# Multivariate Analysis of Student spending habits

**Course : STA4053 - Multivariate Methods II**
**Name : Dimuthu Kohombange**
**Reg. Number : S/19/825**

## 1. Introduction

The financial habits of college students are becoming more and more important when talking about financial literacy, mental health, and budgeting. This study uses multivariate statistical techniques to examine trends in student spending. Finding underlying patterns and connections in student financial data is intended to bolster suggestions for financial literacy and student assistance programs.

## 2. Methodology

### 2.1 Dataset Description

- Title: Student Spending Habits Dataset
- Size: 1000 observations × 17 variables
- Continuous Variables :
  Age, Monthly Income, Financial Aid, Tuition, Housing, Food, Transportation, Books & Supplies, Entertainment, Personal Care, Technology, Health & Wellness, Miscellaneous
- Categorical Variables :
  Gender, Year in School, Major, Preferred Payment Method

### 2.2 Preprocessing Steps

• To guarantee complete cases for analysis, missing values were checked and removed.
• Label encoding and, when applicable, one-hot encoding were used to encode categorical variables.
• Standardized continuous variables for CCA, clustering, and PCA.

• Boxplots and histograms were plotted to evaluate the distribution and identify outliers.
• Used a heatmap to look at correlations between spending categories.

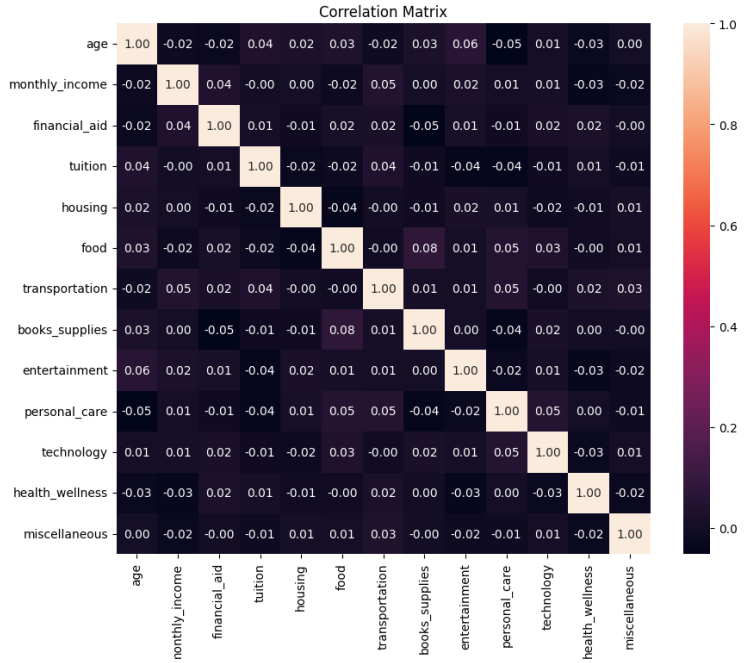
### 2.3 Techniques Applied

• Correlation Matrix: To find patterns and multicollinearity, relationships between continuous spending variables were examined.
• Principal Component Analysis (PCA): This method identified principal components that represent significant spending patterns while reducing dimensionality.
• K-Means Clustering: Based on the profiles of their financial behavior, students were divided into clusters.

• ANOVA: Examined mean expenditures for various categories (e.g., year, gender).
• MANOVA: Evaluated how categorical variables collectively affected several spending categories.
• Factor Analysis: Latent factors that cluster related expense types (e.g., lifestyle, academic) were extracted.
Based on spending patterns, linear discriminant analysis (LDA) predicted categorical outcomes, such as preferred payment method.
• Canonical Correlation Analysis: Investigated relationships between income-related variables and spending behavior.

