MULTIVARIATE STATISTICAL ANALYSIS OF GLOBAL STUDENT COSTS IN HIGHER EDUCATION

Course : STA4053 - Multivariate Methods II

Name : T.S.W.Pathirana

Reg. Number : S19836

1. **Introduction**

The cost of international education is a critical concern for students and families worldwide. This study aims to analyze the key financial components and underlying patterns in the costs incurred by students pursuing higher education abroad. By applying multivariate statistical techniques-including Principal Component Analysis (PCA), Factor Analysis, Discriminant Analysis, and Canonical Correlation Analysis-this report seeks to uncover the main cost drivers, groupings, and relationships among financial variables, program characteristics, and institutional factors.

1. **Methodology** 
   1. **Dataset**

The analysis uses a structured dataset containing geographic (country, city, university), program (name, level, duration), and financial details (tuition, living cost index, rent, visa fees, insurance, exchange rate), all standardized to USD.

The dataset comprises 907 student records, each with 12 variables

* 1. **Preprocessing Steps**
* Checked for and dropped missing values
* Applied log transformations to variables with high skewness to normalize distributions
* Plotted histograms for each numerical variable to visualize distributions and identify skewness.
* Used the Interquartile Range (IQR) method to identify and remove outliers from the continuous variables
* Computed the total cost for each observation by summing tuition, rent (annualized), insurance (annualized), and visa fees
* Created a categorical variable (cost\_cat) to classify total cost into Low, Medium, and High bins using fixed ranges.
* Used label encoding to convert categorical variables (such as Country, City, University, Program, Level) into numeric codes

**2.3 Statistical Techniques Applied**

* Principal Component Analysis (PCA): Reduced dimensionality and identified main components explaining variance in costs.
* Factor Analysis: Identified latent factors underlying financial variables.
* Discriminant Analysis: Classified programs into cost categories and evaluated predictor importance.
* Canonical Correlation Analysis (CCA): Explored relationships between program characteristics and cost variables.

1. **Results and Discussion** 
   1. **Principal Component Analysis (PCA)**

**A screenshot of a computer screen

Description automatically generatedA graph with a red line

Description automatically generated**

The first four principal components explain approximately 89.7% of the total variance:

* PC1: 46.87%
* PC2: 15.80%
* PC3: 13.90%
* PC3: 13.14%

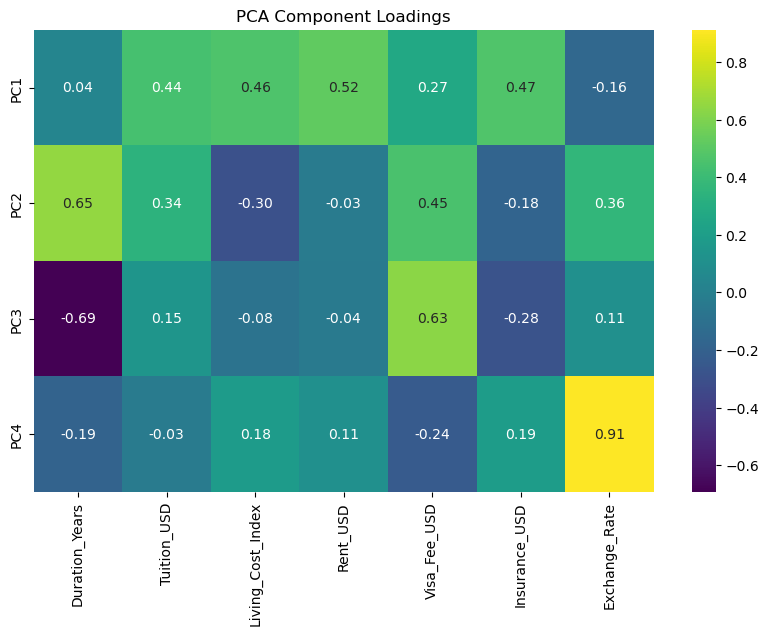
This cumulative explained variance indicates that the majority of the information in the original variables is retained within the first four components. According to statistical best practices, retaining components that together explain at least 70–80% of the variance is generally considered sufficient for dimensionality reduction. Here, the threshold is exceeded, suggesting that the reduced set of components provides a comprehensive summary of the data structure.

