****

**VIRGINIA COMMONWEALTH UNIVERSITY**

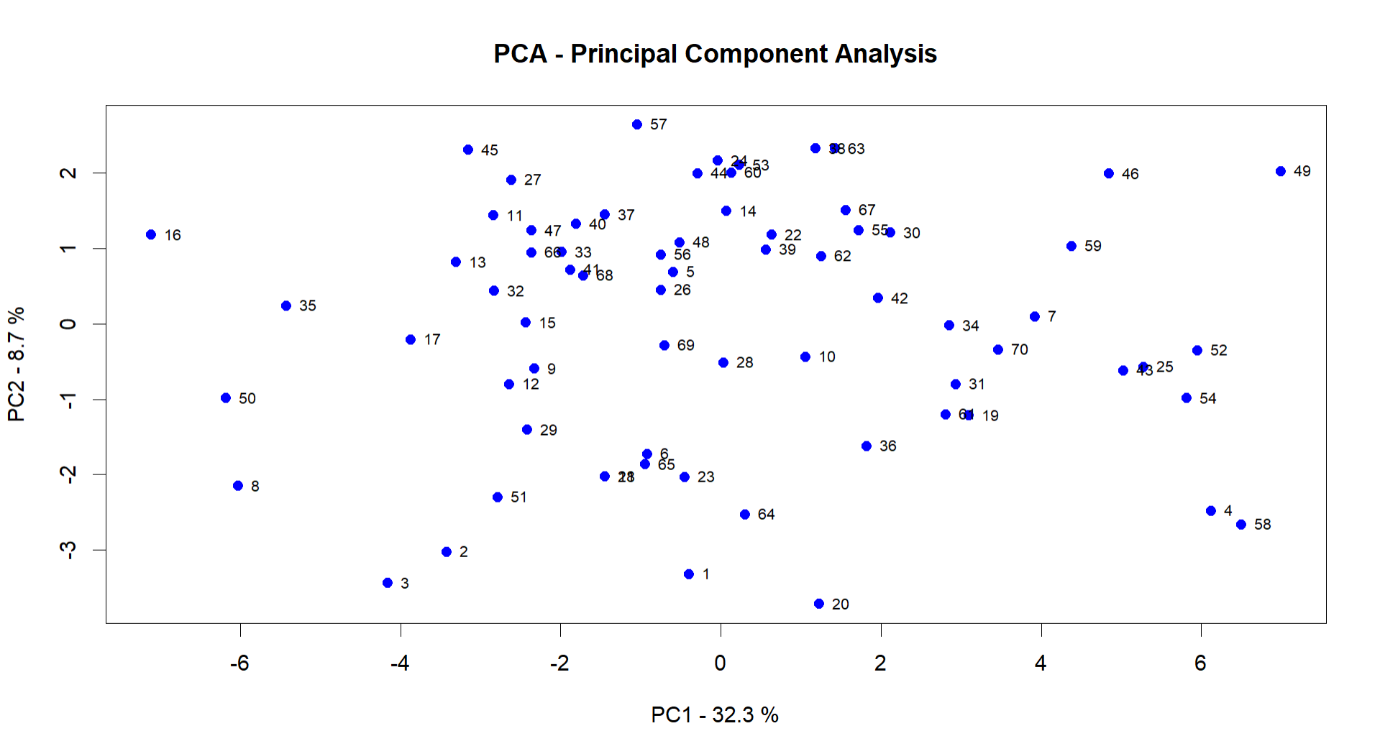
**Statistical analysis and modelling (SCMA 632)**

