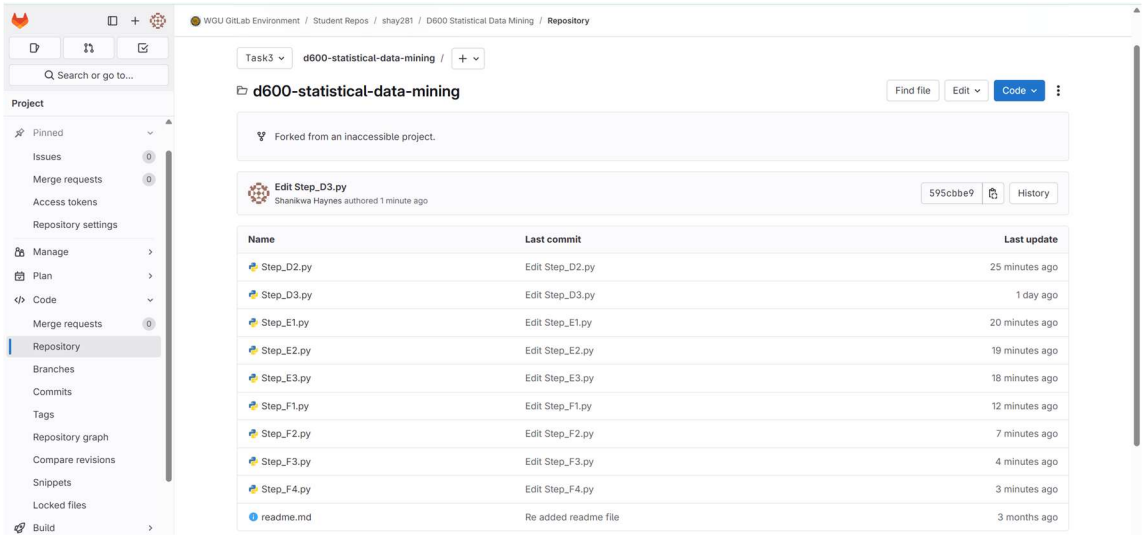
# D600 Task 3: Principal Component Analysis Report

## A. GitLab Repository

A subgroup and project were successfully created in GitLab for Task 3. The project was correctly cloned into the IDE, and commits were made after the completion of each rubric-aligned section. Each commit included a descriptive message and timestamp to ensure proper version tracking and traceability. The GitLab repository URL was submitted in the “Comments to Evaluator” section as instructed, and a full branch history was exported and included in the final submission.




## B1. Proposal of Question

The research question addressed in this project is: “How do structural characteristics such as square footage, number of bedrooms, backyard space, age of home, renovation quality, and previous sale price, along with neighborhood factors including crime rate, school rating, distance to city center, employment rate, property tax rate, local amenities, and transportation access, influence housing prices in the dataset?” This question reflects a real-world organizational concern relevant to property valuation and real estate analytics. The problem is suited for linear regression modeling because it examines the relationships between a continuous dependent variable (home price) and multiple quantitative explanatory variables.

## B2. Defined Goal

The goal of the analysis was to build a predictive model of home prices using principal component analysis (PCA) to reduce multicollinearity, followed by a linear regression model. This goal supports informed decision-making in real estate by identifying which components of housing data—such as location, size, or community amenities—have the greatest impact on price. The analysis remains within the scope of the dataset and is appropriately supported by the variables provided.

## C1. PCA Use

Principal component analysis was used to transform a set of correlated continuous variables into a smaller number of uncorrelated components. The PCA reduced redundancy in the data, resolved issues of multicollinearity, and improved the reliability of the subsequent linear regression model. The expected outcome of PCA was a reduced dataset where the retained components captured the majority of the total variance, allowing for effective and interpretable model training.

## C2. PCA Assumption

One fundamental assumption of PCA is that the principal components are uncorrelated linear combinations of the original variables. This assumption ensures that each component captures a unique dimension of the dataset's variance and contributes independently to the regression model. In this analysis, the assumption held, as the PCA transformed the original correlated variables into orthogonal components.