**Component Interpretation**

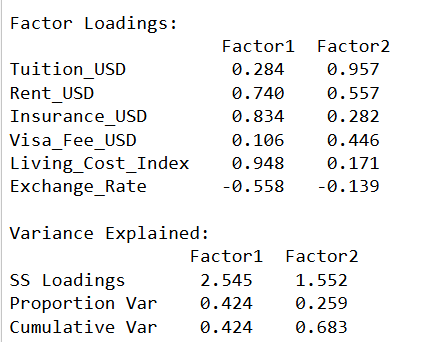
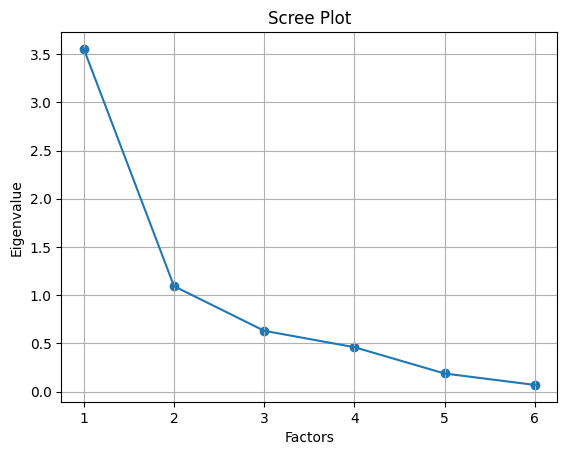
* PC1(General Cost Burden):  
  This component has high positive loadings on Living\_Cost\_Index (0.46), Rent\_USD (0.52), and Insurance\_USD (0.47), as well as a moderate loading on Tuition\_USD (0.43). PC1 represents the overall cost burden, with a particular emphasis on living and accommodation costs.
* PC2 (Program Duration and Visa Fee):  
  PC2 is characterized by a strong loading on Duration\_Years (0.65), as well as Visa\_Fee\_USD (0.45) and Tuition\_USD (0.34). This component reflects the influence of program length, visa and tuition-related expenses.
* PC3 (Academic Duration vs. Visa):  
  PC3 is driven by a very high loading on Duration\_Years (0.69) and a strong negative loading on Visa\_Fee\_USD (-0.63), suggesting a dimension contrasting program duration with visa-related costs.
* PC4 (Exchange Rate)

PC4 is capturing the variability in this dataset that is most strongly associated with the Exchange\_Rate variable, while other variables play a much smaller role

These findings allow institutions and students to focus on the most influential factors affecting the total cost of studying abroad.



* 1. **Factor Analysis**

****

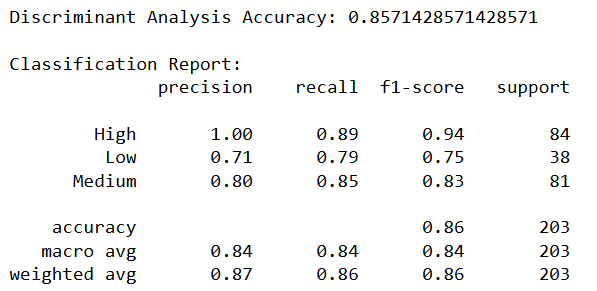
The aim was to reduce the dimensionality of the data and identify latent cost factors that explain the patterns of correlations among these variables.

**Assumption Testing**

* **Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy:** 0.68  
  This value indicates the data are adequate for factor analysis, though not excellent.
* **Bartlett’s Test of Sphericity:** χ² = 2817.63, p < 0.001  
  This highly significant result confirms that the correlation matrix is not an identity matrix, and factor analysis is appropriate.
* The number of factors was determined using the eigenvalue-greater-than-one rule and confirmed by visual inspection of the scree plot.
* Two factors were extracted, together explaining approximately **68% of the total variance** in the financial variables.
* **Factor 1 (Living Costs):**  
  High loadings for Living\_Cost\_Index (0.95), Insurance\_USD (0.83), and Rent\_USD (0.74) indicate this factor represents general living expenses.
* **Factor 2 (Academic Costs):**  
  High loadings for Tuition\_USD (0.96), Visa\_Fee\_USD (0.45), and Rent\_USD (0.56) indicate this factor represents institutional or academic costs.

This structure simplifies the complexity of international education expenses and provides a clearer framework for comparing study destinations and planning budgets.

* 1. **Discriminant Analysis**

****

Discriminant analysis was conducted to classify programs into cost categories (Low, Medium, High) based on financial and program characteristics. The analysis identifies which variables best distinguish between these categories and assesses how accurately the model can predict group membership.

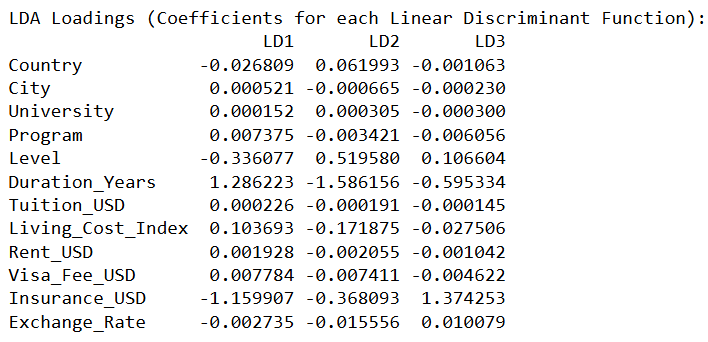
Categorized Total Cost into **High(>= 200000), Medium(50000-200000), Low(<50000)**

**Overall Accuracy:**

The model correctly classified **85.7%** of the cases into the correct total cost category (High, Medium, Low). This is a strong result for a three-class problem.

