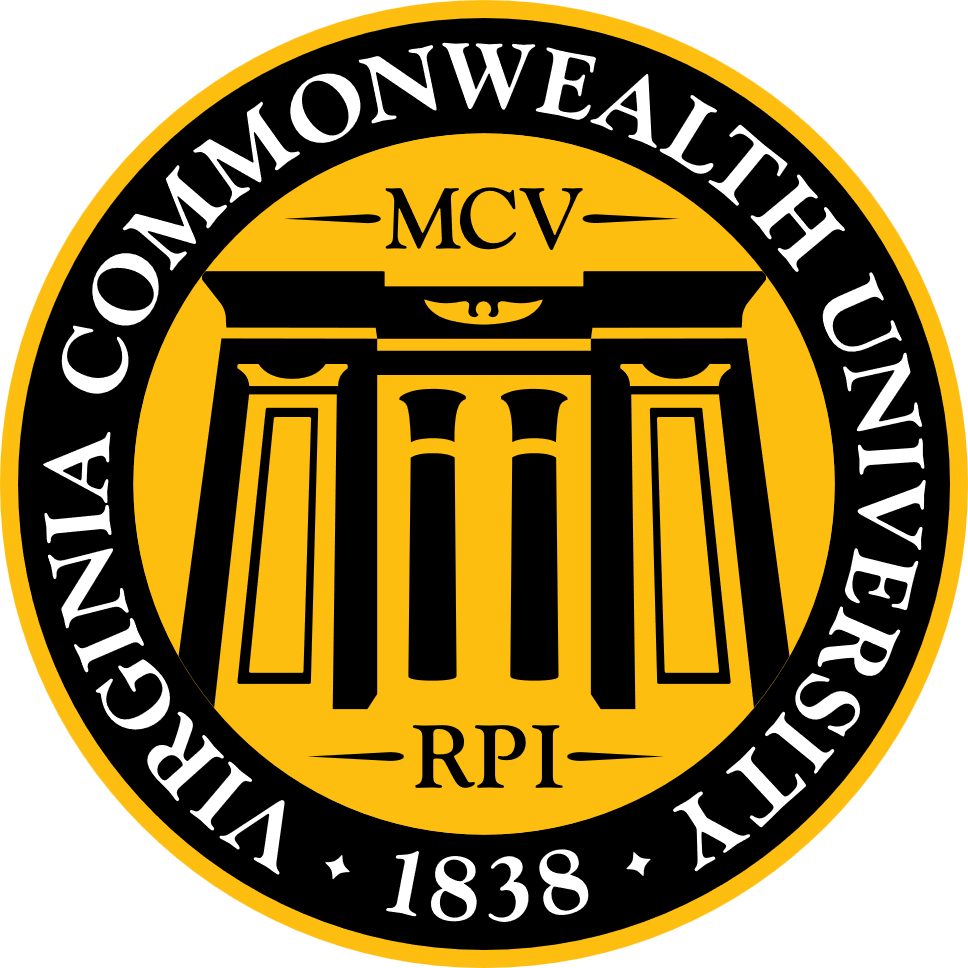
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A4A- Perform Principal Component Analysis and Factor Analysis to identify data dimensions**

**Mithilesh Gurusamy Sivaraj**

**V01107530**

**Date of Submission: 08-07-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | 1 |
| **2.** | Objectives | **1** |
| **3.** | Business Significance | **1-2** |
| **4.** | Results and analysis | **2-19** |
| **5.** | Conclusion |  |
| **6.** | References |  |

**Introduction**

The dataset originates from a survey designed to explore various background variables and their influence on respondent behavior and perceptions. This dataset, which includes a mix of demographic information, preferences, and opinions, provides a comprehensive snapshot of the respondents. The goal of this analysis is to uncover underlying patterns and groupings within the data, utilizing advanced statistical techniques to deepen our understanding.

**Objective**

The primary objectives of this analysis are:

* **Principal Component Analysis (PCA) and Factor Analysis:** To reduce the dimensionality of the dataset and identify the most significant data dimensions that explain the variability in responses. This step will simplify the dataset without losing critical information.
* **Identify Principal Components:** To pinpoint the principal components that capture the maximum variance in the data.
* **Determine Underlying Factors:** To explain the correlations among the observed variables through underlying factors.
* **Simplify the Dataset:** To reduce the number of dimensions while preserving significant information.
* **Provide Insights:** To highlight the key dimensions that influence the data.

