Detailed EDA Report for [Housing.csv]

INTRODUCTION:

This report presents a detailed analysis of the dataset containing information about housing conditions across various states and districts in India. We aim to explore the distribution of different types of residences, identify key patterns, and uncover insights that can inform housing policy and urban planning.

Overview of Data File

1. Source: Healthcare Data\housing.csv

2. Rows: 1908 entries, containing data for multiple states and districts

3. Columns: 156 columns

4. Datatypes:

- float64 (142 columns)/
- int64 (9 columns)
- object (5 columns, likely including State Name)

Techniques Used Pre-Analysis on Dataset

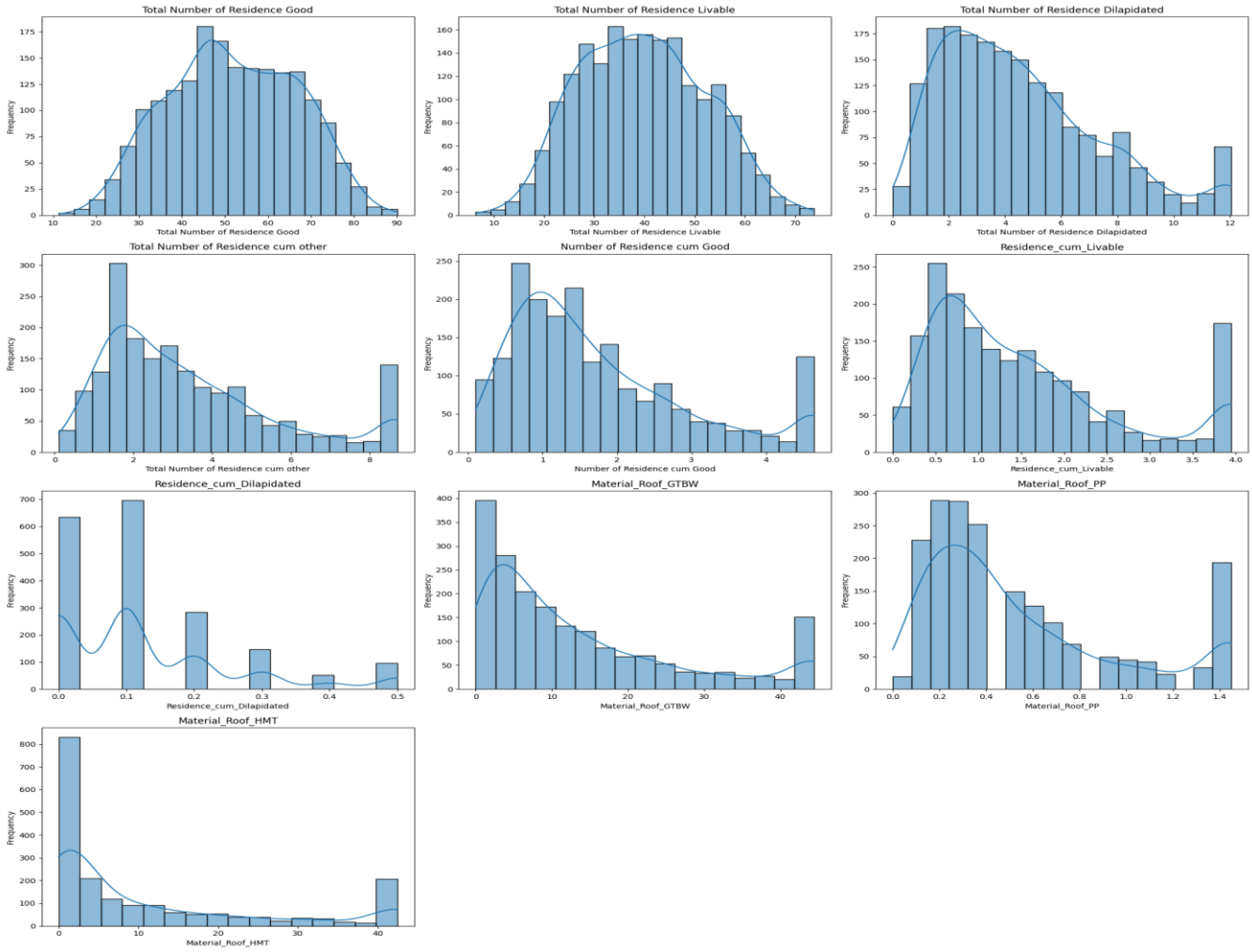
- Data loading and preview
- Initial data type assessment
- Column identification
- Principal Component Analysis (PCA)
- Correlation analysis
- Univariate analysis

