****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

**A4a- Perform Principal Component Analysis and Factor Analysis to identify data dimensions**

**N V SATYANARAYAN**

**V01108247**

**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 |  |

**\*NOTE- PYTHON AND R CODES WTH RESULT ADDED IN GITHUB-** [Satyanaldiga (github.com)](https://github.com/Satyanaldiga)

**Introduction**

The dataset at hand originates from a survey aimed at understanding various background variables and their impact on respondent behavior and perceptions. With a mix of demographic information, preferences, and opinions, the dataset provides a comprehensive overview of the respondents. This analysis aims to uncover underlying patterns and groupings within the data, leveraging advanced statistical techniques to enhance 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 help in simplifying the dataset without losing critical information.
* Identify the principal components that capture the maximum variance in the data.
* Determine the underlying factors that explain the correlations among the observed variables.
* Simplify the dataset by reducing the number of dimensions without losing significant information.
* Provide insights into the key dimensions that influence the data.

**Business Significance**

Understanding the underlying dimensions and respondent segments holds substantial business significance:

* Targeted Marketing and Personalization: By identifying key segments, businesses can tailor their marketing strategies to address the specific needs and preferences of each group, leading to more effective and personalized marketing campaigns.
* Product Development and Innovation: Insights from PCA and factor analysis can highlight 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.

The dataset contains 70 entries with 50 columns, encompassing a mix of demographic information, preferences, and opinions related to house purchasing decisions. Here is a summary of the key columns and their data types:

* **Categorical Variables**: City, Sex, Age, Occupation, Monthly Household Income, Planning to Buy a new house, Time Frame, Reasons for buying a house, what 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 principal function in psych package**

When you run the principal function, you get 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 to make interpretation easier. 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, it means 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 makes it clearer which variables are associated with which components.

**Output from PCA function in FactoMineR package**

When you run the PCA function and use summary(pca), you get:

1. **Eigenvalues and Variance Explained**: Similar to the psych package, showing how much variance each PC explains.
2. **Variable Contributions**: This tells you 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**: Look at the contribution of each variable to understand which ones are driving each component.
* **Individuals Factor Map**: This helps visualize 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, giving 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 lot of variance.
* **Factor Loadings**: Similar to PCA loadings, but here you'll see loadings on both general and specific factors. 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, showing 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. Outliers can indicate unique patterns.

**CODES AND EXPLAINATION**

* A function is defined to install and load required R packages if they are not already installed.

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)

+ }

+ }

* The specified packages are installed and loaded.

# List of packages to install and load

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

* The survey data is read from a CSV file into a data frame named survey\_df.
* survey\_df<-read.csv('C:\\Users\\SPURGE\\Desktop\\SCMA\\A4\\Survey.csv',header=TRUE)
* The dimensions, column names, first few rows, and structure of the data are examined to understand its format and contents.