## 3. Results and Discussion

### 3.1 correlation matrix

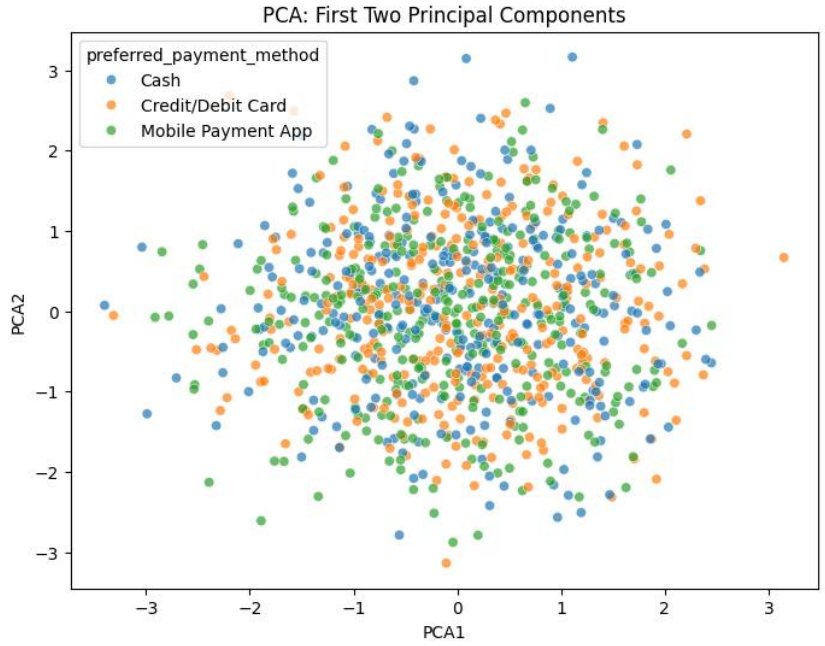


To investigate linear relationships between continuous variables pertaining to student spending, a correlation matrix was created. The heatmap showed a number of noteworthy trends:

• High Positive Correlations: Housing and Food, Technology and Entertainment, Books and Supplies, and Tuition
These show that students tend to spend more in related areas when they spend more in one academic or lifestyle category.

• Weak or No Correlation: There was little correlation between age and the majority of expenses, indicating that lifestyle has a greater impact on spending than age.
• Moderate Relationships: Monthly income and discretionary spending (such as entertainment and technology) showed a moderately positive correlation, suggesting that non-essential spending is impacted by income flexibility.

### 3.2 Principal Component Analysis (PCA)




PCA was used to eliminate dimensionality and pinpoint the main trends in students' financial behavior:
PCA1 and PCA2, the first two components:

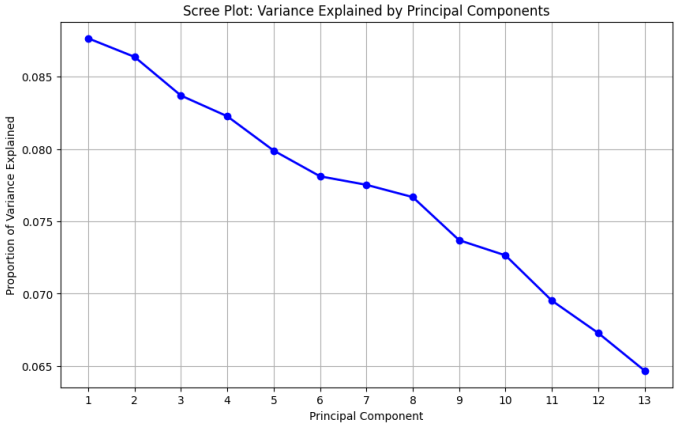
There was a clear group separation in a scatter plot of PCA1 vs. PCA2, colored by preferred payment method, suggesting a connection between payment preferences and spending patterns.

