Principal Component Analysis

Trayvonious Pendleton

WGU

Course Number: D600

STUDENT ID: 011205284

***B1: Research Question***

“How does housing square footage impact property prices?”

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

This analysis aims to determine the relationship between housing square footage and property prices using linear regression. Understanding this relationship will help real estate investors, home buyers, and policymakers make informed decisions regarding pricing strategies and market trends.

***C1: How PCA prepares the Dataset for Regression Analysis***

PCA is useful when dealing with datasets that contain many correlated variables, as it helps:

* Reduce dimensionality, making the model more efficient
* Remove multicollinearity, which can negatively impact regression results.
* Improve computational efficiency by reducing the number of variables while retaining most of the information.
* Enhance model interpretability by focusing on key factors that explain the variance.

Expect Outcomes of PCA:

* The dataset will be transformed into a set of uncorrelated principal components.
* The most important components will be retained.
* The regression model will be based only on relevant components, reducing noise and redundancy.

***C2: Summarize One Assumption of PCA***

Key Assumption of PCA:

Linearity Assumption:

* PCA assumes that the relationships between variables are linear
* If the dataset contains nonlinear relationships, PCA may not be effective.

***Part D1: Identify Continuous Dataset Variables***

Continuous Variable:

* Price
* SquareFootage
* NumBathrooms
* NumBedrooms
* BackyardSpace
* CrimeRate
* SchoolRating
* AgeOfHome
* DistanceToCityCenter
* EmploymentRate
* PropertyTaxRate
* RenovationQuality
* LocalAmenities
* TransportAcces
* PreviousSalePrice

***D3: Descriptive Statistics***

***A screenshot of a computer

AI-generated content may be incorrect.***

***E1: Matrix of all Principal Components***

A screenshot of a computer

AI-generated content may be incorrect.

***E2 Total Number of Principal Components***

***A graph with a line and a number of principal components

AI-generated content may be incorrect.E3: Variance of Each Principal Component***

***A screenshot of a computer program

AI-generated content may be incorrect.***

***E4: Summarizing PCA Results***

***A screenshot of a computer program

AI-generated content may be incorrect.***

***F2: Training and Optimization***

***A screenshot of a computer

AI-generated content may be incorrect.***

A screenshot of a computer

AI-generated content may be incorrect.

***F3: Compute MSE on Training Data***

***A screenshot of a computer program

AI-generated content may be incorrect.***

***F4: Run Predictions and Compute MSE on the training setA screenshot of a computer program

AI-generated content may be incorrect.***

***G1: List of Python Packages and Justifications***

The following libraries were used to conduct the analysis:

* Pandas: Used for data manipulation, cleaning, and storage in CSV files.
* Numpy: Used for mathematical operations like calculating variance and cumulative variance.
* matplotlib.pyplot: Used to visualize the scree plot for PCA interpretation.
* sklearn.preprocessing: Used to normalize data before PCA to ensure all variables have equal influence.
* sklearn.decomposition: Used to reduce dimensionality, improving model efficiency.
* sklearn.model\_selection: Used to split data into training and testing sets.
* statsmodels.api: Used to build the regression model and analyze coefficient significance.
* sklearn.metrics: Used to evaluate model accuracy using MSE.

Justification:

* These packages streamline data processing, reduce redundancy, and improve model accuracy.
* PCA reduces overfitting, and OLS regression identifies significant predictors.

***G2: Model Optimization Approach and Justification***

**Method Used: Backward Stepwise Elimination**

* The initial model was trained using all 10 principal components.
* Predictors with high p-values (p > 0.05) were removed step-by-step.
* The final model retained only statistically significant variables.

**Justification**:

* Backward elimination ensures only relevant predictors remain, improving accuracy.
* Removes weak predictors (e.g., PC9, PC5, PC6) to avoid overfitting.
* Increases interpretability and prevents unnecessary complexity.

G3: Verification of Assumptions

The following assumptions were verified before finalizing the regression model:

* Linearity Assumption: Verified using R² = 0.801, confirming that PCA components explain price variations.
* No Multicollinearity: PCA was used to reduce multicollinearity, ensuring independent components.
* Homoscedasticity: Residual plots showed constant variance, supporting the assumption.
* Normality of Residuals: Jarque-Bera test (p < 0.05) confirmed near-normal residuals.

***G4: Regression Equation and Coefficient Estimates***

**The final regression equation is:**

^Y=0.0025 +0.4289(PC1) – 0.1834(PC2)- 0.1775(PC3) + 0.0225(PC4)+ -0.0152(PC7) -0.0287(PC8) +0.0645(PC10)

**Key Insights on Coefficients:**

* PC1 (0.4289, p < 0.000) → Largest positive impact on price.
* PC2 (-0.1834, p < 0.000) → Negative impact.
* PC3 (-0.1775, p < 0.000) → Negative impact.
* PC10 (0.0645, p < 0.000) → Smaller positive impact.
* Intercept is NOT significant (p = 0.663) → Price is mainly determined by the principal component

***G5: Model Metrics***

**Training Set Metrics**

* R² = 0.801 (80.1% variance explained)
* Adjusted R² = 0.8007 (slightly lower due to adjustments)

**Comparison of MSE:**

* Training Set MSE: 0.2020
* Test Set MSE: 0.1996
* The slight difference between train & test MSE suggests good model generalization.

***G6: Results and Implications of the Prediction Analysis:***

The model predicts house prices with high accuracy.

The PCA-transformed variables effectively capture price variations.

Minimal overfitting means the model works well on unseen data.

**Implications**:

* PC1 has the most potent effect on house prices.
* Crime rate, school rating, and transport access (original variables contributing to PC1) are key determinants.
* This model can predict future home prices for buyers and investors.

***G7: Recommend Course of Action***

**Based on the results, the following actions are recommended:**

* For real estate investors: Focus on properties with favorable PC1-related factors (square footage, crime rate, school rating).
* For urban planners: Improving transport access and school quality will positively impact home values.
* For homeowners: Renovations and upgrades related to PC1 factors can increase property value.
* For policymakers: Reducing crime rates and improving local amenities could raise property prices in low-value areas.