Peyton Bailey

May 15th, 2025

D600 – Task 3

B. Describe the purpose of this data analysis by doing the following:

1. **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.**

How effectively can principal components derived from housing characteristics predict home prices, and which underlying factors have the most decisive influence?

2. **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.**

To develop a regression model using principal components that explains the variance in housing prices while reducing dimensionality, creating a more interpretable and efficient pricing model for real estate valuations.

C. Explain the reasons for using PCA by doing the following:

1. **Explain how PCA can be used to prepare the selected dataset for regression analysis. Include expected outcomes.**

Principal Component Analysis (PCA) is a statistical technique that transforms a set of correlated continuous variables into a smaller set of uncorrelated variables called principal components (Source). For this housing dataset, PCA can be applied to the standardized continuous independent variables, such as square footage, age, previous sale price, etc., to address multicollinearity issues and reduce dimensionality before performing linear regression. Through PCA, the original variables are combined into principal components that capture the majority of

the variance in the data while being uncorrelated with each other. By selecting only the top principal components, I expect to see a regression model that contains predictors that do not overlap in the information they provide while being efficient and less prone to overfitting. It should be capable of explaining a substantial portion of the variance in the dependent variable while also simplifying the model and effectively interpreting the results.

3. **Summarize one assumption of PCA.**

One assumption of PCA is linearity in the dataset. The variables combine in a linear manner to form the dataset and exhibit relationships among themselves (Vadapalli, 2025).

D. Summarize the data preparation process for linear regression analysis by doing the following:

1. **Identify the continuous dataset variables that you will need to answer the research question proposed in part B1.**

The independent continuous variables for this research question are Square Footage, Backyard Space, School Rating, Crime Rate, Age of Home, Renovation Quality, Previous Sale Price, and Local Amenities. The dependent continuous variable is Price.

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

3. 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.

# SquareFootage	
count	6972.0
mean	1050.4571658060813
std	425.9887899385284
min	550.0
25%	663.0350000000001
50%	999.67
75%	1344.0925
max	2874.7

```
count   6,972.00
mean       31.24
std        18.03
min         0.03
25%        17.40
50%        30.39
75%        43.67
max        99.73
Name: CrimeRate, dtype: float64
```

```
count   6,972.00
mean       46.85
std        31.79
min         0.01
25%        20.81
50%        42.69
75%        67.25
max       178.68
Name: AgeOfHome, dtype: float64
count   6,972.00
mean        5.01
std         1.97
min         0.01
25%         3.67
50%         5.03
75%         6.36
max        10.00
Name: RenovationQuality, dtype: float64
```

```
count   6,972.00
mean        5.93
std         2.66
min         0.00
25%         4.00
50%         6.04
75%         8.05
max        10.00
Name: LocalAmenities, dtype: float64
count       6,972.00
mean      308,119.47
std       149,890.38
min        85,000.00
25%       192,851.43
50%       280,477.46
75%       392,435.11
max     1,046,675.64
Name: Price, dtype: float64
```

```
count       6,972.00
mean      285,664.88
std       185,207.47
min           22.80
25%       142,929.55
50%       262,872.06
75%       396,922.51
max     1,296,606.69
Name: PreviousSalePrice, dtype: float64
```

```
count   6,972.00
mean      511.38
std       280.15
min         0.39
25%       300.72
50%       495.92
75%       704.20
max     1,631.36
Name: BackyardSpace, dtype: float64
count   6,972.00
mean        6.95
std         1.89
min         0.22
25%         5.66
50%         7.01
75%         8.37
max        10.00
Name: SchoolRating, dtype: float64
```