* **High cost group:**  
  The model is highly effective for identifying high-cost cases (F1 = 0.94), with high precision and recall.
* **Medium cost group:**  
  Performance is also strong (F1 = 0.83).
* **Low cost group:**  
  The model has lower precision and recall here (F1 = 0.75), likely due to fewer low-cost cases or overlap with other groups.

Key Predictors Identified from LDA Loadings:



* Variables with highest absolute values are most influential in separating cost categories:
  + Insurance\_USD (LD1: -1.16, LD3: 1.37)
  + Duration\_Years (LD1: 1.29, LD2: -1.59)
  + Level (LD2: 0.52, LD3: 0.11)
  + Visa\_Fee\_USD (LD1: 0.01)

**Interpretation:**

Insurance cost, duration of study are the strongest predictors of total cost group.

Categorical variables (Country, City, University, Program) have much smaller coefficients, so they have less impact on classification.

**Conclusion:**

The discriminant analysis model shows high accuracy in classifying total cost categories, with insurance and program duration being the most influential predictors. This insight can help students and institutions focus on the most impactful cost components when planning for international education.

* 1. **Canonical Correlation Analysis**

Canonical correlation analysis (CCA) was performed to examine the relationship between two sets of variables.

**X set (program characteristics)** : Duration\_Years, Level, Program, University

**Y set (cost variables)** : Tuition\_USD, Rent\_USD, Insurance\_USD, Visa\_Fee\_USD, Living\_Cost\_Index

**Canonical Coefficients (Loadings)**

The canonical coefficients for each set show how much each original variable contributes to each canonical variate (U₁–U₄ for X, V₁–V₄ for Y).

* First canonical variate for X:

U₁ = -0.6829×Duration\_Years + 0.1248×Level + 0.7035×Program + 0.1525×University

* First canonical variate for Y:

V₁ = -0.5653×Tuition\_USD + -0.5247×Rent\_USD + -0.0678×Insurance\_USD + -0.0596×Visa\_Fee\_USD + 0.6301×Living\_Cost\_Index

**Correlations with Canonical Variates**

* Within X: Program and Duration\_Years have the highest correlations with U₁ (0.7069 and -0.6470), indicating they are most influential in defining the first canonical variate for X.
* Within Y: Tuition\_USD and Rent\_USD have the strongest negative correlations with V₁ (-0.921 and -0.6081)

**Cross-Set Correlations (Reinforcing the Results)**

* Correlations between each X variable and all canonical variates for Y (V₁–V₄):

All correlations are relatively low (e.g., max 0.2449 for Program with V₁), X variables do not strongly predict the Y canonical variates on their own.

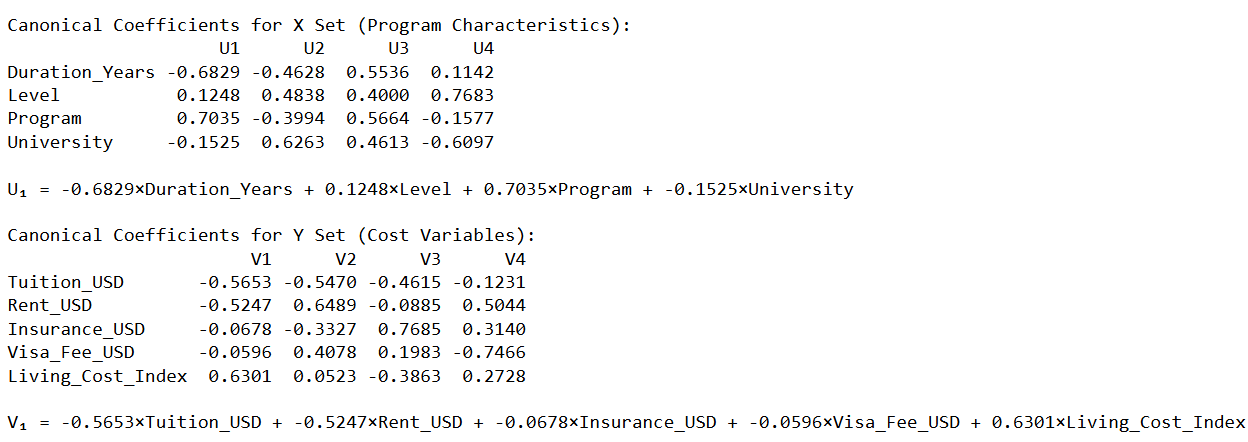
* Correlations between each Y variable and all canonical variates for X (U₁–U₄):

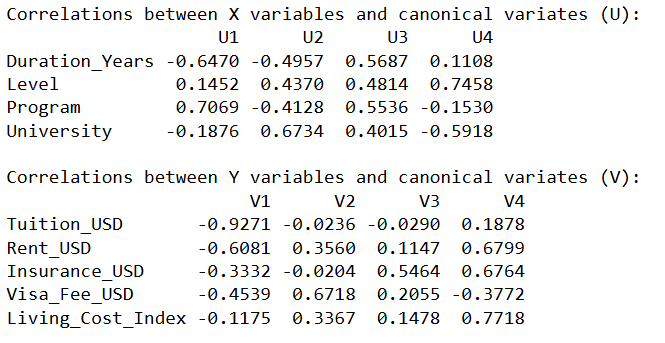
Similarly, the highest is Tuition\_USD with U₁ (-0.3211), which is moderate at best.

