

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

**A4-Principal Component Analysis, Factor Analysis**

**Azim Ziyan A V01102412**

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

# TABLE OF CONTENTS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1. Introduction ----- 3 | | |  | |
| 2. Objectives ----- 3 | | |  |  |
|  | | |  |  |
| 3. Business Significance ----- 3 | | |  |  |
|  | | |  |  |
| 4. Results and Interpretations ------ 4 | | |  |  |
|  | | |  |  |
|  | | |  |  |
|  | | |  |  |
|  | | |  |  |
|  | | |  |  |
|  | | |  |  |
|  | | |  |  |
|  | | |  |  |
|  | | |  |  |
|  |  |
|  |  |
|  | | |  |  |
|  | | |  |  |
|  | | |  |  |

1. INTRODUCTION

This assignment uses the "Survey.csv" dataset, which includes a variety of socio-demographic and housing-related parameters like age, income, number of rooms, and proximity to facilities, to investigate Principal Component Analysis (PCA) and Factor Analysis (FA). Reducing the dataset's dimensionality and locating underlying data structures are the main goals. In order to facilitate data visualization and comprehension, PCA will be used to reduce the original variables into a smaller collection of uncorrelated components that preserve the majority of the variation. Simultaneously, FA will provide latent factors that provide a deeper comprehension of the inherent patterns in the data by explaining observed correlations among variables. By adopting these methods, the study hopes to improve data interpretability, streamline complicated datasets, and offer useful insights into housing and sociodemographic choices. Through the rigorous application of PCA and FA, this analysis not only elucidates key data dimensions but also demonstrates the utility of these methods in extracting meaningful information from multifaceted datasets.

# OBJECTIVES

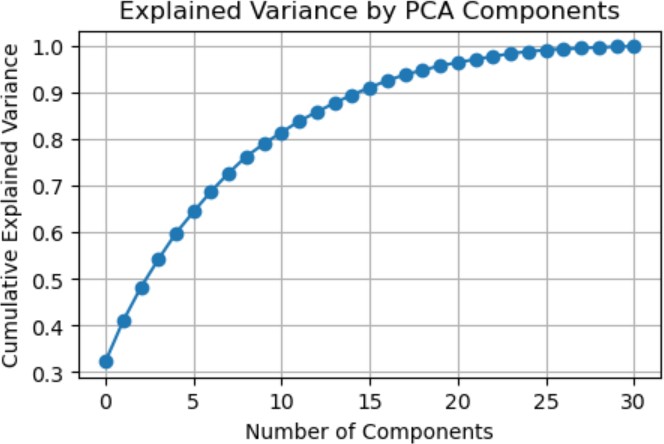
* Dimensionality Reduction
* Identify Underlying Structures
* Enhance Data Interpretability
* Visualize Data Components
* Extract Meaningful Insights