**Business Significance**

Understanding the underlying dimensions and respondent segments is crucial for several business applications:

* **Targeted Marketing and Personalization:** By identifying key segments, businesses can customize their marketing strategies to cater to the specific needs and preferences of each group, leading to more effective and personalized campaigns.
* **Product Development and Innovation:** Insights from PCA and factor analysis can reveal the critical factors influencing customer preferences and behavior, guiding product development to better meet market demands.
* **Customer Relationship Management (CRM):** Clustering respondents based on background variables allows for the development of targeted CRM strategies, enhancing customer satisfaction and loyalty by addressing unique customer profiles.
* **Strategic Decision-Making:** A deeper understanding of the customer base enables more informed strategic decisions, from resource allocation to market expansion, ultimately driving business growth and competitiveness.

**Dataset Overview**

The dataset contains 70 entries with 50 columns, encompassing a mix of demographic information, preferences, and opinions related to house purchasing decisions. Key columns include:

* **Categorical Variables:** City, Sex, Age, Occupation, Monthly Household Income, Planning to Buy a New House, Time Frame, Reasons for Buying a House, Type of House, Influence Decision, etc.
* **Numerical Variables:** Income, Number of Rooms, Size of House, Budget, EMI, Proximity to City, Proximity to Schools, etc.

**Principal Component Analysis (PCA)**

**Output from the Principal Function in the Psych Package**

When using the principal function, you obtain several key pieces of information:

1. **Eigenvalues:** These indicate the amount of variance explained by each principal component (PC). Typically, eigenvalues greater than 1 are considered significant.
2. **Loadings:** These show how much each variable contributes to each PC. High loadings (close to 1 or -1) indicate strong contributions.
3. **Rotated Component Matrix:** After rotation (e.g., Promax), the loadings are adjusted for easier interpretation. Variables will ideally load highly onto one component and low on others.

**Interpretation:**

* **Eigenvalues:** Focus on components with eigenvalues > 1. If the first 2-3 components have eigenvalues > 1, they explain most of the variance.
* **Loadings:** Identify which variables load highly on each component. For example, if several related questions about "customer satisfaction" load highly on the first component, you can interpret this component as "Customer Satisfaction".
* **Rotated Matrix:** This matrix clarifies which variables are associated with which components.

**Output from the PCA Function in the FactoMineR Package**

Running the PCA function and using summary(pca) yields:

1. **Eigenvalues and Variance Explained:** These show how much variance each PC explains.
2. **Variable Contributions:** This indicates which variables contribute the most to each component.
3. **Individuals Factor Map:** A plot showing how individual observations (rows) are represented in the new component space.

**Interpretation:**

* **Variance Explained:** The first few components should ideally explain a substantial portion of the variance (e.g., > 60% combined).
* **Variable Contributions:** This helps understand which variables are driving each component.
* **Individuals Factor Map:** This visualization helps identify how observations cluster based on the new components, indicating potential groupings or patterns in the data.

**Factor Analysis**

The omega function provides a hierarchical factor model, offering insights into general and specific factors:

1. **Omega Hierarchical (ωh):** Indicates the general factor saturation.
2. **Omega Total (ωt):** Reflects the total common variance explained by all factors.
3. **Factor Loadings:** Show how each variable loads onto general and specific factors.

**Interpretation:**

* **ωh and ωt:** High ωh suggests a strong general factor, while high ωt indicates that the factors together explain a significant amount of variance.
* **Factor Loadings:** Similar to PCA loadings, these show loadings on both general and specific factors, helping interpret the factors based on the variables with high loadings.

**Biplot Visualization**

Using fviz\_pca\_biplot, you get a biplot showing:

1. **Variables:** Represented as arrows, indicating their direction and strength of influence on the components.
2. **Observations:** Plotted as points, showing how they score on the components.

**Interpretation:**

* **Variable Arrows:** Arrows pointing in the same direction indicate variables that are positively correlated. The length of the arrow indicates the strength of the contribution to the component.
* **Observation Points:** Clustering of points indicates similarity in their responses, while outliers can indicate unique patterns.