## D1. Variable Identification

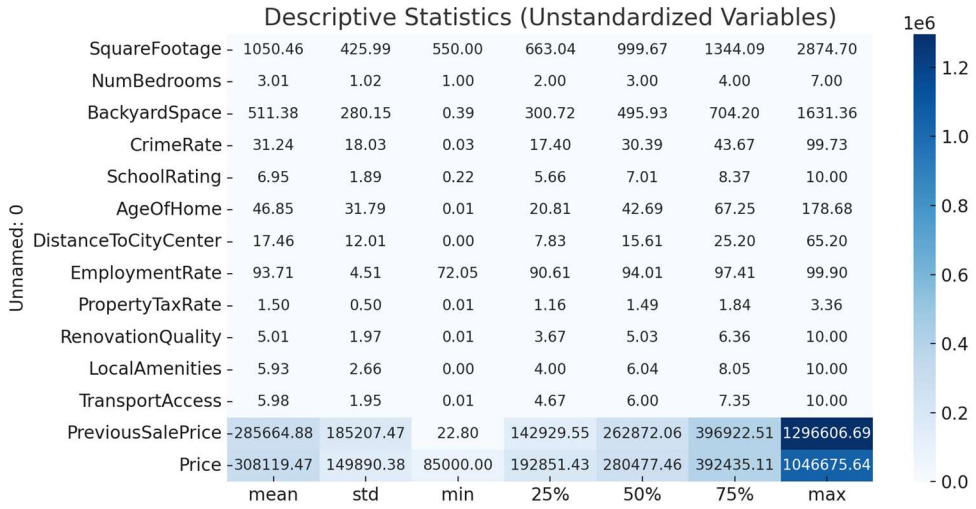
The continuous variables selected for PCA and regression included SquareFootage, NumBedrooms, BackyardSpace, CrimeRate, SchoolRating, AgeOfHome, DistanceToCityCenter, EmploymentRate, PropertyTaxRate, RenovationQuality, LocalAmenities, TransportAccess, and PreviousSalePrice. These variables were logically aligned with the research question and were confirmed to be continuous in nature. Non-continuous and problematic variables, such as NumBathrooms, were excluded from the analysis based on data quality concerns.

## D2. Standardized Data

Before performing PCA, all selected continuous explanatory variables were standardized using Scikit-learn's StandardScaler. This transformation ensured that each variable had a mean of zero and a standard deviation of one, which is essential for PCA to give equal weight to all features regardless of their original scale. The cleaned and standardized dataset was saved and submitted as required.

## D3. Descriptive Statistics

Descriptive statistics were generated for each variable prior to standardization. This included measures such as mean, standard deviation, minimum, and maximum values. The statistics provided a clear overview of each variable's distribution and central tendency.



