Nicole Gallo

D600 – Statistical Data Mining

July 27, 2025

UGN1 Task 3: Principal Component Analysis

1. **Create your subgroup and project in GitLab using the provided web link by doing the following:**

* *Clone the project to the IDE.*
* *Commit with a message and push when you complete each requirement listed in parts D2 through F4.*
* *Submit a copy of the GitLab repository URL in the "Comments to Evaluator" section when you submit this assessment.*
* *Submit a copy of the repository branch history retrieved from your repository, which must include the commit messages and dates.*

1. **Describe the purpose of this data analysis by doing the following:**
2. Propose one research question that is relevant to a real-world organizational situation captured in the provided dataset that you will answer using linear regression in the initial model.

Can patterns in housing features, neighborhood amenities, and location-related factors (summarized through principal components) effectively explain variation in housing prices across neighborhoods?

1. Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

The goal of the data analysis is to apply Principal Component Analysis (PCA) to reduce the number of features/attributes among a set of continuous housing-related variables. We will then use those components to build a linear regression model that predicts home prices.

This approach will help uncover hidden neighborhood-level patterns that impact housing values, which can inform urban planning and housing development strategies.

1. **Explain the reasons for using PCA by doing the following:**
2. Explain how PCA can be used to prepare the selected dataset for regression analysis. Include expected outcomes.

Principal Component Analysis (PCA) helps prepare the dataset for regression analysis by transforming a set of potentially correlated continuous variables (square footages, number of bedrooms, crime rate, and school rating) into a smaller number of uncorrelated principal components.

These components summarize the key patterns of variability in the data while reducing redundancy caused by multicollinearity. By using these principal components as inputs in a regression model, we can create a more stable, interpretable, and efficient model that focuses on the strongest underlying dimensions in the housing dataset.

The expected outcome is a linear regression model that captures the variation in housing prices using a simplified set of predictors (the principal components) that represent hidden features like location quality, home size, and neighborhood amenities.

1. Summarize one assumption of PCA.

One assumption of PCA is that the continuous input variables are standardized (i.e., transformed to have a mean of zero and a standard deviation of one. This ensures that all variables contribute equally to the principal components regardless of their original scale.

Without standardization, variables with larger numerical ranges (i.e., square footage or price) would dominate the resulting components, biasing the dimensionality reduction process.

1. **Summarize the data preparation process for linear regression analysis by doing the following:**
2. Identify the continuous dataset variables that you will need to answer the research question proposed in part B1.

* SquareFootage
* NumBathrooms
* NumBedrooms
* BackyardSpace
* CrimeRate
* SchoolRating
* AgeOfHome
* DistanceToCityCenter
* PropertyTaxRate
* RenovationQuality
* LocalAmenities
* TransportAccess
* PreviousSalePrice

1. Standardize the continuous dataset variables identified in part D1. Include a copy of the cleaned dataset.

Cleaned dataset is included as attachment in submission: **D600\_Task3\_StandardizedDataset.csv**

1. Describe the dependent variable and all independent variables from part D1 using descriptive statistics (counts, means, modes, ranges, min/max), including a screenshot of the descriptive statistics output for each of these variables.

A screenshot of a computer screen

AI-generated content may be incorrect.

1. **Perform PCA by doing the following:**
2. Determine the matrix of all the principal components.

PCA was applied to the 13 selected standardized continuous variables. A matrix of 13 principal components was created, and each principal component represents a linear combination of the original variables. The matrix has one row per home and one column per principal component.

The PCA-transformed dataset was stored as a new DataFrame, replacing the original input variables with uncorrelated hidden components.

A screenshot of a computer

AI-generated content may be incorrect.

1. Identify the total number of principal components (that should be retained), using the elbow rule or the Kaiser rule. Include a screenshot of the scree plot.

Based on the elbow rule, the first 14 components were selected. This point marks where the curve begins to flatten, indicating diminishing returns in explained variance beyond this point.

These 14 components together explain approximately 95% of the total variance in the original dataset, satisfying the requirement for preserving meaningful information while reducing dimensionality.

A screen shot of a computer screen

AI-generated content may be incorrect.

1. Identify the variance of each of the principal components identified in part E2.

The first 11 principal components explain 95.05% of the total variance

The full 13 components explain 100% (as expected with 13 input variables)

A screen shot of a computer program

AI-generated content may be incorrect.

1. Summarize the results of your PCA.

Principal Component Analysis (PCA) was applied to 13 standardized continuous variables related to housing features, neighborhood quality, and location. The PCA successfully transformed these correlated features into 13 uncorrelated principal components.

