025

[3]: ### STEP 1: Prepare workstation
     # Confirm working directory
     import os # Import os module
     # print(os.getcwd()) # Print working directory

[4]: # Import packages
     import pandas as pd # Pandas for data manipulation
     import numpy as np # Numpy for numerical computations and arrays
     import seaborn as sns # Seaborn for Data visualization
     import matplotlib.pyplot as plt # Matplotlib for data visualization

[7]: ### STEP 2: Explore dataset
     # Read in Housing dataset as df
     df = pd.read_csv("25SD-MIS581/HousingData.csv")
     # print(df.columns) # Display column names
     # print(df.head())  # Display top of df
     # df.info() # Display data info (Note: Missing values present)
```

Figure 8. **Prepare Environment.** Code to import packages and load dataset.



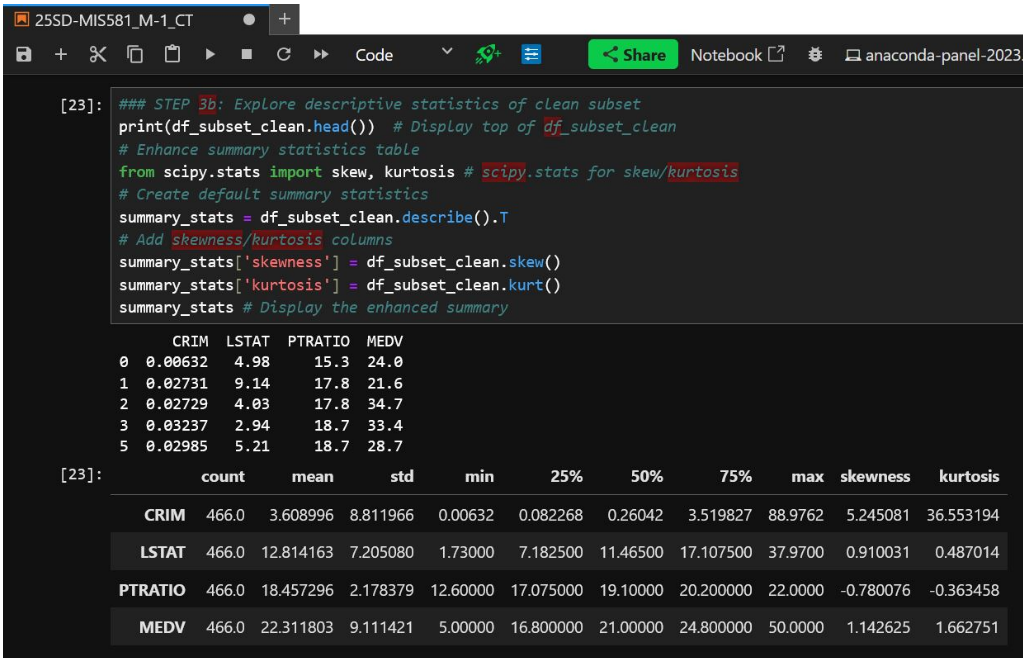
[8]:

```
### STEP 3: Clean data
# Create subset with reduced variables
df_subset = df[['CRIM', 'LSTAT', 'PTRATIO', 'MEDV']]
# Create subset df with removed rows containing missing values
df_subset_clean = df_subset.dropna()
# Create data dictionary for df_subset_clean
df_subset_clean_dictionary = pd.DataFrame({
    'Variable': df_subset_clean.columns,
    'Data Type': df_subset_clean.dtypes.astype(str),
    'Missing Values': df_subset_clean.isnull().sum(),
    'Count': df_subset_clean.notnull().sum()
})
df_subset_clean_dictionary  # Display the cleaned data dictionary
```

[8]:

	Variable	Data Type	Missing Values	Count
CRIM	CRIM	float64	0	466
LSTAT	LSTAT	float64	0	466
PTRATIO	PTRATIO	float64	0	466
MEDV	MEDV	float64	0	466

Figure 9. **Data Dictionary.** Code and output to clean data and display variable summary.



[23]:

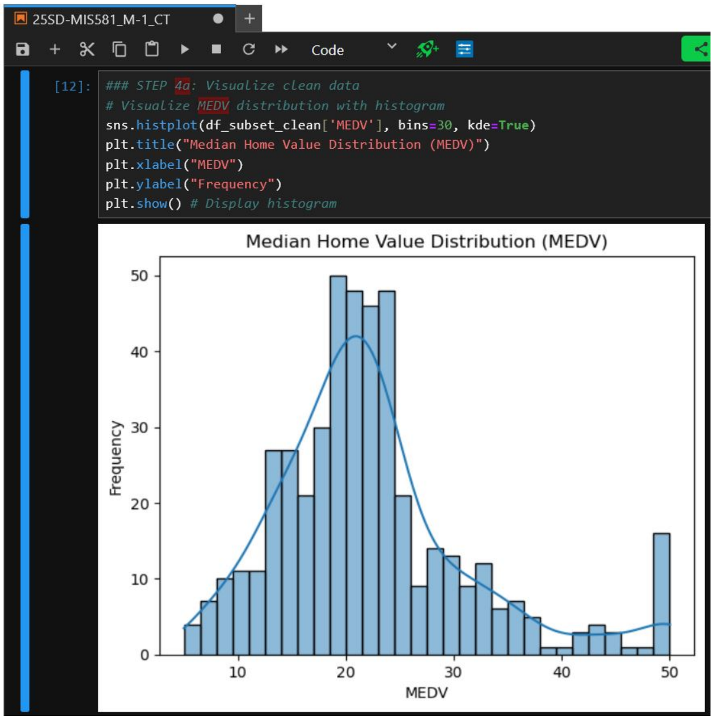
```
### STEP 3b: Explore descriptive statistics of clean subset
print(df_subset_clean.head())  # Display top of df_subset_clean
# Enhance summary statistics table
from scipy.stats import skew, kurtosis # scipy.stats for skew/kurtosis
# Create default summary statistics
summary_stats = df_subset_clean.describe().T
# Add skewness/kurtosis columns
summary_stats['skewness'] = df_subset_clean.skew()
summary_stats['kurtosis'] = df_subset_clean.kurt()
summary_stats # Display the enhanced summary
```