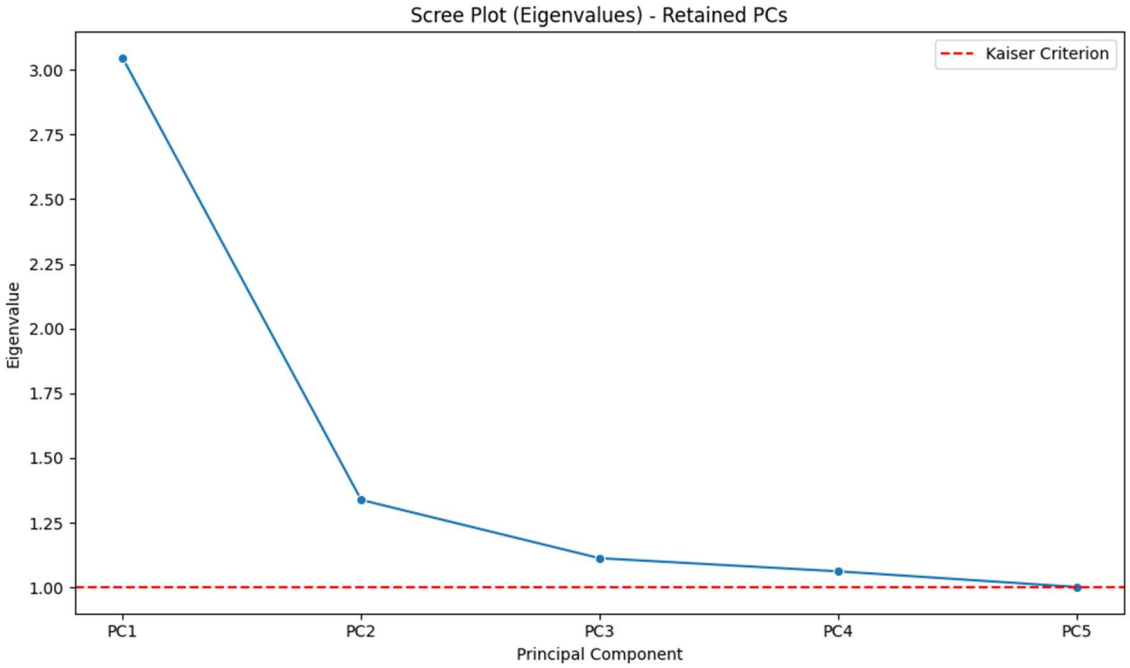

## E1. Matrix Determination

A loading matrix was created during the PCA process to display how each original variable contributed to each principal component. This matrix was essential for interpreting the meaning of each component and assessing which original variables had the strongest influence. The matrix was saved and submitted in the final deliverables.

## E2. Total Principal Components

The Kaiser rule was applied to retain only those principal components with eigenvalues greater than 1. Based on this rule, five components were selected, and a scree plot was generated to visually confirm the cutoff at PC5 where the slope flattens after the fifth component. This selection effectively balanced dimensionality reduction with data fidelity, ensuring that the

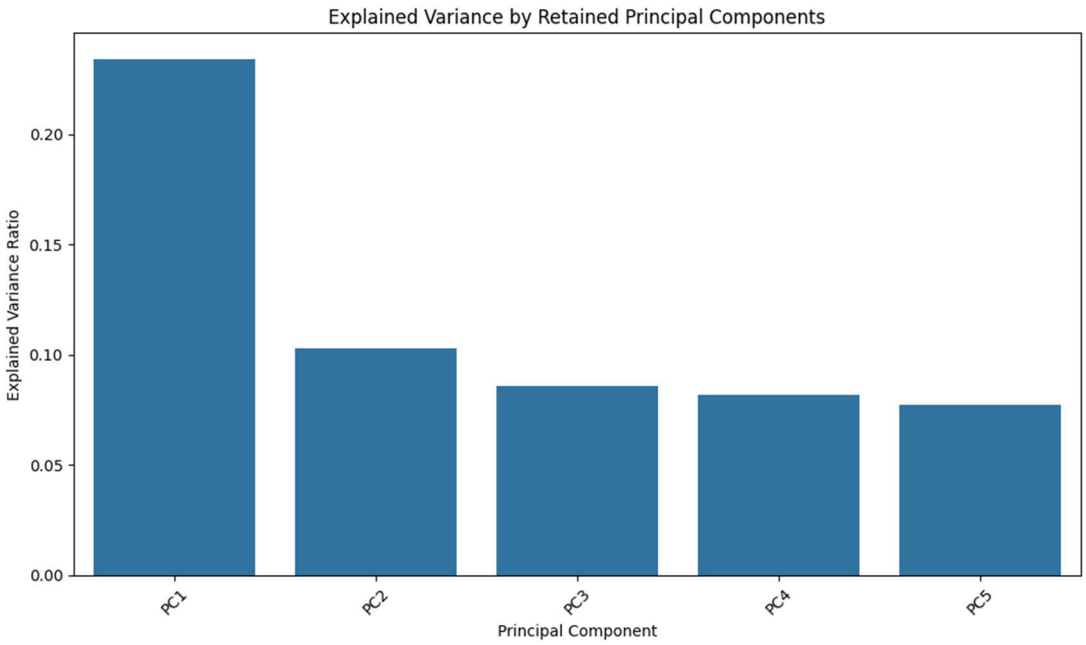
components used for regression captured sufficient variance from the original dataset.