E. Perform PCA by doing the following:

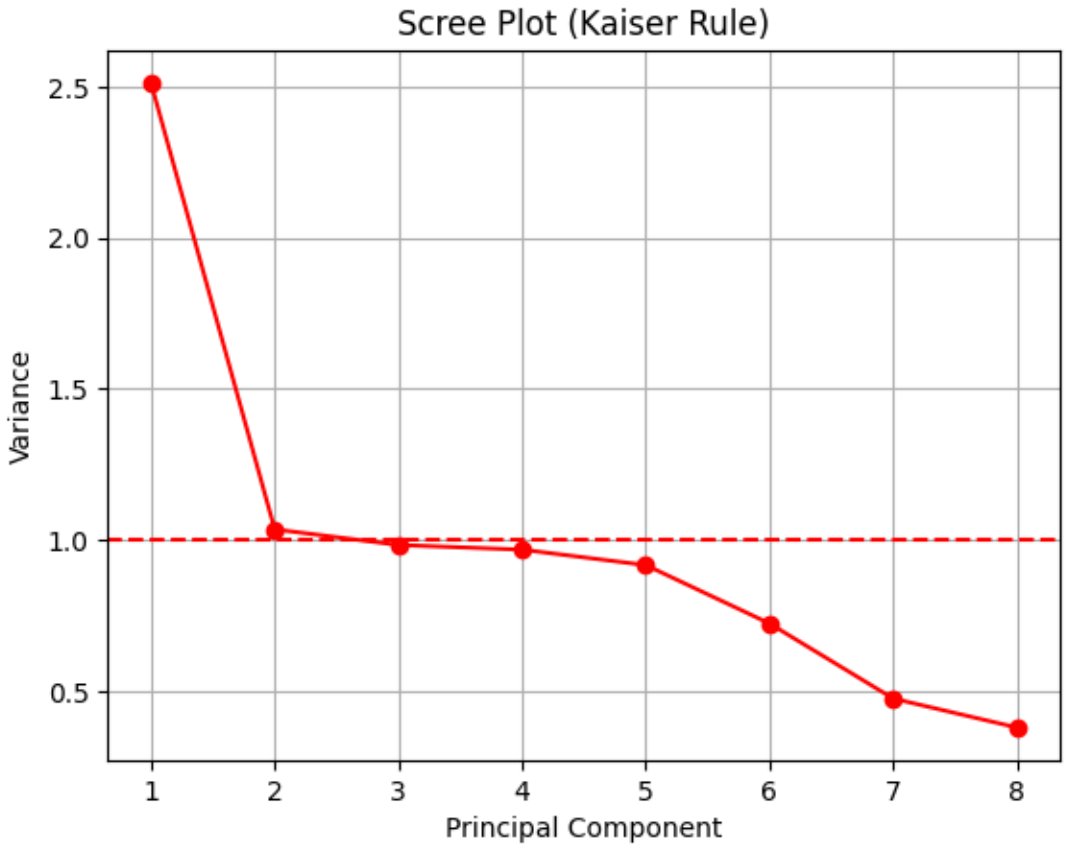
1. Determine the matrix of *all* the principal components.

The loadings matrix represents the correlation between each original variable and each principal component. The higher absolute value indicates a strong influence of that variable on the component, and the sign (positive or negative) indicates the direction of that relationship.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
SquareFootage	0.44	0.11	0.44	-0.14	-0.17	0.49	0.52	0.21
BackyardSpace	0.35	-0.27	-0.13	0.74	0.36	0.07	0.23	-0.24
SchoolRating	0.05	0.82	-0.25	0.41	-0.21	-0.05	-0.00	0.21
CrimeRate	0.12	0.47	0.06	-0.35	0.65	0.02	0.03	-0.47
AgeOfHome	-0.11	-0.03	-0.01	0.00	0.61	0.06	-0.06	0.78
RenovationQuality	-0.55	0.11	0.58	0.36	0.06	0.37	-0.22	-0.15
PreviousSalePrice	-0.23	-0.05	-0.61	-0.12	-0.01	0.74	0.03	-0.09
LocalAmenities	-0.54	0.02	-0.06	-0.00	0.04	-0.26	0.79	-0.03

2. 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.

The Kaiser rule is a widely used method in PCA to determine how many principal components to retain. It states that only components with eigenvalues greater than one should be retained. Eigenvalues represent the amount of variance explained by each principal component (Jolliffe, 2002). Components with an eigenvalue greater than one explain more variance than an original variable, while those with an eigenvalue less than one explain less variance than an original variable, therefore being deemed insignificant. As shown on the scree plot, the first two components meet this requirement of being greater than one. The third and fourth components are just shy of one. However, I decided to include these two in my initial model. This gave me a total of four principal components to retain.



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

PC1- 2.51

PC2 – 1.04

PC3 – 0.98

PC4 – 0.97

5. **Summarize the results of your PCA.**

Principal Component Analysis revealed that the first four components capture the majority of variance in the data at roughly 69%. Using the Kaiser Rule, I observed the variance of each principal component and retained the first four. The first two met the requirement; however, PC3 and PC4 were just underneath the threshold of 1. I decided to keep these two regardless, since I could do further optimization once I began performing linear regression. PC1 contained high positive loadings for property size (0.44) and backyard space (0.35) with high negative loadings for renovation quality (-0.55) and local amenities (-0.54). PC2 primary drivers were the positive loadings for school rating (0.82) and crime rate (0.47). PC3 had a high positive loading for renovation quality (0.58) and a high negative loading for previous sale price (-0.61). Finally, the most significant driver for PC4 was the high positive loading for backyard space (0.74). These components were used as predictors in subsequent regression analysis to model housing prices.