```
      CRIM  LSTAT  PTRATIO  MEDV
0  0.00632   4.98     15.3  24.0
1  0.02731   9.14     17.8  21.6
2  0.02729   4.03     17.8  34.7
3  0.03237   2.94     18.7  33.4
5  0.02985   5.21     18.7  28.7
```

[23]:

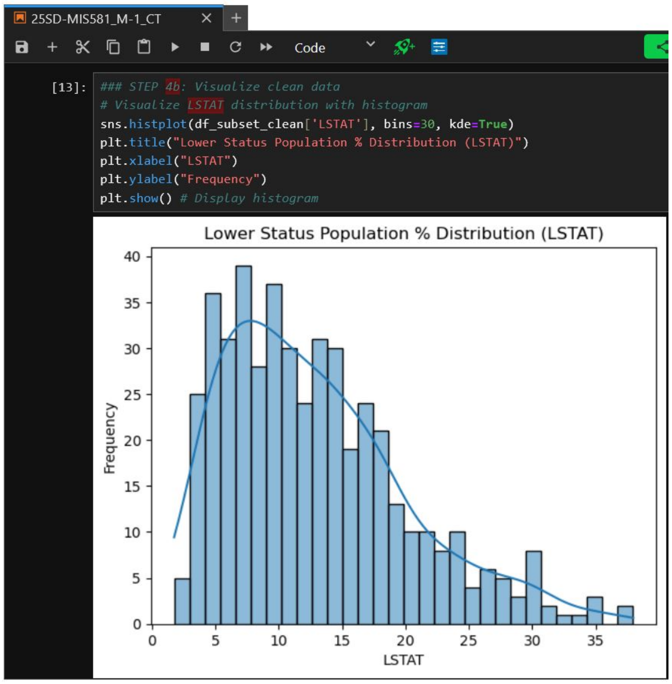
	count	mean	std	min	25%	50%	75%	max	skewness	kurtosis
CRIM	466.0	3.608996	8.811966	0.00632	0.082268	0.26042	3.519827	88.9762	5.245081	36.553194
LSTAT	466.0	12.814163	7.205080	1.73000	7.182500	11.46500	17.107500	37.9700	0.910031	0.487014
PTRATIO	466.0	18.457296	2.178379	12.60000	17.075000	19.10000	20.200000	22.0000	-0.780076	-0.363458
MEDV	466.0	22.311803	9.111421	5.00000	16.800000	21.00000	24.800000	50.0000	1.142625	1.662751

Figure 10. **Descriptive Statistics.** Code and output for summary stats and subset preview.



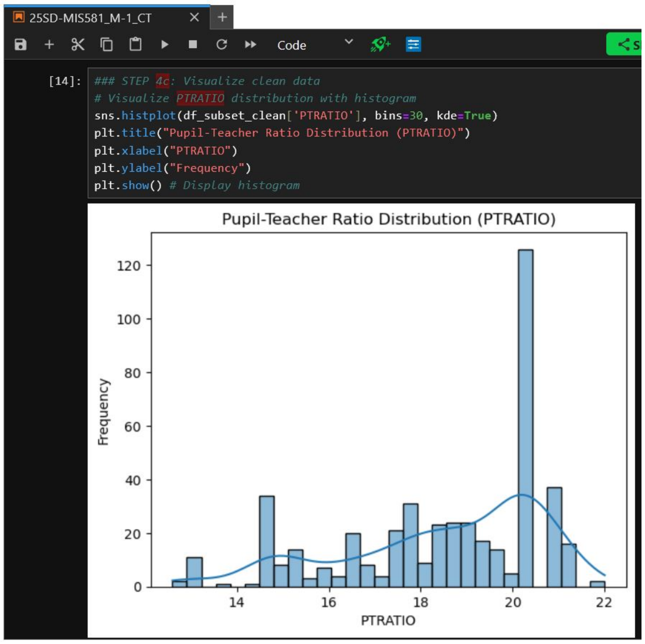
```
### STEP 4a: Visualize clean data
# Visualize MEDV distribution with histogram
sns.histplot(df_subset_clean['MEDV'], bins=30, kde=True)
plt.title("Median Home Value Distribution (MEDV)")
plt.xlabel("MEDV")
plt.ylabel("Frequency")
plt.show() # Display histogram
```

Figure 11. **MEDV Distribution.** Code and histogram of Median Home Value.



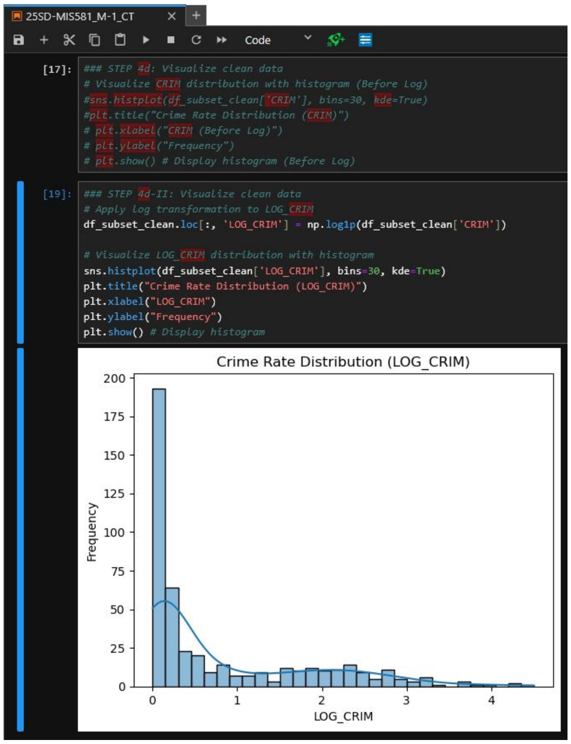
```
### STEP 4b: Visualize clean data
# Visualize LSTAT distribution with histogram
sns.histplot(df_subset_clean['LSTAT'], bins=30, kde=True)
plt.title("Lower Status Population % Distribution (LSTAT)")
plt.xlabel("LSTAT")
plt.ylabel("Frequency")
plt.show() # Display histogram
```

Figure 12. **LSTAT Distribution.** Code and histogram of Lower Status Population.



```
[14]: ### STEP 4c: Visualize clean data
      # Visualize PTRATIO distribution with histogram
      sns.histplot(df_subset_clean['PTRATIO'], bins=30, kde=True)
      plt.title("Pupil-Teacher Ratio Distribution (PTRATIO)")
      plt.xlabel("PTRATIO")
      plt.ylabel("Frequency")
      plt.show() # Display histogram
```

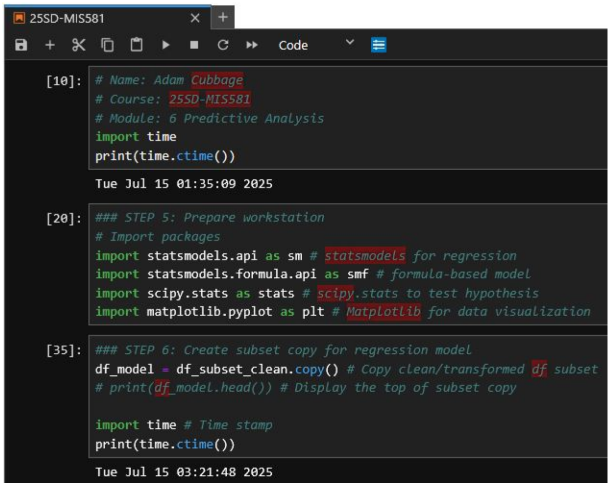
Figure 13. **PTRATIO Distribution.** Code and histogram of Pupil-Teacher Ratio.