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

**Rohan Arpit Dungdung**

**V01108247**

**Date of Submission: 15-07-2024**



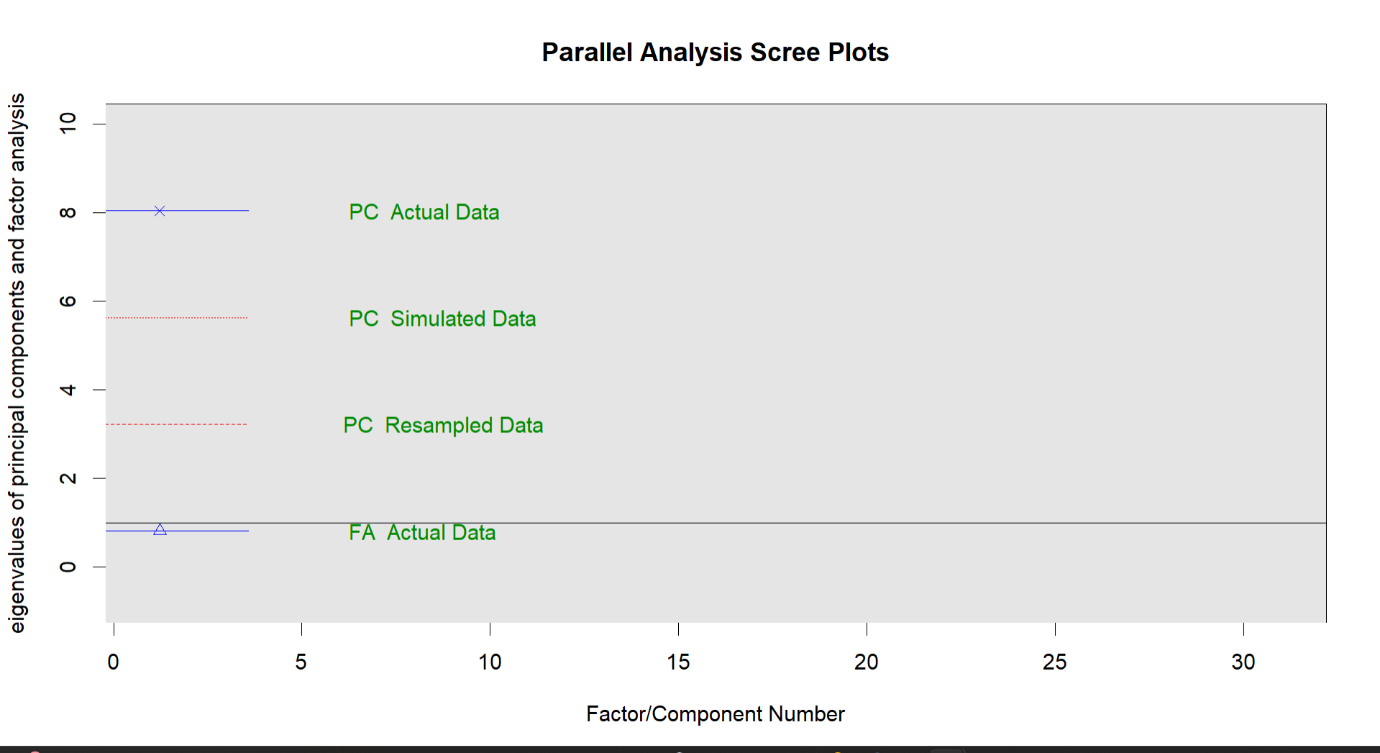
Interpretation

The PCA plot shows how each observation (respondent) is positioned in a two-dimensional space defined by the first two principal components. This helps to visualize the variability and underlying structure in the data, reducing complexity while retaining the most significant information.

The distribution of points suggests that some respondents have similar characteristics, as indicated by their proximity to each other in the PCA plot. For example, observations near the center (around PC1 = 0 and PC2 = 0) likely share common features, while those farther apart (e.g., respondents 4, 49, and 36) are more distinct from the rest.

PCA helps reduce the dimensions of the dataset by transforming the original correlated variables into a set of linearly uncorrelated variables called principal components. This reduction is essential for simplifying data visualization and subsequent analyses, such as clustering.

The PCA plot can also be a preliminary step to identify potential clusters in the data, as clusters might appear as distinct groups of points. In this plot, however, there isn't a clear separation between distinct clusters, which suggests that further clustering methods (like K-means) may be needed to identify any subgroups.



