🌟 ***Detailed EDA Report for The Given Data*** 🌟

**Welcome to the Data Adventure**:

Embark on an insightful journey through Archie's dataset. This report leverages advanced exploratory data analysis (EDA) techniques to uncover hidden patterns, address complexities, and provide actionable insights.

🔍 **Data Overview**:

1. Dataset Class: `pandas.core.frame.DataFrame`
2. Entries: `640`
3. Features: `118`
4. Data Types:
   - `float64`: 115 columns
   - `int64`: 1 column
   - `object`: 2 columns
5. Memory Footprint: `590.1+ KB`

🔧 **Pre-Processing Techniques**:

1. **Null Value Handling:**
   - Numerical Columns: KNN imputation
   - Categorical Columns: Mode imputation

2. **Outlier Handling:**
   - Methods: Z-score and IQR

## 📊 *Phase 1: Column-Based Analysis*

**1️⃣ Univariate Analysis:**

- Frequency Distribution:
  - Visualizes the spread and frequency of individual values within columns.
  - Key Insight: Most distributions are right-skewed, indicating disparities.

- Insightful Highlights:
  - Distribution Shapes:
    - Predominantly right-skewed, with higher frequencies at lower values.
    - Anomalies at the tails do not significantly affect the core distribution.

**2️⃣ Bivariate Analysis:**

- Scatter Plots:
  - Explores relationships between pairs of columns.
  - Key Patterns: Clusters, non-linear relationships, and outliers.

- Trends & Insights:
  - Data Clustering: Highlights categorical or regional differences.
  - Non-linear Relationships: Indicates complex interactions.
  - Ceiling and Floor Effects: Points cluster at maximum or minimum values.
  - Correlation Strengths: Varies from strong to weak correlations.
  - Density Gradients: Shows majority cases and outliers.

**3️⃣ Multivariate Analysis:**

- Correlation Heatmap:

- Visualizes the correlation matrix.

- Key Insight: High correlation among variables indicates redundancy.

### 🌐 ***Identifying Issues:***

1. **Complexity and High Dimensionality**:

   - Large number of variables contribute to overall variance.

   - Visualizations are complex, challenging interpretation.

2. **Correlation and Redundancy**:

   - Significant correlations among variables.

   - Multiple variables capture similar information, leading to redundancy.

3. **Non-Normal Distributions**:

   - Skewed distributions complicate statistical analysis.

4. **Visualization Overload**:

   - Numerous plots overwhelm the analysis process.

### 🧩 ***Introducing PCA (Principal Component Analysis):***

- Purpose:

  - Dimensionality Reduction: Simplifies dataset by combining variables.

  - Eliminates Redundancy: Converts correlated variables into uncorrelated components.

  - Normalizes Data: Principal components often exhibit normality.

  - Enhanced Visualization: Fewer dimensions make it easier to interpret.

PCA Implementation:

1. Explained Variance:

- Variance Ratios: `[0.569, 0.097, 0.059, 0.028, 0.024, 0.017, 0.015, 0.013, 0.012, 0.010]`
- Key Features per Component:
  - PC1: Households, Age_Group_30_49, Male_Workers
  - PC2: Drinking Water Source (Handpump/Tubewell/Borewell), Bathing Facility, Marginal Workers
  - PC3: Female_ST, ST, Male_ST
  - PC4: ST, Male_ST, Female_ST
  - PC5: Drinking Water Source (Spring/River/Canal/Tank/Pond/Lake)

2. Variance Insights:

- PC1 explains `56.92%` of the variance.
- PC1 and PC2 together explain `66.62%`.
- First 5 components capture `78.24%` of total variance.

## 🔬 ***Phase 2: Post-PCA Analysis***

### 1️⃣Univariate Analysis:

- Normal Distribution:
  - Histograms of principal components show approximate normality.
  - Key Insight: Simplified interpretation and statistical analysis.

- Dominant Structure:
  - First principal component explains over `56%` of variance.
  - Highlights the importance of demographics and household characteristics.

