Gaby Masak

D600 – Statistical Data Mining

Task 3: Principal Component Analysis

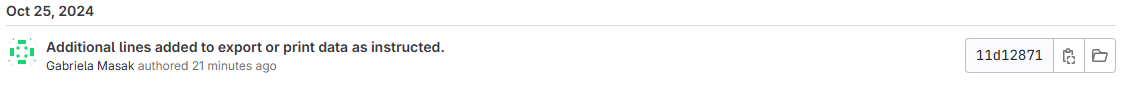
10/5/2024

**Part 1: GitLab Repository URL**

The GitLab repository for this task can be found following this link:

<https://gitlab.com/wgu-gitlab-environment/student-repos/gmasak/d600-statistical-data-mining/-/commit/11d128717cbfb74780101536f455ef3f9aa06bac>

The repository history primarily represents the upload, noting that technical difficulties with both building a pipeline and generating a token from the WGU GitLab led to the majority of the coding to be developed before the initial commit could be made. A secondary submission lead to the following changes:



**Part 2: Purpose of Principal Component Analysis**

Data analytics has many practical applications in every industry. For the real estate industry, linear regression analysis can be used to predict housing prices. It is common knowledge that various features can influence pricing, but exactly which predictors can be hard to ascertain. Analysis can determine not only which variables are influential but can also determine approximately how influential. Linear regression can be used to quantify this question and make future estimates, provided the right set of variables are selected or gathered initially, appropriate analysis steps are taken, and appropriate training and test samples are derived. That said, using too many variables in linear regression can cause overfitting, which is where a dimension reduction technique like Principal Control Analysis is beneficial. Ultimately, analysis can provide answers for real estate research questions of which and how different components affect a house’s price, whether the data is complete or more factors should be assessed, which can then be used practically to meet financial goals by identifying areas where real estate developers should focus time, energy, and funds more efficiently.

PCA can be used to prepare the selected dataset for regression analysis by reducing the dimensionality of the data while retaining most of the original variability. This is achieved by applying orthogonal transformation to the original variables, transforming them into a new set of uncorrelated variables called principal components. These components are ordered such that the first few retain most of the variation present in the original dataset (Aishwarya, 2018). By using these principal components as predictors in the regression model, one can mitigate issues related to multicollinearity and improve the model’s performance. The expected outcomes of using PCA include a more stable and interpretable regression model, reduced overfitting, and potentially improved predictive accuracy. It should be noted that one assumption of PCA is that the principal components are linear combinations of the original variables. This means that PCA assumes the relationships between variables can be captured through linear transformations, which may not always hold true in cases where the underlying data relationships are nonlinear.

**Part 3: Summary of Data Preparation Process**

Twelve variables were selected to calculate whether and how each affects housing prices: price, square footage, number of bathrooms, backyard space, crime rate, school rating, age of home, distance to city center, employment rate, renovation quality, local amenities, and transport access representing the continuous variables from the real estate data. Notably, while number of bedrooms was used in linear regression analysis, this variable features discrete numerical data, which is not appropriate for PCA, and was therefore removed.

Table 1: Descriptive Statistics of Variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Count | Mean | Mode | Range | Max/Min |
| Price | 7000 | 307282 | N/A | 961675.6 | 1046676/85000 |
| Square Footage | 7000 | 1048.95 | 550 | 2324.7 | 2824.7/550 |
| Number of Bathrooms | 7000 | 2.1314 | 1 | 4.81 | 5.80723/1 |
| Backyard Space | 7000 | 511.51 | 300, 418.29, 516.29 | 1630.97 | 1631.36/0.39 |
| Crime Rate | 7000 | 31.23 | 34.01 | 99.7 | 99.73/0.03 |
| School Rating | 7000 | 6.94 | 10 | 9.78 | 10/0.22 |
| Age of Home | 7000 | 46.8 | 18.18, 19.15 | 178.67 | 178.68/0.01 |
| Distance to City Center | 7000 | 17.48 | 8.29 | 65.2 | 65.2/0 |
| Employment Rate | 7000 | 93.71 | 99.9 | 27.85 | 99.9/72.05 |
| Renovation Quality | 7000 | 5 | 10 | 9.99 | 10/0.01 |
| Local Amenities | 7000 | 5.93 | 10 | 10 | 10/0 |
| Transport Access | 7000 | 5.98 | 10 | 9.99 | 10/0.01 |

