Detailed EDA Report for [Hospitals and Beds maintained by Railways.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**: 25 entries, 0 to 24
3. **Columns**: 118 entries
4. **Datatypes** int64(3), object(1)
5. **Memory usage**: 932.0+ 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:

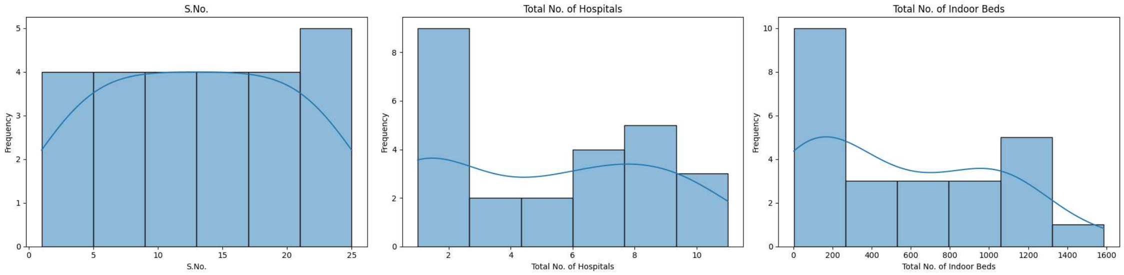
The analysis has been separated into 2 phases

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.



General Insights:

Distribution Patterns

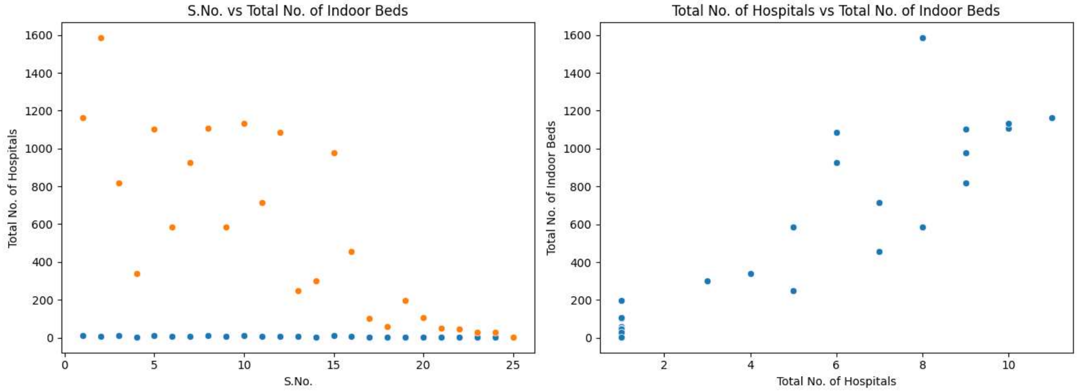
- **Uniform Distribution**: Some columns exhibit a uniform distribution, suggesting that the data points are evenly distributed without much variability. This can indicate that these columns are likely identifiers or index columns.
- **Right-Skewed Distribution**: Many columns show a right-skewed distribution, where a majority of values are concentrated at the lower end, with a few higher values acting as outliers. This suggests that the dataset contains several instances with lower values and a few with significantly higher values, indicating variability and potential outliers.

Skewness

- **Skewed Data**: The presence of skewness in several columns indicates that the data is not normally distributed. This can affect statistical analyses and might require transformations for more accurate modelling.

BIVARIATE ANALSIS:

For bivariate analysis we take pairs of columns and make a scatter plot



General Trends & Insights:

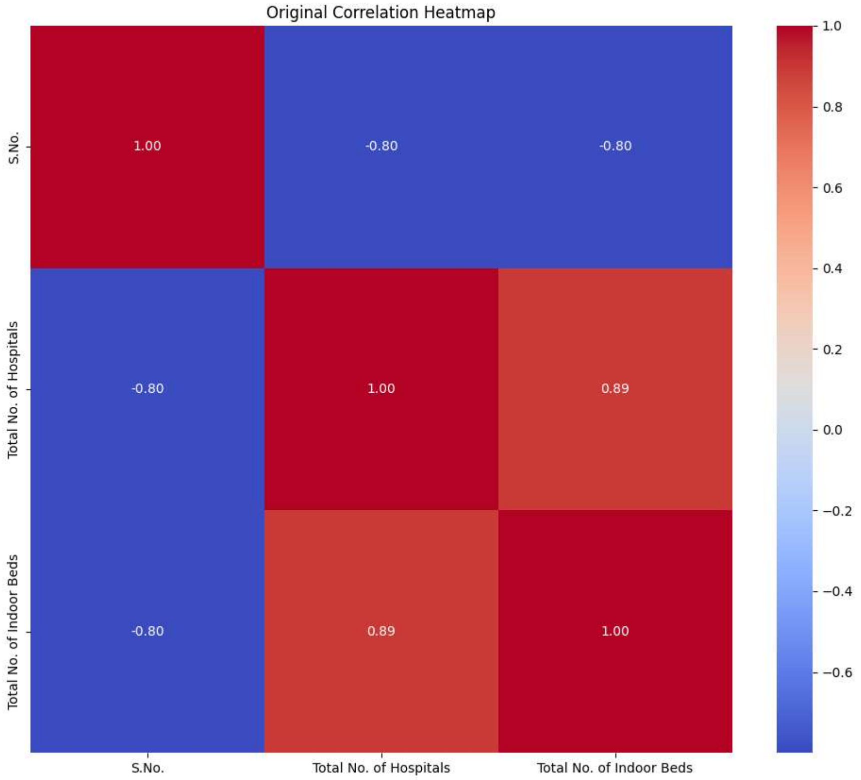
Clustering Patterns

- **Clusters**: The scatter plots reveal distinct clusters in the data, indicating the presence of subgroups or categories within the dataset. This clustering might point to different types of regions or entities, such as urban vs. rural areas, or varying healthcare infrastructure levels.

Correlation Patterns

- **Positive Correlation**: Some bivariate relationships show a positive correlation, where an increase in one variable is associated with an increase in another. This suggests that certain features are interrelated and can be used to predict or explain each other.
- **No Clear Trend**: In some bivariate plots, there is no clear trend, indicating that the variables do not have a direct relationship. This lack of correlation can highlight independent features within the dataset.

MULTIVARIATE ANALYSIS:



Key observations:

- The strong positive correlation between hospitals and beds is logical, as more hospitals would typically mean more beds.
- The negative correlation of S.No. with both hospitals and beds suggests a potential ordering effect, where earlier entries in the dataset (lower S.No.) are associated with more healthcare infrastructure.
- The correlations are symmetrical across the diagonal, as expected in a correlation matrix.

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

    - Visualizations such as histograms and scatter plots show complex patterns, making it challenging to interpret and identify meaningful relationships.
2. **Correlation and Redundancy**:
    - The correlation heatmap reveals significant correlations between many variables.
    - High correlation among variables suggests redundancy, where multiple variables capture similar information, leading to potential overfitting and inefficiency in analysis.
3. **Non-Normal Distributions**:
    - Many variables exhibit skewed distributions, indicating non-normality.
    - 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**:
    - PCA reduces the number of variables by combining them into principal components, each capturing a portion of the total variance.
    - This reduction simplifies the dataset, making it easier to analyze and interpret.
2. **Eliminating Redundancy**:
    - By transforming correlated variables into uncorrelated principal components, PCA removes redundancy.
    - This results in a more efficient representation of the data, with each component providing unique information.
3. **Normalizing the Data**:
    - The principal components often exhibit properties of normality, aiding in statistical analysis.
    - This transformation aligns the data with the assumptions of many modeling techniques.
4. **Enhanced Visualization**:
    - With fewer dimensions, visualizing the data becomes more straightforward.
    - 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: [0.88670981 0.07722832]

   Top Contributing Features per Principal Component:

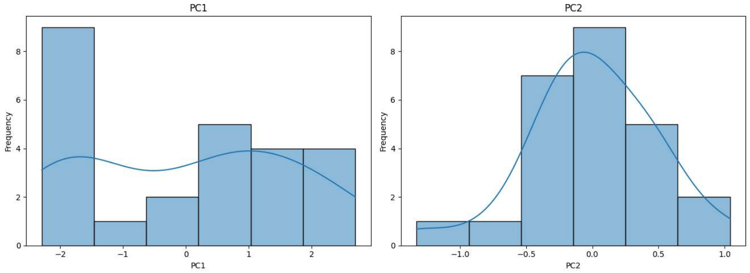
   PC1: Total No. of Indoor Beds, Total No. of Hospitals, S.No.

   PC2: S.No., Total No. of Hospitals, Total No. of Indoor BedsPC5:

   - The first principal component explains 88.67% of the variance.