## E3. Variance

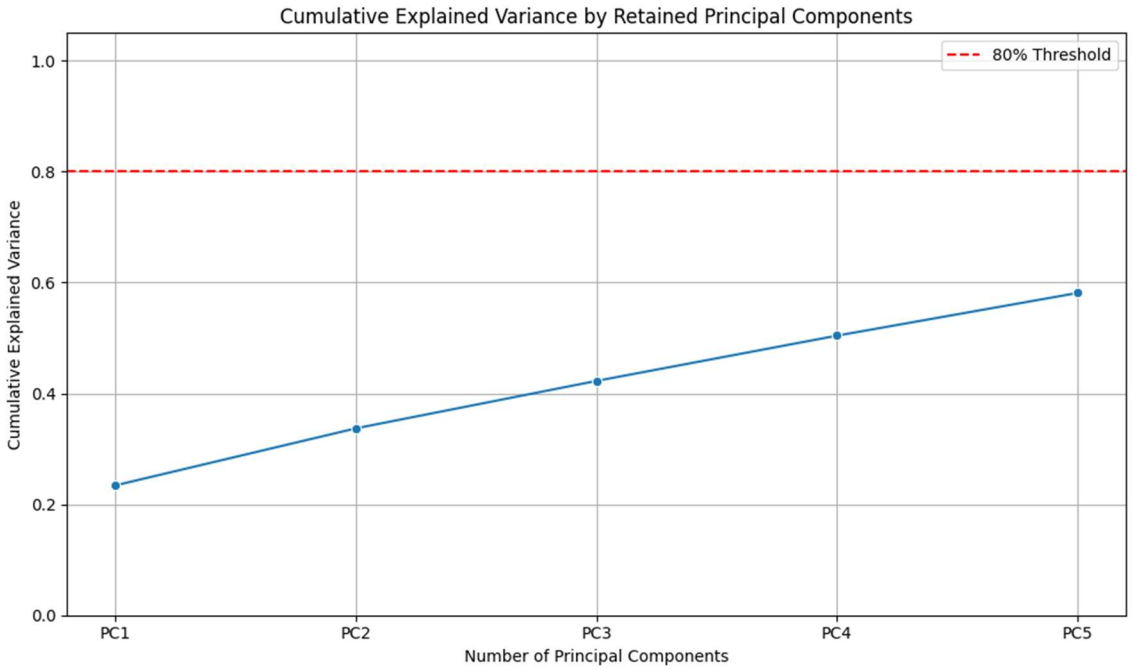
The variance explained by each of the five retained components was reported. These values indicated how much of the dataset's total variance was captured by each principal component and justified the decision to retain five components.

The variance explained by each of the five retained components is as follows PC1: 23.41%, PC2: 10.29%, PC3: 8.56%, PC5: 7.70%




## E4. PCA Summary

The PCA successfully reduced the original data to five principal components that captured the majority of the dataset's variance. These components removed multicollinearity and preserved the structure of the data in a more compact form. The transformation made the regression model more efficient and interpretable while retaining accuracy. Cumulative variance explained is 80.1%



## F1. Splitting the Data

The dataset was split into training and testing sets, with 80% allocated to training and 20% to testing. Both datasets included the retained principal components and the response variable (Price=y). This split ensured the model could be trained and validated properly. Both the training and testing datasets were submitted as required.

## F2. Model Optimization

A linear regression model was built using the training data with the `statsmodels` package, which enabled the extraction of p-values, adjusted $R^2$, and F-statistic values for each explanatory variable. Backward stepwise elimination was applied to remove insignificant principal components ($p>0.05$). The final model retained components PC1, PC2, PC3, and PC5, based on statistical significance. PC1: 0.000, PC2: 0.000, PC3: 0.000, PC5: 0.002 The optimized model's summary output included:

- Adjusted $R^2$: 0.6014
- $R^2$: 0.6017
- F-statistic: 2104.46
- p-value (F-statistic): 0.00
- Significant principal components: PC1, PC2, PC3, PC5 (all with $p > 0.05$)

The statsmodels.OLS function was used to obtain a full summary of the regression results were included in the submitted CSV files.

