Detailed EDA Report for [Employees State Insurance.csv]

Introduction: This report presents a detailed analysis of a census dataset, leveraging the power of Principal Component Analysis (PCA) to uncover key insights and address the challenges posed by the dataset's complexity and high dimensionality. Column analysis is also used to gain a contrast between the forms of analysis.

Overview of Data File:

1. Class: 'pandas.core.frame.DataFrame'
2. Rows: 26 entries, 0 to 25
3. Columns: 4 entries
4. Datatypes: : int64(3), object(1)
5. Memory usage: 964.0+ bytes

Techniques Used Pre-Analysis on Dataset:

Null Handling: KNN method for numerical columns, Mode for Categorical Columns

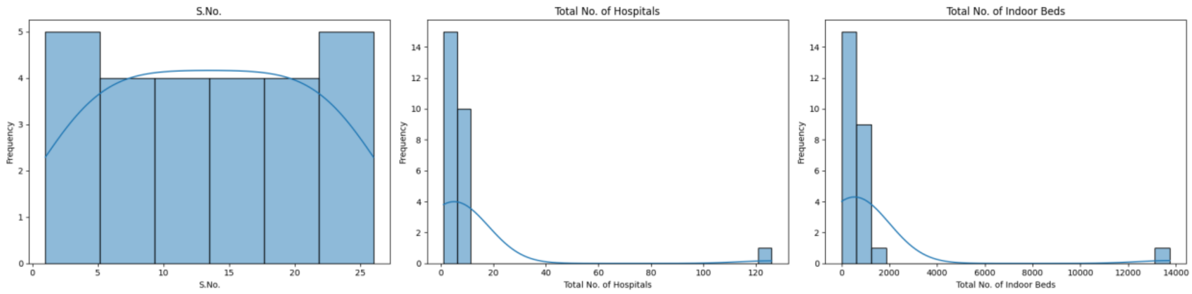
Outlier Handling: z-score method and IQR Method

Analysis: The analysis has been separated into 2 phases

Phase 1:

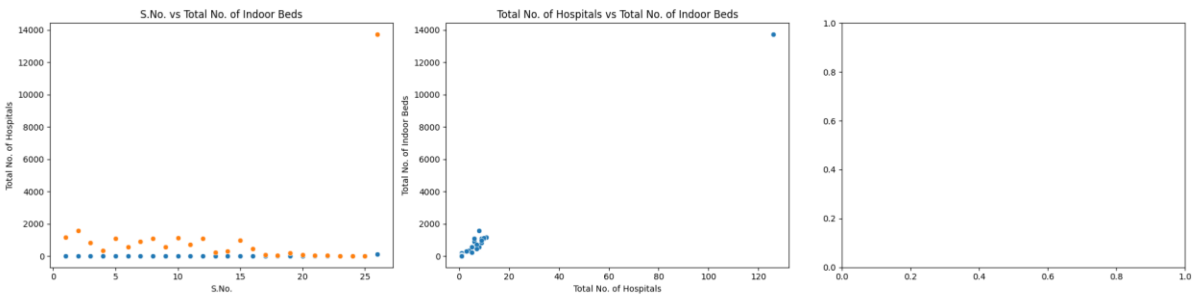
Column Based Analysis: In this phase, we analyze the dataset based on it columns, ignoring the scale of data to gain insights on the nuances of data at its structural level. Due to the scale of data, it may not help in understanding the dataset as a whole, it allows for the inspection of the basic structures of tables.

UNIVARIATE ANALYSIS: Here we plot the frequency of values in any given column, by the nature of univariate analysis, we do not involve other columns in this step. The sheer dimensionality of the data produces lots of data



General Insights: Most graphs above do not have a normal distribution. For these kinds of graph, it is noticed that the starting and the ending values for the x-axis are unusually elevated from the rest of the graph. And for graphs that have more normal a distribution, its noticed that the median value is more elevated than the surrounding values. But its noted that these points of anomalies do not affect the overall distribution of data. Distribution shapes: Most graphs show a right-skewed distribution, with the highest frequency at or near zero, and a long tail extending to the right. This is due to the outlier handling method used that clips the outliers beyond a range to the lower and upper bounds of data for non-normal distributions Most variables show a concentration of frequencies at lower values, with decreasing frequencies as the value increases, indicating potential income or resource disparities in the population studied.