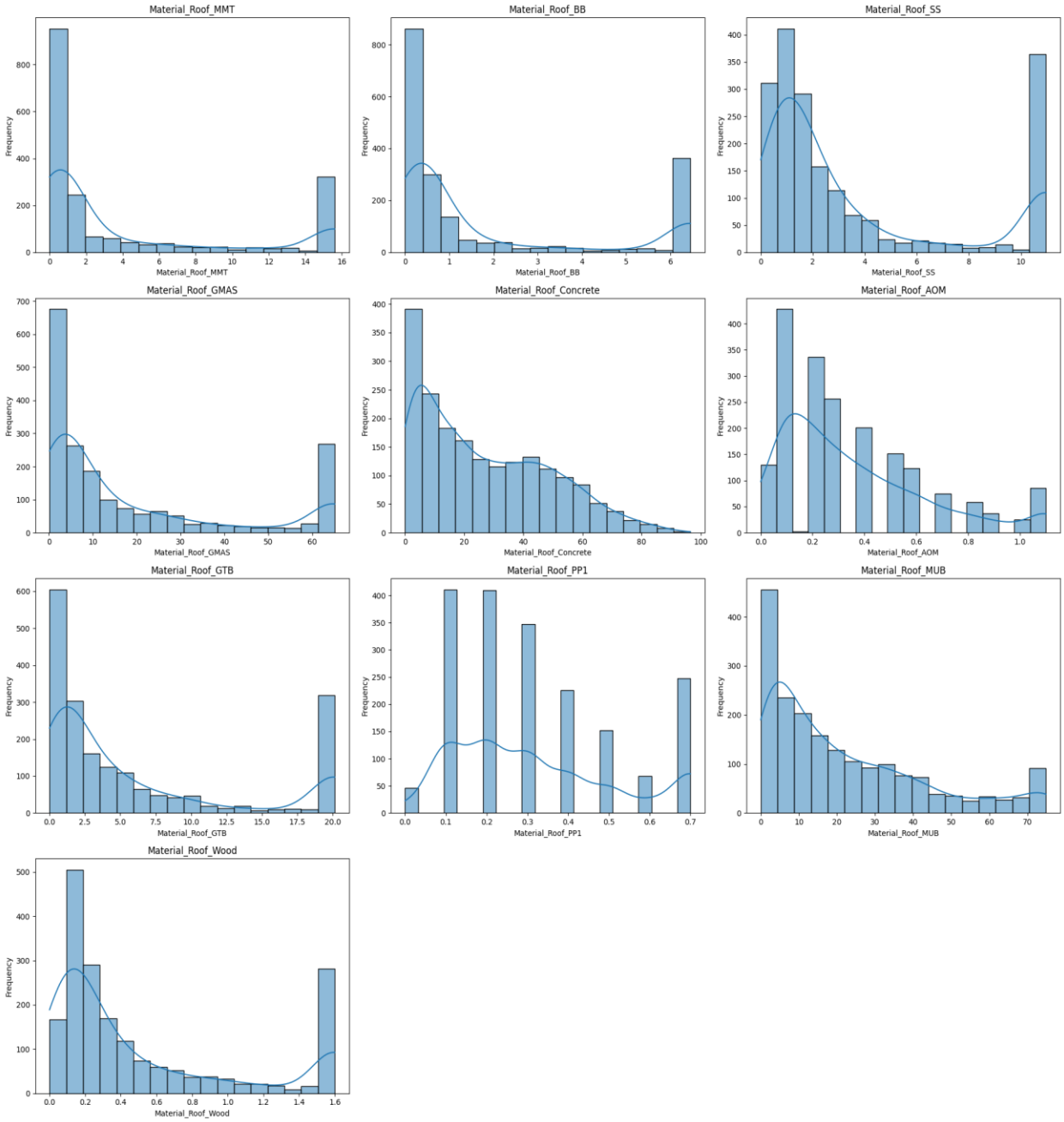
Analysis:

Phase 1: Column Based Analysis

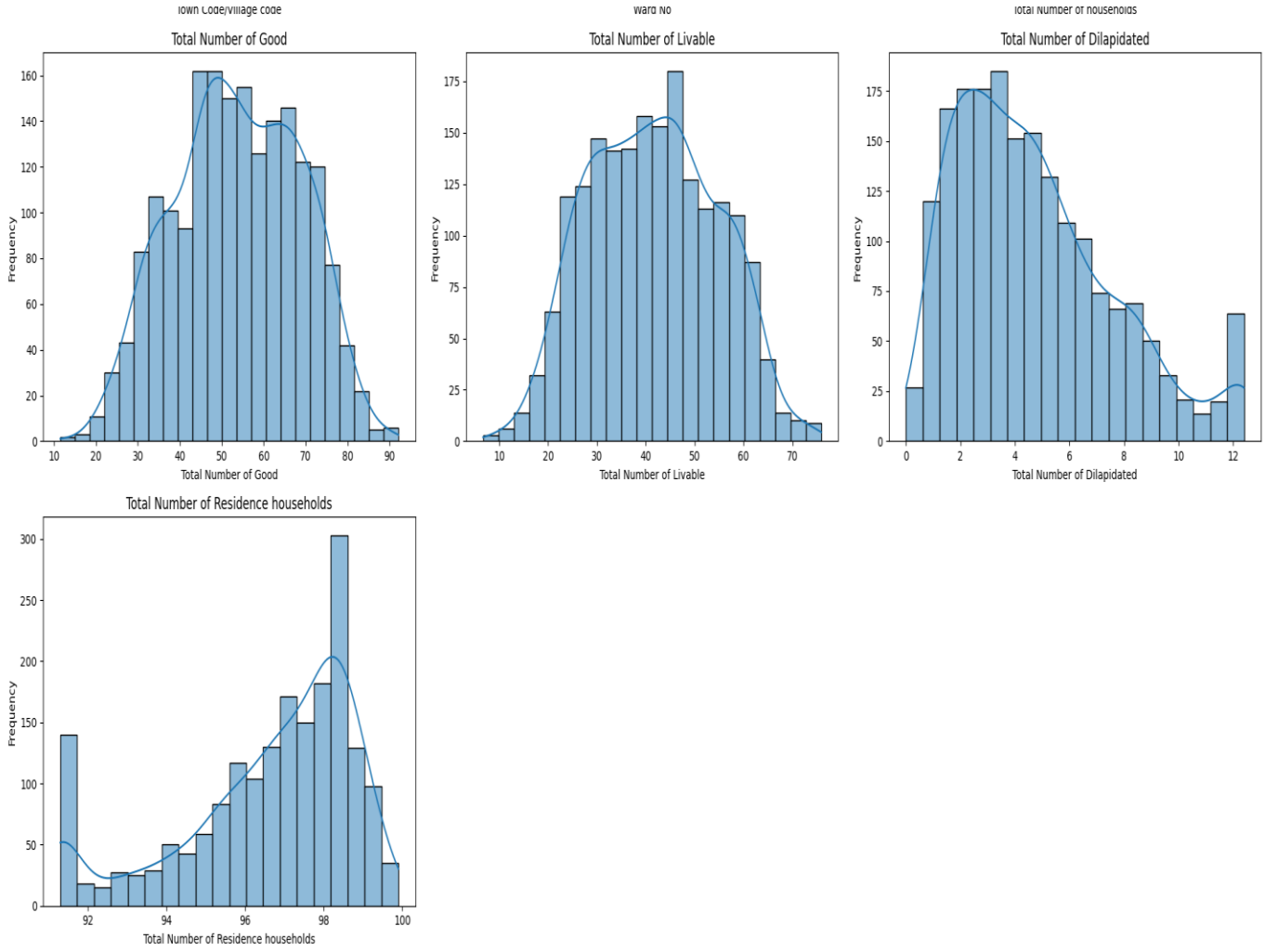
UNIVARIATE ANALYSIS

Due to the large number of variables, we focused on key columns related to housing conditions.




