Detailed EDA Report for [census_2011.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**: 640 entries, 0 to 639
3. **Columns**: 118 entries
4. **Datatypes**: float64(115), int64(1), object(2)
5. **Memory usage**: 590.1+ KB

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:

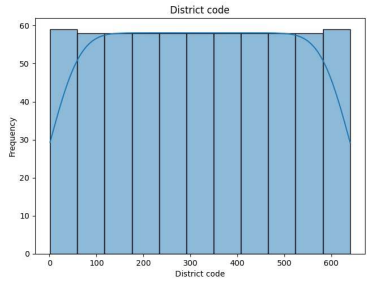
The analysis has been separated into 2 phases

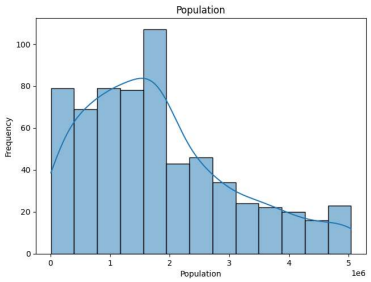
Phase 1: Column Based Analysis:

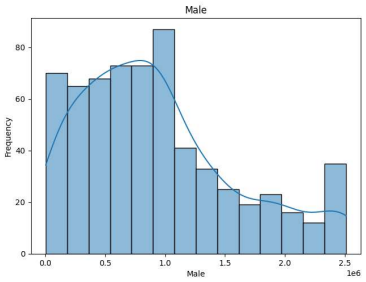
In this phase, we analyze the dataset based on it columns, ignoring the scale of data to gain insights on the nuances of data at its structural level. Due to the scale of data, it may not help in understanding the dataset as a whole, it allows for the inspection of the basic structures of tables.

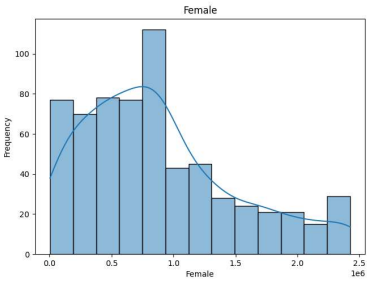
UNIVARIATE ANALYSIS:

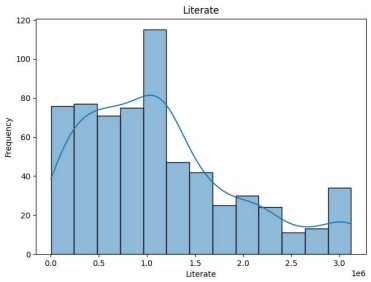
Here we plot the frequency of values in any given column, by the nature of univariate analysis, we do not involve other columns in this step. The sheer dimensionality of the data produces lots of data