```
[17]: ### STEP 4d: Visualize clean data
      # Visualize CRIM distribution with histogram (Before Log)
      #sns.histplot(df_subset_clean['CRIM'], bins=30, kde=True)
      #plt.title("Crime Rate Distribution (CRIM)")
      # plt.xlabel("CRIM (Before Log)")
      # plt.ylabel("Frequency")
      # plt.show() # Display histogram (Before Log)

[19]: ### STEP 4d-II: Visualize clean data
      # Apply log transformation to LOG_CRIM
      df_subset_clean.loc[:, 'LOG_CRIM'] = np.log1p(df_subset_clean['CRIM'])

      # Visualize LOG_CRIM distribution with histogram
      sns.histplot(df_subset_clean['LOG_CRIM'], bins=30, kde=True)
      plt.title("Crime Rate Distribution (LOG_CRIM)")
      plt.xlabel("LOG_CRIM")
      plt.ylabel("Frequency")
      plt.show() # Display histogram
```

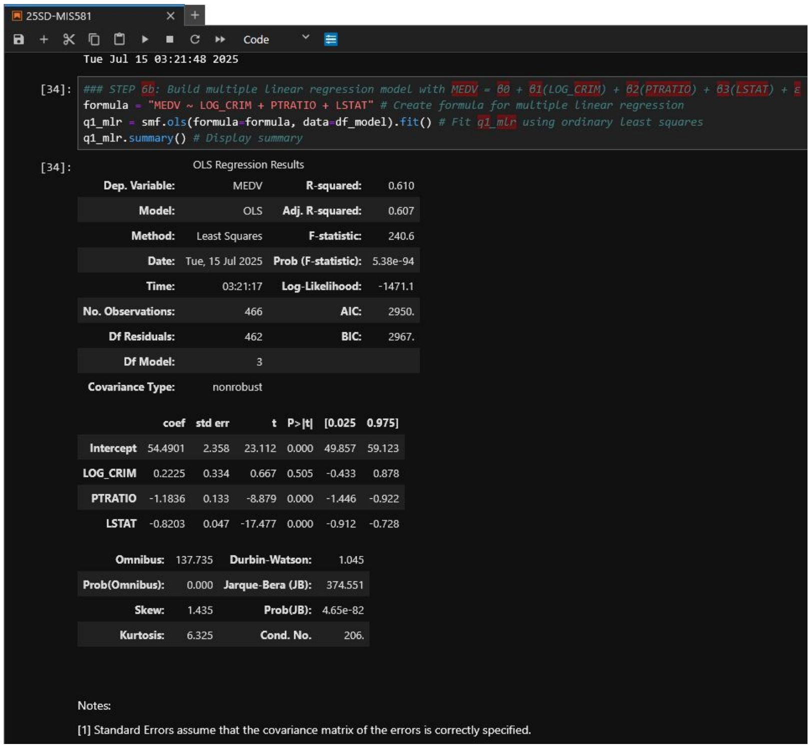
Figure 14. **LOG_CRIM Distribution.** Code and histogram of log-transformed Crime Rate.

Appendix B: RQ1: Predicting Housing Prices — Predictive Analysis Script



```
[10]: # Name: Adam Cubbage
      # Course: 25SD-MIS581
      # Module: 6 Predictive Analysis
      import time
      print(time.ctime())

      Tue Jul 15 01:35:09 2025

[20]: ### STEP 5: Prepare workstation
      # Import packages
      import statsmodels.api as sm # statsmodels for regression
      import statsmodels.formula.api as smf # formula-based model
      import scipy.stats as stats # scipy.stats to test hypothesis
      import matplotlib.pyplot as plt # Matplotlib for data visualization

[35]: ### STEP 6: Create subset copy for regression model
      df_model = df_subset_clean.copy() # Copy clean/transformed df subset
      # print(df_model.head()) # Display the top of subset copy

      import time # Time stamp
      print(time.ctime())

      Tue Jul 15 03:21:48 2025
```

***Figure 15*. RQ1 Prepare Environment.** Code to import packages and load dataset.



```
      Tue Jul 15 03:21:48 2025

[34]: ### STEP 6b: Build multiple linear regression model with MEDV = β0 + β1(LOG_CRIM) + β2(PTRATIO) + β3(LSTAT) + ε
      formula = "MEDV ~ LOG_CRIM + PTRATIO + LSTAT" # Create formula for multiple linear regression
      q1_mlr = smf.ols(formula=formula, data=df_model).fit() # Fit q1_mlr using ordinary least squares
      q1_mlr.summary() # Display summary

[34]:
```

OLS Regression Results

Dep. Variable:	MEDV	**R-squared:**	0.610
Model:	OLS	**Adj. R-squared:**	0.607
Method:	Least Squares	**F-statistic:**	240.6
Date:	Tue, 15 Jul 2025	**Prob (F-statistic):**	5.38e-94
Time:	03:21:17	**Log-Likelihood:**	-1471.1
No. Observations:	466	**AIC:**	2950.
Df Residuals:	462	**BIC:**	2967.
Df Model:	3		
Covariance Type:	nonrobust		

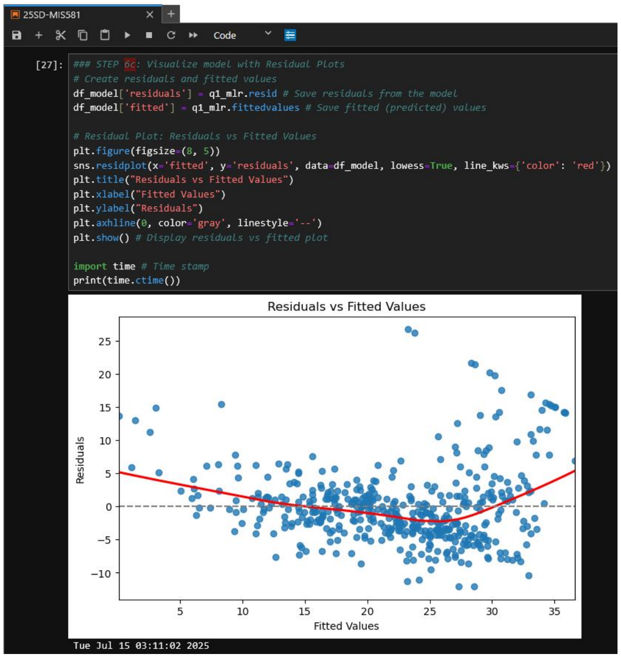
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	54.4901	2.358	23.112	0.000	49.857	59.123
LOG_CRIM	0.2225	0.334	0.667	0.505	-0.433	0.878
PTRATIO	-1.1836	0.133	-8.879	0.000	-1.446	-0.922
LSTAT	-0.8203	0.047	-17.477	0.000	-0.912	-0.728

Omnibus:	137.735	**Durbin-Watson:**	1.045
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	374.551
Skew:	1.435	**Prob(JB):**	4.65e-82
Kurtosis:	6.325	**Cond. No.**	206.

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

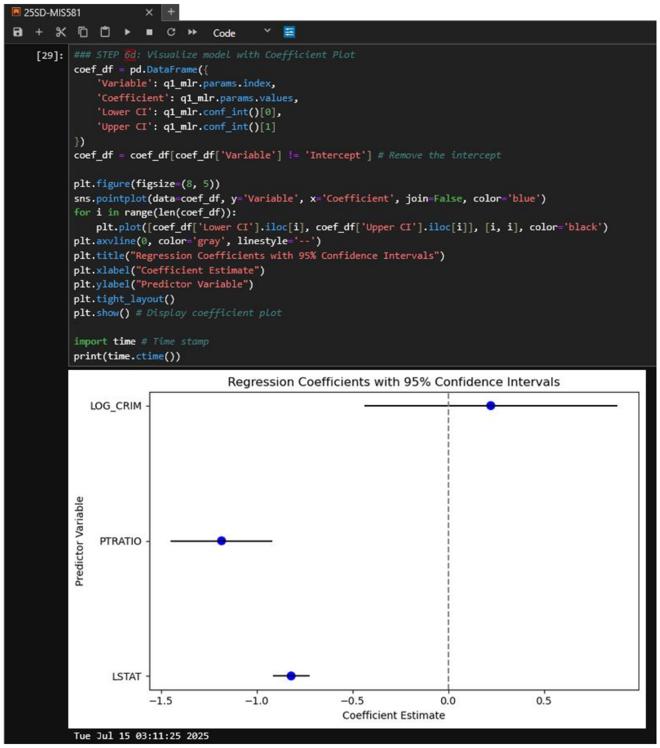
***Figure 16*. RQ1 Multiple Linear Regression Output.** OLS summary for MEDV ~ LOG_CRIM + PTRATIO + LSTAT.