Material_Roof_MMT
Material_Roof_BB
Material_Roof_SS
Material_Roof_GMAS
Material_Roof_Concrete
Material_Roof_AOM
Material_Roof_GTB
Material_Roof_PP1
Material_Roof_MUB
Material_Roof_Wood
Frequency

1. Residence Types:



- The dataset includes information on various residence types, including dilapidated, other, and good residences.

- The distribution of these residence types varies across states and districts.

- Histograms show that good residences are most common, followed by livable, and then dilapidated residences.

2. State and District Distribution:

- The dataset covers multiple states, with Jammu & Kashmir (State Code 1) appearing in the preview.
- Districts are identified by District Codes, allowing for district-level analysis.
- State Code distribution is uneven, suggesting some states have higher representation in the dataset.
- District Code distribution is relatively uniform, indicating balanced representation across districts.

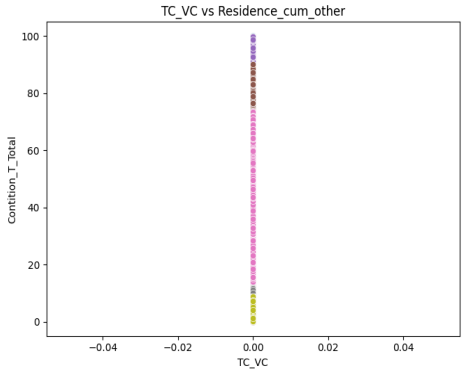
3. Administrative Divisions:

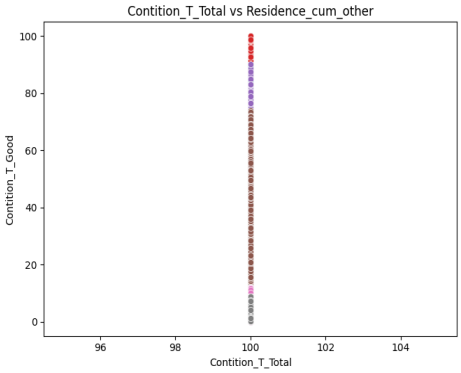
- Tehsil Code, Town Code/Village code, and Ward No show limited variation, possibly due to the aggregation level of the data.

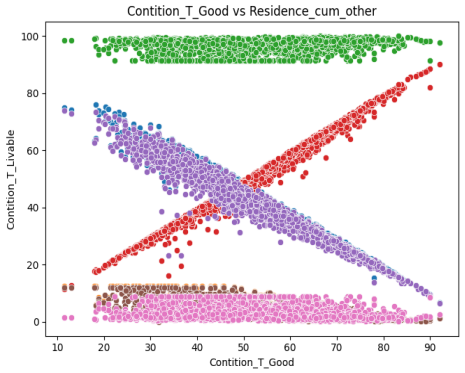
4. Household Distribution:

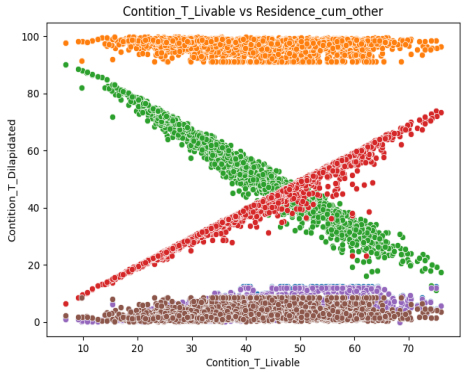
- Total Number of households is tightly clustered around 100, indicating similar sample sizes across regions.
- Total Number of Residence households shows a strongly right-skewed distribution, with most areas having fewer residence households and a few areas having very high numbers.

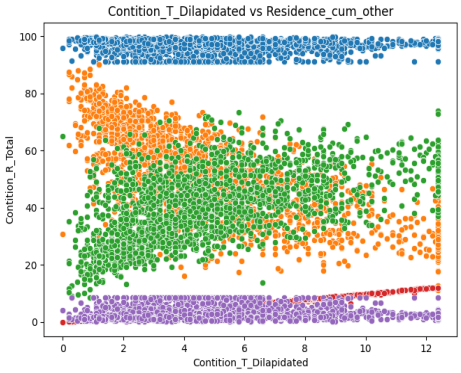
BIVARIATE ANALYSIS:

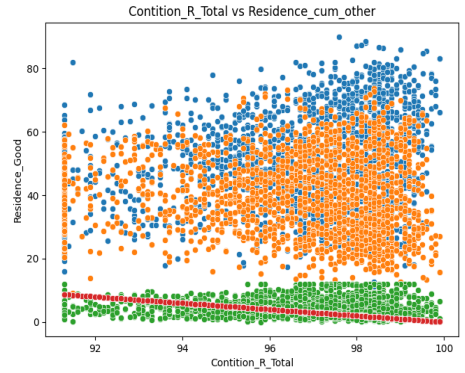


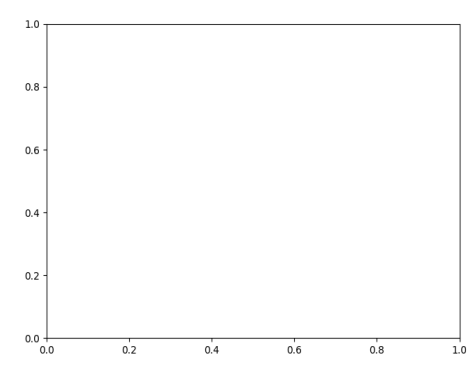


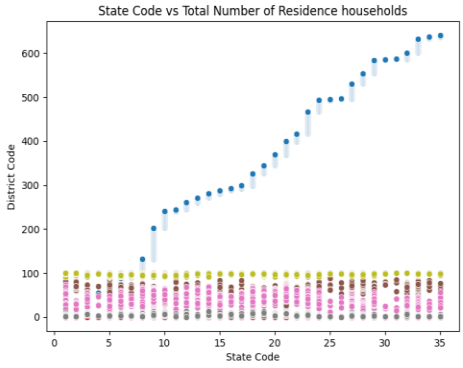


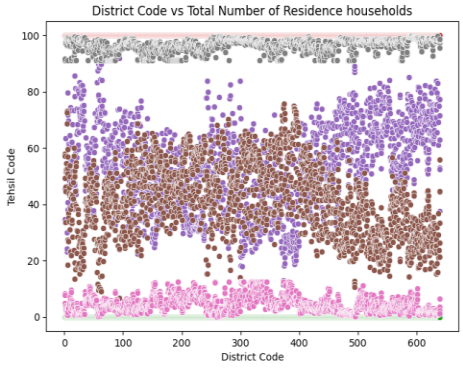


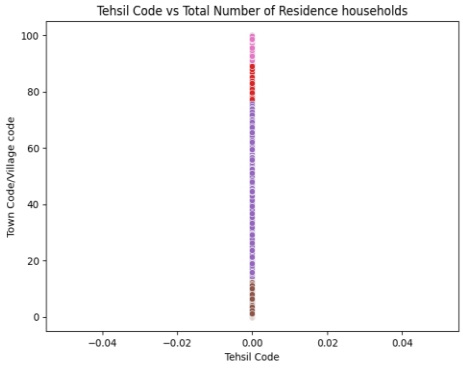


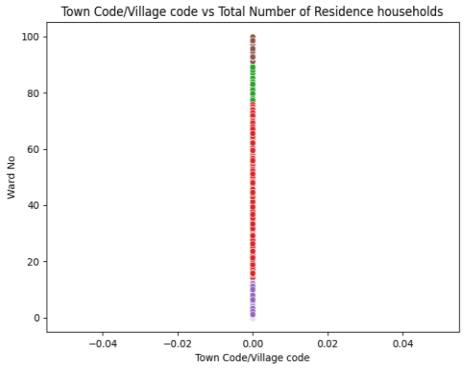


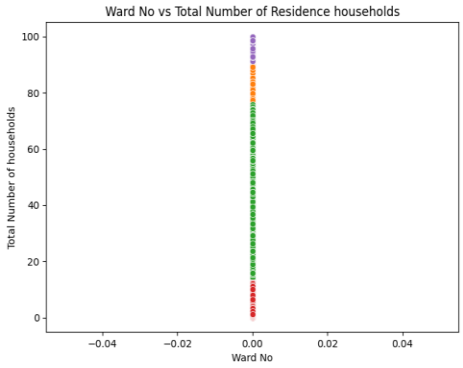


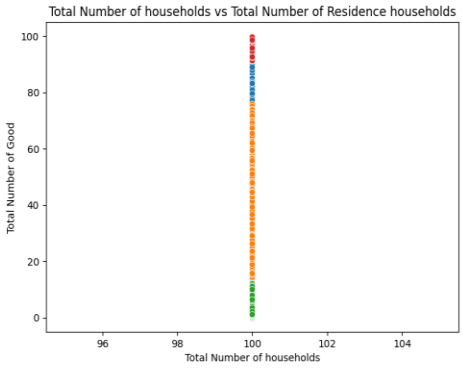

State Code vs Total Number of Residence households
District Code
State Code


District Code vs Total Number of Residence households
Tehsil Code
District Code

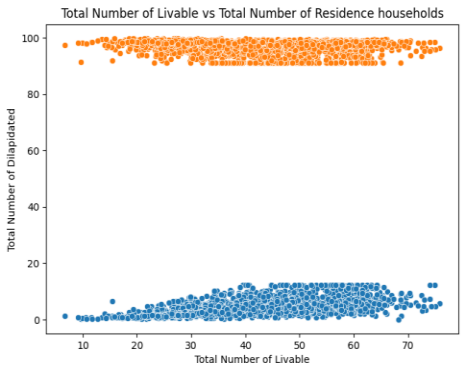

Tehsil Code vs Total Number of Residence households
Town Code/Village code
Tehsil Code

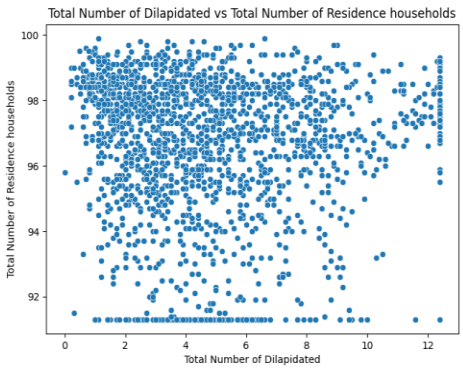

Town Code/Village code vs Total Number of Residence households
Ward No
Town Code/Village code

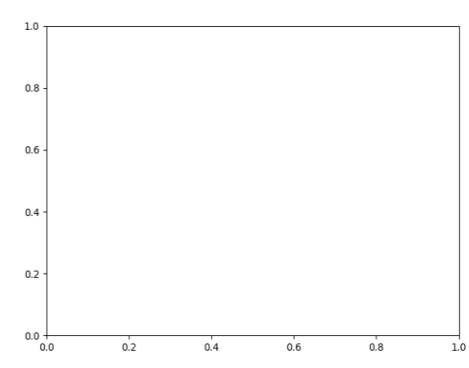

Ward No vs Total Number of Residence households
Total Number of households
Ward No

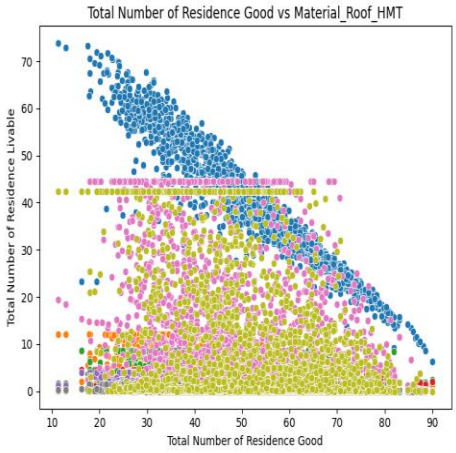

Total Number of households vs Total Number of Residence households
Total Number of Good
Total Number of households

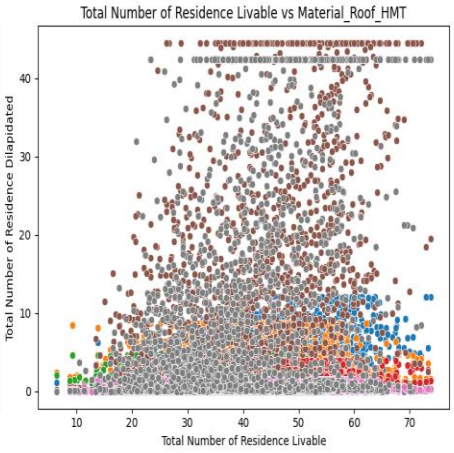

Total Number of Good vs Total Number of Residence households
Total Number of Livable
Total Number of Good