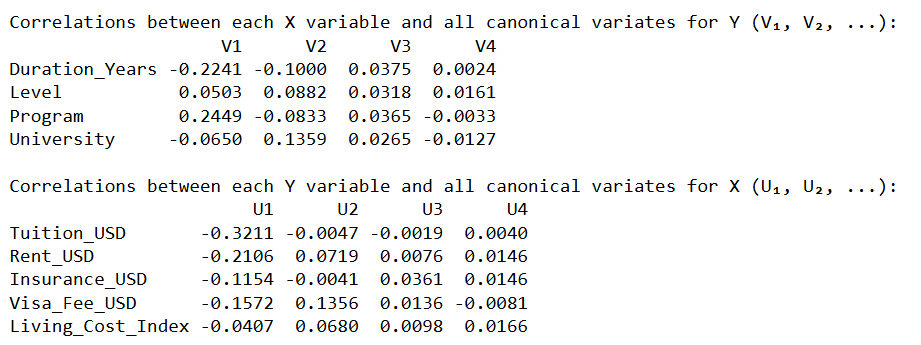
The first canonical variate pair (U₁, V₁) captures the strongest relationship between the two sets, as seen by the relatively high loadings and correlations for certain variables. Variable Importance:

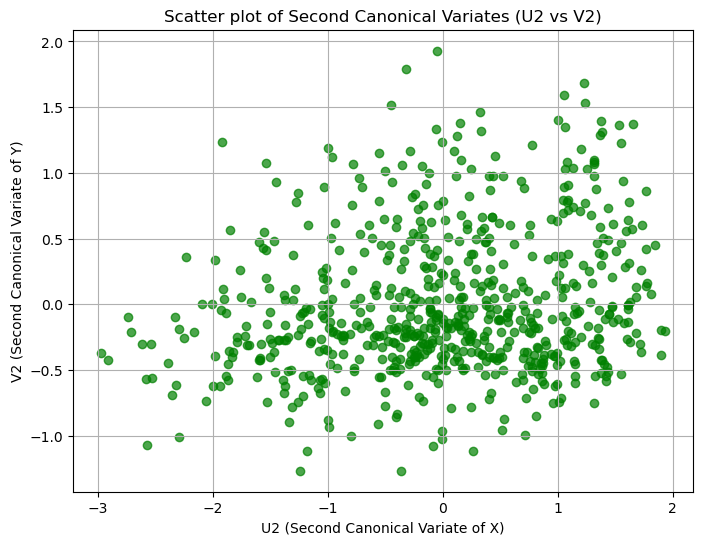
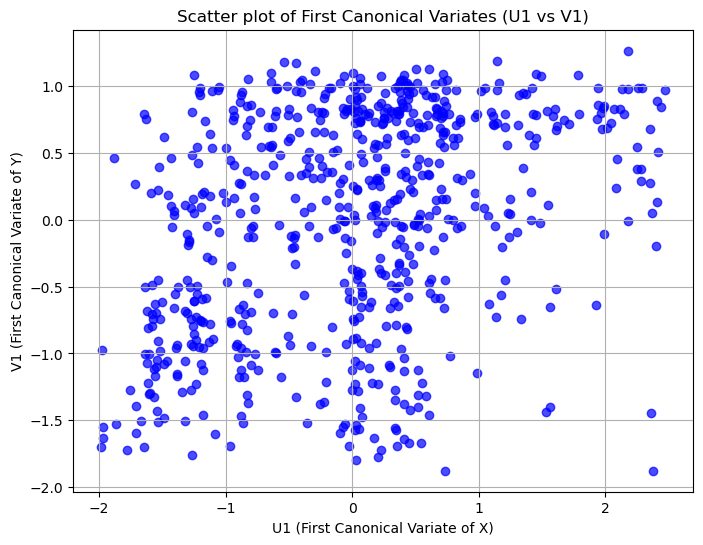
In X, Program and Duration\_Years are most influential for U₁.

In Y, Tuition\_USD and Rent\_USD are most influential for V₁.









* The scatter is even more diffuse and dispersed.
* This indicates the second canonical correlation is even weaker than the first, and there is little to no relationship between these second canonical variates.
* The canonical variate pairs do not show a strong relationship-the main information is that, in my dataset, the optimal linear combinations of program characteristics and cost variables are not highly correlated.