**2️⃣ Bivariate Analysis:**

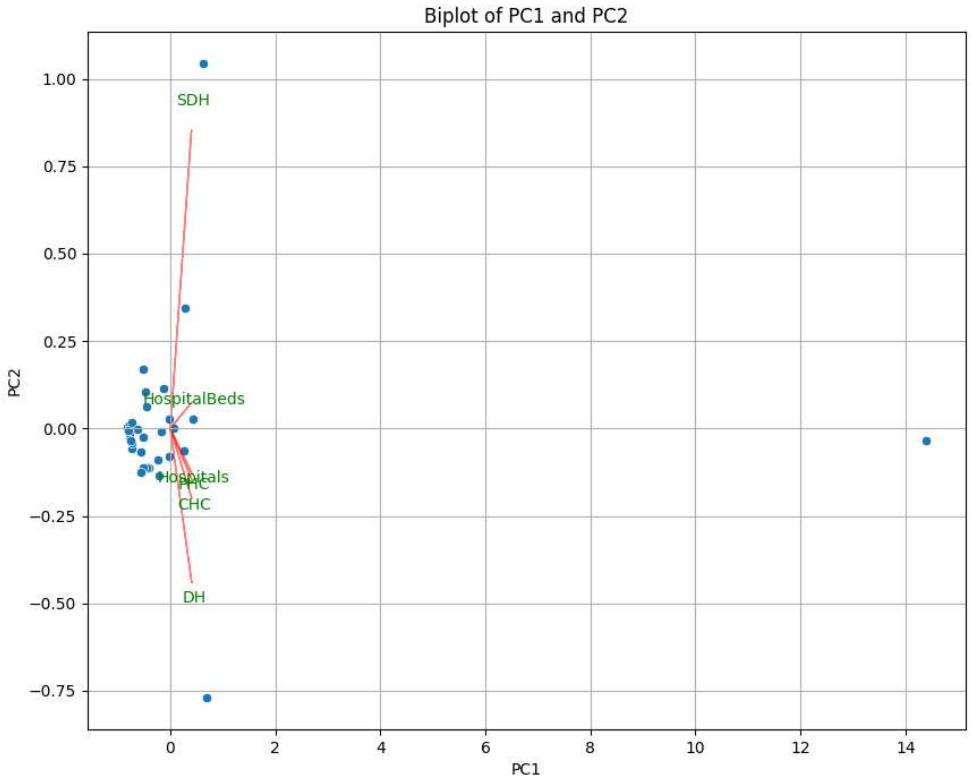
- Oval-shaped Scatter Plots:
  - Densely packed points indicate concentration around the mean.
  - Key Insight: Significant relationships captured by principal components.
  - Residual Analysis: Good model fit, minimal outliers, unbiased residuals.

## Key Takeaways:

- First 5 components capture `78.24%` of variance.
- PCA identifies key contributing columns: Households, Age_Group_30_49, Male_Workers, etc.
- PCA removes redundancy and normalizes data for better modeling.