Interpretation

PC Actual Data represents the eigenvalues of the actual data from the PCA. Each eigenvalue indicates the amount of variance explained by a particular component

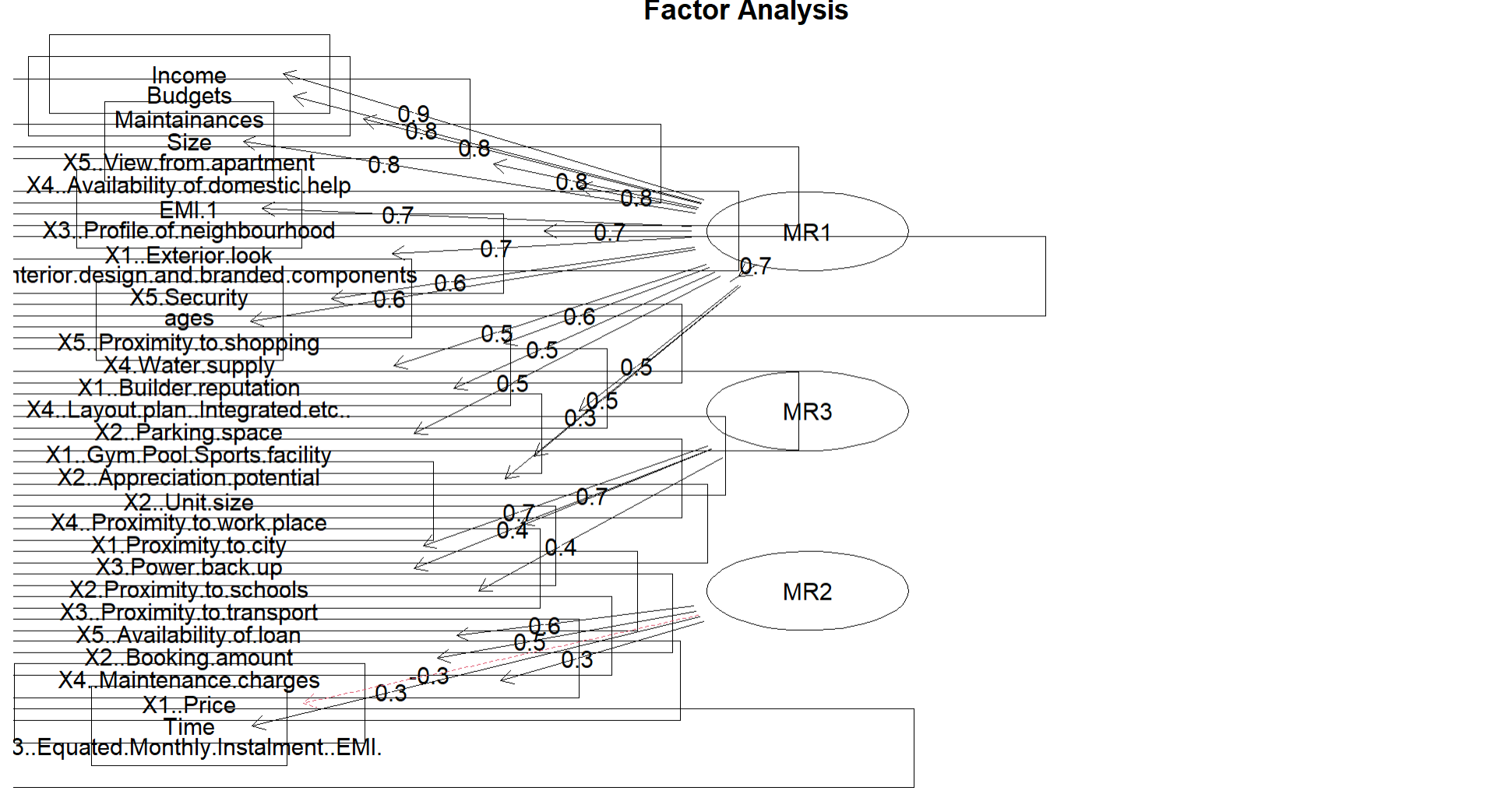
PC Simulated Data represents the average eigenvalues obtained from simulated data, which serve as a threshold for determining significant components.

PC Resampled Data sometimes included for further validation, showing the eigenvalues from bootstrapped samples of the actual data.

FA Actual Data represents the eigenvalues from Factor Analysis.