1. **Conclusion and Recommendation**

This analysis of the Cost of International Education dataset provides a data-driven understanding of the financial landscape faced by international students. By applying a suite of multivariate techniques, I uncovered the main cost structures, identified the most influential financial variables, and clarified how program and institutional characteristics relate to overall student expenses.

**Key Insights:**

* **Distinct Cost Structures:**  
  Factor analysis revealed that student expenses are primarily organized around two central themes:
  1. **Living Costs** (dominated by living cost index, rent, and insurance)
  2. **Academic Costs** (driven by tuition and visa fees)  
     This distinction helps students and stakeholders focus on the most impactful financial categories when planning or comparing study abroad options.
* **Dimensionality Reduction Success:**  
  PCA demonstrated that nearly 90% of the variance in financial data can be explained by the first four principal components, confirming that a small number of underlying factors capture the majority of cost variation across programs and locations.
* **Predictive Classification:**  
  Discriminant analysis showed strong performance (85% accuracy) in classifying programs into low, medium, and high total cost categories. The most important predictors were insurance cost, program duration, and exchange rate, highlighting their outsized influence on overall affordability.
* **Program-Cost Linkages:**  
  Canonical correlation analysis indicated that program features-particularly program type and duration-are moderately associated with cost variables, especially tuition and living expenses. This suggests that both academic choices and geographic context play a meaningful role in shaping students’ financial commitments.

**Limitations:**

* **Variable Scope:**  
  The dataset does not include all possible student expenses (e.g., food, transportation, personal spending), nor does it account for scholarships or financial aid.
* **Static Snapshot:**  
  The analysis reflects costs at a single point in time and may not capture fluctuations due to exchange rates, inflation, or policy changes.
* **Data Source Variation:**  
  Differences in how universities and countries report costs could introduce inconsistencies.

**Recommendations:**

* When comparing study destinations, consider both living and academic costs, as both substantially affect total financial requirements.
* Offer transparent, detailed breakdowns of all major cost components, and update them regularly to reflect market changes.
* Use these insights to design targeted financial support or information campaigns, particularly for high-cost programs or locations.
* Future studies should incorporate additional variables-such as scholarships, personal expenses, and local economic indicators-to further enhance the explanatory power of multivariate models.

1. **References**

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson Prentice Hall.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis* (3rd ed.). Wiley.

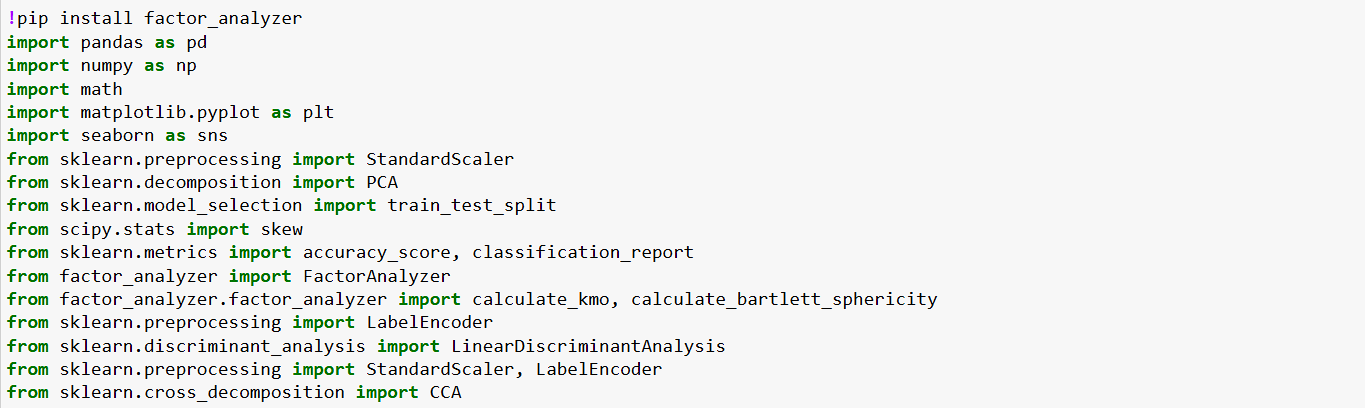
**6. Appendices**

* Dataset : Cost of International Education: Comparative Financial Dataset for Global Study

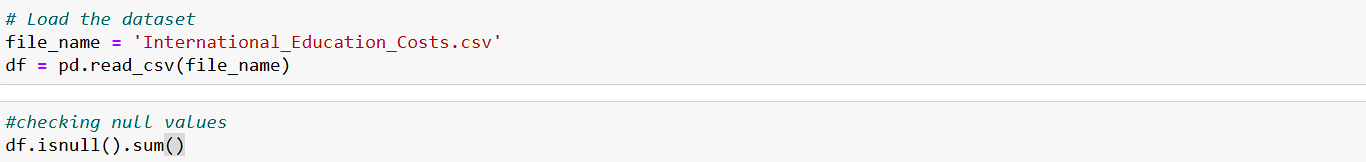
<https://www.kaggle.com/datasets/adilshamim8/cost-of-international-education/data>