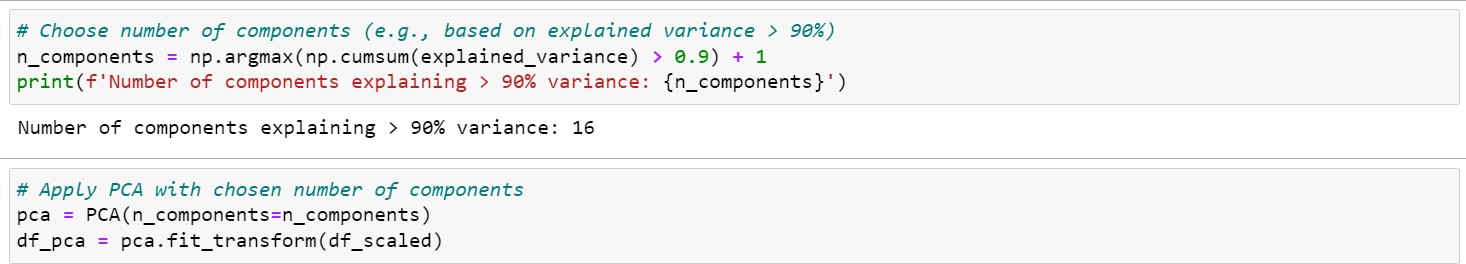
# BUSINESS SIGNIFICANCE

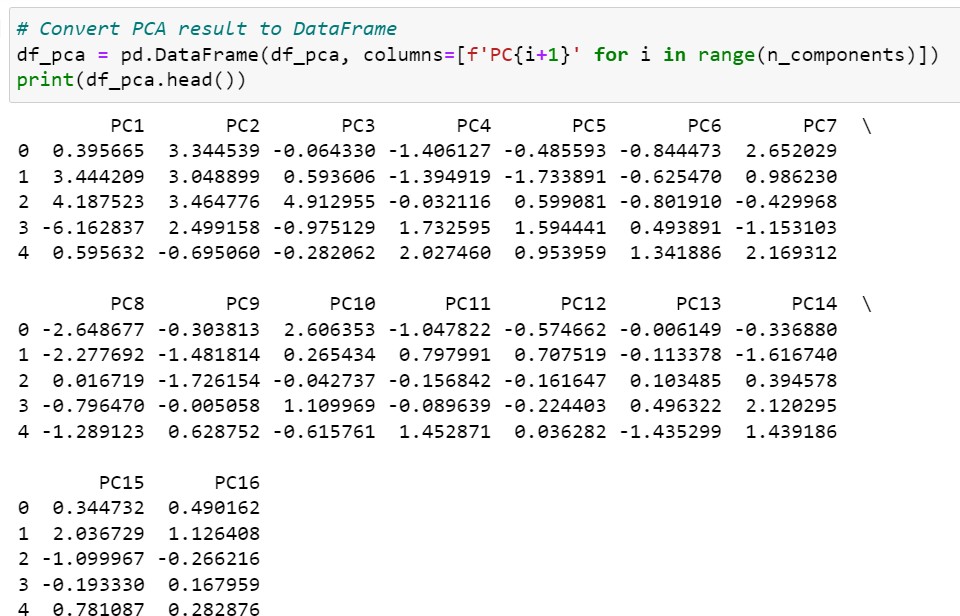
Organizations can extract valuable business insights from complicated socio-demographic and housing-related data by applying Principal Component Analysis (PCA) and Factor Analysis (FA) to the "Survey.csv" dataset. These methods simplify the data and increase its accessibility for strategic decision-making by lowering its dimensionality. Businesses may more precisely understand consumer preferences and behaviors by identifying underlying causes and primary components. This helps them develop marketing strategies that are more focused, enhance their product offers, and allocate resources more efficiently. In the end, using data-driven insights, this analytical method improves the capacity to meet market demands, raise customer happiness, and create competitive advantage.