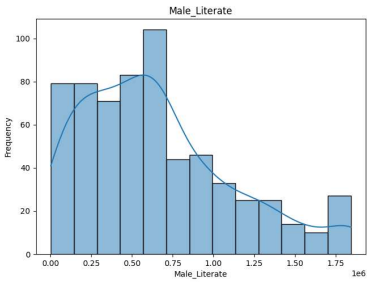

District code
Frequency
District code

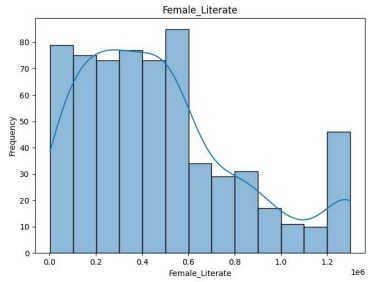

Population
Frequency
Population
1e6

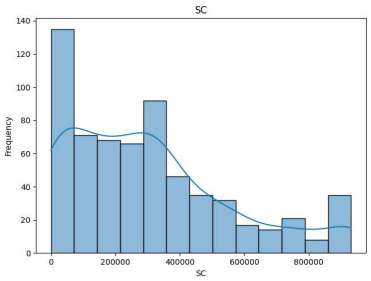

Male
Frequency
Male
1e6

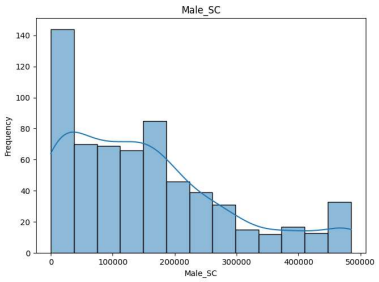

Female
Frequency
Female
1e6

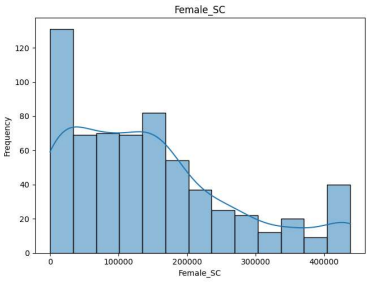

Literate
Frequency
Literate
1e6

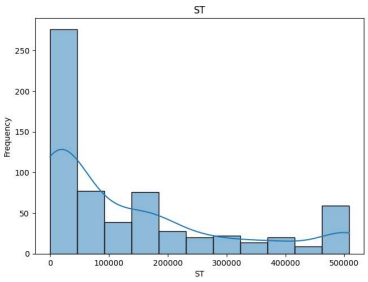

Male_Literate
Frequency
Male_Literate
1e6

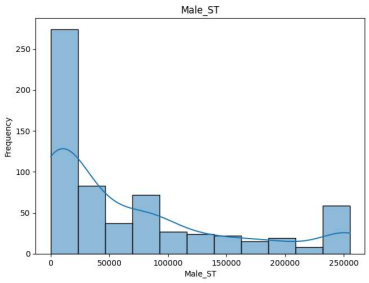

Female_Literate
Frequency
Female_Literate
1e6

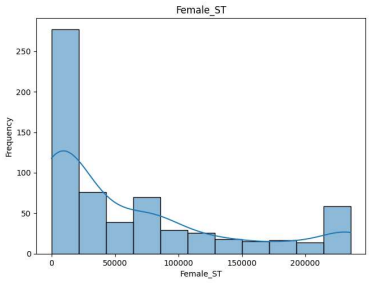

SC
Frequency
SC

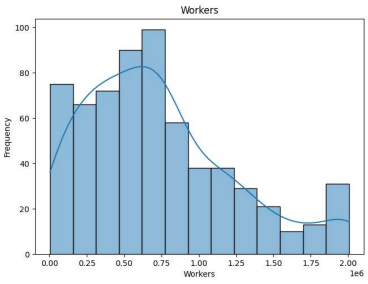

Male_SC
Frequency
Male_SC

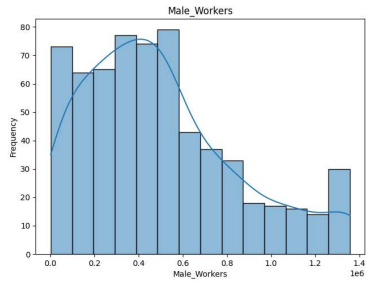

Female_SC
Frequency
Female_SC

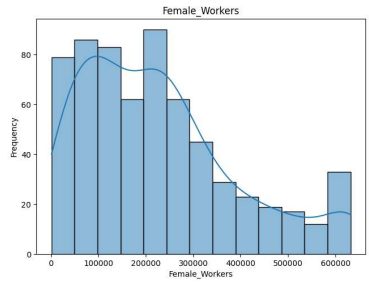

ST
Frequency
ST

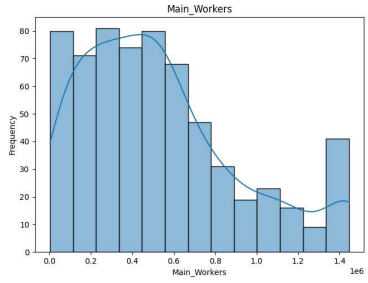

Male_ST
Frequency
Male_ST

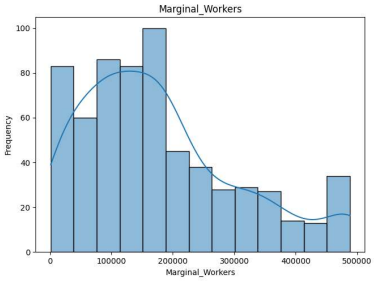

Female_ST
Frequency
Female_ST

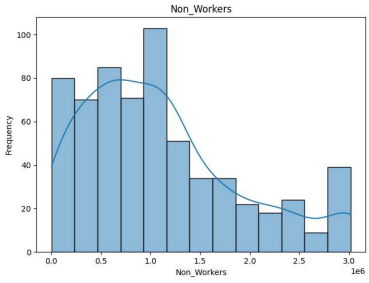

Workers
Frequency
Workers
1e6

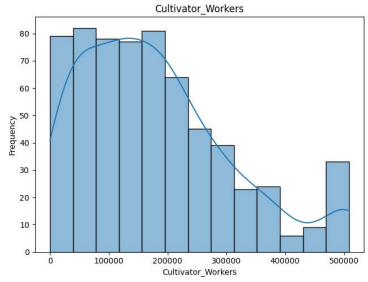

Male_Workers
Frequency
Male_Workers
1e6

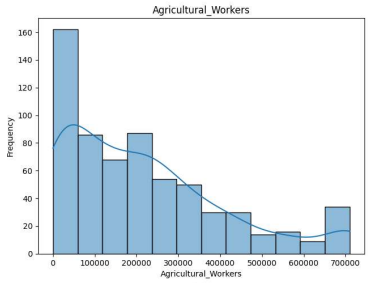

Female_Workers
Frequency
Female_Workers

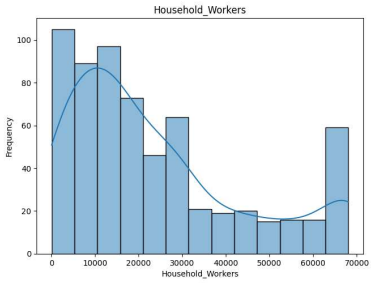

Main_Workers
Frequency
Main_Workers
1e6

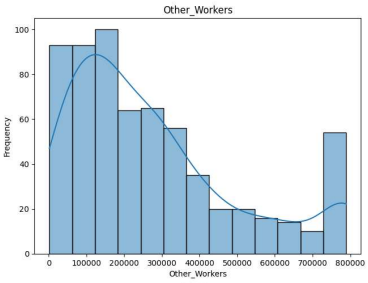

Marginal_Workers
Frequency
Marginal_Workers

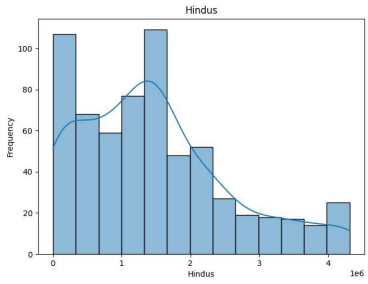

Non_Workers
Frequency
Non_Workers
1e6

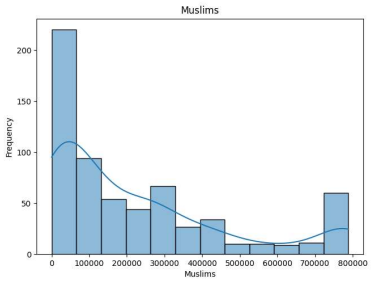

Cultivator_Workers
Frequency
Cultivator_Workers

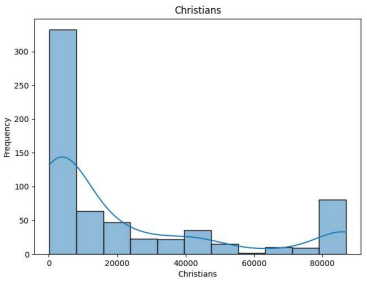

Agricultural_Workers
Frequency
Agricultural_Workers

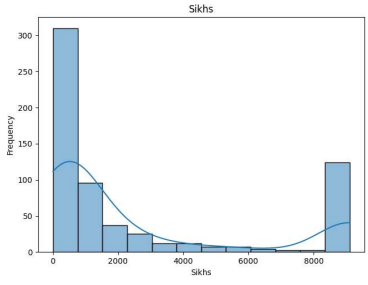

Household_Workers
Frequency
Household_Workers

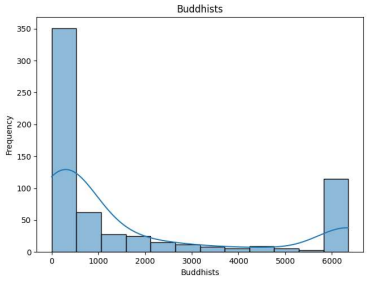

Other_Workers
Frequency
Other_Workers

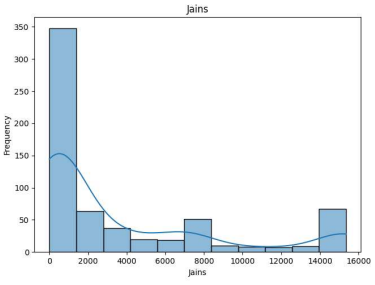

Hindus
Frequency
Hindus
1e6

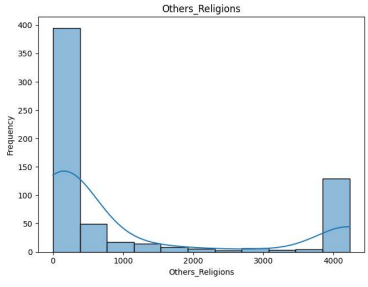

Muslims
Frequency
Muslims

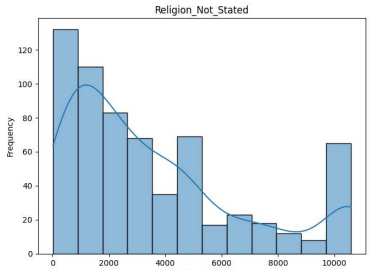

Christians
Frequency
Christians

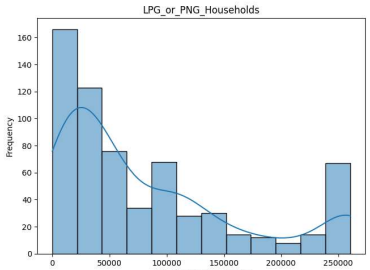

Sikhs
Frequency
Sikhs

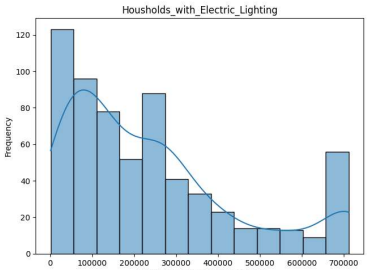

Buddhists
Frequency
Buddhists

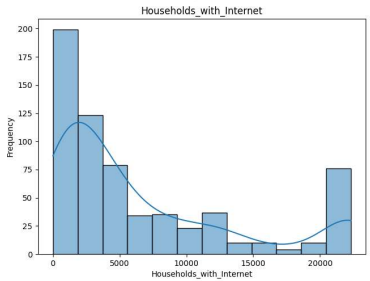

Jains
Frequency
Jains

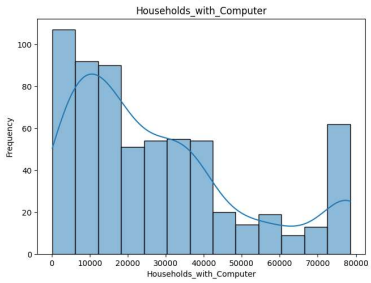

Others_Religions
Frequency
Others_Religions

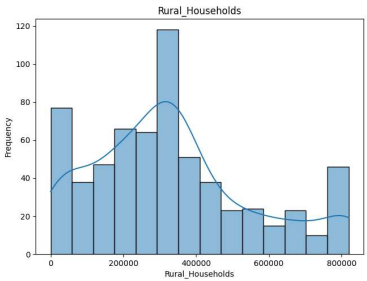

Religion_Not_Stated
Frequency
Religion_Not_Stated

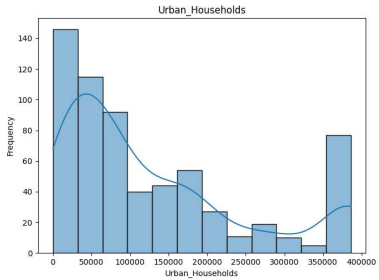

LPG_or_PNG_Households
Frequency
LPG_or_PNG_Households

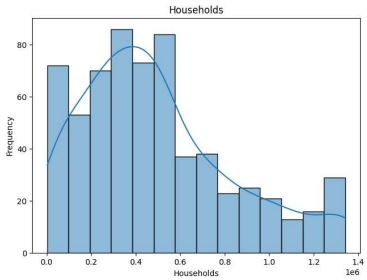

Housholds_with_Electric_Lighting
Frequency
Housholds_with_Electric_Lighting

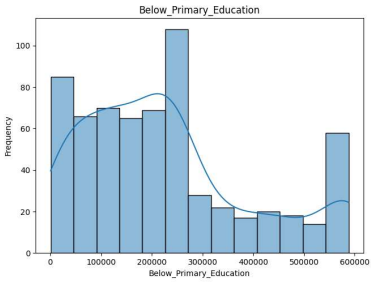

Households_with_Internet
Frequency
Households_with_Internet

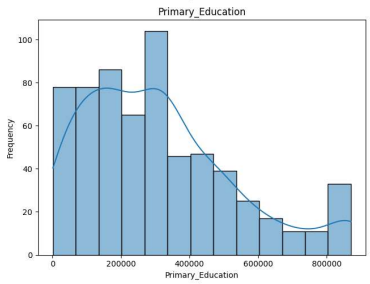

Households_with_Computer
Frequency
Households_with_Computer

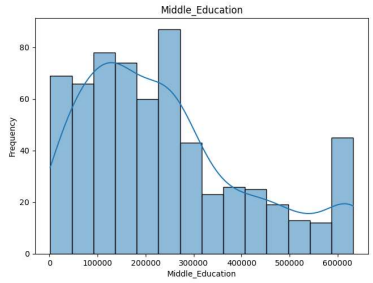

Rural_Households
Frequency
Rural_Households

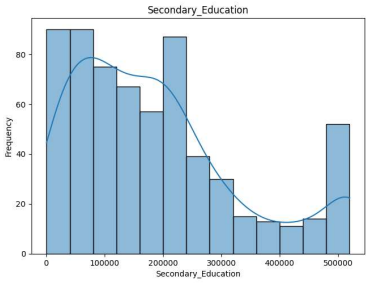

Urban_Households
Frequency
Urban_Households

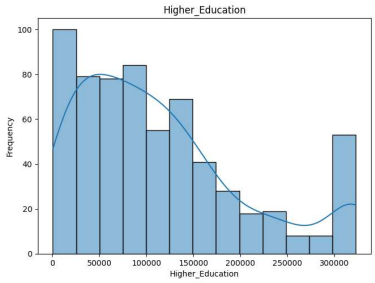

Households
Frequency
Households
1e6

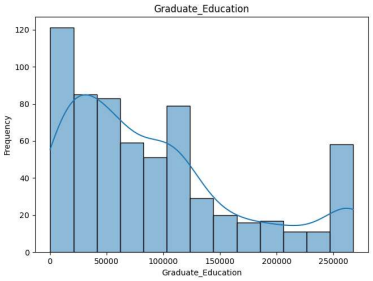

Below_Primary_Education
Frequency
Below_Primary_Education

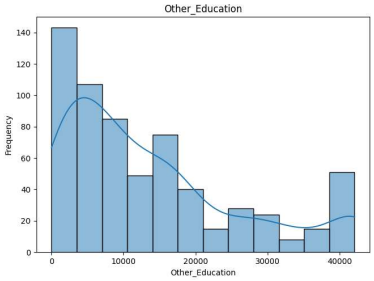

Primary_Education
Frequency
Primary_Education

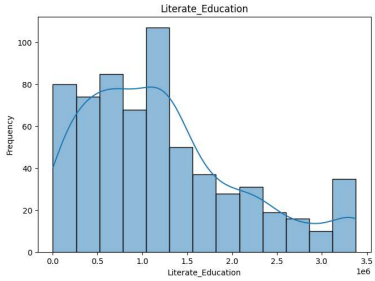

Middle_Education
Frequency
Middle_Education

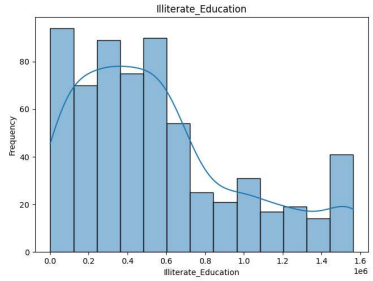

Secondary_Education
Frequency
Secondary_Education

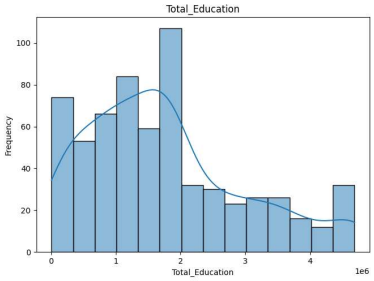

Higher_Education
Frequency
Higher_Education

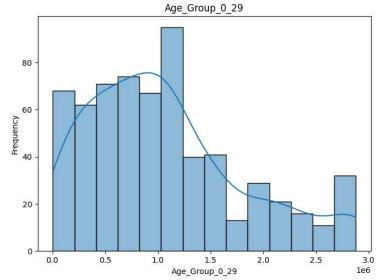

Graduate_Education
Frequency
Graduate_Education

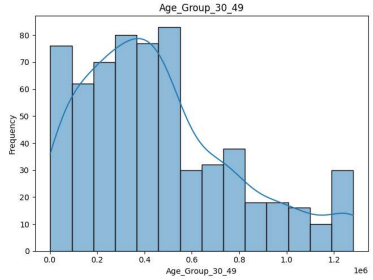

Other_Education
Frequency
Other_Education

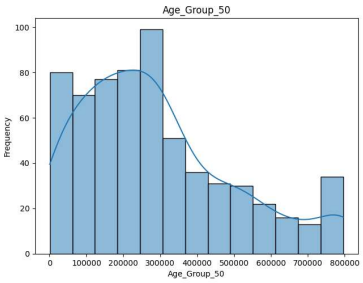

Literate_Education
Frequency
Literate_Education
1e6

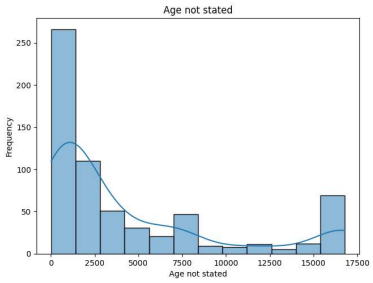

Illiterate_Education
Frequency
Illiterate_Education