* Python Code :

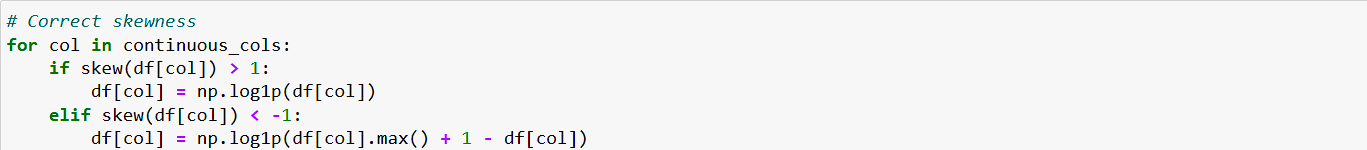
**Import Libraries**



**Exploratory Data Analysis**







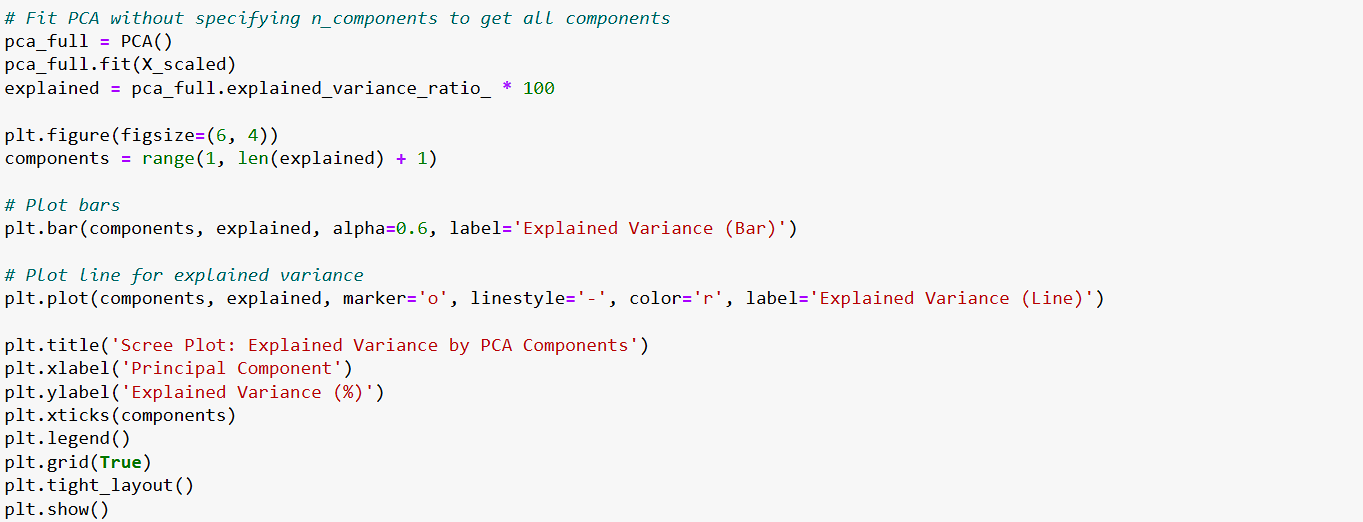
A close-up of a white background

Description automatically generated



**A computer screen shot of a computer code

Description automatically generatedPrincipal Component Analysis**