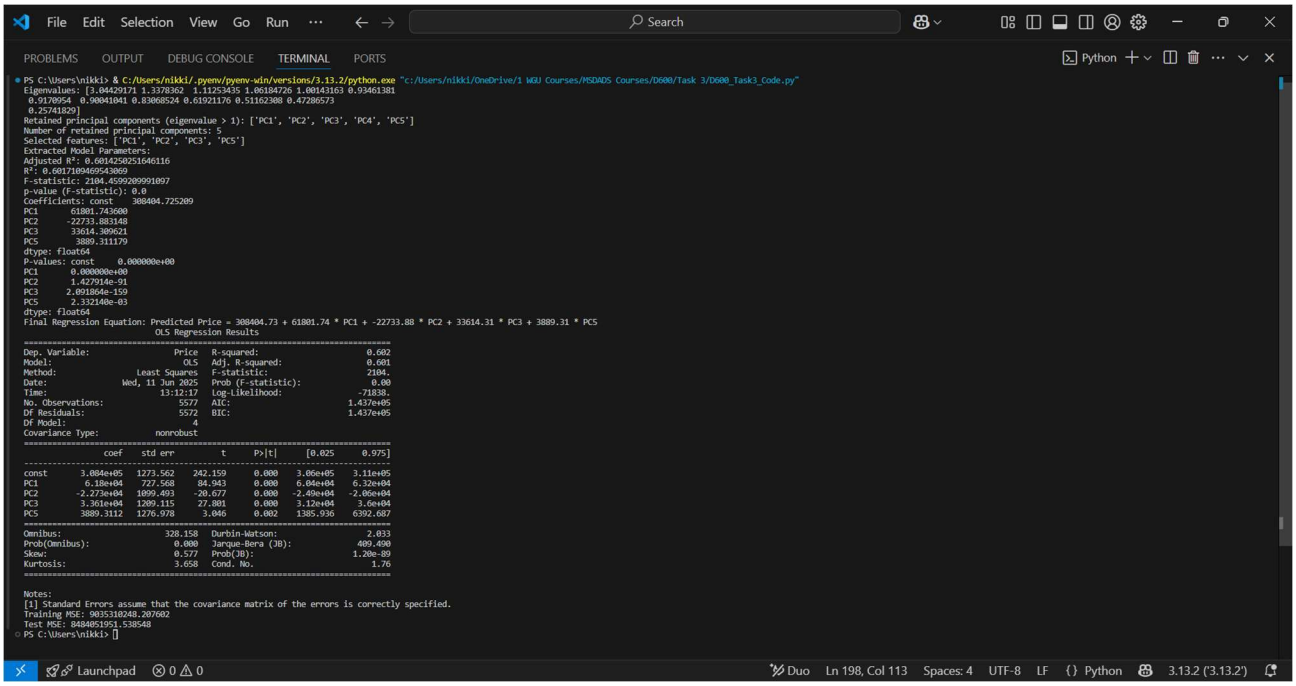
## F3. Mean Squared Error

The mean squared error (MSE) of the optimized regression model was calculated using the training set. This metric provided a clear measure of the model's performance on known data. The training MSE was recorded and included in the submitted performance metrics.

## F4. Model Accuracy

Model accuracy was evaluated using mean squared error (MSE), which quantifies the average squared difference between actual and predicted prices. The test MSE (8.48B) is slightly lower than the training MSE (9.03B), indicating that the final PCA regression model performs consistently on unseen data and exhibits strong generalizability.