|  |
| --- |
| dim(survey\_df)  [1] 70 50  > names(survey\_df)  [1] "City"  [2] "Sex"  [3] "Age"  [4] "Occupation"  [5] "Monthly.Household.Income"  [6] "Income"  [7] "Planning.to.Buy.a.new.house"  [8] "Time.Frame"  [9] "Reasons.for.buying.a.house"  [10] "what.type.of.House"  [11] "Number.of.rooms"  [12] "Size.of.House"  [13] "Budget"  [14] "Finished.Semi.Finished"  [15] "Influence.Decision"  [16] "Maintainance"  [17] "EMI"  [18] "X1.Proximity.to.city"  [19] "X2.Proximity.to.schools"  [20] "X3..Proximity.to.transport"  [21] "X4..Proximity.to.work.place"  [22] "X5..Proximity.to.shopping"  [23] "X1..Gym.Pool.Sports.facility"  [24] "X2..Parking.space"  [25] "X3.Power.back.up"  [26] "X4.Water.supply"  [27] "X5.Security"  [28] "X1..Exterior.look"  [29] "X2..Unit.size"  [30] "X3..Interior.design.and.branded.components"  [31] "X4..Layout.plan..Integrated.etc.."  [32] "X5..View.from.apartment"  [33] "X1..Price"  [34] "X2..Booking.amount"  [35] "X3..Equated.Monthly.Instalment..EMI."  [36] "X4..Maintenance.charges"  [37] "X5..Availability.of.loan"  [38] "X1..Builder.reputation"  [39] "X2..Appreciation.potential"  [40] "X3..Profile.of.neighbourhood"  [41] "X4..Availability.of.domestic.help"  [42] "Time"  [43] "Size"  [44] "Budgets"  [45] "Maintainances"  [46] "EMI.1"  [47] "ages"  [48] "sex"  [49] "Finished.Semi.Finished.1"  [50] "Influence.Decision.1"  > head(survey\_df)  City Sex Age Occupation Monthly.Household.Income Income  1 Bangalore M 26-35 Private Sector 85,001 to105,000 95000  2 Bangalore M 46-60 Government/PSU 45,001 to 65,000 55000  3 Bangalore F 46-60 Government/PSU 25,001 to 45,000 35000  4 Bangalore M 36-45 Private Sector >125000 200000  5 Bangalore M 26-35 Self Employed 85,001 to105,000 95000  6 Bangalore F 36-45 Private Sector 65,0001 to 85,000 75000  Planning.to.Buy.a.new.house Time.Frame Reasons.for.buying.a.house  1 Yes 6M to 1Yr Residing  2 Yes 6M to 1Yr Investment  3 Yes <6 Months Rental Income  4 Yes <6 Months Investment  5 Yes 1-2 Yr Residing  6 Yes <6 Months Investment  what.type.of.House Number.of.rooms Size.of.House Budget  1 Apartment 2BHK 1001-1400 65.1 to 80L  2 Apartment 2BHK 601-1000 25.1 to 40L  3 Apartment 1BHK <600 <25L  4 Apartment 3BHK 1401-1800 95.1 to110L  5 Apartment 2BHK 601-1000 40.1 to 65L  6 Apartment 2BHK 601-1000 40.1 to 65L  Finished.Semi.Finished Influence.Decision Maintainance EMI  1 Semifurnished Site visits 2001to 4000 35.1K to 50K  2 Semifurnished Newspaper <2000 20.1K to 35K  3 Semifurnished Hoarding <2000 <20K  4 Furnished Electronic/Internet 6001 to 8000 >65K  5 Semifurnished Electronic/Internet 2001to 4000 35.1K to 50K  6 Customized Site visits 2001to 4000 35.1K to 50K  X1.Proximity.to.city X2.Proximity.to.schools X3..Proximity.to.transport  1 3 5 5  2 3 5 5  3 1 2 5  4 4 5 3  5 4 2 3  6 3 2 4  X4..Proximity.to.work.place X5..Proximity.to.shopping  1 2 1  2 3 1  3 2 1  4 5 4  5 4 3  6 4 2  X1..Gym.Pool.Sports.facility X2..Parking.space X3.Power.back.up  1 2 5 3  2 1 4 2  3 4 3 2  4 5 5 4  5 2 4 3  6 3 4 4  X4.Water.supply X5.Security X1..Exterior.look X2..Unit.size  1 5 3 2 4  2 4 3 1 4  3 4 5 1 4  4 5 5 4 4  5 4 4 4 3  6 4 3 3 2  X3..Interior.design.and.branded.components X4..Layout.plan..Integrated.etc..  1 4 4  2 4 2  3 3 2  4 5 5  5 4 4  6 4 3  X5..View.from.apartment X1..Price X2..Booking.amount  1 4 5 1  2 2 5 1  3 2 4 2  4 5 5 2  5 4 4 2  6 3 5 2  X3..Equated.Monthly.Instalment..EMI. X4..Maintenance.charges  1 4 3  2 4 4  3 5 4  4 4 2  5 3 4  6 4 3  X5..Availability.of.loan X1..Builder.reputation X2..Appreciation.potential  1 3 4 5  2 4 5 4  3 2 4 4  4 2 5 4  5 4 4 3  6 3 5 4  X3..Profile.of.neighbourhood X4..Availability.of.domestic.help Time Size  1 4 1 9 1200  2 3 2 9 800  3 4 4 3 400  4 5 5 3 1600  5 4 3 18 800  6 4 3 3 800  Budgets Maintainances EMI.1 ages sex Finished.Semi.Finished.1  1 72.5 30000 42500 30.5 M Semifurnished  2 32.5 120 27500 53.0 M Semifurnished  3 12.5 10000 10000 53.0 F Semifurnished  4 102.5 70000 80000 40.5 M Furnished  5 52.5 30000 42500 30.5 M Semifurnished  6 52.5 30000 42500 40.5 F Customized  Influence.Decision.1  1 Site visits  2 Newspaper  3 Hoarding  4 Electronic/Internet  5 Electronic/Internet  6 Site visits |
|  |
| |  | | --- | | > | |

* The number of missing values in the dataset is counted.
* # Check for missing values
* > sum(is.na(survey\_df))
* [1] 0
* Columns 20 to 46 are selected for analysis.