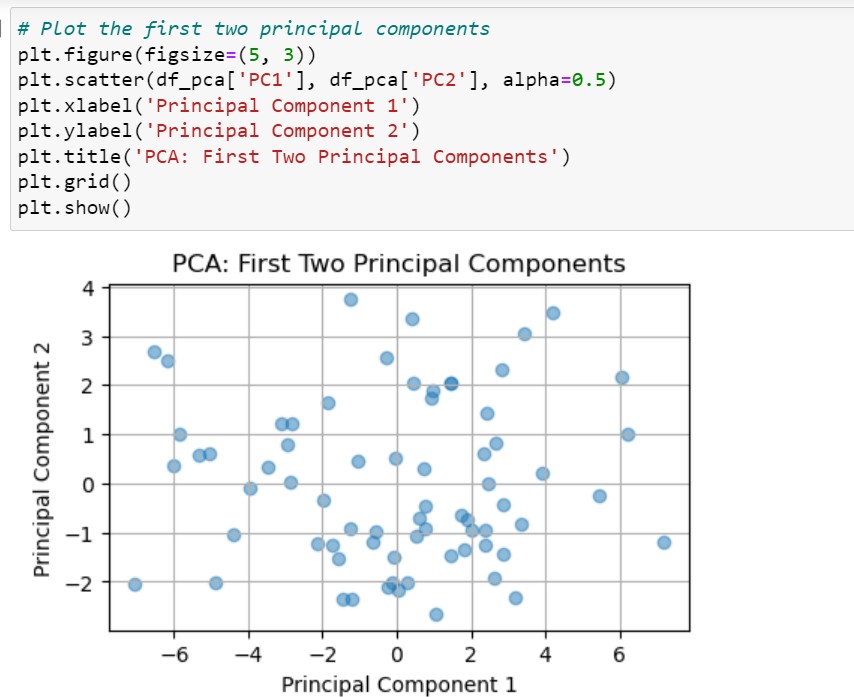
# RESULTS AND INTERPRETATIONS

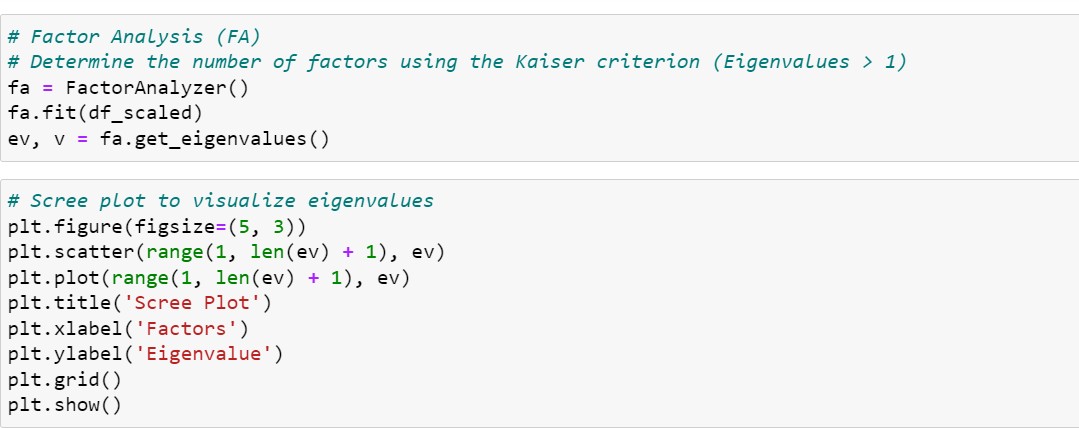
* + - Python