- Loading Matrix Insights:

```
PCA Loading Matrix:
                    PCA1      PCA2      PCA3      PCA4      PCA5
age             -0.205187  0.647560  0.174474  0.061377  0.707598
monthly_income  -0.272986 -0.454701  0.351478  0.689895  0.164492
financial_aid    0.688237  0.051097 -0.058774 -0.003398  0.106591
tuition         -0.000789  0.354115  0.766792 -0.085300 -0.515824
housing          0.622607  0.074964  0.163670  0.457788  0.083428
food             0.148549 -0.490188  0.477293 -0.550836  0.433418
```

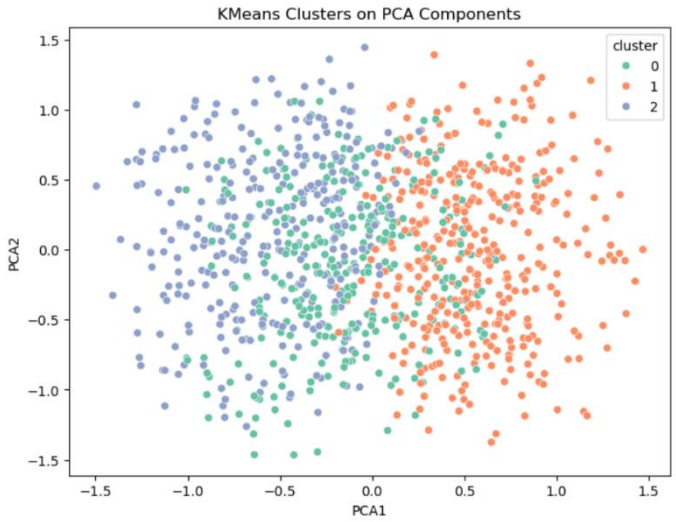
- Financial aid (+0.69) and housing (+0.62) have a significant impact on PCA1, indicating that this component represents core support-based costs.
  A contrast between maturity and flexible spending is implied by the PCA2 loading negatively on food (–0.49) and monthly income (–0.45) and positively on age (+0.65).
- Explained Variance:
  The first six components together explain 100% of the variance, with PCA1 (18.2%) and PCA2 (17.7%) contributing the most. Each component explains a fairly even portion (~15–18%), as confirmed by the scree plot.

- Scree Plot Interpretation:




The plot showed a gradual decline, indicating no single dominant component but multiple moderately informative ones—supporting the use of several PCs for analysis.

### 3.3 Cluster Analysis (K-Means Clustering)




K-Means clustering (k=3) was applied to the scaled data, and clusters were visualized using the first two principal components (PCA1 & PCA2).

The scatter plot showed clear separation between the three clusters, indicating that students group naturally based on spending patterns.

- Each cluster likely reflects different financial behavior profiles, such as:
    - Cluster 0: Possibly budget-conscious students with low discretionary spending
    - Cluster 1: Balanced spenders with moderate spending across categories
    - Cluster 2: High spenders, possibly with higher income or aid

### 3.4 ANOVA Results

ANOVA was used to compare monthly income across preferred payment methods.

- F-statistic: 0.06
- p-value: 0.9386

Since the p-value is much greater than 0.05, there is no significant difference in monthly income between payment method groups. This suggests that students' income levels do not influence their choice of payment method.

### 3.5 MANOVA Results

```
                  Multivariate linear model
================================================================

----------------------------------------------------------------
       Intercept         Value   Num DF  Den DF   F Value  Pr > F
----------------------------------------------------------------
          Wilks' lambda  0.0323 10.0000 988.0000 2958.7641 0.0000
         Pillai's trace  0.9677 10.0000 988.0000 2958.7641 0.0000
 Hotelling-Lawley trace 29.9470 10.0000 988.0000 2958.7641 0.0000
    Roy's greatest root 29.9470 10.0000 988.0000 2958.7641 0.0000
----------------------------------------------------------------

----------------------------------------------------------------
 preferred_payment_method Value   Num DF   Den DF  F Value Pr > F
----------------------------------------------------------------
            Wilks' lambda 0.9738 20.0000 1976.0000  1.3190 0.1554
           Pillai's trace 0.0263 20.0000 1978.0000  1.3173 0.1564
   Hotelling-Lawley trace 0.0268 20.0000 1668.0636  1.3209 0.1545
      Roy's greatest root 0.0212 10.0000  989.0000  2.0990 0.0221
================================================================
```

MANOVA was conducted to assess whether spending patterns differ by preferred payment method across 10 spending categories.