F. Perform the data analysis and report on the results by doing the following:

1. **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).**
2. **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 R^2**
- **R^2**
- **F statistics**
- **Probability F statistics**
- **coefficient estimates**
- **p-value of each independent variable**

```
=== Final Model Summary ===
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.561
Model:                            OLS   Adj. R-squared:                  0.560
Method:                 Least Squares   F-statistic:                     3555.
Date:                Thu, 15 May 2025   Prob (F-statistic):               0.00
Time:                        13:07:53   Log-Likelihood:                -72112.
No. Observations:                5577   AIC:                         1.442e+05
Df Residuals:                    5574   BIC:                         1.443e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3.087e+05   1337.482    230.837      0.000    3.06e+05    3.11e+05
PC1         6.856e+04    841.908     81.432      0.000    6.69e+04    7.02e+04
PC2         2.908e+04   1314.768     22.119      0.000    2.65e+04    3.17e+04
==============================================================================
Omnibus:                      350.827   Durbin-Watson:                   2.041
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              443.667
Skew:                           0.598   Prob(JB):                     4.56e-97
Kurtosis:                       3.690   Cond. No.                         1.59
==============================================================================
```

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

Mean Squared Error on Training Set: 9968972085.3838

4. 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).

The Mean Squared Error on the Test Set is 9,580,565,184.8850, only a 3% difference from the training set MSE. This slight difference indicates the model generalizes well to unseen data, showing good predictive performance without overfitting.

G. Summarize your data analysis by doing the following:

1. List the packages or libraries you have chosen for Python or R and justify how each item on the list supports the analysis.

Pandas – This package allowed me to import and manipulate the dataset within a data frame. This includes isolating each variable, calculating its descriptive statistics, and creating summary tables for the categorical variables.

Numpy – This fundamental library for numerical and scientific computing aided my analysis.

Matplotlib – This package supports the creation of visualizations. I used this package to create the spree plot.

Statsmodels – This package directly supports regression analysis. This allowed me to add a constant to my independent variables, input the X and Y variables, and fit them into a model using Ordinary Least Squares. This model provided a summary that included the coefficients for each independent variable, the R-squared, adjusted R-squared, F-statistic, AIC, and BIC. Furthermore, this package also consists of the variance inflation factor (VIF), which I used to detect multicollinearity before fitting the model.

Sklearn – This package also supports regression analysis, emphasizing machine learning. Through this package, I could implement PCA and all of its components and import a train-test split feature, allowing me to split my data into training and testing data, which allowed me to evaluate the accuracy of the prediction model. This package also allowed me to calculate the mean squared error between the predicted vs. actual values.

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

I used backward stepwise elimination to eliminate components that were not significant to the prediction of the dependent variable. This was also done one at a time using the p-value as a reference. A component with a p-value over 0.05 was considered insignificant and thus removed from the model. The third and fourth components were removed due to high p-values, which deemed them insignificant, supporting the Kaiser rule that I did not follow strictly.

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

PCA assumes orthogonality, which means the principal components should be uncorrelated with each other. This was verified by checking the VIF on each component and verifying that it was under 10. PCA automatically eliminates multicollinearity when creating the principal components.

4. Provide the regression equation and discuss the coefficient estimates

```
=== Multiple Linear Regression Equation ===
Price = 308739.60 + 68557.83*PC1 + 29081.13*PC2
```

The intercept of 308,739.60 represents the average housing price when all principal components equal zero. The PC1 coefficient of 68,557.83 indicates the amount the price increases per one unit increase in PC1. Likewise, the PC2 coefficient of 29,081.13 indicates the amount the price

increases per one unit increase in PC2. Even though PC1 has double the impact on price than PC2, they both play a significant role in predicting price.

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

- **the R^2 and adjusted R^2 of the training set**

The R^2 is 0.561, meaning the model explains 56.1% of the variance in housing prices. The adjusted R^2 is 0.560. The minimal difference between these parameters confirms that the model is not overfitted.

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

The training MSE is 9,968,972,085, and the test MSE is 9,580,565,185. The test MSE is only 3% lower than the training MSE, which is a positive sign. It suggests the model does not overfit and performs even better on unseen data.

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

My PCA-based regression model provided several key insights. Using eight original variables, PCA effectively reduced the dimensions to just two significant principal components while maintaining predictive power. Using these two principal components, the model captured 56.1% of housing price variance, ultimately demonstrating that substantial pricing information is contained within these composite variables. PC1 has more than twice the impact on housing prices compared to PC2. However, both components are significant. There was a slightly better performance on the test data, confirming the model's reliability on unseen data. Finally, the model explains 56.1% of price variance, which is substantial. However, 43.9% remains unexplained.

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

Based on the analysis results, I recommend using the initial linear regression model developed in Task 1. While I see the benefits of using a PCA analysis in some instances, the optimized linear regression explained more of the price variance. The best use case for the PCA-based model is prioritizing the original variables that load heavily on PC1 when making real estate investment decisions. This model could also be deployed as a baseline automated tool to provide initial price estimates. Then the linear regression model could be used when accuracy is the priority. Ultimately, the PCA model is a decent tool that can be used for real estate professionals and home buyers alike.

REFERENCES

Jolliffe, I. (2002) Principal component analysis, 2nd ed.

http://cda.psych.uiuc.edu/statistical_learning_course/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf

Vadapalli, P. (2025). PCA in machine learning: Assumptions, steps to apply & applications.

https://www.upgrad.com/blog/pca-in-machine-learning/