1e6

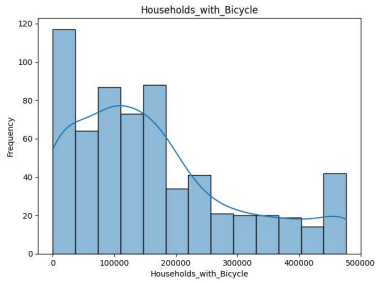

Total_Education
Frequency
Total_Education
1e6

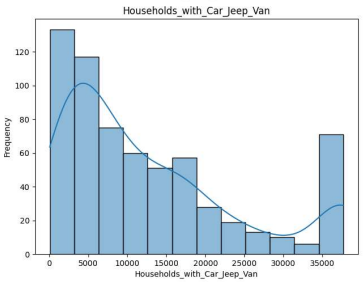

Age_Group_0_29
Frequency
Age_Group_0_29
1e6

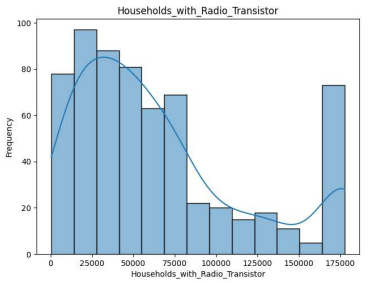

Age_Group_30_49
Frequency
Age_Group_30_49
1e6

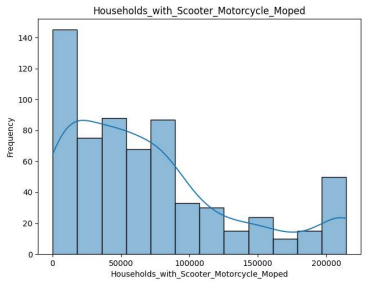

Age_Group_50
Frequency
Age_Group_50

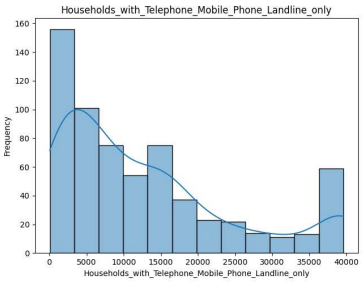

Age not stated
Frequency
Age not stated

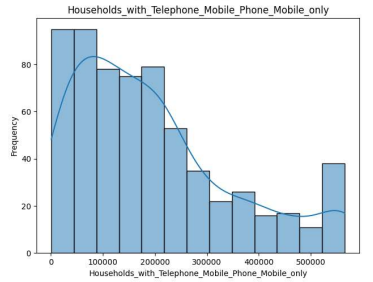

Households_with_Bicycle
Frequency
Households_with_Bicycle

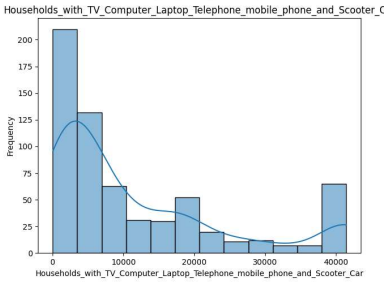

Households_with_Car_Jeep_Van
Frequency
Households_with_Car_Jeep_Van

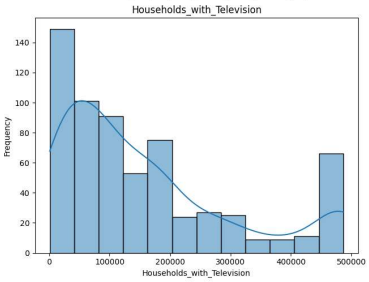

Households_with_Radio_Transistor
Frequency
Households_with_Radio_Transistor

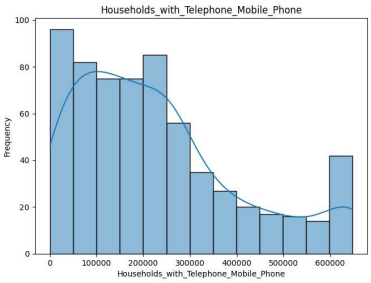

Households_with_Scooter_Motorcycle_Moped
Frequency
Households_with_Scooter_Motorcycle_Moped

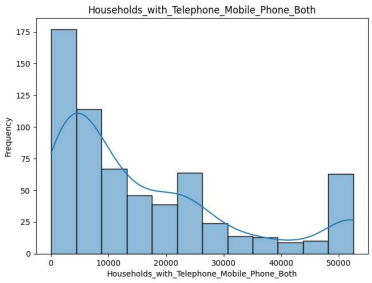

Households_with_Telephone_Mobile_Phone_Landline_only
Frequency
Households_with_Telephone_Mobile_Phone_Landline_only

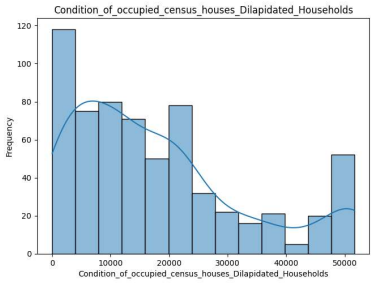

Households_with_Telephone_Mobile_Phone_Mobile_only
Frequency
Households_with_Telephone_Mobile_Phone_Mobile_only

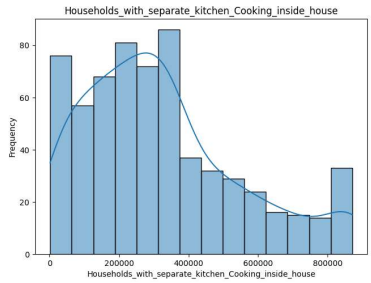

Households_with_TV_Computer_Laptop_Telephone_mobile_phone_and_Scooter_C
Frequency
Households_with_TV_Computer_Laptop_Telephone_mobile_phone_and_Scooter_Car

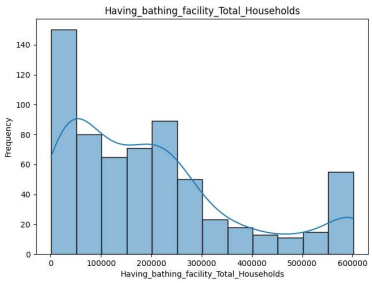

Households_with_Television
Frequency
Households_with_Television

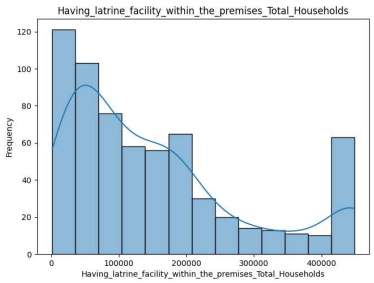

Households_with_Telephone_Mobile_Phone
Frequency
Households_with_Telephone_Mobile_Phone

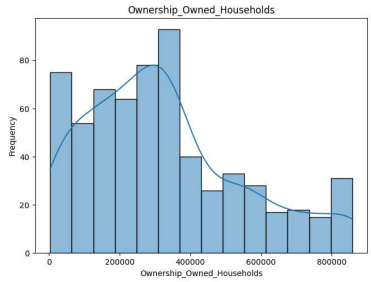

Households_with_Telephone_Mobile_Phone_Both
Frequency
Households_with_Telephone_Mobile_Phone_Both

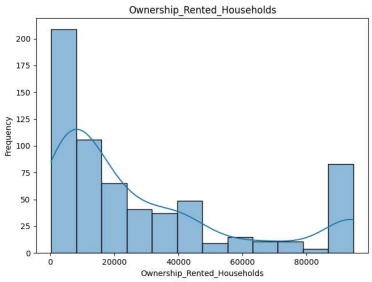

Condition_of_occupied_census_houses_Dilapidated_Households
Frequency
Condition_of_occupied_census_houses_Dilapidated_Households

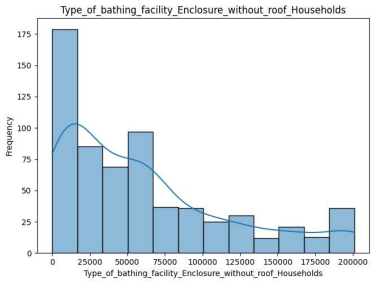

Households_with_separate_kitchen_Cooking_inside_house
Frequency
Households_with_separate_kitchen_Cooking_inside_house

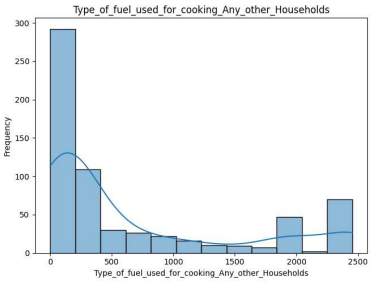

Having_bathing_facility_Total_Households
Frequency
Having_bathing_facility_Total_Households

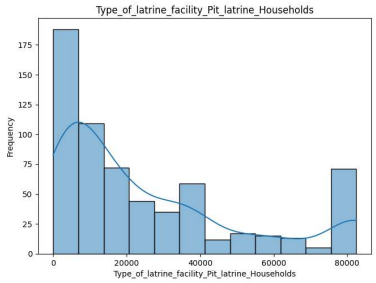

Having_latrine_facility_within_the_premises_Total_Households
Frequency
Having_latrine_facility_within_the_premises_Total_Households

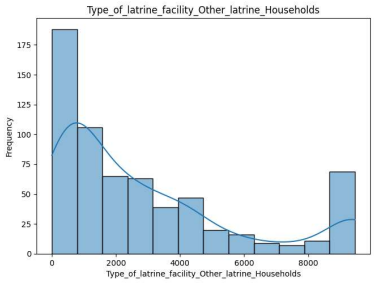

Ownership_Owned_Households
Frequency
Ownership_Owned_Households

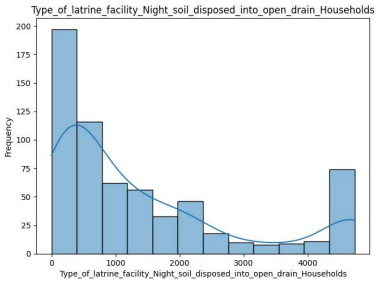

Ownership_Rented_Households
Frequency
Ownership_Rented_Households

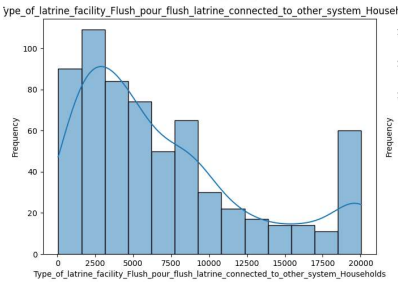

Type_of_bathing_facility_Enclosure_without_roof_Households
Frequency
Type_of_bathing_facility_Enclosure_without_roof_Households

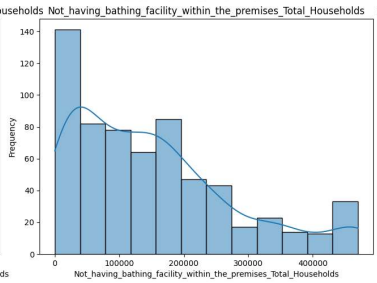

Type_of_fuel_used_for_cooking_Any_other_Households
Frequency
Type_of_fuel_used_for_cooking_Any_other_Households

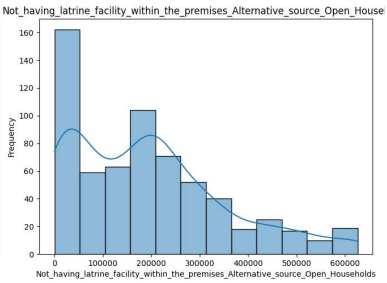

Type_of_latrine_facility_Pit_latrine_Households
Frequency
Type_of_latrine_facility_Pit_latrine_Households

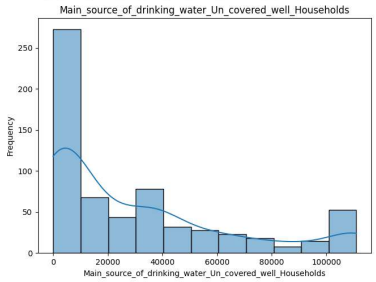

Type_of_latrine_facility_Other_latrine_Households
Frequency
Type_of_latrine_facility_Other_latrine_Households

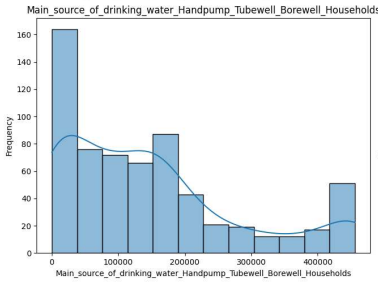

Type_of_latrine_facility_Night_soil_disposed_into_open_drain_Households
Frequency
Type_of_latrine_facility_Night_soil_disposed_into_open_drain_Households

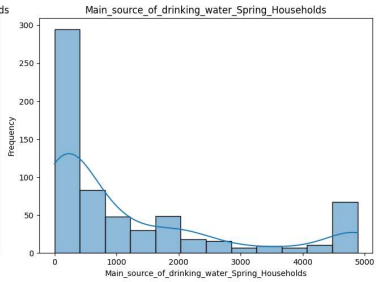

ype_of_latrine_facility_Flush_pour_flush_latrine_connected_to_other_system_Households
Frequency
Type_of_latrine_facility_Flush_pour_flush_latrine_connected_to_other_system_Households

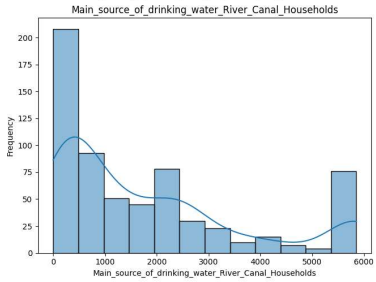

Not_having_bathing_facility_within_the_premises_Total_Households
Frequency
Not_having_bathing_facility_within_the_premises_Total_Households