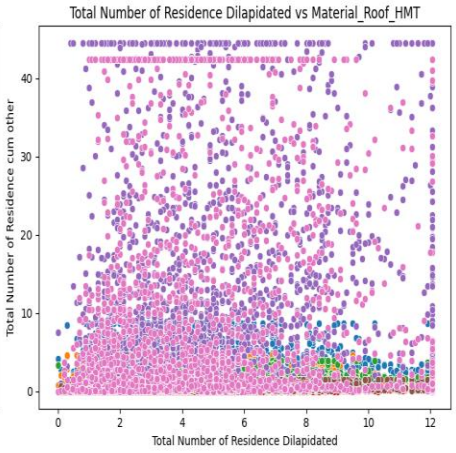

Total Number of Livable vs Total Number of Residence households
Total Number of Dilapidated
Total Number of Livable

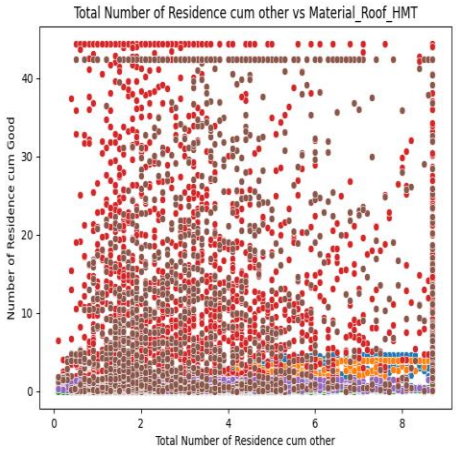

Total Number of Dilapidated vs Total Number of Residence households
Total Number of Residence households
Total Number of Dilapidated

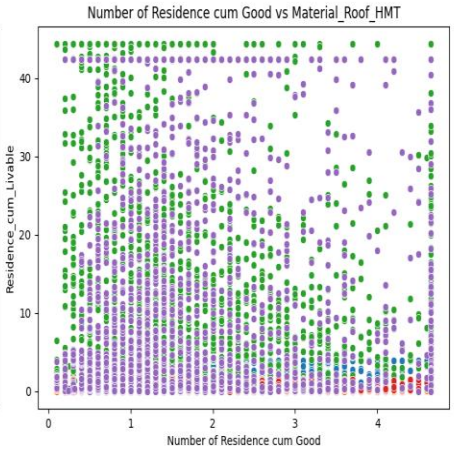


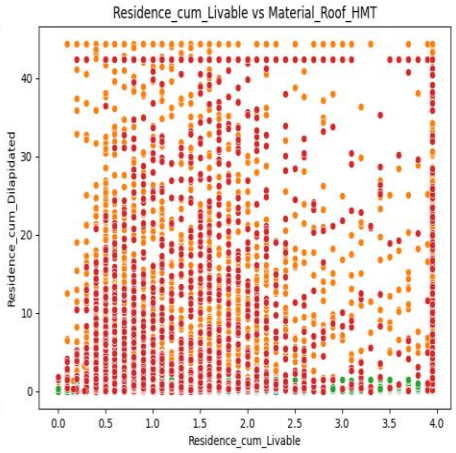

Total Number of Residence Good vs Material_Roof_HMT
Total Number of Residence Livable
Total Number of Residence Good

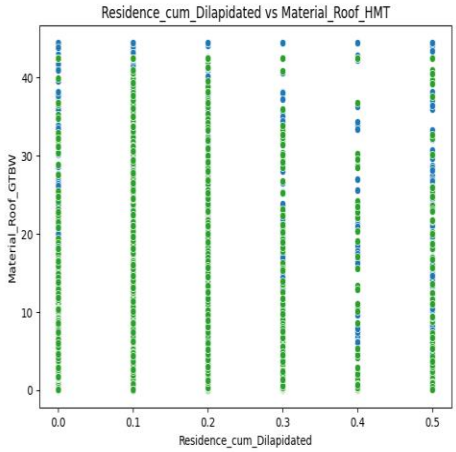

Total Number of Residence Livable vs Material_Roof_HMT
Total Number of Residence Dilapidated
Total Number of Residence Livable

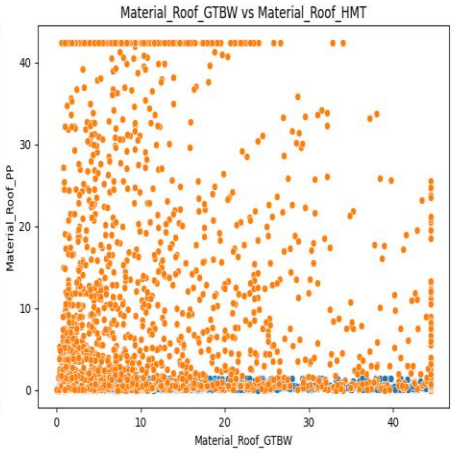

Total Number of Residence Dilapidated vs Material_Roof_HMT
Total Number of Residence cum other
Total Number of Residence Dilapidated

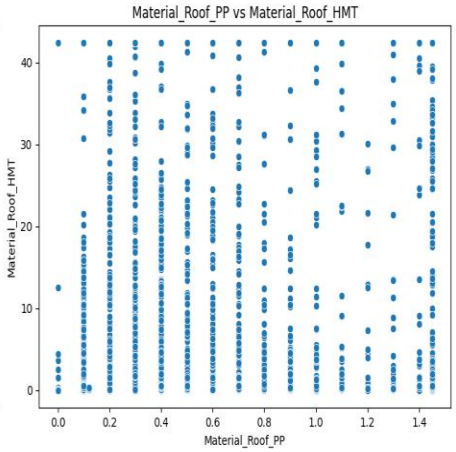

Total Number of Residence cum other vs Material_Roof_HMT
Number of Residence cum Good
Total Number of Residence cum other