**CODES**

# Function to auto-install and load packages

install\_and\_load <- function(packages) {

for (package in packages) {

if (!require(package, character.only = TRUE)) {

install.packages(package, dependencies = TRUE)

}

library(package, character.only = TRUE)

}

}

# Set CRAN mirror

options(repos = c(CRAN = "https://cran.r-project.org"))

# Install necessary packages if not already installed

install.packages(c("FactoMineR", "factoextra"))

# Load necessary libraries

library(FactoMineR)

library(factoextra)

# List of packages to install and load

packages <- c("dplyr", "psych", "tidyr", "GPArotation", "FactoMineR", "factoextra", "pheatmap")

# Call the function

install\_and\_load(packages)

survey\_df<-read.csv("E:/BOOTCAMP/ASSIGNMENTS/SCMA/Survey.csv",header=TRUE)

dim(survey\_df)

names(survey\_df)

head(survey\_df)

str(survey\_df)

#A)Do principal component analysis and factor analysis and identify the dimensions in the data.

is.na(survey\_df)

sum(is.na(survey\_df))

sur\_int=survey\_df[,20:46]

str(sur\_int)

dim(sur\_int)

library(GPArotation)

pca <- principal(sur\_int,5,n.obs =162, rotate ="promax")

pca

om.h<-omega(sur\_int,n.obs=162,sl=FALSE)

op<-par(mfrow=c(1,1))

om<-omega(sur\_int,n.obs=162)

library(FactoMineR)

pca<-PCA(sur\_int,scale.unit = TRUE)

summary(pca)

biplot(pca, scale = 0)

str(sur\_int)

dim(sur\_int)

show(sur\_int)

#Factor Analysis

factor\_analysis<-fa(sur\_int,nfactors = 4,rotate = "varimax")

names(factor\_analysis)

print(factor\_analysis$loadings,reorder=TRUE)

fa.diagram(factor\_analysis)

print(factor\_analysis$communality)

print(factor\_analysis$scores)

**Interpretation of Principal Components Analysis (PCA) Output**

**1. Standardized Loadings (Pattern Matrix)**

* **Loadings:** These numbers indicate the contribution of each variable to each principal component (e.g., RC1, RC2).
  + For instance, "X3..Proximity.to.transport" has a high loading on RC3 (0.77), indicating it significantly contributes to this component.
  + "X4..Proximity.to.work.place" has high loadings on RC4 (0.82) and a negative loading on RC5 (-0.46), suggesting a strong influence on RC4 and a moderate inverse influence on RC5.
* **Communalities (h2):** This column shows the proportion of each variable's variance explained by the five components.
  + For example, "X3..Proximity.to.transport" has a communalities value of 0.58, meaning 58% of its variance is explained by the five components.

**2. Uniqueness (u2) and Complexity (com)**

* **Uniqueness (u2):** This indicates the variance not explained by the components. For "X3..Proximity.to.transport", 42% of its variance is unexplained by the components.
* **Complexity (com):** This measures how many components each variable loads onto. Values closer to 1 indicate that the variable primarily loads on one component.
  + For example, "X3..Proximity.to.transport" has a complexity of 1.2, indicating it mostly loads on a single component.

**3. Sum of Squared Loadings (SS Loadings)**

* **SS Loadings:** These represent the total variance explained by each component.
  + RC1 explains 5.69 units of variance, RC5 explains 4.47 units, etc.
* **Proportion Var:** The proportion of total variance explained by each component.
  + RC1 explains 21% of the total variance.
  + RC5 explains 17% of the total variance.
* **Cumulative Var:** The cumulative variance explained by the components up to that point.
  + By the first component, 21% of the variance is explained.
  + By the second component, 38% (21% + 17%) of the variance is explained.
* **Proportion Explained:** This represents the proportion of explained variance by each component.
  + RC1 explains 35% of the variance among the components.
  + RC5 explains 27% of the variance among the components.
* **Cumulative Proportion:** This represents the cumulative proportion of explained variance.
  + By the first component, 35% of the explained variance is accounted for.
  + By the second component, 62% of the explained variance is accounted for, and so on.

**4. Component Correlations**

* These show the correlations between the components.
  + For instance, RC1 and RC5 have a correlation of 0.50, indicating a moderate positive relationship.
  + RC2 and RC3 have a correlation of -0.19, indicating a weak negative relationship.