```
[27]: ### STEP 6c: Visualize model with Residual Plots
      # Create residuals and fitted values
      df_model['residuals'] = q1_mlr.resid # Save residuals from the model
      df_model['fitted'] = q1_mlr.fittedvalues # Save fitted (predicted) values

      # Residual Plot: Residuals vs Fitted Values
      plt.figure(figsize=(8, 5))
      sns.residplot(x='fitted', y='residuals', data=df_model, lowess=True, line_kws={'color': 'red'})
      plt.title("Residuals vs Fitted Values")
      plt.xlabel("Fitted Values")
      plt.ylabel("Residuals")
      plt.axhline(0, color='gray', linestyle='--')
      plt.show() # Display residuals vs fitted plot

      import time # Time stamp
      print(time.ctime())
```

```
Tue Jul 15 03:11:02 2025
```

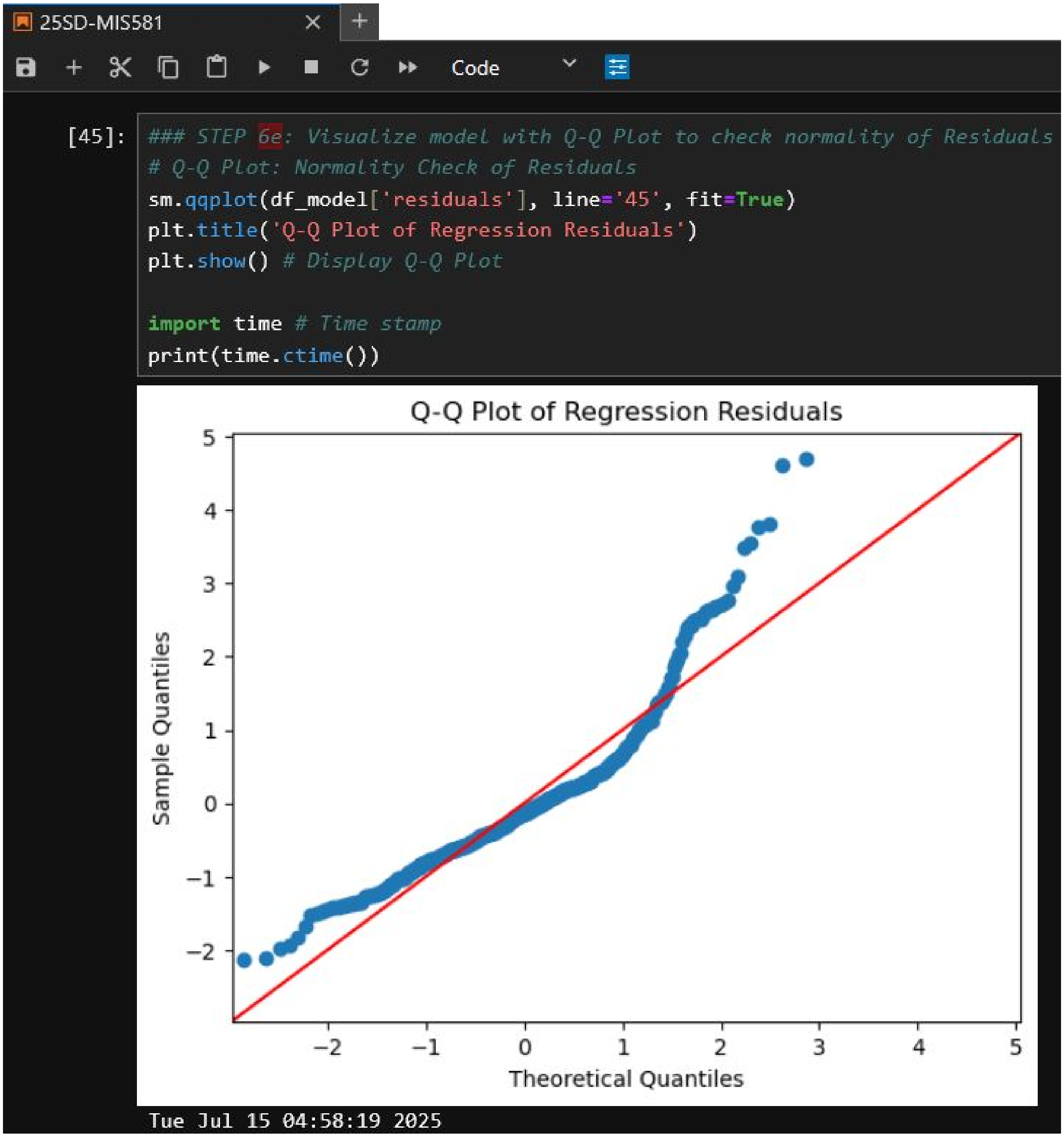
***Figure 17*. RQ1 Residual Plot II.** Residuals vs. fitted values with LOWESS curve.



```
[29]: ### STEP 6d: Visualize model with Coefficient Plot
      coef_df = pd.DataFrame({
          'Variable': q1_mlr.params.index,
          'Coefficient': q1_mlr.params.values,
          'Lower CI': q1_mlr.conf_int()[0],
          'Upper CI': q1_mlr.conf_int()[1]
      })
      coef_df = coef_df[coef_df['Variable'] != 'Intercept'] # Remove the intercept

      plt.figure(figsize=(8, 5))
      sns.pointplot(data=coef_df, y='Variable', x='Coefficient', join=False, color='blue')
      for i in range(len(coef_df)):
          plt.plot([coef_df['Lower CI'].iloc[i], coef_df['Upper CI'].iloc[i]], [i, i], color='black')
      plt.axvline(0, color='gray', linestyle='--')
      plt.title("Regression Coefficients with 95% Confidence Intervals")
      plt.xlabel("Coefficient Estimate")
      plt.ylabel("Predictor Variable")
      plt.tight_layout()
      plt.show() # Display coefficient plot