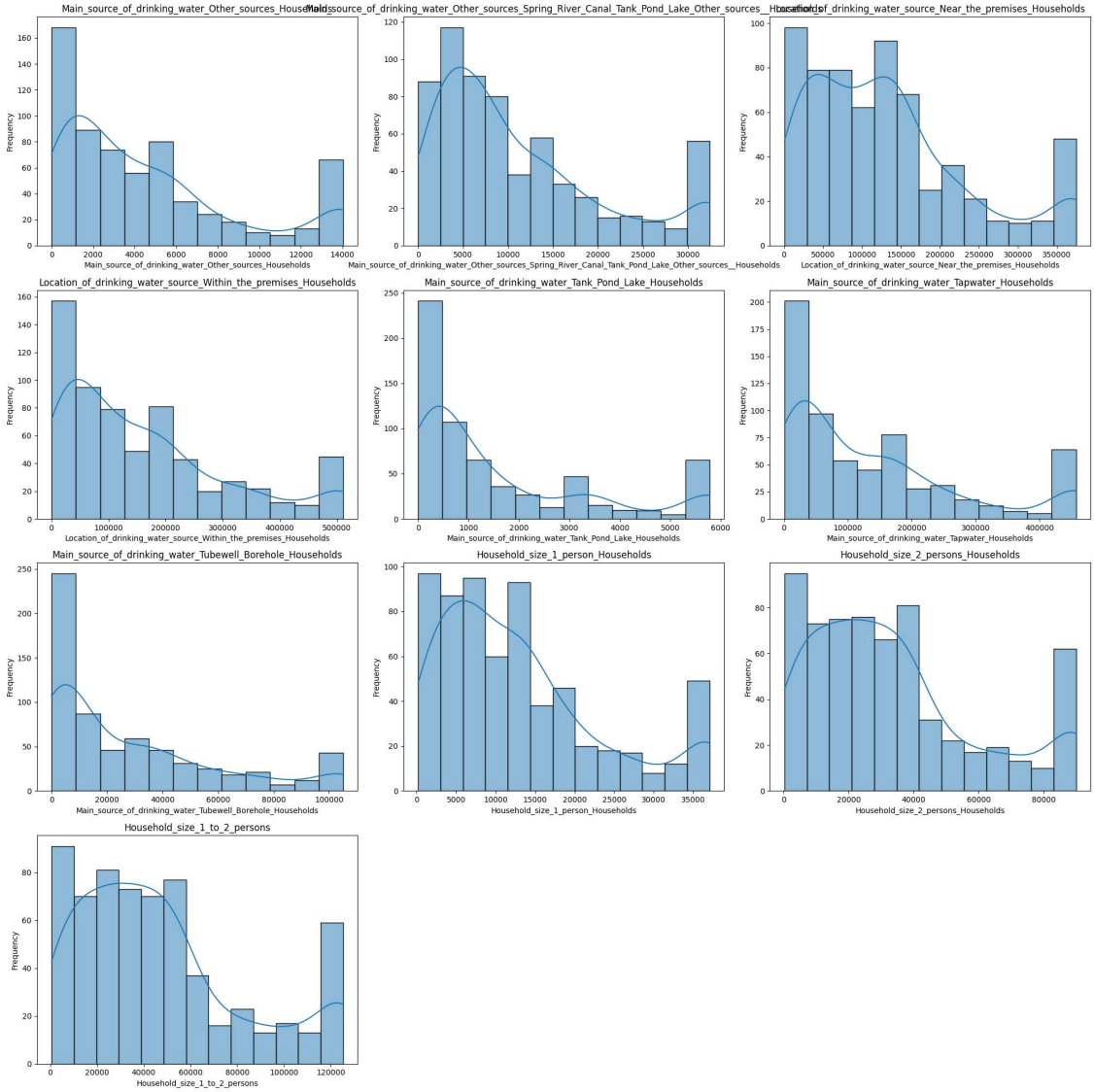

Not_having_latrine_facility_within_the_premises_Alternative_source_Open_House
Frequency
Not_having_latrine_facility_within_the_premises_Alternative_source_Open_Households

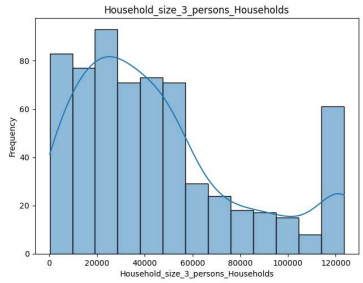

Main_source_of_drinking_water_Un_covered_well_Households
Frequency
Main_source_of_drinking_water_Un_covered_well_Households

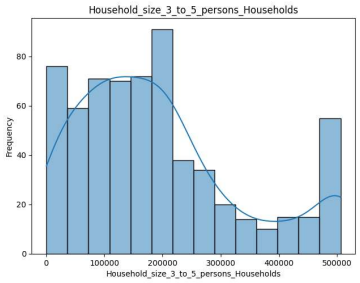

Main_source_of_drinking_water_Handpump_Tubewell_Borewell_Households
Frequency
Main_source_of_drinking_water_Handpump_Tubewell_Borewell_Households

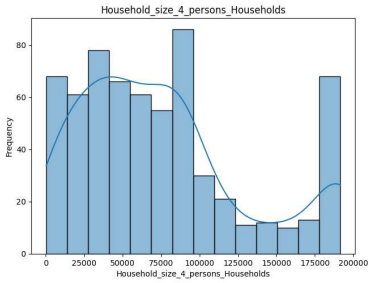

Main_source_of_drinking_water_Spring_Households
Frequency
Main_source_of_drinking_water_Spring_Households

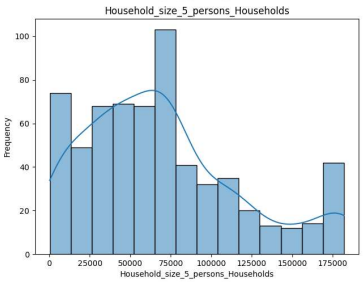

Main_source_of_drinking_water_River_Canal_Households
Frequency
Main_source_of_drinking_water_River_Canal_Households

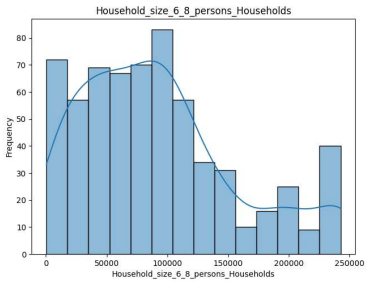

Main_source_of_drinking_water_Other_sources_Households
Frequency
Main_source_of_drinking_water_Other_sources_Households
Main_source_of_drinking_water_Other_sources_Spring_River_Canal_Tank_Pond_Lake_Other_sources__Households
Frequency
Main_source_of_drinking_water_Other_sources_Spring_River_Canal_Tank_Pond_Lake_Other_sources__Households
Location_of_drinking_water_source_Near_the_premises_Households
Frequency
Location_of_drinking_water_source_Near_the_premises_Households
Location_of_drinking_water_source_Within_the_premises_Households
Frequency
Location_of_drinking_water_source_Within_the_premises_Households
Main_source_of_drinking_water_Tank_Pond_Lake_Households
Frequency
Main_source_of_drinking_water_Tank_Pond_Lake_Households
Main_source_of_drinking_water_Tapwater_Households
Frequency
Main_source_of_drinking_water_Tapwater_Households
Main_source_of_drinking_water_Tubewell_Borehole_Households
Frequency
Main_source_of_drinking_water_Tubewell_Borehole_Households
Household_size_1_person_Households
Frequency
Household_size_1_person_Households
Household_size_2_persons_Households
Frequency
Household_size_2_persons_Households
Household_size_1_to_2_persons
Frequency
Household_size_1_to_2_persons

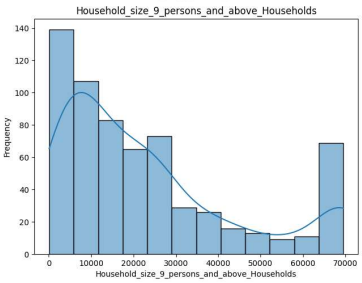

Household_size_3_persons_Households
Frequency
Household_size_3_persons_Households

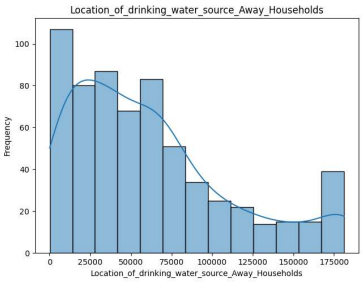

Household_size_3_to_5_persons_Households
Frequency
Household_size_3_to_5_persons_Households

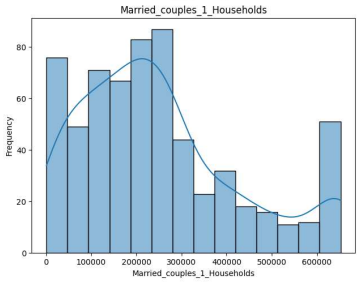

Household_size_4_persons_Households
Frequency
Household_size_4_persons_Households

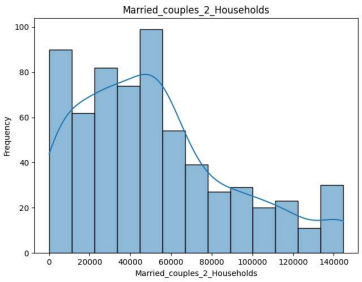

Household_size_5_persons_Households
Frequency
Household_size_5_persons_Households

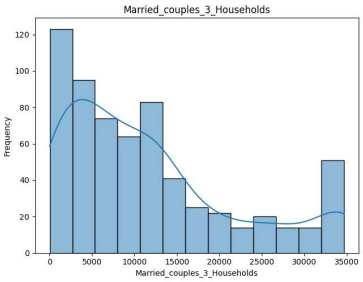

Household_size_6_8_persons_Households
Frequency
Household_size_6_8_persons_Households

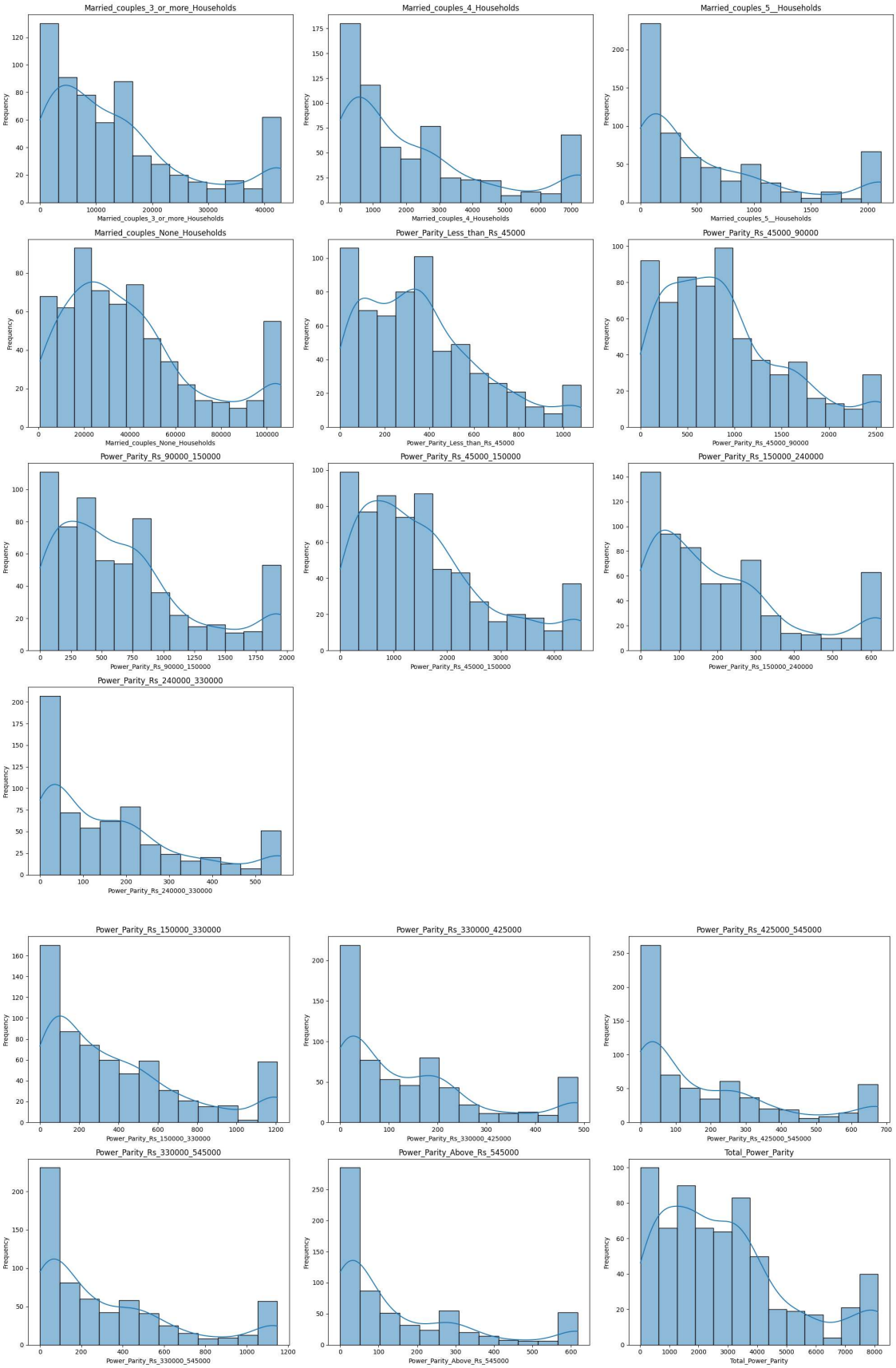

Household_size_9_persons_and_above_Households
Frequency
Household_size_9_persons_and_above_Households