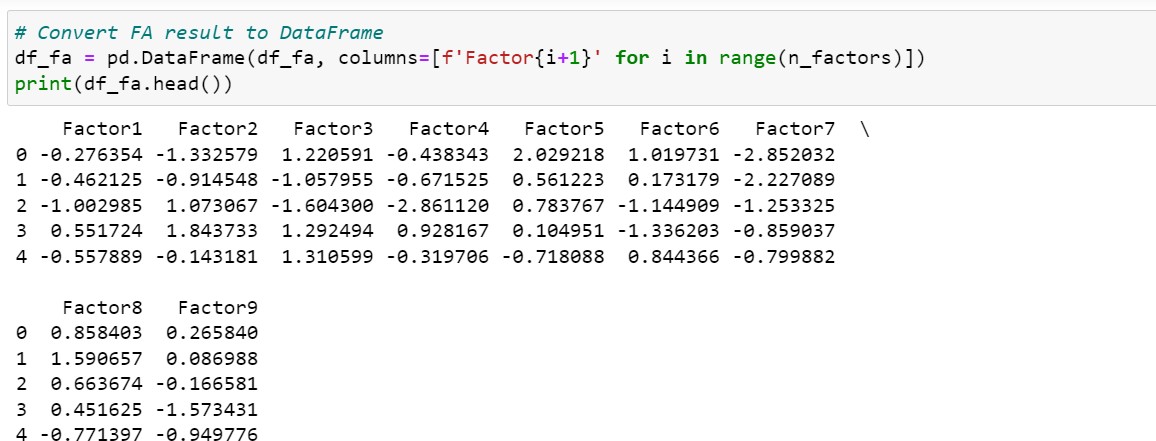
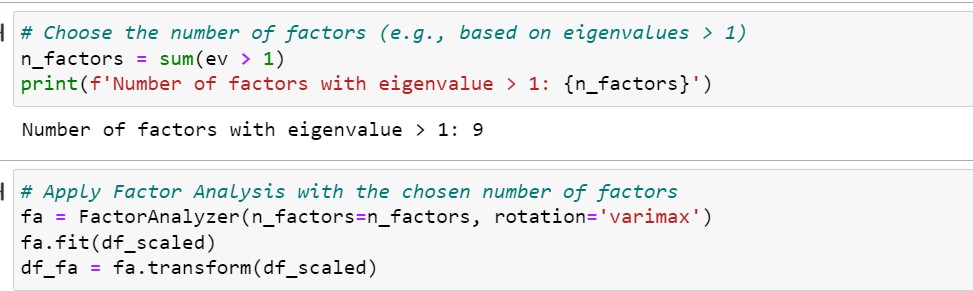
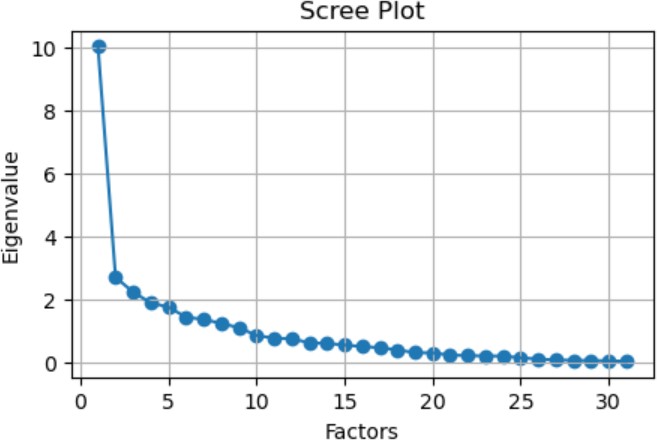