      import time # Time stamp
      print(time.ctime())
```

```
Tue Jul 15 03:11:25 2025
```

***Figure 18*. RQ1 Coefficient Plot II.** Coefficient estimates and 95% CIs for model predictors.

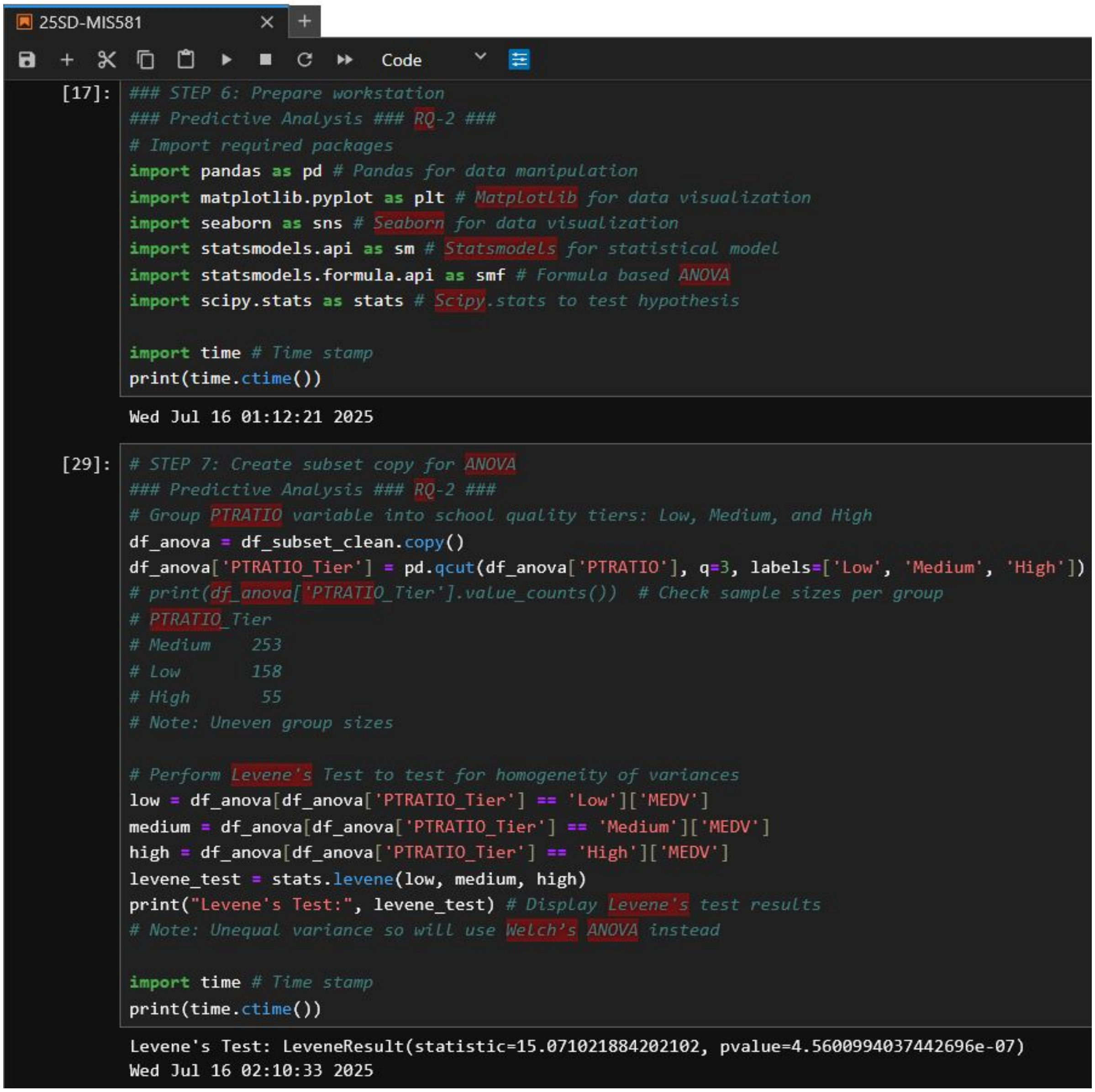


```
### STEP 6e: Visualize model with Q-Q Plot to check normality of Residuals
# Q-Q Plot: Normality Check of Residuals
sm.qqplot(df_model['residuals'], line='45', fit=True)
plt.title('Q-Q Plot of Regression Residuals')
plt.show() # Display Q-Q Plot

import time # Time stamp
print(time.ctime())
```

```
Tue Jul 15 04:58:19 2025
```

***Figure 19*. RQ1 Q-Q Plot of Residuals II.** Normal Q-Q plot for regression residuals.

Appendix C: RQ2: Education Quality Effect — Predictive Analysis Script



[17]:

```
### STEP 6: Prepare workstation
### Predictive Analysis ### RQ-2 ###
# Import required packages
import pandas as pd # Pandas for data manipulation
import matplotlib.pyplot as plt # Matplotlib for data visualization
import seaborn as sns # Seaborn for data visualization
import statsmodels.api as sm # Statsmodels for statistical model
import statsmodels.formula.api as smf # Formula based ANOVA
import scipy.stats as stats # Scipy.stats to test hypothesis

import time # Time stamp
print(time.ctime())
```

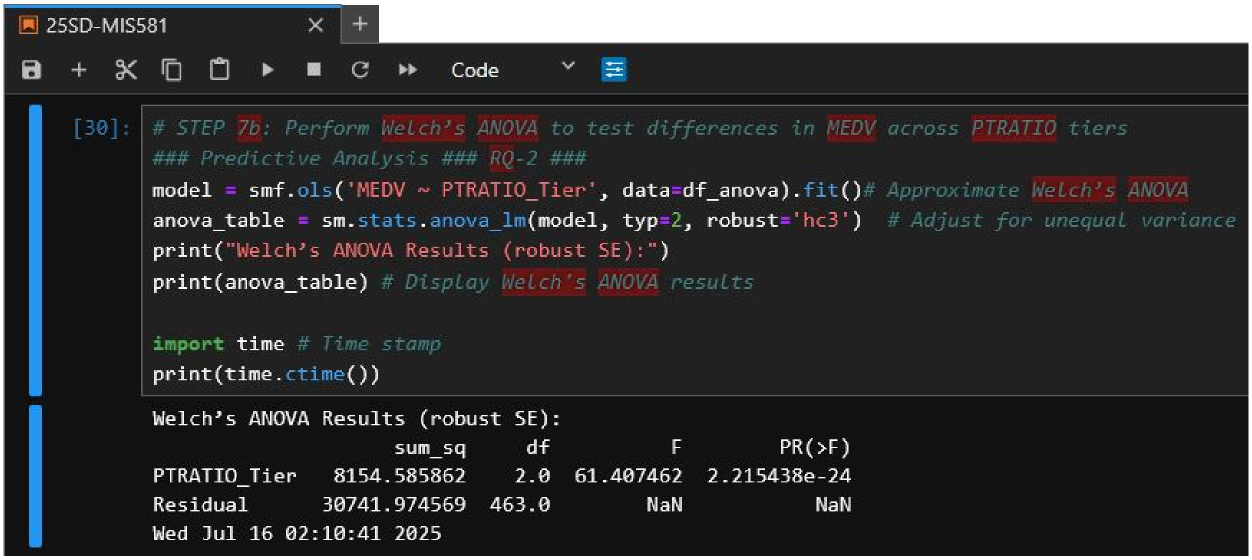
```
Wed Jul 16 01:12:21 2025
```

[29]:

```
# STEP 7: Create subset copy for ANOVA
### Predictive Analysis ### RQ-2 ###
# Group PTRATIO variable into school quality tiers: Low, Medium, and High
df_anova = df_subset_clean.copy()
df_anova['PTRATIO_Tier'] = pd.qcut(df_anova['PTRATIO'], q=3, labels=['Low', 'Medium', 'High'])
# print(df_anova['PTRATIO_Tier'].value_counts())  # Check sample sizes per group
# PTRATIO_Tier
# Medium    253
# Low       158
# High       55
# Note: Uneven group sizes

# Perform Levene's Test to test for homogeneity of variances
low = df_anova[df_anova['PTRATIO_Tier'] == 'Low']['MEDV']
medium = df_anova[df_anova['PTRATIO_Tier'] == 'Medium']['MEDV']
high = df_anova[df_anova['PTRATIO_Tier'] == 'High']['MEDV']
levene_test = stats.levene(low, medium, high)
print("Levene's Test:", levene_test) # Display Levene's test results
# Note: Unequal variance so will use Welch’s ANOVA instead

import time # Time stamp
print(time.ctime())
```

```
Levene's Test: LeveneResult(statistic=15.071021884202102, pvalue=4.5600994037442696e-07)
Wed Jul 16 02:10:33 2025
```

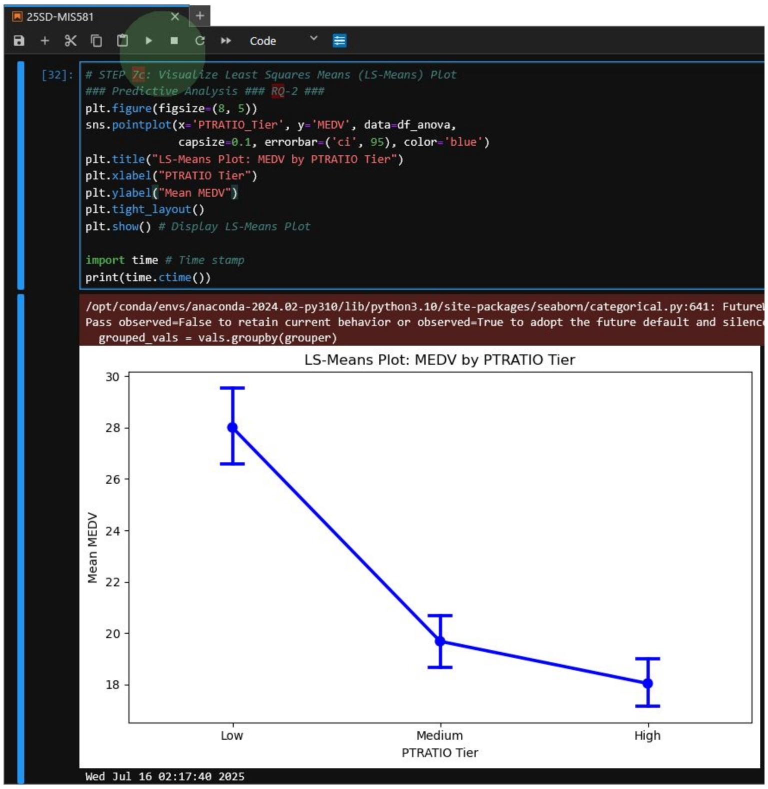
***Figure 20*. RQ2 Prepare Environment.** Import packages and prepare dataset and perform Levene’s Test.



```
[30]: # STEP 7b: Perform Welch’s ANOVA to test differences in MEDV across PTRATIO tiers
### Predictive Analysis ### RQ-2 ###
model = smf.ols('MEDV ~ PTRATIO_Tier', data=df_anova).fit()# Approximate Welch’s ANOVA
anova_table = sm.stats.anova_lm(model, typ=2, robust='hc3')  # Adjust for unequal variance
print("Welch’s ANOVA Results (robust SE):")
print(anova_table) # Display Welch’s ANOVA results

import time # Time stamp
print(time.ctime())
```

```
Welch’s ANOVA Results (robust SE):
                    sum_sq     df          F        PR(>F)
PTRATIO_Tier   8154.585862    2.0  61.407462  2.215438e-24
Residual      30741.974569  463.0        NaN           NaN
Wed Jul 16 02:10:41 2025
```

***Figure 21*. RQ2 Welch’s ANOVA Output.** OLS model using robust SE to test MEDV by PTRATIO.

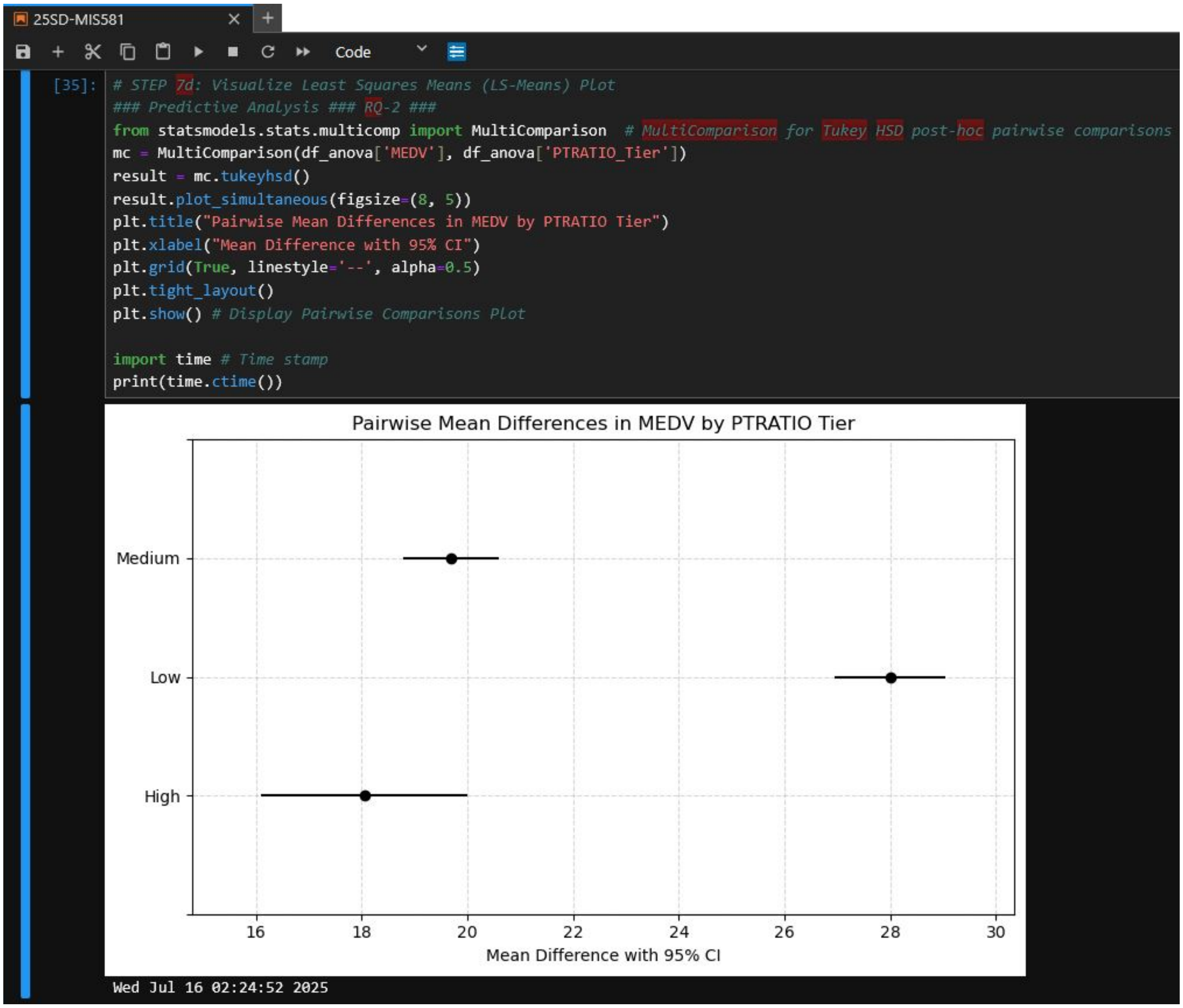