Location_of_drinking_water_source_Away_Households
Frequency
Location_of_drinking_water_source_Away_Households


Married_couples_1_Households
Frequency
Married_couples_1_Households


Married_couples_2_Households
Frequency
Married_couples_2_Households


Married_couples_3_Households
Frequency
Married_couples_3_Households


Married_couples_3_or_more_Households
Frequency
Married_couples_3_or_more_Households
Married_couples_4_Households
Frequency
Married_couples_4_Households
Married_couples_5__Households
Frequency
Married_couples_5__Households
Married_couples_None_Households
Frequency
Married_couples_None_Households
Power_Parity_Less_than_Rs_45000
Frequency
Power_Parity_Less_than_Rs_45000
Power_Parity_Rs_45000_90000
Frequency
Power_Parity_Rs_45000_90000
Power_Parity_Rs_90000_150000
Frequency
Power_Parity_Rs_90000_150000
Power_Parity_Rs_45000_150000
Frequency
Power_Parity_Rs_45000_150000
Power_Parity_Rs_150000_240000
Frequency
Power_Parity_Rs_150000_240000
Power_Parity_Rs_240000_330000
Frequency
Power_Parity_Rs_240000_330000
Power_Parity_Rs_150000_330000
Frequency
Power_Parity_Rs_150000_330000
Power_Parity_Rs_330000_425000
Frequency
Power_Parity_Rs_330000_425000
Power_Parity_Rs_425000_545000
Frequency
Power_Parity_Rs_425000_545000
Power_Parity_Rs_330000_545000
Frequency
Power_Parity_Rs_330000_545000
Power_Parity_Above_Rs_545000
Frequency
Power_Parity_Above_Rs_545000
Total_Power_Parity
Frequency
Total_Power_Parity

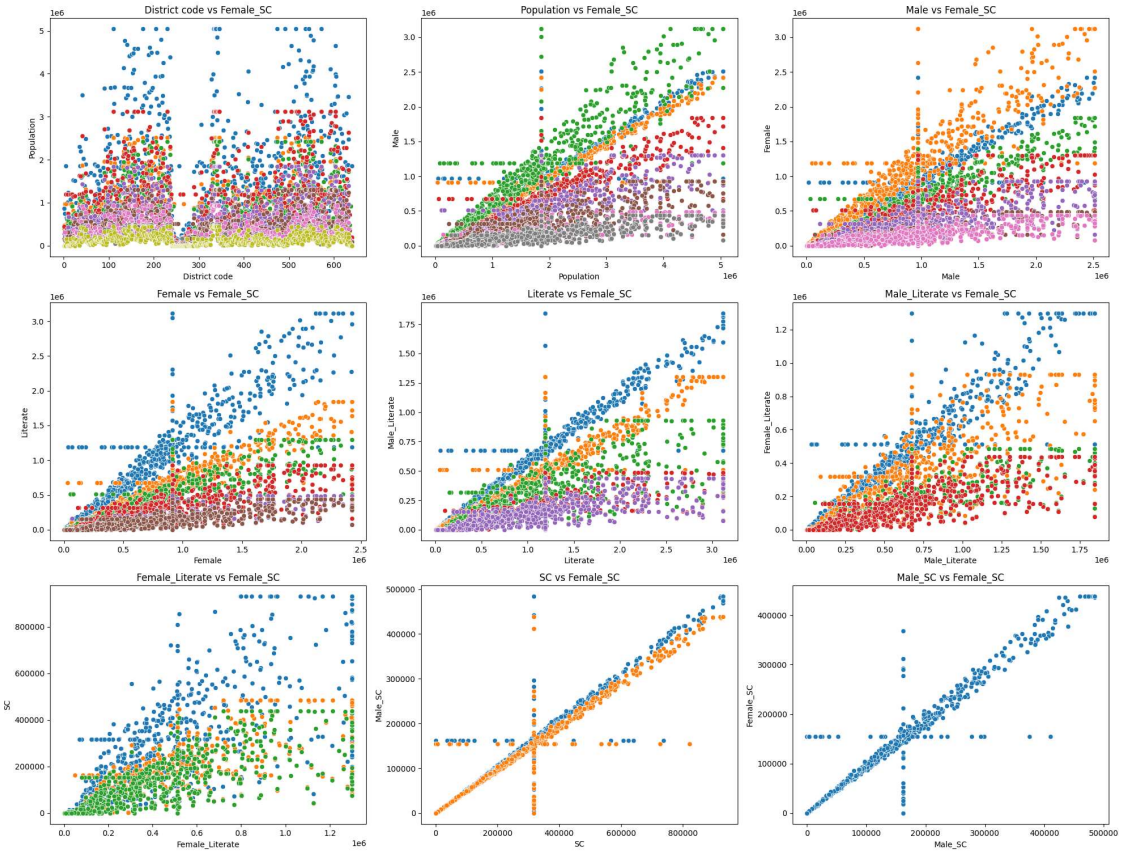
General Insights: Most graphs above do not have a normal distribution. For these kinds of graph, it is noticed that the starting and the ending values for the x-axis are unusually elevated from the rest of the graph. And for graphs that have more normal a distribution, its noticed that the median value is more elevated than the surrounding values. But its noted that these points of anomalies do not affect the overall distribution of data.

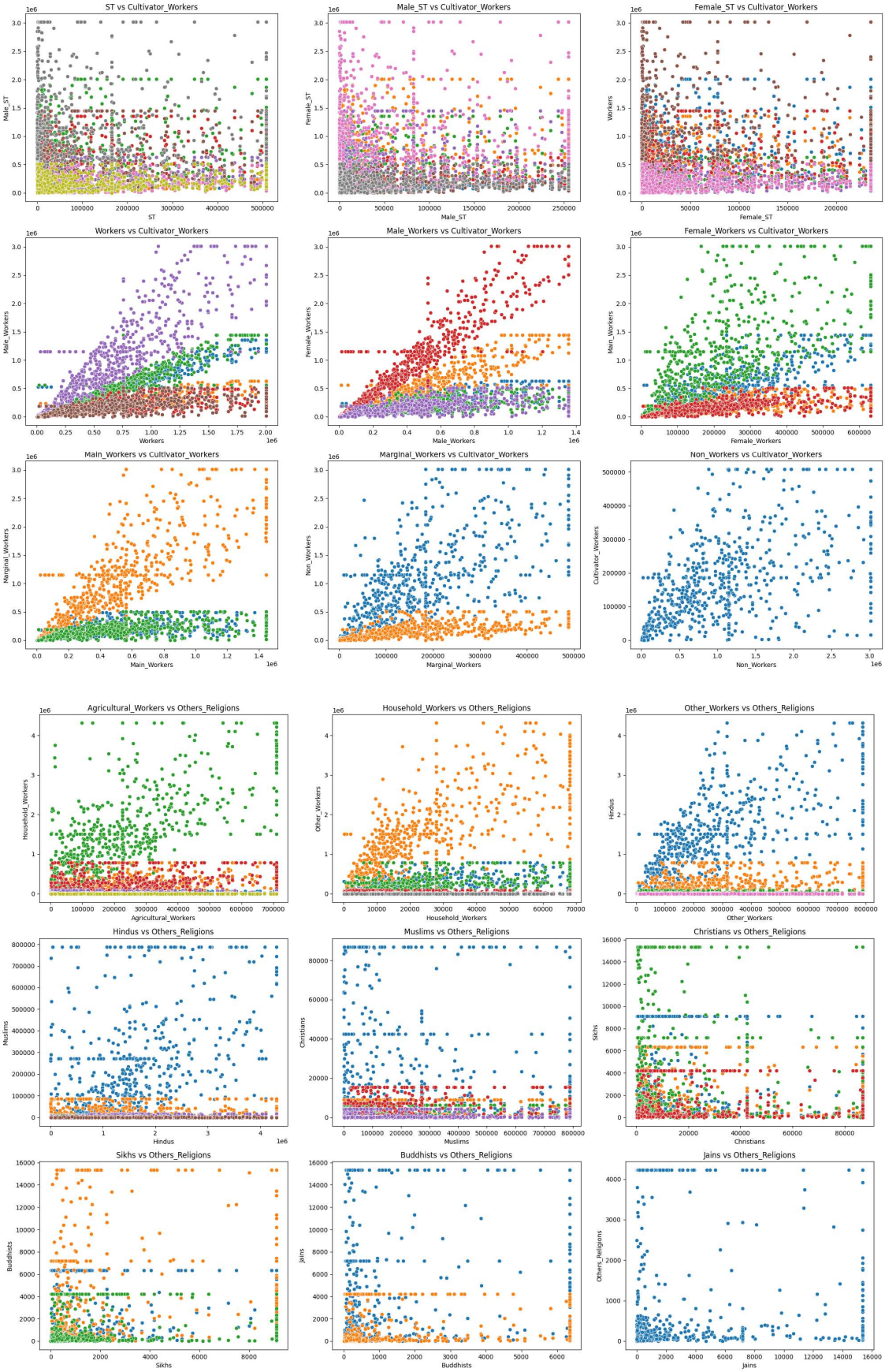
Distribution shapes: Most graphs show a right-skewed distribution, with the highest frequency at or near zero, and a long tail extending to the right. This is due to the outlier handling method used that clips the outliers beyond a range to the lower and upper bounds of data for non-normal distributions

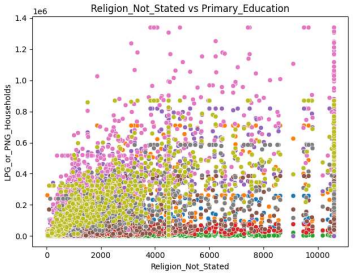
Most variables show a concentration of frequencies at lower values, with decreasing frequencies as the value increases, indicating potential income or resource disparities in the population studied.