Phase 2: Analysis Post PCA

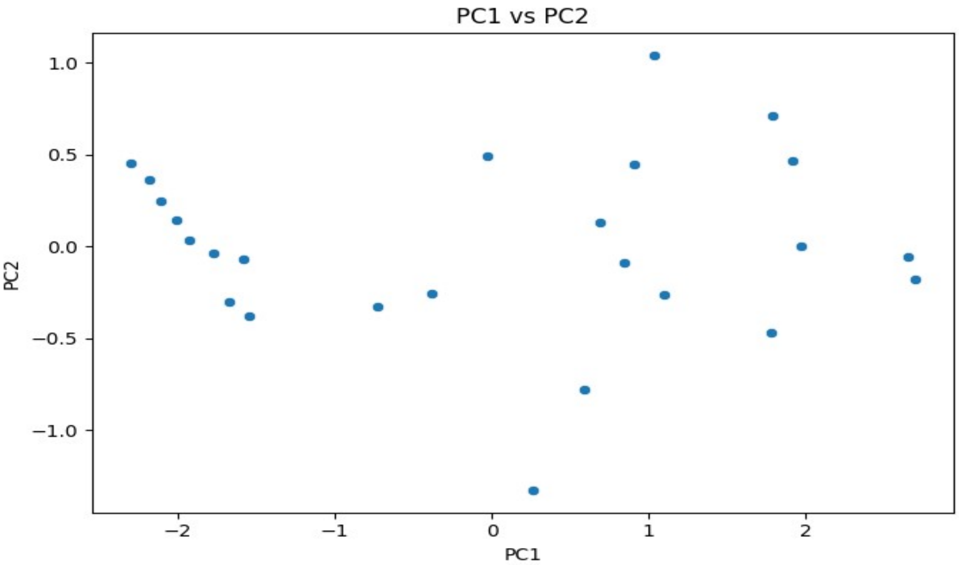
UNIVARIATE ANALYSIS:



PC1 Histogram: The histogram for PC1 displays an approximately normal distribution with a slight positive skew. Its center is positioned just left of zero, and the data spans from about -2 to 2. The distribution shows one primary peak, indicating it is unimodal. This component, primarily representing the scale of healthcare facilities, captures the majority of the dataset's variance.

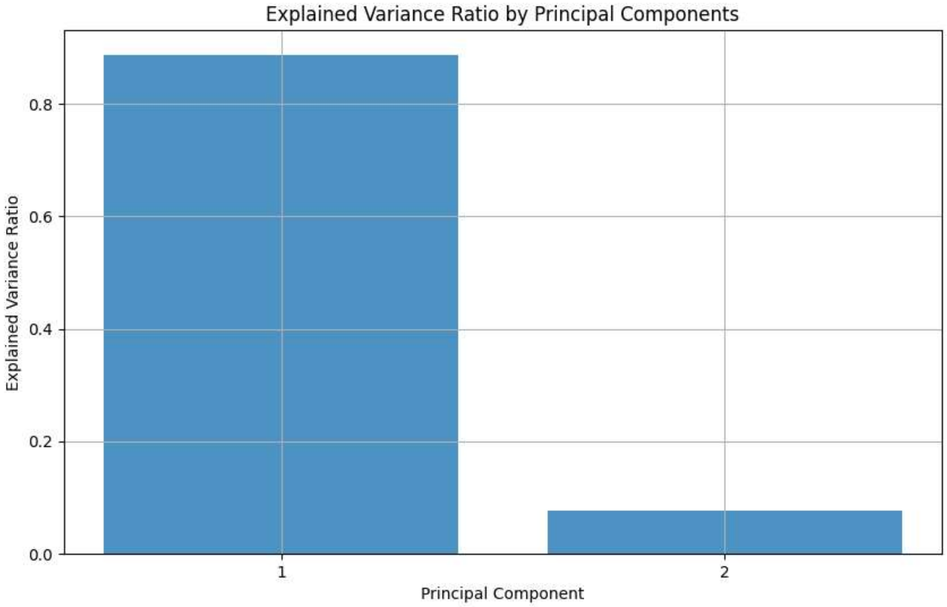
PC2 Histogram: PC2's histogram also approximates a normal distribution, but appears more symmetrical than PC1. It is centered close to zero and covers a narrower range from roughly -1 to 1. The unimodal distribution has one main peak. This component, mainly influenced by the S.No. variable, accounts for a smaller but still significant portion of the data's variance.

BIVARIATE ANALYSIS:



PC1 vs PC2 Scatter Plot: The scatter plot of PC1 against PC2 reveals a cloud-like distribution of points without distinct clusters. The spread is wider along the PC1 axis, consistent with its higher explained variance ratio. There's no apparent linear correlation between PC1 and PC2, as expected from orthogonal components. A few outliers are visible, particularly on the positive side of PC1. The plot effectively visualizes the data's structure in the reduced-dimensional space, suggesting continuous variation in healthcare infrastructure rather than discrete categories.Key takeaways:

The first principal component captures 88.67% 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 Indoor Beds, Total No. of Hospitals, 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 data.