# Subset the relevant columns for analysis (assuming columns 20 to 46 are relevant)

> sur\_int=survey\_df[,20:46}

* The structure and dimensions of the subset data are checked again, and the GPArotation package is loaded.

# Verify the structure and dimensions of the subset data

> str(sur\_int)

'data.frame': 70 obs. of 27 variables:

$ X3..Proximity.to.transport : int 5 5 5 3 3 4 4 4 5 4 ...

$ X4..Proximity.to.work.place : int 2 3 2 5 4 4 4 3 5 2 ...

$ X5..Proximity.to.shopping : int 1 1 1 4 3 2 3 1 1 2 ...

$ X1..Gym.Pool.Sports.facility : int 2 1 4 5 2 3 4 1 3 4 ...

$ X2..Parking.space : int 5 4 3 5 4 4 5 2 3 4 ...

$ X3.Power.back.up : int 3 2 2 4 3 4 5 3 3 3 ...

$ X4.Water.supply : int 5 4 4 5 4 4 5 4 4 3 ...

$ X5.Security : int 3 3 5 5 4 3 4 1 3 3 ...

$ X1..Exterior.look : int 2 1 1 4 4 3 4 1 3 4 ...

$ X2..Unit.size : int 4 4 4 4 3 2 3 3 3 3 ...

$ X3..Interior.design.and.branded.components: int 4 4 3 5 4 4 5 3 3 4 ...

$ X4..Layout.plan..Integrated.etc.. : int 4 2 2 5 4 3 5 4 3 4 ...

$ X5..View.from.apartment : int 4 2 2 5 4 3 4 1 2 4 ...

$ X1..Price : int 5 5 4 5 4 5 5 5 4 5 ...

$ X2..Booking.amount : int 1 1 2 2 2 2 2 3 2 1 ...

$ X3..Equated.Monthly.Instalment..EMI. : int 4 4 5 4 3 4 5 4 4 5 ...

$ X4..Maintenance.charges : int 3 4 4 2 4 3 4 4 3 4 ...

$ X5..Availability.of.loan : int 3 4 2 2 4 3 4 3 4 4 ...

$ X1..Builder.reputation : int 4 5 4 5 4 5 5 4 4 5 ...

$ X2..Appreciation.potential : int 5 4 4 4 3 4 5 3 4 4 ...

$ X3..Profile.of.neighbourhood : int 4 3 4 5 4 4 4 3 3 4 ...

$ X4..Availability.of.domestic.help : int 1 2 4 5 3 3 3 2 3 2 ...

$ Time : int 9 9 3 3 18 3 9 3 18 3 ...

$ Size : int 1200 800 400 1600 800 800 1600 300 800 1600 ...

$ Budgets : num 72.5 32.5 12.5 102.5 52.5 ...

$ Maintainances : int 30000 120 10000 70000 30000 30000 50000 10000 30000 50000 ...

$ EMI.1 : int 42500 27500 10000 80000 42500 42500 80000 10000 42500 80000 ...

> dim(sur\_int)

[1] 70 27

* PCA is performed on the subset data using the principal function from the psych package, with 5 components and "promax" rotation. The results of the PCA are printed.