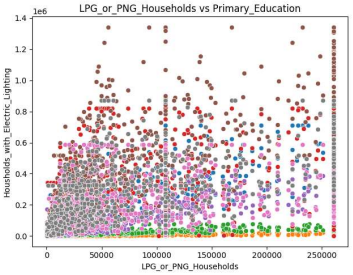
BIVARIATE ANALSIS:

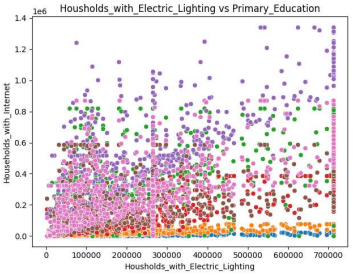
For bivariate analysis we take pairs of columns and make a scatter plot

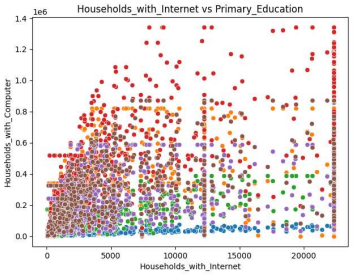


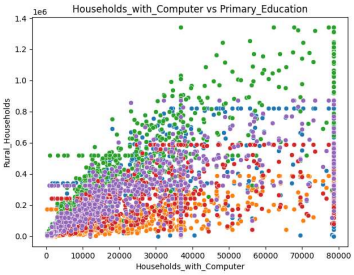

ST vs Cultivator_Workers
Male_ST vs Cultivator_Workers
Female_ST vs Cultivator_Workers
Workers vs Cultivator_Workers
Male_Workers vs Cultivator_Workers
Female_Workers vs Cultivator_Workers
Main_Workers vs Cultivator_Workers
Marginal_Workers vs Cultivator_Workers
Non_Workers vs Cultivator_Workers
Agricultural_Workers vs Others_Religions
Household_Workers vs Others_Religions
Other_Workers vs Others_Religions
Hindus vs Others_Religions
Muslims vs Others_Religions
Christians vs Others_Religions
Sikhs vs Others_Religions
Buddhists vs Others_Religions
Jains vs Others_Religions

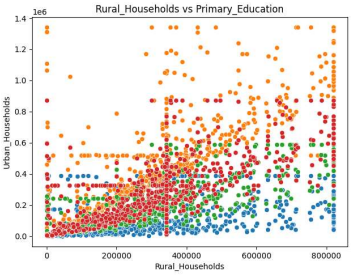

Religion_Not_Stated vs Primary_Education
LPG_or_PNG_Households
Religion_Not_Stated
1e6

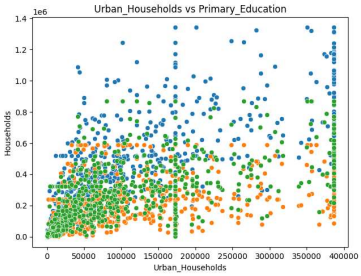

LPG_or_PNG_Households vs Primary_Education
Housholds_with_Electric_Lighting
LPG_or_PNG_Households
1e6

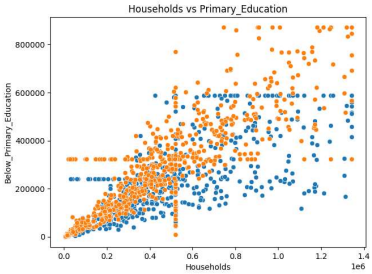

Housholds_with_Electric_Lighting vs Primary_Education
Households_with_Internet
Housholds_with_Electric_Lighting
1e6

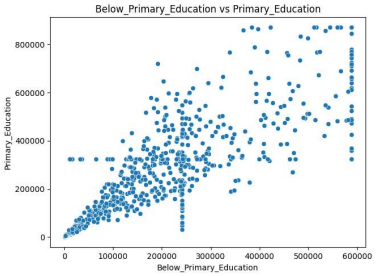

Households_with_Internet vs Primary_Education
Households_with_Computer
Households_with_Internet
1e6

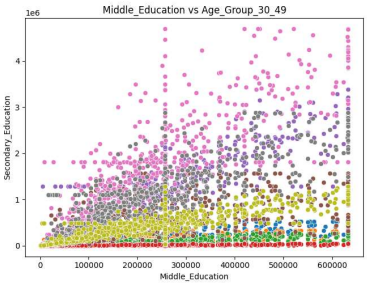

Households_with_Computer vs Primary_Education
Rural_Households
Households_with_Computer
1e6

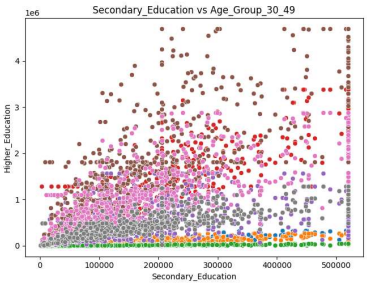

Rural_Households vs Primary_Education
Urban_Households
Rural_Households
1e6

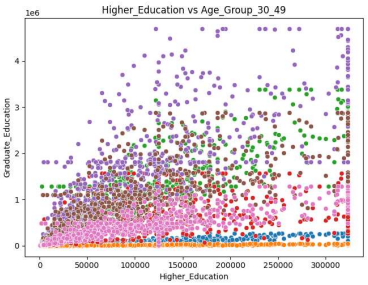

Urban_Households vs Primary_Education
Households
Urban_Households
1e6

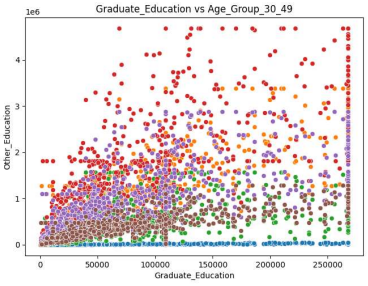

Households vs Primary_Education
Below_Primary_Education
Households
1e6

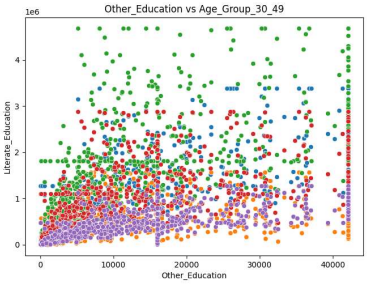

Below_Primary_Education vs Primary_Education
Primary_Education
Below_Primary_Education

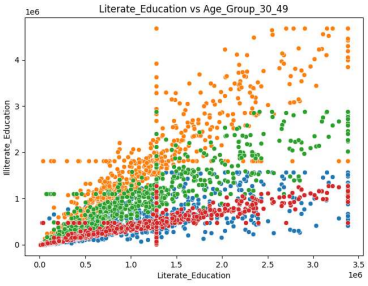

Middle_Education vs Age_Group_30_49
Secondary_Education
Middle_Education
1e6

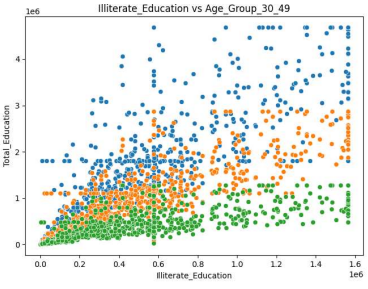

Secondary_Education vs Age_Group_30_49
Higher_Education
Secondary_Education
1e6

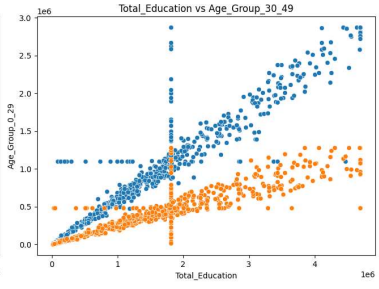

Higher_Education vs Age_Group_30_49
Graduate_Education
Higher_Education
1e6

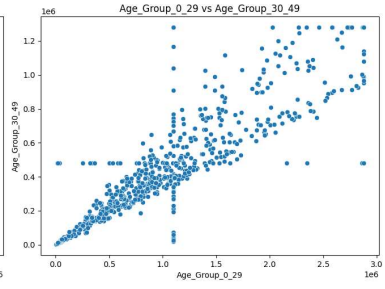

Graduate_Education vs Age_Group_30_49
Other_Education
Graduate_Education
1e6


Other_Education vs Age_Group_30_49
Literate_Education
Other_Education
1e6

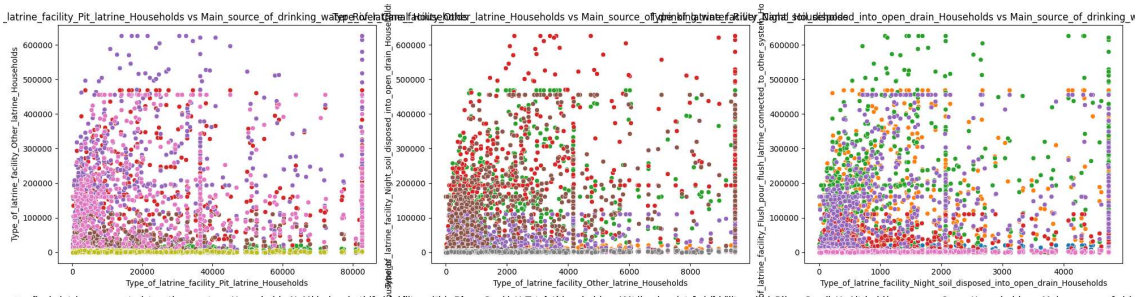

Literate_Education vs Age_Group_30_49
Illiterate_Education
Literate_Education
1e6