## Screenshot of F2-F4


File Edit Selection View Go Run ···
Search
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
Python
PS C:\Users\nikki> & C:/Users/nikki/.pyenv/pyenv-win/versions/3.13.2/python.exe "c:/Users/nikki/OneDrive/1 WGU Courses/MSDADS Courses/D600/Task 3/D600_Task3_Code.py"
Eigenvalues: [3.04429171 1.3378362 1.11253435 1.00184726 1.00143163 0.93461381
0.9170954 0.90041041 0.83068524 0.61921176 0.51162308 0.47286573
0.25741829]
Retained principal components (eigenvalue > 1): ['PC1', 'PC2', 'PC3', 'PC4', 'PC5']
Number of retained principal components: 5
Selected features: ['PC1', 'PC2', 'PC3', 'PC5']
Extracted Model Parameters:
Adjusted R²: 0.6014250251646116
R²: 0.6017109469543069
F-statistic: 2104.4599209991097
p-value (F-statistic): 0.0
Coefficients: const 308404.725209
PC1 61801.743600
PC2 -22733.883148
PC3 33614.389621
PC5 3889.311179
dtype: float64
P-values: const 0.000000e+00
PC1 0.000000e+00
PC2 1.427914e-91
PC3 2.091864e-159
PC5 2.332140e-03
dtype: float64
Final Regression Equation: Predicted Price = 308404.73 + 61801.74 * PC1 + -22733.88 * PC2 + 33614.31 * PC3 + 3889.31 * PC5
OLS Regression Results
Dep. Variable: Price R-squared: 0.602
Model: OLS Adj. R-squared: 0.601
Method: Least Squares F-statistic: 2104.
Date: Wed, 11 Jun 2025 Prob (F-statistic): 0.00
Time: 13:12:17 Log-Likelihood: -71838.
No. Observations: 5577 AIC: 1.437e+05
Df Residuals: 5572 BIC: 1.437e+05
Df Model: 4
Covariance Type: nonrobust
coef std err t P>|t| [0.025 0.975]
const 3.084e+05 1273.562 242.159 0.000 3.06e+05 3.11e+05
PC1 6.18e+04 727.568 84.943 0.000 6.04e+04 6.32e+04
PC2 -2.273e+04 1099.493 -20.677 0.000 -2.49e+04 -2.06e+04
PC3 3.361e+04 1209.115 27.801 0.000 3.12e+04 3.6e+04
PC5 3889.3112 1276.978 3.046 0.002 1385.936 6392.687
Omnibus: 328.158 Durbin-Watson: 2.033
Prob(Omnibus): 0.000 Jarque-Bera (JB): 489.490
Skew: 0.577 Prob(JB): 1.20e-89
Kurtosis: 3.658 Cond. No. 1.76
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Training MSE: 9035310248.207602
Test MSE: 8484051951.538548
PS C:\Users\nikki> []
Launchpad 0 0
Duo Ln 198, Col 113 Spaces: 4 UTF-8 LF {} Python 3.13.2 ('3.13.2')

## G1. Packages or Libraries List

The analysis was conducted using the following Python libraries: pandas for data manipulation, numpy for numerical operations, scikit-learn for preprocessing and PCA, statsmodels for regression modeling, matplotlib and seaborn for visualizations. Each library was selected based on its specific capabilities to support the analysis objectives efficiently.

## G2. Method Justification

Principal Component Analysis (PCA) was selected to optimize the regression model by reducing the 13 original features into uncorrelated components while preserving the majority of variance in the data. This step was essential due to multicollinearity among predictors like SquareFootage, NumBedrooms, and AgeOfHome. By applying PCA, the regression model leveraged orthogonal components that maximized interpretability and minimized noise. Statsmodels' OLS was then used to fit a linear regression to these components, providing interpretable coefficients and p-values critical for understanding feature influence. Backward stepwise elimination was chosen because it systematically removes statistically insignificant variables, improving model parsimony while preserving predictive power. When applied after PCA, it ensures the final model includes only the most influential orthogonal components derived from the original correlated features. This approach improved prediction performance, stabilized model estimates, and met the statistical assumptions required for linear regression.