The primary criterion for retaining components is that the eigenvalues from the actual data (PC Actual Data) should be greater than the corresponding eigenvalues from the simulated data (PC Simulated Data). In this plot, it appears that only the first component (or two) has an eigenvalue above the threshold indicated by the simulated data, suggesting that one or two components should be retained.

The first principal component (PC1) has a significantly higher eigenvalue than the others, indicating that it explains a substantial portion of the variance in the dataset. Subsequent components have eigenvalues that rapidly decline and fall below the simulated data's threshold, indicating diminishing returns in terms of explained variance.



Interpretation

MR1, MR2, and MR3 represent the extracted factors from the Factor Analysis. These are latent variables that explain the patterns of correlations among the observed variables. Each factor groups together variables that are correlated, suggesting they measure similar underlying constructs.

The observed variables are the specific characteristics or questions from the survey data. They are connected to the factors by arrows that denote their loadings.

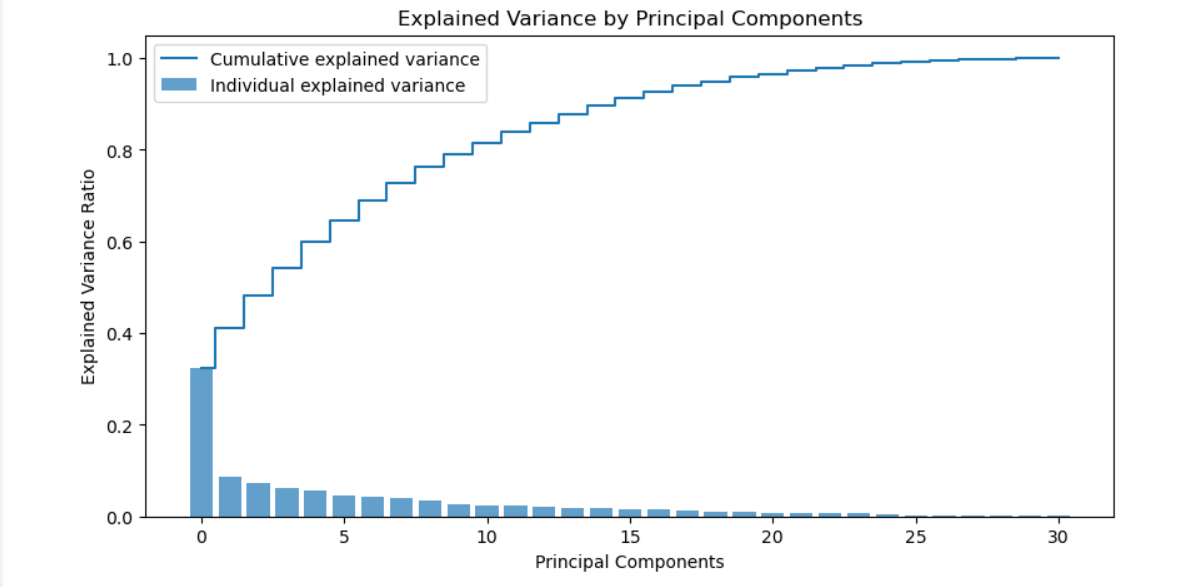
The numbers next to the arrows represent the factor loadings. These loadings indicate how strongly each observed variable is associated with a factor. A higher absolute value indicates a stronger relationship.

For example, variables like "Income," "Budgets," and "Size" have high loadings on MR1 (0.9, 0.8), suggesting that MR1 could represent a dimension related to financial capacity or budget considerations.

Similarly, "X1.Exterior.look," "X5.Security," and "X4.Availability of domestic help" load highly on MR2, which might represent concerns related to security, convenience, and aesthetic aspects of housing.

MR1 (Factor 1): This factor might represent financial considerations or economic capacity, as it includes variables like "Income," "Budgets," and "Size" with high positive loadings.

MR2 (Factor 2): This factor seems related to security, appearance, and amenities, as indicated by variables like "X1.Exterior.look," "X5.Security," and "X4.Availability of domestic help."