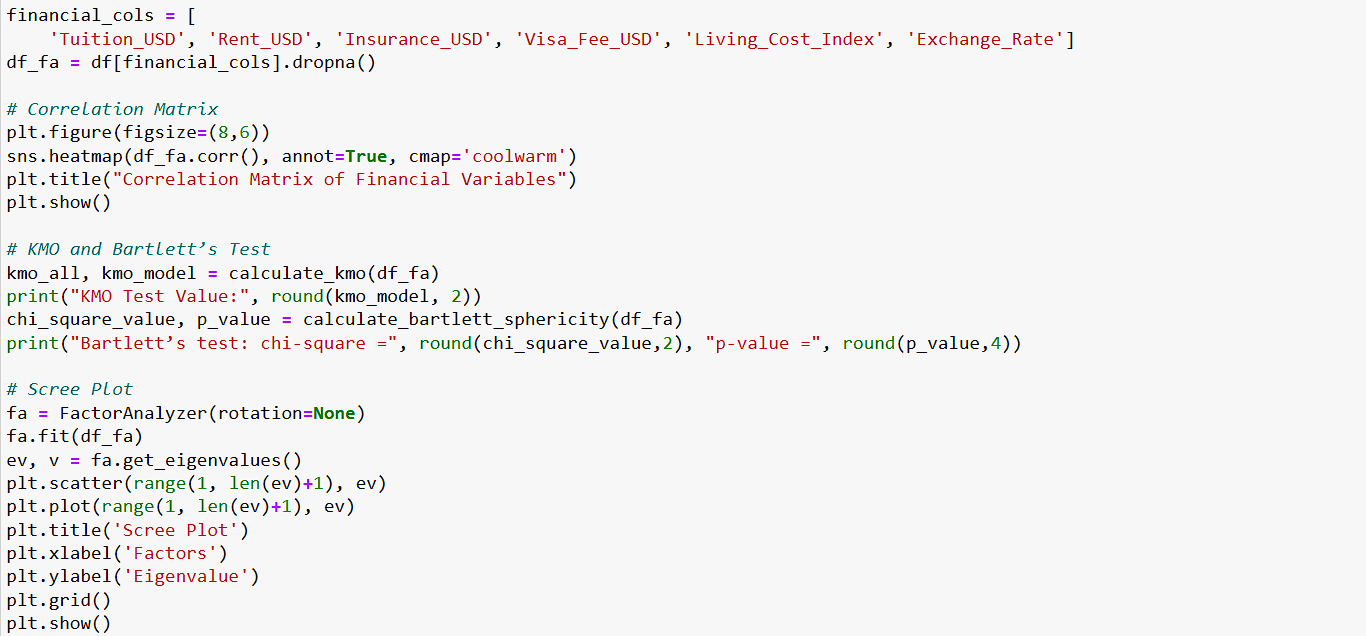
****

**A close-up of a computer code

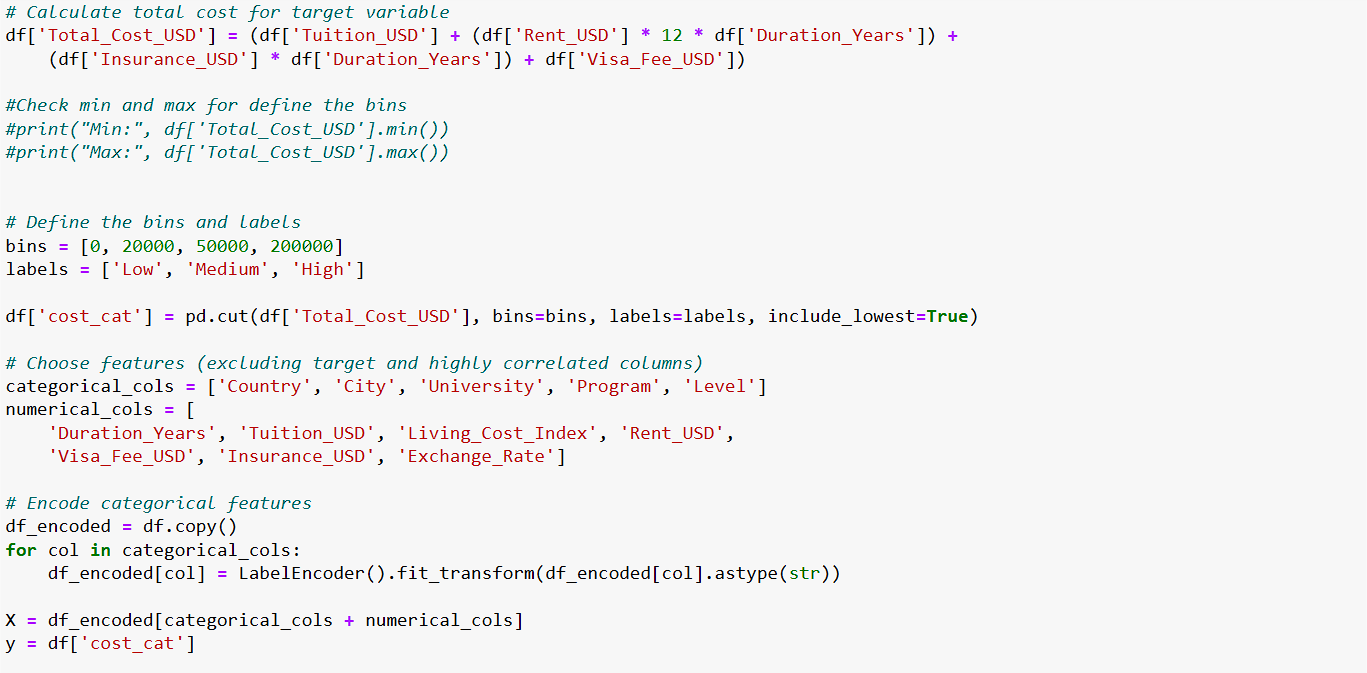
Description automatically generated**

**A group of text on a white background

Description automatically generatedFactor Analysis**