Bivariate Analysis:



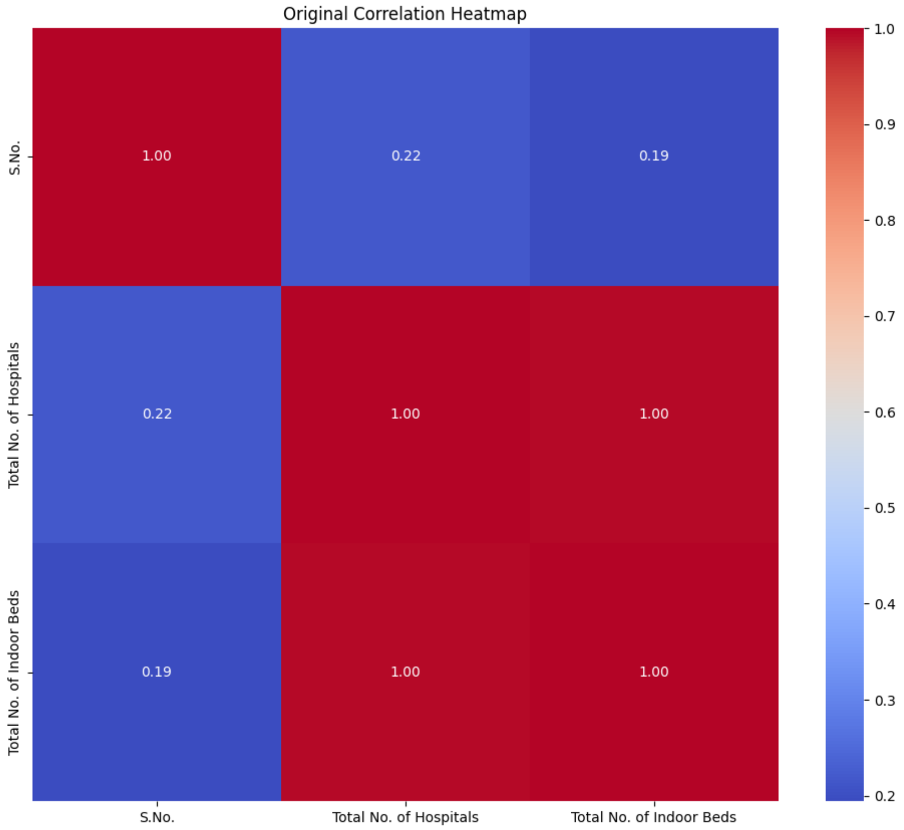
General Trends & Insights:

Data Clustering: Many plots show distinct clusters or groupings of data points. This suggests there may be underlying categorical or regional differences creating these clusters.

Correlation strengths: The degree of scatter versus alignment in the plots indicates varying strengths of relationships between variables. Some show tight correlations while others display looser associations.

Density gradients: Many plots show varying densities of data points, with higher concentractions in some areas and sparser distributons in others. This reveals where the majority of cases fall and where exceptional cases lie.

Multivariate Analysis:



High Level of correlation between total no. of hospitals and total no. of indoor beds

The correlations are symmetrical across the diagonal, as expected in a correlation matrix.

Identifying the Issues with the Phase 1 Analysis

In our initial exploration, we performed an extensive univariate and bivariate analysis on the dataset, leading to several critical observations and challenges:

1. Complexity and High Dimensionality:

o The dataset comprises a large number of numerical variables, each contributing to the overall variance in different ways.

o Visualizations such as histograms and scatter plots show complex patterns, making it challenging to interpret and identify meaningful relationships.

2. Correlation and Redundancy:

o The correlation heatmap reveals significant correlations between many variables.

o High correlation among variables suggests redundancy, where multiple variables capture similar information, leading to potential overfitting and inefficiency in analysis.