```
[32]: # STEP 7c: Visualize Least Squares Means (LS-Means) Plot
### Predictive Analysis ### RQ-2 ###
plt.figure(figsize=(8, 5))
sns.pointplot(x='PTRATIO_Tier', y='MEDV', data=df_anova,
              capsize=0.1, errorbar=('ci', 95), color='blue')
plt.title("LS-Means Plot: MEDV by PTRATIO Tier")
plt.xlabel("PTRATIO Tier")
plt.ylabel("Mean MEDV")
plt.tight_layout()
plt.show() # Display LS-Means Plot

import time # Time stamp
print(time.ctime())
```

```
/opt/conda/envs/anaconda-2024.02-py310/lib/python3.10/site-packages/seaborn/categorical.py:641: Future
Pass observed=False to retain current behavior or observed=True to adopt the future default and silenc
  grouped_vals = vals.groupby(grouper)
```

```
Wed Jul 16 02:17:40 2025
```

***Figure 22*. RQ2 LS-Means Plot II.** Estimated MEDV means by PTRATIO tier with 95% CIs.

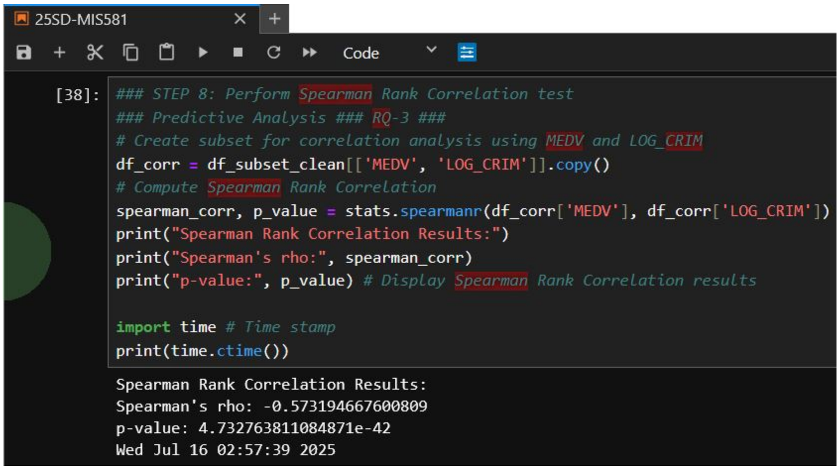

25SD-MIS581
Code
Pairwise Mean Differences in MEDV by PTRATIO Tier
Medium
Low
High
16
18
20
22
24
26
28
30
Mean Difference with 95% CI

```
[35]: # STEP 7d: Visualize Least Squares Means (LS-Means) Plot
      ### Predictive Analysis ### RQ-2 ###
      from statsmodels.stats.multicomp import MultiComparison  # MultiComparison for Tukey HSD post-hoc pairwise comparisons
      mc = MultiComparison(df_anova['MEDV'], df_anova['PTRATIO_Tier'])
      result = mc.tukeyhsd()
      result.plot_simultaneous(figsize=(8, 5))
      plt.title("Pairwise Mean Differences in MEDV by PTRATIO Tier")
      plt.xlabel("Mean Difference with 95% CI")
      plt.grid(True, linestyle='--', alpha=0.5)
      plt.tight_layout()
      plt.show() # Display Pairwise Comparisons Plot

      import time # Time stamp
      print(time.ctime())
```

```
Wed Jul 16 02:24:52 2025
```

***Figure 23*. RQ2 Pairwise Comparisons Plot II.** Tukey HSD mean differences in MEDV across tiers.

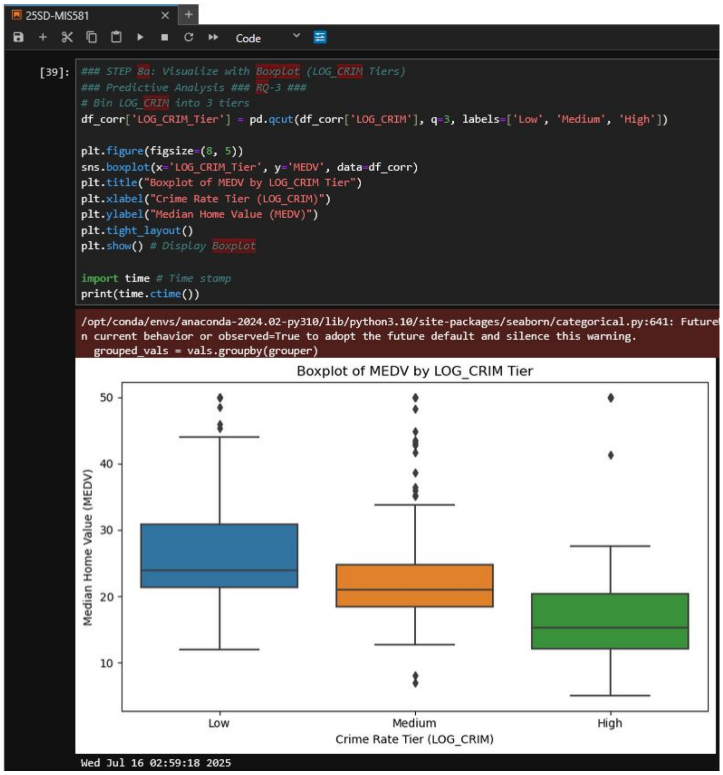
Appendix D: RQ3: Crime Rate Influence — Predictive Analysis Script