# Perform Principal Component Analysis (PCA)

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

> pca

Principal Components Analysis

Call: principal(r = sur\_int, nfactors = 5, rotate = "promax", n.obs = 162)

Standardized loadings (pattern matrix) based upon correlation matrix

RC1 RC5 RC2 RC4 RC3 h2

X3..Proximity.to.transport -0.07 0.06 0.11 -0.17 0.77 0.58

X4..Proximity.to.work.place 0.31 -0.46 0.11 0.82 -0.09 0.65

X5..Proximity.to.shopping 0.06 0.64 0.25 0.19 -0.12 0.66

X1..Gym.Pool.Sports.facility 0.05 0.49 -0.16 0.20 0.23 0.45

X2..Parking.space 0.13 0.50 -0.18 0.19 -0.01 0.46

X3.Power.back.up 0.06 0.23 0.11 0.69 -0.07 0.64

X4.Water.supply 0.38 0.24 0.01 0.10 0.63 0.72

X5.Security -0.16 0.91 -0.18 -0.14 0.33 0.74

X1..Exterior.look 0.31 0.53 0.24 -0.11 -0.36 0.78

X2..Unit.size 0.49 -0.14 -0.17 -0.51 -0.15 0.45

X3..Interior.design.and.branded.components 0.45 0.39 -0.06 0.12 -0.10 0.60

X4..Layout.plan..Integrated.etc.. 0.65 0.02 -0.04 0.24 -0.21 0.59

X5..View.from.apartment 0.33 0.64 -0.05 -0.07 -0.08 0.71

X1..Price 0.61 -0.26 0.04 0.08 0.48 0.54

X2..Booking.amount 0.09 0.00 0.64 -0.06 -0.12 0.47

X3..Equated.Monthly.Instalment..EMI. -0.03 -0.05 0.68 0.01 0.42 0.53

X4..Maintenance.charges -0.13 0.02 0.42 -0.09 0.01 0.22

X5..Availability.of.loan -0.01 -0.20 0.89 0.24 0.00 0.76

X1..Builder.reputation 0.86 -0.18 -0.09 -0.17 0.18 0.67

X2..Appreciation.potential 0.41 0.08 0.37 -0.21 0.08 0.35

X3..Profile.of.neighbourhood 0.43 0.47 -0.21 -0.16 0.25 0.67

X4..Availability.of.domestic.help 0.06 0.83 -0.05 -0.34 -0.11 0.71

Time -0.08 0.23 0.46 -0.05 0.16 0.27

Size 0.74 0.20 0.07 0.04 0.02 0.76

Budgets 0.81 0.16 0.05 0.03 0.05 0.81

Maintainances 0.72 0.20 0.07 0.16 0.08 0.79

EMI.1 0.77 0.13 -0.02 0.18 -0.04 0.81

u2 com

X3..Proximity.to.transport 0.42 1.2

X4..Proximity.to.work.place 0.35 2.0

X5..Proximity.to.shopping 0.34 1.6

X1..Gym.Pool.Sports.facility 0.55 2.1

X2..Parking.space 0.54 1.7

X3.Power.back.up 0.36 1.3

X4.Water.supply 0.28 2.0

X5.Security 0.26 1.5

X1..Exterior.look 0.22 3.1

X2..Unit.size 0.55 2.6

X3..Interior.design.and.branded.components 0.40 2.3

X4..Layout.plan..Integrated.etc.. 0.41 1.5

X5..View.from.apartment 0.29 1.6

X1..Price 0.46 2.3

X2..Booking.amount 0.53 1.1

X3..Equated.Monthly.Instalment..EMI. 0.47 1.7

X4..Maintenance.charges 0.78 1.3

X5..Availability.of.loan 0.24 1.3

X1..Builder.reputation 0.33 1.3

X2..Appreciation.potential 0.65 2.7

X3..Profile.of.neighbourhood 0.33 3.2

X4..Availability.of.domestic.help 0.29 1.4

Time 0.73 1.9

Size 0.24 1.2

Budgets 0.19 1.1

Maintainances 0.21 1.3

EMI.1 0.19 1.2

RC1 RC5 RC2 RC4 RC3

SS loadings 5.69 4.47 2.42 1.88 1.91

Proportion Var 0.21 0.17 0.09 0.07 0.07

Cumulative Var 0.21 0.38 0.47 0.54 0.61

Proportion Explained 0.35 0.27 0.15 0.12 0.12

Cumulative Proportion 0.35 0.62 0.77 0.88 1.00

With component correlations of

RC1 RC5 RC2 RC4 RC3

RC1 1.00 0.50 -0.08 0.16 0.00

RC5 0.50 1.00 0.08 0.29 -0.06

RC2 -0.08 0.08 1.00 -0.16 -0.19

RC4 0.16 0.29 -0.16 1.00 0.09

RC3 0.00 -0.06 -0.19 0.09 1.00

Mean item complexity = 1.8

Test of the hypothesis that 5 components are sufficient.