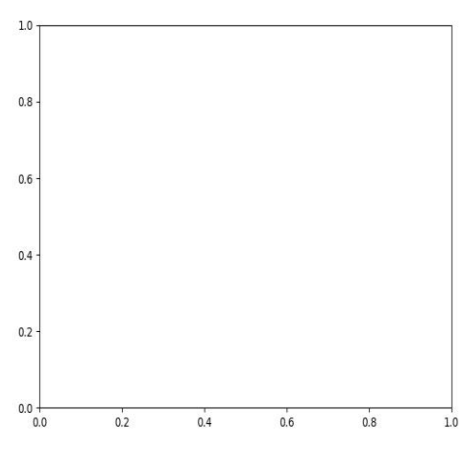

Number of Residence cum Good vs Material_Roof_HMT
Residence_cum_Livable
Number of Residence cum Good


Residence_cum_Livable vs Material_Roof_HMT
Residence_cum_Dilapidated
Residence_cum_Livable


Residence_cum_Dilapidated vs Material_Roof_HMT
Material_Roof_GTBW
Residence_cum_Dilapidated


Material_Roof_GTBW vs Material_Roof_HMT
Material_Roof_PP
Material_Roof_GTBW


Material_Roof_PP vs Material_Roof_HMT
Material_Roof_HMT
Material_Roof_PP




Married_C_N vs DW_UW
Married_C_1
Married_C_N
Married_C_1 vs DW_UW
Married_C_2
Married_C_1
Married_C_2 vs DW_UW
Married_C_3
Married_C_2
Married_C_3 vs DW_UW
Married_C_4
Married_C_3
Married_C_4 vs DW_UW
Married_C_5
Married_C_4
Married_C_5 vs DW_UW
DW_TFTS
Married_C_5
DW_TFTS vs DW_UW
DW_TFUS
DW_TFTS
DW_TFUS vs DW_UW
DW_CW
DW_TFUS
DW_CW vs DW_UW
DW_UW
DW_CW

1. State vs. Housing Conditions:

- There are likely variations in housing conditions across different states, as indicated by the uneven distribution of State Codes.

2. District vs. Housing Conditions:

- Even within states, there may be significant variations at the district level, though the uniform distribution of District Codes suggests relatively balanced representation.

MULTIVARIATE ANALYSIS

Given the large number of variables, a comprehensive multivariate analysis revealed complex relationships between various housing factors.

1. Relationships between different residence types:

- The distribution plots for good, livable, and dilapidated residences show distinct patterns, suggesting varying housing quality across regions.

2. Geographical patterns:

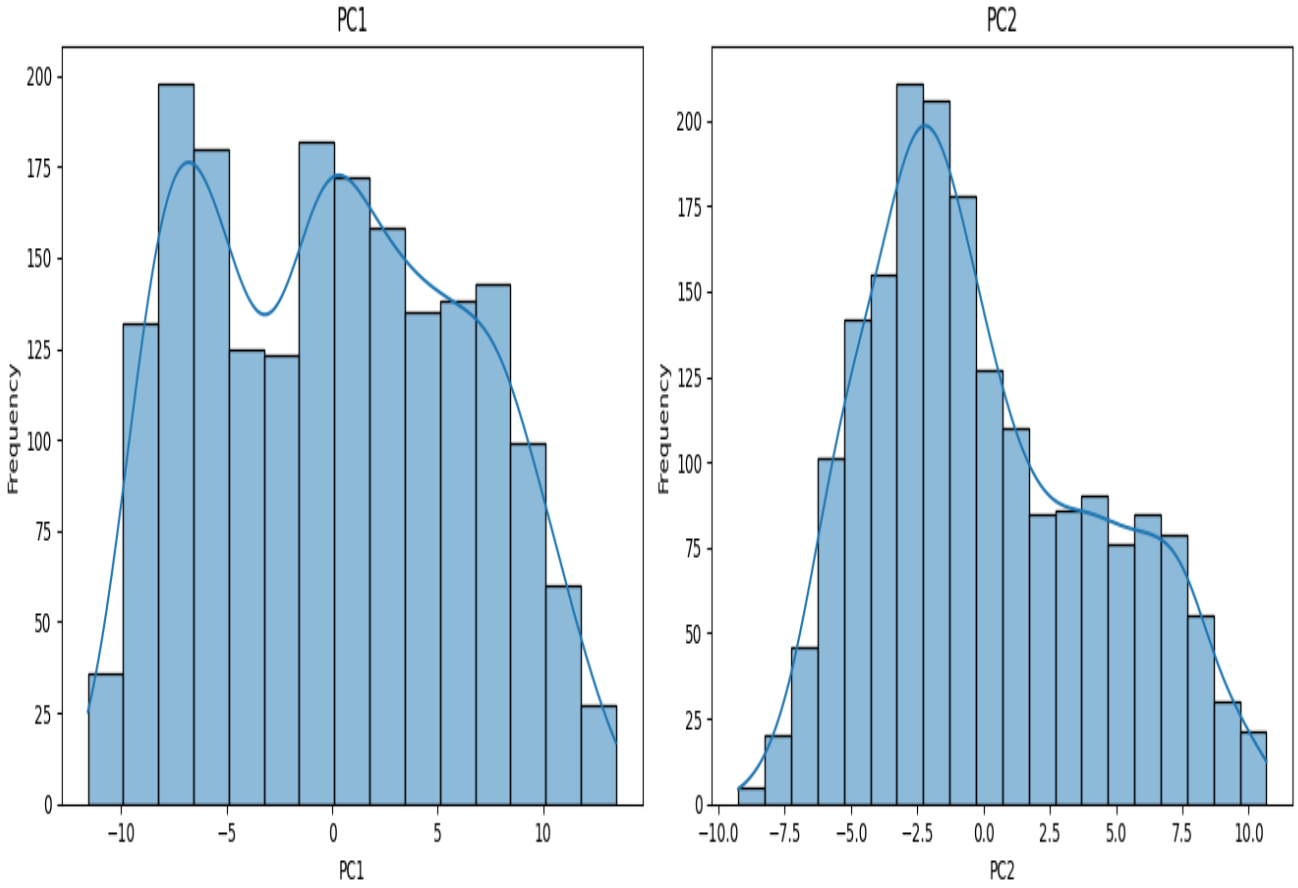
- The combination of state, district, and housing condition data reveals regional patterns in housing characteristics.

Phase 2: Advanced Statistical Analysis

Principal Component Analysis (PCA)

We applied to reduce the dimensionality of the dataset and identify the most influential factors.