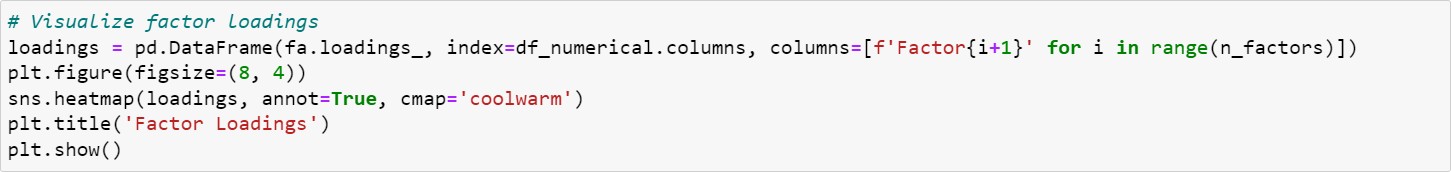


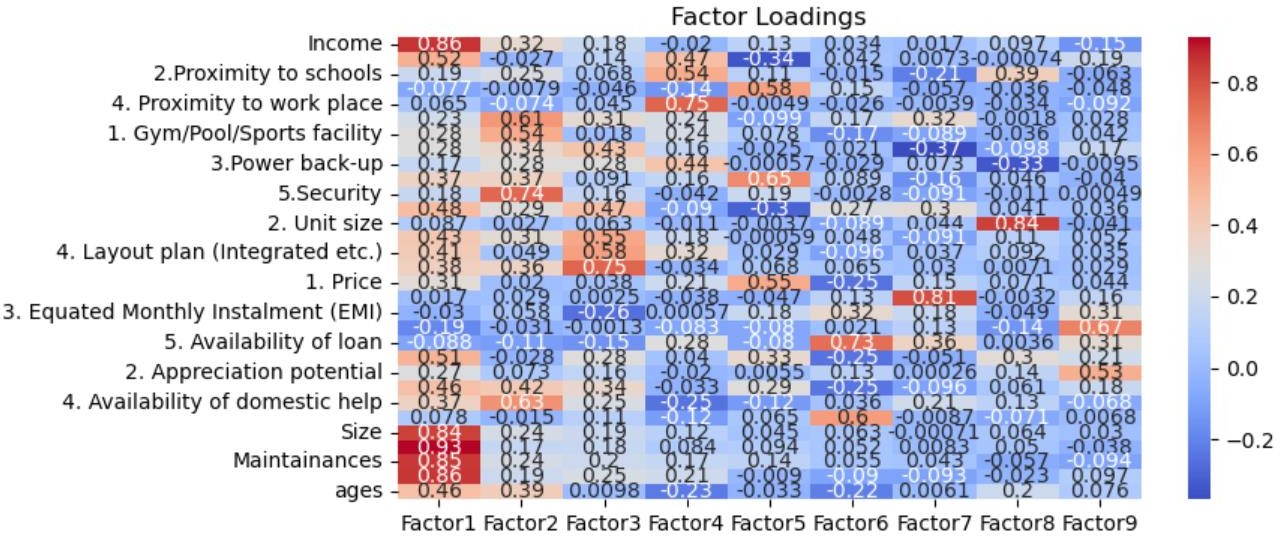




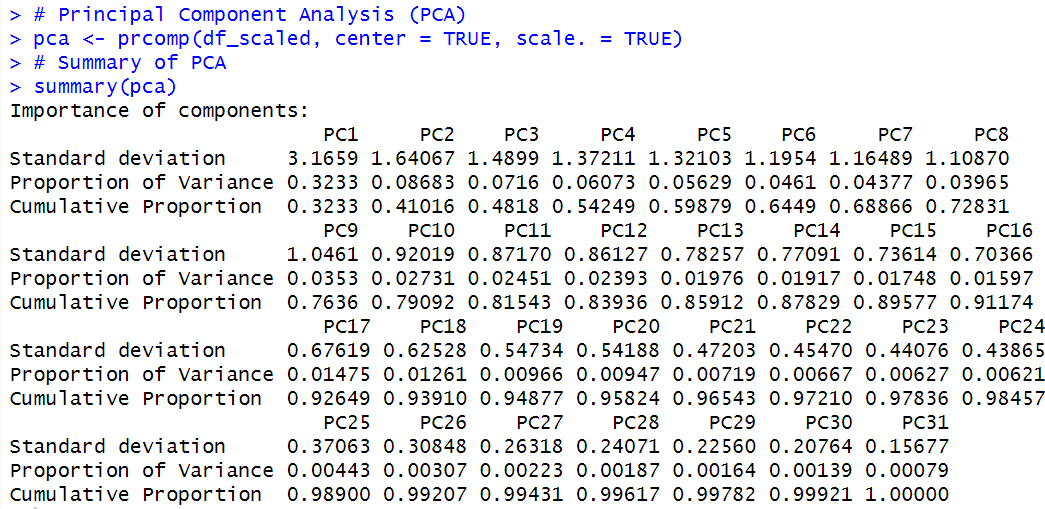


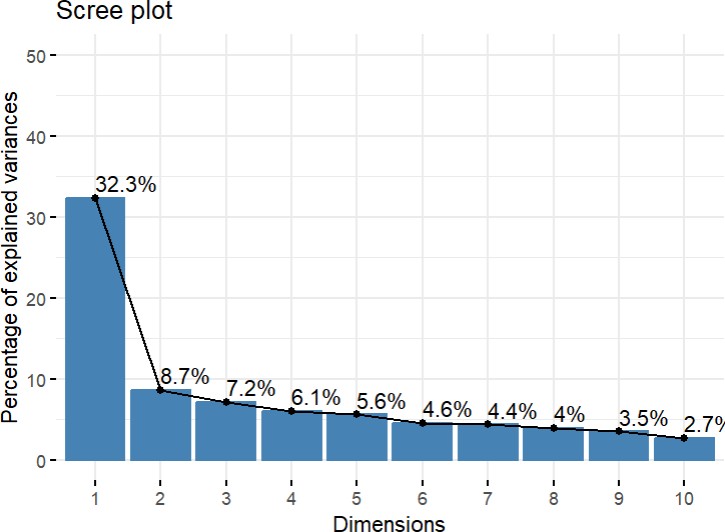


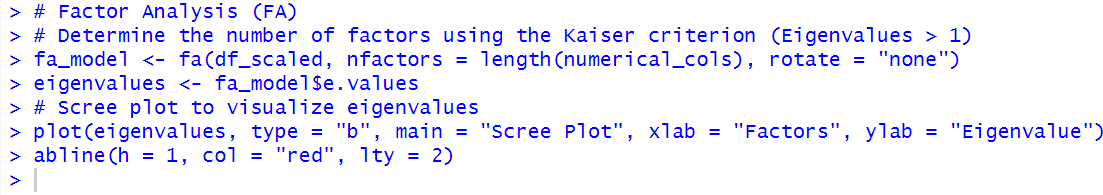
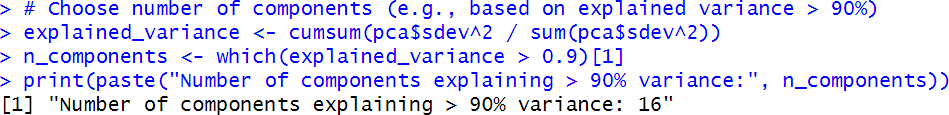


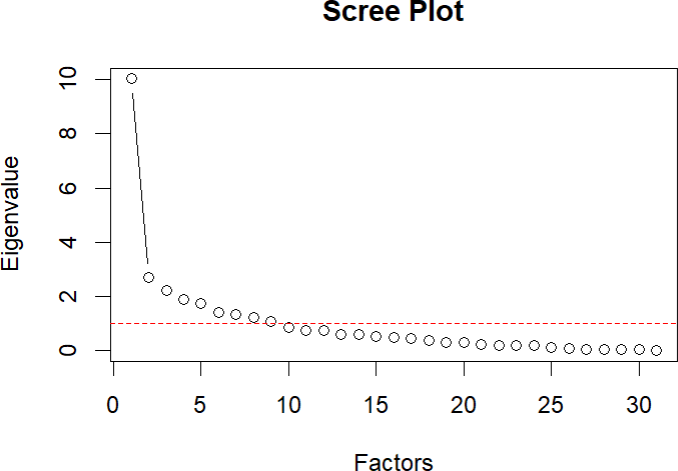


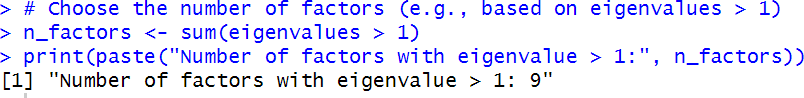
* + - R











Interpretation

Both Python and R's principle Component Analysis (PCA) results show that fewer principle components are needed to adequately summarize the dataset. The cumulative explained variance plot in Python exhibits a consistent upward trend, with roughly 16 components accounting for more than 90% of the variance. With each observation represented by its component scores, the PCA decomposition into 16 components exposes discrete patterns that allow for dimensionality reduction while preserving important variance information. In a similar vein, the PCA summary in R highlights that 16 principle components together account for over 90% of the variance, demonstrating the stability of the dimensionality reduction strategy on both platforms.

However, factor analysis (FA) in R and Python finds underlying factors that explain the observed correlations between the variables in the dataset. The Python scree plot supports the Kaiser criterion by clearly showing a decline in eigenvalues after about nine factors. Varimax rotation is used to turn the variables into nine factors, each of which captures a unique element of the original variables, improving interpretability. Similarly, in R, the eigenvalue analysis and scree plot confirm that nine components, which represent significant latent constructs that streamline the comprehension of socio-demographic and housing-related features, appropriately describe the information. Overall, these analyses in R and Python show strong methods for deciphering crucial information from huge datasets and simplifying them, providing insightful information for strategic decision-making in a variety of business