**5. Root Mean Square of the Residuals (RMSR) and Chi-Square Test**

* **RMSR:** This value indicates the average residuals (differences) between observed and reproduced correlations. A lower RMSR indicates a better fit. An RMSR of 0.07 suggests a good fit.
* **Chi-Square Test:** This tests the hypothesis that 5 components are sufficient. A higher p-value (e.g., prob < 0.11) suggests that the hypothesis is not rejected, indicating that 5 components may be sufficient to explain the data.

**6. Fit Based on Off-Diagonal Values**

* **Fit:** This indicates how well the model fits the off-diagonal values of the correlation matrix. A value of 0.95 suggests a very good fit.

**Summary**

* **Component Interpretation:**
  + **RC1:** Variables like "Size", "Budgets", and "Maintainances" load highly on RC1, suggesting this component might represent financial or size-related factors.
  + **RC5:** Variables like "X5.Security" and "X4..Availability.of.domestic.help" load highly on RC5, indicating this component might represent security and convenience.
  + **RC2:** Variables like "X2..Booking.amount" and "X5..Availability.of.loan" load highly on RC2, indicating this component might represent financial accessibility.
  + **RC4:** Variables like "X4..Proximity.to.work.place" and "X3.Power.back.up" load highly on RC4, suggesting this component might represent location convenience.
  + **RC3:** Variables like "X3..Proximity.to.transport" load highly on RC3, suggesting this component might represent transport convenience.
* **Variance Explained:**
  + The first five components explain 61% of the total variance, with RC1 explaining the most (21%).
* **Component Correlations:**
  + Components are generally weakly correlated, indicating distinct underlying factors.
* **Model Fit:**
  + The low RMSR (0.07) and high fit based on off-diagonal values (0.95) suggest a good model fit.

**Interpretation of Factor Loadings Matrix**

The factor loadings matrix indicates how each observed variable (such as various features or attributes) correlates with each of the four extracted factors (MR1, MR2, MR3, and MR4). These numbers represent the relationship between the variables and the factors.

* **High Loadings:** High absolute values (e.g., > 0.5) show strong relationships between variables and factors.
* **Low Loadings:** Low absolute values (e.g., close to 0) show weak relationships between variables and factors.

**Breakdown of Loadings**

* **X3..Proximity.to.transport:** Loads highly on MR3 (0.539).
* **X4..Proximity.to.work.place:** Loads moderately on MR4 (0.282).
* **X5..Proximity.to.shopping:** Loads highly on MR1 (0.691) and moderately on MR4 (0.288).
* **X1..Gym.Pool.Sports.facility:** Loads moderately on MR1 (0.467) and MR3 (0.232).
* **X2..Parking.space:** Loads moderately on MR1 (0.520) and MR4 (0.249).
* **X3.Power.back.up:** Loads moderately on MR1 (0.362) and MR4 (0.238).
* **X4.Water.supply:** Loads moderately on MR1 (0.347) and MR4 (0.361), and highly on MR3 (0.660).
* **X5.Security:** Loads highly on MR1 (0.753) and MR3 (0.385).
* **X1..Exterior.look:** Loads highly on MR1 (0.671) and MR4 (0.302), and negatively on MR3 (-0.344).
* **X2..Unit.size:** Loads moderately on MR4 (0.150).
* **X3..Interior.design.and.branded.components:** Loads highly on MR1 (0.612) and MR4 (0.432).
* **X4..Layout.plan..Integrated.etc..:** Loads moderately on MR1 (0.405) and highly on MR4 (0.554).
* **X5..View.from.apartment:** Loads highly on MR1 (0.756) and MR4 (0.329).
* **X1..Price:** Loads moderately on MR4 (0.407) and MR3 (0.438).
* **X2..Booking.amount:** Loads highly on MR2 (0.516) and negatively on MR3 (-0.138).
* **X3..Equated.Monthly.Instalment..EMI.:** Loads highly on MR2 (0.520) and moderately on MR3 (0.249).
* **X4..Maintenance.charges:** Loads moderately on MR2 (0.303) and negatively on MR4 (-0.141).
* **X5..Availability.of.loan:** Loads highly on MR2 (0.872).
* **X1..Builder.reputation:** Loads moderately on MR4 (0.578) and MR1 (0.204).
* **X2..Appreciation.potential:** Loads moderately on MR1 (0.231), MR4 (0.228), and MR2 (0.244).
* **X3..Profile.of.neighbourhood:** Loads highly on MR1 (0.590), MR4 (0.352), and negatively on MR2 (-0.204).
* **X4..Availability.of.domestic.help:** Loads highly on MR1 (0.741).
* **Time:** Loads moderately on MR2 (0.362).
* **Size:** Loads moderately on MR1 (0.510) and highly on MR4 (0.701).
* **Budgets:** Loads moderately on MR1 (0.476), MR4 (0.769), and MR3 (0.109).
* **Maintainances:** Loads moderately on MR1 (0.509), MR4 (0.728), and MR3 (0.146).
* **EMI.1:** Loads moderately on MR1 (0.488) and highly on MR4 (0.775).