****

**Discriminant Analysis**

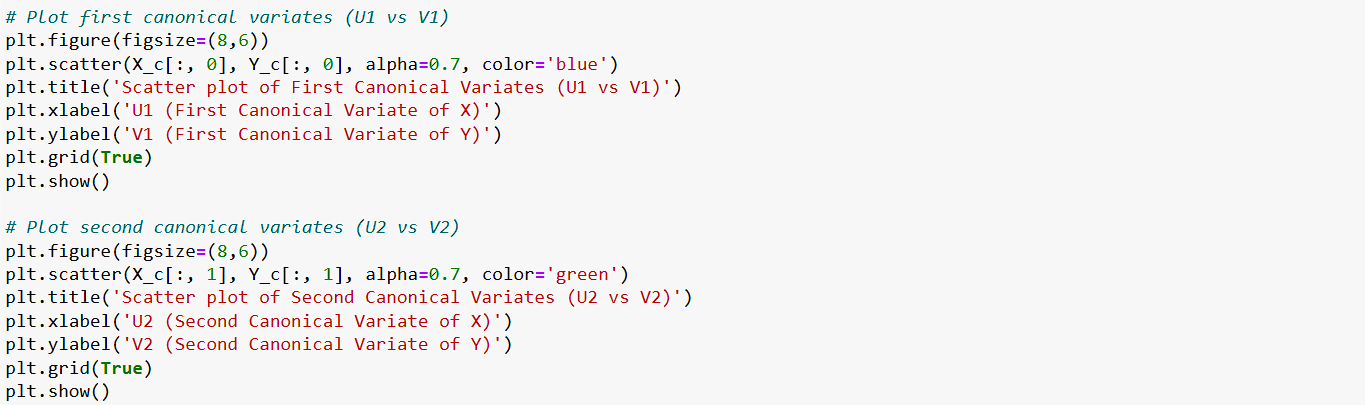
****

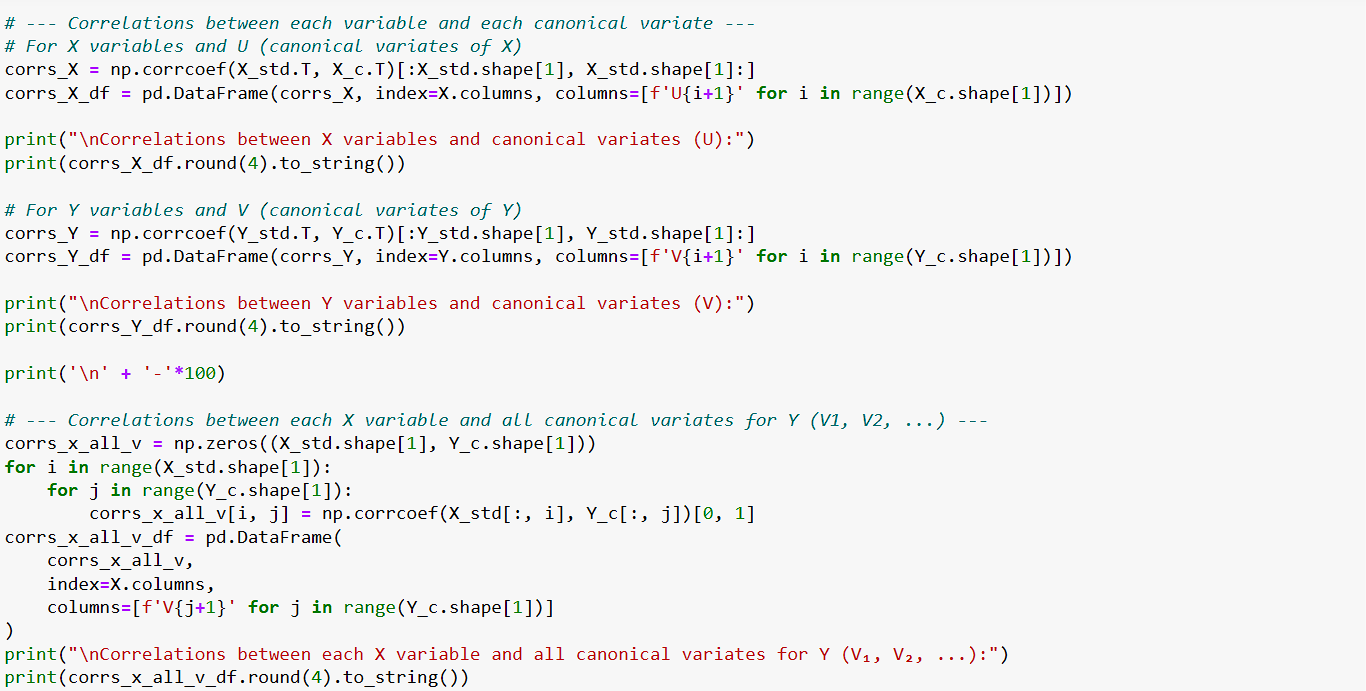
****

**Canonical Correlation Analysis**

**A screenshot of a computer code

Description automatically generated**

**A computer screen shot of a code

Description automatically generated**