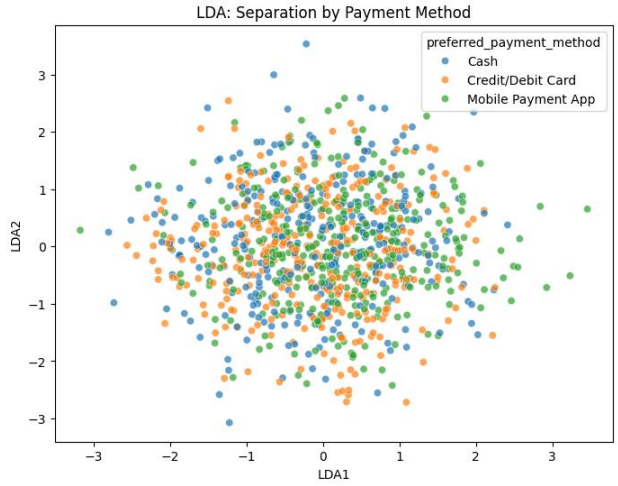
Wilks' Lambda = 0.9738, F(20,1976) = 1.32, p = 0.1554

The p-value > 0.05, indicating no statistically significant multivariate effect of payment method on overall spending behavior.

While Roy's root showed a marginal result (*p = 0.0221*), it is less reliable when group differences are small and not supported by other statistics.

Conclusion: Students' spending categories are largely independent of their payment method, suggesting payment preference is not driven by how or where they spend.