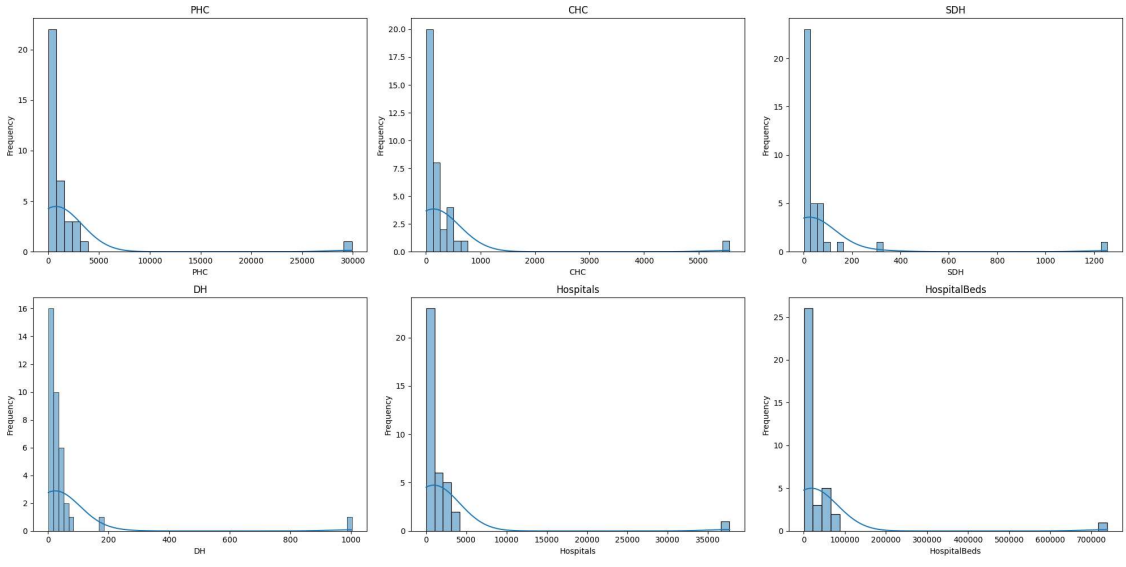
## 📈 ***Column Analysis:***

- Revealed distribution patterns and effects of outlier and null handling.
- Noted elevation in mode and tail values due to Z-score and IQR methods.

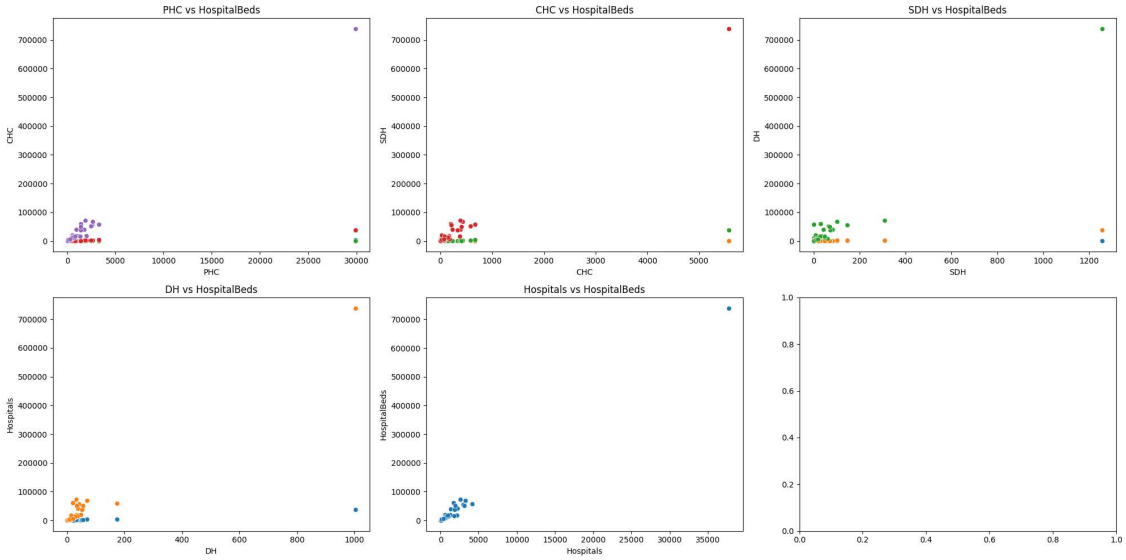



Before pca

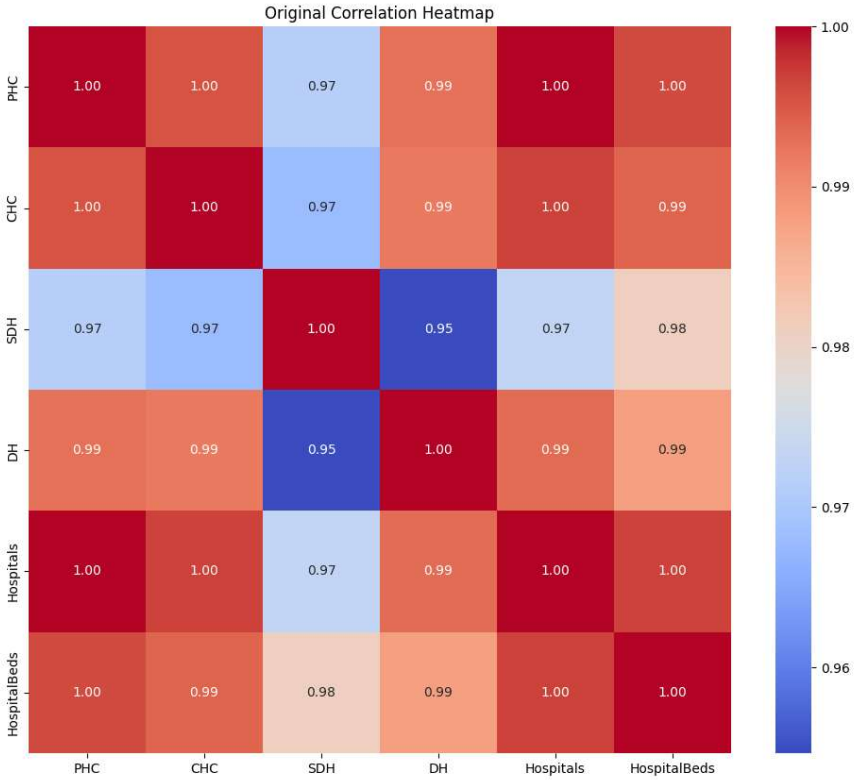
Uni:



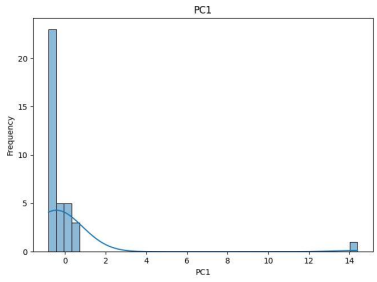

Bi:

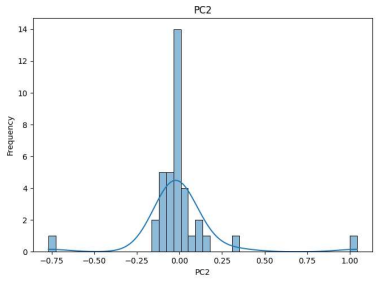


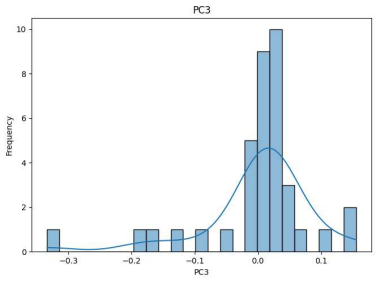

Mul:

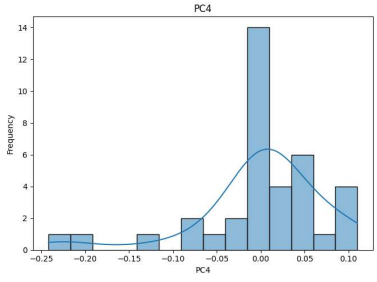


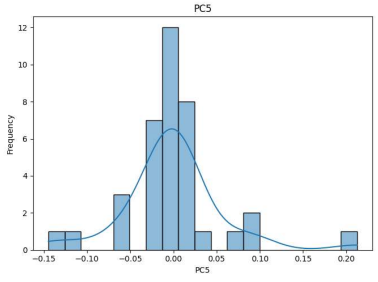

After pca

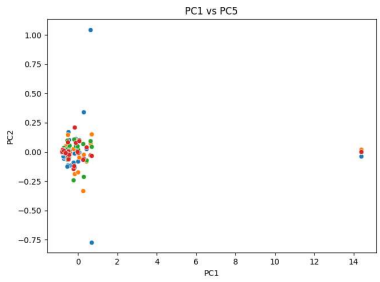

PC1
Frequency
PC1

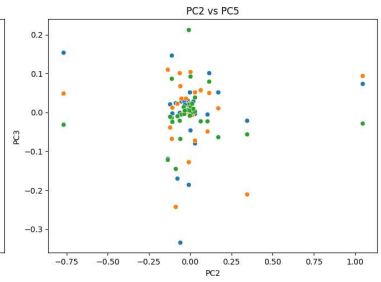

PC2
Frequency
PC2

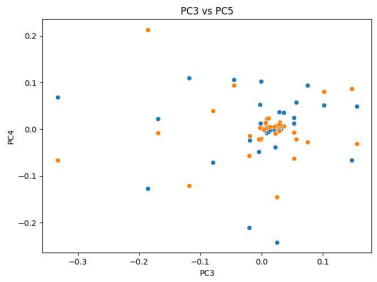

PC3
Frequency
PC3

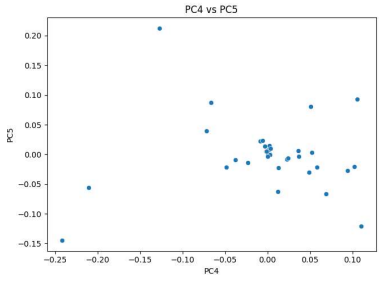

PC4