**Factor Statistics**

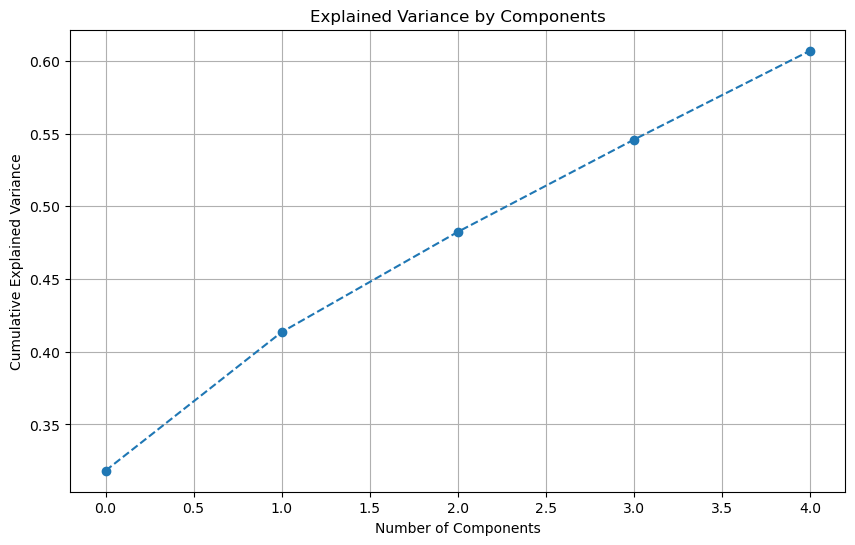
* **Sum of Squared Loadings (SS Loadings):** Sum of squared loadings for each factor.
  + MR1: 5.386
  + MR2: 4.022
  + MR3: 1.908
  + MR4: 1.554
* **Proportion of Variance (Proportion Var):** Proportion of the variance explained by each factor.
  + MR1: 0.199 (19.9%)
  + MR2: 0.149 (14.9%)
  + MR3: 0.071 (7.1%)
  + MR4: 0.058 (5.8%)
* **Cumulative Variance (Cumulative Var):** Cumulative variance explained by the factors.
  + MR1: 0.199 (19.9%)
  + MR2: 0.348 (34.8%)
  + MR3: 0.419 (41.9%)
  + MR4: 0.477 (47.7%)

**Interpretation**

The factors MR1, MR2, MR3, and MR4 represent underlying dimensions or constructs that explain the relationships among the observed variables. For example:

* **MR1:** Likely represents factors related to overall amenities and quality of life (e.g., security, exterior look, gym/pool/sports facility).
* **MR2:** Likely represents financial aspects (e.g., booking amount, EMI, availability of loan).
* **MR3:** Likely represents practical conveniences (e.g., proximity to transport, water supply).
* **MR4:** Likely represents structural and layout features (e.g., layout plan, unit size).