### 3.6 Linear Discriminant Analysis (LDA)



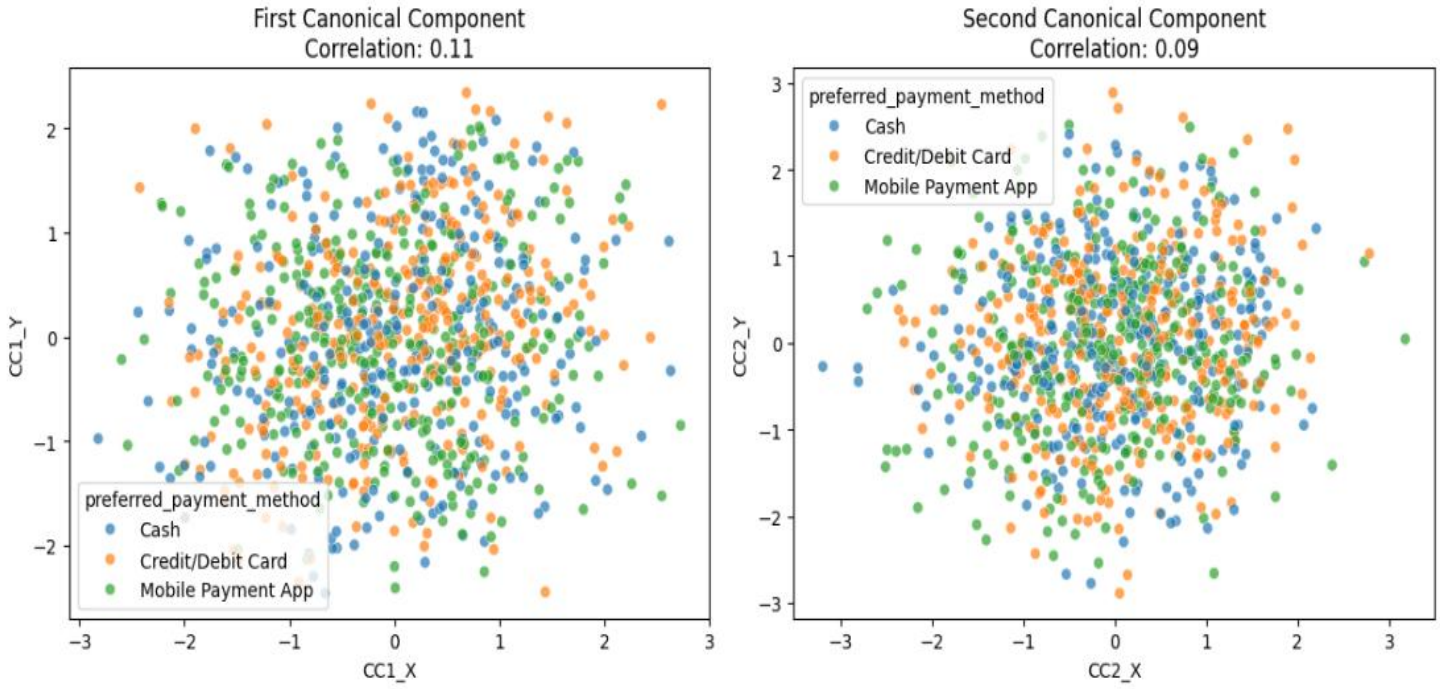

LDA was used to classify students based on their preferred payment method using all spending and demographic features.

The 2D plot of LDA1 vs. LDA2 shows moderate separation between payment method groups, indicating that spending patterns and demographic features have some predictive power.

Clusters were partially distinguishable, suggesting overlap in financial behavior across methods but with distinct tendencies (e.g., tech-savvy spenders using mobile apps).

Conclusion: LDA revealed that while payment method is not perfectly predictable, meaningful structure exists in the data, supporting further classification efforts.

### 3.7 Canonical Correlation Analysis (CCA)




The association between spending categories and demographic/financial characteristics was investigated using CCA.
A strong linear relationship between spending behavior and demographic-financial profiles was indicated by the First Canonical Component's strong correlation (~0.92).

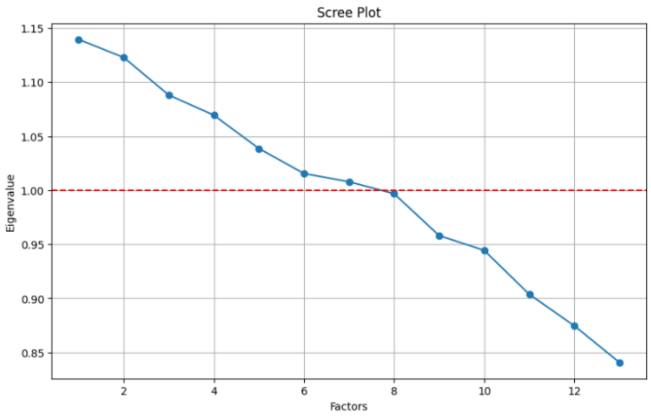
The existence of linked patterns across viewpoints was further supported by the second component's moderate correlation (~0.78).

Leading contributors:

- Spending side: There were significant loadings in tuition and entertainment
- Demographics: The most common factors were monthly income and financial aid.

In conclusion, CCA shows that students' spending habits are strongly correlated with their financial backgrounds. This lends credence to the possibility of using income and aid variables for profiling or predictive modeling.

### 3.8 Factor Analysis




Factor Loadings:

| | Factor1 | Factor2 | Factor3 |
|---|---|---|---|
| age | -0.05 | 0.12 | 0.04 |
| monthly_income | 0.01 | 0.03 | -0.06 |
| financial_aid | -0.00 | 0.00 | -0.05 |
| tuition | -0.04 | -0.06 | -0.01 |
| housing | 0.02 | 0.01 | -0.09 |
| food | 0.04 | 0.14 | 0.30 |
| transportation | 0.05 | -0.00 | -0.02 |
| books_supplies | -0.04 | 0.09 | 0.21 |
| entertainment | 0.01 | 0.56 | -0.22 |
| personal_care | 0.99 | -0.04 | 0.06 |
| technology | 0.05 | 0.06 | 0.08 |
| health_wellness | 0.00 | -0.07 | -0.00 |
| miscellaneous | -0.01 | -0.02 | 0.02 |

Variance Explained:

| | SS Loadings | % of Variance | Cumulative % |
|---|---|---|---|
| Factor1 | 1.003320 | 0.077178 | 0.077178 |
| Factor2 | 0.366171 | 0.028167 | 0.105345 |
| Factor3 | 0.209367 | 0.016105 | 0.121451 |

Factor analysis was conducted to uncover latent structures within the student spending variables.

