Linear Regression Analysis

Trayvonious Pendleton

WGU

Course Number: D599

STUDENT ID: 011205284

***B1: Research Question***

How do Crime Rate and School Rating impact housing prices in a given area?

This question is relevant to urban planners, real estate agents, and home buyers who want to understand how safety (Crime Rate) and educational quality (School Rating) influence property values.

***B2: Goal of the Data Analysis***

This analysis aims to evaluate the relationship between Crime Rate, School Rating, and housing prices, helping stakeholders make informed decisions about property investments and urban development. This goal ensures that we analyze whether areas with lower crime rates and higher school ratings have significantly higher property values, which is crucial for pricing strategies and investment planning.

***C1: Identifying Dependent and Independent Variables***

**Dependent Variable:** Price (House price) → This is what we are predicting.

**Independent Variables:**

* CrimeRate (Crime rate in the area) → Used to evaluate if higher crime rates lead to lower housing prices.
* SchoolRating (School rating in the area) → Used to determine if better schools lead to higher housing prices.

**Justification for Variable Selection:**

* Crime rates and school ratings influence housing prices in real-world scenarios.
* Higher crime rates often correlate with lower property values.
* Better school ratings attract families, increasing demand and housing prices.

***C2: Descriptive Statistics***

A screenshot of a computer

AI-generated content may be incorrect.

***C3 Univariate and Bivariate Visualizations***

***Univariate Distribution***

***A graph of distribution of crime rate

AI-generated content may be incorrect.***

***Bivariate Scatterplots***

***A blue dots on a white background

AI-generated content may be incorrect.***

***D2: Training the Multiple Linear Regression Model***

***A screenshot of a computer

AI-generated content may be incorrect.***

A screenshot of a computer

AI-generated content may be incorrect.

***D3: Mean Squared Error***

***A screenshot of a computer

AI-generated content may be incorrect.***

***D4 Predictions on the Test Dataset***

***A screenshot of a computer program

AI-generated content may be incorrect.***

***E1: List and Justify Libraries Used***

The following libraries were used for the analysis:

**pandas** – for data manipulation and cleaning

**statsmodels** – for performing linear regression and statistical analysis

**scikit-learn** – for data splitting and evaluating model performance

**matplotlib (optional)** – for visualizations (used earlier in C3)

**Justification:**

**statsmodels** was essential for stepwise regression and statistical testing.

**sklearn** was used to split the dataset and compute model evaluation metrics like MSE.

**pandas** were used to manipulate and structure the dataset efficiently.

***E2: Method Used for Model Optimization***

The model was optimized using Backward Stepwise Elimination, where:

1. CrimeRate was removed due to its high p-value (0.3227, greater than 0.05).
2. The final model retained only SchoolRating, which significantly affected Price (p-value = 2.04e-199).

Why This Was Done:

* Removing CrimeRate improved model reliability by ensuring all variables were statistically significant.
* The final model was refitted after removing the insignificant variable.

***E3: Verification of Assumptions***

Assumptions checked:

* Linearity – Scatter plots in C3 confirmed a roughly linear relationship.
* Multicollinearity – Only two predictors were used, reducing multicollinearity risks.
* Homoscedasticity – The variance of residuals should be constant across all fitted values
  + Test Performed: A residuals vs fitted values plot was examined to detect patterns in residual variance.
  + No major increasing or decreasing spread was observed.
  + A diagram of a graph

    AI-generated content may be incorrect.
  + Conclusion: Residual variance is constant, meaning homoscedasticity holds.
* Normality of residuals – The residuals should follow a normal distribution to validate statistical inference.
  + Tests Performed: Histogram of Residuals and Q-Q Plot(Quantile-Quantile Plot) to visually assess normality.

A graph of a normality check

AI-generated content may be incorrect.

A graph with a red line

AI-generated content may be incorrect.

* + Conclusion: The residuals are approximately normal but exhibit slight skewness

***E4: Regression Equation***

After Optimization, the final regression equation is:

Price = 93227.33 + (31022.05 x SchoolRating)

Interpretation:

* A 1 unit increase in SchoolRating increases house price by $31,022
* CrimeRate was removed, meaning it did not have a statistical impact on the price

***E5: Model Metrics***

The final model performance is:

|  |  |
| --- | --- |
| *Metric* | *Value* |
| *R^2(training Set)* | ***0.1497*** |
| *Adjusted R^2* | ***0.1495*** |
| *F-Statistic* | ***985.57*** |

MSE Comparison:

|  |  |
| --- | --- |
| Dataset | MSE |
| Training Set | 19,465,568,630.58 |
| Test Set | 18,049,936,441.20 |

Key Takeaways:

* R^2 is 0.1497, meaning only 15% of price variance is explained by School Rating alone
* MSE is slightly lower on the test set, suggesting a reasonable model fit

***E6: Results and Implications***

* School Rating is a strong predictor of house price, with each unit increase raising prices by ~$31,022.
* Crime Rate was not a significant predictor, suggesting that other factors, such as location amenities or tax rates, may have a stronger influence.
* The model explains only 15% of the variance, indicating missing factors significantly impacting house prices.

***E7: Recommendation for the Organization***

1. Prioritize investments in areas with high-rated schools, as they have a strong influence on housing prices.
2. Gather additional data on property taxes, local amenities, and economic trends to improve model performance.
3. Ignore Crime Rate as a pricing factor based on the data, but consider verifying the dataset for accuracy.
4. Expand the model by adding additional features like distance to the city center, employment rates, or property tax rates to improve predictive power.