**PYTHON INTREPRETATION**

****

### Cumulative Explained Variance Plot

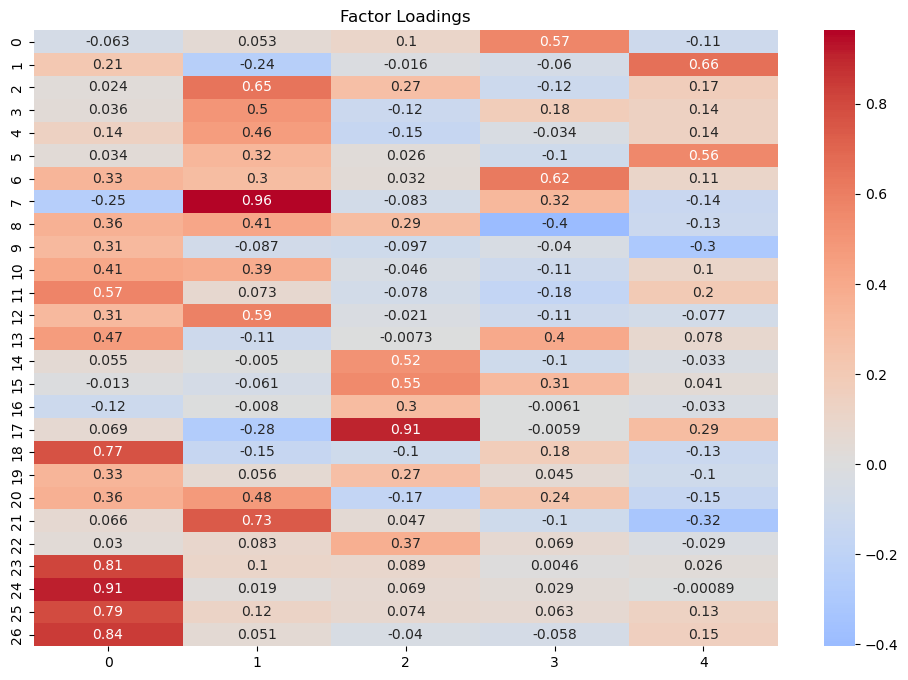
1. **X-Axis (Number of Components)**: This axis represents the number of principal components included in the model, ranging from 1 to the total number of components considered (in this case, 5).
2. **Y-Axis (Cumulative Explained Variance)**: This axis shows the cumulative explained variance, which indicates the proportion of the dataset's total variance that is captured by the selected principal components.
3. **Data Points**: Each point on the graph represents the cumulative explained variance for a given number of components. For instance, the first point shows the variance explained by the first principal component alone, the second point shows the total variance explained by the first two components together, and so on.
4. **Line (Dashed)**: The dashed line connects these data points, illustrating how the cumulative explained variance increases as more components are included.

### Interpretation

* **Starting Point**: The plot starts at zero components, which naturally explains 0% of the variance.
* **Increasing Trend**: As you move to the right (increasing the number of components), the cumulative explained variance increases. This is because each additional principal component accounts for more of the variance in the data.
* **Slope of the Line**: The slope of the line provides insight into the additional variance explained by each new component. If the line is steep, each additional component adds significant new information (variance). If it starts to flatten, additional components contribute less and less new information.

### Specific Observations from the Graph

* **Component 1**: The first component explains about 35% of the variance.
* **Component 2**: Adding the second component increases the cumulative explained variance to about 47%, suggesting the second component explains roughly 12% more of the variance.
* **Component 3**: With three components, around 52% of the variance is explained.
* **Component 4**: Four components explain approximately 57% of the variance.
* **Component 5**: All five components together explain about 60% of the total variance.

****

### Factor Loadings Heatmap

1. **Rows and Columns**:
   * **Rows**: Each row represents a variable from your dataset (labeled 0 to 26, which likely correspond to your survey questions).
   * **Columns**: Each column represents a factor (labeled 0 to 4), which are the underlying latent variables identified by the factor analysis.
2. **Color Scale**:
   * The color scale on the right indicates the magnitude and direction of the factor loadings.
   * **Red**: Positive loadings, with deeper reds indicating stronger positive correlations.
   * **Blue**: Negative loadings, with deeper blues indicating stronger negative correlations.
   * **White/Light Colors**: Loadings near zero, indicating weak or no correlation.
3. **Factor Loadings**:
   * Each cell in the heatmap represents the loading of a particular variable on a particular factor.
   * Loadings can be interpreted as the correlation between the variable and the factor.
   * Higher absolute values indicate a stronger relationship between the variable and the factor.