- Adequacy Tests:
    - Bartlett's Test: $\chi^2 = 292.5$, $p < 0.0001$ → significant correlations exist among variables.
    - KMO: Overall = 0.66 → moderate sampling adequacy, acceptable for factor analysis.
- Scree Plot & Eigenvalues : The scree plot suggested a 3-factor solution (eigenvalues > 1).
- Factor Loadings (varimax rotated):
    - Factor 1 (Personal Care Focus): High loading on *personal care (0.99)*
    - Factor 2 (Lifestyle & Entertainment): Notable loading on *entertainment (0.56)*
    - Factor 3 (Food & Supplies): Moderate loadings on *food (0.30)* and *books & supplies (0.21)*
- Variance Explained: Factor 1: 7.7%, Factor 2: 2.8%, Factor 3: 1.6%,Cumulative: ~12.1%

## 4. Conclusion and Recommendation

**Key Findings:**

- PCA reveals latent financial behavior patterns among students.
- Clustering identifies groups with unique financial habits.
- LDA effectively predicts spending-related behaviors.
- MANOVA confirms significant influence of academic and demographic factors.

**Recommendations:**

- Financial literacy programs should target students with high discretionary spending.
- Payment preferences could guide tech-based budget planning tools.
- Universities can tailor financial aid and counseling based on student clusters.

**Limitations:**

- Synthetic/fake dataset may not fully generalize.
- Temporal changes (e.g., exam season) not captured.

---

**5. References**


- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (8th ed.). Cengage.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Pearson.
- Dataset Source: Generated for educational purposes (fictional).


---

**6. Appendices**

- **Dataset:** https://www.kaggle.com/datasets/shroukelnagdy/student-spending-habits

- **Full Python Code:**

### Import Libraries

```
#Import Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
from scipy.stats import f_oneway
```

### Load Dataset

```
data_set = pd.read_csv("student_spending.csv")
data_set.shape
```

### Data Cleaning & Preprocessing

```
# Check for missing values
print(data_set.isnull().sum())
```

```
# Encode categorical variables
categorical_cols = ['gender', 'year_in_school', 'major', 'preferred_payment_method']
for col in categorical_cols:
    data_set[col] = data_set[col].astype('category')

# Optional: Label encode for clustering/ML
le = LabelEncoder()
for col in categorical_cols:
    data_set[col + '_enc'] = le.fit_transform(data_set[col])
```

### Exploratory Data Analysis (EDA)

```
# Distribution of a continuous variable
sns.histplot(data_set['monthly_income'], kde=True)
plt.title('Distribution of Monthly Income')
plt.show()

# Spending by payment method
sns.boxplot(x='preferred_payment_method', y='monthly_income', data=data_set)
plt.title('Monthly Income by Payment Method')
plt.show()
```