MR3 (Factor 3): This factor could be related to practical and locational preferences, including variables like "X1.Gym.Pool.Sports.facility," "X2.Parking space," and "X4.Proximity to work place."

Interpretation

The bars represent the variance explained by each principal component individually. The first component (PC1) explains the most variance, followed by a sharp drop for the second component (PC2), and so on. This indicates that the first few components capture the most significant patterns in the data.

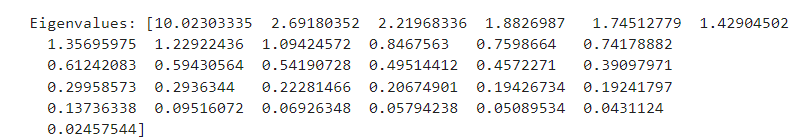
The line shows the cumulative explained variance as more principal components are added. It starts steeply with the first few components and then gradually flattens out, approaching 1.0 (or 100%) as all components are considered.

The first principal component (PC1) alone explains a substantial portion of the variance (approximately 32%, as per the previous data), highlighting its importance. The second component adds a smaller, but still significant, amount of variance explained, and so on.

This rapid initial increase in the cumulative explained variance suggests that the dataset's dimensionality can be significantly reduced while still retaining most of the important information.

One practical approach is to choose the number of components that together explain a sufficient amount of variance, often 80-90%, depending on the application. From the cumulative explained variance line, you can see that about 10 components explain approximately 90% of the variance, suggesting that these 10 components capture most of the information in the data.

The point where the cumulative explained variance line starts to flatten, known as the "elbow point," is often considered a good cutoff for dimensionality reduction. This is around 10 components in this case.



Interpretation

Each eigenvalue corresponds to a principal component and indicates the amount of variance in the dataset that component explains. Higher eigenvalues indicate components that explain more variance.

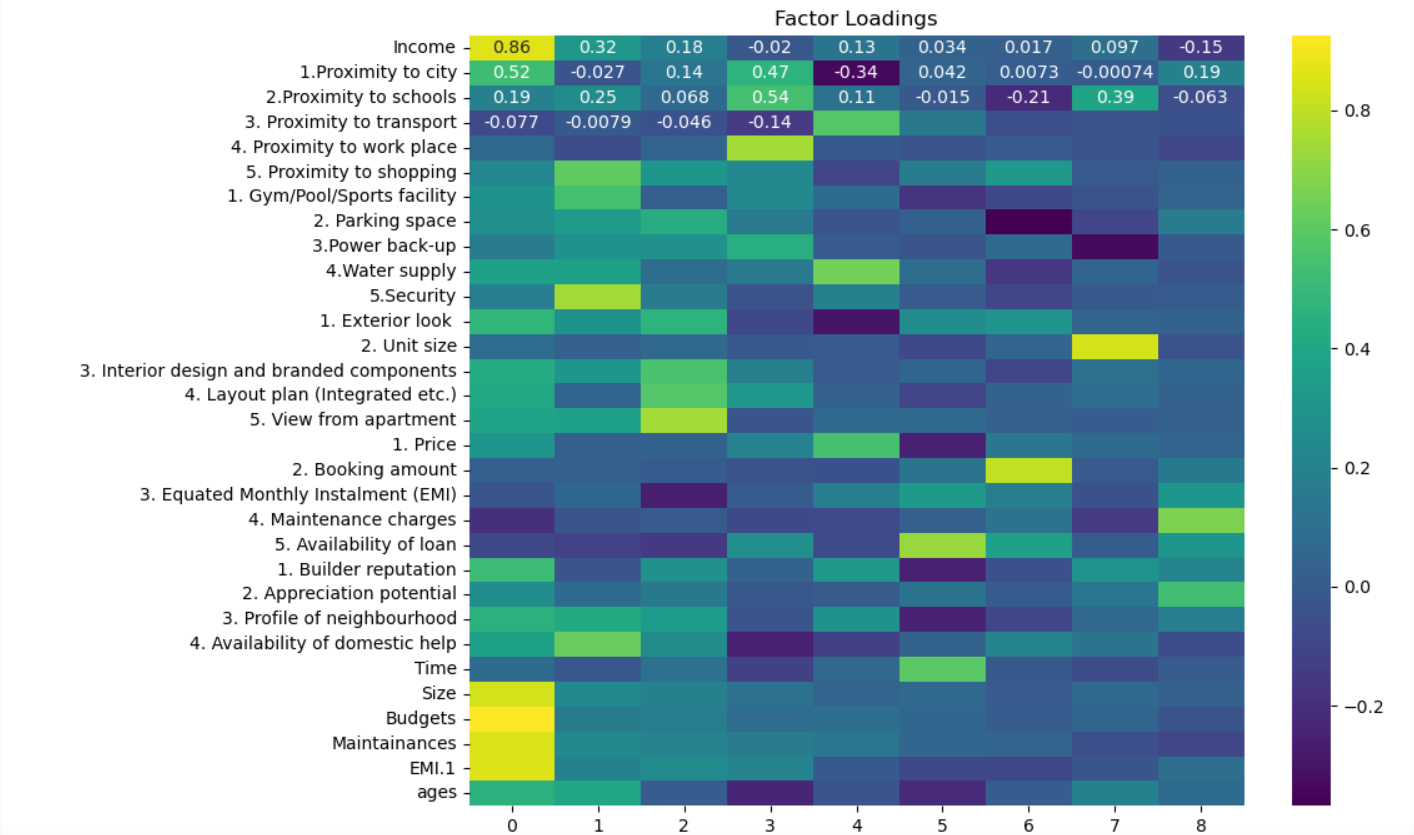
The list starts with the largest eigenvalue, 10.02303335, which corresponds to the first principal component (PC1). This component captures the most variance in the data.