### Interpretation

1. **Strong Loadings**:
   * Variables with high positive loadings on a factor suggest that they are closely related to that factor. For example, variable 8 has a strong positive loading of 0.96 on factor 1.
   * Variables with high negative loadings suggest an inverse relationship with that factor. For example, variable 7 has a loading of -0.25 on factor 1.
2. **Weak Loadings**:
   * Variables with loadings close to zero have little to no relationship with that factor. For instance, variable 4 has a loading of 0.018 on factor 1, indicating a weak relationship.
3. **Grouping of Variables**:
   * Variables that cluster together (similar color patterns) across factors may represent similar underlying constructs. For example, variables 24, 25, and 26 all have strong positive loadings on factor 0, indicating they may measure a similar construct.

### Detailed Observations

* **Factor 0**: Variables 24, 25, and 26 have very high positive loadings, suggesting they strongly define this factor.
* **Factor 1**: Variables 7, 8, and 17 have strong positive loadings, indicating they are key variables for this factor.
* **Factor 2**: Variables 0, 3, and 11 have moderate to high positive loadings, but no extremely strong defining variables.
* **Factor 3**: Variables 4 and 10 have moderate positive loadings, with variable 10 showing a negative loading on other factors.
* **Factor 4**: Variable 0 has a moderate positive loading, and variable 11 has a moderate negative loading.

Eigenvalues:

[8.59214681 2.57211281 1.8601171 1.71218596 1.65005587 1.35764516

1.28618509 1.04216385 0.93653127 0.79863128 0.73574426 0.61407821

0.55277452 0.53292558 0.50862436 0.40155421 0.33881518 0.29791177

0.27864365 0.23229769 0.20266837 0.14516714 0.11998814 0.09516032

0.06000245 0.04985677 0.02601221]

1. **Eigenvalues**:
   * Each eigenvalue represents the amount of variance in the data that is explained by a corresponding factor or principal component.
   * Higher eigenvalues indicate that the factor/component explains a larger portion of the variance.
2. **List of Eigenvalues**:
   * The provided list shows the eigenvalues in descending order.
   * The first few eigenvalues are significantly larger than the rest, indicating that the corresponding factors/components explain more variance.

### Interpretation

1. **Total Variance**:
   * The sum of all eigenvalues equals the total number of variables (in this case, 27), since each variable contributes one unit of variance.
   * The total variance is partitioned among the factors/components based on the eigenvalues.
2. **Significant Eigenvalues**:
   * Generally, factors/components with eigenvalues greater than 1 are considered significant, as they explain more variance than a single variable.
   * In this list, the first eight eigenvalues are greater than 1, suggesting that up to eight factors/components may be significant.

### Detailed Observations

1. **First Eigenvalue (8.59)**:
   * The first factor/component explains the most variance (8.59 units out of 27), indicating it is the most significant.
2. **Subsequent Eigenvalues**:
   * The second eigenvalue is 2.57, the third is 1.86, and so on, showing that the subsequent factors/components explain progressively less variance.
3. **Drop in Eigenvalues**:
   * There is a notable drop after the first few eigenvalues, indicating that the first few factors/components capture most of the significant variance.
4. **Eigenvalues Less Than 1**:
   * Eigenvalues less than 1 indicate that the corresponding factors/components explain less variance than a single variable, and are usually not considered significant in many contexts.

### Example Calculation

Let's sum the first eight eigenvalues to see the total variance they explain:

8.59+2.57+1.86+1.71+1.65+1.36+1.29+1.04≈20.078.59 + 2.57 + 1.86 + 1.71 + 1.65 + 1.36 + 1.29 + 1.04 \approx 20.078.59+2.57+1.86+1.71+1.65+1.36+1.29+1.04=20.07

This sum indicates that these eight factors/components together explain about 20.07 units of variance out of 27, which is approximately 74% of the total variance. This cumulative explained variance helps in deciding how many factors/components to retain for further analysis.

### Conclusion

* **Number of Factors/Components**:
  + Based on the eigenvalues, the first 8 factors/components (with eigenvalues > 1) are likely the most significant and explain a substantial portion of the variance in the dataset.
* **Cumulative Explained Variance**:
  + To decide on the exact number of factors/components to retain, one could look at the cumulative explained variance (which was plotted in the earlier PCA graph) and decide based on a threshold (e.g., 70-80% of total variance).