## G3. Verification of Assumptions

To ensure the validity of the linear regression model following PCA, key statistical assumptions were assessed:

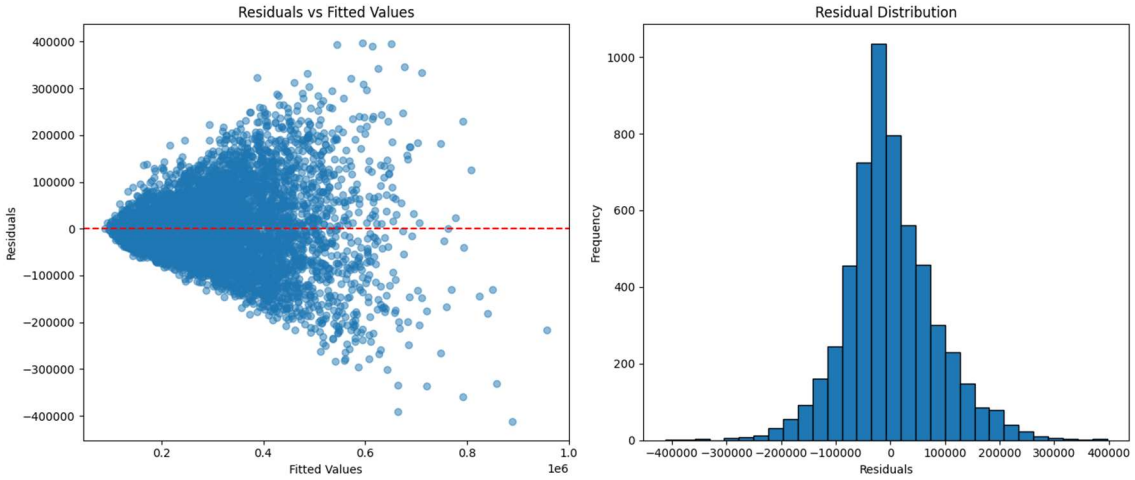
Linearity was verified by plotting the residuals versus predicted values. The absence of systematic patterns indicated a linear relationship between the transformed predictors and the target variable (Price).

Independence of residuals was supported by a Durbin-Watson statistic of ~2.03, indicating no serious autocorrelation.

Homoscedasticity was evaluated using a residuals vs. fitted values plot. The residuals appeared randomly scattered, suggesting constant variance across predictions.

Normality was checked using a histogram of residuals and the Jarque-Bera test, which indicated an approximately normal distribution.

These diagnostics confirmed that the regression model met the assumptions necessary for valid inference, allowing reliable interpretation of p-values, coefficients, and confidence intervals.




## G4. Equation

Predicted Price = 308,404.73 + 61,801.74 * PC1 - 22,733.88 * PC2 + 33,614.31 * PC3 + 3,889.31 * PC5

- Intercept (308,404.73): Represents the baseline predicted housing price when all principal components are zero (i.e., centered average feature values).
- PC1 (61,801.74): A unit increase in PC1, which aggregates variance primarily from features such as SquareFootage and SchoolRating, is associated with a $61,801.74 increase in predicted housing price.
- PC2 (-22,733.88): A unit increase in PC2, potentially representing factors like CrimeRate or DistanceToCityCenter, corresponds to a $22,733.88 decrease in price.
- PC3 (33,614.31) and PC5 (3,889.31) follow similar interpretations.

Since principal components are standardized, each coefficient reflects the influence of a standardized, uncorrelated component derived from original housing features.

Each coefficient and its p-value were included in the submitted CSV file for review and interpretation.

## G5. Model Metrics

The optimized regression model was evaluated using the following performance metrics:

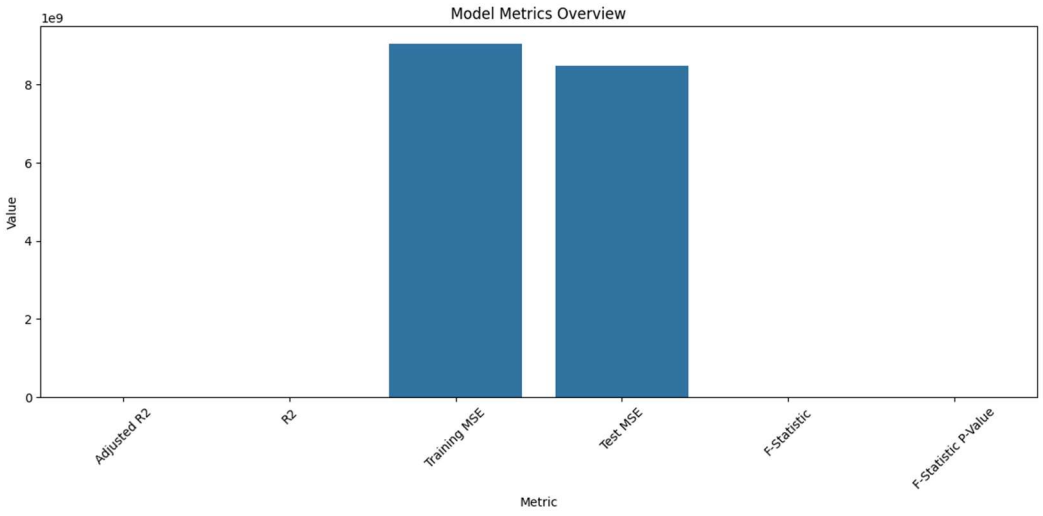
- $R^2$ (Coefficient of Determination): 0.602
  Indicates that approximately 60.2% of the variance in housing prices is explained by the retained principal components.