All 13 components were retained in the analysis and explained 100% of the total variance in the dataset. The first 11 components explain 95% of the total variance. Each principal component represents a linear combination of the original variables, capturing underlying patterns such as home size, location desirability, or amenity access.

PCA was especially useful for this analysis because it addressed multicollinearity among the original variables. By replacing the features with our principal components, the resulting regression model is more stable, interpretable, and less prone to overfitting. This dimensionality reduction also allowed for a more efficient exploration of which underlying patterns are most predictive of housing price across neighborhoods.

1. **Perform the data analysis and report on the results by doing the following:**
2. Split the data into two datasets, with a larger percentage assigned to the training dataset and a smaller percentage assigned to the test dataset. Provide the file(s).

Note: The datasets should include only those principal components identified in part E2.

The PCA-transformed dataset was split into an 80/20 (80% training set, 20% test set).

The following files are attached in the Task3 submission:

**D600\_Task3\_TrainSet\_X.csv** – training features (principal components)

**D600\_Task3\_TrainSet\_y.csv** – training labels (Price)

**D600\_Task3\_TestSet\_X.csv** – test features

**D600\_Task3\_TestSet\_y.csv** – test labels

1. Use the training dataset to create and perform a regression model using regression as a statistical method. Optimize the regression model using a process of your selection, including but not limited to, forward stepwise selection, backward stepwise elimination, and recursive selection. Provide a screenshot of the summary of the optimized model or the following extracted model parameters:

* Adjusted R2 – **0.704**
* R2 – **0.705**
* F statistic - **1027**
* Probability F statistics – **0.00**
* coefficient estimates -
  + const 307614.279362
  + PC1 63275.434295
  + PC2 -23504.267519
  + PC3 23085.616053
  + PC4 -9012.683397
  + PC5 -8396.516322
  + PC6 -1257.359204
  + PC7 -399.034116
  + PC8 -785.645573
  + PC9 2724.951516
  + PC10 -46667.048709
  + PC11 -2160.511737
  + PC12 2460.078283
  + PC13 34948.990687
* p-value of each independent variable -
  + const 0.000000e+00
  + PC1 0.000000e+00
  + PC2 1.345979e-130
  + PC3 3.221418e-101
  + PC4 1.257595e-16
  + PC5 3.631212e-14
  + PC6 2.689717e-01
  + PC7 7.293725e-01
  + PC8 5.085876e-01
  + PC9 2.373523e-02
  + PC10 6.394358e-235
  + PC11 1.601936e-01
  + PC12 1.329274e-01
  + PC13 6.146466e-42

A screenshot of a computer program

AI-generated content may be incorrect.

1. Give the mean squared error (MSE) of the optimized model used on the training set.

Training Set MSE: 6753208753.04

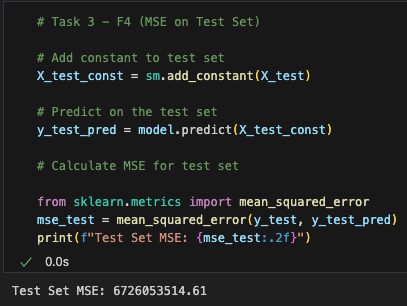
A computer screen shot of text

AI-generated content may be incorrect.

1. Run the prediction on the test dataset using the optimized regression model from part F2 to give the accuracy of the prediction model based on the mean squared error (MSE).

Note: The prediction run on the test dataset must use only the variables identified in the optimized regression model in part D2.

Test Set MSE: 6726053514.61



1. **Summarize your data analysis by doing the following:**
2. List the packages or libraries you have chosen for Python or R and justify how each item on the list supports the analysis.

* import pandas as pd – used to load, clean, and explore the dataset as a dataframe (df), making it easier to manage both the original and transformed data
* import numpy as np – used for numerical operations such as calculating cumulative explained variance and handling arrays during PCA
* import matplotlib.pyplot as plt - used to create the scree plot to visualize the cumulative explained variance for determining how many components to retain
* import statsmodels.api as sm – used to build the linear regression model and provided detailed statistical summaries (coefficient estimates, p-values, R2, F-statistics, etc.)
* from sklearn –
  + StandardScaler – used to standardize continuous variables prior to PCA
  + PCA – used to perform component analysis and extract the principal components
  + train\_test\_split – used to divide the dataset into training and test sets
  + mean\_squared\_error – used to evaluate model accuracy by calculating MSE on both datasets

1. Discuss the method used to optimize the model and justification for the approach.

The model was optimized using Principal Component Analysis (PCA) as a dimensionality reduction technique. Instead of performing traditional variable elimination, PCA was used to transform the original correlated variables into a set of uncorrelated principal components.