Eigenvalues greater than 1 are considered significant as they suggest that the component explains more variance than a single original variable in the dataset. In this list, there are several components with eigenvalues greater than 1, which suggests these components are significant in explaining the dataset's variance.

The eigenvalues decrease as we move from PC1 to subsequent components. This is expected as each new component explains less variance than the previous ones.

The number of components with eigenvalues greater than 1 can guide the number of principal components to retain. For example, in this list, we have 10 eigenvalues greater than 1. This suggests that the first 10 components capture a significant amount of the dataset's variance.

The smaller eigenvalues towards the end of the list (e.g., 0.02457544) indicate that these components contribute very little to explaining the variance in the data and could likely be omitted without significant loss of information.



Interpretation

Factor loadings indicate how much each variable contributes to the respective factor. Higher absolute values (close to 1 or -1) suggest a stronger relationship between the variable and the factor, while values near 0 indicate a weak relationship.

Positive loadings suggest a direct relationship, while negative loadings suggest an inverse relationship.

Factor 0: This factor has a strong loading with the variable "Income" (0.86), indicating that "Income" is a key variable for this factor. Similarly, "Proximity to city" (0.52) also has a significant loading with Factor 0.

Factor 1: The variables "Proximity to work place" (0.54) and "Gym/Pool/Sports facility" (0.54) have strong loadings, suggesting these features are central to this factor.

Factor 2: Variables like "Interior design and branded components" (0.43) and "Layout plan (Integrated etc.)" (0.47) load heavily on this factor, indicating a focus on the aesthetic and design aspects of housing.

Factor 3: This factor has significant loadings for "Booking amount" (0.47) and "Price" (0.34), likely reflecting financial considerations related to housing costs.

Factor 4: The variable "Proximity to schools" (0.25) has a notable loading, suggesting this factor may relate to educational facilities.

The heatmap helps identify which variables are most associated with each factor. For instance, "Income" is predominantly associated with Factor 0, while "Proximity to work place" is significant in Factor 1.

Factors can often be interpreted based on the variables with the highest loadings. For example:

Factor 0 might be interpreted as an "Economic status" factor, heavily influenced by "Income" and "Proximity to city."

Factor 1 might represent "Convenience and Amenities," focusing on access to work and sports facilities.

Factor 2 could be "Aesthetic and Design," emphasizing interior design and layout aspects.

The colour scale ranges from purple (negative or low factor loading) to yellow (high factor loading), with green indicating moderate values. This helps visually identify the strength and direction of each variable's contribution to the factors.