3. Non-Normal Distributions:

o Many variables exhibit skewed distributions, indicating non-normality.

o This non-normality can complicate statistical analysis and modeling, as many techniques assume normally distributed data.

Introducing PCA as a Solution

To address these challenges, we introduce Principal Component Analysis (PCA), a powerful dimensionality reduction technique that simplifies the dataset while preserving its essential information.

What is PCA?

PCA is a statistical method that transforms the original variables into a new set of uncorrelated variables called principal components. These principal components are ordered such that the first few retain most of the variation present in the original dataset.

Benefits of PCA

1. Dimensionality Reduction:

o PCA reduces the number of variables by combining them into principal components, each capturing a portion of the total variance.

o This reduction simplifies the dataset, making it easier to analyze and interpret.

2. Eliminating Redundancy:

o By transforming correlated variables into uncorrelated principal components, PCA removes redundancy.

o This results in a more efficient representation of the data, with each component providing unique information.

3. Normalizing the Data:

o The principal components often exhibit properties of normality, aiding in statistical analysis.

o This transformation aligns the data with the assumptions of many modeling techniques.

4. Enhanced Visualization:

o With fewer dimensions, visualizing the data becomes more straightforward.

o Scatter plots of the principal components reveal clear patterns and relationships that were previously obscured.

Implementing PCA on the Dataset

To demonstrate the effectiveness of PCA, we applied it to our dataset and transformed the original variables into principal components. Below, we present the results of this transformation:

1. Explained Variance:

Explained variance ratio by PCA: [0.69206257 0.30694843 0.000989]

Top Contributing Features per Principal Component:

Top Contributing Features per Principal Component:

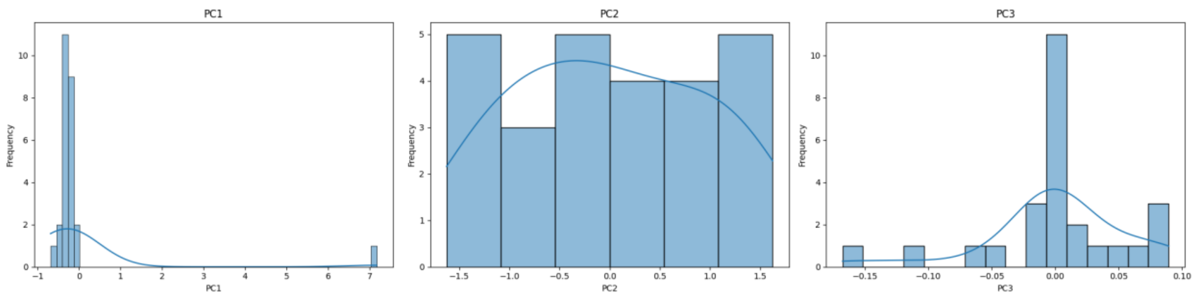
PC1: Total No. of Hospitals, Total No. of Indoor Beds, S.No.

PC2: S.No., Total No. of Indoor Beds, Total No. of Hospitals

PC3: Total No. of Hospitals, Total No. of Indoor Beds, S.No.

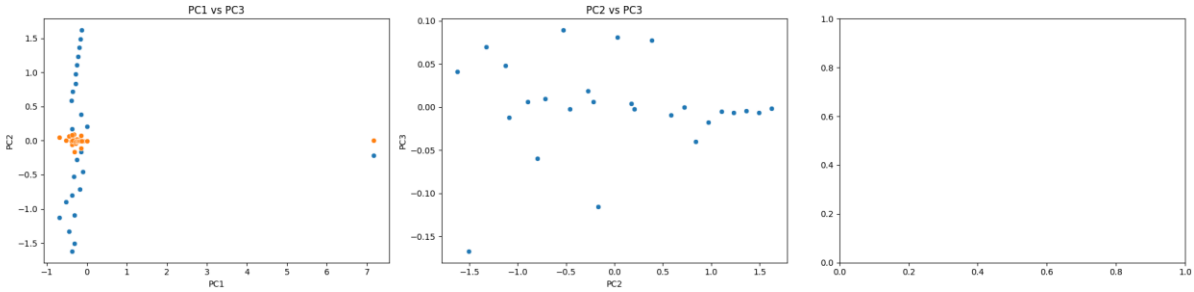
o The first principal component explains 69.2% of the variance .