The root mean square of the residuals (RMSR) is 0.07

with the empirical chi square 252.24 with prob < 0.11

Fit based upon off diagonal values = 0.95

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

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

* **Loadings**: These numbers indicate how much each variable contributes to each principal component (RC1, RC2, etc.).
* For instance, "X3..Proximity.to.transport" has a high loading on RC3 (0.77), suggesting it contributes significantly to this component.
* "X4..Proximity.to.work.place" has high loadings on RC4 (0.82) and negative loading on RC5 (-0.46), suggesting a strong influence on RC4 and a moderate influence in the opposite direction on RC5.
* **Communalities (h2)**: This column shows how much of each variable's variance is 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 loads primarily 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**

**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.

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

> names(factor\_analysis)

[1] "residual" "dof" "chi" "nh"

[5] "rms" "EPVAL" "crms" "EBIC"

[9] "ESABIC" "fit" "fit.off" "sd"

[13] "factors" "complexity" "n.obs" "objective"

[17] "criteria" "STATISTIC" "PVAL" "Call"

[21] "null.model" "null.dof" "null.chisq" "TLI"

[25] "RMSEA" "BIC" "SABIC" "r.scores"

[29] "R2" "valid" "score.cor" "weights"

[33] "rotation" "hyperplane" "communality" "communalities"

[37] "uniquenesses" "values" "e.values" "loadings"

[41] "model" "fm" "rot.mat" "Structure"

[45] "method" "scores" "R2.scores" "r"

[49] "np.obs" "fn" "Vaccounted"

**FACTOR ANALYSIS**

|  |
| --- |
| factor\_analysis$loadings,reorder=TRUE)  Loadings:  MR1 MR4 MR2 MR3  X3..Proximity.to.transport 0.539  X4..Proximity.to.work.place 0.282  X5..Proximity.to.shopping 0.691 0.143 0.288  X1..Gym.Pool.Sports.facility 0.467 0.164 -0.125 0.232  X2..Parking.space 0.520 0.249 -0.143  X3.Power.back.up 0.362 0.238  X4.Water.supply 0.347 0.361 0.660  X5.Security 0.753 -0.101 0.385  X1..Exterior.look 0.671 0.294 0.302 -0.344  X2..Unit.size 0.150 -0.108  X3..Interior.design.and.branded.components 0.612 0.432  X4..Layout.plan..Integrated.etc.. 0.405 0.554  X5..View.from.apartment 0.756 0.329  X1..Price 0.407 0.438  X2..Booking.amount 0.516 -0.138  X3..Equated.Monthly.Instalment..EMI. 0.520 0.249  X4..Maintenance.charges -0.141 0.303  X5..Availability.of.loan -0.146 0.872  X1..Builder.reputation 0.204 0.578 -0.157 0.234  X2..Appreciation.potential 0.231 0.228 0.244  X3..Profile.of.neighbourhood 0.590 0.352 -0.204 0.322  X4..Availability.of.domestic.help 0.741  Time 0.111 0.362  Size 0.510 0.701  Budgets 0.476 0.769 0.109  Maintainances 0.509 0.728 0.146  EMI.1 0.488 0.775  MR1 MR4 MR2 MR3  SS loadings 5.386 4.022 1.908 1.554  Proportion Var 0.199 0.149 0.071 0.058  Cumulative Var 0.199 0.348 0.419 0.477 |
|  |
| |  | | --- | | > | |

### Factor Loadings Matrix

The factor loadings matrix shows how each observed variable (in this case, various features or attributes) loads onto each of the four extracted factors (MR1, MR2, MR3, and MR4). The numbers represent the correlation between the variables and the factors.