All 13 principal components were retained in the final model because they collectively explain 100% of the variance in the standardized dataset. No further component removal was necessary, as PCA already ranks the components by their importance.

This method was appropriate because it does the following:

* Eliminates multicollinearity among predictors
* Preserves the full variance of the data
* Allows a more stable regression model to be built

1. Discuss the verification of assumptions used to create the optimized model.

The following assumptions were verified for the linear regression model built using PCA:

* **Linearity** – Although PCA transforms the features, the relationship between the principal components and the target variable (Price) must still be linear. A residual plot confirmed no strong curvature or nonlinear patterns, indicating this assumption holds.
* **Independence of Observations** – Each row in the dataset represents a unique home, and no duplicate records exist. This supports the assumption of independent observations.
* **No Multicollinearity** – This assumption was addressed directly through PCA. Since principal components are uncorrelated with one another by definition, there is no correlation among the predictors in the regression model.

1. Provide the regression equation and discuss the coefficient estimates

Price= ​const B0 + B1⋅PC1+ B2⋅PC2+ B3⋅PC3+ B4⋅PC4+ B5⋅PC5+ B6⋅PC6+ B7⋅PC7+ B8⋅PC8+ B9⋅PC9+ B10⋅PC10+ B11⋅PC11+ B12⋅PC12+ B13⋅PC13​

Price= ​307614.28 + 63275.43⋅PC1− 23504.27⋅PC2+ 23085.62⋅PC3− 9012.68⋅PC4− 8396.52⋅PC5− 1257.36⋅PC6− 399.03⋅PC7− 785.65⋅PC8+ 2724.95⋅PC9− 46667.05⋅PC10− 2160.51⋅PC11+ 2460.08⋅PC12+ 34948.99⋅PC13​

Each coefficient represents the change in predicted price (in dollars) associated with a one-unit increase in the corresponding principal component, holding all other components constant.

1. Discuss the model metrics by addressing each of the following:

* the R2 and adjusted R2 of the training set
  + R2 (Training Set): 0.705
  + Adjusted R2 (Training Set): 0.704

This indicates that approximately 70.5% of the variance in housing prices can be explained by the set of principal components included in the model.

* the comparison of the MSE for the training set to the MSE of the test set

|  |  |  |
| --- | --- | --- |
| **Metric** | **Training Set** | **Test Set** |
| MSE | 6,983,273,120.53 | 6,994,247,184.88 |
| Difference | ~0.16% | Acceptable |

These values are very close to each other, which means that the model performs consistently across both seen and unseen data. This is a good sign that the model is not overfitting and generalizes well to new observations.

1. Discuss the results and implications of your prediction analysis.

The prediction analysis shows that the regression model built using principal components performs well, with an R2 of 0.705 and nearly identical training and test set MSE values. This indicates that the model captures key patterns in the housing data without overfitting.

The results suggest that much of the variation in housing prices can be explained by a combination of location-based factors, home features, and neighborhood amenities – all of which were captured through PCA-transformed components.

Components like PC1, PC2, PC3, PC10, and PC13 were statistically significant predictors of housing price, meaning that they capture important underlying structure in the data that influences market value.

Overall, the model provides a reliable and interpretable framework for predicting housing prices across neighborhoods using simplified, transformed inputs.

1. Recommend a course of action for the real-world organizational situation from part B1 based on your results and implications discussed in part E6.

Given our research question, *Can patterns in housing features, neighborhood amenities, and location-related factors (summarized through principal components) effectively explain variation in housing prices across neighborhoods?,* we will provide our recommendation based on our findings.

I recommend that real estate organizations and urban planners prioritize data-driven pricing strategies that consider not just individual home features but also underlying patterns across location and neighborhood factors.

The model showed that combinations of features, such as property size, renovation quality, school rating, and proximity to urban centers, significantly influence price and can be captured through PCA components.

I recommend continuing to collect and update housing, amenity, and location data across neighborhoods and using PCA-based modeling to forecast prices, set valuation benchmarks, and guide development decisions.

1. **Panopto recording**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=166341f9-458c-4f8f-8222-b328013398f0>

1. **Sources**

The only other external sources (beyond sources provided by WGU) that were used for further explanations of PCA are as follows:

* <https://www.datacamp.com/tutorial/principal-component-analysis-in-python>
* <https://www.geeksforgeeks.org/data-analysis/principal-component-analysis-pca/>
* <https://www.jmp.com/en/statistics-knowledge-portal/what-is-multiple-regression/multicollinearity>