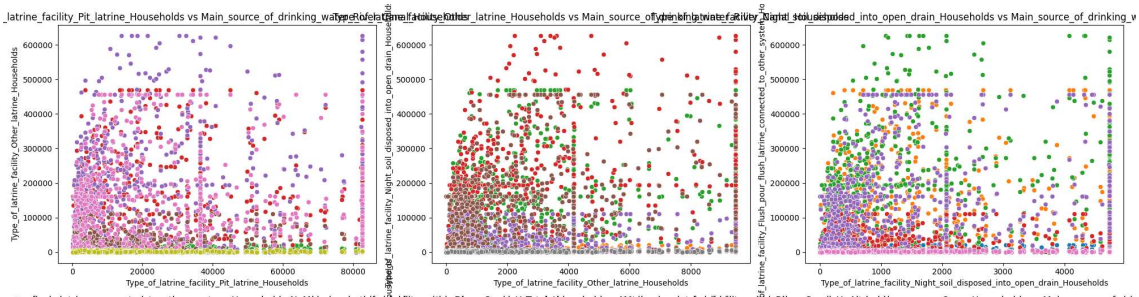

Illiterate_Education vs Age_Group_30_49
Total_Education
Illiterate_Education
1e6

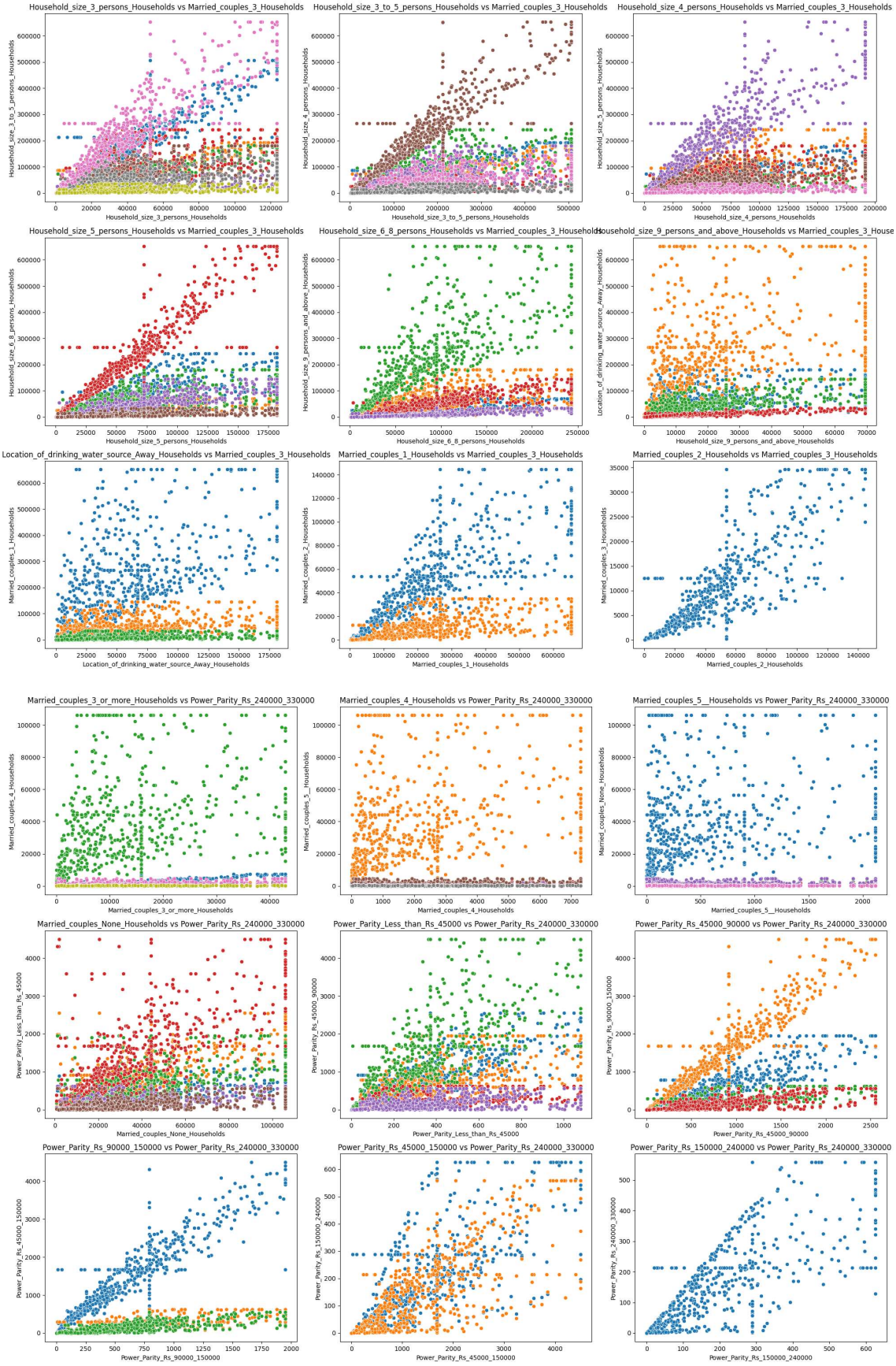

Total_Education vs Age_Group_30_49
Age_Group_0_29
Total_Education
1e6

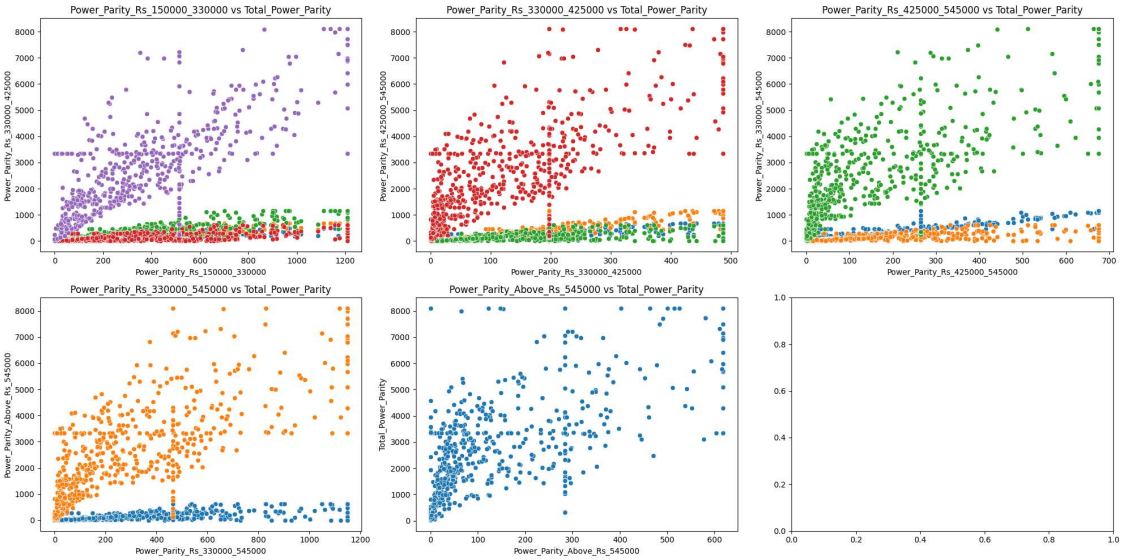

Age_Group_0_29 vs Age_Group_30_49
Age_Group_30_49
Age_Group_0_29
1e6


Age_Group_50 vs Households_with_Television
Age_Group_50
Age not stated
Age not stated vs Households_with_Television
Age not stated
Households_with_Bicycle
Households_with_Bicycle vs Households_with_Television
Households_with_Bicycle
Households_with_Car_Jeep_Van
Households_with_Car_Jeep_Van vs Households_with_Television
Households_with_Car_Jeep_Van
Households_with_Radio_Transistor
Households_with_Radio_Transistor vs Households_with_Television
Households_with_Radio_Transistor
Households_with_Scooter_Motorcycle_Moped
Households_with_Scooter_Motorcycle_Moped vs Households_with_Television
Households_with_Scooter_Motorcycle_Moped
Households_with_Telephone_Mobile_Phone_Landline_only
Households_with_Telephone_Mobile_Phone_Landline_only vs Households_with_Television
Households_with_Telephone_Mobile_Phone_Landline_only
Households_with_Telephone_Mobile_Phone_Mobile_only
Households_with_Telephone_Mobile_Phone_Mobile_only vs Households_with_TV_Computer_Laptop_Telephone_mobile_phone_and_Scooter_Car
Households_with_Telephone_Mobile_Phone_Mobile_only
Households_with_TV_Computer_Laptop_Telephone_mobile_phone_and_Scooter_Car
Households_with_TV_Computer_Laptop_Telephone_mobile_phone_and_Scooter_Car vs Households_with_Television
Households_with_Television
Households_with_Telephone_Mobile_Phone
eholds_with_Telephone_Mobile_Phone vs Type_of_fuel_used_for_cooking_Any_other_Households
Households_with_Telephone_Mobile_Phone
Households_with_Telephone_Mobile_Phone_Both
Households_with_Telephone_Mobile_Phone_Both vs Type_of_fuel_used_for_cooking_Any_other_Households
Households_with_Telephone_Mobile_Phone_Both
Condition_of_occupied_census_houses_Dilapidated_Households
Condition_of_occupied_census_houses_Dilapidated_Households vs Type_of_fuel_used_for_cooking
Condition_of_occupied_census_houses_Dilapidated_Households
Households_with_separate_kitchen_Cooking_inside_house
with_separate_kitchen_Cooking_inside_house vs Type_of_fuel_used_for_cooking_Any_other_Households
Households_with_separate_kitchen_Cooking_inside_house
Having_bathing_facility_Total_Households
Having_bathing_facility_Total_Households vs Type_of_fuel_used_for_cooking_Any_other_Households
Having_bathing_facility_Total_Households
Having_latrine_facility_within_the_premises_Total_Households
Having_latrine_facility_within_the_premises_Total_Households vs Type_of_fuel_used_for_cooking
Having_latrine_facility_within_the_premises_Total_Households
Ownership_Owned_Households
Ownership_Owned_Households vs Type_of_fuel_used_for_cooking_Any_other_Households
Ownership_Owned_Households
Ownership_Rented_Households
Ownership_Rented_Households vs Type_of_fuel_used_for_cooking_Any_other_Households
Ownership_Rented_Households
Type_of_bathing_facility_Enclosure_without_roof_Households
Type_of_bathing_facility_Enclosure_without_roof_Households vs Type_of_fuel_used_for_cooking
Type_of_bathing_facility_Enclosure_without_roof_Households
Type_of_fuel_used_for_cooking_Any_other_Households


Type_of_latrine_facility_Pit_latrine_Households
Type_of_latrine_facility_Other_latrine_Households
Type_of_latrine_facility_Night_soil_disposed_into_open_drain_Households
Type_of_latrine_facility_Flush_pour_flush_latrine_connected_to_other_system_Households
Not_having_bathing_facility_within_the_premises_Total_Households
Not_having_latrine_facility_within_the_premises_Alternative_source_Open_Households
Main_source_of_drinking_water_Un_covered_well_Households
Main_source_of_drinking_water_Handpump_Tubewell_Borewell_Households
Main_source_of_drinking_water_Spring_Households
Main_source_of_drinking_water_River_Canal_Households


Type_of_latrine_facility_Pit_latrine_Households
Type_of_latrine_facility_Other_latrine_Households
Type_of_latrine_facility_Night_soil_disposed_into_open_drain_Households
Type_of_latrine_facility_Flush_pour_flush_latrine_connected_to_other_system_Households
Not_having_bathing_facility_within_the_premises_Total_Households
Not_having_latrine_facility_within_the_premises_Alternative_source_Open_Households
Main_source_of_drinking_water_Un_covered_well_Households
Main_source_of_drinking_water_Handpump_Tubewell_Borewell_Households
Main_source_of_drinking_water_Spring_Households
Main_source_of_drinking_water_River_Canal_Households


Household_size_3_persons_Households vs Married_couples_3_Households
Household_size_3_to_5_persons_Households vs Married_couples_3_Households
Household_size_4_persons_Households vs Married_couples_3_Households
Household_size_5_persons_Households vs Married_couples_3_Households
Household_size_6_8_persons_Households vs Married_couples_3_Households
Household_size_9_persons_and_above_Households vs Married_couples_3_House
Location_of_drinking_water_source_Away_Households vs Married_couples_3_Households
Married_couples_1_Households vs Married_couples_3_Households
Married_couples_2_Households vs Married_couples_3_Households
Married_couples_3_or_more_Households vs Power_Parity_Rs_240000_330000
Married_couples_4_Households vs Power_Parity_Rs_240000_330000
Married_couples_5__Households vs Power_Parity_Rs_240000_330000
Married_couples_None_Households vs Power_Parity_Rs_240000_330000
Power_Parity_Less_than_Rs_45000 vs Power_Parity_Rs_240000_330000
Power_Parity_Rs_45000_90000 vs Power_Parity_Rs_240000_330000
Power_Parity_Rs_90000_150000 vs Power_Parity_Rs_240000_330000
Power_Parity_Rs_45000_150000 vs Power_Parity_Rs_240000_330000
Power_Parity_Rs_150000_240000 vs Power_Parity_Rs_240000_330000



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.

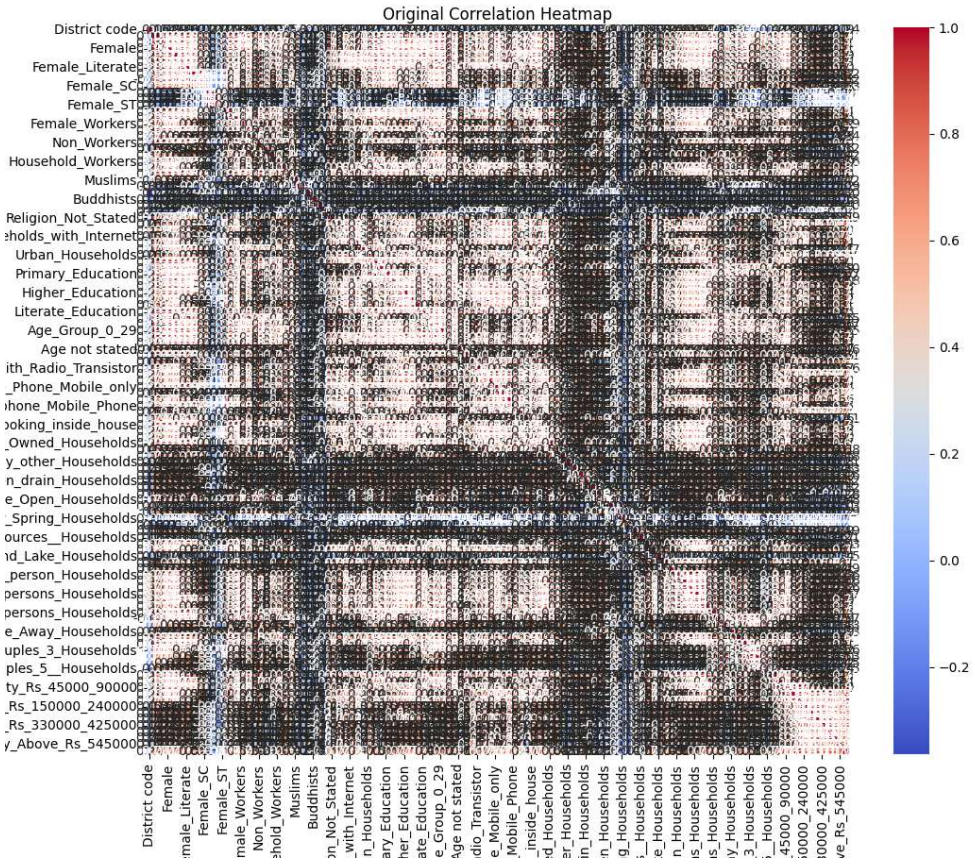
Non-linear relationships: While many relationships appear broadly linear, there are numerous instances of non-linear patterns. This is particularly evident in plots with curved or fan-shaped distributions, indicating more complex interactions between variables.

Ceiling and floor effects: Several plots, especially those involving percentages or ratios, show clear upper and lower bounds. This creates distinctive patterns where data points cluster along maximum or minimum values.

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 concentrations in some areas and sparser distributions in others. This reveals where the majority of cases fall and where exceptional cases lie.

MULTIVARIATE ANALYSIS:



Correlation Coefficients: The colour bar on the right indicates the range of correlation values from -1 (perfect negative correlation, shown in blue) to +1 (perfect positive correlation, shown in red). Values closer to 0 (shown in white) indicate no linear relationship.

Dense Matrix: The heatmap appears quite dense, indicating a large number of variables. Each cell represents the correlation between the variable on the y-axis and the variable on the x-axis.

High Correlation Blocks: There are several blocks of high correlation (dark red) and high negative correlation (dark blue), indicating groups of variables that are strongly positively or negatively correlated with each other

Clustering of Variables: The heatmap shows clusters where certain groups of variables are highly correlated with each other. This can indicate underlying patterns or relationships in the data.