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

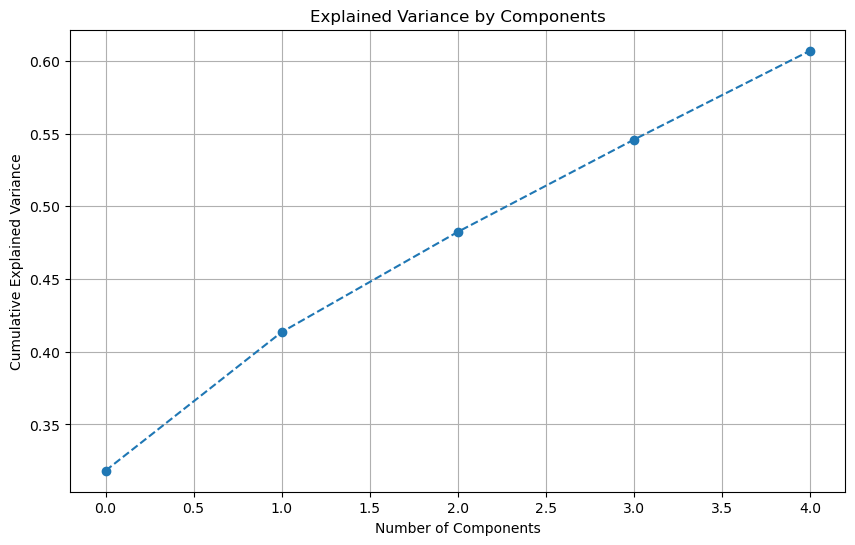
* **SS loadings**: Sum of squared loadings for each factor.
  + MR1: 5.386
  + MR2: 4.022
  + MR3: 1.908
  + MR4: 1.554
* **Proportion Var**: Proportion of the variance explained by each factor.
  + MR1: 0.199 (19.9%)
  + MR2: 0.149 (14.9%)
  + MR3: 0.071 (7.1%)
  + MR4: 0.058 (5.8%)
* **Cumulative Var**: 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**: Might represent factors related to overall amenities and quality of life (e.g., security, exterior look, gym/pool/sports facility).
* **MR2**: Might represent financial aspects (e.g., booking amount, EMI, availability of loan).
* **MR3**: Might represent practical conveniences (e.g., proximity to transport, water supply).
* **MR4**: Might represent structural and layout features (e.g., layout plan, unit size).

**PYTHON INTREPRETATION**

****

### Cumulative Explained Variance Plot

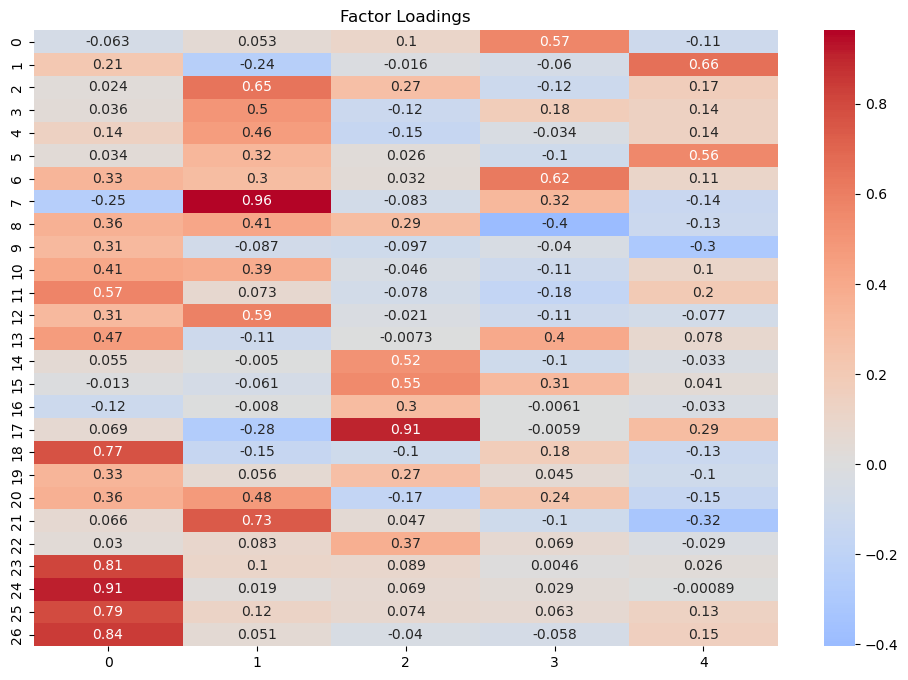
1. **X-Axis (Number of Components)**: This axis represents the number of principal components included in the model, ranging from 1 to the total number of components considered (in this case, 5).
2. **Y-Axis (Cumulative Explained Variance)**: This axis shows the cumulative explained variance, which indicates 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).