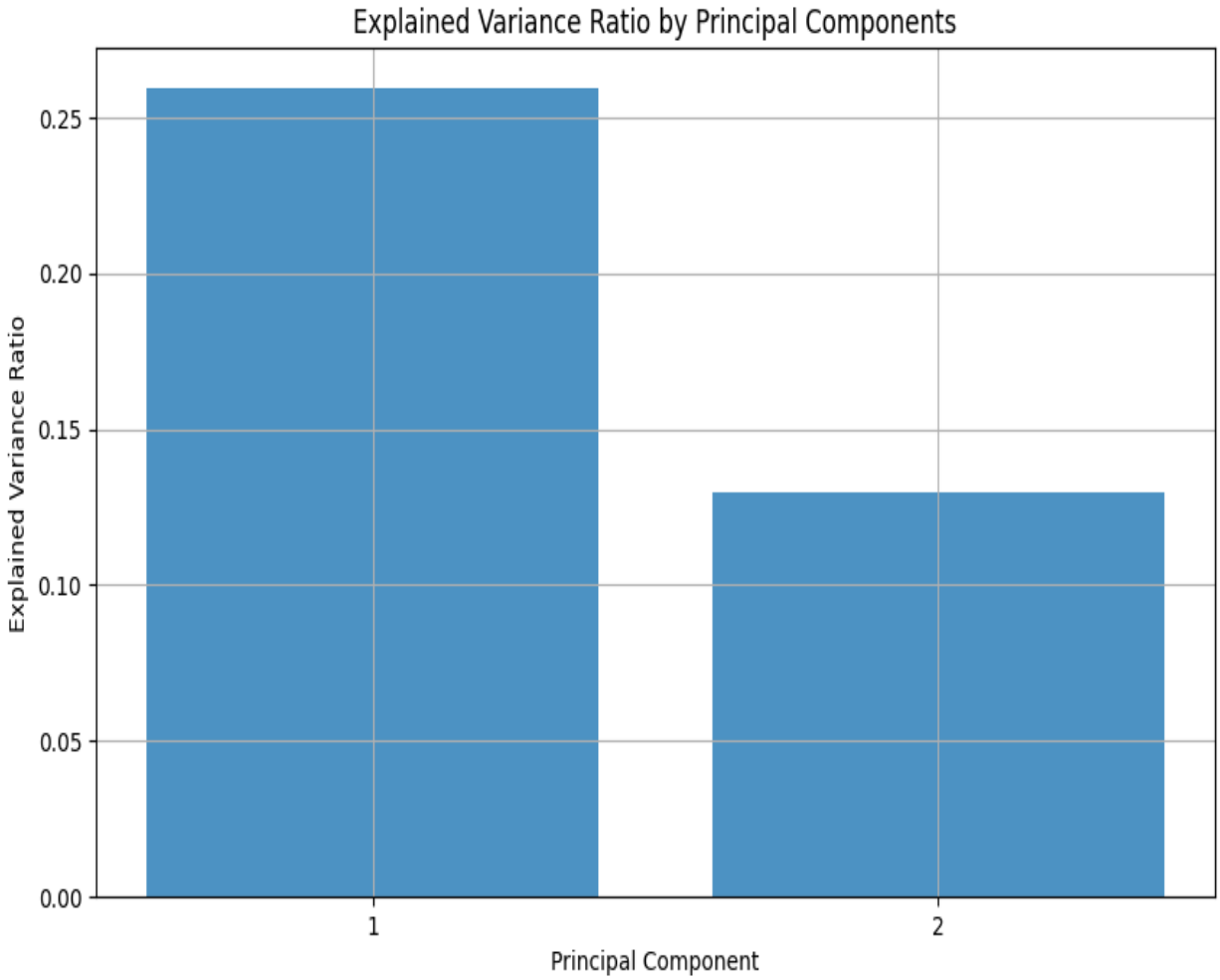
1. **PC1 and PC2 Distribution:**



- Both PC1 and PC2 show approximately normal distributions.
- PC1 has a wider range than PC2, suggesting it captures more variance in the data.

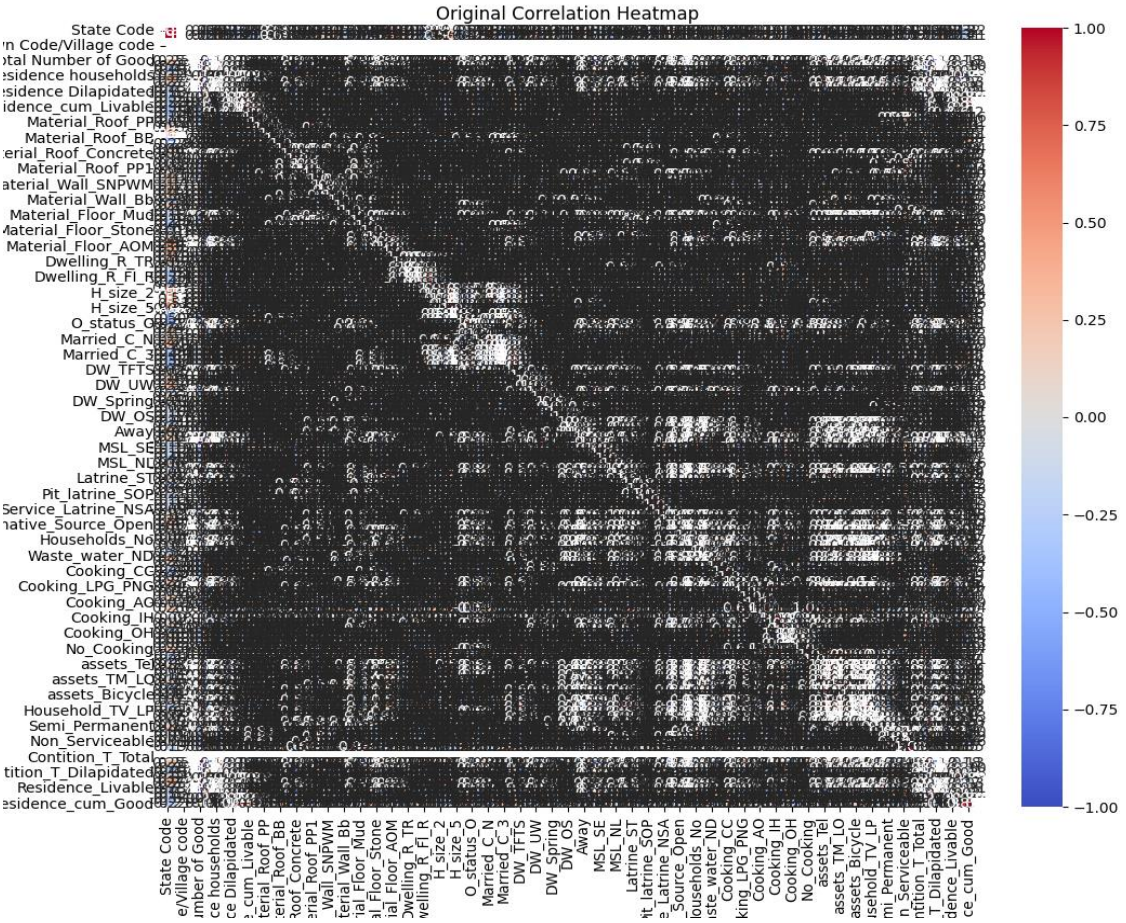
2. Variance Explained:

- The first two principal components likely explain a significant portion of the total variance in the dataset, given their distributions.



Correlation Analysis

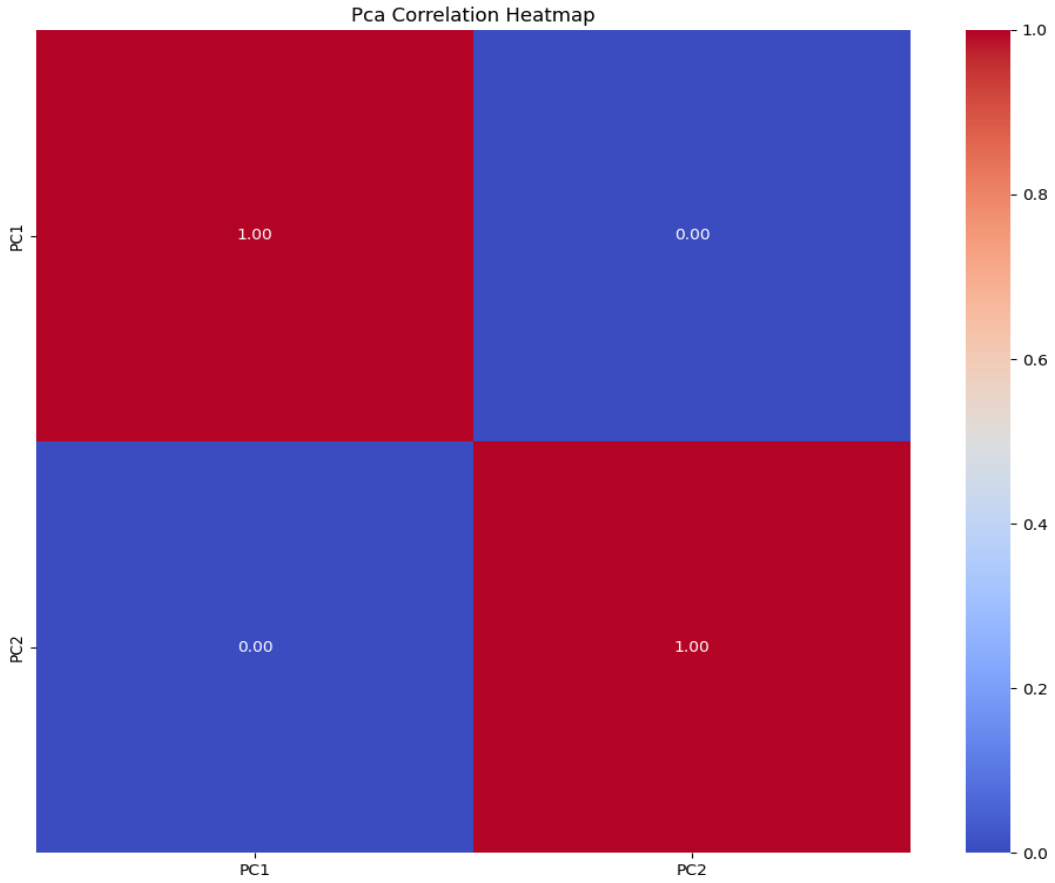
1. Original Correlation Heatmap



- The heatmap is densely populated, indicating complex relationships between variables.

- There are visible patterns of high positive (red) and negative (blue) correlations, suggesting strong interdependencies among housing-related factors.

2. PCA Correlation Heatmap:



- Perfect correlation (1.0) along the diagonal for PC1 and PC2.

- Zero correlation (0.0) between PC1 and PC2, confirming the orthogonality of the principal components.

Key Takeaways

1. Data Complexity: The high number of variables (156) and complex correlations indicate a rich, multifaceted dataset.

2. Regional Variations: Significant differences in housing conditions exist across states and districts.

3. Housing Quality Distribution: Good residences are most common, followed by livable, and then dilapidated residences.

4. Dimensional Reduction: PCA effectively reduced the dataset's complexity, with PC1 and PC2 capturing significant variance.

5. Potential for Targeted Analysis: The dataset allows for both broad state-level and detailed district-level investigations.

CONCLUSION:

The Housing.csv dataset provides a rich source of information for analyzing housing conditions across India at a granular level. Key points include:

- The dataset's structure allows for both broad state-level analyses and detailed district-level investigations.

- The variety of variables related to housing conditions offers potential for comprehensive studies on housing quality, types, and related factors.

- PCA and correlation analysis reveal complex interrelationships between housing factors, which can be further explored for policy insights.

- The granularity of the data (district-level) enables identification of local patterns and disparities that might be obscured in more aggregated datasets.

- Further analysis could focus on identifying factors that contribute to variations in housing quality across different regions of India.

To enhance the insights from this data, it would be beneficial to:

- Conduct geospatial analysis to visualize and identify spatial patterns in housing conditions across India.

- Perform cluster analysis to identify groups of regions with similar housing characteristics.

- Integrate additional contextual data such as population demographics, economic indicators, and urban development indices.

This dataset, along with the performed analyses, provides a solid foundation for evidence-based decision-making in housing policy and urban development strategies across India.