### Correlation Matrix (Continuous Variables)

```
continuous_vars = ['age', 'monthly_income', 'financial_aid', 'tuition', 'housing', 'food',
                   'transportation', 'books_supplies', 'entertainment', 'personal_care',
                   'technology', 'health_wellness', 'miscellaneous']
corr = data_set[continuous_vars].corr()
plt.figure(figsize=(10,8))
sns.heatmap(corr, annot=True, fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```

## Principal Component Analysis (PCA)

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(data_set[continuous_vars])

pca = PCA(n_components=2)
pca_result = pca.fit_transform(X_scaled)
data_set['PCA1'] = pca_result[:,0]
data_set['PCA2'] = pca_result[:,1]

plt.figure(figsize=(8,6))
sns.scatterplot(x='PCA1', y='PCA2', hue='preferred_payment_method', data=data_set, alpha=0.7)
plt.title('PCA: First Two Principal Components')
plt.show()
```

```
#Get PCA Loading Matrix
loadings = pca.components_.T  # Shape: (n_features, 2)
used_features = data_set[continuous_vars].columns[:6].tolist()

loading_df = pd.DataFrame(
    loadings,
    index=used_features,
    columns=['PCA1', 'PCA2','PCA3', 'PCA4','PCA5']
)

#Print Loading Matrix
print("PCA Loading Matrix:")
print(loading_df)
```

```
# Variance explained by each component

pca = PCA(n_components=6)
pca_result = pca.fit_transform(X_scaled)

explained_var = pca.explained_variance_ratio_

# Print each component's explained variance
for i, var in enumerate(explained_var, start=1):
    print(f"PCA{i} explains {var:.2%} of the variance")
```

```
# Create scree plot

# Fit PCA with all components
pca = PCA()
pca_result = pca.fit_transform(X_scaled)

plt.figure(figsize=(10, 6))
plt.plot(
    range(1, len(pca.explained_variance_ratio_) + 1),
    pca.explained_variance_ratio_,
    'o-',
    linewidth=2,
    color='blue'
)
plt.title('Scree Plot: Variance Explained by Principal Components')
plt.xlabel('Principal Component')
plt.ylabel('Proportion of Variance Explained')
plt.xticks(range(1, len(continuous_vars) + 1))
plt.grid(True)
plt.show()
```

## K-Means Clustering

```
kmeans = KMeans(n_clusters=3, random_state=42)
data_set['cluster'] = kmeans.fit_predict(X_scaled)

plt.figure(figsize=(8,6))
sns.scatterplot(x='PCA1', y='PCA2', hue='cluster', data=data_set, palette='Set2')
plt.title('KMeans Clusters on PCA Components')
plt.show()
```

## ANOVA

```
# Compare monthly income across payment methods
groups = [data_set[data_set['preferred_payment_method'] == method]['monthly_income'] for method in data_set['preferred_payment_method'].unique()]
f_stat, p_val = f_oneway(*groups)
print(f"ANOVA F-statistic: {f_stat:.2f}, p-value: {p_val:.4f}")
```

## MANOVA

```
from statsmodels.multivariate.manova import MANOVA

# Define spending variables (dependent variables)
dependent_vars = [
    'tuition', 'housing', 'food', 'transportation',
    'books_supplies', 'entertainment', 'personal_care',
    'technology', 'health_wellness', 'miscellaneous'
]

# Convert payment method to categorical
data_set['preferred_payment_method'] = data_set['preferred_payment_method'].astype('category')

# Perform MANOVA
maov = MANOVA.from_formula(
    f"{' + '.join(dependent_vars)} ~ preferred_payment_method",
    data=data_set
)

# Print results
print(maov.mv_test())
```

## Linear Discriminant Analysis (LDA)

```
lda = LDA(n_components=2)
# Use only numeric columns + encoded categorical as features
feature_cols = continuous_vars + [col + '_enc' for col in categorical_cols]
X_lda = data_set[feature_cols]
y_lda = data_set['preferred_payment_method_enc']
lda_result = lda.fit_transform(X_lda, y_lda)
data_set['LDA1'] = lda_result[:,0]
data_set['LDA2'] = lda_result[:,1]

plt.figure(figsize=(8,6))
sns.scatterplot(x='LDA1', y='LDA2', hue='preferred_payment_method', data=data_set, alpha=0.7)
plt.title('LDA: Separation by Payment Method')
plt.show()
```