Frequency
PC4

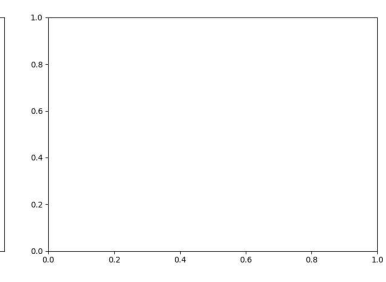

PC5
Frequency
PC5

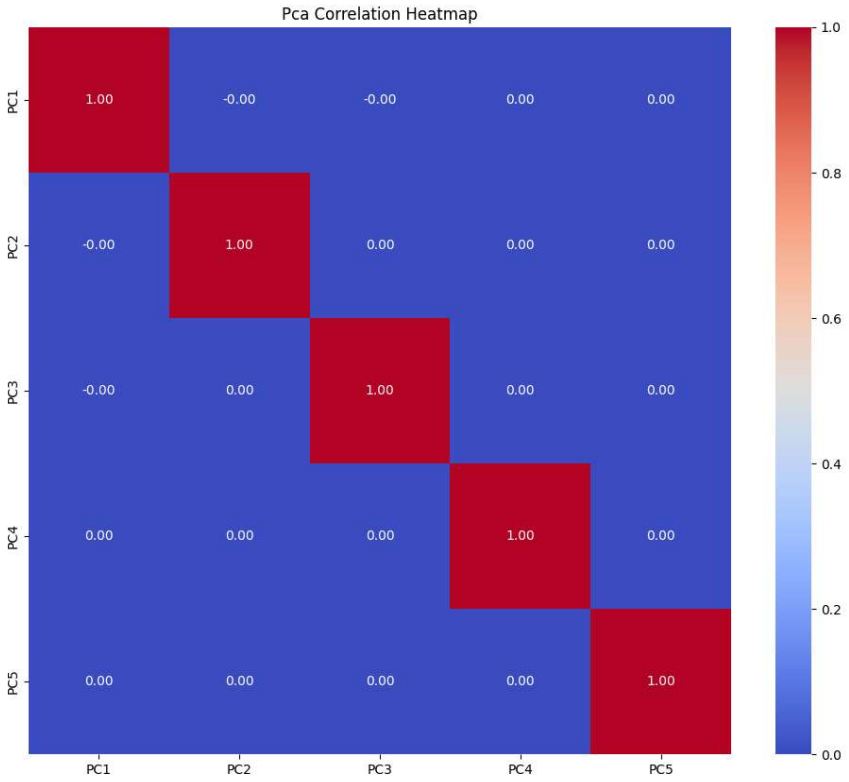

PC1 vs PC5
PC2
PC1

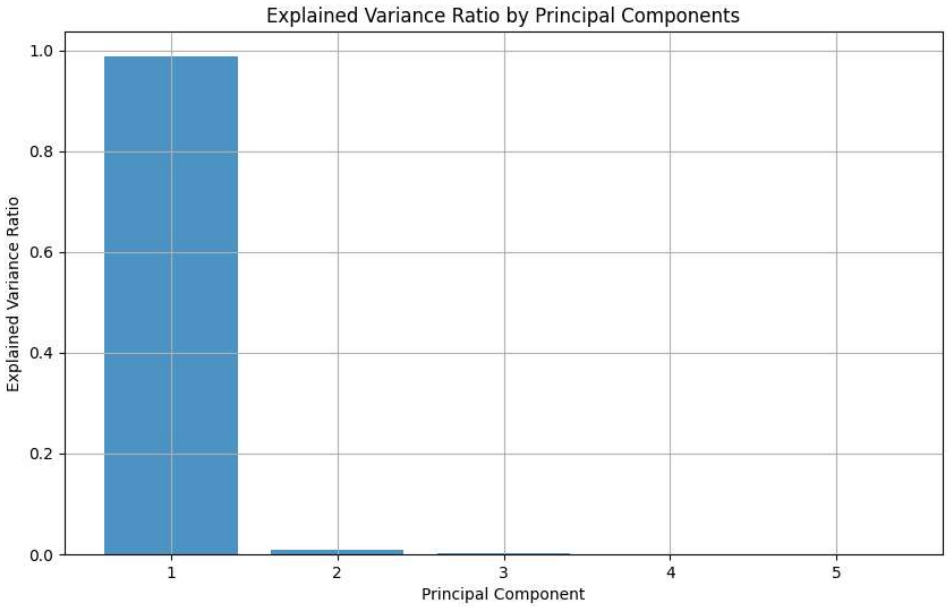

PC2 vs PC5
PC3
PC2


PC3 vs PC5
PC4
PC3


PC4 vs PC5
PC5
PC4




Pca Correlation Heatmap
PC1
PC2
PC3
PC4
PC5
1.00
-0.00
-0.00
0.00
0.00
-0.00
1.00
0.00
0.00
0.00
-0.00
0.00
1.00
0.00
0.00
0.00
0.00
0.00
1.00
0.00
0.00
0.00
0.00
0.00
1.00
1.0
0.8
0.6
0.4
0.2
0.0



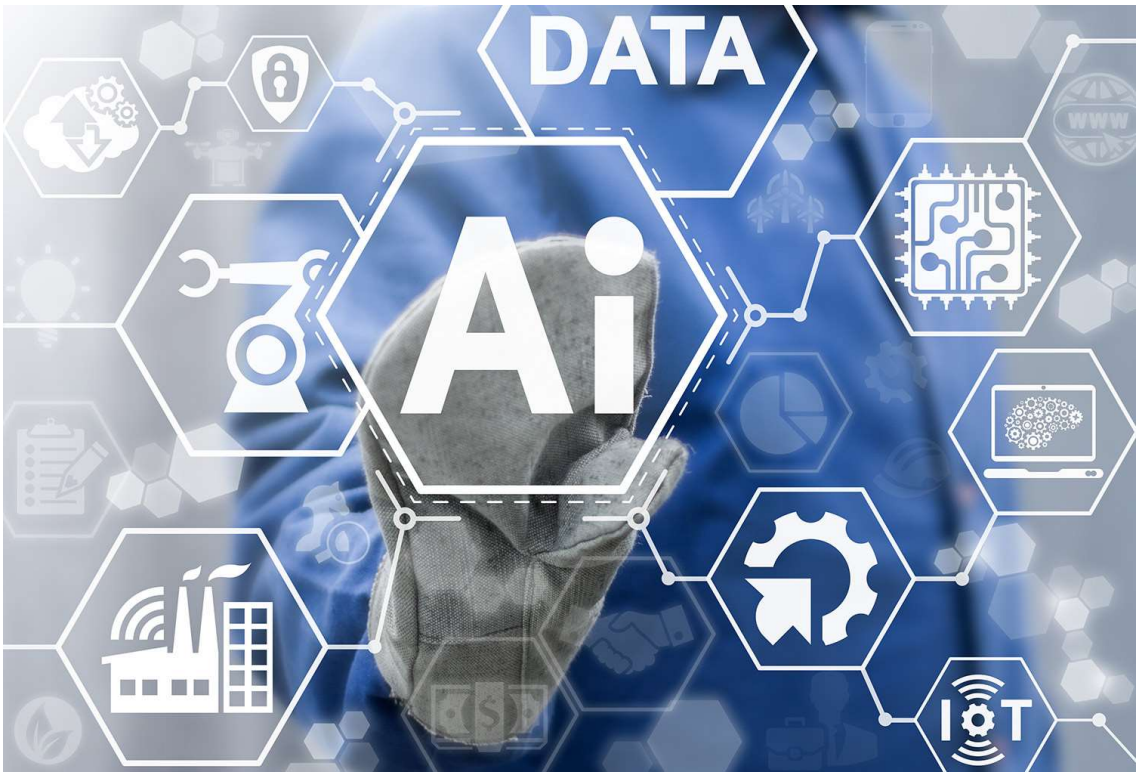

## 🏁 ***Conclusion:***

- Column analysis and PCA provide valuable insights into Archie's dataset.

- PCA effectively addresses complexity, enhances visualization, and simplifies analysis.

- This approach ensures a comprehensive understanding and actionable insights from the data.

***Thank You for Exploring with Us!***




*“Data are just summaries of thousands of stories—tell a few of those stories to help make the data meaningful.” — Dan Heath*