A black text on a white background

Description automatically generated

Figure 1: Screenshot of Descriptive Statistics of Price

A close-up of a number

Description automatically generated

Figure 2: Screenshot of Descriptive Statistics of Square Footage

A white background with black text

Description automatically generated

Figure 3: Screenshot of Descriptive Statistics of Number of Bathrooms

A black text on a white background

Description automatically generated

Figure 4: Screenshot of Descriptive Statistics of Backyard Space

A white background with black text

Description automatically generated

Figure 5: Screenshot of Descriptive Statistics of Crime Rate

A white background with black text

Description automatically generated

Figure 6: Screenshot of Descriptive Statistics of School Rating

A black text on a white background

Description automatically generated

Figure 7: Screenshot of Descriptive Statistics of Age of Home

A close-up of a number

Description automatically generated

Figure 8: Screenshot of Descriptive Statistics of Distance to City Center

A black text on a white background

Description automatically generated

Figure 9: Screenshot of Descriptive Statistics of Employment Rate

A close-up of a white background

Description automatically generated

Figure 10: Screenshot of Descriptive Statistics of Renovation Quality

A close-up of a number

Description automatically generated

Figure 11: Screenshot of Descriptive Statistics of Local Amenities

A close-up of a sign

Description automatically generated

Figure 12: Screenshot of Descriptive Statistics of Transport Access

**Part 4: Principal Control Analysis**

A screenshot of a computer screen

Description automatically generated

Figure 13: Screenshot of Condensed Matrix of All Principal Components

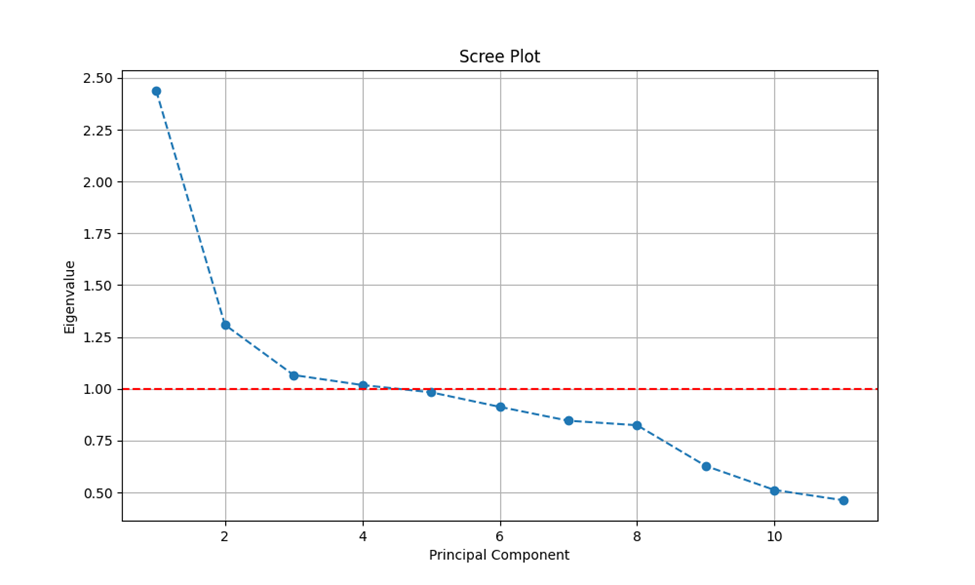


Figure 14: Scree Plot

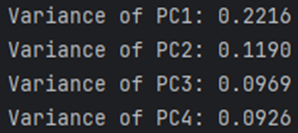


Figure 15: Screenshot of Variance of 8 Principal Components

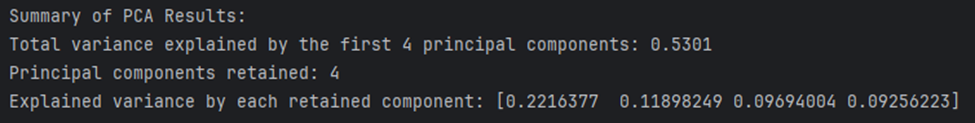


Figure 16: Screenshot of Summary of PCA Results

PCA involved standardizing the data from the independent variables, then transforming the variables into principal components, which are linear combinations of the original variables that capture the maximum variance in the data. The variance of these principal components is then weighed, with the application of the Kaiser rule determining that only the principal components whose Eigenvalues are greater than one should be retained, which reduced the number of principal components from 11 to 4. Figure 14 shows a red dashed line on the y-axis, signifying the appropriate cut off of components. These four principal components above the line are then used as the independent variables in linear regression, after they are separated into training and test sets.

The variance of the PCs are shown in Figure 15. A summary of the results of PCA is shown in Figure 16, also explained by the caption as well as further below in the results section. The variance of the PCs are respectively: 0.2216, 0.1190, 0.0969, and 0.0926. These values are derived from the Eigenvalues seen in the matrix in Figures 13 and 16, further explained in the results section below as explaining the variance of the data along the orthogonal axes.