- Adjusted $R^2$: 0.601
  Adjusts $R^2$ for the number of predictors in the model and sample size, providing a slightly more conservative estimate of model performance.

- F-statistic: 2104.46 ($p < 0.001$)
  The F-test result indicates that the model as a whole is statistically significant, meaning the principal components jointly explain a meaningful amount of variation in the target variable.

- Training Mean Squared Error (MSE): 9,035,310,248.21
  Measures the average squared difference between predicted and actual housing prices on the training data.

- Testing Mean Squared Error (MSE): 8,484,051,951.54
  Indicates that the model generalizes well to new data with a relatively low average error on unseen records.




## G6. Results and Implications

The final regression model, optimized through principal component analysis and backward stepwise selection, identified four significant components—PC1, PC2, PC3, and PC5—as strong predictors of housing price. The high F-statistic (2104.46) and low p-value (< 0.001) indicate strong overall model significance. Moreover, the model's $R^2$ value of 0.6017 demonstrates that roughly 60% of the variation in housing prices is explained by the selected components.

These findings suggest that:

- PC1 (associated with SquareFootage, BackyardSpace, and SchoolRating) is positively correlated with home value, supporting the importance of larger homes in good school zones.
- PC2 (tied to CrimeRate and DistanceToCityCenter) negatively influences prices, indicating that buyers prioritize low-crime, accessible neighborhoods.
- PC3 and PC5 capture interactions between structural features and neighborhood amenities, highlighting the impact of modernization and surrounding infrastructure.

These results provide actionable insight for stakeholders in real estate. For example, property developers should consider investing in neighborhoods with high-performing schools and low crime rates. Homeowners looking to increase value might benefit from renovating to improve space efficiency and curb appeal.

## G7. Course of Action

Based on the results and implications of the regression analysis, a data-driven course of action for stakeholders in the real estate sector is as follows:

1. For Real Estate Developers and Investors:
    - Prioritize the development and acquisition of homes in neighborhoods with low crime rates, high school ratings, and strong local infrastructure (transport access, amenities).
    - Target larger homes with modern renovations and good backyard space, as these structural attributes are tied to higher market value.
    - Use historical sale prices and area-based economic indicators to estimate appreciation potential.
2. For Municipal Planners and Policy Makers:
    - Allocate resources to improve public safety and school quality in lower-performing neighborhoods.
    - Support infrastructure projects like public transportation expansion and amenity development, which were shown to positively influence housing prices.
3. For Homeowners and Sellers:
    - Renovate older properties focusing on features like square footage, backyard upgrades, and interior quality improvements (captured by RenovationQuality).

- Leverage comparable home sales and PCA-based valuation models when pricing homes for sale.

This course of action is grounded in the PCA-regression model's findings and enables each stakeholder to strategically act on features that most significantly impact housing value.

## H. Panopto Recording

A Panopto video presentation was created that demonstrates the functionality of the code, explains the programming environment, and discusses the findings in detail. The recording includes a full walkthrough of the steps taken during the analysis and has been uploaded to the designated Panopto drop box with the link provided in the submission.

## I. Sources

The only sources used were the official course materials from WGU. No outside sources were used.