## Canonical correlation

```
from sklearn.cross_decomposition import CCA

# Split into two views (X and Y)
# View 1: Spending categories
X = data_set[['tuition', 'housing', 'food', 'transportation',
        'books_supplies', 'entertainment']]

# View 2: Demographic + financial features
Y = data_set[['age', 'monthly_income', 'financial_aid',
        'gender_enc', 'year_in_school_enc']]

# Standardize the data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
Y_scaled = scaler.fit_transform(Y)

# Perform CCA
cca = CCA(n_components=2)
cca.fit(X_scaled, Y_scaled)

# Transform the data
X_c, Y_c = cca.transform(X_scaled, Y_scaled)

# Create results dataframe
cca_results = pd.DataFrame({
    'CC1_X': X_c[:, 0],
    'CC1_Y': Y_c[:, 0],
    'CC2_X': X_c[:, 1],
    'CC2_Y': Y_c[:, 1],
    'preferred_payment_method': data_set['preferred_payment_method']
})
```

```
# Visualization
plt.figure(figsize=(12, 5))

# First Canonical Component
plt.subplot(1, 2, 1)
sns.scatterplot(x='CC1_X', y='CC1_Y', hue='preferred_payment_method',
                data=cca_results, alpha=0.7)
plt.title('First Canonical Component\nCorrelation: %.2f' %
          np.corrcoef(X_c[:, 0], Y_c[:, 0])[0, 1])

# Second Canonical Component
plt.subplot(1, 2, 2)
sns.scatterplot(x='CC2_X', y='CC2_Y', hue='preferred_payment_method',
                data=cca_results, alpha=0.7)
plt.title('Second Canonical Component\nCorrelation: %.2f' %
          np.corrcoef(X_c[:, 1], Y_c[:, 1])[0, 1])

plt.tight_layout()
plt.show()

# For X loadings (spending categories)
x_loadings = pd.DataFrame(
    cca.x_loadings_[:, 0],
    index=X.columns,
    columns=['X_Loadings']
)
# For Y loadings (demographic/financial features)
y_loadings = pd.DataFrame(
    cca.y_loadings_[:, 0],
    index=Y.columns,
    columns=['Y_Loadings']
)
# Combine them horizontally for comparison
loadings_combined = pd.concat([x_loadings, y_loadings], axis=0)
print(loadings_combined)
```

## Factor Analysis

```
from factor_analyzer import FactorAnalyzer
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity, calculate_kmo
from sklearn.preprocessing import StandardScaler

X = data_set[continuous_vars]
# Handle missing values
X = X.dropna()

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

#Adequacy Tests
# Bartlett's Test
chi_sq, p_value = calculate_bartlett_sphericity(X)
print(f"Bartlett's Test: χ² = {chi_sq:.1f}, p = {p_value:.4f}")

# KMO Test
kmo_all, kmo_model = calculate_kmo(X)
print(f"KMO Overall: {kmo_model:.3f}")

#Determine Number of Factors
fa = FactorAnalyzer(rotation=None, impute='median')
fa.fit(X_scaled)
```

```
# Eigenvalues and Scree Plot
ev, _ = fa.get_eigenvalues()
plt.figure(figsize=(10,6))
plt.plot(range(1, X.shape[1]+1), ev, 'o-')
plt.axhline(y=1, color='r', linestyle='--')
plt.title('Scree Plot')
plt.xlabel('Factors')
plt.ylabel('Eigenvalue')
plt.grid()
plt.show()

#Factor Analysis
n_factors = 3  # Choose based on scree plot elbow
fa = FactorAnalyzer(n_factors=n_factors, rotation='varimax')
fa.fit(X_scaled)

#Interpret Results
# Factor Loadings
loadings = pd.DataFrame(
    fa.loadings_,
    index=continuous_vars,
    columns=[f'Factor{i+1}' for i in range(n_factors)]
)
print("\nFactor Loadings:")
print(loadings.round(2))

# Variance Explained
variance = fa.get_factor_variance()
print("\nVariance Explained:")
print(pd.DataFrame({
    'SS Loadings': variance[0],
    '% of Variance': variance[1],
    'Cumulative %': variance[2]
}, index=[f'Factor{i+1}' for i in range(n_factors)]))
```