```
[38]: ### STEP 8: Perform Spearman Rank Correlation test
      ### Predictive Analysis ### RQ-3 ###
      # Create subset for correlation analysis using MEDV and LOG_CRIM
      df_corr = df_subset_clean[['MEDV', 'LOG_CRIM']].copy()
      # Compute Spearman Rank Correlation
      spearman_corr, p_value = stats.spearmanr(df_corr['MEDV'], df_corr['LOG_CRIM'])
      print("Spearman Rank Correlation Results:")
      print("Spearman's rho:", spearman_corr)
      print("p-value:", p_value) # Display Spearman Rank Correlation results

      import time # Time stamp
      print(time.ctime())
```

```
Spearman Rank Correlation Results:
Spearman's rho: -0.573194667600809
p-value: 4.732763811084871e-42
Wed Jul 16 02:57:39 2025
```

***Figure 24*. RQ3 Prepare Environment.** Create subset with MEDV and LOG_CRIM for and perform Spearman Test.

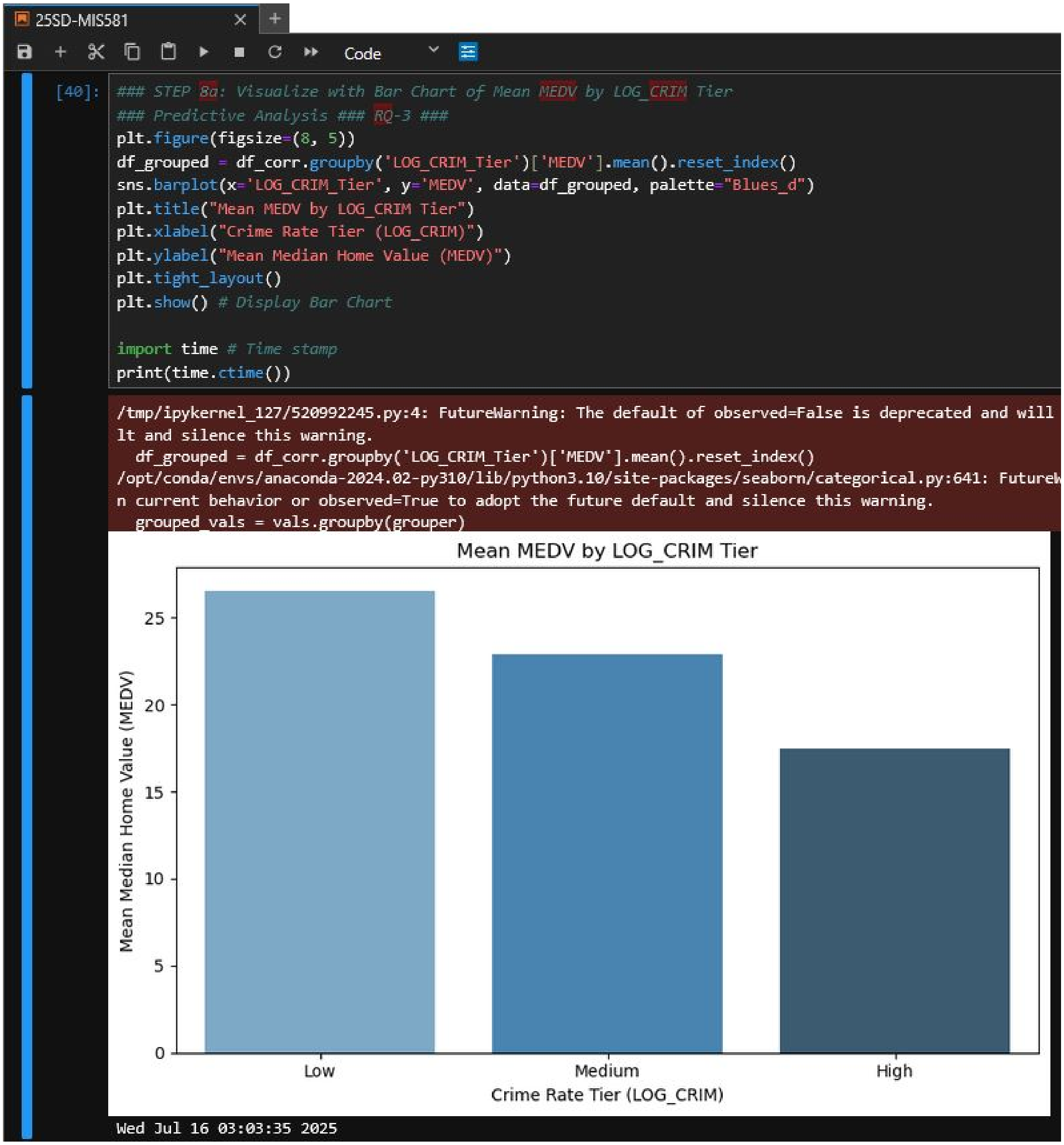


```
[39]: ### STEP 8a: Visualize with Boxplot (LOG_CRIM Tiers)
      ### Predictive Analysis ### RQ-3 ###
      # Bin LOG_CRIM into 3 tiers
      df_corr['LOG_CRIM_Tier'] = pd.qcut(df_corr['LOG_CRIM'], q=3, labels=['Low', 'Medium', 'High'])

      plt.figure(figsize=(8, 5))
      sns.boxplot(x='LOG_CRIM_Tier', y='MEDV', data=df_corr)
      plt.title("Boxplot of MEDV by LOG_CRIM Tier")
      plt.xlabel("Crime Rate Tier (LOG_CRIM)")
      plt.ylabel("Median Home Value (MEDV)")
      plt.tight_layout()
      plt.show() # Display Boxplot

      import time # Time stamp
      print(time.ctime())
```

```
/opt/conda/envs/anaconda-2024.02-py310/lib/python3.10/site-packages/seaborn/categorical.py:641: Future
n current behavior or observed=True to adopt the future default and silence this warning.
  grouped_vals = vals.groupby(grouper)
```

```
Wed Jul 16 02:59:18 2025
```

***Figure 25*. RQ3 Boxplot II.** Distribution of MEDV across LOG_CRIM tiers.



```
### STEP 8a: Visualize with Bar Chart of Mean MEDV by LOG_CRIM Tier
### Predictive Analysis ### RQ-3 ###
plt.figure(figsize=(8, 5))
df_grouped = df_corr.groupby('LOG_CRIM_Tier')['MEDV'].mean().reset_index()
sns.barplot(x='LOG_CRIM_Tier', y='MEDV', data=df_grouped, palette="Blues_d")
plt.title("Mean MEDV by LOG_CRIM Tier")
plt.xlabel("Crime Rate Tier (LOG_CRIM)")
plt.ylabel("Mean Median Home Value (MEDV)")
plt.tight_layout()
plt.show() # Display Bar Chart

import time # Time stamp
print(time.ctime())
```

```
/tmp/ipykernel_127/520992245.py:4: FutureWarning: The default of observed=False is deprecated and will
lt and silence this warning.
  df_grouped = df_corr.groupby('LOG_CRIM_Tier')['MEDV'].mean().reset_index()
/opt/conda/envs/anaconda-2024.02-py310/lib/python3.10/site-packages/seaborn/categorical.py:641: FutureW
n current behavior or observed=True to adopt the future default and silence this warning.
  grouped_vals = vals.groupby(grouper)
```

Mean MEDV by LOG_CRIM Tier
Mean Median Home Value (MEDV)
0
5
10
15
20
25
Low
Medium
High
Crime Rate Tier (LOG_CRIM)

```
Wed Jul 16 03:03:35 2025
```

***Figure 26*. RQ3 Bar Chart II.** Mean MEDV by LOG_CRIM tier highlighting trend.