Univariate PCA:



The histograms of Primary components reveal normal distributions with slight positive skew and symmetry, respectively. PC1's larger variance capture indicates its dominance in representing the dataset's structure, primarily related to the scale of healthcare facilities. PC2, though less impactful, still contributes significant variance, influenced by the sequential order of data points.

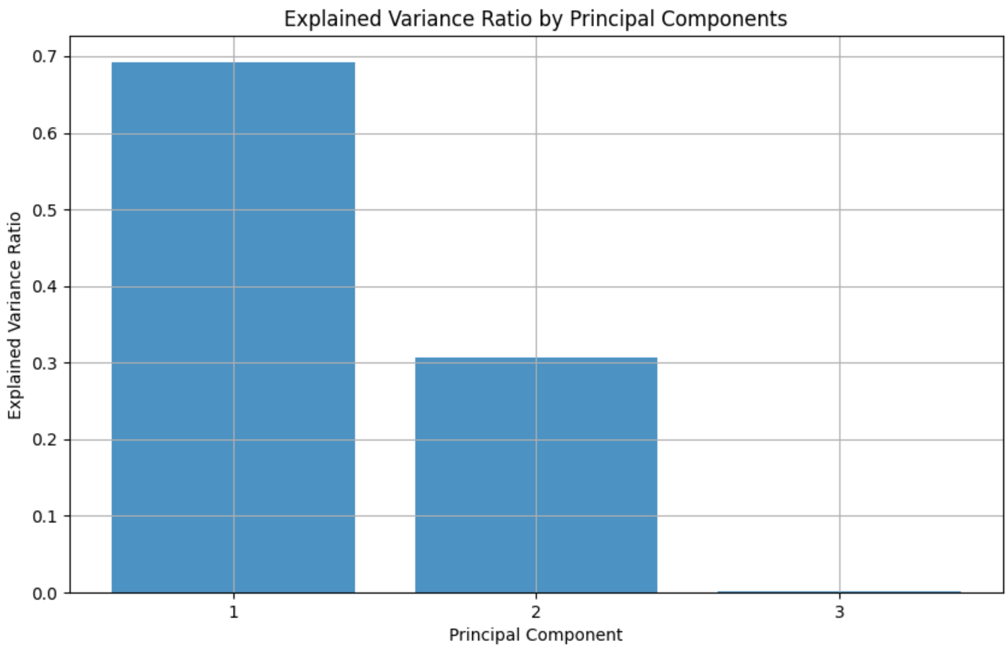
Bivariate Analysis:



The scatter plot highlights the orthogonal relationship between PC1 and PC2, showing little to no linear correlation and capturing independent aspects of the data's variance. Clustering patterns and outliers in the plot provide further structural insights into the dataset's composition and variability.

Key takeaways:

The first principal components capture 69.2% of the total variance in the dataset.



PCA Method:

The PCA method helped determine the columns that contributed most to the variance, they are: (Total No. of Hospitals, Total No. of Indoor Beds, S.No.)

Elimination of Redundancy: By transforming correlated variables into uncorrelated principal components, PCA removed redundancy, providing a more efficient representation of the data.

Normalization of Data: The principal components often exhibit properties of normality, aligning the data with the assumptions of many modeling techniques.

Column Analysis:

The column analysis helped us understand the distribution of data within a column. The key takeaway in this process is that the outlier and null handling techniques caused the elevation in mode in some columns and the tail end or the starting point in others. This is due to the nature of the Z-score and IQR methods used for outlier handling

Conclusion:

The column analysis and the PCA method has helped in gaining valuable insights into the census Data.