**Part 5: Analysis and Results**

The 7000 rows were separated randomly into a training set with 5600 rows and a testing set with 1400 rows. These sets can respectively be found in the attached files D600Task3TrainingSet.csv and D600Task3TestSet.csv. After separating the data into training and testing sets, the data can be analyzed.

A screenshot of a computer screen

Description automatically generated

Figure 17: Screenshot of Regression Results



Figure 18: Screenshot of Training and Test MSE

**Part 5: Summary of Data Analysis**

1. Libraries used in Python and R
   1. Python:
      1. pandas: Used to analyze and manipulate large datasets.
      2. numpy: Used to manipulate arrays and matrices.
      3. statsmodels.api: Used for classes and functions of various statistical methods for statistical tests and data exploration.
      4. sklearn.model\_selection: Used for model selection, analyzing data, training and testing models.
         1. train\_test\_split: Used to randomly split dataframe into training and testing sets, specifying the percentage of test data, and grounding the results.
      5. sklearn.linear\_model: Used to predict values using a linear regression model between dependent and independent variables.
         1. LinearRegression: Used to apply ordinary least squares (OLS) linear regression, minimizing the sum of squares.
      6. sklearn.metrics: Used to calculate various prediction errors.
         1. mean\_squared\_error: Used to calculate the mean squared error between real and predicted values.
      7. Scipy.stats: Library used for statistical tasks like probability distributions, correlations, and density estimation.
         1. Shapiro: Used to perform the Shapiro-Wilk test for normality
      8. Sklearn.preprocessing: Used for scaling, centering, and normalizing data.
         1. StandardScaler: Used to scale to unit variance.
      9. Sklearn.decomposition: Library used to apply matrix decomposition algorithms.
         1. PCA: Used to apply Singular Value Decomposition as a linear dimensionality reduction method.
      10. Matplotlob.pyplot: Used to generate plots, like the skee plot.
   2. R:
      1. readxl: Used to import data from Excel and CSV files.
      2. psych: Used to handle calculate descriptive statistics of variables.
2. Method used to Optimize Model and Justification

In order to properly use linear regression for analysis, it is important to make sure that data is not under- or overfitted. First, it is important to make sure that there are enough data points, of which the 7000 included in the dataset for this task are more than enough. Secondly, for a model with multiple variables, it is important to optimize the data by only including relevant inputs, otherwise overfitting can occur. Optimizing the model is an effective means of weighing whether each variable should be included. This was achieved by applying PAC, the elbow rule, then forward stepwise selection which ultimately concluded that only seven of the eight retained principal components were necessary to forecast price: PC1, PC2, PC3, PC4, PC5, PC6, and PC8. PCA involved standardizing the data from the independent variables, then transforming the variables into principal components, which are linear combinations of the original variables that capture the maximum variance in the data. The variance of these principal components is then weighed, with elbow selection determining that only the principal components that explain 80% of the total variance should be retained, which reduced the number of principal components from 11 to 8, as seen in Figure 14. These principal components are then used as the independent variables in linear regression, which is first optimized by forward stepwise selection.

Forward stepwise selection involves iteratively adding variables to the model one at a time based on their statistical significance. At each step, the independent variable that improves the model the most is added, and this process continues until no further significant improvement is observed. This method helps in identifying the most relevant predictors while avoiding overfitting by excluding variables that do not contribute meaningfully to the model’s predictive power. By using forward stepwise selection, the model becomes more robust and interpretable, ensuring that only the most impactful variables are included in the final analysis.

1. Verification of Assumptions Used to Create the Optimized Model

Several assumptions must be met while optimizing a PCA model for validity and reliability: linearity, large sample size, normalization, and independence of principal components. Tackling the assumption of linearity between the independent and dependent variables is verified visually in Task 1, where the scatterplots display a positive linear correlation between price and the respective independent variable. Secondly, the assumption that the sample size is large enough is validated due to the respective training and test sample sizes, which are 5400 and 1600 respectively, sizes that are more than large enough for PAC. Thirdly, the assumption of normalization, which was automated with Python code in order to ensure that each variable had a mean of 0 and a standard deviation of 1. Lastly, the assumption of the independence of principal components is inherently achieved by the PCA process by transforming the principal components into orthogonal variations of one another.

