Problem Statement

PCA FH (FT): Primary census abstract for female headed households excluding institutional households (India & States/UTs - District Level), Scheduled tribes - 2011

PCA for Female Headed Household Excluding Institutional Household

The Indian Census has the reputation of being one of the best in the world. The first Census in India was conducted in the year 1872. This was conducted at different points of time in different parts of the country. In 1881 a Census was taken for the entire country simultaneously. Since then, Census has been conducted every ten years, without a break. Thus, the Census of India 2011 was the fifteenth in this unbroken series since 1872, the seventh after independence and the second census of the third millennium and twenty first century. The census has been uninterruptedly continued despite several adversities like wars, epidemics, natural calamities, political unrest, etc. The Census of India is conducted under the provisions of the Census Act 1948 and the Census Rules, 1990. The Primary Census Abstract which is important publication of 2011 Census gives basic information on Area, Total Number of Households, Total Population, Scheduled Castes, Scheduled Tribes Population, Population in the age group 0-6, Literates, Main Workers and Marginal Workers classified by the four broad industrial categories, namely, (i) Cultivators, (ii) Agricultural Laborers, (iii) Household Industry Workers, and (iv) Other Workers and also Non-Workers. The characteristics of the Total Population include Scheduled Castes, Scheduled Tribes, Institutional and Houseless Population and are presented by sex and residence (rural-urban). Census 2011 covered 35 States/Union Territories containing 640 districts which in turn contained 5,924 sub-districts, 7,935 towns and 6,40,867 villages.

The data collected has so many variables making it difficult to find useful details without using Data Science Techniques. You are tasked to perform detailed EDA and identify Optimum Principal Components that explains the most variance in data. (**Use Sklearn only)**.

Data file - PCA India Data Census.xlsx

Sample Solution

Read the data and perform basic analysis such as printing a few rows (head and tail), info, data summary, null values duplicate values, etc.

We will start analyzing the data by going thru the basic steps like:

1. Check head
2. Check info
3. Check summary
4. Check nulls
5. Check duplicates

Let us start by reading the data and extracting basic information:

Table 1: headfirst 5 rows of the dataset

State Code	Dist.Code	State	Area Name	No_HH	TOT_M	TOT_F	M_06	F_06
1	1	Jammu & Kashmir	Kupwara	7707	23388	29796	5862	6196
1	2	Jammu & Kashmir	Kadgam	6218	19585	23102	4482	3733
1	3	Jammu & Kashmir	Leh(Ladakh)	4452	6546	10964	1082	1018
1	4	Jammu & Kashmir	Kargil	1320	2784	4206	563	677
1	5	Jammu & Kashmir	Punch	11654	20591	29981	157	4587

(not all columns are shown)

Checking Info about the data:

Table 2: Dataset info

int64	59
object	2

There are **640 rows** and **61 columns** in the dataset where the 59 columns have Integer data type and 2 columns have object data type.

Checking summary:

Table 3: Dataset Summary

	count	mean	std	min	25%	50%	75%	max
State Code	640							