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.
4. **Visualization Overload**:
    - Scatter plots and histograms for each variable and their interactions result in a large number of visualizations.
    - The sheer volume of plots can overwhelm the analysis, making it difficult to draw clear, actionable insights.

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: [0.56922621, 0.09695484, 0.05885575, 0.02791398 0.02446779, 0.01657046, 0.01467139, 0.01305113, 0.01187066 ,0.01041794]

   Top Contributing Features per Principal Component:

   PC1: (Households, Age_Group_30_49, Male_Workers)

   PC2: (Main_source_of_drinking_water_Handpump_Tubewell_Borewell_Households, Not_having_bathing_facility_within_the_premises_Total_Households, Marginal_Workers)

   PC3: (Female_ST, ST, Male_ST)

   PC4: (ST, Male_ST, Female_ST)

   PC5:
   (Main_source_of_drinking_water_Other_sources_Spring_River_Canal_Tank_Pond_Lake_Other_sources__Households,
   Main_source_of_drinking_water_Tank_Pond_Lake_Households,
   Main_source_of_drinking_water_River_Canal_Households)

   PC6: (Others_Religions,
   Main_source_of_drinking_water_Tank_Pond_Lake_Households, Cultivator_Workers)

   PC7: (Sikhs, Type_of_latrine_facility_Other_latrine_Households, District code)

   PC8: (Power_Parity_Rs_150000_240000, Sikhs, Jains)

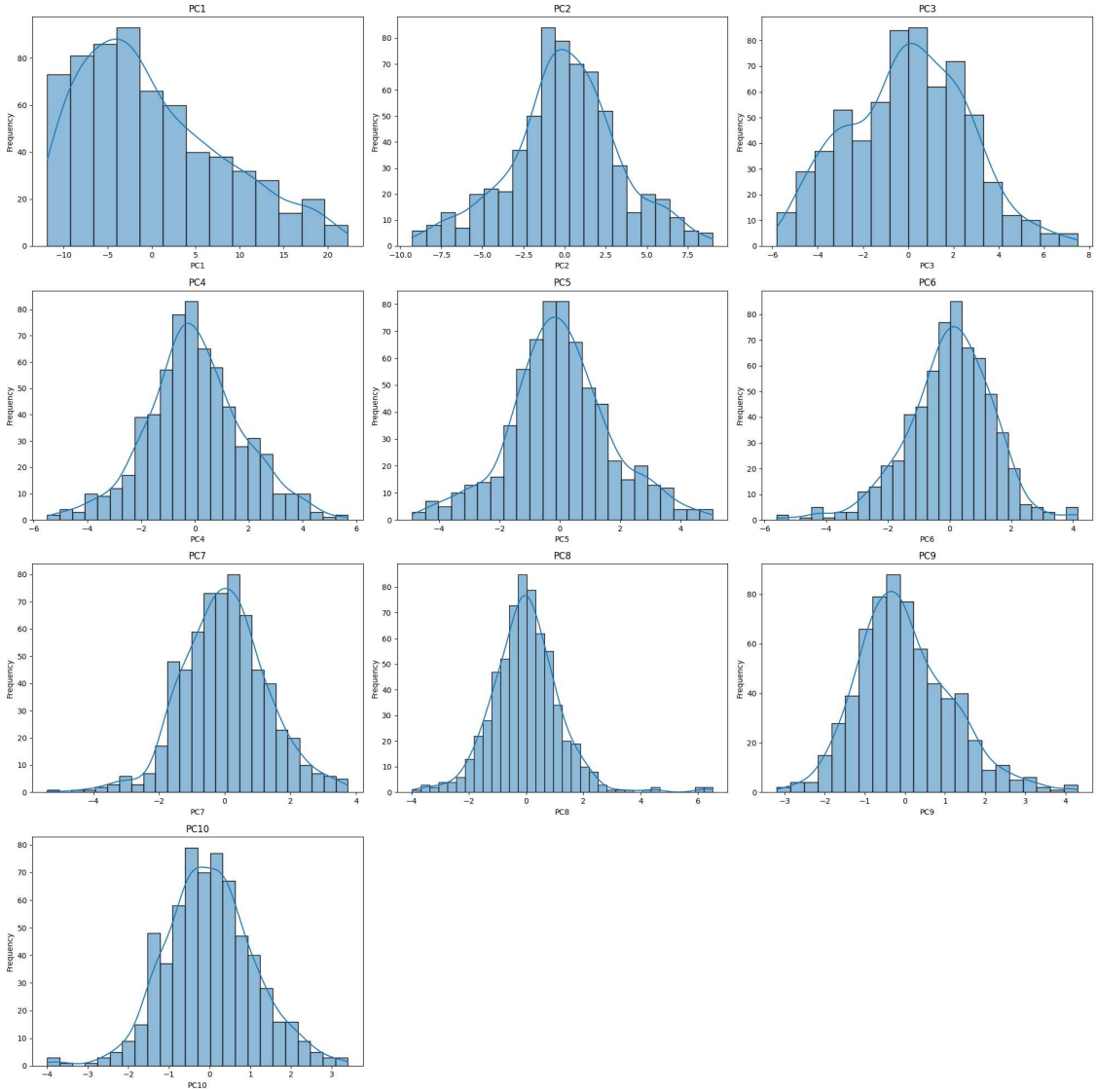
   PC9: (Buddhists, Type_of_latrine_facility_Pit_latrine_Households, Type_of_latrine_facility_Flush_pour_flush_latrine_connected_to_other_system_Households)

PC10: (Type_of_latrine_facility_Pit_latrine_Households,
Type_of_latrine_facility_Other_latrine_Households,
Main_source_of_drinking_water_Un_covered_well_Households)

- The first principal component explains 56.92% of the variance, while the first two components together explain 66.62%.
- By retaining the first five principal components, we capture 78.24% of the total variance, significantly reducing the dimensionality while preserving most of the information.

Phase 2: Analysis Post PCA

UNIVARIATE ANALYSIS:



Normal Distribution:

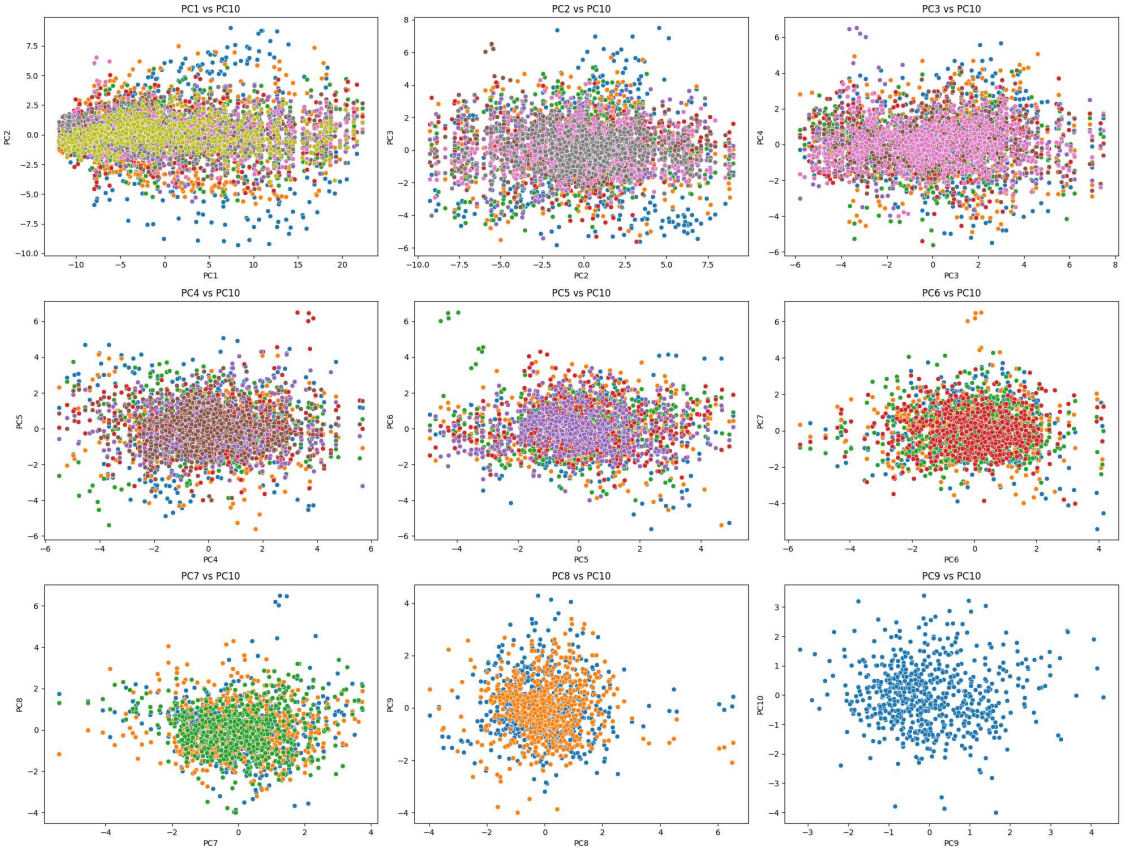
Histograms of the principal components show approximately normal distributions, indicating successful transformation.

These distributions facilitate easier interpretation and further statistical analysis

Dominant Structure by Demographics and Households:

The explained variance ratio indicates a strong underlying structure in the data, with the first principal component (PC1) explaining over 56% of the total variance. This suggests that demographics and household characteristics play a dominant role in shaping the data. This is further supported by the top contributing features for PC1, which include "Households," "Age_Group_30_49," and "Male_Workers."

BIVARIATE ANALYSIS:



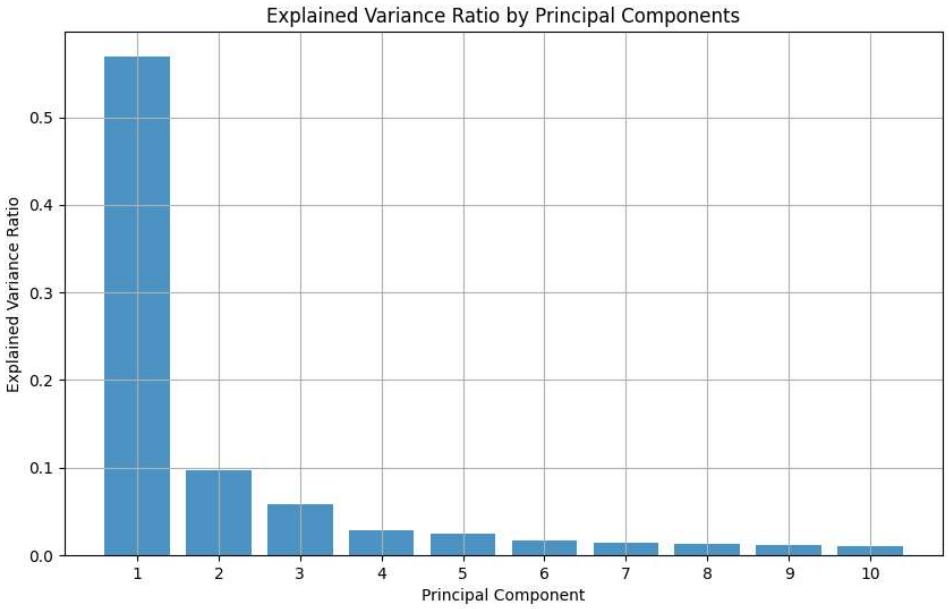
Oval-shaped Scatter Plots:

- The points are densely packed in the middle, suggesting that the majority of the data points are concentrated around the mean, with fewer points as you move away from the center.
- The oval-shaped, densely packed plots indicate that the principal components capture the most significant relationships in the dataset.
- Most of the residuals fall close to zero on the y-axis, which suggests that the model fits the data well.

- There are a few outliers in some of the plots, which means that there are some data points that the model did not fit well.
- The residuals do not show any clear patterns, which suggests that the model is unbiased.

Key takeaways:

The first 5 principal components capture 78.24% of the total variance in the dataset



PCA Method:

The PCA method helped determine the columns that contributed most to the variance, they are: (Households, Age_Group_30_49, Male_Workers, Main_source_of_drinking_water_Handpump_Tubewell_Borewell_Households, Not_having_bathing_facility_within_the_premises_Total_Households, Marginal_Workers)

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.