1. Provide the Regression Equation and Discuss the Coefficient Estimates

The Regression Equation can be viewed below:

Price is predicted using four independent variables, the principal components, and a constant. The weight of each predictor variable on price is expressed by its coefficient estimate. The constant, or intercept, is baseline price when all other variables are zero. It represents the starting point of the regression line. The coefficients of the principal components (PC) provide insights into how different combinations of the original variables (like square footage, number of bathrooms, crime rate, etc.) affect housing prices. The positive coefficient of PC1 indicates that as PC1 increases by one unit, the price increases by $58,286.79. PC1 typically captures the largest variance in the data, so it likely represents a combination of factors that significantly influence housing prices. This negative coefficient of PC2 suggests that an increase in PC2 by one unit decreases the price by $22,948.90. PC2 captures the second largest variance and might represent factors that negatively impact housing prices, such as higher crime rates or lower school ratings. Similarly, this negative coefficient indicates that an increase in PC3 by one unit decreases the price by $31,981.63. PC3 might capture another set of factors that negatively influence prices, such as older homes or greater distance from the city center. PC4’s positive coefficient means that an increase in PC4 by one unit increases the price by $18198.89. PC4 might represent factors like better renovation quality or local amenities that positively impact housing prices.

1. Discuss the Model Metrics:
   1. R2 and adjusted R2 of the training set

As seen in Figure 17, the R² and adjusted R² of the training set are 0.459 and 0.459 respectively. It can be inferred that the model explains approximately 45.9% of the variance in housing prices based on the principal components derived from the original variables. This indicates a moderate level of explanatory power, suggesting that while the model captures a significant portion of the variability in housing prices, there are still other factors not accounted for by the principal components that influence the prices. The close values of R² and adjusted R² also imply that the model is not overly complex and does not suffer from overfitting, as the adjusted R² accounts for the number of predictors used in the model. However, there is still a significant portion of the variance that is unexplained, indicating that there may be other factors influencing the dependent variable that are not captured by the current model. As previously mentioned, in answering the specific organizational goal of the PAC model, this suggests that while eleven initial independent variables would presumably contribute to a robust, or even overfitted model, other factors should ultimately be considered too for more accurate analysis.

* 1. Comparison of the MSE for the training set to the MSE of test set

As seen in Figure 18, the MSE values of the training and the test sets are 12380093052.6506 and 11751394519.6545 respectively. This means that the model performs consistently on both the training and test datasets, indicating that it has not overfitted to the training data. The close similarity in MSE values suggests that the model generalizes well to new, unseen data, providing reliable predictions. However, the relatively high MSE values also indicate that there is room for improvement in the model’s accuracy, possibly by incorporating additional relevant features or by fine-tuning the existing model parameters.

1. Discuss Results and Implications for Prediction Analysis

The goal of the PCA was to determine whether the captured factors could determine the price. After optimizing the model by standardizing the data, applying the elbow rule, and further applying ordinary least squares regression analysis to find a regression equation that could predict home prices, it can be determined that seven factors are primarily responsible. The coefficients highlight the importance of features like square footage, number of bathrooms, school rating, and renovation quality in determining housing prices. The model provides a reasonable prediction of housing prices, capturing key factors that influence property values. The moderate R² suggests that there are other factors influencing housing prices not captured by the model. Further exploration and inclusion of additional relevant variables could improve the model’s performance, such as assessing whether data about the selling or buying agent influenced prices. That said, the model was accurate enough that real estate organizations can use this model to make data-driven decisions regarding property valuation, marketing strategies, and investment opportunities. Overall, the model offers valuable insights into the factors driving housing prices, providing a foundation for further refinement and application in real estate analytics.

1. Recommend Course of Action for Real-World Organization

In Economics, value is observed as being the highest price a buyer is willing to pay for an item. In real estate, this concept is especially important as current prices are not always easy to estimate as the pool of demand is limited, and previous data provides a better opportunity for analysis. Listing a house at an appropriate price is imperative as listing too high will result in poor demand, with the listing growing stale and affecting future potential, or listing too low, where the seller risks leaving money on the table and not realizing the full potential value of their property. Therefore, accurate pricing strategies are crucial to balance demand and maximize returns. This knowledge can also be applied to more than simply estimating price. For example, the analysis completed above shows that other factors should be included and weighed that were not originally present in the data. For example, whether or not the house has a pool, whether or not the garage is fenced, age of roof, age of water heater, and whether or not the home is staged. Ultimately, the real estate data is invaluable for making informed decisions that align with market demands and optimize profitability. By leveraging such insights, developers and agents can strategically plan their projects and listings to meet buyer preferences and maximize returns on investment.

**Part 6: Panopto Audiovisual Presentation**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=4c92d0bc-3ff3-4bbe-b2df-b2140155cf52>

**Sources Cited**

1. Aishwarya. (2018, July 7). *Principal Component Analysis(PCA)*. GeeksforGeeks. https://www.geeksforgeeks.org/principal-component-analysis-pca/#